

Second Edition

ENCYCLOPEDIA OF ATMOSPHERIC SCIENCES

Edited by
Gerald R. North **Editor-in-Chief**
John Pyle and Fuqing Zhang



6 VOLUME SET

ENCYCLOPEDIA OF **ATMOSPHERIC SCIENCES**

SECOND EDITION

VOLUME 1

This page intentionally left blank

ENCYCLOPEDIA OF **ATMOSPHERIC SCIENCES**

SECOND EDITION

EDITOR-IN-CHIEF

GERALD R NORTH

Texas A&M University, College Station, TX, USA

EDITORS

JOHN PYLE

Cambridge University, Cambridge, UK

FUQING ZHANG

Pennsylvania State University, University Park, PA, USA

VOLUME 1



Amsterdam • Boston • Heidelberg • London • New York • Oxford
Paris • San Diego • San Francisco • Singapore • Sydney • Tokyo
Academic Press is an imprint of Elsevier



Academic Press is an imprint of Elsevier
32 Jamestown Road, London NW1 7BY, UK
225 Wyman Street, Waltham, MA 02451, USA
525 B Street, Suite 1800, San Diego, CA 92101-4495, USA

Copyright © 2015 Elsevier Ltd. unless otherwise stated. All rights reserved.

No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means electronic, mechanical, photocopying, recording or otherwise without the prior written permission of the publisher

Permissions may be sought directly from Elsevier's Science & Technology Rights Department in Oxford, UK: phone (+44) (0) 1865 843830; fax (+44) (0) 1865 853333; email: permissions@elsevier.com. Alternatively you can submit your request online by visiting the Elsevier web site at <http://elsevier.com/locate/permissions>, and selecting *Obtaining permission to use Elsevier material*

Notice

No responsibility is assumed by the publisher for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions or ideas contained in the material herein, Because of rapid advances in the medical sciences, in particular, independent verification of diagnoses and drug dosages should be made

British Library Cataloguing in Publication Data

A catalog record for this book is available from the British Library

Library of Congress Catalog Number:

A catalog record for this book is available from the Library of Congress

ISBN (print): 978-0-12-382225-3

For information on all Elsevier publications
visit our website at store.elsevier.com

Printed and bound in the United Kingdom

15 16 17 18 19 10 9 8 7 6 5 4 3 2 1

		Working together to grow libraries in developing countries
www.elsevier.com • www.bookaid.org		

<i>Acquisitions Editor:</i> Simon Holt <i>Project Manager:</i> Michael Nicholls <i>Associate Project Manager:</i> Marise Willis <i>Designer:</i> Matthew Limbert

DEDICATION

This second edition of the *Encyclopedia of Atmospheric Sciences* is dedicated to the memory of James Holton who was editor-in-chief of the first edition. He was a great researcher and colleague inspiring an entire generation of atmospheric scientists.

This page intentionally left blank

CONTENTS

<i>List of Contributors</i>	xxvii
<i>Preface to the First Edition</i>	xxxix
<i>Preface to the Second Edition</i>	xli
<i>Editor Biographies</i>	xlili
<i>Guide to Using the Encyclopedia</i>	xlvi

VOLUME 1

BASIC ATMOSPHERIC STRUCTURE AND CONCEPTS	1
Beaufort Wind Scale	1
<i>L Hasse</i>	
Wind Chill	7
<i>M Bluestein</i>	
Standard Atmosphere	12
<i>W W Vaughan</i>	
AEROSOLS	17
Aerosol–Cloud Interactions and Their Radiative Forcing	17
<i>U Lohmann</i>	
Aerosol Physics and Chemistry	23
<i>M Kalberer</i>	
Climatology of Stratospheric Aerosols	32
<i>L W Thomason and J-P Vernier</i>	
Climatology of Tropospheric Aerosols	40
<i>N Bellouin and J Haywood</i>	
Dust	48
<i>I N Sokolik</i>	
Observations and Measurements	53
<i>P H McMurry</i>	
Role in Radiative Transfer	66
<i>G A Ban-Weiss, and W D Collins</i>	

Role in Climate Change <i>N Bellouin</i>	76
Soot <i>P Chylek, S G Jennings, and R Pinnick</i>	86
Agricultural Meteorology and Climatology <i>E S Takle</i>	92
ARCTIC AND ANTARCTIC	98
Antarctic Climate <i>J Turner</i>	98
Arctic Climate <i>M C Serreze</i>	107
Arctic Haze <i>L M Russell and G E Shaw</i>	116
AIR SEA INTERACTIONS	
Freshwater Flux <i>J Schulz</i>	122
Momentum, Heat, and Vapor Fluxes <i>P K Taylor</i>	129
Sea Surface Temperature <i>W J Emery</i>	136
Surface Waves <i>A Benilov</i>	144
AVIATION METEOROLOGY	153
Aircraft Emissions <i>R R Friedl</i>	153
Aircraft Icing <i>M K Politovich</i>	160
Aviation Weather Hazards <i>A J Bedard, Jr</i>	166
Clear Air Turbulence <i>G P Ellrod (Retired), J A Knox, P F Lester, and L J Ehernberger (Retired)</i>	177
BIOGEOCHEMICAL CYCLES	187
Sulfur Cycle <i>P Brimblecombe</i>	187
Bromine <i>R von Glasow and C Hughes</i>	194
Heavy Metals <i>T D Jickells and A R Baker</i>	201

Iodine	205
<i>L J Carpenter</i>	
BOUNDARY LAYER (ATMOSPHERIC) AND AIR POLLUTION	220
Overview	220
<i>P J Mason and D J Thomson</i>	
Air Pollution Meteorology	227
<i>X-M Hu</i>	
Coherent Structures	237
<i>F T M Nieuwstadt and J C R Hunt</i>	
Complex Terrain	242
<i>J J Finnigan</i>	
Convective Boundary Layer	250
<i>M A LeMone</i>	
Microclimate	258
<i>M W Rotach and P Calanca</i>	
Modeling and Parameterization	265
<i>A A M Holtslag</i>	
Observational Techniques <i>In Situ</i>	274
<i>E F Bradley</i>	
Observational Techniques: Remote	284
<i>W M Angevine and C J Senff</i>	
Ocean Mixed Layer	290
<i>L Kantha and C A Clayson</i>	
Stably Stratified Boundary Layer	299
<i>L Mahrt</i>	
Surface Layer	305
<i>G L Geernaert</i>	
Urban Heat Islands	310
<i>J C Luvall, D A Quattrochi, D L Rickman, and M G Estes, Jr</i>	
Diurnal Cycle	319
<i>A Betts</i>	
CHEMISTRY OF THE ATMOSPHERE	324
Chemical Kinetics	324
<i>R P Wayne</i>	
Ion Chemistry	333
<i>J L Fox</i>	
Isotopes, Stable	348
<i>C A M Brenninkmeijer</i>	
Laboratory Kinetics	356
<i>D J Donaldson and S N Wren</i>	

Methane <i>E Dlugokencky, and S Houweling</i>	363
Observations for Chemistry (<i>In Situ</i>): Ozone Sonde <i>H G J Smit</i>	372
Observations for Chemistry (<i>In Situ</i>): Particles <i>T Deshler</i>	379
Observations for Chemistry (<i>In Situ</i>): Water Vapor Sonde <i>J B Smith</i>	387
Observations for Chemistry (Remote Sensing): IR/FIR (Satellite, Balloon and Ground) <i>H Fischer and F Hase</i>	401
Observations for Chemistry (Remote Sensing): Lidar <i>G Vaughan</i>	411
Observations for Chemistry (Remote Sensing): Microwave <i>J Waters</i>	418
Principles of Chemical Change <i>R P Wayne</i>	429
Radioactivity: Cosmogenic Radionuclides <i>D Lal</i>	437
Volcanoes: Composition of Emissions <i>M T Coffey and J W Hannigan</i>	446
Tracers <i>K A Boering</i>	450

VOLUME 2

CLIMATE AND CLIMATE CHANGE	1
Overview <i>D L Hartmann</i>	1
Carbon Dioxide <i>C L Sabine and R A Feely</i>	10
Climate Feedbacks <i>A E Dessler and M D Zelinka</i>	18
Climate Prediction: Empirical and Numerical <i>S Hastenrath</i>	26
Climate Variability: Decadal to Centennial Variability <i>D G Martinson</i>	33
Climate Variability: Nonlinear and Random Effects <i>M Ghil</i>	38
Climate Variability: North Atlantic and Arctic Oscillation <i>J W Hurrell</i>	47
Climate Variability: Seasonal and Interannual Variability <i>D S Gutzler</i>	61

Energy Balance Climate Models <i>G R North and K-Y Kim</i>	69
Global Impacts of the Madden–Julian Oscillation <i>C Zhang</i>	73
Greenhouse Effect <i>G R North</i>	80
History of Scientific Work on Climate Change <i>S Weart</i>	87
Intergovernmental Panel on Climate Change <i>K E Trenberth</i>	90
Nuclear Winter <i>A Robock</i>	95
Radiative–Convective Equilibrium Climate Models <i>N O Renno and X Huang</i>	102
Volcanoes: Role in Climate <i>A Robock</i>	105
CLOUDS AND FOG	112
Cloud Modeling <i>W-K Tao and M Moncrieff</i>	112
Contrails <i>P Minnis</i>	121
Cloud Microphysics <i>D Lamb</i>	133
Classification of Clouds <i>A L Rangno (Retiree)</i>	141
Climatology <i>S Warren, R Eastman, and C J Hahn</i>	161
Measurement Techniques <i>In situ</i> <i>D Baumgardner, J-F Gayet, A Korolev, C Twohy, and J Fugal</i>	170
Fog <i>P J Croft and B Ward</i>	180
Noctilucent Clouds <i>G E Thomas</i>	189
Stratus and Stratocumulus <i>R Wood</i>	196
CRYOSPHERE	201
Glaciers, Topography, and Climate <i>A B G Bush and M P Bishop</i>	201
Permafrost <i>T E Osterkamp and C R Burn</i>	208

Sea Ice	217
<i>M C Serreze, F Fetterer, and W F Weeks (Retired)</i>	
Snow (Surface)	227
<i>M Sturm</i>	
DATA ASSIMILATION AND PREDICTABILITY	237
Data Assimilation	237
<i>A C Lorenc</i>	
Ensemble-Based Data Assimilation	241
<i>Z Meng and F Zhang</i>	
Ensemble Prediction	248
<i>R Buizza</i>	
Predictability and Chaos	258
<i>L A Smith</i>	
DYNAMICAL METEOROLOGY	265
Overview	265
<i>J R Holton</i>	
Acoustic Waves	272
<i>K E Gilbert</i>	
Atmospheric Tides	287
<i>J Oberheide, M E Hagan, A D Richmond, and J M Forbes</i>	
Balanced Flow	298
<i>M E McIntyre</i>	
Baroclinic Instability	304
<i>R Grotjahn</i>	
Coriolis Force	313
<i>D W Moore</i>	
Critical Layers	317
<i>P Haynes</i>	
Hamiltonian Dynamics	324
<i>T G Shepherd</i>	
Hydraulic Flow	332
<i>R B Smith</i>	
Inertial Instability	334
<i>J A Knox</i>	
Kelvin–Helmholtz Instability	343
<i>P G Drazin</i>	
Kelvin Waves	347
<i>B Wang</i>	
Kinematics	353
<i>D D Houghton</i>	

Laboratory Geophysical Fluid Dynamics <i>R L Pfeffer</i>	360
Lagrangian Dynamics <i>I Roulstone</i>	369
Potential Vorticity <i>M E McIntyre</i>	375
Primitive Equations <i>A A White and N Wood</i>	384
Quasigeostrophic Theory <i>H C Davies and H Wernli</i>	393
Rossby Waves <i>P B Rhines</i>	404
Solitary Waves <i>J P Boyd</i>	417
Static Stability <i>J A Young</i>	423
Stationary Waves (Orographic and Thermally Forced) <i>S Nigam and E DeWeaver</i>	431
Symmetric Stability <i>H B Bluestein</i>	446
Vorticity <i>J R Holton</i>	451
Wave-CISK <i>C S Bretherton</i>	455
Wave Mean-Flow Interaction <i>M Juckes</i>	458
Waves <i>J R Holton</i>	464

VOLUME 3

ELECTRICITY IN THE ATMOSPHERE	1
Global Electrical Circuit <i>E R Williams</i>	1
Ions in the Atmosphere <i>K L Aplin and R G Harrison</i>	9
Lightning <i>M B Baker</i>	14
Sprites <i>W A Lyons</i>	20
Forensic Meteorology <i>L E Branscome</i>	28

GENERAL CIRCULATION OF THE ATMOSPHERE	33
Overview	33
<i>J M Wallace, D W J Thompson, and P Beresford</i>	
Angular Momentum of the Atmosphere	43
<i>D A Salstein</i>	
Energy Cycle	51
<i>R Grotjahn</i>	
Weather Regimes and Multiple Equilibria	65
<i>F Molteni</i>	
Mean Characteristics	73
<i>R Grotjahn</i>	
Teleconnections	90
<i>S Nigam and S Baxter</i>	
 GLOBAL CHANGE	 110
Climate Record: Surface Temperature Trends	110
<i>P D Jones</i>	
Sea Level Change	121
<i>R S Nerem</i>	
Upper Atmospheric Change	128
<i>R G Roble</i>	
Biospheric Impacts and Feedbacks	132
<i>B A Hungate and G W Koch</i>	
 GRAVITY WAVES	 141
Overview	141
<i>D C Fritts</i>	
Buoyancy and Buoyancy Waves: Optical Observations	153
<i>M J Taylor and W R Pendleton, Jr</i>	
Buoyancy and Buoyancy Waves: Theory	160
<i>T J Dunkerton</i>	
Gravity Waves Excited by Jets and Fronts	164
<i>R Plougonven and F Zhang</i>	
Convectively Generated Gravity Waves	171
<i>T P Lane</i>	
 HYDROLOGY, FLOODS AND DROUGHTS	 180
Overview	180
<i>R C Bales</i>	
Deserts and Desertification	185
<i>V P Tchakerian</i>	
Drought	193
<i>S Quiring</i>	

Flooding	201
<i>C A Doswell III</i>	
Groundwater and Surface Water	209
<i>S Ge and S M Gorelick</i>	
Modeling and Prediction	217
<i>Z Yu</i>	
Palmer Drought Severity Index	224
<i>L Nkemdirim</i>	
Soil Moisture	232
<i>A Robock</i>	
LAND-ATMOSPHERE INTERACTIONS	240
Overview	240
<i>R E Dickinson</i>	
Canopy Processes	244
<i>P D Blanken</i>	
Trace Gas Exchange	256
<i>J N Cape and D Fowler</i>	
LIDAR	262
Atmospheric Sounding Introduction	262
<i>P S Argall and R Sica</i>	
Backscatter	270
<i>C M R Platt and R L Collins</i>	
Differential Absorption Lidar	277
<i>S Ismail and E V Browell</i>	
Doppler	289
<i>R M Hardesty</i>	
Raman	296
<i>D N Whiteman</i>	
Resonance	305
<i>C S Gardner and R L Collins</i>	
Magnetosphere	309
<i>G K Parks</i>	
MESOSCALE METEOROLOGY	316
Overview	316
<i>D J Parker</i>	
Cloud and Precipitation Bands	323
<i>R M Rauber and M Ramamurthy</i>	
Gust Fronts	331
<i>R Rotunno</i>	

Hail and Hailstorms	334
<i>C Knight, N Knight, and H E Brooks</i>	
Mesoscale Convective Systems	339
<i>A Laing</i>	
Microbursts	335
<i>R M Wakimoto</i>	
Severe Storms	361
<i>C A Doswell III</i>	
Waterspouts	369
<i>J H Golden</i>	
Bow Echoes and Derecho	384
<i>M L Weisman</i>	
Density Currents	395
<i>P G Baines</i>	
Convective Storms: Overview	401
<i>M L Weisman</i>	
MESOSPHERE	411
Atomic Species in the Mesopause Region	411
<i>M G Mlynczak and L A Hunt</i>	
Ionosphere	422
<i>M C Kelley</i>	
Metal Layers	430
<i>J M C Plane</i>	
Polar Summer Mesopause	436
<i>R H Varney and M C Kelley</i>	
VOLUME 4	
MIDDLE ATMOSPHERE	1
Planetary Waves	1
<i>A K Smith and J Perlwitz</i>	
Polar Vortex	12
<i>M R Schoeberl and P A Newman</i>	
Quasi-Biennial Oscillation	18
<i>T J Dunkerton, J A Anstey, and L J Gray</i>	
Semiannual Oscillation	26
<i>K Hamilton</i>	
Stratospheric Sudden Warmings	30
<i>A O'Neill, A J Charlton-Perez, and L M Polvani</i>	
Transport Circulation	41
<i>S E Strahan</i>	

Zonal Mean Climatology <i>P Braesicke</i>	50
MOUNTAIN METEOROLOGY	57
Overview <i>R B Smith</i>	57
Cold Air Damming <i>B A Colle</i>	62
Downslope Winds <i>D R Durran</i>	69
Katabatic Winds <i>T R Parish</i>	75
Land and Sea Breezes <i>R A Pielke, Sr</i>	80
Lee Vortices <i>C C Epifanio</i>	84
Lee Waves and Mountain Waves <i>D R Durran</i>	95
Orographic Effects: Lee Cyclogenesis <i>C Schär</i>	103
Valley Winds <i>D Zardi</i>	114
NUMERICAL MODELS	135
Chemistry Models <i>M P Chipperfield and S R Arnold</i>	135
Coupled Ocean-Atmosphere Models: Physical Processes <i>M Zhang</i>	144
General Circulation Models <i>C R Mechoso and A Arakawa</i>	153
Methods <i>J Thuburn</i>	161
Model Physics Parameterization <i>D J Stensrud, M C Coniglio, K H Knopfmeier, and A J Clark</i>	167
Parameter Estimation <i>A Aksoy</i>	181
Parameterization of Physical Processes: Clouds <i>R Forbes, C Jakob, and M Miller</i>	187
Parameterization of Physical Processes: Gravity Wave Fluxes <i>M J Alexander</i>	194
Parameterization of Physical Processes: Turbulence and Mixing <i>A Beljaars</i>	200

Spectral Models <i>F Baer</i>	212
Mesoscale Atmospheric Modeling <i>R A Pielke, Sr</i>	219
Cloud-System Resolving Modeling and Aerosols <i>W-K Tao and T Matsui</i>	222
Large-Eddy Simulation <i>C-H Moeng and P P Sullivan</i>	232
Regional Prediction Models <i>B W Golding</i>	241
Convective Storm Modeling <i>M D Parker</i>	246
OBSERVATIONS PLATFORMS	255
Balloons <i>J-P Pommereau</i>	255
Buoys <i>J M Hemsley</i>	264
Kites <i>B B Balsley</i>	268
Radiosondes <i>W F Dabberdt and H Turtiainen</i>	273
Rockets <i>M F Larsen</i>	285
OCEANOGRAPHIC TOPICS	290
General Processes <i>N C Wells</i>	290
Surface/Wind Driven Circulation <i>R X Huang</i>	301
Thermohaline Circulation <i>R X Huang</i>	315
Water Types and Water Masses <i>W J Emery</i>	329
OPTICS, ATMOSPHERIC	338
Optical Remote Sensing Instruments <i>G G Shepherd</i>	338
Airglow Instrumentation <i>M Conde</i>	346

OZONE DEPLETION AND RELATED TOPICS	353
Long-Term Ozone Changes <i>N R P Harris</i>	353
Ozone as a UV Filter <i>J E Frederick</i>	359
Ozone Depletion Potentials <i>D J Wuebbles</i>	364
Photochemistry of Ozone <i>G K Moortgat and A R Ravishankara</i>	370
Stratospheric Ozone Recovery <i>D J Hofmann and R Müller</i>	380
Surface Ozone Effects on Vegetation <i>M Ashmore</i>	389
Surface Ozone (Human Health) <i>M Lippmann</i>	397
PALEOCLIMATOLOGY	404
Ice Cores <i>E J Steig</i>	404
Varves <i>R Gilbert</i>	411
RADAR	415
Cloud Radar <i>T Uttal</i>	415
Incoherent Scatter Radar <i>M P Sulzer</i>	422
Mesosphere—Stratosphere—Troposphere and Stratosphere—Troposphere Radars and Wind Profilers <i>G Vaughan and D Hooper</i>	429
Meteor Radar <i>N J Mitchell</i>	438
Polarimetric Doppler Weather Radar <i>R J Doviak and R D Palmer</i>	444
Precipitation Radar <i>S E Yuter</i>	455
Synthetic Aperture Radar (Land Surface Applications) <i>R K Vincent</i>	470
VOLUME 5	
RADIATION TRANSFER IN THE ATMOSPHERE	1
Radiation, Solar <i>Q Fu</i>	1

Absorption and Thermal Emission <i>R M Goody and X Huang</i>	5
Cloud-Radiative Processes <i>Q Fu</i>	13
Non-local Thermodynamic Equilibrium <i>M López-Puertas and B Funke</i>	16
Scattering <i>M Mishchenko, L Travis, and A Lacis</i>	27
Ultraviolet Radiation <i>K Stamnes</i>	37
Ultraviolet, Surface <i>R McKenzie and S Madronich</i>	45
SATELLITES AND SATELLITE REMOTE SENSING	51
Aerosol Measurements <i>R A Kahn</i>	51
Earth's Radiation Budget <i>N G Loeb and B A Wielicki</i>	67
GPS Meteorology <i>S S Leroy</i>	77
Measuring Ozone from Space — TOMS and SBUV <i>R D McPeters and R S Stolarski</i>	87
Orbits <i>S Q Kidder</i>	95
Precipitation <i>G Liu</i>	107
Remote Sensing: Cloud Properties <i>P Yang and B A Baum</i>	116
Research <i>M D King</i>	128
Surface Wind and Stress <i>W T Liu</i>	138
Temperature Soundings <i>A Dudhia</i>	145
Water Vapor <i>J E Harries</i>	157
SOLAR SYSTEM/SUN, ATMOSPHERES, EVOLUTION OF ATMOSPHERES	163
Evolution of Earth's Atmosphere <i>Y L Yung, M L Wong, and E J Gaidos</i>	163
Planetary Atmospheres: Mars <i>R M Haberle</i>	168

Planetary Atmospheres: Venus <i>P J Gierasch and Y L Yung</i>	178
Solar Terrestrial Interactions: Climate Impact <i>J D Haigh</i>	183
Solar Winds <i>S T Suess and B T Tsurutani</i>	189
Meteors <i>P Jenniskens</i>	195
STATISTICAL METHODS	201
Data Analysis: Empirical Orthogonal Functions and Singular Vectors <i>C S Bretherton</i>	201
Data Analysis: Time Series Analysis <i>G R North</i>	205
STRATOSPHERIC CHEMISTRY TOPICS	211
Overview <i>J A Pyle</i>	211
Halogens <i>D Toohey</i>	215
Halogen Sources, Anthropogenic <i>A McCulloch and P M Midgley</i>	221
Halogen Sources, Natural (Methyl Bromide and Related Gases) <i>S Yvon-Lewis and J H Butler</i>	228
HO _x <i>T F Hanisco</i>	233
Hydrogen Budget <i>J E Harries</i>	238
Reactive Nitrogen (NO _x and NO _y) <i>Y Kondo</i>	242
Stratospheric Water Vapor <i>K H Rosenlof</i>	250
STRATOSPHERE/TROPOSPHERE EXCHANGE AND STRUCTURE	257
Global Aspects <i>J R Holton</i>	257
Local Processes <i>J F Lamarque and P Hess</i>	262
Tropopause <i>M Dameris</i>	269

SYNOPTIC METEOROLOGY	273
Anticyclones <i>S J Colucci</i>	273
Forecasting <i>D Mansfield</i>	280
Weather Maps <i>R Reynolds</i>	289
Cyclogenesis <i>G J Hakim</i>	299
Extratropical Cyclones <i>A Joly</i>	304
Fronts <i>D M (David) Schultz and W Blumen</i>	337
Fronts in the Lower Stratosphere <i>A L Lang</i>	344
Frontogenesis <i>D M (David) Schultz</i>	353
Jet Streaks <i>P Cunningham and D Keyser</i>	359
Lake-Effect Storms <i>P J Sousounis</i>	370
Polar Lows <i>I A Renfrew</i>	379
Thermal Low <i>R H Johnson</i>	386
 THERMODYNAMICS	 391
Humidity Variables <i>J A Curry</i>	391
Moist (Unsaturated) Air <i>J A Curry</i>	394
Saturated Adiabatic Processes <i>J A Curry</i>	398
 Thermosphere <i>S C Solomon and R G Roble</i>	 402
 VOLUME 6	
TROPICAL CYCLONES AND HURRICANES	1
Overview and Theory <i>R A Tomas and P J Webster</i>	1

Hurricane Dynamics <i>Y Wang</i>	8
Hurricane Predictability <i>J A Sippel</i>	30
Hurricanes: Observation <i>F D Marks</i>	35
Tropical Cyclogenesis <i>Z Wang</i>	57
Tropical Cyclones and Climate Change <i>T R Knutson</i>	65
Tropical Cyclones in the Western North Pacific <i>J C L Chan</i>	77
Tropical Cyclones: Secondary Eyewall Formation <i>C-C Wu and Y-H Huang</i>	85
TROPICAL METEOROLOGY AND CLIMATE	91
El Niño and the Southern Oscillation: Observation <i>N Nicholls</i>	91
El Niño and the Southern Oscillation: Theory <i>P Chang and S E Zebiak</i>	97
Equatorial Waves <i>M C Wheeler and H Nguyen</i>	102
Hadley Circulation <i>J Lu and G A Vecchi</i>	113
Intertropical Convergence Zone <i>D E Waliser and X Jiang</i>	121
Intraseasonal Oscillation (Madden–Julian Oscillation) <i>R A Madden</i>	132
Madden–Julian Oscillation: Skeleton and Conceptual Models <i>A J Majda and S N Stechmann</i>	137
Monsoon: Overview <i>J Slingo</i>	146
Monsoon: Dynamical Theory <i>P J Webster and J Fasullo</i>	151
Monsoon: ENSO–Monsoon Interactions <i>K-M Lau</i>	165
Tropical Climates <i>S Hastenrath</i>	170
Walker Circulation <i>K-M Lau and S Yang</i>	177

TROPOSPHERIC CHEMISTRY AND COMPOSITION	182
Aerosols/Particles <i>J H Seinfeld</i>	182
Aliphatic Hydrocarbons <i>J Rudolph and O Stein</i>	188
Aromatic Hydrocarbons <i>I Barnes</i>	204
Biogenic Hydrocarbons <i>A Guenther</i>	214
Cloud Chemistry <i>P Herckes and J L Collett, Jr</i>	218
H ₂ <i>U Schmidt and T Wetter</i>	226
Hydroxyl Radical <i>K C Clemitshaw</i>	232
Mercury <i>J Munthe and J Sommar</i>	239
Oxidizing Capacity <i>D H Ehhalt, F Rohrer, and A Wahner</i>	243
Peroxyacetyl Nitrate <i>H B Singh</i>	251
Sulfur Chemistry, Organic <i>I Barnes</i>	255
Volatile Organic Compounds Overview: Anthropogenic <i>R G Derwent</i>	265
TURBULENCE AND MIXING	268
Overview <i>P Haynes</i>	268
Turbulence, Two Dimensional <i>P Bartello</i>	273
Turbulent Diffusion <i>A Venkatram and S Du</i>	277
WEATHER FORECASTING	287
Marine Meteorology <i>L Xie and B Liu</i>	287
Operational Meteorology <i>D R Novak</i>	293
Seasonal and Interannual Weather Prediction <i>J P Li and R Q Ding</i>	303
Severe Weather Forecasting <i>D J Stensrud, H E Brooks, and S J Weiss</i>	313

Wildfire Weather <i>J Coen</i>	323
Inadvertant Weather Modification <i>S A Changnon</i>	332
<i>Appendix 1: Physical Constants</i>	337
<i>Appendix 2: Units and their SI Equivalents</i>	339
<i>Appendix 3: Periodic Table of the Elements</i>	340
<i>Appendix 4: The Geologic Time Scale</i>	341
<i>Index</i>	343

This page intentionally left blank

LIST OF CONTRIBUTORS

A. Aksoy
*University of Miami, Miami, FL, USA; and NOAA
Hurricane Research Division, Miami, FL, USA*

M.J. Alexander
*NorthWest Research Associates (NWRA), Boulder,
CO, USA*

W.M. Angevine
*CIRES, University of Colorado; and NOAA Earth
System Research Laboratory, Boulder, CO, USA*

J.A. Anstey
University of Oxford, Oxford, UK

K.L. Aplin
University of Oxford, Oxford, UK

A. Arakawa
University of California, Los Angeles, CA, USA

P.S. Argall
*The University of Western Ontario, London, ON,
Canada*

S.R. Arnold
University of Leeds, Leeds, UK

M. Ashmore
University of York, York, UK

F. Baer
University of Maryland, College Park, MD, USA

P.G. Baines
University of Melbourne, Melbourne, VIC, Australia

A.R. Baker
University of East Anglia, Norwich, UK

M.B. Baker
University of Washington, Seattle, WA, USA

R.C. Bales
University of Arizona, Tucson, AZ, USA

B.B. Balsley
University of Colorado, Boulder, CO, USA

G.A. Ban-Weiss
*Lawrence Berkeley National Laboratory, Berkeley, CA,
USA; and University of Southern California,
Los Angeles, CA, USA*

I. Barnes
University of Wuppertal, Wuppertal, Germany

P. Bartello
McGill University, Montréal, QC, Canada

B.A. Baum
University of Wisconsin–Madison, Madison, WI, USA

D. Baumgardner
*Universidad Nacional Autónoma de México, Mexico
City, D.F., Mexico*

S. Baxter
University of Maryland, College Park, MD, USA

A.J. Bedard, Jr.
*National Oceanic and Atmospheric Administration,
Boulder, CO, USA*

A. Beljaars
*European Centre for Medium-Range Weather Forecasts,
Reading, England*

N. Bellouin
University of Reading, Reading, UK

A. Benilov
Acute Solutions, Highlands, NJ, USA

P. Beresford
*European Centre for Medium-Range Weather Forecasts,
Reading, UK*

A. Betts
Atmospheric Research, Pittsford, VT, USA

M.P. Bishop
Texas A&M University, College Station, TX, USA

P.D. Blanken
University of Colorado at Boulder, Boulder, CO, USA

H.B. Bluestein
University of Oklahoma, Norman, OK, USA

M. Bluestein
Indiana University – Purdue University, Indianapolis, IN, USA

W. Blumen[†]
University of Colorado Boulder, Boulder, CO, USA

K.A. Boering
University of California – Berkeley, Berkeley, CA, USA

J.P. Boyd
University of Michigan, Ann Arbor, MI, USA

E.F. Bradley
CSIRO Land and Water, Canberra, ACT, Australia

P. Braesicke
Karlsruhe Institute of Technology, Karlsruhe, Germany

L.E. Branscome
Climatological Consulting Corporation, FL, USA

C.A.M. Brenninkmeijer
Max Planck Institute for Chemistry, Mainz, Germany

C.S. Bretherton
University of Washington, Seattle, WA, USA

P. Brimblecombe
University of East Anglia, Norwich, UK

H.E. Brooks
National Oceanic and Atmospheric Administration, Norman, OK, USA

E.V. Browell
STARS II Affiliate, NASA Langley Research Center, Hampton, VA, USA

R. Buizza
ECMWF, Reading, UK

C.R. Burn
Carleton University, Ottawa, ON, Canada

A.B.G. Bush
University of Alberta, Edmonton, AB, Canada

J.H. Butler
NOAA Earth System Research Laboratory, Boulder, CO, USA

P. Calanca
Agroscope Reckenholz-Taenikon, Zurich, Switzerland

J.N. Cape
Edinburgh Research Station, Midlothian, UK

L.J. Carpenter
University of York, York, UK

J.C.L. Chan
City University of Hong Kong, Hong Kong

P. Chang
Texas A&M University, College Station, TX, USA

S.A. Changnon
University of Illinois, IL, USA

A.J. Charlton-Perez
University of Reading, Earley Gate, Reading, UK

M.P. Chipperfield
University of Leeds, Leeds, UK

P. Chylek
Dalhousie University, NS, Canada

A.J. Clark
University of Oklahoma and National Oceanic and Atmospheric Administration, Norman, OK, USA

C.A. Clayson
Woods Hole Oceanographic Institution, Woods Hole, MA, USA

K.C. Clemitshaw
Imperial College of Science, Technology, and Medicine, Ascot, UK

J. Coen
National Center for Atmospheric Research, Boulder, CO, USA

M.T. Coffey
National Center for Atmospheric Research, Boulder, CO, USA

B.A. Colle
Stony Brook University – SUNY, Stony Brook, NY, USA

J.L. Collett, Jr.
Colorado State University, Fort Collins, CO, USA

R.L. Collins
University of Alaska Fairbanks, Fairbanks, AK, USA

W.D. Collins
Lawrence Berkeley National Laboratory, Berkeley, CA, USA

S.J. Colucci
Cornell University, Ithaca, NY, USA

M. Conde
University of Alaska Fairbanks, Fairbanks, AK, USA

M.C. Coniglio
National Oceanic and Atmospheric Administration, Norman, OK, USA

[†] Deceased.

- P.J. Croft
Kean University, Union, NJ, USA
- P. Cunningham
Florida State University, Tallahassee, FL, USA
- J.A. Curry
Georgia Institute of Technology, Atlanta, GA, USA
- W.F. Dabberdt
Vaisala Company, Boulder, CO, USA
- M. Dameris
Deutsches Zentrum für Luft- und Raumfahrt, Institut für Physik der Atmosphäre, Oberpfaffenhofen, Wessling, Germany
- H.C. Davies
Institute for Atmospheric and Climate Science, Swiss Federal Institute of Technology (ETH), Zürich, Switzerland
- R.G. Derwent
rdscientific, Newbury, UK
- T. Deshler
University of Wyoming, Laramie, WY, USA
- A.E. Dessler
Texas A&M University, College Station, TX, USA
- E. DeWeaver
University of Wisconsin, Madison, WI, USA
- R.E. Dickinson
University of Texas at Austin, Austin, TX, USA
- R.Q. Ding
Institute of Atmospheric Physics, Chinese Academy of Sciences, Beijing, China
- E. Dlugokencky
NOAA Earth System Research Laboratory, Boulder, CO, USA
- D.J. Donaldson
University of Toronto, Toronto, ON, Canada
- C.A. Doswell, III
University of Oklahoma, Norman, OK, USA
- R.J. Doviak
National Severe Storms Laboratory, Norman, OK, USA
- P.G. Drazin[†]
University of Bath, England, UK
- S. Du
California Air Resources Board, Sacramento, CA, USA
- A. Dudhia
University of Oxford, Oxford, UK
- T.J. Dunkerton
Northwest Research Associates, Bellevue, WA, USA
- D.R. Durran
University of Washington, Seattle, WA, USA
- R. Eastman
University of Washington, Seattle, WA, USA
- L.J. Ehernberger
National Aeronautics and Space Administration, Dryden Flight Research Center, Edwards, CA, USA
- D.H. Ehhalt
Forschungszentrum Jülich, Jülich, Germany
- G.P. Ellrod
National Oceanographic and Atmospheric Administration/National Environmental Satellite, Data, and Information Service, Granby, CT, USA
- W.J. Emery
University of Colorado, Boulder, CO, USA
- C.C. Epifanio
Texas A&M University, College Station, TX, USA
- M.G. Estes
Universities Space Research Association, Huntsville, AL, USA
- J. Fasullo
University of Colorado – Boulder, Boulder, CO, USA
- R.A. Feely
NOAA Pacific Marine Environmental Laboratory, Seattle, WA, USA
- F. Fetterer
University of Colorado, Boulder, CO, USA
- J.J. Finnigan
CSIRO Atmospheric Research, Black Mountain, ACT, Australia
- H. Fischer
Institute of Meteorology and Climate Research (IMK), Karlsruhe Institute of Technology, Karlsruhe, Germany
- J.M. Forbes
University of Colorado, Boulder, CO, USA
- R. Forbes
European Centre for Medium-Range Weather Forecasts, Reading, UK
- D. Fowler
Edinburgh Research Station, Midlothian, UK

[†] Deceased.

J.L. Fox
Wright State University, Dayton, OH, USA

J.E. Frederick
The University of Chicago, Chicago, IL, USA

R.R. Friedl
California Institute of Technology, Pasadena, CA, USA

D.C. Fritts
GATS Inc., Boulder, CO, USA

Q. Fu
University of Washington, Seattle, WA, USA

J. Fugal
Max Planck Institute of Chemistry, Mainz, Germany

B. Funke
Instituto de Astrofísica de Andalucía, CSIC, Granada, Spain

E.J. Gaidos
University of Hawaii at Manoa, Honolulu, HI, USA

C.S. Gardner
University of Illinois at Urbana-Champaign, Urbana, IL, USA

J.-F. Gayet
Université Blaise Pascal, Clermont Ferrand, France

S. Ge
University of Colorado, Boulder, CO, USA

G.L. Geernaert
US Department of Energy, Washington, DC, USA

M. Ghil
Ecole Normale Supérieure, Paris, France; and University of California, Los Angeles, CA, USA

P.J. Gierasch
Cornell University, Ithaca, NY, USA

K.E. Gilbert
University of Mississippi, University, MS, USA

R. Gilbert
Queen's University, Kingston, ON, Canada

J.H. Golden
Forecast Systems Laboratory, NOAA, Boulder, CO, USA

B.W. Golding
Met Office, Exeter, UK

R.M. Goody
Harvard University (Emeritus), Cambridge, MA, USA

S.M. Gorelick
Stanford University, Stanford, CA, USA

L.J. Gray
University of Oxford, Oxford, UK

R. Grotjahn
University of California, Davis, CA, USA

A. Guenther
Pacific Northwest National Laboratory, Richland, WA, USA

D.S. Gutzler
University of New Mexico, Albuquerque, NM, USA

R.M. Haberle
NASA/Ames Research Center, Moffett Field, Mountain View, CA, USA

M.E. Hagan
National Center for Atmospheric Research, Boulder, CO, USA

C.J. Hahn
University of Arizona, Tucson, AZ, USA

J.D. Haigh
Blackett Laboratory, Imperial College London, London, UK

G.J. Hakim
University of Washington, Seattle, WA, USA

K. Hamilton
University of Hawaii, Honolulu, HI, USA

T.F. Hanisco
Harvard University, Cambridge, MA, USA

J.W. Hannigan
National Center for Atmospheric Research, Boulder, CO, USA

R.M. Hardesty
NOAA Environmental Technology Laboratory, Boulder, CO, USA

J.E. Harries
Imperial College of Science, Technology and Medicine, Blackett Laboratory, London, UK

N.R.P. Harris
University of Cambridge, Cambridge, UK

R.G. Harrison
The University of Reading, Reading, UK

D.L. Hartmann
University of Washington, Seattle, WA, USA

F. Hase
Institute of Meteorology and Climate Research (IMK), Karlsruhe Institute of Technology, Karlsruhe, Germany

- L. Hasse
Universität Kiel, Kiel, Germany
- S. Hastenrath
University of Wisconsin, Madison, WI, USA
- P. Haynes
University of Cambridge, Cambridge, UK
- J. Haywood
Met Office, Exeter, UK
- J.M. Hemsley
National Data Buoy Center, Stennis Space Center, MS, USA
- P. Herckes
Arizona State University, Tempe, AZ, USA
- P. Hess
National Center for Atmospheric Research, Boulder, CO, USA
- D.J. Hofmann[†]
NOAA Climate Monitoring and Diagnostics Laboratory, Boulder, CO, USA
- J.R. Holton
University of Washington, Seattle, WA, USA
- A.A.M. Holtslag
Wageningen University, Wageningen, The Netherlands
- D. Hooper
Science & Technology Facilities Council (STFC), Didcot, UK
- D.D. Houghton
University of Wisconsin-Madison, Madison, WI, USA
- S. Houweling
SRON Netherlands Institute for Space Research, Utrecht, The Netherlands
- X.-M. Hu
University of Oklahoma, Norman, OK, USA
- R.X. Huang
Woods Hole Oceanographic Institution, Woods Hole, MA, USA
- X. Huang
University of Michigan, Ann Arbor, MI, USA
- Y.-H. Huang
National Taiwan University, Taipei, Taiwan
- C. Hughes
University of York, York, UK
- B.A. Hungate
Northern Arizona University, Flagstaff, AZ, USA
- J.C.R. Hunt
University College London, London, UK
- L.A. Hunt
Science Systems and Applications Incorporated, Hampton, VA, USA
- J.W. Hurrell
National Center for Atmospheric Research, Boulder, CO, USA
- S. Ismail
Science Directorate, NASA Langley Research Center, Hampton, VA, USA
- C. Jakob
Monash University, VIC, Australia
- S.G. Jennings
National University of Ireland, Galway, Ireland
- P. Jenniskens
SETI Institute, Moffett Field, CA, USA
- X. Jiang
University of California, Los Angeles, CA, USA
- T.D. Jickells
University of East Anglia, Norwich, UK
- R.H. Johnson
Colorado State University, Fort Collins, CO, USA
- A. Joly
Centre National de Recherches Météorologiques – Groupe d'étude de l'Atmosphère Météorologique, Météo-France and CNRS, Toulouse, France
- P.D. Jones
Climatic Research Unit, University of East Anglia, Norwich, UK
- M. Jukes
University of Oxford, Oxford, UK
- R.A. Kahn
NASA Goddard Space Flight Center, Greenbelt, MD, USA
- M. Kalberer
University of Cambridge, Cambridge, UK
- L. Kantha
University of Colorado, Boulder, CO, USA
- M.C. Kelley
Cornell University, Ithaca, NY, USA

[†] Deceased.

D. Keyser
*University at Albany, State University of New York,
Albany, NY, USA*

S.Q. Kidder
*Colorado State University, Fort Collins,
CO, USA*

K.-Y. Kim
Seoul National University, Seoul, Korea

M.D. King
University of Colorado, Boulder, CO, USA

C. Knight
*National Center for Atmospheric Research, Boulder,
CO, USA*

N. Knight
*National Center for Atmospheric Research, Boulder,
CO, USA*

K.H. Knopfmeier
*University of Oklahoma; and National Oceanic and
Atmospheric Administration, Norman, OK, USA*

J.A. Knox
University of Georgia, Athens, GA, USA

T.R. Knutson
*NOAA Geophysical Fluid Dynamics Laboratory,
Princeton, NJ, USA*

G.W. Koch
Northern Arizona University, Flagstaff, AZ, USA

Y. Kondo
The University of Tokyo, Tokyo, Japan

A. Korolev
*Meteorological Service of Canada, Toronto, ON,
Canada*

A. Lacis
*Goddard Institute for Space Studies, New York,
NY, USA*

A. Laing
*National Center for Atmospheric Research, Boulder,
CO, USA*

D. Lal
*Scripps Institution of Oceanography, La Jolla,
CA, USA*

J.F. Lamarque
*National Center for Atmospheric Research, Boulder,
CO, USA*

D. Lamb
*The Pennsylvania State University, University Park,
PA, USA*

T.P. Lane
*The University of Melbourne, Melbourne, VIC,
Australia*

A.L. Lang
*University of Albany – State University of New York,
Albany, NY, USA*

M.F. Larsen
Clemson University, Clemson, SC, USA

K.-M. Lau
*NASA/Goddard Space Flight Center, Greenbelt, MD,
USA*

M.A. LeMone
*National Center for Atmospheric Research, Boulder,
CO, USA*

S.S. Leroy
*Harvard School of Engineering and Applied Sciences,
Cambridge, MA, USA*

P.F. Lester
San Jose State University, San Jose, CA, USA

J.P. Li
*Institute of Atmospheric Physics, Chinese Academy of
Sciences, Beijing, China*

M. Lippmann
New York University, Tuxedo, NY, USA

B. Liu
North Carolina State University, Raleigh, NC, USA

G. Liu
Florida State University, Tallahassee, FL, USA

W.T. Liu
*California Institute of Technology, Pasadena,
CA, USA*

N.G. Loeb
NASA Langley Research Center, Hampton, VA, USA

U. Lohmann
ETH Zurich, Zürich, Switzerland

M. López-Puertas
*Instituto de Astrofísica de Andalucía, CSIC, Granada,
Spain*

A.C. Lorenc
The Met Office, Bracknell, Berkshire, UK

J. Lu
*Pacific Northwest National Laboratory, Richland, WA,
USA*

J.C. Luvall
*National Space Science and Technology Center,
Marshall Space Flight Center, Huntsville, AL, USA*

- W.A. Lyons
FMA Research Inc., Fort Collins, CO, USA
- R.A. Madden
National Center for Atmospheric Research, Boulder, CO, USA
- S. Madronich
National Center for Atmospheric Research, Boulder, CO, USA
- L. Mahrt
Oregon State University, Corvallis, OR, USA
- A.J. Majda
New York University, New York, NY, USA
- D. Mansfield
National Meteorological Center, Bracknell, UK
- F.D. Marks
Hurricane Research Division, Miami, FL, USA
- D.G. Martinson
Columbia University, Palisades, NY, USA
- P.J. Mason
Met Office, Bracknell, UK
- T. Matsui
NASA/Goddard Space Flight Center, Greenbelt, MD, USA; and University of Maryland, College Park, MD, USA
- A. McCulloch
University of Bristol, Bristol, UK
- M.E. McIntyre
University of Cambridge, Cambridge, UK
- R. McKenzie
National Institute for Water and Atmospheric Research, Lauder, Central Otago, New Zealand
- P.H. McMurtry
University of Minnesota, Minneapolis, MN, USA
- R.D. McPeters
NASA Goddard Space Flight Center, Greenbelt, MD, USA
- C.R. Mechoso
University of California, Los Angeles, CA, USA
- Z. Meng
Peking University, Beijing, China
- P.M. Midgley
M & D Consulting, Leinfelden Metternich, Germany
- M. Miller
European Centre for Medium-Range Weather Forecasts, Reading, UK
- P. Minnis
Science Directorate, NASA Langley Research Center, Hampton, VA, USA
- M. Mishchenko
Goddard Institute for Space Studies, New York, NY, USA
- N.J. Mitchell
The University of Bath, Bath, UK
- M.G. Mlynarczyk
NASA Langley Research Center, Hampton, VA, USA
- C.-H. Moeng
National Center for Atmospheric Research, Boulder, CO, USA
- F. Molteni
Abdus Salam International Centre for Theoretical Physics, Trieste, Italy
- M. Moncrieff
National Center for Atmospheric Research, Boulder, CO, USA
- D.W. Moore
Pacific Marine Environmental Laboratory, Seattle, WA, USA
- G.K. Moortgat
Max-Planck-Institute for Chemistry, Mainz, Germany
- R. Müller
Institute for Energy and Climate Research (IEK-7), Forschungszentrum Jülich, Jülich, Germany
- J. Munthe
IVL Swedish Environmental Research Institute, Göteborg, Sweden
- R.S. Nerem
University of Colorado, Boulder, CO, USA
- P.A. Newman
NASA Goddard, Space Flight Center, Greenbelt, MD, USA
- H. Nguyen
Centre for Australian Weather and Climate Research, Bureau of Meteorology, Melbourne, VIC, Australia
- N. Nicholls
Bureau of Meteorology Research Centre, Melbourne, VIC, Australia
- F.T.M. Nieuwstadt
Delft University of Technology, Delft, The Netherlands
- S. Nigam
University of Maryland, College Park, MD, USA

L. Nkemdirim
University of Calgary, Calgary, AB, Canada

G.R. North
Texas A&M University, College Station, TX, USA

D.R. Novak
Weather Prediction Center, College Park, MD, USA

A. O'Neill
University of Reading, Earley Gate, Reading, UK

J. Oberheide
Clemson University, Clemson, SC, USA

T.E. Osterkamp
University of Alaska, Fairbanks, AK, USA

R.D. Palmer
University of Oklahoma, Oklahoma, OK, USA

T.R. Parish
University of Wyoming, Laramie, WY, USA

D.J. Parker
University of Leeds, Leeds, UK

M.D. Parker
North Carolina State University, Raleigh, NC, USA

G.K. Parks
University of Washington, Seattle, WA, USA

W.R. Pendleton
Utah State University, Logan, UT, USA

J. Perlwitz
University of Colorado, Boulder, CO, USA

R.L. Pfeffer
Florida State University, Tallahassee, FL, USA

R.A. Pielke, Sr.
University of Colorado at Boulder, CO, USA

R. Pinnick
US Army Research Laboratory, Adelphi, MD, USA

J.M.C. Plane
University of Leeds, Leeds, UK

C.M.R. Platt
Colorado State University, Fort Collins, CO, USA

R. Plougonven
Ecole Polytechnique, Palaiseau, France

M.K. Politovich
National Center for Atmospheric Research, Boulder, CO, USA

L.M. Polvani
Columbia University, New York, NY, USA

J.-P. Pommereau
LATMOS, CNRS, Guyancourt, France

J.A. Pyle
University of Cambridge, Cambridge, UK

D.A. Quattrochi
National Space Science and Technology Center, Marshall Space Flight Center, Huntsville, AL, USA

S. Quiring
Texas A&M University, College Station, TX, USA

M. Ramamurthy
University Corporation for Atmospheric Research, Boulder, CO, USA

A.L. Rangno (Retiree)
University of Washington, Seattle, WA, USA

R.M. Rauber
University of Illinois at Urbana-Champaign, Urbana, IL, USA

A.R. Ravishankara
Colorado State University, Fort Collins, CO, USA

I.A. Renfrew
University of East Anglia, Norwich, UK

N.O. Renno
University of Michigan, Ann Arbor, MI, USA

R. Reynolds
University of Reading, Reading, UK

P.B. Rhines
University of Washington, Seattle, WA, USA

A.D. Richmond
National Center for Atmospheric Research, Boulder, CO, USA

D.L. Rickman
National Space Science and Technology Center, Marshall Space Flight Center, Huntsville, AL, USA

R.G. Roble
National Center for Atmospheric Research, Boulder, CO, USA

A. Robock
Rutgers University, New Brunswick, NJ, USA

F. Rohrer
Forschungszentrum Jülich, Jülich, Germany

K.H. Rosenlof
Earth System Research Laboratory, National Oceanic and Atmospheric Administration, Boulder, CO, USA

- M.W. Rotach
University of Innsbruck, Innsbruck, Austria
- R. Rotunno
National Center for Atmospheric Research, Boulder, CO, USA
- I. Roulstone
University of Surrey, Guildford, UK
- J. Rudolph
York University, Toronto, ON, Canada
- L.M. Russell
Scripps Institution of Oceanography, University of California, San Diego, La Jolla, CA, USA
- C.L. Sabine
NOAA Pacific Marine Environmental Laboratory, Seattle, WA, USA
- D.A. Salstein
Atmospheric and Environmental Research, Inc., Lexington, MA, USA
- C. Schär
Atmospheric and Climatic Science ETH, Zürich, Switzerland
- U. Schmidt
Johann Wolfgang Goethe-University, Frankfurt am Main, Institut für Meteorologie und Geophysik, Frankfurt am Main, Germany
- M.R. Schoeberl
Science and Technology Corporation, Lanham, MD, USA
- D.M. (David) Schultz
University of Manchester, Manchester, UK
- J. Schulz
Meteorological Institute, University of Bonn, Bonn, Germany
- J.H. Seinfeld
California Institute of Technology, Pasadena, CA, USA
- C.J. Senff
CIRES, University of Colorado; and NOAA Earth System Research Laboratory, Boulder, CO, USA
- M.C. Serreze
University of Colorado, Boulder, CO, USA
- G.E. Shaw
Geophysical Institute, University of Alaska, Fairbanks, AK, USA
- G.G. Shepherd
York University, Toronto, ON, Canada
- T.G. Shepherd
University of Toronto, Toronto, ON, Canada
- R. Sica
The University of Western Ontario, London, ON, Canada
- H.B. Singh
NASA Ames Research Center, Mountain View, CA, USA
- J.A. Sippel
National Aeronautics and Space Administration (NASA), Greenbelt, MD, USA
- J. Slingo
University of Reading, Reading, UK
- H.G.J. Smit
Research Centre Jülich, Jülich, Germany
- A.K. Smith
National Center for Atmospheric Research, Boulder, CO, USA
- J.B. Smith
Harvard University, Cambridge, MA, USA
- L.A. Smith
London School of Economics, Centre for the Analysis of Time Series, London, UK
- R.B. Smith
Yale University, New Haven, CT, USA
- I.N. Sokolik
Georgia Institute of Technology, Atlanta, GA, USA
- S.C. Solomon
National Center for Atmospheric Research, Boulder, CO, USA
- J. Sommar
Göteborg University, Göteborg, Sweden
- P.J. Sousounis
AIR Worldwide, Boston, MA, USA
- K. Stamnes
Stevens Institute of Technology, Hoboken, NJ, USA
- S.N. Stechmann
University of Wisconsin–Madison, Madison, WI, USA
- E.J. Steig
University of Washington, Seattle, WA, USA
- O. Stein
IEK 8: Troposphere, Research Center Juelich, Juelich, Germany

D.J. Stensrud
*National Oceanic and Atmospheric Administration,
Norman, OK, USA*

R.S. Stolarski
Johns Hopkins University, Baltimore, MD, USA

S.E. Strahan
*Universities Space Research Association, NASA Goddard
Space Flight Center, Greenbelt, MD, USA*

M. Sturm
*US Army Cold Regions Research & Engineering
Laboratory-Alaska, Fort Wainwright, AL, USA*

S.T. Suess
*NASA Marshall Space Flight Center, Huntsville, AL,
USA*

P.P. Sullivan
*National Center for Atmospheric Research, Boulder,
CO, USA*

M.P. Sulzer
Arecibo Observatory, Arecibo, PR, USA

E.S. Takle
Iowa State University, Ames, IA, USA

W.-K. Tao
*NASA/Goddard Space Flight Center, Greenbelt, MD,
USA*

M.J. Taylor
Utah State University, Logan, UT, USA

P.K. Taylor
*Southampton Oceanography Centre, Southampton,
UK*

V.P. Tchakerian
Texas A&M University, College Station, TX, USA

G.E. Thomas
University of Colorado, Boulder, CO, USA

L.W. Thomason
NASA Langley Research Center, Hampton, VA, USA

D.W.J. Thompson
Colorado State University, Fort Collins, CO, USA

D.J. Thomson
Met Office, Bracknell, UK

J. Thuburn
University of Exeter, Exeter, UK

R.A. Tomas
University of Colorado – Boulder, Boulder, CO, USA

D. Toohey
University of Colorado Boulder, Boulder, CO, USA

L. Travis
*Goddard Institute for Space Studies, New York,
NY, USA*

K.E. Trenberth
*National Center for Atmospheric Research, Boulder,
CO, USA*

B.T. Tsurutani
Jet Propulsion Laboratory, Pasadena, CA, USA

J. Turner
British Antarctic Survey, Cambridge, UK

H. Turtiainen
Vaisala Company, Helsinki, Finland

C. Twohy
Oregon State University, Corvallis, OR, USA

T. Uttal
NOAA, Boulder, CO, USA

R.H. Varney
Cornell University, Ithaca, NY, USA

G. Vaughan
University of Manchester, Manchester, UK

W.W. Vaughan
*University of Alabama in Huntsville, Huntsville,
AL, USA*

G.A. Vecchi
GFDL/NOAA, Princeton, NJ, USA

A. Venkatram
*University of California – Riverside, Riverside,
CA, USA*

J.-P. Vernier
*Science Systems and Applications, Inc., Hampton,
VA, USA*

R.K. Vincent
*Bowling Green State University, Bowling Green,
OH, USA*

R. von Glasow
University of East Anglia, Norwich, UK

A. Wahner
Forschungszentrum Jülich, Jülich, Germany

R.M. Wakimoto
*National Center for Atmospheric Research, Boulder,
CO, USA*

D.E. Waliser
California Institute of Technology, Pasadena, CA, USA

J.M. Wallace
University of Washington, Seattle, WA, USA

- B. Wang
University of Hawaii, Honolulu, HI, USA
- Y. Wang
University of Hawaii at Manoa, Honolulu, HI, USA
- Z. Wang
University of Illinois at Urbana-Champaign, Urbana, IL, USA
- B. Ward
Public Works and Natural Resources, Longmont, CO, USA
- S. Warren
University of Washington, Seattle, WA, USA
- J. Waters
California Institute of Technology, Jet Propulsion Laboratory, Pasadena, CA, USA
- R.P. Wayne
University of Oxford, Oxford, UK
- S. Weart
Center for History of Physics, American Institute of Physics, College Park, MD, USA
- P.J. Webster
Georgia Institute of Technology, Atlanta, GA, USA
- P.J. Webster
University of Colorado – Boulder, Boulder, CO, USA
- W.F. Weeks
University of Alaska Fairbanks, Fairbanks, AK, USA
- M.L. Weisman
National Center for Atmospheric Research, Boulder, CO, USA
- S.J. Weiss
National Oceanic and Atmospheric Administration, Norman, OK, USA
- N.C. Wells
University of Southampton, Southampton, UK
- H. Wernli
Institute for Atmospheric and Climate Science, Swiss Federal Institute of Technology (ETH), Zürich, Switzerland
- T. Wetter
Johann Wolfgang Goethe-University, Frankfurt am Main, Institut für Meteorologie und Geophysik, Frankfurt am Main, Germany
- M.C. Wheeler
Centre for Australian Weather and Climate Research, Bureau of Meteorology, Melbourne, VIC, Australia
- A.A. White
University of Surrey, Guildford, UK
- D.N. Whiteman
NASA Goddard Space Flight Center, Greenbelt, MD, USA
- B.A. Wielicki
NASA Langley Research Center, Hampton, VA, USA
- E.R. Williams
Massachusetts Institute of Technology, Cambridge, MA, USA
- M.L. Wong
California Institute of Technology, Pasadena, CA, USA
- N. Wood
Met Office, Exeter, UK
- R. Wood
University of Washington, Seattle, WA, USA
- S.N. Wren
University of Toronto, Toronto, ON, Canada
- C.-C. Wu
National Taiwan University, Taipei, Taiwan
- D.J. Wuebbles
University of Illinois, Urbana, IL, USA
- L. Xie
North Carolina State University, Raleigh, NC, USA
- P. Yang
Texas A&M University, College Station, TX, USA
- S. Yang
NOAA/NWS/NCEP, Climate Prediction Center, Camp Springs, MD, USA
- J.A. Young
University of Wisconsin, Madison, WI, USA
- Z. Yu
College of Hydrology and Water Resources, Hohai University, Nanjing, China; and University of Nevada Las Vegas, Las Vegas, NV, USA
- Y.L. Yung
California Institute of Technology, Pasadena, CA, USA

S.E. Yuter
*North Carolina State University, Raleigh,
NC, USA*

S. Yvon-Lewis
*Texas A&M University, College Station,
TX, USA*

D. Zardi
University of Trento, Trento, Italy

S.E. Zebiak
*International Research Institute for Climate Prediction,
Palisades, NY, USA*

M.D. Zelinka
*Lawrence Livermore National Laboratory, Livermore,
CA, USA*

C. Zhang
University of Miami, Miami, FL, USA

F. Zhang
*Pennsylvania State University, University Park,
PA, USA*

M. Zhang
*Stony Brook University, Stony Brook,
NY, USA*

PREFACE TO THE FIRST EDITION

A half century ago the American Meteorological Society published the *Compendium of Meteorology*, which in a single volume of 1334 pages summarized the state of understanding of the atmosphere at that time. A perusal of the contents of that volume indicates that although a broad range of topics was covered, the vast bulk of the volume was devoted to traditional meteorological topics such as atmospheric dynamics, cloud physics, and weather forecasting. Barely 4 percent of the volume was devoted to articles related to atmospheric chemistry or air pollution and, of course, none of the volume was devoted to techniques such as satellites and remote sensing.

As Sir John Mason aptly notes in his foreword to the present work, the atmospheric sciences have expanded in scope enormously over the past 50 years. Topics such as atmospheric chemistry and global climate change, of only marginal interest 50 years ago, are now central disciplines within the atmospheric sciences. Increasingly, developing areas within the atmospheric sciences require students, teachers, and researchers to familiarize themselves with areas far outside their own specialties. This work is intended to satisfy the need for a convenient and accessible references source covering all aspects of atmospheric sciences. It is written at a level that allows undergraduate science and engineering students to understand the material, while providing active researchers with the latest information in the field.

More than 400 scientists, from academia, government, and industry have contributed to the 330 articles in this work. We are very grateful to these authors for their success in providing concise and authoritative summaries of complex subjects. As editors, we have benefited from the chance to learn from these articles, and we believe that all students and active scientists who want to increase their knowledge of the atmosphere will benefit enormously from access to this work.

We are also grateful to the 31 members of the Editorial Advisory Board who have guided us in our coverage of the very broad range of topics represented in this encyclopedia. Their willingness to suggest topics and authors, and to carefully review draft articles has contributed significantly to our success.

The production of this multivolume encyclopedia would not have been possible without the dedicated work of the staff of the Major Reference Works group at Academic Press. We are especially grateful to the Major Reference Work Development Manager, Colin McNeil, who has worked closely with us during the entire process. Finally, we appreciate the liberal use of color figures in the printed encyclopedia.

James R Holton, Judith A Curry, and John Pyle

This page intentionally left blank

PREFACE TO THE SECOND EDITION

Since the publication of the first edition of the *Encyclopedia of Atmospheric Sciences*, significant advances in research have been achieved all across the broad and expanding spectrum of the field and related disciplines. In particular, climate science with primary input from the atmospheric research emerges as a new field and integrator of interlocking peripheral disciplines over the last decade. These events have demanded the solicitation of new and updated articles for the 2003 edition. Some articles from the earlier publication were judged to be of such a fundamental and enduring nature that they did not require modification. But huge amounts of new information from Earth-orbiting satellite observatories have brought much new insight to the field. In addition there are new findings in many areas such as the latest simulations of meteorological and climatic processes of interest as well as simulations and observations of the composition and interaction of the field's chemical constituents. While interest in the ozone hole and its ramifications may have reached a plateau, ever more understanding of the stratosphere and its role in climate change emerges. The study of past climates provides new means of testing climate models and theories. In weather prediction we see new progress on how data are to be better assimilated for much improved initialization of the forecast model leading to the promise of more accurate predictions of severe weather and tropical cyclones over longer lead times. These are just a few of the new features of the second edition.

The editors of the second edition are greatly indebted to our predecessors in the first edition. They set the outline of topics and solicited the original authors, while establishing a high standard for the content of this publication. In many cases we decided to reprint those articles or request only minor updates. Nevertheless, many articles in this edition are entirely original, based on which we also made significant reorganization of the content. We are proud of our product and hope it provides the same assistance to students, researchers, and practitioners throughout the science and engineering communities.

Editors of the second edition
Gerald R North
Fuqing Zhang
John Pyle

This page intentionally left blank

EDITOR BIOGRAPHIES



Gerald R North received his PhD in theoretical physics from the University of Wisconsin in 1966. After postdoctoral research at the University of Pennsylvania he became a faculty member in physics at the University of Missouri–St. Louis. He shifted his research focus to climate science research during his sabbatical year at the National Center for Atmospheric Research, where he won the Outstanding Paper Award in 1975. He moved to NASA Goddard Space Flight Center (GSFC) in 1978 where he was awarded the NASA Medal for Research Excellence. During his stay at GSFC, he was the proposer and first study scientist for the Tropical Rainfall Measuring Mission, which was launched in 1997 and is still orbiting in 2014.

He moved to Texas A&M University in 1986 as a university distinguished professor of atmospheric sciences where he served as department head from 1995 to 2003. He has served as editor-in-chief of the *Reviews of Geophysics* and is recognized as one of the most cited authors in geosciences (Web of Science). He has chaired and/or served on a number of national committees and is a Fellow of the American Geophysical Union, American Meteorological Society (AMS) and the American Association for the Advancement of Science, and winner of the Jule Charney Award for Research (AMS).

He has published about 150 refereed papers not including many book chapters and reviews. His books include *Paleoclimatology*, co-authored with Thomas Crowley, and *An Introduction to Atmospheric Thermodynamics* co-authored with Tatiana Erikhimova.

North's interests are focused on the use of mathematical and statistical tools to solve climate problems over a wide range of issues including: analytical solutions of simplified energy balance climate models, use of random field techniques in representing and interpreting climate data and model simulations, detection of deterministic signals in climate change, statistical analysis satellite remote sensing for mission planning and analysis of data, paleoclimate problems using simplified climate models.



John Pyle obtained a BSc in Physics at Durham University before moving to Oxford where he completed a DPhil in Atmospheric Physics, helping to develop a numerical model for stratospheric ozone studies. After a short period at the Rutherford Appleton Laboratory he moved to a lectureship at Cambridge University in 1985. In 2000 he was appointed professor of atmospheric science and since 2007 has been the 1920 professor of physical chemistry. He is a Professorial Fellow at St Catharine's College. He has been a codirector of Natural Environment Research Council's National Centre for Atmospheric Science, where he is currently Chief Scientist.

His research focuses on the numerical modelling of atmospheric chemistry. Problems involving the interaction between chemistry and climate have been addressed; these range from stratospheric ozone depletion to the changing tropospheric oxidizing capacity and have included the environmental impact of aviation, land use change, biofuel technologies, and the hydrogen economy. He has studied palaeochemistry problems as well as the projected atmospheric composition changes during the current century. He has published more than 250 peer reviewed papers.

He played a major role in building an EU stratospheric research programme in the 1990s, coordinating several major field campaigns. He has contributed to all the WMO/UNEP assessments on stratospheric ozone since the early 1980s and is now one of the four international cochairns on

the Scientific Assessment Panel, responsible for these assessments. He was a convening lead author in the IPCC Special report "Safeguarding the ozone layer and the global climate system," published in 2006.

He was elected Fellow of the Royal Society in 2004 and an American Geophysical Union Fellow in 2011. He was awarded the Cambridge ScD degree in 2012. Other honours and awards include membership of Academia Europaea (1993), Royal Society of Chemistry (Interdisciplinary award, 1991, and John Jeyes lectureship, 2008), and the Royal Meteorological Society Adrian Gill Prize, in 2004.



Fuqing Zhang is a professor with tenure in the Department of Meteorology at the Pennsylvania State University, with a joint appointment in the Department of Statistics, along with an endowed position as the E Willard & Ruby S Miller Faculty Fellow at the College of Earth and Mineral Sciences at the Pennsylvania State University. His research interests include atmospheric dynamics and predictability, data assimilation, ensemble forecasting, tropical cyclones, gravity waves, mountain plains and sea-breeze circulations, warm-season convection, and regional-scale climate. He earned his BS and MS in meteorology from Nanjing University, China in 1991 and 1994, respectively, and his PhD in atmospheric science in 2000 from North Carolina State University.

He spent seven years as an assistant and then associate professor at Texas A&M University before coming to Penn State University as a full professor in 2008. In 2000, he spent a year and a half as a postdoctoral fellow at the National Center for Atmospheric Research. He also held various visiting scholarship appointments at various academic and research institutions including the National

Center for Atmospheric Research in Boulder, Colorado; the Navy Research Laboratory in Monterey, California; NOAA/AOML Hurricane Research Division, Miami, Florida; Peking University and Nanjing University, China; the Chinese State Key Laboratory of Severe Weather in Beijing, China; and Laboratoire de Meteorologie Dynamique, École Normale Supérieure in Paris, France.

He has authored/co-authored about 130 peer reviewed journal publications and has given more than 160 keynote speeches or invited talks at various institutions and meetings. He has served as principal investigator/co-principal investigator for 30 federal or state-sponsored research grants. He has chaired/cochaired more than 10 scientific meetings or workshops. He also served on various review or advisory panels for numerous organizations that include National Science Foundation, Office of Naval Research, NASA, NOAA, and National Academies. He has also served as editor of several professional journals including *Monthly Weather Review*, *Science China*, *Atmospheric Science Letter*, *Acta Meteorologica Sinica*, and *Computing in Science & Engineering*.

He has also received numerous awards for his research and service. Notably, in 2007 he received the **Outstanding Publication Award** from the National Center for Atmospheric Research. In 2009, was the sole recipient of the American Meteorological Society's 2009 **Clarence Leroy Meisinger Award** "for outstanding contributions to mesoscale dynamics, predictability, and ensemble data assimilation." Most recently, he received the 2014 American Meteorological Society's **Banner Miller Award** "for valuable insights into incorporating real-time airborne Doppler radar measurements via ensemble data assimilation, leading to improvements in forecasts of tropical cyclone track and intensity."

GUIDE TO USING THE ENCYCLOPEDIA

Structure of the Encyclopedia

The material in the encyclopedia is not arranged by ordinary alphabetical order, but by alphabetical order according to 49 principal topic areas taken to allow all papers belonging to each principal topic to appear together in the same volume. Within each principal subject, article headings are also arranged alphabetically, except where logic dictates otherwise. For example, overview articles appear at the beginning of a section.

There are four features that help you find the topic in which you are interested:

- i. the contents list
- ii. cross-references to other relevant articles within each article
- iii. a full subject index
- iv. contributors

i. Contents List

The contents list, which appears at the front of each volume, lists the entries in the order that they appear in the encyclopedia. It includes both the volume number and the page number of each entry.

ii. Cross-references

All of the entries in the encyclopedia have been cross-referenced. The cross-references, which appear at the end of an article as a See also list, serve four different functions:

- i. To draw the reader's attention to related material in other entries

- ii. To indicate material that broadens and extends the scope of the article
- iii. To indicate material that covers a topic in more depth
- iv. To direct readers to other articles by the same author(s)

Example

The following list of cross-references appears at the end of the article.

See also: **Biogeochemical Cycles:** Biogeochemistry of Iodine. **Stratospheric Chemistry Topics:** HO_x; Halogen Sources, Natural (Methyl Bromide and Related Gases); Halogens; Hydrogen Budget; Overview; Reactive Nitrogen (NO_x and NO_y).

iii. Index

The index includes page numbers for quick reference to the information you are looking for. The index entries differentiate between references to a whole article, a part of an article, and a table or figure.

iv. Contributors

At the start of each volume there is list of the authors who contributed to that volume.

This page intentionally left blank

BASIC ATMOSPHERIC STRUCTURE AND CONCEPTS

Contents

Beaufort Wind Scale

Wind Chill

Standard Atmosphere

Beaufort Wind Scale

L Hasse, Universität Kiel, Kiel, Germany

© 2003 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 1, pp 189–195, © 2003, Elsevier Ltd.

Introduction

For ships at sea, it has been widespread practice to include weather information into ship's logbooks to document the situation during its operations. For sail ships, wind information is most important. Beaufort adopted a scale to estimate the 'force of the wind' in 1805 when he was commanding officer of a man of war. The scale was formulated in terms of the effect of wind on sail ships of a certain type, but was subsequently used for other sail ships and steamers too.

Beaufort's scale of wind force was devised for use in the marine environment. However, since 1874, the Beaufort scale has been used in the international telegram code to transmit wind information from both sea and land. While Beaufort had used the behavior of sail ships in a given wind to define for what conditions common language terms like 'gentle breeze,' 'moderate gale,' 'whole gale,' or 'storm' should be applied, a different definition was required for land surfaces. Observers used certain indicators, e.g., behavior of flags, trees or drag plates, and feeling of wind in the face.

With today's knowledge of boundary layer meteorology, attempts to estimate the wind force over land appear questionable. One would need to use indicators of a known drag coefficient at a prescribed height in an open, level area without obstructions to the flow. Even then, correction for the roughness of the underlying terrain and for stability (e.g., day-night) would be required in order to make estimates comparable. Obviously, anemometer operation on land is a more direct method to determine wind speed instead of estimation. Use of the Beaufort scale to determine wind speed on land is not recommended and will not be discussed in the following.

Anemometer measurements on ships are difficult due to flow distortion. Also, only relative wind speed and direction are

measured. True wind needs to be calculated by vector addition of the independently measured course and speed of a ship, a source of additional error variance. Estimates of Beaufort force have therefore remained a tool for an extended period of time.

With improved types of rigging and transition from sail ships to steamers, the original definition of Beaufort forces was endangered. The appearance of the sea, wind effects on rigging, whistling of the wind, and other phenomena may have helped to pass on a tradition from experienced observers to younger colleagues, though not coded in words. Ship officers in the transition time likely had sufficient training on sail ships to estimate the wind force even on board steamers.

In 1927, a description of the Beaufort scale in terms of sea state was formulated by Petersen as a result of many years of experience (Table 1). His description was added later to the wind code by the International Meteorological Organization (IMO), the predecessor of the World Meteorological Organization (WMO). Obviously, IMO did not view reference to sea state as a redefinition, but rather as a written account of an existing practice. It is common habit to call the redefined scale the Beaufort scale of wind force too.

The use of Beaufort numbers for coding purposes had initiated early investigations into wind speed equivalents to Beaufort numbers, often called a 'Beaufort equivalent scale.' Measured wind speeds presumably could ensure consistent use of Beaufort numbers over land and would help to alleviate the difficulties resulting from change of ship types with time. Also, estimated wind force numbers do not really fit into the concepts of theoretical meteorology, where wind velocity is used in the basic Navier–Stokes equations.

Table 1 Definition of Beaufort force in terms of sea state

Beaufort number	Common name	Definition
0	Calm	Sea like a mirror
1	Light air	Ripples with the appearance of scales are formed, but without foam crests
2	Light breeze	Small wavelets, still short but more pronounced; crests have a glassy appearance and do not break
3	Gentle breeze	Large wavelets; crests begin to break; foam of glossy appearance; perhaps scattered white horses
4	Moderate breeze	Small waves, becoming longer; fairly frequent white horses
5	Fresh breeze	Moderate waves, taking a more pronounced long form; many white horses are formed (chance of some spray)
6	Strong breeze	Large waves begin to form; the white foam crests are more extensive everywhere (probably some spray)
7	Near gale	Sea heaps up and white foam from breaking waves begins to be blown in streaks along the direction of the wind
8	Gale	Moderately high waves of greater length; edges of crests begin to break into the spin drift; the foam is blown in well-marked streaks along the direction of the wind
9	Strong gale	High waves; dense streaks of foam along the direction of the wind; crests of waves begin to topple, tumble, and roll over; spray may affect visibility
10	Storm	Very high waves with long overhanging crests; the resulting foam, in great patches, is blown in dense white streaks along the direction of the wind; on the whole, the surface of the sea takes a white appearance; the tumbling of the sea becomes heavy and shock like; visibility affected
11	Violent storm	Exceptionally high waves (small- and medium-sized ships might be for a time lost to view behind the waves); the sea is completely covered with long white patches of foam lying along the direction of the wind; everywhere the edges of the wave crests are blown into froth; visibility affected
12	Hurricane	The air is filled with foam and spray; sea completely white with driving spray; visibility very seriously affected

Reproduced from World Meteorological Organization, 1970. The Beaufort Scale of Wind Force. Reports on Marine Science Affairs No. 3. World Meteorological Organization, Geneva.

From these introductory remarks, it can be seen that there are at least two tasks:

1. selection of a suitable relation between the force of wind and the wind velocity; and
2. correction for inhomogeneities in the time series of winds at sea.

Relation between Beaufort Force and Wind Velocity: A Problem of Physics and Regression Interpretation

A sail ship is not a well-defined tool to measure wind speed. The same is true for the appearance of the sea surface that depends to some degree on the history of both the wind field and the wave field and even on turbulence in the atmospheric boundary layer over the sea. In retrospect, it appears wise that only a coarse scale was devised for a measure of wind force. As only a statistical relation can be expected, requirements on such a scale need to be defined.

Following Lindau, an optimal Beaufort equivalent scale is required to convert Beaufort estimates into wind speeds such that derived climatological quantities like means and variances agree with respective quantities from unbiased wind measurements.

The attempts to derive a Beaufort equivalent scale differ in the meteorological setup of the experiment and in the subsequent statistical interpretation. Difficulties exist in both parts. It appears that misunderstandings in the interpretation of statistical methods have hampered the development of and agreement on a Beaufort equivalent scale more than questions of measurements and exposure.

One-Sided Regressions

Derivations of Beaufort equivalent scales typically use regression techniques. For a pair of variables (x , y), two

regression curves can be derived, depending on which of the variables are considered 'independent.' A common notion of regression sounds like: "The regression of y on x gives the best estimate (by the method of least squares) of y for a given independent x , and, similarly, the regression x on y gives the best estimate of x as a function of independent y ." The one-sided regression implies the 'independent' variable as nonrandom and attributes all random variation to the other, as can be seen if one follows the derivation, e.g., for a linear fit. The same is true in nonlinear regression, where typically means of the dependent variable are calculated for intervals of the independent variable.

The physical problem of one-sided regressions is seen in the following. If an equivalent scale from data including error variances in both variables is determined, the spread of the independent variable is increased by its random errors and the resulting regression line has too small a slope. Applying this regression to climatological variables (e.g., monthly mean wind speed and variance), the derived variances will deviate systematically from respective quantities determined from unbiased measured winds.

In the development of Beaufort equivalent scales, two approaches have been used:

1. We take the Beaufort number as well determined and calculate the mean velocities for each interval of Beaufort number; i.e., a nonlinear regression of wind velocity is used on Beaufort number.
2. We argue that the average wind velocity is well measured and the random variations stem from imperfect estimation of Beaufort numbers. In this case, the regression of Beaufort force on wind velocity is the given choice.

The pitfall is seen in the following: in one-sided regression, the independent variables are treated as nonrandom and all variability from both axes is ascribed to the dependent variable.

The first approach was applied in early investigations listing wind speeds averaged for each Beaufort number, although difficulties of both Beaufort observations and the exposure and calibration of wind instruments were discussed.

However, there are good arguments to use the second approach, the regression of estimates to measured speeds. It is known that Beaufort numbers are estimates only, which are influenced by different external conditions and by uncertainties in the observer's judgment. Also, the quantization error is larger for estimates due to the larger intervals of the Beaufort scale compared to speeds in meters per second. It is not unreasonable to infer the error of Beaufort number estimation to be much larger than the uncertainty of anemometer measurements. Köppen in 1888 already advocated the averaging of Beaufort estimates for intervals of measured speeds in order to establish an equivalent scale.

IMO adopted a scale for official use, corresponding to today's WMO code 1100, that was established in 1906 by Simpson, Meteorological Office, London, using regression of Beaufort estimates to measured speeds as recommended by Köppen.

Two-Sided Regressions

It is evident that both Beaufort force estimates and velocity measurements contain random errors, and a two-way regression would be the appropriate tool. The best relation between two random variables lies between the two one-sided regressions. For linear regressions, in the absence of information on the respective error variances, the best choice would be the bisector of the angle between the two one-sided regression lines. It minimizes the orthogonal distances of observed points from the line. This is equivalent to assuming the error variances to be the same fraction of total variance for each of the variables. This so-called orthogonal regression is certainly better than a one-sided regression, when both variables are subject to random errors. However, in a given case, the fraction of error in the total variance of each variable need not be the same and an improved technique is required.

For illustration, consider the observations at the Ocean Weather Station (OWS) K in Figure 1. The two one-sided regression lines and the orthogonal regression line are plotted together with the monthly means of anemometer measurements at OWS K and Beaufort estimates from ships in the vicinity. The fit through the monthly means deviates from the orthogonal regression in the direction toward the one-sided regression of Beaufort numbers on measured speeds. This implies that the uncertainty of Beaufort estimates is a larger fraction of the total variance than that of anemometer measurement.

The orthogonal regression yields a better Beaufort equivalent scale than any of the one-sided regressions – in the sense to better reproduce the means of wind speed from Beaufort estimates. The agreement is even better than one would expect from error estimates of wind speed measurements and Beaufort estimates. The reason for this is seen in the natural variability between measurements colocated in space and time that enters as 'unexplained' variance in the regression too.

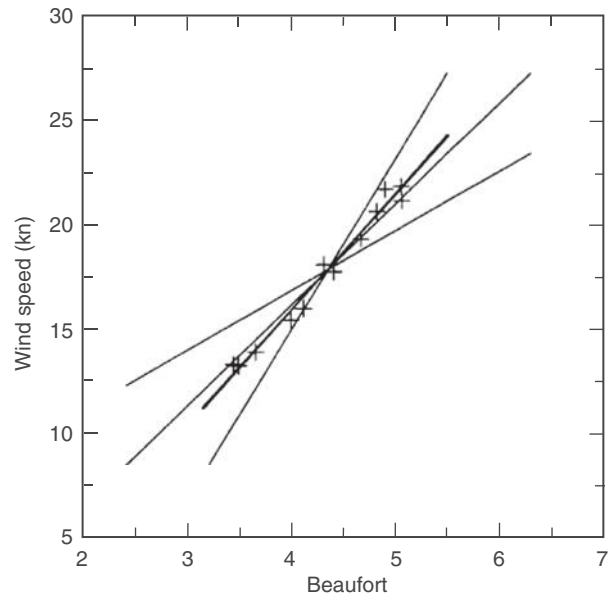


Figure 1 Relation between Beaufort force and wind speed (kn), based on observations from OWS K and ships within 500 km from OWS K, period 1960–71. One-sided and 'orthogonal' linear regressions (thin lines) are compared with monthly means. The linear fit of monthly means (full line) deviates from the best linear regression calling for explicit consideration of error variances. Reproduced from Lindau, R., 1994. Eine neue Beaufort-Äquivalentskala. Berichte Institut für Meereskunde Kiel, Kiel No. 249.

An Optimized Beaufort Equivalent Scale

Wind speed measurements at OWSs and Beaufort estimates from passing voluntary observing ships (VOSs) were used in several attempts to derive an improved equivalent scale. Some averaging is needed to reduce errors of measurements in respective observations and also to account for the natural variability that enters because VOSs pass OWSs at some distance.

Lindau developed a sophisticated method to determine effective variances. From simultaneous observations of pairs of VOSs, differences were obtained and variances calculated as a function of distance between the ships. Extrapolating to fictitious zero distance, the error variances of VOSs were obtained. Using the same technique on pairs of OWSs and VOSs, the error variances of OWSs were also determined.

Knowing the variances, influence of errors and natural variability could be reduced by averaging in time for OWSs and space for VOSs. A suitable area around an OWS was selected to contain the same variance in space from VOS than at the OWS in time for 1 day. This radius and the appropriate number of VOS observations were determined separately for each OWS and each season.

With the error variance at OWS and VOS reduced by averaging the appropriate number of observations, and the natural variability being assured to be the same, the 'orthogonal' regression yields the correct relation. Instead of straight lines, Lindau used the method of cumulative frequencies in order to admit nonlinearities. Anemometer heights at weather ships are near 25 m. Wind measurements at OWSs were reduced to 10 m

Table 2 Selection of Beaufort equivalent scales, given in m s^{-1}

Beaufort	0	1	2	3	4	5	6	7	8	9	10	11	12
IMO (1926) 6 m ^a	0.0	1.1	2.5	4.3	6.3	8.6	11.1	13.8	16.7	19.9	23.3	27.1	
Code 1100 10 m ^a	–	0.8	2.4	4.3	6.7	9.4	12.3	15.5	18.9	22.6	26.4	30.5	~34.8
Lindau 10 m ^b	0.0	1.2	2.7	4.6	7.2	9.7	12.1	14.6	17.3	20.2	23.4	27.1	31.4
CMM-IV 18 m ^a	0.8	2.0	3.6	5.6	7.8	10.2	12.6	15.1	17.8	20.8	24.2	28.0	~32.2
Lindau 25 m ^c	0.1	1.2	2.8	4.9	7.7	10.5	13.1	15.9	18.9	22.2	26.0	30.3	35.4
Kaufeld 25 m ^d	0.4	1.9	4.1	6.4	8.7	11.0	13.4	15.9	18.7	21.8	25.1	28.6	32.4
Reduction factor 25–10 m			0.96	0.94	0.94	0.92	0.92	0.92	0.92	0.91	0.90		

Often knots ($1 \text{ kn} = 0.5144$; $1 \text{ m s}^{-1} = 1.944 \text{ kn}$) are used as the unit. The last line gives the reduction factor to compare equivalent scales for reference heights of 25 and 10 m, derived for the ensemble of measurements at North Atlantic OWSs.

^aReproduced from World Meteorological Organization, 1970. The Beaufort scale of wind force. Reports on Marine Science Affairs No. 3. World Meteorological Organization, Geneva.

^bReproduced from Lindau, R, 2001. Climate Atlas of the Atlantic Ocean. Springer Verlag, Berlin, Heidelberg.

^cReproduced from Lindau, R, 1995. A new Beaufort equivalent scale. In: Proceedings of the International COADS Winds Workshop. Berichte aus dem Institut für Meereskunde, Kiel, No. 265 [available from Institut für Meereskunde, 24105 Kiel, Germany or NOAA, Environmental Research Laboratories-CDC, Boulder, CO 80303, USA].

^dReproduced from Kaufeld, L., 1981. The development of a new Beaufort equivalent scale. Meteorologische Rundschau 34, 17–23.

height using a diabatic wind profile. The resulting equivalent scale is thus applicable to 10 m height. The results are given in Table 2 and depicted in Figure 2.

Discussion of Heights for Equivalent Wind Velocities

Beaufort estimates per se have no natural height above the sea surface. However, equivalent scales give wind speeds. Because of the approximately logarithmic wind profile in the atmospheric boundary layer, for applications of the Beaufort equivalents the corresponding height needs to be known.

Code 1100 in use from 1926 to 1946 was believed to refer to 6 m height. Since the required standard height of anemometers had changed from 6 to 10 m, a slightly changed

code 1100 was introduced in 1946 by WMO, supposedly applicable to 10 m height. Both versions of code 1100 were based on data sets into which observations from the Scilly Islands entered in a relatively large number. However, wind speeds were taken at the small island of St Mary's that features several heights reaching 30–50 m above mean sea level. It is uncertain to what height in undisturbed flow over water the measurements would correspond.

Lindau's improved equivalent scale settled the problem of reference height by using anemometer measurements of known heights and reducing winds to heights of 10 m individually with the aid of the diabatic wind profile. He could also show that for the WMO code 1100 of 1946 a reference height of 10 m is reasonable. The anemometer measurements at OWSs are not corrected for flow distortion. One can hope that the exposure of instruments at OWSs and the mode of ship operation at the station will make this an acceptable error.

Often a reduction of velocity measurements from anemometer height to reference height is necessary. Typically, a constant reduction factor, derived from the neutral wind profile, is used. However, slightly unstable conditions prevail at most parts of the oceans, approaching near-neutral conditions at higher wind speeds. For the mix of stabilities at the North Atlantic OWS, a reduction factor of 25–10 m decreases with wind speeds from ~0.94 to ~0.90; see Table 2.

Discussion of Other Scales

The matter of Beaufort equivalent scales found renewed interest in the second half of the twentieth century. Especially within the wave modeling community, still in 1990 the opinion prevailed that the WMO Beaufort equivalent scale (code 1100) is in error. For example, the WMO Commission of Marine Meteorology produced a scale – known as CMM-IV scale – using regression of anemometer measurements on Beaufort estimates, similar to the scale of Cardone. Regrettably, in the new derivations the variable with small error variances was regressed on a variable with obviously much larger error variances, leading to biased scales.

Fortunately, the governing bodies of WMO adhered to code 1100, though admitted application of CMM-IV and similar scales for scientific purposes, so-called scientific

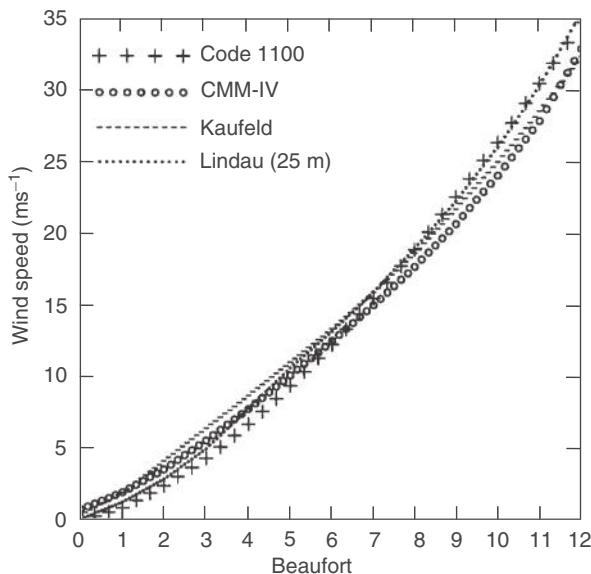


Figure 2 Different Beaufort equivalent scales: code 1100 (10 m reference height), crosses; CMM-IV (25 m height), circles; Kaufeld (25 m height), dashes; and Lindau (25 m height), dots. Code 1100 determined by regression of Beaufort on wind speed, CMM-IV by wind speed on Beaufort, Kaufeld, and Lindau from cumulative frequencies.

equivalent scales. In retrospect, the scientific scales are the wrong choice. They give biased climatological means. The use of 'scientific' scales in wave modeling is even more questionable since this practice is at variance to the use of code 1100 in operational weather analysis and forecasts. Seen in the light of the correct derivation by Lindau, it turns out that the insight of Köppen and Simpson around 1900 resulted in a scale, WMO code 1100, that is less biased than some scientific scales of 70 years later. (It might be noted that the description of Köppen's method in the WMO Report on marine science affairs of 1970 is reversed to what Köppen advocated and used, i.e., regression of Beaufort on wind speed.)

The regression technique of cumulated frequencies applied by Lindau had been used by Kaufeld before him. It has the advantage to account for nonuniform distribution of observations at the tails of the frequency distribution. This technique can be seen as the nonlinear equivalent to the 'orthogonal' two-sided regression. Though the different error variances of the two variables were not accounted for, Kaufeld's scale is certainly preferable to the so-called scientific scales derived by one-sided regression of wind speed on Beaufort force.

Kent and Taylor reviewed a selection of Beaufort equivalent scales and the techniques used in their derivation, and concluded that the Beaufort equivalent scale of Lindau is to be preferred when creating a homogeneous monthly mean wind data set from anemometer and visual winds in Comprehensive Ocean Atmosphere Data Set (COADS).

Time Dependence of Beaufort Estimates

Increasing interest in long time series of climate data inevitably leads to the question of whether the reported winds from the oceans are homogeneous in time. The slow change of observing practices was mentioned. There have been changes in coding practices too. Originally, Beaufort forces were used in transmitting data. Effective in 1948, WMO changed from Beaufort forces to transmit wind speeds in knots (1 knot equals 1 nautical mile per hour or 0.51 m s^{-1}). For a short time of transition, the erroneous use of codes may have influenced reported winds, but no long-term trend is expected.

There are an increasing number of ships carrying anemometers. Determination of true wind from anemometer measurements requires vector subtraction of course and speed of the ship, certainly an additional source for errors and inaccuracies. Also, the code indicating the data as either measured (anemometer) or estimated (Beaufort scale) is known to be less reliable. Peterson found frequency distributions of *estimated* winds to show significant secular changes and ships carrying anemometers to report higher estimated winds than ships without anemometers. The latter is most embarrassing, since no single reason and simple cure can be given.

On the other hand, a trend toward higher wind speeds could well be an indication of climate change. Growing amounts of greenhouse gases change the radiation balance. At the oceans, surplus energy can fuel atmospheric circulations, e.g., midlatitude and tropical storms. In fact, plotting the recorded wind speeds at face value shows significant trends.

Ramage showed that reported ocean wind speeds exhibit secular changes. He studied monthly mean wind speeds at the

Singapore–South China route, taking along the track pressure differences for reference. He attributed the decrease of wind speed prior to World War II to transition from sail to steam and the increase thereafter to a growing portion of ships carrying anemometers; a true secular trend could not be isolated.

Taking the evidence together suggests deriving time-dependent corrections for the Beaufort equivalent scale. The obvious tool is to use independently measured pressure gradients as reference, presuming the relation between surface wind and pressure field to be invariant. Unfortunately, ships tend to follow certain lines providing pressure differences rather than full horizontal pressure gradients. Lindau developed a statistical method to derive wind velocities relative to pressure gradients. He used simultaneous wind vectors and pressure differences from pairs of ships within reasonable distance. Pressure differences were sorted according to the relative wind direction and the magnitude of pressure gradient obtained by fit (Figure 3). Errors in wind direction estimated at VOSs are equivalent to a smoothing of the pressure differences and have been accounted for.

Taking 1960–71 as the base period, time-dependent corrections for estimated wind speeds have been derived for the North Atlantic between 20° N and 60° N . From the uncorrected data, one would identify long-term trends of different sign before and after World War II, while corrected for drift of scale the apparent increase of wind speed after 1950 disappears (Figure 4). Considering the full time range from 1890 to 1990,

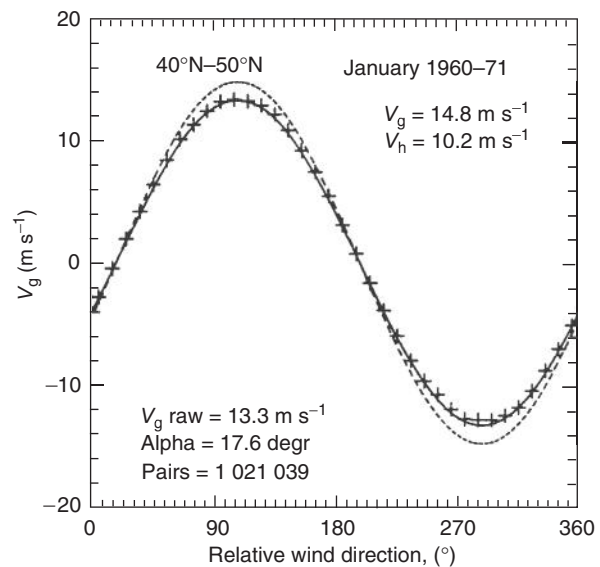


Figure 3 Determination of geostrophic wind V_g as reference for time-dependent calibration of Beaufort estimates, based on pressure differences from pairs of ships. Pressure differences are shown as a function of wind direction relative to the direction of paired observations. The dashed line is obtained considering the error variance of wind direction. Based on data from the North Atlantic between 40° N and 50° N , January, period 1960–71. Reproduced from Lindau, R., 1995. Time dependent calibration of marine Beaufort estimates using individual pressure differences. In: Proceedings of the International COADS Winds Workshop. Berichte Institut für Meereskunde, Kiel, No. 265 [available from Institut für Meereskunde, 24105 Kiel, Germany or NOAA, Environmental Research Laboratories-CDC, Boulder CO 80303, USA].

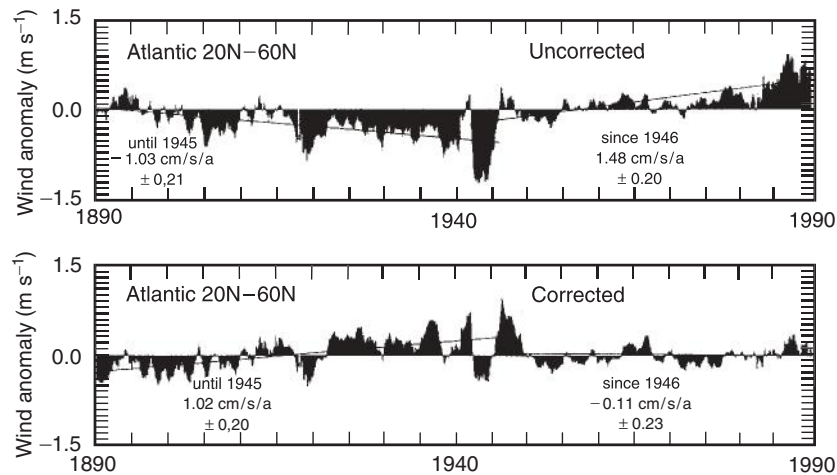


Figure 4 Mean wind speed anomalies for the North Atlantic Ocean between 20° N and 60° N, converted from Beaufort estimates using the Lindau scale for 10 m height. Upper panel uncorrected; lower panel corrected with reference to pressure gradients. Reproduced from Lindau, R., 2001. *Climate Atlas of the Atlantic Ocean*. Springer Verlag, Berlin, Heidelberg.

there appears to be no physically significant trend over the North Atlantic Ocean.

Unresolved Issues

The low and high velocity ends of the Beaufort scale are less well determined. This is not important for the low wind speeds. For the high wind speeds, obviously too few observations are available to establish Beaufort equivalents. In typical listings, speed ranges for the Beaufort numbers are given as well as averages. Beaufort number 12 is usually taken as unlimited, i.e., with no equivalent speed assigned. Bortkovskii fitted a Maxwell distribution to high wind speeds from winter Atlantic OWS observations for the description of storminess in air–sea interaction studies. This method could be used to establish a mean equivalent velocity for Beaufort 12, under the premise of stationarity of the time series.

Beaufort equivalent scales have been determined mainly from observations at the North Atlantic but accepted by WMO for international use. If sorted by ship call sign (corresponding roughly to national origin of data), there are differences in mean results. Noting differences in time as well as in national use, it would be desirable to check the equivalence scale for consistency at other oceans too. The indication that anemometer-carrying ships report estimated winds higher than ships without an anemometer calls for further study too.

Collecting marine surface observations has been a habit of generations. In the beginning, national services collected logbooks. Later, data were transferred to punched cards and collected sets (referred to as ‘decks’) binationally exchanged. Multinational collections finally resulted in the COADS. Wind observations at sea were collected more systematically since about 1860 – welcome information for climate studies. It is hoped that the methods available now will be adopted by the

international community to provide for the production of a homogeneous, unbiased time series of ocean wind fields to be used in future research.

Further Reading

- Cardone, V., Greenwood, J., Cane, M., 1990. On trends in historical marine wind data. *Journal of Climate* 3, 113–127.
- Cheng, C.-L., Van Ness, J.W., 1999. *Statistical Regression with Measurement Error*. Arnold, London, UK.
- Dobson, F.W., 1980. Review of Reference Height for and Averaging Time of Surface Wind Measurements at Sea. Marine Meteorology and Related Oceanographic Activities, Report No. 3. World Meteorological Organisation, Geneva.
- Isemer, H.J., Hasse, L., 1991. The scientific Beaufort equivalent scale: Effects on wind statistics and climatological air-flux estimates in the North Atlantic Ocean. *Journal of Climate* 4, 819–836.
- Kent, E., Taylor, P., 1997. Choice of a Beaufort equivalent scale. *Journal of Atmospheric and Oceanic Technology* 14, 228–242.
- Kinsman, B., 1969. Historical notes on the original Beaufort scale. *Marine Observer* 39, 116–124.
- Lindau, R., 1994. Time dependent calibration of marine Beaufort estimates using individual pressure differences. In: Diaz, H.F., Isemer, H.-J. (Eds.), *Proceedings International COADS Workshop*. Berichte Institut für Meereskunde, Kiel, Germany No. 265.
- Lindau, R., 2001. *Climate Atlas of the Atlantic Ocean*. Springer Verlag, Berlin, Heidelberg.
- Lindau, R., 2002. Rapport on Beaufort Equivalent Scales. Advances in the Applications on Marine Climatology – The Dynamic Part of the WMO Guide to the Applications of Marine Climatology. JCOMM Technical Report No. 13, WMO/TD-No. 1081. World Meteorological Organization, Geneva.
- Petersen, P., 1927. Zur Bestimmung der Windstärke auf See. *Annalen der Hydrographie* 55, 69–72.
- Peterson, E.W., Hasse, L., 1987. Did the Beaufort-scale or the wind climate change? *Journal of Physical Oceanography* 17, 1071–1074.
- Woodruff, S.D., Diaz, H.F., Elms, J.D., Worley, S.J., 1998. COADS release 2 data and metadata enhancements for improvements of marine surface flux fields. *Physics and Chemistry of the Earth* 23, 517–526.
- World Meteorological Organization, 1970. *The Beaufort Scale of Wind Force*. Reports on Marine Science Affairs No. 3. World Meteorological Organization, Geneva.

Wind Chill

M Bluestein, Indiana University – Purdue University, Indianapolis, IN, USA

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

Wind Chill is the name given to the effect of increased heat loss from the skin exposed to cold, windy weather. It is characterized by the wind chill temperature (WCT) defined as that air temperature with no appreciable wind effecting the same heat loss rate as the actual temperature and wind. This article provides a history of how the WCT was originally developed, what flaws were found in that research, and how the current chart of WCT versus temperature and wind speed was arrived at. The chart is provided in both Fahrenheit and Celsius units.

Introduction

All bodies exposed to colder air will lose heat because of the temperature difference between their skin and the air. The rate of heat loss in still air is a function of that difference as well as the amount of exposed area of the body, the conditions at the body's surface, and whether there is radiant energy available. These are well-known principles of the branch of science and engineering called 'thermodynamics.'

It is also known that if the air is moving across the body's surface, the rate of heat loss increases. With increases in air speed, or wind, the heat loss rate increases, a process known to engineers as forced convection. This increase in heat loss rate affects humans exposed to cold, windy weather and has become known as wind chill. The human body feels colder, particularly at the unclothed areas of skin, when there is wind than when there is no appreciable wind at the same temperature. Accelerated heat loss rates can lead to frostbite if the skin remains exposed long enough. The greater the wind speed, the greater the heat loss rate, and the faster will frostbite set in as the skin temperature drops.

Meteorologists have long sought to characterize these effects of the wind on humans. The forced convection process described above has been called the 'wind chill factor.' What has been desired is a quantitative measure of this effect that can be used in weather forecasting and reporting. A number of such measures have been utilized. The heat loss rate per unit area of exposed skin was utilized in Canada until recently. A proposal to report the skin temperature has been presented by researchers in Israel. By far the most accepted measure is the wind chill temperature (WCT) used in the United States and other countries. The WCT is defined as that air temperature with no appreciable wind that would effect the same heat loss rate from exposed skin as that due to the actual dry bulb temperature with wind. The term 'no appreciable wind,' often referred to as 'still air,' is used because the body is considered to be in motion at all times relative to the air. After all, everyone is at least breathing, and most likely moving to get out of the wind on a cold day.

A WCT chart is currently utilized by the US National Weather Service (NWS) and the Canadian weather service, Environment Canada (EC). For a given air temperature and wind speed as reported from a weather station, the WCT may be read from a chart such as that shown in [Figure 1](#), which includes versions in both English and metric units. This new

chart, established in November 2001, includes warnings of a risk of frostbite based on time of exposure.

Numerous researchers had been critical of the original chart, which was developed from research done by Siple and Passel in Antarctica during the 1940s and was in use by the NWS from 1973 to 2001. The origin of the first chart will be reviewed and criticism of its accuracy will be discussed. The basis for the new chart will be described. Since wind chill results in individual discomfort, a degree of subjectivity will always be present in such a concept. Many physical factors may also play a part in an individual's response to wind chill, such as the amount of sunshine present, the wind direction and variability, moisture in the air, and the presence of ground structures that modify the wind.

Origin of the Wind Chill Temperature Calculation

The need to properly clothe American servicemen for the rigors of the European winters of World War II gave rise to the first formal study of wind chill. Two US Army researchers, Paul Siple and Charles Passel (hereafter referred to as S & P), investigated the effect of wind on the heat loss from a body during the Antarctic winter of 1941. The 'body' was a plastic cylinder containing 250 g of water. The cylinder was made of cellulose acetate with dimensions 14.9 cm long, 5.7 cm in diameter, and 0.3 cm thick. A therm ohm, a device that measures temperature via electrical resistance, was immersed in the water and the cylinder was sealed. The cylinder was suspended above the recording station with a thermohm next to it to record dry bulb temperatures. Nearby, a cup anemometer recorded the wind speed. As the cylinder lost heat, the water would begin to freeze. Knowing the heat of fusion of water, the time the water spent at the freezing point would yield the heat loss rate. That time was determined by monitoring the water temperature. The heat loss rate was then modified by the container area exposed and the difference between the container temperature (assumed to be 0 °C) and the air temperature to arrive at a heat transfer coefficient. Such coefficients are necessary if one is to relate the experimental results to other bodies with other temperatures. The units of heat loss rate employed by S & P were kilogram-calories (kcal) per hour, with heat transfer coefficients in units of kcal per hour per square meter per degree Celsius. Average air temperatures ranged from −56 °C to −9 °C, and wind speeds from still air to 12 ms^{−1} (27 miles h^{−1}). Heat transfer

		Temperature (°F)																
Wind (mile h ⁻¹)	Calm	32	30	25	20	15	10	5	0	-5	-10	-15	-20	-25	-30	-35	-40	-45
	5	27	25	19	13	7	1	-5	-11	-16	-22	-28	-34	-40	-46	-52	-57	-63
	10	24	21	15	9	3	-4	-10	-16	-22	-28	-35	-41	-47	-53	-59	-66	-72
	15	22	19	13	6	0	-7	-13	-19	-26	-32	-39	-45	-51	-58	-64	-71	-77
	20	20	17	11	4	-2	-9	-15	-22	-29	-35	-42	-48	-55	-61	-68	-74	-81
	25	19	16	9	3	-4	-11	-17	-24	-31	-37	-44	-51	-58	-64	-71	-78	-84
	30	18	15	8	1	-5	-12	-19	-26	-33	-39	-46	-53	-60	-67	-73	-80	-87
	35	17	14	7	0	-7	-14	-21	-27	-34	-41	-48	-55	-62	-69	-76	-82	-89
	40	16	13	6	-1	-8	-15	-22	-29	-36	-43	-50	-57	-64	-71	-78	-84	-91
	45	15	12	5	-2	-9	-16	-23	-30	-37	-44	-51	-58	-65	-72	-79	-86	-93
Wind (mile h ⁻¹)	50	14	12	4	-3	-10	-17	-24	-31	-38	-45	-52	-60	-67	-74	-81	-88	-95
	55	14	11	4	-3	-11	-18	-25	-32	-39	-46	-54	-61	-68	-75	-82	-89	-97
	60	13	10	3	-4	-11	-19	-26	-33	-40	-48	-55	-62	-69	-76	-84	-91	-98

(a)

Frostbite may occur in 30 min or less

		Temperature (°C)												
Wind (km h ⁻¹)	Calm	10	5	0	-5	-10	-15	-20	-25	-30	-35	-40	-45	-50
	10	8.6	2.7	-3.3	-9.3	-15.3	-21.1	-27.2	-33.2	-39.2	-45.1	-51.1	-57.1	-63.0
	15	7.9	1.7	-4.4	-10.6	-16.7	-22.9	-29.1	-35.2	-41.4	-47.6	-53.7	-59.9	-66.1
	20	7.4	1.1	-5.2	-11.6	-17.9	-24.2	-30.5	-36.8	-43.1	-49.4	-55.7	-62.0	-68.3
	25	6.9	0.5	-5.9	-12.3	-18.8	-25.2	-31.6	-38.0	-44.5	-50.9	-57.3	-63.7	-70.2
	30	6.6	0.1	-6.5	-13.0	-19.5	-26.0	-32.6	-39.1	-45.6	-52.1	-58.7	-65.2	-71.7
	35	6.3	-0.4	-7.0	-13.6	-20.2	-26.8	-33.4	-40.0	-46.6	-53.2	-59.8	-66.4	-73.1
	40	6.0	-0.7	-7.4	-14.1	-20.8	-27.4	-34.1	-40.8	-47.5	-54.2	-60.9	-67.6	-74.2
	45	5.7	-1.0	-7.8	-14.5	-21.3	-28.0	-34.8	-41.5	-48.3	-55.1	-61.8	-68.6	-75.3
	50	5.5	-1.3	-8.1	-15.0	-21.8	-28.6	-35.4	-42.2	-49.0	-55.8	-62.7	-69.5	-76.3
Wind (km h ⁻¹)	55	5.3	-1.6	-8.5	-15.3	-22.2	-29.1	-36.0	-42.8	-49.7	-56.6	-63.4	-70.3	-77.2
	60	5.1	-1.8	-8.8	-15.7	-22.6	-29.5	-36.5	-43.4	-50.3	-57.2	-64.2	-71.1	-78.0

(b)

Frostbite may occur in 30 min or less

Figure 1 (a) The new wind chill temperature (WCT) Index chart, with T = air temperature in °F and V = wind speed in miles h⁻¹ at 33 ft elevation. WCT (°F) = $35.74 + 0.6215 T - 35.75 V^{0.16} + 0.4275 TV^{0.16}$. (b) The new WCT Index chart, with T = air temperature in °C and V = wind speed in km h⁻¹ at 10 m elevation. WCT (°C) = $13.12 + 0.6215 T - 11.37 V^{0.16} + 0.3965 TV^{0.16}$.

coefficients were plotted against wind speed and the original data are shown in Figure 2. The authors mistakenly called the coefficient as the cooling rate in this figure. The relationship between the heat transfer coefficient and the wind speed was determined by a best-fit line through the recorded data points. This line, shown solid in the figure, is represented by eqn [1], where h = heat transfer coefficient, in kcal h⁻¹m⁻² °C⁻¹ and v = wind speed, in m s⁻¹.

$$h = 10\sqrt{v} + 10.45 - v \quad [1]$$

The heat loss rate for a body would then be given by eqn [2], where Q = heat loss rate, in kcal h⁻¹, A = surface area exposed, in m², T_s = surface temperature, in °C, and T_a = air temperature, in °C.

$$Q = hA(T_s - T_a) \quad [2]$$

Under the Antarctic conditions, the heat loss from the container surface would have been due to forced convection created by the air moving over the surface plus radiation from the surface to the surrounding air because of the

temperature difference. Equation [1] indicates that heat loss would increase with wind speed up to a value of 25 m s⁻¹, or about 56 miles h⁻¹, and decrease thereafter.

Armed with a relationship between wind speed and the heat transfer coefficient, the authors then set about to apply this to humans. Using the concept of a WCT as defined above, the actual heat loss rate due to air temperature and wind speed is set equal to the rate when the air would be at the WCT and still. This is expressed mathematically as eqn [3], where h_0 = heat transfer coefficient when the air is still (kcal h⁻¹ m⁻² °C⁻¹).

$$hA(T_s - T_a) = h_0A(T_s - \text{WCT}) \quad [3]$$

The authors assumed a constant skin temperature of 33 °C for T_s . It was further assumed that the heat transfer coefficient in still air would be that value found by eqn [2] for an air speed of 1.79 ms⁻¹, (~4 miles h⁻¹). This had become the accepted value for still air. Equation [3] may be solved for WCT for given values of dry bulb temperature and air speed, since the areas on either side of the equation are the same.

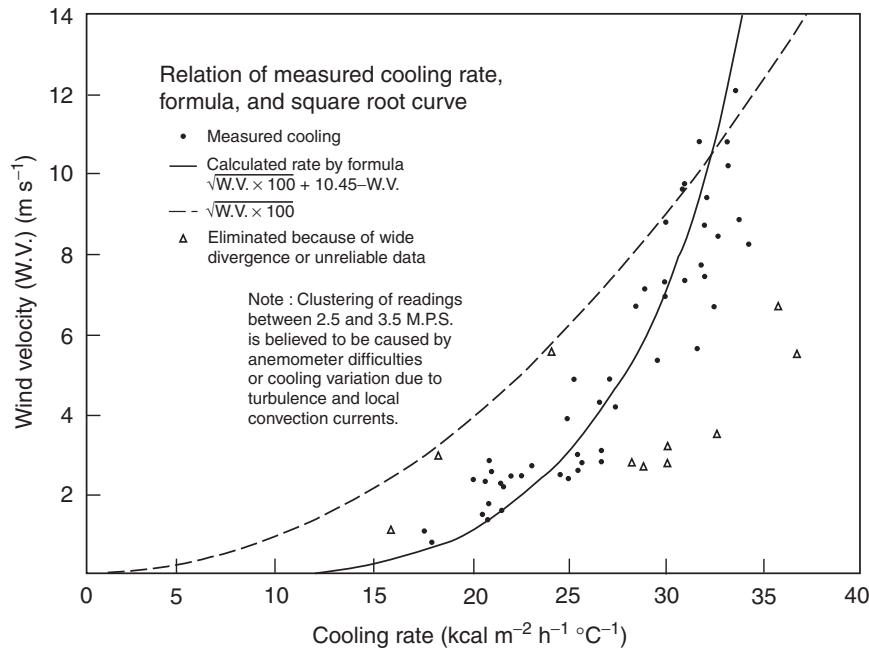


Figure 2 Wind velocity versus total heat transfer coefficient from Siple and Passel. (Reproduced with permission from Siple and Passel (1945).)

In Canada, only the left side of eqn [3], equal to the heat loss rate, was utilized. Since the body surface area can vary, the heat loss rate was expressed per unit area. The value of h is determined from eqn [2]. The air temperature is subtracted from a skin temperature of 33 °C. The result was called the wind chill index (WCI) and the units were converted to watts per square meter. Understandably, most citizens of Canada had difficulty with such an index.

In the dissemination of the original wind chill chart, the NWS had stated that no specific rules exist for determining when wind chill becomes dangerous. As a general rule, they stated that the threshold for potentially dangerous wind chill conditions was about -29 °C (-20 °F). Mitigating circumstances such as strong sunshine would require colder threshold temperatures.

Problems with the Wind Chill Calculation

A number of researchers expressed concern about the procedure used to calculate the WCT. Most of these related to the method used by S & P. The water in their container would not have frozen at the same rate throughout; using only one thermohm could create variability in the data. The data shown in Figure 2 are quite variable. In addition, the assumption of a container surface temperature of 0 °C ignored the thermal resistance of the container. Heat transfer theory shows that the surface temperature would have been below freezing. Perhaps the greatest problem with S & P's approach is the choice of 33 °C for human skin temperature. Human skin exposed to very cold temperature drops well below this value rapidly. If eqn [3] is solved for WCT, one obtains eqn [4].

$$WCT = \frac{h}{h_0}(T_a - T_s) + T_s \quad [4]$$

Since h is always greater than h_0 , as T_s is lowered, WCT is increased. This gave rise to a consensus among researchers that the first WCT chart exaggerated the effect of the wind by providing WCTs that were lower (colder) than they should be.

Another problem with that WCT chart had a more modern root. Wind speeds are recorded at official weather stations 10 m above the ground. Wind speeds at ground level are significantly less, owing to surface effects and structures such as buildings and trees. This means that humans are exposed to less wind than is reported and thus further results in higher WCTs.

Other problems cited include extrapolation of the results for a small container to the larger human body, and the difference between the thermal resistance of the plastic container and that of human skin.

Even with all these problems, it is fair to say that the WCT and the WCI have served a worthwhile purpose for many years. They have made the public aware of the effect of wind on winter comfort and have encouraged the citizenry to respect the dangers of low temperatures combined with high wind speeds.

While the wind chill concept has made the public aware of the dangers of high wind speed combined with cold temperatures, many persons misunderstand what the WCT actually represents. Many believe that if an object such as their automobile is left outside for a long period it will reach the WCT. The public should be led to understand that a low WCT means the object will get to the actual air temperature faster: If the air temperature is above freezing, the skin will not freeze even though the WCT is well below freezing. As a possible remedy,

the current NWS protocol calls for reporting the WCT only when the actual air temperature is at freezing or below.

Proposals to Improve the Wind Chill Determination

A major advance in the attempts to improve the accuracy of the WCT or WCI was made by Oszcewski in Canada in 1995. He created an instrumented model of the human head, surmising that the head is the most often exposed part of the body in winter. The model included a power source to maintain internal temperature similar to that of a human head and temperature sensors at the model's surface. The part of the model considered as representing the human face was turned into a wind generated in a wind tunnel. The resultant heat transfer effect was similar to that of an upright cylinder in a crosswind. Such a cylinder has been extensively studied in the heat transfer literature. Bluestein and Zecher used a similar approach but with a theoretical analysis only. They used a full cylinder to represent the adult head and produced WCTs that were higher than the original National Weather Service chart.

Oszcewski felt that the face rather than the entire head should be the area of interest in determining wind chill effects. Thus, only the front side of the equivalent cylinder would be considered in evaluating the WCI and WCT. Heat transfer experiments have shown that at low wind speeds, effectively the calm condition, the heat loss from the front of such a cylinder (the surface facing the wind) is greater than that from the rear surface. Thus, when the face rather than the entire head is considered (a half-cylinder rather than a whole cylinder), the value of h_0 in eqn [3] tends to be higher. This makes the WCT higher (less cold). When coupled with a more realistic skin surface temperature, the resultant WCT values were higher than those in the chart derived from the work of S & P and higher than those of Bluestein and Zecher.

As noted above, wind speeds are recorded at 10 m above the ground; modern researchers have sought to take this factor into account in modifying the WCT. A formula developed by Steadman has shown that wind speed at the level of the typical adult face in an open area is about two-thirds of that reported at the weather station; forests and urban settings have an even greater reduction. Thus, calculated WCTs should use a corrected wind speed, resulting in still higher values for the WCT.

Another consideration in the improvement of the determination of the WCT is the addition of the sunshine factor. Clearly, bright sun adds a warming effect to exposed skin and should perhaps be included in the WCT determination. Even partial sun, or partly sunny (or cloudy) conditions, can affect the heat loss from the skin. This effect would depend on the time of year and the latitude, and so would add some complexity to the calculation of the WCT. Current meteorological capability exists to evaluate this factor and may help the public to better evaluate environmental conditions.

Many researchers believe that wind chill should be evaluated for a whole-body human model, including clothing and, in some cases, an assumed metabolic activity rate. One of the most complete analyses has been done by Steadman but his formula is complicated. His work results in an Apparent Temperature (AT), which has the advantage of being applicable

for both cold and hot weather conditions. This is certainly a valuable research tool but may be too complex for use in weather forecasts for the general public.

Recent publications have put forth new charts of WCT as a function of air temperature and wind speed as recorded at 10 m height. Quayle has composed graphs to show how the data from Bluestein and Zecher, Oszcewski, and Steadman compare with the original NWS chart. One typical graph is shown in Figure 3 for -6.7°C (20°F) air temperature. The Oszcewski and Steadman results have the warmest WCT, with Bluestein and Zecher's results somewhat midway between those and the NWS values. These comparative results, with all other measures warmer than the S & P values, are the same for all combinations of air temperature and wind speed. Regardless of which of these new charts is considered, it seems clear that the original NWS chart exaggerated the effect of the wind.

Development of the Current WCT Chart

As result of an Internet conference organized by Canada in the spring of 2000, most meteorological organizations recognized the need to improve the method of wind chill determination. The NWS convened a joint action group to study temperature indices (JAG/TI) and to make a recommendation on the WCT. The weather service has indicated its desire to implement any improvements within 2 years. In addition, the International Society of Biometeorology (ISB) has formed a task force to study temperature indices as applicable to the international community. Some members of the JAG/TI are also members of this task force. Two international meetings were held in 2001 to determine the appropriate way to deal with wind chill.

The international community will be studying temperature indices with particular concentration on the following issues:

1. Can a temperature index be developed for the entire range of heat exchange?
2. Will the index be valid for all climates, seasons, and scales?

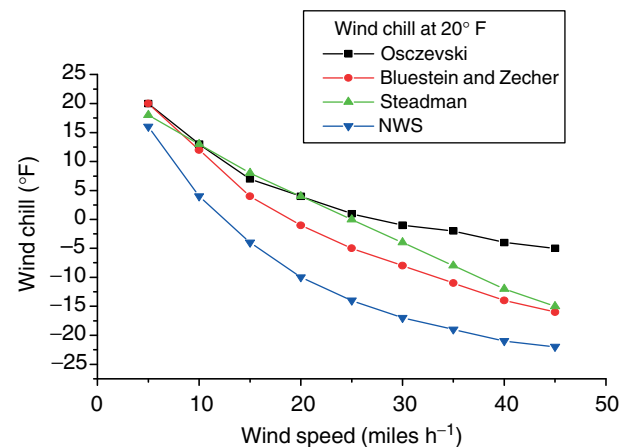


Figure 3 Comparison of published wind chill temperature scales for 20°F air temperature.

3. Will the index be useful in forecasting weather and other applications?
4. Will the index be independent of personal characteristics such as clothing and work activity?

A question to be answered is whether the WCT should be utilized separately from an inclusive temperature index.

The JAG or TI group charged Osczevski and Bluestein to develop a new chart as a compromise between their original models. Human trials were conducted in Canada with 12 instrumented volunteers in a wind tunnel. This resulted in a more accurate determination of skin conduction resistance. The researchers then applied modern heat transfer theory to a model of a half-cylinder representing the face turned into the wind, with a new still air condition of 1.34 ms^{-1} , (4.8 km h^{-1}), a more realistic walking speed. Heat is conducted from the internal body temperature of 37°C through the skin and then leaves the skin by convection and radiation to the ambient air. As a worst-case situation, the person is assumed to be walking at 1.34 ms^{-1} into the wind.

Using the method of Churchill and Bernstein, the rate of heat loss for a given air temperature and wind speed can be determined; an iterative solution using progressive estimates of the skin temperature is required. A WCT is then determined that will yield this same heat loss rate under still air conditions. To prepare the WCT charts, a regression equation was developed to match the data points at increments of 5° and five speed differentials. These equations are shown with the resultant charts in [Figure 1](#). The WCTs herein are higher (warmer) than those of the original chart and are deemed to be more accurate. Note that a correction for wind speeds measured at 10 m elevations has been included.

The new WCT charts include a warning of frostbite by time of exposure under severe conditions. The warning is based on exposed skin reaching -4.8°C , at which the risk is for frostbite in at least 5% of a population. This is based on the work of Danielsson.

See also: Weather Forecasting: Operational Meteorology; Severe Weather Forecasting.

Further Reading

- Bluestein, M., Zecher, J., 1999. A new approach to an accurate wind chill factor. *Bulletin of the American Meteorological Society* 80, 1893–1899.
- Court, A., 1997. Comfort temperatures. In: *McGraw-Hill Encyclopedia of Science and Technology*, vol. 4. McGraw-Hill. pp. 217–218.
- Dixon, J.C., Prior, M.J., 1987. Wind-chill indices — a review. *Meteorological Magazine* 116, 1–16.
- Driscoll, D.M., 1985. Human health. In: Houghton, D.D. (Ed.), *Handbook of Applied Meteorology*. Wiley, New York, pp. 778–814.
- Osczevski, R.J., 2000. Windward cooling: an overlooked factor in the calculation of wind chill. *Bulletin of the American Meteorological Society* 81, 2975–2978.
- Quayle, R.G., Steadman, R.G., 1998. The Steadman wind chill: an improvement over present scales. *Weather Forecasting* 13, 1187–1193.
- Siple, P.A., Passel, C.F., 1945. Measurements of dry atmospheric cooling in sub-freezing temperatures. Reports on Scientific Results of the United States Antarctic Service Expedition, 1939–1941. *Proceedings of the American Philosophical Society* 89, 177–199.
- Steadman, R.G., 1971. Indices of wind chill of clothed persons. *Journal of Applied Meteorology* 10, 674–683.
- Steadman, R.G., 1995. Comments on “Wind chill errors”. *Bulletin of the American Meteorological Society* 9, 1628–1630.
- Wyon, D.P., 1989. Wind-chill equations predicting whole-body heat loss for a range of typical civilian outdoor clothing ensembles. *Scandinavian Journal of Work, Environment and Health* 15 (supplement 1), 76–83.

Standard Atmosphere

WW Vaughan, University of Alabama in Huntsville, Huntsville, AL, USA

© 2003 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 5, pp 2107–2113, © 2003, Elsevier Ltd.

Introduction

A 'standard atmosphere' is a vertical description of atmospheric temperature, pressure, and density that is usually established by international agreement and taken to be representative of the Earth's atmosphere. The first 'standard atmospheres' established by international agreement were developed in the 1920s primarily for the purposes of pressure altimeter calibrations and aircraft performance calculations. Later, some countries, notably the United States, also developed and published 'standard atmospheres.' The term 'reference atmosphere' is used to identify vertical descriptions of the atmosphere for specific geographical locations or globally. These were developed by organizations for specific applications, especially as the aerospace industry began to mature after World War II. The term 'standard atmosphere' has in recent years also been used by national and international organizations to describe vertical descriptions of atmospheric trace constituents, the ionosphere, aerosols, ozone, atomic oxygen, winds, water vapor, planetary atmospheres, and so on.

A standard unit of atmospheric pressure is defined as that pressure exerted by a 760 mm (or 29.22 inch) column of

mercury at standard gravity at 45.5425° N latitude and sea level (9.80665 m s^{-2}) at a temperature of 0 °C (32 °F). The recommended unit for meteorological use is 1013.25 hPa (1 hPa = 1 mb). Standard temperature is used in physics to indicate a temperature of 0 °C (32 °F), the ice point, and a pressure of one standard atmosphere (1013.25 hPa). In meteorology, the term standard temperature has no generally accepted meaning, except that it may refer to the temperature at zero pressure–altitude in the standard atmosphere (15 °C) with a density of 1.2250 g m^{-3} . The standard sea-level values of temperature, pressure, and density that have been used for decades are temperature of 288.15 K, 15 °C, or 59 °F; pressure of 1013.25 mb, 760 mmHg, or 29.22 inches Hg; and density of $1.225.00 \text{ g m}^{-3}$ or $0.076474 \text{ lb ft}^{-3}$.

In 1925, the US National Advisory Committee for Aeronautics (NACA) Standard Atmosphere (or US Standard Atmosphere) was published. In 1952, the International Civil Aeronautical Organization (ICAO) produced the ICAO Standard Atmosphere, and in 1964 an extension to 32 km. Subsequently, there have been a succession of 'Standard and Reference Atmospheres,' some extending to altitudes above

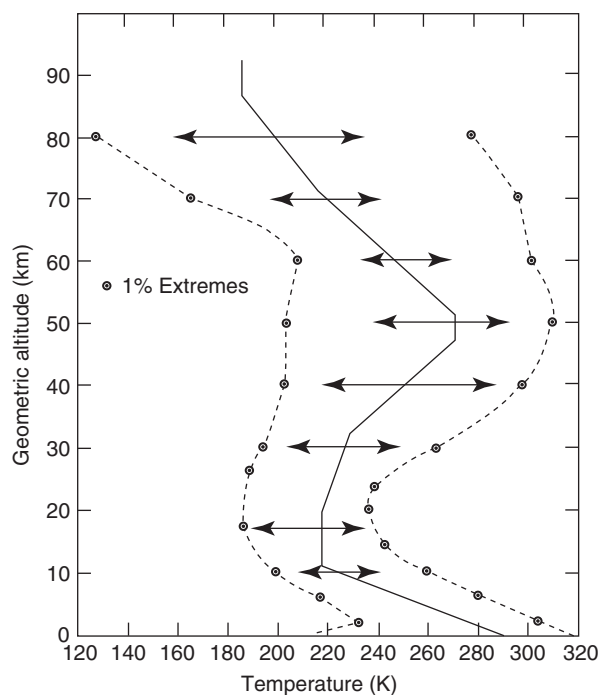


Figure 1 Range of systematic variability of temperature around the US Standard Atmosphere, 1976. Reproduced from Sissenwine, N., Dubin, M., Teweles, S., (COESA Co-Chairmen), 1976. US Standard Atmosphere, 1976, Stock No. 003-017-00323-0. US Government Printing Office, Washington, DC.

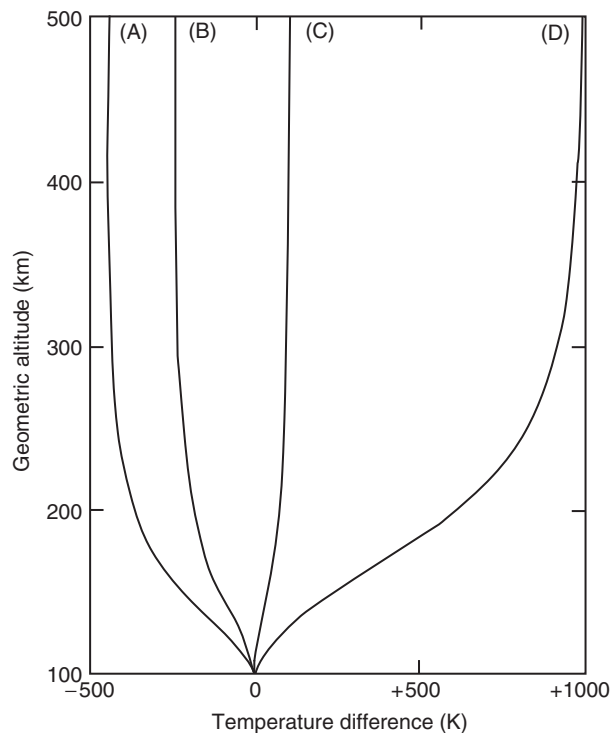


Figure 2 Departures of the temperature–altitude profiles from that of the US Standard Atmosphere, 1976, for various degrees of solar activity. Reproduced from Sissenwine N, Dubin M, Teweles S (COESA Co-Chairmen), 1976. US Standard Atmosphere, 1976, Stock No. 003-017-00323-0. US Government Printing Office, Washington, DC.

Table 1 Summary of reference and standard atmospheres

<i>Model (page no.)</i>	<i>Geographic region</i>	<i>Altitude range (km)</i>	<i>Parameters</i>	<i>Species included</i>	<i>Temporal variation</i>	<i>Output data present</i>	<i>Principal application</i>
CIRA, 1972 (1)	Northern latitude Global	25–120, 110–2000	T, p, d , composition winds	N_2, O_2, O, A, He, H	Seasonal, diurnal, solar activity, magnetic activity	Tables, figures	Aerospace vehicle design and evaluation, atmospheric reference
CIRA, 1986 (3)	Global	130–2000	T, p, d , composition		Seasonal, solar activity, geomagnetic activity	Tables, figures, computer code	Aerospace vehicle design and evaluation, atmospheric reference
New Middle Atmosphere, 1985 (5)	Global 80° S–80° N	20–80	T, p, d , zonal	–	Monthly, interannual, tidal, planetary wave	Tables, figures	Aerospace vehicle design and evaluation, atmospheric reference
ISO Reference Atmosphere, 1982 (7)	Annual – 15° N Seasonal – 30° N, 45° N, 60° N, 80° N Cold/warm middle atmosphere – 60° N, 80° N	0–80	T, p, d	Data on water vapor	Seasonal, diurnal, daily	Tables, figures	Aerospace vehicle and aircraft design and performance studies, atmospheric reference
ISO Standard Atmosphere, 1975 (9)	45° N	–2 to 80	T, p, d , composition, sound speed, coll. freq. mfp, viscosity, spec. wt, scale ht, therm. cond.	–	–	Tables only	Aerospace vehicle design and performance studies, atmospheric reference
Monthly Mean Global Climatology, 1988 (11)	Global	0–120	T, p , zonal winds	–	–	Tables only	Reference Climatology, numerical model initialization, instrumental design, scientific studies
GRAM-95 (13) (Current Edition: GRAM-99)	Global coverage	0–2500	T, p, d , wind velocity, wind shear, composition	$H_2O, N_2, O, CH_4, N_2, O, He, O_3, CO, CO_2, O_2, A, H$	Random perturbation, monthly	Computer code NASA–MSFC and COSMIC	Aerospace vehicle design and simulation studies, space vehicle reentry, atmosphere reference for scientific studies
US Standard, 1962 (16)	Midlatitudes (45°)	–5 to 700	T, p, d , composition, part. speed, coll. freq., mfp, mean mol. wt, viscosity, therm. cond., sound speed	–	–	Tables, figures	Aerospace vehicle design, atmospheric reference
US Standard, 1966 Supplement (18)	Midlatitudes with variation	–5 to 1000	Same as USS 1962	O_2, N_2, O, He, H	Seasonal, diurnal, solar activity, magnetic activity	Tables, figures	Illustrate atmospheric variability
US Standard, 1976 (19)	Midlatitudes (45°)	–5 to 1000	Same as USS 1962	Some data on N_2, O_2, H, He, O	Diurnal, seasonal, solar cycle	Tables, figures	Aerospace vehicle design, atmospheric reference

(Continued)

Table 1 Summary of reference and standard atmospheres—cont'd

<i>Model (page no.)</i>	<i>Geographic region</i>	<i>Altitude range (km)</i>	<i>Parameters</i>	<i>Species included</i>	<i>Temporal variation</i>	<i>Output data present</i>	<i>Principal application</i>
International Tropical Reference Atmosphere 1987 (21)	Tropics	–5 to 1000	T, p, d , composition, part, speed, coll. freq., mean mol. wt, viscosity, therm, cond., sound speed	N ₂ , O ₂ , O, Ar, He	None	Tables, figures	Aerospace vehicle design studies, atmospheric reference
Reference Atmosphere for Indian Equatorial zone, 1985 (23)	Tropics	0–80	T, p, d	–	Monthly, annual	Tables, figures	Design of aerospace vehicles, science applications
Reference Model Middle Atmosphere Southern Hemisphere 1987 (24)	South 0–70° S	20–80	T, p, d , zonal winds	–	Monthly, latitudinal	Tables, figures	Aerospace vehicle design, atmospheric reference
AFGL (Phillips Laboratory) Atmospheric Constitution Profiles, 1986 (26)	Global coverage	0–120	Number density, aerosol properties	H ₂ O, CO ₂ , N ₂ O, O ₃ , CH ₄ , CO, O ₂ , N ₂ , 20 others, aerosols	None	Tables, figures, computer code	Design and performance evaluation, scientific studies
Extreme Envelope of Climate Elements 1973 (28)	60° S–90° N	0–80	Climatic elements: T, p , humidity, wind shear, etc.	–	Monthly	Tables, figures	Systems design
Profiles of Temperature and Density, 1984 (30)	Global except Antarctic	0–80	T, d	–	Monthly	Tables, figures	Systems design
Global Reference Atmosphere, 1985 (32)	Global	18–80	T, p, d , number density, scale ht Wind velocity	–	Monthly	Tables, figures	Reference model for scientific studies
Earth's Upper Atmosphere Density Model (Russia), 1984 (33)	>120 km solar flux dependent	0–1500	d	–	Solar flux, geomagnetic activity, daily and semiannual effects	Tables, computer code	Aerospace vehicle design and orbital lifetimes
Jacchia J70 (34)	Mean global	90–2500	T, p, d , scale ht	N ₂ , O ₂ , O, Ar, He, H	Diurnal, seasonal, geomagnetic activity	Tables	Design and simulation, lifetime analysis
Jacchia J71 (35)	Mean global	90–2500	T, p, d , scale ht	N ₂ , O ₂ , O, Ar, He, H	Diurnal, seasonal, geomagnetic activity	Tables, some computer code	Design and simulation, lifetime analysis
Jacchia J77 (36)	Mean global	90–2500	T, p, d , scale ht	N ₂ , O ₂ , O, Ar, He, H	Diurnal, seasonal, geomagnetic activity	Tables, some computer code	Design and simulation, lifetime analysis
Model of Atmospheric Structure, 1987 (38)	Global	70–130	T, p, d	–	Monthly latitudinal, solar activity, magnetic activity	Tables	Connect Phillips Lab (AFGL) profiles of T, p to MSIS-86

NASA MSIS-86 (39) (Current Edition: NRL-MSIS-00)	Global coverage	85–2000	T, p, d , composition	$N_2, O_2, O, He, H,$ Ar, N	Diurnal, semiannual, latitudinal longitudinal solar activity, magnetic activity	Computer code (NSSDC), floppy disk	General scientific and engineering studies
NASA Marshall Engineering Thermospheric Model, 1988 (41) (Current Edition: Version 2.0)	Global	90–2500	T, p, d , mean mol. wt, scale ht, spec. heat	N_2, O_2, O, Ar, He, H	Solar activity, magnetic activity, seasonal, diurnal	Computer code (NSSDC), floppy disk	Orbital vehicle design and simulation, lifetime analysis
Range Reference Models of the Atmosphere, 1982 (43)	Specific locations (e.g., Cape Canaveral, FL; Kwajalain, MI, etc.)	0–70	T, p, d , wind velocity	Water vapor	Monthly, seasonal, means, monthly, parameter variations	Tables, figures	Site-related engineering analyses
Reference Atmosphere for Edwards AFB, CA, 1975 (46)	Edwards/Dryden only	← Same as Reference Atmosphere for Patrick AFB →					
Hot and Cold Atmosphere for Edwards AFB, CA, 1975 (47)	Edwards/Dryden only	← Same as Hot and Cold Atmosphere for Kennedy Space Center →					
Hot and Cold Atmosphere for Kennedy Space Center, FL, 1971 (48)	Kennedy Space Center only	0–90	T, p, d	–	Seasonal	Tables, figures	Engineering studies
Reference Atmosphere for Patrick AFB, FL, 1963 (49)	Cape Kennedy only	0–700	T, p, d , composition, mean mol. wt, sound speed, viscosity, etc.	–	–	Tables, figures	Engineering studies
Reference Atmosphere for Vandenberg AFB, CA, 1971 (50)	Point Arguello only	← Same as Reference Atmosphere for Patrick AFB →					
Hot and Cold Atmosphere for Vandenberg AFB, 1973 (51)	Arguello only	← Same as Hot and Cold Atmosphere for Kennedy Space Flight Center →					
Mars-GRAM, 1996 (52)	Global	0 to ~1000	T, p, d , winds	–	Seasonal, diurnal, altitudinal longitudinal	Tables, computer code	Spacecraft design, atmospheric entry, orbital drag
Venus International Reference Atmosphere (VIRA), 1985 (53)	Global	0–3500	T, p, d , composition	<100 km $CO_2, N_2,$ Ar, Ne, Kr, $O_2, H_2,$ H_2O, SO_2, D, NH_3 >100 km $CO_2, O,$ CO, He, N, $N_2, H,$ O_2, D, C	<100 km latitudinal solar zenith angle, diurnal >100 km solar zenith angle, decimal, latitudinal, solar activity	Tables, figures	Spacecraft design, atmospheric entry, orbital drag

T = kinetic temperature; p = pressure; d = mass density; mfp = mean free path; part. speed = particle speed; coll. freq. = collision frequency; mean mol. wt = mean molecular weight; therm. cond. = thermal conductivity; scale ht = scale height; spec. wt = specific weight; and spec. heat = specific heat.

CIRA: COSPAR (Committee on Space Research) International Reference Atmosphere; ISO: International Organisation for Standardization; GRAM: Global Reference Atmosphere Model; AFGL: Air Force Geophysics Laboratory; NASA: National Aeronautics and Space Agency; MSIS: Mass Spectrometer and Incoherent Scatter; and NRL: Naval Research Laboratory.

Reproduced from Vaughan, W.W., Johnson, D.L., Justus, C.G., et al., 1996. Guide to Reference and Standard Atmosphere Models, Document ANSI/AIAA G-003A-1996. American Institute of Aeronautics and Astronautics, Reston, VA.

1000 km, produced by the US Committee on Extension to the Standard Atmosphere (COESA), Committee on Space Research (COSPAR), Comitet Standartov (USSR), International Standardization Organization (ISO), US Air Force Research and Development Command (ARDC), US Range Commanders Council (RCC), and US National Aeronautics and Space Administration (NASA), plus others.

In 1975, the International Standards Organization published a Standard Atmosphere for altitudes from -2 to 50 km that is identical to the ICAO Standard Atmosphere from -2 to 32 km. Subsequently, the ISO published in 1982 a family of five Reference Atmospheres for Aerospace Use for altitudes up to 80 km and latitudes of 15° N, 30° N, 45° N, 60° N, and 80° N.

Figure 1 provides an illustration of the temperature–height profiles to 100 km of the COESA US Standard Atmosphere, 1976, and the lowest and highest mean monthly temperatures obtained for any location between the Equator and Pole. The portion of the US Standard Atmosphere up to 32 km is identical with the ICAO Standard Atmosphere, 1964, and below 50 km with the ISO Standard Atmosphere, 1973.

For altitudes above approximately 100 km, significant variations in the temperature, and thus density, occur due to solar and geomagnetic activity over the period of a solar cycle. Variations in the temperature–height profiles for various degrees of solar and geomagnetic activity are presented in **Figure 2**. Profile (a) gives the lowest temperature expected at solar cycle minimum; profile (b) represents average conditions at solar cycle minimum; profile (c) represents average conditions at a typical solar cycle maximum; and profile (d) gives the highest temperatures to be expected during a period of exceptionally high solar and geomagnetic activity.

Currently, some of the most commonly used Standard and Reference Atmospheres include the following:

- ICAO Standard Atmosphere, 1952/1964;
- ISO Standard Atmosphere, 1973;
- US Standard Atmosphere, 1976;
- COSPAR International Reference Atmosphere; (CIRA), 1986; and
- NASA Global Reference Atmosphere Model (GRAM), 1999.

In 1996, the American Institute of Aeronautics and Astronautics (AIAA) published a *Guide to Reference and Standard Atmosphere Models*. This document provides information on the principal features for a number of global, regional, middle atmosphere, thermosphere, test range, and planetary atmosphere models. Summary information on these reference and standard atmosphere models is given in **Table 1**.

See also: Dynamical Meteorology: Static Stability. Solar System/Sun, Atmospheres, Evolution of Atmospheres: Evolution of Earth's Atmosphere.

Further Reading

- Champion, K.S.W., 1995. Early Years of Air Force Geophysics Research Contributions to Internationally Recognized Standard and Reference Atmospheres. Technical Report PL-TR-95-2164. Air Force Phillips Laboratory, Hanscom AFB, MA.
- Sissenwine, N., Dubin, M., Teweles, S., (COESA Co-Chairmen), 1976. US Standard Atmosphere. Stock No. 003-017-00323-0. US Government Printing Office, Washington, DC.
- Vaughan, W.W., Johnson, D.L., Justus, C.G., et al., 1996. Guide to Reference and Standard Atmosphere Models. Document ANSI/AIAA G-003A-1996. American Institute of Aeronautics and Astronautics, Reston, VA.

AEROSOLS

Contents

Aerosol–Cloud Interactions and Their Radiative Forcing

Aerosol Physics and Chemistry

Climatology of Stratospheric Aerosols

Climatology of Tropospheric Aerosols

Dust

Observations and Measurements

Role in Radiative Transfer

Role in Climate Change

Soot

Aerosol–Cloud Interactions and Their Radiative Forcing

U Lohmann, ETH Zurich, Zürich, Switzerland

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by S M Kreidenweis, volume 1, pp 40–47, © 2003, Elsevier Ltd.

Synopsis

Aerosols are essential for cloud formation. Every cloud droplet needs an aerosol particle, called cloud condensation nucleus, for activation. Likewise ice crystals either form on a subset of aerosol particles that act as ice nuclei or form by homogeneous freezing of supercooled solution drops. Anthropogenic aerosols cause an indirect radiative forcing by modifying cloud properties. This indirect radiative forcing has uncertainties that are larger than for any other forcing.

Introduction

Aerosols are essential for cloud formation. Every cloud droplet needs an aerosol particle, called cloud condensation nucleus (CCN), for activation. Likewise ice crystals either form on a subset of aerosol particles that act as ice nuclei (IN) or form by homogeneous freezing of supercooled solution drops (liquid aerosols that took up water). In the absence of aerosols, several 100% relative humidity (RH) would be necessary for cloud droplets to form by homogeneous nucleation from supersaturated water vapor. Due to the ubiquitous presence of CCN, the RH in water clouds hardly exceeds 101%. The situation is different for ice clouds because IN are sparse, and formation of cirrus clouds is dominated by homogeneous freezing of solution droplets. This takes place at relative humidities below 100% with respect to water but well above 100% with respect to ice (at $-60\text{ }^{\circ}\text{C}$ the RH for homogeneous freezing is 150%). At the same time, removal of aerosols by clouds and precipitation is the largest sinks for aerosols with diameters $<1\text{ }\mu\text{m}$. Thus the lifetime of aerosols is strongly linked to that of clouds.

Atmospheric Aerosols

An aerosol is defined as a disperse system with air as the carrier gas and a solid or liquid disperse phase or a mixture of both. In atmospheric science, it is common to use the term ‘aerosol’ just for the solid or liquid particles and neglect the carrier gas. Aerosol particles range from 1 nm to several hundred micrometers in diameter. Aerosol particles can be as large as cloud droplets or ice crystals. Whereas cloud droplets or ice crystals only occur in isolated patches, aerosol particles, especially the smaller ones, are rather homogeneously mixed in the atmosphere.

Aerosols have a variety of different formation mechanisms. They are divided into primary and secondary particles. Primary particles are already emitted as aerosol particles (either liquid or solid) into the atmosphere. Primary aerosols can result from bulk-to-particle conversion, such as wind blown dust from arid regions, or emissions of pollen and spores by plants or can originate from combustion processes. Mineral dust is mainly emitted in arid regions. Its emission rate depends on wind speed, soil moisture and the bare soil fraction. Sometimes

liquid-to-particle conversion is used for the formation of sea-salt aerosols. Sea-salt aerosols originate from bursting of air bubbles that reach the ocean surface. The bursting bubbles leave water droplets containing sea salt behind. Upon evaporation of the water droplet, a sea-salt particle is released to the atmosphere. Primary aerosols are mainly large particles of natural origin and thus account for the largest aerosol mass in the atmosphere.

Primary particles are distinguished from secondary aerosols, which form within the atmosphere from precursor substances (gases) by gas-to-particle conversion and form small aerosol particles. Examples of secondary aerosols in the atmosphere are sulfates, nitrates and secondary organic aerosols. Secondary particles have both natural and anthropogenic sources. Gas-to-particle conversion requires a nucleation process because a new phase (liquid or solid) is formed from a supersaturated gas phase. The most common form of particle nucleation is binary nucleation where the aerosol forms from two gas phase precursors. Homogeneous self-nucleation of a single species does not take place in the atmosphere as that would require that the vapor pressure of a single species is supersaturated. For binary nucleation to take place, each species can be subsaturated if the mixture is supersaturated. A prominent example of binary nucleation is the reaction of n moles water vapor H_2O (g) with m moles of gaseous sulfuric acid H_2SO_4 (g), resulting in the nucleation of liquid sulfuric acid aerosol particles: $(\text{H}_2\text{O})_n(\text{H}_2\text{SO}_4)_m$ (aq), where aq denotes the aqueous phase.

A schematic representation of aerosol binary nucleation of H_2O and H_2SO_4 with subsequent growth to larger sizes is shown in Figure 1. The first step is that a cluster of H_2O and H_2SO_4 is formed in the gas phase. Once the cluster exceeds its critical size, nucleation occurs and a stable sulfuric acid aerosol particle is formed. The critical size of a cluster is the size at which the droplet can exist in equilibrium with its vapor phase. Once the particle is formed, also other species, such as organics of low volatility can condense onto the particle and participate in its growth. Low volatility is

necessary as otherwise the organics would not remain in the condensed phase and evaporate again. The aerosol particle also grows by coagulation with other particles. The newly formed aerosol particle needs to increase 1–2 orders of magnitude in size before it can act as a CCN and be involved in cloud formation.

Hygroscopic Growth

Condensation of water vapor on aerosol particles (water uptake) is important for aerosol particles with an affinity for water vapor, i.e., for hydrophilic or hygroscopic aerosols, such as sulfate, nitrate, sea-salt, or mineral dust particles. This swelling of aerosol particles needs to take place before aerosol particles can act as CCN. Water uptake can cause a phase change if the soluble aerosol was solid before. If the aerosol is already liquid, then water uptake just leads to further growth and dilution of the salt solution. The phase change of the solid soluble aerosol to a liquid aerosol is called deliquescence and does not require a nucleation process. The opposite process, the formation of a solid aerosol, is called efflorescence or crystallization and requires a nucleation process. Nucleation requires overcoming an energy barrier, which can only be achieved if prior to nucleation the nucleating substance is in a supersaturated state as in the case of binary nucleation of H_2O and H_2SO_4 .

The growth of the aerosol is described in terms of the growth factor, which is the ratio of the actual aerosol diameter to its diameter at 0% RH. It is only defined for RHs below 100%. Some aerosols such as sulfuric acid (H_2SO_4) remain liquid at all relative humidities (Figure 2) at atmospheric temperatures and pressures and just change their size according to RH. Other aerosols, predominantly salts such as ammonium sulfate or ammonium bisulfate, change their phase. At which RH this phase change occurs depends on the direction of the change in RH. In case of ammonium sulfate, the sudden increase in diameter by 50% occurs at 80% RH. This increase in

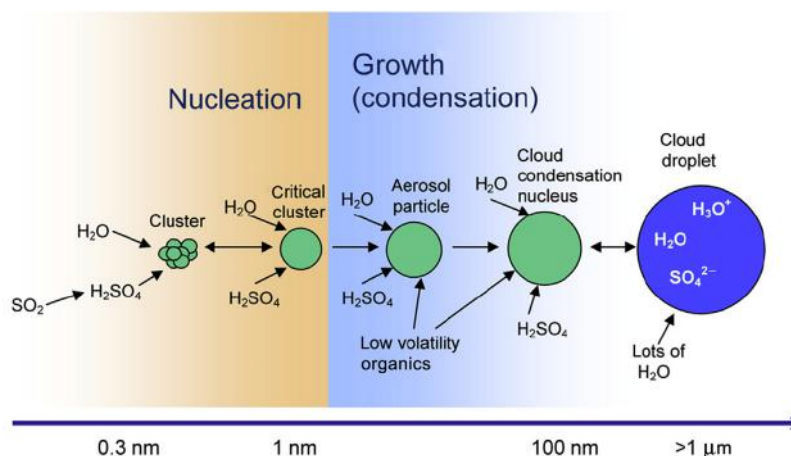


Figure 1 Schematic representation of the nucleation and subsequent growth process for atmospheric binary homogeneous nucleation of H_2SO_4 and H_2O . Once stable clusters are formed, also other substances such as organics can take part in the growth process. Particles may grow to sizes large enough to act as CCN on which cloud droplets may form eventually. Figure reproduced with courtesy from Curtius, J., 2006. Nucleation of atmospheric aerosol particles. *C. R. Phys.* 7, 1027–1045.

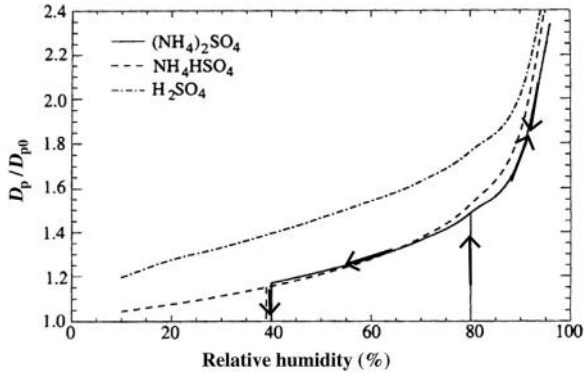


Figure 2 Water uptake of sulfuric acid (H_2SO_4), ammonium bisulfate (NH_4HSO_4), and ammonium sulfate ($(\text{NH}_4)_2\text{SO}_4$) aerosols expressed in terms of their growth factor, which is the ratio of the actual diameter D_p to the dry diameter at 0% RH (D_{p0}), as a function of RH. Figure reproduced with courtesy from Seinfeld, J.H., Pandis, S.N., 1998. *Atmospheric Chemistry and Physics: From Air Pollution to Climate Change*. Wiley, 1326 pp.

diameter goes hand in hand with a phase change from solid to liquid and is referred to as the deliquescence relative humidity. Here the salt dissolves in water.

If RH is decreased, ammonium sulfate remains in the liquid phase until 40% RH, where it suddenly solidifies. This RH is called the crystallization relative humidity. The growth factor of ammonium sulfate exhibits a hysteresis, meaning the value of the growth factor is not unambiguously determined by RH, but depends also on the history of the particle. Between 40 and 80% RH the liquid phase of ammonium sulfate is metastable. Energetically, the aerosol would prefer to be in the solid state but it is prevented from doing so because it first needs to surpass the energy barrier for nucleation.

Cloud Droplet Formation

Cloud droplet formation is not a nucleation process because cloud droplets form on soluble or hydrophilic aerosol particles, which already have taken up large amounts of water below 100% RH. Cloud formation is best described by the Koehler equation:

$$e^*(r)/e_s(\infty) = S = 1 + a/r - b/r^3 \quad [1]$$

The Koehler equation describes the ratio between the equilibrium vapor pressure over a solution droplet $e^*(r)$ with radius r to the saturation vapor pressure over a plane surface of water $e_s(\infty)$ with a and b as described below. $e_s(\infty)$ depends exponentially on temperature (T) as described by the Clausius–Clayperon equation:

$$de_s/dT = L_v e_s / (R_v T^2) \quad [2]$$

where R_v is the specific gas constant of water vapor ($R_v = 461.5 \text{ J K}^{-1} \text{ kg}^{-1}$) and L_v is the latent heat of vaporization. The Koehler equation has two contributions, the increase in vapor pressure that is associated with the formation of the surface (Kelvin equation) and the decrease in vapor pressure due to soluble substances in water (Raoult's law). The Kelvin equation is given as:

$$e_s(r)/e_s(\infty) = S = \exp(2\sigma/(r \rho_w R_v T)) \quad [3]$$

where σ is the surface tension between water and air. At 273.2 K, σ is 0.0756 N m^{-1} and ρ_w , the water density, is 1000 kg m^{-3} . The terms of the Kelvin equation other than the radius of the droplet are summarized in term a of the Koehler equation: $a = 2\sigma/(r \rho_w R_v T)$. Kelvin's equation describes that the equilibrium vapor pressure is larger over a droplet with radius r than over a plane or bulk surface. It inversely relates the critical radius for droplet formation to the necessary supersaturation. At $S = 1.01$, a typical value found in the atmosphere, the critical radius of the droplet needs to be $0.12 \text{ }\mu\text{m}$. This can never occur by chance as it involves 0.25 million water vapor molecules. A more reasonable number of 20 water vapor molecules forming a cluster with a critical radius of 0.5 nm requires a saturation ratio S of 10, i.e., an RH of 1000%, which does not exist in Earth's atmosphere.

The second contributor to the Koehler equation, Raoult's law, is given for a plane surface of water as follows:

$$e^*(\infty)/e_s(\infty) = n_0/(n + n_0) \quad [4]$$

where e^* is the equilibrium vapor pressure over a solution consisting of n_0 water molecules and n solute molecules. This equation shows that if the vapor pressure of the solute is less than that of water and if the total number of molecules remains constant, the vapor pressure over the solution is reduced in proportion to the amount of solute present. The vapor pressure reduction arises from solute molecules at the surface that limit the exchange of water molecules between the water surface and the overlying vapor to those places where water molecules occupy the surface. For dilute solutions and applied to droplets of size r , Raoult's law can be approximated as:

$$e^*(r)/e_s(r) = 1 - b/r^3, \quad \text{where } b = 3 i M_s m_w / (4\pi m_s \rho_w) \quad [5]$$

The competition between Kelvin's equation and Raoult's law is summarized in the Koehler equation. **Figure 3** displays the Koehler curves, i.e., the equilibrium vapor pressure as a function of the droplet radius, for droplets containing different amounts of salt. Note that every salt particle with a dry radius r_d has its own individual Koehler curve. Because

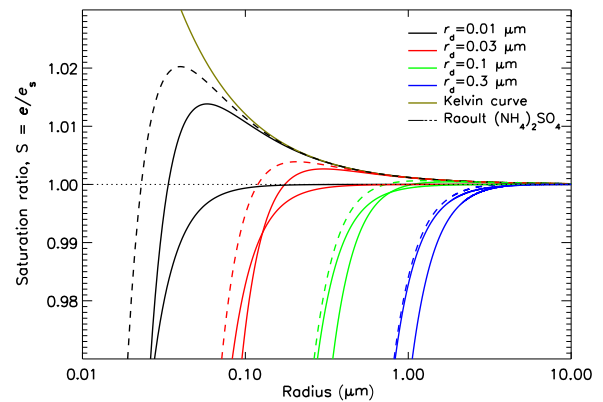


Figure 3 Koehler curves for NaCl (solid lines) and $(\text{NH}_4)_2\text{SO}_4$ (dashed lines) for droplets originating from salt particles with different dry radii r_d . In addition, the Kelvin curve (common for all particles) is shown in olive and Raoult's law for $(\text{NH}_4)_2\text{SO}_4$ as dot-dashed lines.

Raoult's law applies to the droplet volume, it dominates over the Kelvin effect, that depends on droplet size, at small radii and lowers the equilibrium vapor pressure at small radii below $S = 1$. The Kelvin effect is negligible for droplets larger than $1\ \mu\text{m}$ where all curves approach $S = 1$.

The peak in the saturation ratio called critical saturation ratio (S_c) and the corresponding critical radius (r_c) can be obtained by differentiating the Koehler equation with respect to r and setting the derivative to 0:

$$r_c = (3b/a)^{1/2}; \quad S_c = 1 + (4a^3/27b)^{1/2} \quad [6]$$

S_c of a specific Koehler curve is the minimum saturation ratio that is required for the corresponding solution droplet to grow to cloud droplet size. Theoretically, it could grow even larger, but the growth becomes increasingly slower the larger the droplet and more efficient growth mechanisms such as growth by collision-coalescence with other droplets will take over (see **Clouds and Fog: Cloud Microphysics**). If the droplet has grown to $r > r_c$ it is called an activated droplet. All droplets that have a critical saturation ratio $S_c < S$, where S is the supersaturation reached in the ambient air, can thus be activated. If $S < S_c$, a deliquesced salt particle can only grow to the radius at which the Koehler curve takes the value S .

The Koehler curve represents equilibrium conditions and therefore has some limitations of its applicability. Large particles have large equilibrium radii and may have insufficient time to grow to their equilibrium size in clouds that do not last long, such as convective clouds. **Figure 3** shows that the higher S , the more and the smaller aerosols can be activated. It also shows that the largest aerosols are the best CCN because they require the smallest supersaturations to be activated. However, there are only few of them available. Also, because the diffusional growth depends inversely on the size of the droplets with the smallest droplets growing the fastest (see **Clouds and Fog: Cloud Microphysics**), the large aerosols may not reach their critical size by diffusional growth. On the other hand, the smallest aerosols require higher supersaturations to activate than exist in the atmosphere. Thus, these aerosols are not efficient for cloud formation either. Mainly accumulation and coarse mode, and partly Aitken mode aerosols act as CCN and get activated into cloud droplets.

Ice Crystal Formation

Once a cloud extends to altitudes where the temperature is below $0\ ^\circ\text{C}$, ice crystals may form either by freezing of a cloud droplet or by direct deposition of the vapor to the solid phase. Both, freezing of cloud droplets and deposition of vapor to the solid phase, are nucleation processes because a stable cluster of the new ice phase has to form within the parent phase (vapor or liquid). Both homogeneous and heterogeneous nucleations are possible for the formation of ice crystals and have been observed in the atmosphere. Homogeneous freezing of liquid aerosols is the dominant freezing process in cirrus clouds. These liquid aerosols, also called supercooled solution droplets, are in equilibrium with the ambient RH below 100% with respect to water, but are not activated as cloud droplets. Heterogeneous ice nucleation, which requires the presence of

an IN, is the dominant ice nucleation process in mixed-phase clouds. Mixed-phase clouds are clouds that exist between 0 and $-40\ ^\circ\text{C}$ and consist of a mixture of supercooled cloud droplets and ice crystals.

IN are aerosol particles that provide a surface onto which water molecules can stick, bond together, and form aggregates with an ice-like structure. Differently from CCN, only a small fraction (one in $100\,000$ – $1\,000\,000$) of all aerosol particles can serve as IN at temperatures warmer than $-40\ ^\circ\text{C}$. Therefore, the criteria for aerosols to act as IN are less well understood as compared to the CCN ability of aerosol particles.

Homogeneous deposition nucleation of vapor to form ice crystals is analogous to homogeneous cloud droplet nucleation. Also here the vapor pressure increase over an ice sphere is increased as compared to bulk ice, the more the smaller the crystal is. It can thus be described with Kelvin's equation (see above) except that the parameters related to liquid water need to be replaced with those of ice. Homogeneous deposition nucleation is even less likely than homogeneous cloud droplet formation because of the higher surface tension between ice and air than between water and air and because the ice density is lower than the water density. Both differences increase the vapor pressure over an ice sphere more than that over a cloud droplet (see **eqn [3]**).

On the other hand, homogeneous freezing of ice within a liquid drop is observed in the atmosphere. It occurs when statistical fluctuations of the molecule arrangement of water produce a stable, ice-like structure, called ice germ. The formation of the surface of an ice-like structure requires an energy barrier to be overcome analogous to the crystallization of salts discussed above. If the ice germ exceeds the energy barrier for nucleation, it can grow spontaneously and cause the entire droplet to freeze. Experimental data on the freezing of pure water show that cloud droplets below a radius of $5\ \mu\text{m}$ will freeze spontaneously at temperatures below $-38\ ^\circ\text{C}$ **Pruppacher and Klett (1997)**. Larger droplets freeze at slightly warmer temperatures. Homogeneous freezing of pure water only occurs at or above water saturation (**Figure 4**) since only at these conditions pure water droplets can exist in equilibrium.

In cirrus clouds, the RH is not high enough to activate cloud droplets. The RH is, however, high enough for unactivated solution droplets to exist, which also can freeze homogeneously. Homogeneous freezing of solution droplets is the dominant pathway of ice crystal formation in cirrus clouds. Solution droplets refer to sulfuric acid or other liquid aerosols that have taken up water. The presence of salt causes the freezing point to be colder than that for pure water (freezing point depression). If the RH is not sufficiently high, the salt is too concentrated inside the solution droplets for freezing to occur. Therefore, homogeneous freezing of solution droplets occurs only at or above the dashed line that starts at $-35\ ^\circ\text{C}$ in **Figure 4**. Homogeneous freezing of solution droplets deviates from the water saturation more and more with decreasing temperature below $-35\ ^\circ\text{C}$ because solution droplets can freeze with higher salt concentrations and concentrated droplets can exist in equilibrium at lower relative humidities.

Heterogeneous freezing initiates freezing with the help of an IN. An IN favors freezing over homogeneous freezing because it reduces the energy barrier to the formation of a critical ice germ.

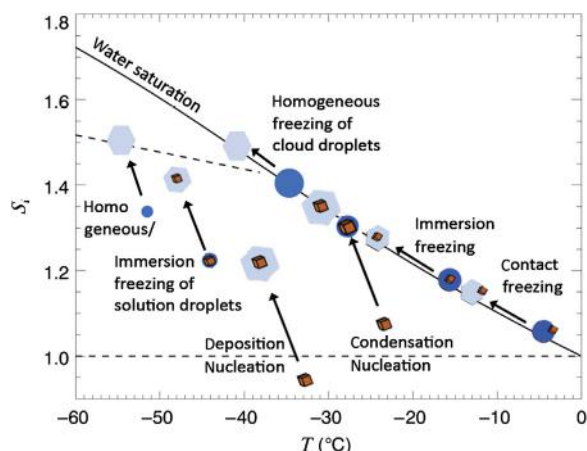


Figure 4 Schematic of the main freezing processes as a function of temperature and the saturation ratio with respect to ice (Hoose, C., Möhler, O., 2012. Heterogeneous ice nucleation on atmospheric aerosols: a review of results from laboratory experiments. *Atmospheric Chemistry and Physics* 12, 9817–9854). The solid line refers to saturation with respect to water. The dashed line that starts at -35°C refers to the homogeneous freezing line of solution droplets according to Koop, T., Luo, B., Tsias, A., Peter, T., 2000. Water activity as the determinant for homogeneous ice nucleation in aqueous solutions. *Nature* 406, 611–614.

Four heterogeneous freezing modes are distinguished in the literature (Figure 4). Immersion freezing refers to the freezing that is initiated from within the droplet. It requires that the IN is already immersed in the cloud droplet at warmer temperatures. Upon cooling freezing is initiated (Figure 4). Sometimes condensation freezing is distinguished from immersion freezing. It is thought that condensation freezing refers to a different pathway in which the aerosol containing the IN starts from subsaturated conditions. When water saturation is exceeded, a liquid phase is formed, which is large enough for an ice germ to form inside of it. This ice germ within the liquid phase then initiates the freezing (Figure 4). Contact freezing refers to the collision of an IN with a supercooled cloud droplet. It requires the presence of a cloud droplet and is therefore shown on the water saturation line in Figure 4. Deposition nucleation refers to the direct deposition of vapor onto an IN. It requires that the air is supersaturated with respect to ice ($S_i > 1$). Deposition nucleation is important for cirrus clouds, when vapor is deposited for instance onto mineral dust particles that act as IN. It does not seem to be important for mixed-phase clouds, because observations reveal that the majority of clouds with 0 and -40°C have liquid water droplets at the cloud base, suggesting that immersion or contact freezing are the most important freezing mechanisms in mixed-phase clouds.

The criteria for aerosols to act as IN are thought to be (1) solid state, (2) size, (3) lattice structure, (4) molecular bindings with water, and (5) active sites. There is, however, some dispute about the importance of each of them and how many criteria are required. In order to reduce the energy barrier of heterogeneous nucleation below that of homogeneous nucleation, an ice germ has to form on a solid surface. A solid surface is necessary and is readily fulfilled for the most commonly occurring IN such as mineral dust, biological particles, and

soot. But also crystalline ammonium sulfate and some organics that are solid at certain conditions of temperature and RH can act as IN. Similarly to CCN also for IN size matters. The larger the IN surface, the larger the probability that a critical cluster forms on it. Also the larger the IN surface, the larger the probability that a good active site (see below) can be found onto it. A crystalline structure, which is similar to the ice lattice, is preferable for an IN. This is the case for silver iodide, which has been found to nucleate ice at temperatures up to -6°C . However, organics and mineral dust particles can also act as IN, although their lattice structures differ from that of ice. Thus, it is conceivable that the ability to form hydrogen bonds with water is more important than the lattice structure. Active sites refer to imperfections on the surface of the IN. They can be thought of as crevasses or steps in the lattice structure. The critical ice germ needs a smaller mass to reach the critical ice germ radius in crevasses or in corners of steps than on a plane surface, i.e., the energy barrier is particularly small in imperfections. Once ice is formed, additional water molecules can readily be attracted. Active sites are also used as an explanation why at a given temperature, RH and size of a given aerosol compound, only a fraction of a given aerosol species acts as IN. In summary, some combinations of size, lattice structure, molecular binding, and low interfacial energy with ice as promoted by active sites accounts for the IN ability of a substance.

Radiative Forcing of Anthropogenic Aerosols due to Aerosol–Cloud Interactions

Aerosol particles can affect the climate by scattering and absorption of radiation and thus exert a radiative forcing due to aerosol–radiation interactions (RF_{ari}). In addition, aerosol particles can cause a radiative forcing by acting as CCN and IN, i.e., due to aerosol–cloud interactions (RF_{aci}). Both RF_{ari} and RF_{aci} have partially offset the greenhouse gas warming since preindustrial times. Whereas radiative forcing means that the atmospheric state remains constant, the cloud lifetime, coverage, or phase may change when perturbed by aerosols from anthropogenic activity. Adjustments in clouds occur on much faster time scales than the warming due to greenhouse gases. Therefore, it has been proven useful to also introduce an effective radiative forcing where macroscopic adjustments (cloud height, lifetime, and cover) to microphysical perturbations are considered as well.

RF_{aci} refers to an increase in cloud droplet number concentration resulting from an increase in anthropogenic aerosols. If the cloud water cloud remains constant, then the surface area of the cloud is larger and therefore more solar radiation is reflected back to space. This effect was previously referred to as the Twomey effect. RF_{aci} is reported as the global annual mean change in the net top-of-the-atmosphere radiation since preindustrial times due to the aerosol-induced changes in cloud optical properties. RF_{aci} has been estimated mainly from global climate models to have caused an RF_{aci} of -0.7 W m^{-2} with a range between -0.3 and -1.8 W m^{-2} (Forster et al., 2007). Evidence for RF_{aci} can be seen in so-called ‘ship tracks’ that leave behind white lines in satellite pictures. They are caused by the increase in albedo due to the injection of pollution aerosols from the ships that increase CCN and thus the number of cloud droplets. However, ship tracks cannot be

seen all the time because of counteracting processes. This includes a faster evaporation of the more but smaller cloud droplets or increased entrainment of warm dry air from above the boundary layer into these ship tracks. Buffering processes are also the reason why ERF_{aci} is even more uncertain than R_{Faci} itself. The total anthropogenic aerosol effect due to aerosol–radiation and aerosol–cloud interactions (ERF_{aci-ari}) has been estimated from global climate models and satellite studies and amounts to -1.2 W m^{-2} with a range between -0.2 and -2.3 W m^{-2} since preindustrial times (Denman et al., 2007). Not only are the adjustment processes less well known, but they are also not well represented or are even missing in current climate models (Stevens and Feingold, 2009).

One source of uncertainty is the ice phase. Compared to warm clouds and CCN activation, aerosol effects on mixed-phase and ice clouds are much less understood (Lohmann and Feichter, 2005). While there is a consensus that mineral dust particles are good IN because they initiate ice formation at rather warm temperatures/low supersaturations, studies disagree about the importance of carbonaceous aerosols to act as IN (Hoose and Möhler, 2012). Terrestrial biogenic aerosols such as bacteria, pollen, and fungal spores have been identified as being good IN (Hoose and Möhler, 2012), and there are indications that marine planktonic diatoms also act as IN. These biological IN initiate ice formation at higher temperatures than mineral dust, but it is not clear yet if sufficiently large concentrations can be found in the atmosphere to substantially influence mixed-phase and ice clouds. Consequently, the contribution of anthropogenic emissions on IN concentrations is not yet determined. If more IN exist due to anthropogenic activity, supercooled clouds would freeze more readily. Because the vapor pressure of ice is lower than that of water, the ice crystals would grow at the expense of the cloud droplets (see Clouds and Fog: Cloud Microphysics). Glaciated clouds precipitate more readily which decreases their lifetime. This could partly offset the aerosol effect on warm clouds. However, anthropogenic emissions may deactivate IN due to coating with soluble material, such as sulfuric acid. Whether the faster glaciation or the deactivation

of anthropogenic IN is more important is still an open question. It differs in different climate models and is therefore partly responsible for the large range in ERF_{aci-ari} (Lohmann et al., 2010). In summary, aerosol–cloud interactions are not well known yet and their radiative and adjusted forcings will remain uncertain for some time to come.

See also: Aerosols: Climatology of Stratospheric Aerosols; Climatology of Tropospheric Aerosols; Observations and Measurements; Role in Climate Change; Role in Radiative Transfer. Clouds and Fog: Cloud Microphysics.

References

- Curtius, J., 2006. Nucleation of atmospheric aerosol particles. *Comptes Rendus Physique* 7, 1027–1045.
- Denman, K.L., Brasseur, G., Chidthaisong, A., et al., 2007. Couplings between changes in the climate system and biogeochemistry. In: Solomon, S., Qin, D., Manning, M., Chen, Z., Marquis, M., Averyt, K.B., Tignor, M., Miller, H.L. (Eds.), *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, Cambridge, United Kingdom and New York, NY.
- Forster, P., Ramaswamy, V., Artaxo, P., et al., 2007. Radiative forcing of climate change. In: Solomon, S., Qin, D., Manning, M., Chen, Z., Marquis, M., Averyt, K.B., Tignor, M., Miller, H.L. (Eds.), *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, Cambridge, United Kingdom and New York, NY, pp. 129–234.
- Hoose, C., Möhler, O., 2012. Heterogeneous ice nucleation on atmospheric aerosols: a review of results from laboratory experiments. *Atmospheric Chemistry and Physics* 12, 9817–9854.
- Koop, T., Luo, B., Tsias, A., Peter, T., 2000. Water activity as the determinant for homogeneous ice nucleation in aqueous solutions. *Nature* 406, 611–614.
- Lohmann, U., Feichter, J., 2005. Global indirect aerosol effects: a review. *Atmospheric Chemistry and Physics* 5, 715–737.
- Lohmann, U., Rotstayn, L., Storelvmo, T., et al., 2010. Total aerosol effect: radiative forcing or radiative flux perturbation? *Atmospheric Chemistry and Physics* 10, 3235–3246.
- Pruppacher, H.R., Klett, J.D., 1997. *Microphysics of Clouds and Precipitation*. Kluwer Academic Publishers, Dordrecht, The Netherlands, 954 pp.
- Seinfeld, J.H., Pandis, S.N., 1998. *Atmospheric Chemistry and Physics: From Air Pollution to Climate Change*. Wiley, NJ, 1326 pp.
- Stevens, B., Feingold, G., 2009. Untangling aerosol effects on clouds and precipitation in a buffered system. *Nature* 461, 607–613.

Aerosol Physics and Chemistry

M Kalberer, University of Cambridge, Cambridge, UK

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by J Heintzenberg, volume 1, pp 34–40, © 2003, Elsevier Ltd.

Synopsis

A broad overview is given on physical and chemical properties and processes of tropospheric aerosol particles. Primary and secondary, natural and anthropogenic particle sources are described followed by a description of key physical properties and processes, such as particle size distribution, formation, and removal processes. Particle properties governing particle water uptake and the formation of cloud droplets and ice crystals are described. The particle composition is discussed, with an emphasis on the organic aerosol fraction, especially secondary organic aerosols. Concepts are presented to describe heterogeneous reactions between atmospheric trace gases and aerosol particles.

Introduction

An aerosol is defined as an ensemble of solid or liquid particles suspended in a gas, e.g., in air. In atmospheric science the term ‘aerosol’ is often used to describe the particles of an aerosol only, which can lead to confusion. Aerosol particles have been recognized since decades to be of key importance for many processes throughout the atmosphere: they critically influence directly the radiative balance of the Earth’s atmosphere, affect cloud formation, and are also one of the main air pollutants contributing to a variety of respiratory and cardiovascular diseases. Despite this importance, aerosol particles are relatively poorly characterized with respect to their concentration, temporal and spatial distribution, and physical and chemical properties. This large uncertainty is mainly caused by the variable and insufficiently understood sources, formation and transformation processes, and composition of atmospheric particles.

This article focuses on tropospheric particle properties and processes and does not discuss particles present at higher atmospheric altitudes.

Aerosol Particle Sources

In the lower atmosphere, the troposphere, there are a large number of aerosol particles sources, which are usually divided into two main categories: natural and man-made (anthropogenic). Distinguishing between natural and anthropogenic particle sources is especially important when technical or regulatory measures are considered to reduce atmospheric particle concentrations and their effects in the atmosphere because only anthropogenic sources can be effectively influenced. Both categories can be subdivided into primary particles, defined as particles directly emitted into the atmosphere, and secondary particles, defined as particles (or particle mass) that are formed within the atmosphere through physical or chemical processes. The identification of secondary particles is especially challenging because this requires an accurate knowledge of the atmospheric chemistry leading to the formation of the compounds that partition from the gas to the particle phase. Due to the highly complex chemical composition of tropospheric particles this is currently one of

the major factors limiting our understanding of particle effects in the atmosphere.

Natural Particle Sources

Main primary natural sources are sea salt, mineral dust, primary biological material, and volcanic particles. Mineral dust particles, emitted predominantly in deserts and semiarid regions, are the most abundant particle type (on a mass basis) in the atmosphere on a global level (Textor et al., 2006). Mineral dust particles can be transported from major source regions such as the Sahara and Gobi desert, the Arabian Peninsula or the U.S. Southwest during large dust storms over long, intercontinental distances. For remote open ocean areas, mineral dust deposition are a significant nutrient source, especially of trace elements such as iron or phosphorous. Equally important on a global level are sea-salt particles, which are generated through bubble bursting processes and breaking waves at the sea surface and are strongly correlated with wind speed.

Primary biological particles include a wide variety of particles such as pollen, fungal spores, bacteria, and plant debris and in marine environments also organic material present in the ocean surface layer (Despres et al., 2012). Primary particles have been identified as efficient cloud condensation and ice nuclei and therefore may significantly affect particle–cloud interactions. However, the number or mass of primary organic particles is difficult to quantify mainly due to their low concentration and diversity. Volcanic primary particle emissions are by nature sporadic and are composed of SiO_2 , Al_2O_3 , and Fe_2O_3 . Due to the rather large particle size, their effects are mostly local or regional.

Natural, secondary particles are formed from a number of nitrogen- and sulfur-containing and also organic source gases. On a global scale dimethyl sulfide (DMS) is an important precursor, which is emitted by marine phytoplankton. In the atmosphere DMS is oxidized to SO_2 and finally to sulfuric acid. Gaseous sulfuric acid forms aerosol particles very quickly or condenses onto existing particles due to its low vapor pressure. Nitrate is another important inorganic particle component, which is formed mainly through oxidation of NO_x and N_2O . Lightning and soil emissions (bacterial activities) are the main sources of these precursors. The largest natural secondary

particle fraction is formed from gaseous organic precursors such as terpenes and sesquiterpenes, which are emitted by plants. Upon oxidation in the atmosphere, mainly with OH, O₃, or NO₃, a large range of reaction products is formed. Some of these reaction products with low vapor pressures condense onto particles. The chemical composition of secondary organic particles is highly complex and several hundred compounds can be found in secondary organic aerosol particles formed from one single organic precursor gas (Reinhardt et al., 2007).

Anthropogenic Particles Sources

Primary particles generated due to human activities include biomass burning, fossil fuel combustion, and industrial activities. Although wild fires can have purely natural causes, they are often linked to anthropogenic activities. In recent years it was also recognized that a significant fraction of the organic particle mass present in the atmosphere is due to biomass burning from residential heating and cooking activities, not only in developing but also in industrialized countries. Soot (i.e., graphite-like carbon), organic material, and inorganic salts are the main components of biomass burning particle and can vary significantly depending on the type of biomass and the burning conditions: initial combustion stages and poor ventilation conditions can cause orders of magnitude larger particle emissions than well established flaming conditions. Fossil fuel combustion is also a significant source of soot and organic particle mass in the atmosphere. In addition, a number of large-scale industrial processes can emit significant particle mass into the atmosphere such as mining, cement manufacturing, or metal processing industry. These sources are often of local importance.

Secondary, anthropogenic sources include inorganic and organic components that are formed via gas–particle conversion processes. Fossil fuel consumption and incomplete combustion as well as agricultural and industrial activities are the major sources. The main inorganic source gases for anthropogenic secondary particle mass are NO_x, SO₂, and NH₃, emitted by combustion processes (NO_x, SO₂) and agricultural activities (NH₃). Their oxidation products such as nitrate, sulfate, and ammonium account for roughly half the tropospheric aerosol particle mass (Jimenez et al., 2009). Similarly, oxidation of small, gaseous aromatic compounds leads to the formation of low-volatility oxidation products, which contribute to the organic aerosol mass (see also Section Chemical Composition).

Physical Properties

Size Distribution

Aerosol particles in the atmosphere usually cover a size range of several orders of magnitudes, from about 1 nm to 100 μm. These limits are not based on a rigorous concept but are often operationally defined and measured size distributions depend on and vary with analyses techniques. Below this size range particles are usually described as clusters as they are composed of only a limited number of molecules. The particle size distribution is usually divided into fine mode particles, consisting of nucleation mode (diameter <10 nm) and Aitken

mode particles (<100 nm), accumulation mode particles (100 nm–2 μm), and coarse-mode particles (>2 μm), which includes cloud droplets. The maximum in the number size distribution in urban areas is generally found in the nucleation mode. However, these smallest particles contribute only very little to the total aerosol surface or mass (Figure 1). The high number concentration observed in the nucleation mode particles in urban areas are mostly due to traffic-related sources. The maximum surface area, which is the important parameter for light scattering and for many heterogeneous reaction rates, usually lies in the accumulation mode. The total particle number concentration varies greatly in the troposphere with <20 particles cm⁻³ for the polar regions and >10⁵ particles cm⁻³ for an urban aerosol. The number concentration of the nucleation mode particles becomes generally less important at location far away from anthropogenic sources. The number size distribution of marine or desert aerosol is dominated by coarse-mode particles.

For air monitoring purposes, particle mass is usually measured for particles with a size smaller than 2.5 μm or 10 μm (abbreviated as PM_{2.5} and PM₁₀, respectively). This is mainly motivated by the size-dependent deposition of inhaled particles within the respiratory tract, where only particles smaller than 10 μm reach bronchia and lower parts of the lung where it is assumed that they cause most harm. Epidemiological studies

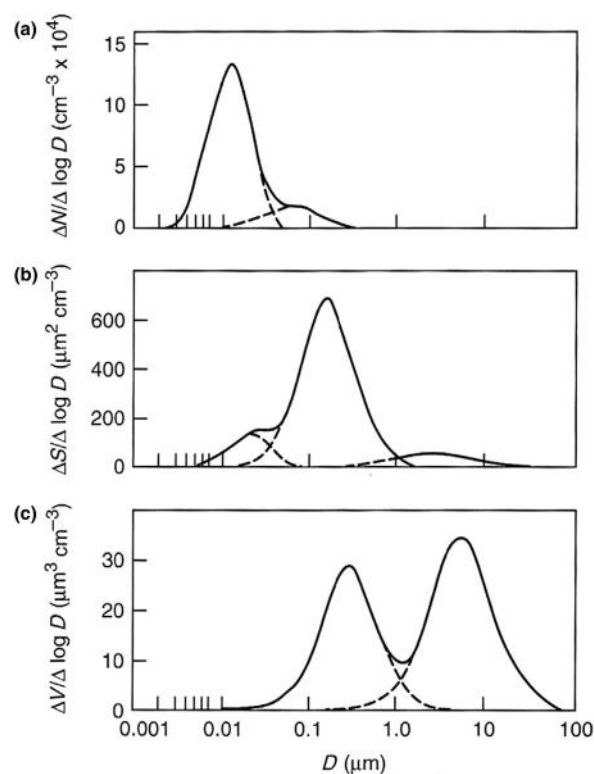


Figure 1 Idealized particle number (a), surface (b), and volume (c) size distribution for an urban location. Nucleation mode particle clearly dominates the total particle number whereas the total volume is dominated by accumulation and coarse-mode particles. Finlayson-Pitts, B.J., Pitts, J.N., 2000. Chemistry of the Upper and Lower Atmosphere. Academic Press, San Diego.

show strong correlations between negative health outcomes such as respiratory diseases and PM_{2.5} or PM₁₀ and less clear effects if the total particle mass is considered (Dockery et al., 1993).

Formation Processes

In the smallest size fraction (nucleation mode), particles are formed via homogeneous nucleation, which describes the formation of new particles through gas-to-particle conversion. Particle nucleation proceeds via initial molecule clusters, which involve water and sulfuric acid and possibly basic compounds such as ammonia, organic amines, or other organic compounds. Under favorable conditions these clusters grow over time to aerosol particles with tens of nm diameter as illustrated in Figure 2, with a typical growth rate of about 1–20 nm h^{−1}. Low-volatility organic compounds are thought to be mainly responsible for the growth of clusters into accumulation mode particles. Particle nucleation is usually observed when the pre-existing particle surface area is low but observations have been also made under highly polluted conditions or in coastal areas where nucleation is initiated by iodine oxides formed from biological sources (Zhang et al., 2012).

Particles in the accumulation mode typically arise from either the partitioning of low-volatility gases onto existing smaller particles, or from coagulation of particles. Chemical modification (heterogeneous reactions) following condensation of semivolatile components from the gas phase onto existing particles is another process, which can increase the mass of accumulation mode particles. Gas phase compounds partitioning into the particle phase are mainly formed in gas phase oxidation reactions leading to reaction products with a lower vapor pressure resulting in the partitioning of these reaction products into the particle phase.

The largest size fraction (i.e., coarse-mode particles with diameters larger than about 2 μm) is mainly generated in mechanical processes. Industrial or traffic-related processes are the main anthropogenic sources but wind-blown dust is by far the largest source of coarse-mode particles on a global level.

Removal Processes

The concentration of nucleation mode particles with diameters <10 nm is rapidly reduced through coagulation, resulting in

a lifetime in the order of minutes for these smallest particles. The coagulation rate of particles with size 1 and 2 can be described by eqn [1]

$$\frac{dN_{1,2}}{dt} = -K_{1,2}N_1N_2 \quad K_{1,2} = \pi(d_1 + d_2)(D_1 + D_2) \quad [1]$$

where $K_{1,2}$ is the coagulation coefficient, D the diffusion constant, and d the particle diameter. Values for $K_{1,2}$ are given in Table 1 for a range of particle sizes. Most efficient coagulation is observed between small particles with a high diffusion constant and large particles with a high surface area.

The main removal process for accumulation mode particles is incorporation into cloud droplets and subsequent wet deposition. Wet deposition is also described as scavenging, which can be distinguished in in-cloud and below-cloud scavenging. The wet deposition rate of particles depends largely on the particle size and the chemical composition (see below).

Particles of all sizes also undergo dry deposition. The efficiency of this process is strongly size dependent. The flux of particle dry deposition to the ground, F , can be described as the product of the particle concentration, c , and the deposition velocity, v_g (eqn [2]).

$$F = v_g \cdot c \quad [2]$$

Particles below about 50 nm diameter behave essentially like gases with respect to dry deposition and are efficiently removed due to Brownian diffusion. Coarse-mode particles in

Table 1 Coagulation coefficients for particles of different sizes ($K_{1,2} \times 10^{-10} \text{ cm}^3 \text{ s}^{-1}$) under standard temperature and pressure

d_2 (nm) \ d_1 (nm)	0.01	0.1	1.0	10.0
0.01	67			
0.1	180	8.6		
1.0	1700	24	3.5	
10.0	16000	220	10.3	3.0

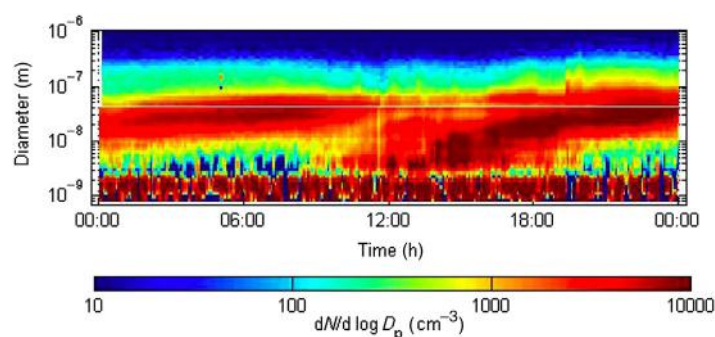


Figure 2 Particle nucleation. Evolution of the particle size distribution over the course of a day. A nucleation event is observed at a boreal forest site in Finland around midday. Particles grow through condensation over several hours to sizes of about 50 nm (Manninen, H.E., et al., 2009. Charged and total particle formation and growth rates during EUCAARI 2007 campaign in Hyytiälä. Atmospheric Chemistry and Physics 9, 4077–4089). The color code indicates the particle concentration at a particular size.

the range of 2–20 μm are mostly deposited due to inertial impaction and very large particles $>20 \mu\text{m}$ are removed due to gravitational settling (sedimentation). All the dry deposition processes are less effective for particles in the accumulation mode (50 nm–2 μm), and thus these particles have the longest atmospheric lifetime (i.e., minimum deposition velocity).

Water Uptake and Particle Phase

As mentioned above, incorporation of particles into clouds is one of the most efficient particle removal processes. In addition, particle–cloud interactions strongly affect the formation of clouds and precipitation and thus are key processes in the hydrological cycle. Via light scattering of clouds this also affects indirectly the radiative budget of the atmosphere.

The ability of a particle to take up water (hygroscopicity) and to grow eventually to a cloud droplet (cloud condensation nuclei) is mainly governed by the initial dry particle size, its chemical composition, and the relative humidity. The Köhler theory describes the water vapor pressure, S , over an aqueous droplet as described in eqn [3],

$$S = a_w \exp\left(\frac{2\sigma M_w}{RT\rho_w D}\right) \quad [3]$$

where a_w is the water activity, σ the surface tension of the surface–air interface, M_w the molecular weight of water, R the gas constant, T the temperature, ρ_w the density of water, and D the diameter of the droplet.

Two processes have to be considered to describe water uptake of particles, the Kelvin effect, the exponential term in eqn [3], describes the vapor pressure due to the curvature of the droplet surface compared to a flat interface and the Raoult's effect, which considers the influence from solutes, described by the water activity. The Kelvin effect tends to increase the water

vapor pressure while the Raoult's effect is mainly decreasing the water vapor pressure.

The effects of the hygroscopicity of solutes on the water activity can be described with a single parameter, κ , and is referred to as κ -Köhler theory (Petters and Kreidenweis, 2007). This allows modeling and comparing the hygroscopicity of single component particles with complex ambient particles of unknown composition.

At low droplet diameters the solute effect is often dominant (Figure 3). Droplet sizes in the raising part of the Köhler curve are stable and in equilibrium with the relative humidity of their surrounding atmosphere. If the relative humidity is higher than the critical supersaturation (i.e., the maximum of the Köhler curves in Figure 3) then a particle will grow to cloud droplet sizes of many micrometers as described by the unstable, decreasing part of the Köhler curve. The growth of a particle into a cloud droplet in a supersaturated atmosphere is called activation. Due to the Kelvin effect the critical supersaturation is higher for smaller particles and therefore larger particles are more efficiently activated to cloud droplets.

Particle hygroscopicity usually increases with increasing age of a particle in the atmosphere, which is due to condensation of oxidized, water-soluble gases onto the particle. This highlights the important link between chemical processing of particles and their physical properties. Inorganic salts such as ammonium sulfate or ammonium nitrate (both secondary particle compounds) are among the most hygroscopic components in atmospheric particles, whereas particles with high organic content or mineral dust particles have a low hygroscopicity.

Aerosol particles are essential not only for cloud droplet formation but also for ice crystal formation. At temperatures below about -15°C mineral dust particles are effective ice nuclei and some bacteria are efficient ice nuclei already at -3°C (Murray et al., 2012). Field experiments showed that

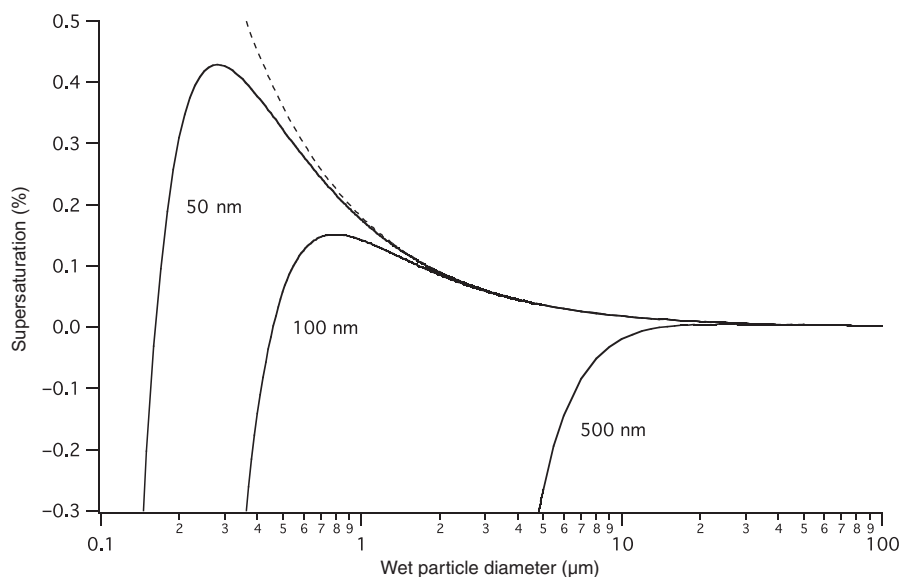


Figure 3 Köhler curves for ammonium sulfate particles with dry diameters of 50, 100, and 500 nm respectively, illustrating the dominant effect of the dry particle diameter on the water vapor supersaturation required to activate a particle to a cloud droplet. The dotted line indicates the Kelvin effect only.

these two particles types are the most abundant atmospheric ice nuclei (**Figure 4**) (Pratt et al., 2009).

Two processes are likely the most important ice nucleation processes in a cloud: contact freezing, which describes ice particle formation due to the collision of an aerosol particle with a subcooled liquid droplet and immersion freezing, where particles are already present in a liquid droplet and served as ice nuclei.

Particle composition does not only affect their ability to serve as cloud condensation or ice nuclei but also determines the phase of an aerosol particle. For particles composed of inorganic salts the water content determines whether they are crystalline or liquid. Particles primarily composed of organic matter were traditionally assumed to be liquid under ambient tropospheric conditions. However, in recent years it became increasingly clear that organic particles in the atmosphere might be present in a semisolid or glassy state (Virtanen et al., 2010). The presence of a glassy state (compared to a liquid particle) has significant consequences with respect to water uptake kinetics and diffusivity of oxidants and particle components into the particle bulk and thus might affect composition and particle–cloud interactions (Koop et al., 2011). Particle-phase reactions (i.e., aging) leading to high-molecular weight compounds are a key factor that increases the particle viscosity.

Optical Properties

Particles are indirectly affecting the radiative balance of the atmosphere via cloud formation. In addition, particles are also directly scattering and/or absorbing light. These processes, depending on the particle size, shape, and composition, lead

to a direct cooling (scattering) or warming (absorption) of the atmosphere.

Chemical Composition and Reactivity

Besides physical properties, information about the chemical particle composition is critical to assess comprehensively aerosol effects in the atmosphere. Composition is closely linked to cloud condensation or ice nucleation activities of particles and their optical properties. In addition, only a detailed knowledge of the chemical composition allows establishing the quantitative contribution of the various particle sources because physical properties such as size are rarely source-specific. This is especially of concern for air quality issues, because particle health effects are most likely related to particle composition and because only the anthropogenic particles fraction can be easily regulated.

The chemical composition of aerosols particles is highly complex and variable in space and time and several thousand of compounds have been separated with chromatographic or mass spectrometry techniques (Hamilton et al., 2004). Inorganic salts, minerals, and organic compounds are the major fractions in tropospheric particles. The detailed composition depends, e.g., on the relative influence of anthropogenic vs natural particle sources. In addition the particle composition constantly changes during the entire atmospheric lifetime of the particle due to chemical and physical processes. With increasing atmospheric age condensation of inorganic salts (like nitrate, ammonium, and sulfate) and organics will gradually lead to a more uniform chemical composition of particle composition and relevant physical properties such as hygroscopicity compared to freshly emitted particles.

Major inorganic components are ammonium nitrate and ammonium sulfate, which are often of anthropogenic origin and are therefore dominant particle components over continental regions where they account for roughly 50% (Jimenez et al., 2009) of the total mass for particles below 2.5 μm diameter, as illustrated in **Figure 5**. In contrast, in the marine atmosphere sea-salt components NaCl, KCl, CaSO_4 , and Na_2SO_4 are dominant. Sea-salt particles also contain a large number of organic compounds, which are present at the ocean surface layer and which are the most abundant in the smallest, submicrometer particle fraction of primary marine particles.

Metals have natural as well as anthropogenic sources. Natural sources include mineral dust emitted from (semi-) arid regions in large quantities. These particles are mainly composed of oxides and carbonates of Si, Al, Ca, Fe. The composition and mineralogy of mineral dust bears a close resemblance to the average crustal composition and may be used to identify its source region. Metals from anthropogenic activities are related to industrial processes or traffic sources but account for only a small fraction of the total particle mass.

Water is an abundant but also a highly variable component in atmospheric aerosol particle (see discussion above about hygroscopicity and cloud condensation nuclei). Thus in most studies on particle composition, water is not considered. When particle mass concentrations are reported, e.g., for air monitoring purposes, the particle weight is determined at 50% relative humidity.

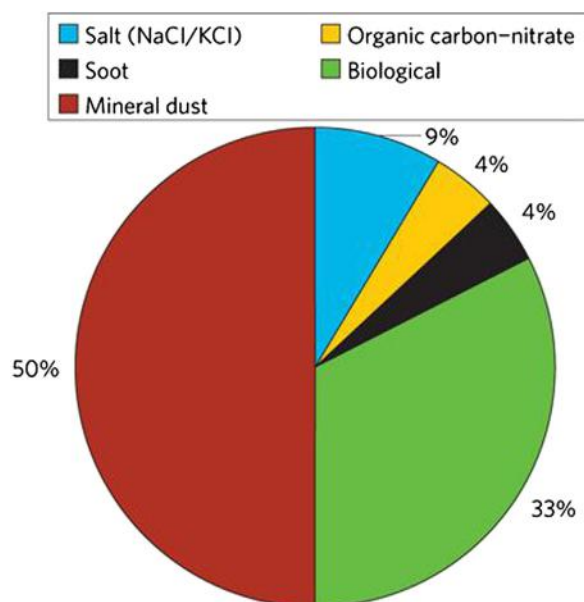


Figure 4 Chemical composition of ice nuclei as observed in field experiments. Mineral dust particles and primary biological particles such as bacteria or pollen and spores are the most abundant categories of ice nuclei. Pratt, K.A., et al., 2009. In situ detection of biological particles in cloud ice-crystals. *Nature Geoscience* 2, 398–401.

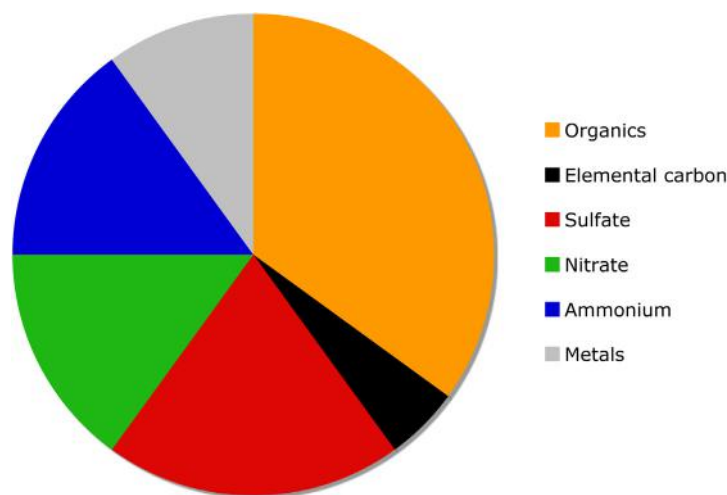


Figure 5 Typical chemical composition of continental urban aerosol particle (PM_{2.5}).

Carbonaceous Aerosols

The carbonaceous particle mass can be divided into an organic and an inorganic carbon fraction (inorganic minerals containing carbonates are not considered here). Separating the two fractions is important because the inorganic, graphite-like carbon mass is the main light absorbing fraction in tropospheric particles and thus a reliable quantification is crucial to assess radiative properties of the particles. However, the quantitative determination of the inorganic carbon mass is challenging mainly because of the difficulty to separate it efficiently from all organic material. There are a number of analytical methods used to quantify the two fractions using combustion methods followed by a quantification of CO₂ (or CH₄ after an additional reduction step) or using optical methods measuring the total adsorption at different wavelengths (Schmid et al., 2001). In an alternative method (single particle soot photometer) the particles are vaporized in an intense laser beam and the emitted incandescent light over a broad wavelength range is measured. The incandescence is empirically related to particle composition and mass (Slowik et al., 2007). Depending on the method a number of terms such as black carbon, elemental carbon, or soot are used in the literature.

Besides inorganic carbonaceous particle components, organic compounds are a major but very poorly understood fraction of tropospheric particles. The large complexity and variability of the organic particle composition, both spatially and temporarily and the usually low amount of sample available for analysis are the main reasons which limit the use of many analytical techniques. Not only the large number of organic compounds is a challenge for most chemical analytical techniques with several thousand compounds that have been separated and categorized, but also their large range of physical properties such as molecular weight distribution, volatility, polarity, or functional group distribution. Studies attempting to characterize the organic aerosol fraction on a molecular level are often able to identify only 10–30% of the total organic mass (Decesari et al., 2006). Thus, the chemical nature of the vast majority of organic compounds is only very poorly understood

and therefore it is also difficult to clearly identify the main sources of the organic particle fraction.

Nonoxidized hydrocarbons (such as alkanes and polyaromatic hydrocarbons) or only slightly oxidized compounds (e.g., fatty acids) are emitted by primary particle sources. The largest fraction of organics, however, are oxidized compounds and have variety of oxygen functional groups such as carboxylic acids, ketones, alcohols, esters, or peroxides with often multiple functional groups such as polycarboxylic acids. Besides compounds with oxygen-containing functional groups, there is also a large range of nitrogen- and sulfur-containing compounds found in organic particles. The presence of such compounds is more prominent at locations with higher anthropogenic influence (i.e., in polluted air) indicating that SO₂ and NO_x are gaseous precursors of these compounds.

Secondary Organic Aerosol

A majority of these oxidized compounds are not directly emitted but are formed in the atmosphere, either in the gas phase followed by condensation or in the particles by oxidation or condensation reactions. The organic aerosol mass formed through these processes is defined as secondary organic aerosol (SOA) mass.

There are a number of volatile organic compound classes that are efficiently forming SOA, such as aromatics (mainly emitted into the atmosphere due to fossil fuel consumption) and terpenes or sesquiterpenes (biogenic precursors of SOA). The atmospheric oxidation of these SOA-precursor compounds leads to the formation of hundreds of organic oxidation products, some of which are highly oxidized and have a low vapor pressure. The particle formation potential of an SOA precursor depends on their chemical structure and on oxidant concentrations. Especially under intense sunlight and high oxidant conditions SOA can be the dominant fraction of the total organic particle mass. Due to the highly complex mixture present in SOA, it is challenging to describe SOA formation on a fundamental level. The partitioning of an individual organic compound between the gas and the particle phase can be described by its gas–particle partitioning coefficient, K_p , which

mainly depends on the compound's vapor pressure and its activity coefficient in the particle phase as described in eqn [4],

$$K_i = \frac{760 RT f_{om}}{MW_{om} 10^6 \alpha_i p_{L,i}^0} \quad [4]$$

where R is the ideal gas constant, T temperature, f_{om} the fraction of the total particle mass into which the organic components can partition into, MW_{om} the mean molecular weight of the absorbing organic particle mass, α_i the activity coefficient of the compound in the organic particle matrix and $p_{L,i}^0$ the (subcooled) liquid vapor pressure of the compound (Pankow, 1994).

The organic mass formed from an individual compound (aerosol yield, Y_i), or a number of compounds, can be described by eqn [5], where M_0 is total organic mass, K is the gas/particle partitioning coefficient and α is the molar yield of compound i .

$$Y_i = M_0 \sum_i \left(\frac{\alpha_i K_i}{1 + K_i M_0} \right) \quad [5]$$

Because the vapor pressure and activity coefficient of many organic aerosol components are not known which are needed to calculate K_i (see eqn [4]) and because the structure of the vast majority of all particle components is unknown it is not possible to describe SOA formation using these fundamental relationship. However, the particle yield of an individual SOA-precursor gas can be parameterized using the SOA yield equation (eqn [5]). It has been shown experimentally that for most SOA precursors it is sufficient to assume only two particle-phase reaction products with different reaction yield and vapor pressure for a quantitative SOA yield description rather than accounting for the hundreds of compounds actually present in the particle (Odum et al., 1996). This simple model further assumes a linear combination of SOA yield when different SOA precursors are present in the atmosphere. Recently it has been recognized that this simple assumption might not be valid under all conditions.

In addition, in recent years it became clear that model estimates using this approach severely underestimate the (secondary) organic aerosol mass observed in the ambient atmosphere (Volkammer et al., 2006). Potential reasons for this discrepancy are unknown organic SOA precursors or particle-phase reactions that change the chemical mass, composition, or lifetime of SOA in the atmosphere. Particle-phase reactions such as oligomer formation and continuous oxidation are decreasing the volatility of particle components and thus enhancing the particle stability in the atmosphere (Kalberer et al., 2004). Such processes further highlight that SOA particle composition cannot simply be described with compounds formed in the gas phase, which condense into the particle phase but that particle-phase reactions are continuously changing the composition of SOA during the entire atmospheric lifetime of the particle.

Tracers for Aerosol Sources

Despite the difficulties to characterize the particle composition on a molecular level, there have been many attempts to identify

tracer molecules or elements for specific particle sources. A quantitative understanding of source-specific tracers directly at the particle source allows determining the contribution of such a source in the complex ambient particle mixture.

Metals have been frequently used as tracers in source apportionment studies because they do not react in the atmosphere or evaporate as it might be the case with organic compounds. However, metal profiles may overlap between particle sources and thus source profiles are rarely unambiguous for a single particle source. In recent years, metal source profiles combined with organic tracers are being increasingly used to obtain more source-specific tracer profiles (Pant and Harrison, 2012).

A large range of organic tracers has been proposed for a variety of mainly primary sources. One of the most used organic tracers is levoglucosan, which is a tracer for biomass burning particles together with long-chain alkanes, hopanes, and PAHs (Simoneit, 2002). Patterns of PAHs and hopanes, together with steranes and elemental carbon, were used to derive specific emission profiles of diesel and gasoline vehicle exhaust. Other sources identified with organic markers include primary particles from vegetation, food cooking, road dust, and industrial processes (Schauer et al., 1996). The major drawback of organic tracers is their potential reactivity in the atmosphere mainly with OH or NO₃ radicals and a careful characterization of the atmospheric chemistry of potential organic tracers is needed to estimate their suitability.

An alternative technique to identify carbonaceous particle sources and quantify the fraction originating from fossil fuel sources is radiocarbon (¹⁴C) analyses. Because ¹⁴C has a half-life of about 5730 years all ¹⁴C decayed in fossil fuel and thus the identification of ¹⁴C levels in atmospheric aerosols allows for an unambiguous quantification of the carbonaceous particle fraction, which originates from fossil fuel sources. The remaining carbon particle fraction is a mixture of natural and anthropogenic sources, including natural biological particles and biomass burning. Using the ¹⁴C technique it has been shown that elemental carbon is generated mainly from fossil sources. However, in rural areas or developing countries, where wood burning is the dominant source for domestic heating and cooking, 50–70% of the black carbon originates from biomass burning (Gustafsson et al., 2009). Such quantitative information about particle sources is crucial to develop future mitigation strategies to decrease the effect of particles on climate and health.

Heterogeneous Reactions

As mentioned above the chemical composition of aerosol particles changes continuously during their atmospheric lifetime and can significantly influence their effects on climate and health. Chemical reactions of a gas phase compound occurring at the surface or in the bulk of particles are defined as heterogeneous reactions. These reactions can take place in aqueous cloud droplets, in liquid organic particles, or on solid particle surfaces.

Heterogeneous chemical reactions in the atmosphere involve a series of physical processes, including diffusion in the gas phase, accommodation (sticking), and evaporation from the surface. In the case of liquids, diffusion and dissolution will

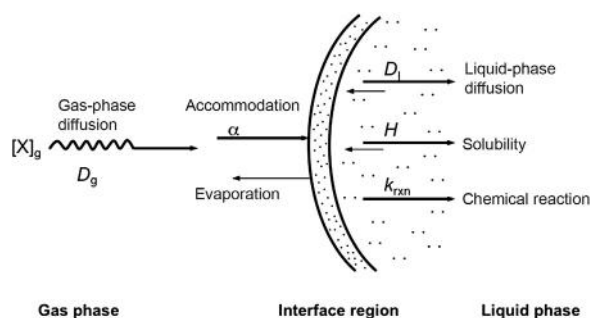


Figure 6 Physical and chemical processes involved in the transfer of molecules from the gas to the liquid particle phase.

occur in the condensed phase. Chemical reaction may also happen at the gas–particle interface or in the condensed phase. All these processes must be treated together to understand the overall rate of removal of a compound from the gas phase, as any one process can act in a rate determining way. These processes, illustrated for a liquid particle, are represented in **Figure 6**.

The rate of a heterogeneous process is usually expressed in terms of the uptake coefficient, γ , which is defined as the fraction of molecules colliding with the surface that are permanently lost from the gas phase (eqn [6]).

$$\gamma = \frac{\text{number of molecules lost from the gas phase per unit time}}{\text{number of molecules colliding with the surface per unit time}} \quad [6]$$

According to the kinetic theory of gases, the rate of collision of a species A at the surface, per unit area, is $[A] \bullet c/4$, where $[A]$ is the concentration (molecules cm^{-3}) of species A and c the mean molecular speed of the gas molecules.

The rate of removal of molecules at the surface can therefore be expressed as a first-order process with rate (eqn [7])

$$-\frac{d[A]}{dt} = k_{\text{het}}[A] \quad [7]$$

with k_{het} defined as (eqn [8])

$$k_{\text{het}} = \gamma \frac{c}{4} S [\text{cm}^3 \text{ s}^{-1}] \quad [8]$$

where S is the surface area of the atmospheric particles. As well as providing data on the rate of removal of trace gases by aerosol particles, uptake coefficient measurements can be used to deduce mechanistic aspects of the heterogeneous processes.

A useful approach for dealing with general problems of gas uptake rates into clouds and aerosols is the resistance model. This is a simplified representation based on an electrical circuit analogue as shown schematically in **Figure 7**. Each process is expressed as a resistance term and thus can be combined in series or in parallel to obtain the overall rate of the heterogeneous process.

Resistances are expressed as the inverse of the dimensionless uptake coefficients. The resistances $1/\gamma_g$ and $1/\alpha$ represent the

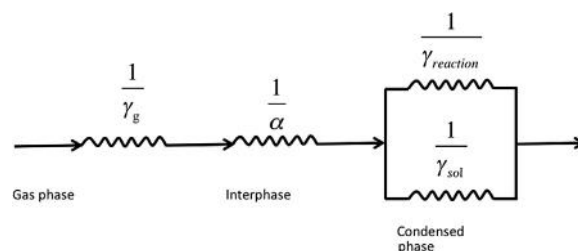


Figure 7 The resistance model describing gas uptake into the condensed phase.

resistance due to gas transport and surface accommodation, respectively, and γ_{sol} and γ_{rxn} describe liquid solubility and reaction, respectively. On the basis of the model illustrated in **Figure 7**, the overall uptake coefficient is expressed as (eqn [9]):

$$\frac{1}{\gamma} = \frac{1}{\gamma_g} + \frac{1}{\alpha} + \frac{1}{(\gamma_{\text{sol}} + \gamma_{\text{rxn}})} \quad [9]$$

where γ is the experimentally determined overall net gas uptake coefficient.

An example for a heterogeneous reaction limited by accommodation is the hydrolysis of nitrogen pentoxide N_2O_5 leading to the formation of nitric acid, a major removal process of nitrogen oxides in the atmosphere. In contrast, the oxidation

of SO_2 to H_2SO_4 by ozone and hydrogen peroxide in cloud droplets is so fast that the overall reaction is controlled by diffusion of SO_2 in the gas phase to the cloud droplet surface. A large number of other heterogeneous atmospheric reactions have been investigated (Crowley et al., 2010).

In recent years, models describing the kinetics of heterogeneous reactions in more detail have been developed, which offer a generalized framework that allows for detailed process description of gas phase, surface and particle bulk physical transport processes and reactions (Pöschl et al., 2007). In addition, models have been developed to specifically describe in detail aqueous phase chemical reactions in cloud droplets to estimate aqueous phase reactions such as N_2O_5 hydrolysis, carboxylic acid formation, or halogen chemistry (Herrmann et al., 2005).

References

- Crowley, J.N., et al., 2010. Evaluated kinetic and photochemical data for atmospheric chemistry: volume V – heterogeneous reactions on solid substrates. *Atmospheric Chemistry and Physics* 10, 9059–9223.
- Decesari, S., et al., 2006. Characterization of the organic composition of aerosols from Rondonia, Brazil, during the LBA-SMOCC 2002 experiment and its representation through model compounds. *Atmospheric Chemistry and Physics* 6, 375–402.
- Despres, V.R., et al., 2012. Primary biological aerosol particles in the atmosphere: a review. *Tellus B* 64, 15598. <http://dx.doi.org/10.3402/tellusb.v64i0.15598>.

- Dockery, D.W., et al., 1993. An association between air pollution and mortality in six U.S. cities. *New England Journal of Medicine* 329, 1753–1759.
- Finlayson-Pitts, B.J., Pitts, J.N., 2000. *Chemistry of the Upper and Lower Atmosphere*. Academic Press, San Diego.
- Gustafsson, O., et al., 2009. Brown clouds over South Asia: biomass or fossil fuel combustion? *Science* 323, 495–498.
- Hamilton, J.F., et al., 2004. Partially oxidised organic components in urban aerosol using GCXGC-TOF/MS. *Atmospheric Chemistry and Physics* 4, 1279–1290.
- Herrmann, H., et al., 2005. Towards a more detailed description of tropospheric aqueous phase organic chemistry: CAPRAM 3.0. *Atmospheric Environment* 39, 4351–4363.
- Jimenez, J.L., et al., 2009. Evolution of organic aerosols in the atmosphere. *Science* 326, 1525–1529.
- Kalberer, M., et al., 2004. Identification of polymers as major components of atmospheric organic aerosols. *Science* 303, 1659–1662.
- Koop, T., et al., 2011. Glass transition and phase state of organic compounds: dependency on molecular properties and implications for secondary organic aerosols in the atmosphere. *Physical Chemistry Chemical Physics* 13, 19238–19255.
- Manninen, H.E., et al., 2009. Charged and total particle formation and growth rates during EUCAARI 2007 campaign in Hyytiälä. *Atmospheric Chemistry and Physics* 9, 4077–4089.
- Murray, B.J., et al., 2012. Ice nucleation by particles immersed in supercooled cloud droplets. *Chemical Society Reviews* 41, 6519–6554.
- Odum, J.R., 1996. Gas/particle partitioning and secondary organic aerosol yields. *Environmental Science and Technology* 30, 2580–2585.
- Pankow, J.F., 1994. An absorption model of gas/particle partitioning of organic compounds in the atmosphere. *Atmospheric Environment* 28, 185–188.
- Pant, P., Harrison, R.M., 2012. Critical review of receptor modelling for particulate matter: a case study of India. *Atmospheric Environment* 49, 1–12.
- Petters, M.D., Kreidenweis, S.M., 2007. A single parameter representation of hygroscopic growth and cloud condensation nucleus activity. *Atmospheric Chemistry and Physics* 7, 1961–1971.
- Pöschl, U., et al., 2007. Kinetic model framework for aerosol and cloud surface chemistry and gas-particle interactions – Part 1: general equations, parameters, and terminology. *Atmospheric Chemistry and Physics* 7, 5989–6023.
- Pratt, K.A., et al., 2009. In situ detection of biological particles in cloud ice-crystals. *Nature Geoscience* 2, 398–401.
- Reinhardt, A., et al., 2007. Ultra-high mass resolution and accurate mass measurements as new tools to characterize oligomers in secondary organic aerosol. *Analytical Chemistry* 79, 4074–4082.
- Schauer, J.J., et al., 1996. Source apportionment of airborne particulate matter using organic compounds as tracers. *Atmospheric Environment* 30, 3837–3855.
- Schmid, H., et al., 2001. Results of the “carbon conference” international aerosol carbon round robin test stage I. *Atmospheric Environment* 35, 2111–2121.
- Simoneit, B.R.T., 2002. Biomass burning – a review of organic tracers for smoke from incomplete combustion. *Applied Geochemistry* 17, 129–162.
- Slowik, J.G., et al., 2007. An inter-comparison of instruments measuring black carbon content of soot particles. *Aerosol Science and Technology* 41, 295–314.
- Textor, C., et al., 2006. Analysis and quantification of the diversities of aerosol life cycles within AeroCom. *Atmospheric Chemistry and Physics* 6, 1777–1813.
- Virtanen, A., et al., 2010. An amorphous solid state of biogenic secondary organic aerosol particles. *Nature* 467, 824–827.
- Volkamer, R., et al., 2006. Secondary organic aerosol formation from anthropogenic air pollution: rapid and higher than expected. *Geophysical Research Letters* 33, L17811. <http://dx.doi.org/10.1029/2006gl026899>.
- Zhang, R., et al., 2012. Nucleation and growth of nanoparticles in the atmosphere. *Chemical Reviews* 112, 1957–2011.

Climatology of Stratospheric Aerosols

LW Thomason, NASA Langley Research Center, Hampton, VA, USA

J-P Vernier, Science Systems and Applications, Inc., Hampton, VA, USA

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

Herein, we discuss the significance and measurement of stratospheric aerosols. Stratospheric aerosols play an important role in the chemistry and radiative influence of the middle atmosphere. In the past several decades, their levels have been highly variable and primarily controlled by a series of low latitude eruptions. Large events, such as the 1991 Pinatubo eruption, reflect incoming solar radiation back into space with a net cooling of the entire Earth though the impact on sensible weather is highly variable across the Earth. Stratospheric aerosols also play a significant role in ozone chemistry.

Introduction

The most basic definition of an aerosol is particles (liquid or solid) suspended in a gas. Generally, when referring to properties of the stratospheric aerosol, it is common to neglect the gaseous media (the atmosphere) and focus only on the particulate material. Unlike the troposphere, where aerosol composition can be highly diverse, in the stratosphere that vast bulk of aerosol is composed of sulfuric acid and water mixtures. The proportion between these two components is dependent on the ambient temperature and available water vapor and most aerosol particles are between 50 and 85% sulfuric acid by mass. In addition to the sulfate–water mixture, many stratospheric aerosols have a nonvolatile core consisting of a wide variety of compositions including meteoric dust, soot, and other materials.

During the past 40 years, stratospheric aerosol levels have been primarily controlled by a series of low latitude (tropical) eruptions which increased aerosol levels by factors as large as 1000 at some locations and required many years to relax to preeruption levels. The relevance of volcanic stratospheric aerosol to climate was graphically demonstrated in the early nineteenth century by weather effects associated with 10 April 1815 eruption of Tambora in Indonesia. The eruption measured 7 on the volcanic explosivity index and is possibly the largest volcanic event of the current era. The eruption injected large amounts of aerosol and aerosol precursors into the stratosphere, which subsequently were dispersed throughout the global stratosphere. This aerosol reflected a portion fraction of the incoming solar radiation back into space with a net cooling of the entire Earth. However, this cooling was not uniformly distributed across the planet but instead influenced weather patterns such that some regions experienced substantial cooling effects while other regions experienced no impact or even warmer than normal weather. Famously, this event produced the well-known ‘Year without a Summer’ in eastern North America where snow fell in New England (United States) in June 1816 and temperatures as low as 4 °C (40 °F) were reported in July as far south as Richmond in the southern United States. Similarly, weather in Europe remained cold and damp throughout the Summer of 1816. Throughout North America and Europe widespread crop failures and poor harvests ultimately lead to increased food prices

and widespread hunger. Given the demonstrated efficient cooling produced by major volcanic eruptions, some scientists have proposed that directly injecting sulfate aerosol or its precursors into the stratosphere (i.e., Geoengineering) could counterbalance global warming.

In addition to these episodic large events, stratospheric aerosol plays a significant role in stratospheric chemistry by providing the surface for a class of processes known as heterogeneous chemical reactions. These reactions play a critical role in ozone chemistry and, as a result, the total surface area density (SAD) provided by stratospheric aerosol is a crucial parameter in understanding changes to ozone. In addition, aerosol even at the lowest levels observed in the measurement record play a subtle role in the radiative balance of climate and may modulate the efficacy of the transport of material from the troposphere into the stratosphere. Some of these materials, such as water vapor, are active greenhouse gases and thus changes to stratospheric aerosol may play an indirect role in modulating climate. While there is debatable evidence for a direct human influence on stratospheric aerosol levels, recent studies on the impact of smaller volcanic events demonstrates that even modest changes have an impact on climate. As a result the need for monitoring the stratospheric aerosol for evidence of human-induced change remains critical.

Measurements of Stratospheric Aerosol

Unlike measurements of stratospheric gas components of the stratosphere, measurements of stratospheric aerosol are not straightforward concentration or mixing ratio measurements. In order to produce a complete depiction of stratospheric aerosol a number of parameters must be measured, inferred indirectly or assumed including composition, size distribution, and the presence of a solid nucleus (and its composition). Most stratospheric aerosol measurements provide limited optical parameters that result from the detailed character of the particle which remains masked to some extent. Ultimately, it is difficult to separate what we know about aerosol from how it is measured.

In general, measurements of stratospheric aerosol can be categorized into two categories: *in situ* and remotely measured. Compared to remote sensing measurements, *in situ*

(‘in position’) measurements are much closer to measuring fundamental aerosol properties. While, by the nature of the measurements, *in situ* datasets do not have the spatial-temporal scope that space-based measurements have, they nonetheless provide crucial validation to the direct optical measurements made from space. They provide estimates of crucial aerosol bulk properties which are far less model dependent than space-based measurements. Space-based measurements provide global coverage on scales of 1-day to 1-month depending on the measurement strategy employed. However, they are limited to bulk optical measurements from which other properties must be inferred. These measurements have limited information content with regard to the details of aerosol properties and thus are dependent on making reasonable assumptions regarding both composition and size distribution. Both approaches are necessary given the strengths and limitations that both exhibit.

***In Situ* Measurements of Aerosol**

The *in situ* measurement of stratospheric aerosol has been conducted using a variety of optical particle counters (OPC). OPCs produce a size distribution by segregating particles into bins based on measuring light scattered from individual particles at a given scattering angle and/or wavelength where amount of light scattered is related to the particle size. The number of bins and the robustness of the corresponding aerosol size distributions is roughly correlated with the platform from which they are deployed. Beyond OPCs, *in situ* stratospheric aerosol measurements have been made employing diverse strategies including particle impactors, quartz crystal microbalances, and backscatter sonde.

Balloon-based optical particles such as the long-lived series operated through the University of Wyoming counters have a nearly 40-year history and important climatological information. These OPCs provide profiles of stratospheric aerosol to as high as 30 km and do not necessarily require recovery of the instrument after flight but provide only a relatively coarse size distribution. The earliest measurements often had only three size bins (0.01, 0.15, and 0.25 μm) and as such inferring bulk aerosol properties like mass or SAD was highly dependent on assumptions regarding the underlying size distribution. Later instruments have more size bins but generally do not fully explore the smaller particle ranges ($<0.1 \mu\text{m}$) and remain dependent on assumptions regarding aerosol size distributions to infer particle’s other properties. Nonetheless, these measurements have the longest consistent measurement dataset and represent a crucial part of the stratospheric measurement ensemble. Some OPCs include modes where the volatile portion of the aerosol is removed and the presence of nonvolatile, presumably solid, components are detected without direct inference of composition.

Airborne OPC instruments can be larger and more complex instruments and are sustainable for multiple flight opportunities. As a result they are able to provide far more detailed size distribution measurements than their balloon-borne cousins. Some of these instruments have extensive field campaign datasets including those associated with the focus cavity aerosol spectrometer (FCAS), the forward scattering spectrometer probe, and the multiangle aerosol spectrometer

probe. Employing a different strategy, the Civil Aircraft for the regular investigation of the atmosphere based on an instrument container OPC flies aboard commercial Lufthansa flights and has produced an extensive set of detailed aerosol and gas species measurements in the upper troposphere and lower stratosphere (UTLS). The inference of aerosol properties from measurements like the FCAS is clear-cut where the measured detailed size distribution can be combined with an appropriate kernel (for SAD, extinction at a given wavelength, etc.) to produce measurements of bulk properties of aerosol without controlling assumptions regarding the form of the underlying aerosol size distribution. However, the composition of aerosol in the upper troposphere must be assumed and is far more complex than in the stratosphere where assuming sulfate aerosol is reasonable under most conditions. Comparisons of aerosol extinction measured by space-based instruments and those from the aircraft-based OPCs generally show an excellent level of agreement. Since all OPCs pump air through their instrument (some accelerating aerosol to aircraft velocities), care must be made to minimize kinetic heating effects and account for the portion which cannot be completely mitigated. This is generally well understood and, where comparisons can be made, it appears that most *in situ* sensors account for these effects well.

The most complete understanding of the composition of aerosol comes from a laser ionization mass spectrometer like the particle analysis by laser mass spectrometry. Data from this instrument confirm the generally accepted inference that stratospheric aerosol at least in the lower stratosphere is primarily sulfate aerosol but also indicates that the presence of organics in the upper troposphere/lower stratosphere is far more common than generally considered by the remote sensing community and that a substantial fraction of stratospheric aerosol contain meteoritic material most likely transported from above 75 km where micrometeorites vaporize and recondense as less than 20-nm-sized particles that either serve as nuclei for stratospheric aerosol or are at least eventually incorporated in the existing aerosol.

Remote Sensing of Stratospheric Aerosol

Space-based measurements form the vast bulk of the global depiction of stratospheric aerosol over the past 4 decades. The data from these instruments provide extensive spatial and temporal coverage not possible from ground or airborne systems. The vertical resolution of these datasets is quite variable from instrument to instrument but in most cases (with the possible exception of the UTLS) adequate for most applications with resolution running from 0.5 to 2.5 km. The most extensive dataset employ the solar occultation technique. In this approach the Sun is observed during each crossing of the terminator (spacecraft sunrise and sunset) and the line-of-sight transmission is measured as a function of altitude. **Figure 1** shows an image demonstrating this measurement geometry. By measuring this at several wavelengths, vertical profiles of gas species like ozone and water vapor along with aerosol extinction coefficient profiles at several wavelengths can be derived. The advantages of this approach are that the Sun is a bright target enabling a small field of view which leads to high vertical resolution. In addition, the geometry is also conducive to high

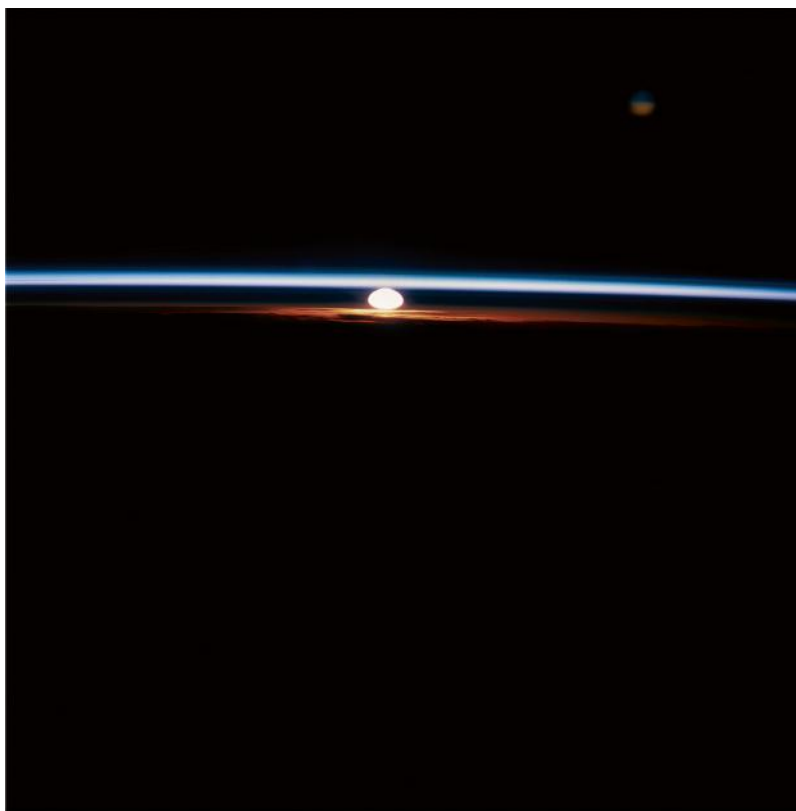


Figure 1 This photograph from NASA Space Shuttle Atlantis (Mission STS-45) shows sunrise over New Mexico during March 1992. The illuminated layer above the Sun is primarily due to the enhanced aerosol layer from the June 1991 Mt. Pinatubo eruption and vividly demonstrates the signal used by limb-scatter instruments like SCIAMACHY and OSIRIS to infer stratospheric aerosol extinction coefficient profiles. Typically, photographs of sunrise or sunset from the shuttle require substantial filtering to avoid saturation. In this photograph, the Sun is not filtered demonstrating the impact on line-of-sight atmospheric transmission that forms the basis of the solar occultation technique.

quality aerosol measurements as the long path lengths through the atmosphere lead to measurable aerosol values over a wide range of aerosol levels while the stratosphere is also reasonably homogeneous on these horizontal scales in most circumstances. On the other hand, data rates are slow with only ~ 30 profiles per day from most orbits and a full coverage of latitude can take 20–40 days for midinclination orbits while measurements are only made at high latitudes for instruments in sun-synchronous orbits (which are the preferred orbit for most nadir-viewing instruments).

There is a long history of solar occultation instruments with the foremost being the Stratospheric Aerosol and Gas Experiment (SAGE) series, which includes the Stratospheric Aerosol Measurement (SAM II, 1978–93), SAGE (1979–81), SAGE II (1984–05), and SAGE III/Meteor 3M (2002–05). These made aerosol extinction coefficient measurements at 1–9 wavelengths from 385 to 1545 nm with all having a channel near 1000 nm where aerosol is a dominant component (relative to molecular scatter and gas absorption) for even the lowest aerosol levels observed in the stratosphere over the past 30 years. On the down side, while these measurements have been used successfully to infer many aspects of aerosol and stratospheric morphology, the visible/near infrared measurements by these measurements have limited information about the underlying composition and size distribution of the aerosol

in the stratosphere. The wavelength dependence primarily provides information about the largest aerosol present but is insensitive to the smallest aerosol present particularly those with radii below 100 nm. As a result, inferences of size distributions and crucial aerosol bulk properties like SAD are particularly model dependent (most often assuming a single mode log-normal size distribution) and can have substantial uncertainties and may tend to underestimate SAD at low aerosol levels (depending on the model used). Other solar occultation instruments include the Polar Ozone and Aerosol Measurement (POAM) series. Like SAM II, these instruments were in sun-synchronous orbits and made valuable polar vortex measurements of ozone and aerosol during their missions (POAM II, 1993–1996; POAM III, 1998–2005).

The Halogen Occultation Experiment (1991–2005) was also a solar occultation instrument but operated in the thermal infrared where it measured ozone and other gas species. It also inferred aerosol extinction coefficient at four wavelengths. The information content of these aerosol measurements is quite different from the visible/near infrared measurements made by the SAGE and POAM series as sulfate aerosol is an absorber beyond a wavelength of about $2.8 \mu\text{m}$ as opposed to a scatterer at shorter wavelengths. This difference fundamentally modifies the information content of these measurements such that while infrared aerosol measurements have little information about

aerosol size they are almost directly proportional to total aerosol volume including even the smallest aerosol present. While this limits the information content of these measurements by themselves, it is clear that the combination of visible/near infrared and thermal infrared aerosol extinction coefficient measurements substantially reduces the uncertainties in the inference of aerosol bulk properties such as SAD and this combined approach is a necessary step in improving the utility of space-based aerosol measurements. Following the 1991 eruption of Mt. Pinatubo, the space shuttle-based Atmospheric Trace Molecule Spectroscopy Experiment instrument was able to resolve features caused by sulfate aerosol absorption in the infrared in a manner analogous to the measurement of gas species. This is typically not possible except when aerosol levels are substantially enhanced. An additional source of thermal infrared measurements of stratospheric aerosol employs the limb emission technique. In this approach, the emission of the atmosphere is observed in the limb direction and the wavelength dependence of this emission is primarily used to infer profiles of gas species. In this case, aerosol extinction coefficient can be inferred as a slowly varying component in regions of the infrared spectra where the atmosphere is not strongly emitting/absorbing. This is most easy to accomplish during enhanced periods as was demonstrated by cryogenic limb array etalon spectrometer, which operated during the peak of the post-Pinatubo eruption period (1991–93). During lower level aerosol periods, this measurement is more challenging though Michelson Interferometer for Passive Atmospheric Sounding (MIPAS) has demonstrated some ability to infer aerosol extinction profiles.

Some additional solar occultation measurements that continue past the 2005 aerosol instrument Armageddon include aerosol measurements by the Atmospheric Chemistry Experiment instruments Fourier transform spectrometer Sun imager and the Measurement of Aerosol Extinction in the Stratosphere and Troposphere Retrieved by Occultation remain in the developmental stage but may eventually continue the solar occultation aerosol measurement ensemble.

Since the demise of the SAGE series of instruments in 2005, the stratospheric aerosol record has become dependent on instrument using a variety of techniques which provide their own strengths and weaknesses. The most similar to solar occultation is the global ozone monitoring by occultation by stars (2002–12), which employed stellar occultation to measure ozone and aerosol extinction at visible wavelengths. The instrument was capable of performing occultations using a number of stars during the night portion of each orbit and thus produced profile data at a significantly higher rate than are possible for solar occultation. On the other hand, the relatively dimmer target and the effects of stellar scintillation (twinkling) complicate the data analysis such that aerosol data below 20 km are difficult and substantial care must be made while using these data. Currently, data from this instrument are only available at a single wavelength but are undergoing a major revision that may dramatically improve the utility of this instrument's aerosol data in the future.

The most likely heirs to the solar occultation heritage of stratospheric aerosol measurements are limb scatter and space-based lidar. In the past decade, three limb-scatter instruments have provided measurements of ozone and other gas

species: Odin spectrometer and infrared imager system (OSIRIS, 2002–present), scanning imaging absorption spectrometer for atmospheric chartography (SCIAMACHY, 2002–12), and ozone mapping and profiler suite (OMPS, 2011–present). More recently, there has been a concerted effort to similarly derived aerosol extinction coefficient profile information from these data. This was initially done simply to account for the impact of aerosol on scattering of solar radiation in the stratosphere in the derivation mission critical products such as ozone. However, the presence of a clear aerosol dependence on the measurements makes an aerosol measurement possible. Limb-scatter aerosol produces complete global coverage in a single day compared to approximately a month for solar occultation. **Figure 1** shows an image demonstrating this measurement geometry.

The limb-scatter instruments observe the limb of the Earth's atmosphere (away from the Sun) and measure the sunlight scattered by the atmosphere into the instrument. The amount of light is dependent on absorption along the path by gas species, scattering out of the line of sight by molecules and aerosol and scattering into the field of view of the instrument by molecules and aerosol. The angle between the position of the Sun, a point in the atmosphere, and the instrument influences the amount of light scattered into the instrument. While this scattering efficiency is well known for molecular scatter, the scattering by stratospheric aerosol is dependent on the size distribution. In the past, the aerosol phase function was inferred from other data and provided to the instrument analysis as a priori data. However, more recent work to use the observed wavelength dependence to infer the aerosol scattering function holds substantial promise to remove the dependence on this external information. Limb-scatter measurements are made over as similar domain as solar occultation information and have the same lack of information on the details of the underlying size distribution. The inference of an effective size distribution (one that reproduces the observed measurements) appears to be adequate to enable good quality aerosol extinction coefficient measurements though concomitant inferences of aerosol bulk properties have the same limitations as those from solar occultation. As a result the next generation of aerosol datasets from these instruments should be substantially improved over earlier versions and are applicable to science applications. However, it will be important to demonstrate that data quality is independent of scattering angle and overlap between limb instruments and solar occultation data in 2002 and 2005 and beginning again in 2014 with the launch of SAGE III/ISS will provide important calibration between these datasets that is crucial to maintaining a trendable aerosol extinction climatology dataset.

The use of laser ranging and detection (lidar) to measure stratospheric aerosol from space is unique compared to any other approach in that it is an active measurement as opposed to a passive measurement (using ambient sources of radiation like the Sun or atmospheric thermal emission). There is a substantial history of ground-based lidar measurements of the stratospheric aerosol dating back as far as the 1960s. Despite the limited spatial coverage they are an important resource in understanding the long-term variability of stratospheric aerosol. More recently, space-based lidar has used a laser to probe the atmosphere and produce high

vertical resolution profiles of aerosol backscatter coefficient. Nominally, space-based lidar can continually produce stratospheric aerosol analyses throughout an orbit as was demonstrated by the first space-borne lidar, Lidar In-space Technology Experiment in September 1994. In practice, the signal is generally small from the stratosphere and with the additional noise associated with daytime measurements; only data from the dark half of the orbit can be used to depict the stratosphere and thus is somewhat analogous to limb-scatter measurements.

The Cloud-Aerosol Lidar with Orthogonal Polarization (CALIOP) has produced lidar data since 2006. This has been a relatively low aerosol-loading period in the stratosphere and substantial averaging of the individual profiles is required to produce a meaningful stratospheric product. In addition, given the low aerosol levels, the calibration of the instrument is a crucial element of a stratospheric product. Typically, lidar is calibrated using a level in the atmosphere that is very low in aerosol treated as aerosol free. In the case of CALIOP, altitudes between 36 and 39 km are used for calibration though the aerosol is most likely on the order of a few percent (1–2%) of the molecular value. Since stratospheric backscatter levels are only on the order of 5–20% enhanced relative to molecular value, the potential for bias in the measurements remains. Nonetheless, CALIOP has been very useful in observing stratospheric aerosol and polar stratospheric clouds (which occur in polar winter). CALIOP has observed a number of volcanic plumes in the UTLS regions as well as aerosol enhancements in the upper troposphere associated with the Asian monsoon convection. Ground-based observations from fixed locations had inferred a stratospheric aerosol trend above 20 km which had been attributed to Chinese sulfur emissions but, as CALIOP data demonstrated, was actually the result of small volcanic events and their transport within the broad stratospheric circulation. While these measurements are challenging, they clearly bring an exciting new capacity to the monitoring of stratospheric aerosol.

The Lifecycle of Stratospheric Aerosol

In many ways, the processes controlling stratospheric aerosol levels, their distribution within the stratosphere, and how they exit the stratosphere can be divided into two categories: periods following significant volcanic injections of aerosol and its major precursor SO_2 in the tropics; and periods well away from these events, where the source of aerosol is more complex and generally less well understood. Figure 2 shows a schematic of the stratospheric aerosol lifecycle. Since the advent of global measures in 1979 the former category can be assigned to a highly volcanic period stretching between the 1982 eruption of El Chichón (Mexico) and the end of the recovery of the 1991 Mt. Pinatubo eruption in the late 1990s. The less volcanic, though not free of volcanic effects on the stratosphere, can be assigned to the period before 1982 and more extensively to the period following 2000 up to the present. Episodic and rare extreme fire events can also occasionally reach the stratosphere as was observed after the mass Black Saturday fire in Australia in February 2009.

During the volcanic period, the global stratospheric aerosol processes were controlled by five or six eruptions. In addition to Pinatubo and El Chichón, they are Nevado del Ruiz (1985, Colombia), Nyamuragira (1986, Democratic Republic of the Congo), Kelut (1990, Indonesia), and Cerro Hudson (1991, Argentina). The 1991 Cerro Hudson eruption is the only event occurring outside of low latitudes and occurred shortly after the Pinatubo eruption which had the largest stratospheric impact of any eruption in the twentieth century. As a result, it was only distinctly discernible in the measurement dataset for a few months before being masked by the massive Pinatubo event. The low latitude events can have the most long-lasting influence on stratospheric aerosol levels depending on the altitude to which the volcanically injected SO_2 is injected. This injection is observed by a number instrument such as the total ozone mapping spectrometer and similar nadir-viewing ultraviolet spectrometers since SO_2 has an absorption feature in the

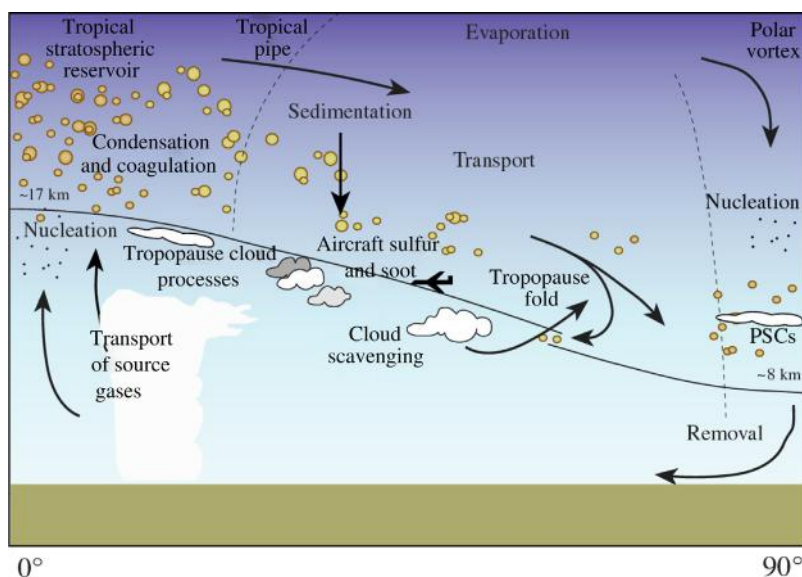


Figure 2 A figure showing the lifecycle of stratospheric aerosol and the primary processes controlling it. From SPARC (2006).

ultraviolet bands where ozone is measured. Aura/MLS has demonstrated a capacity for SO_2 profiling of stratospheric volcanic SO_2 plumes. In the case of Kelut, the impact was mostly felt between the tropopause and 20 km and aerosol from this event was transported from the tropics to higher latitudes within a few months from which it was transported into the troposphere where it is eventually washed out of the atmosphere. On the other hand, the other low latitude events had significant SO_2 injections at altitudes above 20 km into a region of the stratosphere referred to as the 'leaky tropical pipe.' This is a region in which transported to higher latitudes is suppressed and where a volcanic enhancement can be maintained for many years. In the case of Pinatubo, the tropical pipe remained clearly enhanced as late as 1999 or more than 7 years following the eruption. For this reason, the tropical pipe is sometimes colloquially referred to as the tropical aerosol reservoir as it clearly acts as such particularly in the volcanic period. Figure 3 shows the stratospheric optical depth (vertically integrated aerosol extinction coefficient) between 1979 and 2005 and shows the persistence of an enhanced tropical aerosol reservoir following the Pinatubo eruption.

Once the SO_2 is injected it undergoes a photochemical processing that eventually produces gaseous sulfuric acid (H_2SO_4). Since sulfuric acid has a low saturation vapor pressure, it readily condenses onto existing aerosol or nucleates new aerosol using homogeneous (without a condensation nuclei) or using ambient solid particles as nuclei including associated volcanic ash, meteoric dust, and organic particles. Once this aerosol has formed, particle number densities can be much higher than that typically observed in the stratosphere ($\sim 10 \text{ cm}^{-3}$) and the aerosol actively coagulates into a progressively larger aerosol. As a result, it is commonly observed that aerosol size is strongly correlated with bulk aerosol levels (measured by mass or other quantity).

In the immediate aftermath of a major eruption, sedimentation of the largest aerosol can transport significant amounts of aerosol out of the stratosphere and into the upper tropical troposphere where it can be permanently removed from the atmosphere by washout. While the upper tropical troposphere is generally a very low aerosol level region, after the Pinatubo eruption, the level of aerosol in this region increased many fold

primarily as the result of sedimentation. As aerosol levels recover toward preeruption levels, large aerosols are preferentially removed by this process and as aerosol levels decline *in toto*, the particle size also decreases and sedimentation becomes an ineffective mechanism for the removal of aerosol from the stratosphere. Aerosol within the tropical pipe is then removed by two mechanisms: meridional transport to higher latitudes and vertical transport to higher altitudes.

While transport from the tropical pipe is suppressed, it is still a critically important mechanism for the transport of material to higher latitudes. It is generally most efficient into the winter hemisphere at all altitudes but particularly below 20 km. At higher latitudes, the efficacy of this transport is governed by the phase of the quasi-biennial oscillation (QBO) which is a periodic oscillation of the equatorial zonal wind between easterlies and westerlies with a mean period slightly longer than 2 years. In the easterly phase of the QBO, transport out of the tropics is suppressed while in the westerly phase, transport is somewhat more efficient. As a result, the rate at which aerosol is transported to higher latitudes shows a pronounced QBO signal away from tropical latitudes. The efficacy of transport as driven by the phase of the QBO can cause a delay in the arrival of aerosol from even a large event, which can lead to some confusion regarding the source of aerosol observed by ground-based site months or perhaps more than a year after an eruption.

Within the stratosphere, air is generally transported upward in the tropics and then poleward where it is transported downward eventually crossing the tropopause from the stratosphere into the troposphere. This circulation is known as the Brewer–Dobson circulation and this circulation plays a significant role in defining the distribution of aerosol in the stratosphere. The QBO modulates the generally upward Brewer–Dobson vertical transport in the tropical stratosphere. The vertical transport is more efficient during the easterly phase of the QBO and suppressed during the westerly phase. Following a major eruption like Pinatubo, heating of the stratosphere due to the presence of the aerosol can increase vertical transport leading to enhanced aerosol levels to altitudes as high as 35 km. Aerosol is not a passive tracer of vertical transport as sedimentation continues to play a role such that

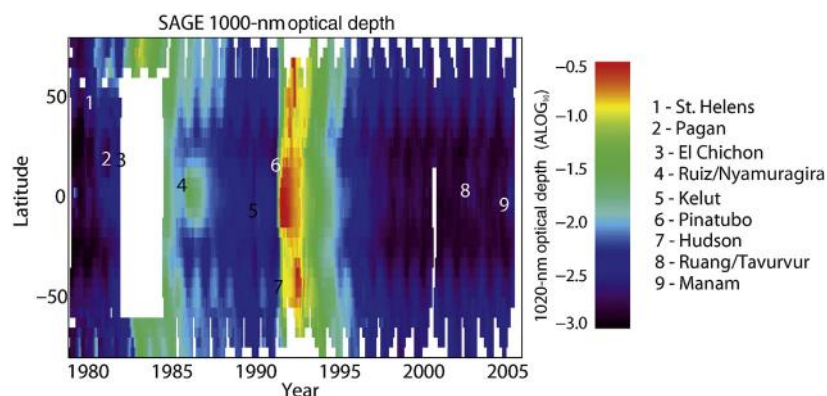


Figure 3 This figure shows the SAM II/SAGE I/SAGE II 1020-nm stratospheric aerosol optical depth record for 1979–2005 with the times of significant stratospheric injections denoted on the figure. The role of the leaky tropical pipe as a reservoir of aerosol from intense events in the tropics is clearly visible particularly in the period between 1985 and 2000.

aerosol is transported upward more slowly than would be expected by the observed vertical velocities. In addition, since the stratosphere warms with increasing altitude, its capacity to hold gaseous sulfuric acid increases and the sulfate aerosol evaporates more with increasing altitude and the volatile part of the aerosol is essentially completely evaporated by 30–35 km where the ‘top’ of the sulfate aerosol is dependent on the overall aerosol levels. This general morphology is shown in Figure 4. The combination of sedimentation and evaporation gives rise to the observed aerosol layer within the stratosphere. This is often referred to as the ‘Junge layer’ after the scientist who first observed the layer. The gaseous sulfuric acid undergoes photochemical modification and returns to an SO_2 reservoir. This gas rather than material is transported to higher latitudes, where it is transported downward to the lower stratosphere, where it returns first to gaseous sulfuric acid and then recondenses to sulfate aerosol. While the downward transport at high latitudes occurs year-around, it is most effective in the polar winter when diabatic cooling forms the polar vortex (the region in which polar stratospheric clouds form and the ozone hole forms). Since the interior of the polar vortex contains air with very little sulfate aerosol and isolated from aerosol-bearing midlatitude air, the aerosol formation process is often noted near 25 km in wintertime *in situ* aerosol datasets.

Aerosol is removed at high and midlatitudes by a combination of processes. In addition to the downward motions across the tropopause associated with Brewer–Dobson circulation, sedimentation can still be effective once the aerosol is in the vicinity of the tropopause. Bulk stratosphere to troposphere exchange, such as that occurs around tropopause folds, can also remove aerosol. Once in the troposphere, most of the aerosol is scavenged by clouds and precipitation and permanently removed from the atmosphere.

Stratospheric aerosol levels even during the nominally ‘nonvolcanic’ periods are still heavily influenced by small injections of volcanic materials into the tropical UTLS and by direct injection of aerosol into the midlatitude and lowermost polar stratosphere. In the tropics, in addition to periodic volcanic influences, there are also nonvolcanic sources of sulfur-bearing gases which help to define aerosol levels during these relatively low aerosol periods. The most important of these is carbonyl sulfide (OCS) whose source is a combination of natural (primarily oceanic) and human-derived sources. OCS is long-lived in the troposphere and there is significant transport of OCS to the stratosphere across the tropical tropopause. Once in the stratosphere, it primarily remains unchanged until it reaches an altitude of ~ 25 km where it is photolyzed and eventually its sulfur is converted to sulfate aerosol. The lifetime

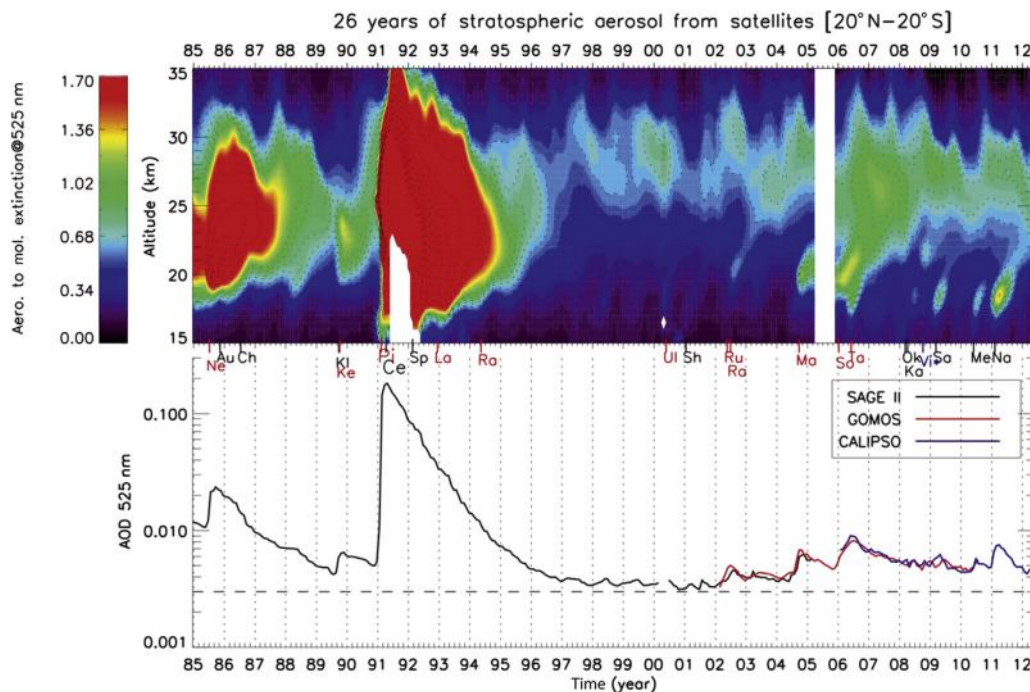


Figure 4 (a) Monthly mean extinction ratio (525 nm) profile evolution in the tropics (20°N – 20°S) from January 1985 to November 2012 derived from (left) SAGE II extinction in 1985–2005 and (right) CALIOP scattering ratio in 2006–2012, after removing clouds below 18 km based on their wavelength dependence (SAGE II) and depolarization properties (CALIOP) compared to aerosols. Black contours represent the extinction ratio in log-scale from 0.1 to 100. The position of each volcanic eruption occurring during the period is displayed with its first two letters on the horizontal axis, where tropical eruptions are noted in red. The eruptions were Nevado del Ruiz (Ne), Augustine (Au), Chikurachki (Ch), Kliuchevskoi (KI), Kelut (Ke), Pinatubo (Pi), Cerro Hudson (Ce), Spur (Sp), Lascar (La), Rabaul (Ra), Ulawun (UI), Shiveluch (Sh), Ruang (Ru), Reventador (Ra), Manam (Ma), Soufrière Hills (So), Tavurvur (Ta), Chaiten (Ch), Okmok (Ok), Kasatochi (Ka), Fire/Victoria (Vi*), Sarychev (Sa), Merapi (Me), and Nabro (Na). Updated from Figure 1 of Vernier et al. (2011). (b) Mean stratospheric aerosol optical depth in the tropics (20°N – 20°S) between the tropopause and 40 km since 1985 from the Stratospheric Aerosol and Gas Experiment (SAGE II) (black line), the global ozone monitoring by occultation of stars (GOMOS) (red line), and CALIOP (blue line). Updated from Figure 5 of Vernier et al. (2011).

of tropospheric SO₂ (from either human-derived or natural sources) is extremely short primarily due to its solubility in water. However, it is produced in prodigious quantities and some of this material is transported into the stratosphere. Observations by *in situ* sensors suggested that substantial new particle formation occurs in the upper tropical troposphere and so it is possible that much of the SO₂ entering the stratosphere has already been converted into aerosol. While substantial uncertainties surround the amount of SO₂ directly entering the stratosphere, it appears at this time that SO₂ and OCS each account for about half of the flux of sulfur into the stratosphere.

Understanding the transport of sulfur across the tropical tropopause is difficult. This is primarily due to difficulty in measuring SO₂ at background, nonvolcanically enhanced levels at the tropopause. This measurement is currently only possible using airborne instrumentation that strongly limit our knowledge of the sulfur budget in this climate-influential region. This is due to the fact that SO₂ lacks strong spectral features (which are not masked by other absorbing gases) and measurements possible during large volcanic events are not possible for SO₂ levels 1000 times or more smaller. Recent advances by the MIPAS science team show that the basic knowledge of SO₂ morphology is correct. Enhanced SO₂ is observed up to the tropical tropopause above which it declines rapidly as it is apparently converted to aerosol. Between 25 and 30 km it is also enhanced apparently by the conversion of OCS to SO₂ on its way to becoming sulfate aerosol. Above 35 km, it is again enhanced as the evaporation processes leads to the photochemical conversion of sulfuric acid to SO₂. The ability to produce these measurements is a significant achievement. There are a number of other gases that may contribute to the flux of sulfur into the stratosphere. These include dimethyl sulfide and CS₂. While they are unlikely to enter the stratosphere directly in any significant amount, they can be converted into both OCS and SO₂ and contribute to the stratospheric sulfur budget in a small way in those forms.

A variety of organic and inorganic compounds contribute to the overall aerosol loading particularly in the lower tropical stratosphere where they may form a significant fraction of the observed aerosol. Recent *in situ* measurements in the upper troposphere during several field campaigns (TROCCINOX, SCOUT-AMMA, SCOUT-O3) have revealed the presence of newly formed particles at and below the tropopause (350–370 K). In relation to those findings, aerosol composition measurements using mass spectrometers suggest that a large fraction of this aerosol population could be composed of organic materials. This modifies substantially the idea that aerosol near the tropopause and in the lower stratosphere are primarily sulfate. More research is needed to fully understand

the composition and lifetime of those aerosols, which can have a profound impact on the lifetime of thin cirrus clouds in the upper troposphere and climate. In addition, fire products particularly from biomass burning by natural human-directed sources has been shown to contribute to the overall aerosol levels. This has particularly been shown in the CALIOP data following the Victoria (Australia) megafire of 2009 when products of the fire were seen to gradually ascend into the stratosphere to at least 20 km. This has also been observed at high latitudes where boreal fires create pyrocumulonimbus clouds capable of penetrating the lower stratosphere. The long-term significance of these injections is not well known. On the other hand, measurements in the lower stratosphere show that up to half of all aerosols show evidence of the presence of meteoritic material. This material enters the stratosphere from the mesosphere and may be a key nuclei for the formation of sulfate aerosol.

See also: **Aerosols:** Climatology of Tropospheric Aerosols; Role in Climate Change. **Climate and Climate Change:** Volcanoes: Role in Climate.

Further Reading

- Deshler, T., Hervig, M.E., Hofmann, D.J., Rosen, J.M., Liley, J.B., 2003. Thirty years of *in situ* stratospheric aerosol size distribution measurements from Laramie, Wyoming (41° N), using balloon-borne instruments. *Journal of Geophysical Research* 108 (D5), 4167. <http://dx.doi.org/10.1029/2002JD002514>.
- Junge, C.E., Chagnon, C.W., Manson, J.E., 1961. Stratospheric aerosols. *Journal of Meteorology* 18, 81–108.
- Murphy, D.M., Thomson, D.S., Mahoney, M.J., 1998. *In situ* measurements of organics, meteoritic material, mercury, and other elements in aerosols at 5 to 19 kilometers. *Science* 282, 1664–1669.
- Reeves, J.M., Wilson, J.C., Brock, C.A., Bui, T.P., 2008. Comparison of aerosol extinction coefficients, surface area density, and volume density from SAGE II and *in situ* aircraft measurements. *Journal of Geophysical Research* 113 (D10202), 8. <http://dx.doi.org/10.1029/2007JD009357>.
- Rinsland, C.P., Yue, G.K., Gunson, M.R., Zander, R., Abrams, M.C., 1994. Mid-infrared extinction by sulfate aerosols from the Mt. Pinatubo eruption. *Journal of Quantitative Spectroscopy and Radiative Transfer* 52, 241–252.
- SPARC. 2006. Assessment of Stratospheric Aerosol Properties (ASAP), SPARC Report No. 4, WCRP-124, WMO/TD-No. 1295, Feb. 2006, L. Thomason and Th. Peter, Eds.
- Thomason, L.W., 2012. Toward a combined SAGE II-HALOE aerosol climatology: an evaluation of HALOE version 19 stratospheric aerosol extinction coefficient observations. *Atmospheric Chemistry and Physics* 12, 8177–8188. <http://dx.doi.org/10.5194/acp-12-8177-2012>. www.atmos-chem-phys.net/12/8177/2012/.
- Trepte, C.R., Hitchman, M.H., 1992. Tropical stratospheric circulation deduced from satellite aerosol data. *Nature* 355, 626–628.
- Vernier, J.P., Thomason, L.W., Pommereau, J.P., Bourassa, A., Pelon, J., Garnier, A., Hauchecorne, A., Blanot, L., Trepte, C., Doug Degenstein, Vargas, F., 2010. Volcanic origin of the recent stratospheric aerosol trend. *Geophysical Research Letters* 38, L12807. <http://dx.doi.org/10.1029/2011GL047563>, 2011.

Climatology of Tropospheric Aerosols

N Bellouin, University of Reading, Reading, UK

J Haywood, Met Office, Exeter, UK

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by J L Gras, volume 1, pp 13–20, © 2003, Elsevier Ltd.

Synopsis

Aerosols are solid or liquid particles either directly emitted into the atmosphere or converted from gaseous precursors. Both natural processes and human activities generate aerosols. They affect visibility, air quality, formation of clouds, and the energy budget of the Earth. Aerosols remain in the troposphere for up to 2 weeks, where they experience chemical transformation and long-range transport.

Introduction

Atmospheric aerosols are solid and liquid particles in suspension in the air, ranging in size from a few nanometers to tens of microns (10^{-9} – 10^{-5} m), with the exception of cloud droplets and ice crystals. Natural sources of aerosols include sea salt generated from breaking waves, mineral dust blown from the surface by wind, and organic aerosols from biogenic emissions. Artificial, also called anthropogenic, aerosols include sulfate, nitrate, and carbonaceous aerosols, and are mainly from fossil fuel combustion sources. The first aerosol studies were motivated by visibility, and in 1880 the British scientist John Aitken correctly postulated that aerosol particles act as nuclei for the condensation of water vapor to form fog and clouds. In the second part of the twentieth century, aerosol particles were linked to an increase in human respiratory diseases and acid rain, prompting the need to improve air quality by reducing emissions of aerosols and their precursors. At the same time, scientists started quantifying the impact of aerosols on the Earth's energy budget and simple aerosol representations were introduced in weather forecast and climate models. Climatologies of aerosols in the troposphere have been developed over time, tailored to the diverse interests in aerosol properties. Climatologies of aerosol surface concentrations and chemical composition originate from air quality monitoring networks such as the European Monitoring and Evaluation Programme and the Clean Air Status and Trends Network of the United States Environmental Protection Agency. Dedicated instrumented aircraft campaigns provide regional snapshots of aerosol composition and size across the lower atmosphere, including information on aerosol microphysical properties and vertical profiles. Networks of remote sensing instruments have been deployed at selected sites worldwide, forming for example the Aerosol Robotic Network (AERONET). AERONET sun photometers provide measurements of the column-integrated extinction of visible or near-infrared radiation by aerosols, called optical depth, and these measurements can be inverted to provide an estimate of aerosol size. The launch of dedicated satellite instruments beginning in the late 1990s provides a global view of aerosol distributions of optical depth. More recently, vertical aerosol distribution has been routinely observed by lidars, located on the ground, aboard aircrafts, or on satellite platforms. In parallel to improved observational capabilities, numerical models have developed and provide daily aerosol forecasts, increasingly often initialized from

assimilation of satellite observations, as well as centennial simulations for climate projections. Numerical models vary in the number of aerosol species they include and in the number of aerosol characteristics they are able to represent depending on the need for computational speed.

In spite of those diverse sources of data, there are a number of difficulties associated with obtaining the aerosol climatology that fits a given application. By definition, aerosols encompass a large range of particle sizes and chemical compositions. Their sources are very diverse and unevenly distributed across the globe. Aerosols are removed from the atmosphere primarily by spatially inhomogeneous precipitation processes and, to a smaller extent, turbulence. Thus, aerosols have a relatively short residence time in the troposphere of typically up to 1 or 2 weeks. Consequently, aerosol concentrations vary greatly in time and space, both horizontally and vertically, in contrast to well-mixed greenhouse gases like carbon dioxide. Well-defined aerosol plumes can often be identified in observations, with aerosol number and mass varying by orders of magnitudes within a few kilometers in the horizontal and a few hundred or even a few tens of meters in the vertical across the plume. Measurements at the surface may therefore not be indicative of aerosol concentrations and properties higher up in the atmosphere. There are also several ways to characterize aerosols. Since aerosols are typically present in large numbers in the atmosphere, their population is described by its size distribution, which gives the number, surface area, or volume of the particles as a function of their radii. As depicted in [Figure 1](#), a typical tropospheric aerosol size distribution exhibits several maxima, called modes. The smaller particles, with radii smaller than 50 nm, belong to the nucleation and Aitken modes and provide most of the total aerosol number. The accumulation mode covers particles with radii up to 0.5 μm , which dominate the aerosol surface area. Both Aitken and accumulation modes are often called fine modes to contrast with the coarse mode, which comprises particles larger than 0.5 μm and makes up the bulk of aerosol volume. Minima in the size distribution arise from coagulation processes, where smaller particles of favorable sizes aggregate into larger particles. Air quality networks routinely report aerosol concentrations for particles with aerodynamic diameters smaller than 2.5 and 10 μm , labeled Particulate Matter (PM) 2.5 and PM10, respectively. Air quality regulations impose thresholds on both the annual average and daily average of PM concentrations in many countries. In Europe, daily average PM10

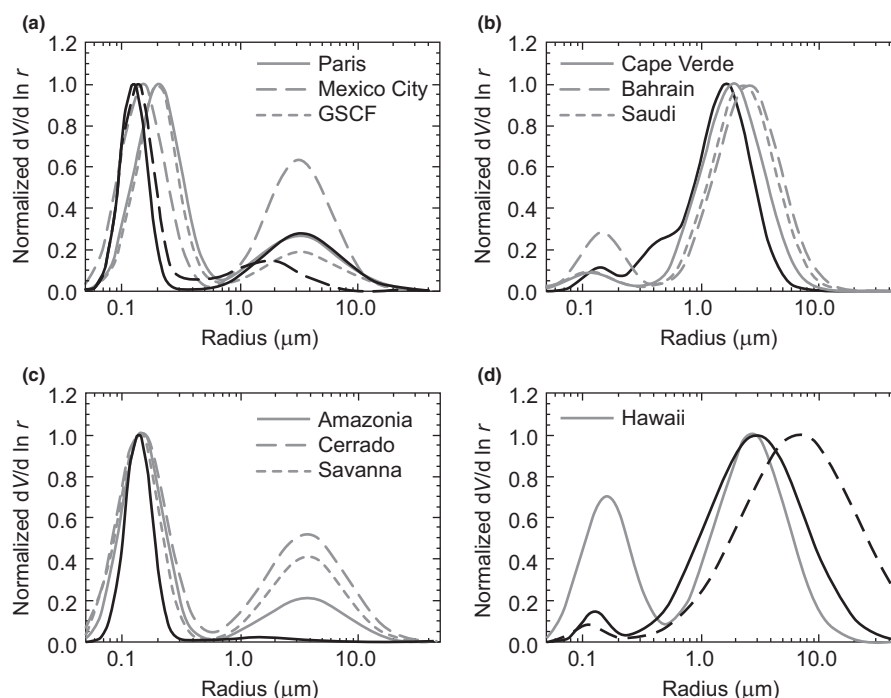


Figure 1 Distribution of the volume of aerosol particles as a function of their radii, as sampled by the Facility for Airborne Atmospheric Measurements aircraft (black lines) and inverted from ground-based sun-photometer measurements of the Aerosol Robotic Network (gray lines). Panels gather measurements of generic aerosol types at different locations: (a) Industrial aerosol, (b) Mineral dust aerosol, (c) Biomass-burning aerosol, and (d) Marine aerosol. Osborne, S.R., Haywood, J.M., 2005. Aircraft observations of the microphysical and optical properties of major aerosol species. *Atmospheric Research* 73, 173–201. © Crown.

concentrations should not exceed $50 \mu\text{g m}^{-3}$ for more than 35 days per year. In newly industrialized countries, where air quality regulation is weaker, PM10 concentrations routinely exceed $100 \mu\text{g m}^{-3}$ on an annual average.

Aerosols are also characterized by their production processes and chemical composition. Primary aerosols are emitted into the atmosphere directly as particles. They include the carbonaceous by-products of incomplete combustion processes and wind-blown mineral dust aerosols. In contrast, secondary aerosols originate from gas-to-particle conversions. Gaseous precursors, such as sulfur dioxide and ammonia, are oxidized in the dry atmosphere or in clouds and become soluble in water. Eventually, a large fraction of the emitted precursor enters the aerosol phase, dissolved in a droplet. As such, water represents a sizable fraction of the total aerosol mass. Gas-to-particle conversion processes also mean that aerosols are part of the chemistry of the atmosphere and indirectly affect the concentrations of gaseous species such as ozone and methane. Aerosol populations are typically of diverse chemical composition and their mixing state can be both internal, where the mixed composition lies within the same particle, and external, where each particle has a chemically distinct composition. Particle size can be a simplified marker of the production process. Fine-mode aerosols often originate from gas-to-particle conversion or combustion processes. Coarse-mode particles typically originate from mechanical processes such as wind friction. The shape of the particles also depends on the production process: aerosol solutions are spherical droplets. Combustion aerosols, grains

of mineral dust, and pollens have more complex, irregular shapes.

Aerosols can finally be characterized by their natural or anthropogenic origin. This distinction is difficult to obtain from observations alone, but is important for climate studies. Aerosols emitted into the atmosphere by human activities since the start of the industrial revolution have scattered and absorbed additional solar and thermal radiation. They also have modified cloud droplet sizes and number, again affecting solar radiation, albeit indirectly. In effect, anthropogenic aerosols have changed the radiative balance of the Earth in a process called radiative forcing. Although uncertain, this forcing is estimated to be negative, and the climate system responds by cooling. Estimates for the year 2000 suggest that aerosol forcing compensates two-thirds of the positive forcing by long-lived greenhouse gases. Local climate changes driven by anthropogenic emissions can also affect emissions of natural aerosols, desertification being a typical example. In that case, the anthropogenic or natural classification of the aerosol is not obvious.

Sources of Tropospheric Aerosols

In this section, we distinguish four aerosol categories that can be identified in observations, and are represented in numerical models: fine-mode anthropogenic aerosols, mineral dust, marine aerosol, and fine-mode natural aerosols. Emissions of tropospheric aerosols are estimated by national inventories,

Table 1 Approximative present-day range of anthropogenic and natural emission rates of primary aerosols and aerosol precursors. Units are Tg per year, as mass of sulfur for sulfur dioxide and DMS, mass of nitrogen for ammonia, and mass of carbon for VOCs and carbonaceous aerosols

Primary aerosol or precursor	Anthropogenic emission rate	Natural emission rate
Sulfur dioxide (SO ₂)	50–90	10
Ammonia (NH ₃)	20–50	10
VOCs	5–40	80–200
Carbonaceous aerosols from fossil fuels	20–50	
Carbonaceous aerosols from biomass burning	50–90	20–40
Mineral dust aerosols	40–130	1000–3000
Sea-salt aerosols		2000–10 000
DMS		10–60
Biogenic bacteria and pollens		100–1000

VOCs, Volatile organic compounds; DMS, Dimethylsulfide.

derived from satellite observations, or inverted from numerical models with assimilation of satellite observations. Those estimates are uncertain and typically within a factor two to three of each other. [Table 1](#) summarizes the current emissions of aerosol precursors and primary aerosols. For most species, anthropogenic emission rates dominate total emissions. Consequently, anthropogenic aerosols dominate aerosol mass in most industrialized regions. They also contribute largely to aerosol number: polluted cities can exhibit more than 10^5 particles cm^{-3} , compared to 100 cm^{-3} in oceanic background. Aerosols and their precursors are injected into the atmosphere from the surface, with the exception of

chimney-level industrial emissions, forest fires, and volcanic emissions, which are injected at several hundreds of meters because of buoyancy effects or local topography. Gas-to-particle conversions can however happen in the free troposphere, where nucleation of new aerosol particles has been observed. Emission rates only give a partial view of aerosols: their residence time in the atmosphere and other properties relevant to different aerosol impacts are also important. In climate studies, for example, the strength of aerosol interaction with radiation, and the ability of specific aerosol species to form cloud condensation nuclei must be taken into account.

Fine-Mode Anthropogenic Aerosols

[Figure 2](#) shows the distribution of aerosol optical depth attributed to fine-mode anthropogenic aerosols in a numerical model using data assimilation of satellite aerosol retrievals. North America, Europe, and Asia are associated with a sizable increase in optical depth between winter and summer months. Ground measurements show that those regions are associated with primary aerosols and aerosol precursors that are by-products of industrial and agricultural activities, and the combustion of fossil fuels. Industrial aerosols typically reside in the atmosphere for 3–5 days.

Anthropogenic aerosol precursors are sulfur dioxide, ammonia, and volatile organic compounds (VOCs). Sulfur dioxide oxidizes into sulfuric acid and, in most of the troposphere, ammonium sulfate aerosols. This photolytic process is more efficient during the summer months. An emission inventory compiled in 2010, summarized in [Figure 3](#), suggests that 42% of anthropogenic sulfur dioxide emissions come from power plants, 29% from industries,

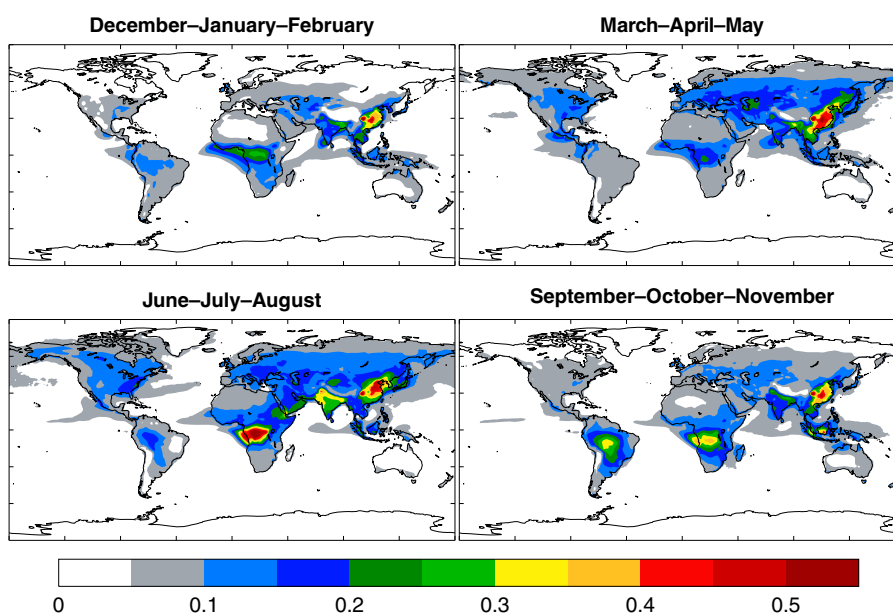


Figure 2 Seasonal distributions (2003–10) of aerosol optical depth at $0.55 \mu\text{m}$ attributed to fine-mode anthropogenic aerosols in the Monitoring Atmospheric Composition and Climate reanalysis of the European Centre for Medium-Range Weather Forecast forecast model with assimilation of Moderate-resolution Imaging Spectroradiometer total aerosol optical depth retrievals.

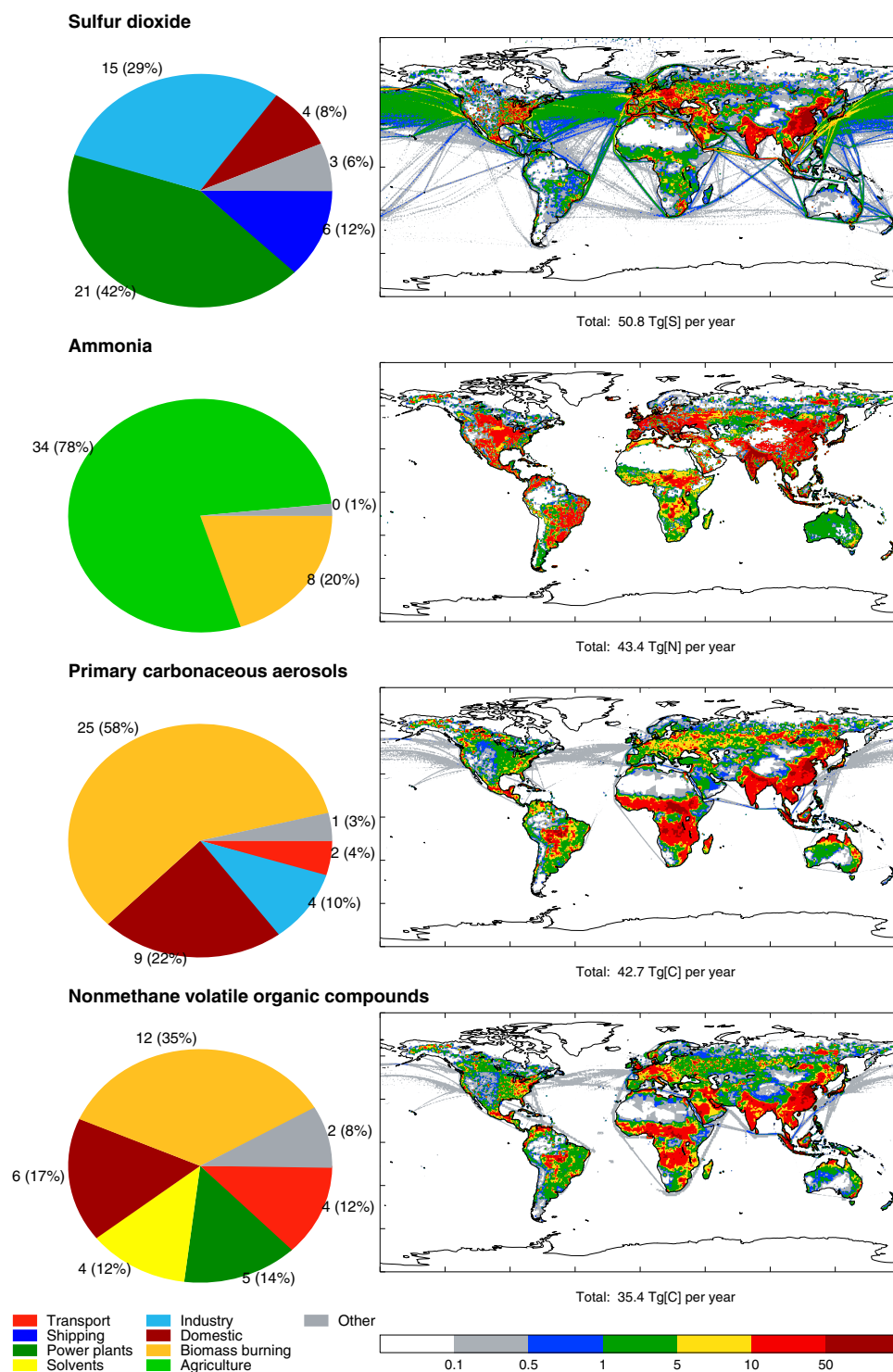


Figure 3 Anthropogenic emissions of aerosols and aerosol precursors for the year 2010, according to the datasets developed as part of the Climate Model Intercomparison Project Phase 5. Left column: sector-based charts, in Tg per year. Right column: Distributions of total emissions, in $\text{ng m}^{-2} \text{s}^{-1}$. Emissions of sulfur dioxide, ammonia, and carbonaceous aerosols and volatile organic compounds are given as masses of sulfur, nitrogen, and carbon, respectively.

and 12% from shipping, with the remainder mostly provided by the transport and domestic sectors. Other inventories may provide different sector-based partitions and total emissions. Ammonia and NO_x oxidize into ammonium nitrate aerosols. Ammonia emissions are estimated at 78% by the agricultural sector, from livestock and fertilizers (Figure 3). VOCs are precursors of secondary organic aerosols. Anthropogenic VOC emissions originate from the transport, energy, solvents, and domestic sectors. The combustion of fossil fuel and biofuel also emits primary carbonaceous aerosols in the atmosphere. Those aerosols are a mixture of pure soot particles, named black carbon aerosols, and particles made of carbon bound in an organic compound, named organic carbon aerosols. Due to the use of cooking stoves, black carbon aerosols are particularly abundant in India. Fossil fuel carbonaceous emissions come mainly from the domestic, industrial, and transport sectors (Figure 3).

Anthropogenic aerosols in South America, Africa, and Indonesia are different and dominated by particles associated with the incomplete combustion of biomass. The seasonality of those aerosols is driven by the clearance of forested areas and the fertilization of crop fields at the start of the growing season. The biomass-burning aerosol season starts in December in western Africa, then gradually moves southward, ending around September in the Congo basin and southern Africa. In South America and Indonesia, the season tends to start in late August, peaks in September, and ends in October, although there can be significant year-year variability in the timing. Aerosol concentrations close to large fires can be very high, up to 1 mg m^{-3} . Annual emissions of biomass-burning aerosols mainly include primary emissions of black and organic carbon aerosols, with the former representing about 10% of total emissions, but also sizable emissions of aerosol precursors such as ammonia and VOCs

(Figure 3). Biomass-burning emissions partly depend on political aspects and exhibit large interannual variability, especially in South America. According to numerical models, biomass-burning aerosols typically reside in the atmosphere for 6–7 days.

Mineral Dust Aerosols

Mineral aerosols are produced by the friction of near-surface winds onto bare soils, and are mainly coarse-mode particles. When wind speeds exceed a critical threshold, mineral particles are emitted from the surface. Smaller particles remain suspended in the atmosphere, while larger particles impact the surface and release smaller particles into the air in a process known as saltation. In addition to the seasonality of wind regimes, dried and less vegetated soils are more prone to mineral dust production. Preferential sources for mineral dust are low-lying regions of silts and sediments, such as dry river and lake beds like the Bodele depression in Chad. Dust emissions exhibit large daily and interannual variability. Figure 4 shows the seasonal distributions of mineral dust optical depth. Emissions from deserts in the Sahara, Arabian Peninsula, and Iran peak in the summer. Deserts in Inner China emit in the spring. Smaller sources include deserts in southern Africa, Argentina, and Australia. The Sahara desert is the main contributor to global emissions. Cement production and related industrial activities also release mineral dust into the atmosphere, thus representing a small, anthropogenic source of mineral aerosols. Residence time of mineral aerosols is strongly dependent on their size: although the larger particles only reside for 1–2 days, smaller particles can remain in the atmosphere for over a week as evidenced by the deposition of mineral dust from the Sahara in the Caribbean and South America.

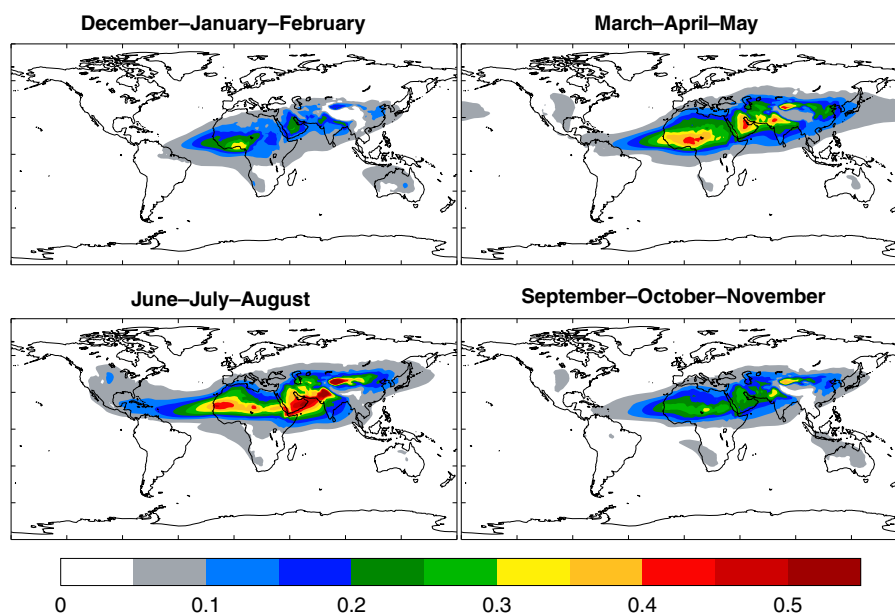


Figure 4 As Figure 2, but for mineral dust aerosols.

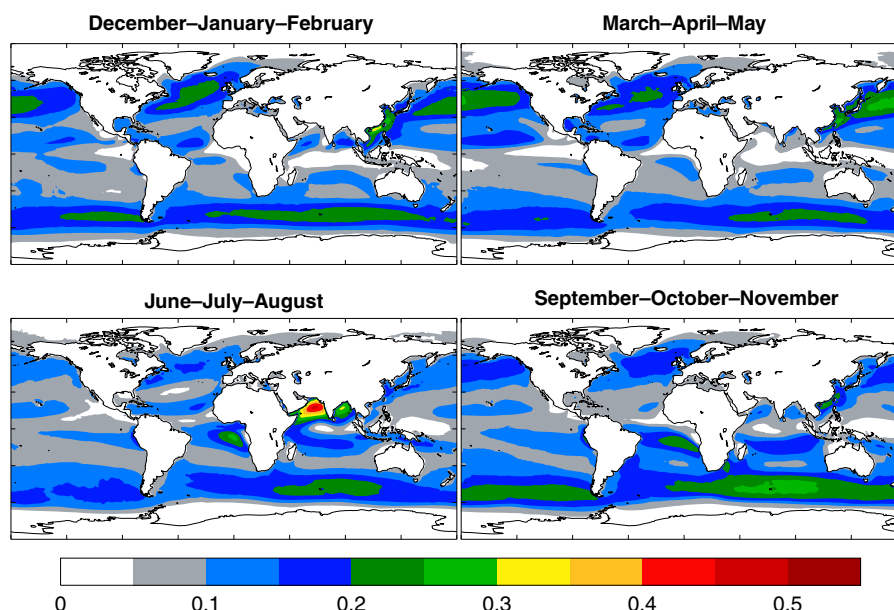


Figure 5 As Figure 2, but for marine aerosols.

Marine Aerosols

Aerosols produced at the ocean–atmosphere interface are dominated by sea-salt particles. Those aerosols are generated from breaking oceanic waves that are themselves driven by near-surface winds. As the seawater droplets emitted in the atmosphere evaporate, the salt concentration of the solution increases. Figure 5 shows the seasonal distributions of marine aerosol optical depth. In the Southern Hemisphere, the latitude band between 40 and 50° S is associated with the strong westerly winds called Roaring Forties and is a region of large sea-salt emissions. Sea-salt aerosol residence time in the atmosphere is typically short, less than a day, but can be longer in cloud-free oceanic regions with low precipitations.

Other aerosols are emitted at the air–sea interface. Firstly, some species of phytoplankton increase the concentrations of dimethylsulfide (DMS) in the ocean. This occurs in regions of high primary productivity, sufficient solar radiation, and shallow mixed-layer depth: high-latitude oceans of both hemispheres during the corresponding summer. DMS makes its way into the atmosphere and may be oxidized to sulfur dioxide, which in turn may form sulfate aerosols. Secondly, the ocean is a strong natural source of ammonia. Thirdly, sources of organic- and iodo-carbon aerosols from phytoplankton blooms and algae have recently been postulated, although their global emissions are currently unknown.

Fine-Mode Natural Aerosols

A variety of natural processes can emit aerosol precursors and fine-mode aerosols into the atmosphere. First, wild fires started by lightning emit biomass-burning aerosols, although those emissions are often counted together with the anthropogenic emissions due to the difficulty of determining the origin of forest fires. Fires are more likely in the local dry

season, and can be large in the boreal forests of Canada and Russia.

Vegetation produces pollens and bacteria that are in effect primary aerosol particles. The strength of those emissions is highly uncertain, but may represent several hundred teragrams per year. Plants also emit classes of VOCs called terpenes (from needle-leaf forests) and isoprenes (from broad-leaf forests) that can oxidize into secondary organic aerosols. The strength of those sources is also uncertain. Their seasonality is linked to the growing cycle of vegetation, with peak emissions from April to September in the Northern Hemisphere. Land-based natural ecosystems also emit ammonia and DMS.

Finally, although volcanoes are often viewed as a source of stratospheric aerosols, this is only true for the major explosive eruptions: Degassing volcanoes are a continuous source of sulfur dioxide in the troposphere, and minor eruptions emit volcanic ash plumes that can be a hazard to aviation.

Historical and Future Variations in Sources

Estimating past aerosol emissions is challenging, as there is virtually no observational basis. Empirical relationships in socio-economic, sector-based models are used to link aerosol emissions to changes in population and technologies. For future emissions, those models are used with prescribed scenarios of emission changes, experimenting with different assumptions on technological shifts, industrialization of developing countries, and population growth.

The industrial revolution has had a large impact on the sources of anthropogenic aerosols. From the middle of the nineteenth century in Europe and North America, emissions of sulfur dioxide and carbonaceous aerosols have increased steadily, at increased rates after the World War II. In Asia,

anthropogenic emissions started to increase in the middle of the twentieth century, and present emissions exceed European and American levels for some species like fossil fuel carbonaceous aerosols. Due to concerns in decreasing air quality and the impact of acid rains, emission controls for sulfur dioxide have been successfully implemented in most industrial countries, leading to decreases in sulfate aerosol concentrations. The decline of hazy conditions in Europe since the late 1970s has been linked to those aerosol decreases. However, they have been partially compensated by increases in nitrate aerosols driven by the increased use of fertilizers in agriculture. In the future, it is expected that stronger emission standards will decrease anthropogenic emissions further, although the industrialization of South America and Africa may lead to local increases.

Past biomass-burning emissions are likely to have been larger than present levels, as agricultural practices switched to synthetic fertilizers, and wild fires are now more efficiently suppressed. However, in countries where forest clearing dominates the emissions, those are assumed to have increased in line with the increase in population and are therefore larger in present days. Future changes in anthropogenic biomass-burning emissions are uncertain and depend on trends in population and agricultural technology changes.

Past and future emissions of natural aerosols depend on the variation of their driving factors like wind speed, vegetation and bare soil areas, temperature, soil moisture, and ocean biogeochemistry. Mineral dust emissions are expected to increase with desertification. Sea-salt emissions are expected to increase with the retreat of sea ice and the corresponding increase in areas of open ocean. Changes in ocean biogeochemistry and land vegetation and their impact on aerosol emissions are driven by near-surface temperature changes and are strongly linked to the hydrological cycle.

Long-Range Transport of Tropospheric Aerosols

Aerosols are transported away from source regions during the few days when they reside in the atmosphere. Consequently, a given region is affected not only by its local aerosol sources, but also by sources upwind. Long-range transport depends on the amount of aerosol emitted, weather systems, and aerosol chemistry, especially aging processes.

Transport occurs predominantly within the troposphere, either at low level in the boundary layer or aloft in the free troposphere. In midlatitude regions, transport is typically eastward with a smaller poleward component. Transport is driven by weather systems in winter and, especially over land, convective processes in summer. Both winds and precipitation are larger in winter than in summer, and influence long-range transport in opposite ways: stronger winds promote longer transport distances while larger precipitation reduces the aerosol residence time. Similar opposing effects occur when comparing low- and high-level transport. Winds are stronger aloft, especially in the jet streams, and accelerate the transport. However, wet deposition is also more powerful above the boundary layer, reducing residence times. Aerosol precursors can be transported away from their source, undergo oxidation during or at the end of the transport, and be received as secondary aerosols. The colder and drier conditions of the upper troposphere can enhance or inhibit the chemical reactions that form secondary aerosols, depending on the precursor. In the Tropics, vertical transport is driven by the rising and descending branches of the Hadley cell, and trade winds export the aerosols westward. Because atmosphere circulation is organized in relatively isolated cells, transport from the Tropics to the midlatitudes, and between the two hemispheres, is small. Figure 6 summarizes the winter and summer aerosol transport patterns.

In the Northern Hemisphere, North American aerosols are transported across the Atlantic toward Europe, European aerosols toward Central Asia and the Arctic, and aerosols from China are transported across the Pacific Ocean. Although local sources dominate anthropogenic aerosol concentrations, import from other regions due to long-range transport can still represent 30–40%.

Mineral dust transport from the Sahara is associated with a specific air layer, called the Saharan Air Layer, located at an altitude of about 6 km and characterized by warm temperatures and dry conditions. Mineral dust aerosols in that layer cross the Atlantic Ocean toward South America in winter and spring and, as the intertropical convergence zone moves northward, toward the Caribbean in summer. Mineral aerosols deposition in the ocean and the Amazon provides a large amount of nutrients (phosphorus, potassium, and iron) to marine and land ecosystems. On an annual average, deposition has been estimated at about 250 Tg in the Atlantic,

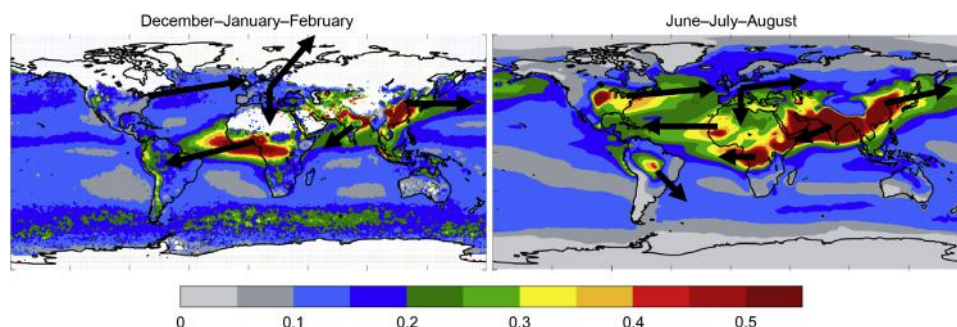


Figure 6 Summary of long-range aerosol transport as black arrows from the main aerosol sources, for north hemisphere winter (left) and summer (right). Background shows total aerosol optical depth at $0.55\ \mu\text{m}$ as derived from observations by the satellite-borne Moderate-resolution Imaging Spectroradiometer (left) and modeled in the Hadley Centre climate model HadGEM2-ES (right).

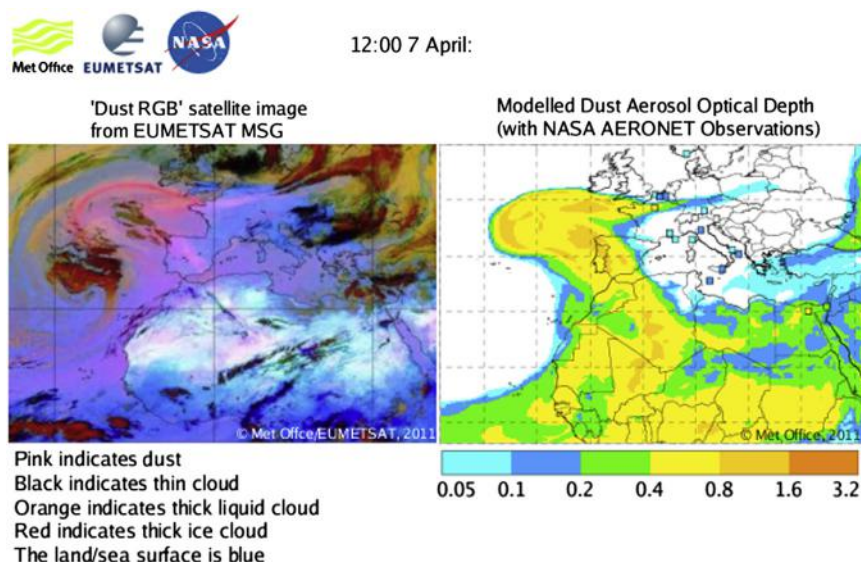


Figure 7 Transport of mineral dust aerosol from the Sahara to Western Europe on 7 April 2011 at 12 a.m. GMT. The left-hand side panel shows an analysis of Meteosat Second Generation satellite observations where mineral dust is highlighted in pink. The right-hand side panel shows the corresponding forecast of mineral dust optical depth by the UK Met Office numerical forecast model. © Crown.

20–30 Tg in the Caribbean, and 15–25 Tg in the Amazon. Less frequently, mineral dust aerosols are transported across the North Atlantic or the Mediterranean toward Europe, such as depicted on [Figure 7](#).

See also: **Aerosols:** Aerosol–Cloud Interactions and Their Radiative Forcing; Climatology of Stratospheric Aerosols; Role in Climate Change; Role in Radiative Transfer. **Biogeochemical Cycles:** Sulfur Cycle. **Chemistry of the Atmosphere:** Observations for Chemistry (*In Situ*): Particles. **Climate and Climate Change:** Volcanoes: Role in Climate. **Clouds and Fog:** Cloud Microphysics. **Satellites and Satellite Remote Sensing:** Aerosol Measurements. **Tropospheric Chemistry and Composition:** Aerosols/Particles.

Further Reading

Dentener, F., Keating, T., Akimoto, H. (Eds.), 2010. Hemispheric Transport of Air Pollution 2010, Part A: Ozone and Particulate Matter. Air Pollution Studies No. 17. United Nations, New York and Geneva.

Forster, P., et al., 2007. Changes in atmospheric constituents and in radiative forcing. In: Solomon, S., Qin, D., Manning, M., Chen, Z., Marquis, M., Averyt, K.B., Tignor, M., Miller, H.L. (Eds.), *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, Cambridge, UK and New York, NY.

Jimenez, J.L., et al., 2009. Evolution of organic aerosols in the atmosphere. *Science* 326, 1525–1529.

Kaufman, Y., Tanré, D., Boucher, O., 2002. A satellite view of aerosols in the climate system. *Nature* 419, 215–223.

Moss, H., et al., 2010. The next generation of scenarios for climate change research and assessment. *Nature* 463, 747–756.

Vautard, R., Yiou, P., van Oldenborgh, G.J., 2009. Decline of fog, mist and haze in Europe over the past 30 years. *Nature Geoscience* 2, 115–119.

Relevant Websites

aeronet.gsfc.nasa.gov – Aerosol Robotic Network.

cmip-pcmdi.llnl.gov/cmip5 – Coupled Model Intercomparison Project Phase 5.

www.emep.int – European Monitoring and Evaluation Programme.

modis-atmos.gsfc.nasa.gov – MODIS Atmosphere.

www.gmes-atmosphere.eu – Monitoring Atmospheric Composition and Climate.

www.htap.org – Task Force on Hemispheric Transport of Air Pollutants.

Synopsis

Mineral aerosols (often referred to as dust) are ubiquitous in the atmosphere. Several sources contribute to the dust loading such as soil-derived wind-blown (aeolian) dust, industrial processes, and volcanic eruptions. This article deals with the major impacts of mineral dust as well as its sources and properties, emphasizing the dust radiative effects.

Introduction

Atmospheric mineral aerosol, often referred to as dust, plays multiple important roles in the Earth systems, affecting the radiative energy balance, hydrological and biogeochemical cycles, land and ocean ecosystems, and human health. This article discusses the sources, properties, and major impacts of mineral dust, with an emphasis on the dust radiative effects.

Dust Sources and Production Mechanisms

Covering about 33% of the global land area, arid and semiarid regions are the major source of wind-blown (aeolian) dust. The Sahara Desert in North Africa is the single world's largest dust producing region, while deserts in East Asia (China and Mongolia) are the second largest source. Other important dust source regions are located in the Middle East, Central Asia, Australia, southwestern United States, and northwestern India. Although there are various similarities between arid and semiarid regions, active dust sources exhibit significant spatial heterogeneity and differ in dust emission strength. To date, there is no general theory of why some regions are more efficient active dust sources than other regions.

The dust emission is the complex, nonlinear process, which is controlled by a combined effect of meteorological characteristics (especially near surface winds) and land surface properties. Vegetation, roughness, soil texture, mineralogy, and moisture of the land surface combine and determine the threshold wind speed required to initiate the motion of dust particles. Advanced parameterizations of the dust emission process allow predicting the threshold wind speed as a function of the land surface characteristics. However, a single value of the threshold wind speed, for instance 6.5 m s^{-1} at 10 m height, is often used in the general circulation models or regional models, mainly because of the lack of input data on surface properties at the global scale.

Once winds are higher than a threshold value, the movement of dust particles can be initiated either by aerodynamic forces (called suspension) or by the impact of saltating soil grains (a process known as bombardment). Only the finest particles with diameters below about $60 \mu\text{m}$ can be suspended directly and transported upward by turbulent eddies. Particles in this size range, however, make up only a small fraction of soil grains, most of which are present in the form of aggregates of larger sizes due to strong interparticle cohesive forces. The

aggregates typically have diameters from about 60 to $80 \mu\text{m}$ to several hundred micrometers. The bombardment-induced breakage of aggregates results in the release of fine particles, which is thought to be a major contributor to aeolian dust entrainment into the atmosphere.

The dependence of dust emission on the land surface characteristics suggests that the modification of land surface (for instance, by human activities) may affect the dust emission and hence alter dust load in the atmosphere. Various human activities (such as agriculture, construction, deforestation, etc.) lead to land surface disturbances. Disturbed surfaces are likely to be more efficient dust sources than natural undisturbed lands. The dust fraction produced as a result of human activities is called anthropogenic dust. Recent estimates show that anthropogenic dust could be as much as 30–50% of total dust load depending on a region, but this remains highly uncertain. In addition, climate variability or climate change may alter dust emissions. Frequent droughts and strong winds increase the generation of dust, while rains tend to inhibit it.

Dust Loadings in the Atmosphere

Dust is a major component of atmospheric aerosols in many parts of the world. Both observations and model simulations indicate that dust load exhibits a complex spatial (horizontal and vertical) and temporal distribution in the atmosphere. The atmospheric loading of dust is controlled by the source strength, dust production mechanism, and removal processes acting upon dust particles during the atmospheric transport.

First of all, initial dust emission has complex spatial and temporal variations because dust sources have a dispersed geographical distribution and their strengths vary with time. The duration of dust storms is typically a few days having pulses of strong winds (gusts) of several hours. In addition, the variability of atmospheric transport and size- and composition-dependent removal processes further contribute to the variability of dust load.

Long-range dust transport exhibits substantial seasonal and interannual variations that are controlled by both the atmospheric circulation and the hydrological cycle over a certain geographical region. In particular, precipitation, which is highly variable from year to year, may affect the atmospheric dust cycle by increasing (or decreasing) soil moisture and vegetation cover and suppressing (or promoting) dust emission. Precipitation events occurring along dust transport

pathways are the most efficient mechanism for removing dust particles from the atmosphere.

Despite the varying frequency and strength of dust events, they do occur each year. Also some seasonal dust emission and transport features are repeatable from year to year for certain dust source regions. For instance, the Sahara is an active dust source the whole year, while the transport of Saharan dust over the North Atlantic Ocean shifts northward from about 5° N during winter to about 20° N during summer. Asian dust outbreaks reach their maximum in spring when Asian dust can be transported as far as North America.

Much of our understanding of the large-scale dust distribution and transport rests largely on satellite remote sensing data. Dust plumes are readily observed in ultraviolet (UV), visible, and infrared (IR) channels of satellite instruments. Satellites provide the primary means of obtaining a global perspective of areal extent of dust plumes as well as some dust properties such as optical depth and effective particle size. [Figure 1](#) presents true color images taken by NASA Moderate Resolution Imaging Spectroradiometer, which illustrate representative dust outbreaks originating in Northern Africa (upper panel) and in Asia (lower panel).

Satellite passive instruments provide a column-integrated view of dust loadings in the atmosphere. Complementary space and ground-based lidar observations allow for characterization of the vertical distribution of dust. Measurements reveal that dust plumes often exhibit a complex multilayered structure. For instance, transport of Saharan dust occurs at higher altitudes in a layer that typically reaches 5–6 km, although one or several layers might be present below it. Concentrations aloft are usually several times greater than those in the marine boundary layer. Dust layers can be intermixed with layers of other aerosols or clouds.

Dust Properties

The mineralogical composition and particle size are the major properties of mineral aerosol governing its impacts upon the Earth's systems. Initially, the composition of soil grains and generation processes determine the particle size distribution of airborne dust particles, their composition, and the degree of particle aggregation. Both the particle size distribution and the composition can be altered during dust transport in the atmosphere, called dust aging.

The main species found in mineral dust are quartz, various clays (e.g., kaolinite, illite, montmorillonite), carbonates, feldspars, chlorites, and iron oxides (e.g., hematite, goethite) among others. These minerals are characterized by very different physical and chemical properties. For instance, different minerals have distinct abilities to adsorb water vapor and other chemically important atmospheric gases. Each mineral has distinct spectral optical constants (or refractive indices), which determine how dust particles scatter and absorb the electromagnetic radiation. Consequently, the properties of dust as a mixture are determined by the relative abundance of various minerals and their aggregates. However, numerous climate and remote sensing studies have considered dust as a single generic species. This is partly due to the complexity of quantitative determination of the mineralogical composition and a lack of data.

The dust composition is thought to vary from source to source. For instance, dust in the Sahelian region is characterized by a high Fe/Al ratio due to the abundance of ferrallitic soils. In contrast, soils in the semiarid regions of Central Asia contain less Fe. The difference in the amount of iron oxides is of special importance because they control primarily the ability of dust particles to absorb sunlight.

Dust particles cover a wide range of sizes from about a few tenths to several hundreds of micrometers. Coarse particles are quickly removed from the atmosphere, and the ones transported over long distances usually have sizes below 20 µm. Several size modes are commonly introduced to characterize the dust particle size distribution. Size modes may have different compositions. In general, clay particles tend to be smaller in size than those made of quartz or carbonates.

Near the source regions, dust concentrations may be as high as a few thousand particles per cubic centimeter during the dust storm. Gravitational settling and rainout are major removal processes affecting dust concentration and particle size distribution during long-range transport. Other important processes, shaping particle spectra and composition, are heterogeneous chemistry on dust particle surfaces, cloud processing of dust particles, and interactions with other atmospheric aerosols. All these processes control the lifetime (or residence time) of dust particles in the atmosphere, which is up to 2 weeks on average. Despite the relatively short lifespan, dust can be carried to great distances up to a few thousand kilometers, affecting large geographical regions. Frequently, Saharan dust plumes reach the Caribbean, the Gulf of Mexico, and the southeast coast of the United States, while Asian dust can be transported as far as the west coast of the United States. Although it is believed that long-range transported dust particles are mainly in the size range from about 0.05 to 20 µm, coarse particles of about 100 µm diameter have been measured at distances of several thousand kilometers from dust sources.

Dust particles exhibit a large variety of shapes, often occurring as irregular (nonspherical) minerals or aggregates of several minerals. To date, there is no generally accepted classification of dust particle morphologies and data are limited. As a necessity, a simplified assumption that dust particles have spherical or spheroidal shapes is often made.

The concentration, composition, size distribution, and morphology of dust particles determine their optical properties. Dust particles can scatter and absorb electromagnetic radiation in a wide range of wavelengths from UV to IR. The optical characteristics needed for radiative effect assessments are the optical depth, single scattering albedo, and scattering phase function. These characteristics are also functions of the location and time, because of varying properties of transported dust.

The optical depth of dust plumes is largest near the dust source and it decreases farther from the source being controlled largely by the dust concentrations. Over the oceans, the highest optical depths occur in regions influenced by dust transport. At visible wavelengths, optical depths as high as 10 have been measured during dust storms. It has a weak dependence on the wavelengths in the visible region, but various spectral features occur in the thermal IR. Observations of the dust optical depth in the thermal IR window region suggest that it is about 2–10 times smaller than in the visible region.

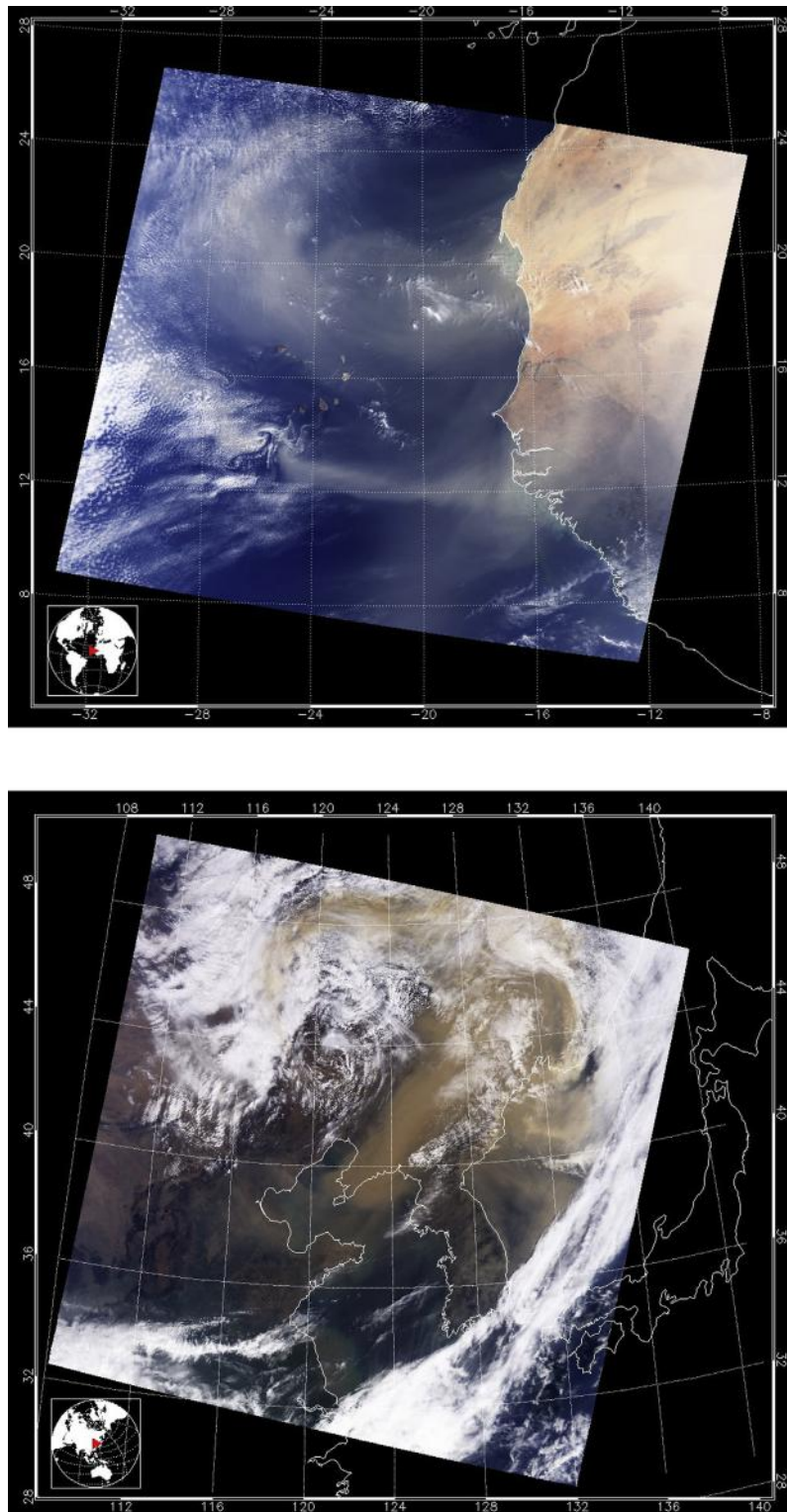


Figure 1 Representative dust outbreaks originating in Northern Africa (upper panel) and in Asia (lower panel). Images taken by the NASA Resolution Imaging Spectroradiometer.

The single scattering albedo, which is defined as the ratio of scattering and extinction coefficients, does not depend on particle concentration but rather on the particle composition, state of mixing, and sizes. It is a key optical characteristic that governs the heating or cooling impact of dust aerosols upon the Earth system. The single scattering albedo of dust is in the range from about 0.75 to 0.99 in the solar and is characterized by strong wavelength dependence, increasing from UV to near-IR. The single scattering albedo may vary during the transport as dust ages. If dust particles are coated by other aerosol species, they may have drastically different properties from those that are evident at the dust source.

Although the scattering phase function of dust particles is crucial for remote sensing applications, there are only a few actual measurements. Therefore, this and other optical characteristics of dust are often computed by applying Mie theory for a certain size distribution and refractive index. More advanced numerical codes allow optical computation of nonspherical shapes. The T-matrix method is the most common approach for computing the optical characteristics of spheroidal shapes, while the discrete dipole approximation models for particles of any morphology.

Dust Effects

Mineral dust particles, by virtue of their chemical, physical, and optical properties, cause various impacts upon the atmosphere, environment, human, and overall Earth's system. [Table 1](#) lists some major impacts of dust and their importance. Many important impacts involve interactions of mineral aerosol with solar and IR radiation. Dust-induced perturbations to the radiative energy distribution have a significant impact on the surface energy balance and the planetary (global) energy balance, with implications for the regional and global climate. In contrast to greenhouse gases, dust may cause either a positive or negative total (solar plus IR) direct radiative forcing at the top of the atmosphere, depending on dust optical properties, atmospheric conditions, and surface reflectance. The radiative forcing and

heating/cooling rates of dust may be of different signs in the solar and IR regions. Dust reduces the solar irradiance reaching the surface and increases longwave radiation by a lesser amount, so the net surface radiative balance decreases.

The direct radiative forcing of dust has a complex geographical distribution because the dust sources and sinks are not uniformly distributed and the lifetime of mineral particles is short in the atmosphere. Thus, the presence of dust may enhance greenhouse gas warming in some regions and oppose it in others. In addition, dust may affect Earth's radiation balance indirectly by altering the properties, amount, and distribution of clouds. These are termed as indirect radiative effects. Dust particles can serve as cloud condensation nuclei (CCN) or ice nuclei (IN), affecting water, ice, and mixed-phase clouds. The ability of dust particles to serve as CCN or IN strongly depends on the mineralogical composition and particle size. Additional influence of dust on the energy and hydrological cycle proceeds through dust deposition on snow-covered surfaces, resulting in decreased surface reflectivity while promoting snowmelt.

Dust also affects the radiatively important gases and other aerosol species through several physicochemical mechanisms. By altering UV radiation, dust influences photolysis rates and hence the photochemical formation of such a radiatively important species like ozone. Dust particles have a high surface area per unit mass, which makes them ideal sites for heterogeneous chemical reactions. Some atmospheric gases (e.g., nitric acid) can directly condense on dust particles. These mechanisms provide a plausible explanation of the elevated sulfate and nitrate levels associated with dust particles that have been observed over both the Pacific Ocean and Atlantic Ocean. Overall, the impact of dust on atmospheric chemistry depends on dust particle composition and size, concentrations of gaseous species, as well as ambient conditions (e.g., temperature and relative humidity).

Dust has a considerable impact on the chemical and biological processes occurring in land and ocean ecosystems. Light is a vital factor governing plant photosynthetic activities. By altering the photosynthetically active radiation (400–700 nm), dust affects the terrestrial vegetation functioning, with implications for the carbon cycle. Deposited dust may supply

Table 1 Major impacts caused by dust aerosol and their importance

<i>Impact</i>	<i>Importance</i>
Direct radiative impacts	
Causes radiative forcing at the top of the atmosphere	Affects energy balance of the Earth's climate system
Causes radiative forcing at the surface	Affects surface temperature and surface–air exchange processes
Causes radiative heating/cooling within a dust layer	Affects temperature profile and atmospheric dynamics
Affects visibility	Decreases visibility, posing a traffic hazard
Impacts on clouds	
Serves as ice nuclei	Affects the properties and amount of ice clouds
Serves as cloud condensation nuclei	Affects the properties and amount of water clouds
Impacts on atmospheric composition and chemistry	
Alters actinic flux	Affects photolysis rates and photochemistry
Provides particle surfaces for heterogeneous chemical reactions	Alters the abundance of radiatively important atmospheric gases and aerosols
Impacts on land and ocean ecosystems	
Alters photosynthetic active radiation	Affects plant photosynthesis and growth
Alters surface energy balance	
Transports nutrients (iron and phosphorous)	Affects biogeochemical processes

nutrients to ecosystems. Soluble iron supplied by dust particles deposited in the oceans is essential for the growth of phytoplankton. The transport of micronutrients (such as phosphate and potassium) associated with dust has been recognized for its importance to terrestrial ecosystems.

The degradation of visibility and health problems caused by dust make it an important air quality issue. Dust particles are toxic or can act as the transport mechanism for toxic species, which can be inhaled and pose a health threat.

Although the importance of dust in the Earth's system has been well recognized, quantitative assessment of the dust impacts still poses significant challenges. Improved understanding of the dust properties and governing processes will be needed to decrease the current uncertainties. This has been an active field of research, especially in the past few decades.

See also: **Aerosols:** Aerosol Physics and Chemistry; Climatology of Stratospheric Aerosols; Observations and

Measurements; Role in Radiative Transfer. **Chemistry of the Atmosphere:** Observations for Chemistry (*In Situ*); Particles; Principles of Chemical Change. **Hydrology, Floods and Droughts:** Soil Moisture. **Lidar:** Atmospheric Sounding Introduction. **Optics, Atmospheric:** Optical Remote Sensing Instruments. **Radiation Transfer in the Atmosphere:** Absorption and Thermal Emission; Radiation, Solar. **Satellites and Satellite Remote Sensing:** Aerosol Measurements. **Tropospheric Chemistry and Composition:** Aerosols/Particles.

Further Reading

- Formenti, P., et al., 2010. Recent progress in understanding physical and chemical properties of mineral dust. *Atmospheric Chemistry and Physics Discussions* 10, 31187–31251. <http://dx.doi.org/10.5194/acpd-10-31187-2010>.
- Ravi, S., et al., 2011. Aeolian processes and the biosphere. *Reviews of Geophysics* 49, RG3001. <http://dx.doi.org/10.1029/2010RG000328>.

Observations and Measurements

PH McMurry, University of Minnesota, Minneapolis, MN, USA

© 2003 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 1, pp 20–34, © 2003, Elsevier Ltd.

Introduction

Atmospheric aerosols consist of a suspension of solid or liquid particles in air. These particles range from molecular clusters less than one nanometer in diameter to pollens, wind-blown soil dust and sea salt of 10 μm or larger. Figure 1 shows an idealized atmospheric aerosol size distribution with the characteristic modes in which particles tend to be found. Homogeneous nucleation leads to the clustering of molecules that grow to particles several nanometers in diameter at the small end of the spectrum. Aitken or nuclei mode particles in the 10 to 100 nm diameter range often contain 'primary' particles that are directly emitted into the atmosphere by combustion processes. Accumulation mode particles contain most of the submicron mass. Most of this mass is typically 'secondary' in origin (i.e., formed in the atmosphere by chemical transformations). Sulfates, nitrates, and organics typically comprise most of the accumulation mode mass, and this mass can accumulate in different submodes, depending on the chemical mechanisms by which they are formed. Coarse-mode particles tend to be produced by mechanical processes.

(Figure 2(a)–(c)) show photographs of atmospheric particles obtained by electron microscopy. The chain-agglomerate soot particle in (Figure 2(a)) is typical of particles emitted by diesel engines or other combustion sources. (Figure 2(b) and (c)) illustrate particles of mixed composition. The salt in (Figure 2(b)) was probably emitted directly into the atmosphere as sea spray, while the sulfate crystals attached to it may

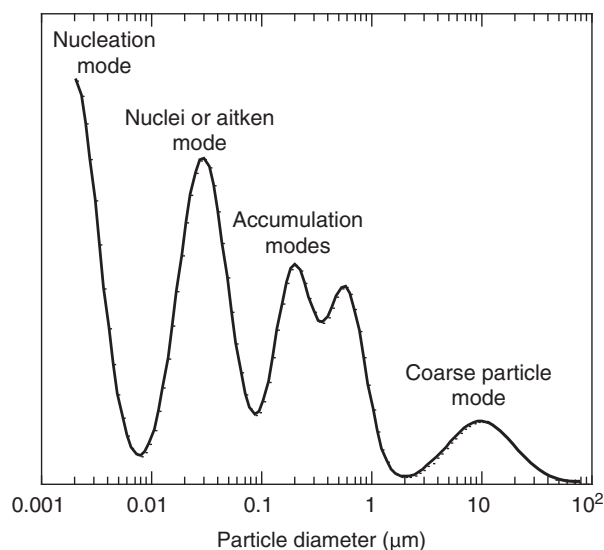


Figure 1 Idealized atmospheric aerosol size distribution. Particles tend to be found in modes that reflect the different processes by which they are formed in the atmosphere by chemical transformations or are emitted directly into the atmosphere from sources.

have been produced in the atmosphere by chemical transformations. The rutile in (Figure 2(c)) probably originated as wind-blown dust and may have accumulated the sulfate coating as a result of atmospheric processing. The variabilities in shape and composition illustrated by the photographs in (Figure 2(a)–(c)) introduce ambiguities in aerosol measurements since the responses of particle measuring instruments depend on such properties. In addition, particles often contain volatile or semivolatile species such as water, nitric acid, ammonia, and various organic compounds that are also present in the gas phase. This adds another challenge to obtaining accurate measurements, since such species undergo continuous exchange between the condensed and gas phases, and the gas/particle distribution can be perturbed during sampling.

This article summarizes the most important methods for measuring atmospheric aerosols. It first discusses aerosol sampling and chemical analysis of aerosol, then concludes with methods used to measure physical properties of particles such as particle size and number concentration. Throughout, discussion focuses on *in situ* particle measurements; a great deal of valuable information on global distributions of aerosols has been obtained with satellites, but such measurements are not discussed here (see **Satellites and Satellite Remote Sensing: Aerosol Measurements**).

Sampling

Accurate aerosol measurements require the collection of a representative sample. Obtaining representative samples of particles smaller than 1 μm is usually not too difficult. Under typical sampling conditions particles in this size range travel with the gas. Difficulties may be encountered with particles smaller than about 10 nm in diameter since, because of their high diffusivities, they can be lost on the walls of sampling tubes during transport to the detector. It is usually possible to design sampling systems that minimize this effect.

Supermicron particles are difficult to sample. Such particles have so much inertia that they tend not to follow the gas flow as it enters the sampler. Figure 3 shows trajectories of large particles as they are drawn into an aerosol sampler. When the wind speed exceeds the velocity of the air flow into the sampling tube (subisokinetic sampling), the sampled concentration exceeds the true ambient concentration (Figure 3(a)), while the reverse is true when the wind speed is less than the sampling speed (superisokinetic sampling; Figure 3(b)). An ideal *isokinetic* sampler is aligned with the wind and draws in air at a speed equal to the wind speed. In practice, wind speeds and directions vary, so achieving this ideal with samplers at a fixed location is impractical.

It might appear that an aircraft would provide an ideal platform for sampling coarse particles. Isokinetic sampling

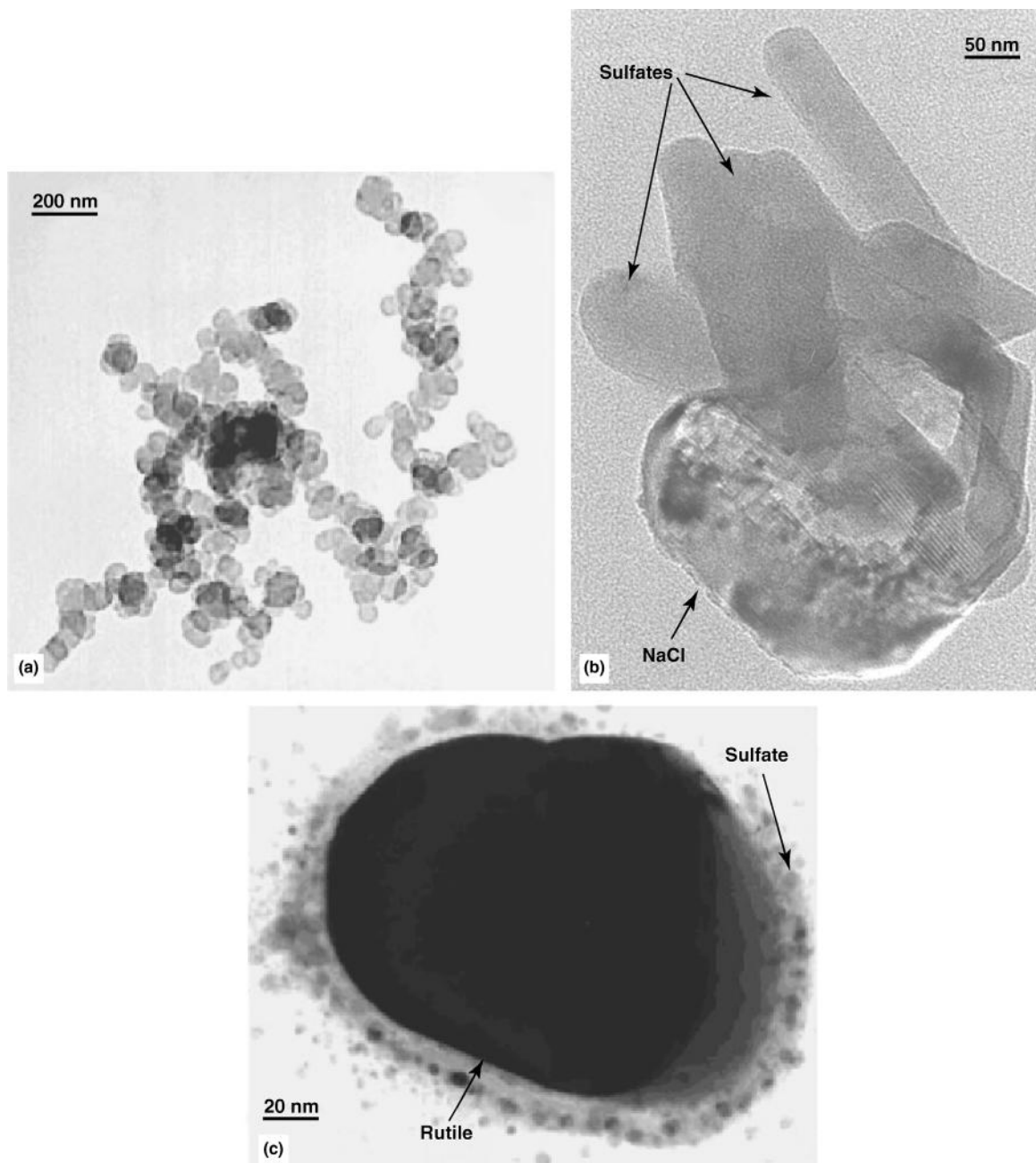


Figure 2 Photographs of atmospheric particles obtained by transmission electron microscopy. (a) Chain agglomerate soot particle. (b) Sulfate attached to a sodium chloride particle. (c) Mineral dust particle coated with sulfate. Reproduced with permission (a) and (c) from Buseck PR, Jacob DJ, Posfai M, Li J, and Anderson JR (2000) Minerals in the air: an environmental perspective. *International Geology Review* 42, 577–593, © V.H. Winston Son, Inc. 360 South Ocean Boulevard, Palm Beach, FL 33480, USA. All rights reserved. (b) From Buseck PR and Posfai M (1999) Airborne minerals and related aerosol particles: effects on climate and the environment. *Proceedings of the National Academy of Sciences of the United States of America* 96: 3372–3379. © 1999, National Academy of Sciences, USA.

would be achieved automatically if the sampler pointed straight ahead and sampled at a velocity equal to the aircraft velocity. Unfortunately, this is not the case. Even if the probe sampling velocity equals the aircraft speed, variations in the aircraft orientation lead to misalignment between the probe and the sampled airstream, and the high sampling velocities required lead to turbulent deposition of particles within the

sampling probe. Thus, it can be difficult to transport particles through the probe to the measuring instrumentation. The design of aircraft sampling inlets that deliver a known fraction of larger particles is a current area of research.

Omnidirectional sampling inlets are typically used for fixed-point sampling. Such inlets draw aerosol from all directions through a horizontal annular opening. Sampling efficiencies

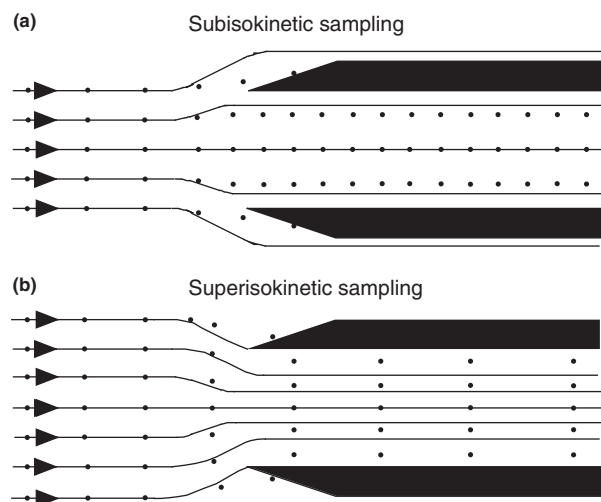


Figure 3 Trajectories of large particles as they are drawn into an aerosol sampling probe. (a) Subisokinetic sampling occurs when the velocity of the sampled aerosol at the probe entrance is less than the velocity of the wind relative to the probe. In this case the concentration of 'large' particles in the sampled aerosol is higher than the ambient concentration. (b) For superisokinetic sampling, the sampled concentration is less than the true concentration.

for omnidirectional inlets are independent of wind speed, and can be nearly 100% for particles up to 10 μm .

In summary, obtaining representative aerosol samples for particles smaller than a few micrometers in diameter is usually straightforward. Obtaining representative samples for particles larger than this is difficult since, owing to their inertia, large particles tend not to follow air as it flows into a sampler.

Sample Collection and Analysis: Filters and Impactors

Filters are the device most commonly used for collecting particles for analysis. Samples are collected by removing particles from a known volume of air, and are analyzed for particulate mass or species concentrations. Electron micrographs of several types of filter media commonly used to sample atmospheric particles are shown in (Figure 4(a)–(c)). All of these media contain passages through which the gas flows and solid substrates on which particles are collected.

Filters collect particles by three different mechanisms, as illustrated in (Figure 5). Very small particles (diameter less than about 0.1 μm) are collected by diffusion, and because diffusivities increase with decreasing size, particle collection efficiencies increase with decreasing size. Large particles (diameter $> \sim 1 \mu\text{m}$) are collected when, due to their inertia, they are unable to follow the gas as it flows around a solid surface (e.g., a fiber) and impact on the fiber. Because inertia increases with increasing particle size, filter collection efficiencies increase with size in this range. Intermediate-sized particles are collected by interception when the gas flow carries them to within one particle radius of the filter medium. Note that 'sieving' is not a significant particle collection mechanism for

air particle filters; as can be seen in (Figure 4), the dimensions of the air passages through the filter are typically much larger than the sizes of the collected particles. All filters have a 'most penetrating size' that falls typically in the 0.05 to 0.3 μm diameter range, where particle collection by diffusion, interception, and impaction reaches a minimum. For filters typically used for air filtration, however, collection efficiencies even at this size usually exceed 99%. Therefore, it is generally reasonable to assume that all particles delivered to the filter are collected. Filters are typically operated downstream of an inlet that removes all particles larger than a specified size.

Substantial effort has been made to develop diffusion denuder filter samplers for measuring the gas and particle phase concentrations of semivolatile species. The design of such samplers for inorganic species including ammonium nitrate is well developed, and work is underway to extend this methodology to various organic compounds. Sampling such compounds presents a challenge because they can volatilize from the filter during sampling, thereby leading to an underestimate of the particle-borne concentration. A schematic of a diffusion denuder sampler is shown in Figure 6. The aerosol is first drawn through an annular denuder, the walls of which are coated with a chemical that reacts readily with the gas to be removed. Basic compounds such as sodium carbonate are often used to remove acidic gases such as sulfur dioxide and nitric acid, and acidic compounds such as citric acid are used for ammonia. Owing to their high diffusivities, gases diffuse to the coated walls and are removed with nearly 100% efficiency. Particle diffusivities are orders of magnitude lower than those for gases, so particles pass through the denuders to a filter where they are collected with nearly 100% efficiency. Volatilization from the collected particles occurs during sampling, and a suitable collector is located downstream of the filter to collect these volatilized compounds. For example, nylon filters are known to collect nitric acid vapor with nearly 100% efficiency, and porous polyurethane foam is used to collect certain organic vapors. The particle phase concentration of the semivolatile compound is found by adding the loadings collected by the filter and the downstream adsorber. Several different approaches are used to measure the gas phase concentration. In some cases the collected gas can be extracted from the diffusion denuder and analyzed directly. Alternatively, two samplers can be operated in parallel, one with a diffusion denuder and one without. The sampler without the diffusion denuder collects the total gas-plus-particle concentration. The gas phase is then determined by difference.

Cascade impactors are used to collect size-segregated samples for subsequent analysis. A schematic diagram of a cascade impactor is shown in Figure 7. The aerosol is first accelerated through a large nozzle towards a collection substrate that deflects the air and collects the largest particles by inertia. The aerosol is then transported to a stage with a slightly smaller nozzle that collects smaller particles due to the higher speed of the gas through the nozzle. Impactors can provide a very sharp size separation down to sizes as small as 10 nm. Measurements of solid particles can be adversely affected by bounce, which can be reduced by greasing the collection substrates. A limitation of impactors is that no strategy has been developed to measure evaporative losses of semivolatile compounds from the individual impactor stages.

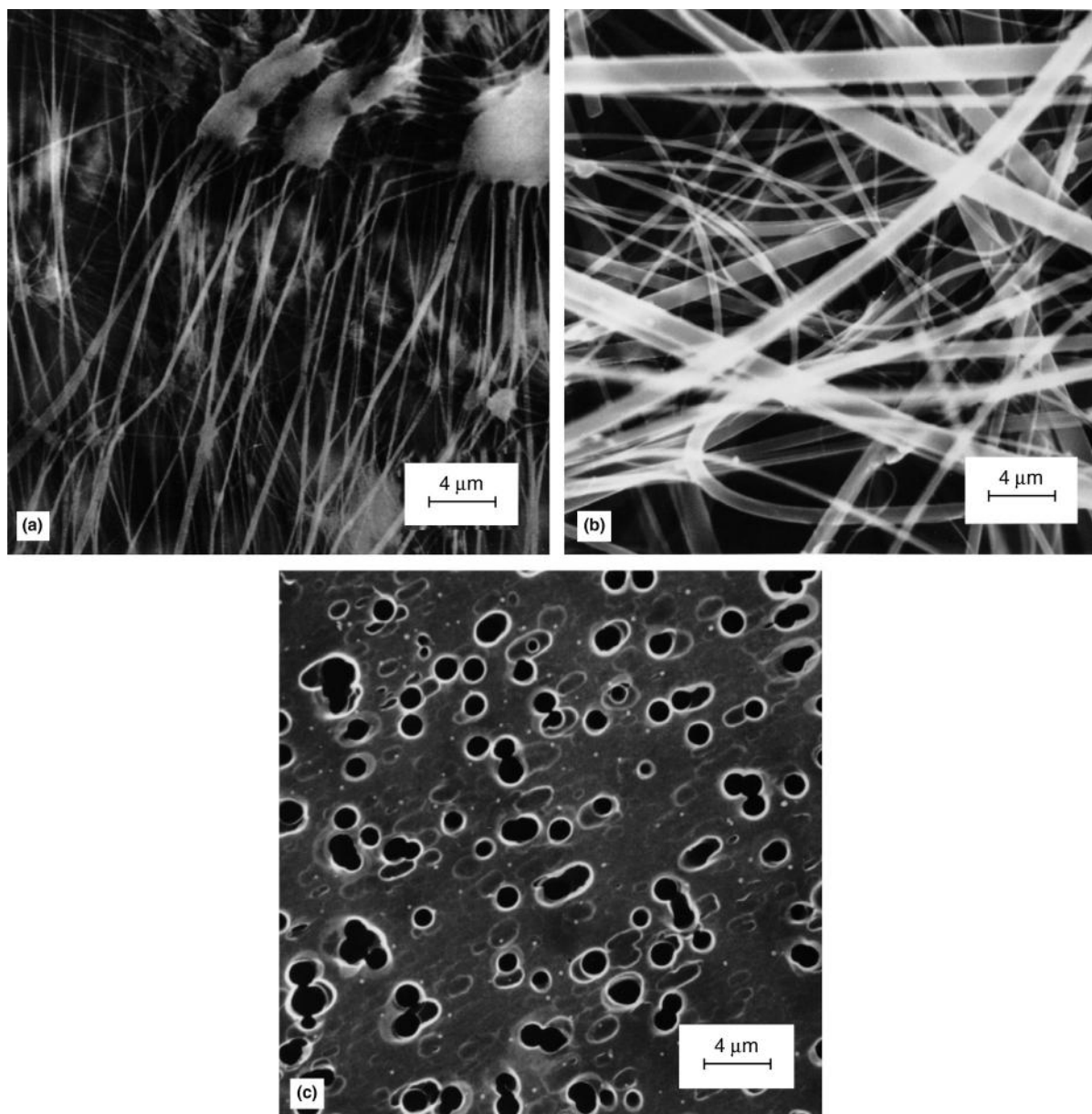


Figure 4 Electron micrographs of several types of filter media commonly used to sample atmospheric particles. The filter media illustrated in these figures include (a) a 3.0 μm Teflon membrane filter, (b) a glass fiber filter, and (c) a 1.0 μm polycarbonate membrane nuclepore filter. Courtesy of Professor Benjamin Y. H. Liu, Department of Mechanical Engineering, University of Minnesota, Minneapolis, MN.

Virtual impactors replace the collection substrate with a receiving tube that draws a small fraction of the flow and into which all particles larger than the cut size are projected. Virtual impactors avoid bounce and use conventional filter substrates, but have the disadvantages that they cannot be designed with cut points that are as sharp or as small as can be achieved with conventional impactors. Cascade impactors segregate particles according to aerodynamic diameter, which depends on shape, physical size, and particle density. Samples collected with filters or impactors are typically analyzed to determine particulate mass or species concentrations. [Table 1](#) provides a summary of typical minimum detection limits that can be achieved by commonly used analytical methods.

Real-Time, *In-Situ* Analysis of Aerosol Chemical Properties

Filtration with subsequent analysis in the laboratory is by far the most commonly used approach for determining the composition of atmospheric aerosols. Considerable progress has been made in the last few years, however, on techniques that provide real-time information on aerosol composition. For example, particles can be collected directly in liquid water and analyzed by ion chromatography to determine concentrations of major aqueous ions, including sulfate, nitrate, ammonium, chloride, etc. When employing denuders these methods can also be used to measure gas phase concentrations

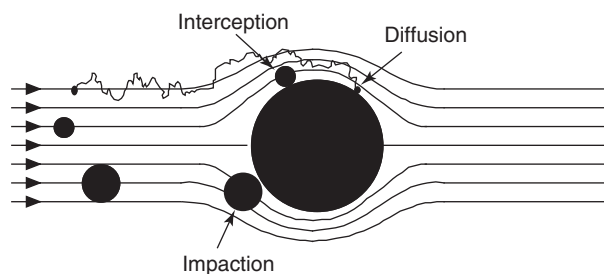


Figure 5 Filters collect particles by three different mechanisms. Diffusion is primarily responsible for the collection of particles smaller than about $0.1\ \mu\text{m}$. Inertial impaction, which causes particles to cross air streamlines and deposit on the filter substrate, is responsible for the collection of particles larger than about $1\ \mu\text{m}$. Interception, which occurs when the air streamline brings a particle to within one particle radius of a filter substrate, is important for particles in the intermediate size range.

of semivolatile compounds. Other approaches involve collecting particles by impaction on a substrate from which they can be volatilized by rapid heating and analyzed by a conventional gas analyzer. These semicontinuous samplers provide species concentrations with a time resolution down to 3 min and show that aerosol concentrations often vary significantly over time scales required for filter samples.

A tremendously exciting new development in aerosol measurement involves the use of mass spectrometry to measure the composition of individual particles in real time. Particles are drawn through an inlet into a low-pressure chamber where they are volatilized and ionized by a high-intensity laser pulse. The ions are analyzed by time-of-flight mass spectrometry to determine particle composition. A schematic diagram of one such instrument is shown in (Figure 8). It is not yet clear that instruments of this type will be able to provide quantitative information on species mass concentrations. They do, however, provide rich new insights into the origins of various particle types and also provide insights into chemical transformations that occur on and within particles in the atmosphere. For example, (Figure 9) illustrates the type of information that can be obtained from such measurements. In this study nearly half of the elements in the periodic table were detected in particles sampled from an aircraft between 5 and 19 km altitude.

Measurement of Physical Properties

Instrumentation for measuring aerosol physical properties can be divided into two major categories: those used to measure a single overall property of the aerosol and those used to

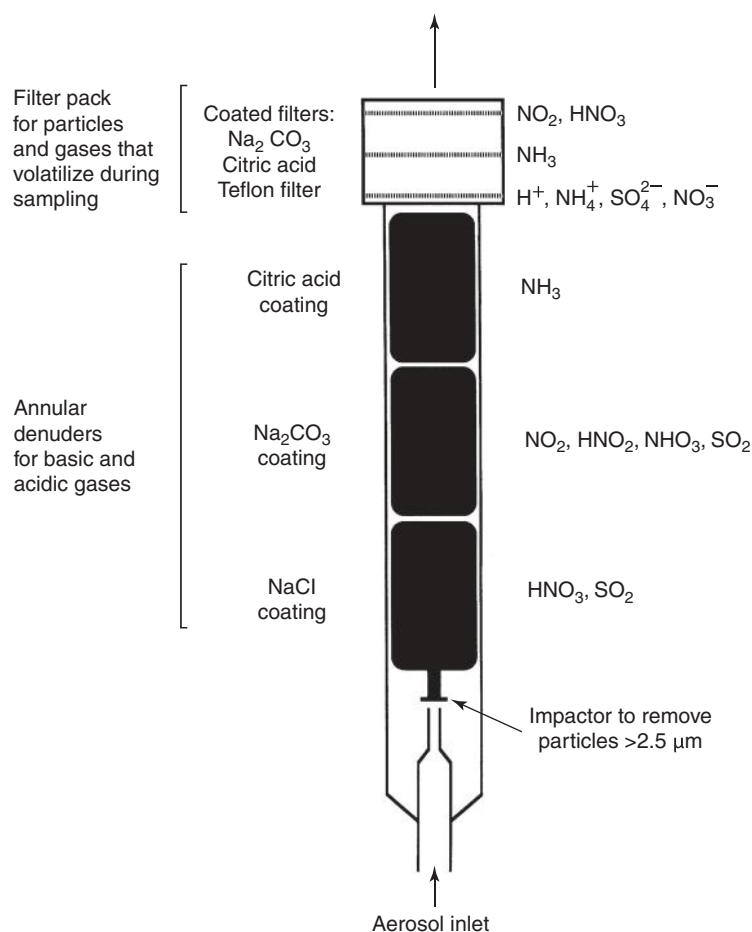


Figure 6 Schematic of a diffusion denuder sampler for collecting the gas and particle phases of semivolatile compounds.

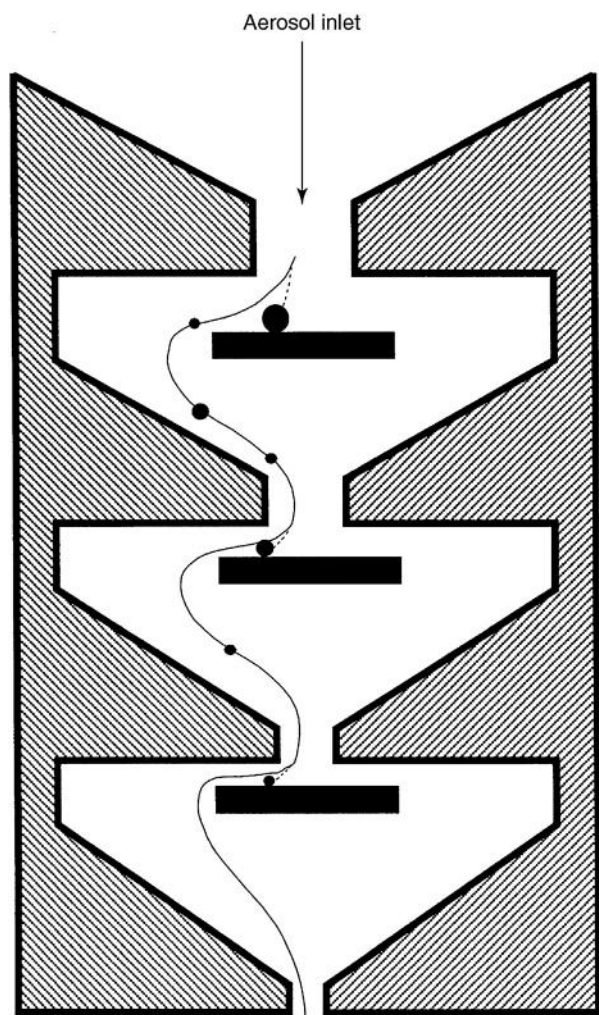


Figure 7 Schematic diagram of a cascade impactor. These instruments separate particles into several different size fractions for subsequent measurement of size-resolved mass or species concentrations.

measure aerosol size distributions. Examples of overall properties (also referred to as integral properties) are the total number concentration of particles of all sizes or the total amount of light scattered by particles of all sizes. Instruments that measure integral properties are often easy to operate and provide data that are straightforward to interpret. A limitation of such measurements is that they cannot be used in a simple way to determine other aerosol properties. In principle, a detailed and accurate measurement of the aerosol size distribution can be used to determine any integral property of the aerosol. Such data, however, are difficult to acquire and analyze and are subject to uncertainties owing to the dependence of instrument response on aerosol properties including shape, density, and refractive index, which are often not known precisely.

Measurements of Integral Properties

Condensation particle counters (CPCs) have been used since John Aitken invented them in the late 1800s to measure total

concentrations of aerosol particles. (CPCs are also referred to as condensation nuclei counters (CNCs) and Aitken nuclei counters.) They function by exposing particles to a supersaturated vapor which causes them to grow by condensation to a size greater than $1\text{ }\mu\text{m}$, which can easily be detected by light scattering. Concentrations are then determined either by counting the pulses of light scattered by individual droplets as they flow through the scattering volume of an optical detector (**Figure 10**) or from the light attenuation or scattering produced by the cloud of droplets following condensation (**Figure 11**). An advantage of the single particle counting instruments is that concentrations are determined directly from the counting rate and the aerosol flow rate through the detector; no empirical calibration is required. The optical attenuation or scattering methods require an empirical calibration to determine the relationship between attenuation and concentration. The largest particle size that can be detected by a CPC is determined by transport into the condenser: because of their inertia, particles bigger than a few micrometers in diameter tend to be lost as they flow into the instrument. CPCs have been designed that detect particles as small as 3 nm , although a minimum detectable size of 5 to 10 nm is more typical. Recent work has shown that gaseous ions (i.e., molecules or molecular clusters that carry an elementary charge) smaller than 1 nm can be detected with a CPC.

Another aerosol integral property of considerable interest is the total amount of light scattered by particles. Light scattering is responsible for the hazy appearance of air in polluted regions (**Figure 12**) and is the most sensitive indicator of air pollution. Gases also scatter light by a process referred to as Rayleigh scattering. Rayleigh scattering is responsible for the blue color of the sky, and in a particle-free atmosphere would limit visibilities to about 300 km at sea level. On a per unit mass basis, however, particles are orders of magnitude more effective than gases at scattering light, and light scattering efficiency (the amount of light scattered per unit mass) reaches a maximum for particles between 0.1 and $1\text{ }\mu\text{m}$ where most secondary pollutant particles accumulate (**Figure 13**). This juxtaposition of light scattering efficiencies with the peak in the accumulation mode mass distribution (see **Figure 1**) causes light scattering to be the most sensitive indicator of polluted air. Integrating nephelometers are used to measure the total amount of light scattered by aerosol particles. The response of integrating nephelometers depends on the frequency distribution of the illuminating radiation, on the frequency response of the optical detector, and on the solid angle over which light scattering is collected.

Current air quality standards for particulate matter are based mostly on the total mass concentration of particles smaller than a specified size. For example, since 1987 the 'PM₁₀' National Ambient Air Quality Standard for particulate matter in the United States has stipulated that mass concentrations of particles smaller than $10\text{ }\mu\text{m}$ aerodynamic diameter cannot exceed $75\text{ }\mu\text{g m}^{-3}$ on an annual basis or $260\text{ }\mu\text{g m}^{-3}$ for 24 hours. The usual approach for measuring such mass concentrations is to determine the change in the mass of a filter that has removed particles from a known volume of air. Sampling times ranging from hours to a day or more are typically required for such gravimetric analyses, so such

Table 1 Typical minimum detection limits (three times standard deviation of blank) for selected elements and species measured by various techniques (ng m^{-3})

Species	Electronic balance	INAA	XRF	PIXE	Flame AAS	Graphite furnace AAS	ICP	AC	IC	TOR
Mass	900									
Cl^-									50	
NH_4^+								50		
NO_3^-									50	
SO_4^{2-}									50	
EC										100
OC										100
Al		24	5	12	30	0.01	20			
Ca		94	2	4	1	0.05	0.04			
Cd		4	6		1	0.003	0.4			
Cl		5	5	8						
Fe		4	0.7	2	4	0.02	0.5			
Hg			1		500	21	26			
K		24	3	5	2	0.02				
Mg		300		20	0.3	0.004	0.02			
Pb			1	3	10	0.05	10			
S		6000	2	8			10			

Abbreviations: INAA – instrumental neutron activation analysis; XRF – X-ray fluorescence analysis; PIXE – proton induced X-ray analysis; AAS – atomic absorption spectrophotometry; ICP – inductively coupled plasma emission spectroscopy; AC – automated colorimetry; IC – ion chromatography; TOR – thermal/optical reflectance analysis. Reproduced with permission from Chow JC (1995) Measurement methods to determine compliance with ambient air quality standards for suspended particles. Journal of the Air and Waste Management Association 45: 320–382.

measurements provide no information on variabilities in mass concentrations that occur during the sampling period. Furthermore, in regions with significant concentrations of semivolatile compounds such as ammonium nitrate and certain particulate organic compounds, filter measurements can be compromised by volatilization losses. For these reasons aerosol scientists have attempted to develop alternative

measurement techniques. These approaches include the measurement of the attenuation of 0.01 to 0.1 μeV beta particles through particulate deposits collected on filters, the change in the natural frequency of a piezoelectric crystal on which particles have been collected, the change in pressure drop across a filter on which particles are being collected, and the change in the natural frequency of a harmonic oscillating

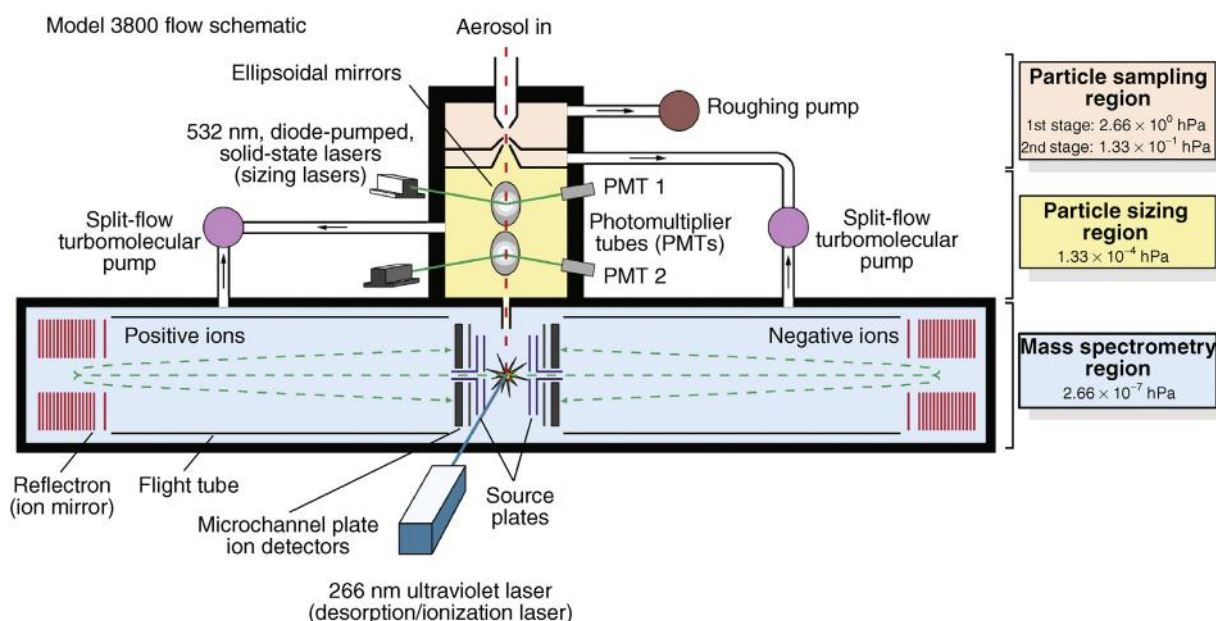


Figure 8 Schematic diagram of a mass spectrometer for determining the composition of individual aerosol particles. Particle size is determined by measuring particle velocity in the particle sizing region of the instrument. Size depends on velocity because large particles, owing to their inertia, are not accelerated as rapidly as small particles as they flow from the aerosol inlet into the low-pressure regions of the instrument. Particle composition is determined by using time-of-flight mass spectrometry to measure the mass spectra of ions produced when the particles are volatilized by a high intensity pulsed laser. Courtesy of TSI, Inc.

MostManyCommonSomeRareMaybe																		He
H																	Ne	
Li	Be											B	C	N	O	F	Ne	
Na	Mg											Al	Si	P	S	Cl	Ar	
K	Ca	Sc	Ti	V	Cr	Mn	Fe	Co	Ni	Cu	Zn	Ga	Ge	As	Se	Br	Kr	
Rb	Sr	Y	Zr	Nb	Mo	Tc	Ru	Rh	Pd	Ag	Cd	In	Sn	Sb	Te	I	Xe	
Cs	Ba	*	Hf	Ta	W	Re	Os	Ir	Pt	Au	Hg	Tl	Pb	Bi	Po	At	Rn	
Fr	Ra	*																
La	Ce	Pr	Nd	Pm	Sm	Eu	Gd	Tb	Dy	Ho	Er	Tm	Yb	Lu				
Ac	Th	Pa	U															

Figure 9 Elements detected by individual particle mass spectrometry at altitudes between 5 and 19 km. Almost every particle was found to contain some H, C, O, N, and S. It was concluded that in this study about half of the particles measured in the stratosphere were of meteoric origin. These particles contained Na, Mg, Al, K, Ca, Cr, Fe, and Ni. Reproduced with permission from Murphy DM, Thomson DS, and Mahoney MJ (1998) *In situ* measurements of organics, meteoritic, and other elements in aerosols at 5 to 19 kilometers. *Science* 282: 1664–1669. © 1998, American Association for the Advancement of Science.

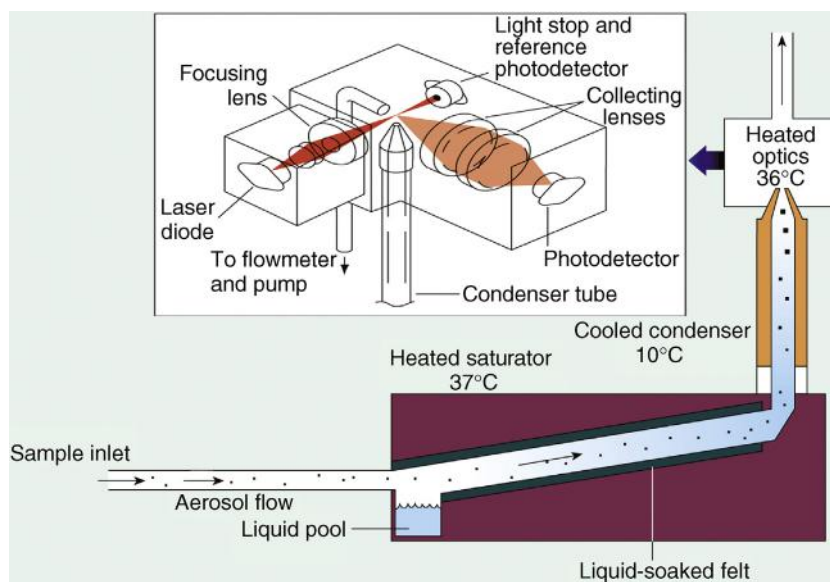


Figure 10 Continuous flow condensation particle counter (CPC). The aerosol is saturated with a vapor (butanol is commonly used) before it flows through a cooled cylindrical tube where the vapor becomes supersaturated and condenses on particles, causing them to grow to a size that is easily detected by light scattering. Courtesy of TSI, Inc.

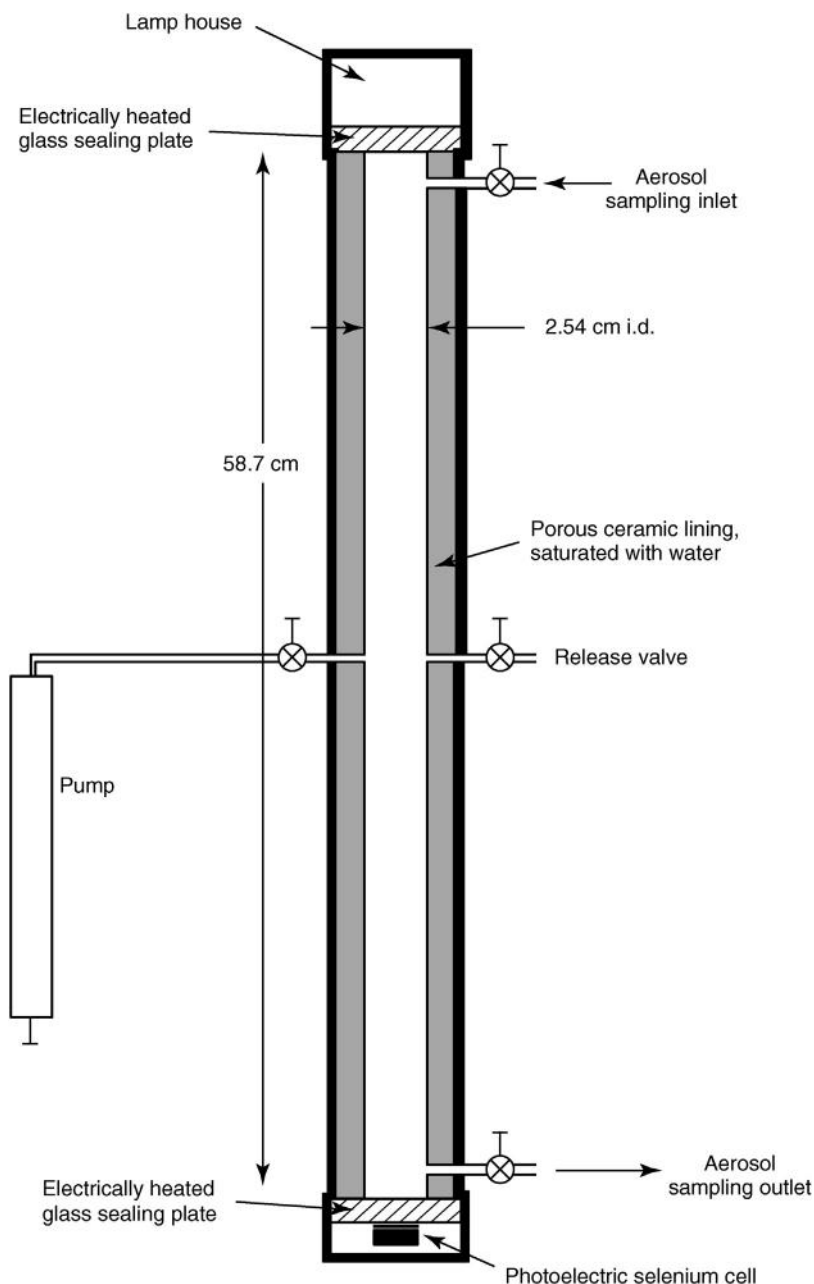


Figure 11 Schematic diagram of the Model 1957 Pollak expansion-type condensation nucleus counter, a classic instrument of its type. An aerosol sample is drawn into the cylindrical chamber and brought with the pump to a pressure 1.21 times local atmospheric pressure. After allowing about 45 s for the aerosol to become saturated with water vapor from the wetted walls, the release valve is opened, permitting the aerosol to expand. The cooling caused by this adiabatic expansion leads to supersaturation of the water vapor and condensational growth of particles. The particle concentration is determined from the attenuation of light through the instrument caused by the aerosol cloud.

element on which a filter for collecting particles is mounted (Figure 14). Each of these techniques has limitations, but they all provide valuable real-time information that cannot be obtained with traditional gravimetric methods.

Measurements of Size Distributions

Instrumentation to measure aerosol size distributions down to 3 nm diameter has advanced significantly in the past two

decades. The most common sizing methods involve classification according to electrical mobility, the measurement of the amount of light scattered by individual particles, and the measurement of the terminal velocity to which particles are accelerated as they flow through a nozzle.

As a rule of thumb, a given instrument can measure the size distribution of particles ranging over about a factor of ten in particle diameter. One reason for this limitation is that concentrations of particles vary so significantly with size. This



Figure 12 View to the north from Keck Laboratory at the California Institute of Technology on a clear and a smoggy day. Photographs by the author.

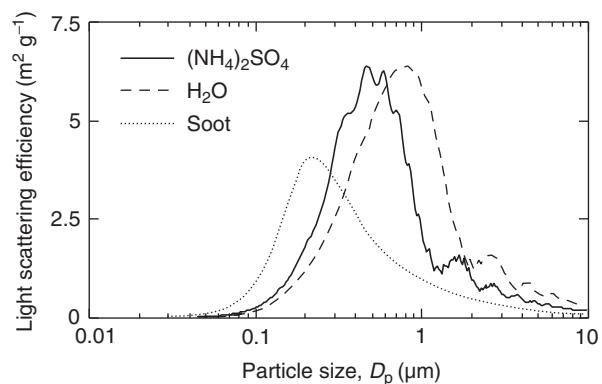


Figure 13 Light scattering efficiency (i.e., the amount of light scattered per particle mass) versus particle size for several important atmospheric aerosol species. Courtesy of Dr William D. Dick.

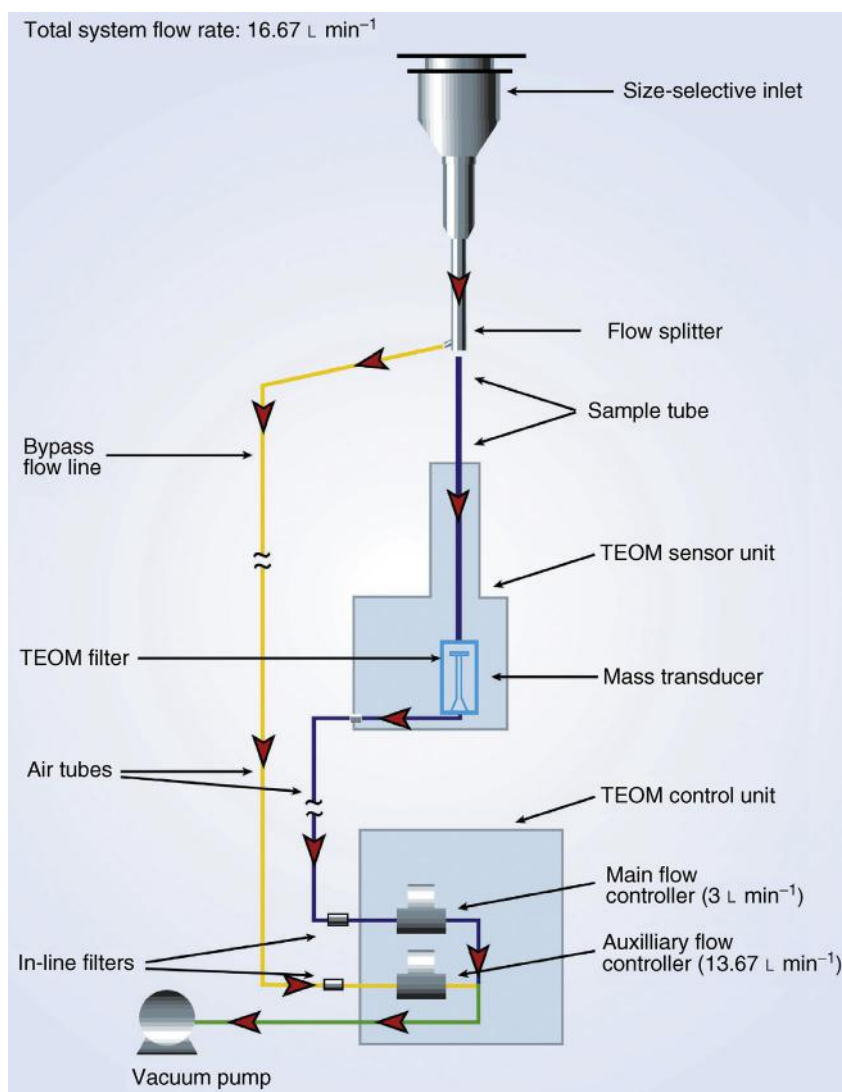


Figure 14 Tapered Element Oscillating Microbalance (TEOM) for continuous measurements of particle mass concentrations. Aerosol is drawn through the filter, which increases in mass as particles are collected. As the filter mass increases, the natural oscillating frequency of the tapered element changes. The mass of particles on the filter is inferred from measurements of this frequency. Courtesy of Rupprecht Patashnick Co., Inc.

Table 2 Nominal particle counting rates in different size ranges for an instrument sampling at 1 liter per minute for a typical urban Los Angeles aerosol

Size range	Nominal sampling rate (particles s^{-1})	Sampling time required for 1% accuracy (s)
0.05–0.1 μm	10^6	0.01
0.1–0.2 μm	10^5	0.1
0.2–0.5 μm	10^4	1.0
0.5–1.0 μm	10^3	10
1.0–2.0 μm	10^2	100
2.0–5.0 μm	10^1	1000

can best be illustrated by considering, as an example, the typical urban Los Angeles aerosol. (Table 2) shows the rates at which particles in several size ranges would have been drawn into an instrument that samples air at 1 liter per minute. Because small particles are so much more abundant than large ones, they are sampled at a much higher rate. Therefore, the time required to collect a statistically significant sample increases sharply with size. The measurement protocol might call for measurements with a time resolution of, say, one minute. This can be achieved best by measuring large particles with high-flow instruments and small particles with low-flow instruments so that statistically significant samples can be achieved for all sizes in comparable time periods. Furthermore, because detection sensitivity depends on size, it is unlikely that a given detection scheme can be used over the entire size range of interest.

Electrical mobility analyzers are used to measure size distributions of particles ranging from about 3 nm to 0.5 μm .

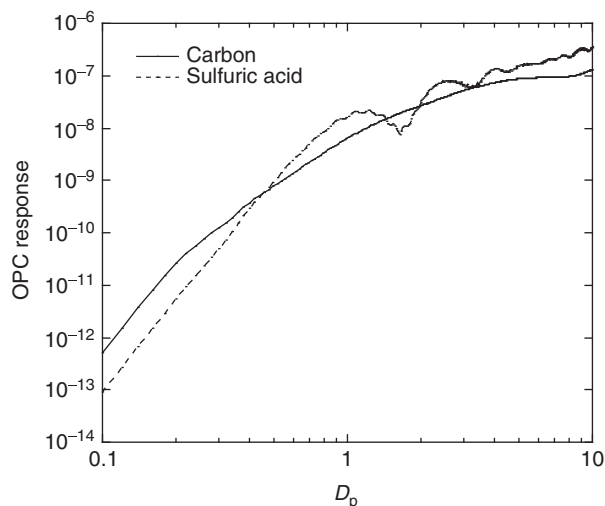


Figure 16 Numerically calculated responses of a laser optical particle counter (OPC) to spherical carbon and sulfuric particles as a function of size. Notice that above 0.7 μm , sulfuric acid particles of three different sizes can produce the same instrument response. The different responses to carbon and sulfuric acid are due to the effect of refractive index on light scattering. Courtesy of Professor W. Szymanski, Department of Physics, University of Vienna, Vienna, Austria.

A schematic of a scanning mobility particle spectrometer (SMPS) commonly used for such measurements is shown in (Figure 15). The aerosol is brought to a Boltzmann equilibrium charge distribution by exposing particles to a high concentration of mixed positive and negative gaseous ions. At Boltzmann equilibrium the most common charge state is

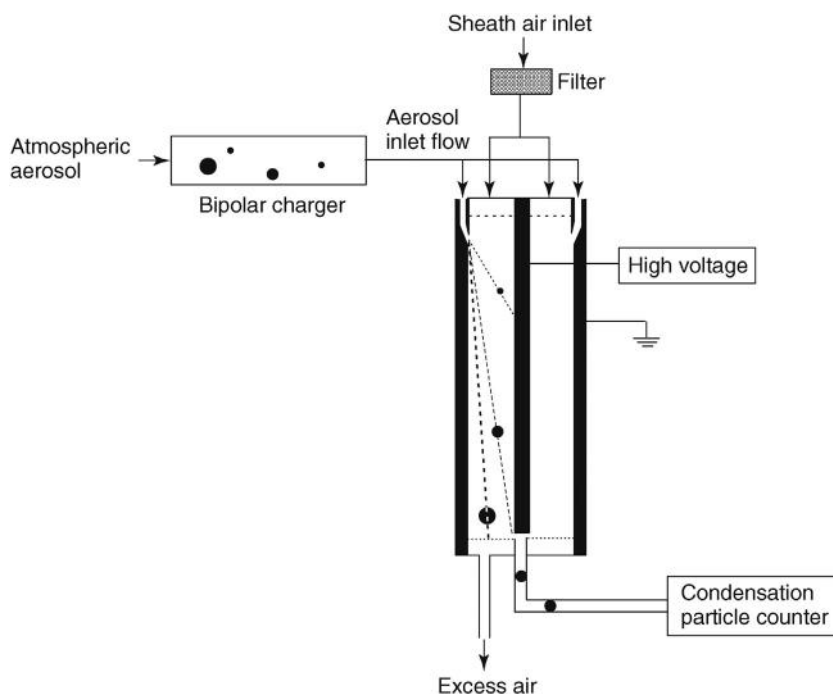


Figure 15 Schematic of a scanning mobility particle spectrometer (SMPS) used to measure size distributions in the ~ 10 nm to 500 nm diameter range. The particle size in the sample flow at the bottom of the instrument is changed by varying the voltage applied to the center cylindrical rod. Size distributions are found by measuring the concentration of classified particles over a range of classifying voltages.

neutral. However, a statistically predictable fraction of particles contains ± 1 , ± 2 , ± 3 , etc. charges, and this distribution varies with particle size. Particles smaller than roughly 50 nm contain very few multiply charged particles, while particles of 0.5 μm contain more multiply charged than singly charged particles.

After the aerosol charge distribution has been adjusted the aerosol flows into the differential mobility analyzer (DMA), which is the heart of the SMPS. The DMA classifies particles according to their electrical mobility by flowing the aerosol through an annular gap between two coaxial cylindrical electrodes. The laminar flow between these electrodes includes the aerosol-containing flow, which enters along the inner wall of the outer cylinder and typically accounts for about 10% of the total flow, and particle-free sheath flow, which occupies the inner portion of the annulus. The outer electrode is maintained at ground while a voltage is applied to the inner electrode. If the center rod is positively charged then negatively charged particles migrate radially towards the inner electrode as the flow draws them axially through the annulus. Particles with high

electrical mobility deposit on the inner electrode, while those with low electrical mobility exit the DMA with the excess air. Particles in a narrow range of electrical mobilities exit the DMA through a narrow gap on the inner electrode and travel downstream to a detector, typically a condensation particle counter. Size distributions are obtained by measuring the concentration downstream of the DMA over a range of classifying voltages so as to cover a range of electrical mobilities. Electrical mobility depends on particle charge, geometric size, and shape; the electrical mobility size of a spherical particle equals its geometric size. Uncertainties in measured size distributions occur when a significant fraction of the measured particles are nonspherical. 'Inverting' the raw data to obtain size distributions requires taking into account the multiplicity of sizes that are obtained at any given classifying voltage, since particles of a given mobility can contain one or more elementary charges.

Optical particle counters (OPCs) are used to measure size distributions of particles as small as 50 nm, although a lower size limit of 0.1 to 0.3 μm is more typical. OPCs function by

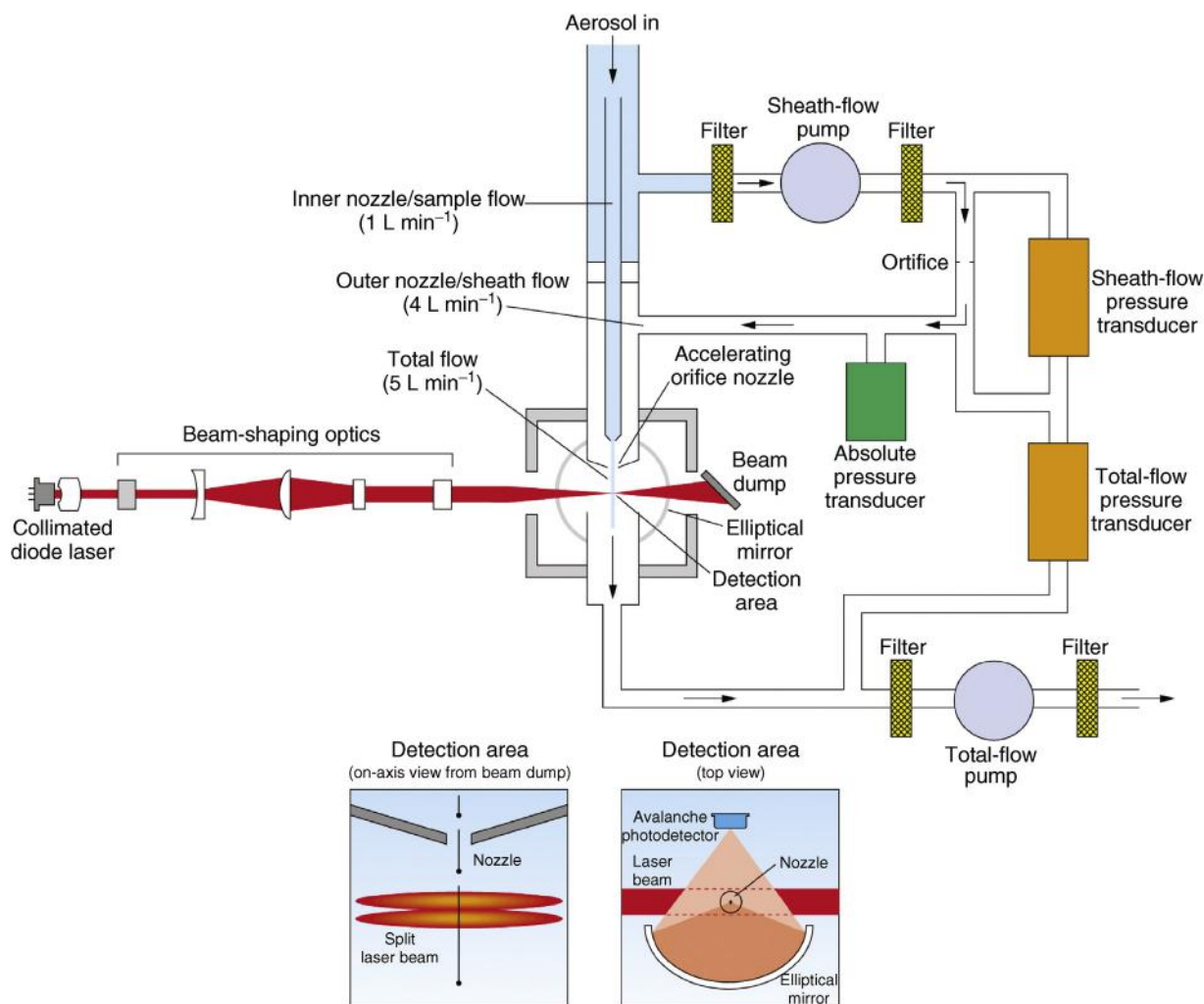


Figure 17 Schematic of an aerodynamic particle sizer (APS). Particles are accelerated through a nozzle to a velocity that decreases with increasing particle size due to particle inertia. Size is inferred from velocity, which is determined from the time required for the particle to travel between the two scattering regions formed from the split laser beam. Courtesy of TSI, Inc.

passing particles through a small 'scattering volume' into which an intense source of light has been focused. As particles pass through the scattering volume they scatter light, which is collected by mirrors and sent to an optical detector which converts the scattered light to a voltage pulse that varies in proportion to the intensity of the scattered light. Size distributions are obtained by first establishing the relationship between pulse height and particle size for particles of known size and composition. Using this calibration, measured pulse height distributions for unknown aerosols can be 'inverted' to obtain size distributions.

The amount of light scattered by a particle depends on its refractive index and on its geometric size and shape. For any given particle, the amount of light scattered can depend on the solid angle over which scattered light is collected in a complicated manner, and particles of different sizes can produce the same OPC response. The response of a laser OPC to spherical carbon and sulfuric acid particles of known geometric size is shown in [Figure 16](#). The multivalued relationship between response and particle size seen for the sulfuric acid (e.g., the same response occurs for particles of 0.9, 1.4, and 1.8 μm) is less pronounced when particles scatter incandescent light, since the optical resonances that lead to the oscillations shown in [Figure 16](#) are wavelength-dependent and lead to a more nearly monotonic response when averaged over a broad distribution of wavelengths.

Aerodynamic particle sizers (APSs) ([Figure 17](#)) also use light scattering to determine particle size. The size measuring principle, however, is quite different from that employed with an OPC. In an APS, particles are accelerated through a nozzle and then pass through two focused laser beams that intersect the particle beam at right angles. A light-scattering pulse is obtained as particles pass through each of the lasers, and the particle velocity is determined from the known distance between the lasers and the time required to traverse this distance. Because of their inertia, large particles tend to lag behind the carrier gas as it accelerates through the nozzle, and they achieve a terminal velocity less than the gas-velocity. As particle size decreases, particles accelerate more readily with the gas. Therefore, the terminal particle speed approaches that of the gas. APSs infer particle size from the measurement of the particle's terminal speed. As with impactors, which also classify particles according to

aerodynamic size, measurements depend on the particle's geometric size, shape, and density.

Summary

Recent interest in the effects of environmental aerosols on radiative transfer through the atmosphere and on human health have led to significant advances in instrumentation for measuring aerosol physical and chemical properties. It is now possible, for example, to measure physical size distributions down to 3 nm diameter with a time resolution of several minutes, and instruments for measuring the composition of individual particles in real time are being developed and used for atmospheric studies. While most measurements of aerosol composition to date have involved the analysis of samples collected on filters, instruments are currently being developed that provide real-time information on aerosol composition with a much better time resolution than can be obtained with filters. These advances are leading to significant refinements in our understanding of the origins of atmospheric particles and of the role they play in atmospheric chemistry.

See also: **Aerosols:** Aerosol Physics and Chemistry; Aerosol-Cloud Interactions and Their Radiative Forcing; Climatology of Tropospheric Aerosols; Role in Radiative Transfer. **Satellites and Satellite Remote Sensing:** Aerosol Measurements. **Tropospheric Chemistry and Composition:** Aerosols/Particles.

Further Reading

- Chow, J.C., 1995. Measurement methods to determine compliance with ambient air quality standards for suspended particles. *Journal of the Air and Waste Management Association* 45, 320–382.
- Hinds, W.C., 1999. *Aerosol Technology: Properties, Behavior, and Measurement of Airborne Particles*, second ed. Wiley, New York.
- McMurry, P.H., 2000. A review of atmospheric aerosol measurements. *Atmospheric Environment* 34, 1959–1999.
- Quinn, P.K., Anderson, T.L., Bates, T.S.D., et al., 1996. Closure in tropospheric aerosol-climate research: direct shortwave radiative forcing. *Beitrag zur Physikalischen Atmosphäre* 69, 547–577.
- Willeke, K., Baron, P.A., 1993. *Aerosol Measurement: Principles, Techniques, and Applications*. Van Nostrand Reinhold, New York.

Role in Radiative Transfer

GA Ban-Weiss, Lawrence Berkeley National Laboratory, Berkeley, CA, USA and University of Southern California, Los Angeles, CA, USA
WD Collins, Lawrence Berkeley National Laboratory, Berkeley, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by W Collins, volume 1, pp 48–53, © 2003, Elsevier Ltd.

Synopsis

Atmospheric aerosols (i.e., particles suspended in air) interact with shortwave radiation from the sun and longwave thermal radiation from the Earth. These interactions profoundly influence the energy budget and climate of the planet. In this article, we first present a theoretical framework for quantifying aerosol–radiation interactions. We then discuss the global quantification of aerosol properties using sensors on satellites that measure upwelling radiation, along with metrics that can be used to estimate global impacts of aerosol–radiation interactions. Finally, we conclude with a discussion on new frontiers in global measurement of aerosol properties relevant for better understanding the role of aerosols on climate.

Introduction

Aerosols suspended in the Earth's atmosphere interact with light incident from the Sun and with terrestrial thermal radiation. The study of these interactions is important for understanding the present and future climate system. Naturally occurring aerosols reflect some of the incident solar radiation back to space before it can be absorbed and converted to heat: these include dust, sea salt, and stratospheric sulfate aerosols formed during volcanic eruptions. Anthropogenic aerosols released by the combustion of biomass and fossil fuels increase both the reflection of solar radiation back to space and its absorption within the Earth's atmosphere. It is necessary to include the radiative effects of aerosols in modern climate models in order to simulate recent changes in global surface temperature with reasonable fidelity.

For radiation incident on an aerosol particle, the two fundamental interactions are scattering and absorption. The combination of scattering and absorption is termed extinction.

Scattering is the deflection of an electromagnetic wave from its original direction of propagation. In geometrical optics, scattering occurs when the wave propagates through a mixture of media with different indices of refraction, for example, aerosols suspended in air. The index of refraction is the ratio of the speed of light in vacuum to that in a given medium. In a more familiar setting, the difference in the indices of refraction for air and glass is responsible for the focusing effect of lenses.

Absorption is the conversion of electromagnetic energy into internal motions of molecules within the aerosol. The molecules can rotate and vibrate in discrete modes determined by quantum mechanics. The amount of rotational or vibrational energy in each mode corresponds to a particular frequency of light. The frequency of light is the number of oscillations in the electromagnetic wave per unit time, and for visible light the frequency corresponds to color. When the frequency of radiation incident on an aerosol corresponds to the quantum modes of the aerosol molecules, some of the radiation is converted into internal energy. Aerosols can also emit thermal radiation following Kirchhoff's laws for the equilibrium between matter and radiation at constant temperature. The amount of thermal energy emitted from an aerosol particle at each frequency is governed by Planck's theory. Due to the moderate temperatures

in the Earth's atmosphere, almost all of the thermal radiation is emitted at infrared wavelengths. The emission of light from the Sun is also governed by Planck's theory, although the emission temperature is approximately 5770 K, corresponding to higher frequency (and lower wavelength) radiation than that of terrestrial radiation.

The interaction of aerosols with sunlight and terrestrial radiation can be calculated using Maxwell's equations for electromagnetic radiation. These equations may be written in an approximate form for radiation propagating through a mixture of gases, aerosols, and clouds. The equations are supplemented with Planck's theory for the emission of light, the optical properties of atmospheric constituents, and data on the distribution and concentrations of these constituents. The solutions quantify the effects of natural and anthropogenic aerosols on the net energy balance of the planet, the radiant energy absorbed in the atmosphere, and the sunlight and thermal radiation incident on the Earth's surface.

Scattering and Absorption of Light

Beer's Law for Direct Sunlight

Solar radiation incident at the top of atmosphere (TOA) can be treated as a traveling plane wave normal to the line between the centers of the Earth and the Sun. The solar radiation propagating through the atmosphere can be partitioned into direct and diffuse radiation. Direct radiation is the beam of sunlight that has not interacted with the constituents of the atmosphere or the Earth's surface. Diffuse radiation is sunlight that has been scattered by atmospheric constituents and/or by the surface. Direct radiation is converted to diffuse radiation when it first scatters from an atmospheric component or from the surface.

As direct radiation propagates from the tenuous upper atmosphere to the dense lower atmosphere, its intensity is reduced by both scattering and absorption. The basic principle governing the extinction of the direct beam is known as Beer's law or as the Beer–Bouguer–Lambert law. The principle states that the extinction through an optical medium is linear in the intensity of radiation incident upon that medium. For simplicity, only the effects of aerosols will be considered. Let ν denote the frequency of light in units of inverse time. Let $I_d(\nu)$

represent the intensity of direct solar radiation at a frequency ν . It is expressed in units of energy per unit frequency interval, time, area, and solid angle. The solid angle is a measure of the directions in which the wave is propagating. Let z denote the distance from the TOA. In mathematical form, Beer's law is given by eqn [1].

$$\frac{dI_d(\nu)}{dz} = -\rho k_e(\nu) I_d(\nu) \quad [1]$$

Here ρ is the density of the aerosols in units of mass per unit volume of air and $k_e(\nu)$ is the specific extinction in units of area per unit mass of aerosol.

The specific extinction is determined by the chemical composition of the aerosols and by the number of aerosol particles of different sizes per unit volume (a.k.a. the aerosol size distribution). For spherical aerosol particles, the extinction may be computed using Mie theory. Mie theory is often applied to nonspherical particles as a first approximation, for example, dust particles and complex chains of carbon compounds. The extinction is the sum of the effects of scattering and absorption denoted by $k_s(\nu)$ and $k_a(\nu)$, respectively (eqn [2]).

$$k_e(\nu) = k_s(\nu) + k_a(\nu) \quad [2]$$

Optical depth is a unitless measure of how much the radiation is reduced between two points, in this case from the TOA to an altitude z . The optical depth is equal to the integral of extinction over altitude (eqn [3]).

$$\tau_\nu(z) = \int_z^\infty \rho k_e(\nu) dz \quad [3]$$

Beer's law given in eqn [1] can be rewritten as eqn [4] using the definition of optical depth. Equation [4] shows that an increase in optical depth over some altitude interval is equal to the fractional reduction of intensity in that interval.

$$\frac{dI_d(\nu)}{I_d(\nu)} = -d\tau_\nu \quad [4]$$

The solution for the direct radiation at an altitude z is simply given by eqn [5], where $I_{d,0}(\nu)$ is the incident intensity of sunlight at the TOA.

$$I_d(\nu) = I_{d,0}(\nu) \exp[-\tau_\nu(z)] \quad [5]$$

The exponential term is the transmission, which for the direct beam is the fraction of incident radiation reaching a particular altitude. The transmission is always between 0 and 1.

When the optical depth at some altitude is much greater than unity, Beer's law shows that nearly all the direct beam is extinguished before reaching that altitude. When the optical depth is much less than unity, the direct radiation is reduced by a fractional amount almost equal to $1 - \tau_\nu(z)$. The solution of the radiative transfer equations to lowest order in $\tau_\nu(z)$ is known as the single-scattering approximation. For example, the global annual-mean aerosol optical depth is roughly 0.1 for the present-day atmosphere. Beer's law predicts that aerosol extinction reduces the direct beam at the surface to $\exp(-0.1)$ of its incident value. This reduction is nearly identical to the estimate from the single-scattering approximation, which

predicts that the direct beam is reduced to $1 - 0.1 = 90\%$ of its incident value.

Local optical depths vary considerably about the global annual-mean value. There are regions where the aerosol optical depth is frequently large enough to invalidate the single-scattering approximation. Nonetheless, the approximation is sufficiently accurate for calculating the global reflection, absorption, and transmission of solar radiation by aerosols. The estimates of these properties will apply to the entire atmosphere and solar spectrum, so for simplicity the explicit dependence on altitude and frequency will be omitted from the equations and results.

Changes in Reflection, Transmission, and Absorption

The amount of radiation extinguished from the direct beam is approximately $\tau_\nu I_{d,0}(\nu)$. Some of the extinguished radiation is absorbed in the atmosphere, and the remainder is converted to diffuse radiation. This diffuse radiation can be further scattered or absorbed in the atmosphere, although these effects are omitted in the single-scattering approximation. One of the main effects of aerosols on climate is in changing the amount of sunlight reflected back to space. The change in the planetary reflectivity, or albedo, from aerosols is due mostly to radiation being reflected out of the atmosphere; this reflected sunlight leaves the atmosphere as diffuse radiation. (Absorption of sunlight by aerosols such as soot also alters planetary albedo.)

The process of scattering sends diffuse radiation in all directions. The mathematical formalism for treating the subsequent propagation of diffuse radiation is relatively complex. However, the main effects of aerosols can be estimated using the two-stream approximation to radiative transfer. In this simplification, all the radiation propagating downward toward the surface is combined in a single downward flux of energy, or 'stream.' This stream includes the direct beam and any diffuse radiation scattered toward the Earth's surface. The downward stream is represented by the symbol $F^\downarrow$. All the diffuse radiation propagating upward is combined into the upwelling stream denoted by $F^\uparrow$. The fluxes are in units of energy per unit area per unit time and are conventionally expressed in watts per square meter (W m^{-2}). For example, the global annual-mean solar flux incident upon the TOA is $F_0^\downarrow = 342 \text{ W m}^{-2}$.

If the optical depth τ of the aerosols is small and if the radiation reflected by the surface is negligible, the upwelling stream is given by eqn [6].

$$F^\uparrow = \omega b \tau F_0^\downarrow \quad [6]$$

The factor ω is the ratio of scattering to total extinction and is known as the single-scattering albedo. The single-scattering albedo is the fraction of incident energy scattered each time sunlight interacts with an aerosol particle. In the two-stream approximation, the factor b , known as the backscattering ratio, is the fraction of scattered energy that is reflected upward. For typical aerosols, b is roughly 0.3. The single-scattering albedo determines the total amount of radiation that is scattered in interactions with aerosols, and the backscattering ratio determines how much of the scattered radiation is directed toward space.

For example, if $\varpi = 0.95$, then 95% of sunlight is scattered and only 5% is absorbed. If $\varpi = 1$, no energy is removed from the sunlight by absorption, and the aerosol is called a conservative scatterer. For sulfate and sea-salt aerosols, $\varpi \approx 1$ for visible light. However, ϖ can be 0.3 or lower in visible wavelengths for black carbonaceous aerosols generated by combustion. Aerosols with $\varpi < 0.9$ in visible frequencies are considered to be quite absorptive.

From eqn [2], ϖ can be written in terms of the specific extinctions as eqn [7].

$$\varpi = \frac{k_s}{k_e} \quad [7]$$

The fractional absorption in each interaction of sunlight with aerosols is known as the coalbedo and is given by $1 - \varpi$. Even if the coalbedo is very close to 0 for a single interaction, the amount of absorption *accumulates* each time a beam of light interacts with aerosols. The radiant energy is reduced by a factor of $1 - \varpi$ after each interaction. After N interactions, the energy remaining is proportional to ϖ^N . In the single-scattering approximation, only one interaction is considered. Therefore, the total amount of energy absorbed is proportional to the coalbedo (eqn [13]).

The change in the planetary albedo ΔR from aerosols is the ratio of the upwelling reflected radiation $F^\uparrow$ to the incident radiation $F_0^\downarrow$ (eqn [8]).

$$\Delta R = \frac{F^\uparrow}{F_0^\downarrow} = \varpi b \tau \quad [8]$$

To provide a realistic estimate of the change in planetary albedo, the extinction by the rest of the atmosphere should be included in the calculation. The change in albedo is estimated using $\Delta R_a = f_a \Delta R$. The empirical factor f_a is approximately equal to 0.2 and includes the effects of atmospheric gases, water vapor, and clouds. For a nonabsorbing aerosol with $\varpi \approx 1$, $b = 0.3$, and $\tau \approx 0.1$, $\Delta R_a \approx 0.0006$. Since the global planetary albedo is approximately 0.3, the aerosols increase the planetary albedo by roughly 2% in relative terms. The corresponding increase in upwelling flux is approximately 2 W m^{-2} .

If the optical depth τ of the aerosols is small and if the radiation reflected by the surface is negligible, the downwelling stream at the Earth's surface is given by eqn [9].

$$F^\downarrow = (1 - \tau)F_0^\downarrow + (1 - b)\varpi\tau F_0^\downarrow \quad [9]$$

Aerosols have two effects on the downwelling solar radiation at the surface corresponding to the terms on the right-hand side of eqn [9]. The effects have opposite sign and therefore partially offset one another. The first effect is the reduction of the direct beam by scattering and absorption. The second effect is the addition of diffuse radiation that represents sunlight scattered from the direct beam toward the surface. Since b is the fraction of radiation scattered upward, $1 - b$ is the fraction of radiation scattered downward. The reduction is equal to $-\tau F_0^\downarrow$ and the addition is equal to $(1 - b)\varpi\tau F_0^\downarrow$. The ratio of the diffuse flux addition and direct-beam flux reduction is as given by eqn [10], which shows that the increase in diffuse flux is always less than the reduction in direct flux.

$$0 \leq (1 - b)\varpi \leq 1 \quad [10]$$

Therefore, the net effect of aerosols is to reduce the total amount of sunlight reaching the Earth's surface. The change in transmission is the ratio of the change in the downwelling flux to the flux incident on the TOA. The expression for the change in transmission is given by eqn [11].

$$\Delta T = \frac{F^\downarrow - F_0^\downarrow}{F_0^\downarrow} = (1 - b)\varpi\tau - \tau \quad [11]$$

The amount of sunlight absorbed by aerosols can be derived using the principle of conservation of energy, which states that the energy in a closed system is constant. The atmosphere is an example of an open system since it can exchange energy and material with the surface and with space. However, the combination of the atmosphere, the fluxes to the surface, and the fluxes to space comprises a closed system. In this system, the albedo determines the flux to space, the transmission determines the flux to the surface, and the absorption determines the conversion of light to internal energy within the atmosphere. The solar energy incident on the atmosphere does not change when aerosols are present. Therefore, the sum of changes in the albedo, transmission, and absorption must equal 0 (eqn [12]).

$$\Delta A + \Delta R + \Delta T = 0 \quad [12]$$

This condition can be solved for the absorption ΔA as in eqn [13].

$$\Delta A = -(\Delta R + \Delta T) = (1 - \varpi)\tau \quad [13]$$

Effects of Surface Reflection

When an aerosol layer is present above a nonreflective surface, it always increases the atmospheric albedo. As eqn [8] shows, ΔR is positive regardless of the optical properties of the aerosol. The surface reflectivity R_s is the fraction of radiation incident on the surface that is reflected back into the atmosphere. Under overhead Sun, the ocean may be treated as a nonreflective surface because the reflectivity is just 3%. However, there are surface types that reflect at least 50% of the incident sunlight, for example, sea ice, glaciers, snow, and some desert areas. The effects of surface reflectivity can easily be calculated using the two-stream and single-scattering approximations.

When the effects of surface reflectivity are included, the albedo ΔR computed in eqn [8] assuming $R_s = 0$ must be corrected in the following three ways:

1. Incoming radiation interacts with the aerosol as it propagates toward the surface. A fraction $1 + \Delta T$ is transmitted to the surface, and some of it is reflected back to space.
2. Radiation reflected from the surface interacts with the aerosol as it propagates toward space. A fraction $1 + \Delta T$ is transmitted to space.
3. Radiation reflected from the surface interacts with the aerosol as it propagates toward space. A fraction ΔR is reflected back to the surface, where some of it is reflected upward and is transmitted to space.

These effects are illustrated in Figure 1.

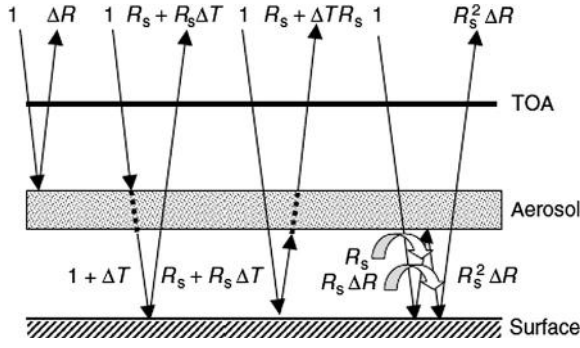


Figure 1 Interactions between incident sunlight, an aerosol layer, and a reflective surface. The incident solar flux is $F_0^{\downarrow} = 1$. The albedo and transmission of the layer are ΔR and ΔT , respectively, and the reflectivity of the surface is R_s . Solid black arrows point along the direction of the sunlight, and dashes indicate where the sunlight is attenuated by the aerosol layer.

In mathematical form, the change $\Delta R'$ in planetary albedo for aerosols above a reflective surface is given by eqn [14].

$$\Delta R' - \Delta R = (\Delta T)R_s + R_s(\Delta T) + R_s(\Delta R)R_s \quad [14]$$

The three terms on the right-hand side of eqn [14] correspond to the three effects discussed above. $\Delta R'$ can be simplified using the conservation of energy (eqn [12]) to rewrite ΔT in terms of ΔA and ΔR . After some algebraic manipulation, the resulting expression for the albedo is as given by eqn [15].

$$\Delta R' = (1 - R_s)^2 \Delta R - 2R_s \Delta A \quad [15]$$

It is easy to show that $\Delta R' < \Delta R$ whenever $R_s > 0$. This inequality states that surface reflectivity always reduces the effects of aerosols on the planetary albedo. There are two mechanisms for the reduction, corresponding to the two terms on the right-hand side of eqn [15]. First, when most of the sunlight incident on the surface is reflected, the sunlight reflected by aerosols is a relatively small component of the upwelling flux. If the surface is completely reflective, all solar radiation incident on the atmosphere is either scattered back to space or absorbed in the atmosphere. Additional scattering by aerosols cannot increase the diffuse radiation emitted to space. Therefore, $\Delta R' \leq 0$ when $R_s = 1$. Second, multiple scattering of diffuse radiation between the aerosol and the surface increases the fraction of incident radiation that is absorbed. Because energy is conserved, the absorption reduces the amount of sunlight that can be reflected to space.

Under certain conditions, the presence of aerosols will either reduce the planetary albedo or leave it unchanged. If the albedo is reduced, then conservation of energy requires that the absorption within the combined surface-atmosphere system must increase. More absorption results in greater internal energy, or heat, within the Earth's atmosphere or surface. Consequently, when $\Delta R' \leq 0$ there is additional heating within the Earth-atmosphere system. The conditions for heating follow from substitution of ΔR and ΔA (eqns [8] and [13]) into the inequality $\Delta R' \leq 0$. The criterion, which was first discovered by Chylek and Coakley, is given as [16].

$$\frac{(1 - \varpi)}{\varpi b} \geq \frac{(1 - R_s)^2}{2R_s} \quad [16]$$

This condition is independent of the total aerosol optical depth (or mass) present in the atmosphere. It depends only upon intrinsic optical properties of the aerosol. Regardless of the value of R_s , the condition can always be satisfied with sufficiently small values of either the backscatter fraction b or the single-scattering albedo ϖ . Assuming $R_s \cong 0.12$, for aerosols with a typical value of $b = 0.3$, the criterion indicates that aerosols heat the climate system if $\varpi \leq 0.51$. Since measured values of ϖ often do not fall below 0.8, $\Delta R' > 0$. (An exception is soot aerosol, which is sometimes observed with lower ϖ .) Thus, aerosols cool the global climate system by reflecting sunlight back to space before it can be absorbed. Over regions with persistent ice or snow cover at high latitudes, however, R_s can easily exceed 0.5. When $R_s \cong 0.5$, the criterion for heating is $\varpi \leq 0.93$. Carbonaceous particles released by combustion frequently satisfy this condition on single-scattering albedo. Polar regions are therefore susceptible to heating when aerosols are only moderately absorptive.

Interactions with Thermal Radiation

Aerosols interact with thermal infrared radiation from Earth's surface and atmosphere, and aerosols can emit additional thermal radiation. For aerosols near the Earth's surface, the effects on thermal radiation are generally much less significant than the effects on solar radiation. However, the thermal effects of thick dust plumes and stratospheric aerosols are sufficiently large to register on satellite instruments.

At infrared frequencies, the extinction by aerosols is mostly due to absorption. To a reasonable approximation, scattering may be omitted from calculations of infrared fluxes. In addition, the atmosphere and surface both act as sources of thermal radiation. The flux of infrared radiation from the Sun is actually much smaller than these terrestrial sources and can be neglected in most applications.

The Stefan-Boltzmann law states that the thermal flux from a perfect thermal source at temperature T is σT^4 . The constant σ is the Stefan-Boltzmann constant and is approximately equal to $5.67 \times 10^{-8} \text{ W m}^{-2} \text{ K}^{-4}$. A perfect thermal source, or 'black body,' is an optical medium that absorbs all thermal radiation incident upon it. Aerosol layers are rarely thick enough to qualify as perfect thermal sources. Instead, the aerosols may be treated as 'gray bodies' that emit a thermal flux equal to $\varepsilon \sigma T^4$. The factor ε is known as the emissivity, and is a unitless number between 0 and 1. The closer ε is to 1, the more the gray body resembles a perfect thermal source.

Suppose the gray body is surrounded by a perfect thermal medium at the same temperature T . From the laws of thermodynamics, the gray body will be in thermal equilibrium with the surrounding medium. In equilibrium, the internal energy of the gray body is constant. Since the gray body is emitting a flux of thermal energy equal to $\varepsilon \sigma T^4$, it must also absorb energy at the same rate. The flux incident on the gray body from the medium is σT^4 and a fraction ΔA of this flux is absorbed. Because thermal equilibrium requires that the absorbed and emitted fluxes are equal, $\Delta A = \varepsilon$. The transmission can be derived from the conservation of energy (eqn [12]). If the gray body does not

reflect thermal radiation and $\Delta R \cong 0$, then the transmission is given by eqn [17].

$$\Delta T \cong -\Delta A = -\epsilon \quad [17]$$

To simplify the analysis, various components of the atmosphere may also be treated as gray bodies. The emissivity of the stratosphere will be denoted by ϵ_* .

The thermal effects of aerosols in the stratosphere and near the surface are illustrated in Figures 2 and 3. Stratospheric aerosols are often created by volcanic eruptions, and aerosols near the surface result from natural and anthropogenic processes. The infrared fluxes can be analyzed with the same two-stream techniques applicable to solar radiation. Let $F_t^\uparrow$ represent the upwelling flux from the troposphere incident on the bottom of the stratospheric aerosol layer. Let $F_a^\downarrow$ represent the downwelling flux from the atmosphere above the surface aerosol layer incident on the top of the layer. The temperature and emissivity of the stratospheric aerosol are T_v and ϵ_v , and the temperature and emissivity of the surface aerosol are T_s and ϵ_s respectively.

The change in the upwelling flux at the TOA has two components. The first represents the reduction in $F_t^\uparrow$ by partial transmission through the aerosol layer. The second represents the additional thermal emission from the aerosol layer. The total change in upwelling flux is given by eqn [18].

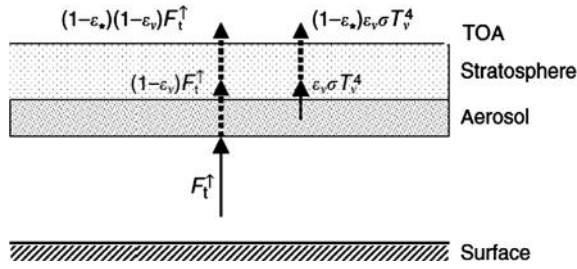


Figure 2 Interactions of thermal infrared radiation with an aerosol layer adjacent to the stratosphere. The temperature and emissivity of the aerosol layer are T_v and ϵ_v , respectively, and the emissivity of the stratosphere is ϵ_* . The upwelling infrared flux from the troposphere incident upon the bottom of the aerosol layer is $F_t^\uparrow$. Arrows point along the direction of the radiation, and dashes indicate where the radiation is attenuated by the aerosol layer and the stratosphere.

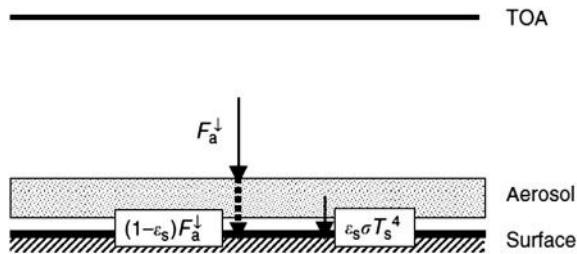


Figure 3 Interactions of thermal infrared radiation with an aerosol layer adjacent to the surface. The temperature and emissivity of the aerosol layer are T_s and ϵ_s , respectively. The downwelling infrared flux from the atmosphere incident upon the top of the aerosol layer is $F_a^\downarrow$. Arrows point along the direction of the radiation, and dashes indicate where the radiation is attenuated by the aerosol layer.

$$\Delta F_v^\uparrow = \epsilon_v (\sigma T_v^4 - F_t^\uparrow) (1 - \epsilon_*) \quad [18]$$

A similar analysis of the surface aerosol layer shows that the change in downwelling thermal flux is as given by eqn [19].

$$\Delta F_s^\downarrow = \epsilon_s (\sigma T_s^4 - F_a^\downarrow) \quad [19]$$

Reasonable values for the aerosol parameters for aerosols located in the tropics are given in eqn [20].

$$\begin{aligned} T_v &= 195 \text{ K} & \epsilon_v &= 0.005 \\ T_s &= 295 \text{ K} & \epsilon_s &= 0.05 \end{aligned} \quad [20]$$

For the purposes of illustration, the unperturbed atmospheric parameters are given in eqn [21].

$$\begin{aligned} \epsilon_* &= 0.088 \\ F_t^\uparrow &= 240 \text{ W m}^{-2} \\ F_a^\downarrow &= 380 \text{ W m}^{-2} \end{aligned} \quad [21]$$

The resulting changes in the upwelling and downwelling thermal infrared fluxes are given in eqn [22].

$$\begin{aligned} \Delta F_v^\uparrow &= -0.7 \text{ W m}^{-2} \\ \Delta F_s^\downarrow &= 2.5 \text{ W m}^{-2} \end{aligned} \quad [22]$$

In this illustrative calculation, aerosols in the stratosphere decrease the outgoing infrared flux, while aerosols near the surface increase the downwelling infrared flux. In more realistic calculations, the thermal infrared flux at the TOA will generally decrease and the flux at the bottom will generally increase. The basic reason is that the global infrared flux emitted to space is equal to the radiation from a perfect thermal source at temperature $T = 255.5 \text{ K}$. This atmospheric radiation is not in equilibrium with aerosol layers at temperatures different from T . The absence of equilibrium means that the colder stratospheric aerosols will absorb some of the upwelling atmospheric flux, thereby reducing the flux emitted to space. Similarly, the warmer stratospheric aerosols emit more radiation to the surface than the atmosphere alone.

Global Quantification of Aerosol–Radiation Interactions

As has been presented, interactions of atmospheric aerosols and radiation can be described by aerosol optical depth τ , single-scattering albedo ω , backscattering fraction b , and the reflectivity of the Earth surface R_s . These properties can be globally quantified using sensors on satellites. The sensors are designed to measure radiation reflected and emitted by the atmosphere and Earth surface in discrete narrow wavelength bands. With knowledge of how the Earth surface and atmosphere reflects, absorbs, and emits radiation in these narrow bands, properties of the Earth system can be derived (see **Satellites and Satellite Remote Sensing: Aerosol Measurements; Research**). The last decade has seen major advances in satellite technologies including the deployment of instruments specifically intended to measure aerosol properties. The future promises additional advances to improve global characterization of atmospheric aerosols and associated climate effects.

Aerosol Optical Depth

Total column optical depth τ is the most routinely globally retrieved aerosol quantity. Regions with high aerosol optical depths have strong aerosol–radiation interactions. Aerosol optical depth is currently being measured by multiple satellite sensors. As an example, [Figure 4\(a\)](#) shows τ from the moderate resolution imaging spectroradiometer (MODIS) on board NASA's Terra satellite. Though MODIS measures τ for all aerosol types, additional processing was performed such that this figure shows τ for only a portion of the fine-mode aerosols, mainly smoke and pollution particles. In general, these

aerosols are primarily carbonaceous (black and organic carbon) and sulfur-containing species. Larger particles that are often naturally produced such as dust and sea salt are not included in this figure. Plumes with high aerosol optical depth can be observed in industrialized regions with heavy fossil fuel combustion (e.g., North America, Europe, and Eastern Asia) and areas with biomass burning (e.g., Central Africa, eastern Siberia, western and northern United States, and Canada). The plumes disperse over land and ocean regions following circulation patterns, ultimately dying off via transformation and removal processes downwind of the sources. The global mean

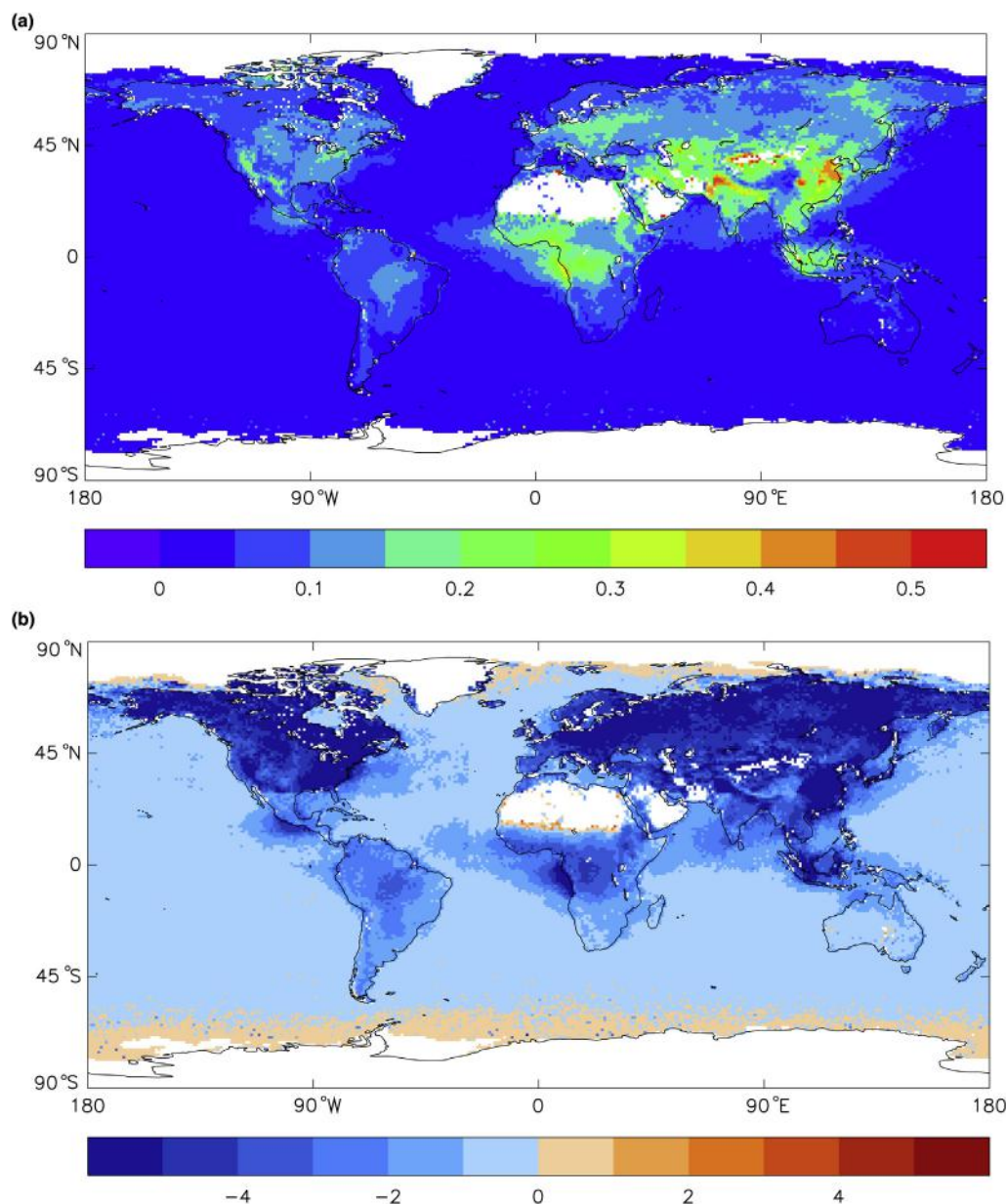


Figure 4 Distributions of (a) aerosol optical depth at 0.55 μm and (b) shortwave direct radiative forcing (DRF) from a portion of fine-mode aerosols. Distributions are derived from MODIS collection 5 and represent annual means for 2002. White indicates areas without data. Reprinted from Bellouin, N., Jones, A., Haywood, J., Christopher, S., 2008. Updated estimate of aerosol direct radiative forcing from satellite observations and comparison against the Hadley Centre climate model. *Journal of Geophysical Research* 113. <http://dx.doi.org/10.1029/2007JD009385>.

optical depth τ shown in Figure 4(a) is 0.043. Optical depth is 0.021 and 0.107 over ocean and land regions, respectively.

Though MODIS can retrieve aerosol optical depth at most locations around the globe, the presence of clouds and/or highly reflective surfaces (e.g., ice or desert sand) precludes its calculation (see, e.g., the lack of AOD values in Northern Africa in Figure 4). This is because the contribution to total upwelling radiation from these highly reflective surfaces can far outweigh the contribution from the atmospheric aerosol. However, other satellite instruments such as ozone monitoring instrument (OMI), multiangle imaging spectroradiometer (MISR), and polarization and directionality of Earth's reflectances (POLDER) are capable of retrieving aerosol optical depth above somewhat higher reflectance backgrounds. OMI does this by measuring aerosol absorption in the ultraviolet (UV); scattering of UV light is highly efficient by molecules in the atmosphere and thus minimal UV reaches the surface. (Moreover, most surfaces on Earth are dark in the UV.) Since minimal UV radiation reaches the surface of Earth, aerosol optical depth as measured by OMI is less sensitive to aerosols at lower altitudes than higher altitudes. MISR works by observing the Earth system simultaneously at nine different viewing angles and four wavelengths ranging from the blue to near-infrared. POLDER separates reflected light from the atmosphere and the surface by measuring its polarization at multiple angles.

Direct Radiative Forcing

Ultimately we are interested in diagnosing how aerosol–radiation interactions affect Earth's climate. Since aerosols reflect sunlight and thus partially offset warming from greenhouse gases, accurately quantifying aerosol–radiation interactions is critical to understanding anthropogenic global climate change. Direct radiative forcing (DRF) is a frequently used metric to describe the change in net radiation at the TOA caused by scattering and absorbing aerosols. Aerosol-induced changes

in clouds that indirectly alter net radiation are not included in DRF but are discussed in Aerosols: Aerosol–Cloud Interactions and Their Radiative Forcing.

Various methods are used to calculate DRF using satellite products, most of which require some modeling. For example, with knowledge of the global distribution of τ from aerosols (Figure 4(a)), DRF can be derived using radiative transfer modeling. Figure 4(b) shows such an estimate of shortwave DRF. Other than τ , the radiative transfer model takes as inputs the aerosol optical properties already discussed, the aerosol vertical profile, and the surface albedo. However, until recently the aerosol optical properties and vertical profiles have not been measured at global scale. Thus, aerosol optical properties have often been derived with surface measurements using sun-photometers, assuming that the measurements at these point locations are representative of wider areas. (See the next section on recent advances for globally measuring τ with satellites.) Further, while the vertical profile of aerosols is an important input to the radiative transfer model, the lack of global measurements has generally required the use of simplified assumptions. (See the next section on recent advances in measuring aerosol vertical profiles with satellites.) Lastly, the surface albedo can be globally quantified using satellite sensors such as MODIS (Figure 5).

It can be observed in Figure 4(b) that DRF is highest in areas with high values of τ (Figure 4(a)). The global mean shortwave DRF shown in this figure is -1.30 W m^{-2} , while values over ocean and land are -0.60 and -3.27 W m^{-2} , respectively. Note that shortwave DRF does not include changes in thermal (longwave) radiation. Changes in thermal radiation are expected to be small for the fine-mode but may be of importance for larger naturally produced aerosols (Section Longwave Forcing). To get a sense of the magnitude of the aerosol DRF, the increase in atmospheric abundance of greenhouse gases from preindustrial times to the year 2005 caused a radiative forcing of approximately 2.9 W m^{-2} .

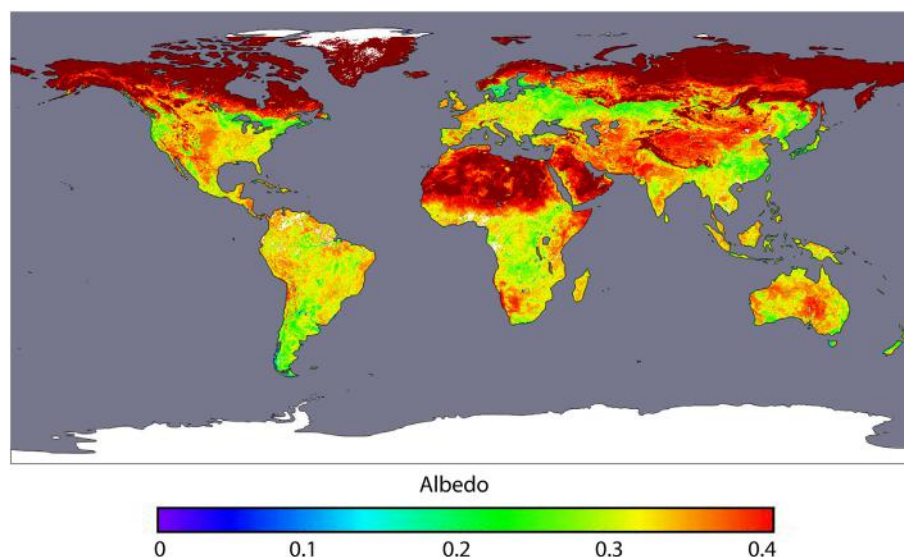


Figure 5 Albedo of the land surface during April 2002 derived using MODIS. White indicates regions without available data. Albedo of the ocean is not shown but is generally low (<0.1). Quantifying aerosol optical depth using satellites is more challenging over areas with high surface albedo (shown in red) such as deserts and snow/ice covered parts of Earth. Original image by Crystal Schaaf. <http://earthobservatory.nasa.gov>.

New Frontiers in Global Measurements of Aerosols

As discussed in the previous section, accurately quantifying DRF using radiative transfer models requires accurate estimates of the global distribution of aerosol optical properties and vertical profiles. Until recently these estimates have been mostly based on point measurements or assumed. While point measurements can be very accurate, they are not always representative of global distributions. Recent advances in satellites are making such measurements possible.

Single-Scattering Albedo

Recall that ω conveys the relative amount of scattering vs absorption of radiation by a particle. Aerosols with low ω like soot can reduce sunlight at the Earth surface but also heat Earth's atmosphere, leading to changes in the hydrological cycle in areas with heavy loading. Single-scattering albedo is currently being measured by both OMI and MISR. The OMI, which is on board NASA's AURA satellite, has been retrieving single-scattering albedo since 2005. **Figure 6** shows annual mean ω for 2007. Note that this figure includes ω from both fine-mode and larger aerosols. Single-scattering albedo is generally high (>0.86) and thus solar radiation is mostly scattered rather than absorbed by aerosols around the globe. Regions with increased carbonaceous aerosol from combustion have lower ω than remote ocean regions containing mostly sea-salt aerosol. Regions with high dust loading (e.g., Northern Africa) also show relatively lower ω . It should be noted that retrievals of single-scattering albedo have relatively high uncertainty. For OMI, the high uncertainties stem in part from assumptions that are required about the vertical profile of

aerosol. Additionally, results are derived using only UV radiation even though ω is a wavelength-dependent property, and for radiative forcing, its value in the visible part of the spectrum is most important. New methods are being used to estimate single-scattering albedo in the visible using OMI.

The future may bring new techniques for quantifying ω . One possibility is the use of hyperspectral satellites. Historically, most satellite sensors have measured at most tens of spectral bands. But new hyperspectral sensors have thousands of channels, improving the fidelity at which the atmosphere can be characterized. Additionally, new algorithms for existing satellite sensors are continually being developed. For example, an algorithm has been proposed to derive ω based on sun glint, which is a phenomenon that occurs when the satellite sensor captures mirrorlike reflection of sunlight off the ocean. Sun glint portions of satellite images can sometimes reveal atmospheric properties that are not ordinarily retrievable. In this case, sun glint regions are used as a background for viewing aerosol absorption properties. Though attempts are being made to use this method with existing satellites such as MODIS, a satellite designed to continually observe glint regions would provide better sampling statistics. As a second example, a new algorithm is currently in development for OMI that allows for characterizing ω (and τ) above clouds.

Aerosol Vertical Profiles

Global characterization of aerosol vertical profiles is one of the most significant recent advances in remote sensing of particles. Knowledge of the vertical profile of particles is important in determining its influence on radiation, but also its atmospheric lifetime. Both are closely related to the vertical position of the

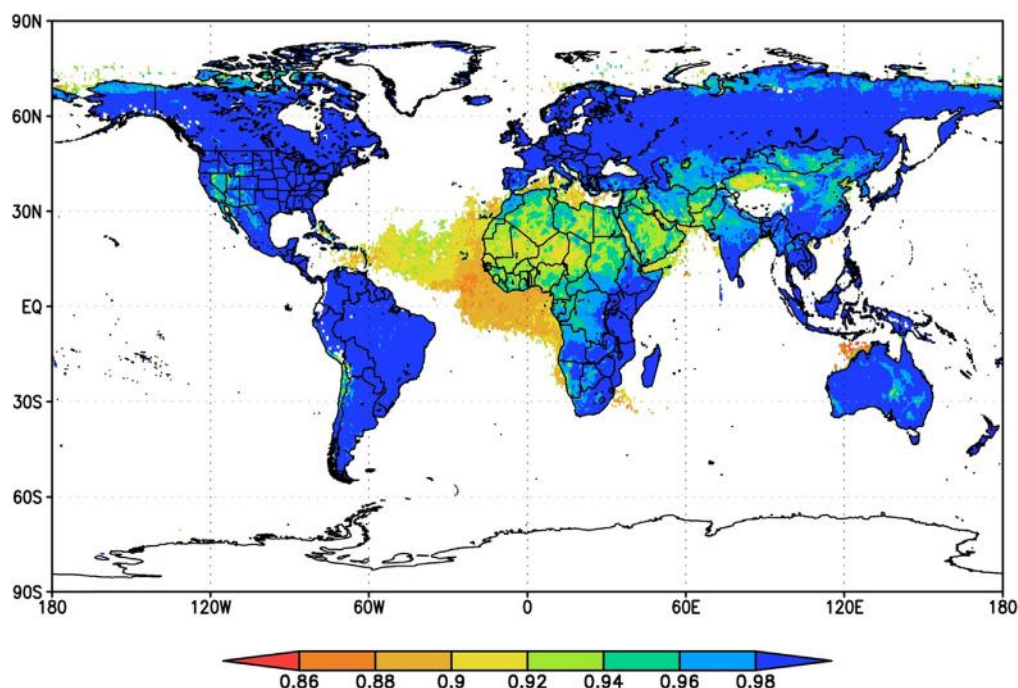


Figure 6 Single-scattering albedo ω derived using the UV channel of the ozone monitoring instrument (OMI), which was launched in 2005. This is the annual mean for 2007. Figure courtesy Omar Torres.

aerosol relative to clouds. Aerosols below clouds will experience lower incident radiation and are more likely to rain out via wet deposition. Aerosols above clouds will experience higher incident radiation (i.e., both downwelling radiation from the sun and upwelling radiation that has been reflected by the clouds below) and will have longer lifetimes due to the lack of rainfall from above. Aerosols with low τ that are positioned above clouds can absorb a significant amount of radiation and thus warm the surrounding air.

The cloud-aerosol lidar and infrared pathfinder satellite observation (CALIPSO) satellite, launched in 2006, is providing new insight into aerosol layer thickness, composition, and altitude. This satellite contains the first 'active' light detection and ranging (LIDAR) in space that is optimized to characterize aerosols (and clouds). LIDAR is similar to radio detection and ranging (RADAR) but uses shorter wavelength energy from laser beams rather than radio energy. Given the active nature of LIDAR, CALIPSO can characterize both the sunlit and dark parts of Earth. (Passive sensors that measure reflected shortwave radiation from the sun cannot characterize the atmosphere at night.) Vertically resolved measurements from CALIPSO can now be used for characterization of Earth's atmosphere, comparison to global climate models, and as inputs to calculations of climate relevant metrics such as shortwave DRF.

An example of an aerosol vertical profile measured by CALIPSO is shown in Figure 7. The red line in the inset of the

figure shows that the location of the transect is off the west coast of Africa where biomass burning aerosols are frequently encountered. The color scale of the vertical profile represents the intensity of light (532 nm) from the LIDAR that was backscattered. The blue and purple colors indicate minor backscattering from air molecules in the clean atmosphere, the light green colors indicate moderate scattering by an elevated layer of biomass burning aerosol, and the red colors indicate heavy backscattering by a layer of marine clouds, which are frequently observed in this area. Backscattering values underneath the cloud are not always reliable because the LIDAR signal cannot fully penetrate the optically thick cloud.

Longwave Forcing

In general DRF includes changes in TOA radiation caused by altering both shortwave and thermal (longwave) radiation. We are often concerned only with DRF from anthropogenic aerosol, which in most cases has negligible longwave DRF. Interactions of aerosols with thermal radiation are significant only for particles with diameter similar in magnitude to wavelengths in the thermal region of the electromagnetic spectrum; generally only naturally occurring aerosols such as dust and sea-salt have sufficiently large diameter. Also, these large diameter aerosols can have a more significant effect on the thermal energy budget of Earth when at high altitudes vs low altitudes. This is because high altitude aerosols are cold. Thus, the

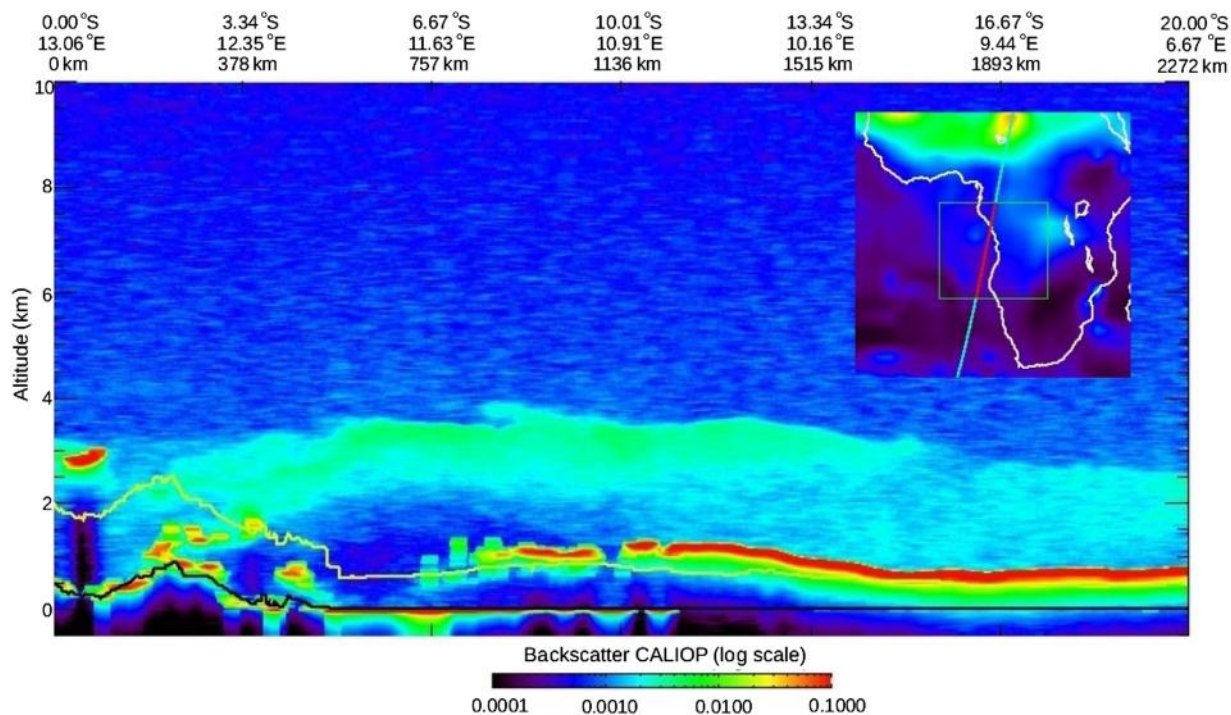


Figure 7 Vertical profile of backscatter signal ($\text{st}^{-1} \text{km}^{-1}$) at 532 nm from the CALIPSO LIDAR (named CALIOP) for 9 July 2009. The red line in the inset figure shows the location of the transect off the west coast of Africa. The black line shows ground level and the yellow line shows the height of the boundary layer (from another data source). On the vertical profile, blue colors are associated with weak backscattering by molecules in the clean upper atmosphere, light green colors show moderate backscattering by an elevated aerosol layer presumably from biomass burning, and red colors show strong backscattering by a layer of marine clouds. Reprinted from Costantino, L., Breon, F.M., 2010. Analysis of aerosol–cloud interaction from multi-sensor satellite observations. *Geophysical Research Letters* 37. <http://dx.doi.org/10.1029/2009GL041828>.

thermal radiation absorbed by the high altitude aerosol is reemitted at cold temperatures. Consequently the thermal energy that escapes to space is reduced, leading to an increase in the energy absorbed by the Earth system. Dust storms are capable of lifting large diameter dust particles to high altitudes. Recent research has shown that longwave DRF from dust storms can be comparable in magnitude to greenhouse gas forcing in certain regions. The vertical fidelity of CALIPSO sensors allow for effects of natural aerosols on thermal radiation (i.e., longwave DRF) to be better characterized.

Conclusions

Accurately quantifying aerosol forcing globally is one of the most challenging aspects of evaluating past climate and predicting future climate. While detailed monitoring of aerosols has occurred in certain regions (largely for characterizing air pollution), we are just beginning to have a more detailed understanding of aerosols globally. This global quantification has been one of our biggest challenges and is a key for both quantifying aerosol–radiation interactions and evaluating global models. The future holds new scientific challenges to be solved related to reducing uncertainties in global measurements of aerosol properties. A further practical challenge for the future is to ensure that the satellites that make global characterization of aerosols possible are sustained. Ultimately, our ability to characterize aerosol–radiation interactions, and consequently climate effects of aerosols, will hinge on our ability to globally characterize the properties discussed in this article.

See also: **Aerosols:** Aerosol Physics and Chemistry; Aerosol–Cloud Interactions and Their Radiative Forcing; Observations and Measurements. **Climate and Climate Change:** Volcanoes: Role in Climate. **Radiation Transfer in the Atmosphere:** Absorption and Thermal Emission; Radiation, Solar. **Satellites and Satellite Remote Sensing:** Aerosol Measurements; Research. **Solar System/Sun, Atmospheres, Evolution of Atmospheres:** Solar Terrestrial Interactions: Climate Impact. **Tropospheric Chemistry and Composition:** Aerosols/Particles.

Further Reading

- Bellouin, N., Jones, A., Haywood, J., Christopher, S., 2008. Updated estimate of aerosol direct radiative forcing from satellite observations and comparison against the Hadley Centre climate model. *Journal of Geophysical Research* 113. <http://dx.doi.org/10.1029/2007JD009385>.
- Charlson, R.J., Schwartz, S.E., Hales, J.M., et al., 1992. Climate forcing by anthropogenic aerosols. *Science* 255, 423–430.
- Charlson, R.J., Wigley, T.M.L., 1994. Sulfate aerosol and climate change. *Scientific American* 270, 48–57.
- Chin, M., et al., 2009. *Atmospheric Aerosol Properties and Climate Impacts. Synthesis and Assessment Product 2.3. Report by the U.S. Climate Change Science Program and the Subcommittee on Global Change Research.*
- Costantino, L., Breon, F.M., 2010. Analysis of aerosol–cloud interaction from multi-sensor satellite observations. *Geophysical Research Letters* 37. <http://dx.doi.org/10.1029/2009GL041828>.
- d'Almeida, G.A., Koepke, P., Shettle, E.P., 1991. *Atmospheric Aerosols: Global Climatology and Radiative Characteristics.* A. Deepak, Hampton, VA.
- Harshvardhan, A., 1993. Aerosol–climate interactions. In: Hobbs, P.V. (Ed.), *Aerosol–Cloud–Climate Interactions.* Academic Press, New York, pp. 75–95.
- Haywood, J.M., Boucher, O., 2000. Estimates of the direct and indirect radiative due to tropospheric aerosols: a review. *Reviews of Geophysics* 38, 513–543.
- Houghton, J.T., Ding, Y., Griggs, D.J., et al. (Eds.), 2001. *Climate Change 2001: The Scientific Basis.* IPCC/Cambridge University Press, Cambridge.
- Houghton, J.T., Meira Filho, L.G., Callander, B.A., et al. (Eds.), 1995. *Climate Change, 1995: The Science of Climate Change.* IPCC/Cambridge University Press, Cambridge.
- Huang, J., Fu, Q., Su, J., Tang, Q., Minnis, P., Hu, Y., Yi, Y., Zhao, Q., 2009. Taklimakan dust aerosol radiative heating derived from CALIPSO observations using the Fu-Liou radiation model with CERES constraints. *Atmospheric Chemistry and Physics* 9, 4011–4021.
- Kahn, R., 2012. Reducing the uncertainties in direct aerosol radiative forcing. *Surv. Geophys.* 33, 701–721. <http://dx.doi.org/10.1007/s10712-011-9153-z>.
- Kaufman, Y., Martins, J., Remer, L., Schoeber, M., Yamasoe, M., 2002. Satellite retrieval of aerosol absorption over the oceans using sunglint. *Geophysical Research Letters* 29. <http://dx.doi.org/10.1029/2002GL015403>.
- Kiehl, J.T., Briegleb, B.P., 1993. The relative roles of sulfate aerosols and greenhouse gases in climate forcing. *Science* 260, 311–314.
- Kondrat'ev, K.I., 1999. *Climatic Effects of Aerosols and Clouds.* Springer-Verlag, New York.
- McPherson, C., Reagan, J., Ferrare, R., Hostetler, C., Hair, J., 2009. Methods of analysis of atmospheric aerosols from future spaceborne high spectral resolution LIDAR data. In: *Geoscience and Remote Sensing Symposium. IEEE Conference Publications, Cape Town, pp. 485–488.* <http://dx.doi.org/10.1109/IGARSS.2009.5417626>.
- Minnis, P., Harrison, E.F., Stowe, L.L., et al., 1993. Radiative climate forcing by the Mount-Pinatubo eruption. *Science* 259, 1411–1415.
- Seinfeld, J.H., Pandis, S.N., 1998. *Atmospheric Chemistry and Physics: From Air Pollution to Climate Change.* Wiley, New York.
- Stier, P., Seinfeld, J., Kinne, S., Boucher, O., 2007. Aerosol absorption and radiative forcing. *Atmospheric Chemistry and Physics* 7, 5237–6261.

Role in Climate Change

N Bellouin, University of Reading, Reading, UK

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

The atmospheric aerosols emitted by human activities interact with radiation, clouds, and snow-covered surfaces, thus exerting a perturbation of the Earth's energy budget called radiative forcing (RF). Aerosol RF is strongly variable in time and space, but the net global effect is to offset the positive RF by long-lived greenhouse gases and counteract the associated global warming. In turn, changes in temperature and precipitation affect aerosols, giving rise to feedbacks that strengthen or oppose the initial climate change. Finally, aerosols can be used as climate engineering tools in efforts to mitigate climate change.

Introduction

Aerosols, the liquid and solid particles in suspension in the atmosphere, matter to the scientific study of climate change for three reasons. First, they are optically active species, interacting with solar and terrestrial radiation directly by scattering and absorption, and indirectly by modifying the microphysical properties of clouds and the radiative properties of surfaces covered by snow and ice. Perturbations to atmospheric distributions of aerosols therefore translate into perturbations of the Earth's radiative budget, which can enhance or counteract the radiative perturbations exerted by changes in the concentrations of long-lived greenhouse gases such as carbon dioxide and methane.

Second, human activities have profoundly perturbed aerosol distributions. Since the start of the Industrial Revolution in the eighteenth century, road transport, aviation, maritime shipping, industries, power plants, domestic cooking stoves, and forest-clearing fires have all emitted aerosols and their gaseous precursors, increasing the amount of particles in the atmosphere and modifying their chemical composition. Figure 1(a) shows a time series of globally averaged emissions of the main anthropogenic aerosol and precursor species since the year 1850. According to those estimates, anthropogenic emissions of sulfur dioxide, the gaseous precursor to sulfate aerosols, have

risen from a global total of 2 Tg for the year 1850 to more than 120 Tg per year at the end of the twentieth century. Over the same period, primary emissions of carbonaceous aerosols from fossil fuel combustion, and emissions of ammonia, the precursor of nitrate aerosols, increased by a factor 3. Increases in emissions increase aerosol mass over the source region and transport pathways, as illustrated by Figure 1(b), which shows changes in surface concentrations of sulfate aerosols over the industrial period. Increases do not cover the whole globe because strong sinks, dominated by washout by precipitation, prevent aerosols from accumulating in the atmosphere. The very heterogeneous distribution of aerosol concentrations and compositions in the atmosphere complicates the study of their climatic role compared to that of well-mixed and long-lived greenhouse gases such as carbon dioxide.

The third and last reason why aerosols play an important role in climate change is the dependence of their sources and sinks on the state of the climate. As climate responds to an initial radiative perturbation, the distributions of natural and anthropogenic aerosols and their gaseous precursors change, giving rise to radiative feedbacks that either oppose or reinforce the initial radiative perturbation. The Introduction section of this article introduces the three important concepts of radiative forcing (RF), fast adjustments, and climate response, which are relevant to the multiple roles of aerosols in climate change. Section

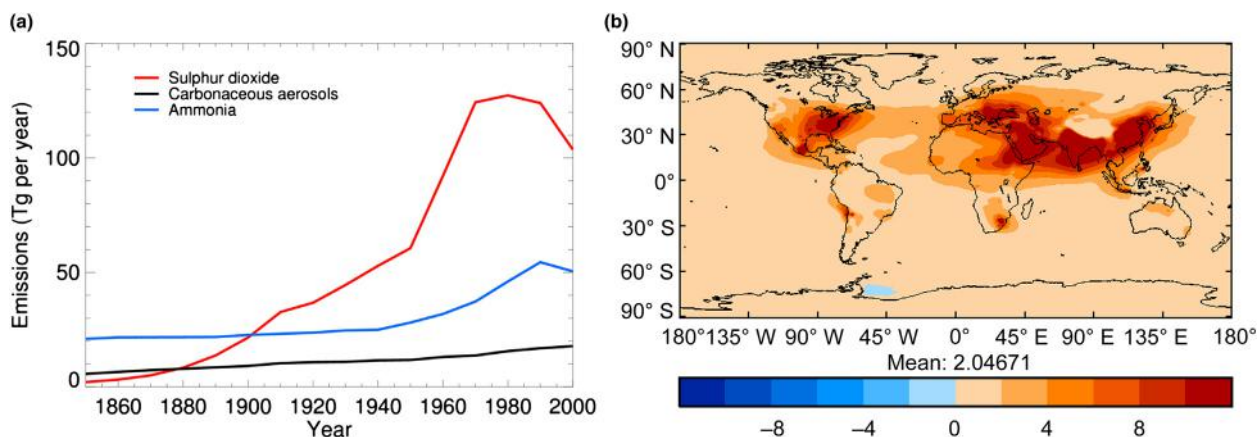


Figure 1 (a) Left panel: Time series of global, annual anthropogenic emissions of sulfur dioxide, ammonia, and primary carbonaceous aerosols over the period 1850–2000 according to the CMIP5 data set. (b) Right panel: Modeled changes in column-integrated concentrations of sulfate aerosols, in mg[S] m^{-2} , due to changes in anthropogenic emissions of sulfur dioxide over the period 1850–2000.

Instantaneous and Effective RF, and Climate Response then reviews the mechanisms through which aerosols perturb the Earth's RF, while Section Mechanisms of Aerosol RF discusses how aerosols are in turn influenced by climate change. Finally, the proposed use of aerosols as tools for climate mitigation is discussed in Section Climate Response and Aerosol Feedbacks.

Instantaneous and Effective RF, and Climate Response

RF quantifies, in units of W m^{-2} , the change in the energy available to the climate system caused by a change in one or more components of the Earth's radiative budget. The concept of RF finds its usefulness from climate modeling studies that found that the long-term change in globally averaged surface temperature is proportional, within 20%, to the RF initially applied. More formally, RF is defined as the instantaneous perturbation in net radiative flux at the tropopause exerted by a change in a component of the radiative budget, with surface temperature and tropospheric state maintained in their unperturbed state (Figure 2). By convention, negative RFs denote a decrease in net radiative flux, or a loss of energy from the climate system. Conversely, positive RFs denote a gain of energy to the climate system. In climate research, RF is typically defined with respect to the year 1750, taken as representative of unperturbed conditions because that year predates the Industrial Revolution. It should be noted however that in terms of aerosols, the year 1750 already experienced anthropogenic perturbations from the stoves used in domestic cooking and the forest and field clearing fires that were common agriculture practices at the time. Those emissions are however taken as being part of the preindustrial (PI) climate system.

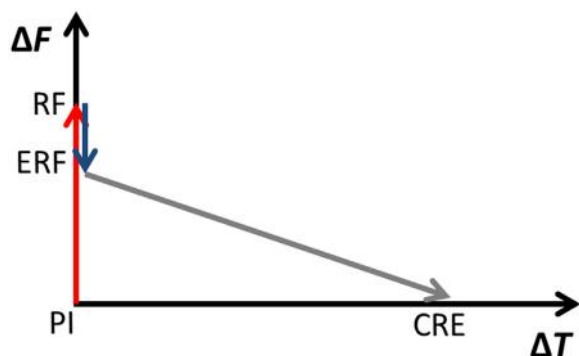


Figure 2 Diagram showing the relationship between radiative perturbation, noted ΔF , and change in globally averaged surface temperature, noted ΔT , to illustrate the concepts of instantaneous RF, effective radiative forcing (ERF), and climate response. Instantaneous RF is the initial radiative imbalance, chosen here positive, with respect to a reference state, typically PI conditions. ERF is the sum of RF and the subsequent fast atmospheric adjustments that do not affect globally averaged surface temperature. The climate system then returns to the reference state (PI) through a change in globally averaged surface temperature. To a good approximation, the surface temperature change is proportional to the ERF, with the constant of proportionality called climate sensitivity. The new surface temperature at equilibrium is called climate response at equilibrium.

The climate system results from the coupling of the atmosphere and the ocean. Because the ocean represents 90% of the heat capacity of the climate system, its response to an instantaneous RF is much slower, on the order of decades, than that of the atmosphere, which adjusts within weeks. It is therefore useful to distinguish two timescales in the response to a RF: fast adjustments and a slow climate response. Fast adjustments (Figure 2) directly depend on the RF and operate on atmospheric timescales, affecting the vertical distribution of temperature and moisture, thus the formation of clouds. ERF is the sum of RF and its fast adjustments. ERF is weaker than RF when fast adjustments counteract the initial forcing, as represented in Figure 2. Conversely, ERF is stronger than RF when fast adjustments enhance the initial forcing. Fast adjustments are particularly important for carbonaceous aerosols, and ERF is then a better predictor of the subsequent long-term change in surface temperature than instantaneous RF. The slow climate response (Figure 2) depends on changes in globally averaged surface temperatures and therefore operates on oceanic timescales. To return to radiative balance, the globally averaged surface temperature increases if the ERF is positive or decreases if the ERF is negative. This change of temperature is determined by the sensitivity of the climate system, which is a property of the atmosphere–ocean system and is, to a good first approximation, independent of the RF mechanism.

Mechanisms of Aerosol RF

Anthropogenic aerosols exert a RF through three mechanisms, represented schematically in Figure 3: aerosol–radiation interactions, aerosol–cloud interactions, and aerosol–surface interactions. Those three categories were introduced by the Fifth Assessment Report of the Intergovernmental Panel on Climate Change to unify the terminology of aerosol RF.

Aerosol–Radiation Interactions

The radiative forcing due to aerosol–radiation interactions (RFari), also termed aerosol direct RF in the scientific literature, stems from the additional scattering and absorption of radiation that follows the increase in aerosol mass provided by anthropogenic emissions of aerosols and gaseous precursors (Figure 3(a)). Anthropogenic aerosols originate in majority from gas-to-particle conversion processes and from primary emissions during combustion of fossil fuels or biomass. They are therefore mostly composed of particles with sizes smaller than $1 \mu\text{m}$ and consequently interact more effectively with wavelengths in the solar spectrum, with much weaker influences on terrestrial radiation. RFari depends on the horizontal and vertical distributions of aerosol concentrations and scattering and absorption properties, which in turn depend on aerosol sizes and chemical composition. Environmental factors, such as the solar zenith angle, and the reflectance of the surface or cloud underlying the aerosol layer, also play important roles. One crucial aspect of RFari is that it can be either negative or positive at the top of atmosphere: its sign depends on the competition between aerosol scattering and absorption processes, which is affected by the reflectance of the surface below the aerosol layer. For a given aerosol size, the

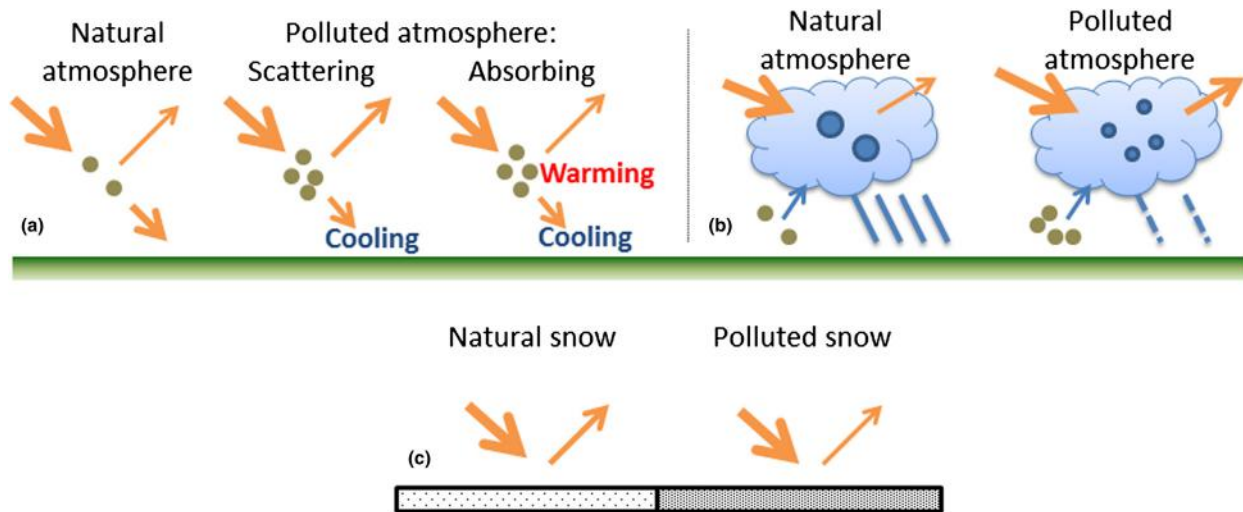


Figure 3 Mechanisms of aerosol RF and fast adjustments. (a) Compared to a natural atmosphere, anthropogenic aerosols (brown circles) scatter and absorb additional solar radiation (orange arrows), reflecting energy back to space or absorbing it in this atmosphere, cooling the surface locally, and warming the aerosol layer for absorption aerosols. The new vertical temperature gradient then affects atmospheric stability and cloud formation. (b) Compared to a natural atmosphere, anthropogenic aerosols provide additional cloud condensation nuclei, leading to the formation of more, smaller cloud droplets. More droplets reflect more solar radiation back to space. Smaller droplets may not easily form large raindrops, potentially decreasing precipitation and increasing cloud lifetime. (c) Deposition of anthropogenic aerosols onto ice and snow surfaces decreases their albedo.

likelihood of a positive top-of-atmosphere RFari increases with aerosol absorption and reflectivity of the surface. A purely scattering anthropogenic aerosol, such as sulfate, exerts a negative RFari over most regions. The more absorbing carbonaceous aerosols exert a negative RFari over dark surfaces, such as vegetated and oceanic surfaces, and a positive RFari over brighter surfaces, such as deserts and icy regions. This point is well illustrated by the satellite imagery shown in **Figure 4**, where a plume of moderately absorbing carbonaceous aerosols emitted by forest fires in Portugal in August 2003 is clearly visible, which in itself is evidence for the ability of aerosols to perturb the radiative signal reaching the space-borne instrument. Over land and cloud-free oceans, the aerosol plume is brighter than the underlying surface, which indicates that aerosols increase the outgoing solar radiative flux at the top of the atmosphere, causing a negative RFari. In contrast, the northwestern section of the aerosol plume overlies bright clouds and appears darker, which indicates that aerosols decrease the outgoing solar radiative flux at the top of the atmosphere, causing a positive RFari.

Approximate formulas to quantify RFari in cloud-free sky have been derived, such as the following, which is valid for both scattering and absorbing aerosols:

$$\text{RFari} \approx ST^2 \varpi_0 \beta \Delta\tau [(1 - R_s)^2 - 2R_s(1 - \varpi_0)] / (\beta \varpi_0) \quad [1]$$

where S is the solar constant, in W m^{-2} , T is the dimensionless transmittance of the atmosphere above the aerosol layer, and R_s is the dimensionless reflectance of the surface. Aerosols are characterized by three dimensionless parameters:

- the change in optical thickness, $\Delta\tau$. The optical thickness, τ , measures the strength of aerosols scattering and absorption of radiation at a given wavelength and integrated over the atmospheric column;

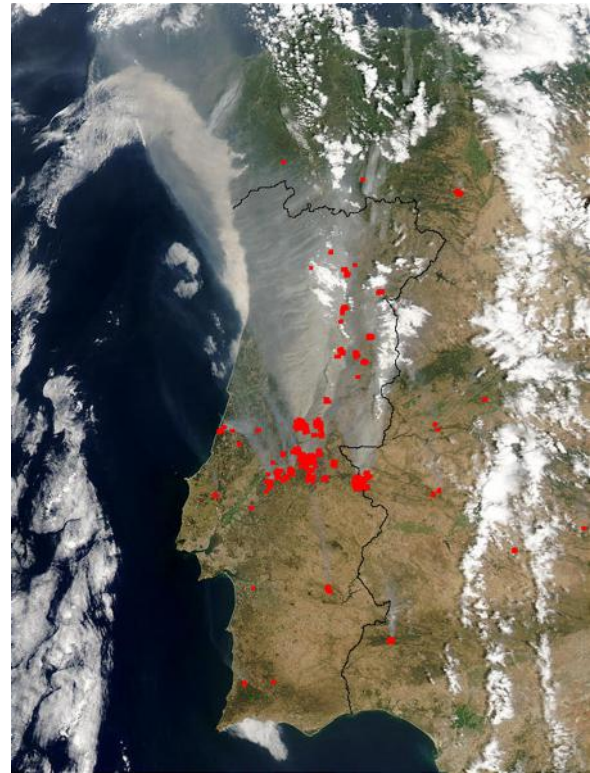


Figure 4 Forest fires over Portugal on 3 August 2003, observed by the Moderate Resolution Imaging Spectroradiometer satellite instrument aboard NASA's Terra platform. Red dots mark the location of individual fires.

- their single-scattering albedo, ω_0 , which is the fraction of the optical thickness due to scattering processes, the remainder being due to absorption; and
- their upscatter fraction, β , which quantifies the fraction of radiation that is scattered upward with respect to the particle's horizontal plan.

Equation [1] highlights the fact that the sign of RFari depends on the aerosol absorption properties and the reflectance of the surface, as discussed above. It also shows that for a given set of optical properties, RFari depends linearly on the change in aerosol optical thickness or, equivalently, aerosol concentrations. The use of such simplified expressions has now been replaced by more sophisticated calculations based on radiative transfer numerical models, which are able to fully represent the strong dependences of RFari with the wavelength of the radiation and the solar zenith angle. Table 1 gives the globally and annually averaged estimate of RFari, obtained from global numerical models with representations of the life cycle and optical properties of aerosols. RFari is estimated at -0.4 W m^{-2} , with an uncertainty range of $\pm 0.5 \text{ W m}^{-2}$. Note the opposite signs of RFari exerted by scattering aerosols, such as sulfate, and absorbing aerosols, such as black carbon.

RFari triggers fast adjustments, which have historically been termed semidirect effects. The absorption of solar radiation by carbonaceous aerosols warms the aerosol layer while reducing the amount of downward radiation, cooling the surface. The combination of cooling at the surface and warming aloft modifies the stability of the atmosphere and the vertical profile of moisture. Cloud formation is in turn modified, which perturbs solar and terrestrial radiative fluxes, giving rise to the effective radiative forcing due to aerosol–radiation interactions (ERFari). The strength of ERFari compared to that of RFari depends in a complex way on the relative position of absorbing aerosols and clouds, and on the cloud regime. When absorbing aerosols overlie clouds, they stabilize the atmosphere and suppress convection, favoring stratocumulus cloud formation but hindering cumulus cloud formation. Absorbing aerosols located within low clouds reduce relative humidity and increase evaporation, reducing low cloud cover. Finally, when absorbing aerosols underlie clouds, they destabilize the atmosphere, promoting formation of convective clouds. Fast adjustments due to aerosol–radiation interactions are difficult to observe because aerosols and clouds are naturally correlated: aerosols are involved in cloud formation, clouds act as aerosol

sources via aqueous oxidation of aerosol precursors, and both aerosols and clouds are transported within the same air masses, thus correlating with meteorology. Estimates from global numerical models remain uncertain because of difficulties in representing convective processes at coarse resolution. Table 1 gives the globally and annually averaged estimate of ERFari, which at $-0.5 \pm 0.5 \text{ W m}^{-2}$ suggests that fast adjustments strengthen RFari on a global average.

Anthropogenic aerosols are not the only aerosol species that exert a RF by interacting with radiation under the definition of RF given in the Introduction section: because of their sporadic nature, aerosols emitted by volcanic eruptions are also considered external to the natural climate system. Large eruptions, such as those of Mount El Chichon in 1982 and Mount Pinatubo in 1991, inject sulfur dioxide directly into the stratosphere, where it is oxidized into sulfate aerosols. Aerosol residence time in the stratosphere is around 2 years, much longer than in the troposphere because of the absence of precipitation and slow air exchanges through the tropopause. Mount Pinatubo injected 20 Tg of sulfur dioxide, increasing stratospheric aerosol optical depths at $0.55 \mu\text{m}$ from their background level of less than 0.01 to more than 0.1. The resulting RFari in the solar spectrum has been estimated between -4 and -5 W m^{-2} on a global average. However, stratospheric aerosols also exert an RFari in the terrestrial spectrum, which is positive because they radiate energy at temperatures that are colder than that of the surface. For Pinatubo, the net RFari has therefore been closer to -2 W m^{-2} .

Aerosol–Cloud Interactions

The radiative forcing due to aerosol–cloud interactions (RFaci), termed alternatively aerosol first indirect RF, cloud albedo forcing, or Twomey forcing in the scientific literature, arises from the role aerosols play in the hydrological cycle as cloud condensation nuclei. The small sizes and soluble nature of most anthropogenic aerosols, especially sulfate, nitrate, and organic carbon aerosols, make them good nuclei upon which liquid cloud droplets form, although they are not good nuclei for ice crystal formation. In contrast to RFari, which is exerted by an increase in aerosol mass, RFaci is exerted by an increase in the number of aerosols that can act as cloud condensation nuclei. The distribution of a fixed amount of cloud water over those more numerous nuclei results in more, smaller cloud

Table 1 Globally and annually averaged best estimate and range of anthropogenic aerosol instantaneous RF (W m^{-2}) and ERF (W m^{-2}) over the period 1750–2000 due to aerosol–radiation interactions (ari), aerosol–cloud interactions (aci), aerosol on snow, and contrail formation

Anthropogenic aerosol species	RFari	ERFari	ERFari + aci	Aerosols on snow	Contrails
Total	-0.35 ± 0.5	-0.55 ± 0.5	$-0.9 (-1.9 \text{ to } -0.1)$	$+0.04 (0 \text{ to } +0.09)$	$+0.05 (+0.02 \text{ to } +0.15)$
Sulfate	$-0.34 (-0.61 \text{ to } -0.13)$				
Black carbon	$+0.40 (+0.05 \text{ to } +0.8)$				
Organic carbon	$-0.09 (-0.16 \text{ to } -0.03)$				
Biomass burning	$-0.00 (-0.2 \text{ to } +0.2)$				
Secondary organic aerosol	$-0.03 (-0.27 \text{ to } -0.02)$				
Nitrate	$-0.11 (-0.3 \text{ to } -0.03)$				

Standard deviations and ranges indicate the 95% confidence level.

Estimates are taken from the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, and are based on global modeling.

droplets (Figure 3(b)). More cloud droplets have a larger surface area, which increases the albedo of the cloud. Brighter clouds increase the outgoing solar radiation at the top of the atmosphere, exerting a negative RFaci. RFaci is exerted in the solar spectrum only, since by definition of the RF cloud albedo is the only factor that changes. An approximate formula for the computation of RFaci is:

$$\text{RFaci} \approx -S \cdot f \cdot \delta\alpha / \delta N_d \cdot \delta N_d / \delta N_a \cdot dN_a \quad [2]$$

where S is the solar constant, f is the fractional cloud cover, dN_a is the change in cloud condensation nuclei due to anthropogenic activities, $\delta N_d / \delta N_a$ is the sensitivity of cloud droplet number, N_d , to a change in cloud condensation nuclei, and $\delta\alpha / \delta N_d$ is the susceptibility of cloud albedo, α , to a change in cloud droplet number. RFaci is not linear with the change in aerosol number, and saturates at large aerosol number concentrations: a given change in aerosol number therefore exerts a stronger RFaci in pristine regions than in polluted regions. Numerous field studies have reported a dependence of cloud-base droplet size distribution with aerosol size distribution, demonstrating that the physical mechanisms behind RFaci occur in the atmosphere (Figure 5). However, the cloud liquid water content and geometric thickness never remain fixed in reality, making RFaci an abstract concept. Modeling studies have also challenged the assumption that anthropogenic emissions can only lead to an increase in the number of aerosols and cloud condensation nuclei. They propose that gas-phase precursors condense onto the additional aerosol particles provided by anthropogenic sources instead of creating new aerosols via nucleation. If confirmed, this mechanism opens the way for a decrease in cloud condensation nuclei in response to anthropogenic emissions, at least in regions where the number of particle available for condensation is low. RFaci would then be positive. The lack of observational constraints and the difficulty to adequately represent in numerical models the multiple scales involved in aerosol–cloud interactions make RFaci very uncertain. Because of the abstract nature of

RFaci and the difficulty of separating it from its fast adjustments, discussed below, recent publications do not estimate RFaci by itself, but include its fast adjustments.

Aerosol–cloud interactions in ice clouds are poorly known, both at the process level and because the characteristics and distribution of aerosols that make good ice nuclei remain unclear. However, emissions by aviation yield a particular kind of interactions with ice clouds, which deserve a specific mention. The aerosol and moisture emitted into the atmosphere by aircraft engines increase cloudiness by forming cirrus-like linear clouds called contrails. Contrail formation occurs when aerosols are emitted into an atmosphere, which is supersaturated with respect to ice. Depending on conditions, contrails evaporate or evolve into cirrus clouds, undistinguishable from naturally occurring cirrus, which complicates observational estimates of effective radiative forcing due to aerosol–cloud interactions (ERFaci) by contrail formation. Because high-level clouds are thin but radiate at a lower temperature than the surface, the contrail ERF is dominated by its terrestrial spectrum component, which is positive. Climate modeling studies estimate the contrail ERFaci at less than $+0.1 \text{ W m}^{-2}$ on a global, annual average. This weak value hides stronger regional forcings over the busiest aircraft corridors, especially the North Atlantic, where local temperatures may be affected.

The ERFaci is the sum of RFaci and its fast adjustments. Adjustments in liquid clouds posit that smaller cloud droplets take more time to reach the sizes required for the conversion of cloud water into rainwater, a process called autoconversion. A smaller autoconversion rate would then increase the lifetime of clouds and/or their vertical development. Observational evidence for fast adjustments is mixed, with observations showing either a reduction or enhancement of precipitation in regions with high anthropogenic aerosol loading, depending on cloud regime. Aerosol-driven fast adjustments are unlikely to influence those clouds whose lifetime is not regulated by precipitation, such as nonprecipitating stratocumulus clouds. At the other end of the precipitation spectrum, clouds where accretion of raindrops dominates over autoconversion are also not likely to be influenced by aerosol-driven changes in cloud droplet size distribution. Aerosol fast adjustments would then be limited to specific cloud regimes, or to situations where increases in cloud condensation nuclei forces a change in cloud regime, such as the transition from open- to closed-cells cumulus clouds. Because of natural correlations between aerosols and clouds, fast adjustments to aerosol–cloud interactions are difficult to observe unequivocally. In numerical models, estimates of ERFaci suffer from the difficulty in representing the cloud-scale processes involved. ERFaci is therefore very uncertain, with estimates ranging from -1.2 to 0 W m^{-2} .

Aerosol–Surface Interactions

Aerosols are removed from the atmosphere by turbulence and precipitation, and deposit onto the surface. In most cases, this deposition does not change the radiative properties of the surface. But the situation with snow- or ice-covered surfaces is different, as absorbing aerosols embedded into the snowpack reduce its albedo by up to 0.1 for large aerosol

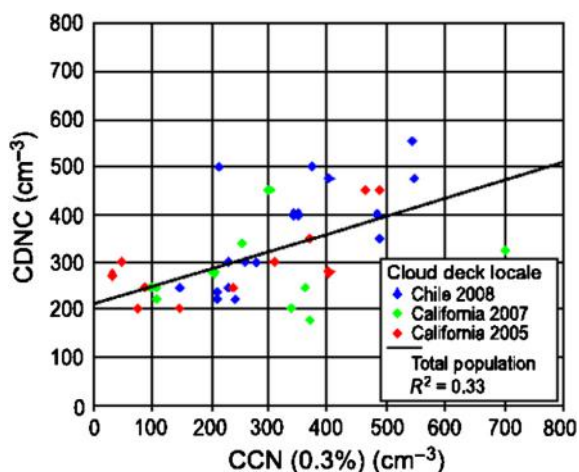


Figure 5 Cloud droplet number concentrations (CDNC, in cm^{-3}) as a function of cloud condensation nuclei (CCN, in cm^{-3}) from observations of stratocumulus clouds off the coasts of California and Chile. Figure 2 of Hegg et al., 2012. Atmospheric Chemistry and Physics.

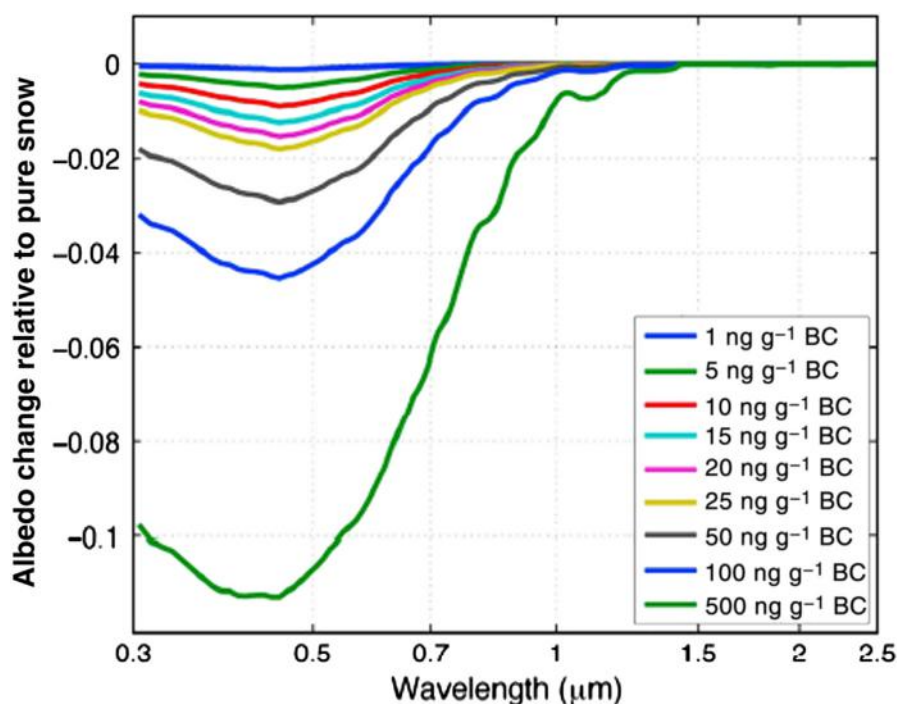


Figure 6 The reduction in wavelength-dependent snow albedo caused by increasing concentrations of black carbon aerosols in snow, for a snowpack with effective grain size of 200 μm . Figure 27, Bond, T.C., Doherty S.J., Fahey D.W. et al., 2013. Bounding the role of black carbon in the climate system: a scientific assessment. *Journal of Geophysical Research* 118, 5380–5552. doi:10.1002/jgrd.50171.

concentrations (Figure 6). Anthropogenic carbonaceous aerosols, and specifically black carbon, therefore exert a positive RF by darkening snow. This RF triggers amplifying fast adjustments: the warmer snowpack leads to increases in the mean grain size of snow, further decreasing surface albedo, while snow melts earlier in the spring, exposing the darker surface under the snow, again decreasing surface albedo. The ERF of aerosol–snow interactions is estimated between 0 and $+0.1 \text{ W m}^{-2}$ on a global, annual average for anthropogenic black carbon aerosols (Table 1). This weak global value hides large regional variations, with estimates of snow albedo RF being strong in the cold regions that are under sizeable influences from transported anthropogenic aerosols, most notably the Arctic and the Himalaya.

Another RF mechanism due to aerosol–surface interactions, albeit very indirect and poorly quantified, stems from competing aerosol radiative effects on vegetation productivity. On the one hand, aerosols decrease the amount of photosynthetically active radiation (a subset of the solar spectrum, ranging from 0.4 to 0.7 μm), which is used as a source of energy by plants. Aerosols are therefore detrimental to plant productivity. On the other hand, aerosol scattering of radiation makes solar radiation reaching the surface more diffuse, improving the distribution of sunlight between leaves under direct insolation and those in the shade. This effect favors vegetation growth. A modeling study of the carbon cycle obtained, by including those two effects, a 25% enhancement in vegetation productivity because of anthropogenic changes in aerosol concentrations over the period 1960–99. The associated increase in the land-based sink of carbon dioxide is however small, so the resulting RF is likely to be weak.

Total Aerosol RF

Estimates of the total aerosol RF and fast adjustment range from -1.9 to -0.1 W m^{-2} (Table 1) and a modeled distribution of aerosol RF is shown in Figure 7. The large uncertainty range highlights our lack of knowledge of present distributions of aerosol amounts and optical properties, and our limited understanding of aerosol RF and adjustment mechanisms, especially for aerosol–cloud interactions. PI aerosol distributions are also poorly constrained, affecting the reference state against which aerosol RF is defined. Nevertheless, it appears

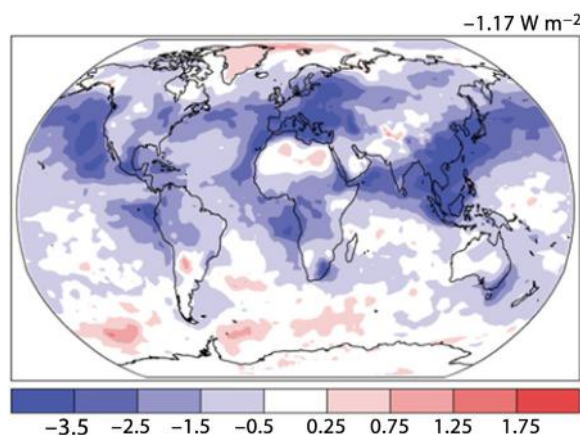


Figure 7 Distribution of aerosol ERF, in W m^{-2} , obtained from numerical simulations based on changes in aerosol emissions over the industrial area. Figure 15 of Shindell et al., 2013. *Atmospheric Chemistry and Physics*.

that aerosols compensate a potentially large fraction of the RF exerted by increases in long-lived greenhouse gases and tropospheric ozone, which is estimated at $+2.6 \text{ W m}^{-2}$. How large that fraction really is has implications for estimating climate sensitivity to a doubling of carbon dioxide concentrations, which is an important property of the climate system. Because an increase in globally averaged surface temperatures is observed over the twentieth century, it follows that greenhouse gas RF is stronger, in absolute value, than aerosol forcing. Climate would need to be very sensitive to carbon dioxide increases to overcome a strong aerosol RF. Conversely, low climate sensitivity would be sufficient to overcome a weak aerosol forcing. Unfortunately, the large uncertainties in aerosol RF limit the usefulness of this method of estimating climate sensitivity.

Climate Response and Aerosol Feedbacks

Climate Response

The RF of aerosols and long-lived greenhouse gases triggers a change in the globally averaged surface temperature, which theoretically continues until a new radiative balance is reached. Although the temperature response of the climate system depends on the climate sensitivity, which is a property of the system, there is a small dependence on the forcing mechanism, because of the existence of fast adjustments. It is therefore useful to take the ratio between globally averaged surface temperature at equilibrium and RF, and compare that value to the corresponding ratio for a doubling of carbon dioxide concentrations. The resulting quantity is called efficacy:

$$\text{Efficacy} = (\Delta T_i / \Delta F_i) / (\Delta T_{2xCO_2} / \Delta F_{2xCO_2}) \quad [3]$$

Climate model estimates of efficacies for aerosol–radiation, aerosol–cloud, and aerosol–surface interactions are given in [Table 2](#) for selected aerosol species. Efficacies for absorbing aerosols are smaller than one because fast adjustments counteract the RF. In contrast, the RF due to darkening of snow from aerosol deposition has a large efficacy, which reflects how the location of the snow darkening RF, in the snowpack itself, promotes snowmelt compared to carbon dioxide, whose forcing is located, in effect, at the tropopause. The resulting surface albedo feedback, whereby the darker surfaces exposed by snowmelt further increase surface temperatures, is then stronger. Efficacies are useful on a global average, but hide differences in the geographical distribution of surface temperature changes ([Figure 8](#)). The positive temperature response to the RF by carbon dioxide exhibits maxima at high latitudes, because of increased poleward transport of heat and the surface albedo feedback from melting snow and ice. The pattern of the

negative temperature response to the RF by aerosols is different, being predominantly located over and downwind the anthropogenic aerosol source regions. Consequently, the temperature response to aerosol forcing is mainly located in the Northern Hemisphere. The different patterns of temperature response to long-lived greenhouse gases and aerosols are useful in detection and attribution studies.

Changes in surface temperatures are not the only manifestation of the climate response to anthropogenic forcing: precipitation and atmospheric dynamics also respond. The precipitation response is driven by the redistribution of energy within the atmosphere and at the surface. The existence of fast adjustments will therefore affect the precipitation response, and strongly absorbing aerosols such as black carbon are unique in their ability to decrease precipitation in spite of increasing surface temperatures. Dynamical responses are caused by changes in net radiative fluxes reaching the surface, and the gradient in RF between the two hemispheres. Anthropogenic aerosols are primarily located in the Northern Hemisphere, where they exert their radiative cooling. The Southern Hemisphere becomes relatively warmer, and modeling studies have repeatedly shown that this temperature gradient perturbs the Hadley circulation and causes a southward shift in the position of the intertropical convergence zone (ITCZ) and its band of strong precipitation. This mechanism may then have been a contributing factor to the severe droughts that affected the Sahel region of Africa in the 1980s. Studies using numerical models of the atmosphere–ocean system have also suggested links between the RF of aerosols and changes in dynamical features as diverse as the Indian monsoon or the multidecadal oscillation of North Atlantic sea surface temperatures.

Large volcanic eruptions have been useful opportunities to study the climate response to an aerosol RF. After the eruption of Mount Pinatubo, surface temperatures have decreased by 0.4 K globally for 2 years, which compares to a global increase of 0.6 K over the industrial period. Global reductions in precipitation have also been observed. In addition, the eruption of Mount El Chichon provided important information on the response of the ITCZ, because it preferentially loaded the Northern Hemisphere stratosphere, causing an interhemispheric RF and cooling gradient that shifted the ITCZ southward. The associated changes in rainfall may have exacerbated the Sahel drought of 1982 and 1983, as observed by monitoring stations and modeled numerically ([Figure 9](#)).

Aerosol Feedbacks

Natural and anthropogenic aerosol distributions are perturbed by the response of the atmosphere and ocean to anthropogenic RF, and this perturbation translates into further changes in the Earth's radiative budget, called aerosol radiative feedbacks. Feedbacks that exacerbate the initial RF are called positive feedbacks, while those that oppose the forcing are called negative feedbacks. For anthropogenic aerosols, feedbacks primarily arise from changes in wet removal by precipitation. The sign of those feedbacks remains unclear. Because precipitation is expected to increase in a warmer climate, one would anticipate an increase in wet removal and a decrease in the residence time of aerosols in the atmosphere. They would

Table 2 Efficacy of aerosol RF for selected species

<i>Aerosol species</i>	<i>Efficacy</i>
Sulfate (RFari + aci)	~1
Biomass burning (RFari + aci)	0.8–0.9
Black carbon (RFari)	<0.8
Black carbon (deposition on snow)	>2

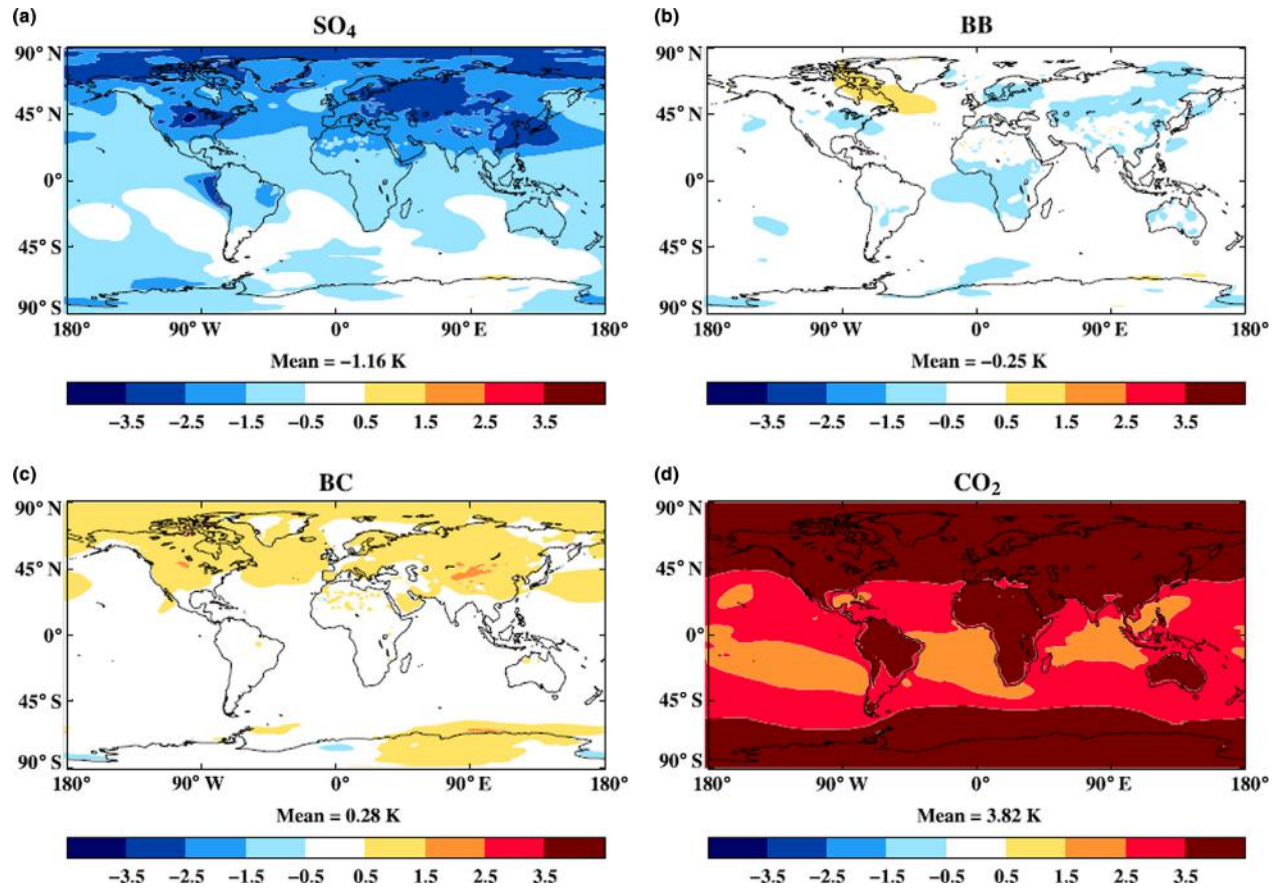


Figure 8 Modeled patterns of surface temperature responses, in K, to PI increases in (a) sulfate, (b) biomass burning, and (c) black carbon aerosols, compared to (d) a doubling of CO₂ concentrations. Figure 2 of Jones et al., 2007. *Journal of Geophysical Research*.

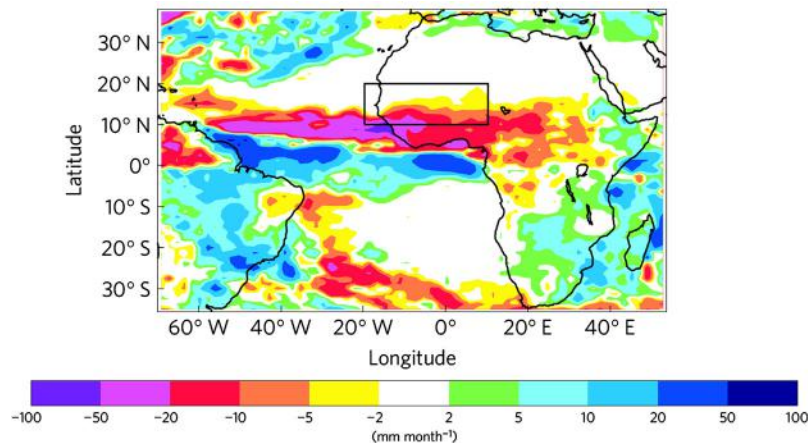


Figure 9 Precipitation response to the El Chichon volcanic eruption, as modeled by the HadGEM2 climate model for June–October 1982. Figure 1 of Haywood et al., 2013. *Nature Climate Change*.

therefore have less time to interact with radiation and clouds. However, models where low-cloud cover is suppressed by global warming suggest a feedback of opposite sign, because aerosols are then less likely to interact with clouds and be removed by precipitation. In addition, increase in convective

activity would loft aerosols higher up in the atmosphere, contributing to an increase in residence time in the atmosphere. Current climate models simulate weak feedbacks driven by changes in residence time, in the order of $\pm 0.2 \text{ W m}^{-2}$ per unit change in globally averaged surface temperature.

Feedbacks involving natural aerosols are potentially stronger, and their mechanisms are:

- Changes in vegetation cover: Fertilization by increased atmospheric concentrations of carbon dioxide is expected to increase vegetation productivity. Temperature-driven increases in occurrences of wild fires would have the opposite effect. Because vegetation emits biogenic volatile organic compounds that oxidize into secondary aerosols, changes in vegetation cover would translate into changes in natural aerosol levels and their interactions with radiation, clouds, and ultimately vegetation productivity, closing the feedback loop. In addition, the import of nutrients, such as phosphorus from transported mineral dust aerosols, is important to rain forest ecosystems, and will change in the future as the sources and sinks of those aerosols respond to climate change. The sign of net aerosol-vegetation feedbacks is currently unknown.
- Changes in ocean biogeochemical activity: Ocean phytoplankton is a strong source of dimethylsulfide (DMS), a gas that is partly oxidized into sulfur dioxide, itself a gaseous precursor of sulfate aerosols. Global warming of sea-surface temperatures has competing effects on DMS production. On the one hand, it reduces ocean nutrients and decreases phytoplankton productivity and DMS emissions. On the other hand, it reduces the depth of the ocean mixed layer, increasing seawater DMS concentrations and DMS emissions into the atmosphere. Climate models simulate a net increase in sulfate aerosols produced by ocean-based DMS emissions with increased temperatures. The aerosol-radiation and aerosol-cloud interactions exerted by those additional sulfate aerosols would be a negative feedback, termed CLAW feedback after the initials of the scientists who proposed the feedback loop, Robert Charlson, James Lovelock, Meinrat Andreae, and Stephen Warren. The numerical models that include the CLAW mechanisms, however, suggest a feedback weaker than initially thought, because of the presence of multiple bottlenecks between increased DMS emissions and increased aerosol radiative effects.
- Changes in sea-ice cover: Increased surface temperatures lead to a retreat of sea-ice, which opens the ocean for emission of DMS precursors and primary sea-salt aerosols. Those aerosols then interact with radiation and clouds to produce a negative feedback.
- Changes in bare-soil cover: Desertification leads to increased emissions of mineral dust aerosols, although the magnitude of this response also depends on changes in near-surface wind speed. Such increases would exert a radiative feedback because mineral dust aerosols interact with solar and terrestrial radiation, and are good ice-cloud nuclei. Mineral dust aerosols also supply nutrients to vegetation and oceanic phytoplankton, and changes in mineral dust deposition to those ecosystems would affect their productivity.

Current estimates of those natural aerosol feedbacks suffer from our lack of knowledge of present-day natural aerosol sources, and the limited representation in numerical models of the multiple mechanisms involved. Consequently, uncertainties are large, but natural aerosol feedbacks are potentially strong.

Aerosol-Based Mitigation of Climate Change

Anthropogenic aerosols provide means of mitigating climate change. Since scattering aerosols exert a negative RF, maintaining the associated anthropogenic emissions at their present levels offsets the positive RF by long-lived greenhouse gases, and slows down the rate of increase in globally averaged surface temperatures. Unfortunately, the same anthropogenic emissions decrease air quality and are detrimental to human health, which is strong argument for decreasing emissions further, to the expense of a reduced aerosol-driven climate cooling. Indeed, most scenarios of future changes in anthropogenic aerosol emissions assume strong decreases because of the enforcement of increasingly ambitious air quality standards. From this point of view, black carbon aerosols become attractive targets for climate mitigation: decreasing their emissions, in particular from combustion of fossil fuels, would mitigate climate change, because they exert a positive RF, and benefit air quality, because they are pollutants. The short residence time of black carbon aerosols in the atmosphere also means that those benefits would manifest quickly. The situation is, however, complicated by the fact that fossil fuel combustion also emits sulfate and organic carbon aerosols, which have a cooling effect on climate. Accounting for those co-emitted species severely decreases the net RF of black carbon, from a strong $+1.1 \text{ W m}^{-2}$ to a weaker $+0.2 \text{ W m}^{-2}$ according to a scientific review.

Aerosols are also involved in climate engineering proposals, where the radiative budget of the Earth is deliberately altered to counteract the positive RF by long-lived greenhouse gases. One such proposal, publicized by Nobel Prize winner Paul Crutzen in 2006, is to replicate large volcanic eruptions by injecting sulfur dioxide in the stratosphere. The eruption of Mount Pinatubo has provided a useful analogue and showed that such an injection would be able to reduce global surface temperatures. Another proposal relies on aerosol-cloud interactions by injecting sea-spray aerosols into stratocumulus decks in the hope of making those clouds brighter. Aerosol-based climate engineering proposals have the advantage to be less costly than adapting to climate change, and can be calibrated to offset positive RFs of varying strengths by varying the amount and size of the aerosols emitted. They are however imperfect, because they do not address nonradiative consequences of increases in carbon dioxide concentrations, such as ocean acidification. Climate engineering techniques are also known to trigger unwanted regional changes, which are potentially worse than nonengineered climate change. Those unwanted changes result from the difficulty of perfectly offsetting the greenhouse gas RF in terms of geographical, vertical, spectral, and temporal distributions. Changes in precipitation patterns are especially problematic, although aerosol climate engineering could be calibrated to minimize precipitation changes by allowing a small amount of global warming. In the event that carbon dioxide emissions cannot be reduced, humanity would enter into a long commitment to climate engineering, because climate modeling studies have shown that switching off climate engineering leads to a rapid warming of surface temperatures, which are back to

nonengineered levels within a decade. Specific techniques also have their own drawbacks. For example, it remains unclear how stratospheric ozone would respond to long-term injections of sulfur dioxide, and what the effects of a large-volcanic eruption on an engineered stratosphere would be.

See also: **Aerosols:** Aerosol–Cloud Interactions and Their Radiative Forcing; Climatology of Stratospheric Aerosols; Climatology of Tropospheric Aerosols; Role in Radiative Transfer. **Biogeochemical Cycles:** Sulfur Cycle. **Chemistry of the Atmosphere:** Observations for Chemistry (*In Situ*): Particles. **Climate and Climate Change:** Volcanoes: Role in Climate. **Clouds and Fog:** Cloud Microphysics. **Satellites and Satellite Remote Sensing:** Aerosol Measurements. **Tropospheric Chemistry and Composition:** Aerosols/Particles.

Further Reading

- Boucher, O., Randall, D., Artaxo, P., et al., 2013. Chapter 7: clouds and aerosols. In: Stocker, T.F., Qin, D., Plattner, G.-K., et al. (Eds.), Working Group I Contribution to the Intergovernmental Panel on Climate Change Fifth Assessment Report (AR5), Climate Change 2013: The Physical Science Basis. Cambridge University Press, Cambridge, United Kingdom, New York, NY, USA.
- Seinfeld, J., Pandis, S., 2006. Atmospheric Chemistry and Physics: From Air Pollution to Climate Change. John Wiley & Sons, Hoboken, NJ, 1232 pp. ISBN: 978-0-4711-7816-3.

Soot

P Chylek, Dalhousie University, NS, Canada

SG Jennings, National University of Ireland, Galway, Ireland

R Pinnick, US Army Research Laboratory, Adelphi, MD, USA

© 2015 Elsevier Ltd. All rights reserved.

Introduction

Soot (also often called black carbon, charcoal, elemental carbon, or graphitic carbon) is produced by incomplete combustion of carbonaceous materials. Soot is found everywhere on Earth, including the atmosphere, oceans, sediments, soil, and ice sheets. It is also found in meteorites, may be present in asteroids and comets, and is believed to be responsible for dark absorption bands observed in stellar spectra. Soot is even suspected of participating in the initiation of life processes.

In the atmosphere, particularly in the boundary layer, soot is a major component of aerosols that strongly absorb solar radiation. Soot particles, when combined with sulfates, nitrates, sea salt, and organic particulate carbon present in the atmosphere, can serve as cloud condensation nuclei. Soot particles inside cloud droplets increase the absorption of solar radiation by droplets and modify droplet size distribution. Soot also provides a suitable surface and serves as a catalyst for various atmospheric heterogeneous chemical reactions. Thus soot is an important constituent of the atmosphere that affects atmospheric chemical composition and atmospheric radiation balance through both its direct effects (absorption and scattering of solar radiation) and its indirect ones (modifying the formation and lifetime of clouds and the size distribution of droplets).

Soot contributes to atmospheric pollution. It reduces visibility and is also blamed for a variety of adverse health effects including a long list of respiratory diseases and various cancers.

The main sources of soot in the atmosphere are biomass burning and fossil fuel combustion. Soot is the only material suspended in the atmosphere with a long residence time (up to 10 days) that strongly absorbs electromagnetic radiation of all wavelengths. Other atmospheric aerosols have either a very low absorption (sulfates, nitrates, sea salt and organic particulate matter) or a moderate absorption (soil and mineral dust) at visible wavelengths. From this follows the unique role of soot in the atmosphere as the only component of the atmospheric aerosol which strongly absorbs the visible part of solar radiation.

Carbon is a major component of all living material. Carbonaceous particles produced by biomass burning or fossil fuel combustion span a large range of sizes. Particles with diameter over 10 μm are subject to fast gravitational settling and are removed from the atmosphere within a short distance from their sources. On the other hand, submicrometer-sized particles remain suspended in the atmosphere for several days and are transported over long distances. Black carbon (soot) has been found at all places on the globe, including the most remote areas in Antarctica.

Carbon has atomic number $Z = 6$. There are two stable carbon isotopes, ^{12}C and ^{13}C , and four radioactive ones: ^{10}C , ^{11}C , ^{14}C , and ^{15}C . The ^{14}C isotope is used for carbon dating. The carbon atom has 6 electrons, and 4 of them are in the outer (2S and 2P) electron shells. These 4 valence electrons are available to form stable covalent bonds (shared pairs of electrons between 2 neighboring atoms) with other carbon atoms or atoms of other elements. Carbon atoms can thus form chains or rings of high complexity. If all 4 electrons are used in covalent bonds, the resulting materials are generally transparent in the visible part of the electromagnetic spectrum. On the other hand, if not all valence electrons are used for covalent bonds then the unused electrons can form a cloud of non-localized electrons, as in graphite, and the material will start showing a definite degree of absorption and anisotropic electric conductivity.

There is an enormous variety of organic compounds of carbon. They are compounds primarily of carbon with oxygen, hydrogen, and nitrogen, although compounds with a number of other elements, including sulfur, phosphorus, and halogens, are also formed. When heated, all organic substances have one thing in common: they always produce, in addition to steam and carbon dioxide, a black material commonly called the char, soot, black, or elemental carbon. This is due to the fact that almost all of the combustion processes taking place are incomplete (oxygen-deficient): they do not provide sufficient oxygen for the full oxidation of the fuel, and generally some of the carbon will end up in a condensed phase rather than in gaseous oxide form.

Graphite and Black Carbon

The two basic forms of solid elemental carbon are diamond and graphite. They differ from each other in the form of lattice structure into which carbon atoms are arranged. This difference leads to vast dissimilarities in physiochemical and optical properties between the two carbon forms.

Graphite has a structure of a planar hexagonal lattice with 4 carbon atoms per primitive cell. Within the lattice plane each carbon atom is bound to 3 neighboring atoms by strong covalent bonds. The 4 valence electron of each atom contributes to a relatively weak bond between planes of hexagonal lattices. These electrons are not bound to any particular carbon atom (nonlocalized electrons) and they can move relatively freely within a periodic potential formed by a hexagonal lattice of carbon atoms in graphite. Nonlocalized electrons are responsible for good electric conductivity of graphite in the basal plane of the hexagonal structure and for its absorption properties in the visible part of the electromagnetic spectrum.

Separation between lattice planes is about 2.4 times the nearest neighbor distance (about 0.142 nm) within the plane. Graphite is often called a semimetal, indicating that it has some properties similar to that of metals. However, its effective density of charge carriers is of order 10^{18} cm^{-3} ; several orders of magnitude below that of typical good metals (10^{22} cm^{-3}).

In the diamond lattice, the 4 nearest neighbors form the vertices of a regular tetrahedron; all 4 valence electrons of each carbon atom are used to form strong covalent bonds with the four nearest neighbors. There is no planar anisotropy and no free electrons. A diamond is an extremely hard, high-density, transparent nonconductor.

The basic characteristic of graphite and the basis of its high absorption in the visible part of the spectrum is its planar hexagonal structure. Whenever a sufficiently large number of carbon atoms get arranged in the form of a planar hexagonal lattice, some electrons will be only weakly bound to their respective atoms; they will form almost a free electron cloud and the material will manifest an increased conductivity and light absorption. This happens even if there are other atoms involved with carbon, as long as the number of other atoms is relatively small (usually below 20% by mass). Such a material, which is not a pure graphite, but at least partially manifests the basic graphite characteristics (elevated conductivity and increased absorption at visible wavelengths), is generally referred to as black carbon. There are two ways in which carbon atoms can be induced into the planar hexagonal lattice: either through a process of soot formation (high-temperature combustion) or by charring (lower-temperature burning).

Particulate Emission by Fossil Fuel and Biomass Burning

Particulate material ejected into the atmosphere during combustion processes contains soot, charcoal and ash. Ash originates from an inorganic component of fuel. Its mass is usually small (around 1%) compared with the mass of other forms of particulate matter ejected to the atmosphere. Unburned hydrocarbons react with atmospheric oxides of nitrogen and solar radiation to form smog.

Soot

Soot production generally proceeds through condensation of vaporized organic matter, usually through a number of polycyclic aromatic hydrocarbons (PAH). This is a complex process involving, first, the production of benzene and acetylene from the original biomass of fossil fuel. It is believed that most fuels break down into the same species at the beginning of the sooting process. In the second step of soot formation, the acetylene and benzene are transformed into the phenyl, a simple aromatic hydrocarbon with just one ring. The chain of aromatic rings then grows through a fast polymerization process (replacement of hydrogen atoms by C_2H_2 groups). With an increasing number of aromatic rings, a nucleus of a soot particle is formed. Some models consider four rings to be sufficient for soot nucleation. Thus the soot is produced by gas to particle conversion. A typical size of a soot nucleus is a few nanometers. The nucleus grows by additional condensation

and coagulation. Freshly produced soot particles are almost spherical and they have tendency to form, through coagulation, chainlike structures with fractal geometry (Figure 1), with a fractal dimension of about 1.8. Soot production takes place at high temperatures, above 1000°C , during the fossil fuel combustion or during the flaming stage of biomass burning.

A mature soot particle is typically composed of a stack of layers, each of them having a graphite-like hexagonal structure. Not all layers are arranged in a parallel fashion. In addition to carbon, soot contains remnants of other elements present in the original fuel. A typical carbon content of soot is between 90 and 98%. Soot X-ray analyses indicates the presence of a regular graphite structure throughout the soot volume.

Generally, fuels with higher C/H (carbon-to-hydrogen) ratio produce more soot. For a given amount of fuel, the variable flow conditions produce more soot than a steady-flow regime. Soot can also be produced by the oxidation of almost pure elemental carbon. At high temperatures, a carbon vapor is formed, which in colder regions, away from the flame, condenses to form solid carbon structures. In this way graphite-like soot as well as the famous fullerenes C_{60} and C_{70} are formed.

Charcoal

Charring of organic materials starts at temperatures considerably lower than that of soot formation. Burning of food during cooking (i.e., the production of nicely black toast) is an example of low-temperature charring. At temperatures above about 300°C , most of the organic materials undergo a slight thermal decomposition; hydrogen and other noncarbon elements are stripped from carbon chains and rings and the carbon condenses into a graphite-like structure. The density of black porous residuum depends on the mass ratio of carbon to other elements in the original material.

X-ray analysis confirms that at temperatures above 300°C the hexagonal, graphite-like structure begins to form. This structure becomes more evident and more regular with an increasing temperature of oxidation. As hydrogen and other elements (e.g., nitrogen and sulfur) are released to the atmosphere and carbon atoms start forming the basic planar hexagonal structures, the optical properties of the material

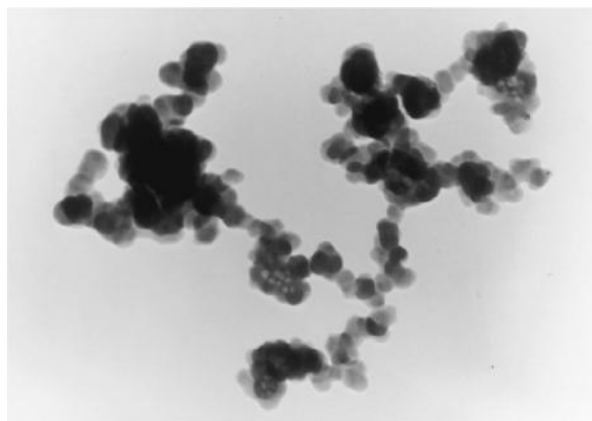


Figure 1 Morphology of freshly produced soot, showing a characteristic chain-like structure of nanometer-size soot particles.

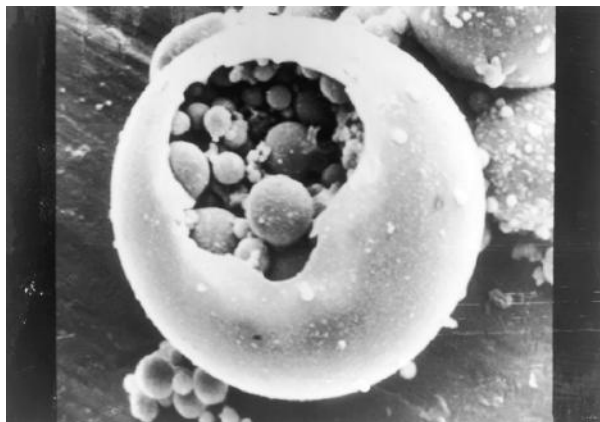


Figure 2 A typical black carbon (charcoal) particle structure from a coal-fired power plant. Photograph by R. Cheng; reproduced with permission from Chylek P, Ramaswami V, Pinnick R, and Cheng R (1981). Optical properties and mass concentration of carbonaceous smokes. *Applied Optics* 20: 2980–2985.

undergo a drastic transformation. With an increasing number of planar hexagonal rings, there is an increasing number of nonlocalized, almost free, electrons, and the material starts showing some of the graphite characteristics, especially an increased absorption of visible electromagnetic radiation. Original organic material becomes dark brown or black.

In the case of biomass burning the charring process takes place during the smoldering phase. Fossil fuel combustion often produces carbonaceous particles that are in a form of hollow spheres (cenospheres) or of spongy spherical structures (Figure 2 and Figure 3).

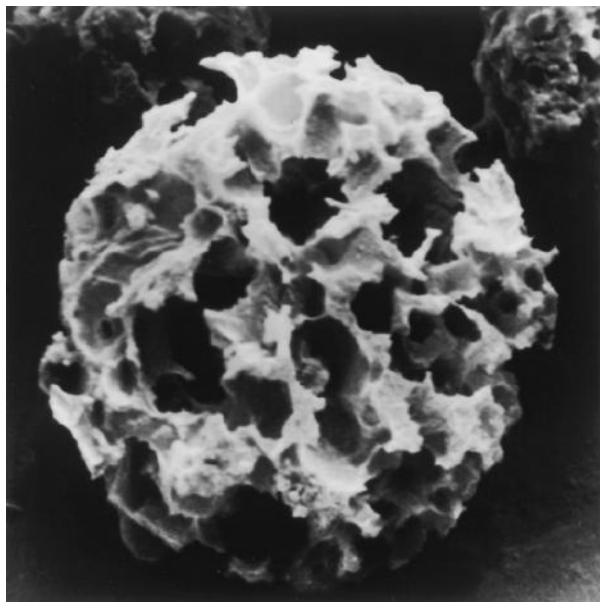


Figure 3 Black carbon (charcoal) particle structures from an oil-fired power plant. Photograph by R. Cheng; reproduced with permission from Chylek P, Ramaswami V, Pinnick R, and Cheng R (1981). Optical properties and mass concentration of carbonaceous smokes. *Applied Optics* 20: 2980–2985.

The sizes of particles produced by charring are from sub-micrometer to several hundred micrometers. Smaller sizes are uplifted during the turbulent conditions produced by localized heating. They reside in the atmosphere for an extended time and are transported over long distances.

Organic and Black Carbon

Regarding aerosol radiative effects, the total carbon in atmospheric aerosols (excluding inorganic carbon in the form of carbonates as a part of soil and mineral dust particles) is usually divided into so-called organic and black carbon. This division is based not on aerosol chemistry, but rather on the aerosol optical properties. Carbon of atmospheric carbonaceous aerosols that absorbs visible radiation strongly is called black carbon; the remaining carbon (carbon of nonabsorbing carbonaceous particulate matter) is organic carbon. The black carbon defined in this way contains pure graphite (elemental carbon), soot, and charcoal as well as their internal mixtures and their mixtures with organic carbon. Black carbon generally resists oxidation at temperatures below about 400 °C, while organic carbon is oxidized easily at lower temperatures.

The separation of total carbon into organic and black carbon is not unambiguously defined chemically. Some of more complex organic compounds may show a substantial absorption in the range of visible wavelengths. Should they be a part of organic or of black carbon? If we are interested in radiative effects of carbonaceous aerosols then all absorbing material should be kept in a category of black carbon. On the other hand, if we are interested in chemical reactions of organic aerosols then we may keep even absorbing organic compounds in the inventory of organic (rather than black) carbon. From the point of view of absorption of solar radiation in the atmosphere, it is reasonable to divide the total carbon into organic and black carbon, even if this division is not chemically well defined.

Black Carbon Measurements

Black carbon (soot) properties, such as density, absorption coefficient, size, and morphology are highly variable. They depend on conditions of generation, source strength, atmospheric transport, transformations due to mechanisms such as catalytic surface reactions, and their degree of mixing of black carbon with other atmospheric particles as well as of their removal due to wet and dry deposition processes.

Measurements of mass concentration, absorption, and size distribution of black carbon are relatively sparse up to the late 1970s, owing mainly to lack of suitable instrumentation.

An increased interest in the role of soot in the atmosphere brought about the development and evaluation of new analytical methods and measuring techniques.

Mass Concentration and Size Distribution

Most soot size distribution measurements have been obtained from filter samples using multistage impactors combined with either conventional or transmission and scanning electron

Table 1 Summary of black carbon (soot) measurements

Region	Extinction coefficient (m^{-1})	Mass concentration ($\mu\text{g m}^{-3}$)
Remote (Antarctic/Arctic)	1×10^{-8}	0.001
Mid troposphere	$1\text{--}3 \times 10^{-8}$	0.001–0.003
Marine	$1\text{--}5 \times 10^{-7}$	0.01–0.05
Rural/continental	$1\text{--}5 \times 10^{-6}$	0.1–0.5
Urban	$1\text{--}4 \times 10^{-5}$	1.0

microscopy. More recent techniques include the use of an optical scattering aerosol sizing probe equipped with a heated intake.

Soot particle size resides predominantly in the sub-micrometer accumulation mode regime, with a geometric mean diameter in the range 0.05–0.2 μm and with a geometric standard deviation 1.33 to 2.0. The average particle size increases with time during long-range atmospheric transport.

Typical soot mass concentration values (Table 1) range from about 1 ng m^{-3} for remote Antarctic locations to more than 1 $\mu\text{g m}^{-3}$ for polluted urban air. The number concentration varies from about 0.1 to $>100 \text{ cm}^{-3}$.

The mass extinction coefficient in polluted urban environments has typical values in the range 10^{-3} to 10^{-4} m^{-1} , while representative values for a more remote atmosphere are $\leq 10^{-5} \text{ m}^{-1}$.

Black Carbon in Precipitation

The removal of black carbon from the atmosphere is believed to be primarily by wet deposition. However, there are only a few measurements of black carbon concentration in rain and snow. The method used consists of the filtering of collected precipitation through quartz fiber filters, followed by a thermo-optical method of determination of the amount of black carbon on the filter. The range of black carbon concentration measured in rainwater and in snow is summarized in Table 2.

Black Carbon in Ice Cores

Ice cores preserve the information concerning the state of the atmosphere at the time of snow deposition (analysis of ice and aerosols) and at the time of enclosure of air bubbles (analysis of gases trapped in bubbles). Black carbon concentration in ice

Table 2 Black carbon concentration in cloud water and in precipitation

Type of cloud or precipitation	Black carbon ($\mu\text{g kg}^{-1}$)
Marine Stratus, North Atlantic	8–60
Stratocumulus, eastern Pacific	20–80
Rain Water, eastern Canada	1–11
Rain Water, Seattle	3–400
Snow, eastern Canada	1–32
Snow, New Mexico and western Texas	5–16
Snow, Cascade Mountains	22–59
Snow, Camp Century, Greenland	2–3
Snow, Antarctica	0.2

cores can be used to deduce the information concerning the past climate and the effect of man's activities on the atmosphere.

The black carbon concentration changes in Alpine glaciers indicate the increase of atmospheric black carbon concentration due to an increase in the regional industrial activities. On the other hand, no increase in black carbon concentration has been found in several analyses of Greenland ice cores. A comparison of black carbon (soot) concentrations found in the Greenland Summit GISP2 (Greenland Ice Sheet Project 2) ice core dated to around the years 320–330 AD and recent (1989–1990) snow from the same location suggests the same average concentration of about 2 ng g^{-1} (Figure 4).

Optical Constants of Black Carbon (Soot)

Determination of the complex refractive index $m = n + ik$, where i is an imaginary unit, n and k are real and imaginary parts of refractive index, respectively, is a difficult task for soot or atmospheric black carbon. A number of different approaches have been made to determine the refractive indices of soot carbon.

One of the principal methods used has involved the measurement of reflectivity of electromagnetic radiation from polished soot-like materials. Reflectivity methods have been applied to soot material, which has been compressed into pellets with nearly specular surfaces. The compression does not result in a uniform carbon material, but contains voids, which have to be considered in the determination of the optical constants.

A combination of transmission and reflection has been used on an amorphous thin film of carbon. Another approach has involved extinction measurements for a suspension of carbon particles (of mean diameter 75 nm), which overcomes uncertainties associated with purity, crystal microstructure variations and void fraction of the sample. Indirect determination of the refractive indices of flame soot has been carried out *in situ* using

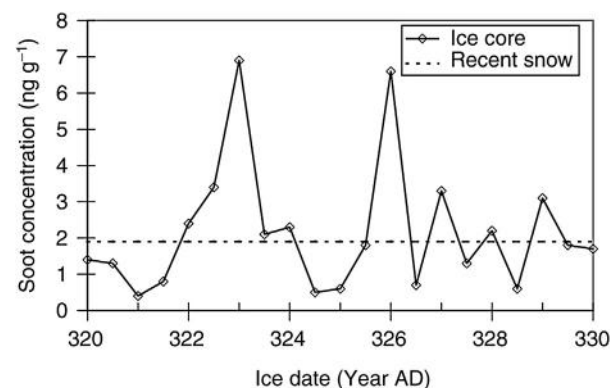


Figure 4 Comparison between soot concentrations (ng g^{-1}) in the Greenland Summit GISP2 ice core dated about 320–330 AD with that in recent snow (1989–1990) from the same location. There is no change of an average soot concentration in remote Greenland location between the current snow and the ice core more than 1700 years old. Large, ancient forest fires somewhere in the Northern Hemisphere are represented by peaks in soot concentrations around the years 324 and 326 AD.

Table 3 Black carbon optical constants (real and imaginary part of refractive index) in the 0.35–1.5 μm wavelength range

Material	Real part	Imaginary part	Method of measurement
Amorphous carbon	1.85–2.8	1.2–0.9	Transmission, reflection
Carbon black	1.92	0.95	Extinction
Polycrystalline graphite	2.24	1.04	Fresnel reflection
Coal samples	1.6–2.1	0.3–0.5	Reflectance
Soot	1.5–1.9	0.4–0.8	Reflectance

light scattering combined with extinction measurements. However, the soot particle size and number concentrations were not measured directly.

A summary of measured optical constants of soot is presented in Table 3. The variability in the data can be attributed largely to factors such as degree of sample homogeneity, compositional change such as C/H ratio, density, sample preparation, etc. Recommended values for the refractive index of black carbon within the wavelength range from 0.3 to 1.5 μm (measurements indicate that the parameters do not greatly change with wavelength in the solar spectrum region) are: $m = (1.9 \text{ to } 2.0) + i(0.7 \text{ to } 1.0)$.

Effect of Soot on Radiative Properties of Aerosols and Clouds

When soot gets incorporated inside cloud droplets or within a composite aerosol particle (to form an internally mixed aerosol) it modifies their radiative properties. The main effect of soot is to increase the absorption by droplets and aerosol particles. Since soot exhibits a strong absorption at all wavelengths from UV to far infrared, while liquid water has a strong absorption only in the infrared region, it is mainly the absorption of the visible and UV radiation that is enhanced by the presence of soot. Consequently, the presence of soot will decrease the single scattering albedo, ω , at visible wavelengths of cloud droplets and aerosol particles.

The intensity of the electromagnetic field within a water droplet or sulfate aerosol is higher than the intensity in free space, owing to the focusing effect of the droplet. On average, a soot particle within a droplet, or as an internally mixed aerosol, will absorb more than twice as much of the incoming radiation than (externally mixed) soot in the free atmosphere.

Effective Medium Approximation

The single scattering albedo (the ratio of the scattering to the sum of the scattering and absorption cross-sections) of a spherical aerosol particle or water droplet can be calculated using the standard Mie scattering formalism. Mie scattering calculations require as input the size of a spherical particle and its refractive index. A refractive index of a composite particle (a droplet of given material with soot inclusions inside) can be calculated using an effective medium approximation. For soot inclusions considerably smaller than the wavelength of a considered radiation, the Maxwell–Garnett effective medium

approximation with soot inclusions surrounded by water matrix (or other material of the original particle) is an appropriate form of an effective medium approximation. The effective refractive index, m_{eff} , of a composite droplet is given by

$$m_{\text{eff}}^2 = m_0^2 \left(1 + \frac{4f(m_c^2 - m_0^2)}{m_c^2 + 2m_0^2 - 2f(m_c^2 - m_0^2)} \right) \quad [1]$$

where m_0 and m_c are refractive indices of a matrix material (water or sulfate) and soot inclusion, and f the soot volume fraction. The single scattering albedo of a composite water–soot or aerosol–soot particle is then obtained by applying the Mie scattering formalism to a homogeneous particle whose optical properties are described by an effective refractive index.

Soot and Direct Radiative Effect of Aerosols

Soot incorporated within an aerosol particle will increase the particle's absorption in the visible part of solar spectrum and thus it will decrease the particle's single scattering albedo. The direct top of the atmosphere radiative forcing, ΔF , of an optically thin aerosol layer is given by

$$\Delta F = -\frac{S_0}{4} T_{\text{atm}}^2 (1 - N) [(1 - a)^2 2\beta\tau_{\text{sc}} - 4a\tau_{\text{abs}}]. \quad [2]$$

where S_0 is the solar constant, N the fraction of sky covered by clouds, T_{atm} the transmittance of the atmosphere above the aerosol layer, a the surface albedo, β the fraction of the scattered radiation that is scattered into the upper hemisphere, and τ_{sc} and τ_{abs} the scattering and the absorption optical thickness of an aerosol layer.

The negative value of radiative forcing implies cooling of the system, while a positive value implies heating. For non-absorbing aerosol $\tau_{\text{abs}} = 0$, and eqn [2] implies always a cooling effect. When soot is present within an aerosol, aerosol absorption increases and the direct aerosol effect will be either cooling or heating, depending on the relative magnitudes of the terms inside the bracket on the right-hand side of eqn [2]. For an optically thin aerosol layer, $\omega = \tau_{\text{sc}}/(\tau_{\text{sc}} + \tau_{\text{abs}})$. The critical single scattering albedo, ω_{cr} , which determines whether an aerosol will heat or cool the system, is derived from eqn [2] in the form

$$\omega_{\text{cr}} = \frac{2a}{\beta(1 - a)^2 + 2a} \quad [3]$$

For given surface albedo, a , and backscattering fraction, β , an aerosol with single scattering albedo $\omega > \omega_{\text{cr}}$ will cool the system, while aerosols with $\omega < \omega_{\text{cr}}$ will cause heating.

Thus the sign of a direct top-of-the-atmosphere aerosol forcing depends – in addition to the fraction of radiation scattered into the upward hemisphere and the albedo of an underlying surface – on the amount of soot within an aerosol particle (which determines the single scattering albedo ω). Most aerosols will cause cooling over the ocean and heating over fresh snow. Thus, the soot heating effect will be especially significant over clouds, ice, and snow.

Soot and Absorption of Solar Radiation by Clouds

Soot within cloud droplets will again increase the droplets' absorption of electromagnetic radiation and decrease their

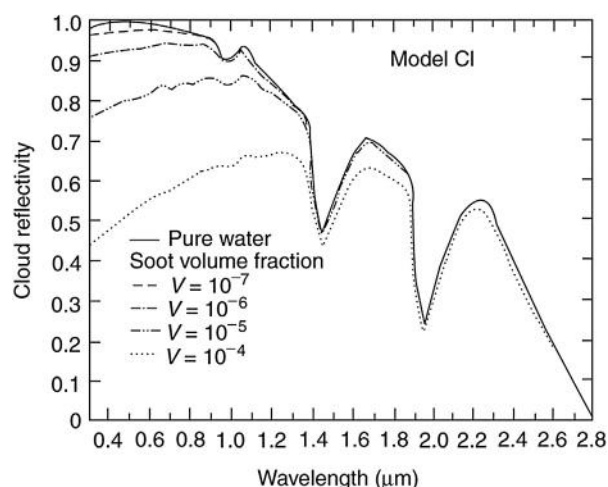


Figure 5 Cloud reflectivity as a function of radiation wavelength for an optically thick (semi-infinite) cumulus cloud. The cases of pure water cloud droplets and for varying soot volume fractions in cloud droplets are shown. Adapted with permission from Chylek P, Ramaswamy V, and Cheng RJ (1984). Effect of graphitic carbon on the albedo of clouds. *Journal of the Atmospheric Sciences* 41: 3076–3084. A significant reduction of cloud reflectivity at visible and near-infrared wavelengths is obtained for soot volume fractions at and above 10^{-6} .

single scattering albedo. This leads to an increased absorption of solar radiation within a cloud layer, to heating, and to a possible increased rate of evaporation of cloud droplets.

A small amount of soot, of the order of 10^{-9} to 10^{-7} by volume, in cloud droplets has little effect on cloud optical

properties. However, soot in highly polluted regions, produced by industrial activities or biomass burning, can affect cloud absorption. Soot in cloud water concentration of the order of 10^{-6} and above will increase cloud absorption significantly. The effect of soot volume fraction, varying from 10^{-7} to 10^{-4} , on the reflectivity of cloud is quite pronounced at visible wavelengths, as shown in Figure 5. Most accumulation-size soot particles can propagate up to several thousands miles away from their sources without a significant decrease in soot concentration. Thus, for example, an extensive biomass burning can affect cloud absorption and regional climate in regions several hundred miles away from source.

See also: **Aerosols:** Aerosol–Cloud Interactions and Their Radiative Forcing; Role in Radiative Transfer. **Aviation Meteorology:** Aircraft Emissions. **Boundary Layer (Atmospheric) and Air Pollution:** Overview. **Clouds and Fog:** Cloud Microphysics. **Tropospheric Chemistry and Composition:** Cloud Chemistry.

Further Reading

- Flanner, Mark, G., 2013. Arctic climate sensitivity to local black carbon. *Journal Of Geophysical Research-Atmospheres* 118 (4), 1840–1851.
- Jacobson, Mark, Z., 2012. Investigating cloud absorption effects: Global absorption properties of black carbon, tar balls, and soil dust in clouds and aerosols. *Journal of Geophysical Research-Atmospheres* 117 (D06205). <http://dx.doi.org/10.1029/2011JD017218>.
- Chylek, P., Ramaswamy, V., Cheng, R.J., 1984. Effect of Graphitic Carbon On the Albedo of Clouds. *Journal of the Atmospheric Sciences* 41 (21), 3076–3084.

Agricultural Meteorology and Climatology

ES Takle, Iowa State University, Ames, IA, USA

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

Agricultural meteorology draws on basic physical and biological sciences to discover, define, and apply knowledge of weather and climate to production of food-, feed-, fiber-, and bio-based products. Agricultural meteorology is based on fundamental principles of radiation and surface aerodynamics and thermodynamics. Models of atmospheric interactions with plants and soil, made more applicable by expanding historical databases, find increased application in risk management and climate change. A highly trained workforce in agricultural meteorology is needed to address future needs for global food security in a changing global climate.

Introduction

Agricultural meteorology is an interdisciplinary science concerned with discovering, defining, and applying knowledge of the interactions between meteorological and hydrological factors and biological systems to practical use in agriculture. An ultimate goal of agricultural meteorology is to extend and fully deploy knowledge of atmospheric and related processes to optimize agricultural production, and hence to increase profitability, decrease risk, contribute to biofuel production, and feed an expanding global population. A second goal that is taking on increased importance is to help conserve natural resources and protect our soil, plant, and water resources.

Environmental interactions of a wide range of agriculturally related organisms are of interest to agricultural meteorologists. Although most attention has been focused on agricultural and horticultural crops and forests, this segment of atmospheric science also includes environmental interactions with animals grown to provide food and fiber, insects, plant and animal pathogens, and aquaculture species. Agricultural meteorology, like the entire field of meteorology, has its roots in the study of temperate (midlatitude) regions of the northern hemisphere. In parallel with its parent discipline, agricultural meteorology has more recently intensified its focus on tropical agriculture, with some of the same difficulties of paucity of data faced by tropical meteorology.

The vagaries of weather always have been a leading cause of variability in agricultural production, but the technological era has increased this vulnerability even as it has provided some means of insulating agriculture from adverse conditions. So, for instance, disease-resistant crops, wide availability of soil amendments and chemicals for pest control, and efficient tillage, planting, and harvesting equipment have reduced agricultural vulnerability and increased yields; however, larger fields and wide use of monocultures have exposed crops to vector-borne diseases and insects and exposed soil to erosion by wind and water. The use of chemicals, new varieties, and genetically modified organisms has brought new weather dependencies.

Agriculture is arguably the most weather-sensitive sector of society. Forty percent of the land surface of the Earth is classed as arid, semiarid, and dry subhumid but is home to millions of people, particularly in developing countries. For some of these areas, frequent crop failure due to adverse weather must be

a part of interannual planning by farmers and regional and state policy makers. International agricultural meteorologists, therefore, have significant concern for famine and food security because of their close link to interannual variability of weather and climate. Even in areas having what are considered ideal climates for crops, water management is a major concern, particularly in regions where competing uses of fresh water put increased pressure on agricultural uses of water.

The historical focus of agriculture to produce food and fiber for an expanding global population has been supplemented by a new thrust at the beginning of the twenty-first century. Agriculturists now also have roles in managing soil and landscapes to regulate flows of carbon, nutrients, soil amendments, and pesticides. Atmospheric transport of pesticides, spores, and pollens (particularly those originating from genetically modified plant materials) must be quantified with increased accuracy. Although uncertainty remains large, consensus estimates of sources and sinks of greenhouse gases such as carbon dioxide, methane, and nitrous oxide reveal that agriculture has a significant role. For instance, agricultural sources of methane from ruminant animals, rice production, and biomass burning are comparable to, or may exceed, natural emissions on a global scale.

Agriculture may play a significant role in moving society from its fossil-fuel base for energy and materials to one that relies more heavily on bioenergy and biobased materials. Tillage practices on natural prairie lands have reduced soil carbon by up to 50% in the US Midwest. Opportunities for agricultural recapture of soil carbon by use of high-yield plant varieties, reduced tillage, and improved management of crop residues, fertilization, and irrigation are under consideration.

Most biological and chemical processes in the biosphere are highly temperature and moisture dependent, and meteorology is the study of underlying physics and chemistry that governs these processes. Emerging recognition of the importance of biocomplexity and ecosystem services and the need for sustainable methods of agriculture and economic development are creating new roles for agricultural meteorology. Agricultural meteorologists, therefore, can be expected to play an increasingly larger role in working with scientists from many disciplines to meet the challenges of these new environmental concerns.

Fundamental Principles

Radiation

Agricultural meteorology is built on a foundation of fundamental physical laws with applications to the plant, animal, and soil environments. The principles of radiation describe how radiant energy received directly from the sun and in diffuse form from the atmosphere is made available to plants for photosynthesis and converted by solid and liquid surfaces into other forms of energy. Approximately 40% of the radiation emitted by the Sun is in the visible wavelength band from 0.4 to 0.7 μm , sometimes referred to as the shortwave band or, for biological applications, as the band of photosynthetically active radiation (PAR). Radiation with wavelengths just shorter than those in the PAR zone is called ultraviolet radiation (UV-A from 0.32 to 0.40 μm and UV-B from 0.28 to 0.32 μm). Ultraviolet radiation is not used by plants for photosynthesis, but it can damage living tissue of plants and animals, particularly simple organisms.

Visible radiation may arrive at a leaf surface either directly from the Sun or indirectly by reflection from atmospheric molecules, clouds, or solid or liquid surfaces (including other plant leaves). Leaf orientations and solar zenith angle cause the amount of radiation received to vary over the course of the day. The fraction of plan area of leaves to ground area covered by the plant (including only one side of the leaf) is called the leaf area index and is used to describe the area of the plant available for photosynthesis.

Radiation of wavelengths just larger than visible light is called longwave or infrared radiation, with the band from about 0.7 to 1.5 μm being referred to as the near-infrared region, which accounts for about 40% of the solar spectrum. Wien's displacement law (eqn [1]) relates the wavelength of radiated energy to temperature.

$$\lambda = \frac{2897}{T} \quad [1]$$

In eqn [1], λ is the wavelength (in μm) and T is the temperature in K. Wien's law can be used to show that most terrestrial surfaces emit radiation of 8–12 μm , with most growing plants radiating at about 10 μm . The amount of dry matter produced by a plant per unit absorbed PAR is a measure of its light-use efficiency, with typical values being from 1.5 to 4.5 g dry matter per megajoule. Chlorophyll in leaves makes plants much less reflective in the PAR region than in the near-infrared (by a factor of 4 for corn and soybeans), a fact that allows remote assessment of photosynthesis by use of the normalized difference vegetation index derived from satellite observations.

Heat Balance

Thermodynamic principles provide the basis for relations among atmospheric pressure, temperature, and density (ideal gas law), as well as the transfer and conversion of energy (first law of thermodynamics). A primary focus of agricultural meteorology is the balance of energy (conservation of energy) for the system being studied, such as a metabolizing organism or a plant-covered or soil surface. For an organism, we can describe the steady state-heat balance by eqn [2].

$$R_n + M = C + LE + G \quad [2]$$

In eqn [2], R_n is the net gain of heat from radiation; M is the net gain of heat from metabolism; C is the loss of sensible heat to air by convection; LE is the loss of latent heat by evaporation of water; and G is the loss of heat to ground and vegetation by conduction. E is the evaporation rate (flux of water vapor per unit area) and L is the latent heat of vaporization. All quantities are considered to be averaged values per unit area. For applications to animal agriculture, M is likely to be a significant factor, but for a soil or vegetated surface, the metabolism contribution is negligible. The radiant component of energy consists of absorbed incoming shortwave energy less net emitted longwave energy.

A plant canopy uses a portion of the shortwave component of this net incoming radiant energy for photosynthesis. In a balanced condition, the plant uses its evapotranspiration capacity to regulate its temperature by converting excess sensible heat to latent heat. Most agricultural animals, like humans, also have the capacity to rid themselves of excess heat by means of evaporation.

Surface Aerodynamics

The aerodynamics of plant interactions with the atmosphere provides a basis for understanding how plants exchange moisture, trace gases, and heat energy with and extract momentum from the free atmosphere through turbulent processes. Descriptions of the movement of pollen, spores, insects, and chemical sprays also require information about mean and turbulent flow processes on scales of centimeters to hundreds of kilometers.

Simple representations of atmosphere-surface interactions are given by drag coefficient formulations of vertical fluxes of quantity S from a surface as given in eqn [3].

$$F_s = -U_t(S_t - S_{\text{surf}}) \quad [3]$$

U_t is the transport velocity for the interface and the values of S are taken at height t and at the surface. The transport velocity at the Earth's surface is usually parameterized by use of a drag coefficient (C_{Ds}) for the quantity S and the mean wind speed at some level (usually taken to be 10 m), i.e., $\{V_{10}\}$ (eqn [4]).

$$U_t = C_{Ds}\{V_{10}\} \quad [4]$$

Drag coefficients depend on atmospheric stability but are typically in the range 1×10^{-3} to 5×10^{-3} (dimensionless).

Concepts of gradient or Fickian diffusion have been used to describe fluxes by measuring vertical gradients and using assumptions or additional measurements to estimate transfer coefficients. Under this approach, the turbulent flux of a quantity is proportional to the vertical gradient of its mean quantities above the surface, eqn [5].

$$F_s = -K_s \frac{\partial s}{\partial z} \quad [5]$$

where K_s is the turbulent diffusion transfer coefficient for variable s , usually estimated to be ku_*z (k being von Karman's constant (0.4) and u_* being the friction velocity) with an additional stability correction factor and constant for each variable s . Equation [5] with an assumed form of K_s is used to derive vertical profiles of temperature and horizontal wind speed over homogeneous surfaces. Profiles inside crop

canopies are more complicated and are usually specified by empirical relations.

Evaporation and Precipitation

Agriculture is practiced over large regions of the Earth where water excess or water deficit is a major limitation for successful crops. Therefore, a major focus of agricultural meteorology and climatology is the study of precipitation and evaporation. The heat balance equation can be used to provide an estimate of the evaporation rate for a surface from knowledge of other components of the heat budget by a type of Penman–Monteith equation (eqn [6]).

$$LE = \frac{\Delta(R_n - H_G) + F_W}{\Delta + \gamma} \quad [6]$$

In eqn [6], $\Delta = RH_s \partial q_s / \partial T$, where RH_s is the saturation relative humidity and q_s is the saturation specific humidity; H_G is the soil heat flux; $F_W = C_E \beta (RH_p - RH_a)$, where C_E is the bulk transfer coefficient for moisture, $\beta = u_* / C_D$, RH_p is the relative humidity at the plant or soil surface, RH_a is the ambient relative humidity, u_* is the friction velocity, and C_D is the drag coefficient; and $\gamma = c_p / L$, where c_p is the specific heat capacity of air at constant pressure. In some implementations, Δ and F_W are replaced by factors that include canopy and atmospheric resistances to flow of heat and momentum.

Both amount and timeliness of precipitation and evaporation are of critical importance to agriculture. Irrigation scheduling requires reliable climate information as well as good weather forecasts, particularly with increased competition for fresh water due to increased population and expanded uses of water. Food security exacerbates the vulnerability of many precipitation-deficient developing countries to interannual variability of precipitation and raises the urgency of improved seasonal to interannual forecasts of weather and climate.

Instrumentation, Measurements, and Networks

Agricultural climatology relies on records of basic meteorological measurements having been taken over extensive areas and significantly long periods of time. These records form the basis for understanding climate variability and change and also for extracting statistically significant relationships between meteorological variables and soil and plant processes, plant, animal, and pest development, and seasonal yield. In addition to standard atmospheric measurements, agriculturists need measurements of soil temperature and soil moisture. These measurements are less widely recorded although they (especially soil moisture) are being recognized for their role in climate memory and hence seasonal forecasting. More such measurements and networks for measurements are needed, particularly in developing countries where use of technology to reduce vulnerability to climate variability is severely limited.

The central role of the surface energy balance in agricultural meteorology calls for accurate methods of evaluating fluxes of heat, momentum, moisture, and trace gases from crop, soil, and forest surfaces. Unfortunately, this is not an easy task for

heterogeneous surfaces typically encountered in agricultural applications. Estimates of surface fluxes can be made by drag coefficient formulations and gradient diffusion estimates or by eddy correlation methods.

The most direct measurement of vertical fluxes is accomplished by using eddy correlation methods, which have seen increased use due to wider availability of improvements in sensors and recording and in data archiving equipment and methods. Eddy correlation methods are based on the principle that turbulent flow near the Earth's surface can lead to vertical fluxes of heat, moisture, momentum, or trace gases in the absence of a mean vertical flux of dry air. We express the vertical flux of quantity s as $F_s = c_s \rho w(t)s(t)$, with a time-averaged value given by $\{F_s\} = c_s \rho \{w(t)s(t)\}$, where c_s is a constant for the particular quantity being transported, ρ is the dry air density, and w is the vertical wind speed. We can express w as a sum of a time-independent mean and a time-dependent turbulent component, $w(t) = w_0 + w'(t)$, and similarly for s , $s(t) = s_0 + s'(t)$. We can then write eqn [7]:

$$\begin{aligned} F_s &= c_s \rho (w_0 + w'(t))(s_0 + s'(t)) \\ &= c_s \rho [w_0 s_0 + w_0 s'(t) + w'(t)s_0 + w'(t)s'(t)] \end{aligned} \quad [7]$$

After time averaging, this becomes eqn [8].

$$\begin{aligned} \{F_s\} &= c_s \rho \{w_0 s_0\} + c_s \rho \{w_0 s'(t)\} + c_s \rho \{w'(t)s_0\} \\ &\quad + c_s \rho \{w'(t)s'(t)\} \end{aligned} \quad [8]$$

The first two terms on the right-hand side of eqn [8] are zero because the mean vertical wind speed is zero. The third term vanishes because, by definition, the mean fluctuation of the vertical wind is zero. The last term can be nonzero, however, if the fluctuation of the vertical wind has correlation other than zero with the fluctuating part of s . The time-averaged turbulent flux of s then reduces to $\{F_s\}$ given by eqn [9], which can be computed by combining measured w' and s' taken from simultaneous recordings of fast response measurements of $w(t)$ and $s(t)$.

$$\{F_s\} = c_s \rho \{w'(t)s'(t)\} \quad [9]$$

The Bowen ratio is defined as the ratio of heat flux to moisture flux near the surface (eqn [10]).

$$B = \frac{C}{LE} = \frac{c_p \{w'T'\}}{L \{w'q'_s\}} \quad [10]$$

From eqn [2], ignoring metabolic contributions, we can express the sensible heat flux and latent heat flux, respectively, from the surface as in eqns [11] and [12].

$$C = \frac{B(R_n - G)}{1 + B} \quad [11]$$

$$LE = \frac{R_n - G}{1 + B} \quad [12]$$

Flux measurements by the eddy correlation method present challenges that can lead to uncertainty of 5–30%. For a particular situation being sampled, an appropriate averaging time must be selected to be long enough to ensure a sufficiently large sample but not so long as to mix turbulent processes with phenomena of longer time scales. Perhaps a more serious problem is the 'representativeness' issue: Fluxes at a point over

an agricultural field are not completely vertical, particularly where inhomogeneities exist in the field. It can be difficult to specify the 'surface footprint' from which the surface flux emerged for situations having changing wind directions, terrain irregularities, changing levels of atmospheric stability, or inhomogeneities of surface vegetation, soil moisture, or soil type. Yearlong measurements of CO₂ flux over a mixed-species forest in irregular terrain, for instance, would require considerably more care in interpretation than daytime measurements over a flat field of corn. Despite the additional expense and care needed in conducting measurements and additional effort for analysis, eddy correlation measurements are increasingly being used for evaluating surface fluxes of CO₂ and other trace gases and moisture.

Measurement networks have been established by some local, state, federal, and international agencies to provide both an expanding climate database and a basis for near-term and seasonal agricultural decision making. There is an urgent need to expand these networks to meet the increasing food needs, particularly in developing countries. A US national measurement network of soil moisture and temperature, as recommended by a National Academies of Science report, would benefit both production agriculture and climate science through improvement in seasonal weather forecasts. Remote sensing by satellite is finding expanded use in providing large-scale data of relevance to agriculture, but its use for individual farmers is limited.

Modeling and Theory

Modeling of plant interactions with the atmosphere has emerged from at least two directions: global climate modelers seeking more accurate representation of energy, momentum, and moisture budgets at the Earth's surface and crop modelers seeking ways of understanding plant responses to climate and of projecting yields of agricultural crops. Climate modelers use the so-called soil-vegetation-atmosphere transfer (SVAT) models as 'surface packages' to which they supply meteorological data at each surface grid point at each model time step (a few minutes to hours). The SVAT model then calculates the response of the soil and plants (e.g., evaporation or transpiration, temperature change, soil moisture content, moisture uptake by roots, rain or dew held on leaves, precipitation runoff, and momentum extracted) and returns to the climate model the surface fluxes of heat, moisture, and momentum consistent with these soil- and plant-based changes. Computational constraints limit the detail to which plant processes can be described, but, as simplistic as they are, the models provide a conceptual framework for eventual coupling of more detailed crop, forest, and ecosystem models.

Crop models may be physiologically based or statistically based. Crop growth models are built on plant biophysical processes of agricultural crops and their relationship to environmental factors. They predict growth, development, and yield based on complex interactions between weather, soil characteristics, nutrients, and plants. A practical application of crop growth models is to estimate agricultural production as a function of weather and soil conditions under alternative management conditions. Basic meteorological information

needed to drive these models includes air temperature, precipitation, and solar radiation (or sunshine hours). More advanced models might additionally use dew-point temperature, wind speed, and soil temperature. Statistically based crop models provide large-area yield predictions based on correlations of past yields with regional average weather conditions. These models tend to be much less computationally intensive but also more location specific and hence less transferable to other regions.

The fate of fugitive agricultural chemicals and movement of insects and pollen are addressed by models of atmospheric flow on scales of turbulent eddies to mesoscale meteorology. Large-eddy simulation models and models developed for use in air pollution regulation are sometimes adapted for simulating transport of agriculturally related materials. Recent advances in numerical simulation of turbulent flow through vegetation have been used to understand the aerodynamic functioning of agricultural shelterbelts. Extensions of these models to simulate the complete microclimate provide opportunities for exploring, by use of first principles, complex physical relationships in heterogeneous ecosystems and landscapes.

Concern for national and international food securities has prompted the need for models of seasonal yield of various food crops. Private organizations as well as governmental agencies have developed yield models based on long-range weather conditions. The Food and Agriculture Organization of the United Nations has developed agrometeorological models that forecast yield on the basis of cumulative weekly or 10-day crop water balances for providing early warning of potential food security problems in developing countries. The Agricultural Model Intercomparison and Improvement Project (AgMIP) is an international effort to improve and couple crop and economic models with the next generation of climate impact projections for improved assessment of food security.

Manipulating Microclimates to Enhance Productivity and Reduce Risk

Agriculturists have a long history of enhancing crop growth by manipulating soil and plant microclimates, through use of irrigation, glasshouses, shelterbelts and windbreaks, snow fences, wind machines, surface mulches, certain tillage practices, alley cropping, and agroforestry. The design and operation of such modifications require considerable information on the mean, extremes, and interannual variability of climate at the specific location where the practice is implemented.

Horticulture crops, which typically have a higher value per unit area than grain crops, are sensitive to small changes in microclimate. Also in contrast to grains, horticulture crops are more sensitive to weather-induced reduction in product quality or market value. For instance, the desirable red coloration on some fruits is sensitive to optimal amounts of solar radiation at a critical stage. Manipulation of microclimates for horticultural crops is more cost-effective than for cereals because of both their high value and their sensitivity of quality to microclimate. Weather extremes may have multiyear impacts on agricultural crops grown as perennials (e.g., fruits, nuts, and grapes), which raises the cost-effectiveness of microclimate modifications for reducing such extremes.

Agriculture Meteorology Forecasts

Agriculturists can use weather forecasts with valid times of a few minutes to several months. Weather forecasts are used for planning tillage and planting operations, seed purchase, chemical application, frost suppression, grain harvesting, transport and storage, pest and disease management, and marketing, as well as crop growth calculations and long-range planning.

Major improvements over the past 10 years in our understanding of the El Niño/Southern Oscillation (ENSO)-related phenomena have enhanced prospects for seasonal to interannual forecasts of agriculture-sensitive climate information. Such information now is being used in early warning systems for planning, management, and operations in some tropical areas. In regions where the climate correlation with ENSO is strong, projected ENSO factors have been used to create projections of stress indices. Current research on this and related areas may offer future progress in seasonal to interannual forecasts.

Climate Data

Agricultural climatologists use long-term records of standard meteorological data to compute derived agriculturally related variables such as growing degree days, heat stress units, frost-free days, Palmer drought index, and temperature–humidity index. These variables have been developed from physiological concepts to correlate with crop development stage, crop yield, weight gain in meat animals, daily milk production, or other agriculturally important parameters.

Education and Training

Basic education for agricultural meteorologists is usually acquired by supplementing a conventional meteorology, physics, or environmental science curriculum with courses on plant, soil, or animal science, forestry, or horticulture. Only a few US and European universities offer undergraduate or graduate degrees specifically in agricultural meteorology. Most courses of study with emphasis in agricultural meteorology are connected with more traditional programs in agriculture such as agronomy. India has taken a more coordinated approach than almost any other country to the university education of agricultural meteorologists. Increased needs in the developing world have expanded the definition of agricultural meteorology to include more frequent weather and climate disasters that threaten production systems. The World Meteorological Organization recently and in applications to developing countries has cited the socioeconomic aspects such as irrigation, storage, agroforestry, floods, drought, erosion and desertification, frost, wind protection, simple artificial growth conditions, sustainable farming, and related farmers' income as emerging priorities in agricultural meteorology.

The practical applications of agricultural meteorology have created needs for training programs aimed at changing the knowledge, skills, and behavior of personnel to achieve the objectives of the organizations they work for. The World Meteorological Organization provides in-service training

through regional meteorological training centers that offer specialized courses in basic agricultural meteorology, database management, agricultural meteorology modeling, and hydro-meteorology. These short courses tend to be task oriented, focusing on improving and standardizing the practice of agricultural meteorology, particularly in relation to observations and data management.

Rapid advances in the field of meteorology have increased the need for more lifelong learning opportunities both in basic education and in training. Increased interest in global observing networks for monitoring a wider range of environmental variables exacerbates this need. The Internet offers a potential means of delivering standardized and authoritative educational and training materials to larger fractions of the global agricultural meteorology community.

Future Issues

The potential impact of global climate change on agriculture has been a subject of intense study in recent years. Uncertainties in projections of future climates at regional scales limit the accuracy with which agricultural impacts can be estimated. However, there is high confidence that increases in atmospheric carbon dioxide will have a beneficial effect on crops both through direct fertilization and through increased water use efficiency. The mean yield increase for C_4 crops (e.g., maize, sugar cane, millet, and sorghum) under a doubling of atmospheric CO_2 is minimal, whereas increases for C_3 crops (most other plants) may be up to 30%, other factors being equal. Loss of soil organic matter, leaching of soil nutrients, and salinization and erosion of soils will occur in some climatic zones, which will call for more effective agricultural land use practices. Crop yields and productivity will vary considerably across different climate zones under climate change, with low-latitude and low-income countries being most negatively affected and some high-latitude countries experiencing more favorable crop growing conditions than under the current climate.

Advances in our knowledge in the traditional areas of agricultural meteorology – surface fluxes of energy, moisture, and trace gases and the study of precipitation and evaporation processes – will be urgently needed for coping with interannual variability and long-term change of future climates. And practical applications of this new knowledge require timely, efficient, and worldwide distribution networks. Emergence of social media fostered widespread use of cell phones and internet provides rich opportunities for democratizing the dissemination of time-sensitive weather and climate information for agricultural decision making.

Better understanding of basic agricultural micrometeorology and associated plant and soil processes will allow for continuing advances in applied agricultural meteorology as well as at larger scales of meteorology. Measurements of heterogeneities in soil and microclimate across a field are increasingly being used for site-specific management of plant environments and yield improvement. At larger scales, the subtle processes regulating the exchange of moisture and energy of plants and soil with the atmosphere are key to improvements in numerical models of mesoscale meteorology and global climate.

See also: Boundary Layer (Atmospheric) and Air Pollution: Surface Layer. **Data Assimilation and Predictability:** Data Assimilation; Ensemble Prediction. **General Circulation of the Atmosphere:** Energy Cycle. **Hydrology, Floods and Droughts:** Soil Moisture. **Numerical Models:** Large-Eddy Simulation. **Radiation Transfer in the Atmosphere:** Ultraviolet Radiation. **Synoptic Meteorology:** Forecasting. **Weather Forecasting:** Seasonal and Interannual Weather Prediction; Severe Weather Forecasting.

Further Reading

- Carbone, R.E., Block, J., Boselly, S.E., Carmichael, G.R., Carr, F.H., Chandrasekar, V., Gruntfest, E., Hoff, R.M., Krajewski, W.F., Lemone, M.A., Purdom, J., Schlatter, T.W., Takle, E.S., Titlow, J., 2009. Observing Weather and Climate from the Ground Up: A Nationwide Network of Networks. National Academies Press, Washington, DC.
- Doraiswamy, P.C., Pasteris, P.A., Jones, K.C., Motha, R.P., Nejedlik, P., 2000. Techniques for methods of collection, database management and distribution of agrometeorological data. *Agricultural and Forest Meteorology* 103, 83–97.
- Hanks, J., Ritchie, J.T. (Eds.), 1991. Modeling Plant and Soil Systems. American Society of Agronomy.
- Hoogenboom, G., 2000. Contribution of agrometeorology to the simulation of crop production and its applications. *Agricultural and Forest Meteorology* 103, 137–157.
- International Rice Research Institute, 1989. Climate and Food Security, International Symposium on Climate Variability and Food Security in Developing Countries. International Rice Research Institute, Manila, Philippines.
- Lomas, J., Milford, J.R., Mukhala, E., 2000. Education and training in agricultural meteorology: current status and future needs. *Agricultural and Forest Meteorology* 103, 197–208.
- Maracchi, G., Péronaud, V., Kleschenko, A.D., 2000. Applications of geographical information systems and remote sensing in agrometeorology. *Agricultural and Forest Meteorology* 103, 119–136.
- Monteith, J.L., 2000. Agricultural meteorology: evolution and application. *Agricultural and Forest Meteorology* 103, 5–9.
- Monteith, J.L., Unsworth, M.H., 1990. Principles of Environmental Physics, second ed. Edward Arnold, London.
- Ogallo, L.A., Boulahya, M.S., Keane, T., 2000. Applications of seasonal to interannual climate prediction in agricultural planning and operations. *Agricultural and Forest Meteorology* 103, 159–166.
- Parry, M.L., Canziani, O.F., Palutikof, J.P., van der Linden, P.J., Hanson, C.E., 2007. Climate Change 2007: Impacts, Adaptation and Vulnerability. Contribution of Working Group II to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press, Cambridge, UK.
- Salinger, M.J., Stigter, C.J., Das, H.P., 2000. Agrometeorological adaptation strategies to increasing climate variability and climate change. *Agricultural and Forest Meteorology* 103, 167–184.
- Sivakumar, M.V.K., Gommers, R., Baier, W., 2000. Agrometeorology and sustainable agriculture. *Agricultural and Forest Meteorology* 103, 11–26.
- Stigter, C.J., Sivakumar, M.V.K., Rijks, D.A., 2000. Agrometeorology in the 21st century: workshop summary and recommendations on needs and perspectives. *Agricultural and Forest Meteorology* 103, 209–227.
- Strand, J.F., 2000. Some agrometeorological aspects of pest and disease management for the 21st century. *Agricultural and Forest Meteorology* 103, 73–82.
- Stull, R., 1988. An Introduction to Boundary Layer Meteorology. Kluwer Academic, Dordrecht.
- Wang, H., Takle, E.S., Shen, J., 2001. Shelterbelts and windbreaks: mathematical modeling and computer simulations of turbulent flows. *Annual Review of Fluid Mechanics* 33, 549–586.
- Walthall, C.L., Hatfield, J., Backlund, P., Lengnick, L., Marshall, E., Walsh, M., Adkins, S., Aillery, M., Ainsworth, E.A., Ammann, C., Anderson, C.J., Bartomeus, I., Baumgard, L.H., Booker, F., Bradley, B., Blumenthal, D.M., Bunce, J., Burkey, K., Dabney, S.M., Delgado, J.A., Dukes, J., Funk, A., Garrett, K., Glenn, M., Grantz, D.A., Goodrich, D., Hu, S., Izaurralde, R.C., Jones, R.A.C., Kim, S.-H., Leaky, A.D.B., Lewers, K., Mader, T.L., McClung, A., Morgan, A.J., Muth, D.J., Nearing, M., Oosterhuis, D.M., Ort, D., Parmesan, C., Pettigrew, W.T., Polley, W., Rader, R., Rice, C., Rivington, M., Rosskopf, E., Salas, W.A., Sollenberger, L.E., Srygley, R., Stöckle, C., Takle, E.S., Timlin, D., White, J.W., Winfree, R., Wright-Morton, L., Ziska, L.H., 2012. Climate Change and Agriculture in the United States: Effects and Adaptation. USDA Technical Bulletin 1935, Washington, DC.
- Whitmore, J.S., 2000. Drought Management on Farmland. Kluwer Academic, Dordrecht.
- World Meteorological Organization, 2009. Guidelines for the Education and Training of Personnel in Meteorological and Operational Hydrology. WMO-No. 258. In: Guidelines for Curricula in Agricultural Meteorology, vol. I, Supplement No. 2. Secretariat of the World Meteorological Organization, Geneva, Switzerland.

ARCTIC AND ANTARCTIC

Contents

Antarctic Climate

Arctic Climate

Arctic Haze

Antarctic Climate

J Turner, British Antarctic Survey, Cambridge, UK

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

Very different climatic regimes are found across the Antarctic, from dry, very cold conditions with few depressions over the interior plateau, to windy and wet maritime environments toward the tip of the Antarctic Peninsula. Although well removed from the other continents, the climate of Antarctica is affected by conditions across the other parts of the Earth, and especially the ocean temperatures in the tropical Pacific Ocean. The climate of the Antarctic has a large interannual variability as a result of interactions between the atmosphere, ocean, and ice. Over the last 50 years the Antarctic Peninsula has experienced a surface warming as large as any in the Southern Hemisphere. This warming extends into West Antarctica, but the rest of the continent has experienced little change. The loss of stratospheric ozone (the ozone hole) has had a major impact on the climate of the Southern Ocean, increasing the surface winds by about 15%. The ozone hole is expected to recover by 2060–2070, but if greenhouse gas concentrations continue to rise there will be a surface warming of several degrees across Antarctica, more precipitation, and a loss of sea ice.

Introduction

Antarctica is the highest, coldest, windiest, and driest continent on Earth, with a climate that varies from extremely cold and dry on the high plateau of East Antarctica to maritime across the northern part of the Antarctic Peninsula. Although remote from the major centers of population, it plays a crucial role in the global climate system and is closely coupled to conditions at lower latitudes via the oceanic and atmospheric circulations.

The Antarctic continent is about 40% larger than the United States, covering an area of $14 \times 10^6 \text{ km}^2$, which is about 10% of the land surface of the Earth. The Antarctic ice sheet contains about $30 \times 10^6 \text{ km}^3$ of ice or about 70% of the world's fresh water, which is equivalent to about 60 m of sea level. The ice sheet is made up of three distinct zones, consisting of East Antarctica (covering an area of $10.35 \times 10^6 \text{ km}^2$), West Antarctica ($1.97 \times 10^6 \text{ km}^2$), and the Antarctic Peninsula ($0.52 \times 10^6 \text{ km}^2$) (Figure 1).

The orography of the continent has a profound effect on the climate of high southern latitudes, limiting the extent to which depressions can penetrate into the interior and giving rise to the katabatic (downslope) winds that are a major feature of the coastal zone. The elevation of the ice surface rises very rapidly inland from the coast, and the continent has a domed profile,

with much of it being above 2000 m in elevation and some parts over 4000 m.

The other landmasses of the Southern Hemisphere are well north of the Antarctic, so that the oceanic and atmospheric flow is much more zonal than in the Northern Hemisphere. However, the highest parts of the ice sheet are found in East Antarctica and are slightly offset from the South Pole, which has implications for the atmospheric circulation around the continent.

The Broad-Scale Synoptic Environment

The Antarctic coastal region is a zone of strong, horizontal thermal gradients (baroclinicity) where cold katabatic winds flow off the continent and meet temperate, maritime air, resulting in the development of many depressions (cyclogenesis). It is also the area where many depressions spiraling south from midlatitudes become slow moving and decline (cyclolysis). The depressions carry warm (cold) air southward (northward) on their eastern (western) flanks and play an important part in the poleward transport of heat.

The large number of depressions in the 60–70° S zone results in a low-pressure belt around the continent known as

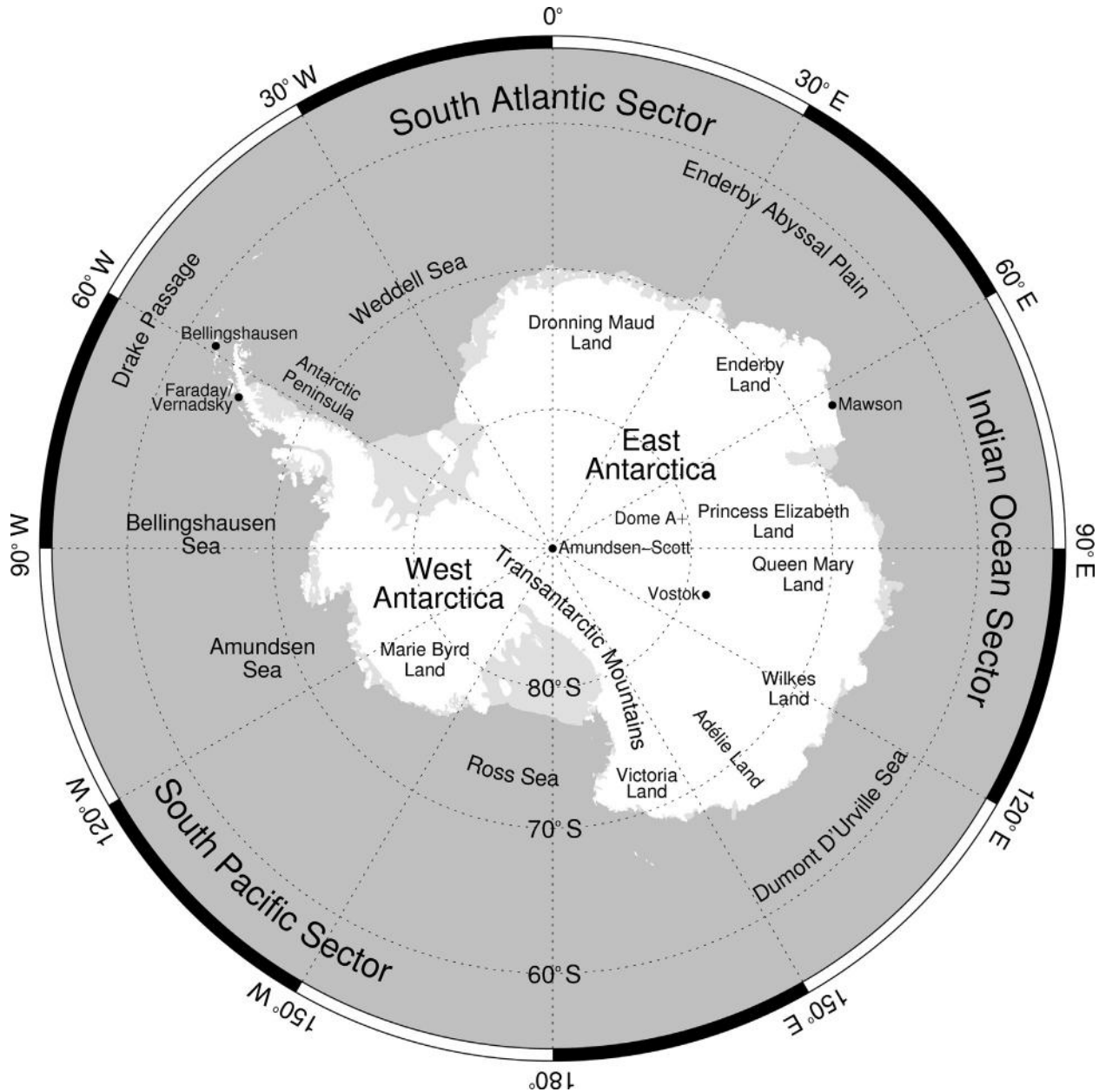


Figure 1 A map of the Antarctic showing regions, topographic features, and the locations of selected research stations.

the circumpolar trough, which is apparent on the mean sea level pressure (MSLP) (**Figure 2**). Within the trough, depressions mainly move toward the east, with the clockwise flow around these systems giving a climatological easterly wind along the Antarctic coast and westerly winds north of the trough.

The circumpolar trough is present throughout the year, and in the mean fields it has an approximate wave number 3 pattern with low-pressure centers close to 30° E, 90° E, and 150° W. This pattern affects a number of aspects of the Antarctic climate, such as the northward extension of sea ice close to the Greenwich meridian, as a result of the climatological southerly flow at this longitude. Because of the distribution of land-masses in the Southern Hemisphere, the atmospheric planetary

waves have a smaller amplitude than their counterparts in the north, so the depressions play a greater role in the poleward transport of heat than in the Northern Hemisphere.

The most marked climatological low-pressure center around the continent is at 150° W and is often referred to as the Amundsen Sea Low. The presence of this low is responsible for the north-to-northwesterly flow on the western side of the Antarctic Peninsula and the relatively mild temperatures that are experienced there. It also gives a mean southerly flow off the Ross Ice Shelf and over the Ross Sea, resulting in this area being a major sea ice production region.

Figure 2 shows that MSLP values within the circumpolar trough are lowest during the spring and autumn and are higher during the summer and winter. This semiannual

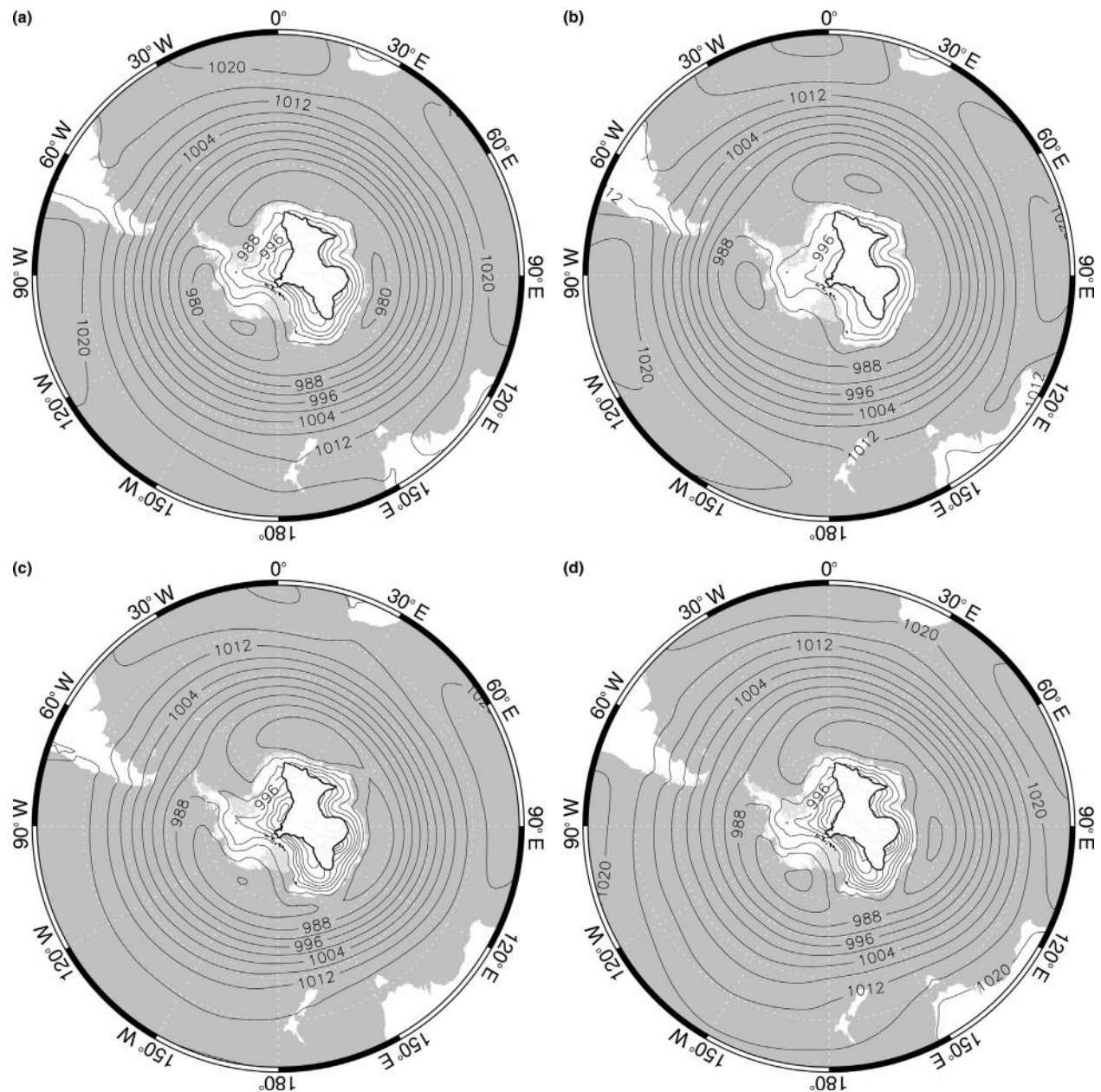


Figure 2 Average mean sea level pressure (hPa) for the four seasons for the period 1979–2010: (a) spring (September–November), (b) summer (December–February), (c) autumn (March–May), and (d) winter (June–August).

oscillation can be seen in the MSLP observations from the coastal stations and also in the number of reports of precipitation. The oscillation is a result of changes in the position and depth of the circumpolar trough over the year, with it being further south (north) and deeper (weaker) in autumn and spring (summer and winter). The oscillation is present because of the phase differences between the seasonal cycles of temperature over the Antarctic continent and Southern Ocean. Across the Antarctic, temperatures drop very rapidly at the start of winter, while over the ocean the minimum is in late winter and early spring. This results in the movement of mass between high latitudes and midlatitudes, giving rise to the semiannual oscillation.

Satellite imagery has revealed that over the Southern Ocean, there are a large number of mesoscale low-pressure systems, which are also known as mesocyclones or polar lows. These have a horizontal length scale of less than 1000 km and a life-time of less than 1 day, so they are difficult to represent and forecast in numerical weather prediction systems. However, they can have a major impact on the weather experienced at coastal sites and so are important in forecasting processes. At the moment, they tend to be predicted using a 'nowcasting' approach, with the systems being identified on satellite imagery and advected with the low- to medium-level tropospheric flow. Although mesocyclones are rare over the high Antarctic Plateau, they are a common feature on the ice shelves. Here,

there is low-level convergence of air that has descended from the plateau, which aids the spin-up of vortices, coupled with the presence of mild, oceanic air masses that provide moisture for cloud formation.

Because of the rapid increase in elevation inland of the coast, few major weather systems penetrate far into the interior of the continent. However, satellite imagery does show that some frontal bands associated with depressions in the circumpolar trough can be identified on the plateau, although automatic weather station data suggest that the pressure signals across these features are small. The conditions that favor depressions having an impact in the interior are amplified planetary waves and strong northerly steering flow aloft. Under such conditions, mild air masses over the plateau can give relatively large falls of precipitation, resulting in a significant fraction of the year's accumulation falling in 1 or 2 days. When the planetary waves are strongly amplified, maritime air masses can affect the South Pole and even Vostok Station on the high plateau of East Antarctica, but such conditions are rather rare.

The Role of Sea Ice

The presence and extent of sea ice across the Southern Ocean have a major impact on the climate of the Antarctic. Unlike in the Arctic, most of the sea ice melts by the late summer, so by February there is on average only about 3.5×10^6 km² of ice, most of which is located over the western Weddell Sea and along the coast of West Antarctica. Through the autumn and winter, the sea ice advances in a divergent fashion around the whole continent, reaching a maximum in September, when the mean extent is about 19×10^6 km².

The Antarctic sea ice is generally about 1 m thick, with some multiyear ice being 2 m or more in thickness where it has been subject to ridging and rafting. The ice provides an effective cap on the upper layers of the ocean, limiting the fluxes of heat and moisture into the lower layers of the atmosphere. However, the effects of the many weather systems over the Southern Ocean on the sea ice is to open up linear cracks (leads) or larger areas of open water (polynyas), which can provide local sources of heat or moisture, resulting in cloud. This can be important for the climates of the coastal stations during the winter months when the opening up of coastal leads and polynyas can significantly increase the temperature and humidity, sometimes leading to fog formation.

An area that is particularly sensitive to the presence or absence of sea ice is the western side of the Antarctic Peninsula. Here, the sea ice passes north–south along the coast during its annual cycle, and years of extensive (limited) sea ice are notably colder (warmer).

Temperature

Much of the Antarctic is extremely cold because of the combined effects of the long period of winter darkness; the high albedo of the snow surface, which results in the reflection of much of the summer incoming solar radiation back to space; and the high elevation, which limits the penetration of maritime air masses into the interior. The Antarctic atmosphere is

characterized by a very strong surface temperature inversion, with temperatures increasing with height over the lowest few hundred meters of the atmosphere. The temperature inversion is strongest in winter on the high plateau and is the result of the intense radiative cooling of the surface and the low wind speeds that give little vertical mixing. The mean strength of the winter inversion (i.e., the temperature difference between the surface and the maximum temperature in the lower troposphere) varies from about 5 °C in the coastal region to more than 25 °C over the highest parts of East Antarctica.

Across the Antarctic, there is a very large range of annual mean surface air temperatures, although it is only in the northernmost part of the Antarctic Peninsula that mean summer temperatures rise above freezing. Over the Antarctic Peninsula and along the coast of East Antarctica, the annual cycle of temperature is similar to those found in midlatitudes, with a broad summer maximum and a minimum in July or August. However, at more southerly latitudes, the cycle is different, with a sharp summer maximum and a 'coreless' winter, during which temperatures vary by only a small amount. This form of the annual cycle comes about for a number of reasons, including the abrupt change in solar radiation at the start and end of the period of austral darkness, the effects of the semiannual oscillation on the annual cycle of advection of warm air into the Antarctic, and the heat reservoir effect of the Antarctic snow pack.

The plateau of East Antarctica experiences the lowest temperatures on Earth, with Vostok Station (78.5° S, 106.9° E, 3488 m elevation) having an annual mean temperature of −55.4 °C. The station has recorded the lowest temperature measured at the surface of the Earth, when on 21 July 1983 the temperature dropped to −89.6 °C. This occurred during a period when there was very little cloud, the wind speed was very low, and the winds blew around the station, limiting the advection of warmer maritime air. A station has recently been established at Dome Argus at an elevation of 4083 m above sea level, and early observations suggest that it is typically 5–6 °C colder than Vostok. It therefore has the potential to record an even lower extreme surface temperature than Vostok.

At higher levels in the troposphere the Antarctic atmosphere is strongly stratified, much more so than in the midlatitude areas of the Southern Hemisphere. This is the case in all seasons, with the stability being strongest below about 4 km during the winter. Temperature data from radiosonde ascents usually show a tropopause in summer, but it can become very indistinct during the winter when the stratosphere cools rapidly.

The Wind Field

The strong, persistent, and directionally constant near-surface winds recorded at a number of sites around the Antarctic are one of the most remarkable features of the continent's climate. Many of the winds at stations around the coast of East Antarctica are katabatic in origin and occur because of the drainage of cold, dense air at low levels from the interior plateau to the coast. [Figure 3](#) shows the near-surface streamlines across the continent derived from the output of a high-resolution weather-forecasting model. These show that much of the flow originates in the higher parts of East Antarctica and flows

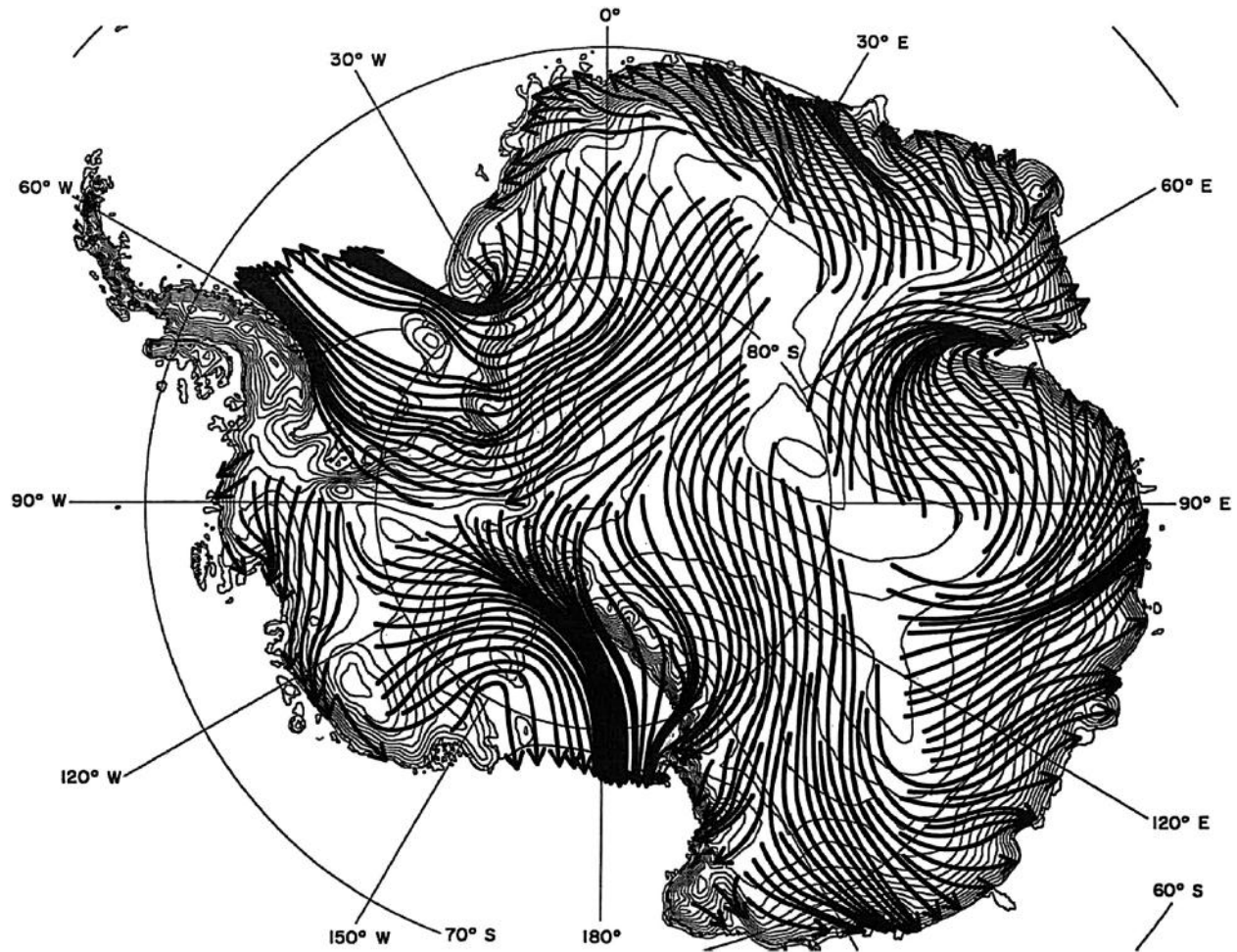


Figure 3 Near-surface mean streamlines for the period June 2003–May 2004 from the Antarctic Mesoscale Prediction System. Reproduced from Parish, T.R., Bromwich, D.H., 2007. Reexamination of the near-surface airflow over the Antarctic continent and implications on atmospheric circulations at high southern latitudes. *Monthly Weather Review* 135, 1961–1973. Used by permission of the American Meteorological Society.

toward the coast, often converging toward certain preferred coastal locations. The katabatic winds are most pronounced during winter, when there is no incoming solar radiation, and a large pool of cold air over the interior is formed to feed the katabatic flow.

Surface winds over the interior show a high directional constancy, indicating that they are dictated by the local orography. The wind speeds are closely related to the slope of the orography, with the strongest winds being measured at stations on the coastal escarpment and the weakest on the parts of the plateau with the smallest orographic gradient. Along the coast of Adélie Land, the orography channels the katabatic flow onto a small stretch of coast, resulting in very strong and persistent winds with a very high directional constancy. It was in this area that Douglas Mawson's 1912–13 expedition recorded the Earth's highest annual mean wind speed of 19.4 m s^{-1} and gale-force winds on all but one of 203 consecutive winter days.

As the katabatic winds descend from the plateau, they interact with the synoptic-scale weather systems within the circumpolar trough. The northerly winds to the east of the lows

tend to suppress the southerly katabatic flow, while the katabatics are enhanced by the southerly flow to the west of the depressions. The Coriolis force also affects the katabatic winds, tending to turn them to the left so that they merge with the coastal easterlies on the southern side of the circumpolar trough. The near-surface flow therefore appears as an anticyclonic vortex, with cold air outflow from the continent. In some parts of the coastal region, such as south of the Weddell Sea, the coastal easterly flow comes up against high orography, and the cold, stably stratified air at low levels does not have the kinetic energy to cross the barrier. The air is then dammed up against the barrier until a pressure gradient develops that results in the air moving north as a 'barrier wind.' With the strong static stability encountered at low levels in the Antarctic, barrier winds are relatively common in the coastal areas of the continent.

Clouds and Precipitation

Clouds are very important in the Earth's climate system as they can reflect a high proportion of incoming solar radiation back

to space. However, since the surface of the Antarctic already has a high albedo by virtue of its year-round snow cover, clouds over the continent tend to have less of an effect on the incoming solar radiation because the surface and cloud have similar albedos. Nevertheless, clouds play a very important part in controlling surface temperatures through their effect on the long-wave radiation budget. In cloud-free conditions, the dry atmosphere allows most of the emitted terrestrial radiation to escape to space, resulting in very low temperatures. However, when thick cloud cover is present, surface temperatures are much higher because of the downward long-wave radiation emitted from the cloud.

Since most of the research stations are located in the coastal region, it is difficult to get an accurate picture of the cloud distribution across the continent. However, using in situ data and satellite imagery, climatologies of cloud cover have been prepared. These suggest that the highest fractional cloud cover is found over the ocean areas north of the edge of the continent, with about 85% cloud cover throughout the year near 60° S. In

the coastal area near 70° S, the surface observations indicate that the total cloud cover is about 45–50%, with little seasonal variability and only a small decrease during the winter months. Inland of the coast, the amounts of thick cloud decrease rapidly, since few synoptic-scale weather systems are found over the interior. However, the inland areas are characterized by extensive, very thin cirrus cloud, which gives a semipermanent veil of ice crystals. This type of cloud causes problems for observers, who have to decide whether to report no cloud or 100% cloud cover. The mean annual percentage cloud cover at the South Pole is 45%, but anyone using such statistics has to be aware of the nature of the cloud that occurs there and the problems facing observers of how to report the thin cirrus.

The amount of precipitation across the Antarctic generally follows the distribution of thick cloud. In other words, the highest precipitation totals are found in the coastal region, with a rapid decrease inland. **Figure 4** shows the mean annual net accumulation (precipitation–evaporation) across the continent as estimated from ice cores. These glaciological

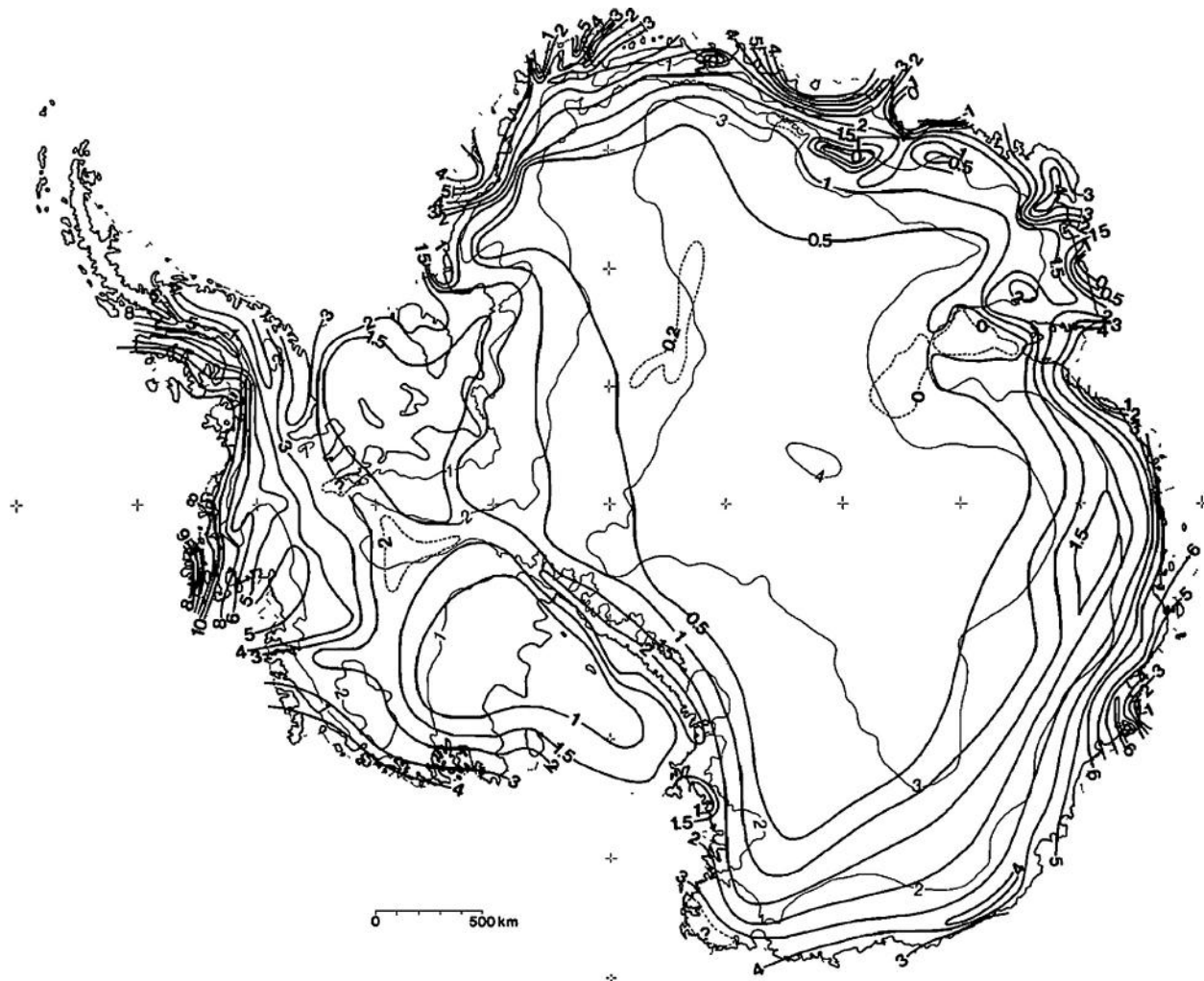


Figure 4 An estimate of snow accumulation across the Antarctic based on ice core data. Isopleths are in units of $100 \text{ kg m}^{-2} \text{ year}^{-1}$ (or, equivalently, 100 mm year^{-1}). Reproduced with permission from Bromwich, D.H., 1988. Snowfall in high southern latitudes. *Reviews in Geophysics* 26, 149–168. © American Geophysical Union.

measurements of accumulation are very similar in magnitude to those of precipitation, since there is little evaporation in the interior. However, they are not identical, because of the effects of blowing snow and summer melt in some areas. But with so few in situ measurements of precipitation, they have been used extensively as a proxy for precipitation. In Figure 4, it can be seen that no data are presented for the northern part of the Antarctic Peninsula because precipitation varies so rapidly in this area. The area of greatest precipitation is along the coast of the southern Bellingshausen Sea, where there is over 1 m water equivalent per year. This peak is found because of the high frequency of northerly airstreams bringing mild, moist air onto the coast. Other areas of high precipitation are found where there is frequent cyclonic activity, such as north of Enderby Land and along the coast of East Antarctica. Some of the smallest amounts of accumulation are found on the low-lying Ross and Ronne Ice Shelves. Inland of the coast, the amounts of accumulation drop very rapidly, so over most of East Antarctica there is less than 50 mm of accumulation per year.

A number of estimates have been made of the mean and total snow accumulation across the whole of the Antarctic ice sheet using glaciological data gathered in situ. Studies suggest that the mean accumulation is about 160 mm water equivalent per year, which is equivalent to a total input of approximately 2205 Gt year⁻¹.

The mechanisms behind precipitation are different across the Antarctic, with most precipitation in the coastal area coming from synoptic-scale weather systems. In the interior, most falls in the form of clear-sky precipitation, also known as 'diamond dust.' This is an almost continuous fallout of ice crystals from a thin veil of cirrus cloud covering the sky. Clear-sky precipitation has not been investigated extensively, but is thought to result from the cooling of air over the plateau and the formation of ice crystals as the precipitation descends into the cold near-surface layer. Just inland of the coast, there is a zone where both synoptic-scale weather systems and clear-sky precipitation play a role. Over Queen Maud Land, studies have shown that clear-sky precipitation falls on most days, but that a few major weather systems can give a significant fraction of the year's accumulation in a few days.

Climate Variability and Change

The high-latitude areas exhibit a greater degree of interannual and interdecadal climate variability than locations in the

tropics or midlatitudes. This is a result of the complex interactions between the atmospheric circulation and the cryosphere, including a number of positive-feedback mechanisms that amplify climate variability. However, our understanding of climate variability and change is limited in the Antarctic because of the shortness of the in situ records and the fact that most research stations are on the coast, with only the Vostok and Amundsen-Scott stations providing long records from the interior.

Standard deviations of the annual mean surface air temperatures for a number of stations are given in Table 1. These stations are located in different climatic regimes: at the South Pole (Amundsen-Scott Station), on the high interior plateau (Vostok), on the coast of East Antarctica (Mawson), and on the Antarctic Peninsula (Faraday/Vernadsky and Bellingshausen). Most of the stations have very similar variability of temperature, which is perhaps surprising considering the very different environments in which they are located. Variability of depression activity in the coastal area is to be expected, which would vary the temperatures. However, the figures show that conditions in the interior also vary from year to year as a result of changes in atmospheric circulation. But the station with the largest variability is Faraday/Vernadsky station on the western side of the Antarctic Peninsula. This station is located close to an area of large sea ice extent variability, and small changes in ice extent are amplified into much larger surface temperature variations.

The primary mode of Antarctic climate variability is the southern annular mode (SAM), which consists of synchronous pressure anomalies of opposite signs in midlatitudes and high latitudes. Thus, the SAM can be considered an index of the strength of the midlatitude westerlies. When pressures are below (above) average over Antarctica, the SAM is said to be in its high (low) index or positive (negative) phase. The SAM contributes a significant proportion of Southern Hemisphere climate variability (typically ~35%) from high-frequency to very low-frequency timescales. The SAM shows a high level of intrinsic variability, but is also affected by the amount of volcanic aerosol in the atmosphere, the concentration of greenhouse gases, and the Antarctic 'ozone hole.' The SAM has shown significant positive trends during autumn and summer over the past few decades, resulting in a strengthening of the circumpolar westerlies by about 15%. It has been suggested that the more positive SAM since about 1980 has mainly been a result of the 'ozone hole,' although during the first decade of the twenty-first century, the SAM

Table 1 Mean temperature data for selected Antarctic stations

Station	Latitude	Longitude	Elevation (m)	Period	Mean annual temperature (°C)	Standard deviation of the annual mean temperature (°C)	Mean January temperature (°C)	Mean July temperature (°C)
Vostok	78.5° S	106.9° E	3488	1958–2010	−55.3	0.9	−32.1	−66.8
Amundsen-Scott	90.0° S	—	2800	1957–2010	−49.4	0.7	−28.1	−60.0
Mawson	67.6° S	62.9° S	16	1955–2010	−11.2	0.7	+0.1	−18.0
Faraday/Vernadsky	65.4° S	64.4° W	11	1951–2010	−3.7	1.6	+0.7	−8.7
Bellingshausen	62.2° S	58.9° W	16	1969–2010	−2.3	0.8	+1.5	−6.5

became more neutral at a time when the ozone hole was still showing no clear sign of recovery.

The near-surface air temperature trends since 1951 at selected Antarctic stations are presented in Figure 5. The data show a strong dipole of change, with significant warming across the Antarctic Peninsula but with very small trends across the rest of the continent. The largest warming trends in the annual mean temperatures are found on the western and northern parts of the Antarctic Peninsula. Here,

Faraday/Vernadsky station has experienced the largest statistically significant ($<1\%$ level) trend in annual mean temperature of $+0.54\text{ }^{\circ}\text{C}$ per decade for the period 1950–2009. The rate of warming decreases away from Faraday/Vernadsky, with the long record from Orcadas on Signy Island, South Orkney Islands, having experienced a warming of only $+0.20\text{ }^{\circ}\text{C}$ per decade. However, it should be noted that this record covers a 100-year period rather than the 60 years for Faraday/Vernadsky.

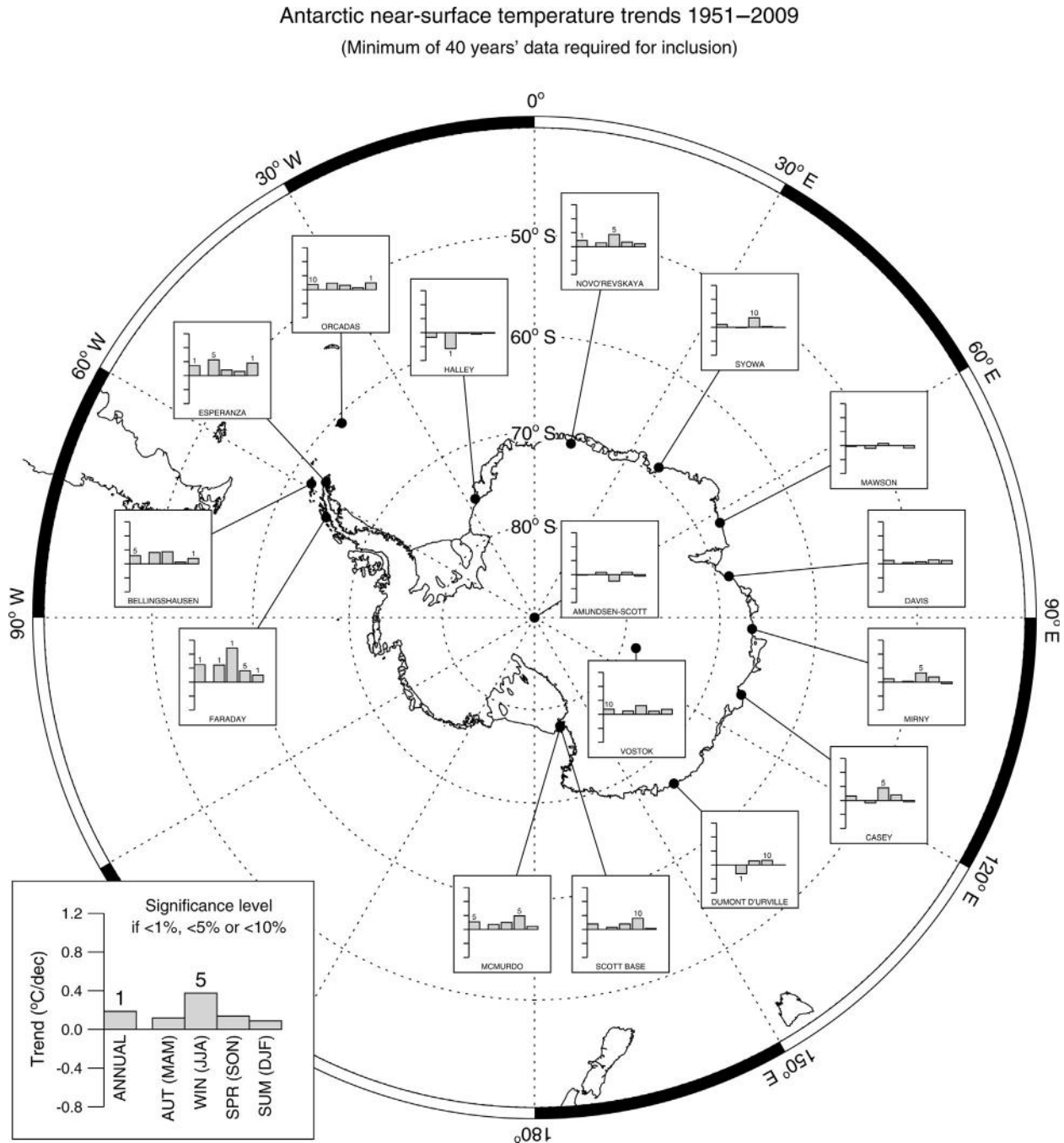


Figure 5 Antarctic near-surface temperature trends for 1951–2009. The scale in the bottom left corner indicates the warming or cooling over the full length of the station records. The numbers above the bars indicate the statistical significance by percentage.

The warming at Faraday/Vernadsky has been largest during the winter, with the temperatures increasing during that season by $+1.03^{\circ}\text{C}$ per decade over 1950–2009. In this area, there is a high anticorrelation during the winter between the sea ice extent and the surface temperatures, suggesting that more sea ice was present during the 1950s and 1960s, with a progressive reduction since that time.

Temperatures on the eastern side of the Antarctic Peninsula have risen most during the summer and autumn months, with Esperanza having experienced a summer increase in annual mean temperature of $+0.42^{\circ}\text{C}$ per decade over 1945–2009. This temperature rise has been linked to a strengthening of the westerlies as the SAM has shifted into a more frequent positive phase. Stronger winds have resulted in more relatively warm, maritime air masses crossing the peninsula and reaching the low-lying ice shelves on the eastern side.

Estimating temperature trends across the remote interior of the Antarctic is difficult because of the lack of staffed stations, and there has been an active debate over how far into West Antarctica the warming observed on the Antarctic Peninsula extends. With no long-term in situ records available, attempts have been made to estimate trends here using data from the coastal stations and knowledge of the spatial pattern of temperature variability. At the moment, it is thought that there has been a small warming across West Antarctic since the 1950s, but the magnitude is smaller than on the Antarctic Peninsula.

The Antarctic radiosonde temperature profiles suggest that there has been a warming of the troposphere and cooling of the stratosphere over the last 30 years, which is the pattern of change that would be expected from increasing greenhouse gas concentrations. However, the midtroposphere has warmed more in winter than anywhere else on Earth at this level. The radiosonde data show that regional midtropospheric temperatures have increased most around the 500 hPa level, with statistically significant changes of $0.5\text{--}0.7^{\circ}\text{C}$ per decade. The exact reason for such a large midtropospheric warming is not known at present. However, it has been suggested that it may, at least in part, be a result of greater amounts of polar stratospheric cloud during the winter.

Possible Future Change

The twenty-first century will be a period when we expect the Antarctic ‘ozone hole’ to recover, and we will possibly see stratospheric ozone levels returning to normal levels by 2060–70, but with greenhouse gas concentrations increasing. The ‘ozone hole’ has in many ways shielded the Antarctic from the impact of greenhouse gas concentration increases, but this will diminish over the coming decades.

Various scenarios of how greenhouse gas concentrations will increase during the twenty-first century have been considered by the Intergovernmental Panel on Climate Change (IPCC), but here we will examine how the Antarctic climate might evolve if CO_2 concentration increases to

720 ppm by 2100, which is one of the most frequently considered scenarios.

With a doubling of CO_2 over the twenty-first century, we expect the SAM to be even more predominantly positive in the future during all four seasons, further increasing the speed of the westerly winds over the Southern Ocean. Although a positive SAM has resulted in little warming around the coast of East Antarctica in recent decades, a doubling of greenhouse gas concentrations would result in a general warming across the continent and Southern Ocean. Estimates from the output of IPCC Fourth Assessment report models suggest that the surface warming averaged over the continent would be of the order of $3\text{--}4^{\circ}\text{C}$, which is approximately the same magnitude as that over other land areas of Earth. However, the models suggest that the largest warming in the Antarctic will be across the high-latitude areas of the Southern Ocean in winter as a result of the loss of sea ice and the greater fluxes of heat into the atmosphere. Here, the models suggest a temperature increase in excess of 0.5°C per decade.

Although the extent of Southern Hemisphere sea ice has increased slightly in recent decades, it is expected to decrease markedly during the coming century. Modeling studies have suggested that the ice extent could decrease by about 25% for the year as a whole, with a loss of around 50% in March and 20% in September.

If temperatures rise across the Antarctic over the next century, the air will be able to hold a greater amount of moisture. Since air masses are advected southward and forced up onto the Antarctic Plateau by depressions over the Southern Ocean, we can expect greater amounts of precipitation, especially in the coastal region. Estimating the increase in precipitation is difficult, but models suggest this could be of the order of 20% by 2100.

Further Reading

- Bromwich, D.H., 1988. Snowfall in high southern latitudes. *Reviews in Geophysics* 26, 149–168.
- King, J.C., Turner, J., 1997. *Antarctic Meteorology and Climatology*. Cambridge University Press, Cambridge. pp. 409.
- Parish, T.R., Bromwich, D.H., 2007. Reexamination of the near-surface airflow over the Antarctic continent and implications on atmospheric circulations at high southern latitudes. *Monthly Weather Review* 135, 1961–1973.
- Schwerdtfeger, W., 1984. *Weather and Climate of the Antarctic*. Elsevier, Amsterdam.
- Thompson, D.W.J., Solomon, S., 2002. Interpretation of recent Southern Hemisphere climate change. *Science* 296, 895–899.
- Turner, J., Anderson, P.S., Lachlan-Cope, T.A., Colwell, S.R., Phillips, T., Kirchgassner, A., Marshall, G.J., King, J.C., Bracegirdle, T.J., Vaughan, D.G., Lagun, V., Orr, A., 2009. Record low surface air temperature at Vostok Station, Antarctica. *Journal of Geophysical Research: Atmospheres* 114, D24102. <http://dx.doi.org/10.1029/2009JD012104>.
- Turner, J., Colwell, S.R., Marshall, G.J., Lachlan-Cope, T.A., Carleton, A.M., Jones, P.D., Lagun, V., Reid, P.A., Iagovkina, S., 2005. Antarctic climate change during the last 50 years. *International Journal of Climatology* 25, 279–294.
- Turner, J., Marshall, G.J., 2011. *Climate Change in the Polar Regions*. Cambridge University Press, Cambridge. pp. 434.

Arctic Climate

MC Serreze, University of Colorado, Boulder, CO, USA

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

Key features of the Arctic, such as its low mean annual air temperature, stable boundary layer, sea ice cover, permafrost, and snow cover, largely result from limited solar radiation receipts as compared to lower latitudes. The high albedo of snow and ice helps to maintain the Arctic in a low thermal energy state. However, regional features of the atmospheric and ocean circulation and surface modify primary latitudinal controls to result in a variety of climate conditions across the Arctic. Recent decades have seen pronounced changes in the Arctic, including reductions in sea ice extent, largest in September, and rises in surface air temperature, largest in autumn and winter.

Physical Features of the Arctic

The Arctic is defined as the region lying north of the Arctic Circle (66.5622° N latitude). In the north of the Arctic Circle, the sun remains above the horizon for 24 h (polar day) at least 1 day per year and below the horizon for 24 h (polar night) at least 1 day per year. On the Arctic Circle those events occur at the June and December solstices, respectively. The North Pole (90° N) experiences 6 months of polar day and 6 months of polar night.

Most of the region north of 70° N is occupied by the Arctic Ocean. Except for the sector straddling the date line between about 20° E and 20° W, the ocean is surrounded by land. Because of its largely landlocked nature, the Arctic Ocean is sometimes referred to as a Mediterranean-type sea. The dominant feature of the ocean surface is its floating sea ice cover. Northern Hemisphere sea ice extent, defined as all areas with an ice concentration (a fractional ice coverage) of at least 15%, waxes and wanes with the seasons (**Figure 1(a) and 1(b)**), typically ranging from upward of 15×10^6 km² in March to 7×10^6 km² or lower in September. These figures include seasonal ice in areas such as the Sea of Okhotsk, the Bering Sea, and Hudson Bay Ice that lie south of the Arctic Circle. The ice cover can be divided into first-year ice that is formed in a single ice growth season, and multiyear ice, which is ice that has survived one or more summer melt seasons. Any first-year ice present at the end of the melt season in September hence gets promoted to multiyear ice. Sea ice thickness ranges widely from a thin veneer to locally over 10 m. The probability distribution of sea ice thickness has a peak of about 3 m. While multiyear ice tends to be thicker than first-year ice, ridging and rafting can result in very thick first-year ice. Because of brine rejection during its formation, sea ice is nearly freshwater. For ice thicker than 1 m, a salinity of 2–6 parts per thousand is typical. Due to subsequent brine drainage by gravity, salinities for multiyear ice are considerably lower.

Apart from areas of land fast ice, the sea ice cover is in near-constant motion. The large-scale mean annual drift pattern is characterized by the clockwise Beaufort Gyre, centered over the Beaufort Sea north of Alaska, and the Transpolar Drift Stream, a motion of ice from the Siberian coast, across the pole and

through Fram Strait (the strait at about 80° N separating Greenland from the Svalbard Archipelago) (**Figure 2**). This pattern reflects roughly equal contributions by winds and surface ocean currents, the latter ultimately wind driven to a large extent. The mean annual sea ice circulation hence broadly resembles the mean annual sea level circulation of the atmosphere. Mean annual drift speed ranges from 1 to 5 cm s⁻¹ in the Beaufort Gyre and tends to increase along the Transpolar Drift Stream and through Fram Strait, where mean values may exceed 10 cm s⁻¹.

Snow cover atop the sea ice cover of the central Arctic Ocean is generally present for 10 months of the year. Most of the surrounding land surface of the Arctic is snow covered from October through May, with the duration of snow cover increasing with latitude. However, precipitation is generally scant in the Arctic. Some of the land, such as in the Canadian Arctic Archipelago, is classified as polar desert, often with less than 5% plant cover. In lower Arctic latitudes, the tundra commonly includes shrub vegetation of birch and willow. Permanent land ice is primarily restricted to the Greenland ice sheet (containing about 7 m of global sea level equivalent) and the ice caps and glaciers of the northeastern Canadian Arctic Archipelago, and the archipelagos of Svalbard, Novaya Zemlya, Severnaya Zemlya, and Franz-Josef Land. However, most Arctic land is underlain by perennially frozen ground (permafrost), overlain by an active layer exhibiting seasonal thaw. Permafrost acts as an impermeable barrier. As a result, many areas are covered by shallow thaw lakes in summer.

Atmospheric Circulation

Large-Scale Features

The primary feature of the northern high-latitude mid-tropospheric circulation is the polar vortex. The vortex is strongly asymmetric during winter (**Figure 3(a)**) with major troughs over eastern North America and eastern Eurasia and a weaker trough over western Eurasia (the Urals trough). A strong ridge is located over western North America. The lowest tropospheric pressure heights in winter are located over northern Canada. These features are related to orography, land-ocean distribution, and

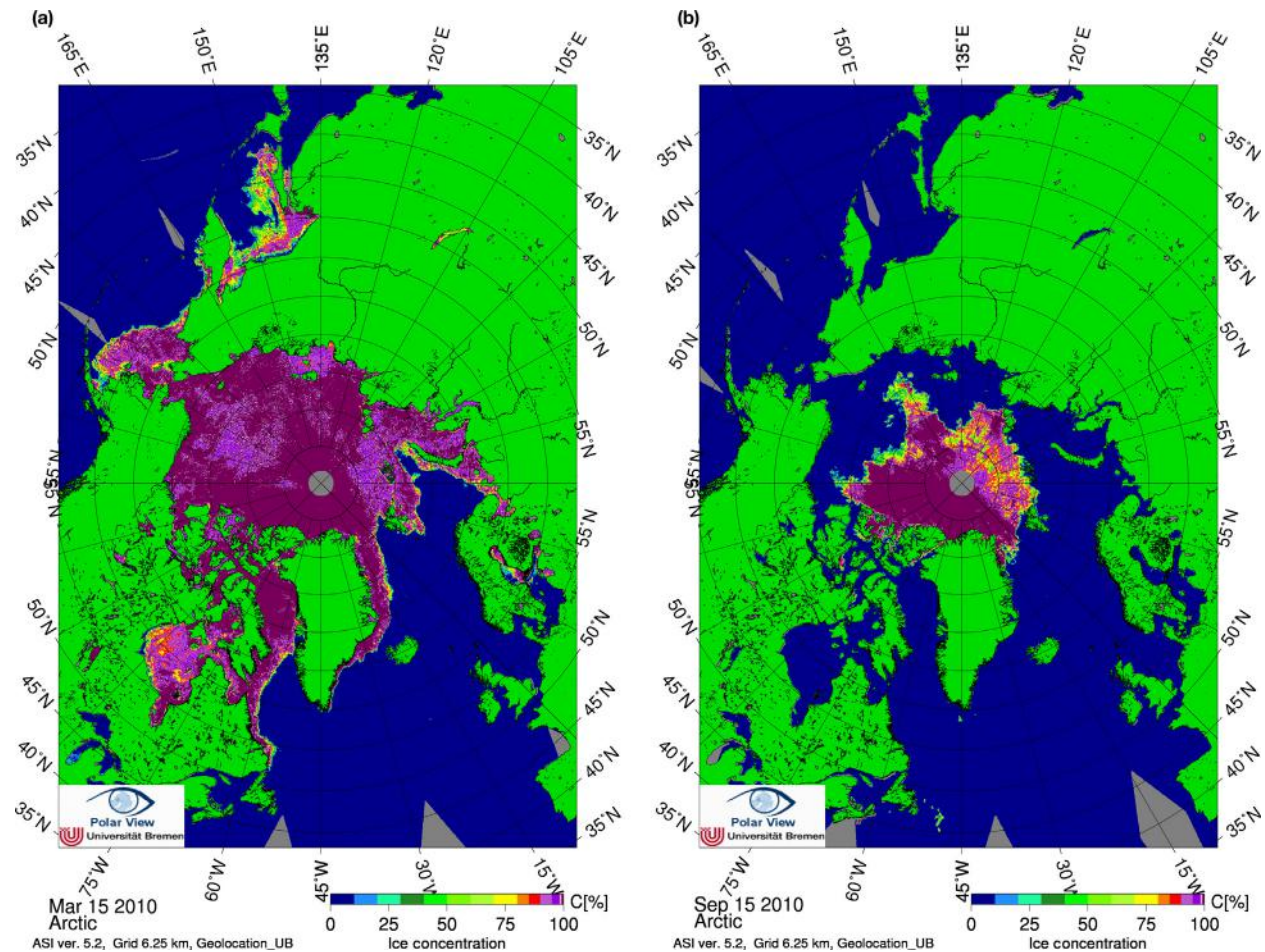


Figure 1 Arctic sea ice concentration for (a) 15 March 2010 and (b) 15 September 2010 based on data from the Advanced Microwave Sounding Radiometer (AMSR-E) aboard the NASA Aqua satellite (University of Bremen, Bremen, Germany).

radiative forcing. The polar vortex is much weaker during summer and is more symmetric than its winter counterpart, with the lowest pressure heights centered roughly over the pole (Figure 3(b)). This shift of the vortex core to over the pole is consistent with the presence of a melting sea ice cover, which strongly inhibits heating of the overlying atmosphere.

The mean annual sea level pressure field (Figure 2) masks large seasonal variability. The dominant sea level features of the mean winter circulation (Figure 4(a)) are the Icelandic Low off the southeast coast of Greenland, the Aleutian Low in the north Pacific basin, and the Siberian High over central Eurasia. In comparison to Figure 2, note the absence of a closed anticyclone over the Beaufort Sea; in winter the region is instead part of a saddle of high pressure extending from the Siberian High across the ocean and into northwestern Canada. The Beaufort Sea high is actually best expressed during spring. The Icelandic and Aleutian Lows are maintained by low-level thermal effects of the comparatively warm and largely ice-free underlying ocean, and position downstream of the major mid-tropospheric stationary troughs where eddy activity is favored. Regional cyclone development processes are also prominent in the vicinity of the Icelandic Low (see

Extratropical Cyclone Activity and Polar Lows). The Siberian High is a cold, shallow feature. The Icelandic and Aleutian Lows are much weaker during summer as compared to winter (Figure 4(b)). Summer also sees replacement of the Siberian High by mean low pressure, related in part to strong seasonal heating of the land surface. A weak high-pressure cell is found in the southern Beaufort Sea. An area of mean low pressure is also found centered near the pole.

Extratropical Cyclone Activity and Polar Lows

Winter cyclone activity is most prominent over the Atlantic side of the Arctic (Figure 5(a)). Atlantic-side cyclones typically take a northerly to easterly track and collectively represent part of the North Atlantic cyclone track. Activity peaks in the vicinity of the Icelandic Low. Cyclone development is favored because of temperature contrasts between the warm, northward flowing North Atlantic drift current and the cold, southward flowing East Greenland current, proximity to strong horizontal temperature gradients along the sea ice margin (see Figure 1(a)), and distortions to the atmospheric flow induced by the topography of the Greenland ice sheet. Synoptic events in the vicinity of the

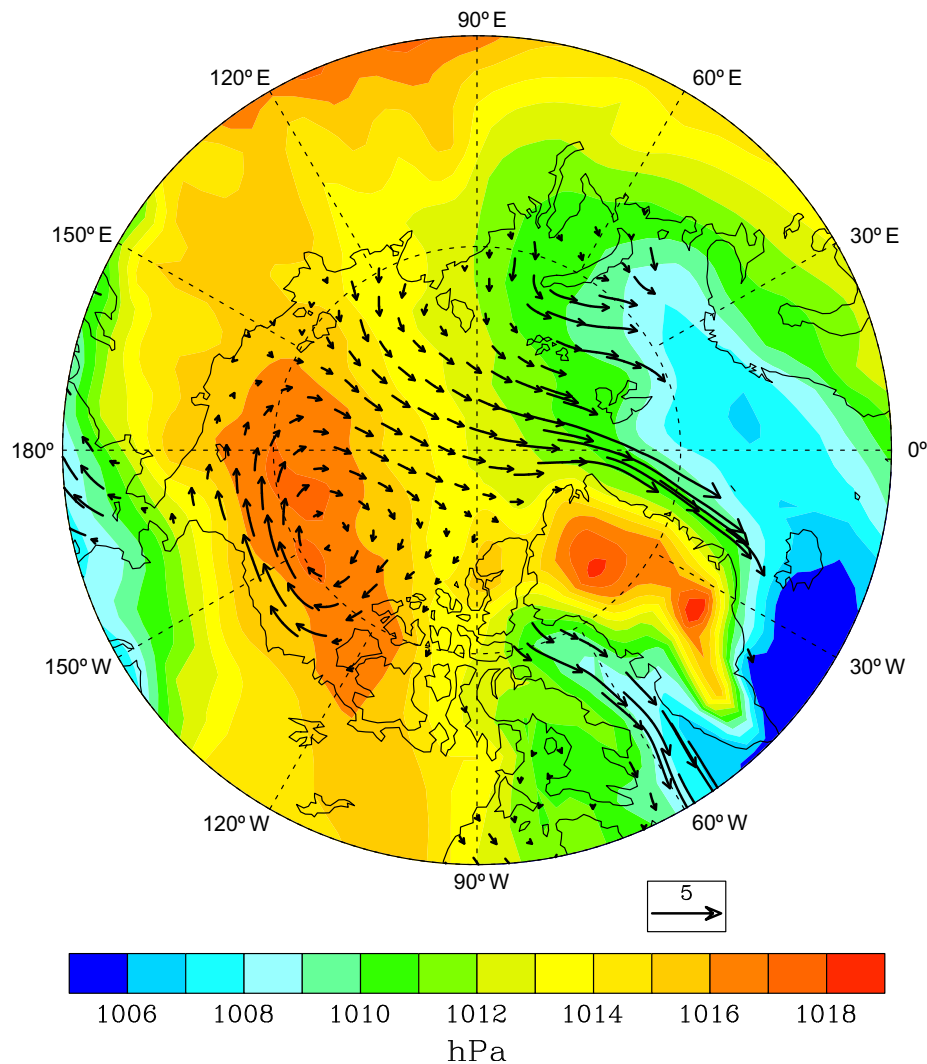


Figure 2 Mean annual sea ice drift (cm s^{-1}), based on data from drifting buoys, manned and unmanned camps and mean annual sea level pressure (hPa, see color scale). Reproduced from Serreze, M.C., Barrett, A.P., 2011. Characteristics of the Beaufort Sea high. *Journal of Climate* 24, 159–182.

Iceland Low include splitting (bifurcation) of cyclones at the southern tip of Greenland, with one center tracking northward along the west side of Greenland and the other tracking east of Greenland, orographic cyclogenesis in the lee of Greenland, and rapid deepening of existing systems that have migrated into the region from the south.

The North Atlantic cyclone track is weaker in summer, but cyclone activity increases over land (**Figure 5(b)**). Summer cyclogenesis is favored over northern Eurasia and over Alaska and extending southeast. A summer cyclone maximum is also found over the central Arctic Ocean, centered near the North Pole in the long-term mean. The summer cyclone pattern is associated with the influx of lows generated over the Eurasian continent and cyclogenesis over the Arctic Ocean itself. Systems entering the central Arctic Ocean from the outside, or formed within the region, migrate around the 500 hPa polar vortex, and decay within the cyclone maximum region or in close proximity. These processes help to maintain the weak

mean low-pressure cell centered near the pole seen in **Figure 4(b)**.

Polar Lows are cold season mesoscale systems that form within or at the leading edge of polar air streams. Polar Lows are particularly common in the Nordic Seas, the Labrador Sea, the Bering Sea, the Gulf of Alaska, and the Sea of Japan. Polar Lows are typically less than 500 km in diameter. They may intensify rapidly and surface wind speeds can reach hurricane force, but they also tend to be short-lived, existing 3–36 h. When moving over land or the sea ice cover, they tend to rapidly dissipate. They can be thought of as hybrid systems, typically having features of both baroclinic and convective in nature. A common feature of Polar Lows seen in satellite imagery is a spiral cloud (comma cloud) signature. Some systems develop a clear eye at the center similar to tropical cyclones. Rapid intensification of an intense Polar Low seems to require some element of convection. Preferred areas for Polar Low development mentioned above are those that are

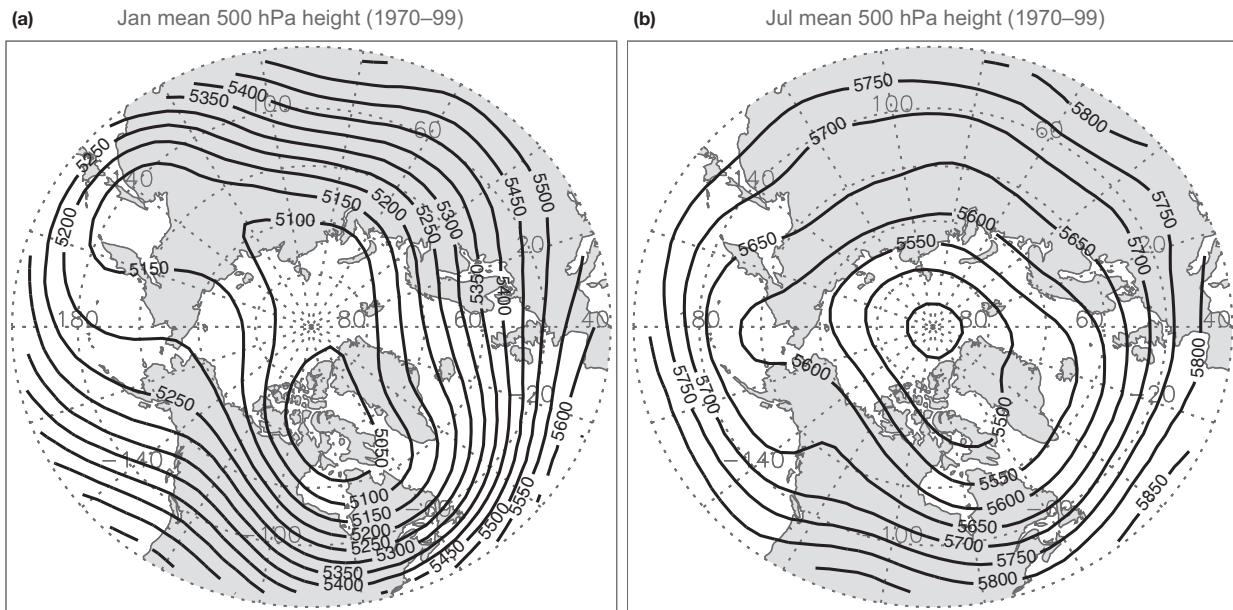


Figure 3 Mean 500 hPa height fields (m) for (a) January and (b) July based on NCEP/NCAR reanalysis data for the period 1970–99.

commonly subject to cold polar outbreaks, where cold continental air is advected over relatively warm open water – conditions favoring convection. This helps explain why Polar Lows are essentially cold season phenomena.

Frontal Activity

The concept of preferred geographical regions of frontal activity in northern high latitudes distinct from frontal activity in

middle latitudes (termed the ‘Arctic Frontal Zone’) has a long history. A maximum in frontal frequencies is found during summer along northern Eurasia from about 60–70° N, best expressed over the eastern half of the continent. A similar relative maximum is found over Alaska, which although best expressed in summer is present year-round. These features are clearly separated from the polar frontal zone in the middle latitudes of the Pacific basin. While some separation between high and middle latitude frontal activity is observed in all

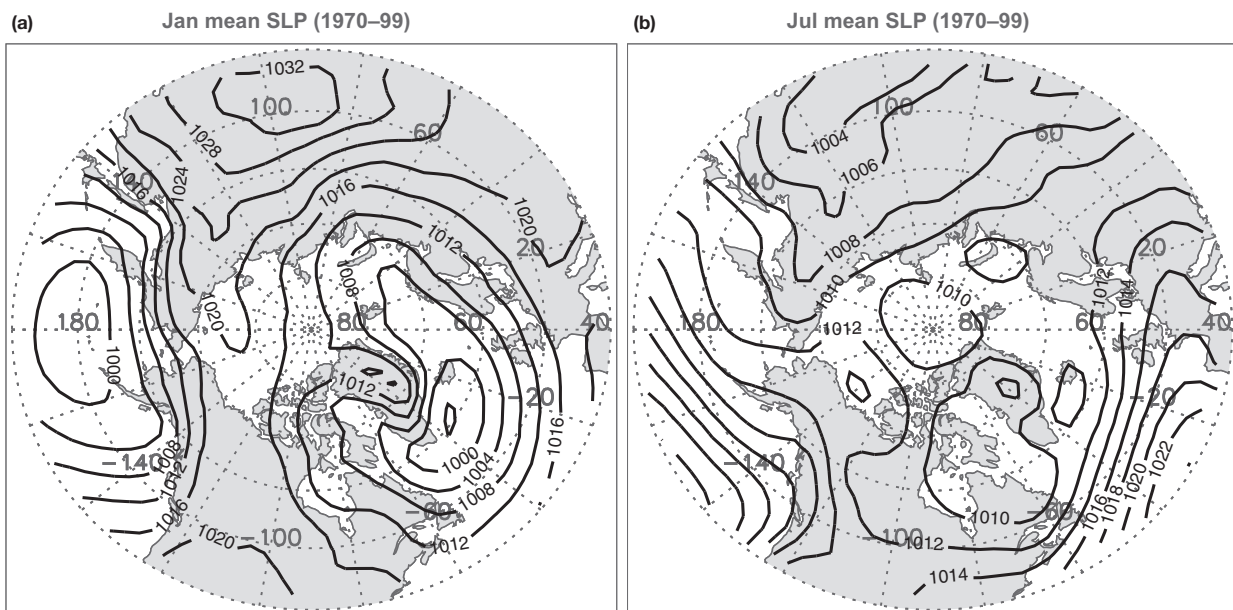


Figure 4 Mean sea level pressure fields (hPa) for (a) January and (b) July based on NCEP/NCAR reanalysis data for the period 1970–99. Note the different contour interval for January (4 hPa) compared to July (2 hPa).

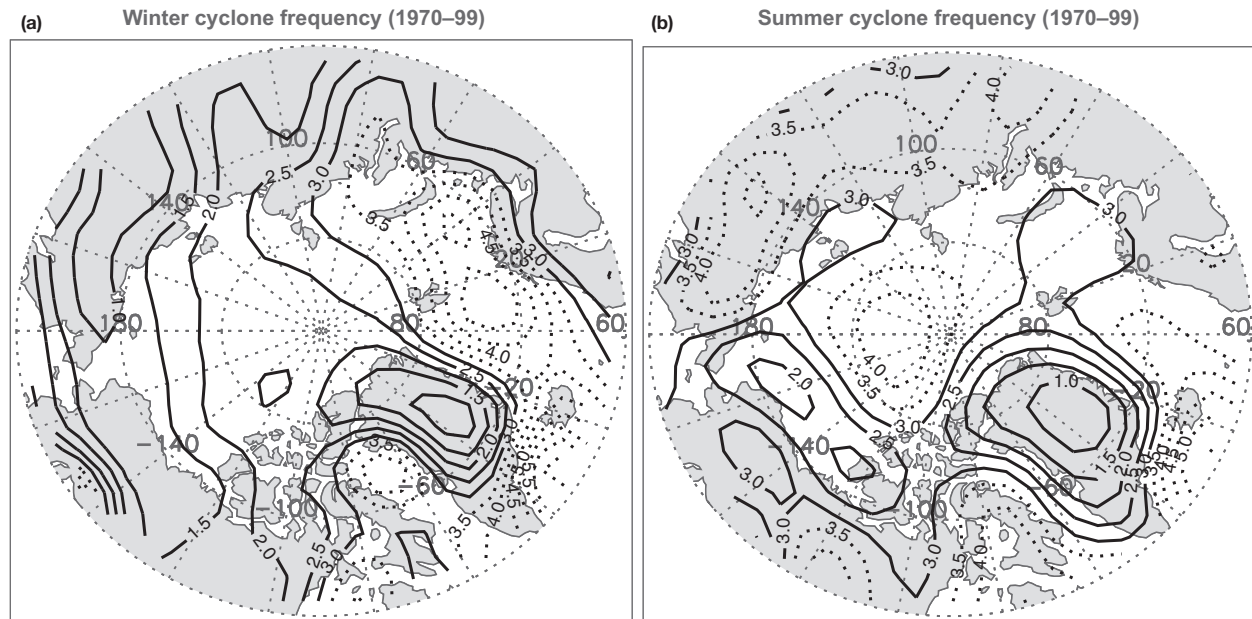


Figure 5 Average seasonal number of extratropical cyclone centers for (a) winter and (b) summer. Results are based on an automated cyclone identification algorithm applied to 6-h sea level pressure fields from the NCEP/NCAR reanalysis for the period 1970–99. Dotted contours are used to highlight areas with more than 3.5 systems per season.

seasons, the summer season is distinguished by the development of a mean baroclinic zone aligned along the Arctic Ocean coastline and associated wind maxima in the upper troposphere. While it has been postulated that the frontal zone arises from contrasts in energy balance between the tundra and boreal forest, it appears that coastal baroclinicity and focusing of the baroclinicity by orography play stronger roles. Regions of maximum summer frontal frequency correspond to preferred areas of summer cyclogenesis over Eurasia and Alaska.

Surface Energy Budget and Cloud Cover

Figure 6 shows typical monthly values of surface radiative flux components for the central Arctic Ocean. The outgoing longwave flux from the surface decreases from about 320 W m^{-2} in summer (when the sea ice surface is melting) to about 200 W m^{-2} in winter. The incoming longwave flux varies between 160 W m^{-2} in winter to 300 W m^{-2} in July. For all months, the net longwave flux is directed away from the surface. The downwelling shortwave (solar) flux is zero during the winter period of polar darkness, rising to about 300 W m^{-2} in June. Because of the high surface albedo (exceeding 0.80 when covered with fresh snow) comparatively little of the solar flux is absorbed by the surface. A fraction of the incoming solar radiation (typically 15%) penetrates into the snow and ice. Net all-wave radiation (net shortwave plus net longwave) is directed away from the surface from October through March and peaks in June at about 80 W m^{-2} . During winter there is a conductive heat flux through the sea ice to the surface. On an annual basis, the sensible and latent heat fluxes together balance 20–50% of the net radiation. During summer, the bulk of the net radiation is used to melt snow and ice. Locally, over areas of open water or

thin ice where strong temperature gradients are formed in the boundary layer in winter, upward sensible heat fluxes may reach 600 W m^{-2} . Condensate plumes emanating from wide ($>10 \text{ km}$) open water areas (leads) that extend to 4 km in the atmosphere have been observed in winter.

The fundamental difference between the surface energy budgets of the Arctic Ocean, glaciers, and tundra in summer is the portion of net radiation used to melt snow and ice. Once the snow is melted from the tundra, energy can be used in sensible heating of the atmosphere and evaporation (turbulent heat fluxes). The consumption of heat through melt on the ocean and glaciers is about four to six times larger than on the tundra. Consequently, sensible heat is transferred from the atmosphere to the surface of the oceans and glaciers, while it is carried from the surface to the atmosphere in the tundra. Evaporation is, on average, the most significant heat sink on tundra and is considerably larger than on the ocean and glaciers.

A key control on Arctic surface energy budgets is cloud cover. Winter cloud fractions range from 40 to 70%, greatest over the Atlantic side where extratropical cyclone activity is most common. Total cloud fractions rise to 70–90% in summer. There is a rapid increase between April and May, characterized by the development of extensive low-level stratus over the ocean. The seasonality of low-level stratus appears to be strongly controlled by the temperature-dependent formation of atmospheric ice. At temperatures below freezing, the saturation vapor pressure over ice is lower than over liquid water, such that ice particles grow at the expense of supercooled droplets. The concentration of ice crystals is smaller than that of cloud condensation nuclei. Hence a given mass of frozen condensate is distributed among smaller numbers of larger nuclei that grow rapidly to precipitable sizes when the

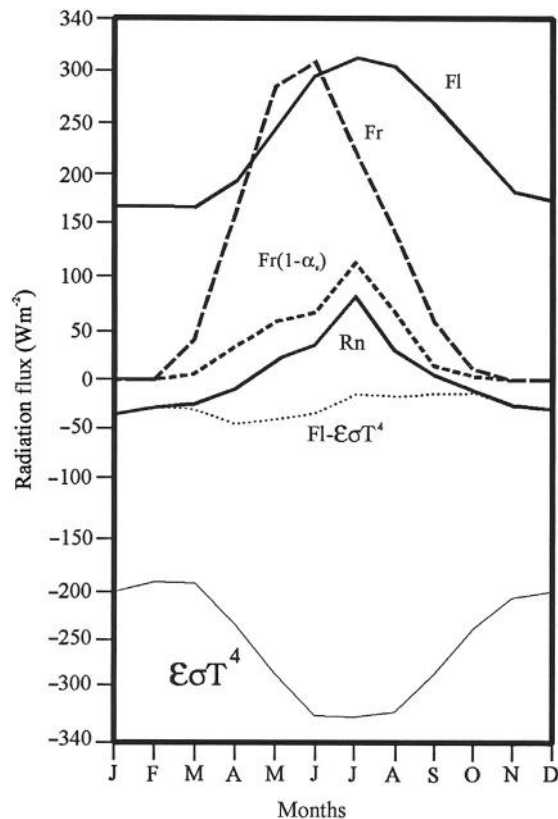


Figure 6 Monthly radiation balance components (W m^{-2}) for the central Arctic Ocean. Notation is as follows: Fl , incoming longwave radiation; Fr , incoming solar radiation; $Fr(1 - \alpha_s)$, solar radiation absorbed at the surface (α_s is surface albedo); $Fl - \epsilon\sigma T^4$, net longwave radiation; $\epsilon\sigma T^4$, outgoing (upward) longwave radiation; and R_n , net radiation. Reproduced from Barry, R.G., Serreze, M.C., Maslanik, J.A., Preller, R.H., 1993. The Arctic sea-ice climate system: Observations and modeling. *Reviews of Geophysics* 31, 397–422.

environment is supersaturated with respect to ice, disfavoring the development of stratus. This can be viewed as a preemptive dissipation of stratus, as the process can prevent the humidity of clear air from reaching saturation. This idea is supported by observations of ice crystal precipitation during the winter months. During summer, when temperatures are higher, the ice crystal scavenging processes are less effective and stratus is more likely to form and persist.

Except for a short period during summer, the net cloud radiative forcing is positive, meaning that clouds have a warming effect at the surface (net radiation at the surface is higher in the presence of clouds). This is basically because the increase in the downwelling longwave radiation flux due to the high emissivity of clouds exceeds the reduction in the downwelling solar radiation flux due to high cloud albedo and to lesser extent cloud absorption. However, cloud radiative forcing is a complex issue, with the sign and magnitude of the forcing depending on the solar flux above the clouds, cloud albedo, optical thickness and temperature, surface albedo, and multiple reflections between the surface and cloud base. Of course during winter, with little or no solar radiation, cloud radiative forcing is always positive.

Air Temperature and Boundary Layer

Winter surface air temperatures decrease sharply from the northern North Atlantic to the central Arctic Ocean (Figure 7(a)). The high temperatures over the Atlantic sector arise from poleward ocean heat transport, which keeps the region free of sea ice, and extensive cloud cover. The lowest winter air temperatures are found over east-central Eurasia in association with the Siberian High. Comparatively higher mean temperatures over the central Arctic Ocean reflect the effect of heat fluxes through areas of open water and thin ice. Low temperatures over the Greenland ice sheet reflect elevation.

Higher summer temperatures over land as compared to the ocean manifest latitude and the transformation of solar radiation into heating the atmosphere from turbulent and upward longwave radiation heat fluxes, as opposed to sea ice melt and seasonal replenishment of the ocean's sensible heat content. These differences in the surface energy budget over land versus ocean (see Surface Energy Budget and Cloud Cover) account for the pronounced temperature gradients along the coast (Figure 7(b)) which are in turn an expression of the summer Arctic frontal zone.

A characteristic feature of the Arctic atmosphere in winter, when there is little or no solar radiation, is a strong surface-based temperature inversion. Away from the Atlantic sector, winter inversions are typically 1000 m deep, with a temperature difference across the inversion layer of 10–12 °C. The winter Arctic inversion can be viewed in terms of an approximate longwave radiative equilibrium. The surface has a longwave emissivity close to one, while the longwave emissivity of the atmosphere is less than one. In longwave equilibrium, the atmosphere must then radiate at a higher physical temperature than the surface, which requires a temperature inversion. However, given that the surface and the atmosphere both radiate to space, maintaining the inversion requires a transport of heat from the south. This provides only a first-order view. Inversion depth and strength vary widely in response to local topographic conditions, winds, cloud cover, and surface fluxes. For example, inversions over the central Arctic Ocean tend to be weaker than over land due to heat fluxes through areas of open water and thin ice. Inversions are also common in summer, although they are weaker than their winter counterparts and are typically separated from the surface by a mixed layer. However, over the Arctic Ocean, shallow surface-based inversions are very common due to sea ice melt, which keeps the surface skin temperature at the melting point, hence leading to the downward sensible heat flux noted earlier.

Hydrologic Budget

Precipitation and Precipitation Minus Evaporation ($P - E$)

Precipitation in the Arctic is difficult to measure accurately because of gauge undercatch of blowing snow, changes in and differences between countries in instrument types and reporting practices, and the sparse precipitation monitoring network. Figure 8 shows the distribution of annual precipitation based on a gridded climatology compiled from several data sources. The highest totals are found off the southeast coast of Greenland (locally >2400 mm) with amounts decreasing

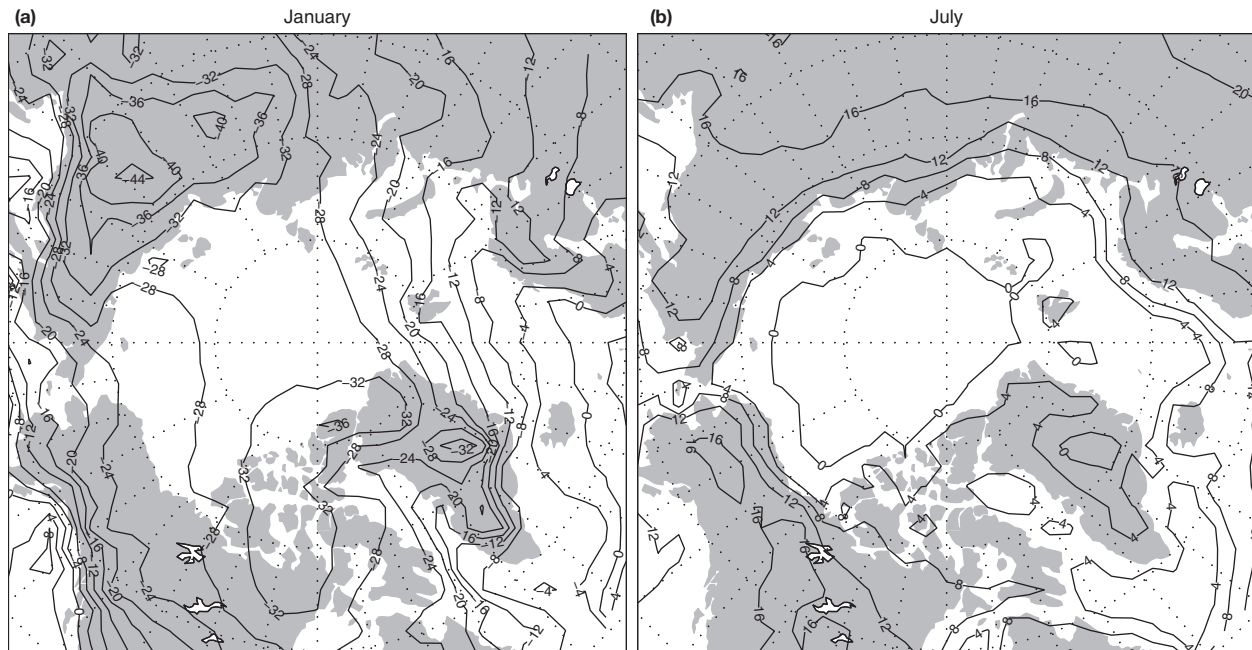


Figure 7 Mean surface air temperature ($^{\circ}\text{C}$) for (a) January and (b) July, based on the University of Washington International Arctic Buoy Programme/ Polar Exchange at the Sea Surface (IABP/POLES) data set. Reproduced from Rigor, I.G., Colony, R.L., Martin, S., 2000. Variations in surface air temperature observations in the Arctic, 1979–1997. *Journal of Climate* 13, 896–914.

northeast to about 400 mm in the Kara Sea. This pattern manifests the pattern of cyclone activity shown in [Figure 5\(a\)](#). High totals are also found over southern Alaska. The lowest annual totals (<200 mm) encompass the Beaufort Sea and the Canadian Arctic Archipelago; these islands are primarily classified as polar desert. The winter pattern is qualitatively similar

to that seen in the annual mean. For example, mean January precipitation ranges from over 200 mm in the northern North Atlantic to less than 10 mm over northern Canada and east-central Eurasia. Summer precipitation is more uniform across the Arctic with markedly higher totals as compared to winter over land areas. This is consistent with seasonal changes in synoptic activity (compare [Figure 5\(a\)](#) and [5\(b\)](#)). Convective precipitation is not uncommon over Arctic land areas during summer.

Winter precipitation is largely stored in the snowpack. Maximum spring snow depths are highly variable due to differences in precipitation, temperature, topographic setting, and redistribution by wind. Values of 20–50 cm over the Arctic Ocean and 40–70 cm over the subarctic can be considered typical. Mean hydrographs for Arctic rivers exhibit a late spring to early summer peak in discharge due to melt of the snowpack.

Direct estimates of evaporation are scanty. However, large-scale estimates of precipitation minus evaporation ($P - E$) (net precipitation) can be obtained through evaluation of the atmospheric vapor flux convergence. Estimated mean annual $P - E$ ([Figure 9](#)) is typically 150–300 mm over land, 200 mm over the central Arctic Ocean and over 1000 mm in the vicinity of the Icelandic Low. Although precipitation over much of the land area peaks in summer, $P - E$ for this season (not shown) tends to be small or even negative. Negative values imply net drying.

Freshwater Budget

The freshwater balance of the Arctic Ocean is a topic of continuing interest. Freshwater storages and transports have typically been calculated with respect to a reference salinity of 34.8 parts per thousand, which approximates the average bulk salinity of the Arctic Ocean. Water lower than this reference

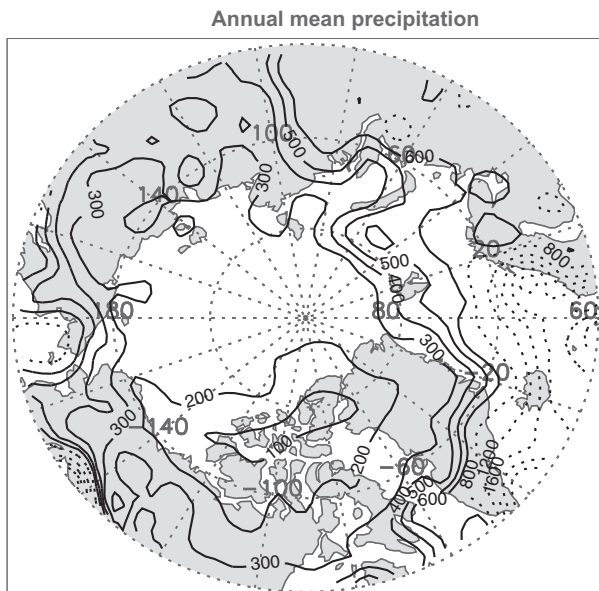


Figure 8 Mean annual precipitation (mm) with estimated adjustments for wind-induced gauge undercatch, changes in instrument types and differences in observing methods. Dotted contours are used to highlight areas with precipitation exceeding 800 mm. Compiled from data provided by Groisman, P., Yang, D., Eischeid, Willmott, C.

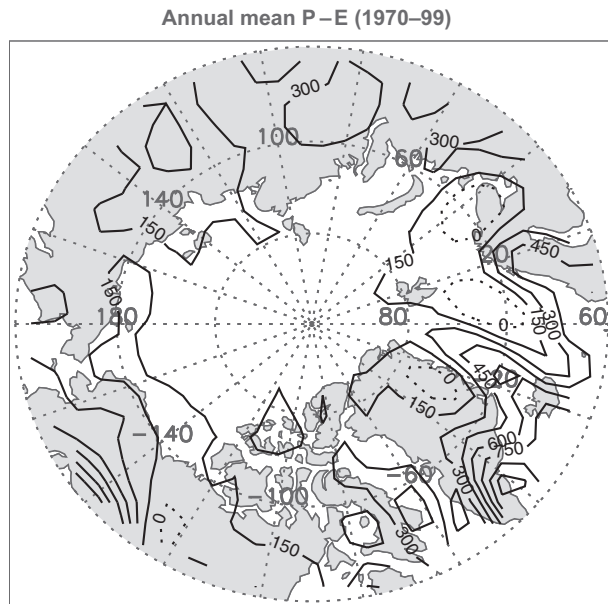


Figure 9 Mean annual precipitation minus evaporation ($P - E$) (in mm) based on calculations of the vapor flux convergence using NCEP/NCAR reanalysis data for the period 1970–99. Areas with negative $P - E$ are indicated with dotted contours.

salinity is said to contain freshwater; the lower the salinity, the greater the freshwater content per unit volume.

The total annual freshwater input to the Arctic Ocean is estimated at about 8500 km^3 . The primary inputs are river runoff (38%), the import of fairly low-salinity water through Bering Strait (30%) and positive $P - E$ over the Arctic Ocean itself (24%). The Arctic Ocean is unique in receiving runoff from four of the world's major rivers (the Ob, Yenisei, and Lena in Eurasia and the Mackenzie in North America). These riverine, atmospheric, and oceanic freshwater sources collectively help to maintain a fairly fresh surface layer that extends down to about 200 m, which is often well-mixed down to about 50 m. Relatively warm and salty waters of Atlantic origin are found between 200 and 900 m depth, which if brought to the surface would quickly melt the sea ice cover. However, at low water temperatures of the Arctic Ocean, the density structure is determined by salinity. Hence the fresh surface layer suppresses vertical mixing with the Atlantic layer, and allows sea ice to form readily in winter.

Freshwater export out of the Arctic Ocean and into the North Atlantic is primarily via Fram Strait in the form of low-salinity sea ice (25%) and liquid water (26%), and through the channels of the Canadian Arctic Archipelago, primarily in liquid form (25%). Freshwater export through Fram Strait in particular is believed to impact on the overturning cell of the global ocean through influencing convection in the subarctic gyres which in turn feed the North Atlantic.

Mean annual freshwater input to the Arctic Ocean is roughly an order of magnitude less than the mean freshwater storage in the Arctic Ocean of about $84\,000 \text{ km}^3$. This storage is represented by both sea ice and fairly low salinity near surface waters. Assuming steady state, the relative magnitudes of the mean

storage versus total inputs (or outputs) imply a mean residence time of freshwater in the Arctic Ocean of about a decade.

Variability and Change in Arctic Climate

Arctic climate exhibits pronounced variability on interannual to decadal scales. A major source of variability is associated with the phase of the North Atlantic Oscillation (NAO), which describes mutual strengthening and weakening of the Azores High and the Icelandic Low. Under the positive mode of the NAO (a deep Icelandic Low), positive temperature anomalies are found over the Eurasian Arctic with negative anomalies over northeastern Canada and the northern North Atlantic. In turn, the North Atlantic cyclone track extends deeper into the Arctic Ocean. Broadly opposing anomalies are associated with negative NAO states. The NAO can be viewed as the North Atlantic component of the Arctic Oscillation (AO) (also called the Northern Annular Mode (NAM)). The AO represents the leading empirical orthogonal function of monthly sea level pressure anomalies poleward of 20° N . Pressure variability associated with the AO is characterized by a primary center of action over the Arctic Ocean, strongest in the vicinity of the Icelandic Low, and opposing anomalies in midlatitudes of the Pacific and Atlantic basins. Changes in the AO index hence manifest a transfer of atmospheric mass between the Arctic and middle latitudes. Time series of the AO and NAO are highly correlated. Arctic climate variability is also linked to the El-Niño Southern Oscillation (ENSO), particularly as it relates to variability in the strength and location of the Aleutian Low, the Pacific North American (PNA) teleconnection, and other modes of variability such as the North Pacific Oscillation (NPO).

The period from the 1970s onward has seen pronounced change in the northern high-latitude environment. Increases in surface air temperature encompass all seasons but are strongest in autumn and winter, and are larger than increases observed for the globe as a whole. Paleoclimate evidence suggests that Arctic temperatures of the late twentieth century and onward are the highest of at least the past 2000 years. Analysis of satellite data available since 1979 documents negative linear trends in sea ice extent for all months, with the strongest trend in September, the end of the summer melt season (13% per decade over 1979–2012 relative to 1979–2000 mean; see <http://nsidc.org/>) (Figure 10). A key driver of this seasonal asymmetry in trends is that the spring ice cover is increasing dominated by relatively thin first-year ice, with less of the generally thicker multiyear ice. As it takes less energy to melt out thin ice, the thinner the ice in spring, the lower the ice extent at the end of summer. Thin spring ice also fosters earlier exposure of dark open water areas in summer which strongly absorb solar energy, leading to a warming ocean that fosters more melt. Other changes include increased vegetation growth, with areas of tundra replaced by shrub vegetation, warming and thawing of permafrost, increased discharge of Arctic-draining rivers in Eurasia, and increased coastal erosion due to a combination of sea ice retreat that allows for stronger wave action, and a warming ocean.

Climate models project that the effects of rising concentrations of atmospheric greenhouse gases will be especially

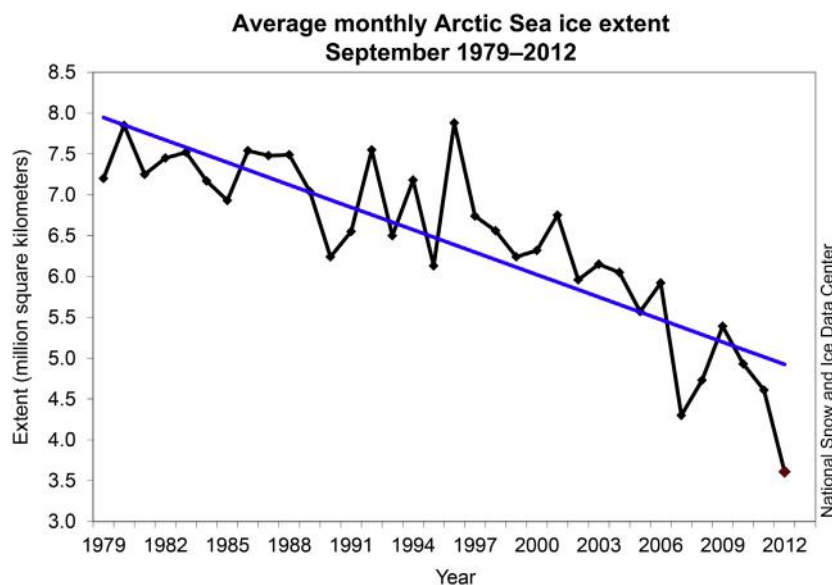


Figure 10 September sea ice extent, 1979–2012, with linear least squares fit (blue line), based on analysis of satellite passive microwave data distributed by the National Snow and Ice Data Center, Boulder, Colorado.

strong in the Arctic due to feedbacks in which variations in sea ice and snow extent, the stability of the lower troposphere, and thawing of permafrost play key roles. However, regional patterns of Arctic warming differ greatly among simulations. Projected warming is generally strongest for autumn and winter, largely because of the delayed growth of sea ice, which allows for large fluxes of heat from the ocean to the atmosphere. Retreat of snow cover and sea ice is accompanied by increased winter precipitation. Observed changes in the Arctic environment over the past decade are generally viewed as reflecting the combined influences of natural variability in patterns of atmospheric and oceanic circulation having strong regional expressions, superposed upon a general warming linked to fossil fuel burning.

See also: **Cryosphere:** Permafrost; Sea Ice; Snow (Surface). **Climate and Climate Change:** Climate Variability: North Atlantic and Arctic Oscillation. **Global Change:** Climate Record: Surface Temperature Trends. **Middle Atmosphere:** Polar Vortex. **Radiation Transfer in the Atmosphere:** Cloud-Radiative Processes. **Synoptic Meteorology:** Extratropical Cyclones; Fronts; Polar Lows.

Further Reading

ACIA, 2005. Impacts of a Warming Arctic: Arctic Climate Impact Assessment. Cambridge University Press, Cambridge, UK.

- Beesley, J.A., Moritz, R.E., 1999. Toward an explanation of the annual cycle of cloudiness over the Arctic Ocean. *Journal of Climate* 12, 395–415.
- Bekryaev, R.V., Polyakov, I.V., Alexeev, V.A., 2010. Role of polar amplification in long-term surface air temperature variations and modern arctic warming. *Journal of Climate* 23, 3888–3906.
- Cullather, R.I., Bromwich, D.H., Serreze, M.C., 2000. The atmospheric hydrologic cycle over the Arctic basin from reanalyses. Part I: Comparison with observations and previous studies. *Journal of Climate* 13, 923–937.
- Curry, J.A., Rossow, W.B., Randall, D., Schramm, J.L., 1996. Overview of Arctic cloud and radiation characteristics. *Journal of Climate* 9, 1731–1764.
- Dimitrenko, I.A., Polyakov, I.V., Krillov, S.A., et al., 2008. Toward a warmer Arctic Ocean: Spreading the early 21st century Atlantic Water warm anomaly along the Eurasian Basin margins. *Journal of Geophysical Research* 113, C05023. <http://dx.doi.org/10.1029/2007JC004158>.
- Kaufman, D.S., Schneider, D.P., McKay, N.P., et al., 2009. Recent warming reverses long-term Arctic cooling. *Science* 325, 1–4.
- National Snow and Ice Data Center <http://nsidc.org/> [accessed 23.03.12].
- Ohmura, A., 1984. Comparative energy balance study for Arctic tundra, sea surface, glaciers and boreal forests. *GeoJournal* 8, 221–228.
- Renfrew, I.A., 2003. Polar Lows. In: Holton, J.R., Curry, J.A., Pyle, J.A. (Eds.), *Encyclopedia of Atmospheric Sciences*. Academic Press, London, UK/San Diego, CA, pp. 1761–1768.
- Screen, J.A., Simmonds, I., 2010. The central role of diminishing sea ice in recent Arctic temperature amplification. *Nature* 464, 1334–1337.
- Serreze, M.C., Barrett, A.P., Slater, A.G., et al., 2006. The large-scale freshwater cycle of the Arctic. *Journal of Geophysical Research* 110, C11010. <http://dx.doi.org/10.1029/2005JC003424>.
- Serreze, M.C., Barry, R.G., 2005. *The Arctic Climate System*. Cambridge University Press, Cambridge, UK.
- Serreze, M.C., Lynch, A.H., Clark, M.P., 2001. The Arctic frontal zone as seen in the NCEP/NCAR reanalysis. *Journal of Climate* 14, 1550–1567.
- Stroeve, J., Serreze, M., Drobot, S., et al., 2008. Arctic sea ice extent plummets in 2007. *EOS. Transactions, American Geophysical Union* 89, 13–14.
- Thompson, D.W.J., Wallace, J.M., 2001. Regional climate impacts of the Northern Hemisphere Annular Mode. *Science* 293, 85–89.

Arctic Haze

LM Russell, Scripps Institution of Oceanography, University of California, San Diego, La Jolla, CA, USA

GE Shaw, Geophysical Institute, University of Alaska, Fairbanks, AK, USA

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

Arctic haze is composed of aerosols – particles and gases – that have increased concentrations in the lower atmosphere across the Arctic regions in late winter and spring. Arctic haze is derived from industrial and wildfire emissions released in continental regions surrounding the Arctic and then is transported to the Arctic over spatial scales of several thousands of kilometers. The haze particles contain sulfate, sea salts, crustal materials, organic components, black carbon, and trace amounts of heavy metals, some of which are characteristic of industry-specific pollution sources.

Introduction

Arctic haze is composed of aerosols – particles and gases – that have increased concentrations in the lower atmosphere across the Arctic regions in late winter and spring. Arctic haze is derived from industrial and wildfire emissions released in continental regions surrounding the Arctic and then is transported into the Arctic over spatial scales of several thousands of kilometers. The gases and particles that constitute arctic haze reside mainly near the meteorological boundaries of the arctic front, a system surrounding the North Pole, reaching its southernmost extent in late winter (February–March). The Arctic is highly sensitive to warming from greenhouse gases and is integral to the global temperature gradient that drives the main circulation of the atmosphere. As opposed to greenhouse gases, the arctic haze includes components that warm and cool the atmosphere. But the haze processes are very complex and the relative importance of the arctic haze remains unclear.

During late winter the system of arctic haze is roughly as large in areal extent (the region encircled in [Figure 1](#)) as the African continent. After the spring polar sunrise, the ‘haze’ can be visible to the eye, especially when viewed edgewise from an aircraft. The haze reduces direct solar radiation at the surface, whitens the sky, and causes slight warming of the Earth–atmosphere system.

The haze was first noted by Murray Mitchell Jr, in 1956, when flying on ‘Ptarmigan’ weather reconnaissance missions in the Arctic. Mitchell recognized that the constituents were probably of the same order as the wavelength of visible light. He speculated that the origin of the unknown haze particles might be quite distant since small submicrometer particles have long lifetimes in the atmosphere. Investigations of chemical composition and optical properties of arctic haze were conducted in the early 1970s and led to the recognition of an arctic-wide pollution phenomenon. The general picture of arctic haze being pervasive, extensive, and caused by circum-polar industrial emission unfolded during research carried out in the 1980s and has continued through the 2010s with more specific characterization of chemical components in particles and more detailed attribution of contributing sources and meteorology.

If one considers the geographic location of pollution sources and the nature of atmospheric circulation at high

latitudes (see the next section), it is apparent ([Figure 1](#)) that the more numerous and more northern Eurasian pollution sources are likely to contribute larger amounts of pollution to the Arctic than the sources in North America. The lack of sunlight, coupled with the snow-covered surface that reflects the little sunlight received back to space, results in especially strong cooling of the surface layers and buildup of surface-based temperature inversions. Mixing is greatly inhibited under such circumstances. In much of the Arctic, most especially around the Siberian High, cloudiness is thin and sparse. These factors increase the residence time of pollutants in the Arctic, resulting in prolonged high concentrations of particles in the arctic atmosphere.

Seasonal and Geographic Variations of the Arctic Haze and Meteorological Transport

In winter, the arctic air mass extends throughout the high Arctic and down over Eurasia and North America. Arctic haze has been described as having a ‘dome,’ 7–8 km deep over the pole with shallow tongues of air, 0–5 km deep spilling southward over the land masses. This air mass is statically stable because of the strong temperature inversions and has relatively small cloud water or ice content. There is a close connection between outbreaks of arctic haze and transport along or within anticyclonic pathways in the atmosphere. In the polar regions since the arctic air is stable, particles may remain airborne for weeks. The connection between arctic haze and anticyclonic conditions suggests that a reason why haze lasts so long is partly due to the lower frequency of rain and snow cleansing mechanisms.

Transport of pollution episodes into the Arctic is often controlled by the midnorthern Eurasian high, extending over Kamchatka and another cell over the Alaska area, or steered by cyclonic systems in the Barents Sea. Studies of the meteorological patterns indicate that the transport of arctic haze is associated with a quasi-stationary ‘blocking’ pattern in the atmosphere. This provides conditions for poleward transport of midlatitudinal air pollution, particularly from the European and northern Asian sectors (Eurasia). The seasonal variation of this blocking, along with scavenging and other removal processes, is important for understanding the annual cycle of the arctic air pollution.

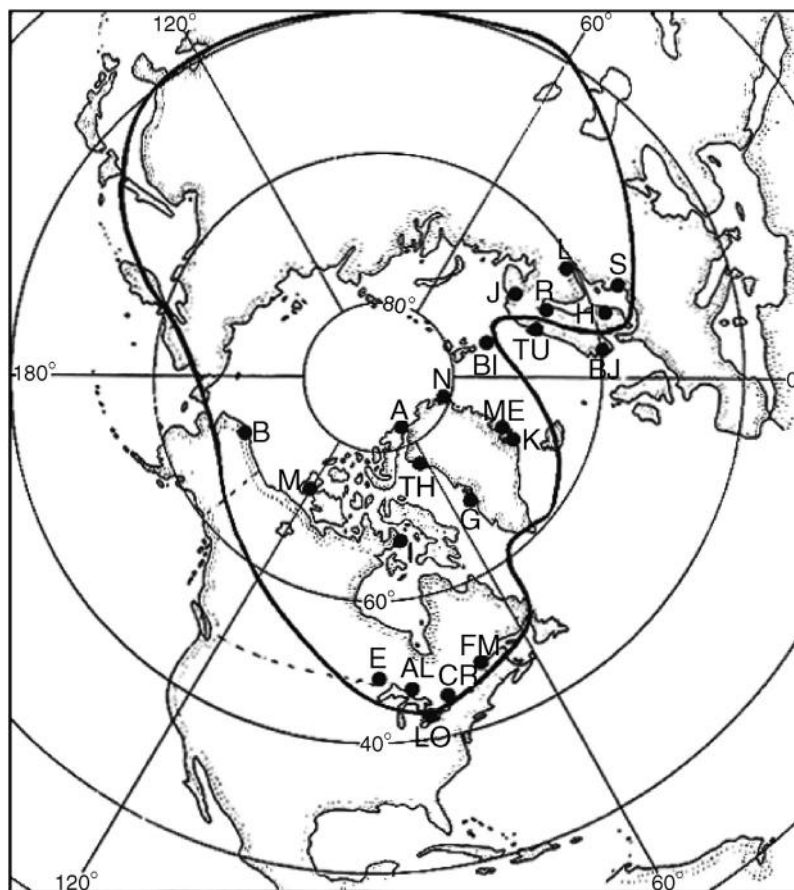


Figure 1 Locations of large sources of sulfur dioxide emissions (in units of millions of metric tons) in the high-latitude Northern Hemisphere that influence the Arctic. Map courtesy of Leonard Barrie.

Arctic pollution in midwinter has substantial contributions from sulfur dioxide. This gas is oxidized to form sulfate (SO_4^{2-}) in particles by both photochemical processes in spring and particle processes in winter. Combining the increased residence time with springtime sunlight makes the polar atmosphere work like a large chemical reactor, producing increasing mass of particles that scatter light to form the persistent haze.

Low-level liquid cloud cover becomes increasingly important throughout the Arctic as the summer progresses and areas of open water increase. Cloud decks form when warm moist air flows over cold icepack; drizzle often accompanies the cloud. The summer arctic stratus is one of the most pervasive and persistent cloud systems on the planet. The turnover and almost constant drizzle is instrumental in cleansing the arctic atmosphere, reducing concentrations of arctic haze to insignificant levels. As a result, the concentration of arctic haze undergoes a strong seasonal variation with maximum in spring and minimum in summer. **Figure 2** shows the strong seasonal variation of sulfate aerosol concentration sampled in the Canadian Arctic over nearly 30 years. **Figure 3** shows similar seasonal cycles and long-term trends for Barrow, Alaska. Notable is the reduction in the contribution of sulfate. This 'arctic air mass' picture provides a rough understanding of the more persistent and larger features of arctic air pollution.

The map of annual emissions of sulfur dioxide with superimposed arctic air mass (**Figure 1**) helps identify the major sources and currents of pollution-derived material affecting the Arctic. Thus, on the basis of the relatively strong source region in the central and western Eurasia sector, the occurrence of a deep lobe of the arctic air mass over much of this source, the occurrence of a poleward flowing circulation over this source area, and the absence of precipitation, clouds, and turbulence along the pathway, it is clear that Eurasia is of greater importance than North America as a source region for the arctic haze.

Chemical Composition of the Arctic Haze

The chemical composition of aerosol has been measured at various arctic monitoring sites since the 1970s in order to identify the origin of the hazy layers. Although sulfate contributes a major fraction of particle mass, the haze also contains sea salts, crustal materials, organic components, black carbon, and trace amounts of heavy metals, some of which are characteristic of industry-specific pollution sources (**Figure 4**). Chemical measurements through 2008 have continued to show the strong seasonal variation (discussed in the last section), with maximum occurrence of the arctic haze in the

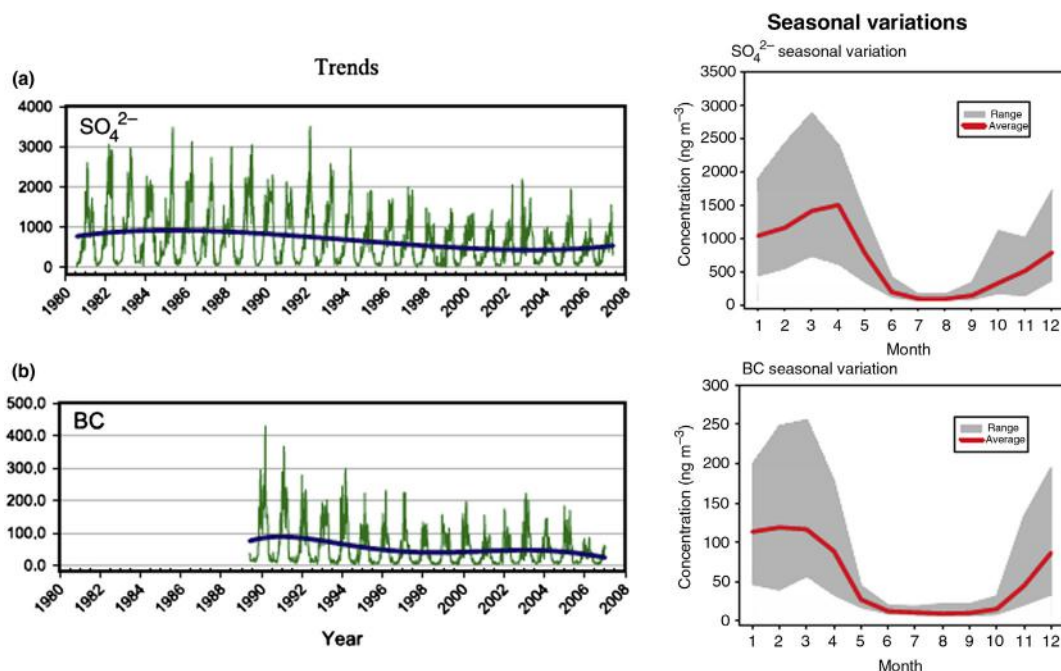


Figure 2 Trends and seasonal variations of (a) sulfate and (b) black carbon (BC) aerosols observed at the Dr Neil Trivett Global Atmospheric Watch Observatory, Alert, Nunavut, Canada, in ng m^{-3} . Sulfate is from the ion chromatography analysis of weekly integrated high-volume filter samples, and the BC is from weekly averaged light absorption measurements (aethalometer) adjusted to elemental carbon concentrations derived from a thermal method. Reprinted from Gong, S. L., Zhao, T. L., Sharma, S., Toom-Saunty, D., Lavoue, D., et al., 2010. Identification of trends and inter-annual variability of sulfate and black carbon in the Canadian High Arctic: 1981–2007. *Journal of Geophysical Research – Atmosphere* 115, doi:10.1029/2009jd012943.

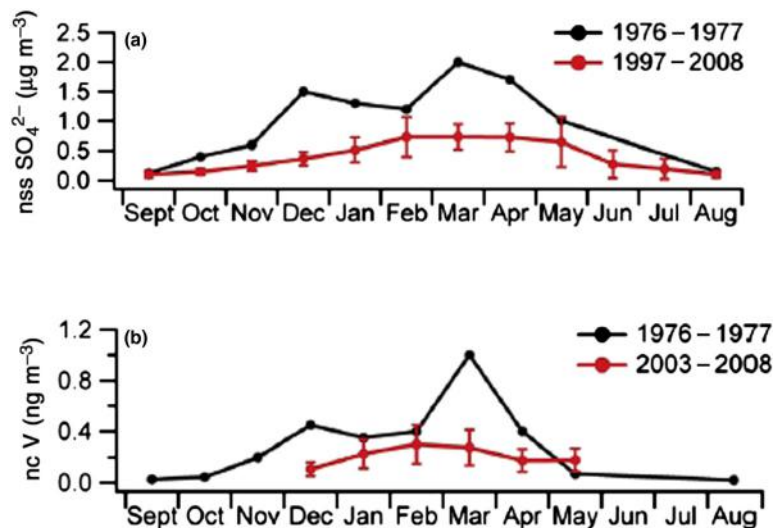


Figure 3 Monthly averaged values from 1976 to 1977 and 1997–2008 at Barrow, Alaska for (a) non-sea-salt (nss) sulfate and (b) non-crustal (nc) Vanadium. The 1976–1977 data are from Rahn and McCaffrey (1980). Vertical bars for the 1997–2008 data are the standard deviation of the monthly average. Reprinted from Quinn, P. K., Bates, T. S., Schulz, K., Shaw, G. E., 2009. Decadal trends in aerosol chemical composition at Barrow, Alaska: 1976–2008. *Atmospheric Chemistry and Physics* 9 (22), 8883–8888. doi:10.5194/acp-9-8883-2009.

late winter and early spring, consistent with meteorological drivers for this pattern.

A large fraction of the arctic haze is sulfate. Note from Figure 1 that the Eurasian sulfur dioxide emissions in areas liable to influence the Arctic are about a factor of 2–4 times larger than for

North America. Note also that the major Eurasian sulfur dioxide emission sources are $5\text{--}10^\circ$ higher in latitude in comparison to those in North America. This, along with the fact that the arctic air mass extends down further over Eurasia, suggests that the major contribution to arctic haze is from the Eurasian sources.

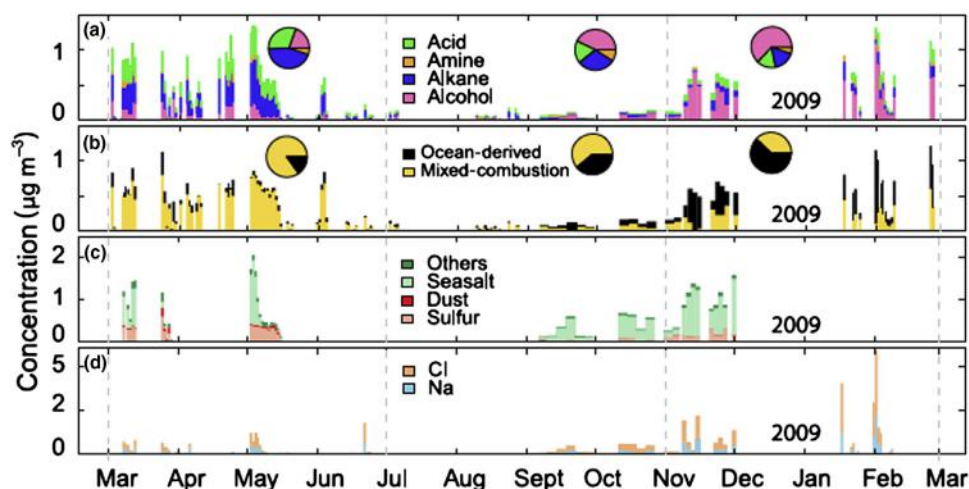


Figure 4 Time series of measured atmospheric aerosol component concentrations at Barrow, Alaska. (a) organic functional groups; (b) source-related factors; (c) sea salt, other trace elements, sulfur (represents approximately one-third of the sulfate mass), and crustal components; and (d) sea salt. Bar widths correspond to durations of collected filters. Inset pies indicate time-weighted seasonal averages. Reprinted from Shaw, P. M., Russell, L. M., Jefferson, A., Quinn, P. K., 2010. Arctic organic aerosol measurements show particles from mixed combustion in spring haze and from frost flowers in winter. *Geophysical Research Letters* 37, doi:10.1029/2010gl042831.

The emission sources of the arctic haze were determined to be of Eurasian origin using a chemical fingerprint based on the ratio of vanadium to manganese concentrations in the collected filtered samples of the arctic haze. This simple tracer system immediately suggested that the greatest fraction of arctic haze aerosol was derived from Eurasian industrial sources, especially in the eastern sectors. The reason is that the former Soviet Union and eastern European nations have been using coal-burning as major power sources, while the western hemisphere nations are heavy users of petroleum products, laced with vanadium used as a catalyst in the cracking process. The arctic haze value of V/Mn was very low, consistent with a coal-burning source. Concentrations of black carbon were also elevated, consistent with the dirtier combustion that takes place in inefficient coal power plants of the type used in the former Soviet Union. This basic picture of the sources and transport mechanisms for the arctic haze was recognized quite clearly during the 1970s.

Aircraft-based measurement programs, including the Arctic Gas and Aerosol Sampling Programs, the German-Russian aircraft 'Arctic Haze/Merise' campaigns of 1993–95, the Arctic Study of Tropospheric Aerosol campaign, the European Arctic Aerosol Study, and Arctic Research of the Composition of the Troposphere from Aircraft and Satellites campaigns, have directly sampled the arctic haze, and the changes in the composition of the arctic haze have been tracked for more than 40 years. Sulfur in the form of sulfate continues to be a substantial fraction of particle mass identified in the arctic haze, even at the higher altitudes, but recent measurements at Barrow confirm an average decrease of non-sea-salt sulfate emissions of a factor of four over 30 years (Figure 5). By comparison, the observations from Alert (Figure 2) have shown a decrease in sulfate of about a factor of two over approximately the same period.

By combining the meteorological trajectory models with geographic information and inventories of chemical emissions,

recent investigations have identified a number of specific geographic locations in Eurasia that have high potential as major emission sources. For example, a Ni–Cu smelting complex at Norilsk is probably one of the major contributors to the haze. But in addition to sulfur-containing particles and heavy metals, carbonaceous particles will also reach the Arctic. During polar winter, the contributions of organic compounds consist largely of primary marine-derived components (Figure 4). After polar sunrise the organic fraction can increase photochemically so that during spring and summer organic carbon is a major aerosol constituent, accounting for as much if not more submicron particle mass than the sulfate.

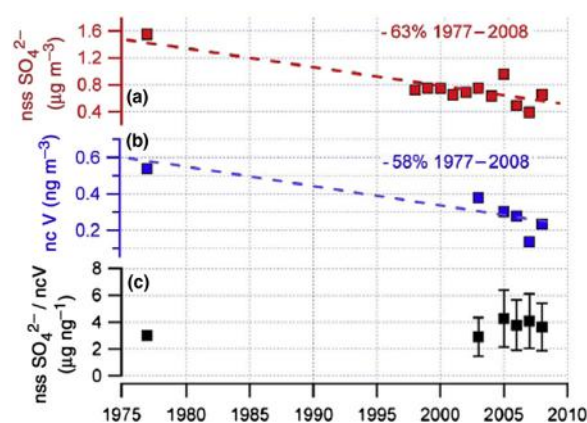


Figure 5 Concentrations of (a) non-sea-salt (nss) sulfate and (b) non-crustal (nc) Vanadium averaged over the haze season (January–April) and (c) the nss sulfate (SO_4^{2-}) to ncV ratio. The 1976–1977 data are from Rahn and McCaffrey (1980). Reprinted from Quinn, P. K., Bates, T. S., Schulz, K., Shaw, G. E., 2009. Decadal trends in aerosol chemical composition at Barrow, Alaska: 1976–2008. *Atmospheric Chemistry and Physics* 9 (22), 8883–8888. doi:10.5194/acp-9-8883-2009.

Recent studies indicate that combustion emissions account for more than 60% of the organic mass (OM) that make up arctic haze particles, but the Eastern and Western regions of the Arctic are largely influenced by two different urban and industrial centers with different fuel usage resulting in particles with different chemical compositions. The high-sulfur coal and wood burning in northeastern Europe (Poland and the Russian Kola Peninsula) contribute to the haze in the eastern (European) Arctic, and the oil-burning and forest fires in northeastern Asia (Siberia) influence the haze in the western (North American) Arctic (Figure 6). The organic components of haze include several types of organic molecules. Organic acid and alcohol group contributions measured by infrared spectroscopic techniques account for almost 25% of the combustion organic functional groups, reflecting the composition expected for photochemical oxidation of incompletely combusted fuels. Aerosol mass spectrometric measurements confirm the highly oxidized organic composition of arctic haze particles, similar to that observed for combustion emissions in the subpolar source regions. One interesting chemical consequence of the regional differences is that the higher coemissions of sulfate from coal-burning in northeastern Europe produce significant concentrations of another type of organic component – organosulfate functional groups – which account for as much as 10% of organic particles over the Barents and Greenland Seas. Long-lived organic compounds such as polychlorinated aromatics and pesticides have also been measured in both arctic air pollution and snowpack.

Optical Transparency and Climatic Effects

In the early 1970s, unexpectedly high values for atmospheric turbidity were reported at the McCall Glacier in the Brooks Range in Alaska. In trying to understand the physical cause of the high turbidity, additional measurements were made of the wavelength dependence of optical extinction caused by the haze and the angular distribution of sky brightness. These measurements confirmed Mitchell's earlier suspicions that the winter arctic haze consisted mainly of small aerosols. During the early 1970s, arctic haze over Alaska was found to be layered by flying a sun photometer aboard a light (Cessna) aircraft. The AGASP experiments of the 1980s demonstrated that the layers consist of aerosol particles that can have climatic effects by interacting with the solar radiation field all across the Arctic. In addition to the submicrometer haze, there is often another source of haziness in the Arctic from precipitations of small ice crystals, which sometimes form in clear air. Such 'diamond dust,' so-called because of its sparkling appearance, can also reduce visibility and interact with the radiation field.

Scattering and absorption of sunlight by the arctic haze was shown to have a slight warming effect on the Earth-atmosphere column, of magnitude about 0.1 °C per day, because the light absorbing properties of black carbon particles lower the albedo. Since the surface albedo of sea ice is so high, this absorption tends to be more important than the slight cooling at the surface caused by the scattering of light. In the winter months, there may be a slight warming of the arctic

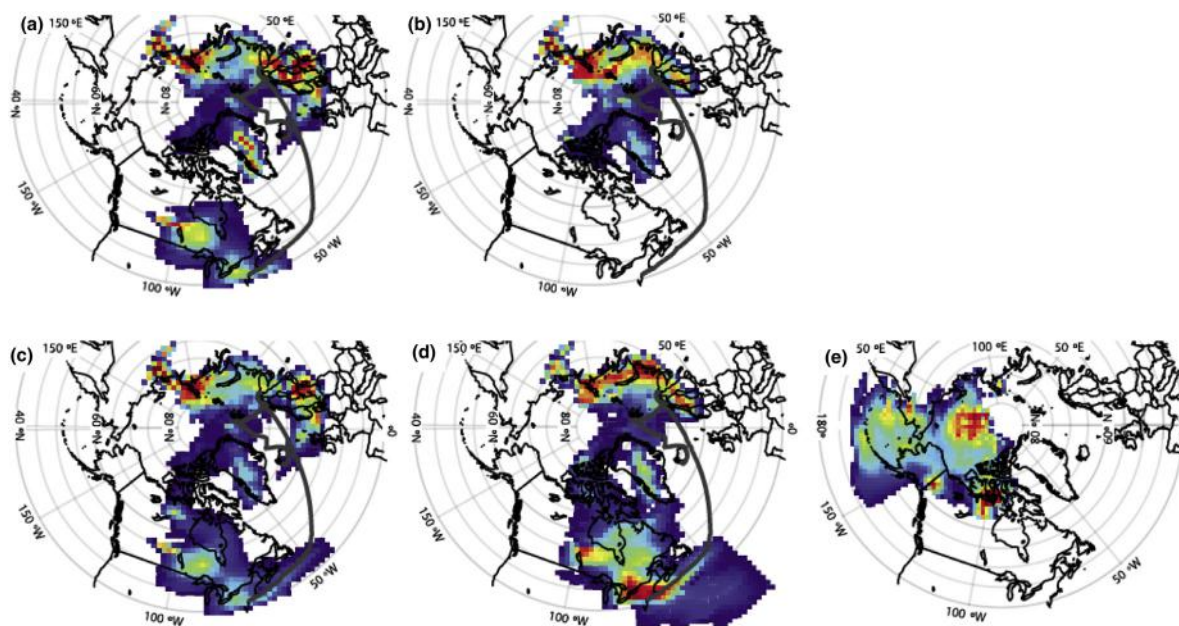


Figure 6 Potential source contribution functions of (a) sulfate, (b) organosulfate functional group concentrations, (c) the total OM1 measured, and (d) the combustion-derived OM measured during a cruise in the North Atlantic and Arctic Oceans in 2008, as well as (e) the combustion-derived OM measured at Barrow in 2009. The higher potential source regions are indicated by red and the lower potential source regions are indicated with blue. Reprinted from Figures 5 and 6 of Frossard, A. A., Shaw, P. M., Russell, L. M., Kroll, J. H., Canagaratna, M. R., et al., 2011. Springtime Arctic haze contributions of submicron organic particles from European and Asian combustion sources. *Journal of Geophysical Research – Atmosphere* 116, doi:10.1029/2010jd015178.

atmosphere caused by the interaction of the aerosol with outgoing infrared radiation.

However, in addition to the slight climate effects caused by absorption and scattering of light by the arctic haze aerosols, there may be subtle and so far not well-evaluated influences on climate from indirect radiative effects. These may result from the modification of cloud properties, since the arctic haze introduces new sources of cloud condensation nuclei and, possibly causes dehydration.

Ecological Implications

Compounds such as pesticides, polychlorinated biphenyls, persistent organics, as well as trace metals are detected throughout the arctic basin in the atmosphere and also in the surface land and sea domains and scattered throughout the region's biome. Most of the atmospheric pollution and some of the surface and biological pollution is undoubtedly caused from the arctic haze phenomena. But it should also be recognized that in places in the Arctic the surface concentrations of pollutants can be extremely small, even during times when the air is quite contaminated. This of course is due to the long residence time of the haze.

In addition to simple transport of material through the atmosphere, another depositional mechanism was suggested that involves a fractional distillation. The surprisingly high concentration of organic material is determined in part by the temperature-dependent partitioning of the low volatility compounds. Substances with low vapor pressure preferentially accumulate in the polar regions.

Aerosols and gases are scrubbed out of the atmosphere in precipitating air masses. Many of these are over oceanic regions in the North Atlantic, the Norwegian Sea, and

possibly in the Bering Sea. These are important world fisheries and the consequences of dumping industrial pollutants that had congregated in the arctic atmosphere are unknown.

See also: Satellites and Satellite Remote Sensing: Aerosol Measurements.

Further Reading

- Curry, J.A., 1995. Interactions among aerosols, clouds, and climate of the Arctic-Ocean. *Science Total Environment* 160–161, 777–791. [http://dx.doi.org/10.1016/0048-9697\(95\)04411-s](http://dx.doi.org/10.1016/0048-9697(95)04411-s).
- Frossard, A.A., Shaw, P.M., Russell, L.M., Kroll, J.H., Canagaratna, M.R., et al., 2011. Springtime Arctic haze contributions of submicron organic particles from European and Asian combustion sources. *Journal of Geophysical Research – Atmosphere* 116. <http://dx.doi.org/10.1029/2010jd015178>.
- Gong, S.L., Zhao, T.L., Sharma, S., Toom-Saunty, D., Lavoue, D., et al., 2010. Identification of trends and interannual variability of sulfate and black carbon in the Canadian High Arctic: 1981–2007. *Journal of Geophysical Research – Atmosphere* 115. <http://dx.doi.org/10.1029/2009jd012943>.
- Quinn, P.K., Bates, T.S., Schulz, K., Shaw, G.E., 2009. Decadal trends in aerosol chemical composition at Barrow, Alaska: 1976–2008. *Atmospheric Chemistry and Physics* 9 (22), 8883–8888. <http://dx.doi.org/10.5194/acp-9-8883-2009>.
- Russell, L.M., Hawkins, L.N., Frossard, A.A., Quinn, P.K., Bates, T.S., 2010. Carbohydrate-like composition of submicron atmospheric particles and their production from ocean bubble bursting. *Proceedings of the National Academy of Sciences USA* 107 (15), 6652–6657. <http://dx.doi.org/10.1073/pnas.0908905107>.
- Shaw, G.E., 1995. The Arctic haze phenomenon. *Bulletin of the American Meteorological Society* 76 (12), 2403–2413. [http://dx.doi.org/10.1175/1520-0477\(1995\)076<2403:tahp>2.0.co;2](http://dx.doi.org/10.1175/1520-0477(1995)076<2403:tahp>2.0.co;2).
- Stohl, A., 2006. Characteristics of atmospheric transport into the Arctic troposphere. *Journal of Geophysical Research – Atmosphere* 111 (D11306). <http://dx.doi.org/10.1029/2005JD006888>.

AIR SEA INTERACTIONS

Contents

Freshwater Flux

Momentum, Heat and Vapor Fluxes

Sea Surface Temperature

Surface Waves

Freshwater Flux

J Schulz, Meteorological Institute, University of Bonn, Bonn, Germany

© 2003 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 1, pp 75–84, © 2003, Elsevier Ltd.

Introduction

The world ocean is a key element of the physical climate system. The ocean contains 97% of the world's water and covers an area of 71% of the globe. As a reservoir, the ocean supplies water vapor to the atmosphere that brings rain and snow over land surfaces. About one-third of the precipitation over land originates from water evaporated from the ocean. The water vapor in the atmosphere is the most important gaseous absorber for solar and terrestrial radiation and accounts for about half of the atmosphere's natural greenhouse effect. The process of evaporation accounts for approximately half of the surface cooling balancing the heating by absorption of solar radiation.

Because of the asymmetric insolation of the Earth's surface by solar radiation, the oceans act as a large energy and heat transport system from the Equator to the poles. The deep-ocean circulation that is critical for this transport is mostly driven by variations in the density of sea water. Ocean salinity is an important contributor to these variations and varies with latitude in the upper layers of the oceans. The surface salinity depends on the fresh water flux at the ocean surface and is relatively high in the subtropics where evaporation exceeds precipitation, whereas it is relatively low in the tropics and middle and high latitudes where precipitation dominates.

The most comprehensive publication about the world water balance was written by Baumgartner and Reichel in 1975. They assembled different estimates of the water balance over continents, river basins, and oceans to calculate a global water balance. This pioneering work resulted in global maps of evaporation and precipitation that even today are widely used by meteorologists, oceanographers, climatologists, and hydrologists. The definition of the water balance and its

components used here follows the work of Baumgartner and Reichel.

Assuming that the amount of water on Earth is not changing with time, the long-term average of the water balance for a unit area of the Earth's surface can be written as eqn [1].

$$P = E + D \quad [1]$$

In eqn [1], P is precipitation, E is evaporation, and D is discharge or river runoff. This balance states that water added to the surface by precipitation is partitioned between E and D . Although globally precipitation and evaporation are balanced, the large differences in the components P , E , and D over land and over ocean produce the world's water cycle, shown schematically in Figure 1. The units used throughout the article for E , P , and $E-P$ are mm d^{-1} . In general, total evaporation exceeds precipitation over oceans, which is compensated for by the runoff of rivers from the continents, where precipitation exceeds evaporation.

In the following sections the focus will be on the different methods used to determine the freshwater flux at the oceans surface: E , P , $E-P$ will be considered in detail. The heat transport within the oceans is beyond the scope of this article (see **Air–Sea Interaction: Momentum, Heat, and Vapor Fluxes**). The second section gives an overview of different techniques that can be used to determine the fresh water flux, followed by a description of the fresh water flux climatology as derived from satellite data. This is followed by a short consideration of the role of river runoff, and finally some conclusions are presented.

Methods for Determining E , P , and $E-P$

Basically, there exist three different methods for determining the components E and P as well as $E-P$. Traditional

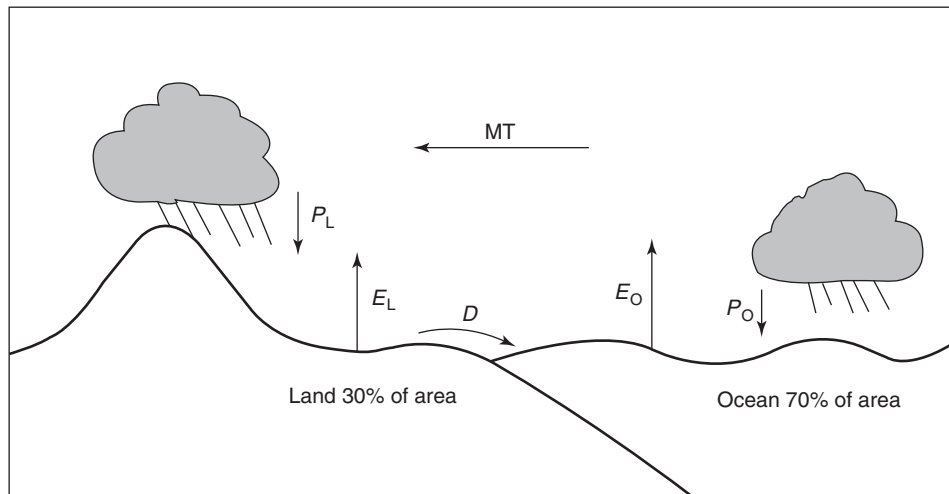


Figure 1 Schematic representation of the world water cycle E and P denote evaporation and precipitation over oceans and land (denoted by subscript O and L , respectively). D is the discharge or river runoff of water from the continents to the oceans and MT is the moisture transport in the atmosphere.

estimates of E are based on *in situ* measurements of surface wind speed U , specific humidity of air Q_a , and sea surface temperature T_s . These are used within the so-called bulk aerodynamic formula to parametrize the evaporation. P is estimated by analyzing actual weather reports using an empirical parametrization. The resulting estimates are interpolated and extrapolated to construct global maps of E and P .

Recently, satellite data have been used to derive the same basic state variables near and at the ocean surface using empirical and physical retrieval schemes. Again these quantities are used to parametrize E using the bulk aerodynamic formula. Many algorithms for estimating rainfall using almost the whole electromagnetic spectrum have been developed during the last 20 years (see **Satellites and Satellite Remote Sensing: Precipitation**). These satellite algorithms are used alone and in conjunction with *in situ* data and model results to give best estimates.

The third method is the so-called moisture budget method, which make use of global-scale analyzed water vapor fields or measurements of atmospheric water vapor by rawinsondes in the form of four-dimensional data assimilation (see **Statistical Methods: Data Analysis: Time Series Analysis**). The global distribution of $E-P$ is then computed from the residuals of water vapor transport in the atmosphere using large-scale numerical models.

Traditional Estimates from *in situ* Measurements

Most of our present knowledge of fresh water fluxes is derived from weather observations on special weather ships, buoy data, and also data from merchant ships participating in the Voluntary Observing Ship system. Many of these data have been organized into the Comprehensive Ocean Atmosphere Data Set which has been used to derive climatologies of the energy fluxes and the fresh water flux at the sea surface. The major disadvantage of ship-based estimates of E and P is that the

observation base is not very good for either parameter. The coverage is mostly obtained along shipping lanes, which may be sufficient in the North Atlantic, North Pacific, and the Mediterranean but is not sufficient in the Tropics and all southern oceans. Additionally, the concentration along shipping lanes can introduce a fair weather bias, since ships try to avoid bad weather.

Whereas the measurement quality for the basic state variables used for the parametrization of E is relatively good on ocean weather ships and research quality buoys, it is less good on the voluntary observing ships. Although much effort has been put into correcting errors on the basis of individual ship measurements during the last few years, the global heat balance has not been closed, mostly because of the low observation density and deficiencies in the bulk aerodynamic formula. The determination of precipitation is even more difficult. It is largely based on the observed actual weather and parametrizations that convert the weather code into rainfall amount. The conversion schemes were developed from data that are not representative of the global oceans, so it became necessary to correct under-estimated precipitation in the Tropics by empirical temperature-dependent corrections. Incorporation of P measurements from islands in data-sparse regions is also very difficult, because of the influence of island terrain on the rainfall. Comparisons of these precipitation fields with satellite-derived fields exhibit large differences even at the climatological scale.

Remote Sensing of $E-P$

Remote sensing of evaporation is based mostly upon the derivation of basic state variables, wind speed, sea surface temperature, and near-surface atmospheric specific humidity, and the parametrization of the evaporation using the bulk aerodynamic formula. Wind speed can be obtained from either passive or active microwave systems. The active system relates the backscattered energy to the wind speed at a reference level

over sea surface (e.g., 10 m) and is also able to deliver wind direction information. The passive systems rely on the surface emission change due to wind-induced sea surface roughness and partial foam coverage. The root-mean-square (rms) errors on an instantaneous time scale for both systems are on the order ± 1.3 to $\pm 2 \text{ m s}^{-1}$. Remote sensing methods for the near-surface specific humidity make use of a vertically integrated water vapor content (obtained from a passive microwave instrument such as the Special Sensor Microwave/Imager (SSM/I)) as a predictor. Several techniques ranging from linear regression analysis to neural networks have been used to deduce the near-surface humidity with rms errors of $\sim \pm 0.7 \text{ g kg}^{-1}$ on the monthly time scale. Estimates of sea surface temperature are deduced from passive infrared sensors like the Advanced Very High Resolution Radiometer (AVHRR). The largest problem with this method is the cloud clearance of the satellite scene, because otherwise the determined T_s would be negatively biased. Accounting for the surface skin effect, rms errors for the best satellite methods are $\sim \pm 0.2 \text{ K}$. Estimates of sea surface temperature with not much less accuracy are also possible employing passive microwave measurements at frequencies between ~ 50 – 10 GHz that are available from TRMM's TMI and will shortly be available from the new Advance Multifrequency Scanning Radiometer onboard the Aqua and ADEOS-II satellites. The big advantage of those estimates is the much better coverage because clouds are almost transparent at those frequencies allowing an undisturbed view of the ocean surface. However, infrared estimates of sea surface temperature remain of high importance for the computation of evaporation climatology since estimates of sea surface temperature from SSM/I measurements were not possible with sufficient accuracy. Recently, some new methods have been developed that circumvent the bulk formula in the retrieval process by relating the satellite data directly to an existing flux data set, e.g., re-analyses of fields derived from *in situ* data that are assumed to be true. The accuracy of all methods is comparable at a level of $\pm 30 \text{ W m}^{-2}$ and $\pm 15 \text{ W m}^{-2}$ at weekly and monthly time scales, respectively.

As mentioned in the previous section, there are not sufficient conventional and surface-based radar rainfall estimates over the oceans for the derivation of rainfall fields. A reasonable alternative is the use of satellite remote sensing. Remote sensing of rainfall from satellites started with the statistical analysis of the reflectivity and emissivity of the upper cloud layers at visible and infrared wavelengths, respectively. Because of the small physical correlation between the signal and the rainfall at the surface, this technique leads to acceptable results only if the derived rainfall is integrated over space and time. Owing to the strong variability of rainfall, rainfall climatologies derived using this technique with data from geostationary satellites with their high repetitive cycle build are still the backbone of today's rainfall analyses.

Over water surfaces, passive microwave radiometers deliver a much better information base. The signal at frequencies below 30 GHz is mostly determined by the emission from rain water, which leads to a strong increase of the brightness temperature over the cold background of the sea surface. For higher frequencies, the brightness temperature decreases owing to scattering by ice particles. This information can also be used to estimate the rainfall rate at the surface.

Since the launch of the first SSM/I onboard the satellites of the Defense Meteorological Satellite Program (DMSP) in 1987, a continuous time-series of data exists from at least one satellite. Many algorithms have been developed to analyze rainfall using these measurements. On the basis of numerous algorithm intercomparison projects, it has proved almost impossible to find a so-called standard algorithm that performs best for most of the situations investigated. The variability of the cases analyzed showed the quality of some algorithms under certain conditions, but not of one prevailing algorithm. In many cases it was found that the accuracy of the validation data was not sufficient to classify the quality of the satellite algorithms. A prominent data set using combinations of geostationary satellite data, passive microwave data, and rain gauge data is that produced by the Global Precipitation Climatology Project (GPCP).

With the launch of the satellite of the Tropical Rainfall Measuring Mission (TRMM) in 1997, for the first time a space-borne radar can be used to derive the three-dimensional structure of rainfall and the surface rainfall. The combination of instruments onboard the TRMM satellite can be considered as a reference for methods applied to other instruments in space. A calibration of rainfall estimates from other satellites like the SSM/I then delivers an optimal combination of accuracy and temporal/spatial sampling and subsequently consistent rainfall distributions.

Moisture Budget Methodology

The moisture budget methodology tries to compute $E-P$ as a residual from the large-scale atmospheric transports of water vapor using global analyses and re-analysis data sets produced with four-dimensional data assimilation schemes. This technique has a long history, although it usually makes use of rawinsonde data directly. $E-P$ is computed from eqn [2], which is the vertically integrated (from the top of the atmosphere to the surface) equation for the conservation of water vapor.

$$E - P = \frac{\partial W}{\partial t} + \nabla \cdot \frac{1}{g} \int_0^{p_s} q \mathbf{v} dp \quad [2]$$

W is the total precipitable water, q is the specific humidity, p is pressure (with p_s being the surface pressure), $\mathbf{v}$ is the velocity vector, and g is the standard gravity. Many comparison studies between precipitation fields produced routinely and the GPCP data set found discontinuities in the analyses due to changes in the data assimilation system. Another problem was that rainfall maxima in the analyses were often in the wrong place and too strong. In general, the assimilation systems have been much improved during the last ten years and are more or less consistent with satellite-derived data sets. However, use of this method with different analysis and re-analysis products from the European Center for Medium-range Weather Forecasts (ECMWF), the National Centers for Environmental Prediction (NCEP), and the National Aeronautics and Space Administration (NASA) as input and comparisons to pure model-computed $E-P$ exhibited large differences. The most critical part of this method is the dependence of the moisture budget on the divergence of the velocity field. This is of special importance

in the Tropics, where the divergence field is not very well known.

River Inflow

The inflow of fresh water from rivers is not included in most ocean surface fresh water data sets derived from satellites and is also neglected in the residual approach, but it is significant to the global fresh water balance of the ocean. Baumgartner and Reichel estimated from a global hydrological balance calculation that the contribution of river runoff to the balance is as high as 10% of the contribution of precipitation. It might be thought that the impact of the fresh water inflow from the rivers on the buoyancy would be local in comparison to the size of the ocean basins. However, the impact of major rivers like the Amazon or Congo is observable several hundred kilometers away from the mouths of the rivers. Neglecting this contribution would increase the average salinity of the upper Atlantic ocean (the first 50 m) by 1.5 psu after 10 years of integration of a numerical ocean model.

Two major runoff data sets are the Global River Discharge Catalogue published by the International Hydrography Program (IHP) and that issued by the Global Runoff Data Center, and these form the backbone of information on river runoff. IHP data consists of a selection of monthly discharges at 949 stations over six continents (Africa, Asia, North America,

South America, Europe and Australia/Oceania), but only 219 stations are listed as corresponding to a direct discharge into an ocean basin. The length of the individual data sets is 19.3 years on average, but varies from 1 year to 100 years. Additionally, a great disparity exists between the different continents, with Europe and North America presenting the longest records. Not included in either data set is the runoff from Arctic and Antarctic regions and the inflow of fresh water from ground water sources.

Analysis of the monthly mean climatology of the direct contribution of rivers to the fresh water flow into the ocean, considering only rivers for which time records longer than two years exist, sums to ~ 0.57 Sv ($1 \text{ Sv} = 10^6 \text{ m}^3 \text{ s}^{-1}$). This estimate is much lower than that of Baumgartner and Reichel, who found 0.73 Sv, which also includes contributions from regions beyond the polar circles, which they estimated from other sources.

Climatology of $E-P$ Derived from Satellite Data

Figure 2, Figure 3 and Figure 4 show the seasonally averaged global maps of evaporation, precipitation, and $E-P$ derived from AVHRR and SSM/I data, on a grid with $1^\circ \times 1^\circ$ resolution. The data set was constructed by averaging instantaneous estimates of the components and the flux over 11 years (1987–98).

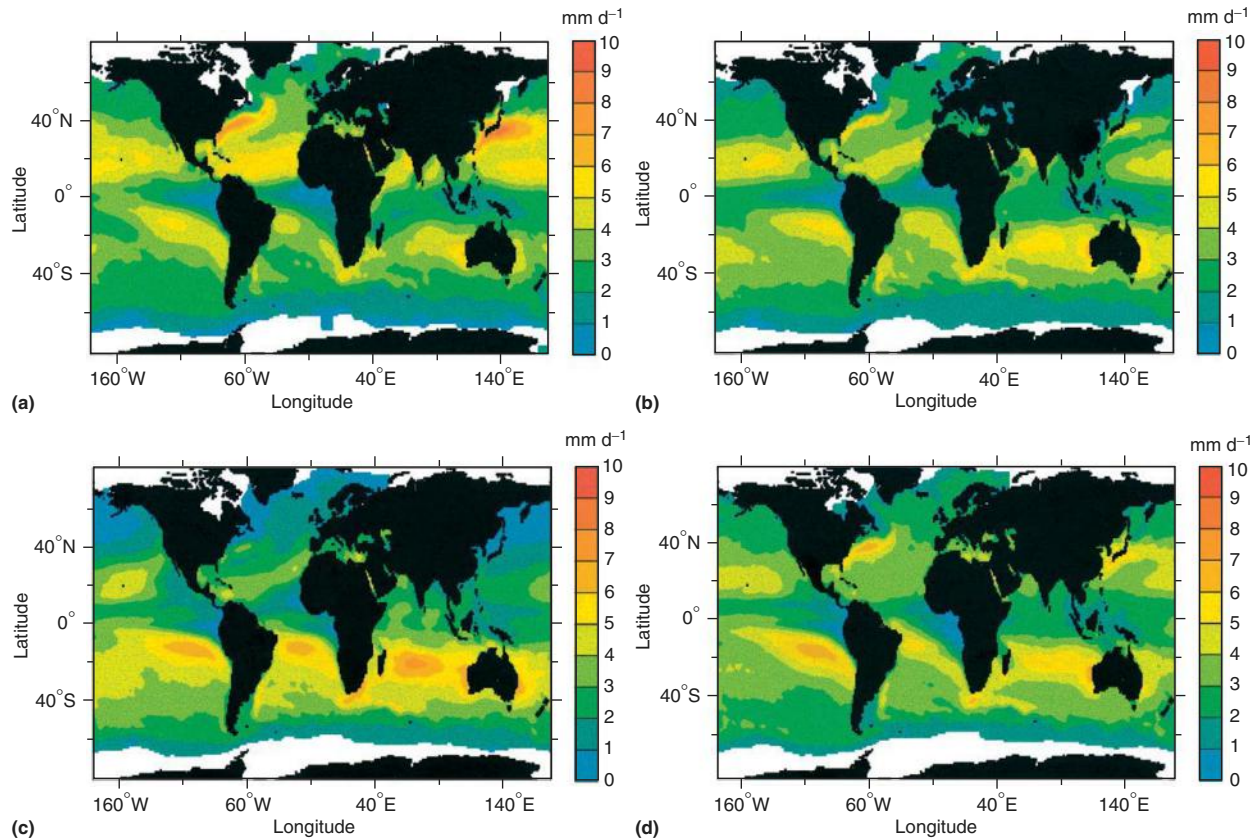


Figure 2 Climatological seasonal averages of evaporation in mm d^{-1} as derived from satellite data: (a) December, January, February; (b) March, April, May; (c) June, July, August; (d) September, October, November.

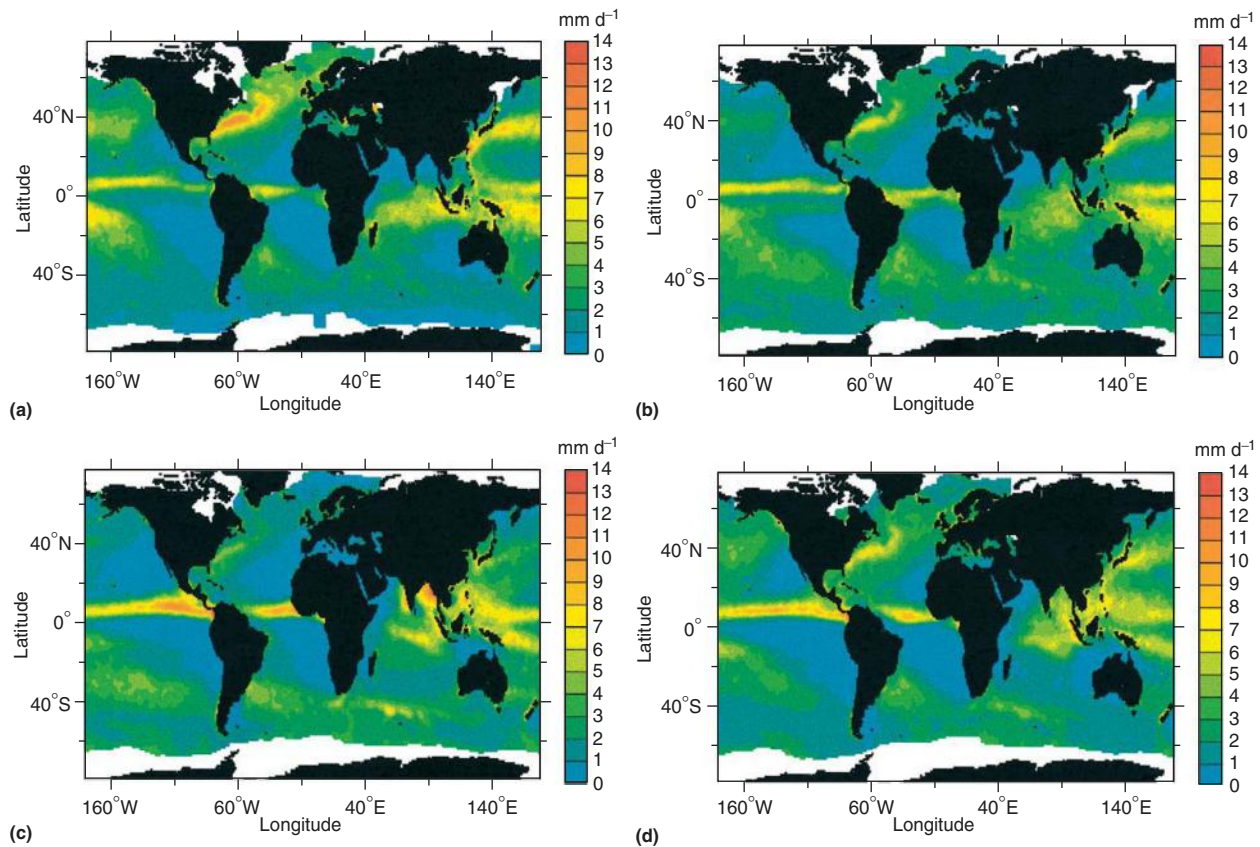


Figure 3 Climatological seasonal averages of rainfall in mm d^{-1} as derived from satellite data: (a) December, January, February; (b) March, April, May; (c) June, July, August; (d) September, October, November.

Maximum values of evaporation up to 9 mm d^{-1} are observed over the Kuroshio and Gulf Stream regions in winter (Figure 2(a)) and minimum values below 1 mm d^{-1} are seen in the eastern equatorial Pacific and Atlantic during all months. Large areas with high evaporation rates of $5\text{--}6 \text{ mm d}^{-1}$ are found in the main Trade Wind belts between about 10° and 40° latitude in both the Northern and Southern hemispheres. These high evaporation regions are the major sources of atmospheric water for the global hydrological cycle. Whereas during Northern Hemisphere winter and spring the maximum extent and the highest values are found north of the Equator (Figure 2(a) and 2(b)), maximum evaporation is observed in the southern Indian, Atlantic, and Pacific Oceans during Northern Hemisphere summer and fall (Figure 2(c) and 2(d)).

The global precipitation pattern is dominated by a strong band of precipitation circling the globe just north of the Equator. This is the region where the northern and southern Hadley circulation cells meet, forming a region of strong surface convergence known as the Intertropical Convergence Zone (ITCZ). Where the maximum precipitation on an annual scale exceeds 6 mm d^{-1} . Another convergence zone in the western tropical Pacific, known as the South Pacific Convergence Zone (SPCZ), is somewhat broader, with precipitation values similar to those in the ITCZ. It extends from the region of Indonesia and the Philippines south-east across the southern Pacific. With the onset of the summer monsoon, the ITCZ, which was earlier

located in the southern Indian Ocean, shifts to its northernmost location and merges with the Monsoon trough giving rise to copious rainfall over the Indian subcontinent and the adjacent seas, namely the Arabian Sea and Bay of Bengal (Figure 3(c)).

On an annual scale, the region off the Indonesian islands receives a maximum rainfall of more than 10 mm d^{-1} . Outside the two convergence zones, precipitation rates are significantly lower, with the exception of two regions. Precipitation rates are quite high over the Gulf and Kuroshio Streams, with values as high as in the ITCZ during the period November to March (Figure 3(a)). This feature has not been recognized in rainfall climatologies derived from routine weather observations.

At tropical and subtropical latitudes, between about 15° N and 40° N and between 5° S and 30° S , the eastern parts of the Pacific and the Atlantic ocean are regions where precipitation is below 1 mm d^{-1} . A similar region is found in the Indian Ocean along the east coast of Africa and Saudi Arabia during all months, with a maximum extent over the whole Arabian Sea during winter (Figure 3(a)). In the Southern Hemisphere, west of Australian coast, another minimum is observed, which has its maximum extension during the Southern Hemisphere spring (Figure 3(d)).

The patterns of precipitation and evaporation exhibit quite different spatial distributions. Precipitation maxima occur in the global convergence regions, while evaporation maxima

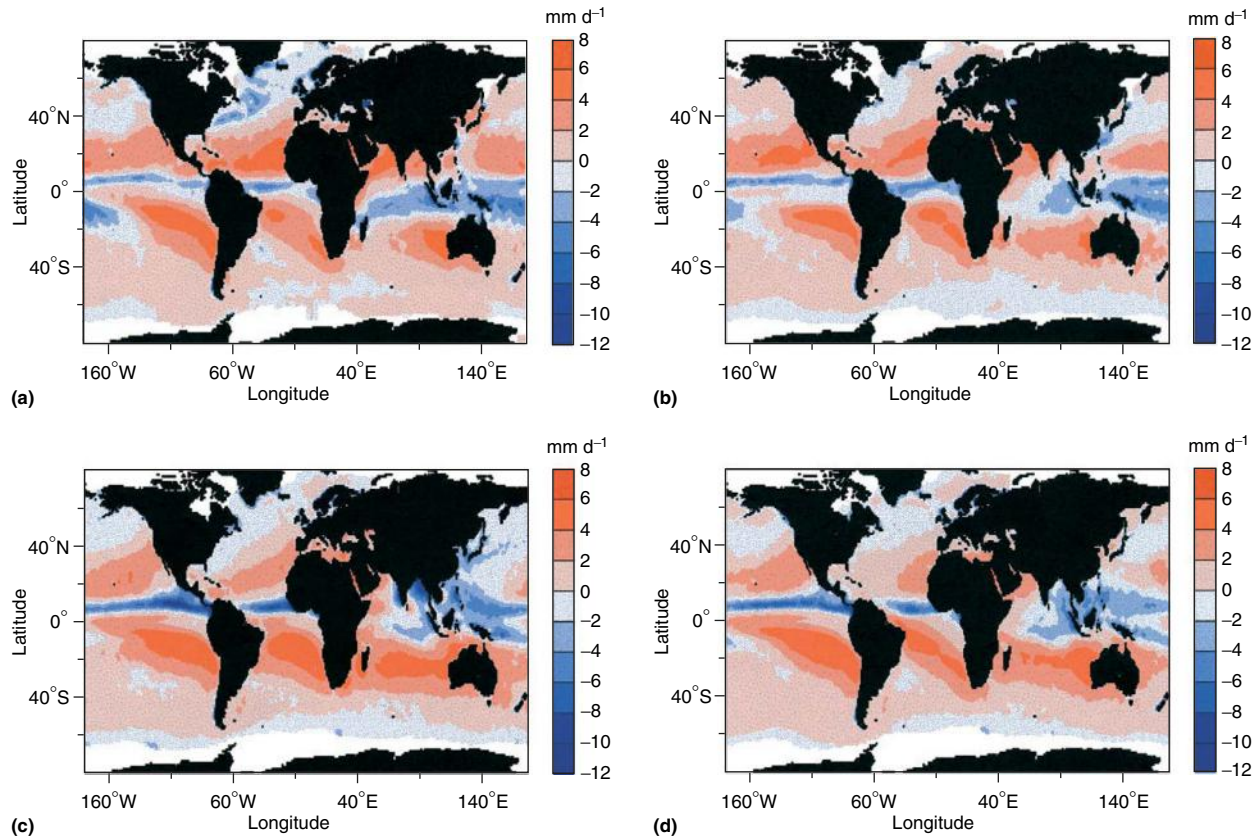


Figure 4 Climatological seasonal averages of $E-P$ in mm d^{-1} as derived from satellite data: (a) December, January, February; (b) March, April, May; (c) June, July, August; (d) September, October, November.

occur in regions of high surface humidity gradient and wind speed. On a monthly time scale, values of E exhibit much less spatial structure than fields of P . From this it is clear that the $E-P$ monthly temporal variability is dominated by variations in location and intensity of rainfall and the spatial structure in $E-P$ is also dominated by the P field.

However, on the climatological time scale, fields of $E-P$ consist of signatures of both evaporation and precipitation fields. The ITCZ and SPCZ appear prominently as regions of fresh water supply to the ocean. In these regions the fresh water flux from atmosphere to ocean is generally larger than 4 mm d^{-1} . With the exception of the SPCZ, precipitation decreases rapidly with latitude to the north and to the south of the ITCZ, while evaporation remains strong or even increases, causing positive values of the fresh water flux.

The strongest gradients in the fresh water flux fields occur in the boundary regions between the negative values of 4 mm d^{-1} within the ITCZ and the strong positive flux regions to the north and south, with values up to 6 mm d^{-1} . Poleward from the evaporation regions, the fresh water flux is relatively small. The evaporation fields generally decrease toward the poles primarily as a result of the decrease in the humidity difference, except during wintry arctic cold air outbreaks which often lead to very high evaporation rates and therefore to positive fresh water fluxes. Although evaporation is quite high, large negative values of $E-P$ can

be found in the Gulf Stream and Kuroshio regions, below -4 mm d^{-1} during the winter owing to high precipitation, while in all other months evaporation almost balances precipitation.

An analysis of the fresh water flux on a seasonal scale (Figure 4(a)–(d)) reveals that the eastern parts of the Arabian Sea, the Bay of Bengal, and the South China Sea all have negative values of fresh water flux during summer (Figure 4(c)) and autumn (Figure 4(d)). Further, on an annual scale, it can be seen that the eastern equatorial Indian Ocean, the Bay of Bengal, and the Kuroshio and the Gulf Stream regions all exhibit negative fluxes. The regions of positive flux are over the north-west Arabian Sea and the southern Indian Ocean south of 20° S . Also, the North Atlantic and the South Atlantic exhibit positive fluxes on both sides of the ITCZ.

Discussion and Conclusion

Table 1 shows how existing estimates of the fresh water flux and its components differ from some examples of estimates from different sources found in the literature. Whereas the older estimates from observations are comparable to the results from General Circulation Models, the satellite estimates differ considerably from all of them. However, today

Table 1 Global climatological averages for E , P , and $E-P$ (in mm d^{-1}) over oceans, from different studies

Source	E	P	$E-P$
Baumgartner and Reichel (1975)	1177	1066	111
Chahine (1992)	1202	1088	114
ECHAM4 ^a (Todini and Dümenil, 1999)	1246	1147	99
ECMWF ^b +GPCP ^c rain (Oki, 1999)	1194	1083	111
HOAPS ^d (Grassl <i>et al.</i> , 2000)	1086	908	178

^aECHAM4: Climate model of the Max-Planck Institute for Meteorology, Hamburg.^bECMWF: European Centre for Medium-range Weather Forecasts.^cGPCP: Global Precipitation Climatology Project.^dHOAPS: Hamburg Ocean Atmosphere Parameters and Fluxes from Satellite Data.

there is no agreed true value for $E-P$. Current results of global analyses seem not to be very reliable, but centers like ECMWF are improving the assimilation of rainfall estimates from satellite data, and these will be operational in a few years. Satellite data sets have great potential to be improved in the future by using sophisticated methods of intercalibration between different satellites. In the case of basic state variables U , Q_a , and T_s , improvements are expected from intercomparison of the satellite estimates with high-quality surface-based measurements.

See also: **Aerosols:** Climatology of Tropospheric Aerosols. **Air Sea Interactions:** Momentum, Heat, and Vapor Fluxes; Sea Surface Temperature. **Clouds and Fog:** Climatology. **Data Assimilation and Predictability:** Data Assimilation. **Radiation Transfer in the Atmosphere:** Ultraviolet Radiation. **Satellites and**

Satellite Remote Sensing: Precipitation. **Statistical Methods:** Data Analysis: Time Series Analysis. **Thermodynamics:** Humidity Variables.

Further Reading

- Baumgartner, A., Reichel, E., 1975. *The World Water Balance*. R. Oldenbourg Verlag, München Wien.
- Chahine, M.T., 1992. The hydrological cycle and its influence on climate. *Nature* 359, 373–380.
- Grassl, H., Jost, V., Schulz, J., et al., 2000. A Climatological Atlas of Satellite-derived Air–Sea Interaction Parameters over the World's Ocean. Max-Planck Report No. 312. Max-Planck Institute for Meteorology, Hamburg. URL: <http://www.mpimet.mpg.de/Depts/Physik/HOAPS>.
- Hartmann, D.L., 1994. *Global Physical Climatology*. Academic Press, San Diego.
- Josey, S., Kent, E.C., Taylor, P.K., 1999. New insights into the ocean heat budget closure problem from analysis of the SOC air–sea flux climatology. *Journal of Climate* 12, 2856–2880.
- Oki, T., 1999. The global water cycle. In: Browning, K.A., Gurney, R.J. (Eds.), *Global Energy and Water Cycles*. Cambridge University Press, Cambridge, pp. 10–27.
- Taylor PK (ed.) (2000) Intercomparison and validation of ocean–atmosphere energy flux fields. Final report of the Joint WCRP/SCOR Working Group on Air–Sea fluxes, WCRP-112, WMO/TD No. 1036. URL: <http://www.soc.soton.ac.uk/JRD/MET/WGASF>
- Todini, E., Dümenil, L., 1999. Estimating large-scale runoff. In: Browning, K.A., Gurney, R.J. (Eds.), *Global Energy and Water Cycles*. Cambridge University Press, Cambridge, pp. 265–277.
- Trenberth, K.E., Guillemot, C.J., 1999. Estimating evaporation-minus-precipitation as a residual of the atmospheric water budget. In: Browning, K.A., Gurney, R.J. (Eds.), *Global Energy and Water Cycles*. Cambridge University Press, Cambridge, pp. 236–246.

Momentum, Heat, and Vapor Fluxes

PK Taylor, Southampton Oceanography Centre, Southampton, UK

© 2003 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 1, pp 93–100, © 2003, Elsevier Ltd.

Introduction

The maintenance of the Earth's climate depends on a balance between the absorption of heat from the Sun and the loss of heat through radiative cooling to space. For each 100 W of the Sun's radiative energy entering the atmosphere nearly 40 W is absorbed by the ocean – about twice that absorbed in the atmosphere and three times that absorbed by land surfaces. Much of this heat is transferred to the atmosphere by the local sea to air heat flux, a major component of which is caused by the evaporation of water vapor. Although about 90% of the evaporated water falls back into the sea (*see Air Sea Interactions: Freshwater Flux*), the remainder represents about one-third of the precipitation which falls over land. The geographical variation of the atmospheric heating drives the weather systems and their associated winds. The wind transfers momentum to the sea, causing waves and the wind-driven currents. Major ocean currents transport heat poleward and at higher latitudes the sea to air heat transfer significantly ameliorates the climate. Thus the heat, water vapor, and momentum fluxes through the ocean surface form a crucial component of the Earth's climate system.

Having defined the various fluxes and their order of magnitude, this article will review methods of flux measurement and the sources of flux data. The regional and seasonal variation of the fluxes will be summarized. Following a discussion of the accuracy of our present flux estimates, the potential for future improvements will be considered.

Definition of the Fluxes

The momentum flux is the downward transfer of horizontal momentum caused by the drag of the sea surface on the wind. The wave-covered sea surface is continually in motion and, compared with typical land surfaces, appears remarkably smooth to the air flow. For gale force winds the waves may be 10 m or more in height, but the momentum flux is no more than that which would occur over a flat plain. As a result, wind speeds over the ocean tend to be greater than those over land.

The total heat transfer through the ocean surface, the net heat flux, is a combination of several components. The *shortwave radiative flux* (wavelength 0.3–3 μm) is the heat input from the Sun. Around noon on a sunny day this flux may reach about 1000 W m^{-2} but, when averaged over 24 hours, a typical value is 100–300 W m^{-2} , varying with latitude and season. Depending on the solar elevation and the sea state, about 6% of the incident radiation is reflected from the sea surface. Most of the remainder is absorbed in the upper few meters of the ocean. In calm weather, with winds less than about 3 m s^{-1} , a shallow layer may form during the day in which the sea has been warmed by a few degrees Celsius (a 'diurnal thermocline'). However, under stronger winds or at night the absorbed heat becomes mixed down through several tens of meters. Thus, in

contrast to land areas, the typical day to night variation in sea and air temperatures is small, typically less than 1°C. Both the sea and the sky emit and absorb long-wave radiative energy (wavelength 3–50 μm). Because, under most circumstances, the radiative temperature of the sky is colder than that of the sea, the downward long-wave flux is usually smaller than the upward flux. Hence the *net longwave flux* acts to cool the sea surface, typically by 30 W m^{-2} (if cloudy) to 80 W m^{-2} (clear skies).

The turbulent fluxes of sensible and latent heat also typically transfer heat from sea to air. The sensible heat flux is the transfer of heat caused by the difference in temperature between the sea and the air. Over much of the ocean this flux cools the sea by perhaps 10–20 W m^{-2} . However, where cold wintertime continental air flows over warm ocean currents, for example, the Gulf Stream region off the eastern seaboard of North America, the sensible heat flux may reach 100 W m^{-2} . Conversely, in regions like the summertime North Pacific Ocean, warm winds blowing over a colder ocean may result in a small sensible heat flux into the ocean. Under most weather conditions the evaporation of water vapor from the sea surface results in a water vapor flux from the sea to the air. The latent heat flux is the heat absorbed on vaporization of the water. This heat is released to warm the atmosphere when the vapor condenses to form clouds. Usually the latent heat flux is significantly greater than the sensible heat flux, being on average 100 W m^{-2} or more over large areas of the ocean. Over regions such as the Gulf Stream, latent heat fluxes of several hundred W m^{-2} are observed. In foggy conditions, with the air warmer than the sea, condensation may occur on the sea surface, and the vapor flux and latent heat flux are directed from air to sea. In summertime over the fog-shrouded Grand Banks off Newfoundland, the mean monthly latent heat transfer is directed into the ocean, but this is a very exceptional case.

Measuring the Fluxes

For the radiative fluxes, the standard method is to measure the voltage generated by a thermopile exposed to the incident radiation. A pyranometer, mounted in gimbals for use on a ship or buoy, is used to measure the incoming shortwave radiation (*Figure 1*). For better accuracy the direct and diffuse components should be determined separately but at present this is rarely done over the sea. The reflected short-wave radiation is normally calculated using tabulated values of the albedo for different solar elevations. Pyrgeometers, used for longwave radiation, are similar to pyranometers but have a coated dome to filter out the shortwave radiation. For these the use of gimbals is less important, but a clear sky view is required and corrections for the dome temperature and short wave leakage are needed. Again, only the downward component is normally measured; the upward component is calculated from the temperature and the emissivity of the sea surface.



Figure 1 A pyranometer used for measuring short-wave radiation. The thermopile is covered by two transparent domes. Photograph courtesy of Southampton Oceanography Centre.

The turbulent fluxes may be measured directly in the near surface atmosphere using the eddy correlation method. For example, if upward-moving air in the turbulent eddies is on average warmer and moister than the downward-moving air, then there is an upward flux of sensible heat and water vapor and hence also of latent heat. Similarly, the momentum flux may be determined from the correlation between horizontal and vertical wind fluctuations. Accurate eddy correlation measurements over the ocean are difficult. Since a large range of eddy sizes may contribute to the flux, fast-response sensors capable of sampling at 10 Hz or more must be exposed for periods of order 30 minutes for each flux determination. For instrumentation mounted on a buoy or ship the six components of the wave-induced motion must be measured and removed in the signal processing procedure. The distortion both of the turbulence and of the mean wind by ship, buoy, or fixed tower must be minimized and, if possible, corrected for. While three-component ultrasonic anemometers (Figure 2) are relatively robust, the sensors for measuring fluctuations in temperature and humidity have previously been fragile and, in the marine atmosphere, prone to contamination by salt particles and sea spray. Improved sonic thermometry, and water vapor instruments using microwave refractometry or differential infrared absorption, are relatively recent developments. Thus eddy correlation measurements are not routinely obtained over the ocean; rather they are used in air–sea interaction experiments to calibrate other flux estimation methods.

In the inertial dissipation method, fluctuations of the wind, temperature, or humidity at a few hertz are measured and related, through turbulence theory, to the fluxes. This method is less sensitive to flow distortion or platform motion but relies on various assumptions regarding the formation and dissipation of turbulent quantities that may not always be valid. It has been implemented on some research ships to increase the range of available flux data.

The bulk (aerodynamic) formulas are the most commonly used method of flux estimation. The flux is determined from the difference between the temperature, humidity or wind at some measurement height, z , and the value assumed to exist at the sea surface – respectively the sea surface temperature, 98% saturation humidity (to allow for salinity effects), and zero wind (or any non-wind-induced water current). Thus the flux F_x



Figure 2 The sensing head of a three-component ultrasonic anemometer. The wind components are determined from the different times taken for sound pulses to travel in either direction between the six ceramic transducers. Photograph courtesy of Southampton Oceanography Centre.

of some quantity x is given by eqn [1], where ρ is the air density and U_z is the wind speed at the measurement height.

$$F_x = \rho U_z C_{xz} (x_z - x_0) \quad [1]$$

While appearing intuitively correct (for example, blowing over a hot drink will cool it faster) these formulas can also be derived from turbulence theory. The value for the transfer coefficient, C_{xz} , characterizes both the surface roughness applicable to x and the relationship between F_x and the vertical profile of x . The transfer coefficient varies with the atmospheric stability, which itself depends on the momentum, sensible heat, and water vapor fluxes as well as the measurement height. Thus, although it may appear simple, eqn must be solved by iteration, initialized using the equivalent neutral value of C_{xz} at some standard height (normally 10 m), C_{x10n} . Typical neutral values are shown in Table 1.

Many research problems remain. For example: the increase of C_{D10n} at higher wind speeds must depend on the varying sea state, but can the latter be successfully characterized by the ratio of the predominant wave speed to the wind speed (the wave age), or by factors like the wave height and steepness, or is a complete spectral representation of the wave field required? What are the effects of waves propagating from other regions at varying angles to the wind (i.e., swell waves)? What is the exact behaviour of C_{D10n} in low wind speed conditions? Since C_{E10n} and C_{H10n} are not well defined by the available experimental

Table 1 Typical values (with estimated uncertainties) for the transfer coefficients. Neither the low wind speed formula for C_{D10n} nor the wind speed below which it should be applied, are well defined by the available, very scattered, experimental data. It should be taken simply as an indication that, at low wind speeds, the surface roughness increases as the wind speed decreases

Flux	Transfer coefficient	Typical values
Momentum	Drag coefficient, C_{D10n}	$0.61 (\pm 0.05) + 0.063 (\pm 0.005) U_{10n} (U_{10n} > 3 \text{ ms}^{-1})$ $0.61 + 0.57/U_{10n} (U_{10n} < 3 \text{ ms}^{-1})$
Sensible heat	Stanton no., C_{H10n}	$1.1 (\pm 0.2) \times 10^{-3}$
Latent heat	Dalton number, C_{E10n}	$1.2 (\pm 0.1) \times 10^{-3}$

data, recent implementations of the bulk algorithms have used theoretical models of the ocean surface (known as surface renewal theory) to predict these quantities from the momentum roughness length.

Sources of Flux Data

Until recent years the only routinely available sources of data were the weather reports from merchant ships. Organized as part of the World Weather Watch system of the World Meteorological Organization, these voluntary observing ships (VOS) are asked to transmit coded weather messages at 00.00, 06.00, 12.00 and 18.00 GMT daily, and to record a more detailed observation in a weather logbook. The very basic instruments used normally include a barometer and a means of measuring air temperature and humidity: wet and dry bulb thermometers mounted in a hand-swung sling psychrometer or in a fixed, louvered Stevenson screen. Sea temperature is obtained using a thermometer and an insulated bucket, or by reading the temperature of the engine cooling water intake. Depending on which country recruited the VOS, an anemometer and wind vane might be provided, or the ship's officers might be asked to estimate the wind velocity from the sea state using a tabulated Beaufort scale. Because of the problems of adequately siting an anemometer and maintaining its calibration, these visual estimates are not necessarily considered inferior to anemometer-based values.

The bulk formulas are used to calculate the turbulent fluxes from the VOS observations. However, in many cases the accuracy is poor. In particular, a large ship can produce significant changes in the local temperature and wind flow. The radiative fluxes must be estimated from the observer's estimate of the cloud amount plus, for short wave, the solar elevation, or for long wave, the sea and air temperature and humidity. The unavoidable observational errors and the crude form of the radiative flux formulas imply that large numbers of reports are needed, and correction schemes must be applied, before satisfactory flux estimates can be obtained. Though there are presently nearly 7000 VOS, they tend to be concentrated in the main shipping lanes. While coverage over most of the North Atlantic and North Pacific is adequate to provide monthly mean flux values, elsewhere data is mainly restricted to relatively narrow, major trade routes. For most of the Southern Hemisphere the VOS data are only capable of providing useful values if averaged over several years, and reports from the Southern Ocean are very few indeed. These problems must be borne in mind when studying the flux distribution maps

shown in marine climatological atlases, of which examples are presented below.

Satellite-borne sensors can overcome these sampling problems. 'Passive' sensors measure the radiation emitted from the sea surface and the intervening atmosphere at visible, infrared, or microwave frequencies; 'active' sensors transmit microwave radiation and measure the returned signal. The problem is to develop methods of determining the fluxes from the various satellite data that can be obtained. For example, sea surface temperature has been routinely determined using visible and infrared radiometers since about 1980. However, the data must be frequently checked against ship and buoy values to avoid errors due to changes in atmospheric aerosol content that may follow volcanic eruptions. Satellite-derived fields of net surface shortwave radiation are available; values for the net surface longwave radiation are less accurate. The surface wind velocity can be determined to good accuracy by active scatterometer sensors by measuring the microwave radiation back-scattered from the sea surface. The determination of near-surface air temperature and humidity from satellites is hindered by the relatively coarse vertical resolution of the retrieved data. Thus the radiation emitted by the near-surface air is dominated by that originating from the sea surface. Statistically based algorithms for determining the near-surface humidity have been successfully developed. More recently, neural network techniques have been applied to retrieving both air temperature and humidity; however, there is presently no routinely available product. Thus the satellite flux products for which useful accuracy has been demonstrated on a global basis are presently limited to momentum, short-wave radiation, and latent heat flux.

Numerical weather prediction (NWP) models (as used in weather forecasting centers) estimate values of the air-sea fluxes as part of the calculations. Assimilating much of the available data from the World Weather Watch system, including satellite data, radiosonde profiles, and surface observations, NWP models are potentially the best source of flux data. However, there are a number of problems. The vertical resolution of these models is relatively poor and many of the near-surface processes have to be represented in terms of larger-scale parameters. Improvements to NWP models are judged on the resulting quality of the weather forecasts, not on the accuracy of the surface fluxes, which may become worse. Indeed, the continual introduction of model changes results in time discontinuities in the output variables. Thus the determination of interannual variations is difficult and, for that reason, centers such as the European Centre for Medium Range Weather Forecasting (ECMWF) and the US National Centers for Environmental Prediction (NCEP) have recently reanalyzed the past weather

over several decades. The surface fluxes from these reanalyses are receiving much study. Those presently available appear less accurate than fluxes derived from VOS data in regions where there are many VOS reports; in sparsely sampled regions the model fluxes are more accurate. Particular weaknesses for the models are in the shortwave radiation and latent heat fluxes. New reanalyses are underway and efforts are being made to improve the flux estimates; eventually these reanalyses will provide the best source of flux data for many purposes.

The Regional and Seasonal Variation of the Momentum Flux

The main features of the wind regimes over the global oceans have long been recognized and descriptions are available in

many books on marine meteorology. The major features of the wind stress variability derived from ship observations from the period 1980 to 1993 are summarized here, using plots for January and July to illustrate the seasonal variation (Figure 3).

In Northern Hemisphere winter (Figure 3(a)) large wind stresses due to strong mid-latitude westerly winds occur in the North Atlantic and the North Pacific west of Japan. To the south, the extratropical high-pressure zones result in low wind stress values, and south of these is the belt of north-east trade winds. The very light winds of the Intertropical Convergence Zone (ITCZ) lie close to the Equator in both oceans. In the summertime Southern Hemisphere the south-east trade wind belt is less pronounced. The extratropical high-pressure regions are extensive but, despite it being summer, high winds and significant wind stress exist in the mid-latitude Southern

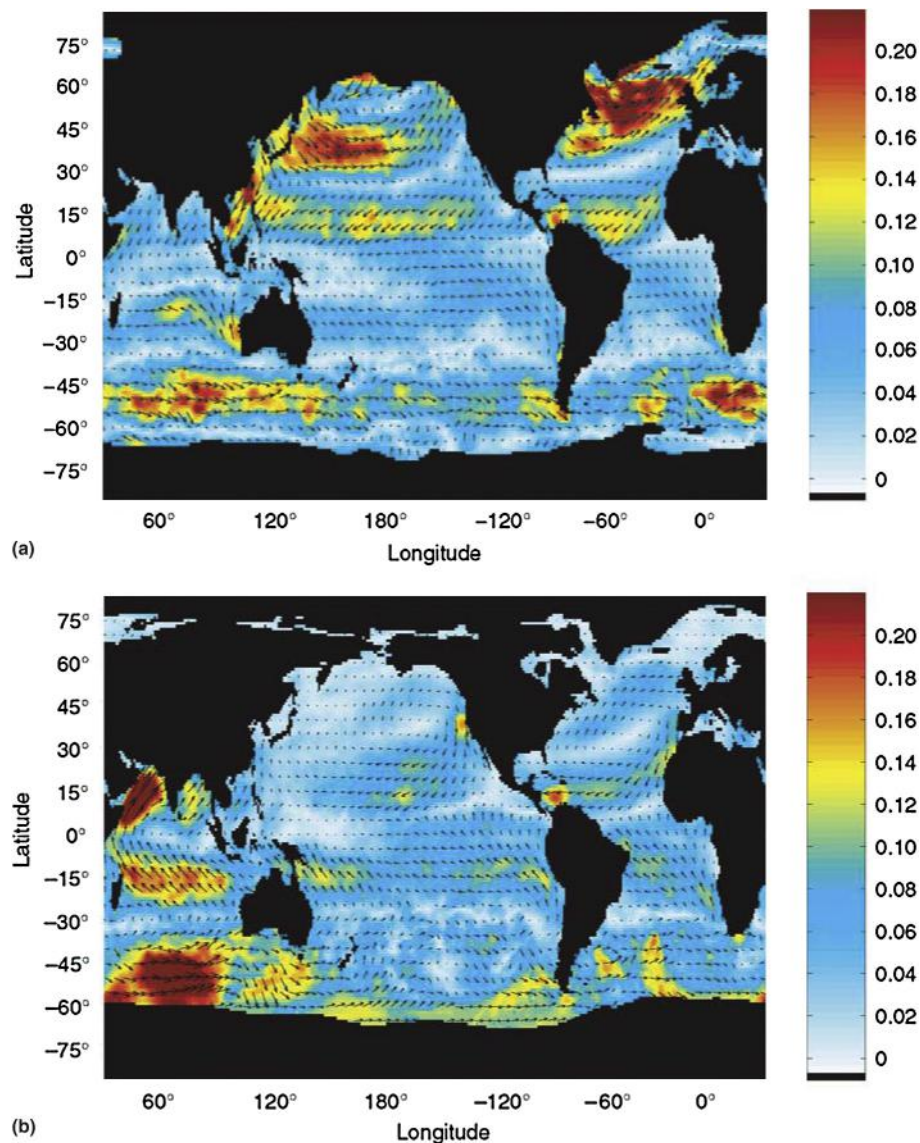


Figure 3 Monthly vector mean wind stress (N m^{-2}) for (a) January and (b) July calculated from voluntary observing ship weather reports for the period 1980 to 1993. Adapted from Josey SA, Kent EC, and Taylor PK (1998). The Southampton Oceanography Centre (SOC) Ocean–Atmosphere Heat, Momentum and Freshwater Flux Atlas, SOC Report No. 6.

Ocean. The north-east monsoon dominates the wind patterns in the Indian Ocean and the South China Sea (where it is particularly strong). In the latter regions the ITCZ is a diffuse area south of the Equator with relatively strong south-east trade winds in the eastern Indian Ocean.

In Northern Hemisphere summer (Figure 3(b)) the wind stresses in the mid-latitude westerlies are very much decreased. Both the north-east and the south-east trade wind zones are evident respectively to the north and south of the ITCZ, which mainly lies north of the Equator. The south-east trades are particularly strong in the Indian Ocean and feed air across the Equator into a very strong south-westerly monsoon flow in the Arabian Sea. These ship data indicate very strong winds in the Southern Ocean south-west of Australia. Such winds are also evident in satellite scatterometer data, which suggest that the winds in the Pacific sector of the Southern Ocean, while still strong, are somewhat less than those in the Indian Ocean sector. In contrast, the ship data appear to show light winds. The reason is that in wintertime there are practically no VOS observations in the far South Pacific. The analysis technique used to fill in the data gaps has, for want of other information, spread the light winds of the extratropical high-pressure region farther south than is realistic; this is a good example of the care needed in interpreting the flux maps available in many atlases.

The Regional and Seasonal Variation of the Heat Fluxes

The global distribution of the mean annual net heat flux is shown in Figure 4(a). Averaged over the year, the ocean is heated in equatorial regions and loses heat in higher latitudes, particularly in the North Atlantic. However, this mean distribution is somewhat misleading as the plots for January (Figure 4(b)) and July (Figure 4(c)) illustrate. The ocean loses heat over most of the extratropical winter hemisphere and gains heat in the extratropical summer hemisphere. It is only because the tropical oceans are heated throughout the year, and atmospheric moisture from trade wind zones converges in the ITCZ, that the tropics are so important with regard to driving the atmospheric circulation. The major regions of ocean cooling occur in winter over the Gulf Stream and the Kuroshio currents. However, in summer the long period of day-light in these mid-latitude regions results in mean short-wave radiation values similar to or larger than those observed in equatorial regions. Thus the mean monthly short-wave flux is greater than the cooling induced by the combined latent heat and net long-wave fluxes. At a more typical mid-latitude site the ocean cools in winter and warms in summer, in each case by around 100 W m^{-2} . The annual mean flux is small, of order 10 W m^{-2} , but cannot be neglected because of the very large ocean areas involved.

Considering now the interannual variation of the surface fluxes, the major large-scale feature over the global ocean is the El Niño Southern Oscillation system in the equatorial Pacific Ocean. In the eastern equatorial Pacific the change in the net heat flux under El Niño conditions is around 40 W m^{-2} . For extratropical and mid-latitude regions the interannual variability of the summertime net heat flux is typically

about $20\text{--}30 \text{ W m}^{-2}$, being dominated by the variations in latent heat flux. In winter the typical variability increases to about $30\text{--}40 \text{ W m}^{-2}$, although in particular areas (such as over the Gulf Stream) variations of up to 100 W m^{-2} may occur. The major spatial pattern of interannual variability in the North Atlantic is known as the North Atlantic Oscillation (NAO). This represents a measure of the degree to which mobile depressions, or alternatively near-stationary high-pressure systems, occur in the mid-latitude westerly zone.

Discussion: Accuracy of Flux Estimates and Future Trends

We have seen that, although the individual flux components are of the order of hundreds of W m^{-2} , the net heat flux and its interannual variability over much of the world ocean is of the order of tens of W m^{-2} . Furthermore, it can be shown that a flux of 10 W m^{-2} over one year would, if stored in the top 500 m of the ocean, heat that entire layer by about 0.15°C . Temperature changes on a decadal time scale are at most a few tenths of a degree, so the global mean budget must balance to better than a few W m^{-2} . For these various reasons there is a need to measure the flux components, which vary on many time and space scales, to an accuracy of a few W m^{-2} . Given the available data sources and methods of determining the fluxes described above, it is not surprising that this accuracy at present cannot be achieved.

To take an example, in calculating the flux maps shown in Figure 4 many corrections were applied to the VOS observations in an attempt to remove biases caused by the observing methods. For example, air temperature measurements were corrected for the 'heat island' caused by the ship heating up in sunny, low-wind conditions. The wind speeds were adjusted depending on the anemometer heights on different ships. Corrections were applied to sea temperatures calculated from engine room intake data. Despite these and other corrections, the global annual mean flux showed about 30 W m^{-2} excess heating of the ocean. Previous climatologies calculated from ship data had shown similar biases and the fluxes had been adjusted to remove the bias, or to make the fluxes compatible with estimates of the meridional heat transport in the ocean. However, comparison of the unadjusted flux data with accurate data from air-sea interaction buoys showed good agreement between the two. This suggests that adjusting the fluxes globally is not correct and that regional flux adjustments are required; however, the exact form of these corrections is presently not known.

In the future, computer models are expected to provide a major advance in flux estimation. Recently, coupled numerical models of the ocean and of the atmosphere have been run for many simulated years, during which the modeled climate has not drifted. This suggests that the air-sea fluxes calculated by the models are in balance with the simulated oceanic and atmospheric heat transports. However, it does not imply that at present the flux values are realistic. Errors in the short-wave and latent heat fluxes may compensate one another; indeed, in a typical simulation the sea surface temperature stabilized to a value that was, over large regions of the ocean, a few degrees different from that

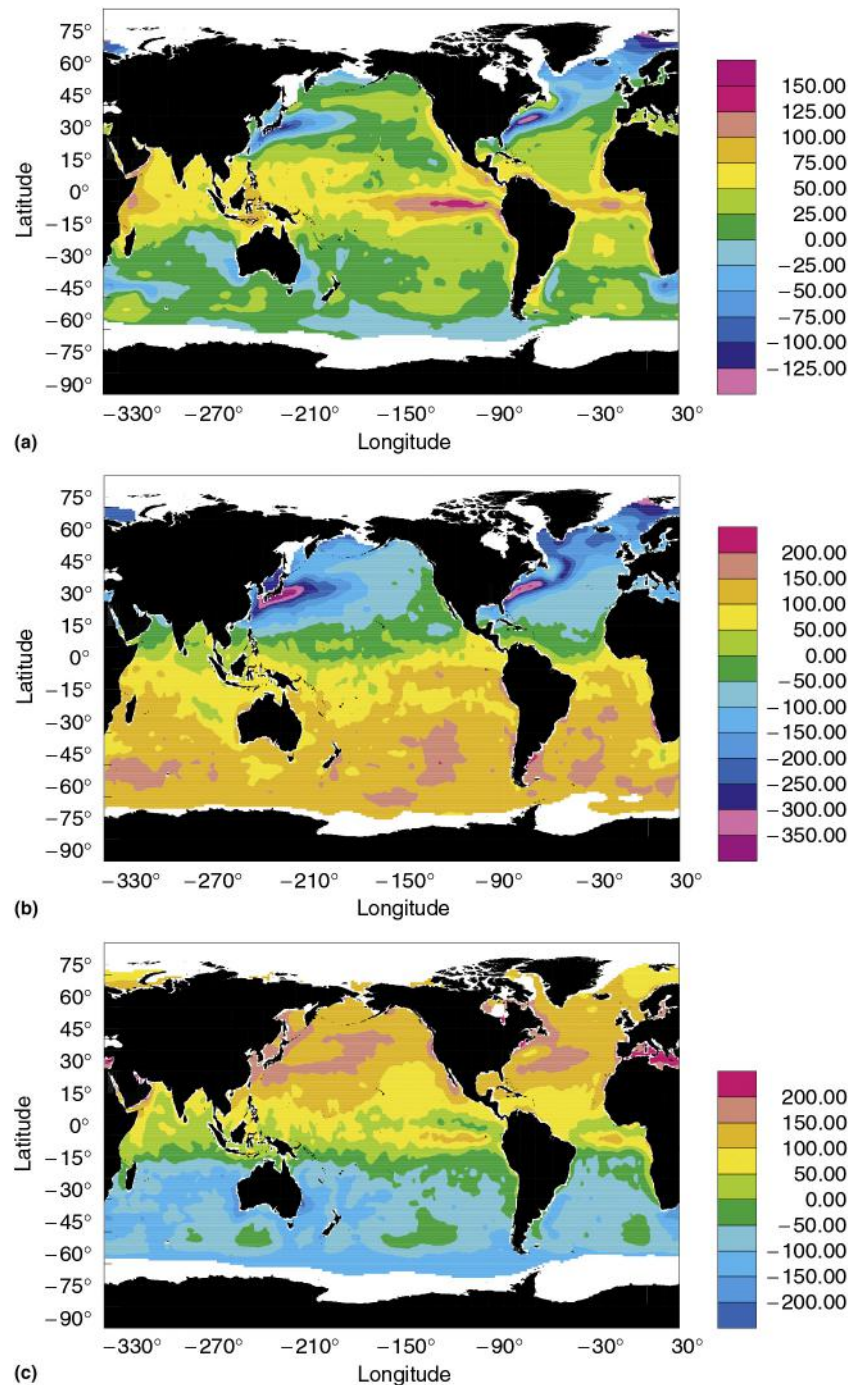


Figure 4 Variation of the net heat flux over the ocean; positive values indicate heat entering the ocean: (a) annual mean; (b) January monthly mean; (c) July monthly mean. Adapted from Josey SA, Kent EC, and Taylor PK (1998). The Southampton Oceanography Centre (SOC) Ocean–Atmosphere Heat, Momentum and Freshwater Flux Atlas, SOC Report No. 6.

which is observed. Nevertheless, the estimation of flux values using climate or NWP models is a rapidly developing field and improvements will doubtless have occurred by the time this article has been published. There will be a continued need for in-situ and satellite data for assimilation into the models and for model development and verification. However, it seems very likely that in future the

most accurate routine source of air–sea flux estimates will be from numerical models of the coupled ocean–atmosphere system.

See also: **Aerosols:** Observations and Measurements; Role in Radiative Transfer. **Air Sea Interactions:** Freshwater Flux; Sea

Surface Temperature; Surface Waves. **Boundary Layer (Atmospheric) and Air Pollution:** Observational Techniques *In Situ*, Observational Techniques: Remote; Surface Layer. **Climate and Climate Change:** Climate Variability: North Atlantic and Arctic Oscillation. **Gravity Waves:** Buoyancy and Buoyancy Waves: Optical Observations. **Numerical Models:** Regional Prediction Models. **Tropical Cyclones and Hurricanes:** Hurricanes: Observation.

Further Reading

Browning, K.A., Gurney, R.J. (Eds.), 1999. Global Energy and Water Cycles. Cambridge University Press, Cambridge.
 Dobson, F., Hasse, L., Davis, R. (Eds.), 1980. Air–Sea Interaction, Instruments and Methods. Plenum Press, New York.

Garraff, J.R., 1992. The Atmospheric Boundary Layer. Cambridge University Press, Cambridge.
 Geernaert, G.L., Plant, W.J., 1990. Surface Waves and Fluxes. In: Current Theory, vol. 1. Kluwer, Academic, Dordrecht.
 Isemer, H.-J., Hasse, L., 1987. The Bunker Climate Atlas of the North Atlantic Ocean. In: Air–Sea Interactions, vol. 2. Springer-Verlag, Berlin.
 Josey SA, Kent EC and Taylor PK (1999) The Southampton Oceanography Centre (SOC) Ocean–Atmosphere Heat, Momentum and Freshwater Flux Atlas, SOC Report No. 6, p. 30 1figs. (Available from The Library, Southampton Oceanography Centre, European Way, Southampton, SO14 3ZH, UK.)
 Kraus, E.B., Businger, J.A., 1994. Atmosphere–Ocean Interaction, second ed. Oxford University Press, New York.
 Stull, R.B., 1988. An Introduction to Boundary Layer Meteorology. Kluwer Academic, Dordrecht.
 Wells, N., 1997. The Atmosphere and Ocean: A Physical Introduction, second ed. Taylor and Francis, London.

Sea Surface Temperature

WJ Emery, University of Colorado, Boulder, CO, USA

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

Sea surface temperature (SST) is the interface between the ocean and the overlying atmosphere. As such, it controls the exchange of heat and gases between the atmosphere and ocean. It is also the longest and most widely measured parameter in the ocean. Traditionally measured by inserting a mercury-in-glass thermometer in a bucket sample of surface water from a sailing ship or from a coastal observing station, it is now widely directly measured by drifting and moored buoys or from engine-cooling water on ships. The need for global SST observations has driven the use of infrared satellite data to estimate SST, which has now extended to the use of passive microwave data as well. The satellite measurements distinguish themselves from the in situ measurements in that they measure the skin of the ocean rather than the subsurface 'bulk' SST measured by ships and buoys.

Introduction

As the controlling variable of heat, momentum, salt, and gas fluxes between the ocean and the atmosphere, the sea surface temperature (SST) has always been a topic of interest to scientists. It is also the easiest oceanographic parameter to observe, and observations of some form of SST extend back to the time of the early Greek scientists. In addition to its relationship to ocean-atmosphere fluxes, the SST also relates directly to a number of human concerns. For example, the success of fisheries and fishermen can be enhanced by knowledge of the SST pattern. For many years, SST was measured by taking a 'bucket sample' of the surface waters and then measuring its temperature. As ships evolved to powered vessels, this practice of SST bucket sampling had to be abandoned, and the practice of using measurements of the temperature of the cooling water coming to the ship's engines was used. These were known as 'ship injection temperatures' (as the SST sensor was 'injected' into the water stream), and these data suffer from many basic problems (engine room heating, depth of the water intake, etc.). The advent of satellite-tracked drifting buoys introduced a platform that could measure the SST and then report it in near-real time. Thought to be less noisy and more accurate than ship SSTs, the drifting buoy SST data became the standard for both the calibration and validation of satellite infrared estimates of SST. Over the years, these buoy SSTs have been used to calculate the infrared algorithm coefficients for the computation of the SST. The problem is that a buoy cannot measure the temperature of the 10 μm thin 'skin of the ocean,' which is the layer that radiates out into space. Thus, satellite infrared measurements of SST are of this skin temperature and not of the deeper bulk SST measured by the ships and buoys. It is the difference between the skin and bulk temperature that is directly related to the wind speed and the net air-sea heat flux. In the past, the bulk SST has been used in the computation of air-sea heat flux terms using empirical 'bulk' formulas. These continue to be used, but research efforts are underway to transform these computations into a satellite-only calculation. These efforts will explicitly involve skin and bulk SSTs, giving a more physical basis to the connections between heat flux and the skin and bulk SSTs. These improvements are critical for the modeling of climate change, since the SST is the boundary that

connects the model atmosphere to the model ocean. Any incorrect specification of the global SST patterns will lead directly to errors in the model simulations.

History of SST Measurements and Applications

One of the earliest uses of SST was when Benjamin Franklin mapped the position of the Gulf Stream from SST measurements made from the mail packet ships that he rode to and from Europe. Fortunately for Franklin, the Gulf Stream has a very sharp thermal contrast on its western edge, where the cold shelf waters coming from the north meet the warm Gulf Stream waters advecting to the northeast. Together with Timothy Folger, he published a map ([Figure 1](#)) that described the position of the Gulf Stream, advising ship captains not to sail against this current, which was strong enough to hold a sailing vessel still. This practical application of SST made it possible to reduce the crossing times for ships sailing from the Americas to Europe.

As one of the easiest measurements in oceanography, SST became widely observed, whether as a sample taken from a pier or beach or as a bucket sample from a ship on the open ocean. The collection of a bucket of water whose temperature was measured as an estimate of SST became common practice among the sailing ships carrying the world's commerce. When the ships' logs became the source of global information on winds and currents, the SST information was also compiled from the ships' logs. This information was routinely published, along with the currents and winds from the ships' logs, as part of 'sailing directions' put out first by the US Navy and later by the US Coast Guard. The same practice was introduced later in Europe and became an operational reality for most oceangoing vessels.

The collection of a bucket sample from sailing ships that travel at speeds between 5 and 15 knots was easy. For powered ships traveling at 20–30 knots, it was no longer possible to collect a bucket sample for SST measurement. Instead, the temperature of the water used to cool the ship's engines was measured as the SST. This temperature was called the 'injection' temperature because, as mentioned in the Introduction, the sensor was 'injected' into the pipe carrying this cooling water to



Figure 1 The Poupard version of the Franklin–Folger Gulf Stream chart published in Philadelphia in 1786 with Franklin's *Maritime Observations*.

the heat exchangers. One of the fundamental problems with these injection temperatures was the location of the inlet pipe for this cooling water, which was positioned far down the ship's hull, generally collecting water from about 5 m depth – far away from the sea surface. In addition, the heat in the engine room was known to heat the cooling water and the associated thermometer, resulting in a warm bias for the SSTs. Other sources of error were the reading of the analog gauge, the hand recording of the data in the ship's log, and the radio broadcast of the 'SST' to be included in the SST database.

Satellite SST

In spite of all these limitations, ship SSTs were initially used to match and adjust infrared satellite measurements of the SST. Later, satellite-tracked drifting buoys were equipped with temperature ports sticking out from the hull to measure the SST as the buoy traveled around the ocean. Most of these buoy SST sensors were initially calibrated to $\pm 0.1^\circ\text{C}$, but since the buoys are considered expendable there is no postdeployment calibration, and it is not known how well the buoys retain their calibration. There are about three different hull types used today in these buoys, with slightly different configurations of the hull SST sensors. All of these buoys float in the active near-surface layer and move up and down with the wave field. As a consequence, the buoy SST represents the temperature between the surface and 1–2 m depth. The best estimate of the accuracy of these data can be made by considering the mean

difference and variability of the difference between contemporary SST measurements. This value is 0.4°C , which can be taken as an overall error limit for buoy-measured bulk SST.

It is important to realize that past practice has been to compute the satellite infrared SST algorithm coefficients from regression with nearly coincident buoy SST data. In this approach, one assumes not only that the satellite and buoys measure the same SST, but also that the buoy SSTs have no errors themselves. That is not to say that people assume the buoy SSTs have no error, but the practice of using them to find the algorithm coefficients through linear regression implies that the buoy SSTs are error-free. It appears that the source of this conflict is the fact that satellite infrared systems and their calibration systems do not retain their calibrations over long periods of time and that regression to in situ SST is required to overcome these drifts. Since buoy SSTs are the best of the present in situ SSTs, they are used to supply the in situ SST observations in spite of the fact that they do not and cannot measure the skin SST, as any direct physical contact with the skin layer will temporarily disrupt it. Observations of the effects of breaking waves on the presence of the skin layer have shown that while a breaking wave does indeed destroy the skin layer, the skin forms again after just a few seconds. So, in spite of strong winds and breaking waves, the skin SST layer is present most of the time and must be considered in the remote sensing of SST. Before various groups and agencies will transform their computations to skin SST, there must be a source of in situ skin SST to replace the buoy SSTs that are used at present. We will discuss this point further in this article.

SST Climatology

One of the benefits of being relatively easy to observe is the abundance of data that exists both geographically and over time. This excellent data coverage makes it possible to compute a climatology of SST conditions, which is a long-term average map of SST conditions (usually over the 12-month period of an annual cycle). Many different climatologies have been computed for SST, starting with strictly ship measurements before about 1970 and followed by a mix of satellite, ship, and buoy SSTs since then. One must always be careful to determine exactly which periods are covered by such a climatology and what data went into the computation of the climatology.

For this article, we will use a climatology that includes satellite infrared, buoy, and ship SST data. As part of this analysis, satellite SSTs were filtered to eliminate obvious outliers. The input buoy and ship SST data sets were also filtered to remove large errors. All three of these input data sets were to save space. We only present four images to represent the 12-month annual cycle. The first is a global map of SST for the month of January (**Figure 2**), which represents SST conditions for the Northern Hemisphere winter.

In this map, a number of basic features are readily obvious. The warmest temperatures are in the tropics, particularly in the Pacific and Indian Oceans; the warmest temperatures are found in the 'warm pool' of the western tropical Pacific. There are cold features along the west coasts of South America and South Africa, which correspond to upwelling events most active in the

austral summer. Cold temperatures extend a bit farther north in the Southern Ocean than they do in the Northern Hemisphere, which is primarily a consequence of the open character of the Southern Ocean compared with the geographically restricted waters of the Arctic.

The representative Northern Hemisphere spring SST map in **Figure 3** shows a modest shift in this temperature distribution. The warmest temperatures have increased, particularly in the equatorial Indian Ocean, which has increased from about 28 °C to about 30 °C. The Western Pacific Warm Pool has expanded slightly and increased in temperature. There is now a distinct band of warm temperatures in the equatorial Pacific. The upwelling zones off southwestern South America and South Africa have decreased in size. The colder SSTs have remained largely the same.

Turning to the Northern Hemisphere summer, we look at the July SST climatology in **Figure 4**. In this case, the maximum temperatures in the tropics have actually cooled slightly, reflecting the decrease in solar insolation on a global level. The warm band in the tropical Atlantic has weakened, as has the equatorial warm band in the tropical Indian Ocean. Even the warm pool in the western tropical Pacific has weakened in both magnitude and areal coverage. The upwelling zones off western South America and western South Africa have again expanded, and both show warm tongues that extend westward from the northernmost extent of the colder upwelling water. There is a corresponding cold upwelling region off North America consistent with the seasonal shift to northerly upwelling winds off that coast. The same is true off northwest Africa.

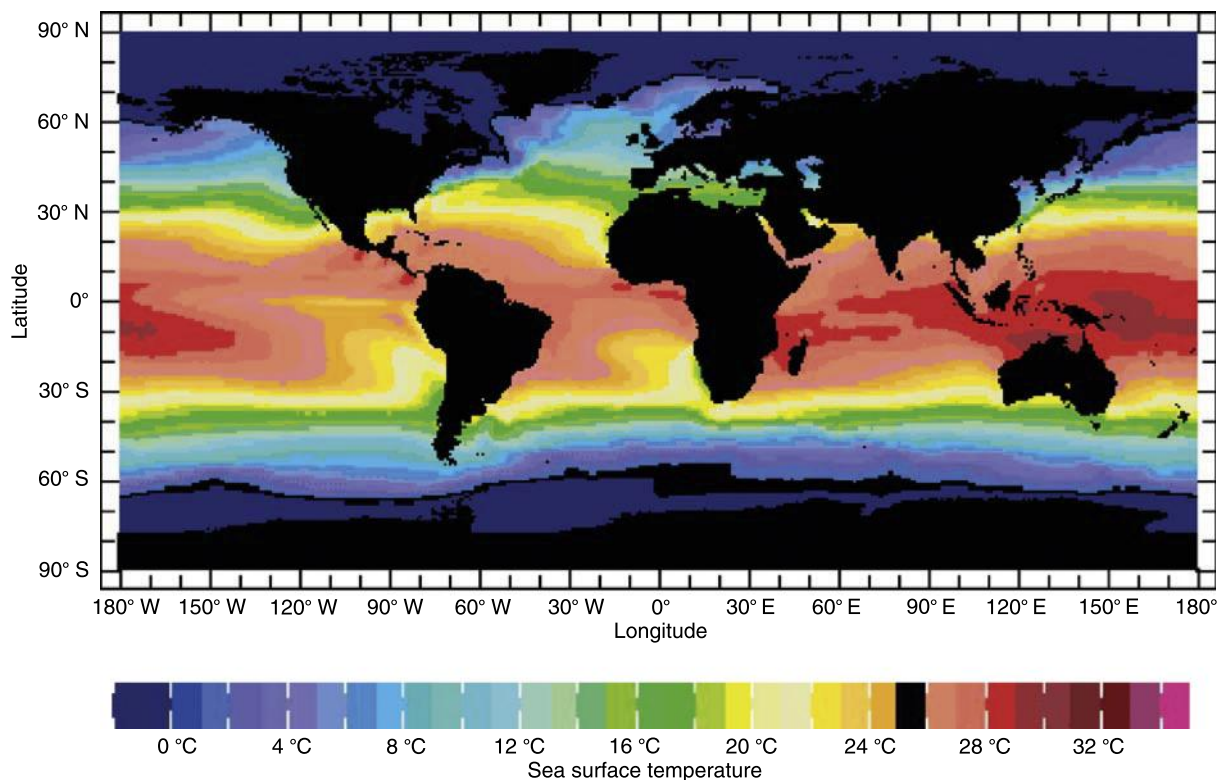


Figure 2 SST climatology for January. Reproduced with permission from Reynolds, R.W., Smith, T.M., 1995. A high resolution global sea surface temperature climatology. *Journal of Climate* 8, 1571–1583.

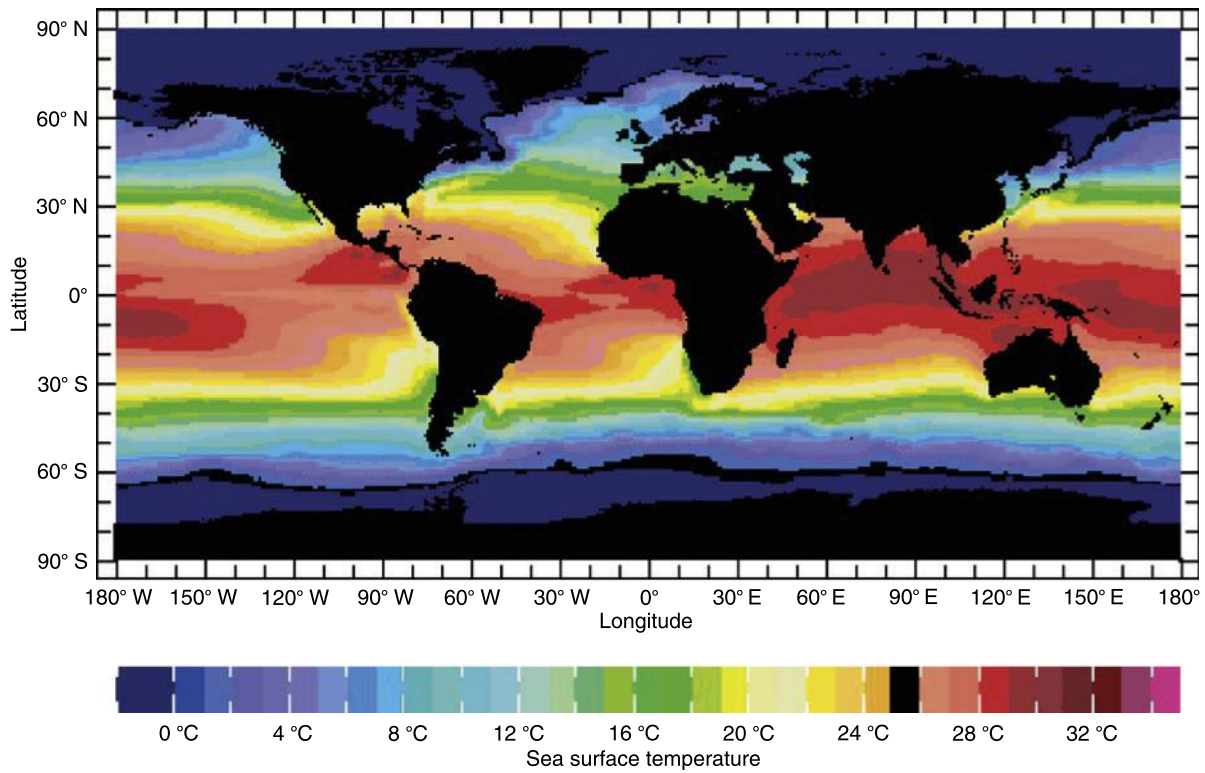


Figure 3 April SST climatology. Reproduced with permission from Reynolds, R.W., Smith, T.M., 1995. A high resolution global sea surface temperature climatology. *Journal of Climate* 8, 1571–1583.

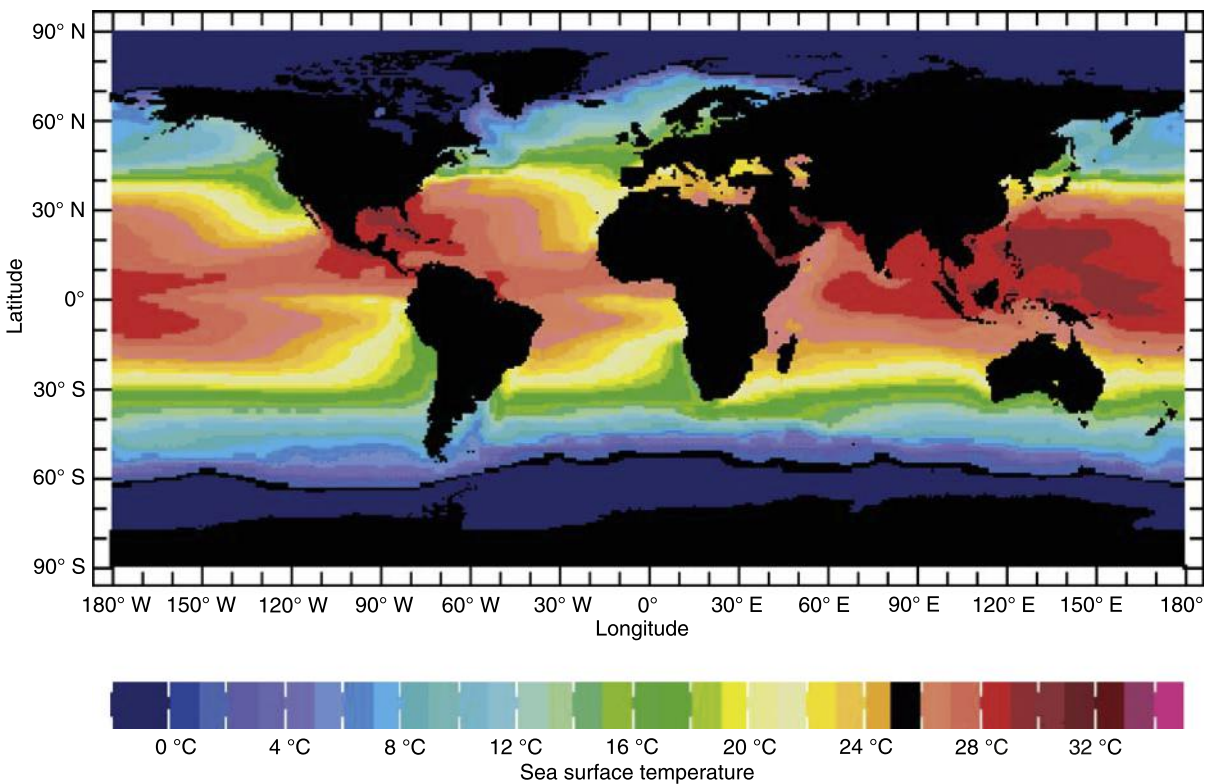


Figure 4 Mean July SST. Reproduced with permission from Reynolds, R.W., Smith, T.M., 1995. A high resolution global sea surface temperature climatology. *Journal of Climate* 8, 1571–1583.

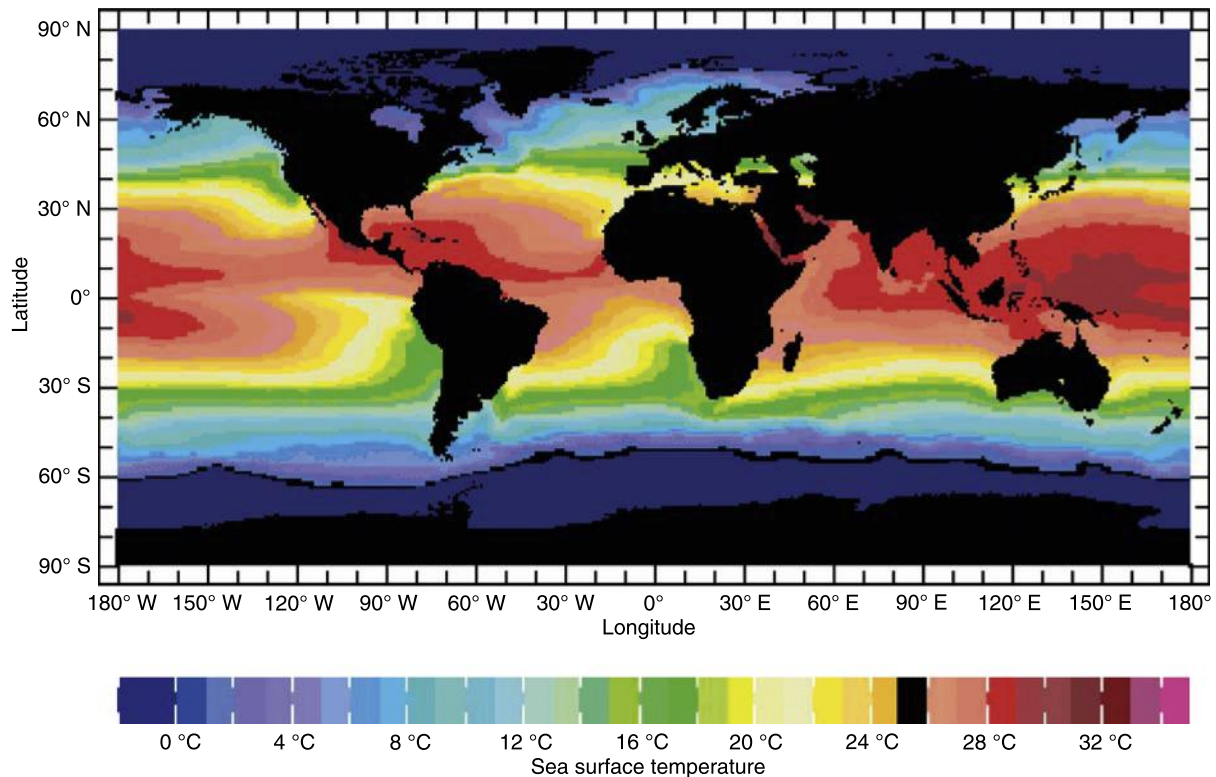


Figure 5 Mean October SST. Reproduced with permission from Reynolds, R.W., Smith, T.M., 1995. A high resolution global sea surface temperature climatology. *Journal of Climate* 8, 1571–1583.

To complete the cycle, we look at the Northern Hemisphere fall (October) map in [Figure 5](#). In this map, all of the west coast upwelling regions have weakened slightly. In particular, the upwelling region of northwest Africa has disappeared. In the Southern Hemisphere, the upwelling regions appear to have stretched farther to the west. The equatorial warm regions have weakened in magnitude, although they are about the same in geographic coverage as they were in July. The colder waters have not changed much since the last season (summer).

All of these maps have used a mixture of satellite infrared measurements, drifting and moored buoy SST measurements, and ship SST measurements. In this application, the satellite skin SST is adjusted to match coincident drifting buoy SSTs, as introduced in this article. Thus, these maps really represent a ‘pseudobulk SST’ due to the overwhelming number of infrared SST observations as compared to the in situ buoy and ship observations. Still, the general seasonal pattern of the SST is clearly apparent, and it is not likely that an adjustment to skin SST would show any substantially different SST patterns. SST patterns computed from various satellite SST algorithms all look very similar; it is the absolute temperature value that is different, and the skin SST must be considered when addressing questions such as air–sea heat and gas exchange.

Skin SST

Due to its very high emissivity, the ocean is considered to very nearly approximate a ‘blackbody.’ This long-wave heat emission

is directly proportional to the skin SST, which is the only SST that interacts with the overlying atmosphere. Having a thickness of between 5 and 10 μm , the skin of the ocean can easily be destroyed by breaking ocean waves. When this happens, the skin reforms within 3–6 s, which means that the skin of the ocean is most generally present. This skin layer is the molecular boundary between the turbulent atmosphere and the turbulent ocean. It is this skin layer that transfers heat, gases, and momentum between the atmosphere and the ocean. The temperature of this ultrathin layer can be measured only radiometrically, since any contact with the skin layer will disturb it. Thus, this layer cannot be measured directly by drifting or moored buoys or by any ship. Equipped with infrared radiometers, ships could measure this skin layer temperature radiometrically without the attenuating atmosphere in between the ship radiometer and the sea surface. Such measurements could provide validation and possibly calibration information for satellite-based infrared radiometric measurements of skin SST.

An important question is ‘What is the precise definition of bulk SST?’ Drifting buoys measure a temperature somewhere between 0.5 and 1.5 m below the sea surface. This is a result of the buoy’s interaction with the waves at the sea surface. As described in this article, ship cooling water intake ports may be located anywhere from 2 to 5 m beneath the waterline. So is the bulk SST located at 2–5 m beneath the surface? If we consider the temperature profile in the shallow upper layer of the ocean ([Figure 6](#)), we can see that the skin SST is always slightly colder than the

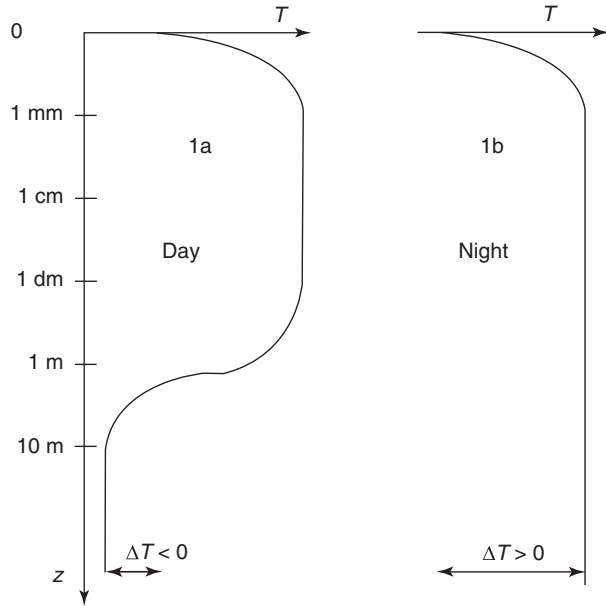


Figure 6 Temperature profile in the shallow upper layer of the ocean.

temperature just below it. Note the logarithmic abscissa in this plot. At night, this temperature slightly below the cooling skin layer is isothermal to a depth of a few meters. In this case, the bulk temperature could be measured at any of the depths to 5 m. During the day, however, solar insolation can heat a shallow layer sufficiently so that the temperature of the ‘cool skin’ is a little higher than the isothermal layer below the warm diurnal layer. Often referred to as ‘warm-skin SST,’ this condition can exist only during daytime under relatively clear-sky conditions when the shallow surface layer is heating.

The relationship between the skin and the bulk temperatures depends on two forcing factors: the wind and the net air-sea heat flux. A good example of this is shown in **Figure 7**, which is taken from two different oceanographic research cruises. Here, we have plotted the difference between the skin SST and the bulk temperature (taken as a temperature at a depth between 2 and 5 m) as a function of wind speed and net air-sea heat flux. The wind speed was observed from the research vessel, while the net air-sea heat flux was calculated from the traditional bulk formulas using routine meteorological measurements also made from the ship. At the top is a series from a cruise on the FS Meteor in the North Atlantic. Here, interestingly enough, the ΔT between the skin SST and the bulk temperature both rises and falls with increasing wind speed. At low heat flux levels, ΔT decreases as wind speed increases. This is consistent with traditional wisdom that increased mixing due to wind stirring will homogenize the upper layer, eventually making skin and bulk temperatures the same. At higher heat flux values, however, ΔT actually increases as the wind speed increases. In the lower figure, a data set was collected from the tropical Pacific. A much more limited change in ΔT is found for this tropical sample, but there is still a decrease in ΔT with increasing wind.

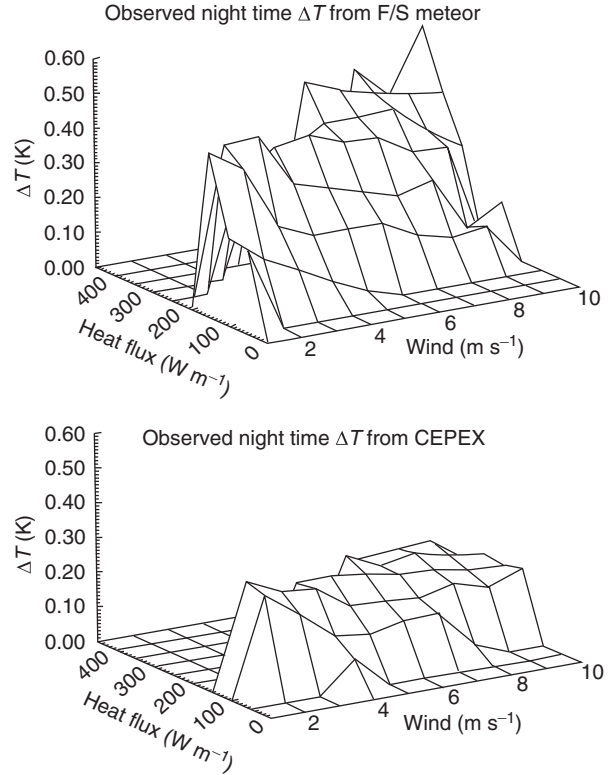


Figure 7 Skin and bulk temperatures as a function of wind speed and net air-sea heat flux.

These effects can be modeled with a semiempirical formula that includes the effects of wind and net heat flux on the ΔT difference between skin SST and bulk temperature:

$$\Delta T = \frac{Q_N}{\rho_w c_p \kappa^{1/2}} \left\{ C_{\text{shear}} \left(\frac{\nu Z_0}{u^{*3}} \right)^{1/2} + \left[C_{\text{conv}} \left(\frac{\nu \rho_w c_p}{\alpha g Q_N} \right)^{1/2} - C_{\text{shear}} \left(\frac{\nu Z_0}{u^{*3}} \right)^{1/2} \right] e^{-Rf_{\text{cr}}/Rf_0} \right\}^{1/2} \quad [1]$$

where Q_N is the net heat flux, ρ_w is the density of water, c_p is the specific heat capacity of seawater at a constant pressure, κ is the thermal diffusivity of water, C_{shear} is the proportionality constant for shear-driven time scale, C_{conv} is the proportionality constant for convective-driven time scale, ν is the kinematic viscosity, Z_0 is the momentum roughness length, u^* is the friction velocity of water, g is the acceleration due to gravity, α is the thermal expansion coefficient, Rf_0 is the surface Richardson number, and Rf_{cr} is the critical Richardson number.

This equation connects the skin SST with the temperature just below it, but does not account for the further transfer of heat downward or upward within the water column. To do this, a ‘mixed-layer’ model must be added to the skin-bulk parametrization. Such a model has been employed to show the development of the full upper-layer temperature profile. Using this model combination, it was possible to trace the temperature history of the upper layer for a couple of consecutive days of field measurements. There is excellent agreement between

the observed bulk–skin (ΔT) temperature difference and that modeled using the wind speed and net air–sea heat flux. Thus, it appears that this combination of models can go a long way to simulating the temperature behavior of the upper layer of the ocean. It should be noted that these models do not account for surface wave effects, or the effects of turbidity and foam on the sea surface. It is likely that these effects are smaller in magnitude than those driven by the wind and net heat flux. Much more sophisticated models will be required to include these phenomena in modeling ΔT .

In Situ Measurement of Skin SST

Since infrared satellite sensors are able to measure radiation only from the skin of the ocean, the challenge is to provide in situ measurements of the skin SST that can be used to calibrate and validate the satellite radiances in terms of temperature. As mentioned in this article, ships can be equipped with radiometers to measure directly the skin SST without the atmosphere attenuating the infrared signal. In principle, these same radiometers could be installed on moored buoys to continuously measure skin SST. The problem with both of these installations is that the radiometer optics must be protected from sea spray, which is difficult to do in an autonomous installation. At least on the ship, the radiometer can be examined each day and cleaned off to maintain a clear optical path.

Another requirement of these radiometers is that they are very well calibrated. The best approach to fulfilling this requirement is to equip the radiometers with two blackbodies, one at ambient temperature and the other is heated about 10 degrees above ambient. To correct for reflected infrared sky radiation, these radiometers must look up and down to view the sky, the ocean, and both blackbodies for every few scans. The easiest way to implement these requirements is to use a rotating mirror to channel the radiation to the detector, as

seen in Figure 8. In this setup, the sensor is a low-cost thermal infrared ‘microbolometer’ with a supplemental rotating mirror to collect radiation from the sky through the top hole, the sea surface through the bottom hole, and both of the blackbodies to the sides. The microbolometer collects 2D array thermal infrared image that is then calibrated regularly by the onboard blackbodies.

It is difficult to predict just how many ships need to be operating to provide the global coverage needed to regularly calibrate and validate the satellite radiometers. It is clear that the need for continuing sampling means that the ships must be merchant ships traveling long routes on a regular basis. A plot of the present ship coverage for a year (Figure 9) reveals that there are plenty of ships to choose from in the Northern Hemisphere, but in the Southern Hemisphere the available selection is much more restrictive. Ship tracks that make the long transits to Australia and New Zealand are good candidates, as are those routes from Asia to southern Chile. Only experience will reveal just how many ships over which routes will be required to supply the in situ skin SSTs needed to calibrate and validate the satellite infrared radiances.

Summary and Conclusions

SST is and has been one of the most measured variables in the ocean, and as such it has received much scientific attention. Using a combination of ship SSTs, moored and drifting buoy SSTs, and satellite infrared SSTs, we have characterized the global and seasonal patterns of SST. Realizing that only the satellite infrared SSTs can represent the SST of the 10 μm thick skin layer of the ocean, future efforts should be aimed at separating the satellite skin SSTs from the 1–5 m deep bulk SSTs measured by the buoys and ships. The skin layer is simply the molecular layer that interfaces between a turbulent ocean and a turbulent atmosphere. The temperature difference between the skin and the bulk temperatures is directly related to the wind

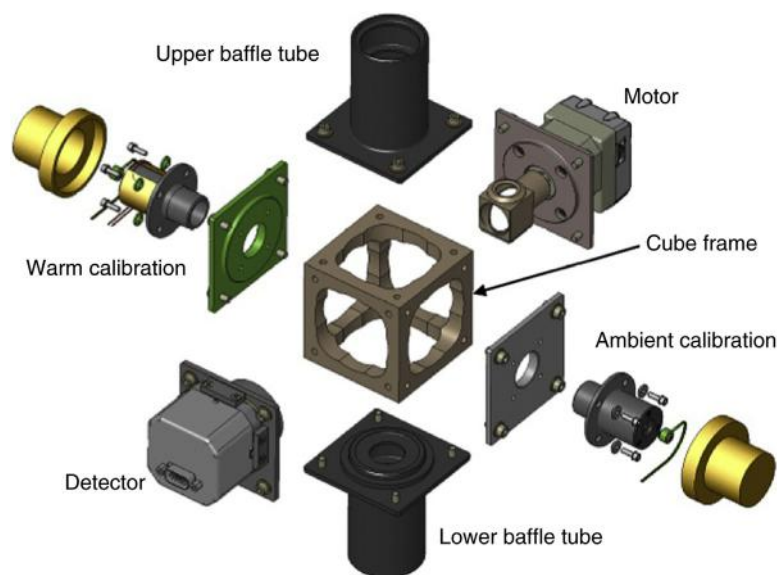


Figure 8 Modular microbolometer-based infrared radiometer (the ball experimental SST or BESST).

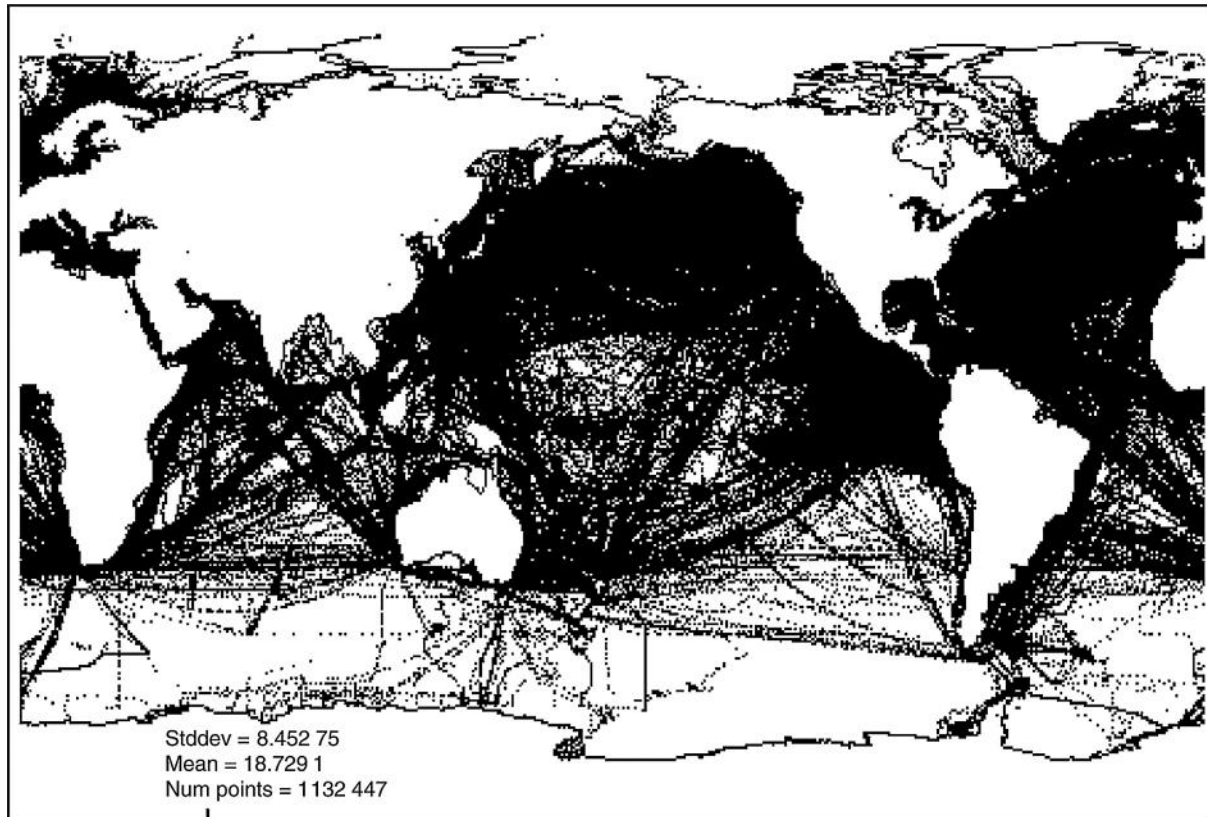


Figure 9 Global merchant ship routes for 1996.

speed and net air–sea heat flux. Thus, an improved understanding of this relationship can help us to better resolve the net heat and momentum fluxes between the ocean and the atmosphere. In part, this understanding and the shift to the computation of skin SST from satellite infrared data depend on the creation of a network of ‘ship of opportunity’ based skin SST radiometers collecting global and continuous samples of skin SST ‘ground truth’ data (i.e., without an intervening atmosphere).

See also: **Air Sea Interactions:** Freshwater Flux; Momentum, Heat, and Vapor Fluxes; Surface Waves. **Chemistry of the Atmosphere:** Observations for Chemistry (Remote Sensing): IR/FIR (Satellite, Balloon and Ground). **Global Change:** Climate Record: Surface Temperature Trends. **Observations Platforms:** Buoys. **Satellites and Satellite Remote Sensing:** Temperature Soundings.

- Emery, W.J., Yu, Y., 1997. Satellite sea surface temperature patterns and absolute values. *International Journal of Remote Sensing* 18, 323–334.
- McClain, E.P., Pichel, W.G., Walton, C.C., 1985. Comparative performance of AVHRR-based multichannel sea surface temperatures. *Journal of Geophysical Research* 90, 11587–11601.
- McClain, E.P., Pichel, W.G., Walton, C.C., Ahmad, Z., Sutton, J., 1983. Multi-channel improvements to satellite-derived global sea surface temperatures. *Advance in Space Research* 2, 43–47.
- Reynolds, R.W., Smith, T.M., 1994. Improved global sea surface temperature analysis using optimum interpolation. *Journal of Climate* 7, 929–948.
- Reynolds, R.W., Smith, T.M., 1995. A high resolution global sea surface temperature climatology. *Journal of Climate* 8, 1571–1583.
- Richardson, P.L., 1979. The Benjamin Franklin and Timothy Folger charts of the Gulf Stream. In: Sears, M., Merriman, D. (Eds.), *Oceanography: the Past*. Springer-Verlag, New York, Heidelberg, Berlin, pp. 703–717.
- Walton, C.C., Pichel, W.G., Sapper, J.F., May, D.A., 1998. The development and operational application of nonlinear algorithms for the measurement of sea surface temperatures with the NOAA polar-orbiting environmental satellites. *Journal of Geophysical Research* 103, 27999–28012.
- Wick, G.A., 1995. Evaluation of the Variability and Predictability of the Bulk–Skin Sea Surface Temperature Difference with Application to Satellite-measured Sea Surface Temperature. PhD thesis, University of Colorado, Boulder, CO.

Further Reading

- Emery, W.J., Baldwin, D.J., Schuessel, P., Reynolds, R.E., 2001. Accuracy of in situ sea surface temperatures used to calibrate infrared satellite measurements. *Journal of Geophysical Research* 105, 2387–2405.

Surface Waves

A Benilov, Acute Solutions, Highlands, NJ, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by L Ly, A Benilov, volume 1, pp 118–127, © 2003, Elsevier Ltd.

Synopsis

Wind waves relate directly to interactions at the air–water interface that affect turbulent regimes in the atmospheric surface layer and the ocean’s upper layer and finally result in exchange of momentum, energy, heat, and moisture. The turbulent boundary layers around the air–sea interface and the interface itself are the unified dynamical system. The air–sea heat–mass–energy exchange is regulated by the turbulence of boundary layers around the air–sea interface. Part of the wind energy and momentum is transferred directly from atmosphere to drift currents, while another part goes into growing surface waves and ocean turbulence.

Introduction

The total area of the world ocean surface, including inner seas, is equal to $361 \times 10^3 \text{ km}^2$ that is 70.8% of the area of Earth surface. Most motions of the ocean water (apart from the tides, currents near estuaries, and the infrequent motions of water mass produced by tectonic or volcanic activities at the ocean floor) are the results of atmospheric impacts. In fact, the world ocean surface is permanently covered by surface waves associated with local wind conditions, and this affects all air–sea interaction processes. The thermal state of the ocean and the seawater stratification significantly depend on atmospheric influences. In particular, these influences are responsible for the strong hydrostatic stability of the ocean, and heating and cooling the ocean’s upper layer. The controlling influence of oceans, particularly in forming weather, is turbulence in the atmospheric and oceanic boundary layers and microinteraction of water and air flow in vicinity of their interface. The main manifestations of the air–sea interaction in the atmosphere are the long-term variability of weather and the formation of climate.

Wind waves relate directly to interactions at the air–water interface that affect turbulent regimes in the atmospheric surface layer and the ocean’s upper layer and finally result in exchange of momentum, energy, heat, and moisture. The character of this transport is regulated by the turbulence of boundary layers around the air–sea interface. These influences are very complex and still poorly understood, although there have been numerous experimental and theoretical investigations. There are not sufficient observational data to specify completely the quantitative impact of surface waves on the characteristics of air–sea boundary layers. One of the major difficulties in air–sea interaction problems is the correct description of surface-wave effects. Here, the distinctive feature is the oscillation of the air–water interface, so that the classical theory of wall turbulence is generally inapplicable.

Surface waves are the most visible manifestation of time-dependent ocean dynamics and have been of interest for as long as man has been involved in fishing, navigation, and trade. Surface waves are important components of the air–sea system. Part of the wind energy and momentum is transferred directly from atmosphere to drift currents, while another part

goes into growing surface waves. Surface waves are the ocean’s responses to wind and pressure changes at the air–sea interface. Surface waves include capillary waves, which undulate within a fraction of a second over a distance less than a centimeter and planetary waves with periods of years and wavelengths of thousands kilometers. For air–sea interaction, surface waves are more important and we will focus on wind-generated waves at the air–sea interface. These waves determine the small-scale configuration of the air–sea interface and that affects the turbulent transfer. These waves depend on the state of the upper layers of the ocean.

Surface waves have been studied actively theoretically and experimentally. Although waves have been studied a long time, existing theories have been able to predict ocean waves with acceptable accuracy not too long ago. Complicated nonlinear interactions between wind, waves, current, and turbulence make both theoretical and observational studies more difficult. Surface waves have relatively short periods, mostly within the range of 1–30 s. The surface waves that determine sea conditions in coastal regions and break on the shore may have been generated either in open ocean and traveled coastward or in the coastal region itself by local winds. The initial wave generation process is similar in the open ocean and coastal regions, but the shallower water depth begins to affect the waves when it is less than about half a wavelength. The longer waves are affected first when a spectrum of waves travels into shallow region. Currents such as tidal currents also affect waves and these currents are often stronger in coastal regions. The reflection of waves, due to the reduction in the velocity of propagation in coastal region, also has an important influence on the waves near the shore. The classical theory of surface waves deals with waves of uniform amplitude, wavelength, and period, traveling in a fixed direction. In reality, ocean surface waves have very different heights, periods, wavelengths, and characters. Surface waves are studied by defining average values (height, period, and wavelength), the distribution of values around mean value, and a wave spectrum, representing the actual sea surface as a superposition of a large number of waves with various periods, amplitudes, and directions.

Remote sensing has evolved such that many parameters of the ocean surface may now be measured. In terms of air–sea interactions the ocean parameters include surface currents, wave directional spectra, surface wind and heat flux, and

temperature. Atmospheric quantities include profiles of winds, temperature, humidity, and chemical species. The remote sensing of surface waves and fluxes has been based, in most part, on empirical fits of infrared and active remote sensing signatures to ground truth measurements. Microwave remote sensing has provoked an active interest in the shortest waves. Short gravity waves and capillary waves have recently received attention as the progenitors of the backscatter cross section observed by microwave remote sensors. The backscatter cross section is a measurement of the amount of microwave reflection from capillary and short gravity waves with wavelengths in the range 1–6 cm. On the smallest scales, the capillary waves and short gravity waves respond quickly to the wind, and thereby provide a linkage to the remote microwave sensors. Synthetic aperture radar provides high-resolution imagery of the sea surface by measuring Bragg scattered radar pulses from the sea surface.

Equation of Motion

The approximate conservation equations of momentum and mass may be written:

$$\frac{\partial \mathbf{u}}{\partial t} + (\mathbf{u} \cdot \nabla) \mathbf{u} + f \cdot [\mathbf{n} \times \mathbf{u}] + \frac{\nabla p}{\rho} = \nu \Delta \mathbf{u} + \mathbf{g} + \mathbf{Fr}, \quad [1]$$

$$\nabla \cdot \mathbf{u} = 0. \quad [2]$$

Here, t is time, $\mathbf{u} = \{u_1, u_2, u_3\}$ is the velocity of wind in the atmospheric boundary layer or the velocity of ocean current in the oceanic boundary layer with the horizontal components u_1, u_2 and vertical u_3 , ∇ is the gradient operator, p is pressure, ρ is the density of air in the atmospheric boundary layer or the density of seawater in the oceanic boundary layer (in boundary layers around the air-sea interface the density variations $\delta\rho$, which is the air density variation $\delta\rho_a$ in the atmosphere and the seawater density variation $\delta\rho_w$ in the ocean, are very small regarding to the reference density $\rho_0 = (\rho_{0a}, \rho_{0w})$ with the air density $\rho_{0a} \approx 1.23 \text{ kg m}^{-3}$ at a pressure of 1013 hPa and a temperature of 15 °C, and the seawater density $\rho_{0w} \approx 1000 \text{ kg m}^{-3}$ at a pressure of 1013 hPa and a temperature of 15 °C, $|\delta\rho/\rho_0| \ll 1$), ν is the kinematic molecular viscosity (at a temperature $T = 20$ °C, $\nu = \nu_a = 1.5 \times 10^{-5} \text{ m}^2 \text{ s}^{-1}$ for air and $\nu = \nu_w = 1 \times 10^{-6} \text{ m}^2 \text{ s}^{-1}$ for the water), $\mathbf{g}$ is the gravitational acceleration ($|\mathbf{g}| = 9.81 \text{ m s}^{-2}$, the direction of $\mathbf{g}$ defines the local vertical), $f = 2\Omega \sin \phi$ is the Coriolis parameter (where $\Omega = 1.161 \times 10^{-5} \text{ Hz}$ is the angular velocity of rotation of the Earth and ϕ is latitude), $\mathbf{n}$ is the unit vector directed along the z -axis, ν is the kinematic viscosity of air or seawater consequently. The forcing term $\mathbf{Fr}$ with components $\text{Fr}_x, \text{Fr}_y, \text{Fr}_z$ represents any forces, other than pressure or gravity or viscous friction, acting in the body of the water. The momentum equations state that a change in the velocity following the motion of a water parcel is caused by an imbalance of the Coriolis force (rotation of the Earth), the force by the pressure gradient, the viscous forces, and the external forces. The assumption that the density element of fluid does not change (though it may differ for different elements) reduces the equation of mass conservation to the incompressibility condition [2], also known as the continuity equation, states that the divergence of the horizontal velocity must be balanced by vertical motion.

Surface Wave Dynamics

The Governing Equations

Surface wave theory assumes that the influences of Coriolis force, viscous friction and buoyancy on water motions are negligible, and that the wave motion is irrotational, i.e., there is no vorticity or the curl of the velocity vector is zero ($\text{curl} \mathbf{u} = 0$). A velocity potential φ is defined as $\mathbf{u} = \mathbf{v} = \nabla \varphi$, and from eqn [2], there follows Laplace's equation:

$$\Delta \varphi = 0. \quad [3]$$

At fixed surface, such as an impermeable seabed, the normal velocity component vanishes. The free surface is specified as $z = \zeta(x, y, t)$ at all times, then the kinematic free-surface condition follows:

$$w_\zeta = d\zeta/dt = (\partial\varphi/\partial z)_\zeta = \partial\zeta/\partial t + (\nabla_h \varphi)_\zeta \cdot (\nabla_h \zeta), \quad [4]$$

where $\nabla_h = (\partial/\partial x, \partial/\partial y)$ is the horizontal gradient operator, the suffix ζ indicates a quantity at the free-surface $\zeta(x, y, t)$. In this case, eqn [1] reduces to Bernoulli's equation, and then the pressure p in the water at the free surface is given by

$$\begin{aligned} p/\rho_w + g\zeta + (\partial\varphi/\partial t)_\zeta + 0.5(\nabla\varphi \cdot \nabla\varphi)_\zeta &= 0, \\ p &= p_{a0} - \gamma(\nabla_h \cdot \nabla_h)\zeta, \end{aligned} \quad [5]$$

where p_{a0} is the atmospheric pressure on the Earth's surface, and γ , in the term of the pressure by the water surface tension, is the ratio of surface tension to water density ($\gamma = 0.074 \text{ N m}^{-1}$ at $T = 15$ °C).

Small-Amplitude Waves

The approach of small-amplitude waves corresponds well to the real waves because the ratio of the wave amplitude a to the wavelength λ is small ($a/\lambda \ll 1$). In this case, nonlinear terms in eqns [4]–[5] will be negligible. At the surface, using kinematic boundary condition and boundary condition for pressure, solutions show that the water particles move in circular orbit of radius ae^{kz} where depth $z \leq 0$ and the wave number $k = 2\pi/\lambda$. The presence of a restoring force (gravity and surface tension) establishes a relationship between the wave frequency ω and the wave number vector $\mathbf{k}$ ($\mathbf{k} = (k_x, k_y)$) is directed along the direction of the wave propagation), which is known as a dispersion relation and, in the case of the wave propagation on the surface of a water layer of thickness H , is

$$\omega^2 = (gk + \gamma k^3/\rho_w) \tan h(kH). \quad [6]$$

The velocity of wave propagation, phase velocity, is defined by $c = \omega k/k^2$. The energy density is the energy of a wave field in a unit of volume, which is proportional to the mean square amplitude of all waves in the volume. Wave radiation is a transport or flux of energy that is caused by the waves. Surface waves in reality come in finite packets that involve a whole set of waves with different wave numbers and frequencies. The rate of energy transmission by these wave packets is the group velocity $c_g = \nabla_k \omega$. The group velocity can have various values in various parts of the spectrum. This means that different parts of a wave group propagate generally with different velocities. This process causes the energy to the group to be dispersed over large volumes or areas. The transfer of energy between different

components of a wave spectrum involves nonlinear interactions in eqns [4] and [5].

The waves with the wavelengths λ in the range $\lambda \gg \lambda_\gamma = 2\pi(\gamma/\rho_w g)^{1/2}$ ($\lambda_\gamma \approx 0.0173$ m) are not affected by the surface tension and they are the gravity surface waves with the dispersion relation:

$$\omega^2 = gk \tan h(kH), \quad [7]$$

while the waves with the wavelengths of the range $\lambda_\gamma \gg \lambda$ are the capillary waves, or ripples, with the dispersion relation:

$$\omega^2 = (\gamma k^3 / \rho_w) \tan h(kH). \quad [8]$$

If the depth H is much larger than the wavelength λ , $kH \gg 1$ and $\tan h(kH) \approx 1$, then the dispersion relation [7] reduces to $\omega^2 = gk$ (deep-water waves) and phase velocity magnitude is defined by $c = (g/k)^{1/2} \gg c_{\min} = (4\gamma g / \rho)^{1/4} \approx 0.23 \text{ m s}^{-1}$. The displacements, the particle velocities, and the oscillatory part of the pressure decrease exponentially with increasing depth, with a factor $\exp(2\pi z / \lambda)$. At a depth equal to half a wavelength, the wave motion is reduced to about 1/23 of its surface amplitude.

If the gravity waves are long and the water is relatively shallow, $kH \ll 1$ and $\tan h(kH) \approx kH$, the dispersion relation [7] reduces to $\omega^2 = gk^2 H$ (shallow-water or long waves including tidal waves) and the magnitude of phase velocity is $c = (gH)^{1/2}$. In shallow-water waves, the vertical velocity decreases linearly from the surface to bottom while the horizontal velocity and the pressure perturbation do not change with depth, and barometric pressure relation is satisfied.

Second-Order Waves

The wave steepness parameter $ak = 2\pi a / \lambda$ defines the maximum surface slope associated with a particular harmonic wave. Waves cannot persist if $ak > 1$ or $a/H > 1$ because the local, horizontal water velocity at the wave crest will be larger than the wave phase speed c . The wave must break in this case. Wave steepness can be related also to the ratio of the maximum vertical acceleration to the restoring force. This ratio can be considered as a wave Froude number Fr (ratio of inertial to gravitational acceleration), which, in the case of the deep water gravity waves, is $Fr = \max[(v_3 \partial v_3 / \partial z)_{z=0}] / g = (ak)^2$. Stokes (1847) first established the second-order nonlinear effects for periodic plane gravity wave on deep water. The actual shape of this wave profile is a curve known as a trochoid: the crests are steeper and the troughs are flatter. This feature becomes accentuated as the wave amplitude increases. The dispersion relation includes a second-order correction in the form $\omega^2 = gk(1 + a^2 k^2)$ known as amplitude (or nonlinear) dispersion. The modified phase velocity expression is $c = [(g/k)(1 + a^2 k^2)]^{1/2}$. The investigations of Stokes wave stability finally established that Stokes waves on deep water are unstable. The Stokes waves become unstable, and must break, when a wave develops a sharp crest with an interior angle of 120° . The nonlinear effects also associate with Stokes drift.

Momentum and Energy of Wave Motion

Wind waves are generated by the transmission of momentum and energy from the air to the water. Wave possesses both

potential energy, arising from the displacement of water particles vertically from their undisturbed water level, and kinetic energy due to the orbital velocity of the particles. The wave momentum per unit surface area of any particular wave is

$$M_{kw} = \rho(ca \times ak)/2. \quad [9]$$

The total wave momentum of all wave numbers is

$$M_w = \rho \iint c \Phi(\mathbf{k}) k dk_1 dk_2, \quad k = \sqrt{k_1^2 + k_2^2}, \quad [10]$$

where $\Phi(\mathbf{k}) = \Phi(k, \alpha)$ is the space with polar coordinates (k, α) . The wave energy of any particular constituent wave per unit surface is

$$E_{kw} = c \cdot M_{kw} = \rho(c^2 a \times ak)/2. \quad [11]$$

The specific energy of the complete wave field is obtained by integration of over all wave numbers or frequencies:

$$E_w = \rho g \iint \Phi(k, \alpha) k dk d\alpha = \rho g \int \Psi(\omega) d\omega, \quad [12]$$

and the wave elevation variance is

$$\sigma_\zeta^2 = \overline{\zeta(x, y, t)^2} = \iint \Phi(k, \alpha) k dk d\alpha = \int \Psi(\omega) d\omega, \quad [13]$$

where $\Phi(k)$ and $\Psi(\omega)$ are the spectra in 2D- k -space and ω -space, respectively.

Wave energy budget for a particular wave number $\mathbf{k}$ in the presence of a mean current with velocity $\mathbf{U}$ and energy source/sink S_k is written as

$$\partial E(\mathbf{k}) / \partial t + ((c_g + \mathbf{U}) \cdot \nabla) E(\mathbf{k}) = S_k, \quad E(\mathbf{k}) = \rho g \Phi(\mathbf{k}). \quad [14]$$

The left hand side describes energy propagation through a nonhomogeneous medium (currents, variable depth). The right hand side consists of source terms describing energy input by wind (S_{in}), sinks (whitecapping dissipation S_{ds} and bottom terms), and a nonlinear interaction term S_{sn} . S_{sn} is a very intensive redistributor of energy, momentum, and wave action along the spectrum. Due to S_{sn} , the direct cascades of energy and momentum as well as inverse cascades of wave action are formed. These processes govern the evolution of the wind-wave spectral peak, characterized by the spectral peak phase speed $c = c_0 = [(g/k_0)]^{1/2} = g/\omega_0$, and play a central role in the formation of the universal spectrum behind the spectral peak. Most popular form of these spectra are Phillips spectrum $\Psi(\omega) = \beta_p g^2 \omega^{-5}$, $\beta_p \approx 0.01$; Pierson-Moskovitz's spectrum $\Psi(\omega) = 1.25 \beta_p g^2 \omega^{-5} \exp(-1.25(\omega_0/\omega)^4)$; JONSWAP wave spectrum; and the Kolmogorov-Zakharov wave spectra $\Psi(\omega) \sim \omega^{-4}$, an exact solution of the stationary kinetic equation.

Wave Prediction

The principles of wave prediction were well known at least by 1960s. None of the wave models developed in the 1960s and 1970s computed the wave spectrum from the full energy balance equation. The first models in the 1960s and 1970s assumed that the wave components suddenly stopped growing as soon as they reached a universal saturation level. Here, the saturation spectrum was represented by Phillip's one

dimensional ' ω^{-5} ' frequency spectrum and an empirical equilibrium directional distribution. Today it is generally recognized that a universal high-frequency spectrum (region between 1.5 and 3 times the peak frequency) does not exist because the high-frequency region of the spectrum not only depends on whitecapping but also on winds and on the low-frequency regions of the spectrum through nonlinear transfer. Therefore, the first-generation wave models exhibit basic shortcomings by overestimating the wind input and disregarding nonlinear transfer.

The second-generation wave models attempted to simulate properly the overshoot phenomenon and the dependence of the high-frequency region of the spectrum on the low frequencies. However, the nonlinear transfer parameterization requires the spectral shape of the wind sea spectrum to be prescribed. These models suffered basic problems on the treatment of wind sea and swell.

The wave spectrum was computed by the integration of the energy balance equation [8,13] without any prior restriction on the spectral shape in the third-generation wave models. At present the state of the art in wave models are WAM (a third-generation wave model) and WAVEWATCH, which were developed by international groups and are based on a detailed physical description of air-sea interaction. A prediction of significant wave height requires a prediction of the wave spectrum. Based on specified surface-wind stress (and current) fields at all times, the wave energy balance is solved to obtain the spectra (and the significant wave heights) in wave predictions. However, the fact that today two of the most popular contemporary models, WAM and WAVEWATCH, operational at two of the most prominent meteorological centers, use different approaches to the problem is an indication that a single 'best' solution has not yet been accepted.

Rogue Wave

Freak, rogue, or giant waves correspond to large-amplitude waves surprisingly appearing on the sea surface ('wave from

nowhere') have been a part of marine folklore for centuries. According to the definition of freak waves they are too high, too asymmetric, too steep, and their height exceed the significant wave height by 2–2.2 times. The data of marine observations as well as laboratory experiments demonstrate that freak waves may appear in deep and shallow waters. The probabilistic theoretical approach shows that numerical simulations of freak waves should be made for very long times on large spatial domains and large number of realizations. Theoretical studies show the main features of the physical mechanisms of rogue wave phenomenon. The prediction of rogue wave is based on data analysis with use of statistical methods, but it is not too productive due to the rare character of the rogue waves.

Marine and Oceanic Surface Layers

An important feature of the atmospheric boundary layer over the ocean is the expenditure of momentum and energy of the wind on the generation of waves and currents. This constitutes a fundamental difference of the atmospheric surface layer not only from the ordinary layer above smooth or rough solid surfaces. The thickness of the marine surface layer, **Figure 1**, is of a few tens of meters that heights z above the mean water surface are in the range $H_{E,atm} \gg L_{atm} > z > L_{w-t,atm}$. A distinctive property of this layer is that within it, the vertical fluxes of momentum, heat, moisture, and various types of passive impurities vary little with height and at the horizontal scales much greater than its thickness. These peculiarities allow reducing eqns [1] and [2] to the form of horizontally uniform turbulent boundary layer ('wall turbulence') where these fluxes do not depend on spatial coordinates. The turbulent boundary layer of the atmosphere is stratified; turbulent processes within it are influenced by buoyancy, which in turn results from changes in air density owing to fluctuations in temperature and humidity. The turbulence of the main interval of the atmospheric surface layer at a distance above sea level greater than a few times the height of the largest waves is well described by the Monin-Obukhov (MO) theory. The MO theory establishes

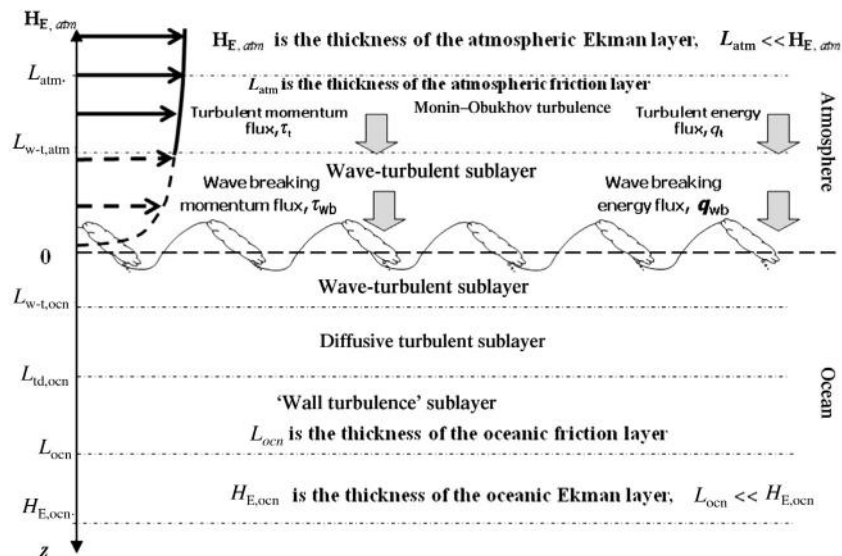


Figure 1 Diagram of dynamic air-sea interaction and dynamical layering of the atmospheric and oceanic boundary layers.

that the turbulent flow is governed by mechanical and thermal forcing, and the statistical characteristics of turbulence are determined by the height above the surface, z , the buoyancy parameter, $\beta = g/S^0$, the friction velocity, u_{*r} , and the vertical turbulent heat flux $Q_T = \rho_{a0} c_{pa} \overline{T' u'_3}$ (T' , u'_3 are turbulent fluctuations of air/seawater temperature and vertical velocity; $u_* = u_{*a}$ in the atmosphere or $u_* = u_{*w}$ in the ocean; S^0 is either air reference potential temperature θ_{a0} or seawater reference density ρ_{w0} ; c_p is the specific heat of air at constant pressure $c_{pa} = 1004 \text{ J K}^{-1} \text{ kg}^{-1}$ or seawater $c_{pw} = 4218 \text{ J K}^{-1} \text{ kg}^{-1}$). Conclusions derived from the MO theory are the principal tool for calculating the characteristics of the turbulent atmospheric surface layer. The atmospheric MO-length scale,

$$L_{MO} = -\rho_{a0} c_{pa} u_{*a}^3 / \kappa \beta Q_T, \quad [15]$$

characterizes buoyancy influence in the MO-turbulent boundary layer (Figure 1), $\kappa = 0.4$ is the Karman constant. The sign of the MO-similarity variable $\zeta = z/L_{MO}$ indicates that the stratification of the turbulent boundary layer is stable if $\zeta > 0$, or unstable if $\zeta < 0$ (the z -axis is positive in upward direction in the atmosphere, Figure 1). The buoyancy influence on turbulence is most significant when $|\zeta| \gg 1$ (thermal convection if $\zeta < 0$, or strong stable stratification if $\zeta > 0$; in thermal convection case, the friction velocity does not contribute into turbulent statistical characteristics except the vertical profile of mean wind). The buoyancy does not affect the turbulence if $|\zeta| \ll 1$, and then the turbulence is similar to the 'wall turbulence' in a fluid with constant density (neutral stratification) or the 'logarithmic' turbulent boundary layer. Thus the mean vertical profiles of wind speed, air temperature, and humidity well follow to their theoretical logarithmic form which depends on an additional parameter z_0 , the roughness parameter of underlying surface or the roughness length scale of underlying surface, and, in our case, is $z_0 \ll L_{w-t,atm}$. The formal definition z_0 states that z_0 is the height above the mean underlying surface at which the mean wind velocity, defined by the 'logarithmic' law, is zero. A characteristic of an actual immobile surface is the mean height h_0 of the actual protrusions on it, and z_0 is in a proportion to h_0 . The underlying surface is dynamically smooth if h_0 is small as $h_0 < 36z_v$, or it is dynamically completely rough if $h_0 > 540z_v$ ($z_v = \nu_a/9 u_{*a}$ is the viscous friction length scale of the dynamically smooth surface, $z_0 = z_v$), and, hence, z_0 of any immobile surface characterizes h_0 if $z_0 \geq z_v$. The sea surface roughness z_0 , in difference to immovable surfaces, varies in a very wide range ($10^{-7} \text{ m} < z_0 < 0.1 \text{ m}$), and in some situations is even less than z_v ($z_0 < z_v$). Field observations and theoretical studies established that z_0 predominantly depends on the surface wave age $\varsigma = c_0/u_{*a} = [(g/u_{*a}^2 k_0)]^{1/2} = g/u_{*a} \omega_0$. The field measurements at height 10 m presented in the form $\ln(z_0/z_v)$ versus the wave age $\varsigma = c_0/u_{*a}$ can be approximated as

$$\begin{aligned} 11.5 \geq \ln(z_0/z_v) &= 14 - 0.516\varsigma + 2.74 \times 10^{-3}\varsigma^2 \\ &\geq -10.25, \quad 5 \leq \varsigma \leq 90, \end{aligned} \quad [16]$$

where $z_0 = z_v$ at $\varsigma = \varsigma_* = 32.87$. This empirical formula allows dividing the range of ς onto two main regions: the region of developing waves $5 \leq \varsigma \leq \varsigma_*$, which corresponds to the rough sea surface $z_0 \geq z_v$, and the region $\varsigma_* \leq \varsigma \leq 90$, which corresponds to the 'oversmoothed' sea surface $z_0 \leq z_v$ when the wind

is weak and a swell propagates in the same direction or in a situation of slackening wind but with already developed waves. The wind waves are the developed waves when $\varsigma \approx \varsigma_*$. This classification is common in physical oceanography. In the case of developing waves ($5 \leq \varsigma \leq \varsigma_*$), the wind transfers its momentum and energy to waves, the sea surface is rougher for young waves (small values ς) and becomes, in aerodynamic sense, 'smooth' ($z_0 \approx z_v$) for mature, or developed, waves when $\varsigma \rightarrow \varsigma_*$. In the case of the 'oversmoothed' sea surface ($z_0 \leq z_v$), the waves momentum and energy transfer to the wind. Hence, in difference to the roughness of hard immovable surfaces, the sea surface roughness has the dynamical nature.

Charnock's formula (Elison, 1956; Charnock, 1955) $z_0 = m u_{*a}^2/g$, where m is the Charnock's constant, is a well-known expression for calculating z_0 of sea surface. Numerous estimates of the numerical value of m from field measurements and laboratory experiments point out that m varies in the broad range as $10^{-3} < m < 1$ (data by Donelan et al., 1993). It is not a universal constant but a function of the wave age ς in the case of developing waves ($5 \leq \varsigma \leq \varsigma_*$). The function $m \approx 355\varsigma^{-3.64}$ roughly approximates the data of this range of ς .

In the layer of air between the logarithmic sublayer and the water surface – the airwave-turbulent sublayer in Figure 1, the theory of similarity cannot be applied because the disturbances, induced by waves, cause strong influence on the dynamics of this layer. The effect of this factor shows up in all characteristics of turbulence and is still poorly understood. The effective mechanism is redistribution of the vertically invariant flux of momentum between the turbulent and wave components of momentum. In this sublayer, about 10 m thick, waves make a significant contribution to the mean and fluctuating fields. The mean velocity profile deviates from logarithmic law, these deviations are positive or negative depending on the wave age with maximum of measured values of about $\pm 0.5 \text{ m s}^{-1}$ at the heights (accessible for measurements) closest to the sea surface (+ for developing waves ($\varsigma - \varsigma_* < 0$), – for waves with $\varsigma - \varsigma_* > 0$). The waves increase the intensity of the fluctuations and change the nature of the correlation between the fluctuation characteristics. The peaks produced by waves are clearly distinguishable in the fluctuation spectra. The waves also make a considerable contribution to the total fluxes of momentum, heat, and moisture. The mathematical procedure for filtering stationary random processes can be used to reconstruct the contributions of waves and turbulence.

Air–Sea Fluxes and Surface Waves

Momentum, Energy Fluxes, and Breaking Waves

As it was noted above, the surface atmospheric boundary layer is the boundary layer with the constant vertical fluxes of momentum, energy, heat, and humidity over its entire thickness including the wave-turbulence sublayer (Figure 1). Therefore, the vertical fluxes in the logarithmic sublayer can be accepted as the actual fluxes and all mean characteristics of the boundary layer should be taken at the height $z_a = 10 \text{ m}$ as the standard reference height where the wave disturbances are negligibly small.

The turbulent fluxes of momentum ($\tau_{1,3} = -\rho_{a0} \overline{u'_1 u'_3} = \rho_{a0} u_*^2$), sensible heat ($Q_T = \rho_{a0} c_{pa} \overline{T' u'_3}$), and humidity

($Q_E = \rho_{a0} \overline{q' u_3'}$, q' is the specific humidity fluctuation) have traditionally been estimated by using bulk form:

$$\begin{aligned} \tau_{1,3}/\rho_{a0} &= u_*^2 = C_u U_{10}^2, Q_T/\rho_{a0} c_{pa} \\ &= -C_T U_{10} (T_{a,10} - T_{ws}), p_{a0} Q_E/\rho_{a0} \\ &= -C_E U_{10} (E_{10} - E_{ws}), \end{aligned} \quad [17]$$

where C_u is the drag coefficient of sea surface, C_T is the heat transfer coefficient, and C_E is the evaporation coefficient; U_{10} , $T_{a,10}$, E_{10} are the mean values of wind speed, air temperature in °C and humidity at the height $z_a = 10$ m; T_{ws} is the water surface temperature in °C; here the specific humidity $q \approx 0.622 \times E/p_{a0}$, where E is the partial pressure of water vapor and the atmospheric pressure on the Earth's surface p_{a0} in hPa, $p_{a0} = 1013$ hPa; E_{ws} is the humidity of saturated vapor at water surface temperature T_{ws} and can be found with the help of the known expression:

$$E_{ws} = 6.1 \exp(7.45 T_{ws} (235 + T_{ws})^{-1}). \quad [18]$$

The latent heat fluxes Q_E can be found using the formula:

$$Q_E = -\rho_a A_e p_{a0}^{-1} \times 0.622 \times C_E \times U_{10} (E_{10} - E_{ws}), \quad [19]$$

where $A_e = 2.45 \times 10^6$ J kg⁻¹ is the specific latent heat of evaporation.

The variability of the coefficients C_u , C_T , C_E constitutes one and half orders of magnitude; measurement data place their values in the range $0.5 < (C_u, C_T, C_E) \times 10^3 < 10$. In the case of developing waves ($\zeta - \zeta_* < 0$) and wind conditions $U_{10} > 3$ m s⁻¹, the drag coefficient C_u is a decreasing function of the wave age ζ . The empirical data and theoretical calculations – mutual wind-wave adaptation model – well follow to the approximation $C_u(\zeta) \approx 0.5[1 + 1.3 \times \tan h(\zeta/\zeta_*)] \zeta^{-2}$ for the wave age range of $10 < \zeta \leq \zeta_*$. Examples of C_u variations versus the wind speed U_{10} – one of traditional forms of description the drag coefficient variability – are shown in Figure 2 where numerical modeling results are shown along with observation data.

The variabilities of the heat transfer coefficient C_T and the evaporation coefficient C_E are very similar to the variability of

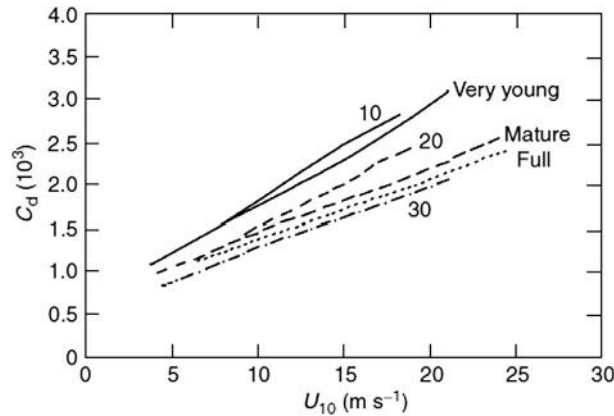


Figure 2 Dependence of model drag coefficients at $z = 10$ m and $\zeta = 10, 20, 30$. Three other curves with 'very young,' 'mature,' and 'full' from observed data by Ly and Garwood (2000).

the drag coefficient C_u . The dependence C_T versus C_u is given by the empirical expression

$$C_T(\zeta) \approx A_1 [1 - 17 \sqrt{C_u(\zeta)}] C_u(\zeta) \quad [20]$$

with the constant $A_1 \sim 1$. This formula agrees well with the conclusions from the theoretical analysis of the transformation of thermal regime of the surface atmospheric layer in the presence of wind-wave interaction. It is recommended in a literature that the averaged values of $C_T \approx C_E$.

In the case of developing waves, the characteristics of the wave field c_0 and σ_ζ can be found as functions of the wave age ζ using the approximation for the drag coefficient C_u and the Phillips spectrum ' ω^{-5} .' These estimates are

$$\begin{aligned} \frac{c_0}{U_{10}} &\approx \sqrt{0.5 \times [1 + 1.3 \times \tan h(\zeta/\zeta_*)]}, \quad \frac{g\sigma_\zeta}{U_{10}^2} \approx 0.5 \beta_p^{1/2} \zeta^2 C_u \\ &\approx \frac{\beta_p^{1/2}}{4} [1 + 1.3 \times \tan h(\zeta/\zeta_*)]. \end{aligned} \quad [21]$$

$$10 < \zeta \leq \zeta_*$$

They fit to the observations and the theory of mutual wind-wave adaptation. If the wave field can be considered as the developed wave field, i.e., $\zeta \approx \zeta_*$, then for $\zeta_* = 32.87$,

$$C_u(\zeta_*) \approx 0.92 \times 10^{-3}, \quad \frac{c_0(\zeta_*)}{U_{10}} \approx 1, \quad \frac{g\sigma_\zeta}{U_{10}^2} \approx 5 \times 10^{-2}. \quad [22]$$

In the case of very weak wind condition or nearly calm weather, when $U_{10} < 3$ m s⁻¹, the estimates of sensible and latent heat fluxes by standard meteorological measurements can be found by formulas:

$$Q_T = -\frac{0.14 \rho_a c_{pa}}{\text{Pr}_T^{1/3}} \Delta T \left[\left(\frac{\Delta T}{T} + 0.61 \times \left(0.622 \times \frac{\Delta E}{p_{a0}} \right) \right) g D_T \right]^{1/3}, \quad [23]$$

$$\begin{aligned} Q_E &= -\frac{0.14 \rho_a A_e}{\text{Pr}_E^{1/3}} \times \left(0.622 \frac{\Delta E}{p_{a0}} \right) \\ &\times \left[\left(\frac{\Delta T}{T} + 0.61 \times \left(0.622 \times \frac{\Delta E}{p_{a0}} \right) \right) g D_E \right]^{1/3}, \end{aligned} \quad [24]$$

where $T = 273 + T_{a,10}$ is the reference temperature of air at measurement height $z_a = 10$ m where $T_{a,10}$ in °C, $\text{Pr}_T = \nu_a/D_T = 0.72$ and $\text{Pr}_E = \nu_a/D_E = 0.6$ are the Prandtl numbers, D_T is the thermal molecular diffusivity, D_E is the molecular diffusivity for water vapor, $\Delta T = T_{a,10} - T_{ws}$, $\Delta E = E_{10} - E_{ws}$.

Spray

The presence of spray in the atmospheric surface layer influences dynamic and thermodynamic processes in the surface layer. Wind gives rise to sea spray through various mechanisms, but bubble bursting is the primary one. Increasing wind speed causes a corresponding increase in the production of white caps at the ocean surface. These white caps form bubbles in the ocean which, when breaking at the surface, produce sea spray. The sea spray can be observed at heights of more than 10 m above the ocean surface in the atmospheric surface layer. The wave motion and turbulent mixing act as lifting forces on the

sea spray. Because of the difference in the motions of small and large droplets, the particle size distribution changes with height, with relatively more small droplets in the upper layers. Some theoretical calculations indicate an increase in the drag of the surface and the rate of evaporation. The water drops are also a cause of scattering and absorption of the radiation.

Wave Breaking and the Ocean Upper Layer Turbulence

The velocity field $u(x, t) = (v(x, t) + w(x, t))$, of the upper ocean constitutes of two components: the turbulent component $w(x, t)$, ($\omega(x, t) = \text{curl}u(x, t) = \text{curl}w(x, t)$) and the wave component $v(x, t)$, ($v(x, t) = \nabla\varphi(x, t)$). The mean currents, or the mean velocity $u(x, t)$, is mostly associated with the turbulent component $w(x, t)$, i.e., $\overline{u(x, t)} = \overline{w(x, t)}$ is the wind-driven current with the surface drift speed U_s . The potential wave field, in a sense of accumulated kinetic energy, is a leading member of the energy contributors in the upper part of the Ekman layer (Figure 1, the depth $z > 0$ is downward). In terms of characteristic scales of velocities, they are allocated in the order $|v(x, t)| > U_s \geq |\overline{u(x, t)}| > |\overline{w'(x, t)}|$. The speed of surface drift has simple empirical estimate as $U_s \approx k_d U_{10}$ with the surface drift coefficient $k_d \approx 1/30$. The estimate of the kinetic energy of the wave motions $k_v = \overline{v \cdot v}/2$ over the entire thickness of the constant friction layer, assuming the horizontal uniformity of the wave field and using the linear theory of waves, is

$$k_v = 0.5\beta_P c_0^2 \left(1 + \sqrt{z/L_{*v}}\right) \exp\left(-\sqrt{z/L_{*v}}\right), \quad L_{*v} = c_0^2/24g \\ = \sigma_\zeta/12\sqrt{\beta_P}. \quad [25]$$

In the case of developing waves when $10 \leq \zeta \leq \zeta_*$, the wave kinetic energy in eqn [25] yield to a function of the depth z and the wave age ζ in the form:

$$k_v(z, \zeta) = U_{10}^2 \frac{\beta_P}{4} \left[1 + 1.3 \times \tan h\left(\frac{\zeta}{\zeta_*}\right)\right] \times \left[1 + \sqrt{\frac{12z\sqrt{\beta_P}}{\sigma_\zeta(\zeta)}}\right] \\ \times \exp\left[-\sqrt{\frac{12z\sqrt{\beta_P}}{\sigma_\zeta(\zeta)}}\right]. \quad [26]$$

The wave kinetic energy increases with $\zeta \rightarrow \zeta_*$; however, $k_v(z, \zeta)$ at the wave age of the range $10 \leq \zeta \leq \zeta_*$ always satisfies to the inequality:

$$k_v(z, \zeta) \leq k_v(z, \zeta_*) \\ = U_{10}^2 \beta_P \times \left(0.5 + \sqrt{\frac{6gz}{U_{10}^2}}\right) \times \exp\left(-2\sqrt{\frac{6gz}{U_{10}^2}}\right). \quad [27]$$

The turbulent energy production by the wave breaking exceeds significantly the mean shear effect in the vicinity of the ocean surface. The general dynamic structure of the turbulent upper layer is produced by wave breaking and turbulent diffusion of the wave kinetic energy. According to the Longuet-Higgins' parameterization, the waves lose simultaneously some part of the obtained momentum and energy because the wave

breaking produces the momentum and energy fluxes τ_{wb} and q_{wb} . These fluxes, or the rate of the wave momentum M_w and energy E_w decay per the characteristic wave cycle ω_0 , can be expressed for the wave age $\zeta \leq \zeta_*$ as

$$\tau_{wb} = \gamma_1 \omega_0 M_w = \frac{\gamma_1 \beta}{3} \rho_w c_0^2, \quad q_{wb} = \gamma_2 \omega_0 E_w = \frac{\gamma_1 \beta}{3} \rho_w c_0^3, \quad [28]$$

where $\gamma_1(\zeta) \leq \gamma_1(\zeta_*) \approx 3 \times 10^{-4}$, $\gamma_2 = (4/3) \times \gamma_1$ are the dimensionless constants representing the fractions of the wave momentum and energy spent in the wave breaking.

The model equations (k - ε group model) take into account the presence of potential surface waves. In this model, $k_T = \overline{w'w'}/2$ is the turbulent kinetic energy, $\nu_T = c_v \varepsilon^{-1} k_T^2 = \ell k_T^{1/2}$ is the traditional form of the turbulent viscosity, where ε is the dissipation rate, ℓ is the length scale of turbulent mixing, and $c_v = 0.09$ is the Kolmogorov's closure constant. The turbulent kinetic energy budget contains an extraterm, $0.5 \overline{w'_3 w'}$, which is the turbulent flux of the wave kinetic energy.

Wave-Turbulent Sublayer

The wave-turbulent sublayer (Figure 1) is a layer where the turbulent flux of the wave kinetic energy, $0.5 \times \overline{w'_3 w'}$, causes dominative influence in the dynamic of this layer. The given definition can be presented in the formal form:

$$|\partial_z k_T| \ll \sigma |\partial_z k_v|, \quad \nu_T (\partial_z \overline{U})^2 \ll \varepsilon, \quad 0 \leq z \leq L_{w-v}, \quad [29]$$

where $\sigma \approx 0.03$ is the closure constant in the diffusive approximation of the turbulent flux of the wave kinetic energy, and $0.5 \times \overline{w'_3 w'}$, $\overline{U}$ is the velocity of the mean current, $L_{w-v} \approx 5.1 \times L_{*v}$ is the thickness of the wave-turbulent sublayer. The principal statistical characteristics of turbulence in the wave-turbulence sublayer are given by the following equations:

$$k_T = k_T(0) \frac{3 \left(1 + \sqrt{z/L_{*v}}\right) \exp\left(-\sqrt{z/L_{*v}}\right)}{2 + \left(1 + \sqrt{z/L_{*v}}\right)^3 \exp\left(-3\sqrt{z/L_{*v}}\right)}, \quad [30]$$

$$k_T(0) \approx 1.3(\gamma_1 \beta_P)^{2/3} c_0^2 = 2.7 \times 10^{-4} c_0^2,$$

$$\varepsilon = \varepsilon(0) \frac{3^{3/2} \left(1 + \sqrt{z/L_{*v}}\right)^{1/2} \exp\left(-\frac{3}{2}\sqrt{z/L_{*v}}\right)}{\left[2 + \left(1 + \sqrt{z/L_{*v}}\right)^3 \exp\left(-3\sqrt{z/L_{*v}}\right)\right]^{3/2}},$$

$$\varepsilon(0) \approx 4\gamma_1 \beta g c_0 = 1.2 \times 10^{-5} g c_0,$$

$$\ell = \ell(0) \left(1 + \sqrt{z/L_{*v}}\right), \quad \ell(0) = \sigma_\zeta/15\sqrt{\beta_P} = c_0^2/30g, \quad [32]$$

$$\nu_T = \nu_T(0) \frac{3^{1/2} \left(1 + \sqrt{z/L_{*v}}\right)^{3/2} \exp\left(-\frac{1}{2}\sqrt{z/L_{*v}}\right)}{\left[2 + \left(1 + \sqrt{z/L_{*v}}\right)^3 \exp\left(-3\sqrt{z/L_{*v}}\right)\right]^{1/2}},$$

$$\nu_T(0) = 0.038(\gamma_1 \beta_P)^{1/3} \frac{c_0^3}{g}. \quad [33]$$

The dissipation rate in the form of eqn [31] is well supported by the field measurements.

Diffusive Turbulent Sublayer

The diffusive turbulent sublayer is a transition zone between the wave-turbulent subsurface layer and the layer where the mean shear flow controls the turbulent regime. The depths of its location are in the range $L_{W-v} \leq z \leq L_W$, where L_W is the lower boundary of the transitional diffusive sublayer. In this region the turbulent diffusion still exceeds the mean shear contribution in the turbulent kinetic energy budget but the wave motion effect becomes insignificant, or in a formal form as

$$|\partial_z k_T| \ll \sigma |\partial_z k_V|, \quad \nu_T (\partial_z \bar{U})^2 \ll \varepsilon, \quad L_{W-v} \leq z \leq L_W, \quad [34]$$

where L_W is the lower boundary of the transitional diffusive sublayer. Taking the coordinate z as $z = L_{W-v} + z_1$, the principal turbulent characteristics of the diffusive turbulent sublayer can be found as a solution of the equations in the $k - \varepsilon$ -turbulence form:

$$\begin{aligned} 0 \leq z_1 \leq L_W - L_{W-v}, \quad k_{T1} &= \frac{k_{T1}(0)}{(1 + z_1/L_{*1})^{n_k}}, \quad \varepsilon_1 \\ &= \frac{\varepsilon_1(0)}{(1 + z_1/L_{*1})^{n_\varepsilon}}, \quad \nu_{T1} = \frac{\nu_{T1}(0)}{(1 + z_1/L_{*1})^{n_{VT}}}, \quad \ell \\ &= \ell(0)(1 + z_1/L_{*1}), \end{aligned} \quad [35]$$

$$n_k = \frac{4}{\mu - 7}, \quad n_\varepsilon = \frac{\mu - 1}{\mu - 7}, \quad n_{VT} = \frac{9 - \mu}{\mu - 7}, \quad [36]$$

$$7 < \mu = \sqrt{1 + 24c_2 \text{Pr}_\varepsilon / \text{Pr}_k} \leq 9, \quad 2 < c_2 \text{Pr}_\varepsilon / \text{Pr}_k \leq \frac{10}{3}, \quad [37]$$

$$\begin{aligned} k_{T1}(0) &= [q_{W-v}(L_{W-v})\varepsilon(L_{W-v})L_{*1}\text{Pr}_k(\mu - 7)/4c_v]^{1/3}, \quad \varepsilon_1(0) \\ &= \varepsilon(L_{W-v}), \end{aligned} \quad [38]$$

$$\begin{aligned} \nu_{T1}(0) &= \frac{c_v k_{T1}(0)^2}{\varepsilon(L_{W-v})\text{Pr}_k}, \quad \ell(0) = \frac{c_v k_{T1}(0)^{3/2}}{\varepsilon(L_{W-v})}, \quad L_{*1} \\ &= \frac{6q_{W-v}(L_{W-v})}{(\mu - 7)\varepsilon(L_{W-v})}, \end{aligned} \quad [39]$$

where c_2 , Pr_ε , Pr_k are the closure constants in the $k - \varepsilon$ -turbulence with the recommended numerical values $c_2 \approx 1.92$, $\text{Pr}_\varepsilon \approx 1.3$, $\text{Pr}_k \approx 1$. Pr_k and Pr_ε are the turbulent Prandtl numbers for the turbulent kinetic energy and the dissipation rate consequently. However, observation data support for the dissipation rate decay the law ' $\varepsilon \sim z^{-4}$ '. Then $n_k = 2$, $n_\varepsilon = 4$, $n_{VT} = 0$, and the turbulent viscosity $\nu_{T1} = \nu_{T1}(0) = \text{const}$ in the turbulent diffusive sublayer. Also, because the values $c_2 = 1.92$ and $\text{Pr}_k = 1$ are most popular their numerical estimates, the Prandtl number for the dissipation rate is $\text{Pr}_\varepsilon = 10\text{Pr}_k/3c_2 \approx 1.736$ instead of the recommended $\text{Pr}_\varepsilon \approx 1.3$. The 'wall turbulence' sublayer provides an additional equation to find Pr_k and Pr_ε , which are consistent with the turbulent diffusive sublayer law ' $\varepsilon \sim z^{-4}$ ' and the 'wall turbulence' sublayer.

'Wall Turbulence' Sublayer

Following after the turbulent diffusive sublayer, there is a layer, located at the depths z in the range $L_W \leq z \leq L$ (Figure 1, L is the lower border of the constant friction layer), where the mean

shear production of turbulent energy dominates. The formal conditions for the layer existence in the steady case take a form:

$$|\text{Pr}_k^{-1} d_{z_2} \nu_{T2} d_{z_2} (k_{T2} + \sigma k_V)| < \nu_{T2} (d_{z_2} \bar{U})^2, \quad 0 \leq z_2 \leq L - L_W \quad [40]$$

where the coordinate z is $z = L_W + z_2$. The principal turbulent characteristics of the 'wall turbulence' sublayer can be found at the depth $0 \leq z_2 \leq L_W - L_{W-v}$ as a solution of the equations in the $k - \varepsilon$ -turbulence form:

$$\begin{aligned} \bar{U}(L_W) - \bar{U}(z_2) &= \frac{u_{*W}}{\kappa} \ln \left(\frac{z_2 + L_{*2}}{L_{*2}} \right), \quad k_{T2} = \frac{u_{*W}^2}{\sqrt{c_v}} \\ &= \text{const}, \quad \varepsilon = \frac{\varepsilon_1(L_W)L_{*2}}{z_2 + L_{*2}}, \end{aligned} \quad [41]$$

$$\nu_{T2} = \kappa(L_{*2} + z_2)u_{*W}, \quad \ell = \kappa \sqrt{c_v}(L_{*2} + z_2), \quad [42]$$

$$L < L_{*2} [\exp(\kappa \bar{U}(L_W)/u_{*W}) - 1], \quad L_{*2} = u_{*W}^3 / \kappa \varepsilon(L_W), \quad [43]$$

$$\kappa = \sqrt{\text{Pr}_\varepsilon \sqrt{c_v}(c_2 - c_1)}. \quad [44]$$

The numerical values for constants c_v , c_1 , c_2 , and Pr_ε , given in textbooks as the most popular estimates ($c_v = 0.09$, $c_1 = 1.44$, $c_2 = 1.92$, $\text{Pr}_k = 1$, $\text{Pr}_\varepsilon = 1.3$), give the Karman constant $\kappa \approx 0.433$ instead of the classical numerical value $\kappa \approx 0.4$. If the closure constants $c_v = 0.09$, $c_1 = 1.44$, $c_2 = 1.92$, and $\kappa = 0.4$ represent their better established numerical values then the law ' z^{-4} ' of the dissipation decay in the diffusive sublayer with eqn [44] of the 'wall turbulence' sublayer define the turbulent Prandtl numbers Pr_k and Pr_ε as

$$\text{Pr}_k = \frac{3c_2 \kappa^2}{10 \sqrt{c_v}(c_2 - c_1)} \approx 0.64, \quad \text{Pr}_\varepsilon = \frac{\kappa^2}{\sqrt{c_v}(c_2 - c_1)} \approx \frac{10}{9}, \quad [45]$$

which are consistent to both sublayers.

Bubbles

Breaking waves at the ocean's surface inject bubbles into the water column. Gas bubbles near the surface of the ocean are important in underwater sound propagation, meteorology, sea surface chemistry, and air-sea gas exchange. Depending on their concentrations and size distribution, the entrained bubbles can significantly change the optical properties of water. Once air is entrapped at the sea surface, there is a rapid development stage resulting in a cloud of bubbles. Some bubbles will be several millimeters in diameter, but the majority will be less than 0.1 mm in size. Each bubble is buoyant and will tend to rise toward the sea surface, but the upper ocean is highly turbulent and bubbles may be dispersed to depths of several meters. Small particles and dissolved organic compounds very often collect on the surface of a bubble while it is submerged. Gas will also be slowly exchanged across the surface of bubbles, resulting in a continual evolution of the size and composition of each bubble. The additional pressure at depth in the ocean will compress bubbles and will tend to force the enclosed gases into solution. Some bubbles will be forced entirely into solution, but generally the majority of the bubbles will

eventually surface carrying their coating and altered contents. At the surface, a bubble will burst, generating droplets that form most of the sea salt aerosol suspended in the lower marine atmosphere.

See also: **Air Sea Interactions:** Momentum, Heat, and Vapor Fluxes; Sea Surface Temperature. **Boundary Layer (Atmospheric) and Air Pollution:** Convective Boundary Layer; Modeling and Parameterization; Ocean Mixed Layer; Stably Stratified Boundary Layer; Surface Layer. **Dynamical Meteorology:** Overview. **Numerical Models:** Coupled Ocean-Atmosphere Models: Physical Processes. Parameterization of Physical Processes: Turbulence and Mixing. **Oceanographic Topics:** Surface/Wind Driven Circulation. **Optics, Atmospheric:** Optical Remote Sensing Instruments. **Satellites and Satellite Remote Sensing:** Surface Wind and Stress. **Tropical Cyclones and Hurricanes:** Hurricanes: Observation. **Weather Forecasting:** Marine Meteorology; Operational Meteorology.

Further Reading

- Babanin, A.V., 2011. *Breaking and Dissipation of Ocean Surface Waves*. Cambridge University Press, 480 pp.
- Benilov, A., Ly, L.N., 2002. Modeling of surface waves breaking effects in the ocean upper layer. *Mathematical and Computer Modelling* 35, 191–213.
- Cavaleri, L., Alves, J., Ardhuin, F., Babanin, A., Banner, M., Belibassakis, K., Benoit, M., Donelan, M., Groeneweg, J., Herbers, T., Hwang, P., Janssen, P., Janssen, T., Lavrenov, I., Magne, R., Monbaliu, J., Onorato, M., Polnikov, V., Resio, D., Rogers, W., Sheremet, A., McKee Smith, J., Tolman, H., Vledder, G., Wolf, J., Young, I., 2007. *Wave Modeling – The State of the Art*. In: *Progress in Oceanography*, vol. 75. Elsevier, 603–674.
- Charnock, H., 1955. Wind stress on a water surface. *Quarterly Journal of the Royal Meteorological Society* 81 (350), 639–640.
- Curry, J.A., Webster, P.J., 1999. *Thermodynamics of Atmospheres and Oceans*. Academic Press, p. 467.
- Donelan, M.A., Dobson, F.W., Smith, S.D., Anderson, R.J., 1993. On the dependence of sea surface roughness on wave development. *Journal of Physical Oceanography* 23 (9), 2143–2149.
- Ellison, T.H., 1956. Atmospheric turbulence. *Surveys in mechanics*. In: Batchelor, G.K., Davies, R.M. (Eds.). Cambridge University Press, pp. 400–430.
- Kagan, B.A., 1995. *Ocean–Atmosphere Interaction and Climate Modeling*. Cambridge University Press, 377 pp.
- Kantha, L., Clayson, C., 2000. *Small Scale Processes in Geophysical Fluid Flows*. Academic Press, 883 pp.
- Kharif, C., Pelinovsky, E., 2003. Physical mechanisms of the rogue wave phenomenon. *European Journal of Mechanics B Fluids* 22, 603–634.
- Kinsman, B., 1965. *Wind Waves, Their Generation and Propagation on the Ocean Surface*. Prentice Hall, Inc., 676 pp.
- Kitaigorodskii, S.A., 1973. *The Physics of Air–Sea Interaction*. Israel Program for Scientific Translations, Jerusalem, 236 pp.
- Komen, G.J., Cavaleri, L., Donelan, M., Hasselmann, K., Hasselman, S., Janssen, P.A.E.M., 1994. *Dynamics and Modelling of Ocean Waves*. Cambridge University Press, 532 pp.
- Landau, L.D., Lifshitz, E.M., 1987. *Fluid Mechanics*, second ed. Pergamon Press, 539 pp.
- Lewis, E.R., Schwartz, S.E., 2004. *Sea Salt Aerosol Production: Mechanisms, Methods, Measurements, and Models – A Critical Review*. In: *Geophysical Monograph Series*, vol. 152. AGU, Washington, DC, 413 pp.
- Lokenath, D., 1994. *Nonlinear Water Waves*. Academic Press, 544 pp.
- Ly, L.N., Garwood Jr., R.W., 2000. Numerical modeling of wave-enhanced turbulence in the oceanic upper layer. *Journal of Oceanography* 56 (4), 473–483.
- Monin, A.S., Yaglom, A.M., 1987. *Statistical Fluid Mechanics: Mechanics of Turbulence*, vol. 1. The MIT Press, 769 pp. vol. 2. 874 pp.
- Phillips, O.M., 1977. *The Dynamics of the Upper Ocean*, second ed. Cambridge University Press, 336 pp.
- Sloviev, A.V., Lukas, Rodger, 2006. *The Near-Surface Layer of the Ocean: Structure, Dynamics and Application*. Springer, 572 pp.
- Stewart, R.H., 1985. *Methods of Satellite Oceanography*. University of California Press, Berkeley, 352 pp.
- Stokes, G.G., 1847. On the theory of oscillatory waves. *Transactions of the Cambridge Philosophical Society* 8, 441–455.
- Thorpe, S.A., 1982. On the clouds of bubbles formed by breaking wind-waves in deep water, and their role in air–sea gas transfer. *Philosophical Transactions of the Royal Society of London A* 304, 155–210.

Contents

Aircraft Emissions

Aircraft Icing

Aviation Weather Hazards

Clear Air Turbulence

Aircraft Emissions

RR Friedl, California Institute of Technology, Pasadena, CA, USA

© 2003 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 1, pp 60–68, © 2003, Elsevier Ltd.

Introduction

Human society is becoming increasingly dependent on aircraft for long-distance travel and shipping. Among transportation modes, aviation is the fastest-growing; the current passenger growth rate is approximately 4% per year and the average growth rate since 1960 has been nearly 9% per year. The fraction of transport fuel use by aviation has risen steadily to about 13% currently. Because of the robust growth rate, concern has been expressed over possible environmental impacts of future aircraft operation. Vigorous science and technology programs have been pursued over the last decade to define potential atmospheric impacts and identify technological strategies to reduce specific exhaust emissions. Environmental compatibility issues have also been central to efforts to develop future aircraft technologies such as high-speed (i.e., supersonic) civil transport.

Jet aircraft burning hydrocarbon-based fossil fuels transport the bulk of air passengers and freight. Currently there are over 15 000 aircraft serving nearly 10 000 airports worldwide and burning nearly 140 Tg of fuel per year. By the year 2015, fuel burn by aviation is forecast to increase to approximately 300 Tg per year. As with other fossil fuel transportation technologies, jet aircraft operation results in gaseous and particle combustion byproducts. Aircraft engines emit principally carbon dioxide (CO_2) and water (H_2O) with minor contributions from nitrogen oxides (NO_x), sulfur oxides (SO_x), unburned hydrocarbons (HC), and soot. All of these exhaust species are atmospheric pollutants. CO_2 and H_2O are greenhouse gases that affect the Earth's climate directly. NO_x and HC are reactive gases that affect atmospheric ozone and methane levels. Soot, SO_x , HC, and H_2O are aerosol and cloud precursors that affect ozone and climate.

A major difference between aviation and other transportation modes is in the atmospheric placement of the combustion exhaust products. Unlike autos and trucks, by far the greater part (>85%) of aircraft exhaust is released above the planetary boundary layer (>2 km) and a large fraction (~70%) of it is released in the upper troposphere (UT) and lower stratosphere (LS) between 9 and 13 km. Consequently, the major polluting effects of aircraft are expected to occur in the UT/LS region of the atmosphere.

The dynamics of the UT/LS region differ from those of the boundary layer in that there is less vertical mixing and less diurnal variation in wind direction. Because of these differences, pollutants emitted into the UT/LS reside there longer and can spread over considerable longitudinal and, in some cases, latitudinal distances. Although aircraft exhaust is released in geographically narrow flight routes and corridors, its injection into the UT/LS means that the polluting effects of aircraft will be felt on regional and, perhaps, global scales. The longer residence times also enable some pollutants, such as NO_x , to spend extended times cycling through catalytic chemical reaction sets that create or destroy ozone. Because of such enhanced catalytic chemical cycling in the UT/LS, the impact of a given amount of aircraft emissions on atmospheric ozone and climate may be much greater than the same amount of emissions from ground transportation sources.

There are also important dynamical and chemical differences between the UT and LS regions that complicate any analysis of aircraft effects. For example, the lifetime of ozone and the chemical mechanisms controlling its concentration are sensitive functions of altitude in the vicinity of the tropopause. Because of this altitude dependence, the sign of the ozone response to injections of NO_x shifts from positive (net ozone formation) to negative (net ozone destruction)

at altitudes slightly above the tropopause (i.e., the transition between the stratosphere and troposphere). The partitioning of aircraft exhaust between the UT and LS is difficult to define (estimates differ by factors of two) because of the high variability and latitudinal dependence of the tropopause height.

Aircraft Exhaust Products

Jet engines on modern aircraft are composed of three essential elements: compressors that increase the pressure and temperature of the entering air, combustors that mix and burn fuel with incoming air, and turbines that convert the hot gas energy, through compressor activity, to bypass airflow that propels the aircraft. The fuel-to-air ratio in modern combustors is approximately 1:9, hence large quantities of ambient air are processed in aircraft engines. Jet fuel is composed predominately of high-weight (C_{12} – C_{15}) alkanes, with substantially smaller quantities of alkenes and aromatics present. An important trace species in the fuel is sulfur, which can represent up to 0.3% (by weight) of the fuel content.

Combustion of the fuel hydrocarbons to produce CO_2 and H_2O is nearly complete (>99.5%) in commercial aircraft engines. In addition, the fuel sulfur is converted to sulfur dioxide and sulfuric acid, although the precise mechanism for this process remains uncertain. The small fraction of incompletely combusted fuel hydrocarbons give rise to CO and various smaller gaseous hydrocarbons (HC) such as ethene, ethine, and formaldehyde. Under fuel-rich combustor conditions, breakdown of the fuel hydrocarbons leads to formation of soot particulates composed primarily of carbonaceous material. The rate-limiting process in soot formation appears to involve the oxidation of C_2 species such as acetylene (C_2H_2). Decomposition of ambient nitrogen and oxygen also occurs in the high-temperature portions of the combustor, giving rise to the important atmospheric pollutants nitric oxide (NO) and nitrogen dioxide (NO_2) (i.e., NO_x) (Table 1).

Aircraft Technology Considerations

Aircraft engine and airframe technologies have undergone dramatic improvements over the last 30 years. One result of these improvements has been a 70% reduction in fuel burned per passenger seat from early to current jets. Gains in fuel efficiency are of benefit both economically and in environmental terms by reducing fuel costs and uniformly lowering CO_2 , H_2O ,

and SO_x emissions. These gains have derived primarily from increasing gas temperatures and pressures inside the engines. Without concomitant changes in engine design, increasing engine temperature leads to increasing NO_x emissions.

Concern over urban pollution has led to increasingly stringent standards being adopted by the International Civil Aviation Organization (ICAO) regarding emissions of smoke, CO, HC, and NO_x . Aircraft smoke refers to visible particulates in the aircraft plume and presumably includes the large diameter (>1 μm) part of the soot population. The ICAO standards have both reflected and motivated improvements in engine design and manufacture. However, because the service lifetime of an individual aircraft is between 25 and 40 years, the current fleet consists of a combination of older and newer technologies.

Measurement of aircraft cruise emissions is an important facet of assessing impacts and documenting technological advances. These difficult measurements are made either in altitude simulation test cells or by in-flight measurements utilizing target and chase aircraft.

Aircraft Operational Considerations

Airline traffic patterns are highly inhomogeneous, with the bulk of current traffic located inside well-defined ‘flight corridors’ in the Northern Hemisphere (Figure 1). The chemical lifetimes of aircraft exhaust products such as NO_x , soot, and sulfate injected in the UT/LS are comparable to atmospheric mixing times. Consequently, a number of the aircraft chemical perturbations are expected to be localized in regions around the flight corridors. A great deal of work has been done to compile accurate inventories of aircraft emissions. These efforts have involved development of aircraft movement databases based on simplifying assumptions about the airframe–engine combinations used and the paths flown between various city pairs. Combining these movement databases with information on individual aircraft emission rates enables construction of global emissions inventories. For atmospheric modeling purposes, the aircraft emission databases are divided into spatial bins that are 1° longitude $\times$ 1° latitude $\times$ 1 km altitude.

Impacts on Carbon Dioxide and Water

Although they are the most prevalent exhaust products, emissions of CO_2 and H_2O from aircraft represent relatively small sources of these species compared with the many other large natural and anthropogenic sources. Given past and current emission rates, aircraft are responsible for increasing atmospheric CO_2 levels by approximately 1 ppmv or 2% over the last 50 years. Because CO_2 is very long-lived in the atmosphere and is well mixed, it is impossible to distinguish the CO_2 emitted from aircraft from any other source. Perturbations due to aircraft H_2O emissions are far less than 1% globally. These small perturbations are impossible to detect on the large scale because water vapor has a short (days to weeks) tropospheric residence time and its ambient concentrations are highly variable. At very small spatial scales, H_2O perturbations from aircraft are substantial and can lead to contrail and cirrus cloud

Table 1 Approximate emission index levels for cruise level operation of current commercial jet aircraft

Species	Emission index ($g\ kg^{-1}$)
CO_2	3160
H_2O	1240
CO	2
HC	1
NO_x (as NO_2)	12
SO_x (as SO_2)	0.8
Soot	0.04

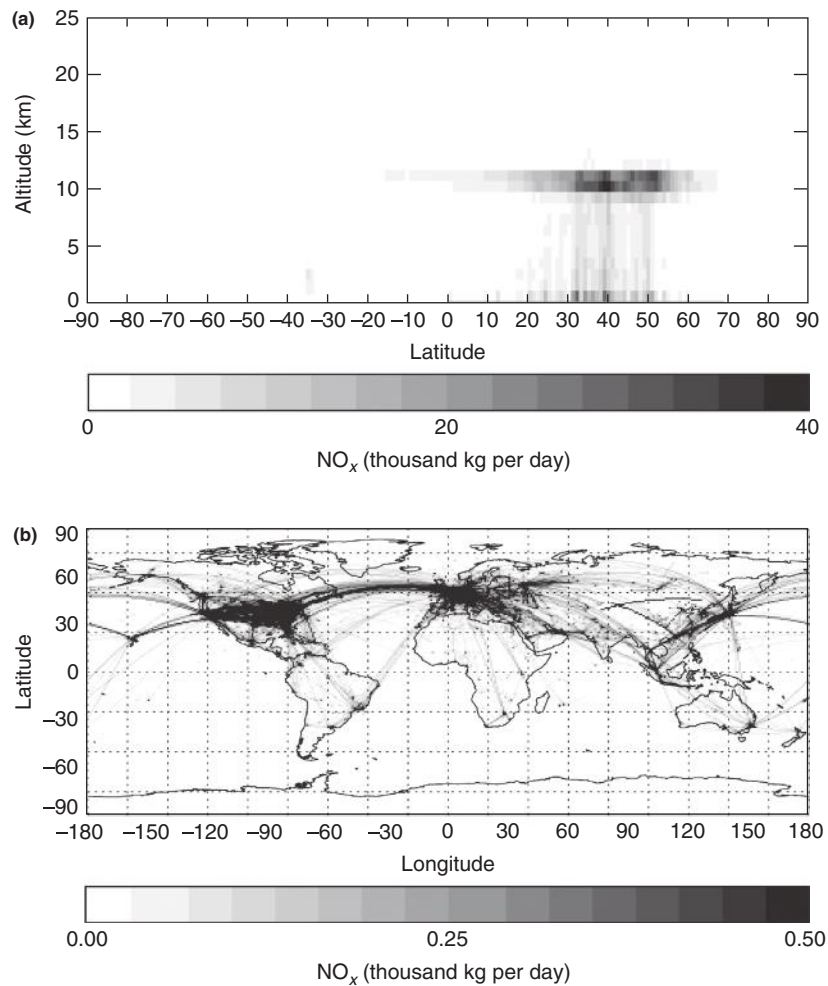


Figure 1 Calculated NO_x emissions for all aircraft traffic in May 1992 as a function of altitude and latitude, summed over longitude (a), and as a function of latitude and longitude summed over altitude (b). Values greater than the range maximum are plotted as black. (From NASA reference publication 1400.)

formation. These effects have important climate consequences that will be discussed below.

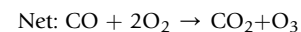
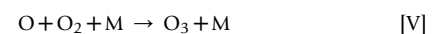
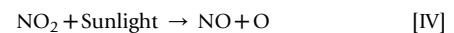
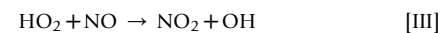
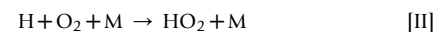
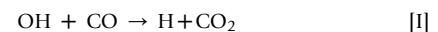
Impacts on Ozone and Methane

Ozone chemistry throughout the stratosphere and troposphere is driven by solar-initiated free radical reactions. Aircraft emit a number of species (i.e., NO_x , SO_x , H_2O , CO, and soot) that participate in ozone-controlling reactions of free radicals and free radical precursors. The relationship between aircraft exhaust products and ozone is complex and depends on the balance between a number of ozone-forming and -depleting chemical processes. These processes are summarized in the next two sections along with observational evidence that addresses the magnitude of the aircraft effect on ozone.

Atmospheric Chemistry

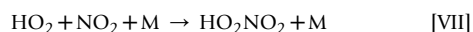
Nitrogen oxides in the UT/LS participate in both ozone-forming and ozone-depleting reaction cycles. The balance

between these processes, and their response to changes in ambient NO_x levels, are sensitive functions of altitude. In the UT region, the primary influence of NO_x is on the production of ozone from CO and CH_4 oxidation. The CO cycle involves the following reactions:

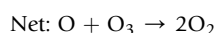
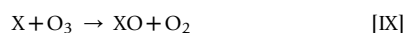


(where M represents a gaseous third body such as N_2 or O_2). An analogous mechanism, which includes the reaction between NO and CH_3O_2 , exists for CH_4 oxidation.

The overall rate of ozone production from CO and CH₄ oxidation decreases generally with height because of decreasing ambient concentrations of CO and CH₄. However, as can be seen from the above reaction sequence, the production rate depends also on the ambient concentrations of NO and OH. For example, increasing OH and NO will increase the rates of reactions [I] and [III], respectively, thereby increasing the ozone production rate. At high enough concentrations of NO_x (>500 pptv) the ozone production rate begins to decrease owing to the increasing importance of NO₂ reactions that remove HO_x species, i.e.,



In the LS region, the primary influence of NO_x is on destruction of ozone by the following radical-catalyzed processes:

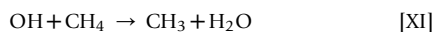


(where X = NO, Cl and OH). Increases in ambient NO_x due to aircraft emissions will enhance ozone destruction for the case where X = NO but will decrease ozone destruction for the case X = Cl by removing ClO through reaction [X]:



The exact balance between these contrary effects depends on the background concentrations of NO_x and ClO_x. Throughout most of the year, the net effect of increasing LS NO_x levels through aircraft emissions will be to increase ozone concentrations. An exception is at high latitudes in summer, when NO_x levels are high. In that case, addition of NO_x will decrease ozone.

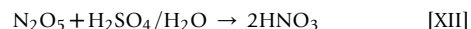
Increasing NO_x levels due to aircraft exert an indirect effect on atmospheric CH₄ concentrations. This effect is initiated by the formation of OH radicals in reaction [III]. The increased levels of OH in the air traffic corridors lead to decreases in carbon monoxide (CO) through reaction [I]. Because CO lifetimes are longer than NO_x lifetimes, the region of decreased CO concentrations spreads out from air traffic corridors to a much greater extent than do the aircraft NO_x emissions. The CO perturbation spreads all the way to tropical and subtropical regions where much of the global oxidation of CH₄ takes place through its reaction with OH:



As CO levels are lowered in the tropics, OH levels are raised correspondingly. The higher levels of OH serve to lower CH₄ concentrations, which, in turn, lead to a further increase in OH. As a result of the complex interplay (i.e., atmospheric feedback cycle) between NO_x, OH, CO, and CH₄, an increase in NO_x will lead to an amplified decrease in CH₄. The amplification factor is approximately 1.5.

Aircraft emissions of SO_x, H₂O and soot also effect atmospheric ozone concentrations by serving as aerosol precursors. In the UT, sulfate- and water-ice-containing aerosols promote ozone decreases by acting as surfaces for

heterogeneous removal of the ozone precursors NO_x and HO_x. A major identified heterogeneous reaction involves conversion of the temporary NO_x reservoir species nitrogen pentoxide (N₂O₅) into the longer-term reservoir nitric acid (HNO₃).



HNO₃, along with a number of other nitrogen and hydrogen acids and peroxides (e.g., HNO₄ and H₂O₂) are absorbed onto sulfate and water-ice. The absorbed species can be removed from the UT by sedimentation.

In the LS, sulfate- and water-ice-containing aerosol particles not only remove HO_x and NO_x species but also liberate ozone-destroying ClO_x by heterogeneous reactions such as



The net effect of the heterogeneous processes is to decrease ozone in the LS and UT. However, the effect of the aircraft-derived aerosols on LS/UT ozone offsets only partially the effect of the NO_x emissions.

Much less is known about the effect of soot particulates on ozone. Ozone is observed to react directly on laboratory soot surfaces, but the reaction slows as the surface is modified. Heterogeneous reactions of NO_x and nitrogen reservoir species also occur on soot surfaces – in some cases the reactions lead to more reactive species, in others to less reactive ones. Consequently, the effect of aircraft soot on atmospheric ozone concentrations is poorly determined at present.

According to the current scientific understanding, the overall effect of aircraft emissions in the UT/LS is to increase ozone levels. Model calculations indicate that aircraft have increased ozone by about 6% in heavy traffic areas, with an associated 0.4% increase in the total ozone column. In terms of climate effects, the radiative forcing changes due to increased ozone appear to be largely offset by the predicted decreases in methane. Considerable uncertainty is attached to these calculations, however (see Figure 2).

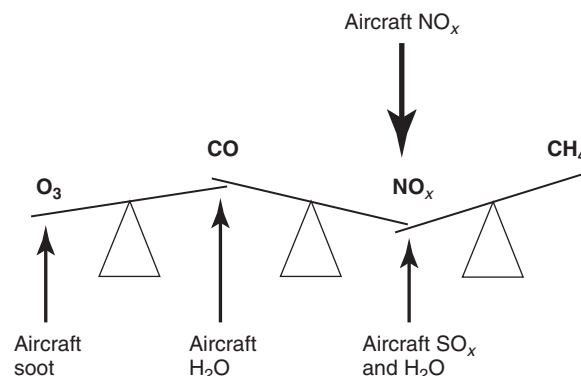
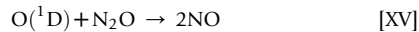
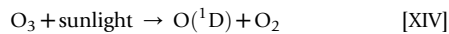


Figure 2 Influence of aircraft emissions on chemical balance in the UT/LS region. Atmospheric chemical reactions couple together O₃, CO, NO_x, and CH₄. Among aircraft emissions, NO_x is calculated to have the greatest effect on the coupled species, acting to increase ambient NO_x and O₃ levels and decrease CO and CH₄.

Observing Ozone Impacts

Dense air traffic in Northern Hemisphere flight corridors will give rise to distinct geographical perturbations of NO_x , aerosols and ozone under two conditions. First, large-scale dispersion of the exhaust must be slower than the chemistry that removes and/or links these emissions to ozone. Second, the strength of the aircraft emissions must be significant relative to other natural and anthropogenic sources of NO_x and aerosols. The total NO_x emission from current global aviation is approximately 0.5 Tg per year, of which roughly 60% is released into the upper troposphere and 15% is released into the lower stratosphere.

The major source of NO_x in the lower stratosphere is chemical oxidation *in situ* of nitrous oxide (N_2O):



The global production rate of NO from N_2O (~ 12 Tg per year) far exceeds that from current subsonic aircraft emissions in the lower stratosphere. Hence there is no expectation, nor observational evidence, that current aircraft are significantly perturbing stratospheric NO_x levels.

In the upper troposphere, the major non-aircraft sources of NO_x include fossil fuel combustion (autos, trucks, etc.), biomass burning, soil emissions, lightning, and N_2O oxidation. Of these, only lightning deposits NO_x directly into the UT. The fractions of NO_x transported into the UT from sources at the Earth's surface or in the stratosphere are small, occurring only during convective events, such as frontal activity or thunderstorms or during stratosphere-troposphere exchange events triggered by meteorological features such as extratropical cyclones. Source strength estimates for the various NO_x sources are listed in (Table 2). As shown in the table, aircraft emissions into the UT are of comparable strength to other sources and contribute a significant fraction of UT NO_x .

Chemical sampling of the UT in and around traffic corridors has revealed each individual aircraft perturbs ambient NO_x levels substantially for distances of several kilometers behind it. At larger spatial scales, aircraft signatures have not been discerned, owing to the high variability of background NO_x . Likewise, there have been no identifiable spatial patterns in ozone concentrations that unambiguously point to production by aircraft NO_x . Long-term ozone trend observations at specific measuring stations (e.g., Hohenpeissenburg, Germany, and Wallops Island, USA), do not correlate with the

growth rate of air traffic from 1970 to the present, indicating that aircraft emissions are not a major factor in the observed upper-tropospheric trends.

Impacts on Clouds

Trails of ice particles – contrails – are the most readily identifiable exhaust signatures of aircraft (*see Clouds and Fog: Contrails*). Contrails often form, even under clear-sky conditions, because aircraft H_2O emissions raise the relative humidity of the air near the exhaust plume above 100%. Water vapor in the supersaturated air subsequently condenses on aircraft-derived soot and sulfate nuclei and freezes to form ice. If the surrounding air is very dry and/or warm, contrails may be short-lived or may not form at all. In either case, the emitted soot and sulfate nuclei will remain in the atmosphere for days and weeks and possibly promote natural ice (cirrus) cloud formation in locations far from the initial aircraft plume. These same nuclei, upon contact with cirrus clouds, may change properties of the cloud particles such as size distribution, number density, and chemical composition.

Like other naturally occurring clouds, contrails and aircraft-induced (or modified) clouds impact the Earth's climate by affecting the radiation balance. For typical particle properties, cirrus clouds trap surface outgoing long-wave radiation more effectively than they reflect solar incoming short-wave radiation. As a result, cirrus clouds tend to warm the climate. However, the magnitude and even the sign of a cloud's radiative effect on climate is a sensitive function of cloud particle size and shape as well as altitude and geographical location.

Cloud Formation Processes

Clouds or contrails can form when air moisture becomes supersaturated with respect to ice. The ice formation process takes place by one of several mechanisms. At higher supersaturations, low-volatility gas phase species will cluster together to form liquid particles. These liquids can subsequently freeze in a process known as homogeneous freezing if the air is cooled by upward dynamical motion. At lower supersaturations, the freezing process may be aided by the presence of a solid particle surface in a process termed heterogeneous freezing. The liquid and solid particles that readily promote freezing and ice crystal growth are typically in the 0.05–1 μm diameter size range; they are referred to as cloud condensation nuclei (CCN) and ice nuclei (IN), respectively. Aircraft emissions may enhance the frequency of these freezing events by increasing the abundances of CCN and IN.

Aircraft soot emissions have attracted attention as a possible source of IN in the UT. The median size of a fresh aircraft soot particulate is approximately 0.02 μm . In order for a soot particulate to become an IN it must be activated (i.e., become more hydrophilic) by reaction with suitable species. Lab studies have shown that acids such as H_2SO_4 induce this activation but that others such as HNO_3 do not. Hydrated samples of soot have been obtained from

Table 2 Present-day sources of NO_x in the troposphere and their approximate strengths

Source	Emission rate (Tg yr^{-1}) Total	Emission rate (Tg yr^{-1}) 9–13 km altitude band
Aviation	0.5	0.3
Fossil fuel combustion	22	0.7
Biomass burning	8	0.2
Soil emissions	7	0.2
Lightning	5	1
N_2O oxidation	12	0.6

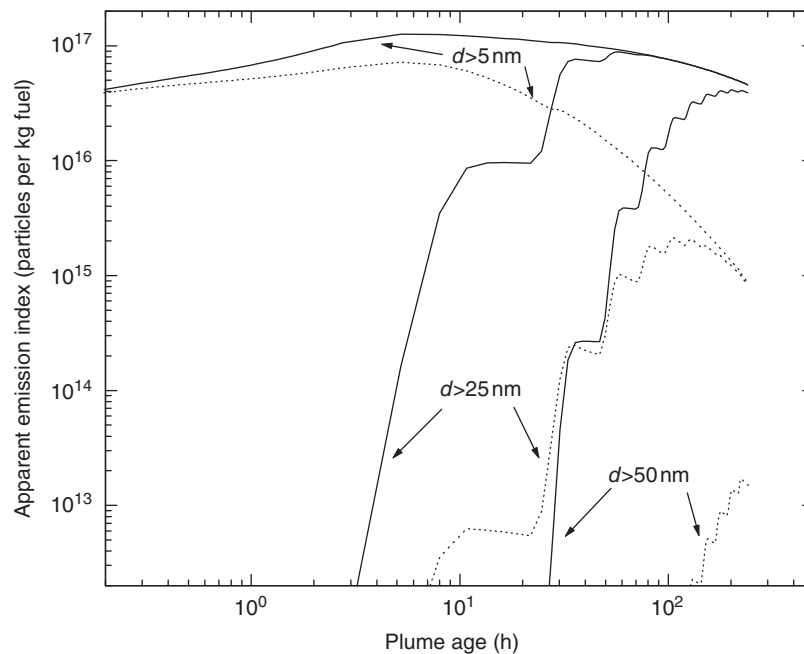


Figure 3 Calculated time evolution of the ‘apparent’ emission indices of aircraft-generated particles for various size thresholds. Solid and dashed lines are for low and high ambient aerosol conditions, respectively. Appreciable concentrations of CCN size particles (>50 nm) are predicted only for low ambient aerosol conditions. Adapted with permission from Yu F and Turco RP (1999) *Geophysical Research Letters* 26: 1703–1706. Washington, DC: American Geophysical Union.

non-sulfur-containing flames, indicating the presence of other, as yet unidentified, activating species.

Aircraft emissions of condensable gases such as sulfur oxides and oxygenated hydrocarbons can contribute to CCN formation. In an aircraft plume, large numbers ($\sim 10^{16}$ particles per kilogram of fuel) of small particles (<0.01 μm radius) are formed from nucleation of sulfuric acid and water. The formation and subsequent growth of these particles may be accelerated by chemi-ions that are emitted into the plume following their production in high-temperature reactions occurring in the combustor. As the plume expands and is diluted by entrainment of ambient air, the small plume particles may continue to grow by uptake of additional gaseous species or they may be scavenged by larger ambient particles. The competition between these two processes depends on a number of environmental variables such as air temperature, relative humidity, and background aerosol concentration. Under low background aerosol conditions, such as exist during wintertime, a significant number of plume particles are expected to survive long enough to grow to CCN size (see [Figure 3](#)).

Addition of aircraft-derived CCN and IN to the UT will increase cirrus cloud occurrence in areas where the air is supersaturated with respect to ice but crystal growth is limited by a lack of sufficient numbers of nuclei. Relative humidity measurements taken in the UT reveal that ice supersaturation occurs in more than 10% of the clear-sky cases examined. Cloud growth in these regions should be particularly susceptible to aircraft IN and CCN. In areas of developing cirrus clouds, aircraft-derived CCN and IN may influence the properties of the cloud particles in one of several ways. If the aircraft particles are larger and function as more active growth

nuclei than ambient particles, they may compete effectively for the available water vapor and induce growth of larger ice crystals at the expense of crystal number density. If, on the other hand, the aircraft particles increase the number of CCN and IN, but do not change the overall rate of crystal growth, then increases in the crystal number density are expected, with concomitant decreases in average crystal size. The radiative properties of the resulting clouds will be altered, but the magnitudes and characteristics of these modifications are uncertain.

Observing Cloud Impacts

Cirrus cloud coverage, as documented by surface and satellite observations, has been increasing over a number of regions in the last two decades, with the largest increases observed over regions of heavy air traffic in the United States and the North Atlantic. Growth of cloud cover in air corridor regions has been approximately 1–2% per decade greater than in other areas; attributing this growth rate to aircraft impacts implies that there has been an overall 5% increase in traffic route cloud cover during the last 30 years of air travel. Contrails are clearly a significant part of this increase. For instance, analysis of satellite images has indicated that contrail coverage over Europe is on the order of 1%.

Indirect effects of aircraft-derived aerosols on cloud formation may be responsible for the rest of the observed increase. However, a number of other natural and anthropogenic causes such as changes in UT temperature and humidity, greenhouse gas concentrations, and upper-atmosphere dynamics may be

Table 3 Ozone and climate impacts of present day aviation. Radiative forcing is used as a measure of the climate impact

Impact	Ozone column (% change)	Radiative forcing ($W\ m^{-2}$)
CO ₂		+0.016
NO _x	+0.5	+0.024
CH ₄		−0.015
H ₂ O		+0.002
Sulfate aerosol	−0.1	−0.003
Soot particulate	−0.01	+0.003
Contrails		+0.021
Cirrus clouds		≤+0.04

Numbers are taken from IPCC (1999).

contributing to the increased cloudiness. Currently, there is little observational information from which to determine if aircraft are significant sources of CCN and IN. In a few cases, sampled contrail and cirrus cloud particles have contained significant amounts of soot and metals, suggestive of an aircraft influence. But the few direct measurements of CCN in aircraft wakes have yielded concentrations ranging from very low to very high values. Similar types of measurement for IN have suggested that aircraft wakes do not contain large numbers of these nuclei.

Impact Summary: Present and Future

Scientific progress on aviation impacts has progressed to a point where it is providing important guidance to technology and policy decision-makers. Aircraft CO₂ emissions are known to be small but significant contributors to the observed rise in ambient levels, accounting for approximately 2% of the current total anthropogenic CO₂ emissions. In addition, it is now clear that the overall radiative forcing by aircraft is substantially larger, by a factor of 2 to 5, than the forcing by aircraft CO₂ alone. In terms of possible mitigation strategies, this finding underscores the need to reduce many of the aircraft emissions simultaneously or consider tradeoffs in aircraft designs and operational measures. Examples of tradeoffs include the development of more efficient (i.e., emitting lower CO₂) engines at the expense of higher NO_x emissions and/or promotion of contrails, and increases in average flight altitudes to reduce contrail formation at the expense of increased O₃ sensitivity to NO_x. Consideration of these tradeoffs is tempered by the fact that the impacts from aircraft CO₂ are far better characterized than the impacts from NO_x, contrails, and aerosols (Table 3).

The calculated impacts of the current aircraft fleet on climate and ozone are small relative to the natural atmospheric variability and have not been observed. Like climate and ozone impacts of other individual sectors, aircraft emissions represent a small fraction of the total anthropogenic impact. Consequently, detection of the aircraft-specific contribution to climate and ozone change is not possible at present. Future growth of aircraft emissions is related to economic and population growth. Under some plausible high-growth scenarios, aircraft emissions may increase fivefold over the next 50 years and contribute larger fractions of the anthropogenic climate forcing and ozone change. In addition, the possible

introduction of a supersonic aircraft fleet may contribute to significant changes in stratospheric water vapor and ozone. The potential impacts of supersonic aircraft emissions have been studied in the 1970s and again in the 1990s in concert with industry interest in building a fleet of such aircraft. From these studies, the higher altitude (i.e. 16–20 km) release of supersonic NO_x and H₂O emissions is predicted to generally enhance ozone-depleting cycles (reactions VIII and IX) relative to ozone-forming ones (reactions IV), and lead to a net decrease of ozone. The water vapor emissions from supersonic aircraft, occurring predominately in the relatively dry lower stratosphere, are also predicted to contribute to climate warming.

Even if these future subsonic and supersonic scenarios come to pass, detection of specific aircraft climate and ozone impacts will continue to be difficult. Consequently, technology and policy decision-making will have to rely on observations of overall global change along with a robust scientific understanding of aviation effects.

Acknowledgements

I dedicate this article to my wife Myrna; her long battle with illness ended sadly during its preparation. I thank Drs S. Baughcum and F. Yu for use of their data in Figure 1 and Figure 3. Preparation of the article was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under contract with the National Aeronautics and Space Administration.

See also: Aerosols: Soot. Chemistry of the Atmosphere: Chemical Kinetics; Principles of Chemical Change. **Clouds and Fog: Contrails. Mesoscale Meteorology:** Overview. **Ozone Depletion and Related Topics:** Long-Term Ozone Changes. **Stratospheric Chemistry Topics:** HO_x; Reactive Nitrogen (NO_x and NO_y); Stratospheric Water Vapor. **Tropospheric Chemistry and Composition:** Aerosols/Particles; Sulfur Chemistry, Organic.

Further Reading

- Brasseur, G., Cox, R.A., Hauglustaine, D., et al., 1998. European Scientific Assessment of the Atmospheric Effects of Aircraft Emissions. In: Brasseur, G., Amanatidis, G.T., Angeletti, G. (Eds.), 1998. *Atmospheric Environment*, 32, pp. 2327–2418.
- Brasseur, G.P., Orlando, J.J., Tyndall, G.S. (Eds.), 1999. *Atmospheric Chemistry and Global Change*. Oxford University Press, New York.
- Friedl, R.R., Anderson, B.E., Baughcum, S.L., et al., 1997. *Atmospheric Effects of Subsonic Aircraft: Interim Assessment Report of the Advanced Subsonic Technology Program*. NASA Reference Publication 1400. NASA Goddard Space Flight Center, Greenbelt, MD.
- Intergovernmental Panel on Climate Change, 1999. *Aviation and the Global Environment*. In: Penner, J.E., Lister, D.H., Griggs, D.J., Dokken, D.J., et al. (Eds.). Cambridge University Press, Cambridge.
- Turco, R.P., 1997. *Earth Under Siege, From Air Pollution to Global Change*. Oxford University Press, New York.
- SONEX/POLINAT Special Section, 2000. *Geophysical Research Letters* 26, 3053–3084. Washington, DC: American Geophysical Union.
- SONEX/POLINAT Special Section, 2000. *Journal of Geophysical Research* 105, 3595–3892. Washington, DC: American Geophysical Union.
- Wayne, R.P., 1991. *Chemistry of Atmospheres*, second ed. Clarendon Press, Oxford.

Aircraft Icing

MK Politovich, National Center for Atmospheric Research, Boulder, CO, USA

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

Aircraft icing is described in terms of its effect on an airplane, its physical characteristics, associated meteorology, and means to detect and diagnose the icing environment. Diagnosing and forecasting icing require understanding of the physical processes governing icing products, recognizing how these processes relate to observable phenomena, and combining information from as many sources as possible to accurately depict relevant characteristics of the icing environment.

Introduction

Aircraft icing is the accretion of supercooled water onto an airplane during flight. Accreted ice adversely affects flight, thus, it is an important component of an aviation weather forecast. Meteorology associated with inflight icing begins with the microscale, addressing growth of water drops and their collision with and adhesion to airframes. Cloud-scale and meso-scale processes control the amount and distribution of water drops while synoptic weather patterns, which produce what was generally referred to as 'weather', govern the location and movement of icing environments. Diagnosing and forecasting inflight icing involve the development and use of numerical weather prediction models as well as *in situ* and remote sensors. Carburetor icing and precipitation or frost adhering to the wings of an airplane prior to takeoff are not covered here.

Effect on an Airplane

Meteorologists, aerospace engineers, and pilots need and want information about icing because it can adversely affect the flight characteristics of an aircraft. Icing can increase drag, decrease lift, and cause control problems. The added weight of the accreted ice is generally a factor only for light aircraft.

Aircraft can fly safely in icing conditions, but to do so legally they must complete a certification process. To certify a particular type of airplane, it must be flown in a range of natural icing conditions and demonstrated that these conditions result in no significant effect on the airplane's ability to fly. The range was developed from measurements obtained in the 1940s and is illustrated in the accompanying figure, which was designed to envelop 99.9% of icing conditions found in stratiform clouds (Figure 1). More recent studies have confirmed that these provide reasonable limits for certification, although they do not address the problem of large supercooled drops (such as freezing drizzle or rain) or mixed-phase (supercooled liquid drops and ice crystals) conditions.

Certified aircraft are commonly equipped with devices that either prevent ice from adhering to the airframe or remove it once it has been adhered. Such anti- or de-icing equipment can either be deployed manually or through an automatic system triggered by an icing detection probe and includes pneumatic 'boots', heat, and liquids. These are usually applied to the leading edges of the wings and tail and occasionally to the propellers.

Tailplane icing is a subset of icing and refers to icing that accretes on the vertical and horizontal stabilizers of specific airplane types. It is not necessarily caused by unique atmospheric conditions, but is usually referred separately because it results in vastly different response of the airplane than does icing which affects the wings. Thus, pilots require special training for this hazard.

Icing tends to affect general aviation more than commuter or air carrier operations. The smaller aircraft included in the general aviation category tends to fly at lower altitudes where icing is more prevalent. Those aircraft may have less deicing capability and reserve power in case of encountering icing conditions, and their pilots may have less experience operating icing. Air carriers tend to quickly penetrate icing-bearing clouds on ascent and descent from airports and cruise at altitudes far above those where icing resides. Commuter aircraft are caught in the middle both in terms of their ability to handle ice and the altitudes they fly.

Icing Severity

Icing is currently classified into four severity categories: trace, light, moderate, and severe. The most important atmospheric parameters determining severity are the liquid water content, outside air temperature, and drop size. The more water there is, the more is available to accrete on the airframe, thus higher liquid water contents are usually associated with more severe conditions. Temperature influences what happens to that liquid once it impacts the airframe – either it freezes in place or

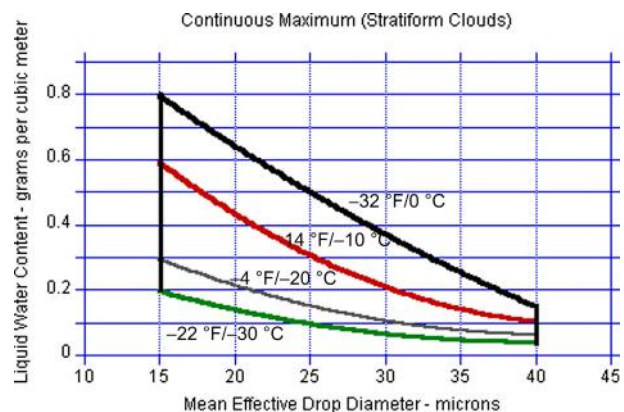


Figure 1 Icing envelopes defined by liquid water content, drop size, and temperature. From FAA Federal Aviation Regulations Part 25, Appendix C.

runs aft along the surface before freezing to unprotected areas. Drop size controls the collection efficiency of those drops onto the airframe. Drop size is not as important as liquid water content or temperature in determining severity until drops reach drizzle sizes, with diameters exceeding $\sim 50 \mu\text{m}$. Research is being conducted to determine appropriate limiting values for these parameters to define severity categories. The definitions must relate atmospheric conditions to observable information as well as effect on flight in order to be useful.

Types of Icing

There are two main physical types of icing: glaze and rime. Mixed icing is a combination of the two. Rime ice is brittle and opaque and tends to grow into the airstream. It is formed as the drops freeze immediately upon impact. Glaze icing, sometimes referred to as clear icing, can be nearly transparent and has a smoother surface, sometimes with a waxy appearance. It is formed when the drops deform and/or flow along the surface prior to freezing. Glaze icing can be more serious to the aircraft than rime since it tends to run back along the airframe, covering more surface area than rime icing, perhaps flowing and adhering to unprotected areas. Glaze icing also can be hard to see from inside the aircraft and the pilot may be unaware of ice buildup. Mixed icing often occurs in layers, as a transition from rime to clear conditions is encountered. These icing types are illustrated in Figure 2.

The type of icing is related to the air temperature, the liquid water content, and the size of the drops. Glaze (rime) is generally associated with higher (lower) temperatures, higher (lower) liquid water contents, and larger (smaller) drops. There are also effects dependent on the airplane itself, including its wing shape and airspeed.

Location and Frequency of Icing Conditions

Icing-related fatal aircraft accidents average ~ 20 – 40 per year in the United States, with the highest incidence in the winter months. Alaska has by far the highest accident rate, followed by the Northwest, Great Lakes, western Pacific states, and the central United States. In North America, icing conditions are most common along the Pacific Coast from Alaska to Oregon and in a large swath from the Canadian Maritimes to the Midwest. Prime locations migrate seasonally, moving south in the summer and retreating to the north in winter. Figure 3 shows latitudes and altitudes conducive to icing in mid-winter in the northern latitudes; those altitudes at which small general aviation and commuter aircraft fly have more exposure to these environments than transport-category aircraft with higher cruise altitudes.

The average altitude of pilot report of icing is $\sim 10\,000$ ft MSL, with few encounters above $20\,000$ ft. Frequency of icing encounters by aircraft based on time of day is a direct reflection of the frequency of flights, with few reports overnight. The

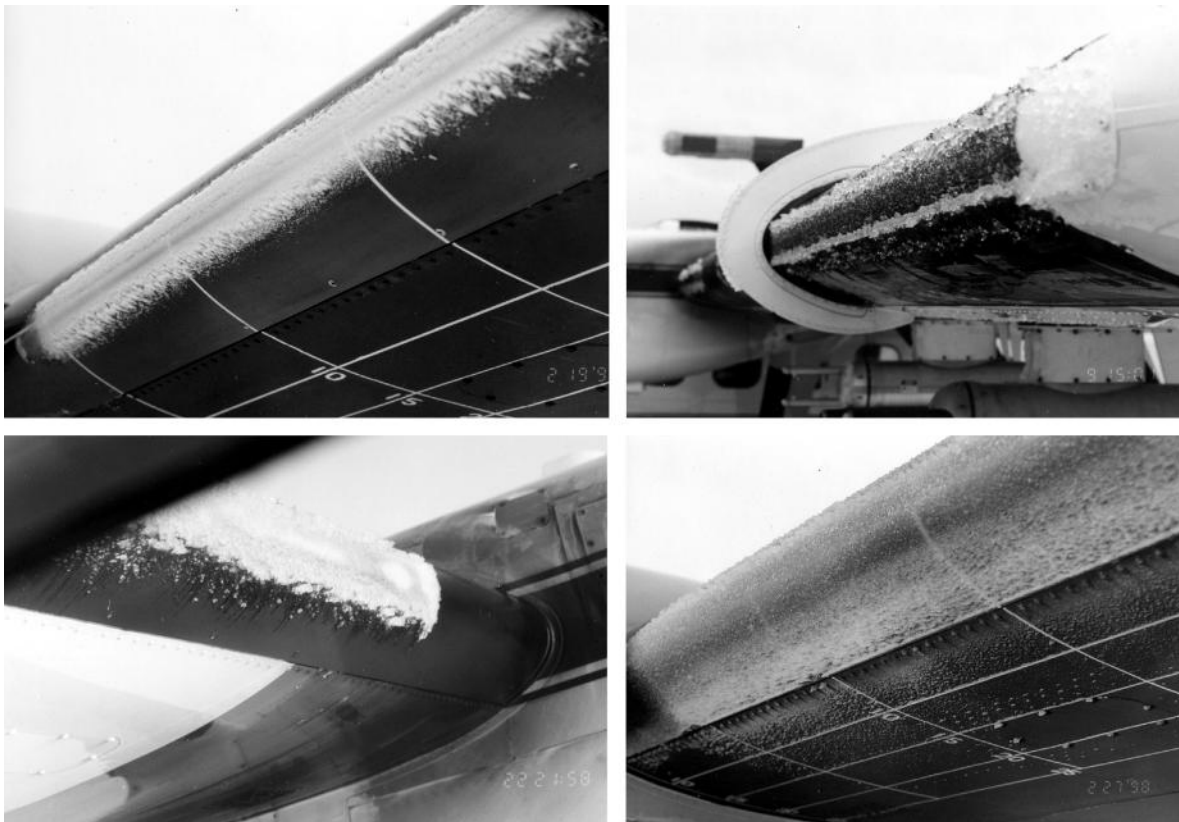


Figure 2 Post-flight photographs of ice encountered by the NASA Glenn Research Center's instrumented Twin Otter aircraft. The leading edge of the left wing is shown in each photograph. Top left – light rime ice; Top right – severe glaze ice; Bottom left – moderate mixed ice; Bottom right – SLD ice. Photos courtesy of NASA Glenn Research Center.

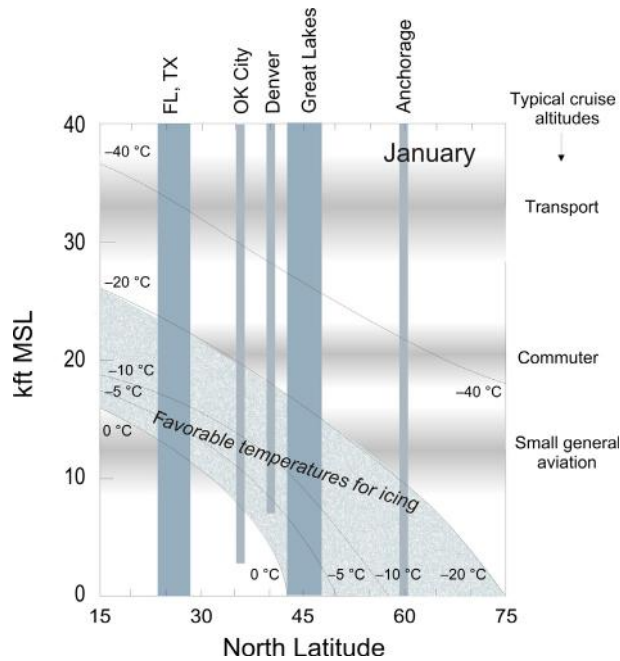


Figure 3 Temperature versus latitude and altitude, based on a Standard Atmosphere for January. Geographic locations, typical flight altitudes and favored temperature for icing conditions are indicated.

weekly pattern also follows air traffic trends, with most reports on Tuesday through Thursday.

Light icing is the most frequent severity category reported by pilots, ~60–70% of all reports. Severe icing, which indicates a condition which cannot sustain flight, is reported only a few percent of the time. Rime icing is reported much more frequently than glaze or mixed, comprising ~70–75% of reports. For both icing type and severity, the largest joint frequency is for light rime icing, which covers nearly half of all reports.

Relation of Icing to Weather Features

The presence of supercooled liquid water in clouds results from production and depletion processes.

- **Production:** Liquid water is produced by bringing the air to >100% relative humidity (supersaturation), which usually results from cooling the air by lifting.
- **Depletion:** Liquid is depleted by mechanisms that erode the cloud, such as entrainment of dry air, or through precipitation processes. In the majority of mid-latitude clouds and storms, the precipitation process is heavily dependent upon the ice phase.

Regions ahead of or near surface warm fronts are favorable icing regions since they provide widespread lifting of generally moist air. Cold fronts also provide opportunities for icing, with narrower bands of more intense lift near the surface front. Moist, maritime air masses are associated with higher frequencies of icing conditions, whereas continental air masses, especially those well behind arctic fronts, have fewer reports. Topography also influences icing, providing a local source of

lift. For example, cold fronts progressing southward through the central United States often provide widespread icing conditions along the front range of the Rocky Mountains from Wyoming through New Mexico. In these situations, which may also occur along the Appalachians in the eastern United States, cold, moist air is forced up the gentle slope leading to the steep mountain range. Or, orographic clouds may be isolated and associated with mountain peaks and ridges.

Precipitation-forming processes tend to deplete liquid water from clouds. Once drops reach a diameter of $\sim 20 \mu$, they begin to collide with one another and coalesce. The increased mass of the resulting larger drop subsequently increases its fall speed, leading to more collisions, etc., depleting the liquid that existed in small drops as it falls out of the cloud. Similarly, once ice crystals are formed in a cloud, they also fall and collect cloud drops in a process referred to as riming. This tends to deplete liquid – much like the accretion of the supercooled drops onto an airplane.

Although it is often assumed that icing will not be present aloft where there is significant precipitation at the surface, examinations of pilot reports of icing have not borne this out. Chances are about even that one of these reports will be associated with surface snow or rain as opposed to no precipitation. If lift within a cloud is strong enough, water can still be condensed onto existing drops, or new drops may be formed, thus continuing the icing condition.

Icing from 'Cloud-Sized' Drops

Measurements of the microphysical characteristics of icing environments have been obtained from the 1940s through the present. Even though a variety of instruments have been used for these measurements, the data sets form a fairly consistent picture of the icing environment. Temperatures range from 0°C to less than -25°C , with a mean around -10°C . Few icing encounters (≤ 4 –10% depending on location) occur at temperatures below -0°C . At temperatures $\geq -5^\circ\text{C}$, adiabatic compression may increase the actual temperature along the leading edges of the airframe to above freezing (typical dynamic heating corrections are 1 – 2°C for small, slow aircraft, to as much as 6 – 8°C or more for large, faster-flying air carriers).

In convective clouds, 90% of the liquid water contents are <0.5 – 0.7 g m^{-3} and for stratiform cloud <0.3 – 0.5 g m^{-3} . Maximum values are typically ~ 1.2 – 1.3 g m^{-3} but can reach higher values in deep convective clouds. Variations occur for different data sets used in the reported analyses.

Drops are typically small, with average mean diameter or median volume diameter (both are used in icing characterization) usually between 10 and $20 \mu\text{m}$. Maximum values for drop sizes are 30 – $50 \mu\text{m}$, depending on the data set. Cumuliform clouds tend to have larger drops than stratiform clouds, and clouds in continental areas have smaller drops than maritime areas. Liquid water content and drop size generally increase with altitude in single cloud layers, but the behavior is less predictable in multi-layered clouds. These are general guidelines since individual clouds vary considerably from one another and variations within clouds occur.

It is not the case that larger cloud drops, and thus greater icing hazards, are usually found lower in the cloud. The cloud top, in most cases, will have the greatest icing potential. In a rising parcel of air containing cloud drops, the drops grow with time and altitude if supersaturation is maintained. Generally, well above cloud base no new drops are formed so any excess vapor condenses on the available drops and they grow. Thus, larger cloud drops and greater liquid water content are generally found higher in the cloud although there will be exceptions based on the thermodynamic structure of the atmosphere and the lifting mechanism.

Supercooled Large Drop Icing

Supercooled large drops (SLDs), which are those with diameters exceeding 50 μm , can pose an especially serious threat to flight. Their larger size means that they are not as likely as small drops to be carried around the airframe with the airstream but will more readily impact on the airframe. They can impact farther aft than small drops, or flow along the aircraft surface before freezing, which means that they may accrete onto areas not usually protected by de- or anti-icing devices. Roughness resulting from this type of ice accretion can create a high amount of drag. Cases of increased performance degradation due to flight in SLD conditions are well documented from research aircraft.

There are two general situations for formation of SLD. The first is the classic freezing rain process, by which snow forms aloft, falls into an intruding warm ($T > 0^\circ\text{C}$) layer, melts, continues to fall into lower cold air ($T < 0^\circ\text{C}$), and becomes supercooled, ready to adhere to an airplane. This is a relatively easy forecast problem since it requires a specific thermodynamic profile. The other general case is the formation of SLD by coalescence of liquid drops and is not so easily recognized using operationally available data sets. Wind shear (differences in wind speed and/or direction) at cloud top in stratiform clouds may encourage the formation of SLD there. There is some evidence that minimum thresholds of liquid water content must be exceeded for drizzle formation to occur; 0.2–0.25 g m^{-3} in continental and around 0.1 g m^{-3} in maritime clouds. This difference emphasizes the need for the inclusion of realistic regional microphysics parameterizations in numerical weather prediction models.

The observation of freezing precipitation – freezing drizzle, freezing rain or ice pellets – at the ground can provide an important clue for SLD conditions aloft. This makes physical sense since all three are supercooled (or already frozen) large drops – if they are present at the surface, they must be present for some depth above the surface. The more difficult part of using this to diagnose SLD conditions aloft is to determine how far aloft the SLD extend. Knowledge of the moisture and thermal structure of the atmosphere are needed to infer this depth.

Mixed-Phase Conditions

The data used to construct the icing certification envelopes include little information about whether they were collected in purely liquid conditions or in mixed-phase (both liquid and ice) conditions. More recent data sets indicate that a large fraction of

icing encounters are in mixed-phase conditions and this appears to be the norm rather than the exception. This has implications for remote detection and forecasting. Mixed-phase conditions are usually thought to represent a transitory state as ice crystals will tend to grow at the liquid drops' expense. However, in a case with sufficient moisture supply and updraft speed, enough condensate can be produced for both deposition on ice and condensation on drops to occur.

It is not well understood whether the nature of the ice accretion process is different in liquid or mixed-phase cases. Limited data from research flights do not suggest any difference in effect on flight. Thermal deicing systems may be affected. Even so, environmental data collected in mixed-phase conditions are being analyzed to determine if there is an intrinsic difference between these and the current envelopes. The anticipation is that this characterization of the mixed-phase icing atmosphere will be used to create new certification envelopes for use in these conditions.

High Ice Water Content

Aircraft jet engines may suffer power loss from ingest of high amounts of ice crystals. From information on over 100 engine weather-related power loss events, it was concluded that these events were due to flight through areas of high ice water content associated with deep convective clouds. In addition to the obvious safety concerns, these events can also lead to costly engine repairs. Research has recently begun on characterizing the microphysical properties of these clouds, determining how they may be forecast, and investigating the physical process of accretion in the engine during flight. This information will be used to provide guidance to manufacturers and to develop a new certification rule for engine performance in high ice mass environments.

Detecting Icing Conditions

Pilots generally have a poor view of the aircraft's wings so they commonly use the ice accreting on windshields, wipers, or pitot tubes near the nose of the aircraft to assess the presence and amount of ice. The pilot can also notice changes in aircraft performance due to icing. Onboard icing detectors warn the pilot when ice is accreting on the aircraft. In some cases, these instruments are sensitive enough to provide an early warning before the ice becomes noticeable to the pilot. These airframe-mounted detectors are a fairly mature technology, although new systems are still being developed. Examples of detector types are those that can be flush-mounted on the wing and detect differences in capacitance on the surface or use a vibrating rod protruding into the airstream, which detects the difference in resonant frequency as ice accretes. The advantage of *in situ* systems is that they provide a positive detection of icing conditions. However, they have the drawback that the aircraft must necessarily be immersed in the icing environment, and in many cases that is not a desirable place to be.

The use of remote sensors for detecting icing remains in a relatively young stage of research and development.

Operational weather radar systems such as TDWR (Terminal Doppler Weather Radar: C-band, 5 cm) and NEXRAD (Next Generation Weather Radar: S-band, 10 cm) were not specifically designed for icing detection, but may provide information that, when combined with that from other sources such as numerical weather prediction models, satellite imagery, or surface observations, provides clues to the location and intensity of icing.

Short-wavelength radars (such as K-band, 0.86 cm or W-band, 3 mm) have potential for detecting icing conditions, especially in non-precipitating clouds. Dual- or triple-wavelength systems, using combinations of W-, K-, X-, and longer wavelengths, take advantage in the differences in attenuation of microwave radiation by atmospheric liquid at the different wavelengths.

Dual-polarization radars transmit radio wave pulses with both horizontal and vertical orientations. Comparisons of the orientation of the returned signals provide information on hydrometeor shape and that information can be used to infer whether the cloud contains ice crystals, drizzle, rain, or cloud drops. Used alone, this information can be somewhat ambiguous, especially in the case of mixed-phase clouds.

It has been noted that the regions of freezing drizzle tend to have a more uniform look or texture when viewed by radars. This information could be useful in discriminating between possibly hazardous freezing drizzle (SLD) and benign light snow.

Multichannel microwave radiometers, which passively detect radiation emitted from atmospheric constituents, have also been shown to be useful in identifying icing aloft. Their drawback is that they do not identify the altitudes at which icing exists and whether the detected liquid is supercooled. As with radar data, combining radiometer-based information with that from other instruments can help the forecaster gain insight into the nature of the icing environment.

Multispectral weather satellites have shown to be useful icing diagnosis tools. Algorithms have been developed which use combinations of wavelengths to determine locations of supercooled liquid cloud tops. These algorithms would not diagnose all icing conditions – for example, ice-bearing cirrus may overlie a supercooled liquid cloud and prevent its detection, or a supercooled liquid layer may be present above an all-ice cloud. Methods have also been devised to use microwave data from satellites to quantify the total integrated amount of atmospheric liquid water content over oceans. When combined with numerical weather prediction model outputs, a three-dimensional icing diagnosis can be provided.

Forecasting Icing Conditions

Forecasting inflight icing is the same as predicting supercooled liquid water in clouds – not exactly on the list for undergraduate weather forecast lab! Following a ‘forecast funnel’ process, the forecaster seeks

- Clouds or precipitation
- Right temperature regime ($<0^{\circ}\text{C}$, $\geq -20^{\circ}\text{C}$)
- Lift to create liquid
- Lacking significant ice to encourage glaciation

Understanding the processes that create and deplete liquid, in combination with information about where clouds are expected and the temperature structure of the atmosphere,

should provide a reliable and geographically robust prediction. These concepts can also be incorporated into automated systems, which provide the forecaster with initial guidance – or the non-meteorologist with a reasonable ‘final answer’ of where to expect icing.

With the advent of improved numerical weather prediction models with explicit cloud liquid, there is not only a chance to determine where icing really exists (as opposed to inferring it from smoothed temperature and humidity fields) but to quantify the hazard in terms of icing type and severity. Generally, microphysical parameterizations are first developed on research models. Concepts are then coded and tested for use in the operational versions deployed by the National Centers for Environmental Prediction (NCEP). In most operationally run models, the microphysical characteristics are heavily parameterized with idealized size distributions, and conversion of cloud liquid water to drizzle and rain is based on total liquid mass thresholds. New methods to explicitly represent physical processes with computational efficiency are being tested; the resulting size distributions of drops will increase the accuracy of forecasting the icing environment. Additionally, current operationally available models use spatial resolutions that do not resolve potentially significant fine-scale features. These larger scales (typically at least 20 km with hourly output) may work well for strategic preflight planning, but are not effective for tactical use such as ‘weather-in-the-cockpit’ displays. Considerable effort is being put into increasing resolution to 2–3 km, which will allow more realistic cloud physics treatments thus, presumably, leading to more precise icing forecasts.

No one observational tool or weather prediction model provides people with all they need to know about where icing is or is not located nor any of its attributes such as type or severity. Each information source provides one part of the picture. Forecasters combine this information to get the complete story on icing; it makes sense to develop automated algorithms to accomplish the same goal. Automated versions of this human technique have been developed and to date have proved quite successful in diagnosing where icing conditions reside, their expected severity, and the probability of encounter. The key to successful icing forecasting lies in understanding the physical processes resulting in supercooled liquid water production, how these processes relate to observable phenomena, and how to combine information from as many sources as possible to gain the most complete picture of the icing situation.

See also: Clouds and Fog: Cloud Microphysics; Cloud Modeling. Mesoscale Meteorology: Cloud and Precipitation Bands. Mountain Meteorology: Overview. Synoptic Meteorology: Forecasting. Weather Forecasting: Operational Meteorology.

Further Reading

Early Work on Inflight Icing

For many years, these works formed a basis for knowledge of the characteristics of the inflight icing environment. Lewis’ work especially formed the foundation for the icing ‘certification envelopes’ contained in the Federal Aviation Regulations Part 25, Appendix C, which are still used today.

Jones, A.R., Lewis, W., 1949. Recommended Values of Meteorological Factors to be Considered in the Design of Aircraft Ice Prevention Equipment, NACA Tech. Note 1855.

Lewis, W., 1951. "Meteorological Aspects of Aircraft Icing." Compendium of Meteorology. American Meteorological Society, Boston, MA.

Newton, D.W., 1978. An integrated approach to the problem of aircraft icing. *Journal of Aircraft* 15, 374–380.

Forecasting Methods

These publications are placed in numerical order and provide a brief historical record of the evolution of forecasting techniques for inflight icing conditions and for characteristics of cloud drops pertinent to forecasting. The Bernstein et al. (2005) paper describes an automated icing diagnosis algorithm and the science behind it; this algorithm is run by the National Weather Service and output is available to the public for flight planning.

Air Weather Service, 1980. Forecasters' guide on aircraft icing. Air Weather Service Report AWS/TR-80/001, 58 pp. [Available from U.S. Air Force, Scott AFB, IL 62225.]

Bernstein, B.C., McDonough, F., Politovich, M.K., Brown, B.G., Ratvasky, T.P., Miller, D.R., Wolff, C.A., Cuning, G., 2005. Current icing potential: algorithm description and comparison with aircraft observations. *Journal of Applied Meteorology* 44, 969–986.

Schultz, P., Politovich, M.K., 1992. Toward the improvement of aircraft icing forecasts for the continental United States. *Weather and Forecasting* 7, 491–500.

Thompson, G., Bullock, R., Lee, T.F., 1997. Using satellite data to reduce spatial extent of diagnosed icing. *Weather and Forecasting* 12 (1), 185–190.

Tafferner, A., Hauf, T., Leifeld, C., Hafner, T., Leykauf, H., Voigt, U., 2003. ADWICE: advanced diagnosis and warning system for aircraft icing environments. *Weather and Forecasting* 18, 184–203.

The Icing Environment

Characteristics of inflight icing environments are described in these papers, which cover varying geographic areas. *In situ* measurements using research aircraft and comparisons of locations of icing reported by pilots to numerical weather prediction model or observational data are used to provide the environmental information.

Bernstein, B.C., Omeron, T.A., McDonough, F., Politovich, M.K., 1997. The relationship between aircraft icing and synoptic-scale weather conditions. *Weather and Forecasting* 12, 742–762.

Cober, S.G., Isaac, G., Strapp, J.W., 1995. Aircraft icing measurements in east coast winter storms. *Journal of Applied Meteorology* 34, 88–100.

Jeck, R.K., 2008. Distance-scaled water concentrations versus mass-median drop size, temperature, and altitude in supercooled clouds. *Journal of Atmospheric Science* 65, 2087–2106.

Stankov, B.B., Bedard Jr., A.J., 1994. Remote sensing observations of winter aircraft icing conditions: a case study. *Journal of Aircraft* 31, 79–89.

Supercooled Large Drops

Supercooled large drops, those drops having diameters exceeding 50 μm , can be extremely hazardous for some types of aircraft. These papers describe the hazard, weather conditions conducive to large drop icing, and the weather associated with the accident. The accident described by Marwitz et al. (1995) was a turning point in icing research as it found supercooled large drop icing was a factor in the accident, which killed 68 people. This opened up a new area of research focused on characterizing the environment and initiated work on new regulations regarding certification of aircraft for flight into such conditions.

Cober, S.G., Isaac, G.A., Strapp, J.W., 2001. Characterizations of aircraft icing environments that include supercooled large drops. *Journal of Applied Meteorology* 40, 1984–2002.

Marwitz, J.D., Politovich, M.K., Bernstein, B.C., Ralph, F.M., Neiman, P.J., Ashenden, R., Bresch, J.F., 1995. Meteorological conditions associated with the ATR72 aircraft accident near Roselawn, Indiana, on 31 October 1994. *Bulletin of the American Meteorological Society* 78, 41–52.

Politovich, M.K., 1989. Aircraft icing caused by large supercooled droplets. *Journal of Applied Meteorology* 28, 856–868.

Ice Accretion and Effect on Flight

The physics of ice accretion on aircraft is both complex and fascinating. These papers summarize ice accretion and its subsequent effect on the aircraft's flight characteristics.

Hansman, R.J., 1985. Droplet size distribution effects on aircraft ice accretion. *Journal of Aircraft* 22, 503–508.

Lynch, F.T., Khodadoust, A., 2001. Effects of ice accretions on aircraft aerodynamics. *Progress in Aerospace Sciences* 37, 669–767.

Aviation Weather Hazards

AJ Bedard, Jr., National Oceanic and Atmospheric Administration, Boulder, CO, USA

© 2003 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 1, pp 166–177, © 2003, Elsevier Ltd.

Introduction

A broad range of aviation weather hazards affect operations from takeoff and landing to in-route flight at high altitudes. A goal here is to provide an overview of key meteorological processes disrupting flight, reducing lift, increasing drag, influencing instrument readings, or reducing visibility. These atmospheric factors include gravity currents including thunderstorm gust fronts and sea breeze fronts. Because the motions and decay of aircraft wake vortices are controlled by local winds, turbulence, and stability, these dangerous wake effects are also discussed. Hazard types covered appear in [Table 1](#), together with brief descriptions of their potential impacts on flight. [Figure 1](#) summarizes some of these meteorological disturbances, indicating typical flow strengths and altitudes affected. Encounters with turbulence aloft can disrupt flight paths and cause injuries to crews and passengers. Such strong encounters can result from organized instabilities of limited duration or extent, such as breaking gravity/shear waves. On the other hand, more random turbulence aloft, when of long duration and covering extended areas, can contribute to structural fatigue and reduce aircraft operating lifetimes.

For selected hazards, atmospheric causal processes are reviewed and key properties such as dimensions and strengths are summarized. Also, discussions of efforts at hazard prediction, detection, and warning illustrate the progress that has been made in mitigating atmospheric impacts on the aviation system.

Atmospheric Gravity Currents

One important type of atmospheric gravity current is the outflow from a distant thunderstorm downdraft. For these systems, because the downdraft region is usually quite large in

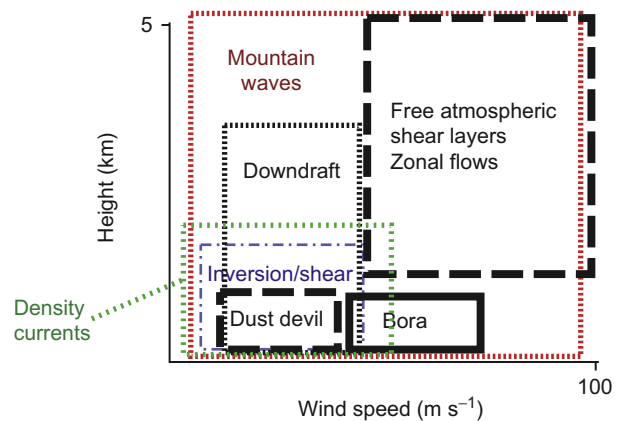


Figure 1 Summary of the altitude impact ranges of meteorological hazards and typical wind speed strengths involved.

diameter (typically 10 km or more), the wind shears near the downdraft can be relatively weak. However, the wind speed and wind direction change that accompany the leading edge can cause significant relative air speed changes for aircraft.

Sudden aircraft performance changes caused by the atmosphere (whether increasing or decreasing performance) are problematical. [Figure 2](#) is a conceptual view of a gust front from a distant thunderstorm crossing an airport. Although at times these boundaries are clearly visible because of entrained dust, more often the boundary will occur invisibly in clear air at large distances from the originating thunderstorm.

The speed of motion, c , of a gravity current with no ambient wind can be estimated from the density current equation [1].

$$c = Fr \left(\frac{\Delta T}{T} gh \right)^{1/2} \quad [1]$$

Table 1 A summary of aviation weather hazards and their areas of impact upon flight operations

Hazard	Areas of impact
Atmospheric gravity current wind shears (e.g., thunderstorm gust front and sea breeze front)	● Relative air speed changes ● Require runway changes
Microburst wind shears	● Flows can exceed performance capabilities of modern aircraft
Vertical wind shear	● Deviations from glide slopes
Gravity/shear waves	● Flight disruption and structural fatigue ● Increased drag and reduced lift
Icing	● Reduced stall angle
Terrain-induced disturbances (e.g., lee waves, rotors, bora)	● Flight disruption ● Deviations from assigned flight altitudes ● Structural damage
Vicinity of thunderstorms (e.g., hail, funnels, obstacle flows)	● Flows can exceed performance capabilities of modern aircraft ● Deviations from flight altitudes ● Structural damage
Aircraft wake vortices transported to unexpected locations	● Roll moments disrupting flight of following aircraft
Altimeter errors	● Deviations from assigned or expected flight altitudes

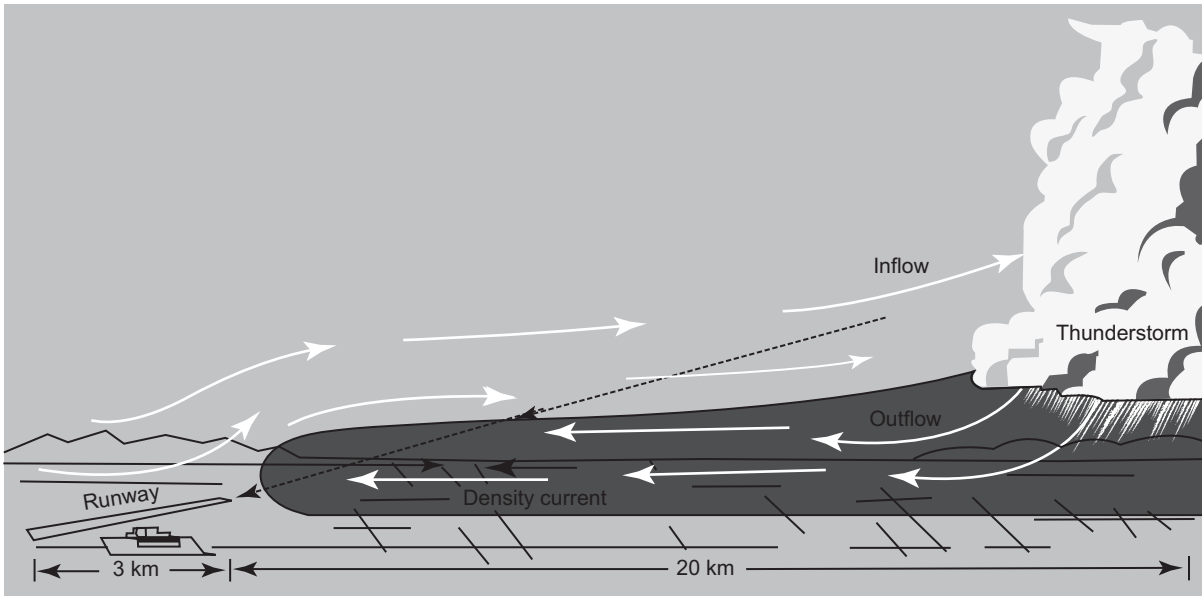


Figure 2 Conceptual view of a gust front from a distant thunderstorm crossing an airport.

Here, Fr is the Froude number (~ 1), ΔT is the temperature change in the gust front air relative to the environmental air, T is the mean temperature, g is local gravity, and h is the height of the outflow boundary. Corrections for the ambient wind can be made. [Table 2](#) summarizes statistics for gust fronts measured in the Denver, Colorado, region during an intensive field program during the summer of 1982. Data from this and other experiments demonstrate that the density current equation applies quite well. The leading edge is usually accompanied by a temperature drop and pressure rise, unless complicated by the existence of a ground based inversion layer. [Figure 3](#) shows an acoustic sounder display of a gust front propagating on top of a ground-based inversion.

Another important aspect of thunderstorm outflow boundaries is that a preferred region for the initiation of new convection is near the leading edges where vertical forcing of ambient air occurs. This is especially true where boundaries collide. Microbursts also have a tendency to occur near outflow boundary regions. Fortunately, Doppler radar can detect these gust front boundaries effectively and provide an invaluable,

allweather resource to guide airport operations when gravity currents are approaching.

Microbursts

Between 1964 and 1985, over 30 commercial aircraft crashes resulted from microbursts. A microburst is defined as a downdraft region with a scale size less than 4 km. The resulting strong outflows usually do not travel radially outward for long distances (>10 km), and the durations are short (often less than 10 min). However, in the vicinity of a microburst, strong winds ($> 50 \text{ m s}^{-1}$) and rapid wind direction changes of 180° can occur. When microbursts descend near or on runways, they constitute an extreme flight hazard. Microburst flows are analogous to those produced when squirting a water hose on a flat surface. The downflow jet interacts strongly with the surface, producing strong radially directed flows. The large spatially concentrated horizontal wind vector changes and the downdraft can produce increasing performance/decreasing performance couplets that are difficult to predict and handle. For example, an aircraft flying through a microburst that has impacted the approach end of a runway will first encounter a head wind, increasing performance and causing excursions above the glide slope. As the pilot corrects for this, the aircraft enters the downdraft region, followed by an outflow region, rapidly degrading performance. Depending upon the timing and relative positions of the aircraft and microburst to the runway, this scenario can be catastrophic. The timing is so critical that even landing differences of several minutes can be important. The statistics of microbursts measured in the vicinity of Denver, Colorado, are presented in [Table 3](#). Microbursts were identified by the winds clearly radiating outward from a center, as distinct from the essentially linear gust front winds. To date, the largest wind speed documented for

Table 2 Gust front statistics derived from surface meteorological stations in the Denver, Colorado region during the summer of 1982. Number of events = 99

Parameter	Average	Minimum value	Maximum value
Wind speed change	9.4 m s^{-1}	3.0	17.0
Wind vector change	13.7 m s^{-1}	5.0	37.0
Temperature change	-1.9°C	-5.1	+3.1 ^a
Pressure change	0.67 hPa	-0.4 ^a	3.3
Rain rate	24 mm h^{-1}	0	3.0

^aThese anomalous readings of temperature increase and pressure decrease resulted from the passage of a gust front on top of and eroding a ground-based inversion.

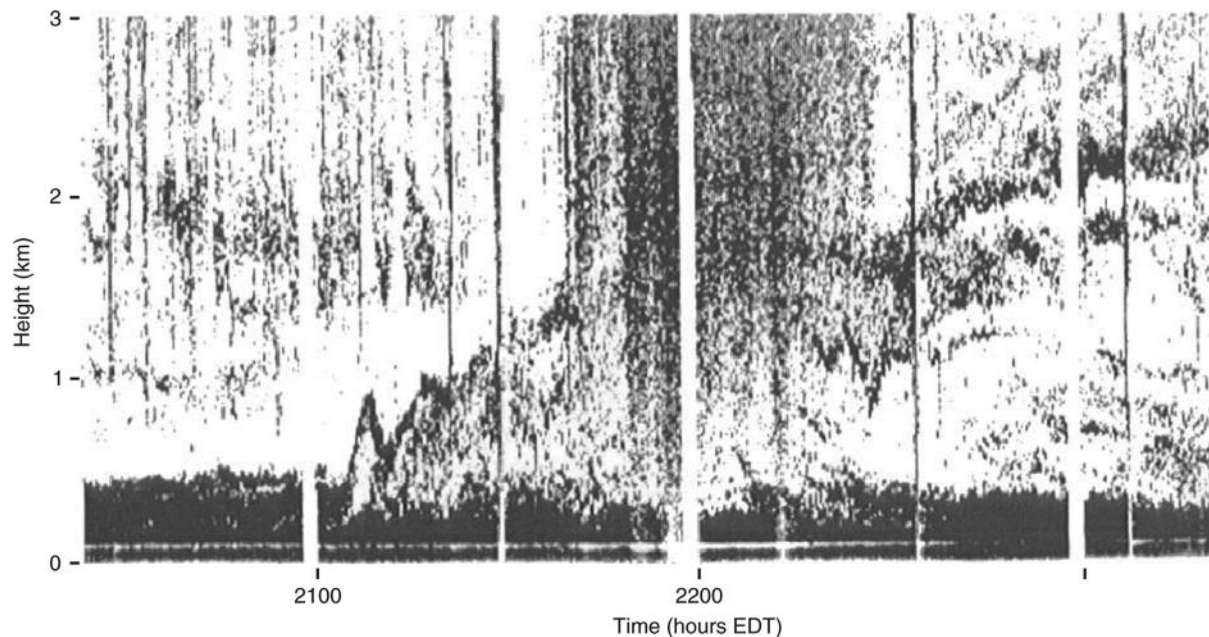


Figure 3 Acoustic sounder detection of a gust front propagating on a ground-based inversion.

a microburst was that related to a 'near miss' of Air Force One with President Reagan on board when it was on the ground near Washington, DC, on 1 August 1983 (a wind speed surge over 60 m s^{-1}). The microburst occurred five minutes after the plane landed. Newspaper accounts said that a secret service officer jumped on top of the president to protect him as the winds buffeted the aircraft.

There are two extreme types of microbursts: 'dry' and 'wet'. Dry microbursts are especially hazardous because the visible virga (raindrops or a snow plume descending from cloud base as in the photograph in Figure 4) related to the microburst initiation process evaporates and becomes invisible as it approaches the surface. Since the downdraft descent time takes about 5 min, it can be difficult to relate an observation of virga to a resulting microburst. Conversely, a wet microburst has a strong rain shaft and is easily seen if not obscured by rain from a surrounding storm (Figure 5).

Doppler radar can detect microbursts once the radial outflow is established by the intense downdraft penetrating to the surface. Also, the concentrated rain shafts for wet microbursts can be detected. The uses of arrays of airport wind sensors, Doppler radars, and improved controller/pilot training

(both to recognize visual clues and to respond in the best possible way if a microburst is encountered) have helped to reduce microburst-related accidents. Also, the fact that the lapse rate between 500 and 700 hPa is correlated with microburst probability provides forecasting potential for dry microburst likelihood. Dry microbursts are more probable when the lapse rate is $> 8^\circ\text{C/km}$.

Vertical Wind Shear and Gravity/Shear Waves

Whereas the thunderstorm gust front and microburst hazards result primarily from horizontal changes in wind speed, vertical changes in wind speed and direction can also present a hazard, especially for lower-level flight operations. A ground-based inversion is often accompanied by calm winds near the surface and strong winds just above the cooler, stable near-surface air. Aircraft descending or ascending through such layers can encounter strong wind shear-produced performance changes and turbulence, and rapid fluctuations associated with gravity/shear waves. These waves have scale sizes from tens to hundreds of meters, resulting in aircraft interaction times of seconds or less. Figure 6 is a conceptual view of such a situation in the vicinity of mountains. Figure 7 is a Doppler lidar display of gravity/shear waves. Vertical wind shear conditions can be especially important for general aviation airports if relative airspeed is suddenly reduced on a low-level approach or during takeoff in the vicinity of terrain. At airports where wind shear above stable air is a frequent problem, boundary layer wind profilers or acoustic sounders can provide valuable real-time monitoring capabilities.

The presence of gravity/shear wave activity often complicates flight through layers of vertical wind shear. A pure shear layer in a neutrally buoyant atmosphere may be modeled as a vortex sheet, highly unstable to disturbances. If wind shear

Table 3 Impacting microburst statistics derived from surface meteorological stations in the Denver, Colorado region during the summer of 1982. Number of events = 33

Parameter	Average	Minimum value	Maximum value
Wind speed change	13.5 m s^{-1}	2.5	27.5
Wind vector change	20.7 m s^{-1}	10.0	37.5
Temperature change	-1.5°C	-9.0	$+5^a$
Pressure change	0.66 hPa	-1.5^a	2.0
Dew point change		-7°C	$+7^\circ\text{C}$
Rain rate	16.4 mm h^{-1}	0	2.75



Figure 4 Photograph showing virga descending from cloud base.



Figure 5 Photograph of a 'wet' downdraft.

occurs in conjunction with a stable layer, gravity provides a restoring force, and such a system will support wave motion. Hence, the term gravity shear wave. Several questions naturally follow from this situation.

- Under what conditions will the upper-level wind shear start to erode the stable air and turbulence grow?
- At what rates do such stable pools of air get removed?
- What are some examples of situations where these processes are important for flight operations?

The Froude number, Fr , is a measure of the relative importance of inertial and gravity forces. For the situation of flow above a stable layer, Fr is given by eqn [2], where U is the flow speed above the inversion, g is local gravity, h is the height

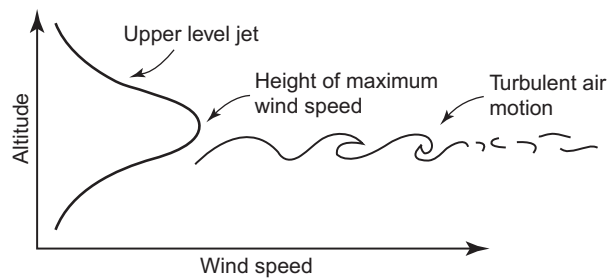


Figure 6 Conceptual view of gravity/shear waves in the vicinity of mountains.

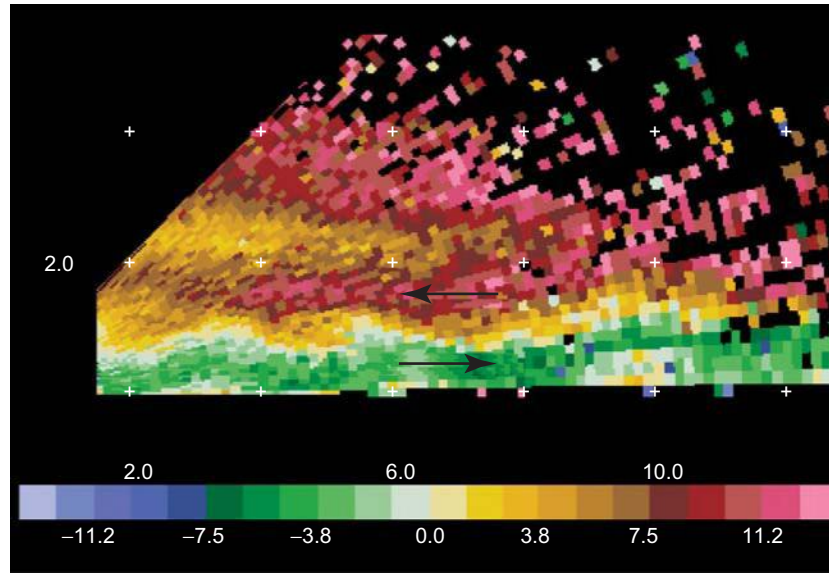


Figure 7 Doppler lidar display of gravity/shear waves in the vicinity of mountains, showing the accompanying wind speed changes. The numbers below the color bar are the radial wind speeds in meters per second. The numbers above the color bar are the distance in kilometers from the lidar. Courtesy L. Darby, NOAA.

of the inversion, $d\rho$ is the density difference between the two layers, and ρ is the mean density.

$$Fr = \frac{U}{\left(d\rho \frac{gh}{\rho}\right)^{1/2}} \quad [2]$$

In field and laboratory experiments, the start of disturbances and waves takes place when the Froude number exceeds about 0.6. Thus, if the height of the inversion and a temperature profile are available, the threshold speed U can be estimated. Once the erosion process starts, it can often continue at a slow and approximately constant rate. Values of vertical erosion rates near 10 cm s^{-1} have been measured near complex terrain.

On the positive side, the time scales are of the order of hours for changes, in contrast with minutes for microbursts. Thus, vertical wind profilers for monitoring and knowledge of local climatologies can be quite valuable, particularly for mountain valleys and the lee sides of complex terrain.

Another important dimensionless number is the Richardson number, Ri , which is an important index for turbulence. This number depends upon the gradients of both temperature and wind speed (eqn [3]).

$$Ri = \frac{g}{\theta} \frac{d\theta}{dz} \bigg/ \left(\frac{dU}{dz} \right)^2 \quad [3]$$



Figure 8 Photograph showing gravity/shear waves visualized by cloud formations.

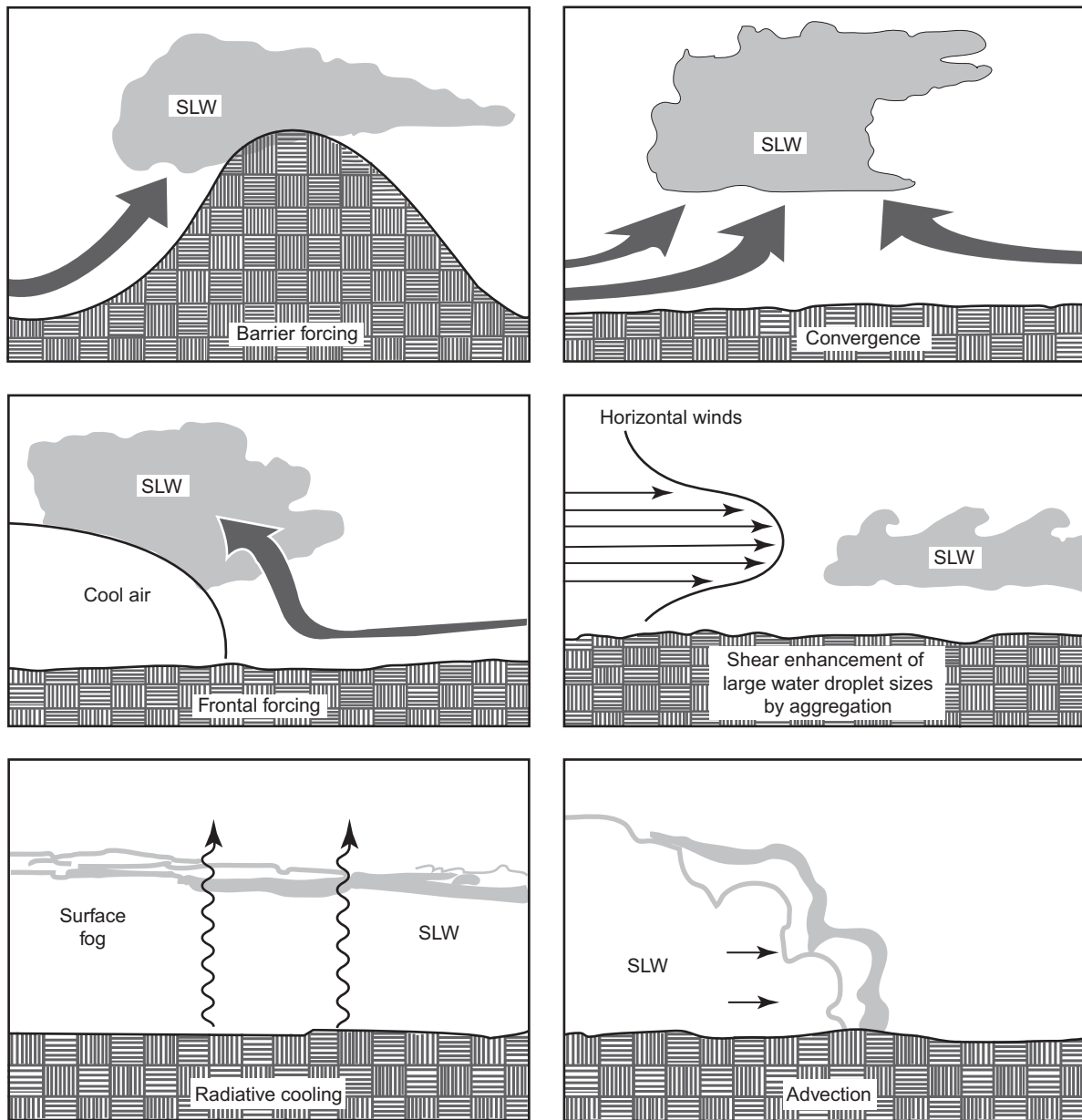


Figure 9 A summary of situations contributing to aircraft icing events. SLW stands for supercooled liquid water.

In eqn [3], g is local gravity, θ is the mean potential temperature, $d\theta/dZ$ is the change in potential temperature with height, and dU/dZ is the change of wind speed with height. The criterion for turbulence, $Ri < 1/4$, has been shown to be a valuable index for aircraft turbulence when temperature and wind speed profiles are available to make the comparison between predicted turbulent altitudes and actual turbulence reports from pilots. A challenge is either to predict the temperature and wind speed gradients or to measure them with sufficient accuracy to produce reasonable estimates of Ri values. Wind profiling radars used in conjunction with the Radio Acoustic Sounding System (RASS) to obtain temperature profiles will be valuable for such applications. Pilot reports as well as the use of visual clues are also an

invaluable component for avoiding regions of turbulence aloft. Figure 8 depicts gravity/shear waves that are revealed by cloud formations.

Large-amplitude, long-lived gravity waves represent another aspect of the hazard. Such waves have been observed to propagate rapidly (at $35\text{--}40\text{ m s}^{-1}$) away from the region of a cyclone where they were generated. The waves in this case traveled through eastern New England in the United States. The high-speed waves were accompanied by precipitation and wind surges. For example, at Boston the wind increased from less than 10 knots to 57 knots over about 5 min. Such discontinuities can represent significant hazards, since they are unexpected weather features producing wind vector changes on scales that can affect flight, especially during takeoff and landing.

Icing

Icing potential depends upon the probability of drops of supercooled liquid impacting aircraft surfaces. Larger droplets are more likely to strike an airfoil, since they do not easily follow the flow streamlines and pass around an obstacle as do smaller droplets. The icing hazard can be insidious because of two factors:

1. Only a small amount of ice deposition can have large, deleterious effects upon lift and drag, thus reducing aircraft performance.
2. Icing and the degradation of performance can increase slowly and imperceptibly, until an emergency exists.

Two situations account for most reported in-flight icing encounters. Convective activity involves relatively large amounts of supercooled liquid water, and the hazard can extend to higher altitudes with the potential of significant ice accumulation in a short time.

The other frequent situation involves flight through layered cloud decks. Although the supercooled liquid content can be lower than in convective situations, aircraft typically spend more time in these layered clouds, accumulating significant ice on fuselage and control surfaces. [Figure 9](#) depicts situations contributing to this form of icing, which is especially important for commuter and general aviation aircraft operating at lower altitudes. The wind shears and turbulence often accompanying such systems can be regions where larger drops are concentrated by mixing. Sometimes, by making small altitude changes, a plane can avoid these regions of enhanced icing potential.

Combinations of remote sensors have value for the monitoring and forecasting of icing situations. Polarimetric radars, dual-wavelength radars, wind profilers, RASS, and passive

radiometric measurements can, in concert, indicate the locations of supercooled liquid. Aircraft can also accumulate icing on the ground in freezing rainstorms. This problem is treated at major airports by application of de-icing fluids before takeoff (see [Aviation Meteorology: Aircraft Icing](#)).

Terrain-Induced Turbulence

Terrain effects can occur at all flight levels, with some disturbances affecting the stratosphere. The flow situations can range from lee waves, bora flows (a form of density current), and rotors, to mechanically induced turbulence. [Figure 10](#) is a Doppler lidar display showing the roll-up of a vortex sheet in the Colorado Springs area. At times, organized instabilities can occur in the forms of vertical or horizontal axis vortices. These obstacle-involved situations can be exceptionally complex when the terrain flows interact with other meteorological factors (such as lee-side inversions). Since 1964, there have been 15 major accidents and incidents in the vicinity of complex terrain ([Table 4](#)).

One study indicated that the general aviation accident rate was 40% higher for US mountain states than for all other continental states, and the rate was 150% higher for a selected group of mountain airports relative to a group of nonmountain airports.

[Table 4](#) indicates a pattern of sporadic encounters of aircraft with severe or extreme turbulence in the vicinity of mountains. In many cases, aircraft preceding or following the aircraft involved in the event encountered some turbulence, but not the extreme turbulence of the encounters (which often exceeded structural limits). Thus, the regions of severe or extreme turbulence may, at times, be spatially concentrated and short-lived. This makes predicting the time and location of these events more difficult. There is a great need to define the properties of mountain-related hazards, improve short-term

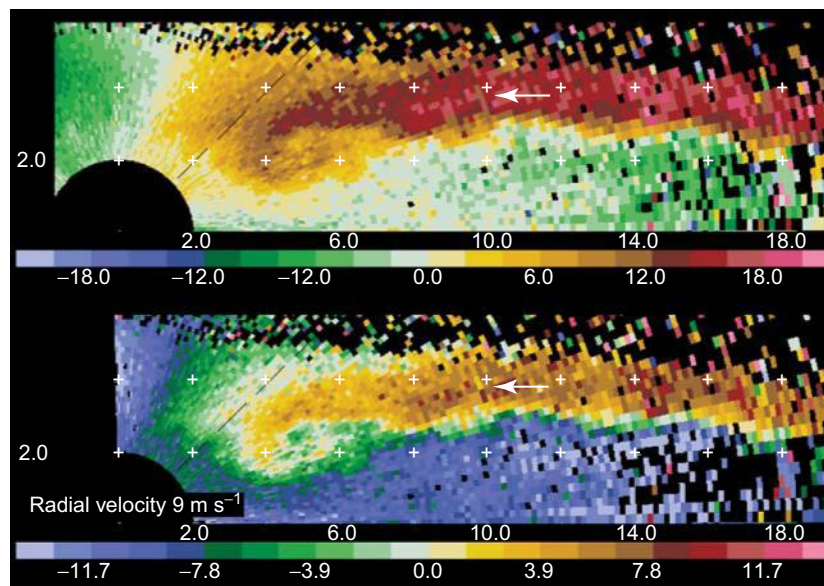


Figure 10 Doppler lidar display showing the rollup of a vortex sheet in the lee of a mountain range. The numbers below the color bar are the radial wind speeds in meters per second. The numbers above the color bar are the distance in kilometers from the lidar. Courtesy L. Darby, NOAA.

Table 4 Turbulence-related accidents and incidents occurring in the vicinity of mountains

Event	Date	Location	Comment
Accident	31 Mar. 1993	Anchorage, Alaska	Turbulence, 747 lost engine
Accident	22 Dec. 1992	West of Denver, Colorado	Loss of wing section and tail assembly, 2-engine cargo plane, lee waves present
Accident	9 Dec. 1992	West of Denver, Colorado	DC-8 cargo plane, loss of engine, lee waves present
Unknown cause accident	3 Mar. 1991	Colorado Springs, Colorado	737 crash
Accident	12 Apr. 1990	Vacroy Island, Norway	DC-6 Crash
Severe turbulence	24 Mar. 1988	Cimarron, New Mexico	767, 1.7G, Mountain wave
Severe turbulence	22 Jan. 1985	Over Greenland	747, +2.7G
Severe turbulence	24 Jan. 1984	West of Boulder, Colorado	Saberliner, +0.4G to -0.4G
Severe turbulence	16 Jul. 1982	Norton, Wyoming	DC-10, +1.6G to -0.6G
Severe turbulence	3 Nov. 1975	Calgary, Canada	DC-10, +1.6G
Accident	2 Dec. 1968	Pedro Bay, AK	Fairchild F27B, wind rotor suspected
Accident	6 Aug. 1966	Falls City, Nebraska	BAC 111, wind rotor suspected
Accident	5 Mar. 1966	Near Mt. Fuji, Japan	BOAC 707, wind rotor suspected
Accident	1 Mar. 1964	Near Lake Tahoe, Utah	Paradise Air Constellation, strong lee wave
Accident	10 Jan. 1964	East of Sangre de Cristo mountains in New Mexico	B52, wind rotor suspected

forecasting of these events, improve pilot training resources, and develop detection methods.

Major field experiments have addressed this problem, which requires three-dimensional sampling of large volumes of the atmosphere as a function of time, documenting both surface and near-surface effects, as well as upper tropospheric and stratospheric effects. Both physical scale models and numerical models have guided the execution of field programs studying mountain flows, which require the commitment of considerable scientific and measurement resources. Figure 11 shows an example of the changes in flows encountered by a research aircraft traversing a lee wave in the Rocky Mountain region of the US. Similar strong changes were also encountered in temperature, pressure, and vertical wind speed, with great

differences as a function of horizontal distance and flight level. Doppler lidars have the ability to detect these hazards near the surface in clear air. Figure 12 is a Doppler lidar display illustrating the complexity possible for such upper-level mountain flows.

Thunderstorms

The most extensive study of thunderstorms was the Thunderstorm Project. This study included surface meteorological stations, soundings, radar, and numerous aircraft flights. This study not only treated the larger-scale thunderstorm structure, circulation, and surface effects, but also properties important for

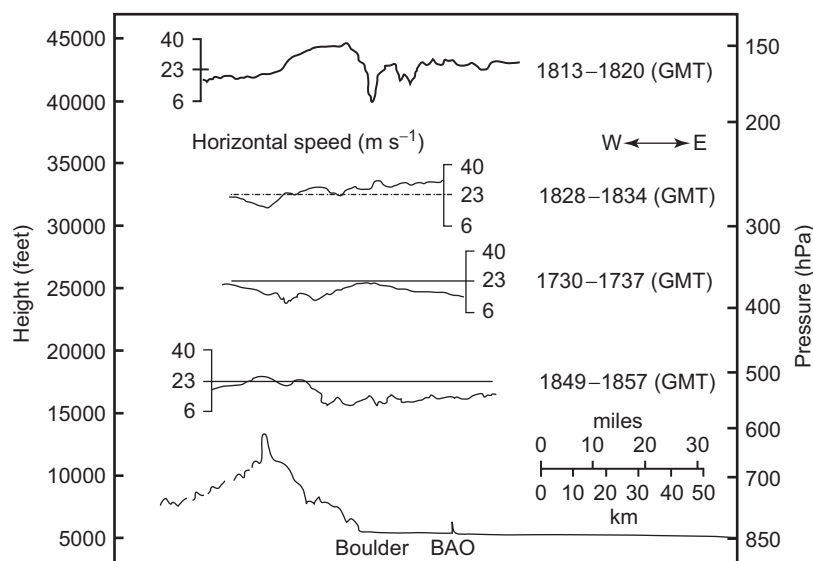


Figure 11 Horizontal wind speed changes encountered by a research aircraft in a lee wave situation on 25 January 1984. Data from flight legs at four altitudes are shown.

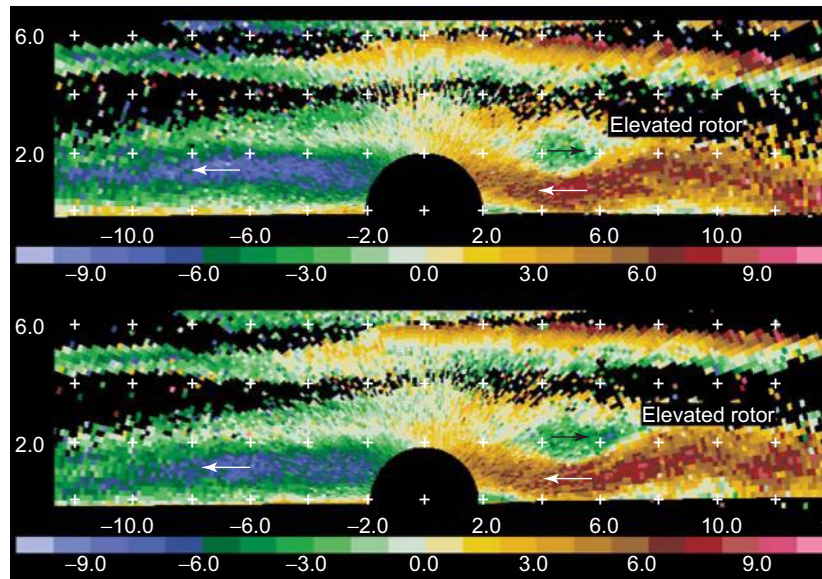


Figure 12 Doppler lidar display showing rotor circulation associated with lee waves. The numbers below the color bar are the radial wind speeds in meters per second. The numbers above the color bar are the distance in kilometers from the lidar. Courtesy L. Darby, NOAA.

flight. These thunderstorm effects influencing aircraft included turbulence, hydrometeors, downdrafts, and updrafts, providing statistics for these thunderstorm parameters. The report also discussed the thunderstorm structure as detected by radar and geographical/topographical effects on thunderstorm development. This report by Byers and Braham is an important resource addressing all aspects of thunderstorms (see [Further Reading](#)).

Radar guidance permits aircraft to navigate gaps between growing thunderstorm cells, and it is presumed here that pilots will avoid penetrating such violent severe weather directly. On the other hand, the clear air around cells or near benign-looking flattened anvil clouds often appears safe for flight, but these regions should be approached with caution if they cannot be avoided. [Table 5](#) outlines these regions near thunderstorms and some of the potential hazards.

Aircraft Wake Vortices

The strong vortex pairs behind heavy aircraft flying at low speeds result from the lift being generated. The tangential winds near the cores can be greater than 100 m s^{-1} (measured

at over 130 m s^{-1} for a Boeing 757). In an ideal atmosphere (neutrally stable, with no winds or turbulence), these vortices move downward and then outward, diverging in a predictable way as they approach the surface of the Earth.

However, with light wind speeds that are close to the transport speeds of the vortices, complex vortex trajectories can result. [Figure 13](#), from numerical simulations, shows how vortex transport can be modified by ambient winds and for some situations can remain stationary over runways. Such motions have been documented in field experiments. In addition, atmospheric stability also affects transport, perhaps causing the vortex pair to remain at a higher altitude than expected. The level of atmospheric turbulence is an important factor in controlling vortex decay times. Although generated by aircraft rather than being a true atmospheric meteorological phenomenon, it is the variability of the atmosphere that magnifies the hazard potential.

Concern about the wake vortex hazard to following aircraft is a limiting factor for the most efficient use of airport operations. This is because safety is currently ensured through the use of increased spacing between heavy aircraft and following aircraft. In spite of this, between 1983 and April 2002 there

Table 5 Atmospheric hazards near thunderstorms

Potentially hazardous region	Comment
1. Clear air regions to the lee of thunderstorms acting as obstacles to ambient winds	Aircraft upsets have been caused by lee waves around thunderstorms
2. Funnels may occur in the vicinities of anvil regions	There have been funnel-related aircraft accidents
3. Near or in the tops of weak-looking convective cells when there is wind shear present	These situations can strengthen regions of concentrated vorticity
4. Lightning strikes in and near thunderstorms	Frequent occurrence; damage can affect flight performance and cockpit instruments. Composite material airframe structures are at greater risk
5. The clear air above thunderstorm tops is a location for sprites, elves, blue jets and other electromagnetic phenomena	The hazard potential is not known. This may become a factor for aircraft manufactured from composite materials

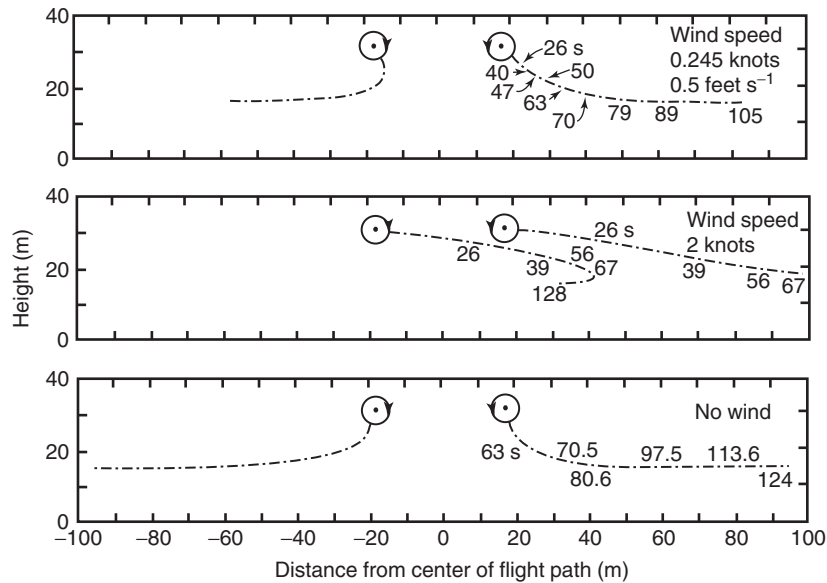


Figure 13 Computations of aircraft wake vortex transport affected by atmospheric winds.

were nine documented accidents with wake vortices as the cause. Although most of these accidents involved general aviation aircraft encountering the wakes of heavy commercial aircraft during approach or departure, some involved in-route wake upsets between two heavy aircraft. For example, on 2 September 1999 a Boeing 737 out of Santa Barbara, California flew through severe turbulence from the wake of an MD-11, with 15 injuries. This occurred in spite of a separation of 11 miles. In the future, the use of remote sensing technologies to track the positions of vortices could permit more effective use of runways. Also, boundary layer wind profilers and RASS could provide detailed real-time wind and temperature data as inputs to vortex transport prediction models.

Altimeter Errors

There are several critical ways in which altimeter errors can affect the aviation system. Pilots may be at any of three distinct altitudes:

- Where they are assigned
- Where they think they are
- Where they are

Several factors can contribute to these three places not being the same. There could be errors in calibrating aircraft altimeters. There could be malfunctions of aircraft instruments (e.g., clogged pressure ports).

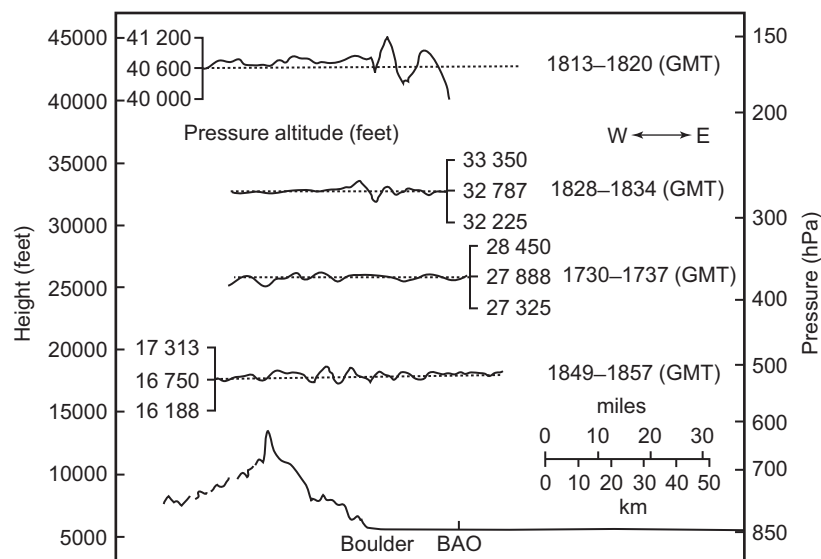


Figure 14 Altitude changes encountered by a research aircraft traversing a lee wave on 25 January 1984. Data from flight legs at four altitudes are shown. These measurements are for the same flight shown in [Figure 11](#).

Table 6 Remaining areas of special need

- Improved knowledge of terrain-induced disturbances and the development of mitigation techniques
- Improved warnings for ‘free air’ turbulence, especially better knowledge of risks in the vicinity of thunderstorms
- Earlier detection techniques for dry microbursts and increased understanding concerning their initiation processes
- Improved warnings of icing situations
- Better predictions of the transport and decay of aircraft wake vortices
- Broader application of site-specific remote sensing technologies for hazard prediction and monitoring applications

Also, the surfaces of constant pressure could be highly perturbed (e.g., by strong lee wave conditions or in the vicinities of frontal boundaries). Atmospheric winds around buildings housing altimeters can cause significant differences between true static atmospheric pressure and a measurement contaminated by dynamic pressure effects. For example, winds between 10 and 25 m s⁻¹ can induce dynamic pressure errors as large as about 2 hPa at an airport barometer (equivalent to altitude errors between 3 and 20 m). The development of new static pressure probe designs and improved calibration procedures is helping with these problems, which can be critical for mountain flying situations.

In the future, deviations from assigned flight altitudes in the vicinity of pressure surface perturbations will become more of a concern. As air traffic continues to increase, so will the motivation to allow smaller vertical separations of flight altitudes. There is a need to document the statistics of these excursions as well as the conditions under which they occur. The significant altitude changes involved with an aircraft traversing a lee wave on 25 January 1984 provide an example of the importance of such encounters (Figure 14).

Concluding Remarks

The early literature of aeronautical meteorology contains vivid accounts of the frightening experiences of balloonists within thunderstorms. Fortunately, the evolution of modern flight has made modern aircraft largely immune to such experiences, but they are susceptible to other hazards, such as microbursts and wake vortices (in part because of our technological advances). On the other hand, modern lighter-than-air vehicles, both

tethered and untethered, can find themselves similarly challenged because of limited agility. The future uses of lighter-than-air craft may be expanded as our ability to make reliable long-range forecasts improves. Table 6 outlines challenges we still face in continuing to mitigate aviation weather hazards. The application of remote sensing technologies, better knowledge of atmospheric dynamics, improved numerical models, and accurate longer-term forecasts will be critical to success.

Acknowledgments

I am grateful to L. Darby of the Optical Remote Sensing Division of the Environmental Technology Laboratory of the National Oceanic and Atmospheric Administration for making available the Doppler lidar images used in Figures 7, 10 and 12. I am also grateful for the valuable suggestions of an anonymous reviewer. G. Salottolo of the US National Transportation Board provided information about aircraft accidents attributed to wake vortices.

See also: **Aviation Meteorology:** Aircraft Icing. **Gravity Waves:** Overview. **Lidar:** Doppler. **Mesoscale Meteorology:** Gust Fronts; Microbursts. **Mountain Meteorology:** Lee Waves and Mountain Waves. **Radar:** Polarimetric Doppler Weather Radar. **Turbulence and Mixing:** Overview.

Further Reading

- Byers, H.R., Braham Jr., R.R., 1949. The Thunderstorm: Report on the Thunderstorm Project. United State Printing Office, Washington, DC.
- Bosart, L.F., Bracken, W.E., Seimon, A., 1998. A study of cyclone mesoscale structure with emphasis on a large-amplitude inertia-gravity wave. *Monthly Weather Review* 126, 1497–1527.
- Fujita, T.T., Caracera, F., 1977. An analysis of three weather related aircraft accidents. *Bulletin of the American Meteorological Society* 58, 1164–1181.
- Fujita, T.T., 1985. The Downburst. The University of Chicago Press, Chicago.
- Gregg, W.R., 1930. *Aeronautical Meteorology*. The Ronald Press, New York.
- Horne, T.A., 1999. *Flying America's Weather*. Aviation Supplies and Academics, Newcastle, WA.
- Lenschow, D.H. (Ed.), 1986. *Probing the Atmospheric Boundary Layer*. American Meteorological Society, Boston, MA.
- Scorer, R.S., 1978. *Clouds of the World*. Lothian Publishing Company, Melbourne.
- Scorer, R.S., 1978. *Environmental Aerodynamics*. Ellis Horwood, Chichester.

Clear Air Turbulence

GP Ellrod (Retired), National Oceanographic and Atmospheric Administration/National Environmental Satellite, Data, and Information Service, Granby, CT, USA

JA Knox, University of Georgia, Athens, GA, USA

PF Lester, San Jose State University, San Jose, CA, USA

LJ Ehernberger (Retired), National Aeronautics and Space Administration, Dryden Flight Research Center, Edwards, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

The development of the scientific understanding of clear air turbulence (CAT) as an aviation hazard is described in this article. Topics range from the discovery of CAT during World War II to modern techniques to observe and predict conditions conducive to CAT formation. Important physical processes leading to Kelvin–Helmholtz instability (KHI), a primary producer of CAT, are explained. The frequency of occurrence of KHI is shown to be a maximum near synoptic scale upper level frontal zones near jet streams, with mountain waves, and above the tops of severe thunderstorms. Several current physical and statistical approaches to CAT prediction based on this knowledge are described. Finally, climatologies derived from both aircraft observations of CAT and numerical indices correlated with CAT are used to illustrate the global and seasonal distribution of CAT potential.

Introduction

Since the first aircraft flight, pilots have been aware of in-flight turbulence. Because the known turbulence of the time was tied to strong, low-level winds, rough terrain, and convection, some of the early pilots predicted that, with the exception of thunderstorms, the ability to attain higher flight altitude with pressurized cabins would be accompanied by a marked decrease in turbulence. This was not to be. In the 1940s, as fighter aircraft attained tropopause altitudes, they experienced a previously unknown phenomenon: clear air turbulence (CAT), so-called because initial encounters occurred in areas devoid of clouds. As aircraft were designed to fly higher and faster during the last half of the twentieth century, CAT became the focus of many organized research efforts. The knowledge of CAT has grown substantially as a result.

Traditional aircraft turbulence intensity scaling (including CAT) differs from turbulence metrics used in fluids, as they are usually conceived for other applications. Aircraft turbulence is defined in terms of the aircraft response magnitudes scaled to subjective intensity descriptors as perceived by pilots and airplane passengers. Classical turbulence is defined for most applications in terms of the state of the fluid eddy dissipation rate (EDR) or the ratio of turbulent flow speed fluctuations to the fluid flow velocity. Simply put, aircraft turbulence is ‘bumpiness in flight.’ The difference is critical in that bumpiness depends on aircraft design, weight, speed, and pilot input, in addition to the state of the atmosphere (gusts, wind shears, or waves), which may or may not be turbulent in the classical sense.

CAT is now defined as aircraft turbulence that occurs at altitudes of 5.6 km (atmospheric pressure of about 500 hPa) or higher, either in cloud-free conditions or within stratiform clouds. The critical influence of CAT is on flight safety. One of the oldest schemes for the characterization of turbulence intensity is based on deviations in vertical acceleration from the normal acceleration of gravity (1g). These categories and their threshold deviations are light, moderate, severe, and extreme (Table 1).

The physical impact of CAT on crew and passengers varies from discomfort for the lighter turbulence categories to loss of

flight control during the rare extreme turbulence event. In the most intense episodes, injuries, and in some very rare cases, fatalities have occurred. Unrestrained crew and passengers are especially vulnerable. Flight through turbulent conditions also produces stresses on the airframe. Repeated turbulence encounters over the lifetime of the aircraft may lead to metal fatigue and, in extremely rare cases, structural failure.

Causative Mechanisms and Characteristics

The discussion that follows considers atmospheric phenomena that contribute to CAT on a wide range of scales. This section examines briefly the primary physical characteristics and causes of microscale CAT. It then considers macroscale and mesoscale forcing mechanisms that create a favorable environment for CAT development.

Kelvin–Helmholtz Waves and CAT

The understanding of the production of CAT and its characteristics is rooted in theoretical studies of fluid mechanics, laboratory and numerical experiments, and field studies. The evidence from these investigations is that Kelvin–Helmholtz instability (KHI) episodes are the cause of a large fraction of CAT. KHI produces shearing gravity waves with typical horizontal wavelengths of a few tens of meters to a few kilometers, precisely the range of eddy sizes to which most aircrafts will have the maximum response. KHI arises from micro- and mesoscale wind shear intensification when smooth, wavelike oscillations within a sheared, statically stable layer grows in amplitude to the point where the wave crests overturn, or ‘break.’ Wave breaking at wavelengths of hundreds of meters is highly effective in producing CAT, with a rapid cascade of energy from the KHI to smaller scale turbulence and dissipation. With respect to CAT, some of the more important characteristics of KHI are the following:

1. The typical lifetime of an individual KHI is about 5 min.

Table 1 Turbulence categories in terms of vertical acceleration, aircraft response, passenger experience, and EDRs for large aircraft

Turbulence category	Aircraft vertical acceleration magnitude (deviations from g , $m\ s^{-2}$)	Aircraft response	Passenger experience	Approximate (EDR) ^{1/3} ($m^{2/3}\ s^{-1}$) for B737, B757 aircraft
Light	0.2–0.5	Momentarily causes slight, erratic changes in altitude and/or attitude.	A slight strain against seat belts. Unsecured objects may be displaced slightly. Walking possible with little difficulty.	0.1–0.3
Moderate	0.5–1.0	Changes in altitude, attitude, and/or airspeed occur.	Definite strain against seat belts. Unsecured objects are dislodged. Walking is difficult.	0.3–0.5
Severe	1.0–2.0	Large, abrupt changes in altitude, attitude, and/or airspeed. Aircraft may be momentarily out of control.	Occupants are forced violently against seat belts. Unsecured objects are tossed about. Walking is impossible.	0.5–0.7
Extreme	>2.0	The aircraft is violently tossed about and is practically impossible to control. May cause structural damage.	Truly frightening.	>0.7

Reproduced from Lane, T.P., Sharman, R.D., Trier, S.B., Fovell, R.G., Williams, J.K., 2012. Recent advances in the understanding of near-cloud turbulence. *Bulletin of the American Meteorological Society*, 93, 499–515.

- The length of the dominant wave (most unstable KHI mode) is proportional to the depth of the sheared layer (i.e., about six times the depth). However, as KHI-induced turbulence and mixing modify the local wind shear structure and stability stratification, variations in the KHI wavelengths can be expected.
- The intensity of the turbulence produced in individual KHI eddies is proportional to the peak local wind shear generated as the wave amplifies and breaks. If the turbulent mixing caused by KHI sufficiently weakens the background wind shear, the turbulence will decay and the flow will again become laminar. However, if turbulent mixing strengthens the wind shears near the boundaries of the old turbulence layer, then new KHI may develop.

Another relevant result from studies of KHI has been the development of a basic dynamic instability principle for a simple model of shearing gravity waves. Based on linear theory, the Miles–Howard criterion states that unstable wave modes resulting from vertical shear are likely to occur when the local gradient Richardson number (Ri) is less than 0.25. In contrast, if (Ri) becomes 1.0 or more, and KHI is present, they will decrease in amplitude. The use of (Ri) as a practical index for CAT diagnosis and prediction is discussed in a later section.

Although the mesoscale sheared layers in which shorter wavelength KHI occur are, typically, less than 1 km in depth, horizontal dimensions are much larger (i.e., from ~ 10 km to a few hundred kilometers, often elongated in the direction of the wind). Thus, an individual KHI develops over a considerably smaller horizontal scale than the sheared stable layers in which it is embedded. This observation has four important ramifications:

- An aircraft flying in a sheared layer may experience CAT over a distance much greater than the scale of an individual KHI.
- CAT encounters are often burstlike (highly intermittent) as an aircraft crosses a thin sheared layer or as an aircraft flying within a sheared layer intercepts individual KHI elements in different stages of development.

- Regular, comprehensive observations of CAT are difficult to acquire because of its small scale and intermittent nature.
- Conditions conducive to the development of CAT are rooted in larger scale processes that produce and perturb the extensive, sheared stable layers.

Internal Gravity Waves (IGW) and CAT

IGW with horizontal scales of a few kilometers to a few hundred kilometers are also mechanisms for CAT production. IGW may become significant in the production of CAT in several ways. Wave amplitude variations with height are caused by the decrease in density with altitude and by the variation of background stability and wind with height. Aircraft intersecting large-amplitude IGW may be exposed to ‘sharp-edged’ gusts or periodic vertical motions which may be interpreted as CAT. Vertical displacements due to gravity wave motions will modulate background wind shear, leading to the production of microscale KHI. Wave breaking may produce intense turbulence near or somewhat below (~ 1 – 2 km) the ‘critical layer,’ where phase speed and background wind speed are the same. Finally, the occurrence and intensity of CAT are also affected by the excitation of IGW from different sources, as well as by resonant nonlinear interactions between different IGW modes and between IGW and KHI in turbulent layers (see **Dynamical Meteorology: Overview**. **Gravity Waves: Overview**).

Macroscale Forcing

Any atmospheric circulation system that produces deformation and/or convergence in the flow field, differential horizontal temperature advection, or secondary (vertical) circulations has the potential to create and strengthen sloping stable layers and their associated atmospheric stability and wind shear. Because of thermal wind requirements, such layers (also called baroclinic layers or frontal zones) are vertically sheared. Frontal zones are particularly common with extratropical cyclones and associated upper tropospheric jet streams, thus presenting optimal conditions for the development of KHI and CAT.

Additionally, developing extratropical cyclones and jet streaks, as well as sharply curved anticyclonic flows, are characterized by unbalanced (ageostrophic) flow. Under these conditions, accelerations generate a broad spectrum of not only IGW but also longer, inertia-gravity waves which may also be effective in producing CAT by modulating shallow shear layers, as well as by their own instabilities. Because the jet stream environment is optimum for the production of sheared stable layers, about two-thirds of CAT occurrences are found near the jet stream. As a function of height, CAT frequencies reach a maximum near the tropopause, and in the baroclinic zone below the jet core.

Macroscale flow patterns conducive to CAT are shown schematically in Figure 1. They can be classified into four basic types: (1) col pattern, (2) sharp trough, (3) ridge, and (4) negatively tilted trough associated with baroclinic instability. Based on observational studies, flow conditions resulting in anticyclonic shear or curvature (such as pattern (3), Figure 1) produce CAT most frequently, while cyclonic conditions (pattern (2), Figure 1) produce the most intense CAT. While patterns (2) and (4) are both associated with upper-level troughs, their orientations result in significant differences in both the locations of CAT and prevalent cloud conditions.

A northeast to southwest-oriented trough axis (pattern (2), Figure 1) results in CAT to the rear of the upper trough in relatively cloud-free conditions. A northwest to southeast trough (pattern (4), Figure 1), which is typically associated with cyclogenesis, results in CAT in advance of the trough, often embedded in thick cirrus clouds. As suggested by Figure 1, sharply curved jet stream segments associated with upper-level troughs or ridges are typically more turbulent than most zonal (predominately westerly) jet streams (see **Dynamical Meteorology: Overview; Inertial Instability; Kelvin–Helmholtz Instability. Synoptic Meteorology: Jet Streaks; Frontogenesis**).

Mesoscale Forcing

Mountain waves

The processes by which gravity waves can produce CAT have been studied extensively for a well-known type of IGW known as a mountain lee wave (typical wavelength: ~ 10 km). Lee waves are generated when stable airflow passes over a topographical barrier. On the lee side of the ridge, the airflow initially descends and then rebounds to generate wave updrafts. As the flow reaches its peak altitude and begins to descend again, a wave crest is

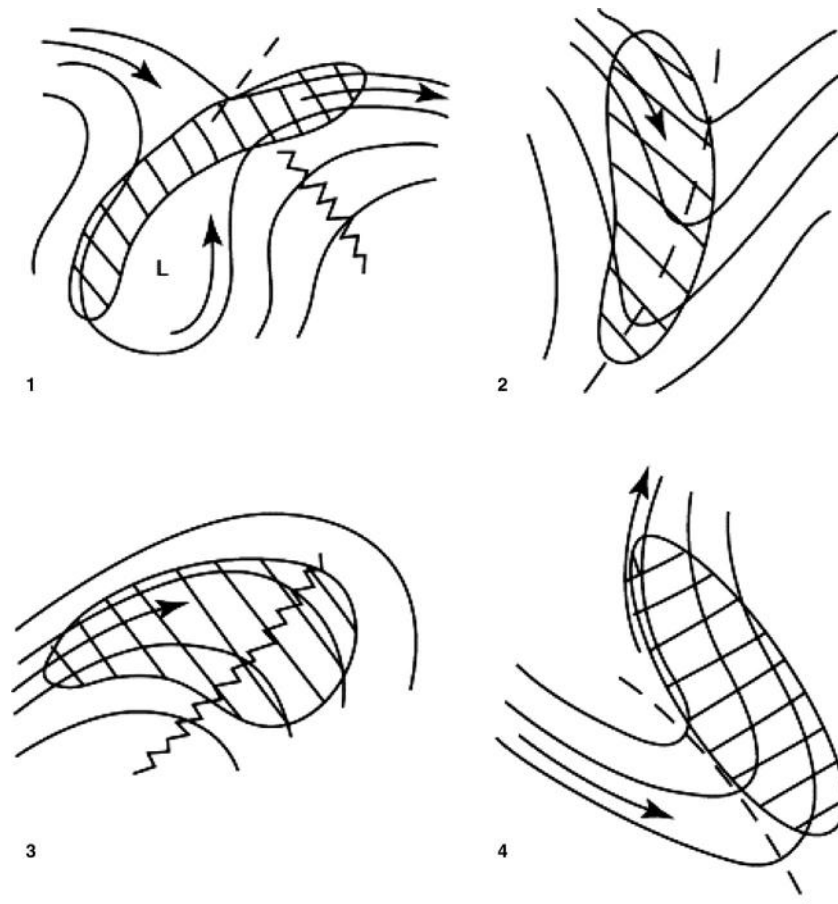


Figure 1 Idealized streamlines at jet stream level showing macroscale flow patterns that are most conducive to the occurrence of CAT. Hatched areas show where CAT is most likely. Lines with arrows represent approximate jet stream locations. The patterns are defined as (1) col or deformation zone, (2) sharp trough, (3) ridge, and (4) negatively tilted trough (associated with baroclinic instability). With (4), the CAT would most likely occur in or near dense cirrus clouds, since this is a favorable flow pattern for cyclogenesis. In day-to-day situations, there are many possible hybrid combinations of these patterns.

formed which may be made evident by the formation of lenticular clouds. Under the crests of strong waves, turbulent rotor circulations are often found in lee wave systems. Occasionally, with large mountain ranges, rotor-produced turbulence will reach the lower altitude range of CAT. In most cases associated with strong turbulence, energy from these waves is trapped at or below a strong temperature inversion near mountaintop level.

In some situations, depending on the wavelength and the distribution of wind velocity and temperature with altitude, lee wave energy can propagate to great heights to produce CAT by one or more of the processes described earlier. Atmospheric conditions typically associated with such higher altitude mountain waves are (1) a high, cold tropopause, (2) a thermal inversion near the mountaintop with (3) weak stability above, (4) low-level winds $>15 \text{ m s}^{-1}$ at mountaintop level, with (5) weak positive vertical wind shear aloft.

The increase of lee wave amplitude with height leads to greater vertical displacement and tilting of the stable layers, and often reduces the local (Ri) enough in local sectors of the wave pattern to produce strong CAT. Furthermore, very strong wave forcing occasionally produces a lee wave 'hydraulic jump' condition, resulting in a deep, extremely turbulent layer in the lee of the mountain, which may extend from near the surface through the tropopause.

An example of an aircraft encounter with severe mountain wave turbulence is shown in [Figure 2](#).

In some cases, a large mountain range will give rise to longer wavelength $\sim 50 \text{ km}$ or more, nearly hydrostatic lee waves. When favorable atmospheric wind and temperature profiles exist, shorter wavelength perturbations induced by these waves may destabilize, owing to partial reflection of lee wave energy from levels near the tropopause. Aircraft flying through such wave action may experience extreme vertical gusts.

Strong cold frontal systems, and jet streams occurring over midlatitude mountainous areas provide conditions favorable for the development of mountain lee waves. The enhancement of CAT-producing mechanisms by mountain waves accounts for the higher frequency of CAT over midlatitude mountains

than elsewhere. However, owing to the set of conditions required for breaking lee waves mentioned earlier, mountain wave-related CAT is not usually found beneath the jet stream core, but mainly on the anticyclonic side (see **Dynamical Meteorology: Overview. Mountain Meteorology: Lee Waves and Mountain Waves**).

Convection-Initiated CAT (CIT)

Another phenomenon that may produce CAT is deep mesoscale convection embedded within moderate to strong winds aloft. If convective cloud elements penetrate a vertically sheared, capping stable layer such as the tropopause, enhancement of those shears will result in the reduction of the local (Ri) and the production of KHI, leading to CIT. In these cases, longer wavelength IGW may also be produced. If stability and wind shear conditions are favorable, then IGW will propagate to their critical levels and break, contributing to CIT as high as several kilometers above the cloud top. Based on numerical model simulations, this latter process may occur even in cases of moderate cloud top winds ($\sim 15 \text{ m s}^{-1}$) and wind shears.

In cases of exceptionally strong convection, tops of thunderstorm cells penetrate the tropopause with vertical velocities as high as $20\text{--}30 \text{ m s}^{-1}$. As a result, IGW develop in the stable stratosphere and propagate away from the thunderstorm, generating CIT. If strong winds are present, flow regimes similar to mountain lee waves develop in the stratosphere, over and downwind of the convection. Therefore, at tropopause levels, an aircraft flying near thunderstorm tops is vulnerable not only to turbulence produced within a thunderstorm but also to CAT outside the thunderstorm. [Figure 3](#) is a record of vertical acceleration and altitude from a commercial jetliner for a CIT encounter near Hannibal, Missouri, on 4 April 1981 that resulted in injuries to passengers.

CAT near thunderstorms in midlatitudes occurs more often during spring and early summer in late afternoon and evening. However, large mesoscale convective systems (MCS) that form

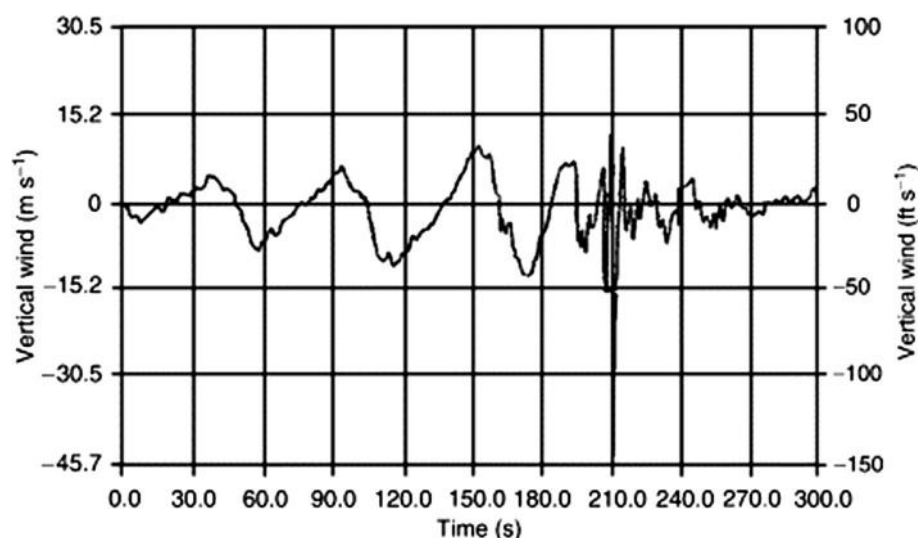


Figure 2 Time history of vertical velocities associated with mountain waves and severe turbulence as measured by an aircraft at 12.4 km (39 000 ft) altitude near Morton, WY, on 16 July 1982. Long-period, relatively smooth mountain waves along the flight path moving upwind (from left to right) deteriorate into extreme CAT conditions associated with KHI closer to the Wind River mountain range ridgeline (near time 210.0 s).

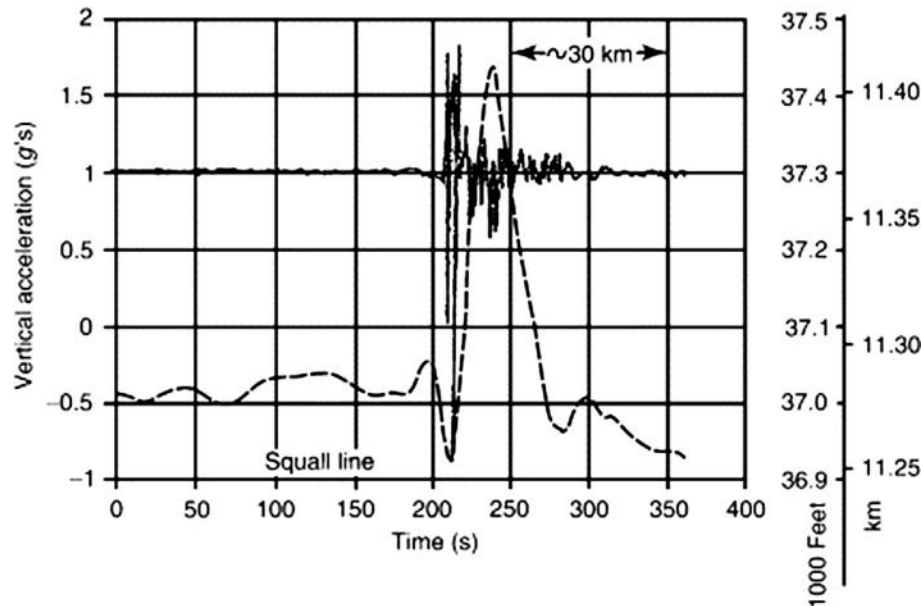


Figure 3 Time history of vertical accelerations (solid) and altitude (dashed) along the flight path of a commercial jet aircraft near Hannibal, MO, from 01.21 to 01.27 UTC on 4 April 1981. The point at which the aircraft flew downwind over the line of thunderstorms is labeled 'Squall Line.'

typically near the axis of an upper-level ridge at night tend to enhance the strength of the jet stream on the poleward side of the MCS. This often results in strong shears and unstable flows, often indicated by transverse cirrus cloud bands, with a significant area of CAT downstream that may continue for several hours after sunrise. Similar conditions may occur on the poleward side of tropical cyclones as they encounter strong westerly winds in the midlatitudes (see **Aviation Meteorology: Aviation Weather Hazards. Numerical Models: Convective Storm Modeling**).

Observations of CAT

Observations of CAT are critical for research, diagnosis, and prognosis of CAT. However, adequate resolution of CAT requires microscale measurements. Such measurements are not regularly available via the standard surface and upper air weather observation networks. Occasionally, instrumented and radar-tracked balloons or ground-based radars have been used for turbulence measurements, but aircraft have been the most frequent platforms of choice for the direct measurement of CAT.

Aircraft Measurements

Subjective pilot reports (PIREPs) include a description of CAT intensity, aircraft position and altitude, and appropriate remarks. Reports of turbulence intensity for most aircrafts are based on the pilot's estimate of flight control difficulty, of the movement of objects within the aircraft, or readings from available instrumentation, for example: (1) airspeed fluctuations, (2) rate of climb in otherwise level flight, and (3) g-meter excursions (see [Table 1](#)). Bias is frequently introduced in these reports as a function of aircraft type, suddenness of onset, and pilot experience, among several factors. Despite these

problems, PIREPs of turbulence are an important day-to-day source of direct CAT measurements.

Improvements in instrumentation and communications have made it possible for automated PIREPs from some commercial airliners to be acquired very quickly by international aviation weather forecast centers, increasing the timeliness and volume of CAT reports.

Some aeronautical and atmospheric research aircrafts are equipped to directly measure true turbulence gust velocities with high accuracy and sample rates (e.g., about 0.3 m s^{-1} and 50 samples per second, respectively). These measurements require specially calibrated sensors, high-capacity data recording systems, and judicious, postflight engineering analysis.

Most commercial airliners carry onboard inertial navigation and digital recording systems that are capable of sampling vertical and horizontal accelerations plus several other parameters including aircraft attitude, engine status, position, and altitude. Such onboard measurement systems provide information for research into the nature and impact of CAT; also, they have laid the foundation for the development of onboard turbulence metrics that lend themselves to automated and standardized CAT reporting, minimizing the bias of subjective CAT reports. Two metrics which suit these requirements are the turbulent kinetic energy (TKE) dissipation rate and the integrated TKE. The latter metric is derived from that part of the TKE spectrum where aircraft have the greatest sensitivity to turbulence. New *in situ* turbulence measurements available from some US airlines provide reports of the cube root of the EDR (see [Table 1](#)) on a 1-min basis.

Beyond using an aircraft as a platform for instruments to measure CAT directly, an ideal airborne CAT detection system should be able to detect the location and intensity of CAT far enough ahead of the aircraft so that the pilot has sufficient time to take evasive flight path deviations and/or to warn cabin crew and passengers to fasten seat belts to minimize the effects of the

turbulence. Although airborne remote sensing of CAT can be achieved partially with forward-looking infrared (IR) radiometers and lidars, the best approach may be the integration of data from multiple detection systems.

Ground-Based Measurements

Although lacking the mobility and limited in number, ground-based systems have generally had the advantage over airborne systems in size, power, and data processing capability to enable the sensing of turbulence and turbulence-related parameters at greater ranges.

Ground-tracked sounding balloon systems, such as radiosondes, are used ordinarily for the determination of indirect turbulence indicators such as stability, wind velocity, wind shear, and (Ri). An ongoing problem with the twice daily (00.00 coordinated universal time (UTC) and 12.00 UTC) radiosonde observations is the poor detection of thin, potentially turbulent layers, which may vary rapidly in time and space. A few specialized sounding balloon systems, such as those used for the support of the launch of space vehicles (e.g., radar tracking of the rigid, Jimsphere balloons at Cape Canaveral), are capable of resolving layers of 50 m or less. Some studies have used balloon rate-of-rise fluctuations to find a direct correlation between large-amplitude IGW vertical motions and turbulence in CAT regions.

Sensitive, vertically pointing, scanning Doppler radars, known as wind profilers, are used to acquire time sequences of wind speed and direction as a function of altitude. Ultra high frequency (UHF) or very high frequency (VHF) wind profilers are capable of making a complete sounding every 30–60 s with samples at 250 m intervals over an altitude range of 0.5–16 km above ground level (AGL). Typically averaged over a period of an hour to eliminate noise, profiler data are most useful for determining the altitude, intensity, and temporal behavior of larger scale features associated with CAT, such as shear zones, trough lines, and jet streams.

Very powerful ground-based research radars are capable of detecting clear air echoes associated with KHI. Modern Doppler weather radar systems can measure turbulence in clear air, based on the variance (or 'spread') of the velocity distribution, although that capability is generally limited to the boundary layer. Most operational Doppler radar systems may also be used to determine higher altitude wind information similar to that from wind profilers.

Ground-based research light detection and ranging (lidar) has been used to accurately measure wind profiles and observe violent wave-breaking episodes at jet stream altitudes. However, such observations cannot be acquired under all sky conditions; this is because the air at the turbulence locations must have enough particles to reflect some of the lidar energy, and the intervening path between the lidar and the turbulence must not have so many particulates or droplets that these would block the lidar beam or its return signal.

Remote Sensing from Space

The advent of the geostationary meteorological satellites in the later half of the twentieth century provided an additional tool for monitoring regions of potential turbulence. Although pixel

resolutions are only marginal for this purpose (1–5 km), the image frequency (5 min–1 h) and spatial coverage (global except near the poles) of geostationary satellites can be quite useful for the detection and short-range forecasting of CAT, especially when combined with other data. Spectral bands most useful for CAT detection are visible, IR, and water vapor (WV) (Table 2). Specific applications of these images will be described in more detail in the next section. Cloud motion vectors derived from a sequence of geostationary satellite images can provide valuable data over remote data-sparse regions for assimilation into numerical prediction models.

Visible and IR images from polar orbiting satellites are available less frequently (every 2–6 h, depending on latitude), but can be used to identify CAT cloud patterns associated with long-lasting, large-scale systems, and to corroborate features observed in geostationary satellite data. They are especially useful at high latitudes where geostationary coverage is poor because of extreme parallax.

Analysis of satellite imagery, compared with colocated PIREPs, indicates that CAT is found not only in clear air but also in cirrus clouds, and along borders of large-scale convective cloud systems. Zones of turbulence associated with the subtropical jet stream and convective outflow are often denoted by pronounced cirrus cloud bands that are oriented nearly perpendicular (transverse) to the flow. These clouds bands are possibly caused by shallow convection, similar to horizontal convective rolls in the boundary layer, and/or by other processes such as gravity waves or inertial instability. An example of this cloud feature associated with a subtropical jet is shown by the Geostationary Operational Environmental Satellite (GOES) IR image in Figure 4. Wider, thicker transverse cloud bands have been associated with a strong likelihood of moderate to severe CAT.

WV images from geostationary satellites are sensitive to moisture in the middle and upper levels of the troposphere. Pronounced warming (drying) observed over a period of a few hours in a series of WV images has been associated with strong subsidence and tropopause 'folds' in the vicinity of upper-level fronts, and a corresponding increase in the risk of CAT. These regions are usually associated with horizontal deformation zones (pattern (1), Figure 1), sharp upper troughs (pattern (2), Figure 1), or in the 'dry slot' region of intensifying cyclones. These synoptic patterns also correlate with high concentrations of stratospheric ozone and large values of potential vorticity. Objective methods are being developed to identify tropopause folds automatically using 6.7 μm WV imagery.

Table 2 Spectral channels used in detecting CAT-related phenomena from space

Channel type	Spectral range (μm)	Applications
Visible	0.5–1.0	Small-scale (≤ 1 km) wave clouds
IR	10–12	Cloud top temperatures (heights), mesoscale (up to 10^3 km) cloud borders, cloud bands, and wave clouds
WV	6–7	Regions (10 – 10^3 km) of sinking/rising associated with mountain waves, upper-level fronts, deformation zones, and cyclones

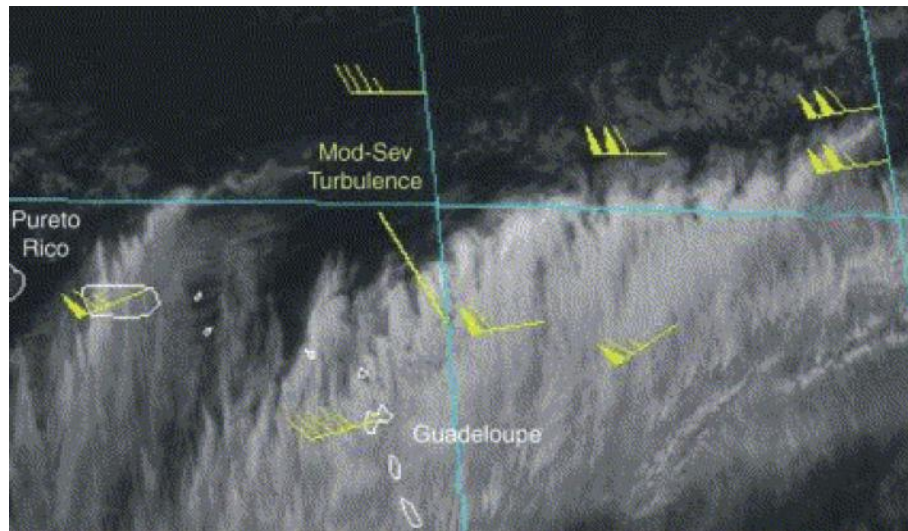


Figure 4 IR image from the GOES-8 at 00.45 UTC on 3 March 2000, showing transverse cirrus cloud bands on the equatorward side of the subtropical jet stream near the Leeward Islands. Moderate-to-severe turbulence was reported by a B767 aircraft at the location shown. Wind barbs are from radiosonde sites for 250 hPa at 00.00 UTC, or aircraft reports for 8.9–11.8 km (28 000–37 000 ft) between 00.00 and 03.00 UTC.

Satellite images in several spectral bands (visible, IR, and WV) may also show classic ‘washboard’ cloud patterns associated with mountain waves (Figure 5). The WV images often depict a greater area coverage of mountain wave conditions than either visible or IR, owing to their greater sensitivity to moisture. Warm subsidence zones oriented along and just downstream from the mountain ridges and slightly upstream from the lee cirrus plume, sometimes referred to as ‘Föhn gaps,’ are indicative of possible intense turbulence at high altitudes (upper left of panel a in Figure 5).

High-resolution visible imagery (0.5–1 μm) can sometimes detect very small-scale (≤ 1 km wavelength) wave cloud patterns (referred to as ‘billows’) that correspond to areas of KHI. Billow clouds are embedded typically within or near large-scale cloud systems, or in the vicinity of convective storms when wind shears are present (see **Satellites and Satellite Remote Sensing: Temperature Soundings; Surface Wind and Stress**).

CAT Prediction Techniques

As illustrated in Figure 1, certain synoptic-scale upper flow patterns have been empirically related to CAT occurrence through many years of operational experience. By applying these pattern types to predicted flow patterns from numerical forecast models, estimates can be made of the likelihood of CAT in certain regions for longer forecast time periods.

Objective prediction of CAT has become more commonplace and accurate, owing to increased computing speed, diminished cost of high-speed computers, improved numerical model physics, and better techniques for assimilation of wind and temperature data from aircraft and satellites into the numerical models.

A number of diagnostic and predictive indices for CAT have been developed and used over the years. The nondimensional (Ri) (see eqn [1]) is adequate in many situations, although it often exhibits a much wider range of critical values

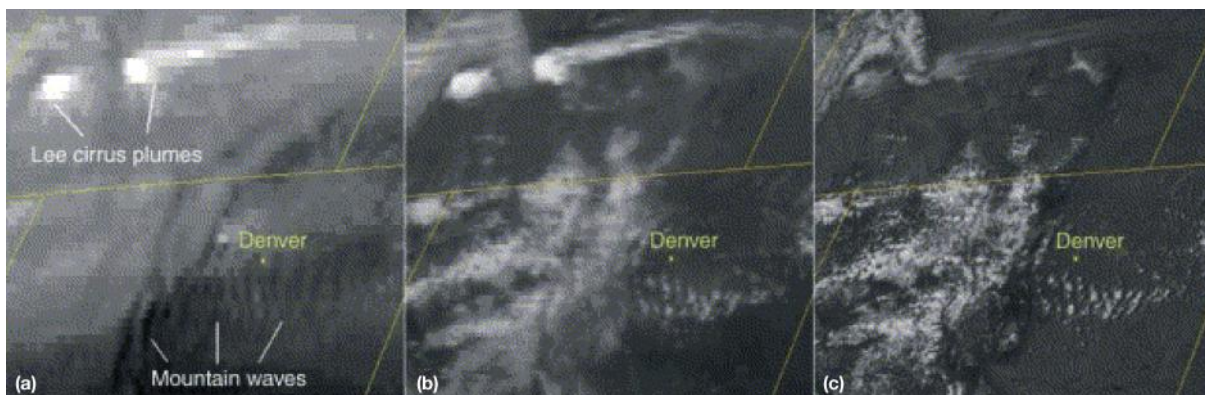


Figure 5 Mountain wave patterns east of the Rocky Mountains observed by GOES-8 in the (a) WV, (b) IR, and (c) visible band at 19.15 UTC on 5 April 2000. Considerable severe turbulence was reported below 6.4 km (20 000 ft). The WV image depicts more extensive coverage of mountain waves than the other two images. Cirrus plumes observed in upper left are indicative of high-altitude mountain waves downwind of the Salt River and Wind River ranges in Wyoming.

in the free atmosphere than in the laboratory. (Ri) is most reliable when generated from high-resolution observational data; therefore, it is principally used as a diagnostic tool. (Ri) is defined as:

$$(Ri) = \frac{(g/\theta)(\partial\theta/\partial z)}{(\partial V/\partial z)^2} \quad [1]$$

where θ is the potential temperature (K), V the vector horizontal wind (m s^{-1}), g the acceleration of gravity (m s^{-2}), and z the height (m).

A more successful index for CAT prediction employs the (Ri) tendency equation. Significant, prolonged turbulence is most likely in regions of the atmosphere where larger scale processes are continually acting to decrease (Ri) , despite the influence of turbulent mixing. The primary processes include horizontal deformation, convergence, and differential thermal advection. Specifically, the (Ri) tendency technique calculates the large-scale forcing necessary to overcome the kinetic energy dissipation resulting from the turbulence process in order to maintain the (Ri) at values $\leq 1/4$. A simple formulation of (Ri) tendency equation is

$$\frac{d\ln(Ri)}{dt} = -\Phi - \varepsilon \quad [2]$$

where Φ is the nonturbulent, large-scale forcing by deformation and ε is the turbulent dissipation.

A similar approach can generate a diagnostic or predictive index by simply calculating the product of horizontal, resultant deformation, and the vertical vector wind shear at each grid point. The basis of this index is similar to (Ri) tendency in that it considers large-scale forcing from frontogenesis, in addition to the presence of vertical shear in the mesoscale. One difficulty with this approach is that deformation can also lead to frontolysis in some cases, which tends to reduce vertical shear. This turbulence index (TI) is defined as:

$$TI = \frac{[(\partial u/\partial x - \partial v/\partial y)^2 + (\partial v/\partial x + \partial u/\partial y)^2]^{1/2}}{\text{Resultant deformation} \times \partial V/\partial z} \quad [3]$$

Vertical shear

where u and v are wind components (m s^{-1}), V the vector wind, and z the height (m). A recent improvement to TI incorporates temporal changes in divergence that captures a wider range of CAT events, with a minimal cost in overprediction.

The TKE technique for the prediction of CAT attempts to show areas of turbulence generation through the processes of production by vertical shear, production or destruction by buoyancy, energy advection, and loss via dissipation. The simplified TKE equation can be expressed as:

$$\frac{d(\text{TKE})}{dt} = \underbrace{-u'w'\frac{\partial u}{\partial z} - v'w'\frac{\partial v}{\partial z} + w'\theta\frac{\varepsilon}{\theta}}_{\text{Energy production Shear production Buoyancy} + A - \varepsilon} \quad [4]$$

Advection Energy dissipation

where $\text{TKE} = (u'^2 + v'^2 + w'^2)/2$, total kinetic energy per unit mass, u' , v' , and w' are perturbation (gust) velocities, and A is an advection term.

The TKE approach can account for turbulence from a wide variety of mechanisms such as convection and mountain

waves, as well as jet stream CAT, and provides a direct estimate of possible turbulence intensity through the energy dissipation rate. While TKE presents the most rigorous depiction of turbulent processes of all the objective approaches, it is more effective when numerical models have very high vertical (~ 100 m) and horizontal resolutions (~ 10 km). In solving eqn [4] using such high-resolution models, the need for accurate observations is critical, and the best analysis and prediction results are attained in regions of dense reporting networks that incorporate aircraft data, chiefly over continental areas.

The indices described above do a reasonably good job of predicting large outbreaks of CAT, but tend to predict CAT over a region that is larger than the area of actual occurrence (also known as 'overprediction'). Their inadequacies are due largely to the following factors: (1) they only account for large-scale conditions favorable for CAT, (2) they do not consider triggering mechanisms, and (3) most numerical models cannot accurately account for the intense, subgrid-scale vertical shears and strong horizontal forcing present during severe CAT. More specifically, model problems are due to resolution limitations and systematic underforecasting of maximum wind speeds within the jet stream.

A new direction in CAT forecasting that is based on a specific triggering mechanism employs the Lighthill–Ford (LF) theory of spontaneous generation of gravity waves. Scaled for midlatitude synoptic-scale flows, it can be expressed to second-order accuracy as a shallow water wave equation for inertia–gravity waves with the following wave forcing terms:

$$LF \sim f\mathbf{u} \cdot \nabla \zeta - \frac{\partial}{\partial t} J(u, v) + 2Df\zeta + f\left(v\frac{\partial D}{\partial x} - u\frac{\partial D}{\partial y}\right) \quad [5]$$

in which f is the Coriolis parameter, ζ is the relative vorticity, and D is the horizontal divergence.

Initial experimental results with this index coupled to a first-order TKE closure technique have indicated a high degree of skill, particularly for moderate or greater CAT. While differences exist regarding the applicability to CAT occurrences and its operational implementation, this new approach has the potential to significantly improve CAT forecasting (see **Gravity Waves: Overview. Turbulence and Mixing: Overview; Turbulence, Two Dimensional; Turbulent Diffusion**).

Statistical Approaches to CAT Prediction

Because of the sporadic, microscale nature of CAT, it would seem that statistical approaches would be useful for forecasting its occurrence. The first such efforts were completed in the United Kingdom in the late 1970s. Turbulence data from about 4500 aircraft reports, compared with 11 colocated numerical parameters derived from a coarse-resolution prediction model, revealed that the best correlation was between CAT and vertical and horizontal wind shears. Similar studies were completed in the United States in the 1980s using higher-resolution numerical model data that showed CAT to be highly correlated with horizontal deformation and scalar wind speed.

A technique that statistically integrates information from many numerical turbulence indices, known as the Graphical Turbulence Guidance (GTG), was developed in the late 1990s, initially for use over North America. GTG first assigns a score to

each diagnostic index, based on comparison with available PIREPs at the initial time period throughout the numerical grid domain. A weighted sum is then determined from all of the indices to arrive at a final GTG index value at each grid point. The same weights are then assigned to each index in deriving predictions for each forecast time period. This process (known as 'dynamic weighting') is repeated with every forecast cycle. GTG algorithms have been developed for both upper level (above 6.1 km (20 000 ft)) and mid-level (3.0–6.1 km (10 000–20 000 ft)). Ten indices are used for the upper level, while nine are applied for mid-level forecasts. The mid-level GTG uses some of the same diagnostic indices as the upper level, but with a few that are different. GTG is also under evaluation for use in other regions of the globe, such as East Asia.

To be successful, an objective TI must provide high probability of detection for turbulence (POD_y), coupled with an acceptably low false alarm rate (FAR). The POD for smooth conditions (POD_n) is preferred over FAR in turbulence and icing verification because POD_n is less sensitive to the number of PIREPs in the data sample than FAR. A metric often used in CAT forecast verification that combines POD_y and POD_n is the True Skill Statistic ($TSS = POD_y - POD_n - 1$).

A verification method that has gained favor is the relative operating characteristic (ROC), derived from signal detection theory. In an ROC graph (such as shown in Figure 6), the performance of CAT forecasts for various thresholds of a given index is plotted in terms of $1 - POD_n$ (x-axis) and POD_y (y-axis). The metric $1 - POD_n$ can be interpreted as the proportion of smooth reports that are incorrectly classified. This creates a curve that traces out the index's performance for all choices of values that trigger a 'yes' forecast for turbulence, from high thresholds (low POD_y , low $1 - POD_n$) to low thresholds (high POD_y , high $1 - POD_n$). The area under the ROC curve is a quantitative estimate of skill; values greater than 0.5 indicate that an index's performance is better than sheer chance.

The GTG procedure minimizes forecast errors due to uncertainties in individual turbulence diagnostic indices and their

respective thresholds. Verification using PIREPs has shown that GTG output based on the dynamic weighting system described above is generally superior to the individual indices based on POD_y , POD_n , TSS, and ROC statistics (Figure 6). The LF method, which is still under evaluation, has shown the potential to provide even better skill in ROC statistics.

Although useful, statistical approaches to CAT prediction are not always able to correctly depict turbulence events accurately because of limitations in numerical model data or scarcity of PIREPs at the initial time period. Climatologically determined weights developed for GTG hold some promise for use in data-sparse regions. The best results for CAT prediction in the long term will likely stem from the development and improvement of physically based techniques, such as LF, that are then incorporated into increasingly sophisticated statistical schemes.

CAT Climatology Studies

Some advances in the knowledge of the global distribution of CAT along heavily traveled airways have been derived from programs that collected large numbers of PIREPs, such as one conducted by the International Civil Aviation Organization in the mid-1960s. With the rapid increase in high-altitude flights since then, PIREP-based CAT climatology studies have become more meaningful, although they still lack sufficient data in many regions. For example, a recent study over the United States based on more than two million PIREPs over a 12-year period found that regional maxima in CAT frequency occur over the central Rocky Mountains, the South, Southeast, and along the North Atlantic seaboard. However, there were some large regions in the southwest near the Mexico border that had very little data, probably due to flight restrictions over military operation areas.

The improvement of numerical prediction models over the past several decades has resulted in their viable use in global CAT climatology studies. More comprehensive global 'climatologies' of large-scale, upper-level conditions favorable for CAT have been constructed by averaging numerical model indices (such as the TI described in CAT Prediction Techniques) to infer the global distribution of CAT activity. This approach usually describes CAT produced by jet streams and upper fronts (e.g., cyclogenesis), but does not account for mountain waves and convection. It also does not provide an indication of the likelihood of severe turbulence.

Based on a 44-year average of TI for example, regions of relatively high CAT risk have been identified over Southern and Eastern Asia, Southwestern and Central United States, and coastal Northeast United States and Canada (Figure 7). A short period (2 years) study using TI in the Southern Hemisphere, which has relatively little land mass, has indicated that maximum CAT is likely near Southern Australia and New Zealand (not shown).

The seasonal variation of the mean TI (Figure 7) also shows that the zone of maximum CAT occurrence tends to weaken and shift poleward during the warm season in each hemisphere (Figure 7), as one would expect due to the annual migration and weakening of the jet stream and upper-level fronts.

Large year-to-year changes have also been observed in these data that are likely related to interannual oscillations, such as the

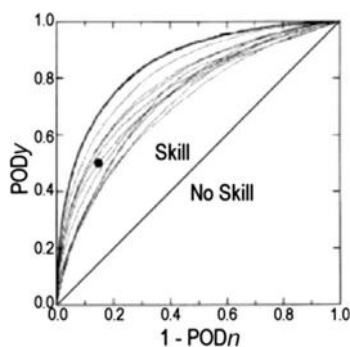


Figure 6 ROC curves of the performance of 6-h forecasts of the upper-level version of GTG for 2003, validated using nearly 50 000 PIREPs. Individual CAT diagnostics are shown in thin gray lines; the GTG combination is the heavy, solid curved line; and the dot represents the average performance of Airman's Meteorological Information (AIRMETs) (with amendments). The thin diagonal line through the origin represents the no-skill line. Reproduced from Sharman, R., Tebaldi, C., Wiener, G., Wolff, J., 2006. An integrated approach to mid- and upper-level turbulence forecasting. *Weather and Forecasting* 21, 268–287.

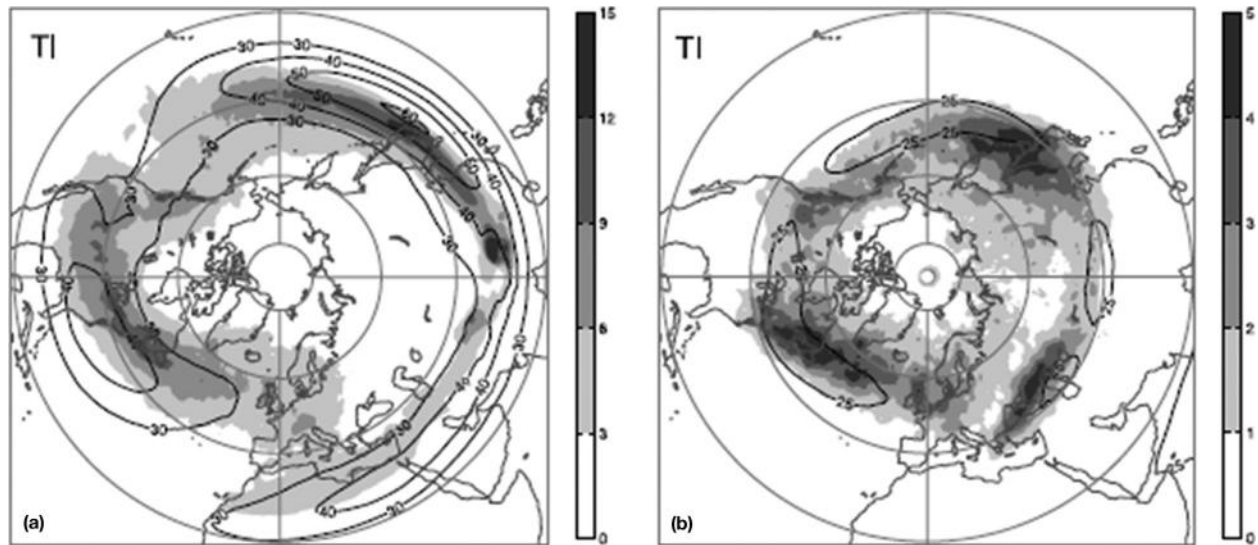


Figure 7 Mean frequency (%) of high values ($>12 \times 10^{-7} \text{ s}^{-2}$) of the CAT TI (shaded areas) for the Northern Hemisphere winter (a) (Dec., Jan., and Feb.) and summer (b) (Jun., Jul., Aug.) based on 44 years of data (1958–2001). Contours show mean wind speed (m s^{-1}). The index was derived from the European Center for Medium Range Weather Forecasts (ECMWF) reanalysis data set (ERA40). Turbulence is likely to be underestimated, since the index does not account for mountain waves or deep convection. Reproduced from Jaeger, E.B., Sprenger, M., 2007. A Northern Hemispheric climatology of indices for clear air turbulence in the tropopause region derived from ERA40 reanalysis data. *Journal of Geophysical Research* 112, D20106, doi:10.1029/2006JD008189.

El Niño Southern Oscillation (ENSO) phenomenon. During strong El Niño conditions, the subtropical jet stream in the Northern Hemisphere is typically much stronger than normal, resulting in a higher potential for CAT occurrence in regions where it converges with the polar jet stream (such as over North Africa and the Southwest United States).

See also: **Aviation Meteorology:** Aviation Weather Hazards. **Dynamical Meteorology:** Inertial Instability; Kelvin–Helmholtz Instability; Overview. **Gravity Waves:** Overview. **Mountain Meteorology:** Lee Waves and Mountain Waves. **Numerical Models:** Convective Storm Modeling. **Satellites and Satellite Remote Sensing:** Surface Wind and Stress; Temperature Soundings. **Synoptic Meteorology:** Frontogenesis; Jet Streaks. **Turbulence and Mixing:** Turbulent Diffusion; Overview; Turbulence, Two Dimensional.

Further Reading

- Anderson, R.K., 1995. Synoptic scale cloud and moisture patterns; Clouds generated by mountains. In: Bader, M.J., Forbes, G.S., Grant, J.R., Lilley, R.B., Waters, A.J. (Eds.), *Images in Weather Forecasting: A Practical Guide for Interpreting Satellite and Radar Imagery*. Cambridge University Press, Cambridge, UK, pp. 70–137. 472–477.
- Cornman, L.B., Morse, C.S., Cunniff, G., 1995. Real-time estimation of atmospheric turbulence severity from in-situ aircraft measurements. *Journal of Aircraft* 32 (1), 171–177.
- Durrant, D.R., 1986. Mountain waves. In: Ray, P. (Ed.), *Mesoscale Analysis and Forecasting*. American Meteorological Society, Boston, MA, pp. 472–492.
- Dutton, J.A., Panofsky, H.A., 1970. Clear air turbulence: a mystery may be unfolding. *Science* 167, 937–944.

- Ellrod, G.P., Knapp, D.I., 1992. An objective clear-air turbulence forecasting technique: verification and operational use. *Weather and Forecasting* 7, 150–165.
- Ellrod, G.P., Knox, J.A., 2010. Improvements to an operational clear-air turbulence diagnostic index by addition of a divergence trend term. *Weather and Forecasting* 25, 789–798.
- Holton, J.R., 1992. *Introduction to Dynamic Meteorology*. Academic Press, New York, NY.
- Jaeger, E.B., Sprenger, M., 2007. A Northern Hemispheric climatology of indices for clear air turbulence in the tropopause region derived from ERA40 reanalysis data. *Journal of Geophysical Research* 112, D20106. doi:10.1029/2006JD008189.
- Keller, J.L., 1990. Clear air turbulence as a response to meso- and synoptic-scale dynamic processes. *Monthly Weather Review* 118, 2228–2242.
- Knox, J.A., 1997. Possible mechanisms of clear-air turbulence in strongly anticyclonic flow. *Monthly Weather Review* 125, 1251–1259.
- Knox, J.A., McCann, D.W., Williams, P.D., 2008. Application of the Lighthill-Ford theory of spontaneous imbalance to clear-air turbulence forecasting. *Journal of the Atmospheric Sciences* 65, 3292–3304. doi: 10.1175/2008JAS2477.1.
- Lane, T.P., Sharman, R.D., Trier, S.B., Fovell, R.G., Williams, J.K., 2012. Recent advances in the understanding of near-cloud turbulence. *Bulletin of the American Meteorological Society* 93, 499–515.
- Lester, P.F., 1994. *Turbulence: A New Perspective for Pilots*. Jeppesen, Englewood, CO.
- Pao, Y., Goldberg, A., 1969. *Clear Air Turbulence and Its Detection*. Plenum Press, New York, NY.
- Schwartz, B., 1996. The quantitative use of PIREPs in developing aviation weather guidance products. *Weather and Forecasting* 11, 372–384.
- Sharman, R., Tebaldi, C., Wiener, G., Wolff, J., 2006. An integrated approach to mid-and upper-level turbulence forecasting. *Weather and Forecasting* 21, 268–287.
- Trier, S.B., Sharman, R.D., Fovell, R.G., Frehlich, R.G., 2010. Numerical simulation of radial cloud bands within the upper-level outflow of an observed mesoscale convective system. *Journal of the Atmospheric Sciences* 67, 2990–2999.
- Vinnichenko, N.K., Pinus, N.Z., Shmeter, S.M., Shur, G.N., 1980. *Turbulence in the Free Atmosphere*. Consultants Bureau, New York, NY.
- Wolff, J.K., Sharman, R.D., 2008. Climatology of upper-level turbulence over the contiguous United States. *Journal of Applied Meteorology and Climatology* 47, 2198–2214.

BIOGEOCHEMICAL CYCLES

Contents

Sulfur Cycle

Bromine

Heavy Metals

Iodine

Sulfur Cycle

P Brimblecombe, University of East Anglia, Norwich, UK

© 2003 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 1, pp 213–220, © 2003, Elsevier Ltd.

Introduction

The sulfur cycle transfers enormous amounts of this biologically important element through the atmosphere every year. Despite the size of natural reservoirs, human activities have had a remarkable effect on the sulfur cycle, perhaps most notably in the production of acid rain. The Earth's surface is a source of windblown dusts, but this is influenced by grazing activities and desertification. Large amounts of sulfur are mobilized by mineral extraction and fuel use. Fossil fuel refining and combustion are the major human emissions to the atmosphere, which exceed the natural sources. The reduction in anthropogenic emissions from North America and Europe is countered by growing emissions from Asia and may represent a special challenge for the twenty-first century.

Sulfur

The sulfur cycle is one of the major elemental cycles. Sulfur has an enormous span of oxidation states: from sulfides in the $-II$ state through to sulfates in the $+VI$ state. Both sulfides and sulfates are frequently found as minerals, which include the innumerable sulfides of metals or sulfates such as that of calcium (gypsum) and/or barium (barite). Sulfur is also found in its elemental state. Rhombic and monoclinic crystalline forms both contain S_8 rings. Liquid sulfur undergoes sharp changes in properties (color, viscosity, specific heat) as the temperature rises above the melting point (112.8°C for rhombic sulfur), with complex changes in the ring structure. Sulfur is essential to living organisms because it is incorporated into amino acids, proteins, and other biomolecules.

There is a naturally occurring radioisotope, sulfur-35, produced by the interaction of cosmic rays with argon in the upper atmosphere. It has a radiochemical half-life of 87 days and is rapidly oxidized to sulfur dioxide (SO_2) and sulfate, which is ultimately transferred to the lower atmosphere and removed in rain.

Evolution of the Sulfur Cycle

Sulfur is a common element with a high cosmic abundance half that of silicon (a frequent marker element). Concentrations of sulfur are much reduced in the Earth's crust because of downward separation when it is segregated in heavy sulfide magmas. Places where sulfur is plentiful are usually the result of intrusions that occurred in the earliest stages of the Earth's history. The element remains very much associated with volcanic activity. The Earth formed through gradual accretion of infalling material during Hadean time (4.5 to 3.8 billion years ago). Accretion left the mantle extensively molten. Gradually the Earth's surface cooled and solid rock formed, probably prior to 3.8 billion years ago. Metallic iron influenced the oxidation state of outgassed volatiles and hydrogen sulfide would have made an important contribution to the early sulfur budget. This differs from the current dominance of volcanic gases by sulfur dioxide. On the Archean Earth (3.8 to 2.5 billion years ago) hydrogen sulfide could be removed from the atmosphere through dissolution and precipitation as iron sulfide from the oceans or destruction by photochemical processes or lightning in the atmosphere.

Within the early oceans the concentration of sulfate would have been controlled by the provision of oxidizing materials from the atmosphere. Sulfate could not have risen to high

concentrations, while the hydrogen content of the atmosphere was maintained in excess of 0.0001 atm through rapid volcanic input.

Atmospheric oxygen remained considerably less abundant through the Archean (3.4 to 2.8 billion years ago) than today. Nevertheless at the beginning of the Archean, anaerobic photosynthesis provided a weak source of sulfate to the global ocean. Photodissociation of water provided a limited amount of oxygen that was supplemented with oxygen from oxygenic photosynthesis about 2.8 billion years ago. This would probably lead to atmospheric oxygen concentrations high enough to oxidize the sulfur in pyrite to sulfate. Bacterial reduction of sulfate to sulfides in marine sediments is a good indicator of increasing sulfate in sea water, with early indications found between 2.7 and 2.5 billion years ago. The sedimentary sulfur isotope record suggests that there were low concentrations of seawater sulfate and oxygen in the early Archean and atmosphere, respectively. Sulfate concentrations in sea water were probably less than a millimolar until the substantial oxygen increases of the early Proterozoic (2.5 to 0.54 billion years ago).

Volcanic Sources of Sulfur

Although elemental sulfur occurs in volcanic gas, perhaps as the dimer, most is found as either SO_2 or hydrogen sulfide (H_2S). The equilibrium calculations of the composition of volcanic gases show oxidized systems that contain SO_2 , and smaller amounts of S_2 and H_2S . More reduced systems may have COS and H_2S in greater relative abundance.

Some fluxes of sulfur gases from volcanoes occur as fairly continuous small releases from fumaroles into the atmospheric boundary layer. Violent volcanic eruptions represent a more sporadic but potentially much larger source of sulfur. The Tambora eruption of 1815 released some 50 Tg of sulfur, but this was exceptionally large (by comparison the Pinatubo eruption of 1991 released perhaps only one-fifth as

much). Such explosive volcanism can inject materials into the atmosphere with such force that the sulfur compounds can be driven into the stratosphere (Figure 1 and Figure 2). These gases are slowly converted to sulfate aerosols with a residence time of about a year in the stratosphere. Larger ash particles sediment out more quickly. Sulfur condenses as sulfur crystals on rocks near volcanic vents. This has provided a source of sulfur since ancient times and the risky activity of mining volcanoes for sulfur continues to the present day.

The variability and the shifting balance between fumarolic and explosive injections make globally averaged emission estimates difficult. Over the last quarter of the twentieth century the annual volcanic emissions have varied by a factor of three, but on average they were 7 Tg(S) year⁻¹ as SO₂ and about 2.6 Tg(S) year⁻¹ as H₂S. The volcanic emissions of carbon disulfide and carbonyl sulfide are an order of magnitude lower than H₂S.

Recently it is clear that the sulfuric acid droplets in the stratosphere (Figure 2) can enhance ozone depletion by providing sites for heterogeneous reactions. This can lead to a reduction in the thickness of the ozone layer after volcanic eruptions.

Marine Sulfur Sources

Present-day sea water has very high concentrations of sulfur, largely because of its high solubility as the sulfate. Sulfate is found at 0.028 mol l^{-1} and is thus the most abundant anion apart from chloride. Hydrothermal processes at midocean ridges are able to reduce sulfates to sulfides. As sea water evaporates it first precipitates calcium and magnesium carbonates, and calcium sulfate, gypsum. When sea water is trapped within basins in arid climates, it can lead to extensive deposits of evaporite minerals. In the geological past these shallow basins were more extensive than at present and gave rise to very extensive evaporite beds.

Sea spray can be blown directly from the sea surface during high winds. The particles of salt from this process tend to be

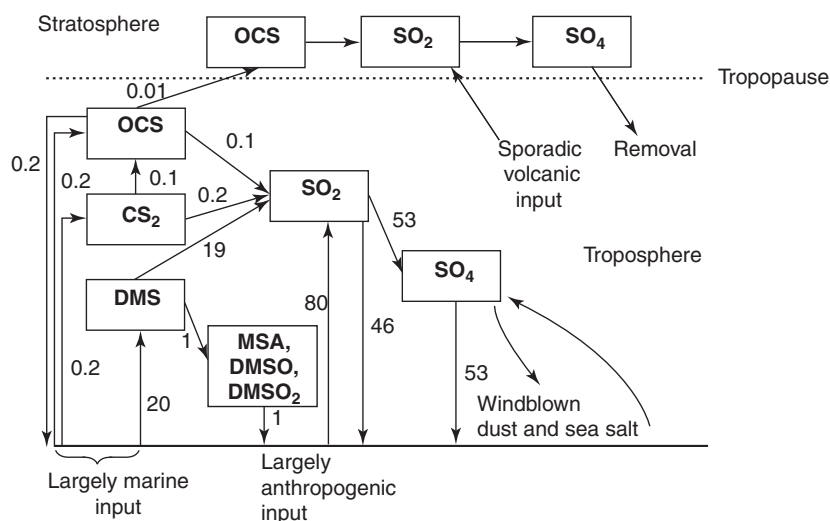


Figure 1 Atmospheric sulfur cycle with fluxes in Tg(S) year^{-1} . The reduced sulfur compounds DMS, OCS, and CS_2 come largely from marine processes. In years with much explosive volcanism there can be large sources of SO_2 directly to the stratosphere. Adapted with permission from Rodhe (1999).

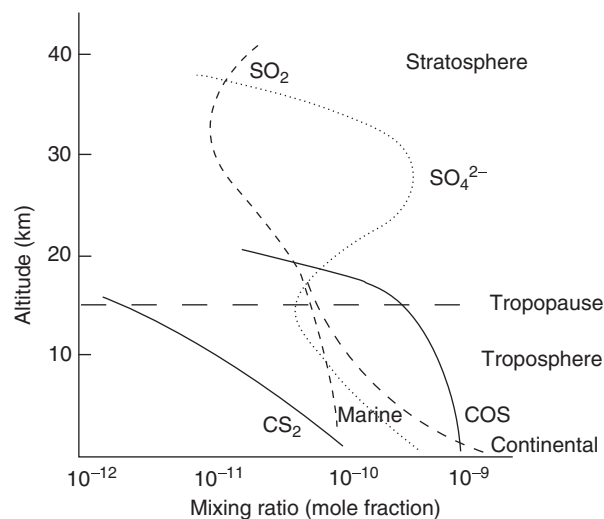
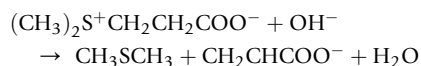


Figure 2 Vertical profile of sulfur species. The sulfate aerosol layer in the stratosphere has important effects on temperatures and ozone depletion. The marine and continental profiles for sulfur dioxide are different in the troposphere.

large and short-lived. Smaller particles are produced during bubble bursting. The very finest arise when the cap of a surfacing bubble shatters, giving film drops that dry into salt crystals of 5–50 pg. Along with these are larger jet drops that emerge as the bubble cavity collapses and give particles of about 0.15 μg . Such processes lead to the production of around a gigagram of sea salt particles each year, of which 30 Tg would be sulfur in the form of sulfate. Thus high chloride and sulfate concentrations are found in coastal rainwater.

H_2S is extensively produced from sulfate reduction in tidal flats and marshes. This is not surprising given the high concentration of sulfate in sea water. H_2S had long been associated with biological activities and was always assumed to be an important source of reduced sulfur to the atmosphere. Anaerobic environments are protected from invasion of oxygen through burial or lying underwater. This limited exchange of oxygen can also restrict the loss of biogenic H_2S to the atmosphere. Concentrations are typically in the range 30–100 ppt in remote air, although on tidal mud flats or in volcanic regions they can be much higher. H_2S oxidizes rapidly in sea water (c. 10 min), so any of this gas found in marine air would have to be produced close to the sea surface.

A preoccupation with H_2S has meant that early studies overlooked other reduced sulfur gases. The most notable was dimethylsulfide (DMS or CH_3SCH_3) identified first by James Lovelock in the 1970s. DMS is the largest natural source of reduced sulfur gases. It is produced through the activities of phytoplankton and most arises from dimethylsulfonium propionate ($[(\text{CH}_3)_2\text{S}^+\text{CH}_2\text{CH}_2\text{COO}^-]$ or DMSP) which leaks from aging or ruptured cells. The bacterial decomposition of DMSP in sea water cleaves it into DMS and acrylic acid:



High concentrations of DMS are found in areas where micro-organisms are abundant, but the amounts released vary with

species and the stage of their life cycle. Seasonal maps of DMS in sea water in the oceans have become quite detailed, reflecting its biological origin. DMS is released from sea water as it is relatively insoluble and readily transfers to the atmosphere. The fluxes to the atmosphere are dependent not only on concentration but also on wind speed. However, there remains an uncertainty of a factor of two with the estimated emissions of 20 Tg(S) year^{-1} . Some DMS in sea water is oxidized, probably to dimethylsulfoxide (DMSO, CH_3SOCH_3). This may be a photosensitized or biological oxidation, while some DMSO is produced directly from biological activities.

DMSO is more soluble than DMS, and so is unlikely to degas from sea water in large amounts. Other alkyl sulfides are more likely to be released from the ocean, but their abundances and thus fluxes are much smaller. Methanethiol (CH_3SH) and dimethyldisulfide (CH_3SSCH_3) are perhaps the best known of these.

The oceans are also a source of carbon disulfide (CS_2) which is released from the ocean at about 0.2 Tg year^{-1} . CS_2 is produced by photosensitized reactions in sea water. Laboratory irradiations have confirmed that the amino acids cysteine and cystine are efficient precursors of CS_2 . Biological production of CS_2 has also been observed in some marine phytoplankton.

CS_2 can be easily oxidized and contributes to the production of carbonyl sulfide (OCS). There are also photochemical mechanisms for the production of OCS from organic matter in the oceans. The emissions show a diurnal cycle due to the photochemical origin, but the fluxes from the ocean are much smaller than DMS. Nevertheless OCS is the most abundant reduced sulfur gas in the atmosphere as it is very stable in the troposphere (25 year residence time) and thus able to build up to substantial concentrations (0.5 ppb).

The low-temperature oxidation products of atmospheric DMS, most notably DMSO, dimethylsulfone ($\text{CH}_3\text{SO}_2\text{CH}_3$), and methanesulfonic acid (MSA, $\text{CH}_3\text{SO}_3\text{H}$), have been found in the Antarctic atmosphere typically at 1.5, 1.3, and 30 ppt, respectively. They may all be found in rainwater, but MSA is especially soluble and so is largely associated with aerosol particles. At times MSA is almost as abundant as non sea salt sulfate (NSS), but under warmer conditions it may only be a 50th that of NSS, the major oxidation product.

These marine sources of reduced sulfur gases can be important as a source of sulfur to the continents. The gypsum accumulations of the hyper-arid Central Namib Desert derive from non sea salt sulfur, in particular oxidation products of marine DMS.

Terrestrial Sulfur Sources

Windblown dusts are important as a source of sulfate in the atmosphere. The largest of these often occur from evaporitic deposits. Dusts from the margins of the Aral Sea (0.1–1 Tg year^{-1}) and arid states of the southern United States are notable source regions.

Biogenic sulfur emissions from land have also been recognized as making an important contribution to the global budget, but the fluxes have been more difficult to establish. Many of the gases that originate from biological activity in the

Oceans are also emitted from the continents, although the quantities are far less certain. Growing crops and biochemical activities in wetlands, especially in salty coastal environments, give predominantly DMS, H₂S, and methyl sulfide, with generally much lower amounts of OCS and CS₂. Tundra, while seemingly a likely source, has been found to be a low emitter of reduced sulfur because of low sulfate amounts in the water.

Biological Sinks for Sulfur

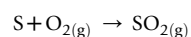
Vegetation can remove sulfur compounds from the atmosphere. Carbonyl sulfide removal by plants is the best known and probably represents a dominant sink for tropospheric OCS. The fate of some oxidation products of DMS is not clear. Some biochemical studies suggest that they may degrade biologically. Facultatively methylotrophic species of *Hyphomicrobium* and *Arthrobacter* seem to be able to produce enzymes necessary for a reductive/oxidative pathway for DMSO and dimethylsulfone (although the former appears to be more generally utilizable). Methanesulfonic acid is stable to photochemical decomposition in the atmosphere and its fate on land has been puzzling. A range of terrestrial methylotrophic bacteria that appear common in soils may mineralize MSA to carbon dioxide and sulfate.

Anthropogenic Sulfur Sources

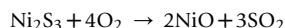
The most pernicious source of human sulfur emissions has been from the combustion of high-sulfur coals. This began in thirteenth-century London with the depletion of nearby wood supplies and increased through the centuries that followed until it reached its peak within the early twentieth century in Europe and North America. In Asia, particularly, emissions have continued to grow with the enormous pressure of industrial development. In the North Atlantic sector there has been a general decline in sulfur emissions with a shift away from coal as a fuel, in all but extremely large industrial plants, as a response to the problems of urban air pollution and acid deposition. It is estimated that in south and eastern Asia more than 10⁷ km² of land will be subjected to sulfur deposition greater than 1 g(S) m⁻² year⁻¹ by AD 2020.

Sulfur in coal exists half as pyrites, which is relatively easy to remove, but the rest is organically bound. This makes it hard to remove except by methods such as coal gasification or in stacks after emission. Sulfur can be removed by catalytic hydrodesulfurization from residual oil, but it leads to fuels that tend to become waxy at low temperature.

The combustion of fuels leads to the release of SO₂ in a simple but effective oxidation:



Many refining and extractive processes release large amounts of air pollution containing SO₂. For example, sulfide ores have been roasted in the past with uncontrolled emissions:



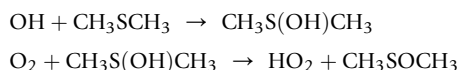
The SO₂ released often destroyed large tracts of vegetation downwind from smelters.

Various techniques are used to remove sulfur from stack gases. The use of lime (calcium hydroxide) or limestone (calcium carbonate) slurries to absorb SO₂ is widely adopted. The main product, calcium sulfate, is not seen as an environmentally hazardous by-product, although it can be contaminated with trace metals. However, the amounts of lime required can be extremely large. There are often problems where these are mined from attractive sites of great ecological and recreational value. Regenerative desulfurization processes such as the Wellman-Lord procedure absorb SO₂ into sodium sulfite solutions converting them to sodium bisulfite. The SO₂ is later degassed and can be used as a feedstock for the chemical industry (i.e., the production of sulfuric acid).

Catalytic converters in vehicles have been used to remove nitrogen oxides, carbon monoxide, and hydrocarbons from exhaust streams. However, under fuel-rich driving cycles (i.e., lots of accelerating and decelerating), hydrogen gas is produced in the exhaust. Three-way catalysts contain cerium dioxide, which stores sulfur from gasoline under driving conditions as cerium sulfate. Reduction of the cerium sulfate by hydrogen gas allows the formation of H₂S. This can create a noticeable odor, where traffic is heavy.

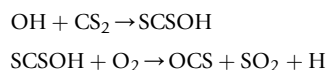
Atmospheric Sulfur Chemistry

Typically the gases we have been discussing are in a somewhat reduced form in the atmosphere and undergo oxidation. DMS is oxidized in an oxygen-rich atmosphere via OH attack:

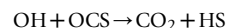


DMSO can yield dimethylsulfone. Methanesulfonic acid is an important product of DMS oxidation although the process is not entirely clear. It has been suggested that this arises from CH₃SO₂ produced in a sequence of reactions that follow hydrogen abstraction by OH from DMS. Once fragmented, the S from DMS can oxidize to SO₂ and sulfuric acid. Lower temperatures as found at high latitude appear to favor the production of larger proportions of DMSO and MSA rather than the more complete oxidation through to SO₂. The oxidation of DMS typically takes place at time scales of about a day. These processes are an important source of sulfuric acid over the ocean and contribute to the production of cloud condensation nuclei. Sulfur in this form is usually called non sea salt sulfate to distinguish it from the sulfate from sea salt.

CS₂ reacts in the atmosphere with a 7 day residence time.

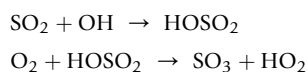


OCS oxidizes only slowly:



However, in the stratosphere the molecule can photofragment into CO and S which can ultimately oxidize to contribute to sulfate particles.

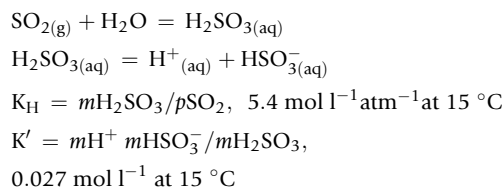
Homogeneous oxidation of SO_2 takes place through OH attack:



The sulfur trioxide reacts very rapidly with water to form sulfuric acid. This can be a significant pathway, but if liquid water is present heterogeneous mechanisms yield most of the sulfate.

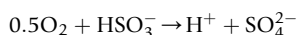
Chemistry of Sulfur in Rain and Cloud Water

The interaction of SO_2 with water is responsible for much of its removal from the atmosphere. SO_2 is not especially soluble in water, but subsequent equilibria increase the partition into cloud water at typical pH values.

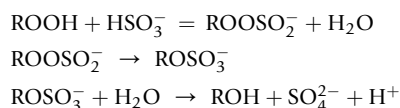


where m and p refer to the molarity and pressure of aqueous and gaseous species, respectively.

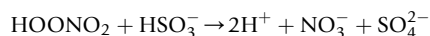
SO_2 dissolves in water as $(\text{SO}_2)_{\text{aq}}$ 'sulfurous acid' (H_2SO_3). It is usually sufficient to treat only the first dissociation to the bisulfite anion (HSO_3^-) as the subsequent dissociation to the sulfite ion is usually small in cloud droplets. The concentrations expressed here in molar or molal units should be as activities, but in rainwater the deviations from ideality are small. At pH 5.4 in a typical cloud with a gram of liquid water in each cubic meter, SO_2 will partition equally into both phases. Once dissolved, sulfur dioxide can be oxidized, but oxidation by dissolved oxygen is typically rather slow. The production of sulfuric acid, which is much stronger, leads to acidification.



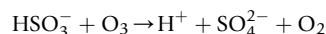
This simple oxidation can be catalyzed by iron, manganese, and other transition metals in the atmosphere, but in unpolluted rainfall the metal concentrations are likely to be low. Here more typically the reaction proceeds with oxidants such as dissolved hydrogen peroxide (or other atmospheric peroxides) and ozone. The hydrogen peroxide route is a particularly significant one as the reaction is faster in acid solution. This means that it does not slow as the system becomes more acidic with the production of sulfuric acid. This oxidation can be represented as follows:



Hydrogen peroxide is an especially important droplet phase oxidant because the gas is very soluble in water and readily produced in atmospheric solutions by photochemical mechanisms. Other oxidants produced in solution, such as peroxynitric acid, appear to be important:



The ozone-driven oxidation is written as



and although ozone is abundant in the atmosphere it is not very soluble.

In general, the overall oxidation in the remote atmosphere takes 2–4 days, but under some conditions it can be much faster. In urban areas plentiful supplies of transition metal catalysts and oxidants can enhance the rates considerably, especially if there are alkaline materials from fly ash or ammonia to neutralize the growing acidity of droplet phases which limits SO_2 solubility.

Land surfaces, particularly those covered by vegetation, represent a further removal process as SO_2 is dry deposited to these surfaces. Although we use the term dry deposition, the vegetated surfaces act as if they are wet and gas exchange takes place with effectively wet surfaces on or inside leaves. Such dry deposition is most significant where SO_2 concentrations are high (polluted regions). The dissolution of SO_2 in water, most particularly sea water (which is alkaline), represents a further important sink.

Acidification

The production of sulfuric acid in air masses over industrialized continents led to the acid rain problem. This acidification was recognized by the mid-nineteenth century, initially as emissions of hydrochloric acid from the alkali and soap industry, but later coal combustion was also seen as an important source of acid. By the 1950s, long-term records of rainfall chemistry showed increasing fluxes of sulfuric acid to the ground. In Scandinavia, poorly buffered soils began to respond to the effects of acid rain and the pH of lakes declined, resulting in a loss of fish. Through the 1970s and more particularly the 1980s, acid rain became a keenly fought environmental issue, which led to a range of protocols in North America and Europe that served to limit the release of sulfur into the atmosphere.

Acids accumulated over the winter are concentrated into the first meltwaters of spring and send an acid shock through the environment. Young fish are readily injured, but amphibians such as frogs and salamanders also suffer. More generally the delicate ecological web that depends on the freshwater systems is affected. The debate over acid rain also examined the effects on forests. Forest damage arises from a complex array of factors that include air pollution (ozone, SO_2 , acid rain, etc.) along with climate stress (i.e., drought), forest management practices, the age of stands, etc. Acidified rain could change the nutrient balance of soils (magnesium, potassium, and calcium) and mobilize toxic metals (aluminum, cadmium, zinc, etc.).

Damage to buildings is frequently seen as an important impact of acid rain, perhaps because we confront our urban fabric almost every day. It is exposed to the ravages of high concentrations of air pollution, which means that buildings have long been disfigured by sooty crusts. Hidden underneath these crusts damage takes place through the oxidation of deposited SO_2 . Stones build up thick gypsum (calcium sulfate) layers and metals are also prone to attack. In general, the deterioration of metal work and, to a lesser extent, stone has declined as SO_2 concentrations have decreased in urban air.

There is now a decrease in the sulfur deposited in rain over Europe. This has led to complaints that some agricultural crops are now sulfur deficient. Although acid rain provided sulfur for crops, it is obviously important to remember its large-scale harmful effects before regretting its gradual elimination. Even though sulfur deposits have decreased in Europe, the acidity of rain has not always fallen to match this decline. Increases in nitric acid in air (from nitrogen oxides produced during combustion) and a decline in the alkaline content of the atmosphere may account for this. Today the acidification of rain has shifted geographically to tropical regions with fast-growing economies. Here there are new problems on ecosystems and land types quite different from those affected in Europe and North America.

Aerosol Sulfur and Climate

Sulfate in the atmosphere is present some of the time in acid aerosols, most commonly sulfuric acid or ammonium bisulfate. The sulfuric acid aerosol is typically present at diameters of 2–3 μm . This makes the particles excellent cloud condensation nuclei (CCN), but also a potential threat to human health when they are found concentrated in urban air. Sulfates, nitrates, and methanesulfonates are frequently found at enhanced concentrations on sea salt aerosols as these strong acids can displace as much as 90% of the chloride as hydrogen chloride, which degases to the atmosphere.

Sulfate aerosols appear to have an important role in controlling climate. The CO_2 -driven greenhouse effect provides a link between increases in global temperature and rising emissions of CO_2 from fossil fuels. SO_2 is also a product of much fossil fuel use and results in the presence

of sulfate particles in the atmosphere. These particles are likely to increase the number of CCN and hence the reflectivity of clouds. This would enhance the albedo of the Earth and lessen the temperature increase imposed by anthropogenic CO_2 emissions. The cooling effect, although difficult to estimate, is only a fraction of the increase from greenhouse gases.

Aerosols are incorporated into the well-known Gaia hypothesis, which sees the Earth as a self-regulator entity, particularly in terms of its temperature. One of the main ways in which atmospheric processes have been interpreted within the Gaia concept has been the potential for DMS emissions to stabilize temperature, against changes imposed by shifts in solar irradiance (Figure 3). This is known as the CLAW hypothesis, after Charlson, Lovelock, Andreae, and Warren, who first suggested it in 1987. It sees long-term climate as part of a feedback loop in which warmer temperatures would increase phytoplanktonic activity in the oceans. This in turn would increase emissions of DMS. Higher DMS concentrations in the atmosphere would increase sulfate containing CCN and thus cloudiness and albedo. This would lead to lower temperatures. It is an attractive and popular vision of the biological mediation in global climate, although proof from ice cores has yet to offer convincing proof for this concept of climate control.

Despite the size of the natural sulfur reservoirs, combustion of fossil fuels, mining, agriculture, and the drainage of large saline basins such as the Aral Sea have had a major impact on the sulfur cycle. Windblown sulfate dusts and acid rain have increased, and although emissions may have shifted away from the North Atlantic sector in recent decades, anthropogenic impacts on sulfur cycles will remain an important consideration in developing regions for some time to come.

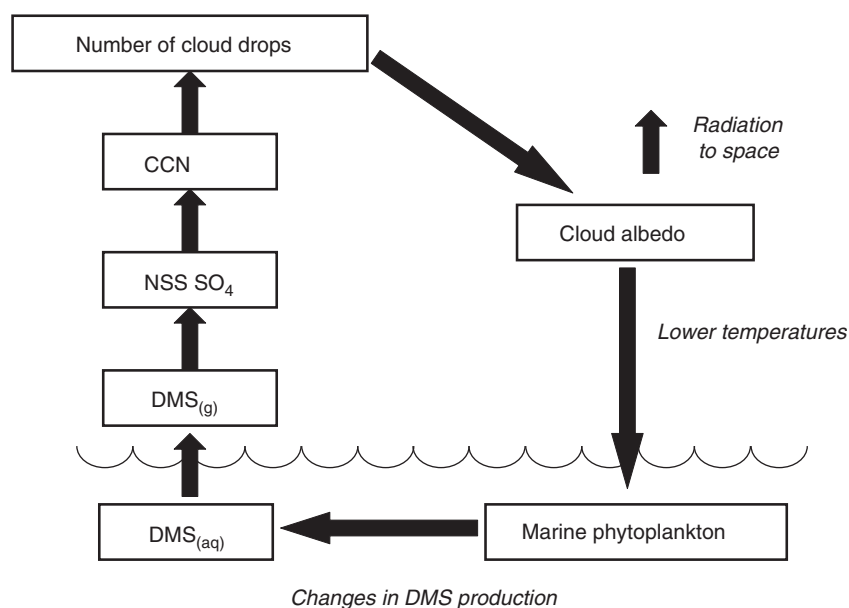


Figure 3 Schematic representation of the CLAW hypothesis. Here phytoplanktonic activity in the oceans affects the emissions of DMS. Its oxidation can form sulfate-containing CCN and increasing cloudiness and albedo, which would lead to lower temperatures. If these lower temperatures cause the phytoplankton to produce less DMS, climate could effectively be controlled.

Further Reading

- Brimblecombe, P., Hammer, C., Rodhe, H., Ryaboshapko, A., Boutron, C.F., 1989. Human influence on the sulfur cycle. In: Brimblecombe, P., Lein, A.Y. (Eds.), *Evolution of the Global Biogeochemical Sulphur Cycle*. Wiley, Chichester.
- Canfield, D.E., Raiswell, R., 1999. The evolution of the sulfur cycle. *American Journal of Science* 299, 697–723.
- Canfield, D.E., Habicht, K.S., Thamdrup, B., 2000. The Archean sulfur cycle and the early history of atmospheric oxygen. *Science* 288, 658–661.
- Charlson, R.J., Lovelock, J.E., Andreae, M.O., Warren, S.G., 1988. Oceanic phytoplankton, atmospheric sulphur, cloud albedo and climate. *Nature* 326, 655–661.
- Findlayson-Pitts, B.J., Pitts, J.N., 2000. *Chemistry of the Upper and Lower Atmosphere*. Academic Press, San Diego.
- Robock, A., 2000. Volcanic eruptions and climate. *Reviews of Geophysics* 38, 191–219.
- Rodhe, H., 1999. Human impact on the atmospheric sulfur balance. *Tellus* 551A–B, 110–122.
- Schlesinger, W.H., 1997. *Biogeochemistry: An Analysis of Global Change*. Academic Press, San Diego.
- Warneck, P., 1999. *Chemistry of the Natural Atmosphere*. Academic Press, San Diego.

Bromine

R von Glasow, University of East Anglia, Norwich, UK

C Hughes, University of York, York, UK

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

Bromine compounds cycle between the oceans, atmosphere, and land and in the atmosphere reactive bromine can lead to substantial destruction of ozone, both in the troposphere and in the stratosphere. Furthermore, bromine compounds play an important role for the oxidation of dimethylsulfide and mercury. The sources, sinks, and atmospheric chemistry of bromine are described in this article.

Overview and Significance of Bromine Cycle

Bromine is the second element of the halogen group and occurs naturally in the environment mainly as bromide salts in small amounts in crustal rock from where it has been leached and then accumulated in the oceans. It is commercially extracted mainly from seawater and Dead Sea waters. It further occurs naturally in soils and is taken up by vegetation and also from ocean water by marine algae, which synthesize a large range of organic bromine compounds which can be released to the ocean and atmosphere. Bromine cycles between the compartments of ocean, atmosphere, and land. In the atmosphere, inorganic bromine quickly cycles between the gas and particulate phase, however many details of this cycling are still ill-quantified. Organic bromine compounds are used commercially as flame retardants, pesticides (fumigants, mainly CH_3Br) and bromine was also used as an antiknocking agent in unleaded fuels. Many of these compounds (especially halons) play an important role in stratospheric ozone depletion (see **Stratospheric Chemistry Topics**: Halogens; Reactive Nitrogen (NO_x and NO_y); HO_x ; Hydrogen Budget; Halogen Sources, Anthropogenic; Halogen Sources, Natural (Methyl Bromide and Related Gases); Overview). In the troposphere bromine plays a role in ozone destruction, the oxidation of dimethylsulfide and elemental mercury. Global models suggest that bromine chemistry is responsible for 5–15% of the loss of tropospheric ozone. The chemistry of bromine is closely linked with that of the halogens chlorine and iodine. Chlorine and bromine, and to a lesser extent iodine, are often coemitted from the oceans and vegetation. Cross-reactions between chlorine, bromine, and iodine speed up their atmospheric cycling and amplify their relevance for atmospheric chemistry. See **Figure 1** for an overview of the most important atmospheric bromine processes.

Sources

The anthropogenic sources of bromine-containing compounds (halons, anthropogenic CH_3Br) are mainly of interest for the stratosphere, and are discussed in detail in the articles, **Stratospheric Chemistry Topics**: Halogens; Reactive Nitrogen (NO_x and NO_y); HO_x ; Hydrogen Budget; Halogen Sources, Anthropogenic; Halogen Sources, Natural (Methyl Bromide

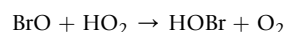
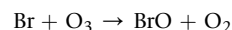
and Related Gases); Overview; therefore, in this section we focus on natural sources. **Table 1** summarizes this information.

Sea Spray

Bromine is present as a trace compound in ocean water (about 0.85 mmol l^{-1}). It is supplied by riverine input and has a residence time of about 10^8 years. Sea spray aerosol is formed by wave action and bursting of gas bubbles that become entrained into the water at the sea surface. This results in the production of small particles that contain seawater. As bromide is present in seawater, production of sea spray aerosol is globally the main source of bromine to the atmosphere and accounts for $3.44\text{--}25.6 \text{ Tg year}^{-1}$ (range due to uncertainty in sea salt flux). The lifetime of particulate bromine in the atmosphere is determined by two factors: the lifetime against deposition of sea spray aerosol and the timescale of photochemical release of bromine to the gas phase. Bromide has been measured in marine aerosol since the 1960s and very often the ratio of $\text{Br}^-:\text{Na}^+$ is smaller in airborne particles than it is in the ocean. This indicates loss of bromine to the gas phase as the exchange of bromine between the gas and particulate phase is well documented from laboratory studies but no mechanism for the release of Na^+ from particles is known. Particulate bromine measurements show very large bromine depletions (10–100%) in aerosol particles with diameters greater than $1 \mu\text{m}$. Interestingly, smaller particles often show an enrichment of bromine compared to seawater which is not understood yet.

Polar Regions

In the early 1980s, sudden dramatic decreases in ozone concentrations were measured in the Canadian Arctic spring boundary layer. Soon they were linked to bromine chemistry but the exact bromine liberation mechanism is still somewhat unclear. These events have since been found to be very widespread both in the Arctic and Antarctic boundary layer in spring. What could be established is that a very efficient autocatalytic reaction cycle leads to the liberation of bromine from condensed phase bromide:



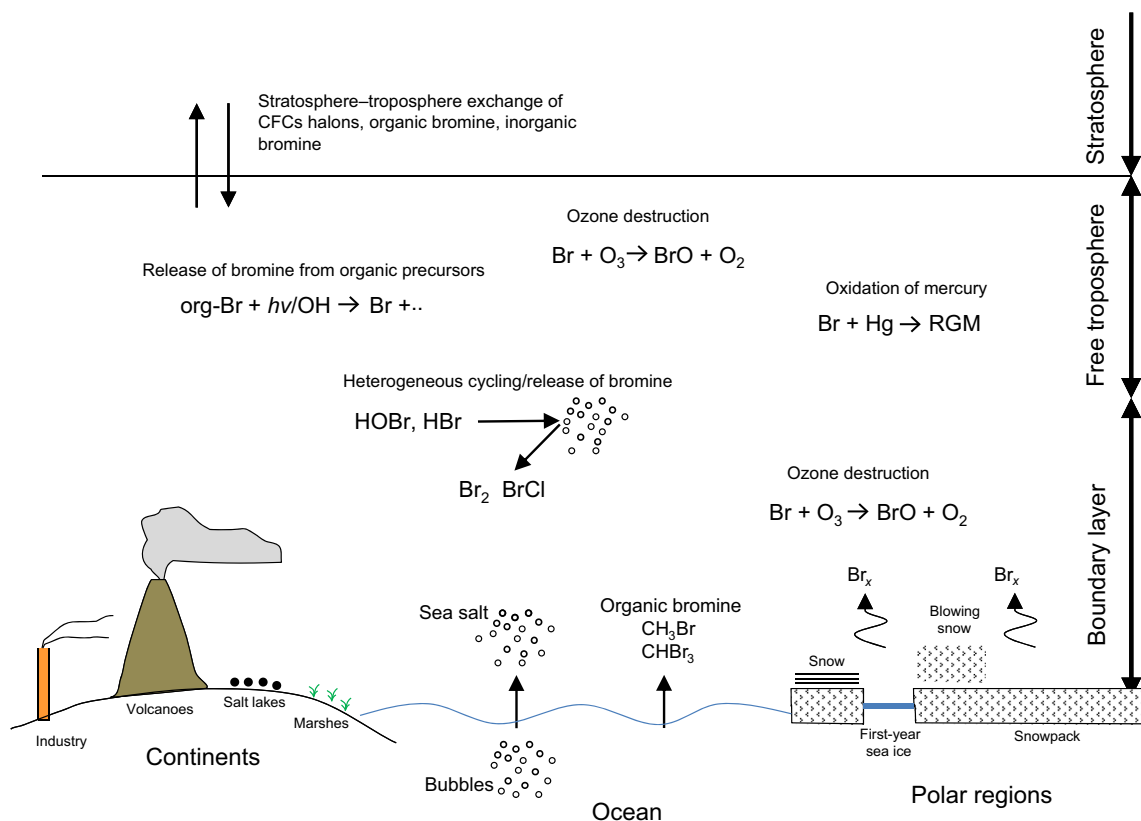
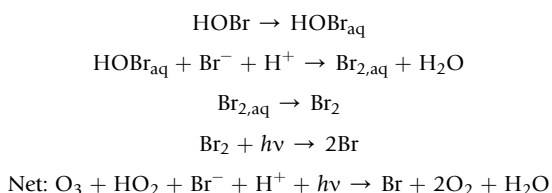


Figure 1 Overview of the most important atmospheric processes involving bromine. RGM stands for reactive gaseous mercury (Hg^{II}) and Br_x for photolyzable bromine compounds. Modified from Saiz-Lopez, A., von Glasow, R., 2012. Reactive halogen chemistry in the troposphere. Chemical Society Reviews 41, 6448–6472.

Table 1 Natural sources of bromine to the atmosphere

	Global source strength (Tg (Br) year^{-1})	Comment
Sea salt aerosol	3.44–25.6	Estimated from Br/Na ratio in seawater and sea salt fluxes from literature
Volcanoes	0.005–0.015	Pyle and Mather (2009)
Ocean exchange of CH_3Br	Source: 0.035 Sink: 0.041 Net: –0.006	WMO (2011)
Ocean exchange of CH_2Br_2	0.057–0.28	WMO (2011)
Ocean exchange of CHBr_3	0.12–1.4	WMO (2011)
Land-based vegetation, incl. salt marshes (CH_3Br)	0.014	WMO (2011)
Biomass burning (CH_3Br)	0.024	WMO (2011)
Fumigation (CH_3Br)	0.012	WMO (2011), Ziska et al. (2013), see text

Local lifetimes: $\text{CH}_3\text{Br} \sim 1$ year; $\text{CH}_2\text{Br}_2 \sim 120$ days; $\text{CHBr}_3 \sim 24$ days.



In this reaction cycle, ozone is destroyed very rapidly. Furthermore the uptake of one bromine atom (in the form of

HOBr) leads to the release of two bromine atoms (in the form of Br_2), hence this autocatalytic cycle has been named the ‘bromine explosion’ mechanism due to the ‘explosive’ growth of gas phase bromine concentrations. The source of the bromide that is involved in this reaction cycle is the ocean. Initially marine biogenic organic bromine gases were suspected to be the source of bromine, but this was soon refuted. Under some conditions however biogenic bromine gases can play a role as an additional source of bromine (see below). Upon freezing of seawater the salts contained in it are expelled from

the ice lattice and largely drain into the ocean, but are also present as deposits on top of first-year sea ice. It is possible that the bromine explosion happens on these deposits. Comparisons of satellite images of BrO and first-year sea ice have shown a strong spatial correlation. Furthermore, release from surface and blowing saline snow has been suggested as an important source of bromine in polar regions. Snow and ice as well as airborne particles can also act as substrates where the bromine explosion takes place. These condensed phases furthermore act to convert less reactive bromine compounds (e.g., HBr) to photolyzable bromine species (e.g., Br₂). Reactive bromine compounds have also been measured in the polar free troposphere pointing to vertical transport either upwards from the boundary layer into the free troposphere (e.g., through convection over regions of relatively warm open ocean, so-called open leads) or downwards from the stratosphere. The chemistry of bromine and chlorine is closely coupled and a role for chlorine in bromine release has been suggested to be of relevance at least under some conditions. Figure 2 summarizes the sources of bromine in the spring polar boundary layer.

Salt Lakes

In salt lakes large areas of salt are exposed to the atmosphere either as evaporites or highly saline surface waters. The bromide content and acidity of salt lakes varies considerably. Very strong bromine explosions have been found in the Dead Sea area with BrO mixing ratios of up to 200 ppt. Model calculations were able to reproduce the measured BrO values when they assumed the 'bromine explosion' mechanism to take place in large areas of exposed salt as well as the Dead Sea surface waters; bromine release from saline aerosol is not enough to reproduce the measured BrO values. Smog chamber experiments were able to reproduce the release of

photolyzable bromine from dry salt pans. BrO was also measured above other salt lakes such as the Salar de Uyuni in Bolivia. BrO was not detected above some salt lakes where measurements were made, possibly hinting at the bromine release being a function of the pH of the soil/salt lake (low pH leading to more release). Salt lakes might also play a role as source of natural halogenated organic gases, but this remains to be quantified.

Salt lakes provide unique 'natural laboratories' to study tropospheric halogen chemistry and bromine chemistry in particular but a more than regional impact on tropospheric chemistry is unlikely.

Volcanoes

The emission of HBr from volcanoes, especially arc volcanoes, has been known for a fairly long time, but the conditions in volcanic plumes were assumed to be rather unreactive except for an equilibration between the gas and particulate phase. In the early 2000s, however, BrO was detected in the plume of the passively degassing volcano Soufrière Hills on Montserrat in the Caribbean. Since then BrO was detected in the plumes of other volcanoes, both from passive degassing (e.g., Etna, Sicily; Masaya, Nicaragua; Villarica, Chile; Nyiragongo, Congo; Mt Erebus, Antarctica) but also from explosive eruptions (e.g., Kasatochi, Alaska; Eyjafjallajökull, Iceland). Satellite investigations were able to detect BrO in the plumes of many but not all probed volcanoes, which can likely only partly be explained by the coarse ground pixels and other instrumental limitations.

In order to account for the dilution of the plume with background air, often the ratio of a reactive compound, such as BrO, and a very unreactive compound such as SO₂ is used to investigate chemical processing in volcanic plumes as taking

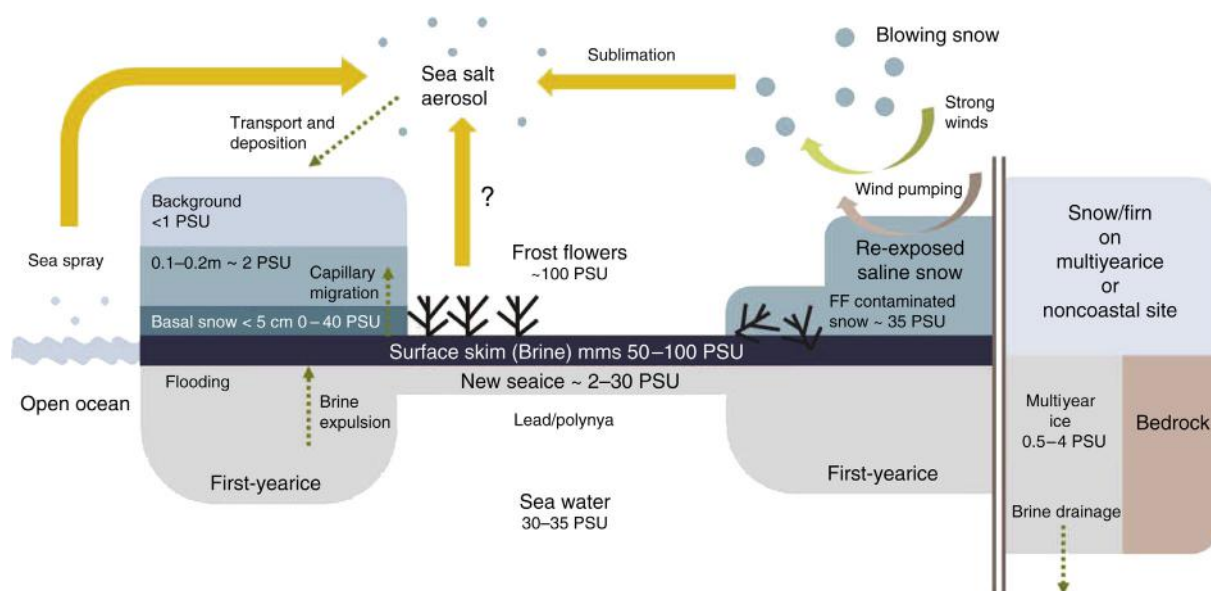


Figure 2 Sources of halides in the polar boundary layer. A range of salinity values is indicated for the various reservoirs. PSU stands for practical salinity unit. From Abbatt, J.P.D., Thomas, J.L., Abrahamsson, K., Boxe, C., Granfors, A., Jones, A.E., King, M.D., Saiz-Lopez, A., Shepson, P.B., Sodeau, J., Toohey, D.W., Toubin, C., von Glasow, R., Wren, S.N., Yang, X., 2012. Halogen activation via interactions with environmental ice and snow. *Atmospheric Chemistry and Physics* 12, 6237–6271, reproduced with permission of the authors.

this ratio effectively cancels out dilution in the atmosphere. At several volcanoes the BrO/SO₂ ratio in the volcanic plume was very low near the crater rim but increased with distance (order of several 10 km). This shows that BrO is being produced in the early plume. Model calculations confirmed that the bromine explosion mechanism involving volcanic HBr and acidic aerosol particles can explain the observed BrO and the rise in BrO/SO₂ ratio. These studies also showed that bromine cannot be present only as HBr when emitted by the volcano but that some divergence from thermodynamic equilibrium producing e.g., Br₂ must occur, likely within the crater where oxygen-rich air is mixed with hot volcanic volatiles, in order to explain the very rapid rise in the BrO/SO₂ ratio.

The global source of bromine from volcanoes is rather small compared to other sources (see Table 1); however, the emissions from most volcanoes are directly into the free troposphere or stratosphere where the atmospheric lifetime of inorganic bromine is much longer than in the boundary layer.

Ocean: Organic Gases

It is well established that bromine-containing organic gases (e.g., CH₃Br, CHBr₃, CH₂Br₂, CH₂IBr) are produced in the marine environment. Once in the atmosphere these gases are broken down by photolysis or chemical reactions, which release reactive bromine. The organic bromine gases have a wide range of lifetimes in the atmosphere ranging from a year for CH₃Br to weeks for CHBr₃ and about 1 h for CH₂IBr. Information on CH₃Br, the longest lived of the organic bromine gases, can be found in Section Stratospheric Chemistry and Composition. The most abundant short-lived organic gases are CHBr₃ and CH₂Br₂ and the formation pathways for these compounds are believed to be principally biological. Marine seaweeds and diatoms are known sources of CHBr₃ and CH₂Br₂. The main production pathway is believed to involve vanadium- or heme-dependent haloperoxidases. These enzymes catalyze the breakdown of H₂O₂ through the two-electron oxidation of bromide yielding hypobromous acid (HOBr), which can then brominate electron-rich organic substrates. In marine environments the oxidation of bromide is primarily catalyzed by bromoperoxidases. The physiological role of haloperoxidase activity in marine organisms is yet to be established, but it has been proposed that they are involved in defense against oxidative stress and/or their inorganic or organic products act as grazing deterrents. A functional bromoperoxidase has recently been identified in strains of the marine cyanobacterium *Synechococcus* sp. so these organisms may also be involved in organic bromine gas formation.

The organic bromine gases are typically found at picomolar concentrations in seawater with CHBr₃ generally found at the highest concentrations (up to 10³ pM). CHBr₃ and CH₂Br₂ have been found to be supersaturated over large areas of the oceans with their concentrations and emission rates highly variable over spatial and temporal scales. Mean background tropospheric mixing ratios of CHBr₃ and CH₂Br₂ are in the range of 1–2 ppt. Emission source strengths for the organic bromine gases from the oceans currently have large uncertainty with estimates ranging from 120 to 1400 Gg Br year⁻¹ for CHBr₃ (note that the upper limit is likely too high) and 57–280 Gg Br year⁻¹ for CH₂Br₂. On a global scale this is

considerably lower than that from sea salt aerosol (Table 1), but the contribution from organic bromine gases to total inorganic bromine in the atmosphere is likely to be more important on a regional scale.

Anthropogenic Sources

The main anthropogenic uses of bromine are as flame retardant (mainly halons), pesticide (mainly CH₃Br), and additive to petrol as an antiknocking agent. Due to its role in stratospheric ozone depletion the use of halons and anthropogenic CH₃Br has been limited and declined substantially. This is described in detail in articles **Stratospheric Chemistry Topics: Halogens; Reactive Nitrogen (NO_x and NO_y); HO_x; Hydrogen Budget; Halogen Sources, Natural (Methyl Bromide and Related Gases); Overview**. Further anthropogenic sources of bromine include water purification. In some countries bromine is used as an agent for water purification, but even if drinking water is mainly chlorinated, bromine will be present as impurities. The same is true for emissions from cooling towers where cooling water is treated with halogens to limit algal growth.

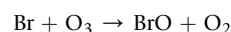
Loss

As mentioned above, inorganic bromine compounds can cycle very quickly between the gas phase and particulate phase so that uptake of the fairly unreactive HBr on to particles is not necessarily the end of the atmospheric life cycle as it can be released again from particles in reaction cycles such as the bromine explosion mechanism. Eventually the loss of bromine compounds however is either through dry deposition of gases or bromine-containing particles to the surface or wet deposition (wash out) of particulate bromine or scavenging of soluble bromine gases by precipitation. Bromine can then be stored in soils or vegetation, from where it can again be released to the atmosphere, or is transported to the oceans through rivers.

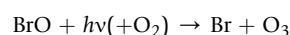
The exchange of brominated compounds between the troposphere and stratosphere is often referred as sink for one and source for the other atmospheric compartment.

Tropospheric Chemistry

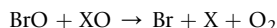
The tropospheric chemistry of bromine is closely linked to that of chlorine and iodine. Furthermore large similarities are present to stratospheric bromine chemistry. The cycling between the gas and condensed phase (such as aerosol particles, cloud droplets, saline solutions on the surface) is an important element of halogen cycling. The 'bromine explosion' mechanism was already explained; in the following further important reaction cycles are shown. X symbolizes a second halogen atom (Cl, Br, I). Bromine atoms react very quickly with O₃ producing BrO:



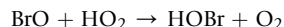
Even though this reaction destroys O₃, the most likely fate of BrO is photolysis:



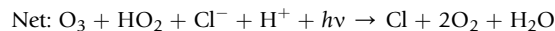
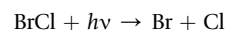
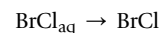
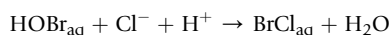
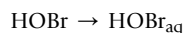
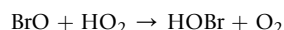
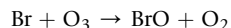
From this reaction sequence it becomes obvious that it is more helpful to think in terms of the odd oxygen family ($O_x = O_3 + O + O(^1D) + NO_2 + 2NO_3 + 3N_2O_5 + HNO_4 + ClO + 2Cl_2O_2 + 2OCLO + BrO + IO + 2OIO$) when quantifying ozone loss. Odd oxygen is only lost if the product of a reaction involving odd oxygen, such as BrO , does not produce another member of the odd oxygen family. Hence the following reactions are 'real' ozone, or, more precisely, odd oxygen loss reactions:



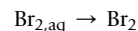
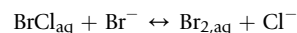
or



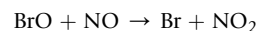
Note that in the first reaction two members of O_x are being destroyed. This reaction also shows the link between various halogens; the reaction rates of the interhalogenoxide reactions are very fast, leading to very rapid ozone loss. Halogen-induced ozone loss has been suggested to be responsible for about 30% of photochemical ozone loss in the marine boundary layer and 5–20% in the global troposphere. Another cross-reaction sequence, similar to the bromine explosion, was shown to occur in the aqueous phase:



Instead of degassing, $BrCl_{aq}$ can also undergo the following sequence, depending on the relative concentrations of Cl^- and Br^- in the condensed phase, leading to the release of Br_2 :



In regions with high NO_x , $BrONO_2$ can constitute a significant fraction of gas phase bromine. It can be photolyzed, decompose or react on surfaces producing $HOBr$, $BrCl$, or Br_2 depending on the surface it reacts on. Under high NO_x conditions the reaction



can shift the NO to NO_2 ratio. Even though this reaction does not constitute a sink of O_x , it shortcuts the formation of O_3 as under most other circumstances the oxidation of NO to NO_2 is a gain of O_x . $HOBr$, which is produced in the reaction of BrO with HO_2 , can in addition to uptake on a surface also be photolyzed:

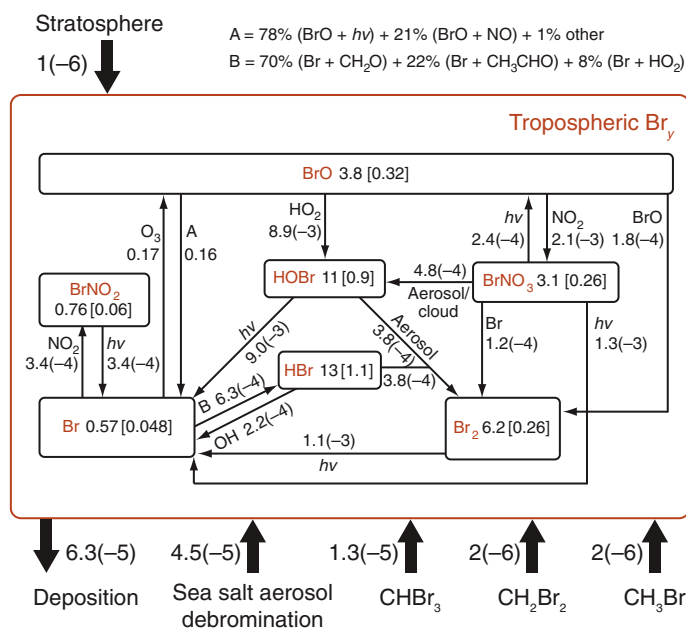
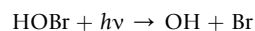
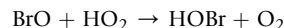


Figure 3 Global annual mean budget of tropospheric inorganic bromine (Br_y) in a state-of-the-art global atmospheric chemistry transport model (GEOS-Chem). The main reactions are indicated. Inventories are given as masses (Gg Br), with mixing ratios ($pmol\ mol^{-1}$) in square brackets. Rates are given in units of $Gg\ Br\ s^{-1}$. Read 6.3(-5) as 6.3×10^{-5} . HBr accounts for 55% of Br_y loss by deposition. Sea salt aerosol debromination is the dominant global source of Br_y , but is mainly confined to the marine boundary layer, where Br_y has a short lifetime against deposition. It accounts for 48% of the Br_y source in the global free troposphere. From Parrella, J.P., Jacob, D.J., Liang, Q., Zhang, Y., Mickley, L.J., Miller, B., Evans, M.J., Yang, X., Pyle, J.A., Theys, N., Van Roozendaal, M., 2012. Tropospheric bromine chemistry: implications for present and pre-industrial ozone and mercury. *Atmospheric Chemistry and Physics* 12, 6723–6740, reproduced with permission of the authors.

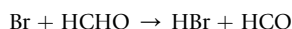
Table 2 Summary of tropospheric measurements of reactive bromine species

Species	Location	Typical mixing ratios (pmol/mol)	Max mixing ratio (pmol/mol)
BrO	Arctic BL	5–20	41
	Arctic free troposphere	5	10
	Summit, Greenland, BL	1–5	5
	Antarctic BL	5–20	20
	Mace Head (Ireland)		6.5
	Brittany, France		7.5
	North Atlantic		2.4
	Cape Verde Islands	2.5	5.6
	Mauritanian Coast		10
	Dead Sea, Israel	20–120	200
	Great Salt Lake, Utah		6
	Salar de Uyuni, Bolivia		~22
	Tropospheric volcanic plumes		~1000
	High- and midlatitudes FT	1	1.5
HOBr	Arctic BL	10	~260
Br ₂	Arctic BL	13	~140
	Antarctic BL	6	45
BrCl	Coastal California, USA	2	19
	Arctic BL	5–15	~35
	Antarctic BL	1	6
Soluble bromine	Arctic BL	10–30	40
	Summit, Greenland		4
Br _{inorg}	Arctic BL	0–20	80
	Off West Africa		~6.4
	Hawaii	4	9
	Atlantic	2–20	64
CH ₃ Br	Marine BL	6.5–8	12
	Free troposphere	6.5–8	12
CH ₂ Br ₂	Marine BL	1.1	1.5
	Free troposphere	0.86	1.2
CHBr ₃	Marine BL	1.6	30
	Free troposphere	0.5	1.2

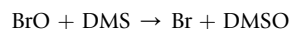
BL; boundary layer. After Saiz-Lopez, A., von Glasow, R., 2012. Reactive halogen chemistry in the troposphere. Chemical Society Reviews 41, 6448–6472. See this article for details and references.

This sequence changes the OH to HO₂ ratio, which is also important for ozone formation.

Unlike chlorine atoms, bromine atoms cannot efficiently attack alkanes so only the reactions with aldehydes such as HCHO are important links between (oxygenated) volatile organic compounds and bromine:

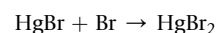
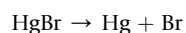
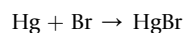


BrO is an efficient oxidant of dimethylsulfide (CH₃SCH₃, DMS), a biogenic gas released from the oceans:



The breakdown products of DMS can lead to the growth of existing particles in the atmosphere and under certain conditions the nucleation of new particles. Further links between the bromine and sulfur cycles have been suggested involving the oxidation of S(VI) to S(VI) in the aqueous phase.

Strong links between the cycles of mercury and bromine have been found. Mercury is a toxic gas but in its main atmospheric reservoir, elemental Hg (Hg⁰), it is mostly harmless and has a lifetime of about a 1 year. However, if oxidized by OH or bromine:



mercury is rather soluble and the atmospheric lifetime is reduced to days. Once deposited it can be methylated and enter the food chain where it accumulates and causes most damage in species that are higher up in the food chain.

Summary

Bromine has very strong natural (e.g., sea spray aerosol, salty surfaces in polar spring, volcanoes, salt lakes, organic brominated gases) and anthropogenic (e.g., halons, CHBr₃) sources. Its chemistry is linked to that of chlorine and iodine. In the troposphere bromine plays a role for ozone destruction, DMS and mercury oxidation and in the stratosphere it plays a key role in ozone destruction.

Figure 3 shows the tropospheric budget of bromine according to a recent global modeling study. Table 2 provides an overview of tropospheric measurements of reactive bromine species.

For stratospheric chemistry, often use is made of the terms product gas (PG) and source gas (SG). SGs are those that have not undergone any chemical change since emission whereas

PGs are breakdown products. The inorganic gases mentioned here are all PGs but transport into the stratosphere is limited due to their usually rather large solubilities especially when compared to organic SGs and PGs. See the Stratospheric composition and chemistry section for more details.

See also: **Biogeochemical Cycles:** Biogeochemistry of Iodine.
Stratospheric Chemistry Topics: HO_x; Halogen Sources, Natural (Methyl Bromide and Related Gases); Halogens; Hydrogen Budget; Overview; Reactive Nitrogen (NO_x and NO_y).

Further Reading

- Abbatt, J.P.D., Thomas, J.L., Abrahamsson, K., Boxe, C., Granfors, A., Jones, A.E., King, M.D., Saiz-Lopez, A., Shepson, P.B., Sodeau, J., Toohey, D.W., Toubin, C., von Glasow, R., Wren, S.N., Yang, X., 2012. Halogen activation via interactions with environmental ice and snow. *Atmospheric Chemistry and Physics* 12, 6237–6271.
- Parrella, J.P., Jacob, D.J., Liang, Q., Zhang, Y., Mickley, L.J., Miller, B., Evans, M.J., Yang, X., Pyle, J.A., Theys, N., Van Roozendaal, M., 2012. Tropospheric bromine chemistry: implications for present and pre-industrial ozone and mercury. *Atmospheric Chemistry and Physics* 12, 6723–6740.
- Pyle, D.M., Mather, T.A., 2009. Halogens in igneous processes and their fluxes to the atmosphere and oceans from volcanic activity: a review. *Chem. Geol.* 263, 110–121.
- Saiz-Lopez, A., von Glasow, R., 2012. Reactive halogen chemistry in the troposphere. *Chemical Society Reviews* 41, 6448–6472.
- Simpson, W.R., von Glasow, R., Riedel, K., Anderson, P., Ariya, P., Bottenheim, J., Burrows, J., Carpenter, L.J., Frieß, U., Goodsite, M.E., Heard, D., Hutterli, M., Jacobi, H.-W., Kaleschke, L., Neff, B., Plane, J., Platt, U., Richter, A., Roscoe, H., Sander, R., Shepson, P., Sodeau, J., Steffen, A., Wagner, T., Wolff, E., 2007. Halogens and their role in polar boundary-layer ozone depletion. *Atmospheric Chemistry and Physics* 7, 4375–4418.
- WMO Scientific Assessment of Ozone Depletion, 2011. World Meteorological Organization Global Ozone Research and Monitoring Project—Report No. 52, 2011.
- Ziska, F., Quack, B., Abrahamsson, K., Archer, S., Atlas, E., Bell, T., Butler, J., Carpenter, L.J., Harris, N.R.P., Hepach, H., Heumann, K., Hughes, C., Kuss, J., Kruger, K., Liss, P., Moore, R., Orlikowska, A., Raimund, R., Reeves, C.E., Reifenhäuser, W., Tanhua, T., Teigtmeier, S., Turner, S.M., Wang, L., Wallace, D., Williams, J., Yamamoto, Y., Yvon-Lewis, S., Yokouchi, Y., 2013. Global sea-to-air flux climatology estimates of bromoform, dibromomethane and methyl iodide. *Atmospheric Chemistry and Physics* 13, 8915–8934.

Heavy Metals

TD Jickells and AR Baker, University of East Anglia, Norwich, UK

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

Metals, other than mercury (0), travel in the atmosphere associated with aerosols. Emission sources include soil dust and sea spray, as well as anthropogenic emissions. Source strength estimates are presented and these show that the relative importance of anthropogenic and natural sources are very different for different metals. This difference, along with atmospheric cycling discussed, leads to differences in aerosol size distribution, atmospheric residence time, solubility in rainwater and environmental impacts. Hg(0) is gaseous, behaves very differently to other metals with a much longer atmospheric residence time.

Introduction

'Heavy metals' is a widely used but poorly defined term. This article will consider all metallic elements in the periodic table except those in groups 1 and 2. This definition will include metals, such as aluminum, which are not heavy in terms of their atomic weight but are emitted, transported, transformed, and deposited via the atmosphere in a similar way to other true heavy metals. Mercury is the only metal to behave very differently since it exists in the atmosphere in the gas phase rather than on aerosols. In the context of global biogeochemical cycling, atmospheric transport of heavy metals is a major transport route and one that has for some metals been significantly perturbed by human activity. The following sections will briefly describe the sources of some heavy metals to the atmosphere, their transport and cycling through the atmosphere, and their subsequent deposition, and will also consider the biogeochemical significance of the atmospheric transport rate. Because of its fundamentally different behavior, a separate section is devoted to mercury.

Sources

Metals are emitted to the atmosphere from a wide range of sources. Some of these emissions are entirely natural, such as from volcanoes or from biological emissions from land and the oceans. Others arise from natural processes that may have been

changed by human activities such as the formation of wind-blown dust or biomass burning. Similarly, sea salt formation is a natural process, but its significance as a source of metals to the atmosphere may have been modified by human perturbation of the concentrations of these metals in seawater, as will marine gaseous mercury emissions. Other sources of metals to the atmosphere are essentially entirely anthropogenic, arising from combustion processes and industrial activity.

Table 1 lists best estimates of fluxes to the atmosphere from the known sources for a group of metals whose fluxes are thought to have been significantly modified by human activity. In general fossil fuel combustion and metal production dominate the anthropogenic emission sources of these metals. The table is not intended to represent a comprehensive list of sources. Indeed, for many metals we do not know sources well. Furthermore, fluxes from human activity can change dramatically with time. This is particularly evident in the case of lead. Emissions of lead worldwide, for instance, have increased more than 25-fold. Even in remote areas of Europe, it is possible to show increasing lead concentrations in sediments dating back 2000 years, associated with its use by the Romans. Concentrations increased further as populations and industrialization developed over the last 1000 years. The biggest increase, from the 1950s, was predominantly due to the use of lead additives in automobile fuels. This source has declined very rapidly from the 1980s onward as a result of the removal of such additives. Measurements of the isotopic composition of lead have been a particularly powerful tool for tracing the changes in lead

Table 1 Worldwide emissions of trace metals (10^9 g per year)

Source	Arsenic	Cadmium	Copper	Lead	Selenium	Zinc
Wind-blown soil	2.6	0.2	8	3.9	0.2	19
Sea spray	1.7	0.1	3.6	1.4	0.5	0.4
Volcanoes	3.8	0.8	9.4	3.3	0.9	9.6
Forest fires	0.2	0.1	3.8	1.9	0.3	7.6
Continental biogenic emissions	1.6	0.2	2.9	1.5	3.7	5.1
Marine biogenic emissions	2.3	0.05	0.4	0.2	4.7	3.0
Total natural emissions	12	1.3	28	12	10.3	45
Total anthropogenic emissions in the mid-1990s	5	3	26	119	4.6	57

Natural emissions based on Nriagu, J.O., 1989. A global assessment of natural sources of atmospheric trace metals. *Nature* 338, 47–49 and anthropogenic emissions on Pacyna, J.M., Pacyna, E.G., 2001. An assessment of global and regional emissions of trace metals to the atmosphere from anthropogenic sources worldwide. *Environmental Reviews*. 9, 269–298.

sources. The use of isotopic measurements in other metals, to trace their sources, transport, and transformation processes, is beginning to be developed, with recent results published for Fe, Nd, and Hf.

Thus **Table 1** serves only to estimate the wide range of specific natural sources that are important for at least some elements, and the scale of perturbations of the total fluxes arising from human activity. The elements listed in **Table 1** have been selected to represent those for which we know that human perturbations are large. For many other metals such as iron, aluminum, and manganese, perturbations appear to be much smaller. This is illustrated in **Table 2**. Here the data are based on direct measurements of the metals in the atmosphere, which are expressed as enrichments of the observed concentrations over those predicted, if soil dust was the only source. As is evident from **Table 1**, this assumption of a soil source is simplistic, but large enrichments such as seen in **Table 2** for As, Cd, Cu, Pb, Se, and Zn do demand that there are large sources in addition to soil dust. In the case of a few metals such as selenium, it is likely that other natural sources, such as biological emissions of alkylated selenium compounds from seawater, are responsible for the enrichments. However, in most cases it is generally accepted that the large enrichments reflect the fact that anthropogenic emissions are now of comparable magnitude to natural sources for some metals as illustrated in **Table 1**.

Transport and Transformation

Metals emitted to the atmosphere with soil dust retain this association and hence, like soil dust itself, are associated with relatively large aerosols of an equivalent aerodynamic mass median radius of 0.5–2 μm during long-range atmospheric transport. Large-sized particles will be found close to sources or occasionally during highly energetic long-range transport. Metals derived from bursting of bubbles in seawater are associated with sea salt particles that typically are again associated with larger aerosol particles and have radii of 1–5 μm . By

contrast to these metals associated with aerosol particles formed directly by physical processes, metals released as gases (either of biogenic origin or arising from hot combustion processes) condense onto existing particles on the basis of surface area. Hence, such metals will be associated with finer modes of aerosol.

During aerosol transport, cloud cycling can induce changes in size distribution, as different aerosol particles are incorporated into cloud droplets that subsequently evaporate to produce a mixed aerosol particle. Despite this complication, it is remarkable that the aerosol size distribution appears to be very effectively retained over very long transport distances, as illustrated by data from Enewetak Atoll in the central Pacific Ocean shown in **Table 3**, where the size distributions of metals with predominantly crustal and anthropogenic origins are still clearly different from each other, but similar to those of these metals close to sources, even at this remote location, thousands of kilometers from their terrestrial sources.

During atmospheric transport, aerosols containing metals will become hydrated and the metals may become solubilized. This solubilization process is highly pH-dependent, with all metals being more soluble at low pH. Crustal aerosol itself is often alkaline and hence high dust loadings may suppress solubility. In contrast, anthropogenic emissions of gases such as SO_2 and NO/NO_2 can acidify atmospheric aqueous solutions and thus promote solubility. Indeed, some trace metals are known to be able to catalyze oxidation of SO_2 , and hence a positive feedback enhancing solubility is possible. During long-range transport, aerosols will be cycled through clouds a number of times, thereby encountering a range of pH values that may include highly acidic cloud waters. There is evidence to suggest that such pH cycling is not completely reversible and that repeated cycles enhance the solubility of metals at a particular final pH relative to the initial aerosol solubility at the same pH.

In the case of iron, photochemical reactions can lead to the formation of OH radicals (eqn [I]), which are powerful oxidizing agents that link these metals to the cycling of many other components in the atmosphere. These reactions can also modify iron solubility because of the much greater solubility of Fe(II) compared to Fe(III).

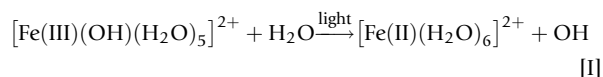


Table 2 Average enrichments of metals in aerosols over the value expected, if soil dust is the dominant source based on samples collected from many sites around the world

<i>Metal</i>	<i>Enrichment</i>
Aluminum	1
Arsenic	190
Cadmium	1100
Cobalt	1.9
Copper	25
Iron	1.3
Lead	320
Manganese	1.5
Selenium	3500
Silicon	0.8
Zinc	50

Based on Wiersma, G.B., Davidson, C.I., 1986. Trace metals in the atmosphere of rural and remote areas. In: Nriagu, J.O., Davidson, C.I. (Eds.), *Toxic Metals in the Atmosphere*. Wiley, New York.

Table 3 Mass median radius of some aerosol trace metals at Enewetak in the remote North Pacific

<i>Metal</i>	<i>Mass median radius (μm)</i>
Sodium	4.3
Aluminum	1.0
Iron	1.1
Manganese	1.3
Copper	0.3–0.4
Lead	0.3–0.6
Zinc	0.5–1.0

Based on Arimoto, R., Duce, R.A., Ray, B.J., Unni, C.K., 1985. Atmospheric trace elements at Enewetak Atoll: 2. Transport to the ocean by wet and dry deposition. *Journal of Geophysical Research*. 90, C22391–C22408.

A further process that may significantly modify the behavior of trace metals during atmospheric transport and transformation is organic complexation. This process has been known to be important for many years in soil, freshwaters, and marine waters. It is now becoming clear that organic material represents a substantial component of many aerosols. This organic material is poorly characterized, but it is clear that some of the organic compounds, including relatively simple organic molecules such as oxalic acid as well probably as some more complex organic material, can chemically bind metals and significantly modify the metals' solubility, bioavailability, and photochemistry.

Deposition

Trace metal removal processes from the atmosphere are similar to those of other aerosol species, and involve wet and dry deposition processes. These are discussed elsewhere in this Encyclopedia and will not be described in detail here. However, a few important points arise that are of specific relevance to metals. First, wet and dry deposition processes are dependent on aerosol particle size. Metals associated with coarser material will therefore be removed more rapidly from the atmosphere than metals associated with finer aerosol particles. The component of the metals in aerosol arising from gas-to-particle conversion, predominantly the anthropogenic component, is therefore less efficiently removed than those components associated with soil dust or sea spray. Thus anthropogenic emissions in general not only increase emissions of metals to the atmosphere but also promote their long-range transport. This is probably one reason why the enrichments seen in Table 2 are greater than predicted from the ratio between natural and anthropogenic emissions (Table 1).

The effect of pH on metal solubility was noted earlier. This results in marked changes in solubilities of metals over rather narrow pH ranges that vary for all metals (Figure 1). This is important because the impact of the metals on the environment varies markedly depending on whether the metal is soluble or insoluble. However, this solubility will also depend on the pH of the receiving media, which will generally be less acidic than rainwater.

Mercury

Mercury is unique among metals because it exists predominantly in the gas phase in the atmosphere as Hg^0 . In this form it has long residence time of about 1 year compared to only a few

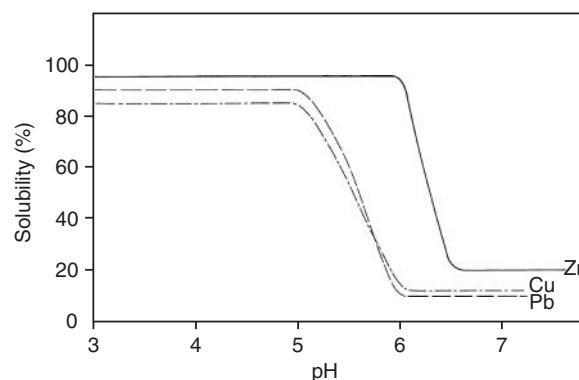


Figure 1 Effect of pH on the solubility (mass/mass) of copper, lead, and zinc in rainwater. Based on Jickells, T.D., 1997. Atmospheric inputs of some chemical species to the North Sea. *Ger. J. Hydrogr.* 49, 111–118.

days for aerosol-bound metals (including oxidized forms of mercury). Mercury has substantial, natural, terrestrial, biological (2.5×10^9 g per year), and marine (2.7×10^9 g per year) sources, together with a large anthropogenic emission (2.3×10^9 g per year). Atmospheric oxidation reactions slowly convert Hg^0 to $\text{Hg}(\text{II})$, in which form it is essentially nonvolatile and is removed rapidly to aerosols and deposited.

Biogeochemical Significance of Atmospheric Transport of Trace Metals

As noted earlier, the emissions of many metals to the atmosphere have been estimated to be increased markedly by human activity and are now a major route for global transport (Table 4). Atmospheric emissions, particularly fine-mode aerosols, lead to very effective long-range transport, as is evident from the enrichment of several trace metals in even the remote atmosphere (Table 2). The complex record of increasing atmospheric concentrations arising from industrialization is preserved in the ice caps, particularly of the Northern Hemisphere, and in corals in central ocean regions remote from riverine sources (Figure 2). Thus it is clear that atmospheric transport has resulted in contamination of remote environments by trace metals, though the evidence for deleterious effects arising from their contamination, and hence for pollution, is equivocal. Recent studies have documented that measures to reduce emissions of lead via its elimination from vehicle fuels have been very successful with concentrations now

Table 4 A comparison of atmospheric and riverine fluxes (10^9 g per year) of some trace metals to the oceans

Metal	Fluvial dissolved flux	Fluvial particulate flux	Atmospheric dissolved flux	Atmospheric particulate flux
Iron	1100	110 000	3200	29 000
Copper	10	1500	14–45	2–7
Lead	2	1600	80	10
Zinc	6	3900	33–70	11–60

Note particulate material is deposited in estuaries and hence the best measure of comparison in terms of impacts on the open ocean is probably of the total atmospheric flux and the riverine dissolved flux.

Based on Duce, R.A., Liss, P.S., Merrill, J.T., et al., 1991. The atmospheric input of trace species to the world ocean. *Global Biogeochemical Cycles* 5, 193–259.

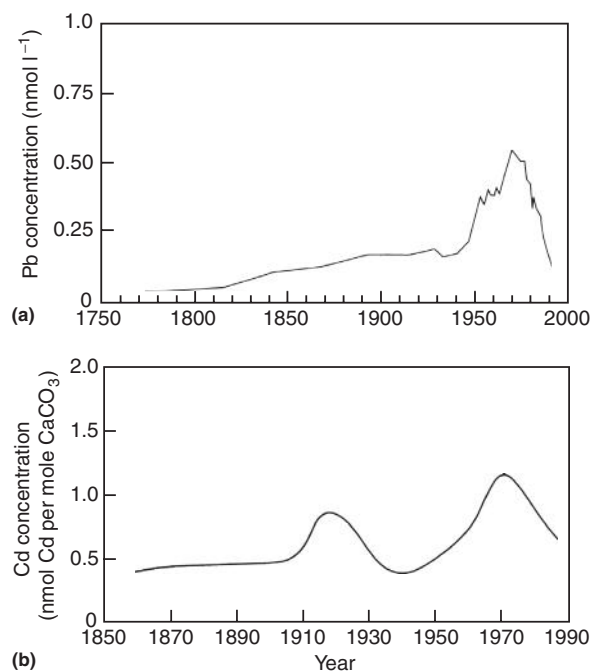


Figure 2 (a) Concentration of lead in Greenland snow. (Reproduced from Wolff, E., 1995. In: Wolff, E., Bales, R.C. (Eds.), *Chemical Exchange between the Atmosphere and Polar Snow*. Springer-Verlag, Berlin, pp. 1–18.) (b) Concentration of cadmium in coral from Bermuda in the Sargasso Sea, Central Atlantic. (Reproduced from Shen, G.T., Boyle, E.A., Lea, D.W., 1987. *Nature* 328, 794–796.) Note high concentrations during the early twentieth century associated with industrialization with limited emission abatement then reduced emissions following basic emission control with a subsequent rise with increasing industrialization before a fall post-1970 with improved emission control.

falling to levels, which are equivalent to those of the early 1900s. For other metals such as cadmium, which do not have a dominant single source, as was the case for lead, there is evidence of reducing concentrations as a result of improving emission control at source, although this is balanced against increasing societal use of resources.

While atmospheric transport of some trace metals has increased due to societal emissions, that of others such as iron has been little changed. Iron fluxes have changed on longer timescales, being higher during the last glaciation owing to increased aridity and stronger winds. Over the last few years the role of atmospheric iron transport has attracted increasing interest as it has become clear that iron is a key nutrient for phytoplankton in the oceans and that in some areas remote from desert regions (the dominant sources of dust), phytoplankton growth may be limited by iron availability. Higher dust loadings during the last glaciation may have promoted high marine productivity and contributed to the lower atmospheric CO₂ levels and hence to a cooler climate. Furthermore, higher productivity may increase emissions of dimethyl sulfide

from the oceans, which would increase the acidity of clouds and promote increasing bioavailability of iron as well as increase global albedo and hence influence climate. This emphasizes the interconnected nature of global cycles and it has become clear over recent years that atmospheric cycling of trace metals is a very important component of the global biogeochemical cycle.

See also: **Aerosols:** Aerosol Physics and Chemistry.
Hydrology, Floods and Droughts: Deserts and Desertification.
Paleoclimatology: Ice Cores. **Tropospheric Chemistry and Composition:** Mercury.

Further Reading

- Arimoto, R., Duce, R.A., Ray, B.J., Unni, C.K., 1985. Atmospheric trace elements at Enewetak Atoll: 2. Transport to the ocean by wet and dry deposition. *J. Geophys. Res.* 90, C22391–C22408.
- Boyd, P.W., Ellwood, M.J., 2010. The biogeochemical cycle of iron in the ocean. *Nat. Geosci.* 3, 675–682.
- Duce, R.A., Liss, P.S., Merrill, J.T., et al., 1991. The atmospheric input of trace species to the world ocean. *Global Biogeochem. Cycles* 5, 193–259.
- Jickells, T.D., 1985. Atmospheric inputs of metals and nutrients to the oceans: their magnitude and effects. *Marine Chem.* 48, 199–201.
- Jickells, T.D., 1997. Atmospheric inputs of some chemical species to the North Sea. *Ger. J. Hydrogr.* 49, 111–118.
- Jickells, T.D., et al., 2005. Global iron connections between desert dust, ocean biogeochemistry and climate. *Science* 208, 65–71.
- Kelly, A.E., Reuer, M.K., Goodkin, N.F., Boyle, E.A., 2009. Lead concentrations and isotopes in corals and water near Bermuda, 1780–2000. *Earth Planet. Sci. Lett.* 283, 93–100.
- Majestic, B.J., Anbar, A.D., Herckes, P., 2009. Stable isotopes as a tool to apportion atmospheric iron. *Environ. Sci. Technol.* 43, 4327–4333.
- Nriagu, J.O., 1989. A global assessment of natural sources of atmospheric trace metals. *Nature* 338, 47–49.
- Pacyna, J.M., Pacyna, E.G., 2001. An assessment of global and regional emissions of trace metals to the atmosphere from anthropogenic sources worldwide. *Environ. Rev.* 9, 269–298.
- Pirrone, N., et al., 2010. Global mercury emissions to the atmosphere from anthropogenic and natural sources. *Atmos. Chem. Phys.* 10, 5951–5964.
- Renberg, I., Brännvall, M.L., Bindler, R., Emteryd, O., 2000. Atmospheric lead pollution during four millennia (2000 BC to 2000 AD) in Sweden. *Ambio* 29, 150–156.
- Rickli, J., Frank, M., Baker, A.R., Aciego, S., de Souza, G., Georg, R.B., Halliday, A.N., 2010. Hafnium and neodymium isotope distribution in surface waters of the eastern Atlantic Ocean: implications for sources and inputs of trace metals to the ocean. *Geochim. Cosmochim. Acta* 74, 540–557.
- Spokes, L.J., Jickells, T.D., 2002. Speciation of metals in the atmosphere. In: Ure, A.M., Davidson, C.M. (Eds.), *Chemical Speciation in the Environment*, second ed. Blackwell, Oxford, pp. 161–187.
- Turner, D.R., Hunter, K.A. (Eds.), 2001. *The Biogeochemistry of Iron in Seawater*. Wiley, Chichester.
- Wiersma, G.B., Davidson, C.I., 1986. Trace metals in the atmosphere of rural and remote areas. In: Nriagu, J.O., Davidson, C.I. (Eds.), *Toxic Metals in the Atmosphere*. Wiley, New York.
- Zhuang, G., Yi, Z., Duce, R.A., Brown, P.R., 1992. Link between iron and sulphur suggested by detection of Fe(II) in remote marine aerosols. *Nature* 355, 537–539.

Iodine

LJ Carpenter, University of York, York, UK

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

Iodine was discovered as a new element two centuries ago. It is essential to human health, yet a substantial proportion of the world's population are at risk from a lack of dietary iodine. Natural processes within the oceans are responsible for the majority of atmospheric iodine. In the lower atmosphere, iodine undergoes a complex array of photochemical reactions that lead to ozone destruction and, in some cases, new particle formation, before being scavenged by aerosol. The nature and mechanisms for oceanic iodine emissions are still uncertain.

Discovery

The discovery of iodine was by a French scientist, Bernard Courtois (1777–1838), in 1811 in the process of making saltpeter (potassium nitrate, a major component of gunpowder) for Napoleon's army. Potassium carbonate, required for saltpeter production, had been previously extracted from the ashes of willow wood. However, the Napoleonic Wars had gone on for so long that supplies were exhausted, and seaweed was suggested as an alternative feedstock to wood ashes. After adding concentrated sulfuric acid to burnt seaweed, Courtois noticed a violet vapor cloud, which he suspected to be a new element (Courtois, 1813). Soon after, several other scientists including Désormes (1777–1862), Clément (1779–1841), Gay-Lussac (1778–1850), Ampère (1775–1836), and the English chemist Sir Humphry Davy (1778–1829) performed scientific investigations into this substance. Davy came to the conclusion that it was an element in its own right, with properties similar to those of chlorine. This new element was named iodine after a Greek word for 'violet colored' (Figure 1).

Commercial and Medicinal Importance of Iodine

Later, iodine gathered commercial interest for its antiseptic qualities, its use in photography, and also in the treatment of thyroid enlargement (goiters). As a component of the thyroid hormones thyroxine and triiodotyronine, iodine is central to

thyroid function in vertebrates. Severe iodine deficiency also causes reproductive and neurological damage, the latter termed cretinism and associated with developmental delays. The introduction of iodized salt since the early 1900s has eliminated iodine deficiency disorders in many countries, however, even today it continues to be a serious threat to global human health and the greatest preventable cause of mental retardation (The Lancet, 2008). According to the World Health Organization, in 2007 nearly 2 billion individuals, a third being of school age, had insufficient iodine intake. Individuals living inland (with a lack of marine, iodine-rich foods in their diet) and in third world nations are especially at risk, however, a trend toward lower discretionary salt use and consumption of more processed foods (using noniodized salt) means that there is a widespread risk of obtaining insufficient iodine from food (Figure 2).

Besides its use in iodized salt, iodine is required in numerous products including in solar cells as a component of the conducting electrolyte, in X-ray contrast media (substances which enable visualization of soft tissues in X-ray examination), as an effective broad spectrum bactericide to disinfect wounds and to sanitize surface water for drinking (as tincture of iodine or more recently povidone iodine which is less irritating to the skin and other tissues), in the industrial production of acetic acid and nylon fiber, and as a component of the polarizing film of liquid crystal displays. Iodine is also frequently utilized as a reagent in organic synthesis, employed in iodinations, oxidations, and as Lewis acid.

Production of iodine from kelps (brown seaweeds), which contain very high concentrations, was a major economic activity in coastal regions of France, Scotland, and Ireland in the nineteenth century. Today, nearly 90% of global iodine production is based in either Chile, from the mining and leaching of nitrate ores (caliches), or in areas of Japan from iodine-rich brines found in some natural gas and oil fields.

Sources of Iodine to the Environment

Anthropogenic releases of atmospheric iodine (such as fossil fuel combustion) are believed to be negligible on a global scale compared to natural sources. The latter are mainly of marine origin, with a lesser contribution from land sources. Over 110 iodine-containing natural products have been identified, most of which originate from marine organisms (Dembitsky, 2006).

Atomic weight	126.90447
Atomic number	53
Melting point	113.7 °C
Boiling point	184.3 °C

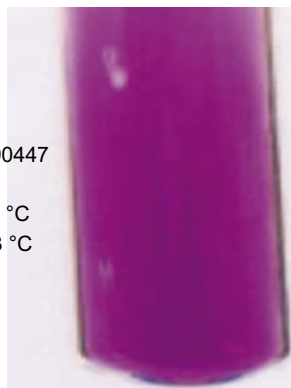


Figure 1 Iodine.

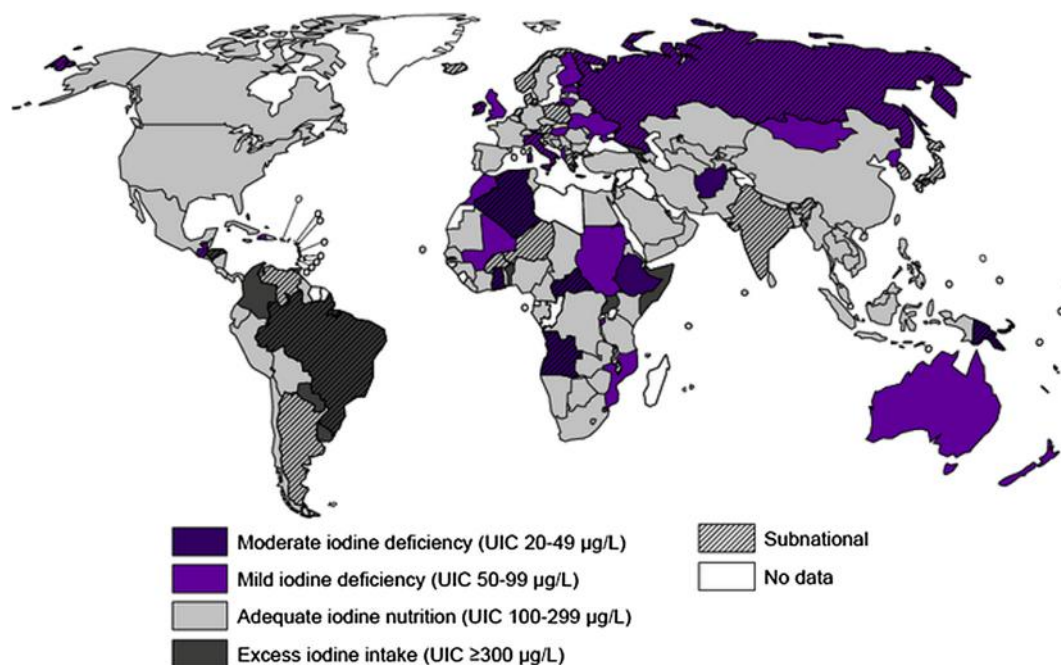


Figure 2 National iodine status in 2013. International Council for the Control of Iodine Deficiency Disorders, www.iccidd.org (accessed January 2014).

Iodine from Marine Algae and Microbes

Biological pathways for the production of iodine compounds by microalgae (phytoplankton) and macroalgae have been identified. Production of mono- and polyhalogenated compounds is believed to occur via the enzyme-catalyzed oxidation of halide ions within the algal tissue. Mono-halogenated organics are produced via the enzyme methyl transferase, present in both micro- and macroalgae (Wuosmaa and Hager, 1990; Itoh et al., 1997). Production of polyhalogenated organic species is enabled by haloperoxidase enzymes, present in both macroalgal (Theiler et al., 1978) and microalgal (Moore et al., 1996) cultures. Haloperoxidases catalyze the oxidation of halides by hydrogen peroxide (H_2O_2), released as part of normal cell metabolism and during oxidative stress, and the resulting products react with available organic material within the apoplast (cell wall space) to form volatile organohalogenes (Theiler et al., 1978). Under conditions of oxidative stress, e.g., at elevated temperatures/light or when exposed to grazing, H_2O_2 would otherwise build up to high levels. Thus, the mechanism is essential for the health of the organism. As a consequence of these reactions, a large variety of volatile organic iodine compounds (VOICs) are emitted from both microalgae (Moore et al., 1996) and macroalgae (Laternus et al., 2000; Carpenter et al., 2000). Enhanced emissions have been found in some studies when the algae are subject to high illumination (Moore et al., 1996; Carpenter et al., 2000; Laternus et al., 2004). In contrast, Hughes et al. (2006) found no effect of high light stress on iodocarbon release by three species of marine phytoplankton.

The kelp *Laminaria* and related brown macroalgae are the strongest iodine accumulators among all living systems (Küpper et al., 1998), and produce substantial volatile iodine

emissions in coastal regions (Carpenter et al., 1999; McFiggans et al., 2004; Küpper et al., 2008). X-ray absorption spectroscopy shows that the accumulated form of iodine in kelps is iodide, which serves as a simple inorganic antioxidant, protecting the apoplast from high levels of reactive oxygen species (ROS) such as H_2O_2 (Küpper et al., 2008). Upon oxidative stress, a strong efflux of accumulated iodide occurs, resulting not only in organic iodine compounds, but also in molecular iodine (I_2) when the seaweed is exposed to air (McFiggans et al., 2004; Saiz-Lopez and Plane, 2004; Palmer et al., 2005). This mechanism occurs via a rapid heterogeneous reaction of gas-phase atmospheric ozone with liquid iodide (millimolar concentrations) on the kelp surface, leading to the release of I_2 at rates up to about 5 orders of magnitude higher than the combined iodocarbon emissions (Küpper et al., 2008). This phenomenon could be widespread over seaweed beds, since release of iodide during oxidative stress is observed in both brown and red algae (Truesdale, 2008; Chance et al., 2009) (Figure 3).

In the open ocean, biological mechanisms for iodine production include methylation of iodine by marine microalgae (phytoplankton) (Itoh et al., 1997) and by a wide variety of aerobic marine bacteria (Amachi et al., 2001; Amachi, 2008), production of polyiodinated organic compounds from phytoplankton containing haloperoxidase enzymes (Moore et al., 1996), and production from natural marine aggregates comprising a host of material including phytoplankton and bacteria (Hughes et al., 2008). Iodide-oxidizing bacteria, which oxidize iodide to I_2 and organic iodine have also been found in iodide-rich environments (Fuse et al., 2003; Amachi et al., 2005), but appear to be absent or present in very low concentrations in natural seawaters and terrestrial soils (Amachi et al., 2005).

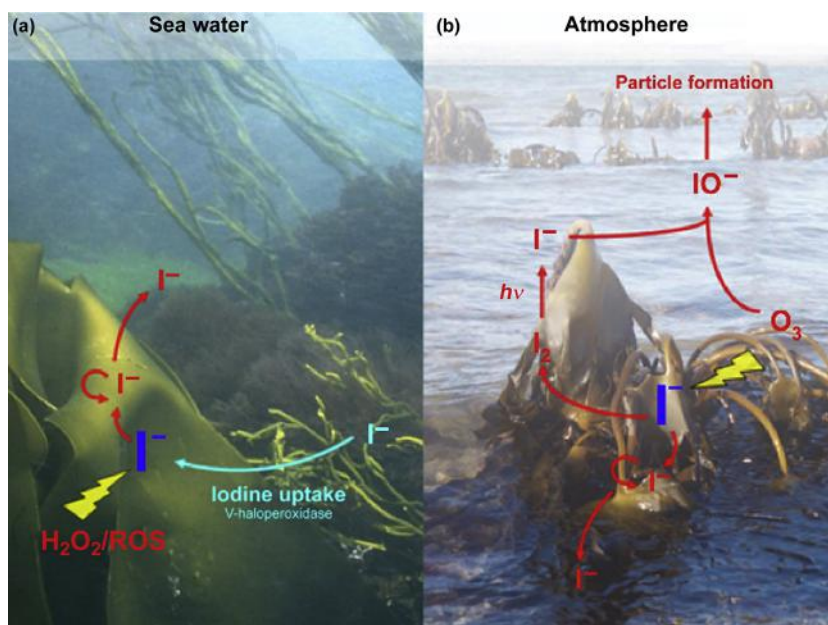


Figure 3 A simple model of iodine metabolism in *Laminaria*. (a) When submerged and unstressed, the alga accumulates iodide from seawater mediated by vanadium haloperoxidase (turquoise). When oxidative stress occurs (red), iodide is released to detoxify ROS (including both aqueous H_2O_2 and gaseous O_3) at the kelp surface. During oxidative stress at high tide, iodide is released into the surrounding seawater. (b) At low tide, the ozone-scavenging reactivity of iodide on kelp surface results in the release of molecular iodine, whose photolysis initiates a series of reactions producing bursts of aerosol particles. From Küpper, F.C., Carpenter, L.J., McFiggans, G.B., 2008. Iodide accumulation provides kelp with an inorganic antioxidant impacting atmospheric chemistry. *Proc. Natl. Acad. Sci. U.S.A.* 105 (19), 6954–6958.

The biological significance of iodine accumulation has been proposed to be linked to the antioxidant properties of I^- (Küpper et al., 2008), and it is well known that I_2 is a highly active oxidizing agent that has strong bactericidal, fungicidal, and sporicidal activities (McDonnell and Russell, 1999). Whether the evolved volatile iodine species produced by marine algae have a biological function is currently unknown. They may simply be a by-product of the intra- and intercellular removal of the H_2O_2 from algal tissue (Manley and Barbero, 2001), alternatively they may also act to deter the grazing of zooplankton. These biological phenomena have, however, significant impacts on atmospheric chemistry on scales from the local and global (see Section *Reactive Iodine Species in the Atmosphere*).

Current 'bottom-up' global estimates of biogenic organoiodine fluxes to the atmosphere from macroalgae, extrapolated from incubation studies, are of the order of 10^8 – 10^{10} g of iodine per year (Manley et al., 1992; Laturnus et al., 2000; Carpenter, 2003). Similar extrapolation of microalgal production deduced from laboratory culture studies results in a contribution of 10^6 – 10^9 g of iodine per year in the form of methyl iodide (CH_3I) (Manley and de la Cuesta, 1997). Smythe-Wright et al. (2006) attributed extremely high CH_3I levels in seawater to biological production by the picoplankton *Prochlorococcus marinus*, suggesting this source alone could contribute 5.3×10^{11} g of iodine per year. In contrast, Brownell et al. (2010) found that the production of CH_3I by *P. marinus* can account for only a small fraction of the estimated global oceanic production. Bacterial production of CH_3I is estimated as 1.6×10^9 g of iodine per year (Amachi, 2008).

Estimates of the global organoiodine burden from sea–air flux measurements are much higher than typical bottom-up estimates. Extrapolating from coastal, shelf, and open ocean summer data, Jones et al. (2010) estimated a total global iodine flux of 1.1×10^{12} g of iodine per year, of which about half is from CH_3I (Butler et al., 2007; Jones et al., 2010), with the remainder dominated by chloroiodomethane (CH_2ICl) and diiodomethane (CH_2I_2) (both $\sim 20\%$). In the Atlantic shelf regions, the major single contributor to the flux appears to be CH_2ICl (~ 40 – 50% of the total organoiodine flux of ~ 30 – $40 \mu\text{mol (I)}$ per square meter per year in summer) (Jones et al., 2010; Archer et al., 2007), whereas CH_3I appears as the most important contributor in open ocean waters of the Atlantic (45 – 65% of the total flux; Jones et al., 2010). For both gases, fluxes tend to be higher over biologically active regions, as demonstrated by the English Channel shelf measurements of Archer (Archer et al., 2007) and the Irish continental shelf and Mauritanian upwelling measurements of Jones et al. (2010). Figure 4 shows measurements of CH_3I and CH_2ICl emissions from several studies in the open Atlantic Ocean, calculated from the measured partial pressure difference of the gas across the sea surface flux = $K_T (C_{\text{water}} - C_{\text{air}}/H)$, where H is the Henry's law coefficient and K_T is the wind-dependent gas transfer velocity (Nightingale et al., 2000).

Thus, it appears that either incubation studies fundamentally underestimate marine emissions, and/or that other oceanic sources are at play. Some evidence for the former comes from Hughes et al. (2008) who observed high monoiodinated organic production rates, approximately an order of magnitude greater than that reported for bulk natural water sample in natural marine aggregates (comprising a host of materials

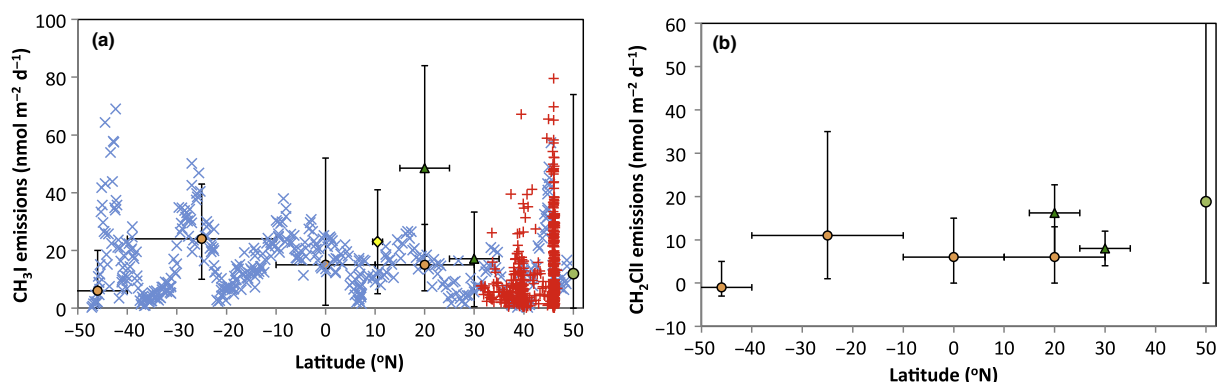


Figure 4 Emissions of CH₃I and CH₂ClI from the open Atlantic Ocean. (a) CH₃I: blue cross (X) – BLASTII data Oct–Nov 1994 Northeast and Southwest Atlantic (Butler, J.H., King, D.B., Lobert, J.R.M., Montzka, D.S., Yvon-Lewis, S.A., Hall, B.D., Warwick, N.J., Mondeel, D.J., Aydin, M., Elkins, J.W., 2007. Oceanic distributions and emissions of short-lived halocarbons. *Glob. Biogeochem. Cycles* 21, GB1023); red plus (+) – GASEX98 data May–Jul 1998 East–West transect across North Atlantic (Butler, J.H., King, D.B., Lobert, J.R.M., Montzka, D.S., Yvon-Lewis, S.A., Hall, B.D., Warwick, N.J., Mondeel, D.J., Aydin, M., Elkins, J.W., 2007. Oceanic distributions and emissions of short-lived halocarbons. *Glob. Biogeochem. Cycles* 21, GB1023); orange circles with error bars – mean and range of (Chuck, A.L., Turner, S.M., Liss, P.S., 2005. Oceanic distributions and air-sea fluxes of biogenic halocarbons in the open ocean. *J. Geophys. Res. Oceans* 110, C10022, doi:10.1029/2004JC002741) Northeast and Southwest Atlantic Sept–Oct 2000; yellow diamond with error bars – mean and range of (Richter, U., Wallace, D.W.R., 2004. Production of methyl iodide in the tropical Atlantic Ocean. *Geophys. Res. Lett.* 31 (23). Art. No. L23S03) tropical East Atlantic Oct–Nov 2002; green filled circle with error bars – annual mean and range of (Archer, S.D., Goldson, L.E., Liddicoat, M.I., Cummings, D.G., Nightingale, P.D., 2007. Marked seasonality in the concentrations and sea-air flux of volatile iodocarbon compounds in the western English Channel. *J. Geophys. Res.* 112, C08009. doi:10.1029/2006JC003963) English Channel; green diamonds – mean and range of (Jones, C.E., Hornsby, K.E., Sommariva, R., Dunk, R.M., von Glasow, R., McFiggans, G., Carpenter, L.J., 2010. Quantifying the contribution of marine organic gases to atmospheric iodine. *Geophys. Res. Lett.* 37 (18), L18804) – Northeast (June 2006) and tropical East (June 2007) Atlantic, excluding coastal data. (b) CH₂ClI: Orange circles with error bars – mean and range of (Chuck, A.L., Turner, S.M., Liss, P.S., 2005. Oceanic distributions and air-sea fluxes of biogenic halocarbons in the open ocean. *J. Geophys. Res. Oceans* 110, C10022, doi:10.1029/2004JC002741) Northeast and Southeast Atlantic Sept–Oct 2000; green diamonds – mean and range of (Jones, C.E., Hornsby, K.E., Sommariva, R., Dunk, R.M., von Glasow, R., McFiggans, G., Carpenter, L.J., 2010. Quantifying the contribution of marine organic gases to atmospheric iodine. *Geophys. Res. Lett.* 37 (18), L18804) – Northeast (June 2006) and tropical East (June 2007) Atlantic, excluding coastal data; green filled circle with error bars – annual mean and range of (Archer, S.D., Goldson, L.E., Liddicoat, M.I., Cummings, D.G., Nightingale, P.D., 2007. Marked seasonality in the concentrations and sea-air flux of volatile iodocarbon compounds in the western English Channel. *J. Geophys. Res.* 112, C08009. doi:10.1029/2006JC003963) English Channel.

including diatoms, bacteria, and phytodetritus), possibly because of the copresence of high bacterial activity and organic material. These authors suggest that such detrital particles could be hotspots of production in the marine environment.

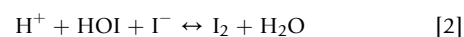
In coastal regions, atmospheric concentrations of both organoiodine compounds (Carpenter et al., 1999) and I₂ (Saiz-Lopez and Plane, 2004; Saiz-Lopez et al., 2006) are highest at low tide, consistent with emissions occurring from seaweed. A footprint model of I₂ emissions from seaweeds estimated the release of up to 510 μmol (I) per square meter per year of I₂ at low tide at Roscoff, France (Leigh et al., 2010; R. Leigh, personal communication), based upon species-specific I₂ emission rates parameterized from the laboratory study of Ball et al. (2010). Agreement between modeled and observed I₂ concentrations could only be obtained during the day if the model included an efficient recycling mechanism for I₂ in the atmosphere, equivalent to approximately 95% slower photolysis of I₂. This is contrary to current understanding and suggests either a substantial missing recycling or source component for I₂, or that unmodeled local emissions were present during this field study.

Abiotic Mechanisms in the Marine Environment

There is also evidence that nonbiological iodine sources are significant in the global inventory of iodine. The major source of CH₃I appears to be from sunlit irradiation of seawater (Moore and Zafriou, 1994; Bell et al., 2002). The suggested

mechanism is photochemical generation of methyl and iodine radicals from, respectively, dissolved organic matter and iodide present in the surface waters. Measurements in the North and tropical Atlantic have confirmed that surface saturation anomalies of CH₃I are correlated with light intensity and that the photochemical source of CH₃I is abiotic, and suggest that this route can support at least half of the average sea-to-air flux of 23 nmol m⁻² day⁻¹ (Happell and Wallace, 1996; Richter and Wallace, 2004). An additional dust-stimulated abiotic emission of CH₃I from the ocean or marine aerosol was suggested by Williams et al. (2007).

An additional abiotic contribution to the atmospheric iodine budget may arise from the reaction of atmospheric O₃ deposited at the sea surface with iodide to evolve I₂ (reactions [1] and [2]), as first proposed by Garland et al. (1980) and Garland and Curtis (1981). Recent work suggests that this could be an efficient mechanism for iodine emission from the open ocean (Carpenter et al., 2013). This is also the mechanism proposed to occur on kelp surfaces (Palmer et al., 2005), as discussed earlier.



Martino et al. (2009) proposed that reactions [1] and [2], followed by the reaction of HOI or I₂ with dissolved organic carbon (DOC) in the sea surface, also results in the production

of reactive organoiodine compounds including CH_2I_2 , CHClI_2 , and CH_3I . Production of polyhalomethanes from such reactions is known to occur during the disinfection of natural waters.

Another potential route to direct ocean surface production of small halogen molecules is via oxidation of halogen anions to their radical forms by photosensitizers such as chlorophyll or aromatic ketones, a known component of marine DOC, which in turn will lead to the formation of organic halogens in the presence of organic compounds in the sea surface microlayer (Jammoul et al., 2009). The oxidation is enhanced in the presence of atmospheric O_3 , which acts as an electron acceptor, thus promoting the cationic form of the photosensitizer.

Although suggestive of additional iodine sources, extrapolating these laboratory studies to the environment is not straightforward. However, some evidence for production of I_2 in surface seawater is provided from voltammetry measurements of $\sim 10^{-9} \text{ mol dm}^{-3} \text{ I}_2$ in some samples of surface coastal seawater (Möller et al., 1996).

Marine Destruction

Chemical depletion of volatile iodine compounds in the oceans can potentially play an important role in the extent to which they contribute to tropospheric photochemistry. Nucleophilic substitution by chloride ions (Cl^-) via an $\text{S}_{\text{N}}2$ mechanism is the major chemical sink of CH_3I , $\text{C}_2\text{H}_5\text{I}$, and $1\text{-C}_3\text{H}_7\text{I}$ in seawater (Zafriou, 1975; Zika et al., 1984; Elliott and Rowland, 1993; Jones and Carpenter, 2005), and is competitive with sea-air transfer for CH_3I . Chlorination reactions are however strongly temperature-dependent and at low temperatures hydrolysis (Elliott and Rowland, 1993; Jones and Carpenter, 2007) of CH_3I becomes competitive. In the case of $\text{C}_2\text{H}_5\text{I}$ and $1\text{-C}_3\text{H}_7\text{I}$, the Cl^- ion reaction occurs at a rate substantially slower than volatilization, except in waters of very high temperature ($\sim 25^\circ\text{C}$) (Jones and Carpenter, 2007). At equivalent rates, $2\text{-C}_3\text{H}_7\text{I}$ is depleted with respect to its reaction with Cl^- ions and hydrolysis, implying an $\text{S}_{\text{N}}1$ mechanism (Jones and Carpenter, 2007). In most oceanic conditions, abiotic chemical destruction of $2\text{-C}_3\text{H}_7\text{I}$ is competitive with loss from the ocean due to volatilization, i.e., chemical

breakdown in the oceans significantly limits the efficiency of $2\text{-C}_3\text{H}_7\text{I}$ transfer to the marine boundary layer (MBL). In regions where the seawater temperature is below $\sim 10^\circ\text{C}$ such as in polar waters, the majority of monoiodinated alkanes produced in the ocean will ultimately end up being emitted to the MBL to participate in tropospheric photochemistry. However, abiotic chemical destruction has the potential to significantly limit the oceanic emission of CH_3I and $2\text{-C}_3\text{H}_7\text{I}$ to the MBL in warm waters.

In contrast to the monoiodinated alkanes, polyhalogenated organoiodides undergo significant photolysis in surface marine waters (Martino et al., 2005; Jones and Carpenter, 2005; Jones et al., 2010). Surface photolysis results in a reduction of average daily sea-air fluxes by up to about 25% (although 20–80% at midday) for CH_2I_2 in tropical Atlantic waters, but has very little effect on the loss of CH_2ICl and CH_2IBr (Jones et al., 2010). About 30% of photolyzed CH_2I_2 produces CH_2ICl in seawater (Martino et al., 2005; Jones and Carpenter, 2005), representing a potentially significant route to this compound (Carpenter et al., 2007).

Lifetimes in seawater against chlorination, hydrolysis, and photolysis for temperatures of 0, 15, and 30°C , respectively, are shown in Table 1.

Terrestrial Halogenated Source Gas Emissions

Methyl halides are produced by a large number of terrestrial ecosystems, crops, and biota, including coastal salt marshes, freshwater wetlands, peatlands, forest soils, tropical forests, rice paddies, and wood-rotting fungi (Dimmer et al., 2001; Harper, 1985; Redeker et al., 2000; Rhew et al., 2001, 2000; Varner et al., 1999; Yokouchi et al., 2002). Terrestrial sources of CH_3I are together believed to comprise up to 80–110 Gg of iodine per year (Bell et al., 2002; Sive et al., 2007), comprising 18–50 Gg of iodine per year from rice paddies (Muramatsu and Yoshida, 1995; Redeker et al., 2000), ~ 6 Gg of iodine per year from natural wetlands (Dimmer et al., 2001), ~ 8 Gg of iodine per year from biomass burning (Bell et al., 2002), and 30 Gg of iodine per year from midlatitude terrestrial biomes (Sive et al., 2007). Keppler et al. (2000) proposed an abiotic route for alkyl halide production in soils and sediments from halide ion alkylation during the oxidation of organic matter

Table 1 Lifetimes (to the nearest year, month, week, or day) with respect to total chemical destruction of CH_3I , $\text{C}_2\text{H}_5\text{I}$, $1\text{-C}_3\text{H}_7\text{I}$, and $2\text{-C}_3\text{H}_7\text{I}$ at a range of seawater temperatures. Lifetimes with respect to simultaneous chlorination and hydrolysis are in bold, chlorination only in brackets, and photolysis only in italics. Photolysis lifetimes are calculated at 1 July, 12 p.m. in 54°N

Iodoalkane	Lifetime (0°C)	Lifetime (15°C)	Lifetime (30°C)
CH_3I	11 months (3 years)	7 weeks (11 weeks)	7 days (8 days)
$\text{C}_2\text{H}_5\text{I}$	11 months (11 years)	9 weeks (7 months)	10 days (2 weeks)
$1\text{-C}_3\text{H}_7\text{I}$	7 months (24 years)	8 weeks (1 year)	2 weeks (4 weeks)
$2\text{-C}_3\text{H}_7\text{I}$	4 months	12 days	1 day
CH_2ICl	<i>9 h $\pm$ 2 h</i>	<i>9 h $\pm$ 2 h</i>	<i>9 h $\pm$ 2 h</i>
CH_2IBr	<i>4.5 h $\pm$ 40 min</i>	<i>4.5 h $\pm$ 40 min</i>	<i>4.5 h $\pm$ 40 min</i>
CH_2I_2	<i>9 min $\pm$ 1 min</i>	<i>9 min $\pm$ 1 min</i>	<i>9 min $\pm$ 1 min</i>

From Jones, C.E., Carpenter, L.J., 2005. Photolysis of CH_2I_2 , CH_2ICl and CH_2IBr in water and seawater by solar radiation. Environ. Sci. Technol. 39, 6130–6137; Jones, C.E., Carpenter, L.J., 2007. Chemical destruction of CH_3I , $\text{C}_2\text{H}_5\text{I}$, $1\text{-C}_3\text{H}_7\text{I}$, and $2\text{-C}_3\text{H}_7\text{I}$ in saltwater. Geophys. Res. Lett. 34 (13).

by an electron acceptor such as Fe(III). The authors proposed that production of C1–C4 alkyl iodides from soils containing Fe(III) and iodide could be significant globally, although their data did not allow for an estimate of emission. Terrestrial sources for the di- and triiodated compounds have not been identified.

Summary of Iodine Sources

It is clear that of the few teragrams of iodine believed to be emitted per year to the atmosphere, the majority comes from the ocean. The volatile fraction of iodine in seawater is low: it exists predominantly as iodate, iodide, and nonvolatile dissolved organic iodine, with a total concentration of around 0.45 mM (Wong, 1991). Sea-salt aerosol particles, arising from evaporation of sea spray, are generally enriched in iodine by 2–3 orders of magnitude compared to concentrations in seawater (Sturges and Barrie, 1988; Baker, 2004). Thus sea-salt aerosol is a net sink of gaseous iodine (in contrast, it represents the main source of chlorine and bromine in the MBL), though it plays an important role in recycling gas-phase iodine. Photooxidation of emitted volatile organic and inorganic iodine species in air to soluble inorganic products allows a significant fraction of the iodine to partition to the aerosol phase, which is eventually removed by wet or dry deposition to the land or ocean. Thus the cycle provides a vital route for terrestrial uptake of iodine, an essential component of mammalian health as discussed in Section [Discovery](#).

The current estimate of global organoiodine emissions of ~1 Tg of iodine per year (Jones et al., 2010) (which equates to an average sea–air flux of ~20 μM of iodine per square meter per year) is rather higher than the depositional flux of I^- and IO_3^- in rainwater and marine aerosol into the southern North Sea of 6.3–9.2 μM of iodine per square meter per year (Baker et al., 2001). The total sea-to-air flux of iodine is likely to be higher due to chemical (Carpenter et al., 2013) and photochemical (Jammoul et al., 2009) mechanisms for production

of inorganic iodine from the ocean surface. The depositional flux does not include soluble organically bound iodine, which can make a significant or even major contribution to precipitation and marine aerosol iodine (Baker, 2005; Gilfedder et al., 2008). The contribution of different chemical and biological sources to gaseous iodine in the marine atmosphere is currently highly uncertain and requires further attention.

Reactive Iodine Species in the Atmosphere

Photoproduction of Iodine Atoms

Current understanding of the main features of iodine photochemistry in the MBL is shown in [Figure 5](#).

The cycle is initiated by photolysis of iodine compounds with lifetimes ranging from several days (CH_3I , $\text{C}_2\text{H}_5\text{I}$, $\text{C}_3\text{H}_7\text{I}$: Rattigan et al., 1997; Roehl et al., 1997), several hours (CH_2ICl : Rattigan et al., 1997; Roehl et al., 1997), an hour or less (CH_2IBr , Mössinger et al., 1998), about 5 min (CH_2I_2 : Roehl et al., 1997; Mössinger et al., 1998) to less than 10 s (I_2 – Saiz-Lopez et al., 2004) at midday in the lower troposphere. The more photoreactive iodo-carbons with two chromophores, e.g., CH_2I_2 and CH_2ICl , have been shown to be the most important organic iodine sources in the MBL (Jones et al., 2010). Although the lifetimes of the polyhalomethanes are controlled almost entirely by photodissociation, OH- and Cl-radical-initiated attack could account for 10–20% of the removal of CH_3I and compete with photolysis for removal of the propyl iodides (Cotter et al., 2001).

The iodine atoms arising from molecular or organoiodine breakdown almost exclusively react with O_3 forming the iodine monoxide (IO) radical and undergoing subsequent reactions, which alter atmospheric chemistry and physics, as described below. The ultimate fate of iodine is accumulation in particles in the form of iodide and iodate (Pechtl et al., 2006), typically occurring by the uptake of inorganic iodine (mostly INO_x , HIO_3 , and I_xO_y).

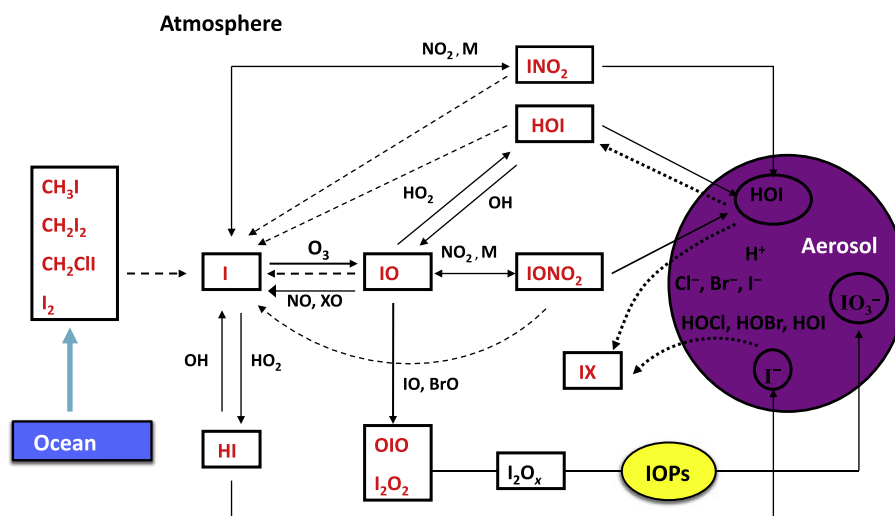


Figure 5 Atmospheric chemistry of iodine. Dashed lines represent photolysis reactions. Dotted lines represent volatilization from aerosol. IOPs, iodine oxide particles; X, Cl, Br, or I.

Atmospheric Observations of Gas-Phase Reactive Iodine Species

Coastlines

It has become clear that coastal regions offer a unique iodine-rich environment through direct emissions of very reactive molecular iodine (I_2) from seaweed (Dixneuf et al., 2009; Palmer et al., 2005; Saiz-Lopez and Plane, 2004; Huang et al., 2010). It was long established that seaweeds, particularly kelps, emit VOIC (Gschwend et al., 1985; Manley and Dastoor, 1987; Nightingale et al., 1995; Carpenter et al., 2000); however, it now appears that the inorganic iodide efflux leading to I_2 formation after an oxidative burst is some 3–5 orders of magnitude higher than organic iodine emissions (Küpper et al., 2008). Mixing ratios of I_2 of up to 93 pptv (peaking at night due to the fast photolysis of I_2) have been made at Mace Head and other coastal locations (Saiz-Lopez and Plane, 2004; Peters et al., 2005; Bitter et al., 2005; Saiz-Lopez et al., 2006; Huang et al., 2010). Such volatilization of I_2 from kelps exposed to air appears to explain the observations of coastal ‘bursts’ of iodine-containing ultrafine aerosol particles at low tide during the day (see Section *Aerosol Formation and Iodine Oxide Particles*) (O’Dowd et al., 1999, 2002b).

Differential Optical Absorption Spectroscopy (DOAS) measurements of IO were first reported at the coastal site of Mace Head, Ireland by Alicke et al. (1999), and since then numerous studies have indicated that IO is ubiquitous in the air above kelp-rich coastlines (Wada et al., 2007; Whalley et al., 2007; Mahajan et al., 2009; Commene et al., 2011). IO levels tend to peak during the day at low tide, consistent with production via the photolysis of seaweed iodine emissions. Nighttime IO mixing ratios of up to 3 pptv have also been observed (Saiz-Lopez and Plane, 2004) and attributed to the reaction of I_2 with the nitrate radical (NO_3) to give I atoms (Chambers et al., 1992), which are subsequently oxidized to IO by O_3 .

Point measurements of IO by laser-induced fluorescence (Whalley et al., 2007; Commene et al., 2011) and by cavity ring-down spectroscopy (CRDS) (Wada et al., 2007), of I_2 by broadband CRDS (Bitter et al., 2005; Leigh et al., 2009), and using diffusion denuders (Huang et al., 2010) reveal considerable spatial heterogeneity in coastal mixing ratios of reactive iodine species. IO mixing ratios measured by point techniques are typically an order of magnitude larger than those reported by long-path DOAS (LP-DOAS). Model simulations of such observations are consistent with I_2 emissions concentrated over the intertidal region of a few hundred meters (Leigh et al., 2009).

In addition to IO and I_2 , there have been several observations of OIO, and more recently I atoms at coastal sites. Using a resonance fluorescence technique, Bale et al. (2008) measured ambient I atom levels of up to 22 pptv during the day at Mace Head. At both Mace Head and a similar site with abundant seaweed at Roscoff, France, mixing ratios of OIO detected by LP-DOAS peaked at nighttime at 9–10 pptv and were below the instrument detection limit of ~ 2 pptv during the day, consistent with rapid photolysis (Gómez Martín et al., 2009; Mahajan et al., 2009). In contrast, high daytime OIO levels (>20 pptv) at the Gulf of Maine have been reported (Stutz et al., 2007).

Open Ocean

Evidence for the widespread occurrence of iodine chemistry, away from the influence of seaweed, comes from observations of IO off the north coast of Tenerife, Canary Islands (LP-DOAS; up to 3 pptv IO) (Allan et al., 2000), the northeast coast of Sao Vicente in the Cape Verde archipelago (LP-DOAS; 1–3 pptv IO) (Read et al., 2008; Mahajan et al., 2010a), and the Eastern Pacific upwelling region (multiaxis DOAS slant columns of IO) (Volkamer et al., 2010). In addition to these ground-based measurements, satellite slant columns indicate widespread levels of IO over tropical upwelling regions of the Pacific (Schönhardt, 2009; Schönhardt et al., 2008).

The presence of reactive iodine species can activate the release of bromine and chlorine via heterogeneous reactions on sea-salt aerosol (Vogt et al., 1999) and the combined presence of halogens in the marine environment acts synergistically to catalyze O_3 destruction (see Section *Ozone Destruction*). At Cape Verde, the result is a $\sim 50\%$ increase in photochemical O_3 destruction rates (compared to a hypothetical situation without halogens), with iodine responsible for about 2/3 of the halogen-related loss (Read et al., 2008; Mahajan et al., 2010a; Jones et al., 2010). Observed ozone depletion in this region cannot be explained in the absence of halogen compounds (Figure 6). Earlier measurements of O_3 in the remote MBL have also invoked halogen chemistry to explain measured diurnal variations (Dickerson et al., 1999; Galbally et al., 2000), albeit without direct evidence of the presence of halogen oxides.

Despite evidence of its effect on atmospheric chemistry, the sources of IO in the open ocean marine environment are unclear. Measurements of sea-air fluxes of organic iodine compounds in the region of Cape Verde (Jones et al., 2010) cannot be reconciled with the observations of IO, even given uncertainties in kinetic parameters, and suggest a significant additional source of iodine. Molecular iodine released via the heterogeneous reaction of O_3 on the sea surface has been invoked (Mahajan et al., 2010a; Jones et al., 2010). If indeed present as I_2 , the resulting atmospheric levels (up to 7 pptv at night) are close to current instrument detection limits (Mahajan et al., 2010a), thus it is not yet possible to verify whether an I_2 production mechanism operates efficiently over the open ocean.

Polar Boundary Layer

High levels of IO in the coastal Antarctic boundary layer have been inferred from differential slant columns (Frieß et al., 2001; mixing ratios up to 10 pptv) and directly from DOAS measurements (Saiz-Lopez et al., 2007; mixing ratios over 20 pptv). Satellite measurements show that IO is abundant over widespread areas of the Antarctic coast and continent (Saiz-Lopez et al., 2007; Schönhardt et al., 2008). These studies reveal a seasonal cycle in IO with a spring maximum followed by a decrease during summer and then a secondary maximum during autumn (Saiz-Lopez et al., 2007; Schönhardt et al., 2008). Though very similar in atmospheric behavior and mixing ratio to bromine oxide (BrO), the presence of IO was predicted to increase the rate of O_3 loss by a factor of 4 compared to the presence of BrO alone (Saiz-Lopez et al., 2007).

Gas-phase iodine chemistry in the Arctic appears to be less active and on a more localized scale than in the

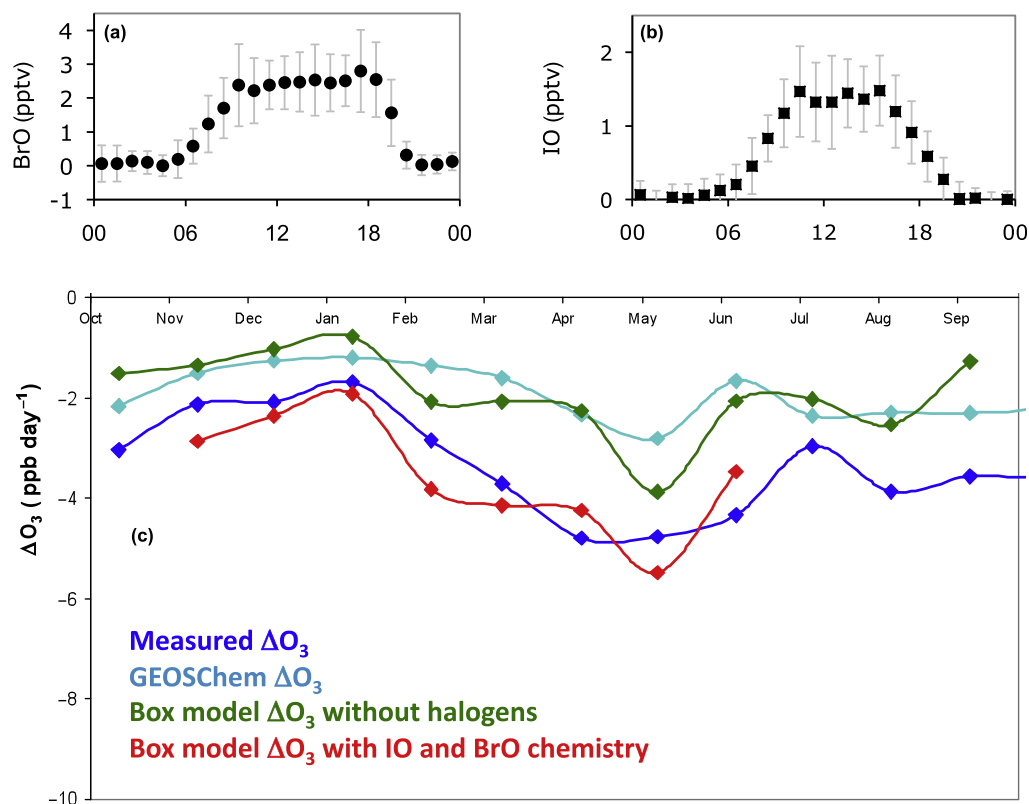


Figure 6 Average hourly (a) BrO, (b) IO mixing ratios measured by LP-DOAS at the Cape Verde Atmospheric Observatory during November 2006–June 2007. Errors (1σ) are indicated as gray lines, and (c) measured and modeled average monthly values of ΔO_3 , defined as the observed daily O_3 destruction at Cape Verde in ppbv (parts per 10^9 by volume). The average loss over the period shown was 3.3 ± 2.6 ppbv per day between 09:00 and 17:00 h UT. The light blue line and symbols represent simulations using the global tropospheric chemistry transport model GEOSChem (with no halogen chemistry). Box model simulations with halogens (constrained using the IO and BrO measurements shown) and without halogens show that IO and BrO chemistry can account for the additional O_3 loss required within GEOSChem. Adapted from Figures 2 and 3 of Read, K.A., Mahajan, A.S., Carpenter, L.J., Evans, M.J., Faria, B.V.E., Heard, D.E., Hopkins, J.R., Lee, J.D., Moller, S., Lewis, A.C., Mendes, L., McQuaidm, J.B., Oetjen, H., Saiz-Lopez, A., Pilling, M.J., Plane, J.M.C., 2008. Extensive halogen-mediated ozone destruction over the tropical Atlantic Ocean. *Nature* 453 (7199), 1232–1235.

Antarctic. At Ny-Ålesund, Spitzbergen, IO was never measured above the detection limit, though total gaseous iodine (measured by neutron activation) reached 8 pptv (Martinez et al., 1999). At Alert, Canada, in springtime, measurements of IO were above the detection limit only on one occasion (Hönninger, 2002). However, IO was recently detected at Hudson Bay in the Canadian Arctic, with levels of up to 3.4 pptv of IO in air that had recently passed over open leads in the sea ice (Mahajan et al., 2010b). Calculations suggested that at the levels observed, IO could enhance the O_3 depletion potential of bromine by a factor of 3. Theoretical studies of the effect of iodine on mercury also suggest that atmospheric mercury depletion events in typical mixtures of bromine can be enhanced significantly by the presence of small amounts of iodine-containing compounds (Calvert and Lindberg, 2004).

The source(s) of reactive iodine to the polar atmosphere and the reason for apparently much more active iodine chemistry in Antarctica remain unknown, but a variety of mechanisms have been proposed. Recent Antarctic measurements suggest very high IO concentrations present in snow interstitial air (Frieß et al.,

2010), with an initial source of iodine postulated to be CH_3I produced in the marginal ice zone (MIZ) and deposited to the snow pack during transport inland. Mahajan et al. (2010b) use modeling studies to suggest that the emission of biogenic VOIC from coastal leads in the sea ice (polynyas) can account for the observed atmospheric concentrations of IO. Polynyas are known to be associated with high rates of biological activity and the dominant phytoplankton group is typically large diatoms, which are known to be an iodocarbon source. Inorganic sources of iodine (I_2 or HOI) to the polar atmosphere have also been suggested. Hill and Manley (2009) recently determined very high production rates of HOI and HOBr from polar marine diatoms (abundant under sea ice), much greater than the previously measured organic halogen rates of release. It is not yet clear how much, if any, of the reactive halogens (HOX and X_2) escape to the atmosphere, but Shaw et al. (2011) show that brine channels in consolidated (even fresh) sea ice offer a negligible trace gas diffusion pathway, though emission from open leads could be important.

Antarctica offers a more biologically active environment than the Arctic with thinner and more porous sea ice (e.g.,

Thomas and Dieckmann, 2003), which may explain the more active iodine chemistry there. At both poles, the occurrence of polynyas, open leads and the MIZ is likely to increase as sea ice continues to thin and retreat, which could enhance the rate of iodine flux to the atmosphere.

Terrestrial Environments: Salt Lakes and Volcanoes

Salt lakes and volcanoes represent environments containing very high halide levels, and both appear to be sources of gas-phase iodine (as well as bromine). Up to 10 pptv of IO have been observed over the surface of the Dead Sea (Zingler and Platt, 2005). The authors proposed the release of iodine from salt deposits via heterogeneous inorganic (IX where X = I, Br, Cl) iodine release induced by liquid phase ozone reactions or catalytic HOX interactions. From volcanoes, a global source strength of 0.11 Gg (range: 0.04–6.6) of iodine per year has been calculated from extrapolation of emissions of acidic iodine species from Mount Etna (Aiuppa et al., 2005).

Stratospheric Iodine

Calculations suggest that only relatively small amounts of iodine (compared with bromine and chlorine) transported to the tropical lower stratosphere, e.g., by tropical deep convection of organic iodine compounds, could make an important contribution to O₃ depletion there (Solomon et al., 1994). However, a number of measurements, made using both ground-based and balloon-borne instruments, have indicated that iodine plays only a minor stratospheric role with reported upper limit mixing ratios in the upper troposphere and lower stratosphere of below 0.2 pptv total inorganic iodine (I_T) (Wennberg et al., 1997; Berthet et al., 2003; Bösch et al., 2003; Pundt et al., 1998; Butz et al., 2009). In contrast, retrievals of differential slant column densities of IO measured using ground-based zenith-sky spectroscopy (Wittrock et al., 2000) suggest mixing ratios up to 0.8 pptv of IO in the high-latitude winter stratosphere technique. This has yet to be confirmed by *in situ* measurements of active iodine.

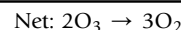
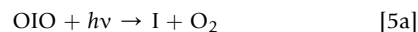
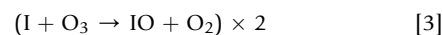
Atmospheric Chemistry of Iodine

The following sections summarize the current literature on atmospherically relevant reactions of iodine. For comprehensive details, the reader is referred to the 2003 and 2012 reviews in *Chemical Reviews* (Carpenter, 2003; Saiz-Lopez et al., 2012).

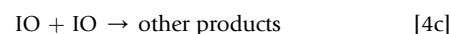
Ozone Destruction

Unlike chlorine and bromine atoms, which are reactive to a range of organic molecules, iodine atoms do not react with either saturated or unsaturated organic compounds. Reaction with O₃ forming the IO radical is their major fate. As noted in Sections [Open Ocean](#) and [Polar Boundary Layer](#), iodine is believed to be responsible for significant photochemical O₃ destruction in the marine and polar lower troposphere. This impact occurs mainly through three catalytic cycles, all of which regenerate I atoms from IO without concomitant O atom formation (which would lead to a null cycle for O₃). At low levels of nitrogen oxides (NO_x), IO self-reaction or reaction with BrO radicals (Cycle 1) and reaction of IO with hydroperoxy radicals (HO₂) (Cycle 2) dominate.

Cycle 1

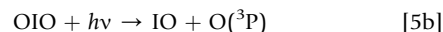
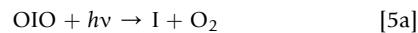


Note that the IO self-reaction has several channels with the IO dimer-forming channel [4b] (~50%) and the OIO-forming channel [4a] (~40%) being dominant (Bloss et al., 2001; Harwood et al., 1997; Sander, 1986; Gómez Martín et al., 2007) at atmospheric pressure:



Experimental data and theoretical calculations indicate that the lifetime of the IO dimer against dissociation to I + OIO is around 1 s under atmospheric conditions, thus OIO is the main product of reactions [4a], [4b], and [4c] (Gómez Martín and Plane, 2009; Kaltsoyannis and Plane, 2008). As well as its self-reaction, IO can react with BrO in a cross reaction that produces OIO and Br (80%) and I + Br (20%) (Sander et al., 2006).

The photolysis rate and pathways of OIO are critical in determining its ozone destruction potential; only channel [5a] will result in net loss of O₃.

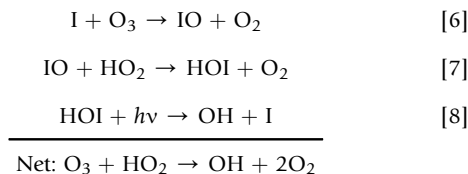


While measurements of the absorption cross sections of OIO have converged (Bloss et al., 2001; Gómez Martín et al., 2005; Joseph et al., 2005; Tucceri et al., 2006), the photolysis yields are still under debate. A number of experimental (Ingham et al., 2000; Cox et al., 1999; Joseph et al., 2005; Tucceri et al., 2006) and computational studies (Misra and Marshall, 1998) have all indicated a high photochemical stability for OIO and/or very low branching ratios for channel [5a]. However, high-resolution spectroscopy of OIO provided evidence for a very short-lived excited state (~200 fs) and indicated that channel [5a] dominates in the visible region (Ashworth et al., 2002). This is consistent with a recent experimental study, utilizing simultaneous measurements of OIO and atomic I (Gómez Martín et al., 2009), which determined an I atom quantum yield of unity from channel [5a]. This results in an atmospheric photolysis rate of $J(OIO) = 0.4 \text{ s}^{-1}$ at noon during summer solstice at 40° N (Saiz-Lopez et al., 2012) and indicates that OIO should be present at only very low levels during daytime. The available field data so far do not help to consolidate these laboratory studies. Daytime OIO observations have been reported in some cases (Stutz et al., GRL, 2007) while in others OIO is only seen at night (Peters et al., 2005; Mahajan et al., 2009) (see Section [Coastlines](#)). Rapid photolysis of OIO enhances the ozone-depleting potential of Cycle 1 but limits the potential role of OIO in new particle formation (see Section [Recycling of Iodine through Aerosol](#)).

Cycle 1 is believed to dominate O₃ destruction when relatively high mixing ratios of halogen oxides are present

(typically, more than 2 pptv of IO and BrO). At lower halogen oxide concentrations, reaction with HO₂ radicals (Cycle 2) becomes important:

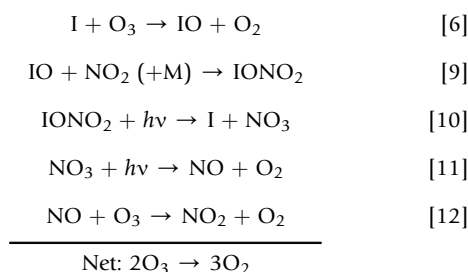
Cycle 2



Note that while Cycle 2 is well understood, the reaction of IO with CH₃O₂ is uncertain both in terms of rate (differing by a factor of 30; Bale et al., 2005; Enami et al., 2006; Dillon et al., 2006, 2010) and in products, with potential channels representing both radical cycling and radical sinks. Bloss et al. (2010) showed that the range of rate constants/products for this reaction alone alters the predicted OH levels by up to 35% in the Antarctic boundary layer.

In semipolluted atmospheres, i.e., NO₂ > 1 ppbv, IONO₂ is a major gas-phase iodine species (Mahajan et al., 2009) and Cycle 3 can contribute to O₃ depletion.

Cycle 3



Photolysis of IONO₂ (which is much faster during the day than thermal dissociation; Kaltsoyannis and Plane, 2008; Saiz-Lopez et al., 2012) via channel [10] of Cycle 3 yields I atoms thus representing an O₃-depleting cycle. The NO₃ fragment produced, however, photolyses to form NO₂ + O with a branching ratio of 0.9 (Atkinson et al., 2009), leading to further ozone formation and reducing the overall O₃-depleting efficiency of this channel. Further, photolysis of the IONO₂ formed is in competition with its removal from the gas phase though uptake to aerosol, although the uptake coefficients and subsequent condensed phase chemistry of IONO₂ are highly uncertain.

In the presence of a few parts per billion of NO₂, models (von Glasow et al., 2002; Pechtl et al., 2006) have predicted very small levels of IO and OIO, with most IO_y converted to IONO₂. However, a number of recent field studies have shown that IO levels are essentially unaffected by several parts per billion of NO_x (Stutz et al., 2007; Whalley et al., 2007; Mahajan et al., 2009). Stutz et al. (2007) postulated that a fast recycling mechanism efficiently regenerates IO from IONO₂; this was later proposed by Mahajan et al. (2009) to occur via an autocatalytic reaction of IONO₂ with atomic I (reaction [13]), in accord with theory (Kaltsoyannis and Plane, 2008):



The I atoms produced from I₂ photolysis will react with IONO₂ rather than O₃ if the ratio [IONO₂]/[O₃] is >0.01 (Kaltsoyannis and Plane, 2008). These reactions thus represent an autocatalytic cycle that limits the build up of IONO₂ and maintain active iodine chemistry and consequent particle formation (see Section *Aerosol Formation and Iodine Oxide Particles*) even in a relatively high NO_x environment.

The forward reaction of I₂ with NO₃ radicals (reaction [13]) represents an additional IO production pathway at night. This reaction has been postulated as the source of nighttime IO observed at Mace Head in association with high NO₃ mixing ratios (Saiz-Lopez et al., 2006).

Recycling of Iodine through Aerosol

The net transfer of iodine from the gas to the condensed phase is reflected by the factor of 100- to 1000-fold enrichment of I in fine-fraction marine aerosol by comparison to the I/Na ratio in seawater (Duce and Hoffman, 1976; Sturges and Barrie, 1988; Baker et al., 2000). While aerosol then represents a net sink for iodine, the interaction of iodine with sea-salt aerosol can accelerate gaseous halogen release and provide a mechanism for maintaining gaseous iodine levels away from emission sources.

The first mechanisms proposed for the release of reactive halogens from aerosol required significant concentrations of nitrogen oxides (Finlayson-Pitts, 1983; Zetzsch et al., 1988). Vogt et al. (1996) suggested an autocatalytic cycle for bromide and chloride release in low NO_x environments and later (Vogt et al., 1999) proposed that the cycling of iodine in sea-salt aerosol could accelerate halogen release, mainly through acid-catalyzed aerosol scavenging of HOI, formed in Cycle 2:



Release of ICl and IBr after interaction of HOI on salt surfaces has been verified in laboratory experiments (Holmes et al., 2001; Mössinger and Cox, 2001) and is believed to significantly increase the gas-phase halogen reservoir (Vogt et al., 1999).

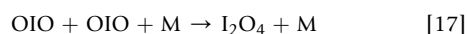
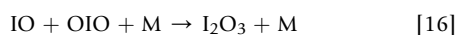
Similarly, the uptake of IONO₂ formed on sea-salt aerosol enhances the release of chlorine and bromine from sea-salt particles into the gas phase, which can then cause further O₃ depletion (McFiggans et al., 2002). Further, the uptake and hydrolysis of IONO₂ on aerosol is a potentially important removal pathway for NO_x in the remote troposphere (McFiggans et al., 2000), leading indirectly to increased rates of ozone destruction (Stutz et al., 1999). Thus, although the direct O₃-depleting potential of Cycle 3 is small (because NO₃ generated from channel [10] produces O₃), the indirect effects of enhanced halogen release from sea-salt aerosol and removal of atmospheric NO_x may be important in reducing boundary layer ozone concentrations.

Aerosol Formation and Iodine Oxide Particles

Bursts of ultrafine particle production events linked to low tide and solar radiation were first observed at the coastal site of Mace Head, Ireland (O'Dowd et al., 1998, 2002). Similar peaks in iodine-containing compounds at low tide (Alicke et al.,

1999; Carpenter et al., 1999, 2001) led to speculation as to the involvement of iodine in the particle formation mechanism. Analysis of new ultrafine particles at Mace Head confirmed that iodine is an important component (Mäkelä et al., 2002) and numerous laboratory investigations have shown that in the presence of ozone, photodissociation of iodine precursors leads to rapid new particle formation composed mainly of iodine oxide particles (IOPs) (Burkholder et al., 2004; Cox and Coker, 1983; Cox et al., 1999; Hoffmann et al., 2001; Jimenez et al., 2003; Saunders and Plane, 2005, 2006; Saunders et al., 2010). IOPs could provide condensation nuclei for other condensable vapors and grow to the point of becoming cloud condensation nuclei (CCN), which would impact on the radiative balance of the atmosphere and hence on climate (Burkholder et al., 2004; Hoffmann et al., 2001; Jimenez et al., 2003; O'Dowd et al., 1998, 2002a; McFiggans et al., 2010).

Although strong links between iodine, new particles, and particle growth in the coastal MBL have now been established, the chemistry of the particle-forming higher oxides of iodine is still poorly understood. IOPs are thought to be initiated by the release of I_2 from kelps (see Section [Coastlines](#)) and consequent formation of IO, followed by recombination reactions of IO and OIO (formed from the IO self-reaction and $IO + BrO$ cross reactions) to form higher oxides (Gómez Martín et al., 2007):



Theoretical work indicates that I_2O_3 is stable at MBL temperatures (Kaltsoyannis and Plane, 2008), but there is little quantitative information on the higher iodine oxides. As discussed earlier, recent work suggests a short photolysis lifetime for OIO, increasing the ozone depletion potential but decreasing the potential for IOP formation.

Solid IOPs are mostly likely comprised of I_2O_4 and I_2O_5 (Saunders and Plane, 2005; Saunders et al., 2010). Saunders and Plane (2005) speculated that the oxidation of I_2O_4 by O_3 was responsible for production of gas-phase I_2O_5 , which then polymerized to produce IOPs. However, a more recent study (Saunders et al., 2010), utilizing N_2O photolysis to form O atoms and hence IO by reaction with I_2 , has demonstrated that IOPs can form in the absence of O_3 and are thus most likely initiated by the spontaneous polymerization of I_2O_3 and I_2O_4 , formed in reactions [16] and [17]. The dry particles are believed to restructure in the solid phase to I_2O_5 and I_2 . In humid marine environments, the IOPs are likely to hydrolyze to HIO_3 (iodate) (Saiz-Lopez et al., 2012).

Once IOPs form, they will only have an impact on climate if they grow to a sufficient size to either scatter and absorb solar radiation (direct impact) or to enhance CCN concentrations (indirect impact). Even in coastal environments with high macroalgal I_2 emissions, the supply of iodine oxides is likely to be limited, thus the IOPs will only grow by condensation of other vapor species such as water, sulfuric acid, or ammonia. In fact, the accommodation of H_2SO_4 vapor on IOP has recently been shown to be very efficient, particularly at high relative humidities (Saunders et al., 2010). In the coastal MBL, uptake of H_2SO_4 onto IOPs is likely to be accompanied by H_2O and NH_3 (Kulmala and Kerminen, 2008; Kulmala et al., 2002; O'Dowd and De Leeuw, 2007). So far, it is not established

whether particles originating from IOPs can act efficiently as CCN at realistic marine supersaturations nor whether, if this does occur, that the regions affected are sufficiently large to induce significant radiative impacts.

Radioactive Iodine

Radioactive isotopes of iodine are produced in fission of uranium and plutonium. Radioactive iodine can be inhaled as a gas or ingested in food or water and is of concern for human health, with a direct link to thyroid cancer. Radioactive I-131 (I-131; half-life 8.02 days) is assumed to cause the majority of the excess thyroid cancers seen after bomb fallout and severe nuclear reactor accidents. Children are at greater risk from radioactive iodine exposure than adults, and thyroid cancers attributed to I-131 exposure during childhood continue to occur throughout adulthood (Brenner et al., 2011). The accident at the Chernobyl nuclear power plant in 1986 was the most severe in the history of the nuclear power industry, releasing huge quantities of radionuclides over large areas of Belarus, Ukraine, and the Russian Federation and causing several thousand cases of thyroid cancer (Williams, 2008). The release of radioactive material to the atmosphere following the Japanese Fukushima nuclear accident on 11 March 2011 has been estimated to be approximately 10% of the Chernobyl accident (International Atomic Energy Agency). Contamination of the marine environment occurred through atmospheric fallout or washout with precipitation and through discharges of contaminated water into the sea.

Iodine-129 (I-129; half-life 15.7 million years), mainly released in the past from above-ground nuclear testing and now from nuclear reprocessing facilities, is one of the most persistent radionuclides, and can participate in the biogeochemical cycling of iodine, as well as potentially accumulating in human thyroid glands (Hou et al., 2000). Historical and ongoing release of I-129 has caused a significant increase of this isotope since the prenuclear era within the ocean mixed layer, rivers, and lakes (Reithmeier et al., 2005, 2010), in marine sediments (Fehn et al., 1986; Lopez-Gutierrez et al., 2004), in glacier ice (Wagner et al., 1996; Reithmeier et al., 2006), and in moss (Sumerling, 1984; Rucklidge et al., 1994). In Europe, the majority of I-129 is currently released in liquid form from the reprocessing facilities at La Hague (France) and Sellafield (UK) (Hou et al., 2007; Reithmeier et al., 2005, 2010) to the Irish Sea and to the English Channel where it is present mostly in the oceans' upper water layer (Reithmeier et al., 2010). Chemical speciation of I-129 in the ocean is mainly as iodide and iodate, although the iodide/iodate ratio for I-129 appears to be higher than that for the natural isotope I-127 (Hou et al., 2007). Although the annual reemission rate of I-129 from the ocean to the atmosphere is reported to be less than about 0.3% of its load in the upper layer of the ocean, this contribution is comparable to that of the combined gaseous release from La Hague and Sellafield (Reithmeier et al., 2006). I-129 originating from gaseous release from reprocessing facilities, and from the fraction of I-129 that escapes the oceans' surface water, is primarily deposited by rain and, to a much smaller amount, by dry deposition (Lopez-Gutierrez et al., 2001). Although the current environmental levels of I-129 do not

present a significant radiation risk for human health, continued release at the present level could present a long-term radiological risk (Hou et al., 2000).

See also: Biogeochemical Cycles: Bromine; Heavy Metals; Sulfur Cycle.

References

- Aluppa, A., Federico, C., Franco, A., Giudice, G., Gurrieri, S., Inguaggiato, S., Liuzzo, M., McGonigle, A.J.S., Valenza, M., 2005. Emission of bromine and iodine from Mount Etna volcano. *Geochim. Geophys. Geosyst.* 6 (8), Q08008.
- Alicke, B., Hebestreit, K., Stutz, J., Platt, U., 1999. Iodine oxide in the marine boundary layer. *Nature* 397 (6720), 572–573.
- Allan, B.J., McFiggans, G., Plane, J.M.C., Coe, H., 2000. Observations of iodine monoxide in the remote marine boundary layer. *J. Geophys. Res.* 105 (D11), 14363–14369.
- Amachi, S., 2008. Microbial contribution to global iodine cycling: volatilization, accumulation, reduction, oxidation, and sorption of iodine. *Microbes Environ.* 23 (4), 269–276.
- Amachi, S., Kamagata, Y., Kanagawa, T., Muramatsu, Y., 2001. Bacteria mediate methylation of iodine in marine and terrestrial environments. *Appl. Environ. Microbiol.* 67, 2718–2722.
- Amachi, S., Muramatsu, Y., Akiyama, Y., Miyazaki, K., Yoshiki, S., Hanada, S., Kamagata, Y., Ban-nai, T., Shinoyama, H., Fujii, T., 2005. Isolation of iodide-oxidizing bacteria from iodide-rich natural gas brines and seawaters. *Microb. Ecol.* 49, 547–557.
- Archer, S.D., Goldson, L.E., Liddicoat, M.I., Cummings, D.G., Nightingale, P.D., 2007. Marked seasonality in the concentrations and sea–air flux of volatile iodocarbon compounds in the western English Channel. *J. Geophys. Res.* 112, C08009. <http://dx.doi.org/10.1029/2006JC003963>.
- Ashworth, S.H., Allan, B.J., Plane, J.M.C., 2002. High resolution spectroscopy of the IO radical: implications for the ozone-depleting potential of iodine. *Geophys. Res. Lett.* 29.
- Atkinson, R., Baulch, D.L., Cox, R.A., Crowley, J.N., Hampson, R.F., Hynes, R.G., Jenkin, M.E., Rossi, M.J., Troe, J., Evaluated kinetic and photochemical data for atmospheric chemistry. <http://www.iupac-kinetic.ch.cam.ac.uk/> (accessed January 2014).
- Baker, A.R., 2004. Inorganic iodine speciation in tropical Atlantic aerosol. *Geophys. Res. Lett.* 31 (23), L23S02. <http://dx.doi.org/10.1029/2004GL020144>.
- Baker, A.R., 2005. Marine aerosol iodine chemistry: the importance of soluble organic iodine. *Environ. Chem.* 2, 295–298.
- Baker, A.R., Thompson, D., Campos, M.L.A.M., Parry, S.J., Jickells, T.D., 2000. Iodine concentration and availability in atmospheric aerosol. *Atmos. Environ.* 34, 4331–4336.
- Baker, A.R., Tunnicliffe, C., Jickells, T.D., 2001. Iodine speciation and deposition fluxes from the marine atmosphere. *J. Geophys. Res. Atmos.* 106 (D22), 28743–28749.
- Bale, C.S.E., Canosa-Mas, C.E., Shallcross, D.E., Wayne, R.P., 2005. A discharge–flow study of the kinetics of the reactions of IO with CH₃O₂ and CF₃O₂. *Phys. Chem. Chem. Phys.* 7, 2164–2172.
- Bale, C.S.E., Ingham, T., Commane, R., Heard, D., Bloss, W., 2008. Novel measurements of atmospheric iodine species by resonance fluorescence. *J. Atmos. Chem.* 60 (1–102), 51–70.
- Ball, S.M., Hollingsworth, A.M., Humbles, J., Leblanc, C., Potin, P., McFiggans, G., 2010. Spectroscopic studies of molecular iodine emitted into the gas phase by seaweed. *Atmos. Chem. Phys.* 10 (13), 6237–6254.
- Bell, N., Hsu, L., Jacob, D.J., Schultz, M.G., Blake, D.R., Butler, J.H., King, D.B., Lobert, J.M., Maier-Reimer, E., 2002. Methyl iodide: atmospheric budget and use as a tracer of marine convection in global models. *J. Geophys. Res.* 107 (D17), 4340.
- Berthet, G. I., Renard, J.B., Chartier, M., Pirre, M., Robert, C., 2003. Analysis of OBrO, IO, and OIO absorption signature in UV-visible spectra measured at night and at sunrise by stratospheric balloon-borne instruments. *J. Geophys. Res.* 108.
- Bitter, M., Ball, S.M., Povey, I.M., Jones, R.L., 2005. A broadband cavity ringdown spectrometer for in-situ measurements of atmospheric trace gases. *Atmos. Chem. Phys.* 5 (9), 2547–2560.
- Bloss, W.J., Rowley, D.M., Cox, R.A., Jones, R.L.J., 2001. Kinetics and products of the IO self-reaction. *J. Phys. Chem.* 105 (33), 7840–7854.
- Bloss, W.J., Camredon, M., Lee, J.D., Heard, D.E., Plane, J.M.C., Saiz-Lopez, A., Bauguittie, S.J.B., Salmon, R.A., Jones, A.E., 2010. Coupling of HO(x), NO(x) and halogen chemistry in the Antarctic boundary layer. *Atmos. Chem. Phys.* 10 (21), 10187–10209.
- Bösch, H., Camy-Peyret, C., Chipperfield, M.P., Fitzenberger, R., Harder, H., Platt, U., Pfeilsticker, K., 2003. Upper limits of stratospheric IO and OIO inferred from center-to-limb-darkening-corrected balloon-borne solar occultation visible spectra: implications for total gaseous iodine and stratospheric ozone. *J. Geophys. Res.* 108.
- Brenner, A.V., Tronko, M.D., Hatch, M., Bogdanova, T.I., Olynyk, V.A., Lubin, J.H., Zablotska, L.B., Tereschenko, V.P., McConnell, R.J., Zamotaeva, G.A., O’Kane, P., Bouville, A.C., Chaykovskaya, L.V., Greenebaum, E., Paster, I.P., Shpak, V.M., Ron, E., 2011. I-131 dose-response for incident thyroid cancers in Ukraine related to the Chernobyl accident. *Environ. Health Perspect.* 119 (7). <http://dx.doi.org/10.1289/ehp.1002674>.
- Brownell, D.K., Moore, R.M., Cullen, J.J., 2010. Production of methyl halides by *Prochlorococcus* and *Synechococcus*. *Glob. Biogeochem. Cycles* 24 (2).
- Burkholder, J.B., Curtius, J., Ravishankara, A.R., Lovejoy, E.R., 2004. Laboratory studies of the homogeneous nucleation of iodine oxides. *Atmos. Chem. Phys.* 4 (1), 19–34.
- Butler, J.H., King, D.B., Lobert, J.R.M., Montzka, D.S., Yvon-Lewis, S.A., Hall, B.D., Warwick, N.J., Mondeel, D.J., Aydin, M., Elkins, J.W., 2007. Oceanic distributions and emissions of short-lived halocarbons. *Glob. Biogeochem. Cycles* 21, GB1023.
- Butz, A., Boesch, H., Camy-Peyret, C., Chipperfield, M.P., Dorf, M., Kreyer, D.S., Kritten, L., Prados-Roman, C., Schwaerzle, J., Pfeilsticker, K., 2009. Constraints on inorganic gaseous iodine in the tropical upper troposphere and stratosphere inferred from balloon-borne solar occultation observations. *Atmos. Chem. Phys.* 9, 7229–7242.
- Calvert, J.G., Lindberg, S.E., 2004. The potential influence of iodine-containing compounds on the chemistry of the troposphere in the polar spring. I. Ozone depletion. *Atmos. Environ.* 38, 5087–5104.
- Carpenter, L.J., 2003. Iodine in the marine boundary layer. *Chem. Rev.* 103 (12), 4953–4962.
- Carpenter, L.J., Sturges, W.T., Penkett, S.A., Liss, P.S., Alicke, B., Hebestreit, K., Platt, U., 1999. Short-lived alkyl iodides and bromides at Mace Head, Ireland: links to biogenic sources and halogen oxide production. *J. Geophys. Res.* 104 (D1), 1679–1689.
- Carpenter, L.J., Malin, G., Kuepper, F., Liss, P.S., 2000. Novel biogenic iodine-containing trihalomethanes and other short-lived halocarbons in the coastal East Atlantic. *Glob. Biogeochem. Cycles* 14, 1191–1204.
- Carpenter, L.J., Hebestreit, K., Platt, U., Liss, P.S., 2001. Coastal zone production of IO precursors: a 2-dimensional study. *Atmos. Chem. Phys.* 1, 9–18.
- Carpenter, L.J., Wevill, D.J., Palmer, C.J., Michels, J., 2007. Depth profiles of volatile iodine and bromine-containing halocarbons in coastal Antarctic waters. *Marine Chem.* 103 (3–4), 227–236.
- Carpenter, L.J., Fleming, Z.L., Read, K.A., Lee, J.D., Moller, S.J., Hopkins, J.R., Purvis, R.M., Lewis, A.C., Muller, K., Heinold, B., Herrmann, H., Wadinga Fomba, K., van Pinxteren, D., Muller, C., Tegen, I., Wiedensohler, A., Muller, T., Niedermeier, N., Achterberg, E.P., Patey, M.D., Kozlova, E.A., Heimann, M., Heard, D.E., Plane, J.M.C., Mahajan, A., Oetjen, H., Ingham, T., Stone, D., Whalley, L.K., Evans, M.J., Pilling, M.J., Leigh, R.J., Monks, P.S., Karunaharan, A., Vaughan, S., Arnold, S.R., Tschritter, J., Pöhler, D., Frieb, U., Holla, R., Mendes, L.M., Lopez, H., Faria, B., Manning, A.J., Wallace, D.W.R., 2011. Seasonal characteristics of tropical marine boundary layer air measured at the Cape Verde Atmospheric Observatory. *J. Atmos. Chem.* <http://dx.doi.org/10.1007/s10874-011-9206-1>.
- Carpenter, L.J., MacDonald, S.M., Shaw, M.D., Kumar, R., Saunders, R.W., Parthipan, R., Wilson, J., Plane, J.M.C., 2013. Atmospheric iodine levels influenced by sea surface emissions of inorganic iodine. *Nature Geosci.* 6 (2), 108–111.
- Chambers, R.M., Heard, A.C., Wayne, R.P., 1992. Inorganic gas-phase reactions of the nitrate radical: iodine + nitrate radical and iodine atom + nitrate radical. *J. Phys. Chem.* 96 (8), 3321–3331.
- Chance, R., Baker, A.R., Küpper, F.C., Hughes, C., Kloareg, B., Malin, G., 2009. Release and transformations of inorganic iodine by marine macroalgae. *Estuarine Coastal Shelf Sci.* 82 (3), 406–414.
- Commane, R., Seitz, K., Bale, C.S., Bloss, W.J., Buxmann, J., Ingham, T., Platt, U., Ohler, D.P., Heard, D.E., 2011. Iodine monoxide at a clean marine coastal site: observations of high frequency variations and inhomogeneous distributions. *Atmos. Chem. Phys. Discuss* 11 (13), 6721–6733.
- Cotter, E.S.N., Booth, N.J., Canosa-Mas, C.E., Wayne, R.P., 2001. Release of iodine in the atmospheric oxidation of alkyl iodides and the fates of iodinated alkoxy radicals. *Atmos. Environ.* 35 (12), 2169–2178.
- Courtois, B., 1813. Découverte d’une substance nouvelle dans le Vareck. *Annales de Chimie de Physique* 88, 304–310.

- Cox, R.A., Coker, G.B., 1983. Absorption cross section and kinetics of IO in the photolysis of CH_3I in the presence of ozone. *J. Phys. Chem.* 87, 4478–4484.
- Cox, R.A., Bloss, W., Jones, R.L., Rowley, D.M., 1999. OIO and the atmospheric cycle of iodine. *Geophys. Res. Lett.* 26.
- Dembitsky, V.M., 2006. Biogenic iodine and iodine-containing metabolites. *Nat. Prod. Commun.* 1, 139–175.
- Dickerson, R.R., Rhoads, K.P., Carsey, T.P., Oltmans, S.J., Burrows, J.P., Crutzen, P.J., 1999. Ozone in the remote marine boundary layer: a possible role for halogens. *J. Geophys. Res. Atmos.* 104 (21), 385–395.
- Dillon, T.J., Blitz, M.A., Heard, D.E., 2006. Determination of the rate coefficients for the reactions $\text{IO} + \text{NO}_2 + \text{M} (\text{Air}) \rightarrow \text{IONO}_2 + \text{M}$ and $\text{O}(^3\text{P}) + \text{NO}_2 \rightarrow \text{O}_2 + \text{NO}$ using laser-induced fluorescence spectroscopy. *J. Phys. Chem. A* 110 (21), 6995–7002.
- Dillon, T.J., Tucceri, M.E., Crowley, J.N., 2010. Rate coefficients for the reaction of iodine oxide with methyl peroxy radicals. *ChemPhysChem* 11 (18), 4011–4018.
- Dimmer, C.H., Simmonds, P.G., Nickless, G., Bassford, M.R., 2001. Biogenic fluxes of halomethanes from Irish peatland ecosystems. *Atmos. Environ.* 35, 321–330.
- Dixneuf, S., Ruth, A.A., Vaughan, S.R., Varma, M., Orphal, J., 2009. The time dependence of molecular iodine emission from *Laminaria digitata*. *Atmos. Chem. Phys.* 9, 823–829.
- Duce, R.A., Hoffman, E.J., 1976. Chemical fractionation at the air/sea interface. *Annu. Rev. Earth Planet. Sci.* 4, 187.
- Elliott, S., Rowland, F.S., 1993. Nucleophilic substitution rates and solubilities for methyl halides in seawater. *Geophys. Res. Lett.* 20 (11), 1043–1046.
- Enami, S., Yamanaka, T., Hashimoto, S., Kawasaki, M., Nakano, Y., Ishiwata, T., 2006. Kinetic study of IO radical with RO_2 ($\text{R} = \text{CH}_3$, C_2H_5 , and CF_3) using cavity ring-down spectroscopy. *J. Phys. Chem. A* 110 (32), 9861–9866.
- Fehn, U., Holdren, G.R., Elmore, D., Brunelle, T., Teng, R., Kubik, P.W., 1986. Determination of natural and anthropogenic ^{129}I in marine sediments. *Geophys. Res. Lett.* 13 (1), 137–139.
- Finlayson-Pitts, B.J., 1983. Reaction of NO_2 with NaCl and atmospheric implications of NOCl formation. *Nature* 306, 676–677.
- Frieß, U., Wagner, T., Pundt, I., Pfeilsticker, K., Platt, U., 2001. Spectroscopic measurements of tropospheric iodine oxide at Neumayer station, Antarctica. *Geophys. Res. Lett.* 28.
- Frieß, U., Deutschmann, T., Gilfedder, B.S., Weller, R., Platt, U., 2010. Iodine monoxide in the Antarctic snowpack. *Atmos. Chem. Phys.* 10 (5), 2439–2456.
- Fuse, H., Inoue, H., Murakami, K., Takimura, O., Yamaoka, Y., 2003. Production of free and organic iodine by *Roseovarius* spp. *FEMS Microbiol. Lett.* 229, 189–194.
- Galbally, I.E., Bentley, S.T., Meyer, C.P., 2000. Mid-latitude marine boundary-layer ozone destruction at visible sunrise observed at Cape Grim, Tasmania, 41 degrees S. *Geophys. Res. Lett.* 27 (23), 3841–3844.
- Garland, J.A., Curtis, H., 1981. Emission of iodine from the sea surface in the presence of ozone. *J. Geophys. Res.* 86.
- Garland, J.A., Elzerman, A.W., Penkett, S.A., 1980. The mechanism for dry deposition of ozone to seawater surfaces. *J. Geophys. Res.* 85.
- Gilfedder, B.S., Lai, S., Petri, M., Biester, H., Hoffmann, T., 2008. Iodine speciation in rain, snow and aerosols. *Atmos. Chem. Phys.* 8, 6069–6084.
- Gómez Martín, J.C., Plane, J.M.C., 2009. Determination of the O-IO bond dissociation energy by photofragment excitation spectroscopy. *Chem. Phys. Lett.* 474, 79–83.
- Gómez Martín, J.C., Spietz, P., Burrows, J.P., 2005. Spectroscopic studies of the I_2/O_3 photochemistry: Part 1: Determination of the absolute absorption cross sections of iodine oxides of atmospheric relevance. *J. Photochem. Photobiol. A* 176 (1–3) (In Honour of Professor Richard P. Wayne), 15–38.
- Gómez Martín, J.C., Spietz, P., Burrows, J.P., 2007. Kinetic and mechanistic studies of the I_2/O_3 photochemistry. *J. Phys. Chem. A* 111 (2), 306–320.
- Gómez Martín, J.C., Ashworth, S.H., Mahajan, A.S., Plane, J.M.C., 2009. Photochemistry of OIO: laboratory study and atmospheric implications. *Geophys. Res. Lett.* 36, L09802. <http://dx.doi.org/10.1029/2009GL037642>.
- Gschwend, P.M., Macfarlane, J.K., Newman, K.A., 1985. Volatile halogenated organic compounds released to seawater from temperate marine macroalgae. *Science* 227, 1033–1035.
- Happell, J.D., Wallace, D.W.R., 1996. Methyl iodide in the Greenland/Norwegian seas and the tropical Atlantic Ocean: evidence for photochemical production. *Geophys. Res. Lett.* 23, 2105–2108.
- Harper, D.B., 1985. Halomethane from halide ion – a highly efficient fungal conversion of environmental significance. *Nature* 315 (6014), 55–57.
- Harwood, M.H., Burkholder, J.B., Hunter, M., Fox, R.W., Ravishankara, A.R., 1997. Absorption cross sections and self-reaction kinetics of the IO radical. *J. Phys. Chem. A* 101 (5), 853–863.
- Hill, V.L., Manley, S.L., 2009. Release of reactive bromine and iodine from diatoms and its possible role in halogen transfer in polar and tropical oceans. *Limnol. Oceanogr.* 54, 812–822.
- Hoffmann, T., O'Dowd, C.D., Seinfeld, J.H., 2001. Iodine oxide homogeneous nucleation: an explanation for coastal new particle production. *Geophys. Res. Lett.* 28 (10), 1949–1952.
- Holmes, N.S., Adams, J.W., Crowley, J.N., 2001. Uptake and reaction of HOI and IONO_2 on frozen and dry NaCl/NaBr surfaces and H_2SO_4 . *Phys. Chem. Chem. Phys.* 3 (9), 1679–1687.
- Hönninger, G., 2002. Halogen Oxide Studies in the Boundary Layer by Multi Axis Differential Optical Absorption Spectroscopy and Active Longpath-DOAS (Ph.D. thesis). University of Heidelberg, Heidelberg, Germany.
- Hou, X., Dahlgard, H., Nielsen, S.P., Ding, W., 2000. Iodine-129 in human thyroids and seaweed in China. *Sci. Total Environ.* 246 (2–3), 285–291. [http://dx.doi.org/10.1016/S0048-9697\(99\)00466-0](http://dx.doi.org/10.1016/S0048-9697(99)00466-0).
- Hou, X., Aldahan, A., Nielsen, S.P., Possnert, G., Nies, H., Hedfors, J., 2007. Speciation of ^{129}I and ^{127}I in seawater and implications for sources and transport pathways in the North Sea. *Environ. Sci. Technol.* 41 (17), 5993–5999. <http://dx.doi.org/10.1021/es070575x>.
- Huang, R.J., Seitz, K., Buxmann, K., Poehler, D., Hornsby, K.E., Carpenter, L.J., Platt, U., Hoffmann, T., 2010. In situ measurements of molecular iodine in the marine boundary layer: the link to macroalgal macroalgae and the implications for O_3 , IO, OIO and NO_x . *Atmos. Chem. Phys.* 10, 4823–4833.
- Hughes, C., Malin, G., Nightingale, P.D., Liss, P.S., 2006. The effect of light stress on the release of volatile iodocarbons by three species of marine microalgae. *Limnol. Oceanogr.* 51 (6), 2849–2854.
- Hughes, C., Malin, G., Turley, C.M., Keely, B.J., Nightingale, P.D., Liss, P.S., 2008. The production of volatile iodocarbons by biogenic marine aggregates. *Limnol. Oceanogr.* 53, 867–872.
- Ingham, T., Cameron, M., Crowley, J.N., 2000. Photodissociation of IO (355 nm) and OIO (532 nm): quantum yields for $\text{O}(^3\text{P})$ and $\text{I}(^2\text{P}_{1/2})$ production. *J. Phys. Chem. A* 104 (34), 8001–8010.
- Itoh, N., Tsujita, M., Ando, T., Hisatomi, G., Higashi, T., 1997. Formation and emission of monohalomethanes from marine algae. *Phytochemistry* 45 (1), 67–73.
- Jammoul, A., Dumas, S., D'Anna, B., George, C., 2009. Photoinduced oxidation of sea salt halides by aromatic ketones: a source of halogenated radicals. *Atmos. Chem. Phys.* 9 (13), 4229–4237.
- Jimenez, J.L., Bahreini, R., Cocker, D.R., Zhuang, H., Varutbangkul, V., Flagan, R.C., Seinfeld, J.H., O'Dowd, C.D., Hoffmann, T., 2003. New particle formation from photooxidation of diiodomethane (CH_2I_2). *J. Geophys. Res.* 108.
- Jones, C.E., Carpenter, L.J., 2005. Photolysis of CH_2I_2 , CH_2ICl and CH_2IBr in water and seawater by solar radiation. *Environ. Sci. Technol.* 39, 6130–6137.
- Jones, C.E., Carpenter, L.J., 2007. Chemical destruction of CH_3I , $\text{C}_2\text{H}_5\text{I}$, $1\text{-C}_3\text{H}_7\text{I}$, and $2\text{-C}_3\text{H}_7\text{I}$ in saltwater. *Geophys. Res. Lett.* 34 (13).
- Jones, C.E., Hornsby, K.E., Sommariva, R., Dunk, R.M., von Glasow, R., McFiggans, G., Carpenter, L.J., 2010. Quantifying the contribution of marine organic gases to atmospheric iodine. *Geophys. Res. Lett.* 37 (18), L18804.
- Joseph, D.M., Ashworth, S.H., Plane, J.M.C., 2005. The absorption cross-section and photochemistry of OIO. *J. Photochem. Photobiol. A* 176, 68–77.
- Kaltsayannis, N., Plane, J.M.C., 2008. Quantum chemical calculations on a selection of iodine-containing species (IO, OIO, INO_3 , $(\text{IO})_2$, I_2O_3 , I_2O_4 and I_2O_5) of importance in the atmosphere. *Phys. Chem. Chem. Phys.* 10 (13), 1723–1733.
- Keppler, F., Eiden, R., Niedan, V., Pracht, J., Schöler, H.F., 2000. Halocarbons produced by natural oxidation processes during degradation of organic matter. *Nature* 403 (6767), 298–301.
- Kulmala, M., Kerminen, V.M., 2008. On the formation and growth of atmospheric nanoparticles. *Atmos. Res.* 90, 132.
- Kulmala, M., Korhonen, P., Napari, I., Karlsson, A., Berresheim, H., O'Dowd, C.D., 2002. Aerosol formation during PARFORCE: ternary nucleation of H_2SO_4 , NH_3 , and H_2O . *J. Geophys. Res. Atmos.* 107 (D19), 8111.
- Küpper, F.C., Schweigert, N., ArGall, E., Legendre, J.M., Vilter, H., Kloareg, B., 1998. Iodine uptake in Laminariales involves extracellular, haloperoxidase-mediated oxidation of iodide. *Planta* 207, 163–171.
- Küpper, F.C., Carpenter, L.J., McFiggans, G.B., 2008. Iodide accumulation provides kelp with an inorganic antioxidant impacting atmospheric chemistry. *Proc. Natl. Acad. Sci. U.S.A.* 105 (19), 6954–6958.
- Laturnus, F., Giese, B., Wiencke, C., Adams, F.C., 2000. Low-molecular-weight organiodine and organobromine compounds released by polar macroalgae – the influence of abiotic factors. *Fres. J. Anal. Chem.* 368 (2–3), 297–302.
- Laturnus, F., Svensson, T., Wiencke, C., Öberg, G., 2004. Ultraviolet radiation affects emission of ozone-depleting substances by marine macroalgae: results from a laboratory incubation study. *Environ. Sci. Technol.* 38 (24), 6605–6609.

- Leigh, R.J., Ball, S.M., Whitehead, J., Leblanc, C., Shillings, A.J.L., Mahajan, A.S., Oetjen, H., Dorsey, J.R., Gallagher, M., Jones, R.L., Plane, J.M.C., Potin, P., McFiggans, G., 2009. Measurements and modelling of molecular iodine emissions, transport and photodestruction in the coastal region around Roscoff. *Atmos. Chem. Phys.* 10 (23), 2010, p11823–p11838.
- Leigh, R.J., Ball, S.M., Whitehead, J., Leblanc, C., Shillings, A.J.L., Mahajan, A.S., Oetjen, H., Lee, J.D., Jones, C.E., Dorsey, J.R., Gallagher, M., Jones, R.L., Plane, J.M.C., Potin, P., McFiggans, G., 2010. Measurements and modelling of molecular iodine emissions, transport and photodestruction in the coastal region around Roscoff. *Atmos. Chem. Phys.* 10, 11823–11838.
- Lopez-Gutierrez, J.M., Garcia-Leon, M., Schnabel, C.H., Suter, M., Sinal, H.-A., Szidat, S., 2001. Wet and dry deposition of ^{129}I in Seville (Spain) measured by accelerator mass spectrometry. *J. Environ. Radioact.* 55, 269–282.
- Lopez-Gutierrez, J.M., Santos, F.J., Garcia-Leon, M., Schnabel, C.H., Sinal, H.A., Ernst, T.H., 2004. Levels and temporal variability of ^{129}I concentrations and $^{129}\text{I}/^{127}\text{I}$ isotopic ratios in atmospheric samples from southern Spain. *Nucl. Instrum. Methods Phys. Res. B*, 223–224, 495–500.
- Mahajan, A.S., Oetjen, H., Saiz-Lopez, A., Lee, J.D., McFiggans, G.B., Plane, J.M.C., 2009. Reactive iodine species in a semi-polluted environment. *Geophys. Res. Lett.* 36, L16803. <http://dx.doi.org/10.1029/2009GL038018>.
- Mahajan, A.S., Plane, J.M.C., Oetjen, H., Mendes, L., Saunders, R.W., Saiz-Lopez, A., Jones, C.E., Carpenter, L.J., McFiggans, G.B., 2010a. Measurement and modelling of tropospheric reactive halogen species over the tropical Atlantic Ocean. *Atmos. Chem. Phys.* 10, 4611–4624.
- Mahajan, A.S., Shaw, M., Oetjen, H., Hornsby, K.E., Carpenter, L.J., Kaleschke, L., Tian-Kunze, X., Lee, J.D., Moller, S.J., Edwards, P., Commane, R., Ingham, T., Heard, D.E., Plane, J.M.C., 2010b. Evidence of reactive iodine chemistry in the Arctic boundary layer. *J. Geophys. Res.* 115, D20303. <http://dx.doi.org/10.1029/2009JD013665>.
- Mäkelä, J.M., Hoffmann, T., Holze, C., Väkevä, M., Suni, T., Mattila, T., Aalto, P.P., Tapper, U., Kauppinen, E.I., O'Dowd, C.D., 2002. Biogenic iodine emissions and identification of end-products in coastal ultrafine particles during nucleation bursts. *J. Geophys. Res. Atmos.* 107 (D19), 8110.
- Manley, S.L., Barbero, P.E., 2001. Physiological constraints on bromoform (CHBr_3) production by *Ulva lactuca* (Chlorophyta). *Limnol. Oceanogr.* 46 (6), 1392–1399.
- Manley, S.L., Dastoor, M.N., 1987. Methyl halide (CH_3X) production from the giant kelp, *Macrocystis*, and estimates of global CH_3X production by kelp. *Limnol. Oceanogr.* 32, 709–715.
- Manley, S.L., de la Cuesta, J.L., 1997. Methyl iodide production from marine phytoplankton cultures. *Limnol. Oceanogr.* 42 (1), 142–147.
- Manley, S.L., Goodwin, K., North, W.J., 1992. Laboratory production of bromoform, methylene bromide, and methyl-iodide by macroalgae and distribution in nearshore southern California waters. *Limnol. Oceanogr.* 37 (8), 1652–1659.
- Martinez, M., Arnold, T., Perner, D., 1999. The role of bromine and chlorine chemistry for Arctic ozone depletion events in Ny-Alesund and comparison with model calculations. *Ann. Geophysicae* 17, 941–956.
- Martino, M., Liss, P.S., Plane, J.M.C., 2005. The photolysis of dihalomethanes in surface seawater. *Environ. Sci. Technol.* 39 (18), 7097–7101.
- Martino, M., Mills, G.P., Woeltjen, J., Liss, P.S., 2009. A new source of volatile organoiodine compounds in surface seawater. *Geophys. Res. Lett.* 36, L01609.
- McDonnell, G., Russell, A.D., 1999. Antiseptics and disinfectants: activity, action, and resistance. *Clin. Microbiol. Rev.* 12, 147–179.
- McFiggans, G., Plane, J.M.C., Allan, B.J., Carpenter, L.J., Coe, H., O'Dowd, C., 2000. A modeling study of iodine chemistry in the marine boundary layer. *J. Geophys. Res.* 105, 14371–14385.
- McFiggans, G., Cox, R.A., Mossinger, J.C., Allan, B.J., Plane, J.M.C., 2002. Active chlorine release from marine aerosols: roles for reactive iodine and nitrogen species. *J. Geophys. Res.* 107 (D15). Art. no. 4271.
- McFiggans, G., Coe, H., Burgess, R., Allan, J., Cubison, M., Alfarra, M.R., Saunders, R., Saiz-Lopez, A., Plane, J.M.C., Wevill, D.J., Carpenter, L.J., Rickard, A.R., Monks, P.S., 2004. Direct evidence for coastal iodine particles from *Laminaria* macroalgae – linkage to emissions of molecular iodine. *Atmos. Chem. Phys.* 4, 701–713.
- McFiggans, G., Bale, C.S.E., Ball, S.M., Beames, J.M., Bloss, W.J., Carpenter, L.J., Dorsey, J., Dunk, R., Flynn, M.J., Furneaux, K.L., Gallagher, M.W., Heard, D.E., Hollingsworth, A.M., Hornsby, K., Ingham, T., Jones, C.E., Jones, R.L., Kramer, L.J., Langridge, J.M., Leblanc, C., LeCrane, J.P., Lee, J.D., Leigh, R.J., Longley, I., Mahajan, A.S., Monks, P.S., Oetjen, H., Orr-Ewing, A.J., Plane, J.M.C., Potin, P., Shillings, A.J.L., Thomas, F., von Glasow, R., Wada, R., Whalley, L.K., Whitehead, J.D., 2010. Iodine-mediated coastal particle formation: an overview of the reactive halogens in the marine boundary layer (RHaMBLE) Roscoff coastal study. *Atmos. Chem. Phys.* 10, 2975–2999.
- Misra, A., Marshall, P., 1998. Computational investigations of iodine oxides. *J. Phys. Chem. A* 102, 9056.
- Möller, A., Lovric, M., Scholz, F., 1996. Evidence for the occasional appearance of molecular iodine in seawater. *Int. J. Environ. Anal. Chem.* 63, 99–106.
- Moore, R.M., Zafiriou, O.C., 1994. Photochemical production of methyl iodide in seawater. *J. Geophys. Res.* 99.
- Moore, R.M., Webb, M., Tokarczyk, R., Wever, R., 1996. Bromoperoxidase and iodoperoxidase enzymes and production of halogenated methanes in marine diatom cultures. *J. Geophys. Res.* 101, 20899–20908.
- Mössinger, J.C., Cox, R.A., 2001. Heterogeneous reaction of HOI with sodium halide salts. *J. Phys. Chem. A* 105 (21), 5165–5177.
- Mössinger, J.C., Shallcross, D.E., Cox, R.A., 1998. UV–VIS absorption cross-sections and atmospheric lifetimes of CH_2Br_2 , CH_2I_2 and CH_2BrI . *J. Chem. Soc. Far. 94*, 1391–1396.
- Muramatsu, Y., Yoshida, S., 1995. Volatilization of methyl-iodide from the soil-plant system. *Atmos. Environ.* 29 (1), 21–25.
- Nightingale, P.D., Malin, G., Liss, P.S., 1995. Production of chloroform and other low-molecular-weight halocarbons by some species of macroalgae. *Limnol. Oceanogr.* 40 (4), 680–689.
- Nightingale, P.D., Malin, G., Law, C.S., Watson, A.J., Liss, P.S., Liddicoat, M.I., Boutin, J., Upstill-Goddard, R.C., 2000. In situ evaluation of air-sea gas exchange parameterizations using novel conservative and volatile tracers. *Glob. Biogeochem. Cycles* 14 (1), 373–387.
- O'Dowd, C.D., De Leeuw, G., 2007. Marine aerosol production: a review of the current knowledge. *Philosophical Transactions of the Royal Society a – Mathematical Physical and Engineering Sciences* 365 (1856), 1753.
- O'Dowd, C.D., Geever, M., Hill, M.K., Jennings, S.G., Smith, M.H., 1998. New particle formation: nucleation rates and spatial scales in the clean marine coastal environment. *Geophys. Res. Lett.* 25, 1661.
- O'Dowd, C.D., McFiggans, G., Creasey, D.J., Pirjola, L., Hoell, C., Smith, M.H., Allan, B.J., Plane, J.M.C., Heard, D.E., Lee, J.D., Pilling, M.J., Kulmala, M., 1999. On the photochemical production of new particles in the coastal boundary layer. *Geophys. Res. Lett.* 26, 1707–1710.
- O'Dowd, C.D., Hameri, K., Makela, J.M., Pirjola, L., Kulmala, M., Jennings, S.G., Berresheim, H., Hansson, H.C., de Leeuw, G., Kunz, G.J., Allen, A.G., Hewitt, C.N., Jackson, A., Viisanen, Y., Hoffmann, T., 2002a. A dedicated study of new particle formation and fate in the coastal environment (PARFORCE): overview of objectives and achievements. *J. Geophys. Res. Atmos.* 107.
- O'Dowd, C.D., Jimenez, J.L., Bahreini, R., Flagan, R.C., Seinfeld, J.H., Hameri, K., Pirjola, L., Kulmala, M., Jennings, S.G., Hoffmann, T., 2002b. Marine aerosol formation from biogenic iodine emissions. *Nature* 417 (6889), 632–636.
- Palmer, C.J., Anders, T.L., Carpenter, L.J., Küpper, F.C., McFiggans, G., 2005. Iodine and halocarbon response of *Laminaria digitata* to oxidative stress and links to atmospheric new particle production. *Environ. Chem.* 2, 282–290.
- Pechtl, S., Lovejoy, E.R., Burkholder, J.B., von Glasow, R., 2006. Modeling the possible role of iodine oxides in atmospheric new particle formation. *Atmos. Chem. Phys.* 6 (2), 505–523.
- Peters, C., Pechtl, S., Stutz, J., Hebestreit, K., Hönninger, G., Heumann, K.G., Schwarz, A., Winterlik, J., Platt, U., 2005. Reactive and organic halogen species in three different European coastal environments. *Atmos. Chem. Phys.* 5, 3357–3375.
- Pundt, I., Pommereau, J.-P., Phillips, C., Lateltin, E., 1998. Upper limit of iodine oxide in the lower stratosphere. *J. Atmos. Chem.* 30 (1), 173–185.
- Rattigan, O.V., Shallcross, D.E., Cox, R.A., 1997. UV absorption cross-sections and atmospheric photolysis rates of CF_3I , CH_3I , $\text{C}_2\text{H}_5\text{I}$ and CH_2I_2 . *J. Chem. Soc. Faraday Trans. 93*, 2839–2846.
- Read, K.A., Mahajan, A.S., Carpenter, L.J., Evans, M.J., Faria, B.V.E., Heard, D.E., Hopkins, J.R., Lee, J.D., Moller, S., Lewis, A.C., Mendes, L., McQuaid, J.B., Oetjen, H., Saiz-Lopez, A., Pilling, M.J., Plane, J.M.C., 2008. Extensive halogen-mediated ozone destruction over the tropical Atlantic Ocean. *Nature* 453 (7199), 1232–1235.
- Redeker, K.R., Wang, N.Y., Low, J.C., McMillan, A., Tyler, S.C., Cicerone, R.J., 2000. Emissions of methyl halides and methane from rice paddies. *Science* 290 (5493), 966–969.
- Reithmeier, H., Lazarev, V., Kubo, F., Ruhm, W., Nolte, E., 2005. ^{129}I in precipitation using a new TOF system for AMS-measurements. *Nucl. Instrum. Methods Phys. Res. B* 239 (3), 273–280.
- Reithmeier, H., Lazarev, V., Rühm, W., Schwikowski, M., Gäggeler, H.W., Nolte, E., 2006. Estimate of European ^{129}I releases supported by ^{129}I analysis in an Alpine ice core. *Environ. Sci. Technol.* 40 (19), 5891–5896.
- Reithmeier, H., Lazarev, V., Ruhm, W., Nolte, E., 2010. Anthropogenic ^{129}I in the atmosphere: overview over major sources, transport processes and deposition pattern. *Sci. Total Environ.* 408, 5052–5064.

- Rhew, R.C., Miller, B.R., Weiss, R.F., 2000. Natural methyl bromide and methyl chloride emissions from coastal salt marshes. *Nature* 403 (6767), 292–295.
- Rhew, R.C., Miller, B.R., Vollmer, M.K., Weiss, R.F., 2001. Shrubland fluxes of methyl bromide and methyl chloride. *J. Geophys. Res.* Atmos. 106 (D18), 20875–20882.
- Richter, U., Wallace, D.W.R., 2004. Production of methyl iodide in the tropical Atlantic Ocean. *Geophys. Res. Lett.* 31 (23), Art. no. L23S03.
- Roehl, C.M., Burkholder, J.B., Moortgat, G.K., Ravishankara, A.R., Crutzen, P.J., 1997. Temperature dependence of UV absorption cross sections and atmospheric implications of several alkyl iodides. *J. Geophys. Res.* 102 (12), 819–829.
- Rucklidge, J., Kilius, L., Fuge, R., 1994. 129I in moss down-wind from the Sellafield nuclear reprocessing plant. *Nucl. Instrum. Methods Phys. Res. B* 92, 417–420.
- Saiz-Lopez, A., Plane, J.M.C., 2004. Novel iodine chemistry in the marine boundary layer. *Geophys. Res. Lett.* 31, L04112.
- Saiz-Lopez, A., Saunders, R.W., Joseph, D.M., Ashworth, S.H., Plane, J.M.C., 2004. Absolute absorption cross-section and photolysis rate of I-2. *Atmos. Chem. Phys.* 4, 1443–1450.
- Saiz-Lopez, A., Plane, J.M.C., McFiggans, G., Williams, P.I., Ball, S.M., Bitter, M., Jones, R.L., Hongwei, C., Hoffmann, T., 2006. Modelling molecular iodine emissions in a coastal marine environment: the link to new particle formation. *Atmos. Chem. Phys.* 6, 883–895. <http://dx.doi.org/10.5194/acp-6-883-2006>.
- Saiz-Lopez, A., Mahajan, A.S., Salmon, R.A., Bauguette, S.J.B., Jones, A.E., Roscoe, H.K., Plane, J.M.C., 2007. Boundary layer halogens in coastal Antarctica. *Science* 317 (5836), 348–351.
- Saiz-Lopez, A., Plane, J.M.C., Baker, A., Carpenter, L.J., von Glasow, R., Gomez Martin, J.C., McFiggans, G., Saunders, R.W., 2012. Atmospheric chemistry of iodine. *Chem. Rev.* 112 (3), 1773–1804.
- Sander, S.P., 1986. Kinetics and mechanism of the IO + IO reaction. *J. Phys. Chem.* 90 (10), 2194–2199.
- Sander, S.P., Orkin, V.L., Kurylo, M.J., Golden, D.M., Huie, R.E., Kolb, C.E., Finlayson-Pitts, B.J., Molina, M.J., Friedl, R.R., Ravishankara, A.R., Moortgat, G.K., Keller-Rudek, H., Wine, P.H., 2006. Chemical Kinetics and Photochemical Data for Use in Atmospheric Studies. Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA.
- Saunders, R.W., Plane, J.M.C., 2005. Formation pathways and composition of iodine oxide ultra-fine particles. *Environ. Chem.* 2 (4), 299–303.
- Saunders, R.W., Plane, J.M.C., 2006. Fractal growth modelling of I₂O₅ nanoparticles. *J. Aerosol Sci.* 37 (12), 1737.
- Saunders, R.W., Mahajan, A.S., Gómez Martín, J.C., Kumar, R., Plane, J.M.C., 2010. Studies of the formation and growth of aerosol from molecular iodine precursor. *Zeitschr. f. Phys. Chem. Munich* 224 (7–8), 1095–1117.
- Schönhardt, A., 2009. DOAS Measurements of Iodine Monoxide from Satellite (Ph.D. thesis). University of Bremen, Bremen, Germany.
- Schönhardt, A., Richter, A., Wittrock, F., Kirk, H., Oetjen, H., Roscoe, H.K., Burrows, J., 2008. Observations of iodine monoxide columns from satellite. *Atmos. Chem. Phys.* 8 (3), 637–653.
- Shaw, M.D., Carpenter, L.J., Baeza-Romero, M.T., Jackson, A., 2011. Thermal evolution of diffusive transport of atmospheric halocarbons through artificial sea-ice. *Atmos. Environ.* 45 (35), 6393–6402.
- Sive, B.C., Varner, R.K., Mao, H., Blake, D.R., Wingenter, O.W., Talbot, R., 2007. A large terrestrial source of methyl iodide. *Geophys. Res. Lett.* 34, L17808.
- Smythe-Wright, D., Boswell, S.M., Breithaupt, P., Davidson, R.D., Dimmer, C.H., Eiras Diaz, L.B., 2006. Methyl iodide production in the ocean: implications for climate change. *Glob. Biogeochem. Cycles* 20.
- Solomon, S., García, R.R., Ravishankara, A.R., 1994. On the role of iodine in ozone depletion. *J. Geophys. Res.* 99, 20491–20499.
- Sturges, W.T., Barrie, L.A., 1988. Chlorine, bromine and iodine in Arctic aerosols. *Atmos. Environ.* 22 (6), 1179–1194.
- Stutz, J., Hebestreit, K., Alicke, B., Platt, U., 1999. Chemistry of halogen oxides in the troposphere: comparison of model calculations with recent field data. *J. Atmos. Chem.* 34, 65–85.
- Stutz, J., Pikel'naya, O., Hurlock, S.C., Trick, S., Pechtl, S., von Glasow, R., 2007. Daytime IO in the Gulf of Maine. *Geophys. Res. Lett.* 34, L22816. <http://dx.doi.org/10.1029/2007GL031332>.
- Sumerling, T.J., 1984. The use of mosses as indicators of airborne radionuclides near a major nuclear installation. *Sci. Total Environ.* 35, 251–265.
- The Lancet., July 12, 2008. Iodine deficiency—way to go yet. *Lancet* 372 (9633), 88. [http://dx.doi.org/10.1016/S0140-6736\(08\)61009-0](http://dx.doi.org/10.1016/S0140-6736(08)61009-0). PMID 18620930.
- Theiler, R., Cook, J.C., Hager, L.P., Studa, J.F., 1978. Halohydrocarbon synthesis by bromoperoxidase. *Science* 202, 1094–1096.
- Thomas, D.N., Diekmann, G.S. (Eds.), 2003. *Sea Ice: An Introduction to its Physics, Chemistry, Biology and Geology*. Wiley-Blackwell.
- Truesdale, V.W., 2008. The biogeochemical effect of seaweeds upon close-to-natural concentrations of dissolved iodate and iodide in seawater – preliminary study with *Laminaria digitata* and *Fucus serratus*. *Estuarine Coastal Shelf Sci.* 78, 155–165.
- Tucceri, M.E., Hölscher, D., Rodriguez, A., Dillon, T.J., Crowley, J.N., 2006. Absorption cross section and photolysis of OIO. *Phys. Chem. Chem. Phys.* 8, 834–846.
- Varner, R.K., Crill, P.M., Talbot, R.W., 1999. Wetlands: a potentially significant source of atmospheric methyl bromide and methyl chloride. *Geophys. Res. Lett.* 26 (16), 2433–2435.
- Vogt, R., Crutzen, P.J., Sander, R., 1996. A mechanism for halogen release from sea-salt aerosol in the remote marine boundary layer. *Nature* 383, 327–330.
- Vogt, R., Sander, R., Von Glasow, R., Crutzen, P.J., 1999. Iodine chemistry and its role in halogen activation and ozone loss in the marine boundary layer: a model study. *J. Atmos. Chem.* 32, 375–395.
- Volkamer, R., Coburn, S.C., Dix, B.K., Sinreich, R., 2010. Sources of reactive halogen species and oxygenated VOC over biologically active upwelling reactions of the tropical Pacific Ocean. *Geochim. Cosmochim. Acta* 74, A1085.
- von Glasow, R., Sander, R., Bott, A., Crutzen, P.J., 2002. Modeling halogen chemistry in the marine boundary layer 1. Cloud-free MBL. *J. Geophys. Res.* 107 (D17), 4341–4356.
- Wada, R., Beames, J., Orr-Ewing, A., 2007. Measurement of IO radical concentrations in the marine boundary layer using a cavity ring-down spectrometer. *J. Atmos. Chem.* 58 (1), 69–87.
- Wagner, M.J.M., Dittich-Hannen, B., Synal, H.A., Suter, M., Schotterer, U., 1996. Increase of 129I in the environment. *Nucl. Instrum. Methods Phys. Res. B* 113, 490–494.
- Wennberg, P.O., Brault, J.W., Hanisco, T.F., Salawitch, R.J., Mount, G.H., 1997. The atmospheric column abundance of IO: implications for stratospheric ozone. *J. Geophys. Res.* 102.
- Whalley, L., Furneaux, K., Gravestock, T., Atkinson, H., Bale, C., Ingham, T., Bloss, W., Heard, D., 2007. Detection of iodine monoxide radicals in the marine boundary layer using laser induced fluorescence spectroscopy. *J. Atmos. Chem.* 58, 19–39.
- Williams, D., 2008. Twenty years' experience with post-Chernobyl thyroid cancer. *Best Pract. Res. Clin. Endocrinol. Metab.* 22 (6), 1061–1073. <http://dx.doi.org/10.1016/j.beem.2008.09.020>.
- Williams, J., Gros, V., Atlas, E., Maciejczyk, K., Batsaikhan, A., Scholer, H.F., Forster, C., Quack, B., Yassaa, N., Sander, R., Van Dingenen, R., 2007. Possible evidence for a connection between methyl iodide emissions and Saharan dust. *J. Geophys. Res.* 112, D07302.
- Wittrock, F., Müller, R., Richter, A., Bovensmann, H., Burrows, J.P., 2000. Measurements of iodine monoxide (IO) above Spitsbergen. *Geophys. Res. Lett.* 27.
- Wong, G.T.F., 1991. The marine geochemistry of iodine. *Rev. Aquatic Sci.* 4 (1), 45–73.
- Wuossmaa, A.M., Hager, L.P., 1990. Methyl chloride transferase: a carbocation route for biosynthesis of halometabolites. *Science* 249, 160–162.
- Yokouchi, Y., Ikeda, M., Inuzuka, Y., Yukawa, T., 2002. Strong emission of methyl chloride from tropical plants. *Nature* 416 (6877), 163–165.
- Zafirou, O.C., 1975. Reaction of methyl halides with seawater and marine aerosols. *J. Mar. Res.* 33 (1), 75–81.
- Zetzsch, C., Pfahler, G., Behnke, W., 1988. Heterogeneous formation of chlorine atoms from NaCl aerosols in smog chamber. *J. Aerosol Sci.* 19, 1203–1206.
- Zika, R.G., Gidel, L.T., Davis, D.D., 1984. A comparison of photolysis and substitution decomposition rates of methyl iodide in the ocean. *Geophys. Res. Lett.* 11 (4), 353–356.
- Zingler, J., Platt, U., 2005. Iodine oxide in the Dead Sea Valley: evidence for inorganic sources of boundary layer IO. *J. Geophys. Res.* 110.

Further Reading

- Küpper, F.C., Feiters, M.C., Olofsson, B., Kaiho, T., Yanagida, S., Zimmermann, M.B., Carpenter, L.J., Luther, G., Lu, Z., Jonsson, M., Kloo, L., 2009. Purple fumes: the importance of iodine. *Sci. School* 27, 45–53. www.scienceinschool.org/2013/issue27/iodine.
- Küpper, F.C., Carpenter, L.J., Feiters, M.C., Kaiho, T., Kloo, L., Lu, Z., Luther, G., Olofsson, B., Yanagida, S., 2011. Commemorating two centuries of iodine research: an interdisciplinary overview of current research. *Angew. Chem. Int. Ed.* 50 (49), 11598–11620. <http://dx.doi.org/10.1002/anie.201100028>.
- Saiz-Lopez, A., von Glasow, R., 2012. Reactive halogen chemistry in the troposphere. *Chem. Soc. Rev.* 41 (19), 6448–6472.

BOUNDARY LAYER (ATMOSPHERIC) AND AIR POLLUTION

Contents

Overview

Air Pollution Meteorology

Coherent Structures

Complex Terrain

Convective Boundary Layer

Microclimate

Modeling and Parameterization

Observational Techniques *In Situ*

Observational Techniques: Remote

Ocean Mixed Layer

Stably Stratified Boundary Layer

Surface Layer

Urban Heat Islands

Diurnal Cycle

Overview

PJ Mason and DJ Thomson, Met Office, Bracknell, UK

Crown Copyright © 2003 Published by Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 1, pp 221–228, © 2003, Elsevier Ltd.

Introduction

Away from the surface the atmosphere is mainly statically stable with little small-scale turbulent mixing occurring outside frontal regions, areas with moist convective updrafts, and regions with breaking gravity waves generated by hills and mountains. In contrast, adjacent to the surface the flow is nearly always turbulent, with the turbulence generated by the action of wind shear and/or buoyant convection. This layer adjacent to the surface in which vertical mixing is especially important is termed the boundary layer. In fact, turbulent boundary layers involving shear or buoyancy effects are a key feature of all bounded fluid flows at high Reynolds number. However, the atmospheric boundary layer involves a variety of features and processes in addition to shear and buoyancy. For example, there is often a distinct interface at the height where the mixing reaches the base of the stable free atmosphere above. Other dynamical and thermodynamical processes that affect the atmospheric boundary layer are Coriolis forces produced by the planetary rotation, and factors such as the formation of clouds and radiative heat transfer.

The ocean is also mainly statically stable and has boundary layers at its top and bottom. The boundary layer at the bottom of the ocean, which is generally much thinner than the atmospheric boundary layer, is called the benthic boundary layer. Here fluxes of heat and buoyancy influences are weak and the boundary layer is usually shear-driven. The surface of the oceans, in contrast, can have substantial fluxes of buoyancy as well as significant shear arising from the influence of the wind at the surface.

The term boundary layer is also sometimes used in association with flows and currents that are concentrated against the sides of basins or orography. These features, such as the Gulf Stream in the North Atlantic Ocean, arise from the flow dynamics without mixing playing a critical role. They are not discussed further here.

Role in the Overall Atmosphere

The boundary layer has an important influence on the behavior of the atmosphere as a whole, and activities involving the representation of the atmosphere such as climate modeling and numerical weather prediction cannot

succeed without the boundary layer being represented in some detail. The main influences on the atmosphere as a whole are as follow:

- Over level terrain the boundary layer determines the drag between the atmosphere and the surface and this drag is the main mechanism by which the energy in the large-scale motion is dissipated. With small-scale orography the drag also occurs through pressure forces whose magnitude is influenced by the boundary layer turbulence.
- It provides a buffer between the surface and the atmosphere, thus influencing the transfer of heat and moisture between the surface and the atmosphere and, in particular, the way surface solar heating is partitioned into sensible and latent heat fluxes.
- It is critical in determining the properties of air entering the base of clouds that form the roots for moist convection extending into the atmosphere above the boundary layer.
- It plays a central role in determining the occurrence of low-level cloud within the boundary layer and the consequential effects on radiation budgets.
- It tends to retain aerosols and pollutants from the surface, with the transfer of such polluted air to the free troposphere being limited mainly to moist convection and frontal motions which, through washout, leave the main atmospheric air free of such material.

Role in the Local Atmospheric Environment

The boundary layer is of particular significance to human activities and natural processes occurring on the Earth's surface. Here prediction and understanding of the local environment requires an understanding of the boundary layer. In particular, the boundary layer is important for predicting a range of parameters such as

- the near surface wind and turbulence;
- daily maximum and minimum temperatures;
- visibility and fog;
- the dispersion of pollutants and other material.

Role in the Oceans

As in the atmosphere, boundary layers in the oceans play an important part in the overall and local ocean circulations:

- In the deep ocean, the bottom benthic boundary layer has only a weak dissipating influence on the ocean circulation, although it plays a more important role on the continental shelf and in coastal regions.
- The surface boundary layer plays a key role in influencing the rate of exchange of heat and momentum between the atmosphere and the ocean and is consequently critical to the ocean circulation.

Theoretical Framework

In turbulent flow, the fluxes of flow variables are caused by the differing properties of air 'parcels' moving in different

directions relative to the mean flow (the mean flow usually being defined, from a practical perspective, as an average over a period of order an hour). For example, if the rising air parcels are warmer than those descending, then there will be a vertical flux of heat H . This heat flux H is equal to $\rho c_p \langle w'T' \rangle$, where ρ and c_p are the density and specific heat capacity of air and $\langle w'T' \rangle$ is the covariance of w' and T' , the turbulent fluctuations in vertical velocity and in temperature about their means. Similarly, the covariance between the horizontal and vertical velocities gives rise to a downward vertical flux of horizontal momentum τ which is given by $-\rho \langle u'w' \rangle$, where u' is the fluctuation in the horizontal component of velocity. Such a momentum flux is often referred to as a turbulent stress.

These turbulent fluxes have a major effect on the flow. For example, if we consider the average flow in a horizontally homogeneous boundary layer over an area of the earth surface, the governing equation of the horizontal momentum balance is given by eqn [1].

$$\frac{d\mathbf{U}}{dt} = -\frac{1}{\rho} \text{grad } p - f\mathbf{k} \times \mathbf{U} + \frac{1}{\rho} \frac{d\tau}{dz} \quad [1]$$

Here p is the mean pressure, f is the Coriolis parameter, $\mathbf{k}$ is a unit vertical vector, $\mathbf{U}$ is the (horizontal) mean velocity vector, and z the vertical coordinate, with the adjective 'mean' being used to indicate quantities where the turbulent fluctuations have been averaged out. Above the boundary layer where τ is zero, the flow will tend to adjust to a steady state with a balance between the pressure and Coriolis forces, leading to a geostrophic flow at right angles to the pressure gradient. Within the boundary layer, however, there is a rotation of the flow away from the geostrophic direction, with the wind having a component directed down the pressure gradient. Similar equations can be written for other variables such as temperature or humidity. These equations will contain a balance between the rate of change and the flux gradient, together with any source terms such as – in the case of the temperature equation – the radiative transfer divergence.

Field experiments have yielded many useful data on flow covariances such as $\langle u'w' \rangle$. In a theoretical description the key issue is to estimate the covariance from the flow properties. This is an impossible problem to resolve fully, because the flow eddies determining the flux evolve through complex, nonlinear and turbulent interactions, making it intractable to obtain exact solutions for the fluxes in terms of the mean flow variables. In many cases, however, progress is possible through simple closure approximations (which relate the turbulence statistics to the local mean flow) and/or through consideration of bulk models.

The simplest closure method is the eddy-viscosity approach in which the flux is assumed proportional to local mean gradients through a turbulent 'eddy viscosity' or 'eddy diffusivity', in the same way as fluxes due to molecular motions are related to gradients via molecular viscosities and diffusivities. In the case of the velocity covariance in the momentum balance equation, this leads to the approximation of eqn [2].

$$-\langle u'w' \rangle = K \frac{dU}{dz} \quad [2]$$

K is the eddy viscosity, which varies in space and time. K itself is usually determined by a combination of dimensional

considerations and empirical measurements. It can usefully be thought of as a turbulent velocity scale multiplied by a length scale, by analogy with the way molecular mixing is related to the size of molecular velocities and the length of the mean free path. Close to the surface, dimensional considerations imply a dependency on the distance from the surface (physically this is because eddy sizes are restricted by the presence of the ground and so tend to be smaller close to the ground) while toward the middle of the boundary layer some length scale related to the boundary layer depth will prevail. A separate estimate of velocity scale is needed. This may be obtained from local or bulk considerations. A local determination can be made by considering a mixing length model (see below) or by considering the turbulence energy equation with consideration of both shear and buoyancy effects.

Although eddy viscosity approaches and other closure models are only an approximate representation of the turbulence properties, under appropriate conditions model-independent results can be derived through dimensional considerations combined with an empirical determination of dimensionless constants or functions. Such similarity descriptions have extensive application in a region near the ground where the flow depends on only a few variables. This region, called the surface layer, occupies the part of the flow where the height z is much greater than the roughness elements but much less than the depth of the boundary layer. The structure of the flow in this region is thought to be characterized by a few parameters, namely, the friction velocity u_* (defined as $\sqrt{(\tau/\rho)}$ where τ is the magnitude of the surface stress), the surface flux of sensible heat H , and the height z above the surface. In neutral stability conditions, with $H = 0$, dimensional analysis then leads to eqn [3].

$$\frac{dU}{dz} = \frac{u_*}{kz} \quad [3]$$

k is the von Karman constant (which has an empirically determined value of about 0.4) and U is the magnitude of the mean wind. Integrating this equation leads to the log law (eqn [4]).

$$U = \frac{u_*}{k} \log \frac{z}{z_0} \quad [4]$$

Here z_0 , the constant of integration, is related to the height of the roughness elements and is called the roughness length. It is typically about one-tenth of the actual height of the roughness elements.

The eddy viscosity approach can provide an exact match to these similarity relations if appropriate empirical choices for the coefficients and the various scales are used. As a simple illustration, consider the situation in the surface layer for the case of neutral stability. Here we can write eqn [5], where l_0 and u_0 are the length and velocity scale determining the eddy viscosity K .

$$\langle u'w' \rangle = K \frac{dU}{dz} = l_0 u_0 \frac{dU}{dz} \quad [5]$$

To match the similarity relation [3], we can choose $u_0 = u_*$ with l_0 equal to kz . If we also note that $\langle u'w' \rangle$ varies only slowly within the surface layer (the surface layer is often as a result referred to as the constant-stress layer) and that it can be approximated by the surface stress u_*^2 , then we again obtain the

log law. The choice of u_* as the correct velocity scale can be difficult to understand (both here and in the pure similarity approach above) and one way to motivate this is to consider a mixing length model. We assume air is mixed in the vertical over a length scale l_0 (the mixing length) and assume that the air starts off with the local mean horizontal velocity and retains this velocity as it travels. At the end of its journey it has a horizontal velocity that differs from the local mean by $u' \sim l_0(dU/dz)$. Taking this as our velocity scale u_0 we obtain eqn [6].

$$u_*^2 = \langle u'w' \rangle = l_0 \left(l_0 \left| \frac{dU}{dz} \right| \right) \frac{dU}{dz} = (kz)^2 \left| \frac{dU}{dz} \right| \frac{dU}{dz} \quad [6]$$

These surface relations can be refined to incorporate buoyancy effects. From dimensional considerations, the velocity scale w_f at a height scale z due to a heat flux H is given by eqn [7].

$$w_f^3 = \frac{kgHz}{\rho c_p T} \quad [7]$$

This order of magnitude estimate of the local buoyancy-driven velocity scale could have been more physically derived by considering the energy equation or buoyancy accelerations subject to the key recognition of distance from the surface as the relevant length scale. From similar dynamical considerations, we can derive an important length scale, the Monin–Obukhov length L , which is the height scale at which the shear and buoyancy velocity scales are equal. It is defined by eqn [8].

$$L = \frac{-zu_*^3}{w_f^2} \quad [8]$$

This is a key parameter in the surface layer, and z/L is a measure of the relative role of shear and buoyancy in the production of turbulence. The flow in the surface layer is then determined by eqn [9].

$$\frac{dU}{dz} = \frac{u_*}{kz} \varphi_M \left(\frac{z}{L} \right) \quad [9]$$

Here φ_M is an empirical function which has been estimated experimentally. Similar relations hold for other variables, and by integration stability-dependent bulk formulas can also be obtained. φ_M is greater than unity in stably stratified conditions ($L > 0$) where turbulence transfer is weaker (i.e. for fixed dU/dz the momentum flux is reduced) and less than unity in convective conditions ($L < 0$). At heights much less than $|L|$, buoyancy effects are unimportant and $\varphi_M \sim 1$, with U approaching the log law appropriate for neutral conditions.

In a model of the boundary layer, the net impacts can be described by bulk drag and transfer coefficients that relate the surface stress to the wind U at some finite height via eqn [10].

$$\tau = C_D |U| U \quad [10]$$

It is trivial to relate this bulk formula to the similarity result – for example, in neutral conditions C_D is given by $C_D = (k/\log(z/z_0))^2$.

Although the classical surface layer theory outlined above has proved to be one of the cornerstones of boundary layer theory, it is important to realize that it may not be exactly correct. For example, the presence of large eddies that fill the boundary layer will cause horizontal fluctuations in wind

near the ground whose properties depend on the boundary layer depth. This is particularly true in convective conditions, but may also play a role in other types of boundary layer.

Types of Boundary Layer

In both the atmosphere and the oceans, boundary layers occur in forms that can be complex, involving a diverse mixture of processes and space and time dependencies. In spite of this general complexity, boundary layer properties can often be understood in terms of the properties of a number of idealized cases, and these cases provide an overview of the range of possible behavior. The simplest cases are those with horizontally homogeneous and steady conditions, and these are best classified by stability.

In the neutral boundary layer, where the surface heat flux is negligible, the wind increases steadily with height (although at a decreasing rate) and the wind direction varies throughout the boundary layer. Such neutral conditions, where shear flow turbulence generation dominates, occur as often because the wind speed is large because the heat flux is actually close to zero. The relevant criterion is that the boundary layer depth is smaller than the magnitude of the Monin–Obukhov length L . Turbulence levels scale on the friction velocity u_* and decrease steadily with height. With a geostrophic wind speed of 10 m s^{-1} , typical values of u_* and of fluctuations in wind speed are about 0.3 m s^{-1} and 1 m s^{-1} respectively, depending of course on the surface roughness. If there are no thermal effects at all, then the boundary layer depth is of order $0.3u_*/f$. However, the term neutral boundary layer usually refers only to the absence of a heat flux at the surface, and stratification and subsidence above the boundary layer almost invariably play an important role in restricting the actual boundary layer depth.

In convective boundary layers the surface heating is sufficient to make the Monin–Obukhov length very much less than the boundary layer depth. The vertical mixing is stronger and this tends to produce nearly uniform profiles of velocity and temperature, with variations largely confined to near the surface and the boundary layer top. With heat being continually added at the surface, a completely steady situation is impossible and the boundary layer usually grows slowly as the stable stratification above the boundary layer is eroded. The nature of this erosion often acts to sharpen gradients at the boundary layer top and frequently there is a strong temperature inversion extending over only a few tens of meters. Vertical turbulent velocities tend to peak some way above the ground (at about a height of $d/3$ where d is the boundary layer depth), with lower values near the ground and boundary layer top. Horizontal turbulent velocities, however, are more uniform as these are not blocked by the presence of the ground. The typical velocity scale is linked to the boundary layer depth and is of order the convective velocity scale w^* which is given by eqn [11].

$$w_*^3 = \frac{kgHd}{\rho c_p T} \quad [11]$$

A typical value for the daytime convective boundary layer in mid-latitudes is about 1 or 2 m s^{-1} . The largest eddies extend

throughout the depth of the boundary layer. In summer in mid-latitudes with clear skies, convective boundary layers will reach a depth of order 2 km by late afternoon. Pollutants emitted into convective boundary layers are dispersed rapidly. This leads to small ground level concentrations from near-surface emissions. However, convective conditions can lead to high near-surface concentrations from releases from elevated stacks because the rapid mixing can bring material down to the ground quickly before it has been much diluted. In these conditions plumes are often seen to ‘loop’ up and down as they are distorted by boundary layer eddies that are much larger than the plume.

In stable boundary layers that generally occur at night or over other cool surfaces, mixing is greatly reduced and temperature and wind vary across the whole depth of the boundary layer at a more constant rate than in neutral conditions. Three-dimensional turbulent eddies tend to be smaller than in the neutral or convective cases and turbulence levels tend to be low (although they scale with u_* as in the neutral case, u_* itself tending to be smaller in stable cases). In addition to the three-dimensional turbulence, slow lateral meandering motions are often present. Stable boundary layers tend to be much shallower than convective or neutral boundary layers, with a typical depth of order 100 – 200 m . Often, however, wind shear is present above the true stable boundary layer due to the legacy from the daytime boundary layer and/or inertial oscillations. In very stable cases it seems that the turbulence can be completely suppressed at times and becomes intermittent. This regime is very hard to understand and predict because small influences that have a negligible effect on neutral and convective boundary layers can become important – examples are slight slopes, variations in the thermal properties of the ground, and spatial variations in cloud cover. Pollutants emitted into stable boundary layers are dispersed slowly. This can lead to high ground level concentrations for near-surface sources. However, ground level concentrations from elevated sources can be small because the plume is mixed down to the ground only very slowly, and may be released at such a height or with sufficient buoyancy that it is carried completely above the boundary layer. Typical profiles of wind and temperature for neutral, convective, and stable boundary layers are shown in Figure 1.

In the real atmosphere, conditions always vary with time, but usually the time variation is slow enough for the main characteristics of the ideal boundary layers to prevail. A primary cause of variation over land is the diurnal cycle. The boundary layer starts at night with a shallow, stable layer. As the sun rises and heats the ground, convective turbulence is generated and a convective boundary layer is formed that gradually erodes the stable layer above the boundary layer to reach its maximum depth by late afternoon. Then the sun sets and the ground starts to cool, with thermal effects now tending to suppress the turbulence. As a result, the turbulence decays rapidly except in a shallow layer near the ground, which forms a new stable boundary layer. Although this is a useful idealized description, it should be noted that there are many situations in which it does not apply. For example, in strong winds, the ‘mechanical’ effects may dominate the thermal effects in the boundary layer energy balance, leading to an effectively neutral boundary layer. Also, at high latitudes the sun may not be strong enough

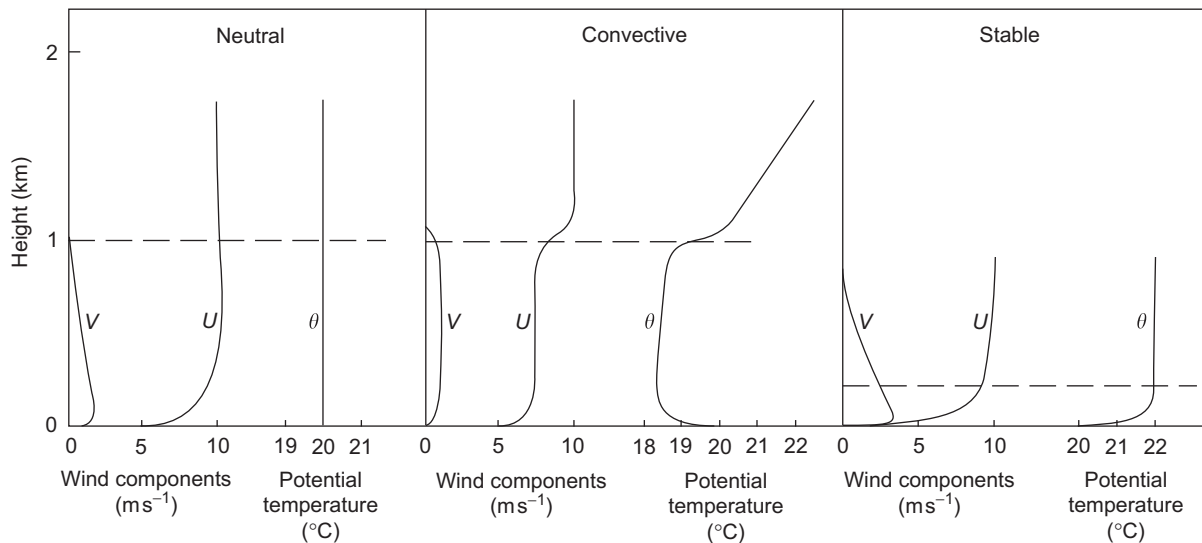


Figure 1 Illustration of typical profiles of mean wind (U , V) and potential temperature (θ) in neutral, convective, and stable atmospheric boundary layers, for a geostrophic wind of 10 m s^{-1} and a surface temperature of 20°C . The dashed lines indicate the boundary layer top, which is generally well defined only in the convective case. The profiles are based on the results of large-eddy simulations for idealized conditions but, for reasons of illustration, have been adjusted to indicate the typical relative depths of the three types of boundary layer. Real profiles will show significant differences due to a range of influences and to the presence of turbulent fluctuations that will not be completely removed even if the profiles are averaged over a period of an hour.

to make H positive. Over the sea the diurnal cycle is almost absent owing to the greater heat capacity of the sea and hence the smaller changes in surface temperature.

In this short introduction it is especially difficult to do justice to the full range of complexities that can have some influence upon the boundary layer, but we can note some key ones. Clouds and fog within the boundary layer do not release enough rain or drizzle to be significant net sources of latent heating but rather have influence through acting to redistribute heat and moisture within their circulations and through the effect of radiative cooling on the cloud tops. Such cooling can be the main source of turbulence energy production in situations with fog and stratocumulus-capped boundary layers. Entrainment is a key process in the heat and moisture budgets of such boundary layers, and the development of cloudy boundary layers is sensitive to this and the surface fluxes. The entrainment can be driven by the action of the boundary layer eddies and is also influenced by wind shear and the potential for dry entrained air to give evaporative cooling of the cloud.

The matching of the description of the turbulent flow to the surface properties is a key issue because the surface fluxes are critical to the overall boundary layer properties. At the surface a great range of complex issues prevail as the flow interacts with the obstacles and vegetation that comprise the surface. The availability of empirical data is usually essential to quantitative prediction. Over homogeneous surfaces with level terrain we have noted that the momentum transfer properties of the surface can be represented by the roughness length z_0 . Corresponding roughness lengths can also be used for heat and moisture and they are generally much smaller because scalar fluxes do not involve pressure forces. They are also less accurate in use owing to variations with plant properties as they interact with varying soil moisture, humidity, temperature, and

radiative fluxes. It is common in models to refine the treatment of vegetated surfaces with plant canopy models that seek to represent key processes such as those of the plant stomata, and even flow within the plant canopy.

The task of representing the surface is made more difficult by the presence of heterogeneity in the surface. The boundary layer tends to respond to an area average of the surface properties. This is somewhat the inverse of the way in which a pollutant from a point source would occupy a large area at heights some distance above the surface. The methods of forming area average properties involve consideration of boundary layer dynamics. Orography on a scale of several kilometers or less is another important surface feature. The consequent pressure forces on the flow give a large surface drag with high levels of turbulence and mixing.

Over the sea it is usual to use either roughness lengths or the equivalent bulk transfer relations. The values are not constant but are mainly functions of wind speed. The momentum transfer occurs mainly through pressure forces on the waves and varies with wind speed as a result of its influence on the wave height. The values of roughness length are generally smaller than over land surfaces. The scalar transfer coefficients increase only weakly with wind speed and spray generation may be one cause of this. For waves in equilibrium with the wind there are reasonable empirical descriptions of the transfer coefficients. More refined approaches seek to use wave models to allow for the waves being out of equilibrium with the wind. There also remains uncertainty over transfer in very strong wind conditions when severe spray makes measurements impossible. These surface fluxes match into the oceanic surface boundary layer, which also receives the influence of any radiative transfer and input of fresh water from precipitation. The oceanic surface boundary layer is subject to dynamics

comparable with those of the basic neutral, stable, and convective atmospheric boundary layers. However, the time scales for the variations in the mean structure of the temperature and salinity profiles differ and derive from a combination of the seasonal heating and cooling modulated by the daily weather cycles. The main areas of the ocean have a stable thermocline somewhat analogous to the stable cap of most atmospheric boundary layers. Deep penetrating convective motion tends to occur in more limited areas where strong cooling occurs. In the atmosphere, vertical movement of air through a significant fraction of the depth of the troposphere is similarly restricted in its area, but the areas of occurrence are linked to moist convection in deep cumulus clouds or frontal zones and not directly to persistent heating.

Practical Boundary Layer Models PBL

Models of a variety of types are used to represent and predict boundary layers. Even the simplest work well in some cases, while even the most complex is not able to give reliable results in all circumstances. Fortunately we have a fair understanding of the reasons for success and failure and can at least anticipate these.

The simplest of these models are 'bulk' models that estimate, for example, the drag and the boundary layer depth directly from the external parameters such as the wind above the boundary layer and the roughness length. More sophisticated models attempt to estimate the various fluxes that occur in the equations for the mean boundary layer properties. This is done either directly and locally within the flow in terms of the mean quantities (mixing length models are the main example here) or by deriving equations for the fluxes and then attempting to estimate the unclosed terms in those equations. The full equations for the fluxes can be simply derived from the equations of motion. Just as in the averaged equations of motion, where a closure model is required for the fluxes, so in these new equations a closure is needed for the higher-order terms that appear. The terms requiring closure involve triple moments and correlations of velocity and pressure. In a so-called second-order closure, a closure assumption is made for these terms. Generally this closure is, as in mixing length models, in terms of local flow properties. Such models usually only address the very difficult issue of nonlocal influences through a more questionable derivation of the local turbulence length scale and through diffusion of the second-order quantities. Typically, in situations where mixinglength models are reasonable, higher-order closure models will do better, and are quite successful in capturing the extra flow details. Neither model can usually deal at all well with the details of flows such as convective ones with strong nonlocal influences. Higher-order closures such as third-order ones have shown success in some cases but they seem unreliable when applied over a wide range of flows.

More sophisticated again are large-eddy simulation (LES) models. These models attempt to simulate the detailed evolution of the flow eddies within the boundary layer, although of course, because the smallest structures have a size of order 10^{-3} m, it is impossible to describe the entire flow in detail. Instead, the largest eddies are simulated, with the effects of the smallest scales parametrized (e.g., with a mixing length model).

The hope and evidence are that results will be insensitive to the treatment of these small scales and so results will not be compromised by errors in the parametrization. In using the LES technique it is not expected that the instantaneous details of the flow will correspond to a particular real case. Instead, the model results are used to derive mean quantities which, by averaging out the unpredictable turbulent variability, should be comparable with reality. In this sense the model is used more in the way climate modelers use large-scale meteorological models than in the way weather forecasters use them. The results from LES show great promise for the future and with adequate numerical resolution they offer a basis both to make direct predictions and to provide flow details with which to develop other closure methods. It can be argued that in the interior of the boundary layer little more is needed from the subgrid parametrization than to allow dissipation to occur explicitly at the smallest resolved scales. The only particular difficulty with LES is that it depends on achieving a correct parameterization near the surface. At the surface the flow eddies become too small for the LES to describe them explicitly. The LES solutions depend upon the closure in this region and in the transition from this region to the flow interior where the resolved eddies dominate the flow.

Models for predicting the dispersion of material in the boundary layer also have a similar range of degrees of sophistication. The simplest models estimate the width of any plume spread directly from a few parameters, e.g., wind speed, time of day and year (to estimate the solar elevation), and cloud cover (which affects the surface radiation budget). Some models involve mixing-length type assumptions or higher-order closure models. For a dispersing plume, the horizontal variations are critical as well as the vertical ones and so, unlike boundary layer models that are often one-dimensional, a fully three-dimensional description is required. A more sophisticated approach is to simulate the motion of many elements of the pollutant in a 'stochastic Lagrangian particle' model. Here the statistical properties of the flow (e.g. mean flow and velocity variances and covariances) are assumed known and a stochastic model for the random dispersion of pollutant 'particles' is constructed to be consistent with these flow statistics. Finally, the LES technique described above can be applied to calculate dispersion as well as the mean flow and turbulence, either by tracking particles within the LES flow or by solving an equation for the evolution of the concentration field.

Summary and Challenges

Our understanding of the boundary layer has progressed greatly over recent decades. Much still remains to be done, however, and progress will depend critically on the development of theory, numerical computations, and the continuing refinement of empirical data. As theory and numerical computations gain in potential it will remain essential to ensure that the underpinning empirical factors that are so critical in turbulence continue to receive good attention.

See also: Boundary Layer (Atmospheric) and Air Pollution: Modeling and Parameterization; Ocean Mixed Layer; Stably

Stratified Boundary Layer; Surface Layer. **Numerical Models:** Parameterization of Physical Processes: Turbulence and Mixing.

Further Reading

- Arya, S.P.S., 1988. Introduction to Micrometeorology. Academic Press, San Diego, CA.
- Garratt, J.R., 1992. The Atmospheric Boundary Layer. Cambridge University Press, Cambridge.
- Kaimal, J.C., Finnigan, J.J., 1994. Atmospheric Boundary Layer Flows – Their Structure and Measurement. Oxford University Press, New York.
- Kraus EB (Ed.), 1977. Modelling and Prediction of the Upper Layers of the Ocean. Pergamon Press, New York.
- Lumley, J.L., Panofsky, H.A., 1964. The Structure of Atmospheric Turbulence. Wiley, New York.
- Monin, A.S., Yaglom, A.M., 1971. Statistical Fluid Mechanics, vol. 1. MIT Press, Cambridge, MA.
- Nieuwstadt, F.T.M., van Dop, H., (Eds.), 1982. Atmospheric Turbulence and Air Pollution Modelling. Reidel, Dordrecht.
- Oke, T.R., 1987. Boundary Layer Climates, second ed. Routledge, London.
- Panofsky, H.A., Dutton, J.A., 1984. Atmospheric Turbulence – Models and Methods for Engineering Applications. Wiley, New York.
- Pasquill, F., Smith, F.B., 1983. Atmospheric Diffusion, third ed. Ellis Horwood, Chichester.

Air Pollution Meteorology

X-M Hu, University of Oklahoma, Norman, OK, USA

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

Meteorological conditions play important roles in modulating the ambient concentrations of pollutants through different ways. For example, meteorological variables exert important influences on the formation and fate of pollutants such as ozone and aerosols; meteorological processes (e.g., advection and vertical mixing) dictate the dispersion of pollutants. Thus, accurately simulating meteorological conditions is critical for correctly simulating pollution events. Uncertainties are associated with model treatments for various processes in air quality models. Methods such as ensemble simulation and parameter estimation could potentially improve air quality simulations.

Description of Pollutants in the Air

The United States Environmental Protection Agency (U.S. EPA) set National Ambient Air Quality Standards (NAAQS) for six principal pollutants in 1990s to provide protection for public health and the environment. These six principal pollutants are called 'criteria' pollutants, which include carbon monoxide (CO), lead, nitrogen dioxide (NO₂), ozone (O₃), sulfur dioxide (SO₂), and particles (or aerosols). The NAAQS are periodically reviewed and revised to be in line with the updated science. Among the six criteria air pollutants, the formation of O₃ and aerosols involves the most complicated processes. Ozone is a secondary pollutant, produced from oxides of nitrogen and reactive organic gases in the presence of sunlight. When O₃ reaches critical levels, adverse environmental effects are expected for human health, crops, and natural vegetation. The adverse environmental effects of O₃ were first reported in Los Angeles during 1940s. Now it is realized that O₃ pollution is no longer confined to Los Angeles and it affects major urban locations in the world. Elevated O₃ concentrations were also reported in rural and even remote regions. Aerosols range from nanometers to hundreds of micrometers (μm), coming from both primary and secondary sources. Elevated concentrations of aerosols can cause or enhance respiratory, cardiovascular, infectious, and allergic diseases. The primary parameters that determine the environmental and health effects of aerosols are their concentration, size, structure, and chemical composition, which are spatially and temporally highly variable. The formation and fate of O₃ and aerosols have received extensive attention by the research community.

Structure of the Atmospheric Boundary Layer and Its Relationship with Plume Behaviors

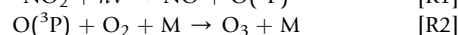
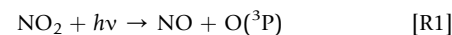
The atmospheric boundary layer is defined as the lowest part of the troposphere that is directly influenced by the presence of the earth's surface, and responds to surface forcing within a timescale of about an hour or less. The variation of the boundary layer plays a critical role for dictating the dispersion of pollutants since most pollutants are emitted or formed in the boundary layer. The boundary layer depth may vary from hundreds of meters to a few kilometers. Over oceans, the

boundary layer depth varies relatively slowly in space and time due to little change of ocean water temperature. The boundary layer depth varies dramatically over the land. During the daytime, due to the surface heating from shortwave radiation, turbulence is generated in the lower 1–2 km above the ground, which is called the mixed layer. The turbulence tends to mix heat, momentum, moisture, and pollutants uniformly in the mixed layer. A stable layer at the top of the mixed layer restrains the vertical extent of turbulence. This layer is called the entrainment zone because entrainment into the mixed layer occurs at this layer. After sunset, turbulence decays in the formerly mixed layer. The upper portion of the formerly mixed layer becomes the residual layer, in which the state variables and concentrations of pollutants remain mostly invariant. The lower portion of the formerly mixed layer is transformed into a stable boundary layer, which is characterized by statically stable air with weaker, sporadic turbulence. After sunrise of the following day, the mixed layer starts to grow again.

A large portion of anthropogenic emissions is released in the form of plumes. Due to higher temperatures compared with the ambient air, these plumes are typically highly buoyant. Depending on the thermal conditions of plume itself and the ambient atmosphere, the plumes can rise to different heights. Different characteristics of the atmospheric boundary layer dictate the way the plumes are dispersed. Plumes released in the mixed layer may loop up and down initially and become uniformly distributed vertically eventually. Plumes released in the stable boundary layer fan out in the horizontal with little vertical dispersion. Plumes in the residual layer spread with an almost equal rate in the vertical and horizontal, exhibiting a conelike shape.

Tropospheric O₃ Chemistry and Aerosol Processes

The overall tropospheric photochemical O₃ formation mechanism is well known. In the troposphere, O₃ is produced from the photolysis of nitrogen dioxide ([R1]) and the subsequent reaction of the ground state oxygen atoms, O(³P), with molecular oxygen ([R2]).



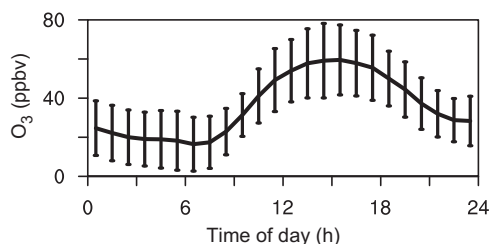
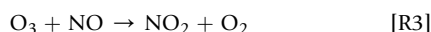


Figure 1 Mean diurnal variation of O_3 and its standard deviation in Beltsville, Maryland, United States, during August 2010.

Once formed, O_3 may be removed from the atmosphere through dry deposition process and NO titration reaction [R3].



Ozone has different variability in different regions. In the continental atmospheric boundary layer, where the concentration of nitrogen oxides (NO_x) is relatively high, local photochemical production of O_3 will contribute to the O_3 maximum in the afternoon in the presence of sunlight and high temperature. Figure 1 shows the diurnal variation of O_3 in Beltsville, Maryland, United States, during August 2010. During the night, O_3 mixing ratios decreased due to NO titration ([R3]) and dry deposition. Photochemical reactions contributed to the O_3 formation during daytime. In the marine boundary layer, where the concentration of NO_x is relatively low, air chemistry can lead to the destruction of O_3 . Figure 2 shows the diurnal variation of O_3 mixing ratio at Kwajalein Atoll situated in the equatorial Pacific Ocean in summer 1999. During the daytime, O_3 photolysis, hydroperoxyl radicals (HO_2), hydroxyl radicals (OH), and bromine atoms (Br) contributed to the destruction of O_3 , which led to the observed minimum O_3 levels in the afternoon. The entrainment of O_3 -rich air from the free troposphere to the local marine boundary layer provided a recovery mechanism of surface O_3 during nighttime.

Aerosol particles are ubiquitous in the atmosphere with diameters ranging from a few nanometers to around hundred micrometers. The ambient aerosol particles are characterized by two modes, the fine mode (with a diameter $\leq 2.5 \mu m$) and the coarse mode (with a diameter $> 2.5 \mu m$). Fine mode aerosol particulate mass is referred to as $PM_{2.5}$, which is believed to pose the largest health risks. The aerosol particulate matter may

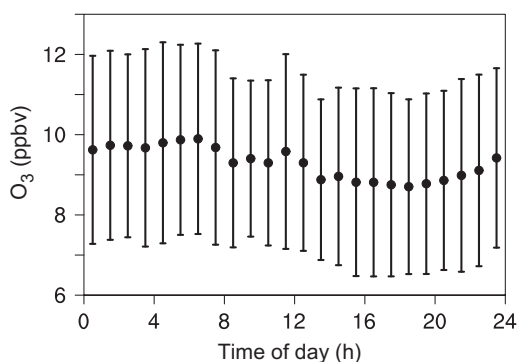


Figure 2 Mean diurnal variation of O_3 mixing ratio and its standard deviation at Kwajalein during July, August, and September 1999.

be either directly emitted or formed in the atmosphere, which are referred to as primary and secondary sources, respectively. Primary sources of aerosol include combustion, windblown dust, pollen, plant fragments, and sea salt. Secondary aerosols are produced in the atmosphere by photochemical processes and added to the preexisting particles through the gas/particle mass transfer process. Most of the mass of $PM_{2.5}$ is composed of secondary aerosol. In some cases more than 90% of the $PM_{2.5}$ mass may be attributed to secondary aerosol.

Effects of Meteorology on Air Pollution

Effects of Meteorology on Biogenic Emissions

Biogenic emissions play an important role in regional air quality and global atmospheric chemistry. Isoprene (C_5H_8) is the predominant volatile organic compound (VOC) emitted by vegetation. It plays a key role in contributing to the formation of O_3 and affecting the lifetime of other species. Biogenic emissions are controlled by ambient environmental variables, most notably temperature and light. Increases in temperature normally lead to increased isoprene emissions. Future climate change (e.g., temperature increase) is expected to increase biogenic emissions, which will likely influence regional air quality.

Effects of Meteorological Variables on the Formation of O_3 and Aerosol

Episodes of high concentrations of surface O_3 usually occur during the summertime in stagnant air under dry, sunny weather conditions. On a local scale, intense solar radiation favors photochemical O_3 production. Ozone generally increases with increasing temperature and decreases with increasing relative humidity. Warmer temperature enhances O_3 production through affecting photochemical rate constants and biogenic emissions. Water vapor affects O_3 abundance through its consumption of $O(^1D)$ via the reaction $O(^1D) + H_2O \rightarrow 2OH$.

Meteorological variables affect the formation of precursors of aerosols through modulating the reaction efficiency. Once the semivolatile precursors of aerosols are formed, meteorological variables also affect the partitioning of those species between the gas phase and aerosols. The partitioning of semivolatile species depends highly on temperature and relative humidity. Low temperature and high relative humidity favor the partitioning of semivolatile species into the aerosol phase while high temperature and low relative humidity favor the partitioning of semivolatile species into the gas phase.

Effects of Meteorology on Dispersion of Pollutants

Meteorological processes (e.g., advection and vertical mixing) dictate the dispersion of pollutants. A good example of a favorable meteorological condition for heavy pollution is the horizontal advection induced by the coastal breeze in Los Angeles, California. The highly populated Los Angeles is surrounded by mountains on three sides and opens to the Pacific Ocean to the west and southwest. Pollutants accumulated over the urban areas in the stagnant morning air are regularly

transported downwind with the onset of the westerly sea breeze in the morning. The air mass is moved back after the onset of land breeze in the evening. The back and forth flow of the air is constrained by the surrounding mountains which allows the air to become highly enriched with pollutants, which likely leads to pollution episodes. Land/sea breezes are also reported to contribute to the elevated O_3 events in other regions such as Houston, Texas and Hampton, Virginia.

Vertical mixing events are reported to affect the variation of O_3 . Figure 3 shows observed O_3 , NO_x , wind vector, and temperature at Beltsville, Maryland, on 11 August 2010. During the period from 1:00 to 4:00 local time (LT), surface O_3 mixing ratio increased by about 30 parts per billion on a volume basis (ppbv) while NO_x mixing ratio decreased by ~ 25 ppbv (Figure 3). During the night of 10–11 August and most of 11 August 2010 the air mass came from the north. A cold front passed the research site, traveling from north to south on 11 August 2010. If O_3 increases resulted from the advection of an upstream polluted plume then mixing ratios of other pollutants such as CO, NO_x would be likely higher. However, NO_x level decreased as O_3 increased. The duration of this nocturnal O_3 increase (several hours) was similar to the 'leaky inversion' event, which was caused by the vertical exchange of air between the surface stable boundary layer and the residual layer above it. Since the residual layer had higher O_3 and lower concentrations of other pollutants, the vertical exchange of trace gases allowed decreases in surface NO_x and increases in surface O_3 .

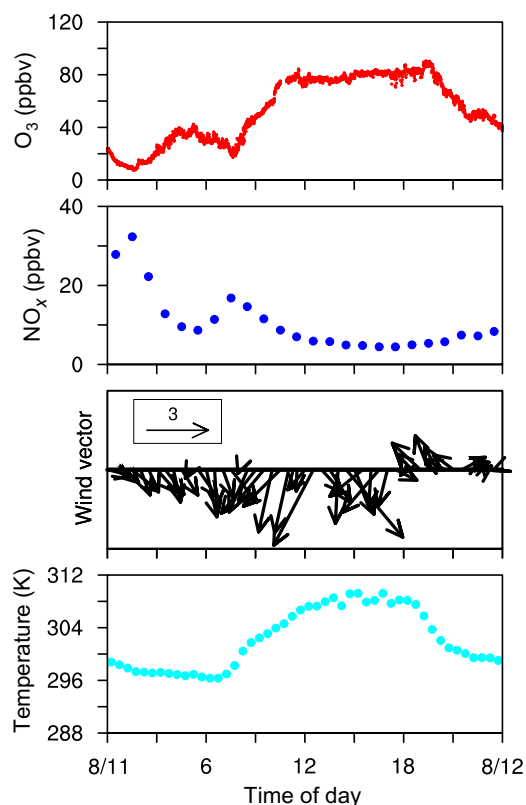


Figure 3 Observed (top to bottom) O_3 , NO_x , wind vector, and temperature at Beltsville, Maryland, on 11 August 2010.

Vertical mixing could be also induced from wind shear and impair air quality near the surface. Nighttime low-level jets (LLJs) occur frequently in the Great Plains. The strong wind shear associated with LLJs was observed to affect the vertical redistribution of O_3 . On the night of 24–25 July 2003, nocturnal O_3 maxima were noticed in the presence of an LLJ (Figure 4). Such phenomenon occurred in Oklahoma quite frequently. Under a calm condition, surface O_3 normally decreased to very low levels (< 10 ppbv) between 20:00 and 22:00 LT due to dry deposition and NO titration. However, on the night of 24–25 July 2003, surface O_3 around Oklahoma city (OKC) increased after 20:00 LT and O_3 concentration stayed at elevated levels (> 40 ppbv) during most of the night. Such nocturnal O_3 maxima (or elevated O_3 concentration) were unlikely due to advection. On the nights of 24–25 July 2003 southerly wind persisted. If the nocturnal surface O_3 maxima were due to southerly advection of O_3 -rich air mass, the upwind site, Goldsby would have experienced higher O_3 maxima than the other downwind sites, which was clearly not the case (Figure 4). Instead, turbulence can be induced by the wind shear associated with the LLJs; the resulted vertical mixing can transport O_3 -rich air mass downward to the surface, thus explaining the frequently observed nocturnal O_3 maxima associated with the LLJs.

Vertical mixing could be also induced from cloud-top radiative cooling and affect the variation of pollutants in the boundary layer. During the springtime, anomalously low O_3 mixing ratios are frequently observed in the Arctic region. Such phenomena are called O_3 depletion events (ODEs). A few mechanisms are proposed to be responsible for the termination of the ODEs. One of the mechanisms is related to the vertical mixing induced from cloud-top radiative cooling. Downdrafts and compensating updrafts induced by the cloud-top radiative cooling can be sufficiently strong to reach the surface. The averaged vertical velocity in the presence of clouds may be as large as 0.6 m s^{-1} in the mixing layer. The vertical mixing associated with cloud updrafts and downdrafts

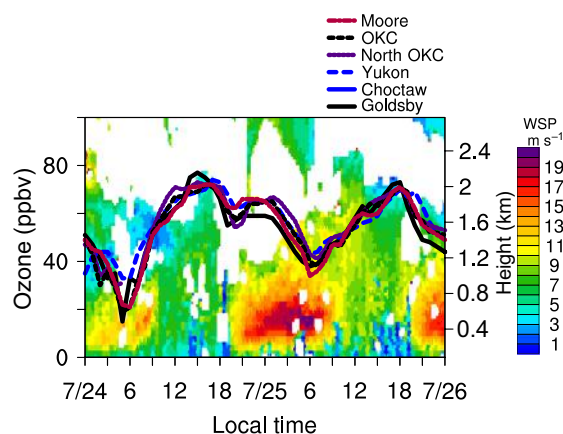


Figure 4 Time–height diagram of wind speed (WSP) in the atmospheric boundary layer and time series of surface O_3 observed on 24–25 July 2003 in the OKC metropolitan area. The left y-axis is O_3 mixing ratio while the right y-axis is the height for wind speed. The six EPA sites in the OKC metropolitan area where O_3 is observed are Moore, OKC, North OKC, Yukon, Choctaw, and Goldsby.

triggered by the clouds can mix the free tropospheric O₃-rich air downward to replenish the O₃ near the surface, thereby terminating the ODEs.

Air Pollution Modeling

Review of Three-Dimensional Air Quality Models

Air quality models (AQMs) integrate our knowledge on how physical and chemical processes affect pollutant concentrations and serve as a numerical laboratory. AQMs could be divided into two categories based on their reference frame: Lagrangian and grid-based Eulerian models. The Lagrangian models have simple structures and low requirements for computational resources, but have limitations in their formulation. The grid-based Eulerian models involve the least restrictive assumptions and are potentially the most powerful.

The grid-based Eulerian models can be further divided into two types based on the coupling between the meteorology (or dynamics) and chemistry in the model: offline vs online. In the off-line models, the meteorology representation/simulation is conducted prior to the simulation of chemistry. The off-line models use either prescribed observation-based or model-based meteorological fields at discrete times (normally hourly) to drive the chemistry models. The meteorological fields are interpolated to the appropriate transport time and coordinate frame for the use of the chemistry simulation in the off-line models. The off-line models are computationally efficient, and one set of meteorological data can be used to drive multiple chemistry simulations; however, they are incapable of simulating chemistry feedback to meteorology. In the real atmosphere, chemical and meteorological processes closely interact through climate–chemistry–aerosol–cloud–radiation feedbacks. Such feedbacks include: (1) aerosols may warm or cool the atmosphere by directly absorbing and scattering solar and terrestrial radiation; (2) aerosols may also indirectly affect cloud microphysics such as formation, albedo, lifetime, and precipitation through serving as cloud condensation nuclei. Those feedbacks are important; however, they cannot be simulated in the off-line models because of the decoupled treatments of meteorology and chemistry. The off-line models are traditionally used for air quality modeling due to the historical separation of meteorology and air quality communities, as well as our limited understanding of the climate–chemistry–aerosol–cloud–radiation feedbacks. In the online models, the meteorology and chemistry are simulated simultaneously in one coordinate frame. The online models make it possible to simulate the complex climate–chemistry–aerosol–cloud–radiation feedbacks given their closely coupled structure and treatments. In addition, the online models can utilize the detail meteorological process to drive chemistry simulation without interpolation.

Eulerian AQMs developed thus far can be categorized into three generations. The first-generation models are relatively simple models, in which dry deposition processes are not included. The second-generation models include deposition processes and have expanded chemical mechanisms in terms of both chemical species and reactions, and use improved numerical integration schemes. Both the first-generation and second-generation AQMs are off-line models. The off-line models

developed in the United States include the Sulfur Transport Eulerian Model, the Regional Oxidant Model (ROM), the Urban Airshed Model with Carbon Bond IV (UAM-IV), the Urban Airshed Model with Variable Grid (UAM-V), the Acid Deposition and Oxidant Model, the Regional Acid Deposition Model, the Community Multiscale Air Quality (CMAQ) Modeling System, SJVAQS/AUSPEX Regional Model Adaptation Project air quality model, the Multiscale Air Quality Simulation Program, and the Urban-to-Regional Multiscale model. ROM, UAM-IV, UAM-V, and CMAQ have been recommended by the U.S. EPA for regulatory applications. Several European regional AQMs are also off-line models, including the European Monitoring and Evaluation Programme model, the European Air Pollution Dispersion model, the Long-Term Ozone Simulation model, and the Regional Eulerian Model with three different chemistry schemes. It was envisioned that the third-generation models should be online models, in which the interactions among chemistry and meteorology will be treated.

Some simple online models were developed since 1960s; however, in most of those, the coupling between meteorology and chemistry was largely incomplete and only limited to a very few prognostic gaseous species. The first online fully coupled meteorology–air quality model is the Gas Aerosol, Transport, and Radiation AQM with a Mesoscale Meteorological and Tracer Dispersion model (referred to as GATOR/MMTD), which solves meteorological and chemical processes simultaneously and considers the two-way feedback between air quality and meteorological parameters. The first community online coupled meteorology–air quality model in the United States, Weather Research and Forecasting Model with Chemistry (WRF/Chem), is developed through the collaborative effort of several research institutes, agencies, and universities. In WRF/Chem, the air quality component is fully consistent with the meteorology component, both of which use the same transport scheme, same gridding, and same physics. The coupling between meteorology and chemistry components in WRF/Chem is relatively comprehensive. Various applications of online AQMs have shown their advantages over the off-line AQMs in many ways. For example, significant model errors of off-line AQMs for chemistry prediction may come from the low updating frequency of meteorological inputs, since some meteorological processes occur in a short timescale (<1 h). Such model errors are found rectified in online AQMs.

Since most of the PM_{2.5} is composed of secondary aerosol, simulating gas/particle mass transfer is essential for accurately predicting aerosol size/composition distributions. The treatment of gas/particle mass transfer in three-dimensional (3-D) AQMs, however, represents one of the major challenges for air quality simulations. Three approaches, equilibrium, kinetic (or dynamic), and hybrid, have been used to simulate gas/particle mass transfer in AQMs. The equilibrium approach assumes equilibrium between gas and aerosol. An aerosol thermodynamic model is used alone to determine the partitioning of semivolatile species between gas phase and aerosol phase. The kinetic approach does not rely on the equilibrium assumption. In this approach, the gas/particle mass transfer due to the difference between the ambient gas concentration and equilibrium gas concentration is explicitly simulated for each size section. The kinetic approach provides the most accurate solution when an appropriate numerical solver and a sufficiently fine

size resolution are used. However, its computational demands hinder its wide applications in 3-D AQMs. The hybrid approach provides a compromise between accuracy and efficiency by using the equilibrium approach for fine particles and the kinetic approach for coarse particles. The hybrid and kinetic gas/particle mass transfer approaches have been implemented in the Model of Aerosol Dynamics, Reaction, Ionization, and Dissolution (MADRID). The improved MADRID has been incorporated into WRF/Chem. The resulted modeling system, WRF/Chem-MADRID, will be used for the following case study.

Case Study of Air Pollution Modeling

Testbed, model configurations, and simulation design

WRF/Chem-MADRID is applied to a 5-day (28 August–2 September) episode from the 2000 Texas Air Quality Study (TexAQS-2000). During this episode more than twenty 1-h O_3 exceedances were observed in the Houston–Galveston–Brazoria area, among which six of them exceeded 150 ppbv. TexAQS-2000 was an intensive field study for air quality issues in the Houston–Galveston area. Houston is the fourth most populous city in the United States with a population of 4 million. Traffic and other anthropogenic activities result in high emission rates of NO_x and VOCs in this area. The biogenic emissions from the forested regions in the northeast of Houston also contribute to the total VOC emissions in the Houston area, depending on the wind direction. Another distinct characteristic of Houston as compared with other large cities in the United States is the numerous petrochemical industries in its surrounding area. Forty percent of the world's production capacity for low molecular weight alkenes is found in the Houston–Galveston area. The O_3 mixing ratios in Houston often exceed the former 1-h NAAQS of 120 ppbv (Note: As of 15 June 2005, EPA revoked the 1-h O_3 standard in all areas except the 8-h O_3 nonattainment Early Action Compact Areas). Under favorable meteorology conditions, the O_3 formation in Houston is rather rapid and some high O_3 events are observed even when the background O_3 mixing ratio is modest, making the O_3 problem in Houston quite unique. Annual mean $PM_{2.5}$ in southeastern Texas is close to the NAAQS of $15 \mu g m^{-3}$ and tends to be higher near urban and industrial areas of Houston. In this area direct emissions contribute approximately 40–50% of $PM_{2.5}$ while secondary sources account for 50–60% of $PM_{2.5}$ and inorganic species dominate the secondary PM. The southeastern Texas is affected by sea-salt emissions from the Gulf of Mexico as well as the anthropogenic emissions. Certain gas/particle mass transfer approaches may fail to predict the distribution of semi-volatile species for the areas where anthropogenic emissions are mixed with sea-salt emission.

The TexAQS-2000 was conducted to improve the understanding of the formation and transport of the pollutants along the Gulf Coast of the southeastern Texas. Intensive measurements of gaseous, particulate, and other hazardous air pollutants were made in the eastern Texas. Significant research efforts are devoted to investigate the pollution issues around the Houston–Galveston area. The modeling study in this work will focus on the effects of different gas/particle mass transfer approaches on the performance of 3-D air quality predictions over the eastern Texas.

The model inputs are set up for a region of $1056 \times 1056 km^2$ with a 12-km horizontal grid spacing and 56 layers vertically

from surface to 100 millibar (mb). Physics options used for the WRF/Chem-MADRID simulations include the Goddard short-wave radiation scheme, the rapid and accurate radiative transfer model for longwave radiation, the Yonsei University (YSU) boundary layer scheme, and the National Center for Environmental Prediction, Oregon State University, Air Force, and Hydrologic Research Lab's land surface scheme. Microphysics is turned off since there was no precipitation around Houston–Galveston area during this episode. The emissions of gas phase species were provided by the Texas Commission on Environmental Quality. The particulate matter emission was obtained from the EPA's 1999 National Emissions Inventory version 3. Eight size sections are used to represent the aerosol size distribution. To test the effects from the improvements of gas/particle mass transfer approaches, simulations are conducted with three mass transfer approaches, i.e., equilibrium, hybrid, and kinetic approaches.

Evaluation of meteorological predictions

Meteorological processes (e.g., sea breeze, low-level jets) play a vital role in O_3 events in Houston. Without properly capturing these meteorological processes, it is unlikely for the model to accurately predict the O_3 events in terms of time of occurrence, location, and peak values. The meteorological predictions are therefore first evaluated before the presentation of the chemical predictions.

Simulated temperature at 2 m (T_2) is evaluated at 32 observational sites. The observed mean temperature during the simulation period is as high as $31.2^\circ C$. Maximum surface temperature exceeded $40^\circ C$ on several days at certain sites during this episode. High temperatures accelerate chemical reactions and favor rapid production of secondary pollutants such as O_3 . WRF/Chem-MADRID captures the diurnal variation of temperature quite well for most sites (with a high correlation of 0.92 with the observation) and only overpredicts T_2 by $0.15^\circ C$.

On the average, wind speed is overpredicted by 8.7%. The mean observed wind direction is south-southwestly while the simulated mean wind is biased by 25° to be more westly. WRF/Chem-MADRID captures the diurnal variations of the wind fairly well at most sites. In addition to the overall statistics, the statistics for wind speed are calculated for nighttime and daytime separately. The performance of wind speed at night (with a correlation coefficient of 0.374, and a Normalized Mean Bias (NMB) of 39.8%) is worse than that during daytime (with a correlation coefficient of 0.492, and an NMB of -10.9%). The worse performance during nighttime may be due to the well-known model deficiency in accurately simulating nocturnal turbulent mixing near the surface, which is a common problem for all numerical weather prediction models.

Planetary boundary layer (PBL) height determines the vertical extent of dilution of pollutants and significant uncertainties are associated with the estimation of PBL height in current AQMs. Evaluation of PBL height at five radar wind profiler sites shows that overall the PBL height is overpredicted by 72.1%.

The high O_3 mixing ratios in the Houston–Galveston area are at times associated with the occurrence of sea breezes. Sea breeze circulations were clearly observed on 29–31 August 2000 over this area. WRF/Chem-MADRID reproduced the observed sea breeze development sequence fairly well even

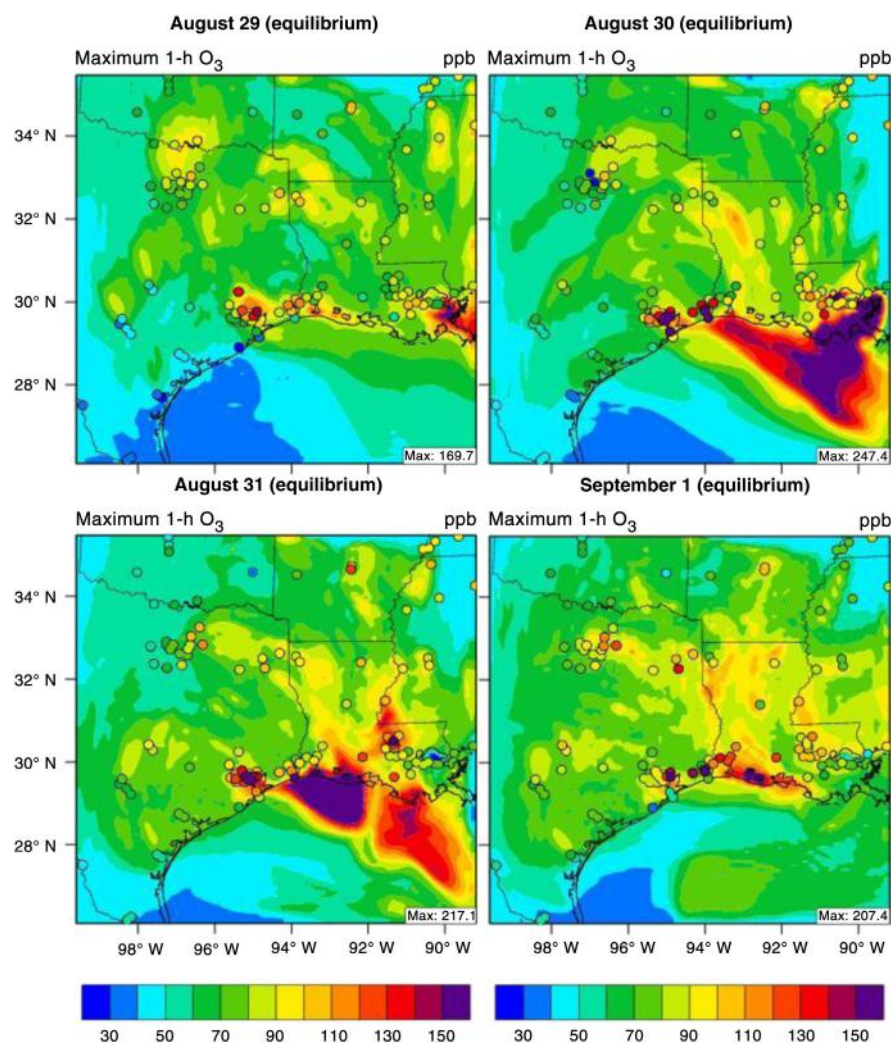


Figure 6 Spatial distributions of maximum 1-h O_3 on 29, 30, 31 August, and 1 September 2000 predicted by WRF/Chem-MADRID with the equilibrium mass transfer approach. The observed values are indicated by shaded circles.

region. The sea breeze continues in the afternoon and the sea breeze front reaches more inland until 18:00 LT.

Results of chemical predictions

Figure 6 shows the spatial distributions of daily maximum 1-h O_3 . There is a cluster of sites experiencing high O_3 mixing ratios around Houston on all 4 days. Substantial photochemical production of O_3 following the emission of precursors (e.g., VOCs and NO_x) from the Houston area leads to prominent O_3 -rich plumes on 30 and 31 August. Driven by the large-scale westerly wind, the O_3 -rich plumes are eastward moving. South part of Louisiana is affected by the O_3 -rich plumes. Such a regional transport pattern was reported to impair the air quality in Louisiana at times.

WRF/Chem-MADRID captures the diurnal variations of surface O_3 at 60 observational sites quite well (Figure not shown). The hourly O_3 throughout the simulation period is underpredicted by 4.6% (1.8 ppbv) from WRF/Chem-MADRID with the equilibrium mass transfer approach (Note: Different mass transfer approaches had negligible effects on O_3

predictions). The daytime hourly O_3 is only underpredicted by 0.4% (0.2 ppbv). This indicates that WRF/Chem-MADRID captures the O_3 formation mechanisms reasonably well.

The differences of daily average $PM_{2.5}$ predicted by WRF/Chem-MADRID with the three different mass transfer approaches are trivial over the inland area while the differences are more prominent over the sea and some coastal areas (Figure 7). The $PM_{2.5}$ concentrations in the plume originating from Houston predicted by the equilibrium approach are much higher than those predicted by the hybrid and kinetic approaches, and the differences are mainly due to nitrate (NO_3^-). The equilibrium approach predicts more fine nitrate (NO_3^- in particles with diameter less than $2.5 \mu m$) and less coarse nitrate (NO_3^- in particles with diameter larger than $2.5 \mu m$) than the hybrid and kinetic approaches (Figure 8). The spatial distribution of simulated nitrate plume matches that of the sodium (Na^+) plume quite well (Figure not shown). Sodium is a tracer of sea-salt aerosol and it is emitted together with chloride (Cl^-) from the ocean. Most sodium chloride is emitted in the coarse mode. Sodium stays in the aerosol phase

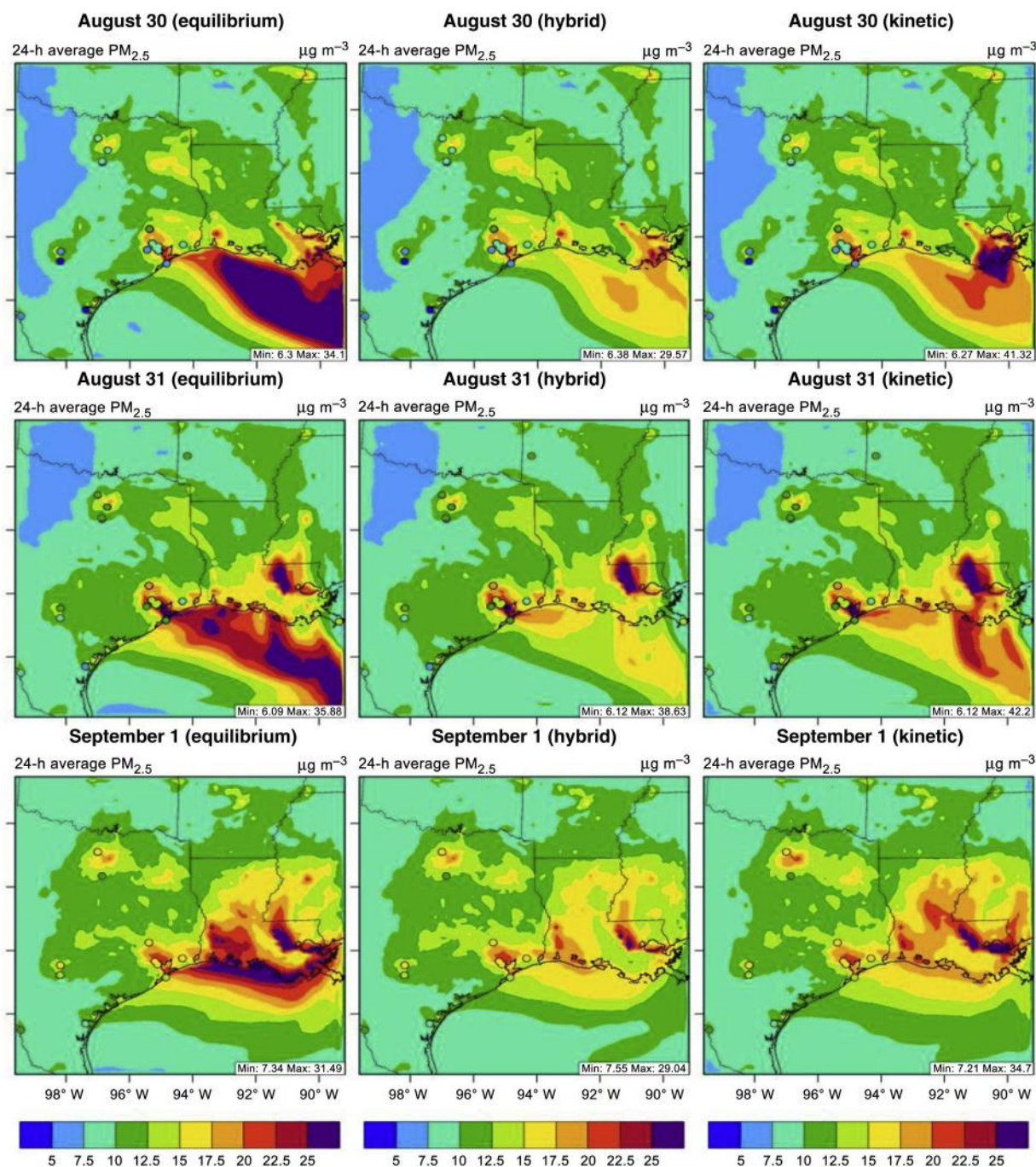
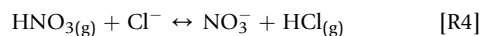


Figure 7 Spatial distributions of daily average $\text{PM}_{2.5}$ mass concentrations simulated with (left to right) the equilibrium, hybrid, and kinetic mass transfer approaches on (top to bottom) 30, 31 August, and 1 September 2000. The measured values are indicated by shaded circles.

since it is nonvolatile, while chloride may exchange between aerosol and gas phase. Nitrate enters aerosol phase through the chloride depletion process ([R4]):



Since the hybrid and kinetic mass transfer approaches both solve the mass transfer for coarse particles kinetically, the chloride depletion process is correctly simulated when anthropogenic pollutant plumes (which contain plenty of

nitric acid, HNO_3) mix with sea salt (which contains sodium chloride). However in the equilibrium approach, the aerosol phase is treated together to equilibrate with gas phase. Even though [R4] can still be simulated by the equilibrium approach, the transferred mass into aerosol phase is redistributed among each section based on initial sulfate (SO_4^{2-}) distribution. Since most sulfates are in the fine mode, the transferred nitrate from the chloride depletion process is erroneously redistributed mostly to the fine mode by the

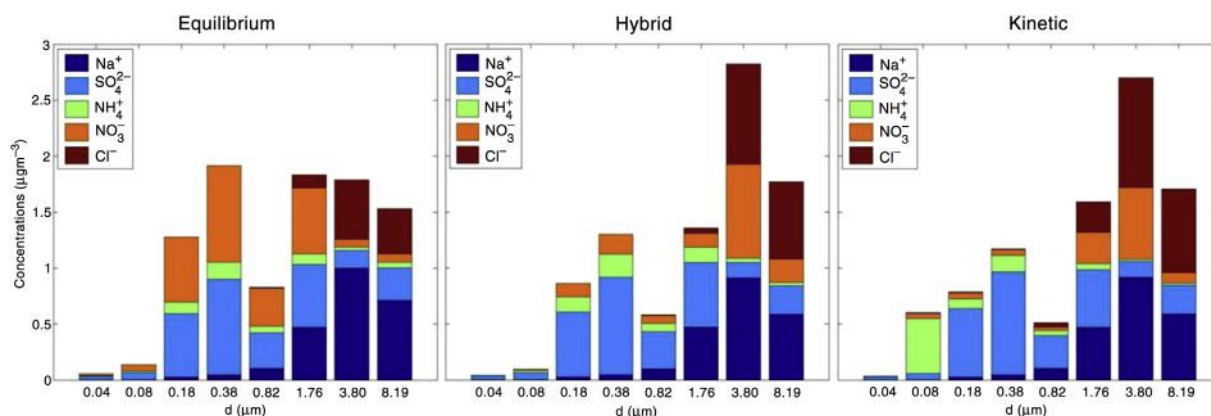


Figure 8 Aerosol size/composition distributions at a coastal site, i.e., Galveston (GALC) on 30 August 2000 predicted by WRF/Chem-MADRID with (left to right) the equilibrium, hybrid, and kinetic gas/particle mass transfer approaches. The x -axis is the diameter of aerosol.

equilibrium approach. Thus WRF/Chem-MADRID with the equilibrium approach predicted higher $PM_{2.5}$ in the coastal area. Due to the prevailing large-scale westerly wind, the $PM_{2.5}$ concentrations in southern Louisiana are enhanced by the pollution plumes originated from Houston (Figure 7).

Improvement of Simulation of Air Pollution Meteorology

Uncertainties of current AQMs

The accuracy of air quality simulations is affected by the uncertainties associated with the model treatments for various processes, including vertical mixing, dry deposition, and chemical mechanisms. Vertical mixing is handled by the PBL scheme in the models. The uncertainties associated with the PBL schemes remain one of the main sources of inaccuracies in air quality simulations. Three PBL schemes in the WRF model, i.e., the YSU scheme, the asymmetric convective model, version 2, (ACM2) scheme, and the Mellor–Yamada–Janjic (MYJ) scheme were evaluated in a modeling study covering the domain of the South Central United States. In that study, the WRF simulations underpredict temperature and overpredict moisture near the surface. Use of the local-closure MYJ scheme produces the largest bias. The YSU and ACM2 schemes both lead to predictions of higher temperature and lower moisture, and thus smaller biases, than the MYJ scheme in the lower atmosphere during daytime because of their stronger vertical mixing. Stronger vertical mixing causes stronger entrainment at the top of PBL, which helps warm and dry the PBL. In the local-closure MYJ scheme, the only entrainment that develops must come from local mixing. Entrainment from penetrating plumes or large eddies is not accounted for. Underestimated entrainment is shown to at least partially cause the colder PBL predicted by the WRF model with the MYJ scheme. During nighttime the WRF model with the YSU PBL scheme produces higher temperatures and lower moisture than with the other two schemes in the lower atmosphere because of its enhanced mixing during nighttime.

Parameter estimation

The accuracy of model simulations is dictated by the uncertainties associated with model treatments for various processes.

Parameter estimation offers a way to improve the accuracy of those model treatments. Parameter estimation is a technique for determining the best value of certain model parameters through data assimilation or similar techniques. When applied to parameterizations of meteorological processes, one hopes to identify optimal parameter values within a given parameterization, with ‘optimal’ defined over some appropriate domain in space and time.

Advanced data assimilation methods, e.g., variational approaches and the ensemble Kalman filter (EnKF), are capable of extracting from observations significant information about the model parameters in addition to the model state. They have been used to counter model errors due to incorrect parameters by calibrating those parameters simultaneously with the model state during the analysis process. The EnKF was applied to estimate the flow-dependent optimal values of two parameters fundamental to the performance of a PBL scheme in the WRF model. Parameter estimation EnKF results in a significant reduction in the model biases of both wind and temperature. Also, deterministic forecasts with updated parameters outperform forecasts with standard parameter settings.

Ensemble simulation

There are dramatic uncertainties in the air quality simulations due to uncertainties in the initial meteorological/chemical conditions and model treatments of physical and chemical processes. Even with plentiful observations, analysis inaccuracies are unavoidable, so a single-minded pursuit of improved initial conditions is inadvisable. Given the uncertainties associated with various model treatments under various conditions, improving model performance in a single deterministic simulation through pursuing perfect model treatments for all the conditions is also unlikely. Instead, ensemble simulations should be utilized to span the range of possible outcomes on a given day, and that uncertainty should be incorporated into the regulatory analysis. As with simulations, deterministic photochemical forecasts are similarly unlikely to be successful. An ensemble forecasting system, incorporating as many sources of error as possible, can provide guidance on not only the most likely evolution of pollutants but also the range of possibilities.

Acknowledgment

Fuqing Zhang, John W. Nielsen-Gammon, and David C. Doughty provided valuable comments to improve the manuscript.

Further Reading

- Hass, H., Builtjes, P.J.H., Simpson, D., Stern, R., 1997. Comparison of model results obtained with several European regional air quality models. *Atmospheric Environment* 31, 3259–3279.
- Hu, X.-M., 2008. Incorporation of the Model of Aerosol Dynamics, Reaction, Ionization, and Dissolution (MADRID) into the Weather Research and Forecasting Model with Chemistry (WRF/Chem): Model Development and Retrospective Applications (Ph.D. dissertation). N. C. State University, Raleigh. July.
- Hu, X.-M., Zhang, Y., Jacobson, M.Z., Chan, C.K., 2008. Coupling and evaluating gas/particle mass transfer treatments for aerosol simulation and forecast. *Journal of Geophysical Research* 113, D11208. <http://dx.doi.org/10.1029/2007JD009588>.
- Hu, X.-M., Nielsen-Gammon, J.W., Zhang, F., 2010. Evaluation of three planetary boundary layer schemes in the WRF model. *Journal of Applied Meteorology and Climatology* 49, 1831–1844.
- Hu, X.-M., Zhang, F., Nielsen-Gammon, J.W., 2010. Ensemble-based simultaneous state and parameter estimation for treatment of mesoscale model error: a real-data study. *Geophysical Research Letters* 37, L08802. <http://dx.doi.org/10.1029/2010GL043017>.
- Hu, X.-M., Sigler, J.M., Fuentes, J.D., 2010. Variability of ozone in the marine boundary layer of the equatorial Pacific Ocean. *Journal of Atmospheric Chemistry* 66, 117–136.
- Hu, X.-M., Zhang, F., Yu, G., Fuentes, J.D., Wu, L., 2011. Contribution of mixed-phase boundary layer clouds to the termination of ozone depletion events in the Arctic. *Geophysical Research Letters* 38, L21801. <http://dx.doi.org/10.1029/2011GL049229>.
- Hu, X.-M., Doughty, D., Sanchez, K.J., Joseph, E., Fuentes, J.D., 2012. Ozone variability in the atmospheric boundary layer in Maryland and its implications for vertical transport model. *Atmospheric Environment* 46, 354–364.
- Jacobson, M.Z., 1994. Developing, Coupling, and Applying a Gas, Aerosol, Transport, and Radiation Model to Study Urban and Regional Air Pollution (Ph.D. dissertation). Department of Atmospheric Sciences, UCLA, 436 pp.
- Nielsen-Gammon, J.W., Hu, X.-M., Zhang, F., Pleim, J.E., 2010. Evaluation of planetary boundary layer scheme sensitivities for the purpose of parameter estimation. *Monthly Weather Review* 138, 3400–3417.
- Peters, L.K., et al., 1995. The current state and future direction of Eulerian models in simulating the tropospheric chemistry and transport of trace species: a review. *Atmospheric Environment* 29, 189–222.
- Russell, A.G., Dennis, R., 2000. NARSTO critical review of photochemical models and modeling. *Atmospheric Environment* 34, 2283–2324.
- Russell, M., Allen, D.T., 2004. Seasonal and spatial trends in primary and secondary organic carbon concentrations in southeast Texas. *Atmospheric Environment* 38, 3225–3239.
- Stull, R.B., 1988. *An Introduction to Boundary Layer Meteorology*. Kluwer, Norwell, Mass.
- Zhang, F., et al., 2007. Impacts of meteorological uncertainties on ozone pollution predictability estimated through meteorological and photochemical ensemble forecasts. *Journal of Geophysical Research* 112, D04304. <http://dx.doi.org/10.1029/2006JD007429>.
- Zhang, Y., 2008. Online coupled meteorology and chemistry models: history, current status, and outlook. *Atmospheric Chemistry and Physics* 8, 2895–2932.

Coherent Structures

FTM Nieuwstadt, Delft University of Technology, Delft, The Netherlands
JCR Hunt, University College London, London, UK

© 2003 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 1, pp 228–233, © 2003, Elsevier Ltd.

Introduction

From the earliest times writers and artists have given us their verbal and pictorial images of the characteristic and repeating patterns of the irregular eddying motions in the lower part of the atmosphere. One can see their form in long streaks of snow (called by one writer 'rivers of wind'), vortices spurring from the desert surfaces in sandstorms, billowing fog-banks and towering cumulus clouds reaching 10 km upwards to the tropopause (e.g. Scorer 1978). The scientific study of these coherent structures began in 1802–03 when Lamarck in Paris and Luke Howard in London noted that there are only a few types of clouds which can therefore be usefully classified. Howard, whose nomenclature (cumulus, stratus, cirrus) has only been slightly changed by later research, had the insight to realize that those structures define the essential dynamics governing the flow. The importance of his insight was recognized by Osborne Reynolds in 1895 when he was establishing the fundamental statistical laws of the 'new' subject of turbulence.

It is now generally accepted that the comprehensive study of any type of turbulent flow, especially atmospheric flows, requires combining the statistical approach with the description and analysis of coherent eddy structures, i.e. motions driven by the unstable inertial and buoyancy forces in the flow that have a characteristic, repeatable and persistent form yet are internally unpredictable and occur randomly in space and time. Development of the dual approach has improved the understanding of the approximations in the statistical approach (as Prandtl indeed first pointed out in 1925), while the statistical analysis also provides a quantitative model for the coherent structures (e.g. Holmes et al. 1996).

When the only systematic measurements were obtained from time series recorded by isolated instruments, it was natural that the study of atmospheric motions was essentially statistical, analyzing correlations and spectra (Panofsky and Dutton 1984) and then using the governing equations of fluid flow to calculate the mean dynamics and energetics, following Reynold's analysis. However, the new technology and capability of multipoint tower and radar measurements (e.g. Kaimal and Finnigan 1994) together with the computational capacity to perform numerical simulations (at varying degrees of accuracy) have now provided nearly complete descriptions of the instantaneous and time development of coherent structures in different atmospheric conditions. These broadly confirm earlier concepts based on observations, clouds, bird flight and dust patterns.

Structures tend to be particularly well defined where one type of structure dominates the boundary layer, such as thermal plumes in unstable boundary layers or Kelvin–Helmholtz billows (which e.g. occur in the stable layer), the undulating top of the mixed layer and sea breeze fronts. When there are

more types of structures or where these are evolving from one form to another, there are no dominant structures and their descriptions are less precise. Quantitative evaluation of their velocity and pressure fields have shown that structures contribute significantly to the overall dynamics of the flow. That is why they tend to determine the form of the velocity spectra, e.g. of the vertical and horizontal components, especially in the distorted eddy structure near the ground. Just as important in practice, structures also affect the statistics of extreme events, such as very large gusts or high downbursts of pollution.

This survey article reviews the various coherent structures in atmospheric turbulence. First the horizontally homogeneous boundary layer is considered, then the structure of the boundary layer over inhomogeneous and rough terrain and equally importantly over the oceans.

The Stationary, Horizontally Homogeneous Atmospheric Boundary Layer Over Flat Terrain

In **Figure 1** are classified the horizontally homogeneous atmospheric boundary layer in terms of the ratio h/L_{MO} between the boundary-layer height h and the Monin–Obukhov length L_{MO} and the ratio z/h , where z is the height above the surface. The Monin–Obukhov length characterizes the height above the ground where buoyancy starts to dominate shear. Since the forms of eddy structures depend on the balance of forces within them and on the proximity to boundaries, the type of coherent structure is characterized by the two parameters h/L_{MO} and z/h . Near the surface, when $h/|L_{MO}| \leq 1$, the main statistical features of the boundary layer tend to be determined locally in terms of z and local fluxes of momentum and buoyancy. However, in the upper part of the boundary layer the large-scale eddying motion are also affected by interactions between the troposphere and the free atmosphere above the boundary layer if these are significant. This means that the typical frequency of the large eddies in the boundary layer, e.g. u_*/h , where u_* is the friction velocity, becomes comparable with the buoyancy frequency of the stable troposphere N_T or the Coriolis parameter f .

The Unstable or Convective Boundary Layer ($h/L_{MO} < 0$)

An unstable or convective atmospheric boundary layer usually occurs when the boundary layer is heated at the surface, which is being warmed by solar radiation. Convective layers can also form when the top of the boundary layer consists of a closed cloud deck, e.g. stratocumulus clouds. In that case the top of the boundary layer cools owing to long-wave radiation from the clouds. The cool air descends which sets up convective motions in the boundary layer. Here, however, the discussion is

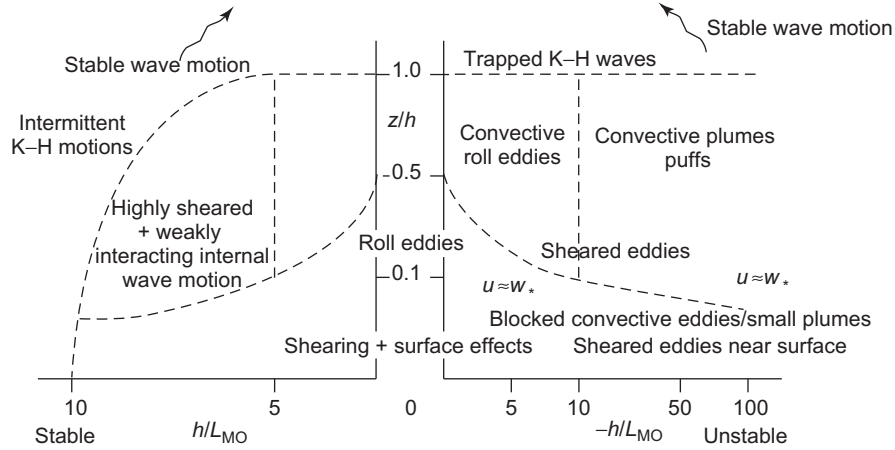


Figure 1 Diagram of the various coherent structures in the atmospheric boundary layer in term of the dimensionless height z/h and the dimensionless stability parameter h/L_{MO} . Adapted with permission from Holtslag, A.A.M., Nieuwstadt, F.T.M., 1986. Scaling the atmospheric boundary layer. *Boundary Layer Meteorology* 36, 201–209.

restricted to the convective boundary layer developing above a surface.

The principle of convection applies to a layer of fluid with a thickness h subjected to a vertical temperature difference ΔT with the high temperature at the bottom and the cold temperature at the top. The layer may become unstable when the buoyancy forces exceed the damping force of viscosity and the tendency of heat to diffuse away from convecting elements. This criterion is usually expressed in terms of the Rayleigh number, defined as

$$Ra = \frac{\beta \Delta T h^3}{\kappa \nu} \quad [1]$$

where β is the buoyancy parameter, κ the heat conduction, and ν the kinematic viscosity. In air, $\beta = g/T_0$, with g the acceleration of gravity and T_0 the absolute mean temperature of the fluid layer.

The critical value of Ra when fluid motions first appear lies in the neighborhood of $Ra \approx 2000$. The first instability mode just beyond this critical value is a flow pattern of two-dimensional rolls. When the Rayleigh number becomes larger, other modes come into play and the flow assumes a pattern of hexagonal cells with a width about equal to the layer of the fluid. These are also known as Rayleigh–Bénard cells.

In the atmosphere the Rayleigh number is typically $Ra \approx 10^{18}$, which is so high that the convective flow is highly turbulent with several types of flow structures. Hexagonal cloud patterns can sometimes be observed, e.g. when air is transported from a cold sea surface over a land surface at higher temperature which sets up strong convection. However, the horizontal scale of these cloud patterns is always much larger than the boundary layer depth, say from 50 to 100 km, which makes them quite different from the Rayleigh–Bénard cells. Generally, the structure of convective eddies, as well as the profiles of the statistical properties of the turbulence, are determined by the ratio of the surface buoyancy flux $F_B = \beta w' \theta'$ to the rate of transfer of mechanical energy (u_*^3/h) caused by turbulent shear stress. Alternatively, this ratio may be expressed as h/L_{MO} , where $L_{MO} = -u_*^3/F_B$. Note that this ratio

also indicates the relative strength of the convective turbulent eddies, with a typical velocity w_* , and the shear-dominated eddies, with a typical velocity u_* , since $h/|L_{MO}| = (w_*/u_*)^3$.

Let us now concentrate on the flow patterns found in the convective atmospheric boundary layer, which in **Figure 1** is the region with $-h/L_{MO} > 10$. In the middle and upper part of the boundary layer, i.e. $z/h > 0.1$, the flow organizes itself in large-scale plume- or puff-like structures in which there is a strong flow upwards carrying the warm air from the surface to the top of the boundary layer. Plumes tend to form when there is a constant heat flux at the surface, for example by strong thermal radiation. The upward motion is compensated by a weak downward motion in the area outside the plumes. The consequence of this velocity structure is that the area-averaged vertical velocity fluctuations have a positive skewness.

Near the top of the boundary layer, i.e. $z/h > 0.8$, where there is usually a temperature inversion, the plume structures impinge on this stable layer. As a result the strong vertical motions are converted into horizontal velocity fluctuations which generate Kelvin–Helmholtz type instabilities in the velocity profile near the boundary layer top. Owing to these instabilities, air from above the inversion is mixed with air of the convective layer. This is called entrainment, and it causes the boundary layer to grow during the day. As a result of impinging convective eddies, wave motions are set up which are partly trapped in the inversion layer and partly propagate into the free atmosphere.

Near the surface, in the so-called surface layer, i.e. $z/h < 0.1$ heating produces structures with a size that increases with distance z above the ground. Owing to their mutual entrainment or coalescence, these small plumes organize themselves into larger structures. When $-h/L_{MO} > 10$ these take the approximate form of hexagonal spoke patterns. The lateral extent of large-scale structures in the surface layer is determined mainly by the blocking of the relatively wide downdrafts as they impact on the ground. This mechanism produces a horizontal wind shear close to the surface and generates sheared eddies with a typical velocity u_* , as also occurs in unstratified and stable boundary layers.

When $1 < -h/L_{MO} < 10$ the convective boundary layer occurs in the combination with a mean wind. In that case the

resulting shear is able to organize the plumes into rolls with a horizontal separation of the order of magnitude close to the boundary layer height. These can sometimes be observed as cloud streets.

Stable Boundary Layer

A stable boundary layer occurs when the vertical (potential) temperature gradient in the boundary layer is positive. This is usually associated with the heat flux and the associated buoyancy flux F_B being negative at the surface, i.e. $h/L_{MO} > 0$. However, stable boundary layers may also occur in spatially developing flows, such as in cold fronts when the boundary layer is significantly cooler than the air above. In all cases the Richardson number $Ri > 0$ with Ri defined as

$$Ri = \beta \frac{\partial \bar{\Theta} / \partial z}{(\partial \bar{U} / \partial z)^2}$$

where $\bar{\Theta}$ is the mean potential temperature and $\bar{U}$ the mean velocity. The coherent structures are not as energetic as in convective and neutral flows (see the following subsection on the shear-driven or neutral boundary layer) because buoyancy forces suppress the vertical displacements of fluid elements. By decorrelating vertical and horizontal motions when Ri is greater than its critical value of about $\frac{1}{4}$, the buoyancy forces effectively suppress the input of energy from the mean flow to turbulence. But in such situations wave motions in the upper part of the boundary layer can induce turbulence with significant energy. Turbulence can survive only if sufficient turbulent energy is produced locally by the breakdown of such waves or by the mean shear. The influence of the stable stratification forces reduces the vertical dimensions of the eddy structures in relation to their horizontal dimensions to such an extent that, as clouds and chimney plumes reveal, the eddies take the shape of pancakes or 'blinis'. This is why the turbulent structures in the middle of the stable boundary layer, i.e. $z/h > 0.1$, are not closely coupled to motions at the surface. As with the convective boundary layer, small sheared eddy structures, e.g. longitudinal vortices with diameters of a few centimeters, are observed near the surface.

Experiments show that the turbulent pancake structures persist even when the stable density gradient is so great that Ri exceeds unity. In this situation the turbulence decays, but periodically it tends to be reenergized by overturning internal waves of local shear-driven instabilities as layers move over each other. In the quiescent periods between these events the turbulence tends to collapse into layered chaotic motions in horizontal planes that have some resemblance to two-dimensional turbulence, as observed in the ocean and the laboratory, and as produced by numerical simulations.

The Shear-Driven or Neutral Boundary Layer

As the heat flux to and from the surface decreases and the wind speed increases, i.e. $h/|L_{MO}| \ll 1$, progressively a greater proportion of the eddy motion in the boundary layer becomes determined by wind shear, and buoyancy forces become negligible. Then the boundary layer is neutrally stable and is similar to those found in engineering flows where the governing similarity parameter in the Reynolds number $Re = u\ell/\nu$,

where u and ℓ are the characteristic velocity and length scale of the flow. However, there are also differences because the Reynolds number in an atmospheric boundary layer ($Re \approx 10^6$) is in general much larger than the Reynolds number in an engineering flow ($Re \leq 10^5$). In the atmosphere, as observation of 'cat's-paws' moving over the sea surface show very clearly, the large-scale eddies interact with the surface by impinging downward and by 'scraping' along the surface. This top-down production of turbulence differs from the bottom-up instability or 'bursting' processes which occur for low- and moderate-Reynolds-number turbulence. The consequence is that the eddy motion at different heights of the boundary layer are well correlated and the variance of horizontal velocity fluctuations decreases only slowly with height in the atmospheric surface layer, in contrast with the quite rapid variation as a function of height in engineering boundary layers. Another effect is that the length scale of the horizontal velocity fluctuation is much larger in the high-Reynolds number atmospheric case.

Above the surface layer, i.e. $z/h > 0.1$, if the troposphere above the boundary layer is very weakly stratified then the wind velocity profile and eddy structure is influenced by Coriolis acceleration f due to the earth's rotation, and this results in a change of direction of the velocity with distance from the surface. This means that the structure is (weakly) dependent on the ratio $u_*/(fz_0)$ and the boundary layer depth is proportional to u_*/f . Owing to this change in wind direction the velocity profile in the 'Ekman' layer (named after its discoverer) becomes three-dimensional and contains an inflection point where the velocity gradient is maximum. Because of both the instability of this profile and the anisotropy of Reynolds stresses, eddy structures develop in the form of rolls, approximately directed along the wind direction. These longitudinal rolls enhance the vertical transport of momentum to such an extent that, except in ideal neutral conditions, the change in mean wind direction with height is usually less than predicted by Ekman's theory and quite often in the opposite direction.

Eddy Structure in Boundary Layers Over Very Rough and Inhomogeneous Terrain

Over snowfields, deserts and prairies, the land surface is effectively made up from small, regular 'roughness elements' such as snow and sand particles or grass and bushes. But over more irregular terrain and inhabited areas there are larger obstacles such as mountains, buildings and trees that affect the mean flow and the eddy structure significantly. Both types of surface roughness slow down the flow and affect the mass and heat transfer to the surface. Note that even if these obstacles are smoothly shaped, such as hills and valleys with low slopes (say less than $\frac{1}{3}$) or water waves in case of air flow over the sea, they have a substantial effect on the mean structure of the atmospheric boundary layer, especially in stably stratified conditions. But when the obstacle slopes are large (such as mountains with slopes more than $\frac{1}{3}$) or when they are shaped with sharp angles (such as buildings) they are described as bluff and they generate a quite distinct eddy structure.

For the former case, of smoothly shaped obstacles in neutral stratification, it is found that in the 'innerlayer' (which over a hill or wave is typically less than about $\frac{1}{20}$ of the length of the

hill) the turbulence and eddy structures are similar to those over a flat surface. However in the outer region the eddy structures are significantly distorted, leading for example to a amplification of vertical turbulence at the top of the hill and a slight reduction of the horizontal fluctuations. When several hills lie perpendicular to the wind, the distortion of the mean shear by the undulating motion leads to large-scale rolls of streamwise vortices.

There are special characteristics of eddies generated by the flow over bluff obstacles, depending on whether their height δ is much smaller than the boundary layer depth h , such as in the case of buildings, or when δ is comparable with h , as in the case of mountains. Roller eddies (or Kelvin–Helmholtz billows) are produced in the intense shear layer separating the fast-moving stream of air over the top of the obstacles from the slow-moving wakes on the downwind side. These roller eddies tend to transform downwind into longitudinal eddy structures over a distance of a few obstacle heights. There tends to be a distinct boundary between the eddies produced by the obstacle shear layers and those of the oncoming boundary layer which is visible when smoke is released into the wake regions. Notable features of isolated bluff bodies are the horseshoe vortices which are wrapped around the obstacle and then extend downwind. These are visible by an indented region on a sandy or snow-covered surface. They interact unsteadily with the vortical eddy structures shed from the obstacle.

By contrast convective turbulence over hills and other rough terrain is broadly similar to that over level terrain, although the eddy structures have some weak correlation with the surface elevation, surface albedo, and temperature. When the slopes become quite large, the plumes tend to develop preferentially on the slopes and may converge into a single large plume at the top of the mountain. These phenomena are well known to glider pilots, who make use of ascending air currents to stay aloft.

In the stably stratified flow over sloping terrain and smoothly shaped obstacles, such as valley flow at night, the mean flow pattern is determined by a complex interaction between the geostrophic flow and the buoyancy-driven flow down the slope. Since the flow direction tends to vary with height, the eddy structures are of small scale and highly sheared. They are also quite intermittent, because the rate of cooling and the slope-driven flow varies from place to place. Such slope flows tend to have a billowing eddy structure characteristic of gravity currents with a strong local vertical mixing at the head of the current and weaker mixing downwind.

Coherent Structures at the Small Scales

Many measurements of spectra and correlations confirm hypotheses of G. I. Taylor and A. N. Kolmogorov that at the smallest scales of motion the basic statistical structure of turbulence is isotropic. However, even qualitative observations, for example of particle motion and smoke patterns, show that when examined closely, even at the smallest scales, turbulence contains structures with a distinct geometrical form that is elongated and far from isotropic. These two findings are

nevertheless consistent, because these anisotropic structures have no particular orientation, so that their statistical distribution is isotropic. These regions consist of long, elongated threads of high vorticity (parallel to the threads). They are sometimes called ‘worms’. These relatively long-lived structures are the residual motions resulting from the rolling up of short-lived vortex sheets. The dynamics of small-scale isotropic turbulence is determined by the vorticity production and dissipation in these vortex sheets, which are formed by strong straining or deforming motions. The duration of this process is typically short, say of the order of the Kolmogorov time scale (seconds in the atmosphere), and is therefore intermittent.

Despite their small size these structures play a crucial role in environmental processes. The straining affects mixing and chemical reactions, while the vortices affect processes that involve small particles, such as aerosols or cloud droplets. The swirling motion in these vortices or worms cause particles heavier than air to be spun out of them. In other words the worms act like centrifuges, and this results in flow regions with few particles and regions with many particles, also called preferential concentrations.

The effect of preferential concentration may have two consequences in the process of warm-cloud rain formation. The first is related to the fact that in regions where there is a small concentration of water droplets almost no condensation of water vapor on droplets can take place. As a result the water vapor density may grow and become supersaturated again, which implies that new cloud droplets can be formed, resulting in a broad distribution of cloud droplets.

The second effect is related to the process of collision and coalescence of cloud droplets. For cloud droplets in the range of 20–50 μm , the strong vortices may influence the trajectory of the droplets such that they preferentially fall along the side of the vortex with the downward-moving velocity. As a result the settling speed is increased and may reach a factor 80% more than the settling speed of a cloud droplet in quiescent air. Together with the increase in concentration this may result in a larger collision probability and thus in droplet growth due to coalescence. When the droplet becomes larger that about 50 μm the vortices can no longer deflect the droplet from its vertical fall trajectory. In that case the vortices will decrease the fall velocity somewhat with respect to the fall velocity in quiescent air.

Significance of Coherent Structures for Practical Problems

Coherent structures in the atmosphere need to be understood and described in order to deal more effectively with engineering and environmental problems. Examples are wind energy, wind loading on structures, aircraft operation, blowing of dust and snow, propagation of electromagnetic waves, wind shelter design, and dispersion of air pollution. Study of coherent structures is also helpful in interpreting statistical data (such as spectra, correlations, and probability distributions) and can also be used for interpolation when data are not available or for extrapolation to more complex situations. A few examples are now given.

In the convective boundary layer, the structure of the thin updrafts and the broad downdrafts causes the position of the maximum surface concentration of an elevated source of pollution to be much closer to the source than in the neutral boundary layer. This effect was not included in early atmospheric dispersion models, in which the turbulence was assumed to have a Gaussian distribution and which therefore made substantially incorrect predictions. The interaction between the large, buoyancy-dominated eddy structure in the upper part of the layer and the shear structure near the surface also has to be understood when estimating the variation of the dispersion from low-level sources.

The eddy structure in the stable boundary layer is also of importance in the dispersion of air pollutants, especially at night and in cold winter conditions. The way the turbulence is unsteady and very sensitive to a slope (even as low as 10^{-2}) helps explain anomalous effects and the limitations of simple predictive models. For agriculture in valleys where frost pockets form, these intermittent mixing events need to be predicted and if possible avoided by artificial mixing (e.g. by fans or burners).

When there are very strong winds, the boundary layer is neutrally stratified. The calculation of wind energy and wind loads on structures needs information about the turbulent spectra. Spectra show that eddies near the ground are larger in the flow direction than those higher in the boundary layer. This surprising result, widely used in engineering calculations, can be understood in terms of elongated eddies very close to the ground. But near buildings, these eddies are broken up and the spectra change. The rapid change with height of the eddy motion over hills has an effect on the performance of wind turbines, and this should be taken into account in an assessment of the viability of wind energy projects.

See also: **Aviation Meteorology:** Clear Air Turbulence. **Boundary Layer (Atmospheric) and Air Pollution:** Complex Terrain; Modeling and Parameterization; Overview. **Clouds**

and Fog: Classification of Clouds. **Dynamical Meteorology:** Kelvin–Helmholtz Instability. **Mesoscale Meteorology:** Waterspouts. **Turbulence and Mixing:** Turbulent Diffusion.

Further Reading

- Cantwell, B., 1990. Future directions in turbulence research and the role of organized motions. In: Lumley, J.L. (Ed.), *Whither Turbulence? Turbulence at the Crossroads*. Springer-Verlag, New York, pp. 97–131.
- Etling, D., Brown, R.A., 1993. Roll vortices in the planetary boundary layer: a review. *Boundary Layer Meteorology* 65, 215–248.
- Holmes, P., Lumley, J.L., Berkooz, G., 1996. *Turbulence, Coherent Structures, Dynamical Systems, and Symmetry*. Cambridge University Press, New York.
- Holtzlag, A.A.M., Nieuwstadt, F.T.M., 1986. Scaling the atmospheric boundary layer. *Boundary Layer Meteorology* 36, 201–209.
- Hunt, J.C.R., Sandham, N.D., Vassilicos, J.C., et al., 2001. Developments in turbulence research: a review based on the 1999 Programme of the Isaac Newton Institute, Cambridge. *Journal of Fluid Mechanics* 436, 353–392.
- Kaimal, J.C., Finnigan, J.J., 1994. *Atmospheric Boundary Layer Flows*. Oxford University Press, New York.
- Panofsky, H.A., Dutton, J.A., 1984. *Atmospheric Turbulence*. Wiley, New York.
- Koschmieder, E.L., 1993. *Bénard Cells and Taylor Vortices*. Cambridge University Press, Cambridge.
- Robinson, S.K., 1991. Coherent motions in the turbulent boundary layer. *Annual Review in Fluid Mechanics* 23, 601–639.
- Scorer, R.S., 1978. *Environmental Aerodynamics*. Ellis Horwood, Chichester.
- Plate, E.J., Federovich, E.E., Viegas, D.X., Wyngaard, J.C. (Eds.), 1997. *Buoyant convection in geophysical flows*. NATO ASI series. Kluwer, Dordrecht.

Complex Terrain

JJ Finnigan, CSIRO Atmospheric Research, Black Mountain, ACT, Australia

© 2003 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 1, pp 234–244, © 2003, Elsevier Ltd.

Introduction

The boundary layer is the layer of the atmosphere that is influenced directly by the roughness and energy balance of the surface. Much of the character of the boundary layer, therefore, is impressed upon it by the particular nature of the underlying surface. Hence we consider the kinds of boundary layers that develop over surfaces that are inhomogeneous either because the surface cover is changing or because the surface is not flat. Although, more often than not, natural surfaces exhibit both topography and changing surface cover, it is more instructive to deal with these two elements of complexity separately.

Changing Surface Cover

Here we are concerned with horizontal inhomogeneity, ranging from simple changes between one surface type and another to continual changes such as might be seen in farms with fields planted with different crops. Atmospheric flow over such terrain is characterized by the appearance of internal boundary layers over each new surface. If the new surface continues sufficiently far downstream without further change, the new internal boundary layer replaces the old boundary layer and eventually a new geostrophic balance is struck between the surface and the synoptic flow above the boundary layer. If the surface character changes continually, however, the impact of each internal layer only extends up to some blending height, above which the total boundary layer behaves as if it were flowing over a surface with properties that are some average of the different patches.

Topography

Hills and valleys affect boundary layer flow because the pressure field that develops as the atmosphere flows over them accelerates and decelerates the near-surface flow. In a relatively thin layer near the surface, analogous to an internal boundary layer, changes in turbulent stresses strongly affect the mean flow, but at higher levels the changes in mean wind speed are essentially inviscid. The pressure field that develops about any given hill is strongly dependent upon the stratification of the atmosphere flowing over it, which can be characterized by a Froude number. Hence, the scale of the topography profoundly affects the resultant boundary layer flow patterns. Those over a very large hill, whose pressure field is largely determined by the displacement of the stratified synoptic flow above the boundary layer, are quite different from those over a smaller hill, where flow displacement is confined within the neutral or unstable boundary layer. Here we will confine our attention to smaller hills.

We will be concerned with the boundary layers that develop over terrain with these two kinds of complexity, due to surface cover and to topography, and will concentrate especially on

two aspects of their description: the windfields that we observe within them and the surface stresses beneath them. Our introduction of spatially averaged equations below suggests one motivation for this. Mathematical model used for climate or weather prediction have horizontal resolutions $50 \times 50 \text{ km}^2$ so that the windspeed averaged over grid cells $500 \times 500 \text{ km}^2$ in area has to be related to some average of the surface properties within the cell. At the same time we want to know how to relate measurements of windspeed and other variables at points in an evolving boundary layer to the surrounding landscape.

Notation

We use a right-handed rectangular Cartesian coordinate system, $x_i(x, y, z)$ with $x_1(x)$ aligned with the mean velocity at the surface and $x_3(z)$ normal to the ground surface. Velocity components are denoted by $u_i(u, v, w)$ with $u_1(u)$ the streamwise component and $x_3(w)$ the vertical component. Time averages are denoted by an overbar (e.g., $\bar{u}$) and departures from the time mean by a prime (e.g., u'). Area averages over the x - y plane are represented by angle brackets (e.g., $\langle \bar{u} \rangle$).

Changing Surface Cover

We will look first at simple changes of surface roughness such as those between bare soil and an irrigated crop. Once we have established the nature of simple transitions from one type of surface to another we will be in a position to describe the boundary layer over patchy surfaces.

Local Advection: The Wind Field

Local advection refers to situations where the effects of surface changes do not propagate above δ_{ASL} , the depth of the atmospheric surface layer. Imagine an equilibrium surface layer flow, characterized by a logarithmic profile with roughness length z_{01} and displacement height d_1 , eqn [1].

$$\bar{u}_1(z) = \frac{u_{*1}}{\kappa} \ln\left(\frac{z - d_1}{z_{01}}\right) \quad [1]$$

$u_{*1} = \sqrt{\tau_{01}}$ is the friction velocity, defined as the square root of the kinematic surface stress τ_0 , and κ is von Karman's constant. This flow encounters a new surface with roughness length z_{02} and displacement height d_2 . We will assume that the boundary is perpendicular to the surface wind vector. As the air flows over the new surface it either slows down because of increased surface friction (smooth-rough change, $z_{02} > z_{01}$) or speeds up because the surface friction falls (rough-smooth: $z_{02} < z_{01}$). The effect of this acceleration or deceleration, which is initially confined to the air layers in contact with the new

surface, is diffused vertically by turbulence, and the effect of the change is felt through a steadily growing internal boundary layer of depth $\delta_i(x)$ (see Figure 1(a)).

The effects of the change are also transmitted by pressure forces that are associated with any change in streamline height that follows if d_1 is not equal to d_2 and this pressure perturbation is not confined to the internal boundary layer. Its effect is negligible, however, except very close to the transition, even when the change in displacement height is significant and for the rest of this section we will ignore it. From now on we will also avoid writing $z - d$, assuming that the origin of the z coordinate is adjusted appropriately to include d .

The strength of the roughness change can be characterized by the ratio of the roughness lengths, M^* , or its logarithm, M (eqn [2]).

$$M^* = \frac{z_{01}}{z_{02}} \quad M = \ln\left(\frac{z_{01}}{z_{02}}\right) = \ln(z_{01}) - \ln(z_{02}) \quad [2]$$

Within the internal boundary layer the flow displays characteristics of the downstream surface. Outside it, apart from the small perturbation caused by the pressure pulse at the transition, the flow field is identical to that upwind (see Figure 1(b)). The internal boundary layer depth $\delta_i(x)$ is usually defined, therefore, as the height at which the downwind velocity $\bar{u}_2(z)$ attains a fixed fraction (e.g., 99%) of its upwind value at the same height.

The growth of the internal boundary layer is caused by turbulent diffusion and, if we take the characteristic diffusion velocity as $u_{*2} = \sqrt{\tau_{02}}$, the downstream friction velocity, then we can write eqn [3].

$$\frac{d\delta_i}{dx} = \frac{Bu_{*2}}{\bar{u}_2(z)} \quad [3]$$

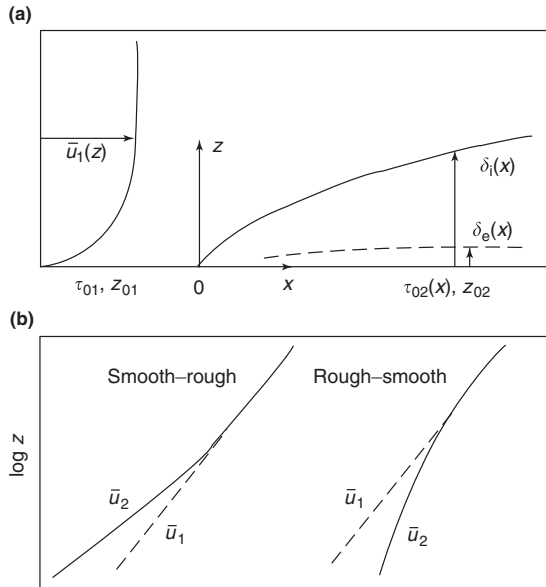


Figure 1 (a) Schematic diagram of internal boundary layer growth. The inner equilibrium region is marked by the dashed curve. This region is not expected to begin until some distance after the roughness change. (b) Logarithmic velocity profiles after a roughness change. The upwind profile is denoted by a dashed line.

To integrate eqn [3] we need an expression for $\bar{u}_2(z)$; for $\delta_i < \delta_{ASL}$ the simplest assumption is that shown in eqn [4].

$$\bar{u}_2(z) = \frac{u_{*2}}{\kappa} \ln\left(\frac{z}{z_{02}}\right) \quad [4]$$

Then stipulating that $\delta_i(x) = 0$ at $x = 0$, and locating the origin of coordinates at the roughness change, from eqns [3] and [4] we obtain eqn [5].

$$\frac{\delta_i(x)}{x} \left[\ln\left(\frac{\delta_i(x)}{z_{02}}\right) - 1 \right] = B\kappa \quad [5]$$

Equation [5] provides a qualitative description of the growth of the internal boundary layer and, with the experimentally determined constant $B \approx 1.25$, provides a good quantitative measure of $\delta_i(x)$ for smooth-rough changes and also for rough-smooth transitions if M is less than 2. When the rough-smooth change is larger ($M \geq 2$), eqn [5] tends to underestimate the growth in $\delta_i(x)$ because then diffusion downstream of the roughness change is controlled for some distance by the slowly decaying upstream turbulence.

To obtain eqn [5] we assumed that the velocity profile within the internal boundary layer was logarithmic all the way up to $\delta_i(x)$. This is a gross oversimplification, however. In Figure 1(a) we have identified an inner equilibrium layer, $\delta_e(x)$ at the bottom of the internal boundary layer. Only in this layer has the flow attained local equilibrium with the new surface with the shearing stress $\tau_2(z)$ approximately constant with height and the velocity profile $\bar{u}_2(z)$ obeying eqn [4]. An estimate for $\delta_e(x)$ can be obtained by first writing an approximate equation for the streamwise momentum balance that ignores any pressure perturbation at the roughness change and also assumes that the changes in the flow field are small

$$\bar{u}_1(z) \frac{\partial \Delta u}{\partial x} \sim \frac{\partial \Delta \tau}{\partial z}$$

where $\Delta u = \bar{u}_2(z) - \bar{u}_1(z)$ and $\Delta \tau = \tau_2(z) - \tau_1(z)$. If we now insist that for local equilibrium to obtain below $\delta_e(x)$, the integral from $z = 0$ to δ_e of the advection term on the left-hand side of this equation must be negligible compared to the perturbation in surface stress, $\tau_{02} - \tau_{01} = u_{*2}^2 - u_{*1}^2$, we obtain

$$\frac{\delta_e(x)}{x} \sim 2 \left(\frac{u_{*2}}{\bar{u}_2} \right)^2$$

Whence

$$\frac{\delta_e(x)}{x} \ln^2\left(\frac{\delta_e(x)}{z_{02}}\right) \sim 2\kappa^2 \quad [6]$$

For the kinds of roughness changes often studied in micrometeorology, the slope $\delta_i/x \approx 1/10$ while $\delta_e/x \approx 1/100$. Hence δ_e corresponds to the height-to-fetch requirements traditionally adopted as a rule of thumb by researchers who wish to apply one-dimensional formulas downwind of a change in surface cover.

For $\delta_i > z > \delta_e$ we have a blending region, where the velocity profile changes smoothly between $\bar{u}(z) = u_{*2}/[\kappa \ln(z/z_{02})]$ and $\bar{u}(z) = u_{*1}/[\kappa \ln(z/z_{01})]$. In this region and downwind of the immediate vicinity of the transition, the velocity and shear stress perturbations are self-preserving, that is, they can be

written as functions of a velocity scale u_0 and a dimensionless height $\eta(x) = z/\delta_i(x)$: eqns [7] and [8].

$$\Delta u(z) = \bar{u}_2(z) - \bar{u}_1(z) = \frac{u_0}{\kappa} g(\eta) \quad [7]$$

$$\Delta \tau = \tau_2(z) - \tau_{*1}^2 = [u_{*2}^2 - u_{*1}^2] h(\eta) \quad [8]$$

A good choice for the velocity scale is $u_0 = |u_{*2} - u_{*1}|$ and the functions $g(\eta)$ and $h(\eta)$ can be found by substituting eqns [7] and [8] into the equations of motion and making a closure assumption to relate $\tau_2(z)$ to $\bar{u}_2(z)$. Several theories have been developed in this way and we will encounter one of them when we consider continually changing surfaces.

Typical examples of the velocity profiles that develop following smooth-rough ($M = -4.8$) and rough-smooth ($M = +4.8$) changes are illustrated in Figure 2(a) and (b). In each case we see the internal boundary layer deepening with downstream distance and the velocity profile slowing in the smooth-rough case and accelerating in the rough-smooth case. In both cases, the lower part of the internal boundary layer is occupied by a logarithmic profile in equilibrium with the new surface, although the true depth of the equilibrium region is exaggerated by the logarithmic height scale. Measured in terms of physical distance, the equilibrium region appears to be established more slowly in the rough-smooth case, but in terms of dimensionless distance x/z_0 there is little difference between the two transitions in the rate at which equilibrium is reached.

Local Advection: Surface Stress

In Figure 3(a) and (b) are plotted measurements of surface shearing stress from the experiment that furnished the velocity profiles of Figure 2(a,b). These results are typical of those from experiments at a range of scales. Two features are noteworthy: the overshoot in stress at the transition and the rapid attainment of a new equilibrium. The overshoot phenomenon is easily explained. In the case of a smooth-rough transition, the

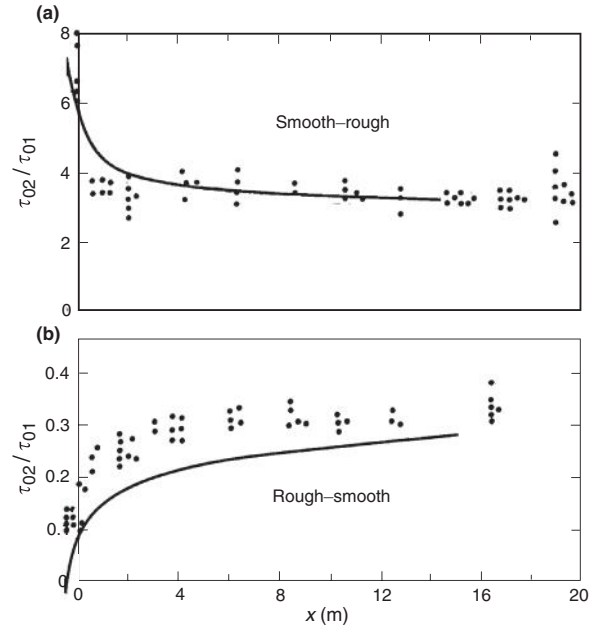


Figure 3 Surface shear stress development after roughness changes (data taken from Bradley (1968), see Kaimal and Finnigan (1994) for details). (a) Smooth-rough change: $z_{01} = 0.02$ mm, $z_{02} = 2.5$ mm, $M = -4.8$. (b) Rough-smooth change: $z_{01} = 2.5$ mm, $z_{02} = 0.02$ mm, $M = +4.8$. The solid line represents eqn [9] with $\delta_i(x)$ calculated using eqn [5].

air stream, traveling relatively rapidly over the smooth surface, generates a high stress on first encountering the increased roughness. As the region of decelerated flow thickens into an internal boundary layer, the velocity of the air in contact with the surface slows and the surface stress falls. In a rough-smooth transition, we see a stress undershoot with a relatively slow airstream generating lower stress when the surface roughness falls but the stress then rising as the flow accelerates.

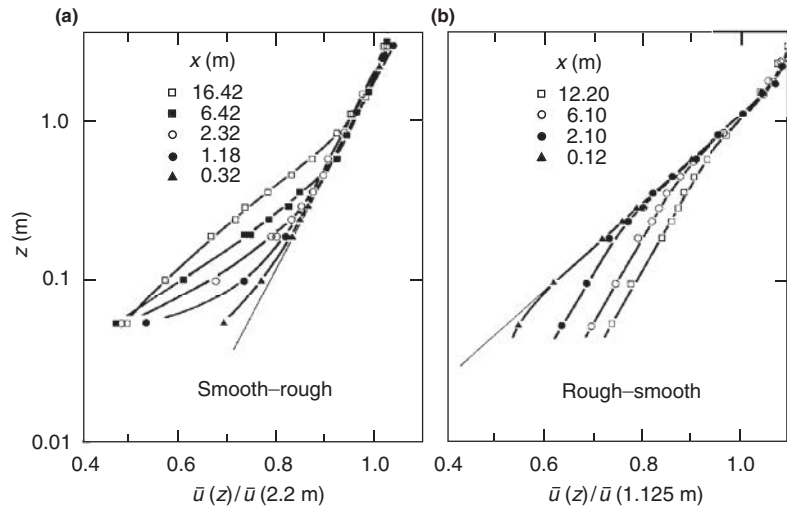


Figure 2 The development of logarithmic velocity profiles after a roughness change (data taken from Bradley (1968) see Kaimal and Finnigan (1994) for details). (a) Smooth-rough change: $z_{01} = 0.02$ mm, $z_{02} = 2.5$ mm, $M = -4.8$. (b) Rough-smooth change: $z_{01} = 2.5$ mm, $z_{02} = 0.02$ mm, $M = +4.8$.

Although sophisticated models of the magnitude of the stress change have been developed, a simple expression can be derived by assuming that the velocity profile obeys eqn [4] with $u_* = u_{*2}$ and $z_0 = z_{02}$ for the full depth of the inner region and then obeys eqn [1] with $u_* = u_{*1}$ and $z_0 = z_{01}$ above a sharp discontinuity at $z = \delta_i$. Matching the two layers leads directly to eqn [9].

$$\frac{\tau_{02}}{\tau_{01}} = \left[1 - \frac{M}{\ln(\delta_i/z_{02})} \right]^2 \quad [9]$$

The result of eqn [9] is plotted on top of the data points in Figure 3(a) and (b) and it is clear that it performs quite well in the smooth-rough case but underestimates the stress change for the rough-smooth transition. Equation [9] relies on an accurate expression for δ_i and we have already noted that eqn [5], which is used to generate the curves in Figure 3(a) and (b) underestimates the growth rate of $\delta_i(x)$ in the rough-smooth case because it discounts the influence of the energetic upstream turbulence on the diffusion of the new internal boundary layer.

Advection on Larger Scales

The formulas we have derived above and the reasoning behind them strictly apply to internal boundary layers that are no deeper than δ_{ASL} , the depth of the atmospheric surface layer, because we have assumed that the mean velocity $\bar{u}_1(z)$ may be described by the logarithmic law. Above δ_{ASL} , both the characteristic velocity and length scales of the turbulence change. The length scale becomes $O(z_i)$, the depth of the whole boundary layer, while the velocity scale depends upon whether the boundary layer is neutrally or unstably stratified. In a neutral boundary layer, the turbulent velocity scale is u_* and, at higher levels, $\bar{u}(z)$ changes more slowly with height than in the logarithmic surface layer. More usually, the surface layer is capped by a convective mixed layer, where the turbulent velocity scale is $w_* = [g/T_0(\overline{w'\theta'})_0 z_i]^{1/3}$ with $(\overline{w'\theta'})$ the surface heat flux. In the mixed layer the mean velocity $\bar{u}(z) = U_M$ is approximately constant with height.

Inserting constant values for the turbulent velocity scale (u_* or w_*) and advection velocity ($\bar{u}(z)$ or U_M) into eqn [3], we see that we can expect $\delta_i(x)$ to grow linearly above the surface layer with a slope between $Bu_*/\bar{u}(z)$ and Bw_*/U_M as the boundary layer varies between neutral stratification and convective mixing. There are relatively few measurements in this regime, but those that exist suggest that the surface layer value $B \approx 1.25$ remains applicable.

The early attainment of a near equilibrium value of surface stress that is shown in Figure 3(a) and (b) masks the continual slow adjustment of this quantity as the internal boundary layer grows out of the surface layer. The new internal boundary layer replaces the old boundary layer when $\delta_i(x)$ equals the old boundary layer depth. This occurs at downstream distances of order $x/z_{02} = 10^6$ in neutral conditions, but possibly much less in a convective boundary layer with a weak mean wind. Current understanding of the magnitude of the geostrophic drag coefficient u_*/G , where G is the geostrophic wind speed, suggests that in the smooth-rough case illustrated in Figure 4, the early equilibrium value of $\tau_{02}/\tau_{01} \approx 3.5$ will fall to $\tau_{02}/\tau_{01} \approx 2.0$ as the new

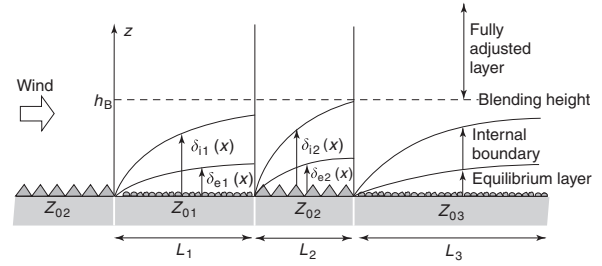


Figure 4 Schematic drawing of the flow structure over a series of surface patches with different roughness lengths z_{0i} and streamwise extents L_i .

boundary layer attains geostrophic balance. For the neutral case, this occurs between the point at which the new boundary layer replaces the old at $x/z_0 \sim 10^6$ and $x/z_0 \sim 10^8$. Attaining a new balance between the surface drag and the geostrophic wind will also change the geostrophic departure, the angle between the surface and geostrophic wind direction. This angle will increase in a smooth-rough change and decrease in a rough-smooth change.

Patchwork Surfaces

Natural surfaces rarely consist of simple changes between two types; rather the surface cover changes continuously. To describe flow over these surfaces we generalize the concept of the internal boundary layer to define the blending height, h_B . Figure 4 illustrates a hypothetical surface consisting of a set of N patches of different surface cover, each occupying a plan area a_i with streamwise extent L_i and having roughness lengths and displacement heights z_{0i} and d_i , respectively. Over each surface an internal boundary layer grows and reaches a depth $\delta_i(L_i)$ by the end of the patch. From the definition of the internal boundary layer we know that above δ_{iMAX} , the height of the deepest internal boundary layer, the velocity profile $\bar{u}(z)$ no longer varies horizontally but attains a spatially averaged value, so we can identify the blending height with δ_{iMAX} (eqn [10]).

$$h_B = \delta_{iMAX} \quad [10]$$

If δ_{iMAX} is smaller than the depth of the surface layer δ_{ASL} , then for $\delta_{ASL} > z > h_B$ the velocity profile will be logarithmic with the form given in eqn [11], where $\langle \rangle$ here denotes an average over the x - y plane.

$$\langle \bar{u}(z) \rangle = \bar{u}(z) = \frac{\langle \tau_0 \rangle^{1/2}}{\kappa} \ln \left(\frac{z}{z_0^{eff}} \right) \quad [11]$$

A central problem over natural surfaces is to find an expression for the effective roughness length z_0^{eff} in terms of patch level roughness lengths z_{0i} and other accessible parameters such as the windspeed above δ_{iMAX} so that the area-averaged surface momentum flux $\langle \tau_0 \rangle$ can be inferred from windspeed measurements or parameterized in models that are unable to resolve the individual patches.

One approach to finding z_0^{eff} is to assume once again that the flow within each internal boundary layer, rather than being self-preserving (eqns [7] and [8]), can be represented by

logarithmic profiles with local roughness lengths and then to average the profiles across the x - y plane (eqn [12]).

$$\langle \bar{u}(z) \rangle = \left\langle \frac{\tau_{0i}^{1/2}}{\kappa} \ln \left(\frac{z}{z_{0i}} \right) \right\rangle = \frac{\langle \tau_{0i} \rangle^{1/2}}{\kappa} \ln \left(\frac{z}{z_0^{\text{eff}}} \right) \quad [12]$$

Whence we obtain

$$z_0^{\text{eff}} = \frac{\langle u_{*i} \ln(z_{0i}) \rangle}{\langle \tau_{0i} \rangle^{1/2}} \text{ and } u_{*i} = \sqrt{\tau_{0i}} \quad [13]$$

Equation [13] is not a very useful formula because we do not, in general, know the stress τ_{0i} on each patch. The simplest recourse is to ignore the correlation between stress and roughness length and to write eqn [14].

$$z_0^{\text{eff}} \approx z_0^M = \langle \ln(z_{0i}) \rangle \quad [14]$$

Because, as we have seen, z_{0i} and τ_{0i} are positively correlated, z_0^M will always be an underestimate of z_0^{eff} , but it forms a useful reference value and provides a first estimate of z_0^{eff} when the variation in roughness length between patches is small.

A more accurate formula for z_0^{eff} has been derived by exploiting the fact that the flow in the internal boundary layers making up the blending region between δ_{ei} and h_B is self-preserving and assuming that in this region a simple mixing-length expression is adequate to express the relationship between shear stress $\tau(z)$ and velocity shear $\partial \bar{u} / \partial z$. These two assumptions allow the shear stress and velocity at the blending height to be related to the local values within the thin equilibrium layer over each patch so that the value of τ_{0i} required to weight the local roughness length z_{0i} in eqn [13] can be inferred. The result is a formula for the effective roughness length that is most simply expressed as in eqn [15], where $h_{ei} = \delta_{ei}(L_i)$ and $h_i = \delta_i(L_i)$.

$$\left\langle \frac{\left[\ln \left(\frac{h_B}{z_0^{\text{eff}}} \right) + \frac{h_{ei}}{h_i - h_{ei}} \ln \left(\frac{h_i}{h_{ei}} \right) - 1 \right]^2}{\left[\ln \left(\frac{h_i}{z_{0i}} \right) + \frac{h_{ei}}{h_i - h_{ei}} \ln \left(\frac{h_i}{h_{ei}} \right) - 1 \right]^2} \right\rangle = 1 \quad [15]$$

Values for h_B , h_i and h_{ei} are readily obtained using eqns [10], [5] and [6].

Equation [15] provides a much better estimate of z_0^{eff} than z_0^M but it also starts to underestimate the momentum-absorbing capacity of a heterogeneous surface when the streamwise length scale of the patches, L_i becomes small. This is because in deriving eqn [15] it is assumed that the equilibrium value of stress, τ_{0i} , applies over an entire patch a_i and the overshoots and undershoots in stress at the roughness transitions that we saw in Figure 4 have been ignored. These strong perturbations in stress just following the change are asymmetrical, the stress increases following a smooth-rough change being generally greater than the decrease after a rough-smooth transition. By ignoring this asymmetry we can underestimate the average stress when the L_i become very small.

Larger-Scale Surface Variability

When the scale of individual surface patches L_i becomes much larger than a kilometer, the blending height will be greater than the depth of the surface layer. At much larger scales ($L_i \gg 10$ km), the new internal boundary layer will replace the

entire planetary boundary layer and the regional surface stress can be calculated by averaging the contributions of essentially independent patches. At the intermediate scale, where $10 \text{ km} > L_i > 1 \text{ km}$, the blending height will be above the surface layer and formulas based on assumptions of logarithmic velocity profiles are inappropriate. Currently there are no simple descriptions of this scale of heterogeneity. Numerical models that can accommodate the diabatic influences that are usually important above the surface layer have been used in particular cases and the average surface stress can be considered to be bounded by the values appropriate to small- and large-scale heterogeneity.

Topography

Air Flow over Isolated Hills

As we did in considering changing surfaces, we will first describe the flow over an isolated hill and then go on to consider how the boundary layer adjusts to continuously hilly terrain. We will confine our attention to hills sufficiently small that the flow perturbations they cause are confined within the boundary layer. In practice this means that the hill height H and the hill horizontal lengthscale L satisfy $H \ll z_i$ and $L \ll h^*$, where h^* , the 'relaxation length' of the boundary layer, is defined as $h^* = z_i U_0 / u_*$ or $z_i U_0 / w^*$ according to whether the flow is neutrally stratified or convectively unstable. The velocity scale U_0 is defined below. The horizontal length scale L is defined as the distance from the hill crest to the half-height point. In continuously hilly terrain it can be more appropriate to use a characteristic wavelength λ as the horizontal length scale. For sinusoidal terrain, $L = \lambda/4$.

In Figure 5 we have sketched the main features of the velocity field about an isolated hill. The figure could represent flow approaching an axisymmetric hill or a 2D ridge at right angles. Close to the surface, the flow decelerates slightly at the foot of the ridge before accelerating to the summit. In the axisymmetric case the deceleration is replaced by a region of lateral flow divergence at the foot of the hill. The wind reaches its maximum speed above the hill top and then decelerates on the lee side. If the hill is steep enough downwind, a separation bubble forms in which the mean flow reverses direction. Whether the flow separates or not, a wake region forms behind

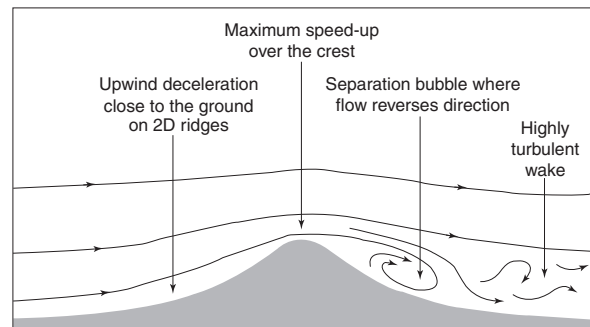


Figure 5 Schematic drawing of the flow over a 2D ridge showing the formation of a downstream separation region when the ridge is steep enough. On an axisymmetric hill, the upwind deceleration region is replaced by a region of lateral flow divergence.

the hill with a marked velocity deficit extending for at least $10 H$ downwind.

The same information is made more concrete in Figure 6, where velocity profiles well upwind, over the hill top and in the wake are plotted. The vertical coordinate z measures height above the local surface. In Figure 6 it is made dimensionless with the inner layer height l , defined below. Upwind we have a standard logarithmic profile, but on the hill top the profile is accelerated with the maximum relative speed-up occurring quite close to the surface at $z/l \sim 0.3$. In the wake we see a substantial velocity deficit extending to at least $z = H$.

Much of the understanding we now have about the dynamics of flow over hills derives from linear theory, which assumes that the mean flow perturbations caused by the hill are small in comparison to the upwind flow. Although, strictly, linear theory is limited to hills of low slope, $H/L \ll 1$, its insights are applicable to much steeper hills. Linear theory supposes a division of the flow field into two main regions, an inner region of depth l and an outer region above, which are distinguished by essentially different dynamics (Figure 7). The balance between advection, streamwise pressure gradient, and the vertical divergence of the shear stress can be expressed in an approximate linearized momentum equation (eqn [16]).

$$U(z) \frac{\partial \Delta u}{\partial x} + \frac{\partial \Delta p}{\partial x} \sim \frac{\partial \Delta \tau}{\partial z} \quad [16]$$

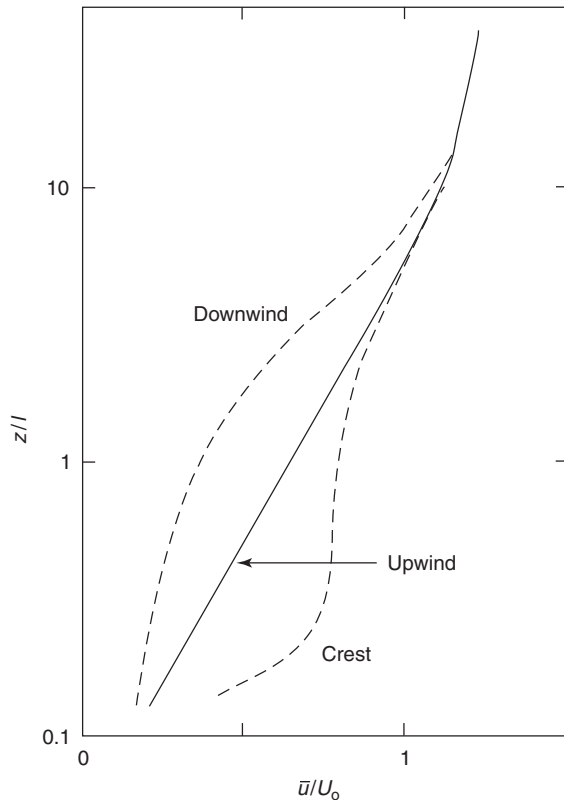


Figure 6 Profiles of mean velocity observed upwind, on the crest, and in the wake region of a hill. The vertical scale is made dimensionless with the inner layer depth, l . Note the position of the maximum speed-up on the crest at $z \sim l/3$.

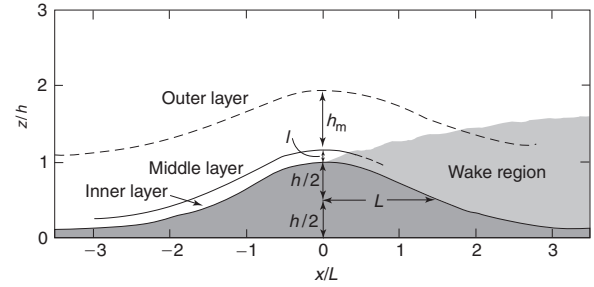


Figure 7 The different regions of the flow over an isolated hill, comprising: inner, middle, outer and wake layers and their associated length scales.

Δu , Δp , $\Delta \tau$ are the perturbations in streamwise velocity, kinematic pressure, and shear stress that are induced by the hill, and $U(z)$ denotes the undisturbed flow upwind of the hill. Well above the surface, perturbations in stress gradients are small and advection and pressure gradient are essentially in balance. Close to the surface an imbalance develops between these terms as the perturbation stress gradient grows. The inner layer height is defined as the level at which the left-hand side of eqn [16] equals the right-hand side.

A second interpretation of l is as the height at which the time taken for a turbulent eddy to be advected over the hill is equal to the eddy turnover time, that is, the typical lifetime before the eddy, generated by interaction with the mean flow, is dissipated. This interpretation tells us that for $z/l \ll 1$ the turbulence will be approximately in local equilibrium, that is, that production and dissipation of turbulent kinetic energy balance locally so that the relationship between the shear stress and the mean flow can be described by a mixing length or eddy-diffusivity. For $z/l \gg 1$, in contrast, the turbulence will experience rapid distortion, where changes to the turbulent stresses will depend on the cumulative straining of eddies by the mean flow as they are advected over the hill. In particular, for $z/l \gg 1$, the response of the mean flow is essentially inviscid because changes in turbulence moments have negligible effect on the mean flow. Hence, although the vertical structure of the undisturbed velocity profile $U(z)$ is entirely the result of turbulent stresses, perturbations to this profile are governed by inviscid dynamics except within the thin inner layer l (as long as $L \ll h^*$).

If the undisturbed upwind profile is taken as logarithmic, $U(z) = u_* / [\kappa \ln(z/z_0)]$, and we adopt a mixing length parametrization to relate the shear stress $\tau(x, y, z)$ to the velocity field over the hill, the first definition for the inner layer depth given above leads to an implicit expression for l (eqn [17]).

$$\frac{l}{L} \ln\left(\frac{l}{z_0}\right) = 2\kappa^2 \quad [17]$$

Equation [17] is very similar to the expression we found for the internal boundary layer height, eqn [5]. This is no accident, as yet a third interpretation of l is as the height to which new vorticity, which is generated at the surface at a rate equal to the streamwise gradient of perturbation pressure, diffuses over the hill.

The pressure field that develops over the hill deflects the entire boundary flow over the obstacle. Its magnitude is

determined, therefore, by the inertia of the faster-flowing air in the outer region and is also related to the steepness of the hill so we expect eqn [18] to hold.

$$\Delta p \sim \frac{H}{L} U_0^2 \quad [18]$$

Scaling arguments reveal that the appropriate definition of U_0 is given by eqn [19].

$$U_0 = U(h_m) \frac{h_m \ln^{1/2} \left(\frac{h_m}{z_0} \right)}{L} \sim 1 \quad [19]$$

The middle layer height h_m divides the outer region into a middle region between l and h_m where shear in the approach flow exerts an important influence on the flow dynamics, and an upper region, where the perturbations are described by potential flow. For a hill with $L = 200$ m, $u_* = 0.3$ m s⁻¹, and $z_0 = 0.02$ m, typical sizes of these scales are $l = 10$ m, $h_m = 70$ m, $U_0 = 6$ m s⁻¹ and $U(l) = 4.5$ m s⁻¹. Note that in the linear theory, the vertical extent of the regions influenced by the hill depends only on the hill length scale, L . The hill height enters only through the influence of steepness H/L on the pressure perturbation that drives all other changes in the flow field.

The pressure perturbation falls to a minimum at the hill top and then rises again behind the hill and propagates essentially unattenuated to the surface. Its scaling gives a strong clue as to why the relative speed-up peaks in the inner layer. Referring again to eqn [16], except very close to the surface the momentum balance is dominated by the pressure gradient and the advection so that $U(z)\Delta u(x, z)/L \sim (H/L)U_0^2/L$. Within the inner layer, as the background flow $U(z)$ becomes much smaller than U_0 , the velocity perturbation Δu must grow to compensate. Eventually, at the bottom of the inner layer the stress gradient dominates the momentum balance and reduces Δu so that the peak in speed-up is found at about $z \sim l/3$. The effects of this shifting balance can be clearly seen in the expression for the relative speed-up derived from linear theory, eqn [20].

$$\frac{\Delta \bar{u}(x, z)}{U(z)} = \frac{H}{L} \left[\frac{U^2(h_m)}{U(l)U(z)} \right] \zeta(x, z_0) \quad [20]$$

$\zeta(x, z_0)$ is a function that factors the precise shape of the hill and the influence of surface roughness into the equation while both the dependence of the driving perturbation pressure gradient on hill slope H/L and the amplification of the speed-up by the ratio of background velocities across the middle layer, $U(h_m)/U(l)$ is evident.

Drag Force on Isolated Hills

We are particularly interested here in the drag force exerted by the hill upon the atmosphere. This is almost entirely a result of the asymmetry in the pressure field about the hill, which results in a net form drag or pressure drag. The hill also produces a negative perturbation in the surface shear stress, equal to the balance between the increase in stress as the wind accelerates to the hill crest and the extended region of reduced stress on the downslope and sheltered wake region. Even on low hills, however, this net reduction in τ_0 is an order of magnitude smaller than the increase in form drag. Over steep hills, flow separation ensures that the pressure on the lee side of the hill

does not recover to its upstream value, leading to a net form drag, but over lower hills without separation the mechanism is more subtle.

The effect of the shear stress in the inner layer on the streamwise velocity perturbation Δu is to displace its peak value slightly upwind of the hill crest and to thicken the inner layer on the downslope side so that streamlines are not symmetrically disposed about the hill but are farther from the surface on the downslope side. The asymmetric perturbation in vertical velocity Δw that accompanies Δu is amplified by the shear in the middle layer and acts to force an asymmetric component in the pressure perturbation Δp , which is determined primarily by the flow at $z = h_m$ and above. When integrated over the hill, this asymmetry in Δp results in form drag. Put more simply, the flow well above the hill acts like an inviscid flow over a surface defined by the streamlines at the top of the inner layer and it is this 'inviscid' flow that determines the pressure field acting on the surface. From this viewpoint it is easy to see that separation will set an upper limit on flow speed-up and pressure drag because, as the hill gets steeper, the upper level flow 'sees' the hill plus separation bubble as the lower boundary condition, effectively increasing L in eqn [20] and limiting H/L .

Linear theory gives an exact expression for the drag on a 2D sinusoidal ridge of wavelength λ as eqn [21], where F_P is the streamwise pressure drag.

$$\begin{aligned} |F_P| &= 2 \left[\frac{U_0}{U(l)} \right]^4 \left[\frac{\pi H}{\lambda} \right]^2 u_*^2 \lambda \\ &= \frac{\pi^2}{2} \left[\frac{U_0}{U(l)} \right]^4 \left[\frac{H}{L} \right]^2 u_*^2 L \end{aligned} \quad [21]$$

Comparing eqns [21] and [20] we can see that the pressure asymmetry is proportional to the square of the velocity perturbation and so is proportional to the square of the hill slope H/L and to the fourth power of the shear amplification factor $U_0/U(l)$.

This formula has been successfully extended to 3D hills by generalizing the hill slope H/L to $\sqrt{A/S_h}$, where A is the frontal area of the hill and S_h its base area.

Furthermore, the same dependence of the pressure drag of a hill upon the square of the hill slope and the fourth power of the ratio of an inner to an outer level velocity scale appears to hold when we go to very steep hills well outside the compass of linear theory. Care must be taken, however, because experimental determinations of hill drag are exceedingly rare and such generalizations of eqn [21] have only been tested against numerical models so far. Nevertheless, they provide a point of departure for a consideration of the effect of hilly terrain on the whole boundary layer.

Effective Roughness Length of Hilly Terrain

Derivations of the logarithmic law in the ASL proceed by an asymptotic matching argument that applies only in the height range $z_i \gg z \gg z_s$, where z_s is a characteristic size of the surface roughness elements. When most of the drag force on the surface is due to the form drag on hills, we expect that $z_s \sim H$ and, if H is a sensible fraction of z_i , we might not expect to observe a logarithmic region at all, the ASL being squeezed between a roughness sublayer, where the flow depends on the

character of the individual hills, and the outer layer flow. Nevertheless, numerical models of flow over ranges of hills suggest that the area-averaged velocity, $\langle \bar{u} \rangle(z)$ generally does have a logarithmic dependence through a layer that can occupy a much greater fraction of the planetary boundary layer depth than the classic surface layer. The experimental evidence is less convincing, partly because area averages are very difficult to measure, but they do not directly contradict the model results.

It seems reasonable then, in analogy with eqn [11], to represent the flow over hilly terrain in the approximate range $z_i/2 > z > 2H$ as in eqn [22].

$$\langle \bar{u} \rangle(z) = \frac{u_*^{\text{eff}}}{\kappa} \ln \left(\frac{z-d}{z_0^{\text{eff}}} \right) \quad [22]$$

Hence,

$$\left(u_*^{\text{eff}} \right)^2 = \frac{\kappa^2 \langle \bar{u} \rangle(z)}{\left\{ \ln \left[(z-d)/z_0^{\text{eff}} \right] \right\}^2} = (-\langle F_P \rangle - \langle F_V \rangle) \quad [23]$$

where $\langle F_V \rangle$ is the surface friction counterpart of $\langle F_P \rangle$. Now since the perturbations in $\langle F_V \rangle$ induced by a hill are much smaller than $|F_V|$, we can write eqn [24].

$$u_*^{\text{eff}} \approx (-\langle F_P \rangle + u_*^2)^{1/2} \quad [24]$$

The area-averaged form drag is obtained by dividing eqn [21] or its 3D equivalent by S_D , the plan area of the domain under consideration, so $\langle F_P \rangle = F_P/S_D$, and we approximate the undisturbed stress u_*^2 by assuming that it can be related to the area-averaged velocity at h_m through the undisturbed roughness length z_0 (eqn [25]).

$$u_*^2 = \frac{\kappa^2 \langle \bar{u} \rangle^2(h_m)}{\ln(h_m/z_0)} \quad [25]$$

This is not a bad assumption when $H \ll h_m$, while over steeper hills the pressure drag term is dominant and we can ignore $\langle F_V \rangle$ for all practical purposes. Equations [23], [24] and [25] can now be combined with [21] or its 3D equivalent to give an expression for the effective roughness length of a hilly landscape. We can take the displacement height d as the average level of the terrain. Since we should not apply eqn [22] too close to $z = H$, the precise specification of d is not critical.

The resulting expression for z_0^{eff} has been compared to mathematical models of the whole boundary layer and of the surface layer and it corresponds to within better than 15% for slopes $\sqrt{A/S_h} < 0.5$. Comparison with field data is more difficult, but the expression matches the available data reasonably well. z_0^{eff} is an increasing function of slope $\sqrt{A/S_h}$ for low to moderate slopes and z_0^{eff} for a range of 2D ridges is substantially larger than for a range of close-packed 3D axisymmetric hills of the same wavelength, a result confirmed by numerical model studies that give values of $z_0^{\text{eff}2D}/z_0^{\text{eff}3D}$ between 3 and 6. An idea of the actual change that hills induce in the effective roughness of a natural surface can be gained by noting that, for 2D ridges with $z_0 \approx 0.1$ m and $\sqrt{A/S_h} \approx 0.5$, we find that $z_0^{\text{eff}}/z_0 \approx 15$, while for close-packed 3D axisymmetric hills with the same z_0 and A/S_h , we find $z_0^{\text{eff}}/z_0 \approx 4$.

Summary

We have seen both similarities and differences between boundary layers over two kinds of complex surface. Over continuously changing horizontal surfaces we observe a blending region where turbulence properties and scaling depart from standard equilibrium ASL forms. The depth of this zone is related to the size of the surface patches and to the rate at which turbulence can diffuse information about the surface changes vertically. For sufficiently large patches the whole boundary layer readjusts to the new surface. Except very close to the edges of distinct patches, pressure effects are negligible but, if the surface is covered with scattered bluff objects like wind-breaks or buildings, these can be responsible for a large fraction of the total drag of the landscape.

Over natural hilly landscapes we also find a region of altered mean flow extending up to $z \sim L$, where L is the horizontal length scale of the hills, but the depth of this region is determined by the pressure field that develops around the hill. Turbulent diffusion, in contrast, affects only a shallow surface layer. Increased momentum absorption by the hills is almost entirely the result of pressure drag and, even for very shallow hills, this substantially exceeds the drag of a flat surface with the same surface texture.

See also: Boundary Layer (Atmospheric) and Air Pollution: Coherent Structures; Convective Boundary Layer; Modeling and Parameterization; Observational Techniques *In Situ*; Observational Techniques: Remote; Overview; Stably Stratified Boundary Layer; Surface Layer.

Further Reading

- Belcher, S.E., Hunt, J.C.R., 1998. Turbulent flow over hills and waves. *Annual Review of Fluid Mechanics* 30, 507–538.
- Blumen, W., 1990. Atmospheric Processes over Complex Terrain. In: Blumen, W. (Ed.), *AMS Meteorological Monographs*, vol. 23. American Meteorological Society, Boston, MA no. 45.
- Bradley, E.F., 1968. A micrometeorological study of velocity profiles and surface drag in the region modified by a change in surface roughness. *Quarterly Journal of the Royal Meteorological Society* 94, 361–379.
- Garratt, J.R., 1992. *The Atmospheric Boundary Layer*. Cambridge University Press, Cambridge.
- Goode, K., Belcher, S.E., 1999. On the parameterisation of the effective roughness length for momentum transfer over heterogeneous terrain. *Boundary Layer Meteorology* 93, 133–154.
- Kaimal, J.C., Finnigan, J.J., 1994. *Atmospheric Boundary Layer Flows: Their Structure and Measurement*. Oxford University Press, New York.
- Schmid, H.P., Bunzli, B., 1995. The influence of surface texture on the effective roughness length. *Quarterly Journal of the Royal Meteorological Society* 121, 1–22.
- Wood, N., Mason, P., 1993. The pressure force induced by neutral turbulent flow over hills. *Quarterly Journal of the Royal Meteorological Society* 119, 1233–1267.
- Xu, D., Taylor, P.A., 1995. Boundary-layer parameterisation of drag over small-scale topography. *Quarterly Journal of the Royal Meteorological Society* 121, 433–443.

Convective Boundary Layer

MA LeMone, National Center for Atmospheric Research, Boulder, CO, USA

© 2003 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 1, pp 244–253, © 2003, Elsevier Ltd.

Introduction

The boundary layer is that layer of a fluid that is directly influenced by the surface. For the convective boundary layer (CBL), the surface influence is felt through rising, buoyant air currents. Over land, the air currents are buoyant because they have been warmed by the surface. Over the ocean, the air currents are buoyant from a combination of warming at the surface, and evaporation of water, which lightens the resulting gas mixture because of its lower molecular weight. These currents rise and cool dry-adiabatically, until they lose their vertical momentum and buoyancy through mixing and reaching heights where the environmental air is warmer. Between the rising currents, air sinks and warms dry-adiabatically. The depth of the dry convective boundary layer is thus determined by the temperature of the updrafts and the change of temperature with height in the environment. Within the CBL, the vertical motions reach values of $\pm 1\text{--}2\text{ m s}^{-1}$, strong enough to support soaring birds or to make flying in aircraft uncomfortable.

Aircraft, radar, and lidar observations of the convective boundary layer suggest a flow structure as shown in [Figure 1](#). Small buoyant elements rise from the warm surface. Some (plumes) remain attached to the surface, others (bubbles) detach as they grow vertically. We will use the word 'thermals' to describe both types of buoyant element. As the thermals rise, they merge. Large eddies, which span the depth of the convective boundary layer, collect the smaller thermals into their upwelling regions. The top of the convective boundary layer is uneven – typically higher where large-eddy motions are upward. While plumes or bubbles can occur anywhere, on average there are fewer thermals in the downwelling portions of the larger eddies, which are thus less turbulent than the upwelling regions. The interfacial or entrainment layer lies within the range of heights occupied by both the CBL air and the atmosphere above at different locations or times.

The convective boundary layer over land forms and dies during the daylight hours ([Figure 2](#)) of sunny days. Starting at sunrise, air rising from the warming ground breaks down the nocturnal inversion. Within 2–3 hours, the CBL grows to about 100 m. Rapid growth typically continues until the CBL reaches a depth of 1–2 km about 11:00 h to 12:00 h, local time. The CBL is in a quasi-steady state for the next 4–5 h, with a slowly changing depth. A few hours before sunset, the ground begins to cool, forming a stable layer near the ground. Plumes no longer rise from the ground, and the turbulence decays in the dry adiabatic layer above. This layer, decoupled from the ground, is often called the residual layer.

We live within the CBL, and its evolution is closely tied to the environmental conditions we experience. Vertical mixing decreases air pollution near the ground from nearby sources, but it can also transport to the ground pollution transported above the CBL from more distant sources. On average, the wind speeds over land are lowest during the night, when the stable nocturnal inversion decouples the surface from the wind overhead; and are largest during the day, when the momentum from above is mixed downward by the CBL. When the CBL grows deep enough for rising thermals to reach their condensation levels, clouds form. The evolution of cloud formation and the depth of the boundary layer affects the maximum surface temperature and the formation of showers.

Mean Vertical Structure

Horizontally averaged profiles of CBL potential temperature (Θ), moisture (Q), and the wind components U and V are controlled by the strong vertical mixing ([Figure 3](#)). The wind components U and V are in a right-handed coordinate system with U parallel to the surface-layer wind and positive downwind. In the dry CBL, temperatures of the rising and sinking air parcels closely follow the dry-adiabatic lapse rate, 9.8 K km^{-1} .

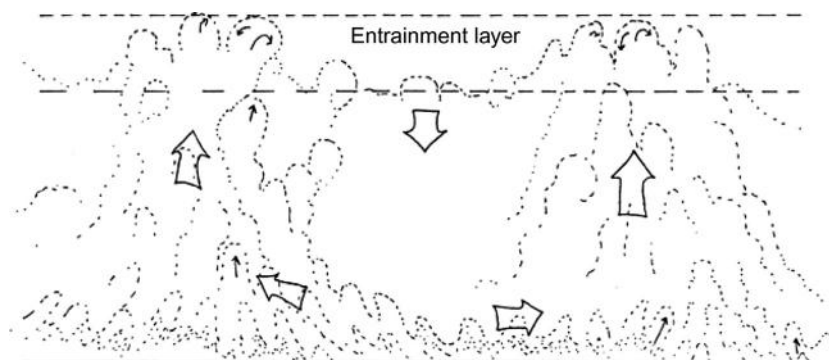


Figure 1 Idealized cross-section of the convective boundary layer. Eddies are outlined. Broad arrows denote flow in large eddies; small arrows denote flow in smaller eddies.

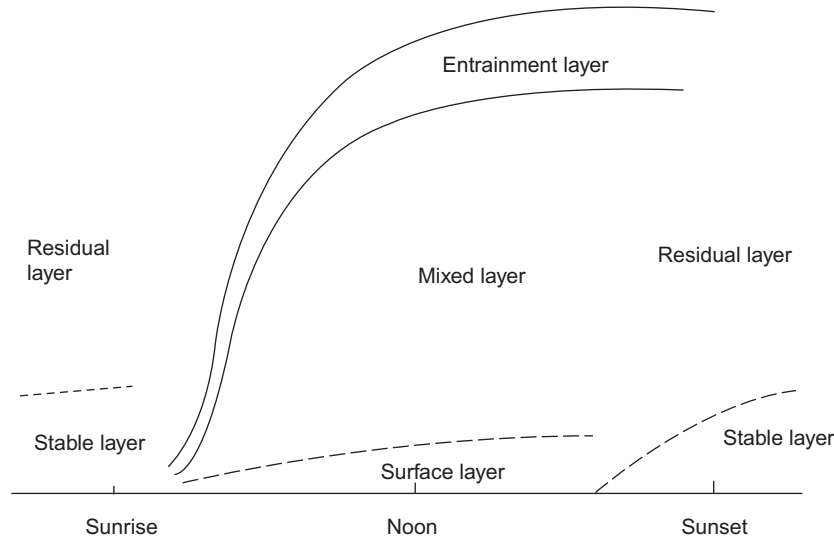


Figure 2 Diurnal variation of CBL over land.

Thus it is convenient to use potential temperature, which is conserved during dry-adiabatic processes, to describe the thermal stratification:

$$\theta = T \left(\frac{1000}{p} \right)^{R/c_p} \quad [1]$$

In eqn [1], which is derived from the first law of thermodynamics and the hydrostatic relation, p is pressure, T is temperature (K), R is the gas constant, and c_p is the specific heat at constant pressure for air. Similarly, the water-vapor quantities conserved in the absence of condensation or evaporation are used to define the mean state of the CBL, namely specific humidity ($q = \rho_w/\rho_a$, where ρ_w is the mass density of water

vapor and ρ_a is the mass density of air) or mixing ratio (ρ_w/ρ_d , where $\rho_d = \rho_a - \rho_w$ is the mass density of dry air).

The convective boundary layer can be divided into three parts (Figure 3), the mixed layer (heights z_s to h_1), where Θ , Q , U , and V are nearly constant, the surface layer (heights less than z_s , where Θ and Q decrease and U increases rapidly with height, and the entrainment or interfacial layer (heights h_1 to h_2), where all four variables change rapidly with height from their mixed-layer values to those of the atmosphere above. In this layer, the vertical changes in Θ , Q , U , and V reflect the fractional area at each height occupied by turbulent convective boundary layer air. The inversion height z_i is drawn roughly halfway through the entrainment layer, where vertical gradients

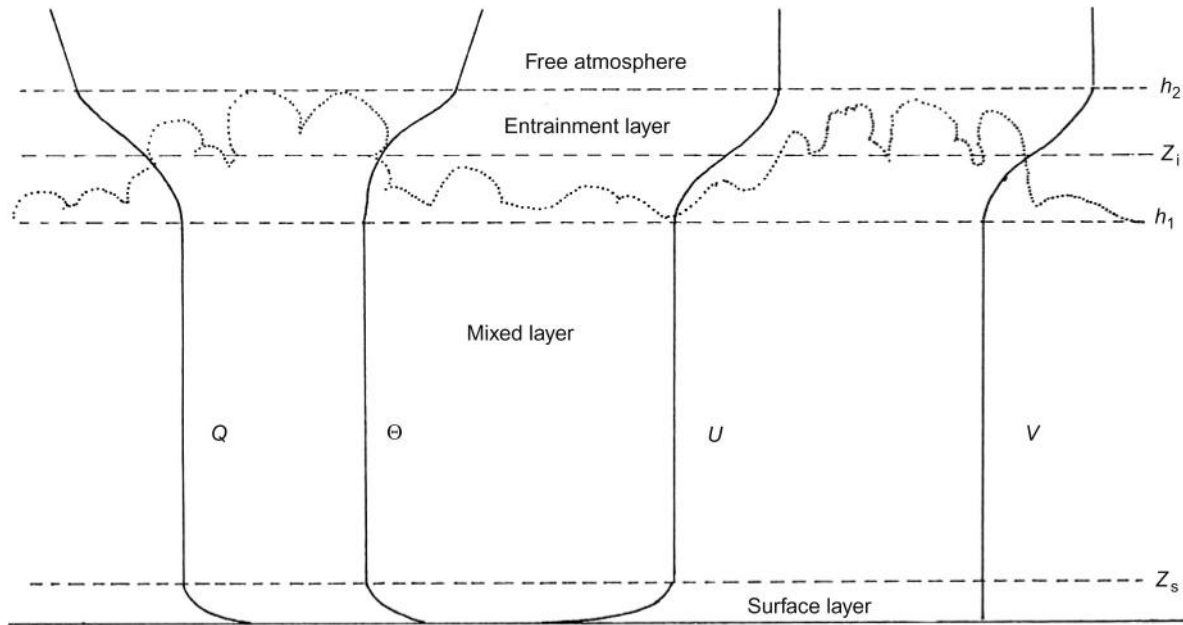


Figure 3 Idealized profiles of horizontally averaged potential temperature Θ , specific humidity Q , and wind components U and V in the convective boundary layer. Values increase to the right. $V \approx 0$ in the CBL.

are a maximum. Above the interfacial layer, the stratification is that of the free atmosphere.

If the water-vapor content is substantial, as is common over the tropical oceans, the stratification of the CBL is often defined in terms of the virtual potential temperature, θ_v , defined by:

$$\theta_v = \theta(1 + 0.61q) \quad [2]$$

since water vapor contributes significantly to buoyancy ($= (g/\Theta_v)(\theta_v - \Theta_v)$), which dominates CBL growth and mixing. In eqn [2], specific humidity is in units of g kg^{-1} .

Momentum is likewise well mixed through most of the CBL. The momentum is determined by a balance between the pressure gradient force $-(1/\rho_a)\nabla P$, the Coriolis force $f\mathbf{k} \times \mathbf{V}$, and surface friction Fr . At equilibrium

$$f\mathbf{k} \times \mathbf{V} = -\frac{1}{\rho_a}\nabla P + Fr \quad [3]$$

The wind vector $\mathbf{V}$ is parallel to the isobars to a close approximation above the CBL, where friction is negligible. Within the CBL, the Coriolis force $f\mathbf{k} \times \mathbf{V}$ is reduced as a result of frictional slowing of the wind, and the air flows down the pressure gradient toward lower pressure. In the northern hemisphere CBL, the mixed-layer wind U flows in a direction ~ 20 – 30 degrees to the left of the wind above the CBL. The rotation of the wind from the surface-layer direction to the free atmosphere direction is concentrated in the entrainment layer.

Vertical Fluxes in the CBL

The Mixed Layer

To describe the behavior of the turbulent air in the mixed layer, we separate out the fluctuations (denoted by primes) from the means, denoted by upper case. Thus, for specific humidity,

$$q' = q - Q$$

where Q is the average of q at a given height. The vertical flux of q , $\overline{w'q'}$ at a height z is given by the average of the product of the fluctuating vertical velocity w' and q' ,

$$\overline{w'q'}(z) = \frac{1}{N} \sum_{i=1}^N w'_i(z) q'_i(z) \quad [4]$$

where N is the number of observations. Positive values of $\overline{w'q'}$ indicate upward transport of water vapor (or downward transport of dry air).

In the mixed layer, the profiles of the vertical fluxes of θ_v , q , and u , are linear (Figure 4). The vertical q -flux is depicted decreasing with height, but $\overline{w'q'}$ can also increase with height. The decrease of $\overline{w'\theta'_v}$ with height to about -0.2 times the surface flux at z_i has been commonly observed. From eqn [2], the profile of θ -flux $\overline{w'\theta'}$ should also be linear:

$$\frac{\partial \overline{w'\theta'}}{\partial z} = (1 + 0.61Q)^{-1} \left(\frac{\partial \overline{w'\theta'_v}}{\partial z} - \Theta \frac{\partial \overline{w'q'}}{\partial z} \right) \quad [5]$$

Over the tropical oceans in fair weather, $\overline{w'q'}$ is large and $\overline{w'\theta'}$ is small or even negative through most of the CBL, particularly if the humidity flux remains large at the CBL top. This is illustrated in Figure 5, which is based on data from the Global Atmospheric Research Programme's (GARP) Atlantic Tropical Experiment (GATE). On Day 253 (bottom), the temperature flux is negative through half the CBL. When this profile is combined with the humidity flux profile using eqn [5], the result is a $\overline{w'\theta'_v}$ -profile close to the idealized profile shown in Figure 4. On Day 258 (top), the temperature flux is so small that the humidity flux accounts for most of the virtual-temperature flux, whose profile again resembles that in Figure 4.

The temperature and moisture fluxes are typically expressed in K m s^{-1} and $\text{g kg}^{-1} \text{m s}^{-1}$, as in Figure 5 or, when the energy balance is of interest, in terms of sensible heat flux

$$H = \rho_a c_p \overline{w'T'} \approx \frac{T}{\Theta} \rho_a c_p \overline{w'\theta'}$$

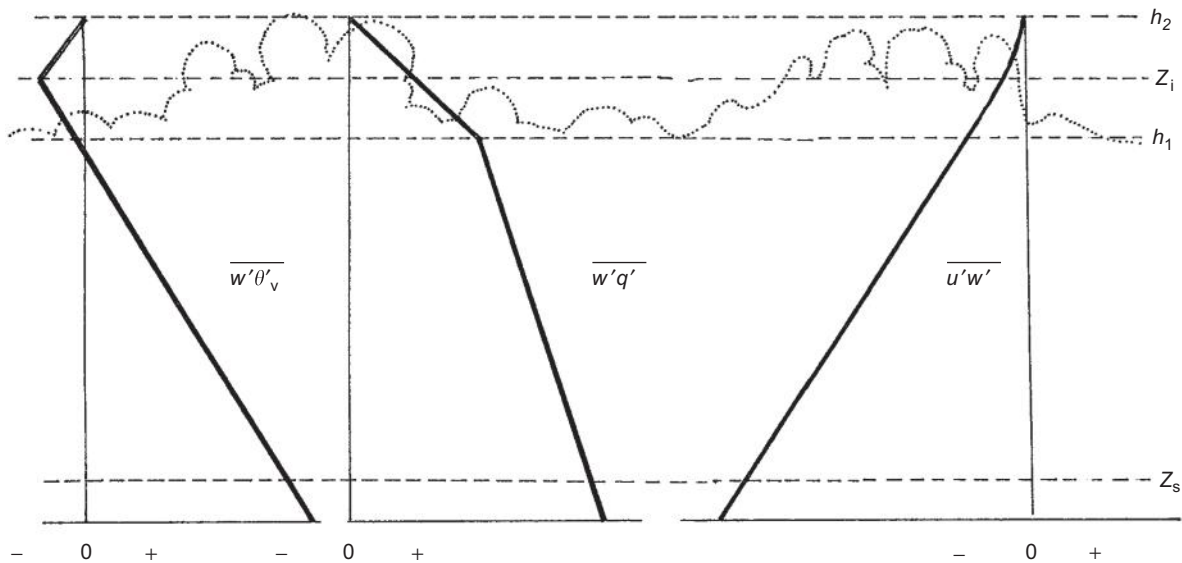


Figure 4 Idealized vertical profiles of the vertical flux of virtual-potential temperature flux $\overline{w'\theta'_v}$, specific humidity, $\overline{w'q'}$, and the u -momentum flux $\overline{u'w'}$ (parallel to the mean surface-layer wind).

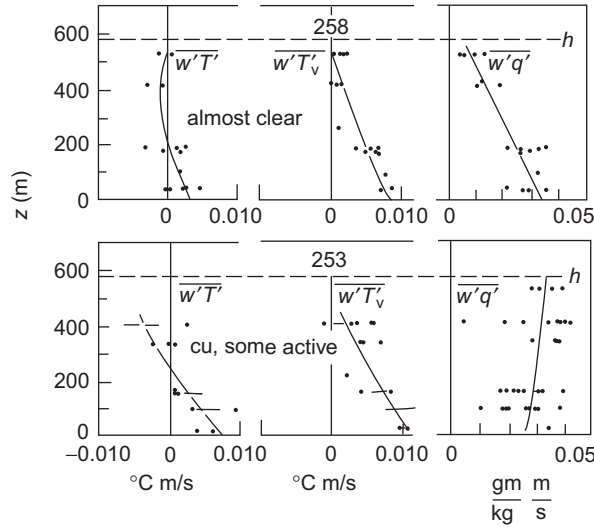


Figure 5 Vertical fluxes of T , T_v , and q , for two days in GATE with different q -flux and T -flux profiles. Note that the shape of the T_v -flux profile remains the same. Figure adapted from Nicholls and LeMone, 1980. *Journal of Atmospheric Science* 37: 2051–2067.

and latent heat flux $LE = \rho_a L_w \overline{w'q'}$ where L_w is the latent heat of vaporization. H and LE have units of $W m^{-2}$. The maximum values of $H + LE$ in fair weather rarely exceed the incoming solar radiation. A summer noontime mid-latitude value of $\sim 1000 W m^{-2}$ is equivalent to $\overline{w'T'} \sim 0.8 K m s^{-1}$ or $\overline{w'q'} \sim 0.3 g kg^{-1} m s^{-1}$ for $\rho_a = 1.27 kg m^{-3}$, $c_p \sim 1000 m^2 s^{-2} K^{-1}$ and $L_w = 2.5 \times 10^6 m^2 s^{-2}$. Typical fair-weather values of temperature fluxes over the ocean range from less than $0.01 K m s^{-1}$ over the tropical and subtropical oceans in fair weather to $0.20 K m s^{-1}$ for cold air flowing over warm water; over land at noontime values reach $0.4 K m s^{-1}$. Water vapor fluxes over the ocean lie in the range 0.015 – $0.05 g kg^{-1} m s^{-1}$ in the fair-weather tropical boundary layer to $0.25 g kg^{-1} m s^{-1}$ for cold air flowing over warm water; noontime values over land reach $0.25 g kg^{-1} m s^{-1}$.

The linear flux profiles for θ_v , q , u , and v can be shown to be consistent with their mean profiles through their tendency equations. The tendency equation for a quantity C takes the form:

$$\frac{\partial C}{\partial t} \approx -\frac{\partial}{\partial z} \overline{w'C'} - V \cdot \nabla C + S + \sum F_i \quad [6]$$

where $\partial C / \partial t$ is the local time tendency, $-(\partial / \partial z) \overline{w'C'}$ is the change in C due to vertical divergence of the vertical c -flux, $-V \cdot \nabla C$ is advection, S represents sources and sinks, and $\sum F_i$ represents forces (for $C =$ momentum). The horizontal divergence of horizontal fluxes is negligibly small in the fair-weather CBL and is neglected. For Θ_v and Q , $\sum F_i = 0$. If there is no horizontal advection or sources/sinks (approximately true with no clouds and if radiative effects are negligible)¹ then

$$\frac{\partial (\Theta_v, Q)}{\partial t} \approx -\frac{\partial}{\partial z} \overline{w'(\theta_v, q)'} \quad [6a]$$

¹ Typical maximum warming or cooling from radiation is of the order of $0.05 K hr^{-1}$ in the unpolluted CBL.

From eqn [6a], a linear flux profile is consistent with the constant mean profiles in Figure 3, since $\partial C / \partial t$ must be independent of height for C to remain independent of height.

Equations [3] and [6] can be combined for momentum to obtain, neglecting advection

$$\frac{\partial U}{\partial t} \approx -\frac{1}{\rho_a} \frac{\partial}{\partial z} \overline{\rho_a u'w'} + fV - \frac{1}{\rho_a} \frac{\partial P}{\partial x} \quad [6b]$$

In the absence of horizontal temperature gradients, the pressure gradient force $-(1/\rho_a)(\partial P / \partial x)$ is nearly constant with height. U and V are nearly constant in the mixed layer from Figure 3 and hence $\partial U / \partial t$ must be nearly constant with height. Thus $-(1/\rho_a)(\partial / \partial z)(\overline{\rho_a u'w'})$ is constant with height, and the profile of $u'w'$ is nearly linear. Figure 6 shows $u'w'$ and $v'w'$ as a function of height based on a Large Eddy Simulation (LES) and observations (for u only) for constant $(1/\rho_a)\nabla P$. Surface values of $u'w'$ are of the order of $-1 \times 10^{-6} U^2 m^2 s^{-1}$, with higher values over rougher surfaces. The vertical flux of u -momentum remains small.

Comparison of Figure 3 and Figure 4 shows that vertical fluxes are large even where the mean vertical gradient is zero. This results from the mixing elements extending through the CBL (Figure 1). For example, updrafts ($w' > 0$) carry low values of u ($u' < 0$) upward from near the surface, and downdrafts ($w' < 0$) carry high values ($u' > 0$) from near the top of the CBL, resulting in the observed negative vertical flux of u -momentum. When vertical fluxes are determined by vertical gradients over a deep layer rather than locally, we call this nonlocal transport or nonlocal mixing.

The Surface Layer

Vertical gradients of Θ , Q , U , and V are substantial in the surface layer. The behavior of the surface layer has been

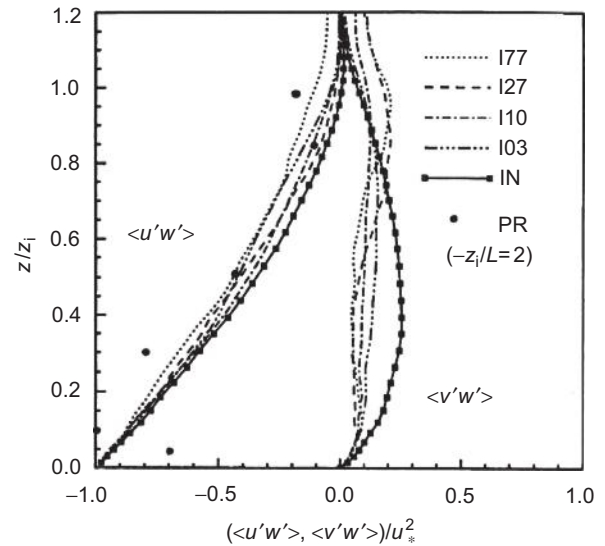


Figure 6 LES vertical fluxes of u and v for $-(1/\rho)\nabla P$ constant with height and CBLs with $-z_i/L$ ranging from 7.7 (I77) to 0.3 (I03). IN represents the neutral boundary layer. Circles are data from north of Puerto Rico ($-z_i/L \sim 2$). Figure adapted from Brown and Grant, 1997. *BLM*, 84: 1–22; with data from Pennell and LeMone, 1974. *Journal of Atmospheric Science* 31: 1308–1323.

successfully described in terms of similarity theory, for which the relevant physical parameters are identified and then combined in a dimensionally consistent way. For example, under neutral stability in the atmosphere, where viscosity is negligible, it can be assumed that the vertical shear of the mean wind is a function related to distance from the ground and surface stress $u_*^2 = ((u'w')_0^2 + (v'w')_0^2)^{1/2} \approx (u'w')_0$ for u defined as parallel to the mean surface-layer wind. That is,

$$\frac{\partial U}{\partial z} \propto \frac{u_*}{z} \quad [7]$$

where u_* is called the friction velocity. Hence $U \sim \ln(z/z_0)$, where z_0 is the height at which $U = 0$. Typical values of z_0 over open land are of the order of 10^{-2} m.

Monin–Obukhov similarity theory includes the effects of buoyancy as well as stress. Thus the relevant parameters are u_* , g/T_v , and virtual temperature flux, $\overline{w'T'_v}$. The scaling height in the surface layer is the Monin–Obukhov scaling length, L , given by:

$$L = -\frac{u_*^3}{(kg/T_v)\overline{w'T'_v}} \quad [8]$$

where g is the acceleration of gravity, and k is the von Karman constant (about 0.4). Within the surface layer, u_* and $\overline{w'T'_v}$ are assumed constant. The depth of the surface layer is typically of the order of $0.1z_i$.

From dimensional reasoning and numerous field measurements, the profiles of a quantity $c = \theta, q$, or u in the surface layer are represented in the form:

$$C = C_0 + c_* \left[\ln \frac{z}{z_c} + \Psi_c \left(\frac{z}{L} \right) \right] \quad [9]$$

where $\Psi_c(z/L)$ is a stability correction that increases as surface buoyancy flux increases ($-L$ decreases), and $c_* \equiv \overline{w'c'}/u_*$. The height z_c is defined so that $C(z_c) = C_0$, the value of C just above the surface. Typical values of z_c are of the order of a few centimeters or smaller.

The Entrainment Layer

The entrainment layer is occupied by both boundary layer and free-atmosphere air, and the virtual temperature flux is negative. The varying CBL top is not a material surface: the more-turbulent air tends to engulf the air above, incorporating it into the CBL. Downward motion at the edge of thermals can draw free-atmosphere air downward into the CBL, where it is mixed in. Strong shear at the top of the CBL can generate shear-buoyancy waves which can break and mix the air at the interface. This mixed air, being less buoyant than free-atmosphere air, can then be more easily mixed downward into the CBL.

The growth of the CBL is determined by how fast new air is engulfed into the CBL, itself a function of the energy of the engulfing eddies, the resistance of the free-atmosphere air to engulfment (a function of the virtual temperature difference ΔT_v between the CBL and the free atmosphere), and the mean largescale vertical motion at the CBL top. For simplicity, we collapse the entrainment layer to zero thickness so that $h_1 = h_2 = z_i$. For this simple case,

$$\frac{\partial z_i}{\partial t} = -\frac{(\overline{w'T'_v})_{z_i}}{\Delta T_v} + W_{z_i} \quad [10]$$

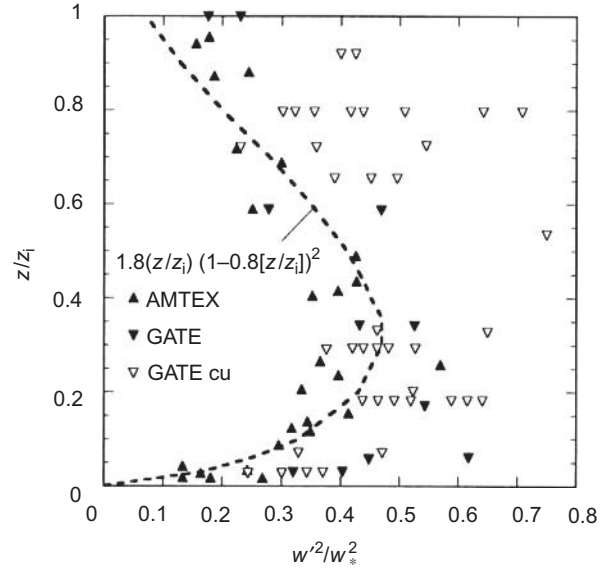


Figure 7 Vertical velocity variance $\overline{w'^2}$, normalized by CBL velocity scale w_*^2 , from GATE and AMTEX aircraft data. Best-fit quadratic from Lenschow, et al., 1980.

In eqn [10], $-(\overline{w'T'_v})_{z_i}/\Delta T_v$ is the entrainment velocity (w_e), or rate of growth of the CBL in the absence of subsidence (W_{z_i}). $(\overline{w'T'_v})_{z_i}$ varies with the energy expended in entrainment; ΔT_v is related to the stratification of the air above the CBL. It is commonly assumed that

$$(\overline{w'T'_v})_{z_i} \approx -0.2(\overline{w'T'_v})_0 \quad [11]$$

and this is widely supported by observations.² However, the coefficient might be much larger than -0.2 if there is large vertical shear of the horizontal wind at the CBL top.

Turbulence Structure in the CBL

In his 1972 paper describing the first large eddy simulation of the convective boundary layer, Deardorff found that the proper scaling parameters for the CBL are

$$w_* \equiv \left(\frac{g}{T_v} (\overline{w'T'_v})_0 z_i \right)^{1/3} \quad [12a]$$

$$\theta_v^* \equiv \frac{(\overline{w'T'_v})_0}{w_*} \quad [12b]$$

where z_i is defined in Figure 3.

The success of the Deardorff scaling velocity w_* is illustrated in Figure 7, which combines data from field programs in very different environments, GATE and the Air Mass Transformation Experiment (AMTEX). For AMTEX (6 days), w_* averaged 1.72 m s^{-1} , with a range from 1.06 to 2.51 m s^{-1} ; for GATE (4 days of data), w_* averaged 0.53 m s^{-1} , with a range from 0.50 to

² It is important that virtual temperature be used here. If the humidity flux is substantial, $-(\overline{w'T'_v})_{z_i}$ is typically much larger than $0.2(\overline{w'T'_v})_0$, as shown in Figure 5.

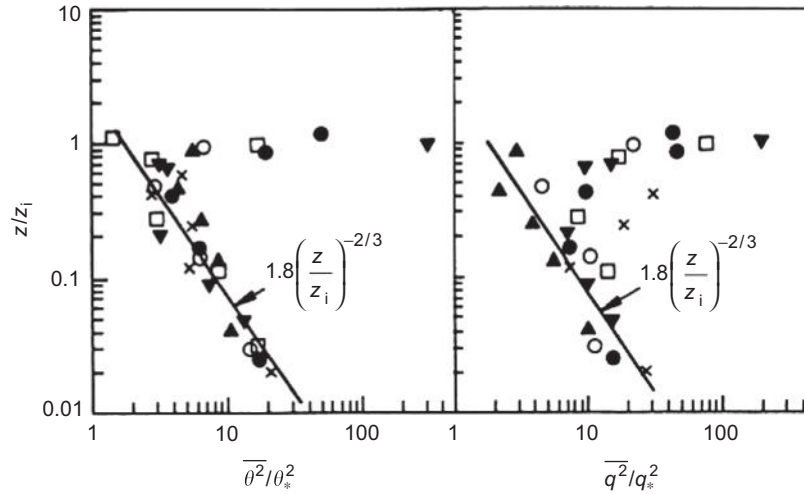


Figure 8 Profiles of $\overline{q'^2}/q_*^2$, $\overline{\theta'^2}/\theta_*^2$, from AMTEX aircraft data. Adapted from Lenschow, et al., 1980. Journal of Atmospheric Science 37: 1313–1326.

0.59 m s^{-1} . This amounts to a factor of 25 range in normalizing factors. The outlined inverted triangles are for GATE days with enhancement of vertical velocities originating from release of latent heat in active cumulus clouds. Normalization with w_* does not account for this process, so it is not surprising that these points are associated mostly with higher vertical velocities than the normalized curve would predict.

Similarly, $\overline{q'^2}/q_*^2$, where $q_* = \overline{w'q'}/w_*$, and $\overline{\theta'^2}/\theta_*^2$ both collapse onto a curve whose shape is based on local free convection similarity arguments, for which z , z_i , g/T_v , and $\overline{w'T'_v}$ are the relevant variables, in the lower CBL (Figure 8), at least for $z = z_i < 0.1$, but not for the upper CBL, especially for $\overline{q'^2}/q_*^2$.

The deviations of the normalized q - and θ -variances from the curves in Figure 8 are related to entrainment processes at the top of the CBL. The deviations can be accounted for if one assumes a separation of the corresponding fluxes into 'top-down' fluxes associated with entrainment processes, which decrease linearly from a maximum value at z_i to zero at the surface, and 'bottom-up' fluxes, which decrease linearly from the surface flux to zero at z_i . The variance of a scalar c can be written:

$$\overline{c'^2} = \overline{c_t'^2} + 2\overline{c_t'c_b'} + \overline{c_b'^2} \quad [13]$$

where the c_t -fluctuations are associated with 'top-down' processes, and c_b -fluctuations with 'bottom-up' processes. Each component can be normalized using the corresponding flux:

$$f_t = \overline{c_t'^2} \left[\frac{w_*}{(\overline{w'c_t'})_1} \right]^2 \quad [14a]$$

$$f_b = \overline{c_b'^2} \left[\frac{w_*}{(\overline{w'c_b'})_0} \right]^2 \quad [14b]$$

$$f_{tb} = \overline{c_t'c_b'} \left[\frac{w_*^2}{(\overline{w'c_t'})_1(\overline{w'c_b'})_0} \right] \quad [14c]$$

where $(\overline{w'c_t'})_1$ is the top-down scalar flux at the CBL top, $(\overline{w'c_b'})_0$ is the bottom-up scale flux at the surface, and f_t , f_b , and f_{tb} are dimensionless functions that can be determined using a large

eddy simulation (LES). From eqns [13] and [14], the nondimensional variance of θ can be written:

$$\frac{\overline{\theta'^2}}{\theta_*^2} = f_b + 2Rf_{tb} + R^2f_t \quad [15]$$

where $R = \overline{w'\theta'_1}/\overline{w'\theta'_0}$ accounts for the top-down processes. The resulting function is in better agreement with AMTEX data (Figure 9).

Large Eddy Structure in the CBL

Figure 1 shows a CBL inhabited by a spectrum of eddies ranging in size from tiny plumes near the surface to large eddies

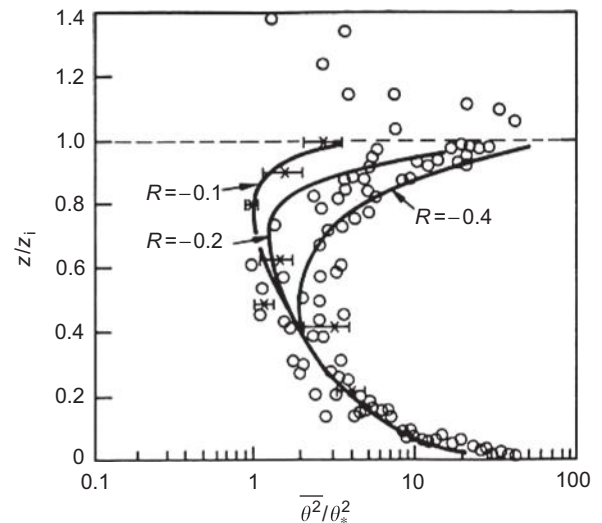


Figure 9 Profile of $\overline{\theta'^2}/\theta_*^2$ from AMTEX aircraft data. LES-derived curves accounting for top-down and bottom-up processes are drawn for a plausible range of $R = \overline{w'\theta'_1}/\overline{w'\theta'_0}$ based on AMTEX data. From Wyngaard and Moeng, 1989. Journal of Atmospheric Science 14: 2313–2330.

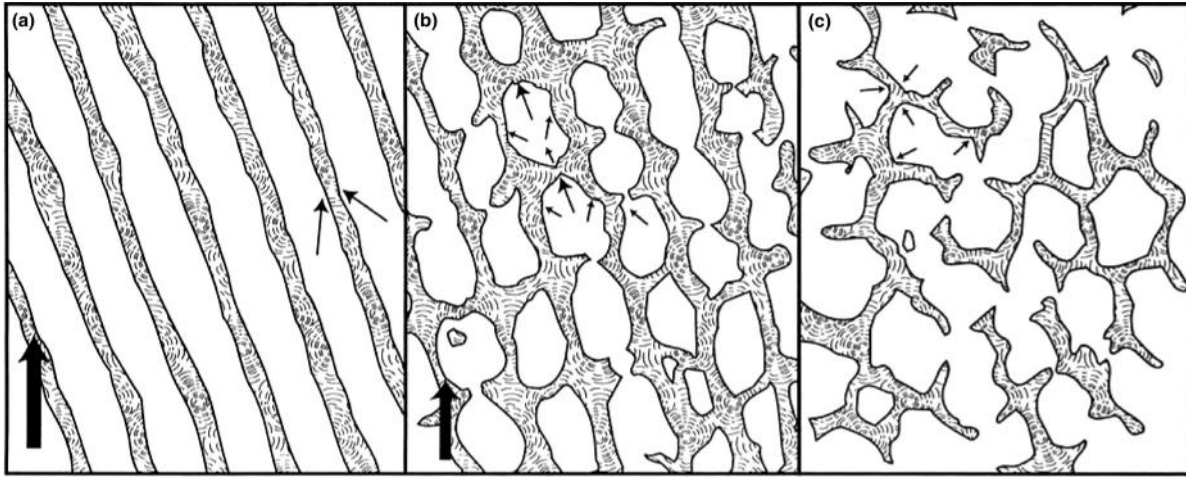


Figure 10 Horizontal variation of CBL large-eddy structure with $-z_i/L$: (a) $-z_i/L$ near 1.5, (b) $-z_i/L$ near 10, (c) large $-z_i/L$ (near-calm winds). Shading indicates upwelling regions at a level in the upper half of the CBL; air sinks between. Small arrows illustrate near-surface flow converging into upwelling regions; flow diverges at CBL top. For rolls, each upwelling region is fed by two counter-rotating vortices. The large arrows represent the wind vector above the CBL.

spanning the CBL, with horizontal scale (updraft+downdraft) of the order of 1.5–3 times the CBL depth. Smaller eddies and turbulence are concentrated in the large-eddy updrafts.

The variation of large-eddy structure with $-z_i/L$ is illustrated in **Figure 10**. Roll vortices occur in the CBL for $-z_i/L$ between some small but unknown positive value and ~ 25 ,³ with more cellular or random convection at higher values. The rolls (**Figure 10(a)**) are horizontal counter-rotating helical circulations that extend through the CBL. They are oriented close to parallel to the mean CBL wind, or between 15 and 30 degrees to the left of the geostrophic wind (or wind above the CBL) in the middle latitudes of the northern hemisphere. Lines of cumulus clouds, or ‘cloud streets’, sometimes lie above their upwelling regions. The small angle between rolls and the mean CBL wind means that it can take a half-hour, an hour, or even longer for a roll to pass over a given point on the surface. As $-z_i/L$ increases, the rolls become more three-dimensional, looking more like lines of cells (**Figure 10(b)**). The relatively slow advection time for rolls makes it possible to separate them from the other large eddies using instruments fixed to the ground, even though the large eddies may have horizontal scales comparable to that of rolls in the cross-wind direction. The linear structures disappear at high $-z_i/L$, and convection becomes more cellular or random (**Figure 10(c)**).

The dependence of large-eddy structure on $-z_i/L$ has been revealed by observational studies and LES. The higher ratios are associated with observational studies; LESs give upper limits around 1.5–2.6. This is partially a result of the three-dimensionality of rolls at higher $-z_i/L$. With limited horizontal resolution and distortions resulting from periodic boundary conditions, such rolls would be hard to identify in an LES. Further, the definition of ‘roll’ becomes ambiguous as they

become more three-dimensional, even in observations. Finally, rolls formed over land in the early morning (small $-z_i/L$) often persist as $-z_i/L$ gets large.

Departures from the Idealized CBL

The CBL profiles presented apply strictly to a quasisteady-state boundary layer in horizontally homogeneous conditions with negligible influence from clouds. Clouds were present when AMTEX (stratocumulus) and GATE (cumulus) measurements shown in **Figure 5** were taken, but their thickness or coverage was too small to produce a measurable effect (the open triangles representing the exceptions). Over land, strong diurnal variation restricts ‘quasi-steady-state’ CBLs to roughly 4–5 h starting around solar noon. Further, terrain, surface vegetation, and soil moisture are rarely homogeneous, leading to uneven heating, evaporation, and stress. The marine CBL in fair weather is close to steady state; winds of a few m s^{-1} can mix the upper levels of the ocean sufficiently that the diurnal variation of temperature at the surface is less than 1 K, with correspondingly small changes in the CBL. Even in near-calm fair weather conditions, the diurnal variation of sea surface temperature rarely exceeds 2.5 K. Cloudless marine boundary layers are rare, but the subcloud layer follows CBL scaling in marine CBLs topped with small cumulus and relatively weak inversions. Marine CBLs with strong surface heating and some shallow stratocumulus whose vertical growth is halted by a strong inversion scales with z_i which also corresponds to cloud top. These two types of CBL remain sufficiently steady over several hours that they can be well sampled. Thus it is not surprising that AMTEX (some stratocumulus) and GATE (clear to small cumulus) data are commonly used to test CBL similarity theories and parametrization schemes.

Because nonideal fair weather CBLs are so common, some of the features to expect are summarized.

³ The ratio $u_*'/w_*' = (-0.4L/z_i)^{1/3}$ is also used to separate rolls from more three-dimensional CBL convection.

Baroclinic CBLs

The wind profiles in the quasi-steady-state baroclinic CBL remain constant within the mixed layer. From eqn [6b], a linearly varying horizontal pressure gradient with height will be reflected in quadratic vertical profiles of $\overline{\rho_a u' w'}$ and $\overline{\rho_a v' w'}$.

Rapidly Entraining CBLs

Such CBLs occur during the morning in fair weather, or when strong surface heating is combined with a weak temperature inversion at the CBL top. Mean profiles of specific humidity and wind vary with height in the mixed layer under such conditions. However, the profile of Θ (or Θ_v in humid conditions) remains nearly constant.

CBLs with Nonprecipitating Cumulus

Once clouds become deep enough, latent heating or interaction of the clouds with vertical shear will develop vertical pressure forces that enhance the exchange of air across the cloud–subcloud layer interface. This increases $\overline{w'^2}$ at mixed-layer top as in Figure 7; the resulting vertical mass exchange across the cloud–subcloud layer interface also produces vertical variation in Q , U , or V and enhances fluxes and variances, especially in the upper CBL. Surprisingly, GATE data showed that the shape of the virtual-temperature flux profile did not change in the presence of deeper cumulus clouds. Deeper clouds can also ‘break up’ roll vortex circulations.

CBLs Over Nonhomogeneous Surfaces

Moistened surfaces with growing vegetation produce more vapor flux and less heat flux than dry surfaces with dormant vegetation or concrete. The fluxes may ‘blend’ or lose their clear relationship to individual surface features at a certain height if land use variability takes place at small scales. However,

larger-scale nonuniformities (>10 km) can produce mesoscale variability, with more vigorous turbulence and higher z_i over warmer surfaces, and less vigorous turbulence and lower z_i over cooler surfaces. Mesoscale circulations with upwelling regions over warmer surfaces may also occur in light winds, but these are difficult to document. Elevated terrain and slopes facing the sun are favored places for updrafts in light winds. Flow is deflected over and around hills, locally producing stronger than ambient winds and turbulence.

See also: Agricultural Meteorology and Climatology. **Air Sea Interactions:** Momentum, Heat, and Vapor Fluxes. **Boundary Layer (Atmospheric) and Air Pollution:** Coherent Structures; Modeling and Parameterization; Overview; Stably Stratified Boundary Layer; Surface Layer. **Clouds and Fog:** Classification of Clouds. **Dynamical Meteorology:** Overview. **Global Change:** Climate Record: Surface Temperature Trends. **Numerical Models:** Convective Storm Modeling; Large-Eddy Simulation; Parameterization of Physical Processes: Turbulence and Mixing. **Thermodynamics:** Humidity Variables. **Turbulence and Mixing:** Turbulent Diffusion.

Further Reading

- Deardorff, J.W., 1972. Numerical investigation of neutral and unstable planetary boundary layers. *Journal of Atmospheric Science* 29, 91–115.
- Garratt, J.R., 1992. *The Atmospheric Boundary Layer*. Cambridge University Press, Cambridge.
- Haugen, D.A. (Ed.), 1973. *Workshop on Micrometeorology*. American Meteorological Society, Boston, MA.
- Kaimal, J.C., Finnigan, J.J., 1994. *Atmospheric Boundary Layer Flows: Their Structure and Measurement*. Oxford University Press, Oxford.
- Stull, R.B., 1988. *An Introduction to Boundary Layer Meteorology*. Kluwer, Dordrecht.
- Wyngaard, J.C. (Ed.), 1980. *Workshop on the Planetary Boundary Layer*. American Meteorological Society, Boston, MA.

Microclimate

MW Rotach, University of Innsbruck, Innsbruck, Austria
P Calanca, Agroscope Reckenholz-Taenikon, Zurich, Switzerland

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

Microclimatology is concerned with the study of the processes by which the local surface properties affect the lowest layer of the atmosphere. Often the latter refers to the so-called roughness sublayer, but in some cases attention is more properly limited to the so-called canopy layer or then, conversely, the whole surface layer. Microclimates are described in terms of climatic variables, their temporal and vertical variability, as established by the balance equations that govern the exchange of radiation, heat, water, and other atmospheric constituents.

Introduction

The word 'climate' stems from the Greek *klíma* (slope, zone), with its roots in *klínein* (to slope), and originally denoted 'a zone of equal latitude,' thus referring to the effects of latitude on the availability of solar energy. In fact, on a global scale the climate is above all determined by the latitudinal and seasonal distribution of incoming solar radiation as given by Earth's orbital parameters.

On very small scales, however, the climate is primarily shaped by the local surface properties that control the exchange of radiation, energy, momentum, and water between the ground and the atmosphere. The *microclimate* of a particular location can hence be defined as the collection of statistics describing the thermal and dynamical conditions prevailing in the atmospheric layer directly affected by the underlying surface. Accordingly, 'descriptive microclimatology,' can be identified as the study of the long-term average and typical variability of climate variables in the lowest layer of the atmosphere, while 'physical microclimatology' can be defined as the study of the processes by which the lowest layer of the atmosphere responds to surface boundary conditions.

As is customary in climatology, the various aspects that concur in creating the microclimate of a particular location are identified with a number of so-called climate variables. Common climate variables include radiation, temperature, humidity, wind speed, and pressure (density), but, depending on the research focus, other variables need to be taken into account. For example, the health and comfort of the ever-increasing number of people living in cities are directly related to the concentration and distribution of air pollutants, which are therefore required to characterize the microclimate of these particular environments.

The surface properties determining a microclimate are seldom constant in time. Long-term variations may arise naturally or as a consequence of human-induced changes in land use. At the seasonal scale, variations in surface properties may be brought about by the presence of a snow cover or through the life cycle of plants. Short-term variations may, for example, be caused by the dynamic effects of the wind on the structure of the surface elements. Ultimately, it is this variability in surface properties that, along with the variability of the large-scale atmospheric forcing, is responsible for the

characteristic temporal patterns observed in microclimatic records.

The Lowest Layers of the Atmosphere

The atmospheric layer affected by the local surface properties is called the 'planetary boundary layer' (PBL). It has a characteristic depth (z_i) on the order of 1000 m and can be divided into an upper or outer layer (the uppermost 90%) and an inner or surface layer (SL) (the lowest 10%) (Figure 1). However, a *direct* influence of the surface characteristics on the atmospheric state is observed only in the *lowest part* of the SL, in the immediate vicinity of the roughness elements. If this layer of influence has any discernible thickness, it is because of a non-negligible vertical extension of the 'roughness elements' (stones, vegetation, tress, and buildings). Therefore, this layer is usually called the 'roughness sublayer.' The upper part of the SL is then referred to as the 'inertial sublayer.' Over a relatively smooth surface such as short grass or sand, the roughness sublayer becomes very thin and the inertial sublayer is often associated with the entire SL (Figure 1).

Based on these considerations, 'microclimatology' can also be defined as the study of the climatic state of the roughness sublayer or specific entities therein, even though the notion is most often extended to the examination of the entire SL or even PBL. Micrometeorology is then the study of the dynamics and thermodynamics of the SL or PBL.

The Roughness Sublayer

In the roughness sublayer, the flow is affected by the individual roughness elements and hence is fully three-dimensional in nature. The upper boundary of the roughness sublayer, z_r , is the level at which the horizontal variability associated with the roughness elements vanishes and the flow properties become horizontally homogeneous. Properly scaled 'profiles' of either mean flow characteristics or turbulence statistics will then merge to one curve whose shape is characteristic for the underlying surface. The depth of the roughness sublayer depends on the height and distribution of the roughness elements. For most surfaces, $2z_h < z_r < 5z_h$ covers the range of estimates, where z_h is the *average* height of roughness elements.

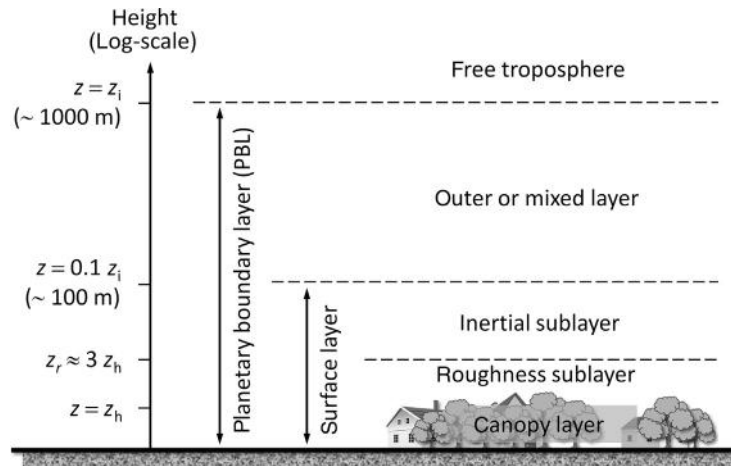


Figure 1 Conceptual sketch and terminology for the lowest layers of the atmosphere over a rough surface. Note the logarithmic height scale. Level z_i refers to the boundary layer height, z_r to the height of the roughness sublayer, and z_h to the (average) height of roughness elements.

The Canopy Layer

The lowest part of the roughness sublayer is the canopy layer. It ranges from the surface to z_h . In this layer, form (or pressure) drag and viscous drag on any individual roughness element are both significant and lead to a retardation of the flow. In addition, the material and orientation of the obstacles give rise to large variations in the energy balance, primarily through variations in the radiation balance, and also as a result of the distribution of sources and sinks of sensible heat, water vapor, or trace gases.

Figure 2 shows the height ranges covered by the roughness sublayer and its lower part, the canopy layer, in dimensionless

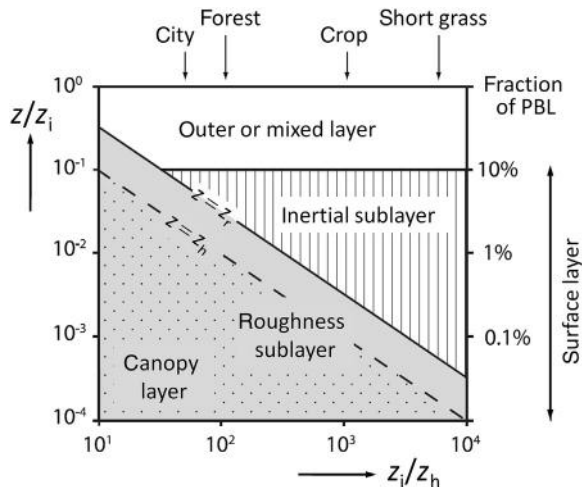


Figure 2 Sketch of the vertical extension of the various layers over rough surfaces and their variation with the nondimensional quantities z/z_i and z_i/z_h . Here, z denotes the (physical) height, z_i refers to the boundary layer height, and z_h is the (average) height of roughness elements. A value of $z_r/z_h = 3$ is assumed for the height of the roughness sublayer. The arrows labeled with 'city,' 'forest,' 'crop,' and 'short grass' are based on typical values for the height of the roughness elements z_h and the boundary layer height z_i . Redrawn from Rotach, M.W., 1999. On the urban roughness sublayer. *Atmospheric Environment* 33, 4001–4008.

form. Arrows indicate how surfaces covered with short grass, crops, trees, and houses, respectively, fit into this scheme. It can be seen that the vertical range, to which the study of microclimate is confined, can extend up to several tens of meters or more in the case of an urban surface. However, in the case of a shallow boundary layer (small z_i and large roughness elements), no inertial sublayer may be present at all.

Internal Boundary Layers

The above considerations are valid for horizontally homogenous surfaces. In the presence of a pronounced small-scale variability of surface properties, internal boundary layers may form downwind of each major change in the surface characteristics. Internal boundary layers may be 'thermal' if they are primarily prompted by an abrupt change in surface temperature, as for example across a shoreline. Alternatively, if changes are primarily in the surface roughness, internal boundary layers are called 'mechanical.' Most common is, of course, the development of combined thermal and mechanical internal boundary layers.

Internal boundary layers increase in depth with distance downstream from the surface discontinuities causing their development. Within a developing internal boundary layer, profiles of climate variables exhibit a continuous transition from conditions adjusted to the new surface characteristics close to the ground, to conditions still nearly in equilibrium with the conditions upstream of the leading edge. At still greater heights, horizontal gradients associated with surface heterogeneities become negligibly small due to the effects of turbulent mixing. The concept of a blending height has been introduced to denote the level at which the flow becomes independent of the local surface condition. In its strictest sense, the study of microclimate over heterogeneous surfaces is then confined to the airspace below the blending height.

Surface Characteristics

The surface characteristics that determine the microclimate can be organized according to which aspect of the

atmospheric dynamics and thermodynamics they affect for the most part:

- ‘Radiative properties’ determine the radiation budget. These are the albedo (α), or surface reflectivity in the shortwave band of the radiation spectrum; the emissivity of the effective surface (ϵ); and the geometric properties that influence the receipt and loss of radiation, such as the sky-view factor in the case of an urban surface or the canopy extinction coefficient (κ_e) in the case of a vegetated surface.
- ‘Aerodynamic properties’ determine the momentum budget. These are the surface roughness length (z_0), the zero-plane displacement (d), and the drag coefficient (C_d).
- ‘Thermal properties’ determine the heat flow in the underlying substrate–roughness elements. These are the thermal conductivity (κ_g), the heat capacity (c_g), and the thermal diffusivity ($\lambda_g = \kappa_g/c_g$). In soils, all of these properties strongly depend on the soil water content (SWC).
- ‘Soil hydraulic properties and vegetation characteristics’ affect soil moisture, the availability of water for evaporation and transpiration, and hence the partitioning between the turbulent energy fluxes and the conductive ground heat flux in the surface energy balance.
- ‘Vegetation properties’ control the distribution of radiation and the transfer of water vapor and trace gases within a canopy. In many modeling studies, these are primarily the Leaf Area Index (LAI) and stomatal conductance (γ_{st}), but they should more properly encompass all botanical, physiological, and geometric characteristics as well as species composition.
- ‘Water properties’ control the thermal state and the phase transitions in water bodies, in particular during the seasonal development of snow covers and ice bodies at high latitudes and elevated altitudes. These include, among others, the volumetric heat capacities of water (c_w) and ice (c_i), and the latent heat of fusion (L_f) and vaporization (L_v).

Although, in general, all properties concur in determining the local microclimate, their relative importance may vary

depending on the specific circumstances. Table 1 provides examples of parameters that are particularly significant in relation to specific surface types and situations. In many instances, the effective surface cannot be treated as static but undergoes dynamic changes. The examples in Table 1 emphasize the dynamic effects of the wind, either in the case of bending roughness elements (vegetated surfaces) or in the presence of water (waves) and snow (drift) surfaces.

In complex terrain, the topographic characteristics that control the development of ‘mesoscale circulation systems,’ such as land and sea breezes or mountain and valley winds, need also to be taken into account to understand the nature of microclimates.

The Budget Equations

The main role of the surface properties discussed in this article is to regulate the exchange of radiation, heat, and mass (water and carbon or other trace gases) between the ground and the atmosphere, establishing corresponding balances defined by the following budget equations:

- Radiation:

$$Q^* = K\downarrow - K\uparrow + L\downarrow - L\uparrow \quad [1]$$

- Energy:

$$Q^* = Q_E + Q_H + Q_G + \Delta Q_S \quad [2]$$

- Water (mass):

$$P = E + R + I_n + \Delta W \quad [3]$$

In these equations, Q^* denotes the net (all-wave) radiation flux at the surface; $K\downarrow$ and $K\uparrow$ are the incoming and

Table 1 Relevant microclimatological parameters for broad categories of characteristic surfaces and for selected aspects of the establishment of local microclimates

Surface type	Relevant parameters	Particular aspects
Bare soil	α , ϵ , z_0 , c_g (SWC), κ_g (SWC)	Usually considered the simplest case of surfaces, providing in many cases a reference; soil thermal properties depend on SWC, which may also affect the albedo.
Short vegetation (crops)	α , ϵ , z_0 , d , LAI, γ_{st} , κ_e	Flexible roughness elements; zero-plane displacement is relatively small and in many cases negligible; in crops there is pronounced seasonal development, with the growth stage determining the LAI and eventually the exchange of radiation, energy, and water.
Tall vegetation (forest)	α , ϵ , z_0 , d , LAI, γ_{st} , κ_e	Strong dynamic interactions with the atmospheric flow; often nonuniform vertical distribution of leaf density; importance of the storage terms in the balance equations.
Water	α , z_0 , C_d	Albedo is a function of solar elevation (diurnal variations) and momentum exchange, as controlled by z_0 or C_d ; strongly dependent on wind regime and resulting wave height (Charnock’s roughness length model).
Snow and ice	α , ϵ , z_0 , c_w , c_i , L_f , L_v	Albedo depends on snow or ice age (seasonal variability) and is a function of solar elevation (diurnal variability); especially with low snow densities, the surface roughness of snow covers depends on drift conditions; melting leads to well-defined surface temperature.
Urban	α , ϵ , z_0 , d , C_d , SVF	Stiff roughness elements; strong influence of the thermal properties of building material; geometry and distribution of roughness elements are of paramount importance for radiation, energy, and momentum exchange.

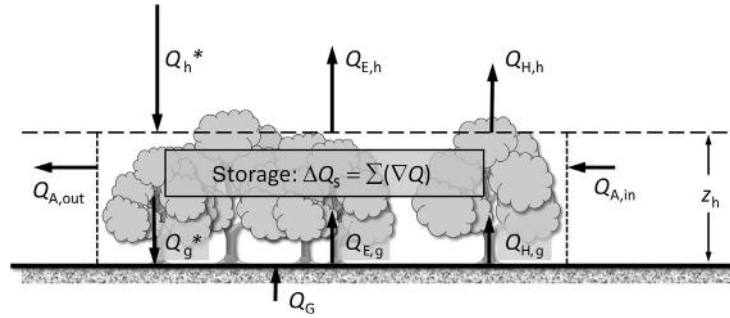


Figure 3 Energy fluxes appearing in the energy budget equation for a layer of finite thickness. Subscripts g and h refer to the ground level and the average canopy height z_h , respectively. Q_A denotes the contribution from horizontal advection to energy storage (ΔQ_s) in the volume. The latter is given by the divergence of all vertical and horizontal fluxes.

outgoing shortwave radiation fluxes, respectively; $L\downarrow$ and $L\uparrow$ are the incoming and outgoing longwave radiation fluxes, respectively; Q_E and Q_H are the turbulent fluxes of latent and sensible heat, respectively; Q_G is the conductive heat flux to the substrate; ΔQ_s is the change in heat storage associated with the divergence of horizontal or vertical energy fluxes that arise from the finite size of the volume under consideration; P is the water input through precipitation; E is the loss of water vapor through evapotranspiration; R is the surface runoff; I_n is the vertical infiltration into the ground; and ΔW is the change in water storage. Moreover, on the right-hand side of the energy and water budget equations, fluxes directed away (toward) the surface are considered positive (negative).

In the presence of snow and ice, transitions between the solid and liquid phases (melting and freezing) significantly contribute to both the energy and mass budgets. In urban environments, anthropogenic heat release also adds another contribution to the energy budget. In the case of vegetated surfaces, the net photosynthetic heat uptake or release appears as an additional term in the energy budget equation. Furthermore, in this latter case, the water budget equation needs to be modified to account for the interception of precipitation and the buildup of an interception store, from which water can evaporate.

Another important remark should be made in relation to the budget equations. Although it is customary to refer to

eqns [1]–[3] as the ‘surface balance equations,’ it is essential to recall that in most cases they are established with respect to a layer of finite thickness (Figure 3). This makes it necessary to include the storage terms that account for changes in the energy and mass content of this layer (ΔQ_s and ΔW , respectively). The storage terms become particularly important when dealing with the microclimates of vegetative canopies or urban areas.

Profiles of Mean Quantities and Turbulence Characteristics

One of the fundamental characteristics of climate variables in the atmospheric layer directly affected by the surface is their pronounced vertical variability. The associated ‘vertical profiles’ again depend in a systematic way on the surface properties, displaying temporal variations in correspondence to the time scales that characterize the governing budget equations.

Figure 4 exemplarily shows daytime profiles of mean wind speed and potential temperature above a bare soil and a forest. In the panel on the right-hand side of the figure, $d + z_0$ refers to the height above which the shapes of the profiles start to become comparable to those observed over a bare soil. This feature of the profiles implies that they can be formulated using z_0 and d as characteristic length scales. A ‘bare soil’ or relatively

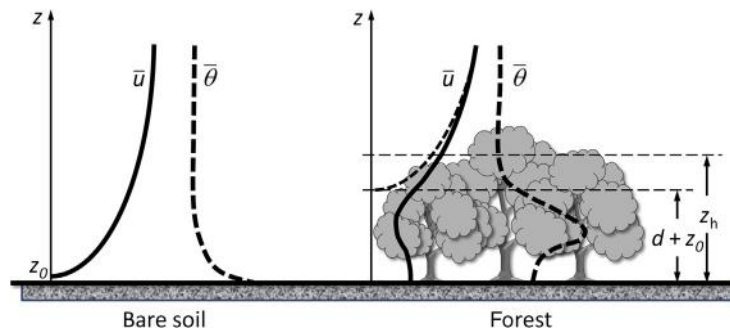


Figure 4 Typical daytime mean vertical profiles of mean wind speed ($\bar{u}$) and potential temperature ($\bar{\theta}$) over a bare soil (left) and a vegetated surface (right). On the right-hand side of the figure, z_h denotes the average height of the canopy, $d \sim 2/3z_h$ is the zero-plane displacement, and $z_0 \sim 0.1z_h$ is the roughness length.

smooth surface (lower-right corner in Figure 2) would then simply be characterized through a very small zero-plane displacement.

The form of the profiles is, in general, complex. Under strongly simplifying assumptions, the profiles can nevertheless be approximated by analytical equations. Well established from theoretical and experimental studies, as well as from laboratory flow experiments, is the 'logarithmic wind profile,' or log-profile (eqn [4]). Strictly speaking, the log-profile is valid over only bare, flat, homogenous terrain and at a sufficiently large distance from the surface (i.e., within the inertial sublayer), and it can be described on the basis of the Monin–Obukhov similarity theory. For negligibly small d , it reads:

$$\bar{u}(z) = \frac{u_*}{k} \left[\ln\left(\frac{z}{z_0}\right) - \Psi_M\left(\frac{z}{L}\right) \right] \quad [4]$$

where u_* is the 'friction velocity' as determined by the Reynolds stress at z_r , $k = 0.4$ is the von Kármán constant, and the effect of atmospheric stability is taken into account through the so-called Obukhov length L and the associated stability function Ψ_M . In the case of a nonnegligible zero-plane displacement, z should be replaced by the reduced height $z - d$.

In practical applications, zero-plane displacement and roughness length can be determined from fitting observed wind speed profiles to eqn [4] or, alternatively, from morphometric properties of the surface. As a rule of thumb, $d \sim 2/3 z_h$ and $z_0 \sim 0.1 z_h$ may be used.

Within the roughness sublayer, the mean wind speed departs from the logarithmic behavior, with a strong retardation at the mean height of the roughness elements (Figure 4) due to form and viscous drag. Under special conditions, the momentum balance equation for the roughness sublayer can be solved analytically to yield an exponential function usually expressed as:

$$\bar{u}(z) = \bar{u}(z_h) \exp \left\{ -a \left(1 - \frac{z}{z_h} \right) \right\} \quad [5]$$

where a is a parameter that depends on the density and character of the roughness elements and is experimentally found to be in the range of 1–4. Note that while stability is a strong factor determining the shape of the wind profile away from the surface, the form, character, and distribution of roughness elements exert a much stronger influence within the roughness sublayer than stability does. For this reason, eqn [5] without stability extension provides a very good first-order approximation.

Variables other than wind speed and temperature also display characteristic vertical profiles. In plant stands, the vertical distribution of net radiation depends on the albedo of the canopy, as well as on the extinction of solar radiation and the absorption and emission of thermal radiation at different levels within the canopy. Owing to the greater absorption in the visible range compared to the near-infrared range, the spectral composition of the solar radiation flux varies from the top to the bottom of a stand.

Profiles of humidity and CO₂ (not shown) in forest stands often display an inflection point within the canopy, as illustrated for potential temperature in Figure 4. This feature can be explained by the divergence of the relevant energy and mass fluxes, the distribution of the sources and sinks of water vapor

and CO₂ (including those at ground level), and advective effects.

In plant stands, turbulence statistics such as velocity variances or turbulent fluxes of sensible heat and momentum exhibit a strong reduction in magnitude from the canopy top to the ground, where very small or vanishing values are observed. The most striking feature of canopy turbulence is probably the fact that it is governed by so-called coherent structures with spatial scales on the order of the canopy height. As a consequence, the turbulent exchange of heat, moisture, or trace gases within canopies is often characterized by countergradient transport. This means that 'K theory,' which is based on the assumption of small dominating eddies and is a well-established concept for the inertial sublayer, is not a useful description of turbulent transport in the roughness sublayer.

Owing to the rough nature of the surface, turbulent mixing is stronger in the upper part of the roughness sublayer than in the inertial sublayer. As for the fluxes of sensible and latent heat, this increase is most pronounced under near-neutral conditions. In this case, the turbulent transport just above the canopy can be up to three times larger than in the inertial sublayer, whereas for momentum the enhancement is of the order of 10%.

Comparison of Urban and Rural Microclimates

The ways in which microclimates are shaped by properties of the underlying surfaces can be illustrated by contrasting rural and urban environments. Radiative energy fluxes over the urban surface are different from those found in rural areas, mainly due to enhanced aerosol load and smaller albedo in the urban environment. The latter reflects the fact that a substantial part of the shortwave incoming radiation is trapped between the buildings, heating up the surface material. During the daytime hours (Figure 5(a)), the urban environment gets more net radiative energy due to the smaller albedo. During night, on the other hand, the net radiation is typically larger over a rural than over an urban surface (Figure 5(b)) owing to differences in the longwave loss.

Spatial heterogeneity in the availability of water for evapotranspiration in an urban fabric typically leads to a substantially different partitioning of the atmospheric energy fluxes across different surface types, which is best illustrated in terms of the Bowen ratio, that is, the ratio of sensible to latent heat, $\beta = Q_H/Q_E$ (Figure 6). For this particular case (a midlatitude European type of city structure), the turbulent flux of sensible heat (Q_H) dominates the latent heat flux (Q_E) over the urban surface ($\beta \approx 2$), while the situation is reversed ($\beta \approx 0.5$) over the nearby rural surfaces, with the suburban sites exhibiting an intermediate behavior. For all types of surfaces, the partitioning of the available energy between latent and sensible heat undergoes a pronounced daily cycle, reflecting changes in the vertical profiles of wind speed, temperature, humidity, and, in the case of the urban environment, heat storage (ΔQ_s).

The characteristic energy flux partitioning seen in Figure 6 results in an enhanced near-surface temperature within an urban area that usually is referred to as the 'urban heat island'

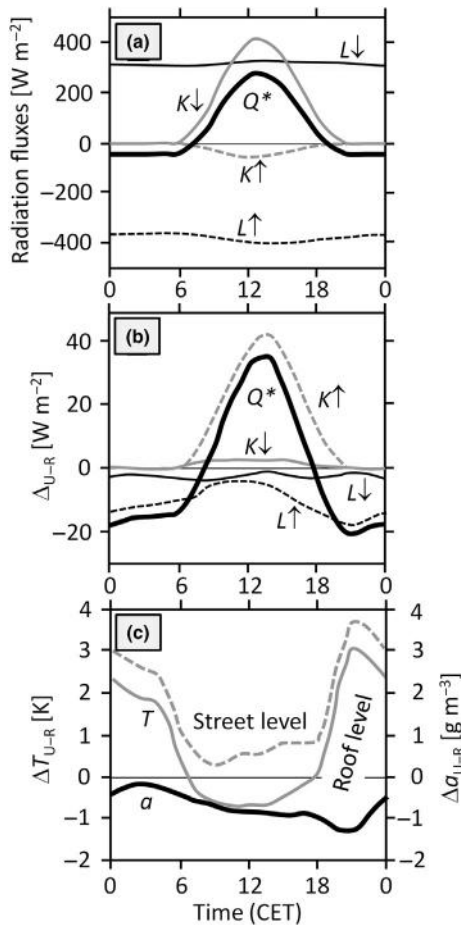


Figure 5 (a) Average daily cycles of the radiation balance components at an urban site in the city of Basel, Switzerland (data from September 2001 to August 2002); (b) daily cycle of the corresponding urban–rural difference (ΔQ_{U-R}) for these mean radiation fluxes ('urban' data are averaged over two urban sites, and 'rural' data are averaged over three rural sites around Basel); and (c) daily cycle of urban–rural difference in temperature (gray lines, left scale) and absolute humidity (black line, right scale). Note that ΔT_{U-R} represents a measure of the so-called urban heat island (see text for details); the data represent averages from two sites (urban) and three sites (rural), respectively, and stem from one summer month of the full-year data of (a) and (b). Composed from figures in Christen, A., 2005. *Atmospheric Turbulence and Surface Exchange in Urban Environments*, PhD dissertation, University of Basel, Switzerland, published as *stratus* 11, p. 140.

(UHI). The UHI is often associated with a drier urban environment, and its strength is a strong function of time and position within the roughness sublayer (Figure 5(c)).

Apart from being related to the size of the urban population (i.e., the city), the maximum strength of the UHI has been shown in other studies to be associated with geometric and surface properties of the urban environment (Figure 7).

Within and above urban canopy layers, temperature, humidity, and wind speed exhibit characteristic profiles similar to those shown in Figure 4 (right) for a forest stand. These profiles are the result of turbulent exchange processes, which are strongly determined by the character of the roughness elements (form and building material) and their density and

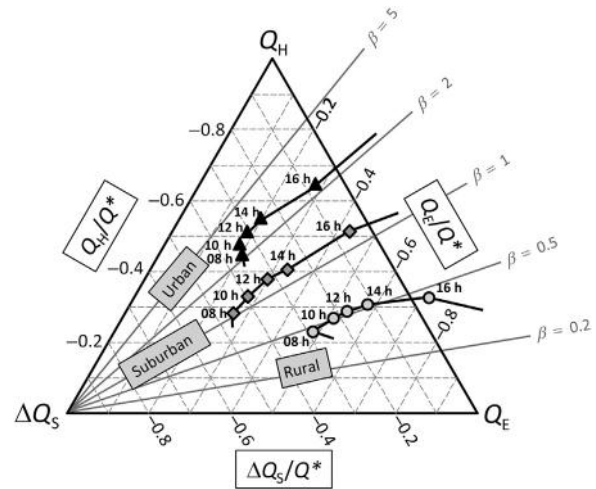


Figure 6 Ternary plot contrasting atmospheric energy flux partitioning over urban, suburban, and rural sites in and around Basel, Switzerland. Data cover one summer month in 2002. Modified after Christen, A., Vogt, R., 2004. Energy and radiation balance of a central European city. *International Journal of Climatology* 24 (11), 1395–1421, with permission from the author.

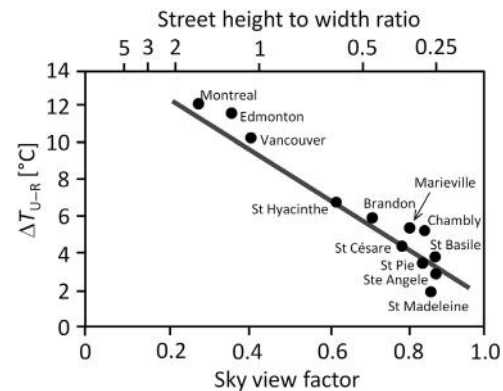


Figure 7 Maximum heat island intensity in Canadian cities and its relation to surface properties (aspect ratio, top scale or sky-view factor, and lower scale). Redrawn after Oke, T.R., 1997. *Urban environments*. In: Bailey, W.G., Oke, T.R., Rouse, W.R. (Eds.). *The Surface Climates of Canada*. McGill-Queen's University Press, Montreal, pp. 303–327.

height distribution. Far away from the roughness elements (i.e., within the inertial sublayer), these turbulent fluxes are essentially constant with height, and this corresponds to the situation over relatively smooth surfaces. Closer to the surface, however, the complicated distribution of sources and sinks yields more complicated profiles (Figure 8). The distribution of sources and sinks of heat and air pollutants, from traffic at the street level and chimneys from domestic heating near rooftops, suggests that in urban environments, profiles of the sensible heat flux, mean temperature, and air pollutants are less systematic than corresponding profiles of the momentum flux, for which the main sink is at ground level. Vertical variations in turbulent fluxes also affect the height distribution of other turbulence-related variables such as velocity or scalar variances. These, in turn, govern the turbulent diffusion of energy, water vapor, and air pollutants.

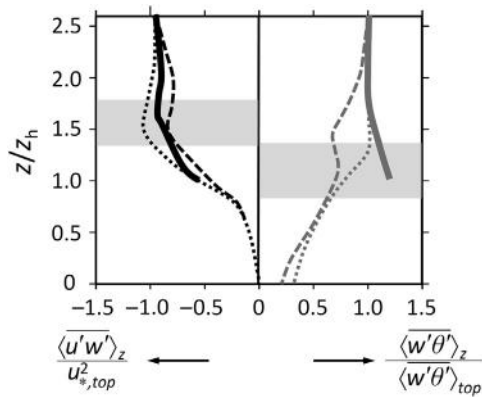


Figure 8 Mean profiles of turbulent fluxes of momentum $\langle u'w' \rangle$ (left) and sensible heat $\langle w'\theta' \rangle$ (right), normalized with their respective values far from the roughness elements. In the case of the momentum flux, normalization was achieved with the square of the friction velocity to preserve the negative sign. The two gray areas indicate the height range where pronounced inflection points appear in the profiles. For both fluxes, the dashed and dotted lines represent profiles taken within the urban area, whereas the thick continuous lines refer to profiles collected in a suburban environment. Composed from figures in Christen, A., 2005. Atmospheric Turbulence and Surface Exchange in Urban Environments, PhD dissertation, University of Basel, Switzerland, published as *stratus* 11, p. 140.

See also: Agricultural Meteorology and Climatology. **Boundary Layer (Atmospheric) and Air Pollution:** Coherent Structures; Complex Terrain; Modeling and Parameterization; Overview; Stably Stratified Boundary Layer; Surface Layer. **Climate and**

Climate Change: Energy Balance Climate Models; Overview. **Numerical Models:** Parameterization of Physical Processes: Turbulence and Mixing. **Turbulence and Mixing:** Turbulent Diffusion.

Further Reading

- Arya, P.S., 2001. Introduction to micrometeorology. In: Holton, J.R. (Ed.), International Geophysics Series, second ed. vol. 71. Academic Press, San Diego, CA., p. 420.
- Bailey, W.G., Oke, T.R., Rouse, W.R., 1997. The Surface Climates of Canada. McGill-Queen's University Press, Montreal, p. 369.
- Brutsaert, W.H., 1982. Evaporation into the Atmosphere. Reidel, Dordrecht, p. 299.
- Cermak, J.E., Davenport, A.G., Plate, E.J., Viegas, D.X., 1995. Wind climates in cities. In: NATO ASI Series E: Applied Sciences, vol. 27. Kluwer Academic Publishers, Dordrecht, p. 771.
- Garratt, J.R., 1992. The Atmospheric Boundary Layer. Cambridge University Press, Montreal, p. 316.
- Geiger, R., Aron, R.H., Todhunter, P., 2009. The Climate Near the Ground, seventh ed. Rowman & Littlefield Publishers, Lanham, MD, p. 623.
- Jones, H.G., 1992. Plants and Microclimate: A Quantitative Approach to Environmental Plant Physiology, second ed. Cambridge University Press, Cambridge, p. 428.
- Monteith, J.L., Unsworth, M.H., 2008. Principles of Environmental Physics, third ed. Academic Press, Burlington, VT, p. 418.
- Oke, T.R., 1987. Boundary Layer Climates, second ed. Methuen, London, p. 372.
- Rosenberg, N.J., Blad, B.L., Verma, S.B., 1983. Microclimate: The Biological Environment, second ed. Wiley Interscience, New York, p. 495.

Modeling and Parameterization

AAM Holtslag, Wageningen University, Wageningen, The Netherlands

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

In this contribution we deal with the representation of the atmospheric boundary layer (ABL) for modeling studies of weather, climate, and air quality. As such we review the major characteristics of the ABL, and summarize the basic parameterizations for the description of atmospheric turbulence and the surface fluxes, where we emphasize the modeling and parameterization of turbulence in the atmospheric boundary layer over land without clouds. The modeling principles are illustrated with the outcome of single-column and mesoscale models for a variety of conditions using field data and fine-scale model results. For readers not familiar with atmospheric turbulence and meteorology, some background and basic definitions are also given.

Introduction

The atmospheric boundary layer (ABL) is the lower part of the atmosphere which is in continuous interaction with the earth's surface due to friction and heating or cooling (Stull, 1988; Garratt, 1992). The ABL is therefore generally characterized by turbulence and a diurnal cycle of temperature, wind, specific humidity, and other tracers in particular over land. Because of its capability to mix air with different properties efficiently, the representation of turbulence is directly relevant for atmospheric and environmental modeling. For instance, turbulence directly impacts on the transfer of momentum, sensible heat, water vapor, ozone, and methane, among many other quantities, between the earth's surface and the atmosphere. Turbulence also defines the mixing of properties inside the atmospheric boundary layer, the transfer of quantities between the boundary layer and the clear or cloudy atmosphere aloft, and the mixing inside clouds.

Turbulence in the atmospheric boundary layer is the three-dimensional, chaotic flow of air with timescales typically between a second and an hour (e.g., Tennekes and Lumley, 1982). The corresponding length scales are from a millimeter up to the depth of the boundary layer (or more in the case of clouds). The depth of the dry ABL can vary over land between tens of meters during night up to kilometers during daytime (Figure 1). Over sea the depth is typically a few hundred meters and rather constant on the timescale of a day. Most of the atmosphere above the ABL is not turbulent, although turbulence can occur throughout the whole atmosphere. For instance, cumulus-type clouds, which may grow into thunderstorms, are always turbulent through convection produced by the heat released due to the condensation of water vapor. Turbulence can also occur in clear air above the ABL; most of this is produced in layers of strong vertical wind shear at the boundary between air masses (so-called 'clear-air turbulence').

Because of the mixing capacity of turbulence, modeling atmospheric boundary layers is relevant for many practical applications. For instance, chimney plumes are diluted and spread over larger volumes than they would be without turbulence. As such, strong local peaks of pollution are prevented and otherwise clean air is polluted (e.g., Nieuwstadt and van Dop, 1982). In practice turbulence may also cause

engineering problems, because it shakes structures such as bridges, towers, and airplanes, causing failure of such systems in extreme cases. Turbulent fluctuations in the horizontal motions during severe storms can be fatal to tall buildings or bridges, particularly if resonance (e.g., forcing of a system at its natural frequency) occurs.

The correct formulation of the overall effects by turbulence, either inside or outside the atmospheric boundary layer, is an essential part of atmospheric models dealing with the prediction and study of weather, climate, and air quality. These models are based on solving the equations dealing with atmospheric behavior. Even with state-of-the-art computers, the number of grid points in atmospheric models is limited. This implies that on the regional and global scale the atmospheric model equations are usually applied to fairly large 'air boxes.' Such boxes are in the order of 10–100 km wide and ten to a few hundred meters thick. In these large boxes, small-scale motions make air parcels interact and mix. For example, if a hot parcel is located next to a cold parcel, turbulent motion at their boundaries will heat the cool and cool the hot parcel. Thus, a closure formulation is needed to reproduce mixing by the turbulent motions into the model-resolved scales using the equations for the large-scale 'mean' motions. It is important to realize that the closure formulation needs to be expressed in terms of variables available

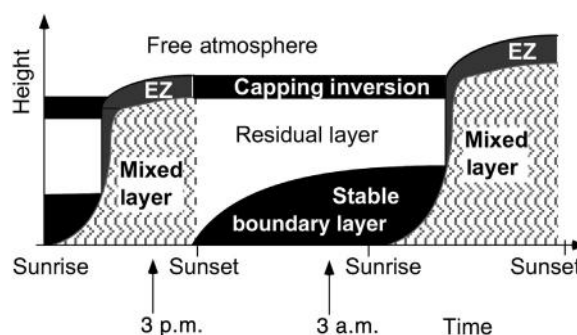


Figure 1 Idealized diurnal evolution of the atmospheric boundary layer over land in fair weather. After Stull, R.B., 1988. *An Introduction to Boundary-Layer Meteorology*. Kluwer Academic Publishers, Dordrecht, 666 pp. (reprinted 1999).

in the modeling context. This is called a 'parameterization' (Stull, 1988; Garratt, 1992; Pielke, 2001).

Background

Atmospheric models for the forecasting and study of weather, climate, and air quality are typically based on integration of the basic equations governing atmospheric behavior. These equations are the gas law, the equation of continuity (mass), the first law of thermodynamics (heat), the conservation equations for momentum (the so-called 'Navier–Stokes equations'), and usually equations expressing the conservation of moisture, trace gases, and air pollutants. At one extreme, atmospheric models may deal with the world's climate and climate change; at the other, they may account for the behavior of local flows at coasts, in mountain-valley areas, or even deal with individual clouds. This all depends on the selected horizontal modeling domain and the available computing resources.

Since there is an enormous range of scales in atmospheric motion and turbulence, there is a need to separate the scales of atmospheric turbulence from large-scale motions. Let C denote an atmospheric variable, such as specific humidity. Then $\bar{C}$ represents a mean or 'smoothed' value of C , typically taken on a horizontal scale of order 10 (or more) km and a corresponding timescale in the order of 10 min–1 h. A local or instantaneous value of C would differ from $\bar{C}$. Thus, we have

$$C = \bar{C} + c. \quad [1]$$

Here c represents the small-scale fluctuations. Note that we use lower case for the latter (often primes are used as well to indicate fluctuations). In principle, the fluctuations around the mean motion also reflect gravity waves and other small-scale motions, in addition to turbulence. Gravity waves often coexist with turbulence or are generated by turbulence. If the wind at the same time is weak, there may be no turbulence at all. Anyhow, if turbulence exists, it is usually more important for most atmospheric applications, because it mixes more efficiently than the other small-scale motions.

To make the mathematical handling of c tractable, it must satisfy the so-called 'Reynolds postulates.' These require, for example, that $\bar{c} = 0$ and that small- and large-scale values must not be correlated. After a quantity has been averaged to create a large-scale quantity, further averaging should produce no further changes, in order for this postulate to apply. The mean of the summation of two variables A and C will produce $\overline{A \pm C} = \bar{A} \pm \bar{C}$. A further condition is that a mean variable $\bar{C}$ must be differentiable, since differentials show up in the atmospheric equations (see below). In practice, not all these conditions are rigorously satisfied. If the Reynolds postulates are fulfilled, then the averaging for the product of two variables provides

$$\overline{AC} = \bar{A}\bar{C} + \overline{ac}. \quad [2]$$

The second term at the right hand side of eqn [2] is known as the turbulent covariance. Similarly, the turbulence variance of a quantity is given by $\overline{C^2} - (\bar{C})^2$ (which is the square of the standard deviation).

If in eqn [2], the variable represents one of the velocity components (U, V, W in the x, y, z direction, respectively), then

$\overline{AC}$ is the total flux of C and the second term at the right hand side of eqn [2] represents a turbulent flux of C . For instance, $\overline{uc}$ and $\overline{wc}$ are the horizontal and vertical turbulent fluxes of the variable C , respectively. Here u and w are the turbulent fluctuations of the horizontal and vertical velocities. Near the surface, the mean vertical wind $\bar{W}$ is usually small, and thus the total vertical fluxes are normally dominated by the turbulent contributions.

Atmospheric Boundary-Layer Structure

Turbulent fluctuations, variances, and fluxes of variables are influenced by the vertical boundary-layer structure. Here the variation of temperature in the atmospheric boundary layer plays an important role. Since pressure decreases with altitude, air parcels, which are forced to rise (sink), do expand (compress). According to the first law of thermodynamics, a rising (sinking) parcel will cool (warm) if there is no additional energy source such as condensation of water vapor. Then this is called a dry adiabatic process.

It can be shown that in the atmospheric boundary layer, the temperature (T) variation with height for a dry adiabatic process is $dT/dz = -g/C_p$ (here g is gravity constant and C_p is specific heat at constant pressure). The value for g/C_p is approximately 1 K per 100 m. An atmospheric layer which has such a temperature variation with height is called neutral for dry air (at least when there is no convection arising from other levels). In that case $\Theta = T + (g/C_p)z$ is constant, where Θ is called the potential temperature (note that the previous definition for potential temperature is not accurate above the boundary layer). Since air normally contains water vapor and because moist air is lighter than dry air, we have to correct for the influence of this on vertical motions. Consequently, a virtual potential temperature is defined as $\Theta_v = \Theta(1 + 0.61q)$, where q is the specific humidity (defined as the mass of water vapor per unit mass of moist air).

In a neutral layer with constant Θ_v , vertical motions of moist (not saturated) air can maintain themselves. If the virtual potential temperature of the atmospheric layer increases with height, vertical displacements are suppressed. This is called a stable condition (or 'inversion'). At the other hand, when the virtual potential temperature decreases with height, vertical fluctuations may be accelerated. Consequently this is called an unstable condition. Thus in considerations with turbulent fluctuations and atmospheric stability, we have to deal with the virtual potential temperature and not with the actual temperature. Similarly, the vertical flux of sensible heat is connected to turbulent fluctuations of (virtual) potential temperature; e.g., it reads as $\overline{w\theta_v}$ (in m K s^{-1}). The latter relates directly to the energy per time and unit area H by $H = \rho C_p \overline{w\theta_v}$ (in W m^{-2}), where ρ is density of the air (in kg m^{-3}).

Figure 2 (after Stull, 1988) provides the typical, idealized, mean vertical profiles for temperature T , potential temperature Θ , specific humidity q , in addition to the horizontal wind M (defined by $M^2 = U^2 + V^2$). These profiles apply for an atmospheric boundary layer over land in clear sky conditions in the afternoon and around midnight. Note that in the free atmosphere the horizontal wind is mostly a result of the acting of the large-scale pressure differences and the Coriolis force due

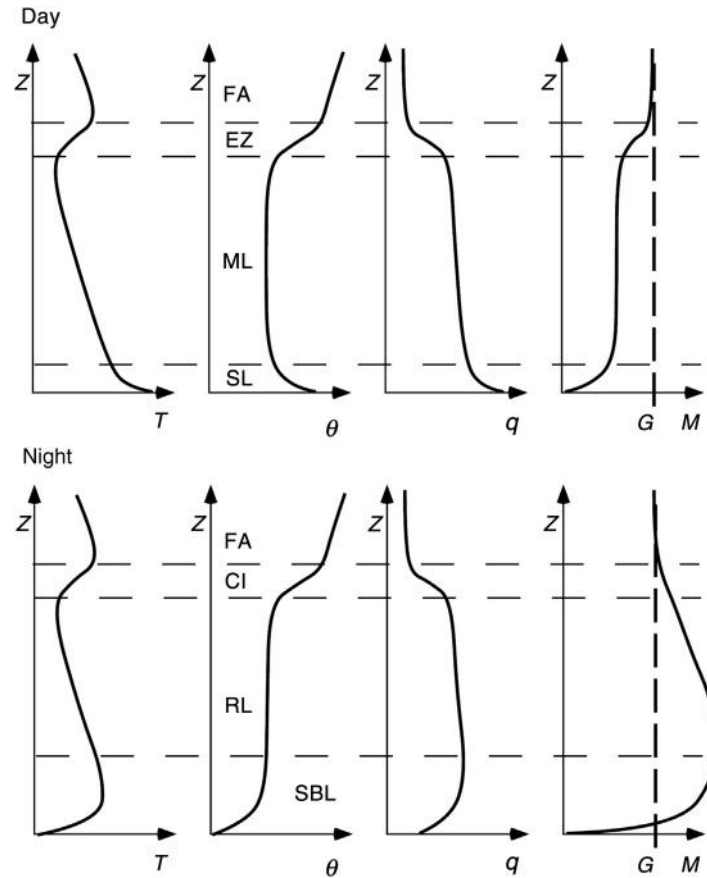


Figure 2 Idealized vertical profiles of mean variables in the atmospheric boundary layer over land in fair weather. See text for additional information. After Stull, R.B., 1988. *An Introduction to Boundary-Layer Meteorology*. Kluwer Academic Publishers, Dordrecht, 666 pp. (reprinted 1999).

to the rotation of the earth (but other effects may play a role as well). The resulting wind is known as the ‘geostrophic’ wind and indicated with G in **Figure 2** (see dashed line). In the daytime boundary layer the actual wind is smaller due to surface friction, while at clear nights the actual wind away from the surface may be substantially stronger than G due to inertial effects (resulting in the so-called ‘low-level jet’).

The temporal variation of the mean boundary-layer profiles over land can be quite substantial due to the strong diurnal variation of solar incoming radiation and the nighttime cooling at the land surface. During daytime the turbulent boundary layer may grow to several kilometers into the nonturbulent ‘free atmosphere’ (indicated as FA in **Figure 2**). At night the turbulent part of the stable boundary layer (SBL) may only extend up to a few hundred meters or less (the lowest dashed line in the lower figure). An idealized picture for the temporal variation of the boundary layer over land is given in **Figure 1** (after Stull, 1988). Here the arrows with local time indications refer to the day and nighttime figures of **Figure 2**.

Figure 2 also indicates that the boundary layer during daytime shows a three-layer structure: an unstable ‘surface layer,’ a ‘well-mixed layer’ with rather uniform (virtual) potential temperature, and a stably stratified ‘entrainment zone.’ In the latter zone, turbulence acts to exchange heat, momentum, water vapor, and trace gases between the boundary layer and the free

atmosphere. During nighttime, often the vertical structure of the previous day persists above the SBL. As such a ‘residual layer’ with sporadic turbulence (remaining from the previous day) can be identified as well as a ‘capping inversion.’

Modeling

The challenge of modeling the atmospheric boundary layer is the prediction of the temporal variation of the vertical and horizontal structures in response to the influence of the major processes acting in the atmosphere and at the earth’s surface. As such the governing equations have to be integrated. In practice, the variables are split into ‘mean’ large-scale motions and small-scale fluctuations as in [eqn \[1\]](#). Inserting this into the basic equations and after averaging this provides a set of equations for the behavior of the large-scale (mean) variables. The large-scale variables are then used explicitly in atmospheric models. This can be demonstrated as follows (Stull, 1988; Garratt, 1992; Pielke, 2001; Holtslag and Duynkerke, 1998).

The general character of any of the budget equations dealing with atmospheric motions is

$$\frac{dC}{dt} = S_i. \quad [3a]$$

Here S_i represents the subsequent sources and sinks for the variable C (such as radiation or chemistry effects). The notation dC/dt represents the total rate of change for the variable C by local changes ($\partial/\partial t$), and changes transported with the fluid motion in the three directions. As such, we have

$$\frac{\partial C}{\partial t} + U \frac{\partial C}{\partial x} + V \frac{\partial C}{\partial y} + W \frac{\partial C}{\partial z} = S_i. \quad [3b]$$

Here U , V , W are the wind speeds in the three directions x , y , z , respectively.

If in the atmospheric motion each variable is split into a mean component and a fluctuation then eqn [3b] provides after Reynolds-averaging, some algebraic manipulations and simplifying assumptions, a budget equation for the mean variable $\bar{C}$. This reads as

$$\frac{d\bar{C}}{dt} = \frac{\partial \bar{C}}{\partial t} + \bar{U} \frac{\partial \bar{C}}{\partial x} + \bar{V} \frac{\partial \bar{C}}{\partial y} + \bar{W} \frac{\partial \bar{C}}{\partial z} = \bar{S}_i - \frac{\partial \bar{u}\bar{c}}{\partial x} - \frac{\partial \bar{v}\bar{c}}{\partial y} - \frac{\partial \bar{w}\bar{c}}{\partial z}. \quad [4]$$

We may note that in the derivation of eqn [4], single terms representing fluctuations have disappeared (as above in eqn [2]). However, terms involving the product of two fluctuations did remain.

Thus because the basic equations are nonlinear, the budget equations for the mean variables contain terms involving small-scale motions. The latter terms are of the form of a divergence of fluxes produced by such motions in the three directions and appear as the last three terms in eqn [4]. These motions are said to be subgrid and consequently, closure formulations or parameterizations are needed to introduce mixing by the small-scale, subgrid motions into the equations for the large-scale motions (as resolved by the model). Note that additional terms may also appear in eqn [4] when the source or sink term S_i incorporates nonlinear effects (such as in the case of chemistry).

A special and simple form of eqn [4] arises for horizontally homogeneous conditions. In such cases the terms including horizontal derivatives are negligible. If in addition the mean vertical wind is small and if there are no other sources and sinks, then eqn [4] provides

$$\frac{\partial \bar{C}}{\partial t} = - \frac{\partial \bar{w}\bar{c}}{\partial z}. \quad [5]$$

This equation is known as the one-dimensional, vertical diffusion equation. It shows that the local time rate of change for the mean of a variable (such as temperature or wind) at a certain height is given by the divergence of the turbulent (corresponding heat or momentum) flux in the vertical direction. As such, information on the turbulent flux may produce a local forecast of the variation of a mean variable (but only under the simplifications mentioned). However, normally the other terms in eqn [4] are also relevant, in particular the terms with mean wind speed (the so-called 'advection terms'). This means that in general the budget equations for momentum, heat, and the various scalars are closely coupled in any atmospheric model.

Equation [4] can also be integrated in the vertical direction to account for the averaged effect of turbulence on the boundary layer development. This is particularly suitable for very unstable boundary layers, which are more or less uniformly mixed by the dominant presence of convection. A well-mixed structure is frequently observed for potential

temperature in daytime convective conditions over land (Figure 2). Integration of eqn [4] provides (e.g., Garratt, 1992)

$$h \frac{d\langle \bar{C} \rangle}{dt} = \bar{w}\bar{c}_0 - \bar{w}\bar{c}_e. \quad [6]$$

Note that we have neglected the horizontal fluxes and the source term S_i here for simplicity. In eqn [6], $\langle \bar{C} \rangle$ is the mean concentration across the boundary layer, h is the mean depth of the turbulent boundary layer, $\bar{w}\bar{c}_0$ is the surface flux, and $\bar{w}\bar{c}_e$ is the flux at the top of the boundary layer. The latter represents the exchange of C through entrainment of air from above the boundary layer into the ABL.

Equation [6] can be used to solve for the time development of $\langle \bar{C} \rangle$, but a separate equation for the boundary layer depth is needed in addition to information on the fluxes at the top and bottom of the ABL. Such an approach is known as 'bulk' or 'mixed-layer' modeling. For clear boundary layers, it appears that the entrainment flux for heat at the top of the boundary layer is proportional to the surface heat flux and to surface friction. However, for cloudy boundary layers the parameterization of the entrainment flux is not that simple.

Before we proceed with more detailed parameterizations for the fluxes in the boundary layer, let us deal with the derivation of the surface fluxes. These fluxes enter as boundary conditions when solving the budget equations for all the relevant mean variables (in any approach). It is important to realize that near the surface, the average wind must vanish because the mean wind is zero at the earth's surface. At the other hand, we know from observations that the fluxes of heat, momentum, and trace gases are nonzero. Consequently, it is convenient to model an 'effective' surface flux $\bar{w}\bar{c}_0$ of a conserved variable due to the combined effect of molecular diffusion and turbulence at the surface. This can be achieved by writing (e.g., Beljaars and Holtslag, 1991)

$$\bar{w}\bar{c}_0 = \beta_t w_t (C_0 - C_a). \quad [7]$$

Here, C_0 and C_a are the values of the transported variable at the surface and in the air, respectively; β_t is a transfer coefficient, and w_t is an effective transport velocity representing the turbulence. For example, in near-neutral conditions the effective transport velocity is well represented by the so-called surface friction velocity u_{*0} (which is related to the surface momentum flux). Then it can be shown that $\beta_t = \kappa / \ln(z/z_0)$, where κ is the 'Von Karman' constant (often specified as $\kappa \cong 0.4$), z is the corresponding height of C_a in the lowest part of the boundary layer, and z_0 is the so-called surface roughness length for the variable C .

Diagnostic Local Mixing Parameterizations

To solve the budget eqn [4] for all the mean atmospheric variables involved, the terms involving turbulent fluxes need to be parameterized. As mentioned before, this means that the fluxes need to be expressed in terms of available mean model quantities, both in the atmosphere and at the surface. Once this has been achieved, the atmospheric model equations can be integrated. Thus, starting with proper initial values, new values can be calculated for the following time step and so on.

The most frequently used parameterization for environmental and atmospheric models is known as first-order closure

or often also called K -theory. In this theory it is assumed that the flux $\overline{wC}$ of a variable C in the vertical direction z is down the vertical gradient of the mean concentration of C per unit mass. Thus

$$\overline{wC} = -K_c \frac{\partial \overline{C}}{\partial z}. \quad [8]$$

Here, K_c is known as the 'eddy-diffusivity' or mixing coefficient for the variable C . Similarly, the horizontal fluxes can be represented in terms of horizontal gradients. Note that the corresponding eddy-diffusivities are typically not constant, but that they generally depend on properties of the flow and the variable of interest. This also means that normally no analytic solutions are possible, not even for the simple case in which eqns [5] and [8] are combined.

We may note that the dimension of an eddy-diffusivity is a length scale l times a velocity scale. These are proportional to the products of effective eddy sizes and eddy velocities in the corresponding directions. Often a diagnostic expression is used for the eddy-diffusivity, on basis of what is called 'mixing length theory' (in analogy with molecular diffusion). The result reads as

$$K_c = l^2 S f(Ri). \quad [9]$$

Here S is vertical wind shear (that is the variation of mean horizontal wind with height). Note that the combination lS in eqn [9] has units of velocity. In eqn [9], $f(Ri)$ denotes a functional dependence on local stability as represented by the gradient Richardson-number Ri defined by

$$Ri = \frac{g}{\Theta_v} \frac{\partial \overline{\Theta_v} / \partial z}{(\partial \overline{U} / \partial z)^2 + (\partial \overline{V} / \partial z)^2}. \quad [10]$$

Here g is the acceleration due to gravity and $\overline{\Theta_v}$ is the mean 'virtual potential temperature.'

The specification of the length scale l is not at all straightforward, except near the surface where the so-called 'surface-layer similarity theory' (Stull, 1988) provides that $l \propto z$. A frequently used form for l is

$$\frac{1}{l} = \frac{1}{\kappa z} + \frac{1}{\lambda}. \quad [11]$$

Here λ is a turbulent length scale, which should be valid for the turbulence far above the surface. We note that the latter has a rather empirical nature and consequently there is no agreement on the specification of λ in the literature.

Equations [9] and [11] are diagnostic equations, which indicate that the eddy-diffusivity varies with height, wind speed, stability, etc. In combination with the flux parameterization of eqn [8], it follows that the flux at a certain height depends on the local gradient of the mean variable involved. Consequently, the approach is referred to as a 'diagnostic local mixing approach.' Such an approach is mostly suitable for relatively homogeneous conditions with neutral and stable stratification, and is not so suitable for cases with convection (see Section Nonlocal Mixing Parameterizations).

Prognostic Mixing Parameterizations

A physically realistic alternative to the diagnostic approach is to relate the eddy-diffusivity of eqn [8] to the actual turbulent

kinetic energy of the flow, by using the prognostic turbulent kinetic energy (TKE) equation and an appropriate choice for the turbulent length scale. It is important to realize that the kinetic energy of atmospheric motion per unit of mass E is given by half of the sum of the velocities squared in the three directions (as in classic mechanics), e.g., $E = (U^2 + V^2 + W^2)/2$. Similar as with respect to eqn [2], we can separate between the mean kinetic energy $\overline{E}$ of the mean atmospheric motions and the Turbulent Kinetic Energy (TKE or e) of the small-scale fluctuating motions by turbulence. Thus e is given by $e = (\overline{u^2} + \overline{v^2} + \overline{w^2})/2$.

The prognostic equation for e reads in its basic form as

$$\frac{de}{dt} = -\overline{uw} \frac{\partial \overline{U}}{\partial z} - \overline{vw} \frac{\partial \overline{V}}{\partial z} + \frac{g}{\Theta_v} \overline{w\theta_v} + D - \varepsilon. \quad [12]$$

Here de/dt is the total variation of e with time (the sum of local variations and those transported with the mean air motion). The two terms at the immediate right hand side of eqn [12] represent the shear production of turbulence. These depend primarily on vertical variations of wind or, near the ground, on wind speed and surface roughness. The terms are almost always positive. The third term in eqn [12] represents the rate of production or breakdown of turbulence by buoyancy effects (such as heat convection). It depends directly on density effects, which can be written in terms of the virtual potential temperature $\overline{\Theta_v}$, and its turbulent flux $\overline{w\theta_v}$. The term D in eqn [12] represents divergence and pressure redistribution terms. These have a tendency to cancel near the surface. Finally, the term ε reflects the molecular dissipation of turbulence into heat and this term is always positive. In fact ε is typically proportional to e/τ , where τ is the characteristic timescale for the turbulent mixing process.

Using eqn [12], TKE can be calculated for given mean profiles when the corresponding fluxes are calculated using eqn [8] for all fluxes involved. In this approach the diffusivities are typically calculated with equations of the form

$$K_c = \alpha_c l \sqrt{e}. \quad [13]$$

Here α_c is a constant depending on the variable of interest. The length scale is typically calculated with a similar type of diagnostic equation as given by eqn [11]. This approach is known as the 'TKE-length scale approach' and it is an example of the so-called 1.5-order closure. Sometimes a prognostic equation is used for the length scale as well, but such an approach is more popular in engineering applications and then in the atmospheric sciences.

It can be shown that eqn [9] is a solution of eqns [12] and [13] in stationary conditions and when other simplifications are made such as the neglect of the influences by advection and turbulence divergence in the TKE equation. A more advanced turbulence scheme is known as 'second-order closure.' In such an approach, prognostic equations are developed for the fluxes and variances themselves. Such equations have a very similar structure as eqn [12] for kinetic energy. Unfortunately, new unknowns are present in these equations. These must be related to the other variables in the model equations, always involving assumptions. Thus, second-order closure involves many more than the original equations and is therefore computationally more time-consuming ('expensive') than first-order and 1.5-order closure.

One may expect that a model with 1.5- or second-order closure would produce more realistic results than a model with a first-order closure. However, in practice this is often not the case, because of complex model interactions and the difficulty of representing all the relevant details with sufficient accuracy (e.g., Steeneveld et al., 2006a,b). That is the reason why diagnostic approaches remain popular. Nevertheless, second-order equations are useful to gain insight in the governing physics, and after simplification useful extensions of the basic parameterizations may be achieved.

The atmospheric model equations can also be applied on much smaller spatial and temporal scales than discussed here, for instance by using vertical and horizontal grid elements of a few to typically 100 m, and time steps of seconds only. It is important to realize that in such cases a significant part of the turbulent fluctuations are resolved by the model equations. This type of modeling is known as 'large-eddy simulation (LES)'. This has become a powerful and popular tool in the last decade to study turbulence in clear and cloudy boundary layers under well-defined conditions. It is important to realize that in the case of LES the simplifying assumptions leading to eqn [2] are normally not valid.

Model Intercomparison for Stable Conditions

As noted above, atmospheric models need to make an overall representation of the small-scale boundary layer and near surface processes. This appears to be more successful during daytime (e.g., Ek and Holtslag, 2004; Holtslag and Ek, 2005; Steeneveld et al., 2011) then during nighttime stable conditions over land (Vogelezang and Holtslag, 1996; Steeneveld et al., 2011; Svensson et al., 2011). The modeling of the SBL over land is rather complex because of the many different physical processes which are 'at work' in stable conditions (e.g., Mahrt, 1999). These small-scale processes are clear air radiation divergence, drainage flow, generation of gravity waves and shear instabilities, fog and dew formation, the occurrence of a low-level jet, and generation of discontinuous or intermittent turbulence (Van de Wiel et al., 2003, 2007). In addition, the phenomenology of stable atmospheric boundary layers is quite diverse, e.g., shallow and deep boundary layers with continuous turbulence through most of their depth, and on the other hand boundary layers with intermittent turbulence or even laminar flow.

The small-scale processes influence the vertical and horizontal exchange of quantities between the surface and the atmosphere (Holtslag et al., 2007), as well as the mixing in the atmosphere on a variety of scales. In addition, it is known that turbulent mixing in stratified flow has an inherent nonlinear character and may, as such, trigger positive feedbacks. These positive feedbacks, in turn, may cause unexpected transitions between totally different SBL regimes (e.g., Van de Wiel et al., 2003).

Having in mind the complexity, one should not be surprised that atmospheric models encounter large forecast errors for stable conditions. One strategy to improve model performance is to provide different models the same forecasting task and analyze which model descriptions are in favor for which atmospheric stability. Recently, such an

intercomparison of boundary-layer schemes for stable conditions was made within the GEWEX Atmospheric Boundary Layer Study. This GEWEX project aims to improve the understanding and the representation of the atmospheric boundary layer in regional and large-scale climate models (Holtslag, 2006; Holtslag et al., 2013). A rather simple case was selected as a benchmark to review the state of the art and to compare the skills of single column (1D) models (Cuxart et al., 2006) and large-eddy simulation models (Beare et al., 2006). In this case an SBL is driven by an imposed, uniform geostrophic wind, with a specified constant surface-cooling rate over (homogeneous) ice. The case is initialized with $\theta = 265$ K for $0 < z < 100$ and a lapse rate of 1 K per 100 m aloft.

It turns out that with the same initial conditions and model forcings, the models indicate a large range of results for the mean temperature and wind profiles. Figure 3 shows the mean profiles for several models after 9 h of constant surface cooling (sufficient to achieve a quasi-steady state). The variable results achieved are strongly related to the details of the boundary-layer mixing schemes (Cuxart et al., 2006). An important finding is that the models in use at operational weather forecast and climate centers (as depicted at the left hand side of Figure 3) typically allow for enhanced mixing resulting in too deep boundary layers, while the typical research models (at the right hand sides) show less mixing in more in agreement with the 'large-eddy simulation' results for this case (Beare et al., 2006).

Because of the enhanced mixing in weather and climate models, these models tend to show a too strong surface drag, too deep boundary layers, and an underestimation of the wind turning in the lower atmosphere (e.g., Svensson and Holtslag, 2009). At the other hand, by decreasing the mixing and surface drag, a direct impact on the atmospheric dynamics ('Ekman pumping') is noted (e.g., Beljaars and Viterbo, 1998). Consequently, cyclones may become too active, corresponding in too high extremes for wind and precipitation, etc.

Nonlocal Mixing Parameterizations

In the previous sections, we have dealt with the most popular turbulence parameterizations in use for modeling atmospheric boundary layers for weather, climate, and air quality. In the literature many more examples can be found of parameterization schemes and of comparison studies like the one of Figure 3. Here we continue our discussion with mixing parameterizations, which have been proposed for boundary layers with atmospheric convection. In such cases, the turbulent flux of a conserved quantity is typically not proportional to the local gradient alone as predicted by eqn [8]. In fact, in a large part of the ABL the mean gradients are small in conditions with dry convection, in particular for potential temperature (Figure 2). Then the fluxes depend mostly on the mixing characteristics of the large eddies across the ABL.

Theories are available, which have modified K -theory to allow for the influence of convection, for example, by including additional terms at the right hand side of eqn [8]. This reads as (e.g., Holtslag and Boville, 1993; Cuijpers and Holtslag, 1998)

$$\overline{w\epsilon} = -K_c \frac{\partial \overline{C}}{\partial z} + \overline{w\epsilon}_{nl}. \quad [14]$$

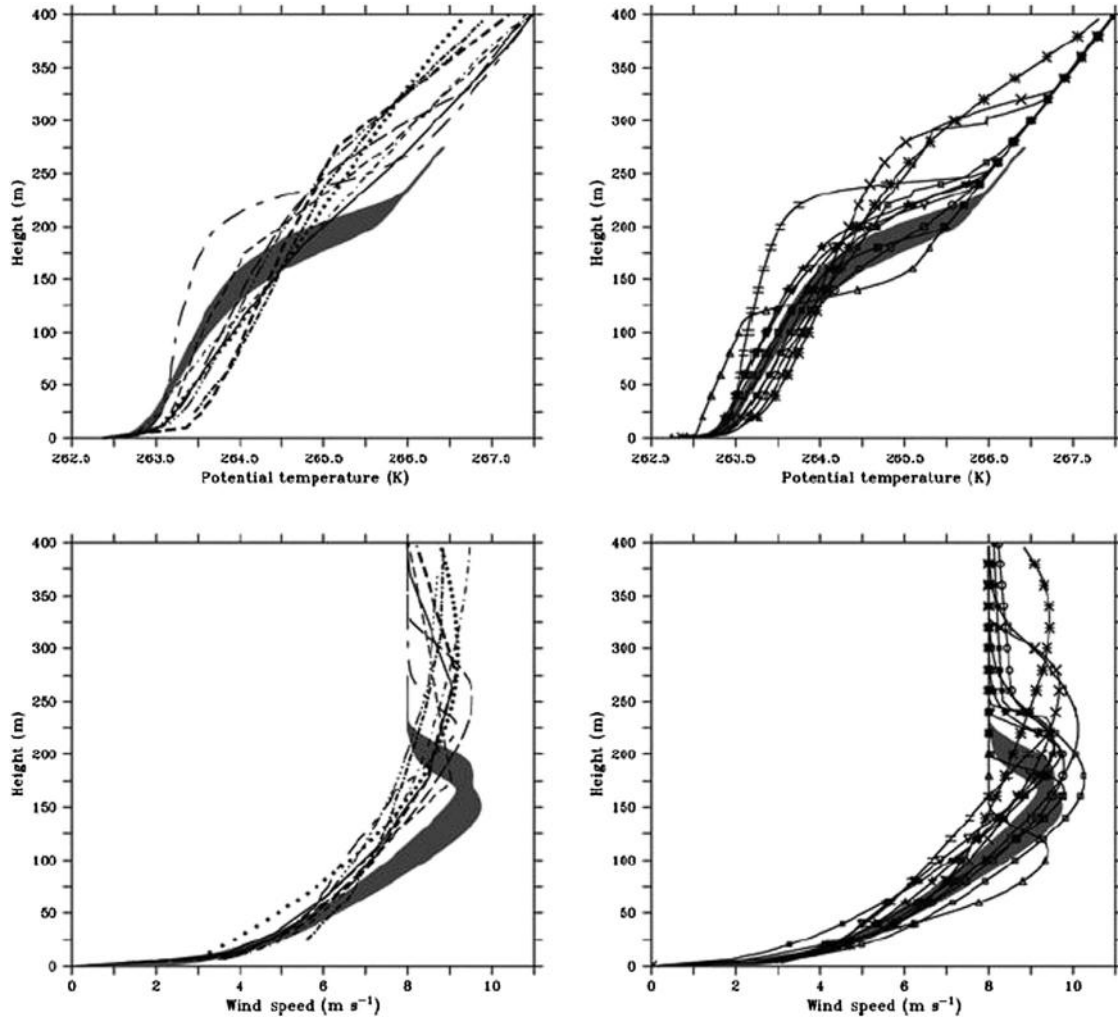


Figure 3 Modeled potential temperature and wind profiles by various single-column models (after 9 h of prescribed surface cooling starting all with the same initial profiles). Gray areas indicate the ensemble of large-eddy simulation results (Beare, R., MacVean, M., Holtslag, A., Cuxart, J., Esau, I., Golaz, J.-C., Jimenez, M., Khairoutdinov, M., Kosovic, B., Lewellen, D., Lund, T., Lundquist, J., McCabe, A., Moene, A., Noh, Y., Raasch, S., Sullivan, P., 2006. An intercomparison of large-eddy simulations of the stable boundary layer. *Boundary-Layer Meteorol.* 118, 247–272.). Left panel shows the results for operational (first-order closure) models and the right panel for research (higher-order closure) models. After Cuxart, J., Holtslag, A.A.M., Beare, R.J., Bazile, E., Beljaars, A., Cheng, A., Conangla, L., Ek, M., Freedman, F., Hamdi, R., Kerstein, A., Kitagawa, H., Lenderink, G., Lewellen, D., Mailhot, J., Mauritsen, T., Perov, V., Schayes, G., Steeneveld, G.J., Svensson, G., Taylor, P., Weng, W., Wunsch, S., Xu, K.-M., 2006. Single-column model intercomparison for a stably stratified atmospheric boundary layer. *Boundary-Layer Meteorol.* 118, 273–303.

Here $\overline{w\epsilon_{nl}}$ is the nonlocal flux representing the influence of the large-eddy mixing in the boundary layer with convection. The formulation of the latter is not so straightforward in the general case. The eddy-diffusivity K_c in eqn [14] is suitably modeled with the 1.5-order approach using the full kinetic energy equation, and a length-scale formulation, which may depend on the actual height and the depth h of the boundary layer.

In the case of a clear, stationary boundary layer dominated by dry convection, it appears that the eddy-diffusivity is also well described by a profile function (Holtslag and Moeng, 1991)

$$K_c = w_* h \left(\frac{z}{h} \right)^{4/3} \left(1 - \frac{z}{h} \right)^2, \quad [15]$$

where w_* is known as the convective velocity scale for the clear boundary layer. The latter is defined as

$$w_* = \left(\frac{g}{\Theta_v} \overline{w\theta_{v0}} h \right)^{1/3}, \quad [16]$$

where $\overline{w\theta_{v0}}$ is the surface sensible heat flux heating the boundary layer from below. For dry convection, the nonlocal flux correction term of eqn [14] is typically proportional to the surface flux of the variable involved. Then

$$\overline{w\epsilon_{nl}} = f(z/h) \overline{w\epsilon_0}, \quad [17]$$

where $f(z/h)$ is a dimensionless function of relative height. As an example, Figure 4 gives the results of two slightly different

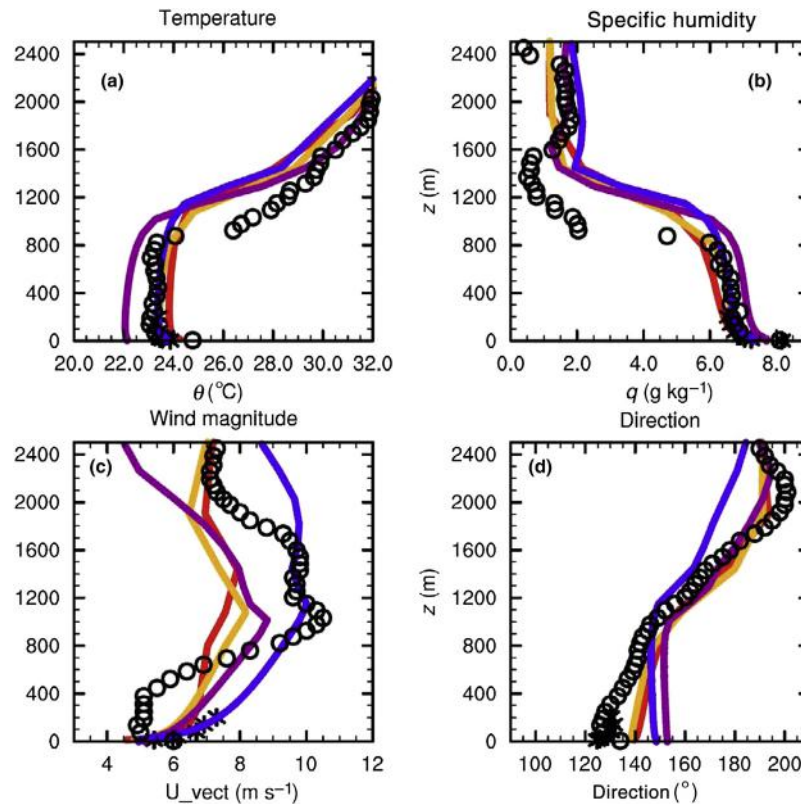


Figure 4 A comparison of observations made in the Netherlands on 12 June 2006 at the Cabauw tower (crosses) and by radio soundings in De Bilt (open circles) with two mesoscale model results each with two versions of a nonlocal boundary layer scheme (various colored lines) for temperature (a), specific humidity (b), wind magnitude (c), and wind direction (d). Results are taken from Steeneveld, G.J., Tol, L., Moene, A.F., Hartogensis, O.K., Peters, W., Holtslag, A.A.M., 2011. Confronting the WRF and RAMS mesoscale models with innovative observations in the Netherlands—evaluating the boundary-layer heat budget. *J. Geophys. Research. D Atmos.* 116, 16.

implementations of boundary layer schemes based on similar equations as eqns [15] and [17] within two atmospheric models in comparison with observations in the Netherlands (after Steeneveld et al., 2011).

Other nonlocal flux parameterizations are also discussed in the literature such as mass flux approaches. In a mass flux approach the turbulent layer is divided into updrafts and downdrafts, each with a typical concentration for the variable of interest. The updrafts and downdrafts can be defined on basis of the sign of vertical velocity or more restrictive definitions (in the case of clouds). Using such an approach, the updraft and downdraft variables are connected to a number of equations, which can be closed when the horizontal exchange ('lateral entrainment') of mass and other variables between the up and downdrafts is treated well (Siebesma and Holtslag, 1996; Neggers et al., 2004; Siebesma et al., 2007). It is noted that the mass flux approach has been quite successful for representing cumulus and stratocumulus in atmospheric models (Emanuel, 1994).

Summary

In this contribution an overview is given of the basic approaches for the modeling and parameterization of

turbulence in the atmospheric boundary layer. Emphasis has been given to the treatment of atmospheric turbulent mixing in models with a horizontal spatial grid distance of order 10–100 km, a vertical grid distance of 10–100 m, and resolving time scales of order 10 min or larger. The treated approaches are in use in atmospheric models for the forecasting and study of weather, climate, and air quality on various domains (from the mesoscale to regional and global domains). Some of the approaches can be adapted for turbulence above the boundary layer, such as turbulence in clouds or in elevated shear layers. However, important exceptions have been documented in the literature when so-called 'mesoscale processes' have their influence. This may occur, for instance, when extensive fields of stratocumulus are present or in the case of deep cumulus clouds and thunderstorms.

Generally still work needs to be done before we have a full understanding of the complexity of atmospheric turbulence and before we have a more unified treatment of turbulence on the different scales which occur in nature (Holtslag et al., 2013). A better understanding of atmospheric turbulence hopefully also contributes to our capability in refining and unifying the turbulence parameterizations for modeling of the atmospheric boundary layer in response to the different type of surfaces which are found in reality. The current text can only be

seen as a rather broad overview of the approaches in use. Details can be found in the items in the References as well as in the scientific literature.

See also: Boundary Layer (Atmospheric) and Air Pollution: Coherent Structures; Convective Boundary Layer; Surface Layer. Dynamical Meteorology: Balanced Flow; Potential Vorticity; Primitive Equations; Static Stability.

Acknowledgments

This contribution is based on related works of the author and discussions with the participants within the GEWEX Atmospheric Boundary Layer Study (GABLS) as well as colleagues at Wageningen University, in particular Dr Gert-Jan Steeneveld.

References

- Beare, R., MacVean, M., Holtslag, A., Cuxart, J., Esau, I., Golaz, J.-C., Jimenez, M., Khairoutdinov, M., Kosovic, B., Lewellen, D., Lund, T., Lundquist, J., McCabe, A., Moene, A., Noh, Y., Raasch, S., Sullivan, P., 2006. An intercomparison of large-eddy simulations of the stable boundary layer. *Boundary-Layer Meteorology* 118, 247–272.
- Beljaars, A.C.M., Holtslag, A.A.M., 1991. Flux parameterization over land surfaces for atmospheric models. *Journal of Applied Meteorology* 30, 327–341.
- Beljaars, A.C.M., Viterbo, P., 1998. Role of the boundary layer in a numerical weather prediction model. In: Holtslag, A.A.M., Duynkerke, P.G. (Eds.), *Clear and Cloudy Boundary Layers*. Published by Royal Netherlands Academy of Arts and Sciences, Amsterdam, 372 pp.
- Cuijpers, J.W.M., Holtslag, A.A.M., 1998. Impact of skewness and nonlocal effects on scalar and buoyancy fluxes in convective boundary layers. *Journal of Atmospheric Science* 55, 151–162.
- Cuxart, J., Holtslag, A.A.M., Beare, R.J., Bazile, E., Beljaars, A., Cheng, A., Conangla, L., Ek, M., Freedman, F., Hamdi, R., Kerstein, A., Kitagawa, H., Lenderink, G., Lewellen, D., Mailhot, J., Mauritsen, T., Perov, V., Schayes, G., Steeneveld, G.J., Svensson, G., Taylor, P., Weng, W., Wunsch, S., Xu, K.-M., 2006. Single-column model intercomparison for a stably stratified atmospheric boundary layer. *Boundary-Layer Meteorology* 118, 273–303.
- Ek, M.B., Holtslag, A.A.M., 2004. Influence of soil moisture on boundary layer cloud development. *Journal of Hydrometeorology* 5, 86–99.
- Emanuel, K.A., 1994. *Atmospheric Convection*. Oxford University Press, 580 pp.
- Garraff, J., 1992. *The Atmospheric Boundary Layer*. Cambridge University Press, 316 pp.
- Holtslag, A.A.M., 2006. GEWEX atmospheric boundary layer study (GABLS) on stable boundary layers. *Boundary-Layer Meteorology* 118, 243–246.
- Holtslag, A.A.M., Boville, B.A., 1993. Local versus nonlocal boundary-layer diffusion in a global climate model. *Journal of Climate* 6, 1825–1842.
- Holtslag, A.A.M., Duynkerke, P.G. (Eds.), 1998. *Clear and Cloudy Boundary Layers*. Published by Royal Academy of Arts and Sciences, Amsterdam, The Netherlands, 372 pp.
- Holtslag, A.A.M., Ek, M.B., 2005. Atmospheric boundary layer climates and interactions with the land surface. In: *Encyclopedia of Hydrological Sciences*. Wiley and Sons.
- Holtslag, A.A.M., Moeng, C.-H., 1991. Eddy diffusivity and counter-gradient transport in the convective boundary layer. *Journal of Atmospheric Science* 48, 1690–1698.
- Holtslag, A.A.M., Steeneveld, G.J., van de Wiel, B.J.H., 2007. Role of land-surface feedback on model performance for the stable boundary layer. *Boundary-Layer Meteorology* 125, 361–376.
- Holtslag, A.A.M., Svensson, G., Baas, P., Basu, S., Beare, B., Beljaars, A.C.M., Bosveld, F.C., Cuxart, J., Lindvall, J., Steeneveld, G.J., Tjernstrom, M., Wiel, B.J.H., van de, 2013. Stable atmospheric boundary layers and diurnal cycles – challenges for weather and climate models (Online first). *Bulletin of the American Meteorological Society*. <http://journals.ametsoc.org/doi/abs/10.1175/BAMS-D-11-00187.1>.
- Mahrt, L., 1999. Stratified atmospheric boundary layers. *Boundary-Layer Meteorology* 90, 375–396.
- Nieuwstadt, F.T.M., van Dop, H. (Eds.), 1982. *Atmospheric Turbulence and Air Pollution Modelling*. Reidel, Dordrecht.
- Neggers, R.A.J., Siebesma, A.P., Lenderink, G., Holtslag, A.A.M., 2004. An evaluation of mass flux closures for diurnal cycles of shallow cumulus. *Monthly Weather Review* 132, 2525–2538.
- Pielke Sr., R.A., 2001. Parameterization of subgrid scale fluxes. In: *Mesoscale Meteorological Modeling*, second ed. Academic Press.
- Siebesma, A.P., Holtslag, A.A.M., 1996. Model impacts of entrainment and detrainment rates in shallow cumulus convection. *Journal of Atmospheric Science* 53, 2354–2364.
- Siebesma, A.P., Soares, P.M.M., Teixeira, J., 2007. A combined eddy diffusivity mass flux approach for the convective boundary layer. *Journal of Atmospheric Science* 64, 1230–1248.
- Steeneveld, G.J., van de Wiel, B.J.H., Holtslag, A.A.M., 2006a. Modeling the arctic stable boundary layer and its coupling to the surface. *Boundary-Layer Meteorology* 118, 357–378.
- Steeneveld, G.J., van de Wiel, B.J.H., Holtslag, A.A.M., 2006b. Modeling the evolution of the atmospheric boundary layer coupled to the land surface for three contrasting nights in CASES-99. *Journal of Atmospheric Science* 63, 920–935.
- Steeneveld, G.J., Tolk, L., Moene, A.F., Hartogensis, O.K., Peters, W., Holtslag, A.A.M., 2011. Confronting the WRF and RAMS mesoscale models with innovative observations in the Netherlands-evaluating the boundary-layer heat budget. *Journal of Geophysical Research: Atmosphere* 116, 16.
- Stull, R.B., 1988. *An Introduction to Boundary-Layer Meteorology*. Kluwer Academic Publishers, Dordrecht, 666 pp. (reprinted 1999).
- Svensson, G., Holtslag, A.A.M., 2009. Analysis of model results for the turning of wind and the related momentum fluxes in the stable boundary layer. *Boundary-Layer Meteorology* 132, 261–277.
- Svensson, G., Holtslag, A.A.M., Kumar, V., Mauritsen, T., Steeneveld, G.J., Angevine, W.M., Bazile, E., Beljaars, A., de Bruijn, E.I.F., Cheng, A., 2011. Evaluation of the diurnal cycle in the atmospheric boundary layer over land as represented by a variety of single-column models: the second GABLS experiment. *Boundary-Layer Meteorology* 140 (2), 177–206.
- Tennekes, H., Lumley, J.L., 1982. *A First Course in Turbulence*, second ed. MIT Press, Cambridge. 300 pp.
- Van de Wiel, B.J.H., Moene, A.F., Hartogensis, O.K., de Bruin, H.A.R., Holtslag, A.A.M., 2003. Intermittent turbulence and oscillations in the stable boundary layer over land, Part III: a classification for observations during CASES99. *Journal of Atmospheric Science* 60, 2509–2522.
- Van de Wiel, B.J.H., Moene, A.F., Steeneveld, G.J., Hartogensis, O.K., Holtslag, A.A.M., 2007. Predicting the collapse of turbulence in stably stratified boundary layers, turbulence. *Flow Turbulence and Combustion* 79, 251–274.
- Vogelezang, D.H.P., Holtslag, A.A.M., 1996. Evaluation and model impacts of alternative boundary-layer height formulations. *Boundary-Layer Meteorology* 81, 245–269.

Further Reading

- Holtslag, A.A.M., Duynkerke, P.G. (Eds.), 1998. *Clear and Cloudy Boundary Layers*. Published by Royal Academy of Arts and Sciences, Amsterdam, The Netherlands, 372 pp.
- Stensrud, D.J., 2009. *Parameterization Schemes: Keys to Understanding Numerical Weather Prediction Models*. Cambridge University Press, 480 p.
- Wyngaard, J.C., 2010. *Turbulence in the Atmosphere*. Cambridge University Press, 406 p.

Observational Techniques *In Situ*

EF Bradley, CSIRO Land and Water, Canberra, ACT, Australia

© 2003 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 1, pp 280–290, © 2003, Elsevier Ltd.

Introduction

We usually refer to *in situ* measurements as those made from some platform anchored to the surface, or in the case of the deep ocean from a ship or buoy. Instruments located on the surface but sensing the atmospheric boundary layer (ABL) remotely are not included. Over land the ABL reaches a height of between 1 and 3 km during the day, collapsing to a few hundred meters at night; the marine boundary layer is typically 500 to 1000 m day and night. The region sampled by *in situ* techniques is mainly in the atmospheric surface layer (ASL). This corresponds roughly to the lowest 10% of the ABL, and its characteristics are influenced strongly by the nature of the surface.

In situ measurements provide data for operational weather forecasting, environmental monitoring, validation for models and remotely sensed observations, and various research topics. The latter include studies of atmospheric structure and transport processes, agricultural and forest meteorology, pollution and climate change. Most applications require measurements of the state variables (temperature, humidity, and pressure), wind speed and direction, incoming and outgoing radiant energy, rainfall, and evaporation. Determination of the fluxes of sensible and latent heat, and of momentum, is also a critical objective of many *in situ* measurement programs.

This article first outlines the theoretical framework of the ASL, which often influences the observations required, and then the sensors commonly used to make the measurements. This is followed by descriptions of a few different types of field site and some practical aspects of deploying instruments and recording the data.

The Atmospheric Surface Layer

To set atmospheric transfer processes within a framework of classical fluid mechanics and thermodynamics, the starting point has been a flat and uniform earth's surface. Many experiments have taken place over such 'ideal' or one-dimensional sites. This article provides a brief and necessarily simplified account of the results, to introduce the quantities whose measurement is considered here. Full details can be found in the works listed under Further Reading.

The defining characteristics of the ASL are its buoyant stability, the vertical variation (or 'profiles') of wind speed, temperature, humidity, and the related vertical fluxes. The latter (so-called eddy fluxes) are given by

$$\tau = -\overline{\rho w' u'} \quad \text{(the flux of momentum, or surface stress),} \quad [1a]$$

$$H = \rho C_p \overline{w' \theta'} \quad \text{(the flux of sensible heat), and} \quad [1b]$$

$$E = \overline{\rho w' q'} \quad \text{(the flux of latent heat, or evaporation)} \quad [1c]$$

ρ and C_p are density and isobaric specific heat of air, w', u', θ' and q' are fluctuations from the mean of vertical and horizontal wind, and of temperature and specific humidity. The overbar indicates a time average of the particular product. Within the ASL over a uniform surface, the fluxes may be assumed constant with height.

The most common measure of thermal stability is based on the relative importance of mechanical and buoyant forces in production of turbulence kinetic energy (TKE); i.e. some function of the ratio of eqn [1a] divided by eqn [1b]. Defining a scaling velocity $u_* = (\tau/\rho)^{1/2}$ and a buoyancy parameter g/T , where g is gravity and T absolute temperature, these form a length scale $L = -u_*^3 / [\kappa(g/T)w'\theta']$; κ is a constant (≈ 0.4). The Monin–Obukhov similarity theory (MOST) adopts the hypothesis that, at height z in the boundary layer, flux–gradient relationships and turbulence parameters are universal functions of z/L . These MOST assumptions have been verified by many field experiments.

Wind increases with height in the surface layer while temperature decreases during daytime and increases at night (the nocturnal inversion) in response to the sign of the surface heat flux. MOST has nondimensional forms for the vertical wind and temperature gradients:

$$\phi_m = \frac{\kappa z}{u_*} \frac{\partial \bar{u}}{\partial z} \quad \text{and} \quad \phi_h = \frac{\kappa z}{\theta_*} \frac{\partial \bar{\theta}}{\partial z} \quad [2]$$

respectively, where ϕ_m and ϕ_h are functions of z/L , and $\theta_* = \overline{w'\theta'}/u_*$ is a scaling temperature. As an example, the variation of wind speed with height is given by integrating the first of these gradient equations, to obtain

$$\bar{u}(z) = (u_*/\kappa) [\ln(z/z_0) - \psi_m] \quad [3]$$

The integration constant, z_0 , is known as the surface roughness length, and ψ_m is a stability function. It follows that the wind variation with height is roughly logarithmic (exactly so in neutral conditions), and this is also true of the scalar quantities, $\bar{\theta}$ and $\bar{q}$. Finally, we note that the relationship between fluxes and gradients can be expressed in terms of a turbulent diffusivity,

$$F_\chi = -K_\chi \frac{\partial \chi}{\partial z} \quad [4]$$

where F_χ is the flux of constituent χ (momentum, heat, or a trace gas) and K_χ is its turbulent exchange coefficient, obviously a function of z/L .

Atmospheric Sensors

General Characteristics

There are often several choices of sensor for each variable, the most suitable for a particular application depending on several factors, including the accuracy and resolution required, frequency response, and overall convenience of operation. Sensors evolve continuously in the research environment,

either testing new physical principles of measurement or else to quantify some newly significant entity (e.g., a 'greenhouse' gas).

Considerations of frequency response highlight two categories of sensor, which we may call slow and fast. Atmospheric variables fluctuate on time scales from below 0.1 s to several hours. Time-averaging (over 15–30 min) is often required to reduce unsteadiness, so slow-response sensors are adequate to determine basic mean variables, such as $\bar{u}$, $\bar{q}$, or $\bar{\theta}$, and their vertical profiles. A sensor responds to a step change exponentially, the time taken to reach $(1 - 1/e; \sim 0.632)$ of the final value being its time response. A slow sensor may have a time response of many seconds. Fast sensors are required when the turbulent fluctuations themselves are of interest, either for studies of atmospheric structure or to determine the surface fluxes by measuring w' , u' , θ' , q' in eqn [1]. For this, frequency response of at least 10 Hz is needed.

The following sections describe first the most common slow-response sensors for atmospheric variables, and then some fast-response instruments. Figure 1 shows an array of instruments mounted at the top of a 30 m tower, part of the Ozflux network (see below). Sensors described in the text are identified by their letter.

Temperature

Atmospheric temperature can be measured with mercury-in-glass thermometers (still used operationally by weather observers), platinum resistance thermometers (PRTs), thermistors, and thermocouples. The last three lend themselves to automatic data logging. PRTs are very stable, and with careful calibration can achieve an accuracy of about 0.01°C. Thermocouple systems have low output voltage, and for absolute measurement require a reference 'cold' junction. Thermistors are semiconductor devices with higher sensitivity to temperature changes than either of the above, but at the expense of

non-linearity and greater self-heating than platinum elements. Formerly, they were prone to uncertainties of stability and calibration, but interchangeability of $\pm 0.1^\circ\text{C}$ is now possible.

Both thermocouples and PRTs can be easily configured for differential measurement, which can improve accuracy when measuring profiles or the wet-bulb depression of a psychrometer (see the next section). To reduce errors from solar radiation, these thermometers are ideally mounted inside a double-wall heat-reflecting shield and have ambient air drawn over them (A in Figure 1).

Humidity

Atmospheric humidity is variously specified by the partial pressure of water vapor (e , in hPa), vapor density ($g\ m^{-3}$), specific humidity (q , g/g of moist air), or relative humidity ($RH = 100e/e_s$). The relationship between them can be found in standard texts. e_s is the liquid water saturation vapor pressure at air temperature, given approximately by the empirical equation

$$e_s = 6.106\{\exp[17.273T/(T + 237.3)]\} \quad [5]$$

where T is Celsius temperature. At a particular ambient humidity, reducing air temperature reaches a point on the $T - e_s$ curve (eqn [5]) where the air is saturated. This gives us the principle of the dew point hygrometer. Its central component is a mirror maintained, by optical and electronic feedback, at the temperature where moisture or ice just condense on its surface. It is often considered an absolute instrument, and used as a secondary standard to other sensors.

The traditional instrument for atmospheric humidity measurement is the psychrometer, consisting of a pair of thermometers, one being covered with a moist wick. Air drawn over the thermometers evaporates the moisture, cooling the wick until the evaporation rate is in equilibrium with the vapor content of the air. For given humidity this wet-bulb depression,



Figure 1 An array of meteorological research instruments at the top of a 30 m tower, part of the 'Ozflux' long-term climate monitoring network in Australia. Photo: © Frank Bradley.

$(T - T_w)$, is known from thermodynamic theory, leading to the psychrometer equation. The form specified by the World Meteorological Organization (WMO) for the Assman psychrometer (a particular design of instrument with mercury-in-glass thermometers) is

$$e = e_s - 0.000\,653 \times (1 + 0.000\,944T_w)(T - T_w)P \quad [6]$$

where e , e_s , and atmospheric pressure, P , are in the same units. With PRTs in a differential bridge, $\pm 0.01^\circ\text{C}$ is possible for $(T - T_w)$. It is crucial to provide adequate airflow over the thermometers to ensure full depression. With care, accuracy of 0.05 g kg^{-1} is possible. For Assman psychrometers, 0.20 g kg^{-1} is more realistic.

Thin-film polymers which absorb or desorb water as the relative humidity changes afford a simpler humidity sensor. The polymer forms the dielectric of a capacitance in a circuit which provides an output proportional to relative humidity. Early instruments often failed at very high humidity, but recent developments have overcome this problem and have greatly improved their accuracy and stability. Conversion to mixing ratio, specific or absolute humidity requires the air temperature around the dielectric, usually measured with a colocated PRT. These temperature/RH sensors are screened and ventilated in the same way as a psychrometer. The best accuracy quoted is around $\pm 0.3\%$ RH (or $\pm 0.05\text{ g kg}^{-1}$ at 20°C and 70% RH).

Wind Speed and Direction

For average wind speed and/or direction over some time period, cup (or propeller) anemometers and wind vanes are usually the most convenient. Operational designs must withstand continuous exposure to stormy conditions, but there are also 'sensitive' instruments intended for research work. Apart from mechanical strength, the difference is reflected in their starting speed and distance constant (response time converted to run of wind). A sensitive cup anemometer will start from rest in a breeze of 0.3 m s^{-1} and have a distance constant less than 1 m .

For best accuracy (typically 1%) cups must be calibrated individually, although calibration in the steady horizontal flow of a wind tunnel can lead to uncertainty. In a gusty wind, cup anemometers overestimate for two reasons: the rotor responds more quickly to an increasing wind than to the reverse, and, in a wind gust with a vertical component, shielding by the upwind cup is reduced. A propeller has poor 'cosine' response (to off-axis wind direction), but the error is usually minimized by mounting it on the front of a wind vane. A cup-anemometer-wind-vane pair are often mounted at opposite ends of a horizontal bar (B and C in Figure 1).

Pressure

A knowledge of mean atmospheric pressure in the ASL is often needed, in eqn [6] for example, to calculate air density in eqn [1], and to convert between the various definitions of humidity. Pressure varies with elevation above sea level and slowly with synoptic changes. For most purposes, the pressure reported at 3-hourly intervals by the nearest weather forecast station is adequate. The WMO target accuracy for pressure measurement is $\pm 0.1\text{ hPa}$. It is read from either the traditional mercury column barometer or an accurate aneroid instrument.

Rainfall

Rainfall, particularly during convective storms, is perhaps the 'patchiest' of all meteorological variables. Single point measurements are generally less relevant than area averaged values, which may be estimated over land with a network of surface rain gauges. At sea this is not feasible, and while spatial rainfall patterns may be obtained by shipmounted radars these must be calibrated by a surface measurement.

Traditional rain gauges measure the rain falling into a funnel of known area. For automatic recording either a weighing system is used or a tipping bucket rain gauge. In this, the funnel discharges to a pair of buckets in a seesaw arrangement which flips over at every 0.1 mm of rainfall. Another system, often used on ships, has a reservoir which fills to its capacity (about 50 mm of rain), when it siphons automatically and starts filling again. An electronic sensor keeps track of the level of water in the reservoir. All funnel gauges lose catch in strong winds, when the gauge deflects airflow so that raindrops are carried past the funnel. The siphon gauge also misses rain while the instrument is siphoning.

Optical rain gauges (ORGs) measure rainrate by detecting raindrops falling through an optical path. One system measures extinction of a light beam by the raindrops; another measures the intensity of scintillations caused by raindrops passing through the beam from a light-emitting diode. Rainfall amount is obtained by integrating the rainrate. ORGs must be calibrated against a funnel gauge in natural or simulated rainfall.

Disdrometers are primarily intended for the measurement of drop size and drop distribution in rainfall. The most usual is an acoustic device which converts the sound of impact of raindrops hitting the sensor surface into an electrical signal related to the size of the drop. Continuous recording of the size and number of drops provides a time series of rainrate and total rainfall by integration.

The above rain gauges can handle rainrates to around 200 mm h^{-1} . This would be an extreme tropical storm; a heavy rainstorm in midlatitudes might produce instantaneous rain rates of 50 to 100 mm h^{-1} , but more commonly rain rates over land are between 1 and 20 mm h^{-1} . They are generally unsuitable for the measurement of precipitation falling as snow. However, because of its importance in the hydrological cycle in many regions, it is usual to estimate the water equivalent of snowfall, for example by measuring the average depth of fresh snow cover and its density.

Trace Gas Measurement

Measurement of minor constituents of the atmospheric boundary layer is increasingly important for studies of climate change (the 'greenhouse effect'), and monitoring of air quality. Applications include measuring emissions from engine exhausts and refineries, industrial stacks, agricultural emissions, and vegetation to atmosphere CO_2 exchange. Species of interest include CO , CO_2 , HF , N_2O , NO , NO_2 , CH_4 , SO_2 , NO_x , NH_3 , and other hydrocarbons and organic compounds. Accurate determination of both concentrations and fluxes are required, the former for regulatory purposes and the latter for studies of the emission process.

Chemical analysis and gas chromatography were some of the earliest techniques used, but are not convenient for

automatic recording. Infrared gas analyzers (IRGAs) have also been used for absolute measurement of CO₂ and some other atmospheric trace gases. The IRGA measures differential infrared absorption between a pair of cells containing the air sample and a reference gas (or between separate air samples for gradient determination). Originally intended as bench instruments, they are sensitive to vibration and in the field must be installed in a shelter with air samples drawn in through tubes.

Spectrographic methods have also been introduced, the most common being Fourier transform infrared spectroscopy (FTIR). Essentially, the instrument is a scanning interferometer which measures infrared spectra at high resolution over a broad spectral range, measuring all frequencies in the signal simultaneously, and resolving those required with the Fourier transform. Also developed as a laboratory technique, it has been adapted for field use, able to monitor species at very low concentrations (parts per billion of volume). Two configurations are employed: either an open atmospheric path is used, often folded to lengths up to 1 km, or the air is drawn through tubes to an internal measurement cell. Advantages of the open path are that it is noninvasive and that it affords spatial averaging. The internal system permits better temperature and pressure control, and reference to calibration spectra.

Trace gas profiles can be obtained by drawing air continuously from several levels through separate tubes, switched with valves sequentially through either the IRGA or FTIR. With the gradient of a particular constituent determined this way, its flux can be inferred from eqn [4], estimating the exchange coefficient from flux and gradient measurements of some easier quantity (e.g. sensible heat).

Fast-Response Instruments

Fast-response instruments are needed for studies of the turbulent structure of the boundary layer, and for eddy flux measurement. For the wind components, heated wire or film sensors have been used in the field, but poor stability limits their accuracy. Nowadays the usual sensor for turbulent wind measurement is the sonic anemometer (sonic for short), which obtains the wind component, V , along a fixed path length, d , by measuring the transit time of acoustic signals traveling in opposite directions (D in Figure 1). Then $V = c^2(t_2 - t_1)/2d$, where c is the velocity of sound in air. The earliest sonics determined this time difference directly or by measuring phase shift, but had the weakness that c^2 depends on temperature and humidity. Most instruments now use an alternative expression, $V = d(1/t_1 - 1/t_2)/2$. Different designs use various arrangements of the three paths needed to determine the total wind vector. Frequency response depends on the size of the sensing volume, a compromise between the conflicting needs for compactness and minimal flow distortion; d is typically 10–20 cm.

An internal microprocessor determines t_1 and t_2 for each path, and outputs wind components on orthogonal axes. Separate determination of t_1 and t_2 also enables fluctuations in virtual temperature to be obtained from the speed of sound. Then momentum and heat fluxes are calculated with eqns [1a] and [1b]. Temperature fluctuations can also be measured independently with a fine wire, usually of platinum, mounted in the yoke of the sonic; however, it is vulnerable to damage from rain and wind-blown debris.

Measurement of humidity fluctuations are usually based on absorption of either infrared (IR) or ultraviolet (UV) radiation by water vapor. IR instruments usually have a broadband source separated from a detector by an open path of physical length similar to that of the sonic vertical path. A rapidly rotating chopper wheel alternately introduces a pair of narrow pass interference filters into the beam, one at a strong H₂O absorption line (e.g. 2.61 μm) and the other at a nearby wavelength (e.g. 3.69 μm) of negligible absorption. This 'differential' scheme helps compensate for instabilities in the system, such as variations in source intensity. Some IR instruments add another filter at a CO₂ absorption line (e.g. 4.22 μm), to obtain both fluxes simultaneously (E in Figure 1).

The UV instruments detect absorption at the Lyman- α emission line of atomic hydrogen (121.56 nm) or from krypton (116.49 and 123.58 nm). These lines are strongly absorbed, providing good signals with a separation of only 1 cm between the source and the detector. Both types have long-term stability problems, but these are not serious for the measurement of fluctuations over periods like 15–30 minutes. Neither instrument is normally used for absolute measurement. The humidity sensor must be mounted as close as practicable to the sonic without increasing flow distortion. Formulae are available to correct for loss of correlation by lateral separation of the two instruments.

Fluctuations of static pressure are of interest in turbulence studies, where correlations like $\overline{w'p'}$ represent the transport of TKE by pressure fluctuations. p' is difficult to measure because any probe inserted in the flow may produce dynamic pressure fluctuations which mask those of static pressure. Instruments for this measurement remain research prototypes.

In principle, if a suitable infrared absorption line and fast-response sensor can be found, trace gas fluxes can be measured by the eddy flux method. Some IRGAs respond rapidly enough to measure eddy fluxes, with allowance for the time lag and frequency damping introduced in the air lines. When measuring trace gases, CO₂ and to some extent H₂O by eddy flux, account must be taken of fluctuations of density due to the presence of heat fluxes (the so-called 'Webb effect'). The corrections can be of the same order as the measured flux.

Radiation Sensors

General

The radiant energy flux to and from the earth's surface comprises the solar (or short-wave) component in the wavelength band 0.3 to 3 μm , and the terrestrial (long-wave or infrared) component from 3 to around 50 μm . They are measured with a pyranometer and a pyrgeometer respectively. These instruments are physically similar, both accepting broadband radiation through a hemispherical dome of appropriate spectral transmissivity.

Radiation observations can be conveniently discussed in the context of the balance of heat energy at the surface over flat terrain. If R_S and R_L are downwelling short-wave and long-wave radiative energy (W m^{-2}), T_S (K) is surface temperature, G_S is heat flux into the surface, and H and λE the turbulent fluxes,

$$(1 - \alpha)R_S + \varepsilon(R_L - \sigma T_S^4) - G_S = H + \lambda E \quad [7]$$

where α and ε are surface albedo and emissivity, σ the Stefan-Boltzman constant, and λ the latent heat of vaporization.

Short-Wave Radiation

Downwelling short-wave radiation at the surface has a component due to the direct solar beam, and a diffuse component scattered from atmospheric constituents and reflected from clouds. Upwelling short-wave radiation comes from reflection at the surface. Both can be measured with the pyranometer, facing either upward or downward (F and G in Figure 1); their ratio is the surface albedo. The pyranometer sensor is a blackened horizontal surface on which the radiation falls, bonded to a thermopile whose reference junction is the instrument body. Accuracy for the instrument is usually quoted as 2%. The direct solar beam can be measured with a pyrreheliometer. These are more accurate, but are generally unsuitable for operational use because they need to track the sun. Their main use is as calibration standards for working instruments.

The diffuse component is obtained with a 'shadow-band', set to shield the sensor from the direct solar beam. The position of the band is changed manually to follow the annual variation of solar elevation. Instruments have also been built with a rotating shadow band which alternately shields and exposes a fast-response radiation sensor, such as a solid state photocell. This system is less accurate, but useful on ships or other moving platforms.

Long-Wave Radiation

Pyrgeometers work by determining the thermal balance of the instrument itself. So as well as the thermopile output, the case and dome temperatures (T_C and T_D) have also to be measured. The long-wave energy passing through the dome is then derived from the pyrgeometer equation

$$R_L = V/s + \sigma T_C^4 + B\sigma(T_C^4 - T_D^4) \quad [8]$$

where V is thermopile output, s its sensitivity, and B an empirically determined constant (~ 3.5). Precision of the instrument using this equation is about 1.5% of the total long-wave flux. The third term on the right represents the effect of dome heating, and the second is the radiative flux contribution from the case of the instrument. The latter is usually the largest term, and is sometimes simulated with an internally generated voltage so that only one signal has to be recorded, instead of three. This, and the necessary assumption that $T_C = T_D$, increases uncertainties to at least 5%.

Net Radiation

Often the net radiant energy (all wavelengths) into or out of the earth's surface is required. This is the algebraic sum of the first two terms of eqn [7]. It can be measured using two pyranometers and two pyrgeometers, or more simply with a net radiometer. This consists of a pair of blackened plates, bonded to the opposite junctions of a thermopile to form a sensor about 25 mm square and 6 mm thick. It is protected by a pair of thin polythene domes, transparent to both short and long wavelengths, kept inflated with a flow of dry air. The symmetrical arrangement is mounted with the sensor

horizontal, measuring the temperature difference produced between the plates by downwelling and upwelling radiative energy (H in Figure 1). The quoted 5% accuracy is poorer than each of the individual instruments, but usually better than their combined uncertainties.

Surface Heat Flux

In eqn [7], G_s is roughly 10% of the net radiation over land and should not be ignored. It is usually measured with a soil heat flux plate, an encapsulated thermopile similar to the net radiometer sensor, buried horizontally at a shallow depth. Heat flux through the plate, assumed to be related to that in adjacent soil, generates a temperature difference between the opposite faces.

Absorption of heat into a water body is more complicated, because incoming short-wave radiation is distributed in depth with an extinction coefficient which depends on turbidity and wavelength. It can penetrate tens of meters into clear water. Incoming long-wave is absorbed within the top millimeter or so. Equation [7] may require additional terms for horizontal and vertical transport of heat (advection) by motion in the water.

Surface Temperature

The surface temperature, T_s , is usually measured with a narrow field of view IR radiometer in the 8–12 μm band. In daytime over land surfaces, T_s will normally vary over quite small distances because of variations in surface cover and topography. Thus, except where the surface is flat and uniform (e.g. bare soil, mown grass) a radiometer measurement from a height of even tens of meters will not be representative of the region. Radiometric measurements of sea surface temperature are more successful, because of the uniformity produced by mixing, and are particularly valuable for the validation of radiometers flown on satellites and aircraft. Corrections must be made for reflected sky radiation, measured with a second radiometer facing skywards, and the emissivity of the water at the particular IR wavelength must be known.

Types of Field Site

General

The physical nature of an experimental site, as well as the purpose of the observations, has a bearing on the instruments and methodology adopted. This section presents some characteristics of various field sites, and the consequences for useful measurement. We envisage midlatitude conditions; the same basic sensors and procedures are used in tropical and polar regions, perhaps modified to suit the extreme conditions encountered.

Flat, Uniform Landscape

As mentioned above, a theoretical framework for the one-dimensional ASL was derived from measurements at 'ideal' sites. These were located in vast regions with little topographic relief (Australian plains, the US Midwest, the Russian steppe),

and the actual observing site was a substantial area of completely uniform surface with short vegetation. Upwind changes in vegetation or soil moisture were avoided. A rule of thumb states that surface uniformity should extend upwind at least 100 times the height of measurement, with no trees or obstacles for many kilometers. Field campaigns in the 1960s rarely used towers higher than about 30 m, so the required fetch was of order 3 km. If application of the results from these experiments to another site is to be strictly valid, it should conform to the same criteria. In practice, useful results may be obtained from less favorable sites, provided the investigator is aware of the implications, and takes appropriate measures. For example, at the field site shown in Figure 2, wind blowing towards the camera would have both profiles and turbulence strongly perturbed by the belt of trees in the background. However the fetch to the left of the picture was clear for some kilometers, so data from that direction were acceptable. The sensor supports face that way and the equipment shelter is located well downwind of the towers.

Plant Canopies and Forests

Some of the earliest studies of the atmospheric surface layer were associated with agriculture and the relationship between climate and plants. Over short ground cover like pasture, the origin of profiles is the surface itself. For tall crops such as corn, orchards and forests, the concept of a zero displacement height, d , evolved, being the level near the top of the plant canopy where the neutral logarithmic wind profile has its origin. In eqn [3], the height z must be replaced by $z - d$. The value of d depends on the structure of the crop, but is usually about $\frac{3}{4}$ of its height.

Within the canopy, branches and leaves are sources and sinks for momentum, heat, moisture and CO_2 , which govern

the turbulent exchange processes. To obtain representative profiles of vertical fluxes and concentrations inside the canopy, measurements must be averaged in space as well as time. Canopy profiles merge with the boundary layer profiles near the top of the crop.

Sensors of the same type are used within the canopy as above, but are often smaller. Because of the three-dimensional wind environment, sonic anemometers are preferred over cups. Solar radiation penetrates the canopy and is absorbed and scattered at plant surfaces. Spatially averaged radiation profiles are sometimes obtained by moving the instruments along tracks at various levels through the vegetation. A wide range of specialized sensors have been developed for detailed study of canopy processes, energy exchange, water use, CO_2 uptake, and stomatal conductance, down to the scale of individual leaves.

Nonuniform Terrain

In the real world, the obstacles and surface nonuniformity so carefully avoided around 'ideal' sites are the normal condition, stimulating efforts to establish general rules to describe such complex terrain. Models have been developed for the two-dimensional internal boundary layers which grow downstream from a change in surface, and a sequence of such changes like the checkerboard pattern of agricultural land. There has also been some success in modeling wind flow over one- and two-dimensional hills.

Over such topography, observational problems escalate. A uniform site can be represented by a single vertical profile of instruments. To document the boundary layer over complex terrain requires many measuring sites and more sets of instruments. Their number, height and location depends on the



Figure 2 A micrometeorological research site in south-eastern Australia, set up by CSIRO Land and Water. The towers carry instruments to measure the profiles of wind, temperature, and humidity, the components of radiation, and eddy fluxes of heat, moisture and CO_2 , as part of a regional climate study. The tallest tower is 22 metres high; a shelter for recording equipment is seen to the right of the picture. Photo: © Greg Heath.

particular aims of the study, and must be to some extent intuitive. For example, is it enough to sample the extremes of the site (a hilltop and a valley), or must we also investigate sites in between?

Some guidance may be gained from simple modeling of the flow field. This, and the analysis of data from hilly sites, is often performed in a coordinate framework based on the streamlines of the flow. At any point in this curvilinear system, the x axis is along the local mean velocity vector, with the other two axes normal to it and to one another.

As the physical scale of topography increases, surface influence propagates higher into the ABL, beyond the range which can be studied exclusively with *in situ* techniques. This is where the combined strengths of *in situ*, remote sensing, and aircraft operations become important.

The Ocean

The ocean surface is notably horizontal, uniform and homogeneous. One-dimensional assumptions, turbulence theory, and the MOST relationships apply to the marine boundary layer. However, exchange between ocean and the atmosphere involves additional physical processes associated with the waves, surface currents, and heat transfer in the water

Most of the sensors described above are used at sea, but designed to survive the more hostile conditions. Wind, temperature, and rainfall measurements can be impaired by the presence of the ship. When practical, meteorological sensors are mounted on a jackstaff at the bow, in relatively clear air and ahead of spray from the ship's bow wave, or high on the mainmast with instruments duplicated on port and starboard sides. Psychrometers are seldom used because of the frequent need to wash salt from the wick. More usually, temperature and humidity are measured with the thin-film polymer package. Radiation instruments are mounted as high as possible, preferably in gimbals. The domes require daily washing to remove salt and particulates from the ship's exhaust. Upwelling radiation components are calculated from sea surface temperature, emissivity, and albedo.

Eddy fluxes are measured with the usual fast-response instruments, the sonic wind signals being corrected for ship motion measured with an inertial navigation system (INS), containing accelerometers, rate sensors, gyros, etc. The inertial dissipation technique of flux measurement (see [Further Reading](#)) is also used at sea, being less affected by flow distortion and ship motion. This also requires the fluctuations w' , u' , θ' , q' , but within the inertial subrange.

Over the ocean, a finite difference form of eqn [4], $F_\chi = C_\chi(\chi_S - \chi_A)$, is commonly used for flux calculation. C_χ is the exchange coefficient for variable χ and subscripts S and A refer to values at the surface and a reference height (usually 10 m). Values of C_χ have progressed from constants or simple functions of wind speed, to sophisticated forms embodying flux-gradient stability relationships and effects of sea state. These require iterative computer codes which calculate momentum, sensible and latent heat fluxes simultaneously (so-called bulk flux algorithms). Radiometric measurement of sea surface temperature is preferred, because skin temperature is the physically correct value to use for air-sea exchange. However, sufficiently accurate IR

radiometers are seldom available, so T_s is measured at some depth below the surface. Some flux algorithms take account of the difference between this (bulk) value and skin temperature.

Long-Term Observation Sites

Any description of *in situ* methods would be incomplete without reference to surface observations made worldwide by national weather services. Their principal purpose is for weather, aviation, and shipping forecasts, and increasingly for climatology, but they also provide data for community use. The variables recorded, and often the sensors used, are the same as those needed for research. Research experiments tend to record continuous time series at high resolution over relatively limited periods. Weather observations feature discrete data sampling (1–3 h, depending on the station), but on a regular and continuing basis.

The value of this data is enhanced by the careful documentation and quality control performed by the forecasting services, and the fact that instruments, sites and observational procedures conform to agreed international practice. Meteorological services do not adopt new instrument types without international comparisons, and avoid too many changes to preserve homogeneity of the climate record. The requirements for instruments (such as accuracy and calibration) and observation procedures are set out in WMO (1986).

Increasingly, automatic weather stations (AWS) are being developed and installed worldwide, taking advantage of technological advances in sensor development, computer capability, and communications. Data transmission from remote sites is possible by satellite or phone. Other advantages include more frequent sampling, exact timing, and avoidance of human error. This not only improves the forecast, but both the real-time and archived data are a quality resource for research.

A global network of stations has also been established for long-term measurement of land-atmosphere carbon, water, and energy exchange. Known as 'flux stations', they are operated by research institutions in various regions: 'Ameriflux', 'Euroflux', 'Ozflux' (Australia), etc. The principal measurement is of the eddy fluxes using the fast-response instruments described above, but a full suite of meteorological instruments is also maintained. Like the AWS, data are transmitted to base by a telemetry link.

Experimental Procedures

Platforms

Over land, instruments are usually mounted on a guyed mast or tower. To avoid distortion of the wind flow, these are as slim as possible without losing rigidity. Sensors are mounted at the end of arms, oriented in the direction of the prevailing wind. To accommodate winds from all directions, sensors are often duplicated on opposite sides of the mast. Equation [3] indicates that the profiles will be best defined with logarithmically spaced sensors.

Up to a few meters in height, the mast can be a simple tube, perhaps only 10 cm diameter. Taller masts are usually the climbable lattice type of triangular cross-section and about

30 cm side. Both types are shown in Figure 2, the near lattice tower being 22 m high; the practical limit is about 60 m. Beyond this, a major structure is necessary with a small elevator to transport equipment and personnel to the upper levels. Several such research towers have been built to heights around 300 m; the best known are by USA/NOAA at Boulder, the Dutch Meteorological Institute at Cabauw, and the Japanese Meteorological Agency at Tsukuba. More elaborate measures must be taken to extend the sensors far enough from these towers to minimize interference.

At shallow water sites, such as lakes and the coastal ocean, masts and research platforms have been built on the seafloor. Over the deep ocean, platforms are either ships or moorings. The difficulties of operating from ships have already been mentioned, particularly distortion of measurements by their bulk. Spar and toroidal buoys moored to the seabed also experience motion, but are less obtrusive. Toroids usually carry a frame for instruments to the height of about 4 m.

Networks of such moorings have been installed in climatically important regions, such as the tropical Pacific. They return data daily via satellite and can operate unattended for periods of six-months or longer. The most serious setback to reliability is not technical, but vandalism.

Figure 3 shows a unique and much-used facility called FLIP (Floating Instrument Platform) developed by the Scripps Institution of Oceanography. In effect a huge spar buoy 300 m long, it is towed on site and the lower part flooded so that it actually flips into the vertical. In this deployment, it is set up to measure the structure of the marine ASL. The photograph highlights the multiple sensors needed to determine variability with height, and the long boom to minimize the influence of the bulky platform.

Combined operation of aircraft with *in situ* platforms has proved highly effective. While surfacemounted instruments perform continuous time sampling, the aircraft complements this with spatial sampling over the surrounding region.



Figure 3 The floating instrument platform (FLIP) operated by Scripps Institution of Oceanography, La Jolla, CA, USA. FLIP is shown here set up for an investigation of the marine boundary layer by the Atmospheric Turbulence Laboratory of the University of California at Irvine. The boom is 20 m long and 12 m above the sea surface. It carries a vertical mast with instruments to measure profiles of wind and temperature, atmospheric turbulence and the fluxes of heat, moisture and momentum. Photo: © Carl Friehe.



Figure 4 A Twin Otter research aircraft operated by the US Department of the Navy, equipped with sensors to study wind, temperature, and humidity structure in the atmospheric boundary layer. Photo: © Carl Friehe.

Missions are often flown only 30 m above the surface, mostly using the sensors described above. **Figure 4** shows an aircraft equipped for boundary layer research; turbulent wind components are measured with pressure probes (C and F in **Figure 4**) corrected for aircraft motion with INS (I), temperature with thermistors or PRTs (A), mean humidity with a dew point hygrometer (B), and humidity fluctuations with Lyman- α or IR sensors (D and G). The latter may also be measured with a refractometer, which detects fluctuations of the radio refractive index of air, which depends on humidity, temperature, and pressure, in a microwave cavity. The advent of miniature electronics and the GPS navigation system has enabled development of miniature autonomous robotic aircraft, which can potentially play a valuable role in boundary layer studies.

Tethered balloons have been used to suspend instruments through the ABL, and may thus provide an alternative *in situ* platform within the height range of tall (300 m) towers. Their main advantage is portability and rapid deployment; an obvious difficulty is their instability. Instruments are attached to the tether cable at various levels although the actual height of measurement fluctuates with balloon motion. Eddy-flux measurements have been made from tethered balloons, the attachment for the flux package being designed to maintain its verticality and orientation. A development of this technique is the kytoon, a kite-shaped balloon which provides aerodynamic lift and better stability.

Sampling and Data Recording

When planning a series of observations it is important to consider the sampling required to achieve the aims of the project. We have seen that nonuniform terrain places demands

on spatial sampling. Sampling in time, even in the one-dimensional case, has several aspects. An instantaneous profile will be uneven because of turbulent fluctuations on many scales. Smooth profiles are obtained by averaging ensembles of many profiles over some period, which should be long enough to include most of the low frequency eddies, but avoid diurnal nonstationarity. In the surface layer, a period of 15–30 minutes is convenient.

This identifies the environmental time constraint to obtain realistic mean values. For sampling rate we consider the time response of the sensor. In the case of a ‘slow’ sensor (e.g. a platinum thermometer) this may be many seconds, much slower than the high-frequency end of the atmospheric spectrum. So the sensor rather than the variable determines the sample time; in practice 2–5 samples within the transducer time constant is sufficient.

Between the energy-containing low-frequency eddies of the atmospheric spectrum, and the high-frequency eddies which dissipate turbulence energy to heat, is a region known as the inertial subrange. This covers the spectral range from about 10 Hz to 0.1 Hz. Thus, to determine turbulence statistics and the eddy fluxes we need not resolve higher frequencies than 10 Hz. According to Shannon’s sampling theorem, to reconstruct the original signal without aliasing, this implies a sampling rate of 20 Hz. The fast-response instruments described above for w' , u' , θ' , and q' have been designed accordingly.

Dramatic improvements in computing speed and recording media, and reduced costs have revolutionized data recording practice. In the past it was usual to sample at the highest frequency required but record only average values. Time series of turbulent fluctuations filled racks of computer tapes. Now that data storage is no longer a restriction, all data can be archived at the original sampling speed, to enable reanalysis at

some later date. Ready availability of CD-ROM writers makes large data sets secure and portable.

See also: Agricultural Meteorology and Climatology. **Air Sea Interactions:** Momentum, Heat, and Vapor Fluxes. **Boundary Layer (Atmospheric) and Air Pollution:** Complex Terrain; Surface Layer. **Chemistry of the Atmosphere:** Observations for Chemistry (Remote Sensing): Microwave. **Land-Atmosphere Interactions:** Canopy Processes. **Statistical Methods:** Data Analysis: Time Series Analysis. **Weather Forecasting:** Operational Meteorology.

Further Reading

American Meteorological Society, 2000. Glossary of Meteorology, second ed. AMS, Boston, MA.
 Lenschow, D.H. (Ed.), 1986. Probing the Atmospheric Boundary Layer. American Meteorological Society, Boston, MA.

Bell, R.J., 1972. Introduction to Fourier Transform Spectroscopy. Academic Press, New York.
 Fritschen, L.J., Gay, L.W., 1979. Environmental Instrumentation. Springer-Verlag, New York.
 Garratt, J.R., 1992. The Atmospheric Boundary Layer. Cambridge University Press, Cambridge.
 Kaimal, J.C., Finnigan, J.J., 1994. Atmospheric Boundary Layer Flows. Oxford University Press, New York.
 Kraus, E.B., Businger, J.A., 1994. Atmosphere–Ocean Interaction. Oxford University Press, New York.
 Panofsky, H.A., Dutton, J.A., 1984. Atmospheric Turbulence. Wiley, New York.
 Stull, R.B., 1991. An Introduction to Boundary Layer Meteorology. Kluwer Academic Publishers, Dordrecht.
 World Meteorological Organization, 1996. Guide to Meteorological Instruments and Methods of Observation, sixth ed. WMO, WMO-No. 8, Geneva.
 World Meteorological Organization, 1979. The Planetary Boundary Layer, Chapter 5, Observational Methods and Programs. In: McBean, G.A. (Ed.). Switzerland, WMO-No. 530, Geneva.

Observational Techniques: Remote

WM Angevine and CJ Senff, CIRES, University of Colorado, and NOAA Earth System Research Laboratory, Boulder, CO, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by W M Angevine, C J Senff, E R Westwater, volume 1, pp 271–279, © 2003, Elsevier Ltd.

Synopsis

This article covers remote sensing instruments including acoustic (sodar), radar wind profilers, radio acoustic sounding systems, and optical instruments (lidar). All these techniques are primarily ground-based. Airborne lidar measurements are also covered. Applications including air quality, research into boundary layer physics, mesoscale and storm scale meteorology, climate monitoring, and aviation weather sensing are considered.

Introduction

The planetary atmospheric boundary layer (ABL) is observed using a great variety of techniques both active and passive, employing radio, optical, and acoustic energy. A variety of instruments have been developed in response to the broad range of temporal and spatial scales of interest in the ABL and the large number of parameters that need to be measured. In this article we cover remote sensing instruments, that is, those that measure parameters of a volume of air at some distance from the instrument itself. Most ABL remote sensing is done from the ground, although some spaceborne or airborne instruments have some boundary layer applications. Generally, however, ABL observations require spatial resolution (horizontal and vertical) that is unavailable from instruments on satellites. We will discuss airborne lidar measurements.

Atmospheric quantities and constituents (winds, temperature, humidity, trace gases, etc.) interact with electromagnetic and acoustic radiation in different ways, but all have in common that they interact only weakly with such radiation. Therefore, a primary challenge in the development of remote sensing instrumentation is deriving useful data at low signal-to-noise ratios.

Remote sensing measurements of the boundary layer are used in a wide variety of applications. Air quality research and operational measurements for air quality, research into basic ABL physics, mesoscale and storm scale meteorology, climate monitoring, and aviation weather sensing are important applications.

Interesting boundary layer motions occur on scales ranging over at least six orders of magnitude, from the dissipative scale of order 1 cm to tens of km, although most measurements are of scales between a few meters and a few kilometers. The boundary layer itself ranges from approximately 100 m to 3 km deep, and the motions that transport energy and momentum are on the scale of the ABL depth. The most important horizontal scales are those of the underlying terrain and variations of pollutant emissions. Different instruments and observational strategies are required to measure different scales even over the range between nocturnal boundary layers (~100 m) and convective boundary layers (~1 km).

The remainder of this article discusses the quantities to be measured in the ABL, then covers the variety of instruments and techniques available. While some ABL remote sensing instruments are commercially available, many are custom-built for research use. An article such as this will necessarily

omit some specific instruments, so we hope to provide a useful overview of instruments and techniques and their underlying principles.

Active and Passive Remote Sensing

Boundary layer remote sensing instruments fall into two major categories, active and passive. Active instruments emit electromagnetic or acoustic radiation and detect the portion of that radiation returned from the target volume of the atmosphere. Passive instruments, on the other hand, detect radiation from natural sources that is scattered or modified by the atmosphere. This article covers only active instruments.

Active instrument categories are radar, lidar, and sodar (q.v.). All are similar in principle, as reflected by their names, which once were acronyms. In the simplest case, radars, lidars, and sodars emit pulses of radio frequency, optical, or acoustic radiation, respectively, and detect the very small fraction of each pulse that is backscattered by the atmosphere. The characteristics of the backscattered radiation carry information about the atmosphere, and that information differs depending on the wavelength and type of radiation and therefore the ways in which the radiation interacts with the atmosphere. These interactions are referred to as *scattering mechanisms*. The primary scattering mechanisms relevant to the ABL are Bragg scatter from the fluctuations of the acoustic and radio refractive index for sodar and radar respectively; Rayleigh scatter from small objects for radar and from air molecules for lidar; Mie scatter from aerosols for lidar; and Raman scatter for Raman lidar. The distance (range) of the volume from which the radiation is scattered is most often measured from the round-trip travel time of the radiation pulses.

Quantities to be Measured

In the boundary layer, measurements of winds, turbulence, temperature, humidity, trace gases, and aerosol content and properties are needed. In addition, derived quantities such as mixing height (ABL depth) are useful. Different instruments and techniques have been developed to address each of these quantities. The graphical presentation of patterns of measured quantities, made possible by the continuous nature of most remote sensing measurements, has led to important insights.

For wind measurements, radar wind profilers, Doppler sodars, and Doppler lidars are used. Given sufficient temporal and spatial resolution, these instruments can also measure turbulence.

Temperature is remotely sensed by radio acoustic sounding systems (RASS), a hybrid radar and acoustic technique, and by rotational Raman lidars. In RASS, a wind profiler measures the speed of a propagating acoustic disturbance and the temperature is derived from that speed of sound.

Humidity or water vapor concentration measurements are made by differential absorption lidar (DIAL) or Raman lidar techniques. The DIAL technique at different wavelengths can also measure concentrations of trace gases such as ozone.

Aerosol backscatter and extinction are measured by backscatter lidars. Most Doppler lidars and DIALs also produce backscatter measurements.

Mixing height or ABL depth, a key quantity for most ABL operations and research, can be detected by several techniques. Wind profiling radar reflectivity exhibits a peak at the ABL top, which has been used extensively to detect mixing height. Aerosol, humidity, trace gas, turbulence, and temperature profiles from lidars can also indicate the ABL depth.

Integration of sensors can yield better results than individual sensors alone. The entire boundary layer cannot always be covered with sufficient resolution by a single sensor, so combining profiles from, for example, a sodar and a wind profiler may be useful. This is especially true for mixing depth over the full diurnal cycle.

Techniques

Sodar

Sodars emit sound waves and detect the backscattered acoustic signal to measure atmospheric structure and, in the case of Doppler sodars, velocity. Sodar was gradually developed during the 1960s and came into routine use in the 1980s. Today sodars are used for operational monitoring of winds above normal tower heights at power plants and other pollution sources, and are commonly included in boundary layer research campaigns. A typical sodar is shown in Figure 1.

The simplest sodars measure only the intensity of backscatter from the turbulent structure of temperature and velocity in the boundary layer. The time and height variations of turbulent structure can be interpreted to elucidate a variety of boundary layer phenomena. Doppler sodars add the capability of measuring the Doppler shift of the backscattered acoustic signal and convert the Doppler shift to a velocity along the direction of the acoustic beam. Beams in different directions are emitted and detected sequentially to produce the full three-dimensional wind vector. Three or five beams are commonly used; in the case of three beams two are separated by 90° in azimuth and aimed $15\text{--}20^\circ$ off the zenith, and the third is a vertical beam; in the case of five beams four are separated by 90° in azimuth and the fifth is again vertical. Different beam directions are generated either by separate antennas or by a phased array allowing electronic beam steering.

Sodars generally measure in the lowest few hundred meters of the atmosphere, although some are capable of maximum ranges of a kilometer or more. The limitation on range is due to



Figure 1 Sodar with three separate antennas at a power plant in Germany. Photo courtesy of METEK Mestechnik GmbH.

the rapid attenuation of the acoustic signal in the atmosphere, which is highly variable and may be as much as 40 dB km^{-1} . Sodars operate at frequencies between a few hundred hertz and a few kilohertz. The operating frequency is chosen to optimize among range resolution, maximum range, and antenna size. The minimum sampling time on each beam is determined by the maximum range and the sensitivity, and is normally a few seconds. A few tens of samples are commonly averaged to improve the quality of the wind estimates, so the 3D wind vector is available every 10–30 min. Range resolution varies from a few meters up to 20 m or so. Some sodars use multiple frequencies.

There are some hundreds of sodars installed at monitoring sites around the world, and several manufacturers provide off-the-shelf systems. In addition, at least tens of systems are in use by research groups.

Quality issues for sodar data include sensitivity to ambient noise and to noise from rain hitting the antenna. Sodars are also sometimes susceptible to ground clutter, strong spurious signals from nearby objects. Siting a sodar can be difficult because of concerns about the noise produced.

Radar

Specialized radars are an important category of boundary layer remote sensing instrumentation. Traditional weather radars rely primarily on scattering from hydrometeors, although modern systems have sufficient sensitivity to sense clear-air returns. Boundary layer remote sensing is most commonly done with wind profiling radars (profilers) (q.v.), which have comparable sensitivity to weather radars at much lower cost, but have reduced scanning flexibility and coarser time resolution. A boundary layer profiler is shown in Figure 2.

Boundary layer wind profilers evolved from mesosphere-stratosphere-troposphere (MST) radars (q.v.), which were developed for investigations of the middle and upper atmosphere. The original impulse for boundary layer profiler development was the need to fill in wind measurements at heights below the minimum range of the MST radar. The first



Figure 2 915 MHz wind profiler of the NOAA Aeronomy Laboratory at Chebogue Point, Nova Scotia for the 1993 North Atlantic Regional Experiment. The cylinders are the RASS acoustic sources. The flaring structure is the clutter fence, designed to reduce ground clutter. The 2×2 m phased array antenna is at the base of the clutter fence inside the structure.

ABL profiler was developed at the NOAA Aeronomy Laboratory by Warner Ecklund and colleagues in the mid-1980s. Since then, a large number of similar profilers have been built and a number of variations have been explored. Most ABL profilers are pulsed Doppler radars, but a few frequency modulation continuous wave (FM/CW) systems are also in use.

Boundary layer profilers operate at about 915 MHz in North America and 1290 MHz in Europe and Asia, the frequency being determined by the availability of bandwidth in the region. The basic beam geometry is similar to that described above for sodars. All profilers have Doppler capability. Beams in different directions are emitted and detected sequentially to produce the full three-dimensional wind vector. Three or five beams are commonly used; in the case of three beams two are separated by 90° in azimuth and aimed $15\text{--}20^\circ$ off the zenith, and the third is a vertical beam; in the case of five beams four are separated by 90° in azimuth and the fifth is again vertical. Different beam directions are generated by separate antennas or by a phased array allowing electronic beam steering. An individual sample is acquired approximately once per second, and 10–50 samples are averaged before going on to the next beam. An estimate of the full 3D wind vector is typically available every 1.5–5 min, but usually these are also averaged to reduce noise, so the most common time resolution is 20–60 min.

Wind profilers measure 3D winds by detecting the motion of refractive index fluctuations in the clear air at the scale of half the radar wavelength. In the boundary layer, the refractive index fluctuations are dominated by the fluctuations of humidity. Therefore the intensity of the backscattered radar signal is essentially proportional to the humidity gradient and the turbulence intensity. This relationship is used to find the height of the convective boundary layer. Wind profilers are generally not suitable for measuring the structure of the stable ABL because their minimum range and range resolution are too large. Estimates of turbulence intensity are made either from

the variance of the time series of wind velocities in the vertical beam, from the width of the Doppler spectrum, or a combination of the two.

Boundary layer profilers generally use range resolution of 60–100 m depending on the choice of operating modes. Better resolution results in less maximum range due to reduced average power. The minimum range of most ABL profilers is 100–150 m, and the maximum range depends strongly on the ambient conditions. The maximum range is always sufficient to see the entire convective boundary layer over land, at least 4 km in summer and in the tropics, but as little as 1 km in continental winter when humidity and turbulence are both lacking. As previously mentioned, time resolution can be as good as a few minutes, but generally is 20 min or more. An example of data from a boundary layer profiler is shown in **Figure 3**.

Most boundary layer profilers are used in field campaigns for basic boundary layer physics and air quality research. Some are in fixed or long-term deployments. Winds from profilers are used to define transport of pollutants and to explore the flow patterns due to coasts and storms. The convective boundary layer height derived from profilers is a key quantity in analysis of pollutant plumes from urban areas and power plants. Winds and ABL height from networks of profilers have been used for basic studies of the spatial and temporal structure of the boundary layer. Profilers have also been deployed for climate monitoring and for monitoring of the winds near airports.

While profilers are designed to detect backscatter from the clear air, they are also sensitive to other types of scatter. Birds, insects, and hydrometeors are important scatterers and sometimes dominate the clear-air signal. Most profiler systems are equipped with hardware and software for quality control, but users of profiler data must take care to avoid misinterpretation of contaminated data. Another quality issue with profiler data is so-called ground clutter, that is, signals from ground-based

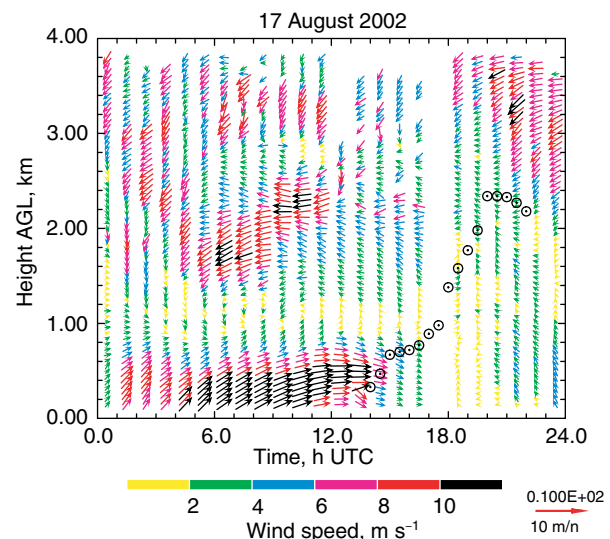


Figure 3 Wind and mixing height data from a boundary layer wind profiling radar at Houston Southwest Airport, Texas. Wind speed is shown by the length and color of the arrows; wind direction is along the arrow with north at the top of the plot. Circled dots are mixing height derived by an observer from the pattern of radar reflectivity.

objects (e.g., trees) that may be seen by a sidelobe of the antenna. As with any radar, interference from other radio sources may be an issue.

In addition to the well-established profilers to which most of the above description pertains, several exotic radar systems have been built for ABL research. Such systems often use spaced-antenna or interferometric techniques. At least one FM/CW system exists that produces extremely high resolution pictures of the boundary layer.

Operational wind profiling radars operating at lower frequencies, including the 449-MHz profilers of the NOAA Profiler Network, have sufficient resolution for some boundary layer studies. Weather radars may also provide some useful boundary layer information.

Radio Acoustic Sounding Systems

RASS (q.v.) are often attached to wind profilers. In RASS, an acoustic signal is emitted with a wavelength one-half the radar wavelength (frequency 2–3 KHz for most ABL profilers). The radar measures the speed of sound, from which the (virtual) temperature can be straightforwardly derived to a reasonable degree of accuracy and precision (<0.5 K accuracy and precision for half-hour averages). A second type of RASS also exists as an accessory to sodars, in which case the radar is a simple continuous-wave type. The range resolution and minimum range of RASS are the same as those of the profiler or sodar to which it is attached, but the maximum range is limited by advection of the acoustic signal out of the radar beam when moderate to strong winds blow, and by the acoustic attenuation, which can be very large at typical RASS frequencies.

Lidar

Lidars are active remote sensing instruments that are well suited for ABL research. They measure many important quantities, such as wind speed, turbulence, aerosol backscatter and extinction, water vapor, other trace gases, and temperature; and they typically cover the entire depth of the ABL. Lidars use lasers to emit short light pulses into the atmosphere. The small fraction of the light backscattered by the atmosphere is detected by a receiver typically consisting of a telescope, photodetectors, and data acquisition electronics. For lidars used in ABL research the relevant scattering mechanisms are Rayleigh and Raman scattering from air molecules and Mie scattering from aerosol particles.

The first lidars were built in the 1960s soon after the invention of the laser and extensive work went into developing lidar remote sensing techniques. In the 1970s and 1980s, with the advent of lasers that could be tuned over a wide range of wavelengths, the application of lidars in atmospheric remote sensing became more widespread. The 1990s saw increased use of solid state lasers as lidar transmitters and a trend toward more compact lidar systems. This trend continued in the 2000s as robust, commercially available optical components, initially developed for the telecom industry, were increasingly utilized in lidar systems. Lidars that are in use today operate at wavelengths from the near ultraviolet to the far infrared regions of the electromagnetic spectrum. Most lidars are still prototype,

research-grade instruments. In recent years, however, a number of lidar systems, in particular aerosol and Doppler wind lidars, have been developed that can be operated unattended over extended time periods and lend themselves to long-term monitoring applications. A specific category of lidars, called ceilometers, are autonomous, turn-key instruments that are commercially available and have seen fairly widespread use for a number of years; however, their application is generally not research oriented. Ceilometers are compact, unattended, low-power lidars that are mostly operated at airports to routinely measure cloud base height.

Many of the lidar systems used for ABL research are installed on mobile platforms, such as trailers, ships, or aircraft, so they can be easily deployed in different locations. [Figure 4](#) shows a lidar system installed in a research aircraft. This aircraft typically flies at 3–4 km above sea level allowing profiling of the entire ABL with the downward-looking lidar.

In response to measurement needs, various types of lidars have been developed that are distinctly different in terms of the detection technique and scattering mechanism they employ. Most ABL lidars detect elastically backscattered light from air molecules and aerosol particles. Among this group of lidars one distinguishes direct detection lidars and heterodyne Doppler lidars. Direct detection lidars measure the intensity of the backscattered light, while heterodyne lidars use coherent mixing with a reference beam to detect the Doppler shift between the emitted and backscattered light. The Doppler shift is a measure of the radial wind speed along the lidar beam direction. By scanning the Doppler lidar beam using specific patterns, profiles of horizontal wind speed, direction, vertical velocity, and turbulence can be measured. Both direct detection and Doppler lidars can be used to infer information about aerosols in the ABL, particularly aerosol backscatter and extinction. A direct detection lidar that is used solely for this purpose is often referred to as a backscatter lidar. A High Spectral Resolution Lidar is a specific type of backscatter lidar that separates the atmospheric particulate and molecular return signals and is capable of providing independent measurements of aerosol backscatter and extinction profiles. DIALs are direct



Figure 4 Airborne, downward-looking ozone lidar of the NOAA Earth System Research Laboratory installed in a research aircraft.

detection lidars that emit laser light at two or more wavelengths. These wavelengths are chosen so that they are absorbed to a different degree by an atmospheric trace gas. If the absorption cross section of the trace gas is known, its concentration can be inferred from the intensity difference of the backscattered light at the various wavelengths. DIALs are most commonly used to measure water vapor and ozone concentration, but other trace gases, such as sulfur dioxide, nitrogen dioxide, ammonia, and various hydrocarbons, have been detected with DIAL as well. As an example of the kind of information a DIAL system can provide, Figure 5 shows a time–height cross section of ozone mixing ratio measured with an airborne ozone DIAL. The high ozone concentrations in the center of the plot indicate the extent of the urban pollution plume that had developed on 12 July 1995 over Nashville, Tennessee. In contrast to direct detection and Doppler lidars, Raman lidars detect the fraction of the emitted light that is Raman scattered by air molecules or trace gases. The Raman scattered light is shifted in wavelength, and the magnitude of the shift depends on the scatterer. Raman lidars are primarily used to measure humidity profiles by detecting the Raman scatter of water vapor molecules in the atmosphere. Raman lidars can also be employed to measure aerosol extinction from the Raman scattered light of oxygen or nitrogen molecules (for which the concentrations are known). Another important application of Raman lidar is the measurement of temperature profiles using the rotational Raman lidar technique.

Time and spatial scales of lidar measurements in the ABL vary depending on the measured variable and the type of lidar used. In the following, time and range resolutions are given that can be achieved with state-of-the-art, research-grade lidar systems. Profiles of aerosol backscatter can be measured with time and range resolutions as low as 0.1 s and 1.5 m, while wind speeds can be retrieved at resolutions of about 1 s and

30 m. The DIAL technique requires taking the derivative of the measured signals with respect to range. Therefore, higher signal-to-noise ratios are necessary than for wind speed or aerosol backscatter measurements, requiring longer averaging times and coarser range resolution. Time and range resolutions of ABL water vapor or ozone measurements that can be achieved with DIAL are on the order of a few seconds and several tens of meters, respectively. Due to the low Raman scattering cross sections, daytime Raman lidar measurements in the ABL generally require averaging over several minutes and yield range resolutions of about 100 m. At night, without interference from solar background light, water vapor measurements using the Raman technique are comparable to DIAL measurements in terms of their resolution. The maximum range of lidars used in ABL research is in most cases sufficient to cover the entire ABL and the portion of the free troposphere immediately above the ABL. The minimum range is typically on the order of a few hundred meters. With a scanning lidar the altitudes below minimum range can be probed by pointing the beam a few degrees above the horizon. Certain meteorological conditions restrict the useful range of lidar. Lidar measurements in or beyond optically thick clouds (such as ABL cumulus or stratus) or in dense fog are not possible because the lidar signals are scattered and absorbed within the first tens of meters. During precipitation events lidar operation is not useful and may be impossible.

The measurement capabilities of lidars make them valuable tools in many ABL research areas. Profiles of aerosol backscatter, humidity, trace gas concentrations, temperature, and turbulence measured with lidar can be used to infer the ABL height, which is a key parameter in ABL research. Due to the high time and range resolution that can be achieved with lidar it is possible to investigate turbulence statistics in the ABL, including third- and fourth-order moments. Their narrow beam width and shallow-angle scanning capabilities

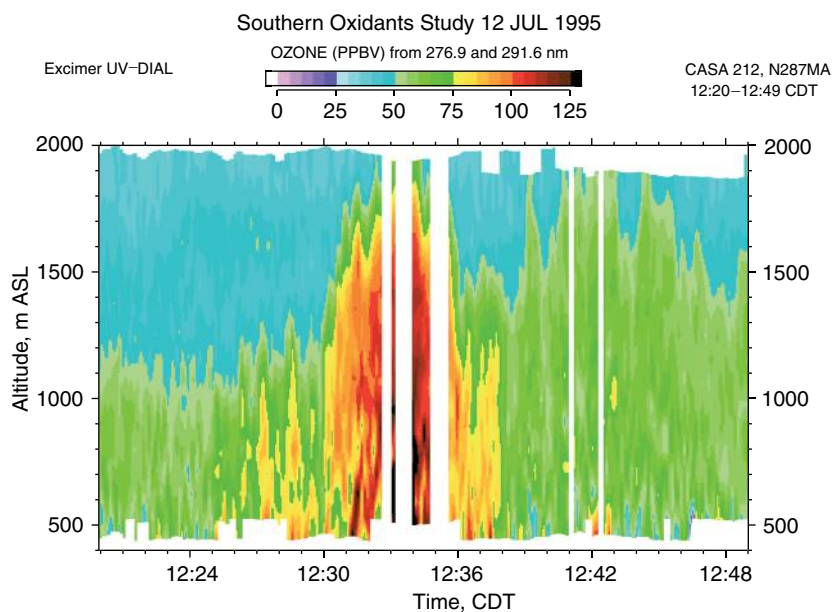


Figure 5 Ozone cross section measured with an airborne ozone DIAL around midday on 12 July 1995 during a transect passing over the city of Nashville, Tennessee.

make lidars well-suited to provide detailed information about the shallow nocturnal ABL. Air quality research, in particular the distribution and transport of pollutants and fine particles, is another important application. Other research areas where lidar has made important contributions include the study of flows in complex terrain, the characterization of the wind field at wind farms, and the measurement of trace gas fluxes by using collocated DIAL and Doppler lidars.

See also: **Aerosols:** Observations and Measurements. **Boundary Layer (Atmospheric) and Air Pollution:** Air Pollution Meteorology. **Chemistry of the Atmosphere:** Observations for Chemistry (Remote Sensing): IR/FIR (Satellite, Balloon and Ground); Observations for Chemistry (Remote Sensing): Lidar; Observations for Chemistry (Remote Sensing): Microwave. **Mesoscale Meteorology:** Cloud and Precipitation Bands; Mesoscale Convective Systems. **Mountain Meteorology:** Overview; Valley Winds. **Numerical Models:** Mesoscale Atmospheric

Modeling. Optics, Atmospheric: Optical Remote Sensing Instruments. **Radar:** Mesosphere–Stratosphere–Troposphere and Stratosphere–Troposphere Radars and Wind Profilers.

Further Reading

- Carter, D.A., Gage, K.S., Ecklund, W.L., Angevine, W.M., Johnston, P.E., Riddle, A.C., Wilson, J., Williams, C.R., 1995. Developments in UHF lower tropospheric wind profiling at NOAA's Aeronomy Laboratory. *Radio Sci.* 30, 977–1001. American Geophysical Union.
- Clifford, S.F., Kaimal, J.C., Lataitis, R.J., Strauch, R.G., 1994. Ground-based remote profiling in atmospheric studies: an overview. *Proceedings of the IEEE* 82, 313–355. Institute of Electrical and Electronic Engineers.
- Hinkley, E.D. (Ed.), 1976. *Laser Monitoring of the Atmosphere*. Springer, New York.
- Lenschow, D.H. (Ed.), 1986. *Probing the Atmospheric Boundary Layer*. American Meteorological Society, Boston, 269 pp.
- Measures, R.M., 1984. *Laser Remote Sensing*. John Wiley & Sons, New York.
- Wilczak, J.M., Gossard, E.E., Neff, W.D., Eberhard, W.L., 1996. Ground-based remote sensing of the atmospheric boundary layer: 25 years of progress. *Boundary-Layer Meteorology* 78, 321–349. Kluwer Academic.

Ocean Mixed Layer

L. Kantha, University of Colorado, Boulder, CO, USA

CA Clayson, Woods Hole Oceanographic Institution, Woods Hole, MA, USA

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

The oceanic mixed layer (OML) mediates the exchange of mass, momentum, heat, and energy between the oceans and the atmosphere. As such, it plays a crucial role in long-term weather and climate. It is also important to the oceanic food chain, since the nutrient uptake into the euphotic zone from nutrient-rich waters below depends on the entrainment at the base of the mixed layer. This section provides an overview of the OML and the associated physical processes.

The oceanic mixed layer (OML) mediates the exchange of mass, momentum, heat, and energy between the oceans and the atmosphere. As such, it plays a crucial role in long-term weather and climate. It is also important to the oceanic food chain, since the nutrient uptake into the euphotic zone from nutrient-rich waters below depends on the entrainment at the base of the mixed layer. This section provides an overview of the OML and the associated physical processes.

Introduction

The OML, the ocean region adjacent to the air–sea interface, is typically tens of meters deep, and due to the fact that it is well mixed, the temperature and salinity (and therefore the density) are fairly uniform. The rapidly changing regions below these uniform regions of temperature, salinity, and density are called the thermocline, halocline, and pycnocline, respectively. The mixing is primarily shear driven, since the wind stress at the surface is the primary mixing agent, although at night significant convective mixing driven by the heat loss to the atmosphere takes place. The OML is heated near the surface by both short wave (SW) and long wave (LW) radiative fluxes, and deeper in the water column from solar radiation in the visible part of the spectrum penetrating into the OML. This solar heating produces a diurnal cycle that varies in importance and magnitude at different latitudes. The cooling, however, is driven from heat and evaporative losses at the surface. Seasonal variation of the OML due to radiative heating is also important, although its importance depends on the latitude.

The OML mediates the exchange of mass, momentum, energy, and heat between the atmosphere and the ocean and hence plays a central role in long-term climate and weather. Because of the high heat capacity of water (2.5 m of the upper ocean has the same heat capacity as the entire troposphere), and because the oceans compose over two-thirds of the surface of the globe, most of the solar heating on Earth passes through the OML. Oceans are heat reservoirs, gaining heat during spring and summer and losing it slowly during fall and winter, and therefore act like a flywheel in matters related to weather on timescales of weeks and longer.

The OML also plays an important role in the oceanic food chain. Primary production by phytoplankton is the first link in this chain. The need for an energy source in producing biomass

restricts primary production to the upper few tens of meters (the euphotic or photic zone), in which the solar insolation is strong enough to assist carbon fixation. The mixing at the base of the OML is also crucial to biological productivity. The OML is normally nutrient-poor and it is the injection of nutrients from the nutrient-rich waters below the seasonal thermocline that permits higher levels of primary productivity. In fact, it is the upwelling regions (which compose just a few percent of the world's oceans), where nutrient-rich waters are forced into the OML and brought into the photic zone, that provide most of the fish catch around the world.

Biological productivity is important from a climatic point of view over timescales of decades or more. Carbon fixing constitutes a biological pathway for removing some of the anthropogenic CO₂ introduced into the atmosphere. There also exists an inorganic pathway since there is a significant uptake of CO₂ in the cold subpolar oceans, some of which are also regions of deep and intermediate dense water formation. The ocean acts as an important CO₂ sink on the globe and accounts for a significant fraction of the 'missing' anthropogenic CO₂ input to the atmosphere. However, quantification of the magnitude of this sink requires accurate OML models coupled to accurate ecosystem and air–sea transfer models.

Finally, the OML constitutes the first link in the chain of oceanic pollution. Most of the pollution in the global oceans takes place in the coastal oceans through the OML, and therefore the fate of any pollutants accidentally or intentionally deposited in the OML depends on the mixing and dispersion in the OML.

Characteristics

An OML can be divided into four parts, the very thin but important molecular sublayer, a few millimeters thick; the wave sublayer, normally 2–6 m thick; the main bulk of the OML, 10–40 m thick; and the entrainment sublayer of about 5–10 m thickness. In deep convective OMLs, where the mixed layer depth is a few hundred meters or more, the fractions of the wave and entrainment sublayers are small. In a shallow diurnal OML, a few meters thick, the wave sublayer can be a large fraction. An active gravity wave field can damp out the diurnal modulation of sea surface temperature (SST) by wave-driven mixing through Langmuir turbulence or wave breaking processes.

The active turbulent mixed layer in the upper ocean is usually bounded below by a strong buoyancy interface, in the form of a layer with either a sharp decrease in temperature (seasonal thermocline) or a sharp increase in salinity (halocline) (or both). In either case, this layer (called a pycnocline) is stably stratified, and here, turbulence is damped by buoyancy forces. The transition region from active turbulent mixing to mostly quiescent layers below can be called a turbocline, in analogy with the thermocline. Normally, the turbocline coincides with the seasonal thermocline or halocline, but not necessarily both. During high precipitation events, a shallow brackish layer can form and the halocline and turbocline are at similar depth but the thermocline is much deeper. In the tropical western Pacific, a similar situation exists, leading to the so-called barrier layer that plays an important role in the transfer of heat from the ocean to the atmosphere in the tropics, by acting as a barrier to mixed layer deepening and entrainment of waters below the halocline.

An OML is mixed from both the top and the bottom. At the top, it is the winds, waves, and buoyancy fluxes that stir the fluid. At the bottom, it is the entrainment driven by large turbulent eddies in the OML that mixes the denser fluid from below into the OML. Wind-driven current in the OML also

causes strong shear at the base of the mixed layer; shear instability ensues, inducing Kelvin–Helmholtz billows, which thicken the buoyancy interface and hence decrease its resistance to erosion by turbulent eddies. In deep OMLs, it is these mechanisms at the bottom that are responsible for a majority of the deepening of the OML. In shallow OMLs, the surface-stirring processes due to gravity wave breaking and cellular motions are also important. Note that turbulent erosion tends to sharpen the pycnocline, while K–H billows tend to make it more diffuse.

Perhaps the most salient aspect of the OML in midlatitudes is its diurnal and seasonal variability. Figure 1 shows the typical seasonal cycle in midlatitudes. This seasonal variability in OML depth and temperature, and hence the heat content of the OML, is a prime factor in the air–sea exchange at these latitudes. The onset of spring warming restratifies the water column and once the shallow spring–summertime thermocline forms, its depth stays roughly constant. However, the formation period is heavily influenced by wind events at the time. Similarly, wind forcing controls the deepening of the OML at the onset of autumn cooling. During this time, the OML deepens episodically during intense storms that pass through the region, with significant assistance from cooling at the surface. Both sensible and latent heat fluxes are important in

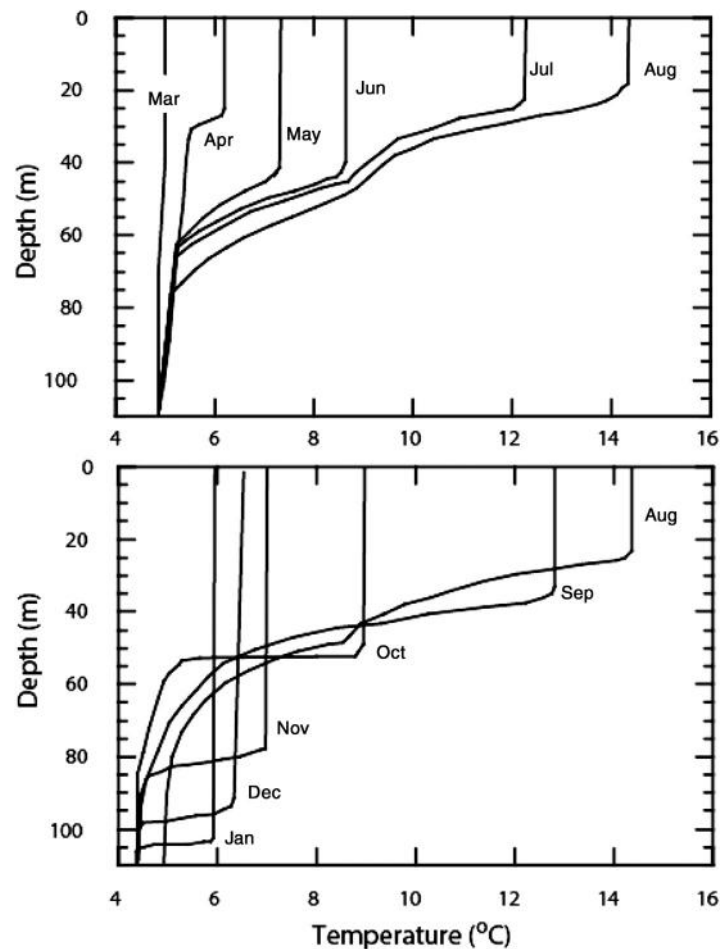


Figure 1 Seasonal evolution of temperature in the midlatitude upper ocean. Shallow warm mixed layers during spring/summer alternate with cold deep ones during fall/winter.

cooling the ocean at midlatitudes, whereas it is principally the evaporative losses that dominate the air–sea exchange at warmer low latitudes and sensible heat loss at colder high latitudes. Precipitation events also play a role in mixing. Wintertime cold air outbreaks along the east side of continents lead to rapid OML deepening, a large heat loss from the ocean, and cyclogenesis in the atmosphere. At higher latitudes, the OML deepens more due to penetrative convection, and in subpolar regions, deep convection occurs. The OML structure in midlatitudes is affected by both salinity and temperature, whereas in subpolar and polar oceans, salinity plays an overwhelmingly important role in mixing, and in the tropics, the thermal structure in general predominates.

Diurnal variability affects the heat exchange on shorter time-scales and may play a role over long timescales as well. The intensity of diurnal modulation of the OML depth and temperature depends on the season. Generally, the modulation is stronger if the solar insolation is strong and winds weak. Longer daylight hours also lead to increased diurnal warming producing a latitudinal dependence. Precipitation, especially during the early morning hours, can lead to a stable fresh layer also enhancing diurnal warming. Thus during summer in subtropics, with longer days, higher insolation values reaching 1200 W m^{-2} , and light winds, can lead to a diurnal modulation of as much as $6\text{--}7^\circ\text{C}$. Under these extreme conditions, the surface layer no longer resembles a shallow mixed layer with very reduced (but still existing) turbulence, but instead an extremely shallow very stable layer. Turbulence itself is completely extinguished and molecular processes in addition to the absorption profile of solar radiation drive the temperature profile. This shallow layer is essentially uncoupled from the deeper residual mixed layer, and a diurnal current can form over the region where this condition occurs, as the surface water ‘slips’ over the deeper mixed layer. The depth of the diurnal mixed layer (and the resulting temperature structure) can also be affected by Langmuir circulation and turbulence. A part of the heat built up in the OML during the day is lost by nocturnal cooling, which drives a vigorous convection and mixing in the water column that normally mixes some of the heat gained into the seasonal mixed layer.

A major factor in OML dynamics in the equatorial regions is the presence of strong background currents in the vicinity of the OML. The Equatorial Undercurrent in the Equatorial Pacific is a typical example. It exists at depths ranging from 50 to 200 m and is an eastward-flowing current that produces a strong vertical shear, which has a major influence on mixing in the upper water column. In contrast, in midlatitude oceans, the principal balance is between the Coriolis terms and the stress divergence, and the currents are not continuously accelerated by a steady wind; instead a steady state is reached and an Ekman-like spiral is produced.

In ice-covered oceans, the ice mediates the exchange of momentum between the atmosphere and the OML. The principal balance in ice is between the Coriolis force, the wind stress at the top, internal stresses in ice, and the shear stress on the ocean at the bottom. This force balance determines the stress available for mixing under ice. In addition, ice growth and melting causes buoyancy fluxes that affect the OML below the ice. Stirring by deep ice keels is an important factor. The ice cover tends to insulate high latitude oceans from the cold atmosphere. However, it is not continuous and even in the

middle of winter, there exist narrow openings in ice, called leads, through which a substantial fraction of winter-time heat loss to the atmosphere at high latitudes takes place. **Figure 2** shows the mixing processes prevalent under leads.

There are striking similarities between the atmospheric boundary layer (ABL) over land and the OML. Under convective conditions, similar scaling laws hold in both turbulent layers. However, the most important difference between the OML and the ABL over land is the presence of surface waves at the air–sea interface that play an active role in its dynamics. The dynamical influence of the ground surface on the ABL is determined by its roughness and topography, which are invariant, whereas it is the effective roughness of the mobile sea surface that is constantly changing with the winds that is important in the OML.

Under neutral stratification, it is possible to find a region where the universal law of the wall scaling would apply: $q \sim u_*$, $l \sim z$, and $\varepsilon \sim z^{-1}$, $v_t \sim z$, where u_* is the friction velocity (square root of the ratio of the wind stress to the water density), q is the turbulence velocity scale, l its length scale, ε is the dissipation rate of turbulence kinetic energy (TKE), v_t the eddy viscosity, and z is the depth. Here, the mean shear is proportional to u_* but inversely proportional to z and therefore the mean velocity is proportional to the logarithm of the distance from the free surface (see references listed under further reading). Indeed this scaling can be found in the upper part of the OML, except close to the surface. Close to the surface, under strong wind conditions, modern measurements have found that the dissipation rate is one to two orders of magnitude larger than that given by the law of the wall (**Figure 3**). This near-surface elevated dissipation rate is due to the influence of surface waves and wave breaking. Wave breaking generates intermittent, shear-free turbulence somewhat akin to the turbulence generated by a stirring grid in a fluid. The turbulence intensity drops off sharply away from the source. Therefore, while the turbulence intensities are elevated above the usual levels during extensive wind-wave breaking, this turbulence is important only to a depth on the order of the amplitude of the breaking waves. Below these depths, the law of the wall can often be found once again. Wave breaking and associated turbulence are likely to be important for the dynamics of OMLs, especially shallow ones; because of the elevated near-surface dissipation rates, they may bring about a higher exchange of gas and heat across the air–sea interface. If it were not for the surface waves, the turbulence near the surface of an OML would behave roughly similar to that adjacent to a solid boundary, such as the ABL over land.

Solar Heating

The solar radiation incident on the ocean surface can be divided into three components: short wavelengths in the ultraviolet part of the spectrum ($<350 \text{ nm}$), the wavelengths available for photosynthesis (photosynthetically available radiation (PAR), $350\text{--}700 \text{ nm}$), and the infrared and near-infrared wavelengths ($>700 \text{ nm}$). The ultraviolet portion is roughly 2%, PAR 53%, and infrared 45% of the total solar insolation. PAR coincides roughly with the visible portion of

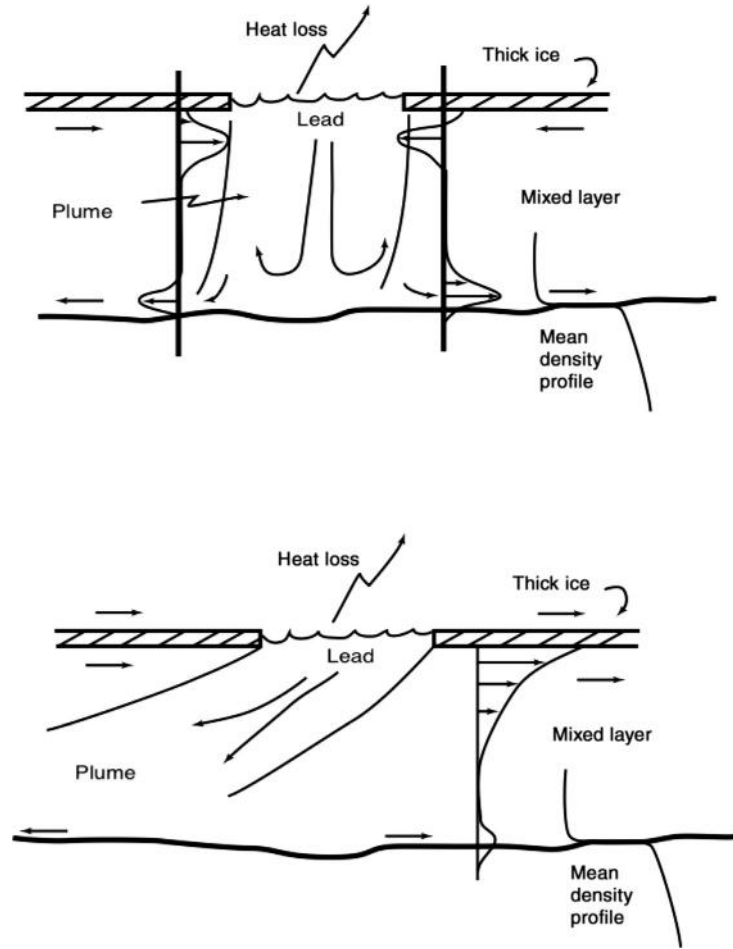


Figure 2 OML driven by freezing and salt extrusion underneath a refreezing lead. Top panel corresponds to stationary ice cover, when free convection dominates, and the bottom panel to the case when the ice is in motion, when forced convection dominates the mixing process.

the spectrum and is the most important of the three portions for biological aspects of the upper ocean. Primary productivity and fixing of carbon by phytoplankton take place only in the euphotic zone, defined as the depth at which respiration and primary production balance, which is where PAR decreases to 1% of its surface value. The ultraviolet part is important to the production of certain photochemicals such as carbonyl sulfide. **Figure 4** shows the spectrum of solar insolation at the top of the ocean and at various depths. Only the visible part remains below about 10 m.

Air–Sea Fluxes

For air–sea exchange purposes, fluxes of momentum, sensible and latent heats, water vapor, and dissolved gases such as CO_2 across the air–sea interface affect both the ABL and the OML. The transfer of momentum from winds to surface waves is usually relevant only in so far as it affects the net transfer to the ocean currents. However, the water vapor and gas fluxes from the oceans to the atmosphere, which are usually net losses to the ocean, are also important. All the fluxes of heat, mass, momentum, and gases are determined by turbulent

processes in the surface layers of the atmosphere and the ocean adjacent to the air–sea interface, with surface waves playing an important role by virtue of their ability to act as sinks of momentum, to determine the ‘roughness’ of the sea surface, and to disrupt the aqueous molecular sublayer responsible for transfer of scalar properties across the interface. At sufficiently high wind speeds, spray and droplets ejected into the atmosphere and air bubbles entrained into the ocean during wave breaking directly affect the fluxes between the two media.

There must be a balance among fluxes of all scalars, including heat, at the interface (**Figure 5**).

$$SW_{\downarrow} + LW_{\downarrow} + H_{pr} - SW_{\uparrow} - LW_{\uparrow} - H_s - H_L - SW_{\downarrow}^p = Q_{net} \quad [1]$$

where $SW_{\uparrow} = \alpha SW_{\downarrow}$; α is the albedo of the ocean surface. There is thus a balance between the downwelling shortwave ($SW_{\downarrow}$) and longwave ($LW_{\downarrow}$) radiative fluxes, and the upwelling $SW_{\uparrow}$ and $LW_{\uparrow}$ fluxes, the sensible (H_s) and latent (H_L) heat fluxes, the heat flux due to any precipitation (H_{pr}), the solar radiative flux penetrating into the ocean ($SW_{\downarrow}^p$) and the net heat flux Q_{net} at the surface. Treating the mixed layer as

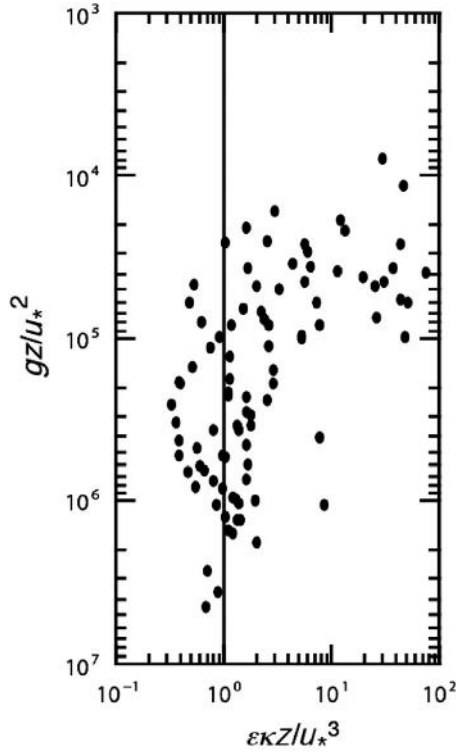


Figure 3 Measured dissipation rate of TKE in the near-surface layers of the OML. The dissipation rate is more than an order of magnitude higher than that expected from the classical law of the wall scaling.

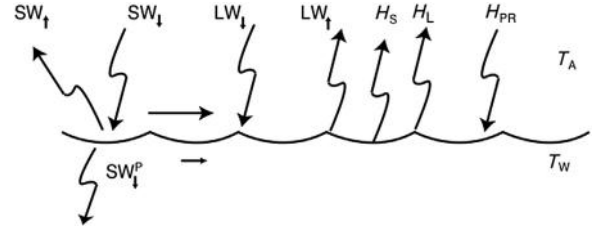


Figure 5 A sketch of the heat fluxes at the air-sea interface.

a slab, the rate of change in temperature of the OML due to the effects of the surface fluxes is given by the upper ocean heat budget:

$$\frac{\partial T}{\partial t} = \frac{Q_{\text{net}}}{\rho C_p D} + (U_{\text{geo}} + U_{\text{Ek}}) \cdot \nabla T + \frac{(W_e + W_{\text{Ek}})(T - T_b)}{D} + \kappa \nabla^2 T \quad [2]$$

where T is the OML temperature, D its thickness, ρ and C_p are the density and specific heat of air, U_{geo} and U_{Ek} are the geostrophic and Ekman currents, W_e is the entrainment velocity associated with a deepening of the mixed layer, W_{Ek} is the Ekman current velocity, T_b is the temperature below the mixed layer, and κ is the diffusivity of temperature. These terms thus describe the temperature change due to surface heat flux, advection, entrainment, and diffusion.

The net salinity flux to the ocean is

$$F_S = (\dot{P}_{\text{r,sn}} - \dot{E})(-S_s); \quad \dot{E} = H_L/L_E \quad [3]$$

where $\dot{P}_{\text{r,sn}}$ is the precipitation (rain or snow) rate (m s^{-1}), $\dot{E}$ is the evaporation rate, S_s is the surface salinity, and L_E is the

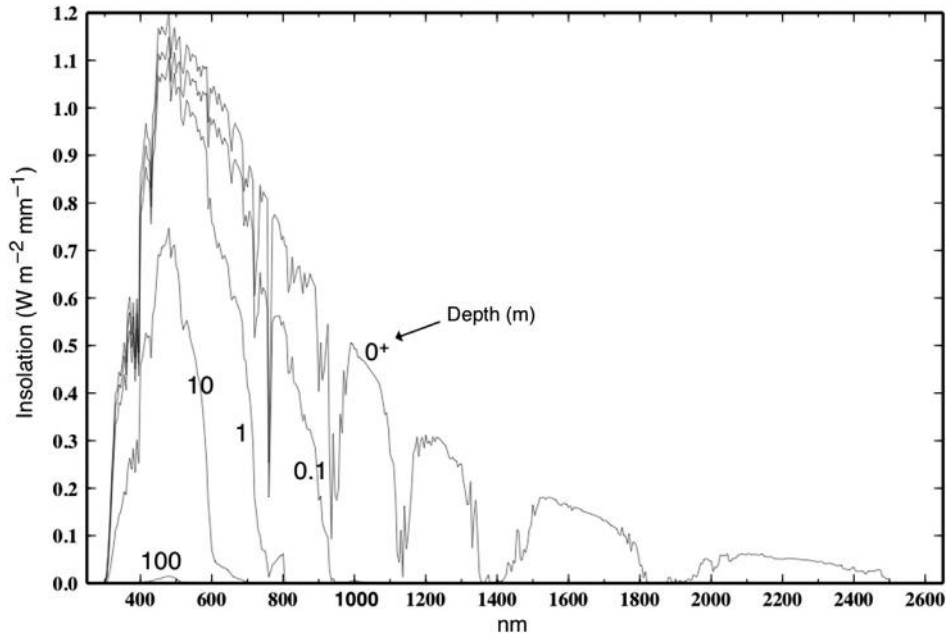


Figure 4 Solar insolation at the ocean surface and below. Note the rapid attenuation of nonvisible components with depth. Only the visible part remains below about 10 m depth.

latent heat of evaporation. Precipitation tends to suppress mixing because of the stabilizing effect of fresh water precipitated onto the salty water of the ocean. The net buoyancy flux due to precipitation is normally stabilizing, since the salinity effect overwhelms any thermal effect. As with temperature, the rate of change of salinity in the OML can be written as:

$$\frac{\partial S}{\partial t} = \frac{(\dot{E} - \dot{P})S}{D} + (U_{\text{geo}} + U_{\text{Ek}}) \cdot \nabla S + \frac{(W_e + W_{\text{Ek}})(S - S_b)}{D} + \kappa \nabla^2 S \quad [4]$$

with the salinity changing as a result of the net surface fresh-water flux, advection, entrainment, and diffusion. Subscripts and symbols are as with the temperature equation.

Momentum is a conserved quantity. Therefore, the momentum flux must also be in balance at the air-sea interface. This just means continuity of stresses at the interface, in particular, tangential stresses,

$$\tau_a + \tau_{\text{pr}} = \tau_w + \tau_{\text{wv}} \quad [5]$$

where τ_a is the air-side stress (the shear stress applied by the atmosphere to the ocean), τ_w is the water-side stress (negative of the drag exerted by the ocean on the atmosphere), τ_{wv} is the momentum flux radiated out by propagating surface waves generated by the wind, and τ_{pr} is the momentum flux due to any precipitation, approximately equal to $\rho_w \dot{P} U_{10}$, where ρ_w is the density of water and U_{10} is the air velocity at anemometric height (10 m). Enormous effort has gone into parameterizing τ_a in terms of the atmospheric variables and τ_{wv} , since they determine the value of τ_w . The momentum flux to the surface waves is a drag exerted on the atmosphere and is therefore important. It is especially important in the initial stages of development of the wave field (meaning short fetches or immediately following a change in the wind), since a considerable fraction of the momentum flux from the atmosphere goes then into generating the waves, with the remainder going directly into ocean currents. For a mature wave field, however, near-equilibrium conditions prevail and most of the momentum flux put into waves is immediately 'lost' and transferred to the currents, and τ_{wv} can be safely neglected. It is difficult in practice to compute τ_{wv} without a wave model, and it is a normal practice to ignore τ_{wv} and put $\tau_w = \tau_a$ in the absence of any precipitation.

Sensible and latent heat fluxes, as well as long wave back radiation at the ocean surface occurs across the skin of the ocean. Within this skin layer, which is on the order of a millimeter in thickness, there is usually a sharp drop in temperature of a few tenths of a degree Celsius (Figure 6). Exchanges of heat, momentum, and mass through this region are by molecular processes. This cool skin plays an important role in air-sea transfer processes, because of its influence on air-sea temperature and humidity differences. Transport of dissolved gases also occurs across a molecular sublayer of similar thickness to the thermal sublayer as do the transports of momentum and mass. Below this thin layer, turbulent processes dominate, driven by momentum and energy exchanges from the atmosphere to the ocean. The presence of the wave sublayer and a mobile interface, whose roughness as felt by both the ocean and the atmosphere is dynamically determined, provides the most important distinction between an ABL over land and that

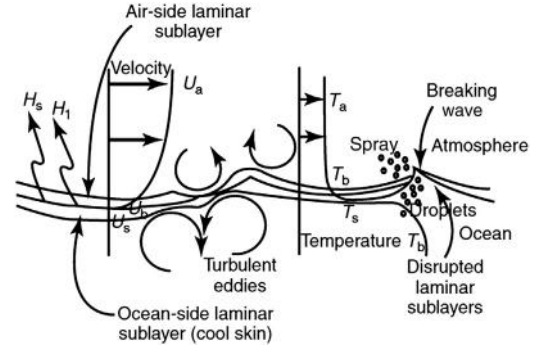


Figure 6 A sketch of the air-sea interface showing the molecular sublayers on the air and the water sides. The heat flux through sublayer

over the ocean. Consequently, surface layer similarity laws derived from the ABL over land are imprecise analogies when applied to the ABL over water and the OML. For the OML, the fraction of the OML affected directly by wave orbital velocities and hence wave dynamics can be large.

However, the wave generated currents and turbulence in the upper ocean do not just simply coexist. Recent advances in wave-turbulence interactions indicate that transfer of energy to turbulence from waves occurs even without wave breaking. The mechanism is the so-called Stokes production of TKE. The Stokes drift current (a small steady Lagrangian residual current in the direction of surface wave propagation which decays exponentially with depth) produced by wind waves augments the mean current produced by the wind stress acting at the surface. The Reynolds stresses acting on the vertical shear of the Stokes drift in the water column extract energy from waves and transfer it to turbulence, in exactly the same manner in which Reynolds stresses extract energy from the mean currents through interaction with the mean shear. The salient parameter is the modified Langmuir number (introduced by Kantha)

$$\text{La}_K = \left[\frac{(\vec{\tau}_w / \rho_w) \cdot (\vec{U}_S)}{u_*^3} \right]^{1/3} = \left(\frac{|\vec{U}_S| \cos \theta}{u_*} \right)^{1/3} \quad [6]$$

where $\vec{U}_S$ is the Stokes drift velocity at the surface and θ the angle between wind stress and the direction of wave propagation. The larger this number, the higher the Stokes production of TKE. Stokes production constitutes an important source of turbulent mixing in the upper ocean (Figure 7) and is now being incorporated into ocean models. This mechanism is also of importance to wind-generated waves over the ocean, since it constitutes an important sink mechanism for surface gravity waves in the presence of turbulence in the OML. The Stokes dissipation rate of wind wave energy in the ocean is more than 2.5 terawatts on the average, which is more than the dissipation rate of wind wave energy in all the surf zones around the ocean margins.

Langmuir Cells

Langmuir cells are organized counter-rotating cells in the surface layer (Figure 8), with axes roughly aligned with the wind. Their

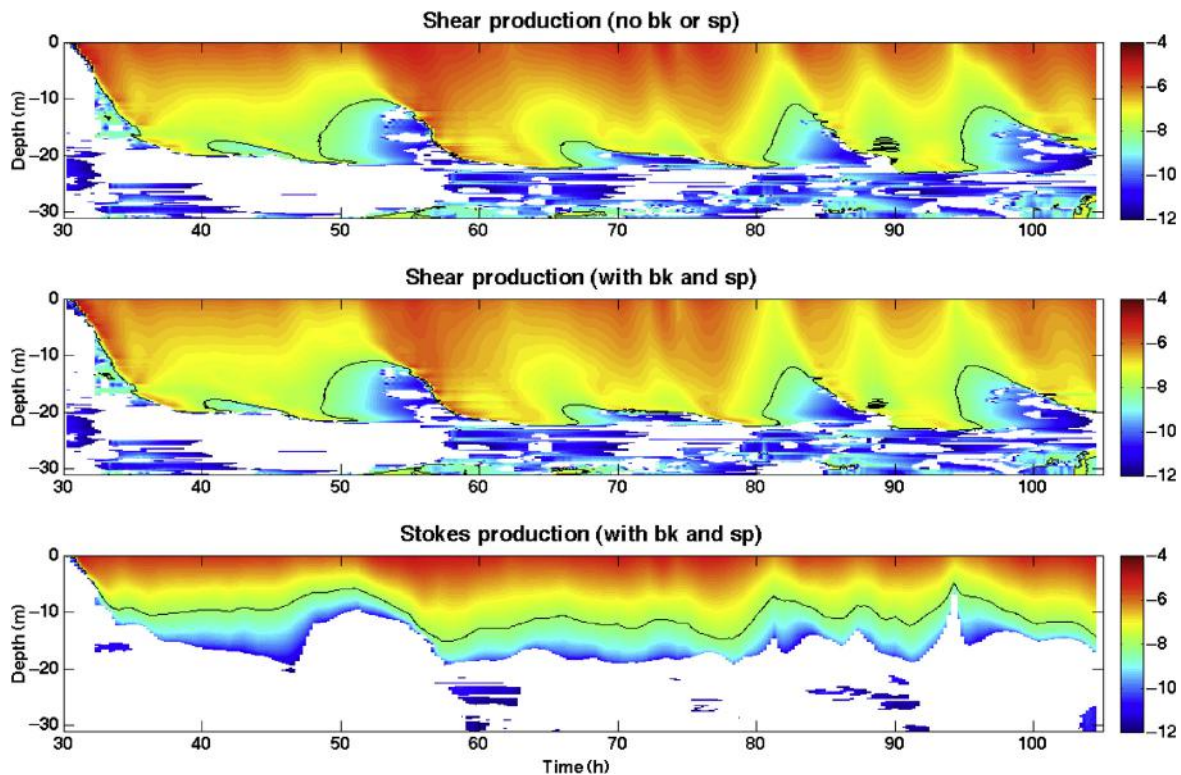


Figure 7 Shear production of TKE (top two panels) and Stokes production of TKE (bottom panel) shown in logarithmic scale (in units of $\text{m}^2 \text{s}^{-3}$) for a station in the Baltic Sea. The top panel is for the case without wave breaking or Stokes production. Middle panel shows including them produces some minor changes in shear production. The bottom panel shows that Stokes production is comparable in magnitude to shear production and hence is important to include in OML models. Reproduced from Kantha, L., Lass, H.U., Prandke, H., 2010. A note on Stokes production of turbulence kinetic energy in the oceanic mixed layer: observations in the Baltic Sea. *Ocean Dynamics*, 60, 171–180.

presence is indicated by surface convergence at the boundary of counter-rotating cells, where seaweed and flotsam accumulate. The convergence region is also made visible by white-capping and bubble entrainment due to breaking of small scale waves

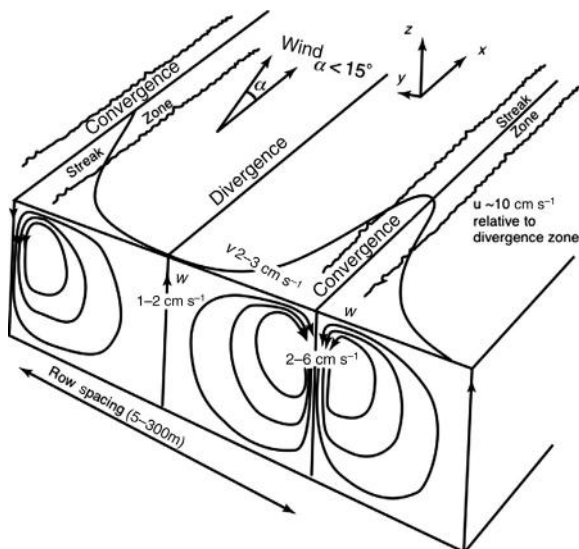


Figure 8 Langmuir cells and the associated velocities. Note the 'wind-rows' and the strong near-surface vertical velocities associated with the convergence zones of a pair of counter-rotating cells.

resulting in parallel white lines roughly aligned with the wind and roughly uniformly spaced. Bubble clouds in water are also concentrated at surface convergence zones and are visible in side-scan sonar observations.

Organized motions in the OML such as those due to large eddies and Langmuir circulations are important to upper ocean mixing and transport. Langmuir cells can be quite vigorous with downward vertical velocity immediately below the convergence zone as high as a few tens of cm s^{-1} . These motions are capable of not only injecting additional energy into the OML for mixing, but also transporting particles such as phytoplankton deep into the OML. They are however transient processes and are not easily quantified. Their existence depends on the presence of a surface wave field with the associated Stokes drift. An instability brought on by the vortex force term that appears in the momentum equations due to the interaction of the Stokes drift with the mean shear in the upper layers leads to the formation of Langmuir cells. Thus they are unique to the oceans since they result from a subtle interaction of the wind-driven turbulence and the Stokes current drift produced by surface gravity waves. Observational programs and advanced computer models such as large eddy simulations are helping us understand such large-scale features of the OML. Langmuir cells can have a dramatic effect on shallow diurnal mixed layers and can wipe out the strong diurnal peaks in the SST that would otherwise be manifest when solar insolation is strong and winds are weak. Their effect on mixing in deep

mixed layers can also be significant, even though the Stokes drift decays rapidly with depth.

The characteristic velocity scale for Langmuir circulation is,

$$V_L = [(\vec{\tau}_w / \rho_w) \cdot (\vec{U}_S)]^{1/3} \quad [7]$$

$$\text{For monochromatic waves, } V_L = [u_*^2 (ka)^2 C \cos \theta]^{1/3} \quad [8]$$

where u_* is the friction velocity, k is the wavenumber, C the phase speed, and a the amplitude of the surface waves. Note that $La_K = \frac{V_L}{u_*}$. Clearly, the strength of the Langmuir cells depends on both the Stokes drift and wind stress, so that strong winds and small waves can have an influence similar to that of weak winds and large waves.

Deep Convection

There exist a few regions in the high latitude oceans where convection is both deep and long lasting. Under the present climatic conditions, in the Greenland Sea, the Labrador Sea, and the western Mediterranean Sea in the northern hemisphere, and in the Weddell Sea in the southern hemisphere, strong, prolonged wintertime cooling occurs at the surface, leading to deep convective layers that extend over most of the water column. More recently, deep convection has been shown to occur along oceanic fronts, particularly in the Japan (East) Sea. Deep convection in the open ocean is the means by which the deep ocean is ventilated and its thermal structure maintained. The resulting meridional thermohaline circulation, and the poleward oceanic heat transport from the low latitudes to the midhigh latitudes associated with it, has a major influence on the climate at these latitudes. In these few deep convection regions, the stable stratification in the water column that normally isolates the abyss from the atmosphere is broken down violently by strong convective cooling at the surface. The associated timescales are much larger than the inertial period ($2\pi/f$, where $f = 2\Omega \sin \varphi$ is the Coriolis parameter, Ω being the angular velocity of Earth, and φ being the latitude) and hence deep convection occurs under the influence of Earth's rotation.

There are three phases of the most common form of deep convection: (1) The preconditioning phase, where the prevailing large-scale circulation brings the weakly stratified deep water masses closer to the surface for the stratification to be gradually eroded by strong sustained surface cooling during early winter. This phase is crucial to the whole process. Deep convection in the open ocean unassociated with oceanic fronts is found only in regions with cyclonic circulation that causes an upward doming of the isotherms. In the Labrador Sea, the cyclonic circulation is due to the West Greenland and Labrador Currents hugging the continental slope. In the Mediterranean, the cyclonic Lions Gyre provides the preconditioning. Strong, sustained cooling is also essential to breakdown the stratification built up in the upper layers during the previous spring and summer. It is interesting that even stronger heat losses ($\sim 1000 \text{ W m}^{-2}$) occur in the oceans during wintertime cold air outbreaks off the east coasts of continents leading to strong cyclogenesis in the atmosphere, but not deep-water formation because of the brevity of the events. In the polar oceans, the air-sea temperature differences during off-ice wind conditions

can reach 30–40 °C, and if sustained long enough can lead to intermediate and deep dense water formation, examples being the Weddell Sea in the Antarctic, the Sea of Okhotsk, and the Arctic shelves. (2) Eventually the stratification breaks down. A strong cooling event lasting several days with heat losses of 500–1000 W m^{-2} , brought on by air-sea temperature differences of 8–12 °C, and strong wind bursts are common. Deep convection ensues with intense plumes a few kilometers in size reaching down to as deep as 2–3 km. (3) The cooling weakens, and the well-mixed water mass in the convective 'chimney' spreads laterally, undergoing baroclinic instabilities in the process, and mixes with the ambient waters. Stratification is restored and the stage is set for the next cycle.

The most relevant parameters in deep convection are the buoyancy flux (due to both sensible and evaporative heat losses) B_0 that can reach values of $1\text{--}3 \times 10^{-7} \text{ m}^2 \text{ s}^{-3}$, the inertial frequency f , the depth of the mixed layer D , 1000–3000 m, and the Rossby radius of deformation $a = \frac{c}{f}$

(where c is the internal gravity wave speed), typically a few kilometers. Rossby radius is indicative of the horizontal scales of motion under the influence of ambient rotation. Under conditions where the rotational effects dominate, the

relevant length and velocity scales are $l_{dc} = \left(\frac{B_0}{f^3}\right)^{1/2}$ and

$$u_{dc} = \left(\frac{B_0}{f}\right)^{1/2}. \text{ The associated Rossby number } Ro = \frac{u_{dc}}{l_{dc}f},$$

characterizing the relative importance of rotation, is unity. The relevant Rayleigh number, a parameter of importance

in thermal instability and free convection, $Ra = \frac{B_0 D^4}{\nu k^2}$,

is very large on the order of 10^{26} . Note that in the atmosphere, where rotational effects are not important, the relevant length scale is D_{ABL} , the thickness of the ABL and the relevant velocity scale is the Deardorff velocity scale $w_* = (B_0 D_{ABL})^{1/3}$, indicative of the typical velocities in a convective ABL.

Numerical Models

OML models can be grouped into roughly two categories: bulk (or slab) models and diffusion models. Bulk models attempt to model the OML in an integral sense. The governing equations are integrated over the mixed layer so that the momentum and heat balance of the entire mixed layer, under the action of momentum and buoyancy fluxes at the ocean surface, can be considered. The major problem in bulk mixed layer modeling arises from the necessity to parameterize the advance and retreat of the OML under the action of surface momentum and buoyancy fluxes. The deepening of the OML is due to the entrainment rate at the base of the OML, which is determined by turbulence processes. The entrainment rate has been a subject of both laboratory and field experiments. It is also necessary to know the depth to which turbulence generated at the surface can penetrate under the action of a stabilizing buoyancy flux at the surface (as for example, during a rainstorm or strong solar heating), in order to determine the OML depth under these conditions. Bulk models parameterize the entrainment (OML deepening) and

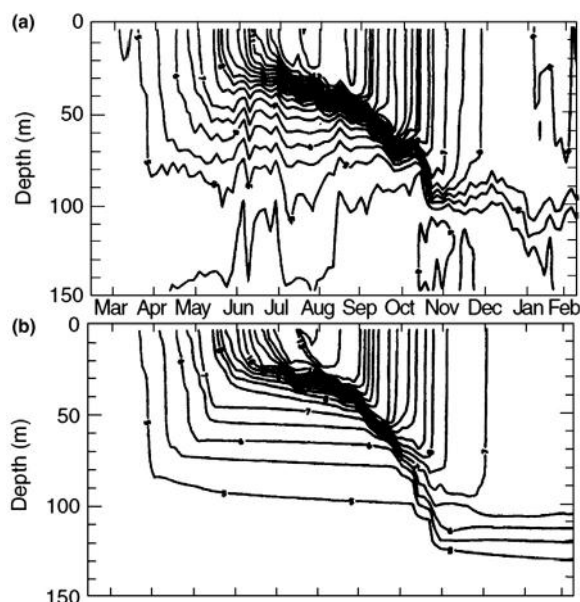


Figure 9 Comparison of (a) observed and (b) one-dimensional mixed layer model-simulated mixed layer temperatures at Ocean Station PAPA in the North-Eastern Pacific. Simulations are remarkably good given the large uncertainties associated with forcing parameters and the neglect of advection effects.

‘detrainment’ (OML retreat) in terms of surface fluxes of momentum and buoyancy, using the well-known properties of turbulence in geophysical mixed layers and/or observational evidence. The K-profile parameterization (KPP) model is a good example. It was built originally by borrowing ideas from the ABL over land and has been routinely included in ocean models used for long-term simulations. The surface wave effects were however absent in the original formulation.

Diffusion models parameterize turbulent mixing in the OML. Those based on higher moments of governing equations close the governing equations for turbulence quantities at some level by judicious modeling of the unknown higher moments and other terms. Once turbulence is thus quantified, it is a straightforward matter to deduce the mixing intensity. In second-moment closure models, the turbulence equations are closed at the second-moment level; conservation equations for turbulence Reynolds stresses, heat fluxes, and scalar variances are solved by modeling the unknown third-moment turbulence quantities and pressure-strain rate and pressure-scalar gradient covariance terms by appealing to physical intuition and/or observational evidence (and lately to Large Eddy Simulations of turbulence). However, the complexity of this approach is at least an order of magnitude more than that of the simpler models cited above, since there is now a need to solve partial differential equations governing second moments in addition to the usual ones for mass, momentum (for U and V), and scalar (for T and S) conservation. Attempts have therefore been made to simplify the set by once again utilizing certain aspects of

turbulence such as its departure from the state of local isotropy. The result is a hierarchy of models, of which the most useful for geophysical applications is the model that consists of one conservation equation for TKE (half the square of the turbulence velocity macroscale q), and a set of algebraic equations for turbulence second-moment quantities. The resulting simplicity and the potential ‘universality’ of application are of particular interest in modeling OMLs. Since the most basic description of turbulence is incomplete without quantifying its macro length scale l , this set is supplemented either by auxiliary information on the turbulence length scale, or by utilizing an equation for a quantity that includes the turbulence length scale such as dissipation rate (q^3/l) or the product of the turbulence length scale and twice the TKE ($q^2 l$). **Figure 9** shows an example of the accuracy attainable with a current generation OML model. For the current state of OML modeling using second moment closure, see the 2011 review article by L. Kantha, listed in the references. See also the book *Marine Turbulence*.

To summarize, there has been a significant improvement in our understanding of the OML in the last decade since the first publication of this Encyclopedia. However, much work remains and any further improvement in understanding and quantification of the state of the OML requires a concerted and combined observational, theoretical, and modeling effort to address the holes in our knowledge. While significant progress has been made in modeling the OML under strong stable stratification, modeling mixing under convective or near-convective conditions, where counter-gradient transport by large eddies is important, is still imprecise.

See also: Air Sea Interactions: Freshwater Flux; Momentum, Heat, and Vapor Fluxes; Sea Surface Temperature; Surface Waves. **Boundary Layer (Atmospheric) and Air Pollution:** Complex Terrain; Modeling and Parameterization; Stably Stratified Boundary Layer; Surface Layer.

Further Reading

- Apel, J.R., 1987. *Principles of Ocean Physics*. Academic Press, San Diego.
- Baumert, H., Simpson, J., Sundermann, J. (Eds.), 2005. *Marine Turbulence*. Cambridge University Press, Cambridge.
- D’Asaro, E.A., 2014. Turbulence in the upper ocean - ocean mixed layer. *Annual Review of Marine Science* 6, 4.1–4.15.
- Kantha, L., 2011. Modeling turbulent mixing in the global ocean: second moment closure models. In: Marcuso, R.J. (Ed.), Chapter 1, *Turbulence: Theory, Types and Simulation*. Nova Publishers.
- Kantha, L.H., Clayson, C.A., 2000. *Small Scale Processes in Geophysical Flows*. Academic Press, San Diego, CA.
- Kantha, L., Lass, H.U., Prandke, H., 2010. A note on Stokes production of turbulence kinetic energy in the oceanic mixed layer: observations in the Baltic Sea. *Ocean Dynamics* 60, 171–180.
- Marshall, J., Schott, F., 1999. *Open-ocean convection: observations, theory and models*. Reviews of Geophysics. American Geophysical Union.
- Soloviev, A., Lukas, R., 2006. *The Near-Surface Layer of the Ocean: Structure, Dynamics and Applications*. Springer.
- Tennekes, H., Lumley, J.L., 1982. *A First Course in Turbulence*, second ed. MIT Press, Cambridge, MA.
- Thorpe, S.A., 2005. *The Turbulent Ocean*. Cambridge University Press, Cambridge.

Stably Stratified Boundary Layer

L Mahrt, Oregon State University, Corvallis, OR, USA

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

The traditional formulation of the stable boundary layer is briefly reviewed including the vertical structure, similarity theory, and stability parameters. A number of complications are considered including very stable conditions where the primary source of the weak turbulence may be detached from the surface, sometimes related to the nocturnal jet. The role of a vegetation canopy and surface heterogeneity are considered. Nonstationarity and meandering of the wind vector for very stable weak wind conditions are discussed.

Introduction

Stable boundary layers are most commonly generated by surface radiative cooling or advection of warm air over a cooler surface. Turbulence transports heat toward the cooler surface. The actual transfer of heat to the surface occurs by thermal conductivity through a thin laminar sublayer, perhaps only a few millimeters thick. However, the rate of this heat conduction is dictated by the downward turbulent transport of heat to the laminar sublayer. Therefore, we concentrate on the turbulent transfer.

The downward transport of heat corresponds to downward buoyancy flux that destroys turbulent kinetic energy. That is, the energy required to push warm, light air downward and lift cold heavy air upward comes at the expense of the turbulence kinetic energy (conversion of turbulence kinetic energy to potential energy). Therefore, for a given wind speed and surface roughness, the turbulence is weaker with downward buoyancy flux compared to the case of no buoyancy flux or upward buoyancy flux. In general, the buoyancy flux is downward when the boundary layer is stably stratified; that is, when the potential temperature increases with height. To simplify the discussion of a very complex subject, we make an idealized distinction between cases where such buoyancy effects are weak or strong and then move on to consider general aspects of the stable boundary layer, which apply to a wide range of stabilities.

Basics

Turbulent fluctuations of a generic variable ϕ are mathematically defined as deviations from a mean value as in eqn [1].

$$\phi' \equiv \phi - \bar{\phi} \quad [1]$$

The overbar defines a time-, space-, or volume-average. ϕ may be potential temperature, one of the three velocity components, or concentrations of gases such as water vapor or carbon dioxide.

We proceed with time averaging of the flow at a fixed point in space. The turbulent vertical flux of quantity ϕ is then computed as expression [2].

$$\overline{w'\phi'} \quad [2]$$

The vertical flux may then be averaged over a longer averaging time to reduce random flux errors. The vertical divergence of the vertical flux appears in the budget equation for ϕ , such as the equations of motion for momentum, provided that the averaging process defined by the overbar satisfies Reynolds averaging. This requirement is satisfied by simple unweighted averaging. With filtering or a moving average, additional flux terms result in the budget equations. This topic is beyond the scope of this article.

Fluxes in the stable boundary layer are often formulated in terms of local vertical gradients through use of an eddy diffusivity as in eqn [3], where K_ϕ is the eddy diffusivity for vertical transfer of variable ϕ .

$$\overline{w'\phi'} = -K_\phi \frac{\partial \bar{\phi}}{\partial z} \quad [3]$$

Such a formulation is likely to be useful in the stable boundary layer where the eddies are locally generated by shear and are of small vertical scale. In contrast, the eddy diffusivity is not likely to be useful in the convective daytime boundary layer where large eddies transport according to a vertical gradient over the depth of the entire boundary layer (see **Thermodynamics: Moist (Unsaturated) Air**).

The eddy diffusivity is often related to a form of the Richardson number. The flux Richardson number is defined in eqn [4], where $\bar{\theta}_v$ is the averaged virtual potential temperature.

$$Rf \equiv - \frac{(g/\bar{\theta}_v) \overline{w'\theta'_v}}{\overline{w'u'}(\partial \bar{u}/\partial z) + \overline{w'v'}(\partial \bar{v}/\partial z)} \quad [4]$$

Often, the influence of moisture fluctuations on the buoyancy are neglected so that θ_v is replaced by θ . The numerator of Rf is the buoyancy-destruction or generation of the turbulence kinetic energy, while the denominator is the shear-generation of turbulence. The following considers positive Richardson numbers where the turbulence is destroyed by buoyancy effects, which can also serve as a definition of the stable boundary layer.

Approximating the fluxes in terms of an eddy diffusivity and rotating the x -coordinate into the direction of the mean vertical shear, the flux Richardson number becomes as shown in eqn [5], where K_m and K_h are the eddy diffusivities for momentum and heat, respectively.

$$Rf = \frac{(g/\bar{\theta}_v) K_h \partial \bar{\theta}_v / \partial z}{K_m (\partial \bar{u} / \partial z)^2} \quad [5]$$

The flux Richardson number then becomes Ri/Pr , where Pr is the eddy Prandtl number, defined as k_m/k_h , and Ri is the gradient Richardson number defined by eqn [6].

$$Ri = \frac{(g/\bar{\theta}_v)\partial\bar{\theta}_v/\partial z}{(\partial\bar{u}/\partial z)^2} \quad [6]$$

In numerical models and with observations, the gradients are approximated with finite differencing between different discrete vertical levels, designated by the δ operator, which leads to the layer Richardson number Ri_L given by eqn [7].

$$Ri_L = \frac{(g/\bar{\theta}_v)\delta\bar{\theta}_v\delta z}{(\delta\bar{u})^2 + (\delta\bar{v})^2} \quad [7]$$

A bulk Richardson number for the entire boundary layer is formed by choosing the two levels to be the surface and the top of the boundary layer; that is, replacing δz in eqn [7] with the boundary layer depth and replacing $\delta\bar{u}$ and $\delta\bar{v}$ with the wind components at the boundary layer top.

In some studies, the turbulence is considered to be first generated when the layer Richardson number decreases to 0.25, referred to as the critical Richardson number. This expectation is based on linear instability theory in terms of the gradient Richardson number. However, turbulence has been observed at much larger values of the layer Richardson number. This could be due in part to nonstationarity where the turbulence decays over a finite period after the layer Richardson increases above the critical value. However, the critical value of the layer Richardson number may increase with the distance between the vertical levels, δz . Apparently, turbulence may occur in sublayers smaller than δz even if the layer Richardson number is large.

The eddy diffusivity, K_ϕ , is sometimes formulated in terms of a Richardson number. Some formulations specify the eddy diffusivities to vanish at the critical Richardson number, while other formulations, recognizing the above complexities, specify the eddy diffusivity to gradually decrease with increasing Richardson number. The eddy Prandtl number K_m/K_h increases with the Richardson number, apparently owing to some momentum transfer by gravity waves.

If the eddy diffusivity represents the vertical turbulent flux over a horizontal area, the concept of a critical Richardson number is not applicable. No matter how large the Richardson number becomes, based on spatially averaged variables, some intermittent turbulence may occur somewhere within the area. Such a horizontal area might correspond to a grid area in a numerical model.

Weak Stability

With weak buoyancy effects, the boundary layer turbulence is generally continuous in time and space. The strength of the turbulence decreases with height and vanishes at the top of the boundary layer. These conditions often occur during the night with significant wind speed or cloudy conditions. The more stable case, corresponding to clear skies and weak winds, is discussed in the next section. Three idealized layers can be defined in the weakly stable boundary layer (Figure 1).

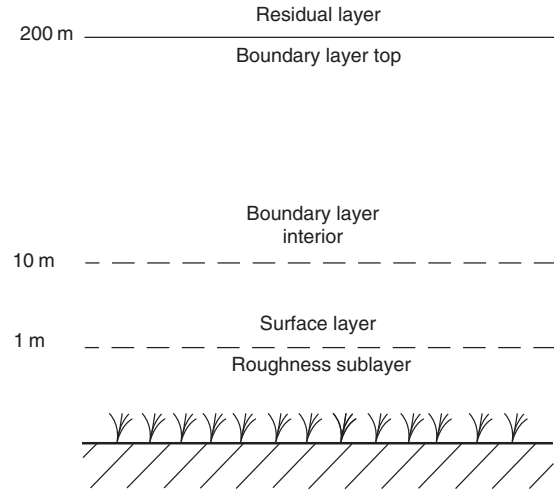


Figure 1 Idealized layering of the stable boundary layer. With strong stability, the vertical structure is less organized. The numerical values of the heights above ground are just examples. For example, the boundary-layer depth may be as small as a few meters or larger than 500 m.

Roughness Sublayer

In the roughness sublayer close to the plant canopy, the time-averaged flow varies spatially on the scale of the plant elements and a universal flux–gradient relationship seems unobtainable. For example, the flux immediately above a plant will be different from the flux at the same level between plants. Although the roughness sublayer has been studied in some detail, the generality of such studies is unknown.

Surface Layer

Above the roughness sublayer, in the surface layer (Figure 1), the individual plant elements no longer induce small-scale horizontal variation of the time-averaged flow. Then, the height above ground and the Obukhov length (defined below) are the only relevant length scales, in which case Monin–Obukhov similarity theory is valid. This theory requires that the flow is stationary and homogeneous and that $z \ll h$ where h is the boundary layer depth and z is the height above ground. Sometimes, the surface layer is assumed to extend upward to about 10% of the boundary layer depth. This condition on depth allows neglect of the height variation of the fluxes in the surface layer.

The Obukhov length is defined as in eqn [8], where κ is the von Karman constant and g is the acceleration of gravity, $(w'\theta'_v)$ the kinematic virtual heat flux (buoyancy flux) in the surface layer, assumed to be a good approximation to the surface flux, and u_* is the surface friction velocity, defined in eqn [9].

$$L \equiv -\frac{\bar{\theta}_v u_*^3}{kgw'\theta'_v} \quad [8]$$

$$u_* \equiv \left(\overline{w'u'^2} + \overline{w'v'^2} \right)^{1/4} \quad [9]$$

In eqn [9] $\overline{w'u'}$ and $\overline{w'v'}$ are the components of the momentum flux in the surface layer, also assumed to be good

approximations to the surface momentum flux. The Obukhov length represents the relative importance of mechanically generated turbulence (shear-generated turbulence associated with the roughness of the surface) to the buoyancy-destruction (or generation) of turbulence. The ratio z/L is the stability parameter, which ranges from zero at neutral conditions (no virtual heat flux) to values on the order of one and greater for very stable conditions.

The surface stress can be related to the wind speed, V , with the bulk formula, written as in eqn [10], where C_D is the drag coefficient.

$$u_*^2 = C_D V^2 \quad [10]$$

The drag coefficient is formulated in terms of Monin-Obukhov similarity theory as in eqn [11].

$$C_D = \left[\frac{k}{\ln(z/z_{om}) - \psi(z/L)} \right]^2 \quad [11]$$

The parameter κ is the von Karman constant, typically taken to be about 0.4. The stability function ψ is specified to be a function of z/L , based on parametrizations available in the suggested Further Reading at the end of the article. For stable conditions, the value of ψ is negative, associated with reduction of turbulent mixing by stratification. Thus, the drag coefficient is small compared to the case of no virtual heat flux ($\psi(z/L) = 0$). The roughness length for momentum z_{om} represents the influence of the surface on the momentum flux-gradient relationship in the surface layer. Generally, a rougher surface corresponds to a larger surface roughness length, which in turn leads to greater turbulence. Corresponding relationships for the transfer coefficients for surface fluxes of heat, moisture, and other scalar quantities are defined in terms of scalar roughness lengths. Over vegetated surfaces, scalar roughness lengths are not as well behaved as that for momentum. The stability functions for scalars can be found in the Further Reading. These references also include completely different similarity theories for prediction of the drag coefficient and transfer coefficients, which are defined in terms of the geostrophic wind.

When evaluating similarity theory from actual data, one must assess the importance of artificial self-correlation. For example, C_D is correlated with ψ even if the fluxes are randomly generated variables. This is because both C_D and ψ are functions of the surface friction velocity u_* . This self-correlation can be eliminated by replacing the stability parameter z/L with the surface bulk Richardson number Ri_b given by eqn [12], which is a form of eqn [7].

$$Ri_b \equiv \frac{(g/\theta_v)(\theta_v(z) - \theta_v(sf))z}{(V(z))^2} \quad [12]$$

Here, $\theta_v(sf)$ is the surface virtual potential temperature, corresponding to the ground surface temperature (often the surface radiation temperature), and $V(z)$ is the wind speed at level z , which is within the surface layer.

Boundary Layer Interior

Above the surface layer, the approximation of height-independent fluxes is no longer valid. The fluxes normally decrease with height and reach small values at the top of the

boundary layer. However, if the Obukhov length is redefined in terms of local fluxes, similarity theory can be restored in the form of local scaling in which z/L is replaced by z/Λ , where Λ is the Obukhov length based on the local fluxes at height z . The length scale Λ reduces to the Obukhov length when z is in the surface layer.

The vertical structure above the surface layer is often expressed as functions of z/h , where h is the boundary layer depth. For example, the eddy diffusivity is sometimes modeled as a function of z/h . The classical Ekman solution is a special case corresponding to a balance between the vertical divergence of the momentum flux, the horizontal pressure gradient, and the Coriolis force. The simplest form of the Ekman solution assumes a height-independent eddy diffusivity.

Boundary Layer Top

Sometimes the top of the stable boundary layer is viewed as an entrainment layer where turbulent eddies engulf air from above the boundary layer and mix it downward. This concept applies with growing stable boundary layers, usually corresponding to windy conditions. During fair weather conditions, the air above the nocturnal boundary layer would have been within the deeper convective boundary layer in the daytime. This layer above the nocturnal boundary layer is referred to as the 'residual layer' (Figure 1). This layer 'remembers' the daytime convective layer in that the stratification usually remains weak. As a result, modest vertical shear generates turbulence in the residual layer.

The depth of the stable boundary layer is sometimes modeled by assuming that the bulk Richardson number for the boundary layer is equal to a critical value or constantly adjusts toward the critical value. Other formulations express the boundary layer depth in terms of length scales such as u_*/f , where f is the Coriolis parameter. The later formulation does not apply at low latitudes.

Strong Stability

The very stable boundary layer occurs with clear nocturnal skies and weak winds or with advection of warm air over a much cooler surface. The very stable boundary layer is of considerable practical importance. The absence of significant mixing allows buildup of high concentrations of contaminants near the surface. Frost damage is most likely to occur in the very stable case where the downward turbulent heat flux from above is suppressed (Figure 2), leading to even greater surface cooling.

With strong buoyancy effects, the turbulence is more likely to be intermittent, with brief episodes of turbulence separated by intervening periods of relatively weak or unmeasurably small fluctuations. The term intermittency is somewhat ambiguous in that all turbulence is considered to be intermittent to the degree that the fine scale structure occurs intermittently within larger eddies. This intermittency within a given large eddy is referred to as fine scale intermittency.

Global intermittency defines the case where eddies on all scales are missing or suppressed on a scale that is large compared to the large eddies. Global intermittency is sometimes viewed as a sequence of events beginning with reduction

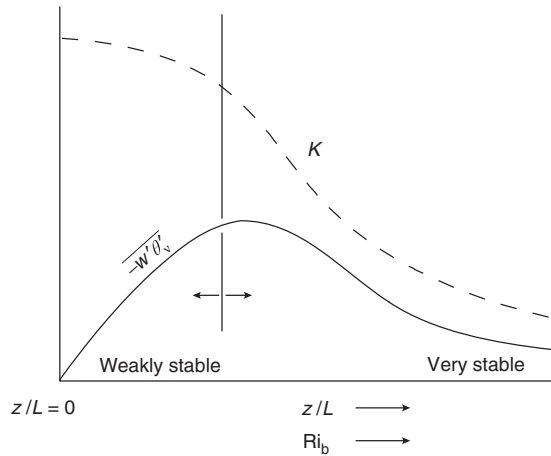


Figure 2 Typical variation of the eddy diffusivity and the virtual heat flux with stability, z/L , or Ri_b .

of the mean shear by mixing and the corresponding increase of the Richardson number to values greater than the critical value. This stage is followed by collapse of the turbulence. Without vertical mixing, the mean shear rebuilds such that the Richardson number falls below the critical value and turbulence is regenerated. With this interplay between the turbulence and mean shear, the Richardson number stays close to the critical value.

Additional characteristics of the very stable case may include significant gravity wave transport, layering of turbulence, and the absence of a well-defined boundary layer top. The boundary layer may assume an upside-down form, where the principal turbulence is detached from the surface. This detachment may be only temporary since decoupling with the surface leads to flow acceleration above the stable layer adjacent to the surface. The resulting development of shear generates turbulence and recoupling with the surface (Figure 3). Thus, the concept of an upside-down boundary layer, top-down transport of turbulence kinetic energy toward the surface and cyclic behavior about a critical value of the bulk Richardson number may all be related.

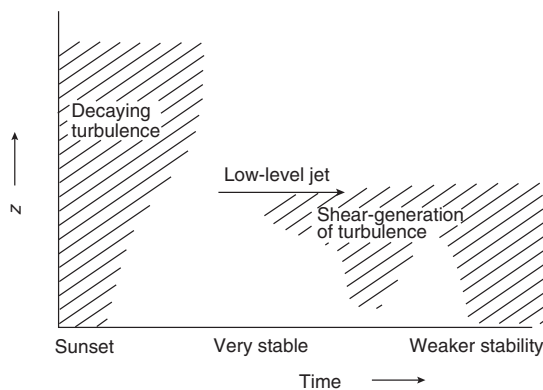


Figure 3 Hypothetical schematic of one type of diurnal variation of the nocturnal boundary layer. Shaded regions indicate significant turbulence.

However, a unified picture or theory of the very stable case does not exist owing to both the difficulty of measuring weak intermittent turbulence with existing instrumentation and the complex multitude of different physical processes with strong stability.

Very Stable Surface Layer

The turbulence and eddy diffusivity near the surface are sometimes observed to decrease rapidly as z/L exceeds a critical value, as qualitatively sketched in Figure 2. This rapid decrease marks the beginning of the very stable regime. Unfortunately, the critical value of z/L varies between observational studies. The downward heat flux seems to reach a maximum value at the beginning of the transition between the weakly and very stable cases. With stronger stability, the vertical velocity fluctuations are suppressed, which reduces the downward heat flux. For weaker stability, the temperature fluctuations are small as a result of small vertical gradients of potential temperature. The downward buoyancy flux vanishes as the vertical gradients of potential temperature vanish (neutral stability).

Virtually all models apply Monin–Obukhov similarity theory to the first model level above the surface to estimate surface fluxes, even for very stable conditions where Monin–Obukhov similarity may apply only to levels below the first model level, or may not apply at all. No other practical formulations exist.

The turbulence in the very stable surface layer may be intermittent, sometimes related to downward bursts from turbulence generated aloft. Clear air radiative flux divergence, transport by nonlinear gravity waves, drainage flows, and surface heterogeneity (see later) may invalidate assumptions required for Monin–Obukhov similarity theory.

Interior of the Boundary Layer

Above the surface layer in very stable conditions, the eddies may no longer ‘feel’ the influence of the ground and the height above ground is no longer a relevant variable. This is referred to as ‘ z -less’ stratification. This flow cannot be modeled in terms of dependences on z/h since z is not a relevant variable and the boundary layer depth may not be definable.

Upside-Down Structure

Some very stable boundary layers take the appearance of an ‘upside-down’ boundary layer where the main source of turbulence is elevated and detached from the surface. In one scenario, turbulence adjacent to the surface collapses in the early evening, while some decaying turbulence remains at higher levels for a finite period of time (Figure 3). Flow above the surface inversion layer is now decoupled from the surface stress and accelerates, often leading to a low-level jet. The associated shear may induce turbulence above the surface inversion layer, which bursts downward toward the surface (Figure 3). Downward bursting of turbulence is one cause of intermittent turbulence at the surface.

With the upside-down structure, the usual concept of a boundary layer breaks down. We retain the term ‘boundary layer’ because the turbulence may be intermittently coupled to

the surface. In addition, the detached turbulence is often generated by shear associated with drainage flows, low-level jets, and gravity waves that are induced by surface processes. The detached turbulence may also be associated with the residual layer above the surface inversion, which in the early evening is characterized by decaying turbulence from the daytime boundary layer.

Similar behavior may occur with the advection of warm air over a much colder surface. For example, in advection of warm daytime air from land over colder water, the cooling of the air in contact with the water surface leads to strong stratification and suppression of the turbulence near the surface.

Low-Level Jet and Elevated Turbulence

Shear-generation of turbulence at the top of the surface inversion layer sometimes results from the formation of a low-level nocturnal jet. A nocturnal maximum in the wind profile can be generated by cooling over sloped terrain or inertial effects, described below.

Cooling over sloping terrain leads to a time-dependent, height-dependent horizontal pressure gradient force, which generates low-level flow. Small-scale sloping terrain leads to nonhydrostatic downslope drainage flows, which can be significant even over weak slopes if skies are clear and the large-scale flow is weak. On larger scales, the pressure field is hydrostatic. On still larger scale slopes, the flow becomes influenced by the Coriolis parameter.

As the second mechanism, nocturnal low-level jets are driven by ageostrophic flow caused by daytime frictional effects in concert with collapse of the daytime boundary layer in late afternoon. The turbulence and stress divergence vanish above the new thin nocturnal boundary layer, leading to an imbalance between the Coriolis and horizontal pressure gradient terms. This in turn induces an inertial oscillation. Spatial variations of the inertial oscillation generate vertical motions and adjustment of the pressure field, which feeds back upon the inertial oscillation.

Mesoscale Variations

Nocturnal mesoscale motions include cold air drainage and meandering motions, which might propagate from outside the local domain. Meandering motions refers to the ‘flopping around’ of the wind vector, most obvious with weak nocturnal airflow. Although meandering has been attributed to a variety of physical mechanisms, its link to these mechanisms has not been established from observations. It is not clear whether such motions systematically degrade Monin–Obukhov similarity theory. These mesoscale motions become relatively more important for the very stable case, partly because the turbulence is weaker.

Mesoscale motions may be locally generated in stably stratified flows by turbulence that decays into mainly horizontal motions, sometimes referred to as one type of meandering motion. Vortex motions in the horizontal plane (vertical vorticity) may merge into larger vortices, corresponding to upscale energy transfer. These theoretical

expectations have been difficult to verify with actual atmospheric data. Such meandering can be augmented by topography. These layered motions randomly enhance the shear between layers, which in turn intermittently regenerates local turbulence.

Vegetation Canopy

Most land surfaces are covered with some type of vegetation canopy, which substantially impacts the structure of the nocturnal boundary layer near the surface. The vegetation causes microscale variability on the scale of individual plants, leading to the roughness sublayer. With a partially open canopy, the subcanopy air is significantly stratified at night, partly owing to direct radiative cooling of the ground surface between the trees (Figure 4(a)). The net radiative cooling under each tree is small owing to downward longwave radiation emitted by the tree. With a closed canopy, the subcanopy flow may be stably stratified in the daytime and less stratified at night when the inversion is concentrated in the upper part of the canopy (Figure 4(b)). Some of the air cooled at the canopy top sinks into the subcanopy, leading to vertical mixing of the subcanopy flow.

Irregularity of the canopy top due to individual trees may induce gravity waves in the nocturnal boundary layer.

Enhanced Influence of Surface Heterogeneity

Since vertical mixing in the very stable boundary layer is due to weak small eddies, small-scale heterogeneity may become important even though the influence of the same surface heterogeneity in the daytime, convectively heated, boundary layers is eliminated by large convective eddies. Such small-scale heterogeneity might include clumps of trees or bushes, isolated buildings, or horizontal variation of heat flux from the soil due to variations of soil type and moisture on

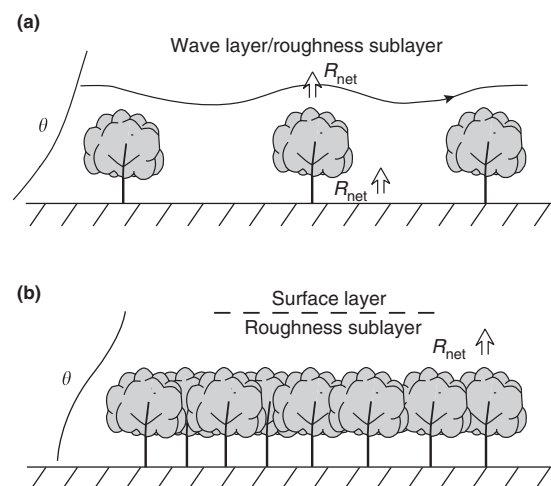


Figure 4 The nocturnal boundary layer for (a) an open canopy and (b) a closed canopy. R_{net} is the net radiation. The potential temperature profile is indicated on the left.

horizontal scales as small as tens of meters. The nocturnal boundary layer is slow to adjust and is therefore easily influenced by surface heterogeneity. On the other hand, mesoscale surface heterogeneity over flat surfaces with stable stratification is less likely to generate its own secondary circulation since stationary vertical motion fields are inhibited by the stratification through hydrostatic pressure adjustments. For example, rising motion induced by a relatively warm surface area generates adiabatic cooling in the stratified atmosphere, which acts to increase the surface pressure, which in turn opposes the horizontal inflow.

See also: Agricultural Meteorology and Climatology. **Boundary Layer (Atmospheric) and Air Pollution:** Complex Terrain; Convective Boundary Layer. **Land-Atmosphere Interactions:** Canopy Processes.

Further Reading

- Arya, S.P.S., 1988. *Introduction to Micrometeorology*. Academic Press, New York.
- Garratt, J.R., 1992. *The Atmospheric Boundary Layer*. Cambridge University Press, New York.
- Geiger, R., 1961. *The Climate near the Ground*. Harvard University Press, Cambridge, MA.
- Hinze, J.O., 1975. *Turbulence*. McGraw-Hill, New York.
- Kaimal, J.C., Finnigan, J.J., 1994. *Atmospheric Boundary Layer Flows: Their Structure and Measurement*. Oxford University Press, New York.
- Oke, T.R., 1987. *Boundary Layer Climates*, second ed. Methuen, London.
- Panofsky, H.A., Dutton, J.A., 1984. *Atmospheric Turbulence – Models and Methods for Engineering Applications*. Wiley, New York.
- Sorbjan, Z., 1989. *Structure of the Atmospheric Boundary Layer*. Prentice-Hall, Englewood Cliffs, NJ.
- Stull, R.B., 1990. *An Introduction to Boundary Layer Meteorology*. Kluwer Academic, Boston, MA.
- Tennekes, H., Lumley, J.L., 1972. *A First Course in Turbulence*. MIT Press, Cambridge, MA.
- Yoshino, M.M., 1975. *Climate in a Small Area*. Tokyo Press, Tokyo.

Surface Layer

GL Geernaert, US Department of Energy, Washington, DC, USA

© Published by Elsevier Ltd.

Synopsis

The atmospheric surface layer represents the medium for nearly all activity on planet Earth. With a depth of order one kilometer, the physics governing profiles of wind, temperature, and gases can be represented in terms of a small number of key turbulent processes framed by Monin–Obukhov similarity (MOS) theory. This article summarizes the assumptions and applications of MOS theory to the surface layer, and key scientific challenges are summarized where MOS theory is exposed to conditions that violate basic assumptions.

Introduction

The lowest layer of the atmosphere, the surface layer, directly influences the daily activities of nearly all life on planet Earth. Extending up to an altitude of order 50–100 m, the surface layer possesses dynamical, physical, and chemical characteristics long recognized as having a controlling influence over a wide range of human and societal interests. For example, even as early as 600 BC, the force of the wind acting against the flow of the Nile was already being used as a secondary process included in flood forecasts in ancient Egypt. About 2000 years ago, simple climatologies of near-surface wind directions were the basis of weather forecasts in the Mediterranean, for use in both military and commercial activities. By the late Middle Ages in both Europe and China, sailing ships were being designed to exploit the greater wind speeds, which regularly were observed at increasing heights above the surface. In fact, the Chinese used kites to study surface layer wind profiles as early as 1400 years ago. More recently, i.e., in the early twentieth century, pollutant dispersion models were developed, based on anthropogenic emissions from point sources (such as smokestacks) and compiled wind statistics, in order to estimate pollutant concentrations at different distances downwind. Some of the greatest challenges we face today in areas such as remote sensing involve understanding the relationship between the surface layer and the underlying terrestrial biosphere and oceans.

The role of the atmospheric surface layer in surface energy exchanges is an essential ingredient in understanding both local and global climatology. The dominant local process governing the surface layer energy balance is a large downward flux of solar radiation, absorbed at the earth's surface during the day that in turn is converted rapidly to heat. While some of the surface heat energy is transferred down into the surface canopy and soil system, a significant fraction of the heat is transferred vertically upward into the surface layer. If the surface layer is highly turbulent (as it is during most afternoons), the upward flux of heat is highly efficient, thus leading to significant vertical mixing. In stark contrast, e.g., during calm clear nights, turbulence is often suppressed, thus leading to minimal vertical transfer of heat and the more extreme cold nights often observed, e.g., in midwinter. Note that if the nocturnal surface layer is otherwise windy, the associated higher turbulence levels can provide an efficient mechanism for transferring heat rapidly down to the surface, thus reducing the rate of nighttime cooling. From both

a scientific and a societal perspective, it is the presence or absence of turbulence in the surface layer which acts as a control over the degree to which humans and surface-based ecosystems are exposed to diurnal and climatic extremes.

Surface Layer: Turbulence, Eddies, and Logarithmic Profiles

In the atmospheric sciences, the depth of the surface layer is, by convention, defined as 10% of the depth of the full planetary boundary layer (PBL). This implies that the surface layer can range from 10 m height up to 150 m height, depending in most part on the mixing rates and degree of large-scale convection or subsidence. Within the surface layer, quantities such as wind speed, temperature, and humidity vary significantly with height; above the surface layer, these quantities tend to approach more constant values with height until one reaches the inversion that overlies the PBL.

The surface layer is also defined as the lowest layer of the atmosphere where the vertical scales of all of the turbulent eddies are limited by proximity to the surface, thus all surface layer turbulent eddies will have length scales that on average are proportional to height above the surface. It is this basic assumption when combined with information on turbulence spectra that allows one to derive the characteristic logarithmic wind, temperature, and humidity profiles, as well as the various sampling methodologies to obtain turbulent fluxes within the surface layer.

The derivation of the wind, temperature, and humidity profiles start with the assumption that the wind vector can be represented with a mean downstream velocity, $\langle U \rangle$, and its fluctuating component, u' , i.e.,

$$U = \langle U \rangle + u' \quad [1]$$

In analogy, one can also write the vertical velocity as,

$$W = \langle W \rangle + w' \quad [2]$$

Since the downstream component of the vertical momentum flux, τ , normalized by density, ρ , may be expressed as $(\tau/\rho) = -UW$, one may combine with eqns [1] and [2] to easily arrive at $(\tau/\rho) = \langle -u'w' \rangle$. For this expression, the averaging time be long enough so that $\langle W \rangle = 0$. Furthermore, in this expression, it is a common practice in the atmospheric sciences to assume that momentum flux is positive downwards. With

this practice, the vertical momentum flux may also be assigned a characteristic friction velocity, u^* , such that:

$$u^{*2} = \langle -u'w' \rangle \quad [3]$$

While field observations have generally shown that the value of u^* is nearly constant with height within the atmospheric surface layer, the reality is that u^* must be allowed to vary by up to 10% of its surface value. In addition, while the mean cross-stream velocity is defined to be 0, there is sometimes a small $\langle -u'v' \rangle$ contribution to u^{*2} , such that the momentum flux vector and wind vector are no longer aligned.

The next step in the derivation of wind, temperature, and humidity profile relations is based on similarity laws applied to engineering studies of flow over plates. Here we invoke the assumption that an eddy viscosity coefficient K_m developed from flows over flat plates may be applied to the atmospheric surface layer, such that

$$K_m \partial U / \partial z = u^{*2} \quad [4]$$

Letting K_m be defined by the Prandtl mixing length hypothesis, i.e., $K_m = \langle h'^2 \rangle |\partial U / \partial z|$, where h' is the turbulent mixing length, and letting h' be proportional to height above the surface, z , eqn [4] may now be rewritten as:

$$\partial U / \partial z = u^* / kz \quad [5]$$

where k is the von Karman constant. Integration of eqn [5] over z leads to the classical logarithmic wind speed profile:

$$U = (u^* / k) \ln [z / z_0] \quad [6]$$

where z_0 is the roughness length. In analogy to eqn [6], the vertical profile of temperature and humidity may also be derived and will exhibit a logarithmic behavior.

The logarithmic nature of profiles has over the years become a defining feature of the atmospheric surface layer, and measurements have corroborated eqn [6] and similar equations for heat and moisture for near neutral stratifications. A major limitation of the early profile relations was a theoretical deficiency for treating nonneutral stratifications, and more recent limitations are associated with the roles that non-stationarity, spatial heterogeneity, discontinuities, internal boundary layers, and local flux divergence.

Surface Layer Profiles, Fluxes, and Stratification

During the 1950s, the Monin–Obukhov similarity (MOS) theory emerged as a major theoretical development able to address the influence of atmospheric stratification on the logarithmic profiles and fluxes (Monin and Obukhov, 1954). These investigators began their analysis by determining which terms of the turbulent kinetic energy (TKE) equation were the most important controls over atmospheric mixing rates, and they used the results of their analyses to construct an appropriate stratification parameter.

Most commonly written in the following form, the TKE budget is:

$$\begin{aligned} \partial e / \partial t = & -U \partial e / \partial x - \langle u'w' \rangle \partial U / \partial z - \partial \langle w'e' \rangle / \partial z \\ & - \rho^{-1} \partial \langle w'p' \rangle / \partial z - (g/T_v) \langle w'T_v' \rangle - \nu \langle (\partial u_i' / \partial x_i) \\ & \times (\partial u_i' / \partial x_i) \rangle \end{aligned} \quad [7]$$

In eqn [7], the quantities u' , v' , and w' represent the fluctuating downwind (x -direction), crosswind (y -direction), and vertical (z -direction) wind velocity components. The parameter e is TKE per unit volume (E), normalized by density; the instantaneous contribution to the TKE may be denoted as $e' = 1/2 u_i' u_i'$. The virtual temperature, T_v , is defined as $T(1 + 0.61q)$, where T is temperature and q is humidity. The first term on the right hand side (r.h.s.) of eqn [7] represents the rate of change of TKE due to advection. The second term represents shear production. The third term is the flux divergence of TKE, while the following term is the divergence of pressure flux. The fifth term on the r.h.s. is TKE production or loss due to buoyancy, and the last term is the loss of TKE due to viscosity, i.e., the dissipation rate. If conditions are steady state and horizontally homogeneous, and letting the dissipation rate be represented by ε , eqn [7] is simplified to:

$$0 = -\langle u'w' \rangle \partial U / \partial z - (g/T_v) \langle w'T_v' \rangle - \varepsilon - R \quad [8]$$

where the imbalance terms are combined as $R = \rho^{-1} \partial / \partial z (\langle w'p' \rangle + \rho \langle w'e' \rangle)$; most investigators assume that the combination of imbalance terms is small enough to be ignored. Substituting eqn [5] into eqn [8], a dimensionless stability parameter was introduced, based on the ratio of the shear-induced TKE to the buoyancy-induced TKE, i.e.,

$$z/L = gkz \langle w'T_v' \rangle / T_v u^{*3} \quad [9]$$

The quantity L is hereinafter referred to as the Monin–Obukhov length.

For wind speed and temperature profiles, estimates of profile gradients could account for variations in atmospheric stratification, i.e.,

$$\partial U / \partial z = (u^* / kz) \phi_U \quad [10a]$$

$$\partial T / \partial z = (T^* / kz) \phi_T \quad [10b]$$

where the functions ϕ_U and ϕ_T are formulated in terms of z/L .

Integration of eqn [10a] and [10b] over height yields logarithmic profiles of wind speed and temperature, i.e.,

$$U - U_0 = (u^* / k) \{k \ln (z/z_0) - \Psi_U\}^{-1} \quad [11a]$$

$$T - T_0 = (T^* / k) \{k \ln (z/z_{0T}) - \Psi_T\}^{-1} \quad [11b]$$

where Ψ_U and Ψ_T are stability functions that are related to ϕ_U and ϕ_T , z_0 and z_{0T} represent roughness lengths for momentum and temperature, and $T^* = \langle w'T' \rangle / u^*$. For neutral stratifications, i.e., where $\partial T_v / \partial z = 0$, the quantities Ψ_U and Ψ_T equal 0; positive/negative values of Ψ_U and Ψ_T correspond to unstable/stable stratifications. The functional relationships between ϕ_U and Ψ_U as well as between ϕ_T and Ψ_T may be found in most references following this article.

The Bulk Aerodynamic Relationships

For many of the more important applications of surface layer theory to, e.g., pollution dispersion, atmospheric and oceanic modeling, loads on structures, and remote sensing inversion algorithms, estimates of the surface fluxes based on easily obtained observations of mean surface layer quantities (such as

wind speed, temperature, etc.) are highly desirable. It was partly to satisfy these needs that similarity theory was developed.

Rearrangement of eqn [11a] and [11b] yields the bulk aerodynamic relations for momentum and temperature fluxes, e.g.,

$$u^{*2} = C_D(U - U_0)^2 \quad [12a]$$

$$u^*T^* = C_H(U - U_0)(T - T_0) \quad [12b]$$

where the quantities C_D and C_H are respectively the drag coefficient and the Stanton number, i.e.,

$$C_D = [k/\ln(z/z_0) - \Psi_U]^2 \quad [13a]$$

$$C_H = [k/\ln(z/z_0) - \Psi_U][k/\ln(z/z_T) - \Psi_T] \quad [13b]$$

A similar relation can be derived for the humidity profile and its coefficient, the Dalton number.

The beauty of the derivation of the drag and other flux coefficients is that they, at least in theory, do not depend on wind speed and are weakly dependent on height, roughness, and typical ranges of atmospheric stratification. We note herein that strongly stable stratifications do, in fact, lead to much smaller flux coefficients, i.e., when compared with neutral or unstable conditions. In the experimental literature, the coefficients are often normalized to values representative of neutral stratifications (e.g., C_{DN} and C_{DH}) and for a standard height of 10 m above the surface so that they may be easily compared to other data sets and/or integrated into models as general parameterizations. Details of the relationships between the flux coefficients and their neutral stratification counterparts are summarized in Geernaert (1999).

Sampling of Fluxes in the Surface Layer

Most modern research into the characteristics of the surface layer has relied on very complicated measurements collected over a wide variety of wind, stratification, and roughness conditions. Flow distortion, platform motions, and the simplifying assumptions associated with specific methods have made both the measurements and subsequent interpretations of data difficult. Until the 1970s, the most common method of gathering surface layer flux data was based on vertical profiles of the average wind speed, temperature, and humidity. This method relied on the rearrangement of eqn [11] into a form where plots of wind speed vs the natural logarithm of height yielded information on the friction velocity and surface roughness simultaneously. As an example, for wind speed, the slope is (u^*/k) and the bias is $[(u^*/k)(\ln\{z/z_0\} + \Psi_U)]$ when one plots data in a form where $(x, y) = (U - U_0, \ln z)$.

During the 1970s and 1980s, the newly developed sonic anemometer and other fast-response instruments for measuring temperature and humidity rapidly replaced the profile technique, thus making possible directly measured turbulent fluxes using eddy correlation. The turbulent fluxes, i.e., $\langle u'x' \rangle$, where x' is either the fluctuating component of wind speed, temperature, or gas, could now be easily computed with field observations, as long as fluctuations at altitudes of order 10 m could be sampled at 5 Hz or more and the

averaging time was long enough that $\langle W \rangle$ is negligible. To guarantee that $\langle W \rangle$ is sufficiently small, it is a routine practice to sample the domain for at least 30 min and much longer during stably stratified conditions.

Much of the recent scientific progress in understanding the dynamics governing turbulence spectra, asymmetric turbulence, coherent structures, intermittency, and internal boundary layers was made possible with rapid sampling techniques that were led by the sonic anemometer and other fast-response instruments. The inertial subrange of surface layer turbulence was also discovered with these methodologies to have uniquely identifiable surface layer properties and scaling laws, and this discovery helped drive a major body of surface layer research starting during the 1970s and 1980s. Relevant for eddy length scales of order 1 cm up to order 10 m, the TKE density, $S(k_\lambda)$, within the inertial subrange was discovered to follow a $-5/3$ wavenumber (k_λ) power law:

$$S(k_\lambda) = \alpha \varepsilon^{2/3} u^{*2} k_\lambda^{-5/3} \quad [14]$$

where the quantity ε is the TKE dissipation rate and α is an accepted value of 0.52.

During the 1980s and 1990s, the application of eqn [14] to the dissipation method became a popular sampling technique especially for overocean research, owing to the relative insensitivity of this method to the platform motions. The dissipation method also requires shorter averaging time than for the eddy correlation technique, thus giving it an added advantage. For the case of momentum fluxes, the dissipation method is easily derived from the combination of eqns [14] and [10a]; by ignoring R and assuming near neutral stratifications, one obtains:

$$S(k_\lambda) = \alpha(kz)^{-2/3} u^{*2} k_\lambda^{-5/3} \quad [15]$$

A simplified form of the TKE budget may now be expressed as $u^{*3} = kz/\Phi_e$. Using empirical functions that describe Φ_e (referring to references at the end of this article), one-dimensional spectra of horizontal wind speeds easily produces estimates of u^* with this technique.

In addition to momentum, the dissipation method has been applied to infer temperature, water vapor, and carbon dioxide fluxes. As techniques begin to emerge for other compounds that have had difficulty sampling with very high frequency, the dissipation method will no doubt gain popularity across a broad range of species and applications. The challenges, however, remains how to deal with imbalance terms in both the TKE budget and the variance budgets for temperature and gaseous compounds, particularly if the domain of interest exhibits substantial spatial variability (e.g., in coastal zones).

Trace Gas Exchange

Simple models for the surface exchange of trace gases exhibit more complexity than the bulk aerodynamic relations described in the previous sections. Not only does turbulent transfer in the surface layer play a key role, but one must consider the reactions of the compound with both other atmospheric species and with the surface. For chemical compounds that have reaction time scales, which are long

compared with surface layer turbulence time scales, the constant flux layer assumption may be invoked. This applies to gases such as CO_2 , O_2 , and SO_2 . However, if the reactions are relatively fast, in particular with respect to the time scales of surface layer turbulence, a flux divergence of the particular compound will occur. Species such as HNO_3 and NH_3 are examples, and the following equation must be solved in this case:

$$U \partial c / \partial x + \partial \langle w'c' \rangle / \partial z = S \quad [16]$$

In eqn [16], S represents a production or loss of chemical concentration, c , based on chemical reactions. Because of the risk of other processes, which are more important than turbulent transport, this balance tests the limits of similarity theory applicable to the surface layer. In spite of this shortcoming, the scientific community has proceeded to produce a body of literature in which many of the limiting assumptions can be relaxed if one can use reference heights which are closer to the surface, i.e., where characteristic turbulence time scales are much smaller.

An alternative approach to measuring gas fluxes is based on the concept of the deposition velocity, v_d . Here one uses various additive resistances associated with the flow of compound, c :

$$v_d = \langle w'c' \rangle / (c - c_s) = (r_a + r_b)^{-1} \quad [17]$$

where r_a is the aerodynamic resistance governing turbulent transport of species c ; and r_b is the surface resistance, governing the diffusion transport over the laminar sublayer. While the surface concentration, c_s , is set to 0 for many gases, there are others such as the biogenic reactive gases (e.g., NH_3) that can exhibit high surface concentrations; in these cases, the surface concentrations need to be estimated based in part on models of surface biological activity. In essence, eqn [17] implies that the turbulent transport associated with r_a represents the maximum possible deposition velocity for any species undergoing air-surface exchange ($=1/C_D U$), and r_b is a correction factor that depends on the properties of the particular compound undergoing air-surface exchange. The surface resistance, r_b , is more difficult to describe. The quantity r_b is based on the two key assumptions: that there exists a relatively homogeneous laminar sublayer at the surface; and that the physical characteristics of the surface and the biological characteristics govern the rate of diffusion.

Flux Profile Relations for Quasi-Inhomogeneous Conditions

Applications of MOS theory generally require that the surface layer be characterized by spatially homogeneous and steady state conditions, and that the vertical fluxes within the surface layer are nearly constant with height. In recent years, however, a new class of scientific and engineering problems has emerged, where some of the requirements, most notably spatial homogeneity, need to be relaxed. Examples of such new problems include understanding the role of internal boundary layers on turbulence profiles associated with offshore wind farms, and evaluating subgrid surface layer variabilities within the much finer resolution atmospheric and oceanic models. Several

modified theories have been introduced to examine MOS theory under spatially heterogeneous or nonstationary conditions, some of which employ results of large eddy simulations while others explore the use of 'flux footprints' and local internal boundary layers.

In a recent study, Geernaert (2010) produced a modified form of MOS theory, applicable to quasi-homogeneous and nonstationary conditions. Assuming that advection and nonstationarity can act as proxies of external forcing that lead to local deviations of similarity theory, eqn [8a] may be rewritten into a more general form:

$$\partial U / \partial z = u^* / kz (\phi_u + H^* + J^*) \quad [18]$$

where H^* and J^* are functions that describes respectively the spatial and temporal variabilities of wind speed. Integration of eqn [18] over z will yield a drag coefficient expression that is much more complicated than eqn [13a]. In this case, the drag coefficient that is normalized for neutral stratification, horizontal homogeneity, and stationarity, i.e., C_{DR} , will be significantly different from the drag coefficient that is normalized only for neutral stratifications, C_{DN} :

$$C_{DR}^{-1/2} = C_{DN}^{-1/2} - \left\{ (z/2U) C_D^{-3/2} \right\} \left\{ \left(1 - \left[\xi C_D^{1/2} / k \right] \right) \partial U / \partial x + U^{-1} \partial U / \partial t \right\} \quad [19]$$

Analogous to eqns [18] and [19], spatial inhomogeneity and nonstationarity introduces more complicated flux profile expressions and flux coefficients for temperature, humidity, and reactive gases. See Geernaert (2010) for a review of the derivations.

Challenges and New Directions

The state of our present knowledge of the surface layer has been built from the similarity paradigm introduced in 1954 by Monin and Obukhov. This theory proved quite successful, particularly for inferring fluxes or profiles averaged over spatial scales which are greater than 25 km and temporal scales of order 1 h or more. During recent years, however, the scientific challenges given by mesoscale and boundary modelers and other applied customers involve scales up to three orders of magnitude smaller than those that the research community explored even two or three decades ago. Obviously, the problems of determining the momentum flux to swell and wind waves, and the problem of estimating fluxes for reactive gases (e.g., ammonia, nitric oxide, ozone, etc.) deposition patterns near, e.g., the edge of sensitive forested ecosystems, are good examples where the basic assumptions of similarity theory have become routinely violated. For this reason, one may expect that studies of surface layer theory will soon enter a new era, where advanced modeling and *in situ* measurement techniques will need to be combined with multiscale *in situ* and remote sensing systems.

See also: Air Sea Interactions: Momentum, Heat, and Vapor Fluxes. **Boundary Layer (Atmospheric) and Air Pollution:** Observational Techniques *In Situ*. **Land-Atmosphere**

Interactions: Trace Gas Exchange. **Numerical Models:** Parameterization of Physical Processes: Turbulence and Mixing.

Further Reading

- Garratt, J.R., 1992. *The Atmospheric Boundary Layer*. Cambridge University Press, Cambridge.
- Geernaert, G.L. (Ed.), 1999. *Air–Sea Fluxes: Physics, Chemistry, and Dynamics*. Kluwer, Dordrecht.
- Geernaert, G.L., 2010. Normalizing air-sea flux coefficients for horizontal homogeneity, stationarity, and neutral stratification. *Journal of Physical Oceanography* 40, 2148–2158.
- Kaimal, J.C., Wyngaard, J.C., Izumi, Y., Cote, O.R., 1972. Spectral characteristics of surface layer turbulence. *Quarterly Journal of the Royal Meteorological Society* 98, 563–589.
- Kitaigorodskii, S.A., 1973. *Physics of Air–Sea Interaction*, Trans. from Russian by Baruch A. Israel Program for Scientific Translations, Jerusalem.
- Lumley, J.L., Panofsky, H.A., 1964. *The Structure of Atmospheric Turbulence*. Wiley, New York.
- Monin, A.S., Obukhov, A.M., 1954. Basic turbulent mixing laws in the atmospheric surface layer. *Tr. Akad. Nauk, SSSR Geophys. Inst.* 24, 163–187.
- Panofsky, H.A., Dutton, J.A., 1984. *Atmospheric Turbulence*. Wiley, New York.
- Stull, R.B., 1989. *An Introduction to Boundary Layer Meteorology*. Kluwer, Boston, MA.
- Wyngaard, J., 2010. *Turbulence in the Atmosphere*. Cambridge University Press, Cambridge.

Urban Heat Islands

JC Luvall, DA Quattrochi, and DL Rickman, National Space Science and Technology Center, Marshall Space Flight Center, Huntsville, AL, USA

MG Estes, Jr., Universities Space Research Association, Huntsville, AL, USA

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

It is estimated that by the year 2025, 80% of the world's population will live in cities. This conversion of the natural landscape vegetation into man-made urban structures such as roads and buildings drastically alter the regional surface energy budgets, hydrology, precipitation patterns, and meteorology. Research studies from many cities have documented that these effects range from decreases in air quality, increased energy consumption, and alteration of regional climate to direct effects on human health.

Introduction

It is estimated that by the year 2025, 80% of the world's population will live in cities. The extent of these urban areas across the world can be seen in an image of city lights from the Defense Meteorological Satellite Program (Figure 1). In many areas of North America and Europe, it is difficult to separate individual cities because of the dramatic growth and sprawl of urbanized areas. This conversion of the natural landscape vegetation into man-made urban structures such as roads and buildings drastically alter the regional surface energy budgets, hydrology, precipitation patterns, and meteorology.

One of the earliest recognized and measured phenomena of urbanization is the urban heat island (UHI) which was reported as early as 1833 for London (Howard, 1833) and 1862 for Paris. The UHI results from the energy that is absorbed by man-made

materials during the day and is released at night resulting in the heating of the air within the urban area. The magnitude of the air temperature difference between the urban and surrounding countryside is highly dependent on the structure of the urban area, amount of solar insolation received during the day, and atmospheric conditions during the night. These nighttime air temperature differences can be in the range of 2–5 °C (Figure 2). Daytime air temperature differences between urban areas and the countryside may exist during the day; however, atmospheric mixing and instability reduce the magnitude. This phenomena is not limited to large urban areas, but also occurs in smaller metropolitan areas. The UHI has significant impacts on the urban air quality, meteorology, energy use, and human health. The UHI can be mitigated through increasing the amount of vegetation and modification of urban surfaces using high albedo materials for roofs and paved surfaces.

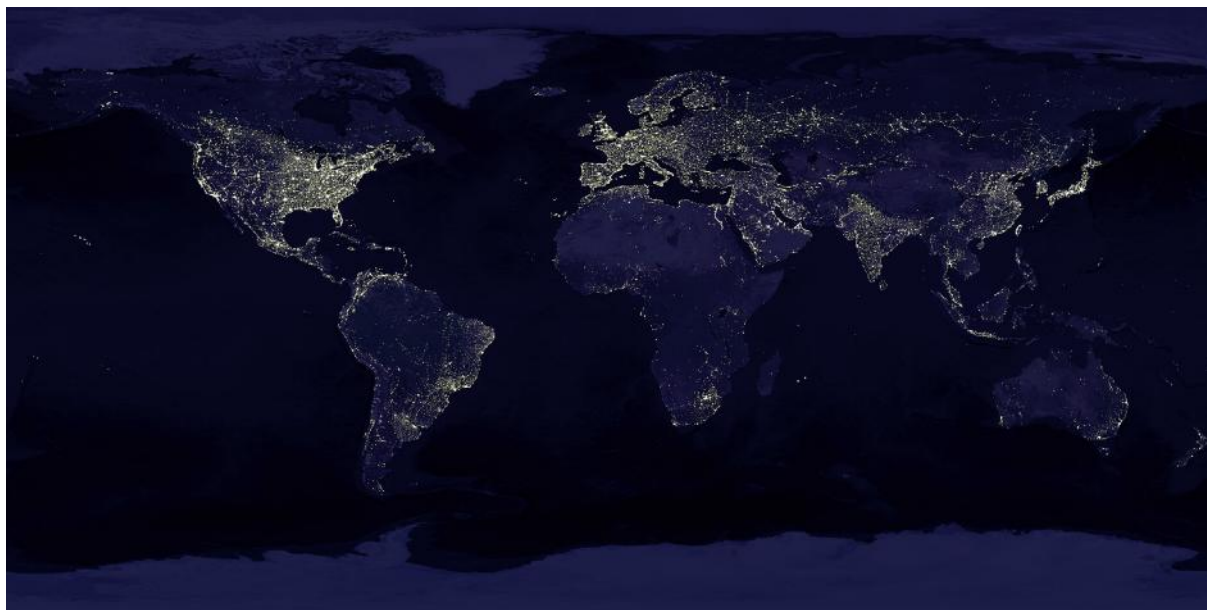


Figure 1 This image of Earth's city lights was created with data from the Defense Meteorological Satellite Program (DMSP) Operational Linescan System (OLS). Originally designed to view clouds by moonlight, the OLS is also used to map the locations of permanent lights on the Earth's surface. Craig Mayhew and Robert Simmon, NASA GSFC, based on DMSP data courtesy Christopher Elvidge, NOAA National Geophysical Data Center: Description.

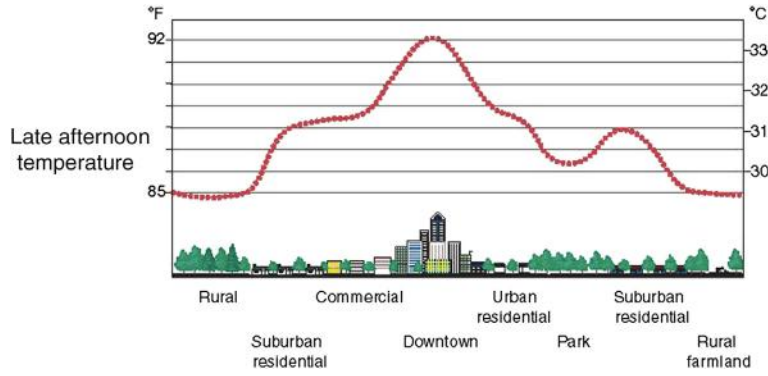


Figure 2 Idealized urban heat island profile. EPA Urban Heat Island Project.

Urban Surface Energy and Radiation Budget

To understand why the UHI phenomena exists, it is useful to define the surface in terms of the surface energy budget. Surface temperature and albedo are major components of the surface energy budget. Knowledge of it is important in any attempt to describe the radiative and mass fluxes that occur at the surface. Use of energy terms in modeling surface energy budgets allows the direct comparison of various land surfaces encountered in an urban landscape, from vegetated (forest and herbaceous) to nonvegetated (bare soil, roads, and buildings). These terms are also easily measured using remote sensing from aircraft or satellite platforms allowing one to examine the spatial variability of the urban surface. The partitioning of energy budget terms depends on the surface type. In natural landscapes, the partitioning is dependent on canopy biomass, leaf area index, aerodynamic roughness, and soil moisture status, all of which are influenced by the regional climate. In urban landscapes, coverage by man-made materials substantially alters the surface energy budget.

The net all-wave radiation balance (W m^{-2}) of landscape canopies can be determined following:

Net solar radiation, K^* , is given by

$$K^* = (1 - \alpha)(K\downarrow) \quad [1]$$

where α = site albedo and $K\downarrow$ = incoming solar radiation.

Albedo is defined as

$$\alpha = \frac{K\uparrow}{K\downarrow} \quad [2]$$

where $K\uparrow$ = reflected solar radiation.

The long wave energy ($L\uparrow$) emitted from a surface is dependent on surface temperature:

$$L\uparrow = \epsilon[\rho T^4] \quad [3]$$

where ϵ = emissivity, ρ = Stefan-Boltzman constant ($5.7 \times 10^{-8} \text{ W m}^{-2} \text{ K}^{-4}$), and T = land surface temperature (K).

The net long wave radiation at the surface is given by

$$L^* = L\downarrow - L\uparrow \quad [4]$$

where $L\downarrow$ = long wave radiation from the atmosphere (mostly due to water vapor).

Therefore, the net all-wave radiation, Q^* , can be given as

$$Q^* = K^* + L^* \quad [5]$$

Net radiation is a particularly useful term because, under most conditions, it represents the total amount of energy available to the land surface for partitioning into nonradiative processes (mass heating, evapotranspiration, biological synthesis, etc.) at the surface. In vegetated areas the amount of net radiation is dependent upon vegetation type and varies with canopy leaf area and structure.

Net radiation may be expressed as the sum of these non-radiative fluxes:

$$Q^* = LE + H \pm G \quad [6]$$

where LE = latent heat flux (both transpiration by plants and evaporation), H = sensible heat flux, and G = energy flux into or out of storage (both vegetation, urban materials, and soil).

The partitioning of LE , H , and G are dependent on the surface composition. Vegetation canopies (leaf stomata) can control transpiration rates over a wide range of soil moisture conditions and atmospheric vapor deficits. Both the physiological control of moisture loss (stomatal resistance) and leaf/canopy morphology for vegetation determines how Q^* is partitioned among LE , H , and G . In urban areas, the combination of both man-made materials and vegetation results in a spatially variable, heterogeneous mixture of surfaces that produce a complex, range of surface albedo values and significant differences in the partitioning of the surface energy budget.

The change in surface temperature as a function of time is an additional property that can be measured using remotely sensed data. Usually a separation of about 30 min results in a measurable change in surface temperature caused by the change in incoming solar radiation. Their ratio can be used to define a surface property called the thermal response number (TRN) (Luvall and Holbo, 1989):

$$\text{TRN} = \frac{\sum_{t_1}^{t_2} (Q^*)(\Delta t)}{\Delta T} \quad [7]$$

where $\sum_{t_1}^{t_2} (Q^*)(\Delta t)$ is total net radiation and ΔT change in surface temperature for time period t_1 – t_2 .

The TRN provides an analytical framework for studying the effects of surface thermal response for large spatial resolution map scales that can be aggregated for input to smaller scales, as

needed by climate models (Comarazamy et al., 2013b). The importance of TRN is that (1) it is a functional classifier of land cover types; (2) it provides an initial surface characterization for input to various climate models; (3) it is a physically based measurement; (4) it can be determined completely from remotely sensed data; and (5) it is a scale-independent measurement that can be examined from a pixel-by-pixel measurement or by extracting a polygon from the landscape feature containing multiples of pixels representing the required element. The TRN can be used as an aggregate expression of both surface properties (vegetation canopy structure and physiological condition; urban structures and material types) and environmental energy fluxes.

The remotely sensed data obtained from the aircraft and satellite, when properly calibrated allows the measurement of important terms in the radiative surface energy budget: $K\uparrow$ and $L\uparrow$ on an urban landscape scale. Additional radiative balance terms $K\downarrow$, $L\downarrow$, and Q^* can be determined by modeling atmospheric radiance properties using radiosonde data and shadow band radiometers with atmospheric models such as MOD-TRAN4 (Berk et al., 1999).

Although satellite data are very useful for analysis of the UHI effect at a coarse scale, they do not lend themselves to developing a better understanding of which surfaces across the city contribute or drive the development of the UHI effect. Analysis of thermal energy responses and energy budgets for specific or discrete surfaces typical of the urban landscape (e.g., asphalt, building rooftops, vegetation) requires measurements at a very fine spatial scale (i.e., <15 m) to adequately resolve these surfaces and their attendant thermal energy regimes. Additionally, very fine scale spatial resolution thermal infrared data, such as that obtained from aircraft, are very useful for demonstrating to planning officials, policy makers, and the general populace the benefits of planning cool communities. These benefits include mitigating the UHI effect, making cities more aesthetically pleasing and more habitable environments, and sustainable communities (Quattrochi et al., 2000).

It is important to understand the partitioning of surface energy budget given in eqns [1] and [6] and its role in determining actual surface temperatures. A significant UHI mitigation strategy is to increase the city's albedo to reduce roof and pavement warming. Equation [6] shows that there are only three pathways, which the energy from net radiation can take: (1) evaporate water (latent heat); (2) warm the air (sensible heat); and (3) reach the ground (storage). Forest vegetation has a low albedo value, similar to many rooftops. Its surface temperature may be much cooler than a roof because the energy absorbed by the surface is partitioned into latent heat (cooling the surface) rather than sensible heat (surface heating). An asphalt surface may not get as hot as a rooftop, but the energy is going into storage and released overnight. These basic principles must be understood in order to make intelligent decisions when incorporating the UHI mitigation strategies to both urban heating and air quality issues.

High spatial resolution visible and thermal data are required to quantify how artificial surfaces within the city contribute to an increase in urban heating and the benefit of cool surfaces. One good example is Atlanta, a large southeastern US city, which illustrates the complexity of surfaces found in a city. A NASA project using an airborne multispectral visible and thermal scanner (ATLAS, airborne thermal and land applications sensor) collected 10 m resolution data over Atlanta, Georgia, on 11 and

12 May 1997. These data were collected around solar noon (maximum surface heating) and about 5 h after sunset. The day was clear and cloud free. The maximum daytime air temperature was 25.5 °C and the nighttime minimum was 14.4 °C. Atlanta central business district albedo (Figure 3) averaged a low value of 0.18. The lower albedo values (0.10–0.13) consisted of vegetation and asphalt roads. A few high albedo areas were white membrane rooftops. Daytime surface temperatures (Figure 4) indicated that the greatest temperatures were the roofs (45 °C), followed by asphalt parking lots and roads. Areas of vegetation and water were the coolest, at around 22 °C. The overall average for the area is 31.7 °C. After sundown (Figure 5) it becomes evident what areas of the city store the sun's energy and slowly release it at night. The roofs are the hottest surfaces during the day, but have little thermal mass and cool off quickly. Asphalt roads and parking lots, and concrete buildings have large thermal mass and cool off much more slowly. The overall average surface temperature about 5 h after sunset was 14.7 °C. We see that the core central business district is much warmer, about at 16.6 °C and a nearby residential area was 15.0 °C. One can understand the importance of how the city is designed will influence greatly the magnitude of the UHI effect, along with the prevailing meteorological conditions.

Table 1 allows the comparison among other US cities for albedo, net radiation, surface temperature, change in surface temperature over time and the TRN for four different land use types. Generally, the industrial areas had the highest albedo and the hottest temperatures, indicating little or no vegetation. In contrast, the albedo of urban parks was low and the surface temperatures were the coolest, indicating a well-watered vegetated surface. The partitioning of Q^* can be evaluated

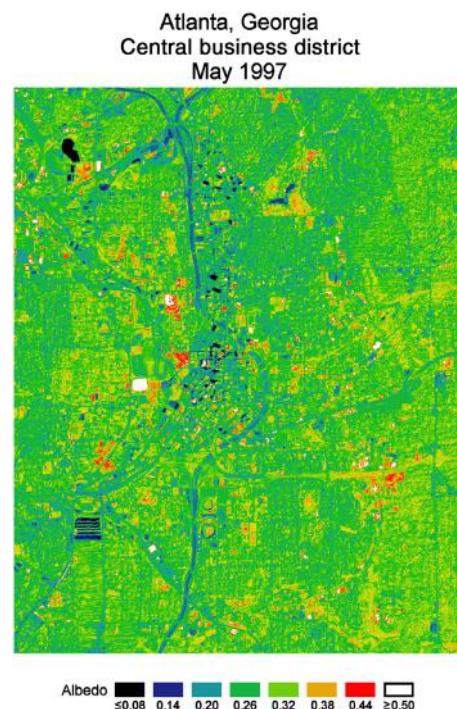


Figure 3 Atlanta, Georgia, surface albedo values produced from NASA's airborne ATLAS scanner 10 m resolution data. NASA Atlanta Urban Heat Island Project.

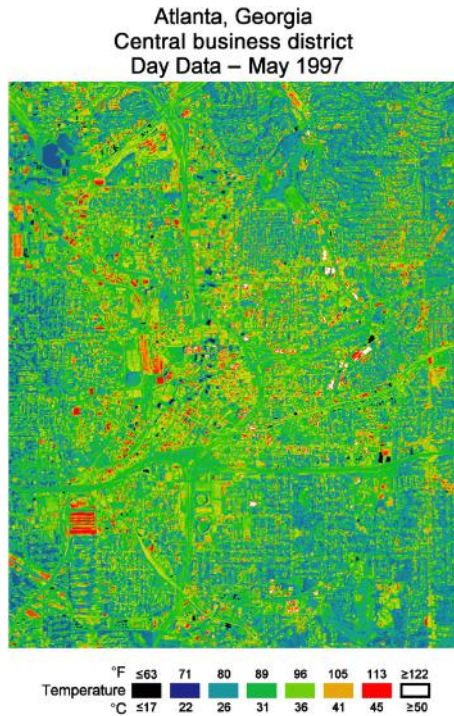


Figure 4 Atlanta, Georgia, daytime surface temperatures ($^{\circ}\text{C}$) values produced from NASA's airborne ATLAS scanner 10 m resolution data. NASA, Atlanta Urban Heat Island Project.

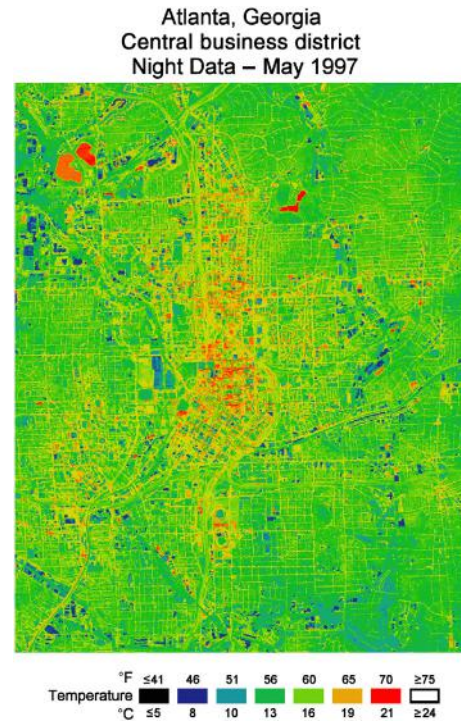


Figure 5 Atlanta, Georgia, nighttime temperatures ($^{\circ}\text{C}$) values produced from NASA's airborne ATLAS scanner 10 m resolution data. NASA, Atlanta Urban Heat Island Project.

Table 1 Net radiation and thermal response numbers (TRNs) calculated from aircraft remotely sensed data for several US cities

	Albedo	Net radiation (W m^{-2})	Average temperature ($^{\circ}\text{C}$)	Delta temperature ($^{\circ}\text{C}$)	TRN ($\text{kJ m}^{-2} ^{\circ}\text{C}$)
Baton Rouge					
Delta time 50.4 min					
Industrial	0.204	284	47.6	4.5	242
Central business district	0.185	281	45.6	2.3	490
Residential	0.162	325	38.4	3.8	335
Park	0.177	291	35.9	0.2	5903
Sacramento					
Delta time 67.5 min					
Industrial	0.273	334	44.7	3.1	428
Central business district	0.245	353	41.2	3.4	414
Residential	0.223	378	36.8	3.6	423
Park	0.195	386	30.0	2.2	701
Salt Lake City					
Delta time 120.6 min					
Industrial	0.272	253	43.8	5.3	183
Central business district	0.256	274	41.1	6.3	166
Residential	0.242	294	37.9	6.0	186
Park	0.260	316	30.8	4.1	299

EPA Urban Heat Island Project.

using the TRN. The greater the TRN the more energy is being used to evaporate water rather than to heat the air. For example, the Park in Baton Rouge has a TRN of $5903 \text{ kJ m}^{-2} ^{\circ}\text{C}^{-1}$ and a small ΔT over the measurement period. In other areas the ΔT is much greater and the TRN is

smaller. In Salt Lake City the park TRN is only $299 \text{ kJ m}^{-2} ^{\circ}\text{C}^{-1}$ compared to $701 \text{ kJ m}^{-2} ^{\circ}\text{C}^{-1}$ for the park in Sacramento and ΔT of $4.1 ^{\circ}\text{C}$ and $2.2 ^{\circ}\text{C}$, indicating that less of the surface energy is being partitioned into latent heat flux and more into heating the air. It appears that the vegetation cover in the Salt

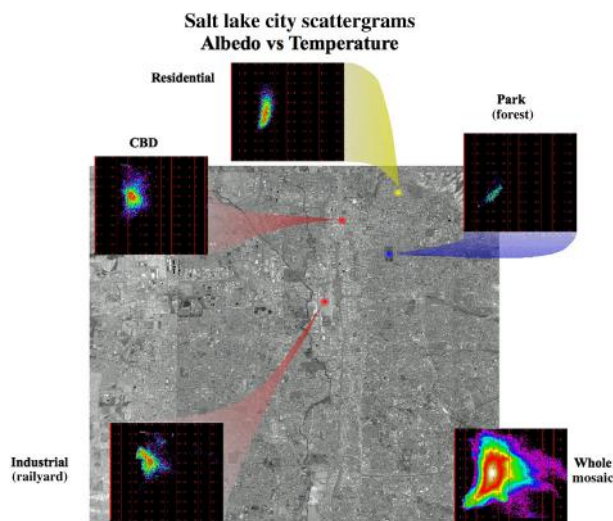


Figure 6 A scattergram of surface temperatures and albedo produced from NASA's airborne ATLAS scanner 10 m resolution data of Salt Lake City, UT.

Lake City park was not transpiring (latent heat flux) as much as the vegetation in Baton Rouge or Sacramento.

Since albedo alone does not truly reflect how the urban surface partitions energy, one needs additional information to access the 'urban fabric' of the city. Including surface temperature provides the needed additional information. Aircraft- or satellite-based remotely sensed data sets provide the needed calibrated and quantifiable data sets in physical units. Since we are working in physical units, the TRN, surface temperature and albedo classifications represents a functional classification of that surface, that can readily be incorporated into the surface parameterization of meteorological and air quality models.

A city has a distinctive 'energy print' that is a characteristic of its surface composition and processing energy (Figure 6). For example a park, i.e., dark and cool, shows up quite nicely in Salt Lake in the lower left corner of the scattergram. These scattergrams become a very powerful classification tool representing the functional classification of urban land surfaces. Within each city, each land use has a unique 'energy print' that is directly physically related to how that surface is processing energy. These 'energy prints' of the land use are unique for each city; i.e., the Sacramento central business district (CBD) scattergram is significantly different than Baton Rouge or Salt Lake City's CBD scattergram. These results again emphasize that classifications based on cover type/land use cannot be applied across a variety of cities, since they cannot represent the true energy partitioning of that surface.

Urban Growth Modeling

Urban growth models provide a tool for cities to project future conditions based on existing trends or policy modifications such as implementation of higher density development policies and preservation of environmentally sensitive areas. These future scenarios are also useful as input to physical models designed to evaluate the effect of urbanization changes to air quality and other environmental conditions.

Growth models may be applied at a variety of scales and are broadly grouped into either empirical or dynamic process simulation models. Empirical models strive to statistically match temporal trends and/or spatial variables with a set of predictor variables, whereas dynamic process models seek to represent the most important interactions among people and the environment. Cellular automata models are examples of process models (Brown et al., 2004). The Prescott Spatial Growth Model (PSGM) is an emerging dynamic process growth model that operates on a geographic information systems platform. The PSGM allows users to design future conditions by developing rules and priorities for the model to determine the future allocation of vacant land for development. The primary drivers of the quantity of future developed land are population and employment projections (Estes et al., 2010).

The PSGM was used to develop future urban scenarios in the Atlanta Metropolitan area and coupled with air quality models to determine the future impact on ozone levels and the potential benefit of proposed mitigation strategies. Results enable policy makers to determine both magnitude and spatial changes in the projected UHI and the environmental benefits of mitigation strategies. Also, future projections will benefit planners in the development of State Implementation Plans required for areas in air quality nonattainment (Quattrochi et al., 2006).

Impacts of the Urban Heat Island

The UHI has significant impacts on our urban environment. Research studies from many cities have documented that these effects range from decreases in air quality, increased energy consumption, and alteration of regional climate to direct effects on human health (Jacob and Winner, 2006; Stone, 2008).

One of the major air quality concerns is summertime ozone levels. Ozone is produced from a photochemical reaction requiring sunlight (UV component), various nitrogen oxide compounds (NO_x , derived from combustion processes), and complexes of volatile organic compounds (derived from vegetation and petrochemical products). The rate of ozone production is sensitive to the increase in air temperature and higher surface temperatures increase the amount of precursors evolved from the surface (Vukovich et al., 1979). Wintertime effects from the UHI on air quality have also been documented in Tokyo, Japan (Yoshikado and Tsuchida, 1996). The winter heat island disrupt the land-sea breeze and results in higher concentrations of air pollutants remaining over the city for longer periods of time by blocking the sea-land breeze. The UHI maintains a temperature gradient down to the inland area, thus tending to reinforce the sea breeze in the shore area and prevent it from penetrating through the urban area. This convergence zone tends to remain over urban areas for several hours.

The UHI results in overall increased energy usage by cities, particularly a large increase in the peak power demands. Calculations done by Lawrence Berkeley Laboratory found that for Los Angeles for every 1 °C increase in air temperature there was a 2.9% increase in power requirements. Since most electrical energy is produced by the combustion of fossil fuels, there is an increase in NO_x released into the urban environment thus adding to the production of ozone (Akbari et al., 1999, 2001; Akbari and Konopacki, 2005; Ashley et al., 2012; Taha, 2007).

Alterations of regional weather patterns caused by the UHI have been well documented. One of the major alterations linked to the UHI is the change in the quantity, location, and timing of precipitation in the region surrounding the urban areas caused by UHI-induced convergence processes (Shepherd et al., 2010). Generally, precipitation is increased downwind from urban areas. Studies from New York City, New York, showed lines of local convective thunderstorms originating in New Jersey during summer afternoons split into two halves that pass north and south of New York City (Loose and Bornstein, 1977). A time series of visible data (1 km resolution) collected from GOES8 for Atlanta during 3 August 1996 show the development of thunderstorms caused by the convergence initiated by the city (Figure 7) (Bornstein and Lin, 2000).

The impact of 'mega cities' in exacerbating regional heat effects can be important. Zhang et al. (2009) using a combination of ultrahigh resolution numerical simulations as fine as 500 m using a multinested (0.5, 1.5, 4.5, and 13.5 km) version of the Weather Research and Forecasting (WRF) model, a single-layer urban canopy model, and the National Land Cover Data found a significant increased UHI contribution resulting from adjacent urbanization 'upstream.' They found (modeled) that when upstream urbanization was replaced by natural vegetation UHI effects could be reduced by more than 25%.

Urbanization in island tropical cities have significant climate impacts in tropical coastal regions with the added complexity of occurring within the context of a warming climate. Comarazamy et al. (2013a) investigated the individual and combined effects of these two factors in tropical islands by use of an integrated mesoscale atmospheric modeling approach (Regional Atmospheric Modeling System) and high resolution ATLAS (10 m) data, using the northeastern region of

Puerto Rico as the test case. An ensemble of climate simulations was performed, combining two land cover land use (LCLU) and global warming scenarios. The authors show that LCLU changes produced the largest near-surface (2-m above ground level.) air temperature differences over heavily urbanized regions and that these changes do not extend into the boundary layer. The simulated influence of the global warming signal induced a positive inland gradient of maximum temperature and increased trade winds that impacted convergence zones and the resulting convection that transport heat and moisture into the boundary layer. Minimum temperatures were also increased along the coastal plains and inland lowlands.

Other meteorological impacts caused by the UHI include suppression of weak tornadoes and an increase in lightning. Studies starting in 1957 from England and Wales found that within 10 km of London there is a suppression of weak tornadoes (Olfe and Lee, 1971; Elsom and Terence Meaden, 1982). Research examining lightning for a 12-year period around Houston, Texas, reported a significant enhancement in lightning for both summer and winter centered over and downwind of the urban area (Orville et al., 2001; Ashley et al., 2012). The result of the UHI encouraging convection and the mechanical effect of the urban heat fabric causing frictional convergence are generally believed to be the most important factors responsible.

Although urbanization of natural vegetated areas is the most significant factor in creating the UHI, local and regional weather patterns determine the intensity of the UHI effect. Hu et al. (in review) using surface meteorological observations coupled with WRF simulations of Oklahoma City determined that low-level jets (LLJs) were critical in moderating the nocturnal UHI intensity and extent. Presence of a strong LLJ increased vertical mixing and instability reducing the UHI

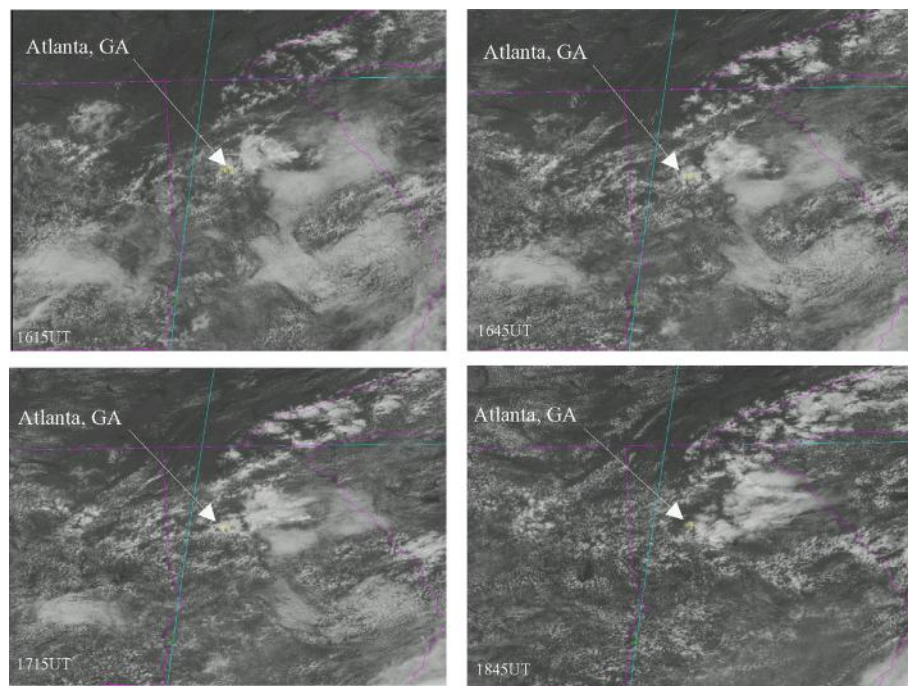


Figure 7 Thunderstorm development caused by Atlanta, Georgia's urban heat island induced convergence. GOES8 images are from 3 August 1996 starting at 1615UT. Images courtesy of Stan Kidder, Cooperative Institute for Research in the Atmosphere, Colorado State University.

intensity. Weak LLJ produced less vertical mixing and a shallow, stable boundary layer which increased the UHI.

Heat stress kills more people worldwide than any other weather phenomenon, including cyclones and floods, according to the World Meteorological Organization (WMO). Deaths rise by over 50% on average during heat waves. The UHI is a significant factor of heat stress experienced by the urban population (Gaffen and Ross, 1998). Every summer in the United States, heat stress directly causes about 2000 deaths. WMO statistics reveal the heat also is indirectly blamed for many other deaths, most from heart attacks, strokes, and other heat-related problems. The death rate during heat waves can triple in cities like New York, where the number of heat-related deaths jumped from an average of 490 to 1260 during a single 24-h period in 1966. In Chicago, the heat wave in the summer of 1995 caused at least 700 deaths (Karl and Knight, 1997). It is readily apparent from Figure 5 how the partitioning of the sun's energy into storage within the urban areas impact nighttime temperatures and increase the heat stress in the urban population.

Mitigation of the UHI

Given that the UHI impacts alteration of city surface-atmosphere energy exchange, air quality, and human health, developing ways to mitigate the UHI are becoming increasingly important to policy- and decision-makers and the general public. There are several ways that actions can be taken to cool urban surfaces, thereby reducing the amount of long wave energy that emanates from urban surfaces to force the UHI. These actions are promoted by the US Environmental Protection Agency as part of its UHI Mitigation initiative (<http://www.epa.gov/heatisland/>) and are increasingly being used by cities around the world in their efforts to reduce urban

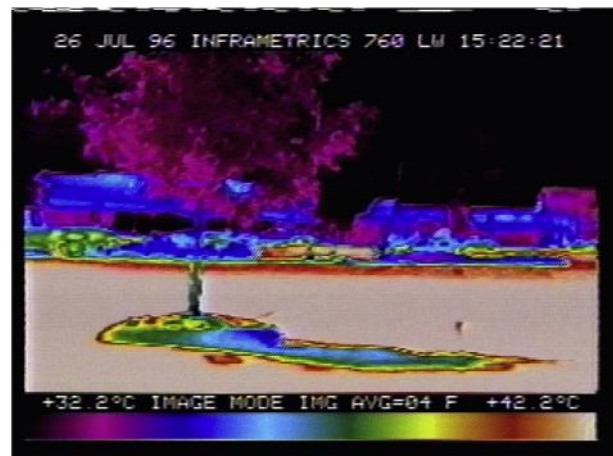


Figure 8 A summer thermal image of a tree growing in an asphalt parking lot in Athens, GA. The reduction in asphalt temperature caused by shading is approximately 10 °C. Luvall, NASA, Marshall Space Flight Center.

temperatures. The methods being promoted to mitigate the UHI are

- planting trees and vegetation,
- installing green roofs, and
- installing cool roofs.

Trees, vegetation, and green roofs can reduce the amount of urban surface thermal energy to mitigate the UHI, and can provide other important benefits such as reducing heating and cooling energy use and associated air pollution and greenhouse gas emissions, remove air pollutants, sequester and store carbon, help lower the risk of heat-related illnesses and deaths, improve storm-water control and water quality, reduce noise levels, create habitats, improve aesthetic qualities, and increase property values (Figure 8) (Rosenzweig et al., 2006). Cool



Figure 9 A 70 000 square foot rooftop meadow complete with trees constitutes the green roof of the Mormon Church Mormon Assembly Hall in Salt Lake City, UT. Luvall, NASA, Marshall Space Flight Center.

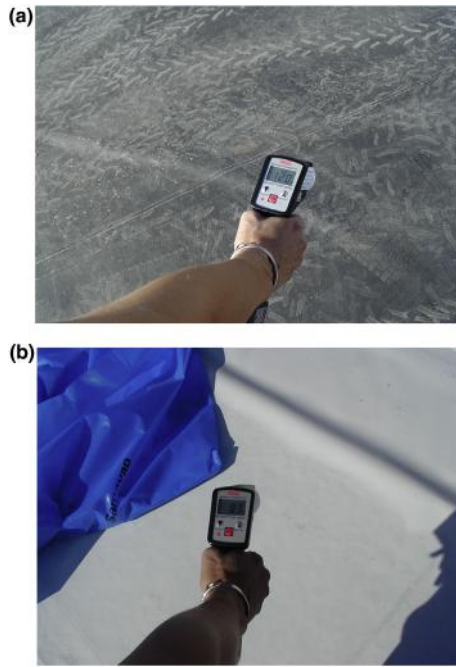


Figure 10 (a and b) The original black membrane roof in the Energy Solutions Arena, Salt Lake City, UT is 48 °C. The replacement white membrane roof is 28 °C. Luvall, NASA, Marshall Space Flight Center.

roofs, either by installing green (vegetated) roofs or by installing highly reflective roofing material, can reduce thermal energy emittance into the lower atmosphere, and lower cooling energy use, peak electricity demand, air pollution, greenhouse gas emissions, and heat-related incidents that stress the human body (Figure 9).

Rooftops in urban areas, particularly in the CBD where there are many buildings in close proximity to one another, can become extremely hot due to the type of roofing materials used. Black membrane roof temperatures on these roofs can be 75 °C or warmer in the summertime. Even during September and several hours after solar noon, these roofs can get very hot. The Energy Solutions Arena rooftop shown in Figure 10(a) had a temperature of 49 °C on 30 September, at 4.45 p.m. This same roof was in the process of being reroofed with high reflectance material (i.e., high albedo) that significantly lowered its temperature to 28 °C (Figure 10(b)).

Cool pavements are also being used to reduce thermal response from the urban surface. The term currently refers to paving materials that reflect more solar energy, enhance water evaporation, or have been otherwise modified to remain cooler than conventional pavements. Conventional paving materials can reach peak summertime temperatures of 48–67 °C, transferring excess heat to the air above them. Cool pavements can be created with existing paving technologies (such as asphalt and concrete) by increasing the porosity of these surfaces for increased water retention, as well as with newer approaches such as the use of coatings or grass paving.

References

- Akbari, H., Konopacki, S., Pomerantz, M., 1999. Cooling energy savings potential of reflective roofs for residential and commercial buildings in the United States. *Energy* 24 (5), 391–407.
- Akbari, H., Pomerantz, M., Taha, H., 2001. Cool surfaces and shade trees to reduce energy use and improve air quality in urban areas. *Solar Energy* 70 (3), 295–310.
- Akbari, H., Konopacki, S., 2005. Calculating energy saving potentials of heat-island reduction strategies. *Energy Policy* 33 (6), 721–756.
- Ashley, W.S., Bentley, M.L., Anthony Stallins, J., 2012. Urban induced thunderstorm modification in the southeast United States. *Climatic Change* 113 (2), 481–498.
- Berk, A., Anderson, G.P., Acharya, P.K., Chetwynd, J.H., Bernstein, L.S., Shettle, E.P., Matthew, M.W., Adler-Golden, S.M., 1999. MODTRAN4 Users Manual. Air Force Research Laboratory, Space Vehicles Directorate, Air Force Material Command, Hanscom AFB, MA 01731-3010.
- Bornstein, R., Lin, Q., 2000. Urban heat islands and summertime convective thunderstorms in Atlanta: three case studies. *Atmospheric Environment* 34, 507–516.
- Brown, D.G., Walker, R., Manson, S., Seto, K., 2004. Modeling land use and land cover changes. In: Gutman, G., Janetos, A.C., Justice, C.O., Moran, E.F., Mustard, J.F., Rindfuss, R.R., Skole, D., Turner II, B.L., Cochrane, M.A. (Eds.), *Land Change Science: Observing, Monitoring and Understanding Trajectories of Change of the Earth's Surface*. Kluwer Academic Publishers, p. 399.
- Comarazamy, D.E., Gonzalez, J.E., Luvall, J.C., Rickman, D.L., Bornstein, R.D., 2013a. Climate impacts of land cover and land use changes in tropical islands under conditions of global climate change. *Journal of Climate* 26, 1535–1550.
- Comarazamy, D.E., Gonzalez, J.E., Luvall, J.C., 2013b. Quantification and mitigation of long-term impacts of urbanization and climate change in the tropical coastal city of San Juan, Puerto Rico. *Int. J. Low-Carbon Technologies* (September 5). <http://dx.doi.org/10.1093/ijlct/ctt059>.
- Elsom, D.M., Terence Meaden, G., 1982. Suppression and dissipation of weak tornadoes in metropolitan areas: a case study of greater London. *Monthly Weather Review* 110 (7), 745–756.
- Estes Jr., M., Al-Hamdan, M., Crosson, W., Quattrochi, D.A., Johnson III, H., Hodgson, J., 2010. Validation and demonstration of the Prescott Spatial Growth Model in Metropolitan Atlanta, Georgia. *Urban Reg. Inf. Syst. J.* 22 (1), 5–21.
- Gaffan, D., Ross, R., 1998. Increased summertime heat stress in the US. *Nature* 396, 529–530.
- Gallo, K.P., Owen, T.W., 1999. Satellite-based adjustments for the urban heat island temperature bias. *Journal of Applied Meteorology* 38 (6), 806–813.
- Howard, L., 1833. *The Climate of London: Deduced from Meteorological Observations Made in the Metropolis and at Various Places around It*. Harvey and Darton, London.
- Hu, X.-M., Klein, P.M., Xue, M., Lundquist, J.K., Zhang, F., Qi, Y., 2013. Impact of low-level jets on the nocturnal urban heat island intensity in Oklahoma city. *Journal of Applied Meteorology and Climatology* 25 (8), August 7, 1779–1802. <http://dx.doi.org/10.1175/JAMC-D-12-0256.1>.
- Jacob, D.J., Winner, D.A., 2006. Effect of climate change on air quality. *Atmospheric Environment* 43 (1), 51–63.
- Karl, T.R., Knight, R.W., 1997. The 1995 Chicago heat wave: how likely is a recurrence? *Bulletin of the American Meteorological Society* 78 (6), 1107–1119.
- Loose, T., Bornstein, R.D., 1977. Observations of mesoscale effects on frontal movement through an urban area. *Monthly Weather Review* 105 (5), 563–571.
- Luvall, J.C., Holbo, H.R., 1989. Measurements of short-term thermal responses of coniferous forest canopies using thermal scanner data. *Remote Sensing of Environment* 27, 1–10.
- Olfe, D.B., Lee, R.L., 1971. Linearized calculations of urban heat island convection effects. *Journal of Atmospheric Sciences* 28 (8), 1374–1388.
- Orville, R.E., Huffines, G., Mielsen-Gammon, J., Zhang, R., Ely, B., Steiger, S., Phillips, S., Allen, S., Read, W., 2001. Enhancement of cloud-to-ground lightning over Houston Texas. *Geophysical Research Letters* 28 (13), 2597–2600.
- Quattrochi, D.A., Estes Jr., M.G., Crosson, W.L., Lapenta, W., Limaye, A., Khan, M., 2006. NASA Air Quality Applications Technical Report: The Application of Satellite-Derived High Resolution Land Use/Land Cover Data to Improve Urban Air Quality Model Forecasts. NASA Technical Publication. November 2006-214710.
- Quattrochi, D.A., Luvall, J.C., Rickman, D.L., Estes, M.E., Laymon, C.A., Howell, B.F., 2000. A decision support information system for urban landscape management using thermal infrared data. *PE&RS* 66 (10), 1195–1207.
- Rosenzweig, C., Solecki, W., Slosberg, R., 2006. *Mitigating New York City's Heat Island with Urban Forestry, Living Roofs, and Light Surfaces*. New York.
- Shepherd, J.M., Stallins, J.A., Jin, M., Mote, T.L., 2010. Urbanization: Impacts on clouds, precipitation, and lightning. In: Aitkenhead-Peterson, J., Volder, A. (Eds.), *Urban Ecosystem Ecology*. American Society of Agronomy, Crop Science Society of America, and Soil Science Society of America, Madison, WI, pp. 1–27.

- Stone, B., 2008. Urban and rural temperature trends in proximity to large U.S. cities: 1951–2000. *International Journal of Climatology* 27 (13), 1801–1807.
- Taha, H., 2007. Urban climates and heat islands: Albedo, evapotranspiration, and anthropogenic heat. *Energy and Buildings* 25 (2), 99–103.
- Vukovich, F.M., King, W.J., Dunn III, J.W., Worth, J.J.B., 1979. Observations and simulations of the diurnal variation of the urban heat island circulation and associated variations of the ozone distribution: a case study. *Journal of Applied Meteorology* 18 (7), 836–854.
- Yoshikado, H., Tsuchida, M., 1996. High levels of winter air pollution under the influence of the urban heat island along the shore of Tokyo Bay. *Journal of Applied Meteorology* 35 (10), 1804–1814.
- Zhang, D.L., Xuan, Y., Dickerson, R.R., 2009. Upstream urbanization exacerbates urban heat island effects. *Geophysical Research Letters* 36 (L24401), 1–5.
- Gallo, K.P., Owen, T.W., 1999. Satellite-based adjustments for the urban heat island temperature bias. *Journal of Applied Meteorology* 38 (6), 806–813.
- González, J.E., Luvall, J.C., Rickman, D.L., Comarazamy, D., Picon, A.J., 2007. Urban heat island identification and climatological analysis in a coastal, tropical city: San Juan, Puerto Rico, pp. 223–252. In: Weng, Q., Quattrochi, D.A. (Eds.), *Urban Remote Sensing*. CRC Press, Boca Raton, London, New York, 412 pp.
- Katsoulis, B.D., Theoharatos, G.A., 1985. Indications of the urban heat island in Athens, Greece. *Journal of Applied Meteorology* 24 (12), 1296–1302.
- Morris, C.J.G., Simmonds, I., Plummer, N., 2001. Quantification of the influences of wind and cloud on the nocturnal urban heat island of a large city. *Journal of Applied Meteorology* 40 (2), 169–182.
- METROMEX Update, 1976. *Bulletin of the American Meteorological Society* 57 (3), 304–308.
- Planning and Urban Design Standards, 2006. American Planning Association. John Wiley and Sons, Hoboken, NJ, pp. 101–106, ISBN - 13: 978-0-471-47581-1.
- Preston-Whyte, R.A., 1970. A spatial model of an urban heat island. *Journal of Applied Meteorology* 9 (4), 571–573.
- Price, J.C., 1979. Assessment of the urban heat island effect through the use of satellite data. *Monthly Weather Review* 107 (11), 1554–1557.
- Project METROMEX, 1974. *Bulletin of the American Meteorological Society* 55 (2), 86–121.
- Quattrochi, D.A., Estes Jr., M.G., Laymon, C.A., Crosson, W.L., Howell, B.F., Luvall, J.C., Rickman, D.L., 2007. Urban heat islands. In: King, M.D., Parkinson, C.L., Partington, K.C., Williams, R.G. (Eds.), *Our Changing Planet*. Cambridge University Press, pp. 298–301. ISBN 978-0-521-82870-3.
- Renee, G., Quattrochi, D.A., Luvall, J.C., 2006. A multi-scale approach to urban thermal analysis. *Remote Sensing of Environment* 104, 123–132.
- Vukovich, F.M., King, W.J., 1980. A theoretical study of the St. Louis heat island: comparisons between observed data and simulation. Results on the urban heat island circulation. *Journal of Applied Meteorology* 19 (7), 761–770.
- Bornstein, R.D., Robock, A.D., 1976. Effects of variable and unequal time steps for advective and diffusive processes in simulations of the urban boundary layer. *Monthly Weather Review* 104 (3), 260–267.
- Changnon Jr., S.A., 1975. Operations of mesoscale networks, illustrated by MET-ROME. *Bulletin of the American Meteorological Society* 56 (9), 971–978.
- Comarazamy, D.E., González, J.E., Luvall, J.C., Rickman, D.L., Mulero, P.J., 2010. A land–atmospheric interaction study in the coastal tropical city of San Juan, Puerto Rico. *Earth Interactions* 14, 1–24.
- Dominguez, A., Kleissl, J., Luvall, J.C., Rickman, D.L., 2011. High-resolution urban thermal sharpener (HUTS). *Remote Sensing of Environment* 115 (7), 1772–1780.
- EPA, 2008. *Reducing Urban Heat Islands: Compendium of Strategies*. <http://www.epa.gov/hiri/resources/compendium.htm>.
- Estes Jr., M.G., Quattrochi, D.A., Stasiak, E., 2003. The Urban Heat Island Phenomenon: How Its Effects Can Influence Environmental Decision Making in Your Community. *Public Management* 85 (3), 8–12.

Diurnal Cycle

A Betts, Atmospheric Research, Pittsford, VT, USA

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

The diurnal cycle over land is driven by solar heating in the daytime, and longwave cooling at night. In summer, the maximum temperature decreases with increasing cloud cover, because clouds reflect sunlight. In winter, the minimum temperature falls steeply under clear skies, because clouds reduce the longwave cooling to space. Over moist soils, increased evaporation reduces the diurnal temperature and humidity ranges. A few hours after sunrise, there is a transition when the nighttime stable layer is eroded by surface heating. Carbon dioxide shows a dawn maximum as nighttime respiration is trapped near the surface, and an afternoon minimum.

Introduction

Near the Earth's surface, many variables have a characteristic *diurnal or daily* cycle, driven by the diurnal cycle of the incoming solar radiation, which is zero at night and peaks at local noon. The atmosphere is relatively transparent to the short-wave radiation from the sun and relatively opaque to the thermal radiation from the Earth. As a result, the surface is warmed by a positive net radiation balance in the daytime in summer, and cooled by a negative radiation balance at night. The surface temperature oscillates almost sinusoidally between a minimum at sunrise and a maximum in the afternoon. This is referred to as the diurnal cycle of temperature. In warm seasons, the daily net radiation balance is positive, and the daily mean temperature is determined by the daily mean surface energy balance, which involves not only the short- and long-wave radiation components, but also heat transfers to the atmosphere.

The magnitude of this diurnal range of temperature is determined by many factors, which we shall discuss. Clouds have a large impact on the surface radiation balance and this differs between warm and cold seasons. The nature of the underlying surface is important, whether land or water, and so is the coupling to the atmosphere above. The phase change of water, particularly evaporation and condensation plays an important role in moderating the diurnal range of temperature, because of the large latent heat of vaporization. (In cold climates, the freezing and thawing of the soil are also important on the seasonal timescale.)

Over the ocean (and large lakes), the diurnal temperature range is small, because the incoming solar energy is mixed downward into the ocean 'mixed layer,' which is usually tens of meters deep. One day of solar heating will warm a layer of water 50 m deep to less than 0.1 K, because of its large thermal capacity. Only in light winds, when the downward mixing is small, does the diurnal range of sea surface temperature reach 1 K. On timescales longer than the diurnal, evaporation of water primarily balances the surface net radiation budget.

Over land, only a small fraction (<20%) of the net radiation at the surface is conducted downward in the daytime, or stored by warming trees on the surface, for example. As a result, the surface temperature rises rapidly after sunrise, until near balance is achieved between the net radiation and the direct

transport of heat to the atmosphere (referred to as the *sensible heat flux*) and evaporation of water (or transpiration from plants), referred to as the *latent heat flux*. If the surface is a desert, then the daytime temperature rise is large, but if water is readily available for transpiration, the daytime rise of temperature is greatly reduced, because most of the net radiation goes into the latent heat of vaporization. The surface sensible and latent heat fluxes have a large diurnal cycle, with a peak near local noon, as they are driven primarily by the incoming solar radiation. The surface temperature peaks a little later in the afternoon, when the surface sensible heat flux goes negative as the surface cools.

Coupling of the Summer and Winter Diurnal Cycle to Clouds

The diurnal cycle over land is driven by the surface net radiation balance, which we may write

$$R_n = SW_n + LW_n \quad [1]$$

where the net radiation, R_n , is the sum of the net shortwave radiation, SW_n a heating term, and the net longwave radiation, LW_n a cooling term. The net shortwave flux is reduced by reflection at the surface, which we call the surface albedo, α_s ; and by the reflection and absorption by the cloud fields above, which we call the effective cloud albedo (ECA). So we can write the reduction of the downward clear-sky flux, $SW_{dn}(\text{clear})$ in the form

$$SW_n = (1 - \alpha_s)(1 - \text{ECA})SW_{dn}(\text{clear}) \quad [2]$$

Grassland or crops have α_s in the range 0.15–0.2, but this rises to 0.6–0.8 with snow cover. The boreal forest has a lower α_s of order 0.1, and this rises to about 0.4–0.5 with snow on the canopy, and falls back to 0.2 once the canopy sheds the snow. This is because snow on the ground is largely in the shade of the trees. The ECA has an even wider range: zero with a clear sky by definition, and >0.9 in overcast conditions with heavy rain. So cloud cover is an important modulator of the diurnal cycle. We will illustrate this using Canadian Prairie data, which have long-term records of opaque or reflective cloud cover that have been calibrated to the ECA.

Figure 1(a) shows that in summer (June–July–August) the amplitude of the diurnal cycle increases as daily mean opaque cloud cover falls from 95 to 5%. In summer, near-surface temperature is minimum at sunrise, but this falls only slightly when cloud cover is small. On the other hand, the afternoon temperature maximum increases steeply as cloud cover decreases, and SW_n increases. Cooled by the net outgoing longwave radiation, LW_n , the surface temperature falls at night from the afternoon maximum back to the sunrise minimum.

Figure 1(b) shows an inverted pattern for winter (December–January–February): it is coolest with near-clear-sky conditions. This shows that there are fundamental differences between the surface radiation balance with clouds between summer and winter. The sun is low in the sky in winter, and the Prairies have a high surface albedo with snow, so SW_n in eqns [1] and [2] becomes small. However, clouds also blanket the Earth and greatly reduce LW_n , the surface cooling to space, and this longwave effect of clouds becomes dominant in winter. We see that the diurnal range of temperature is very small with 95% cloud cover. As cloud cover decreases, the surface cooling increases and the temperature drops. The minimum temperature at sunrise falls the most, because in the daytime the LW_n cooling is partly offset by SW_n heating.

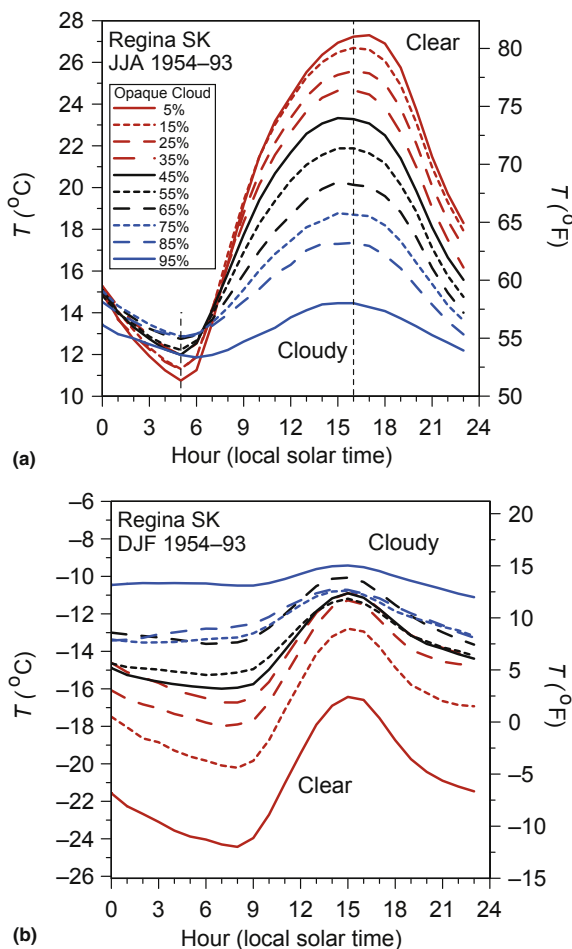


Figure 1 Dependence of diurnal cycle of temperature on opaque cloud cover in (a) summer and (b) winter for Regina, Saskatchewan, Canada.

Dependence of Diurnal Cycle on Evaporation

Figure 2 illustrates this diurnal variation using data from sunny days in midsummer during a 1987 field experiment (with acronym FIFE) conducted over grassland near Manhattan, Kansas. The panels on the left (from top to bottom) show net radiation, R_n , sensible heat flux, H , and latent heat flux, LE . The surface energy balance can be written as

$$R_n = H + LE + G \quad [3]$$

where G is the storage in the ground and vegetation, which we do not show. In addition a small amount of energy goes into photosynthesis, which again we do not show. The time axis is local solar time, which is UTC-6 h. The data have been grouped and averaged based on the percent soil moisture (SM) in the first 10 cm of soil, so that there are three curves (each an average of about 10 days) representing dry, medium, and wet soils. The upper left panel shows that the mean net radiation on these sunny days is very similar. However because SM is a major control on evaporation, the partition of the net radiation into sensible and latent heat is very different. When the soil is wet, the latent heat flux (or 'evaporative energy' flux) is about three times the sensible heat flux, whereas when the soil is dry, these two fluxes are nearly equal. The panels on the right side show the response to the different surface forcing. The upper right panel shows the surface temperature (measured by an infrared radiation thermometer, mounted on a tower and pointed downward at the grass). Although R_n is the same, on days when the soil is dry and water is not readily available for evaporation, the surface gets very hot, as warm as 44 °C near noon. This warm surface temperature drives the large sensible heat flux H and heats the air above the surface. The diurnal range of the surface temperature is more than 20 °C on these days, while for the air at 2 m above the surface in the middle panel, the diurnal range is only 12 °C. As SM increases, the daily maximum surface and air temperature decrease. The upper two panels on the right are similar, except that the amplitude of the surface temperature is larger than that of the air temperature. Both are related to the sensible heat flux H . Note that the air temperature has a broad afternoon maximum, because H is upward as long as the surface is warmer than the air. The surface temperature falls below the air temperature only in late afternoon, H then changes sign, and at night the surface is cooler than the air. The lower right panel shows the diurnal cycle of relative humidity (RH) as a percentage. Over the wetter soils, the RH of the air at 2 m reaches 85% before sunrise, and falls in the daytime as the surface and air warm. The fall of RH is smallest on the days with the greatest evaporation, LE . When evaporation is reduced because the soil is dry, daytime RH falls as low as 30% and even at night, only reaches 72% at sunrise.

Coupling between the Surface Diurnal Cycle and the Atmospheric Mixed Layer

As the land surface is heated during the daytime, a dry convective boundary layer grows in depth. This is called the

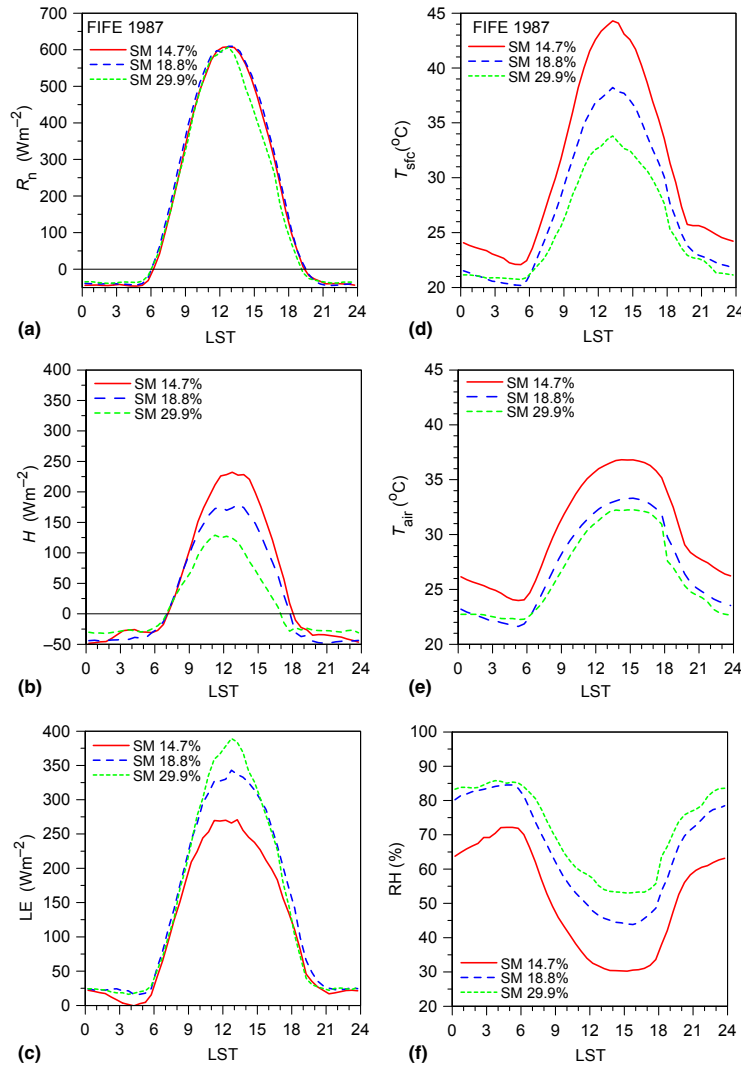


Figure 2 Diurnal cycle of net radiation (R_n), sensible heat flux (H), and latent heat flux (LE) (a–c); surface temperature, air temperature, and RH (d–f), stratified by percent volumetric SM in first 10 cm of the soil.

‘mixed layer,’ because the turbulent dry convection rapidly stirs the layer to one of near-neutral buoyancy and near-constant water vapor mixing ratio. The diurnal cycle of the surface and the mixed layer are tightly coupled. As a result the pre-existing atmospheric structure above the surface at sunrise has a considerable impact on the daytime diurnal cycle, as illustrated in the following figures using surface and sounding data collected over the boreal forest in Saskatchewan, Canada during the Boreal Ecosystem-Atmosphere Study (BOREAS) in 1994. **Figure 3** shows the surface diurnal cycle for 2 days in spring. The upper panel shows for each day the temperature at two levels, an upper level T_U , which is at 21 m, about 5 m above the canopy of a jack pine forest, and a lower level about 5 m above the forest floor. On both days the surface cools strongly at night and rises steeply after sunrise with a greater diurnal range than the near-clear-sky case in **Figure 1**. The diurnal range under the canopy is larger than above it. At night on 26 May, the winds are lighter, and the near-surface nighttime boundary

layer is more stable (see **Figure 4**). The air under the canopy becomes effectively decoupled from the atmosphere above and the stable temperature gradient across the canopy at night reaches 7 K. There is very little evaporation from either the forest, or the cold lakes at this time in spring. The lower panel shows RH measurements above the canopy. In the late afternoon, RH falls as low as 20% on 31 May. Before sunrise on this day, RH above the canopy reaches 90% as T_U falls to a minimum of 4 °C. RH was not measured below the canopy, but the temperatures there are cold enough to saturate the air in the hours before sunrise. The dew point is often used to estimate minimum nighttime temperatures at the surface.

The right-hand scale of the upper panel shows the corresponding dry potential temperature, which is defined as

$$\theta = (T + 273.15)(1000/p)^{.286} \quad [4]$$

where p is the surface pressure (here about 950 hPa, since the observation site is about 500 m above sea level). The potential

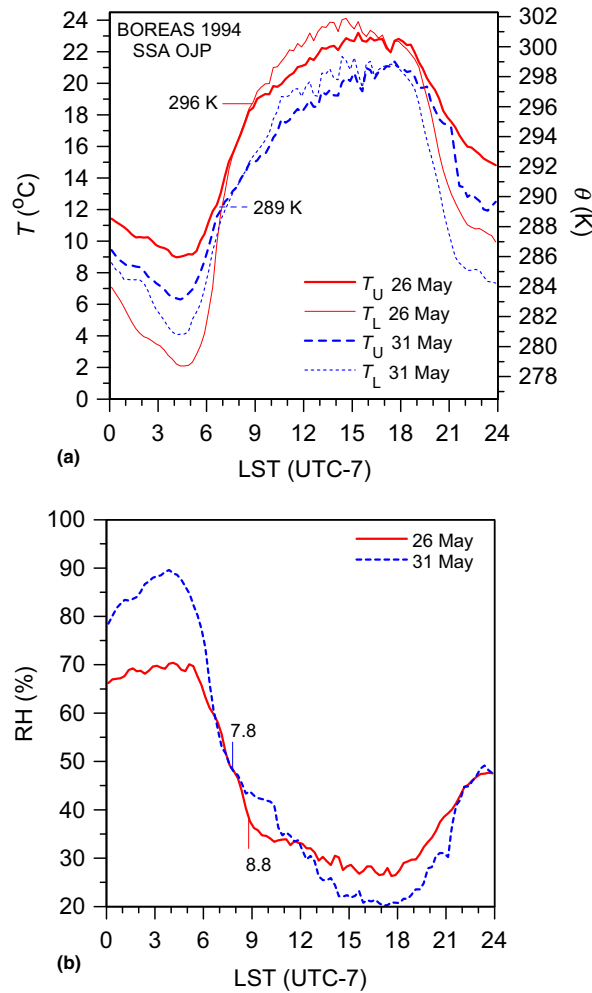


Figure 3 Diurnal cycle of temperature, above and below a boreal forest canopy (a), and of RH above the canopy (b) for 2 days in May 1994.

temperature, θ , is useful as a variable because it allows us to compare the surface and atmosphere above. During the daytime the boundary layer above the surface is mixed to almost constant potential temperature (see Figure 4). The strong radiative cooling of the surface at night generates a stable layer close to the ground, typically only a few hundred meters deep. About 3–4 h after sunrise, the surface has warmed enough to remove this stable surface layer and reconnect to a deeper layer. When this happens, the rate of rise of temperature and the rate of fall of RH decrease sharply. In Figure 3, this occurs on 26 May at a local time of 8.8 h, when θ reaches 296 K; while on 31 May, it occurs at 7.8 h, when $\theta = 289$ K, and on this day the change is smaller.

Figure 4 shows sequences of seven profiles of potential temperature in the lower troposphere, measured by rawinsonde ascents, nominally every 2 h from sunrise to late afternoon on the 2 days. The upper panel shows at sunrise (417 LST, solid) a cold (stable) surface layer only about 25 hPa deep (200 m), with a deep layer above of constant θ , which is the residual or ‘fossil’ mixed layer from the previous

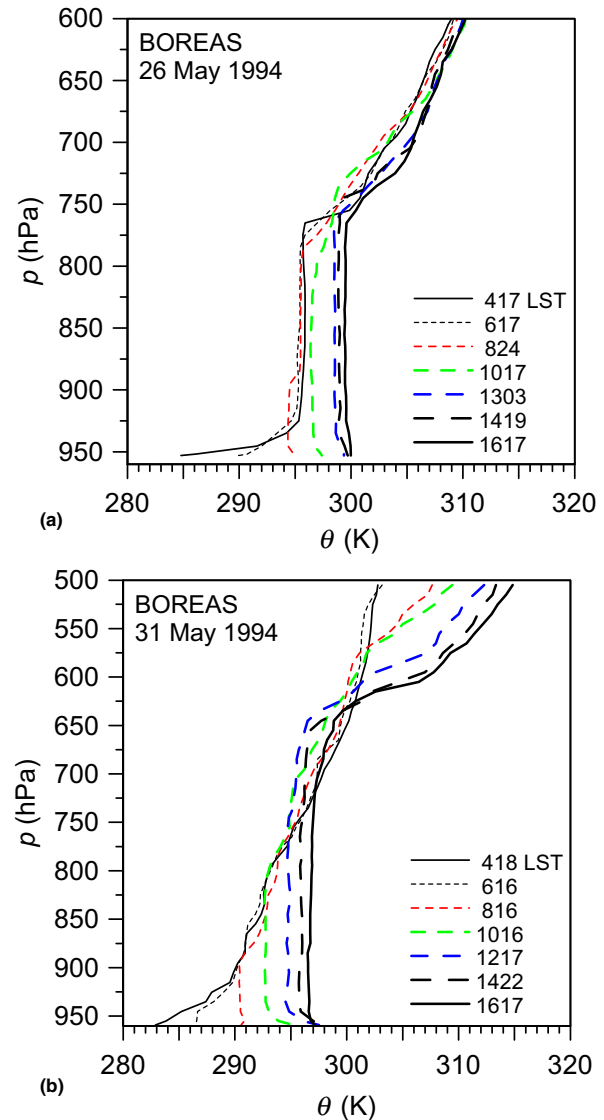


Figure 4 Profiles of potential temperature in the lower troposphere on (a) 26 May and (b) 31 May.

day. At the surface the temperature warms rapidly, as the surface sensible heat flux is trapped in this shallow surface layer. The profile at 0824 LST shows a mixed layer with $\theta = 294.5$ K to 890 hPa. Shortly afterward, when the surface potential temperature reaches $\theta = 296$ K, the new growing boundary layer merges with the deep residual mixed layer. From then on, the surface and mixed layer warm much more slowly, as seen in Figure 3. Even though H exceeds 300 W m^{-2} at all the forest sites for several hours around local noon (not shown), this large heat flux is distributed through a deep layer.

The lower panel shows the time sequence on 31 May. Note that at sunrise (solid), the profile is quite different than on 26 May. Instead of a deep layer of constant θ , produced by dry convection the previous day (a so-called dry adiabatic structure), there is a layer from 920 to 650 hPa in which

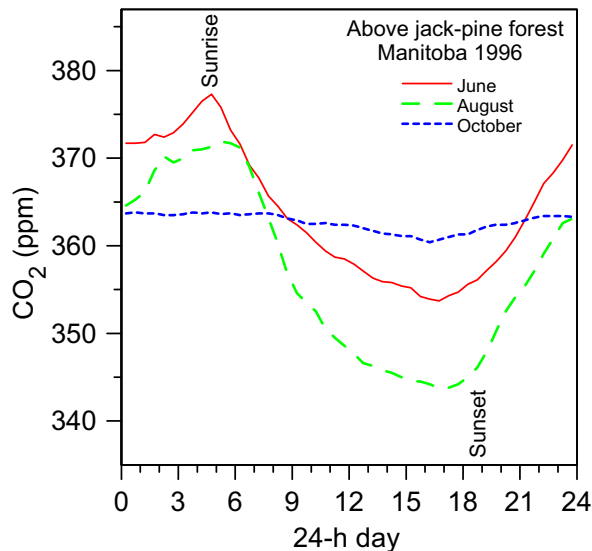


Figure 5 Monthly mean diurnal cycle of CO_2 for June, August, and October for a boreal jack pine site. Data from McCaughey, J.H., Dr., Queens University, Kingston, Ontario.

θ increases steadily with height. In fact, this layer was produced by showers the previous evening (and it has a so-called wet adiabatic structure). The change in slope of the early morning profile at 920 hPa is at $\theta = 289$ K, and hence we see on [Figure 3](#), a change in the rate of warming, once the surface reaches this potential temperature. This change of slope is more dramatic on 26 May, because the change in the vertical profile is also greater. On 31 May, the mixed layer grows steadily all day until it is 300 hPa deep (about 3000 m) in the late afternoon. On both these days, there is some broken cumulus cover in the afternoon at the top of the mixed layer. The rapid warming on 31 May, that is seen between 500 and 600 hPa, is related to the lowering and change in structure of a powerful jet stream above, not by surface processes.

Diurnal Cycle of CO_2

The diurnal cycle of the solar radiation drives a diurnal cycle in CO_2 through photosynthesis and respiration in plants. [Figure 5](#)

shows the mean diurnal cycle over a young jack pine canopy (about 5–6 m tall) near Thompson, Manitoba from the 1996 BOREAS experiment for 3 months, June, August, and October. During the summer months, CO_2 decreases during the daylight hours as it is taken up in photosynthesis, and increases at night as it is released by respiration from both plants and soil. The amplitude of the diurnal cycle increases from June to August, but the monthly mean decreases as there is a net CO_2 uptake by the entire Northern Hemisphere. By October of this year however the diurnal cycle is very small, as temperatures have dropped low enough that both photosynthesis and respiration have almost ceased.

See also: **Clouds and Fog:** Climatology. **Numerical Models:** Clouds. **Radiation Transfer in the Atmosphere:** Radiation, Solar.

Further Reading

- Arya, S.P., 1988. *Introduction to Micrometeorology*. Academic Press, New York.
- Betts, A.K., 2006. Radiative scaling of the nocturnal boundary layer and the diurnal temperature range. *Journal of Geophysical Research* 111, D07105. <http://dx.doi.org/10.1029/2005JD006560>.
- Betts, A.K., Ball, J.H., 1995. The FIFE diurnal cycle climate. *Journal of Geophysical Research* 100, 25679–25693.
- Betts, A.K., Desjardins, R., Worth, D., 2013. Cloud radiative forcing of the diurnal cycle climate of the Canadian Prairies. *Journal of Geophysical Research* 118, 8935–8953. <http://dx.doi.org/10.1002/jgrd.50593>.
- Budyko, M.I., 1974. *Climate and Life*. Academic Press, New York.
- Curry, J.A., Webster, P.J., 1999. *Thermodynamics of Atmospheres and Oceans*. Academic Press, New York.
- Garraff, J.R., 1992. *The Atmospheric Boundary Layer*. Cambridge Univ. Press, Cambridge.
- Geiger, R., 1965. *The Climate near the Ground*. Harvard Univ. Press, Cambridge.
- Hartman, D.L., 1994. *Global Physical Climatology*. Academic Press, New York.
- Simpson, J.E., 1994. *Sea Breeze and Local Wind*. Cambridge Univ. Press, Cambridge.
- Stull, R.B., 1988. *An Introduction to Boundary Layer Meteorology*. Kluwer, Dordrecht.

CHEMISTRY OF THE ATMOSPHERE

Contents

Chemical Kinetics

Ion Chemistry

Isotopes, Stable

Laboratory Kinetics

Methane

Observations for Chemistry (*In Situ*): Ozone Sondes

Observations for Chemistry (*In Situ*): Particles

Observations for Chemistry (*In Situ*): Water Vapor Sondes

Observations for Chemistry (Remote Sensing): IR/FIR (Satellite, Balloon and Ground)

Observations for Chemistry (Remote Sensing): Lidar

Observations for Chemistry (Remote Sensing): Microwave

Principles of Chemical Change

Radioactivity: Cosmogenic Radionuclides

Volcanoes: Composition of Emissions

Tracers

Chemical Kinetics

RP Wayne, University of Oxford, Oxford, UK

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

The basic concepts of order, molecularity, rate coefficient, and the temperature dependence of rate are presented first, before turning to theoretical interpretations of rates of bimolecular reactions in the gas phase such as collision theory and transition-state theory. The modifications needed when long-range forces operate, as in ion–molecule reactions, are indicated. Unimolecular and termolecular reactions may occur via discrete steps, which can often be treated kinetically using the stationary-state hypothesis. The dependence of rate on pressure (and temperature) may be complex. Consideration is given in the last section to the kinetics of processes occurring on and within water droplets and other atmospheric aerosol particles.

Reaction Kinetics and Atmospheric Chemistry

Chemistry in the atmosphere often consists of several consecutive and parallel steps that compete with each other. Interpretation of the rates of chemical change, and the concentrations and lifetimes of atmospheric constituents, requires a knowledge of the rates of the *elementary* reaction steps that make up the complex scheme. The *kinetic data* embodied in this knowledge are best obtained by laboratory experiment, although theory may have to be used if the experiments cannot be performed. Theory is useful in another way as well, because it can provide a rational basis for the extrapolation of laboratory data to temperatures, pressures,

and concentrations that exist in the atmosphere, but that cannot be used in the laboratory studies.

Rate Laws

Consider the hypothetical chemical reaction



Experimentally, the rate is found to be proportional to the concentrations of A, B, ... raised to some power

$$\text{Rate} = -\frac{d[A]}{dt} = -\frac{d[B]}{dt} = k[A]^\alpha[B]^\beta \quad [1]$$

Order, Molecularity, and Rate Constant

The powers α and β are the *order* of reaction with respect to reactants A and B, and $\alpha + \beta$ is the overall order; the constant of proportionality, k , is the *rate coefficient* (rate 'constant'). The *molecularity* of a reaction is the number of reactant molecules written in the stoichiometric equation. Order is thus an experimental quantity, molecularity an arbitrary theoretical one. An elementary reaction step is conceived as one that cannot be split into any chemically simpler processes. For truly elementary steps, order and molecularity are in general identical. Thus, if [reaction \[1\]](#) is elementary, and the only reactants are A and B, it is both bimolecular and overall second-order: first-order in each of the components A and B. However, a special case often arises in atmospheric chemistry. If the second reactant, B, is in great excess over A, then its concentration is effectively constant throughout the reaction. We can then combine the concentration with the rate coefficient, k , and write the rate of reaction as $k'[A]$, where $k' = k[B]$. Such a process is termed a *pseudo-first order* reaction, and k' is the *pseudo-first order rate coefficient*.

The Arrhenius Equation

Many rate constants are found to follow a temperature law embodied in the *Arrhenius expression*

$$k = A \exp\left(-\frac{E_a}{RT}\right) \quad [2]$$

where E_a is the *activation energy* and A is the *preexponential factor*. The Arrhenius equation is entirely empirical, but several theories of kinetics yield expressions for the rate coefficient that are similar in form.

Theories of Elementary Gas-Phase Bimolecular Reaction Steps

Bimolecular processes are probably the most important class of reaction, and, as we shall see later, termolecular, and many very important unimolecular, reactions involve several bimolecular elementary steps. The obvious starting point in discussing the theories of reaction is thus with bimolecular reactions. Two simplifications are commonly adopted in discussions of these theories. The first is the *collision theory* (CT), and the second *transition-state theory* (TST).

Collision Theory

In simple CT, reactant molecules are assumed to be hard spheres (radii r_A and r_{BC} in our example), and reaction is taken to be possible only if two conditions are met: a collision must occur, and the energy of collision along the line of centres must equal or exceed the energy required, E_c , to reach a critical configuration ($ABC^\ddagger$ in Figure 1 in *Chemistry of the Atmosphere: Principles of Chemical Change*). The rate of reaction according to this theory is readily shown to be given by

$$-\frac{dn_A}{dt} = -\frac{dn_{BC}}{dt} = n_A n_{BC} \sigma_c \bar{c} \exp\left(-\frac{E_c}{RT}\right) \quad [3]$$

where σ_c is the *cross-sectional area* (or *collision cross-section*)

$$\sigma_c = \pi(r_A + r_{BC})^2 \quad [4]$$

and $\bar{c}$ is the mean relative speed of molecules for temperature T ,

$$\bar{c} = \left(\frac{8k_B T}{\pi \mu}\right)^{1/2}; \quad \mu = \frac{m_A m_{BC}}{m_A + m_{BC}} \quad [5]$$

The quantities n_A and n_{BC} are the *number densities* of A and BC (concentrations in molecular units such as molecule cm^{-3}), respectively. [Equation \[3\]](#) certainly has the correct concentration dependence for an elementary bimolecular reaction, so that the rate coefficient can be written as:

$$k = \sigma_c \bar{c} \exp\left(-\frac{E_c}{RT}\right) \quad [6]$$

It is clear that the Arrhenius equation ([eqn \[2\]](#)) and [eqn \[6\]](#) closely resemble each other, and E_c is commonly identified with E_a , so that the question may be asked if $\sigma_c \bar{c}$ is to be compared with A . However, it should not be forgotten that $\bar{c}$ is dependent on $T^{1/2}$ (cf [eqn \[5\]](#)), while A , in the simplest formulation, is not temperature dependent. A more telling difficulty concerns the absolute magnitudes of A and $\sigma_c \bar{c}$. For typical atmospheric reactants, with collision radii ~ 400 pm and relative molecular masses ~ 30 , $\sigma_c \bar{c}$ is $\sim 3 \times 10^{-10} \text{ cm}^3 \text{ molecule}^{-1} \text{ s}^{-1}$ at 300 K. The product $\sigma_c \bar{c}$ is called the *collision frequency factor*. Except for the very simplest of reactants, experimental A factors are usually less than, and often much less than, the collision frequency factor. An explanation for the lack of agreement is sought in terms of molecular complexity, with the existence of special geometric arrangements that are needed during the collision to bring reactive parts of the molecules together (*steric requirements*), and of special needs for the distribution of internal energy. That explanation takes us well away from the idea of hard-sphere reactants.

Transition-State Theory

The alternative simplification adopted in the interpretation of bimolecular reactions is that of the TST or *Activated Complex Theory* (ACT). The reactants and the critically configured ABC molecule are assumed to be in 'quasi-equilibrium.' Equilibrium constants can be expressed in statistical thermodynamic terms, and if the formulation is also valid for the quasi-equilibrium, where the system is at a (free) energy maximum rather than minimum, then concentrations of $ABC^\ddagger$ (the *transition state*) may be calculated. Rates of reaction can then be obtained from the rate at which $ABC^\ddagger$ passes to products (as a result of translational or vibrational motions along the reaction coordinate). The resultant rate coefficient, k , is given by

$$k = \frac{k_B T}{h} \frac{q''_{ABC^\ddagger}}{q'_A q'_{BC}} \exp\left(-\frac{E_c}{RT}\right) \quad [7]$$

Partition functions are written as q'_A , q'_{BC} , $q''_{ABC^\ddagger}$ for reactants and transition state, the primes showing that the quantities are volume independent. The double prime on $q''_{ABC^\ddagger}$ indicates that the motion along the reaction coordinate has been factorized out (and a numerical constant introduced). In TST, then, the internal motions neglected in CT are expressly taken into

account through the use of the partition functions. TST concentrates only on that region of the potential energy surface around the transition state (for the calculation of the partition function $q''_{ABC^\ddagger}$) while CT is interested only in the height of the energy barrier at the transition state. It is the calculation of $q''_{ABC^\ddagger}$ that offers most difficulty in the practical implementation of TST. Spectroscopic parameters for the reactant molecules are usually available, so that q'_A and q'_{BC} are readily estimated. However, a knowledge of the shape of, and the forces acting at, the transition state would imply that the potential energy surface is itself known, at least in the region of $ABC^\ddagger$. The usual practice is to make an 'informed guess' at the magnitude of $q''_{ABC^\ddagger}$ based on a hypothetical interaction mechanism and a corresponding model for the transition state. Considerable differences in predicted preexponential factors are obtained from models of the transition state that are, for example, linear, bent, or cyclic. In a more limited way, TST can suggest a sensible order of magnitude for the preexponential factor. The three total partition functions in eqn [7] are each the product of translational, rotational, and vibrational partition functions. The translational parts can all be calculated, and orders of magnitude for rotational and vibrational parts employed in accordance with the number of each of these modes that exist in A, BC, and $ABC^\ddagger$. The temperature dependence for every partition function can be evaluated as a power law, so that eqn [7] can be rewritten as

$$k = A'T^n \exp\left(-\frac{E_c}{RT}\right) \quad [8]$$

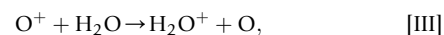
where A' is the temperature-independent part of the preexponential factor, and n some exponent chosen from the nature of the reactants (monatomic, diatomic, etc.) and a model of the transition state. For the hard sphere (CT) case, $n = 0.5$, from eqns [2] and [4]. In the more general case, n can be positive or negative. The most sensible procedure in temperature extrapolation thus seems to be first to predict n from a model of the reaction, and then to fit the experimental data to eqn [8] with that value of n .

Activation Energies and Long-Range Forces

Rates of reaction are, in part, controlled by the energy of a critical (transition-state) configuration, an energy that has as its counterpart the activation energy of experimental kinetics. The energy barrier arises because the reactant molecules are forced close together (closer than the sum of their radii in the hard-sphere collision approximation), and reactant bonds have to be broken while product bonds are made. The energy required is less than that required first to break reactant bonds and then to form product molecules in separate steps. The energy does not decrease at any stage in this picture as the system passes from separated reactants to the transition state. Such a decrease in energy would correspond to long-range attractive forces, and might lead to an increased collision frequency, and to an A factor that exceeded $\sigma_c \bar{c}$. Many examples of this type of behaviour are in fact known, even with neutral reactants, but the effects are strongest and most common with charged reactants. In *ion-molecule* reactions, such as



the ion can induce a dipole in the neutral reactant, and the resultant attractive force can both balance the ordinary chemical activation barrier as well as make the real encounter rate greater than the gas-kinetic collision frequency factor for neutral molecules. Near-zero activation energies are thus often found in this type of reaction, and the preexponential factors (typically $\sim 10^{-9} \text{ cm}^3 \text{ molecule}^{-1} \text{ s}^{-1}$) are several times larger than the values for neutral reactants. Because the long-range attractive forces dominate the potential energy, high velocities of approach are counterproductive in promoting reaction, and some *negative temperature coefficient* of rate constant may be observed. The stronger (or longer range) the interaction, the larger the rate coefficient. For ion reactions with neutral molecules possessing permanent (rather than induced) dipoles, preexponential factors are increased by another two or three times. Thus, charge transfer from O^+ to the dipolar molecule H_2O ,



has a rate coefficient of $2.3 \times 10^{-9} \text{ cm}^3 \text{ molecule}^{-1} \text{ s}^{-1}$ at 298 K, and the activation energy is essentially zero. The long-range interactions are yet larger, of course, for two reactants both of which are charged. Positive ion-negative ion, or positive ion-electron reactions are characterized by rate coefficients three to four orders of magnitude larger than typical gas-kinetic collision frequency factors. For example, the rate coefficient (298 K) for neutralization of NO^+ by an electron



is $4.5 \times 10^{-7} \text{ cm}^3 \text{ molecule}^{-1} \text{ s}^{-1}$.

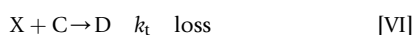
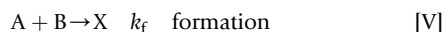
Multistep Reactions and the Stationary-State Hypothesis

Atmospheric chemistry consists of complex interactions of elementary reactions. Some of the processes to be described in the section that follows on unimolecular and termolecular reactions also involve several steps, which is why the subject of multistep reactions is introduced at this stage.

Consecutive and *parallel* steps involve reactive *intermediates* in competitive processes. Reaction intermediates of particular interest include atoms, radicals, ions, and excited species. Most of these intermediates are highly reactive, and, with one or two exceptions, cannot be 'stored' in a laboratory for long periods because they are lost on the walls of the containing vessel, or react with each other. Such intermediates are not necessarily *unstable*, and chemical lifetimes of isolated atoms or radicals in the absence of surfaces can be virtually infinite. Many excited-state species are unstable, since they may possess enough internal energy to fragment, and they may also be able to lose their energy by emission of radiation. An excited species that cannot undergo loss by an allowed radiative transition is said to be *metastable*.

Multistep reaction schemes are interpreted kinetically by writing down the differential equations, such as eqn [1], for all the species of interest, including the intermediates. Solution of these equations then allows prediction of the concentration-time variation of each of the species. Unfortunately, analytical solution of the many simultaneous differential equations is rarely possible. Numerical solution has become a widely

used alternative since the advent of high-speed computers and the development of good techniques for dealing with differential equations. For some highly reactive intermediates, the *Stationary-State Hypothesis* (SSH) (often alternatively called the *Steady-State Hypothesis*) provides a simplification that will permit algebraic solution of the kinetic equations. Consider an intermediate X that is created in a process whose rate is constant, and whose loss rate increases with increased [X]. After the reaction is started, [X] will increase until the rate of loss is equal to the rate of formation. A steady state for [X] has been reached, and $d[X]/dt \rightarrow 0$. To illustrate the stationary-state method, consider the pair of reactions



For simplicity, let [B] and [C] be in great excess, so that we may write pseudo-first-order rate coefficients $k'_f = k_f[B]$ and $k'_t = k_t[C]$. The kinetic equation that describes reactions [V] and [VI] is

$$\frac{d[X]}{dt} = k'_f[A] - k'_t[X] \quad [9]$$

If X is in a stationary state, then we set the differential equal to zero, and

$$[X]_{ss} = \frac{k'_f[A]}{k'_t} \quad [10]$$

where $[X]_{ss}$ indicates a steady-state concentration of X. The problem is to know if the concentration of X calculated using the SSH bears any relationship to actual concentrations. Our two-reaction example has been chosen because it can also be solved analytically. So long as [A] and [C] are independent of time, eqn [9] can be integrated to yield

$$[X] = \frac{k'_f[A]}{k'_t} [1 - \exp(-k'_t t)] \quad [11]$$

where t is the time for which the system has been reacting. This expression for [X] approaches the steady-state expression so long as $k'_t t \gg 1$, the error in applying the SSH being less than 1% for $k'_t t > 4.6$. The SSH can thus be applied so long as [A] and [C] remain constant over a period long enough for this inequality to be reached. It is evident that the circumstances under which the SSH is most likely to be valid are thus those where $k'_t t$ is large: that is, if the species X is highly reactive. One example of a species at steady state in the atmosphere is the highly reactive electronically excited state of atomic oxygen, $O(^1D)$, throughout the troposphere, stratosphere, and probably the mesosphere. Ground-state atomic oxygen, $O(^3P)$, however, cannot generally be treated in the atmosphere by steady-state methods because of its relatively small reactivity.

Theories of Unimolecular and Termolecular Reactions

We are now in a position to consider thermal unimolecular reactions, and their close counterpart, termolecular reactions. If chemical reaction requires collision between, or at least close proximity of, the reactants, then it might seem that all thermal processes ought to be kinetically of second order.

Unimolecular, first-order, elementary processes appear to lack the necessary approach of reactants, and termolecular, third-order, steps suffer from the impossibility of a simultaneous collision between three hard-sphere reactants. The explanation for first- and third-order thermal kinetics shares common ground, and a simple introduction is provided here.

No obstacle exists to understanding how single-step unimolecular and first-order decomposition occurs in a molecule AB that *already* has more than enough energy in it to break one of its bonds. An obvious case is the photodissociation of polyatomic molecules, where optical dissociation or predissociation populates vibrational levels of AB sufficient to cause fragmentation (see **Chemistry of the Atmosphere: Principles of Chemical Change**). The rate of fragmentation may depend on the rate at which energy can accumulate in the bond to be broken, but the reaction will be kinetically of first order. *Chemical activation* offers another route to high vibrational excitation. For example, the reaction of HO_2 with NO can produce a highly excited $HO_2NO^\dagger$ molecule (the dagger representing vibrational excitation)

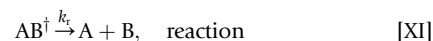
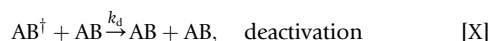


This excited $HO_2NO^\dagger$ can then either split up to the reactants again, or form OH and NO_2



Reaction of the excited $HO_2NO^\dagger$ is an unimolecular, first-order, elementary reaction.

It is in interpreting *thermal* unimolecular reactions that some difficulty arises, since the formation of an excited $AB^\dagger$ molecule involves collisions between the AB species, and might therefore be expected to show *second-order* kinetics. A basic understanding was provided by Lindemann, who suggested that thermal first-order reactions were not true elementary steps, but rather involved at least three elementary processes:



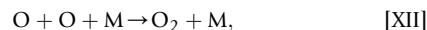
If reaction [X] dominates over [XI] as a loss process for $AB^\dagger$, then the concentration of $AB^\dagger$ is almost at its thermal equilibrium value, while the rate-determining step for reaction is the first-order process [XI]. Overall first-order kinetics follow. It is obvious, however, that at sufficiently low concentrations of AB, there becomes a point at which reaction [IX] is rate limiting, and the kinetic behaviour will be second order. Transition from first- to second-order behaviour is, indeed, seen at low enough pressures in this kind of thermal unimolecular reaction. Quantitative expression of these ideas can be obtained by a steady-state treatment for the concentration of $AB^\dagger$ as described in the previous section. The result for the rate of loss of AB is

$$-\frac{d[AB]}{dt} = k_1[AB] = \frac{k_a k_r [AB]^2}{k_d [AB] + k_r} \quad [12]$$

where k_1 is the experimentally defined pseudo-first order rate coefficient. So long as $k_d [AB] \gg k_r$, the reaction is first order, but if [AB] is reduced to the point at which the reverse inequality

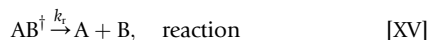
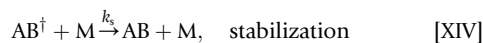
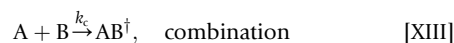
holds, then the reaction becomes second order. At high concentration, the limiting value of k_1 (referred to as k_1^∞) is equal to $(k_a k_t / k_d)$ and is thus truly first order, being independent of $[AB]$. The low-pressure limit, k_1^0 , is equal to $k_a [AB]$ and is itself first order in pressure, or second order overall.

Considerations about high- and low-pressure extrapolations of rate data are most frequently met in atmospheric chemistry in connection with termolecular reactions. As with unimolecular reactions, termolecular processes have orders variable with pressure, being third order at 'low' pressure and second order at 'high' pressure. Such reactions are extremely important in combination processes (sometimes erroneously called 'recombination' processes), and we can see why by first looking at the reaction of two atoms to form a diatomic molecule. A typical case is the combination of two $O(^3P)$ atoms. Curve 'X' in Figure 2 of Chemistry of the Atmosphere: Principles of Chemical Change illustrates this combination. Even if the combining atoms have no relative translational energy, the newly formed O_2 molecule has the $O + O$ combination energy stored in it: that energy is the $O-O$ bond energy, and the O_2 is chemically activated $O_2^\dagger$ at its dissociation limit. Unless some energy is removed within one vibrational period, the molecule will fall apart again as the internuclear distance increases on the first oscillation. Energy can be removed in collisions; the species that dissipates energy is often represented by the symbol M . In the atmosphere of Earth, M is usually the 'bath' mixture of N_2 and O_2 . The overall reaction is now written



which is a termolecular step. The redissociation that has been prevented is the unimolecular dissociation of $O_2^\dagger$ equivalent to step [XI], and the process deactivating $O_2^\dagger$ is the equivalent of [X]. We shall see shortly that internal energy seems to flow fairly freely between different vibrational modes of a polyatomic molecule. If the newly formed molecule is larger than diatomic, there are such modes into which the bond-combination energy can flow. The lifetime of the newly formed molecule can thus correspond to many vibrational periods before the energy flows back to the critical bond. With a large enough polyatomic molecule, the lifetime can be so great that collisional removal of excess energy (*stabilization*) is no longer rate determining, and combination then exhibits second-order kinetics. Because reactions [X] and [XI] are common to both unimolecular and termolecular reactions, the same general considerations about flow of energy apply to both types of process.

The analogue of expression [12] can be derived from the single excitation level kinetic scheme



The result is

$$-\frac{d[AB]}{dt} = k_{II}[A][B] = \frac{k_a k_s [A][B][M]}{k_s [M] + k_t} \quad [13]$$

where k_{II} is the experimentally defined pseudo-second order rate coefficient (analogous to k_1 in eqn [12]). We see straightaway that, if $k_t \gg k_s [M]$, the reaction is third order, with $k_{II}^0 = (k_a k_s / k_t) [M]$. If, however, $k_t \ll k_s [M]$, then the reaction is second order, with $k_{II}^\infty = k_a$. Increased complexity in the molecule AB reduces the value of k_t , because the combination energy is distributed among more vibrational modes. The concentration, or pressure, of third-body M at which third-order behaviour turns over to second-order kinetics is thus lower the more complex the molecule produced. 'Complex' is only a relative term here: combination of two hydrogen atoms to form H_2 is third order up to 10^4 atm, while combination of two CH_3 radicals to form C_2H_6 is second order at all but the lowest pressures. However, it so happens that the reactants in several combination reactions of great atmospheric importance, such as



are of just that molecular size that complex intermediate-order kinetics are displayed at some point in the atmospheric pressure range.

Expressions [12] and [13] represent the variations of experimentally determined rates of reaction with pressure. The pseudo-first- or second-order rate coefficients k_1 or k_{II} can be conveniently expressed in terms of the high- and low-pressure limiting values k_1^∞ , k_1^0 , or k_{II}^∞ , k_{II}^0 . For example, k_{II} in eqn [13] can be expressed as

$$k_{II} = \frac{k_{II}^0 k_{II}^\infty}{k_{II}^0 + k_{II}^\infty} \quad [14]$$

Remembering that k_{II}^0 is itself first order in pressure, it can be seen that eqn [14] represents in outline the variation of k_{II} with pressure that is found experimentally. Figure 1 shows

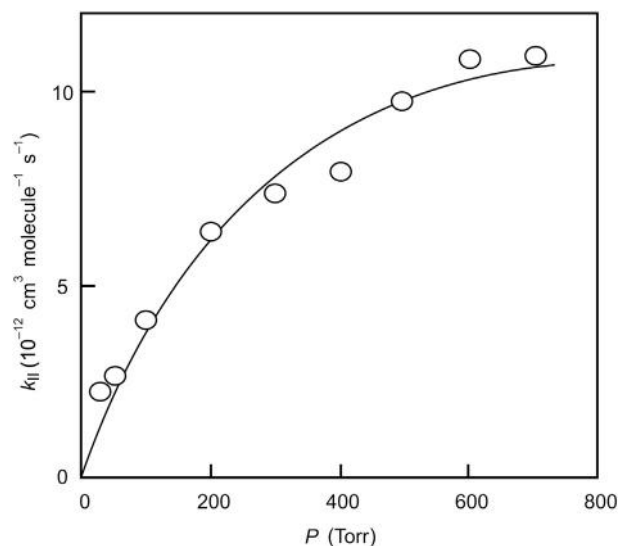


Figure 1 Rate coefficient for the reaction $OH + NO_2 + N_2 \rightarrow$ products as a function of pressure at $T = 298$ K. Experimental data measured by D'Ottone, L., Campuzano-Jost, P., Bauer, D., Hynes, A.J., 2001. A pulsed laser-photolysis-pulsed laser-induced fluorescence study of the kinetics of the gas-phase reaction of OH with NO_2 . J. Phys. Chem. A 105, 10538–10543.

experimental data for reaction [XVI] obtained over the pressure range 30–700 Torr with $M = N_2$. The ordinate is k_{II} as defined by the first two terms of eqn [13]. At the lowest pressures, k_{II} is seen to be a nearly (but not quite) linear function of pressure, as predicted by the right-hand term of the equation, with the slope corresponding to k_{II}^0 . At higher pressures, k_{II} becomes progressively less pressure-dependent, again as predicted by eqn [13] or eqn [14]. Unfortunately, however, eqn [14] does not match experimental data in detail, so that it cannot be applied directly to the calculation of rates at intermediate pressures. The reasons for the failure are known. The reactions and the rate coefficients k_a or k_c , k_d or k_s , and k_r should have been defined for each individual quantized vibrational level of $AB^\ddagger$, and the individual rates summed to give the total rate. It is, perhaps, easy to see that the more energy available (beyond the critical amount needed to break a particular bond), the more rapid will be the fragmentation (i.e., the larger will be k_r). Related to this point is the implication that energy stored in any vibrational mode can be made available to the critical bond. Experimental evidence largely favours the flow of energy between modes as being fairly free, and the distribution as being near statistical. An additional complication involves the interconversions of vibrations and rotations in the fragmenting molecule. The theory has been extended, modified, and manipulated over the years by Rice, Ramsperger, Kassel, and Marcus, and the familiar initials RRKM are used to designate their formulation. With sufficient sophistication of the input information, very good agreement can be obtained between theory and experiment. Correspondingly, one could have confidence in the extrapolation of data obtained in an intermediate concentration regime to either high-pressure (first-order) or low-pressure (second-order) limits. However, application of RRKM theory to real processes of atmospheric importance is in practice rather difficult, and an alternative, much simpler, approach is now almost universally adopted. This approach has its origins in work by Troe on the theoretical prediction of unimolecular reaction rate parameters. However, with k_{II}^∞ , k_{II}^0 known, Troe has shown that a simplification of his theory allows the right-hand side of eqn [14] to be multiplied by a broadening factor, F , that is a function of (k_{II}^0/k_{II}^∞) . For many atmospherically important termolecular reactions F may be calculated from a simple mathematical expression

$$F = F_{\text{cent}}^{\{1 + [\log_{10}(k_{II}^0/k_{II}^\infty)]^2\}^{-1}} \quad [15]$$

where F_{cent} is the broadening factor for the centre of the fall-off curve; typically, $F_{\text{cent}} = 0.6$.

Third-order reactions often show decreasing rate with increasing temperature: they have a *negative temperature coefficient*. The reason is that the larger the thermal kinetic energy possessed by the reactants A and B in process [XIII], the more internal vibrational energy will be stored in the $AB^\ddagger$ molecule produced. As pointed out earlier, the chance of the critical bond energy finding its way back to a breakable bond is thus increased, and k_r is larger. Since k_c and k_s are only slightly affected by temperature, it follows from eqn [13] that the rate of reaction will decrease with increasing temperature. Thermal energy in effect assists the newly formed molecule to split up again, thus slowing the rate of combination. In the third-order limit, k_{II}^0 is inversely proportional to k_r (see above). Theory

suggests that the temperature variations of k_r should be better expressed in terms of a power, T^n , rather than as a conventional activation energy. Hence, experimental measurements of k_{II}^0 as a function of temperature should be fitted against a T^{-n} law to allow rational interpolation or extrapolation to atmospheric temperatures. Typical measured values of n are 2.5–3.1 for reaction [XVI] and 1.7 for reaction [XVII]. Models of the transition state for bond-association reactions also suggest that, at the high-pressure limit, k_{II} should possess a negative exponent of temperature.

Condensed-Phase, Surface, and Heterogeneous Reactions

Reactions within the liquid droplets of clouds and fogs are important in several aspects of tropospheric chemistry, such as the oxidation of sulfur dioxide. Liquid or solid particles can play a critical role in the chemistry of the stratosphere under certain conditions. Particles such as those of sulfate aerosol or clouds formed from water-ice and hydrates of nitric acid (*polar stratospheric clouds*, PSCs) are implicated in such processes. This chemistry may involve surface reactions, or reactions within the bulk material, but the interface between gas and condensed phases is involved in some way, and the reactions are thus known as *heterogeneous reactions*. The types of heterogeneous process that are found in the atmosphere are considered in more detail in article **Chemistry of the Atmosphere: Principles of Chemical Change**.

The possible complexity of the kinetics can be illustrated by considering explicitly the steps involved in a heterogeneous reaction of a gas-phase species with either the bulk constituent of a liquid droplet or with another species that is already dissolved in it. The uptake of gas-phase molecules can be either reactive or nonreactive. Chemical change corresponds to loss of the gas-phase molecule; uptake that is nonreactive can arise from physical dissolution or from reversible chemistry. Figure 2 illustrates some of the most important individual steps that can be envisaged in a reaction where at least one of the reactants and one of the products is a gas-phase species, but where the chemistry takes place within a droplet. The forward steps shown in the diagram are (1) gas-phase transport of the reactant to the surface of the droplet; (2) accommodation at the surface; (3) diffusion into the liquid; (4) chemical reaction; (5) diffusion of both products and unreacted molecules to the surface; and (6) desorption of species from the interface. Characterization of each of these individual steps is obviously a formidable task, although one that may be simplified – as often happens in kinetics – by one of the steps being rate-determining.

In many respects, the kinetics of liquid, surface, and heterogeneous reactions are governed by the same principles that we have established for gas-phase processes. There are, however, some key differences. Reactions occurring *inside* particles are really confined to the liquid phase, since diffusion coefficients within solids are too small to allow significant reaction rates. On the other hand, reactions *on* solid surfaces are thought to be of very considerable atmospheric significance. The PSCs involved in stratospheric chemistry consist, for example, in part, of solid water-ice and solid nitric acid

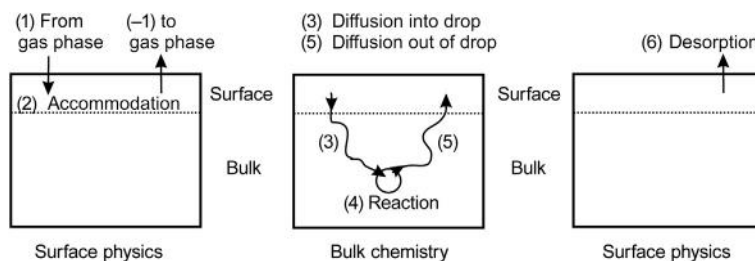


Figure 2 Processes involved in the uptake and chemical reaction of a gas-phase molecule by a liquid droplet, and chemical reaction within it. The individual steps are explained in the text. Diagram based on an idea of Ravishankara, A.R., 1997. Heterogeneous and multiphase chemistry in the troposphere. *Science* 276, 1058–1065.

trihydrate. A convenient starting point in the present discussion will thus be an examination of surface reactions themselves. An added degree of complexity arises when the particle is liquid, as is the case for droplets in the troposphere, and possibly for stratospheric sulfate aerosol, which may be in the form of supercooled liquid sulfuric acid.

Formal kinetic equations for reactions on the surface of atmospheric particles often start from the *Langmuir adsorption isotherm*, which is the simplest of equations that expresses the partitioning of gas between surface and gas phases. The isotherm makes several assumptions, including one that states that all surface sites are equivalent and that there are no interactions between molecules adsorbed on them. With these assumptions, it may be shown that the *surface coverage*, θ_X , is given by the equation

$$\theta_X = \frac{bp_X}{1 + bp_X} \quad [16]$$

where b is a constant equal to the ratio of rate coefficients for adsorption onto the surface and desorption from it. For the low partial pressures of adsorbates present as minor constituents in the atmosphere, θ_X is likely to be very nearly a linear function of p_X , although the full equation might be needed for very strongly adsorbed reactants or for high partial pressures in the atmosphere. An extended treatment for the situation where two different species, X and Y , are adsorbed is straightforward, with the surface coverages θ_X and θ_Y both entering into the equation.

The kinetics of reaction for a single reactant are then developed by including a loss (probably decomposition) of the adsorbed molecules. If the rate coefficient for this first-order loss process is k_L , the rate of chemical change is

$$\text{Rate} = k_L \theta_X \approx k_L b_X p_X, \quad [17]$$

the second (approximate) equality applying for low θ_X . The overall kinetics are first order in p_X (and thus $[X]$). Analogous equations can obviously be developed for the case of X and Y both adsorbed on a surface, and interacting there. If both species are weakly adsorbed (small surface coverage), the kinetics will be second order, with a rate proportional to $p_X p_Y$. If one (or both) of the reactants is adsorbed too strongly to use the low-pressure limiting equation, then the full form of the adsorption isotherm must be employed.

A key parameter in the discussion of surface processes is the *uptake coefficient*, γ , which is the ratio of molecules lost to

a surface to the number of gas–surface collisions that occur. If the rate of collision of a molecule X with an area A of the surface is ω , then the rate of loss of X per unit volume, $-d[X]/dt$, is equal to $\gamma \cdot \omega/V$, where V is the volume of the system. The kinetic theory of gases shows that

$$\omega = \frac{\bar{c}A[X]}{4} \quad [18]$$

so that

$$-\frac{d[X]}{dt} = \frac{\gamma \bar{c}A[X]}{4V} \quad [19]$$

Now the loss of X may also be described in terms of phenomenological rate equations of the type

$$-\frac{d[X]}{dt} = k_S \{S\} [X] \equiv k'_S [X] \quad [20]$$

where $\{S\}$ represents the number of active surface sites per unit area and k_S and k'_S are second-order and the corresponding pseudo-first order rate coefficients for the surface loss process. It follows, from a comparison of eqns [19] and [20], that

$$k'_S = \frac{\gamma \bar{c}(A/V)}{4} \quad [21]$$

Uptake coefficients may be determined by a variety of experimental methods. Regardless of whether the molecule is removed by reaction on or within the particle, or by dissolving in it, eqn [21] provides the link between the kinetics of the uptake process and the uptake coefficient. It will be evident that the reactive uptake coefficient is equivalent to the *reaction probability*.

A complication obviously arises if a molecule does not react irreversibly, but can desorb again from a surface, or come out of solution to reenter the gas phase. In such cases, γ can apparently be time dependent, and the measurement of the variation of γ with time provides one way of examining these reversible processes. In the case of the atmosphere, the most important aspect concerns the partitioning of molecules between gaseous and liquid phases. Solubilities of gases at low solute concentrations obey *Henry's law*

$$[X(s)] = H_X p_X \quad [22]$$

where $[X(s)]$ is the concentration of X in solution, p_X is its pressure in the gas phase, and H_X is the Henry's law coefficient

(which is a function of temperature). Henry's law expresses an equilibrium situation, in which the fluxes of molecules into and out of the liquid are equal. However, it is straightforward to calculate the forward and reverse fluxes, and thus the net flux into the liquid, under nonequilibrium conditions. The first term comes immediately from eqn [20], while the second requires use of the diffusion equation for transport of the molecules from the bulk liquid to the interface. If the coefficient of diffusion for this latter process is D , then it may be shown that

$$\frac{1}{\gamma_t} = \frac{1}{\gamma_0} + \frac{\pi^{1/2} \bar{c}}{4H_X R T D^{1/2}} t^{1/2} \quad [23]$$

where γ_0 and γ_t are the uptake coefficients at time 0 and time t , respectively. The equation shows how H_X can be calculated from measurements of uptake coefficient as a function of time, or, conversely, how the variation of uptake coefficient with time may be estimated from a knowledge of the solubility of the gas. Note that at 'infinite' time, γ_t becomes zero: the system has reached equilibrium.

The material developed so far is applicable to both physical processes – adsorption, absorption, or solution – and chemical change. In the particular case of chemical change, we can envisage two possibilities. Either the reaction may involve an interaction of the gas-phase reactant with the surface or the bulk constituent of the particle, or it may involve reaction with some second species already adsorbed on, or dissolved in, the particle. The concepts set out earlier remain applicable in the second, 'bimolecular,' situation, but the value of $\{S\}$ at the surface (eqn [20]), or the concentration of the partner reactant Y, in solution, will be determined by factors similar to those already determining the adsorption or solubility of Y.

Finally, it is necessary to examine the kinetics of reaction *within* the liquid phase itself. The solvent obviously has the potential to exert a considerable influence on the course of chemistry in the liquid phase. In air at 1 atm pressure, and at ambient temperature, the molecules themselves occupy only roughly 0.2% of the total volume; in liquids, the molecules can make up half the volume. At pressures of 1 atm and below, we have been able to assume that the reactant molecules undergo essentially unhindered motion, and that assumption lies behind the various formulations of kinetics that we have discussed in previous sections. In distinction, in liquids the reactive molecules must squeeze past the solvent molecules (or each other, if one species is also the bulk liquid) if they are to reach each other and undergo reaction. Reactants, activated complexes or intermediates, and products can also all interact with the solvent. One manifestation of the interaction with intermediates is that energy removal in association reactions, such as the combination processes [XVI] or [XVII], is virtually instantaneous, and the systems always display pure second-order kinetics in the liquid phase, in contrast to the behaviour described in the last section for gas-phase reactants. Interactions of the reactants and the solvent (especially water) may make the formation of ions energetically more favourable than in the gas phase. New reaction channels may thus become accessible, and the kinetics of the processes can be influenced by the attractive or repulsive electrostatic interactions between the reactants, among many other factors.

Simple treatments of liquid-phase kinetics often start from the concept of the *encounter pair* of reactants that find themselves together within a *solvent cage*. Two extreme cases can be envisaged. In the first, the two species are very highly reactive towards each other, and undergo chemical transformation within a very few 'collisions' within the cage. The rate-determining process is then the diffusion of the reactants through the solvent to form the encounter pair, and the process is a *diffusion-controlled reaction*. At the other extreme, the activation energy for reaction may require the partners to pick up appreciable amounts of energy as they shake against each other within the cage, so that the kinetics are controlled by the rate of reaction within the cage, rather than by the rate at which they reach it. *Activation-controlled* reaction kinetics then result.

For many of the liquid-phase reactions of interest in atmospheric chemistry, the intrinsic reactivity of the partners is, indeed, very high, leading to diffusion-controlled kinetic behaviour. A very elementary treatment of the diffusion-controlled rate constant, k_d , leads to the equation

$$k_d = 4\pi r_{AB} D_{AB}, \quad [24]$$

where r_{AB} is a hypothetical *encounter distance* at which two partners A and B will react, and D_{AB} is the *diffusion coefficient* for the reactants.

The encounter distance *may* be roughly the sum of the gas-kinetic radii of the partners for neutral reactants, while the appropriate diffusion coefficient *may* be similar to a mean bulk diffusion coefficient of the reactants in the solvent. Making the assumptions that these values can be taken, and with typical values of $r_{AB} = 0.5$ nm and $D_{AB} = 1.3 \times 10^{-9} \text{ m}^2 \text{ s}^{-1}$ (for Na^+ in H_2O), k_d is calculated as approximately $8 \times 10^{-18} \text{ m}^3 \text{ molecule}^{-1} \text{ s}^{-1}$ or, in the units that we have been using for rate coefficients so far, $8 \times 10^{-12} \text{ cm}^3 \text{ molecule}^{-1} \text{ s}^{-1}$. In liquid-phase kinetics, it is more conventional to use molar units for concentrations, so that the equivalent figure is $(6 \times 10^{23} \times 10^3 \times 8 \times 10^{-18}) \approx 5 \times 10^9 \text{ dm}^3 \text{ mol}^{-1} \text{ s}^{-1}$. In whatever units this rate coefficient is expressed, it is evidently about 40 times smaller than the maximum gas-kinetic rate coefficient. In general, a rate coefficient of $>10^9 \text{ dm}^3 \text{ mol}^{-1} \text{ s}^{-1}$ for an aqueous-phase reaction is taken to be indicative of a diffusion-controlled mechanism.

One of the largest known rate coefficients for a condensed-phase process is that for the very important reaction



($1.4 \times 10^{10} \text{ dm}^3 \text{ mol}^{-1} \text{ s}^{-1}$ at 298 K). The magnitude mainly reflects the large diffusion coefficients in water of OH^- and, especially, of H^+ ; the rapid diffusion is itself a consequence of the special mechanisms by which these ions migrate in liquid H_2O .

Although the diffusion coefficient is most important in making reaction [XVIII] so fast, there is another factor operating that may be dominant in other reactions. The positive and negative ions attract each other, so that the effective encounter distance can be much greater than the gas-kinetic collision distance; that is, r_{AB} has to be replaced by r_{eff} in eqn [24]. For $r_{AB} = 0.5$ nm, straightforward electrostatic calculations indicate that r_{eff} ought to be about 0.2 nm for oppositely charged ions

(and 0.7 nm for like charges) in water with a relative permittivity of 78. However, it seems that this bulk permittivity is inappropriate to the highly ordered solvent molecules in the immediate vicinity of the ions, and that more realistic values of r_{eff} ought to be 10 nm and 10^{-9} nm for unlike and like charges. These values mean, of course, that oppositely charged ions will react 20 times faster than their neutral analogues, under similar conditions, while similarly charged ions can be assumed not to react at all.

While this discussion has so far centred on the behaviour of the atmospherically dominant class of diffusion-controlled reactions, some processes of interest are activation controlled. One characteristic of such reactions is that the activation energy may be smaller than for the equivalent gas-phase reaction, because the reactant pair undergoes many individual 'collisions' at each encounter, whereas, in the gas phase, the collision and the encounter are the same thing. A particularly interesting property shown by activation-controlled ionic reactions is that of the *kinetic salt effect*. Rate coefficients are affected by the presence of other ionic species present in the solution that do not themselves participate in the reaction. Interactions between oppositely charged partners are slowed down by the presence of such salts. In the atmosphere such effects may be of significance, since water droplets may contain substantial amounts of sea-salt or other similar species.

See also: Chemistry of the Atmosphere: Laboratory Kinetics; Principles of Chemical Change.

Further Reading

- Arnaut, L.G., Formosinho, S.J., Burrows, H., 2007. *Chemical Kinetics: From Molecular Structure to Chemical Reactivity*. Elsevier, Amsterdam.
- Barker, J.R. (Ed.), 1995. *Progress and Problems in Atmospheric Chemistry*. World Scientific Publishing, Singapore.
- Forst, W., 2003. *Unimolecular Reactions: A Concise Introduction*. Cambridge University Press, Cambridge.
- Huthwelker, T., Ammann, M., Peter, T., 2006. The uptake of acidic gases on ice. *Chemical Reviews* 106, 1375–1444.
- Molina, M.J., Molina, L.T., Kolb, C.E., 1996. Gas-phase and heterogeneous kinetics of the troposphere and stratosphere. *Annual Review of Physical Chemistry* 47, 327–367.
- Moortgat, G.K., Barnes, A.J., Le Bras, G., Sodeau, J.R. (Eds.), 1994. *Low Temperature Chemistry of the Atmosphere*. Springer-Verlag, Berlin.
- Pilling, M.J., 1996. Radical–radical reactions. *Annual Review of Physical Chemistry* 47, 81–108.
- Pilling, M.J., Smith, I.W.M. (Eds.), 1987. *Modern Gas Kinetics*. Blackwell Scientific Publications, Oxford.
- Smith, I.W.M., 2003. Laboratory studies of atmospheric reactions at low temperatures. *Chemical Reviews* 103, 4549–4564.
- Smith, I.W.M., 2008. The temperature-dependence of elementary reaction rates: beyond Arrhenius. *Chemical Society Reviews* 37, 812–826.
- Smith, I.W.M., 2008. *Low Temperatures and Cold Molecules*. Imperial College Press, London.
- Su, T., Bowers, M.T., 1979. Classical ion–molecule collision theory. In: Bowers, M.T. (Ed.), *Gas Phase Ion Chemistry*, vol. 1. Academic Press, New York, pp. 83–118.
- Wayne, R.P., 2000. Photochemistry and Kinetics Applied to Atmosphere. Chapter 3 in *Chemistry of Atmospheres*, third ed. Oxford University Press, Oxford. pp. 97–137.
- Yang, X.-M., 2007. State-to-state dynamics of elementary bimolecular reactions. *Annual Review of Physical Chemistry* 58, 433–459.
- Yang, X.-M., Liu, K. (Eds.), 2004. *Modern Trends in Chemical Reaction Dynamics*, Parts I and II. World Scientific Publishing Co. Pte, Singapore.

Synopsis

In this article, we first summarize the production processes for ions, including solar photoionization, photoelectron-impact ionization, and, mostly for X-ray photons, secondary and further electron impact ionization. In the auroral regions, impact of keV electrons and heavier particles is important. In these regions, very energetic secondary and further electrons are produced, which can further ionize the neutral constituents. We summarize the processes that are important in forming ion-density profiles. Chapman layer theory, while interesting pedagogically and historically, does not truly describe any real ionospheric layer. The *D*, *E*, *F*₁, and *F*₂ layers of the ionosphere are differentiated mostly by the production mechanisms for ions in the regions. Ions in the *F*₁ and *F*₂ regions are produced by EUV photons at roughly the same altitudes, but the *F*₁ region comprises molecular ions and the *F*₂ region mostly comprises atomic ions, which are differentiated by their loss processes: dissociative recombination for molecular ions in the *F*₁ region, and ion–molecule reactions and transport by diffusion for the atomic ions in the *F*₂ region. The chemistry of the *E* and *F* regions is discussed, followed by the chemistry of the *D* region. In the *E* and *F*₁ regions, the terminal ions are those for which the parent neutrals have low ionization potentials. In the *D* region, positive cluster ions, which terminate in proton hydrates, are ultimately formed. The terminal cluster ions are those for which the proton affinities are large. In this region of the ionosphere, there are also metal ion layers that are formed from metal atoms, which have extremely low ionization potentials. Metal atoms are formed in narrow layers by ablation of meteors, and are ionized by photoionization and charge transfer from the molecular ions O₂⁺ and NO⁺. In the lower *D* region, negative ions are formed, and they too may cluster. The terminal ions in this region are those for which the parent neutral has a large electron affinity. Mutual neutralization may be an important loss process in this region.

Introduction

Most of the ionosphere is to be found in the outer reaches of the atmosphere, the thermosphere, where the neutral temperature *T* increases with altitude from a minimum of 160–190 K near the mesopause, which is at about 85 km in the terrestrial atmosphere. At great heights, when the conductivity becomes large, the temperature eventually reaches a constant value, the exospheric temperature (*T*_∞). The value of the exospheric temperature depends on solar activity and ranges from about 700 to 2000 K in the terrestrial thermosphere. Most of the ionosphere consists of equal densities of positive ions and electrons, although in the lowest regions of the ionosphere, below ~75 km, significant densities of negative ions are present. The densities of positive atomic ions attain maximum values at the *F*₂ peak, near 300 km. Positive ions are initially produced by direct ionization of neutrals, the ultimate source of which is largely solar energy. This solar input can take the form of photons and photoelectrons during the daytime, or, in the auroral regions, of particles that have their origin in the magnetosphere: energetic electrons, and less frequently protons or even heavier particles. Some ions are created in the middle and lower atmosphere by cosmic ray bombardment.

The ions that are produced are then transformed by ion–molecule reactions, lost by recombination, or transported to other regions of the atmosphere. In the photochemical equilibrium (PCE) region, the densities of ions are not affected by transport. The upper boundary of the PCE region is the altitude where the time constant for loss of an ion by chemical reactions, *τ*_c is equal to that for transport by diffusion *τ*_D. The chemical time constant or lifetime is given by

$$\tau_c = \frac{n_i}{L_i} \quad [1]$$

where *n_i* is the number density of the ion and *L_i* is the total rate of loss by chemical reactions. *L_i* can be represented by

$$L_i = \sum_l k_l n_i n_l \quad [2]$$

In this equation *k_l* is the rate coefficient for reaction *l*, and *n_l* is the number density of the species that reacts with ion *i*. The time constant for transport by diffusion, *τ*_D, can be estimated as *H_i²/D_i*, where *D_i* is the diffusion coefficient of the ion and *H_i* = *kT_p/m_ig* is the scale height of the ion. In this formula, *k* is Boltzmann constant, *m_i* is the mass of the ion, and *g* is the acceleration of gravity; the plasma temperature *T_p* is given by the sum *T_e* + *T_i*, where *T_i* is the ion temperature and *T_e* is the electron temperature. In general the PCE boundaries vary widely from one ion to another and depend on the available loss processes. Above the PCE region, transport becomes important, and must be accounted for in determining the density profiles of the ions. At very high altitudes, chemistry is not significant, and the ion-density profiles are determined only by transport processes such as diffusion and convection.

The evolution and distribution of the number density of an ion *i* is determined by the continuity equation:

$$\frac{\partial n_i}{\partial t} + \nabla \cdot \Phi_i = P_i - L_i \quad [3]$$

In this equation the production rate *P_i* includes both direct production and production by chemical reactions and *Φ_i* is the flux of species *i*. At steady state, the change in number density with time, represented by the term *∂n_i/∂t*, is zero. If only the vertical direction is considered, the divergence of the flux becomes *∂φ_i/∂z*, where *φ_i* = *n_iw_i* and *w_i* is the vertical component of the velocity. Determining the velocity of an ion is fairly complex, since the motion of an ion is determined by collisions with neutrals and with other ions, which may

themselves be in motion. In addition, the ions and electrons are constrained to move together due to the polarization or charge separation electric field that arises when the electrons, which have larger mobilities and smaller masses, separate from the ions. The presence of external electric and/or magnetic fields further complicates the equations. In general, it is not possible to solve for the ion velocities in closed form. We can, however, write an approximate one-dimensional equation for the vertical velocity of a major ion i diffusing through a stationary neutral background gas

$$w_i = -D_a \sin^2 I \left[\frac{1}{n_i} \frac{\partial n_i}{\partial z} + \frac{m_i g}{k(T_e + T_i)} + \frac{1}{T_e + T_i} \frac{\partial(T_e + T_i)}{\partial z} \right] \quad [4]$$

where I is the magnetic dip angle and the ambipolar diffusion coefficient is $D_a = (1 + T_e/T_i)D_{in}$, where D_{in} is the ion-neutral diffusion coefficient. For a minor ion diffusing through a stationary major ion the velocity is given approximately by

$$w_j = -D_i \left[\frac{1}{n_i} \frac{\partial n_i}{\partial s} + \frac{T_e}{T_i} \frac{1}{n_e} \frac{\partial n_e}{\partial s} + \frac{m_i g}{kT_i} + \frac{1}{T_i} \frac{\partial(T_e + T_i)}{\partial s} \right] \quad [5]$$

where the path s is either along the magnetic field lines or along the vertical in the absence of a magnetic field. The ion diffusion coefficient D_i must account for collisions between the ion and the major ion j , as well as ion-neutral collisions. Equations for the ion velocities in more complicated (and more realistic) situations can be found in, for example, Schunk and Nagy (2009) or Banks and Kockarts (1977).

Ionization Processes

At low and midlatitudes, ionization may be produced by the interaction of solar extreme ultraviolet (EUV) photons, X-ray photons, photoelectrons, and secondary electrons with atmospheric gases. In the auroral regions, energetic electrons, protons, and occasionally heavier positively charged particles such as O^+ , precipitate into the atmosphere along the magnetic field lines

and ionize, excite, and dissociate atmospheric neutral species. Photons in the visible region of the solar spectrum, which are not absorbed appreciably by the atmosphere, arise from the photosphere of the sun. The visible portion of the solar spectrum can be approximated as that of a blackbody with a characteristic temperature of ~ 6000 K. The shorter wavelength ionizing photons, however, arise from parts of the solar chromosphere, the transition region to the corona, and the corona, where the temperatures range from 10^4 to 10^6 K. The solar EUV and X-ray spectra differ substantially from that of a blackbody, and vary markedly from low to high solar activity. An example of the solar photon flux for the wavelength range 0–2000 Å appropriate to low solar activity is shown in Figure 1.

The photon flux of solar radiation, $F_\lambda(z)$, in a small wavelength interval around λ at an altitude z , can, for the most part, be computed from the Beer–Lambert absorption law:

$$F_\lambda(z) = F_\lambda^\infty \exp(-\tau(\lambda, z)), \quad [6]$$

where F_λ^∞ is the solar photon flux outside the atmosphere and $\tau(\lambda, z)$ is the optical depth which, in the plane parallel approximation, is given by

$$\tau(\lambda, z) = \sum_j \int_z^\infty n_j(z') \sigma_j^a(\lambda) \sec \chi \, dz'. \quad [7]$$

Here, $n_j(z')$ is the number density of species j at altitude z' , $\sigma_j^a(\lambda)$ is the absorption cross section of species j at wavelength λ , and χ is the solar zenith angle, the angle of the line of sight to the sun with respect to the local vertical.

For χ greater than about 75° the optical depth must be computed by numerical integration over the path of the radiation in spherical geometry. For $\chi \leq 90^\circ$, the optical depth is

$$\tau(\lambda, z) = \sum_j \int_z^\infty n_j(z') \sigma_j^a(\lambda) \left[1 - \left(\frac{r_o + z}{r_o + z'} \right)^2 \sin^2 \chi \right]^{-0.5} dz'. \quad [8]$$

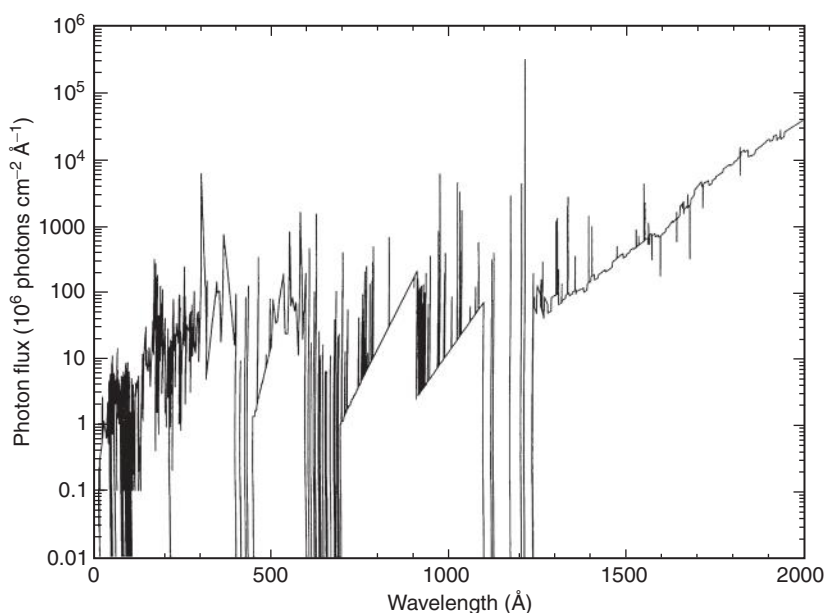


Figure 1 Solar photon flux as a function of wavelength from 18 to 2000 Å. From the SC#21REFW spectrum of Hinteregger (private communication).

For χ larger than 90° the optical depth is given by

$$\tau(\lambda, z) = \sum_j \left\{ 2 \int_{z_s}^{\infty} n_j(z') \sigma_j^a(\lambda) \left[1 - \left(\frac{r_o + z_s}{r_o + z'} \right)^2 \right]^{-0.5} dz' - \int_z^{\infty} n_j(z') \sigma_j^a(\lambda) \left[1 - \left(\frac{r_o + z}{r_o + z'} \right)^2 \sin^2 \chi \right]^{-0.5} dz' \right\} \quad [9]$$

where z_s is the tangent altitude, that for which the solar zenith angle is 90° along the line connecting the sun with the point of interest.

In a one-species atmosphere, the rate of absorption of solar photons of wavelength λ at altitude z is

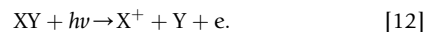
$$q^a(\lambda, z) = F_\lambda(z) \sigma^a(\lambda) n(z). \quad [10]$$

It can be easily shown that, for a given wavelength, the absorption rate maximizes where the optical depth is unity. The altitudes of unit optical depth for photons characterized by wavelengths from the X-ray region to the far-UV at moderate resolution for overhead sun in the terrestrial atmosphere are shown in Figure 2. The shape of the curve reflects closely that of the absorption cross sections $\sigma^a(\lambda)$ of the major atmospheric gases. The absorption threshold of N_2 is about 12.14 eV; therefore, it does not absorb photons longward of about 1021 Å; the absorption threshold of O is 13.618 eV, and it does not absorb photons longward of about 910 Å. Thus in the terrestrial thermosphere and mesosphere, O_2 is the primary absorber between about 1000 and 2200 Å. At longer wavelengths, ~ 2200 –3000 Å, the near ultraviolet photons are absorbed by ozone in the middle atmosphere. Photons in the EUV (100–1000 Å) and parts of the far-UV (1000–1750 Å) are absorbed in the thermosphere and upper mesosphere, and X-rays ($\lambda < 100$ Å) are absorbed in the lower thermosphere and mesosphere. The strong solar Lyman alpha emission line at 1216 Å penetrates through a window in the O_2 absorption cross sections to an altitude of ~ 75 –80 km.

Photoionization of an atom or molecule (X) may be represented as



where e^* is a photoelectron. Photoelectrons produced by EUV photons are released with a range of energies that average less than 20 eV. The primary photoelectron spectrum near 170 km in the terrestrial atmosphere is shown in Figure 3, where it is compared to that at 100 km. The threshold energy at which an atom or molecule can be ionized is called the ionization potential (I_p). Ionization potentials of species commonly found in the ionospheres of the Earth, planets, and satellites are shown in Table 1. For molecules XY, photodissociative ionization may also occur:



In the latter equation, X and Y may be atomic or molecular fragments. The energy of the photoelectron in reaction [11] is given by

$$E_{pe} = h\nu - I_X - E_{ex} \quad [13]$$

and that in reaction [12] is

$$E_{pe} = h\nu - E_d - I_X - E_{ex}, \quad [14]$$

where I_X is the ionization potential of species X, E_d is the dissociation energy of molecule XY, and E_{ex} is the excitation energy of the product species. In photodissociative ionization, E_{ex} also includes the kinetic energy of the fragments, which has been found to be large for many atmospheric molecules.

The rate of ionization of a species j by photons of wavelength λ at an altitude z is given by

$$q_j^i(\lambda, z) = F_\lambda(z) \sigma_j^i(\lambda) n_j(z), \quad [15]$$

where $\sigma_j^i(\lambda)$ is the photoionization cross section. The expression above must be integrated over the solar ionizing spectrum to give the total photoionization rate, and all species must be included in the calculation of the optical depth to obtain the solar photon flux at a given wavelength. In addition, it is often necessary to take into account ionization to different final internal states of the products, so the partial cross sections or branching ratios to the different states are needed. In the atmospheres of magnetic planets, photoelectrons originating on the dayside may travel along the magnetic field lines to the

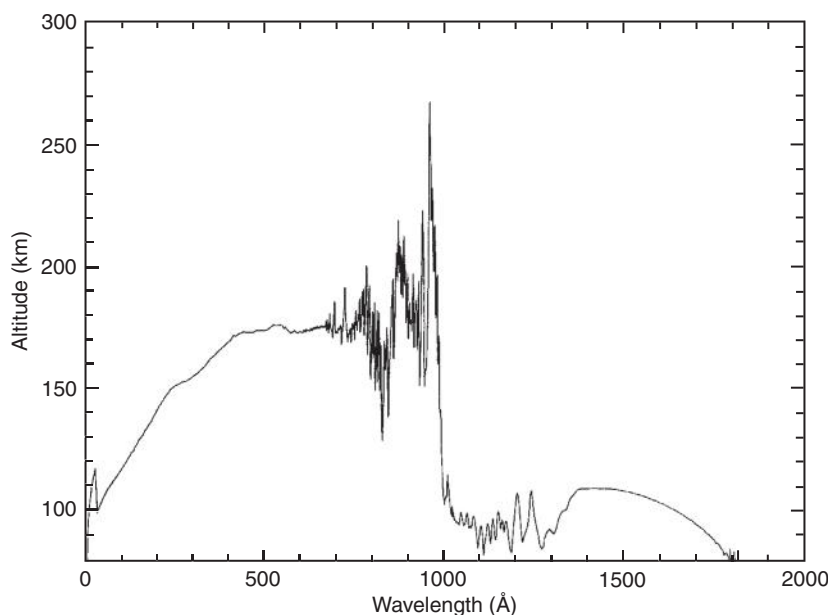


Figure 2 Altitude of unit optical depth in the terrestrial ionosphere as a function of wavelength in the extreme and far-UV wavelength regions.

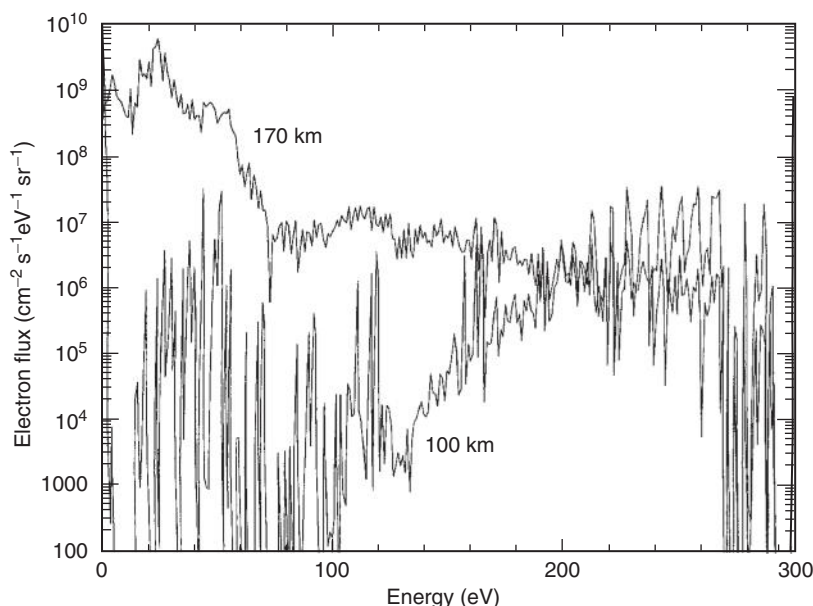


Figure 3 Primary photoelectron flux in the terrestrial atmosphere at 170 km, near the F_1 peak, and at 100 km. Note that the spectrum at 100 km is significantly harder than that at 170 km.

Table 1 Ionization potentials (I_p) of common species in the atmospheres of the Earth and planets^a

High I_p (eV)		Medium I_p (eV)		Ionized by Ly α (eV)	
Species	I_p	Species	I_p	Species	I_p
He	24.59	H ₂ O	12.61	C ₄ H ₂	10.18
Ne	21.56	CH ₄	12.51	CH ₃	9.84
Ar	15.76	SO ₂	12.32	C ₃ H ₆	9.73
N ₂	15.58	CH ₃ CN	12.19	NO	9.264
H ₂	15.43	C ₂	12.11	C ₆ H ₆	9.246
N	14.53	O ₂	12.07	C ₂ H ₃	8.9
CO	14.01	C ₂ H	11.70	Si	8.15
CO ₂	13.77	HC ₃ N	11.64	C ₂ H ₅	8.13
O	13.618	C ₂ H ₆	11.52	HCO	8.10
H	13.598	C ₂ H ₂	11.40	N ₂ H ₄	8.10
HCN	13.60	C	11.26	C ₃ H ₇	8.09
OH	13.00	C ₃ H ₈	10.95	Fe	7.87
		CH	10.64	Mg	7.65
		C ₂ H ₄	10.51	<i>trans</i> -HCNH	7.0 ^b
		H ₂ S	10.45	<i>cis</i> -HCNH	6.8 ^b
		CH ₂	10.40	c-C ₃ H ₃	6.6
		S	10.35	Ca	6.11
				Na	5.14

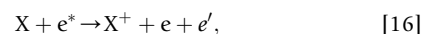
^aComputed with data from Lias, S.G., Bartmess, J.E., Liebman, J.F., Holmes, J.L., Levin, R.D., Mallard, W.G., 1988. Gas phase ion and neutral thermochemistry. Journal of Physical Chemistry Reference Data 17 (Suppl. 1), except as noted.

^bFrom Nesbitt, F.L., Marston, G., Stieff, L.J., 1991. Measurement of the photoionization spectra and ionization thresholds of the H₂CN and D₂CN radicals. Journal of Physical Chemistry 95, 7613–7617.

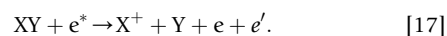
conjugate point on the nightside, where the field lines reenter the atmosphere. These suprathermal electrons may interact with the atmosphere producing ionization, excitation, and dissociation.

Auroral primary electrons have energies of the order of keV, and can ionize, excite, and dissociate atmospheric species. Some typical terrestrial auroral electron spectra are shown in Figure 4. Photoelectrons produced by photons with wavelengths

less than about 500 Å also have enough energy to carry out further ionization through electron-impact ionization:



and electron-impact dissociative ionization



In these reactions, e^* represents the energetic photoelectron or the auroral primary electron, e represents the energy degraded electron, and e' represents the secondary electron. The energy of the secondary electron $E_{e'}$ in an electron-impact ionization process [16] is given by

$$E_{e'} = E_{e^*} - I_X - E_{ex} - E_e, \quad [18]$$

where E_{e^*} is the energy of the initial primary or photoelectron, E_e is the energy of the degraded primary or photoelectron, and, as above, E_{ex} is the total excitation energy of the product ions and/or neutral fragments. For the dissociative ionization process [17], the dissociation energy of the XY molecule must also be subtracted as well. It should be noted that these ionization processes must compete with other processes that energetic electrons can carry out, such as inelastic rotational, vibrational, and electronic excitation, and (for molecules) dissociation; at low energies elastic collisions with the ambient electrons must also be considered.

The ionization rate due to electron impact $q_j^{ei}(z)$ of a species j with ionization potential I_j is given by

$$q_j^{ei}(z) = n_j(z) \int_{I_j}^{\infty} \int_0^{(E-I_j)/2} \frac{d\sigma_j^{ei}(E)}{dW_s} \frac{dF(z, E)}{dE} dW_s dE, \quad [19]$$

where $dF(z, E)/dE$ is the differential flux of primary electrons with respect to energy, and $d\sigma_j^{ei}(E)/dW_s$ is the differential cross section for production of a secondary electron with energy W_s by a primary electron with energy E . The integral over secondary energies W_s terminates at $(E - I_j)/2$ because the

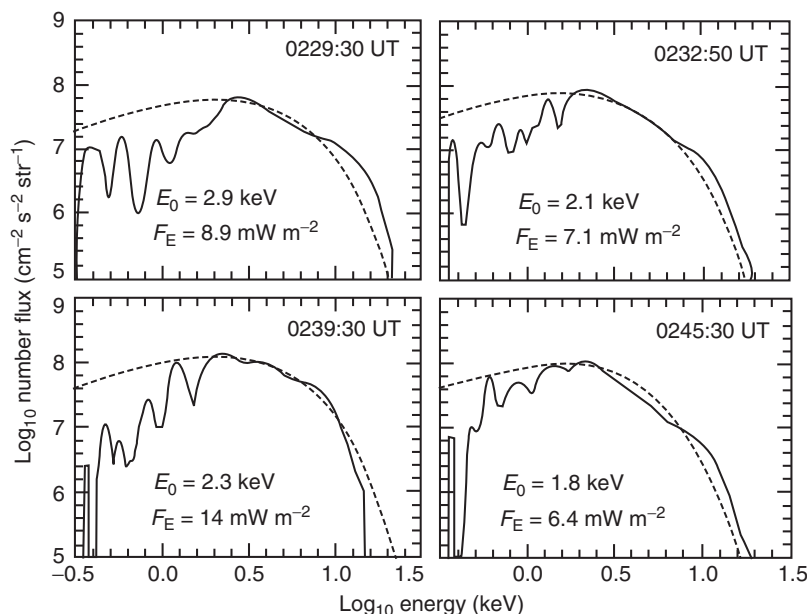


Figure 4 Typical auroral spectra derived from a coupled experiment of ultraviolet images from the POLAR satellite and electron densities derived from the Sondrestrom incoherent scatter radar in 1996. E_0 is the characteristic energy and F_E is the energy flux of the spectrum. Taken from Doe, R.A., Kelly, J.D., Lummerzheim, D., Parks, G.D., Brittnacher, M.J., Germany, G.A., Span, J., 1997. Initial comparison of POLAR UVI and Sondrestrom IS radar estimates for auroral electron energy flux. *Geophysical Research Letters* 24, 999–1002.

secondary electron is by convention considered to be the one with the smaller energy. Since the average energy of photoelectrons produced by EUV photons is less than 20 eV, the error incurred in cutting off the integral in eqn [19] at 300 eV or so, rather than $(E - I_j)/2$ is not serious, although for high energy auroral electrons and for photoelectrons produced by solar X-ray ionization, a larger upper limit must be used.

The differential cross section for the production of a secondary electron with energy W_s is usually adopted from an empirical formula that is normalized so that

$$\sigma_j^{ei}(E) = \int_0^{(E-I_j)/2} \frac{d\sigma_j^{ei}(E)}{dW_s} dW_s, \quad [20]$$

where $\sigma_j^{ei}(E)$ is the measured total ionization cross section at primary electron energy E . One formula in common use is:

$$\frac{d\sigma_j^{ei}(E)}{dW_s} = \frac{A(E)}{1 + (W_s/\bar{W})^{2.1}}, \quad [21]$$

where $A(E)$ is a normalization factor and $\bar{W}$ is an empirically determined constant, which has been found to be equal to within about 50% of the ionization potential for a number of species.

In practice, discrete energy loss of electrons can be easily treated numerically if the electrons can be considered to lose their energy locally, which is generally a good approximation for photoelectrons near and below the peak altitude for production of ions. Since elastic scattering of electrons by neutrals mostly changes the direction of the incident electron, and not its energy, only inelastic processes such as excitation, dissociation, and ionization need to be considered in this approximation. For ionization, of course, the energy distribution of the secondary

electrons must be taken into account, but not the scattering angles of either the primary or secondary electron. Below the lowest thresholds for excitations, energetic electrons lose their energy in elastic collisions with thermal electrons. This process and rotational excitation are often approximated as continuous rather than as discrete excitations.

The slowing down of high-energy auroral primary electrons or energetic photoelectrons arises from both elastic and inelastic scattering processes, and cannot be treated using the local energy loss approximation. In solving the equations for electron transport, the angle through which the primary electron is scattered, as well as the change in energy of the primary electron and the production of any secondaries must be taken into account. Electron-transport methods should also be applied to determine the spectrum of high-altitude photoelectrons, which stream up from near the ionospheric peak. Discussion of the electron-transport equations can be found, for example, in the monographs by Rees (1989) and by Schunk and Nagy (2009).

Several methods for approximating the energy deposition of auroral electrons are currently in use. In the continuous slowing down approximation (CSDA), the electrons are assumed to lose their energy continuously and at the same rate, and to be scattered only in the forward direction. The CSDA provides only a rough approximation to the depth of penetration of the electrons, and to the rates of ion production and other energy loss processes. In the 'two-stream' approximation, the electrons are assumed to be scattered in either the forward or backward direction. Implementation of this method requires only the backscattering probabilities, rather than complete angular differential cross sections. The method has been generalized to multistream models, in which the solid angle range of the electrons is divided into 20 or more intervals, so more or less complete differential cross sections are required. Monte Carlo

methods have also been used to model auroral precipitation and photoelectron energy deposition.

Ion Density Profiles

Density profiles of molecular ions are often approximated as idealized Chapman layers. A Chapman layer is one in which a single ion is produced from a single neutral species in an isothermal atmosphere by interactions with photons of a single wavelength. The positive ions are assumed to be destroyed by recombination with electrons or negative ions. Dissociative recombination (DR) is the process of electron recombination with a molecular ion that leads to the production of neutral fragments:



PCE is assumed to prevail in a Chapman profile.

The photoionization rate q^i in a Chapman layer is given by eqn [15], with the subscripts for various species and wavelengths removed:

$$q^i(z) = F(z)\sigma^i n(z). \quad [23]$$

Here, n is the number density of the single ionizable species, $F = F^\infty \exp(-\tau)$ is the local solar photon flux, and σ^i is the ionization cross section at the single wavelength λ . The acceleration of gravity g is also assumed to be constant, and therefore the density profile of the neutral species in a Chapman layer is given by $n(z) = n_0 \exp(-z/H)$, and the optical depth τ is $nH\sigma^a \sec \chi$, where the constant scale height H of the neutral species is given by kT/mg . In this expression, T is the (assumed) constant thermospheric temperature, and m is the mass of both the single neutral species and the single ion. The ionization cross section is sometimes expressed as

$$\sigma^i = \eta^i \sigma^a, \quad [24]$$

where η^i is the ionization efficiency.

The maximum photoionization rate occurs where the optical depth τ is unity, and therefore where $n_{\max} = 1/(\sigma^a H \sec \chi)$. The maximum photoionization rate is

$$q_{\max, \chi}^i = \frac{F^\infty}{e} \frac{\sigma^i}{\sigma^a H \sec \chi} = \frac{q_{\max, 0}^i}{\sec \chi}, \quad [25]$$

where $q_{\max, 0}^i$ is the maximum ionization rate for $\chi = 0$. Thus the maximum ionization rate decreases with increasing solar zenith angle. We can obtain a simple expression for the ionization rate by defining the altitude of maximum ionization for overhead sun as $z = 0$. Then $n_0 = (\sigma^a H)^{-1}$, and, expressing F^∞ in terms of $q_{\max, 0}^i$, the photoionization rate is

$$q^i(z) = q_{\max, 0}^i \exp\left[1 - \frac{z}{H} - \sec \chi \exp(-z/H)\right]. \quad [26]$$

The third term in the exponent of eqn [26] is significant only near and below the peak. At high altitudes ($z \rightarrow \infty$) the ionization rate decreases exponentially, following that of the neutral species from which it is produced, and below the peak (as $z \rightarrow -\infty$), the ionization rate rapidly approaches zero. As the solar zenith angle increases, the altitude of the production peak rises and the magnitude of the production rate maximum decreases. Figure 5 shows a production profile for an idealized Chapman layer on a semilog plot, which clearly shows the asymmetry with respect to the maximum.

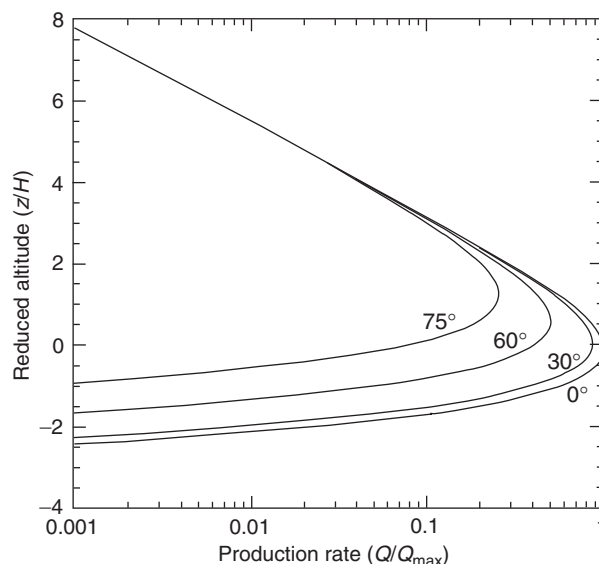


Figure 5 Ideal Chapman absorption (or production) rate profile for solar zenith angles of 0, 30, 60, and 75°. The altitude is scaled so that $z = 0$ is the altitude of maximum absorption for overhead sun.

In a Chapman layer, PCE prevails, and therefore the photoproduction rate of the molecular ion is equal to the loss rate due to DR (process [22]):

$$q^i(z) = \alpha_{\text{DR}} n_i(z) n_e(z) = \alpha_{\text{DR}} n(z)^2 \quad [27]$$

where α_{DR} is the DR coefficient and n_e is the electron density. Therefore, the density of the ion as a function of altitude is given by

$$n(z) = \left(\frac{q^i(z)}{\alpha_{\text{DR}}}\right)^{1/2} = \left(\frac{q_{\max, 0}^i}{\alpha_{\text{DR}}}\right)^{1/2} \exp[1/2 - z/(2H) - 1/2 \sec \chi \exp(-z/H)]. \quad [28]$$

At altitudes above the peak, the scale height of the ion in a Chapman layer approaches $2H$.

Actual ion-density profiles differ from the ideal Chapman profile for several reasons. First, in a real atmosphere, ionization is produced by solar photons over a wide range of wavelengths, from the EUV to the X-ray regions; these photons reach unit optical depth over a wide range of altitudes, as Figure 2 shows. Second, although thermospheric neutral temperatures reach constant values of T^∞ at high altitudes, they are usually characterized by increasing temperatures near the altitude of peak ion production. More important, T_e and T_i are not equal to the neutral temperature T_n , but diverge substantially from T_n near or slightly above the ion peak. The DR coefficient α_{DR} is not constant, but varies with the particular ion and inversely with T_e . The major ion produced is often transformed by ion-molecule reactions before it can undergo DR. Most important, photoionization is supplemented by electron-impact ionization, the peak ionization rate of which occurs slightly below the peak photoionization rate of EUV photons, and greatly below the peaks in the photoionization rate due to the absorption of X-rays. In fact, energetic photoelectron and secondary electron impact are the major sources of ionization in the regions in

which X-rays are absorbed. Energetic electrons are not extinguished, as are photons, in producing an ion. Although precipitation of auroral electrons may be characterized by a production profile that appears similar to that of a Chapman-type layer, the actual ion production and density profiles differ greatly from the idealized profile because of the latter effect and the fact that auroral electron spectra are far from monoenergetic (Figure 4).

All these factors combine to make the shape of real ion-density peaks very much broader than that of an ideal Chapman layer. Chapman theory predicts that peak magnitudes vary as $(\cos \chi)^{1/2}$, and as $F^{1/2}$ but underlying this prediction is the assumption that the neutral atmosphere does not change with either χ or F , which is clearly insupportable. Also, photon fluxes in various regions of the ionizing spectrum vary differently with solar activity. Nonetheless, the idealized concept of the Chapman layer, while not strictly applicable to real ionospheres, is useful for a pedagogical first order understanding of the ion production and density peak shapes, and their behavior as the solar zenith angle and solar fluxes change.

Ionospheric Regions

The division of the ionospheres of the Earth and planets into regions is based on the structure of the terrestrial ionosphere, which consists of overlapping layers of ions. These layers are the result of changes in both the composition of the thermosphere and the sources of the ionization, and are illustrated in Figure 6. The major molecular ion layer is the F_1 layer, which is produced by absorption of EUV photons with wavelengths between about 200 and 1000 Å by the major thermospheric species. The F_1 peak appears where the ion production due to solar EUV absorption as a whole maximizes. The column densities of the neutral absorbing species in the F_1 region are of the order of 10^{17} cm^{-2} . The ions in the F_1 region are mostly molecular and the ions are in PCE. In the terrestrial atmosphere, the F_1 peak appears in the electron density profile only as a ledge in the range 170–200 km. In the ionospheres of Venus and Mars, the F_1 peaks are the absolute maxima of the electron density profiles.

The F_2 region, if it appears, is the highest region in an ionosphere, where the major ions are monatomic. The F_2 region is still considered to be part of the F region because the production rate profile of the atomic ions peak in the same altitude range as those of the F_1 peak. The peak density occurs higher than that of the F_1 peak, however, because atomic ions are not destroyed by DR; even at fairly high altitudes, the major chemical loss processes are reactions with the thermospheric neutrals. The F_2 peak occurs at approximately where the chemical lifetime of the atomic ion is equal to the characteristic time for transport by diffusion, and PCE does not apply to this region.

The E region is below the F_1 region. Here the ions are produced by shorter and longer wavelength solar photons than those that produce the ions in the F_1 region. These photons include soft X-rays ($\lambda = 10\text{--}100 \text{ Å}$), which are characterized by smaller cross sections than are those of the EUV, and which therefore penetrate further into the atmosphere. In the terrestrial ionosphere, E -region ions may also be produced by absorption of solar Lyman β at 1026 Å, which can ionize O_2 and NO .

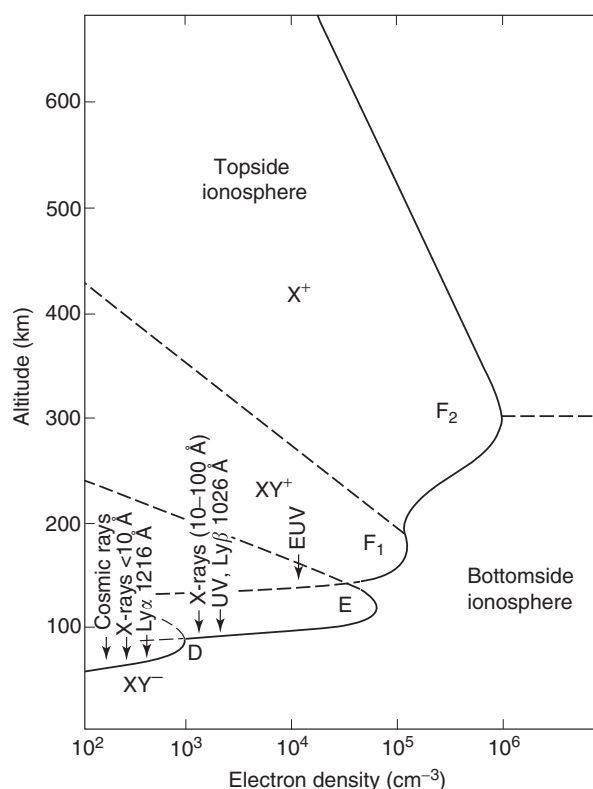


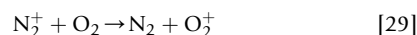
Figure 6 Schematic diagram of the various regions in the terrestrial ionosphere, and the major sources of ion production. From Bauer, S.J., Lammer, H., 2004. Planetary Aeronomy. Springer, Berlin.

High-energy photons in the soft X-ray region of the spectrum produce high-energy photoelectrons, which are capable of further ionization. The secondary electrons produced may themselves produce further ionization. In fact, as noted above, ion production in the E region is dominated by electron impact, rather than by photoionization. A comparison of the primary photoelectron fluxes near the F_1 peak of the Earth with those in the E region near 100 km has been presented in Figure 3. The spectrum near 100 km is considerably harder than that at the F_1 peak near 170 km. The E region is often considered to be a Chapman layer, although the physics and chemistry of this region diverge significantly from those of such an idealized concept. Because photons with wavelengths below 100 Å exhibit larger solar activity variations than those with longer wavelengths in the EUV region, the peaks of the ionospheric E regions tend to exhibit larger solar activity variations than do the F_1 peaks, and are more sensitive to the presence of solar flares.

In the D region, the ion production is dominated by harder solar X-rays, with wavelengths less than 10 Å, and for Earth, by ionization of NO by Lyman alpha photons, which penetrate to low altitudes through an accidental window in the O_2 absorption cross sections. This phenomenon is unique to the terrestrial thermosphere where the EUV absorption is dominated by O_2 . In CO_2 -dominated atmospheres, Lyman alpha does not penetrate to such low altitudes. The ions produced in this region are in PCE, but, as we will show below, the production mechanisms and the chemistry of the ions preclude the description of this region as a Chapman layer.

Chemistry of the *E* and *F* Regions

The major ions produced in the ionospheres of the Earth and planets are usually those from the major neutral thermospheric species. On Earth, the major ions produced by solar photons and photoelectrons are N_2^+ , O_2^+ , and O^+ . In the presence of sufficient neutral densities, however, the ions are transformed by ion–molecule reactions. In the *E* and *F* region, ion–molecule reactions may be either charge-transfer reactions, such as



in which only an electron is transferred between the two species, or ion–atom interchange reactions, such as



which involve rearrangement of the molecular bonds.

Reactions between neutrals and ions are generally faster than neutral–neutral reactions because of the longer range forces involved. The attraction between a charged particle and a nonpolar neutral is produced by the induction or polarization force: that between an ion and the dipole moment that is induced in the neutral by the proximity of the ion. The magnitude of this force depends on the polarizability α of the neutral and the charge q on the ion. The long-range potential between an ion of mass m_i and a neutral with mass m_n at separation r is $-\alpha q^2/2r^4$, and the collision or Langevin rate coefficient for energies less than a few eV is given by

$$k_L = 2\pi q(\alpha/\mu)^{1/2}, \quad [31]$$

where

$$\mu = \frac{m_n m_i}{m_n + m_i} \quad [32]$$

is the reduced mass of the reactants. The Langevin rate coefficient is temperature independent, because the capture cross section is inversely proportional to the velocity. In traditional units, for a singly charged ion, the Langevin rate coefficient can be simply calculated as

$$k_L = 2.34 \times 10^{-9} (\alpha/\mu)^{1/2} \text{ cm}^3 \text{ s}^{-1} \quad [33]$$

where μ is in unified mass units and the polarizability α is given in $\text{\AA}^3$ or 10^{-24} cm^3 . Polarizabilities of common atmospheric species are presented in Table 2. If the neutral has a permanent dipole or quadrupole moment, then a term must be added that represents the force between the charged particle and the dipole or quadrupole to that for the ion-induced dipole (e.g., Su and Bowers, 1979).

In the presence of sufficient neutral densities, there is a tendency for ion–molecule reactions to transform ions whose parent neutrals have high ionization potentials into those whose parent neutrals have low ionization potentials. This is merely a convenient restatement of the requirement that reactions be exothermic, and is rigorously true only for charge-transfer reactions. In practice, however, it is true for most other ion–molecule reactions, such as ion–atom interchange, because dissociation energies of typical atmospheric molecules differ less than ionization potentials. Notable exceptions to this tendency are the reactions

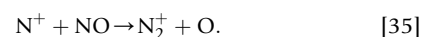


Table 2 Polarizabilities of common atmospheric species

Species	Polarizability, $\text{\AA}^3$
He	0.205
H	0.667
O	0.734
H ₂	0.807
N	1.078
H ₂ O	1.43
O ₂	1.57
Ar	1.642
NO	1.71
N ₂	1.74
C	1.78
CO	1.97
CO ₂	2.93
CH ₄	2.59
Mg	10.6
Na	23.6
Ca	25.0

Data from Mason, E.A., McDaniel, E.W., 1988. Transport Properties of Ions in Gases. Wiley, New York.

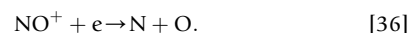
and



Even though the ionization potential of C is 11.3 eV, and that of CO is 14.01 eV, the lower ionization potential is counteracted by the large dissociation energy of CO (11.1 eV), which is twice that of CO₂ (5.45 eV). In reaction [35], the difference between the ionization potentials of N₂ (15.58 eV) and N (14.54 eV) is counteracted by the large dissociation energy of N₂ (9.76 eV) compared to that of NO (6.49 eV). Reaction [35] tends to proceed, however, via charge transfer, producing NO⁺, rather than by abstraction.

A schematic diagram of ion chemistry in the *F* and *E* regions of the ionospheres of the terrestrial planets is shown in Figure 7. In this diagram, the ionization potentials of the parent neutrals decrease downward. Where collisions are sufficiently frequent, near the *F*₁ peak and below, ionization tends to flow from the ions near the top of the diagram to those near the bottom. This principle is illustrated by the chemistry of N₂⁺. N₂ has a very high ionization potential, and N₂⁺ is produced in large quantities in the terrestrial ionosphere by photoionization and electron-impact ionization, but its steady-state densities are small. For N₂⁺, few chemical production mechanisms exist, but many loss processes are possible. Figure 8 shows altitude profiles of the production and loss mechanisms of N₂⁺ for a model of the terrestrial ionosphere.

By contrast, NO has the lowest ionization potential of the major *E*- and *F*-region ionospheric species. There are many mechanisms for the production of NO⁺, but only one significant loss process, DR:



Even in the absence of neutral NO, NO⁺ is produced by ion–atom interchange reactions in oxidizing atmospheres. The major sources of NO⁺ for a model of the terrestrial ionosphere are shown in Figure 9. The lower thermospheres of Venus and

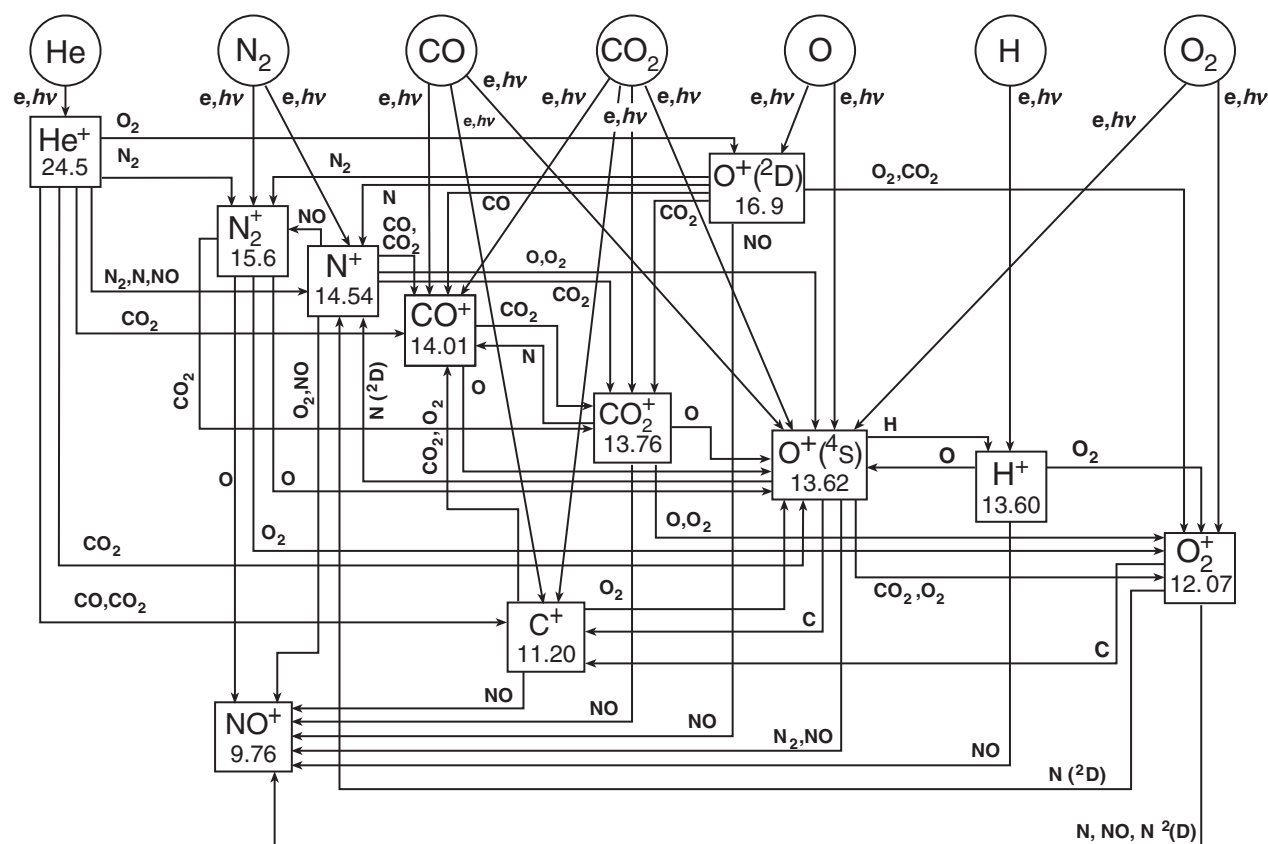


Figure 7 Schematic diagram of the major chemical reactions in the terrestrial ionosphere. The ionization potentials of the parent neutrals, which are shown below each ion, decrease from top to bottom. Where collisions are sufficiently frequent, ionization tends to flow downward on the diagram. From Fox, J.L., 2011. Dissociative recombination data needs for the aeronomy community. In: Guberman, S.L., Orel, A.E. (Eds.), *Dissociative Recombination: Theory, Experiments, and Applications*. Journal of Physics: Conference Series 300, IOP Publishing, jpcs.iop.org, Bristol, 012025. <http://dx.doi.org/10.1088/1742-6596/300/1/012025>.

Mars are dominated by CO_2 , with a small admixture of about N_2 (2–4%); O_2 is only a minor species, comprising less than 1% of either thermosphere. Because of the small ionization potentials of NO and O_2 , however, NO^+ and O_2^+ are the major ions in the E and F_1 regions of Venus and Mars, as well as in those of the Earth. Measured rate coefficients for many ion-molecule reactions have been compiled, for example, by Anicich (1994, 2003) and by Ikezoe et al. (1987). About 200 rate coefficients for reactions in oxidizing thermospheres/ionospheres have been compiled by Fox and Sung (2001). Richards (2011) has discussed the most important ~ 40 ion-molecule and DR reactions in the terrestrial F region for four ions that are in PCE: N_2^+ , O_2^+ , NO^+ , and N^+ .

In the reducing atmospheres of the outer planets and satellites, where hydrogen is abundant, and at low enough altitudes where collisions with neutrals are sufficiently frequent, ionization flows from species formed by protonation of neutrals that have smaller proton affinities to those formed by protonation of neutrals that have large proton affinities. Proton affinities for several major and minor atmospheric species of atmospheric importance are shown in Table 3. There are no *in situ* measurements of the ion composition of the outer planets, but models predict that H_3^+ and hydrocarbon ions dominate F_1 and E regions of these ionospheres. The thermosphere of Titan is composed of N_2 , CH_4 , and

H_2 , and measurements made by the Cassini Ion and Neutral Mass Spectrometer show that the ionosphere is quite rich in hydrocarbon ions with masses up to about 100 Da (unified mass units), the instrumental limit. The ion mass spectra show that the densities of hydrocarbon ions are clustered around masses that are separated by 12–14 Da, and the most important ion in each cluster is that for which the proton affinity is largest. Proton affinities are also important for the H-containing ions in the E and F regions of the terrestrial planets, and to the chemistry of the terrestrial D region, which will be discussed below.

Molecular ions are efficiently destroyed by DR (process [22]) in regions where the electron densities are sufficiently large. The long-range Coulomb forces between the ion and electron lead to DR rate coefficients (α_{DR}) that are large, in the order of a few $\times 10^{-7} \text{ cm}^3 \text{ s}^{-1}$ at thermal values of T_e , and simple theory predicts a negative temperature dependence of the rate coefficient of $T_e^{-0.5}$. In practice, however, the values of α_{DR} of common atmospheric ions are found to vary as $T_e^{-(0.5 \pm 0.25)}$. Some DR coefficients relevant to the ionospheres of the terrestrial planets are given in Table 4. Florescu-Mitchell and Mitchell (2006) have reviewed computed and measured DR coefficients and branching ratios for various sets of products compiled from the literature prior to 2006.

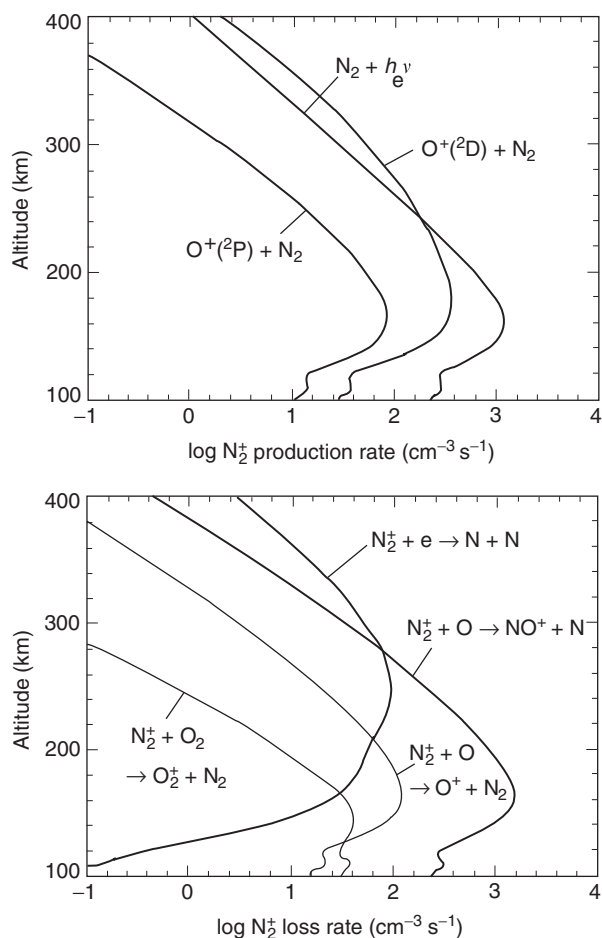


Figure 8 Major production and loss mechanisms for N_2^+ from a model of the terrestrial ionosphere. From Dalgarno, A., Fox, J.L., 1994. Ion chemistry in atmospheric and astrophysical plasmas. In: Ng, C.Y., Baer, T., Powis, I. (Eds.), *Unimolecular and Bimolecular Ion-Molecule Reaction Dynamics*. Wiley and Sons, New York, pp. 1–85.

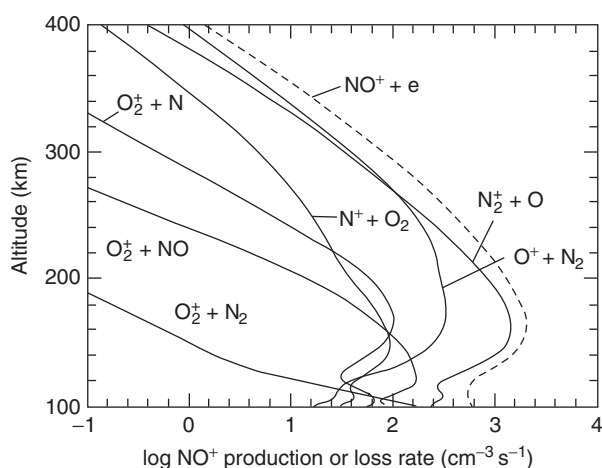


Figure 9 Major production and loss mechanisms for NO^+ in the terrestrial ionosphere. Dissociative recombination is the only significant loss mechanism. From Dalgarno, A., Fox, J.L., 1994. Ion chemistry in atmospheric and astrophysical plasmas. In: Ng, C.Y., Baer, T., Powis, I. (Eds.), *Unimolecular and Bimolecular Ion-Molecule Reaction Dynamics*. Wiley and Sons, New York, pp. 1–85.

Table 3 Selected proton affinities (eV)

Neutral species	Ion produced	Proton affinity
He	HeH ⁺	1.84
H	H ₂ ⁺	2.69
N	NH ⁺	3.37
H ₂	H ₃ ⁺	4.39
O ₂	HO ₂ ⁺	4.40
CO	COH ⁺	4.74
O	OH ⁺	5.04
N ₂	N ₂ H ⁺	5.13
CN	HCN ⁺	5.36
CH ₃	CH ₄ ⁺	5.62
CO ₂	CO ₂ H ⁺	5.67
NO	HNO ⁺	5.69
CH ₄	CH ₅ ⁺	5.72
C ₂ N	C ₂ NH ⁺	5.75
CN	HNC ⁺	5.80
OH	H ₂ O ⁺	6.16
CO	HCO ⁺	6.16
C ₂ H ₆	C ₂ H ₇ ⁺	6.23
C ₂ H ₅	C ₂ H ₆ ⁺	6.42
C	CH ⁺	6.46
C ₃ H ₈	C ₃ H ₉ ⁺	6.50
NH	NH ₂ ⁺	6.62
C ₂ H ₂	C ₂ H ₃ ⁺	6.69
C ₂	C ₂ H ⁺	6.92
C ₃ H ₇	C ₂ H ₈ ⁺	7.03
C ₂ H ₄	C ₂ H ₅ ⁺	7.05
H ₂ O	H ₃ O ⁺	7.24
HCN	HCNH ⁺	7.46
C ₂ H ₃	C ₂ H ₄ ⁺	7.56
HN ₃	H ₂ N ₃ ⁺	7.76
CH ₃ CCH	CH ₃ CCH ₂ ⁺	7.78
C ₄ H ₂	C ₄ H ₃ ⁺	7.79
c-C ₃ H ₆	(c-C ₃ H ₆)H ⁺	7.80
C ₃ N ₂	C ₃ N ₂ H ⁺	7.81
CH ₃ OH	CH ₃ OH ₂ ⁺	7.89
C ₂ H	C ₂ H ₂ ⁺	7.95
NH ₂	NH ₃ ⁺	8.13
HNC	HCNH ⁺	8.14
C ₂ H ₅ OH	C ₂ H ₅ OH ₂ ⁺	8.18
CH ₃ CN	CH ₃ CNH ⁺	8.18
CH ₂ CHCN	CH ₂ CHCNH ⁺	8.23
CH ₃ OCH ₃	CH ₃ OCH ₃ ⁺	8.31
CH ₂	CH ₃ ⁺	8.56
(H ₂ O) ₂	H ₃ O ⁺ (H ₂ O)	8.78 ^a
NH ₃	NH ₄ ⁺	8.85
N ₂ H ₄	N ₂ H ₅ ⁺	8.87
CH ₃ NNCH ₃	CH ₃ NHNCH ₃ ⁺	8.97
CH ₃ CHNH	CH ₃ CHNH ₂ ⁺	9.15
CH ₂ CHNH ₂	CH ₃ CHNH ₂ ⁺	9.37
CH ₂ NCH ₃	CH ₂ NHCH ₃ ⁺	9.41
(H ₂ O) ₃	H ₃ O ⁺ (H ₂ O) ₂	9.73 ^a

^aFrom Kebarle, P., Searles, S.K., Zolla, A., Scarborough, J., Arshadi, M., 1967. *Journal of the American Chemical Society* 89, 6393. Computed from data taken from Lias, S.G., Bartmess, J.E., Liebman, J.F., Holmes, J.L., Levin, R.D., Mallard, W.G., 1988. Gas phase ion and neutral thermochemistry. *Journal of Physical Chemistry Reference Data* 17 (Suppl. 1), except as noted.

Table 4 Selected dissociative recombination (DR) rate coefficients

Reaction	Rate coefficient, $\text{cm}^3 \text{s}^{-1}$
$\text{CO}_2^+ + e \rightarrow \text{CO} + \text{O}$	$4.2 \times 10^{-7} (300/T_e)^{0.75}$
$\text{CO}^+ + e \rightarrow \text{C} + \text{O}^a$	$1.75 \times 10^{-7} (300/T_e)^{0.55}$
$\text{CO}^+ + e \rightarrow \text{C} + \text{O}(^1\text{D})$	$0.33 \times 10^{-7} (300/T_e)^{0.55}$
$\text{CO}^+ + e \rightarrow \text{C}(^1\text{D}) + \text{O}$	$0.216 \times 10^{-7} (300/T_e)^{0.55}$
$\text{O}_2^+ + e \rightarrow \text{O}(^3\text{P}) + \text{O}(^3\text{P})^b$	$4.3 \times 10^{-8} (300/T_e)^{0.70}$ for $T_e \leq 1200 \text{ K}$ $1.63 \times 10^{-8} (1200/T_e)^{0.56}$ for $T_e \geq 1200 \text{ K}$
$\text{O}_2^+ + e \rightarrow \text{O}(^3\text{P}) + \text{O}(^1\text{D})$	$8.2 \times 10^{-8} (300/T_e)^{0.70}$ for $T_e \leq 1200 \text{ K}$ $3.1 \times 10^{-8} (1200/T_e)^{0.56}$ for $T_e \geq 1200 \text{ K}$
$\text{O}_2^+ + e \rightarrow \text{O}(^1\text{D}) + \text{O}(^1\text{D})$	$6.05 \times 10^{-8} (300/T_e)^{0.70}$ for $T_e \leq 1200 \text{ K}$ $2.29 \times 10^{-8} (1200/T_e)^{0.56}$ for $T_e \geq 1200 \text{ K}$
$\text{O}_2^+ + e \rightarrow \text{O}(^1\text{D}) + \text{O}(^1\text{S})$	$9.75 \times 10^{-9} (300/T_e)^{0.70}$ for $T_e \leq 1200 \text{ K}$ $3.69 \times 10^{-9} (1200/T_e)^{0.56}$ for $T_e \geq 1200 \text{ K}$
$\text{N}_2^+ + e \rightarrow \text{N} + \text{N}(^2\text{D})$	$1.01 \times 10^{-7} (300/T_e)^{0.39}$
$\text{N}_2^+ + e \rightarrow \text{N}(^2\text{D}) + \text{N}(^2\text{D})$	$1.01 \times 10^{-7} (300/T_e)^{0.39}$
$\text{N}_2^+ + e \rightarrow \text{N} + \text{N}(^2\text{P})$	$1.76 \times 10^{-8} (300/T_e)^{0.39}$
$\text{NO}^+ + e \rightarrow \text{N}(^2\text{D}) + \text{O}$	$3.40 \times 10^{-7} (300/T_e)^{0.5}$
$\text{NO}^+ + e \rightarrow \text{N} + \text{O}$	$0.60 \times 10^{-7} (300/T_e)^{0.5}$
$\text{H}_3\text{O}^+ + e \rightarrow \text{H}_2\text{O} + \text{H}$	$1.9 \times 10^{-7} (300/T_e)^{0.83}$ for $T_e \leq 1000 \text{ K}$ $7.0 \times 10^{-8} (1000/T_e)^{1.1}$ for $T_e \geq 1000 \text{ K}$
$\text{H}_3\text{O}^+ + e \rightarrow \text{OH} + \text{H}_2$	$1.06 \times 10^{-7} (300/T_e)^{0.83}$ for $T_e \leq 1000 \text{ K}$ $3.9 \times 10^{-8} (1000/T_e)^{1.1}$ for $T_e \geq 1000 \text{ K}$
$\text{H}_3\text{O}^+ + e \rightarrow \text{OH} + \text{H} + \text{H}$	$4.56 \times 10^{-7} (300/T_e)^{0.83}$ for $T_e \leq 1000 \text{ K}$ $1.68 \times 10^{-7} (1000/T_e)^{1.1}$ for $T_e \geq 1000 \text{ K}$
$\text{H}_3\text{O}^+(\text{H}_2\text{O})_n + e \rightarrow$ neutrals	$(0.5 + 2n) \times 10^{-6} (300/T_e)^{0.5}$
$\text{Na}^+(\text{CO}_2) + e \rightarrow$ $\text{N} + \text{CO}_2$	5×10^{-6} at 300 K

^aBranching ratios interpolated at 2500 K from the values measured at collision energies of 0.0, 0.4, 1.0, and 1.5 eV.

^bRecently it has been shown that the branching ratios for O_2^+ DR vary with collision energy (e.g., Peverall, R., Rosen, S., Peterson, J.R., et al., 2001. J. Chem. Phys. 114, 6679–6689) and with vibrational level of the O_2^+ ion (e.g., Petrigiani, A., van der Zande, W.J., Cosby, P.C., et al., 2005. Journal of Chemical Physics 122, 014302). Adapted from Fox, J.L., Sung, K.Y., 2001. Solar activity variations in the Venus ionosphere/thermosphere. J. Geophys. Res. 106, 21305–21335, with additional data from the compilation of Florescu-Mitchell, A.I., Mitchell, J.B.A., 2006. Dissociative recombination. Physics Report 430, 277–374.

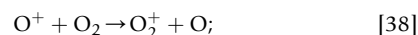
Since atomic ions cannot undergo DR, the only recombination mechanism available to an ion such as O^+ is radiative recombination:



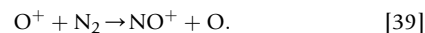
Radiative recombination is the reverse of photoionization, and the spectrum of photons emitted is a continuum with a cutoff at the ionization potential of the product neutral atom. Radiative recombination reactions are very slow, with rate coefficients of the order of $10^{-12} \text{ cm}^3 \text{s}^{-1}$ at 300 K. Because the electron and ion must remain in close proximity long enough for the emission of a photon, such reactions tend to have strong inverse temperature dependences.

Because radiative recombination is slow, it is rarely a major loss process for an atomic ion. In the presence of sufficient neutral densities, atomic ions are transformed into molecular ions via ion–molecule reactions. The molecular ions are then destroyed by DR. For example, O^+ is produced mainly by photoionization and electron-impact ionization of O ; the peak in the production rate of O^+ in the terrestrial ionosphere is near 175 km. In the *E* and *F* regions of the

ionosphere, chemical loss of O^+ is dominated by reaction with O_2



and with N_2



Altitude profiles of the most important sources and chemical sinks of O^+ are shown in Figure 10.

Since the densities of the major neutral species decrease exponentially with altitude, the rates of ion–molecule reactions also decrease with altitude. If these reactions were the only loss process for atomic ions, their densities would increase with altitude indefinitely. At altitudes above that at which the chemical lifetime of the ion, τ_c , is equal to its diffusion time, τ_{Di} , the major loss process for an atomic ion is diffusion. In the terrestrial ionosphere, this effect produces an F_2 peak near 300 km, where the O^+ maximum density of $\sim 10^6 \text{ cm}^{-3}$ is also the absolute maximum for the ionosphere, as shown in Figure 11, where we present ion-density profiles for the *E* and *F* regions of the terrestrial atmosphere from a model.

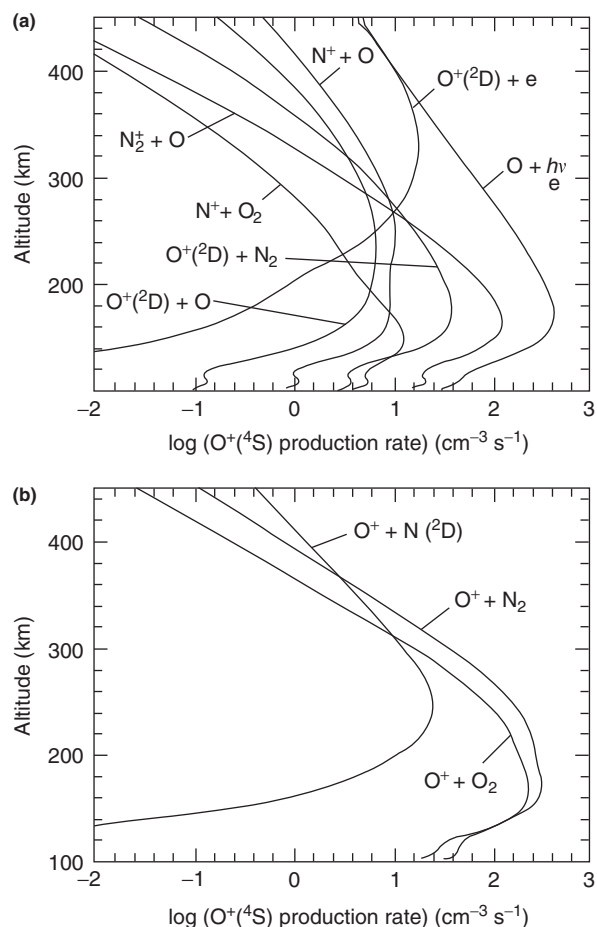


Figure 10 Major production and loss mechanisms for O^+ from a model of the terrestrial ionosphere. From Dalgarno, A., Fox, J.L., 1994. Ion chemistry in atmospheric and astrophysical plasmas. In: Ng, C.Y., Baer, T., Powis, I. (Eds.), Unimolecular and Bimolecular Ion-Molecule Reaction Dynamics. Wiley and Sons, New York, pp. 1–85.

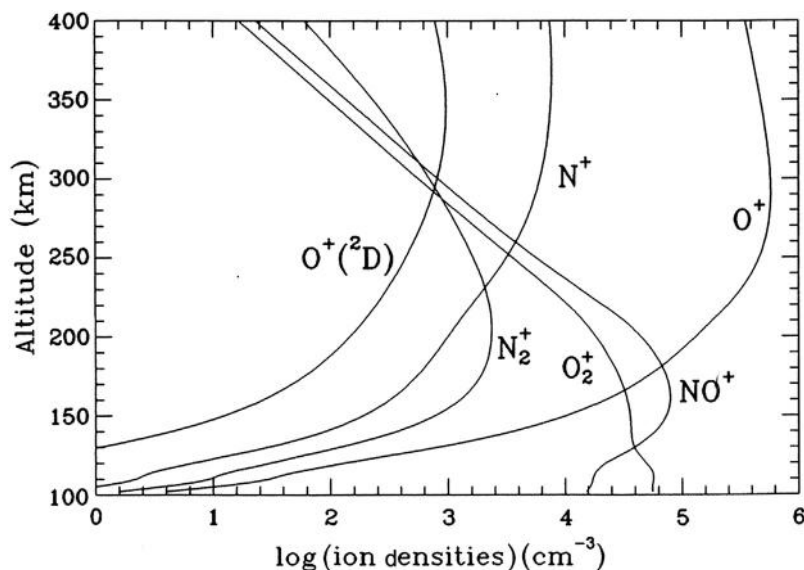


Figure 11 Major ion density profiles for the E , F_1 , and F_2 regions of the terrestrial ionosphere. From Dalgarno, A., Fox, J.L., 1994. Ion chemistry in atmospheric and astrophysical plasmas. In: Ng, C.Y., Baer, T., Powis, I. (Eds.), *Unimolecular and Bimolecular Ion-Molecule Reaction Dynamics*. Wiley and Sons, New York, pp. 1–85.

Such pronounced F_2 peaks do not appear on Mars and Venus for a number of reasons, including smaller mixing ratios of O and O_2 , and smaller neutral scale heights. This combination leads to lower altitude and smaller density peaks in the O^+ profiles, which may not be visible in the electron density profiles. On the outer planets, however, H^+ is predicted to form a prominent F_2 peak, the magnitude and altitude of which is difficult to reconcile with the radio occultation electron density profiles returned from spacecraft.

D-Region Ion Chemistry

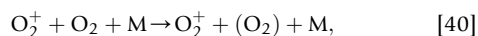
The D region of the terrestrial ionosphere is found in the altitude range of approximately 50–100 km. Photoionization of NO by solar Lyman alpha photons at 1216 Å, and ionization of all neutral species by hard X-rays are supplemented by absorption of cosmic rays, which deposit their energy over a large altitude range. These sources are, however, highly variable. The solar X-ray fluxes depend on the 11-year solar cycle and the presence or absence of flares on the solar disk. The cosmic ray flux is anticorrelated with solar activity. Photoionization of $O_2(a^1\Delta_g)$ has also been suggested as a source of ionization in the D region. $O_2(a^1\Delta_g)$ is a metastable electronically excited state of O_2 , which has an ionization potential of 11.1 eV, and a lifetime against radiation to the ground state of more than 4000 s. It is formed by photodissociation of ozone in the mesosphere. Among FUV photons capable of ionizing $O_2(a^1\Delta_g)$, only those in the limited range 1090–1116 Å and Lyman beta at 1026 Å penetrate below 90 km.

The ions in the D region can be considered to be in PCE. Just as in the E and F_1 regions of the ionosphere, regardless of the species ionized, chemical reactions in the D region tend to transform ions produced, such as N_2^+ for which the parent neutral has a high ionization potential, to the ions O_2^+ and

NO^+ , and to metal ions, for which the parent neutrals have very low ionization potentials. Neutral metal atoms, such as Fe, Mg, Na, and Ca, are produced in the mesosphere by ablation of meteoroids, which enter the atmosphere continuously. The neutral metal atoms form narrow layers at about 90 km that are a few kilometers thick. The metal atoms may react with O and O_3 to form oxides and other reservoir species, and they are ionized by photoionization as well as charge-transfer reactions. At low enough altitudes the metal ions form metal oxide ions or clusters with N_2 , CO_2 , or H_2O . These cluster ions may undergo DR, which releases the neutral metal atoms. The chemistries of each meteoric metal atom and ion, although similar, differ in details; the chemistry of this region of the atmosphere is quite complex. Plane (2002) has presented a detailed overview of the subject.

Figure 12 shows the density profiles of the most important classes of positive ions, including molecular, metal, and cluster ions, between ~80 and ~100 km, as measured by a rocket. At sufficiently low altitudes, where the background neutral densities are large, the terminal ions are not NO^+ or O_2^+ because they cluster efficiently with background molecules, such as N_2 or O_2 . Mass spectrometers carried aboard rockets have shown that below an altitude known as the ‘transformation altitude’, the diatomic positive ions O_2^+ and NO^+ are transformed to large polyatomic cluster ions. The transformation altitude is determined by the temperature, electron densities, and neutral densities. The large measured masses correspond mostly to proton hydrates $H^+(H_2O)_n$ (also expressed as $H_3O^+(H_2O)_{n-1}$) where n is 1, 2, 3, or 4, with $n = 2$, $n = 1$, and $n = 3$ dominating. H_3O^+ , the hydronium ion, has been shown to be a stable species in the gas phase.

The formation of cluster ions may begin with the three-body recombination of O_2^+ with O_2 ,



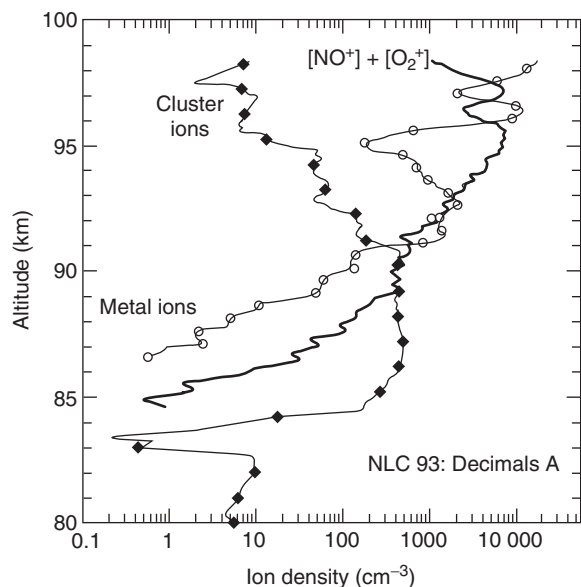
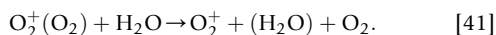
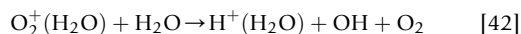


Figure 12 Altitude profiles of the major ion categories: molecular ions, cluster ions, and metal ions. Taken from Kopp, E., Balsiger, F., Murad, E., 1995. Silicon molecular ions in the *D*-region. *Geophysical Research Letters* 22, 3473–3476.

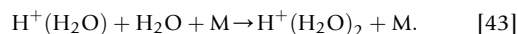
where *M* is an abundant background species. Unless the O_2^+ (O_2) species redissociates, reaction [40] may be followed by the two-body switching reaction to form a cluster with a water molecule, which has a dipole moment:



Since $O_2^+(H_2O)_n$ photodissociates or undergoes collisional dissociation for *n* larger than 1, further hydration of O_2^+ is unlikely. Direct transformation to a proton hydrate

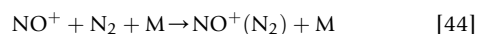


or to a proton hydrate clustered with a hydroxyl radical, $H^+(H_2O)(OH)$, is more probable. The latter species may then undergo switching of the hydroxyl radical with H_2O , forming $H^+(H_2O)_2$. $H^+(H_2O)_2$ may also be formed by a termolecular hydration reaction



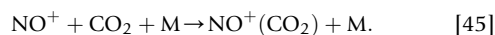
Similar three-body hydration reactions may then continue to occur, producing $H^+(H_2O)_3$, $H^+(H_2O)_4$, and even larger proton hydrates. The terminal cluster ions are determined partly by the proton affinities of the molecules to which they cluster (see Table 3).

The transformation of NO^+ to proton hydrates is less straightforward, and probably does not occur via termolecular reactions with H_2O . Although they are less strongly bound, formation of clusters with more abundant species such as N_2 , in reactions such as

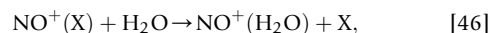


is more likely. Although the density of CO_2 is much smaller than that of N_2 , it is expected to have a larger NO^+ affinity than

N_2 , and to form cluster ions either by a switching reaction with $NO^+(N_2)$ or by three-body association reactions:

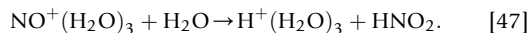


The weakly bound cluster ions $NO^+(N_2)$ and $NO^+(CO_2)$ may then be transformed to the more stable hydrates via switching reactions with water, such as

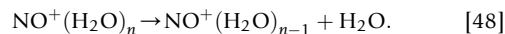


where *X* is either N_2 or CO_2 . Because H_2O has a permanent dipole moment, the NO^+ cluster $NO^+(H_2O)$ is more strongly bound than that with N_2 or CO_2 , but its dissociation energy is still less than 0.8 eV. The NO^+ affinities of neutral species are expected to be correlated with the proton affinities, which are shown in Table 3. NO^+ clusters such as $NO^+(H_2O)(X)$ may be stabilized by the presence of a species *X*, which denotes any of a number of abundant or trace mesospheric species. The most stable clusters are those for which the clustering molecule *X* has a permanent dipole moment. Many of the NO^+ cluster ions are weakly bound, and dissociate collisionally. This tendency leads to a strong temperature dependence for some of the reactions, and of the NO^+ hydration scheme.

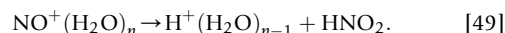
The formation of clusters is much more likely at temperatures near and below 200 K than at room temperature. The equilibrium constants for cluster formation reactions increase by orders of magnitude from 300 to 200 K. The NO^+ water cluster ion, $NO^+(H_2O)_3$, may react with H_2O in a switching reaction to produce a proton hydrate:



Small $NO^+(H_2O)_n$ clusters may undergo unimolecular decay via



For large *n*, the NO^+ clusters are more likely to decay via:



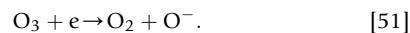
Only the larger proton hydrates are efficiently produced from NO^+ .

A schematic diagram of possible *D*-region positive ion chemistry is shown in Figure 13. Recent model calculations of altitude profiles of the most important positive ions are shown in Figure 14. It should be noted that, in addition to the transformation reactions discussed above, cluster ions are subject to competing loss processes such as photodissociation, collisional dissociation, and DR. Large loosely bound cluster ions are characterized by large DR coefficients, which are of the order of $5 \times 10^{-6} \text{ cm}^3 \text{ s}^{-1}$. A collection of rate coefficients for positive cluster ions involved in *D*-region chemistry has been compiled by Kopp (1996).

The density of free electrons become less than that of negative ions below an altitude of about 70–75 km during the daytime, and about 75–80 km at night. The formation of negative ions begins with the pressure-dependent three-body attachment of an electron to O_2



and to a lesser extent by dissociative attachment



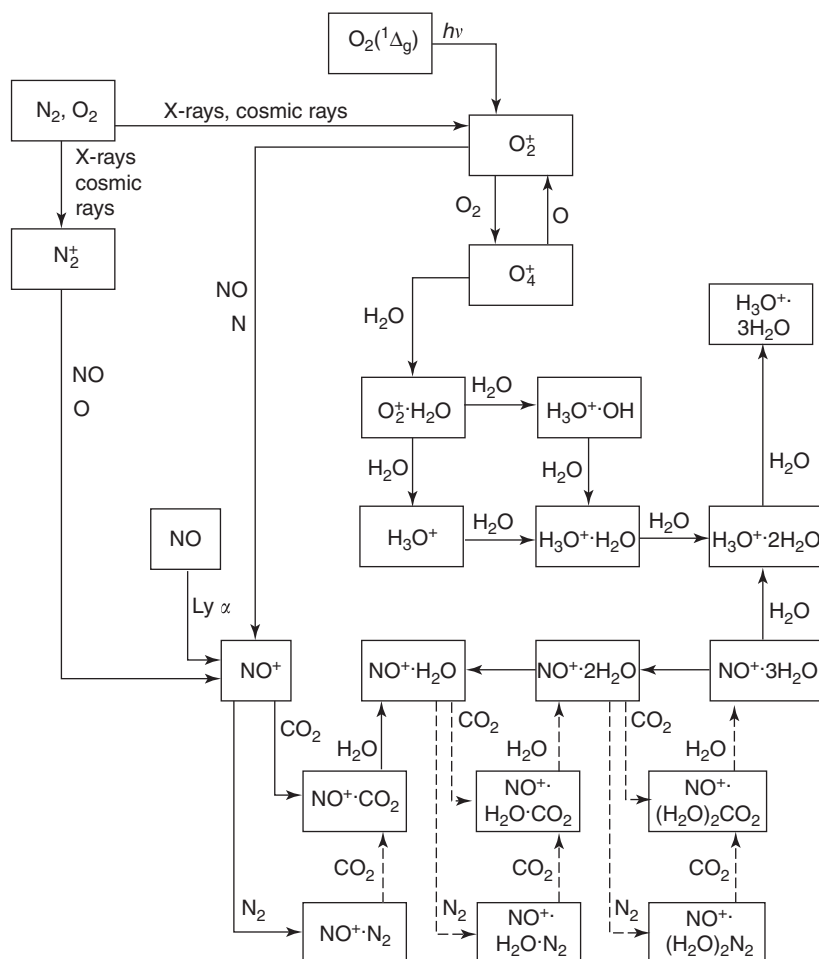


Figure 13 Schematic diagram of positive ion chemistry in the *D* region. Adapted from Ferguson, E.E., Fehsenfeld, F.C., Albritton, D.L., 1979. Ion chemistry in the Earth's atmosphere. In: Bowers, M.T. (Ed.), *Gas Phase Ion Chemistry*, vol. 1. Academic Press, New York.

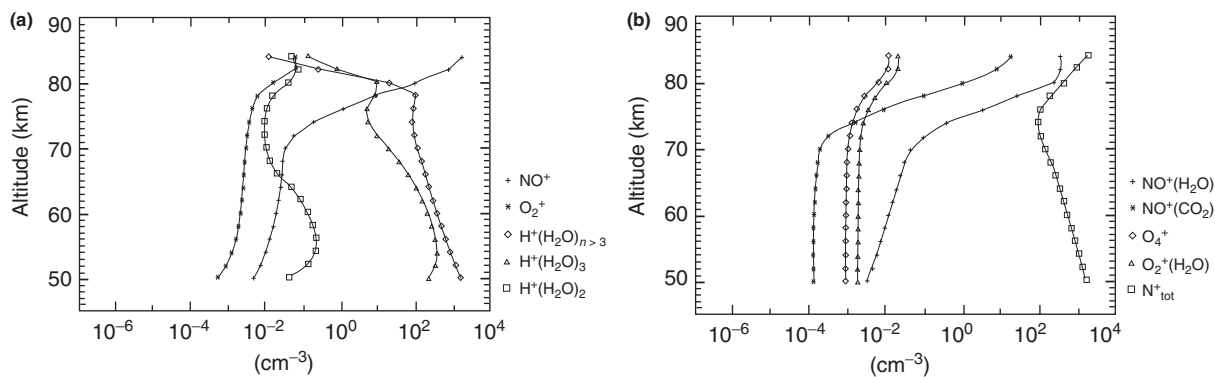


Figure 14 Altitude profiles of selected positive ions from a model appropriate to a latitude of 50°, January and noon. (a) The ions NO^+ , O_2^+ , $\text{H}^+(\text{H}_2\text{O})_2$, $\text{H}^+(\text{H}_2\text{O})_3$, and $\text{H}^+(\text{H}_2\text{O})_{n>3}$ and (b) $\text{NO}^+(\text{H}_2\text{O})$, $\text{NO}^+(\text{CO}_2)$, O_4^+ , $\text{O}_2^+(\text{H}_2\text{O})$, and N_{tot}^+ . Taken from Kull, A., Kopp, E., Granier, C., Brasseur, G., 1997. Ions and electrons of the lower-latitude *D* region. *Journal of Geophysical Research* 102, 9705–9716.

Table 5 Selected electron affinities (eV)

Neutral species	Ion produced	Electro affinity
NO	NO ⁻	~0.026
O ₂	O ₂ ⁻	0.44
H	H ⁻	0.75
CO	CO ⁻	1.33
O	O ⁻	1.46
OH	OH ⁻	1.83
O ₄	O ₄ ⁻	<2
O ₃	O ₃ ⁻	2.1
CO ₃ (H ₂ O)	CO ₃ ⁻ (H ₂ O)	2.1
ClO	ClO ⁻	2.28
NO ₂	NO ₂ ⁻	2.3
HCO ₂	HCO ₂ ⁻	3.17
CO ₃	CO ₃ ⁻	3.26
Cl	Cl ⁻	3.6
HCO ₃	HCO ₃ ⁻	3.67
NO ₃	NO ₃ ⁻	3.9

From Lias, S.G., Bartmess, J.E., Liebman, J.F., Holmes, J.L., Levin, R.D., Mallard, W.G., 1988. Gas phase ion and neutral thermochemistry. *Journal of Physical Chemistry Reference Data* 17 (Suppl. 1) and from NIST Chemistry Webbook (webbook.nist.gov).

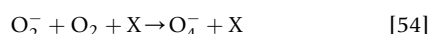
Further transformation of these negative ions competes with photodetachment



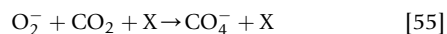
and associative detachment



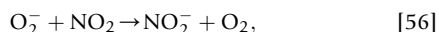
If they are not destroyed, these initial negative ions undergo a series of reactions that produce more stable negative ions, which are those that are formed from parent neutrals that have larger electron affinities. Selected electron affinities are presented in **Table 5**. Transformation reactions include three-body association, in such reactions as



and



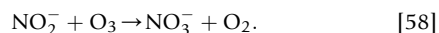
electron-transfer reactions such as



and atom abstraction reactions including

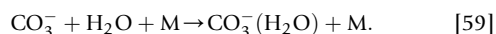


and



These reactions transform the initial negative ions into a plethora of intermediate ions, ending probably in the terminal ions NO₃⁻, CO₃⁻, Cl⁻, and HCO₃⁻, the parent neutrals of which have very large electron affinities.

Termolecular clustering reactions of negative ions may also occur, in such reactions as



Negative ions may also be destroyed by mutual neutralization with positive ions, the rate coefficients for which are of the same

order as DR coefficients, $\sim 10^{-7} \text{ cm}^3 \text{ s}^{-1}$. The terminal ions tend to have longer chemical lifetimes and larger densities than the intermediate ions. Negative ion chemistry is highly dependent on pressure and temperature, and on the densities of minor photochemically produced constituents in the atmosphere, especially O and O₃. Since solar fluxes play a role in electron detachment, there is a strong day–night asymmetry. Compilations of negative ion reactions relevant to the D region have been presented by Kopp (1996) and by Ikezoe et al. (1987).

See also: Chemistry of the Atmosphere: Chemical Kinetics. **Mesosphere:** Atomic Species in the Mesopause Region; Ionosphere; Metal Layers. **Solar System/Sun, Atmospheres, Evolution of Atmospheres:** Planetary Atmospheres: Mars; Planetary Atmospheres: Venus; Solar Terrestrial Interactions: Climate Impact. Thermosphere.

Bibliography

- Anicich, V.G., 1994. Evaluated bimolecular ion-molecule gas phase kinetics of positive ions for use in modeling the chemistry of planetary atmospheres, cometary comae, and interstellar clouds. *Journal of Chemical Physics Reference Data* 22, 1469.
- Anicich, V.G., 2003. An Index of the Literature for Bimolecular Gas Phase Cation-Molecule Reaction Kinetics. JPL Publication, 03–19, NASA.
- Banks, P.M., Kockarts, G., 1977. *Aeronomy*. Academic Press, New York.
- Bauer, S.J., Lammer, H., 2004. *Planetary Aeronomy*. Springer, Berlin.
- Brasseur, G., Solomon, S., 1986. *Aeronomy of the Middle Atmosphere*. D. Reidel, Boston.
- Brown, R., Lebreton, J.P., Waite, J.H. (Eds.), 2009. *Titan from Cassini-Huygens*. Springer-Science + Business Media, Berlin.
- Chapman, S., 1931. The absorption and dissociative or ionizing effects of monochromatic radiation in the atmosphere of a rotating earth. *Proceedings of the Physics Society of London*, 26–45.
- Dalgarno, A., Fox, J.L., 1994. Ion chemistry in atmospheric and astrophysical plasmas. In: Ng, C.Y., Baer, T., Powis, I. (Eds.), *Unimolecular and Bimolecular Ion-Molecule Reaction Dynamics*. Wiley and Sons, New York, pp. 1–85.
- Ferguson, E.E., Fehsenfeld, F.C., Albritton, D.L., 1979. Ion chemistry in the earth's atmosphere. In: Bowers, M.T. (Ed.), *Gas Phase Ion Chemistry*, vol. 1. Academic Press, New York, pp. 45–83.
- Florescu-Mitchell, A.I., Mitchell, J.B.A., 2006. Dissociative recombination. *Physics Report* 430, 277–374.
- Fox, J.L., 2006. *Aeronomy*. In: Drake, G.W.F. (Ed.), *Atomic, Molecular and Optical Physics Handbook*, second ed. American Institute of Physics Press, Woodbury, NY, pp. 1259–1292.
- Fox, J.L., Sung, K.Y., 2001. Solar activity variations in the Venus ionosphere/thermosphere. *Journal of Geophysical Research* 106, 21305–21335.
- Ikezoe, Y., Matsuoka, S., Takebe, M., Viggiano, A., 1987. *Gas-Phase Ion-Molecule Reaction Rate Constants through 1986*. Maruzen Co., Ltd, Tokyo.
- Kopp, E., 1996. Electron and ion densities. In: Dieminger, W., Hartmann, G.K., Leitinger, R. (Eds.), *The Upper Atmosphere: Data Analysis and Interpretation*. Springer, Berlin, pp. 620–630.
- Lias, S.G., Bartmess, J.E., Liebman, J.F., Holmes, J.L., Levin, R.D., Mallard, W.G., 1988. Gas phase ion and neutral thermochemistry. *Journal of Physical Chemistry Reference Data* 17 (Suppl. 1).
- Murad, E., Williams, I.P. (Eds.), 2002. *Meteors in the Earth's Atmosphere*. Cambridge University Press, Cambridge.
- Plane, J.M.C., 2002. Atmospheric chemistry of meteoric metals. *Chemistry Reviews* 103, 4963–4984.
- Rees, M.H., 1989. *Physics and Chemistry of the Upper Atmosphere*. Cambridge University Press, Cambridge.
- Richards, P.G., 2011. Reexamination of ionospheric photochemistry. *Journal of Geophysical Research* 116, A08307. <http://dx.doi.org/10.1029/2011JA016613>.
- Schunk, R.W., Nagy, A.F., 2009. *Ionospheres: Physics, Plasma Physics, and Chemistry*, second ed. Cambridge University Press, Cambridge.
- Su, T., Bowers, M.T., 1979. Classical ion-molecule collision theory. In: Bowers, M.T. (Ed.), *Gas Phase Ion Chemistry*, vol. i. Academic Press, New York, pp. 89–118.
- Wayne, R.P., 2000. *Chemistry of Atmospheres*, third ed. Clarendon Press, Oxford.

Isotopes, Stable

CAM Brenninkmeijer, Max Planck Institute for Chemistry, Mainz, Germany

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

Knowledge about the isotopic composition of a range of atmospheric trace gases has proven useful in better defining their sources and chemical modification, including phase transitions in the case of water vapor. Information on the nomenclature, relevant processes like isotope fractionation is given and the better known applications are reviewed.

Introduction

Chemical reactions, phase transitions, and processes like diffusion induce small but measurable changes in the stable-isotope abundances of the elements in atmospheric gases and in aerosol. These natural variations do not affect chemistry or transport but their investigation deepens our understanding of certain atmospheric processes. The underlying principles are that different sources of trace gases have different isotopic compositions ('signatures,' or 'fingerprints'), and that the isotopic compositions of trace substances are changed by chemical and physical sink processes. Application of isotopic analysis in atmospheric chemistry has revealed a class of new isotope effects, which stimulates fundamental research. Analytical progress in mass spectroscopy allows very small quantities of substances to be analyzed rapidly. Moreover, the precision, which lies at the heart of resolving the intrinsically small isotope effects, is continuously increasing, reducing under special conditions the uncertainties in the isotopic ratio determinations to below 10 ppm. Progress in optical spectroscopy allows orders of magnitudes more measurements to be carried out. Also laboratory research keeps pace with the developments and for an increasing number of atmospherically relevant reactions, isotope effects are being measured and predicted by theoretical models. This facilitates new applications of isotope analysis in atmospheric sciences, while at the same time the links between the atmosphere and the biosphere can be better studied using these isotope tools.

Notation and Nomenclature

With the exception of hydrogen, variations in stable-isotope ratios occurring in nature usually amount to only several percent. Therefore, the isotope abundances of a particular element in a given compound are expressed as per mil (10^{-3}), or sometimes even per meg (10^{-6}) on a relative scale that is defined by one or more standard reference materials. For a given sample (SA), the isotopic value relative to the standard material (ST) is given in parts per thousand (per mil, or ‰). Using carbon-13 as an example with $R = n(^{13}\text{C})/n(^{12}\text{C})$, eqn [1] gives the relative isotopic composition.

$$\delta^{13}\text{C} = (R_{\text{SA}}/R_{\text{ST}} - 1) * 1000 \quad [1]$$

This definition is based on the ratio of the minor to the most abundant isotope. It excludes the use of molecular

ratios, so that, for instance, $\delta^{13}\text{CH}_4$ is not defined and the notation $\delta^{13}\text{C}(\text{CH}_4)$ is used. For the deuterium content of a sample of methane the appropriate equation corresponding to eqn [1] is used. The molecular ratio in the sample, $(\text{CDH}_3)/(\text{CH}_4)$, will be four times larger. Given the low abundance of deuterium, the fraction of methane molecules with two or more deuterium atoms is negligible. The word 'isotopologue' (isotopic analogue) denotes the isotopically substituted molecule, accordingly $^{13}\text{CO}_2$ and $^{14}\text{N}^{15}\text{N}$ are isotopologues of CO_2 and N_2 . The word 'isotopomer' (isotopic isomer) is used to indicate that isotopic substitution is at another site in the molecule, for instance for nitrous oxide we have the two isotopologues $^{15}\text{N}^{14}\text{NO}$ and $^{14}\text{N}^{15}\text{NO}$.

For any two reference materials A and B, the conversion for a sample X is made according to eqn [2].

$$1 + \delta_{\text{A}}(\text{X}) = (1 + \delta_{\text{B}}(\text{X})) / (1 + \delta_{\text{B}}(\text{A})) \quad [2]$$

For mixing a molar fraction f having a delta value δ_{a} into a reservoir with a background value of δ_{b} , the resulting isotopic composition will be given in good approximation by eqn [3].

$$\delta_{\text{r}} = f\delta_{\text{a}} + (1 - f)\delta_{\text{b}} \quad [3]$$

For convenience, Table 1 gives the isotopic abundances of the elements most often analyzed in atmospheric compounds. Reference materials for reporting isotopic composition are available from the International Atomic Energy Agency (IAEA). The International Union of Pure and Applied Chemistry have issued in 2011 a modified periodic table of the elements. This new table gives for a number of elements (H, Li, Be, B, C, N, O, Si, S, and Cl), instead of the standard atomic weights of these

Table 1 Minor isotope abundances

Element	Minor isotopes	Abundance (%)	Main reference materials
H	^2H or D	0.15	V-SMOW ^a
C	^{13}C	1.11	V-PDB ^b
N	^{15}N	0.37	Atmospheric N_2
O	^{17}O	0.037	V-SMOW
	^{18}O	0.20	
S	^{33}S	0.76	V-CDT ^c
	^{34}S	4.2	
	^{36}S	0.014	
Cl	^{37}Cl	24.5	SMOC ^d

^aVienna-Standard Mean Ocean Water.

^bVienna Pee Dee Belemnite (carbonate).

^cVienna Canyon Diablo Troilite (meteorite).

^dStandard Mean Ocean Chloride.

elements, ranges of values. For instance, for sulfur in natural terrestrial compounds the standard atomic weight lies between 32.05 and 32.08. The differences in stable-isotope composition are due to isotopic fractionation, mainly caused by the role of mass differences in a variety of chemical and physical processes. When isotopic fractionation occurs, the fractionation factor α is defined as the isotopic ratio in one compound or phase relative to that of the other. No strict convention is in operation. For instance, water vapor is often assayed relative to the liquid, and $\alpha < 1$ (i.e., depletion). The accompanying fractionation constant ϵ is defined as $\alpha - 1$. For chemical reactions kinetic isotope effect (KIE) is often used instead of the conventional α . For KIE, the fractionation factor is the ratio between the reaction rate constants for the minor and major isotopomers, or vice versa.

Measurement

Measurement of isotopic ratios for the elements mentioned above is based on mainly two principles. One is mass spectrometry. This technique can deal with smallest amounts, down to subnanomolar quantities, which is important as we deal with atmospheric trace gases. Also the highest precision is achieved by mass spectrometry. The analytical principle is mostly electron impact ionization, ion acceleration, magnetic separation of the isotopic masses followed by ion current measurement using Faraday collectors. Conversion of measured ratios to the appropriate reference material is basically done using eqn [2], after corrections, especially ion current corrections, have been made. For instance for CO_2 , mass 45 represents a combination of the isotopologues $^{13}\text{C}^{16}\text{O}^{16}\text{O}$ (about 1%) and $^{12}\text{C}^{17}\text{O}^{16}\text{O}$ (twice about 0.04% because of the two O atoms). To make the correction, one needs information about the $^{17}\text{O}/^{16}\text{O}$ ratio, which is mostly estimated from the respective $^{18}\text{O}/^{16}\text{O}$ ratio (the ratio of mass 46 ion beam current to that of mass 44). Aerosols or trace gases can be oxidized to give CO_2 , which is purified by gas chromatography and injected in a flow of helium into the mass spectrometer. In case very high precisions are required, the gas to be analyzed, for instance CO_2 or SF_6 in case of sulfur isotope analysis, has to be introduced in micromole amounts using vacuum transfer followed by repeated comparisons with a reference gas.

The other measurement principle is based on optical absorption spectroscopy. When using high-resolution spectroscopy, the differences in optical spectra of isotopologues and isotopomers can be resolved. The challenge of the extremely low abundance – for instance N_2^{18}O is only 0.2% of 380 ppbv (N_2O mixing ratio) – is dealt with by using very long absorption path lengths and the cavity ring down method. These optical techniques are extremely powerful for atmospheric applications as they can provide real time, continuous *in situ* data. Also sample pretreatment is in many cases not required at all. Development to increase the precision of these optical techniques is ongoing.

Kinetic and Equilibrium Isotope Effects

A distinction is made between kinetic and equilibrium isotope effects. Using water vapor in contact with water in an enclosed

volume, the depletion of the vapor in D and ^{18}O can be calculated on the basis of the vapor pressure differences. When water evaporates through the stomata in plant leaves into dry air, the additional kinetic effect caused by the diffusion speed differences of H_2^{18}O and H_2^{16}O in air increases the overall fractionation strongly. When CO_2 is in oxygen isotopic equilibrium with H_2O for which carbonate and bicarbonate formation is necessary, its ^{18}O content at 25 °C is 41‰ above that of the water. This effect can be calculated like most other (thermodynamic) equilibrium effects, using statistical mechanics and the partition function.

Kinetic isotope effects are common in chemical reactions. The heavier isotopes have a lower zero point energy which leads to stronger intramolecular bonding. Generally, heavier isotopologues molecules react more slowly. Calculating kinetic isotope effects is more complex and transition state theories with a range of refinements, including tunneling effects, are often used. For the atmospherically relevant reaction between chlorine radicals (Cl) and CH_4 , calculations are based on the intermediate complex Cl-HCH_3 . The theory and experimental results do not always conform. The kinetic isotope effects for all D-based carbon-12 isotopomers of methane are shown in Table 2. These are the largest effects known, although in atmospheric applications only reactions with CH_4 and CDH_3 play a role. The heavier isotopologues react slower, which is a normal isotope effect. A notable exception where the substitution with a heavier isotope does not reduce, but increases reaction speed is the well-known reaction $\text{CO} + \text{OH}$, in which C^{18}O reacts faster than C^{16}O . The various isotope fractionation processes of relevance to atmospheric processes are shown in Table 3. For the biogeochemical cycles on N_2O , nitrates, CH_4 , H_2 , and other atmospheric trace compounds, not only gas-phase reactions play a role, but also enzymatic processes.

Table 2 Systematics of reaction rate constants of methane isotopomers

	<i>OH</i>	<i>Cl</i>
CH_4	6.4 ^a	100
CDH_3	5.2	74
CD_2H_2	3.4	46
CD_3H	2.1	23
CD_4	0.9	6

^aAll given at 298 K in $10^{-15} \text{ cm}^3 \text{ molecules}^{-1} \text{ s}^{-1}$.

Table 3 Fractionation processes

Process	Example
Phase change	Evaporation of H_2O
Diffusion	CO_2 in leaf stomata
Dissolution	CO_2 in ocean water
Gravitational settling	Firn and ice core air studies
Atmospheric escape	H_2
Chemical reaction	$\text{CO} + \text{OH}$
Photolysis	N_2O , HCHO
Biological/enzymatic processes	Nitrification/denitrification

Mass-Independent Fractionation

Stable-isotope studies dealing with atmospheric trace gases have revealed one or more types of fractionation processes that are not completely understood and have been pooled under the name ‘mass-independent fractionation’ (MIF). The rationale for this name was that such an effect, a very large one, was first observed for ozone (O_3), and became attributed to molecular symmetry effects, totally independent of mass. Retrospectively, the name is an unfortunate choice and is at times used along with the expression ‘anomalous fractionation.’ We emphasize that the effect was discovered using oxygen isotopes. There was a strong discrepancy measured in the abundance of ^{17}O in ozone relative to that expected on the basis of a simple relationship between ^{17}O and ^{18}O fractionation. Thus by virtue of oxygen having three stable isotopes, the effect became so clearly detectable. Figure 1 shows the relationship between $\delta^{17}O$ and $\delta^{18}O$ for a wide range of substances. The slope of the mass-dependent fractionation line thus defined is close to 0.52. The theoretical basis for this correlation for a range of physical and chemical processes is well understood. For instance, the equilibration constant for an isotope exchange reaction is proportional to the vibrational frequencies, and therefore the reciprocal masses. For diatomic molecules for O isotopes (^{18}O , ^{17}O , and ^{16}O) one obtains simply $(1/16 - 1/17)/(1/16 - 1/18) = 0.53$. In fact, because of this omnipresent strict mass dependence, almost no attention was paid to the analysis of ^{17}O in oxygen-containing compounds for several decades. Atmospheric research has

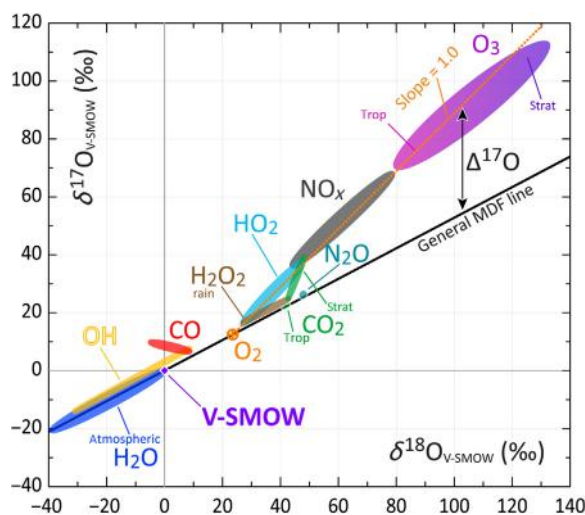


Figure 1 The $\delta^{18}O$ and corresponding $\delta^{17}O$ values for some important reservoirs showing mass-independent fractionation effects as deviations from the ‘mass dependent’ fractionation line of slope 0.52. This slope is however not exactly the same as it depends on the fractionation processes involved, nearly all of which are strictly mass dependent. Note that molecular oxygen has a small negative value $\Delta^{17}O = -0.15\text{‰}$, not visible on this scale. All mass-independent effects shown are due to the anomaly in ozone, with exception of CO which obtains most of its MIF from the reaction $CO + OH$ and a smaller part from the formation of CO by the reaction of ozone with unsaturated hydrocarbons such as isoprene.

changed this and MIF has been detected for O_3 , CO_2 , CO , N_2O , H_2O_2 , SO_2 , and even atmospheric molecular oxygen discussed later. Nearly all of these effects can be traced back to the large anomaly in ozone. A measure for MIF is the deviation in ^{17}O content, in approximation expressed as $\Delta^{17}O = \delta^{17}O - 0.52 \delta^{18}O$.

Atmospheric Applications

General

Isotope effects do not significantly affect atmospheric chemistry. Not only are fractionation effects generally small, but also the abundances of the minor isotopes are low. Even though the isotope effects for deuterium may be large because of the large relative mass difference, its low abundance (Table 1) excludes effects on chemistry. For chlorine, the less-abundant isotope is still rather abundant, but in this case, fractionation effects of only a few percent do not impact on chemistry. The corollary is that isotope effects may give useful information but have no impact. Given a certain atmospheric trace gas, isotopic variations are due to two factors. One is the isotopic composition of its sources. In principle, the relative source strengths can be calculated using eqn [3], provided there are only two isotopically distinct sources. The other factor, which disturbs this simple source-based picture, is isotope fractionation in the removal of the gas, be it deposition or atmospheric chemistry. For an isolated, well-mixed amount of gas being depleted by a loss process (e.g., reaction with OH or photolysis), the isotopic composition will evolve in time according to a Rayleigh distillation process. Using m/m_0 as the ratio of the actual to the initial mixing ratio, eqn [4a] or [4b] apply.

$$\delta + 1 = (m/m_0)^{(1-1/\alpha)} \quad [4a]$$

$$\delta \approx (1 - 1/\alpha) \ln (m/m_0) \quad [4b]$$

Ozone, O_3

In situ isotopic measurements of stratospheric ozone were obtained using balloon-borne mass spectrometry. Moreover cryogenic air sample collection was used. Ozone is formed via $O(^3P) + O_2 + M \rightarrow O_3 + M$ and it has a remarkable isotopic composition throughout the atmosphere. It is considerably enriched in ^{18}O relative to its immediate precursor molecular oxygen by 70–110‰ and surprisingly, by almost the same degree in ^{17}O . No other substance on Earth has found to have such high and anomalous oxygen isotope enrichment. As we will explain later, MIF in ozone affects the isotopic composition of a host of other trace gases that are linked to the ozone cycle. The causes for the oxygen isotope anomaly in ozone are not fully understood. Theoretical work by R.A. Marcus and coworkers published in 2011 deals for instance with Coriolis coupling as a source of non-RRKM effects in the ozone molecule. Also laboratory studies are in progress. Part of the ozone enrichment is linked to the reaction constants for the formation of the various ozone isotopomers and isomers. Table 4, based

Table 4 Relative reaction rates for the formation of ozone from $^{16}\text{O}(\text{O})$, $^{17}\text{O}(\text{P})$, and $^{18}\text{O}(\text{Q})$

Reaction	Relative rate
O + OO	1
P + OO	1.03
Q + OO	0.93
O + PP	1.23
P + PP	1.02
Q + PP	1.03
O + QQ	1.53
P + QQ	1.31
Q + QQ	1.03

on laboratory experiments, shows the surprising fact that isotopic substitution leads to enormous differences in reaction rate constants.

The analysis of the isotopic composition of stratospheric, and particularly tropospheric ozone is extremely difficult, and few data are published. Most work concentrate therefore on laboratory studies, on the formation of ozone, its behavior on surfaces, temperature and pressure dependence, and so on. An interesting venue for atmospheric measurement is however available, namely the reaction of ozone with a compound that can be subsequently analyzed. One can think of silver or a sulfide as reactant. Interestingly, such a technique gives the $^{17}\text{O}/^{16}\text{O}$ and $^{18}\text{O}/^{16}\text{O}$ ratio for the terminal oxygen atom in ozone. There are indications that the enrichment at this site is higher than for the central oxygen atom position.

Carbon Dioxide, CO_2

Troposphere

Quantifying fluxes of atmospheric CO_2 between the Earth's carbon reservoirs is a major task and of great importance. Therefore, isotopic analysis of CO_2 has continuously evolved since the first applications by Keeling and Craig in the 1950s. Currently, ^{13}C and ^{18}O analyses are carried out on CO_2 extracted from air sampled in flasks of the NOAA-ESRL network. The combustion of fossil fuel and deforestation have lead to a decrease in $\delta^{13}\text{C}$ from a preindustrial value (relative to V-PDB) of about -6 to about -8‰ in 2010. Photosynthesis favors the uptake of $^{12}\text{CO}_2$ over that of $^{13}\text{CO}_2$, by which plant organic matter is depleted. The ^{13}C value of fossil fuel carbon is close to the values of the original vegetation and aquatic organisms millions of years ago, and is around -27‰ . The ^{13}C equilibrium between ocean bicarbonate and atmospheric CO_2 has kept the values for the biota at nearly the same level, but the sudden release of enormous amounts of fossil fuel-derived carbon disturb the equilibrium. One application of ^{13}C is to distinguish between CO_2 uptake by the land biota and by the oceans. The overall activity of the biosphere, which is dominated by the Northern Hemisphere biota, induces a seasonal cycle of up to 1‰ at the surface at mid to high latitudes. The seasonal minimum is in spring when CO_2 emissions from respiration, oxidation, and fossil fuel burning have peaked. A useful application of $\delta^{13}\text{C}(\text{CO}_2)$ has been the identification of

a large Northern Hemispheric terrestrial carbon sink in 1992 and 1993. When Charles Keeling, who pioneered high precision CO_2 concentration measurements, plotted $\delta^{13}\text{C}$ against the inverse of the CO_2 concentration of air samples, basically applying eqn [3], what he established is known as the Keeling plot. In the plot, the extrapolation to infinite CO_2 concentrations provides the ^{13}C isotopic composition of the admixing source.

In contrast to ^{13}C , $\delta^{18}\text{O}(\text{CO}_2)$ is determined by the isotopic exchange of CO_2 with the main water reservoirs on Earth. The effect for ^{13}C being a mere source-sink effect, for ^{18}O the gross fluxes of CO_2 into and from plant leaves dominate the isotopic composition. The exchange of ^{18}O in cloud droplets is in comparison negligible. Exchange with water vapor is excluded as the formation of carbonate or bicarbonate is necessary. However, exchange with soil water, the mentioned leaf water, and ocean water is important. In this, CO_2 strives toward isotopic equilibrium accompanied by a temperature-dependent enrichment of about 40‰ . Accordingly, the $\delta^{18}\text{O}$ value of CO_2 is close to 40‰ (Vienna-Standard Mean Ocean Water (V-SMOW)). Note that the standard V-SMOW- CO_2 (CO_2 in isotopic equilibrium with V-SMOW at 25 °C) is often used, thus giving a value of near 0. Because leaf water is enriched owing to the equilibrium and kinetic isotope effects accompanying stomatal evaporation, CO_2 in equilibrium with leaf water is more strongly enriched. The ^{18}O isotope effect of emissions of CO_2 from fossil fuel combustion has only a local influence in relative close proximity to the sources. Internal combustion engines produce CO_2 with ^{18}O values close to that of atmospheric CO_2 . The combustion of coal gives values closer to 0, whereas burning gives a range of values in between. When CO_2 from two sources is mixed, each with a given $^{18}\text{O}/^{16}\text{O}$ ratio, one can calculate the resulting (very low) abundance of $\text{C}^{18}\text{O}^{18}\text{O}$. This calculated ratio does not necessarily correspond with the real $\text{C}^{18}\text{O}^{18}\text{O}$ ratio, because it would require scrambling of ^{18}O between the CO_2 molecules, which does not occur. The deviations from the statistically expected random distribution of isotopomers can be measured by carrying out what is named 'clumped isotope analysis.' By taking CO_2 and heating it to high temperatures in a quartz vial, the distribution of ^{13}C , ^{17}O , and ^{18}O is randomized, and this situation can be frozen by rapid cooling. By comparing the relative abundance of $^{13}\text{C}^{16}\text{O}^{18}\text{O}$ of a CO_2 sample to that of heat-treated CO_2 with great precision using the ion beam of the rare mass 47, deviations have been measured. There are, up to date, few atmospheric applications mainly due to the difficulty of such measurements.

Stratosphere

The seasonal variations of $\delta^{13}\text{C}$ and of CO_2 mixing ratios propagate with delay while being attenuated into the stratosphere. This allows accurate stratospheric measurements to be used as a timer. The most important aspect from the isotope point of view is however that CO_2 can interact with ozone which has this anomalous isotopic composition. The reactions involved are $\text{O}_3 + h\nu \rightarrow \text{O}_2 + \text{O}(^1\text{D})$. The excited oxygen atom can, if not quenched before, form a short-lived complex

CO_3 which enables isotopic exchange, according to $\text{O}(^{17}\text{O}) + \text{CO}_2 \rightarrow (\text{CO}_3)^* \rightarrow \text{O}(^{17}\text{O}) + \text{CO}_2$. In the stratosphere $\Delta^{17}\text{O}(\text{CO}_2)$ can reach values of 17‰ . By this $\Delta^{17}\text{O}$ is a tracer for CO_2 of stratospheric origin. In the troposphere this signal is diluted by CO_2 that has no anomalous ^{17}O content. It is speculated that measurement of $\Delta^{17}\text{O}$ can help to determine the rate of oxygen isotope exchange of CO_2 with the hydrosphere.

Water, H_2O

Troposphere

Because of the many roles of atmospheric water through latent heat, cloud formation and albedo, heterogeneous chemistry, and the radiation budget, various efforts are made to use isotopic analysis. The IAEA coordinates the global monitoring of the isotopic composition of precipitation for hydrological applications (isotope hydrology). There are large seasonal and geographical differences. Water evaporating from the oceans is depleted, but the condensing water is again enriched relative to the vapor. With increasing distance into the continents, D and ^{18}O decrease (distance effect) following the mentioned Rayleigh fractionation (eqn [4b]). With increasing latitude and decreasing temperatures, continued precipitation causes further depletion. In Antarctica, δD can be as low as -400‰ and $\delta^{18}\text{O}$ reaches -50‰ . Isotopic analysis of ice cores is one of the main tools in paleoclimatology. Global precipitation D and ^{18}O isotope values define the 'meteorological water line' $\delta\text{D} = 8\delta^{18}\text{O} + 10\text{‰}$. Hailstones have been assayed layer by layer for study of their formation process. Increasing attention is being paid to tropospheric water vapor analysis. Not only the worsening situation concerning water supplies, but also climate change research drives large efforts to obtain and better use isotopic information of atmospheric water. Isotopic values are included in certain models that incorporate the hydrological cycle. The advent of the *in situ* optical determination of δD and $\delta^{18}\text{O}$ in water vapor heralds great progress for these important applications.

In the section [Mass Independent Fractionation](#), the general mass-dependent line of slope 0.52 was mentioned. This slope varies slightly depending on the molecules and fractionation processes that are involved. B. Luz and E. Barkan have by means of extremely precise measurements established the Global Meteoric Water Line, which is given by $\ln(\delta^{17}\text{O} + 1) = 0.528(\delta^{18}\text{O} + 1) + 0.000033$ ($R^2 = 0.99999$). The meteoric waters have a small excess of ^{17}O relative to ocean water. Also diffusion of water vapor into plant leaves will lead to deviations that can be used to study certain processes.

Stratosphere

In the stratosphere, gas-phase chemical interactions affect the isotopic composition. Through cooling and concomitant condensation, water vapor reaching the tropopause attains values of $\delta\text{D} = -670\text{‰}$ and $\delta^{18}\text{O} = -82\text{‰}$. During transport into the middle atmosphere, isotope exchange and addition of H_2O from methane oxidation take place. Exchange of oxygen is via the HO family of reactions involving oxygen and ozone.

These processes enrich the isotopically depleted vapor imported from the troposphere. Above 40 km, δD reaches about -400‰ and stays constant at that level. $\delta^{18}\text{O}$ increases by as much as 100‰ . There is a clear deviation from the meteoric water line. Furthermore, stratospheric H_2O obtains MIF largely from O via $\text{H} + \text{O}_3 \rightarrow \text{OH} + \text{O}_2$ and the subsequent hydrogenation of OH. The oxygen isotope exchange between OH and H_2O has a rate constant of $6 \times 10^{-17} \text{ cm}^3 \text{ s}^{-1}$ at 250 K and plays no significant role. $\delta^{17}\text{O}$ increases from 0 at the tropopause to 10‰ at 40 km.

Methane, CH_4

Atmospheric methane with its many categories of sources forms a classical example of stable-isotope applications. Its isotopic composition, i.e., δD and $\delta^{13}\text{C}$, and that of its sources are shown in [Figure 2](#). The depletion in ^{13}C is largely the result of the fractionation in the formation of CH_4 by bacterial processes as well during digestion in ruminant, as in rice paddies and wetlands. Deuterium shows a similar pattern, but has a large range. The difference between the input from the combined sources and the actual atmospheric composition is caused by the isotopic fractionation in the removal reaction $\text{CH}_4 + \text{OH}$. The large effect for deuterium is apparent. Also, the small soil sink isotopically enriches the methane that is left in the atmosphere. Despite the considerable spread in the source values, inverse modeling using the isotopic composition has helped to further constrain the methane budget. Both $\delta^{13}\text{C}$ and δD show seasonal cycles with amplitudes of approximately 0.2 and 4‰ respectively, depending on latitude which can be resolved with difficulty. There are presently few data on deuterium available, but the advances in isotope mass spectrometry that have been mentioned change this. Interannual variations in the Southern Hemisphere have been associated with biomass burning. Moreover, in the Southern Hemisphere

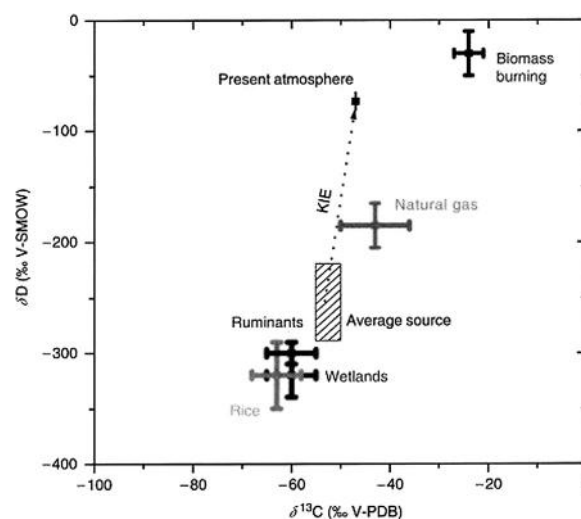


Figure 2 Overview of the ^{13}C and deuterium (^2H) isotope ratios of the main methane sources. The arrow indicates the kinetic isotope effect responsible for the differences between the isotopic composition of the combined sources and that of the atmospheric inventory.

Table 5 Kinetic isotope fractionation factors (KIE) for reaction with CH₄ at 296 K

	¹³ C	D
OH	1.004 ^a	1.294 ^b
Cl	1.065	1.50
O(¹ D)	1.013	1.06

^aValues given for k_{12}/k_{13} .^bValues given for k_H/k_D .

small increases in $\delta^{13}\text{C}$ have been detected that were attributed to the reaction of Cl + CH₄. As Table 5 shows, this reaction has a characteristic large ¹³C kinetic isotope effect. The strong increase in atmospheric CH₄ has induced an isotopic disequilibrium in which the isotopic composition came closer to that of the averaged source. Analysis of CH₄ extracted from firm air reveals a clear minimum δD value about 75 years ago. After a weakening in the increase of CH₄ over the last decade, the equilibrium was approached more closely and δD increased again. Stratospheric chemistry produces large isotope changes in CH₄ because of the greater role of Cl and O(¹D) as sinks. Table 5 shows that O(¹D) has a large isotope effect, which was not expected on the basis of the rapid kinetics of this reaction. The isotopic composition of CH₄ in the stratosphere can be successfully modeled using a 2D model, incorporating the fractionation factors from Table 5. Such results are further used for calculating δD of H₂O and $\delta^{13}\text{C}$ of CO.

Hydrogen, H₂

Scientific interest in the D/H ratio of molecular hydrogen is increased due to the possible large-scale future use of H₂ as an energy carrier to replace fossil fuels. Releases of H₂ into the atmosphere could disturb the OH-based oxidative cycle and also lead to the production of additional H₂O in the stratosphere. Given that the distribution of hydroxyl, OH, and the KIE of H₂ + OH are well known, the uptake of H₂ by soils, which introduces much less fractionation, can be assessed using δD measurements. The Northern Hemisphere has lower H₂ mixing ratio, the Southern Hemisphere has higher $\delta\text{D}(\text{H}_2)$ values. With OH not showing major N–S differences, the effect of the soil sink with much smaller fractionation is evident. Few laboratories embarked on such measurements because of the elaborate procedure of separating H₂ from air. Nevertheless, another issue is that across the tropopause H₂ mixing ratios changes little. The reason is production of H₂ from CH₄ while in contrast to the troposphere the (large) soil sink is missing. In all these, the D/H ratio of H₂ from the photolysis of HCHO has undergone close scrutiny. It is clear that D/H analyses considerably improve H₂ budget estimates and modeling of D has been fully developed. A difficult issue is the D/H ratio of HCHO and that of H₂ from its photolysis.

Hydrocarbons, Formaldehyde, Methyl Chloride, and Other Trace Gases

The recent introduction of coupled gas chromatography isotope mass spectrometry has allowed the analysis of nanomole

amounts of substances, bringing the low-abundance trace gases within reach of isotopic analysis. Values obtained for ethane, ethene, and propene in background air in New Zealand yielded $\delta^{13}\text{C}$ values of -22 to -29‰ , which is typical for organic matter. Methyl chloride was strongly depleted at -43.5‰ . Samples collected south of Japan, between 5 and 35° N, showed the following $\delta^{13}\text{C}$ values: ethane, propane, *n*-butane, *n*-pentane, *i*-butane, and *i*-pentane between 30 and -23‰ ; ethylene and propylene between about -30 and -10‰ ; acetylene between -20 and -2‰ ; and methyl chloride between -40 and -30‰ . There is only one major study on formaldehyde, HCHO, and it deals with urban air. Bearing in mind that even formaldehyde concentration measurements are rare, measuring its isotopic composition is a daunting task indeed. Over a wide range of concentrations $\delta^{13}\text{C}$ values were close to that of organic matter, namely -28‰ . In remote air one could expect HCHO derived from CH₄ oxidation to lead to lower δ values. At the same time δD values ranged from -300 to $+200\text{‰}$ and were weakly anticorrelated with concentration. If considered worthwhile, years of work are needed to make more HCHO isotope measurements. An extreme application was the detection of a 20‰ enrichment of ³⁷Cl in CF₂Cl₂ at stratospheric altitudes where its mixing ratio had dropped from about 550 to 100 pmol mol⁻¹ due to photolysis.

Carbon Monoxide, CO

The shorter lifetime of CO results in considerable concentration and isotopic changes (Figure 3). Concerning the source signatures, CO from high-temperature combustion processes adopts the $\delta^{18}\text{O}$ value of atmospheric oxygen (about $+23.5\text{‰}$) without much fractionation. This gives a useful clear signal for the technological source. Also distinctive is the low $\delta^{13}\text{C}$ value of CO derived from CH₄ oxidation. This value is believed to be about -51‰ , composed of the -47‰ of CH₄ further lowered by the kinetic fractionation of -4‰ in the reaction CH₄ + OH (Table 5). There are problems in closing the isotope budget for ¹³CO in the Southern Hemisphere using this information, because atmospheric $\delta^{13}\text{C}(\text{CO})$ values rarely fall below -32‰ . One has to assume that additional isotope fractionation occurs in the formation of CO from the photolysis of formaldehyde. The yield of CO from CH₄ + OH may be different throughout the troposphere, but low values are not likely. For ¹³C there is a pressure effect. Its KIE maximizes at 5‰ at 1013 hPa, and turns into a negative effect at altitudes above a pressure of 300 hPa. The annual average $\delta^{13}\text{C}$ value in the Southern Hemisphere is approximately -29‰ ; in the midlatitude Northern Hemisphere it is -27‰ . The annual cycles are dominated by the CH₄ source effect and not by the KIE of CO + OH. For ¹⁸O there is a negative isotope effect of nearly -10‰ , almost independent of pressure. Generally, the farther away from CO sources, the more negative the $\delta^{18}\text{O}$ becomes, reaching -10‰ in the lowermost stratosphere. There are only few stratospheric data for CO, but air samples collected during ozone hole conditions in the Antarctic lowermost stratosphere yielded CO with $\delta^{13}\text{C}$ as low as -43‰ . Modeling confirms that CO from CH₄ + Cl causes such anomalous low values.

Another application of the large KIE in CH₄ + Cl is the estimation of free chlorine during Arctic low ozone events at

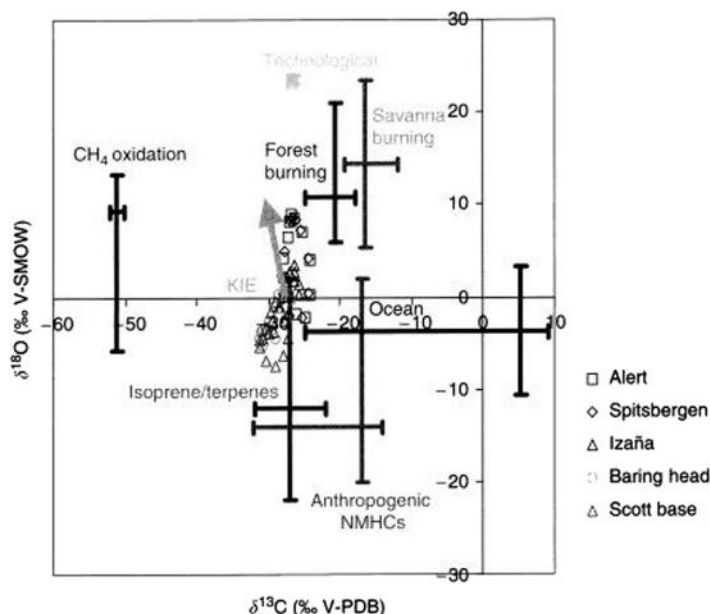


Figure 3 Overview of the ^{13}C and ^{18}O isotope ratios of the main carbon monoxide sources. The arrow indicates the kinetic isotope effect. Observational values for some Northern Hemisphere locations are shown. The spread in the data is mostly due to the seasonal cycle.

the surface. Small downward excursions in $\delta^{13}\text{C}(\text{CO})$ were observed during such events proving the presence of Cl atoms.

After MIF had been detected for O_3 and CO_2 , it also was discovered in CO. MIF in CO has two causes. One is that unsaturated organic compounds react with O_3 . CO from these reactions is a small source of CO enriched in ^{17}O and ^{18}O and exhibiting MIF derived from O_3 . However, the main cause lies in the important reaction $\text{CO} + \text{OH} \rightarrow \text{CO}_2 + \text{H}$. Assuming for the moment that C^{18}O reacts 10‰ faster than C^{16}O , C^{17}O should react about 5‰ faster if the normal mass-dependent fractionation applies. In contrast, C^{17}O reacts nearly as rapidly as C^{16}O . This then results in an excess of ^{17}O in atmospheric CO. There is no complete theoretical explanation for the cause of MIF in $\text{CO} + \text{OH}$, but work by R.C. Markus and coworkers is in progress. Because all major CO sources are mass dependent, $\delta^{17}\text{O}(\text{CO})$ is a unique signal indicating the ‘aging’ of CO by reaction with OH. Atmospheric chemistry transport models have been developed that incorporate the isotope effects of CO and its precursors.

Nitrous Oxide, N_2O

Despite its long lifetime and concomitant small isotopic and abundance variability, the isotopic composition of N_2O is being studied intensively. For this greenhouse gas with a variety of sources closely linked to human activities and having a long lifetime, isotope information is valuable. The $\delta^{15}\text{N}$ and $\delta^{18}\text{O}$ values of tropospheric N_2O show little scatter around approximately 7 and 21‰ relative to atmospheric N_2 and O_2 , respectively. The main sources of N_2O are based on microbial nitrification and denitrification in soils and the ocean, and are generally depleted relative to the atmosphere. The range of δ values is considerable. Soil gases form the

most depleted source type, with averages of about -13 and 10 ‰, respectively. A source identified with enrichment was N_2O from denitrification in upwelling deep water in the Arabian Sea. Enriched N_2O was also found in the Pacific. These sources do not account for the atmospheric enrichment. The enrichment of atmospheric N_2O has been explained on the basis of its main stratospheric sink, i.e., photolysis. N_2O that escapes photolysis and is reimported into the troposphere causes enrichment relative to the average sources. Theory and experiment of photolysis show a qualitative agreement. Photolysis experiments over the entire range of wavelength of interest still have to be performed. Analysis of stratospheric N_2O samples confirms that enrichment increases with altitude. Increased interest in N_2O was sparked by the awareness that in this linear molecule, NNO, fractionation through photolysis is not identical for the two different N atoms.

Instrumental innovation in mass spectrometry now allows the detection of the isotopic ratios for both positions using only small amounts of sample. Also spectroscopic techniques allow such determinations. Figure 4 shows the result of N_2O photolysis for ^{15}N at 193 nm. The fractionation factors for $^{15}\text{N}^{14}\text{NO}$ and $^{14}\text{N}^{15}\text{NO}$ are 10.9 and 35.7‰ respectively. Thus $^{14}\text{N}^{15}\text{NO}$ is less likely to be photolyzed. The same applies for $^{14}\text{N}^{14}\text{N}^{18}\text{O}$, for which at this wavelength the fractionation factor is 17.3‰. For the other important stratospheric reaction, $\text{O}(^1\text{D}) + \text{N}_2\text{O}$, only few data are available. Possible gas-phase sources of N_2O have been proposed in the literature, and the occurrence of MIF, albeit at a low level ($\delta^{17}\text{O} \approx 1$ ‰), has intensified this interest. No confirmation has been provided, and the cause of MIF in N_2O , if a real effect, remains unknown. In the meantime the intramolecular site preference of ^{15}N fractionation in tropospheric N_2O has been determined by careful FTIR analyses to be 19.8‰.

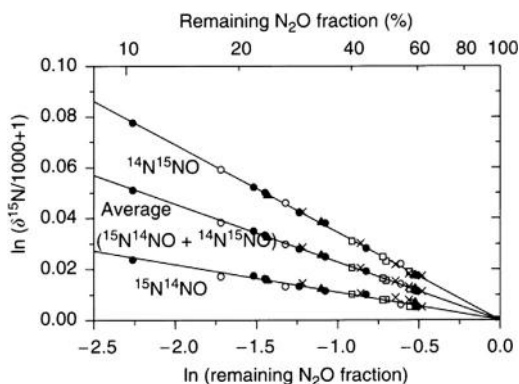


Figure 4 A Rayleigh type of plot for the change in the ^{15}N isotopic composition of N_2O subjected to photolysis at 193 nm. The enrichment for $^{14}\text{N}^{15}\text{NO}$ is considerably larger than for $^{15}\text{N}^{14}\text{NO}$.

Molecular Oxygen, O_2

In view of the sheer size of this reservoir, no usable isotope effects were originally contemplated to occur. Notwithstanding, it has been established that even atmospheric oxygen itself possesses a small degree of MIF $\Delta^{17}\text{O} = -0.15\text{‰}$, which offers interesting applications. The ^{18}O isotopic enrichment of O_2 relative to ocean water (V-SMOW) is about 23.5‰ . The cause for this enrichment is well understood (Dole effect). Respiration processes in plants and soils favor the use of the lighter oxygen isotopologues. In addition to this is that the leaf water in plants, which forms the substrate for the oxygen appearing in photosynthetic products, is enriched. These processes keep atmospheric oxygen enriched relative to the very large reservoir of terrestrial water. Because of this exchange between the large terrestrial water reservoir and atmospheric O_2 , its isotopic composition should be strictly mass-dependent. However, a very small fraction of O_2 is continuously converted to O_3 , which exhibits MIF. Thus a small excess of ^{17}O is continuously being withdrawn. In the stratosphere, this MIF signal is transferred from the O_3 pool via $\text{O}(^1\text{D})$ to become incorporated into CO_2 . The ensuing step in the process is that CO_2 reenters the troposphere, where it isotopically exchanges with H_2O at the surface. In this way, CO_2 loses its MIF signature continuously to the large reservoir of H_2O . Another process transferring MIF is the reaction $\text{O}(^1\text{D}) + \text{H}_2\text{O}$, which also leads to a net loss of ^{17}O . Both processes remove a small excess of ^{17}O from the O_2 reservoir. Given the long lifetime of atmospheric oxygen of roughly 1000 years, it acquires a deficit of ^{17}O , resulting in the value $\delta^{17}\text{O} = -0.15\text{‰}$. Although this value is small, there are applications in limnology and oceanography in which atmospheric oxygen can be distinguished from photosynthetic oxygen.

Aerosols

The oxidation of SO_2 by O_3 or H_2O_2 in the liquid phase leads to a small degree of MIF in atmospheric sulfates. This

can possibly be applied to trace the oxidation pathway of SO_2 . There is one report of SO_2 possessing MIF, for which the reason is not known. Sulfur isotope measurements on marine sulfate aerosol particles support the hypothesis that dimethyl sulfide can be a source of non-sea-salt sulfate. Sulfates remain widely analyzed for their isotopic composition with useful applications. Some research has also been dedicated to carbon but also chlorine isotopes in aerosols. Much work concerns nitrates. It has been shown that the oxygen in aerosol nitrates is mass independently fractionated. The responsible process is the important reaction $\text{NO} + \text{O}_3 \rightarrow \text{NO}_2 + \text{O}_2$. When this NO_2 is further oxidized, much of the ozone MIF signal is transferred to nitrate. The effect has been measured, and modeled, and interesting applications are in ice cores and the pathways of nitrates after deposition.

See also: **Aerosols:** Aerosol Physics and Chemistry. **Chemistry of the Atmosphere:** Methane. **Climate and Climate Change:** Carbon Dioxide. **Land-Atmosphere Interactions:** Overview; Trace Gas Exchange. **Ozone Depletion and Related Topics:** Long-Term Ozone Changes. **Solar System/Sun, Atmospheres, Evolution of Atmospheres:** Evolution of Earth's Atmosphere; Planetary Atmospheres: Mars; Planetary Atmospheres: Venus.

Further Reading

- Barkan, E., Luz, B., 2012. High-precision measurements of $^{17}\text{O}/^{16}\text{O}$ and $^{18}\text{O}/^{16}\text{O}$ ratios in CO_2 . *Rapid Communications in Mass Spectrometry* 26 (23), 2733–2738. <http://dx.doi.org/10.1002/rcm.6400>.
- Ghaderi, N., Marcus, R.A., 2011. Bimolecular recombination reactions: low pressure rates in terms of time-dependent survival probabilities, total J phase space sampling of trajectories, and comparison with RRKM theory. *Journal of Physical Chemistry* 115 (18), 5625–5633. <http://dx.doi.org/10.1021/jp111833m>.
- Gonfiantini, R., Stichler, W., Rozanski, K., 1995. Standards and intercomparison materials distributed by the International Atomic Energy Agency for stable isotope measurements. In: *Reference and Intercomparison Materials Stable Isotopes of Light Elements*, IAEA-Techdoc 825. IAEA, Vienna, pp. 13–29.
- Gromov, S., Jöckel, P., Sander, R., Brenninkmeijer, C.A.M., 2010. A kinetic chemistry tagging technique and its application to modelling the stable isotopic composition of atmospheric trace gases. *Geoscientific Model Development* 3, 337–364.
- Kaye, J.A., (Ed.), 1992. *Isotope Effects in Gas-Phase Chemistry*. Proceedings of a Symposium by the Division of Physical Chemistry, 201st National Meeting of the American Chemical Society, Atlanta, Georgia. ACS Symposium Series 502. American Chemical Society, Washington, DC.
- Richet, P., Bottinga, Y., Javoy, M., 1977. A review of hydrogen, carbon, nitrogen, oxygen, sulphur, and chlorine stable isotope fractionation among gaseous molecules. *Annual Review of Earth and Planetary Sciences* 5, 65–110.
- Röckmann, T., Brenninkmeijer, C.A.M., Saueressig, G., et al., 1998. Mass-independent oxygen isotope fractionation in atmospheric CO as result of the reaction $\text{CO} + \text{OH}$. *Science* 281, 544–546.
- Thiemens, M.H., 1999. Mass-independent isotope effects in planetary atmospheres and the early solar system. *Science* 283, 341–345.
- Wolfsberg, M., van Hook, W.A., Paneth, P. (Eds.), 2010. *Isotope Effects in the Chemical, Geological and Bio Sciences*. Springer, Dordrecht, Heidelberg, London, New York. ISBN 20099423459.

Laboratory Kinetics

DJ Donaldson and SN Wren, University of Toronto, Toronto, ON, Canada

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by DJ Donaldson, volume 3, pp 1090–1097, © 2003, Elsevier Ltd.

Synopsis

Chemical kinetics is the study of the rates of transformation of chemical compounds from reactant species into products. This article describes the basic principles of chemical kinetics. By means of atmospherically relevant reactions, fundamental concepts relating to rate coefficients, reaction order, the rate expression, reaction mechanisms, elementary reactions, third-body reactions, and steady-state concentration analysis are introduced. Physical descriptions of the rate coefficient (and its temperature dependence) are given, derived from both collision theory and statistical mechanics. Approaches for extracting rate parameters from experiments are addressed, with a focus on pseudo-first order kinetic studies and relative rate determinations. The article briefly outlines the basic principles, as well as the advantages and disadvantages, of the three most commonly used experimental techniques for studying reaction kinetics: flow tubes, flash photolysis and smog chambers. Spectroscopic methods for detection of reagents and/or products are also discussed.

Introduction

Chemical kinetics is the study of the rates of transformation of chemical compounds from reactant species into products. The rate of a reaction is defined to be the rate of decrease with time of the reactant concentration (in number of moles or molecules per unit volume) due to chemical reaction(s), or, equivalently, the rate of increase of the product concentration. In the atmosphere, chemical reactions typically involve free radicals, as reactants, products, or both. Free radicals are neutral fragments of molecules containing an unpaired electron and are generally very reactive. Some important atmospheric examples include OH, HO₂, HCO, CH₃, and CH₃O. Stable molecules such as O₃, NO, and NO₂ also contain unpaired electrons, which can influence their reactivity.

Principles of Chemical Kinetics

Overall and Elementary Reactions

The rate is expressed in terms of 1 mol of reaction, so rates of concentration change are normalized to the reaction stoichiometric coefficients, which give the number of moles of each compound appearing in the balanced chemical equation. Thus for the reaction [I], in which *a*, *b*, *c*, and *d* represent the stoichiometric coefficients for their respective compounds in the balanced chemical equation, the rate is defined by eqn [1].



$$\text{Rate} = -\frac{1}{a} \frac{d[A]}{dt} = -\frac{1}{b} \frac{d[B]}{dt} = \frac{1}{c} \frac{d[C]}{dt} = \frac{1}{d} \frac{d[D]}{dt} \quad [1]$$

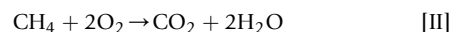
This expression gives the phenomenological rate at which reactants A and B are transformed to products C and D. It contains no information concerning the mechanism of the reaction, nor does it have any predictive utility.

Since most chemical reactions involve an exchange of atom(s) among the reagents, it is intuitive that such processes should proceed via collisions, or at any rate, close approaches

of the reagents to one another. Since the likelihood of such collisions increases with number density (concentration), the rate is often expressed in terms of the concentrations of the compounds involved (eqn [2]), where *k*, called the rate constant or rate coefficient, represents a concentration-independent factor (which may depend on other parameters, particularly the temperature; see below) and the exponents (*w*–*z*) give the order of reaction with respect to each of the species involved. The overall order of reaction is given by the sum of the exponents.

$$\text{Rate} = k[A]^w[B]^x[C]^y[D]^z \quad [2]$$

The exact form of this expression must be determined experimentally; any of the species present in the reaction system may appear and the exponents, *w*–*z* may or may not bear any relationship to the stoichiometric coefficients *a*–*d*. For example, the low temperature oxidation of methane to carbon dioxide and water, whose balanced equation is given by [II], has a rate expression that may be approximated by eqn [3], where [M] represents the total concentration of all species present, and *k* and *k'* are two different rate constants.

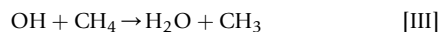


$$\text{Rate} = -\frac{d[\text{CH}_4]}{dt} = k[\text{CH}_4][\text{M}] \left\{ 5 \left(e^{4k'[\text{O}_2]t} - 1 \right) \right\} \quad [3]$$

This rate expression is clearly not intuitively derived from the balanced reaction and indicates that the reaction does not proceed, at a molecular level, by the simultaneous interaction of a methane molecule with two oxygen molecules. A more complex process, involving many individual reaction steps, is responsible for the observed rate expression.

The reaction mechanism is a sequence of elementary chemical reactions, each of which occurs by a single interaction between reagents. This sequence of reactions, taken together, must give rise to the observed rate expression. The rate expressions for elementary reactions are easily written from their balanced chemical equations; for example, the elementary reaction [III], which is the first step in the atmospheric

oxidation of methane, has a rate given by eqn [4], since the reaction takes place by the interaction of *one* OH molecule with *one* methane molecule.



$$\text{Rate} = -\frac{d[\text{CH}_4]}{dt} = k[\text{CH}_4][\text{OH}] \quad \text{[4]}$$

Elementary reactions generally involve atoms, free radicals, or other highly reactive species.

Elementary Chemical Reactions: General Features

If a reaction is known to be elementary, the explicit time dependence of the concentration of all chemical species involved may be obtained through integration of the rate expression. A few instances are of particular interest in atmospheric chemistry.

First-order reactions

These are reactions in which the exponents in the rate expression (eqn [2]) sum to unity. This class of system includes direct photochemical transformations and unimolecular reactions (in the high-pressure limit). For example, the rate expression for photodecomposition of formaldehyde in the near ultraviolet, shown in reaction [IV], is given by eqn [5],



$$\text{Rate} = -\frac{d[\text{H}_2\text{CO}]}{dt} = k(h\nu)[\text{H}_2\text{CO}] \quad \text{[5]}$$

In which the rate constant depends explicitly upon the photon energy, $h\nu$, through the absorption spectrum of formaldehyde. The photolysis rate constant in atmospheric chemistry is denoted as J rather than k , hence the $k(h\nu)$ in eqn [5] for reaction [IV] would typically be written as J_{IV} . This expression is easily integrated to yield eqn [6],

$$[\text{H}_2\text{CO}]_t = [\text{H}_2\text{CO}]_0 e^{-J_{\text{IV}}t} \quad \text{[6]}$$

which represents a simple exponential decay from an initial concentration, $[\text{H}_2\text{CO}]_0$.

Second-order reactions

These include bimolecular reactions between different reagents, as well as self-reactions. The rate expression depends upon the instantaneous concentrations of both species involved in the reaction. For example, the reaction of OH with methane (reaction [III] above) is second order. The time dependence of reactant concentrations is somewhat more complicated than that for first-order reactions, as shown by eqn [7].

$$kt = \frac{1}{[\text{OH}]_0 - [\text{CH}_4]_0} \ln \left\{ \frac{[\text{OH}]_t [\text{CH}_4]_0}{[\text{CH}_4]_t [\text{OH}]_0} \right\} \quad \text{[7]}$$

In laboratory studies of bimolecular reactions, the concentration of one reagent is typically forced to be in great excess, and so remains essentially a constant during the reaction. For example, if $[\text{CH}_4] \gg [\text{OH}]_0$, $[\text{CH}_4]_t \approx [\text{CH}_4]_0$, and the rate of change of the OH concentration can be expressed as in eqn [8].

$$\text{Rate} = -\frac{d[\text{OH}]}{dt} = k[\text{CH}_4]_0[\text{OH}] = k'[\text{OH}] \quad \text{[8]}$$

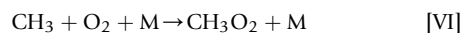
Here, $k' = k[\text{CH}_4]_0$. In this case, the reaction is said to exhibit pseudofirst-order behavior, and the time dependence of the OH concentration is given by eqn [9].

$$[\text{OH}]_t = [\text{OH}]_0 e^{-k't} \quad \text{[9]}$$

The true bimolecular rate constant is then obtained by performing measurements at many different values of $[\text{CH}_4]_0$, then plotting the observed k' as a function of $[\text{CH}_4]_0$. The slope of such a plot yields k . An example of this is presented below.

Third-order reactions

In the atmosphere, many elementary reactions require the participation of three molecules, and are thus strictly third order. Important reactions such as [V] and [VI] require the participation of a 'third body' (usually designated as M) to remove the excess energy in the newly formed product.



In Earth's atmosphere, M is generally N_2 and O_2 , since these gases account for approximately 99% of the atmospheric composition. The rate expressions of such reactions are properly written as eqn [10], in which the participation of the third body is made explicit.

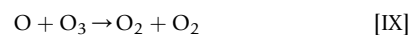
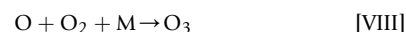
$$\text{Rate} = \frac{d[\text{HONO}_2]}{dt} = k[\text{OH}][\text{NO}_2][\text{M}] \quad \text{[10]}$$

At the higher pressures present in the lower atmosphere, some (but by no means all!) such reactions are at their high-pressure limits: collisions with M are frequent enough that the reaction exhibits pseudosecond-order kinetics, with the concentration of M incorporated into the observed rate constant.

Sequences of Elementary Reactions

As mentioned earlier, the overall reaction mechanism is constructed from a sequence of elementary reactions, which combine to yield the observed reaction rate expression. Often in such a sequence, a species X is formed as the product of one reaction, and consumed in a subsequent reaction. It can be shown that, if the reaction consuming X occurs more rapidly than its formation, the concentration of X will remain small, and almost time invariant. Here, the formation of X is said to be the 'ratelimiting step' in the reaction sequence, since X is consumed as soon as it is formed. Under these circumstances, the concentration of X is said to be in 'steady state.' Because the majority of atmospheric radicals are highly reactive, the analysis of their steady state concentrations becomes important.

Consider the sequence of reactions given below ([VII]–[X]), which represents a simplified version of the 'oxygen only' chemistry of ozone.



For both of the reactive oxygen species (O and O₃), the rates of change of concentration are given by the difference between the rates of formation (e.g., reaction [VIII] for ozone) and destruction (e.g., reactions [IX] and [X] for ozone).

$$\frac{d[\text{O}_3]}{dt} = k_{\text{VIII}}[\text{O}][\text{O}_2][\text{M}] - k_{\text{IX}}[\text{O}][\text{O}_3] - J_{\text{X}}[\text{O}_3] \quad [11]$$

$$\frac{d[\text{O}]}{dt} = 2J_{\text{VII}}[\text{O}_2] + J_{\text{X}}[\text{O}_3] - k_{\text{VIII}}[\text{O}][\text{O}_2][\text{M}] - k_{\text{IX}}[\text{O}][\text{O}_3] \quad [12]$$

If the concentrations of these reactive species are time invariant, eqns [11] and [12] are set to zero and the rates of formation and destruction are equal. The steady state concentrations of each (designated 'ss') may then be calculated by eqns [13] and [14].

$$[\text{O}_3]_{\text{ss}} = \frac{k_{\text{VIII}}[\text{O}][\text{O}_2][\text{M}]}{k_{\text{IX}}[\text{O}] + J_{\text{X}}} \quad [13]$$

$$[\text{O}]_{\text{ss}} = \frac{2J_{\text{VII}}[\text{O}_2] + J_{\text{X}}[\text{O}_3]}{k_{\text{VIII}}[\text{O}_2][\text{M}] + k_{\text{IX}}[\text{O}_3]} \quad [14]$$

The measured atmospheric concentrations of oxygen atoms and ozone are fairly constant over a time scale of several hours, under constant illumination from the Sun. The analysis given above, although crude, provides a useful picture for how these species' concentrations depend upon altitude (through the total pressure, related to [M]) and the altitude dependence of solar ultraviolet intensities, through k_{VII} and k_{X} .

Temperature Dependence of Elementary Chemical Reactions

The rates of elementary reactions depend upon the reagent concentrations, as discussed above, but also on other parameters, most importantly on temperature, T . The T -dependence of reaction rates is incorporated into k , the rate constant (more properly called the rate coefficient). Empirically, it is often found that over the temperature range important in the lower and middle atmosphere k depends on T exponentially, as shown in eqn [15].

$$k = Ae^{-E_a/RT} \quad [15]$$

In this expression (commonly known as the Arrhenius expression), E_a represents the activation energy, R is the gas constant ($8.314 \text{ J K}^{-1} \text{ mol}^{-1}$), T is the temperature in Kelvin, and A is unimaginatively named the preexponential factor. Careful measurements over wide enough temperature ranges reveal that the preexponential factor depends weakly on temperature, so is strictly written as $A(T)$. The dependence is generally much weaker than exponential, however, and a temperature-independent value for A is often used. This empirical expression may be interpreted simply as follows: for a chemical reaction to occur, the reagents must collide (1) with sufficient energy to overcome any energy barrier(s) along the reaction path and (2) in an appropriate geometry to facilitate product formation. The preexponential factor contains information concerning the collision rate (dependent upon the reagent velocities, and hence temperature) and any geometric constraints on the reaction. The exponential term arises from consideration of the fraction of collisions, under conditions of

thermal equilibrium, which possess energy in excess of some threshold value, E_0 . This energetic threshold is assumed to arise from the presence of energetic 'barriers' along the reaction path, due to the energetic cost(s) of rearranging the chemical bonds. The activation energy is closely related to threshold energy; for our purposes they may be taken as equivalent.

An alternate interpretation considers the elementary reaction to occur in two steps: in the first, the crest of the energetic barrier is achieved; an activated complex (or transition state) is reached. From this point, the reaction may continue to products, or return to reactants as an unsuccessful collision. If the activated complexes are held to be in rapid quasiequilibrium with reagent species, with product formation being rate limiting, a statistical mechanical analysis of a reaction between A and B yields the rate constant in terms of the 'partition functions,' Q , of the reactants (Q_A and Q_B) and of the activated complex (designated $Q^\ddagger$). This can be expressed by eqn [16], where k_B is the Boltzmann constant (equal to the gas constant divided by the Avogadro number) and h is the Planck's constant.

$$k = \frac{k_B T}{h} \frac{Q^\ddagger}{Q_A Q_B} e^{-E_0/k_B T} \quad [16]$$

The partition functions describe how thermal energy is partitioned among the available degrees of freedom in a molecule (i.e., translation, vibration, rotation) and depend only upon molecular properties, such as bond lengths and angles and vibrational energies. These may be determined spectroscopically or calculated theoretically, allowing a priori calculation of the rate constant for the reaction using eqn [16]. However, accurately carrying out these calculations is, in practice, quite difficult and hence is often limited to simple reactions.

This dependence of the rate constant on molecular properties provides an explanation as well for observed isotope effects in reaction rates. Often it is found that reactions of chemically identical, but isotopically different, species will exhibit different rates, with the heavier isotope displaying the smaller rate constant. The smaller vibrational frequencies, and thus lower zero-point energy, of the heavier isotope give rise to somewhat larger reagent partition functions and higher values of E_0 , and hence a smaller rate constant. This is known as the kinetic isotope effect, and is often used to elucidate reaction mechanisms.

Methods for Measuring Atmospheric Rate Parameters

Extracting Useful Parameters from Experiments

A host of methods is in current use for measuring gas phase reaction rates within a laboratory setting. The specific technique employed in any particular case depends somewhat on the particulars of the reaction being studied, and also on the available apparatus in any given laboratory.

As may be inferred from the foregoing discussion, the object of almost all kinetics experiments is to determine the reaction rate coefficient, preferably as a function of temperature, and also, in the case of termolecular and unimolecular reactions, of pressure. Very often this is done by following the concentration of some reactant or product as a function of time, generally

under pseudofirst-order conditions. From eqns [6] and [9], one sees that the apparent rate constant, k' , may be obtained from an appropriate fit to a plot of concentration against time. The values of k' thus obtained are then plotted as a function of the concentration of the reagent held in excess, which is generally varied over about an order of magnitude. The slope of this dependence yields the true second-order rate coefficient at a given temperature.

Figure 1 illustrates this situation for a hypothetical reaction between OH and an unspecified hydrocarbon molecule, RH, at room temperature. The reaction of OH with methane, given as reaction [III] above, is an example of such an OH + RH reaction. Figure 1(a) shows three plots of the decay of OH as a function of time, following its creation at time $t = 0$. The OH concentration is at all times much smaller than that of RH, ensuring that pseudofirst-order conditions are maintained. Curve (i) displays the decay for a low concentration of RH; curves (ii) and (iii) display decays for successively higher RH concentrations. All three exhibit single exponential decays, as required for straightforward pseudofirst-order analysis. Figure 1(b) shows the exponential decay constants obtained from data like those in Figure 1(a), plotted as a function of RH concentration; the slope of this linear fit gives the bimolecular rate coefficient at room temperature. Note the existence of a finite (positive) intercept, indicating a finite rate of

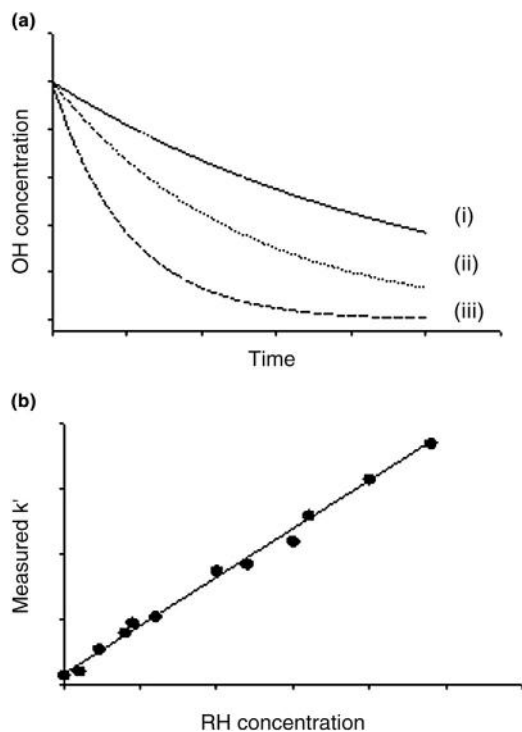


Figure 1 Extraction of a bimolecular rate coefficient from experimental pseudofirst-order reaction conditions, using the OH + RH $\rightarrow$ H₂O + R reaction as an example. (a) Decay of OH as a function of exposure time to RH. Curves (i), (ii), and (iii) display the result for increasing RH concentrations. (b) The pseudofirst-order rate coefficient obtained from fits to data such as that shown in (a), plotted as a function of the (constant) RH concentration. The slope of the linear fit to these data gives the true second-order rate coefficient.

disappearance of OH in the absence of RH. This loss of reagent may be due to a combination of wall reactions, reactions of OH with impurities, and self-reaction of OH.

Modern experiments yield measurement uncertainties of 5–10% on the rate coefficient at any given temperature. Uncertainties are generally larger at lower temperatures, where the reaction rate is smaller for most reactions. The agreement between different laboratories is usually within 10–20% for uncomplicated reactions.

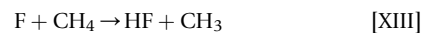
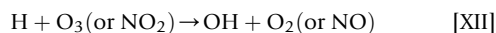
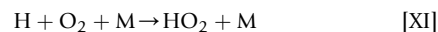
Experimental Techniques

Clearly, the technical issues in measuring reaction rates are: (1) to generate reactive species at some well-defined time $t = 0$; (2) to measure accurately the reagent (or product) concentrations as a function of time, over the course of the reaction; and (3) to minimize losses of reagent due to processes other than the reaction of interest. Implicit in (2) is the absolute identification of the species whose concentration is being followed.

Flow tubes

Two techniques are commonly used to study fairly fast reactions (i.e., those with rate coefficient greater than 10^{-15} – 10^{-16} cm³ mol⁻¹ s⁻¹): flow tube studies and flash photolysis. The flow tube method has several variants, all of which share the same basic principles. A schematic of a typical flow tube apparatus is shown in Figure 2. The flow tube consists of two main components: a temperature regulated outer tube and a concentric inner tube (the moveable injector). One of the reagents, entrained in a fast flow of carrier gas (usually an inert gas such as He or N₂) at low pressure (a few mbar) passes down the flow tube at constant velocity. The second reagent is introduced into the flow via the concentric inner tube, whose longitudinal position may be varied. The zero-of-time is established when the reagents first come into contact with one another; therefore, varying the position of the injector along the tube also varies the amount of exposure of the reagents to one another along the length of the flow tube. Since the reagents move down the flow tube at constant velocity, the contact length, as determined by injector position, is directly proportional to exposure time. The kinetics is studied by monitoring either the disappearance of reagent(s) and/or the appearance of product(s), as a function of injector position (and hence exposure time), in a volume near the end of the flow tube. Optical spectroscopy and mass spectrometry are the two most common methods of detection.

Reactive atomic or radical species are produced continuously by passing a precursor gas through a plasma generated by a microwave or radio frequency discharge (i.e., O and H atoms are formed from O₂ and H₂, respectively). These atomic reagents may be used directly, or they may be converted to different radical species via titration reactions with additional coreactants. Examples of this process are the reactions [XI], [XII], and [XIII] shown here.



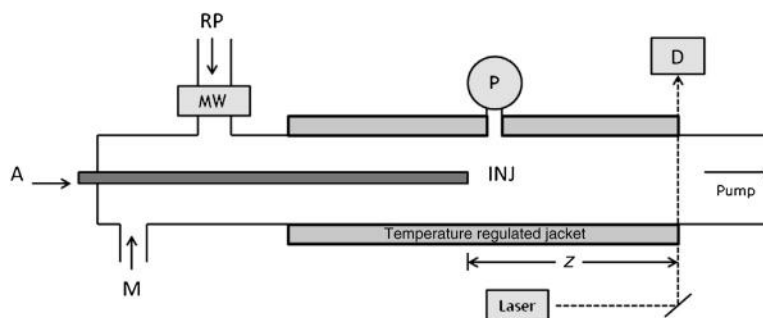


Figure 2 Schematic of a flow tube apparatus. Reactants and a carrier gas (M) are introduced from gas reservoirs into the flow tube. One reactant (A) is introduced through the moveable injector (INJ). Atomic or radical reagents are formed by flowing radical precursors (RP) through a microwave discharge cavity (MW). The total pressure in the flow tube is monitored (at P). The concentration of radical reactant (or of product) is monitored by a detector (D) at the exit of the flow tube (shown here as a laser–detector combination).

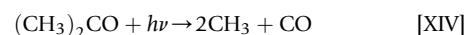
The radical concentrations are typically between one and three orders of magnitude smaller than those of the molecular reagents, so that a pseudofirst-order kinetic analysis is often possible.

The main drawback of the flow tube technique is that heterogeneous chemistry at the reactor walls may interfere with the kinetics (wall losses). There is also a need for well-defined flow parameters, and axial and radial concentration gradients may exist. However, major advantages of the flow tube technique include excellent reactant and detector versatility. The flow tube technique has been particularly well suited to take advantage of improvements in mass spectrometric detection. For example, developments in chemical ion mass spectrometry (CIMS) allow for the fast detection (~ 1 s) of atmospheric trace gases (e.g., reactive nitrogen species such as HNO_3 and NH_3), with sensitivity in the tens of pptv and lower. Proton-transfer reaction mass spectrometry, a variation of CIMS, has sensitivities nearing pptv for the detection of volatile organic compounds. Another recent development in the flow tube technique is the ability to simultaneously measure the concentrations of two radical reagents (and thus to determine radical–radical reaction rates, via analysis such as that shown in eqn [7]). Conventionally, flow tube reactors operate under low pressures and laminar flow regimes. However, atmospheric pressure flow tubes operating under turbulent flow conditions have recently been developed and characterized, allowing for the study of smaller rate constants as well as termolecular reactions. Finally, the flow tube technique may be adapted to study heterogeneous reactions as well (using, for instance, coated-wall flow tubes or entrained aerosol flow tubes).

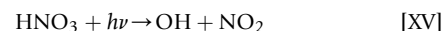
Flash photolysis

The flash photolysis method, which also has several variants, was first developed by Norrish and Porter in the late 1940s. In this technique, a suitable radical precursor species is already present in a well-mixed gas mixture, which also contains the second reagent and generally a buffer gas as well. The radical reagent is generated *in situ* by photodissociating the precursor molecule using a short (≤ 1 μs) pulse from a laser or flashlamp (almost always a laser nowadays). For example, methyl radicals are efficiently generated by the photolysis of acetone using

the 10 ns pulse from a 193-nm-wavelength ArF excimer laser (reaction [XIV]).



Similarly, OH may be formed via the 248-nm KrF excimer laser photolysis of nitric acid (reaction [XV]).



The buffer gas (usually an inert gas such as He or N_2) is present to thermalize the photodissociation products prior to reaction. For flash photolysis experiments, the zero-of-time is established by the pulse of light responsible for creating the reactive species. The decay of radical species concentration (or the growth of product species concentration) is monitored as a function of time following that pulse. Measurements are typically made using time-resolved optical spectroscopic methods such as transient absorption or laser-induced fluorescence (LIF) spectroscopy, which require a second laser pulse. In such experiments (often referred to as ‘pump and probe’ experiments), there is a variable delay time between the firing of the radical generating laser (pump) and the firing of the LIF laser (probe), allowing a variable exposure time of the reagents to one another prior to their interrogation. Following each pump–probe cycle, the reaction volume may be replenished with fresh reactants. Unless radical–radical reactions are being studied, the photolysis conditions are maintained such that the concentration of radicals remains much smaller than that of molecular reagents, allowing for a pseudofirst-order kinetic analysis.

Flash photolysis experiments can generally be carried out over a greater pressure range than flow tube experiments, which makes the determination of smaller rate constants possible. The experiments can also be conducted in a larger reactor, which reduces the influence of heterogeneous chemistry at the reactor walls. The main limitation of this technique is the requirement for photolytic generation of the radical reagent; the flash radiation may lead to the production of unwanted photofragments, thereby restricting the choice of reactants. Since flash photolysis experiments are carried out in real-time rather than steady state (as in the flow tube technique), the method of detection must have high time resolution. Advances in laser spectroscopy (narrowing of the laser pulse width and

advent of ultrafast femtosecond lasers) make it possible to study very fast reactions (and sometimes, reaction intermediates) using this method. Other recent developments include the application of time-resolved Fourier transform techniques, especially Fourier transform infrared spectroscopy (FTIR), with increasingly higher time resolution and sensitivity, and the use of cavity ring-down detection in absorption measurement, with the potential to increase detection sensitivity by many orders of magnitude.

Relative rate determination

Often, it is more convenient to measure the relative rate of reaction of two reagents with a third reactant, rather than perform an absolute rate determination. For instance, slower reactions may suffer from significant artifacts such as wall reactions, but their rate constants may be measured with quite reasonable precision (although not necessarily with great accuracy) in this manner. For a simple set of two competing reactions [XVI] and [XVII], it is possible to derive a simple relationship between the concentrations of the reactants and their rate constants, as shown by eqn [17].



$$\ln\left(\frac{X_0}{X_t}\right) = \left(\frac{k_X}{k_Z}\right) \ln\left(\frac{Z_0}{Z_t}\right) \quad [17]$$

Here, X_0 and Z_0 represent the respective concentrations at $t = 0$, X_t and Z_t give the concentrations at time t , and k_X and k_Z represent the two rate coefficients. If one of the rate coefficients is known independently, the other may be determined in this way. Note that the concentration of reactant A need not be measured at all; a similar relationship may be derived for the two product concentrations, meaning that the experimenter may monitor either of the ratios $[X]/[Z]$ or $[P1]/[P2]$, depending on convenience, detection sensitivity, or other experimental factors.

Chamber experiments

Gaining a better understanding of the chemistry behind photochemical smog formation motivated the development of chamber experiments in the 1970s. Large chambers (often referred to as smog, environmental, or simulation chambers) make it possible to study gas-phase kinetics on longer time scales and using lower, atmospherically relevant reagent concentrations (ppm and sub-ppm levels). Chambers may range in size from as small as 25 l to as large as 250 m³. The smaller of these chambers can operate at low pressures while the large environmental chambers usually operate near atmospheric pressure. The main advantage of using large chambers is that they have a very low surface area-to-volume ratio and thus slow homogeneous reactions gain in importance over fast heterogeneous processes which occur on the chamber walls. Early chambers were constructed out of glass or quartz, but nowadays most chambers are made out of (collapsible) flexible Teflon bags. These chambers are usually housed in a temperature regulated, dark enclosure made out of wood or aluminum. For photochemical experiments, this enclosure may be lined with reflective aluminum foil and mounted with lamps. Some chambers are

built outdoors to take advantage of the natural solar spectrum. In most chamber studies, a relative rate method is employed.

In a typical experiment, the first reagent, X (and the tracer gas, Z, for the relative rate analysis), is introduced in a buffer gas (which may be an inert gas, ambient air, or artificial air). The contents are usually allowed to equilibrate over several hours or a few days. The second reagent, A, is then introduced to the chamber to initiate the reaction; radical species are often produced photolytically from A, X, or Z. Application of the relative rate method is ideal as the rate constants thus determined are less sensitive to the presence of impurities and do not depend on knowing the exact concentration of reagent A (which may not be well characterized due to incomplete or uneven mixing). Rates can also be determined in a more complex matrix with multiple reaction partners. Pressure gauges, hygrometers, and thermocouples are usually coupled to the chamber walls, and a variety of analytical instruments can be interfaced to the chamber for the detection of gas-phase species. The large size of the chambers makes them ideal for long-path absorption spectroscopy (particularly FTIR) and in some cases, the instrumentation may even be placed within the reactor. Since the experiments occur on a longer time scale (up to a few days), instruments with lower time-resolution (such as gas chromatography/flame ionization detection) may be used.

A major disadvantage of chamber experiments is their large size and cost. Another disadvantage of using large chambers is that they may be difficult to clean between experiments and so 'sticky' molecules such as HNO₃ may accumulate on the reactor walls. Since reagent concentrations are already very low, the presence of unwanted species may have a significant impact on the chemistry. Environmental chambers have most recently been used to study complex atmospheric processes such as the formation of secondary organic aerosols from gaseous precursors.

Spectroscopic detection methods

Here, we present a brief overview of the various methods in common use at the time of writing. All spectroscopic methods rely upon the resonant absorption or emission of radiation in a wavelength region characteristic of the species being detected. The absorption and emission of radiation by atoms and molecules only occurs in particular spectral regions, corresponding to the energy differences between quantum levels. The amount of radiation absorbed by a sample of molecules or atoms at a particular wavelength, $A(\lambda)$, depends upon the path length of interaction between the light and the sample, l , the concentration of the absorbing species, c , and the molar absorptivity $\epsilon(\lambda)$ of that species. The molar absorptivity is a proportionality constant that describes how efficiently a specific molecule absorbs light of a given wavelength and it is somewhat temperature dependent. As long as the fraction of incident light that is absorbed remains small (i.e., less than about 20% of the incident intensity), then the fraction of light absorbed can be related to the concentration of absorbers by the Beer-Lambert law (eqn [18]).

$$A(\lambda) = \ln\left(\frac{I}{I_0}\right) = \epsilon(\lambda)lc \quad [18]$$

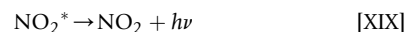
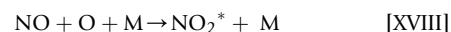
Thus, as long as $\epsilon(\lambda)$ is known, the Beer-Lambert relationship can be exploited to determine the absolute concentrations of reacting species as they change in time during a reaction.

Several implementations are in common use. In direct absorption spectroscopy, $A(\lambda)$ is recorded, either at a specific wavelength or as a function of wavelength, yielding an absorption spectrum. In flow tube methods, this may take place in a sample cell through which the reacting flow passes downstream of the mixing region; in flash photolysis methods, the absorption is measured in the reaction chamber as a function of time after the initial pulse. This variation is commonly known as transient absorption spectroscopy. In its simplest form, a single wavelength, which is chosen to be in resonance with a particular absorption feature, is transmitted through the sample, using resonance lamps or lasers. The time-resolved change in the intensity of this source after it has passed through the sample is measured following the initiation of reaction. Depending upon the exact experimental configuration used, the time resolution can be very good; typically it lies in the range 10^{-6} – 10^{-4} s. This technique is readily coupled to long path-length absorption cells to yield very sensitive concentration measurements. With the recent development of a wide variety of tunable solid state lasers, including tunable diode lasers, quantum cascade lasers, and optical parametric oscillators, coupled with sophisticated long-path absorption cells, such as cavity ring-down systems, it is now possible to measure gas-phase kinetics of a very large number of previously inaccessible species. Under the right conditions, a multiwavelength technique such as FTIR, or broadband cavity-enhanced absorption (IBB-CEAS) may be used to identify many of the reactants and products simultaneously as the reaction proceeds. However, the sensitivity of FTIR is not particularly high, and the wavelength range available in a single IBB-CEAS instrument is limited, so this is not often done, except for instance in environmental chambers, using FTIR.

If a molecule or atom reemits some fraction of the initially absorbed light, fluorescence detection may be used as a sensitive (often zero-background) probe of concentration. This forms the basis for the resonance fluorescence (RF) spectroscopy and laser induced fluorescence (LIF) spectroscopy methods. In RF spectroscopy, the sample is illuminated by an atomic resonance lamp optimized for a particular atom X; resonance emission (i.e., fluorescence) is observed from any atomic X which is present. Reactions of atomic species such as Cl and ground state O have been studied using such lamps. LIF spectroscopy generally utilizes a tunable laser source to scan the excitation wavelength over the absorption spectrum of a molecule of interest. When the excitation wavelength is in resonance with a molecular absorption transition, fluorescence from the excited molecules may be observed. The OH radical is especially well suited to LIF detection, and this is the method of choice in studying its kinetics. LIF spectroscopy is particularly useful in pump–probe flash photolysis experiments. In both RF and LIF spectroscopy, the fraction of absorbed light that is reemitted, depends upon many variables, but remains constant if experimental conditions do not change. The intensity of emitted light (i.e., the fluorescence intensity) is thus a very

sensitive proxy for the amount of absorption, so can be used to track concentrations via a modification of eqn [18].

Some chemical reactions result in the formation of electronically excited products. These excited products may emit light in a process known as chemiluminescence; the emission intensity is related to the concentration of electronically excited products formed in the reaction. Thus, chemiluminescence forms the basis for another detection method based on measuring emission intensity. For example, chemiluminescence is often used to measure NO concentrations. The time dependence of the concentration of NO may be followed by detecting emission (via reaction [XIX]) from electronically excited NO₂, which is formed in the chemiluminescent reaction [XVIII] (here, M represents a bath gas molecule).



Other spectroscopic detection methods, which are no longer in such common use, include laser magnetic resonance and electron spin resonance spectroscopy. Both rely upon the ability of an external magnetic field to perturb the energy levels of an atom or molecule. By varying the magnetic strength, optical transitions may be brought into resonance with a fixed-wavelength light source, allowing light absorption (and hence detection) to occur.

See also: Chemistry of the Atmosphere: Chemical Kinetics. Numerical Models: Chemistry Models. Ozone Depletion and Related Topics: Photochemistry of Ozone.

Further Reading

- Barnes, I., Rudzinski, K.J., 2004. Environmental Simulation Chambers: Application to Atmospheric Chemical Processes. In: IV. Earth and Environmental Sciences, vol. 62. Springer, The Netherlands.
- Finlayson-Pitts, B.J., Pitts Jr., J.N., 2000. Chemistry of the Upper and Lower Atmosphere. Academic Press, San Diego.
- Gierczak, T., Talukdar, R.K., Herndon, S.C., Vaghjiani, G.L., Ravishankara, A.R., 1997. Rate coefficients for the reactions of hydroxyl radicals with methane and deuterated methanes. *Journal of Physical Chemistry A* 101 (17), 3125–3134.
- Howard, C.J., 1979. Kinetic measurements using flow tubes. *Journal of Physical Chemistry* 83 (1), 3–9.
- Huey, L.G., 2007. Measurement of trace atmospheric species by chemical ionization mass spectrometry: speciation of reactive nitrogen and future directions. *Mass Spectrometry Reviews* 26.
- Paulsen, D., et al., 2005. Secondary organic aerosol formation by irradiation of 1,3,5-trimethylbenzene-NO_x-H₂O in a new reaction chamber for atmospheric chemistry and physics. *Environmental Science and Technology* 39.
- Pilling, M.J., Seakins, P.W., 1995. Reaction Kinetics. Oxford University Press, Oxford.
- Seeley, J.V., Jayne, J.T., Molina, M.J., 1993. High pressure fast-flow technique for gas phase kinetics studies. *International Journal of Chemical Kinetics* 25.
- Steinfeld, J.I., Francisco, J.S., Hase, W.L., 1999. Chemical Kinetics and Dynamics, second ed. Prentice Hall, Upper Saddle River, NJ.
- Thorn, R.P., Cronkhite, J.M., Nicovich, J.M., Wine, P.H., 1995. Laser flash photolysis studies of radical–radical kinetics: the O(³PJ) + BrO reaction. *Journal of Chemical Physics* 102 (10), 4131–4142.

Methane

E Dlugokencky, NOAA Earth System Research Laboratory, Boulder, CO, USA

S Houweling, SRON Netherlands Institute for Space Research, Utrecht, The Netherlands

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

This article discusses the role of methane in the atmosphere and its historic and recent variations. Techniques are described for monitoring its atmospheric abundance from the ground, and recently also from space. An overview is given of the main sources and sinks of atmospheric methane, and the large-scale constraints on the size of these processes as provided by the atmospheric measurements. The use of models is discussed, including the use of inverse modeling techniques for obtaining more detailed information on the spatiotemporal variation of methane emissions and sinks.

Introduction

In 2011, less than 2 parts in 10^6 (on a molar basis) of the Earth's atmosphere was methane (CH_4), but this is a factor of 2.5 more than what was present in 1750, prior to the industrial revolution. Even though this abundance is relatively small, CH_4 is the most abundant organic compound present in the atmosphere, and it significantly affects Earth's climate and atmospheric chemistry.

Methane is a greenhouse gas because it absorbs infrared radiation in a region of the terrestrial IR spectrum that is unaffected by water vapor and carbon dioxide (CO_2), the two most important greenhouse gases. Increasing amounts of atmospheric methane contribute to climate change. Methane currently contributes $\sim 0.5 \text{ W m}^{-2}$ to the total direct radiative forcing (estimated to be 2.84 W m^{-2}) caused by increasing atmospheric burdens of long-lived greenhouse gases since 1750. Only CO_2 contributes more to this total ($\sim 1.82 \text{ W m}^{-2}$).

Methane is removed from the atmosphere predominantly by gas-phase oxidation initiated by hydroxyl radical (OH). Changes in the abundance of atmospheric CH_4 affect the oxidizing capacity of the atmosphere. It is predicted that increasing CH_4 will decrease the concentration of OH in the atmosphere, so it will have an indirect effect on climate by affecting the atmospheric residence times of other reduced, long-lived greenhouse gases that are removed by reaction with OH . Under conditions where the concentrations of compounds called nitrogen oxides ($\text{NO}_x = \text{NO} + \text{NO}_2$) exceed a specific value, oxidation of CH_4 produces O_3 , which is itself an oxidant and greenhouse gas. About 5% of CH_4 emitted to the atmosphere is oxidized in the stratosphere, where it produces H_2O vapor, again potentially affecting climate.

Since CH_4 has the potential to impact climate significantly, it has been targeted by the Kyoto Protocol for reduced emissions. Strategies designed to mitigate the potential impact of CH_4 on climate must rely on a detailed understanding of methane's atmospheric budget (i.e., the balance of sources that emit CH_4 to the atmosphere, and sinks that remove it). Since many of the sources that emit CH_4 to the atmosphere are diffuse and highly variable in space and time, estimating total emissions is difficult. (This is in contrast to anthropogenic compounds such as chlorofluorocarbon compounds, whose emissions are determined from production.) Constraining the global CH_4

budget requires a range of studies including background atmospheric measurements, laboratory and field studies of CH_4 emission rates and the factors affecting these emission rates, and computer modeling studies.

Atmospheric Methane Measurements

Ice Core Measurements

Systematic modern measurements of atmospheric CH_4 abundance began in 1978. Our knowledge of CH_4 abundance prior to that comes from analysis of bubbles trapped in cores drilled and extracted from polar ice. It is assumed that, for methane (and many other stable, long-lived species), air trapped in polar ice bubbles accurately represents the atmosphere at the approximate time when the ice was formed. The longest record of atmospheric CH_4 abundance is for the past 800 000 years, based on an ice core from Antarctica. This record shows that prior to industrialization, atmospheric methane varied from about 350 parts per billion (ppb) during glacial times to 700 ppb during interglacial times. (Note: most methane measurements are reported in a dry-air, mole fraction scale; $\text{ppb} \equiv \text{mol}^{-1}$.) There is excellent correlation between CH_4 abundance and indicators of temperature extracted from the ice. **Figure 1** shows the historic variation of global averaged CH_4 mole fraction estimated from CH_4 measurements in Antarctic ice cores, Antarctic firn, and a globally distributed network of air sampling sites (since 1984). Detailed comparisons of measurements from Arctic ice cores with those from Antarctic cores indicate greater emissions in the northern hemisphere than in the southern. A large increase in methane begins ~ 200 years ago, and it is related to increased CH_4 emissions associated with the increased food and energy demands of a rapidly growing human population.

Ambient Air Measurements

Recent trends and year-to-year variations in global atmospheric CH_4 are monitored using a measurement network, consisting of about 100 sites that regularly receive 'background air'. We define 'background air' as air that is well mixed and representative of a large volume of the atmosphere. These observations generally fall into two categories. The first is measurements

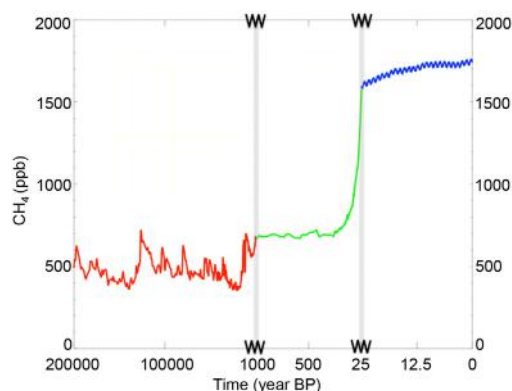


Figure 1 Reconstruction of the globally averaged CH_4 mole fraction in the past 200 000 years determined from ice core, firn, and whole air samples, highlighting variations during: Red, glacial cycles (Dome C Antarctica, courtesy R. Spahni, University of Bern); Green, the industrial revolution (Law Dome Antarctica, courtesy C. MacFarling Meure, CSIRO); Blue, recent decades (South Pole, NOAA–ESRL cooperative air sampling network, available at www.esrl.noaa.gov/gmd).

from discrete samples collected under predetermined meteorological conditions (wind speed and direction) at low frequency (e.g., weekly) in vacuum-tight containers and shipped to a central laboratory for analysis. The second method is measurements *in situ* at a relatively high frequency with an analytical instrument maintained at the sampling site. Gas chromatography with flame ionization detection has been the analytical method most often used for CH_4 measurements, although laser-based optical methods are becoming more common. There are advantages and disadvantages to both sampling strategies. Measurements *in situ* can be costly if multiple measurement sites are desired, and require highly skilled personnel to maintain the analytical instruments and

standard gases. While discrete sampling can achieve extensive geographical coverage at relatively low expense and with greater consistency (since all samples are analyzed on the same analytical system), it cannot match the sampling frequency of measurement *in situ*.

The most cost-effective method for determining large-scale features in the global CH_4 distribution is discrete sampling. Figure 2(a) shows the large-scale variation in methane over the past decades, derived from NOAA's global cooperative air sampling network. Since CH_4 at many sites is variable, particularly in the northern hemisphere, it is useful to smooth the measurements into zonally (i.e., longitudinally) averaged values. To do this, data from the sampling sites are smoothed temporally and as a function of latitude to define a surface of atmospheric methane as a function of latitude and time (Figure 3). The surface is used to calculate zonal averages; examples for the northern and southern hemispheres and global averages are plotted in Figure 2(a). Also plotted in Figure 2(a) are deseasonalized trend curves for each time series. The time derivative of the trend gives the instantaneous growth rate (shown for northern and southern hemispheres only in Figure 2(b)).

Some features in the data are evident. In both hemispheres, methane has been increasing from the start of the measurements in 1984 until 1999. During this period the rate of increase gradually decreased, with large interannual variation, until it came to a halt around the turn of the century. For the first time since the start of industrialization, methane stopped increasing. Between 1999 and 2006 methane levels remained approximately constant, after which the measurements show a renewed increase. The cause of this major growth rate anomaly is not well understood at present. Recent investigations point to changes in natural emissions and sink temporarily compensating for increased anthropogenic emissions.

Superimposed on the trend, the time series in Figure 2(a) show a strong seasonality, with smaller values during summer

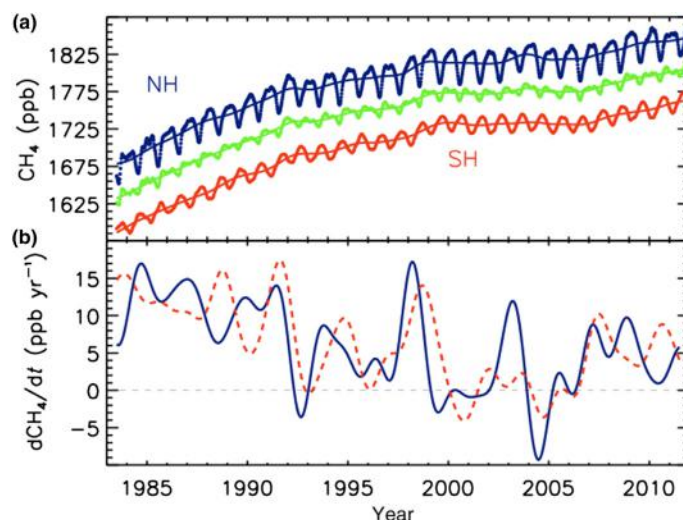


Figure 2 (a) Large-scale average time evolution of methane determined from surface measurements, where the cosine of latitude was used to weight for surface area. Plotted are northern hemisphere, global, and southern hemisphere mole fractions, from top to bottom. Deseasonalized trend lines are also plotted for each. (b) Instantaneous CH_4 growth rates determined as the derivatives of the trend lines for the northern (solid line) and southern (dashed line) hemispheres.

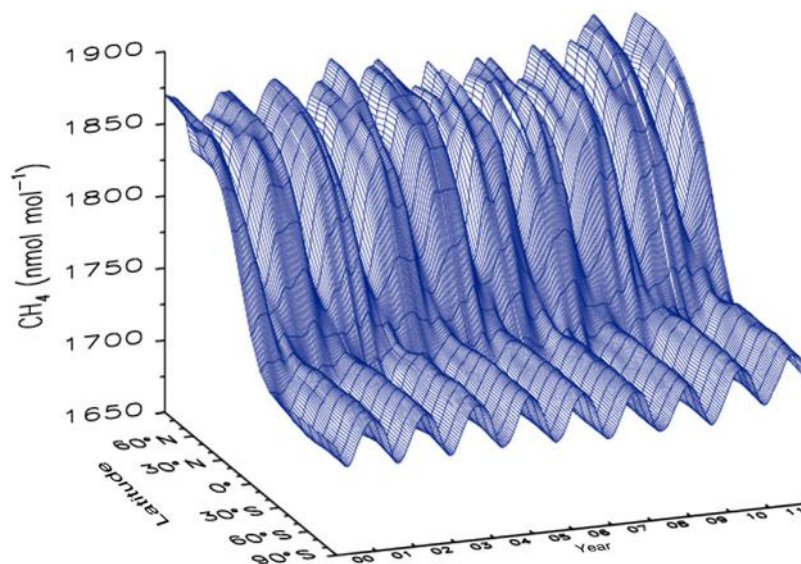


Figure 3 Smoothed, zonally averaged representation of the global distribution of methane in the marine boundary layer for the period 2000–11. The grid spacing is 10° latitude by 1 week. Data from 42 sampling sites were used to construct the surface.

than during winter. In the northern hemisphere, the seasonal cycle amplitude, and CH_4 values in general, are larger than in the southern hemisphere, and show a larger variability. This is because there are many CH_4 sources in northern latitudes, and, depending on the trajectory of an air parcel, it can contain highly variable amounts of CH_4 . Methane mole fractions are on average ~ 90 ppb larger in the northern compared with the southern hemisphere, because most of the anthropogenic CH_4 emissions are in the northern hemisphere and atmospheric mixing between hemispheres (or north–south mixing within a hemisphere) is not rapid enough to homogenize CH_4 . Measurements at sites from comparable latitudes indicate that in addition to N–S gradients, longitudinal (or E–W) gradients are also observed despite the short timescales of zonal mixing. For example, CH_4 values at Mace Head, Ireland (53° N in the eastern Atlantic Ocean) are about 10 ppb lower than at Cold Bay, Alaska (55° N) and Shemya Island (53° N) in the Pacific Ocean. Qualitatively, this makes sense because Cold Bay and Shemya are much closer to strong Siberian sources of CH_4 than Mace Head. Another important constraint on the global CH_4 budget is the vertical gradient. Some measure of the vertical gradient is observable in **Figure 4** where CH_4 measurements are plotted for air samples collected above Carr, a site in northern Colorado at 1740 m above sea level (triangles) and Cape Grim, a site on the northwest coast of Tasmania in Australia (circles). At Carr, CH_4 is, on average, greater at the surface, close to sources, than it is aloft. The low values above 6500 m are due to transport from the stratosphere. Above Cape Grim, CH_4 values are greater aloft, since there are only few sources of CH_4 in the high southern hemisphere, and most CH_4 in that part of the atmosphere has been transported from the northern hemisphere in the upper troposphere.

The global averages in **Figure 2** can be used to calculate the ‘burden’ (mass) of CH_4 in the atmosphere (~ 4989 Tg CH_4 in 2011; where 1 Tg = 10^{12} g). When the burden is combined with an estimate of the globally averaged methane atmospheric

lifetime (9.5 year) and a negligible growth rate between 2000 and 2010, total emissions are calculated as:

$$\frac{\partial \text{CH}_4}{\partial t} = Q - \frac{\text{CH}_4}{\tau} \quad [1]$$

where τ is the CH_4 atmospheric lifetime (years), Q emissions (Tg CH_4), and (CH_4) the atmospheric methane burden (also in Tg CH_4). Rearranging eqn [1] and solving for Q , the total source of CH_4 to the atmosphere is ~ 525 Tg CH_4 year $^{-1}$ (including soil oxidation). Some useful constraints on the global methane budget are also obtained from the hemispheric averages. Assuming an interhemispheric, exchange time of 1 year, a methane lifetime of 9.5 years, and steady state conditions, the observed interhemispheric difference of 90 ppb constrains the fraction of emissions in the northern hemisphere to $\sim 70\%$.

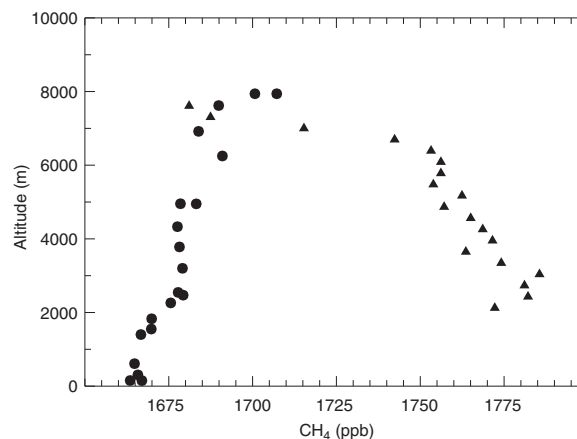


Figure 4 Vertical profiles of CH_4 over Cape Grim, Australia, for 15 March 1995 (circles) and Carr, Colorado, on 7 July 1994 (triangles). Cape Grim data courtesy of R. Langenfelds, CSIRO Division of Atmospheric Research, Australia.

Further tests of our understanding of the methane budget are obtained when measurements are compared with models that include atmospheric transport and chemistry. For example, the observed seasonal cycle of CH_4 constrains seasonality of loss, emissions, and transport. Model studies, described in more detail below, have helped to further constrain the methane budget.

Isotope Measurements

Each source of atmospheric CH_4 has a characteristic range of values for the ratios of ^{14}C and ^{13}C relative to ^{12}C , and D (i.e., ^2H or deuterium) relative to H, and the ratios depend on the mechanistic details of CH_4 production and destruction prior to release to the atmosphere. The ratios are measured relative to a standard and expressed in parts-per-thousand deviations from that standard (in units called 'per mil,' abbreviated ‰). Isotopic measurements of CH_4 provide a constraint on its budget; the mass-weighted isotopic composition of CH_4 emissions must equal the value found in the background atmosphere, after accounting for fractionation during loss processes (chemical oxidation and destruction by soil bacteria). For example, the ^{14}C content of CH_4 from biological sources is about 1 part in 10^{12} ^{14}C , while the CH_4 in natural gas has no ^{14}C . (Note: ^{14}C is a radioactive isotope of C that has a half-life of 5730 years, therefore the CH_4 in fossil fuels contains no ^{14}C .) To first order, comparing a measure of ^{14}C in background atmospheric methane with ^{14}C in CH_4 from biological sources will give an indication of the amount of CH_4 emitted during fossil fuel exploitation. Further examples of the use of isotopes are given later.

Remote Sensing

A recent development is to measure CH_4 using open path spectroscopy, which is generally referred to as remote sensing. The instruments are mounted on various platforms, including Earth orbiting satellites. The latter allows space-born greenhouse gas monitoring. A challenge, which so far has still prevented operational application of this technique, is to reach the required measurement accuracy and precision. While there is the potential to sample the entire Earth within several days, in

practice, the actual coverage that is obtained is limited by certain conditions that have to be met to be able to reach sufficient accuracy. For example, instruments that make use of Earth reflected sunlight require cloud free conditions during day time, and a sufficiently large surface albedo. The latter limits the coverage over sea for instruments that measure in the short wave infrared (SWIR), a problem that can partially be overcome by measuring in the direction of sun glint. Despite these limitations, the coverage that is reached by the exploratory mission SCIAMACHY allowed a new view on CH_4 in the atmosphere (Figure 5). Important urban and remote source regions could be mapped that are difficult to monitor from the ground, such as Southeast Asia and Amazonia. In maps of total column CH_4 , regions of elevated methane show up pointing to regional and local sources, associated, for example, with industrial activities and agriculture in China and tropical wetlands in the Amazon rain forest. Atmospheric models can be used to translate these signals into corresponding emissions, as will be explained later in this article.

The more advanced greenhouse gas observing satellite GOSAT allows an important step forward in measurement accuracy using Fourier transform spectroscopy. Compared with grating spectrometers such as SCIAMACHY the spectral resolution is much better. This is achieved at the price, however, of reduced measurement coverage because of the time it takes to record a single spectrum. As with SCIAMACHY, the major challenge is to account for scattering of light on aerosol and cirrus particles along the optical path. In the case of SCIAMACHY, this problem is solved by measuring the ratio of methane to carbon dioxide. The improved spectral resolution of GOSAT allows independent retrieval of methane and carbon dioxide.

Instead of measuring Earth reflected solar radiation, as utilized by SCIAMACHY and GOSAT in the SWIR, other instruments such as Atmospheric InfraRed Sounder (AIRS) and Infrared Atmospheric Sounding Interferometer (IASI) make use of thermal infrared radiation (TIR) emitted by Earth's surface. This approach is less sensitive to clouds and aerosols and allows extension of the measurement coverage to nighttime conditions (including the polar night). Due to the limited transparency of the Earth's atmosphere to infrared radiation, these measurements are most sensitive to methane in the upper troposphere

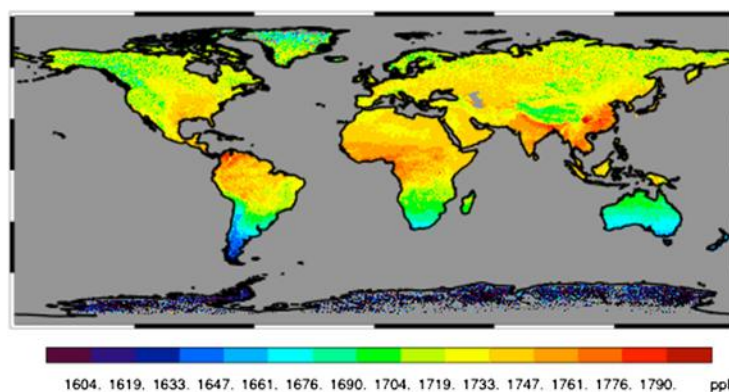


Figure 5 Vertical column averaged methane mole fractions retrieved from the SCIAMACHY satellite instrument (courtesy C. Frankenberg, JPL). This composite map represents the average of measurements collected in over 2003–10.

and lower stratosphere. By combining the SWIR and TIR measurements, that show very different vertical sensitivities, information is obtained about the vertical gradient of methane, which can in theory be used to separate the influences of surface sources from atmospheric sinks. The challenges are to reach a sufficient accuracy in modeling the vertical transport, and it is not yet possible to calibrate retrievals of CH_4 on the same standard scales used for *in situ* measurements, or at all.

Sources of Atmospheric Methane

Sources of methane at Earth's surface can be classified into ~ 10 major source types. Emission rates may vary by orders of magnitude over a few meters in spatial scale, and they are distributed over enormous geographical regions. The estimation of methane emissions at regional to global scales requires extrapolation from relatively few direct flux measurements, which introduces large uncertainties. Depending on the nature of the processes involved, these extrapolations make use of process-based models (mostly for natural sources such as wetlands and biomass burning) or statistical inventories (mostly for industrial and agricultural emissions). Both methods allow integration of the available information and evaluation against various types of available measurements as discussed in the next section.

In contrast to the diversity of sources, CH_4 is produced by only a few fundamental processes:

1. Microbial production: Decomposition of organic matter by methanogenic microbes under anaerobic conditions in, for example, wetlands, flooded soils, sediments of lakes and oceans, sewage, and digestive tracts of ruminant animals.
2. Thermogenic production: At a few kilometers' depth in soils, pressure and temperature become great enough for the decomposition of organic matter by condensation and cracking processes, eventually resulting in the formation of coal, oil, and natural gas.
3. Pyrogenic production: The incomplete combustion of organic material yields many compounds, including CO, formaldehyde, acetonitrile, and methane.
4. Photolytic production: Various biological carbohydrate polymers, such as cellulose, pectine, and lignin release CH_4 upon exposure to UV radiation. Emissions from living plants are believed to be small as they protect themselves against UV. Larger emissions are expected from plant litter. The significance of this process for the global methane budget is uncertain at present, but the most recent estimates point to a relatively minor importance.

The main source types are discussed in further detail below.

Figure 6 shows the estimated trends of anthropogenic CH_4 emissions per source category since 1970.

Natural Wetlands

Natural wetlands are the largest source of methane to the atmosphere, responsible for about 30% of the global and yearly emissions. These emissions show strong seasonal and interannual variations in response to climatic variations, and explain an important fraction of the observed variation of methane in the atmosphere. A wide variety of ecosystems are classified as wetlands, ranging from Arctic peat bogs to tropical river floodplains. Even temporary water saturated soils, which are generally not classified as wetland, can produce methane. Methane in soil and sediment can escape to the atmosphere by diffusion, ebullition, and plant mediated transport. In the oxic surface layer methanotrophic bacteria consume a part of the methane that is transported from deeper soil layers. Under dry conditions these bacteria extract methane from the ambient air, turning the soil into a net sink of methane. Besides inundated soils, methane production also takes place in sediments of lakes, rivers, and oceans. In such environments, usually the oxic section of the overhead water column acts as an efficient transport barrier, as a result of which only a small fraction of the methane that is produced in the sediment can reach the atmosphere. Emissions from fresh water lakes such as hydroelectric reservoirs and thermokarst thaw lakes are difficult to estimate because of the delicate balance between production and oxidation, which shows strong spatiotemporal variations across and between lakes.

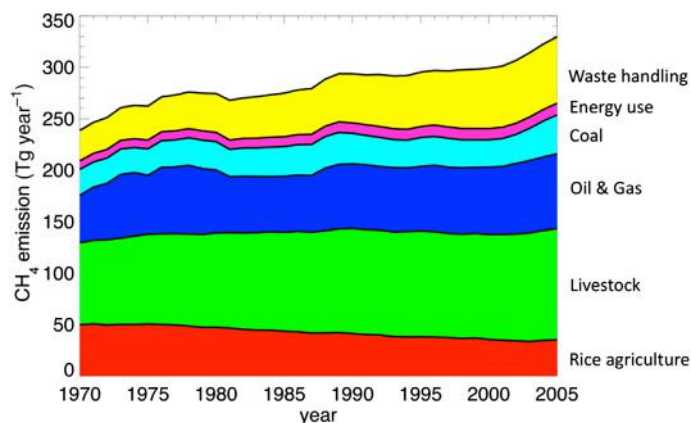


Figure 6 Recent evolution in anthropogenic methane emissions. Data taken from the global emission inventory EDGAR, version 4.1, <http://edgar.jrc.ec.europa.eu>.

Biomass Burning

Biomass burning is a relatively modest source of methane (<10%), but nevertheless important for understanding the observed methane variability because of the high spatiotemporal variability of the emissions. Large amounts of biomass are burned in the tropics in mostly human induced fires related to shifting cultivation, deforestation, burning of agricultural wastes, and the use of biofuels. The fraction of carbon that is released as methane depends on the fuel type and the burning conditions. For example, the efficient burning of dry savannah releases relatively small amounts of CH₄ compared with smoldering forest or peat fires. Since the timing and the extent of tropical biomass burning is closely related to climatic factors (precipitation and temperature), these sources vary strongly with season and contribute to interannual variations in atmospheric methane abundance and its isotopic ratios.

Fossil Fuel Use

The human exploitation of fossil fuel resources leads to important emissions of methane to the atmosphere. At every step along the line of mining, distribution, and use, small amounts of methane are released to the atmosphere. Globally integrated, the contribution of these sources is estimated to be 15–20% of total emissions. Together with agriculture, fossil fuel use is the primary driver of the observed CH₄ increase since preindustrial times. Despite the relatively high level of certainty concerning fuel production and consumption, emissions of methane are still quite uncertain. This is explained by the comparably poor quantification of emission factors and their change over time. In developed parts of the world energy use has become more efficient and clean, motivated in part by economic incentives but also by a growing environmental awareness. For example, the venting of gas from mining shafts is commonly flared to transfer methane into carbon dioxide, which has a lower greenhouse gas warming potential. Overall, emission inventories show declining fossil fuel emissions in Europe and North America, offset by increasing emissions from the growing Asian economies.

Besides human exploitation of fossil fuel, there are also natural emissions of fossil methane from relatively shallow sedimentary basins. Methane escapes to the atmosphere from reservoirs on land and off shore, through volcanic activity (mostly from so called mud volcanoes) and microseepage through faults in tectonically active regions. Large-scale emission estimates are highly uncertain, but might explain a gap between the reported fossil fuel emissions and the global inventory of methane from fossil fuel inferred from radio-carbon measurements.

Agriculture

Agriculture is a major player in the global budget of methane, mostly because of two sectors: animal husbandry and rice production. Methane is produced by methanogenic microbes in the fermentation of feeds in the rumen and lower digestive tract of ruminants. Methanogenesis in ruminants causes a loss of 2–12% of the gross energy of feeds and contributes about 30% of global anthropogenic methane emissions. Ruminant

methane emissions vary with animal productivity, fraction of digestible carbon in the food, and the treatment of animal waste. These in turn vary with farming practices regionally and between countries.

Methane production in rice paddies is in many ways similar to that in natural wetlands, except for the influence of management practices, such as irrigation, fertilization, harvesting, etc. Rice production statistics are a poor indicator of methane emissions. Production continues increasing to satisfy the needs of a growing population, whereas emission estimates show a declining trend. The latter is explained by changes in agricultural practices over time, such as the use of new rice plant varieties, synthetic fertilizers, and intermittent irrigation. Confirmation of the large-scale significance of these trends using atmospheric measurements is complicated by the limited capability to resolve the influences of coexisting sources in densely populated parts of the world.

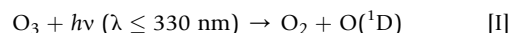
Waste Handling

Decomposition of organic waste leads to methane production in solid waste dumps, such as landfills, and in waste water streams and reservoirs. The amount of methane that is produced depends on the waste production, the biodegradable organic fraction, landfill storage conditions, and escape pathways. Covered landfills provide a suitable anaerobic environment for methane production, which can escape through leaks in the mantle. It is becoming more common, however, to extract and make use of methane produced in covered landfills, reducing the emissions to the environment. The net influence of increasing waste production and decreasing emission factors varies regionally, complicating its quantification.

Methane sources can be characterized by source-specific isotopic signatures. While methanogenesis results in emissions that are depleted of these isotopes, methane consumption by methanotrophs results in enrichment of these isotopes in the methane left behind. As a result, the net effect depends on the balance between microbial production and oxidation of methane. Methane derived from biomass burning retains the isotopic characteristics of the fuel, which is slightly different for C-3 ($\delta^{13}\text{C} \approx -27\text{‰}$) and C-4 plants ($\delta^{13}\text{C} \approx -12\text{‰}$), but it is highly enriched in ¹³C relative to the isotopic composition of background atmospheric methane ($\delta^{13}\text{C} \approx -47\text{‰}$). As mentioned earlier, fossil sources can be distinguished from other sources by the absence of the radioisotope ¹⁴C.

Atmospheric Sinks

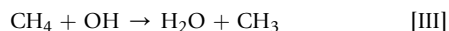
About 90% of atmospheric methane is removed from the atmosphere by reactions initiated by OH. The amount of OH in the atmosphere depends on the concentrations of O₃ and H₂O (vapor) and the UV actinic flux as follows:



Most of the O(¹D) is quenched by reaction with N₂ or O₂ to produce ground-state O(³P), but in the tropics (where up to 3% of air is water vapor) up to 25% reacts as follows:



(Note: the rate coefficient for reaction of $O(^1D)$ with H_2O is 5–10 times faster than the quenching reactions with N_2 and O_2 , so despite the large atmospheric concentrations of N_2 and O_2 , a significant fraction of $O(^1D)$ will result in OH production in the tropics.) The photolysis of O_3 to form $O(^1D)$ depends on sunlight, so OH production varies diurnally. Much of tropospheric O_3 is produced *in situ*. The chemistry that produces O_3 is complex, and the concentration of O_3 is a balance between production and loss. Oxidation of CH_4 is initiated by the following reaction:



Methyl radical (CH_3) adds O_2 to form CH_3O_2 . The subsequent CH_4 oxidation reactions affect O_3 , but the sign of the effect depends on the abundance of oxides of nitrogen (NO_x). In high- NO_x environments (>30 ppt) (ppt = parts per trillion or 10^{12}), O_3 is produced, but with $NO_x < 30$ ppt, O_3 is normally destroyed. The abundance of NO_x also affects OH directly. In NO_x -poor environments OH is destroyed, but with sufficient NO_x present CH_4 oxidation regenerates OH. On average, photochemistry of the troposphere is limited by the amount of NO_x , which means that the oxidation of methane is a net sink of radicals. As a consequence, an increase of methane is expected to decrease OH. Ultimately, most of the CH_4 that reacts with OH produces CO_2 .

Oxidation of CH_4 also influences the isotopic composition of methane. The rate coefficient for reaction of OH with $^{12}CH_4$ is $\sim 0.5\%$ greater than that for reaction with $^{13}CH_4$. This means that CH_4 found in the background atmosphere is enriched in ^{13}C relative to the mass-weighted sum of its sources.

In the stratosphere, reactions with chlorine and electronically excited oxygen atoms, $O(^1D)$, become significant sinks of methane, in addition to OH reaction. In the mesosphere, photolysis also contributes to CH_4 destruction, although this process remains negligible relative to the total sink. Stratospheric methane is relatively enriched in ^{13}C , both because it has been in contact with ^{12}C -depleting OH radicals for a relatively long time and because the Cl reaction leads to a larger fractionation of ^{13}C than the OH radical reaction. Oxidation of methane in the stratosphere is an important source of stratospheric water vapor.

Modeling as a Tool to Constrain the CH_4 Budget

Various types of numerical models have been developed to improve our understanding of the atmospheric methane budget. As discussed earlier, the global burden of methane is relatively well constrained by measurements. Highly simplified models that represent the atmosphere with one or a few boxes (called box models) have been used to interpret observations of atmospheric CH_4 over large spatial scales and long time periods, and to calculate global emissions and sinks. To interpret realistically the measurements on sub-hemispheric scales or at specific air sampling locations, models with finer spatial and temporal resolutions are needed, including detailed parameterizations of atmospheric transport and chemistry.

Box models and three-dimensional (latitude, longitude, and altitude) chemistry transport models typically calculate atmospheric abundances based on a set of hypothetical emissions that are prescribed as boundary conditions to the model. The simulated and observed abundances are then compared to test how realistic these boundary conditions were. It is also possible to compute the emissions and sinks that result in the best possible agreement between model and measurements. This technique, called inverse modeling, is an attempt to reconstruct the emissions and sinks from atmospheric measurements. Here it is assumed that differences between measurements and the ‘forward’ computed abundances can be attributed to emissions and sinks. This requires that inaccuracies in the representation of transport and the errors caused by comparing measurements at a fixed site with model ‘grid-box’ averages (where a grid box may include $200\,000\text{ km}^2$) are small compared with the uncertainties in the sources. Table 1 presents a synthesis of inverse modeling-derived estimates of the global CH_4 budget in the first decade of the twenty-first century including uncertainty ranges (95% confidence limits). As can be seen in the table, anthropogenic sources are estimated to account for 60% of the global emissions, which is a factor of 1.5 larger than total natural emissions and consistent with a 2.5-fold increase of the CH_4 concentration since preindustrial times. The rise in anthropogenic emissions of methane also caused a change in its lifetime because of the feedback of CH_4 on OH. Changes in other precursors of OH such as CO and NO_x also contribute to that change, the net effect of which is not well quantified at present. In addition, emissions from natural sources have changed in response to changes in climate and land use.

Figure 7(a) shows how global emissions varied during 2003–10, comparing the bottom-up and top-down approaches. The inversion-derived fluxes are on average $\sim 10\text{ Tg } CH_4\text{ year}^{-1}$ smaller, and show a temporal variation,

Table 1 Estimates of annual CH_4 emission rates (in Tg CH_4 , where $1\text{ Tg} = 10^{12}\text{ g}$) for the period 2000–10, derived from inverse modeling

Source	Emission rate
Fossil fuels	95 [77–123]
Agriculture & waste	210 [180–240]
Biomass burning	25 [14–45]
Natural wetlands	175 [145–208]
Other natural sources	45 [37–65]
Total	550 [526–569]
Soil oxidation	35 [26–42]
Tropospheric	490 [481–502]
Stratospheric oxidation	20 [12–29]
Total	545 [514–560]

Courtesy Kirschke, S., Bousquet, P., Ciais, P., Saunio, M., Canadell, J.G., Dlugokencky, E.J., Bergamaschi, P., Bergmann, D., Blake, D.R., Bruhwiler, L., Cameron-Smith, P., Castaldi, S., Chevallier, F., Feng, L., Fraser, F., Heimann, M., Hodson, E.L., Houweling, S., Josse, B., Fraser, P.J., Krummel, P.B., Lamarque, J., Langenfelds, R.L., Quéré, C., Naik, V., O'Doherty, S., Palmer, P.I., Pison, I., Plummer, D., Poulter, B., Prinn, R.G., Rigby, M., Ringeval, B., Santini, M., Schmidt, M., Shindell, D.T., J. Simpson, I.J., Spahni, R., Steele, L.P., Strode, S.A., Sudo, K., Szopa, S., van der Werf, G.R., Voulgarakis, A., van Weele, M., Weiss, R.F., Williams, J.E., Zeng, G., 2013. Three decades of global methane sources and sinks, *Nature Geoscience* 6, 813–823, doi:10.1038/ngeo1955.

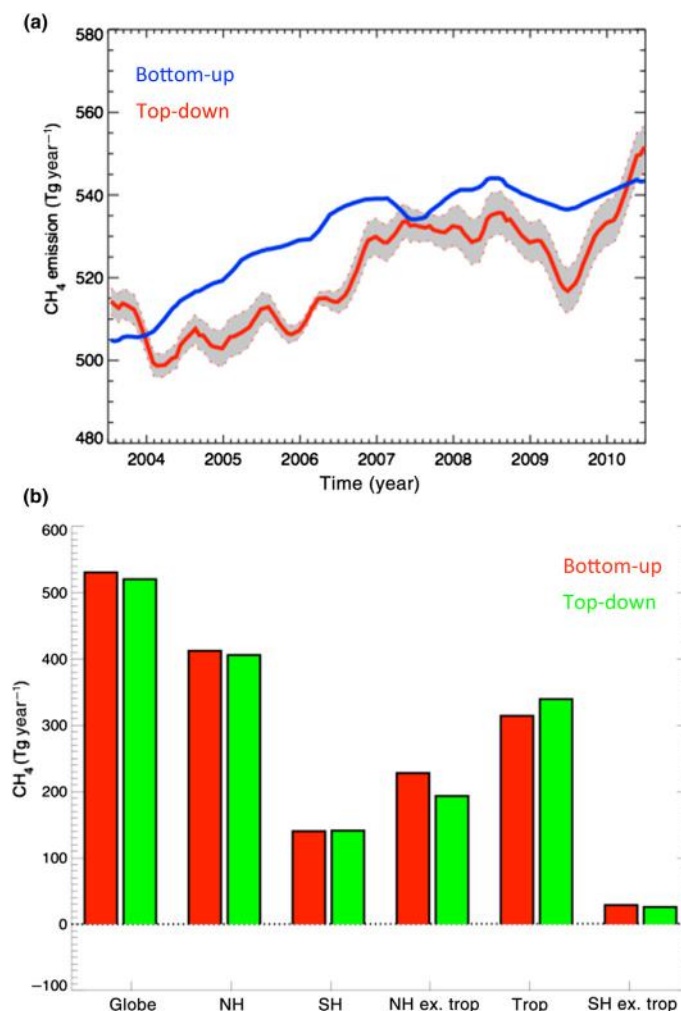


Figure 7 (a) The recent evolution of global methane emissions comparing estimates from the bottom-up and top-down approaches. (b) The large-scale distribution of methane emissions averaged over the period 2003–10, comparing bottom-up and top-down approaches. Top-down estimates are derived from the TM5-4DVAR inverse modeling framework using measurements from 46 sites of the National Oceanic and Atmospheric Administration–Earth System Research Laboratory (NOAA–ESRL) global cooperative air sampling network.

which is in line with the observed increase in global CH₄ since 2007. Large-scale characteristics of the average horizontal distribution of the emissions are presented in [Figure 7\(b\)](#), highlighting the importance of the tropics (30° S–30° N). The atmospheric measurements cause a shift of about 30 Tg CH₄ year⁻¹ from the NH extra-tropics (30–90° N) to the tropics (30° S–30° N).

As illustrated by [Figure 7](#), inverse modeling using background measurements of the global surface monitoring network provides useful constraints on the global methane budget and its variation on interannual timescales and longer. However, the information is generally insufficient to resolve regional sources and sinks since the influence of emissions on atmospheric abundance is quickly attenuated by atmospheric mixing as the distance of the measurements from the sources increases. The solution to regional-scale flux estimation is to make use of high frequency *in situ* measurements from tall towers situated inside the

region of interest. In this case, high-resolution mesoscale models are used that are capable of simulating the large observed variability. Furthermore, a sufficiently dense network of tall towers is required which currently limits the application of this approach to Europe and the USA. An alternative approach to increase flux-resolving power in regions without sophisticated monitoring capabilities on the ground is to make use of satellites. As discussed earlier this approach is still in an exploratory stage because of the challenges to meet the required accuracy and precision. This is true for the measurements as well as for the models. Furthermore, this approach calls for sophisticated optimization methods capable of dealing with large datasets.

See also: Chemistry of the Atmosphere: Chemical Kinetics; Ion Chemistry; Isotopes, Stable; Principles of Chemical

Change. **Climate and Climate Change:** Volcanoes: Role in Climate. **Numerical Models:** Chemistry Models. **Ozone Depletion and Related Topics:** Photochemistry of Ozone. **Paleoclimatology:** Ice Cores. **Statistical Methods:** Data Analysis: Time Series Analysis. **Tropospheric Chemistry and Composition:** Hydroxyl Radical; Oxidizing Capacity.

Fung, I., John, J., Mathews, E., et al., 1991. Three-dimensional model synthesis of the global methane cycle. *Journal of Geophysical Research* 96, 13033–13065.
Khalil, M.A.K. (Ed.), 1993. *Atmospheric Methane Sources, Sinks, and Role in Global Change*. Springer, Berlin.
Khalil, M.A.K. (Ed.), 2000. *Atmospheric Methane: Its Role in the Global Environment*. Springer, Berlin.

Further Reading

Cicerone, R.J., Oremland, R.S., 1988. Biogeochemical aspects of atmospheric methane. *Global Biogeochemical Cycles* 2, 299–327.

Observations for Chemistry (*In Situ*): Ozone Sondes

HGJ Smit, Research Centre Jülich, Jülich, Germany

© 2015 Elsevier Ltd. All rights reserved.

Introduction

Ozone, although a minor constituent, plays a key role in the physics and photochemistry of the atmosphere. As an important absorber of both infrared and ultraviolet (UV) radiations, ozone is of crucial importance for Earth's climate but also as a UV filter for the biosphere. Several methods are available for observation of the vertical distribution of atmospheric ozone. All remote methods, ground-based or space-borne, use optical techniques, while *in situ* methods, (chemical or optical) are made from aircraft, balloon, or rocket platforms. A widely used method for measuring upper-air ozone *in situ* up to altitudes of 30–35 km is through small, lightweight, and compact balloon-borne sondes. The reaction of ozone with potassium iodide in aqueous solution is used to measure the ozone concentration continuously in an electrochemical cell. The sensing device is interfaced to a standard meteorological radiosonde for data transmission to the ground station and can be flown on a small rubber weather balloon (Figure 1).

The electrochemical sondes were developed in the 1960s, primarily to measure ozone in the study of the large-scale dynamics in the stratosphere. Since about 1970, with increased awareness of the photochemical depletion of stratospheric ozone by chlorofluorocarbons and the photochemical



Figure 1 Ozone sounding just after launch at Meteorological Observatory Hohenpeißenberg, Germany. Courtesy of the German Weather Service.

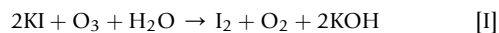
increase of ozone in the troposphere, regular ozone soundings have been made in a global network to observe long-term changes of ozone. Ozone sounding records provide the longest time series of the vertical ozone distribution between the surface and 30–35 km altitude. Up to an altitude of 20 km, ozone sondes constitute the single data source with long-term coverage for the derivation of ozone trends with sufficient vertical resolution, particularly in the altitude region around the tropopause. In addition, ozone sondes are also deployed to study photochemical and dynamical processes in the atmosphere or to validate satellite observations.

Three major types of ozone sondes – Brewer–Mast (BM), electrochemical concentration cell (ECC), and the carbon–iodine (KC96) – are in use since the early 1970s. Each sonde type has its own specific design, and even small differences in instrumental parameters and their uncertainties can have significant effects on the performance of the different sonde types. To assess the performance of the sondes and to quantify any systematic differences between the various sonde types, several intercomparison studies have been carried out since 1970. Most of these studies were based on dedicated short-term intercomparison campaigns in the field and focused on the performance of the sondes in the stratosphere. In addition, comparisons of time series of ozone sonde data with other simultaneously operating ozone profiling devices such as lidar, microwave, or satellites are used to the data quality of long-term ozone sonde records to derive ozone trends.

Instrumental Description

Principle of Operation

The principle of ozone measurement by the electrochemical sonde is based on the titration of ozone in a potassium iodide (KI) sensing solution according to the redox reaction [I]. (An exception is the carbon–iodine ozone sensor, which uses potassium bromide (KBr) instead of KI.) However, the principle of operation is via a reaction mechanism similar to the redox reaction [I].



The titration involves a conductometric method employing electrochemical reaction cells to determine the amount of generated ‘free’ iodine (I_2) per unit time through conversion into an electric current at a depolarizing electrode. Continuous operation is achieved by a small electrically driven gas sampling pump that forces ambient air through the sensing solution of the electrochemical cell. Transported by the stirring action of the air bubbles, the iodine makes contact with a platinum cathode and is reduced back to iodide ions by the uptake of two electrons per molecule of iodine as in reaction [II].



An electric current I_M (μA) generated in the external circuit of the electrochemical cell is directly related, after correction for a background current I_B (μA), to the uptake rate of ozone in the sensing solution. Knowing the gas volume flow rate Φ_P ($\text{cm}^3 \text{s}^{-1}$) of the air sampling pump, its temperature T_P (K) and the conversion efficiency of the ozone sensor η_C , the measured partial pressure of ozone P_{O_3} is determined from eqn [1].

$$P_{\text{O}_3} = \frac{R}{2 \cdot F} \frac{T_P}{(\eta_C \cdot \Phi_P)} \cdot (I_M - I_B) \quad \text{with} \quad \frac{R}{2 \cdot F} = 0.043085 \quad [1]$$

where R is the universal gas constant and F is Faraday's constant. The number 2 originates from the fact that each molecule of iodine (I_2) formed from the reaction of $\text{KI} + \text{O}_3$ will be converted back into 2 iodide (I^-) ions and deliver thereby 2 electrons in the external electrical current circuit to contribute to the measured electrical current I_M . In practice the conversion efficiency η_C is assumed to be unity. In principle, this type of electrochemical ozone sensor is an absolute measuring device. Although the principle of operation is similar for all three sonde types, the instrumental layouts have significant differences.

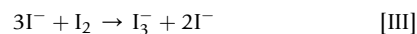
ECC Ozone Sonde

The ECC ozone sonde was developed by Komhyr in 1969. The ECC ozone sensor is an electrochemical cell consisting of two half-cells, made of Teflon, which serve as cathode and anode chambers, respectively (Figure 2). Both half-cells contain platinum mesh electrodes. They are immersed in KI solutions of differing concentrations. The two chambers are linked together by an ion bridge in order to provide an ion pathway and to prevent mixing of the cathode and anode electrolytes.

ECC sensors do not require an external electrical potential, in contrast to the Brewer–Milford type of electrochemical ozone sensor. The difference in the concentrations of the KI

solutions in the cathode and anode chambers, 0.5–1% KI and saturated KI, respectively, delivers the electromotive force for the ECC.

A small chemically inert gas sampling pump made of Teflon forces ozone in ambient air through the cathode cell, containing the KI-sensing solution of lower concentration, and causes an increase of 'free' iodine (I_2) according to the redox reaction [I]. At the surface of the Pt cathode, I_2 will be converted to I^- through the uptake of two electrons (reaction [II]), while at the anode surface, I^- is converted to I_2 through the release of two electrons. The overall cell reaction is shown in reaction [III].



Thus, one ozone molecule causes two electrons to flow in the external circuit. The electric current is thus directly related to the uptake rate of ozone in the cathode chamber. The ECC ozone sonde is displayed in Figure 2. The instrument, size about $8 \times 8 \times 14 \text{ cm}$, is enclosed in a Styrofoam flight box ($19 \times 19 \times 25 \text{ cm}$). The ECC ozone sonde is nowadays the most widely used sonde type.

BM Ozone Sonde

The oldest ozone sonde type still in routine operation is the BM sonde, which evolved from the Oxford–Kew ozone sonde developed by Brewer and Milford in 1960. The BM-ozone sensor consists of a single electrochemical cell with a silver anode and platinum cathode immersed in an alkaline potassium iodide solution. A polarizing potential of 0.41 V is applied between the electrodes such that no current will flow unless free iodine is present. In operation, ozone in the sampled ambient air is forced through the sensing solution in the electrochemical cell (bubbler) to produce free iodine according to the redox reaction [I]. At the surface of the Pt cathode, I_2 will be converted to I^- through the uptake of two

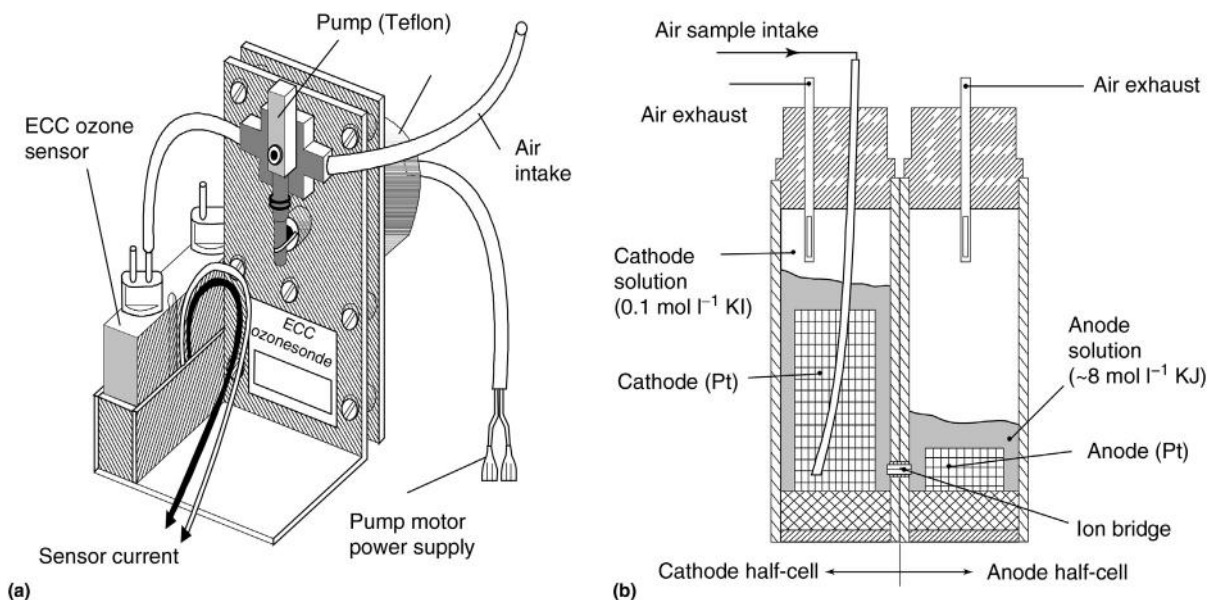
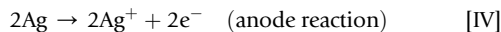


Figure 2 Schematics of the electrochemical concentration cell (ECC) ozone sonde (a) and sensing cell (b).

electrons (reaction [III]), while at the Ag anode surface two electrons are released through the ionization of two silver atoms to form the insoluble silver iodide (reaction [IV]).



In principle, each ozone molecule entering the sensor causes a current of two electrons to flow through the external circuit.

The most recent version (type 730-10) of the original BM sonde is shown in Figure 3. The reaction chamber (bubbler) is made of clear acrylic glass and contains a cylindrical platinum mesh cathode ($\sim 6 \text{ cm}^2$) and a thin silver wire as anode. The bubbler is filled with 2 cm^3 of neutrally buffered aqueous solution of potassium iodide (0.1%). The electrically driven gas sampling pump is mounted at the right side of the bubbler and forces about $220 \text{ cm}^3 \text{ min}^{-1}$ of ambient air through the bubbler. The sonde is protected by a Styrofoam flight box. Older ozone sounding stations mostly used the BM-sonde type. Most of the stations that are still operational have changed to the ECC ozone sonde type. With an exception of the Meteorological Observatory Hohenpeissenberg (Germany) where BM-ozone sondes are flown since the late 1960s.

A hybrid of a Brewer–Milford type ozone sensor made of acrylic and a nonreactive Teflon pump is manufactured and flown by the Indian Meteorological Department.

Carbon–Iodine Ozone Sonde (KC96)

The KC96 ozone sonde type is a modified version of the KC79- and the earlier KC68-sonde, which was developed by Kobayashi and Toyama in 1966. These sonde types are based on the carbon–iodine ozone sensor type developed in the early 1960s. The ozone sensor is a single electrochemical cell containing a platinum gauze cathode and an activated carbon anode immersed in an aqueous neutral potassium bromide solution. Ozone in ambient air is forced through the sensing

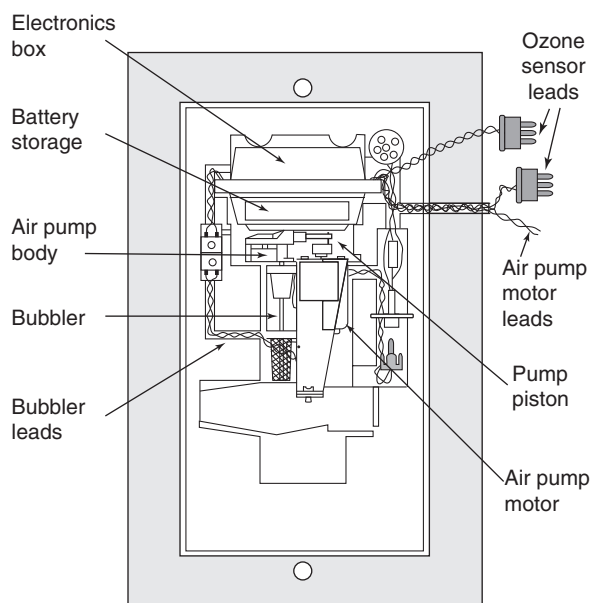


Figure 3 Schematic of the Brewer–Mast (BM) ozone sonde.

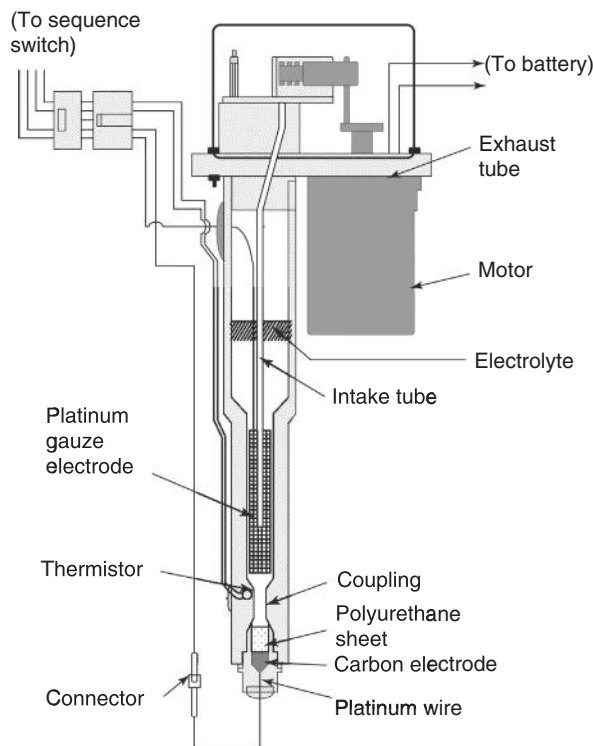
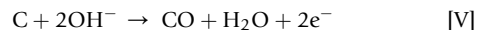


Figure 4 Schematic of the carbon–iodine ozone sonde (KC79).

solution, generating ‘free’ bromine molecules (Br_2), similarly to the redox reaction [I]. At the Pt cathode, the bromine is reconverted into bromide (Br^- ions) by the uptake of two electrons, while at the activated carbon anode with the corresponding release of two electrons reaction [V] takes place.



Accordingly, one ozone molecule produces an electric current of two electrons in the external circuit.

A scheme of the sonde is shown in Figure 4. The gas sampling pump and the electrochemical cell are made of methacrylate resin. The pump flow rate is about $400 \text{ cm}^3 \text{ min}^{-1}$, with the pump motor speed being held constant by a governor. The sonde is enclosed in a Styrofoam flight box. The KC sondes were majorly flown by the Japanese Meteorological Agency in their ozone sonde network until 2011. The network deploys now also ECC sondes.

Vertical Ozone Sounding

The set-up of an ozone sounding operation is shown in Figure 5. During normal flight operation, ozone sondes are coupled via special interfacing electronics with standard meteorological radiosondes for data transmission of the measured sensor current and pump temperature plus additional measurement of aerological parameters such as pressure, temperature, and humidity (and, optionally, wind direction and speed). Using the telemetry of the radiosonde, the data measured by the sonde are transmitted to the ground station for further processing.

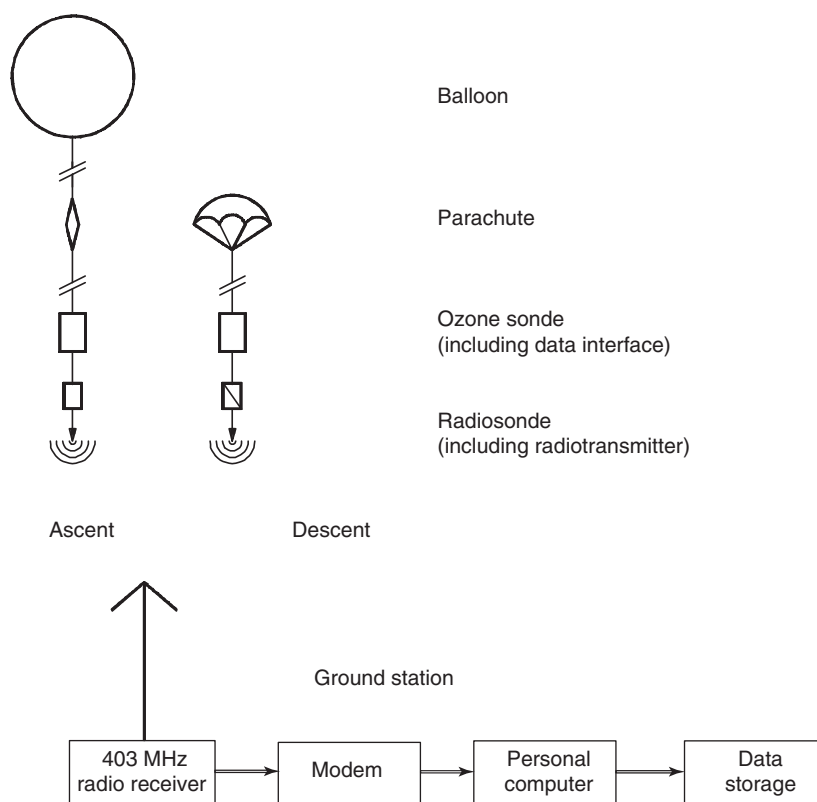


Figure 5 Set-up of ozone sounding system.

The total weight of the flight package is typically about 1 kg and it can be flown on small weather balloons (Figure 1). Normally data are taken during ascent at a velocity of about 5 m s^{-1} to a balloon burst altitude of 30–35 km. The inherent response time of the ozone sonde is 20–30 s, such that the effective height resolution of the measured vertical ozone profile is 100–150 m.

Figure 6 shows some examples of vertical ozone sounding profiles obtained at a midlatitude site (Jülich, Germany), over the tropical Atlantic (RV *Polarstern*), and at the Antarctic (South Pole). Ozone sondes can be flown under almost all weather conditions, even under heavy cloud or in rainy conditions where optical profiling techniques are rather limited in their capabilities.

Factors Influencing Sonde Performance

Each ozone sounding is made with a new instrument, which therefore has to be characterized properly prior to flight. Consistency of instruments with regard to their quality and characteristics, as well as standardization of operating procedures, is a prerequisite to assure consistent sonde measurements. Several instrumental and procedural parameters (see also eqn [1]) and their uncertainties can have a substantial influence on the quality of the ozone sonde measurements. Changes of these parameters through changes in instrument, operating procedures, or environmental conditions can have a significant impact on the long-term

ozone trends derived from ozone sonde measurements. From intercomparisons between different sounding stations using the same sonde type, it has been shown that the observed differences are for the most part due to the differences in the preparation and correction procedures applied at the different sounding stations. Fortunately, some of the instrumental factors with the potential to influence the observed ozone trends involve postflight data processing, and the data may subsequently be reevaluated when the influences of these instrumental factors and their uncertainties are better understood.

Temperature of Gas Sampling Pump

To correct for changes of the air mass flow rate through the sensor due to temperature changes, the actual pump temperature is measured in-flight either inside the pump or in the instrument enclosure (Styrofoam box). Over the course of a sounding, the pump temperature can decline by typically 10–25 °C. However, in most of the older sounding systems used before the 1990s it was not possible to measure the actual pump temperature owing to the limited number of signals the analog operating radiosondes were able to transmit. Therefore, it was a common practice for either a constant pump temperature or an empirical table of the pump temperature as a function of ambient air pressure to be applied. This procedure can introduce uncertainties of 1–7% in the ozone computations, particularly for the highest altitudes.

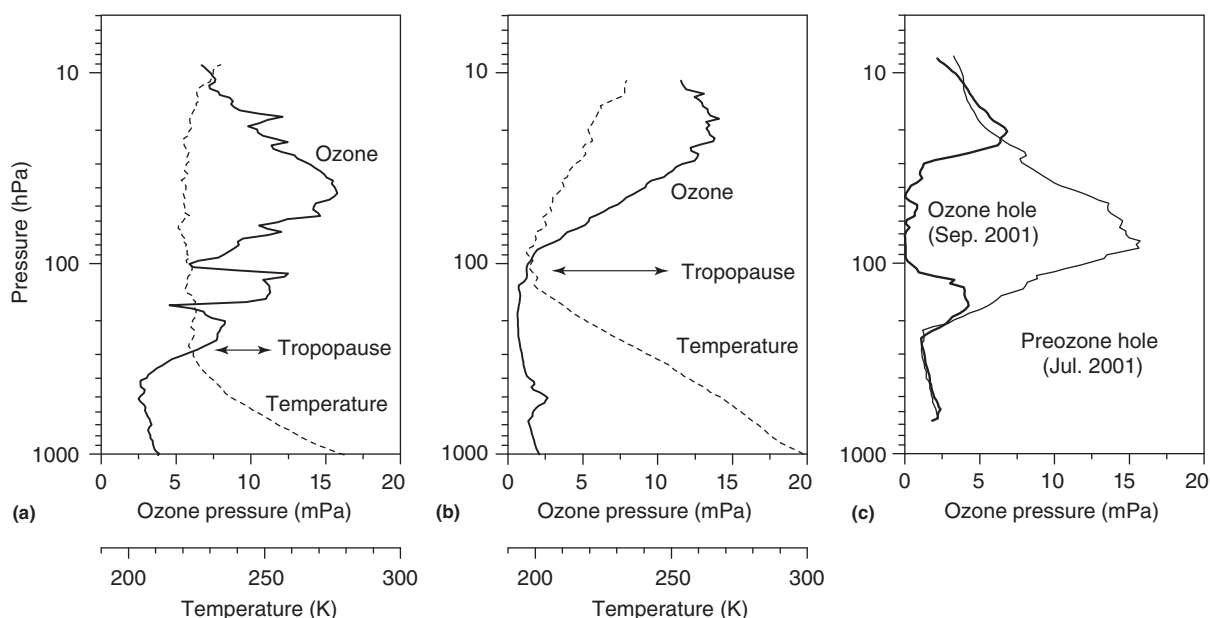


Figure 6 Vertical ozone sounding profiles: (a) midlatitude (Jülich, Germany, 51° N 6° E, March 1994); (b) tropical Atlantic (RV *Polarstern*, 6° N 30° W, September 1988); (c) South Pole (preozone hole in July 2001 and ozone hole in September 2001). Courtesy of NOAA/CMDL.

Pump Flow Efficiency

The volumetric flow of the gas sampling pump of each sonde is measured individually before flight. At ambient air pressures below 100 hPa the efficiency of the gas sampling pump degrades, and this is corrected for by applying a correction table of average pump efficiency as a function of ambient pressure specific for each sonde type. Typical correction factors as a function of ambient pressure for each ozone sonde type are listed in Table 1. The correction tables are based on empirical averages obtained from pump flow efficiency measurements made at different air pressures in the laboratory. The uncertainty of the tabulated correction factors increases substantially at pressures below about 20 hPa, which can contribute significantly to the overall uncertainty of the sonde performance above 25–30 km altitude. The uncertainty of the pump flow efficiency is the major contribution to the overall uncertainty of the sonde measurements above 25 km altitude.

Table 1 Pump flow efficiency correction as function of ambient pressure for the three ozone sonde types as obtained from laboratory measurements

Pressure (hPa)	Correction factor		
	BM sonde	ECC sonde	KC79/96 sonde
1000	1	1	1
200	1	1	1
100	1.01	1.01	1.02
50	1.03	1.02	1.04
30	1.07	1.03	1.07
20	1.09	1.04	1.11
10	1.17	1.06	1.25
7	1.24	1.08	1.40
5	1.30	1.10	1.66

Background Current

Prior to launch of the balloon, the background current of each ozone sonde is determined individually, except in the BM sondes for which prior to flight the sonde readings are electronically compensated for the background current. Background signals determined at the surface are typically in the range of 0.1–0.5 mPa ozone partial pressure equivalent. Ozone soundings are thus sensitive to errors in the background signal correction in regions where the ozone concentration is low, i.e., in the upper troposphere. Such errors have the potential to become large if the background signal is similar in magnitude to the ozone signal. This is particularly the case in the tropics (Figure 6(b)), but also in the Polar Regions when the stratospheric ozone hole occurs (Figure 6(c)). For the conventional method of background correction it is assumed that the background current is proportional to the oxygen partial pressure, so that this offset gradually declines with decreasing pressure and is vanishingly small in the upper troposphere and stratosphere. However, laboratory studies do not show any oxygen dependence of the background current and it is more appropriate to use a constant background current correction throughout the entire sounding profile. The origin of the background signal is not really understood. The timed background measurement during the preflight preparation is directly correlated to the ozone exposure of the sensing cell. Therefore, the background current is most likely not a time invariant property of the electrochemical cell but the result of a minor but still slowly decaying contribution to the measured cell current caused by an additional minor and slow reaction pathway of the chemical oxidation of ozone in the cathode sensing solution. However, more research is needed to resolve the problem of ECC-sensor response with the fast and slow signal component.

Conversion Efficiency

The conversion efficiency includes both the efficiency of absorption of O_3 into the sensing solution and the stoichiometry of the conversion of O_3 into I_2 . During normal operation both parameters are unity, so that the conversion efficiency is also assumed to be unity. However, in the course of a sounding, the uncertainties of the sensor cell characteristics can increase. For example, a certain percentage of the sensing solution evaporates at a rate dependent on the temperature of the cell and the ambient pressure encountered during the sounding. For the ECC sondes, this means that the concentration of the sensing solution increases, which can enhance the sensitivity of the ECC sensor and can increase the uncertainty of the ozone measurement.

For ECC sondes, this increase of sensitivity is significant in presence of the 'pH-buffering' phosphate chemicals. Usually, the sodium-hydrogen phosphate buffer is added to the cathode sensing KI solution to keep the pH neutral at 7.0. However, the buffer can also be the cause for the controversy of yielding a stoichiometric factor (I_2/O_3) larger than unity. The reaction mechanism and stoichiometric factor of the net iodimetric reaction [I] has been studied by many investigators using a variety of KI solutions and pH-buffers. It appears that depending on the concentrations of KI and the pH-buffer, the stoichiometric factor can expand up to 1.25. For BM or KC sondes, this effect is negligible because these sondes operate at much lower concentrations of sensing solution.

When ECC sondes of the same type are operated with the different cathode sensing solution strength that are nowadays applied, it has been shown that significant differences (~ 5 – 10%) in the ozone readings can occur. In addition, the performance characteristics of ECC sondes from different manufacturers can be significantly different ($\sim 5\%$), even when operated under the same conditions.

Another source of uncertainty is the influence of local air pollution, which can have detrimental effect on the conversion efficiency, i.e., on sensor performance. Ozone measurements by a KI method, as in the electrochemical ozone sensor, are sensitive to interferences by oxidizing or reducing agents. However, several laboratory experiments have shown that for moderate polluted air, sulfur dioxide (SO_2) is normally the only trace gas that can produce significant negative interferences in the measurements of ozone. As a reducing agent, SO_2 converts the iodide produced from the ozone-iodine reaction back to iodine. A memory effect can occur if excess of SO_2 is accumulated in the sensing solution of the sensor, which can affect measurements not only in the polluted boundary layer itself but also up to 1–2 km above.

Total Ozone Normalization

Optionally, ozone sonde profiles are normalized so that the integrated ozone column obtained from the sonde profile plus an estimated residual ozone column above the burst altitude is adjusted to an independent nearby total ozone column measurement by, e.g., a Dobson or Brewer spectrophotometer. Standard procedure for calculating the residual ozone column is to assume a residual column with constant mixing ratio

equal to the measured value at the top of the sonde profile. Uncertainties in the sonde readings of pressure and ozone at the burst altitude, and also the assumption of constant mixing ratio, can introduce significant uncertainties in the estimation of the residual ozone column. This can be improved by an alternative method in which a climatology of residual ozone column data obtained from satellite observations is used. However, the influence of uncertainties in pressure readings at the bursting point remains.

The use of normalization to correct the profile conflicts with the fact that the electrochemical ozone sonde is, in principle, an absolute measuring device. However, even if the total ozone normalization factor is not used to correct the sonde profile, it provides a screening test for unreliable soundings using the criterion that the normalization factor may not deviate more than about 10–20% from unity. However, a normalization factor of unity is not a guarantee that the profile is correct. In routine operation, the normalization factors for ECC sondes are in the range of 0.9–1.1, while BM and KC sondes show normalization factors of 0.8–1.2.

Radiosonde Pressure and Temperature

Errors in radiosonde pressure or temperature measurements will imply corresponding errors in calculated geopotential heights, causing measured ozone concentrations to be assigned to incorrect altitudes and pressures. This is potentially an important issue for the derivation of trends, as radiosonde changes may therefore introduce vertical shifts in the ozone profile, and apparent changes in ozone concentration at a given height. A number of different radiosonde designs, from several manufacturers, have been used in the global observing network over the last four decades. At pressures below 50 hPa, significant bias effects of 5–10% in the ozone profile can occur, particularly for the radiosonde types used before 2000. Through the use of the Global Positioning System, modern radiosondes can measure geopotential heights with a higher accuracy. This will reduce any bias effects in the measured pressure, i.e., ozone profile significantly.

Precision and Accuracy

To quantify the precision and accuracy of the three different types of ozone sondes, several comparison studies of sondes with other ozone profiling techniques have been made since 1970. Most of the intercomparisons have been conducted in the field to assess the sonde performance up to 30–35 km altitude. However, short-term intercomparisons are more or less 'snapshots' and may not necessarily reflect the performance of ozone sondes under operational field conditions. Comparison studies of time series of ozone sonde data with other simultaneously operating ozone monitoring methods such as lidar or microwave are more suitable for assessing the data quality of the ozone sonde measurements in regular operation. In addition, intercomparisons conducted in a controlled environmental chamber capable of simulating real sounding conditions, where a UV photometer serves as reference, allow specific questions arising from field intercomparisons to be addressed.

Owing to the much lower concentrations of ozone in the troposphere compared to the stratosphere, the performance of the sondes and the typical instrumental and operational factors determining precision and accuracy are rather different in these two regions of the atmosphere. However, most intercomparison studies, particularly before 1990, focused exclusively on the sonde performance in the stratosphere and only a few intercomparison studies have addressed the sonde performance under typical tropospheric conditions.

Stratospheric Performance

In general, all intercomparison studies, short-term as well as long-term, have indicated that in the lower to middle stratosphere between the tropopause and ~ 28 km, the three different sonde types show consistent results provided the individual measured sonde profiles have been normalized to ground-based total ozone column measurements at the launch site. This can be understood from the fact that the normalization is mainly weighted to the ozone in the lower stratosphere, which contains most of the column ozone. In this altitude range, the precision of the various sonde types is within $\pm 3\%$, while any systematic biases compared to other ozone-sensing techniques are smaller than $\pm 5\%$.

For altitudes above 28 km, the results are not so conclusive and the measurement behavior of the different sondes is different and cannot be generally characterized. The BM sonde used by the established (long-term record) stations (e.g., Hohenpeissenberg) show systematic underestimations of ozone that increase with altitude (-15% at 35 km). For the ECC sondes, there is some evidence suggesting that measurements agree with each other and with reference techniques to within $\pm 5\%$. The Japanese KC68/79/96 tends to overestimate ozone above 30 km. The data quality of sondes above 28 km is strongly influenced by the performance of the air sampling pump and its efficiency decreases at lower pressures. However, most of the intercomparison studies show that the performance of the ECC sondes between 28 and 35 km is still rather good and even tends to overestimate the ozone concentration compared to lidar measurements. There are some experimental indications that for ECC sondes there is probably a compensating effect due to evaporation of the KI-sensing solution, which will cause an increase in concentration and may result in a higher sensitivity of the ozone sensor. However, in general the sonde data above about 28 km are less reliable and should be used with caution, at least for the non-ECC types.

Tropospheric Performance

There is a dearth of sonde validation studies for the troposphere and, because of the small number of comparisons, only estimates of the reliability of the sonde data records below the tropopause can be made. Further, because ozone values are much lower in the troposphere than in the stratosphere, the impact of instrumental errors and variability is larger. Intercomparison campaigns between 1970 and 1990 have shown systematic differences between sonde types, typically varying from 10 to 15%. Campaigns conducted after 1990 have shown that the BM and KC79/96 sondes are less precise than the ECC types and that the ECC sondes deliver much more consistent

results than the other two types of sondes. The precision of the ECC sonde is better than $\pm(5$ to $10)\%$ and shows a small positive bias of about 3%. The BM and KC79/96 sonde showed precisions in the range of $\pm(10$ to $20)\%$, but there are no indications of any bias larger than $\pm 5\%$. Key issues of uncertainty are the background correction and the use of the total ozone normalization factor.

Changes of Instruments or Operating Procedures

Apart from a few exceptions, nowadays, ECC ozone sondes are predominantly flown within the global ozone-sounding network or deployed in scientific field experiments. Laboratory and field experiments have demonstrated that small changes of the ECC-sonde instrument or operating procedures can significantly influence its performance. For example, the performance characteristics of two different ECC-sonde types, manufactured by either Science Pump Corporation (Model type: SPC-6A) or Environmental Science Corporation (Model type ENSCI-Z), can be significantly different, even when operated under the same conditions. Different ozone readings were obtained when sondes of the same type are operated with different cathode sensing solutions. This means that for ozone sounding stations performing long-term measurements, a change of the sensing solution type or ECC-sonde type can easily introduce a sudden step of $\pm 5\%$ or more in their records, affecting the determination of ozone trends. Existing artifacts in long-term sounding records caused by changes of the instrument or operating procedures can be resolved by homogenization either in space (between different stations) or in time (long-term changes). This can be achieved through use of generic transfer functions, which can be derived from intercomparison experiments or dual balloon soundings.

See also: Chemistry of the Atmosphere: Principles of Chemical Change. **Ozone Depletion and Related Topics:** Long-Term Ozone Changes; Ozone Depletion Potentials; Ozone as a UV Filter. **Radiation Transfer in the Atmosphere:** Radiation, Solar. **Satellites and Satellite Remote Sensing:** Measuring Ozone from Space – TOMS and SBUV.

Further Reading

- Brewer, A., Milford, J., 1960. The Oxford Kew ozone sonde. *Proceedings of the Royal Society of London Series A* 256, 470–495.
- Komhyr, W.D., 1969. Electrochemical concentration cells for gas analysis. *Annals of Geophysics* 25, 203–210.
- Schenkel, A., Broder, B., 1982. Interference of some trace gases with ozone measurements by the KI-method. *Atmospheric Environment* 16, 2187–2190.
- Smit, H.G.J., Straeter, W., Johnson, B., Oltmans, S., Davies, J., Tarasick, D.W., et al., 2007. Assessment of the performance of ECC-ozone sondes under quasi-flight conditions in the environmental simulation chamber: Insights from the Jülich Ozone Sonde Intercomparison Experiment (JOSIE). *Journal of Geophysical Research* 112, D19306. <http://dx.doi.org/10.1029/2006JD007308>.
- World Meteorological Organization, 1998. SPARC-IOC-GAW Assessment of Trends in the Vertical Distribution of Ozone, SPARC Report No. 1, WMO Global Ozone Research and Monitoring Project Report No. 43WMO. Vienna.
- World Meteorological Organization, 2013. Quality Assurance and Quality Control for Ozone Sonde Measurements in GAW, WMO Global Atmosphere Watch Report Series, No. 201. World Meteorological Organization, Geneva. http://www.wmo.int/pages/prog/arep/gaw/documents/GAW_201.pdf.

Observations for Chemistry (*In Situ*): Particles

T Deshler, University of Wyoming, Laramie, WY, USA

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

The principles behind the large suite of instruments available for *in situ* particle observations are described. Instruments use condensation, optical scattering, electrical mobility, direct capture, absorption, and dissociation to provide highly resolved space and time measurements of aerosol number, size, mass, composition, optical properties, shape, and nucleating characteristics, when a full suite of instruments is available. Such measurements are required to fully characterize and understand the role of particles in certain aspects of heterogeneous chemistry, radiative transfer, cloud formation, and precipitation.

Introduction

History

The scientific observation, and discovery, of particles or liquids suspended in air, or aerosol, began with Tyndall in the 1850s through his observation of forward scattering from particles in direct sunlight. This led to Tyndall's further investigation of smaller particles by dark field microscopy, establishing the basis for nephelometers, ultra microscopes, and optical particle counters. At nearly the same time the work of Epsy, Coulter, and Aitken led to the realization that particles were required for the condensation of water in air and that this could be used to measure concentration, leading to condensation particle counters. Human interest in aerosol significantly preceded these discoveries, due to the obvious impact of aerosol on life through air pollution: the foul air of Rome, coal burning in London, mortality of hard rock miners. Thus, the first concerted measurements of aerosol, in the early 1900s, were focused on high concentrations of respirable particles which, when captured in alveoli, will lead to various forms of emphysema, depending on the contaminant particles: silicosis, tabacosis, and asbestosis. The natural role of particles in the formation of droplets and ice crystals in clouds, and their possible use in cloud seeding, was apparent by the end of the 1940s. The pivotal role of aqueous particles in mediating the atmospheric concentrations of reactive trace gases, in particular stratospheric ozone, became apparent in the 1980s. The interaction of particles with radiation was described well by the early 1900s; however, the impact of aerosol on the Earth's radiation balance has only recently been of significant interest as we strive to understand, and forecast, the impact of the increasing atmospheric abundance of carbon dioxide on the Earth's future. As the twenty-first century begins, the measurement of aerosol particles appears even more important than at the beginning of the twentieth century, when measurements began.

Aerosol Characteristics

The instruments and tools available today to measure aerosol are impressive: filters, electron microscopy, condensation particle counters, optical particle counters, nephelometers, mobility analyzers, impactors, mass spectrometers, lasers, fast microprocessors, and ample storage facilities. The task, however, is more impressive and no instrument can provide all the information desired. Atmospheric aerosols arise from the Earth's surface,

through combustion, mechanical disturbance, and gas to particle conversions from gaseous organic and volcanic emissions. Particle sizes range from <1 nm to >10 μm , concentrations from $>10^4$ cm^{-3} for small particles to $<10^{-3}$ cm^{-3} for large particles, and mass concentrations from <1 $\mu\text{g m}^{-3}$ to >1 mg m^{-3} . Particle shape varies and particles may be externally or internally mixed (Figure 1). Tropospheric particles contain a large fraction of the nonnoble elements in the upper half of the periodic table (Figure 2). The most common volatiles are water, organics, nitrates, sulfates, and ammonium. The most common crustal materials are aluminum, silica, and iron oxides. Aerosol particles provide cloud condensation and ice nuclei. Removal occurs through sedimentation and cloud processes. Residence times are typically <30 days in the troposphere, less than atmospheric mixing times, leading to large spatial and temporal structure. Stratospheric aerosols are well mixed, within a few years of penetrating volcanic eruptions, with residence times of years.

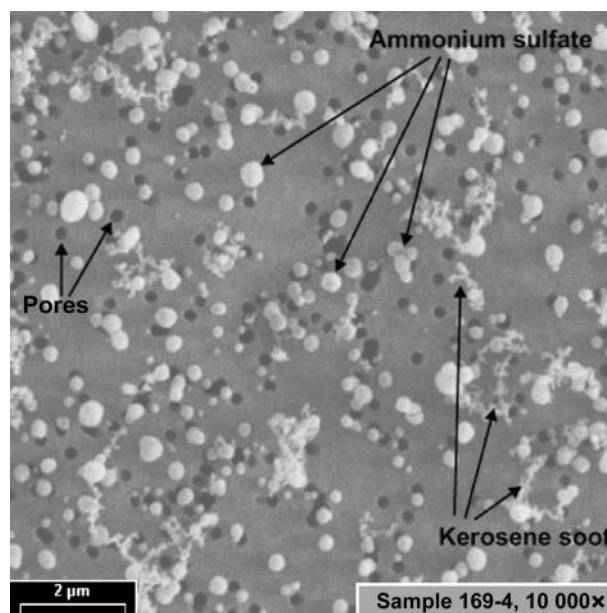


Figure 1 Electron micrograph of externally mixed ammonium sulfate and soot particles. From Sheridan, P., Arnott, W., Ogren, J., et al., 2005. The Reno aerosol optics study: an evaluation of aerosol absorption measurement methods. *Aerosol Science and Technology* 39, 1–16. Copyright 2005. Mount Laurel, NJ. Reprinted with permission.

H																	He				
Li	Be															B	C	N	O	F	Ne
Na	Mg															Al	Si	P	S	Cl	Ar
K	Ca	Sc	Ti	V	Cr	Mn	Fe	Co	Ni	Cu	Zn	Ga	Ge	As	Se	Br	Kr				
Rb	Sr	Y	Zr	Nb	Mo	Tc	Ru	Rh	Pd	Ag	Cd	In	Sn	Sb	Te	I	Xe				
Cs	Ba	*	Hf	Ta	W	Re	Os	Ir	Pt	Au	Hg	Tl	Pb	Bi	Po	At	Rn				
Fr	Ra	*																			
			La	Ce	Pr	Nd	Pm	Sm	Eu	Gd	Tb	Dy	Ho	Er	Tm	Yb	Lu				
			Ac	Th	Pa	U															

Figure 2 Elements observed in aerosol particles above 5 km. Frequencies are approximate because of differing ionization efficiencies, isotope signatures, and spectral interferences. From Murphy, D.M., Thomson, D.S., Mahoney, M.J., 1998. *In situ* measurements of organics, meteoritic material, mercury, and other elements in aerosols at 5 to 19 kilometers. *Science* 282, 1664–1669. Reprinted with permission from AAAS.

To completely characterize atmospheric aerosol would require time and space-resolved measurements of aerosol number, size, mass, composition, optical properties, shape, charge, and nucleating characteristics. No instrument is capable of this; however, individual *in situ* instruments and techniques are available which can measure each of these properties relying on condensation, optical scattering, electrical mobility, direct capture, and dissociation. *In situ* implies measurements at or in the immediate vicinity of an instrument. *In situ* aerosol measurements provide highly resolved space and time measurements, while suffering from severely limited space and time coverage, which are more easily obtained from remote aerosol measurements, such as lidar, limb extinction, or other electromagnetic-radiation-based sensors. *In situ* measurements, however, offer the only way to describe some of the specifics of aerosol populations required to fully characterize/understand heterogeneous chemistry, radiative transfer, cloud formation, and precipitation. *In situ* measurements play vital roles in improving our understanding of each of these fundamental areas.

Uncertainties

In situ instruments measure either an ensemble of particles or single particles, and are either extractive or noninvasive. Extractive instruments pull aerosol samples into a sampling chamber, while noninvasive instruments define a sampling volume external to the instrument, generally optically. Ensemble instruments are sensitive to the entire size distribution at once and thus measure directly integrals of the size distribution: mass, surface area, composition, absorption, scattering, or extinction. Single particle measurements provide size distributions of most of these quantities, and nucleating characteristics. Extractive instruments provide a larger suite of properties and are required for particles $<0.5\text{--}1.0\ \mu\text{m}$. All extractive instruments suffer from biases due to the inlets and tubing

leading to the sampling chamber. An ideal inlet has an internal air speed matching the speed of air flowing past the inlet, is straight, not too long, and does not heat the sample. As real inlets deviate from the ideal they may over- or under- count the edges of the size range; may lose particles in the inlet tube due to turbulent, inertial, electrostatic, or diffusional impaction; and particle size may change due to evaporation. Noninvasive instruments are required for particles $>10\ \mu\text{m}$, and are generally used for cloud particles. The illuminated volume defines the sample volume. Uncertainties arise from particles at the sample volume edge and from fragments of particles broken on the shrouds for the optical source and detector. Aside from uncertainties introduced by the inlet, or sample volume, all instruments have additional sampling biases related to sensitivity to specific size ranges, number concentrations, mass concentrations, and ionization thresholds. Each specific instrument has its own calibration and sampling efficiency challenges, and not all can be fully met, leading to instrumental limitations for many applications; see approximate size limitations in Figure 3. Adequate descriptions of these challenges and limitations cannot be accomplished here and will not be attempted. Instead the general principles underlying the instruments used to measure the wide array of aerosol properties will be focused on. The article will be organized around these aerosol properties. Table 1 summarizes the instruments and properties measured.

Although some aerosol sampling techniques, such as microscopy, spectroscopy, and gravimetry, provide primary data and are thus self-calibrating, most aerosol instruments require calibration against primary standards. Once instrument response functions are established partial calibrations are usually adequate for routine operation. The most fundamental primary standard is electron microscopy, for nonvolatile particles. Commercial primary standards include polystyrene latex beads of specific single sizes and aerosol generation and classification systems capable of supplying reliable monodisperse aerosol samples between 0.01 and $1.0\ \mu\text{m}$ for several aerosol compositions. Sample flow rates must also be carefully measured. However, even with careful calibration, the precision of most aerosol measurements is 20% or more, significantly higher than most gas phase measurements. The reason for this has less to do with the maturity of the instrumentation, than the complexity of aerosol particles compared to gas molecules. For gases, molecular composition, structure, size, shape, and mass are mostly well known. The primary measurement is concentration, and, at levels of parts per trillion, the number concentration of molecules exceeds total aerosol populations by factors of a thousand or more. Thus counting uncertainties are small. For aerosol particles counting uncertainties are large and are dominant for the largest particles, none of the properties listed above are known a priori, and often the particles are not well mixed so particle shape and composition will vary among the particles sampled, Figures 1 and 2, in addition to the normal variations in size and mass.

Chemical Properties

Composition

Aerosol mass spectrometers provide the biggest advances in the past 20 years for aerosol composition measurements. These

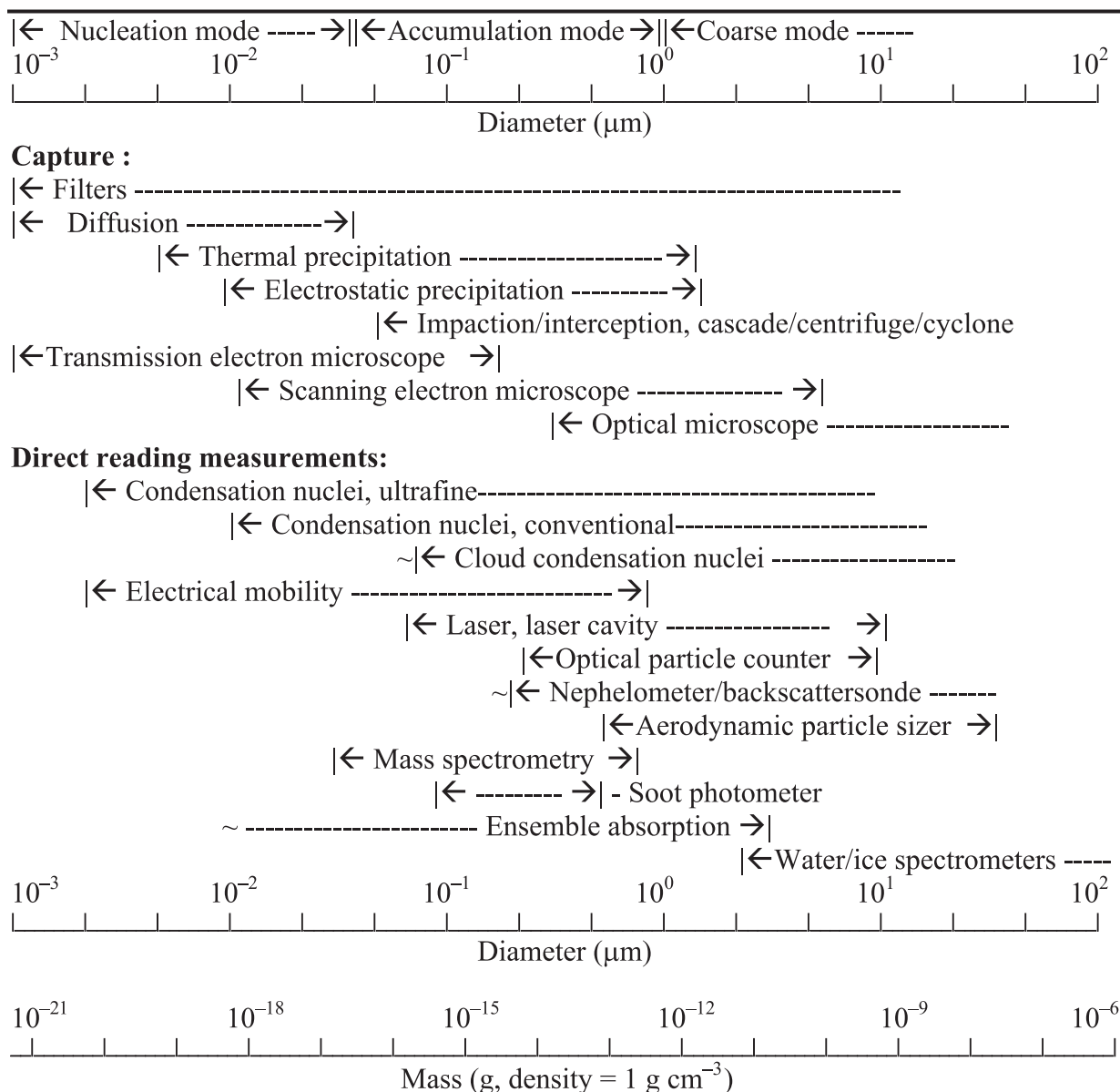


Figure 3 Size range for *in situ* aerosol measurement techniques as a function of diameter (μm). Interdecadal divisions on the log scales are at approximately 2 and 5. The size range for the nucleation, accumulation, and coarse modes typically assumed for atmospheric aerosol are shown at the top. Particle equivalent mass for a density of 1 g cm^{-3} is shown at the bottom.

single particle extractive instruments size and chemically analyze individual aerosol particles in real time, and several commercial instruments are available. Prior to this capability nearly all the chemical information was based on postanalysis of captured particles. Mass spectrometers consist of an evacuated inlet, which focuses particles into a central beam, a size determination, particle volatilization and ionization, and a mass to charge (m/z) detector, [Figure 4](#). The differences in these pieces define the different aerosol mass spectrometers available. Inlets may consist of single orifices, capillaries, or aerodynamic lenses, with the latter two capable of focusing a range of particle sizes. An aerodynamic lens is a series of

decreasing orifices, with each orifice concentrating successively smaller particles onto the central beam and is most common. Sizing is accomplished through either optical scattering or time of flight from the inlet exit to the target. Volatilization is accomplished by heating or laser ablation. Heating vaporizes only the volatile components, whereas laser ablation also provides refractory components. The vaporized molecules are ionized by electron impact, photoionization, or chemical ionization. The efficiency of ionization determines the quantitative nature of the measurement. Electron impact and chemical ionization are more well characterized, leading to more quantitative mass fractions when used. In contrast

Table 1 Aerosol properties measured by *in situ* aerosol instruments

Instrument	Physical properties					Chemical properties			Optical properties		
	Mass	No.	Size	Shape	SA	Crustal/elements	Volatile/ions	Carbon	Scat	Abs	Ext
Captured particles											
Gravimetry/microbalance	X										
Microscopy	I	X	X	X							
Colorimetry						X	X				
X-ray fluorescence						X					
γ -Ray emission						X					
Atomic spectroscopy						X					
Ion chromatography							X				
Infrared Raman spectroscopy							X				
Mass spectrometry						X	X				
Oxidation								X			
Direct reading single particle											
Condensation nuclei		X									
Cloud condensation/ice nuclei		X									
Electrical mobility	I	X	X		I						
Optical particle counter	I	X	X		I				I		
Soot photometer		X	X					X		I	
MS-laser ablation/ionization	I	X	X			X	X				
MS-volatilization, EI ionization	I	X	X				X				
Direct reading ensemble											
Nephelometer/backscattersonde									X		
Aethalometer										X	
Photoacoustic spectrometer										X	
Cavity ring-down spectrometer											X
Epiphaniometer					X						

MS, mass spectrometer; EI, electron ionization; SA, surface area; X, property measured; I, property implied; Scat, scattering; Abs, absorption; Ext, extinction.

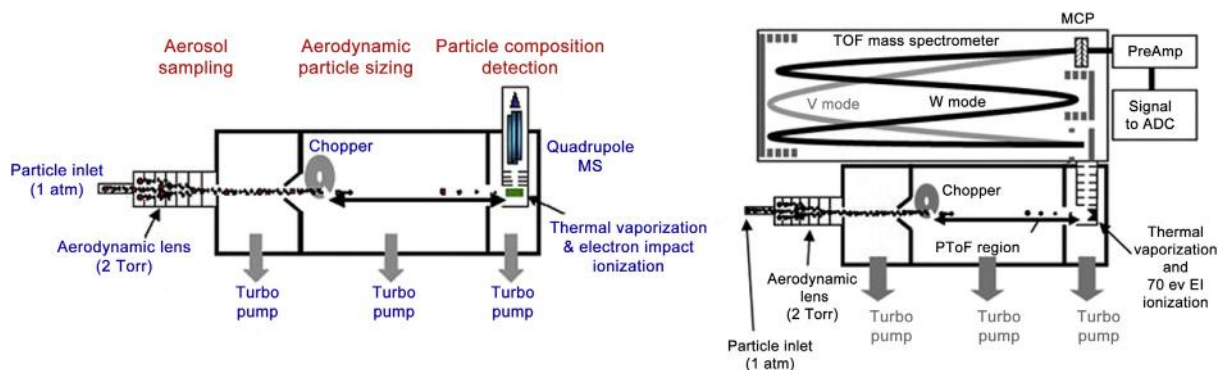


Figure 4 Schematic diagrams of two aerosol mass spectrometers showing an aerodynamic focusing lens, vaporization and ionization region, and mass spectrometer. The mass spectrometer on the left is a quadrupole and on the right a time of flight using reflector plates to produce either a V or W charged particle path. The figure on the left is reproduced with permission from M.L. Alexander, EMSL, Richland, Washington. The figure on the right is reprinted with permission from DeCarlo et al., Field-deployable, high-resolution, time-of-flight aerosol mass spectrometer. Analytical Chemistry 78, 8281–8289. Copyright 2006 American Chemical Society.

photoionization leads to less quantitative results. The ion detectors are primarily either quadrupole or time of flight. Quadrupole detectors focus ions of specific m/z ratios onto the electron detector. Only one m/z ratio can be sampled per particle due to timing limitations. Scanning through the m/z ratios for a series of particle impacts provides a quantitative mass spectra. Time of flight spectrometers use an electric potential to accelerate ions into a drift tube ending at the electron detector. The spectrum of arrival times provides the

mass spectrum since the drift velocity is mass dependent. The drift tube may be straight or bent to increase the length and thus resolution.

Prior to the advances in aerosol mass spectrometry chemical information was primarily limited to chemical analysis of particles captured on filters and more recently in high purity water. These ensemble invasive techniques can determine the elemental, ionic, and carbonaceous composition of aerosol, and are the only possibility for some applications. Elemental

analysis is accomplished by colorimetry, X-ray fluorescence, γ -ray emission, or atomic emission or absorption spectroscopy. Elemental concentrations are identified by spectroscopy of absorption or emission lines. Colorimetry uses wet chemistry to produce a solution whose light absorption is quantified. X-ray fluorescence and γ -ray emission use an energy dispersive detector to measure characteristic X-rays or γ -rays emitted from excited states of the elements. X-ray fluorescence is induced by exposure to X-rays or protons, γ -ray emission by exposure to neutrons. For emission and absorption spectroscopy, samples are dissolved and vaporized. Atoms excited by arcs, inductively coupled plasmas, sparks, or lasers will emit. Alternatively absorption of radiation by samples dissociated into ground state atoms in a flame can be measured.

Ions such as, SO_4^{2-} , NO_3^- , and NH_4^+ , major components of ambient aerosol, can be measured by colorimetry, ion chromatography, selective ion electrodes, infrared or Raman spectroscopy, or mass spectrometry. The first three techniques measure only the aqueous fraction of the aerosol. Ion chromatography, most widely used, separates sample ions within a column using ion exchange. Ion concentrations are measured with conductivity. Particle into liquid samplers inject the aerosol into a region of high water supersaturation forcing water to condense on the aerosol. The water droplets containing the aerosol are collected and inorganic ion fractions determined with ion chromatography. To remove gaseous contaminants the aerosol stream is usually preceded with a denuder.

Infrared or Raman spectroscopy uses infrared photons at energies characteristic of molecular vibrational bands, which are absorbed in proportion to molecular concentration. Visible photons, which inelastically (Raman) scatter from samples display frequency shifts characteristic of molecular vibrational bands. These techniques are also sensitive to organics and oxidized organics. For mass spectrometry, a laser volatilizes and ionizes the collected samples, or an ion beam removes, 'sputters,' ions off particle surfaces. Ion detection then proceeds as described above.

Carbon Content

Measurement of carbon in aerosol is another area with recent significant advancement, while remaining particularly challenging, due to the thousands of organic species, plus black carbon, which appear in particles. Mass spectrometers provide measurements of the volatile organics, but laser ablation is required for black carbon. Specific black carbon instruments are invasive and range from ensemble filter, cavity ring down, and photoacoustic, to single particle absorption. Total carbon can be determined by oxidizing filter samples and measuring the CO_2 . Techniques to separate total carbon into elemental and organic are attempted; however, there is no accepted standard. Techniques include slowly volatilizing organics or graduated chemical extraction. All other techniques rely on the absorption of light by carbon. Continuous measurements of light absorption by aerosol collected on filters provide good relative values but have large uncertainties. Cavity ring-down spectrometers trap photons in a sample chamber with highly reflective mirrors. Light intensity decays as the light is absorbed and scattered by the ensemble of particles in the cavity. The decay time then measures aerosol extinction of the light. To

obtain absorption requires an independent measure of scattering. Photoacoustic spectrometers measure absorption directly by measuring the acoustic waves resulting from heating of aerosol in the sample cavity by the absorption of light. The laser source is modulated at a characteristic frequency of the sample cavity, thus amplifying the sound for detection. The single particle soot photometer measures the incandescent black body radiation from particles, which are vaporized within the cavity of an infrared laser. Only absorbing particles will absorb the laser energy and become heated, with the time to vaporization, and the temperature, dependent on the carbon content. Particle size is measured simultaneously by scattering. This technique can be combined with mass spectrometric measurements of the incandesced particles for more detail.

Index of Refraction

Index of refraction is a function of aerosol composition, and if particle molar ratios are known, index of refraction can be derived. Another approach is to use size distribution measurements to calculate scattering for comparison with scattering measurements. The index of refraction is used to match the calculated and observed scattering. There have been a few attempts to measure particle index of refraction directly using optical instruments with detectors at several angles. Changes in index of refraction will lead to dissimilar changes in angular scattering, which should be sensible. These multiangle approaches have not proven to be very robust.

Physical Properties

Concentration

The first measurements of number concentration, the tyndallometer, and ultramicroscope were limited to particles large enough to scatter visible light, approximately $>0.3\ \mu\text{m}$, which misses the majority of aerosol. These instruments, however, were quickly followed by Aitken's condensation nuclei (CN) counter, variations of which remain the standard technique for number concentration. All CN counters consist of a saturator, condenser, and detector. The working fluid, water, butanol, and ethylene glycol, is evaporated into the aerosol stream in the saturator. Supersaturation occurs in the condenser and is achieved with either expansive or conductive cooling. Expansive cooling and water were originally used to create a cloud, which attenuates a light beam; however, none of these original instruments remain in service. Today, continuous flow instruments with conductive cooling supersaturation chambers are used. Typically air is saturated with *n*-butyl alcohol, at temperatures near $30\ ^\circ\text{C}$, before entering the condenser, held near $0\ ^\circ\text{C}$, where the air becomes rapidly supersaturated. Each particle acts as a condensation site and becomes a droplet large enough to be detected optically. The extent of supersaturation, controlled by the temperature difference between saturator and condenser, determines the minimum particle size, which will be condensed upon and thus be counted. Size distributions can be measured by preselecting particles with a diffusion battery or differential mobility analyzer, or by carefully controlling the supersaturation to activate successively smaller particles. An alternate total concentration measurement can be accomplished by charging

the aerosols with unipolar ions and then measuring the current as the aerosols are collected on a filter in a Faraday cage, assuming each particle carries one electron.

Size Distribution

There are a number of mature instruments to measure size distribution; however, only microscopy provides the physical or geometric size, but is quite laborious. The other methods provide either the optical or aerodynamic size. Optical size is provided by optical particle counters, among the first instruments to provide size distribution, which consist of a light source, dark field scattering chamber, and photodetector. Visible wavelengths were initially used, thus optical particle counters. Particle size is determined from the intensity of scattered radiation using Mie theory. To limit the influence of Mie resonances for particles near the wavelength of light, instruments use either a white light source or a large solid angle for detection of scattered light. Both approaches smooth the counter response function and limit the sensitivity to particle index of refraction. Sensitivity is determined by particle size for small particles and by particle concentration for large particles. Scattered radiation from small particles approaches Rayleigh scattering from molecules and therefore the detection limit. Particles greater than a few micrometers are rare, thus concentration uncertainties are large; however, the signal from one particle is well above instrumental noise. An alternate ensemble approach illuminates a sample with a laser and measures the scattering pattern, which has variations dependent upon average particle size and size distribution. The fluctuations arise from constructive and destructive interference of the scattering from individual particles as they move due to Brownian motion. This technique, known as photo correlation spectroscopy, measures the mean diffusion coefficient of the particles, and hence their mean size and a rough size distribution. The technique is sensitive to nanometer sizes. To achieve sufficient aerosol concentrations, samples are often captured in a liquid, producing a colloidal suspension. If particle concentrations are high enough, such as in flames or smokes, the technique can be used directly on aerosol samples as opposed to hydrosol samples.

Aerodynamic size is measured in several ways. Differential mobility analyzers or electrostatic aerosol spectrometers consist of an aerosol charging region, a drift tube with an applied electric field, and an outlet to a particle sensor. Charging regions usually contain sources of low energy electrons or alpha particles and produce bipolar equilibrium charge distributions which are a function of size. The particles then pass to a large diameter annular opening at the beginning of the drift tube. Particles with the appropriate charge polarity are attracted across a region of filtered air to a central rod held at a specific voltage. The particle drift velocity is inversely proportional to aerodynamic size. Only particles with the appropriate charge and aerodynamic drag to arrive at a small open annulus near the central rod will exit, providing a population of uniformly sized particles. The aerosol detector is usually a CN counter, but could be a filter and Faraday cage. Scanning the center rod across a range of voltages provides a differential size distribution if the size-dependent charge distribution is known, which also accounts for multiply charged particles. If the exit annulus is replaced by a large opening, all charged particles too large for collection on the central rod will

exit, providing a cumulative size distribution measurement. This approach was used by instruments called electrostatic aerosol analyzers. Final size distributions are available after an inversion. The lower size limit is constrained by inefficient charging for small particles, and the upper limit by the length of the drift tube. For supermicrometer particles, size can be measured with the aerodynamic particle sizer, which injects particles from a sample nozzle into a fast airstream. Particles are accelerated toward the airstream velocity based on their inertia and aerodynamic drag. Particles not fully accelerated prior to a velocity measurement with dual laser beams can be differentiated, based on their resistance to the airstream, which depends on both inertia and aerodynamic drag. Fully accelerated particles cannot be differentiated. Preceding both of these techniques were inertial impaction devices, using centrifugal force, cascading orifices, or counterflow for size discrimination. Centrifuges and cyclones collect the particles, which cross streamlines as the air is forced into a graduated rotation. In a cascade impactor successive nozzle diameters decrease thus increasing air velocity and capturing ever smaller particles. Nozzle diameters are designed for the sizes of interest. Counterflow, or virtual impactors only collect particles with enough inertia to resist the counterflow. In all of these techniques, the collected particles must be analyzed using standard filter and microscopy techniques.

Mass

All aerosol mass measurements rely on aerosol capture either on filters or oscillating microbalances. All but one are gravimetric, comparing the weight of exposed and unexposed substrate. The exception uses the attenuation of a beam of β -particles (electrons) irradiating exposed and unexposed substrate. The attenuation is directly proportional to mass. There are several systems using sensitive oscillating devices such as microbalances, which provide direct, real time, mass measurements as opposed to the delayed analysis of gravimetry and β -ray attenuation. These instruments measure changes in oscillation frequency of sensitive elements as mass is accreted. The tapered element oscillating microbalance mounts a replaceable aerosol collector on the tip of the tapered element, which is oscillated by an electric field. The resonant frequency of oscillation changes as mass accumulates on the filter, through which air is flowing. Similar principles apply to piezoelectric crystal microbalances. The resonant frequencies of the metal-coated quartz crystals decrease as mass is deposited. Particles are deposited, by either electrostatic precipitation or inertial impaction onto the metal, which can become inefficient for larger solid particles due to rebound. The frequency of exposed crystals are referenced to an unexposed crystal at the same temperature, pressure, and humidity. Size differentiated mass can be obtained from these instruments by coupling them with inertial impactors or other size selecting techniques. A variation of the quartz crystal microbalance involves a technique to increase oscillation frequency from 10 to 100 MHz. This is called a surface acoustic wave and increases mass sensitivity by a factor of a thousand.

Surface Area and Shape

Aerosol surface area is of primary interest when considering the impact of aerosol on either atmospheric chemistry or radiation.

Most commonly, surface area is derived from other fundamental measurements such as size distribution. There are, however, several direct ways to measure surface area. These rely on a tracer colliding with and sticking to the particles. Since the collision rate depends on surface area, a measure of the number of tracers collected is a measure of the aerosol surface area, at least the 'active' surface area available for either chemical or radiative interactions. Electrical, radioactive, and photon tracers are used. For electrical tracers, a corona discharge is used for a source of unipolar ions in concentrations high enough for their collection by aerosol through Brownian motion. The charged aerosols are then captured on an impactor stage, or filter, in a Faraday cage, and the very low current (fA) measured. The current is directly proportional to the ions collected, and thus surface area. If a cascade impactor is inserted prior to particle collection, then size resolution of the surface area is available. Alternatively number distributions can be determined if the average number of charges collected per particle size is known. The epiphaniometer works similarly but uses the decay of actinium (^{227}Ac) to radon (^{219}Rn) to lead (^{211}Pb) to produce a source of ^{211}Pb atoms dispersed in a chamber through which the particles flow. Downstream of the chamber with the lead atoms, the aerosol is collected on a filter below an α particle detector. The number of ^{211}Pb atoms collected is measured through their α -decay to ^{207}Pb . The photoelectric aerosol sensor uses UV photons to irradiate aerosol samples. UV photons colliding with a particle are energetic enough to create free electrons from electrons near the particle surface. As the particles lose electrons they become charged. The free electrons ionize nearby gas molecules and move away from particles $<1\text{ }\mu\text{m}$. The gas ions are trapped and the charged particles collected on a filter in a Faraday cage from which the very low current is measured.

Solid particle shape is another of the difficulties of aerosol measurement. Shape affects surface area and therefore chemistry, radiational interactions, and aerodynamic drag, and thus any measurement or process depending on these properties. Shape is often accounted for by adding an adjustable shape factor used to reconcile disparate measurements of other properties. One direct measurement is through microscopy, which is useful to present the challenges which shape can provide. The depolarization of polarized light sources from backscattersondes is another option for particles $>\text{about } 1\text{ }\mu\text{m}$ to indicate the extent of nonsphericity of the particles.

Optical Properties

The optical properties of aerosol form a significant part of several of the instruments already discussed; however, aerosol scattering, absorption, and their sum, extinction, are also measured directly with ensemble instruments. Scattering is measured with a nephelometer, which consists of a monochromatic light source, sampling chamber, and photodetector. Light scattered from aerosol and gas in the sampling chamber is collected over an angular range of ~ 7 to $\sim 170^\circ$. Filtered and ambient air are alternated in the sample chamber to account for variations due to air density and illumination. Some instruments include a shutter to block the forward scatter, so that both total scatter and backscatter are measured. Multiwavelength instruments are available. The backscattersonde, usually deployed on balloons

or aircraft, consists of a pulsed light source and receiving optics focused on the light backscattered from 30 to 50 m from the instrument. Light sources can be xenon flash lamps, lasers, or light emitting diodes. Pulsed light sources allow comparisons of stimulated and ambient backscatter and provide instrument background. For broadband sources, detector filters provide the wavelength dependence. For monochromatic sources, several wavelengths are used. For polarized light sources, depolarization provides particle shape information. Measurements of absorption have already been discussed in the descriptions of the filter-based black carbon instruments and the photoacoustic spectrometer, while extinction measurements have been described using cavity ring-down instruments.

Nucleative Properties

Aerosols play pivotal roles in the hydrologic balance of the atmosphere through their role as cloud condensation nuclei (CCN) and ice nuclei. Without particles we would have neither clouds nor atmospheric ice. CCN counters are similar to CN counters, with the exception that the working fluid is water and the supersaturation is much lower and more carefully controlled. Most CCN counters are based on thermal gradient diffusion chambers, which consist of parallel water saturated surfaces held at different temperatures. Since the diffusion rates of temperature and water between the surfaces differ, supersaturated regions develop in the interior. Controlling the wall temperatures controls the chamber supersaturations in the range 0.05–2.0%. Static thermal diffusion chambers operate by holding air samples drawn into the sample chamber at a single supersaturation for approximately 20 s. Scattering from the cloud of droplets is associated with a number concentration through calibration. Continuous flow thermal diffusion chambers can be operated at several supersaturations simultaneously, by maintaining several discrete temperature differences across the plates at different points in the sampling chamber. Aerosol particles are exposed to increasing supersaturations and so increasingly smaller particles will activate and grow. The droplet size at the end of the chamber, measured with a single particle optical counter, is an indication of droplet growth, which is directly related to the critical size and supersaturation of the CCN. These instruments provide a CCN supersaturation spectrum from single aerosol samples. Static diffusion chambers must adjust plate temperatures between samples to provide a supersaturation spectrum.

Because of their highly active role in precipitation, ice nuclei are even more interesting than CCN, are even harder to measure, and are much more rare. While CCN may constitute 0.1 of the ambient aerosol number concentration, ice nuclei constitute usually $<10^{-4}$ of the ambient aerosol population. There are only a few ice nuclei counters available. They all work on the same principle, mixing the aerosol into a water supersaturated environment, cooling the resulting cloud droplets, which now contain the aerosol, and counting the droplets which freeze. Differences occur due to chamber geometry and orientation, and ice particle detection methods ranging from manual to automatic. The first automated instrument detected ice acoustically as the relatively large ice particles exited a small opening. More recent instruments use optical particle counters. All the instruments primarily measure particles, which act as

condensation freezing nuclei. Instruments to measure deposition or contact ice nuclei are not available.

Nucleated Particles

Completing the contribution of particle measurements to our understanding of the hydrologic cycle are instruments to measure cloud droplets and ice particles. These instruments are optical and use either forward Mie scattering for cloud droplets, or geometric optics to create shadowgraphs of the particles, for ice. The sampling volumes must be adjusted for the different size ranges of interest and thus several instruments are required to cover the complete spectrum. These optical instruments have completely replaced earlier approaches, relying on impaction on films and microscopy. Because of the large particles involved they suffer from edge effects, out of focus images, and false images (*see Clouds and Fog: Measurement Techniques In Situ*).

See also: **Aerosols:** Aerosol Physics and Chemistry, Aerosol–Cloud Interactions and Their Radiative Forcing; Climatology of Tropospheric Aerosols; Dust; Observations and Measurements; Role in Radiative Transfer; Soot. **Arctic and Antarctic:** Arctic Haze. **Chemistry of the Atmosphere:** Observations for Chemistry (Remote Sensing): Lidar; Volcanoes: Composition of Emissions. **Clouds and Fog:** Cloud Microphysics; Measurement Techniques *In Situ*. **Optics, Atmospheric:** Optical Remote Sensing Instruments. **Radiation**

Transfer in the Atmosphere: Scattering. **Satellites and Satellite Remote Sensing:** Aerosol Measurements. **Tropospheric Chemistry and Composition:** Aerosols/Particles.

Further Reading

- Baron, P.A., Willeke, K. (Eds.), 2001. *Aerosol Measurement, Principles, Techniques, and Applications*, second ed. Wiley-Interscience, New York.
- Ensor, D.S. (Ed.), 2011. *Aerosol Science and Technology: History and Reviews*. RTI Press, Research Triangle Park.
- Hinds, W.C., 1999. *Aerosol Technology Properties, Behavior and Measurement of Airborne Particles*, second ed. John Wiley and Sons, New York.
- Kulkarni, P., Baron, P.A., Willeke, K. (Eds.), 2011. *Aerosol Measurement, Principles, Techniques, and Applications*, third ed. Wiley-Interscience, New York.
- Liu, B.Y.H., 1976. *Fine Particles Aerosol Generation, Measurement, Sampling, and Analysis*. Academic Press, New York.
- McMurray, P.H., 2000. A review of atmospheric aerosol measurements. *Atmospheric Environment* 23, 1959–1999.
- Murphy, D.M., 2007. The design of single particle laser mass spectrometers. *Mass Spectrometry Reviews* 26, 150–165.
- Noble, C.A., Prather, K.A., 2000. Real-time single particle mass spectrometry: a historical review of a quarter century of the chemical analysis of aerosols. *Mass Spectrometry Reviews* 19, 248–274.
- Vincent, J.H., 1989. *Aerosol Sampling*. John Wiley and Sons, Chichester.
- Vincent, J.H., 2007. *Aerosol Sampling, Science, Standards, Instrumentation and Applications*. John Wiley and Sons, Chichester.
- Willeke, K., Baron, P.A., 1993. *Aerosol Measurement Principles, Techniques, and Applications*. Van Nostrand Reinhold, New York.

Observations for Chemistry (*In Situ*): Water Vapor Sondes

JB Smith, Harvard University, Cambridge, MA, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by E Weinstock, E Hintsa, volume 4, pp 1490–1499, © 2003, Elsevier Ltd.

Synopsis

Water vapor is central to defining the unique properties of the Earth's atmosphere. Despite its importance, it has proven very difficult to measure accurately, especially at the low concentrations encountered in the upper troposphere and lower stratosphere (UTLS). This article offers an explanation of the techniques employed by the current generation of *in situ* sensors designed to measure water vapor in the UTLS. These techniques can be classified into five principal types: frost point hygrometer; photo-fragment fluorescence spectrometer; infrared absorption spectrometer; chemical ionization mass spectrometer; and capacitance hygrometer. Descriptions of a representative set of individual instruments that utilize each technique are also provided.

Introduction

Water vapor is one of the most important atmospheric trace gases. It plays a central role in defining the unique chemical, dynamical, and radiative properties of the Earth's atmosphere. It is also a critical component of Earth's climate system. Water vapor absorbs throughout the infrared (IR) region of the electromagnetic spectrum and as a result is the dominant greenhouse gas. Furthermore, because it can condense into both liquid and solid phases, it is the critical element for aerosol and cloud formation, even at stratospheric altitudes.

Despite comparatively low concentrations in the upper troposphere and lower stratosphere (UTLS), water vapor in this region of the atmosphere has the potential to influence human health and well-being through (1) its response to greenhouse gas forcing and its consequent climate impact, and (2) its effect on the recovery of stratospheric ozone, which moderates the flux of ultraviolet radiation reaching the Earth's surface. Increases in water vapor concentrations in the UTLS lead to radiative cooling at these levels and induce warming at the surface (e.g., Forster and Shine, 1999, 2002; Held and Soden, 2000). The recent analyses of Solomon et al. (2010) and Dessler et al. (2013), for example, suggest that warming at the Earth's surface may be sensitive to sub-ppmv changes in water vapor in the lower stratosphere. Additionally, water vapor is integral to stratospheric chemistry. It is the dominant source of OH in the lower stratosphere (Hanisco et al., 2001), and increases in water vapor concentrations can promote stratospheric ozone loss by raising the reactivity of several key heterogeneous reactions as well as by promoting the growth of reactive surface area, e.g., (Anderson et al., 2012; Drdla and Muller, 2012; Kirk-Davidoff et al., 1999).

This article offers a basic explanation of the principal techniques employed for the *in situ* detection of water vapor in the UTLS, followed by brief descriptions of a representative set of individual instruments that utilize each technique. The measurement techniques can be classified into five principal types: frost point hygrometer; photo-fragment fluorescence spectrometer; IR absorption spectrometer; chemical ionization mass spectrometer; and capacitance hygrometer. In addition to water vapor, some of these instruments have been designed to measure total water, i.e., the combination of water in its gaseous

and condensed phases, as well as the isotopologues of water, e.g., HDO and H₂¹⁸O. Table 1 lists each individual instrument discussed here, and enumerates the technique utilized for water vapor detection, the airborne platform(s), the sampling strategy, target measurement, and a selection of intercomparison campaigns in which each instrument has participated. More thorough coverage of instrumental details, calibration and validation procedures, and analysis of acquired data are available in the cited references. It is important to recognize that no single measurement method is guaranteed to be more accurate or precise than any other, and no method is intrinsically free of systematic errors and uncertainties.

Measurement History

Water in the Earth's atmosphere has proven very difficult to measure accurately. This is largely due to the fact that atmospheric concentrations vary from a few percent at the surface to a few ppmv in the UTLS. The propensity for water to adhere to surfaces also makes it especially challenging, as any surface upstream of the detection region can be a source for water that can contaminate the ambient sample. Such surfaces include aircraft skin, if the instrument inlet is not out of the boundary layer of the aircraft, or water shedding off of a balloon on ascent, for those instruments carried aloft by balloons, as well as instrument ducting. Similarly, any path in a laser absorption experiment between the laser source and the detector that is not explicitly part of the detection volume can contain water that can contaminate or bias the measurement. Finally, the adaptation of instrumentation to moving platforms and the wide range of ambient conditions – pressures ranging from 1000 hPa to less than 50 hPa, and temperatures ranging from ~280 to ~180 K – can introduce sources of uncertainty not encountered in controlled laboratory settings.

Early comparisons of water vapor measurements in the stratosphere exhibited large discrepancies. The World Meteorological Organization report published in 1985, titled Atmospheric Ozone, shows results of what were then state-of-the-art *in situ* and remote water vapor measurements (WMO, 1985). Differences among the *in situ* instrumentation from 14 km to >30 km are on the order of 1.0 ± 0.5 ppmv, equivalent to

Table 1 *In situ* water vapor instruments with their measurement method, respective platform, sampling technique, target measurement, and selection of intercomparison campaigns in which they have participated

Instrument acronym	Measurement principle	Platform	Sampling	Molecule	Comparison campaign ^a
FPH	Frost point chilled mirror	Balloon	Tube	Water vapor	MACPEX, AV-2, balloon
CFH	Frost point chilled mirror	Balloon	Tube	Water vapor	CRAVE, MACPEX, AV-1 & 2, balloon
LMD	Frost point chilled mirror	Balloon	Tube	Water vapor	–
FISH	Fluorescence Lyman- α	Aircraft Geofizika, WB-57	Internal cell (warm)	Total water	MACPEX, AV-1 & 2
HWV	Fluorescence Lyman- α	Aircraft WB-57, ER-2	Internal duct (fast flow)	Water vapor	CRAVE, MACPEX, AV-1 & 2
HTW	Fluorescence Lyman- α	Aircraft WB-57, ER-2	Internal duct (warm)	Total water	–
FLASH	Fluorescence Lyman- α	Balloon	External	Water vapor	AV-1, balloon
HOX	Fluorescence photolysis/LIF	Aircraft WB-57	Internal cell (P, T control)	Total water H ₂ O, HDO	–
JLH	IR absorption 2nd harmonic	Aircraft WB-57, ER-2	Open path	Water vapor	CRAVE, MACPEX, AV-1
DLH	IR absorption 2nd harmonic	Aircraft WB-57, DC-8	Open path	Water vapor	MACPEX, AV-2
ALIAS	IR absorption 2nd harmonic	Aircraft WB-57, ER-2	Internal cell	Total water isotopologues	CRAVE, MACPEX
HHH	IR absorption direct	Aircraft WB-57, ER-2	Internal cell (fast flow)	Water vapor	MACPEX, AV-2
ICOS	IR absorption ICOS	Aircraft WB-57	Internal cell (P, T control)	Water vapor isotopologues	CRAVE
CIMS	CIMS	Aircraft WB-57	Internal cell	Water vapor	MACPEX
MHD	Capacitance hygrometer	Commercial aircraft	Inlet housing	Water vapor	–

Balloon: Refers to a series of balloon intercomparisons over the past decade, see Vömel, H., David, D.E., Smith, K., 2007a. Accuracy of tropospheric and stratospheric water vapor measurements by the cryogenic frost point hygrometer: instrumental details and observations. *Journal of Geophysical Research: Atmospheres* 112 (D8).

CRAVE: Costa Rica Aura Validation Experiment, Airborne campaign, San Jose, Costa Rica, Winter 2006, see Jensen et al., 2008, and Rollins, A.W., et al., 2014. Evaluation of UT/LS hygrometer accuracy by intercomparison during the NASA MACPEX mission. *Journal of Geophysical Research: Atmospheres* 2013JD020817.

MACPEX: Mid-latitude Airborne Cirrus Properties Experiment, Airborne campaign, Houston, Texas, Spring 2011, see Rollins, A.W., et al., 2014. Evaluation of UT/LS hygrometer accuracy by intercomparison during the NASA MACPEX mission. *Journal of Geophysical Research: Atmospheres* 2013JD020817.

FPH, Frost point hygrometer; CFH, Cryogenic frost point hygrometer; FISH, Fast *in situ* hygrometer; HWV, Harvard water vapor; HTW, Harvard total water; FLASH, Fluorescent advanced stratospheric hygrometer; JLH, JPL laser hygrometer; DLH, Diode laser hygrometer; ALIAS, Aircraft laser infrared absorption spectrometer; HHH, Harvard Herriott hygrometer; ICOS, integrated cavity output spectroscopy; CIMS, Chemical ionization mass spectrometry; MHD, MOZAIC humidity device; LIF, Laser-induced fluorescence.

^aAV-1 & 2: Aqua Validation and Instrument Tests (AquaVIT), Laboratory campaigns, Karlsruhe, Germany, October 2007, April 2013, see Fahey, D.W., et al., 2014. The AquaVIT-1 intercomparison of atmospheric water vapor measurement techniques. *Atmospheric Measurement Techniques Discussions* 7, 3159–3251.

approximately 15–50% depending on the ambient mixing ratio. These data were acquired by balloon-borne and aircraft-borne instrumentation during two targeted international water vapor intercomparison launches out of Palestine, Texas, in 1981 and 1983. In both cases, data from Lyman- α photo-fragment fluorescence instruments were biased high relative to data from the frost point sensors.

Since that time, there have been significant advances in instrument design, the development of new measurement technologies, as well as improvements in calibration and validation methodologies. The results of a comprehensive analysis of intercomparison data obtained prior to 2000 are published in Chapter 2 of the Stratospheric Processes and their Role in Climate (SPARC) Assessment of Upper Tropospheric and Stratospheric Water Vapor (SPARC, 2000). The report includes intercomparisons of satellite, aircraft, balloon-borne, and ground-based water vapor instrumentation. Discrepancies in the critical range of 1–10 ppmv in the 60–100 hPa layer (shown in Table 2.7 of the report, and bottom panel of Figure 2.72) exhibit a range for the comparison of the

measurement means of approximately $\pm 15\%$, with the full range extending to approximately $\pm 30\%$. The two instruments at the highest end of the comparisons are aircraft-borne *in situ* instruments (photo-fragment fluorescence and IR absorption hygrometers), with the balloon-borne *in situ* instruments (frost point sensors) positioned near the middle. (SPARC (2000) is an excellent resource. In addition to the intercomparison analysis, it motivates the need for accurate and precise water vapor measurements, provides descriptions of a comprehensive suite of *in situ* as well as remote instrumentation, and summarizes the state of knowledge regarding the factors controlling the distribution and variability of water vapor in the UTLS.)

Despite this apparent improvement in instrument agreement, differences of a larger magnitude, i.e., ~ 1.5 ppmv or equivalently approximately 50–70% at the water vapor minimum near the tropical tropopause, were once again documented during coordinated intercomparisons of balloon and aircraft instrumentation during the Costa Rica Aura Validation Experiment (CRAVE) campaign operated out of

San Jose, Costa Rica, in February 2006 (Jensen et al., 2008; see also Figure 1 of Fahey et al., 2014). These differences, which are greater than the stated instrumental uncertainties, provided the motivation for an international laboratory intercomparison campaign, i.e., the first Aqua Validation and Instrument Tests (AquaVIT-1), which was conducted at the AIDA aerosol and cloud simulation chamber at the Karlsruhe Institute of Technology, Germany, in October 2007. The results of the laboratory campaign among a core group of instruments showed good agreement, of about $\pm 10\%$ ($\pm 1\sigma$) for mixing ratios between 10 and 150 ppmv, and fair agreement of $\pm 20\%$ ($\pm 1\sigma$) for the region of most interest between 1 and 10 ppmv. The implication is that the substantially larger differences observed during some *in situ* intercomparisons may stem from factors associated with instrument integration onto moving platforms or arise in response to the inherent differences between the laboratory and ambient environments (Fahey et al., 2014). (Fahey et al., (2014) offer a thorough overview of the AquaVIT mission goals, an explanation of the structure of the blind intercomparison, which can function as a template for future efforts both in the laboratory and in *in situ*, a discussion of the analysis methods applied and results, as well as descriptions of the participant instruments submitted by the instrument teams.)

The most recent results from *in situ* comparisons of aircraft-borne and balloon-borne instrumentation acquired during the Mid-latitude Airborne Cirrus Properties Experiment (MACPEX) operated out of Houston, Texas, during the spring of 2011, once more showed improved agreement, with differences consistently < 1 ppmv (Rollins et al., 2014). (Rollins et al. (2014) provide the most recent discussion and analysis of an *in situ* water vapor intercomparison effort. Many of the same instruments that participated in AquaVIT-1 were present during the MACPEX mission.) However, the tendency for the frost point balloon instruments to measure lower than independent aircraft instruments persists, and the large variability in agreement over the past two decades is cause for concern.

These discrepancies prevent a rigorous understanding of the processes that control the distribution and phase of water vapor throughout the upper troposphere and stratosphere (Peter et al., 2006). With disagreements among water vapor measurements, on the order of 1–2 ppmv in the UTLS, the atmospheric chemistry and climate community cannot reliably model the chemical and radiative properties of this region, or predict the response of this region to anthropogenic climate forcing. Several studies have shown that the discrepancy among measurements is too large to constrain the role of different hydration and dehydration mechanisms, e.g., cirrus cloud formation and subsequent dehydration in the tropical upper troposphere (Jensen et al., 2005, 2008; Krämer et al., 2009; Weinstock et al., 2009). For example, model validation is dependent upon which data set is chosen. Moreover, the controversy among the *in situ* instruments fails to support a robust satellite data set, and the overall uncertainty is too large to unambiguously detect, quantify, and attribute trends in water vapor concentrations, particularly at the level of $\sim 1\%$ per year (0.04 ppmv per year) (Oltmans and Hofmann, 1995; Oltmans et al., 2000; Scherer et al., 2008; Hurst et al., 2011). These concerns highlight the need for measurements with high precision and unassailable accuracy.

Frost Point Hygrometer

Principle of Operation

Most standard frost point hygrometers use a chilled-mirror technique, in which a layer of dew or frost is formed on the surface of a small mirror in contact with ambient air. Changes in the frost layer are detected optically, with photo-detectors sensing light, typically from a light-emitting diode, that is scattered off the frost layer. The mirror temperature is actively controlled, heated, or cooled as necessary, to maintain a constant frost layer, i.e., a constant signal on the photo detectors. Depending upon the instrument type, the mirror temperature is controlled with either a mechanical (Stirling-type cooler) or a thermoelectric (Peltier) device, or by actively heating a resistive element against passive cooling with a cryogenic bath. In the cryogenic devices, the mirror is connected to the bath by a rod of high thermal conductivity. Assuming the frost layer is held in thermodynamic equilibrium with water vapor in the air passing over the mirror, measurements of the mirror's temperature obtained with a thermistor or platinum resistance thermometer embedded in, or near, the mirror surface are used to determine the partial pressure of water vapor in the sample. One of several variations of the Clausius–Clapeyron equation may be used to derive the equilibrium vapor pressure of water vapor over ice as a function of temperature (The equation has not been standardized in the measurement community. Though the differences in derived mixing ratio are often approximately $< 1\%$, it is important to be aware of which equation is used by each measurement system.) (Murphy and Koop, 2005). Ambient water vapor mixing ratio is calculated from the derived vapor pressure and simultaneous measurements of ambient pressure and temperature.

The accuracy and precision of this technique may depend upon several factors, among them: (1) the integrity of the ambient sample, i.e., that water from the balloon or other surfaces upstream does not contaminate the flow through the device, or alternatively, that water is not condensed out on the walls upstream of the mirror; (2) the assumption that the frost layer achieves equilibrium with ambient water vapor as the sensor ascends and descends through the atmosphere; (3) the assumption that the observed condensation temperature is not altered by atmospheric impurities (e.g., nitric acid or other water-soluble contaminants); (4) the accuracy of the mirror surface temperature, which is dependent upon the calibration of the primary temperature device, as well as the assumption that the measured temperature accurately represents the mirror surface temperature, i.e., that there are no significant temperature gradients; (5) the sensitivity of the optical sensor and feedback electronics to changes in the frost layer; (6) the sensitivity of the optical feedback to solar scatter; and (7) the accuracy of the simultaneous measurements of ambient temperature and pressure that are used to determine mixing ratio. Furthermore, the spatial and temporal resolution of the measurements are influenced by the time response and efficiency of the frost layer/mirror temperature control loop, changes in ambient humidity, and the ascent or descent rate of the balloon, or speed of the aircraft. The specific instruments listed here have successfully addressed these concerns, albeit in different ways, and data from frost point devices constitute one

of the longest continuous records of accurate and precise water vapor data in the UTLS.

Frost point hygrometers do not rely on routine or regular calibration, that is, the empirical determination of their detection sensitivity through comparison with external water vapor standards. However, comparison with a reference standard is recommended for assessing the sensor's susceptibility to sources of systematic error. Ultimately, the accuracy of the frost point measurements is tied to the accurate determination of the mirror temperature. It is imperative, therefore, that the mirror thermistors be carefully calibrated, i.e., that their resistive response to temperature be determined over the relevant temperature range.

The NOAA Cryogenic Frost Point Hygrometer

The National Oceanic & Atmospheric Administration (NOAA) Earth System Research Laboratory's (ESRL) Global Monitoring Division (GMD) frost point hygrometers (FPHs) are balloon-borne instruments that provide continuous measurements of water vapor from near the surface to the middle stratosphere. Their heritage extends back to the Naval Research Laboratory in Washington, DC (Mastenbrook and Dinger, 1961), and they have provided a long history of water vapor profiles in the UTLS and have contributed significantly to understanding stratospheric water vapor (e.g., Hurst et al., 2011, and references therein). Approximately 500 launches of the NOAA FPH from Boulder, Colorado, from April 1980 through February 2010 have yielded more than 300 water vapor vertical profiles with reliable stratospheric data, providing information on water vapor trends and variability throughout the upper troposphere and middle stratosphere (Hurst et al., 2011; Kunz et al., 2013). The present generation has a stated total uncertainty of ± 7 – 10% arising from systematic uncertainties in the determination of the frost temperature and total pressure ($\sim 0.5^\circ\text{C}$ and 0.5 hPa , respectively) (Vömel et al., 1995; Vömel et al., 2007a).

The FPH instruments operate with a cryogenic bath, and regulate the mirror temperature with a resistive heater connected to a microprocessor-controlled feedback loop. The system allows for both fast heating and cooling rates, and thus good time response and vertical resolution. Ambient air is drawn in through a clean 18 cm long stainless steel tube, designed to eliminate contamination. Flow velocities through the sample tube are estimated to be roughly half the velocity of the balloon. For ascent velocities of 5 m s^{-1} mass flows through the tube are ≥ 10 standard liters per minute (slm) up to the tropical tropopause level, and decrease to ~ 5 slm at 20 km. It is expected that self-contamination from the inlet tubes at these velocities and mass flows is minimal (Vömel et al., 2007a). Contamination from the balloon itself can occur on ascent, and is sometimes evident between 20 and 25 km. Data showing evidence of contamination are identified and removed.

The Cryogenic Frost Point Hygrometer

The cryogenic frost point hygrometer (CFH) was developed during 2003–04. It uses the same fundamental measurement principle and design as the NOAA/ESRL/GMD FPH. Like the FPH, it is a small balloon-borne instrument that provides continuous vertical profiles of water vapor from the surface to

approximately 25–28 km. At the time of its development, a number of features were introduced to reduce the power consumption, weight, cost, and operator skill required by the older generation FPH. In its current configuration, which is manufactured and sold by Droplet Measurement Technologies (DMT), the instrument package includes the CFH, an electrochemical concentration cell ozone sonde, a Global Positioning System, and a radiosonde. The radiosonde provides *in situ* temperature and pressure data as well as serving as the data transmitter. Details of the instrument and an analysis of its measurement results are contained in Vömel et al. (2007a). The stated accuracy of the CFH is better than $\pm 4\%$ in the lower troposphere, and less than or equal to $\pm 9\%$ in the vicinity of the tropopause, and $\pm 10\%$ in middle stratosphere (<http://www.dropletmeasurement.com/sites/default/files/Brochures/CFH.pdf>, recovered 27 February 2014).

Comparisons between an older generation FPH and the then newly developed CFH instrument were executed in June 2005. Four soundings in total (two each) were acquired on different days. Comparisons of descent profiles of the two instruments show agreement well within the instrumental uncertainty throughout the stratosphere. The data from these soundings show no evidence of a bias in either instrument. More recently, from June 2008 to February 2009 there were five soundings launched of Boulder, Colorado, with both the CFH and present generation FPH on the same balloon. The mean difference between stratospheric water vapor measurements by the two instruments on these flights was $0.1 \pm 0.3\text{ ppmv}$, or $\sim 2 \pm 6\%$ of the mean stratospheric water vapor mixing ratio. For more comparisons of CFH with other UTLS hygrometers, and a detailed discussion of using CFH in support of Aura Microwave Limb Sounder (MLS) version 1.5 and version 2.2 validation efforts refer to the following references (Read et al., 2007; Vömel et al., 2007a,b).

Laboratory versions of both the FPH and CFH instruments have been developed, equipped with the same or similar thermal, optical, and electronic components as the flight hygrometers, and packaged within a vacuum housing. This modification allows them to interface to a controlled water vapor source operated at relevant atmospheric pressures.

The LMD-CNRS Frost Point Hygrometers

The Laboratoire de Météorologie Dynamique du Center National de Recherche Scientifique (LMD-CNRS) has developed a frost point hygrometer for use on aircraft, traditional balloon-borne payloads, and long duration (Montgolfier Infra-Rouge, MIR) balloon-borne payloads. The principle of operation of the three airborne versions is the same, with differences only in (1) how the frost point temperature is controlled, i.e., Peltier thermoelectric device versus cryogenic cooling, and (2) how air is brought to the mirror surface. Calibration and tests of the instruments are routinely performed with a system that allows the generation of frost point temperatures down to -90°C , from 1000 to 20 hPa, simulating the conditions encountered during balloon and aircraft flights. Comparisons between the frost point produced by the calibration device and the corresponding frost point measured by the flight hygrometers typically show agreement to within the precision of the hygrometer. A discussion of the family of LMD hygrometers, operational details, the

calibration procedure, and a list of publications utilizing LMD data is available online at http://ether.ipsl.jussieu.fr/etherTypo/fileadmin/files/ELHYSA/info_ELHYSA.pdf, recovered 27 February 2014.

Photo-Fragment Fluorescence

Principle of Operation

In the photo-fragment fluorescence technique, water vapor is dissociated with vacuum ultraviolet (VUV) radiation, typically Lyman- α ($\lambda_{\alpha} = 121.6$ nm). Some of the OH photo-fragments are formed in their first excited electronic state (OH^*), and their subsequent fluorescence, near 310 nm, is proportional to the number density of water vapor in the detection volume. The following equation demonstrates this proportionality: $S_{\text{OH}^*} = [\text{H}_2\text{O}] \cdot C_{\text{flr}}$. In this equation, S_{OH^*} is equal to the normalized net fluorescence signal detected, e.g., counts per second, $[\text{H}_2\text{O}]$ represents the water vapor number density in the detection volume with units of ($\# \text{ cm}^{-3}$), and C_{flr} is the empirically determined density-dependent constant of proportionality. At moderate and high ambient pressures, i.e., $\geq \sim 100$ hPa, most of the excited state OH^* radicals are quenched by collisions with air molecules at a rate proportional to air density, $[\text{M}]$ ($\# \text{ cm}^{-3}$). Thus, in the troposphere and lowermost stratosphere the fluorescence signal is almost directly proportional to the ambient volume mixing ratio, e.g., $S_{\text{OH}^*} \sim [\text{H}_2\text{O}]/[\text{M}]$.

The accuracy and precision of this technique depend upon several factors, among them (1) the delivery of uncontaminated ambient air to the detection region; (2) the accurate calibration of the instrument sensitivity, represented by the density-dependent proportionality constant C_{flr} that relates the detected fluorescence counts to the water vapor concentration; (3) the applicability of the calibration, determined in the laboratory, to the *in situ* environment; (4) the accurate measurement of the VUV lamp intensity used in the normalization; (5) the rejection by the detector of background UV radiation from the Sun as well as from the VUV light source in the region of the fluorescence signal (290–350 nm); (6) the assumption that there are no other sources of VUV stimulated fluorescence; and (7) the accuracy of the simultaneous measurements of temperature and pressure that are required to determine mixing ratio. When these criteria are met, photo-fragment fluorescence can provide accurate and precise measurements of water vapor from the middle troposphere to over 20 km with a time resolution of ~ 1 Hz or better.

As noted above, the photo-fragment fluorescence detection method requires calibration, i.e., the empirical determination of the quantitative relationship between normalized net fluorescence counts and known water vapor mixing ratios over a range of temperatures and pressures that are representative of the atmospheric environment. The accuracy of the photo-fragment instruments critically depends upon the accuracy with which the mixing ratios generated in the laboratory are known. Different photo-fragment hygrometers have chosen different reference standards. Again, no single reference standard is fundamentally preferable to another, and the final estimate of uncertainty for the photo-fragment sensors, typically ± 5 –10% in mixing ratio, depends more upon the details

of the laboratory calibration procedure than on the choice of standard.

Fast *In Situ* Hygrometer Instruments

The fast *in situ* stratospheric hygrometer (FISH) instruments, which utilize the Lyman- α photo-fragment fluorescence technique, were developed at the Forschungszentrum Jülich, Germany (Zöger et al., 1999). A modular design has allowed FISH to be deployed on different aircraft (Lear Jet, Geofizika, Falcon, and NASA's WB-57), as well as balloon-borne payloads. Furthermore, with the addition of a heated inlet, FISH can operate in an enhanced total water (vapor + amplified condensed phase water) mode. The accuracy of the current design is typically $\sim \pm 6\%$, $+0.15$ ppmv, with a precision of $< \pm 0.1$ ppmv at 1 Hz (Fahey et al., 2014; Rollins et al., 2014; and references therein).

FISH consists of a closed, vacuum-tight fluorescence cell made of black glass mounted in a larger stainless steel cell designed to eliminate both scattered light and outgassing from dead volumes, a Lyman- α radiation source, a photo-multiplier tube (PMT) in photon-counting mode, two detectors to monitor the VUV lamp intensity, one across from the lamp, and the other positioned equidistant from, but at right angles to the lamp, and a mirror drive to control the measuring cycle, e.g., the determination of the fluorescence signal, the background count rate, as well as the lamp intensity. The water vapor mixing ratio of the sample is determined from the net normalized fluorescence signal and the empirically derived calibration coefficients that define the instrument sensitivity.

The nature of the inlet design varies depending upon the aircraft and whether or not the instrument is measuring vapor only or enhanced total water. However, the inlet is always designed to be well outside the aircraft boundary layer. In the enhanced total water configuration, the inlet is forward facing and samples particles anisokinetically, i.e., the number of particles in the sample is enhanced relative to the ambient particle number density. The amplification factor depends upon particle size and can be calculated. In this configuration, the inlet line is ~ 2.5 m long and consists of ~ 2 m of 1 cm inner diameter heated stainless steel tubing ensuring the complete vaporization of the sampled particles. Flow rates through the instrument are 10–50 l per minute, and decrease with altitude.

Calibration of FISH occurs frequently during mission deployments. The calibration system consists of a humidity generator and a commercial frost point device (MBW-DP30), which serves as the laboratory reference instrument (Zöger et al., 1999; Meyer, 2008). Calibrations are executed over a range of pressures (approximately 80–500 hPa) and mixing ratios (1–500 ppmv) to cover the relevant range of the *in situ* measurements. The MBW-DP30 is calibrated by the manufacturer and is traceable to international standards.

Harvard Water Vapor and Harvard Total Water Instruments

Versions of the Harvard University water vapor Lyman- α photo-fragment fluorescence instrument (HWV) have measured water vapor *in situ* in the UTLS for over 20 years (Schwab et al., 1990; Weinstock et al., 1994; Hintsa et al., 1999). The present design, intended for deployment on NASA's

ER-2 and WB-57 aircraft, was developed in 1998. A companion Harvard total water (vapor + condensed phase) (HTW) instrument was developed in parallel and both instruments have been flying missions since 2001 (Weinstock et al., 2006a,b). The most recent configuration of HWV includes both the Lyman- α instrument discussed here and an independent instrument, the Harvard Herriott Hygrometer (HHH) discussed in the IR absorption section, which is located upstream (Sargent et al., 2013). HWV has an achievable precision of ~ 0.1 ppmv at 1 Hz, and an accuracy of $\sim \pm 5$ –8% based upon laboratory calibrations, with a maximum additional bias of < 0.3 ppmv (Weinstock et al., 2009; Rollins et al., 2014; and references therein).

All versions of the Harvard Lyman- α instrument utilize 121.6 nm radiation from a sealed radio-frequency discharge lamp to photo-dissociate water vapor in an internal duct, with the resulting OH* fluorescence collected at right angles to the duct and the Lyman- α light source through a stack of optical filters and detected with a PMT. A cell containing flowing dry air is used to filter unwanted VUV radiation from the lamp, and a coated magnesium fluoride narrow bandpass filter (incorporated in 2008) is used to preferentially allow 121.6 nm light into the detection volume. This optic eliminates background signal from the lamp at the near UV fluorescence wavelengths. An antireflection coated quartz window within the filter cell is cycled in and out of the Lyman- α lamp beam to enable measurement of any nonfluorescent background signal from either the lamp or solar scatter in the duct. The Lyman- α lamp intensity is continuously monitored with a VUV vacuum photodiode mounted across the duct and opposite the lamp. This signal is used to normalize the net fluorescence counts. Conversion of normalized net counts to water vapor number density in the duct is accomplished with the empirically determined calibration constants. The determination of the ambient mixing ratio is achieved via pressure and temperature measurements acquired by a pitot tube and thermistors mounted immediately downstream of the detection region.

A rear-surface coated magnesium fluoride (MgF_2) mirror adjacent to the diode reflects some of the VUV radiation back across the duct to a second diode, allowing for the simultaneous measurement of water vapor by direct absorption in the mid- to upper troposphere. This provides a means of independently verifying the fluorescence measurement *in situ*.

In the current version of HWV, air is ram-fed through a $\sim 9 \times 10$ cm aluminum duct in the nose of a pod mounted under the wing of NASA's WB-57 or ER-2 aircraft. The inlet extends forward of the pod so as to sample air unperturbed by the aircraft or the wing. The laminar core of the primary flow is sampled downstream through a $\sim 5 \times 5$ cm square secondary duct. The secondary duct directs the flow through an aerodynamically designed light trap to exclude solar scatter, and on to the Lyman- α detection axis. Flow velocities of $\sim 70 \text{ m s}^{-1}$ are maintained through the detection region for fast time response and to prevent contamination of the sample due to water outgassing from walls upstream. Under normal conditions, a downstream throttle valve is fully open to facilitate fast flow. For diagnostic purposes, the valve is periodically throttled to slow the flow and test for potential contamination.

The HTW instrument, which utilizes the same detection method as HWV, integrates in a pallet that flies in the belly of

the WB-57 aircraft. Air is drawn through a 1 cm diameter forward facing inlet positioned about 1 m from the skin of the aircraft to avoid contamination by either the aircraft boundary layer air or ice particles bouncing off the skin of the aircraft. Isokinetic flow through the inlet is maintained via real-time control of a roots pump located downstream of the detection axis. This ensures that air is brought into the instrument without perturbing the ambient particle density. Approximately 600 W of heat distributed both in the flow and along the ducting ensures that ice particles $\leq 50 \mu\text{m}$ in diameter are fully sublimated upstream of the detection region.

The Harvard Lyman- α instruments are calibrated in the laboratory before, during, and after mission deployments by flowing known amounts of water vapor in air through the detection axes. The air–water vapor mixtures are prepared by diluting saturated air, generated using a dual-stage bubbler apparatus, with a primary flow of dry air. The concentration of the water vapor added to the calibration duct is independently verified by a direct absorption measurement at 121.6 nm. Calibrations are executed over a range of mixing ratios, pressures, and temperatures representative of the UTLS. The combination of laboratory calibrations and in-flight diagnostics sets the expected ± 5 –8% accuracy of HWV (Weinstock et al., 1994; Hints et al., 1999; Weinstock et al., 2006a, 2009). The maximum potential bias is empirically constrained to be < 0.3 ppmv by measuring the mixing ratio in nitrogen gas delivered from a liquid nitrogen Dewar or very dry air in the laboratory (Hints et al., 1999).

Fluorescent Advanced Stratospheric Hygrometer Instruments

The fluorescent advanced stratospheric hygrometer for balloon (FLASH-B) instrument is a compact lightweight Lyman- α photo-fragment fluorescence sonde developed at Central Aerological Observatory (CAO), Russia, for balloon-borne water vapor measurements in the UTLS (Yushkov et al., 1998, 2001). FLASH-B has flown successfully in a number of balloon campaigns. Some of these campaigns have included simultaneous measurements by independent instrumentation, e.g., FPH and CFH (Vömel et al., 2007a; Khaykin et al., 2013). The expected precision of FLASH is 5.5% for a 4-s integration time, with a calibration uncertainty of $\sim 4\%$ in the 3–100 ppmv range. The total uncertainty of the measurement is estimated to be $\sim \pm 10\%$ for mixing ratios > 3 ppmv, and $\sim \pm 20\%$ for mixing ratios < 3 ppmv (Fahey et al., 2014; and references therein).

The instrument uses a current-stabilized hydrogen discharge lamp filled with a mixture of hydrogen and helium at a total pressure of 10 hPa to provide Lyman- α radiation for the photo-dissociation of water vapor. The lamp utilizes a specially fabricated MgF_2 window to suppress wavelengths in the 270–320 nm band arising from both stray helium emission and hydroxyl emission from the lamp. A PMT run in photon counting mode detects the resulting OH fluorescence through a narrow band interference filter in the fluorescence spectral region, i.e., 308–316 nm.

FLASH-B uses an open layout where the detection optics look directly at the sky (Khaplanov et al., 1992). Accordingly it is only suitable for nighttime measurements. The intensity of the fluorescent light sensed by the photomultiplier is effectively directly proportional to the water vapor mixing ratio under

stratospheric conditions (30–150 hPa) with small oxygen absorption (3% at 50 hPa). At pressures greater than 150 hPa, the VUV absorption by oxygen and water vapor is taken into account. The background signal caused by the night sky emissions in the absence of fluorescence is detected by modulating the lamp with a 1 kHz square wave with a 1/8 duty cycle, combined with a synchronous demodulation of the received signal.

A FLASH instrument has also been developed for use aboard the Geofizika highaltitude aircraft (Sitnikov et al., 2007). Generally, the airborne system is nearly identical to the balloon-based version. A closed cell design, however, permits measurements during the day as well as at night. The FLASH instruments are routinely calibrated at CAO. An MBW-373L dew point hygrometer serves as the reference instrument. Calibrations cover the range of water vapor mixing ratios encountered in the UTLS at pressures of 10–1000 hPa.

Photo-Fragment Laser-Induced Fluorescence

Principle of Operation

A photo-fragment laser-induced fluorescence instrument was developed at Harvard University for the simultaneous measurement of H₂O and its heavy isotopologue HDO in the UTLS (St Clair et al., 2008; Wennberg et al., 1994). Unlike the Lyman- α instruments discussed previously, the photolysis and detection are separated significantly in space and time. This instrument uses a high pressure Xenon excimer lamp with maximal output at 172 nm to make ground electronic state OH and OD from the photo dissociation of H₂O and HDO. The OH and OD radicals are subsequently detected downstream via laser-induced fluorescence (LIF) using a tunable pulsed laser near 288 nm. Precision tuning of the laser allows for the selective excitation and detection of both OH and OD. In order to compensate for the relative atmospheric abundance of HDO and H₂O, $\sim 3 \times 10^{-4}:1$, the specific spectral region for this instrument was chosen such that the sensitivity to OD is several orders of magnitude greater than that for OH. The excited OH and OD radicals fluoresce via the same mechanism beginning with collisional relaxation of the selected excited vibrational state followed by the radiative relaxation of the excited electronic state that yields the detected fluorescence signal at ~ 309 nm.

The OH and OD concentrations are directly related to the measured fluorescence signal through the following equation: $(OX) = S_{OX} \cdot C_{OX}$. In this relationship, S_{OX} is equal to the signal collected at the PMT, and C_{OX} represents the empirically determined sensitivity or calibration constant for OH or OD detection. C_{OX} includes terms that account for the fluorescence efficiency of the excited state OH and OD radicals, and the optical collection efficiency of the detection system. The measured concentrations of OH and OD are then tied directly to ambient H₂O and HDO concentrations via a second equation that accounts for the photolysis yield of OH and OD and the chemical loss of OH or OD via the self reaction $OH + HO_2 \rightarrow H_2O + O_2$ that occurs during time between photolysis and fluorescence detection. The overall sensitivity of the instrument to H₂O and HDO is determined via frequent laboratory calibrations with a known water vapor source,

where the initial HDO:H₂O ratio of the liquid water source and the isotopic fractionation that occurs during calibration are well characterized.

The accuracy and precision of this technique depend upon several factors, among them (1) the delivery of uncontaminated ambient air to the detection region; (2) accurate calibration, specifically the determination of the photolysis yield, and careful monitoring of the optical collection efficiency in the laboratory and in flight; (3) the applicability of the calibration, determined in the laboratory, to the *in situ* environment; (4) the assumption that there are no sources of chemical or fluorescence interference; and (5) the accuracy of the simultaneous measurements of temperature and pressure that are required to determine mixing ratio. When these criteria are met, photo-fragment LIF detection provides molecular selectivity, high sensitivity and fast time response, for accurate and precise measurements of water vapor and its heavy isotopologue, HDO, throughout the UTLS.

Harvard ‘Hoxotope’

At present, there is only one instrument that utilizes this detection technique. This instrument, developed at Harvard University, utilizes the same technique and much of the same equipment, e.g., ducting, optics, and electronics, that was developed for the detection of OH and HO₂ radicals in the UTLS aboard NASA’s ER-2 aircraft (Wennberg et al., 1994). The ER-2 ‘HO_x’ instrument was modified in 2004 for the *in situ* measurement of water vapor and its isotopologue, HDO, and acquired the name ‘Hoxotope’. The instrument has flown in several missions over the past decade on NASA’s WB-57, and analyses of the simultaneous H₂O and HDO measurements have been influential in demonstrating the potential importance of convection in hydrating the UTLS (Hanisco et al., 2007). At typical stratospheric values, e.g., HDO = 0.8 ppbv and H₂O = 5 ppmv, the precision in 10 s is ± 0.04 ppbv for HDO and ± 0.02 ppmv for H₂O. Adding all the known uncertainties in quadrature yields an overall calibration uncertainty, and in-flight accuracy, of $\pm 5\%$ for both H₂O and HDO (St Clair et al., 2008).

Hoxotope has flown in both a vapor-only sampling configuration (rear-facing inlet), as well as in a total water sampling configuration (forward-facing isokinetic inlet). In both configurations, the inlet is heated. In the first case, this is to prevent water uptake, and in the second case, it facilitates particle sublimation. Temperature sensitivities of the detection scheme are also minimized by warming the airflow to $\sim 25^\circ\text{C}$ prior to photolysis and detection. A scroll pump, with a maximum displacement of ~ 8 l per second draws air through the instrument, and an actively controlled conductance-limiting pinch valve is used to regulate the pressure within the detection region of the instrument. The instrument pressure is maintained at 15.3 hPa regardless of the ambient pressure. Every component was chosen to minimize stagnant volumes that can retain water and contaminate the sample.

Calibrations of Hoxotope are executed in the laboratory by adding air with a known water vapor concentration directly to the instrument inlet. The pressure and temperature within instrument during calibration are identical to those in flight. Known concentrations of water vapor are prepared by diluting

saturated air, generated using a single-stage bubbler apparatus, with a primary flow of dry air. While this is a robust calibration method for water vapor, the saturated flow is subject to isotopic fractionation. The heavy isotopologue, HDO, preferentially remains in the liquid phase, and the saturated flow is correspondingly light. Consequently, a second method was devised in which liquid water drops are added directly to the dry carrier flow, via a droplet microinjector, where they completely evaporate, thus preserving the initial H₂O to HDO ratio. As expected, the two methods yield nearly identical results for water vapor, but the microinjector water vapor contains more deuterated water than the bubbler water vapor.

IR Absorption

Principle of Operation

Because of its bent molecular structure, a broad range of rotational and vibrational energy states are accessible to gaseous water, and transitions between these states are associated with the absorption of radiation at specific wavelengths in the IR region of the electro-magnetic spectrum. Two O–H stretching vibrations (symmetric and asymmetric) lead to an absorption band centered at 2.7 μm , and the H–O–H bending vibration leads to an absorption band centered at 6.3 μm . Combinations of these primary vibrational modes lead to additional absorption bands throughout the IR region. It is this feature of water vapor that makes it the most important greenhouse gas in Earth's atmosphere, and that makes IR absorption spectroscopy a preferred method for water vapor detection.

The fractional absorption of light at a specific wavelength by an absorbing species, e.g., water vapor, in a gas sample is directly related to the concentration of that species by the Beer–Lambert Law, i.e., $I/I_0 = e^{-\sigma\ell[\text{H}_2\text{O}]}$. In this equation, I_0 and I refer to the intensity or power per unit area of incident light detected in the absence and presence of water vapor respectively, σ is the molecular absorption cross section of water at the incident wavelength with units (cm^{-1} per ($\# \text{ cm}^{-3}$)), ℓ is the distance in cm the light travels through the detection region, and $[\text{H}_2\text{O}]$ is the water vapor concentration in ($\# \text{ cm}^{-3}$), in the detection volume.

Tunable diode lasers in the IR region are used as the narrow band light sources for these measurements because they can be rapidly and continuously scanned over individual spectroscopic absorption features and thus provide a measure of the intensity at the peak of the absorption feature, I , as well as in a nearby region where absorption is negligible, I_0 . Laser tuning is achieved by continuously increasing the current through the laser over a specified range. As the current increases the laser temperature increases causing a change in the output wavelength. Different types of laser are chosen for accessing different spectral features in the IR region, where the choice of spectral region is dictated by the desired measurement sensitivity, other target molecules, e.g., HDO and H₂¹⁸O, and avoiding absorption by species other than the target molecule(s). Similarly, different detectors are chosen to match the selected laser wavelength.

The sensitivity of the measurement is related to the ratio, I/I_0 , and is typically higher for higher water vapor concentrations, as long as the feature is not saturated. For low water

vapor concentrations, like those encountered in the UTLS, the sensitivity can be improved by increasing the path the light travels through the detection region. This is typically done by creating a cavity in which the laser light is bounced back and forth several times between reflective mirrors before it encounters the detector. The Herriott cell, in which two spherical mirrors are positioned opposite one another, is a commonly utilized multiple-pass cell for absorption spectroscopy (Herriott et al., 1964).

In the simplest version of these instruments, i.e., direct absorption, the ambient concentration of water vapor, $[\text{H}_2\text{O}]$, depends solely on the measurement of the signal, I , accurate determination of the background signal, I_0 , and knowledge of the pathlength, ℓ , as well as the absorption coefficient σ . The absorption coefficient is, in turn, a function of the strength and the shape of the specific molecular transition. Determination of σ requires knowledge of the temperature and pressure in the detection volume as well as several additional spectral line parameters. The HITRAN database (<http://www.cfa.harvard.edu/hitran/>), which was expressly created for modeling the IR properties of the atmosphere, is the standard reference for these empirically determined spectroscopic constants. The direct absorption method provides a technique that does not necessitate frequent calibration, that is, the empirical determination of the instrument's sensitivity to water vapor. However, comparison with an external standard is essential for identifying systematic errors and validating instrument performance.

An additional technique, frequently employed when high-sensitivity is required, is harmonic detection. With this technique, a small-amplitude sinusoidal modulation is added to the basic laser scan. The modulation frequency, f , is usually of the order of tens of kHz. The detector signal is then demodulated at a multiple of this frequency, most often at twice the sine wave frequency, or $2f$ (May and Webster, 1993; May, 1998). This method minimizes certain types of electronic and optical noise that can limit the detection sensitivity of the instrument. In contrast to the direct absorption method, however, harmonic detection complicates the determination of the water vapor mixing ratio. Uncertainties in several independent parameters can cause pressure and temperature dependent errors that are difficult to quantify in a second-harmonic spectrometer. Analytical laboratory calibrations are therefore used to derive the appropriate data processing matrices used to convert the modulated $2f$ peak-to-peak amplitude signal, normalized by returned laser power, to measured mixing ratio.

IR absorption measurements have excellent molecular specificity, and have demonstrated outstanding precision and fast time response. The accuracy and precision of this technique depend upon several factors, among them: (1) the delivery of uncontaminated ambient air to the detection region; (2) the minimization of contaminant water in the laser path outside of the detection volume; (3) accurate determination of the temperature and pressure in the detection volume; (4) the accuracy of the spectroscopic parameters in the HITRAN database; and (5) the assumption that instrument performance is robust to the environmental change in going from the laboratory to the atmosphere. Because the water vapor concentration is determined by the ratio of (I/I_0), measurement accuracy is expected to be insensitive to changes in mirror reflectivity, scattering by particles in the duct, or long term changes in laser

power. Direct absorption with a narrow line width source allows the water vapor concentration to be calculated directly from the observed line, known spectroscopic parameters, and simultaneous measurements of the temperature and pressure within the cell. Harmonic detection provides substantial improvements in instrument sensitivity, but demands the empirical determination of data processing matrices.

Jet Propulsion Laboratory (JPL) Laser Hygrometer

The NASA Jet Propulsion Laboratory (JPL) laser hygrometer (JLH) instrument uses a near-IR tunable diode laser coupled into an open-path multi-pass Herriott cell, which is mounted external to the aircraft platform and positioned outside the aircraft boundary layer (May, 1998). The instrument utilizes harmonic detection for high precision measurements of atmospheric water vapor in the UTLS. Three versions have been developed at the Jet Propulsion Laboratory, California Institute of Technology, for use on NASA's ER-2, WB-57, DC-8, and Global Hawk aircraft, and JLH has participated in multiple airborne missions from 1997 to the present. The stated precision of JLH is $\sim\pm 0.10$ ppmv for a nominal 1-s integration period, and an accuracy of $\pm 10\%$ (May, 1998; Fahey et al., 2014).

JLH utilizes a near-IR distributed feedback (DFB) tunable diode laser at ~ 1.4 μm . JLH is capable of operating in two modes: harmonic wavelength modulation spectroscopy (May, 1998; May and Webster, 1993), and direct absorption spectroscopy. As discussed in the introductory section, the advantages of harmonic spectroscopy are fast response and high precision, while the advantages of direct absorption measurements are high accuracy, and a verification of overall instrumental performance. The laser is scanned at a 10 Hz repetition rate through a 2 cm^{-1} spectral region that includes the chosen water vapor absorption line. Second-harmonic detection ($2f$) is utilized for signal enhancement by adding a small-amplitude sinusoidal waveform at a frequency of 10 kHz to the laser current. The detector signal is demodulated at 20 kHz to produce a second-harmonic spectrum. Periodic direct absorption measurements serve as an *in situ* validation of the measurements acquired with the calibrated $2f$ detection technique.

The Herriott cell is made up of two opposing gold-coated mirrors that are mounted in the free stream and strategically placed below the aircraft wing or forward fuselage. The mirror separation is maintained to within a tight tolerance over the wide range of ambient temperatures, e.g., from ~ 180 K to 300 K, by invar rods. The laser and detector are enclosed in an evacuated aluminum housing, which is also mounted external to the aircraft, and is temperature stabilized at 15 °C. The laser beam is directed through a hole in the mirror closest to the laser housing. The beam completes ~ 50 passes in the Herriott cell, in total traversing ~ 10 m, before returning through the same hole and reaching the detector. There is an additional short path in ambient air between the entrance/exit mirror and the window to the evacuated laser/detector housing. Spurious absorption due to water vapor in the very short path inside the housing itself is minimized by maintaining very low pressure ($< 10^{-6}$ hPa) in the vessel.

Laboratory calibrations are used to derive the empirical relationship between the $2f$ spectra and water vapor mixing

ratio. JLH is calibrated in the laboratory utilizing a Thunder Scientific 3900 Low-Humidity Generator as both a water vapor source and as a reference. Additional standards, which can monitor the water vapor mixing ratio at the input and/or output of the chamber, include a National Institute of Standards and Technologies (NIST)-traceable chilled mirror hygrometer from General Eastern, and a mid-IR direct absorption measurement at 5316 cm^{-1} (Troy, 2007). The instrument is placed in stainless steel chamber and calibrated over a wide range in both temperature and pressure to map out the instrument response under conditions representative of those in the UTLS.

NASA Langley/Ames Diode Laser Hygrometer

The NASA Langley/Ames diode laser hygrometer (DLH) utilizes a commercially available mid-IR fiber-coupled tunable diode laser coupled to a long external path to measure water vapor throughout the troposphere and lower stratosphere (Diskin et al., 2002; Livingston et al., 2008). Versions of DLH have successfully flown in many recent field campaigns and on several different aircraft that cover a range of altitudes from near the boundary layer to the lower stratosphere, e.g., the P-3, DC-8, WB-57, ER-2, and the Global Hawk. The long path-length combined with harmonic detection provides excellent sensitivity and rapid time response. The precision of DLH is $\sim 1\%$ (1σ) or 50 ppbv at 20 Hz, with an accuracy of $\pm 10\%$ (Podolske et al., 2003).

In order to cover the orders of magnitude change in ambient mixing ratio from the mid-troposphere to the lower stratosphere, DLH targets different absorption lines in the 1.4 μm spectral region. Weaker lines are used at low altitudes (high mixing ratios) and a strong line is used at high altitudes (low mixing ratios). The laser line is locked to the center of the desired absorption feature with a low pressure reference cell. DLH utilizes $2f$ and/or $4f$ harmonic detection. The laser beam is wavelength modulated at ~ 2 kHz, and detection is accomplished by demodulating the return signal at twice or four times the modulation frequency. The return laser power is also measured for normalization.

All versions of the instrument utilize an open-path, two-pass configuration. The path is defined on one end by a laser transceiver and on the other by a panel of high grade retro-reflecting material. The mounting locations of the laser/detector and retro-reflectors depend upon the aircraft. On the Global Hawk, for example, the pressurized laser transceiver housing is mounted in a payload bay within the aircraft fuselage, and the retro-reflecting material is adhered to a fin mounted below the left wing. This configuration leads to an external round-trip path of 12.22 m, with most of the sample volume outside the aircraft boundary layer, and no inlet effects. The internal path inside the pressurized optical housing is of 0.55 m. Maintaining pressure in this volume and it purging with dry air reduces the sensitivity of the measurement to internal moisture.

DLH is calibrated in the laboratory over a representative range of pressures and water vapor concentrations. The calibration data are incorporated into a multiparameter model, which is used to determine the coefficients that convert the demodulated signal, along with independent measurements of ambient temperature and pressure, to water vapor mixing ratio.

Aircraft Laser Infrared Absorption Spectrometer

Aircraft laser infrared absorption spectrometer (ALIAS) is a high-resolution scanning tunable laser absorption spectrometer designed to fly in a spearpod on NASA's ER-2 or WB-57 aircraft (Webster et al., 1994). The instrument has had many configurations utilizing different IR lasers to target a wide array of atmospheric molecules, e.g., HCl and N₂O. From 2002 to the present it has targeted measurements of total water (water vapor + water in the condensed phase), as well as the isotopologues of water, e.g., HDO, H₂¹⁷O and H₂¹⁸O in the gaseous and condensed phases, throughout the UTLS (Webster and Heymsfield, 2003). The accuracy of the ALIAS H₂¹⁶O measurements is ±10% (Rollins et al., 2014).

ALIAS uses a forward-facing isokinetic inlet, which extends from the front of the spear pod and well outside its region of influence on the flow. The isokinetic design ensures that the sampled particle number density is equivalent to that in the ambient environment. 500 W of heat is applied to the inlet to vaporize any water in the condensed phases prior to detection. In 2011, a mesh with 100 μm openings was inserted into the flow to facilitate the evaporation of larger ice particles in the sample (Rollins et al., 2014). The air is then directed into an internal Herriott cell, which is maintained at ~12 °C by upstream heating. The details of the Herriott cell effective path length and volume vary by the generation of the instrument. An increase in the wavelength of the targeted spectral region from 1.88 to 2.65 μm in 2011, allowed for a significant reduction in the cell volume to its present value of 300 cm³, with a path length of 36 m.

The laser is scanned over the spectral region at 10 Hz, and both direct absorption and second harmonic spectra are recorded (Webster and Heymsfield, 2003). For data reduction of the direct absorption data a Voigt line shape is used in conjunction with spectroscopic parameters measured at JPL, which were within 5% of the HITRAN 2008 values (Rothman et al., 2009). The 2f spectra are mapped to the direct absorption spectra in extensive tests executed prior to flight (Webster and Heymsfield, 2003).

The ALIAS measurements are verified in the field after individual flights, as well as in the laboratory before and after each mission. In the field, the addition of synthetic air containing a known mixing ratio is used as a reference standard for evaluating instrument performance. In the laboratory, more extensive calibrations involve sampling synthetic air containing variable concentrations of water vapor with flows that are continuous over several days. A Thunder Scientific 3900 humidity generator is used to generate these flows, with the values cross checked by a Vaisala humidity sensor (Rollins et al., 2014). To obtain accurate measurements of the isotopic ratios of H₂¹⁶O to its heavy isotopologues, calibrations are executed with a certified water sample with known isotopic composition (Webster and Heymsfield, 2003).

Harvard Herriott Hygrometer

The Harvard Herriott Hygrometer (HHH) measures water vapor via direct absorption in the near-IR utilizing a fiber-coupled tunable diode laser and a multipass Herriott cell (Sargent et al., 2013). In its present configuration, the cell is

integrated into the primary inlet duct of the Harvard Lyman-α hygrometer. Fast flow through the instrument duct, ~70 m s⁻¹, ensures a fast flush time and a rapid time response. The combined instrument, which provides two independent measurements of water vapor in a common duct, has flown missions on NASA's ER-2 and WB-57 aircraft from 2011 to the present. Advances in signal processing and data acquisition systems have made it possible to achieve both high precision and accuracy with a direct absorption technique. HHH has a stated precision of ~<0.10 ppmv for a 1-s integration period, an accuracy of ±5–8%, and a bias uncertainty of ±0.2 ppmv for pressures <500 hPa or mixing ratios ~<600 ppmv (Sargent et al., 2013).

HHH uses a fiber coupled ~1.4 μm DFB laser to scan over a strong water vapor absorption feature, in this case a single rotational-vibrational transition at 7181.16 cm⁻¹. The laser beam is split using a fiber beam splitter to two primary detectors. The main beam, which consists of ~90% of the light, is directed through a hole in one of the two gold-coated mirrors that form the Herriott cell. The mirrors, which are embedded opposite each other ~10 cm apart in the walls of the primary duct of the dual-axis HWV instrument, support a 92 pass pattern for a total absorption path of ~10 m. The beam exits the cell through the entrance/exit hole and is detected by an InGaAs photodetector. In addition to the main cell, 5% of the light is directed through a silicon etalon in order to accurately determine the frequency tuning rate of the laser as it is scanned across the absorption feature.

The use of direct absorption distinguishes HHH from most other airborne laser absorption instruments, which rely on second-harmonic detection to achieve the desired precision. With direct absorption, the accuracy of HHH is theoretically tied only to the accuracy of the state variables, i.e., temperature and pressure, the pathlength, and the spectroscopic parameters from the HITRAN database. However, in practice, any field instrument must be validated in the laboratory to constrain potential sources of systematic error, i.e., water in the path between the laser and detector, which are housed in a thermally regulated pressurized. Pressurization and routine purging with dry air prior to flight reduces the sensitivity of the measurement to internal moisture. Furthermore, the HHH instrument is routinely validated in the laboratory by comparison with multiple standards under conditions that replicate the flight environment as closely as possible. The calibration system and choice of standards are identical to those used for the Harvard Lyman-α instrument.

Harvard Integrated Cavity Output Spectroscopy Isotope Instrument

The Harvard integrated cavity output spectroscopy (ICOS) water isotope instrument is a mid-IR spectrometer which utilizes the off-axis ICOS technique to make *in situ* measurements of water vapor and its major isotopologues, HDO and H₂¹⁸O, in the UTLS in an internal cell (Sayres et al., 2009). The long pathlength provided by this technique provides the sensitivity and accuracy necessary to measure trace atmospheric species at concentrations in the ppbv range. The Harvard ICOS isotope instrument flies on NASA's WB-57 high-altitude research aircraft, and has participated successfully in four field

campaigns from the winter of 2004–2005 to the present. The data acquisition rate of the instrument (1/4 Hz) is a convolution of the data rate needed to achieve the desired precision and the flush time of the detection volume. The stated precision and accuracy of Harvard ICOS varies by molecule. For water vapor the precision is 0.14 ppmv in 4 s, the accuracy is $\pm 5\%$, and the bias uncertainty is ± 0.25 ppmv; for H_2^{18}O the precision is 0.16 ppbv in 4 s, the accuracy is $\pm 5\%$ and the bias uncertainty is ± 0.30 ppbv; and for HDO the precision is 0.10 ppbv in 4 s, the accuracy is $\pm 5\%$, and the bias uncertainty is ± 0.06 ppbv (Sayres et al., 2009).

ICOS is a relatively new cavity absorption technique that forces light from a high-power continuous-wave laser through the rear of the first mirror into a high-finesse optical cavity consisting of a pair of highly reflective mirrors ($R \approx .09998$) (Baer et al., 2002; Paul et al., 2001). The light is trapped in the cavity for tens of microseconds. Light bleeding out of the rear of the second mirror of the cavity is captured and focused on a detector. It is the steady state power emitted from the cavity that is recorded as the laser is scanned across the spectral region of interest. In the Harvard instrument a mid-IR quantum cascade laser operating at 1484 cm^{-1} is used to scan over a region that includes individual rotational lines of H_2O , HDO, and H_2^{18}O . Changes in the steady state power are tied to intracavity losses due to molecular absorption, and the resulting spectrum yields nearly simultaneous measurements of the number density of the absorbing gases within the cell, much as in traditional IR absorption spectroscopy. Because light is trapped within the optical cell for tens of microseconds, however, the effective optical pathlength is several kilometers rather than the tens of meters of a typical Herriott cell, and the instrument sensitivity is correspondingly greater. The Harvard ICOS isotope instrument has an effective optical path-length of nearly 4 km.

The ICOS flow system is designed to (1) minimize the residence time of air in the absorption cell (<1 s), (2) minimize contamination of the air from dead volumes and surfaces, and (3) maintain a constant pressure and temperature within the cell. The flow is maintained by an oil-less scroll pump located at the outlet end of the flow system. Ambient air is brought into the system through a rear facing inlet for particle rejection. The inlet is located well outside the aircraft boundary layer and is heated to prevent condensation. Additional heating downstream ensures that the temperature of the sample is near 25°C , which is equivalent to the temperature of the optical cell. Maintaining a well-defined and constant temperature is important for reducing uncertainties in the measurement. A pinch valve, which allows for high conductance without stagnant volumes, is used to regulate the cell pressure at 53 hPa. A proportional-integral feedback loop between the cell pressure and the valve position maintains a constant cell pressure even as the ambient pressure changes.

The system is capable of bringing both ambient air into the optical cell as well as air from an in-flight gas addition system used for *in situ* calibration. The ICOS instrument is equipped with a gas deck that consists of two high pressure gas bottles capable of delivering dry air or air with a small amount of water and methane to the cell. The first bottle is filled with ultra-pure dry air with less than 0.2 ppmv of water, and the second bottle is typically filled with air containing about 20 ppmv of water

vapor and about 5 ppmv of methane. Periodically during flight the instrument samples air from the calibration gas deck to verify instrument performance *in situ*.

Though water vapor mixing ratios can be calculated directly from spectroscopic parameters obtained from the HITRAN database and physical constants of the ICOS system, laboratory calibrations are routinely performed to verify instrument accuracy and to minimize the impact of any systematic errors. ICOS utilizes the same two water addition systems as Hoxtope for performance verification: a bubbler system and a micro-droplet injector system. The use of dual standards provides a rigorous evaluation of uncertainty.

Chemical Ionization Mass Spectrometer

Principle of Operation

Chemical ionization mass spectrometry is a selective and highly sensitive method for molecular detection and has been used successfully for the *in situ* detection of atmospheric trace species, e.g., nitric acid (HNO_3) and hydrochloric acid (HCl) (Neuman et al., 2000), on NASA's high-altitude research aircraft. Chemical ionization mass spectrometer (CIMS) consists of three primary components: an ion source generator, a mass analyzer, and a detector. Typically, chemical ionization relies on the collision of the molecule of interest, i.e., the analyte, with ions of a reagent gas. The reagent ions are often generated by flowing the gas past radioactive elements in the ion source region. The collision of the reagent ions and the analyte yields the desired ion for detection downstream. An extraction system removes the ions from the sample and directs them through a mass analyzer and ultimately to a detector. Different ion fragments produced during the ionization process will have different mass-to-charge ratios (m/z). These differences allow the mass analyzer to separate the ions. Mass spectra, i.e., measurements of the detector intensity as a function of m/z , are recorded over a selected range. The intensity is directly related to the abundance of each ion present, and thus, through careful calibration, to the initial concentration of the molecular species under investigation.

National Oceanic and Atmospheric Administration CIMS- H_2O

The development of the NOAA CIMS is notable for introducing a new technique to the suite of water vapor measurement methods as well as, for integrating a full calibration system into the flight package. The CIMS- H_2O instrument flew aboard NASA's WB-57 during the MACPEX mission in 2011, and participated in the water vapor intercomparisons during that mission (Rollins et al., 2014). It demonstrated excellent precision, $<\pm 2\%$ ($\pm 2\sigma$, 1 s), and high accuracy $\sim \pm 10\%$ for mixing ratios <150 ppmv (Thornberry et al., 2013).

The CIMS- H_2O was adapted from the existing NOAA CIMS instrument, described by Neuman et al., (2000), to target measurements of H_2O at low mixing ratios in the UTLS aboard the WB-57. Modifications included minor changes to the quadrupole mass spectrometer to convert it from measuring negative ions to positive ions, and the development of a new and compact custom ion source. In contrast to the standard

ionization technique described above, no separate reagent ion flow is needed. Instead a low-pressure flow of sample air is directed through the ion source chamber and exposed directly to α particles emitted from a radioactive source (in this case ^{241}Am foil). The resulting cascade of ion–molecule reactions produces protonated water ions (H_3O^+) as a stable end product, with the number of H_3O^+ ions monotonically, but nonlinearly related to the water vapor mixing ratio in the sample. The relationship is determined empirically through frequent calibration in both flight and the laboratory over the range of atmospherically relevant mixing ratios. The production of H_3O^+ ions depends on the pressure and flow through the ion source. As a result, both are tightly controlled in order to maintain consistent measurement sensitivity.

The instrument samples ambient air through an inlet line that is oriented perpendicular to the aircraft flight direction in order to reject condensed phase water. The inlet port extends below the aircraft fuselage and well outside the boundary layer of the aircraft. The tip of the inlet is heated to prevent condensation, and more heat is applied along the length of the inlet line. The pressures inside the inlet system and ion source region, and the resulting flow through the ion source, are held constant within their respective tolerances over a wide range of ambient pressures and temperatures. Sample residence times are <0.1 s at ambient pressures below 500 hPa allowing for data acquisition at a rate of 10 Hz. The data are then averaged to 1 s.

The CIMS- H_2O instrument was intentionally designed to maintain highly regulated internal temperatures, pressures, and flows, and to be insensitive to changes in the ambient environment. Because of these careful design considerations, CIMS- H_2O operates, and is calibrated, identically in flight and in the laboratory. Calibrations are run every 45 min in flight. The CIMS- H_2O uses two independent calibration sources to sample known mixing ratios of water vapor. The standards are (1) a mixture of water vapor in air near 5 ppmv stored in an onboard stainless steel cylinder, and (2) dynamically controlled flows of air with water vapor mixing ratios between 0.5 and 150 ppmv, generated by the Pt-catalyzed oxidation of onboard H_2 standards (Rollins et al., 2011). In both cases the calibration flow is capable of completely displacing the ambient flow. The uncertainty associated with each system is less than $\pm 10\%$, and the two systems have demonstrated consistency with each other both on the aircraft and in the laboratory. The entire calibration system is routinely checked against an NIST-traceable MBW-373LX reference hygrometer on the ground.

Capacitance Hygrometer

Principle of Operation

Capacitive humidity sensors relate measurements of the capacitance between two charged plates to the ambient water vapor mixing ratio, through the measurement of relative humidity (RH). RH is defined as the vapor pressure of water in the air sample ($p_{\text{H}_2\text{O}_{\text{samp}}}$) divided by the saturation vapor pressure with respect to liquid water or ice air at the temperature of the sample ($p_{\text{H}_2\text{O}_{\text{sat}}}$), i.e., $\text{RH} = p_{\text{H}_2\text{O}_{\text{samp}}} / p_{\text{H}_2\text{O}_{\text{sat}}}$. The plates are separated by a polymer membrane that adsorbs and desorbs water in order to reach equilibrium with the vapor pressure of water in the surrounding air. The

capacitance is, therefore, roughly proportional to the RH, with the exact relationship determined by sensor-specific calibration. Given the measurement of RH, the saturation vapor pressure, which can be calculated using one of several versions of the Clausius–Clapeyron equation (Murphy and Koop, 2005), and a measurement of the sample pressure, the mixing ratio of water vapor can be determined. These devices can provide a very compact means of detecting water vapor *in situ* on board balloons and aircraft.

The response time of these devices is tied to the polymer's ability to absorb and desorb water and on the temperature of the sensor itself. Chemical contamination of the polymer by molecular species other than water vapor can reduce its ability to absorb water and lead to a dry bias in the measurement. Furthermore, at the cold temperatures of the UTLS, e.g., less than -60°C , the response time slows to ~ 1 min, limiting the utility of these sensors in this region.

Each individual device requires careful and regular cleaning and calibration to ensure optimal performance and measurement accuracy.

The Measurement of Ozone and Water Vapor by Airbus In-Service Aircraft Humidity Device

The measurement of ozone and water vapor by airbus in-service aircraft (MOZAIC) project deploys these capacitive devices, which operate autonomously, on commercial aircraft to obtain an extensive data set of RH and water vapor measurements throughout the lower to middle troposphere (Marengo et al., 1998). Regular measurements of ozone, carbon monoxide and nitrogen oxides are also obtained. MOZAIC was initiated in 1993, and represents a highly successful collaboration among the European scientific community, aircraft manufacturers, and the airlines. Plans to extend the MOZAIC research project into the future are in place under the European research consortium Integration of routine Aircraft measurements into a Global Observing System (IAGOS) (Petzold et al., 2012).

For water vapor, the MOZAIC humidity device (MHD) is a special airborne capacitive humidity sensor developed by Aerodata (Braunschweig, Germany). The design utilizes a humidity and temperature transmitter manufactured by Vaisala (Helsinki, Finland). The capacitive sensor (Humicap-H, Vaisala) is paired with a resistive temperature device in a special inlet and associated housing, which is mounted to the aircraft. The inlet, produced by Rosemount (United States) is positioned out of the boundary layer of the aircraft skin, and is designed to simultaneously protect the sensors from particles and minimize contamination of the sample by the inlet walls. The Rosemount housing is installed on a removable plate, located on the left side of the fuselage in the nose of the aircraft.

Experience has shown that with regular calibration, i.e., every 500 h of flight operation, the accuracy of the RH measurement is typically better than $\sim \pm 10\%$, with a response time of approximately 4–10 s, for data acquired up to 12 km. Each unit is calibrated against a water vapor reference instrument over the relevant air temperatures and pressures in the environmental simulation chamber at the Forschungszentrum Jülich, Germany (Helten et al., 1998, 1999). As mentioned above, the performance of these devices in the stratosphere is

not suitable for quantitative water vapor measurements (Bange et al., 2013; and references therein). (Airborne Measurements for Environmental Research: Methods and Instruments, edited by Manfred Wendisch and Jean-Louis Brenguier, is another excellent resource.)

Conclusions

The need for accurate measurements of UTLS water vapor has been repeatedly delineated in the peer-reviewed literature, and much progress has been made to meet the stringent requirements set by the scientific objectives. Results from two recent intercomparison campaigns, one executed in the laboratory (AquaVIT) and one *in situ* (MACPEX), demonstrate improved agreement among several aircraft and balloon-borne hygrometers with long histories of measuring in the UTLS (Fahey et al., 2014; Rollins et al., 2014). These sensors represent four of five techniques discussed here, e.g., frost point, photo-fragment fluorescence, IR absorption, and chemical ionization mass spectrometry. The joint intercomparison results establish that each technique has intrinsic skill for measuring atmospheric water vapor, and that the reference standards and calibration procedures for the individual sensors are consistent with each other to within their stated uncertainties. Both intercomparisons yielded comparable results, with good agreement at the level of $\pm 10\%$ ($\pm 1\sigma$) for mixing ratios >10 ppmv, and agreement degrading to $\pm 20\%$ ($\pm 1\sigma$) for the region between 1 and 10 ppmv. Furthermore, the similarity in the results of the two intercomparison campaigns indicates that the gap between agreement found in the laboratory and agreement demonstrated in flight is closing.

Despite these notable improvements, the intercomparison results also show that routine agreement at the level needed to satisfy the scientific objectives still has not been achieved. Indeed, the history of intercomparisons demonstrates that the level of agreement among instruments can vary over time, and regular monitoring is recommended to address issues of long-term accuracy, especially with respect to trend detection. At present, no research instrument has demonstrated beyond a reasonable doubt either an accuracy or long-term precision of $\sim 5\%$ or better, especially at the lowest mixing ratios encountered in the UTLS. Correcting this will require continued effort from the community in setting instrumental requirements, in maintaining an objective approach to examining the different data sets, as well as in planning and funding repeated intercomparison campaigns of the quality demonstrated during AquaVIT and MACPEX. Furthermore, rigorous calibration and validation efforts on the part of individual experimenters should continue, along with efforts to improve instrumentation. Rollins et al. (2014) suggest that the larger differences, evident at the lower mixing ratios in both intercomparisons, are most likely due to variable background artifacts. Thus, future efforts at improving instrument performance should be aimed at addressing potential sources of measurement bias, whether due to sampling or signal errors. Finally, transparency and documentation of these efforts will enable the community to review the results and maximize their utility in assessing long-term precision and measurement accuracy.

See also: **Aerosols:** Climatology of Stratospheric Aerosols. **Climate and Climate Change:** Greenhouse Effect; Overview. **Clouds and Fog:** Cloud Microphysics. **Global Change:** Upper Atmospheric Change. **Middle Atmosphere:** Transport Circulation; Zonal Mean Climatology. **Observations Platforms:** Balloons. **Ozone Depletion and Related Topics:** Photochemistry of Ozone. **Stratosphere/Troposphere Exchange and Structure:** Global Aspects; Tropopause. **Stratospheric Chemistry Topics:** Hydrogen Budget; Overview; Stratospheric Water Vapor. **Thermodynamics:** Humidity Variables.

References

- Anderson, J.G., Wilmouth, D.M., Smith, J.B., Sayres, D.S., 2012. UV dosage levels in summer: increased risk of ozone loss from convectively injected water vapor. *Science* 337 (6096), 835–839.
- Baer, D.S., Paul, J.B., Gupta, J.B., O'Keefe, A., 2002. Sensitive absorption measurements in the near-infrared region using off-axis integrated-cavity-output spectroscopy. *Applied Physics B: Lasers and Optics* 75 (2–3), 261–265.
- Bange, J., Esposito, M., Lenschow, D.H., Brown, P.R.A., Dreiling, V., Giez, A., Mahrt, L., Malinowski, S.P., Rodi, A.R., Shaw, R.A., Siebert, H., Smit, H., Zöger, M., 2013. Measurement of aircraft state and thermodynamic and dynamic variables. In: Wendisch, M., Brenguier, J.-L. (Eds.), *Airborne Measurements for Environmental Research: Methods and Instruments*. Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, Germany, pp. 7–76.
- Dessler, A.E., Schoeberl, M.R., Wang, T., Davis, S.M., Rosenlof, K.H., 2013. Stratospheric water vapor feedback. *Proceedings of the National Academy of Sciences U.S.A.* 110, 18,087–18,091.
- Diskin, G.S., Podolske, J.R., Sachse, G.W., Slate, T.A., 2002. Open-path airborne tunable diode laser hygrometer, in diode laser and applications in atmospheric sensing. In: Fried, A. (Ed.), *Proceedings of SPIE*, vol. 4817, pp. 198–204.
- Drdla, K., Müller, R., 2012. Temperature thresholds for chlorine activation and ozone loss in the polar stratosphere. *Annals of Geophysics Germany* 30 (7), 1055–1073.
- Forster, P.M.D., Shine, K.P., 2002. Assessing the climate impact of trends in stratospheric water vapor. *Geophysical Research Letters* 29 (6).
- Forster, P.M.D., Shine, K.P., 1999. Stratospheric water vapour changes as a possible contributor to observed stratospheric cooling. *Geophysical Research Letters* 26 (21), 3309–3312.
- Fahey, D.W., et al., 2014. The AquaVIT-1 intercomparison of atmospheric water vapor measurement techniques. *Atmospheric Measurement Techniques Discussions* 7, 3159–3251.
- Hanisco, T.F., et al., 2007. Observations of deep convective influence on stratospheric water vapor and its isotopic composition. *Geophysical Research Letters* 34 (4).
- Hanisco, T.F., et al., 2001. Sources, sinks, and the distribution of OH in the lower stratosphere. *Journal of Physical Chemistry A* 105 (9), 1543–1553.
- Held, I.M., Soden, B.J., 2000. Water vapor feedback and global warming. *Annual Review of Energy and the Environment* 25, 441–475.
- Helten, M., et al., 1999. In-flight intercomparison of MOZAIC and POLINAT water vapor Measurements. *Journal of Geophysical Research: Atmospheres* 104, 26,087–26,096.
- Helten, M., et al., 1998. Calibration and performance of automatic compact instrumentation for the measurement of relative humidity from passenger aircraft. *Journal of Geophysical Research: Atmospheres* 103, 25,643–25,652.
- Herriott, D., Kompfner, R., Kogelnik, H., 1964. Off-axis paths in spherical mirror interferometers. *Applied Optics* 3 (4), 523–526.
- Hintsaa, E.J., Weinstock, E.M., Anderson, J.G., May, R.D., 1999. On the accuracy of *in situ* water vapor measurements in the troposphere and lower stratosphere with the Harvard Lyman- α hygrometer. *Journal of Geophysical Research: Atmospheres* 104, 8183–8189.
- Hurst, D.F., Oltmans, S.J., Vömel, H., Rosenlof, K.H., Davis, S.M., Ray, E.A., Hall, E.G., Jordan, A.F., 2011. Stratospheric water vapor trends over Boulder, Colorado: analysis of the 30 year Boulder record. *Journal of Geophysical Research: Atmospheres* 116 (D2), D02306.
- Jensen, E.J., et al., 2005. Ice supersaturations exceeding 100% at the cold tropical tropopause: implications for cirrus formation. *Atmospheric Chemistry and Physics* 5, 851–862. <http://dx.doi.org/10.5194/acp-5-851-2005>.
- Jensen, E.J., et al., 2008. Formation of large (similar or equal to 100 μm) ice crystals near the tropical tropopause. *Atmospheric Chemistry and Physics* 8 (6), 1621–1633.

- Khaplanov, M., Astakhov, V., Lukjanov, A., Kretova, M., Yushkov, V., 1992. Fluorescent hygrometer for middle atmosphere measurements. In: Proc. 19th Annual European Meeting on Atmospheric Studies by Optical Methods, pp. 540–545.
- Khaykin, S.M., et al., 2013. Arctic stratospheric dehydration – Part 1: unprecedented observation of vertical redistribution of water. *Atmospheric Chemistry and Physics* 13, 11503–11517. <http://dx.doi.org/10.5194/acp-13-11503-2013>.
- Kirk-Davidoff, D.B., Hintsa, E.J., Anderson, J.G., Keith, D.W., 1999. The effect of climate change on ozone depletion through changes in stratospheric water vapour. *Nature* 402 (6760), 399–401.
- Krämer, et al., 2009. Ice supersaturations and cirrus cloud crystal numbers. *Atmospheric Chemistry and Physics* 9, 3505–3522. <http://dx.doi.org/10.5194/acp-9-2035-2009>.
- Kunz, A., Müller, R., Homonnai, V., János, I.M., Hurst, D., Rap, A., Forster, P.M., Rohrer, F., Spelten, N., Riese, M., 2013. Extending water vapor trend observations over Boulder into the tropopause region: trend uncertainties and resulting radiative forcing. *Journal of Geophysical Research: Atmospheres* 118 (19), 11,269–11,284.
- Livingston, J.M., Schmid, B., Russell, P.B., Redemann, J., Podolske, J.R., Diskin, G.S., 2008. Comparison of water vapor measurements by airborne sun photometer and diode laser hygrometer on the NASA DC-8. *Journal of Atmospheric and Oceanic Technology* 25 (10), 1733–1743.
- Marengo, A., et al., 1998. Measurement of ozone and water vapor by AIRBUS in-service aircraft: the MOZAIK airborne program, an overview. *Journal of Geophysical Research: Atmospheres* 103, 25631–25642.
- Mastenbrook, H.J., Dinger, J.E., 1961. Distribution of water vapor in the stratosphere. *Journal of Geophysical Research: Atmospheres* 66 (5), 1437–1444.
- May, R.D., 1998. Open-path, near-infrared tunable diode laser spectrometer for atmospheric measurements of H₂O. *Journal of Geophysical Research: Atmospheres* 103 (D15), 19161–19172.
- May, R.D., Webster, C.R., 1993. Data-processing and calibration for tunable diode-laser harmonic absorption spectrometers. *Journal of Quantitative Spectroscopy & Radiative Transfer* 49 (4), 335–347.
- Meyer, J., 2008. Vergleich und Charakterisierung stratosphärisch messender in-situ hygrometer (in German). Diploma Thesis. University Aachen.
- Murphy, D.M., Koop, T., 2005. Review of the vapour pressures of ice and super-cooled water for atmospheric applications. *Quarterly Journal of the Royal Meteorological Society* 131 (608), 1539–1565.
- Neuman, J.A., Gao, R.S., Schein, M.E., et al., 2000. A fast-response chemical ionization mass spectrometer for in situ measurements of HNO₃ in the upper troposphere and lower stratosphere. *Reviews of Scientific Instruments* 71, 3886–3894.
- Oltmans, S.J., Hofmann, D.J., 1995. Increase in lower-stratospheric water-vapor at a mid-latitude northern-hemisphere site from 1981 to 1994. *Nature* 374 (6518), 146–149.
- Oltmans, S.J., Vomel, H., Hofmann, D.J., Rosenlof, K.H., Kley, D., 2000. The increase in stratospheric water vapor from balloon borne, frost point hygrometer measurements at Washington, DC, and Boulder, Colorado. *Geophysics Research Letters* 27 (21), 3453–3456.
- Paul, J.B., Lapon, L., Anderson, J.G., 2001. Ultrasensitive absorption spectroscopy with a high-finesse optical cavity and off-axis alignment. *Applied Optics* 40 (27), 4904–4910.
- Peter, T., Marcolli, C., Spichtinger, P., Corti, T., Baker, M.B., Koop, T., 2006. When dry air is too humid. *Science* 314 (5804), 1399–1402.
- Petzold, A., et al., 2012. IAGOS - in-service aircraft for a global observing system. In: TAC-3 Proceedings, Prien Am Chiemsee.
- Podolske, J.R., Sachse, G.W., Diskin, G.S., 2003. Calibration and data retrieval algorithms for the NASA Langley/Ames Diode Laser Hygrometer for the NASA Transport and Chemical Evolution over the Pacific (TRACE-P) mission. *Journal of Geophysical Research: Atmospheres* 108 (D20), 8792.
- Read, W.G., et al., 2007. Aura microwave limb sounder upper tropospheric and lower stratospheric H₂O and relative humidity with respect to ice validation. *Journal of Geophysical Research: Atmospheres* 112 (D24).
- Rollins, A.W., et al., 2014. Evaluation of UT/LS hygrometer accuracy by intercomparison during the NASA MACPEX mission. *Journal of Geophysical Research: Atmospheres*, 2013JD020817.
- Rollins, A.W., Thornberry, T.D., Gao, R.-S., Hall, B.D., Fahey, D.W., 2011. Catalytic oxidation of H₂ on platinum: a robust method for generating low mixing ratio H₂O standards. *Atmospheric Measurement Techniques* 4, 2059–2064.
- Rothman, L.S., et al., 2009. The HITRAN 2008 molecular spectroscopic database. *Journal of Quantitative Spectroscopy & Radiative Transfer* 110, 533–572.
- Sargent, M.R., et al., 2013. A new direct absorption tunable diode laser spectrometer for high precision measurement of water vapor in the upper troposphere and lower stratosphere. *Reviews of Scientific Instruments* 84 (7).
- Sayres, D.S., et al., 2009. A new cavity based absorption instrument for detection of water isotopologues in the upper troposphere and lower stratosphere. *Reviews of Scientific Instruments* 80 (4).
- Scherer, M., Vomel, H., Fueglistaler, S., Oltmans, S.J., Staehelin, J., 2008. Trends and variability of midlatitude stratospheric water vapour deduced from the re-evaluated Boulder balloon series and HALOE. *Atmospheric Chemistry and Physics* 8 (5), 1391–1402.
- Schwab, J.J., Weinstock, E.M., Nee, J.B., Anderson, J.G., 1990. In situ measurement of water-vapor in the stratosphere with a cryogenically cooled lyman-alpha hygrometer. *Journal of Geophysical Research: Atmospheres* 95 (D9), 13781–13796.
- Sitnikov, N.M., Yushkov, V.A., Afachine, A.A., Korshunov, L.I., Astakhov, V.I., Ulanovskii, A.E., Krämer, M., Mangold, A., Schiller, C., Ravegnani, F., 2007. The FLASH instrument for water vapor measurements on board the high-altitude airplane. *Instruments and Experimental Techniques* 50 (1), 113–131.
- Solomon, S., Rosenlof, K.H., Portmann, R.W., Daniel, J.S., Davis, S.M., Sanford, T.J., Plattner, G.K., 2010. Contributions of stratospheric water vapor to decadal changes in the rate of global warming. *Science* 327 (5970), 1219–1223.
- SPARC, 2000. Assessment of upper tropospheric and stratospheric water vapour. In: Kley, D., Russell III, J.M., Phillips, C. (Eds.), *World Climate Research Programme, WCRP-113*. http://www.atmosp.physics.utoronto.ca/SPARC/WAVASFINAL_000206/WWW_wavas/Cover.html.
- St Clair, J.M., et al., 2008. A new photolysis laser-induced fluorescence instrument for the detection of H₂O and HDO in the lower stratosphere. *Reviews of Scientific Instruments* 79 (6).
- Thornberry, T.D., Rollins, A.W., Gao, R.S., Watts, L.A., Ciciara, S.J., McLaughlin, R.J., Voigt, C., Hall, B., Fahey, D.W., 2013. Measurement of low-ppm mixing ratios of water vapor in the upper troposphere and lower stratosphere using chemical ionization mass spectrometry. *Atmospheric Measurement Techniques* 6 (6), 1461–1475.
- Troy, R.F., 2007. Field Studies of Ice Supersaturations in the Tropical Tropopause Layer. Ph.D. thesis. University of California at Los Angeles, Los Angeles, CA.
- Vömel, H., David, D.E., Smith, K., 2007a. Accuracy of tropospheric and stratospheric water vapor measurements by the cryogenic frost point hygrometer: instrumental details and observations. *Journal of Geophysical Research: Atmospheres* 112 (D8).
- Vömel, H., et al., 2007b. Validation of aura microwave limb sounder water vapor by balloon-borne cryogenic frost point hygrometer measurements. *Journal of Geophysical Research: Atmospheres* 112 (D24).
- Vömel, H., Oltmans, S.J., Hofmann, D.J., Deshler, T., Rosen, J.M., 1995. The evolution of the dehydration in the Antarctic stratospheric vortex. *Journal of Geophysical Research: Atmospheres* 100, 13919–13926.
- Webster, C.R., Heymsfield, A.J., 2003. Water isotope ratios D/H, ¹⁸O/¹⁶O, ¹⁷O/¹⁶O map dehydration pathways in and out of clouds. *Science* 302, 1742–1745.
- Webster, C.R., May, R.D., Trimble, C.A., Chave, R.G., Kendall, J., 1994. Aircraft (ER-2) laser infrared absorption spectrometer (ALIAS) for in-situ stratospheric measurements of HCl, N₂O, CH₄, NO₂, and HNO₃. *Applied Optics* 33 (3), 454–472.
- Weinstock, E.M., et al., 2009. Validation of the Harvard Lyman-alpha in situ water vapor instrument: implications for the mechanisms that control stratospheric water vapor. *Journal of Geophysical Research: Atmospheres* 114.
- Weinstock, E.M., Smith, J.B., Sayres, D., Pittman, J.V., Allen, N., Anderson, J.G., 2006a. Measurements of the total water content of cirrus clouds. Part II: Instrument performance and validation. *Journal of Atmospheric and Oceanic Technology* 23 (11), 1410–1421.
- Weinstock, E.M., et al., 2006b. Measurements of the total water content of cirrus clouds. Part I: Instrument details and calibration. *Journal of Atmospheric and Oceanic Technology* 23 (11), 1397–1409.
- Weinstock, E.M., Hintsa, E.J., Dessler, A.E., Oliver, J.F., Hazen, N.L., Demusz, J.N., Allen, N.T., Lapon, L.B., Anderson, J.G., 1994. New fast-response photofragment fluorescence hygrometer for use on the NASA ER-2 and the perseus remotely piloted aircraft. *Reviews of Scientific Instruments* 65 (11), 3544–3554.
- Wennberg, P.O., Cohen, R.C., Hazen, N.L., Lapon, L.B., Allen, N.T., Hanisco, T.F., Oliver, J.F., Lanham, N.W., Demusz, J.N., Anderson, J.G., 1994. Aircraft-borne, laser-induced fluorescence instrument for the in-situ detection of hydroxyl and hydroperoxyl radicals. *Reviews of Scientific Instruments* 65 (6), 1858–1876.
- WMO, 1985. *Atmospheric Ozone 1985: Assessment of Our Understanding of the Processes Controlling its Present Distribution and Change*, WMO. Global Ozone Research and Monitoring Project, Report No. 16, vol. II. WMO, Geneva.
- Yushkov, V., Merkulov, S., Astakhov, V., 1998. Optical balloon hygrometer for upper stratosphere and stratosphere water vapour measurements. In: Wang, J., Wu, B., Ogawa, T., Guans, Z.-H. (Eds.), *Optical Remote Sensing of The Atmosphere and Clouds*, Proc SPIE, vol. 3501, pp. 439–445.
- Yushkov, V., Sitnikov, N., Zaitcev, I., Pommereau, J.-P., Garnier, A., 2001. Stratospheric water vapor measurements in the winter arctic with optical fluorescence hygrometer on short and long duration balloons. In: Warmbein, B. (Ed.), *Proceedings of the 15th ESA Symposium on European Rocket and Balloon Programmes and Related Research*, Biarritz, France, ESA SP-471, 28–31. ESA, pp. 263–268.
- Zöger, M., et al., 1999. Fast in situ stratospheric hygrometers: a new family of balloon-borne and airborne Lyman α photofragment fluorescence hygrometers. *Journal of Geophysical Research: Atmospheres* 104 (D1), 1807–1816.

Observations for Chemistry (Remote Sensing): IR/FIR (Satellite, Balloon and Ground)

H Fischer and F Hase, Institute of Meteorology and Climate Research (IMK), Karlsruhe Institute of Technology, Karlsruhe, Germany

© 2015 Elsevier Ltd. All rights reserved.

Nomenclature

$L_{v,O}$ Radiance detected by observer located at x_O ($W/(m^{-2} \text{ sr cm}^{-1})$).

$L_{v,S}$ Radiance of background source located at x_S ($W/(m^{-2} \text{ sr cm}^{-1})$).

$B_v(T)$ Radiance emitted by blackbody source of temperature T ($W/(m^{-2} \text{ sr cm}^{-1})$).

k_v Absorption coefficient (m^{-1}).

$\delta_v(x_1, x_2) = \int_{\tilde{x}=x_1}^{x_2} k_v(\tilde{x}) d\tilde{x}$ Optical thickness between locations x_1 and x_2 (-).

$\tilde{x}$ Is the coordinate along the curved path determined by the line-of-sight, directed from the observer to the source ($x_O < x_S$) (m).

Synopsis

This article provides an overview of passive remote sensing techniques in the infrared and far-infrared spectral domain with emphasis on the observation of the atmosphere's chemical composition. At first, the formation of spectral lines is outlined and the radiative transfer equation is presented. This allows for a description of the appearance of atmospheric spectra as a function of chosen observer location and observing geometry. Different types of measurement devices for recording infrared spectra as grating, Fabry-Perot, and Fourier transform spectrometers are introduced. The quantitative analysis of infrared spectra is surveyed: methods applied for the retrieval of trace gas concentrations from spectra are introduced and some inherent limitations of the remote sensing approach are pointed out. The final section provides some examples of past and current sensors, and presents illustrative results achieved with these instruments.

Introduction

Passive remote sensing of the atmosphere by analyzing the ubiquitous infrared (IR) radiation field offers the possibility to measure a wide variety of atmospheric parameters simultaneously. Large sampling volumes can be investigated, and in the case of spaceborne instruments a considerable fraction of the global atmosphere can be observed within a single day.

The impact of the sampling volume on the radiation field depends on its density, its temperature, and on its chemical composition. The volume may also contain aerosols, which change the radiation depending on their optical properties and allowing a constraint of their optical properties allowing to constrain the chemical composition and the volume distribution of the particles.

By offering such a wide set of observables, the passive remote sensing in the IR spectral domain contributes significantly to many key problems of atmospheric research: The deduced global fields of various trace gases help to improve the understanding of physical and chemical processes. In the case of the stratospheric ozone depletion, the simultaneous observation of ozone together with many of the ozone-related species imposes severe constraints to the chemical transport models. The considerable time period since the invention of the method (ground-based atmospheric measurements started in the 1940s) allows to monitor the chemical evolution of the present atmosphere in response to natural and anthropogenic influences. Spectrally resolved measurements of the global IR

flux upwelling from the top of the atmosphere supply important information concerning the Earth's radiation budget and its drivers, such as temperature structure, distribution of water vapor, and various types of clouds.

This outline is organized in the following way: First, we sketch the origin of the spectroscopic features observable in the IR spectral region, and describe the radiative transfer in the atmosphere. Next, the geometric aspects of the observation are outlined. In the following section, the various experimental methods for probing the radiation field are introduced, their spectral coverage and resolution differing by orders of magnitude. Since the sensor has to be adapted carefully to the requirements defined by the desired data products and their accuracy, the problem of analyzing the measured radiances is at the very heart of remote sensing experiments. A further section is devoted to the fundamentals of this analysis procedure. In the end, we present some examples of remote sensing experiments together with illustrative results.

Spectroscopic Features

The photon energies in the IR and far-infrared (FIR) spectral regions correspond to energies involved in rovibrational and pure rotational transitions of molecules, respectively. Observations with sufficient spectral resolution reveal the complexity of atmospheric IR spectra, the molecular signatures comparable to unique fingerprints of the corresponding constituents. Not all

molecules interact with the IR radiation field, only those with a permanent electric dipole moment can efficiently 'serve as an antenna' when rotating, and only those with a variable electric dipole moment correlated with the displacements of the atomic constituents do so when vibrating. In case of the prevalent molecular constituents, the interaction of a magnetic dipole or an electric quadrupole with the radiation field can also be detected.

In the framework of quantum mechanics, the spectral positions of the spectral lines are determined by the energy differences between discrete eigenstates of the molecule, which are designated by a set of quantum numbers. The resulting line strength is determined by the transition probability between the states involved, and by the fraction of molecules populating the initial level. In the spectra of simple molecules with a high degree of symmetry and constituted by only a few atoms, the rotational lines can be resolved, whereas complex molecules show broadband features, with the underlying structure not resolvable, since the spacing of the lines is narrower than the width of individual lines. In any case, the spectral signatures are highly characteristic, and the concentration of each species in a mixture of gases can be deduced from spectra of sufficient spectral resolution unambiguously. Since the moments of inertia as well as the effective masses involved depend on the atomic masses, the spectral features of isotopomers are characteristic and thus they can be discerned by IR and FIR remote sensing. Figure 1 shows a measured atmospheric spectrum in the 700–970 cm^{-1} range.

The spectral characteristics of aerosols, cloud droplets, and ice particles do not show isolated lines. Instead, broadband spectral features are observed due to the overwhelming large number of interacting molecules involved. The resulting emission, absorption, and scattering characteristics depend on size, composition, shape, and orientation of the particle. This does not mean, however, that all these properties are recoverable by remote sensing of an ensemble of particles in the

atmosphere. If strongly simplifying assumptions are made (parameterized size distribution, homogeneity, and sphericity in case of solid particles), then the chemical composition, the mean size, and the total volume of the observed particles can be deduced to some extent.

Atmospheric Radiative Transfer

To relate the observed IR radiances with the state of the atmosphere, the modeling of the radiative transfer is of crucial importance. First, the raypath through the atmosphere has to be determined, as it is bent according to the local density (and water vapor) gradient. The spectrum observed at a chosen location and along a chosen line-of-sight depends on the radiance of a background source (e.g., sun, ground), if present, and on the emission, absorption, and scattering in each atmospheric element along the raypath to the observer. If scattering of radiation into the line-of-sight can be neglected as valid in the IR in the absence of clouds and aerosols, these contributions depend on local properties of each atmospheric path element solely.

Besides temperature, the pressure has a strong impact on the radiative transfer through an atmospheric element of given partial column. Frequent collisions disturb the molecule during the transition, which leads to a pressure broadening of each spectral line. In higher altitudes, the line width is dominated by the Doppler broadening associated with the thermal motions of the molecules. Since the Doppler width is proportional to the frequency of the line, the boundary between the pressure-broadening and Doppler-broadening height regime increases from about 20 km in the near-IR to about 60 km in the FIR region.

According to the assumption of local thermodynamic equilibrium (LTE), the emission and absorption characteristics of each atmospheric element are fully specified by chemical composition, partial pressures, and a single thermodynamic

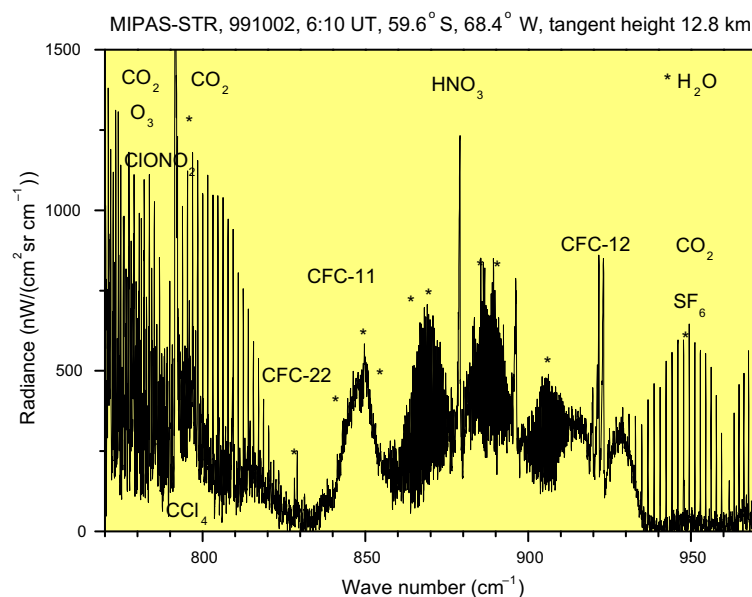


Figure 1 Measured limb spectrum from 770 to 970 cm^{-1} at 12.8 km tangent height using the airborne Fourier transform spectrometer MIPAS-STR. Signatures of various species are marked.

variable, the temperature. If the population of an energy level involved in a transition deviates from the Boltzmann distribution determined by the temperature of the atmospheric element (e.g., due to chemical reactions or due to interaction with energetic solar photons), so-called non-LTE conditions apply for this transition. This is most likely to occur in high altitudes, where the mean free path is large.

If the impact of aerosols cannot be neglected, the calculation of the resulting spectrum may be complicated tremendously. If scattering by the particles into the line-of-sight is neglected, these are treated formally as a further absorbing and emitting component in the air parcel. The total extinction cross section of the particles is used for the calculation of an approximate optical thickness. In a more refined approximation, photons scattered into the line-of-sight are taken into account, but it is assumed that multiple encounters of a given photon with aerosols can be neglected. This requires to calculate radiances from all representative directions at the position of the air parcel. If multiple scattering is taken into account, the resulting radiances and the aerosol distribution become intimately related, and the radiative transfer calculation needs to be performed in an iterative manner or by using Monte Carlo methods.

Without scattering and under LTE conditions, the radiative transfer equation reads

$$L_{v,O} = L_{v,S} \cdot \exp(-\delta_v(x_O, x_S)) + \int_{\tilde{x}=x_O}^{x_S} k_v(\tilde{x}) B_v(T(\tilde{x})) \exp(-\delta_v(x_O, \tilde{x})) d\tilde{x} \quad [1]$$

with

$L_{v,O}$ radiance detected by observer located at x_O ,
 $L_{v,S}$ radiance of background source located at x_S ,
 $B_v(T)$ radiance emitted by blackbody source of temperature T ,
 k_v absorption coefficient,

$\delta_v(x_1, x_2) = \int_{\tilde{x}=x_1}^{x_2} k_v(\tilde{x}) d\tilde{x}$ optical thickness between locations x_1 and x_2 ,

x is the coordinate along the curved path determined by the line-of-sight, directed from the observer to the source ($x_O < x_S$).

Figure 2 shows calculated atmospheric spectra for a spectral interval extending from 783 to 789 cm^{-1} in the mid-IR region.

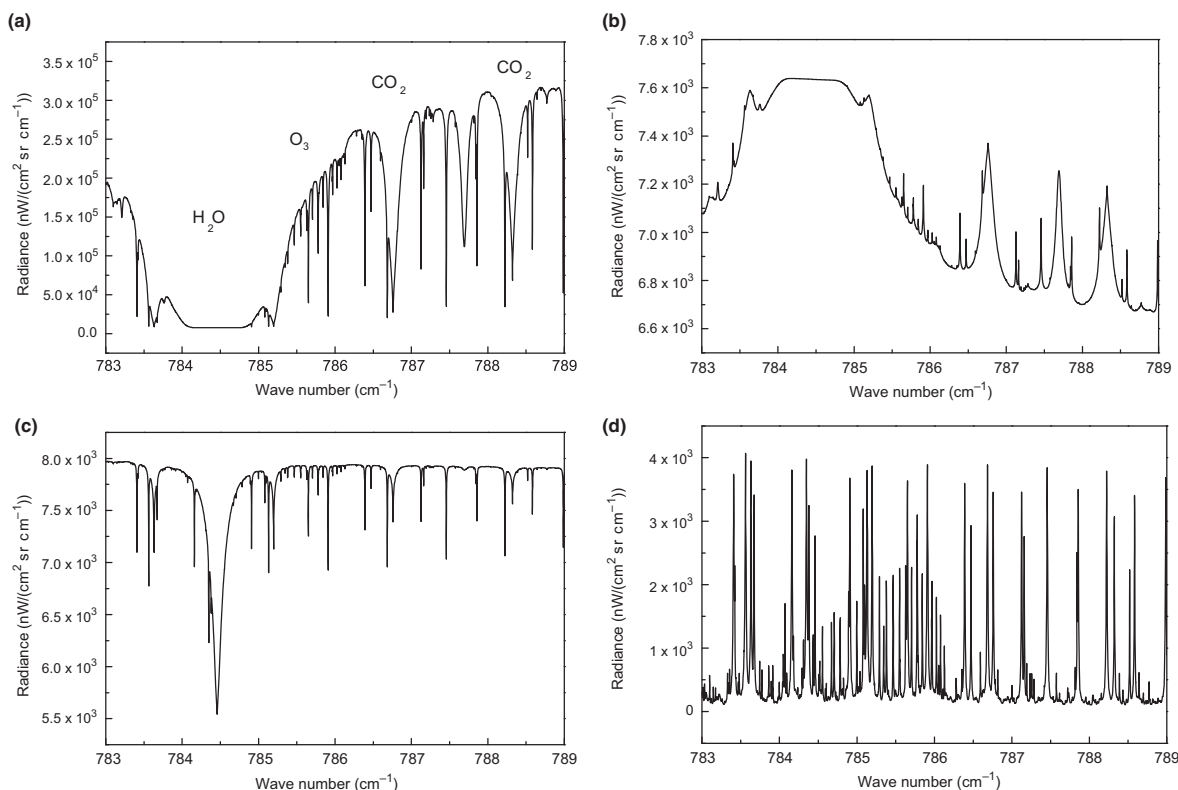


Figure 2 (a)–(d): Calculated atmospheric spectra in the same spectral interval for various observational modes: At top (a) a ground-based solar absorption spectrum is shown. Some of the most striking features are labeled. All narrow absorption lines belong to O_3 , the strong H_2O line around 784.5 cm^{-1} is completely opaque. The second diagram (b) shows the spectrum emitted by the atmosphere as observed from the ground. The spectral features now appear in emission, the O_3 signatures are strongly damped. The saturated H_2O line around 784.5 cm^{-1} reaches the limiting blackbody radiance of the atmospheric temperature near ground. In the third diagram (c), a nadir spectrum according to the view from a satellite is shown. The warmer ground acts as a background source, so the spectral features appear in absorption. Note the changed ordinate scale and the poor contrast between lines and continuum in (b) and (c) as compared to (a). This is, because the temperatures involved are not so different as in case (a), where the hot sun outshines the atmospheric emission. In (d), a limb spectrum for a tangent height of 20 km is shown. The emission of the stratospheric O_3 dominates, whereas the strong H_2O feature seen in (a)–(c) is not detectable, due to the low amount of H_2O in the stratosphere.

According to the radiative transfer equation, the character of the resulting spectra depends on the location of the observer, the line-of-sight, and the temperature of a background source, if present.

Modes of Observation

The observer can choose spectral resolution, location, and line-of-sight of the instrument according to his aims and the justifiable expenditure. Figure 3 shows some possible modes of observation.

A ground-based instrument can be used to observe the atmosphere in emission or absorption (in the latter case the Sun or the Moon serves as a background source). In the FIR region, ground-based observations are impossible, since the atmosphere is by far too opaque due to the strong absorption of water vapor. In the IR, ground-based observations in emission suffer from the fact that, due to the number density and temperature stratification of the atmosphere, the observed signal stems mainly from the lowermost layers of the atmosphere. Therefore, only a few kilometers above the instrument can be investigated by this method. Ground-based observations in absorption with high spectral resolution are in wide use for monitoring of trace gases in the free troposphere and stratosphere, but are also applied to local environmental studies, e.g., plume observations.

An airborne instrument can be operated on an aircraft or balloon. The observations may be performed in absorption or emission. Whereas balloon-borne experiments reach heights of 40 km, even high-altitude aircraft are limited to ceilings of about 20 km. A balloon offers in principle the possibility of long duration flights. On the other hand, an airplane can probe extended atmospheric regions along a preselected flight track. The location of the sensor on an airborne platform permits upward-directed, downward-directed, or tangential line-of-sights. Limb measurements, either by sounding different tangent heights in emission or by observing the rising or

lowering Sun near the horizon, allow to construct vertical profiles with considerable vertical resolution below the flight level, whereas nadir sounding offers high lateral resolution, but poor vertical resolution, since this information has to be deduced from the combination of various spectral intervals with different optical thicknesses, with their contributions to the observed radiances peaking at different height levels. Upward-directed observations from a given flight level have to rely mainly on the pressure broadening and temperature dependence of spectral lines to deduce the vertical distribution of the investigated species as in the case of ground-based measurements, an exception being upward sensing during ascent or descent.

A satellite-borne instrument offers global coverage and continuous operation over several years. The observations are performed in limb emission, nadir, or solar occultation mode during sunset and sunrise. The orbits range from heights of several hundred kilometers with orbital periods of about 1.5 h to geostationary orbits. The oblateness of the Earth can be used to introduce a precession of the inclined orbital plane of the satellite, so that e.g., a sun synchronous orbit can be achieved. In the case of solar occultation, the number of observations, their distribution in local time, and latitude are quite restricted. Limb sounding and occultation observations offer the best sensitivity due to the large effective path lengths probed, but are essentially restricted to tangent heights where the spectral signatures under investigation are not saturated and to line-of-sights that are free from clouds. Nadir sounding is preferable for the investigation of the troposphere, since the impact of clouds is minimized.

Instrumental Techniques

Various experimental devices are in use to measure atmospheric radiances.

If a quite low spectral resolution is sufficient, spectral windows can be defined simply by using an arrangement of optical bandpass filters and/or dichroics (e.g., LIMS Limb

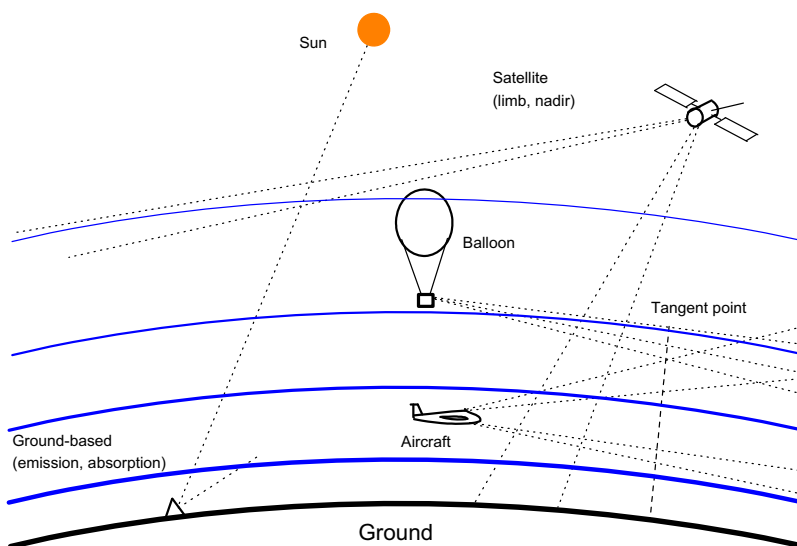


Figure 3 Some examples of remote sensing geometries. For details see text.

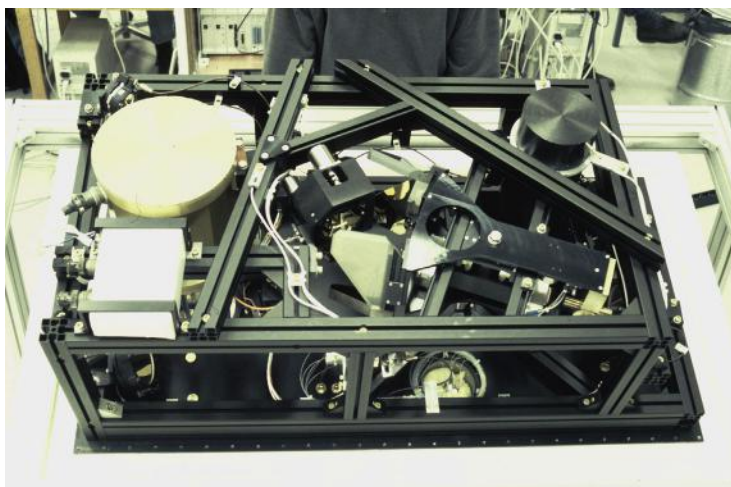


Figure 4 The Fourier spectrometer MIPAS-B. The beamsplitter is located in the quadratic turret near the center of the image. To the right, the pendulum structure generating the path delays can be seen. Some technical details of the experiment are given in the subsection ‘balloon-borne instruments’ in the last section ‘some experiments and results.’

Infrared Monitor of the Stratosphere, launched in 1978). Modern IR filters combine the transmission properties of substrate materials with the Fabry–Perot-like transmission characteristics of a multilayer coating to adjust the desired bandpass. Filters may show some leakage outside the intended spectral interval and may underlie aging effects, their optical properties being not fully stable.

If higher spectral resolving power is needed, grating spectrometers may be favorable. The simplest form is the monochromator, with a single detector element scanning the spectrum sequentially. The grating spectrometer gained considerable attraction due to the increasing availability of detector arrays in the near- and mid-IR spectral region.

The Fourier transform spectrometer generates intensity modulation by varying the optical path difference in a two-beam interferometer. The Fourier transform of the modulated signal yields the spectrum. This technique is capable to reach a very high spectral resolution at a wide spectral coverage, but is interesting also in the case of moderate and low spectral resolving power, because it offers high throughput and compact design. **Figure 4** shows the Fourier transform spectrometer Michelson Interferometer for Passive Atmospheric Sounding (MIPAS)-B, the balloon-borne version of the spectrometer. If the detector noise exceeds the noise associated with the statistics of the incoming photons (most of these may emerge from the instrument itself), the Fourier spectrometer is advantageous over the monochromator, because all spectral elements are measured at the same time. The recent availability of array detectors in connection with increasing data processing capabilities encourages the design of imaging Fourier transform spectrometers. The airborne imaging Fourier spectrometer GLORIA (GLOBal Limb Radiance Imager of the Atmosphere) recorded the first spectra in December 2011. This sensor simultaneously observes 128×128 spectral scenes, covering the $770\text{--}1400\text{ cm}^{-1}$ spectral range at a resolution of 0.1 cm^{-1} .

Fabry–Perot etalons are optical resonators. They combine very high spectral resolution with compact size, but are not in wide use, because the mechanical and optical tolerances are

very demanding, the necessary high reflectivity of the optical resonator is achievable only in quite restricted wavelength regions, and the periodic transmission characteristics makes additional filtering necessary, by either using filters and stages of several etalons, or a predispersing grating.

Gas correlation techniques use the characteristic absorption properties of each gas for its detection: The spectral selection is achieved by a cell filled with the gas under consideration. This cell acts as a highly specific filter for the atmospheric signal. During the measurement, the transmission of the gas probe is modulated by changing the pressure or the path length of the cell. The measured modulation around the average signal level allows to deduce the concentration of the cell gas in the atmosphere. The technique is restricted to a small set of gases with dominating spectral features over a considerable wavelength interval, otherwise it suffers from unacceptable cross-sensitivity to the variability of other species with overlapping spectral features.

In case of the heterodyne detection technique, the signal of a local oscillator is superimposed to the incoming radiation in a mixing element with nonlinear response, a signal oscillating with the beat frequency being generated. The contributions of two symmetrical sidebands above and below the local oscillator are superimposed by this operation, and the undesired sideband can be suppressed by additional filtering before mixing. The signal at the beat frequency is amplified and spectrally analyzed. The technique can reach extremely high spectral resolution, but the spectral coverage is strongly limited. Heterodyne instruments have been realized for frequencies up to the near-IR region.

The Inverse Problem

In remote sensing, the atmospheric variables to be deduced from the measurements are indirectly related to the observed radiances, so an inverse problem has to be solved. Since the radiative transfer in the atmosphere cannot be solved analytically, the analysis relies on the comparison between measured

and computed radiances. A radiative transfer code is required to model the radiances and their derivatives with respect to the wanted variables. The radiative transfer equation (eqn [1]) can be linearized in a vicinity of the current values for the variables, and together with the spectral discretization introduced by the measurement process this allows to apply methods of linear algebra. While the assumption of atmospheric horizontal homogeneity is common in the analysis of spectra, it is not a necessary condition for the radiative transfer modeling and the subsequent inversion. In particular, if the same air volume is probed several times from different positions, a reconstruction of a three-dimensional spatial distribution of variables can be attempted.

The information content of the remote sensing measurement might be insufficient to avoid the propagation of some a priori assumptions on the solution. In some circumstances, the impact of the a priori assumptions on the solution can be profound. If, e.g., the vertical resolution of the remotely sensed measurement is quite poor, it is not permitted to compare the deduced profile with an *in situ* measurement directly. An appropriate smoothing has to be applied to the *in situ* measurement before the intercomparison can be performed.

If the desired variables are not determined by the measurement unambiguously, the solution given by an unconstrained least-squares fit is useless. For illustration, imagine that the elements of the solution vector describe a trace gas profile. If slight changes in the measurement vector lead to tremendous oscillations in the solution vector, the inversion is an ill-posed problem.

Various methods are in use to construct a useful solution: Iterative relaxation methods as the Chahine–Twomey method rely on the fact that oscillatory fine structure in the solution converge more slowly than well-determined broad features. The oscillations in the profile are suppressed by using a finite number of iterations. A truncated singular value decomposition also suppresses unwanted oscillations. The Twomey–Tikhonov method explicitly introduces an additional linear constraint on the solution vector that is weighted vs the fit quality. For example, a smoothness condition on a trace gas profile can be used. Among the most refined inversion methods is the optimal estimation approach: The state of the atmosphere is thought to be represented by an appropriate probability density function. Under the action of the additional spectral information introduced by the measurement, a modified probability density function is generated. The most probable realization of the overall state vector (atmosphere + measurement) might be interpreted as the result of the measurement.

An estimation of errors in remotely sensed data needs an examination of the error mapping from the spectral domain to the set of atmospheric variables deduced from the measurement. To be included into the error budget are the uncertainties of spectroscopic data, the uncertainties of auxiliary atmospheric quantities affecting the spectrum, the impacts of interfering spectral features, instrumental error sources and the impacts of the a priori assumptions. In general, a meaningful error characterization of a remotely sensed set of variables has to take the correlation of errors between the variables into account.

Some Experiments (Past, Present, and Future) and Results

In this section, we present a few examples out of many remarkable remote sensing experiments. The examples are sorted with respect to their platforms, from ground-based to satellite-borne sensors.

Sensors operating in the IR spectral region with sufficiently high spectral resolution can measure more than 30 chemical constituents, as well as temperature. Among these species are O_3 , CO_2 , H_2O , NO , NO_2 , HNO_3 , N_2O_5 , $ClONO_2$, HO_2NO_2 , N_2O , CH_4 , C_2H_6 , SF_6 , CO , $HOCl$, ClO , COF_2 , CFC-11, CFC-12, OCS , H_2CO , and SO_2 . If the experiment relies on the thermal emission of the atmosphere, only the spectral range below approximately 2500 cm^{-1} can be used, therefore some species of interest are not accessible, especially HCl and HF . Due to the limited atmospheric path length as compared to limb sounding, nadir sounding is noticeably less sensitive, and consequently the set of observable species is somewhat limited (H_2O , CH_4 , N_2O , CO_2 , CO , O_3 , HNO_3 , CFC-11, CFC-12, and OCS). The same argument holds for ground-based solar absorption measurements in comparison to tangential absorption measurements.

Sensors operating in the FIR spectral region with sufficient high spectral resolution can measure about 20 chemical constituents including, e.g., O_3 , H_2O , OH , ClO , HCl , HF , HNO_3 , and N_2O .

● Ground-based instruments:

More than 20 ground-based mid-IR Fourier transform spectrometers are operated within the framework of the international organization Network for the Detection of Atmospheric Composition Change (NDACC) at stations all over the world. The spectrometers record solar absorption spectra with about 0.003 cm^{-1} resolution. This is sufficient to resolve the shape of the atmospheric lines. From these spectra, total column amounts of many stratospheric and tropospheric gases, e.g., O_3 , HNO_3 , CO , CH_4 , HCN , and C_2H_6 are deduced. In the case of strong absorbers, such as O_3 and CH_4 , information concerning the vertical concentration profiles can be retrieved in addition.

In Europe, the observations performed by the University of Liege, Belgium, atop the Jungfraujoch ridge document an outstanding tradition in remote atmospheric measurements: Already in 1950, using a grating spectrometer, M. Migeotte inferred the presence of CH_4 and CO from IR absorption features in atmospheric spectra. Since 1975, Fourier transform spectrometers are used for the measurements. Spanning more than three decades in the meantime, the time series deduced from the measurements are well suited for trend analyses. Figure 5 shows trends of several species as deduced from the measurements.

More recently, a second network applying high-resolution ground-based Fourier transform infrared (FTIR) spectrometers has been established. The Total Carbon Column Observing Network performs solar absorption measurements in the near-IR. The oxygen bands observable in this spectral region provide a valuable reference, enabling the measurement of column-averaged dry air mole fractions of greenhouse gases such as CO_2 and CH_4 with unprecedented accuracy.

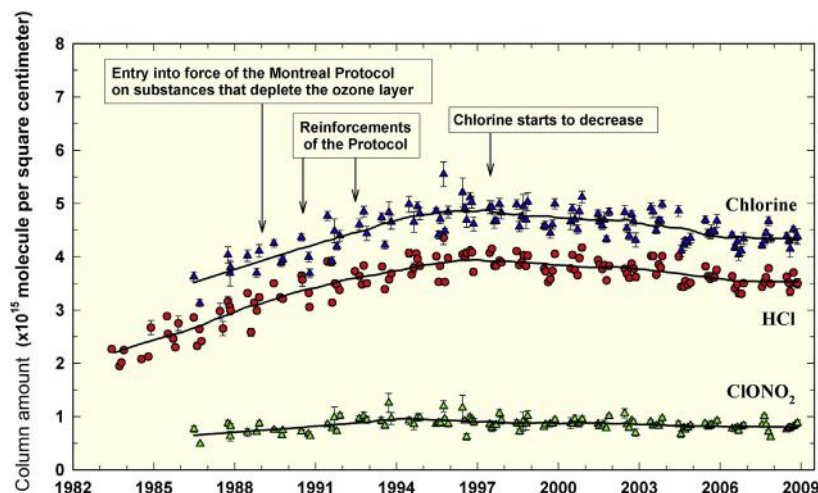


Figure 5 Trends in inorganic chlorine from 1983 to 2009 as deduced from ground-based FTIR measurements at the high-altitude site Jungfraujoch (Suisse). With kind permission of Demoulin, P. University, Liège.

- Airborne instruments:

The SAFIRE-A (Spectroscopy of the Atmosphere using Far-Infrared Emission-Airborne) instrument is operated on a Russian M-55 aircraft capable of reaching a maximum altitude of 21 km. SAFIRE-A, developed at the Institute of Applied Physics (IFAC) in Florence, Italy, is a Fourier transform spectrometer measuring in the FIR region from 10 to 250 cm⁻¹ with a spectral resolution of 0.004 cm⁻¹ (equivalent to 125 cm max. optical path difference). The interferometer comprises a compact pantographic dual slide design. In the wave number region from 10 to 50 cm⁻¹ bolometers are used, whereas the higher frequency channels are equipped with semiconductor detectors. The main data products of SAFIRE-A are O₃, ClO, HNO₃, and N₂O. On the same airplane, a cryogenic Fourier transform spectrometer for the mid-IR region Michelson Interferometer for Passive Atmospheric Sounding-STR atmospheric aircraft (MIPAS-STR) is operated, that is a descendant of the MIPAS-B experiment (see next paragraph). Precursor instruments of SAFIRE-A have been flown successfully aboard stratospheric balloons.

- Balloon-borne instruments:

The MIPAS-B experiment is a cryogenic limb-sounding Fourier transform spectrometer. The instrument was designed and built by the Institut für Meteorologie und Klimaforschung (IMK) in Karlsruhe, Germany. The flights started in 1989 and, up to now, 23 MIPAS-B flights were performed. The whole instrument is cooled to 195 K, so that the thermal self-emission of the atmosphere can be studied with high spectral resolution. The experiment aims in particular to improve our knowledge of the stratospheric ozone loss by probing the chemical composition of the stratosphere. For the observation of the chemical evolution inside the polar vortex, the independence from solar radiation is especially beneficial. The detector system is equipped with four Si:As Blocked Impurity Band detectors, and allows the simultaneous observations in four spectral channels covering 750–1900 cm⁻¹ with a resolution of 0.033 cm⁻¹ (equivalent to 15 cm max. optical path difference). The first

arctic vertical profiles of the species ClONO₂ and N₂O₅ were provided by this experiment, and the partitioning of chlorine and nitrogen species as well as denitrification events inside the Arctic vortex were studied in some detail. Figure 6 shows the NO_y-budget in the polar vortex, as deduced from MIPAS-B measurements.

The MkIV FTIR spectrometer was designed and built at the Jet Propulsion Laboratory in Pasadena, California, in 1984. The instrument records solar absorption spectra in the entire 700–5700 cm⁻¹ spectral range using two detectors. The resolution is 0.01 cm⁻¹. Vertical profiles of more than 30 species are deduced from sunrise/sunset observations at float height or from balloon ascent/descent measurements. The experiment aims to monitor trace gases in order to test chemical transport models, and to perform correlative measurements for other experiments. The MkIV instrument successfully completed 21 balloon flights since 1989, and has been operated on the NASA DC-8 airplane, too. Moreover, it is used for ground-based measurements in the framework of the NDACC.

The Limb Profile Monitor of the Atmosphere (LPMA) spectrometer operated by the Université Pierre et Marie Curie in Paris, France, follows the same measurement principle as the MkIV instrument. The LPMA experiment demonstrates that a commercial FTIR spectrometer (a Bomem DA-2, offering 0.01 cm⁻¹ spectral resolution) can be successfully operated on a balloon. A balloon instrument for nadir observations has been derived from the LPMA spectrometer, to yield test data for the forward model calculations and the retrieval procedure involved in the Infrared Atmospheric Sounding Interferometer (IASI) satellite experiment.

- Instruments in space:

The ATMOS instrument (Atmospheric Trace Molecule Spectroscopy Experiment) represents the outgrowth of spectrometers developed by the Jet Propulsion Laboratory from 1972 on. The ATMOS sensor itself was designed and built by NASA's contractor Honeywell Electro Optics Center, and was flown four times on the space shuttle in the solar occultation

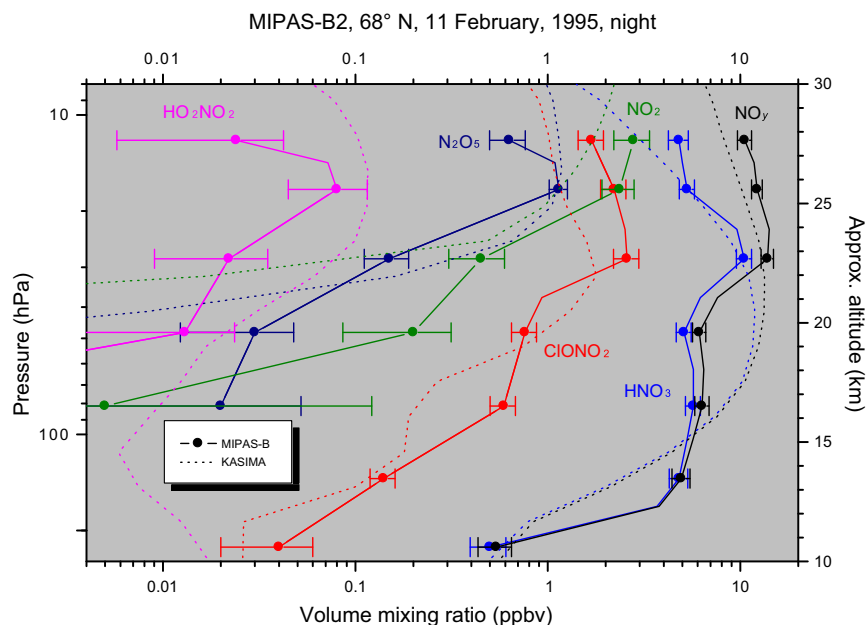


Figure 6 Profiles of the relevant members of the NO_y chemical family and total NO_y in the nighttime polar vortex as deduced from MIPAS-B measurements. The dotted lines refer to the predictions of the chemical transport model KASIMA.

mode since 1985. More than 30 species have been deduced from the spectra covering the range from 600 to 5000 cm^{-1} with a resolution of 0.01 cm^{-1} . **Figure 7** shows a set of trace gas profiles deduced from ATMOS measurements. Moreover, on the basis of essentially pure solar spectra recorded at high tangent heights, an atlas of the solar spectrum was compiled.

A recent satellite-borne spectrometer with similar characteristics is Atmospheric Chemistry Experiment (ACE), being in

orbit since August 2003. The ACE measurements are extending the ATMOS data set with the advantage of performing observations on a strongly inclined orbit over many years.

The Upper Atmospheric Research Satellite was launched in September 1991 and was in operation until September 2001. Ten instruments yielded a comprehensive suite of geophysical parameters. In the given context, three of these instruments deserve special attention: CLAES is the Cryogenic Limb Array

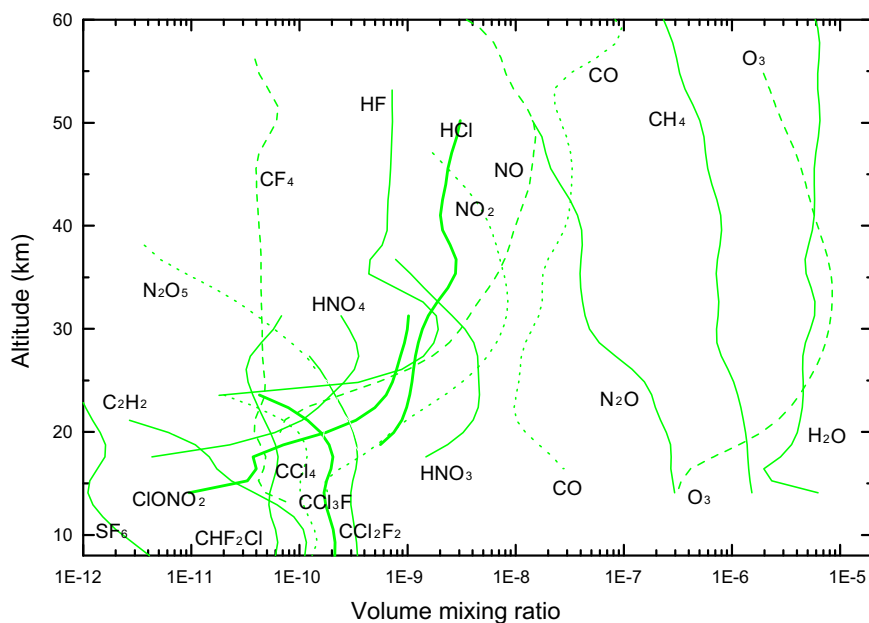


Figure 7 Trace gas profiles deduced from ATMOS solar occultation spectra during the Spacelab 3 mission in 1985. Note that the range of mixing ratios spans about seven orders of magnitude. The extensive set of species underpins the capabilities of remote sensing in monitoring the chemical composition of the atmosphere. ATMOS data with kind permission of Gunson, M. JPL.

Etalon Spectrometer. The instrument was in operation until May 1993, when the cryogenics used to cool the spectrometer were used up. Among the data products were O_3 , ClONO_2 , N_2O , CH_4 , N_2O_5 , HNO_3 , H_2O , CFCl_3 , and CF_2Cl_2 observed in the spectral range from 780 to 2900 cm^{-1} . ISAMS, the Improved Stratospheric and Mesospheric Sounder, is also a limb sounding radiometer. It used a combination of pressure-modulated and wideband channels in the $600\text{--}2500\text{ cm}^{-1}$ range to perform measurements of the species O_3 , CO , HNO_3 , NO , NO_2 , N_2O , H_2O , and N_2O_5 . ISAMS made measurements from September 1991 to July 1992. HALOE (Halogen Occultation Experiment) also combines wideband and gas cell correlation techniques in the spectral range from $1000\text{ to }4000\text{ cm}^{-1}$. It yielded profiles of the species O_3 , HCl , HF , CH_4 , H_2O , NO , and NO_2 on about 30 available solar occultations each day from October 1991 until November 2005.

The CRISTA experiment (CRYogenic Infrared Spectrometers and Telescopes for the Atmosphere) is a liquid-helium-cooled cryogenic limb scanner. Four grating spectrometers cover the $140\text{--}2500\text{ cm}^{-1}$ region. Three telescopes are used simultaneously to improve the horizontal coverage. The sensor is mounted on the free-flying satellite ASTRO-SPAS that is released and recaptured by the shuttle. Up to now, CRISTA has been in orbit two times, 1994 and 1997. CRISTA measures about 20 species in the $15\text{--}150\text{ km}$ altitude range and is able to observe small-scale dynamical structures due to its good spatial resolution.

The Interferometric Monitor for Greenhouse gases (IMG) experiment was part of the Japanese ADvanced Earth Observing Satellite mission. It was developed by the Japan Resources Observation System Organization for the Ministry of International Trade and Industry. The satellite was launched in August 1996, and was lost in June 1997. The IMG nadir sounder offered a spectral resolution of 0.05 cm^{-1} in the spectral region

from $700\text{ to }3000\text{ cm}^{-1}$. The central aim of IMG was to map global features in the distribution of CH_4 and CO to improve our knowledge on natural as well as anthropogenic sources and sinks of these gases.

The MOPITT (Measurements of Pollution in the Troposphere) instrument on board the Terra satellite was launched in 1999 as part of NASA's Earth observing programme. The project is a collaboration of Canada and the USA. It monitors the global tropospheric concentration of CO using correlation spectroscopy in several spectral channels between $2200\text{ and }4500\text{ cm}^{-1}$.

The MIPAS sensor on board the European Environmental Satellite Envisat is a limb sounder aiming at the composition of the upper troposphere, stratosphere, and mesosphere. It was developed by the European Space Agency and belongs to the core experiments on Envisat, and was originally proposed by H. Fischer and colleagues at IMK (see MIPAS-B entry). Envisat was launched in March 2002 and operated until April 2012. MIPAS uses a dual port interferometer and a dual slide design. The spectral range from $650\text{ to }2400\text{ cm}^{-1}$ is covered by four spectral channels. The spectral resolution is 0.025 cm^{-1} (Figure 8).

The TES (Tropospheric Emission Spectrometer) instrument, built by the Jet Propulsion Laboratory, is designed to examine the troposphere. The Fourier spectrometer TES has been launched aboard Aura (NASA's Earth Observing System) spacecraft in July 2004. TES has both nadir and limb modes of observation, primarily nadir measurements are recorded; the limb mode ($0\text{--}37\text{ km}$ coverage) is only used for special observations. In the nadir mode, the observations will be performed with a spectral resolution of 0.1 cm^{-1} , in the limb mode, a resolution of 0.025 cm^{-1} will be applied. Another innovative feature is the use of focal plane arrays to measure spectra from 16 adjacent directions of view at the same time.

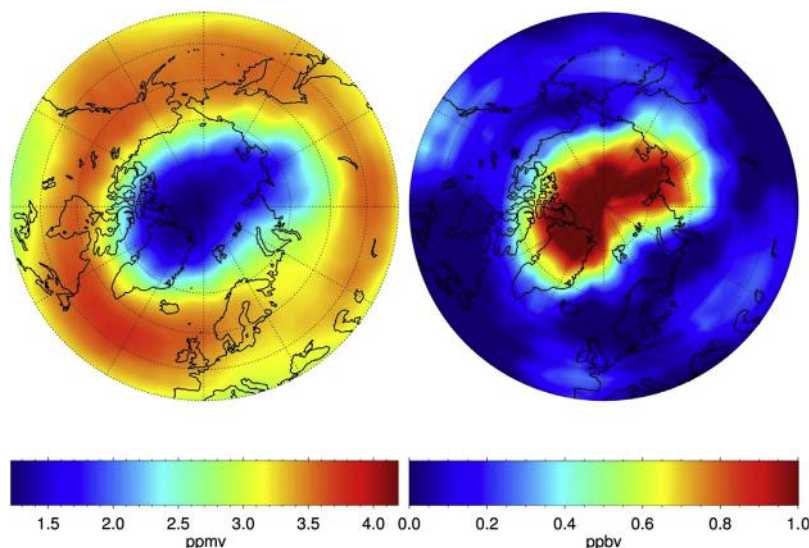


Figure 8 The first Arctic ozone hole as measured by MIPAS in winter 2010/2011, resulting from the prolonged cooling of the Arctic stratosphere. At the same time MIPAS detected elevated concentrations of ClO in the vortex, which proves the conditions for ozone depletion (Sinnhuber et al., 2011. GRL 38, L24814). The left plate shows the ozone volume mixing ratio at 50.0 hPa , the right plate shows the ClO volume mixing ratio at 50.0 hPa measurements taken on 18 March 2011.

Since October 2006 there is a further nadir-looking Fourier spectrometer in orbit. The IASI instrument observes the spectrum between 3.6 and 15.5 μm from a low-altitude sun-synchronous orbit, covering a swath width of 2000 km. The mission objectives are to provide information on temperature, water vapor, and ozone profiles, as well as fractional cloud cover and cloud top temperature. In addition, IASI can also quantify total or partial columns of other atmospheric species as CH_4 , CO , N_2O , HNO_3 , CFCs, and OCS.

See also: **Chemistry of the Atmosphere:** Principles of Chemical Change. **Global Change:** Biospheric Impacts and Feedbacks; Upper Atmospheric Change. **Observations Platforms:** Balloons. **Ozone Depletion and Related Topics:** Long-Term Ozone Changes; Ozone Depletion Potentials; Photochemistry of Ozone; Stratospheric Ozone Recovery. **Radiation Transfer in the Atmosphere:** Absorption and Thermal Emission; Non-Local Thermodynamic Equilibrium; Scattering. **Satellites and Satellite Remote Sensing:** Aerosol Measurements; Research; Temperature Soundings. **Stratosphere/Troposphere Exchange and Structure:** Global Aspects; Local Processes. **Stratospheric Chemistry Topics:** HO_x ; Halogen Sources, Anthropogenic; Halogen Sources, Natural (Methyl Bromide and Related Gases); Halogens; Hydrogen Budget; Overview; Reactive Nitrogen (NO_x and NO_y); Stratospheric Water Vapor.

Further Reading

- Beer, R., 1992. *Remote Sensing by Fourier Transform Spectrometry*. Wiley-Interscience, New York, Chichester, Brisbane, Toronto, Singapore.
- Brasseur, G.P., et al. (Eds.), 1999. *Atmospheric Chemistry and Global Change*. Oxford University Press, New York.
- Carli, B., Carlotti, M., 1992. Far infrared and microwave spectroscopy of the Earth's atmosphere. In: Rao, N., Weber, A. (Eds.), *Spectroscopy of the Earth's Atmosphere and Interstellar Medium*, pp. 1–95.
- Demaison, J., et al. (Eds.), 2001. *Spectroscopy from Space*. NATO Science Series. Kluwer Academic Publishers.
- Dessler, A.E., Burrage, M.D., Groos, J.-U., Holton, J.R., Lean, J.L., Massie, S.T., Schoeberl, M.R., Douglass, A.R., Jackman, C.H., 1998. Selected science highlights from the first 5 years of the Upper Atmosphere Research Satellite (UARS). *Rev. Geophys.* 36 (2), 183–210.
- Fischer, H., 1993. Remote sensing of atmospheric trace gases. *Interdisciplinary Science Reviews* 18 (3), 185–191.
- Houghton, J.T., Taylor, F.W., Rodgers, C.D., 1986. *Remote Sounding of Atmospheres*. Cambridge University Press, London.
- Rodgers, C.D., 2000. *Inverse Methods for Atmospheric Sounding*. World Scientific, Singapore, New Jersey, London, Hong Kong.
- Stephens, G.L., 1994. *Remote Sensing of the Lower Atmosphere*. Oxford University Press, Oxford.
- Siegrist, M.W. (Ed.), 1994. *Air Monitoring by Spectroscopic Techniques*. Wiley-Interscience, New York, Chichester, Brisbane, Toronto, Singapore.
- Visconti, G., et al. (Eds.), 2007. *Observing Systems for Atmospheric Composition*. Springer, New York.

Observations for Chemistry (Remote Sensing): Lidar

G Vaughan, University of Manchester, Manchester, UK

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

The applications of lidar to remote sensing of atmospheric composition are very diverse. Atmospheric particles (aerosols, clouds, and volcanic ash) are readily detected and the altitude of layers is accurately measured, while depolarization provides a means for distinguishing liquid droplets (spherical) from solid particles (irregular). For gaseous composition, two methods are commonly used – Raman scattering and differential absorption, the latter of which has the widest application. Ozone, humidity, and boundary-layer pollutants are commonly measured by lidar from many stations worldwide. Mesospheric metals and radicals may also be measured by resonance fluorescence.

Introduction

This article illustrates applications of lidar to studies of atmospheric chemistry and physics through measurement of trace atmospheric constituents. The fundamentals of lidar, and the different techniques that it can exploit (Rayleigh, Mie, and Raman scattering, differential absorption, Doppler and resonance scattering) are described in articles **Lidar: Atmospheric Sounding** Introduction; Backscatter; Differential Absorption Lidar; Doppler; Raman; Resonance of this Encyclopedia and are not repeated here; the reader is advised to consult those articles if unfamiliar with the lidar technique.

Lidar has a particular strength for atmospheric measurement in its ability to combine the high resolution of *in situ* measurements with the coverage (in space and time) of passive remote sensing. A vertically pointing lidar can measure humidity (for instance) with a vertical resolution of a few tens of meters, and, if atmospheric conditions allow, provide continuous measurements over a period of many hours. An airborne ozone lidar has a lower vertical resolution (a few hundred meters) but will measure a curtain of ozone values ~10 km deep and hundreds or thousands of kilometers long during a typical flight. Such measurements have been crucial in advancing our understanding of the interaction between atmospheric chemistry and dynamics. At the present time, space-borne lidars are being developed, which will extend the benefit of lidar to global measurement, vastly improving the vertical resolution attainable with passive remote sensing techniques, particularly in nadir sounding.

Lidars designed for measurements above the planetary boundary layer usually point only in the vertical direction. In this part of the atmosphere, vertical gradients greatly exceed horizontal gradients and there is no advantage in a scanning capability. For many boundary-layer applications however there is a need to scan the laser beam through a range of zenith angles and azimuths. This is particularly acute for measurements near sources – e.g., monitoring of hydrocarbon emissions from oil refineries or methane leaks from gas pipelines. Some of the most important commercial applications of lidar relate to such problems and to the general issue of pollutant dispersion, so scanning lidar technology is well established in industry. For scientific applications, however, the majority of lidar observations of composition have used a vertically pointing beam.

Aerosol Observations

Aerosols are tiny liquid or solid particles (usually less than 1 µm in diameter) suspended in the atmosphere. They play an important role in nucleating clouds (and so affect climate) and in high concentrations they impact on air quality. The lidar technique is readily applied to aerosol observations since the aerosols provide a direct backscatter signal. This can either be used by itself or converted to the lidar backscatter ratio, R :

$$R = \frac{\text{measured backscattered signal}}{\text{backscattered signal from air alone}}$$

The aerosol-free backscatter can either be measured using rotational or vibrational Raman scattering from atmospheric N₂, or calculated from a coincident temperature profile measured (for instance) by a radiosonde.

One important application of this technique has been in the monitoring of stratospheric aerosol during volcanically perturbed periods (**Figure 1**). Regular lidar observations of stratospheric aerosols began in the late 1970s and have continued until the present day. An example, showing integrated backscatter measured at Garmisch-Partenkirchen (47.5° N, 11° E) is shown in **Figure 2**, and clearly demonstrates the massive perturbations to stratospheric aerosol caused by the eruptions of El Chichon (1983) and Mt Pinatubo (1991), as well as a string of minor perturbations from smaller eruptions. Although the global-scale evolution of the aerosol cloud is best observed from satellites, the ability of lidar to resolve small-scale features and to measure correctly at the lowest levels in the stratosphere makes it an invaluable tool, especially during the early phase of a volcanic event. In the troposphere and lowermost stratosphere volcanic ash presents a hazard to aviation, and it is imperative that the height as well as horizontal extent of an ash cloud be accurately tracked. In 2010, for example, lidar measurements proved invaluable for tracking the ash cloud of the Eyjafjallajökull volcano in Iceland, which caused extensive disruption to air traffic over Europe.

A single laser wavelength is sufficient to monitor volcanic events, but information on the size distribution and composition of aerosols requires multiwavelength lidars. Even with wavelengths spanning the infrared and UV from 300 to 2000 nm, the problem is under-constrained, and modeled theoretical shapes for the size distribution (e.g., lognormal)

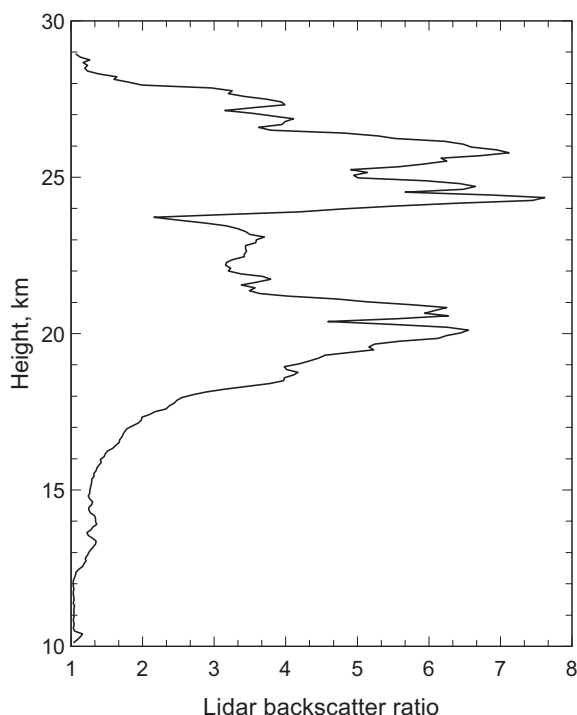


Figure 1 Lidar backscatter ratio profile at 532 nm measured at Aberystwyth, Wales (52.4° N, 4° W) on 27 January 1992, showing stratospheric aerosol caused by the eruption of Mt Pinatubo in June 1991.

must be assumed. Nevertheless, multiwavelength lidars combined with theoretical modeling provide valuable constraints on the properties of stratospheric aerosols.

Further information on the phase of aerosols may be gained by measuring the depolarization of the lidar signal. The laser

beam is usually plane polarized, and atmospheric gases and spherical droplets depolarize either very weakly or not at all. Solid particles, however, depolarize strongly. An important application of this technique to atmospheric chemistry has been in the study of polar stratospheric clouds (PSCs). PSCs come in two forms – ice clouds (PSC type 2), which strongly backscatter and depolarize the laser beam, and non-ice clouds (PSC type 1), which are much smaller, weakly backscattering particles. Based on lidar observations these are further divided into type 1a particles, which depolarize the laser beam, and type 1b particles, which do not. The latter are interpreted as supercooled ternary solutions (liquid droplets) of nitric acid, sulfuric acid, and water, and the former as crystalline nitric acid trihydrate. The distinction is important for atmospheric chemistry since the rates of heterogeneous reactions releasing chlorine are different for the different type 1 PSCs, as is their sedimentation velocity: the type 1b particles are significantly smaller than the type 1a particles. (See the articles on stratospheric composition for further information on the importance of chlorine for ozone destruction). Lidar has played a pivotal role in building this picture of PSC composition and the way in which one type can evolve into another, for instance in a mountain wave.

In the boundary layer, the aerosol population can have a very complex distribution of sizes and composition; this is particularly true for polluted air masses. In such cases the lidar backscatter ratio is difficult to interpret scientifically, and extensive ancillary information is needed to derive, for instance, the aerosol scattering coefficient. For this application, extinction of the laser beam is a more useful measurement. This can be readily accomplished from the N₂ Raman-shifted backscatter, where the signal is from the air alone but the extinction results from both air and aerosol scattering. Currently, the EARLINET network of 28 stations in Europe routinely monitor boundary-layer aerosol by measurements of

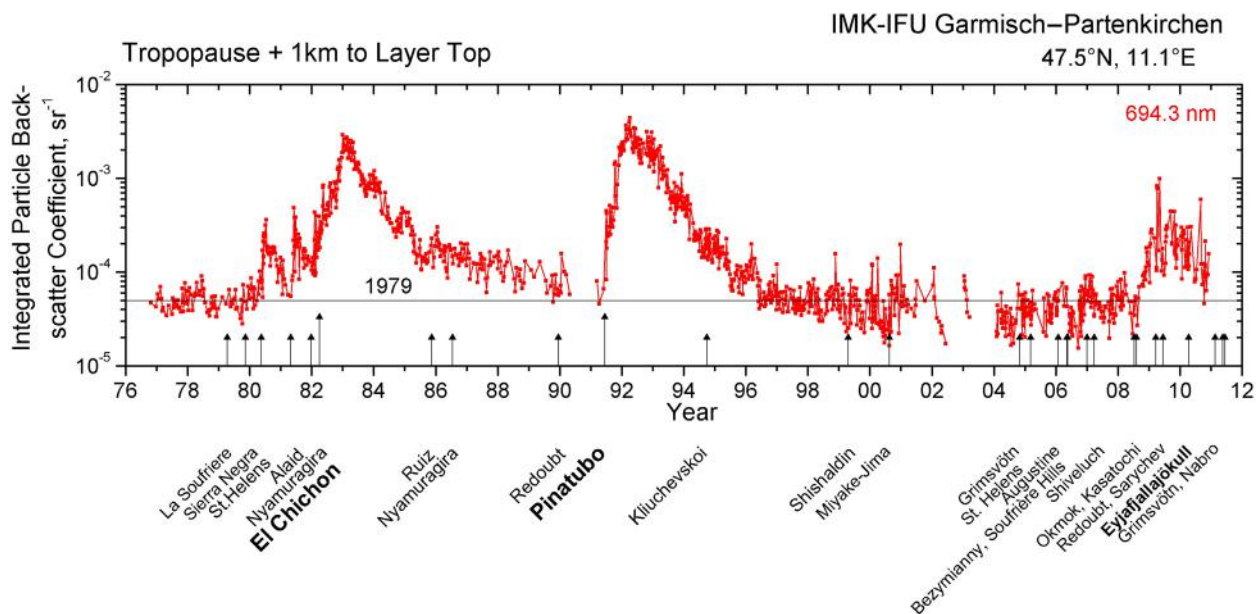


Figure 2 Integrated aerosol backscatter (vertically integrated profile of R-1) at 694.3 nm, measured by lidar at Garmisch-Partenkirchen (47.5° N, 11.1° E) since 1976. Figure courtesy of T. Trickl, IMK-IFU, Karlsruhe, Germany.

backscatter and extinction, building up a climatology of its vertical distribution.

Ozone Observations

One of the main applications of the lidar technique to atmospheric chemistry has been in observations of ozone. Ozone absorbs strongly in the ultraviolet part of the spectrum, and furthermore the absorption coefficient varies rapidly with wavelength. These two facts make the differential absorption (DIAL) method ideal for ozone measurements. The exact choice of wavelengths used in an ozone lidar depends on the atmospheric region to be probed: troposphere or stratosphere. Ozone concentrations in the troposphere are usually less than 150 ppbv, so lidars can operate in the solar-blind region (<300 nm), where the stratospheric ozone layer completely absorbs the sun's radiation. This means that background light levels are very low even in daytime, enabling 24-h lidar operation without sophisticated wavelength selection in the receiver. For stratospheric observations, returns are needed from above the ozone concentration peak, so these lidars must use wavelengths above 300 nm and are not solar-blind.

Most existing ozone lidars use either Nd-YAG or excimer lasers. The fourth harmonic of the Nd-YAG at 266 nm may be used for boundary-layer measurements (typically, three or more wavelengths are needed for accuracy in the boundary layer because of interference by aerosols), but is absorbed too quickly for measurements more than ~ 2 km above the lidar. Stimulated Raman scattering, however, can be used to convert the laser to longer wavelengths, which can cover the entire troposphere – e.g., 289 nm (first Stokes line of deuteride (D_2)), 294 nm (hydrogen deuteride (HD)), and 299 nm (hydrogen (H_2)). The advantages of cost, ease of use, and wavelength stability mean that Nd-YAG lasers are widely exploited for tropospheric ozone lidars.

Excimer lasers have considerable advantages as lidar transmitters, being very bright and with a high pulse repetition frequency. The XeCl laser, with its wavelength of 308 nm, is ideal for stratospheric ozone measurements and is extensively used in ground-based ozone monitoring. The second DIAL wavelength may be provided by another laser (e.g., a third harmonic Nd-YAG at 355 nm), by stimulated Raman scattering from the main laser beam (e.g., 353 nm using H_2) or by using back-scattered Raman-shifted radiation from atmospheric N_2 at 332 nm. Because of the substantial difference between the two wavelengths these configurations are subject to aerosol interference, particularly in volcanically perturbed conditions, which introduces substantial systematic error. One solution to this problem is to use only Raman-shifted backscatter for the DIAL wavelengths – e.g., 332 and 387 nm (the N_2 Raman-shifted wavelength from 355 nm): such a configuration also allows ozone retrieval in PSCs. However, when the stratospheric aerosol concentration is near to background levels, such sophistication is not necessary and a simple two-laser DIAL lidar is sufficient.

There have been a number of intercomparisons between XeCl stratospheric ozone lidars and other techniques such as electrochemical ozonesondes, ground-based microwave, and satellite instruments. Agreement is best between 20 and 30 km, where $\pm 5\%$ between different lidars, and between

lidar and sonde may be expected. Above 30 km the sondes are unreliable and below 20 km systematic errors in the lidars become more apparent. The accuracy of a DIAL lidar tends to be limited by precision at higher levels (typically determined by the photon count rate) and systematic effects such as overloading of the photo multiplier detectors at lower levels. Agreement between different XeCl lidars, and between lidars and ozonesondes, is generally around 10% between the tropopause and 20 km.

Lidars designed to measure in the troposphere and the lowermost stratosphere normally use wavelengths closer together than stratospheric lidars: e.g., the 289/299 nm pair obtained from an Nd-YAG laser, by stimulated Raman scattering. These lidars generally agree with other techniques (ozonesondes, aircraft-borne UV spectrometers) to better than 10%, and are less susceptible to aerosol interference than their stratospheric counterparts. Very high aerosol concentrations in the planetary boundary layer do present a problem, however, and a lidar with three or more wavelengths is needed for accurate ozone measurements in such conditions. In the lowermost stratosphere, under background aerosol conditions, a DIAL wavelength pair such as 299 and 316 nm (the latter is the second Stokes line from the Raman scattering of 266 nm by D_2) can measure profiles with an accuracy better than 5%.

Because of the accuracy and inherently good vertical resolution of the lidar technique, ozone lidars are well suited for routine monitoring of the ozone layer and for validation of satellite measurements. In the troposphere and lower stratosphere the ability to make continuous measurements over time confers an extra advantage. Ozone in this part of the atmosphere is generally considered a tracer, in other words changes in its concentration at a particular location are due to transport rather than chemistry. The distribution of ozone is intimately linked to weather systems, and measurement of the ozone distribution gives an immediate picture of the effect of weather systems on transport in the atmosphere. One of the most successful applications of tropospheric ozone lidars has been in studies of stratosphere–troposphere exchange. Stratospheric ozone concentrations are many times higher than in the troposphere, and a layer of stratospheric air drawn into the troposphere is readily detected by lidars. Furthermore, as the layer advects over the lidar site, its horizontal structure is revealed. An important example is the tropopause fold, a layer of stratospheric air extruded down into the frontal zone beneath a jet stream. The converse – layers of tropospheric air transported into the lower stratosphere by breaking Rossby waves – are also readily detected by lidar.

These advantages of temporal continuity and vertical resolution are exploited to the full in airborne lidars. Rather than wait for the atmosphere to bring an interesting event to a stationary lidar, an aircraft takes the instrument to the event, and has become a standard tool for studies of stratosphere–troposphere exchange and long-range transport of pollution. A cross section of a tropopause fold measured by an airborne lidar is shown in [Figure 3](#); it clearly depicts the morphology and scale of this event. Equally valuable is the ability of a downward-pointing lidar to map out the extent of a pollution plume, either from an individual city or in the outflow from a continent to a nearby oceanic region. An example is shown in [Figure 4](#).

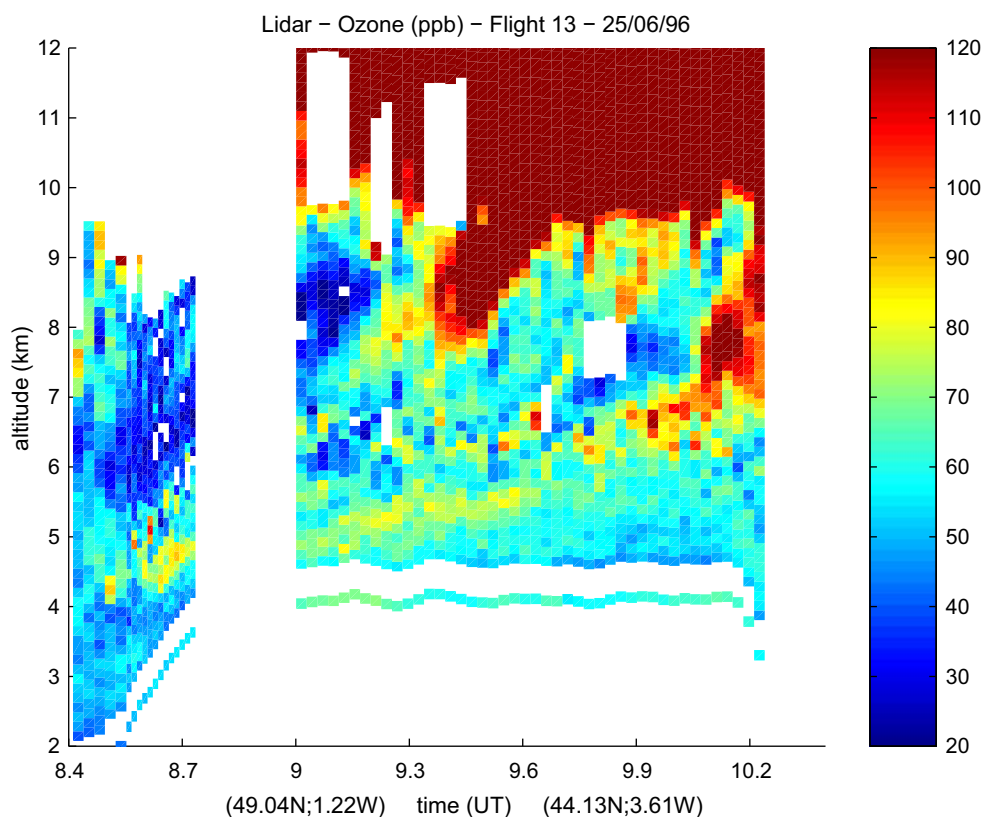


Figure 3 Tropopause folds revealed by airborne ozone lidar on 25 June 1996 over Brittany and the Bay of Biscay, as layers of high ozone sloping downward into the troposphere. Figure courtesy of G. Ancellet, Service d'Aéronomie, Paris.

Humidity Observations

Accurate measurements of atmospheric water vapor, especially in the dry, cold air of the upper troposphere and stratosphere, are notoriously difficult. Two lidar techniques may be used to measure water vapor, which provide a valuable source of observations in their own right and for intercomparison with other techniques.

The most direct measurement of water vapor is by Raman scattering. The vibrational Raman spectrum of a particular molecule provides a unique spectral signature, so that by spectrally analyzing the backscattered radiation from a vertically pointing laser beam the composition of the atmosphere may in principle be determined. In practice the scattering from most minor constituents is too weak to be detected, or is masked by returns from O_2 and N_2 . The two important exceptions are methane (see below) and water vapor. The first vibrational Raman line of water vapor, at 3652 cm^{-1} , is well separated from O_2 at 1556 cm^{-1} and N_2 at 2331 nm^{-1} . The ratio of backscattered intensity from H_2O and N_2 should give a direct measurement of humidity mixing ratio unaffected by aerosol and by most systematic errors in the lidar (e.g., alignment). This is particularly attractive for a reference instrument – it turns out that the largest source of systematic error in a Raman humidity lidar is inadequate knowledge of the scattering cross sections of H_2O and N_2 . An example of a lidar humidity profile compared with a nearby radiosonde is shown

in Figure 5. Long integration times are required to obtain good measurement precision in the upper troposphere.

Vibrational Raman scattering is very weak – thousand times weaker than Rayleigh scattering. A Raman lidar therefore needs a powerful laser and large collecting optics, as well as excellent wavelength selection to avoid interference by elastically scattered and background light. However, a Raman lidar does have a major advantage for ground-based operation: the weak scattering means there is no problem observing the very dry upper troposphere through the relatively moist lower layers. This sharp decrease with altitude of atmospheric humidity militates against the DIAL technique for ground-based profiling. For aircraft lidars, however the situation is reversed – the aircraft can look down through progressively more humid layers. Greater absorption of the laser signal with depth compensates for the reduced signal-to-noise ratio, so ensuring accurate measurements over a range of many kilometers with one wavelength pair. This, together with the high-powered lasers and large collecting optics required for Raman lidars, makes DIAL the favored technique for airborne lidars. Observations in the 730, 815, and 940 nm absorption bands (using alexandrite, Ti:sapphire, and optical parametric amplifier lasers, respectively) have all been made from research aircraft, together covering the entire troposphere and lower stratosphere up to about 17 km. These observations have similar applications to those of airborne ozone lidars – e.g., studying the extent and evolution of tropospheric haze layers and the

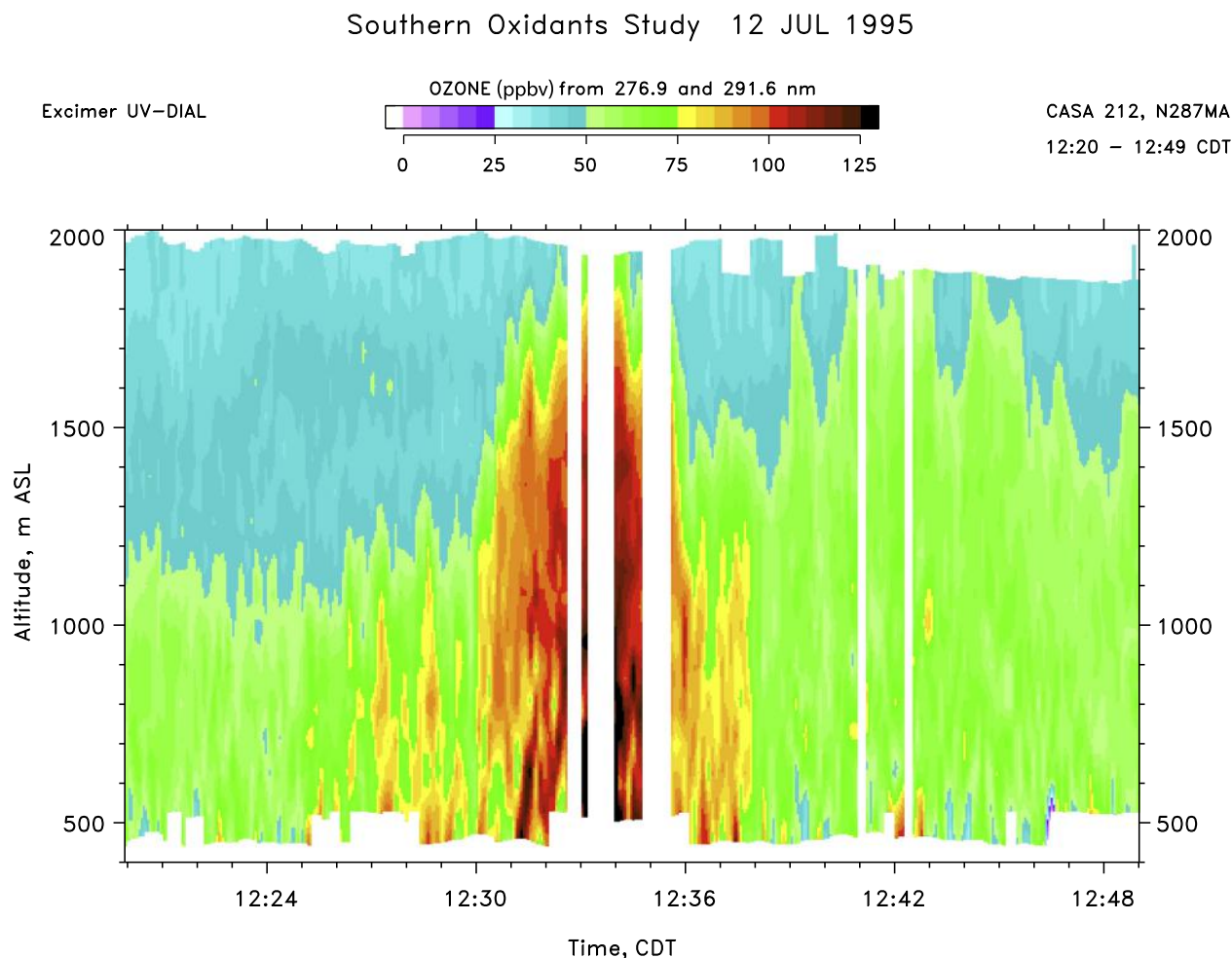


Figure 4 Cross section of ozone concentration in ppbv measured with NOAA/ETL's airborne ozone lidar during the 1995 Southern Oxidants Study field campaign in the Nashville, Tennessee area. The data were taken on 12 July 1995 during a flight leg extending from the northwest to the southeast of Nashville. Note the region of high ozone concentrations over the Nashville urban area indicated by the red and black colors. Figure courtesy of R. Alvarez, NOAA Environmental Technology Laboratory, Boulder, CO 80305, USA.

structure of streamers of low-latitude air in the lower stratosphere.

Pollutant Measurements in the Boundary Layer

One of the most important applications of lidar lies in its ability to make remote measurements of pollutant emissions from industrial sources. In the modern regulatory environment such a capability is of great commercial value, and many lidar systems have been developed around the world to provide it. Typically, such a lidar is mounted in a truck or van and driven to the perimeter of the site under investigation. The lidar performs a two-dimensional scan of the region upwind and downwind of the emission source, mapping the pollutant concentration as a function of height and elevation angle. When combined with coincident wind measurements (e.g., from a meteorological tower), the flux of pollutant being emitted from the site may be estimated. The lidar's ability to make range-resolved measurements makes

it the primary tool for remote sensing investigations of this type.

Early applications of lidar to pollutant monitoring used the Raman technique. Raman has the advantage that it does not require a specific laser wavelength, but it has two main disadvantages as already mentioned: Raman backscatter is very weak, and the rotational-vibrational bands of O_2 and N_2 can mask Raman lines from minor constituents. An advantage of Raman is that good range resolution can be achieved if the back-scattered signal is sufficiently large. A typical application of the Raman technique has been to search for leaks from gas pipelines – concentrations of $\sim 1\%$ methane may readily be detected 1 km away.

Recent work in this field, however, has almost exclusively exploited the DIAL technique. As new and better tunable laser transmitters are developed – especially in the infrared – new applications for DIAL are opened up. The two wavelength regions most commonly exploited are the ultraviolet between 230 and 300 nm, and the infrared between 3 and 5 μm . In the former, gases such as nitric oxide (226 nm),

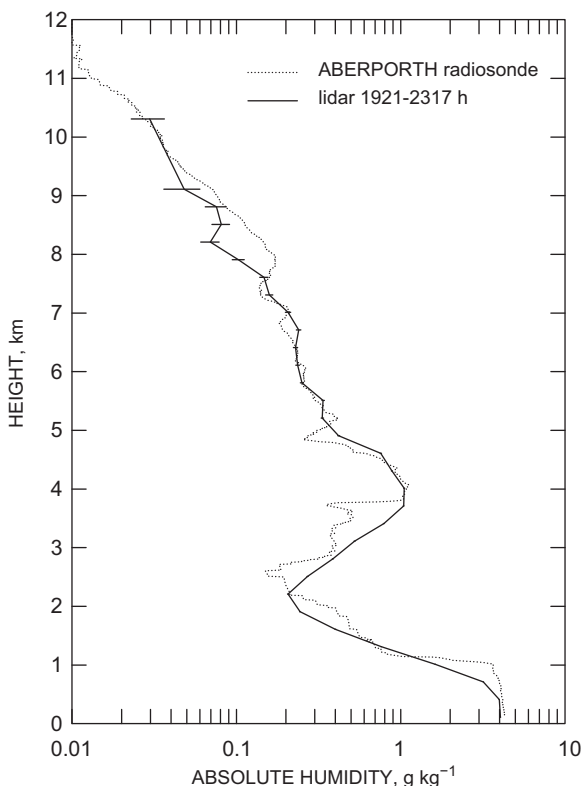


Figure 5 Humidity profile measured by Raman lidar at Aberystwyth, Wales (52.4° N, 4° W) on 19 October 1999, compared with a radiosonde (Vaisala RS80) launched from Aberporth (42 km away) at 2330 h. The lidar profile is an average of four hours' data. Error bars on lidar profile denote 1 standard deviation. Radiosonde profile provided by UK Met Office.

sulfur dioxide (287 nm), toluene (267 nm), and benzene (253 nm) have sharp absorption lines making them suitable for DIAL detection: this spectral region is also solar-blind, with obvious advantages for daytime operation. Dye lasers,

pumped by Nd-YAG and frequency-doubled Dye lasers, pumped by Nd-YAG, were used in many early lidars of this type, but the advent of tunable solid-state lasers such as Ti:sapphire has led to more powerful and compact transmitters. Closely related to these lidars is the NO_2 lidar, typically operating near 450 nm, which is used to detect the emission of this gas from combustion sources.

Another gas that can be detected readily by UV lidar is mercury vapor. Mercury is released from its ore, cinnabar, by roasting it, and in regions where this extraction is practised (particularly the huge cinnabar deposits at Almadén in Spain), there is an obvious concern about the poisonous effects of mercury vapor. Since mercury has a very strong electronic transition at 253.6 nm it is readily detected (in high enough concentration) by DIAL lidar, and indeed much work has been conducted at Almadén on the dispersion of mercury vapor using DIAL.

The advent of high-quality infrared nonlinear optical materials has led to a new class of bright, stable infrared laser beams, usually obtained through nonlinear mixing of Nd-YAG and dye lasers. These infrared sources are tunable and have very narrow line widths (0.1 nm is attainable), opening up the possibility of DIAL measurements in the infrared. Very many molecules of atmospheric interest have vibrational-rotational bands in the near infrared (2–5 μm), and provided that interference with water vapor and CO_2 bands can be avoided, they can be measured with infrared DIALs. Examples are methane (CH_4), acetylene (C_2H_2), ethylene (C_2H_4), and ethane (C_2H_6). The obvious application for such lidars is to measure emissions from petrochemical plants and storage facilities, although they are also useful for tracking plumes for several kilometers downwind of industrial sources.

An example of a mobile lidar designed for pollution measurements is that operated by the UK's National Physical Laboratory, London (Figure 6). This uses an Nd-YAG-pumped dye laser for UV measurements via a second harmonic generator and for IR measurements via a LiNbO_3 difference frequency meter. The gases measurable with this technique



Figure 6 Mobile DIAL lidar system operated by the National Physical Laboratory, London, UK. Figure courtesy of T. Gardiner, National Physical Laboratory.

Table 1 Typical parameters of the National Physical Laboratory DIAL lidar system

Species	Laser wavelength	Measurement sensitivity at 200-m range, ppbv
Nitric oxide, NO	226 nm	5
Nitrogen dioxide, NO ₂	450 nm	10
Sulfur dioxide, SO ₂	300 nm	10
Ozone, O ₃	289 nm	5
Mercury vapor, Hg	254 nm	0.5
Benzene, C ₆ H ₆	253 nm	10
Toluene, C ₇ H ₈	267 nm	10
Methane, CH ₄	3.42 μm	50
Ethane, C ₂ H ₆	3.36 μm	20
Ethylene, C ₂ H ₄	3.35 μm	10
Acetylene, C ₂ H ₂	3.02 μm	40
Hydrogen chloride, HCl	3.42 μm	20
Nitrous oxide, N ₂ O	2.90 μm	100
Methanol, CH ₃ OH	3.52 μm	200

(together with their typical detection limits for a 100-m diameter plume centered at 200 m range) are shown in Table 1. A 12-m mast carries wind, temperature, and humidity sensors, used to convert the number density to mixing ratio and to calculate fluxes across the plane scanned by the lidar. The system is calibrated by reference to standard cells with known concentrations of the measured gases. It has been used to measure volatile organic compound emissions from more than 20 different petrochemical facilities, including oil refineries, retail petrol stations, and an ethylene processing plant, and also for methane emissions from landfill.

Mesospheric Chemistry

The most distinctive aspect of mesospheric chemistry to be studied by the lidar is that of metal atoms and ions of meteoric origin (*see Mesosphere: Atomic Species in the Mesopause Region; Metal Layers; Polar Summer Mesopause. Solar System/Sun, Atmospheres, Evolution of Atmospheres: Meteors*). Meteors ablate in the 90- to 100-km altitude range, releasing their contents as vapors. Metal atoms and ions are characterized by excited electronic states with excitation energies corresponding to visible photon wavelengths. Such is the low density in the upper mesosphere (~ 0.01 Pa) that collisional quenching of these states is generally much slower than their radiative decay. Under these conditions resonance scattering comes into its own as a lidar technique: if the laser beam is tuned to an electronic transition of the desired species, the backscatter cross section is enhanced by an enormous factor – typically of order 10^{16} . This enables concentrations of metal atoms or ions as low as 10^7 m⁻³ to be measured by lidar.

The first and most extensively studied species to be measured in this way was sodium, using the D₂ line at 589 nm. Profiles have also been measured of potassium (770 nm), lithium (671 nm), neutral and ionized calcium (423 and 393 nm, respectively), and iron (372 nm). The altitude distribution of these species corresponds to the meteor ablation region, with a peak concentration at ~ 90 km. In the case of sodium the measured concentrations vary strongly with season

and latitude, being greatest in winter and at high latitudes. A typical sodium atom concentration at 90 km at 69° N is around 5×10^9 m⁻³. Surprisingly, the strong seasonal variation characteristic of sodium is not found for potassium, which changes little over the year. All the metals so far studied exhibit thin layers of greatly enhanced concentration, which appear sporadically and last typically a few hours. Their origin remains something of a mystery, as does their distribution – enhanced sodium layers, for instance, are found mainly at low and high latitudes, whereas iron layers are commonest in midlatitudes. There is also some uncertainty about the relative concentration of different metals. Whereas those of sodium, potassium, and lithium are consistent with their meteoric source, the concentrations of iron and calcium are much lower than expectation. The advent of tunable solid-state lasers promises to simplify considerably the construction of lidars for mesospheric metal measurements, so that rapid advances in this field may be expected in the next few years.

Also measured in the mesosphere by lidar is the hydroxyl radical, OH. In fact this measurement is a serendipitous bonus of stratospheric ozone lidars based on XeCl lasers. The 308-nm laser wavelength falls within an electronic excitation band of OH, which means that resonance scattering is possible, as for metal atoms (although the technique is often called laser-induced fluorescence there is no wavelength shift between transmitted and received radiation). A layer of enhanced backscatter appears at night between 75 and 85 km at 308 nm, corresponding to OH concentrations of $2\text{--}4 \times 10^5$ radicals cm⁻³. The existence around the world of several lidars for stratospheric ozone monitoring means that the temporal and latitudinal variation of mesospheric OH may be followed for little extra cost.

See also: Lidar: Atmospheric Sounding Introduction; Differential Absorption Lidar; Backscatter; Raman; Doppler; Resonance.

Mesosphere: Metal Layers; Atomic Species in the Mesopause Region; Polar Summer Mesopause. **Solar System/Sun, Atmospheres, Evolution of Atmospheres:** Meteors.

Further Reading

- Ansmann, A., Neuber, R., Rairoux, P., Wandinger, U. (Eds.), 1996. *Advances in Atmospheric Remote Sensing with Lidar*. Springer-Verlag, Berlin.
- Measures, R.M., 1984. *Laser Remote Sensing*. Krieger Publishing Company, Malabar, Florida.
- Robinson, R., Gardiner, T., Innocenti, F., Woods, P., Coleman, M., 2011. Infrared differential absorption lidar (DIAL) measurements of hydrocarbon emissions. *Journal of Environmental Monitoring* 13, 2213–2220.
- Sedlacek, A.J., Fischer, K.W. (Eds.), 1999. *Application of Lidar to Current Atmospheric Topics III*. Proceedings of SPIE, vol. 3757. SPIE, Washington, DC.
- Thomas, L., 1995. Lidar methods and applications. In: Clark, R.J.H., Hester, R.E. (Eds.), *Spectroscopy in Environmental Science*. J. Wiley, Chichester, pp. 1–47.
- Wolf, J.-P. (Ed.), 1997. *Lidar Atmospheric Monitoring*. Proceedings of SPIE, vol. 3104. SPIE, Washington, DC.

Relevant Website

<http://www.earlinet.org/>

Observations for Chemistry (Remote Sensing): Microwave

J Waters, California Institute of Technology, Jet Propulsion Laboratory, Pasadena, CA, USA

© 2003 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 4, pp 1516–1528, © 2003, Elsevier Ltd.

Introduction

Microwave remote sensing observations for atmospheric chemistry have been obtained from ground, aircraft, balloon, and satellite platforms. Examples of results are given later in this article, after describing the measurement fundamentals, chemical species having microwave spectra, and instrumentation. The technique has been used, to date, for observations of stratospheric and mesospheric chemistry. It has not yet been applied to tropospheric chemistry, primarily because of difficulties associated with tropospheric H₂O absorption and the larger widths of tropospheric spectral lines. Recent technological advances have removed the second difficulty, and future application to the *upper* troposphere, where there are sufficient spectral windows, is expected. Microwave observations of chemical species can be made in the presence of clouds and volcanic aerosols that limit shorter-wavelength infrared, visible, and ultraviolet techniques.

'Microwave' is used here to denote heterodyne measurements at centimeter, millimeter, and submillimeter wavelengths. The spectral range of interest extends, roughly, from the 1.35 cm wavelength (22 GHz frequency, 1 GHz = 10⁹ Hz) H₂O spectral line through the 0.12 mm wavelength (2.5 THz frequency, 1 THz = 10¹² Hz) OH lines. 'Heterodyne' indicates multiplying a weak input signal by a strong local oscillator signal to translate – without loss of information – the input signal to a portion of the spectrum more convenient for further processing. This process allows measurements of weaker signals, and better spectral resolution, than can generally be obtained with other techniques. Technology for low-noise submillimeter heterodyne measurements has become available only recently, and is advancing rapidly. Far-infrared non-heterodyne techniques made the first measurements of the atmospheric submillimeter spectrum.

Resolved spectral lines providing unique signatures of selected species are measured. Measurements can be either of atmospheric thermal emission or of absorption against a source such as the Sun; most are of thermal emission in order to obtain results at all times of day and night. Temperature is usually obtained from emission by O₂ but can also be obtained from other gases. The rotational states usually sensed are in thermal equilibrium well up through the mesosphere, easing interpretation. The approximately linear temperature dependence of the Planck function at the observed wavelengths and temperatures means that uncertainties in atmospheric temperature generally do not limit accuracy of the measured abundance. The resolved line measurements give robust results: taking the difference between outputs from nearby channels 'on' and 'off' a line can eliminate artifacts while retaining the signal. The 'shape' of output from channels covering the line, which must have a certain form, also provides robustness and information on the altitude profile. Mathematical techniques

for retrieving abundance and temperature from remote measurements are well established.

Figure 1 shows the tropospheric spectrum for the wavelength region considered here. The window at ~200–300 GHz is especially productive for ground-based measurements of stratospheric chemistry. Aircraft, balloon, and satellite platforms usually must be employed for measurements above ~300 GHz owing to tropospheric H₂O absorption.

Measurement Fundamentals

The power in the atmospheric signal measured at frequency ν over incremental range $d\nu$ can be written as eqn [1].

$$dP_\nu = \frac{1}{2} d\nu \int_{\Omega} I_\nu(\theta, \phi) A_e(\theta, \phi) d\Omega \quad [1]$$

$I_\nu(\theta, \phi)$ is the intensity (W Hz⁻¹ m⁻² sr⁻¹) of atmospheric radiation from direction (θ, ϕ) and $A_e(\theta, \phi)$ is the instrument's effective collecting area in that direction. The integration is over all solid angle Ω ; the factor 1/2 appears because only one polarization is received and random polarization is assumed here for the atmospheric radiation. The collecting area of a heterodyne 'radiometer' (instrument for measuring radiative power) has the property [2], where $\lambda = c/\nu$ is the wavelength of the radiation ($c = 2.998 \times 10^8$ m s⁻¹ is the speed of light), and is related to $G(\theta, \phi)$, the instrument's 'antenna gain pattern', by $A_e(\theta, \phi) = \lambda^2 G(\theta, \phi)/4\pi$.

$$\int_{\Omega} A_e(\theta, \phi) d\Omega = \lambda^2 \quad [2]$$

Equation [2] has important practical implications: the instrument's collecting area integrated over solid angle is just the square of the wavelength, independent of its 'physical' size. The power collected from a source that fills the antenna beam, generally true for atmospheric measurements, is not increased by increasing the aperture's physical size. The increase in power that might have been expected from the aperture size increase is, owing to diffraction effects, offset by a decrease in the angular range over which radiation is collected. The instrument field-of-view (FOV) width (in radians) in a particular plane is, approximately, the inverse of the aperture dimension (in wavelengths) in that plane. Because the integrated collecting area is λ^2 , the power per unit frequency interval collected at frequency ν from blackbody radiation is $\frac{1}{2}\lambda^2 B_\nu(T)$, where $B_\nu(T) = (2h\nu^3/c^2)/(e^{h\nu/kT} - 1)$ is the Planck function giving the radiation intensity; $h = 6.626 \times 10^{-34}$ Js is the Planck constant; and $k = 1.38 \times 10^{-23}$ W Hz⁻¹ K⁻¹ is the Boltzmann constant. For temperatures in Earth's atmosphere, $\frac{1}{2}\lambda^2 B_\nu(T)$ is nearly constant with ν up to ~5 THz, beyond which it drops rapidly, as shown in Figure 2. Heterodyne measurements of atmospheric thermal emission are, thus, relatively efficient at

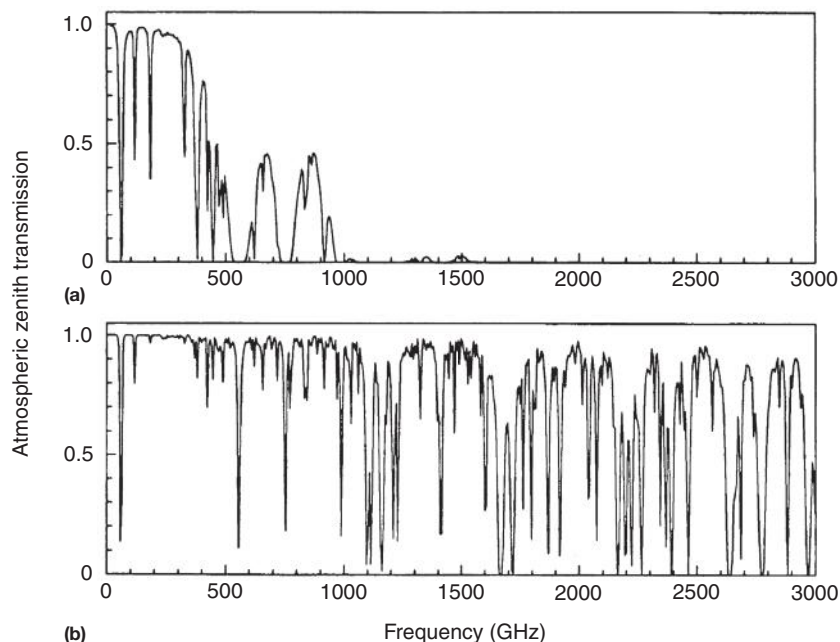


Figure 1 Atmospheric zenith transmission above a 4 km high mountain (a) and above an aircraft at 12 km (b). The absorption features seen here are due to H₂O and, to a lesser extent, O₂. Reproduced with permission from Phillips TG and Keene J (1992) Submillimeter astronomy. Proceedings of the IEEE 80: 1662–1678. © 1992 Institute of Electrical and Electronics Engineers, Inc.

frequencies up to $\sim 5\text{THz}$ ($\lambda > \sim 5 \times 10^{-5}\text{m}$), but not much beyond. Shorter-wavelength (e.g., infrared) heterodyne atmospheric measurements are usually made of solar absorption.

The intensity of radiation from a given direction is given by eqn [3].

$$I_\nu = I_\nu(0)e^{-\tau_\nu(0,L)} + \int_0^{\tau_\nu(0,L)} B_\nu(T)e^{-\tau_\nu(s,L)} d\tau_\nu(s) \quad [3]$$

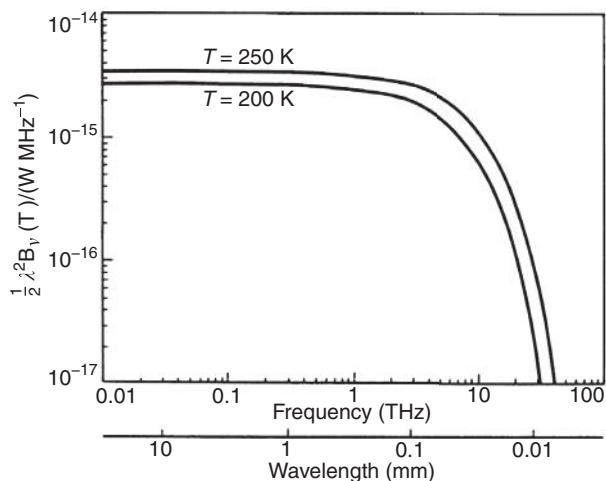


Figure 2 The power per unit frequency interval collected by a heterodyne instrument at frequency ν from a blackbody radiation field at 200 and 250K. The function is plotted in units of W MHz^{-1} , as MHz is the typical unit for widths of microwave spectral lines. Reproduced with permission from Waters JW (1993) Microwave limb sounding. Chapter 8 in Janssen MA (ed.). Atmospheric Remote Sensing by Microwave Radiometry. New York: Wiley. © 1993 John Wiley.

The integral is over the radiation path through the atmosphere with the instrument at L , $I_\nu(0)$ is the radiation at point 0 outside the atmosphere and τ_ν , given by eqn [4], is the 'optical depth' at frequency ν between points s and L on the path, where $\alpha_\nu(s)$ is the 'absorption coefficient' (having dimension 1/length) at s .

$$\tau_\nu(s,L) = \int_s^L \alpha_\nu(s) ds \quad [4]$$

Combining eqns [1], [2] and [3], the power collected from a thermal emission signal (no $I_\nu(0)$ term in eqn [3]) over frequency range $\Delta\nu$ can be written as eqn [5].

$$P_{\text{sig}} = \frac{\lambda^2}{2} \overline{B_\nu(T)} (1 - e^{-\tau_\nu(0,L)}) \Delta\nu \xrightarrow{\tau_\nu(0,L) \ll 1} \frac{\lambda^2}{2} \overline{B_\nu(T)} \tau_\nu(0,L) \Delta\nu \quad [5]$$

In eqn [5] an overbar indicates an appropriate average, and the rightmost expression applies to signals with small optical depth. The microwave signal power is usually expressed as signal 'brightness temperature'. This quantity is proportional to the detected power from the signal, has units of temperature, and for $h\nu \ll kT$ converges to the temperature of blackbody radiation that would give rise to the amount of detected power from the signal (eqn [6] where eqn [5] has been used to obtain the expression for large optical depth after the first arrow).

$$T_{\text{sig}} \equiv \frac{P_{\text{sig}}}{k\Delta\nu} \xrightarrow{\tau_\nu(0,L) \gg 1} \frac{\lambda^2}{2k} \overline{B_\nu(T)} \xrightarrow{h\nu \ll kT} \bar{T} \quad [6]$$

The absorption coefficient for an isolated spectral line is given by eqn [7].

$$\alpha_v^{ul} = n \frac{2\pi^2 g_l \phi_{ul}^2 \mu^2}{3\epsilon_0 h c Q(T)} \times [e^{-E_l/kT} - e^{-E_u/kT}] v b(v, \nu_{ul}) \quad [7]$$

In eqn [7] n is number density of the species with the spectral line, u and l denote the upper and lower quantum states involved in the transition, g_l is the degeneracy and E_l the energy for state l , ϕ_{ul}^2 is the transition matrix element (sometimes g_u is included in ϕ_{ul}^2), μ is the overall dipole (or other) moment coupling to the radiation field, and $\epsilon_0 = 8.854 \times 10^{-12} \text{ F m}^{-1}$ is the permittivity of vacuum. $Q(T) = \sum_{g_l} \exp(-E_l/kT)$ is the partition function, and it is assumed that the quantum states are in thermal equilibrium. The 'line shape function' $b(v, \nu_{ul})$, normalized such that $\int b(v, \nu_{ul}) dv = 1$, gives the probability density that the transition is observed at frequency ν rather than the nominal $\nu_{ul} = (E_u - E_l)/h$.

Atmospheric microwave line shapes are dominated by collision (pressure) broadening at lower altitudes and Doppler (thermal motion) broadening at higher altitudes. The collision linewidth parameter (half-width at half-maximum) can be written $\Delta\nu_c^{ul} = (\Delta\nu_{co}^{ul}) \cdot (\rho/1 \text{ hPa}) \cdot (T/300 \text{ K})^{-x}$ where ρ is atmospheric pressure, and $\Delta\nu_{co}^{ul}$ and x are constants. Typical values are $\Delta\nu_{co}^{ul} = 2.5 \text{ MHz}$, $x = 0.75$. The Lorentz shape, $b(v, \nu_{ul}) = [\Delta\nu_c^{ul}/\pi] / [(v - \nu_{ul})^2 + (\Delta\nu_c^{ul})^2]$, is applicable to the narrow collision-broadened lines measured for atmospheric chemistry. Doppler broadening gives a Gaussian shape $b(v, \nu_{ul}) = (1/\Delta\nu_D^{ul}) (\ln 2/\pi)^{1/2} \exp\{-\ln 2[(v - \nu_{ul})/(\Delta\nu_D^{ul})]^2\}$ with linewidth parameter $\Delta\nu_D^{ul} = (\nu_{ul}/c) \cdot (2 \ln 2 kT/m)^{1/2} = 3.58 \times 10^{-7} \cdot \nu_{ul} \cdot (T/m)^{1/2}$, where m is the molecular mass in g mol^{-1} and T is in K. The Voigt shape (convolution of collision and Doppler shapes) applies when both Doppler and collision broadening are important. Figure 3 shows representative linewidths versus altitude. The collision linewidth for O_2 is noticeably smaller than the others in Figure 3 because collision linewidths generally increase with dipole moment and O_2 has a very small (magnetic) dipole. Doppler broadening is important above $\sim 80 \text{ km}$ for the 22 GHz H_2O line, decreasing to above $\sim 40 \text{ km}$ for the 2.5 THz OH line.

The line shape function at the center of a collision-broadened line has value $1/(\pi \Delta\nu_c^{ul})$. Equation [7] then becomes [8], where $f \equiv n/N$ is the volume mixing ratio of the gas whose spectral line is being measured, N being atmospheric total number density, $\Delta\nu_c^{ul} \propto \rho T^{-x} \propto NT^{1-x}$ has been used, and all transition and frequency-dependent terms, and constants, are placed in $S_{ul}(T)$.

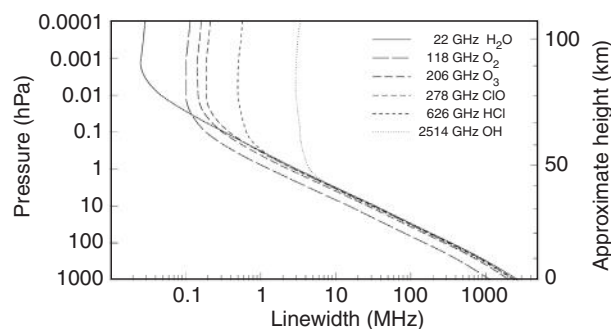


Figure 3 Linewidth (half-width at half-maximum) versus altitude for some representative microwave spectral lines.

$$\alpha_{\nu_{ul}}^{ul} = \frac{f \mu^2}{Q(T)} S_{ul}(T) \quad [8]$$

Equation [8] shows that the intensity at center of a (low optical depth) collision-broadened line is proportional to the volume mixing ratio of the species. Figure 4 shows a spectral line for the same mixing ratio at the top, middle, and bottom of the stratosphere.

Chemical Species and Spectra

Many chemical species have spectral lines in the wavelength region considered here. It is useful to have a 'figure of merit' that, in some sense, indicates the typical strength of a microwave spectral line from a particular species and can provide a first step for exploring measurement feasibility. Extracting factors from eqn [8] that depend only upon the species' overall properties yields $M = f \mu^2 / Q$ as a 'figure of merit'. Abundance of the species is described by f , μ^2 describes its overall interaction strength with radiation, and Q roughly indicates the number of quantum states over which it is spread. Table 1 lists stratospheric species in order of decreasing M for approximate maximum abundances in the stratosphere. Symmetric molecules (e.g., CH_4 and CO_2) with no dipole moment and no microwave spectra are missing, as are complex molecules (e.g., chlorofluorocarbons) having large partition functions that cause their 'figure of merit' to be lower than those included in the table. Examination of a species' spectrum, and – to avoid interference – the spectra of other species having M greater than ~ 10 times below that of the targeted species, is required to determine measurement feasibility and the spectral line(s) best suited. A catalog maintained by the Molecular Spectroscopy Group at the California Institute of Technology Jet Propulsion Laboratory (see World Wide Web site <http://spec.jpl.nasa.gov>) has microwave line parameters for the species in Table 1, and many more. Line frequencies are typically known to $\sim 0.1 \text{ MHz}$ or better; strengths to $\sim 1\%$ or better. Dipole (and other) moments are measured from Stark and Zeeman splitting of

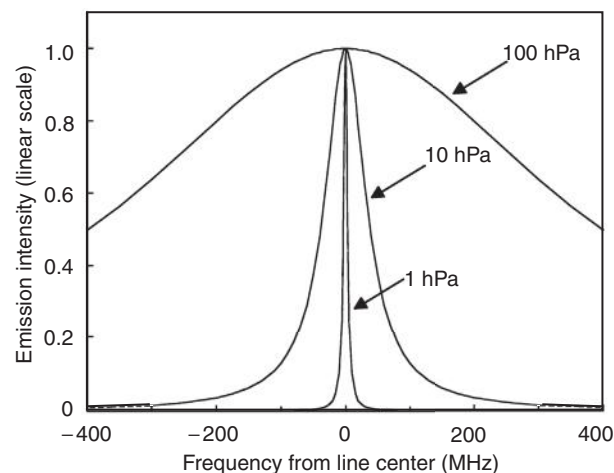


Figure 4 Microwave emission lines for the same mixing ratio of a gas at the bottom (100 hPa, $\sim 15 \text{ km}$), middle (10 hPa, $\sim 30 \text{ km}$), and top (1 hPa, $\sim 50 \text{ km}$) of the stratosphere.

Table 1 Stratospheric species in decreasing order of the microwave 'figure of merit' $M = f\mu^2/Q$, where f is abundance in volume mixing ratio (column 5), μ is the dipole moment (column 3, where 1 Debye = 10^{-18} esu cm = 3.34×10^{-30} C m), and Q is the spin-rotation partition function (column 4, for a typical stratospheric temperature of 225 K). Excited vibrational states, indicated in parentheses (e.g., $2\nu_2$ indicates the second excited state of the second vibrational mode), are treated as separate species. The right column indicates if a microwave spectral line from that species has been detected by instruments from ground 'G', aircraft 'A', balloon 'B', or satellite 'S'. '(S)' indicates a satellite instrument to measure this molecule is planned for the near future. Species having $\log_{10}(M) \geq -14.3$ are included here

Species	$\log_{10}(M)$	Dipole moment (Debye)	Partition function at 225 K	Maximum stratospheric abundance (approximate)	Microwave detected from
O ₂	-6.4	0.0186	1.64×10^2	2.1×10^{-1}	G, A, B, S
H ₂ O	-6.8	1.847	1.16×10^2	5.0×10^{-6}	G, A, B, S
O ₃	-8.9	0.5324	2.23×10^3	1.0×10^{-5}	G, A, B, S
¹⁸ O ¹⁶ O	-9.1	0.0186	3.46×10^2	8.6×10^{-4}	G, (S)
HF	-9.4	1.827	7.96	1.0×10^{-9}	
H ₂ ¹⁸ O	-9.5	1.855	1.17×10^2	1.0×10^{-8}	G, S
H ₂ ¹⁷ O	-10.3	1.855	1.17×10^2	1.9×10^{-9}	(S)
HDO	-10.3	1.732	9.56×10^1	1.5×10^{-9}	G, S
HCl	-10.4	1.109	6.34×10^1	2.3×10^{-9}	A, B, (S)
¹⁷ O ¹⁶ O	-10.6	0.0186	2.02×10^3	1.5×10^{-4}	
OH	-10.6	1.655	6.03×10^1	5.0×10^{-10}	A, B, (S)
O ₂ (ν_1)	-10.7	0.0186	1.64×10^2	1.0×10^{-5}	
N ₂ O	-10.7	0.1608	3.74×10^2	3.0×10^{-7}	G, A, B, S
H ³⁷ Cl	-10.8	1.109	6.34×10^1	7.5×10^{-10}	A, B
O ₃ (ν_2)	-10.9	0.5324	2.23×10^3	1.0×10^{-7}	G, B, (S)
O ₂ ($^1\Delta$)	-11.0	0.0186	1.13×10^2	3.0×10^{-6}	G
CO	-11.1	0.1101	8.17×10^1	5.0×10^{-8}	G, A, S
SO ₂	-11.2 ^a	1.633	3.84×10^3	1.0×10^{-8a}	A, S ^a
HCN	-11.3	2.984	3.18×10^2	2.0×10^{-10}	G, (S)
¹⁶ O	-11.3	0.0186	6.324	1.0×10^{-7}	
H ₂ O(ν_2)	-11.3	1.816	1.16×10^2	1.9×10^{-10}	
HNO ₃	-11.5	1.986	1.82×10^4	1.5×10^{-8}	G, A, B, S
¹⁸ O ¹⁶ O ¹⁶ O	-11.6	0.5324	4.69×10^3	4.0×10^{-8}	G, B, S
O ¹⁸ O ¹⁶ O	-11.6	0.5324	2.29×10^3	2.0×10^{-8}	G
O ₃ (ν_3)	-11.8	0.5324	2.23×10^3	1.3×10^{-8}	G, (S)
O ₃ (ν_1)	-12.0	0.5324	2.23×10^3	9.0×10^{-9}	G
³⁵ ClO	-12.0	1.297	2.31×10^3	1.5×10^{-9}	G, A, B, S
N ₂ O(ν_2)	-12.3	0.1608	3.74×10^2	7.0×10^{-9}	
³⁷ ClO	-12.5	1.297	2.35×10^3	5.0×10^{-10}	
NO	-12.5	0.1587	8.17×10^2	1.0×10^{-8}	G ^b , (S)
H ₂ CO	-12.5	2.331	1.87×10^3	1.0×10^{-10}	G, S
HO ₂	-12.6	1.541	2.84×10^3	3.0×10^{-10}	G, A, B, (S)
OCS	-12.7	0.715	7.72×10^2	3.0×10^{-10}	
O ₃ ($2\nu_2$)	-12.8	0.5324	2.23×10^3	1.3×10^{-9}	
HCN(ν_2)	-12.8	2.942	1.06×10^2	2.0×10^{-12}	
NO ₂	-12.9	0.316	8.76×10^3	1.0×10^{-8}	G
HD ¹⁸ O	-13.0	1.726	9.68×10^1	3.0×10^{-12}	
HNO ₃ (ν_7)	-13.1	1.986	1.82×10^4	3.8×10^{-10}	
¹³ CO	-13.1	0.1105	8.55×10^1	5.0×10^{-10}	
O ¹⁷ O ¹⁶ O	-13.1	0.5337	1.35×10^4	3.8×10^{-9}	(S)
¹⁷ O ¹⁶ O ¹⁶ O	-13.1	0.5337	2.73×10^4	7.6×10^{-9}	(S)
N ¹⁵ NO	-13.1	0.1608	3.74×10^2	1.1×10^{-9}	
¹⁵ NNO	-13.1	0.1608	3.87×10^2	1.1×10^{-9}	
HNO ₃ (ν_9)	-13.2	1.986	4.57×10^4	8.1×10^{-10}	
H ¹³ CN	-13.2	2.984	3.27×10^2	2.2×10^{-12}	
HC ¹⁵ N	-13.2	2.984	1.09×10^2	7.3×10^{-13}	
CH ₃ Cl	-13.2	1.899	1.83×10^4	3.0×10^{-10}	
HNO ₃ (ν_6)	-13.3	1.986	1.82×10^4	2.4×10^{-10}	
¹⁸ OH	-13.3	1.667	6.06×10^1	1.0×10^{-12}	
H ₂ O ₂	-13.4	1.572	5.76×10^3	1.0×10^{-10}	S
N ₂ ¹⁸ O	-13.4	0.1608	3.96×10^2	6.0×10^{-10}	
DF	-13.5	1.819	1.48×10^1	1.5×10^{-13}	
CH ₃ CN	-13.6	3.922	2.36×10^4	5.0×10^{-11}	S
O ¹⁸ O ¹⁶ O(ν_2)	-13.5	0.5324	2.45×10^3	2.2×10^{-10}	

(Continued)

Table 1 Stratospheric species in decreasing order of the microwave 'figure of merit' $M = f\mu^2/Q$, where f is abundance in volume mixing ratio (column 5), μ is the dipole moment (column 3, where 1 Debye = 10^{-18} esu cm = 3.34×10^{-30} C m), and Q is the spin-rotation partition function (column 4, for a typical stratospheric temperature of 225 K). Excited vibrational states, indicated in parentheses (e.g., $2\nu_2$ indicates the second excited state of the second vibrational mode), are treated as separate species. The right column indicates if a microwave spectral line from that species has been detected by instruments from ground 'G', aircraft 'A', balloon 'B', or satellite 'S'. '(S)' indicates a satellite instrument to measure this molecule is planned for the near future. Species having $\log_{10}(M) \geq -14.3$ are included here—cont'd

Species	$\log_{10}(M)$	Dipole moment (Debye)	Partition function at 225 K	Maximum stratospheric abundance (approximate)	Microwave detected from
$^{18}\text{OOO}(\nu_2)$	-13.6	0.5324	4.99×10^3	4.4×10^{-10}	
$\text{HNO}_3(\nu_8)$	-13.6	1.986	1.82×10^4	1.2×10^{-10}	
$\text{O}_3(\nu_2 + \nu_3)$	-13.7	0.5324	2.23×10^3	1.6×10^{-10}	
$\text{CH}_3^{37}\text{Cl}$	-13.7	1.895	1.86×10^4	1.0×10^{-10}	
HD^{17}O	-13.8	1.73	9.6×10^1	5.7×10^{-13}	
HO^{35}Cl	-13.8	1.471	6.18×10^3	5.0×10^{-11}	(S)
H^{81}Br	-13.8	0.828	7.84×10^1	2.0×10^{-12}	
H^{79}Br	-13.8	0.828	7.84×10^1	2.0×10^{-12}	
C^{18}O	-13.8	0.1108	8.58×10^1	1.0×10^{-10}	
$\text{O}_3(\nu_1 + \nu_2)$	-13.9	0.532	2.23×10^3	1.0×10^{-10}	
$\text{N}_2\text{O}(2\nu_2)$	-13.9	0.1608	3.74×10^2	1.7×10^{-10}	
^{18}O	-14.0	0.0186	6.324	2.0×10^{-10}	
H^{15}NO_3	-14.0	1.9	1.8×10^4	5.5×10^{-11}	
HO_2NO_2	-14.0	1.288	8.66×10^4	5.0×10^{-10}	
O^{35}ClO	-14.0	1.792	3.49×10^4	1.0×10^{-10}	
OC^{34}S	-14.1	0.715	7.78×10^2	1.2×10^{-11}	
N_2^{17}O	-14.1	0.16	3.8×10^2	1.1×10^{-10}	
^{79}BrO	-14.1	1.780	2.93×10^3	7.0×10^{-12}	
^{81}BrO	-14.1	1.780	2.94×10^3	7.0×10^{-12}	(S)
HO^{37}Cl	-14.2	1.471	6.29×10^3	2.0×10^{-11}	
$\text{HDO}(\nu_2)$	-14.2	1.7	9.5×10^1	2.0×10^{-13}	
$\text{N}_2\text{O}(\nu_1)$	-14.2	0.16	3.8×10^2	8.0×10^{-11}	
H_2SO_4	-14.4	2.725	1.01×10^5	5.0×10^{-11}	
H_2^{13}CO	-14.5	2.331	1.93×10^3	1.1×10^{-12}	
O^{37}ClO	-14.5	1.792	3.54×10^4	3.3×10^{-11}	
DCI	-14.5	1.103	1.18×10^2	3.0×10^{-13}	
C^{17}O	-14.6	0.1103	8.38×10^1	1.9×10^{-11}	
COF_2	-14.6	0.951	3.99×10^4	1.0×10^{-10}	
O^{13}CS	-14.7	0.715	7.74×10^2	3.3×10^{-12}	
^{17}O	-14.7	0.0186	6.324	3.8×10^{-11}	
$^{35}\text{ClOO}^{35}\text{Cl}$	-14.7	0.72	2.77×10^5	1.0×10^{-9}	
OC^{18}O	-14.8	0.0007	4.2×10^2	1.4×10^{-6}	
$^{35}\text{ClONO}_2$	-14.9	0.72	3.72×10^5	1.0×10^{-9}	

^aAbundance and measurements are for enhanced SO_2 injected into the stratosphere by a volcanic eruption. Background SO_2 abundance in the stratosphere is $\sim 10^{-11}$, corresponding to.

^bThe 250 GHz spectral line due to thermospheric NO has been measured.

Updated from Waters JW (1993) Microwave limb sounding. Ch. 8 in Janssen MA (ed.), Atmospheric Remote Sensing by Microwave Radiometry. New York: John Wiley; Spectroscopic data from 9 Oct 2001 revision of the JPL Submillimeter, Millimeter, and Microwave Spectral Line Catalog, available on the world wide web at <http://spec.jpl.nasa.gov>.

spectral lines, allowing line strengths to be determined without requiring measurement of the gas abundance in a laboratory cell. Linewidth parameters can be measured with $\sim 3\%$ accuracy.

Figure 5 shows, on a very compressed scale, representative spectra including those of radicals involved in the major chemical cycles for stratospheric ozone destruction. Figure 6 shows an expanded region near 625 and 650 GHz where the strongest rotational line of ClO and the first rotational line of HCl occur.

Instrumentation

Figure 7 gives a typical block diagram of a microwave instrument for atmospheric chemistry observations. Atmospheric

signals are collected by an antenna, possibly amplified or filtered, and passed to a mixer. The mixer combines the atmospheric signal with a monochromatic local oscillator (LO) signal in a nonlinear process that reproduces the signal spectra at frequencies that are sums and differences of the atmospheric and LO signal frequencies, and possibly at their harmonics. State-of-the-art mixers are based on planar Schottky diodes (either cooled or room temperature), superconductor-insulator-superconductor (SIS) tunnel junctions, or superconducting hot electron bolometer (HEB) devices. The LO frequency is chosen so that the mixing product of interest (usually at the difference between the LO and signal frequencies) appears at intermediate frequencies (IFs) convenient for further processing. The IF signal, after amplification, is

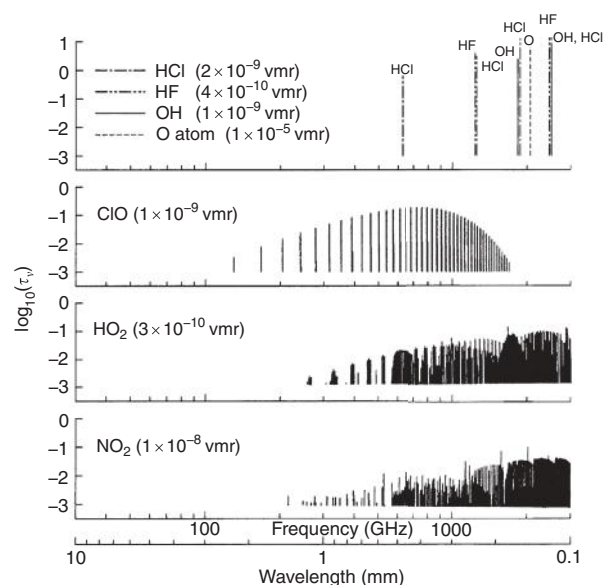


Figure 5 Spectra of some stratospheric molecules and atomic oxygen. Vertical axis is the logarithm of the optical depth for an observation path through the atmospheric limb with the indicated volume mixing ratios (vmr). Reproduced with permission from Waters JW (1993) Microwave limb sounding. Ch. 8 in Janssen MA (ed.). Atmospheric Remote Sensing by Microwave Radiometry. New York: Wiley. © 1993 John Wiley.

passed to a spectrometer that separates it into individual spectral channels with desired resolution. Each channel's signal is then 'detected' (converted to a voltage proportional to its power and digitized), and passed to the data-handling system. Instruments often contain multiple mixers, LOs, IF amplifiers, and spectrometers to allow simultaneous observations of

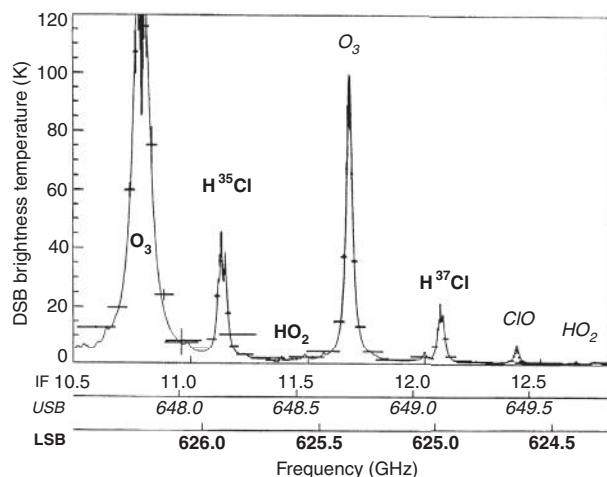


Figure 6 Measured (crosses, bars) and calculated (thin line) stratospheric emission spectrum in bands measured simultaneously near 626 and 649 GHz. 'USB' refers to the frequency covered by the upper sideband of the radiometer, 'LSB' to the lower sideband, and 'IF' to the intermediate frequency (see discussion in 'Instrumentation' section of text). LSB spectral lines are in bold face (e.g., **HO₂**); USB lines are in italic face (e.g., *HO₂*). Adapted from Stachnik RA *et al.* (1992) Submillimeterwave heterodyne measurements of stratospheric ClO, HCl, O₃ and HO₂: first results. Geophysical Research Letters 19: 1931–1934.

several spectral lines. LO frequencies can be changed operationally to change measurements. Radiometers can be 'chopped', to reduce amplifier gain variation effects, by rapidly moving a 'reference' target in and out of the signal path or by frequency-switching the LO. An instrument sufficiently stable between calibrations does not require chopping and has twice the sensitivity of an otherwise identical symmetrically chopped instrument.

Calibration is performed by inserting targets (typically blackbody targets, with 'cold space' generally used for one when possible) in the signal path near the instrument input, ideally before the antenna. 'Double sideband' (DSB) radiometers receive signals in two mixer 'sidebands', at IFs above and below the LO, whereas 'single sideband' (SSB) radiometers receive signals in only one sideband. The DSB thermal calibration signal, coming through both sidebands, is twice that for SSB. A spectral line occurring in one sideband appears half as strong when (thermally) calibrated DSB as when calibrated SSB. Accurate calibration can require measuring and accounting for the responses in the two sidebands.

The sensitivity of a microwave radiometer is usually specified by its 'receiver noise temperature', a quantity – analogous to 'brightness temperature' – that is proportional to the instrument noise power referenced to the instrument input. The root mean square measurement noise, expressed in temperature, for integration time Δt and spectral resolution $\Delta \nu$ is given by eqn [9].

$$\Delta T_{\text{rms}} \approx a \frac{T_{\text{rec}} + T_{\text{sig}}}{\sqrt{\Delta t \Delta \nu}} \quad [9]$$

T_{rec} is the receiver noise temperature, and $a \approx 1$ for a non-chopped system or $a \approx 2$ for a chopped system. T_{sig} , usually much smaller than T_{rec} (except for solar absorption measurements), is present because the signal itself is noisy. The value of T_{rec} is primarily set by noise in the first amplifier and the amount of signal loss preceding it. Advances in technology are rapidly reducing noise and extending spectral bandwidth. Current values of T_{rec} (SSB values, or $2 \times$ DSB values) for the ~ 200 – 300 GHz range with ~ 0.5 GHz bandwidth are ~ 2000 K for room-temperature radiometers and ~ 400 K for SIS radiometers. Bandwidth can be increased at the cost of somewhat increased noise: room temperature radiometers at ~ 200 GHz have been developed with $T_{\text{rec}} \approx 4000$ K and ~ 15 GHz IF bandwidth. The measurement noise with $\Delta \nu = 16$ MHz resolution and $\Delta t = 1$ s integration, for example, is $\Delta T_{\text{rms}} \approx 1$ K for $T_{\text{rec}} \approx 4000$ K and $\Delta T_{\text{rms}} \approx 0.1$ K for $T_{\text{rec}} \approx 400$ K. Increasing the integration time to 1 hour gives $\Delta T_{\text{rms}} \approx 0.02$ K for $T_{\text{rec}} \approx 4000$ K and $\Delta T_{\text{rms}} \approx 0.002$ K for $T_{\text{rec}} \approx 400$ K. Microwave radiometer noise generally increases with frequency, but this can be offset by increases in line strength. A 22 GHz radiometer for ground-based stratospheric H₂O measurement ($T_{\text{sig}} \sim 0.2$ K) has $T_{\text{rec}} \approx 100$ K with ~ 0.5 GHz bandwidth and 20 K cooled first-stage transistor amplifier. A 2.5 THz room temperature radiometer with ~ 15 GHz IF bandwidth developed for satellite OH measurements ($T_{\text{sig}} \sim 100$ K) has $T_{\text{rec}} \approx 20\,000$ K. HEB radiometers at 2.5 THz have been constructed recently with $T_{\text{rec}} \approx 3000$ K and ~ 10 GHz bandwidth; improvements are expected as this technology matures.

Several types of spectrometers are currently used. Filter banks have a set of simultaneously observed spectral filters, usually

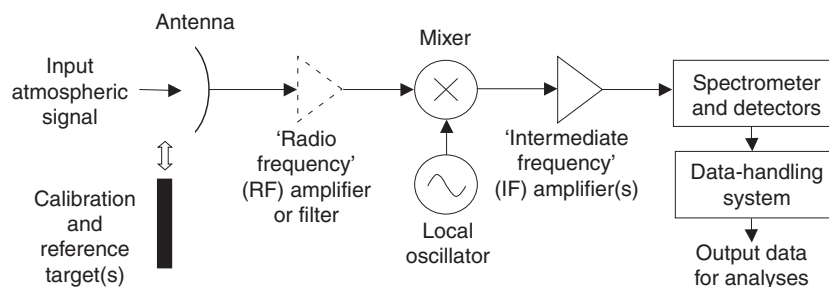


Figure 7 Typical block diagram of an instrument for microwave observations of stratospheric chemistry. The 'radiofrequency' amplifier is currently available only at lower frequencies and does not appear in many systems; a filter is sometimes placed at this position to eliminate unwanted signals in one of the mixer's sidebands. The portion of the instrument between the antenna and spectrometer is called the 'receiver' or 'radiometer'.

made of discrete or distributed capacitive and inductive elements, whose frequencies and widths are set as needed for a particular measurement. Digital autocorrelators measure a signal's autocorrelation simultaneously at many time lags; Fourier transforming the measured autocorrelation function gives the spectrum. Acousto-optic spectrometers use the IF signal to modulate a Bragg cell which, according to the signal's spectral content, diffracts a laser beam to a detector array. Chirp transform spectrometers multiply the IF signal by a frequency-modulated (chirp) waveform and convolve the resulting product with a filter that is appropriately matched to the chirp, and the spectrum appears as a function of time at the output.

Ground-Based Observations

Ground-based observations can provide continuous monitoring at selected sites. Instrumentation can be upgraded and repaired if needed, and can rapidly respond to changing priorities and atmospheric conditions. Vertical resolution is obtained from the spectral line shape and is typically around one atmospheric pressure scale height ($\sim 6\text{--}8$ km) but can be somewhat smaller with good signal to noise. Initial ground-based microwave measurements in the 1970s included stratospheric and mesospheric O_3 from lines near 100 GHz, and high-rotational lines of O_2 (on the edge of its 60 GHz spin-rotation band) that are very sensitive to stratospheric temperatures. Mesospheric CO was first measured by 115 GHz ground-based observations, with many later measurements at 230 GHz. Early measurements of the 22 GHz H_2O line showed a dry stratosphere, in contrast to many previous balloon *in situ* measurements (now thought to have been contaminated) that indicated a very wet stratosphere.

Ground-based microwave measurements have provided important results for understanding stratospheric chlorine chemistry. In 1981 they gave the first definitive remote measurements of ClO, the key chlorine radical involved in ozone depletion. Early results also included the first measurement of ClO diurnal variation, testing crucial aspects of upper stratospheric chlorine chemistry. In 1986 the technique gave the first evidence of greatly enhanced ClO in the Antarctic lower stratosphere, firmly connecting chlorine chemistry to the ozone hole. **Figure 8** shows an example of ClO evolution observed over Antarctica. Measurements of ClO diurnal variation tested chemical models for formation and photolysis of the ClO dimer in the Antarctic lower stratosphere. Enhanced ClO in the

Arctic winter stratosphere also has been measured. HNO_3 , important in processes involving polar stratospheric clouds (PSCs) and quenching of reactive chlorine, has been measured, as has N_2O , which gives information on dynamics.

Additional results include measurement of stratospheric HCN, and studies of mesospheric HO_x chemistry from

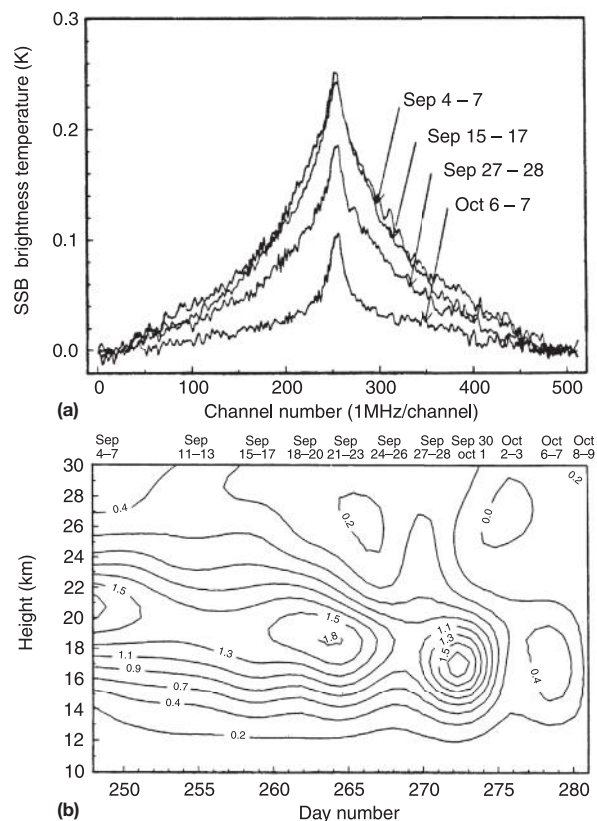


Figure 8 Ground-based 278 GHz measurement of stratospheric ClO over Antarctica. ClO is the key chlorine radical involved in ozone destruction. (a) Day–night differences of the spectral line measured during several days in 1992. (b) A height–time cross-section of the retrieved ClO mixing ratio profile, where contours are in parts per billion by volume. Reproduced with permission from deZafra RL, Reeves JM, and Shindell DT (1995) Chlorine monoxide in the Antarctic spring vortex 1. Evolution of midday vertical profiles over McMurdo Station, 1993. *Journal of Geophysical Research* 100: 13999–14007. © 1995 American Geophysical Union.

measurements of HO₂, O₃, and H₂O. The O₂¹Δ_g excited electronic state in the mesosphere has been measured, including diurnal variation, and implications were obtained for its radiative lifetime and chemistry. Zeeman splitting of mesospheric O₂ lines has been measured. The 250 GHz line from thermospheric NO has been measured. Ground-based microwave instruments are currently used by several research groups, and are deployed in the international Network for the Detection of Stratospheric Change to measure stratospheric ClO, O₃, and H₂O.

Aircraft-Based Observations

Aircraft-based observations can provide measurements with good horizontal resolution along a measurement track over an extended spatial range. Instruments can observe in high-frequency spectral windows where tropospheric H₂O absorption (Figure 1) prevents ground-based measurements and where more species have spectral lines. Vertical resolution is obtained from spectral line shape for measurements above the aircraft altitude, and can be obtained from limb sounding techniques at heights below the aircraft. Initial aircraft measurements in the 1970s included stratospheric H₂O and O₃ from lines near 183 GHz, and an upper limit on stratospheric ClO abundance. Recent measurements include stratospheric HCl, ClO, O₃, HNO₃, N₂O, H₂O, HO₂, BrO, and volcanic SO₂ from a 600 GHz SIS radiometer. Figure 9 shows HCl and ClO results from

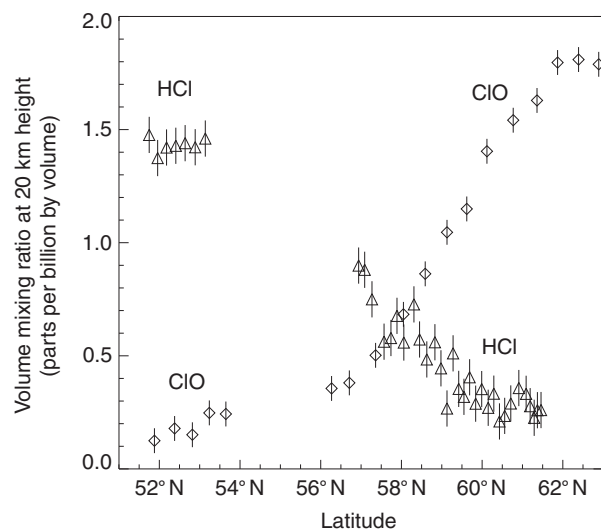


Figure 9 Aircraft measurements of stratospheric HCl (triangles, from the 626 GHz line) and ClO (diamonds, from the 649 GHz line) from a January 2000 flight through the edge of the Arctic vortex. Measurement time for each point is ~1 min for HCl and ~3 min for ClO. These results show the transition of stratospheric chlorine from the relatively inert HCl at lower latitudes to the highly reactive ClO at higher latitudes inside the vortex. Gaps in the measurements around 54–56° N are where the instrument was tuned to measure HNO₃ and N₂O. The instrument and campaign in which these measurements were made are described by Bremer H, *et al.* (2002) Ozone depletion observed by the Airborne Submillimeter Radiometer (ASUR) during the Arctic Winter 1999/2000. *Journal of Geophysical Research* 107, No. D20, 8277, 10.1029/2001JD 000546.

a flight through the edge of the Arctic vortex. OH and H₂O have been measured from aircraft with a room-temperature 2.5 THz radiometer.

Balloon-Based Observations

Balloon-based microwave observations can provide measurements throughout the stratosphere with 2–3 km vertical resolution. The instrument FOV is vertically scanned through the atmospheric limb to observe a long path length and to obtain the vertical resolution. Balloon instruments provide measurements to higher altitudes with better resolution than can be obtained from aircraft or ground, and provide valuable development and tests of techniques to be deployed on satellites. Initial measurements in the 1980s were of ClO and O₃ from lines near 205 GHz. A 600 GHz room-temperature instrument for measuring HCl, ClO, HNO₃, N₂O, O₃, and HO₂ became operational in the early 1990s; some results are shown in Figure 10. Simultaneous measurement of HCl and ClO gives a stringent monitor of stratospheric chlorine chemistry, and first results from the 600 GHz balloon instrument showed that the ClO/HCl ratio in the mid-latitude upper stratosphere could not be explained by chemical models current at the time. Reaction of OH and ClO producing HCl, hypothesized as cause of the discrepancy, has since been measured in the laboratory with a rate that adequately explains the observations. Arctic winter flights have provided information on chlorine partitioning for perturbed chemistry in the vortex. Stratospheric OH has been measured at 2.5 THz from a balloon instrument that is a precursor for satellite observation.

Satellite-Based Observations

Satellite-based observations can provide global coverage on a daily basis. However, instruments cost more and require longer development time than for other platforms. Limb sounding is used for chemistry observations because of its vertical resolution and long path length for observations of small concentrations. Low-orbit (~700 km altitude) satellites have an observation path tangent point ~3000 km from the instrument. Vertical resolution of ~3 km, for example, then requires an antenna having vertical dimension of ~1000 wavelengths. Vertical pointing information is obtained from optically thin O₂ emission and spectral line shapes.

The need for global measurements of ClO motivated initial development of a satellite microwave instrument for stratospheric chemistry. The Microwave Limb Sounder (MLS), operating unchopped in bands around 63, 183, and 205 GHz, was developed to measure ClO, O₃, and H₂O from NASA's Upper Atmosphere Research Satellite (UARS) launched in 1991. A somewhat similar instrument, the Millimeter-wave Atmospheric Sounder (MAS) was flown on three Space Shuttle missions between 1992 and 1994. MLS results showed enhanced ClO filling the lower stratosphere polar winter vortices of both the Antarctic and Arctic. This finding was especially important for the Arctic, demonstrating the effectiveness of localized polar stratospheric clouds in activating chlorine throughout the vortex. Observed ozone loss during

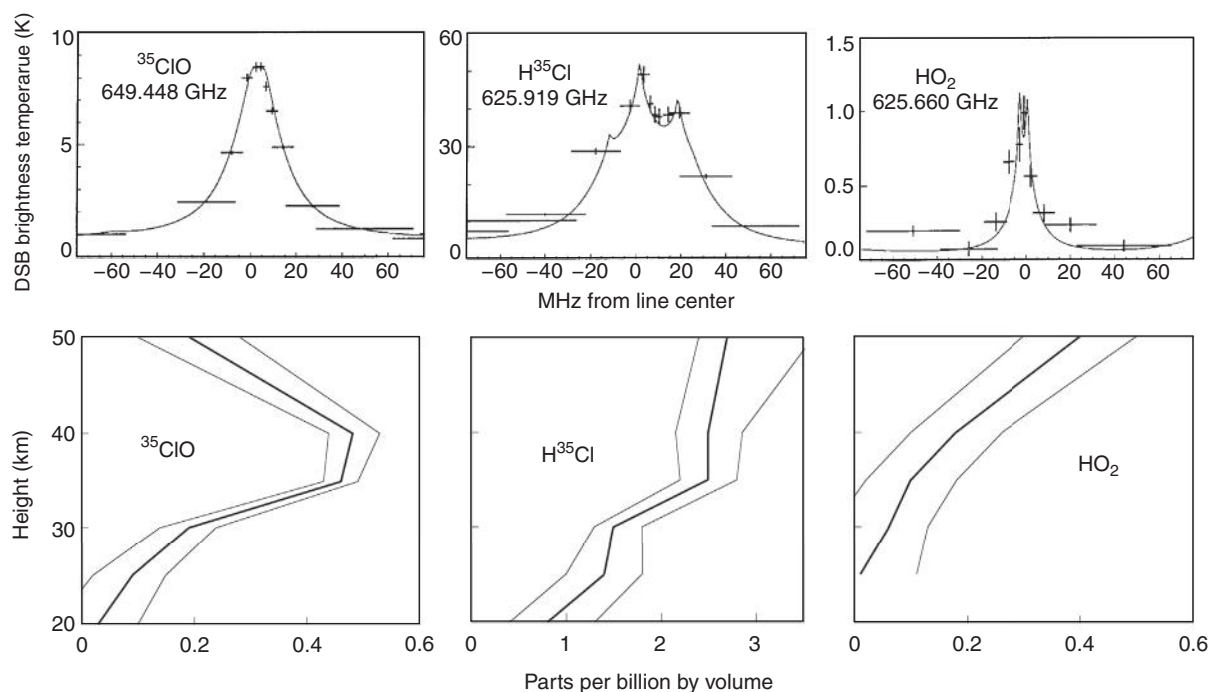


Figure 10 Balloon microwave measurements of ClO, HCl, and HO_2 . Top panels show the spectral lines for a limb observation path through the middle stratosphere: measurements are the horizontal bars whose widths give the spectral resolution of individual filters and smooth lines are calculated. Fine structure features are seen in the HCl line. The bottom panels show retrieved mixing ratio profiles (thick) and uncertainty limits (thin). Adapted from Stachnik RA, *et al.* (1992) Submillimeter wave heterodyne measurements of stratospheric ClO, HCl, O_3 and HO_2 : first results. *Geophysical Research Letters* 19: 1931–1934.

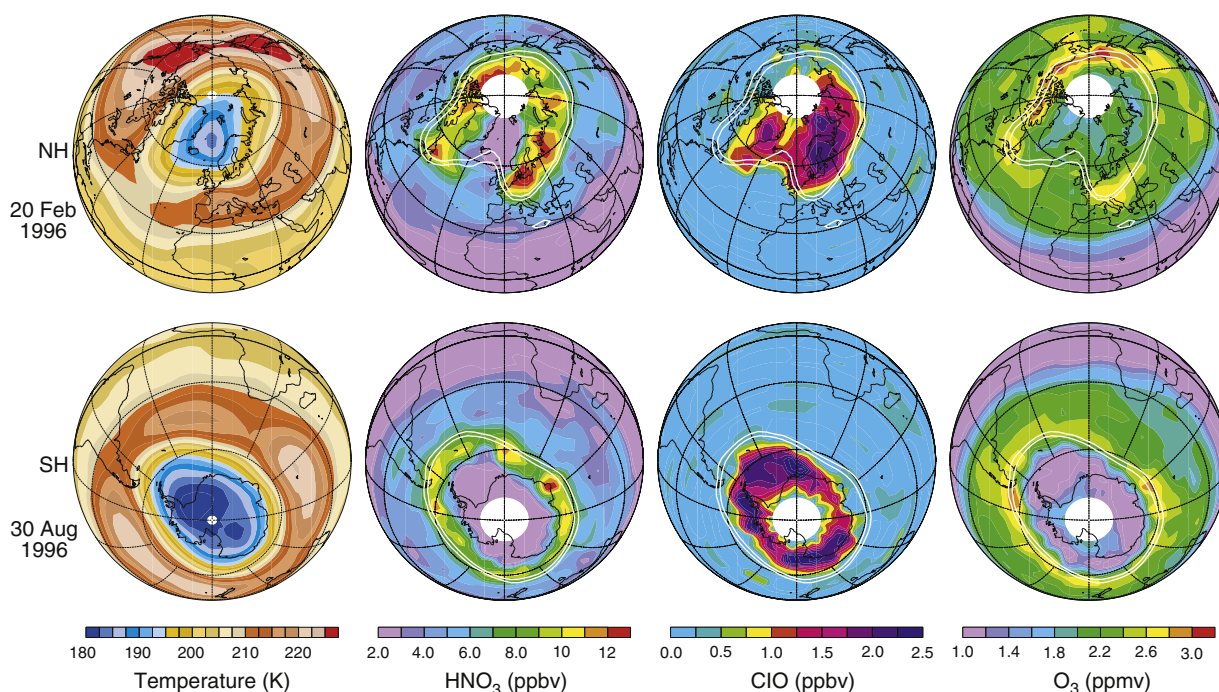


Figure 11 Earth's lower stratosphere in the Northern Hemisphere on 20 February 1996 (top) and in the Southern Hemisphere on 30 August 1996 (bottom). White contours show the dynamical edge of the polar vortices. HNO_3 , ClO and O_3 are from the Microwave Limb Sounder on the Upper Atmosphere Research Satellite (no measurements are made in the white areas near the poles due to orbit limitations). Temperature data are from operational analyses of the US National Center for Environmental Prediction. Temperatures in the blue and violet color ranges allow formation of polar stratospheric clouds from HNO_3 and H_2O ; heterogeneous chemistry on these clouds leads to enhanced ClO that causes chemical depletion of O_3 . HNO_3 also provides a source of NO_x , which quenches ClO and reduces the amount of ozone destruction. Both HNO_3 and O_3 increase in the lower stratospheric vortices during early winter due to downward transport of air rich in these species. The amount of ozone destruction each winter in the polar vortices depends on the duration of enhanced ClO, which is longer for the Antarctic than the Arctic. This difference is traceable to the Antarctic lower stratosphere being colder, and remaining cold for longer, than the Arctic. Reproduced with permission from Waters JW, *et al.* (1999) The UARS and EOS Microwave Limb Sounder experiments. *Journal of the Atmospheric Sciences* 56: 194–218. © 1999 American Meteorological Society.

periods of enhanced ClO is generally consistent with that expected from the ClO amount. MLS found that enhanced ClO appears in the edge of the Antarctic vortex by late May or early June each year and that by late July or early August the depletion of ozone by chlorine starts to dominate increases in ozone by transport. HNO_3 has been measured, providing insights into PSC microphysical processes and quantifying differences in denitrification (removal of nitrogen) between the Arctic and Antarctic. Figure 11 shows examples of daily maps of temperature, HNO_3 , ClO, and O_3 . Latitudinal variation in upper stratospheric zonal mean ClO from MLS appears

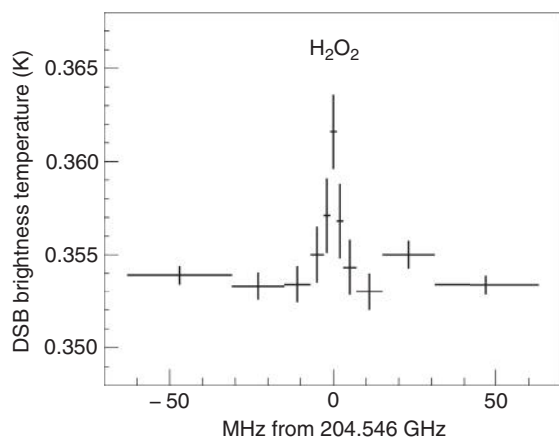


Figure 12 The weak 204 GHz H_2O_2 line from satellite measurements made over a period of 38 days, selected for tangent point pressure between 1 and 0.3 hPa (~ 2 days averaging time for the results shown here). Horizontal bars give the spectral resolution of individual filters and vertical bars give the $\pm 1\Delta T_{\text{rms}}$ measurement uncertainty calculated from eqn [9] for $T_{\text{rec}} = 1000$ K DSB of the unchopped satellite instrument. The line strength corresponds to $\sim 10^{-10}$ H_2O_2 volume mixing ratio in the upper stratosphere and lower mesosphere. The 0.353 K background is emission from the lower atmosphere received through the antenna sidelobes.

qualitatively consistent with that expected from variations in CH_4 . Global ClO variations over a 6-year period have been related to CH_4 variations, increases in total chlorine, and changes in lower stratospheric chemistry associated with the decrease of aerosol from the Mt. Pinatubo volcano. MLS and MAS ClO measurements agree to within ~ 0.1 ppbv where comparisons have been possible.

MLS measurements of H_2O led to the discovery of the atmospheric 'tropical tape recorder' whereby H_2O entering the tropical stratosphere is imprinted with a signature (the corresponding H_2O saturation mixing ratio) of the seasonally varying tropopause temperature. They showed that mid-latitude interhemispheric H_2O differences are not strongly related to the Antarctic winter vortex dehydration. Information has been obtained on various types of atmospheric oscillations and waves that affect the distribution of chemical species. Upper tropospheric H_2O , key for understanding aspects of climate variability, has been measured. The vertical profile of stratospheric SO_2 injected into the stratosphere by the Pinatubo volcano was measured for up to 6 months after the eruption, and its decay rate was shown to be consistent with that expected from reaction with OH. Stratospheric CH_3CN has been measured. Weak signals can require averaging of data taken over long periods, which is feasible from satellites as well as from the ground. Figure 12 shows a weak H_2O_2 line measured by averaging MLS data taken over 38 days.

ODIN, a small Swedish satellite launched in 2001 for time-sharing atmospheric and astronomical observations, has microwave radiometers in bands centered near 118, 495, 550, 557, and 570 GHz to measure stratospheric H_2O (and its minor isotopic variants H_2^{17}O , H_2^{18}O , and HDO), O_3 , ClO, N_2O , HNO_3 , CO, H_2CO , NO, HO_2 , H_2O_2 , and temperature. A next-generation MLS for NASA's Earth Observing System, has radiometers in five broad bands centered near 118, 190, 240, 640 GHz, and 2.5 THz to make measurements throughout the stratosphere and upper troposphere as shown in Figure 13. Japan is developing an SIS instrument for the International

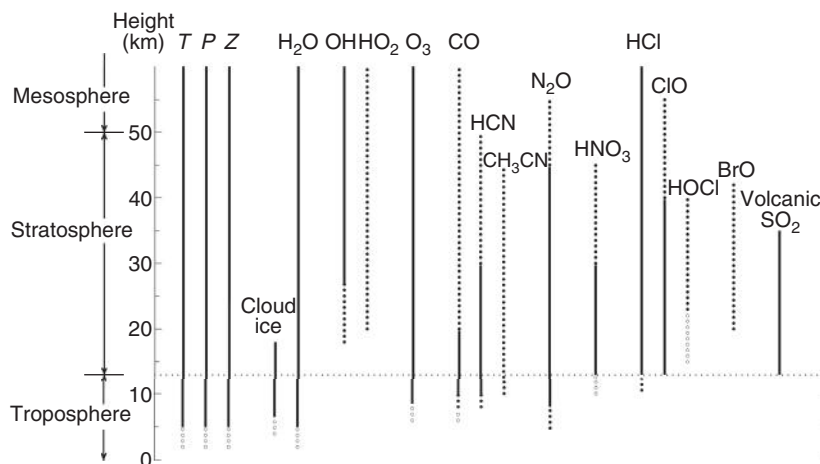


Figure 13 Measurements planned for the Earth Observing System Microwave Limb Sounder scheduled to begin operations in 2003 on NASA's 'Aura' satellite. P is pressure and Z is geopotential height. P, Z, H_2O , O_3 , HCl, OH, and CO measurements extend higher than the 60 km shown here. Closed circles indicate where averages will likely be required, and open circles at lowest altitudes are goals for more difficult measurements. Updated from Waters JW, *et al.* (1999) The UARS and EOS Microwave Limb Sounder experiments. *Journal of the Atmospheric Sciences* 56: 194–218.

Space Station to measure stratospheric O₃ (and its minor isotopes ¹⁸OOO, ¹⁷OOO, O¹⁷OO and excited vibrational states O₃(ν_2), O₂₃(ν_3)) ClO, BrO, H³⁵Cl, H³⁷Cl, HOCl, HO₂, H₂O₂, HNO₃, volcanic SO₂, and temperature in spectral bands near 626 and 650 GHz. Several groups are studying concepts for later experiments.

See also: Climate and Climate Change: Climate Variability: Seasonal and Interannual Variability. **Ozone Depletion and Related Topics:** Ozone Depletion Potentials. **Stratosphere/Troposphere Exchange and Structure:** Global Aspects. **Stratospheric Chemistry Topics:** Halogens; Hydrogen Budget.

Further Reading

- deZafra, R.L., 1995. The ground-based measurement of stratospheric trace gases using quantitative millimeter wave emission spectroscopy. In: Fiocco, G., Visconti, G. (Eds.), *Proceedings of the International School of Physics "Enrico Fermi"*. Course CXXIV. IOS Press, Amsterdam, pp. 23–54.
- Froidevaux, L., Waters, J.W., Read, W.G., et al., 2000. Variations in the free chlorine content of the stratosphere (1991–1997): anthropogenic, volcanic and methane influences. *Journal of Geophysical Research* 105, 4471–4481.
- Janssen, M.A. (Ed.), 1993. *Atmospheric Remote Sensing by Microwave Radiometry*. Wiley, New York (Chapter 8 has derivations of eqn [3] and [7], plots of spectra for most stratospheric species, and a more extensive version of Table 1.)
- Jarnot, R.F., Cofield, R.E., Waters, J.W., Peckham, G.E., Flower, D.A., 1996. Calibration of the microwave limb sounder on the upper atmosphere research satellite. *Journal of Geophysical Research* 101, 9957–9982 (Contains a derivation of eqn [2].)
- Mackenzie, I., Harwood, R.S., Froidevaux, L., Read, W.G., Waters, J.W., 1996. Chemical loss of polar vortex ozone inferred from UARSMLS measurements of ClO during the Arctic and Antarctic late winters of 1993. *Journal of Geophysical Research* 101, 14505–14518.
- Manney, G.L., Froidevaux, L., Waters, J.W., et al., 1994. Chemical depletion of ozone in the Arctic lower stratosphere during winter 1992–93. *Nature* 370, 429–434.
- Mote, P.W., Rosenlof, K.H., McIntyre, M.E., et al., 1996. An atmospheric tape recorder: the imprint of tropical tropopause temperatures on stratospheric water vapor. *Journal of Geophysical Research* 101, 3989–4006.
- Parrish, A., deZafra, R.L., Solomon, P.M., Barrett, J.W., Carlson, E.R., 1981. Chlorine oxide in the stratospheric ozone layer: ground based detection and measurement. *Science* 211, 1158–1161.
- Pickett, H.M., Poynter, R.L., Cohen, E.A., et al., 1998. Submillimeter, millimeter, and microwave spectral line catalog. *Journal of Quantitative Spectroscopy and Radiative Transfer* 60, 883–890.
- Read, W.G., Waters, J.W., Wu, D.L., et al., 2001. UARS microwave limb sounder upper tropospheric humidity measurement: method and validation. *Journal of Geophysical Research* 106, 32207–32258.
- Rodgers, C.D., 2000. *Inverse Methods for Atmospheric Sounding*. World Scientific, Singapore.
- Sandor, B.J., Clancy, R.T., 1998. Mesospheric HOx chemistry from diurnal microwave observations of H₂O, O₃, and H₂O. *Journal of Geophysical Research* 103, 13337–13351.
- Santee, M.L., Read, W.G., Waters, J.W., et al., 1995. Interhemispheric differences in polar stratospheric HNO₃, H₂O, ClO and O₃. *Science* 267, 849–852.
- Solomon, P.M., Connor, B., deZafra, R.L., Parrish, J., Barrett, J., Jaramillo M, A., 1987. High concentrations of chlorine monoxide at low altitudes in the Antarctic spring stratosphere: secular variation. *Nature* 328, 411–413.
- Waters, J.W., Froidevaux, L., Read, W.G., et al., 1993. Stratospheric ClO and ozone from the microwave limb sounder on the upper atmosphere research satellite. *Nature* 362, 597–602.
- Wu, D.L., Waters, J.W., 1997. Observations of gravity waves with the UARS Microwave Limb Sounder. In: Hamilton, K. (Ed.), *Gravity Wave Processes and Their Parameterization In Global Climate Models*, NATO ASI Series, vol. 50. Springer-Verlag, New York, pp. 103–120.

Principles of Chemical Change

RP Wayne, University of Oxford, Oxford, UK

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

The main categories of chemical change important in atmospheric chemistry are homogeneous thermal gas-phase reactions, heterogeneous reactions on surfaces and within droplets, and photochemical reactions driven by solar radiation. Evolution from the reactants to the products is constrained by the energetics of the system, described by a potential-energy surface. The characteristics and mechanisms of unimolecular, bimolecular, and termolecular gas-phase reactions of both neutral and ionic species are presented. Four significant types of heterogeneous process are considered, and differences between gas-phase and condensed-phase behavior are examined. Absorption of visible and ultraviolet solar radiation may electronically excite atoms and molecules; the several possible fates of the excited species determine the pathways of atmospheric photochemistry.

Equilibrium and Change

Chemical change is essentially a reflection of a disequilibrium situation, in which a system employs excess free energy in order to drive spontaneous chemical processes. Ultimately, a condition of minimum free energy is reached, which corresponds to chemical equilibrium. The rates at which change occurs are the subject of *reaction kinetics* (see **Chemistry of the Atmosphere: Chemical Kinetics**). In this article, emphasis is placed on the types of chemical process that occur, with special reference to the atmosphere.

Homogeneous Gas-Phase Reactions

Chemical reactions are conveniently categorized by their *molecularity*. For an *elementary reaction*, the molecularity is the number of 'particles' (the word is used here to represent atoms or molecules) participating in the process of interest (see **Chemistry of the Atmosphere: Chemical Kinetics** for further details).

Bimolecular Processes

Bimolecular processes represent a good starting point for the discussion of chemical change, since they involve a fully developed interaction between two reactant species. When two particles approach each other closely enough, the energy of the system increases as the particles experience repulsive forces. Consider the interaction of an atom A with a molecule BC. The positions of the A, B, and C atoms can be defined in terms of their interatomic distances r_{AB} , r_{BC} and the A...B...C bond angle. That is, four dimensions are required to define energy in terms of the geometric coordinates. For a defined bond angle (for colinear approach, say), the reduced three-dimensional potential-energy surface (PES) can be represented by a two-dimensional contour map, as in **Figure 1(a)**. It is now useful to distinguish three different outcomes of the approach of A and BC, shown schematically in **Figure 1(b)**, **1(c)**, and **1(d)**. In **Figure 1(b)**, the pathway superposed on the PES shows the r_{AB} distance decreasing as A approaches BC, and the energy of the system increases up to the point at which the speed of approach becomes zero. At this point, the particles separate again on the same path along which they approached each other; they have undergone an *elastic collision*, in which all the energy is

conserved as translational motion. The process represented by **Figure 1(c)** is rather different, because the initial energy of interaction allows the particles to reach a part of the PES where there is an alteration of the geometry of the forces acting on BC. As a consequence, the return path is different from the approach path, with the B...C distance undergoing periodic oscillations. Translational energy has been converted to vibrational energy of BC in an *inelastic collision* that has brought about *energy transfer*. **Figure 1(d)** shows the final possibility, that of *reactive collision*. The motions of the particles now take them to the configuration where the distance r_{BC} is much larger than r_{AB} , so that the molecular species is now AB. That is, the reaction



has occurred. The contours of the surface suggest an analogy with mountainous terrain, in which there is a 'valley' that provides the lowest energy approach of the reactants, and the dotted line in the figure is that path. Beyond the point marked ' $\neq$ ', the energy starts to decrease again, and product formation is now energetically favorable. In the mountain analogy, the point $\neq$ is a 'col' or 'pass'; in terms of chemical kinetics, it is a *transition state*, of particular significance in a major theory of chemical kinetics (see **Chemistry of the Atmosphere: Chemical Kinetics**).

The reaction pathway (and the transition state) shown is only one of an infinite number of possibilities. In principle, if the PES is known, it is possible to calculate, using the laws of mechanics, the path followed for any initial 'starting' distance and direction of approach of the reactants A and BC. For a given speed, and internal excitation (vibration, rotation) of the reactants, the fraction of 'trials' leading to product formation can then be assessed, and this fraction is related to the probability of reaction. The ordinary macroscopic rate coefficient, k , (see **Chemistry of the Atmosphere: Chemical Kinetics**) can then be determined from a sum over the distributions of translational velocity, and vibrational and rotational excitation, appropriate to any temperature T : for thermal equilibrium, the *Maxwell-Boltzmann distribution* will be used. Such methods of *molecular dynamics* would be ideal for predicting rate coefficients if only it were possible to calculate from first principles the PESs for all reactions. Unfortunately, it is not yet feasible to perform these a priori calculations of surfaces, except for the very simplest reactions. Modern experimental kinetics can show the probability of passing from one set of reactant states

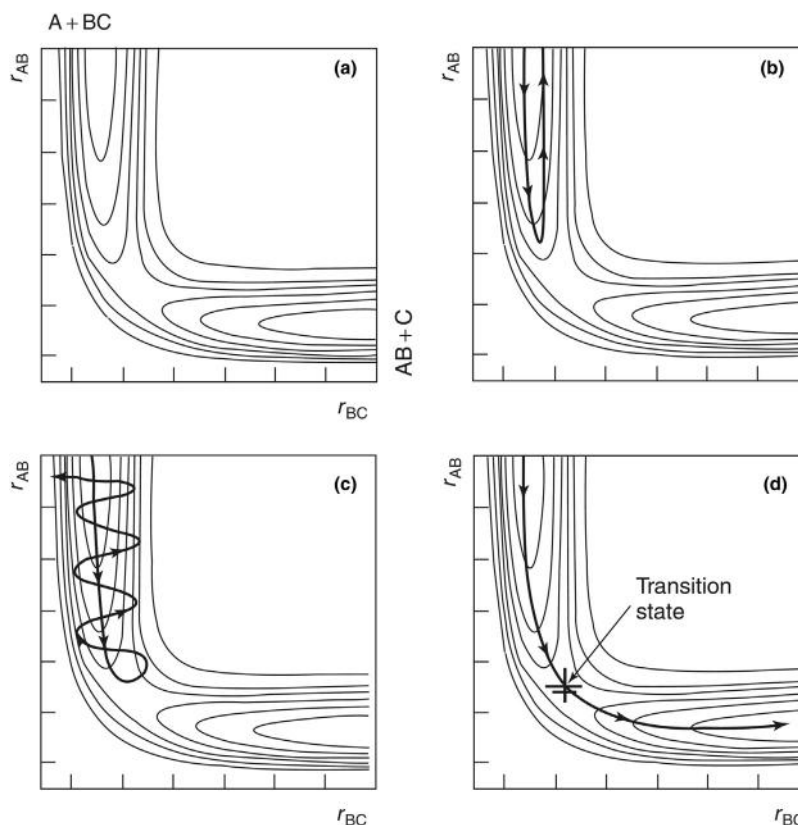


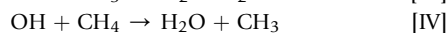
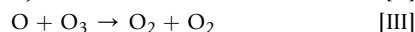
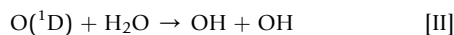
Figure 1 (a) Potential-energy surface for the system $A \cdots B \cdots C$ at a fixed bond angle; (b) elastic collision; (c) inelastic collision; and (d) reactive collision.

(translational, vibrational, rotational, etc.) to one set of product states: so-called *state-to-state kinetics*. The results of such experiments can then be used to test hypothetical PESs. In this case, however, the molecular dynamic calculations have lost their predictive value, and the simplifications described in the article **Chemistry of the Atmosphere: Chemical Kinetics** are usually employed instead.

It is now appropriate to examine the types of reactive bimolecular encounter that can occur in atmospheric chemistry. The most important are

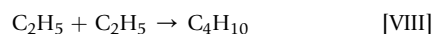
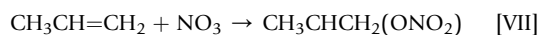
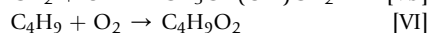
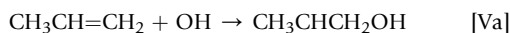
1. atom or group transfer;
2. addition;
3. exchange (metathesis).

Reaction [I] is, of course, formally an example of atom transfer. Real examples of importance in atmospheric chemistry include the processes



Reactions [III] and [IV] are hydrogen-atom transfers, while **reaction [II]** is an oxygen-atom transfer.

Additions can be illustrated for atmospheric chemistry by the processes

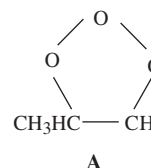


We shall discuss reactions of this type further in connection with termolecular processes.

The third category of bimolecular reaction, the exchange process, is relatively rare in atmospheric chemistry (and, indeed, in gas-phase chemistry in general). Many apparent examples are, in reality, made up of several stages involving, as the bimolecular steps, transfers, and additions. A true exchange reaction requires the production of a multicenter transition state. One example from the atmosphere that illustrates the problem is that of the ozonolysis of a simple alkene



In this case, O_3 initially adds across the double bond of the propene to form A.



This species undoubtedly goes on to form the products given for the two branches of **reaction [IX]**, which appear to be those of an exchange reaction. The species may not, however,

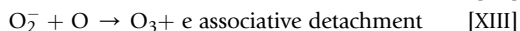
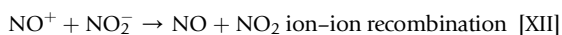
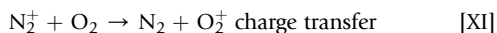
satisfy the formal conditions for being a true transition state, but rather be an energy-rich adduct (ozonide). In this latter case, the reaction is bimolecular addition, followed by unimolecular decomposition (see later) of the adduct.

Reactions of Ions in the Gas Phase

A more convincing example than [reaction \[IX\]](#) of a genuine exchange reaction is afforded by a process of considerable significance in the ionosphere



This is an example of an *ion–molecule reaction*. Although many ion–molecule reactions have their counterparts in neutral chemistry, some others do not. Three bimolecular processes that are peculiar to ion–molecule reactions are the steps



In all three of these examples, it is the charge – positive or negative – that is involved in the unusual behavior.

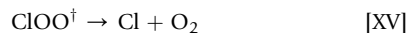
Charge on the reactant species brings with it another influence on ion–molecule reactions, which is that the rate coefficients for reaction can be orders of magnitude larger than those for the comparable reactions for neutral reactants. This question is explored in more detail in the article **Chemistry of the Atmosphere: Chemical Kinetics**. The effect arises because of long-range attractive forces between the reactants. The PES of [Figure 1](#) shows a monotonic increase from large separations of the reactants until the transition state. For charged species, however, the Coulombic interactions may cause the energies to decrease at some separations, with the consequence that long-range attraction draws the species together. The effects are obviously greatest for ion–ion reactions in which the ions have charges of opposite signs, as they do, for example, in [reaction \[XII\]](#). Coulombic forces operate over shorter ranges between ions and neutral molecules possessing a dipole; even nonpolar neutral molecules experience some attraction as a result of dipoles induced as the ionic species approaches.

Unimolecular Reactions

If the species A introduced in connection with [reaction \[IX\]](#) is really an energy-rich adduct (rather than a transition state), then its subsequent decomposition to form the products of steps [\[IXa\]](#) or [\[IXb\]](#) is of the form



which is a *unimolecular decomposition*, since it involves only one chemical species in an elementary reaction step. A typical atmospheric example is the decomposition of the species $\text{ClOO}^\ddagger$

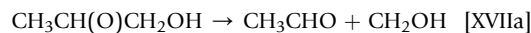


which is important in the chlorine-catalyzed destruction of stratospheric ozone. Although decompositions are the most important atmospheric unimolecular reactions, *isomerization*

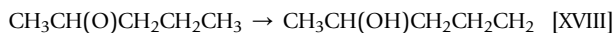


is another unimolecular process that is of atmospheric significance. Alkoxy radicals, in particular, can undergo both

fragmentation and isomerization, depending on their structures. The alkoxy radicals derived from the products of [reaction \[Va\]](#) and [\[Vb\]](#) in the oxidation of propene, for example, fragment to yield carbonyl compounds and further hydroxyl radicals



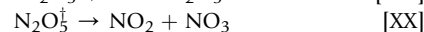
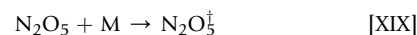
while the 2-pentoxyl radical, which is an intermediate in the oxidation of *n*-pentane, undergoes isomerization



via a cyclic transition state.

The formation of the energy-rich species represented by $\text{AB}^\ddagger$ is easy enough to understand if the energy derives from chemical reaction, as it might do in the formation of the intermediate ozonide in [reaction \[IX\]](#). Photochemical excitation, to be discussed later in this article, is another way in which an energy-rich species might be formed. However, many of the unimolecular processes of interest are *thermal reactions*. The energy that drives them is gained initially by elastic collisions (see above) with other molecules, represented for this purpose by the symbol M, which is called in this context a *third body*.

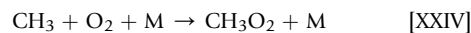
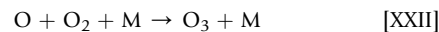
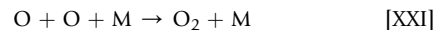
For example, the thermal unimolecular dissociation of N_2O_5 consists of two steps; the first is the *thermal activation* by collision and the second is the real unimolecular decomposition



In the Earth's atmosphere, M is generally, of course, the mixture of the major gases N_2 and O_2 . It is the activation (and a corresponding deactivation) step that gives rise to the special kinetic behavior of unimolecular reactions that is presented in the article **Chemistry of the Atmosphere: Chemical Kinetics**. Experimental evidence has shown clearly that the energy already stored within a molecule (for example, as the thermal equilibrium population of vibration) can contribute to the energy required for dissociation, so that not all the energy has to come from the collisional activation step [\[XIX\]](#). It further appears that the energy can flow fairly freely between vibrational modes in a molecule, and that the rate of dissociation (or isomerization) of a molecule depends on the total excess energy possessed by the molecule over the critical energy required to bring about the chemical change under consideration. It is all these features taken together that must be accommodated in sophisticated theories of the kinetics of unimolecular reactions.

Termolecular Addition Reactions

One class of termolecular reaction is of particular note in atmospheric chemistry, and that is the one involving the combination (often, and usually erroneously, called 'recombination') of atoms and small radicals. The processes



are all of this kind. As explained in the article **Chemistry of the Atmosphere: Chemical Kinetics**, such reactions may be

kinetically third order, as the chemical equations suggest, but may also become second order at high enough pressures. Reaction [XXI] can be used to explain why the third body, M, is essential in many of the reactions, and thus needs to be written into the equation. Figure 2 shows some selected potential-energy curves for the O₂ molecule, of which the lowest ('ground state') is labeled 'X' and is the subject of the current discussion. This curve is equivalent (for BC = O₂) to a section of the PES of Figure 1(a) taken parallel to the abscissa, and at large r_{AB} . Even if the two O atoms only just drift toward each other at near-zero velocity along the asymptote at large O–O separation, the newly formed O₂ molecule nevertheless possesses enough energy to redissociate on the first vibration. Only if enough energy is removed from the O₂[†] during this vibration, is there the possibility that the newly formed entity will be stabilized sufficiently to survive as an O₂ molecule. It is there, of course, that the collision with M comes in. The rate-determining step is this stabilization up to pressures of several hundred bar for the atom + atom system, and the kinetics remain third order until such pressures are reached. The situation is changed if the newly formed molecule is larger than diatomic. The product molecules of reactions [XXI], [XXII], [XXIII], and [XXIV] possess two, three, five, and six atoms, implying zero, two, eight, and eleven internal vibrational modes in addition to vibration in the new bond. The point is that the bond energy of the newly formed molecule can flow into these vibrations to some extent during the first vibration of the new bond. The efficiency of energy flow increases with an increasing number of modes, so that the requirements for removal of the energy by external collisions become increasingly less stringent. Now, the demand is that a collision with M is required to remove energy before the bond energy reassembles in the critical bond. It may be that the collisional removal of energy is no longer rate determining, in which case the reaction becomes kinetically second order. These ideas are all explored in a more quantitative form in the article **Chemistry of the Atmosphere: Chemical Kinetics**. For reactions such as

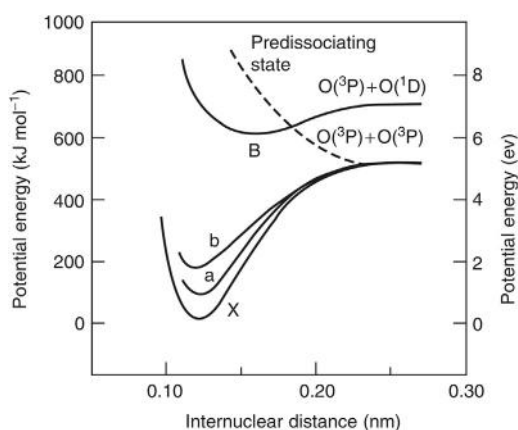
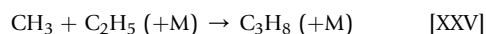


Figure 2 Potential-energy curves for some selected states of molecular oxygen. The solid curves represent bound states, and the labels are here just used as identifiers. The actual electronic states are $X^3\Sigma_g^-$, $a^1\Delta_g$, $b^1\Sigma_g^+$, and $B^3\Sigma_u^-$. The states of atomic oxygen with which the different curves correlate are indicated on the asymptotes at large internuclear distances.

the combination of two polyatomic radicals, for example CH₃ and C₂H₅



the reaction would probably be second order, and apparently bimolecular, throughout the troposphere, the atmospheric region in which it is likely to occur. It is for this reason that the addition reactions [V]–[VIII] were written as simple bimolecular processes, even if, in principle, at a sufficiently low pressure a third body would be required to stabilize the products suggested. What is worth emphasizing is that the combination processes are really just another facet of the unimolecular chemistry discussed earlier, because the stabilization required is that of an energy-rich molecule, here formed by what is really just chemical activation.

The Energetics of Chemical Change

The enthalpy (ΔH_r) of a chemical reaction is the difference between the enthalpies of the products and the reactants, corresponding to the difference in limiting 'heights' between the extreme bottom right and extreme top left of any of the panels of Figure 1, so long as the energy is expressed in appropriate units. All the enthalpies are normally given for a 'standard' pressure or concentration, and at a specified temperature, but for simplicity this aspect is ignored here. The way in which the energy of the system changes as reaction progresses is more easily seen if the potential energy is plotted as a function of a distance coordinate corresponding to the distance traveled along the reactive pathway shown in Figure 1(d). Figure 3 is such a representation of energy as a function of *reaction coordinate*, in this case for an *exothermic* reaction (negative ΔH_r): for passage in the reverse direction (i.e., from AB + C to A + BC), the reaction is *endothermic* (positive ΔH_r). Regardless

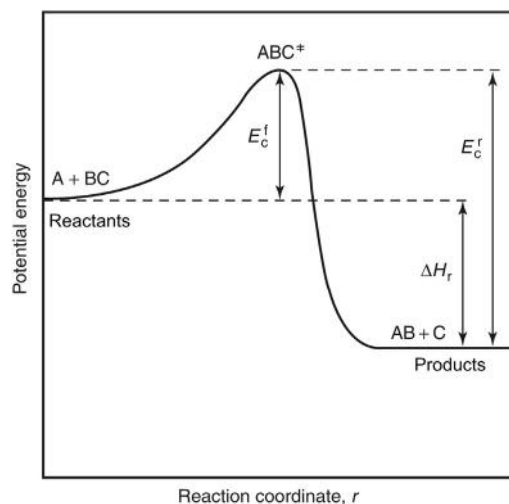


Figure 3 Energy of a reacting system as the reaction proceeds. The energy is represented as a function of reaction coordinate, which is essentially the distance traversed along the reaction path marked in Figure 1(d). The enthalpy of reaction, and the activation energies for forward and reverse reactions are marked on the diagram to clarify the relation between these quantities.

of whether the reaction under consideration is exothermic or endothermic, it can be seen that there is nevertheless a barrier to reaction in either direction, which corresponds to the energy of the transition state relative to that of the starting species. As explained earlier, and discussed in more detail in **Chemistry of the Atmosphere: Chemical Kinetics**, this barrier is thought to be the explanation of the experimentally observed activation energy that results in most thermal bimolecular reactions showing an increase in rate with increasing temperature.

Figure 3 makes clear a simple relationship between the barrier heights (and, therefore, expected activation energies) for forward and reverse reactions and the enthalpy of reaction. If the barrier heights are E_c^f and E_c^r , then

$$E_c^r = E_c^f - \Delta H_r \quad [1]$$

In principle, the vibrational zero-point energies of the reactants and transition state ought to be included in calculating the energy required to pass over the barrier. Indeed, at elevated temperatures, the calculation should provide a weighted mean for all the populated vibrational levels. Experimental evidence clearly supports this theoretical notion, since many reactions exhibit *kinetic isotope effects*, the most important of which is associated with zero-point energy differences. For example, in many H- or D-atom transfer reactions, vibrations that remain in the transition state do not involve the H or D atom. Vibrations in the reactants may, however, include those with -H or -D contributions. The lower vibrational frequency of the D-substituted species compared with the unsubstituted molecule puts the zero-point energy lower, and the barrier height to be surmounted is correspondingly larger. Rate coefficients for the deuterated species (k_D) are thus lower than those for the undeuterated molecule (k_H). Calculations of the vibrational frequencies suggest a maximum value for k_H/k_D of not much more than five at ambient temperature, and experimental observations give similar results for straightforward activated reactions, thus lending confidence in the association of activation energies with the height of the 'col' on the PES.

Heterogeneous Chemistry

The liquid and solid particles suspended in the atmosphere significantly influence the chemistry that can occur. Included in the category of atmospheric *heterogeneous chemical reactions* are not only those occurring at the interface between two phases, but also those in which initial transfer of reactants from the gaseous to a condensed phase is followed by homogeneous chemical change within that condensed phase. Atmospheric chemistry involving clouds, fogs, rain droplets, ice particles, and other solid and liquid aerosol particles requires transfer of gas-phase molecules to the condensed-phase system across the particle interface. Reactions occurring inside particles are confined to the liquid phase, since diffusion coefficients within solids are too small to allow significant reaction rates. On the other hand, reactions on solid surfaces are thought to be of very considerable atmospheric significance. An added degree of complexity arises when the particle is liquid, as is the case for droplets in the troposphere, and possibly for stratospheric sulfate aerosol, which may be in the form of supercooled liquid sulfuric acid.

Four simple categories can be recognized for the chemistry that creates aerosols and droplets, and occurs on and in them:

1. condensation of a single component;
2. reaction of more than one gas to form a new particle;
3. reaction of gases on a preexisting particle; and
4. reactions within the particles themselves.

Category (1), the condensation of a single gaseous component to form a new suspended particle is *homogeneous, homomolecular nucleation*. The most obvious example is the aggregation of sufficient H_2O molecules from the gas phase to produce a droplet of liquid, or particle of solid, water. Category (2) is the analogous process for reacting species involving two or more gases to form a condensable product species in a *homogeneous, heteromolecular process*. A typical example is the reaction between gaseous NH_3 and HNO_3 to form particles of NH_4NO_3 . But the same reaction can occur on a particle that already exists, a process that is a *heterogeneous, heteromolecular reaction* of type (3). Heteromolecular reactive condensation of gas-phase molecules on preexisting particles is sometimes called *aerosol scavenging*. It can have an impact on bulk tropospheric chemistry by providing a sink for nitrogen and hydrogen species such as HNO_3 , NO_3 , N_2O_5 , H_2O_2 , and HO_2 , as well as organic nitrates and peroxides. Clouds and raindrops have a major effect on gas-phase species through the scavenging mechanism. Finally, category (4) includes chemical reactions that occur within the aerosol itself to form particles of changed composition, as in the oxidation of SO_2 to sulfate ions in clouds. These are multiphase processes, since they involve transfer from (and perhaps back to) the gas phase.

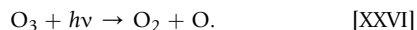
How do gas-phase and condensed-phase reactions differ? The solvent may substantially alter the chemistry of a condensed-phase reaction. To start with, the solvent hinders free motion of the reactants, with the result that the assumptions made in developing theories of gas-phase kinetics are no longer valid. Further, and of potentially great importance, the reactants, activated complexes or intermediates, and products can also all interact with the solvent, perhaps making possible reactions that do not occur in the gas phase. For example, reactions in aqueous solution often involve ionic processes, because the high relative permittivity and polar nature of the solvent make reactant ionization energetically accessible. Another key difference is that an *encounter pair* of reactants may find themselves undergoing multiple collisions within a *solvent cage*, leading to interesting differences in the rates of reaction in gas and condensed environments. We are straying now into the field of chemical kinetics, and further discussion is more appropriately continued in **Chemistry of the Atmosphere: Chemical Kinetics**.

Photochemical Change

The reactions discussed so far have been 'thermal': that is, the driving force has been the exothermicity of the chemical change, and barriers to reaction have been overcome by the thermal translational energy of the reactants, and perhaps by their internal vibrational energy. Another class of reaction of enormous importance is that of *photochemical reactions*, in which the driving force is the absorption of radiation. An atmosphere is a giant photochemical reactor, in which the light

source is the Sun. Radiation, generally in the visible and ultraviolet regions, either fragments the atmospheric constituents to produce atoms, radicals, or ions, or excites the constituents, without chemical change, to alter their reactivity.

Planck's law states that the energy of one photon of frequency is $h\nu$, where h is Planck's constant. Photochemists therefore usually write ' $h\nu$ ' in chemical equations as a shorthand for the photon that is a reactant, as, for example, in the *photolysis* of ozone



Planck's constant is known, so that the energy per photon can be calculated immediately. A useful conversion expresses E in the molar units kJ mol^{-1} and employs the wavelength of radiation (λ in nm) rather than the frequency itself

$$E = 119625/\lambda. \quad [2]$$

Thus the red extreme of the visible spectrum (~ 800 nm) corresponds to about 150 kJ mol^{-1} , and the violet extreme (~ 400 nm) to twice that energy. At shorter wavelengths lies the ultraviolet, conventionally subdivided into 'near' (λ approximately 400–200 nm), 'vacuum' (VUV, λ approximately 200–100 nm), and 'extreme' (EUV, λ approximately 100–10 nm) regions, with successively higher energies. The photon energy of red light is comparable with the bond energies of rather loosely bound chemical species. Of common gaseous inorganic species, ozone is, in fact, the only compound with such a small bond energy (the O–O₂ energy is $\sim 105 \text{ kJ mol}^{-1}$); nitrogen dioxide, with an O–NO bond energy of $\sim 300 \text{ kJ mol}^{-1}$ ($\cong 399$ nm) is more typical. Ionization becomes possible at the shorter wavelengths (e.g., ionization of NO at $\lambda \sim 135$ nm). X-rays, gamma rays, and galactic cosmic radiation constitute the shortest ($\lambda < 10$ nm) wavelengths of the electromagnetic spectrum. The very high photon energies are associated with an ability to penetrate as well as to ionize atmospheric gases. The point to be emphasized is that the visible region contains the lowest energy photons of the entire electromagnetic spectrum that are capable of promoting chemical change in single quantum events. Many more photons arrive each second at a planet at longer wavelengths, but they can only heat the atmosphere up.

The wavelengths at which chemical change becomes possible also correspond roughly to the energies at which *electronic transitions* are excited in atoms and molecules. Longer wavelengths tend to excite molecular vibrations or rotations. Although high vibrational levels are involved in some photochemical processes, electronic excitation is the spectroscopic step most frequently associated with photochemical change. Even very weak, highly forbidden, transitions can contribute to atmospheric absorption because of the large optical paths involved. Small, light chemical species generally show intense electronic absorption at shorter wavelengths than more complex compounds. Molecular oxygen absorbs strongly for $\lambda < 200$ nm, H₂O for $\lambda < 180$ nm, CO₂ for $\lambda < 165$ nm, while N₂ and H₂ absorb significantly only for $\lambda < 100$ nm. It is the limitation on laboratory experiments in air, resulting from the O₂ absorption that has led to the term 'vacuum' ultraviolet for $\lambda < 200$ nm. Atmospheres tend to act as filters cutting out short-wavelength radiation, since the absorptions of their major constituents are generally strong at wavelengths shorter than

the threshold value. As a result, photochemically active radiation that penetrates deeper into an atmosphere is of longer wavelength, and the chemistry characterized by lower energies, than that absorbed higher up. The principle is well exemplified by the chemistry of Earth's atmosphere. Tropospheric photochemistry is dominated by species such as O₃, NO₂, SO₂, and HCHO, which absorb in the near-ultraviolet region. At progressively higher altitudes, photodissociation of O₂ and photoionization phenomena become the most important processes. At ground level, only radiation with $\lambda > 300$ nm ($\cong 400 \text{ kJ mol}^{-1}$) remains, and the peak intensity is at $\lambda \sim 500$ nm ($\cong 240 \text{ kJ mol}^{-1}$).

Photochemical Primary Processes

Absorption of a photon of photochemically active radiation leads to electronic excitation, a process that may be represented symbolically as



Many fates of the excited AB* molecule are recognized, and several of them occur in atmospheres. **Figure 4** summarizes the processes most frequently encountered, and they will be discussed briefly in turn.

Photodissociation, Photoionization, and Intramolecular Energy Transfer

Routes (i) and (ii) lead to *fragmentation* of one kind or another. Fragmentation of a chemical species following absorption of light is one of the most important photochemical processes in atmospheric chemistry. *Photodissociation* may come about when the energy of the absorbed photon exceeds the binding energy of the chemical bond under consideration. That is, the species AB* excited initially in the absorption event [XXVII] can lie energetically above the dissociation threshold in the molecule, and the bond can then rupture in some way. Two main mechanisms are recognized for this photochemical rupture. The potential-energy curves of **Figure 2** illustrate these processes for the photodissociation of the O₂ molecule. The molecule in its ground state, labeled in the diagram as X, can absorb ultraviolet radiation to populate the curve labeled B. If the wavelength of absorbed radiation is short enough that the B state is populated above the dissociation limit of the products O(³P) + O(¹D), then *direct dissociation* (also called *optical dissociation*) can lead to the formation of these fragments.

Process (iii) of **Figure 4** is *intramolecular energy transfer*, and can generate a new electronic state of the same molecule by a *radiationless transition*. The dashed curve in **Figure 2** for O₂ crosses the B state, and is populated by such transitions. This new electronic state of O₂ is *repulsive*, and falls apart to form the fragments O(³P) + O(³P). Both mechanisms are thought to operate in the atmospheric production of O atoms from O₂. The overall dissociative process is referred to as *predissociation*, and leads to the dissociation at lower energies (longer wavelengths) than are needed for direct dissociation. Note that the direct dissociation route produces an (electronically) excited atomic fragment, O(¹D), as well as a ground-state atom. Such excitation in the fragments of photodissociation is of common occurrence. Indeed, the photodissociation of O₃ by ultraviolet

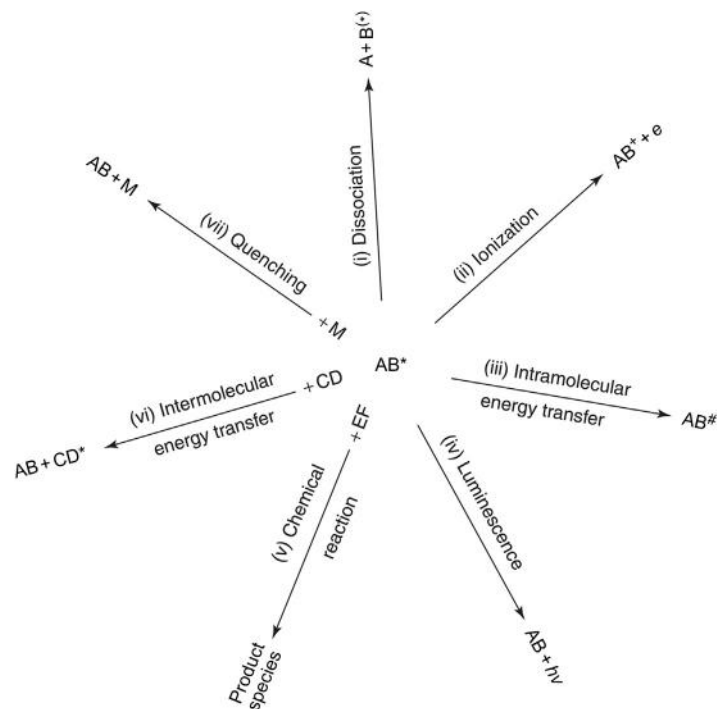
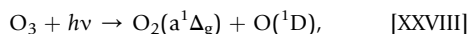


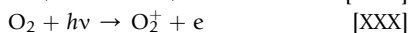
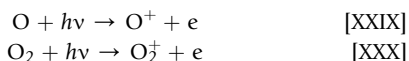
Figure 4 Atmospherically significant fates of an electronically excited species, AB^* . The excitation is usually a result of absorption of visible or ultraviolet radiation. The symbols $*$ and $\#$ indicate different electronically excited states. Although the processes are written for a molecule AB , excited atoms can also decay via any of the pathways except (i) and (iii).

radiation, already represented by equation [XXVI], can produce two excited fragments



and both fragments are of importance in atmospheric chemistry. The curve labeled a in Figure 2 is, in fact, the potential-energy curve for the electronic state of O_2 formed in this atmospherically most important step.

Photoionization (process (ii) in Figure 4) may be regarded as a special case of photodissociation, but one in which the products are a positively charged ion and an electron. The reactions

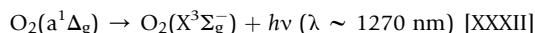
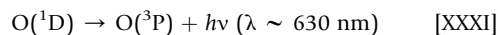


are typical of photoionization. In general, the energies (*ionization potentials*) required to remove an electron from an atom or molecule are larger than those needed to split a molecule into chemical fragments. For example, the ionization process [XXX] requires $\lambda \leq 103$ nm. Mechanisms of photoionization are analogous to those for dissociation, both *direct ionization* and *preionization* (*autoionization*) being recognized. Excited electronic (and, in molecules, vibrational and rotational) states of the ions may be generated.

Luminescence or Emission

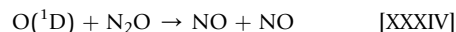
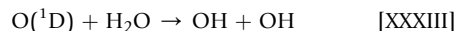
Pathway (iv) of Figure 4 is the emission of radiation, usually termed *luminescence*. A distinction is often made in photochemistry between radiative transitions that are allowed, in which case the phenomenon is termed *fluorescence* and those in

which it is forbidden, in which case the phenomenon is *phosphorescence*. Such emission is responsible for many important features of the *atmospheric airglow*. For example, both the excited species formed in reaction [XXVIII] contribute to the *dayglow*.



Chemical Reaction

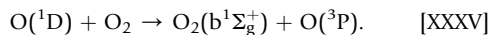
Pathway (v) shown in Figure 4 is that of *chemical reaction*, in which species are formed that are chemically distinct from the starting reactants. It includes all processes where reaction is possible only for, or rates are enhanced with, electronically excited reactants. The chemical reactivity of an electronically excited species may be altered from that of the ground state both as a result of the excess energy that it possesses and because of the altered electronic structure. Two very important examples of the chemical reaction of an excited species in the atmosphere again involve the atom $O(^1D)$. The reactions of ground-state, $O(^3P)$, atoms with H_2O and N_2O are endothermic, and very slow under atmospheric conditions, while the reactions with excited atoms



are exothermic and fast, and make essential contributions to the formation of OH and of NO in various regions of the atmosphere.

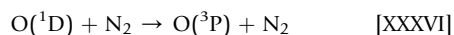
Intermolecular Energy Transfer and Quenching

Intermolecular energy transfer, route (vi) of Figure 4, excites a molecule that is often chemically distinct from the absorbing species. An atmospheric example can be drawn from the airglow of oxygen yet again. Figure 2 shows a curve labeled b; this curve represents the electronic state $O_2(b^1\Sigma_g^+)$ which contributes to the airglow at $\lambda \sim 760$ nm. Some of the $O_2(b^1\Sigma_g^+)$ is populated by direct absorption of solar radiation during the day, but an additional contribution is made by the energy-transfer process



That is, the excited oxygen atom has transferred (some of) its excess energy to the O_2 .

Quenching or *deactivation* (pathway vii) is a special case of intermolecular energy transfer, where electronic excitation is degraded to vibrational, rotational, and translational modes. It competes with the other routes for decay of an electronically excited species. For example, the rapid quenching reaction



imposes a limit on the extent of OH or NO formation in reactions [XXXIII] or [XXXIV], especially in view of the relatively small mixing ratios of H_2O or N_2O compared to N_2 . Since quenching is a bimolecular process, it is pressure- (and therefore altitude-) dependent. Thus the competition with unimolecular processes such as emission means that airglow features often show a marked intensity decrease with decreasing altitude.

See also: **Chemistry of the Atmosphere:** Chemical Kinetics; Laboratory Kinetics.

Further Reading

- Crim, F.F., 2008. Chemical dynamics of vibrationally excited molecules: controlling reactions in gases and on surfaces. *Proceedings of the National Academy of Sciences* 105, 12654–12661.
- Friedman, L., Reuben, B.G., 2007. A review of ion-molecule reactions. In: Prigogine, I., Rice, S.A. (Eds.), *Advances in Chemical Physics*, 19. John Wiley & Sons, Inc., Hoboken, NJ.
- Holloway, A.M., Wayne, R.P., 2010. *Atmospheric Chemistry*. Royal Society of Chemistry, London.
- Houston, P.L., 2006. *Chemical Kinetics and Reaction Dynamics*. Dover Publications, New York.
- Lee, Y.P., 2003. State-resolved dynamics of photofragmentation. *Annual Reviews of Physics and Chemistry* 54, 215–244.
- Levine, R.D., 2005. *Molecular Reaction Dynamics*. Cambridge University Press, Cambridge.
- Moore, G.B., Smith, I.W.M., 1996. State-resolved studies of reactions in the gas phase. *Journal of Physical Chemistry* 100, 12848–12865.
- Pilling, M.J., Seakins, P.W., 1995. *Reaction Kinetics*. Oxford University Press, Oxford.
- Pilling, M.J., Smith, I.W.M. (Eds.), 1987. *Modern Gas Kinetics*. Blackwell Scientific Publications, Oxford.
- Reid, J.P., Sayer, R.M., 2003. Heterogeneous atmospheric aerosol chemistry: laboratory studies of chemistry on water droplets. *Chemical Society Reviews* 32, 70–79.
- Smith, I.W.M., 1980. *Kinetics and Dynamics of Elementary Gas Reactions*. Butterworths, London.
- Vione, D., Maurino, V., Minero, C., et al., 2006. Photochemical reactions in the tropospheric aqueous phase and on particulate matter. *Chemical Society Reviews* 35, 441–453.
- von Clarmann, T., Hase, F., Funke, B., et al., 2010. Do vibrationally excited OH molecules affect middle and upper atmospheric chemistry? *Atmospheric Chemistry and Physics* 10, 9953–9964.
- Wayne, R.P., 1988. *Principles and Applications of Photochemistry*. Oxford University Press, Oxford.
- Wayne, R.P., 2000. *Photochemistry and Kinetics Applied to Atmosphere*. Chapter 3 in *Chemistry of Atmospheres*, third ed. Oxford University Press, Oxford. pp. 97–137.

Radioactivity: Cosmogenic Radionuclides

D Lal, Scripps Institution of Oceanography, La Jolla, CA, USA

© 2003 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 5, pp 1891–1990, © 2003, Elsevier Ltd.

Introduction

The Earth receives a great variety of radiations at the top of the atmosphere, over a very wide range of energies. Among these, the only radiation that is capable of producing significant changes in the isotopic composition of matter is the cosmic radiation. Nuclear interactions of cosmic ray particles with the constituent nuclei of the atmosphere produce several stable and radioactive nuclei. These nuclides (isotopes of elements) are termed cosmogenic nuclides or cosmogenic isotopes. Thus the atmosphere is continually labeled 'naturally' by nuclides, which serve as tracers for studying the nature of large-scale motions of the atmosphere: their distribution allows one to directly determine the nature of the large-scale air circulation, as well as the time scales of their removal from the lower atmosphere by wet precipitation and dry deposition. These tracers have several unique features: (1) the source functions of the tracers in all parts of the atmosphere can be determined precisely as a function of time; (2) several radionuclides with different chemical properties and half-lives varying in the range of an hour to more than a million years are produced by cosmic rays, whereby air transport/mixing and aerosol scavenging processes by wet precipitation can be studied quantitatively, on time scales ranging from hours to years.

The usefulness of tracers in fluids undergoing complex, space- and time-dependent motions cannot be underestimated. They provide space–time integrals of motion on time scales of the mean lives of the nuclides. It is interesting to note that a quantum jump in our understanding of the atmospheric processes occurred in the 1950s when artificial radionuclides were introduced in the atmosphere as a result of nuclear weapons tests. The spring 1954 Castle thermonuclear test introduced a large amount of artificial ^3H (tritium: half-life 12.3 years) into the atmosphere. Studies of its concentration in rains clearly showed that it was removed from the atmosphere in periods as short as weeks. Similarly, the injections of the fission radionuclide ^{90}Sr into the atmosphere in stratospheric tests of nuclear weapons led to the first insight into time scales of removal of aerosols from the stratosphere to the troposphere and finally onto the surface of the Earth, within periods of the order of few years. These were remarkable observations indeed, since conventional physical meteorological studies did not provide even rough estimates of the time scales involved in circulation and self-purging of the atmosphere, although enough was known about the physical state of the atmosphere. However, a model had then been proposed for the grand Equator-to-Poles troposphere–stratosphere–troposphere circulation, namely the Brewer–Dobson model, which remained controversial until recently.

The naturally produced cosmogenic radionuclides ^{14}C (half-life 5730 years) and ^3H had been discovered before the first Castle test in the atmosphere, in 1947 and 1951, respectively. Soon thereafter, in the mid-1950s, the long-lived radionuclide ^{10}Be (half-life 1.5 million years), and several

short-lived cosmogenic radionuclides (^{22}Na , ^{35}S , ^7Be , ^{33}P , and ^{32}P) with half-lives in the range of weeks to years were discovered. Additionally, seven short-lived radionuclides of half-lives ranging from 30 min to 20 h were detected. See Table 1 for half-lives and production mechanisms of the cosmogenic nuclides. Studies of their distribution in the atmosphere and in wet precipitations showed clearly the power of these nuclides as tracers for studying wide-ranging problems in atmospheric dynamics. The great attraction of tracers of different half-lives in such studies lies in the fact that they allow integration of complex fluid motion over very different space and time scales, and the greatest virtue of cosmic ray tracers lies in the fact that they are naturally injected into the atmosphere continuously at a rate that can be determined fairly accurately. Fortunately, tracers of different half-lives and chemical properties are available to answer most of the important questions in atmospheric transport and removal processes.

The importance of the study of atmospheric transport and aerosol scavenging processes cannot be overemphasized. First, it must be considered as the central problem in physical and chemical meteorology, which should in fact appear as a tangible

Table 1 Cosmogenic nuclides^a produced in the Earth's atmosphere with half-lives exceeding 30 min

Nuclide	Half-life	
	(s = stable)	Main target nuclei
^3H	12.3 years	N, O
^3He	s	N, O
^7Be	53 days	N, O
^{10}Be	1.5×10^6 years	N, O
^{14}C	5730 years	N, O
^{20}Ne , ^{21}Ne , ^{22}Ne	s	Ar
^{22}Na	2.6 years	Ar
^{24}Na	15 h	Ar
^{26}Al	7.1×10^5 years	Ar
^{28}Mg	21.2 h	Ar
^{31}Si	2.6 h	Ar
^{32}Si	~ 150 years	Ar
^{32}P	14.3 days	Ar
^{33}P	25.3 days	Ar
$^{34\text{m}}\text{Cl}$	32 min	Ar
^{35}S	87 days	Ar
^{36}Cl	3.0×10^5 years	Ar
^{37}Ar	35 days	Ar
^{38}Cl	37.3 min	Ar
^{38}S	2.9 h	Ar
^{39}Cl	55.5 min	Ar
^{39}Ar	268 years	Ar
^{81}Kr	2.3×10^5 years	Kr
^{129}I	1.6×10^7 years	Xe

^aArranged in order of increasing mass number.

product of studies of the physical structure of the atmosphere. Second, it is a prerequisite in understanding the chemistry of the atmosphere. Finally, with the increased injections of a variety of chemical compounds as a result of industrial activity, including CO_2 , fluorocarbons, CCl_4 , etc., it becomes necessary to develop a quantitative capability to estimate the evolution of their distributions in the atmosphere, in space and time. This is being accomplished using a variety of tracers: those produced naturally and those produced artificially. Radon-222 (^{222}Rn) is a naturally produced tracer. In view of its short half-life (3.8 days) and because it is injected into the atmosphere from the Earth's surface, it is especially useful for characterizing air motions in the lower troposphere. The production rates of cosmogenic nuclides are highest in the stratosphere, but still appreciable in the troposphere. They can therefore be used as tracers throughout the atmosphere. As will become clear later, cosmogenic nuclides show great potential for quantifying atmospheric transport and aerosol scavenging processes, as demonstrated in several experiments carried out since the mid-1960s. However, their full potential has yet to be exploited since no dedicated synoptic effort has been launched to study various cosmogenic nuclides in the air masses sampled. Seasonal, interannual, and decadal measurements would be required to provide robust transport/mixing and scavenging parameters for development of robust three-dimensional atmospheric models. The task is easily feasible with modern techniques of sampling air and measurements of cosmogenic nuclides.

Cosmogenic Nuclides Produced in the Atmosphere

Cosmic rays are composed of high-energy particles of nuclear matter, accelerated in shockwaves of type II supernovas. Typical energies of nuclei in the primary cosmic ray beam, incident at the top of the atmosphere, lie in the range of 1–10 GeV (1 GeV = 10^9 electron volts). The total energy range is of course much greater, covering 0.1– 10^{11} GeV. Most of the cosmic ray particles have velocities close to that of light and have energies sufficient to break nuclei into fragments. As a comparison, the binding energies of nucleons in nuclei are on the order of 10 MeV. During the passage of cosmic ray particles through the Earth's atmosphere, a nucleonic cascade develops as a result of nuclear interactions of the cosmic rays with atmospheric nuclei. As a result, the cosmic ray beam at depth is composed of a greater number of secondary particles that are capable of inducing further nuclear reactions. A variety of stable and radioactive nuclei are produced in the nuclear interactions of primary and secondary cosmic rays with the atmospheric nuclei N, O, Ar, Kr, and Xe. **Table 1** lists cosmogenic nuclides produced in the Earth's atmosphere that have half-lives exceeding 30 min, which are applied as tracers in earth sciences.

Cosmogenic nuclide production rates are strongly dependent on latitude and altitude in the atmosphere. These rates have been estimated rather accurately, and further attempts are under way to improve these estimates. Estimates have been made of the intensities of slow neutrons in the atmosphere, which allow one to estimate the altitude–latitude dependence of the rates of production of ^{14}C and ^{81}Kr in the atmosphere by the capture of slow neutrons by ^{14}N and ^{80}Kr , respectively. Most other isotopes are produced by high-energy spallation (fragmentation) of nuclei; their rates are approximately

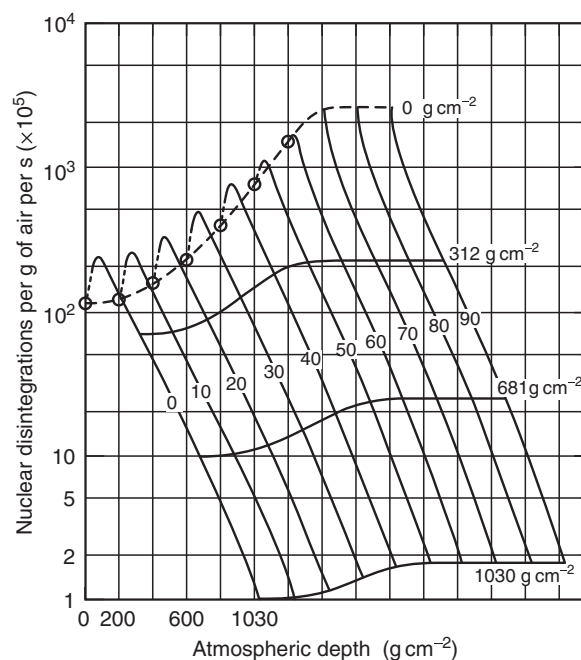


Figure 1 The rate of production of nuclear disintegrations with energy release above 40 MeV is plotted as a function of atmospheric pressure (from Lal and Peters, 1967). To avoid overlapping, curves for different latitudes have been displaced with respect to each other successively by 200 g cm^{-2} along the abscissa. From Lal (1966).

proportional to the rates of nuclear disintegrations, and have been estimated fairly accurately on the basis of measurements of slow neutrons in the atmosphere. **Figure 1** shows the altitude–latitude dependence of the rate of nuclear disintegrations in the atmosphere. For details of (small) departures in the altitude–latitude dependence of nuclide yields from nuclear disintegration rates, refer to Lal (1966) (see [Further Reading](#)). **Figure 1** therefore shows the approximate relative altitude–latitude dependence of all nuclides, except for ^{14}C and ^{81}Kr .

Criteria of Applicability of Cosmogenic Nuclides in Meteorology and Related Sciences

Before discussing the source strengths of cosmogenic nuclides in the atmosphere, and results to date of studies on them, we will briefly review the basis of their application in the study of atmospheric transport and removal processes, and also how one can obtain information about the nature of past atmospheric circulations from observations of the concentrations of these nuclides in continental and oceanic reservoirs.

After their creation in the atmosphere in nuclear interactions, most of the isotopes (**Table 1**) become oxidized; exceptions are ^3He , $^{37,39}\text{Ar}$, and ^{81}Kr . Among the oxides, ^{14}C mixes with the atmospheric CO_2 , whereas others quickly become attached to aerosols (primarily of size $<1 \mu\text{m}$ in diameter). Removal of all nuclides from the atmosphere occurs by radioactive decay and by physical/chemical transfer to other terrestrial reservoirs. In the lower atmosphere, nuclides are removed either by scavenging by droplets formed during condensation, or by gas exchange at air–water interfaces, as applicable for ^3He , $^{37,39}\text{Ar}$, and ^{81}Kr . **Figure 2** is a schematic

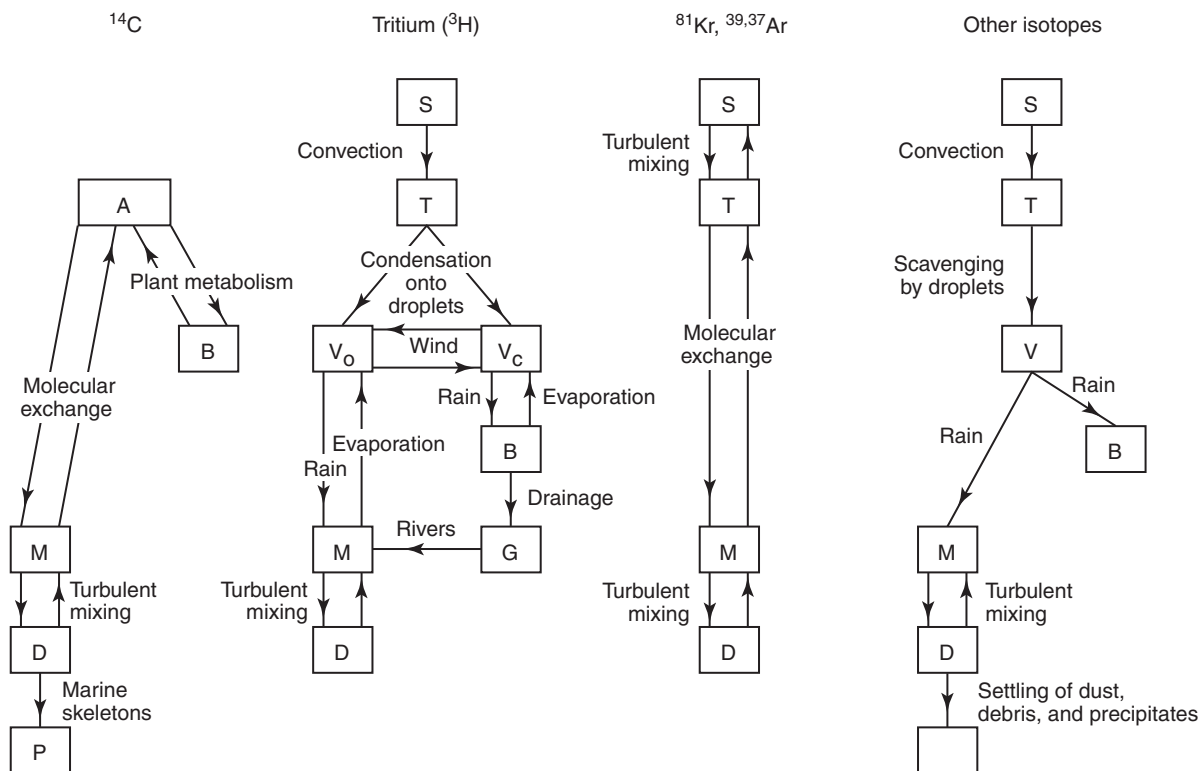


Figure 2 Schematic showing the principal avenues of transfer of different cosmogenic nuclides through the atmosphere to different terrestrial reservoirs. A, atmosphere; S, stratosphere; T, troposphere; V, water vapor; B, topsoil and biosphere; M, ocean mixed layer; D, deep sea. V_o and V_c refer to the vapor reservoirs over the ocean and continents, respectively. From Lal (1966).

of different modes of transport of cosmogenic nuclides classified into four principal categories: atmosphere (A), which is subdivided into stratosphere (S), troposphere (T), water vapor (V); topsoil and biosphere (B); ocean mixed layer (M); and deep sea (D).

The flow of tracers through the atmosphere depends on the chemical nature of the elements corresponding to the nuclides; their resulting concentrations depend on their half-lives and the rate constants for exchange/transfer between the reservoirs. The principal application for the cosmogenic nuclides are thus governed primarily by considerations of their expected transfer/flow through the reservoirs, their half-lives, and their chemical natures.

From the above discussion, it follows that, provided cosmogenic nuclides can be measured in the different reservoirs, they should serve as suitable tracers in certain applications. This has been found to be the case. Table 2 lists principal applications of cosmogenic nuclides formed in the atmosphere in nuclear reactions.

Finally, the distributions of the cosmogenic nuclides in the terrestrial reservoirs are expected to depend on the nature of stratospheric-tropospheric exchange and the nature of circulation within the troposphere, and on tropospheric scavenging processes, which are expected to be climate-dependent. Hence, observations of the concentrations of cosmogenic nuclides in the continental and marine sediments and in the polar ice sheets are expected to be indicative of temporal

changes in the nature of large-scale atmospheric circulation in the past.

Production Rates of Cosmogenic Nuclides and Their Inventories in Terrestrial Reservoirs

Table 3 presents estimates of production rates of cosmogenic nuclides with half-lives exceeding 2 weeks, separately for the troposphere, and their global inventory expected to be in secular equilibrium with their production. Table 4 presents production rates of cosmogenic nuclides with half-lives of 30 min to 24 h.

On the basis of the pathways of the nuclides and the inapproximate residence times within the reservoirs versus times for exchange with neighboring reservoirs, their estimated steady-state fractional inventories in the atmosphere and other reservoirs are presented in Table 5. Absolute inventories can be obtained by normalizing the relative inventories by their production rates as given in Table 3. The sources of these nuclides is in the atmosphere, and obviously their absolute inventories in different reservoirs depend critically on the model used for their circulation within the atmosphere and for scavenging from the troposphere. Information on present-day processes and rates can be obtained from current measurements, whereas those for the past have to be based on the concentrations in the polar ice sheets or in continental and marine sediments. Values in Table 5 are therefore useful only as an approximation to the expected values.

Table 2 Important characteristics and principal applications of selected cosmogenic nuclides produced in the atmosphere

Nuclide	Half-life	Application
Isotopes that do not form compounds		
³ He	Stable	Air–sea exchange; escape of helium from the atmosphere
³⁷ Ar	35 days	Atmospheric circulation and air–sea exchange
³⁹ Ar	268 years	Atmospheric circulation; vertical mixing in oceans
⁸¹ Kr	2.3×10^5 years	Ground water ages, and constancy of cosmic radiation
Isotopes that attach to aerosols/particles		
⁷ Be	53 days	Atmospheric circulation, vertical mixing in surface ocean waters
¹⁰ Be	1.5×10^6 years	Atmospheric circulation; role of particle scavenging in the coastal and open oceans; dating of sediments and accretions
²⁶ Al	7.1×10^5 years	Role of particle scavenging in the coastal and open oceans; dating of marine sediments and accretions
³² Si (HSiO ₃ , SiO ₂)	~ 150 years	Atmospheric circulation; labeling the dissolved oceanic silicon pool; atmospheric circulation
³³ P, ³² P	14.3, 25.3 days	Atmospheric circulation; labeling the dissolved oceanic phosphorus pool; tropospheric circulation
³⁹ Cl, ³⁸ S, ³⁸ Cl, ^{34m} Cl, ³¹ Si, ²⁸ Mg, ²⁴ Mg	0.5 h to 21 h	Cloud formation and dissipation, and aerosol scavenging processes by wet precipitation
Isotopes, of half-lives >2 weeks, that label constituent molecules in the atmosphere and the hydrosphere		
³ H (H ₂ O)	12.3 years	Atmospheric circulation; characterizing water molecules in the atmosphere, hydrosphere and cryosphere
¹⁴ C (CO ₂ , CO ₃ , HCO ₃)	5730 years	Atmospheric circulation; characterization of the carbon cycle reservoirs
³² Si (HSiO ₃ , SiO ₂)	~ 150 years	Atmospheric circulation; biogeochemical cycle of silicon
³³ P, ³² P (DIP, DOP)	14.3, 25.3 days	Atmospheric circulation; biogeochemical cycle of phosphorus

DIP and DOP refer to dissolved inorganic and organic phosphorus, respectively.

Table 3 Production rates of cosmogenic nuclides in the Earth's atmosphere, of half-lives exceeding 2 weeks; arranged in order of decreasing half-lives [#]

Isotope	Half-life	Production rate (atoms cm ⁻² s ⁻¹)		
		Troposphere	Total atmosphere	Global inventory
³ He	Stable	6.7×10^{-2}	0.2	3.2×10^3 tons ^d
¹⁰ Be	1.5×10^6 years	1.5×10^{-2}	4.5×10^{-2}	260 tons
²⁶ Al	7.1×10^5 years	3.8×10^{-5}	1.4×10^{-4}	1.1 tons
⁸¹ Kr ^b	2.3×10^5 years	5.2×10^{-7}	1.2×10^{-6}	8.5 kg
³⁶ Cl	3.0×10^5 years	4×10^{-4}	1.1×10^{-3}	15 tons [*]
¹⁴ C	5730 years	1.1	2.5	75 tons
³⁹ Ar ^c	268 years	4.5×10^{-3}	1.3×10^{-2}	52 kg
³² Si	~ 150 years	5.4×10^{-5}	1.6×10^{-4}	0.3 kg
³ H	12.3 years	8.4×10^{-2}	0.25	3.5 kg
²² Na	2.6 years	2.4×10^{-5}	8.6×10^{-5}	1.9 g
³⁵ S	87 days	4.9×10^{-4}	1.4×10^{-3}	4.5 g
⁷ Be	53 days	2.7×10^{-2}	8.1×10^{-2}	3.2 g
³⁷ Ar	35 days	2.8×10^{-4}	8.3×10^{-4}	1.1 g
³³ P	25.3 days	2.2×10^{-4}	6.8×10^{-4}	0.6 g
³² P	14.3 days	2.7×10^{-4}	8.1×10^{-4}	0.4 g

[#]Based on Lal (1966).

^dThe inventory of this stable nuclide is based on its atmospheric inventory, which includes an appreciable contribution from crustal degassing of ³He.

^bBased on the measured atmospheric ⁸¹Kr/Kr ratio of $(5.2 \pm 0.4) \times 10^{-13}$.

^cBased on the measured atmospheric ³⁹Ar/Ar ratio of (0.107 ± 0.004) dpm/liter Ar (STP).

^{*}Includes a rough estimate of ³⁶Cl produced by the capture of neutrons at the Earth's surface.

Observed Distributions of Short-Lived and Long-Lived Cosmogenic Nuclides

Extensive measurements exist of most of the short-lived and long-lived cosmogenic nuclides in all the terrestrial reservoirs where they are found. Relevant to this discussion are the observed concentration of the nuclides in the

atmosphere and their fallout on the Earth, which are governed by the nature of large-scale atmospheric circulation and scavenging processes. It would be beyond the scope of this article to discuss the whole of the data; instead we discuss below some of the highlights to illustrate the scope of cosmogenic tracers.

Table 4 Production rates of several short-lived isotopes produced by cosmic rays in the Earth's atmosphere, of half-lives 30 min–24 h, arranged in order of decreasing half-lives

Isotope	Half-life	Production rate (atoms cm ⁻² s ⁻¹)	
		Troposphere	Total atmosphere
²⁸ Mg	21.2 h	4.6×10^{-5}	1.7×10^{-4}
²⁴ Na	15.0 h	8.2×10^{-5}	3.0×10^{-4}
³⁸ S	2.9 h	1.7×10^{-5}	4.9×10^{-5}
³¹ Si	2.6 h	1.5×10^{-4}	4.4×10^{-4}
³⁹ Cl	55.5 min	4.9×10^{-4}	1.4×10^{-3}
³⁸ Cl	37.3 min	7.7×10^{-4}	2.0×10^{-3}
^{34m} Cl	32.0 min	1.1×10^{-4}	2.0×10^{-4}

Based on Bhandari *et al.* (1966b) and Lal *et al.* (1968).

The fallout of the short-lived nuclides ³⁵S, ⁷Be, ³³P, and ³²P at latitudes 0–30° is in agreement with their production in the troposphere. The mean scavenging time for the removal of cosmogenic nuclides from the troposphere by dry and wet precipitation is about 30 days, independent of the latitude. This result has been derived using single isotopes as well as pairs of isotopes, such as ⁷Be and ³²P or ³³P and ³²P.

At latitudes of 30–40°, appreciable amounts of stratospheric air mix into the troposphere during spring months, leading to an appreciable enhancement in the concentrations of the nuclides in the air and in wet precipitation. The effect is quite appreciable, and is dramatically supported by observations of a host of fission products released in the stratosphere, which show marked increases in isotope concentrations in air and in wet precipitation at 30–40° latitudes. Stratospheric air builds up to near secular values of concentrations for short-lived isotopes (e.g., ⁷Be and ³²P), corresponding to the higher levels of production rates in the stratosphere (relative to troposphere). When this air descends into the troposphere, the ratios of isotope concentrations in air change dramatically at first for periods comparable to their half-lives (owing to preferential decay of the shorter-lived isotope of a pair), later reaching secular equilibrium values for the troposphere. See, for example, Figure 3, which shows the expected ratios for isotope pairs of ³⁵S, ⁷Be, ³³P, and ³²P.

The fairly extensive measurements of short-lived nuclei in the stratosphere and in the troposphere clearly show their half-life-dependent response to air circulation dynamics. Figure 4 and Figure 5 show measured concentration ratios of ⁷Be and ²²Na/⁷Be, respectively, in the atmosphere. The transport/mixing characteristics clearly show altitude and latitude dependence, as well as seasonal dependence. Detailed measurements of ³²P, ⁷Be, and ²²Na in the same air filters have been presented by Bhandari and colleagues.

Ratios of isotope concentrations are particularly valuable because they are independent of any uncertainties in the volume of air filtered. The observed degree of undersaturation in ²²Na concentrations indicates that the apparent irradiation age of stratospheric air ranges from a few months in the lower layer, just above the tropopause, to ~1–2 years at low latitudes and at altitudes of 18–20 km (Figure 5).

The expected fallout of nuclides that are scavenged by wet precipitation can be deduced fairly accurately from their

estimated production rates in the troposphere and stratosphere (Figure 6), and the observed fallout pattern of ⁹⁰Sr injected in the stratosphere. The resulting distribution is given in Figure 7 for a long-lived nuclide that which does not decay appreciably in the stratosphere.

Constructing Past Histories of Atmospheric Circulation

It seems quite logical to assume that the fallout pattern of cosmogenic isotopes that attach to aerosols would change with changes in the pattern of atmospheric circulation, which in turn would depend on past climates. The paleorecord of this change is partly contained in their concentrations in marine and continental sediments, and in polar ice sheets. This is all one has, to the best of our knowledge, but it is gratifying to see that this record is of a differential type and carries useful information on even high-frequency changes in the atmospheric circulation patterns. As seen from Figure 7, the fallout pattern is latitude-dependent, with varying relative contributions from the troposphere and the stratosphere. Consequently, it should be possible to extract some useful information on any observed paleochange in the fallout of nuclides.

Discussion

We are concerned here with the study of transport and mixing processes within and between the stratosphere and troposphere using tracers, with principal emphasis on cosmic ray-generated nuclides. The injection of radionuclide tracer into the atmosphere as a result of nuclear weapons tests provided fortuitous but valuable information about atmospheric circulation and aerosol scavenging processes, and this directly or indirectly launched similar observations using a variety of other tracers of natural and anthropogenic origin. These include water vapor, carbon dioxide, nitrous oxide, and anthropogenic trace gases: nuclides artificially produced as fission products (e.g., ⁹⁰Sr, and ⁸⁵Kr from weapons tests and due to release from nuclear reactors, respectively), and ¹⁴C produced by capture of weapons-produced neutrons by atmospheric N; and naturally produced radon (²²²Rn) and cosmogenic nuclides. Today we have several atmospheric transport models based on the use of observed distributions of the three categories of tracers to set up and tune three-dimensional models.

Studies of cosmogenic nuclides as atmospheric tracers began soon after the first observations of artificial radionuclides in the atmosphere produced by nuclear weapons tests. This was very fortunate, since aircraft- and balloon-borne programs that were designed specifically to measure fission products released from nuclear weapons tests provided suitable stratospheric and tropospheric air samples for the first measurements of ⁷Be, ³²P, and ²²Na in the atmosphere. These data immediately established the large potential of cosmogenic nuclide as atmospheric tracers.

The value of the cosmogenic radionuclides as tracers lies in the fact that their source functions are fairly well known. This, combined with the fact that several tracers of different chemical properties and half-lives are available, makes them ideally suited for studying spatial and temporal atmospheric

Table 5 Approximate steady-state fractional inventories of cosmic ray produced radioisotopes in exchange reservoirs^a

Exchange reservoir	Radioisotope													
	¹⁰ Be	²⁶ Al	³⁶ Cl	⁸¹ Kr	¹⁴ C	³² Si	³⁹ Ar	³ H	²² Na	³⁵ S	⁷ Be	³⁷ Ar	³³ P	³² P
Atmosphere	2.3×10^{-3}	1.4×10^{-6}	1.1×10^{-6}	0.96	1.9×10^{-2}	2.0×10^{-3}	0.99	7.2×10^{-2}	0.27	0.65	0.71	0.99	0.80	0.84
Land surface	0.29 ^b	0.29 ^b	0.29 ^b	0	4×10^{-2}	0.29 ^b	0	0.27	0.21	0.1	0.08	0	5.6×10^{-2}	4.7×10^{-2}
Mixed oceanic layer	5.7×10^{-6}	1.4×10^{-5}	1.4×10^{-2}	6×10^{-4}	2.2×10^{-2}	3.5×10^{-3}	0	0.35	0.44	0.24	0.20	0	0.13	0.11
Deep oceanic layer	10^{-4}	7×10^{-5}	0.69	3.5×10^{-2}	0.92	0.68	0.01	0.3	8×10^{-2}	4×10^{-3}	2×10^{-3}	0	7×10^{-4}	10^{-4}
Oceanic sediments	0.71	0.71	0	0	4×10^{-3}	2.8×10^{-2}	0	0	0	0	0	0	0	0
Half-life	1.5×10^6 y	7.1×10^5 y	3.0×10^5 y	2.3×10^5 y	5730 y	~ 150 y	268 y	12.3 y	2.6 y	87 d	53 d	35 d	25.3 d	14.3 d

^aApproximate calculations based on Lal (1966). Values given as zero imply very small fractional inventories.^bPart of the inventory may in fact be carried as silt or dust to the oceans before decay.

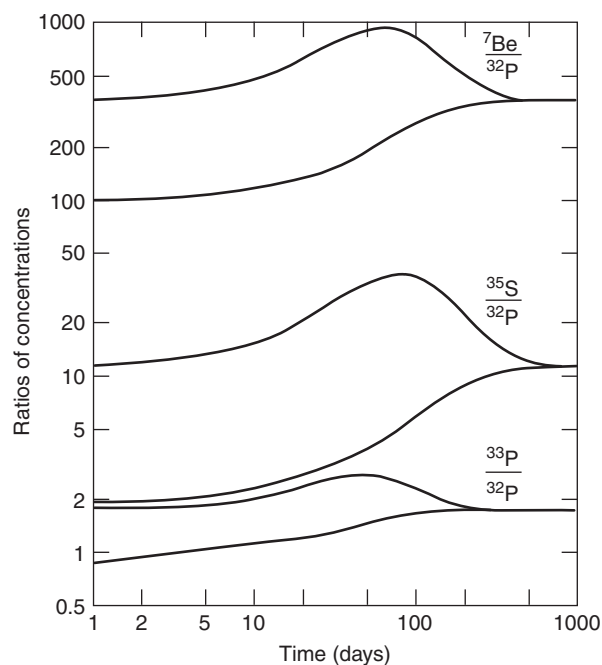


Figure 3 Expected ratios of concentrations (atoms g^{-1} air) of ^{35}S , ^7Be , ^{33}P , and ^{32}P for the hypothetical trajectories of stratospheric air in secular equilibrium, descending in the troposphere. For details see Lal (1966).

processes. Nevertheless, in some respects the transient tracers – those introduced in pulses as a result of test detonation of nuclear weapons – have certain unique advantages that the cosmogenic tracers lack. These include the ability to determine north–south mixing in the equatorial stratosphere, the resulting fallout pattern of nuclides in the troposphere once injected, say, into the equatorial stratosphere, the meridional transport of tracers within the troposphere in the Northern and Southern Hemispheres, and the cross-equatorial transport via the troposphere. The limitations on high accuracy using cosmogenic nuclides arises from the fact that cosmic ray nuclide source functions do not show much latitude dependence in these cases.

The value of transient tracers, is illustrated by two dramatic examples of studies of atmospheric transport rates based on them.

- From combined studies of ^{90}Sr and ^{14}C released in large nuclear weapons tests by the United States and USSR during late 1962, it became possible to determine quantitatively the structure of the troposphere with respect to mixing of air, rates of meridional transport within the two hemispheres, and cross-Equator transport/mixing of air between the two hemispheres. It was shown that the air within the $0\text{--}30^\circ$ and $30\text{--}90^\circ$ tropospheric cells is well mixed on periods of about 1 week, that the meridional mixing within the Northern Hemisphere exhibited a strong seasonal cycle with mean mixing times of 0.5–2.5 months, and that faster mixing occurred during the winter months. It was also determined that the cross-Equator mixing in the troposphere occurred on time scales of 8 ± 3 months.

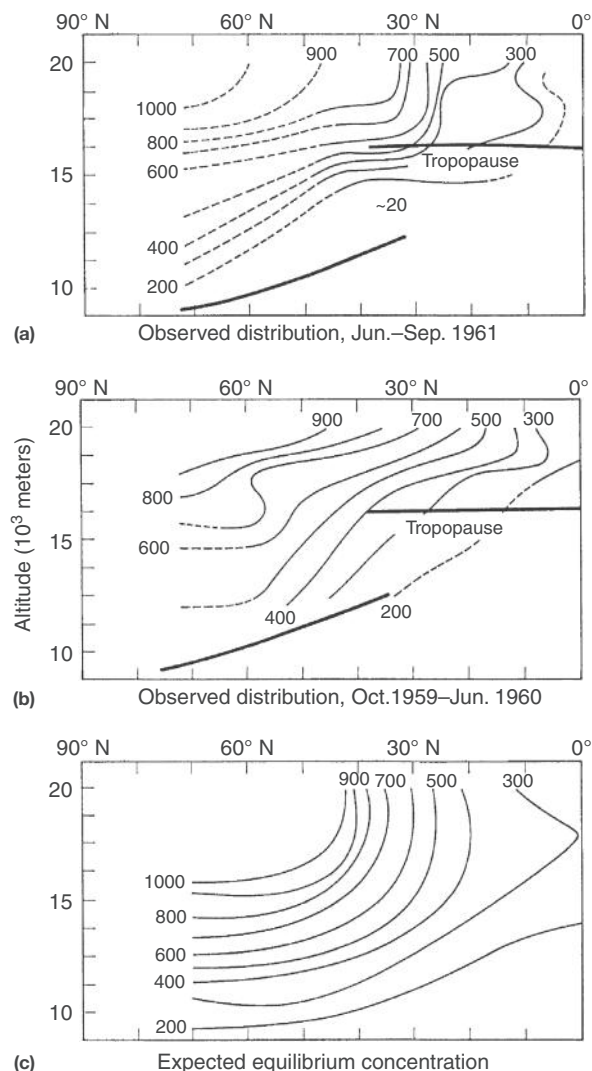


Figure 4 (a, b) Measured concentrations of ^7Be (in $\text{dpm}/28.3 \text{ m}^3$ STP), in air during two periods of observation in the Northern Hemisphere. The contour lines of equal production rates of ^7Be are given in (Figure 3(c)), (in $\text{dpm}/28.3 \text{ m}^3$ STP), corresponding to the expected ^7Be concentrations for a motionless atmosphere. Thick lines show the position of the tropopause. For details see Lal (1966) and Bhandari *et al.* (1966a).

- Analyses of atmospheric distribution of ^{85}Kr released primarily in the Northern Hemisphere by the nuclear industry suggested a mean time of 1.1 years for the inter-hemispheric exchange time in the troposphere, with little seasonal dependence.

To extract the full value of naturally produced or transient tracers, atmospheric tracers, it is clear that one has to integrate the information obtained from them and realize their limitations and potentials. To fully understand the dynamics of the atmosphere, including transports within and between the stratosphere and the troposphere, one has to integrate the tracer-based information with information on the nature of energy sources and sinks and their temporal and spatial

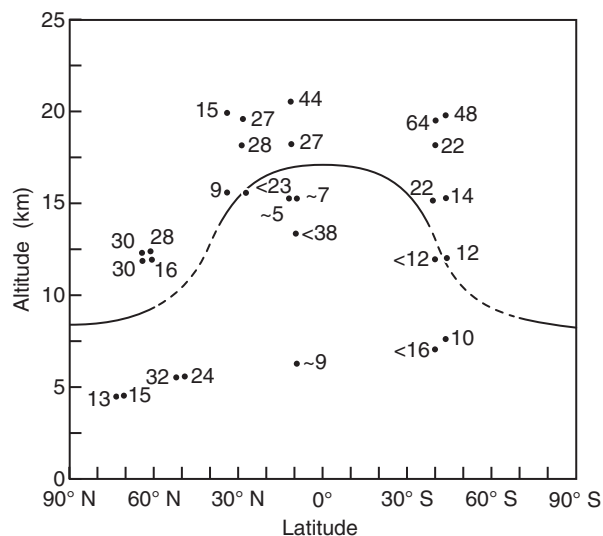


Figure 5 Measured ratios of activities of ^{22}Na and ^7Be in the atmosphere in a north–south section of the atmosphere. The numbers shown in the figure should be multiplied by 10^{-5} to obtain the absolute activity ratios.

variability. A meteorologist interested in atmospheric dynamics is concerned with numerous atmospheric motion-determining parameters – potential temperature, potential vorticity, generation of large-scale and small-scale waves, their propagation in the stratosphere and troposphere, and so on. Tracer geochemists look at all of this basic physics as locked within ‘black boxes’; they look only at the information on the net space- and time-averaged transport fluxes on the Earth. Clearly, attempts have to be made to bring the two approaches together; limited but successful efforts have been made, as documented in the Further Reading

Except for the detailed studies of several cosmogenic nuclides in the atmosphere during 1955–60, recent studies of ^7Be in the atmosphere, and some isolated studies of ^{36}Cl and ^{10}Be , no dedicated synoptic effort has yet been launched to study many cosmogenic nuclides in the air masses sampled. From the data available to date, however, their potential is clearly borne out for providing robust transport/mixing and scavenging parameters for developing robust three-dimensional atmospheric models. The task is easily feasible with modern techniques of air sampling and measurements of cosmogenic nuclides.

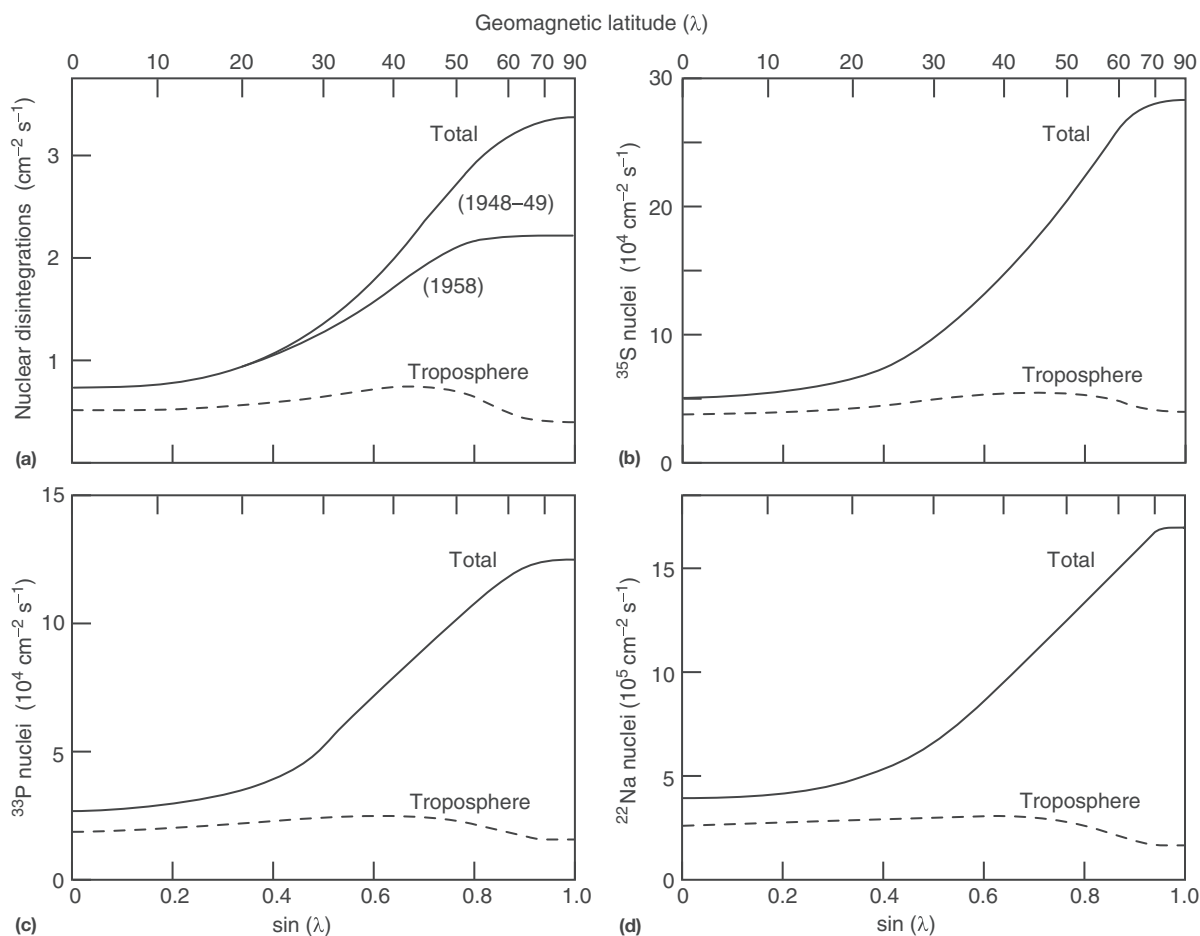


Figure 6 Production rates of nuclear disintegrations and ^{35}S , ^{33}P , and ^{22}Na in an air column with a cross-section of 1 cm^2 plotted against geomagnetic latitude. The lower curves in each figure show the amount produced in the troposphere, the upper curves the total production in the atmosphere. In (a), the dashed curve refers to the total production during 1958, a period of unusually high solar activity. All other curves refer to 1948–49, when solar activity was also quite high. Based on Lal (1966).

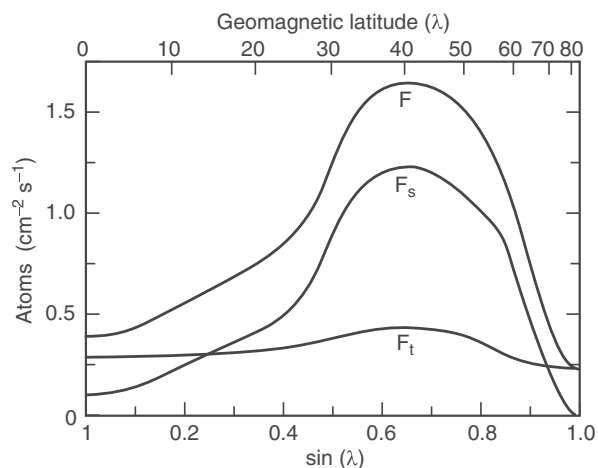


Figure 7 The expected pattern of fallout of a long-lived cosmogenic nuclide whose global average production rate is $1 \text{ atom cm}^{-2} \text{ s}^{-1}$ is given by the curve labeled F, curves F_s and F_t give the contributions to fallout from nuclide production in the stratosphere and troposphere, respectively. Based on Lal (1966).

See also: **Aerosols:** Aerosol Physics and Chemistry; Climatology of Tropospheric Aerosols; Observations and Measurements.

Chemistry of the Atmosphere: Ion Chemistry; Tracers. **Mesoscale**

Meteorology: Overview. **Numerical Models:** General Circulation

Models. **Paleoclimatology:** Ice Cores. **Stratosphere/Troposphere**

Exchange and Structure: Local Processes. **Stratospheric**

Chemistry Topics: Stratospheric Water Vapor. **Tropospheric**

Chemistry and Composition: Aerosols/Particles.

Further Reading

- Appenzeller, C., Holton, J.R., Rosenlof, K.H., 1996. Seasonal variation of mass transport across the troposphere. *Journal of Geophysical Research* 101, 15071–15078.
- Baumgartner, S., Beer, J., Suter, M., et al., 1997. Chlorine 36 fallout in the Summit Greenland Ice Core Project ice core. *Journal of Geophysical Research* 102, 26659–26662.
- Bhandari, N., Lal, D., Rama, 1966a. Stratospheric circulation studies based on natural and artificial radioactive trace elements. *Tellus* 18, 391–406.
- Bhandari, N., Bhat, S.G., Kharkar, D.P., Krishnaswamy, S., Lal, D., 1966b. Cosmic ray produced ^{28}Mg , ^{31}Si , ^{38}S , ^{38}Cl and ^{34}mCl and other short lived isotopes in wet precipitation. *Tellus* 18, 504–515.

- Brewer, A.W., 1949. Evidence for a world circulation provided by the measurements of helium and water vapor distribution in the stratosphere. *Quarterly Journal of the Royal Meteorological Society* 75, 351–363.
- Brost, R.A., Fleichter, J., Heimann, 1991. Three dimensional simulation of ^7Be in a global climate model. *Journal of Geophysical Research* 96, 22423–22445.
- Danielsen, E.F., Hipskind, R.S., Gaines, S.E., et al., 1987. Three dimensional analysis of potential vorticity associated with tropopause folds and observed variations of ozone and carbon monoxide. *Journal of Geophysical Research* 92, 2103–2111.
- Dentener, F., Fleichter, J., Jeuken, A.D., 1999. Simulation of the transport of ^{222}Rn using on-line and off-line global models at different horizontal resolutions: A detailed comparison with measurements. *Tellus* 51B, 573–602.
- Eluszkiewicz, J., 1996. A three dimensional view of the stratosphere-to-troposphere exchange in the GFDL SKYHI model. *Journal of Geophysical Research* 23, 2489–2492.
- Heimann, M., Keeling, C.D., 1989. A Three Dimensional Model of Atmospheric CO_2 Transport Based on Observed Winds. 2. Model Description and Simulated Tracer Experiments. American Geophysical Union Geophysical Monograph No. 55. ACOU, Washington DC. pp. 237–275.
- Holton, J.R., 1995. Stratosphere–troposphere exchange. *Reviews of Geophysics* 33, 403–439.
- Jacob, D.J., Prather, M.J., Wofsky, S.C., McElroy, B., 1987. Atmospheric distribution of ^{85}Kr simulated with a general circulation model. *Journal of Geophysical Research* 92, 6614–6626.
- Joseph, A.B., Gustafson, P.F., Russell, I.R., et al., 1971. Sources of radioactivity and their characteristics. In: *The Radioactivity in the Marine Environment*. National Academy of Sciences, Washington DC, pp. 6–41.
- Junge, C.E., 1963. *Air Chemistry and Radioactivity*. Academic Press, San Diego.
- Lal, D., Peters, B., 1967. Cosmic ray produced radioactivity on the earth. *Handbuch der Physik* 46/2, 551–612.
- Lal, D., Rama, 1966. Characteristics of global tropospheric mixing based on man-made ^{14}C , ^3H and ^{90}Sr . *Journal of Geophysical Research* 71, 2865–2874.
- Lal, D., Suess, H.E., 1968. The radioactivity of the atmosphere and hydrosphere. *Annual Review of Nuclear Science* 18, 407–434.
- Mahlman, J.D., 1997. Dynamics of transport processes in the upper troposphere. *Science* 276, 1079–1083.
- Prather, M., McElroy, M., Wofsky, S., Russell, G., Rind, D., 1987. Chemistry of the global troposphere: Fluorocarbons as tracers of air motion. *Journal of Geophysical Research* 100, 26141–26161.
- Rehfeld, S., Heimann, 1995. Three dimensional atmospheric transport simulation of the radioactive tracers ^{210}Pb , ^7Be , ^{10}Be , and ^{90}Sr . *Journal of Geophysical Research* 71, 2865–2874.
- Reiter, E.R., 1978. *Atmospheric Transport Processes. Radioactive Tracers*. US Department of Energy (TID-27114), Washington DC.
- Warneck, P., 1988. *Chemistry of the Natural Atmosphere*. Academic Press, San Diego.
- Wofsy, S.C., Cohen, R.C., Schmeltekopf, 1994. Overview: The stratospheric photochemistry aerosols and dynamic expedition (SPADE) and airborne arctic stratosphere expedition II (AASE-II). *Journal of Geophysical Research Letters* 21, 2535–2538.
- Wogman, N.A., Thomas, C.W., Cooper, J.A., Engelmann, R.J., Perkins, R.W., 1968. Cosmic ray-produced radionuclides as tracers of atmospheric precipitation processes. *Science* 159, 189–192.
- Yiou, F., Raisbeck, G.M., Baumgartner, S., et al., 1997. Beryllium 10 in the Greenland Ice Core Project ice core at Summit, Greenland. *Journal of Geophysical Research* 102, 26783–26794.

Volcanoes: Composition of Emissions

MT Coffey and JW Hannigan, National Center for Atmospheric Research, Boulder, CO, USA

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

The principal gaseous components of volcanic emissions are water, carbon dioxide, and sulfur (as SO₂, H₂S, and OCS). Only the sulfur gases are significant on a global scale when compared with other natural and anthropogenic sources and represent less than 10% of the input. However, sulfur compounds that reach the stratosphere, when combined with volcanic or meteoric dust particles, are the main precursors of stratospheric aerosols. Those aerosols play a critical role in the regulation of incoming solar radiation and consequently global warming.

Introduction

Volcanic emissions always have been important to the atmosphere of the Earth. Indeed, volcanic and seismic activities are believed to be the source of the initial atmosphere, beginning some 4.6 billion years ago. The principal gaseous components of the early volcanic emissions were, much as they are today, water (H₂O), carbon dioxide (CO₂), nitrogen (N₂), and sulfur (SO₂ and H₂S). The outgassed water condensed to form oceans. CO₂ and the sulfur gases dissolved in the oceans, leaving N₂ as the dominant gas in the atmosphere. Noticeably absent in the emissions is oxygen (O₂), which only appeared some 3.5 billion years ago as organisms developed that could convert CO₂ to organic carbon by photosynthesis thereby producing O₂.

Composition in Magma

Gases comprise approximately 1–4% of the weight of magma ejected by volcanoes. Measurements of the composition of magmatic gases vary widely. Table 1 shows the composition of gases in the magma of a number of major volcanoes. As seen in the table, there can be a wide range of composition, and these values should be considered rather uncertain. The last major eruption was by Mount Pinatubo in 1991. The recent eruption by Eyjafjallajökull in Iceland in April 2010, which was well observed and caused considerable disruption of air travel, was more than an order of magnitude less energetic than the Mount St. Helens (1980) eruption, the least energetic eruption listed in Table 1. Volcanic gas is predominantly water. The second most abundant constituent is CO₂ followed by sulfur gases (SO₂ and H₂S) and hydrogen halides (HBr, HCl, and HF). Smaller

amounts of carbon monoxide (CO), carbonyl sulfide (COS), ammonia (NH₃), hydrogen (H₂), nitrogen (N₂), or other trace gases also may be present.

These gases may react with one another within the magma and the relative composition may change as the magma ascends toward the surface, with an associated change of temperature and pressure. The composition of magmatic gases may also depend on the geologic or tectonic setting of the volcano. Gases emitted from volcanoes at the junction of converging plates have proportionately more H₂O and Cl than gases from divergent plate and hot spot volcanoes. When plates converge, significant amounts of sea water and oceanic crust, rich in H₂O and Cl, can be incorporated into the magma. Hot spots are persistent (10⁷ years or more) areas of hot mantle extrusion at the surface, which may or may not be associated with a plate boundary. The Hawaiian volcanoes are examples of hot spots. The composition of gases emitted by volcanoes may be further modified by interactions near the surface with groundwater systems or local geology.

Clearly, a number of factors may affect the composition of gaseous volcanic emissions, many of which are not well understood. Each volcano must be observed to determine the individual composition of its output.

Volcanic Contribution to Global Sources

Emissions from volcanoes may take the form of a slow, continuous seeping of gas from a vent or fumarole or of an energetic but sporadic eruption. Of the gases listed in Table 1, only the sulfur gases represent a significant contribution by volcanoes to the global budget on an annual basis. It is

Table 1 Gaseous composition of volcanic magma

Volcano year	Magma type	Mass of ejected magma (kg)	Gas composition (volume %)			
			H ₂ O	CO ₂	SO ₂	HCl
Pele 1902	Andesite		80.0	9.0	1.0	0.6
Kilauea 1917	Basalt		79.3	11.6	6.5	0.05
Nyiragongo 1959	Basalt		43.2	40.9	4.4	
Agung 1963	Andesite	2.4 × 10 ¹²	90.0		8.0	1.0
Surtsey 1964	Basalt	2.8 × 10 ¹²	86.1	5.7	2.6	1.0
Mount St. Helens 1980	Andesite	6.5 × 10 ¹¹	91.6	6.6	0.6	
El Chichon 1982	Andesite	1.0 × 10 ¹²	84.0		10.0	1.0
Mount Pinatubo 1991	Andesite	1.0 × 10 ¹⁶	81.6	9.2	4.6	0.3

Table 2 Volcanic contribution to global sources

Gas	Annual volcanic source	Annual natural source (nonvolcanic)	Annual anthropogenic source	Units
H ₂ O	0.0006	423	0.045	Eg H ₂ O yr ⁻¹
CO ₂	0.2	220	35	Pg C yr ⁻¹
H ₂ S, SO ₂	28	229	113	Tg S yr ⁻¹
HCl	0.7	38	6	Tg Cl yr ⁻¹
Hg	0.10	5.2	2.3	Gg Hg yr ⁻¹

estimated that emissions from fumaroles of major volcanoes during quiescent periods may account for more than 10% of the natural sulfur emission. For the other gases, annual emissions are insignificant when compared to other sources, as may be seen in Table 2. In a given year, sulfur emissions during eruptive periods are estimated to contribute less than 1% to the annual, global budget (mostly as H₂S and SO₂). Particular, energetic, sporadic eruptions can, however, represent a significant portion of the annual source for a limited period. The 1991 eruption of Mount Pinatubo was observed to inject a mass of approximately 20 Tg of SO₂ into the atmosphere, approximately 6% of the global annual source for that year. The annual volcanic emission of CO₂ is 100–150 times less than that from human activities.

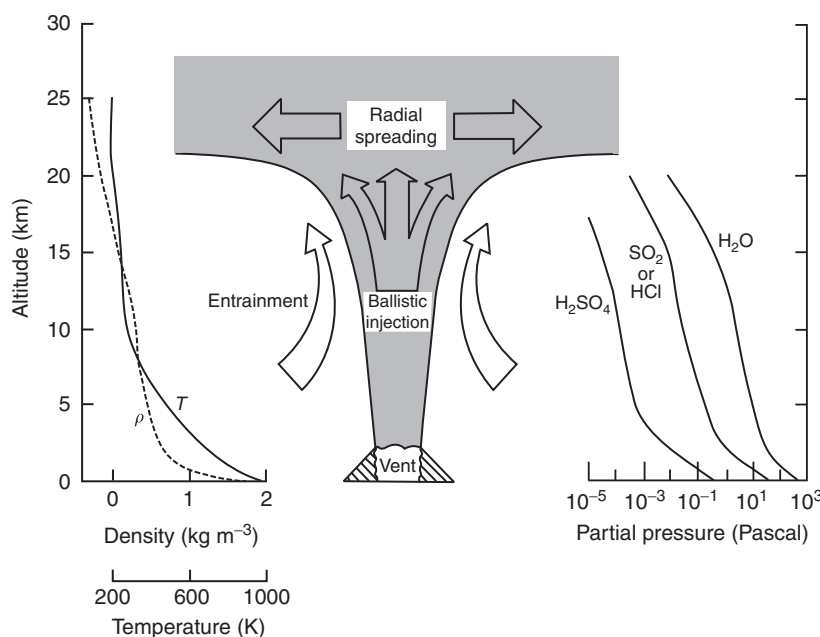
Eruption Dynamics

The form of a volcanic emission is important in assessing its potential impact on the atmosphere. Gases that are slowly vented at the surface may be efficiently oxidized or removed by rainout. H₂S reacts readily with tropospheric OH and then O₂ to produce SO₂, which eventually ends up as sulfuric acid (H₂SO₄); this is removed in the lower atmosphere by rain. High concentrations of SO₂ may substantially reduce the

lifetime of tropospheric OH, a key reactant in many chemical cycles. HBr contained in volcanic emissions can lead to significant bromine-catalyzed destruction of ozone and mercury within a plume. Hydrochloric acid is very soluble in water and is efficiently removed by rain in the troposphere. Thus, vent emissions may acidify the regional rainfall or have a local polluting effect but have much less consequence on the global scale. Figure 1 shows a diagram of an energetic volcanic eruption. In an energetic eruption, magma and gases are discharged from the vent at high velocity (up to 400 m s⁻¹). Entrainment and heating of the surrounding air cause the plume to rise under the effect of buoyancy. At higher altitudes, the plume reaches a neutral buoyancy level and spreads to form an umbrella-shaped cloud. The maximum height attained by the plume depends largely on the thermal flux at the vent, vent geometry, stratification and moisture content of the atmosphere, and volatile content of the magma.

If the form of the emission is sufficiently explosive and has the appropriate geometry, large quantities of gas and particles may be injected directly into the stratosphere, bypassing tropospheric oxidation and rainout. Once in the stratosphere, the volcanic effluent becomes a global feature due to the rapid east-to-west transport of the region and lack of removal mechanisms. Clear examples of such global volcanic events were the eruptions of El Chichon in 1982 and of Mount Pinatubo in 1991, which each injected large quantities of gas and particles into the stratosphere. Within a few weeks, the plumes of these eruptions encircled the globe and evidence of their effects persisted for many years.

Figure 2 shows the integrated backscatter from a lidar operated at NASA Langley Research Center for the period from 1974 to 2002 and more recent aerosol optical depth observations from three satellite-borne instruments. The backscattered and optical depth signals are an indication of the extent of stratospheric aerosol amount. Signatures of some of the more

**Figure 1** Schematic diagram of an energetic volcanic eruption.

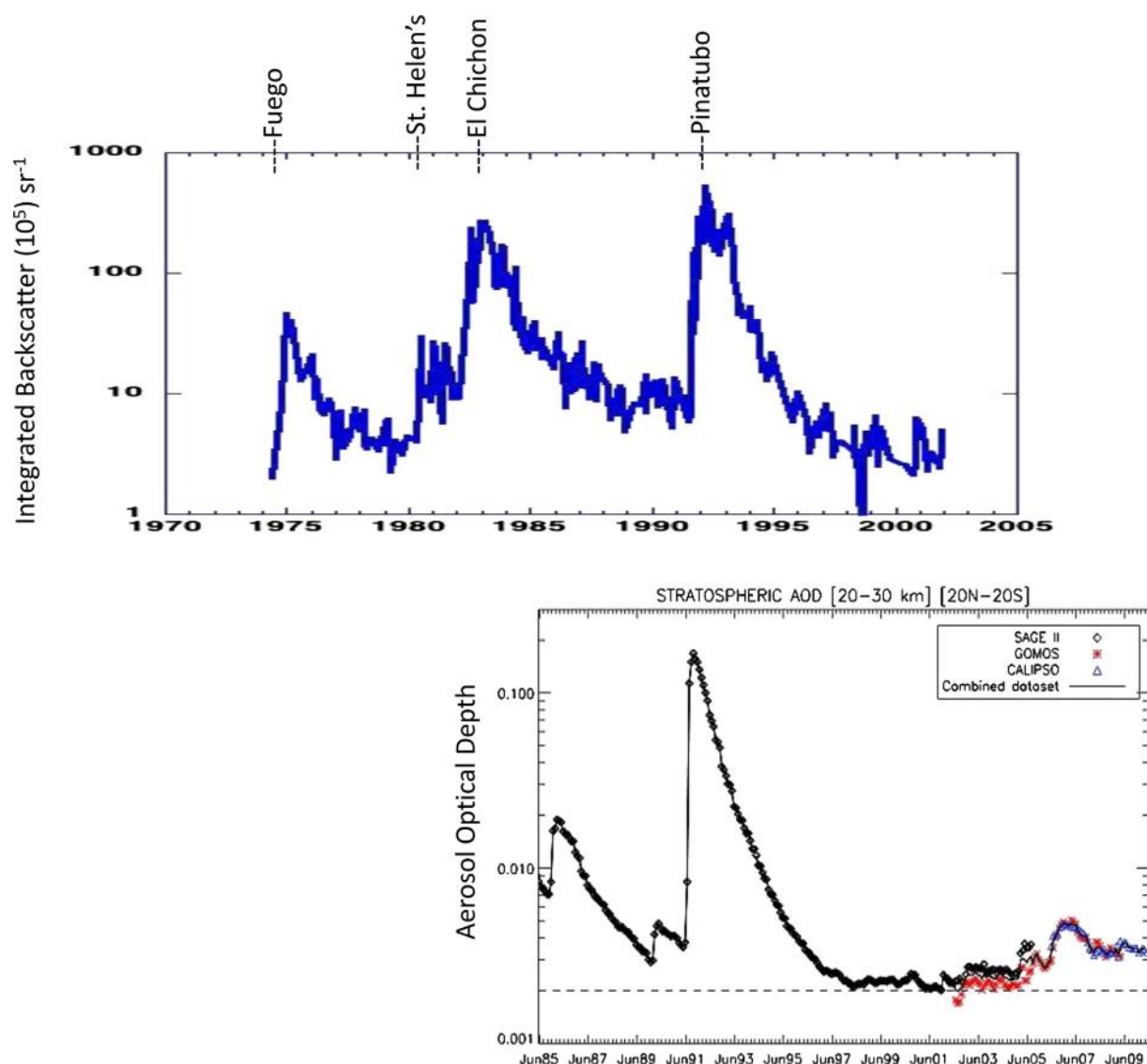


Figure 2 Aerosol Integrated Backscatter from the 48" LIDAR at NASA Langley Research Center (LaRC Atmospheric Science Data Center) and Aerosol Optical Depth from SAGE II, GOMOS, and CALIPSO observations. Vernier J.P. et al., *Geophysical Research Letters*, 38, 10.1029/2011GL047563. Copyright 2011 American Geophysical Union. Reproduced by permission of the American Geophysical Union.

recent major volcanic eruptions may be seen in the record. After a period of having a stratosphere as free of aerosols as has occurred since the beginning of these measurements, there now appears to be a slow increase in stratospheric aerosols due possibly to emissions by a number of moderately energetic volcanoes.

Stratospheric Effects

The most important atmospheric effect of volcanoes is probably the introduction of aerosols to the stratosphere. Aerosols are produced by the reaction of gaseous H_2S and SO_2 with water to form aqueous sulfuric acid (H_2SO_4). Sulfuric acid condenses onto solid particles to form stratospheric aerosols. Energetic eruptions also may inject large amounts of solid particles into the stratosphere. These particles are composed

mostly of silicate dust and ash (less than 2 mm diameter). Larger particles fall out of the atmosphere relatively near the vent; smaller particles may remain in the stratosphere for many months as may be seen in Figure 2.

Volcanic aerosols can affect the radiative heating and cooling in the atmosphere. Introduction of H_2SO_4 aerosols into the stratosphere will increase the albedo of the Earth, which is essentially the reflectivity for incoming solar radiation. Sulfate aerosols are efficient scatterers but only weak absorbers at solar wavelengths. Increasing the proportion of solar radiation reflected back to space should cool the atmosphere. The effectiveness of the sulfate aerosols to heat the stratosphere by absorbing outgoing infrared radiation depends on the size of the aerosol. Observations after the eruptions of Agung, El Chichon, and Mount Pinatubo have shown that the lower stratosphere (16–20 km) was warmed by 1–2 K due to the

presence of volcanic aerosols. This topic is discussed further in the article **Climate and Climate Change: Volcanoes: Role in Climate**.

A second important role of volcanic aerosols that are injected into the stratosphere is to serve as sites for heterogeneous reactions which otherwise would not occur. Heterogeneous reactions on sulfate or water aerosols can release chlorine from reservoir species (such as HCl and ClONO₂) and can convert reactive nitrogen species (NO, NO₂, and N₂O₅) into the more stable HNO₃ reservoir. Both these changes enhance the destruction of ozone.

See also: **Climate and Climate Change: Volcanoes: Role in Climate**.

Further Reading

- Delmelle, P., Stix, 2000. Volcanic gases. In: Sigurdsson, H. (Ed.), *Encyclopedia of Volcanoes*. Academic Press, New York, pp. 803–815.
- Stoiber, R.E., 1995. Volcanic gases from subaerial volcanoes on Earth. In: Ahrens, T.J. (Ed.), *Global Earth Physics*. American Geophysical Union, Washington, DC, pp. 308–319.

Tracers

KA Boering, University of California – Berkeley, Berkeley, CA, USA

© 2003 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 6, pp 2297–2305, © 2003, Elsevier Ltd.

Introduction

The topic of ‘tracers’ is vast, with a rich history that has contributed much to our understanding of the chemistry and dynamics of the atmosphere. A tracer, in the context of atmospheric science, generally refers to a measurable atmospheric species or a parameter that allows one to deduce or infer atmospheric motions and transport, atmospheric chemical reaction pathways, or the magnitude of biogeochemical processes and their influence on the composition of the atmosphere and, hence, on climate. A number of specific tracers are already covered in detail in individual articles in the encyclopedia. The goal of the discussion here is to provide a general overview of the nature and use of measurements of chemical tracers (i.e., chemical species, as opposed to meteorological parameters such as winds, potential temperature, or potential vorticity) to infer the atmospheric circulation and transport of chemical species within and between the stratosphere and troposphere. In many cases, once the influence of transport on chemical species is accounted for, the associated chemical and biogeochemical processes can then be studied.

An Historic Example: The Brewer–Dobson Circulation

A particularly noteworthy and illustrative application of chemical tracers to our fundamental understanding of the circulation of the atmosphere is that of measurements and analysis of the total column ozone by Dobson in the 1920s and 1930s and of stratospheric water vapor and helium by Brewer in the 1940s. Dobson observed that the total amount of ozone measured in the atmospheric column above his ground-based spectrophotometers showed maxima at high latitudes and minima at low latitudes. This dependence on latitude was exactly the opposite of what was expected on the basis of atmospheric chemistry alone, since ozone production is greatest in the upper tropical stratosphere where solar irradiance is most intense. From these observations, Dobson deduced that there must be a global-scale circulation of air from the tropics to high latitudes resulting in the meridional transport of ozone from the tropics to the poles. From the tracers helium and water vapor, Brewer added important vertical information to this picture. The fact that the helium mixing ratio did not change with altitude meant that turbulent transport acted on the stratosphere to prevent gravitational mass fractionation. His observations of very low water vapor mixing ratios in the stratosphere (while studying the formation of condensation trails from aircraft exhaust for defense purposes during World War II) were consistent only with upward transport of air into the stratosphere in the tropics (where the tropopause is high and therefore cold enough to ‘freeze dry’ air as it passed into the stratosphere) and

a downward transport at midlatitudes (where the tropopause is lower and warmer; downwelling would prevent the upward diffusion of wet tropospheric air across the warm extratropical tropopause, which could not freeze dry the air to the observed low stratospheric water vapor levels). This overall picture, derived from chemical tracers, of upwelling in the tropics, meridional transport from the tropics to the poles, and downwelling in the extratropics is known as the Brewer–Dobson circulation. It remains qualitatively correct today. Scientific progress continues through iteration between tracer measurements, dynamical analyses, and chemical–radiative-transport computer models to quantify these rates and to understand the mechanisms driving them. Some of these ‘newer’ efforts using chemical tracers are outlined in the following sections.

Tracers in the Stratosphere

The magnitudes of the mean vertical and meridional winds in the stratosphere are quite small – both in an absolute sense (e.g., $<0.5 \times 10^{-3} \text{ m s}^{-1}$ in the vertical) and in relation to the much faster zonal winds (i.e., around a latitude circle). Thus, they are difficult to measure directly, and tracers must be used to infer these rates and the transport of air that results. These same wind characteristics also make artificial tracer release studies very difficult, as the released material would simply be rapidly redistributed and diluted zonally, yielding little information on vertical and meridional transport, particularly on the time scales of a conceivable tracer release experiment. Thus, we rely on various serendipitous natural or anthropogenic tracers and their time-varying and/or spatially varying characteristics in the troposphere or stratosphere to deduce mean vertical and meridional winds, to understand the efficiency of transport of stratospheric species within and between different regions, such as between the midlatitudes and the polar vortex, for example, and to decouple the effects of transport from the effects of local photochemistry on ozone, for example, once transport has been inferred.

The Morphology and Correlations of Tracer Mixing Ratios in the Stratosphere: CH_4 , N_2O , and CO_2

Methane (CH_4) and nitrous oxide (N_2O) are produced primarily through biological processes in the troposphere and are destroyed by photolysis and photochemistry in the stratosphere on time scales from many years to months, depending on altitude. Observations from satellites have provided a global-scale picture of the distribution of their mixing ratios throughout the stratosphere (Figure 1). The morphologies of their mixing ratio contours share a common shape. The common shape can be understood qualitatively in terms of the

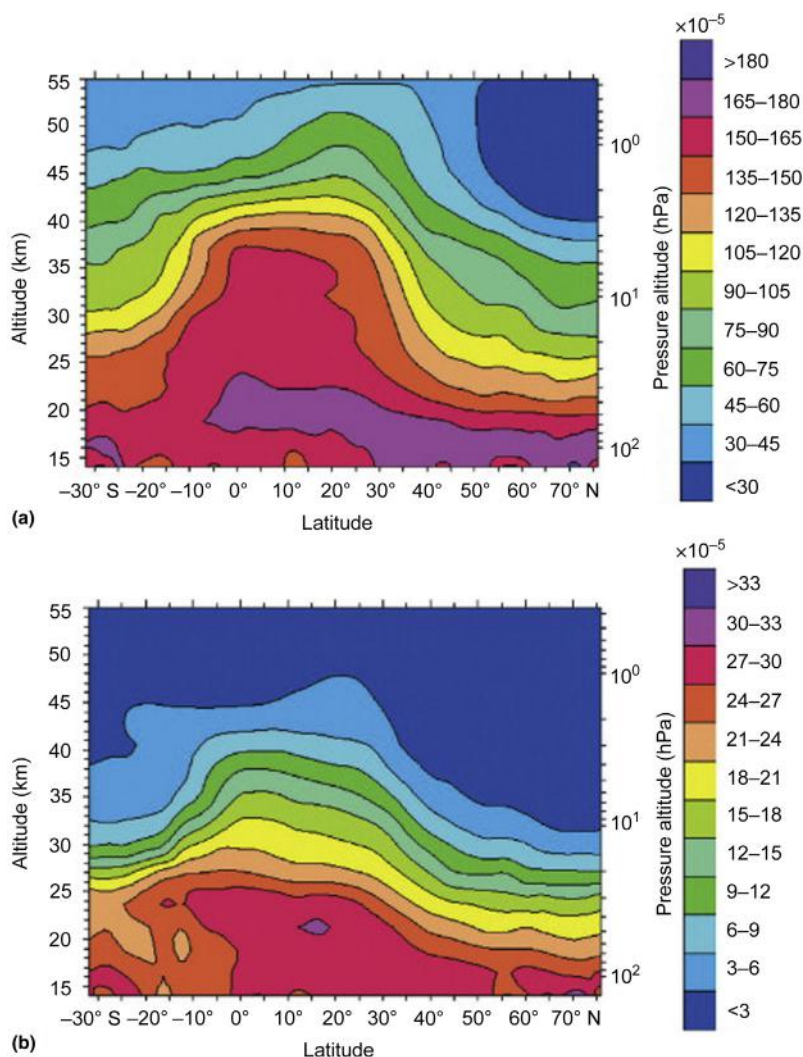


Figure 1 Stratospheric mixing ratios of (a) CH_4 and (b) N_2O versus latitude and pressure–altitude measured on 25 September 1992 from the CLAES instrument aboard the Upper Atmosphere Research Satellite. Adapted with permission from Roche, A.E., et al., 1996. Validation of CH_4 and N_2O measurements from the CLAES instrument on UARS. *Journal of Geophysical Research* 101 (D6), 9679–9710. Copyright 1996 American Geophysical Union.

Brewer–Dobson circulation combined with rapid quasi-horizontal transport and mixing of air induced by the breaking of planetary-scale waves at midlatitudes, as illustrated in Figure 2. Upwelling of air that recently entered the stratosphere from the troposphere in the tropics (with correspondingly high values of CH_4 and N_2O) couples with downwelling of older, photochemically processed air from the upper stratosphere (with low levels of CH_4 and N_2O) at middle and high latitudes to form a ‘bulge’ of high mixing ratio contours (also known as ‘isopleths’) in the tropics. Rapid quasi-horizontal transport and mixing of air between the tropics and extratropics along isentropic surfaces serves to flatten this bulge. Thus, the slope of the isopleths at midlatitudes is controlled by the relative rates of vertical advection (which steepens the isopleths) and quasi-horizontal diffusion (which flattens the isopleths). Relatively rapid changes in slope with latitude separate regions where the dominant transport mechanisms are different, such as between the tropics and the

midlatitudes (where the so-called subtropical boundary indicates a transition between upwelling in the tropics and rapid quasi-horizontal transport in the so-called midlatitude ‘surf zone’) and between the midlatitude surf zone and the polar vortices (where transport across the wind maximum at the vortex edge is limited). Consequently, the magnitudes of the isopleth slopes and their dependence on latitude and how these properties change with season, latitude, and altitude provide important information on transport and details of the underlying dynamics, as well as diagnostics for models.

In addition, theory dictates that any chemical tracers that are long-lived with respect to both vertical and horizontal transports in a region of the stratosphere will share a common morphology in that region. Thus, a plot of simultaneous measurements of one long-lived tracer versus another will yield what is known as a ‘compact’ or tight relationship for which one tracer mixing ratio is well predicted if the other is known. The species are said to be in slope equilibrium. If horizontal

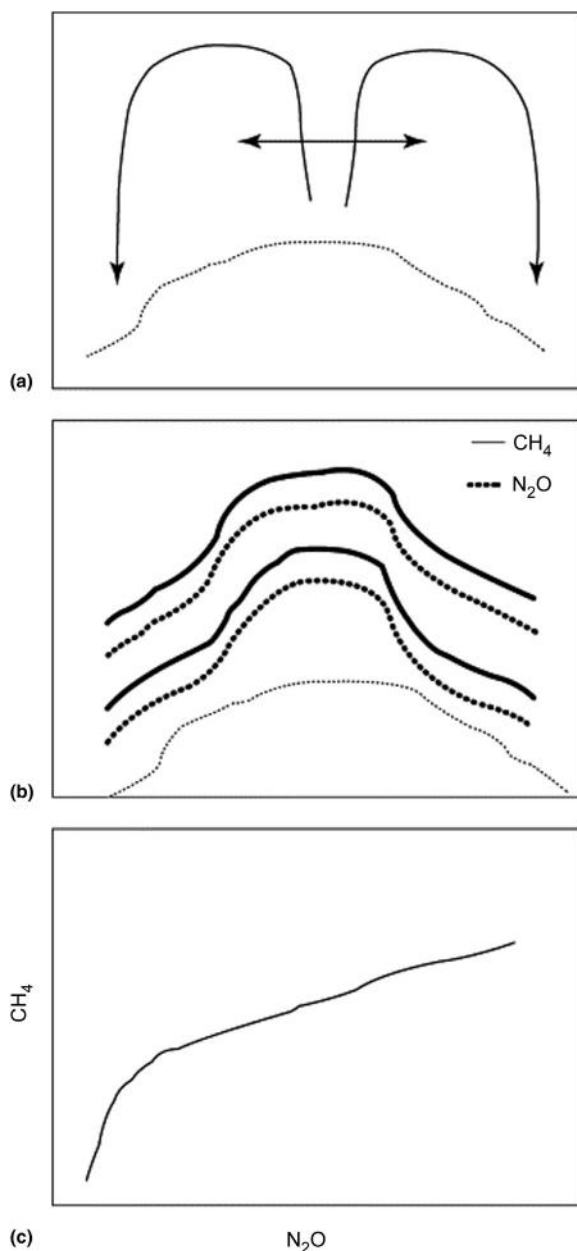


Figure 2 Schematic representations of (a) stratospheric transport for which quasi-horizontal transport (diffusion) is globally effective, (b) the resulting tracer isopleths (surfaces of constant mixing ratio), and (c) the resulting tracer–tracer correlation for tracers that are long-lived with respect to both vertical and horizontal transport rates.

transport is fast and effective globally, then the long-lived tracer isopleths and the tracer–tracer relationships are globally uniform (Figure 2). If horizontal transport from the midlatitudes into the tropics is restricted, the isopleths and the resulting tracer–tracer correlations are different in these different regions (Figure 3). Simultaneous observations of tracer–tracer correlations measured from satellite, balloon, and aircraft instruments in the 1990s (including CH_4 , N_2O , CFCs, O_3 , NO_x , and CO_2 measurements) revealed that the real atmosphere lies somewhere between the two conceptual

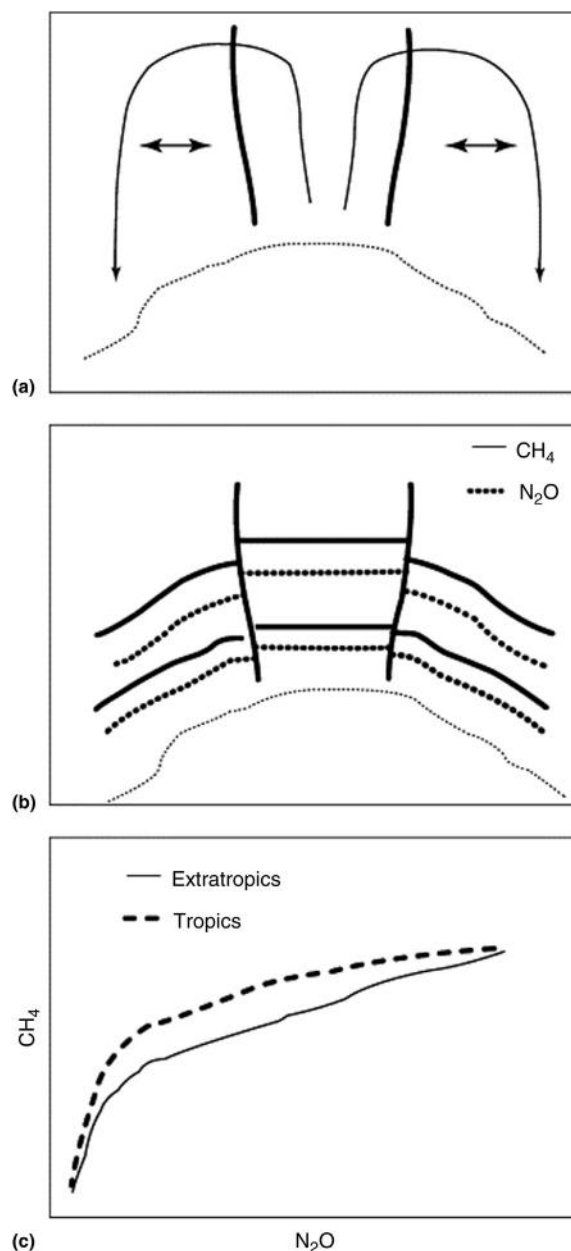


Figure 3 Schematic representations of (a) stratospheric transport for which quasi-horizontal transport from the extratropics into the tropics is slow, (b) the resulting tracer isopleths, and (c) the resulting tracer–tracer correlations.

extremes in Figures 2 and 3. Observed tracer correlations are indeed compact but different in the tropics and extratropics, demonstrating that horizontal transport from midlatitudes into and across the tropics is not rapid enough to result in globally uniform correlations. However, analyses of the correlations indicate that a significant amount of older air from midlatitudes does mix back into the tropical upwelling region at altitudes between the tropopause and about 21 km. Besides being of fundamental interest, knowledge of the degree to which midlatitude air ‘recirculates’ into the tropics is critical for

predicting pollutant build-up in the stratosphere and its redistribution, such as exhaust from a future fleet of commercial supersonic aircraft that would fly at about 20 km mostly in the midlatitudes.

Plotting one long-lived tracer against another also serves another important function: it removes the variability in stratospheric tracer measurements associated with large-scale reversible air displacements and, to some extent, with irreversible mixing. The spatial variability these processes induce on tracer mixing ratios can make it difficult to interpret tracer measurements along an aircraft flight track, for example, and to compare them with numerical model predictions. Simultaneous measurements of CO_2 and N_2O from the NASA ER-2 aircraft illustrate both this simplification and the determination of relative transport rates discussed above. CO_2 is a conserved tracer in the stratosphere; its mixing ratio in air entering the tropical stratosphere exhibits both a long-term increase due to anthropogenic activities and an annual cycle due to the seasonal uptake and release of carbon by the terrestrial biosphere (Figure 4). These variations forced from the troposphere propagate into the stratosphere. In the tropics, the vertical propagation of the maxima and minima of the CO_2 seasonal cycle from the tropopause is easily discernible when the observed CO_2 mixing ratio is plotted against potential temperature (Figure 5(a)). As discussed further below, the preservation of the seasonal cycle in the

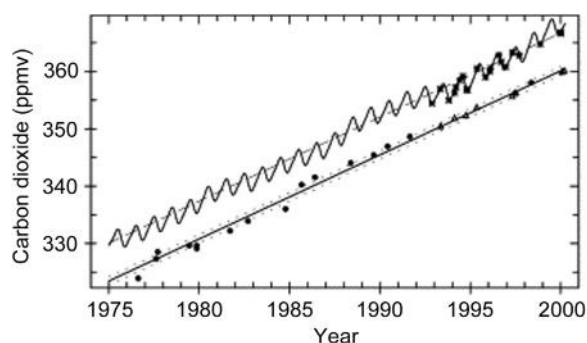


Figure 4 Observations of stratospheric CO_2 . Asterisks are measured CO_2 mixing ratios at the tropical tropopause or in air that has recently entered the stratosphere based on simultaneous measurements of other tracers. Upper curve: Continuous boundary condition for CO_2 mixing ratios entering the stratosphere based on surface measurements from the NOAA Climate Monitoring and Diagnostics Laboratory Cooperative Air Sampling Network, which accounts for seasonal and interannual variations in the CO_2 growth rate. The long-dashed line is a linear fit to the continuous boundary condition. Lower curve (see text): The solid line has the same slope as the linearized boundary condition above but with a time lag of 4.5 years; short dashed lines correspond to delays of 4 and 5 years. Filled circles are balloon measurements of CO_2 made in the mid-latitude stratosphere between about 20–25 and 30 km (a region of near-constant CO_2 mixing ratio) and triangles are aircraft measurements corresponding to N_2O values of 110 ppb, revealing a consistency in mean ages in this region of the stratosphere that dates back to the 1970s. Adapted from Andrews A.E., et al., 2001. Mean ages of stratospheric air from in situ observations of CO_2 , CH_4 , and N_2O . *Journal of Geophysical Research* 106 (D23), 32295–32314. Copyright 2001 American Geophysical Union.

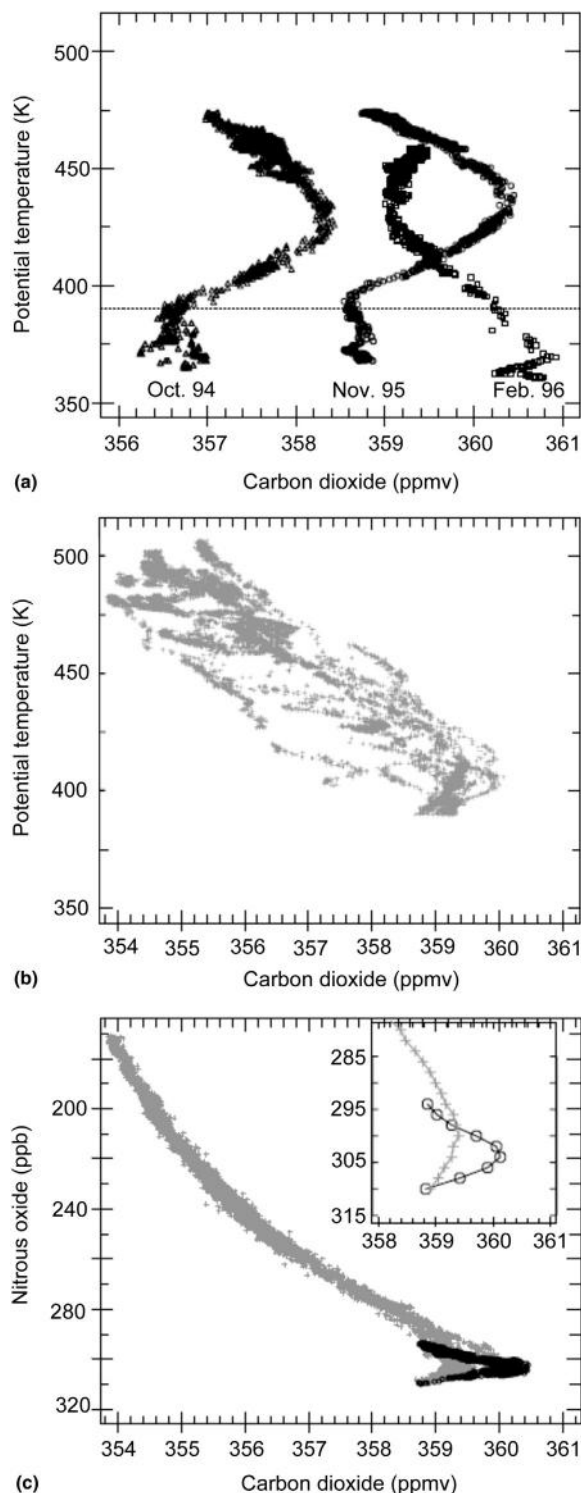


Figure 5 (a) Observations of CO_2 from the NASA ER-2 aircraft in the tropics versus potential temperature. The tropopause is at 390 K. (b) CO_2 versus potential temperature in the extratropics in November 1995. (c) CO_2 versus N_2O in the extratropics (gray) and the tropics (black) in November 1995; inset are the CO_2 observations binned and averaged as a function of N_2O in these two regions. Reprinted with permission from Boering K.A., et al., 1996. Stratospheric mean ages and transport rates from observations of carbon dioxide and nitrous oxide. *Science* 274, 1340–1343. Copyright 1996 American Association for the Advancement of Science.

tropics indicates that vertical advection dominates both vertical diffusion and the mixing in of older air from the extratropics in this region of the stratosphere. At midlatitudes, however, a plot of CO₂ mixing ratio against potential temperature is highly scattered (Figure 5(b)). Plotting CO₂ versus another long-lived tracer, such as N₂O, gives a striking compact relationship (Figure 5(c)). From this compact relationship, we can infer that quasi-horizontal transport in the extratropics is rapid compared to the chemical lifetime of N₂O and to the rate of seasonal changes in the CO₂ mixing ratio at the tropical tropopause. Moreover, because the CO₂:N₂O tracer correlation in the tropics is distinguishable from that in the extratropics (Figure 5(c)), quasi-horizontal transport across the tropics is not efficient enough to result in a globally uniform CO₂:N₂O relationship. The observed attenuation of the CO₂ seasonal cycle in the tropics, however, does indicate that a significant amount of extratropical air has been mixed into the tropical upwelling region, with 30–50% of the air at 19 km of midlatitude origin.

Temporally Increasing Inert Tracers: CO₂ and SF₆

Species that are increasing over time in the troposphere due to anthropogenic activities but that are inert in the stratosphere have been used to infer the mean age of air at various locations and times in the stratosphere. The mean age is the average time since the air for which CO₂ or SF₆ is measured was in contact with the troposphere. A boundary condition for CO₂ or SF₆ mixing ratios at the tropical tropopause, where air predominantly enters the stratosphere, is needed to derive accurate mean ages (e.g., Figure 4). The mean age of an air parcel can be approximated from the time delay between the mixing ratio observed (or inferred) at the tropopause and the mixing ratio measured in the air parcel. From a series of measurements during the 1990s from aircraft and balloon instruments, precise and accurate mean ages have been derived for air over a wide range of latitudes and altitudes (± 0.3 and ± 0.5 years, respectively, for ages derived from CO₂ measurements once the seasonal cycle has damped out at a mean age of ~ 2 years). These mean ages are also consistent with those derived from balloon observations of CO₂ dating back to the 1970s now that a robust boundary condition for CO₂ entering the stratosphere, based on the extensive NASA ER-2 CO₂ observations, can be parameterized from surface CO₂ measurements for those years (e.g., Figure 4). As expected from the Brewer–Dobson circulation, the youngest air is observed just above the tropical tropopause, with age increasing with altitude and latitude to a maximum of about 6 years. Knowledge of the mean ages helps define relative vertical and horizontal transport rates in different regions of the atmosphere and the turnover time for the stratosphere with respect to exchange of air with the troposphere. Mean ages from inert tracers are also serving as quantitative tests of transport, independent of photochemistry, for two- and three-dimensional models of the stratosphere. As of 1998 when the mean ages inferred from CO₂ and SF₆ observations were well-established, most two- and three-dimensional models were underestimating mean ages by 20–100%.

Seasonally Varying Tracers: H₂O, CO₂

Species that enter the stratosphere with a pronounced annual cycle in their mixing ratios have been used to infer vertical ascent rates in the tropics, vertical diffusion in the tropics, and the rates at which air is transported quasi-horizontally between the tropics and midlatitudes. Observations of CO₂ were discussed in the preceding section. Water vapor mixing ratios entering the stratosphere also exhibit an annual cycle, which is in phase with the annual cycle in tropopause temperatures. Exactly what dynamical and/or microphysical processes control the saturation mixing ratios in the tropopause region, however, is still hotly debated; apparently, Brewer's simple freeze-drying mechanism at the tropopause is not sufficient. From observations of the vertical propagation of the seasonal cycles of CO₂ and water vapor from the tropopause (e.g., Figure 5(a)), vertical ascent rates can be derived. The observed attenuations of the seasonal amplitudes with altitude effectively give the rate of horizontal transport of older air from midlatitudes (since vertical diffusion is known to be small from other observations), as noted above for CO₂. Observations of seasonal variations at midlatitudes yield information on quasi-horizontal transport rates out of the tropics. Continuous monitoring of stratospheric water vapor from the tropopause to the stratopause has been achieved through satellite measurements that have a vertical resolution of 1–3 km. The satellite observations show the propagation of the water vapor seasonal cycle to about 28 km before it is damped out. The *in situ* aircraft CO₂ observations are not continuous in time but have high vertical resolution (~ 10 m) in the near-tropopause region between 16 and 20 km where air is entering the stratosphere and the mixing in of midlatitude air is the most rapid. Thus, the CO₂ and H₂O measurements are highly complementary. Interestingly, these two tracers both yield relatively precise seasonally resolved vertical ascent rates that are nearly identical to those derived from calculations of the meridional circulation estimated from heating rates computed with radiative transfer models using satellite and/or climatological data for species active in the infrared. The calculations of radiative heating in the lower tropical stratosphere have large uncertainties because the net heating rates are small, being the difference between two large numbers.

Anthropogenic 'Pulsed' Tracers: Bomb Debris from Atmospheric Nuclear Testing

It was noted above that the release of artificial tracers to study stratospheric transport is not readily feasible, in part owing to the prohibitive magnitude of material that would have to be released. However, the massive amount of material injected into or produced in the stratosphere by atmospheric nuclear testing in the late 1950s and early 1960s served as a 'pulsed' input of radioactive tracers into the stratosphere. Transport within and out of the stratosphere can be followed from observations of tritiated water vapor (HTO), ¹⁴CO₂, and fission products, such as ⁹⁰Sr, that were made from aircraft and balloons from that time until as recently as 1983 for HTO. In particular, observations over a number of years of gaseous ¹⁴CO₂ and of ⁹⁰Sr and other fission products scavenged by aerosol were used to study transport within the stratosphere. For several decades, the ¹⁴CO₂ data were considered unreliable

and too sparse to be useful, at least in part because virtually all atmospheric models of the time dispersed and removed the $^{14}\text{CO}_2$ faster than was consistent with the observations. However, with the problems that contemporary models have had in calculating mean ages of air that are too young compared with those derived from CO_2 and SF_6 , coupled with recent observations of the slow dispersal of sulfuric acid aerosol from the Mt. Pinatubo eruption out of the tropics at altitudes greater than 22 km (see below), the data are now considered less suspect and are undergoing a renaissance of re-examination and comparisons with model predictions. In particular, the pulsed nature of this tracer is complementary to the continuous input of long-lived tracers with tropospheric sources, such as CO_2 , N_2O , CH_4 , and SF_6 noted above. A recent reanalysis of HTO measurements shows an interesting decay that is longer than the oldest mean ages or residence times for stratospheric air, which has been interpreted as the eigentime of the longest-lived mode of the stratospheric transport equations. Thus, even older data that were once thought to be suspect or too sparse to be compared with model results are being reanalyzed and reinterpreted using new insights as our understanding of stratospheric dynamics and tracer transport evolves.

A Natural 'Pulsed' Tracer: Sulfur from the Mt. Pinatubo Eruption

In June 1991, Mt. Pinatubo in the Philippines erupted, injecting tons of SO_2 into the tropical stratosphere. The SO_2 was rapidly oxidized to H_2SO_4 and resulted in sulfate aerosol concentrations 30 times higher than the nonvolcanic background levels. Because this material entered the tropics, its transport via the stratospheric circulation could be followed by satellite, balloon, aircraft, and ground-based aerosol measurements. Observations showed the rate of transport quasi-horizontally out to midlatitudes as a function of altitude. Transport out of the 'tropical reservoir' (so-called because the tropics acted to 'contain' the aerosol for a significant amount of time) to midlatitudes was rapid below altitudes of 22 km but very slow between 22 and 28 km, suggesting, as did the satellite water vapor data mentioned above, a relatively isolated region at $\sim 22\text{--}28$ km where transport was solely by vertical advection with little or no input of air from midlatitudes and only episodic and occasional transport from the tropical region out to midlatitudes.

Other Tracers in the Stratosphere: Cosmogenic Radionuclides and Stable Isotopes

Although a number of pioneering observations of cosmogenic radionuclides and stable isotope compositions of long-lived tracers were made, in both cases relatively new mass spectrometric techniques have been developed that make their measurement in the stratosphere both more practical and more compelling for the study of stratospheric transport processes in the future. They include continuous-flow isotope ratio mass spectrometry for stable isotopic analyses, which requires orders of magnitude smaller whole air sample sizes, and accelerator mass spectrometry, which allows precise ^{10}Be measurements to be made in conjunction with ^7Be to obtain the $^{10}\text{Be}/^7\text{Be}$ ratio, a more robust tracer than ^7Be alone.

Cosmogenic Radionuclides

^{10}Be and ^7Be are produced in the lower stratosphere and upper troposphere by cosmic ray-induced spallation reactions involving nitrogen and oxygen (e.g., $^{16}\text{O}(\text{n},4\text{p}3\text{n})^{10}\text{Be}$). Peak production rates occur at approximately 16–20 km at midlatitudes. Once formed, the ^{10}Be and ^7Be atoms are scavenged by aerosol particles and decay with half-lives of 1.6×10^6 years and 53 days, respectively. In the stratosphere, where submicrometer aerosol residence times are quite long, the ratio of $^{10}\text{Be}/^7\text{Be}$ will increase over time as the ^7Be decays rapidly relative to ^{10}Be . In the troposphere, however, aerosol residence times are short (on the order of 20–40 days), preventing the ratio of $^{10}\text{Be}/^7\text{Be}$ from increasing substantially above the production ratio. Thus, owing to this difference in aerosol residence times, the $^{10}\text{Be}/^7\text{Be}$ ratio can increase significantly above the production ratio only in the stratosphere, and the ratio represents an 'age' since the air mass was last cleansed of its beryllium atoms. In general, then, the distribution of ^{10}Be and ^7Be production in the atmosphere and the 'clock' provided by the difference in their half-lives make observations of the ratio useful for the study of both meridional and diffusive transport processes in the stratosphere and stratosphere–troposphere exchange on time scales of several months to several years. To date, however, there are fewer than 50 published observations of stratospheric $^{10}\text{Be}/^7\text{Be}$ ratios.

Stable Isotopes

Rates of chemical reactions often depend on whether the molecule possesses a heavy or light isotope; the difference in rates, for example, between the reactions $^{13}\text{CH}_4 + \text{Cl}$ and $^{12}\text{CH}_4 + \text{Cl}$ is called an isotope effect and is large at stratospheric temperatures ($\text{rate}(^{12}\text{CH}_4)/\text{rate}(^{13}\text{CH}_4)1.075$ at 223 K). Thus, as CH_4 is oxidized in the stratosphere, the remaining CH_4 becomes progressively enriched in ^{13}C . While the carbon isotopic composition of CH_4 is in some respects just another long-lived tracer, its correlation with CH_4 mixing ratios depends sensitively on transport while its sensitivity to, for example, modeled photochemistry should be the same for $^{13}\text{CH}_4$ as for $^{12}\text{CH}_4$ in a given model. Therefore, stable isotope compositions for CH_4 , N_2O , and CO_2 and other species may provide additional constraints on stratospheric photochemistry and transport in different regions of the stratosphere now that global-scale measurements are becoming more feasible.

Tracers of Stratosphere–Troposphere Exchange

Understanding the rates, patterns, and underlying dynamics of mass exchange between the troposphere and stratosphere has significant implications for chemistry in both regions and for predicting how these characteristics may change as climate changes. Tracers used to investigate troposphere-to-stratosphere transport have included CO_2 and water vapor mixing ratios, short-lived chemical species in the troposphere such as CHBr_3 , and ^{222}Rn , a radionuclide with a half-life of 3.8 days derived from crustal rocks and soils. Tracers used to investigate stratosphere-to-troposphere transport have included cosmogenic radionuclides and fission products from atmospheric nuclear testing.

Troposphere-To-Stratosphere Transport

From the seasonally resolved vertical ascent rates just above the tropical tropopause inferred from the water vapor and CO₂ observations discussed above, mass fluxes into the stratosphere can be calculated. In addition, the stratospheric CO₂ observations show that air entering the tropical stratosphere throughout the year is characteristic of the upper tropical troposphere, as opposed to air at any given surface site from either hemisphere being injected directly by convective systems without extensive mixing along the way. This analysis is predicated on the fact that CO₂ variations in the troposphere vary dramatically with latitude and altitude. Water vapor is a powerful tracer of details of troposphere-to-stratosphere transport since its saturation mixing ratio is very sensitive to the pathway and timing of entry. Dynamical and microphysical mechanisms that quantitatively explain stratospheric dryness are still being sought, however.

New measurements of the isotopic composition of water vapor entering the tropical stratosphere may shed further light on the dehydration mechanism(s), since the relative abundances of deuterium and ¹⁸O in upper tropospheric and lower stratospheric water vapor are very sensitive to the dehydration process. For example, models of slow vertical ascent into the stratosphere yield isotopic values very different from those from models of dehydration (and rehydration) in convective systems reaching or penetrating the tropopause.

Measurements of very short-lived chemical species, such as CHBr₃ (with a global lifetime of 0.1 years) or radioactive tracers with fast decay rates, such as ²²²Rn (with a half-life of 3.8 days) above and below the tropopause have also been pursued in order to provide information on rates of transport of air from the surface to the upper troposphere and the lower stratosphere. Significant vertical gradients exist for these species in the troposphere, as expected from their short lifetimes and surface sources. High spatial and temporal variability is observed in the upper troposphere, with higher mixing ratios or activities observed after recent convective events. Sporadic observations of such short-lived species have been made just above the tropical tropopause, but quantifying the average time scale for input of air across the tropical tropopause and its mechanism (i.e., slow vertical ascent or rapid overshooting convection) or determining from what characteristic altitude region the air comes from has so far remained elusive. In addition, tracers have also been used to test the degree and nature of transport of air across the extratropical tropopause. Studies using observations of CO₂, water vapor, and other tracers, for example, are based on the dependence of these tracers' concentrations with latitude of entry of air: CO₂ varies with latitude owing to the surface distribution of its sources and sinks, and water vapor varies owing to the variation of tropopause temperature with latitude. To date, these measurements have shown that diabatic transport across the midlatitude tropopause is minimal and that isentropic transport from the tropical/subtropical upper troposphere affects stratospheric composition only in the lowermost stratosphere (defined as the region below 380 K potential temperature).

Stratosphere-To-Troposphere Transport

Because most of the production of cosmogenic beryllium nuclides occurs in the stratosphere, ⁷Be has been used as

a qualitative tracer of stratosphere-to-troposphere transport. Measurement of the ¹⁰Be/⁷Be ratio is now preferred, however, since sampling errors cancel out and isotope fractionation does not occur during aerosol removal processes, such as gravitational settling in the stratosphere or wet and dry deposition in the troposphere. Thus, even if only a small fraction of beryllium atoms remain in sampled air, the ratio still retains its 'age' or stratospheric signature. For example, air samples collected weekly at the surface in the Arctic at Alert, Canada, and in some high-latitude upper tropospheric samples showed low ⁷Be concentrations that alone might have suggested no stratospheric influence; however, they have frequently corresponded to high ¹⁰Be/⁷Be ratios, which can only have come from the stratosphere. Estimates of stratospheric input of air to the Arctic lower troposphere of a few percent have been made from 1 year of ground-based observations. Other important tracers of the timing and magnitude of stratosphere-to-troposphere transport are the large data sets of fission products and tritium detected in tropospheric precipitation after the atmospheric nuclear bomb tests. These data show that transport from the stratosphere to the Northern Hemisphere maximizes during northern spring, even though the maximum mass flux into the stratosphere occurs December through February. The fact that the maximum mass flux out of the stratosphere occurs several months after the maximum mass flux in is consistent with the fact that the lowermost stratosphere grows in mass from December to March.

Tracers in the Troposphere

For comparison with the stratospheric and stratosphere-troposphere exchange tracers discussed above, several examples of tracer studies in the troposphere are briefly outlined here. Gradients in the mixing ratios of tracers such as CO₂ that have major sources in the Northern Hemisphere are used to constrain the rate of interhemispheric transport for air at the surface to about 1 year. The disappearance of these gradients in the upper troposphere indicates that air is efficiently mixed between the hemispheres in the intertropical convergence zone. Vertical transport has been investigated, for example, using the short-lived radioactive tracer ²²²Rn and its daughter product ²¹⁰Pb as well as the attenuation of the CO₂ annual cycle in the upper troposphere with respect to the surface. Simultaneous measurements of a set of chemical species with varying photochemical lifetimes have also been used in tropospheric studies, somewhat analogously to the CH₄:N₂O:CFC correlations in the stratosphere but operative on smaller spatial and temporal scales. For example, the ratios of alkanes in an air mass can be used to estimate the age of an industrial plume, since the ratios of the alkane concentrations change as the air mass ages due to their different reaction rates with OH. Tracer release studies are difficult in the troposphere as well as in the stratosphere. Small-scale releases of SF₆ have been used for studies of urban pollution dispersion and of vertical transport within a convective system. Larger-scale tracer release experiments are difficult in part due to the large mass of the troposphere. One attempt to overcome this difficulty was the release in 1984 of fully deuterated methane, ¹³CD₄ and ¹²CD₄, at an altitude of 5 km between Christchurch, New Zealand, and Antarctica, followed by collection of air from

aircraft and ground-based stations for analysis of the isotope ratios with respect to $^{12}\text{CH}_4$. Because the analysis could detect as little as two parts of tracer in 10^{17} parts of air against an essentially zero background, tracer could be detected in surface and mid-troposphere samples from 1 day to more than 3 weeks after the release. Some limited information was derived on tropospheric relative diffusion rates to hemispheric scales from this unique experiment.

See also: **Chemistry of the Atmosphere:** Methane; Radioactivity: Cosmogenic Radionuclides. **Climate and Climate Change:** Carbon Dioxide; Volcanoes: Role in Climate. **Middle Atmosphere:** Transport Circulation. **Ozone Depletion and Related Topics:** Ozone Depletion Potentials. **Satellites and Satellite Remote Sensing:** Measuring Ozone from Space – TOMS and SBUV. **Stratosphere/Troposphere Exchange and Structure:** Global Aspects; Local Processes. **Stratospheric Chemistry Topics:** Overview; Reactive Nitrogen (NO_x and NO_y); Stratospheric Water Vapor.

- Dobson, G.M.B., Harrison, D.N., Lawrence, J., 1929. Measurement of the amount of ozone in the earth's atmosphere and its relation to other geophysical conditions. *Proceedings of the Royal Society of London, Series A* 122, 456–486.
- Dobson, G.M.B., 1968. Forty years' research on atmospheric ozone at Oxford: a history. *Applied Optics* 7 (3), 387–405.
- Holton, J.R., Haynes, P.H., McIntyre, M.E., et al., 1995. Stratosphere–troposphere exchange. *Reviews of Geophysics* 33 (4), 403–439.
- Plumb, R.A., Ko, M.K.W., 1992. Interrelationships between mixing ratios of long-lived stratospheric constituents. *Journal of Geophysical Research* 97, 10145–10156.
- Plumb, R.A., 1996. A 'tropical pipe' model of stratospheric transport. *Journal of Geophysical Research* 101, 3957–3972.
- Prather, M.J., 1998. Time scales in atmospheric chemistry: coupled perturbations to N_2O , NO_y , and O_3 . *Science* 279, 1339–1341.
- Solomon, S., 1999. Stratospheric ozone depletion: a review of concepts and history. *Reviews of Geophysics* 37 (3), 275–316.
- Sparling, L.C., 2000. Statistical perspectives on stratospheric transport. *Reviews of Geophysics* 38 (3), 417–436.

Further Reading

- Brewer, A.M., 1949. Evidence for a world circulation provided by the measurements of helium and water vapor distribution in the stratosphere. *Quarterly Journal of the Royal Meteorological Society* 75, 351–363.
- Dessler, A.E., Burrage, M.D., Grooss, J.-U., et al., 1998. Selected science highlights from the first 5 years of the Upper Atmosphere Research Satellite (UARS) program. *Reviews of Geophysics* 36 (2), 183–210.

This page intentionally left blank

Second Edition

ENCYCLOPEDIA OF ATMOSPHERIC SCIENCES

Edited by
Gerald R. North **Editor-in-Chief**
John Pyle and Fuqing Zhang



6 VOLUME SET

ENCYCLOPEDIA OF **ATMOSPHERIC SCIENCES**

SECOND EDITION

VOLUME 2

This page intentionally left blank

ENCYCLOPEDIA OF ATMOSPHERIC SCIENCES

SECOND EDITION

EDITOR-IN-CHIEF

GERALD R NORTH

Texas A&M University, College Station, TX, USA

EDITORS

JOHN PYLE

Cambridge University, Cambridge, UK

FUQING ZHANG

Pennsylvania State University, University Park, PA, USA

VOLUME 2



Amsterdam • Boston • Heidelberg • London • New York • Oxford
Paris • San Diego • San Francisco • Singapore • Sydney • Tokyo
Academic Press is an imprint of Elsevier



Academic Press is an imprint of Elsevier
32 Jamestown Road, London NW1 7BY, UK
225 Wyman Street, Waltham, MA 02451, USA
525 B Street, Suite 1800, San Diego, CA 92101-4495, USA

Copyright © 2015 Elsevier Ltd. unless otherwise stated. All rights reserved.

No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means electronic, mechanical, photocopying, recording or otherwise without the prior written permission of the publisher

Permissions may be sought directly from Elsevier's Science & Technology Rights Department in Oxford, UK: phone (+44) (0) 1865 843830; fax (+44) (0) 1865 853333; email: permissions@elsevier.com. Alternatively you can submit your request online by visiting the Elsevier web site at <http://elsevier.com/locate/permissions>, and selecting *Obtaining permission to use Elsevier material*

Notice

No responsibility is assumed by the publisher for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions or ideas contained in the material herein, Because of rapid advances in the medical sciences, in particular, independent verification of diagnoses and drug dosages should be made

British Library Cataloguing in Publication Data

A catalog record for this book is available from the British Library

Library of Congress Catalog Number:

A catalog record for this book is available from the Library of Congress

ISBN (print): 978-0-12-382225-3

For information on all Elsevier publications
visit our website at store.elsevier.com

Printed and bound in the United Kingdom

15 16 17 18 19 10 9 8 7 6 5 4 3 2 1

		Working together to grow libraries in developing countries
www.elsevier.com • www.bookaid.org		

<i>Acquisitions Editor:</i> Simon Holt <i>Project Manager:</i> Michael Nicholls <i>Associate Project Manager:</i> Marise Willis <i>Designer:</i> Matthew Limbert

DEDICATION

This second edition of the *Encyclopedia of Atmospheric Sciences* is dedicated to the memory of James Holton who was editor-in-chief of the first edition. He was a great researcher and colleague inspiring an entire generation of atmospheric scientists.

This page intentionally left blank

CONTENTS

<i>List of Contributors</i>	xxvii
<i>Preface to the First Edition</i>	xxxix
<i>Preface to the Second Edition</i>	xli
<i>Editor Biographies</i>	xliii
<i>Guide to Using the Encyclopedia</i>	xlvi

VOLUME 1

BASIC ATMOSPHERIC STRUCTURE AND CONCEPTS	1
Beaufort Wind Scale	1
<i>L Hasse</i>	
Wind Chill	7
<i>M Bluestein</i>	
Standard Atmosphere	12
<i>W W Vaughan</i>	
AEROSOLS	17
Aerosol–Cloud Interactions and Their Radiative Forcing	17
<i>U Lohmann</i>	
Aerosol Physics and Chemistry	23
<i>M Kalberer</i>	
Climatology of Stratospheric Aerosols	32
<i>L W Thomason and J-P Vernier</i>	
Climatology of Tropospheric Aerosols	40
<i>N Bellouin and J Haywood</i>	
Dust	48
<i>I N Sokolik</i>	
Observations and Measurements	53
<i>P H McMurry</i>	
Role in Radiative Transfer	66
<i>G A Ban-Weiss, and W D Collins</i>	

Role in Climate Change <i>N Bellouin</i>	76
Soot <i>P Chylek, S G Jennings, and R Pinnick</i>	86
Agricultural Meteorology and Climatology <i>E S Takle</i>	92
ARCTIC AND ANTARCTIC	98
Antarctic Climate <i>J Turner</i>	98
Arctic Climate <i>M C Serreze</i>	107
Arctic Haze <i>L M Russell and G E Shaw</i>	116
AIR SEA INTERACTIONS	
Freshwater Flux <i>J Schulz</i>	122
Momentum, Heat, and Vapor Fluxes <i>P K Taylor</i>	129
Sea Surface Temperature <i>W J Emery</i>	136
Surface Waves <i>A Benilov</i>	144
AVIATION METEOROLOGY	153
Aircraft Emissions <i>R R Friedl</i>	153
Aircraft Icing <i>M K Politovich</i>	160
Aviation Weather Hazards <i>A J Bedard, Jr</i>	166
Clear Air Turbulence <i>G P Ellrod (Retired), J A Knox, P F Lester, and L J Ehernberger (Retired)</i>	177
BIOGEOCHEMICAL CYCLES	187
Sulfur Cycle <i>P Brimblecombe</i>	187
Bromine <i>R von Glasow and C Hughes</i>	194
Heavy Metals <i>T D Jickells and A R Baker</i>	201

Iodine	205
<i>L J Carpenter</i>	
BOUNDARY LAYER (ATMOSPHERIC) AND AIR POLLUTION	220
Overview	220
<i>P J Mason and D J Thomson</i>	
Air Pollution Meteorology	227
<i>X-M Hu</i>	
Coherent Structures	237
<i>F T M Nieuwstadt and J C R Hunt</i>	
Complex Terrain	242
<i>J J Finnigan</i>	
Convective Boundary Layer	250
<i>M A LeMone</i>	
Microclimate	258
<i>M W Rotach and P Calanca</i>	
Modeling and Parameterization	265
<i>A A M Holtslag</i>	
Observational Techniques <i>In Situ</i>	274
<i>E F Bradley</i>	
Observational Techniques: Remote	284
<i>W M Angevine and C J Senff</i>	
Ocean Mixed Layer	290
<i>L Kantha and C A Clayson</i>	
Stably Stratified Boundary Layer	299
<i>L Mahrt</i>	
Surface Layer	305
<i>G L Geernaert</i>	
Urban Heat Islands	310
<i>J C Luvall, D A Quattrochi, D L Rickman, and M G Estes, Jr</i>	
Diurnal Cycle	319
<i>A Betts</i>	
CHEMISTRY OF THE ATMOSPHERE	324
Chemical Kinetics	324
<i>R P Wayne</i>	
Ion Chemistry	333
<i>J L Fox</i>	
Isotopes, Stable	348
<i>C A M Brenninkmeijer</i>	
Laboratory Kinetics	356
<i>D J Donaldson and S N Wren</i>	

Methane <i>E Dlugokencky, and S Houweling</i>	363
Observations for Chemistry (<i>In Situ</i>): Ozone Sonde <i>H G J Smit</i>	372
Observations for Chemistry (<i>In Situ</i>): Particles <i>T Deshler</i>	379
Observations for Chemistry (<i>In Situ</i>): Water Vapor Sonde <i>J B Smith</i>	387
Observations for Chemistry (Remote Sensing): IR/FIR (Satellite, Balloon and Ground) <i>H Fischer and F Hase</i>	401
Observations for Chemistry (Remote Sensing): Lidar <i>G Vaughan</i>	411
Observations for Chemistry (Remote Sensing): Microwave <i>J Waters</i>	418
Principles of Chemical Change <i>R P Wayne</i>	429
Radioactivity: Cosmogenic Radionuclides <i>D Lal</i>	437
Volcanoes: Composition of Emissions <i>M T Coffey and J W Hannigan</i>	446
Tracers <i>K A Boering</i>	450

VOLUME 2

CLIMATE AND CLIMATE CHANGE	1
Overview <i>D L Hartmann</i>	1
Carbon Dioxide <i>C L Sabine and R A Feely</i>	10
Climate Feedbacks <i>A E Dessler and M D Zelinka</i>	18
Climate Prediction: Empirical and Numerical <i>S Hastenrath</i>	26
Climate Variability: Decadal to Centennial Variability <i>D G Martinson</i>	33
Climate Variability: Nonlinear and Random Effects <i>M Ghil</i>	38
Climate Variability: North Atlantic and Arctic Oscillation <i>J W Hurrell</i>	47
Climate Variability: Seasonal and Interannual Variability <i>D S Gutzler</i>	61

Energy Balance Climate Models <i>G R North and K-Y Kim</i>	69
Global Impacts of the Madden–Julian Oscillation <i>C Zhang</i>	73
Greenhouse Effect <i>G R North</i>	80
History of Scientific Work on Climate Change <i>S Weart</i>	87
Intergovernmental Panel on Climate Change <i>K E Trenberth</i>	90
Nuclear Winter <i>A Robock</i>	95
Radiative–Convective Equilibrium Climate Models <i>N O Renno and X Huang</i>	102
Volcanoes: Role in Climate <i>A Robock</i>	105
CLOUDS AND FOG	112
Cloud Modeling <i>W-K Tao and M Moncrieff</i>	112
Contrails <i>P Minnis</i>	121
Cloud Microphysics <i>D Lamb</i>	133
Classification of Clouds <i>A L Rangno (Retiree)</i>	141
Climatology <i>S Warren, R Eastman, and C J Hahn</i>	161
Measurement Techniques <i>In situ</i> <i>D Baumgardner, J-F Gayet, A Korolev, C Twohy, and J Fugal</i>	170
Fog <i>P J Croft and B Ward</i>	180
Noctilucent Clouds <i>G E Thomas</i>	189
Stratus and Stratocumulus <i>R Wood</i>	196
CRYOSPHERE	201
Glaciers, Topography, and Climate <i>A B G Bush and M P Bishop</i>	201
Permafrost <i>T E Osterkamp and C R Burn</i>	208

Sea Ice	217
<i>M C Serreze, F Fetterer, and W F Weeks (Retired)</i>	
Snow (Surface)	227
<i>M Sturm</i>	
DATA ASSIMILATION AND PREDICTABILITY	237
Data Assimilation	237
<i>A C Lorenc</i>	
Ensemble-Based Data Assimilation	241
<i>Z Meng and F Zhang</i>	
Ensemble Prediction	248
<i>R Buizza</i>	
Predictability and Chaos	258
<i>L A Smith</i>	
DYNAMICAL METEOROLOGY	265
Overview	265
<i>J R Holton</i>	
Acoustic Waves	272
<i>K E Gilbert</i>	
Atmospheric Tides	287
<i>J Oberheide, M E Hagan, A D Richmond, and J M Forbes</i>	
Balanced Flow	298
<i>M E McIntyre</i>	
Baroclinic Instability	304
<i>R Grotjahn</i>	
Coriolis Force	313
<i>D W Moore</i>	
Critical Layers	317
<i>P Haynes</i>	
Hamiltonian Dynamics	324
<i>T G Shepherd</i>	
Hydraulic Flow	332
<i>R B Smith</i>	
Inertial Instability	334
<i>J A Knox</i>	
Kelvin–Helmholtz Instability	343
<i>P G Drazin</i>	
Kelvin Waves	347
<i>B Wang</i>	
Kinematics	353
<i>D D Houghton</i>	

Laboratory Geophysical Fluid Dynamics <i>R L Pfeffer</i>	360
Lagrangian Dynamics <i>I Roulstone</i>	369
Potential Vorticity <i>M E McIntyre</i>	375
Primitive Equations <i>A A White and N Wood</i>	384
Quasigeostrophic Theory <i>H C Davies and H Wernli</i>	393
Rossby Waves <i>P B Rhines</i>	404
Solitary Waves <i>J P Boyd</i>	417
Static Stability <i>J A Young</i>	423
Stationary Waves (Orographic and Thermally Forced) <i>S Nigam and E DeWeaver</i>	431
Symmetric Stability <i>H B Bluestein</i>	446
Vorticity <i>J R Holton</i>	451
Wave-CISK <i>C S Bretherton</i>	455
Wave Mean-Flow Interaction <i>M Juckes</i>	458
Waves <i>J R Holton</i>	464

VOLUME 3

ELECTRICITY IN THE ATMOSPHERE	1
Global Electrical Circuit <i>E R Williams</i>	1
Ions in the Atmosphere <i>K L Aplin and R G Harrison</i>	9
Lightning <i>M B Baker</i>	14
Sprites <i>W A Lyons</i>	20
Forensic Meteorology <i>L E Branscome</i>	28

GENERAL CIRCULATION OF THE ATMOSPHERE	33
Overview	33
<i>J M Wallace, D W J Thompson, and P Beresford</i>	
Angular Momentum of the Atmosphere	43
<i>D A Salstein</i>	
Energy Cycle	51
<i>R Grotjahn</i>	
Weather Regimes and Multiple Equilibria	65
<i>F Molteni</i>	
Mean Characteristics	73
<i>R Grotjahn</i>	
Teleconnections	90
<i>S Nigam and S Baxter</i>	
 GLOBAL CHANGE	 110
Climate Record: Surface Temperature Trends	110
<i>P D Jones</i>	
Sea Level Change	121
<i>R S Nerem</i>	
Upper Atmospheric Change	128
<i>R G Roble</i>	
Biospheric Impacts and Feedbacks	132
<i>B A Hungate and G W Koch</i>	
 GRAVITY WAVES	 141
Overview	141
<i>D C Fritts</i>	
Buoyancy and Buoyancy Waves: Optical Observations	153
<i>M J Taylor and W R Pendleton, Jr</i>	
Buoyancy and Buoyancy Waves: Theory	160
<i>T J Dunkerton</i>	
Gravity Waves Excited by Jets and Fronts	164
<i>R Plougonven and F Zhang</i>	
Convectively Generated Gravity Waves	171
<i>T P Lane</i>	
 HYDROLOGY, FLOODS AND DROUGHTS	 180
Overview	180
<i>R C Bales</i>	
Deserts and Desertification	185
<i>V P Tchakerian</i>	
Drought	193
<i>S Quiring</i>	

Flooding	201
<i>C A Doswell III</i>	
Groundwater and Surface Water	209
<i>S Ge and S M Gorelick</i>	
Modeling and Prediction	217
<i>Z Yu</i>	
Palmer Drought Severity Index	224
<i>L Nkemdirim</i>	
Soil Moisture	232
<i>A Robock</i>	
LAND-ATMOSPHERE INTERACTIONS	240
Overview	240
<i>R E Dickinson</i>	
Canopy Processes	244
<i>P D Blanken</i>	
Trace Gas Exchange	256
<i>J N Cape and D Fowler</i>	
LIDAR	262
Atmospheric Sounding Introduction	262
<i>P S Argall and R Sica</i>	
Backscatter	270
<i>C M R Platt and R L Collins</i>	
Differential Absorption Lidar	277
<i>S Ismail and E V Browell</i>	
Doppler	289
<i>R M Hardesty</i>	
Raman	296
<i>D N Whiteman</i>	
Resonance	305
<i>C S Gardner and R L Collins</i>	
Magnetosphere	309
<i>G K Parks</i>	
MESOSCALE METEOROLOGY	316
Overview	316
<i>D J Parker</i>	
Cloud and Precipitation Bands	323
<i>R M Rauber and M Ramamurthy</i>	
Gust Fronts	331
<i>R Rotunno</i>	

Hail and Hailstorms	334
<i>C Knight, N Knight, and H E Brooks</i>	
Mesoscale Convective Systems	339
<i>A Laing</i>	
Microbursts	335
<i>R M Wakimoto</i>	
Severe Storms	361
<i>C A Doswell III</i>	
Waterspouts	369
<i>J H Golden</i>	
Bow Echoes and Derecho	384
<i>M L Weisman</i>	
Density Currents	395
<i>P G Baines</i>	
Convective Storms: Overview	401
<i>M L Weisman</i>	
MESOSPHERE	411
Atomic Species in the Mesopause Region	411
<i>M G Mlynczak and L A Hunt</i>	
Ionosphere	422
<i>M C Kelley</i>	
Metal Layers	430
<i>J M C Plane</i>	
Polar Summer Mesopause	436
<i>R H Varney and M C Kelley</i>	
VOLUME 4	
MIDDLE ATMOSPHERE	1
Planetary Waves	1
<i>A K Smith and J Perlwitz</i>	
Polar Vortex	12
<i>M R Schoeberl and P A Newman</i>	
Quasi-Biennial Oscillation	18
<i>T J Dunkerton, J A Anstey, and L J Gray</i>	
Semiannual Oscillation	26
<i>K Hamilton</i>	
Stratospheric Sudden Warmings	30
<i>A O'Neill, A J Charlton-Perez, and L M Polvani</i>	
Transport Circulation	41
<i>S E Strahan</i>	

Zonal Mean Climatology <i>P Braesicke</i>	50
MOUNTAIN METEOROLOGY	57
Overview <i>R B Smith</i>	57
Cold Air Damming <i>B A Colle</i>	62
Downslope Winds <i>D R Durran</i>	69
Katabatic Winds <i>T R Parish</i>	75
Land and Sea Breezes <i>R A Pielke, Sr</i>	80
Lee Vortices <i>C C Epifanio</i>	84
Lee Waves and Mountain Waves <i>D R Durran</i>	95
Orographic Effects: Lee Cyclogenesis <i>C Schär</i>	103
Valley Winds <i>D Zardi</i>	114
NUMERICAL MODELS	135
Chemistry Models <i>M P Chipperfield and S R Arnold</i>	135
Coupled Ocean-Atmosphere Models: Physical Processes <i>M Zhang</i>	144
General Circulation Models <i>C R Mechoso and A Arakawa</i>	153
Methods <i>J Thuburn</i>	161
Model Physics Parameterization <i>D J Stensrud, M C Coniglio, K H Knopfmeier, and A J Clark</i>	167
Parameter Estimation <i>A Aksoy</i>	181
Parameterization of Physical Processes: Clouds <i>R Forbes, C Jakob, and M Miller</i>	187
Parameterization of Physical Processes: Gravity Wave Fluxes <i>M J Alexander</i>	194
Parameterization of Physical Processes: Turbulence and Mixing <i>A Beljaars</i>	200

Spectral Models <i>F Baer</i>	212
Mesoscale Atmospheric Modeling <i>R A Pielke, Sr</i>	219
Cloud-System Resolving Modeling and Aerosols <i>W-K Tao and T Matsui</i>	222
Large-Eddy Simulation <i>C-H Moeng and P P Sullivan</i>	232
Regional Prediction Models <i>B W Golding</i>	241
Convective Storm Modeling <i>M D Parker</i>	246
OBSERVATIONS PLATFORMS	255
Balloons <i>J-P Pommereau</i>	255
Buoys <i>J M Hemsley</i>	264
Kites <i>B B Balsley</i>	268
Radiosondes <i>W F Dabberdt and H Turtiainen</i>	273
Rockets <i>M F Larsen</i>	285
OCEANOGRAPHIC TOPICS	290
General Processes <i>N C Wells</i>	290
Surface/Wind Driven Circulation <i>R X Huang</i>	301
Thermohaline Circulation <i>R X Huang</i>	315
Water Types and Water Masses <i>W J Emery</i>	329
OPTICS, ATMOSPHERIC	338
Optical Remote Sensing Instruments <i>G G Shepherd</i>	338
Airglow Instrumentation <i>M Conde</i>	346

OZONE DEPLETION AND RELATED TOPICS	353
Long-Term Ozone Changes <i>N R P Harris</i>	353
Ozone as a UV Filter <i>J E Frederick</i>	359
Ozone Depletion Potentials <i>D J Wuebbles</i>	364
Photochemistry of Ozone <i>G K Moortgat and A R Ravishankara</i>	370
Stratospheric Ozone Recovery <i>D J Hofmann and R Müller</i>	380
Surface Ozone Effects on Vegetation <i>M Ashmore</i>	389
Surface Ozone (Human Health) <i>M Lippmann</i>	397
PALEOCLIMATOLOGY	404
Ice Cores <i>E J Steig</i>	404
Varves <i>R Gilbert</i>	411
RADAR	415
Cloud Radar <i>T Uttal</i>	415
Incoherent Scatter Radar <i>M P Sulzer</i>	422
Mesosphere—Stratosphere—Troposphere and Stratosphere—Troposphere Radars and Wind Profilers <i>G Vaughan and D Hooper</i>	429
Meteor Radar <i>N J Mitchell</i>	438
Polarimetric Doppler Weather Radar <i>R J Doviak and R D Palmer</i>	444
Precipitation Radar <i>S E Yuter</i>	455
Synthetic Aperture Radar (Land Surface Applications) <i>R K Vincent</i>	470
VOLUME 5	
RADIATION TRANSFER IN THE ATMOSPHERE	1
Radiation, Solar <i>Q Fu</i>	1

Absorption and Thermal Emission <i>R M Goody and X Huang</i>	5
Cloud-Radiative Processes <i>Q Fu</i>	13
Non-local Thermodynamic Equilibrium <i>M López-Puertas and B Funke</i>	16
Scattering <i>M Mishchenko, L Travis, and A Lacis</i>	27
Ultraviolet Radiation <i>K Stamnes</i>	37
Ultraviolet, Surface <i>R McKenzie and S Madronich</i>	45
SATELLITES AND SATELLITE REMOTE SENSING	51
Aerosol Measurements <i>R A Kahn</i>	51
Earth's Radiation Budget <i>N G Loeb and B A Wielicki</i>	67
GPS Meteorology <i>S S Leroy</i>	77
Measuring Ozone from Space — TOMS and SBUV <i>R D McPeters and R S Stolarski</i>	87
Orbits <i>S Q Kidder</i>	95
Precipitation <i>G Liu</i>	107
Remote Sensing: Cloud Properties <i>P Yang and B A Baum</i>	116
Research <i>M D King</i>	128
Surface Wind and Stress <i>W T Liu</i>	138
Temperature Soundings <i>A Dudhia</i>	145
Water Vapor <i>J E Harries</i>	157
SOLAR SYSTEM/SUN, ATMOSPHERES, EVOLUTION OF ATMOSPHERES	163
Evolution of Earth's Atmosphere <i>Y L Yung, M L Wong, and E J Gaidos</i>	163
Planetary Atmospheres: Mars <i>R M Haberle</i>	168

Planetary Atmospheres: Venus <i>P J Gierasch and Y L Yung</i>	178
Solar Terrestrial Interactions: Climate Impact <i>J D Haigh</i>	183
Solar Winds <i>S T Suess and B T Tsurutani</i>	189
Meteors <i>P Jenniskens</i>	195
STATISTICAL METHODS	201
Data Analysis: Empirical Orthogonal Functions and Singular Vectors <i>C S Bretherton</i>	201
Data Analysis: Time Series Analysis <i>G R North</i>	205
STRATOSPHERIC CHEMISTRY TOPICS	211
Overview <i>J A Pyle</i>	211
Halogens <i>D Toohey</i>	215
Halogen Sources, Anthropogenic <i>A McCulloch and P M Midgley</i>	221
Halogen Sources, Natural (Methyl Bromide and Related Gases) <i>S Yvon-Lewis and J H Butler</i>	228
HO _x <i>T F Hanisco</i>	233
Hydrogen Budget <i>J E Harries</i>	238
Reactive Nitrogen (NO _x and NO _y) <i>Y Kondo</i>	242
Stratospheric Water Vapor <i>K H Rosenlof</i>	250
STRATOSPHERE/TROPOSPHERE EXCHANGE AND STRUCTURE	257
Global Aspects <i>J R Holton</i>	257
Local Processes <i>J F Lamarque and P Hess</i>	262
Tropopause <i>M Dameris</i>	269

SYNOPTIC METEOROLOGY	273
Anticyclones <i>S J Colucci</i>	273
Forecasting <i>D Mansfield</i>	280
Weather Maps <i>R Reynolds</i>	289
Cyclogenesis <i>G J Hakim</i>	299
Extratropical Cyclones <i>A Joly</i>	304
Fronts <i>D M (David) Schultz and W Blumen</i>	337
Fronts in the Lower Stratosphere <i>A L Lang</i>	344
Frontogenesis <i>D M (David) Schultz</i>	353
Jet Streaks <i>P Cunningham and D Keyser</i>	359
Lake-Effect Storms <i>P J Sousounis</i>	370
Polar Lows <i>I A Renfrew</i>	379
Thermal Low <i>R H Johnson</i>	386
 THERMODYNAMICS	 391
Humidity Variables <i>J A Curry</i>	391
Moist (Unsaturated) Air <i>J A Curry</i>	394
Saturated Adiabatic Processes <i>J A Curry</i>	398
 Thermosphere <i>S C Solomon and R G Roble</i>	 402
 VOLUME 6	
TROPICAL CYCLONES AND HURRICANES	1
Overview and Theory <i>R A Tomas and P J Webster</i>	1

Hurricane Dynamics <i>Y Wang</i>	8
Hurricane Predictability <i>J A Sippel</i>	30
Hurricanes: Observation <i>F D Marks</i>	35
Tropical Cyclogenesis <i>Z Wang</i>	57
Tropical Cyclones and Climate Change <i>T R Knutson</i>	65
Tropical Cyclones in the Western North Pacific <i>J C L Chan</i>	77
Tropical Cyclones: Secondary Eyewall Formation <i>C-C Wu and Y-H Huang</i>	85
TROPICAL METEOROLOGY AND CLIMATE	91
El Niño and the Southern Oscillation: Observation <i>N Nicholls</i>	91
El Niño and the Southern Oscillation: Theory <i>P Chang and S E Zebiak</i>	97
Equatorial Waves <i>M C Wheeler and H Nguyen</i>	102
Hadley Circulation <i>J Lu and G A Vecchi</i>	113
Intertropical Convergence Zone <i>D E Waliser and X Jiang</i>	121
Intraseasonal Oscillation (Madden–Julian Oscillation) <i>R A Madden</i>	132
Madden–Julian Oscillation: Skeleton and Conceptual Models <i>A J Majda and S N Stechmann</i>	137
Monsoon: Overview <i>J Slingo</i>	146
Monsoon: Dynamical Theory <i>P J Webster and J Fasullo</i>	151
Monsoon: ENSO–Monsoon Interactions <i>K-M Lau</i>	165
Tropical Climates <i>S Hastenrath</i>	170
Walker Circulation <i>K-M Lau and S Yang</i>	177

TROPOSPHERIC CHEMISTRY AND COMPOSITION	182
Aerosols/Particles <i>J H Seinfeld</i>	182
Aliphatic Hydrocarbons <i>J Rudolph and O Stein</i>	188
Aromatic Hydrocarbons <i>I Barnes</i>	204
Biogenic Hydrocarbons <i>A Guenther</i>	214
Cloud Chemistry <i>P Herckes and J L Collett, Jr</i>	218
H ₂ <i>U Schmidt and T Wetter</i>	226
Hydroxyl Radical <i>K C Clemitshaw</i>	232
Mercury <i>J Munthe and J Sommar</i>	239
Oxidizing Capacity <i>D H Ehhalt, F Rohrer, and A Wahner</i>	243
Peroxyacetyl Nitrate <i>H B Singh</i>	251
Sulfur Chemistry, Organic <i>I Barnes</i>	255
Volatile Organic Compounds Overview: Anthropogenic <i>R G Derwent</i>	265
TURBULENCE AND MIXING	268
Overview <i>P Haynes</i>	268
Turbulence, Two Dimensional <i>P Bartello</i>	273
Turbulent Diffusion <i>A Venkatram and S Du</i>	277
WEATHER FORECASTING	287
Marine Meteorology <i>L Xie and B Liu</i>	287
Operational Meteorology <i>D R Novak</i>	293
Seasonal and Interannual Weather Prediction <i>J P Li and R Q Ding</i>	303
Severe Weather Forecasting <i>D J Stensrud, H E Brooks, and S J Weiss</i>	313

Wildfire Weather <i>J Coen</i>	323
Inadvertant Weather Modification <i>S A Changnon</i>	332
<i>Appendix 1: Physical Constants</i>	337
<i>Appendix 2: Units and their SI Equivalents</i>	339
<i>Appendix 3: Periodic Table of the Elements</i>	340
<i>Appendix 4: The Geologic Time Scale</i>	341
<i>Index</i>	343

This page intentionally left blank

LIST OF CONTRIBUTORS

A. Aksoy
University of Miami, Miami, FL, USA; and NOAA Hurricane Research Division, Miami, FL, USA

M.J. Alexander
NorthWest Research Associates (NWRA), Boulder, CO, USA

W.M. Angevine
CIRES, University of Colorado; and NOAA Earth System Research Laboratory, Boulder, CO, USA

J.A. Anstey
University of Oxford, Oxford, UK

K.L. Aplin
University of Oxford, Oxford, UK

A. Arakawa
University of California, Los Angeles, CA, USA

P.S. Argall
The University of Western Ontario, London, ON, Canada

S.R. Arnold
University of Leeds, Leeds, UK

M. Ashmore
University of York, York, UK

F. Baer
University of Maryland, College Park, MD, USA

P.G. Baines
University of Melbourne, Melbourne, VIC, Australia

A.R. Baker
University of East Anglia, Norwich, UK

M.B. Baker
University of Washington, Seattle, WA, USA

R.C. Bales
University of Arizona, Tucson, AZ, USA

B.B. Balsley
University of Colorado, Boulder, CO, USA

G.A. Ban-Weiss
Lawrence Berkeley National Laboratory, Berkeley, CA, USA; and University of Southern California, Los Angeles, CA, USA

I. Barnes
University of Wuppertal, Wuppertal, Germany

P. Bartello
McGill University, Montréal, QC, Canada

B.A. Baum
University of Wisconsin–Madison, Madison, WI, USA

D. Baumgardner
Universidad Nacional Autónoma de México, Mexico City, D.F., Mexico

S. Baxter
University of Maryland, College Park, MD, USA

A.J. Bedard, Jr.
National Oceanic and Atmospheric Administration, Boulder, CO, USA

A. Beljaars
European Centre for Medium-Range Weather Forecasts, Reading, England

N. Bellouin
University of Reading, Reading, UK

A. Benilov
Acute Solutions, Highlands, NJ, USA

P. Beresford
European Centre for Medium-Range Weather Forecasts, Reading, UK

A. Betts
Atmospheric Research, Pittsford, VT, USA

M.P. Bishop
Texas A&M University, College Station, TX, USA

P.D. Blanken
University of Colorado at Boulder, Boulder, CO, USA

H.B. Bluestein
University of Oklahoma, Norman, OK, USA

M. Bluestein
Indiana University – Purdue University, Indianapolis, IN, USA

W. Blumen[†]
University of Colorado Boulder, Boulder, CO, USA

K.A. Boering
University of California – Berkeley, Berkeley, CA, USA

J.P. Boyd
University of Michigan, Ann Arbor, MI, USA

E.F. Bradley
CSIRO Land and Water, Canberra, ACT, Australia

P. Braesicke
Karlsruhe Institute of Technology, Karlsruhe, Germany

L.E. Branscome
Climatological Consulting Corporation, FL, USA

C.A.M. Brenninkmeijer
Max Planck Institute for Chemistry, Mainz, Germany

C.S. Bretherton
University of Washington, Seattle, WA, USA

P. Brimblecombe
University of East Anglia, Norwich, UK

H.E. Brooks
National Oceanic and Atmospheric Administration, Norman, OK, USA

E.V. Browell
STARS II Affiliate, NASA Langley Research Center, Hampton, VA, USA

R. Buizza
ECMWF, Reading, UK

C.R. Burn
Carleton University, Ottawa, ON, Canada

A.B.G. Bush
University of Alberta, Edmonton, AB, Canada

J.H. Butler
NOAA Earth System Research Laboratory, Boulder, CO, USA

P. Calanca
Agroscope Reckenholz-Taenikon, Zurich, Switzerland

J.N. Cape
Edinburgh Research Station, Midlothian, UK

L.J. Carpenter
University of York, York, UK

J.C.L. Chan
City University of Hong Kong, Hong Kong

P. Chang
Texas A&M University, College Station, TX, USA

S.A. Changnon
University of Illinois, IL, USA

A.J. Charlton-Perez
University of Reading, Earley Gate, Reading, UK

M.P. Chipperfield
University of Leeds, Leeds, UK

P. Chylek
Dalhousie University, NS, Canada

A.J. Clark
University of Oklahoma and National Oceanic and Atmospheric Administration, Norman, OK, USA

C.A. Clayson
Woods Hole Oceanographic Institution, Woods Hole, MA, USA

K.C. Clemitshaw
Imperial College of Science, Technology, and Medicine, Ascot, UK

J. Coen
National Center for Atmospheric Research, Boulder, CO, USA

M.T. Coffey
National Center for Atmospheric Research, Boulder, CO, USA

B.A. Colle
Stony Brook University – SUNY, Stony Brook, NY, USA

J.L. Collett, Jr.
Colorado State University, Fort Collins, CO, USA

R.L. Collins
University of Alaska Fairbanks, Fairbanks, AK, USA

W.D. Collins
Lawrence Berkeley National Laboratory, Berkeley, CA, USA

S.J. Colucci
Cornell University, Ithaca, NY, USA

M. Conde
University of Alaska Fairbanks, Fairbanks, AK, USA

M.C. Coniglio
National Oceanic and Atmospheric Administration, Norman, OK, USA

[†] Deceased.

- P.J. Croft
Kean University, Union, NJ, USA
- P. Cunningham
Florida State University, Tallahassee, FL, USA
- J.A. Curry
Georgia Institute of Technology, Atlanta, GA, USA
- W.F. Dabberdt
Vaisala Company, Boulder, CO, USA
- M. Dameris
Deutsches Zentrum für Luft- und Raumfahrt, Institut für Physik der Atmosphäre, Oberpfaffenhofen, Wessling, Germany
- H.C. Davies
Institute for Atmospheric and Climate Science, Swiss Federal Institute of Technology (ETH), Zürich, Switzerland
- R.G. Derwent
rdscientific, Newbury, UK
- T. Deshler
University of Wyoming, Laramie, WY, USA
- A.E. Dessler
Texas A&M University, College Station, TX, USA
- E. DeWeaver
University of Wisconsin, Madison, WI, USA
- R.E. Dickinson
University of Texas at Austin, Austin, TX, USA
- R.Q. Ding
Institute of Atmospheric Physics, Chinese Academy of Sciences, Beijing, China
- E. Dlugokencky
NOAA Earth System Research Laboratory, Boulder, CO, USA
- D.J. Donaldson
University of Toronto, Toronto, ON, Canada
- C.A. Doswell, III
University of Oklahoma, Norman, OK, USA
- R.J. Doviak
National Severe Storms Laboratory, Norman, OK, USA
- P.G. Drazin[†]
University of Bath, England, UK
- S. Du
California Air Resources Board, Sacramento, CA, USA
- A. Dudhia
University of Oxford, Oxford, UK
- T.J. Dunkerton
Northwest Research Associates, Bellevue, WA, USA
- D.R. Durran
University of Washington, Seattle, WA, USA
- R. Eastman
University of Washington, Seattle, WA, USA
- L.J. Ehernberger
National Aeronautics and Space Administration, Dryden Flight Research Center, Edwards, CA, USA
- D.H. Ehhalt
Forschungszentrum Jülich, Jülich, Germany
- G.P. Ellrod
National Oceanographic and Atmospheric Administration/National Environmental Satellite, Data, and Information Service, Granby, CT, USA
- W.J. Emery
University of Colorado, Boulder, CO, USA
- C.C. Epifanio
Texas A&M University, College Station, TX, USA
- M.G. Estes
Universities Space Research Association, Huntsville, AL, USA
- J. Fasullo
University of Colorado – Boulder, Boulder, CO, USA
- R.A. Feely
NOAA Pacific Marine Environmental Laboratory, Seattle, WA, USA
- F. Fetterer
University of Colorado, Boulder, CO, USA
- J.J. Finnigan
CSIRO Atmospheric Research, Black Mountain, ACT, Australia
- H. Fischer
Institute of Meteorology and Climate Research (IMK), Karlsruhe Institute of Technology, Karlsruhe, Germany
- J.M. Forbes
University of Colorado, Boulder, CO, USA
- R. Forbes
European Centre for Medium-Range Weather Forecasts, Reading, UK
- D. Fowler
Edinburgh Research Station, Midlothian, UK

[†] Deceased.

J.L. Fox
Wright State University, Dayton, OH, USA

J.E. Frederick
The University of Chicago, Chicago, IL, USA

R.R. Friedl
California Institute of Technology, Pasadena, CA, USA

D.C. Fritts
GATS Inc., Boulder, CO, USA

Q. Fu
University of Washington, Seattle, WA, USA

J. Fugal
Max Planck Institute of Chemistry, Mainz, Germany

B. Funke
Instituto de Astrofísica de Andalucía, CSIC, Granada, Spain

E.J. Gaidos
University of Hawaii at Manoa, Honolulu, HI, USA

C.S. Gardner
University of Illinois at Urbana-Champaign, Urbana, IL, USA

J.-F. Gayet
Université Blaise Pascal, Clermont Ferrand, France

S. Ge
University of Colorado, Boulder, CO, USA

G.L. Geernaert
US Department of Energy, Washington, DC, USA

M. Ghil
Ecole Normale Supérieure, Paris, France; and University of California, Los Angeles, CA, USA

P.J. Gierasch
Cornell University, Ithaca, NY, USA

K.E. Gilbert
University of Mississippi, University, MS, USA

R. Gilbert
Queen's University, Kingston, ON, Canada

J.H. Golden
Forecast Systems Laboratory, NOAA, Boulder, CO, USA

B.W. Golding
Met Office, Exeter, UK

R.M. Goody
Harvard University (Emeritus), Cambridge, MA, USA

S.M. Gorelick
Stanford University, Stanford, CA, USA

L.J. Gray
University of Oxford, Oxford, UK

R. Grotjahn
University of California, Davis, CA, USA

A. Guenther
Pacific Northwest National Laboratory, Richland, WA, USA

D.S. Gutzler
University of New Mexico, Albuquerque, NM, USA

R.M. Haberle
NASA/Ames Research Center, Moffett Field, Mountain View, CA, USA

M.E. Hagan
National Center for Atmospheric Research, Boulder, CO, USA

C.J. Hahn
University of Arizona, Tucson, AZ, USA

J.D. Haigh
Blackett Laboratory, Imperial College London, London, UK

G.J. Hakim
University of Washington, Seattle, WA, USA

K. Hamilton
University of Hawaii, Honolulu, HI, USA

T.F. Hanisco
Harvard University, Cambridge, MA, USA

J.W. Hannigan
National Center for Atmospheric Research, Boulder, CO, USA

R.M. Hardesty
NOAA Environmental Technology Laboratory, Boulder, CO, USA

J.E. Harries
Imperial College of Science, Technology and Medicine, Blackett Laboratory, London, UK

N.R.P. Harris
University of Cambridge, Cambridge, UK

R.G. Harrison
The University of Reading, Reading, UK

D.L. Hartmann
University of Washington, Seattle, WA, USA

F. Hase
Institute of Meteorology and Climate Research (IMK), Karlsruhe Institute of Technology, Karlsruhe, Germany

- L. Hasse
Universität Kiel, Kiel, Germany
- S. Hastenrath
University of Wisconsin, Madison, WI, USA
- P. Haynes
University of Cambridge, Cambridge, UK
- J. Haywood
Met Office, Exeter, UK
- J.M. Hemsley
National Data Buoy Center, Stennis Space Center, MS, USA
- P. Herckes
Arizona State University, Tempe, AZ, USA
- P. Hess
National Center for Atmospheric Research, Boulder, CO, USA
- D.J. Hofmann[†]
NOAA Climate Monitoring and Diagnostics Laboratory, Boulder, CO, USA
- J.R. Holton
University of Washington, Seattle, WA, USA
- A.A.M. Holtslag
Wageningen University, Wageningen, The Netherlands
- D. Hooper
Science & Technology Facilities Council (STFC), Didcot, UK
- D.D. Houghton
University of Wisconsin-Madison, Madison, WI, USA
- S. Houweling
SRON Netherlands Institute for Space Research, Utrecht, The Netherlands
- X.-M. Hu
University of Oklahoma, Norman, OK, USA
- R.X. Huang
Woods Hole Oceanographic Institution, Woods Hole, MA, USA
- X. Huang
University of Michigan, Ann Arbor, MI, USA
- Y.-H. Huang
National Taiwan University, Taipei, Taiwan
- C. Hughes
University of York, York, UK
- B.A. Hungate
Northern Arizona University, Flagstaff, AZ, USA
- J.C.R. Hunt
University College London, London, UK
- L.A. Hunt
Science Systems and Applications Incorporated, Hampton, VA, USA
- J.W. Hurrell
National Center for Atmospheric Research, Boulder, CO, USA
- S. Ismail
Science Directorate, NASA Langley Research Center, Hampton, VA, USA
- C. Jakob
Monash University, VIC, Australia
- S.G. Jennings
National University of Ireland, Galway, Ireland
- P. Jenniskens
SETI Institute, Moffett Field, CA, USA
- X. Jiang
University of California, Los Angeles, CA, USA
- T.D. Jickells
University of East Anglia, Norwich, UK
- R.H. Johnson
Colorado State University, Fort Collins, CO, USA
- A. Joly
Centre National de Recherches Météorologiques – Groupe d'étude de l'Atmosphère Météorologique, Météo-France and CNRS, Toulouse, France
- P.D. Jones
Climatic Research Unit, University of East Anglia, Norwich, UK
- M. Jukes
University of Oxford, Oxford, UK
- R.A. Kahn
NASA Goddard Space Flight Center, Greenbelt, MD, USA
- M. Kalberer
University of Cambridge, Cambridge, UK
- L. Kantha
University of Colorado, Boulder, CO, USA
- M.C. Kelley
Cornell University, Ithaca, NY, USA

[†] Deceased.

D. Keyser
*University at Albany, State University of New York,
Albany, NY, USA*

S.Q. Kidder
*Colorado State University, Fort Collins,
CO, USA*

K.-Y. Kim
Seoul National University, Seoul, Korea

M.D. King
University of Colorado, Boulder, CO, USA

C. Knight
*National Center for Atmospheric Research, Boulder,
CO, USA*

N. Knight
*National Center for Atmospheric Research, Boulder,
CO, USA*

K.H. Knopfmeier
*University of Oklahoma; and National Oceanic and
Atmospheric Administration, Norman, OK, USA*

J.A. Knox
University of Georgia, Athens, GA, USA

T.R. Knutson
*NOAA Geophysical Fluid Dynamics Laboratory,
Princeton, NJ, USA*

G.W. Koch
Northern Arizona University, Flagstaff, AZ, USA

Y. Kondo
The University of Tokyo, Tokyo, Japan

A. Korolev
*Meteorological Service of Canada, Toronto, ON,
Canada*

A. Lacis
*Goddard Institute for Space Studies, New York,
NY, USA*

A. Laing
*National Center for Atmospheric Research, Boulder,
CO, USA*

D. Lal
*Scripps Institution of Oceanography, La Jolla,
CA, USA*

J.F. Lamarque
*National Center for Atmospheric Research, Boulder,
CO, USA*

D. Lamb
*The Pennsylvania State University, University Park,
PA, USA*

T.P. Lane
*The University of Melbourne, Melbourne, VIC,
Australia*

A.L. Lang
*University of Albany – State University of New York,
Albany, NY, USA*

M.F. Larsen
Clemson University, Clemson, SC, USA

K.-M. Lau
*NASA/Goddard Space Flight Center, Greenbelt, MD,
USA*

M.A. LeMone
*National Center for Atmospheric Research, Boulder,
CO, USA*

S.S. Leroy
*Harvard School of Engineering and Applied Sciences,
Cambridge, MA, USA*

P.F. Lester
San Jose State University, San Jose, CA, USA

J.P. Li
*Institute of Atmospheric Physics, Chinese Academy of
Sciences, Beijing, China*

M. Lippmann
New York University, Tuxedo, NY, USA

B. Liu
North Carolina State University, Raleigh, NC, USA

G. Liu
Florida State University, Tallahassee, FL, USA

W.T. Liu
*California Institute of Technology, Pasadena,
CA, USA*

N.G. Loeb
NASA Langley Research Center, Hampton, VA, USA

U. Lohmann
ETH Zurich, Zürich, Switzerland

M. López-Puertas
*Instituto de Astrofísica de Andalucía, CSIC, Granada,
Spain*

A.C. Lorenc
The Met Office, Bracknell, Berkshire, UK

J. Lu
*Pacific Northwest National Laboratory, Richland, WA,
USA*

J.C. Luvall
*National Space Science and Technology Center,
Marshall Space Flight Center, Huntsville, AL, USA*

- W.A. Lyons
FMA Research Inc., Fort Collins, CO, USA
- R.A. Madden
National Center for Atmospheric Research, Boulder, CO, USA
- S. Madronich
National Center for Atmospheric Research, Boulder, CO, USA
- L. Mahrt
Oregon State University, Corvallis, OR, USA
- A.J. Majda
New York University, New York, NY, USA
- D. Mansfield
National Meteorological Center, Bracknell, UK
- F.D. Marks
Hurricane Research Division, Miami, FL, USA
- D.G. Martinson
Columbia University, Palisades, NY, USA
- P.J. Mason
Met Office, Bracknell, UK
- T. Matsui
NASA/Goddard Space Flight Center, Greenbelt, MD, USA; and University of Maryland, College Park, MD, USA
- A. McCulloch
University of Bristol, Bristol, UK
- M.E. McIntyre
University of Cambridge, Cambridge, UK
- R. McKenzie
National Institute for Water and Atmospheric Research, Lauder, Central Otago, New Zealand
- P.H. McMurry
University of Minnesota, Minneapolis, MN, USA
- R.D. McPeters
NASA Goddard Space Flight Center, Greenbelt, MD, USA
- C.R. Mechoso
University of California, Los Angeles, CA, USA
- Z. Meng
Peking University, Beijing, China
- P.M. Midgley
M & D Consulting, Leinfelden Metternich, Germany
- M. Miller
European Centre for Medium-Range Weather Forecasts, Reading, UK
- P. Minnis
Science Directorate, NASA Langley Research Center, Hampton, VA, USA
- M. Mishchenko
Goddard Institute for Space Studies, New York, NY, USA
- N.J. Mitchell
The University of Bath, Bath, UK
- M.G. Mlynarczyk
NASA Langley Research Center, Hampton, VA, USA
- C.-H. Moeng
National Center for Atmospheric Research, Boulder, CO, USA
- F. Molteni
Abdus Salam International Centre for Theoretical Physics, Trieste, Italy
- M. Moncrieff
National Center for Atmospheric Research, Boulder, CO, USA
- D.W. Moore
Pacific Marine Environmental Laboratory, Seattle, WA, USA
- G.K. Moortgat
Max-Planck-Institute for Chemistry, Mainz, Germany
- R. Müller
Institute for Energy and Climate Research (IEK-7), Forschungszentrum Jülich, Jülich, Germany
- J. Munthe
IVL Swedish Environmental Research Institute, Göteborg, Sweden
- R.S. Nerem
University of Colorado, Boulder, CO, USA
- P.A. Newman
NASA Goddard, Space Flight Center, Greenbelt, MD, USA
- H. Nguyen
Centre for Australian Weather and Climate Research, Bureau of Meteorology, Melbourne, VIC, Australia
- N. Nicholls
Bureau of Meteorology Research Centre, Melbourne, VIC, Australia
- F.T.M. Nieuwstadt
Delft University of Technology, Delft, The Netherlands
- S. Nigam
University of Maryland, College Park, MD, USA

L. Nkemdirim
University of Calgary, Calgary, AB, Canada

G.R. North
Texas A&M University, College Station, TX, USA

D.R. Novak
Weather Prediction Center, College Park, MD, USA

A. O'Neill
University of Reading, Earley Gate, Reading, UK

J. Oberheide
Clemson University, Clemson, SC, USA

T.E. Osterkamp
University of Alaska, Fairbanks, AK, USA

R.D. Palmer
University of Oklahoma, Oklahoma, OK, USA

T.R. Parish
University of Wyoming, Laramie, WY, USA

D.J. Parker
University of Leeds, Leeds, UK

M.D. Parker
North Carolina State University, Raleigh, NC, USA

G.K. Parks
University of Washington, Seattle, WA, USA

W.R. Pendleton
Utah State University, Logan, UT, USA

J. Perlwitz
University of Colorado, Boulder, CO, USA

R.L. Pfeffer
Florida State University, Tallahassee, FL, USA

R.A. Pielke, Sr.
University of Colorado at Boulder, CO, USA

R. Pinnick
US Army Research Laboratory, Adelphi, MD, USA

J.M.C. Plane
University of Leeds, Leeds, UK

C.M.R. Platt
Colorado State University, Fort Collins, CO, USA

R. Plougonven
Ecole Polytechnique, Palaiseau, France

M.K. Politovich
National Center for Atmospheric Research, Boulder, CO, USA

L.M. Polvani
Columbia University, New York, NY, USA

J.-P. Pommereau
LATMOS, CNRS, Guyancourt, France

J.A. Pyle
University of Cambridge, Cambridge, UK

D.A. Quattrochi
National Space Science and Technology Center, Marshall Space Flight Center, Huntsville, AL, USA

S. Quiring
Texas A&M University, College Station, TX, USA

M. Ramamurthy
University Corporation for Atmospheric Research, Boulder, CO, USA

A.L. Rangno (Retiree)
University of Washington, Seattle, WA, USA

R.M. Rauber
University of Illinois at Urbana-Champaign, Urbana, IL, USA

A.R. Ravishankara
Colorado State University, Fort Collins, CO, USA

I.A. Renfrew
University of East Anglia, Norwich, UK

N.O. Renno
University of Michigan, Ann Arbor, MI, USA

R. Reynolds
University of Reading, Reading, UK

P.B. Rhines
University of Washington, Seattle, WA, USA

A.D. Richmond
National Center for Atmospheric Research, Boulder, CO, USA

D.L. Rickman
National Space Science and Technology Center, Marshall Space Flight Center, Huntsville, AL, USA

R.G. Roble
National Center for Atmospheric Research, Boulder, CO, USA

A. Robock
Rutgers University, New Brunswick, NJ, USA

F. Rohrer
Forschungszentrum Jülich, Jülich, Germany

K.H. Rosenlof
Earth System Research Laboratory, National Oceanic and Atmospheric Administration, Boulder, CO, USA

- M.W. Rotach
University of Innsbruck, Innsbruck, Austria
- R. Rotunno
National Center for Atmospheric Research, Boulder, CO, USA
- I. Roulstone
University of Surrey, Guildford, UK
- J. Rudolph
York University, Toronto, ON, Canada
- L.M. Russell
Scripps Institution of Oceanography, University of California, San Diego, La Jolla, CA, USA
- C.L. Sabine
NOAA Pacific Marine Environmental Laboratory, Seattle, WA, USA
- D.A. Salstein
Atmospheric and Environmental Research, Inc., Lexington, MA, USA
- C. Schär
Atmospheric and Climatic Science ETH, Zürich, Switzerland
- U. Schmidt
Johann Wolfgang Goethe-University, Frankfurt am Main, Institut für Meteorologie und Geophysik, Frankfurt am Main, Germany
- M.R. Schoeberl
Science and Technology Corporation, Lanham, MD, USA
- D.M. (David) Schultz
University of Manchester, Manchester, UK
- J. Schulz
Meteorological Institute, University of Bonn, Bonn, Germany
- J.H. Seinfeld
California Institute of Technology, Pasadena, CA, USA
- C.J. Senff
CIRES, University of Colorado; and NOAA Earth System Research Laboratory, Boulder, CO, USA
- M.C. Serreze
University of Colorado, Boulder, CO, USA
- G.E. Shaw
Geophysical Institute, University of Alaska, Fairbanks, AK, USA
- G.G. Shepherd
York University, Toronto, ON, Canada
- T.G. Shepherd
University of Toronto, Toronto, ON, Canada
- R. Sica
The University of Western Ontario, London, ON, Canada
- H.B. Singh
NASA Ames Research Center, Mountain View, CA, USA
- J.A. Sippel
National Aeronautics and Space Administration (NASA), Greenbelt, MD, USA
- J. Slingo
University of Reading, Reading, UK
- H.G.J. Smit
Research Centre Jülich, Jülich, Germany
- A.K. Smith
National Center for Atmospheric Research, Boulder, CO, USA
- J.B. Smith
Harvard University, Cambridge, MA, USA
- L.A. Smith
London School of Economics, Centre for the Analysis of Time Series, London, UK
- R.B. Smith
Yale University, New Haven, CT, USA
- I.N. Sokolik
Georgia Institute of Technology, Atlanta, GA, USA
- S.C. Solomon
National Center for Atmospheric Research, Boulder, CO, USA
- J. Sommar
Göteborg University, Göteborg, Sweden
- P.J. Sousounis
AIR Worldwide, Boston, MA, USA
- K. Stamnes
Stevens Institute of Technology, Hoboken, NJ, USA
- S.N. Stechmann
University of Wisconsin–Madison, Madison, WI, USA
- E.J. Steig
University of Washington, Seattle, WA, USA
- O. Stein
IEK 8: Troposphere, Research Center Juelich, Juelich, Germany

D.J. Stensrud
*National Oceanic and Atmospheric Administration,
Norman, OK, USA*

R.S. Stolarski
Johns Hopkins University, Baltimore, MD, USA

S.E. Strahan
*Universities Space Research Association, NASA Goddard
Space Flight Center, Greenbelt, MD, USA*

M. Sturm
*US Army Cold Regions Research & Engineering
Laboratory-Alaska, Fort Wainwright, AL, USA*

S.T. Suess
*NASA Marshall Space Flight Center, Huntsville, AL,
USA*

P.P. Sullivan
*National Center for Atmospheric Research, Boulder,
CO, USA*

M.P. Sulzer
Arecibo Observatory, Arecibo, PR, USA

E.S. Takle
Iowa State University, Ames, IA, USA

W.-K. Tao
*NASA/Goddard Space Flight Center, Greenbelt, MD,
USA*

M.J. Taylor
Utah State University, Logan, UT, USA

P.K. Taylor
*Southampton Oceanography Centre, Southampton,
UK*

V.P. Tchakerian
Texas A&M University, College Station, TX, USA

G.E. Thomas
University of Colorado, Boulder, CO, USA

L.W. Thomason
NASA Langley Research Center, Hampton, VA, USA

D.W.J. Thompson
Colorado State University, Fort Collins, CO, USA

D.J. Thomson
Met Office, Bracknell, UK

J. Thuburn
University of Exeter, Exeter, UK

R.A. Tomas
University of Colorado – Boulder, Boulder, CO, USA

D. Toohey
University of Colorado Boulder, Boulder, CO, USA

L. Travis
*Goddard Institute for Space Studies, New York,
NY, USA*

K.E. Trenberth
*National Center for Atmospheric Research, Boulder,
CO, USA*

B.T. Tsurutani
Jet Propulsion Laboratory, Pasadena, CA, USA

J. Turner
British Antarctic Survey, Cambridge, UK

H. Turtiainen
Vaisala Company, Helsinki, Finland

C. Twohy
Oregon State University, Corvallis, OR, USA

T. Uttal
NOAA, Boulder, CO, USA

R.H. Varney
Cornell University, Ithaca, NY, USA

G. Vaughan
University of Manchester, Manchester, UK

W.W. Vaughan
*University of Alabama in Huntsville, Huntsville,
AL, USA*

G.A. Vecchi
GFDL/NOAA, Princeton, NJ, USA

A. Venkatram
*University of California – Riverside, Riverside,
CA, USA*

J.-P. Vernier
*Science Systems and Applications, Inc., Hampton,
VA, USA*

R.K. Vincent
*Bowling Green State University, Bowling Green,
OH, USA*

R. von Glasow
University of East Anglia, Norwich, UK

A. Wahner
Forschungszentrum Jülich, Jülich, Germany

R.M. Wakimoto
*National Center for Atmospheric Research, Boulder,
CO, USA*

D.E. Waliser
California Institute of Technology, Pasadena, CA, USA

J.M. Wallace
University of Washington, Seattle, WA, USA

- B. Wang
University of Hawaii, Honolulu, HI, USA
- Y. Wang
University of Hawaii at Manoa, Honolulu, HI, USA
- Z. Wang
University of Illinois at Urbana-Champaign, Urbana, IL, USA
- B. Ward
Public Works and Natural Resources, Longmont, CO, USA
- S. Warren
University of Washington, Seattle, WA, USA
- J. Waters
California Institute of Technology, Jet Propulsion Laboratory, Pasadena, CA, USA
- R.P. Wayne
University of Oxford, Oxford, UK
- S. Weart
Center for History of Physics, American Institute of Physics, College Park, MD, USA
- P.J. Webster
Georgia Institute of Technology, Atlanta, GA, USA
- P.J. Webster
University of Colorado – Boulder, Boulder, CO, USA
- W.F. Weeks
University of Alaska Fairbanks, Fairbanks, AK, USA
- M.L. Weisman
National Center for Atmospheric Research, Boulder, CO, USA
- S.J. Weiss
National Oceanic and Atmospheric Administration, Norman, OK, USA
- N.C. Wells
University of Southampton, Southampton, UK
- H. Wernli
Institute for Atmospheric and Climate Science, Swiss Federal Institute of Technology (ETH), Zürich, Switzerland
- T. Wetter
Johann Wolfgang Goethe-University, Frankfurt am Main, Institut für Meteorologie und Geophysik, Frankfurt am Main, Germany
- M.C. Wheeler
Centre for Australian Weather and Climate Research, Bureau of Meteorology, Melbourne, VIC, Australia
- A.A. White
University of Surrey, Guildford, UK
- D.N. Whiteman
NASA Goddard Space Flight Center, Greenbelt, MD, USA
- B.A. Wielicki
NASA Langley Research Center, Hampton, VA, USA
- E.R. Williams
Massachusetts Institute of Technology, Cambridge, MA, USA
- M.L. Wong
California Institute of Technology, Pasadena, CA, USA
- N. Wood
Met Office, Exeter, UK
- R. Wood
University of Washington, Seattle, WA, USA
- S.N. Wren
University of Toronto, Toronto, ON, Canada
- C.-C. Wu
National Taiwan University, Taipei, Taiwan
- D.J. Wuebbles
University of Illinois, Urbana, IL, USA
- L. Xie
North Carolina State University, Raleigh, NC, USA
- P. Yang
Texas A&M University, College Station, TX, USA
- S. Yang
NOAA/NWS/NCEP, Climate Prediction Center, Camp Springs, MD, USA
- J.A. Young
University of Wisconsin, Madison, WI, USA
- Z. Yu
College of Hydrology and Water Resources, Hohai University, Nanjing, China; and University of Nevada Las Vegas, Las Vegas, NV, USA
- Y.L. Yung
California Institute of Technology, Pasadena, CA, USA

S.E. Yuter
*North Carolina State University, Raleigh,
NC, USA*

S. Yvon-Lewis
*Texas A&M University, College Station,
TX, USA*

D. Zardi
University of Trento, Trento, Italy

S.E. Zebiak
*International Research Institute for Climate Prediction,
Palisades, NY, USA*

M.D. Zelinka
*Lawrence Livermore National Laboratory, Livermore,
CA, USA*

C. Zhang
University of Miami, Miami, FL, USA

F. Zhang
*Pennsylvania State University, University Park,
PA, USA*

M. Zhang
*Stony Brook University, Stony Brook,
NY, USA*

PREFACE TO THE FIRST EDITION

A half century ago the American Meteorological Society published the *Compendium of Meteorology*, which in a single volume of 1334 pages summarized the state of understanding of the atmosphere at that time. A perusal of the contents of that volume indicates that although a broad range of topics was covered, the vast bulk of the volume was devoted to traditional meteorological topics such as atmospheric dynamics, cloud physics, and weather forecasting. Barely 4 percent of the volume was devoted to articles related to atmospheric chemistry or air pollution and, of course, none of the volume was devoted to techniques such as satellites and remote sensing.

As Sir John Mason aptly notes in his foreword to the present work, the atmospheric sciences have expanded in scope enormously over the past 50 years. Topics such as atmospheric chemistry and global climate change, of only marginal interest 50 years ago, are now central disciplines within the atmospheric sciences. Increasingly, developing areas within the atmospheric sciences require students, teachers, and researchers to familiarize themselves with areas far outside their own specialties. This work is intended to satisfy the need for a convenient and accessible references source covering all aspects of atmospheric sciences. It is written at a level that allows undergraduate science and engineering students to understand the material, while providing active researchers with the latest information in the field.

More than 400 scientists, from academia, government, and industry have contributed to the 330 articles in this work. We are very grateful to these authors for their success in providing concise and authoritative summaries of complex subjects. As editors, we have benefited from the chance to learn from these articles, and we believe that all students and active scientists who want to increase their knowledge of the atmosphere will benefit enormously from access to this work.

We are also grateful to the 31 members of the Editorial Advisory Board who have guided us in our coverage of the very broad range of topics represented in this encyclopedia. Their willingness to suggest topics and authors, and to carefully review draft articles has contributed significantly to our success.

The production of this multivolume encyclopedia would not have been possible without the dedicated work of the staff of the Major Reference Works group at Academic Press. We are especially grateful to the Major Reference Work Development Manager, Colin McNeil, who has worked closely with us during the entire process. Finally, we appreciate the liberal use of color figures in the printed encyclopedia.

James R Holton, Judith A Curry, and John Pyle

This page intentionally left blank

PREFACE TO THE SECOND EDITION

Since the publication of the first edition of the *Encyclopedia of Atmospheric Sciences*, significant advances in research have been achieved all across the broad and expanding spectrum of the field and related disciplines. In particular, climate science with primary input from the atmospheric research emerges as a new field and integrator of interlocking peripheral disciplines over the last decade. These events have demanded the solicitation of new and updated articles for the 2003 edition. Some articles from the earlier publication were judged to be of such a fundamental and enduring nature that they did not require modification. But huge amounts of new information from Earth-orbiting satellite observatories have brought much new insight to the field. In addition there are new findings in many areas such as the latest simulations of meteorological and climatic processes of interest as well as simulations and observations of the composition and interaction of the field's chemical constituents. While interest in the ozone hole and its ramifications may have reached a plateau, ever more understanding of the stratosphere and its role in climate change emerges. The study of past climates provides new means of testing climate models and theories. In weather prediction we see new progress on how data are to be better assimilated for much improved initialization of the forecast model leading to the promise of more accurate predictions of severe weather and tropical cyclones over longer lead times. These are just a few of the new features of the second edition.

The editors of the second edition are greatly indebted to our predecessors in the first edition. They set the outline of topics and solicited the original authors, while establishing a high standard for the content of this publication. In many cases we decided to reprint those articles or request only minor updates. Nevertheless, many articles in this edition are entirely original, based on which we also made significant reorganization of the content. We are proud of our product and hope it provides the same assistance to students, researchers, and practitioners throughout the science and engineering communities.

Editors of the second edition
Gerald R North
Fuqing Zhang
John Pyle

This page intentionally left blank

EDITOR BIOGRAPHIES



Gerald R North received his PhD in theoretical physics from the University of Wisconsin in 1966. After postdoctoral research at the University of Pennsylvania he became a faculty member in physics at the University of Missouri–St. Louis. He shifted his research focus to climate science research during his sabbatical year at the National Center for Atmospheric Research, where he won the Outstanding Paper Award in 1975. He moved to NASA Goddard Space Flight Center (GSFC) in 1978 where he was awarded the NASA Medal for Research Excellence. During his stay at GSFC, he was the proposer and first study scientist for the Tropical Rainfall Measuring Mission, which was launched in 1997 and is still orbiting in 2014.

He moved to Texas A&M University in 1986 as a university distinguished professor of atmospheric sciences where he served as department head from 1995 to 2003. He has served as editor-in-chief of the *Reviews of Geophysics* and is recognized as one of the most cited authors in geosciences (Web of Science). He has chaired and/or served on a number of national committees and is a Fellow of the American Geophysical Union, American Meteorological Society (AMS) and the American Association for the Advancement of Science, and winner of the Jule Charney Award for Research (AMS).

He has published about 150 refereed papers not including many book chapters and reviews. His books include *Paleoclimatology*, co-authored with Thomas Crowley, and *An Introduction to Atmospheric Thermodynamics* co-authored with Tatiana Erikhimova.

North's interests are focused on the use of mathematical and statistical tools to solve climate problems over a wide range of issues including: analytical solutions of simplified energy balance climate models, use of random field techniques in representing and interpreting climate data and model simulations, detection of deterministic signals in climate change, statistical analysis satellite remote sensing for mission planning and analysis of data, paleoclimate problems using simplified climate models.



John Pyle obtained a BSc in Physics at Durham University before moving to Oxford where he completed a DPhil in Atmospheric Physics, helping to develop a numerical model for stratospheric ozone studies. After a short period at the Rutherford Appleton Laboratory he moved to a lectureship at Cambridge University in 1985. In 2000 he was appointed professor of atmospheric science and since 2007 has been the 1920 professor of physical chemistry. He is a Professorial Fellow at St Catharine's College. He has been a codirector of Natural Environment Research Council's National Centre for Atmospheric Science, where he is currently Chief Scientist.

His research focuses on the numerical modelling of atmospheric chemistry. Problems involving the interaction between chemistry and climate have been addressed; these range from stratospheric ozone depletion to the changing tropospheric oxidizing capacity and have included the environmental impact of aviation, land use change, biofuel technologies, and the hydrogen economy. He has studied palaeochemistry problems as well as the projected atmospheric composition changes during the current century. He has published more than 250 peer reviewed papers.

He played a major role in building an EU stratospheric research programme in the 1990s, coordinating several major field campaigns. He has contributed to all the WMO/UNEP assessments on stratospheric ozone since the early 1980s and is now one of the four international cochairns on

the Scientific Assessment Panel, responsible for these assessments. He was a convening lead author in the IPCC Special report "Safeguarding the ozone layer and the global climate system," published in 2006.

He was elected Fellow of the Royal Society in 2004 and an American Geophysical Union Fellow in 2011. He was awarded the Cambridge ScD degree in 2012. Other honours and awards include membership of Academia Europaea (1993), Royal Society of Chemistry (Interdisciplinary award, 1991, and John Jeyes lectureship, 2008), and the Royal Meteorological Society Adrian Gill Prize, in 2004.



Fuqing Zhang is a professor with tenure in the Department of Meteorology at the Pennsylvania State University, with a joint appointment in the Department of Statistics, along with an endowed position as the E Willard & Ruby S Miller Faculty Fellow at the College of Earth and Mineral Sciences at the Pennsylvania State University. His research interests include atmospheric dynamics and predictability, data assimilation, ensemble forecasting, tropical cyclones, gravity waves, mountain plains and sea-breeze circulations, warm-season convection, and regional-scale climate. He earned his BS and MS in meteorology from Nanjing University, China in 1991 and 1994, respectively, and his PhD in atmospheric science in 2000 from North Carolina State University.

He spent seven years as an assistant and then associate professor at Texas A&M University before coming to Penn State University as a full professor in 2008. In 2000, he spent a year and a half as a postdoctoral fellow at the National Center for Atmospheric Research. He also held various visiting scholarship appointments at various academic and research institutions including the National

Center for Atmospheric Research in Boulder, Colorado; the Navy Research Laboratory in Monterey, California; NOAA/AOML Hurricane Research Division, Miami, Florida; Peking University and Nanjing University, China; the Chinese State Key Laboratory of Severe Weather in Beijing, China; and Laboratoire de Meteorologie Dynamique, École Normale Supérieure in Paris, France.

He has authored/co-authored about 130 peer reviewed journal publications and has given more than 160 keynote speeches or invited talks at various institutions and meetings. He has served as principal investigator/co-principal investigator for 30 federal or state-sponsored research grants. He has chaired/cochaired more than 10 scientific meetings or workshops. He also served on various review or advisory panels for numerous organizations that include National Science Foundation, Office of Naval Research, NASA, NOAA, and National Academies. He has also served as editor of several professional journals including *Monthly Weather Review*, *Science China*, *Atmospheric Science Letter*, *Acta Meteorologica Sinica*, and *Computing in Science & Engineering*.

He has also received numerous awards for his research and service. Notably, in 2007 he received the **Outstanding Publication Award** from the National Center for Atmospheric Research. In 2009, was the sole recipient of the American Meteorological Society's 2009 **Clarence Leroy Meisinger Award** "for outstanding contributions to mesoscale dynamics, predictability, and ensemble data assimilation." Most recently, he received the 2014 American Meteorological Society's **Banner Miller Award** "for valuable insights into incorporating real-time airborne Doppler radar measurements via ensemble data assimilation, leading to improvements in forecasts of tropical cyclone track and intensity."

GUIDE TO USING THE ENCYCLOPEDIA

Structure of the Encyclopedia

The material in the encyclopedia is not arranged by ordinary alphabetical order, but by alphabetical order according to 49 principal topic areas taken to allow all papers belonging to each principal topic to appear together in the same volume. Within each principal subject, article headings are also arranged alphabetically, except where logic dictates otherwise. For example, overview articles appear at the beginning of a section.

There are four features that help you find the topic in which you are interested:

- i. the contents list
- ii. cross-references to other relevant articles within each article
- iii. a full subject index
- iv. contributors

i. Contents List

The contents list, which appears at the front of each volume, lists the entries in the order that they appear in the encyclopedia. It includes both the volume number and the page number of each entry.

ii. Cross-references

All of the entries in the encyclopedia have been cross-referenced. The cross-references, which appear at the end of an article as a See also list, serve four different functions:

- i. To draw the reader's attention to related material in other entries

- ii. To indicate material that broadens and extends the scope of the article
- iii. To indicate material that covers a topic in more depth
- iv. To direct readers to other articles by the same author(s)

Example

The following list of cross-references appears at the end of the article.

See also: **Biogeochemical Cycles:** Biogeochemistry of Iodine. **Stratospheric Chemistry Topics:** HO_x; Halogen Sources, Natural (Methyl Bromide and Related Gases); Halogens; Hydrogen Budget; Overview; Reactive Nitrogen (NO_x and NO_y).

iii. Index

The index includes page numbers for quick reference to the information you are looking for. The index entries differentiate between references to a whole article, a part of an article, and a table or figure.

iv. Contributors

At the start of each volume there is list of the authors who contributed to that volume.

This page intentionally left blank

CLIMATE AND CLIMATE CHANGE

Contents

Overview

Carbon Dioxide

Climate Feedbacks

Climate Prediction: Empirical and Numerical

Climate Variability: Decadal to Centennial Variability

Climate Variability: Nonlinear and Random Effects

Climate Variability: North Atlantic and Arctic Oscillation

Climate Variability: Seasonal and Interannual Variability

Energy Balance Climate Models

Global Impacts of the Madden–Julian Oscillation

Greenhouse Effect

History of Scientific Work on Climate Change

Intergovernmental Panel on Climate Change

Nuclear Winter

Radiative–Convective Equilibrium Climate Models

Volcanoes: Role in Climate

Overview

DL Hartmann, University of Washington, Seattle, WA, USA

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

Climate is defined and the variations of average or expected temperature and precipitation on the Earth as a function of location and season are briefly described. The role of the energy balance and greenhouse effect of the Earth, the distribution of insolation, and the role of oceans and topography in determining climate are explained. Past climate changes are introduced, including much warmer climates millions of years ago and the succession of ice ages in the past several millions of years. The primary determinant of the future climate will likely be the modification of atmospheric composition by human activities.

Introduction

Climate is the composite or generalization of weather conditions of a region as a function of season. It can be expressed in terms of the expected values of meteorological variables such as temperature, precipitation, pressure, humidity, cloudiness, sunshine, and winds. The expected values are usually obtained by averaging observations over a number of years. A complete description of the climate would also include information on the year-to-year variability. Climatology is the scientific study of climate. A complete understanding of climate requires a thorough understanding of the atmosphere and its physical and chemical interactions with the ocean and the land surface.

Life in the ocean and on the land influences the composition of the air, the color of the ocean, and the exchange of energy and moisture between the land surface and the atmosphere.

Climate is important for humanity and life on the Earth, since it has set the context for human evolution and subsequent social, political, and historical developments. In the modern world, it influences agriculture, water resources, human health, and energy use. It continues to play an important role in natural ecology and the interaction of human endeavor with Earth's biological and geochemical resources. Because of the large human population of the Earth and the adoption of technology by societies, humans now have the ability to make relatively rapid changes in Earth's global climate.

Description of Climate

The average surface temperature of the Earth is about 288 K or 15 °C. The global average precipitation is about 1 m year⁻¹. The climate of a given location varies with latitude, altitude, and geographical conditions (Figure 1). Seasonal variations are greater in higher latitudes and in continental rather than maritime areas (Figure 2). Oceans have a large capacity to store heat, so that seasonal variations in surface temperature are tempered by heat exchange with the ocean.

The distribution of precipitation is more complex. Precipitation is greatest near the equator, reduced in subtropical latitudes (15°–30°), and increases again in middle latitudes (Figure 3). The rainfall maximum near the equator is associated with the general convergence of wind at low levels and the

high water vapor content of tropical air. This precipitation maximum moves north and south across the equator following the position of the Sun relative to the equator. The very low precipitation values that occur in the tropics are associated with downward motion and divergence of the surface winds. Precipitation maxima occur in midlatitudes where cyclonic storms are frequent and produce heavy rainfall. Two relatively strong precipitation maxima in the Northern Hemisphere are associated with the storm tracks over the western Pacific and Atlantic Oceans.

Annual variations of temperature and precipitation vary greatly with location. In the tropics, temperature variations are usually modest, but precipitation can vary from complete drought in some seasons to torrential rains in another. Land areas that are downwind of large water masses generally have

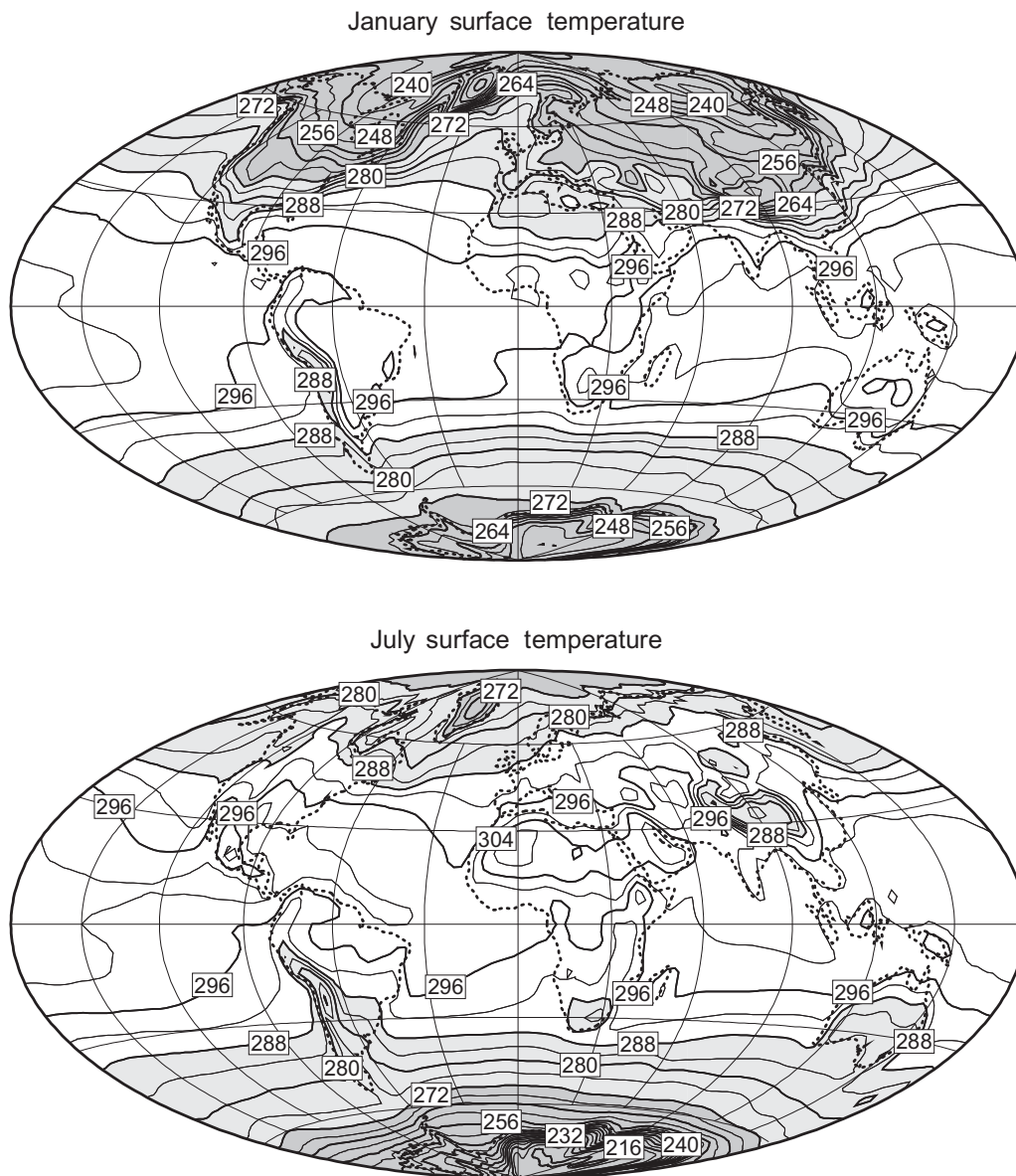


Figure 1 Global surface temperature maps January and July and January minus July. Contour interval is 4 K. Data are 2 m temperatures from the analysis products of The European Centre for Medium-Range Weather Forecasts for the period 1985–94.

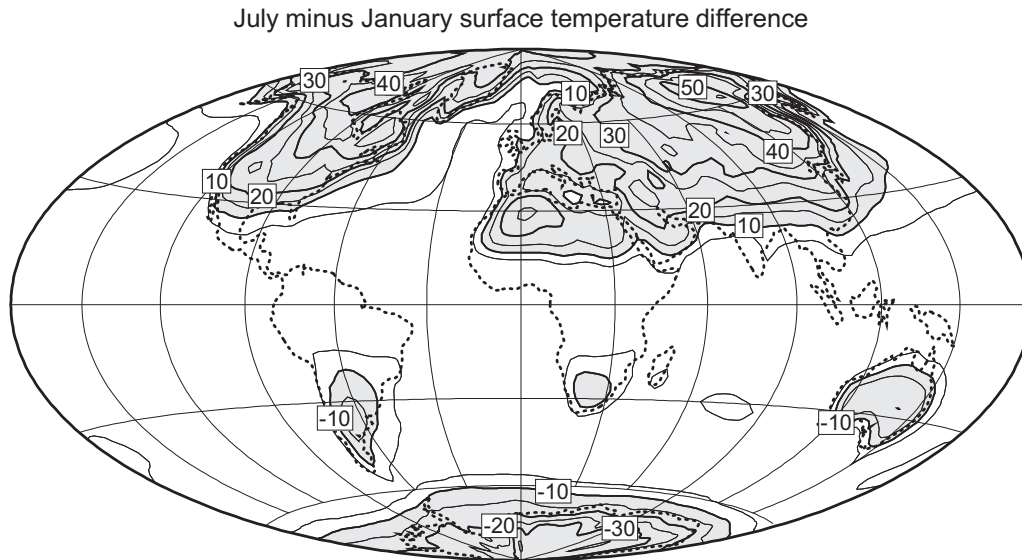


Figure 2 July minus January difference in surface air temperature. Contour interval is 5 K.

smaller seasonal temperature variations than locations in the center of large continents. Good examples of maritime and continental climates are Seattle and Minneapolis (**Figure 4**). Minneapolis has a much larger seasonal variation of temperature. The variations of precipitation in these two regions are also very different. Seattle has a wintertime precipitation maximum associated with the midlatitude storm activity. Minneapolis has a summertime rainfall maximum that comes largely from thunderstorms. The annual temperature variation of New York City is influenced by the large continental upwind to the west and so is quite large. New York gets both wintertime storm precipitation and summertime convective precipitation, so that monthly precipitation is more nearly evenly spread over the year. The annual total precipitation in New York City (120 cm) is greater than that of Seattle (95 cm).

Global Energy Balance

The mean temperature at the surface of the Earth is determined by the flow of energy through the climate system, which consists of the atmosphere, ocean, and land surface. The source of energy for the planet is radiation emitted by the Sun. Although Earth has an internal energy source from radioactive decay, this source is too small to influence the global mean surface temperature. At the average position of the Earth in its orbit about the Sun, the Sun provides about 1360 W m^{-2} of total solar irradiance. The solar irradiance varies as the inverse square of the distance from the Sun. Because the Earth is approximately spherical in shape, the ratio of its shadow area to its surface area is 4. So to get the solar energy available per unit of surface area, the total solar irradiance must be divided by 4, yielding about 340 W m^{-2} or the energy equivalent of about 3.4 100-W light bulbs for each square meter of the Earth. Averaged over the whole Earth, about 70% of this flux is absorbed by the Earth and about 30% is reflected back to space without heating the Earth. The fraction that is reflected is called

the albedo, from a Greek word meaning whiteness. The energy that is absorbed is converted into heat and later emitted back to space as thermal infrared radiation.

The simplest model for the global mean temperature T of the Earth equates the absorbed solar radiation with the emitted terrestrial radiation, assuming that the Earth emits like a blackbody.

$$\frac{S}{4}(1 - \alpha) = \sigma T^4 \quad [1]$$

where S is the total solar irradiance, α is the albedo, and $\sigma = 5.67 \times 10^{-8} \text{ W m}^{-2} \text{ K}^{-4}$ is the Stefan-Boltzmann blackbody emission constant. If this equation is solved for the blackbody emission temperature of the Earth, a value of about 255 K, or -18°C , is obtained which is much less than the global mean surface temperature of 288 K, or 15°C . The emission temperature of the Earth is equal to the average temperature of the atmosphere about 5 km above the surface, and indeed, most of the energy that the Earth emits to space is emitted from the atmosphere rather than the surface.

A diagram showing the energy flow through the global climate system is given in **Figure 5**. Although nearly half of the solar energy that enters the climate system is absorbed at the surface, very little of the infrared radiation emitted from Earth's surface escapes directly to space. The atmosphere absorbs most of the infrared radiation emitted from the surface, primarily by water vapor, clouds, and carbon dioxide gas. Moreover, the atmosphere emits infrared radiation downward toward the surface, and the energy supplied to the surface by this downward infrared flux is nearly twice as great as the amount of energy supplied to the surface from the Sun. The transparency of the atmosphere to solar radiation combined with the opaqueness of the atmosphere to infrared radiation result in a heating effect that raises the surface temperature above the value that it would have in the absence of the atmosphere. This is often called the atmospheric

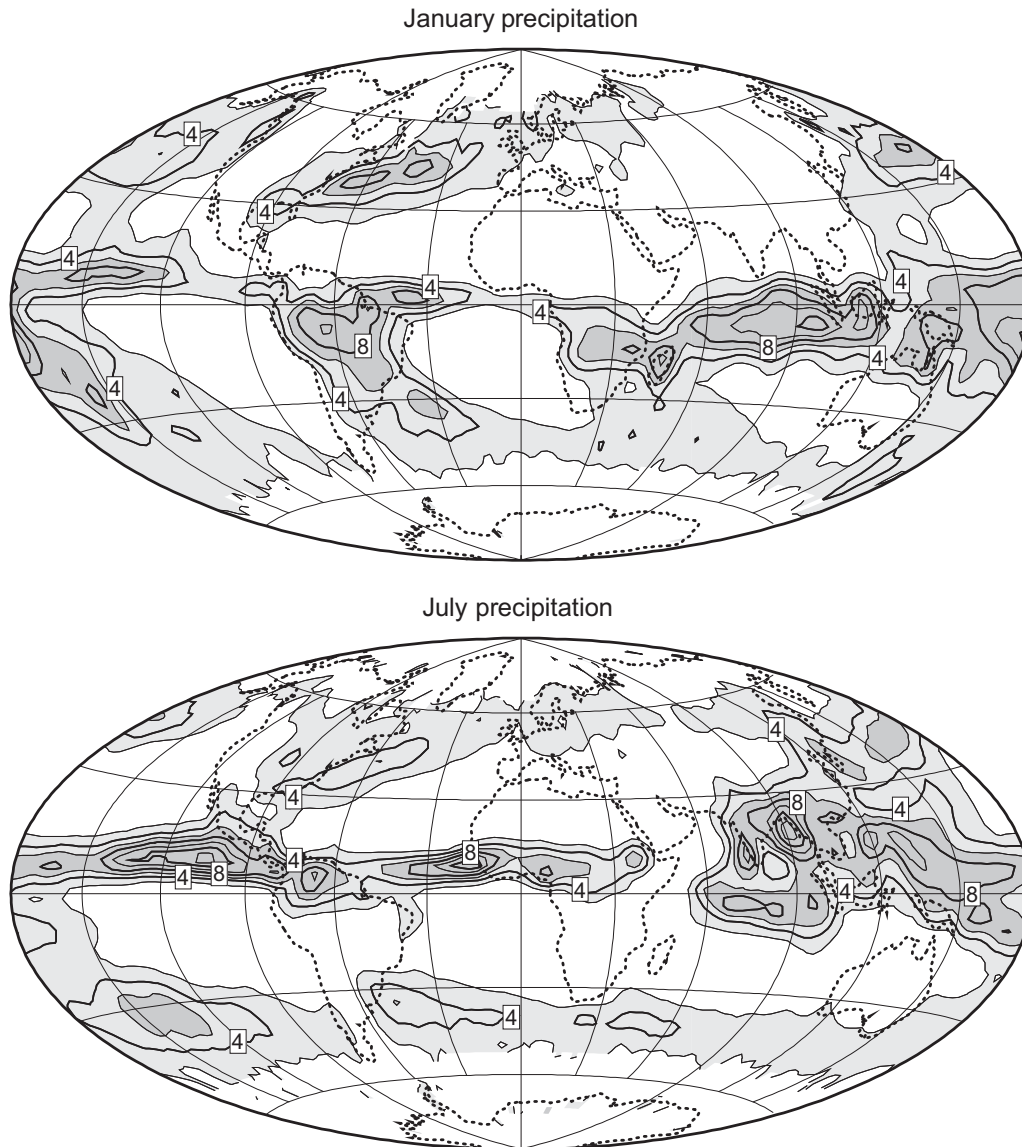


Figure 3 Global precipitation maps for January and July. Contour interval 2 cm month^{-1} . Data are from the Global Precipitation Climatology Project.

greenhouse effect. The greenhouse effect also reduces the amplitude of the diurnal cycle in surface temperature, the daily variation of temperature associated with the rising and setting of the Sun. Because the downward longwave emission from the atmosphere continues after the Sun sets, the fall of temperature after sunset is much less than it would be in the absence of the atmosphere. At high altitudes, the daily variation of temperature is greater because the mass of atmosphere above the surface is less and the greenhouse effect is therefore reduced. Because water vapor is the principle greenhouse gas, the strength of the greenhouse effect increases with temperature and relative humidity.

A net radiative input heats the surface of the Earth, and the net effect of radiation on the atmosphere is to cool it at the rate of about $1.5^\circ\text{C day}^{-1}$. Heat is transferred from the surface to the atmosphere by atmospheric motions that carry heat and moisture upward. The release of latent heat of vaporization

stored in water vapor during condensation is the largest heating term in the atmosphere and offsets atmospheric cooling by radiation emission. The global mean precipitation rate of 1 m year^{-1} corresponds to an atmospheric heat input of 80 W m^{-2} .

Continuous heating of the surface and cooling of the atmosphere by radiative processes drive convective instability and the hydrologic cycle within the Earth's climate system. Evaporation is greater than precipitation over the world's oceans. The excess water is transported to the land areas where the average precipitation exceeds the evaporation. The excess of precipitation over evaporation in land areas returns to the oceans as runoff in rivers (Figure 6). The supply of water from the ocean supports life on the land, and the return of minerals and other elements of life to the ocean in rivers supports life in the ocean. The hydrologic cycle is also a key element of the chemical and biological cycling of carbon through the Earth

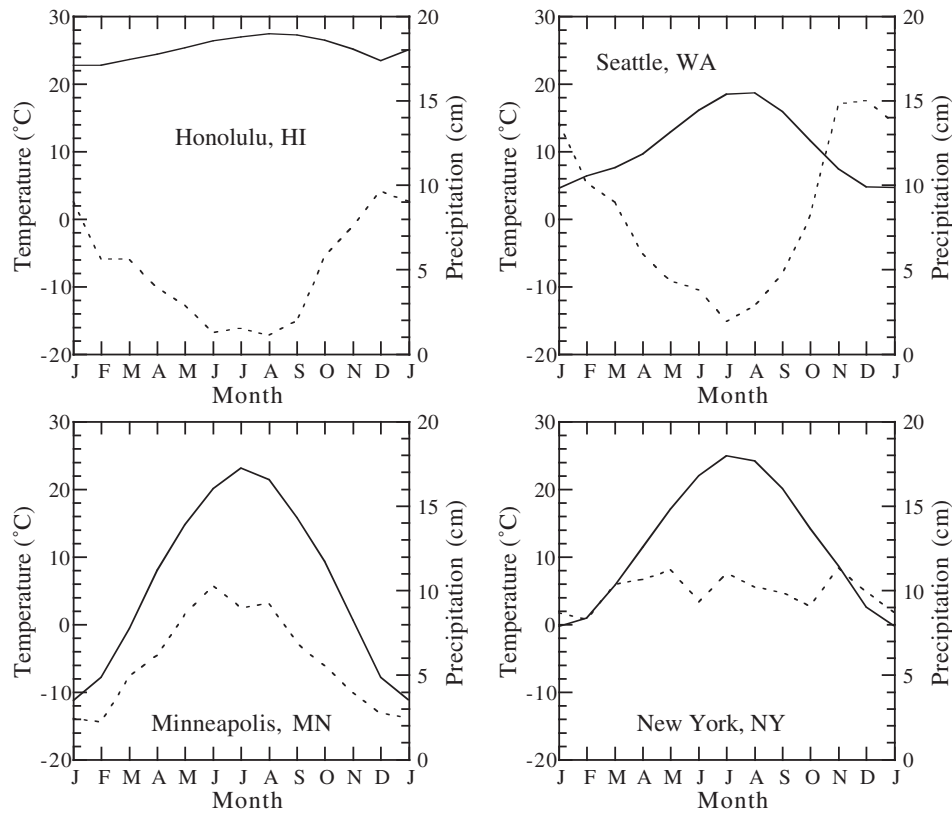


Figure 4 Annual cycles of climatological monthly temperature (solid line) and precipitation (dashed line) at four cities.

system that regulates atmospheric carbon dioxide on timescales of millennia and longer.

Seasonal and Latitudinal Insolation Variations

The seasonal and latitudinal distribution of insolation is an important determinant of climate. The instantaneous insolation per unit of surface area is given by the total solar irradiance times the cosine of the solar zenith angle, the angle between local vertical and the Sun. The daily average insolation available at the top of the atmosphere is given as a function of latitude and season in [Figure 7](#). The insolation decreases with increasing latitude, except in summer, and the annual variation of insolation is greatest near the poles, where 6 months of darkness alternate with 6 months of daylight. In polar regions during summer, the available insolation is greater than that at the equator, because, although the Sun is near the horizon, it shines 24 h a day at the poles during the summer half-year. The insolation available during Southern Hemisphere summer is about 7% greater than that available during Northern Hemisphere summer, because Earth's orbit is not perfectly circular and at the present time, the Earth is closer to the Sun during Southern Hemisphere summer.

On timescales of millennia, the distribution of insolation with latitude and season changes as the parameters of Earth's orbit vary in response to dynamical interactions with the orbits of other planets. The tilt of the axis of rotation with respect to the plane of Earth's orbit varies with a period of

41 000 years. In the last few million years the tilt angle, or obliquity, has varied between 22° and 24.5° . It is currently 23.45° . The eccentricity or degree to which Earth's orbit differs from a perfect circle varies with periods of 100 000 and 400 000 years. The season when the Earth makes its closest approach to the Sun, the perihelion of the orbit, varies with periods near 20 000 years.

Transport of Energy and the Circulation of the Atmosphere and Ocean

Much more solar energy is available to heat tropical latitudes than high latitudes. In the annual average, net radiative energy is input into tropical latitudes and high latitudes lose energy ([Figure 8](#)). The heating of the tropics and cooling of the polar regions drive circulations in the atmosphere and ocean that transport heat from tropical to polar regions. The ocean and the atmosphere have similarly important roles in poleward transport, with the ocean transport larger in subtropical latitudes (20° N and 20° S) and the atmosphere dominating at middle and high latitudes (50° N and 50° S). Transport in the atmosphere comprises latent, thermal, and potential energy transports. In tropical latitudes, upward motion near the equator and downward motion in subtropical latitudes are prominent features of the atmospheric circulation. At low levels, this requires equatorward winds that are turned westward by Earth's rotation to form the trade winds. At upper levels, the poleward flow is turned eastward by Earth's rotation

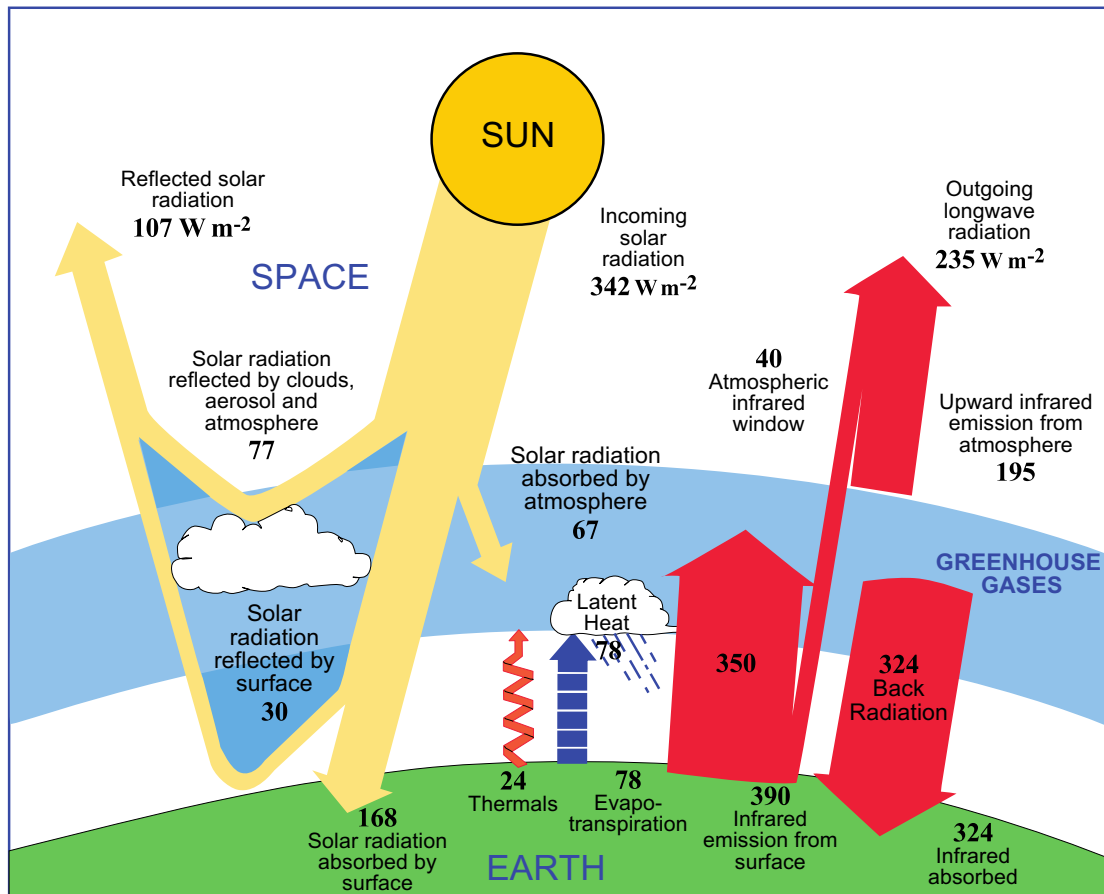


Figure 5 Energy flow through the global climate system.

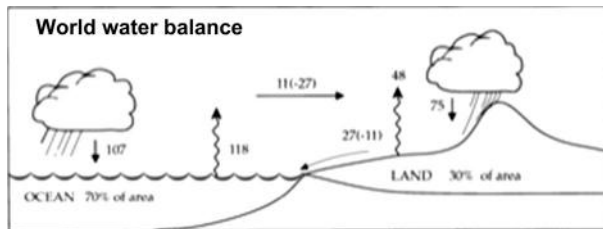


Figure 6 Global cycling of water between ocean and land. Reproduced from Hartmann, D.L., 1994, *Global Physical Climatology*, Academic Press, San Diego, CA.

to form the subtropical jet stream, which is a band of high winds around 30° N and 30° S at around 12 km altitude. The general rising near the equator and sinking in the subtropical latitudes of the winter hemisphere is called the Hadley circulation, after a seventeenth century meteorologist. The Hadley circulation extends only to about 30° of latitude. Beyond that the circulation becomes unstable and breaks down into eddies. These eddies are very efficient at transporting energy poleward, and in middle latitudes, the poleward transport of heat is accomplished mostly by atmospheric eddies or storms. The structure of these storms is such that warm, humid parcels of air move poleward and cold, dry parcels of air move toward

the equator. Storms thus result in a net transport of heat and moisture toward the poles. Because warm, humid parcels within these storms are pushed upward as they are pushed poleward, midlatitude storms also produce heavy precipitation.

Poleward movement of warm water parcels and equatorward movement of cold water parcels produce poleward heat transport in the ocean. The most visible examples of this are the warm western boundary currents of the Atlantic and Pacific oceans – the Gulf Stream and the Kuroshio Current. These are narrow streams of warm water that flow poleward at the western margins of the oceans. The Gulf Stream can carry warm, salty water all the way from the tropics to the vicinity of Northern Norway. These currents account for the relative warmth of the oceans in middle latitudes and have a significant influence on the climate. Because of the shape and far northward extent of the Atlantic Ocean, the Gulf Stream and the associated deep overturning circulation of the Atlantic Ocean have a dramatic effect on the climate and its variability that has a global reach.

The western boundary currents are generally shallow surface currents. In the North Atlantic Ocean, warm, salty water travels very far north, where it is cooled without substantial dilution of the salt content. Because the density of seawater depends mostly on salinity near the freezing point, the cooled Gulf Stream water becomes some of the densest water in the ocean

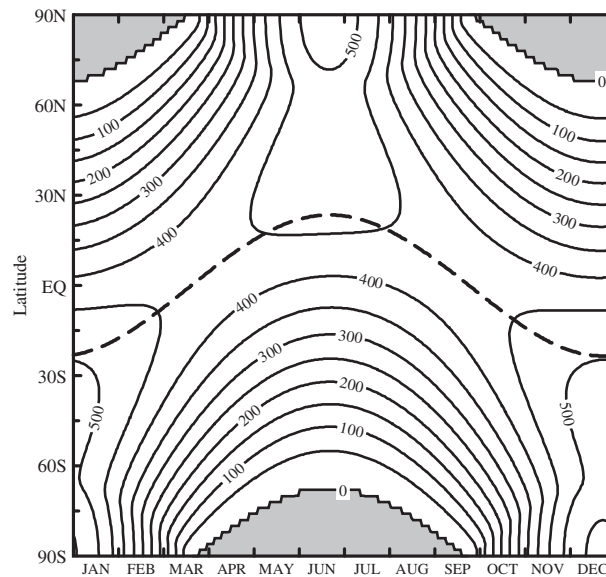


Figure 7 Daily insolation at the top of the atmosphere as a function of season and latitude. Reproduced from Hartmann, D.L., 1994, *Global Physical Climatology*, Academic Press, San Diego, CA.

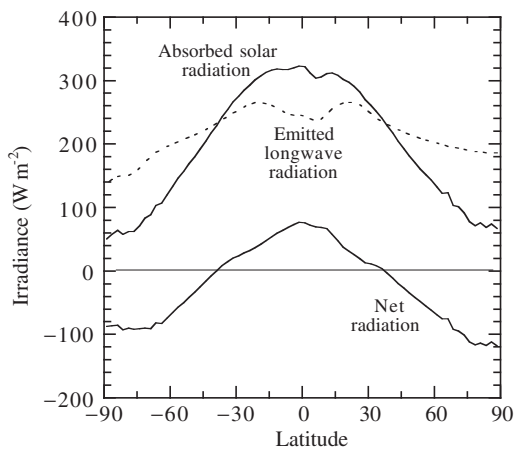


Figure 8 Annually averaged absorbed solar radiation emitted longwave radiation and net radiation at the top of the atmosphere as functions of latitude. Data are from the Earth Radiation Budget Experiment.

and sinks to near the bottom. It then returns equatorward and may travel as far as the North Pacific before it rises to the surface again. The density-driven circulation of the deep ocean is called the thermohaline circulation. The formation of deepwater in the North Atlantic Ocean creates a convergence of surface water, which allows more warm water to flow in from the south and heat the high latitude regions near the North Atlantic. Deep water is not formed in the North Pacific because the Pacific is not as salty as the Atlantic. The salinity of the oceans differs because the atmospheric circulation carries freshwater from the Atlantic basin to the Pacific. Major variations in the intensity of the thermohaline circulation in the North Atlantic are known to have occurred during ice ages. Evidence from Greenland ice cores suggests that the formation of deepwater has switched on and off on timescales of centuries or less during glacial periods.

Topographic Effects

Air pressure and density decrease rapidly with altitude. The air pressure at 5 km above sea level is only half of that at the surface. The air temperature at 5 km above the surface of the ocean is about 30°C colder than the ocean surface. When the land extends higher into the atmosphere, the surface pressure and temperature both decrease. If the land is heated by the Sun, it will be warmer than the air at the same pressure that is not in contact with the land. This results in the generation of buoyancy of the air in contact with the land, which may cause it to rise. A dramatic example of this is the seasonal variation in winds and precipitation over southern Asia. When the Sun heats the Himalayas during Northern Hemisphere summer, air at low levels is drawn toward the Himalayas, resulting in heavy precipitation in the mountains and the adjoining lowlands. During winter, the situation is reversed and air flows from the continent toward the ocean. During the Northern Hemisphere winter, precipitation near the Himalayas is reduced and precipitation over the equatorial Indian Ocean is increased. This seasonal reversal in winds and associated precipitation changes is called the Asian monsoon. A muted form of the same seasonal variation occurs over the tropical and subtropical Americas.

When air approaches a topographic obstruction and is forced to pass over it, much of the moisture in the air is condensed out on the windward side of the topography, and the leeward side can be quite arid if it is consistently blocked from maritime sources of water vapor. The Cascade and Sierra Nevada Mountains, for example, block the flow of moisture from the Pacific Ocean to the Great Basin between the coastal mountains and the Rocky Mountains farther inland. The Himalayas prevent moisture from the Indian Ocean from reaching central Asia.

Topographic barriers also divert the flow of air by generating waves in the atmosphere. During winter, the mountains of

western North America generate a stationary upstream high-pressure system and downstream low-pressure system. This is associated with equatorward flow on the lee of the Rocky Mountains, which makes the center of North America colder than it would be in the absence of topography.

Climate Variability and Change

It is known that the climate of the Earth has varied greatly in the past. During the Cretaceous period 65–135 million years ago, the climate was much warmer. Tropical plants and dinosaurs ranged far north above the Arctic Circle. On these long time-scales, continental drift has altered the arrangement of the continents, which influences the climate. It is likely that atmospheric greenhouse gases such as carbon dioxide or methane were much higher during the Cretaceous period. Changes in atmospheric greenhouse gas concentrations equivalent to a factor of 5 or 10 increase in carbon dioxide seem to be required to explain the similarity of tropical and polar surface temperatures during the Cretaceous and may have been made more likely by the distribution of continents and oceans at that time.

Twenty thousand years ago, much of North America and Europe were covered with sheets of ice several kilometers thick, and a similar glacial age occurred about 140 000 years ago. More recent variations in climate such as the succession of ice ages over the last several million years appear to be triggered by changes in Earth's orbital parameters. A large tilt of Earth's axis of rotation relative to the plane of Earth's orbit about the Sun favors an interglacial climate, because summertime and annual mean insolation in high latitudes increase with larger tilt. When Northern Hemisphere summer solstice occurs, the Earth is closest to the Sun; this also acts to reduce the ice piled on the northern continents and leads to interglacial climate conditions.

The effect of insolation variations is amplified by ice albedo feedback and biogeochemical feedbacks affecting the concentration of carbon dioxide and other greenhouse gases in the atmosphere. As the Earth cools, more of the surface of the Earth is covered with ice. Because ice has a higher albedo than other surfaces, this leads to less solar absorption by the planet and further cooling. The carbon cycle feedbacks are less well known, but changes in the ocean circulation and changes in the supply of trace metals to the ocean have been suggested as means of coupling decreased temperature to decreased carbon dioxide. On timescales of thousands of years, the atmospheric carbon dioxide concentration is tied closely to the carbon dioxide in the ocean. In the surface of the ocean, carbon dioxide is removed by photosynthetic life. Photosynthetic life in the ocean is limited by the amount of nutrients and trace metals that are present in the illuminated zone near the surface. Ice ages are known to be dustier than interglacial epochs like the current one. During ice ages, stronger winds mix the ocean more effectively and can bring more nutrients to the surface from depth. Stronger winds also loft more continental dust containing trace metals and carry it to regions of the ocean far removed from continents where trace metals required for photosynthesis would otherwise be lacking.

Climate also experiences significant year-to-year and decadal variations. These result from the natural internal variability of the climate. The El Niño Southern Oscillation (ENSO) phenomenon is a coupled ocean–atmosphere mode of variability with a timescale from 2 to 7 years, which is centered in the equatorial Pacific. During ENSO warm events, the equatorial SST in the east Pacific rises several degrees above normal and convection that is normally present in the far west Pacific may extend all the way to coastal South America. The rainfall that normally occurs in the western equatorial Pacific and Indonesia follows the warm water eastward into the central Pacific. The movement of the localized convective heating in the atmosphere drives atmospheric waves that may influence weather in middle latitudes. ENSO variability may be coupled to longer term variability in the North Pacific Ocean. Year-to-year variations in surface temperature may be caused by explosive volcanic eruptions that inject large amounts of sulfur-bearing gases into the stratosphere, where fine aerosol particles of sulfuric acid can reflect sunlight.

Decadal variations of climate have been observed, which seem to be related to interactions between atmosphere and ocean. Weather anomalies during winter in middle latitudes can drive thermal anomalies in the oceanic mixed layer and thermocline structure, which are then sealed under a shallow, warm oceanic mixed layer during the following summer. With the onset of surface cooling and turbulent mixing in the following winter, these thermal anomalies can be uncovered and influence the climate, thus giving climate anomalies a year-to-year persistence in middle latitudes. Ocean current circulations in middle latitudes can also transport these anomalies from place to place.

Human-Induced Climate Change

The concentrations of atmospheric greenhouse gases and aerosol particles in the atmosphere have been changing over time in response to human activities. From ice bubbles trapped in ice cores and other evidence, it was known that the concentration of carbon dioxide just prior to the industrial revolution was about 275 ppmv. The concentration has increased at a rate of about 0.5% per year so that the difference between the preindustrial value and the current value is larger than the differences between preindustrial and ice age conditions (Figure 9). Scientists are certain that recent changes in carbon dioxide concentration are caused by fossil fuel use by humans, because the isotopic composition of carbon derived from fossil fuels such as coal, oil, and natural gas is different from carbon that may have come from other sources.

Because carbon dioxide is an important greenhouse gas, its increase can lead to increases in global mean surface temperature. In addition to carbon dioxide, the concentrations of nitrous oxide, methane, and some industrially created greenhouse gases are also increasing in the atmosphere. The effects of these and projected future changes can be studied with global climate models. The known physics of the climate system can be incorporated into these computer models. Because of the wide range of spatial and temporal scales in the climate system, current limitations on computer power, and some gaps in the understanding of the climate system, these projections remain

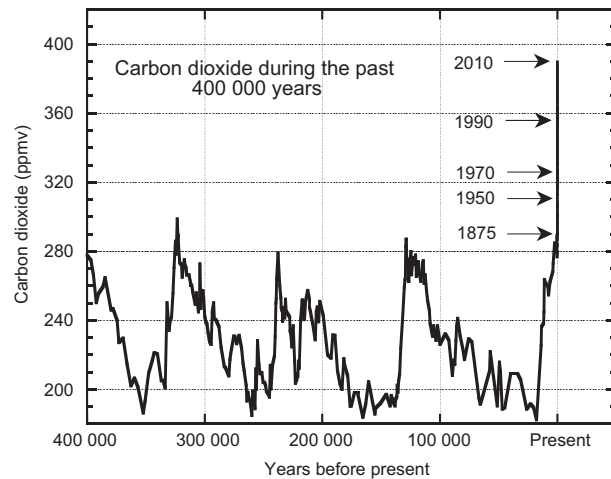


Figure 9 Carbon dioxide for last 400 000 years. Data are from Vostok, Antarctica, and Siple Station, Greenland ice cores and modern instrumental data from Mauna Loa, Hawaii.

somewhat uncertain. Key feedback processes such as the cloud formation need to be treated with very approximate methods. Nonetheless, the basic physics of the greenhouse effect can be calculated precisely and scientists believe that human-induced changes in atmospheric composition have caused much of the recent warming of the surface temperature of the Earth. Many of the warmest years in the global temperature record have occurred in the last two decades, and the land has warmed faster than the ocean (**Figure 10**).

Changes in Earth's climate will occur in the future as a result of human activities. Because of the large heat capacity of the oceans, the warming resulting from greenhouse gases lags several decades behind the greenhouse gas changes. It is estimated that about another 0.5°C of warming should result from changes in the composition of the atmosphere that humans have already caused, and human modification of the atmosphere is continuing at an increasing rate.

See also: **Aerosols:** Role in Climate Change. **Climate and Climate Change:** Carbon Dioxide; Climate Variability: Decadal to Centennial Variability; Climate Variability: Seasonal and Interannual Variability; Greenhouse Effect. **Global Change:** Climate Record: Surface Temperature Trends. **Paleoclimatology:** Ice Cores. **Satellites and Satellite Remote Sensing:** Earth's Radiation Budget.

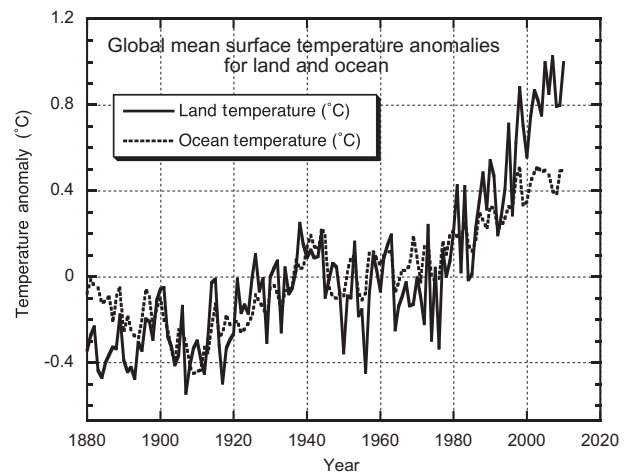


Figure 10 Global mean surface air temperature as a function of time over land and ocean areas. Data are from the National Climatic Data Center. Annual means end in 2010.

Further Reading

- Crowley, T.J., North, G.R., 1991. *Paleoclimatology*. Oxford University Press, Oxford, UK.
- Graedel, T.E., Crutzen, P.J., 1995. *Atmosphere, Climate, and Change*. W.H. Freeman, New York.
- Hartmann, D.L., 1994. *Global Physical Climatology*. Academic Press, San Diego, CA.
- Henson, R., 2011. *The Rough Guide to Climate Change*. Rough Guides, London.
- Imbrie, J., Imbrie, K.P., 1979. *Ice Ages: Solving the Mystery*. Enslow Publishers, Short Hills, NJ.
- Kump, L.R., Kasting, J.F., Crane, R.G., 1999. *The Earth System*. Prentice Hall, Upper Saddle River, NJ.
- Peixoto, J.P., Oort, A.H., 1992. *Physics of Climate*. American Institute of Physics, New York.
- Solomon, S., 2007. *Climate Change 2007: The Scientific Basis: Contribution of Working Group I to the Third Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, Cambridge.
- Trenberth, K.E., 1992. *Climate System Modeling*. Cambridge University Press, Cambridge, UK.
- Washington, W.M., Parkinson, C.L., 1986. *An Introduction to Three-Dimensional Climate Modeling*. University Science Books, Mill Valley.

Carbon Dioxide

CL Sabine and RA Feely, NOAA Pacific Marine Environmental Laboratory, Seattle, WA, USA

© Published by Elsevier Ltd.

Synopsis

Carbon dioxide plays a significant role in the Earth's life cycle and in controlling the global climate. It is a greenhouse gas that makes life on Earth possible, but mankind is currently in the process of altering the global carbon cycle through land use practices and burning of fossil fuels. This article looks at the history of atmospheric CO₂ from the geological past, through the industrial era, and considers possible changes in the future. It looks at the relevant controls on atmospheric CO₂ in the atmosphere, terrestrial biosphere, and in the ocean on timescales of years to centuries.

Introduction

Carbon dioxide (CO₂) is considered a trace gas in the atmosphere, with contemporary concentrations approaching 400 parts per million (ppm) by volume. Despite its low concentrations relative to those of nitrogen or oxygen, CO₂ plays a significant role in the Earth's life cycle and in controlling the global climate.

CO₂ is released as a by-product of aerobic respiration. Plants take up CO₂ and release oxygen as a part of photosynthesis. Variations in the global balance between photosynthesis and respiration result in seasonal variations in atmospheric CO₂ of up to 15 ppm. For example, atmospheric CO₂ concentrations in the Northern Hemisphere are generally lower in the summer, when many plant species enter a new growth stage and photosynthesis predominates over respiration. CO₂ is also a greenhouse gas that absorbs long-wavelength radiation in the atmosphere, attenuating its escape into space. The trapping of radiation by CO₂ and other greenhouse gases (e.g., water vapor, methane, nitrous oxide, and chlorofluorocarbons) helps keep Earth warmer than it would be without an atmosphere. It is this greenhouse warming that makes life, as we know it, possible on Earth.

Mankind is currently in the process of altering the chemistry of the global atmosphere. Atmospheric CO₂ concentrations have been increasing as a direct result of human activities such as deforestation and the burning of fossil fuels (e.g., coal and oil). Over the past 150 years, CO₂ concentrations in the atmosphere have increased by as much as 40% (from 280 to 392 ppm in 2012; [Figure 1](#)). This has been accompanied by an increase in global mean surface temperature of between 0.65 and 1.06 °C. The present rate of increase in CO₂ is unprecedented over the last 22 000 years.

This article briefly describes the complicated role that CO₂ plays on Earth, the different pools where carbon is stored, and the ways in which carbon is transferred between pools over various periods. We will focus on those pools (reservoirs) and transfers (fluxes) with timescales relevant to the human alteration of the natural carbon cycle. In addition, scientific issues relevant to future atmospheric CO₂ concentrations will be discussed along with a brief introduction to some of the global policy issues regarding regulation of future CO₂ emissions.

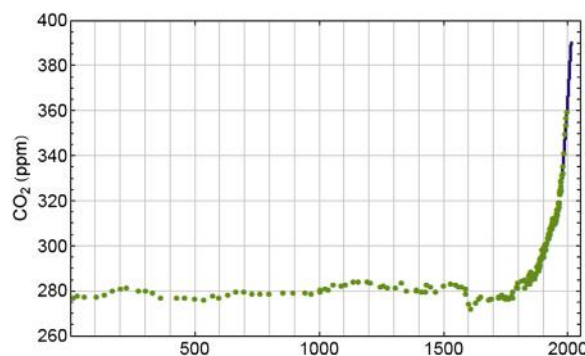


Figure 1 CO₂ concentrations in Antarctic ice cores (symbols) and annual mean concentrations from direct atmospheric measurements (line) for the past millennium. Prior to the industrial revolution atmospheric CO₂ values were very near 280 ppm. For the past 150 years, atmospheric concentrations have been increasing exponentially. Adapted from Ciais, P., Sabine, C., Govindasamy, B., Bopp, L., Brovkin, V., Canadell, J., Chhabra, A., DeFries, R., Galloway, J., Heimann, M., Jones, C., Le Quéré, C., Myneni, R., Piao, S., Thornton, P., 2013. Carbon and other biogeochemical cycles (Chapter 6). In: Stocker, T., Qin, D., Plattner, G.-K. (Eds.), *Climate Change 2013: The Physical Science Basis*. Cambridge University Press, Cambridge.

Geological History of Atmospheric CO₂

Atmospheric CO₂ varied considerably over the Earth's early history. The balance of several geochemical processes, including organic carbon and calcium carbonate burial, silicate rock weathering, and volcanism, controls the concentration of atmospheric CO₂ over long (millennial) timescales. When one process dominates over the others, such as during intensified periods of active volcanism, atmospheric CO₂ concentrations can change significantly over time. For example, during times of high volcanism in the Jurassic period about 200–150 Ma, CO₂ concentrations in the atmosphere rose to more than 3000 ppm. The high CO₂ levels made the rain more acidic by reacting with the water to form carbonic acid. The high acid content of the rainwater eventually caused higher rates of chemical breakdown (weathering) in carbonate and silicate rocks, releasing basic ions (e.g., bicarbonate and silicate). These basic ions traveled via rivers and groundwater to the oceans, eventually increasing the

ocean's capacity to absorb CO_2 from the atmosphere. The patterns of the large-scale geological processes and the responses of the global carbon cycle during the early stages of the Earth's history caused large swings in atmospheric CO_2 .

The relatively high atmospheric CO_2 concentrations present over most of the Earth's history were substantially decreased with the evolution of terrestrial vegetation and the subsequent enhancement of silicate weathering from the decomposition of plant matter in soils. For approximately the past 20 My, the geochemical evidence from the sedimentary record suggests that CO_2 concentrations in the atmosphere remained below 300 ppm until the beginning of the twentieth century. Recent studies of CO_2 trapped in air bubbles preserved in ice cores from Greenland and Antarctica have provided scientists with high-resolution records of atmospheric CO_2 concentrations for the past 800 000 years. The longest record, from the Vostok ice cores in Antarctica, shows that atmospheric CO_2 fluctuations were only about 100 ppm, associated with transitions from glacial to interglacial periods (Figure 2). The Vostok ice core record, which spans four glacial-interglacial cycles, reveals that atmospheric CO_2 was low (~ 180 ppm) during the glacial periods and high (~ 280 ppm) during the interglacial periods. Natural changes in the global carbon cycle between glacial and interglacial periods have maintained the atmospheric CO_2 levels between these two extremes for at least the past million years.

Controls of Modern Atmospheric CO_2

There are only three major reservoirs with exchange rates fast enough to vary significantly on the timescale of decades to centuries: the atmosphere, the terrestrial biosphere, and the oceans. Of this three-component system, approximately 93% of the carbon is located in the oceans. The atmosphere, at about

830 Pg C ($1 \text{ Pg} 10^{15} \text{ g}$), has the smallest total pool of carbon. The annual atmospheric exchange of CO_2 with both the oceans and the terrestrial biosphere is on the order of ~ 200 Pg C per year. Although these fluxes are very large, the mass of carbon in each of these reservoirs may not change over time. Ice core records suggest that the atmospheric CO_2 concentrations were very close to 280 ppm for more than 1000 years prior to the industrial era (see Figure 1). This constancy suggests that the carbon pools were more or less in equilibrium, and the net transfer over sufficiently large areas was close to zero.

Atmosphere

Of the three rapidly exchanging major carbon pools, the atmosphere is the most well mixed. Since most of the current CO_2 sources to the atmosphere are in the Northern Hemisphere and CO_2 exchange across the Equator is estimated to take about 1 year, the CO_2 concentrations in the north are higher than in the southern latitudes (Figure 3). The fact that most of the terrestrial biosphere is located in the Northern Hemisphere also results in larger seasonal variations relative to the Southern Hemisphere. Data from the present Global CO_2 Atmospheric Sampling Stations (Figure 4) indicate that distinct CO_2 sources and sinks are difficult to identify by examining atmospheric concentrations alone because the CO_2 is quickly distributed around the globe by the winds. East-west gradients of atmospheric CO_2 concentration are an order of magnitude smaller than north-south gradients. CO_2 concentrations over the continents are only a few ppm higher than corresponding locations over the oceans.

Terrestrial Biosphere

Every year approximately 18% of the CO_2 in the atmosphere is cycled through terrestrial plants. This amount of carbon, or

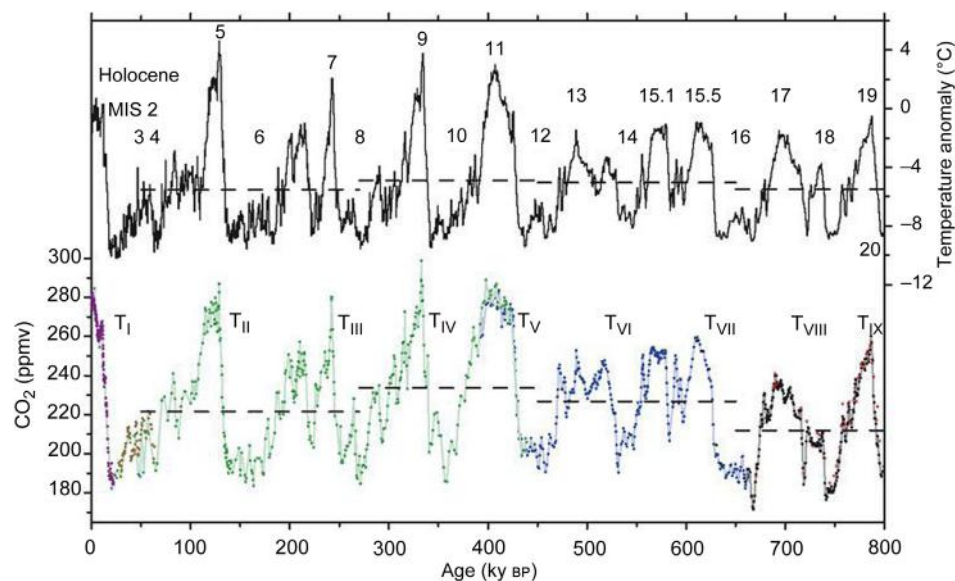


Figure 2 Compilation of the European Project for Ice Coring in Antarctica (EPICA) Dome C CO_2 ice core record and temperature anomaly over the past 800 000 years. After Lüthi, D., Le Floch, M., Bereiter, B., Blunier, T., Barnola, J.-M., Siegenthaler, U., Raynaud, D., Jouzel, J., Fischer, H., Kawamura, K., Stocker, T., 2008. High-resolution carbon dioxide concentration record 650,000–800,000 years before present. *Nature* 453, 379–382. <http://dx.doi.org/10.1038/nature06949>.

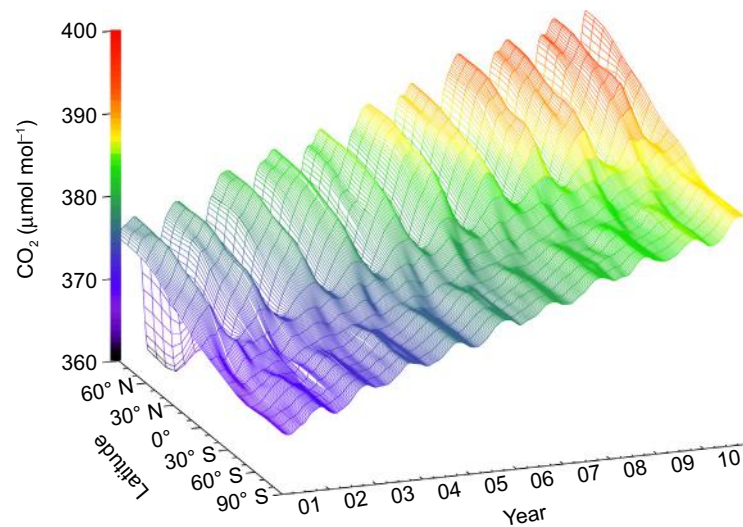


Figure 3 Three-dimensional representation of changes in the latitudinal distribution of atmospheric carbon dioxide as a function of time. CO₂ concentrations increase from south to north. The surface shows greater variability in the northern latitudes because of seasonal changes in the terrestrial biosphere. The increase in CO₂ concentration with time is the result of anthropogenic emissions. Data are from the National Oceanic and Atmospheric Administration's Earth System Research Laboratory (NOAA/ESRL) Global Monitoring Division cooperative air sampling network.

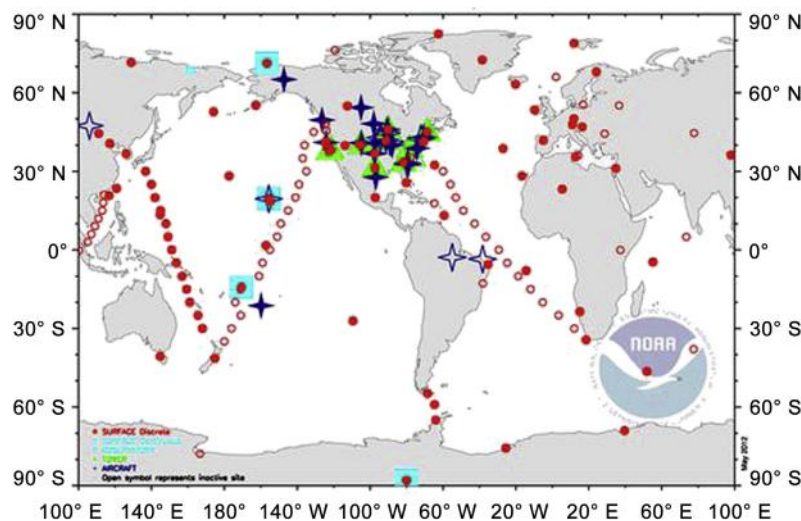


Figure 4 Map of Global CO₂ Atmospheric Sampling Stations organized through NOAA/ESRL Global Monitoring Division. Network and observatory stations provide time series measurements. Tower and aircraft measurements provide vertical resolution. Figure provided by P. Tans, NOAA/ESRL.

about 120 Pg C, is converted every year to carbohydrates and plant material during photosynthesis. About half of this material forms plant tissue (leaves, stems, roots, and wood) and the other half is returned to the atmosphere by autotrophic respiration (CO₂ released by the plants). Global net primary production (NPP, photosynthesis–autotrophic respiration) is estimated to be approximately 60 Pg C per year. Of this amount, approximately 83% is returned to the atmosphere via heterotrophic respiration (decomposition of plant material by bacteria, fungi, and herbivores) and combustion due to forest fires. The carbon that remains is called the net ecosystem

production. Estimates of changes in carbon stocks by the US Forest Service suggest that about 10 Pg C per year is fixed into new plant matter globally every year. The net ecosystem production suffers further losses of carbon from harvesting and erosion of soil carbon and plant materials that are eventually transported via streams to rivers, and to the oceans. What remains behind, the net biome production, is the amount of carbon that is accumulated by the terrestrial biosphere every year. It is this carbon that is ultimately lost from the atmosphere as a terrestrial sink. Estimates of the net biome production averaged 0.2 Pg C per year (i.e., small net carbon

uptake from the atmosphere) during the decade of the 1980s and 1.0 Pg C per year in the decade of the 1990s.

Oceans

The oceans are the largest of the three main CO₂ reservoirs, containing about 50 times more CO₂ than the atmosphere and 19 times more than the terrestrial biosphere. On an annual basis, the two-way exchange of CO₂ between the atmosphere and the surface ocean is approximately 80 Pg C per year. Net exchange of CO₂ occurs by diffusion when there is a difference in the CO₂ partial pressure ($p\text{CO}_2$) between the atmosphere and oceans. For example, when the atmospheric $p\text{CO}_2$ is higher than the surface ocean, CO₂ diffuses across the air-sea boundary into the oceans. The oceans are able to hold much more carbon than other reservoirs because most of the CO₂ that diffuses into the oceans reacts with the water to form carbonic acid and its dissociation products, bicarbonate and carbonate ions. The conversion of dissolved CO₂ to bicarbonate and carbonate ions effectively reduces the $p\text{CO}_2$ in the water, promoting more diffusion from the atmosphere.

The oceans are mixed much more slowly than the atmosphere, so there are large horizontal and vertical gradients in CO₂ concentration. For more than 40 years, marine scientists have been measuring the distribution of $p\text{CO}_2$ in the surface waters of the oceans. A summary of the global data set is presented in Figure 5. CO₂ uptake from the atmosphere was generally found to occur in the high-latitude oceans as a result of two factors. First, CO₂ is more soluble in cold water, so as ocean currents (such as the Gulf Stream) transport water from the tropics to the poles they are cooled and can absorb more

CO₂ from the atmosphere. Second, the high-latitude zones are also regions where intermediate and bottom waters are formed. As the waters are cooled, they become denser and sink into the ocean's interior taking with them the CO₂ accumulated at the surface. This process of transporting CO₂ from the surface ocean to the deep because of the cooling and sinking of water masses is known as the solubility pump.

Another process that transfers CO₂ away from the surface ocean is termed the biological pump. Photosynthetic production of marine plants (phytoplankton) incorporates CO₂ and nutrients from seawater into living plant tissue and detritus. Microscopic marine animals, called zooplankton, consume the phytoplankton and provide the basis for the food web for all animal life in the sea. The gross primary production by marine phytoplankton is estimated to be about 110 Pg C per year. Most of this carbon is recycled in the upper ocean via autotrophic respiration, similarly to what occurs on land. Some of the organic carbon is transformed into dissolved organic carbon that is transported by currents and diffusion to deeper depths and oxidized by marine bacteria. The remainder of the organic carbon sinks as particulate matter. The downward transport of dissolved organic carbon, particulate organic carbon, and detritus makes up the bulk of the downward export flux of carbon into the ocean interior. Estimates of this global export production range from 10 to 20 Pg C per year. Heterotrophic respiration converts most of this organic carbon back into dissolved inorganic carbon (dissolved CO₂, bicarbonate, and carbonate) at depth. Only about 0.1 Pg C per year reaches the seafloor to be buried in the sediments. The CO₂ that is recycled at depth is slowly transported large distances by currents to areas where the waters return to the surface

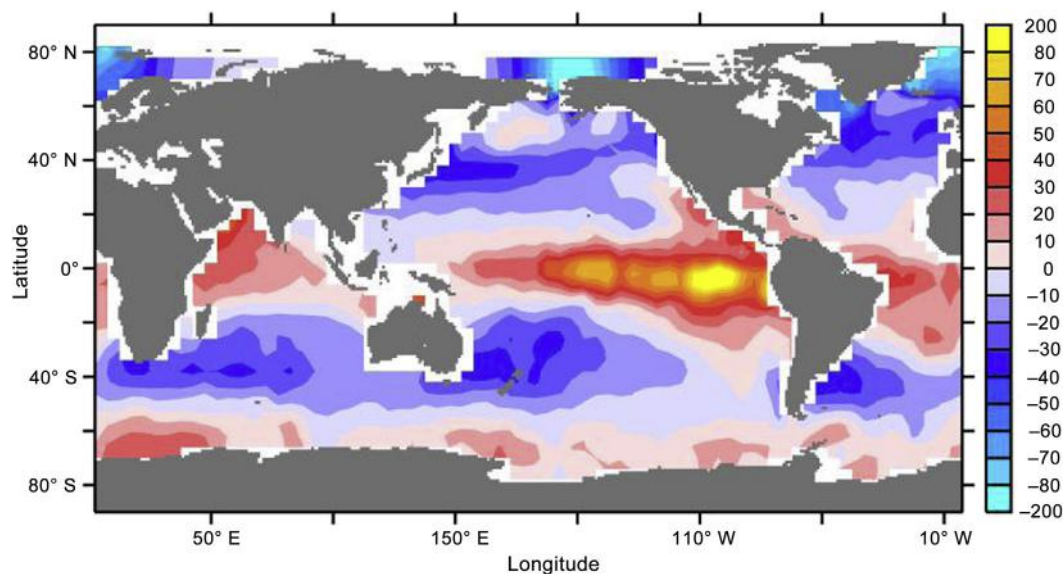


Figure 5 Climatological map of seawater $p\text{CO}_2$ -atmospheric CO₂ (µatm) for the year 2000. Red and yellow regions indicate areas where CO₂ is transferred from the ocean to the atmosphere. Blue and purple regions indicate where CO₂ is entering the ocean. Adapted from Takahashi, T., Sutherland, S.C., Wanninkhof, R., Sweeney, C., Feely, R.A., Chipman, D.W., Hales, B., Friederich, G., Chavez, F., Sabine, C., Watson, A., Bakker, D.C.E., Schuster, U., Metzl, N., Yoshikawa-Inoue, H., Ishii, M., Midorikawa, T., Nojiri, Y., Körtzinger, A., Steinhoff, T., Hopemman, M., Olafsson, J., Arnarson, T.S., Tilbrook, B., Johannessen, T., Olsen, A., Bellerby, R., Wong, C.S., Delille, B., Bates, N.R., de Baar, H.J.W., 2009. Climatological mean and decadal change in surface ocean $p\text{CO}_2$, and net sea-air CO₂ flux over the global oceans. *Deep-Sea Research II* 56 (8–10), 554–577. <http://dx.doi.org/10.1016/j.dsr2.2008.12.009>.

(upwelling regions). When the waters regain contact with the atmosphere, the CO₂ originally sequestered by the phytoplankton is returned to the atmosphere. This re-equilibration process helps to regulate atmospheric CO₂ concentrations over decadal and longer timescales. The major upwelling regions are along the equatorial belt, the Antarctic Circumpolar region, and various localized coastal upwelling regions. It is this biological pump that primarily maintains the strong vertical gradient of dissolved carbon in the oceans.

CO₂ in the Industrial Era

Vast amounts of carbon are stored in the Earth's crust as coal, natural gas, and oil. If not for the mining activities of mankind, this carbon would not be recycled back into the atmosphere for millennia. Combustion of coal, natural gas, and oil released an estimated 5.5 Pg C per year into the atmosphere in the decade of the 1980s, 6.4 Pg C per year in the 1990s and 7.8 Pg C per year during the first decade of this century (Table 1). Other human activities, such as cement manufacturing, also contribute significant quantities of CO₂ into the atmosphere every year. CO₂ generated as a direct result of human activities is called anthropogenic CO₂. The release of anthropogenic CO₂ increased the atmospheric concentrations on average by 2.0 ± 0.1 ppm every year over the last decade. This increase has been well documented since the 1950s by direct atmospheric measurements in locations such as Mauna Loa, Hawaii, that are well away from the primary CO₂ source regions (Figure 6).

Annual assessments of fossil fuel consumption are compiled for each country. By estimating the CO₂ released by each fuel type, one can estimate the total CO₂ released to the atmosphere each year. One intriguing finding of the atmospheric CO₂ sampling program is that only about half of the anthropogenic CO₂ released each year remains in the atmosphere. The exact mechanisms and locations for the remaining anthropogenic CO₂ sinks has been a matter of intense research for the past few decades. It is important to note that, in order to qualify as a sink for anthropogenic CO₂, the fluxes must be enhanced over preindustrial rates. CO₂ fluxes associated with the ocean's biological pump, for example, are 10–20 Pg C per year. There is no conclusive evidence, however, that these rates

have changed in the open ocean as a result of increased atmospheric CO₂. The growth of marine phytoplankton is generally limited by the availability of nutrients in the waters, not by carbon availability. In this case, although the fluxes associated with the biological pump are roughly twice the fossil fuel signal, the biological pump is not believed to be a significant sink for anthropogenic CO₂. The difficulty with determining exactly where the anthropogenic CO₂ is stored is that the Earth systems are very complex and the anthropogenic CO₂ signal is small relative to the natural fluxes. Scientists attribute the majority of the anthropogenic CO₂ removed from the atmosphere to uptake by the terrestrial biosphere and the ocean's solubility pump.

Terrestrial Biosphere

Human activities, particularly in the northern temperate and tropical forests, directly affect net biome production by changing agricultural practices, deforestation and reforestation practices, nitrogen fertilization practices, and CO₂ fertilization of land plants via increasing atmospheric CO₂ concentrations. Estimation of carbon stocks in terrestrial ecosystems requires an accurate knowledge of land cover, carbon density in vegetation and soils, and the fate of the plant carbon (e.g., burning, decomposition). Global estimates of the net terrestrial CO₂ sink via annual carbon stock changes have very large uncertainties. Scientists believe that the terrestrial biosphere is currently a global sink for carbon despite large releases of carbon as a result of deforestation (Table 1).

Deforestation has been responsible for almost 90% of the estimated emissions due to land-use change since 1850, with a 20% decrease of the global forest area. Conversion of natural vegetation to agriculture has also been a major source of CO₂, not only because of plant biomass loss but also from increased decomposition of soil organic matter caused by disturbance and the energy costs of various agricultural practices. However, these losses can be minimized with proper land management and the use of high-yielding plant varieties.

All trees, nearly all plants from cold climates, and most agricultural crops respond to increasing atmospheric CO₂ levels by enhancing their photosynthetic uptake of carbon. Experiments have shown an average increase in NPP of 20–25% for a doubling of atmospheric CO₂. Increased uptake, particularly in temperate climates, can serve to counteract the CO₂ release due to land-use practices like deforestation. However, experiments on some plant species have shown a diminishing or lack of CO₂ fertilization effect in some ecosystems; likely due to nitrogen or phosphorus limitation. Nitrogen limitation appears to be prevalent in temperate and boreal ecosystems while phosphorus limitation dominates in the tropics. Research also suggests that inputs of nitrogen from human activities may also help to increase NPP rates in some areas. Thus, increases in the uptake rate of the terrestrial biosphere due to CO₂ fertilization have large uncertainties at this time.

Ocean Solubility Pump

The constancy of atmospheric CO₂ concentrations in the centuries prior to the industrial revolution suggests that the

Table 1 Global CO₂ budgets (in Pg C per year) for the last three decades based upon atmospheric CO₂ and O₂ data^a

Source or sink	1980–89	1990–99	2000–09
CO ₂ emissions (fossil fuel, cement product)	5.5 ± 0.4	6.4 ± 0.5	7.8 ± 0.6
Atmospheric increase	3.4 ± 0.2	3.1 ± 0.2	4.0 ± 0.2
Ocean–atmosphere flux	–2.0 ± 0.7	–2.2 ± 0.7	–2.3 ± 0.7
Land–atmosphere flux	–0.1 ± 0.8	–1.1 ± 0.9	–1.5 ± 0.9

^aPositive values represent fluxes to the atmosphere; negative values represent uptake away from the atmosphere. The land–atmosphere flux represents the net balance of a positive term due to land use changes and a negative term due to a residual carbon sink.

Source: Ciais, P., Sabine, C., Govindasamy, B., Bopp, L., Brovkin, V., Canadell, J., Chhabra, A., DeFries, R., Galloway, J., Heimann, M., Jones, C., Le Quéré, C., Myneni, R., Piao, S., Thornton, P., 2013. Carbon and other biogeochemical cycles (Chapter 6). In: Stocker, T., Qin, D., Plattner, G.-K. (Eds.), *Climate Change 2013: The Physical Science Basis*. Cambridge University Press, Cambridge.

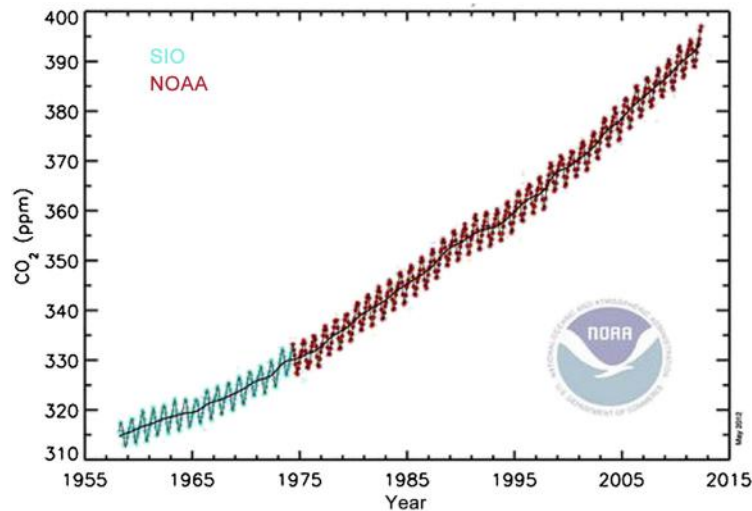


Figure 6 Monthly mean atmospheric CO₂ concentrations from Mauna Loa, Hawaii, sampling station. Data prior to May 1974 are from the Scripps Institution of Oceanography (SIO). Data since May 1974 are from NOAA/ESRL. The monthly trend curve illustrates the magnitude of the seasonal cycle at this latitude. While the seasonal range has been relatively constant, the long-term trend curve clearly shows the annual increase in atmospheric CO₂. Figure provided by P. Tans, NOAA/ESRL.

Oceans were a small net source of CO₂ to the atmosphere to balance the carbon input from rivers. Today, the net oceanic flux has been reversed because of anthropogenic inputs of CO₂ to the atmosphere. A global average air–sea difference of 8 ppm would result in a net flux of approximately 2 Pg C per year. The uptake of anthropogenic CO₂ by the oceans is driven by the thermodynamic forcing from the air–sea difference in $p\text{CO}_2$ and the air–sea transfer velocity. The transfer velocity is related to the surface roughness of the ocean and the wind speed. The thermodynamic forcing is related to the amount of CO₂ that is converted from $p\text{CO}_2$ to other carbon species in the seawater, such as bicarbonate and carbonate ions. This so-called buffer capacity is what allows the oceans to hold so much carbon. The relative concentrations of dissolved CO₂ (1%), bicarbonate ion (91%), and carbonate ion (8%) control the acidity (pH) of the oceans. Since CO₂ is an acid gas, the uptake of anthropogenic CO₂ consumes carbonate ions and lowers the oceanic pH in a process called ocean acidification. The carbonate ion concentration of surface seawater in equilibrium with the atmosphere will decrease by about 28% with a doubling of atmospheric CO₂ from preindustrial levels (from 280 to 560 ppm). As the carbonate ion concentration decreases, the buffer capacity of the oceans and its ability to further absorb CO₂ from the atmosphere is reduced.

Future CO₂ Concentrations

Coupled Atmosphere–Ocean General Circulation Models (AOGCMs) have long been used for making climate projections. Model runs coordinated through the 5th Coupled Model Intercomparison Project (CMIP5) form the core of the projections section of the Intergovernmental Panel on Climate Change (IPCC) 5th Assessment Report. Many of the AOGCMs used in CMIP5 now have an interactive carbon cycle; that is components of land and ocean biogeochemistry respond to changes in the

climate conditions to influence atmospheric CO₂ concentrations. These models provide an important predictive link between fossil fuel CO₂ emissions, future CO₂ concentrations and climate.

The carbon cycle response to future climate and CO₂ changes can be viewed as two strong and opposing feedbacks: climate and CO₂ concentration. Changes in climate generally lead to decreases in carbon storage while increases in atmospheric CO₂ generally lead to larger carbon storage. On a global scale, models agree on the sign of ocean and land response to increasing CO₂ and climate change, but do not agree on the magnitude of the changes.

The models generally agree that tropical ecosystems will store less carbon in a warmer climate while at high latitudes warming will increase storage of carbon in trees. It should be noted, however, that limitations in forest growth due to nitrogen availability and a loss of carbon from the decomposition of permafrost are generally not accounted for in the models. Models suggest that ocean carbon storage will continue to increase as long as atmospheric CO₂ continues to rise. However, warming and circulation changes will reduce the rate of carbon uptake in the ocean, particularly in the Southern Ocean and North Atlantic.

Carbon Sequestration Efforts

Over the long term (millennial timescales), the ocean has the potential to absorb approximately 85% of the anthropogenic CO₂ that is released to the atmosphere. As long as atmospheric CO₂ concentrations continue to rise, the oceans will act as a sink. However, this reaction is reversible. If atmospheric CO₂ were to decrease in the future, the oceans would start to release the accumulated anthropogenic CO₂ back out into the atmosphere. The ultimate sinks for anthropogenic CO₂ must be reactions that bind the CO₂ in a manner that is not easily reversed. Weathering of silicate rocks on land, for example, is a long-term sink for CO₂.

Carbon burial into the sediments can also provide a long-term sink. One proposed approach for enhancing carbon removal from the atmosphere is to enhance phytoplankton production by fertilizing specific regions of the ocean with micronutrients such as iron that are currently limiting production. The hypothesis is that the resulting bloom of oceanic plants would remove CO₂ from the atmosphere then transport that carbon into the deep ocean or sediments, effectively removing it from the short-term budget. The effectiveness of the 'iron hypothesis' is currently being studied. Other carbon sequestration approaches, including direct injection of liquefied CO₂ into the deep ocean, are also being examined. Further research is necessary to determine whether any of these techniques will be effective or economically feasible. Implementation of these approaches may depend, in large part, on policy decisions made on national and international levels.

Efforts to Reduce CO₂ Emissions – The Kyoto Protocol and Beyond

During the Earth Summit in Rio de Janeiro in 1992, the United Nations established the Framework Convention on Climate Change (UNFCCC) to encourage member countries to stabilize greenhouse gas emissions at a level that will prevent dangerous interference with the global climate system. In 1997 a meeting was held in Kyoto, Japan, to establish legally binding commitments for reducing future emissions of CO₂ and other greenhouse gases. The goal was for each member country to set emission commitments so that by the period 2008–12 the overall global emissions will be reduced to 5% below the levels emitted in 1990. The time frame 2008–12 was set for compliance so that member nations could make the transition to efficient lower-emitting carbon technologies.

The Kyoto Protocols stipulated that if some member countries could enhance the Earth's natural carbon sinks, for example by enhancing the uptake of carbon into forests, these additional carbon 'sinks' could be used as part of their emission assessments. For countries with major forests, like the United States, Russia, and Canada, these enhanced carbon sinks could be used to counterbalance some small increases in CO₂ emissions. In addition, the Protocols allow for trading of these so-called carbon credits to take advantage of international trade agreements. Developing countries were assigned no emission commitments and developed countries were required to reduce their emissions by an average of 7% below their 1990 levels. One provision of the protocols allows developing countries to trade emission credits to developed countries in exchange for technology that enhances carbon uptake.

The Protocol entered into force in 2005 despite the fact that the United States signed, but did not ratify it. The Conference of the Parties (COP) is the supreme decision making body of the UNFCCC. They have continued to meet annually to discuss the implementation of the Convention, the Kyoto Protocol and any other legal instruments that the COP adopts. At the COP 17 meeting in Durban, South Africa at the end of 2011, the Protocol was amended to accommodate a second commitment period from 2013 to 2020, a new platform of negotiations

under the Convention was outlined for the period beyond 2020, and the long-term implementation of a global support network to deliver funding and technologies that help developing countries was launched. At the Warsaw Climate Change Conference (COP19) in 2013 decisions to further advance the 'Durban Platform,' the Green Climate Fund and the long-term financing of REDD plus (Reducing Emissions from Deforestation and forest Degradation in developing countries) were adopted.

Summary

Atmospheric CO₂ is an integral part of all life on Earth. It is not only involved in basic photosynthesis and respiration but also helps control the climate of the environment we live in. The global carbon cycle is very complex and not completely understood today. The gross fluxes associated with a number of processes have the potential to quickly alter the composition of the global atmosphere. The positive and negative feedback mechanisms associated with these processes appear to have been relatively well balanced for the period of mankind's existence. Today, however, human activities are serving to alter that balance. Attempts to regulate emissions at the international level have not had great success. Researchers are working to better understand how the global environment might be altered from the addition of anthropogenic CO₂ and other greenhouse gases to the atmosphere. At this point, a doubling of atmospheric CO₂ over preindustrial levels is almost inevitable. Recent changes observed in the global climate are consistent with the predicted response to increasing greenhouse gases in the atmosphere.

See also: Boundary Layer (Atmospheric) and Air Pollution: Ocean Mixed Layer; Surface Layer. **Climate and Climate Change:** Climate Variability: Decadal to Centennial Variability; Climate Variability: Seasonal and Interannual Variability; Greenhouse Effect; History of Scientific Work on Climate Change; Intergovernmental Panel on Climate Change; Overview. **Land-Atmosphere Interactions:** Overview; Trace Gas Exchange. **Paleoclimatology:** Ice Cores.

Further Reading

- Ciais, P., Sabine, C., Goussaud, B., Bopp, L., Broecker, V., Canadell, J., Chhabra, A., DeFries, R., Galloway, J., Heimann, M., Jones, C., Le Quéré, C., Myneni, R., Piao, S., Thornton, P., 2013. Carbon and other biogeochemical cycles (Chapter 6). In: Stocker, T., Qin, D., Plattner, G.-K. (Eds.), *Climate Change 2013: The Physical Science Basis*. Cambridge University Press, Cambridge.
- Friedlingstein, P., Houghton, R.A., Marland, G., Hackler, J., Boden, T.A., Conway, T.J., Canadell, J.G., Raupach, M.R., Ciais, P., Le Quéré, C., 2010. Update on CO₂ emissions. *Nature Geoscience* 3, 811–812. http://dx.doi.org/10.1038/ngeo_1022.
- Le Quéré, C., Andres, R.J., Boden, T., Conway, T., Houghton, R.A., House, J.I., Marland, G., Peters, G.P., van der Werf, G.R., Ahlstrom, A., Andrew, R.M., Bopp, L., Canadell, J.G., Ciais, P., Doney, S.C., Enright, C., Friedlingstein, P., Huntingford, C., Jain, A.K., Jourdain, C., Kato, E., Keeling, R.F., Klein Goldewijk, K., Levis, S., Levy, P., Lomas, M., Poulter, B., Raupach, M.R., Schwinger, J., Sitch, S., Stocker, B.D., Viomy, N., Zaehle, S., Zeng, N., 2013. The global carbon budget 1959–2011. *Earth System Science Data* 5, 165–185. <http://dx.doi.org/10.5194/essd-5-165-2013>.

- Lüthi, D., Le Floch, M., Bereiter, B., Blunier, T., Barnola, J.-M., Siegenthaler, U., Raynaud, D., Jouzel, J., Fischer, H., Kawamura, K., Stocker, T., 2008. High-resolution carbon dioxide concentration record 650,000–800,000 years before present. *Nature* 453, 379–382. <http://dx.doi.org/10.1038/nature06949>.
- Pan, Y., Birdsey, R., Fang, J., Houghton, R., Kauppi, P., Kurz, W.A., Phillips, O.L., Shvidenko, A., Lewis, S.L., Canadell, J.G., Ciais, P., Jackson, R.B., Pacala, S., McGuire, A.D., Piao, S., Rautiainen, A., Sitch, S., Hayes, D., 2011. A large and persistent carbon sink in the world's forests. *Science* 333 (6045), 988–993. <http://dx.doi.org/10.1126/science.1201609>.
- Sabine, C.L., Tanhua, T., 2010. Estimation of anthropogenic CO₂ inventories in the ocean. *Annual Reviews of Marine Science* 2, 175–198. <http://dx.doi.org/10.1146/annurev-marine-120308-080947>.
- Scholes, R.J., Monteiro, P.S., Sabine, C., Canadell, J.G., 2009. Systematic observations of the global carbon cycle. *Trends in Ecology and Evolution* 1098, 1–4.
- Takahashi, T., Sutherland, S.C., Wanninkhof, R., Sweeney, C., Feely, R.A., Chipman, D.W., Hales, B., Friederich, G., Chavez, F., Sabine, C., Watson, A., Bakker, D.C.E., Schuster, U., Metzl, N., Yoshikawa-Inoue, H., Ishii, M., Midorikawa, T., Nojiri, Y., Körtzinger, A., Steinhoff, T., Hopemman, M., Olafsson, J., Arnarson, T.S., Tilbrook, B., Johannessen, T., Olsen, A., Bellerby, R., Wong, C.S., Delille, B., Bates, N.R., de Baar, H.J.W., 2009. Climatological mean and decadal change in surface ocean pCO₂, and net sea–air CO₂ flux over the global oceans. *Deep-Sea Research II* 56 (8–10), 554–577. <http://dx.doi.org/10.1016/j.dsr2.2008.12.009>.

Climate Feedbacks

AE Dessler, Texas A&M University, College Station, TX, USA

MD Zelinka, Lawrence Livermore National Laboratory, Livermore, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

Feedbacks modify an initial warming of the climate system, caused, for example, by increasing carbon dioxide. We discuss here the feedbacks that are of primary importance on decadal timescales: the water vapor feedback, the lapse-rate feedback, the surface albedo feedback, and the cloud feedback. Together, they account for approximately two-thirds of the warming we expect over the twenty-first century. The strongest positive feedback is the water vapor feedback, with the surface albedo and cloud feedbacks being smaller, positive feedbacks. The lapse-rate feedback is a negative feedback that offsets some of the water vapor feedback. The cloud feedback is the most uncertain one, and it is responsible for much of the spread among climate models in predictions of future climate change.

Planetary Energy Balance

Using an energy-balance perspective, we can represent energy flows within the climate system with a simple difference equation:

$$H = \Delta F - \lambda \Delta T \quad [1]$$

where ΔT is the global average surface temperature anomaly (an anomaly is the departure from a predefined reference state) and H is the top-of-atmosphere (TOA) net energy imbalance. ΔF is the radiative forcing, which is an energy imbalance imposed on the Earth's climate system from, for example, changes in the Sun's intensity, the addition of aerosols from a volcanic eruption, or the addition of greenhouse gases to our atmosphere (Forster et al., 2007).

In response to the energy imbalance imposed by the forcing, the Earth's temperature will change in order to reestablish energy balance for the planet. Setting $H = 0$, we find that the equilibrium temperature change in response to the radiative forcing is:

$$\Delta T = \Delta F / \lambda \quad [2]$$

Thus, $1/\lambda$ is how much the climate system must warm per unit of radiative forcing, and it is one measure of what is commonly referred to as the climate sensitivity.

λ can be decomposed into a sum of terms:

$$\lambda = \lambda_0 + \sum \lambda_i \quad [3]$$

where λ_0 is the 'no feedback' response (discussed in The 'No Feedback' Response section) and λ_i are various feedbacks (discussed in Feedbacks section).

The 'No Feedback' Response

In order to define feedbacks, one must first define a 'no feedback' response: the response before any feedbacks have influenced the climate. The most common definition of the 'no feedback' response is a vertically uniform warming of the surface and atmosphere, with everything else – atmospheric water vapor, clouds, surface albedo, etc. – held constant (See **Climate and Climate Change: Greenhouse Effect**). This is frequently referred to as the Planck response (sometimes, it is

also referred to as the Planck feedback, although, technically, it should not be considered a feedback). Given this definition of a 'no feedback' response, changes in water vapor, lapse rate, clouds, and surface albedo in response to a warming surface, are all categorized as feedbacks.

We can make a very simple estimate of the magnitude of λ_0 by taking the derivative of the Stefan–Boltzmann relation, $d(\sigma T^4)/dT = 4\sigma T^3$, and evaluating it at 255 K, the effective radiating temperature of the Earth. This yields $-3.7 \text{ W m}^{-2} \text{ K}^{-1}$ (the sign convention is positive fluxes are downward, so negative numbers indicate that a warmer planet radiates more energy to space). What this means is that, everything being constant, an increase in the temperature of the Earth will increase the power the Earth radiates to space by 3.7 W m^{-2} per degree.

Table 1 lists more sophisticated estimates based on radiative transfer calculations and realistic climate variations. This includes estimates from reanalysis data sets (ERA-Interim (Dee et al., 2011) and NASA's Modern Era Retrospective-analysis for Research and Application (MERRA) (Rienecker et al., 2011)), which are constructed from observations. These mainly represent the system's response to short-term climate variations, such as the El Niño/Southern Oscillation (ENSO). λ_0 from these data are $-3.1 \text{ W m}^{-2} \text{ K}^{-1}$.

Table 1 also includes estimates from an ensemble of 13 fully coupled global climate models (GCMs). The first estimate is from preindustrial control runs of the ensemble – runs in which greenhouse gas abundances and other forcings are held constant at their preindustrial concentrations, so there is no long-term climate change. The climate simulated in this ensemble is also dominated by short-term climate variations, so it is comparable to the calculations based on the reanalyses. This ensemble produces an average λ_0 of $-3.1 \text{ W m}^{-2} \text{ K}^{-1}$, in good agreement with the reanalyses. Values for the individual models can be found in Table S1 of Dessler (2013).

λ_0 can also be calculated from twenty-first century runs of the ensemble driven by the A1B emissions scenario, a moderate greenhouse-gas emissions scenario in which the Earth warms by several degrees Celsius over the twenty-first century. This ensemble produces an average λ_0 of $-3.3 \text{ W m}^{-2} \text{ K}^{-1}$, reasonably close to the value based on short-term climate variations and close to previous calculations (Colman, 2003; Soden and Held, 2006). Given radiative forcing from doubled

Table 1 All quantities are in $\text{W m}^{-2} \text{K}^{-1}$

Feedback	ERA-Interim		MERRA		Control runs		A1B runs	
	Feedback	Uncertainty	Feedback	Uncertainty	Feedback	Uncertainty	Feedback	Uncertainty
<i>Standard breakdown</i>								
Planck feedback	−3.09	0.07	−3.12	0.08	−3.12	0.12	−3.27	0.03
Lapse rate	−0.17	0.46	0.29	0.58	0.03	0.59	−0.86	0.28
Water vapor	1.35	0.35	1.12	0.39	1.38	0.62	2.00	0.21
Albedo	0.28	0.15	0.24	0.15	0.30	0.12	0.28	0.09
Total cloud	0.49	0.69	0.58	0.70	0.81	0.31	0.58	0.44
LW cloud	0.46	0.42	0.24	0.47	0.46	0.40	0.59	0.23
SW cloud	0.02	0.78	0.34	0.74	0.36	0.35	0.00	0.59
Thermal damping rate	−1.15	0.88	−0.89	0.88	−0.60	0.37	−1.26	0.45
<i>Constant-RH breakdown</i>								
Planck/RH	−1.92	0.03	−1.92	0.04	−1.91	0.05	−1.90	0.01
Lapse rate/RH	0.09	0.20	0.26	0.26	0.14	0.27	−0.28	0.13
ΔRH	−0.08	0.27	−0.05	0.30	0.06	0.19	0.06	0.10

For the ERA-Interim and MERRA, the uncertainties are 2σ . For the control and A1B model ensembles, the feedback is the average of the model ensemble and the uncertainty is the standard deviation of the ensemble.

Reproduced from Dessler, A.E., 2013. Observations of climate feedbacks over 2000–10 and comparisons to climate models. *J. Clim.* 26, 333–342. <http://dx.doi.org/10.1175/jcli-d-11-00640.1>.

carbon dioxide of 3.7 W m^{-2} (Myhre et al., 1998), eqn [2] predicts that the amount of warming in the absence of feedbacks is 1.2°C .

Identifying that part of the total climate response that constitutes the ‘no feedback’ response is somewhat arbitrary, Held and Shell (2012) suggested an alternative definition: a uniform warming of the surface and atmosphere, but with water vapor mixing ratio increasing so as to maintain constant relative humidity (RH) (Ingram, 2010). Table 1 shows that this alternative ‘no feedback’ climate response has a magnitude of $-1.9 \text{ W m}^{-2} \text{K}^{-1}$.

This value is less negative than the standard Planck response because the water vapor mixing ratio increases as a planet warms in order to maintain constant RH. Because water vapor is itself a greenhouse gas, this increase in water vapor traps infrared radiation and reduces the ability of the planet to increase radiation as it warms – so it only emits an additional 1.9 W m^{-2} for every degree of warming. Note also that changing the ‘no feedback’ climate response will necessarily change the definition of the climate feedbacks.

Feedbacks

It is worth reviewing at this point the chain of events that produce climate change. It begins with an imposed radiative forcing ΔF , which alters the planetary energy balance. This causes an initial change in the surface temperature ΔT_i , the magnitude of which is set by the ‘no feedback’ response.

In response to this initial warming, changes in the climate system occur that can further influence the climate – these are the climate feedbacks. These feedbacks generate an additional temperature change of $g\Delta T_i$ (g is a measure of the gain or strength of the feedbacks). But feedbacks also operate on the warming $g\Delta T_i$, and this leads to additional warming of

$g(g\Delta T_i) = g^2\Delta T_i$. Also feedbacks operate on this additional warming, leading to an additional warming of $g^3\Delta T_i$, etc. This goes on forever, so the final warming ΔT_f is

$$\Delta T_f = \Delta T_i + g\Delta T_i + g^2\Delta T_i + g^3\Delta T_i + g^4\Delta T_i \dots \quad [4]$$

This infinite series can be rewritten more simply as:

$$\Delta T_f = \frac{\Delta T_i}{(1 - g)} \quad [5]$$

The parameter g can be either positive or negative. Negative values of g correspond to feedbacks that ameliorate an initial temperature increase, and these are referred to as negative feedbacks. Positive values of g correspond to feedbacks that amplify the initial warming, and these are referred to as positive feedbacks. Note that $g > 1$ corresponds to ‘runaway’ greenhouse warming, where the warming from feedbacks outstrips the ability of the ‘no feedback’ response to shed energy (Roe, 2009). In such a situation, warming eventually ceases due to fundamental limitations in the feedbacks. For example, the ice-albedo feedback ceases when all of the ice on the planet has melted away, while the water vapor feedback eventually runs out of steam after the oceans boil away.

Equations [2] and [5] are equivalent if one recognizes that $\Delta T_i = \Delta F/\lambda_0$ and $1 - g = 1 + \frac{1}{\lambda_0} \sum \lambda_i$. λ_i is the strength of feedback i , and it has units of $\text{W m}^{-2} \text{K}^{-1}$. One can think of λ_i as a measure of how much this process alters TOA net flux per unit of surface warming. Putting this together, $g = -\frac{1}{\lambda_0} \sum \lambda_i$, so it is the sum of the individual λ_i 's divided by λ_0 .

The Water Vapor Feedback

The water vapor feedback is the process whereby an initial warming of the planet, caused, for example, by an increase in

atmospheric carbon dioxide, leads to an increase in the humidity of the atmosphere. Because water vapor is itself a greenhouse gas, this increase in humidity traps energy before it can escape to space, leading to additional warming. But not all regions of the atmosphere contribute equally: the water vapor feedback mainly results from changes in humidity in the tropical upper troposphere (Held and Soden, 2000), where temperatures are far below that of the surface and the vapor is mostly above clouds.

It had initially been argued that understanding tropical water vapor would be difficult because of the interaction between large-scale processes (e.g., the Hadley and Walker circulations) and small-scale processes (e.g., microphysics of clouds) (Emanuel and Pierrehumbert, 1995; Rennó et al., 1994). There was even the suggestion that microphysical effects could actually lead to a drying of the free troposphere as the climate warmed (Lindzen, 1990).

Subsequent research, however, demonstrated that the distribution of humidity in the climatologically important tropical upper troposphere is well reproduced by ‘large-scale control’ models, in which air leaves stormy regions in a saturated condition, but with negligible ice or liquid content. Water vapor is thereafter transported by the large-scale circulation, which conserves the specific humidity, except during subsequent saturation events, when loss of water occurs instantaneously to prevent supersaturation. Despite the simplicity of this idea, which entirely neglects detailed microphysics and other small-scale processes, such models accurately reproduce the observed water vapor distribution for the mid and upper troposphere (Sherwood, 1996; Dessler and Sherwood, 2000; Folkins et al., 2002; Pierrehumbert et al., 2007; Dessler and Minschwaner, 2007).

It has long been expected (Manabe and Wetherald, 1967) that the atmosphere’s RH would remain roughly constant – meaning that the specific humidity would exponentially increase with a warming climate. Minschwaner and Dessler (2004) created a simple radiative–convective model with interactive water vapor that showed how the couplings between the moistening effects of convective detrainment and the drying effects from clear-air subsidence, constrained by a balance between adiabatic warming and radiative cooling, produces a near constant, but slightly declining RH as the climate warms. Climate models also reproduce the slight reduction in RH in the tropical mid to upper troposphere as the climate warms (Soden and Held, 2006).

The first observational estimates of the water vapor feedback tested whether the atmosphere did indeed preserve constant RH during climate variations due to volcanic eruptions (Soden et al., 2002), ENSO and associated short-term internal climate variations (e.g., Minschwaner and Dessler, 2004), and decadal climate change (Soden et al., 2005). These analyses all concluded that RH is nearly conserved during these climate variations, confirming a strongly positive water vapor feedback.

More recently, the magnitude of the water vapor feedback, λ_{wv} , has been quantified as the change in TOA net energy flux due to the increase in water vapor per degree of surface warming. Observational estimates of λ_{wv} based on a volcanic eruption (Forster and Collins, 2004) and ENSO variations (Dessler et al., 2008; Dessler and Wong, 2009; Dessler 2013)

indicate that λ_{wv} is in the range 1.0–2.0 W m^{−2} K^{−1}. These latter three studies used off-line radiative transfer calculations (so-called radiative kernels (Soden et al., 2008; Shell et al., 2008)), which have become a popular method of calculating feedbacks.

The water vapor feedback has also been calculated in GCMs in response to short-term climate variations and in response to long-term warming. In response to short-term climate variations, λ_{wv} in GCMs is $\sim 1.1\text{--}1.3$ W m^{−2} K^{−1} (Colman and Power, 2010; Dessler, 2013), while that in response to long-term warming is 1.8–2.0 W m^{−2} K^{−1} (Colman and Power, 2010; Dessler, 2013; Colman, 2003; Soden and Held, 2006). These values are summarized in Table 1.

Thus, the water vapor feedback cancels some of the increased outgoing longwave radiation due to warmer atmospheric and surface temperatures. The net result is that, for a given radiative forcing, the Earth has to warm more in order to reestablish energy balance with the water vapor feedback than without. In this way, the water vapor feedback amplifies an initial warming, making it a positive feedback.

Comparisons between GCMs and observations show that some GCMs have large biases in their simulated present-day water vapor fields (John and Soden, 2007). Thus, it may be surprising that the GCMs generally agree among themselves and with observations on the value of the water vapor feedback (although some differences do exist, as described later). It turns out that the absorption of longwave radiation by water vapor is proportional to the logarithm of the amount of water, so it is the fractional changes in water vapor that determine the strength of the water vapor feedback – and all GCMs simulate similar fractional changes in the water vapor field as the surface warms (John and Soden, 2007). This occurs because tropospheric water vapor in the tropical upper troposphere is set by the temperature of convective detrainment through the Clausius–Clapeyron relation, which is a fundamental physical relation that all climate models incorporate.

Lapse-Rate Feedback

The standard ‘no feedback’ climate response is a uniform warming of the surface and atmosphere. However, the atmosphere’s temperature can also change relative to the surface, and such changes constitute the lapse-rate feedback. For example, the temperature of the tropics generally follows a moist adiabat set by the surface temperature; thus we expect the upper troposphere to warm about 2–3 times as much as the surface (Xu and Emanuel, 1989; Holloway and Neelin, 2007).

The enhanced radiation from the warmer upper troposphere enhances radiative cooling to space compared to the ‘no feedback’ response. Thus, the decrease of the lapse rate tends to reduce the initial temperature perturbation, leading to a negative feedback. In response to long-term warming, λ_{lr} in GCMs is -0.9 W m^{−2} K^{−1} (Colman and Power, 2010; Dessler, 2013; Soden and Held, 2006). Values of λ_{lr} during short-term climate variations (from both models and the reanalyses) are much closer to zero (Colman and Power, 2010; Dessler, 2013). These values are listed in Table 1.

The Connection between the Water Vapor and Lapse-Rate Feedbacks

There is some variation in the magnitude of the water vapor feedback among the GCMs (Colman, 2003; Soden and Held, 2006; Dessler, 2013). The cause of this is differences among the models in the amount of upper tropospheric warming per unit of surface warming (John and Soden, 2007), which fundamentally controls upper tropospheric water vapor through the Clausius–Clapeyron relation (Minschwaner and Dessler, 2004).

Thus, a model whose upper troposphere warms more than the ensemble average will have a positive water vapor feedback that is larger than the ensemble average. The warmer upper troposphere also radiates more power to space, so this model will also have a more negative lapse-rate feedback than the ensemble average. These effects largely cancel – so that the sum of the water vapor and lapse-rate feedbacks shows much less spread among the models than either feedback individually (Colman, 2003; Soden and Held, 2006).

Magnitudes of both the (positive) water vapor and (negative) lapse-rate feedbacks are smaller for short-term variability than for long-term climate change in climate models (Table 1). However, for reasons discussed above, the sum of the water vapor and lapse-rate feedbacks is similar for these two climate variations.

In the alternative breakdown of Held and Shell (2012), the ‘no feedback’ response and lapse-rate feedback both include a constant-RH assumption. Because the atmosphere nearly conserves RH in the global average, assuming constant RH essentially folds the water vapor feedbacks into these other terms. By combining the water vapor and lapse-rate feedbacks, this approach produces a new lapse-rate feedback that is more certain than in the standard breakdown. In addition, by eliminating the water vapor feedback, this alternative breakdown avoids having a large positive feedback (water vapor) canceling a large negative feedback (the lapse rate), which occurs in the conventional breakdown.

Surface Albedo Feedback

As the planet warms, the spatial coverage of ice and snow decreases, exposing the darker underlying surface. The resultant decrease in surface albedo (the ratio of reflected to incident solar radiation flux at the surface) increases the amount of solar radiation absorbed by the planet, thereby generating additional warming. This process therefore represents a positive feedback.

As feedbacks are defined at the TOA, one must relate the temperature-mediated change in surface albedo to that at the TOA (Hall and Qu, 2006):

$$\lambda_{\alpha} = \frac{dSW_{\alpha}}{dT} = SW_{in} \frac{d\alpha_s}{dT} \frac{\partial \alpha_p}{\partial \alpha_s} \quad [6]$$

where SW_{in} is the incident SW flux at the TOA, α_s is the surface albedo, and α_p is the planetary albedo. Thus, the surface albedo feedback depends on the incident solar radiation, the sensitivity of surface albedo to the change in global mean surface temperature, and the sensitivity of planetary albedo to changes in surface albedo. This final term is roughly 0.3 in both observations and models (Donohoe and Battisti, 2011), which

means that roughly 70% of the variations in surface albedo are hidden by overlying cloud cover and therefore do not impact the TOA energy budget.

In climate models, the global mean surface albedo feedback λ_{α} is roughly $0.3 \text{ W m}^{-2} \text{ K}^{-1}$ in response to long-term global warming (Soden et al., 2008; Vial et al., 2013); in response to short-term climate variations, both models and observations show a similar feedback strength (Dessler, 2013) (Table 1). Unsurprisingly, the surface albedo feedback primarily arises from changes in high latitudes (Crook et al., 2011), with about 75% coming from the Northern Hemisphere, with half from melting snow and half from melting sea ice (Winton, 2006). The remaining 25% comes from retreat of sea ice in the Southern Ocean. This is one of the reasons that high latitudes are expected to warm more than the tropics over the twenty-first century.

Intermodel differences in the sensitivity of the surface albedo to surface temperature ($d\alpha_s/dT$) are by far the dominant source of intermodel disagreement in surface albedo feedback, spanning a threefold range across models (Qu and Hall, 2006, 2014; Winton, 2006). Intermodel spread in the snow albedo feedback mainly arises from intermodel spread in the albedo decrease from reductions in areal coverage of snow, which itself is mainly attributable to the intermodel spread in mean effective snow albedo (Qu and Hall, 2007). Thus, models with a greater mean-state snow albedo generally experience a larger snow albedo feedback. Hall and Qu (2006) showed that the modeled response of Northern Hemisphere springtime snow albedo to temperature in the present-day seasonal cycle is an excellent predictor of its response to climate change, implying that this component of surface albedo feedback can be observationally constrained.

In a similar manner to the snow albedo feedback, the sea ice-albedo feedback depends strongly on the areal coverage and albedo of ice. Models with thinner mean-state sea ice tend to have larger sea ice-albedo feedbacks because thinner ice is easier to melt. Similarly, models with larger mean-state sea ice extents have more ice available to lose, and therefore tend to have larger sea ice-albedo feedbacks (Rind et al., 1995; Holland and Bitz, 2003).

Flanner et al. (2011) estimate a Northern Hemisphere surface albedo feedback between 0.3 and $1.1 \text{ W m}^{-2} \text{ K}^{-1}$ based on the observed decrease in surface albedo between 1979 and 2008, suggesting that current models are underestimating this feedback. Moreover, the observed decline in the spatial extent of Arctic sea ice at the end of the melt season between 1953 and 2006 is proceeding at a much faster rate than is predicted by any current climate models, also suggesting that sea ice cover may also be too insensitive to warming in the models (Stroeve et al., 2007).

Cloud Feedback

Clouds have a large impact on the Earth’s TOA energy budget, both in the longwave (LW, wavelengths >4 microns) and in the shortwave (SW, wavelengths <4 microns). These are commonly quantified using cloud radiative effect (CRE; Charlock and Ramanathan, 1985), which is the difference between clear-sky (R_{clear}) and all-sky (R_{total}) TOA upwelling flux:

$$CRE = R_{clear} - R_{total} \quad [7]$$

R_{clear} is the flux that would occur if clouds were instantaneously removed from the atmosphere, but all other parameters stayed the same. Also, note that CRE is sometimes referred to as cloud radiative forcing or CRF.

Breaking a scene into an overcast portion with area f and a clear portion with area $(1 - f)$, one can express the total upwelling radiation as:

$$R_{\text{total}} = fR_{\text{cloud}} + (1 - f)R_{\text{clear}} \quad [8]$$

where R_{cloud} is the TOA flux from the cloudy region. Combining [7] and [8] yields:

$$\text{CRE} = f(R_{\text{clear}} - R_{\text{cloud}}) \quad [9]$$

The LW, SW, and net CRE of idealized overcast scenes characterized by single-layer clouds with tops at various pressures and optical depths are shown in Figure 1. The greater the cloud fraction, or the greater the contrast in upwelling SW flux between clear and overcast locations (which depends on both the albedo contrast and the incident solar radiation), the greater the SW CRE. As cloud albedo is almost always larger than the clear-sky albedo, the SW CRE is negative, with a global and annual mean of roughly -50 W m^{-2} (Allan, 2011). Similarly, the greater the cloud fraction, or the greater the temperature difference between the surface and the cloud top (which depends primarily on the cloud-top height), the greater the LW CRE. One can therefore think of the LW CRE as measuring the greenhouse effect of clouds. Because most clouds are found at locations colder than the surface, the upwelling LW flux from clouds is almost always smaller than that from clear-sky regions, making LW CRE positive. In the annual and global mean, the LW CRE is roughly 30 W m^{-2} (Allan, 2011). Thus, the net effect of clouds is to cool the planet by roughly 20 W m^{-2} .

Cloud feedback is a measure of how much this cooling effect of clouds changes per unit increase in global mean surface temperature. A positive global mean cloud feedback means that the cooling effect of clouds decreases in magnitude as the planet warms. Because the effect of clouds on the SW reflectivity tends to dominate over the effect of clouds on the LW heating of the planet (i.e., net CRE is negative), temperature-mediated decreases in total cloud fraction holding their other properties fixed would represent a positive cloud feedback. All else being equal, temperature-mediated increases in cloud-top altitude (which reduces LW emission from clouds) and decreases in cloud optical depth (which reduces SW reflection more than it increases LW emission) are positive cloud feedbacks. A shift of clouds from regions having large values of insolation (low latitudes) to regions having lower values of insolation (high latitudes), holding all else constant, would also constitute a positive cloud feedback, as the amount of SW radiation reflected to space depends on the amount that is incident on the cloud. Conversely, temperature-mediated increases in total cloud amount, decreases in cloud-top altitude, increases in cloud optical depth, and shifts of clouds toward regions with larger incident solar radiation represent negative cloud feedbacks.

There are a wide variety of possible cloud responses to warming. The cloud feedback is unique among the climate feedbacks in that its net global mean value is the integrated effect of a multitude of processes that may oppose each other

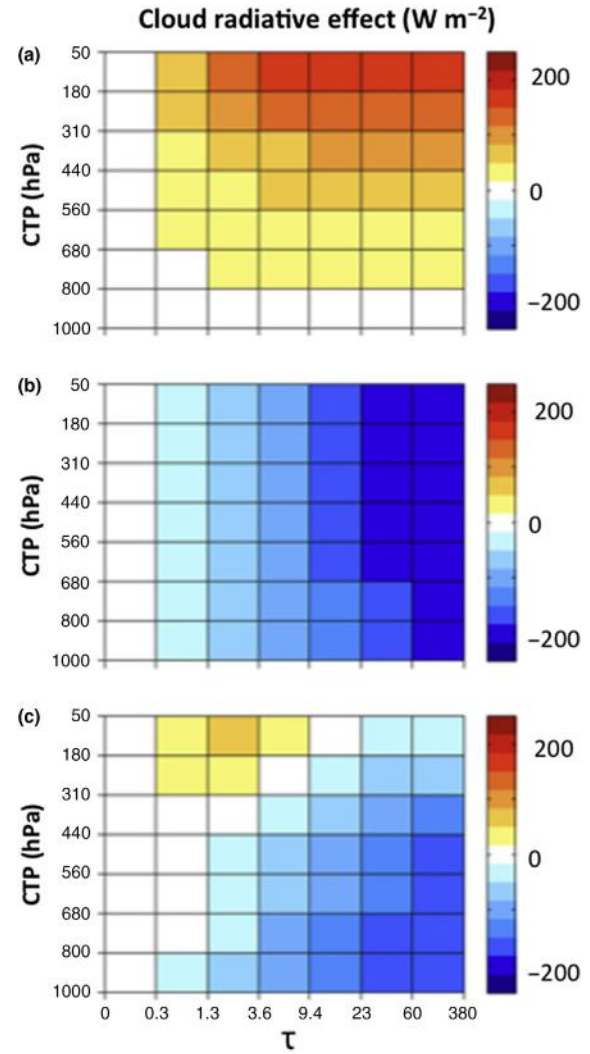


Figure 1 Globally averaged and annually averaged (a) LW, (b) SW, and (c) net cloud radiative effects for completely overcast scenes containing clouds with given properties of cloud-top pressure (CTP) and cloud optical depth (τ). Cloud optical depth is plotted on a logarithmic scale due to the linear relationship between $\log(\tau)$ and albedo. In the LW, τ is a proxy for cloud emissivity. Figure modified from Zelinka, M.D., Klein, S.A., Hartmann, D.L., 2012a. Computing and partitioning cloud feedbacks using cloud property histograms. Part I: cloud radiative kernels. *Journal of Climate* 25, 3715–3735. <http://dx.doi.org/10.1175/jcli-d-11-00248.1>.

horizontally, vertically, in optical depth space, and spectrally. This level of complexity, in addition to the fact that clouds are parameterized rather than explicitly resolved in GCMs, makes it the most uncertain feedback operating in climate models, and persistently so. For a given radiative forcing, most of the spread among climate model predictions of future warming arises from intermodel differences in cloud feedback (Cess and Potter, 1988; Cess et al., 1990; Webb et al., 2006; Dufresne and Bony, 2008; Andrews et al., 2012; Vial et al., 2013).

Despite large intermodel disagreement in the exact magnitude of cloud changes, several features are robust. GCMs systematically predict a fairly spatially uniform increase in high cloud-top altitude that is likely related to deepening of the

well-mixed troposphere as the planet warms, causing an increase in LW trapping and a positive net cloud altitude feedback of roughly $0.3 \text{ W m}^{-2} \text{ K}^{-1}$ (Zelinka et al., 2012b; Zelinka et al., 2013). The tendency for tropical high clouds to systematically rise as the planet warms is grounded fairly strongly in theory (Hartmann and Larson, 2002), and is supported by cloud-resolving model experiments (Tompkins and Craig, 1999; Kuang and Hartmann, 2007; Harrop and Hartmann, 2012), GCM experiments (Zelinka and Hartmann, 2010), and observations (Zelinka and Hartmann, 2011; Li et al., 2012).

GCMs also robustly predict small decreases in cloud optical depth for warm clouds at low latitudes and large increases for cold clouds at high latitudes and high altitudes as the planet warms (Zelinka et al., 2012b; Zelinka et al., 2013). The latter may be related to the temperature-induced increase in total cloud water content, which is more sensitive at the cold temperatures characteristic of high latitude clouds (Betts and Harshvardhan, 1987), or to the preferential increase of liquid relative to ice at warmer temperatures, which results in brighter clouds (Tsushima et al., 2006). The negative feedback due to enhanced SW reflection from optically thicker clouds is negative in the global mean in the vast majority of current models, averaging about $-0.15 \text{ W m}^{-2} \text{ K}^{-1}$. It is the dominant cloud feedback at high latitudes.

In all current models, global average total cloud fraction systematically decreases as the planet warms, contributing a positive feedback of about $0.15 \text{ W m}^{-2} \text{ K}^{-1}$ because the changes in the SW dominate those in the LW (Zelinka et al., 2012b; Zelinka et al., 2013). The tendency for total cloud fraction to decrease at low latitudes and increase at high latitudes represents a positive feedback owing to the overall shift of clouds toward regions with less insolation. A portion of this pattern may be related to the poleward shift of the major circulation features (i.e., the relatively cloud-free subtropical dry zone expands and the cloudy midlatitude storm track region shifts poleward (Hall et al., 1994; Yin, 2005)). Despite most models predicting a decrease in the amount of subtropical marine boundary layer clouds in regions of moderate subsidence, subtle differences among models in the magnitude of this change, when weighted by the expansive coverage of these cloud types, translates into large intermodel differences in cloud feedback and climate sensitivity (Bony and Dufresne, 2005; Soden and Vecchi, 2011; Webb et al., 2013). Their properties are known to be sensitive to a multitude of processes including boundary layer turbulent fluxes, lower tropospheric stability, cloud-top entrainment, free-tropospheric subsidence, and cloud-top radiative cooling, some of which are parameterized, expected to change in warming world, and subject to considerable uncertainty.

Obtaining observational estimates of cloud feedback, though highly desirable, is difficult owing to the dearth of observational data of suitable length and/or quality for identifying the subtle responses of clouds to the limited amount of warming that has been realized in nature. The global mean net cloud feedback in response to short-term (interannual) fluctuations in nature is estimated to be $0.50 \text{ W m}^{-2} \text{ K}^{-1}$ with a 2σ uncertainty of $0.75 \text{ W m}^{-2} \text{ K}^{-1}$ (Dessler, 2010; Dessler and Loeb, 2013), which is in agreement with the ensemble average short-term cloud feedback in climate models (Table 1). The ensemble average cloud feedback in response to long-term

climate change is similar to the average short-term cloud feedback, but little cross-model correlation exists between the long-term and short-term cloud feedbacks (Dessler, 2010), so it is presently unclear whether the short-term cloud feedback has relevance for the feedback in response to long-term warming. While we cannot verify the long-term cloud feedback using observations, the agreement between the models and observations for the short-term cloud feedback provides some confidence in the models' long-term cloud feedback. Moreover, specific aspects of cloud responses to natural climate variability may be useful for testing physical mechanisms of cloud feedbacks to long-term climate change (Tselioudis et al., 1998; Clement et al., 2004; Zelinka and Hartmann, 2011).

Slow Feedbacks

The feedbacks discussed above are the so-called fast feedbacks, meaning that they affect the climate on timescales of a week (water vapor and temperature) to a few years (sea ice). There are other feedbacks, however, that operate on much longer timescales. For example, the melting of the earth's great ice sheets is a positive albedo feedback, but one that occurs on millennial timescales. The timescale and strength of carbon cycle feedbacks – e.g., a warming climate melts permafrost, leading to the release of more carbon dioxide – is unknown, but probably occurs on centennial timescales.

Probably the longest timescale feedback is the weathering thermostat. In this feedback, warmer temperatures lead to higher rates of precipitation and enhanced weathering of rocks. The weathering reaction removes carbon dioxide from the atmosphere and transports it into the ocean. The resulting reduction of carbon dioxide offsets the initial warming, leading to a negative feedback. The long-term cooling since the Eocene (over the last 50 million years) has been hypothesized to be the result of the reduction in carbon dioxide due to enhanced precipitation on rock newly exposed from the uplift of the Himalayas and Tibetan plateau. While this is clearly a powerful feedback, its million-year timescale means that it will not act to ameliorate global warming over the next century.

Overall, there is some evidence, based on events such as the Paleocene–Eocene Thermal Maximum, that these slow feedbacks are net positive (Zeebe et al., 2009). However, given the uncertainty in both the temperature and the forcing data for these deep-time events, such a conclusion should not be considered definitive.

Summary

In response to a forcing of the climate system, the temperature of the planet adjusts until the planet's energy balance is restored. How much warming is required per unit of radiative forcing is a measure of the climate sensitivity. And this climate sensitivity can be broken down into a 'no feedback' response and a set of climate feedbacks that amplify or ameliorate it.

Without any feedbacks, a doubling of carbon dioxide produces a warming of $\sim 1.2^\circ\text{C}$. Including feedbacks increases this to $\sim 3^\circ\text{C}$. Thus, feedbacks are responsible for the bulk of the warming that the Earth will experience over the next

century. Estimates of the magnitude of the individual fast feedbacks (those that respond on timescales of a few years or less) are listed in [Table 1](#). The water vapor feedback is the climate system's primary positive feedback, and by itself it roughly doubles the 'no feedback' warming. The albedo and cloud feedbacks are smaller, positive feedbacks. While the cloud feedback is small compared to the water vapor feedback, it is the most uncertain feedback and is responsible for much of the spread in climate models' predictions of the climate sensitivity. The lapse-rate feedback is negative, meaning that it partially offsets the warming from the 'no feedback' response and the other feedbacks.

There is some evidence that the strength of feedbacks may differ for short-term climate variability and long-term warming. It is the feedbacks in response to long-term warming that interest us the most (because they are key to estimating warming over the next century) – but, unfortunately, the lack of long-term observations means that we cannot calculate the long-term feedbacks from observations.

While uncertainties in our estimates of the feedbacks and the resulting climate sensitivities exist, our confidence in our overall understanding of these processes is generally high. This arises primarily from a convergence between observations and models and theoretical expectations. Confidence also comes from the stability of understanding. Our present understanding of climate feedbacks, as described here, is in good agreement with estimates made throughout the twentieth century, including the first comprehensive modern studies of climate sensitivity and feedbacks ([Hansen et al., 1984](#); [Zhang et al., 1994](#); [Colman and McAvaney, 1997](#)).

Acknowledgments

AED acknowledges support from NSF grant AGS-1012665 to Texas A&M University. MDZ's contribution was supported by the Regional and Global Climate Modeling Program of the United States Department of Energy's Office of Science and was performed under the auspices of the United States Department of Energy by Lawrence Livermore National Laboratory under contract DE-AC52-07NA27344.

See also: **Climate and Climate Change:** History of Scientific Work on Climate Change. **Clouds and Fog:** Classification of Clouds. **Global Change:** Biospheric Impacts and Feedbacks. **Satellites and Satellite Remote Sensing:** Water Vapor.

References

- Allan, R.P., 2011. Combining satellite data and models to estimate cloud radiative effect at the surface and in the atmosphere. *Meteorological Applications* 18, 324–333. <http://dx.doi.org/10.1002/met.285>.
- Andrews, T., Gregory, J.M., Webb, M.J., Taylor, K.E., 2012. Forcing, feedbacks and climate sensitivity in CMIP5 coupled atmosphere–ocean climate models. *Geophysical Research Letters* 39. <http://dx.doi.org/10.1029/2012gl051607>.
- Betts, A.K., Harshvardhan, 1987. Thermodynamic constraint on the cloud liquid water feedback in climate models. *Journal of Geophysical Research* 92, 8483–8485. <http://dx.doi.org/10.1029/JD092iD07p08483>.
- Bony, S., Dufresne, J.L., 2005. Marine boundary layer clouds at the heart of tropical cloud feedback uncertainties in climate models. *Geophysical Research Letters* edited, p. 10.1029/2005GL023851.
- Cess, R.D., Potter, G.L., 1988. A methodology for understanding and intercomparing atmospheric climate feedback processes in general-circulation models. *Journal of Geophysical Research* 93, 8305–8314.
- Cess, R.D., et al., 1990. Intercomparison and interpretation of climate feedback processes in 19 atmospheric general circulation models. *Journal of Geophysical Research* 95, 16601–16615.
- Charlock, T.P., Ramanathan, V., 1985. The albedo field and cloud radiative forcing produced by a general circulation model with internally generated cloud optics. *Journal of Atmospheric Sciences* 42, 1408–1429. [http://dx.doi.org/10.1175/1520-0469\(1985\)042<1408:tafacr>2.0.co;2](http://dx.doi.org/10.1175/1520-0469(1985)042<1408:tafacr>2.0.co;2).
- Clement, A.C., Hall, A., Broccoli, A.J., 2004. The importance of precessional signals in the tropical climate. *Climate Dynamics* 22, 327–341. <http://dx.doi.org/10.1007/s00382-003-0375-8>.
- Colman, R., 2003. A comparison of climate feedbacks in general circulation models. *Climate Dynamics* 20, 865–873.
- Colman, R.A., McAvaney, B.J., 1997. A study of general circulation model climate feedbacks determined from perturbed sea surface temperature experiments. *Journal of Geophysical Research* 102, 19383–19402. <http://dx.doi.org/10.1029/97jd00206>.
- Colman, R.A., Power, S.B., 2010. Atmospheric radiative feedbacks associated with transient climate change and climate variability. *Climate Dynamics* 34, 919–933. <http://dx.doi.org/10.1007/s00382-009-0541-8>.
- Crook, J.A., Forster, P.M., Stuber, N., 2011. Spatial patterns of modeled climate feedback and contributions to temperature response and polar amplification. *Journal of Climate* 24, 3575–3592. <http://dx.doi.org/10.1175/2011jcli3863.1>.
- Dee, D.P., et al., 2011. The ERA-Interim reanalysis: configuration and performance of the data assimilation system. *Quarterly Journal of the Royal Meteorological Society* 137, 553–597. <http://dx.doi.org/10.1002/qj.828>.
- Dessler, A.E., 2010. A determination of the cloud feedback from climate variations over the past decade. *Science* 330, 1523–1527. <http://dx.doi.org/10.1126/science.1192546>.
- Dessler, A.E., 2013. Observations of climate feedbacks over 2000–10 and comparisons to climate models. *Journal of Climate* 26, 333–342. <http://dx.doi.org/10.1175/jcli-d-11-00640.1>.
- Dessler, A.E., Loeb, N.G., 2013. Impact of dataset choice on calculations of the short-term cloud feedback. *Journal of Geophysical Research* 118, 2821–2826. <http://dx.doi.org/10.1002/2012JD018713>.
- Dessler, A.E., Minschwaner, K., 2007. An analysis of the regulation of tropical tropospheric water vapor. *Journal of Geophysical Research* 112, D10120. <http://dx.doi.org/10.1029/2006JD007683>.
- Dessler, A.E., Sherwood, S.C., 2000. Simulations of tropical upper tropospheric humidity. *Journal of Geophysical Research* 105 (D15), 20155–20163.
- Dessler, A.E., Wong, S., 2009. Estimates of the water vapor climate feedback during the El Niño Southern Oscillation. *Journal of Climate* 22, 6404–6412. <http://dx.doi.org/10.1175/2009JCLI3052.1>.
- Dessler, A.E., Yang, P., Zhang, Z., 2008. The water-vapor climate feedback inferred from climate fluctuations, 2003–2008. *Geophysical Research Letters* 35, L20704. <http://dx.doi.org/10.1029/2008GL035333>.
- Donohoe, A., Battisti, D.S., 2011. Atmospheric and surface contributions to planetary albedo. *Journal of Climate* 24, 4402–4418. <http://dx.doi.org/10.1175/2011jcli3946.1>.
- Dufresne, J.L., Bony, S., 2008. An assessment of the primary sources of spread of global warming estimates from coupled atmosphere–ocean models. *Journal of Climate* 21, 5135–5144. <http://dx.doi.org/10.1175/2008jcli2239.1>.
- Emanuel, K.A., Pierrehumbert, R.T., 1995. Microphysical and dynamical control of tropospheric water vapor. In: Crutzen, P.J., Ramanathan, V. (Eds.), *Clouds, Chemistry, and Climate*. Springer, New York, pp. 17–28.
- Flanner, M.G., Shell, K.M., Barlage, M., Perovich, D.K., Tschudi, M.A., 2011. Radiative forcing and albedo feedback from the Northern Hemisphere cryosphere between 1979 and 2008. *Nature Geoscience* 4, 151–155. <http://dx.doi.org/10.1038/ngeo1062>.
- Folkins, I., Kelly, K.K., Weinstock, E.M., 2002. A simple explanation for the 11 to 15 km increase in tropical relative humidity. *Journal of Geophysical Research* 107, 4736. <http://dx.doi.org/10.1029/2002JD002185>.
- Forster, P.M.D., Collins, M., 2004. Quantifying the water vapour feedback associated with post-Pinatubo global cooling. *Climate Dynamics* 23, 207–214.
- Forster, P., et al., 2007. Changes in atmospheric constituents and in radiative forcing. In: Solomon, S., Qin, D., Manning, M., Chen, Z., Marquis, M., Averyt, K.B., Tignor, M., Miller, H.L. (Eds.), *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, United Kingdom and New York, NY, USA.

- Hall, A., Qu, X., 2006. Using the current seasonal cycle to constrain snow albedo feedback in future climate change. *Geophysical Research Letters* 33, L03502. <http://dx.doi.org/10.1029/2005gl025127>.
- Hall, N.M.J., Hoskins, B.J., Valdes, P.J., Senior, C.A., 1994. Storm tracks in a high-resolution GCM with doubled carbon-dioxide. *Quarterly Journal of the Royal Meteorological Society* 120, 1209–1230. <http://dx.doi.org/10.1002/qj.49712051905>.
- Hansen, J., Lacis, A., Rind, D., Russell, G., Stone, P., Fung, I., Ruedy, R., Lerner, J., 1984. Climate sensitivity: analysis of feedback mechanisms. In: *Climate Processes and Climate Sensitivity*. AGU, Washington, pp. 130–163.
- Harrop, B.E., Hartmann, D.L., 2012. Testing the role of radiation in determining tropical cloud-top temperature. *Journal of Climate* 25, 5731–5747. <http://dx.doi.org/10.1175/jcli-d-11-00445.1>.
- Hartmann, D.L., Larson, K., 2002. An important constraint on tropical cloud – climate feedback. *Geophysical Research Letters* 29, 1951. <http://dx.doi.org/10.1029/2002GL015835>.
- Held, I., Shell, K.M., 2012. Using relative humidity as a state variable in climate feedback analysis. *Journal of Climate* 25, 2578–2582. <http://dx.doi.org/10.1175/JCLI-D-11-00721.1>.
- Held, I.M., Soden, B.J., 2000. Water vapor feedback and global warming. *Annual Review of Energy and the Environment* 25, 441–475.
- Holland, M.M., Bitz, C.M., 2003. Polar amplification of climate change in coupled models. *Climate Dynamics* 21, 221–232. <http://dx.doi.org/10.1007/s00382-003-0332-6>.
- Holloway, C.E., Neelin, J.D., 2007. The convective cold top and quasi equilibrium. *Journal of Atmospheric Sciences* 64, 1467–1487.
- Ingram, W., 2010. A very simple model for the water vapour feedback on climate change. *Quarterly Journal of the Royal Meteorological Society* 136, 30–40. <http://dx.doi.org/10.1002/qj.546>.
- John, V.O., Soden, B.J., 2007. Temperature and humidity biases in global climate models and their impact on climate feedbacks. *Geophysical Research Letters* 34, L18704. <http://dx.doi.org/10.1029/2007GL030429>.
- Kuang, Z., Hartmann, D.L., 2007. Testing the fixed anvil temperature hypothesis in a cloud-resolving model. *Journal of Climate* 20, 2051–2057. <http://dx.doi.org/10.1175/jcli4124.1>.
- Li, Y., Yang, P., North, G.R., Dessler, A., 2012. Test of the fixed anvil temperature hypothesis. *Journal of Atmospheric Sciences* 69, 2317–2328. <http://dx.doi.org/10.1175/jas-d-11-0158.1>.
- Lindzen, R.S., 1990. Some coolness concerning global warming. *Bulletin of the American Meteorological Society* 71, 288–299.
- Manabe, S., Wetherald, R.T., 1967. Thermal equilibrium of atmosphere with a given distribution of relative humidity. *Journal of Atmospheric Sciences* 24, 241–259.
- Minschwaner, K., Dessler, A.E., 2004. Water vapor feedback in the tropical upper troposphere: model results and observations. *Journal of Climate* 17, 1272–1282.
- Myhre, G., Highwood, E.J., Shine, K.P., Stordal, F., 1998. New estimates of radiative forcing due to well mixed greenhouse gases. *Geophysical Research Letters* 25, 2715–2718. <http://dx.doi.org/10.1029/98gl01908>.
- Pierrehumbert, R.T., Brogniez, H., Roca, R., 2007. On the relative humidity of the Earth's atmosphere. In: *The Global Circulation of the Atmosphere*. Princeton University Press.
- Qu, X., Hall, A., 2006. Assessing snow albedo feedback in simulated climate change. *Journal of Climate* 19, 2617–2630. <http://dx.doi.org/10.1175/jcli3750.1>.
- Qu, X., Hall, A., 2007. What controls the strength of snow-albedo feedback? *Journal of Climate* 20, 3971–3981. <http://dx.doi.org/10.1175/jcli4186.1>.
- Qu, X., Hall, A., 2014. On the persistent spread of snow-albedo feedback. *Climate Dynamics* 42, 69–81.
- Rennó, N.O., Emanuel, K.A., Stone, P.H., 1994. Radiative–convective model with an explicit hydrologic cycle. 1. Formulation and sensitivity to model parameters. *Journal of Geophysical Research* 99, 14,429–14,441.
- Rienecker, M.M., et al., 2011. MERRA – NASA's Modern-Era retrospective analysis for research and applications. *Journal of Climate* 24, 3624–3648. <http://dx.doi.org/10.1175/JCLI-D-11-00015.1>.
- Rind, D., Healy, R., Parkinson, C., Martinson, D., 1995. The role of sea-ice in $2 \times \text{CO}_2$ climate model sensitivity. 1. The total influence of sea ice thickness and extent. *Journal of Climate* 8, 449–463. [http://dx.doi.org/10.1175/1520-0442\(1995\)008<0449:trosii>2.0.co;2](http://dx.doi.org/10.1175/1520-0442(1995)008<0449:trosii>2.0.co;2).
- Roe, G., 2009. Feedbacks, timescales, and seeing red. *Annual Review of Earth and Planetary Sciences*, 93–115.
- Shell, K.M., Kiehl, J.T., Shields, C.A., 2008. Using the radiative kernel technique to calculate climate feedbacks in NCAR's Community Atmospheric Model. *Journal of Climate* 21, 2269–2282. <http://dx.doi.org/10.1175/2007jcli2044.1>.
- Sherwood, S.C., 1996. Maintenance of the free-tropospheric tropical water vapor distribution. Part II: simulation by large-scale advection. *Journal of Climate* 9, 2919–2934.
- Soden, B.J., Held, I.M., 2006. An assessment of climate feedbacks in coupled ocean–atmosphere models. *Journal of Climate* 19, 3354–3360. <http://dx.doi.org/10.1175/JCLI3799.1>.
- Soden, B.J., Held, I.M., Colman, R., Shell, K.M., Kiehl, J.T., Shields, C.A., 2008. Quantifying climate feedbacks using radiative kernels. *Journal of Climate* 21, 3504–3520. <http://dx.doi.org/10.1175/2007JCLI2110.1>.
- Soden, B.J., Jackson, D.L., Ramaswamy, V., Schwarzkopf, M.D., Huang, X., 2005. The radiative signature of upper tropospheric moistening. *Science* 310, 841–844.
- Soden, B.J., Vecchi, G.A., 2011. The vertical distribution of cloud feedback in coupled ocean–atmosphere models. *Geophysical Research Letters* 38, L12704. <http://dx.doi.org/10.1029/2011gl047632>.
- Soden, B.J., Wetherald, R.T., Stenchikov, G.L., Robock, A., 2002. Global cooling after the eruption of Mount Pinatubo: a test of climate feedback by water vapor. *Science* 296, 727–730.
- Stroeve, J., Holland, M.M., Meier, W., Scambos, T., Serreze, M., 2007. Arctic sea ice decline: faster than forecast. *Geophysical Research Letters* 34. <http://dx.doi.org/10.1029/2007gl029703>.
- Tompkins, A.M., Craig, G.C., 1999. Sensitivity of tropical convection to sea surface temperature in the absence of large-scale flow. *Journal of Climate* 12, 462–476. [http://dx.doi.org/10.1175/1520-0442\(1999\)012<0462:sotcts>2.0.co;2](http://dx.doi.org/10.1175/1520-0442(1999)012<0462:sotcts>2.0.co;2).
- Tselioudis, G., DelGenio, A.D., Kovari, W., Yao, M.S., 1998. Temperature dependence of low cloud optical thickness in the GISS GCM: contributing mechanisms and climate implications. *Journal of Climate* 11, 3268–3281. [http://dx.doi.org/10.1175/1520-0442\(1998\)011<3268:tdolco>2.0.co;2](http://dx.doi.org/10.1175/1520-0442(1998)011<3268:tdolco>2.0.co;2).
- Tsushima, Y., Emori, S., Ogura, T., Kimoto, M., Webb, M.J., Williams, K.D., Ringer, M.A., Soden, B.J., Li, B., Andronova, N., 2006. Importance of the mixed-phase cloud distribution in the control climate for assessing the response of clouds to carbon dioxide increase: a multi-model study. *Climate Dynamics* 27, 113–126. <http://dx.doi.org/10.1007/s00382-006-0127-7>.
- Vial, J., Dufresne, J.-L., Bony, S., 2013. On the interpretation of inter-model spread in CMIP5 climate sensitivity estimates. *Climate Dynamics*. <http://dx.doi.org/10.1007/s00382-013-1725-9>.
- Webb, M.J., Lambert, F.H., Gregory, J.M., 2013. Origins of differences in climate sensitivity, forcing and feedback in climate models. *Climate Dynamics* 40, 677–707. <http://dx.doi.org/10.1007/s00382-012-1336-x>.
- Webb, M.J., et al., 2006. On the contribution of local feedback mechanisms to the range of climate sensitivity in two GCM ensembles. *Climate Dynamics* 27, 17–38.
- Winton, M., 2006. Surface albedo feedback estimates for the AR4 climate models. *Journal of Climate* 19, 359–365. <http://dx.doi.org/10.1175/jcli3624.1>.
- Xu, K.M., Emanuel, K.A., 1989. Is the tropical atmosphere conditionally unstable? *Monthly Weather Review* 117, 1471–1479.
- Yin, J.H., 2005. A consistent poleward shift of the storm tracks in simulations of 21st century climate. *Geophysical Research Letters* 32. <http://dx.doi.org/10.1029/2005gl023684>.
- Zeebe, R.E., Zachos, J.C., Dickens, G.R., 2009. Carbon dioxide forcing alone insufficient to explain Palaeocene–Eocene Thermal Maximum warming. *Nature Geoscience* 2, 576–580. <http://dx.doi.org/10.1038/ngeo578>.
- Zelinka, M., Hartmann, D.L., 2010. Why is longwave cloud feedback positive? *Journal of Geophysical Research* 116, D16117. <http://dx.doi.org/10.1029/2010JD013817>.
- Zelinka, M.D., Hartmann, D.L., 2011. The observed sensitivity of high clouds to mean surface temperature anomalies in the tropics. *Journal of Geophysical Research* 116, D23103. <http://dx.doi.org/10.1029/2011jd016459>.
- Zelinka, M.D., Klein, S.A., Hartmann, D.L., 2012a. Computing and partitioning cloud feedbacks using cloud property histograms. Part I: cloud radiative kernels. *Journal of Climate* 25, 3715–3735. <http://dx.doi.org/10.1175/jcli-d-11-00248.1>.
- Zelinka, M.D., Klein, S.A., Hartmann, D.L., 2012b. Computing and partitioning cloud feedbacks using cloud property histograms. Part II: attribution to changes in cloud amount, altitude, and optical depth. *Journal of Climate* 25, 3736–3754. <http://dx.doi.org/10.1175/jcli-d-11-00249.1>.
- Zelinka, M.D., Klein, S.A., Taylor, K.E., Andrews, T., Webb, M.J., Gregory, J.M., Forster, P.M., 2013. Contributions of different cloud types to feedbacks and rapid adjustments in CMIP5. *Journal of Climate*. <http://dx.doi.org/10.1175/JCLI-D-12-00555.1>.
- Zhang, M.H., Hack, J.J., Kiehl, J.T., Cess, R.D., 1994. Diagnostic study of climate feedback processes in atmospheric general circulation models. *Journal of Geophysical Research* 99, 5525–5537. <http://dx.doi.org/10.1029/93JD03523>.

Climate Prediction: Empirical and Numerical

S Hastenrath, University of Wisconsin, Madison, WI, USA

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

In tropical climate prediction, three categories of approaches can be recognized: (1) empirical methods based on the combination of circulation diagnostics with statistical techniques, (2) numerical modeling, and (3) empirical but purely statistical procedures. It is desirable to pursue in parallel (1) and (2). By way of example, an overview is offered for five target regions: El Niño and the Southern Oscillation (ENSO), the Nordeste of Brazil, the West African Sahel, the boreal autumn rains of East Africa, and the Indian summer monsoon. Comparisons yield no advantage of numerical modeling over empirical approaches. ENSO contributes strongly to interannual climate variability in only limited regions of the tropics. Indispensable are the verification of forecast performance on an independent data set, the documentation and publication of the method, publication of the forecasts in real time, and continuous evaluation of forecasts.

Introduction

Attempts at climate prediction span more than a century, and a renewed interest has developed since the early 1980s. On theoretical grounds, it has been suggested that prospects for the prediction of annual variability should be better for the lower than the higher latitudes. Method development has been directed at a wide variety of regional targets, as illustrated in Figure 1. From experience, it seems more fruitful to pursue empirical and numerical modeling approaches concurrently. A 'code of good conduct' requires documentation of the method, assessment of method performance on an independent data set, and regular verification of real-time forecasts.

Methods

In the endeavors at climate prediction over the past two decades, three broad categories of approaches can be recognized: (1) empirical methods based on the combination of general circulation diagnostics and statistical techniques, (2) numerical modeling, and (3) empirical but purely statistical techniques, although there are combinations of and transitions between the categories. Categories (1) and (2) imply a diagnostic understanding of circulation mechanisms. In the empirically based approaches (1) and (3), a clear distinction must be made between the dependent portion of the record (or training period) from which the method was developed and the independent

data set reserved for the verification of forecast performance (verification period). An independent verification period is essential to protect against noise fitting, a severe risk especially where a large number of predictors is used after excessive screening. In category 2, verification of performance against observations over a couple of decades is also in order.

Approach 1 – general circulation and statistics – is illustrated in Figure 2. Studies of general circulation diagnostics are aimed at understanding the mechanisms of climate anomalies. On this basis, indices are selected that appear promising as predictors. Effective combinations of predictors and formulations of quantitative predictor–predictand relationships are ascertained by appropriate statistical methods, such as stepwise multiple regression (SMR), linear discriminant analysis (LDA), and neural networks. Other statistical procedures widely used include canonical correlation analysis (CCA), singular spectrum analysis (SSA), and empirical orthogonal function (EOF) analysis. The quantitative predictor–predictand relationship represents the prediction model. This then serves for calculating predictand values for a portion of the record that had not been used in the development of the prediction model. Quantitative comparison of the calculated versus the observed values of the predictand provides verification of forecast performance. The evolution from general circulation diagnostics to prognosis is exemplified by a series of papers on Brazil's northeast region, the Nordeste.

An understanding of the general circulation mechanisms of climate anomalies from empirical–diagnostic studies is also basic to approaching numerical modeling. This is exemplified

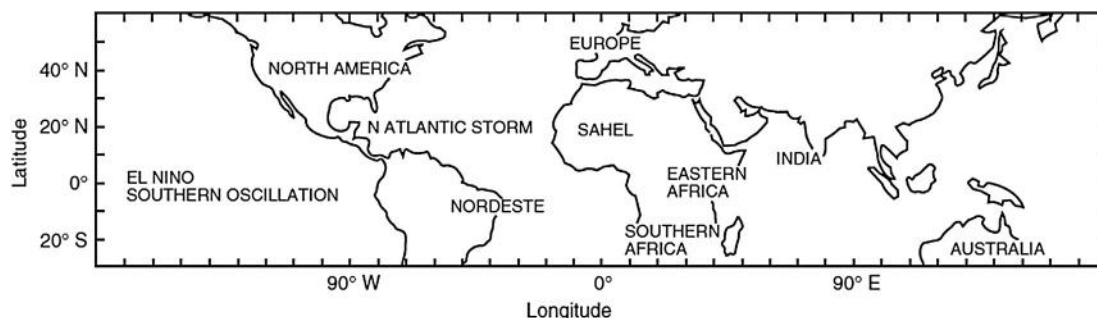


Figure 1 Orientation map showing the location of forecasting targets referred to in this article: India, eastern Africa, southern Africa, Sahel, Northeast Brazil (Nordeste), North Atlantic storms, El Niño and the Southern Oscillation, Australia, North America, and Europe.

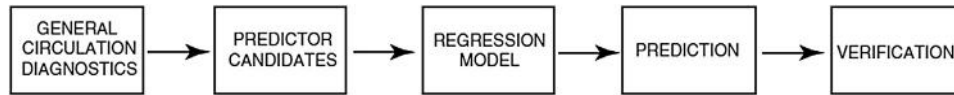


Figure 2 Scheme of climate prediction based on general circulation diagnostics and statistics. Adapted from Hastenrath, S., 1995a. Recent advances in tropical climate prediction. *Journal of Climate* 8, 1519–1532.

for the Pacific Ocean's El Niño phenomenon in the progression from the pioneering empirical oceanographic work over the first diagnostic ocean numerical studies to the application of numerical models for the operational prediction of El Niño. Similarly, the application of general circulation models (GCMs) to seasonal rainfall forecasting for the West African Sahel and Brazil's Nordeste has a broad empirical–diagnostic basis. In approach (2), the prediction model is developed from basic circulation theory rather than from past observations as in approach (1). Accordingly, the separation into dependent and independent portions of the record (as for approaches (1) and (2)) does not arise. Values calculated from previously observed conditions are again compared with the observed predictand values for verification. In principle, there are prospects of coupling the atmosphere, ocean, and land surface portions of the system. Purely statistical techniques (approach (3)) may seem unsatisfactory because of the lack of insight into the circulation mechanisms involved. Some of these endeavors merit attention, however, because of their remarkable practical performance.

Regional Targets

Methods for the forecasting of interannual variability have been developed for numerous regions, especially in the low latitudes. A series of examples are addressed here, namely Indian monsoon, eastern and southern Africa, Sahel, Northeast Brazil (Nordeste), North Atlantic storms, El Niño and Southern

Oscillation (ENSO), Australia, North America, and Europe, as identified in Figure 1.

Indian Monsoon

Empirical–diagnostic research over more than a century has led to a considerable understanding of the general circulation mechanisms of Indian monsoon rainfall anomalies and to the identification of numerous viable predictors. These can be loosely grouped into three classes, pertaining to the upper-air flow over India, to 'heat low' development over southern Asia and the establishment of a meridional pressure gradient and cross-equatorial flow over the Indian Ocean, and to the Southern Oscillation (SO, a high phase defined by anomalously high and low pressure at Tahiti and Darwin), although there are interrelations between the three classes. Figure 3 illustrates two simple prediction models based on variously used predictors and SMR. Conspicuously absent in the prediction of the Indian monsoon rainfall anomalies is the use of numerical models in the operational work of the India Meteorological Department or in the published literature. This is all the more remarkable because the sustained empirical work has afforded a sound understanding of the functioning of the general circulation of the monsoon and has shown Indian monsoon rainfall anomalies to be amenable to prediction. Along with these overall encouraging prospects for Indian monsoon forecasting, the reservation should be noted that the performance of various predictors has been known to vary in

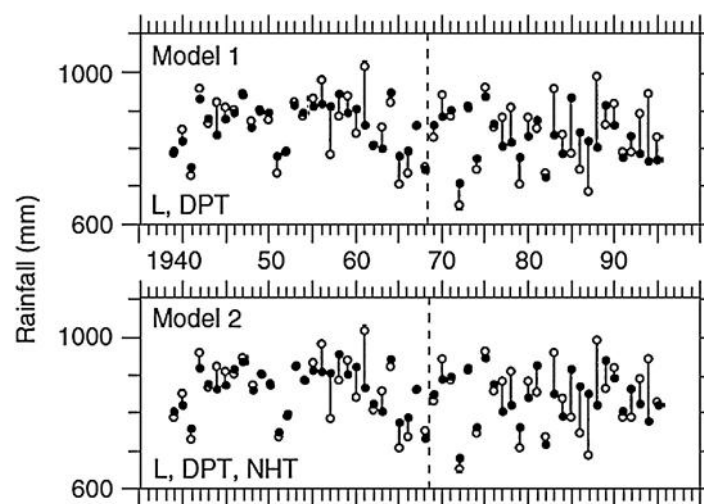


Figure 3 Prediction of all-India summer monsoon rainfall from two models, using as predictors the latitude position of a 500 hPa ridge along 75° E in April (L); April minus January pressure tendency at Darwin (DPT); and an index of January–February Northern Hemisphere temperature (NHT). Solid dots denote regressed values for years up to 1968 and forecast values from 1969 onward, as separated by the vertical broken line. Open circles indicate observed rainfall. Adapted from Hastenrath, S., 1995a. Recent advances in tropical climate prediction. *Journal of Climate* 8, 1519–1532.

the long term. Predictability was found to be high for the 1960s and 1970s, to be followed by drastically reduced performance in the 1980s, as is apparent in [Figure 3](#). Such changes in predictability should be seen in the context of decadal-scale evolutions in the general circulation setting, but the direct causalities await clarification. Numerical model applications to the prediction of Indian monsoon rainfall anomalies will have to face the decadal-scale changes in intrinsic predictability as an added challenge.

Eastern Africa

Eastern Africa features two rainy seasons centered on April–May and October–November, and only the latter is strongly (and inversely) related to the SO. On purely statistical grounds, it has been shown that a portion of the variance of boreal autumn rains is predictable from the Tahiti minus Darwin pressure index alone. The general circulation causes of eastern African rainfall anomalies are related to the SO through a combination of cooperative mechanisms, which most effectively function in the boreal autumn rainy season of eastern Africa: (1) Westerly winds along the Indian Ocean Equator are conducive to lower-tropospheric divergence over equatorial East Africa; in the high SO phase these are accelerated, especially in October–November owing to the anomalous eastward pressure gradient. (2) The Equatorial westerlies drive the Eastward Equatorial Jet in the upper hydrosphere, which entails cold water upwelling in the western extremity of the basin, where sea surface temperature (SST) further hydrostatically affects the zonal pressure gradient and thus feeds back into the equatorial westerly winds. (3) In addition, cold-water anomalies in the western Indian Ocean, most pronounced in October–November during the high SO phase, also suppress convection. (4) In the high SO phase, the Indian summer monsoon tends to be strong, leaving behind an anomalously cold western Indian Ocean, which in turn feeds into the mechanisms (1)–(3). These processes are strong and definite in October and November, when the equatorial westerlies correlate at -0.85 with the rains at the East African coast, but reveal themselves little in effective precursors. Relationships for April and May are weak.

Southern Africa

Various groups endeavor to forecast the austral summer rains of Southern Africa from purely statistical methods and more recently also by using numerical modeling. Rains tend to be more abundant during the high and cold SO phase. A better diagnostic understanding of the circulation mechanisms of rainfall anomalies is needed.

Sahel

The climate problems in the Sub-Saharan zone of West Africa are particularly complex because interannual variability is superimposed on drastic decadal-scale changes of rainfall. These share some common mechanisms involving variations of the meridional SST gradient in the tropical Atlantic Ocean and latitudinal displacements of the Intertropical Convergence Zone (ITCZ), as well as SST changes in the western Indian

Ocean, but the reasons for the persistence of the present drought conditions are not yet understood. Thus, it seems essential to separate year-to-year variations from trendlike developments, a challenge in terms of both general circulation mechanisms and the proper statistical treatment.

Real-time forecasts are published regularly. The Hadley Center of the UK Meteorological Office contributes forecasts from both empirical and numerical modeling approaches. The empirical method entails EOFs of SST, where the inter-hemispheric SST gradient in the tropical Atlantic is most important. Correlation between forecast and observed rainfall is around 0.6, mostly related to persistence. The numerical modeling has global SST as input, and correlations of predicted versus observed rainfall are between 0.3 and 0.6. The National Oceanic and Atmospheric Administration's (NOAA) Climate Prediction Center in the United States reports results from an empirical method using CCA on global SST and gridded rainfall data for Africa. Most important is the inter-hemispheric difference in SST anomalies, particularly in the Atlantic, and ENSO is not the dominating aspect of the relationship; correlation between predicted and observed rainfall is around 0.3. Thus, there are indications for a moderate predictability of Sahel rainfall anomalies. Given the peculiar evolution of the Sahel climate, closer attention to the contributions from year to year compared to trendlike developments seems desirable.

Northeast Brazil

Northern Northeast Brazil has its rainy season narrowly concentrated around March–April, when the near-equatorial trough reaches its southernmost position in the course of the annual cycle. Rainfall anomalies have a severe human impact, their general circulation mechanisms are definite and well understood, and they offer themselves as a prime target of opportunity for climate prediction. During drought years, the southward SST gradient in the tropical Atlantic is enhanced, the near-equatorial low-pressure trough and embedded wind confluence and ITCZ are displaced northward, and the North Atlantic trade winds are reduced while the cross-equatorial flow from the Southern Hemisphere is accelerated. Warm-water anomalies in the equatorial Pacific also tend to be associated with dry conditions in Brazil's Nordeste.

General circulation diagnostics combined with statistical techniques form the basis for effective prediction methods. In the work at the University of Wisconsin, the predictand is an index of the March–June rainfall in the northern Nordeste, constructed from a network of quality-controlled rain gauge stations with continuous record. Effective predictions can be made from observations through January. The predictors are the pre-season rainfall in the Nordeste itself, indices of the fields of the meridional wind component and of SST in the tropical Atlantic, and less importantly, an index of equatorial Pacific SST. This information serves as input to SMR, LDA, and neural networking. An approach by the UK Meteorological Office has as input information SST in the tropical Atlantic and equatorial Pacific. The diagnostic research at the University of Wisconsin also served as basis for numerical modeling experiments by other groups. [Figure 4](#) illustrates the performance of a model using SMR. [Figure 5](#) presents a comparison of performance by

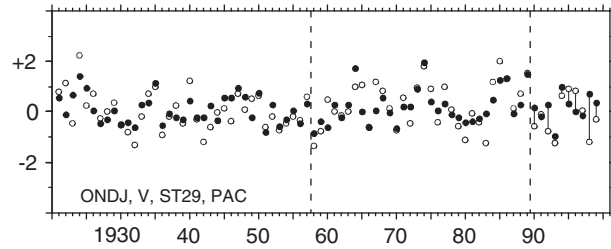


Figure 4 Prediction of an index of March–June rainfall in northern Northeast Brazil from stepwise multiple regression, using as predictors October–January Nordeste rainfall (ONDJ), index of January meridional wind component over the tropical Atlantic (V), index of January SST field in the tropical Atlantic (ST29), and SST anomaly in equatorial Pacific (PAC). Solid dots denote regressed values for years up to 1957 and forecast values from 1958 onward, as separated by the vertical broken line. Open circles indicate observed MAMJ values.

(a) empirical prediction and (b) numerical modeling over the 32 year common period 1968–1999. Verification of forecast performance shows for the numerical modeling larger errors and it captures 30% of the variance as compared to the 59% by the empirical method.

North Atlantic Storms

Since the mid-1980s, seasonal forecasts have been issued for North Atlantic hurricane activity with remarkable success. Input information includes the stratospheric zonal wind component, El Niño developments in the Pacific, pressure over the Caribbean, and rainfall over West Africa.

El Niño and Southern Oscillation

Following the considerable understanding of the underlying atmospheric–hydrospheric mechanism offered by extensive empirical–diagnostic investigations, El Niño has been an early target of efforts in extended-range forecasting. It was the prediction of El Niño to which numerical modeling was first applied. Methods developed from a numerical ocean model driven by recently observed surface winds, to a coupled ocean–atmosphere model using the observed history of the Pacific wind field during the preceding years, to a two-tiered approach in which a model for predicting tropical Pacific SST is used in tandem with an atmospheric GCM. It may be anticipated that even though the upper-air flow patterns seem realistically depicted, the prediction of regional rainfall anomalies remains a challenge.

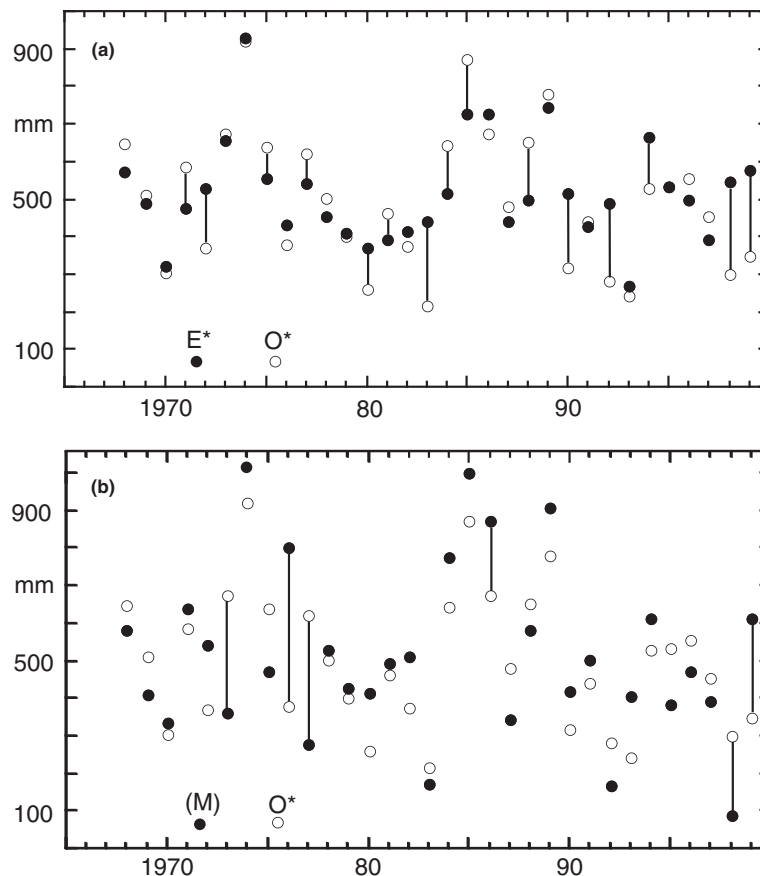


Figure 5 Time series plots of March–June precipitation indices in millimeters. (a) Empirical prediction E* dots, and gridded observations O* open circles; (b) predictions from numerical modeling M dots, with vertical line in lower left indicating correction for negative bias, and gridded observations O* open circles.

Purely statistical techniques have also proven effective for the prediction of El Niño, including CCA of SST and wind fields, SSA of the Tahiti minus Darwin index, and linear inverse modeling of SST.

To the extent that warm water anomalies in extended areas of the equatorial Pacific feed back into the atmospheric circulation, the prediction of El Niño is of interest beyond the Pacific domain proper. It must be noted, however, that the SO contributes substantially to the interannual climate variability only in limited regions of the tropics. With a view to implications for the global atmospheric circulation, at the Climate Prediction Center of NOAA observed surface and subsurface thermal conditions are input to an ocean model and an anomaly-coupling scheme is used for interaction with the atmospheric model. The Pacific El Niño phenomenon is clearly related to the SO, and this is commonly described by the Tahiti minus Darwin pressure index. Signals of the SO are pervasive, but only in limited domains of the tropics does this contribute substantially to the regional climate variability.

Australia

Lying as it does near a dipole of the SO, it is not surprising that the greater Australian region has its climatic variability strongly tied to the phases of this large-scale pressure seesaw. In the high SO phase, surface waters to the north of Australia and in the Indonesian seas tend to be anomalously warm, with low pressure and relatively abundant rainfall. In accordance with the prevailing SST anomalies, tropical storm activity also tends to be enhanced. Such processes in the large-scale circulation setting are being used to advantage for real-time forecasting of rainfall, tropical cyclone activity, and crop yields.

North America

Seasonal forecasts of the spatial patterns of temperature and precipitation anomalies over the United States are regularly issued by the Climate Prediction Center of NOAA, based on statistical methods and numerical modeling. The forecast system consists of tools that forecast the tropical Pacific SST and tools that forecast the US surface temperature and precipitation. Forecasts of tropical Pacific SST are produced from CCA of patterns of global sea level pressure and tropical Pacific SST, and from coupled ocean–atmosphere models run with full coupling in the tropical Pacific region. These statistical and numerical modeling results are then combined into a forecast of tropical Pacific SST. This along with the SST values observed elsewhere in the global ocean serve as boundary conditions to force an atmospheric GCM. Note that this two-tiered numerical modeling system is in principle applicable anywhere over the globe, depending on any skill that could be derived from it, but it is used operationally only for US seasonal forecasts. In addition to this two-tiered numerical modeling system, statistical techniques are also applied, namely optimal climate normals for the region and CCA of the global SST, northern hemispheric 700 hPa patterns, and US temperature and precipitation. A statistical consolidation

of the diverse tools is made to aid the forecasters in objectively combining the information. In the course of the annual cycle, skill is highest at the end of winter and summer, and poorest in between. In this context, lead times are less relevant, so forecasts are issued for the coming year; at certain seasons, skill is low even with the shortest lead times; and, at other seasons, whatever little is known is known well in advance. In a similar vein, experiences with India and Northeast Brazil have shown predictability a couple of months ahead of the season, with no improvement at shorter lead times.

Europe

Attempts at seasonal forecasting of temperature and precipitation by the UK Meteorological Office continue, based on statistical methods and numerical modeling. Wintertime SST anomalies in the North Atlantic serve as inputs to linear regression to predict summer temperature for central England. An extension of this work is devoted to the prediction of summer temperature over Europe. Associations with ENSO are found to be weak.

A major collaborative European effort involving 11 institutions in six countries, Prediction of Climate Variations on Seasonal to Interannual Time-Scales (PROVOST), aims at exploring the potential of numerical modeling for seasonal forecasting for the continent. Tentative results have been reported from the contribution of the UK Meteorological Office. ENSO has not furnished skill for the prediction of the winter or spring conditions. Interest also focuses on the possibility of predicting when a forecast will be skillful. Work continues.

Caveats

It is tempting to believe that the tropical climate prediction problem can be reduced to ENSO and that numerical modeling is the obviously superior tool in the long run. These widely held beliefs are briefly addressed in this section.

Regarding ENSO, it should be realized that interannual variability in the regional climate is directly due to anomalous behavior of the quasipermanent circulation systems in the region; to the extent that the SO is associated with variability in the regional circulation, it may show a correlation with regional climatic conditions. [Figure 6](#) offers some illustrations pertaining merely to diagnostics rather than prognosis. Thus, the October–November rains at the coast of equatorial East Africa are correlated at better than 0.8 with the concurrent surface westerlies over the equatorial Indian Ocean, as compared to correlations of only about 0.6 between the SO and the rains or winds. In a similar vein, the summer rainfall in the West African Sahel has correlations of around 0.6 with indices of the circulation but of only 0.3 with the SO. The rains of Brazil's Nordeste are correlated at about 0.7 with the regional circulation, whereas the SO has correlations of only about 0.3 with either the rain or the circulation.

The development of both empirical and numerical modeling methods for operational climate prediction has been pioneered by two large governmental institutions: the Hadley Center of the UK Meteorological Office and the Climate Prediction Center of

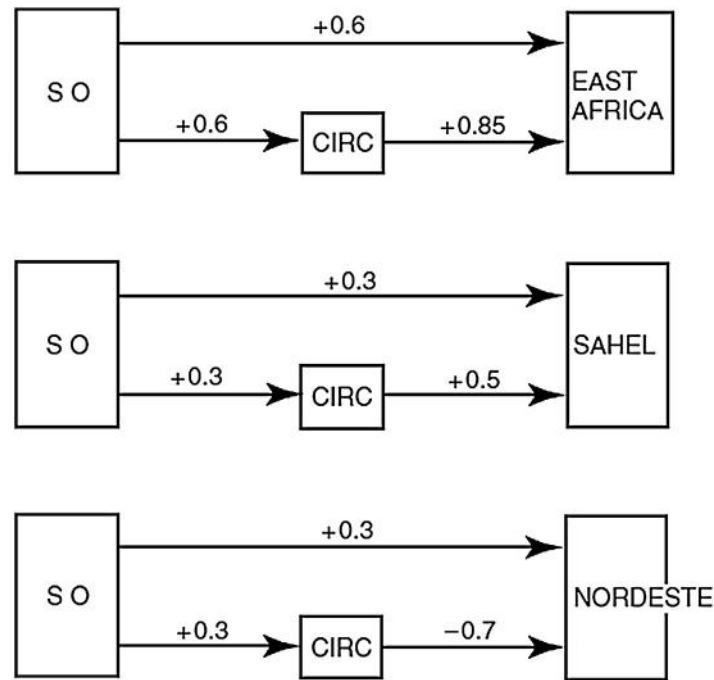


Figure 6 Schemes of long-term mean diagnostic relationships, showing coefficients of concurrent correlation between indices of the Southern Oscillation (SO), the regional circulation (CIRC), and rainfall at the coast of eastern Africa (October–November), in the West African Sahel (July–August), and in Brazil’s Nordeste (March–April).

NOAA. A group from the latter institution has critically examined the potential of the two approaches in review papers half a decade apart. In the earlier review, they found that the two approaches delivered roughly equally skillful forecasts. Five years later, they confirmed this assessment and speculated about the possibility that the numerical models will never be able to outperform empirical models. Such perception is shared by other authors in later publications. Concerning expectations from downscaling in numerical modeling appreciation is in order for precision versus accuracy, an issue meriting further exploration.

In this context, Nicholls’s cautioning in 1999 against cognitive illusions in climate prediction is pertinent. With reference to a recent experience, he noted: “The general view among scientists of the accuracy of forecasts of the El Niño of 1997 appears to illustrate hindsight bias. None of the climate forecast systems predicted anything more than slight warming. Yet the forecasting of the 1997 El Niño with large models is now regarded as a stunning success. Some of the model predictions for the 1997/98 event were very poor. It seems these misfortunes are being ignored when the stunning success of the El Niño model forecasts is assessed. People seek confirmatory evidence and avoid the search for disconfirming evidence”.

Outlook

There has been encouraging progress in the seasonal prediction of interannual climate variability. While forecasting methods are being developed for ever more target regions, the impression remains that climate anomalies may be highly predictable for only limited areas of the tropics where the bulk of the

rainfall is prevalingly derived from a single well-organized quasipermanent circulation system (such as the ITCZ), especially at the extremes of the planetary-scale annual cycle. Fortunately, however, these regions tend naturally to coincide with areas that are most prone to climatic anomalies and where the human impact is most severe. For other vast domains, the prospects of seasonal prediction may be more remote. The extent to which moderate intrinsic predictability may be practically useful in agricultural and economic planning is an issue that remains to be explored in continuing discourse between the forecasting community and the potential user community. In this context, climate monitoring should be cultivated in tandem with climate prediction.

It seems desirable to pursue in parallel the general circulation-based empirical approach and the numerical modeling, because this combination should be mutually fruitful and offer insight into the atmosphere–ocean mechanisms involved. It should be realized that, although the Southern Oscillation signal is pervasive, climate anomalies in many regions are not prevalingly related to the SO. Decadal-scale changes in predictability are poorly understood and are an impediment to prediction regardless of technique. A broad-based effort is needed, combining empirical and modeling approaches and coupling diagnostics with prognosis. Documentation of the method, verification of performance on an independent data set, and evaluation of real-time forecasts are always imperative.

See also: **Aerosols:** Role in Radiative Transfer. **Arctic and Antarctic:** Antarctic Climate; Arctic Climate. **Basic Atmospheric Structure and Concepts:** Wind Chill. **Climate and Climate**

Change: Carbon Dioxide; Climate Variability: Decadal to Centennial Variability; Climate Variability: North Atlantic and Arctic Oscillation; Climate Variability: Seasonal and Interannual Variability; Energy Balance Climate Models; Overview. **Data Assimilation and Predictability:** Predictability and Chaos. **General Circulation of the Atmosphere:** Energy Cycle; Overview. **Global Change:** Climate Record: Surface Temperature Trends. **Numerical Models:** Methods. **Oceanographic Topics:** General Processes. **Tropical Cyclones and Hurricanes:** Hurricanes: Observation; Overview and Theory. **Tropical Meteorology and Climate:** El Niño and the Southern Oscillation: Theory; Walker Circulation. **Weather Forecasting:** Seasonal and Interannual Weather Prediction.

Further Reading

- Anderson, J., Van den Dool, H., Barnston, A., et al., 1999. Present-day capabilities of numerical and statistical models for atmospheric extratropical seasonal simulation and prediction. *Bulletin of the American Meteorological Society* 80, 1349–1361.
- Barnston, A.G., Van den Dool, H., Zebiak, S.E., et al., 1994. Long-lead seasonal forecasts – where do we stand? *Bulletin of the American Meteorological Society* 75, 2097–2114.
- Barnston, A.G., Leetmaa, A., Kousky, V.E., et al., 1999. NCEP forecasts of the El Niño of 1997–98 and its US impacts. *Bulletin of the American Meteorological Society* 80, 1829–1852.
- Carson, D.J., 1998. Seasonal forecasting. *Quarterly Journal of the Royal Meteorological Society* 124, 1–26.
- Cavalcanti, I.F.A., Goddard, L., Kirtman, B., 2006. The future of seasonal prediction in the Americas. *VAMOS Newsletter* 3, 3–7.
- COLA, 1998–99. COLA Experimental Long-Lead Forecast Bulletin, vols. 7–8.
- Druyan, L.M., Fukaleza, M., Loneragan, P., 2002. Dynamic downscaling of seasonal climate predictions over Brazil. *Journal of Climate* 15, 3411–3426.
- Hastenrath, S., 1985. *Climate and Circulation of the Tropics*. Reidel, Dordrecht.
- Hastenrath, S., 1986. On climate prediction in the tropics. *Bulletin of the American Meteorological Society* 67, 692–702.
- Hastenrath, S., 1990. Tropical climate prediction: a progress report 1985–90. *Bulletin of the American Meteorological Society* 71, 819–825.
- Hastenrath, S., 1995a. Recent advances in tropical climate prediction. *Journal of Climate* 8, 1519–1532.
- Hastenrath, S., 1995b. *Climate Dynamics of the Tropics*. Kluwer, Dordrecht.
- Hastenrath, S., Sun, L., Moura, A.D., 2009. Climate prediction for Brazil's Nordeste by empirical and numerical modeling methods. *International Journal of Climatology* 29, 921–926.
- Latif, M., Anderson, D., Barnett, T., et al., 1998. A Review of the Predictability and Prediction of ENSO. *JGR-Oceans*, 14375–14393.
- Lorenz, E.N., 2007. Foreword to Van den Dool, H., *Empirical methods in short-term climate prediction*. Oxford University Press, Oxford, New York.
- Nicholls, N., 1999. Cognitive illusions, heuristics, and climate prediction. *Bulletin of the American Meteorological Society* 80, 1365–1397.
- NOAA-CPC, 1992–97. NOAA-CPC Experimental Long-Lead Forecasting Bulletin, vols. 1–6.
- Palmer, T.N., Anderson, D.L.T., 1994. The prospect for seasonal forecasting – a review paper. *Quarterly Journal of the Royal Meteorological Society* 120, 755–793.
- Peixoto, J.P., Oort, A.H., 1992. *Physics of Climate*. American Institute of Physics, New York.

Climate Variability: Decadal to Centennial Variability

DG Martinson, Columbia University, Palisades, NY, USA

© 2003 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 1, pp 418–424, © 2003, Elsevier Ltd.

Introduction

In the final decade of the twentieth century, recovery of high-resolution paleoclimate records of natural climate variability (see **Paleoclimatology**: Ice Cores; Varves) improved dramatically our perception of the long-term behavior of the Earth's climate. Immediately apparent from these records was the fact that the once implicit notion that the modern-day climate system (following termination of the last ice age) was relatively stable, was no longer tenable. This notion quickly yielded to one in which the Earth's climate is continually changing over all time scales. Presumably it will continue to change, with or without human-induced (anthropogenic) influences.

Climate, as measured by the averaged value of any characteristic of weather, such as temperature or precipitation, can show considerable differences in how it varies through time and across the globe. For example, it might undergo a smooth or abrupt transition from one quasi-stable state to another; it may vary cyclically, not unlike the familiar daily or annual cycles, but with cycles lasting tens or hundreds of years or longer; and it may vary through changes in extreme states (e.g., colder winters), or in the magnitude or degree of fluctuations (e.g., more storms, or larger differences from one year to the next). When such variations occur over, or persist for, tens to hundreds of years, we classify them as decadal-to-centennial variability. For convenience, variability over these time scales is referred to as 'dec-cen' climate variability.

From a practical standpoint, dec-cen climate variability involves climate change that occurs once in a while over the course of a human lifetime (e.g., the 'real' winters your grandfather remembers), or in general it occurs slowly relative to our year-to-year experience of climate. This is distinctly different from shorter-term change that occurs rapidly, and often apparently dramatically within a year or several years (see **Climate and Climate Change**: Climate Variability: Seasonal and Interannual Variability). The difference extends well beyond our perception of the change. It also has considerable implications regarding how the change or variations influence society and how we study them.

Dec-Cen Variability and Society

From a societal perspective, the time scale over which climate variability manifests itself is important from several perspectives. Short-term variability, such as that related to El Niño events (see **Tropical Meteorology and Climate**: El Niño and the Southern Oscillation: Observation; El Niño and the Southern Oscillation: Theory), influences everything from agriculture to recreation. But if such short-term change persists for dec-cen time scales, the implications are considerably different. For example, in the 1990s the Midwestern United States was twice hit with devastating floods (1993 and 1997). As rare and

extreme short-term climate events, society must deal with their impact through disaster relief and flexible adaptive measures. But if they are a consequence of changed climate conditions during the latter part of the twentieth century and represent a regular characteristic of the new climate state, then such flooding must be dealt with through policy decisions and investments in infrastructure.

The latter requires considerable foresight and the best possible information. This is particularly important for dec-cen variability since the magnitude of climate change is often proportional to the length of time over which it operates. Changes on dec-cen time scales can involve potential shifts in agriculture belts, in drought/flooding frequency, magnitude and extent, significant rises in sea level, fundamental adjustments in energy usage, etc. It can also modify the larger background climate state that influences our ability to predict shorter-term climate events, such as El Niño. An ability to recognize or anticipate such change in order to minimize the negative impacts and optimize the positive ones often requires advance action and decisions. Consequently, an ability to forecast such change, recognize its signs, or understand its potential is a fundamental goal of modern climate studies.

Unfortunately, the potentially huge impacts (positive and negative) of dec-cen variations are typically, though not always, realized only slowly with time. Such slow, often imperceptible, change disguises the potential long-term implications of the change, while tending to undermine the immediacy of the problem and our resolve to address it. So too does the knowledge that any experienced change may be part of a longer cycle that will be returning to its previous (acceptable) state in due time. Regardless, climate will change and infrastructure/policy decisions will ultimately be made either in response to the change or in anticipation based on imperfect information. Thus, understanding dec-cen climate change in order to provide sound information regarding the potential or likelihood of change is fundamental to our long-term, social and individual, well-being.

The Study of Dec-Cen Variability

The study of dec-cen climate variability is new, and is at a distinct disadvantage relative to that of shorter-term climate variability. In fact, the differences are significant enough that the paradigm by which climate studies, particularly prediction, have heretofore relied on must be changed to study dec-cen climate variability. Because dec-cen variability involves slow change, long data records are required for model calibrations, predictive skills (the ultimate test of understanding), and observational analyses. That is, records must contain enough occurrences (realizations) of the phenomena being examined to give a reasonable

statistical foundation for analysis or model–data comparison. For dec–cen variability, only the bare minimum of such data sets currently exist.

It will take decades into the future to obtain long and comprehensive enough modern records for dec–cen studies. The historical record from modern instruments does not extend far enough back in time and is typically too sparse in most locations. Invaluable advances are being made in our ability to interpolate these historical data into an internally and dynamically consistent, comprehensive ‘reanalysis’ data set through the use of models. The paleoclimate records, our greatest hope, are still limited in their spatial distribution, and sometimes accuracy and precision, though they are improving rapidly (*see Paleoclimatology: Ice Cores; Varves*).

Contrast this to the remarkably successful paradigm used to advance short-term climate prediction. In that case, numerical models or statistical methods for prediction can be immediately calibrated against the past few decades of observations. They can then be tested against an upcoming year, and soon thereafter modified according to their success or failure. This allows a very effective and rapid advancement in short-term prediction.

For prediction of dec–cen climate variability, this paradigm is clearly impractical except for cases where the change is realized at a fast enough rate to allow some testing in the near distant future (this is the paradigm being used to predict greenhouse warming; *see Global Change: Biospheric Impacts and Feedbacks; Climate Record: Surface Temperature Trends; Upper Atmospheric Change. Ozone Depletion and Related Topics: Long Term Ozone Changes*). Consequently, we are forced into a different paradigm, dependent upon the collection of comprehensive and widespread paleoclimate data sets and the new model reanalysis products. We are equally dependent upon sustained acquisition of observations that will eventually provide that comprehensive data set necessary for quick hindcast calibration and evaluation of models for future generations.

Furthermore, because dec–cen variability involves such long time scales, its study is faced with a number of additional difficulties: (1) Even the relatively high-order physics, those that describe and control processes that do not impose any significant influence over short time scales, have enough time to introduce systematic biases or feedbacks into the system which must be accounted for over long time scales. In models, it is typically necessary that all of the physics of a given order (i.e., relative level of importance) be included since they often work to mitigate or enhance the influence of one another. (2) With longer time scales, changes can be communicated over farther spatial distances and have time to interact with different components of the climate system. For example, predictions of El Niño have been quite successful without taking into consideration the current state of the polar sea ice fields, the level of atmospheric CO₂, or the state of vegetation on land. However, if one wishes to predict how the frequency, intensity, or other characteristics of El Niños may vary over decade-to-century time scales, it is possible that any and all of these may impart some influence that ultimately impacts the tropical Pacific and its evolution.

The implications of these are tremendous. Numerical models examining dec–cen climate variability must therefore include some treatment of detailed physical processes not required in shorter-term climate models. These detailed processes typically operate on the smallest spatial scales, and may have their largest influence in regions quite remote from the region of interest (for example, some aspects of the ocean circulation originate in the subpolar regions as a result of small-scale local interactions, and from there transport heat and salt to the remainder of the world’s oceans). This requires that the models either break the Earth into a great many very small grid cells (*see Numerical Models: Methods*), or include parameterizations of how larger areas may respond in average to the small (‘subgrid’) scale processes that cannot be explicitly resolved. The former allows for better treatment of the physics but at tremendous computational burden (making a computer simulation of the model extremely slow and expensive); the latter relieves some of the computational burden but at the cost of possibly overlooking some of the additional details that may be important. In addition to these detailed processes, the models must contain other parts of the climate system not required in the short-term climate models, such as the biosphere, cryosphere, and atmospheric composition, or more aspects of the ocean, such as the very slow deep ocean circulation. Not only do the additional physics, global scale, and finer model resolution add computational burden, but the very nature of the problem being decade-to-century time scale variability requires that such burdened models run for inordinately long simulation periods as well. Consequently, model simulations are slow, many of the physics missing or uncertain, and the models are so complex that relatively few exist. This limits our ability to assess dec–cen climate variability by multiple experiments under a variety of conditions with many different models involving different parameterizations (though this is still done to the extent possible, but it is greatly hindered by these extreme demands).

Finally, the dec–cen problem is also faced with making predictions based on changes in greenhouse gases (the composition of the atmosphere) that are a function of highly uncertain future emission scenarios. Therefore, even if we succeed in understanding natural dec–cen climate variability, future predictions must be compromised by including uncertain estimates of how the atmospheric composition will change. Here we are forced to depend on multiple forecasts using a variety of different emission scenarios, but again the large computational burden of such comprehensive models limits the practical number of experiments that can be run and examined. Note that it is sometimes taken for granted that the expression climate ‘change’ represents variations in climate due to anthropogenic reasons, where climate ‘variability’ refers to natural variability. In this article, both change and variability are used interchangeably, with the differentiation between natural and anthropogenic change explicitly stated when required.

Modes of Dec–Cen Variability

Climate variability on decade-to-century time scales has manifested itself in historic times through a number of fairly

well-known climate events (see the [Further Reading](#) section). Some of the more notable ones include the prolonged drought of the Great Plains of the United States responsible for the dust bowl of the 1930s, and the crippling drought of the Sahel in northern Africa during the latter decades of the twentieth century, which killed over half a million people in the mid-1970s. The global warming of the twentieth century or the enhanced warming started in the mid-1970s. Changes in the Earth's ozone layer have led to increases in the level of ultra-violet radiation at the Earth's surface in high southern and, more recently, northern latitudes. Dec-cen changes in cloud cover also have led to increased surface radiation in Australia, North America, India, and Europe throughout the twentieth century. The number of major hurricanes varies on dec-cen time scales, as do the number of Nor' Easters ravaging the north-east coast of North America. Sea level has been rising throughout the last century (~ 20 cm) and changes in ecosystems directly related to dec-cen climate have been dramatic in fisheries: the North Atlantic cod and eastern North Pacific salmon. Paleoclimate evidence suggests that dec-cen climate variability is also responsible for the fall of civilizations (e.g., the Classic Maya) and mass migrations of societies (e.g., the Nordestinos of Brazil).

While the dramatic nature of these events makes them rather conspicuous examples of dec-cen variability, much of our focus is directed toward a broader view of dec-cen variability. Specifically, observations made during the last several decades suggest that climate variability over large expanses of the Earth seems to organize itself into patterns that preserve a general shape in space, but whose amplitude may change in time. Such coherent structure is referred to as a mode of variability or, to admit a slightly broader class of structures, they are more generally referred to as climate patterns or spatial-temporal patterns.

Our study of such patterns is relatively new and, because of the lack of historical global data (though, here reanalysis data have proven invaluable), we do not have a comprehensive inventory of the global patterns, nor do we understand their mechanisms, couplings, longevity, or full implications for climate prediction. However, we are encouraged by the study of the most thoroughly investigated pattern: the El Niño Southern Oscillation (ENSO) pattern. El Niño and La Niña are extreme states of the ENSO pattern (see **Tropical Meteorology and Climate: Monsoon: ENSO–Monsoon Interactions**). The ENSO phenomenon is characterized by a pattern of tropical Pacific sea surface temperature (SST) relative to the mean SST. Studies of it revealed that the state of the pattern (e.g., periods when the eastern tropical Pacific SST was abnormally high) was related to regional climate in various regions around the world. Further investigation revealed that the pattern was predictable according to some simple laws involving the interaction of the atmosphere and ocean (over short time scales). This led to considerable insights regarding the nature of climate, the coupling between climate components (in this case, the ocean and atmosphere), scales of influence and, most importantly, our first successful climate predictions (over seasonal to interannual time scales). We are hopeful that additional predictions for other regions of the Earth may be realized in analogous fashion for dec-cen varying patterns.

These other patterns are not as well documented or studied, but indeed they do appear to be related to regional climate. Some are also related to the frequency of hurricanes, Nor' Easters, crop yields, and fisheries. The covariation of two patterns dominated global temperature variations since the mid-1970s. Others display regional or global teleconnections, and they may serve to focus different climate forcings and processes into single coherent responses. Because of these attributes and covarying relationships, it is hoped that their further study may ultimately yield benefits similar to those obtained through the study of ENSO. Patterns also provide an obvious means for breaking the complex climate system down into a finite set of manageable, and hopefully predictable, components.

Most modes are defined by statistical classifications of the observed variability in surface temperature, sea-level pressure, or other quantities. The precise definition may vary according to the statistical methodology employed to define them (see **Statistical Methods: Data Analysis: Empirical Orthogonal Functions and Singular Vectors**). Statistical patterns may ultimately prove to be related to physical laws or to the distribution of land and ocean, of mountains, etc. They may also be artifacts of nature, whereby they are not stable over long periods of time, or they may be statistical artifacts.

The two most important patterns that show dec-cen variability are the North Atlantic Oscillation (NAO; see **Climate and Climate Change: Climate Variability: Seasonal and Interannual Variability**) and the Pacific–North American Teleconnection (PNA; 0400). In addition to these, there is a pattern in the tropical Atlantic (referred to as tropical Atlantic SST variability), in which SST often shows anomalous warmth (referred to as a 'warm pool') in the tropical North Atlantic and a complementary cool pool in the tropical South Atlantic, or vice versa. These seem to vary coherently over decadal time scales, though they vary independently on shorter time scales in these regions. These low-frequency SST phenomena show concurrent anomalies in the rainfall over Brazil and northern Africa. It has also been suggested that the decadal changes in the SST in the subtropical North Atlantic may be responsible for changes in the distribution and intensity of hurricanes in that region.

Likewise there is a decadal ENSO-like pattern, where low-frequency covarying changes in the tropical Pacific atmosphere and ocean strongly resemble the pattern of the interannual ENSO phenomenon, including teleconnected anomalies in the midlatitude atmosphere and ocean of the North Pacific. These decadal ENSO-like anomalies are also teleconnected throughout the tropics, with large concurrent changes in tropical Atlantic and Indian Ocean SST, in addition to the North Pacific.

This anomaly pattern has shown an extended 'warm' phase throughout the last few decades of the twentieth century, which preceded a significant reduction in the alpine glaciers throughout the tropics. The frequency of precipitation, stream flow, and snowpack in the north-west and south-west of North America are also well correlated with this time series describing the decadal ENSO-like climate phenomenon variability.

A number of regional atmospheric patterns have been analyzed such as the North Pacific Oscillation (NPO), West

Pacific Oscillation (WPO), West Atlantic Pattern, and Pacific Decadal Oscillation (PDO). It is not clear how these are related to the PNA or each other (if at all). A completely different kind of pattern, involving sea ice, has been found in the Southern Ocean, and is called the Antarctic Circumpolar Wave (ACW). This pattern is characterized by deviations in the Antarctic sea ice extent from monthly climatological averages, though it is also apparent in surface wind, SST, and sea-level pressure anomalies near the winter ice edge. It is also highly coherent with temporal variations in ENSO and the Indian Ocean monsoons. Other atmospheric patterns have been identified in the Southern Hemisphere, though the data are typically too sparse in time and space to allow more detailed analyses of these.

In addition to the above, there are structures which may, or may not, be considered climate patterns, though they are often related to the other patterns or presented in a similar manner. For example, the Asian monsoon, though predominantly a seasonal signal, is strongly correlated to ENSO and shows decadal variability as indexed by precipitation and wind speeds over India. The global thermohaline circulation (*see Oceanographic Topics: Thermohaline Circulation*) has been tied to distinct changes in the ocean surface conditions and NAO in the North Atlantic Ocean.

Mechanisms of Dec-Cen Variability

The mechanisms responsible for dec-cen climate variability are conveniently separated into those that arise as a consequence of changes in the external forcing of the system and those that arise due to internal variability within the system (independent of changes in the forcing). The external forcing on the Earth's climate consists of solar radiation impinging on the Earth's atmosphere, aerosols (particles or liquid suspended in air) from volcanic eruptions, and the chemical composition of the atmosphere (controlled by natural and anthropogenic sources/sinks of greenhouse gases). The ice ages and the more recent Little Ice Age have been attributed to changes in the intensity of incoming solar radiation. Such changes are typically very small (of the order of a couple percent for the ice age changes associated with changes in the Earth's orbital geometry; and even smaller for the Little Ice Age associated with changes in sunspot activity). Unfortunately, while the observed climate variations of the ice ages and Little Ice Age are consistent with the variations in solar activity, we still do not understand how such minuscule changes can drive such significant responses.

Clearly changes in climate associated with changes in the external forcing require that we can predict the changes in the forcing. Therefore, climate predictions and mechanisms of change associated with changes in external forcing require studies that work under given scenarios of change.

The most fundamental theory regarding internal dec-cen climate variability (i.e., independent of changes in external forcing) was presented in an elegant theory by Hasselmann, whereby the day-to-day weather, representing high-frequency climatic noise, works to drive a slow component of the climate system such as the ocean. Relative to the atmosphere, the ocean has an enormous heat capacity and mixes slowly. These characteristics regulate the speed with which it can

respond to changes in the forcing (thus, it is considered a slow component relative to the atmosphere which is a fast component). The 'muted' response of the ocean serves to integrate the high-frequency atmospheric forcing resulting in an ocean that also varies but over considerably slower time scales. This theory, in combination with the influence of land-sea contrasts and distribution of mountain belts, allows for spatially varying patterns of dec-cen variability. It represents the most basic mechanism of dec-cen variability (effectively, our 'null hypothesis' in the absence of more complex mechanisms).

Other possible internal mechanisms include the interactions between the slow variations of the slow components (e.g., the ocean and the cryosphere), and the coupling of system components that individually may not show slow variability, but together can. A considerable amount of attention has been given to mechanisms of decadal ENSO variability whereby tropical SST anomalies are quickly propagated through the atmosphere to the extratropics and midlatitudes where they introduce local anomalies to the ocean. The ocean retains the anomaly while slowly transporting it back to the tropics (via surface currents, subsurface currents, etc.) where it moderates the tropical SST causing further anomalies and continuation of the cycle. Considerable progress has been made toward identifying potential mechanisms, though more work is still required. Presently, these hypotheses help to focus model experiments and observational studies.

Future Directions and Needs

As stated, the study of decadal-to-centennial climate variability is in its infancy. New discoveries are being made each year, and at the time of this writing we are clearly on the steep slope of the learning curve. The task is onerous given the considerable demands on the models for including all components of the climate system (atmosphere, ocean, cryosphere, biosphere, and land surface), resolving processes operating at the smallest spatial scales to the largest, and having to integrate (run) the models for decades to centuries of model years to realize a single simulation. More powerful computers and their broader availability to climate researchers will aid considerably in this respect, as will close collaboration between scientists of multiple disciplines, and between observationalists, theorists and modelers.

We are faced with changing atmospheric concentrations of radiatively active gases and we need to obtain more accurate records of the actual emission rates to help constrain models simulating past conditions. We have limited observations of dec-cen climate variability, and those from the twentieth century may already be contaminated by anthropogenic climate change (masking the signal of natural variability that is required in order to ultimately recognize an anthropogenic change from the natural variability background). There are additional demands on the data quality required to efficiently study dec-cen variability. This reflects the fact that dec-cen change proceeds at such a small pace on a year-to-year basis, and that any such change is easily lost within the diurnal and seasonal cycle as well as standard interannual variability, all of which are large relative to the annual dec-cen change. Thus, we need

high-precision instruments to pick up dec-cen change as early as possible.

At present, there is no long-term climate observing system for dec-cen variability in place. Consequently, if we are to provide an observational basis from which future generations will be able to more reliably diagnose their model (hindcast) predictions, and analyze comprehensive records of dec-cen variability, we must begin a systematic collection of key variables now. This requires close coordination between research and operational groups.

Despite these impediments, we are encouraged by the fact that climate patterns may ultimately allow us to predict some aspects of dec-cen climate variability. The problem is of considerable importance and will yield invaluable insights regarding the nature and sensitivities of our planet's climate system.

See also: **Climate and Climate Change:** Climate Variability: North Atlantic and Arctic Oscillation; Climate Variability: Seasonal and Interannual Variability. **General Circulation of the Atmosphere:** Teleconnections. **Numerical Models:** Methods. **Oceanographic Topics:** Thermohaline Circulation. **Paleoclimatology:** Ice Cores. **Statistical Methods:** Data Analysis: Empirical Orthogonal Functions and Singular Vectors. **Tropical Meteorology and Climate:** El Niño and the Southern Oscillation: Observation.

Further Reading

- Bradley, R.S., Jones, P.D. (Eds.), 1992. *Climate since AD 1500*. Bradley, Routledge, London.
- Hasselmann, K., 1976. Stochastic climate models. 1. Theory. *Tellus* 28, 473–485.
- Kalnay, E., Kanamitsu, M., Kistler, R., et al., 1996. The NCEP/NCAR 40-year reanalysis project. *Bulletin of the American Meteorological Society* 77, 437–471.
- Mann, M.E., Bradley, R.S., Hughes, M.K., 1998. Global-scale temperature patterns and climate forcing over the past six centuries. *Nature* 392, 779–787.
- National Research Council, 1998. In: Martinson, D.G., et al. (Eds.), *Decade-to-Century-Scale Climate Variability and Change: A Science Strategy*. National Academy Press, Washington, DC.
- National Research Council, 1998. In: Moore, B., et al. (Eds.), *Overview of Global Environmental Change: Research Pathways for the Next Decade*. National Academy Press, Washington, DC.
- National Research Council, 1998. In: Karl, T., et al. (Eds.), *Capacity of US Climate Modeling*. National Academy Press, Washington, DC.
- National Research Council, 1998. In: Dutton, J., et al. (Eds.), *The Atmospheric Sciences Entering the Twenty-First Century*. National Academy Press, Washington, DC.
- National Research Council, 1999. In: Moore, B., et al. (Eds.), *Global Environmental Change: Research Pathways for the Next Decade*. National Academy Press, Washington, DC.
- National Research Council, 1999. In: Karl, T., et al. (Eds.), *Adequacy of Climate Observing Systems*. National Academy Press, Washington, DC.

Climate Variability: Nonlinear and Random Effects

M Ghil, Ecole Normale Supérieure, Paris, France; and University of California, Los Angeles, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

Nonlinear and random effects are pervasive in the atmospheric, oceanic, and climate sciences. This article gives a unified treatment of such effects from the point of view of the theory of dynamical systems and their bifurcations. Energy balance models are used to illustrate multiple equilibria, while multidecadal oscillations in the thermohaline circulation illustrate the transition from steady states to periodic behavior. Random effects are introduced in the setting of random dynamical systems, which permit a unified treatment of both nonlinearity and stochasticity. This treatment is applied to a stochastically perturbed version of the classical Lorenz convection model.

Introduction

The global climate system is composed of a number of subsystems—atmosphere, biosphere, cryosphere, hydrosphere, and lithosphere—each of which has distinct characteristic times, from days and weeks to centuries and millennia. Each subsystem, moreover, has its own internal variability, all other things being constant, over a fairly broad range of timescales. These ranges overlap between one subsystem and another. The interactions between the subsystems thus give rise to climate variability on all timescales.

We outline here the rudiments of the way in which *dynamical systems theory* is starting to provide an understanding of this vast range of variability. Such an understanding proceeds through the study of successively more complex patterns of behavior. These spatiotemporal patterns are studied within narrower ranges of timescales, such as *intraseasonal*, *interannual*, *interdecadal*, and *multimillennial*; each of these frequency bands is covered in a separate article of this Encyclopedia. The main results of dynamical systems theory that have demonstrated their importance for the study of climate variability involve *bifurcation theory* and the *ergodic theory* of dynamical systems. Since the first edition of this encyclopedia, the theory of *random dynamical systems* has made substantial contributions as well, and these are now accounted for here, too.

In the next section, we describe the climate system's overall balance between incoming solar radiation, dominated by shortwaves, and outgoing terrestrial radiation, dominated by long waves. This balance is consistent with the existence of *multiple equilibria* of surface temperatures. Such multiple equilibria are also present for other balances of climatic actions and reactions. Thus, on the intraseasonal timescale, the thermal driving of the midlatitude westerly winds is countered by surface friction and mountain drag. Multiple equilibria typically arise from *saddle-node bifurcations* of the governing equations. Transitions from one equilibrium to another may result from small and random pushes, a typical case of minute causes having large effects in the long term.

In the following section, we sketch the ocean's overturning circulation between cold regions, where water is heavier and sinks, and warm regions, where it is lighter and rises. The effect of temperature on the water masses' density and, hence, motion is in competition with the effect of salinity: density increases, through evaporation and brine formation, compete

further with decreases in salinity and, hence, density through precipitation and river runoff. These competing effects can also give rise to two distinct equilibria.

In the present-day oceans, a *thermohaline* circulation (THC) prevails, in which the temperature effects dominate. In the remote past, about 50 My ago, a *halothermal* circulation may have obtained, with salinity effects dominating. In a simplified mathematical setting, these two equilibria arise by a *pitchfork bifurcation* that breaks the problem's mirror symmetry about the equator. On shorter timescales, of decades-to-millennia, *oscillations* of intensity and spatial pattern in the THC seem to be the dominant mode of variability. We show how interdecadal oscillations in the ocean's circulation arise by *Hopf bifurcation*.

In the final section, we address the way that faster processes, modeled as random effects, can interact with the slower, nonlinear ones. The combined treatment of the nonlinear and stochastic processes can reveal amazingly fine structure in the climate system's behavior, but also—and rather surprisingly—add robustness and predictability to the results. Concluding remarks follow.

Energy Balance Models (EBMs) and the Modeling Hierarchy

The methods of dynamical systems theory have been applied first to simple models of atmospheric and oceanic flows, starting anithybout 50 years ago. More powerful computers now allow their application to fairly realistic and detailed models of the atmosphere, ocean, and the coupled atmosphere–ocean system. The present section starts, therefore, by presenting such a hierarchy of models.

This presentation is interwoven with that of the *successive bifurcations* that lead from simple to more complex solution behavior for each climate model. Useful tools for comparing model behavior across the hierarchy and with observations are provided by ergodic theory. Among these, advanced methods for the analysis and prediction of uni- and multivariate time series play an important role.

Radiation Balance and EBMs

At present, the most highly developed hierarchy is for atmospheric models. Atmospheric models were originally

developed for weather simulation and prediction on the timescale of hours to days. Currently, they serve—in a stand-alone mode or coupled to oceanic and other models—to address climate variability on all timescales.

The first rung of the modeling hierarchy for the atmosphere is formed by zero-dimensional (0D) models, where the number of dimensions, from zero to three, refers to the number of independent space variables used to describe the model domain, i.e., to physical space dimensions. Such 0D models essentially attempt to follow the evolution of global surface air temperature $\bar{T}$ as a result of changes in global radiative balance:

$$c \frac{d\bar{T}}{dt} = R_i - R_o \quad [1a]$$

$$R_i = \mu Q_o \{1 - \alpha(\bar{T})\}, \quad R_o = \sigma m(\bar{T}) \bar{T}^4 \quad [1b, c]$$

Here R_i and R_o are, respectively, incoming solar radiation and outgoing terrestrial radiation. The heat capacity c is that of the global atmosphere, plus that of the global ocean or some fraction thereof, depending on the timescale of interest: one might only include in c the ocean mixed layer when interested in subannual timescales but the entire ocean when studying paleoclimate. The rate of change of $\bar{T}$ with time t is given by $d\bar{T}/dt$, while Q_o is the solar radiation received at the top of the atmosphere, s is the Stefan–Boltzmann constant, and m is an insolation parameter, equal to unity for present-day conditions. To have a closed, self-consistent model, the planetary reflectivity or albedo α and grayness factor m have to be expressed as functions of $\bar{T}$; $m = 1$ for a perfectly black body and $0 < m < 1$ for a gray body like planet Earth.

There are two kinds of one-dimensional (1D) atmospheric models, for which the single spatial variable is latitude or height, respectively. The former are so-called EBM, which consider the generalization of the model eqn [1] for the evolution of surface–air temperature $T = T(x, t)$, say,

$$c(x) \frac{\partial T}{\partial t} = R_i - R_o + D \quad [2]$$

Here the terms on the right-hand side can be functions of the meridional coordinate x (latitude, colatitude, or sine of latitude), as well as of time t and temperature T . The horizontal heat flux term D expresses heat exchange between latitude belts; it typically contains first and second partial derivatives of T with respect to x . Hence, the rate of change of local temperature T with respect to time also becomes a partial derivative, $\partial T/\partial t$.

The first striking results of theoretical climate dynamics were obtained in showing that eqn [2] could have two stable steady-state solutions, depending on the value of the insolation parameter μ , cf. eqn [1b]. This multiplicity of stable steady-states, or physically possible ‘climates’ of our planet, can be explained, in its simplest form, in the 0D model eqns. [1a]–[1c].

The simple explanation resides in the fact that—for a fairly broad range of μ values around $\mu = 1.0$ —the curves for R_i and R_o as a function of $\bar{T}$ intersect in three points. One of these corresponds to the present climate (highest $\bar{T}$ value), and another one to an ice-covered planet (lowest $\bar{T}$ value); both of these are stable, while the third one (intermediate $\bar{T}$ value) is unstable. To obtain this result, it suffices to make two assumptions: (1) $\alpha = \alpha(\bar{T})$ is a piecewise linear function of $\bar{T}$, with high albedo at low temperature, due to the presence of

snow and ice, and low albedo at high $\bar{T}$, due to their absence; and (2) $m = m(\bar{T})$ is a smooth, increasing function of $\bar{T}$ that attempts to capture in its simplest form the ‘greenhouse effect’ of trace gases and water vapor.

The *bifurcation diagram* of such a 1D EBM is shown in Figure 1. It displays the model’s mean temperature as a function of the fractional change μ in the insolation $Q = Q(x)$ at the top of the atmosphere. The ‘S’-shaped curve in the figure arises from two back-to-back saddle-node bifurcations. The *normal form* of the first one is

$$\dot{X} = \mu - X^2 \quad [3]$$

Here X stands for a suitably normalized form of $\bar{T}$ and $\dot{X} \equiv dX/dt$ is the rate of change of X , while μ is a parameter that measures the stress on the system, in particular a normalized form of the insolation parameter.

The uppermost branch corresponds to the steady-state solution $X = +\mu^{1/2}$ of eqn [3] and is stable. It matches rather well Earth’s present-day climate for $\mu = 1.0$, more precisely the steady-state solution $T = T(x; \mu)$ of the full 1D EBM (not shown) matches closely the annual mean temperature profile from instrumental data over the last century.

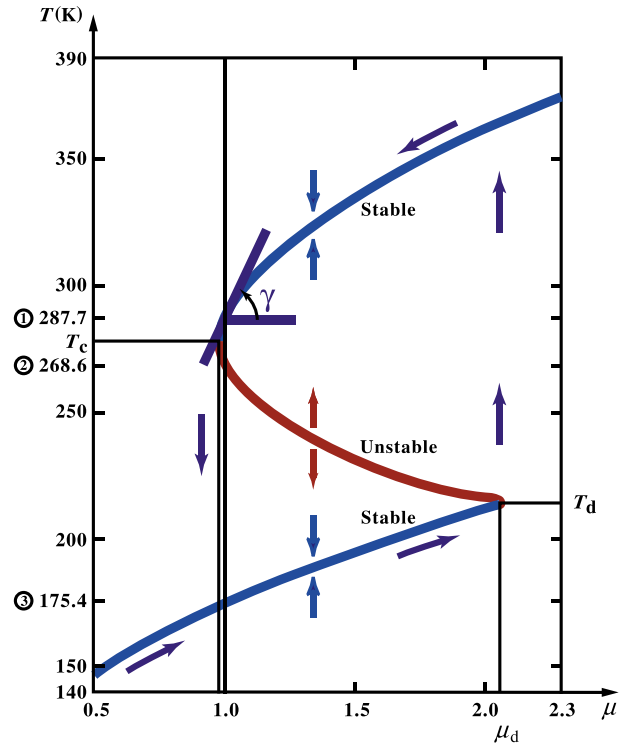


Figure 1 Bifurcation diagram for the solutions of an EBM, showing the annual mean temperature $\bar{T}$ vs fractional change of insolation at the top of the atmosphere μ . The arrows pointing up and down at about $\mu = 1.4$ indicate the stability of the branches: toward a given branch if it is stable and away if it is unstable. The other arrows show the hysteresis cycle that global temperatures would have to undergo for transition from the upper stable branch to the lower one and back. The angle γ gives the measure of the present climate’s sensitivity to changes in insolation. Reproduced from Ghil, M., Childress, S., 1987. *Topics in Geophysical Fluid Dynamics: Atmospheric Dynamics, Dynamo Theory and Climate Dynamics*. New York, NY/Berlin: Springer-Verlag.

The intermediate branch starts out at the left, it corresponds to the second solution, $X = -\mu^{1/2}$, of eqn [3] and it is unstable. This unstable branch blends smoothly into the upper branch of a coordinate-shifted and mirror-reflected versions of eqn [3], say,

$$\dot{X} = \mu - \mu_0 + (X - X_0)^2, \quad \text{where } \mu_0 = 1 \text{ and } X_0 = -\frac{1}{2} \quad [4]$$

This branch, $X = X_0 + (\mu_0 - \mu)^{1/2}$, is also unstable. Finally, the lowermost branch in Figure 1 corresponds to the second steady-state solution of eqn [4], $X = X_0 - (\mu_0 - \mu)^{1/2}$, and it is also stable. This last branch represents an ice-covered planet at the same distance from the Sun as the Earth.

The fact that the upper-left bifurcation point (μ_c, T_c) in Figure 1 is so close to present-day insolation values created great concern in the climate dynamics community in the mid-1970s, when these results were obtained. Indeed, much more detailed computations (see below) confirmed that a reduction of about 2–5% of insolation values would suffice to precipitate the Earth into a ‘deep freeze’. The great distance of the lower-right bifurcation point (μ_d, T_d) from present-day insolation values, on the other hand, suggests that one would have to nearly double atmospheric opacity, say, for the Earth’s climate to jump back to more comfortable temperatures.

Other Atmospheric Processes and Models

The 1D atmospheric models in which the details of radiative equilibrium are investigated with respect to a height coordinate z (geometric height, pressure, etc.) are often called *radiative–convective models*. This name emphasizes the key role that convection plays in vertical heat transfer. While these models preceded historically EBMs as rungs on the modeling hierarchy, it was only recently shown that they, too, could exhibit multiple equilibria. The word (stable) ‘equilibrium’, here and in the rest of this article, refers simply to a (stable) steady state of the model, rather than to a true thermodynamic equilibrium.

Two-dimensional (2D) atmospheric models are also of two kinds, according to the third space coordinate that is not explicitly included. Models that resolve explicitly two horizontal coordinates, on the sphere or on a plane tangent to it, tend to emphasize the study of the dynamics of large-scale atmospheric motions. They often have a single layer or two. Those that resolve explicitly a meridional coordinate and height are essentially combinations of EBMs and radiative–convective models and emphasize therewith the thermodynamic state of the system, rather than its dynamics.

Yet another class of ‘horizontal’ 2D models is the extension of EBMs to resolve zonal, as well as meridional surface features, in particular land–sea contrasts. We shall see in the section on “Bifurcation diagrams for GCMs” how such a 2D EBM is used, when coupled to an oceanic model.

Schneider and Dickinson in 1974, as well as Ghil and Robertson in 2000 have discussed additional types of 1D and 2D atmospheric models. See [Further Reading](#) for references to these and to the types discussed above, along with some of their main applications. Finally, to encompass and resolve the main atmospheric phenomena with respect to all three spatial coordinates, *general circulation models* (GCMs) occupy the pinnacle of the modeling hierarchy.

The dependence of mean zonal temperature on the insolation parameter μ (the normalized ‘solar constant’)—as obtained for 1D EBMs and shown in Figure 1 here—was confirmed, to the extent possible, by using a simplified GCM, coupled to a ‘swamp’ ocean model. More precisely, forward integrations with a GCM cannot confirm the presence of the intermediate, unstable branch. Nor was it possible in the mid-1970s, when this numerical experiment was done, to reach the ‘deep-freeze’ stable branch, because of the GCM’s computational limitations. But the parabolic shape of the upper, present-day like branch near the upper-left bifurcation point in the figure, cf. eqn [3], was well supported by the GCM simulations.

Ghil and Robertson have also described the separate hierarchies that have grown over the last quarter century in modeling the ocean and the coupled ocean–atmosphere system. More recently, an overarching hierarchy of earth system models—that encompass all the subsystems of interest, atmosphere, biosphere, cryosphere, hydrosphere, and lithosphere—has been developing. Eventually, the partial results about each subsystem’s variability, outlined in this section and the next one, will have to be verified from one rung to the next of the earth system modeling hierarchy.

Interdecadal Oscillations in the Oceans’ THC

Theory and Simple Models

Historically, the THC was first among the climate system’s major processes to be studied using a very simple mathematical model. Stommel formulated in 1961 a two-box model and showed that it possessed multiple equilibria.

A sketch of the Atlantic Ocean’s THC and its interactions with the atmosphere and cryosphere on long timescales is shown in Figure 2. These interactions can lead to climate oscillations with multimillennial periods, such as the Heinrich events, and are summarized in the figure’s caption. An equally schematic view of the global THC is provided by the widely known ‘conveyor belt’ diagram. The latter diagram does not commonly include the THC’s interactions with water in both its gaseous and solid phases, which the former does include.

Basically, the THC is due to denser water sinking, lighter water rising, and water mass continuity closing the circuit through near-horizontal flow between the areas of rising and sinking. The effects of temperature and salinity on the ocean water’s density, $\rho = \rho(T, S)$, oppose each other: The density ρ decreases with increasing T and increases with increasing S . It is these two effects that give the THC its name, from the Greek words for T and S . In high latitudes, ρ increases as the water loses heat to the air above and, if sea ice is formed, as the water underneath is enriched in brine. In low latitudes, ρ increases due to evaporation but decreases due to sensible heat flux into the ocean.

For the present climate, the temperature effect is stronger than the salinity effect, and ocean water is observed to sink in certain areas of the high-latitude North Atlantic and Southern Ocean—with very few and limited areas of deepwater formation elsewhere—and to rise everywhere else. Thus, in a *thermohaline* regime, T is more important than and hence comes before S . During some remote geological times, deepwater may

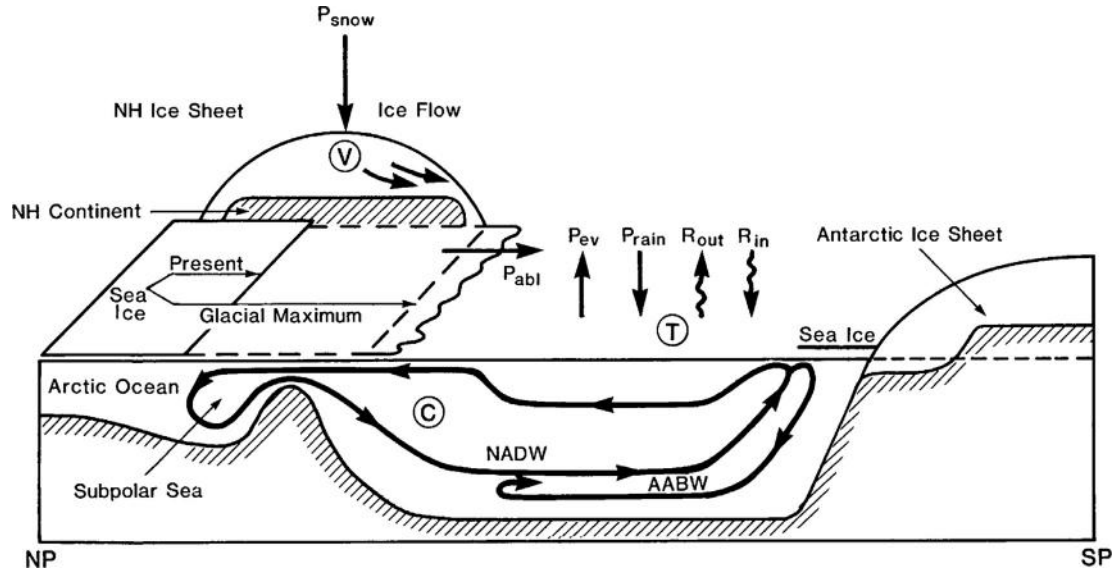


Figure 2 Diagram of an Atlantic meridional cross section from North Pole (NP) to South Pole (SP), showing mechanisms likely to affect the THC on various timescales. Changes in the radiation balance $R_{\text{in}} - R_{\text{out}}$ are due, at least in part, to changes in extent of Northern Hemisphere (NH) snow and ice cover, V , and how they affect the global temperature, T ; the extent of Southern Hemisphere (SH) ice is assumed constant, to a first approximation. The change in hydrologic cycle expressed in the terms $P_{\text{rain}} - P_{\text{evap}}$ for the ocean and $P_{\text{snow}} - P_{\text{abl}}$ for the snow and ice is due to changes in ocean temperature. Deepwater formation in the North Atlantic Subpolar Sea (North Atlantic Deep Water: NADW) is affected by changes in ice volume and extent, and regulates the intensity C of the THC; changes in Antarctic Bottom Water (AABW) formation are neglected in this approximation. Deep-water formation in turn affects the system's temperature, and is also affected by it. Reproduced from Ghil, M., Mullhaupt, A., Pestiaux, P., 1987. Deep water formation and Quaternary glaciations, *Climate Dynamics* 2, 1–10.

have formed in the global ocean near the equator; such an overturning circulation of opposite sign to that prevailing today has been dubbed *halothermal*, *S* before *T*. The quantification of the relative effects of *T* and *S* on the oceanic water masses' buoyancy in high and low latitudes is far from complete, especially for paleocirculations; the association of the latter with salinity effects that exceed the thermal ones is thus rather tentative.

Stommel considered a two-box model, with two pipes connecting the two boxes. He showed that the system of two nonlinear, coupled ordinary differential equations that govern the temperature and salinity differences between the two well-mixed boxes, has two stable steady-state solutions, distinguished by the direction of flow in the upper and lower pipes. Stommel's paper was primarily concerned with distinct local convection regimes, and hence vertical stratifications, in the North Atlantic and the Mediterranean (or Red Sea), say. Today, we mainly think of one box as representing the low latitudes and the other one the high latitudes in the global THC.

The next step in the hierarchical modeling of the THC is that of 2D meridional plane models, in which the temperature and salinity fields are governed by coupled nonlinear partial differential equations with two independent space variables, latitude and depth. Given boundary conditions for such a model that is symmetric about the Equator, as are the equations themselves, one expects a symmetric solution, in which water either sinks near the poles and rises everywhere else (thermohaline) or sinks near the Equator and rises everywhere else (halothermal). These

two symmetric solutions would correspond to the two equilibria of Stommel's box model of 1961.

In fact, symmetry breaking can occur, leading gradually from a symmetric two-cell circulation to an antisymmetric one-cell circulation. In between, all degrees of dominance of one cell over the other are possible. A situation lying somewhere between the two seems to resemble most closely the meridional overturning diagram of the Atlantic Ocean in **Figure 2**.

This symmetry breaking can be described by a *pitchfork bifurcation*:

$$\dot{X} = \mu X - X^3 \quad [5]$$

Here X stands for the amount of asymmetry in the solution, so that $X = 0$ is the symmetric branch, and μ is a parameter that measures the stress on the system, in particular a normalized form of the buoyancy flux at the surface. For $\mu < 0$, the symmetric branch is stable, while for $\mu > 0$, $X = 0$ becomes unstable and the two branches $X = \pm \mu^{1/2}$ inherit its stability.

In the 2D THC problem, the left cell dominates on one branch, while the right cell dominates on the other: for a given value of μ , the two stable steady-state solutions—on the $\{X = +\mu^{1/2}\}$ branch and on the $\{X = -\mu^{1/2}\}$ branch—are mirror images of each other. The idealized THC in **Figure 2**, with the North Atlantic Deep Water (NADW) extending to the Southern Ocean's polar front, corresponds to one of these two branches. In theory, therefore, a mirror image circulation, with the Antarctic Bottom Water (AABW) extending to the North Atlantic's polar front, is equally possible.

Table 1 Thermohaline circulation oscillations

Timescale	Phenomena	Mechanism
Interdecadal	3D, wind-driven + thermohaline circulation	– Gyre advection – Localized surface-density anomalies due to surface coupling
Centennial	Loop-type, Atlantic–Pacific circulation	Conveyor-belt advection of high-latitude density anomalies
Millennial	Relaxation oscillation, with “flushes” and superimposed decadal fluctuations	Bottom-water warming, due to high latitude freshening and its braking effect

Adapted from Ghil, M., 1994. Cryothermodynamics: The chaotic dynamics of paleoclimate. *Physica D* 77: 130–159.

Bifurcation Diagrams for GCMs

F Bryan was the first, in 1986, to document transition from a two-cell to a one-cell circulation in a simplified ocean GCM with idealized, symmetric forcing. Results of coupled ocean–atmosphere GCMs, however, have led to questions about the realism of more than one stable THC equilibrium. The situation with respect to the THC’s pitchfork bifurcation, as illustrated in eqn [5], is thus subtler than it was with respect to [Figure 1](#) for radiative equilibrium. In the previous section, atmospheric GCMs confirmed essentially the EBM results; the results obtained in climbing the rungs of the modeling hierarchy for the THC are still in need of further clarification.

Internal variability of the THC—with smaller and more regular excursions than the huge and totally irregular jumps associated with bistability—was studied intensively in the late 1980s and the 1990s. These studies placed themselves on various rungs of the modeling hierarchy, from box models through 2D models and all the way to ocean GCMs. A summary of the different kinds of oscillatory variability found in the latter appears in [Table 1](#). Such oscillatory behavior seems to match more closely the instrumentally recorded THC variability, as well as the paleoclimatic records for the recent geological past, than bistability.

The (multi)millennial oscillations interact with variability in the surface features and processes, which are shown in [Figure 2](#). Chen and Ghil, in particular, studied some of the interactions between atmospheric processes and the THC. They used a so-called hybrid coupled model, namely a (horizontally) 2D EBM, coupled to a rectangular box version of the North Atlantic rendered by a low-resolution ocean GCM. This hybrid model’s regime diagram is shown in [Figure 3\(a\)](#). A steady state is stable for high values of the coupling parameter λ_{ao} or of the EBM’s diffusion parameter d . Interdecadal oscillations with a period of 40–50 years are self-sustained and stable for low values of these parameters.

The self-sustained THC oscillations in question are characterized by a pair of vortices of opposite sign that grow and decay in quadrature with each other in the ocean’s upper layers. Their centers follow each other anticlockwise through the northwestern quadrant of the model’s rectangular domain. Both the period and the spatiotemporal characteristics of the oscillation are thus rather similar to those seen in a fully coupled GCM with realistic geometry. The transition from a stable equilibrium to a stable limit cycle, via Hopf bifurcation, in this hybrid coupled model, is shown in [Figure 3\(b\)](#).

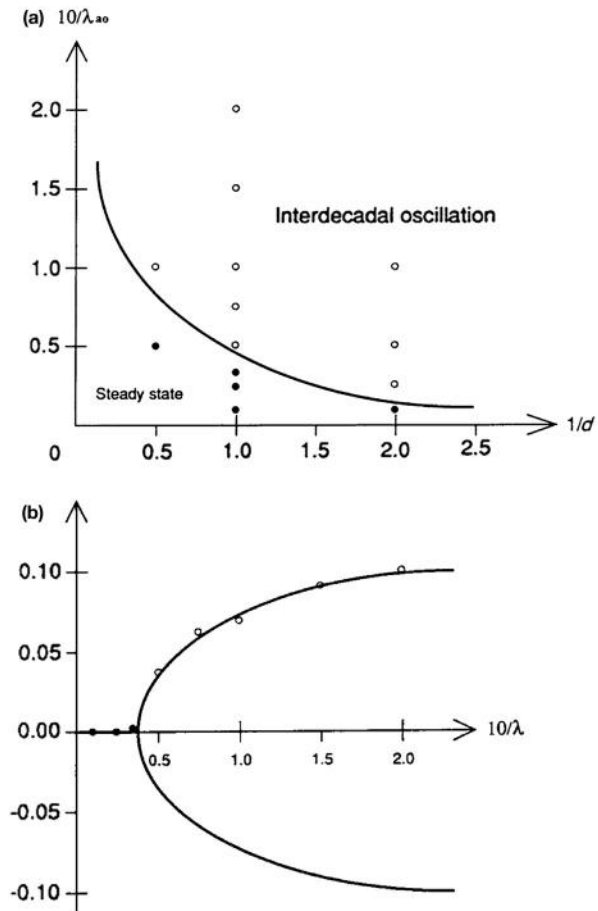


Figure 3 Dependence of THC solutions on two parameters in a hybrid coupled model; the two parameters are the atmosphere–ocean coupling coefficient λ_{ao} and the atmospheric thermal diffusion coefficient d . (a) Schematic regime diagram. The full circles stand for the model’s stable steady states, the open circles for stable limit cycles, and the solid curve is the estimated neutral stability curve between the former and the latter. (b) Hopf bifurcation curve at fixed $d = 1.0$ and varying λ_{ao} ; this curve was obtained by fitting a parabola to the model’s numerical simulation results, shown as full and open circles, for the stationary and the periodic GCM solutions, respectively. Reproduced from Chen, F., Ghil, M., 1996. Interdecadal variability in a hybrid coupled ocean–atmosphere model. *Journal of Physical Oceanography* 26, 1561–1578.

Randomness and Nonlinearity

What to Expect

The geometric and the ergodic theories of dynamical systems represent significant achievements of the twentieth century. The foundations of the stochastic calculus in its second half also led to the birth of a rigorous theory of time-dependent random phenomena. Historically, theoretical developments in climate dynamics have been largely motivated by these two complementary approaches, based on the works of Lorenz and that of Hasselmann, respectively.

It now seems clear that these two approaches complement, rather than exclude each other. Incomplete knowledge of small-, subgrid-scale processes, as well as computational limitations, will always require one to account for these processes in a stochastic way. As a result of sensitive dependence on initial data and on parameters, numerical weather forecasts and climate projections are both expressed these days in probabilistic terms. In addition to the intrinsic challenge of addressing the nonlinearity along with the stochasticity of climatic processes, it is thus more convenient—and becoming more and more necessary—to rely on a model's (or set of models') probability density function (PDF) rather than on its individual, pointwise simulations or predictions.

We summarize here the results on the surprisingly complex statistical structure that characterizes stochastic nonlinear systems. This complex structure does provide meaningful physical information that is not described by the PDF alone; it lives on a *random attractor*, which extends the concept of a *strange attractor* and of its *invariant measures* from deterministic to stochastic dynamics.

What One Finds

On the road to including random effects, one needs to realize first that the climate system—as well as any of its subsystems, and on any timescale—is not closed: it exchanges energy, mass, and momentum with its surroundings, whether other subsystems or the interplanetary space and the solid earth. The typical applications of dynamical systems theory to climate variability so far have only taken into account exchanges that are constant in time, thus keeping the model—whether governed by ordinary, partial, or other differential equations—*autonomous*, i.e., the models had coefficients and forcings that were constant in time.

Succinctly, one can write such an autonomous system as

$$\dot{\mathbf{X}} = \mathbf{f}(\mathbf{X}, \boldsymbol{\mu}) \quad [6]$$

where $\mathbf{X}$ now may stand for any state vector or climate field, while $\mathbf{f}$ is a smooth function of $\mathbf{X}$ and of the vector of parameters $\boldsymbol{\mu}$, but does not depend explicitly on time. The characteristic of being autonomous greatly facilitated the analysis of model solutions' properties. For instance, two distinct trajectories, $\mathbf{X}_1(t)$ and $\mathbf{X}_2(t)$, of a well-behaved, smooth autonomous system cannot pass through the same point in phase space, which helps describe the system's phase portrait. So does the fact that we only need to consider the behavior of solutions $\mathbf{X}(t)$ as we let time t tend to $+\infty$: the resulting sets of points are—possibly multiple—equilibria, periodic solutions, and chaotic sets. In the language of dynamical systems theory, these are called, respectively, *fixed points*, *limit cycles*, and *strange attractors*.

We only know too well, however, that the seasonal cycle plays a key role in climate variability on many timescales, while orbital forcing is crucial on the Quaternary timescales of many millennia, and now anthropogenic forcing is of utmost importance on interdecadal timescales. How can one take into account such time-dependent forcings, and analyze the nonautonomous systems, written succinctly as:

$$\dot{\mathbf{X}} = \mathbf{f}(\mathbf{X}, t; \boldsymbol{\mu}) \quad [7]$$

to which they give rise? In eqn [7], the dependence of $\mathbf{f}$ on t may be periodic, $\mathbf{f}(\mathbf{X}, t + P) = \mathbf{f}(\mathbf{X}, t)$ as in various El Niño–Southern Oscillation (ENSO) models, where the period $P = 12$ months, or monotone, $\mathbf{f}(\mathbf{X}, t + \tau) \geq \mathbf{f}(\mathbf{X}, t)$, as in studying scenarios of anthropogenic climate forcing.

To illustrate the fundamental character of the distinction between an autonomous system like eqn [6] and a nonautonomous one like eqn [7], consider the simple scalar version of these two equations:

$$\dot{X} = -\beta X \quad [8]$$

$$\dot{X} = -\beta X + \gamma t \quad [9]$$

respectively. We assume that both systems are dissipative, i.e., $\beta > 0$, and that the forcing is monotone increasing, $\gamma \geq 0$, as would be the case for anthropogenic forcing in the industrial era. Lorenz in his 1963 paper pointed out the key role of dissipativity in giving rise to strange, but attracting solution behavior, while Ghil and Childress in their 1987 book emphasized its importance and pervasive character in climate dynamics. Clearly, the only attractor for the solutions of eqn [6], given any initial point $X(0) = X_0$, is the fixed point $X = 0$, attained as $t \rightarrow +\infty$.

In the case of eqn [9], though, this *forward-in-time approach* yields blowup as $t \rightarrow +\infty$, for any initial point. To make sense of what happens in the case of time-dependent forcing, one introduces instead the *pullback approach*, in which solutions are allowed to still depend on the time t at which we observe them, but also on a time s from which the solution is started, $X(s) = X_0$; presumably $s \ll t$. With this little change of approach, one can easily verify that

$$|X(s, t; X_0) - A(t)| \rightarrow 0 \text{ as } s \rightarrow -\infty \quad [10a]$$

for all t and X_0 , where

$$A(t) = \gamma(t - 1/\beta)/\beta \quad [10b]$$

One thus obtains, in this pullback sense, the intuitively obvious result that the solutions, if started far enough in the past, all approach the attractor set $A(t)$, which has a linear growth in time, thus following the forcing.

Let us return now to the more general, nonlinear case of eqn [7] and add not only deterministic time dependence $\mathbf{f}(\mathbf{X}, t)$, but also random forcing,

$$d\mathbf{X} = \mathbf{f}(\mathbf{X}, t)dt + \mathbf{g}(\mathbf{X})d\eta \quad [11]$$

where $\eta = \eta(t, \omega)$ represents a Wiener process—while $d\eta$ is commonly referred to as 'white noise'—and ω labels the particular realization of this random process. The case $\mathbf{g}(\mathbf{X}) = \text{const.}$ is the case of *additive noise*, while in the case of $\partial \mathbf{g}(\mathbf{X}) / \partial \mathbf{X} \neq 0$ one speaks of *multiplicative noise*. The distinction

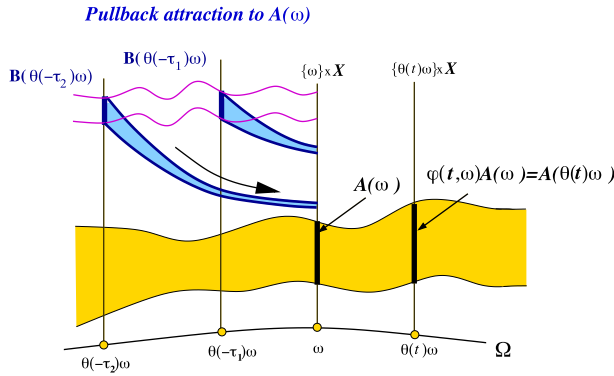


Figure 4 Schematic diagram of a random attractor $A(\omega)$ and of the pullback attraction to it; here ω labels the particular realization of the random process $\theta(t)\omega$ that drives the system. We illustrate the evolution in time t of the random process $\theta(t)\omega$ (light solid black line at the bottom); the random attractor $A(\omega)$ itself (yellow band in the middle) with the ‘snapshots’ $A(\omega) = A(\omega; t = 0)$ and $A(\omega; t)$ (the two vertical sections, heavy solid); and the flow of an arbitrary set B from ‘pullback times’ $t = -\tau_2$ and $t = -\tau_1$ onto the attractor (heavy blue arrows). Reproduced from Ghil, M., Chekroun, M.D., Simonnet, E., 2008. Climate dynamics and fluid mechanics: Natural variability and related uncertainties, invited survey paper for Special Issue on “The Euler Equations: 250 Years On” *Physica D* 237: 2111–2126, doi:10.1016/j.physd.2008.03.036.

between dt and $d\eta$ in eqn [11] is necessary since, roughly speaking and following Einstein’s celebrated 1905 paper on Brownian motion, it is the variance of a Wiener process that is proportional to time and thus $d\eta \propto (dt)^{1/2}$.

In the case of random forcing, the concepts introduced by the simple example of eqns [10a] and [10b] above can be illustrated by the random attractor $A(\omega)$ (yellow band) of Figure 4. In the figure, $d\eta(t, \omega) = \theta(t)\omega$ is the random process that drives the system (light solid black line) and the pullback attraction is depicted by the flow of an arbitrary set B from ‘pullback times’ $t = -\tau_2$ and $t = -\tau_1$ onto the attractor (heavy blue arrows).

More explicitly, one can see in Figure 5 four ‘snapshots’ $\{A_j(\omega) = A(\omega; t = t_j); j = 1, 2, 3, 4\}$ that correspond to the vertical cross sections (heavy solid) in the attractor of Figure 4; a short video, from which these snapshots are taken, is also linked to this article. These snapshots were calculated for the random attractor $A(\omega)$ of a stochastically perturbed Lorenz system, given by

$$\begin{cases} dX = P(Y - X)dt + \sigma_0 X d\eta \\ dY = (rX - Y - XZ)dt + \sigma_0 Y d\eta \\ dZ = (-bZ + XY)dt + \sigma_0 Z d\eta \end{cases} \quad [12]$$

The parameters r , P , and b in eqn [12] have the usual meanings for 2D thermal convection: $r = R/R_c$ is the Rayleigh number R normalized by its critical value R_c at the onset of convection, P is the Prandtl number, and b is a normalized wave number for the most unstable wave at the onset of convection. The noise in this case is multiplicative: its intensity $\sigma_0 = 0.5$ is multiplied in each one of the three coupled, nonlinear equations above by the corresponding variables X , Y , or Z .

Supplementary data related to this article can be found at <http://dx.doi.org/10.1016/B978-0-12-382225-3.00110-9>.

To be precise, what is plotted in Figure 5, and in the associated video at <http://dx.doi.org/10.1016/B978-0-12-382225-3.00110-9>, is the density of the invariant measure $\nu(\omega)$

supported on the random attractor of the stochastically perturbed Lorenz system in eqn [12]. This measure indicates the probability of trajectories winding up in a particular region of phase space and it is very highly concentrated on the attractor, as inferred from the huge range of density values: the color bar in the figure is on a logarithmic scale, and extends over more than 10 orders of magnitude. The situation is thus very different from that expected when studying additive noise; in that case, the noise tends to smear out the fine, Cantor-set-like structure of the deterministic, strange attractor, and the associated PDF is supported on a set of nonzero volume.

It hardly needs saying that additive noise has been studied in climate dynamics much more extensively since it was easier to do so. Studies along these lines were suggested by the simple Brownian motion analogy of ‘weather = water molecules’ and ‘climate = pollen particle’, as proposed by Hasselmann in 1976. Across the hierarchy of climate models discussed in the previous two sections of this article, however, it is clear that small and fast scales of motion do not enter exclusively in an additive manner: they pop up in many, if not all terms of the governing equations, as summarized in eqn [11]. The insights offered, therefore, by Figure 5 and the video are likely to be of interest across the hierarchy of models, all the way up to coupled GCMs and Earth system models.

The invariant measure in Figure 5 exhibits amazing complexity, with fine, very intense filamentation: There is no fuzziness whatsoever in the topological structure of this filamentation, which does evoke the Cantor set foliation of the deterministic attractor. This fine structure strongly suggests that an object of vanishing volume supports this measure, i.e., that the random attractor $A(\omega)$ of eqn [12] has—like the strange attractor of the classical, deterministic version, with $\sigma_0 = 0$ —dimension smaller than 3.

Such complexity, however, should not hide the fact that the theory of random dynamical systems provides robust tools for studying the parameter dependence of a nonlinear, randomly perturbed system’s various ‘metrics’. These metrics can include global quantities, such as mean temperature or total energy, but also much finer functionals of the state of the system, such as regional temperatures or precipitation. In addition, this theory can help improve prediction of future system properties, by relying on a judicious combination of the history of its slow and fast behavior.

Concluding Remarks

A complete theory of climate variability, across the entire range of timescales of interest, is still in the future. We have shown, though, that powerful conceptual and numerical tools exist in order to organize the emerging knowledge so far. The approach described herein relies on applying systematically dynamical systems theory, both deterministic and stochastic, across a hierarchy of models, from the simplest toy models to the most detailed, coupled GCMs. This approach has progressed from its first modest steps, taken almost exactly half a century ago, to the analysis of the behavior of atmospheric, oceanic, and coupled GCMs over

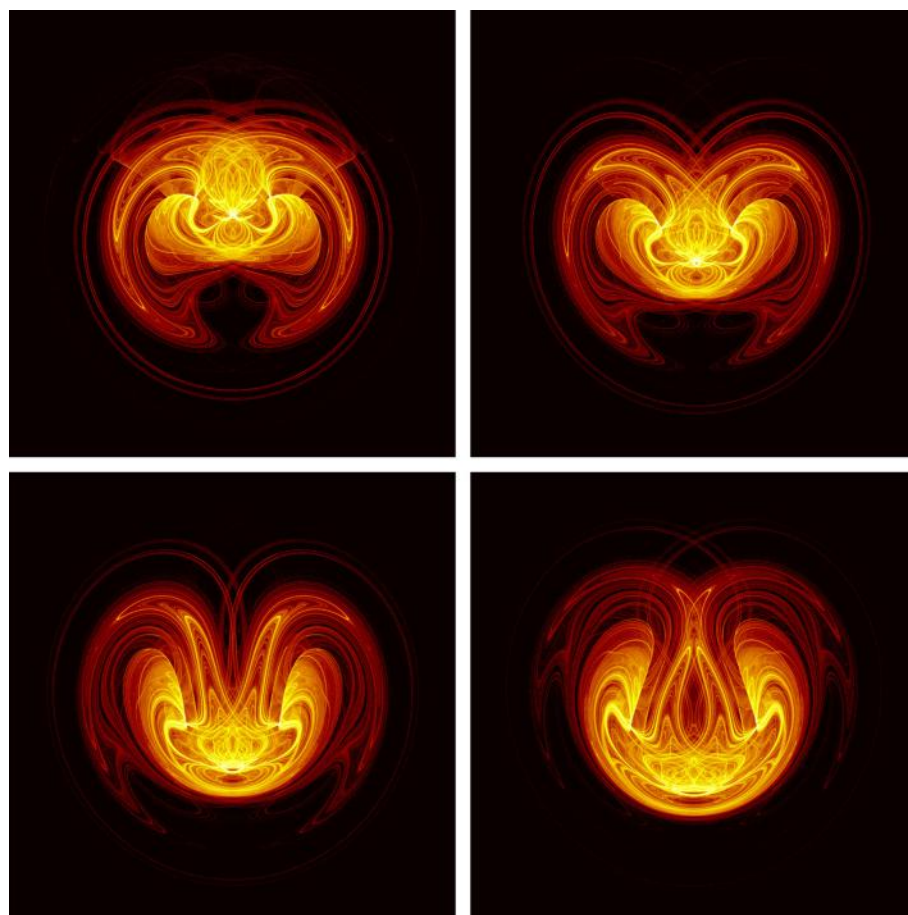


Figure 5 Four snapshots of the stochastically perturbed Lorenz (1963) model's random attractor $A(\omega)$ and the invariant measure $\nu(\omega)$ supported on it. The parameter values are the classical ones—normalized Rayleigh number $r = 28$, Prandtl number $P = 10$, and normalized wave number $b = 8/3$ —while the noise intensity is $\sigma_0 = 0.5$ and the time step is $\delta t = 5 \times 10^{-3}$. The color bar used is on a log-scale and quantifies the probability to end up in a particular region of phase space; shown is a projection of the 3D phase space (X, Y, Z) onto the (X, Z) plane. Notice the complex, interlaced filament structures between highly (yellow) and moderately (red) populated regions. The time interval Δt between two successive snapshots—moving from left to right and top to bottom—is $\Delta t = 0.0875$. Note that the support of the invariant measure $\nu(\omega; t)$ may change quite abruptly, from time t to time $t + \Delta t$; see the related short video at <http://dx.doi.org/10.1016/B978-0-12-382225-3.00110-9>. Weakly populated regions cover an important part of the random attractor and are, in turn, entangled with regions that have near-zero probability (black). Reproduced from Chekroun, M.D., Simonnet, E., Ghil, M., 2011. Stochastic climate dynamics: Random attractors and time-dependent invariant measures. *Physica D* 240 (21), 1685–1700. <http://dx.doi.org/10.1016/j.physd.2011.06.005>.

the last two decades. Especially, interesting strides have been taken over the last decade in studying the interaction of the faster timescales with the slower ones, within a genuinely nonlinear framework.

Cores; Varves. Tropical Meteorology and Climate: El Niño and the Southern Oscillation: Observation; El Niño and the Southern Oscillation: Theory.

See also: **Climate and Climate Change:** Climate Variability: Decadal to Centennial Variability; Climate Variability: North Atlantic and Arctic Oscillation; Climate Variability: Seasonal and Interannual Variability. **General Circulation of the Atmosphere:** Weather Regimes and Multiple Equilibria. **Global Change:** Biospheric Impacts and Feedbacks; Climate Record: Surface Temperature Trends; Sea Level Change; Upper Atmospheric Change. **Oceanographic Topics:** General Processes; Surface/Wind-Driven Circulation; Thermohaline Circulation; Water Types and Water Masses. **Ozone Depletion and Related Topics:** Long-Term Ozone Changes. **Paleoclimatology:** Ice

Further Reading

- Arnold, L., 1998. *Random Dynamical Systems*. Springer-Verlag, New York, NY/Berlin.
- Bryan, F., 1986. High-latitude salinity effects and interhemispheric thermohaline circulations. *Nature* 323, 301–304.
- Chekroun, M.D., Kondrashov, D., Ghil, M., 2011. Predicting stochastic systems by noise sampling, and application to the El Niño–Southern Oscillation. *Proceedings of the National Academy of Sciences of the United States of America* 108 (29), 1766–1771. <http://dx.doi.org/10.1073/pnas.1015753108>.
- Chekroun, M.D., Simonnet, E., Ghil, M., 2011. Stochastic climate dynamics: Random attractors and time-dependent invariant measures. *Physica D* 240 (21), 1685–1700. <http://dx.doi.org/10.1016/j.physd.2011.06.005>.
- Chen, F., Ghil, M., 1996. Interdecadal variability in a hybrid coupled ocean–atmosphere model. *Journal of Physical Oceanography* 26, 1561–1578.

- Dijkstra, H.A., 2005. Nonlinear Physical Oceanography: A Dynamical Systems Approach to the Large-Scale Ocean Circulation and El Niño, second ed. Kluwer Academic Publishers, Dordrecht, The Netherlands/Norwell, MA, pp. 532.
- Eckmann, J.-P., Ruelle, D., 1985. Ergodic theory of chaos and strange attractors. *Reviews of Modern Physics* 57, 617–656 [addendum, *Reviews of Modern Physics*, 57, 1115].
- Einstein, A., 1905. On the movement of small particles suspended in a stationary liquid demanded by the molecular-kinetic theory of heat. *Annalen der Physik (Leipzig)* 17, 549 ff [reprinted in *Investigations on the Theory of the Brownian Movement*, Five Articles by A. Einstein, R. Fürth (Ed.) and A. D. Cowper (transl.), 1956, Dover Publ., New York, NY, pp. 122].
- Ghil, M., 1994. Cryothermodynamics: The chaotic dynamics of paleoclimate. *Physica D* 77, 130–159.
- Ghil, M., Childress, S., 1987. *Topics in Geophysical Fluid Dynamics: Atmospheric Dynamics, Dynamo Theory and Climate Dynamics*. Springer-Verlag, New York, NY/Berlin.
- Ghil, M., Mullhaupt, A., Pestiaux, P., 1987. Deep water formation and quaternary glaciations. *Climate Dynamics* 2, 1–10.
- Ghil, M., Chekroun, M.D., Simonnet, E., 2008. Climate dynamics and fluid mechanics: Natural variability and related uncertainties, invited survey paper for Special Issue on “*The Euler Equations: 250 Years On*,” *Physica D* 237, 2111–2126. <http://dx.doi.org/10.1016/j.physd.2008.03.036>.
- Ghil, M.A., Robertson, W., 2000. Solving problems with GCMs: General circulation models and their role in the climate modeling hierarchy. In: Randall, D. (Ed.), *General Circulation Model Development: Past, Present and Future*. Academic Press, San Diego, CA, pp. 285–325.
- Ghil, M.M., Allen, R.M., Dettinger, D., et al., 2002. Advanced spectral methods for climatic time series. *Reviews of Geophysics* 40 (1), 3.1–3.41. <http://dx.doi.org/10.1029/2000RG000092>.
- Guckenheimer, J., Holmes, P., 1983. *Nonlinear Oscillations, Dynamical Systems and Bifurcations of Vector Fields*. Springer-Verlag, New York, NY/Berlin. 453.
- Hasselmann, K., 1976. Stochastic climate models, Part I: Theory. *Tellus* 28, 473–485.
- Kalnay, E., 2003. *Atmospheric Modeling, Data Assimilation and Predictability*. Cambridge University Press, Cambridge, UK/London, UK. 341.
- Kennett, R.P., Stott, L.D., 1991. Abrupt deep-sea warming, paleoceanographic changes and benthic extinctions at the end of the Palaeocene. *Nature* 353, 225–229.
- Lorenz, E.N., 1963. Deterministic nonperiodic flow. *Journal of Atmospheric Sciences* 20, 130–141.
- Marotzke, J., 2000. Abrupt climate change and thermohaline circulation: Mechanisms and predictability. *Proceedings of the National Academy of Sciences of the United States of America* 97, 1347–1350.
- North, G.R., Cahalan, R.F., Coakley Jr., J.A., 1981. Energy balance climate models. *Reviews of Geophysics and Space Physics* 19, 91–121.
- Ramanathan, V., Coakley, J.A., 1978. Climate modeling through radiative convective models. *Reviews of Geophysics and Space Physics* 16, 465–489.
- Ruelle, D., Takens, F., 1971. On the nature of turbulence. *Communications in Mathematical Physics* 20, 167–192. and 23, 343–344.
- Schneider, S.H., Dickinson, R.E., 1974. Climate modeling. *Reviews of Geophysics and Space Physics* 25, 447–493.
- Solomon, S., et al., 2007. *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the IPCC*. Cambridge University Press, Cambridge, UK.
- Stommel, H., 1961. Thermohaline convection with two stable regimes of flow. *Tellus* 13, 224–230.

Climate Variability: North Atlantic and Arctic Oscillation

JW Hurrell, National Center for Atmospheric Research, Boulder, CO, USA

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

A leading pattern of weather and climate variability over the Northern Hemisphere is the North Atlantic oscillation (NAO). The NAO refers to a redistribution of atmospheric mass between the Arctic and the subtropical Atlantic, and swings from one phase to another to produce large changes in surface air temperature, winds, storminess, and precipitation over the Atlantic as well as the adjacent continents. The NAO also affects the ocean through changes in heat content, gyre circulations, mixed layer depth, salinity, high-latitude deepwater formation, and sea ice cover. A better understanding of how the NAO responds to external forcing, including increasing greenhouse gas concentrations, is thus crucial.

Introduction

Changes in naturally occurring patterns or 'modes' of atmospheric and oceanic variability, such as the El Niño southern oscillation (ENSO) and the North Atlantic oscillation (NAO), orchestrate large variations in weather and climate over much of the globe on interannual and longer timescales. For instance, significant changes in regional temperature and precipitation from one year (or one decade) to the next can be attributed to changes in the phase and amplitude of these two dominant patterns of variability. Moreover, it has been argued that the spatial pattern of the response to anthropogenic forcing may project principally onto such modes of natural climate variability.

Modal variability thus forms a natural subject on which the investigators of climate and climate impact science can collaborate. This collaboration is also required to determine the most relevant effects for society of global change on regional natural and managed ecosystems. This article focuses on the extratropical Northern Hemisphere (NH) and it begins with a basic description of the spatial structure of extratropical climate and climate variability. It includes a discussion of how the NAO is defined and its impacts on surface temperature, precipitation, storms, the underlying ocean, and sea ice. The article concludes with a brief discussion of outstanding issues and future challenges, especially related to the mechanisms that most likely govern NAO variability.

The Spatial Structure of Extratropical Climate and Climate Variability

Climate variability is usually characterized in terms of 'anomalies,' where an anomaly is the difference between the instantaneous state of the climate system and the climatology (the mean state computed over many years representative of the era under consideration). Since the spatial structure of climate variability in the extratropics is strongly seasonally dependent, it is useful to briefly examine the seasonal evolution of the mean state upon which the climate variations are superimposed.

The Mean State and Planetary Waves

Large changes in the mean distribution of sea level pressure (SLP) over the NH are evident from winter (December–

February) to summer (June–August). Perhaps, most noticeable are those changes over the Asian continent related to the development of the Siberian anticyclone during winter and the monsoon cyclone over Southeast Asia during summer (Figure 1). Over the northern oceans, subtropical anticyclones dominate during summer, with the Azores high-pressure system covering nearly all of the North Atlantic. These anticyclones weaken and move equatorward by winter, when the high-latitude Aleutian and Icelandic low-pressure centers predominate.

Because air flows counterclockwise around low pressure and clockwise around high pressure in the NH, westerly flow across the middle latitudes of the Atlantic sector occurs throughout the year (Figure 2). The vigor of the flow is related to the north–south pressure gradient, so the surface winds are strongest during winter when they average near 5 m s^{-1} from the eastern United States across the Atlantic onto northern Europe. These middle latitude westerly winds extend throughout the troposphere and reach their maximum (up to 40 m s^{-1}) at a height of about 12 km. This 'jet stream' roughly coincides with the path of storms (atmospheric disturbances operating on timescales of days) traveling between North America and Europe. Over the subtropical Atlantic the prevailing surface northeasterly trade winds are relatively steady but strongest during boreal summer.

In the middle troposphere ($\sim 5\text{--}6 \text{ km}$), the boreal winter map of the geopotential height field reveals a westward tilt with elevation of the high-latitude surface cyclones and anticyclones (Figure 3). Two low-pressure troughs and two high-pressure ridges are evident: the troughs are located over northeastern Canada and just east of Asia, and the ridges are positioned just west of Europe and over western North America. These strong zonal asymmetries reflect the so-called 'stationary waves' that are forced primarily by the continent–ocean heating contrasts and the presence of the Rocky and Himalayan mountain ranges. In summer, the flow is much weaker and more symmetric, consistent with a much more uniform equator-to-pole distribution of solar radiation.

Although the planetary-scale wave patterns (Figure 3) are geographically anchored, they do change in time either because the heating patterns in the atmosphere vary or because of internal (chaotic) processes. The amplitude and structure of the variability of the seasonal mean 500-hPa geopotential height field (Figure 4) is characterized by a strong longitudinal

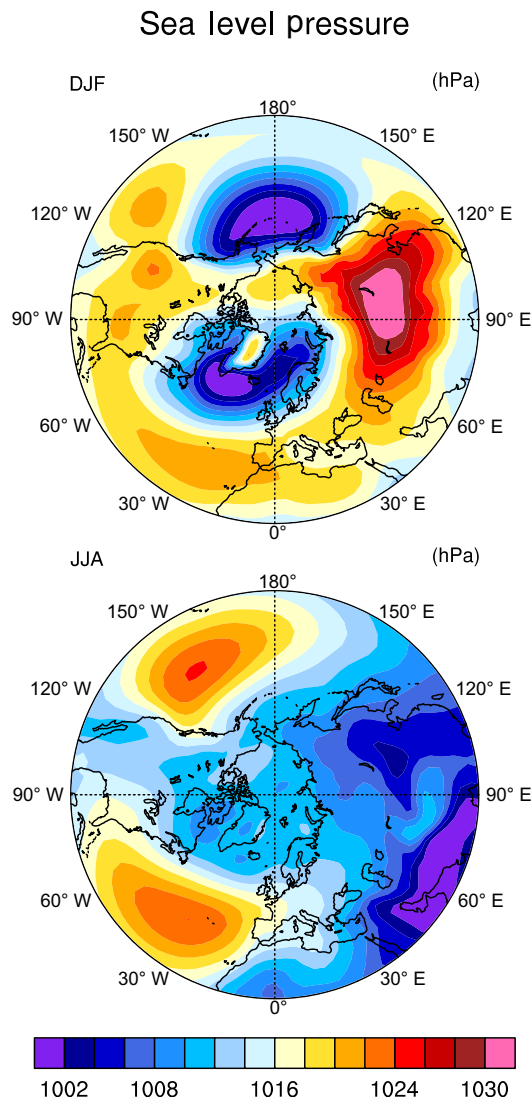


Figure 1 Mean SLP for (top) boreal winter (December–February) and (bottom) boreal summer (June–August). The data come from the NCEP/NCAR reanalysis project over 1958–2006.

dependence with maximum temporal variance over the northern oceans, especially during boreal winter. The frequency dependence of the winter pattern is subtle: maps of the variability of monthly mean data, or data filtered to retain fluctuations within specific frequency bands (e.g., 60–180 days), also exhibit distinct variance maxima at 500 hPa over the Atlantic and Pacific Oceans, although the longitudinal contrasts become increasingly apparent as longer timescales are examined. In comparison, throughout most of the NH, the standard deviations of boreal summer 500 hPa heights are only about half as large as those of the wintertime means (Figure 4).

Teleconnections: the PNA and the NAO

A consequence of the transient behavior of the atmospheric planetary-scale waves is that anomalies in climate on seasonal timescales typically occur over large geographic regions. Some

regions may be cooler or perhaps drier than average, while at the same time thousands of kilometers away, warmer and wetter conditions prevail. These simultaneous variations in climate, often of opposite sign, over distant parts of the globe are commonly referred to as ‘teleconnections’ in the meteorological literature. Though their precise nature and shape vary to some extent according to the statistical methodology and the data set employed in the analysis, consistent regional characteristics that identify the most conspicuous patterns emerge.

Arguably the most prominent teleconnections over the NH are the NAO and the Pacific–North American (PNA) patterns. Both patterns are of largest amplitude during the boreal winter months, and their mid-tropospheric spatial structure is illustrated most simply through one-point correlation maps (Figure 5). These maps are constructed by correlating the 500-hPa height time series at a ‘reference gridpoint’ with the corresponding time series at all gridpoints. The PNA teleconnection pattern has four centers of action. Over the North Pacific Ocean, geopotential height fluctuations near the Aleutian Islands vary out-of-phase with those to the south, forming a seesaw pivoted along the mean position of the Pacific subtropical jet stream (Figure 2). Over North America, variations in geopotential height over western Canada and the northwestern United States are negatively correlated with those over the southeastern United States, but are positively correlated with the subtropical Pacific center. The significance of the locations and the respective phases of the four centers of the PNA is their relation to the mean atmospheric circulation (Figure 3). Variations in the PNA pattern represent variations in the waviness of the atmospheric flow in the western half-hemisphere and thus the changes in the north–south migration of the large-scale Pacific and North American air masses and their associated weather.

On interannual timescales, atmospheric circulation anomalies over the North Pacific, including the PNA, are linked to changes in tropical Pacific sea surface temperatures (SSTs) associated with the ENSO phenomenon. This association reflects mainly the dynamical teleconnection to higher latitudes forced by deep convection in the tropics. The PNA pattern is sometimes viewed, then, as the extratropical arm of ENSO. Significant variability of the PNA occurs even in the absence of ENSO, however, indicating that the PNA is an ‘internal’ mode of atmospheric variability.

Similarly, the NAO does not owe its existence to coupled ocean–atmosphere–land interactions, as is evident from climate model experiments that do not include SST, sea ice, or land surface variability. In contrast to the wavelike appearance of the PNA, the NAO is primarily a north–south dipole characterized by simultaneous out-of-phase height anomalies between temperate and high latitudes over the Atlantic sector (Figure 5). Both the NAO and PNA are also reflected in the spatial patterns of the two leading empirically determined orthogonal functions (EOFs) of NH boreal winter 500-hPa height (not shown). The NAO also dominates the leading EOF of the NH SLP field. Analyzing SLP allows for the longer term behavior of the NAO to be evaluated, as a long series of SLP charts over the NH begin in 1899, in contrast to 500-hPa height charts that are confined to after 1947. Moreover, even longer instrumental records of SLP variations are available, especially from European stations. Thus, in the following, the spatial

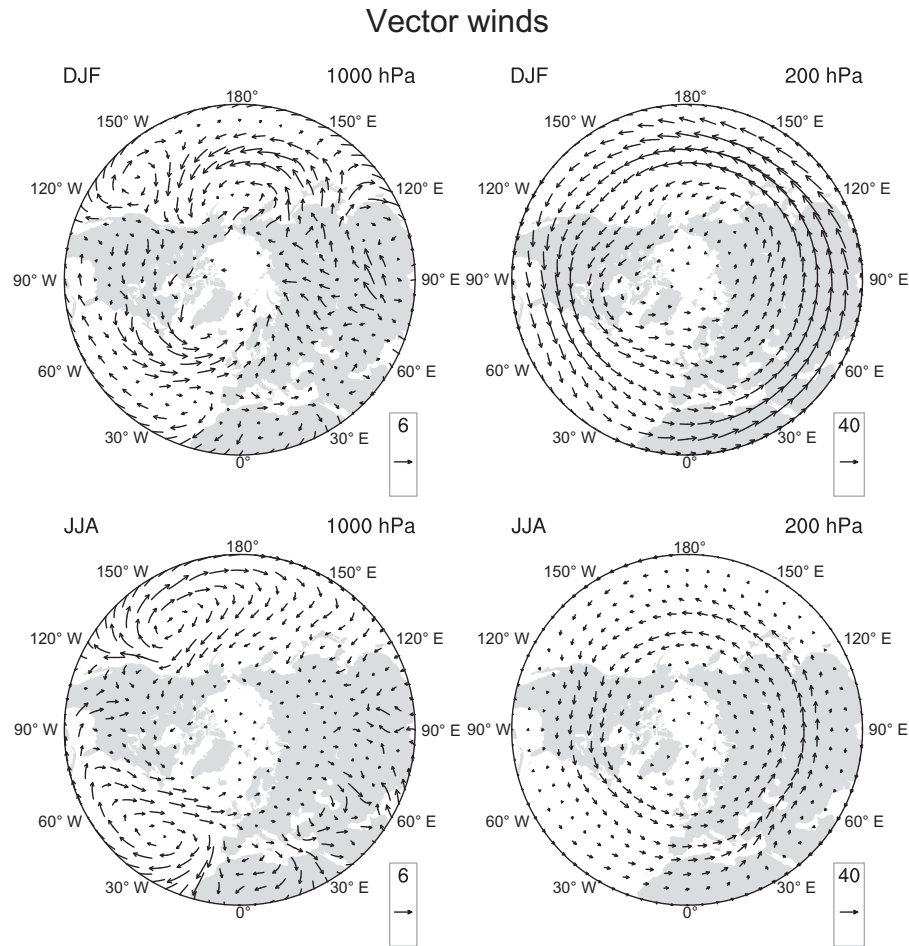


Figure 2 Mean vector winds for (top) boreal winter (December–February) and (bottom) boreal summer (June–August) for (left) 1000 hPa and (right) 200 hPa over 1958–2006. The scaling vectors in meters per second are indicated in the boxes.

structure and time evolution of the NAO are examined in more detail from SLP records.

The Spatial Signature of the NAO

There is no single way to ‘define’ the NAO. One approach is through conceptually simple one-point correlation maps (Figure 5), identifying the NAO by regions of maximum negative correlation over the North Atlantic. Another technique is EOF (or principal component (PC)) analysis. In this approach, the NAO is identified from the eigenvectors of the cross-covariance (or cross-correlation) matrix, computed from the time variations of the gridpoint values of SLP or some other climate variable. The eigenvectors, each constrained to be spatially and temporally orthogonal to the others, are then scaled according to the amount of total data variance they explain. This linear approach assumes preferred atmospheric circulation states come in pairs, in which anomalies of opposite polarity have the same spatial structure. In contrast, climate anomalies can also be identified by cluster analysis techniques, which search for recurrent patterns of a specific amplitude and sign. Clustering algorithms identify weather or climate ‘regimes,’ which correspond to peaks in the probability density

function of the climate phase space. In the following, the spatial patterns of the NAO are compared as estimated from both traditional EOF and clustering techniques.

EOF Analysis of North Atlantic SLP

The leading eigenvectors of the cross-covariance matrix calculated from seasonal (3-month average) SLP anomalies in the North Atlantic sector (20–70° N; 90° W–40° E) are illustrated in Figure 6. The patterns are very similar if based on the cross-correlation matrix (not shown).

The largest amplitude anomalies in SLP occur during the boreal winter months; however, throughout the year the leading pattern of variability is characterized by a surface pressure dipole, and thus may be viewed as the NAO, although the spatial pattern is not stationary. Since the eigenvectors are, by definition, structured to explain maximum variance, it is expected that the ‘center of actions’ of the leading EOFs will coincide with the regions of strongest variability, and the movement of those regions through the annual cycle is reflected in Figure 6.

The NAO is the only teleconnection pattern evident throughout the year in the NH. During the winter season (December–February), it accounts for more than one-third of

500-hPa Geopotential height

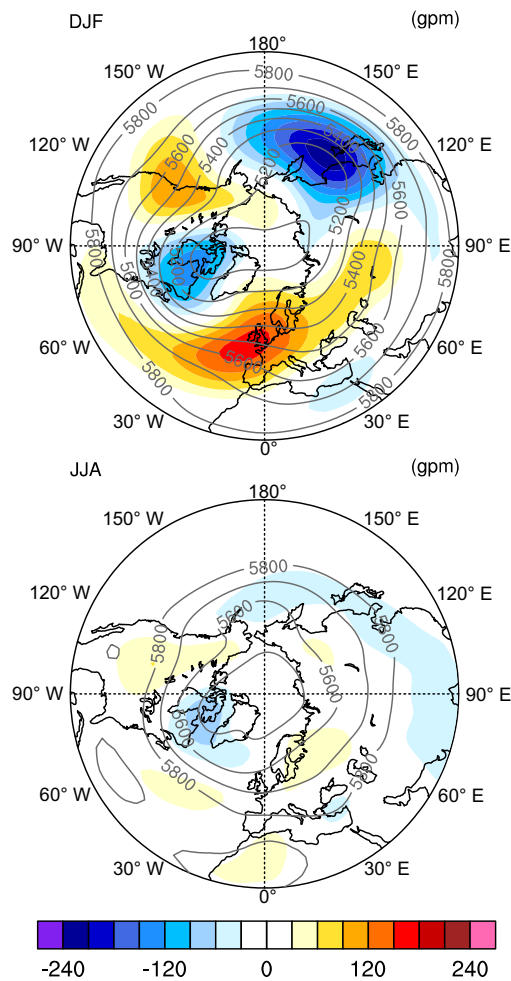


Figure 3 Mean 500-hPa geopotential height for (top) boreal winter (December–February) and (bottom) boreal summer (June–August), indicated by the gray contours for every 120-geopotential meters (gpm), over 1958–2006. The colors indicate departures from the zonal average.

the total variance in SLP over the North Atlantic and appears with a slight northwest to southeast orientation. In the so-called ‘positive phase’ (depicted), higher than normal surface pressures south of 55° N combine with a broad region of anomalously low pressure throughout the Arctic to enhance the climatological meridional pressure gradient (Figure 1). The largest amplitude anomalies occur in the vicinity of Iceland and across the Iberian Peninsula. The positive phase of the NAO is associated with stronger-than-average surface westerlies across the middle latitudes of the Atlantic onto Europe, with anomalous southerly flow over the eastern United States and anomalous northerly flow across the Canadian Arctic and the Mediterranean (Figure 7).

By boreal spring (March–May), the NAO appears as a north–south dipole with a southern center of action near the Azores. Both the spatial extent and the amplitude of the SLP anomalies are smaller than during winter, but not by much, and the leading EOF explains 30% of the SLP variance. The

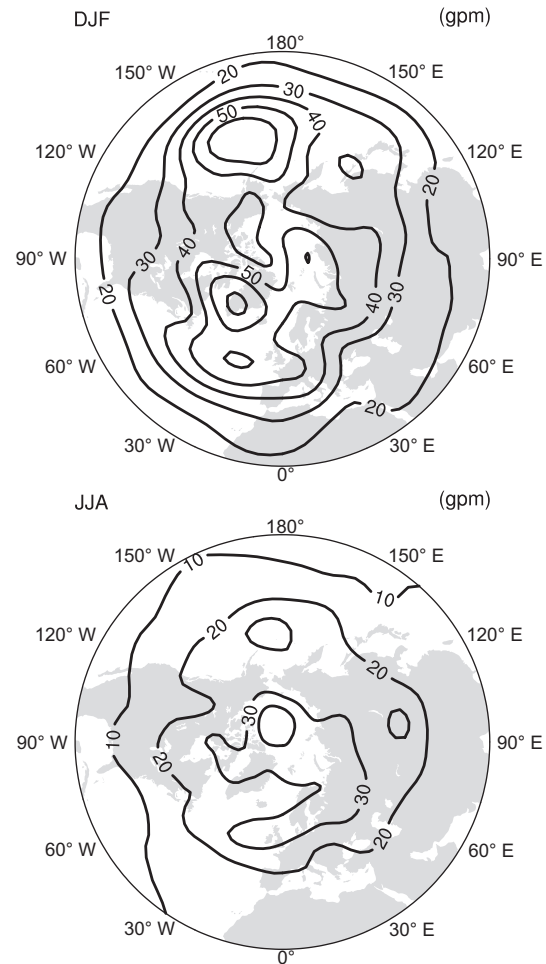
 $\sigma(z)$ 500 hPa

Figure 4 Interannual variability of 500-hPa geopotential height for (top) boreal winter (December–February) and (bottom) boreal summer (June–August) over 1958–2006. The contour increment is 10 gpm.

amplitude, spatial extent, and the percentage of total SLP variability explained by the NAO reach minimums during the summer (June–August) season, when the centers of action are substantially north and east relative to winter. By fall (September–November), the NAO takes on more of a southwest-to-northeast orientation, with SLP anomalies in the northern center of action comparable in amplitude to those during spring.

That the spatial pattern of the NAO remains largely similar throughout the year does not imply that it also tends to persist in the same phase for long. To the contrary, it is highly variable, tending to change its phase from one month to another, and its longer term time-average behavior reflects the combined effect of residence time in any given phase and its amplitude therein.

Most studies of the NAO focus on the NH winter months, when the atmosphere is most active dynamically and perturbations grow to their largest amplitudes. As a result, the influence of the NAO on surface temperature and precipitation is likely to be greatest at this time of year. However, coherent

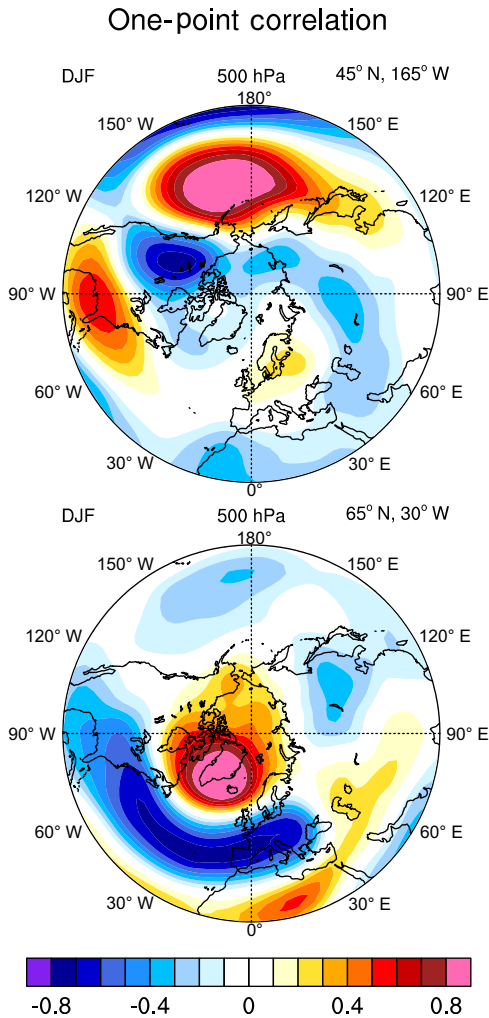


Figure 5 One-point correlation maps of 500-hPa geopotential heights for boreal winter (December–February) over 1958–2006. In the top panel, the reference point is 45° N, 165° W, corresponding to the primary center of action of the PNA pattern. In the lower panel, the NAO pattern is illustrated based on the reference point of 65° N, 30° W.

fluctuations of surface pressure, temperature, cloudiness, and precipitation occur throughout the year over the North Atlantic, and decadal and longer term variability is not confined to winter. Moreover, the vigorous wintertime NAO can interact with the slower components of the climate system (the ocean, in particular) to leave persistent surface anomalies into the ensuing parts of the year that may significantly influence the evolution of the climate.

EOF Analysis of NH SLP

A well-known shortcoming of EOF analysis is that eigenvectors are mathematical constructs, constrained by their mutual orthogonality and the maximization of variance over the entire analysis domain. There is no guarantee, therefore, that they represent physical/dynamical modes of the climate system. An EOF analysis, for instance, will not clearly reveal two patterns that are linearly superposed if those patterns are not

orthogonal. Moreover, the loading values of EOFs do not reflect the local behavior of the data: values of the same sign at two different spatial points in an EOF do not imply that those two points are significantly correlated. This means that the pattern structure of any particular EOF must be interpreted with care. These issues have been at the center of a debate over whether or not the NAO is a regional expression of a larger scale (hemispheric) mode of variability known as the NH annular mode (NAM) or, previously, the Arctic oscillation (AO).

The NAM is defined as the first EOF of NH (20–90° N) winter SLP data (Figure 8, upper panel, based on the cross-covariance matrix). It explains 23% of the extended winter-mean (December–March) variance, and it is clearly dominated by the NAO structure in the Atlantic sector. Although there are some subtle differences from the regional pattern (Figure 8, lower panel) over the Atlantic and Arctic, the main difference is larger amplitude anomalies over the North Pacific of the same sign as those over the Atlantic. This feature gives the NAM an almost annular (or zonally symmetric) structure that reflects a more hemispheric-scale meridional seesaw in SLP between polar and middle latitudes. Some have argued that the NAM is a fundamental structure of NH climate variability and that the ‘regional’ NAO reflects the modification of the annular mode by zonally asymmetric forcings, such as topography and land–ocean temperature contrasts. It would then follow that the annular mode perspective is critical in order to understand the processes that give rise to NAM (or NAO) variations. For instance, the leading wintertime pattern of variability in the lower stratosphere is clearly annular (not shown), but the SLP anomaly pattern that is associated with it is confined almost entirely to the Arctic and Atlantic sectors and coincides with the spatial structure of the NAO.

Cluster Analysis of North Atlantic SLP

The dynamical signature of interannual variability in the North Atlantic domain can also be examined through nonlinear approaches, such as cluster analysis or nonlinear PC analysis. Here the former one was applied to 57 years of daily SLP data from December to March. Briefly, cluster analysis is a multivariate statistical technique that groups together the daily SLP maps into a small number of representative states (or regimes) according to an objective criterion of similarity. Note that by construction, the percentage of occurrence of the identified clusters sums to 100.

The clustering algorithm applied over the Atlantic domain (20–70° N; 90° W–40° E) identifies four winter climate regimes in SLP (Figure 9). Two of them correspond to the negative and positive phases of the NAO, while the third and fourth regimes display strong anticyclonic ridges over Scandinavia (the ‘blocking’ regime) and off western Europe (the ‘Atlantic ridge’ regime). The latter bears some resemblance to another prominent atmospheric teleconnection: the East Atlantic pattern. All four regimes occur with about the same frequency (20–30% of all winter days), although these numbers are sensitive to the period of analysis, reflecting that the dominance of certain regimes over others varies over time (see Section Temporal Variability of the NAO).

In contrast to the typical NAO pattern identified through linear approaches (e.g., Figures 5 and 6), some interesting

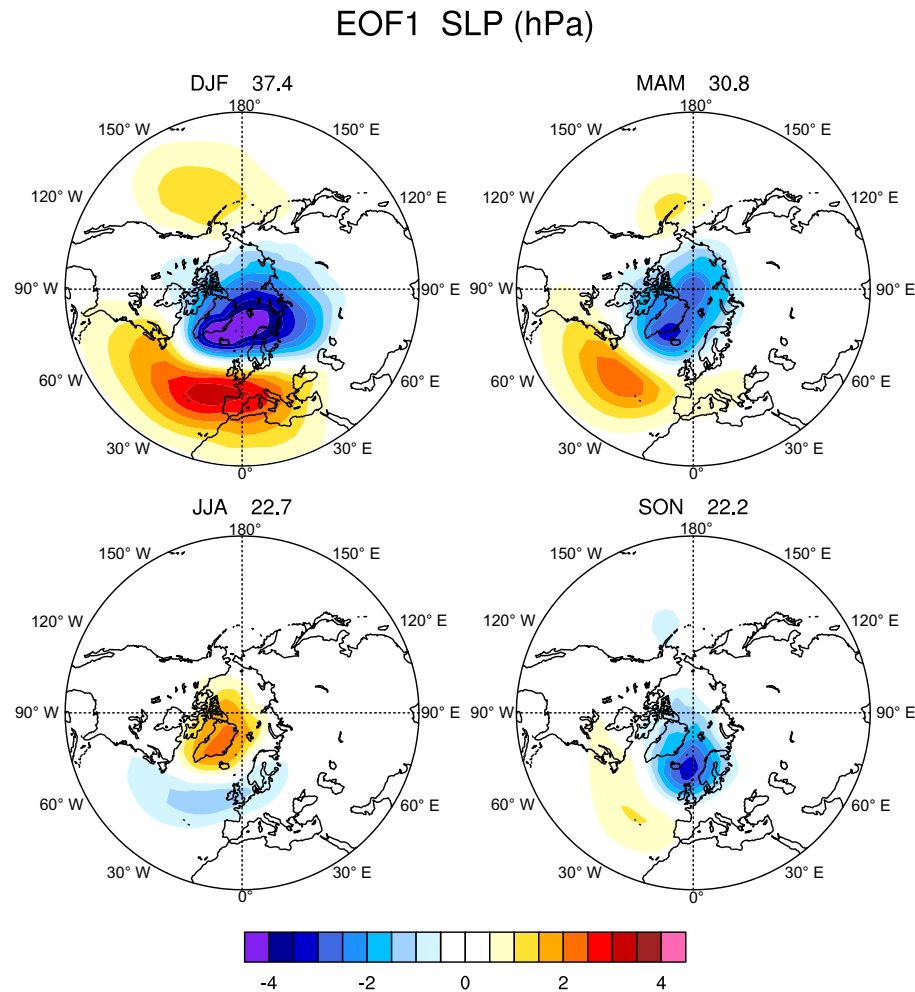


Figure 6 Leading EOF1 of the seasonal mean SLP anomalies in the North Atlantic sector (20–70° N, 90° W–40° E), and the percentage of the total variance they explain. The patterns are displayed in terms of amplitude (hPa), obtained by regressing the hemispheric SLP anomalies upon the leading PC time series. The data cover 1899–2006.

spatial asymmetries are evident in [Figure 9](#). Most striking is the difference in the position of the pressure anomalies between the two NAO regimes: in particular, the eastward shift (by $\sim 30^\circ$ longitude) and northeastward extension of the subpolar SLP anomalies in the positive relative to the negative regime. These spatial asymmetries are not dependent on the analysis period: they are evident in the subperiods of the SLP data set. Similar results, indicating a nonlinearity in NAO variability, are found when the PC time series of the leading EOF of Atlantic SLP is used to define and average together positive and negative index winters (not shown). The robustness of the eastward displacement of the NAO in positive regime months has interesting implications for conclusions drawn from some climate model studies on how increasing greenhouse gas (GHG) concentrations might affect the spatial structure of the NAO. For instance, some studies have concluded that future enhanced GHG forcing might result in an eastward displacement of the NAO centers of action. The results from the regime analysis, however, suggest that longitudinal shifts could arise from the preferential excitement of positive NAO regimes,

which are intrinsically displaced eastward, rather than a static shift of the Atlantic pressure centers.

Temporal Variability of the NAO

Since there is no unique way to define the spatial structure of the NAO, it follows that there is no universally accepted index to describe the temporal evolution of the phenomenon. Most modern NAO indices are derived either from the simple difference in surface pressure anomalies between various northern and southern locations or from the PC time series of the leading (usually regional) EOF of SLP. Many examples of the former exist, usually based on instrumental records from individual stations near the NAO centers of action, but sometimes from gridded SLP analyses.

A disadvantage of station-based indices is that they are fixed in space. Given the movement of the NAO centers of action through the annual cycle ([Figure 6](#)), such indices can only adequately capture NAO variability for parts of the year.

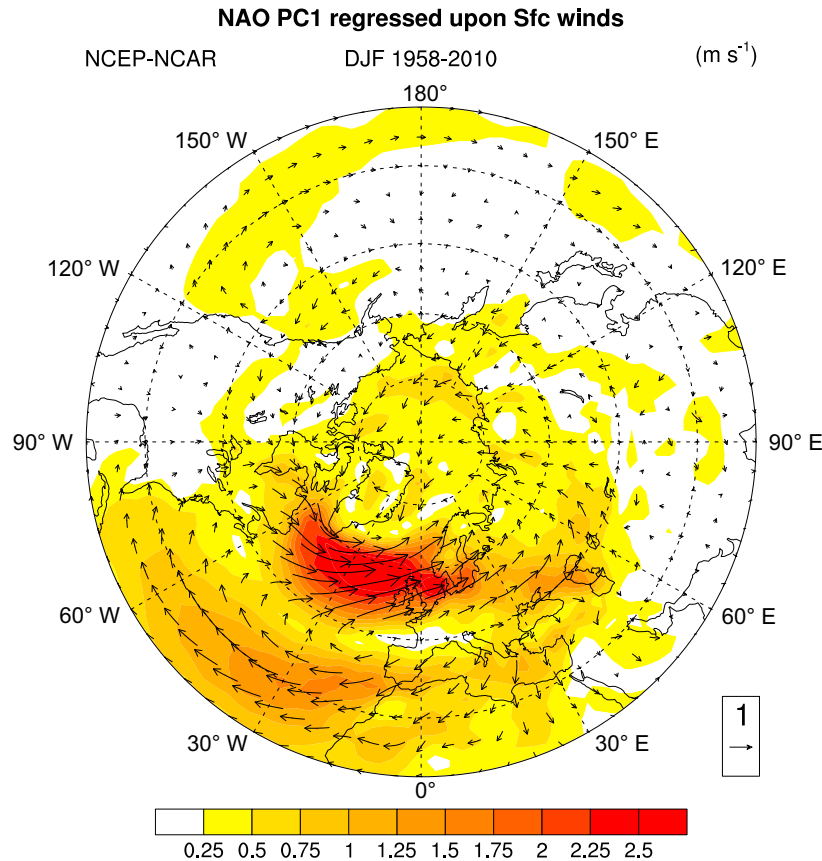


Figure 7 The amplitude of boreal winter (December–February) 1000-hPa scalar wind speed (color) and vector wind (arrows) associated with a one standard deviation change of an NAO index, defined as the PC time series of the leading EOF of Atlantic sector SLP (as in [Figure 6](#)). The plot is constructed from winter data from the NCEP/NCAR reanalyses over 1958–2006. The color scale is in units of meters per second, and the scaling vector (lower right) is 1 m s^{-1} .

Moreover, individual station pressures are significantly affected by small-scale and transient meteorological phenomena not related to the NAO and, thus, contain noise.

An advantage of the PC time series approach is that such indices are more optimal representations of the full NAO spatial pattern; yet, as they are based on gridded SLP data, they can only be computed for short (relative to some station records) periods of time, depending on the data source. Below a station-based index was compared to the PC time series of the leading EOF (PC1) of both Atlantic sector and NH SLP. The latter is the NAM index. The time history of occurrence of the NAO regimes identified in [Figure 9](#) was presented. All comparisons are for the winter (December–March) season.

A widely used winter-mean NAO station index is shown in [Figure 10](#) (upper panel). Positive values of the index indicate stronger-than-average westerlies over the middle latitudes. The station-based index for the winter season agrees well with PC1 of the Atlantic sector SLP: the correlation coefficient between the two is 0.92 over the common period since 1899, indicating that the station-based index adequately represents the time variability of the winter-mean NAO spatial pattern. Moreover, it correlates with the NAM index (lower panel) at 0.85, while the correlation of the two PC1 time series is 0.95. These results again emphasize that the NAO and NAM reflect essentially the

same mode of tropospheric variability. When intraseasonal anomalies are considered by stringing together the individual winter months, the correlation coefficient between the two PC1 time series is reduced slightly to 0.89.

An important conclusion from [Figure 10](#) is that there is little evidence for the NAO to vary on any preferred time scale. Large changes can occur from one winter to the next, as well as from one decade to the next. The power spectra of the indices in [Figure 10](#) are only slightly 'red,' with power increasing with period (not shown). When the spectral characteristics of the NAO are examined using daily data (not shown), its temporal evolution is generally consistent with a stochastic (Markov or first-order autoregressive) process with a fundamental time-scale of about 10 days. This then means that observed inter-annual and longer timescale NAO fluctuations ([Figure 10](#)) could primarily be a statistical remnant of the energetic weekly variability. This 'climate noise paradigm' fails, however, to explain the enhanced interannual NAO variability observed over the last half of the twentieth century, when a role for forcing by other climate system components is likely. Indeed, numerous studies have argued that variations in heat exchange between the atmosphere and ocean, sea ice and/or land systems could modulate NAO variability on seasonal to multi-decadal timescales.

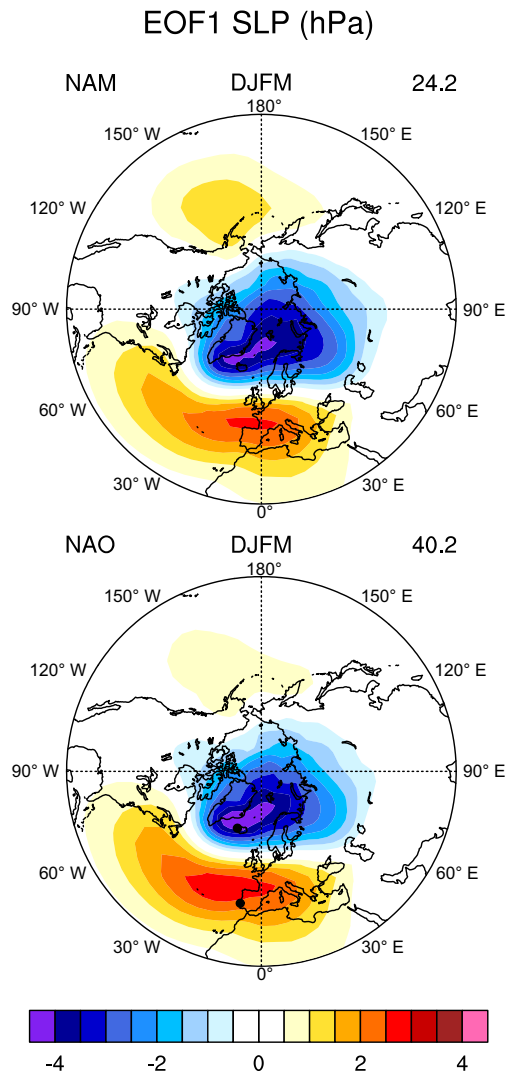


Figure 8 Leading EOF1 of the winter (December–March) mean SLP anomalies over (top) the NH (20–90° N) and (bottom) the North Atlantic sector (20–70° N, 90° W–40° E), and the percentage of the total variance they explain. The patterns are displayed in terms of amplitude (hPa), obtained by regressing the hemispheric SLP anomalies upon the leading PC time series. The data cover 1899–2006. The dots in the bottom panel represent the locations of Lisbon, Portugal, and Stykkisholmur, Iceland used in the station-based NAO index of Hurrell (1995; see [Figure 10](#)).

Another index, the time history of the occurrence of NAO, blocking and Atlantic ridge regimes ([Figure 8](#)), offers a different perspective ([Figure 11](#)). Plotted is the frequency of occurrence of each regime in units of the number days a regime is present within a given winter (December–March) season. For the two NAO regimes, as for the more conventional indices, strong interannual variability is evident, and there are periods when one NAO regime occurs almost to the exclusion of the other. For instance, very few positive NAO regime occurrences are found during the 1960s, while very few negative regime occurrences were observed during the 1990s, consistent with the upward trend in traditional NAO indices over this period ([Figure 10](#)).

The regime analysis also illustrates two other important points. First, there is a large amount of within-season variance in the atmospheric circulation of the North Atlantic. Most winters are not dominated by any particular regime; rather, the atmospheric circulation anomalies in one month might resemble the positive index phase of the NAO, while in another month, they resemble the negative index phase or some other pattern altogether. Since 2001, for instance, more winter days over the North Atlantic have been characterized by circulation anomalies that project onto the Atlantic ridge or blocking patterns than either phase of the NAO ([Figure 11](#)). Moreover, roughly the same numbers of negative and positive NAO index days have occurred over this period, consistent with the small, winter-mean values of conventional NAO indices ([Figure 10](#)). Thus, the second point is that although the NAO is the dominant pattern of atmospheric circulation variability over the North Atlantic, it explains only a fraction of the total variance, and most winters cannot be characterized solely by the canonical NAO pattern in [Figure 6](#).

Impacts of the NAO

The NAO exerts a dominant influence on wintertime temperatures across much of the NH. Surface air temperature and SST across wide regions of the North Atlantic Ocean, North America, the Arctic, Eurasia, and the Mediterranean are significantly correlated with NAO variability. These changes, along with related changes in storminess and precipitation, ocean heat content, ocean currents and their related heat transport, and sea ice cover, have significant impacts on a wide range of human activities as well as on marine, freshwater, and terrestrial ecosystems. In the following, a very brief overview of these impacts is presented.

Storms and Precipitation

Changes in the mean circulation patterns over the North Atlantic associated with the NAO are accompanied by changes in the intensity and number of storms, their paths, and their weather. During winter, a well-defined storm track connects the North Pacific and North Atlantic basins, with maximum storm activity over the oceans. Positive NAO index winters are associated with a northeastward shift in the Atlantic storm activity with enhanced activity from Newfoundland into northern Europe and a modest decrease in activity to the south. Positive NAO index winters are also typified by more intense and frequent storms in the vicinity of Iceland and the Norwegian Sea.

The ocean exhibits a marked response to long-lasting shifts in the storm climate. For instance, the very persistent and positive NAO index winters of the 1990s were associated with increased wave heights over the northeast Atlantic. Such changes have consequences for the regional ecology, as well as for the operation and safety of shipping, offshore industries, such as oil and gas exploration, and coastal development.

Changes in the mean flow and storminess associated with swings in the NAO index are also reflected in pronounced changes in the transport and convergence of atmospheric moisture and, thus, the distribution of precipitation ([Figure 12](#)). Winters tend to be dry over much of Greenland

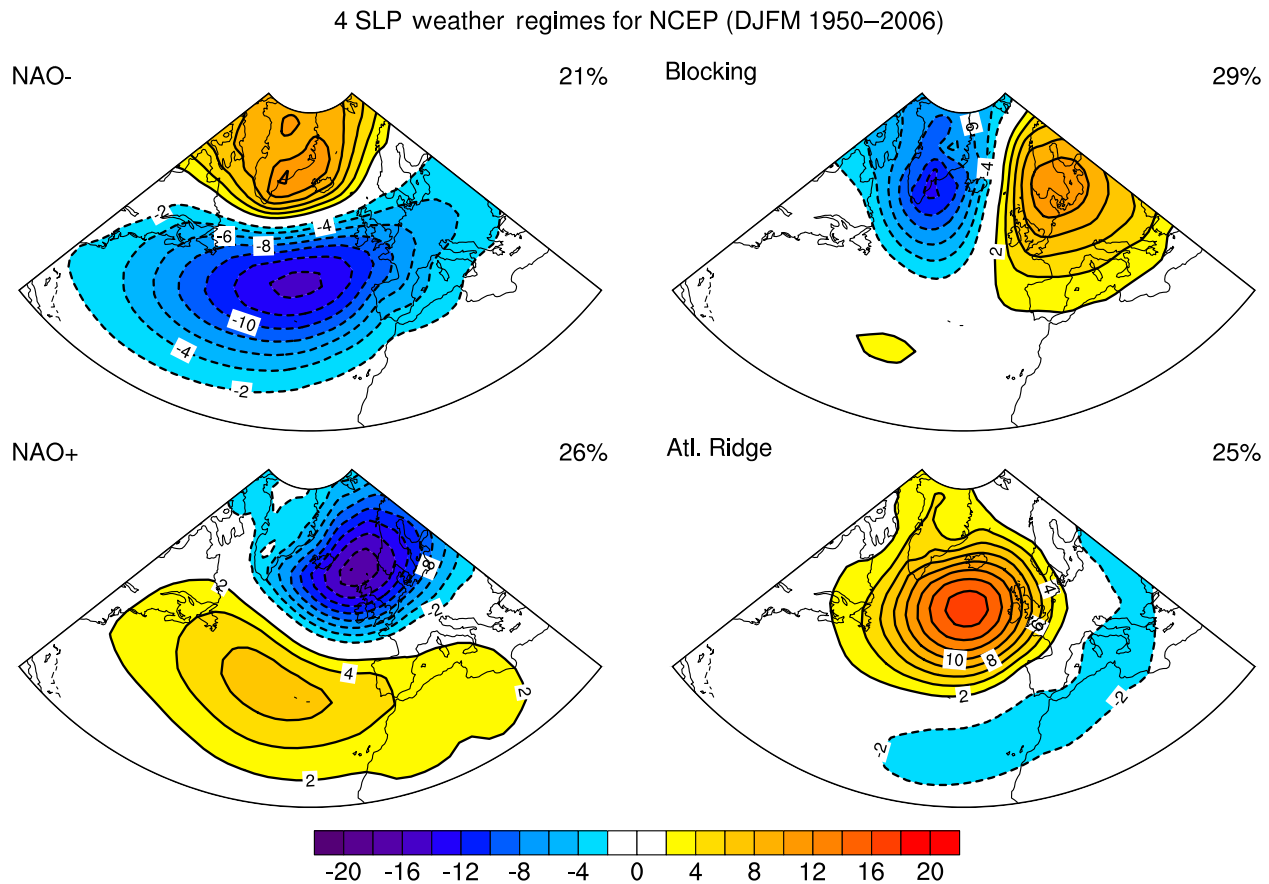


Figure 9 Boreal winter (December–March) climate regimes in SLP (hPa) over the North Atlantic domain (20–70° N, 90° W–40° E) using daily data over 1950–2006. The percentage at the top right of each panel expresses the frequency of occurrence of a cluster out of all winter days since 1950. The contour interval is 2 hPa.

and the Canadian Arctic during high NAO index winters. Drier winter conditions of the same magnitude also occur over much of central and southern Europe, the Mediterranean and parts of the Middle East, whereas more precipitation than normal falls from Iceland through Scandinavia.

Surface Temperature and SST

The NAO exerts a dominant influence on wintertime temperatures across much of the NH. Surface air temperature and SST across wide regions of the North Atlantic Ocean, North America, the Arctic, Eurasia, and the Mediterranean are significantly correlated with NAO variability. Such changes in surface temperature (and related changes in rainfall and storminess) can have significant impacts on a wide range of human activities as well as on marine and terrestrial ecosystems.

When the NAO index is positive, enhanced westerly flow across the North Atlantic during winter moves relatively warm (and moist) maritime air over much of Europe and far downstream across Asia, while stronger northerlies over Greenland and northeastern Canada carry cold air southward and decrease land temperatures and SST over the northwest Atlantic (Figure 13). Temperature variations over North Africa and the Middle East (cooling), as well as North America (warming),

associated with the stronger clockwise flow around the subtropical Atlantic high-pressure center are also notable.

The pattern of temperature change associated with the NAO is important. Because the heat storage capacity of the ocean is much greater than that of the land, changes in continental surface temperatures are much larger than those over the oceans, so they tend to dominate average NH (and global) temperature variability. Given especially the large and coherent NAO signal across the Eurasian continent from the Atlantic to the Pacific (Figure 13), it is not surprising that NAO variability explains about one-third of the NH interannual surface temperature variance during winter.

Over the oceans, it has long been recognized that fluctuations in SST and the strength of the NAO are related. The leading pattern of SST variability during boreal winter (not shown) consists of a tripolar structure marked, in one phase, by a cold anomaly in the subpolar North Atlantic, a warm anomaly in the middle latitudes centered off Cape Hatteras, and a cold subtropical anomaly between the equator and 30° N. This structure suggests that the SST anomalies are driven by changes in the air–sea heat exchanges and surface wind-induced Ekman currents associated with NAO variations. The relationship is indeed strongest when the NAO index leads an index of the SST variability by several weeks, which highlights

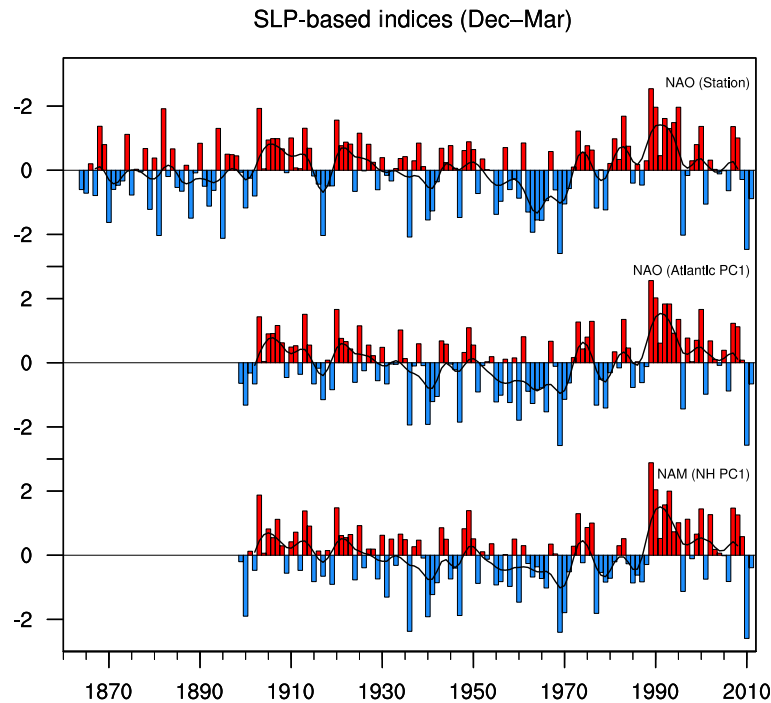


Figure 10 Normalized indices of the mean winter (December–March) NAO through 2011 constructed from SLP data. In the top panel, the index is based on the difference of normalized SLP between Lisbon, Portugal, and Stykkisholmur/Reykjavik, Iceland. The average winter SLP data at each station were normalized by the division of each seasonal pressure by the long-term mean (1864–1983) standard deviation. In the middle and lower panels, the index is the PC time series of the leading EOF of Atlantic sector and NH SLP, respectively. The heavy solid lines represent the indices smoothed to remove fluctuations with periods less than 4 years. The indicated year corresponds to the January of the winter season (e.g., 1990 is the winter of 1989/1990).

the well-known result that large-scale SST over the extratropical oceans responds to atmospheric forcing on monthly and seasonal timescales. Compositing North Atlantic SST on high and low NAO index winters clearly illustrates the aforementioned tripole pattern of SST change (Figure 14).

Over longer periods, persistent SST anomalies also appear to be related to persistent anomalous patterns of SLP, including those associated with the NAO, but the mechanisms whereby the atmosphere forces SST anomalies on decadal and longer timescales are different from those on interannual timescales. On decadal and longer timescales, the ocean adjusts dynamically to the overlying changes in wind stress curl, both locally via Ekman pumping and nonlocally through changes in the gyre-scale circulation. This dynamical adjustment alters the horizontal and vertical oceanic heat transports, which in turn impact SST. It is quite likely, for instance, that sustained NAO forcing results in a basin-wide SST response in which the northern and subtropical parts of the tripolar pattern merge. There is also evidence for a northward shift in the position of the Gulf Stream during the positive phase of the NAO, consistent with Sverdrup adjustment of the ocean gyre circulation.

Subsurface Ocean Changes

Subsurface ocean observations more clearly depict long-term climate variability, because the effect of the annual cycle and month-to-month variability in the atmospheric circulation

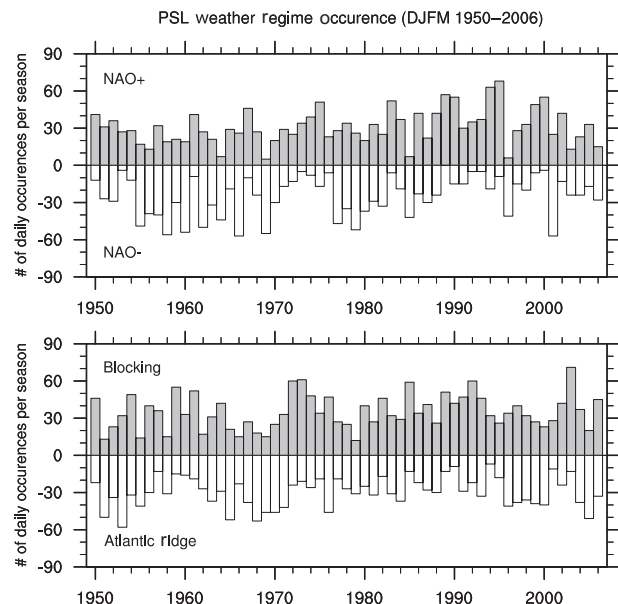


Figure 11 The time history of occurrence of the NAO, Atlantic ridge, and blocking regimes (see Figure 9) over 1950–2006. The vertical bars give the number of days in each winter (December–March) season that the given regime is present. The indicated year corresponds to the January of the winter season (e.g., 1990 is the winter of 1989/1990).

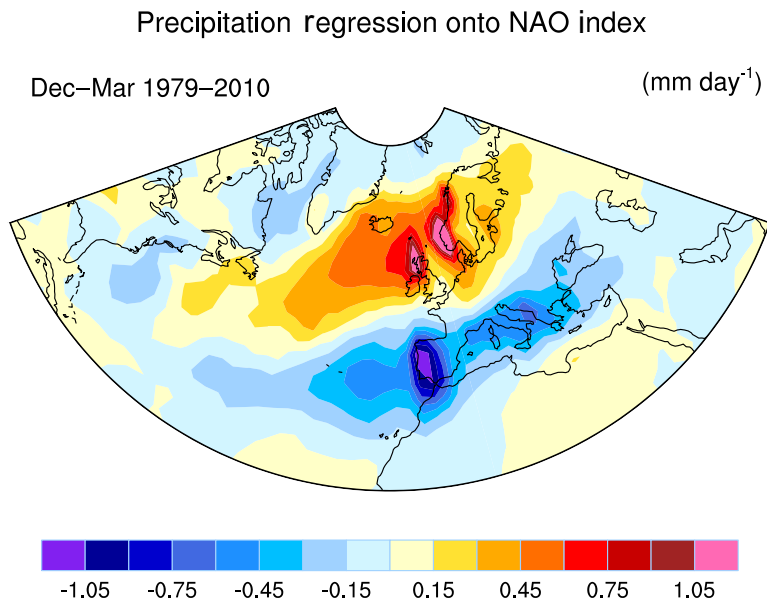


Figure 12 Changes in mean winter (December–March) precipitation (millimeters per day) associated with a one standard deviation change in the NAO index, defined as in the middle panel of [Figure 10](#). The precipitation data cover 1979 through 2010.

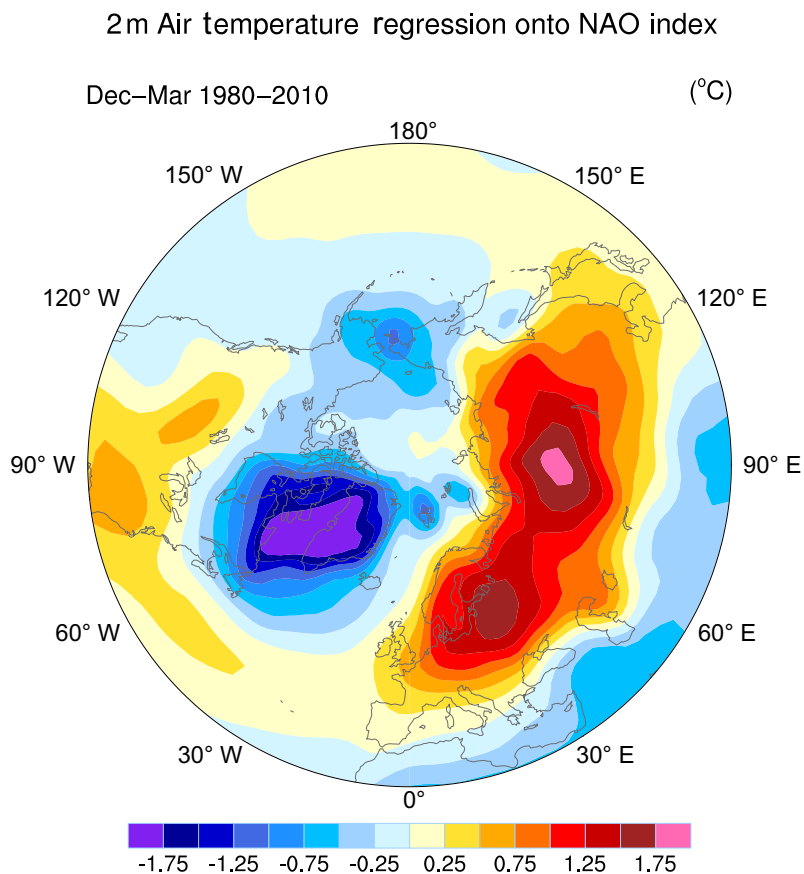


Figure 13 Changes in mean winter (December–March) air temperature (°C) associated with a one standard deviation change in the NAO index, defined as in the middle panel of [Figure 10](#). The temperature data cover 1980 through 2010.

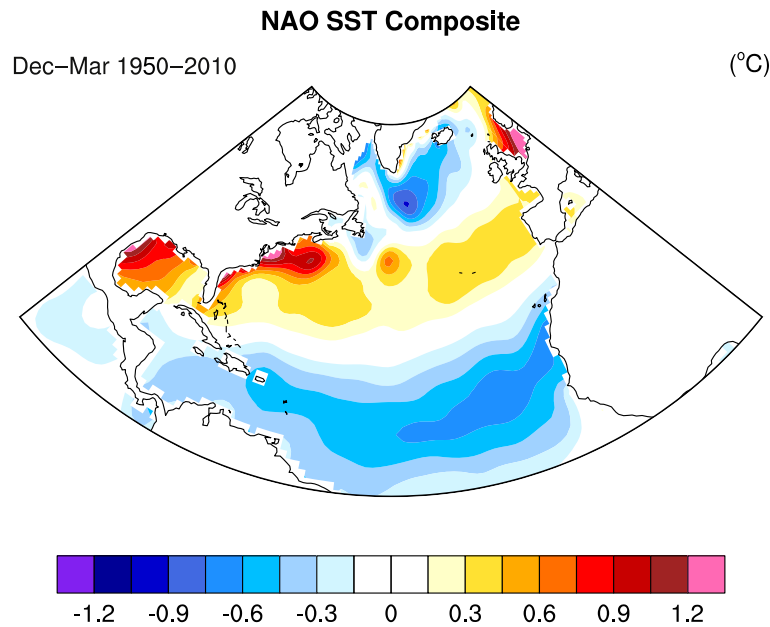


Figure 14 Difference in mean winter (December–March) SST between years when the NAO index exceeds one standard deviation over 1950–2010. The NAO index is defined as in the middle panel of [Figure 10](#).

decays rapidly with depth. These measurements are much more limited than surface observations, but over the North Atlantic they too indicate fluctuations that are coherent with the low-frequency winter NAO index to depths of 400 m.

Oceanic mixed layer depth (MLD) is an important physical factor influencing marine biological productivity and ecosystem dynamics. MLD is influenced by atmospheric and oceanic conditions through wind-induced vertical mixing, heat exchange, and upwelling. The response of the winter MLD to the NAO, estimated by compositing winter months

(December–March) over 1955–2003, consists of a region of negative anomalies (shallower-than-normal mixed layers) extending across much of the middle North Atlantic from the southeastern United States to Spain and a region of weaker positive anomalies to the south. ([Figure 15](#); note that there are insufficient data to determine NAO-related MLD variations in the subpolar North Atlantic.) This pattern of MLD anomalies is similar to the pattern of SST anomalies associated with the NAO ([Figure 14](#)), with negative MLD anomalies corresponding to positive SST anomalies and vice versa.

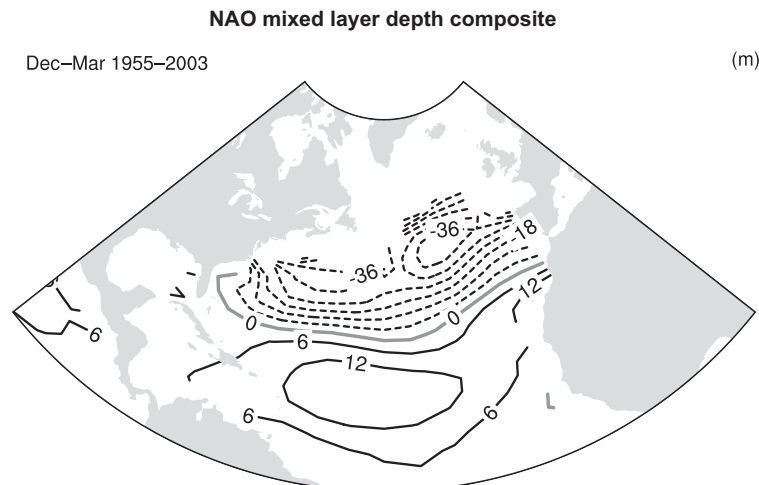


Figure 15 Difference in mean winter (December–March) ocean mixed layer depth (m) between years when the NAO index (defined as in the middle panel of [Figure 10](#)) exceeds one standard deviation over 1955–2003. The contour increment is 6 m and positive (negative) differences are given by the solid (dashed) contours.

The SST tripole pattern (Figure 14) has been shown to recur from one winter to the next with little persistence during the intervening summer. The mechanism for this winter-to-winter memory of the SST anomaly tripole is due to the seasonal cycle of MLD through the so-called ‘reemergence mechanism.’ Briefly, the winter NAO creates a tripole pattern of ocean temperature anomalies that extend down to the base of the deep winter mixed layer. These anomalies persist at depth through spring and summer within the stably stratified seasonal thermocline, insulated from the atmosphere by the formation of a shallow mixed layer in response to increasing solar radiation and weakening stirring due to slackened surface winds. Their sequestration ends in the following fall or early winter when the mixed layer deepens again due to the seasonal intensification of the extratropical atmospheric circulation, and the thermal anomalies created the previous winter become re-entrained into the mixed layer, affecting SST. This re-entrainment thus leads to the ‘reemergence’ of the previous winter’s SST anomalies.

The oceanic response to NAO variability is also evident in changes in the distribution and intensity of winter convective activity in the northern North Atlantic. The convective renewal of intermediate and deepwaters in the Labrador Sea and the Greenland–Iceland–Norwegian (GIN) Seas contributes significantly to the production and export of North Atlantic deep-water and, thus, helps to drive the global thermohaline circulation. The intensity of winter convection at these sites is not only characterized by large interannual variability, but also interdecadal variations that appear to be synchronized with variations in the NAO. Deep convection over the Labrador Sea, for instance, was at its weakest and shallowest in the postwar instrumental record during the late 1960s. Since then, Labrador Sea water has become progressively colder and fresher, with intense convective activity to unprecedented ocean depths (>2300 m) in the early 1990s. In contrast, warmer and saltier deepwaters in recent years are the result of suppressed convection in the GIN Seas, whereas tracer evidence suggests that intense convection likely occurred during the late 1960s.

Sea Ice

The leading pattern of variability of winter Arctic sea ice concentrations exhibits a seesaw in ice extent between the Labrador and Greenland Seas. Strong interannual variability is evident in the sea ice changes, as are longer term fluctuations including a trend over the past 30 years of diminishing (increasing) ice concentration during boreal winter east (west) of Greenland. Associated with the sea ice fluctuations are large-scale changes in SLP that closely resembles the NAO.

When the NAO is in its positive index phase, the Labrador Sea ice boundary extends farther south while the Greenland Sea ice boundary is north of its climatological extent (not shown). This is qualitatively consistent with the notion that the atmosphere directly forces the sea ice anomalies, either dynamically via wind-driven ice drift anomalies or thermodynamically through surface air temperature anomalies. The relationship between the NAO index and an index of the North Atlantic ice variations is strong, although it does not hold for all individual winters, illustrating the importance of the regional atmospheric circulation in forcing the extent of sea ice. In general, the

winters over the past decade have witnessed a retreat of the ice edge throughout the Arctic, even during the recent winters of weak NAO-related atmospheric circulation anomalies.

Summary and Challenges

In this article, we have presented a basic review of modal variability over the North Atlantic. We have focused, in particular, on the NAO, as it is the dominant mode of regional climate variability. It is therefore critical to understand the mechanisms that control and affect the NAO and its temporal evolution.

There is ample evidence that most of the atmospheric circulation variability in the form of the NAO arises from the internal, nonlinear dynamics of the extratropical atmosphere. Interactions between the time-mean flow and synoptic-time-scale transient eddies are the central governing dynamical mechanism. As such, the month-to-month and even year-to-year changes in the phase and amplitude of the NAO are largely unpredictable. But that external forces might nudge the atmosphere to assume a high or low NAO index value over a particular month or season is important: even a small amount of predictability could be useful considering the significant impact the NAO exerts on the weather and climate of the NH.

There are a number of different mechanisms that could influence the detailed state of the NAO. Within the atmosphere itself, changes in the rate and location of tropical heating have been shown to be one way to influence the atmospheric circulation over the North Atlantic and, in particular, the NAO. Tropical convection, in turn, is sensitive to the underlying SST distribution, which exhibits much more persistence than SST variability in middle latitudes. This might lead, therefore, to some predictability of the NAO phenomenon. In addition, climate modeling work has shown that the atmospheric response to the re-emerging North Atlantic SST tripole resembles the phase of the NAO that created the SST tripole the previous winter, thereby modestly enhancing the winter-to-winter persistence of the NAO.

Interactions with the lower stratosphere are also important. This mechanism is of interest because it might also explain how changes in atmospheric composition influence the NAO. For example, changes in ozone, GHG concentrations and/or levels of solar output affect the radiative balance of the stratosphere that, in turn, modulates the strength of the winter polar vortex. Given the relatively long timescales of stratospheric circulation variability (anomalies persist for weeks), dynamic coupling between the stratosphere and the troposphere via wave-mean flow interactions could yield a useful level of predictive skill for the wintertime NAO. One example of the possible role of the lower stratosphere in the wintertime NAO involves Eurasian snow cover conditions. Several studies have suggested a mechanism whereby autumnal snow extent over Eurasia influences the winter NAO via dynamical coupling between the stratosphere and troposphere, with above (below) normal October snow extent leading to a negative (positive) phase of the winter NAO.

One of the most urgent challenges is to advance our understanding of the interaction between GHG forcing and the NAO. It now appears as though there may well be a deterministic relationship, which might allow for moderate low-

frequency predictability and thus needs to be studied carefully. Also, while the predictability of seasonal to interannual NAO variability will most likely remain low, some applications may benefit from the fact that this phenomenon leaves long-lasting imprints on surface conditions, in particular over the oceans.

See also: **Air Sea Interactions:** Momentum, Heat, and Vapor Fluxes; Sea Surface Temperature; Surface Waves. **Arctic and Antarctic:** Arctic Climate. **Boundary Layer (Atmospheric) and Air Pollution:** Overview. **Climate and Climate Change:** Climate Variability: Seasonal and Interannual Variability; Climate Variability: Decadal to Centennial Variability; Volcanoes: Role in Climate. **Cryosphere:** Sea Ice. **Dynamical Meteorology:** Wave Mean-Flow Interaction. **General Circulation of the Atmosphere:** Teleconnections. **Oceanographic Topics:** General Processes; Thermohaline Circulation; Water Types and Water Masses. **Ozone Depletion and Related Topics:** Ozone Depletion Potentials.

Further Reading

- Barnston, A.G., Livezey, R.E., 1987. Classification, seasonality and persistence of low frequency atmospheric circulation patterns. *Monthly Weather Review* 115, 1083–1126.
- Cassou, C., Terray, L., Hurrell, J.W., Deser, C., 2004. North Atlantic winter climate regimes: spatial asymmetry, stationarity with time, and oceanic forcing. *Journal of Climate* 17, 1055–1068.
- Hurrell, J.W., 1995. Decadal trends in the North Atlantic Oscillation, regional temperatures and precipitation. *Science* 269, 676–679.
- Hurrell, J.W., Deser, C., 2009. Atlantic climate variability. *Journal of Marine Systems* 78, 28–41.
- Hurrell, J.W., Kushnir, Y., Visbeck, M., Ottersen, G., 2003. An overview of the North Atlantic Oscillation. In: Hurrell, J.W., Kushnir, Y., Ottersen, G., Visbeck, M. (Eds.), *The North Atlantic Oscillation, Climatic Significance and Environmental Impact*. Geophysical Monograph Series, vol. 134. AGU, pp. 1–35.
- Osprey, S., Ambaum, M., 2011. Evidence for the chaotic origin of Northern Annular Mode variability. *Geophysical Research Letters* <http://dx.doi.org/10.1029/2011GL048181>.
- Thompson, D.W.J., Wallace, J.M., 2000. Annular modes in the extratropical circulation, part I, month-to-month variability. *Journal of Climate* 13, 1000–1016.

Climate Variability: Seasonal and Interannual Variability

DS Gutzler, University of New Mexico, Albuquerque, NM, USA

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

Seasonal variability of climate is principally controlled by changes in the amount of incoming solar radiation hitting different latitudes during Earth's year-long orbit around the Sun. Continents and oceans respond differently to seasonal changes in solar radiation, adding geographical complexity to seasonal variability. Interannual variability in the atmosphere can be organized into very large scale patterns of wind and temperature fluctuations. These patterns are often found to be associated with slow shifts in ocean circulation and temperature. Coupled ocean-atmosphere models have demonstrated increasing skill in simulating and predicting interannual variability.

Introduction

Seasonal variability refers to periodic fluctuations of average temperature, precipitation, and other weather variables resulting directly or indirectly from the annual cycle of incoming solar radiation. The source of seasonal variability is the well-understood annual periodicity of insolation associated with Earth's orbit around the Sun. Seasonal insolation changes directly force temperature fluctuations with periods corresponding to the insolation variability, causing regular 12-month (annual) periodicities in many locations, especially in middle latitudes, and 6-month (semiannual) periodicities in some regions near the equator or the poles. Thermal inertia in the climate system causes the seasonal cycle of temperature to lag somewhat behind the seasonal cycle of insolation, particularly over the oceans. Other climate variables, such as pressure, humidity, and precipitation, vary seasonally as the result of changes in large-scale atmospheric and oceanic circulations that in turn are associated with the seasonal variability of temperature and temperature gradients. Some annually periodic features of the climate system, such as monsoon onset dates, represent abrupt changes in circulation patterns in response to the smoothly varying seasonal insolation cycle. In a few places, notably along the equator across the Pacific Ocean, atmospheric or oceanic circulation systems respond to insolation variability in ways that create seasonal cycles of temperature and precipitation that do not correspond directly to seasonal changes in insolation.

Interannual variability is defined in terms of the year-to-year fluctuations in seasonal (or monthly) averages of temperature, precipitation, or some other climate variable about the average seasonal value determined by a long-term (generally 30-year) climatological record. Interactions between the atmosphere and upper ocean account for considerable interannual variability. Similar interactions between the atmosphere and continental surface anomalies also modulate climate on interannual time scales but probably to a lesser extent. Some forms of intermittent forcing external to the climate system, such as fluctuations of solar brightness or increased aerosol concentrations caused by large volcanic eruptions, also affect the climate system for more than a season and thus contribute to interannual variability.

Seasonal and interannual variability are at the short time scale, or 'short-term,' end of the spectrum of climate variability.

Seasonal and longer fluctuations are considerably slower than typical 'weather' (i.e., much slower than the time scale of growth and decay of individual clouds or synoptic weather systems). The time scales of seasonal and interannual variability are much longer than the deterministic limit of weather prediction determined from chaos theory, which is thought to be several weeks. Therefore, efforts to understand and predict these short-term climate fluctuations must depend on knowledge of persistent or predictable nonatmospheric forcing. The ultimate cause for seasonal variability is well known, but interannual fluctuations can be much more difficult to attribute to individual causes.

Patterns of seasonal and interannual variability of temperature or precipitation are heavily modulated by large-scale atmospheric dynamics. Planetary wave anomalies organize much of the interannual variability into large-scale teleconnection patterns, or zonally symmetric annular modes, which are often defined in terms of pressure or geopotential height anomalies. The existence of fixed geographical patterns of interannual variability provides a means for these fluctuations to be described and monitored in terms of relatively few large-scale indices of the anomalous circulation, and it provides a framework for assessing the potential long-range predictability of interannual fluctuations.

Causes and Magnitude of Seasonal Variability

The seasonal cycle of Earth's atmosphere is a consequence of systematic changes in incoming solar radiation (or 'insolation') associated with Earth's orbit around the Sun. The primary parameter in determining seasons is the tilt of Earth's rotational axis relative to the plane of its orbit, currently about 23.5° . The Northern Hemisphere is tilted toward the Sun during the half-year that is centered on the boreal summer solstice and tilted away from the Sun during the half-year that is centered on the boreal winter solstice. Earth's orbit is nearly circular at the present time, so the small seasonal variation in the distance from the Sun to Earth plays a much smaller role than the tilt in determining the seasons. Very slow changes in the tilt and circularity of Earth's orbit, so-called Milankovitch fluctuations, cause climate changes over millennial time scales, but these orbital characteristics are considered to be essentially fixed for describing seasonal and interannual variability in this article.

The tilt of Earth's rotational axis relative to its solar orbital axis determines the density of insolation impinging on different latitudes, a phenomenon known as 'beam spreading.' The most intense input of solar radiation occurs at the latitude where the Sun is directly overhead at noon local time. On the dates of the spring and autumnal equinoxes, the Sun is overhead at noon on the equator. This latitude shifts northward and southward between 23°27' latitude in either hemisphere as Earth progresses through its annual orbit of the Sun. The northernmost latitude of maximum insolation density, 23°27'N, is denoted the Tropic of Cancer; maximum insolation occurs there on the boreal summer solstice, on or about June 22. The corresponding southern limit of maximum insolation is called the Tropic of Capricorn, where maximum insolation occurs on the austral summer solstice, usually December 22. The band of latitude within 23.5° of the equator is generally denoted the 'tropics.' Insolation density within the tropics, while variable during the year, is generally high year-round, so seasonal variability of temperature is relatively modest compared to that of higher latitudes.

Near Earth's geographic poles, the seasonal variability of insolation is extreme. Earth's tilt means that no direct sunlight at all is received in the core of the winter season poleward of about 66.5°, so seasonal temperature variability is much larger at high latitudes compared to low latitudes. The Arctic and Antarctic circles, at 66°33' in either hemisphere, denote the latitudes in either hemisphere beyond which the sun drops completely below the horizon for at least one day during the winter.

The seasonal variability of insolation is the direct driver of the seasonal variability of temperature, but other factors modify temperature variations. Figure 1 illustrates the magnitude of the seasonal cycle in surface air temperature, expressed here as the annual range of climatological monthly mean temperature during the year. Values generally increase from the equator toward the poles, consistent with the greater seasonal cycle of solar insolation toward the poles. The seasonal cycle of temperature over the world oceans and equatorial continents is generally less than 5°C. Relative maxima are found in

extratropical continental interiors. The largest amplitudes, exceeding 40°C, are found across northern Canada and northeastern Asia, the northernmost continental regions far downwind from the moderating influence of oceans to the west. The seasonal cycle is smaller in amplitude in the Southern Hemisphere relative to the Northern Hemisphere. The phase of the seasonal cycle generally follows the seasonality of insolation, with maximum temperatures lagging the summer solstice of insolation by a few weeks to a few months.

The geographical variability of the amplitude of the seasonal cycle in temperature illustrates the strong thermal buffering effect of the oceans, whose capacity for storing heat (or thermal inertia) is several orders of magnitude greater than the heat capacity of land surfaces on the continents. The difference is due to two characteristics of bodies of water: the specific heat of liquid water is much larger than that of rock or soil, and vertical motions in the upper ocean distribute heat through a much deeper surface layer than occurs on continents. The oceans store huge amounts of heat in summer and release heat to the atmosphere in winter, damping the seasonal cycle of air temperature very considerably. The larger thermal inertia of the oceans relative to the continents also affects the phase of the seasonal cycle: temperatures over the oceans (or over coastal regions, especially along western coasts affected by westerly winds off the ocean) tend to reach their maximum seasonal value somewhat later in the year than at sites in continental interiors.

The seasonal cycle of precipitation is indirectly related to insolation variations via the seasonal variability of atmospheric temperature, and seasonally varying circulation also plays a major role in modulating precipitation. The moisture content of the air, which is strongly controlled by the Clausius–Clapeyron equation such that warmer near-surface temperatures allow for the possibility of larger vapor content, is one factor that leads to a tendency for the warm season to be the rainy season as well across much of the planet. Another factor is the increased instability of an air column that is warmed at the surface, where most of the increased solar insolation in

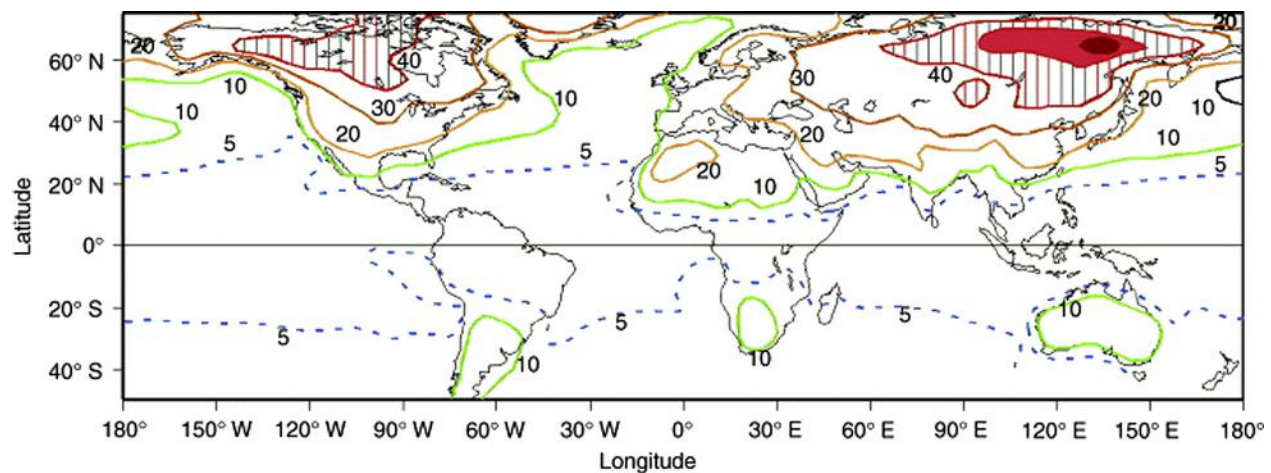


Figure 1 Annual range of monthly climatological surface temperature (°C) calculated from data for the period 1950–79. The annual range is less than 5°C equatorward of the blue dashed line; ranges of 10, 20, and 30°C are delineated by solid contours; regions where range exceeds 40, 50, and 60°C are shown by red hatching and successively darker shades of solid red. Adapted from Shea, D., 1986. *Climatological Atlas: 1950–1979*. Surface Air Temperature, Precipitation, Sea-Level Pressure, and Sea-Surface Temperature. NCAR, Boulder, CO. Technical Note NCAR/TN-269+STR.

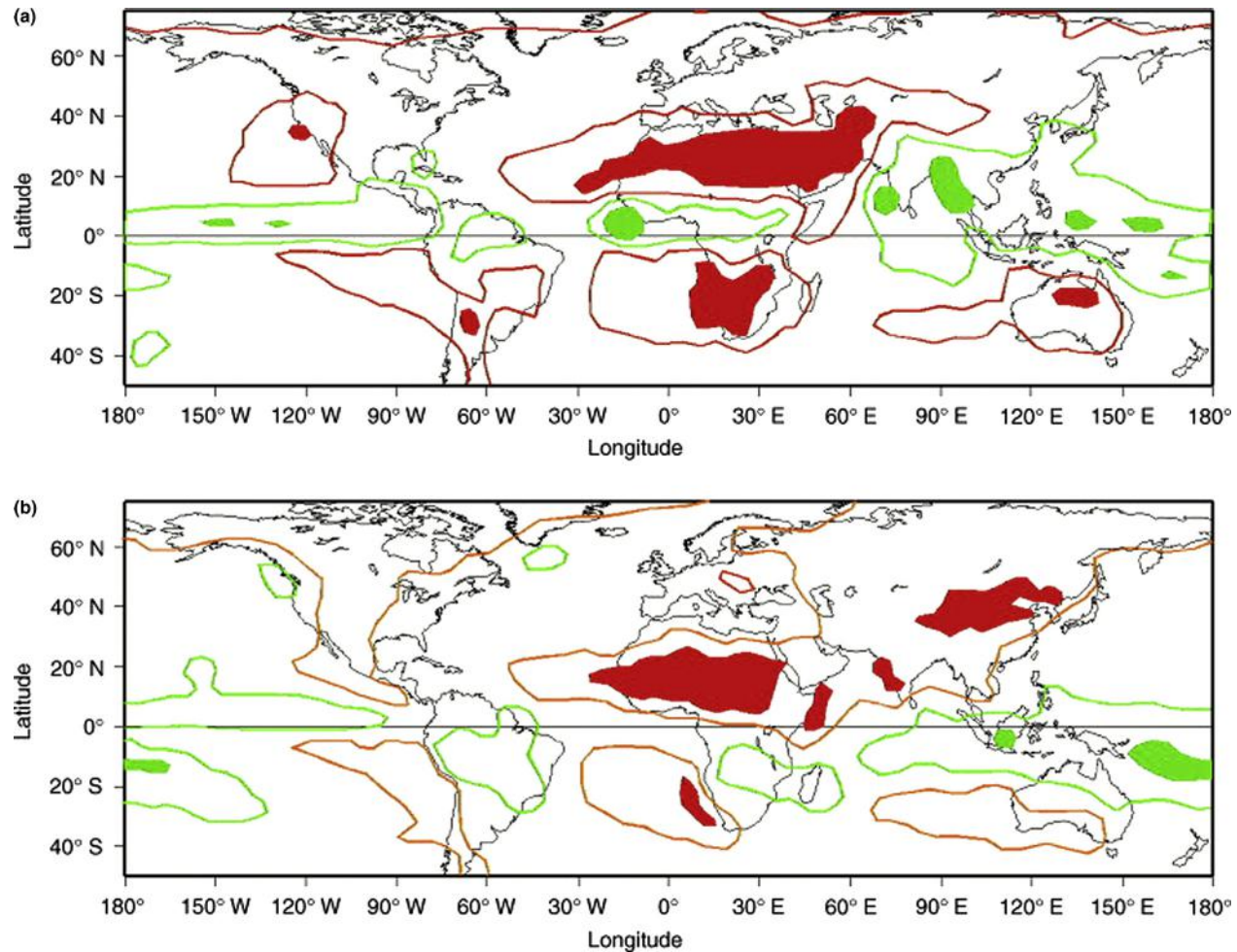


Figure 2 Seasonal climatological surface precipitation (mm) for (a) June–August and (b) December–February, calculated from data for the period 1950–79. Regions with seasonal precipitation < 10 mm are solid red; isohyets of 100 and 500 mm are yellow and green, respectively; and regions with precipitation > 1000 mm are solid green. Adapted from Shea, D., 1986. *Climatological Atlas: 1950–1979. Surface Air Temperature, Precipitation, Sea-Level Pressure, and Sea-Surface Temperature*. NCAR, Boulder, CO. Technical Note NCAR/TN-269+STR.

summer is absorbed. More unstable air promotes the formation of updrafts that cause clouds and precipitation.

Thus, the largest precipitation amounts tend to occur in tropical latitudes (Figure 2). But the largest seasonal cycle amplitudes of precipitation also generally occur in the tropics, despite modest insolation and temperature variability in those latitudes. Small seasonal variations of temperature and upward motion in the tropics can generate large seasonal changes in precipitation. Over tropical oceans, the Intertropical Convergence Zone (ITCZ) tends to shift northward and southward, following the maximum latitude of insolation. This progression causes pronounced wet and dry seasons in the tropics, even if the seasonal cycle of surface temperature is modest.

Tropical continents are often characterized by monsoonal climates, with a pronounced summer rainy season. Monsoons are driven by seasonally reversing continent–ocean temperature gradients. In summer, the lower heat capacity of continents raises the surface temperature above that of the ocean, creating a pressure gradient directed from ocean to continent that drives a moist onshore flow. As the ITCZ shifts poleward in the

summer hemisphere following the maximum latitude of insolation, its shift can be greatly enhanced by the presence of a tropical continent, generating the initiation of a summer monsoon circulation anchored over the continent.

The combination of abundant onshore transport of water vapor and conditionally unstable lapse rates over continents leads to a sharp summertime peak in rainfall, especially in regions where convective precipitation is enhanced by orographic lifting. These ingredients – proximity to a warm tropical ocean, strong continental heating in springtime, and large inland mountain ranges – are present in South Asia, where the world’s most intense monsoon circulation takes place and very large summer rainfall amounts occur (Figure 2(a)). In winter, the monsoon circulation reverses as the continent cools relative to the ocean, and the associated continental climate across South Asia is cold and dry (Figure 2(b)). Similar monsoonal seasonal transitions can be seen in Figure 2 in Africa and northern Australia. Monsoonal circulations also occur in southwestern North America and east of the Andes in South America, but these circulations are less pronounced.

The transition to the rainy season in a strongly monsoonal climate is generally quite abrupt, so an onset date can be defined and monitored on an annual basis. This date is typically very important for agriculture, water, and wildfire management and is identifiable in terms of vegetation greenness and other indices of ecosystem function. The monsoon onset date represents a climatological step change in response to smoothly varying insolation forcing and so is an example of a seasonal variation that is forced by the seasonal cycle in insolation modified by nonlinear processes that affect circulation dynamics. In contrast, the end of the rainy season in monsoonal climates tends to be gradual so that the annual demise of a summer monsoon is much more difficult to characterize than monsoon onset.

The seasonal cycle of precipitation over low-latitude oceans is strongly modulated by seasonal shifts in the Hadley Circulation, which maintains its near-equatorial (rising) branch in the summer hemisphere and its subtropical (sinking) branch in the winter hemisphere. Precipitation is strongly suppressed in the vicinity of the sinking branch, particularly near and to the east of the prominent oceanic subtropical highs in the summer season. Thus, in [Figure 2\(a\)](#), there are extensive regions of low precipitation (<100 mm) across the eastern subtropical Pacific and Atlantic Oceans (20–30°N).

Over the extratropical oceans and western continents, the phase of seasonal precipitation variations is the reverse of that of the tropics: winter tends to be the wet season. This is because summertime convective precipitation over these regions is strongly suppressed by subsidence over the summer subtropical highs. The winter season precipitation maximum results from frontal lifting associated with baroclinic wave activity, which is strongest in winter when temperature gradients are larger, especially in the winter ‘storm track’ regions on the poleward flanks of the principal tropospheric jet streams. Storm track-related boreal winter precipitation maxima, exceeding 500 mm in [Figure 2\(b\)](#), are observed over the northeastern Pacific Ocean and south of Greenland in the Atlantic Ocean.

However, indices of wintertime storminess in these regions do not rise and fall sinusoidally like insolation; instead, mid-ocean storminess in middle latitudes of both the Northern and Southern Hemispheres increases early in the winter and then generally goes through a pause in midwinter, a pattern of variability sometimes described as a ‘coreless winter,’ which projects onto the semiannual (twice-yearly) harmonic of the annual cycle. The midwinter pause in storminess seems to be caused by a nonlinear threshold in jet stream dynamics; several possible specific mechanisms have been proposed to explain this phenomenon.

Equatorward of the Tropics of Cancer and Capricorn, the sun moves directly overhead twice per year, as the latitude of maximum insolation shifts northward before the boreal summer solstice, and again as this latitude shifts southward before the austral summer solstice. Therefore, the seasonal cycle of insolation can drive a semiannual seasonal cycle in the tropics. A semiannual cycle is clearly observed over the tropical Indian and western Pacific Oceans.

At polar latitudes, the absence of insolation for a period of time centered on the winter solstice generates an asymmetric seasonal cycle in insolation. In response, very high latitudes experience a coreless winter in temperature, in which the

seasonal cycle of temperature (like insolation) tends to reach a flat minimum for weeks at a time in a climatological average. Summer temperatures near the poles, like insolation fluctuations, tend to rise and fall in a more sinusoidal manner, reaching an identifiable summertime peak.

In a few regions, the seasonal cycles of temperature and precipitation are determined by regional circulation systems, driven indirectly by insolation variations. An important example is the tropical eastern Pacific Ocean, where a ‘cold tongue’ in surface temperature develops right along the equator in the boreal autumn and weakens in springtime. The spatial and seasonal variations of precipitation are closely tied to ocean temperatures near the equator, with surface wind convergence (and hence precipitation) over the warmest water. Thus, spring is relatively warm and rainy, while autumn is colder and less rainy. The special ocean–atmosphere dynamics that create this unusual seasonal cycle are an intrinsic part of the El Niño cycle of interannual variability.

The geographical variations in seasonal cycles described here have been incorporated into numerous climate classification schemes, which are based in part on the amplitude and timing of seasonal fluctuations of temperature and precipitation. Classification systems are often used to relate different climatic zones to observed geographical variations in flora and fauna because plant species (for example) are typically adapted to limited ranges in temperature and precipitation. Once these associations between climate classifications and ecosystem variables are made, then the classifications can be used to characterize ancient climates where fossil evidence for certain plants or animals can be dated. The same climate–ecosystem relationships are used to formulate estimates of the long-term impacts on ecosystems associated with projected future climate changes.

Causes and Magnitude of Interannual Variability

Monthly and seasonal averages of temperature, precipitation, and other climatic variables are not the same year after year, although across most of the planet the amplitude of the seasonal cycle exceeds by far the deviations from seasonal averages that define interannual variability. [Figure 3](#) describes the interannual variability of Northern Hemisphere winter surface air temperature, expressed as the standard deviation of winter seasonal (December through February) averages. Note that this calculation includes all resolvable fluctuations, including any long-term trends or decadal variability that may be present during these three decades. Comparing this plot with [Figure 1](#), it is easy to see that the interannual variations in northern winter temperature are almost everywhere much smaller than the winter–summer seasonal difference in temperature. The largest values of interannual temperature variability occur over the subpolar continents, including maxima exceeding 5 °C in northwest North America and Siberia. Interannual fluctuations over the oceans are much smaller, rarely exceeding 1 °C.

Interannual temperature variability during other seasons is considerably smaller than in the boreal winter. The local interannual standard deviation of boreal summer temperature is generally less than 1 °C even over the continents. The pronounced winter maximum in interannual temperature variability results from two factors. First, planetary waves are most

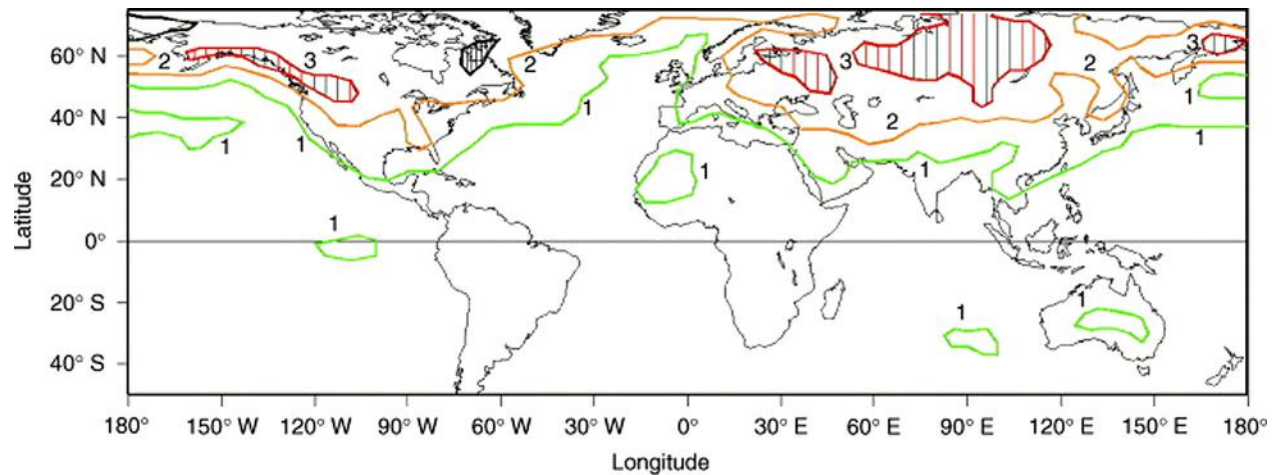


Figure 3 Interannual standard deviation of seasonal mean boreal winter surface temperature ($^{\circ}\text{C}$) calculated from data for the period 1950–79. Contours indicating 1 and 2°C are shown as solid green and orange; regions where standard deviation exceeds 3°C are hatched red. Adapted from Shea, D., 1986. Climatological Atlas: 1950–1979. Surface Air Temperature, Precipitation, Sea-Level Pressure, and Sea-Surface Temperature. NCAR, Boulder, CO. Technical Note NCAR/TN-269+STR.

pronounced and most variable during this season, so anomalous troughs and ridges are large during this season. Second, temperature gradients are large in the northern winter; in combination with the possibility of large anomalous troughs and ridges, these temperature gradients can support very large seasonal temperature advection anomalies in the boreal winter season.

Interannual variability of precipitation is largest during the seasons and over the regions where seasonal precipitation averages are large. Thus, the variability is generally larger across the continents in summer and over the oceans in winter. In regions of low-average precipitation, however, even small seasonal precipitation deficits can cause profound hardship on water-stressed environments and societies. In the tropics, interannual variability is large and can be comparable to the range of seasonal variations.

Interannual variability derives from several sources, including the following:

- **Limited sampling of weather:** Although 1 month or 1 season is generally longer than the lifetime of an individual synoptic weather system, there are entirely random differences in the average weather from 1 year to the next associated with the limited sampling of weather during 1 month or 1 season. Such ‘weather noise’ is quite pronounced in precipitation statistics in arid climates, where just one or two intense precipitation episodes (or their absence) can significantly affect the average for an entire month or season.
- **Internal atmospheric variability:** The atmosphere contains internal modes of dynamical variability that are considerably longer lived than individual weather systems. A prominent example is the stratospheric quasi-biennial oscillation (QBO) in equatorial zonal winds, which results from upward propagating planetary waves forced randomly from below in the troposphere. Equatorial stratospheric wave dynamics set the time scale of the QBO, rather than the time scale of the forcing from below. Long-lived internal variability is generally less pronounced in the troposphere

relative to synoptic-scale weather fluctuations. However, some long-lived fluctuations (such as ‘blocking ridges,’ which are persistent regions of high pressure that perturb the middle latitude westerly flow) project significantly onto interannual variability.

- **Ocean–atmosphere interactions:** Nonseasonal oceanic variability occurs over time scales longer than atmospheric weather. Large-scale waves on the thermocline propagate across ocean basins over months or even years. Surface temperature perturbations associated with these waves act as short-term climatic forcing for the overlying atmosphere by modifying the fluxes of latent and sensible heat into the atmosphere, thereby modulating near-surface air temperature, pressure, and humidity. These atmospheric anomalies in turn affect the distribution of vertical motion and precipitation so that the atmospheric response to slowly varying ocean surface temperature anomalies extends through the troposphere. Large-scale atmospheric dynamics worldwide are thereby affected by large oceanic temperature anomalies.

In particular, much of the interannual variability across large regions of Earth has been shown to correlate with fluctuations of the coupled ocean–atmosphere system known as the El Niño Southern Oscillation (ENSO). ENSO extrema generally are linked to the seasonal cycle, with the largest equatorial Pacific sea surface temperature anomalies occurring in the boreal autumn and winter seasons, when the cold tongue is normally most pronounced. Extrema in the ENSO cycle include huge pools of anomalous warm (El Niño) or cold (La Niña) surface water, extending along the equator from the South American coast westward to the dateline (fully one-quarter of the circumference of Earth). Such a gigantic anomaly in ocean surface temperatures profoundly affects the large-scale distribution of deep convective precipitation across the Pacific, which in turn perturbs the Hadley Circulation, the subtropical jet stream across the Pacific, and the storm tracks

and planetary waves in the Western Hemisphere. Thus, the atmosphere transmits the effects of equatorial Pacific Ocean surface temperatures across much of the globe.

- **Volcanic eruptions:** Occasional explosive volcanic eruptions inject large quantities of aerosols into the stratosphere. In addition to the profound effects of volcanic plumes on the local climate, which tend to decay quickly once the eruption episode is over, the stratospheric aerosols can remain in circulation for several years after the eruption, and their effects are distributed by the stratospheric circulation across much of the planet. Global cooling is generally observed during the period of time that volcanic aerosol plumes are resident in the stratosphere, although some volcanic aerosols also absorb outgoing longwave radiation and can thereby increase the greenhouse effect and act to warm the surface.
- **Solar fluctuations:** The brightness of the sun is not constant, although on the day-to-day time scale associated with weather, insolation is nearly constant. Therefore, solar brightness changes are not considered as part of a daily weather forecast. From year to year, however, the sun's brightness does change, on the order of 0.1% or less. The sun undergoes a nearly regular cycle of brightness of about 11 years, associated with variations in sunspot numbers, which is pronounced enough to cause measurable 11-year cycles in temperature. The brightness of the sun also varies on longer time scales for reasons that are poorly understood but can affect long-term climate.
- **Persistent land surface anomalies:** Land surface anomalies, such as soil moisture fluctuations, vegetation changes, or interannual variations in snowpack, can have effects on the overlying atmosphere analogous to ocean surface temperature anomalies. As described in this article's discussion of the seasonal cycle, however, the heat storage capacity of land surfaces is much smaller than upper ocean heat storage, so land surface anomalies are thought to play a smaller role than large-scale ocean anomalies in forcing atmospheric interannual variability. Nevertheless, modeling studies have shown that prescribed large-scale anomalies in land surface characteristics can affect climate significantly. In particular, lengthy episodes of drought in the interior of a continent are exacerbated by a positive feedback on the energy and hydrologic cycles from the land surface: lack of rainfall leads to dry soil and sparse vegetation, which in turn causes decreases in evaporation and moist static energy in the atmospheric boundary layer. These deficits in near-surface water vapor and energy tend to suppress convective precipitation in the summer, acting to perpetuate drought.

On longer time scales, seasonal and interannual variability may be loosely distinguished from longer time scale (decadal–centennial) fluctuations by the relative importance of the climatic processes involved. In particular, a considerable fraction of the forcing for seasonal and interannual variability can be described by focusing on the upper ocean (above the thermocline). The atmospheric interannual variability associated with continental land surface anomalies is not well understood but is generally thought to be less important than oceanic

forcing because the heat capacity of land surfaces is so much smaller. Explosive volcanic eruptions can affect global climate strongly for a year or two following the eruption, after which most aerosols drop out of the stratosphere. A succession of explosive eruptions would be required to strongly affect the climate for a longer period of time. The abyssal oceanic circulation, including the thermohaline circulation, may play a much more prominent role at longer (decade–century) time scales. The same is true for solar brightness fluctuations. Long-term changes in vegetation patterns, and the relatively slow changes in atmospheric composition, are very important on decade–century time scales but may also modulate interannual variability.

Diagnosis and Prediction of Interannual Variability

Interannual variability does not occur randomly across the planet. Preferred large-scale patterns of circulation anomalies, called teleconnection patterns, have been identified via statistical analysis of the observed interannual variability of sea level pressure, zonal wind, or geopotential heights. These teleconnection patterns tend to be most pronounced in the boreal winter season when planetary wave energy is maximized. The existence of such patterns allows climatologists to describe a large fraction of interannual variability using just a few indices of the anomalous circulation associated with teleconnection patterns. The existence of such patterns was recognized early in the twentieth century by visionary climate data analysts such as Sir Gilbert Walker, but the causes of the patterns and their potential predictability have been the subject of very active recent research.

Among the most prominent teleconnection patterns are the Southern Oscillation over the tropical Pacific (now known to be associated with the equatorial ocean surface temperature anomalies of El Niño), the North Atlantic Oscillation (NAO) of heights and temperatures, and the Pacific–North American pattern of geopotential heights, each of which was recognized by Walker in the 1920s. Seasonal surface temperature and precipitation anomalies are associated with the planetary wave perturbations that characterize these patterns. In the Southern Hemisphere, geopotential heights are observed to vary in a characteristic wavenumber-3 pattern anchored to the distribution of continents and oceans in middle latitudes.

Another class of teleconnection patterns, the zonally symmetric ‘annular modes,’ was also recognized decades ago. Analysis of new global data sets has reconfirmed the utility of a zonally symmetric description of interannual variability, so these modes have been the subject of renewed research in recent years. Annular modes can be identified as fluctuations in the zonally averaged zonal wind in the extratropics of either hemisphere. Alternatively, via the geostrophic wind relationship, annular modes appear as fluctuations in the north–south gradient of geopotential height between the extratropics (around 45° latitude) and high latitudes (around 80°) in either hemisphere.

Various names have been given to these modes of variability over the years. Fluctuations in zonal wind averaged around latitude circles were referred to as the ‘index cycle’ in the

mid-twentieth century. More recently, modes of variability describing much of the same variance have been called the Arctic and Antarctic Oscillation, or the Northern and Southern Annular Modes.

The Northern Annular Mode (NAM) and the NAO present different conceptual paradigms but describe similar sets of interannual surface temperature and precipitation anomalies in the extratropics. Defined as an annular mode, the NAM and its Southern Hemisphere counterpart exhibit pronounced coherent structure extending up into the stratosphere. The NAO, defined early on in terms of surface temperature variability, has naturally been diagnosed in terms of ocean–atmosphere interaction. But using either paradigm, studies of the possible cause of this general package of variability have suggested that interannual fluctuations of the NAM and NAO are probably manifestations of internal atmospheric variability, rather than forced by external causes that could be predictable in advance. Thus, at the present level of understanding, interannual variability associated with the NAM and NAO could serve to limit the overall predictability of interannual variability in extra-tropical latitudes.

Drought episodes, defined as prolonged periods of unusually dry weather, are another form of interannual variability with profound consequences for ecosystems and for people. Droughts are usually defined over continental regions, and they can be characterized in terms of time scale or, in a complementary way, in terms of the impacts of prolonged precipitation deficits on different components of the land–atmosphere system. Thus, ‘meteorological drought’ is a deficit of precipitation from a long-term average, and it can be defined for either a short period (days to weeks) or a much longer period (decades). Prolonged meteorological drought can affect vegetation, crops, and soil moisture (‘agricultural drought’), and stream flows and lake levels (‘hydrological drought’).

Similar to other aspects of interannual variability related to teleconnection patterns, persistent anomalies in ocean surface temperatures have been examined to explain the possible cause of atmospheric circulation changes that can cause long-term drought. North American droughts have been diagnosed in this way, with multiyear episodes of cold equatorial Pacific Ocean temperature anomalies identified as potentially causing severe North American drought. Other tropical ocean temperature anomaly patterns seem to contribute to atmospheric circulation changes conducive to long-term drought as well.

In order to monitor interannual variability, the major weather services of the world calculate monthly and seasonal averages of temperature, precipitation, and geopotential height at the end of each month. Indices of the principal teleconnection patterns and ocean temperatures in the tropical Pacific are a particular focus of real-time monitoring. Drought indices, which can be based on different combinations of precipitation, soil moisture, and vegetation status indicators, are particularly closely monitored and disseminated to the public. Areas of the planet experiencing extreme, persistent anomalies of temperature and precipitation are also monitored closely.

If any of the causes of interannual variability listed here persist longer than a season (or if the causes themselves could

be predicted a season or more in advance), and if the magnitude of interannual variability forced thereby is large compared to unpredictable interannual variability such as weather noise, then in principle it should be possible to make skillful predictions of interannual variability at lead times extending beyond the deterministic limit of a weather forecast. Such seasonal forecasts of short-term climate anomalies cannot predict the evolution of individual weather systems and thus are incapable of predicting the weather on specific days. The goal instead is to forecast how the average weather for a month or a season will differ from the expected value derived from a straightforward climatological average.

Seasonal hurricane forecasts for the northwest Atlantic region, for example, are based on knowledge of the ENSO cycle, Atlantic Ocean temperature anomalies, the phase of the QBO in tropical stratospheric winds, and several other factors that can be observed prior to the start of the hurricane season. At the beginning of the North Atlantic hurricane season (in late summer), forecasts are issued for the number of tropical cyclones that are expected to make landfall in North America based on historical associations between the factors listed here and subsequent hurricane activity. It should be emphasized that no attempt is made to forecast the timing or tracks of individual cyclones, just the aggregate hurricane activity for the entire season.

Interannual variability of step changes in the seasonal cycle, such as an early or late monsoon onset date, can also potentially be predicted based on antecedent forcing. The ENSO cycle, for example, has been shown to have a statistical relationship with onset dates for the summer monsoon in South Asia and southwestern North America. Improving forecast skill for such step changes has proven to be a difficult challenge, at least in part because step changes occur on the fast time scale of weather.

Active operational prediction efforts are aimed at producing skillful and useful estimates of the probabilities of seasonal climate anomalies of temperature and precipitation up to about a year in advance. At present, the principal well-demonstrated source of seasonal climate predictability is the ENSO cycle, which has been shown to correlate with seasonal climate anomalies in many regions throughout the world. Ocean temperatures and surface winds across the tropical Pacific are now observable in real time thanks to the development of an extensive monitoring system that combines in situ and satellite observations. Climate forecasters know that if the equatorial Pacific Ocean and atmosphere develop a significant cold or warm anomaly in the boreal autumn, then it is highly likely that such anomalous conditions will persist through the following winter season. Empirical prediction techniques, based largely on the historical occurrence of systematic seasonal anomalies of temperature or precipitation in previous ENSO-extreme winters, have demonstrated useful skill. More recently, dynamical models driven by ENSO-related ocean temperature anomalies have also demonstrated seasonal prediction skill. The development of ENSO-based seasonal forecast skill using dynamical models has raised optimism that other sources of potential predictability on the seasonal time scale could be explored and exploited using models.

See also: **Aerosols:** Role in Climate Change. **Air Sea Interactions:** Sea Surface Temperature. **Climate and Climate Change:** Climate Prediction: Empirical and Numerical; Climate Variability: Decadal to Centennial Variability; Climate Variability: North Atlantic and Arctic Oscillation; Volcanoes: Role in Climate. **Clouds and Fog:** Contrails. **Data Assimilation and Predictability:** Predictability and Chaos. **Dynamical Meteorology:** Kelvin Waves. **General Circulation of the Atmosphere:** Teleconnections. **Hydrology, Floods and Droughts:** Drought; Palmer Drought Severity Index. **Land-Atmosphere Interactions:** Overview. **Middle Atmosphere:** Quasi-Biennial Oscillation. **Radiation Transfer in the Atmosphere:** Radiation, Solar. **Tropical Meteorology and Climate:** El Niño and the Southern Oscillation: Observation; El Niño and the Southern Oscillation: Theory; Hadley Circulation; Monsoon: Overview; Tropical Climates; Walker Circulation. **Weather Forecasting:** Seasonal and Interannual Weather Prediction.

Further Reading

- Bonan, G.B., 2002. *Ecological Climatology*. Cambridge University Press, Cambridge.
- Hartmann, D.L., 1994. *Global Physical Climatology*. Academic Press, San Diego.
- Peixoto, J.P., Oort, A.H., 1992. *Physics of Climate*. American Institute of Physics, New York.
- Philander, S.G.H., 1990. *El Niño, La Niña, and the Southern Oscillation*. Academic Press, San Diego.
- Schubert, S., Gutzler, D., Wang, H., et al., 2009. A U.S. CLIVAR project to assess and compare the responses of global climate models to drought-related SST forcing patterns: overview and results. *Journal of Climate*. 22, 5251–5272.
- Shea, D., 1986. *Climatological Atlas: 1950–1979*. Surface Air Temperature, Precipitation, Sea-Level Pressure, and Sea-Surface Temperature. NCAR, Boulder, CO. Technical Note NCAR/TN-269+STR.
- Thompson, D.W.J., Wallace, J.M., 2000. Annular modes in the extratropical circulation. Part I: month-to-month variability. *Journal of Climate*. 13, 1000–1016.
- Trenberth, K.E., 1983. What are the seasons? *Bulletin of the American Meteorological Society*. 64, 1276–1282.
- van den Dool, H., 2007. *Empirical Methods in Short-Term Climate Prediction*. Oxford University Press, Oxford.

Energy Balance Climate Models

GR North, Texas A&M University, College Station, TX, USA
K-Y Kim, Seoul National University, Seoul, Korea

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

Energy balance climate models are a class of simplified models of the large-scale climate based upon the energy balance of a column of air/ocean above/below an infinitesimal box at the surface. When integrated over the entire surface area they are the global average models. Zonal average models can be constructed by integrating around latitude circles. Some of the more simplified models can be solved analytically providing special insights to such aspects of climate at its sensitivity to small perturbations of solar brightness or greenhouse gases.

Introduction

Climate models come in a variety of sizes and degrees of complexity. The global energy balance models are at the simplest level of description of the global climate system. These models represent the entire climate system with a single number, the global average temperature. The coupled ocean/atmosphere general circulation models that incorporate the detailed circulation of the atmosphere and the oceans as well as other components, altogether employing millions of numbers to specify the state of the climate, are at the other end of the spectrum of models. The purpose of these models is to gain a better understanding of the climate system, why its current state is what it is, how sensitive it is to external perturbations, and how it compares to the observations in detail. This article considers the models at the very low end of the so-called hierarchy of climate models, the energy balance climate models (EBCMs). Most of these models use the surface temperature field of the Earth as the basic indicator of the climate. The physical principle that constrains or determines this climate is the rate of incoming solar radiation absorbed by a column of air over a particular small box at the surface being balanced by the rate of release of radiation to space from the same column added to the net flux of heat leaving the column to enter surrounding columns. The article will introduce the global models, then proceed to models with latitude dependence and finally to both latitude and longitude dependence.

Globally Averaged EBCMs

The surface of a material body emits electromagnetic radiation proportionally to the fourth power of the Kelvin temperature. The radiation is bunched in wavelengths peaked at a value inversely proportional to the wavelength. This peak is in the visible part of the spectrum (~ 500 nm) for the Sun's radiation and in the infrared (~ 15 μ m) for the Earth's radiation. Over the range of values important to climate, the infrared radiation energy flux (Watts per meter squared = W m^{-2}) can be written approximately as

$$I_0(T) = A_0 + B_0 T \quad [1]$$

where $A_0 = 314.9 \text{ W m}^{-2}$ and $B_0 = 4.61 \text{ W m}^{-2} (\text{°C})^{-1}$ and T is the surface temperature in degrees Celsius. If the Earth had no

atmosphere, the planet would radiate the amount given by the formula for each square meter of surface area (on global averaging). The Sun supplies energy mostly in the visible range such that a square meter receives energy per second perpendicular to the rays from the Sun of about $S_0 \approx 1360.8 \pm 0.5 \text{ W m}^{-2}$ (Kopp and Lean, 2011). The balance of rates can be written as

$$4\pi R^2(A_0 + B_0 T) = \pi R^2 S_0 a_p \quad [2]$$

where R is the radius of the Earth (which along with π cancel out) and a_p is the fraction of sunlight absorbed from space by the Earth system (the planetary co-albedo) – its current value is taken to be 0.70. Inserting the numerical values, it can be computed that $T = -16.7 \text{ °C}$, a value over 30 °C too cold (given the present co-albedo). Satellite data can be used to get estimates of the coefficients (Graves et al., 1993) in the formula for $I(T)$ by using the data from different locations and seasons for T . The approximate results are $A = 206 \text{ W m}^{-2}$ and $B = 2.2 \text{ W m}^{-2} (\text{°C})^{-1}$ (actually, these values are slightly uncertain because of the differences in their values for clear sky versus cloudy skies; values that are consistent with the correct temperature for the global average and within the range of error for the observations are chosen). Using these empirical coefficients yields $T = 14.5 \text{ °C}$, a value very close to that observed. The atmosphere contains gases that absorb in the infrared range then reradiate at colder temperatures at roughly 5 km above the ground. This so-called greenhouse effect is the reason for the warmer planetary surface. Next one can ask how much the temperature would change if the Sun's output were to increase by 1%. The answer is approximately 1.08 °C , which can be calculated by incrementing eqn [2]. The next question is: How much does the equilibrium temperature change if the CO_2 concentration is doubled? This was estimated from the reduction of outgoing radiation caused by such an action. The change in radiation balance at the top of the atmosphere is about 4.0 W m^{-2} (Myhre et al., 1998), based on accurate radiative transfer calculations. Inserting the negative of this value on the right-hand side of eqn [2] gives an increase of T by 1.82 °C . This latter number gives the sensitivity of the global average temperature to a doubling of CO_2 for the global average EBCMs when the empirical formula for $I(T)$ is used, and it is assumed that there are no changes in the planetary albedo during the doubling experiment. By the way, if there were no atmosphere (A_0 and B_0 apply) then the sensitivity to

raising the solar constant by 1% leads to only half the sensitivity as in the current atmosphere. Evidently, the presence of the atmosphere and its effect on outgoing radiation amplifies the sensitivity by about a factor of 2. This increase of sensitivity is due to the feedbacks in the atmospheric system – most of the effect coming from water vapor increases as the planet is warmed, since water vapor itself is a greenhouse gas and hence the strong positive feedback. The water vapor feedback mechanism along with some others such as the so-called lapse rate feedback are thought to be captured in the coefficient B . There can be other feedbacks in the climate system not included in the simple balance that is used, and the nature of these additional feedbacks is a very active research area in climate dynamics. Coupled ocean/atmosphere general circulation models yield a range of sensitivities for doubling CO_2 centered at 3.0 ± 1.0 °C.

One-Dimensional EBCMs

One-dimensional EBCMs were introduced by [Budyko \(1968\)](#) and [Sellers \(1969\)](#). In this case, the zonal averages are considered (averages around latitude belts), equivalently, and it was imagined that the planet to be zonally symmetric. Sunlight is not distributed evenly from one latitude belt to another: more solar energy per unit area strikes the tropics than the polar areas. If the seasonal distributions are averaged out as well an approximate formula can be obtained for the normalized distribution of sunlight as a function of latitude:

$$S(x) = 1 + S_2 P_2(x); \quad \text{where} \quad P_2(x) = \frac{1}{2}(3x^2 - 1); \quad [3]$$

$$\int_0^1 S(x) dx = 1$$

where x is the sine of latitude (or cosine of the polar angle θ), which varies from 0 at the equator to unity at the North Pole, $S_2 = -0.482$, and $P_2(x)$ is the second Legendre polynomial. One advantage of using x as the measure of latitude is that equal increments of x contain equal areas around a latitude belt on the sphere. For a given latitude belt dx , there will be an absorbed incoming flux of sunlight (letting $Q \equiv S_0/4$) equal to $Qa(x)$, where $a(x)$ is the absorptivity (or co-albedo) of the Earth-atmosphere system at sine of latitude x ; the outgoing infrared radiation from the belt may be written as $A + BT(x)$. To close the budget, the net rate of heat that flows into/from the belt per unit area from/into belts on either side needs to be considered.

Satellite observations show ([Graves et al., 1993](#)) that the co-albedo $a(x)$ decreases with latitude because of the zenith angle effect (solar rays from near the horizon tend to be reflected more by the Earth-atmosphere system). The empirical result is

$$a(x) = a_0 + a_2 P_2(x) \quad [4]$$

with $a_0 = 0.679$ and $a_2 = -0.241$ ([Graves et al., 1993](#)). The product of $S(x)$ and $a(x)$ leads to

$$S(x)a(x) \approx 0.702 - 0.535 P_2(x) + \dots \quad [5]$$

after truncating the Legendre polynomial series at the second degree (neglecting a small $P_4(x)$ term that results from the product of the two $P_2(x)$ terms).

The simple model is adopted where the heat flux is proportional to the negative gradient of the temperature field. In spherical coordinates, this can be expressed as

$$\begin{aligned} \text{poleward heat flux} \equiv F_\theta &= -D \frac{1}{R} \frac{\partial T}{\partial \theta} = -D \frac{1}{R} \frac{\partial T}{\partial x} \frac{\partial x}{\partial \theta} \\ &= -D \frac{1}{R} \sqrt{1-x^2} \frac{dT}{dx} \end{aligned} \quad [6]$$

where θ is the polar angle and R is the Earth's radius. The heat lost per unit area in the thin belt is the divergence taken in spherical coordinates, $\frac{1}{R \cos \theta} \frac{\partial \cos \theta F_\theta}{\partial \theta}$. The two factors of R can be absorbed into the definition of D or what is equivalent, setting R to unity.

With this parameterization, the energy balance equation becomes:

$$-\frac{d}{dx} \left(D(1-x^2) \frac{dT}{dx} \right) T + A + BT = Q((Sa)_0 + (Sa)_2 P_2(x)) \quad [7]$$

with $(Sa)_0 = 0.702$, $(Sa)_2 = -0.535$. In general D , A , and B might depend on x and in more complicated models they might depend on $T(x)$ itself. In the case where these phenomenological coefficients are constant, the two-mode solution is the exact solution. By this, it was meant that

$$T(x) = T_0 + T_2 P_2(x); \quad T_0 = \frac{Q(Sa)_0 - A}{B}; \quad T_2 = \frac{Q(Sa)_2}{B + 6D} \quad [8]$$

is the exact solution. Using the observed values $T_0 = 14.5$ °C and $T_2 = -28.0$ °C, one can solve to find that a better value of A is 206 W m^{-2} and $D = 0.724 \text{ W m}^{-2} (\text{°C})^{-1}$. [Figure 1](#) shows that the parabolic fit in x is excellent except near the equator where diffusive heat transport is a poor approximation. [Figure 2](#) shows the total heat flux toward the North Pole as a function of x :

$$\begin{aligned} [\text{total heat flux crossing latitude circle at } x] \\ = -6\pi T_2 D R^2 x(1-x^2) \end{aligned} \quad [9]$$

[Figure 2](#) also shows points from observations for the year 1988 ([Trenberth and Solomon, 1994](#)). The significance of [Figures 1 and 2](#) is that even though there is some arbitrariness in the parameters, the EBCM captures the shape of the curves almost exactly (quadratic in [Figure 1](#) and $x(1-x^2)$ in [Figure 2](#)).

Time Dependence

The EBCM can include time dependence by introducing a heat storage term into the energy balance. Here the problem is complicated by the fact that the effective heat capacity over land area is surely very different from that over ocean surfaces. Consider first the case of a uniform planet with an effective heat capacity per unit area of C . Also stick to the uniform zonal average. The energy balance can be written as

$$\begin{aligned} C \frac{\partial T'}{\partial t} - D \frac{d}{dx} \left((1-x^2) \frac{dT'}{dx} \right) T' + BT' &= F(x, t) \\ &= F_0(t) + F_2(t) P_2(x) \end{aligned} \quad [10]$$

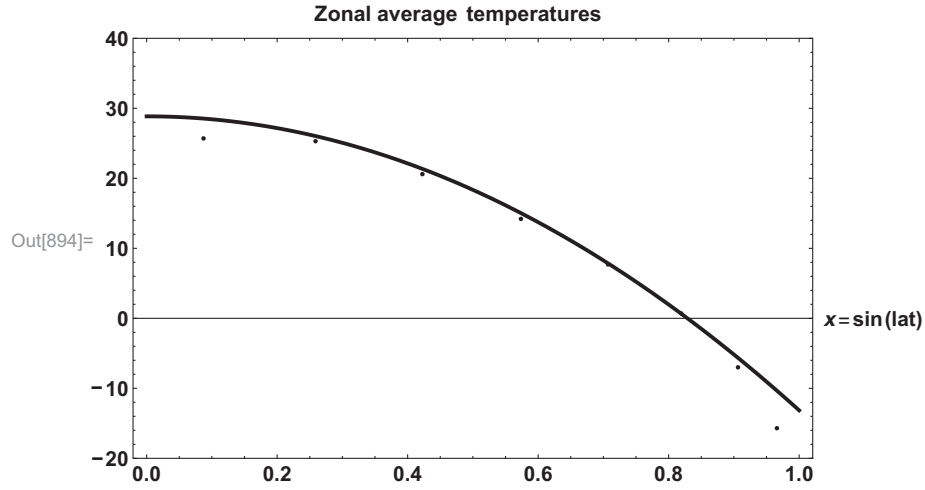


Figure 1 Temperature ($^{\circ}\text{C}$) versus the $x = \sin(\text{latitude})$ (equator is 0, pole is 1). The solid line is the model fit with parameters $S = 1360 \text{ W m}^{-2}$, $A = 206 \text{ W m}^{-2}$, $B = 2.20 \text{ W m}^{-2} (^{\circ}\text{C})^{-1}$, and $D = 0.72 \text{ W m}^{-2}$.

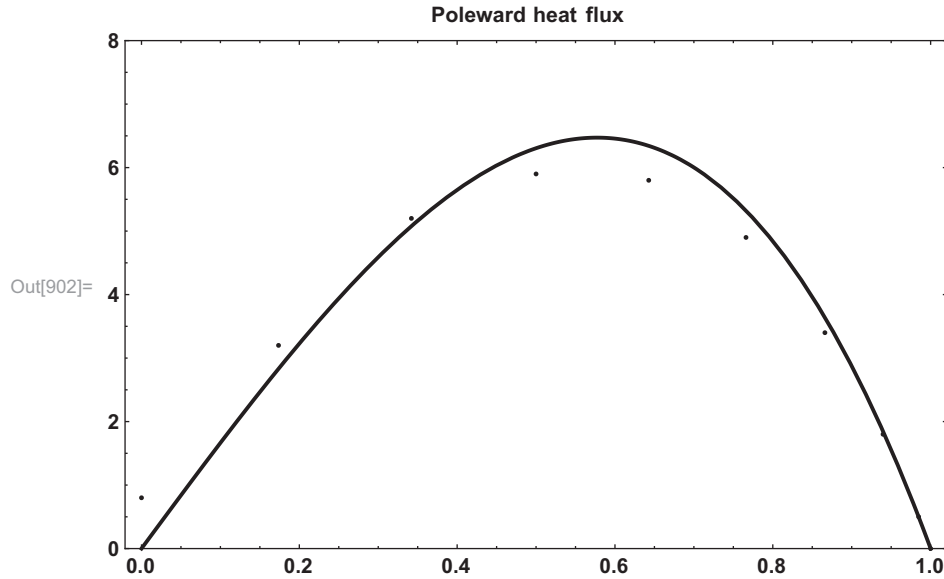


Figure 2 Heat flux across a latitude circle at $x = \sin(\text{latitude})$ toward the pole in petawatts (10^{15} W).

where $T'(x, t)$ is the departure from the steady-state solution of the last section, and $F(x, t)$ is a space-time-dependent 'forcing' containing only x dependence up to two Legendre modes.

If $F(x, t)$ is set suddenly to 0, there is an exponential decay to the steady-state solution in each Legendre mode with time constants $\tau_0 = C/B$ and $\tau_2 = \tau_0 / \left(1 + 6 \frac{D}{B}\right)$. If the planet is all land, the effective heat capacity might be taken to be that of about half the mass of the atmosphere, leading to a decay time of about a month for the global average (T'_0), a value reminiscent of the results from detailed radiative convective models (Manabe and Strickler, 1964), and about 9 days for T'_2 . If the planet is all ocean, the mixed layer of the ocean should be used

for the heat capacity. This leads to a global decay time of a few years. For long-time scales, the deeper ocean has to be taken into account, leading to much longer adjustment times.

Further Extensions and Applications

In this short introduction to EBCMs, only its simplest linear members have been introduced in an effort to show how these models can be used to develop intuition about the real climate system. The above treatment can be extended to two dimensions (2D) on the globe but must take into account the geography in the $C(\hat{r})$ (large over ocean, small over land) term where $\hat{r}$ denotes the location on the Earth's surface. This can be

done, but the solutions require numerical methods and the storage capabilities of a modern computer. The most important application of the 2D case is to the seasonal cycle. The 2D EBCM captures the geographical morphology of the surface thermal field of the seasonal cycle with remarkable fidelity because the amplitude and phase of the field are dominated by the contrasting effective heat capacities of the two surfaces (North et al., 1983).

Another important application is to the larger scale response to the small-scale drumming of weather fluctuations which in this model are represented as white noise in space and time for the driving term $F(\hat{r}, t)$ (Kim and North, 1991). Once again, the geographical pattern of climate variability is faithfully represented even in different frequency bands from months to decades. The 2D EBCMs have proven useful in a number of paleoclimate applications. Two-dimensional models have also been used in the detection and attribution of faint climate signals. These and many other examples can be found in publications by the authors.

Nonlinear terms can be included in the EBCM. For example, the co-albedo can be modified to include the ice-albedo feedback mechanism. One way this can be accomplished is by allowing an ice or snow line to be tied to a particular critical temperature. Budyko (1968) suggested that if the mean annual temperature is below -10°C , then there will be perennial ice poleward of this latitude with its smaller co-albedo. One can also tie the seasonal snow line to the freezing point. One might also employ nonlinear infrared radiation effects as well as in the horizontal transport of heat. The ice-albedo effect has been most interesting in that the nonlinearity induces more than one solution for a given set of conditions such as a particular value of the solar constant. These model solutions can feature jumps from one branch of the solution to another for a continuous change in internal parameters (North et al., 1981).

EBCMs can also be attached to one (vertical)-dimensional ocean models. One example is the upwelling-diffusion ocean model which for global average has been used by Wigley and Raper (2002) and in two horizontal dimensions by Kim et al. (1992) in transient climate scenarios.

See also: Climate and Climate Change: Climate Feedbacks.
Global Change: Climate Record: Surface Temperature Trends.

Bibliography

- Budyko, M.I., 1968. Effect of solar radiation variations on the climate of the earth. *Tellus* 14, 611–619.
- Graves, C.E., Lee, W.-H., North, G.R., 1993. New parameterizations and sensitivities for simple climate models. *Journal of Geophysical Research* 98, 5025–5036.
- Kim, K.-Y., North, G.R., 1991. Surface temperature fluctuations in a simple stochastic climate model. *Journal of Geophysical Research* 96, 18573–18580.
- Kim, K.-Y., North, G.R., Huang, J., 1992. On the transient response of a simple coupled climate system. *Journal of Geophysical Research* 97, 10069–10081.
- Kopp, G., Lean, J.L., 2011. A new, lower value of total solar irradiance: evidence and climate significance. *Geophysical Research Letters* 38, L01706. <http://dx.doi.org/10.1029/2010GL045777>.
- Manabe, S., Strickler, R.F., 1964. Thermal equilibrium of the atmosphere with a convective adjustment. *Journal of Atmospheric Science* 21, 361–385.
- Myhre, G., Highwood, E.J., Shine, K.P., 1998. New estimates of radiative forcing due to well mixed greenhouse gases. *Journal of Geophysical Research* 25, 2715–2718.
- North, G.R., Cahalan, R.F., Coakley, J.A., 1981. Energy-balance climate models. *Rev. Geophys. Space Physics* 19, 91–121.
- North, G.R., Mengel, J.G., Short, D.A., 1983. A simple energy balance model resolving the seasons and the continents: application to the milankovitch theory of the ice ages. *Journal of Geophysical Research* 88, 6576–6586.
- Sellers, W.D., 1969. A climate model based on the energy balance of the earth-atmosphere system. *Journal of Applied Meteorology* 8, 392–400.
- Trenberth, K.E., Solomon, A., 1994. The global heat balance: heat transports in the atmosphere and ocean. *Climate Dynamics* 10, 107–134.
- Wigley, T.M.L., Raper, S.C.B., 2002. Reasons for larger warming projections in the IPCC third assessment report. *Journal of Climate* 15, 2945–2952.

Global Impacts of the Madden–Julian Oscillation

C Zhang, University of Miami, Miami, FL, USA

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

As its convection center moves from the Indian Ocean to the Pacific Ocean, the Madden–Julian Oscillation affects global weather, climate, and many other phenomena in the Earth system.

Introduction

The Madden–Julian Oscillation (MJO) affects global weather and climate mainly through two mechanisms. One is direct by enhancing or suppressing tropical convection and rainfall along its propagation pathway from the Indian Ocean to the Pacific Ocean. The other is indirect by generating atmospheric or oceanic wave disturbances that propagate into regions with little or no direct signals of the MJO itself. MJO's influences on weather and climate are not separated from each other. The degree to which global weather is affected by the MJO often depends on the phases of certain climate modes. Climate phenomena under the influence of the MJO in turn modulate weather events in many regions of the world. In addition, the MJO also affects many important and interesting phenomena in the atmosphere and ocean that are not commonly categorized as either weather or climate. Because of the large number of weather, climate, and other phenomena in the Earth system under the influence of the MJO, only very brief descriptions of them are possible here. The listed further reading materials provide more detailed discussion of the global impact of the MJO.

Severe Weather

Severe weather around the world occurs with or without the MJO. However, the strength, frequency, and spatial and temporal distributions of severe weather can be affected by the MJO. More precisely, these characteristics of severe weather may change depending on the location of the MJO convection center and periods with and without active MJO events.

Tropical Cyclones

Effects of the MJO on Tropical Cyclone (TC) frequency can be visualized from [Figure 1](#), which shows TC tracks and rainfall anomalies in the global tropics at four typical stages of the MJO. The density of the tracks illustrates the TC frequencies. They vary with the longitudinal location of the MJO convection center. Over the southern Indian Ocean, TC frequency is highest when MJO convection center is over the Indian Ocean, but lowest when it is in the Western Hemisphere. Over the Bay of Bengal, TCs are the most frequent when the MJO convection center is over the Maritime Continent, and the least frequent when the MJO convection center is in the Western Hemisphere. Over both the northwestern and southwestern Pacific and

north of Australia, TCs are the most (least) frequent when the MJO convection center is over the western Pacific (Indian Ocean). TCs (hurricanes) over the northeastern Pacific occur most frequently when the MJO convection center is in the Western Hemisphere, and least frequently when the MJO convection center is over the Maritime Continent. Atlantic hurricanes are the most (least) frequent when the MJO convection center is over the Indian Ocean (western Pacific). The MJO may affect TCs through impacts on local vertical wind shear, low-level convergence, midlevel moisture, cyclonic relative vorticity, deep convection, small eddies, and synoptic disturbances serving as embryos for TCs.

Tornados

In boreal spring (March, April, and May), the contiguous US is prone to tornados. Violent tornado outbreaks days, with six or more strong tornados within a 24-h period, are more than twice as frequent when the MJO convection center is over the Indian Ocean than other times. Combined intraseasonal anomalous patterns (an upper tropospheric trough extending eastward from the Western United States, upper tropospheric southwesterly wind anomalies, and low-level southerly wind anomalies over the southern Great Plains) and the seasonal circulation (a ridge over the Southeastern US) provide atmospheric conditions favorable for violent tornado formation.

Extreme Rainfall

Globally, frequencies of extreme rainfall events during periods of the active MJO are about 40% higher than periods of weak or no MJOs. Extreme rainfall events are defined here as record-breaking rain events or total rain amount within a given top percentile of the local climate rainfall probability distribution. While several events that brought tremendous damages to society have been suspected to occur partially because of the MJO, the most convincing evidence of MJO modulation on extreme rainfall is the change of extreme rainfall frequency as a function of the longitudinal location of the MJO convection center. For example, during March–May in the equatorial east Africa, 62% of extreme rainfall events over the highlands occur when MJO convection is over the Indian Ocean, while 72% of extreme events over the coastal region occur when MJO convection is suppressed over the Indian Ocean and Maritime Continent. Extreme rainfall events in the semiarid north-central coastal area of Chile occur normally only 3–5 times during the fall and winter of rainy years; about 80% of them happen when

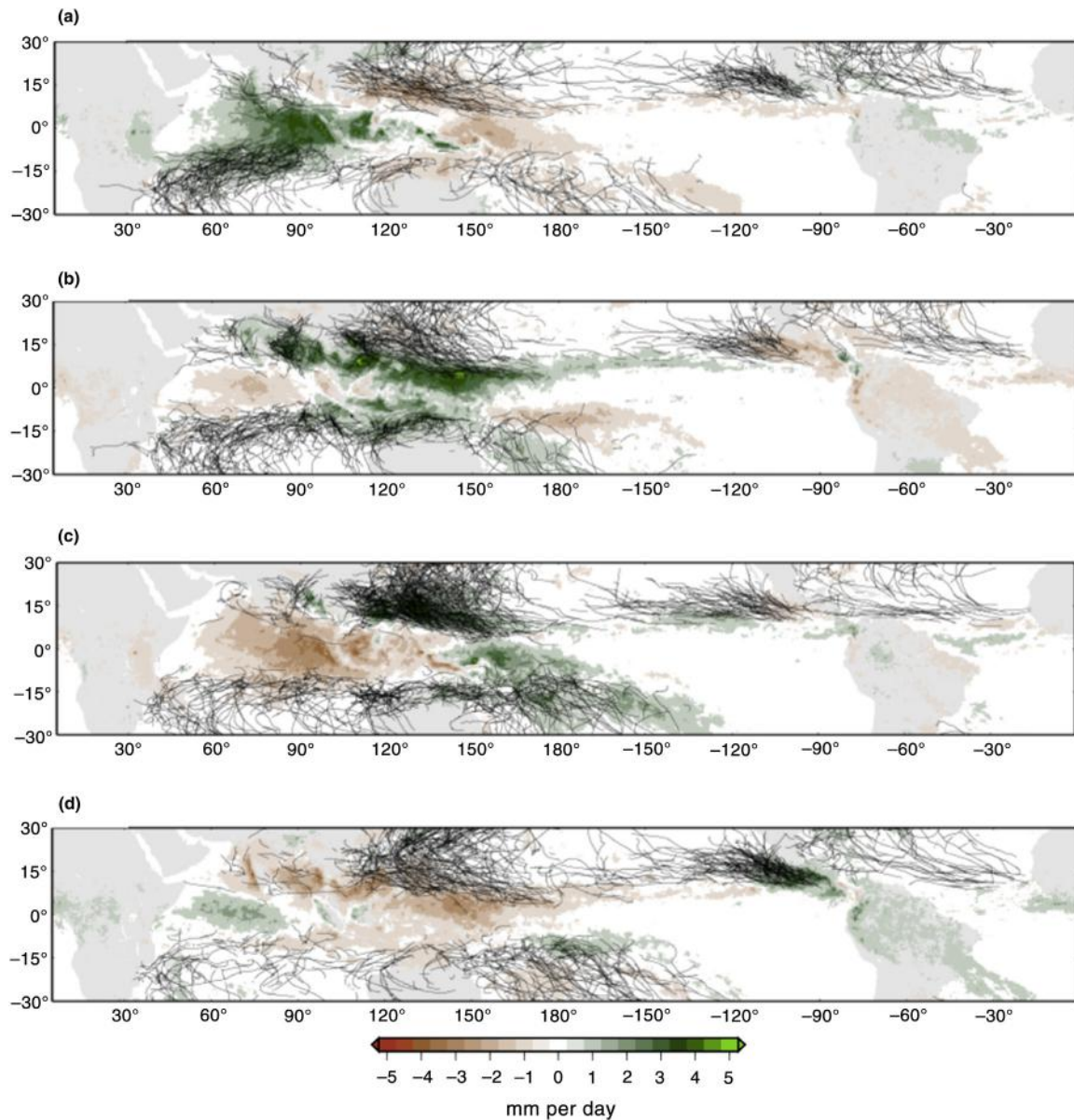


Figure 1 Tropical cyclone (TC) tracks (1975–2011) and precipitation anomalies (1998–2011) when MJO convection center is over the (a) Indian Ocean, (b) Maritime Continent, (c) western Pacific, and (d) Americas and Africa. Adapted from Zhang, C., 2013. Madden-Julian oscillation: bridging weather and climate. *Bulletin of the American Meteorological Society* 94, 1849–1870.

an MJO convection center is in the central equatorial Pacific. In boreal winter over the contiguous United States, extreme rainfall (exceeding 90th percentile of frequency distribution in intensity and spatial coverage) occurs twice as frequently during a period with an active MJO than one without, and is the most frequent when the MJO convection center is over the Indian Ocean.

Flood

It has been suggested that some unusual flood events with large casualties were associated with particular MJO events. However, as for extreme rainfall events, the most evident effect

of the MJO on floods comes from fluctuations of the flood frequency through the life cycle of the MJO or between periods with and without active MJO events. Take 'large flood events' as an example, which is defined as extreme flood events with damages that have been reported with intervals of a decade or longer. Their probability or occurrence frequency is measured as the total number of flood days or flood events (which consists of continuous flood days) divided by the total number of days in a given MJO phase. Frequencies of large flood events in many parts of the world are modulated by the MJO (Figure 2). When the MJO convection center is over the western Pacific, large flood events are most frequent along the west coast of North America, on the Philippine Islands, and over

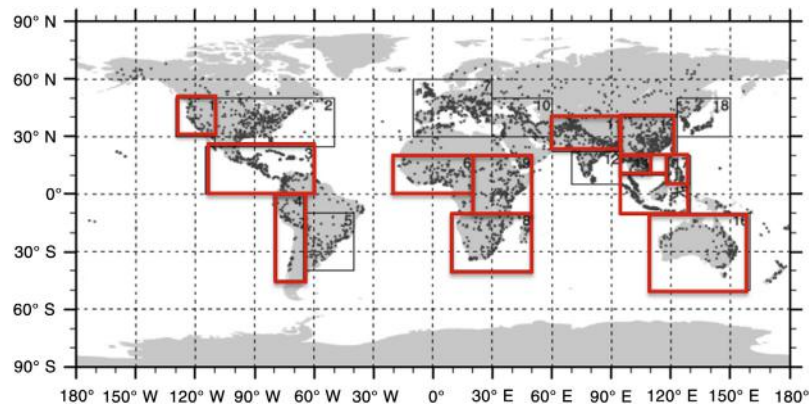


Figure 2 Locations of large flood during 1985–2010 based on the Dartmouth Flood Observatory Global Archive of Large Flood Events at University of Colorado (Brakenridge, G.R. Global Active Archive of Large Flood Events. Dartmouth Flood Observatory at University of Colorado, <http://floodobservatory.colorado.edu/Archives/index.html>). Red boxes mark regions where probabilities of total flood days and/or events are significantly affected by the MJO. Zhang, C., 2013. Madden-Julian oscillation: bridging weather and climate. *Bulletin of the American Meteorological Society* 94, 1849–1870.

South Asia and West Africa. They are the least frequent in these regions when the MJO convection center is over the Indian Ocean. In contrast, large flood events in Central America and East Africa are the most (least) frequent when the MJO convection center is over the Indian Ocean (western Pacific). Over China, frequencies of large flood events tend to be reduced during periods with strong MJO activities in comparison to those without. Similar MJO effects on large flood events can be found in other regions (Australia, the Maritime Continent, the Middle East, Southern Africa, the Andes).

Lightning

By affecting characteristics of deep convection, the MJO modulates the global lightning frequency. In the main region of MJO activity, namely, the tropical Indian and Pacific oceans, very deep (>10 km) lightning-producing convective towers tend to occur immediately prior to the local onset of MJO convective periods, yielding the highest lightning frequency. The MJO also affects lightning outside the tropics. For example, as the MJO convection center moves from the Indian Ocean to the Pacific Ocean, summertime lightning over the United States exhibits a migrating pattern clockwise starting from the Southeast. The MJO effects on lightning project on to the Schumann resonances (SR), which are electromagnetic waves of zonal wave number one in the natural cavity between the Earth and the ionosphere. The intensity of the SR is mainly modulated by fluctuations in the number and intensity of global lightning activity. Intraseasonal perturbations in the SR have been detected and related to the MJO.

Fire

Through its modulation on rainfall and perhaps also wind and temperature, the MJO affects the frequency of wildfire in many regions of the world. Global monthly fire counts are always larger when there are active MJO events than otherwise. **Figure 3** compares global fire frequencies at two stages of the MJO life cycle: one is when MJO convection begins over the

Indian Ocean ('initiation stage'), the other is when MJO convection has propagated across the Maritime Continent and reached the western Pacific ('WP stage'). Fire frequencies are much higher at the MJO initiation stage than the WP stage over the Amazon and northern Australia, whereas they are lower over Alaska and north of Lake Victoria over equatorial Africa. Over East Europe and European Russia, fire is more concentrated to the south at the MJO initiation stage but spreads to the north when the MJO becomes mature. Other regions experience their fire maxima and minima at different MJO stages (phases). Over Siberia, fire is frequent when MJO convection is most enhanced over the eastern Indian Ocean.

Cold Surges

Two-thirds of extreme cold surges with temperature reductions greater than 2 standard deviations in East Asia occur when the MJO convection center is over the Indian Ocean. Although in general, the MJO tends to prevent weak cold surges from penetrating southward into the subtropics and tropics, an MJO event with its convection center stalled over Sumatra resulted in an extreme cold event during a cold year of El Niño Southern Oscillation (ENSO) (February 2008) that broke a 50-year record of minimum daily temperature and duration of large negative temperature anomalies over Southeast Asia. An extreme cold surge with record-breaking snowfall in Korea in the winter of 2009–10 might have resulted from combined effects by the MJO and the Arctic Oscillation (AO).

Climate

Climate phenomena subject to MJO influences include the monsoons and several climate modes such as ENSO, the North Atlantic Oscillation (NAO), the AO and Antarctic Oscillation (AAO), the Pacific North American (PNA) pattern, and the Indian Ocean Dipole (IOD). While these climate modes all feed back to the MJO, discussions in this section focus on MJO effects on them.

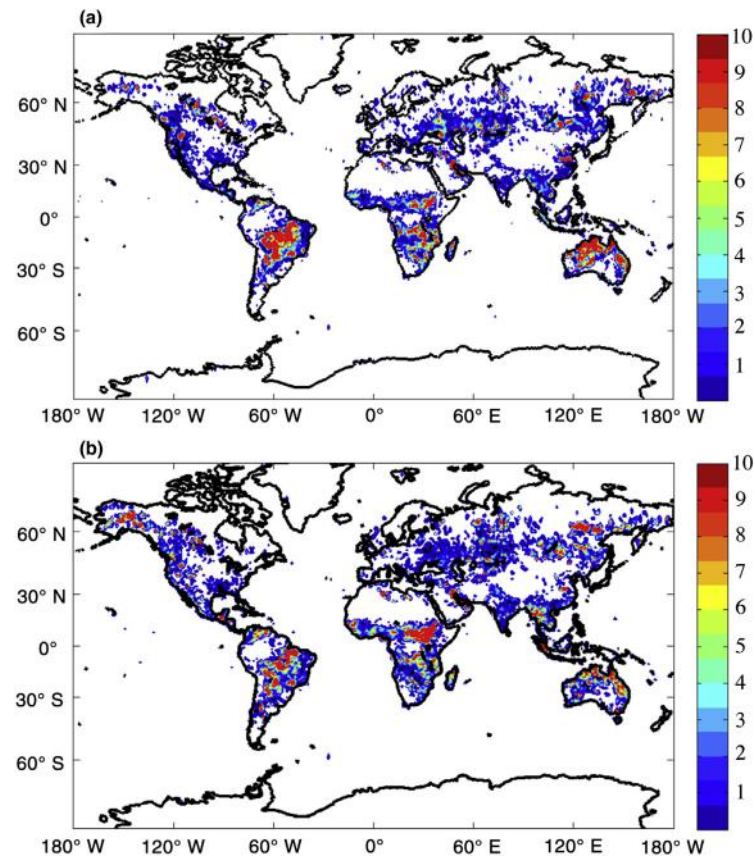


Figure 3 Fire frequency (%) in each $1^\circ \times 1^\circ$ grid box at the (a) MJO initiation stage (MJO convection starts over the Indian Ocean) and (b) MJO WP stage (MJO convection center is over the western Pacific). The fire count data are from the Along Track Scanning Radiometers (ATSR) World Fire Atlas (<http://due.esrin.esa.int/wfa/>).

Monsoons

The MJO is the main source of intraseasonal fluctuations in many, if not all, monsoon systems. About 33–80% of intra-seasonal variability of monsoon rainfall is related to the MJO. Summer monsoon onsets and breaks are often related to the MJO, especially for the Indian and Australian monsoons. Convectively active episodes of the MJO can enhance heavy rainfall events during summer monsoons. The northward propagation of the MJO in boreal summer is as important as its eastward propagation to the Asian monsoon. The MJO affects the American monsoons through two mechanisms: MJO perturbations propagating eastward along the Pacific Inter Tropical Convergence Zone (ITCZ), and Rossby wave trains excited by MJO convection in the tropical Indo-Pacific region that enter South America. Both westward propagating equatorial Rossby waves excited by MJO convection over the Indian Ocean and eastward propagating equatorial Kelvin waves excited over the tropical Atlantic Ocean are considered the main mechanisms for the MJO to affect the African monsoon.

El Niño Southern Oscillation

Strong episodes of the MJO often occur prior to major ENSO warming events (El Niño). The peak of ENSO

warming in the eastern Pacific in boreal winter is usually preceded by enhanced MJO activities in boreal spring. This leads to a significant correlation between ENSO sea surface temperature (SST) and MJO activities with the former lagging the latter by 6–10 months. Oceanic downwelling Kelvin waves forced by westerly wind anomalies of the MJO are the main mechanism for the MJO to influence ENSO. In an ENSO paradigm, the MJO acts as a main source of high-frequency stochastic forcing that drives ENSO. Meanwhile, feedback from ENSO SST to the MJO can be important to this MJO–ENSO connection, making the MJO a ‘multiplicative’ stochastic forcing of ENSO.

North Atlantic Oscillation

Enhanced (suppressed) MJO convection over the central Pacific can considerably amplify the NAO in its negative (positive) phase. Northward momentum transport by Rossby wave dispersion from the tropical Pacific to the extratropical North Atlantic is thought to be the main mechanism for the MJO influence on the NAO. The daily phase of the wintertime NAO can be predicted with a success rate of about 70% at a lead time of 9–13 days by a statistical model based on the connection between the MJO and NAO.

Arctic Oscillation

During winter, the positive (negative) phase of the AO, also known as the Northern Annular Mode (NAM), is twice as likely to occur as the opposite phase when MJO convection is enhanced (suppressed) over the Indian Ocean. When MJO convection is enhanced (suppressed) in the Eastern Hemisphere, especially over the Maritime Continent, the number of days of positive (negative) AO phase becomes large. In November–March, 18–21% of the variance in extratropical 1000-hPa geopotential height is related to the MJO. The MJO influence on the AO is also through Rossby wave trains excited by MJO convection and propagating from the tropical Pacific into the extratropics.

Antarctic Oscillation

The southern hemispheric counterparts of the NAM and AO are the Southern Annular Mode (SAM) and AAO. They are also influenced by the MJO. Negative (positive) phases of the AAO in austral winter tend to occur when MJO convection is enhanced (suppressed) over the central Pacific. The SAM reaches its maximum positive phase immediately after MJO convection peaks over the equatorial Indian Ocean. The Antarctic circumpolar transport can be accelerated by MJO-enhanced surface westerly wind associated with the SAM that covers almost the entire latitude circle at 60° S.

PNA Pattern

The PNA teleconnection pattern undergoes both interannual and intraseasonal variations. On the intraseasonal timescale, about 30% of the emergence of the PNA pattern is related to the MJO. The positive (negative) phase of the PNA pattern is most likely to occur when MJO convection is inactive (active) over the region from the Bay of Bengal to the western Pacific. This MJO–PNA connection explains about 30% of the emergence of the PNA pattern. This MJO–PNA connection is through propagation of Rossby wave trains from anomalous convection of the MJO.

Indian Ocean Dipole

Anomalous upwelling oceanic Kelvin waves forced by easterly anomalies associated with suppressed convection of the MJO over the equatorial Indian Ocean can lead to shoaling of the thermocline in the eastern Indian Ocean and thereby help the onset of positive phases of the IOD. Active convection associated with the MJO may help terminate a positive IOD phase by generating downwelling oceanic Kelvin waves that tend to deepen the thermocline. By the same token, an IOD mature phase can sustain itself only in the absence of strong MJO events.

Upper Ocean

Through its anomalies in surface wind, cloudiness, and rainfall, the MJO strongly disturbs fluxes of momentum, heat, and freshwater at the ocean's surface along its pathway from the Indian Ocean to the Pacific Ocean. The direct consequences are changes in the upper ocean structures of currents, temperature, and salinity, and generation of oceanic waves.

Variability in the upper ocean chemistry and biology induced by the MJO is discussed in Section [Atmospheric–Oceanic Chemistry and Biology](#).

Ocean Mixed Layer

The most commonly observed effect of the MJO on the ocean is the intraseasonal perturbation in SST underneath the MJO pathway in the equatorial Indian and Pacific oceans. The average SST anomaly amplitude is typically 0.25 °C, but can be as large as 1 °C in individual cases. The sea surface is warmed during convectively suppressed period of the MJO when there is relatively less cloud coverage and therefore more insolation. The sea surface is cooled by the MJO mainly through enhanced latent heat flux by westerly wind anomalies and reduced insolation by increased cloudiness in and near its convection center ([Figure 4](#)). The cooling can be intensified by upwelling of cold water from below the ocean mixed layer, when the vertical mixing in the upper ocean is particularly vigorous due to extraordinarily strong surface wind. While surface cooling tends to reduce the buoyancy and enhance vertical mixing, this effect is often compensated by an increase in buoyancy by rainfall of the MJO. Abundant freshwater input into the upper ocean during convectively active period of the MJO may create a barrier layer, defined as an isopycnal layer embedded in a deeper isothermal layer. Such a barrier layer makes it difficult for vertical mixing due to surface perturbations to penetrate deep down and thereby isolates the upper ocean from the thermocline. One consequence is quick warming and cooling of the surface layer and a large diurnal cycle of SST during convectively suppressed periods of the MJO when surface wind is relatively weak.

Surface Currents

Another consequence of the strong surface wind of the MJO, especially its westerly wind anomalies, is the acceleration of surface current. The Wyrtki jets are the most striking example. They are narrow (2° N–2° S), eastward currents at the ocean surface (0–100 m) along the equator across the Indian Ocean generated by seasonal mean westerly surface wind during the transition periods (April–May and October–November) between the two monsoon seasons. Due to surface westerly wind associated with the MJO, the Wyrtki jets fluctuate substantially on intraseasonal timescales with amplitude (0.5–2 m s^{−1}) often greater than their seasonal means. Anomalous Wyrtki jets can also occur during other seasons because of strong westerly forcing by the MJO. The intraseasonal variability of the Wyrtki jets exhibits two spectral peaks. Direct MJO forcing results in a 30- to 60-day peak. Resonant excitation of the second-baroclinic-mode waves by the MJO wind at its spectral end of low frequencies and interference between directly forced and reflected ocean waves lead to a 90-day peak.

The MJO also causes intraseasonal fluctuations in the Indonesian Throughflow (ITF), which is the surface current passing through the Indonesian archipelago and acts as the main artery transporting heat and mass from the Pacific Ocean to the Indian Ocean. The intraseasonal spectral peaks of the ITF at 50–60 days are partially due to direct MJO wind forcing and partially due to MJO-forced oceanic waves. In addition, the MJO also induces intraseasonal fluctuations in other ocean

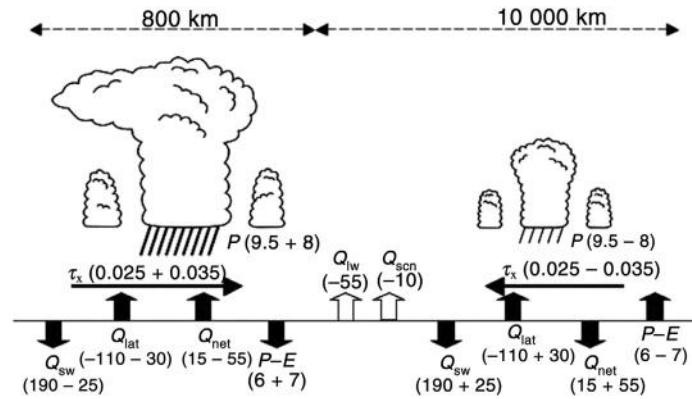


Figure 4 Schematic diagram illustrating the amplitudes of surface forcing of the MJO in terms of the mean plus/minus intraseasonal perturbation in solar radiation flux (Q_{sw}), latent heat flux (Q_{lat}), net heat flux (Q_{net}), precipitation (P), freshwater input ($P - E$), and zonal stress (τ_x), whose direction is marked by horizontal arrows. Downward (upward) pointing arrows and positive (negative) values indicate fluxes into (from) the ocean. The deep (shallow) cloud symbol at the left (right) represents a convectively active (inactive) phase of the MJO, whose zonal scale is indicated at the top. Intraseasonal fluctuations in longwave radiation flux (Q_{lw}) and sensible heat flux (Q_{sen}) are negligibly small and are assumed to be zero. Units are W m^{-2} for the fluxes, mm day^{-1} for precipitation and freshwater input, and N m^{-2} for wind stress. Uncertainties in the net fluxes are 10 W m^{-2} . Zhang, C., 2005. Madden–Julian oscillation. Reviews in Geophysics 43, RG2003. <http://dx.doi.org/10.1029/2004RG000158>.

currents, such as the Somali current and the Indian Ocean south equatorial current.

Waves

The strong surface westerly wind of the MJO often generates downwelling oceanic Kelvin waves that propagate eastward and equatorial Rossby waves that propagate westward along the equatorial waveguide. The downwelling Kelvin waves play an instrumental role in MJO effects on ENSO, IOD, and ITF, among others. When these Kelvin waves reach the eastern boundary of an ocean basin (e.g., Sumatra), they lead to coastal waves that move along the coastlines toward higher latitudes where they cause local impacts, such as large intraseasonal fluctuations in SST near the northwestern coast of Australia.

Sea Level

Sea level fluctuates intraseasonally with amplitudes more than 10 cm in many areas of the Indian and Pacific oceans due to anomalies in MJO-related surface wind and their generated equatorial Kelvin and Rossby waves and coastal waves. Strongest sea level fluctuations in response to the MJO are in the equatorial Pacific, the coastal region of Sumatra, the Bay of Bengal, and the Gulf of Carpentaria. Along the equator, sea level anomalies propagate eastward at the phase speed of the oceanic Kelvin wave. Intraseasonal fluctuations in sea level at higher latitudes are mostly in the coastal region, due to coastal waves originating in the tropics.

Atmospheric–Oceanic Chemistry and Biology

Several types of atmospheric gases undergo intraseasonal fluctuations under the influence of the MJO. Tropospheric ozone shows positive (negative) anomalies in the subtropical regions of the upper-level cyclonic (anticyclonic) circulation east (west) of equatorial MJO convection centers. Tropospheric ozone anomalies associated with the MJO are about 2.5 Dobson units. These

perturbations in tropospheric ozone are related to the vertical displacement of the ozone-rich stratospheric air due to the sinking or rising of the tropopause associated with the cyclonic or anticyclonic circulations generated by MJO convection.

Fluctuations in carbon monoxide (CO) near the tropopause that propagate along with MJO convection have also been observed. Positive (negative) anomalies in CO in the tropopause transition layer (TTL) are colocated with enhanced (suppressed) MJO convection. These CO anomalies are likely generated by vertical transport (or a lack of it) of CO from lower troposphere to the TTL by convective updraft.

MJO-related intraseasonal fluctuations in carbon dioxide (CO_2) are observed in the midtroposphere with an amplitude of about 1 ppmv. The largest positive (negative) CO_2 anomalies are located near enhanced (suppressed) MJO convection. MJO-related vertical motions that transport high (low) CO_2 concentration from the lower (upper) troposphere upward (downward) are likely to be the main mechanism for the midtropospheric CO_2 fluctuations.

The MJO modulates tropospheric aerosol. Positive (negative) aerosol anomalies are found over the equatorial Atlantic Ocean and Africa when MJO convection is suppressed (enhanced) over the equatorial Indian Ocean and suppressed (enhanced) over the equatorial western Pacific. The amplitude of the intraseasonal fluctuations in the aerosol optical thickness is about 0.04. Enhanced or reduced zonal advection by changes in the low-level zonal wind direction associated with the MJO can be a mechanism for the intraseasonal fluctuation in aerosol over the tropical Atlantic Ocean.

There is a regime transition in surface aerosol on intraseasonal timescales over the equatorial Indian Ocean. Prior to an MJO convectively active period, surface aerosol is dominated by submicrometer sulfate aerosol of continental origin. When surface wind becomes strong during and after the MJO convective period with heavy rainfall, the continental aerosol are washed out and sea spray becomes dominant.

The strong surface wind of the MJO can induce entrainment at the base of the ocean mixed layer, which brings nutrient-rich

cold water from below into the mixed layer. In consequence, increases in surface chlorophyll and phytoplankton bloom occur in areas of surface cooling due to the MJO. These have been observed in many tropical and subtropical open oceans and coastal regions, ranging from the Arabian Sea, northern Indian Ocean, Bay of Bengal, Southeast Asia, and Pacific. Anomalies in Pacific surface chlorophyll propagate eastward with the MJO.

Other Impacts

The MJO also impacts many other phenomena of the Earth system. Some of them cannot be labeled as either weather or climate.

Earth's Angular Momentum

The momentum exchange between the atmosphere and the solid Earth is modulated in part by the MJO. A mountain torque is produced when surface Kelvin waves generated by the MJO propagate along the Pacific equator and are intercepted by the Andes and turn into poleward edge waves along the Andes. In consequence, the Earth's angular momentum fluctuates in concert with MJO activity and so does the length of the day (LOD). In periods of strong MJO activity (e.g., boreal winter), the spectrum of LOD shows a 50-day peak. Fluctuations in LOD with the MJO are about 1.5×10^{-3} s.

Tropopause

When MJO convection moves from the Indian Ocean to the western Pacific, the tropical tropopause rises at most longitudes as well as in regions near the MJO convection center. Tropopause (cold-point) temperature drops more than 2 °C. The largest perturbations occur in the subtropical cyclonic vortices on the polar sides of the MJO convection center. In the tropical (10° S–10° N) TTL, mean cirrus fraction doubles, with its maximum over equatorial Africa and South America.

Near the tropopause, water vapor and cloud ice water content fluctuate intraseasonally with the MJO deep convection. Within the range of 120 ppmv, intraseasonal fluctuations in upper-troposphere (215 hPa) moisture propagate eastward with MJO convection center over the Indian and Pacific oceans. Anomalies in moisture coherent with the MJO are out of phase at the 215 and 100 hPa levels. Anomalies in ice water content are about $\pm 2 \text{ mg m}^{-3}$ at 215 hPa and decrease upward. Positive ice water content anomalies are located in the region of enhanced MJO convection.

ITCZ

The MJO influences the breakdown of the Pacific ITCZ. Deep convective clouds in the ITCZ yield to shallow convection when MJO zonal wind switches from the westerlies to the easterlies and the atmospheric environment for deep convection changes from neutral to suppressing. The eastward propagation signal of the MJO over the eastern Pacific is the most robust in the ITCZ. Coincident with the eastward propagation of MJO, there is a tendency for ITCZ convection to move northward. Convective initiation of the MJO over the Indian Ocean interrupts the ITCZ south of the equator. Interaction between the MJO and Atlantic

ITCZ is a potential mechanism for hurricane genesis. Precipitation in the Atlantic ITCZ is enhanced when the MJO convection center is over either the Indian Ocean or the western Pacific Ocean, but reduced when it is over the Maritime Continent.

Diurnal Cycle

Diurnal cycles in both the atmosphere and upper ocean from the tropical Indian Ocean to the western Pacific are modulated by the MJO. The diurnal cycle in convection over continental land and islands is much stronger during periods of suppressed convection of the MJO due to strong diurnal heating at the surface. In periods of active convection of the MJO, daytime insolation is reduced by enhanced cloudiness and mesoscale convective systems become prominent, which may last beyond one day. These tend to overwhelm diurnal forcing of convection and substantially reduce diurnal signal in convection. Over the ocean, a strong diurnal cycle in SST occurs during convectively suppressed periods of the MJO when surface wind is weak, daytime solar heating is strong, and nighttime radiative cooling is effective in inducing buoyancy driven vertical mixing. During convectively active periods of the MJO, strong surface wind generates constant mixing, which smears diurnal heating and cooling, hence a weak diurnal cycle in SST.

Concluding Remarks

The broad impacts of the MJO on global weather, climate, and many other phenomena are a strong testament of the unique role of the MJO in connecting weather and climate and affecting many aspects of the Earth system. Hardly can any other single atmospheric phenomenon be found which is even close to the MJO in this regard. Because of this unique role, improvement of MJO prediction can potentially lead to advancement in prediction of many other phenomena in the Earth system. It has been demonstrated that numerical models that produce more realistic MJO signals are capable of predicting many other phenomena, such as TCs and extratropical circulation patterns, with a longer lead time. It is clear that advancement of Earth system prediction must be achieved with a better understanding and prediction of the MJO.

See also: **Climate and Climate Change:** Climate Variability: North Atlantic and Arctic Oscillation. **Electricity in the Atmosphere:** Lightning. **Synoptic Meteorology:** Extratropical Cyclones. **Tropical Meteorology and Climate:** El Niño and the Southern Oscillation: Observation; El Niño and the Southern Oscillation: Theory; Intertropical Convergence Zone; Intraseasonal Oscillation (Madden–Julian Oscillation); Monsoon: Overview.

Further Reading

- Lau, W.K.M., Walliser, D.E. (Eds.), 2011. *Intraseasonal Variability of the Atmosphere–Ocean Climate System*, second ed. Springer, Heidelberg, Germany, p. 613.
- Zhang, C., 2005. Madden-Julian oscillation. *Reviews in Geophysics* 43, RG2003. <http://dx.doi.org/10.1029/2004RG000158>.
- Zhang, C., 2013. Madden-Julian oscillation: bridging weather and climate. *Bulletin of the American Meteorological Society* 94, 1849–1870.

Greenhouse Effect

GR North, Texas A&M University, College Station, TX, USA

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

The Earth's climate is determined by a balance between the rate of radiation energy from Earth to space and by the rate of its receipt and absorption from the Sun. The radiation rate to space increases rapidly with temperature, so that if the rate of heating exceeds that of cooling to space, the planetary temperature will rise leading to increased radiation emitted until the balance is restored. If there is a deficit, the reverse will happen. When greenhouse gases such as CO₂, CH₄, and H₂O are in the air, they cause the effective altitude of emission to space to be higher in the atmosphere where it is colder. This effect reduces the radiation to space for a given surface temperature. This reduction of emitted radiation means the planetary temperature will increase until balance is reestablished. This article addresses the question of the response of the surface temperature field to a doubling of CO₂, for the case where no feedback mechanisms are operating. The value found is $1.0 \pm 0.1^\circ\text{C}$, a value often quoted in the literature. Various feedback mechanisms can multiply or diminish the response. These include positive feedback mechanisms such as water vapor, ice-albedo; negative feedback mechanisms include lapse rate. Cloud feedback is not yet fully understood, but the balance of evidence suggests that it is positive.

Introduction

The greenhouse effect refers to the complex of phenomena starting with the passage of visible light from the Sun through the atmosphere essentially unattenuated to be absorbed by and warming the surface. The cooling to space by emitted radiation is inhibited by gases present in the atmosphere that absorb radiation and essentially trap the heat from escaping to space as efficiently as would otherwise occur. The effect was first considered by Joseph Fourier, the famous French mathematician and philosopher in the 1820s. Later in that century John Tyndall isolated a number of gases such as H₂O and CO₂ and showed that they are absorbers of infrared radiation (he called it *heat radiation*), but that the dominant (>99.9% by molecule count) constituents of Earth's atmosphere (N₂, O₂, and Ar) are transparent in both the visible and the infrared. Later scientists such as Arrhenius, Calendar, and Plass took an interest in the problem, eventually leading up to the modern study with sophisticated general circulation models of the atmosphere (the most remarkable of which is [Manabe and Wetherald, 1975](#)). The history of this exciting course of research is presented by [Weart \(2009\)](#). The atmospheric greenhouse effect is much more subtle than the analogy with a building with a glass roof that does cause a warmer interior, but increasing the glass thickness has no effect because the glass already absorbs all of the upwelling radiation. Instead in the real atmosphere the upward passage of infrared radiation through the atmosphere is strongly affected by the vertical dependence of temperature of the air and the changes in absorptivity of the greenhouse gases with air pressure. Moreover, the absorption of infrared radiation by greenhouse gases in the atmosphere depends very strongly on the distribution of energy over wavelengths of the emitted infrared radiation. The calculation is complicated and requires accurate and detailed knowledge of infrared absorption spectra for each gas. At this time, we are confident that we have adequate computer algorithms and empirical information to make these calculations for clear skies to an accuracy adequate for predicting the effects of changes in the concentration of the common greenhouse gases. A useful collection of

scientific papers from the recent period is collected in the volume edited by [Archer and Pierrehumbert \(2011\)](#).

The Earth's average surface temperature is about 289 K. For a blackbody at this temperature the emission spectrum peaks at around 10 μm (1 μm denoted $\mu\text{m} = 10^{-6}\text{ m}$), which lies in the infrared portion of the spectrum. The dominant gases in the Earth's atmosphere are N₂ (78.084% by molecule count), O₂ (20.946% by molecule count), and Ar (0.934% by molecule count). None of these species absorb in the infrared portion of the electromagnetic spectrum. On the other hand, there are some trace gases that are well mixed throughout the atmosphere (up to hundreds of kilometers) such as CO₂, CH₄, and N₂O that do absorb and emit radiation in parts of the infrared band of wavelengths. For convenience, let us call these gases GHGs. Ozone (O₃) is a strong greenhouse gas and is variable in the troposphere (since its lifetime in the troposphere is of the order of weeks, too short for homogenization over the globe). It is abundant in the stratosphere where it is produced by photolysis (dissociation of molecular oxygen or other species by ultraviolet or X-rays from the Sun to give free O atoms, leading to their subsequent combination with O₂, releasing heat). In addition, water vapor is a strong absorber and emitter in the infrared, but its concentration or mixing ratio (ratio of its number of molecules per unit volume to that of background air, often given as parts per thousand or parts per million) is not uniformly mixed in the atmosphere. Its mixing ratio is strongly diminished in the vertical because of condensation as the temperature falls below the saturation value for the vapor. Water vapor's short turnover time in the atmosphere and the heterogeneous source locations at the surface lead to spatial nonuniformities that can last for days. The condensation process causes the vertical scale height to be only a few kilometers, compared to 8–15 km (depending on latitude and season) for the dry air and the well-mixed GHGs listed above. Water vapor even in trace amounts in the upper troposphere and stratosphere is a strong absorber and emitter of infrared radiation. So technically water vapor is a GHG as well but it is mainly controlled by the temperature and the availability of liquid water at the local surface.

Earth's Climate without Greenhouse Gases

Suppose there were no GHGs in the atmosphere. If that were the case, we could compute the globally averaged surface temperature from balancing the rate of outgoing infrared radiation to the rate at which solar radiation is absorbed. The amount of radiation to space in this case would be $4\pi R^2 \sigma T^4$ where $\sigma = 5.67 \times 10^{-8} \text{ J m}^{-2} \text{ s}^{-1} \text{ K}^{-4}$, the Stefan-Boltzmann constant, and R is the Earth's radius. The rate of energy absorbed is $\pi R^2 S_0(1 - \alpha_p)$ where S_0 is Total Solar Irradiance (TSI, the amount of radiation energy per second from the Sun impinging on a one square meter surface perpendicular to the solar beam at the annual average distance of Earth to the Sun), which has the value 1360 W m^{-2} (the TSI is also known as the solar constant), and α_p is the planetary albedo (or reflectivity to sunlight). The current value of α_p is about 0.30, but this includes clouds and snow/ice. For a planet with no GHGs there would be no snow/ice or clouds (as we know them, anyway). Hence, we will take a conservative estimate of 0.10 for the albedo. Equating the rates, we obtain a value of $T = 271 \text{ K}$. Using $\alpha_p = 0.2$ yields 263 K . Both are unacceptably cold when compared to the observed present temperature of 289 K . The presence of the GHGs is the reason.

A Gray Greenhouse

To gain some understanding of how the greenhouse effect works consider the planet above, but surround the planet with a thin spherical shell, which is transparent to solar radiation but partially absorbing to terrestrial (infrared) radiation. The surface now receives radiation not just from the Sun but also from downwelling infrared radiation from the shield, which is itself a material surface capable of emitting radiation both upward and downward. The total rate absorbed at the surface is $\pi R^2 S_0(1 - \alpha_p) + 4\pi R^2 \epsilon \sigma T_{\text{sh}}^4$, where $\epsilon(0 \leq \epsilon \leq 1)$ is the shield's emissivity (also its absorptivity, a rule known as Kirchoff's Law) and T_{sh} is the temperature of the shield. The ground radiates upward at a rate $4\pi R^2 T_{\text{surf}}^4$. The shield is now a player as well and needs its own energy balance equation. It radiates both upward and downward at a rate $4\pi \epsilon R^2 T_{\text{sh}}^4$. It absorbs infrared radiation upwelling from the surface at a rate $4\pi \epsilon R^2 T_{\text{surf}}^4$. We have the balance equations for the shield (after canceling common factors):

$$S_0(1 - \alpha_p) = 4\sigma T_{\text{surf}}^4 - 4\epsilon \sigma T_{\text{sh}}^4, \quad \text{surface balance} \quad [1]$$

$$T_{\text{surf}}^4 = 2T_{\text{sh}}^4, \quad \text{shield balance} \quad [2]$$

leading to:

$$T_{\text{surf}} = T_{\text{surf}}^0 \left(\frac{1}{1 - \frac{\epsilon}{2}} \right)^{\frac{1}{4}} \quad [3]$$

where T_{surf}^0 is the surface value at $\epsilon = 0$, the case of a transparent and therefore nonparticipating shield. Note that the surface temperature is largest at $\epsilon \rightarrow 1$. Inserting an 'optically thick' or 'black' infrared shield increases the Kelvin temperature at the surface by a factor of $2^{\frac{1}{4}} = 1.189$, the largest warming a single shield can cause. This is at best a very crude estimate of the greenhouse effect in the real atmosphere. We could add

more gray shields, each in 'radiative equilibrium' in terms of their own temperatures, and thereby make this greenhouse effect stronger. Note how different this multiple shield case is from simply thickening the glass in a greenhouse, which is equivalent to increasing ϵ . One flaw of the gray-atmosphere approach is the fact that the absorption of the GHGs at different wavelengths is hardly constant. In fact, it is highly irregular. In the next section, we look at some illustrations with more realistic details of the infrared spectra emitted by the individual greenhouse gases in an idealized column of air.

Infrared Spectra of Outgoing Radiation

In this section, we go a step further and examine the detailed spectrum of the outgoing radiation. Many decades of laboratory and theoretical research on molecular absorption spectra have led us to a very detailed understanding of the relevant parameters such as line intensities and the pressure and temperature dependences of line widths. The term 'line' is used here to indicate that infrared absorption spectra are mostly composed of series (typically millions) of very narrow but strong spikes as a function of wavelength. The spikes are the result of 'resonant' absorption at the natural frequencies of the molecules as they rotate and vibrate, these frequencies can be understood through the treatment of molecular dynamics via quantum mechanics. Integrals over intervals in wavelength must treat the distribution of these spikes and other discontinuities carefully.

Through the same historical period radiation transfer computer algorithms have been improved to the point that under ideal conditions such as no clouds or aerosols, we can compute the spectrum of outgoing radiation to space with high accuracy. Satellite observations with high-resolution infrared spectrometers have also verified the calculations. Here for demonstration purposes, we use an early version of the program that is available to the public called MODTRAN. A simplified version of it is available in the form of a calculator at the Web site (<http://forecast.uchicago.edu/modtran.html>, developed by David and Jeremy Archer; we refer to this as the Chicago Web site). The program takes a specific latitude belt (in our case the tropics) and specifies sky conditions (in our case *clear* with no precipitation or clouds). The user can specify the (vertically uniform) tropospheric mixing ratios of methane and carbon dioxide and ozone's climatological profile can be scaled up or down by a constant factor. The surface temperature can be given but once entered into the calculator the vertical dependence of temperature (the lapse rate) is constrained to the climatological profile at the latitude belt chosen (here tropical). Constraining the lapse rate implies that some vertical transport of heat is occurring in the atmosphere along with the radiation transfer of energy. The program then assigns a vertical distribution of the greenhouse gases, temperature, and air density appropriate to the latitude and mixing ratios of the GHGs as specified by the user. The vertical distribution of water vapor is adjusted to be in equilibrium with the climatological temperature profile as given in the calculator model (either fixed relative humidity or fixed vapor pressure). We chose the clear-sky tropics for our example because the tropopause is very high allowing very cold temperatures at its highest points

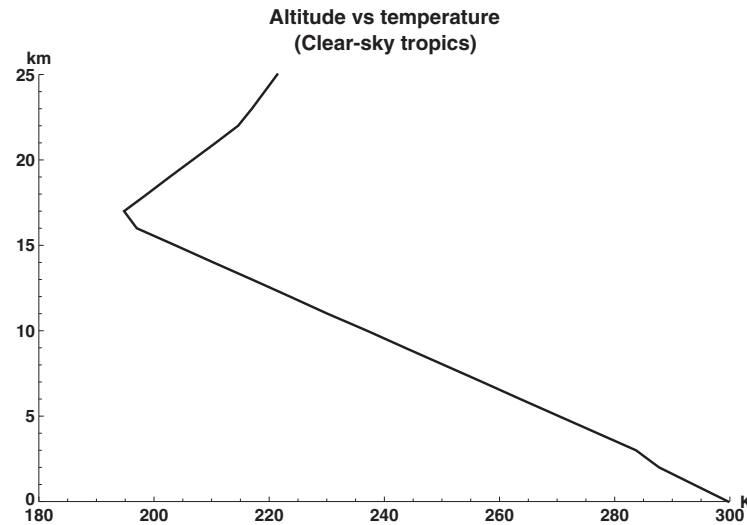


Figure 1 Vertical profile of the temperature in the clear-sky tropics used in this exercise and specified in the MODTRAN calculator model based on the Web site: <http://geoflop.uchicago.edu/forecast/docs/Projects/modtran.orig.html>.

(≈ 195 K) with an average lapse rate of nearly 10 K km^{-1} . **Figure 1** shows the vertical dependence used by the calculator model for the clear-sky tropics. These conditions give a maximum greenhouse effect. A number of other minor greenhouse gases are included in the standard program (such as N_2O , NH_4 , and the chlorofluorocarbons) and are fixed at climatological profiles.

Figure 2 shows the spectrum of outgoing radiation ($\text{W m}^{-2} \text{ cm}^{-1}$) as a function of wave number (cm^{-1}). (Wave number is defined as the inverse of wavelength. It is proportional to the frequency of the electromagnetic radiation, c/λ , where c is the speed of light. Wave number is used in infrared spectroscopy because it is easily related to the molecular structure of the molecules involved.) This first case includes all

greenhouse gases at or near their present concentrations. The infrared flux is for the clear sky in the tropics. Incidentally, all computations are conducted in this version of MODTRAN with spectral resolution of 2 cm^{-1} . (The most accurate calculations are the so-called Line-by-Line (LBL) codes. These take into account each individual line and its width, the latter varying with altitude. Such calculations are very important for benchmarking approximate radiative transfer codes such as MODTRAN and testing against measurements, but LBL codes are too consumptive of computer time to be used directly in climate model calculations.) If there were no carbon dioxide in the atmosphere the spectrum would be the blue curve denoting blackbody radiation at 300 K. Also shown in green is the blackbody curve for an emission temperature of 215 K. The

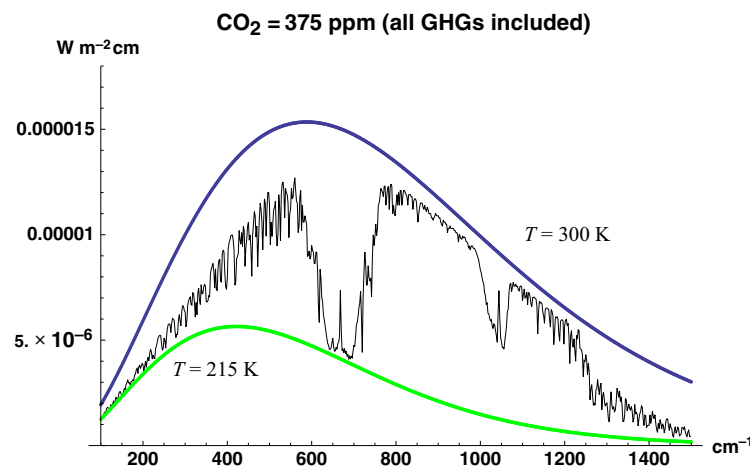


Figure 2 Outgoing infrared spectrum for the Earth's tropical clear-sky atmosphere with all greenhouse gases present. The abscissa is wave number in cm^{-1} , while the ordinate is $\text{W m}^{-2} \text{ cm}^{-1}$. The blue curve is the emission spectrum expected for a 300 K blackbody radiator and the green line is for a 215 K emitter. The area under the curves is proportional to the energy flux density between the horizontal limits on the abscissa. Adapted from calculations based on the Web site: <http://geoflop.uchicago.edu/forecast/docs/Projects/modtran.orig.html>.

large negative departure (the 'ditch') between wave numbers 600 and 800 cm^{-1} is due to the presence of CO_2 . Notice that the emission around 680 cm^{-1} is at an emission temperature of 220 K, a temperature about 25 K above that of the tropopause in the tropics (see Figure 1). The spike in the center (located at 667.4 cm^{-1}) of the band is due to a very strong narrow feature in the CO_2 absorption/emission spectrum. This feature is so strong and hence the result is a spike emerging from the floor of the ditch.

There are a number of interesting features besides the main ditch in Figure 2. For example, there is another, but smaller, ditch at about 1100 cm^{-1} due to the GHG ozone (O_3). Note that the emission surrounding the ozone ditch from around 800 to about 1250 cm^{-1} hugs the blue 300 K emission spectrum for a blackbody whose surface is at 300 K, the surface temperature specified in this simulation. This broad band is called the 'atmospheric window,' a range of wave numbers where water vapor has almost no absorptivity so that the radiation in this band is coming almost directly from the Earth's surface (this turns out to be important when clouds are present). Left of the main CO_2 ditch is a wide band of emission that is coming from cooler emission temperatures. This is mainly coming from layers of water vapor well above the surface. The same holds for wave numbers beyond 1250 cm^{-1} , except that methane is also playing a role there. The main point of Figure 2 is that greenhouse gases reduce the outgoing radiation from that of a blackbody at 300 K by about 60 W m^{-2} .

As a second example consider the planet with only one GHG, CO_2 , at a nominal concentration of 375 ppm. Figure 3 shows this case wherein the same conditions as Figure 2 apply (temperature at the surface, 300K, tropical clear sky), except that there are no other GHGs. We see the CO_2 ditch clearly but no other prominent features. Consider the bottom of the ditch. The bottom follows the 220 K blackbody curve (see the green 215 K curve in Figure 2), except for the spike in the center of the ditch. Why 220 K? This is because the tropopause temperature is at this level in this particular simulation. As shown in Figure 1, the temperature in the troposphere falls off nearly linearly from the surface (300 K) to the tropopause (195 K)

then begins to increase steeply in the lower stratosphere. As we increase the concentration of CO_2 in the air column, the level at which the emission to space occurs rises. As this level rises, the emission temperature lowers until the tropopause is reached. This explains the tilted, flat bottom of the ditch. The spike is due to a very strong CO_2 emission line (actually a convergence of many lines) at the center of the ditch. The emission level of this line is well above the level corresponding to neighboring wavelengths in the ditch and it hits the tropopause when the CO_2 concentration reaches a mere 25 ppm. As the concentration of CO_2 in the air column increases the spike's emission level (being in the stratosphere) actually goes down. In other words, the spike tends to cool the atmospheric column as the concentration is increased. The decrease in outgoing radiation is not at the tilted flat bottom of the ditch but rather in the 'wings' of the ditch.

What determines the emission altitude of the CO_2 ? It varies by wave number as should be expected. Since a good emitter is also a good absorber, we can visualize the emission altitude by considering a downwelling beam at a particular wave number interval. The downwelling beam is attenuated by the CO_2 absorber proportional to its local density and its absorptivity. As one descends into the atmosphere more of the beam is attenuated by CO_2 and when the absorption reaches a value e^{-1} , where e is the base of the natural logarithm ≈ 2.718 . As viewed from above this level is the location of 'unit optical depth.' This level is also roughly the average level of emission (it is called the 'Chapman Level'), but it is smeared out over about one unit optical depth. Now we can see that as the total concentration of CO_2 is increased and because it is constrained to be exponentially distributed in the vertical (well-mixed), the Chapman Level must increase.

Figure 4 shows the result of doubling the concentration of CO_2 from 375 to 750 ppm in the case where no other GHGs are present. The range of wave numbers focusses on the band of wavelengths covering the ditch in order to see more clearly the changes in the outgoing spectrum. As the CO_2 concentration is doubled, the flat portion of the ditch is unaffected, but the wings of the ditch are clearly deepened by a small amount. The

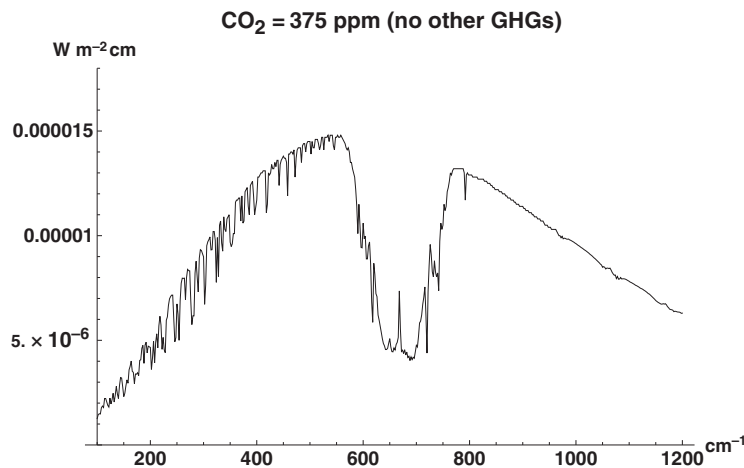


Figure 3 The outgoing infrared radiation spectrum for the Earth's tropical clear-sky atmosphere where the only greenhouse gas present is CO_2 at a mixing ratio of 375 ppm.

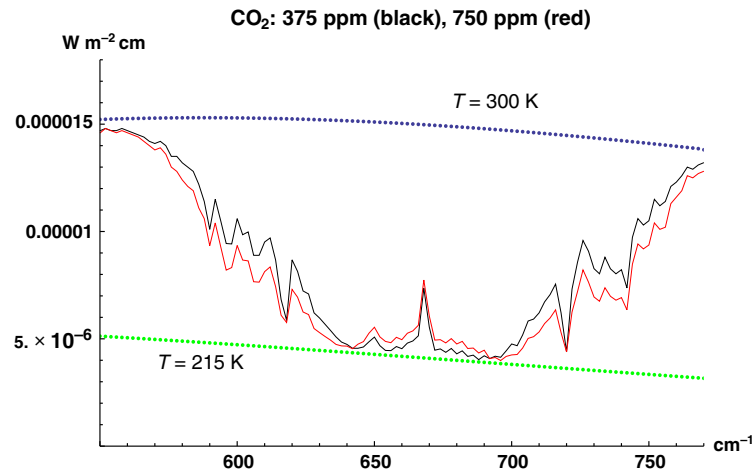


Figure 4 The outgoing infrared spectrum in the range $550\text{--}770\text{ cm}^{-1}$ at mixing ratios of CO_2 at 375 ppm (black) and 750 ppm (red). The blue dotted curve is the emission spectrum expected for a 300 K blackbody radiator and the green dotted line is for a 215 K emitter. Adapted from calculations based on the Web site: <http://geoflop.uchicago.edu/forecast/docs/Projects/modtran.orig.html>.

changes in the wings seem tiny, but the integral over the ditch (and small contributions elsewhere) reveal (the Chicago Web site simulator tells us) that the infrared radiation is reduced by 4.4 W m^{-2} .

We show one more experiment where low concentrations of CO_2 are doubled twice as shown in Figure 5. Again the range of wave numbers in the figure spans the CO_2 ditch. In this case, the emission from the ditch decreases with CO_2 increases and the greenhouse effect spreads all across the ditch including both the floor and the wings. This is to be contrasted with the higher initial concentrations experiment of Figure 4. We can now explain the reversal of change in the floor of the ditch compared to the wings. In the lower concentration cases, the brightness temperature is well below the tropopause (except for the tiny spike in the very center of the ditch where the very

strong absorption feature lies). As the concentration is increased, the strong absorption spike at the center shallows (the spike grows), but the floor of the ditch deepens because the radiation to space is originating from higher and cooler layers of the troposphere. When the concentration becomes large enough the brightness temperature actually reaches a minimum and even turns around as the level of last radiation enters the inverted temperature profile of the lower stratosphere. It is interesting that the brightness temperature never falls to 195 K, which is the minimum in the temperature, as seen in Figure 1. This is because the distribution of radiation origins in the continuous vertical profile is from a 'smear' or integrated aggregate of infinitesimal levels of the order of one optical depth and its effective value is not at the actual minimum of the distribution.

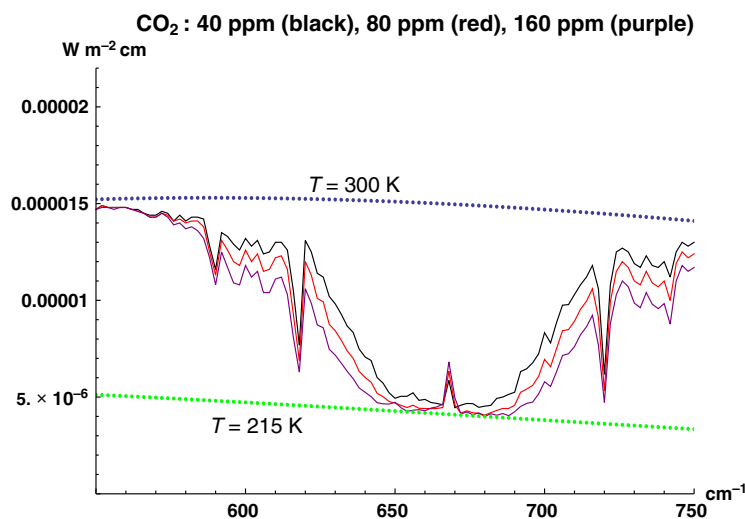


Figure 5 The outgoing infrared spectrum in the range $550\text{--}770\text{ cm}^{-1}$ at mixing ratios of CO_2 at 40 ppm (black), 80 ppm (red), and 160 ppm (purple). The blue dotted curve is the emission spectrum expected for a 300 K blackbody radiator and the green dotted line is for a 215 K emitter. Adapted from calculations based on the Web site: <http://geoflop.uchicago.edu/forecast/docs/Projects/modtran.orig.html>.

So far in these exercises, we have kept the surface temperature fixed. That is, we have not forced the incoming absorbed rate to balance the outgoing. Adjusting the surface temperature to come into exact balance would not be proper because the incoming and outgoing rates do not balance in a finite width latitude belt because transport of heat to (and from) neighboring latitude belts contributes to the energy balance. Although we seek to know the outcome if CO₂ is doubled, we will not be able to do this exactly because it is the entire globe that is to be balanced, not just the tropics. Nevertheless, we can do an approximate job of it by adjusting the temperature upward in the column of air until the outgoing radiation is the same as before. When we adjust the ground temperature, we will hold the water vapor fixed at zero mixing ratio. (Actually, the MODTRAN code used here has some water vapor above about 12 km but it is held fixed during any changes.) We can iteratively change the ground temperature incrementally until the outgoing radiation is restored to the value it had when CO₂ was at 375 ppm. After this adjustment we find that the required temperature increase is about 1.10 K, a value within a few percent for the global average change found in LBL calculations (Myhre et al., 1998).

We can do a few more cases to improve our insight into the changing greenhouse effect due to doubling of CO₂. First consider the MODTRAN simulation when clouds are present (but still no other GHGs, slightly absurd since clouds are made of water droplets). The result is that for cloudy atmospheres in the tropics, we simply replace the broad band ground emission in Figure 3 with the temperature at cloud top – the ditch remains unaltered. In other words, the cloud tops in most cases are much lower than the emission level of the CO₂. The change in surface temperature to restore the outgoing radiation to its lower CO₂ concentration is left unchanged from the cloud-free case if we assume that the changes in temperature aloft are carried all the way through the clouds to the ground. We can use MODTRAN in another series of experiments with summer and winter middle latitude conditions. Perusing these cases (with no GHGs other than CO₂) reveals very little difference from the tropical case. We conclude that given the approximations inherent in MODTRAN the change in temperatures will be of the order of between 0.9 and 1.1 K for a doubling of CO₂ from 375 to 750 ppm, and since the dependence on CO₂ mixing ratio is logarithmic, we can expect the doubling effect to be about the same for doubling from any base level. (Some authors prefer to start with preindustrial levels of 250 ppm.)

Summary of Assumptions

We have executed some thought experiments leading us to believe that the increase in global average temperature due to a doubling of CO₂ is about 1.0(±0.1) °C. We have cautioned that the physical model and the approximations to it are pretty schematic but have pedagogical value. We have to be careful not to apply the quantitative result to the real world without reviewing the assumptions that went into our calculation.

1. We used clear sky only in the calculations. While the presence of fixed (in their fractional coverage and altitudes) clouds (cloud tops are below the emission levels of CO₂ and

water vapor) would not have an appreciable effect on the response to doubling CO₂, the assumption of ‘fixed’ is not likely to hold. Cloud feedback processes are among the most challenging problems facing the climate science community. Much of the difficulty stems from the fact that many cloud processes are at smaller or comparable scales to the grid spacing of our climate models. But clouds are also hooked to the larger scales of the general circulation of the atmosphere, for example, the midlatitude storm belts, which might be undergoing secular change with global warming.

As mentioned earlier, the horizontal distribution of water vapor is nonuniform and temporally variable. The precise effect of this phenomenon is still a matter of ongoing research. For example, how well do we need to incorporate such space–time nonhomogeneities in our climate models.

2. Following the Chicago Web site, we adjusted the temperature ‘rigidly’ up the whole column when we changed the surface temperature. This is probably not what actually happens in the air column. The lapse-rate profile might change as the surface temperature is raised and this is likely to lower the response. This effect is called the lapse-rate feedback and it is probably negative.
3. In increasing the temperature to compensate for the reduced outgoing radiation we ignored that fact that the atmospheric column will now hold more water vapor because of the strong dependence of saturation vapor pressure on air temperature. That roughly constant relative humidity is supported by climatology and model simulations. This positive feedback is likely to be strong, possibly increasing the response by a factor of two. Lapse-rate and water vapor feedbacks are anticorrelated, but water vapor appears to be much stronger based on climate model simulation studies.
4. Other known feedbacks such as those due to snow and ice cover are also ignored. These are positive feedbacks but are thought to be smaller than the combination of water vapor, lapse-rate, and cloud feedbacks.
5. As the CO₂ increases, changes in the temperature of the stratosphere will occur along with those of the troposphere. Generally to maintain balance of air layers in the stratosphere, the temperature there will have to decrease during greenhouse warming below. This comes about because convection mixes the air in the troposphere, but the stratospheric layers are not coupled to the troposphere through convection. Basically, the stratosphere does not know the troposphere is warming as regards convective overturning, so the cooling in the stratosphere due to increased CO₂ leads to a lowering of stratospheric temperatures.
6. We ignored the rest of the planet. Middle and higher latitudes may have quite different sensitivities to GHG concentrations. If the tropics exhibit the largest sensitivity, this will have to be mixed with that of less sensitive latitude belts.
7. There are likely to be slow feedbacks in the system that alter the composition of the atmosphere including its GHGs. These feedbacks may take decades or even centuries to kick in, as permafrost is melted or GHGs are released from sources deep in the oceans and terrestrial biosphere.

8. A final consideration. [Figure 4](#) shows the differential of radiation as a function of wave number across the ditch. Much here depends on the height of the tropopause on the year, month, or even the day of the occurrence. This is one source of variability of the sensitivity of climate. We could also ask whether a steadily increasing altitude of the tropopause due to global warming represents an additional feedback in the system.

Climate feedback mechanisms are considered in more detail in article [Climate and Climate Change: Climate Feedbacks](#) in this Encyclopedia.

While the exercises presented above are instructive, we cannot end the story at this point. The only way to accurately simulate the response to CO₂ is through a general circulation model of the atmosphere. This is a formidable task given that many of the important processes are not yet well represented in the models.

Conclusion

The Earth's surface temperature has been rising steadily over the last century. Most of the potential drivers of climate change (volcanic activity, solar brightness variability, atmospheric aerosols, and greenhouse gas concentration increases) have been examined in great detail in recent years. We now have reasonably good estimates of the strengths and time dependences of these drivers and although much needs to be done in substantiating these assertions, virtually all have been eliminated except for the increasing influence of GHGs. The importance of this driver is also consistent with paleoclimate evidence. We have shown a series of pedagogical computer experiments that provide estimates of the response in the tropics to doubling the CO₂ concentration while excluding all feedbacks. These experiments can be repeated by the reader by going to the Chicago Web site. While virtually all experts on the radiative aspects of climate science would agree with the values we have obtained, we have to acknowledge that many additional physical effects will come into play along with intensification of any of the primary drivers. Some of these additional physical effects (feedbacks) are likely (at this writing) to amplify the response to CO₂ doubling to a value perhaps as much as four times.

One more important effect not considered here is the time dependence of the response. The illustrations used here were from one equilibrium climate state to another equilibrium state. The Earth system has a number of effective heat storage components that have varying effective heat capacities. Among these are the atmospheric column with a response time of about a month, the oceanic mixed layer with a response time of a few years, continental glaciers with centuries of response time, and finally the deep ocean which communicates with the surface waters through small passageways that limit the flow of heat toward the deeper parts of the world ocean. Many parts of the deep ocean have not touched the surface in 800 years. The upshot of this is that while climate sensitivity is an important index for comparing one atmospheric model with another, the global system has many interlocking parts that cause the response of the surface to be delayed in its full response by perhaps hundreds of years. This means that even if we were to stop or reverse the rate of greenhouse gases entering the atmosphere, the reversal of the response is likely to be delayed by these sluggish and nearly inaccessible components.

See also: [Climate and Climate Change: Carbon Dioxide](#); [Climate Feedbacks](#). [Solar System/Sun, Atmospheres, Evolution of Atmospheres](#): [Solar Terrestrial Interactions: Climate Impact](#).

Further Reading

- Archer, D., Pierrehumbert, R.T. (Eds.), 2011. *The Warming Papers*. Wiley Publishers, New York, 432p.
- Manabe, S., Wetherald, R.T., 1975. Effects of doubling the CO₂ concentration on the climate of a general circulation model. *Journal of the Atmospheric Sciences* 32, 3–15.
- Myhre, G., Highwood, E.J., Shine, K.P., Stordal, F., Stordal, F., 1998. New estimates of radiative forcing due to well mixed greenhouse gases. *Geophysical Research Letters* 25, 2715–2718.
- Pierrehumbert, R.T., 2011a. *Principles of Planetary Climate*. Cambridge University Press, New York, 652p.
- Pierrehumbert, R.T., Jan 2011b. Infrared radiation and planetary temperature. *Phys. Today* 64, 33–38.
- Weart, S., 2009. *The Discovery of Global Warming*, second ed. Harvard University Press, Cambridge, MA, 240p.

History of Scientific Work on Climate Change

S Weart, Center for History of Physics, American Institute of Physics, College Park, MD, USA

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

People understood since antiquity that climate could change locally, but nobody imagined that human activity could alter global climate. In 1896 Arrhenius and in the 1930s Callendar promoted a theory of anthropogenic global warming caused by a buildup of carbon dioxide. Few found this plausible until the mid twentieth century, when new theoretical work and observations showed that the concept had to be taken seriously. In the 1970s, complex computer models suggested a risk of several degrees of warming in the following century. The temperature–carbon dioxide connection was supported by the paleontological evidence from deep ice cores. Uncertainties about the effects of aerosol pollutants and clouds complicated the issue. Nevertheless, around the start of the twenty-first century, improved computer and paleontological arguments, plus a prolonged observed warming of the planet, led to consensus statements by international panels that warned of severe impacts.

The nineteenth century discovery of ice ages in the distant past proved that climate could change radically and on a global scale. Could human civilization have any influence on such processes? From ancient times, people had suspected that their activities could change the climate locally. Scholarly theories and folk beliefs speculated that chopping down a forest, irrigating a desert, or draining marshlands might change the temperature and rainfall in the vicinity. Americans in the nineteenth century argued that settlement of the country had brought a less savage climate, and farmers who moved onto the Great Plains boasted that ‘rain follows the plough.’ Some European scientists, however, argued that deforestation made for a drier, not wetter, climate. By the end of the nineteenth century, meteorologists had accumulated enough reliable weather records to test whether rain follows the plough or flees from the axe. Both ideas failed the test. Even the transformation of the entire ecosystem of eastern North America from forest to farmland had apparently made little difference to climate. If the spectacular changes wrought by humankind could not alter a region’s climate, there seemed still less reason to imagine that they could alter the planet as a whole.

Through the first half of the twentieth century, scientists who studied climate treated not only humanity but the entire biological system as passive. Deserts and forests expanded or shrank in helpless response to climate changes. The cause of these changes might be, for example, the diversion of winds when mountain ranges arose or eroded away, or an alteration of ocean currents when some isthmus opened or closed, or variations of the Sun’s heat, or a spate of volcanic eruptions pouring smoke into the air. Whatever the cause, it was surely a force far mightier than anything done by the meter or so of organic matter that covered some patches of the planet’s surface. Still less could the tiny activities of humanity be compared with the grand forces of geophysics. The scale of time was also grand, for scientists thought that ice ages and other significant changes unfolded over many tens of thousands of years. Within the time span of a few human generations, climate was thought to be held stable by a benign ‘balance of nature.’

One theory for the cause of climate changes looked at the composition of the atmosphere.

In memoirs published in 1824 and 1827, Joseph Fourier calculated the temperature of a body at Earth’s distance from the Sun: he got a temperature well below freezing, much colder than the actual planet. The difference, Fourier recognized, was due to the atmosphere. Somehow it kept part of the heat radiation in. He tried to explain this by comparing Earth with its covering of air to a box with a glass cover. Visible light warms the surface, but the atmosphere (like the glass) prevents the invisible heat energy from escaping. This was the effect that would later be called, by an inaccurate analogy, the ‘greenhouse effect.’

John Tyndall, curious about the cause of ice ages, set out to find whether there was in fact any gas in the atmosphere that could trap heat rays (infrared radiation). In 1859, his careful laboratory work identified several gases that did just that. The most important was simple water vapor. Also, effective were carbon dioxide (CO₂) and methane (CH₄). Tyndall likened these gases in the atmosphere to a dam built across a river. Just as the dam made the water deeper, he explained, so the atmosphere, by blocking outgoing radiation, raises the temperature at Earth’s surface.

Some speculated that ice ages began and ended when volcanoes became less or more active in emitting such ‘greenhouse gases.’ In 1896, Svante Arrhenius published laborious calculations, which he believed showed that a severe drop in atmospheric CO₂ would in fact bring a lowering of temperature sufficient to cause an ice age. He went on to a new idea: as humanity burned fossil fuels such as coal, enough CO₂ was added to Earth’s atmosphere to influence the climate. If the concentration of the gas in the atmosphere doubled (which Arrhenius figured would take thousands of years), the planet’s average temperature would be raised by a few degrees. In retrospect it was shown that Arrhenius’ calculations were inadequate; he had not proved that greenhouse gases would significantly change climate, but he had at least shown that this was a possibility.

Arrhenius’ idea was only one of many speculations about climate and not the most convincing. Scientists found good reason to believe that the emissions could not change global climate significantly. For one thing, experiments and theoretical considerations seemed to show that the absorption of

infrared by CO₂ in the atmosphere was saturated: adding more gas would make no difference since the radiation was already fully blocked. For another, the oceans would absorb CO₂ as fast as industry emitted it.

In the 1930s, old people began to notice that the United States and the North Atlantic region had warmed significantly since their childhoods. Meteorological records confirmed it. Scientists supposed this was just a phase of some mild natural cycle, with unknown causes. Only one lone voice, that of the engineer Guy Stewart Callendar, insisted that greenhouse warming was on the way. He had evaluated the old measurements of atmospheric CO₂ concentrations and concluded that over the past hundred years the concentration of the gas had increased by about 10%. This rise, Callendar asserted in 1938, could explain the observed warming. He argued that even if the CO₂ presently in the atmosphere thoroughly blocked outgoing infrared radiation, adding more gas must make for more surface warming, by raising to a higher and colder level the point in the atmosphere from which radiation ultimately escaped. His arguments failed to convince meteorologists.

In the 1950s, Callendar's claims provoked few scientists to look into the greenhouse effect with improved techniques and theories. Roger Revelle and colleagues studied the absorption of CO₂ by the oceans. In 1957, he demonstrated that, with human population and industry both climbing exponentially, the oceans could not in fact absorb the gas fast enough to keep it from building up in the atmosphere. Charles David Keeling drove home the point in 1961 with meticulous measurements showing that the concentration of the gas in the air was in fact rising, year by year, as Callendar had claimed.

Meanwhile, Gilbert Plass took up the challenge of calculating the transmission of radiation through the atmosphere and nailed down the fact that adding more CO₂ would markedly increase the interference with infrared radiation. He calculated that doubling the concentration – which he thought would take a few centuries to happen – would bring a 3–4 °C average rise. But Plass's calculation, like Arrhenius's and others up to that time, did not take into account the possible changes in cloudiness and many other features of the real atmosphere. It could not be accepted as a reliable calculation of the greenhouse effect.

Over the next decade, a few scientists devised simple mathematical models of the climate. The models included feedbacks that made the system surprisingly variable. For example, if a little extra warmth melted northern snow and ice earlier in the spring, the dark soil and water would absorb sunlight and warm up the Arctic still more. Other scientists figured out ingenious ways to retrieve past temperatures by studying ancient pollen in bogs and fossil shells in the clay of sea floors. The data showed that grave climate changes had happened within as little as a few centuries.

This finding was reinforced by computer models of the general circulation of the atmosphere, the fruit of a long effort to learn how to predict the weather. Applying basic physics equations to calculate the climate, by the mid 1960s these models were able to roughly reproduce trade winds, desert regions, Arctic ice, and other features of the planet. A 1967 computation by Syukuro Manabe and colleagues incorporated clouds, convection of heat to upper levels of the atmosphere, and other crucial features neglected in the simple

earlier models. Their results suggested that doubling the concentration of CO₂ – which was now seen to be likely by the end of the twenty-first century – would raise the planet's average temperature by a few degrees. It was understood, however, that much more work would be needed before such results could be relied upon.

In the early 1970s, a series of unusual droughts and food shortages helped turn curiosity about the climate into anxious concern. Alongside greenhouse gases, Reid Bryson and some other scientists pointed out that human activity was putting ever more dust and smog aerosols into the atmosphere. Perhaps this haze would block sunlight and cool the world. Meanwhile, painstaking compilation and analysis of Northern Hemisphere weather statistics showed that a cooling trend had begun in the 1940s. Scientists were uncertain, sometimes predicting a balmy globe with coastal areas flooded as the ice caps melted and sometimes seeing a risk of a catastrophic new ice age. Study panels, first in the United States and then in Europe, warned that some kind of future climate change might pose a severe threat. The only thing most scientists agreed on was that they did not yet understand the climate system well enough to predict what kind of dangerous changes might be in store.

Research activity accelerated, including huge data-gathering schemes that mobilized international fleets of oceanographic ships and orbiting satellites. As computer models improved, by the late 1970s they showed clearly that warming, not cooling, was likely in the future. Inspecting models devised by Manabe's team and a team under James Hansen, a panel of the National Academy of Sciences concluded in 1979 that greenhouse warming was potentially serious. The panel said that they had rather high confidence that a doubling of CO₂ would warm up Earth by about 3 °C, plus or minus 50%: in other words, 1.5–4.5 °C. Because of the complexities of the climate system, subsequent work has been unable to narrow the range. But at the time the range itself was regarded as highly uncertain; many scientists continued to doubt that there would be any greenhouse-effect global warming at all. Hansen and others predicted that the warming would rise above the noise level to become unequivocally visible sometime around the start of the twenty-first century.

Scientists had sought a single master key to climate, but now they were coming to understand that climate is an intricate system responding to many influences. Volcanic eruptions and solar variations were still plausible causes of change, and some argued these would swamp any effects of human activities. Subtle changes in Earth's orbit could also make a difference. Studies of ancient climates, as revealed in cores from the deep sea floor, showed that minor astronomical cycles (calculated in the 1930s by Milutin Milankovitch) had set the timing of the ice ages. To the surprise of many, the climate was apparently so delicately balanced that almost any small perturbation might set off a great shift. According to the new 'chaos' theories, a shift in such a system might be sudden and irreversible. Support for the idea came from ice cores arduously drilled from the Greenland and Antarctic ice sheets. They showed large and disconcertingly abrupt regional temperature jumps in the past, taking not centuries but mere decades.

Computer models confirmed that such severe climate shifts were possible, for example if greenhouse warming caused

a change in the pattern of ocean circulation. Even without that, experts predicted that warming would bring a risk of greater heat waves and droughts, stormy rains and floods, rising sea level, and other harms. However, the modelers had to make many arbitrary assumptions about clouds and the like, and some reputable scientists continued to dispute the reliability of the results.

One unexpected discovery was that the concentrations of certain greenhouse gases such as methane were rising fast. Trace gases derived from human activities were significantly blocking bands of the infrared that had formerly been transparent. According to the calculations published by a group under Veerabhadran Ramanathan in 1980, the cumulative effects could bring on global warming twice as fast as had been expected. Moreover, by 1980 laborious compilations of average global temperatures showed that these had begun to rise again. International panels of scientists began to warn that the world should take active steps to restrain fossil fuel emissions.

Concerns were sharpened by new evidence from the Greenland and Antarctic ice caps. The long cylinders of ice extracted by drillers contained tiny bubbles with samples of ancient air – by good fortune there was this one thing on the planet that preserved CO₂ intact. In 1980, a team published findings that were definite, unexpected, and momentous. In the depths of the last ice age, the concentration of CO₂ in the atmosphere had been as much as 50% lower than in our own warmer times. (These Greenland measurements were later called into question, but the dramatically lower ice-age level was quickly confirmed by other studies.) Pushing forward, by 1985, a French-Soviet drilling team at Vostok Station in central Antarctica produced an ice core 2 km long that carried a 150 000-year record, a complete ice age cycle of warmth, cold, and warmth. They found that the concentration of atmospheric CO₂ had gone up and down in remarkably close step with temperature. These findings marked a turning point, convincing many scientists that greenhouse warming must be taken seriously.

The scientists' worries first caught wide public attention in the summer of 1988, which was the hottest on record till then. (Most years since were hotter.) But the many scientific uncertainties, and the sheer complexity of climate, gave space for a vehement debate. Individuals who opposed government restraints on fossil fuels argued, for example, that the warming seen so far was caused by a coincident rise in solar activity. However, solar activity subsequently leveled off and then declined, but the planet continued to warm up.

In 1988, the world's governments created a novel institution, the Intergovernmental Panel on Climate Change (IPCC), to give them the most reliable possible advice. Following years of discussion among all the world's climate experts, any IPCC conclusions would have to be approved by a consensus of officials appointed by all the world's governments. During the 1990s, the scientists worked through a large number of new studies, notably a proliferation of improved computer models. These now included computation of realistic ocean circulation and the effects of aerosol emissions. By 2001, the modelers had simulated the detailed pattern of geographical and vertical

distribution of atmospheric heating expected from greenhouse warming. The pattern of change was different from the patterns that other influences (for example, changes in solar output) would produce. The computed greenhouse-effect 'signature' roughly matched the actual observational record of recent decades. Warming was greatest in the Arctic, for example, while the stratosphere was cooling as predicted.

Paleoclimate studies were also eloquent. The data from ice cores were reinforced by studies going back hundreds of millions of years. These studies confirmed, by wholly independent means, the computer modelers' finding that a doubling of CO₂ went along with a rise of mean global temperature of 3 °C, give or take a degree or two. Meanwhile, a 1999 reconstruction of a millennium of Northern Hemisphere temperatures by Michael Mann and colleagues showed that the warming of recent decades markedly exceeded anything reliably reported in the historical past. (This 'hockey stick' reconstruction came under severe attack, but the basic finding was confirmed by subsequent studies.)

In 2001, the IPCC managed to establish a meaningful consensus, phrased so cautiously that scarcely any expert dissented. The panel announced that although the climate system was so complex that scientists would never reach complete certainty, it was much more likely than not that our civilization faced severe global warming. At this point, scientists knew the most important things about how the climate could change during the twenty-first century. How the climate actually would change now depended chiefly on what policies humanity would adopt to restrain its emissions.

After 2001, improved computer models and an abundance of data of many kinds strengthened the conclusion that human emissions were very likely to be a main cause of the recent global warming. In 2007, the IPCC warned that, depending on what policies people adopted for emissions, by the end of the twenty-first century, one could expect the planet's average temperature to rise anywhere between about 1.4 and 6.5 °C. Although only a small fraction of this warming had happened so far, the predicted effects were already visible in some regions. The impacts were coming sooner and more severely than expected: more deadly heat waves, worse droughts and floods, and the migration or decline of sensitive species. If the emissions are not restricted, the panel warned that in coming centuries the seas would rise by meters to drown coastal cities and entire ecosystems on which civilization depended would be impoverished.

See also: Clouds and Fog: Climatology.

Further Reading

- Archer, D., Pierrehumbert, R.T. (Eds.), 2011. *The Warming Papers: The Scientific Foundation for the Climate Change Forecast*. Wiley-Blackwell, Hoboken, NJ.
- Fleming, J.R., 1998. *Historical Perspectives on Climate Change*. Oxford University Press, New York.
- For a list of additional historical books and articles, see <http://www.aip.org/history/climate/links.htm#hist>.
- Weart, S., 2008. *The Discovery of Global Warming*, second ed. Harvard University Press, Cambridge, MA.

Intergovernmental Panel on Climate Change

KE Trenberth, National Center for Atmospheric Research, Boulder, CO, USA

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

An outline is given of the IPCC and its role in helping to lay out options and consequences of actions or lack thereof related to anthropogenic climate change. A description is provided of how the IPCC is set up and functions, the scientific assessment process, who participates, and the intergovernmental linkages, all of which lead to the IPCC reports every 5–7 years. The procedures are designed to provide policy relevant but not policy prescriptive scientific advice to policy makers and the public. The main findings are briefly described along with the recent controversy involving IPCC.

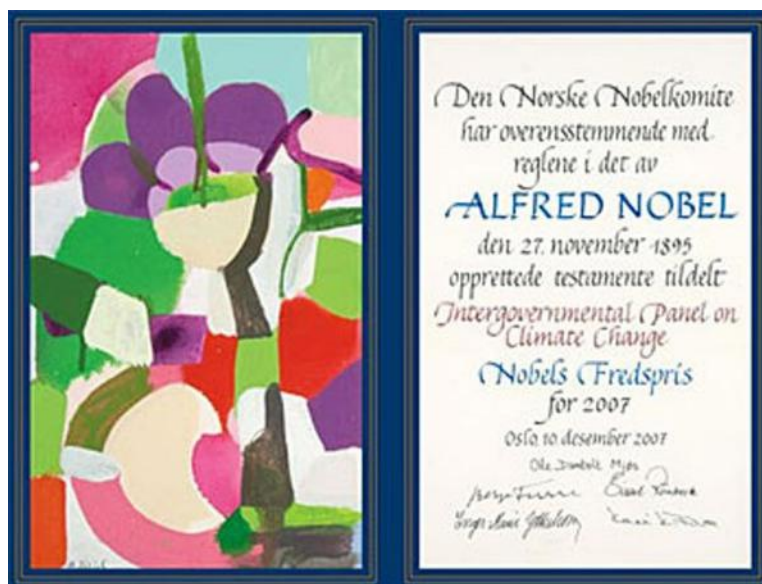
In 2007, the Fourth Assessment Report of the Intergovernmental Panel on Climate Change (IPCC), known as AR4, clearly stated that 'Warming of the climate system is unequivocal' and it is 'very likely' due to human activities. Later in 2007, the IPCC won the Nobel Peace Prize, jointly with Al Gore Jr. 'for their efforts to build up and disseminate greater knowledge about man-made climate change and to lay the foundations for the measures that are needed to counteract such change.' This article describes the IPCC and its role in society and politics.

The IPCC is an international organization that includes a panel of governments and a body of scientists from around the world convened by the United Nations jointly under the United Nations Environment Programme and the World Meteorological Organization (WMO) and initiated in 1988. Its mandate is to provide policy makers with an objective assessment of the scientific and technical information available about climate change, its environmental and socioeconomic impacts, and possible response options. The IPCC reports on all aspects of the science of global climate and the effects of human activities on climate in particular. Major assessments were made in 1990, 1995, 2001, and 2007 and are called the First, Second, and Third Assessment Reports (FAR, SAR, and

TAR) and AR4 for the Fourth Assessment Report. The Fifth Assessment report (AR5) was released in 2013.

Each new IPCC report reviews all the published literature over the previous 5–7 years and assesses the state of knowledge, while trying to reconcile disparate claims and resolve discrepancies, and document uncertainties. There is a 'technical summary' and a short 'summary for policy makers' (SPM) and the volume from each of the three working groups (WGs) has tended to run to about 1000 pages or so.

WG I deals with how the climate has changed and the possible causes. It considers how the climate system responds to various agents of change and the ability to model the processes involved as well as the performance of the whole system. It further seeks to attribute recent changes to the possible various causes, including the human influences, and thus it goes on to make projections for the future. WG II deals with impacts of climate change, vulnerability, and options for adaptation to such changes, and WG III deals with options for mitigating and slowing the climate change, including possible policy options. Each WG has two coauthors, one from a developing country and one from a developed one. Each WG is staffed by a small Technical Support Unit that is hosted by one of the coauthors of the WG.



The IPCC also includes a task force on National Greenhouse Gas Inventories (TFI) to oversee the National Greenhouse Gas Inventories Programme and to develop and refine an internationally agreed methodology and software for the calculation and reporting of national greenhouse gas emissions and removals and to encourage its use by parties of the United Nations Framework Convention on Climate Change (UNFCCC). The task group on Data and Scenario Support for Impacts and Climate Analysis was established to facilitate cooperation between the climate modeling and climate impacts assessment communities by increasing the availability of climate change-related data and scenarios for climate analysis and impacts, adaptation, vulnerability, and mitigation research.

The IPCC is an intergovernmental body. It is open to all member countries of the United Nations and WMO. Currently, 194 countries are members of the IPCC. Governments participate in the review process and the plenary sessions, where main decisions about the IPCC work program are taken and reports are accepted, adopted, and approved. The IPCC bureau members, including the Chair, are also elected by governments during the plenary sessions. The IPCC Secretariat in Geneva coordinates all the IPCC work and liaises with governments. Thousands of scientists from all over the world contribute to the work of the IPCC on a voluntary basis.

The IPCC bureau comprises the IPCC chair, the IPCC vice-chairs, the co-chairs and vice-chairs of the WGs, and the co-chairs of the task force. The IPCC bureau is chaired by the IPCC chair. The purpose of the bureau is to provide guidance to the panel on the scientific and technical aspects of its work, to advise on related management and strategic issues, and to take decisions on specific issues within its mandate, in accordance with the principles governing IPCC work. Coordinating lead authors (CLAs) and lead authors (LAs) are selected by the relevant WG or task force bureau from those experts cited in nominations provided by governments and participating organizations, and other experts as appropriate, known through their publications and works, and taking into account both geography and gender. None of them is paid by the IPCC.

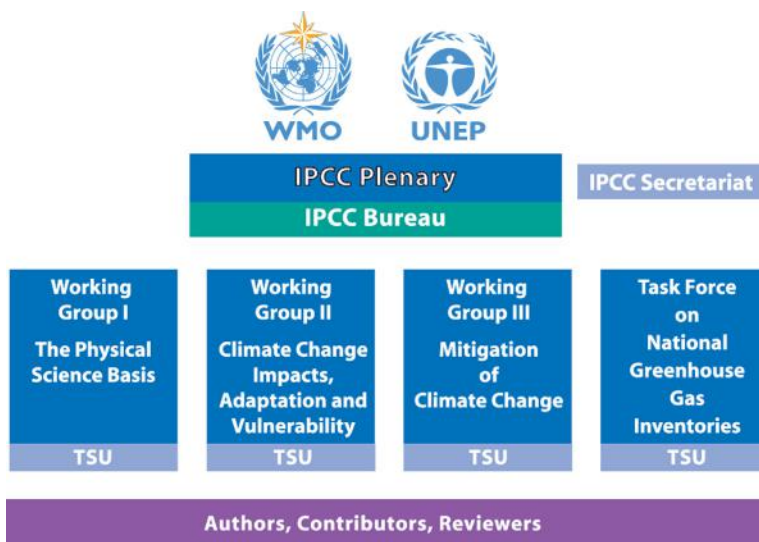
The IPCC starts a new assessment with a scoping process among experts, which leads to the proposed general outline of a report, highlighting also new features and cross-cutting matters. Policy makers and other users of IPCC reports are consulted in order to identify the key policy-relevant issues along with experienced scientists who can attest to the state and capabilities of the science. The outlines are subject to formal approval by the panel before work begins.

CLAs take overall responsibility for coordinating a major section of a report and the chapter that they are responsible for. For the first four reports, there were usually two CLAs per chapter. LAs work in teams to produce the content of the chapter. LAs may enlist contributing authors who provide additional technical information on specific subjects covered by the chapter. LAs are responsible for the production of designated sections of chapters. The essence of the LA task is to synthesize the scientific, technical, and socioeconomic information available in peer-reviewed and internationally available literature and in selected non-peer-reviewed literature.

Along with comprehensive assessment reports, the IPCC has produced several special reports and technical papers on topics of interest, as well as methodology reports. Many of these reports are prepared in response to requests from the UNFCCC or from other international organizations and conventions.

Each WG is made up of participants from the United Nations countries, and for the 2007 AR4 assessment, there were over 450 LAs, 800 contributing authors, and over 2500 reviewers from over 130 countries who provided over 90 000 review comments. The IPCC process is very open. Two major reviews are carried out in producing the report, a first review by experts and a second review by experts and governments. Climate 'skeptics' can and do participate, some as authors. All comments are responded to in writing and result in many changes in the report. The process is overseen by two or more review editors for each chapter.

As an example for the AR4 in WG I, there were 11 chapters and the report was 996 pages plus supplementary material



online. There were 140 LAs, hundreds of contributors, and two or three review editors for each chapter (26). There were also over 700 reviewers. For example, for Chapter 3 on observations, the CLAs were Kevin E. Trenberth and Philip D. Jones; there were 10 other LAs and 66 contributing authors. The published chapter ran to 101 pages plus online supplementary material, 47 figures (126 panels), 8 tables, and 863 references, making it the longest chapter in the report. In the expert scientific review, there were 2231 comments and another 1270 comments in governmental review, for a total of 3501 comments. Every comment and the writer were entered into a huge spreadsheet along with the response and actions taken in terms of changing the text.

Summaries for policy makers are prepared after the first expert review rather than concurrently with the main reports. They undergo one round of expert and government review. Each point undergoes not only the careful scrutiny of the scientists but also government officials and nongovernmental organizations. A second opportunity for comments by governments takes place as input to the approval process. The SPM is then approved line-by-line by governments in a major meeting, which takes place over 3 or 4 days. Those participating include government representatives (typically more than 120 countries are present), several nongovernmental organizations as observers, and typically 40–50 scientists. Simultaneous translation occurs throughout the meeting into English, French, Spanish, Russian, Chinese, and Arabic, per UN practice.

United Nations rules require a unanimous consensus to be sought. Negotiations occur over wording to ensure accuracy, balance, clarity of message, and relevance to understanding and policy. The strength is that it is a consensus report but the process also makes it a conservative report. The rationale is that the scientists determine what can be said, but the governments determine how it can best be said.

The IPCC process is dependent on the good will of the participants in producing a balanced assessment. However, by the time of the SAR, it appeared that there were attempts to blunt, and perhaps obfuscate, the messages in the report by certain nations. This led to protracted debates over wording on even bland and what should be uncontroversial text. The result was that the schedule of the meetings quickly became disrupted so that extra evening sessions were scheduled. During the limited breaks that did occur, authors of the report worked with delegates in side meetings to craft revised text for submission to the plenary. In many ways the meeting became one of endurance. In spite of these trials and tribulations, the result is a reasonably balanced consensus summary. However, the SPM did tend to grow in size relative to the initial draft length during the course of the meeting and usually became more technical than desirable.

The role of the IPCC is to provide policy relevant but not policy prescriptive scientific advice to policy makers and the general public. IPCC scientists, with all kinds of value systems, ethnic backgrounds, and from different countries, gather together to produce the best consensus description of what they jointly understand and with appropriate statements about confidence and uncertainty. The strength of the IPCC report is not just the solid scientific credentials but also the open process by which it is created.

Some Findings

The scientific evidence brought up by the first IPCC Assessment Report of 1990 unveiled the importance of climate change as a topic deserving a political platform among countries to tackle its consequences. It therefore played a significant role in leading to the creation of the UNFCCC, the key international treaty to reduce global warming and cope with the consequences of climate change. The UNFCCC was ratified in 1994 by 194 countries.

The SAR was the time when the IPCC first attracted more focused attention of policy makers in a major way when the statement coming out of the intergovernmental meeting and the SPM was ‘The balance of evidence suggests a discernible human influence on global climate.’ This statement was carefully crafted after much debate and compromise. The SAR provided key input into the adoption of the Kyoto Protocol in 1997. The Kyoto Protocol was finally ratified in February 2005 by 164 countries, and eventually by Australia, but not by the United States.

In the TAR, WG I presented an improved understanding of climate processes, forcing agents, and feedback and addressed the question of human influence on climate. Since the SAR, the evidence became much stronger – from the continuing record warmth, the improved paleorecord that provides context, improved modeling and simulation of the past climate and improved statistical analysis. Projections of future climate were based on new scenarios and a wider range of models. WG II updated impacts, vulnerabilities, coping strategies and adaptation, and implications for sustainable development. WG III assessed mitigation options, their costs and co-benefits as well as barriers, opportunities, and policy instruments. It also placed climate change mitigation in the context of sustainable development. The most contentious paragraph in the TAR WG I SPM was the concluding one on attribution. After much debate the following was carefully crafted in the summary statement for the TAR: “There is new and stronger evidence that most of the warming observed over the last 50 years is attributable to human activities.”

In AR4, WG I provided a new knowledge on human and natural drivers of climate, a detailed assessment of past paleoclimate changes and causes, and stronger evidence on attribution of climate change; it also included some regional aspects and observations of trends in other variables such as ocean warming, temperature extremes, and wind patterns. WG II further assessed observational evidence of impacts of climate changes and identified vulnerable places and people. Projections were made concerning impacts of future warming trends, taking into consideration different possible developments and stresses from other effects of global change. WG III further evaluated emissions trends, mitigation options, and how stabilization of greenhouse gas concentrations in the atmosphere might be achieved, along with associated costs in the near and longer term. Sustainable development policies were considered along with the relationship between mitigation and adaptation. A more consistent evaluation of uncertainty and risk was attempted. AR4 WG I carefully stated that “Warming of the climate system is unequivocal as is now evident from observations of increases in global average air and ocean temperatures, widespread melting of snow and ice, and rising

global average sea level” and Most of the observed increase in global average temperatures since the mid-20th century is very likely due to the observed increase in anthropogenic greenhouse gas concentrations. Here ‘very likely’ is deemed to mean more than 90% probability of occurrence.

AR5 will put greater emphasis on assessing the socioeconomic aspects of climate change and implications for sustainable development, risk management, and the framing of a response through both adaptation and mitigation. It will aim to provide more detailed information on regions, including on climate phenomena such as monsoons and El Niño.

Controversy

Following the SAR in 1995, a controversy broke out after completion of the report because of publicized charges (*Wall Street Journal*) that changes were made by the convening LA of Chapter 8, Dr. Ben Santer, to ‘deceive policy makers and the public into believing that the scientific evidence shows human activities are causing global warming.’ Santer indeed made changes in the chapter as instructed and as an outcome of the intergovernmental approval meeting and associated changes in the SPM. The charges were completely unfounded, but the lack of IPCC oversight of this process led to the introduction of review editors in subsequent reports.

Following the TAR, the main controversy centered on the paleoclimate reconstruction of northern hemisphere temperature for the last millennium that has the shape of an ice hockey stick with the blade as the upturn in the twentieth century. It has been used by some as a symbol of the human influence on climate. Dr. Michael Mann was the leader of the group that produced the ‘hockey stick’ and has come under considerable but unjustified criticism.

After the 2007 IPCC reports, in the lead up to the political Conference Of the Parties (COP) meeting in Copenhagen in December 2009 (COP-15), some controversies broke out that involved the IPCC. Disparagingly called ‘climate gate,’ some emails and personal information about individuals were illegally taken from the University of East Anglia through a hacking incident. Subsequently, there was selective publication of some of the stolen emails taken out of context and charges were made of data tampering and other wrongdoing. The material published was related to the work of the Climatic Research Unit and Professors Phil. Jones and Keith Briffa in particular. As Jones was a CLA of Chapter 3 of WG I in AR4 and both contributed to the chapter on paleoclimate, many of the emails involved IPCC matters and authors. There were several things in the emails that were obviously not for public consumption and possible abuses of the freedom of information act were revealed. Charges were brought against both Jones and Mann. Several official enquires in both the United Kingdom and the United States cleared them both of any wrongdoing.

The 2007 IPCC report itself has been scrutinized along with all of the comments and responses to the comments. Two minor errors have been found: both in WG II. These were on (1) Himalayan glaciers melt (correct in WG I) and (2) the area of the Netherlands below sea level. Unfortunately, these were played up and exaggerated in the media, and the IPCC was

slow to respond and did not handle the charges well by defending its processes.

Scientists have been found innocent of the charges made against them, and the basic science and all the main IPCC conclusions are unaffected. However, the criminals who stole the emails have not been brought to justice and those who misused the information and carried out the disinformation campaign were the real offenders. Unfortunately, their tactics have influenced the political process.

The adverse publicity concerning the IPCC raised concerns in some quarters regarding the continuing credibility of the IPCC assessments themselves and the processes and procedures underlying them and led to an investigation. In August 2010, the prestigious InterAcademy Council (representing scientific academies around the world) issued a report reviewing the procedures of the IPCC and made a number of recommendations for changes. They noted the need for the IPCC to update procedures, the need to have an executive committee to deal with issues such as minor corrections and to respond promptly to matters arising, to upgrade review oversight of reports, improve ways of communicating uncertainty, and improve communication in general.

Summary

The WG I findings might be summarized as follows. Climate changes have occurred in the past naturally, over decades to millennia for various reasons. However, humankind is performing a great geophysical experiment by modifying the Earth’s environment in various ways and changing the climate. Legitimate debates go on about the extent and rate of these changes, and what, if anything, to do about them but that the experiment is underway is not in doubt. The human-induced environmental changes of most relevance are in land use (e.g., farming, building cities), storage and use of water (dams, reservoirs, and irrigation), generation of heat, and combustion of fossil fuels. The latter, in particular, pollutes the atmosphere and alters the balance of radiation on Earth through both visible particulate pollution (called aerosols) and gases that change the composition of the atmosphere. The latter are referred to as greenhouse gases because they are relatively transparent to incoming solar radiation, while they absorb and reemit outgoing infrared radiation, thus creating a blanketing effect, which results in warming. Global warming and associated climate change are expected as a result and indeed are already occurring. The planet is unequivocally warming. Projections into the future indicate potentially dire consequences for ecosystems and human existence and well-being, sooner or later.

Given these findings, a key question is what should be done about them? The options include (1) do nothing, (2) mitigate or stop the problem, (3) adapt to the changes as they happen, or some combination of these. WG II deals with the first and the third option and WG III deals with option 2. Different value systems come into play in deciding what to do, and it is not the role of IPCC to decide, but it is the role of IPCC to lay out the options and likely consequences. Considerations include those of population growth, equity among developed and developing countries,

inter-generational equity, stewardship of the planet, vested interests, environmental concerns, and the precautionary principle (which suggests that it may be better to be safe than sorry). In rationally discussing options, it is helpful to recognize different points of view and that they are all legitimate. This problem is truly a global one, as the atmosphere is a global commons. It cannot be solved by one nation acting alone. IPCC helps to lay out the options and consequences of actions or lack of actions.

The IPCC website is <http://www.ipcc.ch>. Some of the material above comes from the IPCC website. A listing of IPCC procedures is at http://www.ipcc.ch/organization/organization_procedures.shtml#.T9Ipk5iwWP8.

Any opinions, findings, conclusions, or recommendations expressed in this publication are those of the author and do not necessarily reflect those of the National Science Foundation.

The National Center for Atmospheric Research (NCAR) is sponsored by the National Science Foundation.

Further Reading

- Archer, D., Rahmstorf, S., 2010. *The Climate Crisis: An Introductory Guide to Climate Change*. Cambridge University Press, pp. 249.
- Henson, R., 2011. *The Rough Guide to Climate Change*. Penguin, pp. 416.
- Houghton, J., 2009. *Global Warming: The Complete Briefing*, fourth ed. Cambridge University Press, pp. 456.
- IPCC, 2007a. *Climate Change 2007. The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*. In: Solomon, S., Qin, D., Manning, M., Chen, Z., Marquis, M., Averyt, K.B., Tignor, M., Miller, H.L. (Eds.). Cambridge University Press, pp. 996.
- IPCC, 2007b. *Climate Change 2007. Impacts, Adaptation and Vulnerability. Contribution of Working Group II to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*. In: Parry, M.L., Canziani, O.F., Palutikof, J.P., van der Linden, P.J., Hanson, C.E. (Eds.). Cambridge University Press.
- IPCC, 2007c. *Climate Change 2007. Mitigation of Climate Change. Contribution of Working Group III to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change, 2007*. In: Metz, B., Davidson, O.R., Bosch, P.R., Dave, R., Meyer, L.A. (Eds.). Cambridge University Press.
- IPCC, 2007d. *Climate Change 2007. Synthesis Report. Contribution of Working Groups I, II and III to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*. In: Core Writing Team, Pachauri, R.K., Reisinger, A. (Eds.). IPCC, Geneva, Switzerland, pp. 104.

Nuclear Winter

A Robock, Rutgers University, New Brunswick, NJ, USA

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

Nuclear winter is the term that describes the climatic effects of a nuclear war. Massive amounts of smoke from fires from burning cities and industrial areas would absorb sunlight, making it cold, dark, and dry at the Earth's surface and deplete ozone. A nuclear war between the USA and Russia today could produce nuclear winter, threatening the food supply for most of the planet. Even a nuclear war between new nuclear powers such as India and Pakistan could produce climate change unprecedented in recorded human history. The only way to eliminate the possibility of this climatic catastrophe is to eliminate the nuclear weapons.

Introduction

Nuclear weapons have existed for more than 65 years. The use of just one of these weapons would be horrible, as evidenced by Hiroshima or Nagasaki. For most of these 65 years, however, it is now known that mankind has possessed not just the means to destroy cities but the means to destroy the world, a 'Doomsday machine.' Although many people, several hundred million, would die from the immediate effects of nuclear weapons in a full-scale nuclear war, many more would die from the indirect effects, from starvation. Mass starvations in Africa, but without any outside help, now seem a more appropriate model for the world after nuclear war than Hiroshima or Nagasaki. More people could die in India or China from a nuclear war, even if no bombs are dropped there, than would die in the USA and Russia combined!

How could this possibly happen? Massive fires would be started by bombs dropped on cities and industrial targets, and the smoke from these fires would be so thick that it would block out the Sun for years. The resulting cold and dark at the Earth's surface was dubbed 'nuclear winter' by Richard Turco in 1983, because it would get as cold in the spring or summer as it gets in winter. Although these conclusions come from theoretical climate models, analogs on this planet and Mars give strong support to the theory.

Hiroshima

On 6 August 1945, a 15 kt nuclear bomb was dropped on Hiroshima, Japan, killing approximately 150 000 people. Many of these people died from the fires ignited by the bomb, which turned the city into a raging *inferno* – a firestorm – which pumped dense clouds of smoke high into the atmosphere. **Figure 1** shows the remains of the city. Where did all the buildings go? A significant fraction of them went up in smoke.

Many more people would have died if help had not been available immediately from outside the city in the form of medical care, food, water, and shelter. Three days later a 20 kt bomb was dropped on Nagasaki, also killing tens of thousands, and since then nuclear bombs have not been used in warfare.

Current Nuclear Arsenals

When nuclear winter theory first became known in the mid-1980s, at the height of the nuclear arms race, the world had

more than 50 000 nuclear weapons. Now, with the Cold War over, the USA and Russia are reducing their nuclear arsenals. Unfortunately, with more than 20 000 nuclear weapons still deployed in the world, there are still many more than would be necessary to produce nuclear winter. Even after the full implementation of the New Strategic Arms Reduction Treaty, signed on 8 April 2010 between the USA and Russia, with approximately 4000 nuclear weapons in the world by 2017 it will still be possible to produce nuclear winter. The total explosive power of the current nuclear arsenal is equal to about 5000 Mt (5 billion tons) of trinitrotoluene (TNT). 1 kt = 1000 t; 1 Mt = 1000 kt = 1 000 000 t. This large number becomes more meaningful when considered in perspective. There is the equivalent explosive power of almost a 1 t of TNT for each human on the face of the Earth. The Hiroshima bomb had an explosive power of 15 kt. This equals 0.015 Mt, which equals 0.0003% of the current global arsenal. If one Hiroshima-sized bomb had been dropped every *hour* from the end of World War II (WWII), it would have taken 38 years, until 1983, to use up the current arsenal!

The total explosive power of all bombs dropped in all of WWII, during which 50 000 000 people died, including Hiroshima and Nagasaki, was 3 Mt. The total explosive power of all bombs ever used in the history of the world in wars is 10 Mt. Yet, there is now available 500 times this explosive power in the world arsenals. This illustrates the enormity of the current potential to start fires.

While the American and Russian arsenals contain by far the majority of the world's nuclear weapons, there are seven more nuclear powers: France, Britain, China, India, Pakistan, Israel, and North Korea. A war between Pakistan and India could produce 5 Tg or more smoke, which would produce climate change unprecedented in recorded human history. It would not be nuclear winter, but it would be devastating.

How Nuclear Winter Could Be Produced

About one-third of the energy of a nuclear explosion is in the form of light or heat. It is like bringing the Sun to the Earth's surface for a brief moment. Anything close to the explosion will burst into flames. The assumption made in many nuclear winter scenarios is that anything receiving more than 10 calories per square centimeter per minute (about 7000 W m^{-2} – 20 times the average amount of energy received at the top of the Earth's



Figure 1 Hiroshima after a 15 kt bomb was dropped on 6 August 1945. The streets were cleaned before this picture was taken. Where have all the buildings gone? They burned in the resulting fire, pumping thick clouds of black smoke into the atmosphere. From US Air Force Photo Library, Bolling Air Force Base, Washington, DC.

atmosphere from the Sun) will burst into flames, and this was demonstrated in actual tests in Nevada before the atmospheric nuclear test ban.

Following the flash of light comes the blast wave (like thunder following lightning), which will break apart many structures and blow out the flames, but crumpled structures burn more easily and fires would be reignited by burning embers and electrical sparks. Imagine how easily a house would burn with open gas lines, or a filling station with gas pumps knocked over. In fact, there are many flammable sources of fuel for fires in cities, including buildings and their contents, trees, and even asphalt. Modern materials, such as plastics, not only burn with a sooty smoke, but also produce high levels of toxic chemicals.

The climatic, and hence lethal, effects of the use of nuclear weapons depend on the amount of smoke they would generate, and this depends on the targets. Nuclear targeting plans call for not only cities to be targeted but also industrial facilities such as oil refineries and wells. Forests around military targets would also provide fuel. All these targets together would produce clouds of black sooty smoke, which rise into the atmosphere. Less than 1% of current global nuclear arsenals would be enough to produce nuclear winter, if targeted at oil refineries.

The black smoke cloud would rise into the atmosphere and rapidly envelop the entire world. Absorption of sunlight would heat the cloud, cause it to rise, and induce winds, which would blow some of it into the Southern Hemisphere. Although about two-thirds of the smoke would be fairly rapidly cleansed from the atmosphere by rain and snow, about a third would remain for many years, with substantial amounts still in the stratosphere for at least a decade. Initial studies suggested that

the Earth's surface would become dark and cold, as cold in the summer as normal winter temperatures, hence the term 'nuclear winter.' Recent calculations with modern climate models have validated these early results. Intermediate results with less cooling led some critics to suggest that 'nuclear fall' would result and would in some way be more desirable than nuclear 'winter', but they were incorrect.

Climate Model Calculations

In 1982, the Swedish journal *AMBIO* commissioned a special issue on the environmental effects of nuclear war. They composed a scenario for a large-scale nuclear war in which 5000 Mt of weapons were dropped on targets in North America, Europe, and Asia, and invited different scientists to write papers on the effects on various parts of the environment. Paul Crutzen (winner of the Nobel Prize in Chemistry in 1995) and John Birks looked into the effects on atmospheric pollution. One of the most important constituents of air pollution near the Earth's surface is ozone, a major component of smog, which is formed by a photochemical reaction. Sunlight shining on a mixture of chemicals that would be released by explosions of industrial targets would produce ozone. To calculate the reaction rates they needed to know the amount of sunlight, so they did a quick calculation of the amount of smoke that would be produced from the fires that would burn after the bombs exploded. They were amazed to find that there would be so much smoke that virtually no sunlight would reach the surface! They also concluded that ozone would not be a major atmospheric pollutant near the Earth's surface.

This astounding result was quickly studied by climate researchers. During a 1-year period, scientists from many nations, including the USA and the Soviet Union, used many types of climate models to calculate how the Earth's climate system would respond to so much smoke in the atmosphere, and how the temperature, precipitation, and winds would change. The cooperative research between American and Russian scientists was a natural outgrowth of the close working relationship that had developed during the Soviet–American Scientific Exchange Program in climate research.

All the calculations came to the same conclusions. There would be so much smoke that virtually all the sunlight would be absorbed, and the surface of the Earth would become cold and dark for months, if not years. As climate models became more refined, and more and more details were added to the calculations, the precise amount of cooling moved up and down slightly, but no one could find a mechanism that would overcome the basic fact that light cannot go through thick clouds of smoke. Some critics argued that when more of the 'unknowns' were taken into consideration, this drastic result would be disproved, but unknowns, by definition, are *unknown*. While some previously unconsidered factors indeed lessened the cooling, others increased it, and the final consensus is not much different from the original calculations. In fact modern climate models, which incorporate a complete treatment of the ocean and simulate the atmosphere up through the mesosphere up to 80 km (50 miles), found that by absorbing sunlight the smoke would be lofted into the upper stratosphere and last for many years. The climatic effects of nuclear winter would last much longer than previously thought.

Not only would the surface air temperature be affected severely, but other drastic changes would also take place in the atmosphere. Climate model calculations show that precipitation would be reduced substantially over the continents in the years following the soot injection. The summer monsoon precipitation over Asia, the main source of rainfall for the entire year, would be virtually completely missing for years.

The fires would generate a substantial amount of air pollution, especially toxic chemicals (such as dioxins, furans, and polychlorinated biphenyls (PCBs)) from burning plastics and petrochemical facilities. Asbestos fibers would be injected into the air. In addition, radioactive fallout would be widespread, with some 'hotspots' having lethal levels.

The nuclear fireballs from large bombs penetrate up into the stratosphere. The intense heat produces NO and NO₂ (NO_x), which act as a catalyst to destroy ozone. Calculations show reductions of ozone in the stratosphere by as much as 50%, with the effects only gradually ameliorating over a period of years. A trend toward smaller nuclear weapons would lessen the number of fireballs reaching the stratosphere, but the changed atmospheric circulation would still take the NO_x up into the ozone layer and into the Southern Hemisphere, and the stratosphere would be so hot that ozone-destroying reactions would proceed much more rapidly, producing a global 'nuclear winter ozone hole' that would last for years. Even with the smoke, this would admit much more than the current amount of ultraviolet B (UV-B) radiation to the surface, affecting humans and other organisms.

The above results for the climate are for a nuclear war that would take place in the spring or summer. For one that would take place in the fall or winter, the immediate surface temperature effects would be less, since there is less sunlight to block out, but severe effects would still last for years.

Within the past 5 years, new results have shown that even a nuclear war between nuclear powers other than the USA and Russia superpowers could still produce so much smoke that summer temperatures around the world in food producing regions would plummet by several degrees, precipitation would go down, it would be darker, growing seasons would be shorter by weeks, and there would be much more UV-B radiation from ozone depletion. Reduced food production combined with panic in world food trade could spell starvation for a billion or more people. [Figure 2](#) shows the global average surface air temperature changes for 5 Tg smoke (war between India and Pakistan), and for two different smoke injections (50 and 150 Tg) that could result from a USA–Russia war.

Analogs

How can the extreme situations that have been calculated in theoretical models be investigated? The atmosphere cannot be brought into the laboratory and experiments cannot be performed on it. And the experiment cannot actually be performed in nature. Or it could be performed only once, and then it would be too late. However, other occurrences can be examined in the climate system to see if any similar situations have existed that would help learn about what would happen in the event of a nuclear winter. Situations that teach about parts of the interactions discussed above or for the global climate response can be studied, and the study need not have to be confined to this planet.

During WWII, firestorms were produced by conventional bombing in Dresden, Hamburg, Darmstadt, and Tokyo, and by nuclear bombing in Hiroshima and Nagasaki. Therefore, it is known that cities can burn, and in fact produce firestorms – superfires that spread far beyond the initial area of ignition and pump smoke high in the atmosphere. Kurt Vonnegut was a prisoner of war in Dresden during the fire there and in his book *Slaughterhouse 5* he described his experience:

He was down in the meat locker on the night that Dresden was destroyed. There were sounds like giant footsteps above. Those were sticks of high-explosive bombs. The giants walked and walked...

So it goes.

A guard would go to the head of the stairs every so often to see what it was like outside, then he would come down and whisper to the other guards. There was a firestorm out there. Dresden was one big flame. The one flame ate everything that was organic, everything that would burn.

It wasn't safe to come out of the shelter until noon the next day. When the Americans and their guards did come out, the sky was black with smoke. The sun was an angry little pinhead. Dresden was like the moon now, nothing but minerals. The stones were hot. Everybody else in the neighborhood was dead.

So it goes.

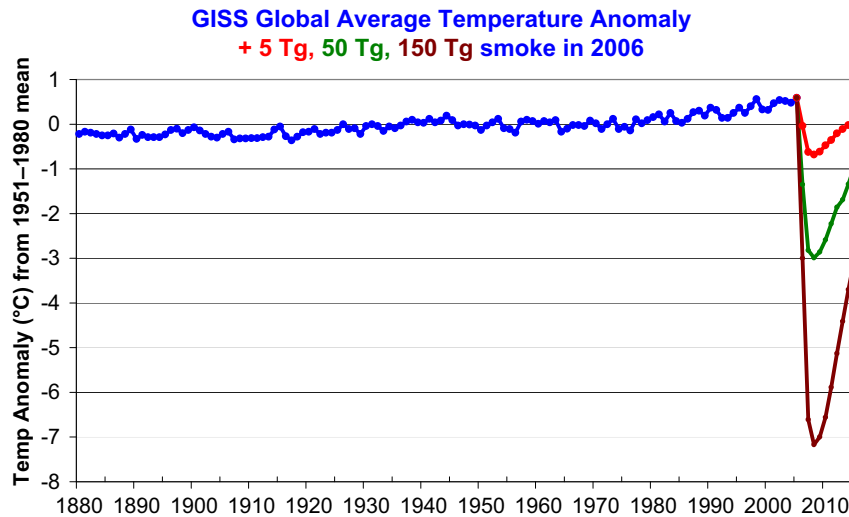


Figure 2 Global average surface air temperature change from the 5 Tg (red), 50 Tg (green), and 150 Tg (brown) cases in the context of the climate change of the past 125 years. Observations are from the National Aeronautics and Space Administration Goddard Institute for Space Studies analysis. Reproduced from Hansen et al. (2001); updated at <http://data.giss.nasa.gov/gistemp/2005/>. Figure 8 from Robock, A., Oman, L., Stenchikov, G.L., 2007b. Nuclear winter revisited with a modern climate model and current nuclear arsenals: Still catastrophic consequences. *Journal of Geophysical Research* 112: doi:10.1029/2006JD008235. With permission from ©American Geophysical Union.

So even one day later, the sky was still so full of the smoke that it almost completely blocked the Sun, and this was from 'conventional' bombing of only one city.

The diurnal cycle (day and night) provides a good analog. Imagine if the Sun did not rise tomorrow, if tonight were followed by another night followed by another night. This would be the situation under thick clouds of smoke in plumes downwind from major fires. It is easy to see how, even in the summer, temperatures could rapidly plummet to below freezing. The seasonal cycle analog, as already mentioned, has given the term, 'nuclear winter.'

Sixty-five million years ago an asteroid or comet smashed into the Earth producing a mass extinction of species. The best-known group of animals that disappeared at that time was the dinosaurs, but in addition, every land animal larger than a cat died and most of the plant species disappeared. In addition to a global dust cloud that blocked out the sunlight producing cold and dark conditions, there were continent-scale forest fires that also produced a thick smoke layer in the atmosphere, exactly what has been proposed for the nuclear winter.

Periodically, clouds of dust are blown up from the Sahara desert and transported all the way across the Atlantic Ocean. From this it is learned that dust particles in the troposphere (the lowest atmospheric layer, from the surface up to 8–18 km) can be spread large distances around the world. It has been observed that under Saharan dust clouds it is colder and there are fewer water clouds and less rain. This suggests a similar reaction to nuclear soot and dust.

There have been large forest fires in recent history, too, although not as large as those implicated in the extinction of the dinosaurs. These can be studied to learn about the properties of the smoke particles and how they affect light and heat radiation going through them. In September 1950 a giant forest fire raged in Western Canada for a week. A week later the smoke cloud covered the Eastern USA and a week after that it

was seen over Europe. Again how far particles can be transported by the wind before getting washed out of the atmosphere is seen. When the smoke was over Washington, DC, weather forecasts for high temperature were as much as 6 °C too high. The actual temperatures were 6 °C lower than were forecast because of the sunlight blocked by the smoke. Even in late September, when the Sun is not very intense, large surface temperature effects can result from a smoke cloud in the atmosphere, one that is much less thick than that calculated for nuclear winter.

This anecdotal report was confirmed in more recent studies that found surface cooling of 2–4 °C in the Midwestern USA under smoke clouds generated by forest fires in British Columbia, Canada, in 1982, and more than that produced by the September, 1988 forest fires in Yellowstone Park. The May 1987 forest fires in Northern China, and more recent ones in Australia and Canada, produced thick smoke clouds that rose all the way into the stratosphere where they resided for a long time, just as predicted by the computer models discussed in the previous section. In a nuclear winter situation, the smoke from burning cities and industrial facilities is expected to be thicker and blacker than forest fire smoke, producing even more cooling.

When the US Mariner 9 spacecraft first flew by Mars to take high-resolution pictures of the surface, the Northern Hemisphere of Mars was covered by a thick cloud of dust. (Mars has an atmosphere, too, but much thinner than the Earth's.) A few weeks later, the entire Martian globe was covered by this dust cloud. The heating of the atmosphere caused by the dust cloud in one hemisphere induced a circulation that transported the dust later into the other hemisphere. This same effect is part of the nuclear winter scenario and implies that regions far removed from the conflict would experience climate changes.

Volcanic eruptions provide several examples that can teach about nuclear winter. Clouds of volcanic dust and sulfuric acid

droplets that get injected into the stratosphere (the layer above the troposphere, where there is no weather and no rain to wash out the particles) have been observed to be spread completely around the globe in 3 weeks and remain for several years. This is the same fate postulated for nuclear smoke that either gets initially injected into the stratosphere or is lofted there by solar heating. Large volcanic eruptions can produce dust clouds in the troposphere immediately after the eruptions. Large surface temperature changes have been observed under these tropospheric volcanic dust clouds following the 1883 Krakatau, 1980 Mt St Helens, and 1991 Pinatubo eruptions in a manner similar to the effects of forest fire smoke.

The long-lasting stratospheric dust clouds also have been observed to produce global climate changes for several years following large volcanic eruptions. One of the largest in recent memory was the eruption of Tambora in 1815, which was followed by such cold weather during the following summer that 1816 has become known as the 'Year Without a Summer.' That summer the famous poet Lord Byron lived by the shore of Lake Geneva, Switzerland, next door to his friends, Percy Bysshe Shelley and his young 18-year-old bride Mary. The weather was so cold and gloomy that they had a contest to see who could write the best ghost story, and Mary Shelley won by writing *Frankenstein*, which begins and ends with frigid images of the monster climbing over ice floes. Byron himself was so depressed by the cold, gray weather that he wrote a poem called *Darkness* which begins:

I had a dream, which was not all a dream.
The bright sun was extinguish'd, and the stars
Did wander darkling in the eternal space,
Rayless, and pathless, and the icy earth
Swung blind and blackening in the moonless air;
Morn came and went—and came, and brought no day,
And men forgot their passions in the dread
Of this their desolation; and all hearts
Were chill'd into a selfish prayer for light:
And they did live by watchfires – and the thrones,
The palaces of crowned kings – the huts,
The habitations of all things which dwell,
Were burnt for beacons; cities were consumed,
And men were gather'd round their blazing homes
To look once more into each other's face; ...

This remarkable description of what the world might be like in a nuclear winter was inspired by a volcanic dust cloud much thinner than the cloud that would produce a nuclear winter!

Biological Consequences

The most important consequence of nuclear winter for humans is the disruption of food supplies. This could come about by two processes. One is the environmental disruptions that reduce or completely wipe out agricultural production. The other is the disruption of the distribution mechanisms. Not only would it be virtually impossible to grow food for several years after the nuclear holocaust, it would also be impossible to

obtain food from other countries. It is hard to imagine that countries with some food production (possibly Argentina, Australia, or New Zealand) would be willing to export it in such an uncertain international situation. For a regional nuclear war with less cooling, but still important disruptions of food production, these market disruptions could produce more shortages than the direct impacts on agriculture. In addition to the disruption of food, there would be many other stresses for any surviving people. These would include the lack of medical supplies and personnel, high levels of pollution and radioactivity, psychological stress, rampant diseases and epidemics, and enhanced UV-B radiation after the smoke clears.

There are many ways that agriculture is vulnerable to nuclear winter. The cold and the dark alone are sufficient to kill many crops. Superimposed on the average cooling would be large variations. During the summer of 1816 in New England, there were killing frosts in each summer month. Only 1 day with the temperatures below freezing is enough to kill rice crops. Colder temperatures mean shorter growing seasons and also slower maturation of crops, the combination results in much lower yields. Most of the grains that are grown in midlatitudes, such as corn, are actually of tropical origin, and will only grow in summerlike conditions. A study done in Canada shows that with summer temperatures only 3 °C below normal, wheat production would halt. Insufficient precipitation would also make agriculture difficult.

The tremendous productivity of the grain belt of the USA and Canada feeds not only those countries but also many in the rest of the world, including Russia, where normal climate variability often results in reduced harvests. This productivity is the result of modern farming techniques that allow less than 2% of the population to produce more than enough for the rest. In order to do this, tremendous energy subsidies are needed. Farmers depend on fuel for their machinery, fertilizer, and pesticides, none of which would be available or distributed in the aftermath of a war. Furthermore, insects have a higher tolerance for radiation and the stresses that would follow than do their predators, such as birds. Whatever might grow would be eaten by pests, already a significant problem in today's production. Also, the seeds that are in use were designed to yield high productivity assuming the current climate and inputs of chemicals and energy as discussed above. These seeds would not grow well in a radically altered growing environment. Mankind's dependence on technology is such that if every human in the USA went out to the fields to try to raise crops with manual labor, and if they knew what they were doing, and if they had enough food to eat, and if they were healthy, they still could not produce what is produced today.

Thus, most of the world's people are threatened with starvation following a nuclear war. The number that would survive depends on how much food is in storage, and how much could be produced locally. Studies of various countries around the world conclude that even with extremely optimistic assumptions of perfect distribution systems within countries, that each person who will survive becomes a vegetarian and eats the minimum needed for survival, and the others waste none of the food, that nations in Asia, Africa, and South America could only last 1–2 months. In many nations, people would be

reduced to a hunter/gatherer existence with nothing to hunt and precious little to gather.

The effects on health would add to the misery. Immune deficiencies can be produced by any of the following: burns and trauma, radioactivity, malnutrition, psychological stress, and UV-B radiation. All of these would be present for the survivors in the target nations. It would in some respects resemble an international epidemic of acquired immune deficiency syndrome (AIDS).

Pollution from dioxins, PCBs, asbestos, and other chemicals will make the air unhealthy to breathe. Severe psychological stress will prevent the survivors from making the efforts to continue to exist.

Would it be possible to survive in the USA or Europe? If you had a fallout shelter that was deep enough or far enough from the immediate effects of blast and fires, and if you stayed in it long enough so that you could safely come out (about a year considering the radioactivity and UV-B radiation) and if you had enough supplies to last you for that long and for several more years until you could produce more food, and if you did not go crazy from being cooped up in a hole in the ground, and if you had enough weapons to keep out all the other hungry people, then you could survive, if you call that surviving.

One might think that the ocean shore would be a good place to be because the temperatures would not fall as much, and there would be plenty of food to catch. Although the ocean would not cool very fast, the darkness would decimate the phytoplankton, which are at the base of the oceanic food chain. That, combined with toxic and radioactive pollution would severely limit the food sources in the oceans. Furthermore, the large temperature contrasts between the oceans and the land would produce strong storms that would make fishing difficult at best.

While it is important to point out the consequences of nuclear winter, it is also important to point out what will not be the consequences. Although extinction of mankind was not ruled out in initial studies by biologists, it now seems that this would not take place. Especially in Australia and New Zealand, humans would be sure to survive. Also, the Earth will not be plunged into an Ice Age. Ice sheets, which covered North America and Europe only 18 000 years ago and were more than 3 km thick, take many thousands of years to build up from annual snow layers, and the climatic disruptions would not last long enough to produce them. The oxygen consumption by the fires would be inconsequential, as would the effect on the atmospheric greenhouse by carbon dioxide production. The consequences of nuclear winter would be extreme enough without these additional effects, however.

Policy Implications

The suicidal nature of the use of nuclear weapons is one of the most important policy implications. If country A used enough weapons only against military targets to prevent country B from retaliating, in what is called a 'first strike,' the climatic consequences could be such that everyone in country A could die. Nuclear weapons, therefore, become an instrument of suicide and not an instrument of defense. A 'limited war' would not in itself be enough to trigger nuclear winter. Furthermore, once nuclear war is started, communications

would be so disrupted, by the smoke in the air and by the electromagnetic pulse (EMP) that would destroy all unprotected electronic gear, and people would be under so much stress, that it is considered unlikely that nuclear war could ever be limited.

Soon after the nuclear winter theory was discovered, Carl Sagan gave a briefing on the subject to Senators, Congressmen, and staff on Capitol Hill. He described how the smoke from burning cities and industrial areas after a nuclear war would be so thick as to block out so much sunlight that the Earth's surface would become so cold, dark, and dry for so long that agriculture would be impossible and most of the people in the world would starve to death. After the presentation, one of them called him aside and said, "Carl, if you think the mere threat of the end of the world is enough to change the way people in Washington and Moscow think, you clearly haven't spent enough time in either place!" Albert Einstein said, after nuclear weapons were invented, that their existence changed everything about the world, except for the way that we think, and thus we drift toward 'unparalleled catastrophe'.

Yet it does seem that nuclear winter has provided a context to reexamine all the existing policy assumptions about nuclear war. People are gradually changing the way they think. And it happened only because scientists have tried to warn the world of the dangers of current policies. The author feels that it is the responsibility of scientists, particularly those who use public money for their research, to warn society when they discover potential dangers, such as from nuclear winter or ozone depletion.

The world seems to be a much safer place now than it was in 1982 and 1983 when the first papers on nuclear winter were published. How much of this change was caused by the realization of the dangers of nuclear winter? Nobel Peace Prize winner Mikhail S Gorbachev, the prime architect of the current good East-West relations, said in an interview in 2000, "Models made by Russian and American scientists showed that a nuclear war would result in a nuclear winter that would be extremely destructive to all life on Earth; the knowledge of that was a great stimulus to us, to people of honor and morality, to act."

The Cold War is over, but many of the nuclear weapons produced during this period remain. The USA and Russia are very slowly reducing the numbers of weapons, but each still maintain an arsenal far larger than necessary to produce nuclear winter. No current leader of the USA or Russia would use nuclear weapons, but their existence alone makes the possibility of nuclear winter in the future possible if a crazy person or computer error or misunderstanding causes their use. The only solution is to reduce the number of weapons to a level that will still provide a deterrent, but will not create a nuclear winter should they ever be used. Reducing these numbers to a level below which they could produce a global climatic catastrophe, as Sagan was fond of saying, is a matter of elementary planetary hygiene. This number is less than a hundred, less than the number of weapons that Britain, France, and China have had in each of their arsenals for decades, a number that Israel, India, and Pakistan have chosen for their arsenals, and a number they have deemed more than sufficient to maintain a credible defense of their countries. This

is also the number Admiral Stansfield Turner, former Director of the CIA, argued for on other grounds in 1997. This problem must be solved so that there is the luxury of spending the time solving the problem of greenhouse warming and working toward nuclear abolition.

This article is dedicated to Carl Sagan, one of the leaders of nuclear winter research who died tragically in December 1996 at the young age of 62. Some of the ideas in his paper come from listening to his public talks on the issue. I thank him for so clearly expressing his concerns about this important issue.

See also: **Aerosols:** Role in Climate Change; Role in Radiative Transfer. **Agricultural Meteorology and Climatology:** Agricultural Meteorology and Climatology. **Boundary Layer (Atmospheric) and Air Pollution:** Diurnal Cycle. **Climate and Climate Change:** Volcanoes: Role in Climate. **General Circulation of the Atmosphere:** Overview. **Global Change:** Biospheric Impacts and Feedbacks.

Further Reading

- Crutzen, P.J., Birks, J.W., 1982. The atmosphere after a nuclear war: Twilight at noon. *Ambio* 11, 115–125.
- Ehrlich, P.R., Sagan, C., Kennedy, D., Roberts, W.O., 1984. *The Cold and the Dark – The World after Nuclear War*. W.W. Norton & Co., Inc., New York, NY.
- Hansen, J.E., Ruedy, R., Sato, M., Imhoff, M., Lawrence, W., Easterling, D., Peterson, T., Karl, T., 2001. A closer look at United States and global surface temperature change. *Journal of Geophysical Research* 106, 23,947–23,963.
- Harwell, M.A., Hutchinson, T.C. (Eds.), 1986. *Environmental Consequences of Nuclear War*, SCOPE 28. Ecological and Agricultural Effects, vol. II. John Wiley & Sons, New York, NY.
- Özdoğan, M., Robock, A., Kucharik, C.J., 2013. Impacts of a nuclear war in South Asia on soybean and maize production in the Midwest United States. *Climatic Change* 116, 373–387.
- Pittcock, A.B., Ackerman, T.P., Crutzen, P.J., MacCracken, M.C., Shapiro, C.S., Turco, R.P. (Eds.), 1986. *Environmental Consequences of Nuclear War*, SCOPE 28. Physical and Atmospheric Effects, vol. I. John Wiley & Sons, New York, NY.
- Robock, A., 1984. Snow and ice feedbacks prolong effects of nuclear winter. *Nature* 310, 667–670.
- Robock, A., 1989. Policy implications of nuclear winter and ideas for solutions. *Ambio* 18, 360–366.
- Robock, A., Toon, O.B., 2010. Local nuclear war, global suffering. *Scientific American* 302, 74–81.
- Robock, A., Oman, L., Stenchikov, G.L., Toon, O.B., Bardeen, C., Turco, R.P., 2007a. Climatic consequences of regional nuclear conflicts. *Atmospheric Chemistry and Physics* 7, 2003–2012.
- Robock, A., Oman, L., Stenchikov, G.L., 2007b. Nuclear winter revisited with a modern climate model and current nuclear arsenals: Still catastrophic consequences. *Journal of Geophysical Research* 112, D13107. doi:10.1029/2006JD008235.
- Rudolf, A., 1984. *Byron's Darkness: Lost Summer and Nuclear Winter*. Menard Press, London, UK.
- Sagan, C., Turco, R., 1990. *A Path where No Man Thought – Nuclear Winter and the End of the Arms Race*. Random House, New York, NY.
- SCOPE ENUWAR Committee, 1987. Environmental consequences of nuclear war: An update; severe global-scale effects of nuclear war reaffirmed. *Environment* 29 (4), 4–5. 46.
- Toon, O.B., Turco, R.P., Robock, A., Bardeen, C., Oman, L., Stenchikov, G.L., 2007. Atmospheric effects and societal consequences of regional scale nuclear conflicts and acts of individual nuclear terrorism. *Atmospheric Chemistry and Physics* 7, 1973–2002.
- Toon, O.B., Robock, A., Turco, R.P., 2008. Environmental consequences of nuclear war. *Physics Today* 61 (12), 37–42.
- Turco, R.P., Toon, O.B., Ackerman, T.P., Pollack, J.B., Sagan, C., August, 1984. The climatic effects of nuclear war. *Scientific American* 251, 33–43.
- Turner, S., 1997. *Caging the Nuclear Genie*. Westview Press, Boulder, CO.
- Vonnegut, K., 1969. *Slaughterhouse-5*. Dell, New York, NY.
- Xia, L., Robock, A., 2013. Impacts of a nuclear war in South Asia on rice production in mainland China. *Climatic Change* 116, 357–372.

Radiative–Convective Equilibrium Climate Models

NO Renno and X Huang, University of Michigan, Ann Arbor, MI, USA

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

Radiative–convective models and key results of simulations with them such as multiple climate equilibria and runaway greenhouse are described. These simple climate models are widely used not only to simulate the thermodynamic structure of the atmosphere of the Earth and other planets but also to study their sensitivity to changes in the concentration of greenhouse gases, convective processes, clouds, and the flux of incoming solar radiation.

Introduction

The terrestrial planets emit as much energy to space in the form of thermal radiation as they receive from the sun in the form of solar radiation. On Earth, the incident solar radiation is absorbed mostly at the surface, but it is also scattered and absorbed by atmospheric gases, aerosols, and clouds. The absorption of solar radiation, its redistribution by dynamic and radiative processes, and the emission of thermal radiation back to space determine the mean surface temperature and the mean vertical thermodynamic structure of the atmosphere. Radiative–convective models simulate these processes and give insights into their effects on the energy budget, the vertical structure, and the stability of planetary atmospheres.

Radiative–convective models are ideal for studying general principles and testing fundamental ideas. Their major drawback is the inability to calculate the feedbacks between the horizontal heat transports and the temperature structure from first principles. Radiative–convective models are widely used not only to simulate the thermodynamic structure of the atmosphere of the Earth and other planets but also to study their sensitivity to changes in the concentration of greenhouse gases, convective processes, clouds, and the flux of incoming solar radiation.

In the classical radiative–convective models developed in the 1960s, the atmosphere's water vapor mixing ratio is either fixed or diagnosed based on the climatological profile of relative humidity. In addition, these models use simple numerical procedures to parameterize the cumulus convection, the processes responsible for the distribution of water vapor into the atmosphere. Since water vapor is the most important greenhouse gas, it is desirable to explicitly calculate its content and vertical distribution. The radiative–convective models developed during the last 2 decades do this by explicitly calculating the hydrological cycle. Furthermore, they employ complex parameterization schemes similar to those that global climate models use to represent the cumulus convection and therefore to distribute the heat and water vapor and to produce the precipitation.

Radiative–Convective Models

Radiative–convective processes are described by the equations:

$$\frac{\partial \theta}{\partial t} = Q + R + F_\theta \quad [1]$$

$$\frac{\partial r}{\partial t} = C + F_r \quad [2]$$

integrated at each model pressure level. Time, t , and pressure, p , are independent variables. The prognostic variables are the potential temperature, θ , and the water vapor mixing ratio, r . Q in eqn [1] represents the heating by cumulus convection and large-scale condensation, while C in eqn [2] stands for the moisture source due to these two processes. These terms are calculated by parameterization schemes representing the net effects of cumulus convection and large-scale condensation. R in eqn [1] represents the net radiative heating. Radiation is calculated using models of wide range of complexity, depending on the specific goals of the simulation.

Vertical diffusion of heat and moisture are represented by

$$F_\theta = g \frac{\partial \tau_\theta}{\partial p} \quad [3]$$

$$F_r = g \frac{\partial \tau_r}{\partial p} \quad [4]$$

where τ_θ and τ_r represent the vertical fluxes of heat and water vapor. In the free atmosphere, these vertical fluxes are represented by

$$\tau_\theta = \rho^2 g k_\theta \frac{\partial \theta}{\partial p} \quad [5]$$

$$\tau_r = \rho^2 g k_r \frac{\partial r}{\partial p} \quad [6]$$

where ρ is the air density, g is the gravity acceleration, and k_θ and k_r are the coefficients of vertical diffusion of heat and water vapor. Surface fluxes are usually calculated using the bulk aerodynamic formulae:

$$\tau_\theta = \rho_a C_D |\mathbf{v}_a| (\theta_s - \theta_a) \quad [7]$$

$$\tau_r = \rho_a C_D |\mathbf{v}_a| (r_s - r_a) \quad [8]$$

where the subscripts 's' and 'a' refer to values at the surface and just above the surface, at a standard height referred to as the anemometer level. $C_D \approx 0.0025$ is the drag coefficient and $\mathbf{v}_a \approx 5 \text{ m s}^{-1}$ is the wind vector at the anemometer level. The radiative–convective equilibrium is not sensitive to the values of these arbitrary parameters because changing them simply changes the difference in potential temperature and water vapor mixing ratio between the surface and the anemometer level. The values of the potential temperature and water vapor

mixing ratio at the anemometer level are calculated by making the potential temperature and the water vapor mixing ratio constant from the model's lowest level to the anemometer level. This represents the effects of adiabatic mixing by either dry convection or turbulence.

A saturated surface with zero heat capacity and an infinite supply of moisture is usually used as the lower boundary condition because it speeds up the approach to equilibrium and does not affect the solution when the supply of water is not a limiting factor. In this case, the net energy flux into the surface is required to be zero at each instant.

Atmospheric Convection

A representation of the basic physical processes governing the transport of water vapor by cumulus convection is crucial for climate simulations because water vapor is the most important greenhouse gas. Manabe and his associates developed a simple parameterization for cumulus convection, referred to as convective adjustment, while pioneering the development of climate models in the 1960s. The Manabe scheme is based on the assumption that when the temperature lapse rate of a saturated area exceeds the moist adiabatic lapse rate, moist convection is always strong enough to mix the air moist adiabatically. Then, these mixed or adjusted layers are kept saturated and the excess water vapor falls out as rain. If a model layer is supersaturated, but the temperature lapse rate does not exceed the moist adiabatic, large-scale condensation occurs and the excess water vapor falls out as rain. After large-scale condensation, the layer is kept saturated, while its temperature is increased by the release of latent heat of condensation. This classical scheme is still widely used in radiative–convective models.

The comprehensive cumulus parameterization scheme developed by Emanuel three decades later lies in the other extreme of complexity by representing the detailed physics of cumulus convection. It assumes that the fundamental entities of cumulus convection are the subcloud-scale drafts rather than the cloud-scale circulations themselves. This scheme assumes that convection occurs whenever the environment is unstable to reversible adiabatic ascent of air parcels originating at any model level. It assumes that vertical transports of heat and water vapor are accomplished by saturated updrafts and downdrafts, by a single unsaturated downdraft driven by the evaporation of falling precipitation, and by the compensating subsidence in the cloud environment. The precipitation efficiencies of each saturated updraft, the fraction of the precipitation that falls through unsaturated air, and the rate of evaporation of falling precipitation are either specified or calculated based on physically based assumptions.

Multiple Climate Equilibria

Simulations with radiative–convective equilibrium climate models produce two linearly stable solutions. The first solution corresponds to an optically thin atmosphere. In this solution, solar radiation is mainly absorbed at the surface, and the bulk of the emission of radiation to space originates in the warm surface and low troposphere (the entropy flux is minimum).

This solution is stable to finite amplitude perturbations when the radiation emitted by an opaque version of the planet is larger than the absorbed solar radiation. The second equilibrium corresponds to an optically thick atmosphere. In this second solution, the atmosphere's opacity to infrared radiation is the largest. The bulk of the emission of radiation to space originates in the cold upper troposphere (the entropy flux is maximum) and the hydrological cycle is extremely active. **Figure 1** indicates that these two linearly stable solutions occur when the value of the net flux of solar radiation at the surface (the net forcing) is within the range for Earthlike planets, that is, the net forcing is between $F_1 \approx 275 \text{ W m}^{-2}$ and $F_3 \approx 295 \text{ W m}^{-2}$. These two solutions are separated by a linearly unstable solution marked by dashed lines in **Figure 1**.

Figure 1 shows that there is an asymptotic limit for the emission of thermal radiation by a convective moist planet in which the main atmospheric constituent is water vapor (or any infrared absorber whose atmospheric concentration is temperature dependent). This indicates that a runaway greenhouse might occur when the net solar forcing is above $F_2 \approx 290 \text{ W m}^{-2}$. When the solar forcing is above a larger, but model-dependent value $F_4 \approx 310 \text{ W m}^{-2}$, equilibrium is not possible and a runaway greenhouse always occurs. The runaway greenhouse is discussed below.

Runaway Greenhouse

The net flux of solar radiation into a planet with equilibrium climate must be balanced by the emission of thermal radiation to space. For a planet without atmosphere or for a hypothetical planet with an atmosphere without infrared absorber, this balance is always possible. In this case, increases in the surface temperature produce monotonic increases in the emission of thermal radiation by the planet's surface that leads to a unique equilibrium solution.

When the main atmospheric constituent is an infrared absorber whose atmospheric content is temperature dependent such as water vapor, increases in the surface temperature lead to

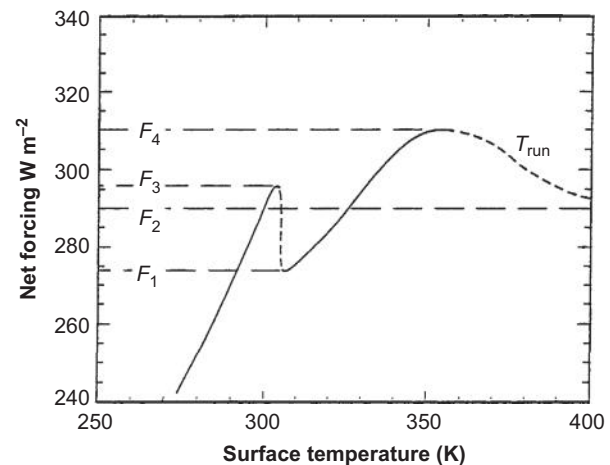


Figure 1 Sketch of the bifurcation for a moist atmosphere in radiative–convective equilibrium. The branches represented by a dashed line are linearly unstable.

an asymptotic value for the emission of thermal radiation to space (represented by F_2 in Figure 1). In this case, equilibrium might not be possible when the net heat input into the planet is larger than the value F_2 . In this case, increases in the surface temperature produce increases in the emission of thermal radiation by the planet up to a maximum value of the surface temperature (about 350 K in Figure 1) at which the emission of thermal radiation to space reaches a peak value. For temperatures above this value, increases in the surface temperature produce decreases in the emission of thermal radiation. This happens because increases in the atmosphere's optical thickness overcompensate for increases in the emission of thermal radiation by the warmer surface. Further increases in the surface temperature lead to increases in the concentration of the infrared absorber and produce the asymptotic value for the emission of thermal radiation.

The runaway greenhouse is caused by the fact that cumulus convection constrains the atmospheric temperature lapse rate to be approximately moist adiabatic and distributes water vapor vertically. When the atmosphere becomes opaque, thermal radiation is emitted to space from layers in the moist adiabatic atmosphere above the surface and therefore a region of constant temperature and opacity is produced. This decouples the emission of thermal radiation from the temperature of the surface and produces an asymptotic value for the emission of thermal radiation. When the net value of the incoming solar radiation is above this asymptotic value, a runaway greenhouse can occur.

See also: **Boundary Layer (Atmospheric) and Air Pollution:** Convective Boundary Layer. **Climate and Climate Change:** Greenhouse Effect. **Gravity Waves:** Convectively Generated Gravity Waves.

Further Reading

- Chamberlain, T.P., Hunten, D.M., 1990. *Theory of Planetary Atmospheres: An Introduction to Their Physics and Chemistry*, vol. 36. Academic Press.
- Emanuel, K.A., 1994. *Atmospheric Convection*. Oxford University Press.
- Goody, R.M., 1949. The thermal equilibrium at the tropopause and the temperature of the lower stratosphere. *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences* 197, 487–505. <http://dx.doi.org/10.1098/rspa.1949.0076>.
- Ingersoll, A.P., 1969. The runaway greenhouse: a history of water on Venus. *Journal of Atmospheric Sciences* 26, 1191–1198.
- Komabayasi, M., 1967. Discrete equilibrium temperatures of a hypothetical planet with the atmosphere and the hydrosphere of one component-two phase system under constant solar radiation. *Journal of the Meteorological Society of Japan* 45, 137–138.
- Manabe, S., Strickler, R.F., 1964. Thermal equilibrium of the atmosphere with a convective adjustment. *Journal of Atmospheric Sciences* 21, 361–385.
- Manabe, S., Wetherald, R.T., 1967. Thermal equilibrium of the atmosphere with a given distribution of relative humidity. *Journal of Atmospheric Sciences* 24, 241–259.
- Nakajima, S., Hayashi, Y.-Y., Abe, Y., 1992. A study of the “runaway greenhouse effect” with a one-dimensional radiative–convective model. *Journal of Atmospheric Sciences* 49, 2256–2266.
- Rasool, S.I., de Berg, C., 1970. The runaway greenhouse and the accumulation of CO_2 in the Venus atmosphere. *Nature* 322, 1037–1045.
- Renno, N.O., 1997. Multiple-equilibria in radiative–convective atmospheres. *Tellus* 49A, 423–438.
- Renno, N.O., Emanuel, K.A., Stone, P.H., 1994. A radiative–convective model with an explicit hydrologic cycle: 1. Formulation and sensitivity to model parameters. *Journal of Geophysical Research* 99, 14429–14441.

Volcanoes: Role in Climate

A Robock, Rutgers University, New Brunswick, NJ, USA

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

Large volcanic eruptions inject sulfur gases into the stratosphere, which convert to sulfate aerosols with an e-folding residence time of about 1 year. The radiative and chemical effects of this aerosol cloud produce responses in the climate system. Volcanic eruptions produce global cooling, and are an important natural cause of interdecadal and interannual climate changes. Regional responses include winter warming of Northern Hemisphere (NH) continents following major tropical eruptions and weakening of summer Asian and African monsoons following tropical and NH high latitude eruptions. The volcanic cloud also produces stratospheric ozone depletion. Very large, but rare, eruptions, such as that of Toba 74 000 years ago, may have caused very large climate changes.

Introduction

Explosive volcanic eruptions affect climate by injecting gases and aerosol particles into the stratosphere. Only if the eruption cloud is rich in SO₂, will the eruption produce a long-lived cloud of sulfate aerosols that forms over the next few weeks. Otherwise, explosive eruptions that only produce large ash particles, such as the 1980 Mt. St. Helens eruption, can produce a large local weather perturbation but do not have long-lasting climatic effects. Some volcanoes, such as Kilauea and Etna, produce large quiescent tropospheric emissions of sulfate aerosols, but only if there is a dramatic change in these emissions will climate be changed. Stratospheric aerosol clouds last for several years, reflecting sunlight and cooling the surface. These clouds also absorb both solar (near infrared) and terrestrial radiation, heating the lower stratosphere. Volcanic aerosols also serve as surfaces for heterogeneous chemical reactions that destroy stratospheric ozone, which lowers ultraviolet (UV) absorption and reduces the radiative heating in the lower stratosphere, but the net effect is still heating. This also allows more UV radiation to reach the surface. As this chemical effect depends on the presence of anthropogenic chlorine, it has become important only in recent decades. Tropical eruptions produce asymmetric stratospheric heating, producing a stronger polar vortex and associated positive mode of the Arctic Oscillation in tropospheric circulation. This pattern is one of enhanced warm advection over Northern Hemisphere (NH) continents in winter, producing winter warming after large tropical eruptions. There is no evidence that volcanic eruptions can produce El Niños, but El Niño–Southern Oscillation (ENSO) variations must be considered when searching the climatic record for volcanic signals, as they have similar amplitudes and timescales.

There have been several large volcanic eruptions in the past 250 years (Table 1), and each has drawn attention to the atmospheric and potential climatic effects. The 1783 Laki eruption in Iceland produced large effects in Europe causing Benjamin Franklin, the United States ambassador to France, to publish the first paper on the subject in more than 1800 years. The 1815 Tambora eruption, combined with the effects of the unknown 1809 eruption, produced the ‘Year Without a Summer’ in 1816 and inspired *Frankenstein*, written by

Table 1 Major volcanic eruptions of the past 250 years

<i>Volcano</i>	<i>Year of eruption</i>
Laki (or Lakagigar), Iceland	1783
Unknown	1809
Tambora, Sumbawa, Indonesia	1815
Cosiguina, Nicaragua	1835
Askja, Iceland	1875
Krakatau, Indonesia	1883
Okataina (Tarawera), North Island, New Zealand	1886
Santa María, Guatemala	1902
Ksudach, Kamchatka, Russia	1907
Novarupta (Katmai), Alaska, United States	1912
Gunung Agung, Bali, Indonesia	1963
Mt. St. Helens, Washington, United States	1980
El Chichón, Chiapas, Mexico	1982
Mt. Pinatubo, Luzon, Philippines	1991

Mary Shelley on the shores of Lake Geneva, Switzerland, that summer. The 1883 Krakatau eruption was the largest explosion ever observed, and the sound wave was tracked on microbarographs for four complete circuits of the Earth, taking almost 2 days for one circuit. The Royal Society report on this eruption published 5 years later remains the most extensive report on the atmospheric effects of a volcanic eruption. The 1963 Agung eruption produced the largest stratospheric dust veil in more than 50 years in the Northern Hemisphere, and inspired many modern scientific studies. The subsequent 1982 El Chichón and 1991 Mt. Pinatubo eruptions produced very large stratospheric aerosol clouds and large climatic effects. Quantification of the size of these eruptions is difficult, as different measures reveal different information. For example, one could examine the total mass ejected, the explosiveness, or the sulfur input to the stratosphere.

Volcanic Emissions

Volcanic eruptions inject several different types of particles and gases into the atmosphere (Figure 1). These volatile inputs can be assessed based on measurements from active, but not explosive, eruptions, and remote sensing of the resulting

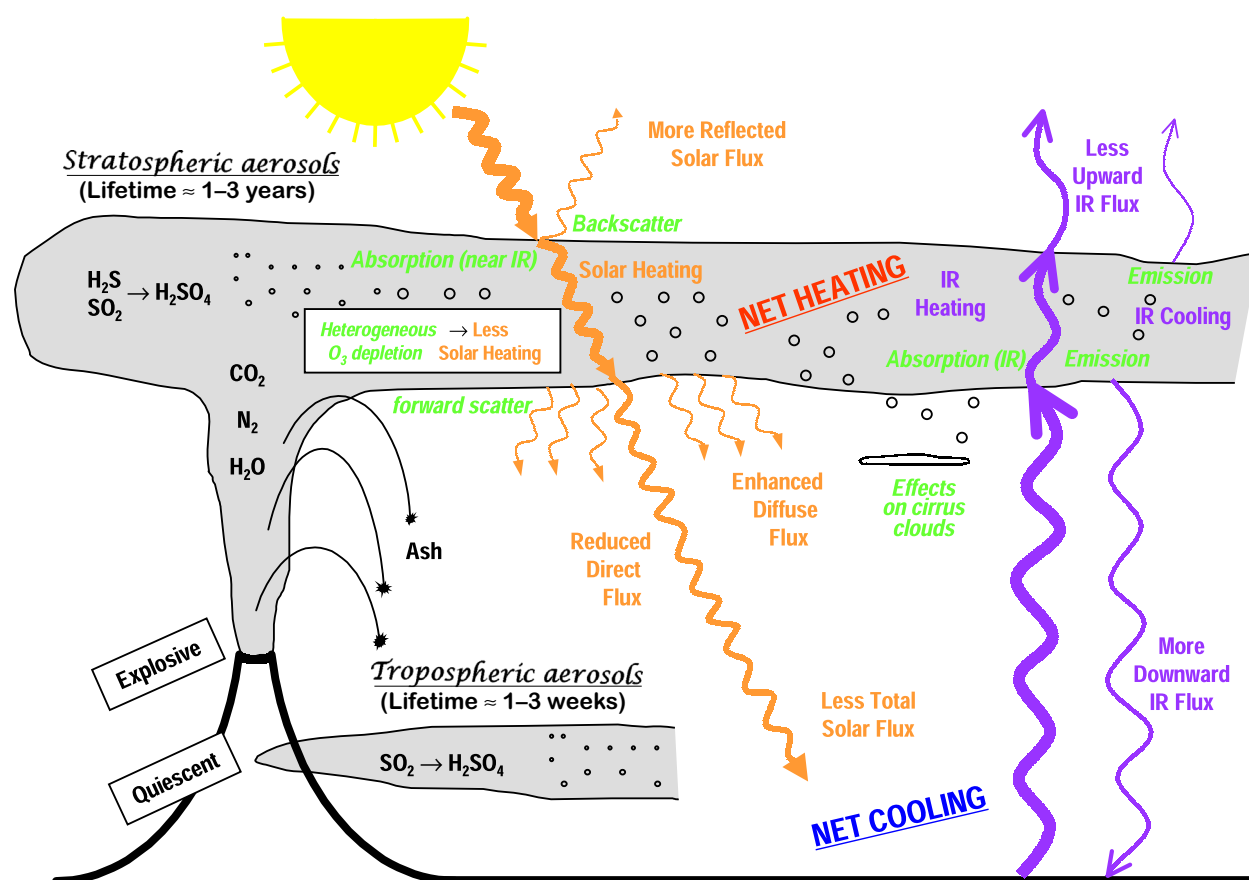


Figure 1 Schematic diagram of volcanic inputs to the atmosphere and their effects. Reproduced from Robock, A., 2000. Volcanic eruptions and climate [Plate 1]. Reviews of Geophysics 38: 191–219 [From which this article was condensed and updated], with permission from American Geophysical Union, copyright American Geophysical Union.

aerosol clouds from lidar, radiometers, and satellites. Several satellites routinely monitor SO_2 , allowing the direct measure of stratospheric injection of gases from eruptions and measure the resulting aerosol clouds.

The major component of volcanic eruptions is the matter which emerges as solid, lithic material, or solidifies into large particles, which are referred to as ash or tephra. These particles fall out of the atmosphere very rapidly, on timescales of minutes to a few days, and thus have no climatic impacts. When an eruption column still laden with these hot particles descends down the slopes of a volcano, this pyroclastic flow can be deadly to those unlucky enough to be at the base of the volcano. The destruction of Pompeii and Herculaneum after the AD 79 Vesuvius eruption is the most famous example.

Volcanic eruptions typically also emit gases, with H_2O , N_2 , and CO_2 being the most abundant. Over the lifetime of the Earth, these gases have been the main source of the planet's atmosphere and ocean, after the primitive atmosphere of hydrogen and helium was lost to space. The water has condensed into the oceans, the CO_2 has been changed by plants into O_2 or formed carbonates which sink to the ocean bottom, and some of the carbon has turned into fossil fuels. Of course, humans eat the plants and animals that eat the plants, drink the water, and breathe the oxygen, so each human is made of volcanic emissions. The atmosphere is now mainly

composed of N_2 (78%) and O_2 (21%), both of which had sources in volcanic emissions.

Of these abundant gases, both H_2O and CO_2 are important greenhouse gases, but their atmospheric concentrations are so large (even for CO_2 at only 392 ppm in 2011) that individual eruptions have a negligible effect on their concentrations and do not directly impact the greenhouse effect. Rather the most important climatic effect of explosive volcanic eruptions is through their emission of sulfur species to the stratosphere, mainly in the form of SO_2 , but possibly sometimes as H_2S . These sulfur species react with H_2O to form H_2SO_4 on a time-scale of weeks, and the resulting H_2SO_4 aerosols produce the dominant radiative effect from volcanic eruptions.

The 1982 El Chichón eruption injected 7 Mt of SO_2 into the atmosphere. There has not been a large stratospheric injection since 1991, when Mt. Pinatubo in the Philippines put about 20 Mt of SO_2 into the lower stratosphere. In 2008 Kasatochi (in the Aleutian Islands of Alaska), and in 2009 Mt. Sarychev (in the Russian Kamchatka Peninsula), and in 2011 Nabro (in Eritrea) each put about 1.5 Mt SO_2 into the lower stratosphere, but that was not enough to have any detectable climatic influence. The Eyjafjallajökull eruption in Iceland in 2010, while very disruptive of air traffic for weeks, had so little SO_2 , and with a short lifetime of a week or so in the troposphere, that it had no impact on climate.

Once injected into the stratosphere, the large aerosol particles and small ones being formed by the sulfur gases are rapidly transported around the globe by stratospheric winds. Observations after the 1883 Krakatau eruption showed that the aerosol cloud circled the globe in 2 weeks. Both the 1982 El Chichón cloud and the 1991 Pinatubo cloud circled the globe in 3 weeks. Although El Chichón (17° N) and Pinatubo (15° N) are separated by only 2° of latitude, their clouds, after only one circuit of the globe, ended up separated by 15° of latitude, with the Pinatubo cloud straddling the Equator and the El Chichón cloud extending approximately from the Equator to 30° N. Subsequent dispersion of a stratospheric volcanic cloud depends heavily on the particular distribution of winds at the time of eruption. For trying to reconstruct the effects of older eruptions, this factor adds a further complication, as the latitude of the volcano is not sufficient information.

Quiescent continuous volcanic emissions also add sulfates to the troposphere, but their lifetimes there are much shorter, although longer than anthropogenic sulfates as they are emitted from the sides of mountains rather than at the surface. The local pollution produced by the emission of the Kilauea crater on the Big Island of Hawaii is called 'vog' (volcanic fog). Global sulfur emission of volcanoes to the troposphere is about 14% of the total natural and anthropogenic emission, producing cooling at the surface. Only if there is a long-term trend in these emissions, will they be important for climate change; nevertheless, they must be considered when evaluating the effects of anthropogenic sulfate emissions.

Radiative Interactions and Climate Forcing

The major effect of a volcanic eruption on the climate system is the effect of the stratospheric cloud on solar radiation (Figure 1). Some of the radiation is scattered back to space, increasing the planetary *albedo* and cooling the Earth-atmosphere system. The sulfate aerosol particles (typical effective radius of 0.5 μm , about the same size as visible light) also forward scatter much of the solar radiation, reducing the direct solar beam, but increasing the brightness of the sky. After the 1991 Pinatubo eruption, the sky around the Sun appeared more white than blue because of this. Figure 2 illustrates this effect using observations from the Mauna Loa observatory in Hawaii. After the El Chichón eruption of 1982 and the Pinatubo eruption of 1991, the direct radiation was significantly reduced, but the diffuse radiation was enhanced by almost as much. Nevertheless, the volcanic aerosol clouds reduced the total radiation received at the surface. The El Chichón radiative effect at Hawaii was larger than that of Pinatubo, because its cloud was centered at the latitude of Hawaii, while only the edge of the larger Pinatubo cloud was monitored.

As the Sun sets, the yellow and red light (because Rayleigh scattering removes the shorter wavelengths in the process that produces the blue sky) is reflected from the bottom of stratospheric volcanic clouds, producing a characteristic yellow and red sky $1/2$ –1 h after the time of sunset (Figure 3). This effect has been used in the past to detect distant eruptions and to estimate the height of the aerosol cloud and its extent.

Climatic Impact of Volcanic Aerosols

Stratospheric aerosol clouds from volcanic eruptions cool the Earth's surface for several years, but produce winter warming over the continents in the NH. These and other effects are summarized in Table 2. Volcanic aerosols can be important causes of temperature changes for several years following large eruptions, and even on a 100-year timescale, they can be important when their cumulative effects are taken into account. This is very significant in analyzing the global warming problem, as the impacts of anthropogenic greenhouse gases and aerosols on climate must be evaluated against a background of continued natural forcing of the climate system from volcanic eruptions, solar variations, and internal random variations from land-atmosphere and ocean-atmosphere interactions. Individual large eruptions produce global or hemispheric cooling for 2 or 3 years, but the winter following a large tropical eruption is warmer over the NH continents, and this counterintuitive effect is due to a nonlinear response through atmospheric dynamics. The winter warming pattern is illustrated in Figure 4, which shows the global lower tropospheric temperature anomaly pattern for the NH winter of 1991–92, following the 1991 Mt. Pinatubo eruption. This pattern is closely correlated with the surface air temperature pattern where the data overlap, but the satellite data allow global coverage. The temperature over North America, Europe, and Siberia was much higher than normal, and that over Alaska, Greenland, the Middle East, and China was lower than normal. In fact, it was so cold that winter that it snowed in Jerusalem, a very unusual occurrence. Coral at the bottom of the Red Sea died that winter, because the water at the surface cooled and convectively mixed the entire depth of the water. The enhanced supply of nutrients produced anomalously large algal and phytoplankton blooms, which smothered the coral. This coral death had only happened before in winters following large volcanic eruptions.

At the tropopause, the boundary between the troposphere and the stratosphere, the strongest winds are found in the midlatitudes in the winter and are called the jet stream or polar vortex. The strength of the jet stream depends on the temperature difference (gradient) between the tropics and the polar region, which is largest in the winter when the polar regions cool. For a tropical eruption, the stratospheric heating from volcanic aerosols is larger in the tropics than in the high latitudes, producing an enhanced Pole-to-Equator temperature gradient, and in the NH winter, a stronger polar vortex and winter warming of NH continents. The stronger jet stream produces a characteristic wind pattern in the troposphere, which warms some regions and cools other ones. This pattern is called the 'Arctic Oscillation' and is the dominant mode of tropospheric variability. Tropical eruption clouds push the atmosphere into the positive phase of this natural variation. This indirect advective effect on temperature is stronger than the radiative cooling effect that dominates at lower latitudes and in the summer.

Because volcanic aerosols normally remain in the stratosphere for no more than 2 or 3 years, the radiative effect of volcanoes is interannual rather than interdecadal in scale. A series of volcanic eruptions could, however, raise the mean optical depth significantly over a longer period and thereby

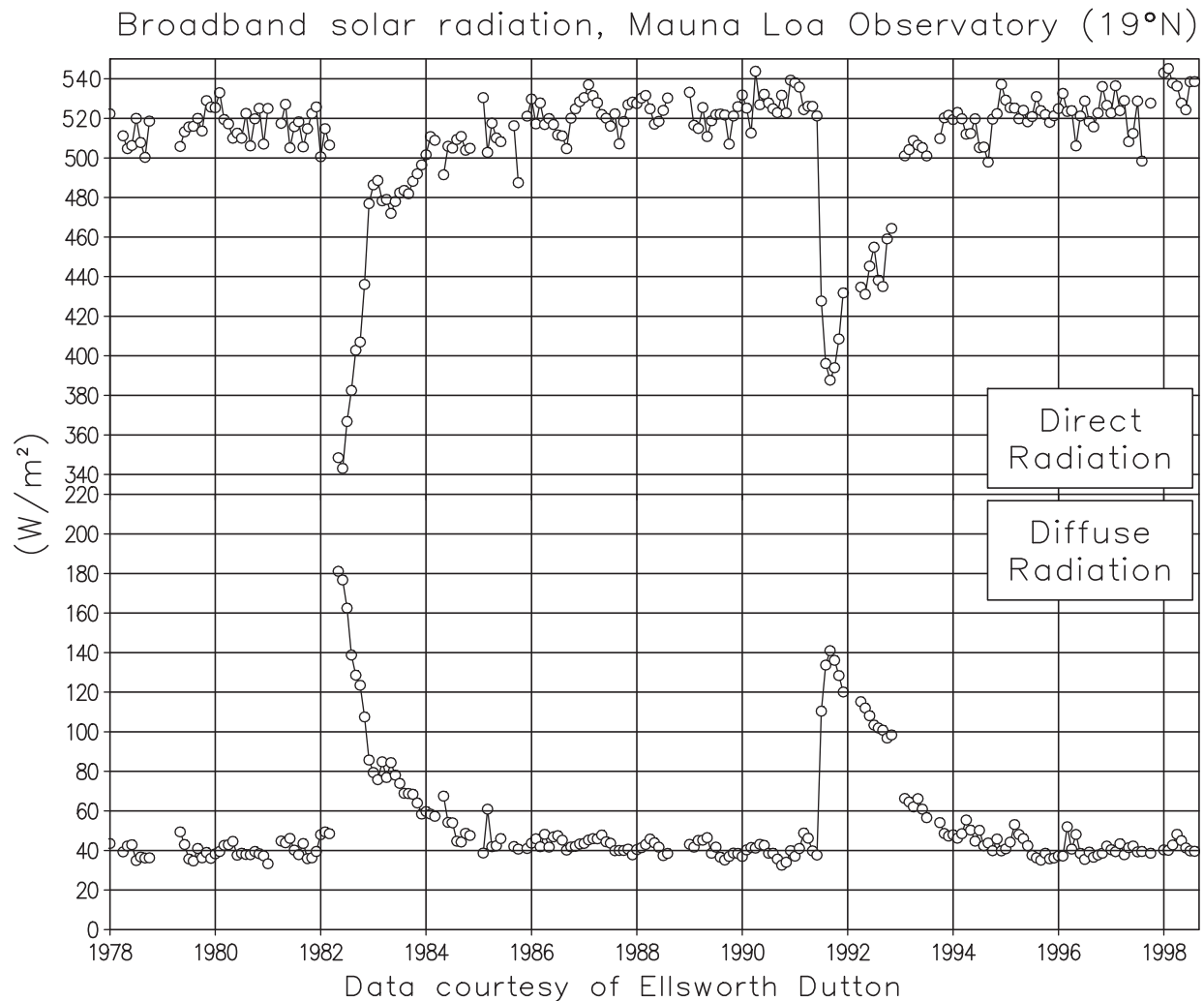


Figure 2 Direct and diffuse broadband radiation measurements from the Mauna Loa observatory, measured with a tracking pyrheliometer and shade disk pyranometer on mornings with clear skies at solar zenith angle of 60°, equivalent to two relative air masses. The reduction of direct radiation and enhancement of diffuse radiation after the 1982 El Chichón and 1991 Pinatubo eruptions are clearly seen. Years on abscissa indicate January of that year. Data courtesy of E Dutton. Reproduced from Robock, A., 2000. Volcanic eruptions and climate. *Reviews of Geophysics* 38: 191–219, with permission from American Geophysical Union, copyright American Geophysical Union.

give rise to a decadal-scale cooling. If a period of active volcanism ends for a significant period, such as for the 51-year period from 1912 to 1963 when global climate warmed, the adjustment of the climate system to no volcanic forcing could produce warming. Furthermore, it is possible that feedbacks involving ice and ocean, which act on longer timescales, could transform the short-term volcanic forcing into a longer term effect. As a result, the possible role of volcanoes in decadal-scale climate change remains unclear. However, the current century is the warmest of the past 10 centuries, with the previous several centuries called the Little Ice Age due to their coldness. Studies show that the interannual and interdecadal variations during this period were strongly affected by both volcanic eruptions and solar variations. The large warming of the past century, however, can only partially be explained by these natural causes, and in fact the second half of the twentieth century would have cooled due to volcanic eruptions if there had been no human emissions. The large warming of this

period can only be explained by including the effects of warming from anthropogenic greenhouse gases.

The Toba eruption 74 000 years ago, the largest volcanic eruption of the past 100 000 years, may have produced such large climate changes that it killed most humans on the planet, producing a genetic bottleneck such that all humans are descended from the same small number of survivors. This eruption on the island of Sumatra left a caldera about 86 km long and 30 km wide, with a large island inside, the resurgent block of the caldera. It erupted 1000 times more rock than the 1980 Mt. St. Helens eruption and injected approximately 100–300 times the amount of SO₂ into the stratosphere than the 1991 Pinatubo eruption. Depending on the assumptions about the properties of the resulting sulfate particles in the stratosphere, climate model simulations produce global average coolings of 3–15°C lasting for a decade or more. The larger values would have certainly been devastating for many species.



Figure 3 Volcanic sunset over Lake Mendota in Madison, Wisconsin, in July 1982, 3 months after the El Chichón eruption. Photograph by Alan Robock. Reproduced from Robock, A., 2000. Volcanic eruptions and climate [Plate 4]. *Reviews of Geophysics* 38: 191–219, with permission from American Geophysical Union, copyright American Geophysical Union.

Table 2 Effects of large explosive volcanic eruptions on climate

<i>Effect/mechanism</i>	<i>Begins</i>	<i>Duration</i>
Stratospheric warming <i>Stratospheric absorption of shortwave and longwave radiation</i>	1–3 months	1–2 years
Global cooling <i>Blockage of shortwave radiation</i>	Immediately	1–3 years
Global cooling from multiple eruptions <i>Blockage of shortwave radiation</i>	Immediately	Up to centuries
Winter warming of NH continents <i>Differential stratospheric heating, dynamical interaction with troposphere</i>	$\frac{1}{2}$ – $1\frac{1}{2}$ years	One or two winters
Reduced tropical precipitation <i>Blockage of shortwave radiation, reduced evaporation</i>	Immediately	~1 year
Reduction of Asian and African summer monsoon <i>Continental cooling, reduction of land–sea temperature contrast</i>	$\frac{1}{2}$ –1 year	One or two summers
Ozone depletion, enhanced UV <i>Dilution, heterogeneous chemistry on aerosols</i>	1 day	1–2 years

Ozone Impacts

Volcanic aerosols have the potential to change not only the radiative flux in the stratosphere, but also its chemistry. The most important chemical changes in the stratosphere are related to O_3 , which has significant effects on UV and longwave radiative fluxes. The reactions that produce and destroy O_3 depend on the UV flux, the temperature, and the presence of surfaces for

heterogeneous reactions, all of which are changed by volcanic aerosols. The heterogeneous chemistry responsible for the ozone hole over Antarctica in October each year occurs on polar stratospheric clouds of water or nitric acid, which only occur in the extremely cold isolated spring vortex in the Southern Hemisphere. Conditions in the NH are now changing and small O_3 depletions are being observed in spring there now, too. Reactions on polar stratospheric clouds make anthropogenic

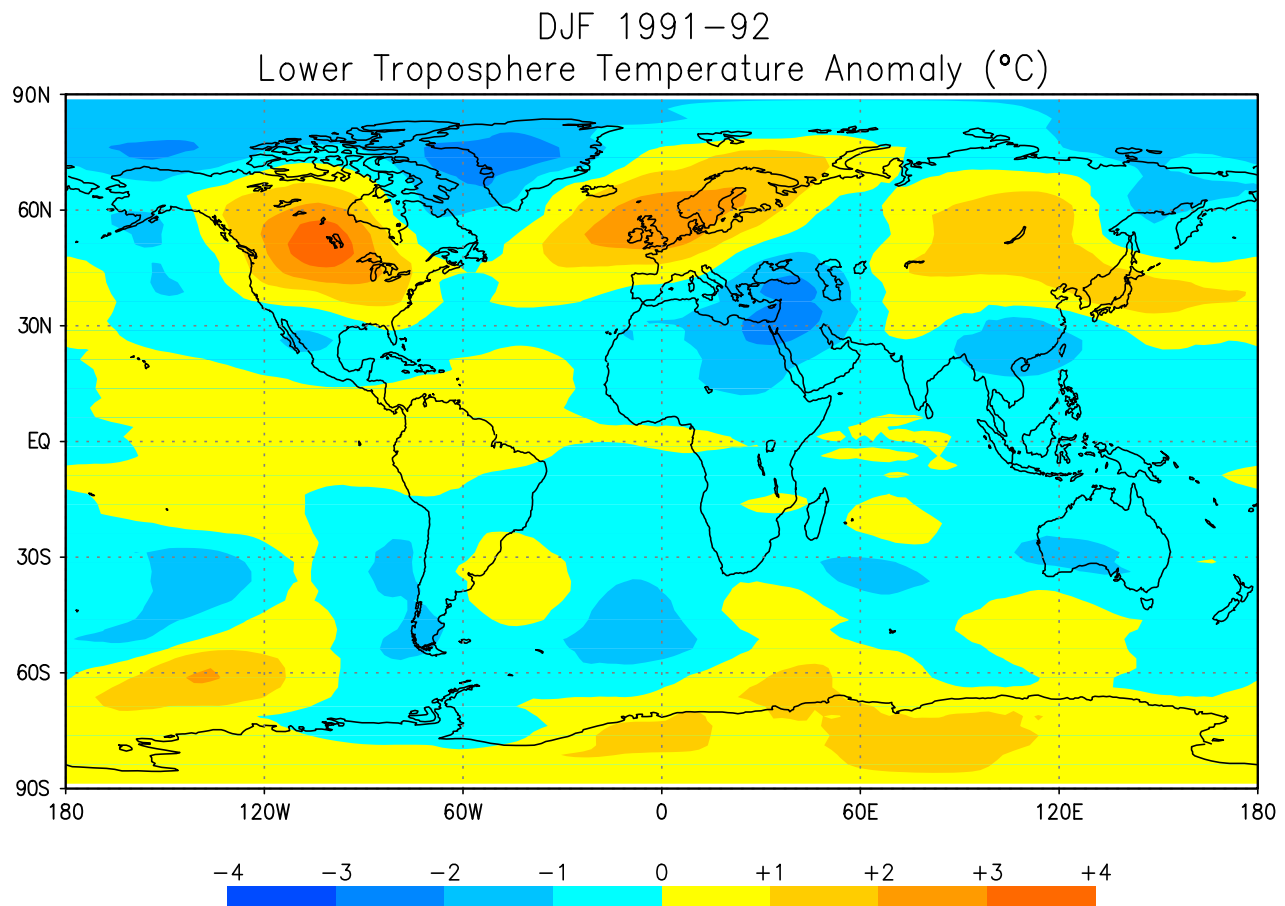


Figure 4 Winter (Dec.–Jan.–Feb. (DJF)) lower tropospheric temperature anomalies (with the nonvolcanic period of 1984–90 used to calculate the mean) for the 1991–92 NH winter (DJF) following the 1991 Mt. Pinatubo eruption. This pattern is typical of that following all large tropical eruptions, with warming over North America, Europe, and Siberia, and cooling over Alaska, Greenland, the Middle East, and China. Data from Microwave Sounding Unit Channel 2R, updated courtesy of J Christy and now called Channel 2LT. Reproduced from Robock, A., 2000. Volcanic eruptions and climate. *Reviews of Geophysics* 38: 191–219 with permission from American Geophysical Union, copyright American Geophysical Union.

chlorine available for chemical destruction of O_3 . However, sulfate aerosols produced by volcanic eruptions can also provide these surfaces at lower latitudes and at all times of the year. In fact, after the 1991 Pinatubo eruption, column O_3 reduction of about 5% was observed, ranging from 2% in the tropics to 7% in the midlatitudes. Therefore, ozone depletion in the aerosol cloud is much larger and reaches 20%. The chemical ozone destruction is less effective in the tropics, but lifting of low ozone concentration layers with the aerosol cloud causes a fast decrease in ozone mixing ratio in the low latitudes.

Decrease of the ozone concentration following volcanic eruptions causes less UV absorption in stratosphere, which modifies the aerosol heating effect. The net effect of volcanic aerosols on the surface UV flux is to increase it, as the aerosols backscatter less UV than the subsequent O_3 depletion allows through. The reduced O_3 absorption of shortwave and longwave radiations reduces the stratospheric heating effect and can affect the winter warming phenomenon described above.

The volcanic effect on O_3 chemistry is a new phenomenon, depending on anthropogenic chlorine in the stratosphere. While there are no observations available, the 1963 Agung

eruption probably did not deplete O_3 , as there was little anthropogenic chlorine in the stratosphere. Due to the Montreal protocol and subsequent international agreements, chlorine concentration has peaked in the stratosphere and is now decreasing. Therefore, for the next few decades, large volcanic eruptions will have effects similar to Pinatubo, but after that, these O_3 effects will go away and volcanic eruptions will have a stronger effect on atmospheric circulation without the negative feedback produced by O_3 depletion.

Discussion

There is no evidence that volcanic eruptions produce El Niño events, but the climatic effects of El Niño and volcanic eruptions must be separated to understand the climatic response to each. It had been suggested that the simultaneous appearance of the large 1982–83 El Niño and the 1982 El Chichón eruption and the 1991 smaller El Niño and the Pinatubo eruption suggested a cause and effect relationship. However, no plausible mechanism has been suggested and further research into the oceanography of El

Niños shows that these started before the volcanic eruptions. Examination of the entire record of past El Niños and volcanic eruptions for the past two centuries also shows no significant correlation.

As volcanic eruptions and their subsequent climatic response represent a large perturbation to the climate system over a relatively short period, observations and the simulated model responses can serve as important analogs for understanding the climatic response to other perturbations. While the climatic response to explosive volcanic eruptions is a useful analog for some other climatic forcings, there are also limitations. For example, successful climate model simulations of the impact of one eruption can help validate models used for seasonal and interannual predictions. But they cannot test all the mechanisms involved in global warming over the next century, as long-term oceanic feedbacks are involved, which have a longer timescale than the response to individual volcanic eruptions. Theory tells that volcanic eruptions also will produce multidecadal impacts on oceanic heat content, but these impacts are small and cannot be separated from other factors in observations.

The theory of ‘nuclear winter,’ the climatic effects of a massive injection of soot aerosols into the atmosphere from fires following a global nuclear holocaust, includes upward injection of the aerosols to the stratosphere, rapid global dispersal of stratospheric aerosols, heating of the stratosphere, and cooling at the surface under this cloud. As this theory cannot be tested in the real world, volcanic eruptions provide analogs that support these aspects of the theory.

Recent suggestions of considering the use of geoengineering to control global climate through the creation of a permanent stratospheric aerosol cloud have used volcanic eruptions as an analog. While volcanic eruptions indeed cool the surface, they also produce ozone depletion and drought, thus raising cautions about the wisdom of such ideas. There are also many other reasons why geoengineering may be a bad idea.

Given the current understanding of the climatic impact of volcanic eruptions, it can be safely predicted that following the next large tropical eruption, there will be global cooling for about 2 years, and winter warming of the NH continents for 1 or 2 years. There will also be reduced summer monsoon precipitation over Asia and Africa. A large NH high latitude

eruption, if it occurs in spring or summer, will also produce a weak summer monsoon.

See also: **Aerosols:** Climatology of Stratospheric Aerosols; Observations and Measurements; Role in Climate Change; Role in Radiative Transfer. **Chemistry of the Atmosphere:** Observations for Chemistry (Remote Sensing); Lidar; Volcanoes: Composition of Emissions. **Climate and Climate Change:** Climate Variability: Decadal to Centennial Variability; Climate Variability: North Atlantic and Arctic Oscillation; Nuclear Winter; Climate Variability: Seasonal and Interannual Variability. **Lidar:** Backscatter. **Satellites and Satellite Remote Sensing:** Aerosol Measurements; Measuring Ozone from Space: TOMS and SBUV. **Stratosphere/Troposphere Exchange and Structure:** Global Aspects. **Stratospheric Chemistry Topics:** Stratospheric Water Vapor. **Weather Forecasting:** Seasonal and Interannual Weather Prediction.

Further Reading

- Forsyth, P.Y., 1988. In the wake of Etna, 44 B.C. *Classical Antiq.* 7, 49–57.
- Franklin, B., 1784. Meteorological imaginations and conjectures. In: *Manchester Literary and Philosophical Society Memoirs and Proceedings*, 2 [Reprinted in *Weatherwise*, 35, p. 262, 1982] 122.
- Genin, A., Lazar, B., Brenner, S., 1995. Vertical mixing and coral death in the Red Sea following the eruption of Mount Pinatubo. *Nature* 377, 507–510.
- GRL Special Issue. *Geophysical Research Letters* 10, 1983, 989–1060 [Studies of the 1982 El Chichón eruption].
- GRL Special Issue. *Geophysical Research Letters* 19, 1992, 149–218 [Studies of the 1991 Mt. Pinatubo eruption].
- Lamb, H.H., 1970. Volcanic dust in the atmosphere; with a chronology and assessment of its meteorological significance. *Philos. Trans. Royal Society London A266*, 425–533.
- Robock, A., 2000. Volcanic eruptions and climate. *Reviews of Geophysics* 38, 191–219 [From which this article was condensed and updated].
- Robock, A., Ammann, C.M., Oman, L., et al., 2009. Did the Toba volcanic eruption of ~74 ka B.P. produce widespread glaciation? *Journal of Geophysical Research* 114, D10107. doi:10.1029/2008JD011652.
- Siebert, L., Simkin, T., Kimberly, P., 2011. *Volcanoes of the World*, third ed. University of California Press, Berkeley, CA, p. 568.
- Stommel, H., Stommel, E., 1983. *Volcano Weather, the Story of 1816, the Year without a Summer*. Seven Seas Press, Newport, RI, p. 177.
- Symons, G.J., 1888. *The Eruption of Krakatoa, and Subsequent Phenomena*. Trübner, London, UK, p. 494.

CLOUDS AND FOG

Contents

Cloud Modeling

Contrails

Cloud Microphysics

Classification of Clouds

Climatology

Measurement Techniques *In Situ*

Fog

Noctilucent Clouds

Stratus and Stratocumulus

Cloud Modeling

W-K Tao, NASA/Goddard Space Flight Center, Greenbelt, MD, USA

M Moncrieff, National Center for Atmospheric Research, Boulder, CO, USA

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

Cloud resolving models (CRMs) use more sophisticated and realistic representations of cloud microphysical processes, and they can reasonably well resolve the time evolution, structure, and life cycles of clouds and cloud systems (with sizes ranging from about 2 to 200 km). CRMs also allow for explicit interaction between clouds, outgoing longwave (cooling) and incoming solar (heating) radiation, and ocean and land surface processes. CRMs can be used to improve our understanding of: (1) convective organization, (2) cloud temperature and water vapor budgets, (3) convective momentum transport, (4) diurnal variation of precipitation processes, and (5) water and energy cycles.

Introduction

Numerical cloud models, which are based on the non-hydrostatic equations of motion, have been extensively applied to cloud-scale and mesoscale processes during the past five decades. Uncertainties stemming from convection that must be parameterized in (hydrostatic) large-scale models are mitigated by cloud models, which simulate the salient small-scale dynamics. Global weather and climate models will need to use the nonhydrostatic framework when their horizontal resolution gets to about 10 km, the theoretical limit of the hydrostatic approximation. This juncture will be reached within a decade.

The earliest kind of cloud model, the one-dimensional entraining bubble or plume that simply parameterizes the lateral entrainment of environmental air, was applied extensively to cloud seeding research, and is still used as a transport module for convective parameterization schemes. In the 1960s,

two-dimensional cloud models were developed to study cloud evolution in idealized environments. Three-dimensional cloud models, developed in the early 1970s, quantified the effects of wind shear on convection; for example, tropical squall lines, and midlatitude supercell thunderstorms associated with tornado genesis.

During the late 1970s and the early 1980s, cumulus ensemble models (also known as cloud-system resolving models) began to simulate the collective effects of moist convection on the large-scale environment, with emphasis on the Tropics. A primary objective was to improve cumulus parameterization, a quest that continues to this day. The effect of ice processes on cloud formation and evolution, stratiform rain processes, and their relation to convective rainfall were focal issues during this period, as was the organizing effect of environmental shear on mesoscale convective systems. The impact of radiative processes on cloud development was investigated in the late 1980s.

In the 1990s, cloud-system resolving models began to quantify interaction between the meso- and large-scale motion, cloud–chemistry interactions, cloud–climate interaction, and surface processes. An important application was in the design of satellite rainfall and latent heat retrieval algorithms for the Tropical Rainfall Measuring Mission (TRMM).

In the 2000s, cloud-system resolving models were used to study the aerosol–cloud and precipitation interactions and they were embedded into global models to study the water and energy cycles from local up to regional scale. **Table 1** summarizes the major highlights of model development over the past five decades.

Theoretical studies have advanced our basic knowledge of convection dynamics, and how convection interacts with the larger scales of motion. These studies enable the complexity of numerically simulated clouds to be reduced to first principles, which is essential to understand the role of moist processes in the Earth's weather and climate.

During the past generation, voluminous data on atmospheric convection accumulated from radar, instrumented aircraft, satellites, and rawinsonde measurements in field campaigns, enabled detailed model evaluations. Improved numerical methods have resulted in more accurate and computationally efficient dynamical cores. Improvements have been made in the parameterizations of microphysical processes, radiation and boundary layer effects, and turbulence; however, microphysical parameterizations remain a major source of uncertainty in all classes of atmospheric model.

Table 1 Highlights of cloud modeling development and application over the past five decades

<i>Major highlights</i>	
1960s	Loading, buoyancy, and entrainment
1970s	Slab- vs axis-symmetric model Cloud seeding
1980s	Cloud dynamics and warm rain Ensemble of clouds – cumulus parameterization Cloud interactions and mergers Ice processes Super cell dynamics Squall line Convective and stratiform interactions Wind shear and cool pool Gravity wave and density current Cloud radiation interaction
1990s	2D vs 3D Cloud-radiation quasi-equilibrium – climate variation implications Cloud transport and chemistry Diurnal variation of precipitation Gewex cloud system study (GCSS) Coupled with microwave radiative and radar model for satellite cloud retrieval (TRMM)
2000s	Land and ocean processes Multiscale interactions Energy and water cycle Cloud aerosol–chemistry interactions Cumulus parameterization improvements

Adapted from Tao, W.-K., 2003. Goddard Cumulus Ensemble (GCE) model: application for understanding precipitation processes. AMS Meteorol. Monogr. Cloud Syst. Hurricanes TRMM, 107–138.

In recent years, exponentially increasing computer power has extended cloud model integrations from hours to months, and the number of grid points from less than 1000 to more than 1 billion. Three-dimensional models are now more prevalent. Cloud models simulate the evolution, structure, and life cycle of cloud systems. They explicitly calculate interactions among clouds, longwave and solar radiation that are difficult, if not impossible, to observationally obtain. Much attention has been on precipitating cloud systems where the crucial meso-scale circulations are resolved by cloud-system resolving models. Observations provide both the initial conditions and data for model validation; model results provide statistical information useful for developing physically based parameterization for climate models and numerical weather prediction models.

Physical Processes in Cloud Models

Cloud-microphysical processes (phase changes of water and precipitation) must be parameterized in cloud models, as must atmospheric turbulence (dissipation of kinetic energy), turbulent processes at oceanic or terrestrial boundaries (latent and sensible heat fluxes into the atmosphere), and radiative transfer processes (complex in the presence of clouds).

Microphysics and Precipitation

Figure 1 depicts the widely used two-class liquid (cloud water and rain droplet) and three-class ice (cloud ice, snow, and graupel/hail) microphysics schemes. The shapes of liquid and ice particles are assumed to be spherical. The warm cloud microphysics assumes that the population of water particles is bimodal, consisting of small cloud droplets whose terminal velocity is minute compared to typical vertical air velocities, and large rain droplets that obey certain size distributions based

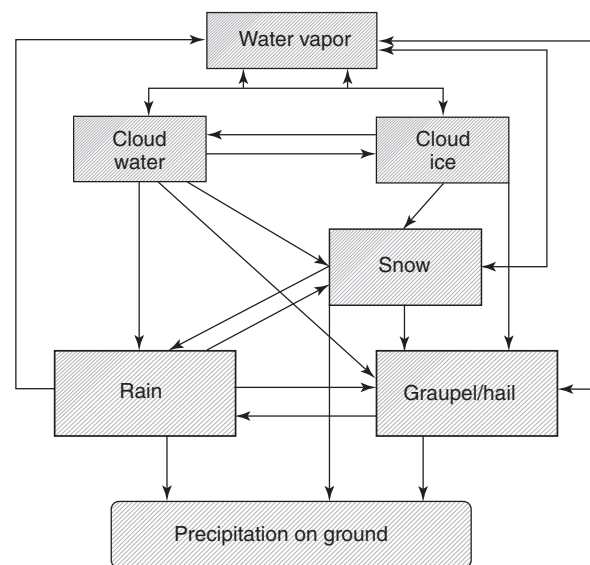


Figure 1 Representation of the three-class ice scheme used in the cloud model.

on limited observations. Condensation, evaporation, and autoconversion/collection processes (from small cloud droplets to large rain droplets) are parameterized. The ice microphysical schemes typically assume three classes of particle: small cloud ice, whose terminal velocity is also minute compared to typical vertical air velocities; snow, whose terminal velocity is about $1\text{--}3\text{ m s}^{-1}$; and large-sized graupel or hail with faster terminal velocities. Graupel has a lower density and a higher intercept (i.e., high number concentration). In contrast, hail has a high density and a small intercept. Graupel is representative of tropical clouds and hail of midlatitude clouds. More than 25 transfer processes between water vapor, liquid, and ice particles are included. These include the growth of ice crystals by riming, the aggregation of ice crystals, the formation of graupel and hail, the growth of graupel and hail by the collection of supercooled raindrops, the shedding of water drops from hail, the rapid growth of ice crystals in the presence of supercooled water, the melting of all forms of ice, and the deposition and sublimation of ice. Only large rain droplets, snow, and graupel/hail fall toward the ground as precipitation.

Only recently have some cloud-system resolving models adopted a two-moment four-class ice scheme that combines the main features of the three-class ice schemes by calculating the mixing ratios of both graupel and frozen drops/hail. Additional model variables include the number concentrations of all ice particles (small ice crystals, snow, graupel, and frozen drops), as well as the mixing ratios of liquid water for each of the precipitation ice species during wet growth and melting for purposes of accurate active and passive radiometric calculations.

In addition, explicit (spectral bin-microphysical) schemes have been developed for the study of cirrus development and cloud precipitation–aerosol interaction. The formulation of the explicit microphysical processes is based on solving stochastic kinetic equations for the size distribution functions of water droplets (cloud droplets and raindrops), and ice particles of different habits (columnar, platelike, dendrites, snowflakes, graupel, and frozen drops). Each type is described by a special size distribution function containing over 30 categories (bins). Nucleation (activation) processes are also based on the size distribution function for cloud condensation nuclei (also over 30 size categories). Because of the numerous interactions involved in spectral bin-microphysical schemes, computational domains are relatively small and simulation times short. These detailed microphysics calculations can provide a useful framework for evaluating and ultimately improving bulk microphysical schemes. Of particular interest in this regard is convectively generated cirrus, which affects the radiative properties of the tropical atmosphere.

Turbulence

Whilst large eddies are resolved in cloud models, eddies much smaller than the grid scale must be parameterized. An implicit assumption is that the small scales approximate to an inertial subrange where the energy spectrum is in statistical equilibrium, with an energy cascade from the resolved scales to the dissipation scales. The most sophisticated turbulence parameterization presently used is a third-moment closure. Typical cloud models used simple k -type (first-order) turbulence

closure or determine the coefficient k , diagnostically or prognostically, from the turbulence kinetic energy (TKE) equation (one-and-a-half order). In the prognostic TKE method, thermodynamic stability, deformation, shear stability, diffusion, dissipation, and transport of subgrid energy are included. In the diagnostic method, deformation and stability are used for computing the k coefficient.

Radiation

Emission and absorption by water vapor and cloud droplets are represented by two-stream long-wave radiative transfer schemes. Broadband methods for long-wave radiation combine the effects of reflection, emission, and transmission by cloud droplets and air molecules. The treatment of short-wave radiation is also based on broadband approximations. One key issue is how to parameterize cloud optical properties (optical thickness), especially in the presence of the ice phase, in view of the important impact of radiative heating and cooling profiles within clouds. Note that only limited observations are available upon which to base parameterizations for ice clouds.

The use of a fully explicit microphysics scheme (liquid and ice) and a fine horizontal resolution provides relatively realistic cloud optical properties, which are crucial for radiation budgets. With high spatial resolution, each atmospheric layer is considered either completely cloudy (overcast) or clear. No partial cloudiness is assumed.

Ocean Surface Processes

Two types of surface exchange (flux) schemes are typically used. The first is a simple bulk aerodynamic formula where the transfer coefficients for momentum, sensible heat, and latent heat fluxes are functions of wind speed only. The second type is more complex but, nevertheless, still a bulk approach. The transfer coefficients for momentum, sensible heat, and latent heat fluxes are based on the Monin–Obukhov similarity theory of the atmospheric surface layer. The parameters, such as the roughness lengths, are closely related to the sea surface characteristics and the turbulence characteristics. At very low wind speeds the similarity profile is singular, a problem addressed by adding a convective velocity to yield nonzero fluxes under windless conditions. The exchange coefficients in the simple bulk aerodynamic formula method and in the second bulk flux algorithm differ in two ways. First, in the lower wind speed regime (less than 4 m s^{-1}), the exchange coefficients in the complex bulk scheme increase with decreasing wind speed in order to account for the convective exchange at low wind speeds. Second, the coefficients in the simple bulk aerodynamic formula linearly increase with respect to the wind speed, but decrease if the wind speed is greater than 4 m s^{-1} in the more complex bulk schemes. These differences in the exchange coefficients can affect rainfall amounts and boundary layer structure.

Land Surface Processes

Detailed interactive land surface process models of the heterogeneous land surface (soil and vegetation) and adjacent near-surface atmosphere have recently been applied in cloud

models to study the effect of soil moisture distribution and atmospheric boundary conditions on cloud structure, rainfall, and soil moisture distribution. A land surface model usually has three elements: (1) a soil module that includes several water reservoirs, i.e., plant internal storage, dew/intercepted precipitation, surface material, a topsoil root layer, a subsoil root layer, and two deeper layers that regulate seasonal and interannual variability of the soil hydrology; (2) a surface slab of vegetation, litter, and other loose material that shades the soil and acts as the source for sensible heat flux, and intercepts precipitation and dew; and (3) the surface layer of the atmosphere (up to the lowest grid level of the model to which it is coupled), within which the fluxes of sensible heat and water vapor are calculated.

Modeling Tropical Convective Systems

While cloud models have been applied in numerous geographic locations, the most comprehensive studies have been conducted in the Tropics, often in association with observational field campaigns. Tropical convection affects the large-scale circulation of the atmosphere in many ways, including atmosphere–ocean coupling. The organization of tropical convection occurs across scales: cloud streets in the shallow trade-wind convection, cumulonimbus, cloud clusters, squall lines, tropical cyclone rainbands, multiscale convection in westerly wind bursts and intraseasonal oscillations, and enhanced convection in the intertropical convergence zone (i.e., the rising branch of the Hadley circulation) and in planetary-scale Walker circulation. Within the next few decades, computers will likely be powerful enough to resolve this multiscale hierarchy. Until this juncture many basic issues must be addressed in parameterization, including the role of convective organization and its implications for the scale-separation assumption at the root of convective parameterization schemes. However, cloud-system resolving models will not make convective parameterization redundant, but move the problem downscale to the planetary boundary layer and demand improved microphysical parameterizations.

Ensemble vs Local Convection

Convective cloud models can be broadly categorized as follows. First, a quasi-statistical approach where clouds of different types and/or in various stages of evolution are simulated in large computational domains (i.e., cumulus ensemble models or cloud-system resolving models). In this article, the terms cloud model, cloud-system resolving model, and cumulus ensemble model will be used interchangeably.

A prime objective is to quantify how convection interacts with the large scales of motion. For example, large-scale advective ‘forcing,’ the primary source of convective available potential energy, is derived from an objective analysis of sounding networks and applied as domain-mean tendencies of temperature and moisture (continuously forced convection). This approach has been used extensively in the west Pacific warm pool region Tropical Ocean Global Atmosphere–Coupled Ocean Atmosphere Response Experiment (TOGA COARE) and the eastern Atlantic Global Atmospheric Research Program’s Atlantic Tropical Experiment (GATE).

Second, in the initial-value approach (e.g., the classical cloud model), convective evolution is simulated over periods of hours. The initiation (or triggering) of convection is important: cold pools, surface fluxes, or stochastic perturbation locally force new convection. Such simulations are useful in their own right, for model development, and for model validation when used in conjunction with field campaigns.

Mesoscale Organization of Moist Convection

The past decades have witnessed advances in the understanding of organized convection, with convection over the tropical oceans being a focus. Figure 2(a) and 2(b) shows the evolution of numerically simulated convective cloud systems in the west Pacific warm pool region and eastern Atlantic region, respectively. In the former, cloud systems travel in one direction and embedded convection in the opposite direction. Synoptic-scale easterly waves modulate convection in the eastern Atlantic because they strongly affect environmental shear and large-scale forcing, and thereby convective organization and intensity. Shallow convection evolves to nonsquall (slow-moving) deep precipitating cloud systems that travel westward, steered by the mean wind. Fast-moving squall systems subsequently develop as the lower-tropospheric shear intensifies. The simulated cloud systems become less organized and produce less surface precipitation as the forcing decreases.

The simulated domain-averaged surface rainfall (mm) and stratiform amount (percentage) for both the west Pacific warm pool and the east Atlantic regions are shown in Table 2. The ratios between evaporation and condensation, sublimation and deposition, and deposition and condensation illustrate the relative importance of liquid vs ice processes and source and sink terms associated with water vapor. The microphysical processes are decomposed according to the regime of convective organization: slow-moving, fast-moving, less organized convective episodes in the east Atlantic, vigorous deep convection to gather with weaker convective events in the western Pacific during the convectively active phase of the intraseasonal oscillation where the stratiform component is larger. The dominance of warm rain processes in the east Atlantic region squall and nonsquall convective systems explains the smaller stratiform rain amounts. The depleted ice processes on 6 and 8 September are indicative of shallow convection. In contrast, ice processes are important for both active and relatively inactive convective periods over the west Pacific.

Figure 3 shows numerically simulated three-dimensional cloud systems over the west Pacific warm pool. Organized mesoscale convective systems consist of families of leading edge, quasi-linear, heavily precipitating cumulonimbus followed by an extensive area of trailing light (stratiform) precipitation. The cumulonimbus slope with height is due to the environmental wind shear. Mesoscale descent behind the leading edge develops mainly at low levels where evaporative cooling is strongest. Ascent occurs above the mesoscale ascent, typically separated by the 0 °C (melting) level. Evaporative-cooled downdraft outflows (density currents) provide localized uplift that triggers convection. Convectively generated gravity waves occur in the upper troposphere.

Cloud-system resolving models with large computational domains quantify how convection organized on mesoscales

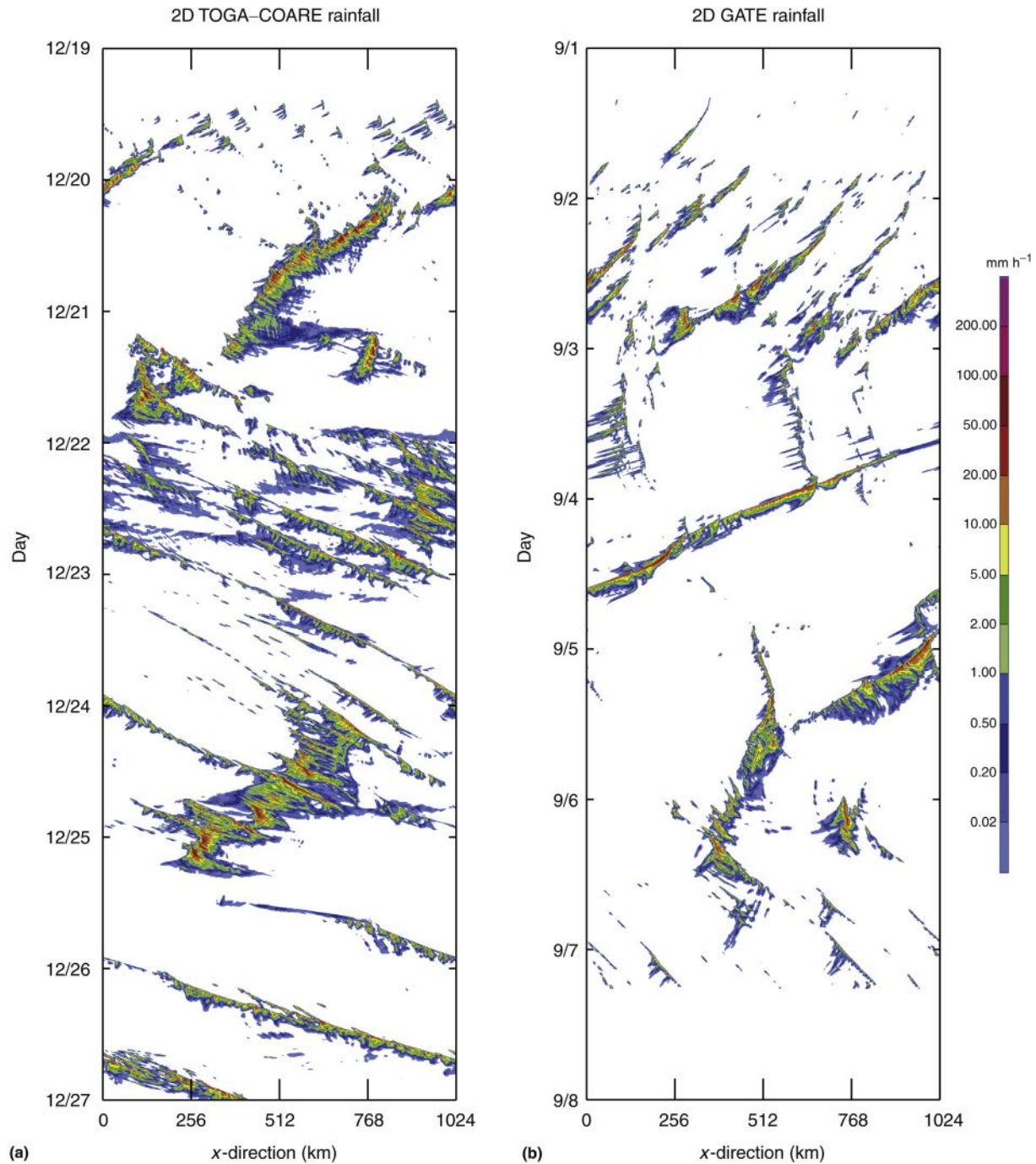


Figure 2 Time sequence of the two-dimensional cloud model estimated domain mean surface rainfall rate (10^{-3} m h^{-1}) for (a) the west Pacific warm pool region and (b) the east Atlantic region. Adapted from Tao, W.-K., 2003. Goddard Cumulus Ensemble (GCE) model: application for understanding precipitation processes. AMS Meteorol. Monogr. Cloud Syst. Hurricanes TRMM, 107–138.

interacts with large-scale tropical waves and intraseasonal oscillations. For example, Figure 4 is a realization of Madden-Julian-like tropical intraseasonal variability. Figure 4(a) shows an eastward-traveling convective envelope while Figure 4(b)–4(d) shows the westward-traveling organized convective systems embedded in the envelope and steered by the mean flow. The fact that multiscale convective

organization occurs even with a uniform sea surface temperature (SST) indicates that it is a self-organizing process.

The above modeling studies have quantified many observed properties of moist convection and its multiscale organization. The challenge now is to understand how this organization relates to the large-scale atmospheric circulation.

Table 2 Cloud model-simulated domain-average surface rainfall (mm), stratiform amount (percentage), and microphysical processes (ratios between evaporation and condensation, sublimation and deposition, and deposition and condensation) for (a) the west Pacific warm pool region and (b) the east Atlantic region. For west Pacific warm pool region, the cloud model results are also separated into subperiods, deep strong convection during 20–23 and 24–25 December and weaker convection prior to, between, and after the deep convection (19–20, 23–24, and 25–26 December 1992). Slow-moving (nonsquall, 2–4 September), fast-moving (squall, 4–6 September), and less-organized (6–8 September) periods for the cloud model-simulated east Atlantic region results are also shown

(a) West Pacific warm pool region (19–26 December 1992)			
	After WWB (2 days)	Pre and during WWB (4 days)	
Total surface rainfall (mm)	50.8	85.1	
Stratiform amount (%)	42.0	48.0	
Evaporation/condensation (%)	62.0	73.0	
Sublimation/deposition (%)	46.0	50.0	
Deposition/condensation (%)	34.0	41.0	

(b) East Atlantic region (2–8 September 1974)			
	Slow-moving	Fast-moving	Random
Total surface rainfall (mm)	43.34	39.62	8.50
Stratiform amount (%)	27.0	26.0	44.0
Evaporation/condensation (%)	58.0	44.0	68.0
Sublimation/deposition (%)	36.0	27.0	24.0
Deposition/condensation (%)	23.0	25.0	9.5

Adapted from Tao, W.-K., 2003. Goddard Cumulus Ensemble (GCE) model: application for understanding precipitation processes. AMS Meteorol. Monogr. Cloud Syst. Hurricanes TRMM, 107–138.

Heat and Moisture Budgets

The collective effects of subgrid convection are represented by parameterizations. The budgets of heat, moisture, convective mass flux, and convective momentum transport can be derived from observations. Heat and moisture transports by convection strongly affect large-scale atmospheric dynamics. Transports are estimated from the objective analysis of sounding networks as thermodynamic budget residuals the apparent sources of heat (Q_1) and moisture (Q_2). Figure 5(a) shows that convective heating has a maximum in the 600–650 hPa layer. In the stratiform region (Figure 5(b)), heating is maximum in the upper troposphere (around 400 hPa), with cooling prevailing below the melting level. The stratiform heating is smoother because convective bursts have a more rapid evolution than the mesoscale processes in the stratiform region. Also, stronger heating occurs aloft, and stronger cooling below in the stratiform region owing to the evaporation of rain produced by melting ice particles.

Figure 5(c) and 5(d) shows the apparent moisture sink Q_2 . Drying in the convective region is caused by the condensation processes associated with cloud updrafts, and its maximum is lower than the apparent heat source. In the stratiform region, there is strong moistening (by evaporative cooling) below the 600-hPa level with weak drying aloft. Cloud models show that the eddy transport of heat is one order smaller than the effects of the microphysical processes, contrasting with the eddy transport of moisture, which is of the same order. These

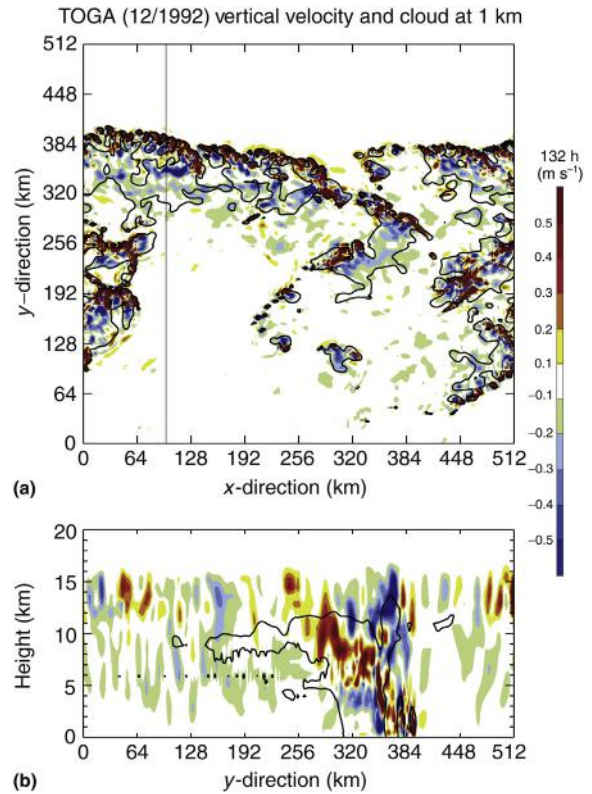


Figure 3 (a) Horizontal and (b) vertical cross-sections of vertical velocity (filled contours) and total cloud mixing ratio (solid contour) taken from a three-dimensional cloud model simulation of the west Pacific warm pool region precipitating system (during a Westerly Wind Burst episode). The location of the vertical cross-section shown by the vertical line in (a).

distinctive heating/moistening patterns are consistent with observed mesoscale convective systems. The small difference (balance) between the cloud processes (response/feedback) and large-scale forcing is indicative of the quasi-equilibrium state of the tropical atmosphere.

Convective Mass Flux

Convective mass flux is an important quantity in the parameterization of convection in large-scale models but difficult to observe accurately; therefore, cloud models are essential. Figure 6 shows the 7-day evolution of simulated cloud mass fluxes (total condensate exceeding 0.1 g kg^{-1}). The larger mass fluxes trace the organized cloud systems (nonsquall clusters, days 2 and 5; and squall line, day 4). Evaporative cooling associated with the downdrafts is about half of the condensational heating in the updrafts (Table 2). Convective updrafts account for approximately 75% of the cloud updraft mass flux, yet occupy a mere 12–14% of the total area; these so-called ‘hot towers’ play a critical role in the heat and moisture budgets in the tropics, despite the small fractional area they occupy. Downdrafts account for about 30% of the mass flux, suggesting they are active over relatively small areas.

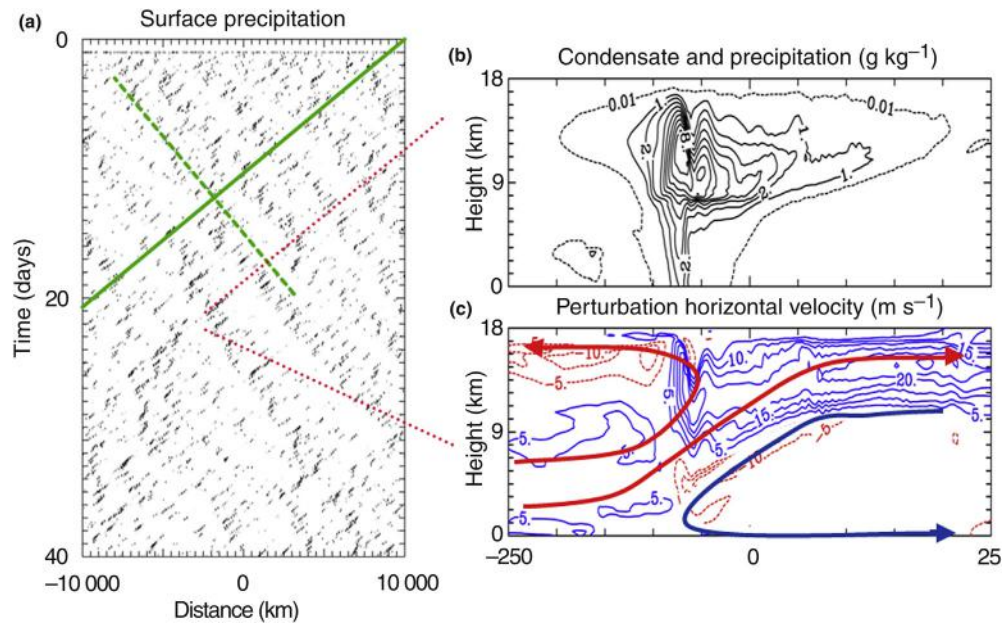


Figure 4 Clockwise from left: (a) Space-time distribution of surface precipitation simulated by a 40-day, two-dimensional 20 000 km domain cloud-system resolving model of tropical convection for a constant sea-surface temperature aquaplanet. Dotted green line indicates the eastward-propagating large-scale convective envelope, and full green line westward-propagating mesoscale convective systems; (b) shows the total condensate and precipitation distribution in a typical mesoscale system; and (c) shows the horizontal component of the system-relative airflow, and the flow organization. From Moncrieff, M.W., 2010. *The Multiscale Organization of Convection at the Intersection of Weather and Climate. Why Does Climate Vary?* AGU Geophysical Monograph Series, vol. 189, pp. 3–26. doi:10.1029/2008GM000838.

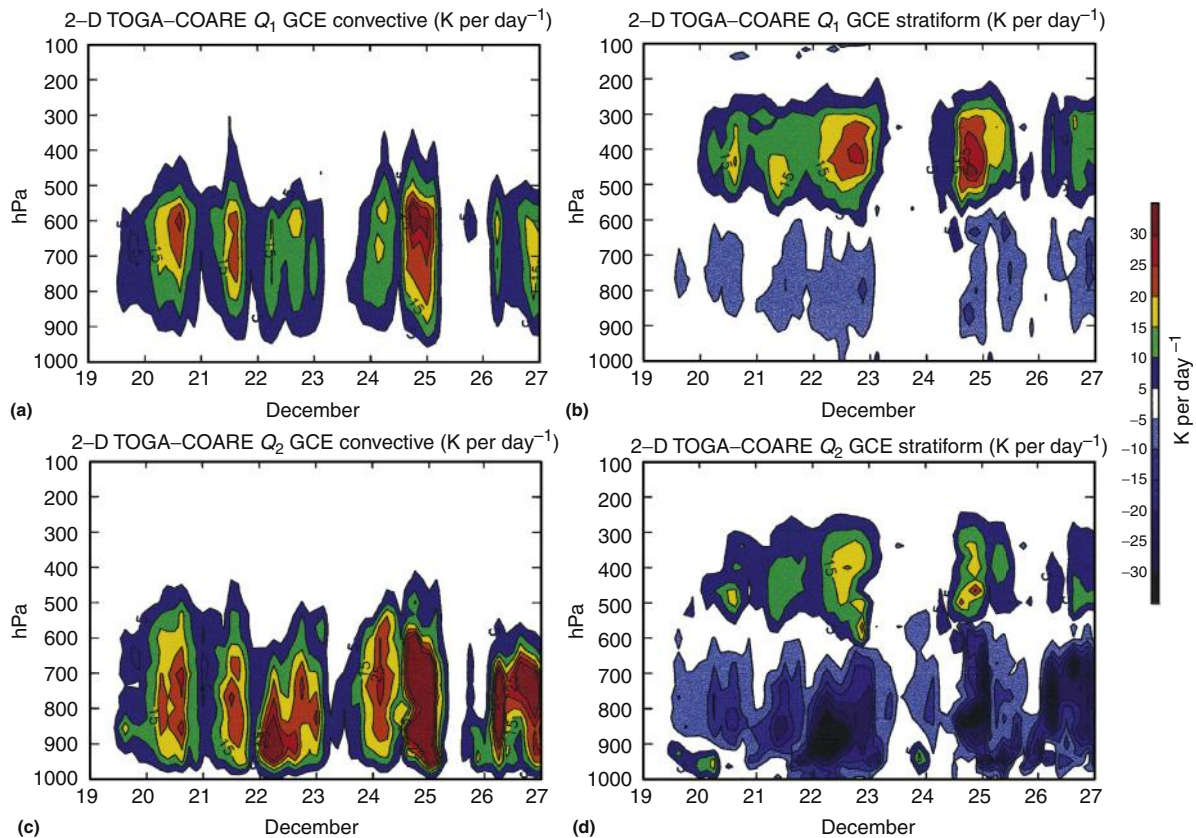


Figure 5 Evolution of the simulated apparent sources averaged over the west Pacific warm pool region for the period 19–27 December 1992. (a) Q_1 for the convective region; (b) Q_1 for the stratiform region; (c) Q_2 for the convective region; and (d) Q_2 for the stratiform region.

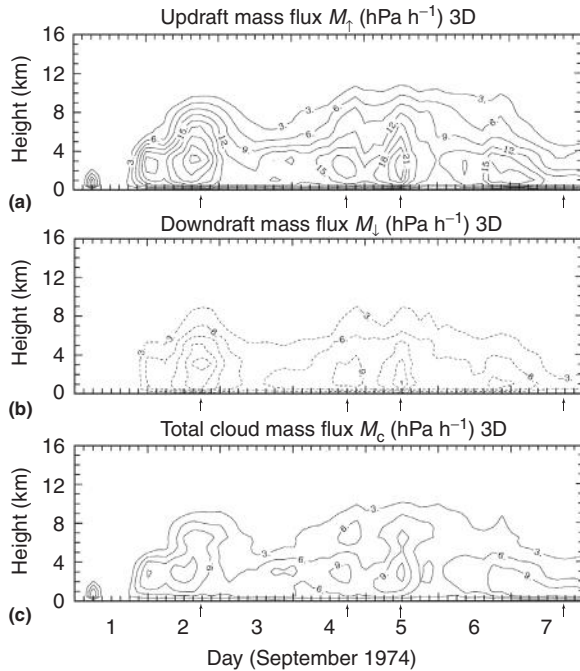


Figure 6 Evolution of domain-averaged mass fluxes (hPa h^{-1}), (a) updraft, (b) downdraft, and (c) total mass fluxes, for simulated east Atlantic cloud systems. Adapted from Wu, X., Moncrieff, M.W., 1996. Recent progress on cloud-resolving modeling of TOGA COARE and GATE cloud systems. Proc. Workshop on New Insights and Approaches to Convective Parameterization, Reading, ECMWF, UK, 128–156.

Convective Momentum Transport

The vertical transport of momentum by convection affects the conversion of kinetic energy from subgrid-scale eddies to the mean flow, the rate of frictional dissipation, and the atmospheric energy spectrum. However, the large-scale effects of momentum transport represented by convective parameterizations are not always applied in large-scale models. The horizontal pressure gradient force within cloud systems affects convective momentum transport, but is difficult to measure accurately from observations mainly because the pressure field is a strong function of convective dynamics. Theoretical models show that convective momentum transport in mesoscales can be either upgradient, which enhances the mean flow, or downgradient in which case it is a mixing process. The existence of these (opposing) effects is supported by observations. Entropy considerations would suggest that, on average, momentum transport must be downgradient, with upgradient transport occurring in special conditions (e.g., in highly organized squall systems). The largest upgradient momentum fluxes occur in organized flow. Cloud models are instrumental for quantifying convective momentum transport and in deriving physically based parameterizations.

Convectively generated gravity waves affect the momentum balance of the atmosphere, and are particularly important in the tropical stratosphere and mesosphere. The wave generation mechanism and its vertical propagation to the deep atmosphere are being studied using cloud models.

Diurnal Variability of Precipitation

Cloud models have quantified the mechanisms responsible for the diurnal cycle of precipitation processes over the tropical oceans associated with the diurnal variation of radiation. The diurnal variation of rainfall can be simulated even when the diurnal variation of SST is suppressed. However, the maximum rainfall is shifted from nighttime (0200 Local Solar Time) to early morning (about 0500 LST) with suppressed diurnal variation of SST. While the diurnal variation of SST modulates rainfall processes, it may play a secondary role in the diurnal variability. Cloud models also indicate that convection is modulated by the diurnal change in available water as a function of temperature and is responsible for the nighttime maximum in rainfall. This implies that the increase (decrease) in surface precipitation associated with long-wave cooling (solar heating) may be due to an increase (decrease) in relative humidity. However, the interaction of radiation with organized convection can affect the diurnal variability of rainfall. Well (less)-organized cloud systems can produce strong (weak) diurnal variations in rainfall, but ice processes enhance the diurnal variation of precipitation.

Prediction of the diurnal cycle of convective precipitation has low skill over land, for reasons not well understood. The development of the diurnal evolution of the convective boundary layer, the role of orography and propagating convective systems, and the effects of land surface processes are all involved. Concerning the diurnal cycle of precipitation over tropical islands, sea breezes and land breezes and their interaction with coastlines and orography are key mechanisms. Cloud models have addressed these aspects in considerable detail.

Water and Energy Cycles

Cloud models are used in studies of the tropical water and energy cycles. In this context, the models typically run for several weeks until the temperature and water vapor fields reach a quasi-equilibrium state. They can produce different quasi-equilibrium states (warm and humid vs cold and dry), even though similar initial thermodynamic profiles and fixed SST are used. Stronger surface winds tend to produce a warmer and more humid thermodynamic equilibrium state. The moist static energy budget further indicates that the large-scale forcing of water vapor is another process responsible for warmer and more humid equilibrium states.

Cloud models have also been used to quantify hypotheses relating to global warming. Key results to date are (1) conversion of ice-phase water into the vapor phase associated with the dissipation of upper-level stratiform/cirrus clouds contributes to upper tropospheric moisture on the same order as moisture transport from deep convection; (2) cloud activity is much more sensitive to convergence in the large-scale atmospheric circulation over an oceanic warm pool than it is to the local SST; and (3) organization of cloud systems modulate the magnitude of upper-level cloudiness and moisture profiles. The above conclusions do not say whether or not global warming is occurring, only that if cloud processes are neglected or poorly

formulated, the consequences could lead to substantial errors in climate hypotheses.

Conclusion

Observations together with cloud models are a promising approach for improving and testing representations of cloud processes in numerical weather prediction models and climate models. Observations provide both the initial conditions and model-validation data. Modern cloud models operate with reasonable, albeit incomplete, microphysical parameterizations and simulate the evolution, structure, and life cycles of cloud systems. They also explicitly calculate interactions between clouds and long-wave and solar radiation that are difficult, if not impossible, to measure observationally.

During the past five decades, cloud models have advanced from simple process models to full multiscale cloud-system simulators that span a dynamic range of 1 km to planetary scale. They are helping improve our understanding of the interaction between convection, radiation, and the large-scale environment. They address basic issues in global and regional prediction, as well as fundamental problems associated with the Earth's water and energy cycles. They are used extensively to improve physical parameterizations. Because of the dynamic range of modern cloud-system models, space-based remote sensing is an ever more necessary part of model validation.

See also: **Air Sea Interactions:** Momentum, Heat, and Vapor Fluxes. **Boundary Layer (Atmospheric) and Air Pollution:** Diurnal Cycle; Modeling and Parameterization. **Clouds and Fog:** Cloud Microphysics. **Mesoscale Meteorology:** Convective Storms: Overview; Density Currents; Mesoscale Convective

Systems; Overview. **Numerical Models:** Mesoscale Atmospheric Modeling; Methods; Parameterization of Physical Processes: Clouds. **Observations Platforms:** Balloons; Radiosondes. **Radar:** Polarimetric Doppler Weather Radar; Precipitation Radar. **Radiation Transfer in the Atmosphere:** Cloud-Radiative Processes. **Satellites and Satellite Remote Sensing:** Precipitation; Remote Sensing: Cloud Properties. **Thermodynamics:** Saturated Adiabatic Processes. **Tropical Cyclones and Hurricanes:** Overview and Theory. **Turbulence and Mixing:** Overview.

Further Reading

- Cotton, W.R., Anthes, R.A., 1989. Storm and cloud dynamics. In: *International Geophysics Series*, vol. 44. Academic Press, San Diego.
- Houze Jr., R.A., 1993. Cloud dynamics. *International Geophysics Series*, vol. 53. Academic Press, San Diego.
- Ludlam, F.H., 1980. *Clouds and Storms: The Behavior and Effect of Water in the Atmosphere*. The Pennsylvania State University Press, University Park.
- Moncrieff, M.W., 2010. The multiscale organization of convection at the intersection of weather and climate: Why does climate vary? In: Sun, D.-Z., Bryan, F. (Eds.), *AGU Geophysical Monograph Series*, vol. 189, pp. 3–26. <http://dx.doi.org/10.1029/2008GM000838>.
- Smith, R.K., 1997. The physics and parameterization of moist atmospheric convection. *NATO Adv. Study Inst. Ser. C Math. Phys. Sci.* 505, Kluwer Academic, pp. 498.
- Tao, W.-K., 2003. Goddard Cumulus Ensemble (GCE) model: application for understanding precipitation processes. *AMS Meteorol. Monogr. Cloud Syst. Hurricanes TRMM*, 107–138.
- Tao, W.-K., Moncrieff, M.W., 2009. Multiscale cloud system modeling. *Reviews in Geophysics* 47, RG4002. <http://dx.doi.org/10.1029/2008RG000276>.
- Wu, X., Moncrieff, M.W., 1996. Recent progress on cloud-resolving modeling of TOGA COARE and GATE cloud systems. *Proc. Workshop on New Insights and Approaches to Convective Parameterization*, Reading ECMWF, UK, 128–156.

Contrails

P Minnis, Science Directorate, NASA Langley Research Center, Hampton, VA, USA

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

Contrails are anthropogenically induced clouds that have the potential to impact climate and reveal aircraft positions. Contrail formation is governed by temperature and humidity fields and is generally limited to temperatures below -40°C . Contrail growth and transformation into cirrus clouds is more complex and must be understood to fully account for contrail effects on climate. Contrail microphysical and optical properties as well as coverage, location, and illumination conditions are also critical for determining contrail radiative forcing, a climate change metric. *In situ* measurements, satellite remote sensing, cloud process models, and climate models are used to study contrails and their potential impacts on global climate.

Introduction

One of the most visible anthropogenic effects on the atmosphere is the condensation trail or contrail. These aircraft-induced clouds have become a common sight since the 1960s because of increasing jet traffic, but were observed as early as 1919. Contrails were frequently seen and filmed during World War II during bombing raids or dogfights. They were briefly studied in Germany during the war but drew little scientific interest again until the early 1950s when the use of jet aircraft by military and commercial aviation accelerated. Interest waned with only sporadic studies until the 1990s when aircraft exhaust and contrails became the foci of numerous research efforts. Concerns over their impact on climate and aircraft visibility have been the primary motivation for the recently intensified research into contrails. Climate researchers are interested in knowing whether they cause warming or cooling of the atmosphere, while military mission planners would prefer that contrails do not reveal the positions or paths of their aircraft. Understanding the contrail effects requires knowledge of their physical and optical characteristics and how, when, and where they form.

Contrail Formation

Contrails are generally composed of ice crystals with trace amounts of exhaust products such as soot and sulfates. The contrail ice crystals form because the relative humidity with respect to liquid water U_w temporarily reaches the saturation point in the plume mixture of ambient air and hot exhaust gases. Tiny droplets develop on background aerosols or on aerosols formed by exhaust compounds. Because the ambient temperatures required for contrails are generally colder than -40°C , the small water droplets instantly freeze and grow via vapor-to-ice deposition as long as the relative humidity with respect to ice U_i remains above 100%. They dissipate via sublimation if U_i is below the saturation point or by precipitation into unsaturated layers below the flight level.

Aerodynamic Contrails

Another type of contrail that forms at higher temperatures is the aerodynamic contrail that forms behind the tips or the

leading edges of aircraft wings. These are commonly seen emanating from subsonic aircraft in humid atmospheres, usually at low levels. In these cases, the ambient air is compressed at the wing tip and then expands quickly during adiabatic expansion within the low-pressure area above the wing tip. The expansion temporarily cools the air sufficiently so that it falls below the dew point resulting in condensation. Aerodynamic contrails are most commonly seen at low altitudes where they form short-lived liquid water contrails. Sometimes, at temperatures slightly greater than -40°C , they can form ice crystal trails that will persist if the air is supersaturated with respect to ice. Modeling studies indicate that they are likely to be important in tropical latitudes where cruising altitudes are more often at temperatures slightly greater than -40°C . Because ice contrails generated from aircraft exhaust are the more common variety, aerodynamic contrails are not considered any further here.

Exhaust Contrails

The basic concepts for determining the conditions for contrail formation were independently developed by E Schmidt in Germany during 1941 and H Appleman in the USA during 1954. The lines in Figure 1 schematically illustrate the ice contrail formation process for several scenarios with the ambient temperatures T_a and water vapor partial pressures e_a indicated by the points at the lower end of each line. Each line

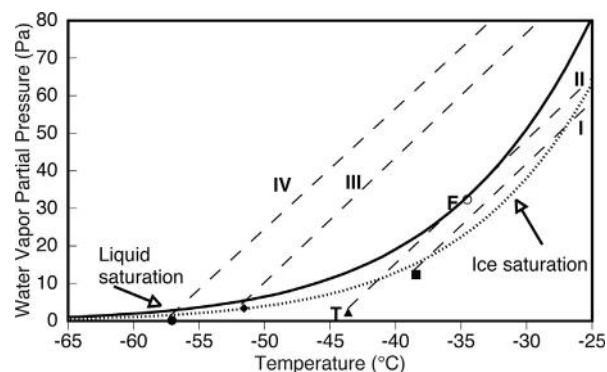


Figure 1 Phase diagram with mixing lines for aircraft exhaust in different ambient conditions.

extends to the temperature T_e and water vapor partial pressure e_e of the exhaust exiting the engine. As the exhaust gases mix with the ambient air, the mean temperature and moisture properties of the mixture follow these mixing lines until they approach the ambient conditions. In cases defined by the lines I, II, and IV, the ambient water vapor pressure is less than the ice saturation partial pressure e_i , while in Case III, $e_a > e_i$. In Case I, the partial pressure exceeds e_i during the mixing but never reaches water saturation and a contrail does not develop. A short-lived contrail would develop in Case II because at point F, the mixture temperature T_F coincides with the liquid water saturation partial pressure e_w . The contrail would form when the plume temperature reached T_F and persists until the plume partial pressure decreased to a value below e_i at approximately -42°C . A long-lived, persistent contrail would form in Case III, because the ambient air is supersaturated with respect to ice. Because saturation conditions cover a greater range of temperatures after initial formation, the contrail formed in Case IV would probably last longer than that in Case II.

Although contrail formation has been observed at temperatures as great as -36°C , it is clear from Figure 1 that contrails form more easily at lower temperatures. The threshold temperature T_T for contrail formation is defined as the warmest ambient temperature that will support contrail formation for a given value of e_a and the exhaust parameters T_e and e_e . The latter quantities determine the mixing line slope G and are a function of engine type, operating conditions, and fuel, while the value of e_a can be determined from vertical profiles of atmospheric and dew point temperatures. In Case II, the ambient temperature at point T is the contrail formation threshold temperature for the given values of e_a and the mixing line slope G . That is, the ambient temperature enabling contrail formation temperature would have to change if either e_a or G varied and, therefore, T_T is unique for each pair of e_a and G . The threshold temperatures are greater than T_a for Cases IV and III, and less than T_a for Case I. To find T_T for a particular slope and e_a , it is necessary to determine the tangent point T_F for a line having slope G with the curve describing the variation of e_w with T . Given a value of G , the threshold temperature can be computed for T_F between -10 and -60°C using

$$T_F = -46.46 + 9.43\ln(G - 0.053) + 0.720[\ln(G - 0.053)]^2 \quad [1]$$

where G is given in Pascals per Kelvin. The threshold temperature for any value of U_w or e_a can be determined iteratively with

$$T_T = T_F - [e_w(T_F) - U_w e_w(T_T)]/G \quad [2]$$

The mixing line slope depends on the specific plume enthalpy h_p and the water vapor mixing ratio q , which, in turn, are related to the emission index EI_w , mass specific combustion heat Q , and the overall engine efficiency η . Specifically,

$$G = \Delta e / \Delta T = [\Delta q / \Delta h_p] p c_p / \varepsilon = \text{EI}_w p c_p / [\varepsilon Q (1 - \eta)] \quad [3]$$

where c_p is the specific heat capacity, p is the pressure, and $\varepsilon = 0.622$. The emission index, the mass of water produced per mass of combusted fuel, accounts for Δq since $e_e \gg e_a$. The enthalpy differential is also determined almost entirely by Q and η because the ambient heat is negligible compared to that produced by the engine. Since Q and EI_w can be determined for a given fuel, then the overall efficiency, the ratio

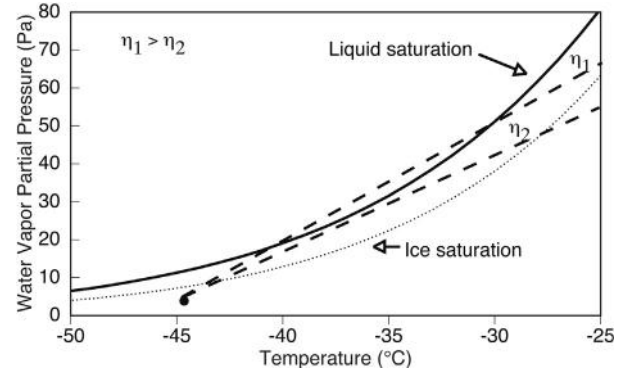


Figure 2 Hypothetical mixing lines for different propulsion efficiencies.

of propulsion energy to total combustion energy, is the primary variable affecting the mixing line slope. The slope of the line increases with increasing efficiency. Each type of engine has a nominal efficiency that is based on stationary operating conditions. The overall efficiency, however, may vary for a given engine because of different airframes, maintenance, and operating conditions. Figure 2 illustrates the impact of efficiency for a given set of ambient conditions. In this instance η_2 is slightly less than η_1 resulting in a contrail from the airplane with η_1 and no contrail from the one with η_2 . Thus, two airplanes flying in the same environment can produce two different results. Similarly, an airplane might produce a contrail when it is cruising but not when it is ascending, depending on the effect of acceleration on the efficiency.

Contrails typically form at a distance of about 30 m or less behind the aircraft engines where the turbulent mixing has sufficiently reduced the temperature. The latest research results indicate that the initial condensation of the supercooled droplets takes place on a wide variety of particles including exhaust products, e.g., sulfate aerosols, soot, and metallic particles as well as ambient mineral aerosols. When the contrails are about 1 min old, the mean particle radius is around $2\ \mu\text{m}$ and increases up to $4\text{--}5\ \mu\text{m}$ after 3 min. A wide variety of particle shapes have been observed in young contrails including hexagonal columns and plates, triangles, irregular forms, and spheroids. Young contrails often appear saw-toothed or appear to have a cellular structure that results from the wake vortices formed by the aircraft. The wake vortex pairs act to extend the contrail vertically with irregularities that can lead to formation of local convective cells or radiative cooling gradients that aid mixing of the contrail with the ambient air.

Contrail Growth and Structure

Once formed, a contrail develops or dissipates in the same manner as a naturally generated cirrus cloud. Contrail growth and spreading depend on the thickness of the supersaturated layer, the degree of ice supersaturation, the nature of the wake vortex, and the wind velocity and shear. When contrails persist, the particles typically grow to lengths from 30 to $1000\ \mu\text{m}$, sizes usually associated with natural cirrus clouds. Ice particle growth is rapid in highly supersaturated layers and results in fall streaks that spread horizontally in lower layers according to

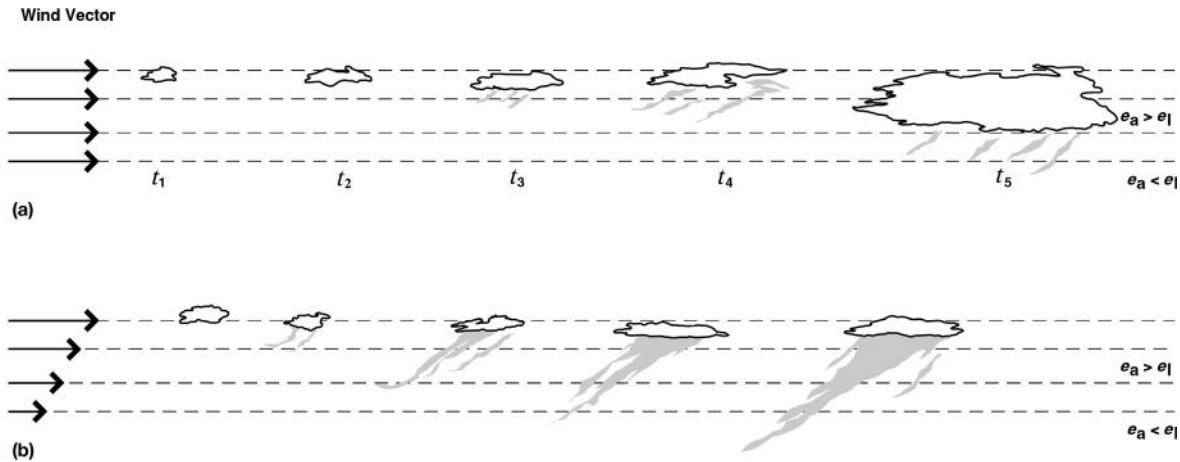


Figure 3 Schematic depiction of contrail spreading in conditions with and without wind shear.

the wind shear. **Figure 3(a)** shows a cross-sectional view of a hypothetical persistent contrail growing and spreading in the absence of vertical wind shear. When wind shear is present (**Figure 3(b)**), the contrail may spread mostly by turbulent mixing induced by the aircraft vortex or by radiative processes. It may also thicken by turbulence or through precipitation. If the crystals fall into supersaturated air below, they will continue to grow or, possibly, split into additional crystals. The linear shape of the contrail will be distorted and the contrail will soon look like a natural cirrus cloud to the observer.

An example of this shear effect is seen in **Figure 4**, which shows the cross-sectional growth of a contrail observed with a surface-based scanning lidar. After 1.1 min (**Figure 4(a)**), when the vortex effect is dissipating, the overall width of the contrail is only ~ 160 m compared to its vertical depth of roughly 260 m. It grows progressively wider during the following 14 min reaching a horizontal extent of 1 km (**Figure 4(d)**). As it widens, the lower part of the contrail starts evaporating (**Figure 4(b)**) in the dry layer below the 10.35-km level, while the upper portion slowly rises due to the buoyancy resulting from the release of latent heat into the air from the condensation-freezing process producing the ice particles. The wind mean vertical wind shear for this case was $\sim 4 \text{ m s}^{-1} \text{ km}^{-1}$.

Contrails that are older than several hours are often indistinguishable from natural cirrus clouds regardless of shear conditions. Most studies indicate that the number of crystals in a contrail remains essentially constant after formation in supersaturated conditions. Thus, if the contrail precipitates, the contrail cloud at flight level might gradually fade as its particles are depleted. If e_a is just above ice supersaturation, then the crystal growth will be limited and little precipitation occurs. In this case, the contrail may spread slowly by diffusion, maintaining its linear shape for a relatively long time. Because the crystals grow by deposition, the amount of ice water in the contrail increases until the particles fall out or equilibrium is reached between the ice water content and e_i . Such equilibrium conditions generally do not last very long and the contrail eventually dissipates. Although most persistent contrails observed from satellites have visible optical depths between 0.10 and 0.4, the values are highly variable ranging between 0.01 and 2. Optically thin contrails may occur frequently but

most cannot be detected in satellite imagery. The lifetimes of contrails are also extremely variable. Short-lived contrails may only last a few seconds while some contrail-generated cirrus

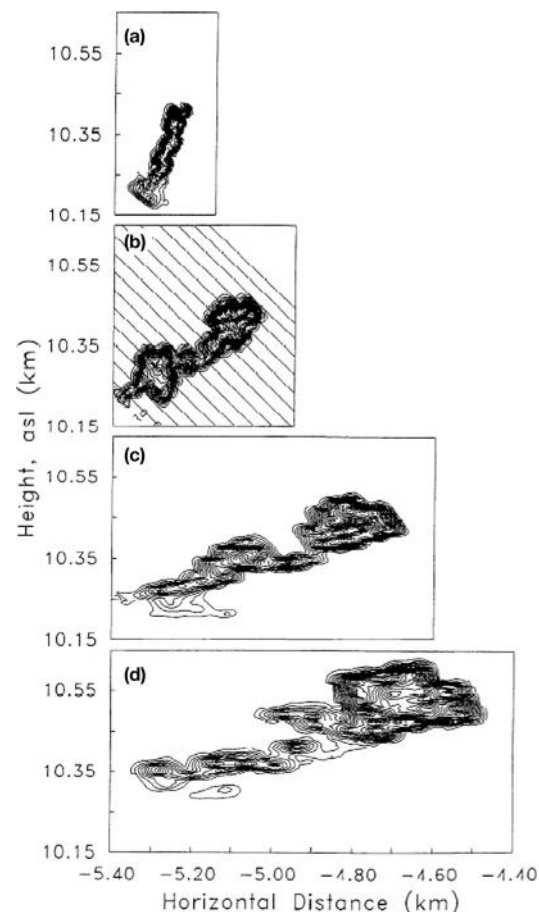


Figure 4 Contour plots of the lidar backscatter signal of cross sections of a contrail. Age of contrail: (a) 1.1 min; (b) 6.2 min; (c) 10.9 min; and (d) 15.5 min. Reproduced from Freudenthaler, V., Homburg, F., Jäger, H., 1995. Contrail observations by ground-based scanning lidar: Cross-sectional growth. *Geophysical Research Letters* 22: 3501–3504.

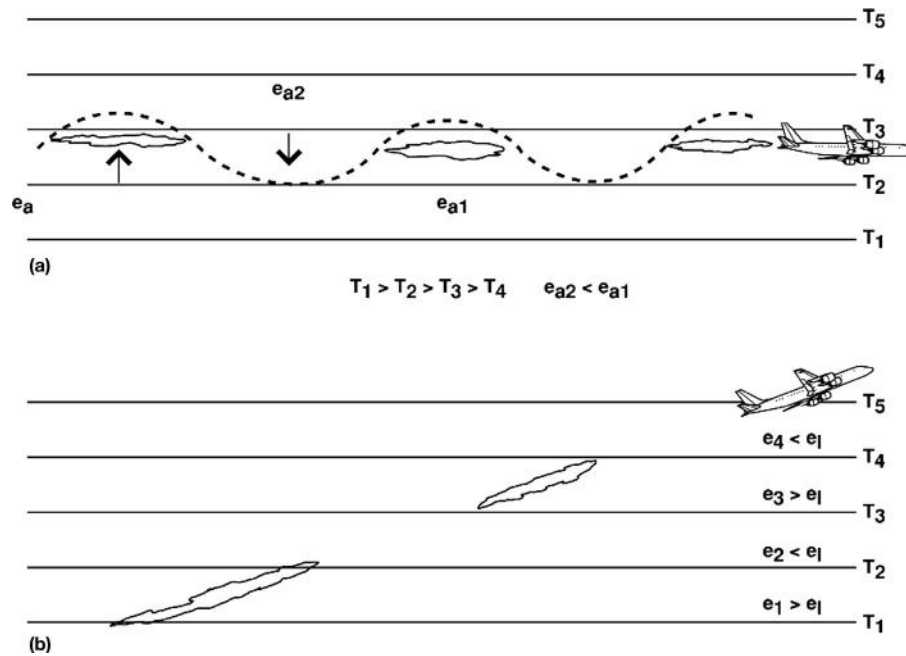


Figure 5 Schematic depiction of contrails forming in an on-off pattern: (a) due to wave action; (b) due to ascent through dry and moist layers.

clouds have been tracked for more than 17 h. The shape, size, optical properties, and life cycle of contrails are highly dependent on their environment, so that a multitude of contrail morphologies can occur. Contrail-cirrus clouds are generally similar to natural cirrus clouds within a few hours after their formation.

Because water vapor and temperature are not homogeneously distributed, even at relatively small scales (~ 100 m), contrails may form or persist in an apparently erratic fashion as shown in Figure 5. For example, an on-off pattern can occur as an aircraft flies through a moist layer disturbed by a vertical wave or even weak convective plumes. The contrails in Figure 5(a) form in the ascending parts of the wave or plume where the temperature of the rising air drops below the threshold temperature, while in the descending portions the air warms and dries resulting in no contrail formation. Similar patterns can result from an airplane ascending or descending through several thin layers that are near saturation but separated

by dry layers as in Figure 5(b). The persistence of a contrail or parts of it depends on the value of e_a relative to e_i along the contrail line. Thus, parts of a contrail may rapidly dissipate while other portions may linger and even grow. The local turbulence induced by the airframe, the atmospheric stability, and the wind vector also affect the morphology of the contrail.

Photographs of the most familiar type, short-lived contrails are shown in Figure 6. In both cases, the trails gradually fade without much spreading. In those situations, e_a is only slightly less than e_i . When e_a exceeds e_i , less familiar shapes can occur. In the top panel of Figure 6, the contrail dissipates several kilometers behind the aircraft. The lower panel shows several intermittent contrails (e.g., Figure 5) that persisted for less than an hour with little spreading. Figure 7 shows examples of contrails at different stages of growth or persistence at the same time in different parts of the sky as seen by an observer at Ft. Monroe, Virginia. East of the observer (Figure 7(a)), the contrails appear to maintain their linearity in the foreground



Figure 6 Short-lived contrails.

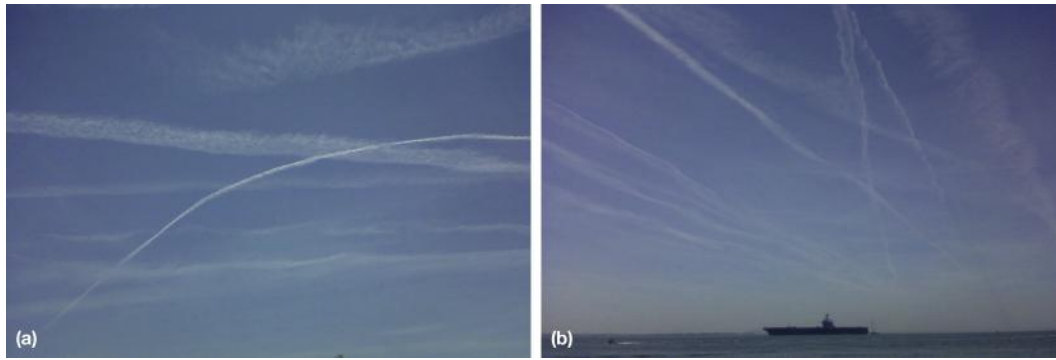


Figure 7 Persistent contrails in various stages of growth and decay as seen: (a) east; (b) south of Ft Monroe, VA at 16.25 UTC, 21 May 2010.

except for the curved contrail crossing many of the other contrails. Toward the horizon, it appears that the contrails have spread out and lost some of their linearity, looking more like natural cirrus. One contrail appears to be segmenting. To the south (**Figure 7(b)**), a succession of slowly spreading contrails extends off to the hazy horizon. Different parts of these contrails spread differently with the sections near the center of the photograph shifted and twisted slightly compared to the rest of the contrail segments. Many of these contrails persisted for at least several hours before advecting southeastward out of view.

Condensation trails often form ahead of advancing fronts in the poleward flow of an upper-level trough where conditions are not quite saturated enough for natural cirrus development. In these instances they can occur at multiple levels in the atmosphere because the formation conditions often cover a large depth of the atmosphere and air traffic uses a wide range of altitudes, but they also form in a variety of other conditions. In **Figure 7**, the contrails formed from moisture advecting westward from the high altitude outflow of an occluding low-

pressure system approximately 1400 km west of the observer. The atmosphere was saturated with respect to ice at temperatures below -40°C in many layers between 9.9 km (32.7 kft) and 13.1 km (43.2 kft). Thus, the formation of crossing contrails, which only occur because airplanes fly at different flight levels depending on flight path direction, is likely for this case because of the depth of the saturated layers.

The contrails seen in **Figure 7** are part of a larger area of contrails that is evident in infrared images (**Figure 8(a)**) from the MODerate-resolution Imaging Spectroradiometer (MODIS) on the Terra satellite. These images were taken around the same time as the photograph in **Figure 7**. Subsequent imagery shows that these contrails dissipated downstream to the east, while additional contrails formed within or beneath the advancing thin cirrus clouds.

Contrails can form within cirrus clouds where they are manifested by reduced particle sizes or local thickening of the cloud. Aircraft exhaust can also affect supercooled liquid water clouds that have temperatures below the threshold value. When an airplane flies through this type of cloud, the dynamic

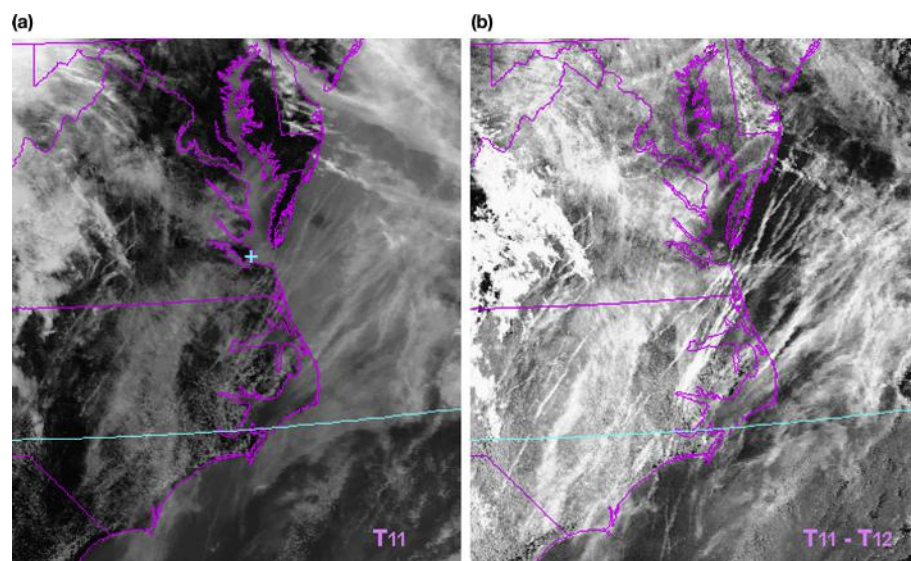


Figure 8 Infrared temperature and temperature difference images of contrails over Maryland, Virginia, and North Carolina, USA from the Terra 1-km resolution MODIS at 16.15 UTC, 21 May 2010. Cross indicates position of observer in **Figure 7**.

pressure changes due to flow over the wing edges or behind propeller tips cause cooling sufficient to freeze the surrounding cloud droplets. The thermodynamic equilibrium shifts from a vapor-to-liquid to vapor-to-ice process causing a rapid depletion of the available water vapor onto the frozen droplets. Newly formed ice crystals quickly grow large enough to fall out of the cloud resulting in a streak below the cloud and a gap within in the cloud.

This gap, called a distrail, hole punch cloud, or canal cloud, is linear when the airplane flies for an extended distance within the cloud or oval shaped when the aircraft is briefly inside the cloud as it ascends or descends. Depending on the conditions, especially the original cloud thickness, a distrail will either persist or be filled in with a new water droplet cloud. In very moist conditions, significant snowfall can occur along the path of the aircraft. Distrails are most frequently observed in altostratus or altocumulus clouds.

Contrail Remote Sensing

Although contrails are most often identified by their linear shapes both from ground observations (Figure 7) and satellite imagery (Figure 8), these man-made clouds can take on other geometric shapes according to the particular flight patterns and winds. For instance, spiral shapes result from an airplane in a circular holding pattern within an advecting supersaturated layer, while a figure eight can form in a similar type of layer if the airplane flies in a linear holding pattern. The linear structure is most common and forms the basis for automated identification of contrails. Because detection of contrails is important for various scientific applications, methods have been developed for differentiating contrails from other linear clouds in satellite imagery, the only plausible data source for studying the spatial and temporal distribution of contrails and their physical characteristics. Automated techniques for contrail detection typically create an image of a parameter most likely to be associated with a contrail, then apply a variety of image processing methods to find linear structures within that image.

Such methods, which are still being researched, usually take advantage of the relatively distinct infrared optical properties of younger contrails to compute a parameter that is likely to yield a distinctive contrail signal for image processing. Because of their relatively small size (mean effective radii between 5 and 25 μm), the ice crystals in contrails have single-scattering albedos that increase more with decreasing wavelength λ in the thermal infrared region (3.5–15 μm) than the single scattering albedos of larger particles typically found in most cirrus clouds (effective radii greater than 15 μm). Thus, contrails usually transmit more radiation at shorter infrared wavelengths than a cirrus cloud of equivalent optical depth, resulting in a signal that often reveals a contrail.

To better understand this effect, consider that the satellite measures a spectral radiance L_λ that is recorded as a brightness temperature T_λ , which is proportional to λ through the Planck function. For a cloud or contrail, the observed radiance can be modeled simply as:

$$L_\lambda = \varepsilon_\lambda L_c + (1 - \varepsilon_\lambda) L_b \quad [4]$$

where L_c is the radiance emitted at the cloud temperature T_c , L_b is the upwelling radiance at the cloud base with an equivalent brightness temperature T_b , and ε_λ is the cloud emissivity. In general, $L_b > L_c$, so that an increase in cloud transmissivity ($1 - \varepsilon_\lambda$) results in more transmission of L_b and a larger value of L_λ . Thus, T_λ will be greater at shorter wavelengths than at longer ones as long as the cloud is optically thin (ε less than 0.9 or so). This effect can be seen in Figure 8, which shows the 11- μm temperature (T11) image and an image of brightness temperature difference (T11 – T12) between the 11- and 12- μm channels of the Terra MODIS. Some of the contrails are readily apparent in the 11- μm image (Figure 8(a)) but are obscured by other cirrus clouds. The temperature difference image in Figure 8(b) reveals many contrails that were not evident in the standard infrared image and highlights others more clearly. Even the curved contrail in Figure 7(a) can be seen crossing over some of the parallel contrails east of the observer's location in Figure 8(b). Because the actual temperature difference contrast depends on the effective particle sizes and optical depths of the surrounding clouds, and those quantities are naturally variable, the contrails are not always detected. Furthermore, other features such as cirrus streaks, coastlines, or cloud edges may produce similar signals.

When T_b is not very different from T_c , such techniques do not reveal the contrails very readily because the signal is so small. Therefore, contrails imbedded in relatively thick cirrus clouds cannot be seen in most temperature difference imagery. However, during the daytime, contrails can often be detected using temperature differences between a channel near 11 μm and one in the shortwave (SW) infrared wavelength range (3.5–4.5 μm). At the shorter wavelengths, the satellite imagers measure an emission component and a solar reflected component. The smaller contrail ice crystals reflect more sunlight than the surrounding cirrus crystals resulting in a relatively large brightness temperature. Figure 9 shows MODIS imagery for the same case as in Figure 8, except that it is shifted westward covering parts of Ohio, Pennsylvania, Virginia, and West Virginia. The 11- μm image in Figure 9(a) shows little sign of contrails, particularly in the thick cirrus over Ohio, Pennsylvania, and West Virginia, while the T11 – T12 image (Figure 9(b)) reveals a few contrails over those same areas. The 11- and 3.7- μm temperature difference (T3.7 – T11) image (Figure 9(c)), however, reveals a large number of linear contrails around the common borders of those states. These contrails are absent or faint in Figure 9(b). Conversely, where thick cirrus is absent (e.g., South Central Virginia), the contrails are more prominent in Figure 9(b) than in Figure 9(c). Thus, the two images tend to complement each other with respect to contrail detection. Further enhancement of the photograph would likely reveal more contrails in the temperature difference images. Again, the ability to detect contrails in a thick cirrus cloud depends on many factors including the contrail age and its relative depth in the cloud as well as the particle sizes in the cirrus cloud, and the viewing and illumination conditions. To date, the T3.7 – T11 information has not been employed to monitor contrails for climate purposes since it is applicable only during the daytime, and contrails imbedded in thick cirrus clouds will likely have minimal radiative impact.

Contrails can also be detected in high-resolution visible and near-infrared imagery in certain conditions. For example, when

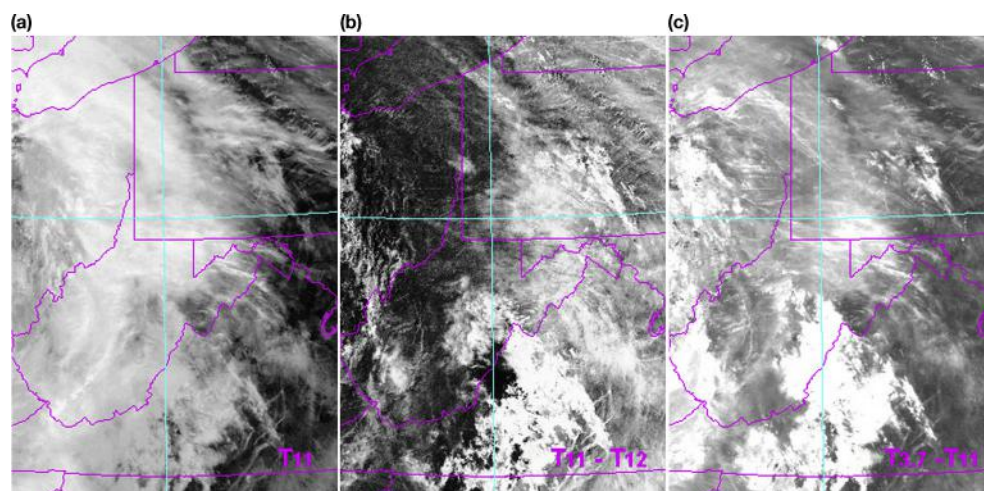


Figure 9 Contrails imbedded in thick cirrus over Ohio, Pennsylvania, Virginia, and West Virginia from the Terra 1-km resolution MODIS at 16.15 UTC, 21 May 2010. (a) 11- μm brightness temperature; (b) 11–12 μm brightness temperature difference; (c) 3.7–11 μm brightness temperature difference images.

not imbedded in cirrus or over lower clouds, a young contrail is often reflective enough to be seen as a bright line in a 1-km visible channel image. Sometimes, the contrail will cast a shadow on lower clouds and can be detected from its shadow. Near-infrared channels on or near water vapor absorption lines, like that near 1.38 μm , often do not receive any significant reflectance from lower clouds so that even very thin high clouds, such as contrails, can be detected.

Figure 10 shows an example of a contrail outbreak over Florida at 16.00 coordinated universal time (UTC), 23 March 2011 as seen in Terra MODIS. Very faint contrails can be seen west of the peninsula in the true color image (Figure 10(a)) with little indication of contrails near the Georgia–Florida border. The 1.38- μm reflectance image (Figure 10(b)) reveals a mass of crossing contrails along the border and fine detail of the contrails and contrail–cirrus over the whole area. The spread of the contrails and cirrus clouds is greater in Figure 10(b) than in the true color image (Figure 10(a)) or the T11 – T12 image (Figure 10(c)). When there are no underlying clouds, the 1.38- μm channel can reveal contrails with very small optical depths, which modeling studies suggest are quite common but are not seen in infrared imagery. Although potentially promising for more accurate monitoring of contrails, the 1.38 μm channel and other near-infrared channels are only applicable during the daytime and are strongly affected by underlying clouds. Thus, thermal infrared channels are mainly used for monitoring contrails.

Contrail properties such as temperature, height, optical depth, and effective particle size can be determined with the same methods used to remotely sense cloud properties from satellite imager data. Such techniques typically require multispectral imagers that can be used to simultaneously solve for ϵ , T_c , and the particle size. When an insufficient multispectral radiance set is available, one or more of the parameters must be assumed to obtain a solution for the other parameters. These methods generally provide results consistent with *in situ* measurements. Satellite analyses have found mean optical depths that average approximately 0.11 over Europe and 0.25 for the USA. Modeling studies show similar discrepancies

between contrails over the USA and Europe, but some have also indicated that the satellite retrievals miss large numbers of contrails having optical depths smaller than 0.05. The average temperature of contrails detected from satellite data over the midlatitudes is -54°C , which corresponds to an average pressure of about 200 hPa.

Contrail–Cirrus

An increase in cirrus cloudiness because of contrail formation has been hypothesized since the beginning of the commercial jet age. The possibility for enhanced cirrus coverage resides in the frequency and extent of areas that are ice supersaturated. *In situ* measurements and numerical model analyses have shown that e_a exceeds e_i 10–20% of the time in air at flight altitudes (8–12 km). Thus, the potential exists for substantial increases in cirrus coverage over areas crossed by air traffic. Because the stratosphere is generally very dry, aircraft flying above the tropopause generate few contrails, especially persistent ones.

The conditions necessary for supporting contrail formation at flight altitudes change with the seasons. Over midlatitude areas, contrail conditions are favorable most often during winter and early spring when the troposphere is coldest. During summer, the temperatures at flight levels are often too high to enable contrail initiation. Over areas poleward of about 50° latitude, the tropopause is often below flight level during winter, so that a significant number of airplanes fly in the stratosphere resulting in contrail suppression. Conditions are more favorable for contrails during the summer and autumn in the subarctic regions. In the tropics, the altitude for contrails is generally above 11 km year-round, so the potential for contrail formation is reduced for many commercial airplanes. However, persistent contrails are likely to occur more frequently in the tropics than at other latitudes at altitudes above 11 km because of more abundant water vapor.

Surface observations over the USA during the 1990s indicate that persistent contrails occur, on average, approximately 9% of the time, but the frequency varies from less than 5% in

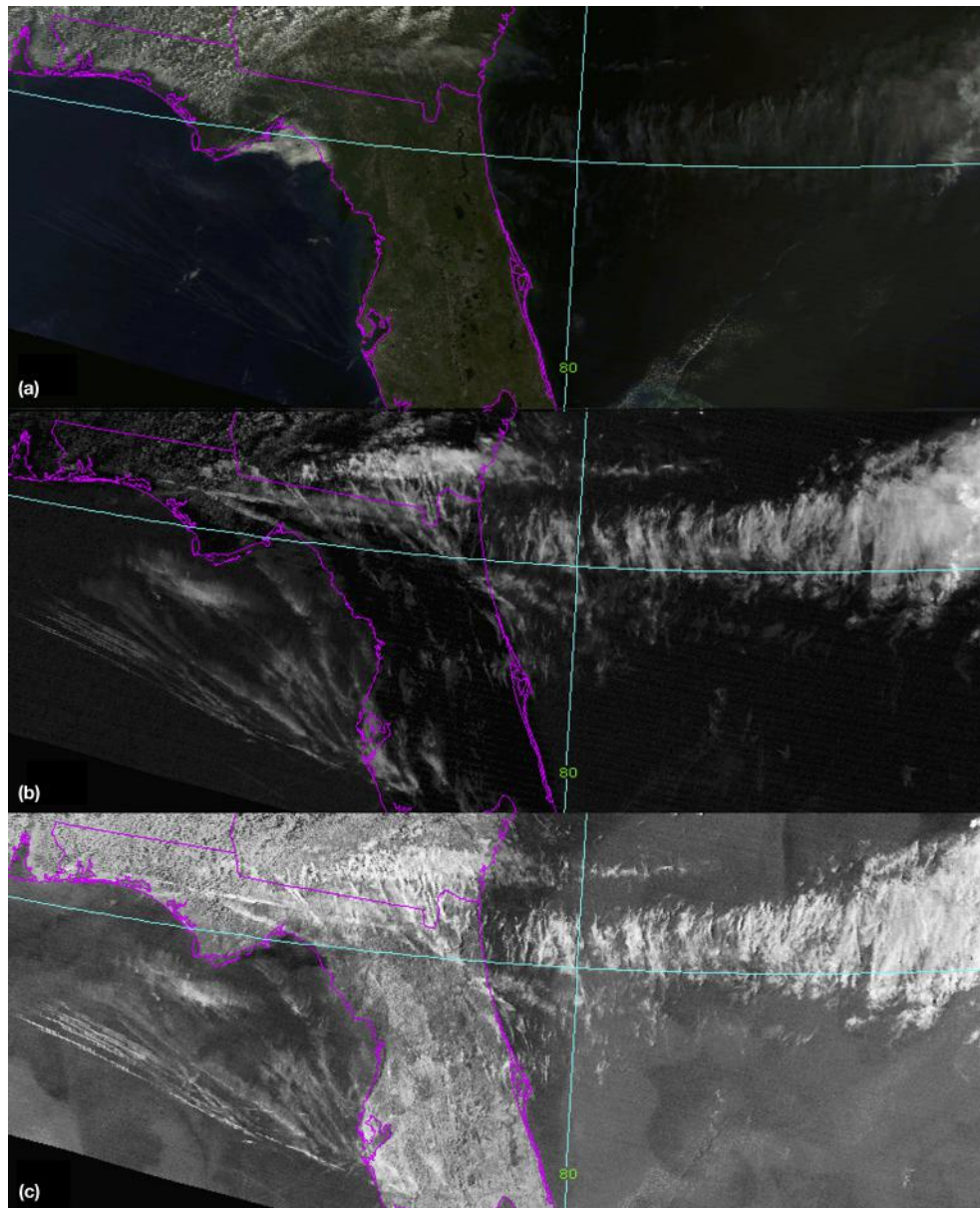


Figure 10 Terra MODIS 1-km imagery taken 16.00 UTC, 23 March 2011. (a) True color; (b) 1.38- μm reflectance; (c) T11 – T12.

low traffic areas to 25% in the main air corridors. Approximately 80% of these persistent contrails are either imbedded in, extending from, or near natural cirrus clouds. Contrail coverage has been derived from satellite imagery only for those contrails that are linear and large enough to detect in 1-km infrared data. Satellite-based estimates of daily mean contrail amounts over central and western Europe, the North Atlantic, the conterminous USA, Japan, and Thailand are approximately 0.5, 1.0, 0.5, 0.3, and 0.1%, respectively, and, for the early 1990s, roughly 0.1% globally. Similar as well as much smaller or larger values have been derived from theoretical calculations using realistic air traffic patterns, numerical analyses of meteorological fields, and specified engine efficiencies. However, uncertainties in detecting contrails unambiguously with automated satellite image analysis techniques are still quite large.

Automated detection and assessment of contrail coverage have been confined to contrails that are identifiable by their linear structure and small particle sizes. Because these identifying features are often lost as the contrails spread, the linear contrail coverage estimates represent the minimum amount of the sky that is covered by contrails. The T11 – T12 images (Figure 11) from the AVHRR on the NOAA-16 and NOAA-19 satellites illustrate the difficulty of distinguishing contrails from cirrus clouds. The large area of cirrus clouds seen east of North Florida in Figure 10(b) began as hundreds of linear contrails north of the Florida peninsula at least 3 h earlier (Figure 11(a)). Those west of Florida were already quite wide at 12.56 UTC and extended as far west as Louisiana. The latter group dissipated by 17.35 UTC (not shown), while the former mass of contrails persisted as a collection of nearly contiguous

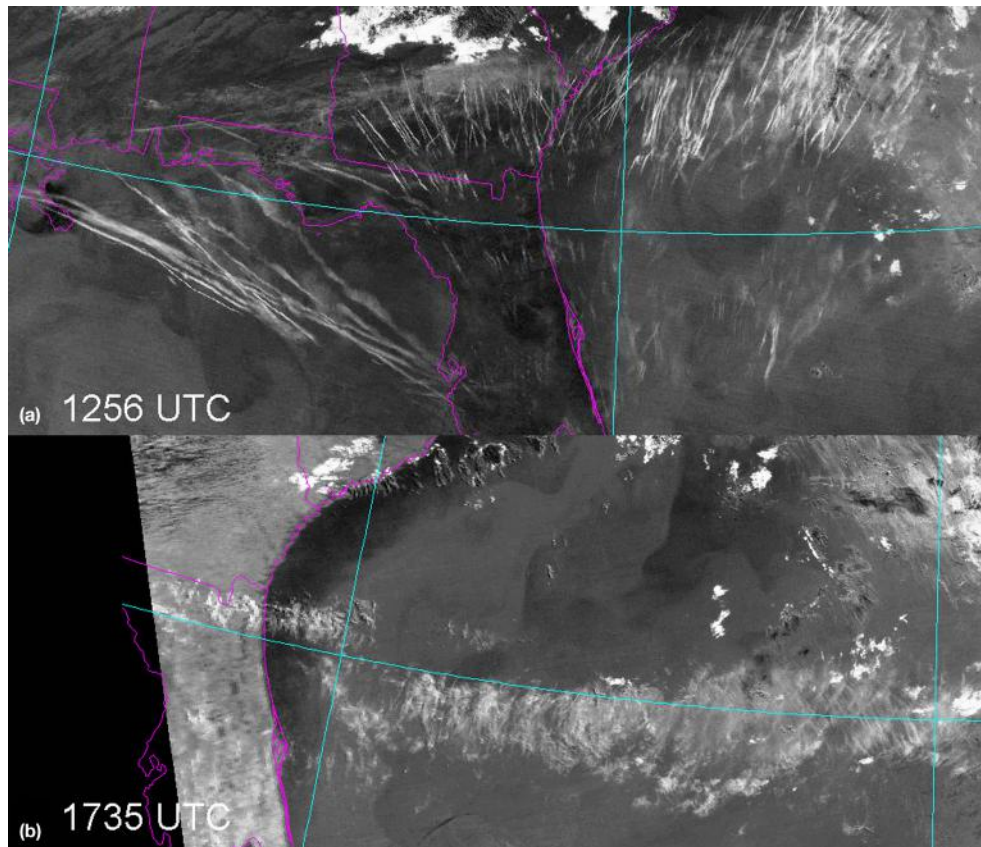


Figure 11 AVHRR T11 – T12 images for 23 March 2011 from (a) NOAA-16; (b) NOAA-19.

cirrus clouds over the Atlantic with a few scattered linear features (**Figure 11(b)**). Without following the linear contrails seen in **Figure 11(a)**, it would be impossible to differentiate the contrail–cirrus over the Atlantic in **Figures 10(b) and 11(b)** as separate from naturally forming cirrus clouds.

Geostationary satellite data can be used to track some contrails as they grow and change in shape and composition. Examination of the geostationary data (not shown) taken during 23 March 2011 revealed that the large contrail mass over the Atlantic actually started forming as early as 11.45 UTC over North Georgia and South Carolina and did not completely dissipate until 23.45 UTC. The long contrails west of Florida started forming over Louisiana around 08.00 UTC and lasted roughly 12 h. Published studies based on geostationary data indicate that the actual cirrus coverage generated by persistent contrails might be as large as a factor of 10 times the coverage estimated for younger, linear contrails. Other studies based on trends in cirrus cloud cover suggest that the coverage including the linear contrails is most likely to be between 1 and 4. Determination of contrail coverage and evaluation of the resulting changes in cirrus cloud amounts are topics of ongoing research.

Another approach to estimating the total change in cloud cover due to contrails is through modeling, which has been undertaken at various levels of complexity and detail. Typically, a model will use flight track distances within an atmospheric specified volume or other measures of air traffic and simulate

the production of contrails either by explicitly forming contrails and spreading them using a parameterization that differentiates them from natural cirrus or by adding cirrus to the atmosphere in proportion to the flight densities. Observations are often used to guide and/or validate these approaches.

Figure 12 shows an example of contrail and contrail–cirrus coverage estimated using a fairly sophisticated model. In this case, the coverage is broken down according to the derived contrail optical depth and age. Contrail–cirrus coverage is greatest over the Northern Hemisphere, especially over the North Atlantic, Europe, and the USA (**Figure 12(a)**), regions with considerable air traffic. Europe is downwind of the Atlantic corridor and receives contrail–cirrus advected from that area. The heavier traffic over the USA and Europe leads to significant coverage by young contrails (<5-h old), as seen in **Figure 12(b)**. The model indicates that most of the contrail–cirrus coverage is due to very thin cirrus having optical depths < 0.02 (**Figure 12(c)**). These very thin contrail–cirrus clouds would be difficult to detect visually from the surface or with passive imagery. Younger contrails comprise a larger proportion of the contrail–cirrus coverage in the Southern Hemisphere than over the north where some areas (e.g., Russia, Northern Canada) have negligible coverage due to young contrails (**Figure 12(d)**). Not shown are the areas where significant decreases in natural cirrus occurred in the model as a result of moisture being locked up in contrail cirrus. For example, the largest drops in natural cirrus coverage occurred

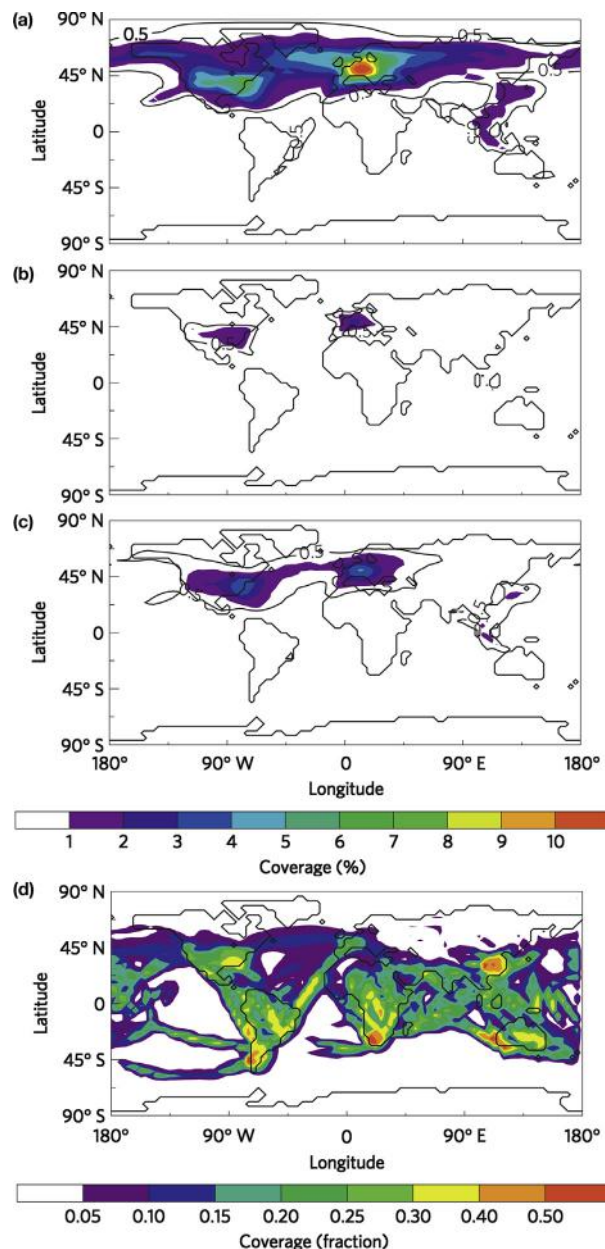


Figure 12 Average 2002 contrail-cirrus and young contrail coverage as simulated by ECHAM4-CCMod. Coverage due to (a) contrail cirrus; (b) persistent young contrails up to 5-h old; (c) visible contrail-cirrus with optical depth > 0.02 ; (d) fraction of contrail-cirrus due to contrails up to 5-h old. Coverages are computed assuming maximum overlap among contrails or contrail-cirrus coverage leads to increased overall cloud coverage. Reproduced from Burkhardt, U., Kärcher, B., 2011. Global radiative forcing from contrail cirrus. *Nature Climate Change* 1: 54–58.

over the Eastern USA, Europe, and the northwest coast of Africa. Although the results discussed here have not been validated to date, they reveal the complexity of the problems faced in quantifying contrail impacts on cloudiness. Other models have produced different results and more research is needed to determine how accurately each model represents the actual atmospheric conditions.

Contrail Climate Effects

Contrails, like other cirrus clouds, can affect both the hydrological and radiation budgets. Many of the possible contrail effects have only been the subjects of educated speculation, although some have been estimated to some degree. Some of these potential effects are mentioned here.

As noted earlier, by freezing out water vapor prior to the natural formation of cirrus clouds, contrails can alter the overall distribution of cirrus. Contrail formation may decrease precipitation in some clouds by reducing the average particle size in the affected clouds. Conversely, the precipitation induced by persistent contrails in otherwise clear air (e.g., [Figure 3\(a\)](#)) may result in moistening of the middle layers of the troposphere and drying of the atmosphere at flight altitudes.

As a thin cirrus cloud, contrails reflect some solar or SW radiation that would otherwise warm the surface and absorb outgoing infrared radiation that cools the surface-atmosphere system. The overall radiative impact depends on the contrast between the contrail and its background, the lifetime of the contrail, and the solar zenith angle when it is present. Depending on the solar zenith angle and the thermal and albedo contrasts between the contrail and background, either surface or lower clouds, the net forcing can result in cooling or warming of the system. For instance, if the contrail forms over a dark background during midday, the amount of reflected sunlight may exceed the amount of infrared radiation blocked and reradiated by the cloud. Conversely, if it develops over a bright hot surface (i.e., desert) during the day, a contrail may reflect little additional solar radiation, but trap a significant amount of infrared because it is so much colder than the surface. Its overall impact would be substantially different from that over the dark surface. A similar effect would occur for a contrail over a warm low cloud deck. At night, contrails warm the atmosphere. However, even during the day when solar and infrared forcings can almost cancel each other, the contrail will still impact the radiation field because most of the blocked sunlight results in cooling of the surface, while much of the infrared or longwave (LW) radiation ‘trapped’ by the contrail warms the upper troposphere and has little immediate impact on the surface. Currently, 60% of air traffic is estimated to occur during the daytime.

These radiative forcing (RF) effects have been estimated with several different models and assumptions resulting in a minor amount of global warming when averaged over a long time period or some slight cooling on an instantaneous basis. [Figure 13](#), which presents results from a general circulation model study, shows the distributions of linear contrail RFs assuming random contrail cloud overlap, an average contrail particle effective diameter of $30 \mu\text{m}$, and an optical depth of 0.2. This estimate is based on air traffic for 2002. The mean RFs for contrails in otherwise clear skies are shown on the left half of the figure, while on the right are those for all skies, including both cloud-free backgrounds and clouds computed within the model. Positive RF indicates warming and conversely, negative forcing denotes a cooling effect. The greatest LW RF for clear skies occurs over the Eastern USA ([Figure 13\(a\)](#)) where it exceeds 100 mW m^{-2} . [Figure 13\(b\)](#) shows that the LW RF is reduced by roughly a factor of 2 when clouds are included in

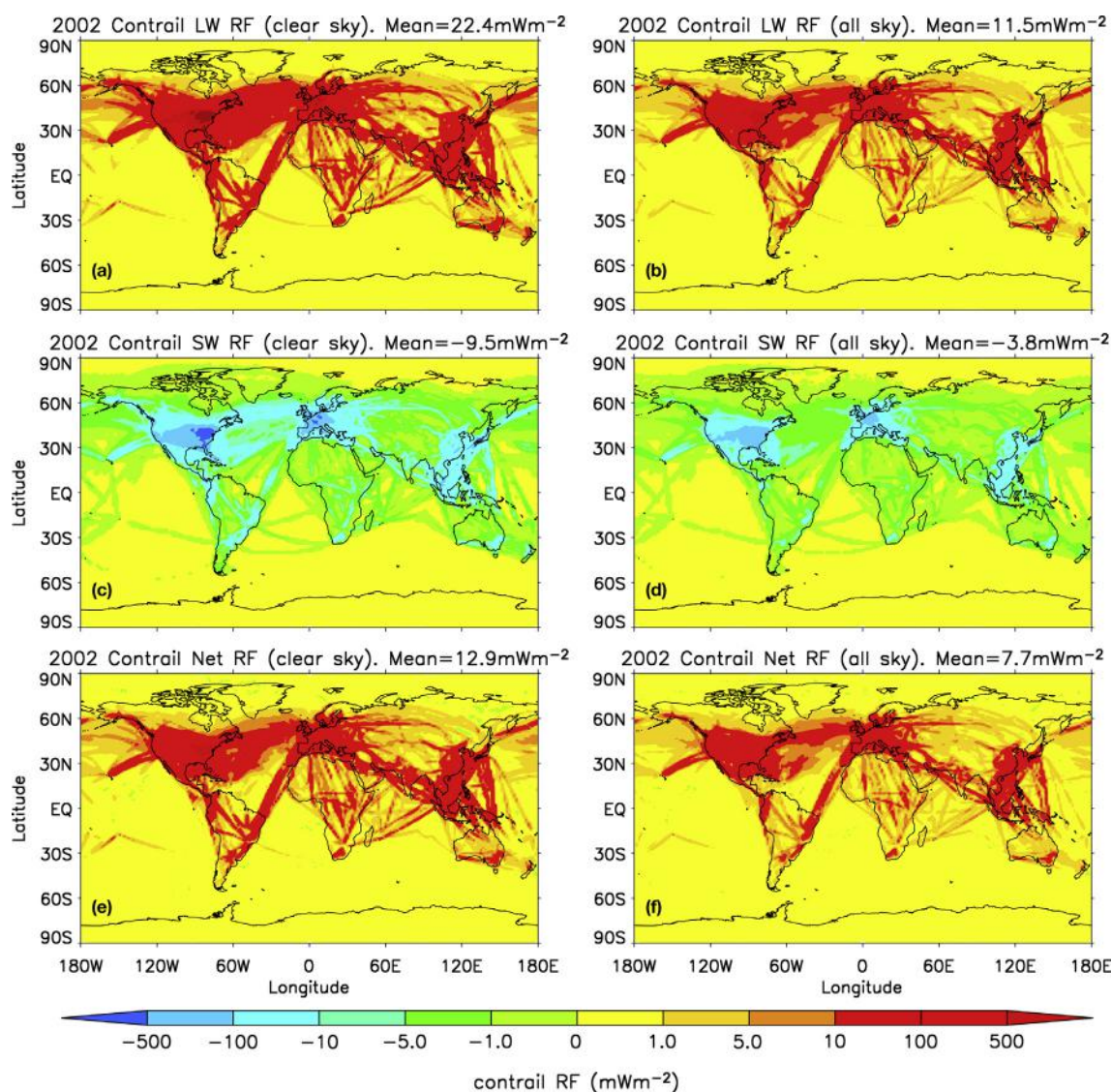


Figure 13 RF (mW m^{-2}) at the top of the atmosphere for 2002 air traffic. (a, b) LW; (c, d) SW; (e, f) net for (left) clear-sky forcings and (right) for all-sky forcings. Reproduced from Rap, A., Forster, P.M., Jones, A., et al. 2010. Parameterization of contrails in the UK Met Office Climate Model. *Journal of Geophysical Research* 115: D10205, doi:10.1029/2009JD012443.

the calculations. The SW RF is negative everywhere and strongest for cloud-free skies (Figure 13(c)) over the Eastern USA and Europe where it exceeds 500 mW m^{-2} . Inclusion of the clouds reduces the global SW RF by 60% to -3.8 mW m^{-2} (Figure 13(d)). Combining the SW and LW forcings yields a net RF of 12.9 mW m^{-2} for clear skies (Figure 13(e)) and 7.7 mW m^{-2} for the all-sky case (Figure 13(f)). While the net RF is positive over most regions, negative net RF is seen over a few areas in the Arctic, central tropical Pacific, and over Northeastern Asia.

This example is one of the many computed in various ways during the last two decades. Various scenarios by other researchers have yielded net RF values between 0.4 and 20.0 mWm^{-2} or more depending on many factors and the ways the model treats contrails and clouds. For example, use of a different approach to account for clouds in the model calculations performed for Figure 13 yields 12.0 mW m^{-2} . The

best estimate of global net RF from linear contrails as of 2009 for the year 2005 is 11 mW m^{-2} with a 90% confidence interval range from 5 to 25 mW m^{-2} . Yet, the scientists producing such estimates recognize that the level of scientific understanding remains low for such estimates. Even less understanding and confidence accompanies estimates of contrail-cirrus RF, which ranges from 11 to 87 mW m^{-2} . To put these results in perspective, the linear contrail RF constitutes roughly 20% of all aviation RFs excluding that from contrail cirrus, including those from ozone carbon dioxide, water vapor, and nitrogen compounds. The 2005 best estimate of RF for all aviation, however, is only about 4% of all estimated anthropogenic RF if contrail-cirrus is not included and could be 5% if it were included.

Since air traffic is expected to increase steadily in the coming decades, the same models used to estimate contrail RF for past or present air traffic have been employed to compute

the contrail RF for the future. Based on several scenarios of technology and air traffic changes, climate model estimates of mean linear contrail net RF range from 37 to 55 mW m⁻² for the year 2050. Inclusion of contrail–cirrus boosts the contrail/contrail–cirrus RF estimates up to as high as 315 mW m⁻². Current uncertainties in contrail coverage, optical depth, lifetimes, overlap with lower clouds, and other factors preclude a definitive assessment of the overall contrail impact. Despite these uncertainties, it is clear that, whatever effect they currently have on climate, it will increase in the future.

The Future

Contrails are difficult to study because of their high altitude, large advection rates, and frequent co-occurrence with natural cirrus. Thus, current estimates of their impact are highly uncertain. Nevertheless, their potential for affecting the global climate and providing military intelligence has spurred more interest and focused research into their formation, dissipation, microphysical and morphological characteristics, and methods for suppressing them. Removal of fuel sulfur or use of liquid hydrogen fuels has been suggested as means for diminishing the number of cloud nuclei and, hence, the number of contrails. Tests and theoretical studies have shown that such measures would probably not reduce the frequency of contrails. Hydrogen fuels would cause larger increases in local relative humidity in the exhaust plume causing higher supersaturations than would occur with hydrocarbon fuels. Thus, liquid hydrogen would probably cause more contrails to form, but possibly with greater particle sizes and fallout rates resulting in shorter lifetimes and less radiative impacts. It is possible that a propulsion source that does not require exhaust of water vapor will be necessary to effectively eliminate the generation of contrails from high flying aircraft.

Other methods to minimize contrail formation would involve changes in flight altitude or path. Contrail coverage could be reduced dramatically by flying in the stratosphere where contrail formation conditions are rare. However, other effects from the exhaust and increased fuel usage may limit the amount of stratospheric traffic. Flying at lower altitudes would diminish the number of contrails in tropical areas, but would cause additional coverage in the midlatitudes and polar regions. Conversely, higher mean flight altitudes would decrease contrails over the poles and temperate zones while causing more contrails in the equatorial areas. Ideally, numerical weather prediction models and contrail formation prognostication programs could be used together with flight planners to map out for each destination a sequence of flight altitudes that best avoids contrail formation conditions. Such planning could theoretically be used to neutralize the warming effect of contrails or even induce a small cooling. Such sophisticated planning would require more accurate temperature and humidity data and contrail prediction schemes than currently available as well as a more complex air traffic control network. Future research may provide the tools to minimize the climatic effects of

contrails, but it is likely that these artificial clouds will be a common feature in the sky for many years to come.

See also: **Aviation Meteorology:** Aircraft Emissions. **Clouds and Fog:** Classification of Clouds. **Numerical Models:** Cloud System Resolving Modeling and Aerosols. **Optics, Atmospheric:** Optical Remote Sensing Instruments. **Radiation Transfer in the Atmosphere:** Cloud-Radiative Processes. **Satellites and Satellite Remote Sensing:** Remote Sensing: Cloud Properties.

Further Reading

- Brasseur, G., Gupta, M., 2010. Impact of aviation on climate. *Bulletin of the American Meteorological Society* 91, 461–463.
- Burkhardt, U., Kärcher, B., 2011. Global radiative forcing from contrail cirrus. *Nature Climate Change* 1, 54–58.
- Duda, D.P., Minnis, P., Nguyen, L., Palikonda, R., 2004. A case study of contrail evolution over the Great Lakes. *Journal of the Atmospheric Sciences* 61, 1132–1146.
- Fahey, D.W., Schumann, U., Ackerman, S., et al., 1999. Aviation-produced aerosols and cloudiness. Chapter 3 of IPCC Special Report: Aviation and the Global Atmosphere. Cambridge University Press, Cambridge, UK. 65–120.
- Forster, P., Ramaswamy, V., Artaxo, P., et al., 2007. Changes in atmospheric constituents and in radiative forcing. Fourth Assessment Report of Working Group I of the Inter-governmental Panel on Climate Change. In: *Climate Change*. Cambridge University Press, Cambridge, UK.
- Freudenthaler, V., Homburg, F., Jäger, H., 1995. Contrail observations by ground-based scanning lidar: Cross-sectional growth. *Geophysical Research Letters* 22, 3501–3504.
- Gierens, K., Kärcher, B., Mannstein, H., Mayer, B., 2009. Aerodynamic contrails: Phenomenology and flow physics. *Journal of the Atmospheric Sciences* 66, 217–226.
- Heymsfield, A.J., Bansemer, A., Thompson, G., et al., 2011. Formation and spread of aircraft-induced holes in clouds. *Science* 333, 77–81.
- Lee, D.S., Pitari, G., Grewe, V., et al., 2010. Transport impacts on atmosphere and climate: Aviation. *Atmospheric Environment* 44, 4,678–4,734.
- Mannstein, H., Spichtinger, P., Gierens, K., 2005. How to avoid contrail cirrus. *Transportation Research D10*, 421–426.
- Meyer, R., Mannstein, H., Meerkötter, R., Schumann, U., Wendling, P., 2002. Regional radiative forcing by line-shaped contrails derived from satellite data. *Journal of Geophysical Research* 107, 1–16.
- Minnis, P., Ayers, J.K., Palikonda, R., Phan, D.N., 2004. Contrails, cirrus trends, and climate. *Journal of Climate* 17, 1671–1685.
- Myhre, G., Stordal, F., 2001. On the tradeoff of the solar and thermal infrared radiative impact of contrails. *Geophysical Research Letters* 28, 3119–3122.
- Rap, A., Forster, P.M., Jones, A., et al., 2010. Parameterization of contrails in the UK Met Office Climate Model. *Journal of Geophysical Research* 115, D10205. doi:10.1029/2009JD012443.
- Ryan, A.C., MacKenzie, A.R., Watkins, S., Timmis, R., 2011. World War II contrails: A case study of aviation-induced cloudiness. *International Journal of Climatology*, 9. doi: 10.1002/joc.2392.
- Sausen, R., Isaksen, I., Hauglustaine, D., et al., 2005. Aviation radiative forcing in 2000: An update on IPCC (1999). *Meteorologische Zeitschrift* 14, 555–561. 10.1127/0941-2948/2005/0049.
- Schumann, U., 1996. On conditions for contrail formation from aircraft exhausts. *Meteorologische Zeitschrift* 5, 4–23.
- Toon, O.B., Mlake-Lye, R.C., 1998. Subsonic aircraft: Contrail and cloud effects special study. *Geophysical Research Letters* 25, 1109–1168.
- Unterstrasser, S., Sölch, I., 2010. Study of contrail microphysics in the vortex phase with a Lagrangian particle tracking model. *Atmospheric Chemistry and Physics* 10, 10,003–10,015.
- Voigt, C., Schumann, U., Jurkat, T., et al., 2010. In-situ observations of young contrails – Overview and selected results from the CONCERT campaign. *Atmospheric Chemistry and Physics* 10, 9039–9056.

Cloud Microphysics

D Lamb, The Pennsylvania State University, University Park, PA, USA

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

The microphysics of a cloud is concerned with the aqueous particles making up the cloud, how they are classified, and how they change with time. The particles, either liquid or solid (ice), vary greatly in size and shape, so mathematical methods are used to characterize them. The growth of cloud particles, as well as any transformation of phase, is determined by considering the physics of condensation, collision-coalescence, and nucleation. The diverse physical processes operate in a complicated dynamical environment to produce the precipitation that falls to the ground.

Introduction

Cloud microphysics is the branch of the atmospheric sciences concerned with the many particles that make up a cloud. Relative to the cloud as a whole, the individual particles are very small and so exist on the ‘microscale,’ that is, range over distances from fractions of a micrometer to several centimeters. The ‘microstructure’ of a cloud is a specification of the number concentrations, sizes, shapes, and phases of the various particles, factors that strongly influence the behavior of the particles and cloud lifetimes. The abilities of clouds in general to produce rain or snow, generate lightning, and alter the radiation balance of the earth, for instance, stem in large part from their individual microstructures working in concert with the local air motions. Cloud physicists characterize the diverse microstructures of atmospheric clouds and try to understand the processes that cause them to change with time.

Clouds typically form in response to changes in atmospheric conditions on scales much larger than those of the particles or, indeed, of the cloud itself. Synoptic-scale wind patterns and convection often force air to rise, which causes the pressure and temperature to decrease. This cooling during ascent leads to lowering of the equilibrium vapor pressures of the liquid and solid phases of water. ‘Excess’ vapor, that amount above the equilibrium value, thus develops and causes a state of disequilibrium to exist in the rising air parcels. Condensation is nature’s way of attempting to restore equilibrium.

A cloud, especially during its early stages of formation, often exhibits the properties of a colloidal system, a suspension of tiny particles that follow the air flow and interact only weakly with one another. Whereas the individual aqueous particles may form, grow, and subsequently disappear, the system as a whole remains microphysically stable for some time. The discipline of cloud microphysics helps us understand the specific mechanisms by which clouds form, break any colloidal stability, and form precipitation.

Microphysical Descriptions

The microstructure of a cloud may be categorized and described statistically in a number of ways. Empirical descriptions, typically derived from *in situ* or remotely sensed measurements of clouds, facilitate communications among

atmospheric scientists and provide the first glimpses of the physical processes operative at the time the measurements were taken. The microstructure may differ substantially from one part of a cloud to another, and it evolves with time in ways that depend on the environmental setting and the specific physical processes that are active at the time. Mathematical and numerical models, depending on their purpose, may employ the empirical descriptions directly, or they may calculate the time evolution of the microstructure if the relevant processes are included.

The particles in a cloud vary in phase, composition, size, and shape. The phase, whether solid or liquid, is the most fundamental descriptor of the aqueous particles, but the nonaqueous aerosol particles are also crucial to cloud formation and evolution. Liquid condensate forms preferentially on soluble ‘cloud condensation nuclei’ (CCN), which typically contain sulfates and nitrates. Ice particles, by contrast, often form on ‘ice nuclei,’ insoluble aerosol particles that contain crustal components or biological matter having crystalline structures related to that of ice. The dominant phase of the aqueous particles forms the basis for classifying clouds as ‘warm,’ when only liquid drops are present, or ‘cold,’ when ice is involved (with or without liquid drops). The ‘mixed-phase’ region of a cloud, throughout which both the liquid and solid (ice) phases of water may be present simultaneously, is that vertical zone between the melting level (0 °C) and the –40 °C isotherm, the practical lower limit for liquid water to exist in the metastable (i.e., ‘supercooled’) state. The relative abundance of each phase in a given cloud depends on the prevailing meteorological conditions and the microphysical processes active during the life cycle of the cloud.

The sizes and shapes of the aqueous particles play important roles in cloud development. Whereas ice particles can and do appear in a wide variety of shapes, all but the largest liquid drops tend to remain spherical because of surface tension effects. **Figure 1** depicts the various categories of liquid drops based on their sizes. Note that the size of a particle is an important determinant of its terminal fall speed and hence of its ability to fall against prevailing updraft speeds ($\sim 10 \text{ cm s}^{-1}$ in stratiform clouds; $\sim 10 \text{ m s}^{-1}$ in convective storms). Drizzle drops represent a transition between the small ‘droplets’ that follow the air motions and the larger drops that sediment and possibly reach the ground as rain. The shapes of bigger raindrops tend to become distorted because of the large dynamic

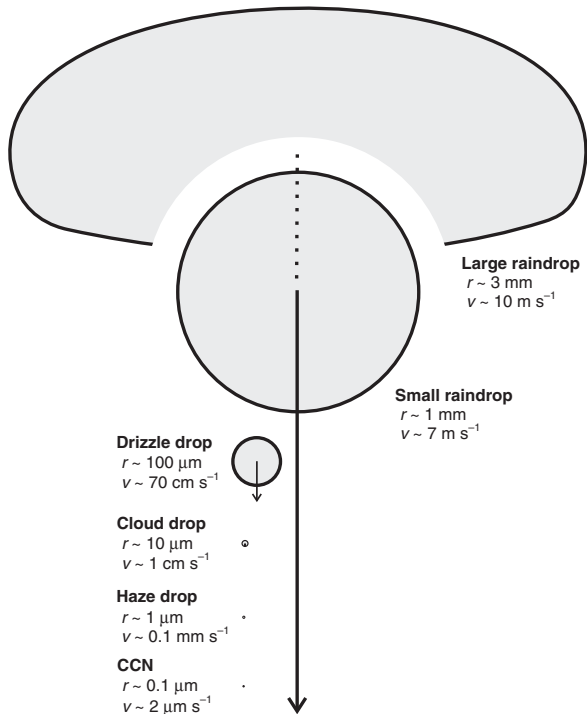


Figure 1 Various categories of liquid drops in clouds. The drops are drawn roughly to scale relative to each other, as are the arrows representing the terminal fall speeds. Representative magnitudes of drop radii (r) and fall speeds (v) are as indicated. The large raindrop is shown distorted to suggest the effect of dynamic pressure on its underside due to its fall through the air. 'CCN' represents a cloud condensation nucleus, a solution droplet that serves as the initial site of condensation.

pressure on the lower side, giving the rough appearance of a ball of dough pressed lightly onto a tabletop. This flattening of the underside causes the aerodynamic resistance to be larger than that of spherical drops of equivalent volume and so limits the fall speeds of raindrops to about 10 m s^{-1} .

The ice particles in a cloud vary enormously in both size and shape. The first ice to appear in many clouds tends to be small ($\sim 10 \text{ μm}$ across) and monocrystalline in structure. Single crystals of ice subsequently grow into hexagonal prisms (each prism being bounded by two 'basal' faces and six 'prism' faces) with axial ratios (length along the principal or ' c ' axis divided by the ' a ' axis, the width across the corners of the hexagon) that depend systematically on the temperature. As shown in **Figure 2**, 'plates' (c/a axial ratios less than unity) are found when the temperature is either between 0 and about -3°C or between about -8 and -22°C . On the other hand, 'columns' ($c/a > 1$) appear in the approximate temperature ranges -3 to -8°C and less than -22°C . Deviations from simple hexagonal prisms are common and depend on the excess vapor density, as suggested by the various symbols in **Figure 2**. In addition to the many single crystals, a number of polycrystalline forms of ice are found in cold clouds. For instance, several to hundreds of single crystals may clump together to form 'aggregates' (i.e., snowflakes), and supercooled cloud droplets may freeze onto ice particles, giving rise to rimed crystals, graupel, and hail.

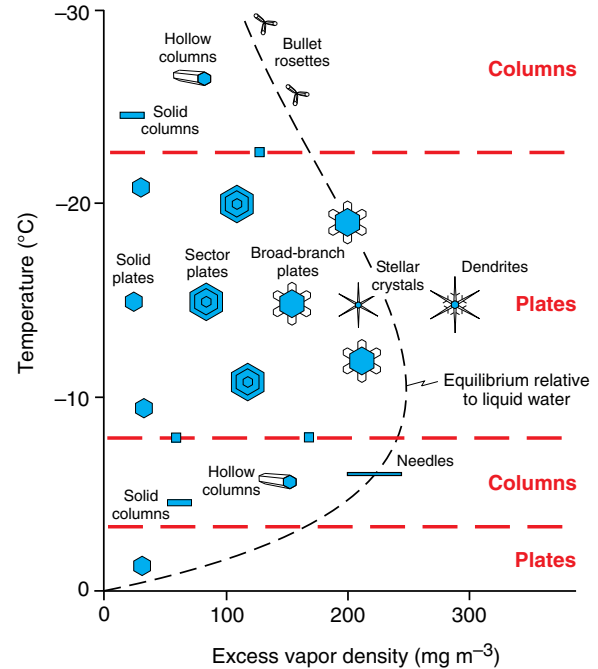


Figure 2 Schematic representation of the wide variations in the shapes ('habits') of ice crystals found in clouds at the indicated temperatures and excess vapor densities (relative to the ice phase). The excess vapor density of air saturated with respect to supercooled liquid water is shown by the dashed curve. The heavy dashed lines identify the approximate temperatures at which the primary habits change between plates and columns. The diagram has been oriented to suggest the normal decrease of temperature with altitude.

Individual crystals seldom grow to more than a few millimeters across, but hailstones can sometimes exceed 10 cm in diameter.

The cloud microstructure is best viewed as a multidimensional specification of the number concentrations of the various particles in a cloud. In the case of liquid drops, one needs to consider only the sizes of the drops and how they vary in space and time. The nonspherical ice particles, by contrast, require additional specification, such as axial ratio, to account for the shapes of the particles. 'Spectral' descriptions of the cloud particles tell us in effect how many of what kinds and sizes of particles are present at given locations within a cloud.

Often, it is useful to characterize the spectra in mathematical terms to minimize the number of variables needed to represent the microstructure. The size distributions of raindrops from convective storms, for example, can be described by analytical functions that have been fitted to observational data by specifying two or three parameters, as shown in **Figure 3**. An exponential function, one subclass of which is referred to as a 'Marshall–Palmer' distribution, has the form

$$n(D) = n_0 \exp(-\lambda D) \quad [1]$$

where n_0 and λ are the parameters fitted to the measured number concentration $n(D)$ of drops within a unit size interval about diameter D . Exponential distributions are used frequently because only two parameters (here, n_0 and λ) need to be specified. However, as the heavy dotted line in **Figure 3** shows, exponential distributions often overestimate the

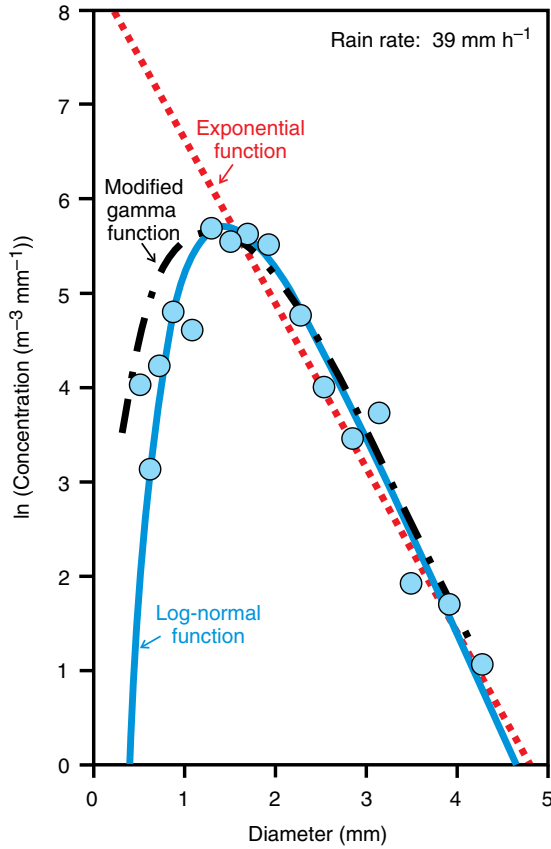


Figure 3 The size distribution of raindrops from a convective storm. Circles: calculated values based on 1-min samples from a 20-channel disdrometer. Dotted line: an exponential fit to the data. Dot-dashed curve: a modified gamma function. Solid curve: log-normal function. Adapted from Feingold, G., Levin, Z., 1986. The lognormal fit to rain-drop spectra from frontal convective clouds in Israel. *Journal of Applied Meteorology and Climatology* 25, 1346–1363.

number of smaller drops. A more general form, the modified gamma distribution,

$$n(D) = n_0 D^\mu \exp(-\lambda D) \quad [2]$$

attempts to correct this deficiency, although at the expense of requiring an additional parameter (μ). At least for the data shown in Figure 3, a log-normal function of the form

$$n(D) = \frac{N_T}{\sqrt{2\pi D \ln \sigma_g}} \exp(-\ln^2(D/D_g)/2 \ln^2 \sigma_g) \quad [3]$$

works well with appropriate choices of the three parameters: total number concentration N_T , geometric median diameter D_g , and geometric standard deviation σ_g . It is important to recognize that empirical size spectra simply describe the cloud microstructure without regard to the mechanisms that produced it.

Warm-Cloud Microphysics

The liquid drops in ‘warm’ clouds evolve spectrally via microphysical processes that interact in complicated ways with the

larger scale cloud environment. The process of condensation, for instance, starts during upward motion of moist air and continues even as the drops interact with themselves and grow into raindrops. As the cloud updraft entrains dry environmental air, ceases, or possibly reverses, evaporation may dominate for a time and change the microstructure in important ways. The ever-changing population of drops often influences the very atmospheric motions that spawned the drops in the first place.

Droplet Formation

The many small droplets in a cloud form initially as excess vapor condenses onto the larger, more soluble aerosol particles (i.e., onto the CCN). The solute in the CCN lowers the equilibrium vapor pressure of the liquid droplets through molecular scale effects, whereas the droplet curvature increases it.

The two opposing effects of solute and curvature are typically combined in Köhler theory to yield the equilibrium saturation ratio S_K as a function of the droplet radius r :

$$S_K = a_w \cdot \exp(A/r) \quad [4]$$

Here, the first factor on the right-hand side of eqn [4] describes the vapor pressure lowering effect of the solute in terms of the water activity $a_w = 1 - ix_s$, where x_s is the mole fraction of nonvolatile solute that effectively dissociates into i molecular or ionic components. The second factor in eqn [4] accounts for the effect of droplet curvature on vapor pressure, where $A = 2\sigma_{LV}/(n_L RT)$ is a function of the physically relevant variables, the liquid–vapor surface free energy σ_{LV} ($= 72 \text{ mJ m}^{-2}$), the liquid water density n_L ($= 5.5 \times 10^4 \text{ mol m}^{-3}$), the universal gas constant R ($= 8.31 \text{ J mol}^{-1} \text{ K}^{-1}$), and the temperature T . Because each droplet grows by the sole addition of water, its total solute content N_s remains constant and the solute mole fraction $x_s = N_s/(n_L V_d)$ decreases as the droplet grows in volume $V_d = 4\pi r^3/3$. Equation [4] can thus be expressed in terms of the supersaturation s_K needed to maintain equilibrium with the solution droplet:

$$s_K \equiv S_K - 1 = \frac{A}{r} - \frac{BiN_s}{r^3} \quad [5]$$

where $B = 3/(4\pi n_L)$ and approximations suitable for initial cloud formation have been made. Equation [5] can be seen to represent a family of ‘Köhler’ curves, each of constant solute content. The competing effects of solute and curvature yield a maximum in the equilibrium saturation ratio, as shown in Figure 4. The solute effect predominates at small sizes, whereas the curvature effect predominates at large sizes. The maximum value, termed the critical supersaturation, must be overcome by the ambient supersaturation before the particle can ‘activate’ and grow spontaneously as a cloud droplet. The critical supersaturation, found mathematically by setting the first derivative of s_K equal to zero, decreases as the solute content increases:

$$s_c = \left(\frac{4A^3}{27BiN_s} \right)^{1/2} \quad [6]$$

The larger the aerosol particle, the smaller is the ‘critical supersaturation’ that needs to be exceeded. Note that the

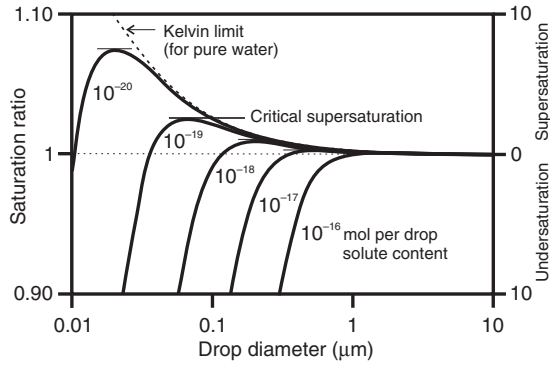


Figure 4 A set of Köhler curves. Each curve represents the saturation ratio of vapor needed to maintain equilibrium with solution drops containing the indicated solute contents.

critical supersaturation, $s_c \propto 1/N_s^{1/2} \propto 1/r^{3/2}$, is especially large for particles with the smallest solute contents, so small aerosol particles may never experience sufficiently high ambient supersaturations to become activated. Such small particles remain as submicron haze droplets interstitially between the much larger cloud droplets.

Growth by Condensation

Each cloud droplet acts as a tiny sink of water vapor during active growth in an updraft. As condensation proceeds, the concentration of vapor immediately over each droplet surface (within a few mean free paths of the air molecules) is reduced relative to the average vapor concentration far from the droplet. The radial gradients of vapor concentration thus established give rise to a net flux of vapor molecules toward the drop by the process of molecular diffusion. The water molecules must also be transported across the liquid–vapor interface, but it is the vapor diffusion step that tends to limit the mass transport under most cloud conditions. Nevertheless, the change of phase from vapor to liquid at the surface results in a slight warming of the droplet due to the added enthalpy of condensation, energy that must be conducted away from the droplet through the air. This energy consequence of condensation raises the equilibrium vapor pressure of the liquid and imposes an additional limitation to the growth rate.

The theory that simultaneously accounts for the exchanges of vapor and energy between a growing droplet and the surrounding air was first developed fully by Maxwell in the nineteenth century. The resulting expression for the linear growth rate is

$$\frac{dr}{dt} = G(s - s_K) \frac{1}{r} \quad [7]$$

where G is a growth parameter that acts as an effective diffusivity and varies slowly with the temperature and pressure. Note that a droplet grows only to the extent that the ambient supersaturation (s) exceeds the equilibrium value (s_K). As r becomes large, $s_K \rightarrow 0$ and the growth rate $dr/dt \propto 1/r$, indicating that the droplets grow relatively more slowly as they become larger.

A population of growing cloud droplets obtains its water from a common supply, namely the vapor initially carried with the rising air parcel. Competition for the available vapor among all the droplets sets up a strong interplay between the condensation kinetics, the aerosol population, and the vapor field. Results from numerical computations of droplet growth within an air parcel rising adiabatically are shown in [Figure 5](#) for the case of a relatively clean maritime environment. The ambient supersaturation (dashed curve) builds up in the rising parcel because of the decrease in temperature with height and the consequent lowering of equilibrium vapor pressure. Aerosol particles with relatively large solute contents, those having low critical supersaturations, activate and become cloud droplets that grow rapidly (seen in [Figure 5](#) by the sudden changes in slope). Condensation onto the activated droplets removes water vapor from the air, so the supersaturation increases progressively more slowly as the droplets grow and deplete the vapor. At some point, usually within several meters above cloud base, the rate of vapor removal exceeds the rate at which excess vapor is generated, and the supersaturation then steadily decreases. The maximum supersaturation, $s_{\max}$, attained in the cloud is influenced strongly by the abundance of aerosol particles and becomes an important determinant of the microstructure of warm clouds. Once the maximum in the supersaturation is reached, no new particles can be activated. The initial number concentration of cloud droplets is thus established low in the cloud and reflects the properties of the aerosol population present at the time of cloud formation.

The interaction between the aerosol population and the vapor field can be understood with the help of [Figure 6](#) (also available as a sequence of PowerPoint slides in the supplementary material). Early in the evolution of supersaturation, say at point 1 in panel (a), the only particles that can activate

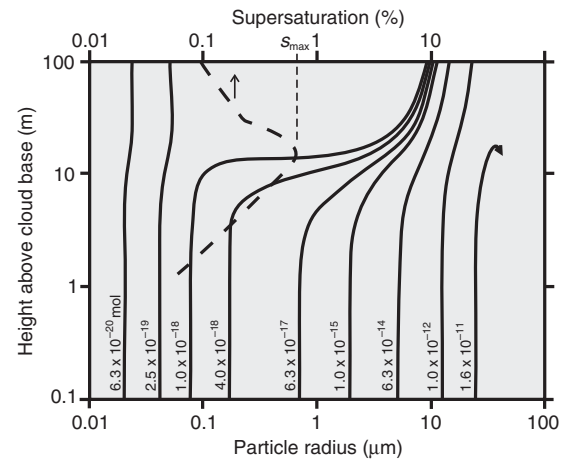


Figure 5 Results from numerical calculations of activation and growth of droplets in an air parcel ascending adiabatically at the rate of 15 cm s^{-1} . The total concentration of particles is 50 cm^{-3} . Solid curves: droplet radii (lower scale) at various heights above cloud base for droplets containing the indicated solute contents (number of moles of nonvolatile solute per particle). Dashed curve: ambient supersaturation (upper scale). Adapted from Mordy, W., 1959. Computations of the growth by condensation of a population of cloud droplets. *Tellus* 11, 16–44.

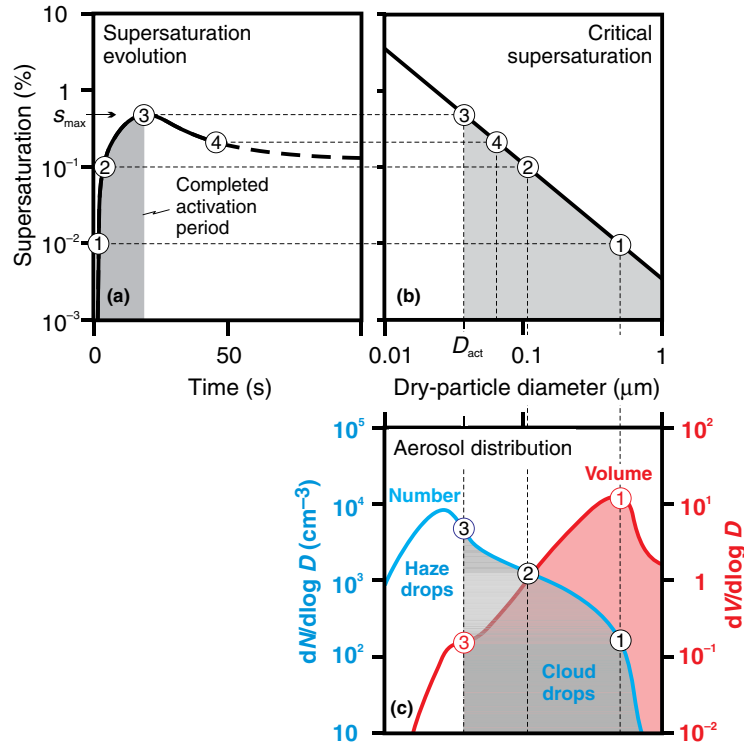


Figure 6 Graphical relationships between the evolution of ambient supersaturation in an adiabatically ascending parcel (panel a), the critical supersaturations of individual aerosol particles (b), and the size distributions of the aerosol number and volume concentrations (c). The circled numbers show specific values of each variable at selected supersaturations; the shaded regions identify portions pertaining to the activated aerosol after the supersaturation has reached its maximum value (s_{max}).

are those with small critical supersaturations (corresponding point in panel (b)). Such particles are relatively large and so have large solute contents. These initially activated particles typically reside in the tail of the aerosol size distribution ('Number' curve in panel (c)) and so represent only a small fraction of the total aerosol population. However, these are the largest particles and so represent a significant fraction of the total solute incorporated into the cloud water (see 'Volume' curve in panel (c)).

As the supersaturation continues to increase (e.g., point 2 in panel (a) of Figure 6), progressively smaller particles (those with increasing critical supersaturations; panel (b)) are activated and added to the population of droplets. The greater the number of growing droplets, the greater are the opportunities for vapor to be consumed. Once the maximum supersaturation has been reached (point 3 in panel (a)), all the particles that can be will have been activated. The limiting diameter of the dry aerosol particles activated, those with critical supersaturations equal to or less than the maximum ambient supersaturation, is called the activation diameter (D_{act}). With additional uplift of the cloud parcel, the supersaturation decreases (e.g., point 4) below the critical supersaturations of the remaining particles in the aerosol population. The inactivated set of particles is the haze droplets and the activated set, the cloud droplets. As long as the air continues to rise, the supersaturation stays positive, and the activated droplets continue to grow.

The close packing of the curves near the top on the right-hand side of Figure 5 is consistent with eqn [7] and indicates

that the droplets tend to bunch together in radius. This narrowing of the drop spectrum is an inherent property of adiabatic condensation and poses a hindrance to the formation of precipitation. Detailed calculations show that individual droplets experiencing a supersaturation of 1% require hundreds of seconds to grow by condensation to radii much beyond $10 \mu\text{m}$. Significant additional growth depends on collisions between particles.

Collisional Interactions

Individual pairs of cloud drops occasionally collide with one another. If two drops 'coalesce' during a particular collision, a single, larger drop replaces the two parent drops in the cloud. Repeated collision-coalescence events eventually lead to large drops that fall rapidly and become raindrops.

The growth of drops through collisional interactions may be quantified by considering the separate probabilities for collision and for coalescence. Most commonly, collisions result when a larger drop (the 'collector' drop) overtakes a smaller drop (the 'collected' drop) during its fall through the air. However, not all drops in the geometrical path of the collector experience collisions, for the simple reason that the air deviating around the collector drop 'pushes' the smallest drops out of the way. The fraction of drops in the path of the collector that do collide with it is the 'collision efficiency' (E), a complicated function of both the collector- and collected-drop sizes. The maximum collision efficiency can approach unity when the

collector-drop radii are greater than about 40 μm and the collected-drop radii exceed about 5 μm . However, for collector-drop radii less than about 20 μm , the collision efficiency becomes very small. The ‘coalescence efficiency’ (ϵ), the fraction of drop–drop collisions actually resulting in the formation of a larger drop, is often less than unity for larger collected drops because of drop distortion and the trapping of air at the point of collision. The ‘collection efficiency’ (E_c) is the product of the collision and coalescence efficiencies, $E_c = E \cdot \epsilon$, and typically reaches a weak maximum at intermediate collected-drop radii. The collection efficiency provides an overall indication of the effectiveness of drop growth by collision–coalescence.

Collision–coalescence becomes a powerful mechanism for generating raindrops under appropriate microphysical conditions. Because of the *de facto* thresholds that exist on both the collected- and collector-drop sizes, collision–coalescence tends to begin in the tail of the cloud-drop size distribution. Initially, only a tiny fraction of the bigger droplets will collide and coalesce with neighboring droplets, yielding slightly larger droplets that then have enhanced probabilities of collecting additional droplets. The growth process accelerates as the collection efficiencies increase and other drops join this favored subset of the drop population. Eventually, a new mode in the drop-size distribution emerges, as shown by the maxima toward the right-hand side of Figure 7. Once the drops in this large-drop mode exceed a few hundred microns, they grow rapidly in mass at a more-or-less continuous rate given by

$$\frac{dm}{dt} = K(r_1, r_s) \cdot \omega_L \quad [8]$$

where $K(r_1, r_s) = \pi(r_1 + r_s)^2 E_c \cdot (v_1 - v_s)$ is the collection kernel and ω_L is the liquid water concentration (mass per unit volume of cloudy air). The collection kernel is best viewed as the effective volume of cloudy air (containing small droplets of radius r_s , each falling at rate v_s) that is swept out in unit time by the collector drop of radius r_1 having fall speed v_1 . In this continuous-growth regime (in which $r_s \ll r_1$, $v_s \ll v_1$, $v_1 \propto r_1$, and $E_c \cong 1$), one finds to first approximation $dm/dt \propto r_1^3 \propto m$. Thus, once the collision–coalescence process gets started, the large drops increase in mass (and size) exponentially with time, until the supply of cloud droplets is exhausted or the drops rupture and disperse as many smaller drops.

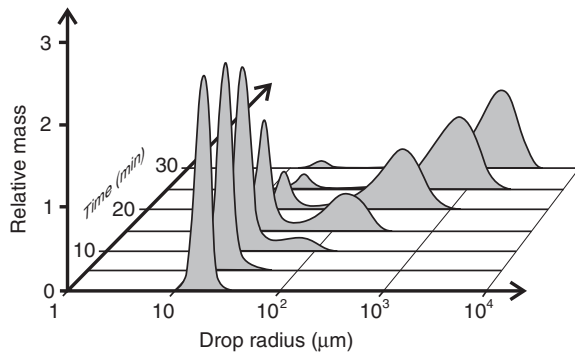


Figure 7 Evolution of the size distribution of drops undergoing collision–coalescence. Adapted from Berry, E. X., Reinhardt, R. L., 1974. An analysis of cloud drop growth by collection: Part II. Single initial distributions. *Journal of Atmospheric Sciences* 31, 1825–1831.

Cold-Cloud Microphysics

Clouds are classified as ‘cold’ once ice particles form and become active players in the cloud microphysics. Whereas ice particles are necessary components of cold clouds, the liquid drops are nevertheless often present and important to the evolution of the cloud microstructure. The mixed-phase zone of a cloud, where the ice particles and liquid drops interact, is microphysically the most active portion of a cloud. ‘Glaciation,’ the transformation of a cloud from supercooled liquid drops to ice particles, is complicated by the diversity of interactions that can take place.

Ice Formation

Ice can form once the liquid drops have become supercooled by at least 5 °C, even though the supercooled state can persist in some clouds to temperatures as low as –40 °C. The first ice particles in a supercooled cloud most commonly appear when the temperature is between –10 and –15 °C following ‘primary nucleation,’ a process by which submicron, insoluble aerosol particles catalyze ice formation by acting as molecular templates for the crystal lattice. Such primary ice particles may form directly from the vapor phase (via ‘deposition nucleation’), but more commonly they arise from the freezing of supercooled cloud droplets (via ‘freezing nucleation’). The freezing of droplets at relatively high temperatures (greater than about –18 °C) tends to yield single crystals that subsequently grow into crystallographically aligned double plates. At lower temperatures, however, the probability of forming multiple crystals within a single droplet becomes large, giving rise to polycrystalline forms, such as bullet rosettes.

Some ice particles form because of the prior existence of other ice particles. Such ‘secondary’ ice particles arise via several mechanisms, none of which are fully understood. In some situations, crystals may ‘fragment,’ such as when the delicate arms of dendrites break off, thereby increasing the ice particle number concentration. When conditions are just right, as when the temperature is between about –3 and –8 °C and ‘graupel’ particles are actively growing by riming (i.e., accreting supercooled cloud water), tiny splinters of ice may be released that subsequently grow into columnar ice crystals. The self-breeding, or ‘multiplication’ of ice seems to be important in clouds that glaciate rapidly.

Growth by Vapor Deposition

Individual ice particles grow initially by the deposition of vapor onto their surfaces. As with the growth of a cloud droplet by condensation, water molecules must first diffuse to the particle surface from the supersaturated vapor field surrounding the particle. However, the transport of molecules across the vapor–solid interface cannot be ignored, if for no other reason than the need to account for facets and the nonspherical shapes of the ice particles. The molecular scale processes on the crystal surface involve migration of molecules across the surface, from the point of first contact to a step that may be a micron or more away. As the step gathers adsorbed molecules, it propagates across the surface, contributing an additional layer (of thickness equal to the step height) of molecules to the lattice.

The rate of advancement of the crystal face is determined largely by the frequency with which the steps are generated, a factor that depends on whether the steps originate from two-dimensional layer nucleation or from the emergence of screw dislocations on the surface.

The aspect ratio of an ice crystal reflects the relative rates of growth of the basal and prism faces. For reasons that remain largely unknown, the linear growth rates vary with the temperature in complicated ways, giving rise to the observed alternation of primary habit between plates and columns with temperature (as depicted in [Figure 2](#)). At relatively large supersaturations with respect to ice, the vapor gradients in the vicinity of a given crystal face become important, leading to the bunching of steps, 'hollowing' of the face near its center, and a myriad of secondary habit features superimposed on the primary habit. Good physical reasons exist why one seldom if ever finds any two ice crystals alike in nature.

Riming and Aggregation

Riming and aggregation are both processes involving collection. As in warm clouds, collisions between particles must first occur and then the colliding particles must stick together to form a combined particle. In the case of riming, an ice particle collects supercooled droplets that freeze on contact with the ice surface. By contrast, aggregation involves the collisional interaction of two ice particles, with no change of phase. Aggregation is a complicated process, in part because snowflakes fall erratically and because two solid particles may simply bounce apart after colliding. Snow crystals are most likely to stick together at temperatures within a few degrees of the melting point (because of sintering), or when the arms of dendrites can interlock.

Growth via the riming process occurs by stages that depend on particle size and the rate at which supercooled droplets are accreted. Initially, during the 'crystal stage,' the rate of accretion is slow, the collision efficiency becoming appreciable only once the vapor-grown crystal attains an *a*-axis dimension of about 150 μm for plates, 25 μm for columns. The crystal becomes lightly to moderately rimed, but the crystal morphology remains identifiable. Such growth is termed 'dry' because each droplet freezes rapidly at the spot of impingement. The 'graupel stage' begins once the crystal identity becomes obscured by the shroud of 'dry' rime ice on the particle. During this and the subsequent 'hail stage,' the ice particle grows in mass at rates described reasonably well by [eqn \[8\]](#), with suitable adjustments in the parameters. The hail stage is distinguished from the graupel stage by the formation of one or more layers of clear ice, which results when the rate of accretion exceeds the ability of the particle to dissipate the enthalpy added by the freezing of the supercooled water. Such 'wet' growth occurs when the surface temperature rises to 0 $^{\circ}\text{C}$ and the accreted liquid spreads across the surface before freezing. Hailstones represent one extreme to which the microstructure of clouds can evolve.

Precipitation

Precipitation, whether in the form of rain, snow, sleet, or hail, generally results once the aqueous particles in a cloud

have grown sufficiently large to fall against the local updraft. In the case of a stratiform cloud, one characterized by rather weak and uniform updrafts over a broad area, the precipitating particles may simply fall out through the base of the cloud, in the process depleting condensate from the cloud and depositing it on the ground. On the other hand, in convective storms, in which large local updraft speeds can aerodynamically support big particles, the precipitation itself may influence the motions of air through the cloud. The large mass associated with the precipitation commonly initiates a downdraft along the edge of the updraft, causing the microphysical and dynamical aspects of cloud evolution to become intertwined in complicated ways. As an aid to the discussion below, [Figure 8](#) offers a summary of the various processes operating in the 'warm' and 'cold' parts of a representative convective cloud during precipitation formation.

In the warm parts of clouds, large drops can emerge out of the stable population of cloud droplets only through collisional interactions. Condensational growth alone is too slow, but it must be recognized as a necessary process, for the numerous cloud droplets serve as the feedstock for the growth of the larger drops. The collision-coalescence process becomes an effective mechanism for breaking the colloidal stability of the cloud once the threshold size ($\sim 25 \mu\text{m}$ diameter) for collection has been overcome. The needed 'coalescence embryos' can arise from the droplet population itself (most commonly in clean, maritime environments or during turbulent mixing), alternatively from 'giant nuclei' in dusty regions. Once the warm-rain mechanism is established, the raindrops grow rapidly by sweeping out the smaller cloud droplets until they themselves become unstable and rupture into fragments during their fall to earth.

In the cold parts of a cloud, the colloidal stability of the cloud is broken once the ice phase has been nucleated in the presence of supercooled droplets. This 'ice-crystal' mechanism, often termed the Bergeron process, arises from the inherent difference in the equilibrium vapor pressures of liquid and solid water. The relatively low vapor pressure of ice compared with that of the droplets at any given temperature gives the ice crystals a growth advantage by causing water vapor to transfer (via diffusion) from the many cloud droplets to the fewer ice crystals. The process proceeds rapidly, especially in the temperature range between -12 and -18°C , permitting the ice crystals to attain sizes sufficient to initiate the other cold-cloud growth mechanisms, such as aggregation and riming. These large ice particles eventually fall into the warm part of the cloud, where they may melt and join the population of raindrops formed by the collision-coalescence process. The ice process can be an effective initiator of precipitation in both stratiform clouds and summer thundershowers.

The efficiencies with which clouds develop precipitation depend partly on the types of microphysical processes that are active and partly on the environmental settings in which the clouds form. For instance, the relative ease with which 'maritime' clouds release precipitation compared with 'continental' clouds most likely stems from the differences in aerosol abundance found in the different air masses. The relative absence of active sources of aerosol particles over the open

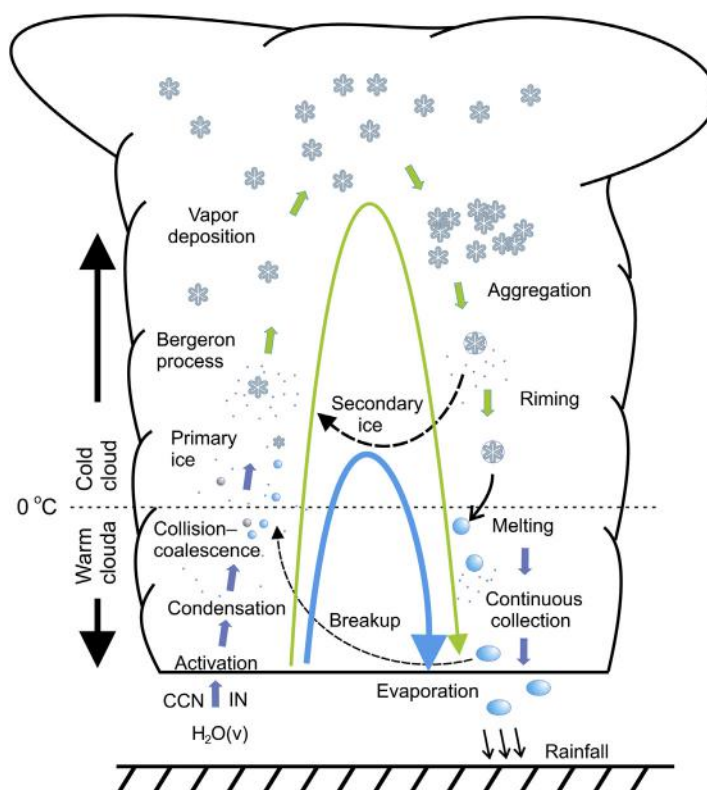


Figure 8 Summary of microphysical processes operating inside a deep convective cloud.

oceans leads to low CCN concentrations, which in turn means that those few droplets that do form in maritime clouds tend to grow faster on average than do those in continental clouds. The collision-coalescence process thus gets started early in the life cycle of maritime clouds, providing such clouds with a decisive microphysical mechanism for developing precipitation. Within a given climatic regime, storm organization on the mesoscale seems to be an important contributor to precipitation efficiency. Small magnitudes of vertical wind shear at the time of cloud formation tend to favor vertically erect storms with high precipitation efficiencies, presumably because the incipient precipitation particles can then fall directly through the condensate-rich inflow of the storm. At the same time, however, such systems tend to be short lived and yield relatively small total amounts of precipitation. Storms that form in environments in which the wind varies modestly with height in both speed and direction last longer and yield the most precipitation, for then synergism arises between the dynamical time scales of the storm and the time scales for the microphysical processes to operate effectively.

Supplementary data related to this article can be found at <http://dx.doi.org/10.1016/B978-0-12-382225-3.00111-0>.

See also: Numerical Models: Parameterization of Physical Processes: Clouds. Tropospheric Chemistry and Composition: Cloud Chemistry.

Further Reading

- Lamb, D., 1999. Atmospheric ice. Update 1. In: Trigg, G. (Ed.), *Encyclopedia of Applied Physics*. Wiley-VCH, Weinheim, pp. 3–25.
- Lamb, D., 2001. Rain production in convective storms. no. 50. In: Doswell, C.A. (Ed.), *Severe Convective Storms*, Meteorological Monographs, vol. 28. American Meteorological Society, Boston, pp. 299–321.
- Lamb, D., Verlinde, J., 2011. *Physics and Chemistry of Clouds*. Cambridge University Press, Cambridge.
- Pruppacher, H.R., Klett, J.D., 1997. *Microphysics of Clouds and Precipitation*, second ed. Kluwer Academic Publishers, Dordrecht.
- Rogers, R.R., Yau, M.K., 1989. *A Short Course in Cloud Physics*, third ed. Pergamon Press, Oxford.

Classification of Clouds

AL Rangno (Retiree), University of Washington, Seattle, WA, USA

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

The history of how clouds came to be named and what those names are is reviewed. The typical visual and microstructural attributes of the 10 cloud genera and their most commonly seen species or varieties are accompanied by a short cloud pictorial atlas.

Introduction

Official synoptic weather observations, available for more than 70 years at weather observing stations around the world, have contained information on the coverage of various types of clouds based on a classification system that was largely in place by the late 1890s. Cloud observations have had increased value in recent years besides their traditional role in helping to assess the current condition of the atmosphere; those observations are now also seen as providing a long-term record from which changes in cloud coverage and type associated with climate change might be discerned that are not detectable in the relatively short record of satellite data. This article discusses what a cloud is, the origin of the classification system of clouds that is used today, and contains photographs of the most commonly seen clouds.

What Is a Cloud?

As defined by the World Meteorological Organization (WMO, 1969), a cloud is 'an aggregate of minute, suspended particles of water or ice or both that are in sufficient concentrations to be visible'; a collection of 'hydrometeors,' a term that also includes in some cases, due to perspective, the precipitation particles that fall from them.

Clouds are tenuous and transitory; no single cloud element, even within an extensive cloud shield, exists for more than a few hours, and most small clouds in the lower atmosphere exist for only a few minutes. And, in precise numbers, the demarcation between a cloud and clear air is hard to define: How many cloud drops per liter constitute a cloud? When are ice crystals and snow termed 'clouds' rather than precipitation? When are drops or ice crystals too large to be considered 'cloud' particles but rather 'precipitation' particles?

These questions are difficult for scientists to answer in unanimity because the difference between cloud particles and precipitation particles, for example, is not black and white rather they represent a continuum of fall speeds. For some scientists, a 50- μm diameter drop represents a 'drizzle' drop because it must have formed from collisions with other drops, but for others, it may be termed a 'cloud' drop because it falls too slowly to produce measurable precipitation and evaporates almost immediately after exiting the bottom of the cloud. Also, the farther an observer is from falling precipitation, the more it appears to be a 'cloud' due to perspective. For example, many of the higher 'clouds' seen above, such as cirrus and altostratus

clouds, are composed mainly of ice crystals and even snowflakes that are settling toward the earth; they would not be considered a 'cloud' by an observer inside them on Mount Everest, for example, but rather a very light snowfall.

Origin of the Present-Day Cloud Classification System

The classification system for clouds is based on what was seen above. The process of classifying objectively the many shapes and sizes of something as ephemeral as a cloud was first accomplished at about the same time at the turn of the nineteenth century by an English chemist, [Luke Howard in 1803](#) and a French naturalist, Jean Baptiste Lamarck in 1802. Both published systems of cloud classifications. However, because Howard used Latin descriptors of the type that scientists were already using in other fields, his descriptions appeared to resemble much of what people saw, as did Lamarck's, and because he published his results in a relatively well-read journal, *Tilloch's Philosophical Magazine*, Howard's system became accepted and was reproduced in books and encyclopedias soon afterward.

Howard observed, as had Lamarck before him, that there were three basic cloud regimes. There were fibrous and wispy clouds that Howard called 'cirrus' (Latin for hair); sheet-like laminar clouds that covered much or all of the sky that he referred to as 'stratus' (meaning flat); and clouds that were less pervasive but had a strong vertical architecture that he called 'cumulus' (meaning 'heaped up'). Howard used an additional Latin term 'nimbus' (Latin for cloud) meaning in this case, a cloud or system of clouds from which precipitation fell. Today, nimbus itself is not a cloud but rather a prefix or suffix to denote the two main precipitating clouds, nimbostratus and cumulonimbus. The question over clouds and their types generated such enthusiasm among naturalists in the nineteenth century that an ardent observer and member of the British Royal Meteorological Society, Ralph Abercromby, took two voyages around the world to make sure that no cloud type had been overlooked!

The emerging idea that clouds preferred just two or three levels in the atmosphere was supported by measurements using theodolites and photogrammetry to measure cloud height at Uppsala, Sweden as well as at sites in Germany and in the United States in the 1880s. These measurements eventually led H. Hildebrandsson, Director of the Uppsala Observatory, and Abercromby to place the 'low,' 'middle,' and 'high' cloud groupings of Howard more systematically in their

own 1887 cloud classification. At this time, cumulus and cumulonimbus clouds were placed in a fourth distinct category representing clouds with appreciable updrafts and vertical development.

Howard's modified classification system was reexamined at the International Meteorological Conference at Munich in 1891 followed by the publication of a color cloud atlas in 1896. At this point, the definitions of clouds were close to their modern forms. Additional international committees made minor modifications to this system in 1926 that were realized with the publication of the 1932 International Cloud Atlas. Little change has been made since that time. Perhaps the most noticeable change was to move thick layers of altocumulus (opacus), a cloud comprised of mostly or completely of droplets, out of the altostratus domain and into the altocumulus domain. Altostratus was then confined to clouds mostly or completely comprised of ice, but unlike cirriform clouds are generally thick enough to produce shading. Altostratus, like cirrostratus, usually covers most of the sky if not all of it.

The most comprehensive version of the cloud classification system was published by the WMO in 1956 in two volumes entitled 'The International Cloud Atlas, Volumes I and II.' Volume I contained the cloud morphology, while Volume II consisted of photographs. An abridged *International Cloud Atlas* was published in 1969. This Atlas combined morphology and photographs. The cloud morphology volume (Volume I) was published again in 1975 by the WMO. In 1987, a revised Volume II of photographs was published that included photographs of clouds from more disparate places other than in the previous volumes. However, there was some lack of consistency in classifying clouds compared with the earlier Volume II. Care should be taken when using the 1987 revised version.

The Classification of Clouds

There are 10 main categories or 'genera' into which clouds are classified for official observations. These 10 categories are cirrus, cirrostratus, cirrocumulus, altostratus, altocumulus, nimbostratus, stratocumulus, stratus, cumulus, and cumulonimbus. **Table 1** is a partial list of the nomenclature used to describe the most commonly seen species (defined by overall shapes) and varieties (defined by arrangements of subelements of the clouds) of these genera. **Figures 1–25** illustrate these most frequently seen forms.

Within these 10 categories are three cloud base altitude regimes. They are 'high' clouds, those with bases generally above 7 km above ground level (AGL); 'middle-level' clouds, those with bases between 2 and about 7 km AGL; and 'low' clouds, those with bases at or below 2 km AGL. The word 'about' is used because clouds with certain visual attributes that make them, for example, a 'middle-level' cloud may actually have a base that is above 7 km. Similarly, in winter-time or in the arctic, 'high' clouds with cirriform attributes (fibrous and wispy) may be found at heights below 7 km. Also, some clouds that are still considered low clouds (e.g., cumulus clouds) can have bases that are a kilometer or more above the general 'low cloud' upper base limit of 2 km AGL. Therefore, these cloud base height boundaries should be

Table 1 The 10 cloud types and their most common species and varieties. After the World Meteorological Organization, 1975. The letters in parentheses denote accepted abbreviations

Genera	Species	Varieties
Cirrus (Ci)	Uncinus, fibratus, spissatus, castellanus	Intortus, radiatus, vertebratus
Cirrostratus (Cs)	Nebulosus, fibratus	
Cirrocumulus (Cc)	Castellanus, floccus, lenticularis	Undulatus
Alto cumulus (Ac)	Castellanus, floccus, lenticularis	Translucidus, opacus, undulatus, perlucidus
Altostratus (As)	None	Translucidus, opacus
Nimbostratus (Ns)	None	None
Stratocumulus (Sc)	Castellanus, lenticularis	Perlucidus, translucidus opacus
Stratus (St)	Fractus, nebulosus	
Cumulonimbus (Cb)	Calvus, capillatus	
Cumulus (Cu)	Fractus, humilis, mediocris, congestus	

considered somewhat flexible. Note, too, that what is classified as an altocumulus layer when seen from sea level, will be termed a stratocumulus layer when the same cloud is seen by an observer at the top of a high mountain because the apparent size of the cloud elements, part of the definition of these clouds, becomes larger, the nearer one is to the cloud layer.

The classification of clouds is also dependent on their composition. This is because the composition of a cloud, all liquid, all ice, or a mixture of both, determines many of its visual attributes on which the classifications are founded (e.g., luminance, texture, color, opacity, and the level of detail of the cloud elements). For example, an altocumulus cloud cannot contain too many ice crystals and still be recognizable as an altocumulus cloud. It must always be composed largely of water drops to retain its sharp-edged compact appearance. Thus, it cannot be too high and cold. On the other hand, fibrous trails of ice crystals comprising cirrus clouds cannot be too low (and thus, too warm). Therefore, having the ability to assess the composition of clouds (i.e., ice vs. liquid water) visually can help in the determination of a cloud's height.

Other important attributes for identifying a cloud are How much of the sky does it cover? Does it obscure the sun's disk? If the sun's position is visible, is its disk sharply defined or just a bright spot? Does the cloud display a particular pattern such as small cloud elements, rows, billows, or undulations? Is rain or snow falling from it? If so, is the rain or snow falling from it concentrated in a narrow shaft, suggesting towering cloud tops above, or is the precipitation widespread with little gradation, a characteristic that suggests uniform cloud tops? Answering these questions will allow the best categorization of clouds into their 10 basic types.

High Clouds

Cirrus, cirrostratus, and cirrocumulus clouds (**Figures 1–5**, respectively) comprise 'high' clouds. By WMO definition, they



Figure 1 Cirrus uncinus.



Figure 2 Cirrus spissatus.

are not dense enough to produce shading except when the sun is near the horizon, with the single exception of a thick patchy cirrus species called cirrus spissatus ([Figure 2](#)) in which gray shading is allowable. (Many users of satellite data refer to

'cirrus' or 'cirriform' those clouds with cold tops in the upper troposphere without regard to whether they produce shading as seen from below. However, many such clouds so described would actually be classified as 'altostratus' clouds by ground



Figure 3 Cirrostratus nebulosus (smooth regions); Cirrostratus fibratus (region with lines).



Figure 4 Cirrocumulus (lower warmer version, right third of photo above and to the right of the palm tree, droplet composition).

observers due to the gray shading they produced.) Cirrus (Figures 1 and 2) and cirrostratus (Figure 3) clouds are composed of ice crystals with, perhaps, a few momentary exceptions at formation when the temperature is higher than

−40 °C. In these cases, droplets may be briefly present at the instant of formation. The ‘bases’ or visual bottoms of cirrus and cirrostratus clouds are composed of generally low concentrations of ice crystals that are about to evaporate. They are usually



Figure 5 Cirrocumulus (high and very cold; composed of ice soon after formation, as has occurred above and to the right of the antenna).

colder than -20°C . The coldest cirriform cloud tops (i.e., cirrus and cirrostratus) can be -80°C or lower in deep storms with high cloud tops such as in anvils associated with exceptional thunderstorms.

Cirrus and cirrostratus clouds are fibrous, wispy, and diffuse because the concentrations of ice crystals that comprise them are relatively low (from less than 1 per liter to tens per liter) compared with particle concentrations in other clouds. An exception to this is at the moment of formation when a spec or small, hard looking tuft of cirrus can have many thousands per liter of tiny quasispherical ice crystals which then gradually disperse after the moment of formation.

The long, usually curved filaments that often comprise cirriform clouds are caused by the growth of larger ice crystals that fall out into regions of changing wind speeds and directions below the parent cloud. Due to the slow settling of ice crystals, and depth of moist air below the formation level, mature cirrus and cirrostratus clouds are often 1 km or more thick though the sun may not be appreciably dimmed.

Haloed are usually seen with cirrostratus clouds, and occasionally, partial haloed are seen with cirrus clouds. Cirrus

clouds do not usually produce full haloed due to their patchy nature. In contrast, thicker ice clouds, such as altostratus (Figures 11 and 12), cannot produce haloed when seen from the ground. This is because while cirrus and cirrostratus clouds usually contain small, hexagonal ‘prism’ crystals such as thick plates, short solid columns – simple crystals that refract the sun’s light as it passes through them, deeper altostratus clouds generally have larger more complicated crystals and snowflakes that do not permit simple refraction even when the sun’s position is plainly evident.

The appearance of cirrostratus clouds in wintertime in the middle and northern latitudes, with its typical halo, has long been identified as a precursor to steady rain or snow.

Cirrocumulus (Figure 4 shown with altocumulus for comparison) clouds are patchy, finely granulated clouds. The largest of the visible cloud elements in cirrocumulus can be no larger than the width of a finger held skyward when observed from the ground; if larger, the cloud is classified as an altocumulus. Due to a definition that allows no shading, cirrocumulus clouds are very thin (less than 200 m thick), and usually very short-lived, often appearing and disappearing in minutes.



Figure 6 Altocumulus opacus.



Figure 7 Altocumulus perlucidus.

Cirrocumulus clouds, contrary to many descriptions of them found elsewhere, are generally composed mostly of or completely of water droplets, not ice crystals. This is because they generally occur at higher temperatures and lower

altitudes than cirrus and cirrostratus clouds. The fine granulation of cirrocumulus in isolation in the sky often leads to a misperception of much greater height of that cloud than where they are actually located. In fact, they are often



Figure 8 Altocumulus castellanus.



Figure 9 Altocumulus floccus virgae.

located (as shown in [Figure 4](#)) in what otherwise would be termed the 'midlevels' rather than a cirrus or 'high levels' as often thought, and altocumulus clouds are even present at the same level!

The liquid phase of cirrocumulus clouds can usually be deduced by the extremely sharp edges of the individual elements with no sign of fallout from them, and when they are near the sun, a corona or irisation (also called 'iridescence') is



Figure 10 Altostratus lenticularis.



Figure 11 Altostratus translucidus.

produced due to the diffraction of sunlight by the cloud's tiny ($<10\text{ }\mu\text{m}$ diameter) droplets.

However, there are many cirrocumulus clouds that do form at low temperatures ($<-30\text{ }^{\circ}\text{C}$) and altitudes where cirrus and

cirrostratus form in the high troposphere (Figure 5). These quickly evolve into fibrous cirriform masses within a few minutes, which destroy the granulated appearance required for the label for a cirrocumulus cloud to be applied (see adjacent



Figure 12 Altostratus opacus. It is not raining but close.



Figure 13 Nimbostratus. It is raining lightly.

regions to the right of the cirrocumulus cloud in [Figure 5](#) where the granulation has been replaced by fibrous elements having miniature fall streaks). Therefore, cirrocumulus clouds that transit from liquid to ice become forms of cirrus clouds.

Middle-Level Clouds

Altostratus, altostratus, and nimbostratus clouds ([Figures 6–13](#), respectively) are considered ‘middle-level’ clouds



Figure 14 Stratocumulus.



Figure 15 Stratus.

because their bases are located between about 2 and 7 km AGL (see discussion concerning the highly variable base of what would be called, nimbostratus clouds, aka, 'rain clouds,' below).

All of these clouds are the product of slow updrafts (centimeters per second) often taking place in the middle troposphere over an area of thousands of square kilometers or more. Gray shading is expected in altostratus and is generally



Figure 16 Cumulus humilis.

present in altocumulus clouds. Nimbostratus clouds by definition are dark gray and the sun's location cannot be detected. It is this property of shading that immediately differentiates these clouds from 'high' clouds, ones that with the single exception of cirrus spissatus, have no shading.

Altostratus and altocumulus are different from one another in the same way that cirrus and cirrostratus clouds are different from cirrocumulus clouds. With altocumulus clouds, droplets predominate and that give them a crisp, sharper-edged look as it does with cirrocumulus. In altostratus clouds, ice crystals and snowflakes dominate or comprise the entire cloud giving it a diffuse, fibrous look.

Altocumulus clouds are distinguished from cirrocumulus because they are generally lower and their cloud elements are several times larger than those in cirrocumulus clouds. For example, the elements of an altocumulus cloud are typically the width of three fingers held skyward from the ground. Too, shading toward the center of the thicker elements is usually present in altocumulus clouds, a property that is not allowed in the classification of cirrocumulus clouds.

Altocumulus clouds are distinguished from stratocumulus because they are higher above ground level than stratocumulus (at least 2 km) and because the individual cloud elements in altocumulus are, or appear to be from the ground, smaller than those in stratocumulus.

In spite of its name, altocumulus clouds (Figures 6 and 7, examples of opacus and translucidus, respectively) are overall rather flat clouds that strongly resemble a higher layer of stratocumulus clouds. An exception to this overall laminar architecture is in those species of altocumulus called castellanus (Figure 8) and floccus (Figure 9). In these forms,

altocumulus clouds resemble miniature, lofted cumulus clouds; they truly are 'altocumulus' in the full sense of the name. These forms also usually occur in rows or patches rather than in widespread layers.

In spite of the gray shading that may be present in altocumulus clouds, they are rarely more than 1 km thick. This gray shading is because the concentrations of drops in them are relatively high (typically 50 to several hundreds of thousands per liter) compared with fibrous ice clouds whose particle concentrations may only be several to a few hundreds per liter and are relatively transparent though usually deeper. The concentration of droplets in altocumulus clouds is usually sufficient to produce an 'optical depth' of 4 or more in which the sun's disk is obscured.

Altocumulus clouds sometimes sport patchy 'virga' (Figure 9). Virga is light precipitation that falls from a cloud but does not reach the ground due to evaporation. Because virga is almost always due to falling snow, it appears fibrous, often with striations or long filaments that often far surpass the depth of the cloud from which it is falling (Figure 9). Altocumulus clouds with virga are predominantly those clouds whose temperatures are lower than -10°C . However, at the same time, they are rarely colder than about -30°C . This is because at very low temperatures, they are likely to take on the attributes of ice clouds such as cirrus or its thicker icy brethren, altostratus and as such would no longer be classified as altocumulus.

The species of altocumulus clouds called altocumulus castellanus and floccus (Figures 8 and 9) have always had a special significance in meteorology because these clouds reveal a 'conditionally' unstable lapse rate in the middle



Figure 17 Cumulus humilis shedding ice (areas of where a thin veil-like cloud is present). These clouds are almost certainly colder than -15°C .

troposphere. Conditional instability has been viewed as an indicator for likely releases of deeper convection in the hours ahead such as in thundery weather. Occasionally, altocumulus castellanus clouds themselves group and deepen upward into high-based cumulus and cumulonimbus clouds with significant rain.

When the winds are relatively strong aloft (greater than about 20 m s^{-1}) and moderately moist, but stable lapse rate conditions are present, a species of altocumulus called 'lenticularis' (lens or almond shaped) clouds may form over or downwind of mountains (Figure 10). Altocumulus lenticularis clouds can hover over the same location for minutes to hours while expanding and shrinking in response to fluctuations in the relative humidity of the air mass being lifted over the terrain. Because the conditions under which these clouds form are most often associated with advancing short wave troughs in the middle and upper atmosphere and their accompanying regions of low pressure, lenticularis clouds are usually precursors to deteriorating weather.

With altostratus clouds (Figures 11 and 12), the dominance of ice causes a diffuse, amorphous gray, dull appearance with striations or falls streaks (virga) at the bottom. What an observer

at the ground is viewing at the bottom of altostratus are relatively low concentrations of ice crystals and snowflakes rather than a 'cloud' per se. Those particles at the bottom are evaporating before reaching the ground.

Altostratus clouds are also thick rarely less than 2 km thick and often have tops at the same heights as cirrus and cirrostratus clouds. Because of this great altitude range, they are considerably colder and span a much greater temperature range than do altocumulus clouds. By definition, they cover the entire sky or at least wide portions of it; they are not patchy clouds!

Precipitation is usually imminent when altostratus clouds are moving in because they are a sign that a widespread slab of air, several kilometers deep is rising above you, usually that due to an approaching cyclone and its frontal system. An exception to this rule is in desert regions where the altostratus may be the only cloud that passes over even if a low pressure area and a front moves through.

The relatively low concentrations of large particles in some altostratus clouds (often tens per liter) can allow the sun's position to be seen as though looking through 'ground' or 'fogged' glass. That is, the sun's position is apparent, but the outline of its disk is not. The thin variety of altostratus is termed



Figure 18 Cumulus mediocris.



Figure 19 Cumulus congestus.

altostratus translucidus (Figure 11). In spite of being able to see where the sun is, they are rarely less than 2 or more kilometers thick. Figure 12 shows altostratus opacus, dense, and dull clouds announcing that rain is on the doorstep.

Altostratus translucidus to opacus is a normal progression prior to the onset of precipitation.

When the tops of an altostratus clouds are warmer than about -30°C , the top is, surprisingly, often composed of



Figure 20 Cumulonimbus calvus.



Figure 21 Cumulonimbus capillatus.

a thin droplet cloud virtually identical to an altocumulus cloud layer, but one that produces the initial ice that is found below it. The growth, aggregation, and fragmentation of ice crystals spawned by the topmost liquid clouds over a great

depth usually prevent ground observers from detecting the altocumulus-like cloud at the top. These kinds of situations were dubbed 'the upside-down storm' when first noticed in the mid-1950s because the coldest part of the cloud (the top)



Figure 22 Cumulonimbus capillatus, but one that is entirely above freezing in temperature but exhibits a fibrous structure similar to ‘cold-topped’ cumulonimbus capillatus clouds.

was liquid and the warmer regions below were comprised of ice.

Optical phenomena seen from the ground with altostratus clouds are limited to parhelia (‘sun dogs’ or ‘mock suns’). These are only observed in the variety, altostratus translucidus and normally when the sun is low in the sky. Parhelia are bright, colored highlights that sometimes rival the brightness of the sun that are located 22° from the sun’s position.

Since the composition of the uppermost regions of the deepest altostratus clouds are virtually identical to cirriform ice clouds having simpler, smaller ice crystals, haloes are often observed in the uppermost portions of deep altostratus clouds, for example, as an aircraft ascends through it.

Nimbostratus clouds (Figure 13) are virtually identical to altostratus clouds in their composition except that their bases are usually perceived from the ground as lower than in altostratus, the layer from which it has usually derived due to a downward thickening. In addition, steady precipitation is falling from it. Therefore, they often appear somewhat darker than altostratus clouds and, by definition, do not allow the sun to be seen through them. The perceived base of nimbostratus is found in the melting level where snowflakes are melting into raindrops. This apparent

‘base’ of the cloud is because the greater opacity of snow particles gives the impression of a ‘bottom’ or sharp increase in thickness of the cloud. Therefore, the so-called ‘base’ of a nimbostratus cloud might be perceived at ‘midlevels’ on a day when the freezing level is high (>2 km) such as in southern latitudes or the tropics or be perceived as low when the freezing level is low as in northern latitudes in the winter.

Nimbostratus clouds produce relatively steady precipitation that often continues for hours at a time. They are not clouds responsible for passing showers with periods of sun in between. The tops of dark, steadily precipitating nimbostratus clouds can be as shallow as 2–3 km and even be above freezing in temperature or they may reach into the upper regions of the troposphere (to cirriform cloud levels) and be as cold as -80°C . At altitudes above the freezing level, nimbostratus is largely composed of ice crystals and snowflakes, though embedded thin supercooled droplet cloud layers similar to altocumulus clouds are relatively common. Also, similar to altostratus clouds, when the temperature at the top of nimbostratus clouds is higher than about -30 to -35°C , a thin droplet cloud layer may be found in which the ice crystals form and settle out.



Figure 23 Cumulonimbus capillatus incus, the granddaddy of them all.



Figure 24 Rainshaft from a cumulonimbus.

Though it is not present in [Figure 13](#) nimbostratus from which light rain was falling, a broken-to-overcast layer of shallow stratocumulus ([Figure 14](#)) or stratus ([Figure 15](#)) clouds is usually present at the bottom of them. While usually

not precipitating themselves, these lower cloud layers are important in enhancing the amount of rain or snow that falls from nimbostratus clouds. This enhancement occurs because of the collection of cloud drops in those lower clouds by the



Figure 25 Cumulonimbus mammatus.

precipitation that falls through them in the following way: imagine a cup of water being dumped in the top of a stratocumulus cloud ([Figure 14](#)); more than a cup would fall out at the very bottom of that cloud – if you could catch it all! This kind of precipitation increase is especially evident in hilly or mountainous regions compared with flat regions upwind where there may not be lower clouds.

However, even the existence of lower clouds with drops too small to be collected by precipitation can result in more precipitation than areas without any clouds. Why? This is because where there are no lower clouds, the precipitation is likely to be subject to a bit of evaporation, and the drops or snowflakes are then slightly smaller in comparison to those locations where clouds exist and no evaporation occurs while they are falling through the cloud. Again, imagine dropping a cup of water at the top of a small cloud with tiny droplets, droplets too small to be collected. At the very bottom, you could potentially retrieve your exact cup of water since there was no evaporation inside the cloud. In an area devoid of clouds, even though it is raining, dropping that same amount of water over the same depth as before would lead you to collect less than the full cup you started with. Even though the relative humidity be high, it would be that bit less than 100%, the value required to exactly preserve your cup of water. So, where there are lower clouds, the rain is heavier than without them.

Cumulonimbus clouds (see [Convective Clouds](#)) may also be embedded in nimbostratus clouds. The presence of such clouds within nimbostratus is evident by sudden gushes of much heavier rain and sometimes lightning within a context of relatively steady rain.

Low Stratiform Clouds

Stratocumulus and stratus clouds ([Figures 14 and 15](#), respectively) are low-based (below 2 km AGL), shallow ‘stratiform’ clouds. They are almost always less than 1 km thick. They are composed of droplets unless the cloud top is cooler than about -5 to -10 °C in which case ice crystals may form. Stratocumulus clouds ([Figure 14](#)) differ from stratus clouds ([Figure 15](#)) because they have an obvious rather lumpy appearance at cloud base with darker and lighter regions due to embedded weak convection. These changes in shading represent variations in the liquid water content of the clouds, with the darker regions representing higher amounts of liquid water. Also, the bases of stratocumulus clouds tend to be higher and more irregular in height than those of stratus clouds. Stratus clouds present a smoother, lower, more uniform sky than does stratocumulus clouds because the internal convective overturning that produces lighter and darker regions in stratocumulus is nil in these clouds.

Drizzle precipitation (defined as ‘fine,’ less than $500\text{ }\mu\text{m}$ but greater than $100\text{-}\mu\text{m}$ diameter drops that are also close together and nearly float in the air) often falls from these clouds when the cloud droplet concentrations are lower than about 100 cm^{-3} . In these cases, a broad droplet spectrum that is one to where the cloud droplets have reached beyond about $30\text{ }\mu\text{m}$ in diameter are frequently present in the upper portions of such clouds. When cloud droplets attain sizes larger than this, they begin to stick together when they bump into each other, producing a much larger drop, one that can fall out as a drizzle drop. Due to the requirement for ‘clean’ conditions, drizzle is common from both stratus and

stratocumulus clouds in regions with few CCN such as at sea and especially along and slightly inland of western coastlines in periods of onshore flow.

However, recent measurements have also shown that drizzle and light rain can also develop in these shallow clouds in surprisingly far inland locations in which occasionally very clean conditions are observed such as in the interior of the United States in winter. Since these latter clouds are often supercooled in wintertime and affect populated areas with airports, they pose a severe potential for aircraft icing when freezing rain or drizzle forms in them.

Convective Clouds

Cumulus and cumulonimbus clouds (Figures 16–25, respectively) illustrate convective clouds, those clouds brought about when the temperature decreases rather rapidly with increasing height above the ground. Differential heating and converging air currents in this vertical temperature structure can therefore send plumes of warmer air skyward with relative ease since those plumes will likely be warmer than the air around them.

Convective clouds are limited in coverage compared with stratiform clouds and, except for the anvil portions of cumulonimbus clouds, rarely cover the entire sky or do so only for short periods. This coverage characteristic differentiates cumulus clouds, for example, from stratocumulus clouds because the latter have linked cloud bases covering large portions of the sky. Similar-sized cumulus clouds must, by definition be relatively scattered into isolated clouds or small clusters with large sky openings.

Cumulus clouds are roughly divided into species by their depths. For example, cumulus fractus, those first cloud shreds that appear at the top of the convective boundary layer, may be less than 100 m thick. Cumulus humilis (Figure 16), the next larger size, should not be more than 1 km thick and looks more like a fat pancake than a heaped up cloud. Figure 17 shows that in very cold situations, that even cumulus humilis clouds can form ice. Cumulus mediocris clouds (Figure 18) show clouds that are beginning to be humped up and resemble the profiles presented by mountain ranges. They are around 1–2 km thick. The largest cumulus species is cumulus congestus (Figure 19), always more than about 2 km deep to several kilometers deep and generally much taller than they are wide.

The tops of these larger cumulus clouds, mediocris and congestus, are marked by sprouting subelements referred to as turrets that appear as noticeable protuberances at the top. Turrets are also crenelated on their surface with dozens of lesser, tuft-like ‘units’ perhaps tens of meters wide. Turrets are generally one to a few kilometers wide. However, in severe thunderstorms, individual turrets may coalesce into groups of many turrets to form a large, tightly packed, and hard-appearing cauliflower mass that roils upward with little turret differentiation.

Cumulus clouds are, with rare exceptions discussed below, composed solely of droplets. They have the highest liquid water concentrations of any clouds in their upper portions where the moist air has been lifted the highest. To be a purely cumulus cloud, very few precipitation-sized particles are in them, though they may be imminent. There is no definite rainshaft,

an appendage requiring the use of the modifier, ‘nimbus,’ Latin for rain.

The development of extensive precipitation in cumulus clouds is one in which a cumulus congestus is also becoming a cumulonimbus cloud and leaving the cumulus category. The depth of this transition is different for different aerosol regimes. In clean conditions, cumulus congestus can migrate to a cumulonimbus having a pronounced rainshaft when they reach depths of only 1.5–3 km thick, such as over the oceans. However, in polluted situations, the depth must be much greater, about 3 km thick over land. The precipitation that falls from cumulonimbus clouds can be either due to collisions with coalescence of cloud drops to form raindrops (a process termed the ‘warm rain process’) or it may be due to the formation of ice particles which then collect water through riming or through the aggregation of many ice crystals into snowflakes that melt on the way down. In the cases where the warm rain process is the sole producer of raindrops, the cumulonimbus clouds are necessarily shallower than those requiring an ice process for a strong rainshaft since their tops will be only around or below the freezing level altitude.

Often times in maritime locations and also for clouds in continental locales with warm bases (above about 10 °C), both processes, the warm rain and the ice process, are active.

In wintertime, even small to moderate cumulus clouds with tops colder than about –10 to –15 °C can produce virga, snow flurries, or even accumulating amounts of snow (Figure 17). These kinds of small, cold-based, and precipitating cumulus clouds are found in wintertime in such locations as the Great Lakes of the United States, off the east coasts of the continents, over high mountains or deserts regions.

When significant precipitation develops in cumulus congestus clouds, the visual attributes begin to change noticeably. In the first stage of this change, often very subtle and hard to detect without a practiced eye, the cloud is called a cumulonimbus calvus (‘bald,’ Figure 20). Often, a strong precipitation shaft is seen below cloud base with a cloud top that has softened from a hard, crenelated appearance. In some cases, the rainshaft has not yet appeared or is just emerging, an exciting moment! The ‘soft,’ fibrous, fraying, or wispy transition, sometimes compared to the look of ‘cotton candy,’ is due to the lowering of the concentrations of the particles from 50 to hundreds of thousands per liter of small cloud droplets (<50 µm diameter), to only tens to hundreds per liter of much larger diameter particles that are fractions of a millimeter to greater than millimeter sized. These precipitation particles can be raindrops or ice particles or briefly, both. These larger particles tend to fall in filaments and often produce a striated appearance.

This process from the ‘hard’ to ‘soft’ appearance of a cloud top takes just a few minutes, typically around five or so, that is, to the point that most observers can recognize that ‘something has changed’ at cloud top from the time the cloud was a congestus. If there is already a strong shaft, it is likely that you are viewing the upwind side of a cumulonimbus cloud where new turrets are forming and are going through the complete glaciation cycle and the fibrous appearance which was normally expected to see with fully developed rainshafts is hidden from view and downwind from the observer.

When the fibrousness of the upper portion of the cloud is readily apparent, the cumulonimbus cloud has transitioned from a 'calvus' to a capillatus (Latin, 'hair' – grew 'hair' after being 'bald!'). At this point, when the capillatus stage is reached, all of the upper portion of the cumulonimbus clearly consists of ice crystals and snowflakes (Figure 21). In the tropics or in warm humid air masses, this visual transformation also occurs but can be due solely to the evaporation of the smaller drops leaving the much lower concentrations of drizzle and raindrops that result in a softening of the clouds appearance (Figure 22).

Hail or graupel (small, soft hail) are usually found, if not at the ground, then aloft in virtually all cumulonimbus clouds that reach above the freezing level.

If a pronounced flattening of the top of these clouds develops into a spreading anvil, then the cloud has achieved the status of a cumulonimbus capillatus incus (incus, Latin for 'anvil,' Figure 23). With the familiar anvil, it is perhaps the most recognizable form of a cumulonimbus cloud. The flattening at top usually indicates that the updraft has reached the tropopause, and therefore, these cumulonimbus clouds are more likely than 'capillatus' versions to be severe storms. And watch out if a mound or towering dome of cloud appears above the flattened top! That dome represents an updraft that has overshot the troposphere and entered the stratosphere, the sign of an exceptionally strong updraft within that thunderstorm. So, a dome above the anvil is a very good sign of an especially severe thunderstorm, one that you would not want to drive your car under.

Updrafts may reach tens of meters per second in cumulus and cumulonimbus clouds, particularly in warm air masses. The greatest updraft speed that has been measured by an aircraft was an astounding 40 m s^{-1} ! These strong updrafts lead to large amounts of condensation and liquid water content in the upper regions of these clouds, and often at temperatures far below freezing, even to -30°C !

Depending on how warm cloud base is, the middle and upper developing portions of deep cumulus clouds might contain $1\text{--}5 \text{ g m}^{-3}$ of condensed water in the form of cloud droplets and raindrops. Supercooled water concentrations of these magnitudes are sufficient to cause a buildup of ice on an airframe of about 1 cm or more of for every 1–2 min in cloud. No aircraft can fly for long with such an accumulation of ice and therefore, cumulus and cumulonimbus clouds are avoided by aircraft.

Cumulonimbus clouds are the only clouds that produce lightning. If lightning is observed, the cloud type producing, it is therefore designated as a cumulonimbus. Unlike other clouds, too, the bases of cumulonimbus clouds are marked by strong rainshafts, a feature that differentiates them from nimbostratus clouds. Therefore, when strong rainshafts are observed (Figure 24), the cloud producing, it is also assumed to be a cumulonimbus even if lightning is not observed.

Occasionally, too, large downward projecting protuberances are seen in connection with cumulonimbus clouds. These formations are called 'mammatus,' Figure 25. Contrary to popular opinion, mammatus formations may or may not be associated with a strong thunderstorm. Mammatus are also seen in such tame clouds as altostratus and cirrus.

Cumulonimbus clouds span a wide range of depths, from miniature versions only about 2 km deep in polar air masses over the oceans, ones that never produce lightning, to as much as 20 km in the most severe thunderstorms in equatorial regions, the plains of eastern China, in Brazil and Argentina, and the plains and southeast regions of the United States. However, the most lightning on earth occurs in the interior of central Africa along the Equator.

See also: **Aerosols:** Aerosol–Cloud Interactions and Their Radiative Forcing. **Aviation Meteorology:** Aircraft Icing. **Clouds and Fog:** Climatology; Cloud Microphysics; Stratus and Stratocumulus. **Electricity in the Atmosphere:** Lightning. **Mesoscale Meteorology:** Cloud and Precipitation Bands; Convective Storms: Overview. **Numerical Models:** Parameterization of Physical Processes: Clouds. **Radar:** Cloud Radar. **Radiation Transfer in the Atmosphere:** Cloud–Radiative Processes. **Synoptic Meteorology:** Lake Effect Storms. **Thermodynamics:** Saturated Adiabatic Processes. **Tropospheric Chemistry and Composition:** Aerosols/Particles.

Further Reading

- British Meteorological Office, 1982. *Cloud Types for Observers*. Her Majesty's Stationery Office, London. Met. 0.716.
- Brooks, C.F., 1951. *The Use of Clouds in Forecasting*. Compendium of Meteorology. American Meteorological Society, Boston.
- Clause, R., Facy, L., 1961. *The Clouds*. Evergreen Profile Books, New York.
- Cunningham, R.M., 1957. *A Discussion of Generating Cell Observations with Respect to the Existence of Freezing or Sublimation Nuclei*. Artificial Stimulation of Rain. Pergamon Press, New York.
- Day, J., 2003. *The Book of Clouds*. Silver Lining Books, New York.
- Grant, H.D., 1944. *Cloud and Weather Atlas*. George G. Harrap and Company, Ltd., London.
- Hamblyn, R., 2001. *The Invention of Clouds*. Farrar, Straus and Giroux, New York.
- Heymsfield, A.J., 1993. *Microphysical Structures of Stratiform and Cirrus Clouds*. Aerosol–Cloud–Climate Interactions. Academic Press, New York.
- Hildebrandsson, H., Riggenbach, A., et Teisserenc de Bort, L., 1896. *Atlas international des nuages*. Comité Météorologique International.
- Hobbs, P.V., Rangno, A.L., 1985. Ice particle concentrations in clouds. *Journal of Atmospheric Sciences*, 42, 2523–2549.
- Houze, R.A. Jr., 1993. *Cloud Dynamics*. Academic Press, New York.
- Howard, L., 1803. On the modifications of clouds, and on the principles of their production, suspension, and destruction. *Philosophical Magazine*, 16, Royal Society of England, London.
- Howell, W.E., 1951. *The Classification of Cloud Forms*. Compendium of Meteorology. American Meteorological Society, Boston.
- Ludlum, D.M., 1991. *The Audubon Society Field Guide to North American Weather*. Alfred A. Knopf, New York.
- Ludlum, F.H., 1980. *Clouds and Storms: The Behavior and Effect of Water in the Atmosphere*. University Park, The Pennsylvania State University Press, Pennsylvania.
- Plank, V.G., Atlas, D., Paulsen, W.H., 1955. The nature and detectability of clouds and precipitation by 1.25 cm radar. *Journal of Meteorology*, 12, 358–378.
- Pretor-Pinney, G., 2011. *The Cloud Collector's Handbook*. Chronicle Books, San Francisco.
- Rangno, A.L., Hobbs, P.V., 1994. Ice particle concentrations and precipitation development in small continental cumuliform clouds. *Quarterly Journal of the Royal Meteorological Society*, 120, 573–601.
- Rasmussen, R.M., Geresdi, I., Thompson, G., Manning, K., Karplus, E., 2005. Freezing drizzle formation in stably stratified layer clouds: the role of radiative cooling of cloud droplets, cloud condensation nuclei, and ice initiation. *Journal of Atmospheric Sciences*, 59, 837–860.
- Rauber, R.M., Tokay, A., 1991. An explanation for the existence of supercooled liquid water at the top of cold clouds. *Journal of Atmospheric Sciences*, 48, 1005–1023.
- Scorer, R.S., 1972. *Clouds of the World*. Stackpole Books, Harrisburg.

- Scorer, R.S., Verkaik, A., 1989. *Spacious Skies*. David & Charles, Brunel House.
- Wallace, J.M., Hobbs, P.V., 2006. *Atmospheric Science: An Introductory Survey*, second ed. Academic Press, New York.
- Warren, S.G., Hahn, C.J., London, J., 1991. *Analysis of Cloud Information from Surface Weather Reports*. World Meteorological Organization, Geneva. World Climate Research Programme Report on Clouds, Radiative Transfer, and the Hydrological Cycle.
- World Meteorological Organization, 1956. *International Cloud Atlas (Complete Atlas)*, vol. I. World Meteorological Organization, Geneva.
- World Meteorological Organization, 1956. *International Cloud Atlas (Complete Atlas)*, vol. II. World Meteorological Organization, Geneva.
- World Meteorological Organization, 1969. *International Cloud Atlas (Abridged Atlas)*. World Meteorological Organization, Geneva.
- World Meteorological Organization, 1975. *International Cloud Atlas*. In: *Manual on the Observations of Clouds and Other Meteors*, vol. I. World Meteorological Organization, Geneva.
- World Meteorological Organization, 1987. *International Cloud Atlas*, vol. II. World Meteorological Organization, Geneva.
- Zeng, Z., Yuter, S.E., Houze, R.A., Kingsmill, D.E., 2001. Microphysics of the rapid development of heavy convective precipitation. *Monthly Weather Review*, 129, 1882–1904.

Climatology

S Warren and R Eastman, University of Washington, Seattle, WA, USA
CJ Hahn, University of Arizona, Tucson, AZ, USA

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

A cloud climatology describes the time-averaged geographical distribution of cloud properties and their diurnal, seasonal, and interannual variations. Visible and infrared radiation measurements from space are used to obtain cloud top height, cloud optical thickness, and droplet sizes. Visual reports from surface observers on land and sea are used to obtain the frequency of occurrence and the amount of sky covered by each of the various cloud types such as cumulus, stratus, and cirrus. This article summarizes results from both analyses, emphasizing the half-century period of record available from the surface-based climatology.

Introduction

Clouds are an important component of the Earth's climate system. They reflect solar radiation back to space, they absorb thermal infrared radiation emitted from below, and they produce rain and snow. A cloud climatology describes the time-averaged geographical distribution of cloud properties and the diurnal, seasonal, and interannual variations of these properties.

Cloud climatologies are used to determine the radiative effects of clouds on climate and to determine the extent to which interannual and multidecadal changes in the Earth's radiation budget can be attributed to changes in clouds. Cloud climatologies also find applications in assessing the prediction of clouds by climate models, assessing the significance of chemical reactions in clouds, quantifying climatic feedbacks involving clouds, estimating the radiative forcing by anthropogenic aerosols, selecting sites for astronomical observatories and atmospheric field experiments, and assessing the potential for solar energy development.

The properties of clouds most important for climate are those that affect radiation and precipitation, namely cloud height, thickness, horizontal extent and horizontal variability, water content, phase (liquid or ice), and droplet and crystal sizes. It is therefore important to distinguish different types of clouds. The climatic effects of clouds further depend on the geographical location of the clouds, the albedo and temperature of the underlying surface, the season, and the time of day. The effect of clouds on the Earth's radiation budget, called the 'cloud radiative effect,' is generally negative in the daytime but positive at night (i.e., clouds cool the surface in the day but warm the surface at night), so an accurate determination of the diurnal cycle of each cloud type is an important component of a cloud climatology.

Cloud Types

Clouds are classified according to their form and height. Low clouds, with bases in the atmospheric boundary layer less than 2 km above the surface, are influenced by their proximity to the surface. Solar heating of the surface can initiate convection and cause cumulus (Cu) clouds to form at the lifting condensation level. Cu clouds are small and may develop further into large cumulonimbus (Cb) clouds. Cb can extend vertically to the tropopause and often contain ice

crystals in their upper parts. Cb are associated with thunder, lightning, and showery rain or snow.

Stratus (St) and stratocumulus (Sc) are both horizontally extensive low clouds. They are distinguished in that Sc contains convective elements but St does not. Fog is a cloud at the ground surface, usually in the form of Sc. St and Sc cover large regions of the oceans. In the subtropics, they are found over the eastern parts of the oceans, where subsidence is occurring in the free atmosphere above the boundary layer.

Nimbostratus (Ns) clouds are much thicker than Sc and St, extending vertically through several kilometers of the atmosphere. Ns clouds form as a result of large-scale uplift of moist air near frontal boundaries in synoptic-scale storms at middle and high latitudes, and they precipitate rain and snow.

Clouds with bases 2–6 km above the surface are termed 'middle' clouds, and are classified as altostratus (As) or altocumulus (Ac) by their resemblance to St or Cu. Clouds with bases between 6 km and the tropopause are the 'high' clouds: cirrus (Ci), cirrostratus (Cs), and cirrocumulus (Cc). They consist of ice crystals and as a group are called 'cirriform' clouds. They can result from gradual uplift in large-scale storms in midlatitudes, or can be sheared off the tops of Cb in the tropics.

Clouds above the tropopause are rare, but they can occur in the polar regions in the stratosphere at 15–25 km height as polar stratospheric clouds (PSCs; nacreous clouds), and in the mesosphere at 80 km height as polar mesospheric clouds (PMCs; noctilucent clouds). These two types of clouds are discussed in other articles in the encyclopedia; this article is concerned only with tropospheric clouds.

Satellite Observations

Cloud climatologies have been developed from two kinds of data: (1) using radiances measured by satellites in polar and geostationary orbits; and (2) using visual observations of clouds from the Earth's surface, as coded in weather reports from stations on land and ships in the ocean. Satellites detect clouds principally at visible and thermal infrared wavelengths. At visible wavelengths, cloudy scenes appear brighter than cloud-free scenes when viewed from above. Clouds are usually colder than the underlying surface, so the emission of thermal infrared radiation to space is less than for clear scenes. During the daytime clouds can be detected in both wavelength regions, but at night only in the thermal infrared. The altitude of the cloud

top is inferred by relating the infrared emission temperature to the vertical profile of temperature obtained from radiosondes (carried by weather balloons) or satellite sounders. Cloud optical thickness (opacity) is inferred from reflectance in the visible channel. If a second solar channel (in the near infrared) is available, then the vertically integrated liquid water content, and the effective radius of the droplets, can also be inferred.

Measurements from satellites can be used to produce a cloud climatology if the following criteria are satisfied: (1) pixel size is at most a few kilometers, (2) temporal sampling is conducted at regular intervals throughout the day and night, (3) the coverage is global, and (4) a long period of record (many years) is maintained. To satisfy these requirements, the International Satellite Cloud Climatology Project (ISCCP) uses five geostationary satellites that hover over the equator at five longitudinal locations, and two polar-orbiting satellites. That project began in 1983 and is still continuing.

More detailed information about clouds can be obtained from satellite instruments with finer spatial resolution (e.g., Landsat) and from satellites with more spectral channels (e.g., the Moderate Resolution Imaging Spectroradiometer (MODIS) on the Earth Observing Satellites (EOS)). Three-dimensional information about clouds can be obtained from satellites that look at the same scene from different angles (e.g., the Multi-angle Imaging Spectroradiometer (MISR)). Cloud climatologies are also being developed using satellite-borne radar (CloudSat) and lidar (Cloud Aerosol Lidar with Orthogonal Polarization (CALIOP), Cloud Aerosol Lidar and Infrared Pathfinder Satellite Observations (CALIPSO)). These instruments are useful for studying cloud properties but do not yet offer sufficiently long periods of record to produce a climatology (see **Satellites and Satellite Remote Sensing: Remote Sensing: Cloud Properties**). The principal satellite cloud climatology in use now is therefore that of ISCCP.

Surface Observations

The surface observations of clouds are made less frequently than satellite observations in many areas, and they have variable spatial density, but they offer a useful adjunct to satellite observations for the following reasons.

1. The surface observer views clouds from below, and thus can observe the low clouds, which are often hidden from the satellite's view by higher clouds. Multiple cloud layers often occur together, so the views from above and below are complementary.
2. Some clouds are difficult to detect from satellites (clouds over snow, low clouds at night), because they provide little contrast in albedo or temperature to the underlying surface.
3. The surface observers are close to the clouds, so they can identify clouds by type, including clouds smaller than a satellite's pixel size, which is typically at least 1 km.
4. The cloud types defined morphologically by surface observers are directly related to meteorology and cloud processes, whereas the satellite climatology defines cloud types by their radiative properties.
5. Weather reports of clouds are available for several decades with no change in official observing instructions, so inter-decadal variations and trends can be studied.

This article emphasizes the climatology obtained from surface observations, because that is the subject of the authors' own research.

Cloud Information in Surface Weather Reports

Cloud observations are coded into weather reports using the 'synoptic code' of the World Meteorological Organization (WMO). In some countries, the observations are reported both in the synoptic code and in another code used locally. Reports in these other codes have been used to develop climatologies in some individual countries, but the synoptic code is the only one used worldwide. The information about clouds in the synoptic weather report consists of total cloud cover, low or middle cloud amount, low cloud type, middle cloud type, high cloud type, present weather, and base height of the lowest cloud.

About 6500 land stations routinely report cloud observations in the synoptic code. Usually, they report every 3 h beginning at 00.00 coordinated universal time (UTC), but about one-quarter of them report only every 6 h. About 20% of the stations do not make observations at night. The average spacing of land stations is about 180 km, but it is far from uniform. Europe has more stations than needed for a cloud climatology, and Antarctica has too few. Some parts of the Sahara Desert and Western Australia are also inadequately sampled.

Most ships make weather observations four times per day; the observations are recorded in logbooks and also transmitted by radio to world meteorological centers. In a recent typical year, reports from an average of 1150 ships were received at 00.00, 06.00, 12.00, 18.00 UTC and from 160 ships at 03.00, 09.00, 15.00, 21.00 UTC. Most of these are merchant ships with volunteer weather observers; some are military ships and research ships, and a few (less than 10) are dedicated weather ships. Unlike on land, there is little tendency for fewer observations at night, but the nighttime observations may not be transmitted promptly by radio, so it is important to have the complete logbook records. The average spacing of ships that report clouds is 600 km, much greater than for weather stations on land, but the ships are moving so they do sample most parts of the ocean. A project to compile ship-based weather observations from all maritime nations, including many logbook reports, has resulted in the International Comprehensive Ocean-Atmosphere Data Set (ICOADS), which is being used for research on air-sea interaction and climatic change throughout the world ocean.

In many parts of the ocean, the accuracy of computed mean cloud amounts is limited by the scarcity of observations. This is not the case on land, where the random error due to inadequate temporal sampling is very small. Statistical tests performed on observations from weather ships indicate that 100 observations taken at random times during a 3-month period will represent the seasonal mean cloud cover to within 3% in an oceanic grid box of size 5° latitude by 5° longitude. If 1% accuracy is desired, then 1000 observations are needed.

The synoptic code was defined in 1929, but changed in 1949; the reporting procedures became adopted worldwide in the early 1950s. Synoptic observations are available with global coverage for all oceans since 1954 and for all continents since 1971, about 550 million observations to date.

Computation of Average Cloud Amounts

For low clouds, the computation of average amount is straightforward, but for middle and high clouds the question of overlap must be considered. The 'amount' of a cloud type is defined as the fraction of the sky covered by that type, whether visible or hidden behind another cloud. The time-averaged amount can be obtained as the product of frequency-of-occurrence (fraction of weather observations in which a cloud of this type is present, whether visible or hidden) and amount-when-present (the average fraction of the sky covered by this cloud type when it is present, whether visible or hidden). For example, if Cu is present in 30% of the weather observations from a station, and if it covers on average 40% of the sky when it is present, then the average amount of Cu at that station is 12%.

The amount, or even the presence, of a middle or high cloud may be indeterminate when a lower cloud nearly or completely covers the sky. The average amounts of middle and high cloud types can be estimated by assuming that the frequency and amount-when-present are the same in observations where they cannot be calculated and as in observations where they can be calculated. Also, to obtain amount-when-present the clouds at different levels are assumed to be randomly overlapped. The amounts directly visible from below (the 'nonoverlapped' amounts) may also be calculated.

For the climatology, the Earth is divided into an array of boxes on a geographical grid, and cloud cover is computed for each box. There are several possible biases which may affect computed cloud cover but which may be reduced or eliminated with appropriate analysis procedures. Two small biases that oppose each other and are unique to ship observations are the fair-weather bias (the tendency for more ships to enter a grid box on days of fair weather) and the foul-weather bias (the tendency of ships to oversample stormy or foggy weather because they are traveling more slowly). Two other biases that may affect both ship and land data are the diurnal sampling bias (somewhat more reports are transmitted by ships during the daytime than at night, and some land stations in a box with several stations may not report at night) and the trend bias (a box may be sampled by more ships in later years than in earlier years, or a land station may change location during the period of record). These situations can cause biases if the cloud amount undergoes a diurnal cycle or exhibits a long-term trend, but such biases can be eliminated by appropriate analysis procedures.

The most serious bias, on both land and ocean, is the 'night-detection bias.' Visual observations of clouds are hindered at night due to inadequate illumination of the clouds. This usually leads to an underestimation of the average cloud cover at night, especially for the amounts of middle and high clouds, in climatologies based on surface observations. The diurnal cycles of cloud amounts, if based on all the surface observations, are therefore in error, but the cycles can be obtained more accurately if the nighttime observations are screened to select those made under sufficient moonlight or twilight. A criterion for adequacy of moonlight or twilight has been established; it permits the use of about 38% of the nighttime observations. By this criterion, adequate illumination is provided by a full moon at an elevation angle of 6° or a partial moon at higher elevation, or twilight from the sun less than 9° below the horizon.

A complete description of the climatology of clouds is the subject of atlases such as those in the bibliography, which give the average amounts of each cloud type for each season in grid boxes of 5° latitude by 5° longitude, as well as their diurnal cycles and interannual variations. A few illustrative examples from the climatology are shown in this article. Updated plots and descriptions are available at the climatology website: www.atmos.washington.edu/CloudMap.

Global Averages

The annual average total cloud cover as determined from surface observations is summarized in Table 1. Average cloud cover is greater over the ocean than over land. Cloud cover over land tends to be greater in daytime than at night, but the ocean shows little day-night difference.

Cloud properties from 8 years (1986–93) of the ISCCP are summarized in Table 2. The average cloud cover and the day-night differences are slightly different than those obtained from surface observations (Table 1). The optical thickness (opacity) and cloud water path (vertically integrated liquid water content) inferred from the satellite radiances are smaller than those usually obtained from aircraft in field experiments. This difference is probably due to horizontal inhomogeneity of the clouds; ISCCP's optical thickness is an effective optical thickness for a hypothetical horizontally homogeneous cloud.

Global average amounts for nine different cloud types defined in the surface observations are shown in Table 3. Globally, the most common types are Sc, Ac, and high (cirriform) clouds. All the low cloud types are more common over the ocean than over land. The middle cloud types As and Ac together cover the same fraction of the sky over land as over ocean; cirriform cloud is the only type that is less common over ocean.

For the low clouds, Table 3 also shows the observers' estimate of the height of the cloud base above the ground surface. The bases are on average twice as high over land as over ocean, and the heights increase with distance inland from the ocean.

Table 1 Annual average cloud cover from surface observations (1971–96, land; 1954–97, ocean)

	<i>Land</i>	<i>Ocean</i>	<i>Globe</i>
Average total cloud amount (%)	54	69	64
Day–night difference (%)	4	–1	0

Table 2 Annual average cloud properties from the ISCCP (1986–93)

	<i>Land</i>	<i>Ocean</i>	<i>Globe</i>
Average total cloud amount (%)	58	72	68
Day–night difference (%)	+5	–2	0
Cloud top temperature (°C)	–20	–7	–11
Day–night difference (K)	+12	+2	+5
Cloud top pressure (mbar)	490	620	580
Cloud optical thickness	4	4	4
Cloud water path (g m ^{–2})	76	61	66

Data from Rossow, W.B., Schiffer, R.A., 1999. Advances in understanding clouds from ISCCP. Bulletin of the American Meteorological Society 80, 2261–2287, Boston, MA: American Meteorological Society.

Table 3 Cloud type amounts and heights from surface observations

Cloud type	Annual average amount (%)		Base height (meters above surface)	
	Land	Ocean	Land	Ocean
Fog	1	1	0	0
St	5	13	500	400
Sc	12	22	1000	600
Cu	5	13	1100	600
Cb	4	6	1000	500
Ns	5	5		
As	4	6		
Ac	17	18		
High (cirriform)	22	12		
Clear sky (frequency)	22	3		

Land values for the years 1971–96. Ocean values are for the years 1954–2008, from Table 9 from: www.atmos.washington.edu/CloudMap/Atlases/DistOcean.pdf. The amounts of all the cloud types add up to more than the total cloud cover because of overlap.

Geographical Variations

What the averages of **Table 3** cannot show is that there are striking geographical variations. Fog is rare over most of the globe but its frequency exceeds 10% over the North Atlantic and North Pacific oceans in summer poleward of 40° N, and reaches 20–40% in the Sea of Okhotsk. Ns is likewise rare in the tropics but common in middle and high latitudes. Cb amount exceeds 10% in a narrow band along the intertropical convergence zone (ITCZ) near the equator and over a much broader region of warm water in the Western Pacific called the ‘warm pool.’

Completely clear sky, also given in **Table 3**, is common over land but rare over the ocean. Over most of the open ocean there

are almost always some low clouds visible from ships; the reports of clear sky are mostly confined to coastal regions. The Red, Mediterranean, and Arabian Seas are the most cloud-free parts of the world ocean.

The geographical distribution of total cloud cover for Dec.–Jan.–Feb. (DJF) is shown in **Figure 1**. This is the only season in which complete global coverage is available from surface observations, because ships avoid the Antarctic Ocean in other seasons when it is ice-covered. The largest cloud amounts are found in the high latitude oceans, particularly in summer, exceeding 90% in the sub-Antarctic in DJF. The North Atlantic and North Pacific cloud amounts reach similarly high values during the northern summer (Jun.–Jul.–Aug. (JJA)). Values of cloud cover below 40% are found in the deserts of Australia, Central Asia, Arabia, North Africa, Southern Africa, and Mexico. Most of the Indian subcontinent has cloud cover 20–30% on this map, during the winter monsoon dry season. The Eastern Sahara is the clearest large region on the Earth, with cloud cover below 20% on this map for DJF, but it is even clearer in summer when a large area of less than 5% cloud cover extends on across Northern Arabia. A complete set of maps (for all types in all seasons) is available at the website: www.atmos.washington.edu/CloudMap.

The ITCZ appears in **Figure 1** as a latitudinal maximum near the equator in the Atlantic and north of the equator in the Eastern Pacific, then south of the equator through Indonesia, and at about 10° S across the Indian Ocean, Africa, and South America. This is in agreement with the location of the ITCZ as determined by measurement of winds and pressure.

The total cloud cover averaged around latitude zones is shown in **Figure 2** for the two extreme seasons. The figure shows that the average cloud cover is less over land than over ocean, and that the latitudinal variation of cloud cover is greater over

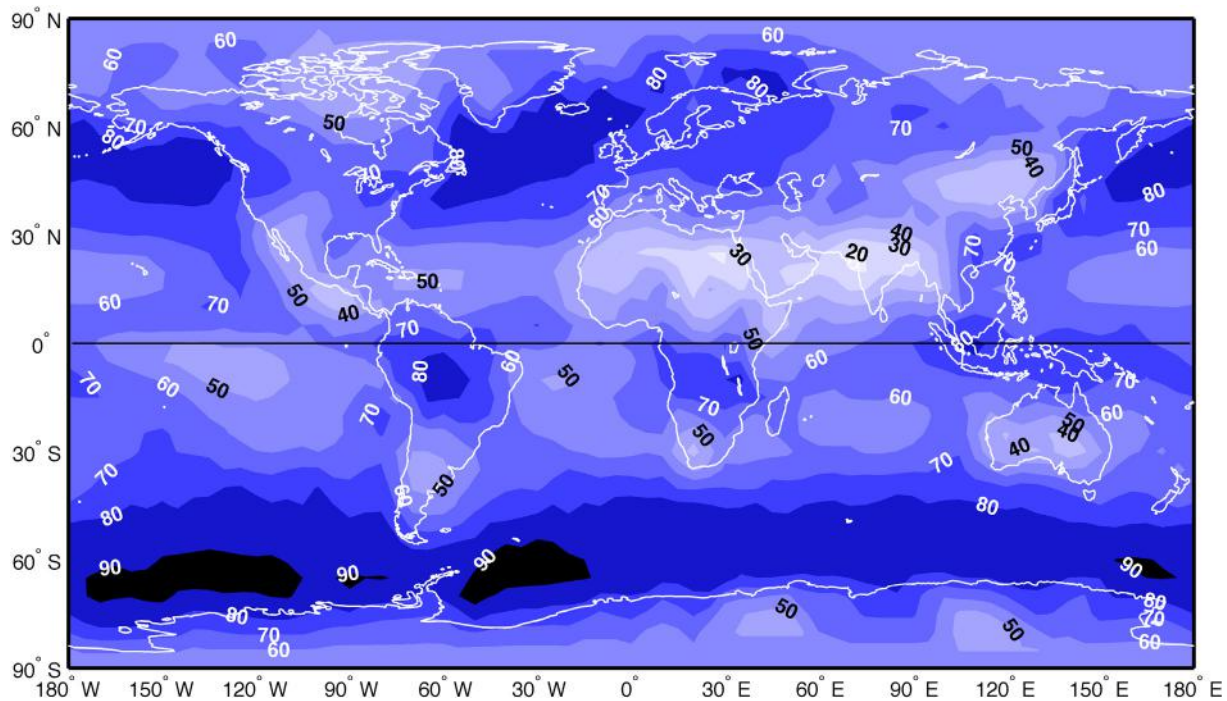


Figure 1 Percent total cloud cover for DJF from surface observations (weather stations on land, ships in the ocean) for the 26-year period 1971–96 over land and the 44-year period 1954–97 over the ocean.

land than over ocean. The peak cloudiness in the ITCZ moves from 7° N in JJA only to 2° N in DJF over the ocean, but to as far as 12° S over land. The latitudes of maximum cloud cover near 60° N, 60° S, and the equator correspond to the latitudes of maximum precipitation, and the latitudes of minimum cloud cover on land are the latitudes of the great deserts.

Diurnal Variations

The amounts of many cloud types vary from day to night. Two examples of diurnal variations for oceanic regions are shown in Figure 3. The central North Pacific in winter exhibits no diurnal variation, with total cloud cover averaging about 82% at all hours. The largest oceanic diurnal variations are in the Sc regions of the eastern subtropical Atlantic and Pacific. The region displayed, in the Atlantic Ocean west of Namibia, exhibits a strong diurnal cycle in total cloud cover with a peak of 80% at 04.00 and minimum of 55% at 16.00. This cycle is paralleled by the diurnal cycle of low stratiform clouds, indicating that these cloud types are the types responsible for the diurnal cycle here. These boundary layer clouds develop during the night and dissipate during the day under the influence of solar heating.

Figure 4 shows an example of diurnal cycles on land, in Central America during the summer rainy season. Solar heating of the surface begins at sunrise, leading to convection which produces Cu clouds in the morning. In the afternoon, many of these clouds further develop into Cb, which continue precipitating into the evening.

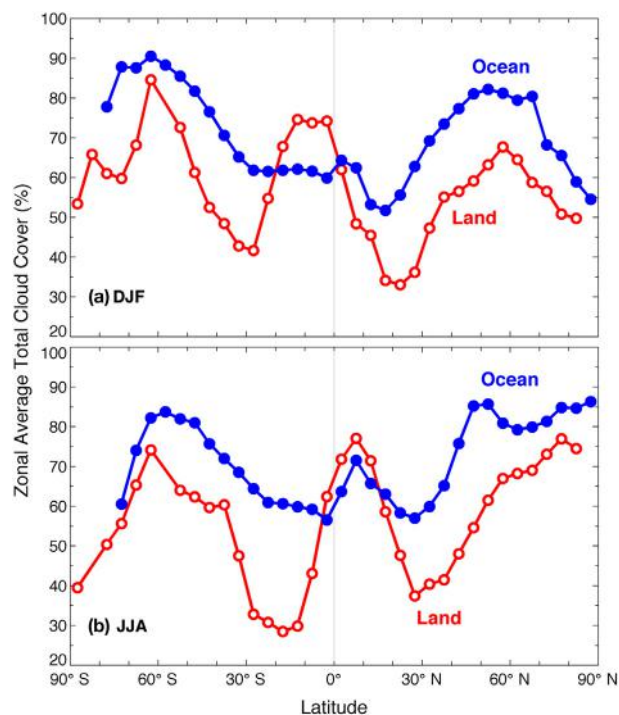


Figure 2 Zonal average total cloud cover (average of day and night) for 5° latitude zones. Separate averages are formed for the land and ocean parts of each zone. (a) DJF and (b) JJA. Data span the periods 1971–96 over land, and 1954–97 over the ocean.

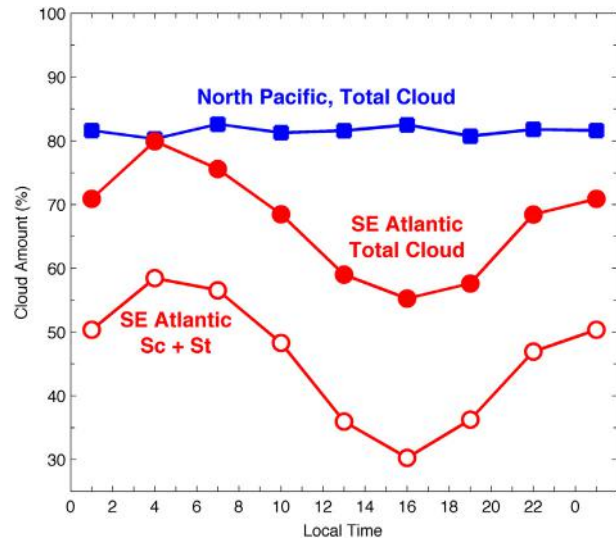


Figure 3 Diurnal cycles of oceanic cloud, from ship observations in DJF (1954–97). North Pacific: 40–50° N, 170–150° W; Southeast Atlantic: 20–30° S, and 0–20° E.

Seasonal Variations

The largest seasonal variations of cloud cover are associated with the subtropical monsoons of Africa, South America, India, and Australia. Cloud variations in the Indian Ocean region are shown in Figure 5(a). In Southwest India, the average total cloud cover increases from 16% in February to 89% in July. During India's dry winter, Northern Australia experiences its cloudy and rainy summer.

In contrast to the sinusoidal pattern of the Indian and Australian monsoons, the Central Arctic Ocean (Figure 5(b)) exhibits a peculiar boxlike seasonal cycle, in which cloud cover increases rapidly during May. The greater cloud cover from June

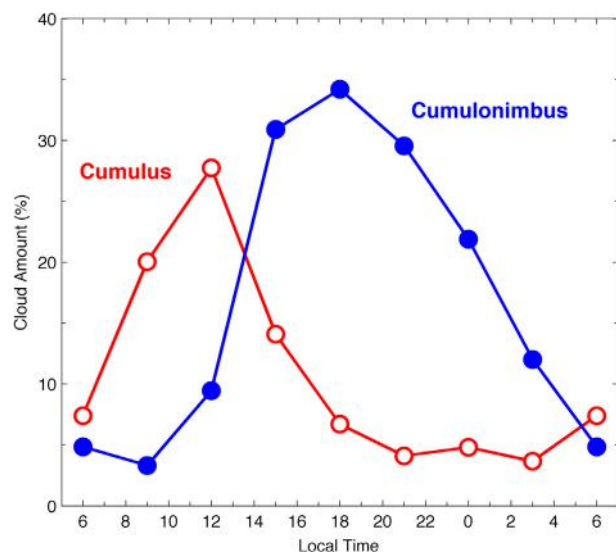


Figure 4 Diurnal cycles of Cu and Cb amounts reported from weather stations in Central America (10–15° N, 85–90° W) in summer (JJA 1971–96).

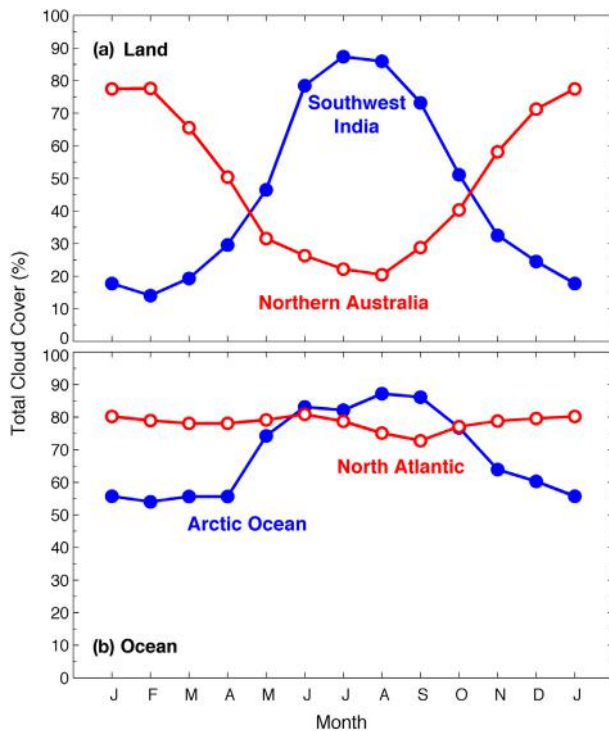


Figure 5 Examples of seasonal cycles of total cloud cover from surface observations. (a) Land (1971–96): Southwest India, 15–20° N, 70–75° E; Northern Australia, 10–15° S, 130–135° E. (b) Ocean (1954–97): North Atlantic, 40–50° N, 20–30° W; Arctic Ocean, 80–90° N. Observations in the Arctic Ocean were made from drifting stations established on perennial sea ice.

to September is due mainly to the low thin 'Arctic summer St' clouds that form over the perennial sea ice during the melting season. A still different pattern is exhibited in parts of the North Atlantic (Figure 5(b)) which have almost no seasonal variation of cloud cover.

Figure 6 shows a map of the amplitude of the seasonal cycle of total cloud cover over both land and ocean. With the exception of the North Atlantic, the regions shown in Figure 5 appear in Figure 6 as darker areas, indicating a large amplitude. Other areas showing a large seasonal cycle include tropical South America, Central America, the Sahel, and Southern Africa. This figure also illustrates the tendency for ocean areas to have a less pronounced seasonal cycle than land areas.

Interannual Variations and Trends

Clouds interact with other components of the climate system, so changes in cloud amounts can be expected to accompany changes in other climatic variables, and also to feed back on those other variables. The magnitudes, and even the nature, of the possible climatic feedbacks involving clouds are not well understood, but the long historical climatic record may help to identify them. The degree to which the actual variations of the amounts of the different cloud types are faithfully recorded in the analysis of visual observations is itself variable, depending on the spatial and temporal density of observations, the ability to detect and remove biases, and the spatial scale of the analysis. Real interannual variations of cloud amount in a $10^\circ \times 10^\circ$ grid box, for example, are often large enough to overwhelm any subtle progressive changes in observing procedure. However, interannual variations of zonal average cloud amount are smaller than

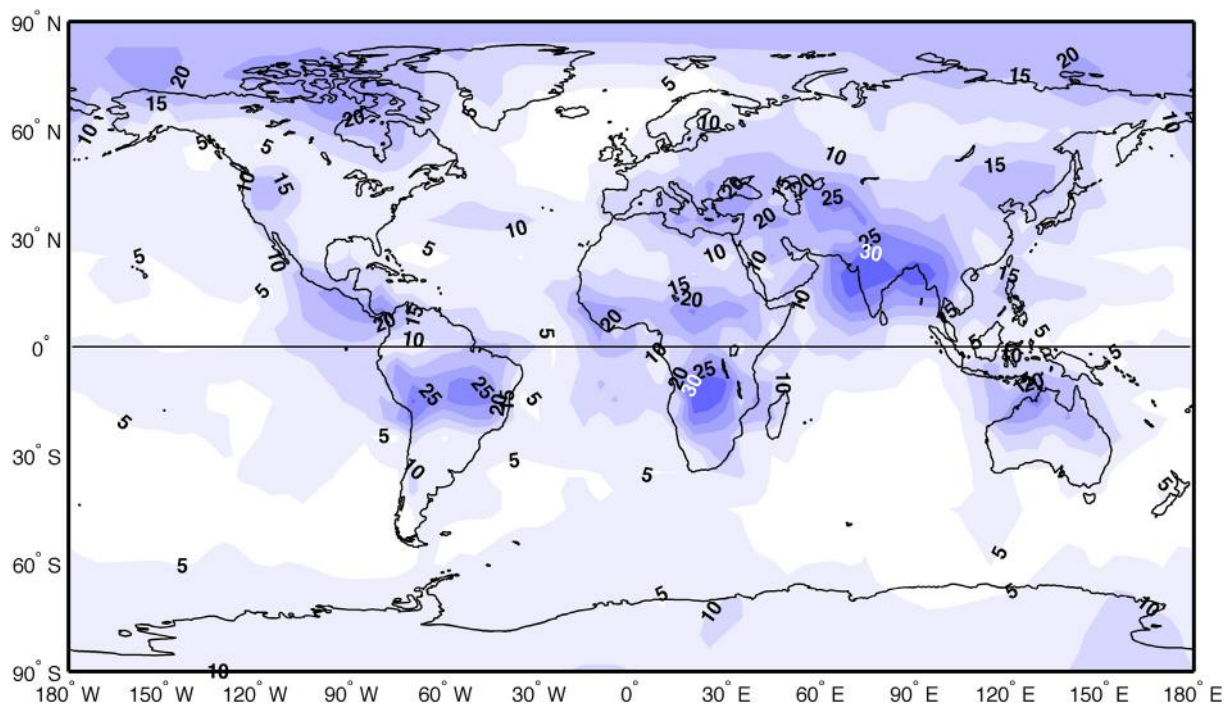


Figure 6 Amplitude (percent) of the seasonal cycle of total cloud cover over land (1971–96) and ocean (1954–97).

those of grid box cloud amount because of partially compensating positive and negative changes in different parts of the zone. For zonal averages, it is therefore more difficult to dissect the observed changes into climatic and nonclimatic causes.

A powerful way to assess the validity of observed cloud changes is to identify likely causes (e.g., changes in sea surface temperature (SST) or atmospheric circulation) and effects (e.g., diurnal temperature range) of the cloud changes and to correlate these related climatic variables with the cloud changes. Some examples are shown in Figure 7. Interannual variations of the amount of marine stratiform cloudiness ($St + Sc$) commonly correlate negatively with interannual variations of SST. Figure 7 shows how $St + Sc$ and SST (both measured aboard ships, but by different methods) covary at two locations in the Pacific Ocean. Many of these interannual variations are related to cycles of El Niño and the Southern Oscillation. Frame b shows downward spikes in stratiform cloud cover coinciding with high SST (plotted as increasing downward) during strong El Niño years: 1972–73, 1982–83, and 1997–98. That clouds and SST are measured differently but correlate well argues for the reality of both time series. The strong correlations in Figure 7 also suggest that the error in a seasonal mean due to

random sampling of weather situations during a particular season is small. In other grid boxes traversed by fewer ships, seasonal means suffer from sampling error and the correlation of $St + Sc$ with SST is not as strong.

Cloud cover changes may accompany the global warming brought on by anthropogenic increases in greenhouse gases. Regional changes in cloud cover may also be expected from anthropogenic sulfate aerosols that can act as cloud condensation nuclei. Figure 8 shows trends (0.1%/decade) in total cloud cover for grid boxes from 1971 through 1996. Decreases in total cloud cover are observed over South America, Southern Africa, and in a large area between Southern Australia and Northern China. Decreases in cloud cover are also seen in all eastern subtropical ocean basins. Increasing cloud cover is observed in the Central Equatorial Pacific, Western Africa, and Arctic North America. The causes of these observed changes in cloud cover are still under investigation.

Individual time series for selected boxes (shown in Figure 8) showing significant trends in total cloud cover are plotted in Figure 9. Time series are plotted as seasonal anomalies (the departure from the multiyear seasonal mean cloud cover). Figure 9(a) shows the time series over the Equatorial

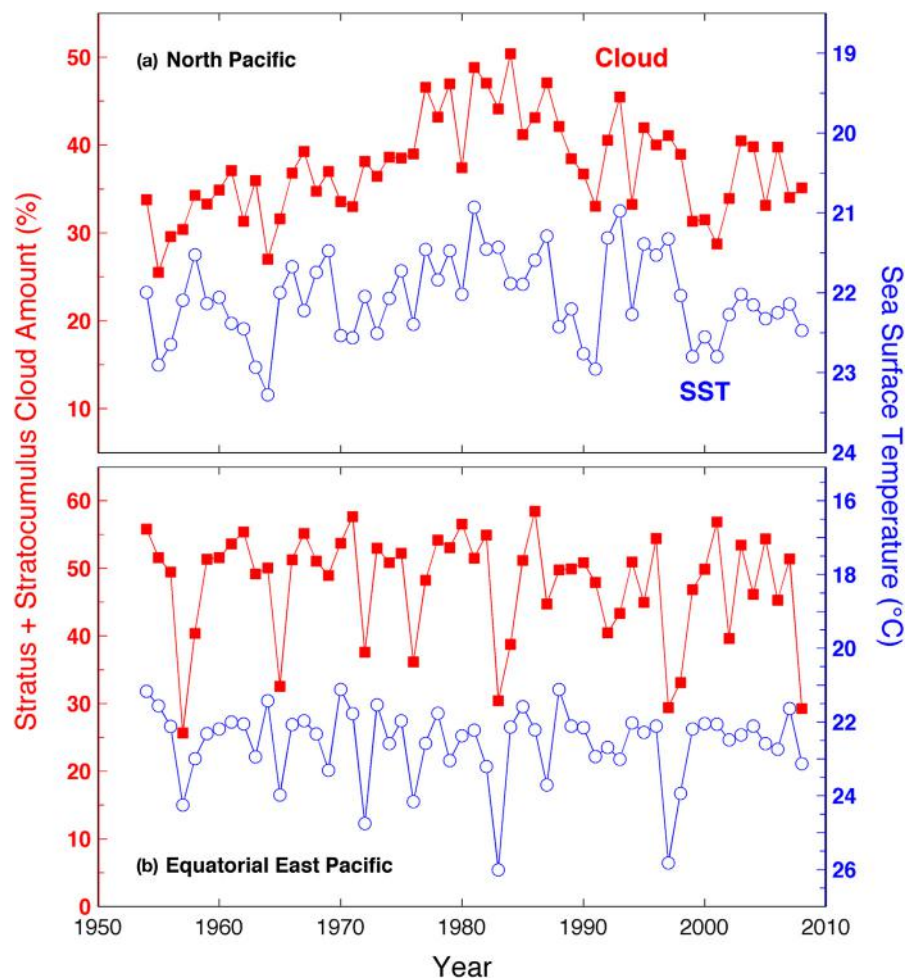


Figure 7 Seasonal average daytime amounts of St plus Sc , and seasonal average SST for two grid boxes in the Pacific Ocean. (a) JJA, $30\text{--}40^\circ\text{ N}$, $160\text{--}180^\circ\text{ W}$. (b) JJA $0\text{--}10^\circ\text{ S}$, $80\text{--}100^\circ\text{ W}$. The SST is plotted on a reversed scale to illustrate the correlation. (SST data provided by the Hadley Centre HadISST1 product.)

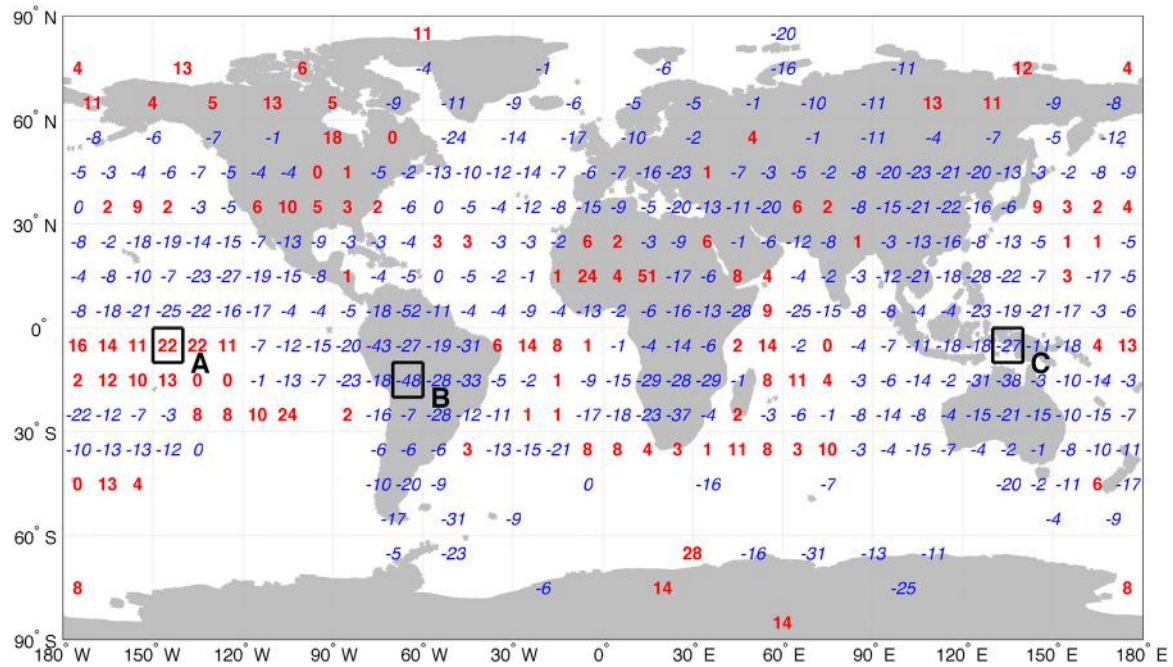


Figure 8 Yearly average trends (0.1%/decade) in total cloud cover over land and ocean (1971–96). Boxes A, B, and C used in **Figure 9** are labeled.

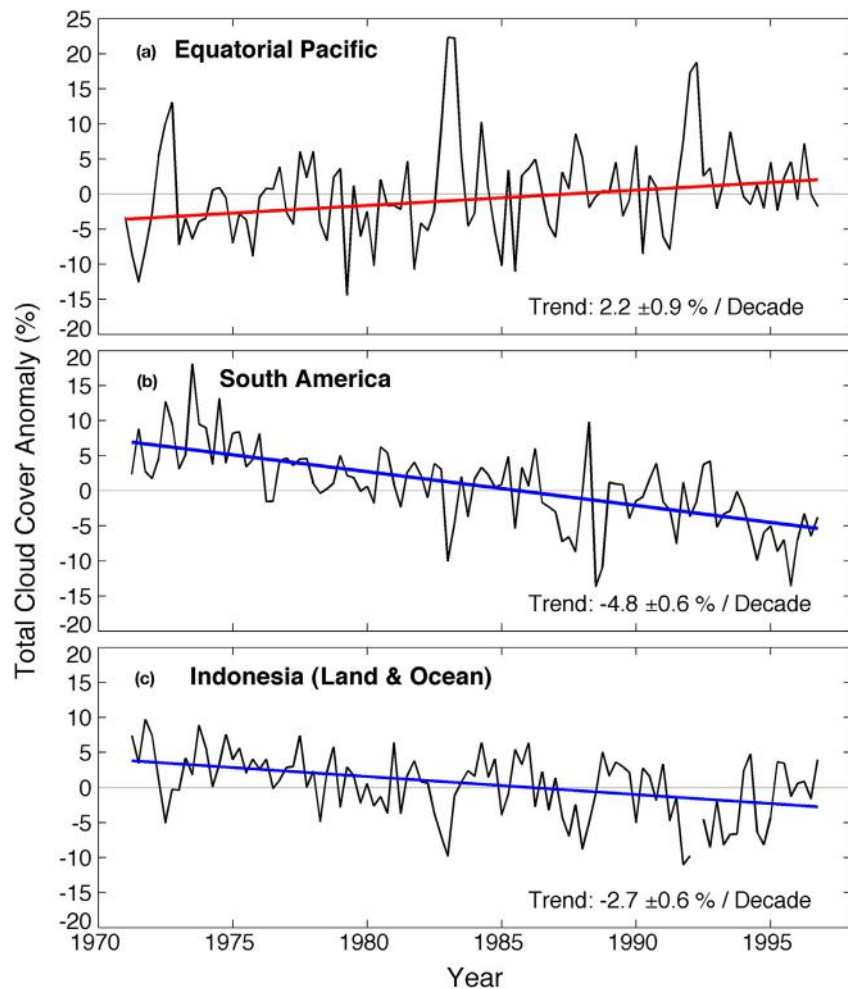


Figure 9 Time series of anomalies (percent departure from seasonal average) of total cloud cover in the three grid boxes outlined in **Figure 8**: (a) Equatorial Pacific, (b) South America, and (c) Indonesia (using both land and ocean data). Each point represents one year's seasonal anomaly. Trend lines are fit to the time series and shown in red (increasing) or blue (decreasing).

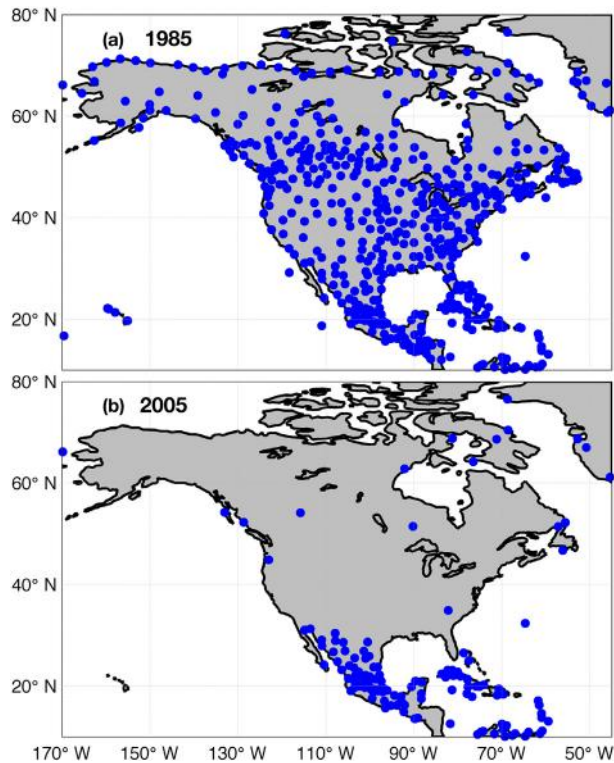


Figure 10 North American weather stations contributing cloud observations to the global climatology in (a) 1985 and (b) 2005.

Pacific. While an increasing trend of about 6% over 26 years can be seen, the year–year variation ranges up to 20% with three distinct spikes, likely associated with El Niño activity. Frames b and c show decreasing cloud cover over boxes in South America and Indonesia, respectively. The year–year variations in these boxes are also greater in magnitude than the trends, which show a decrease of roughly 10% over 26 years. This figure shows that while significant trends in these areas are apparent, the trends are not drastic and it is still normal to see above or below average cloud cover during any given season.

At present, it is difficult to obtain reliable multiyear trends of cloud amounts from satellite observations because of the short lifetime of individual satellites and the difficulty of intercalibrating instruments on different satellites, especially because the spectral response of the radiation detectors may change from one satellite to the next. However, efforts are underway to address these problems, and in the future more use will be made of satellite observations to detect long-term changes of cloud amounts.

The Future of Cloud Observations

The long time span of cloud reports, covering the transition period from a time of perhaps minimal human impact on climate in the 1950s to the anthropogenically altered climate of the future, is a valuable resource that is appreciated by many national meteorological agencies. There has been remarkable worldwide international cooperation in reporting weather

observations in compliance with the WMO regulations. Except for brief periods of political instability (Iran in 1979, Zaire in the late 1990s), essentially all nations have been contributing their weather observations reliably. Recently, however, three nations (United States, Canada, and New Zealand), in conjunction with the automation of their weather stations, have essentially ceased reporting of visual cloud observations in the WMO synoptic code. Changes of codes or changes of observational methods (laser ceilometers in place of the human eye), or even changes of station location, make it difficult to infer reliable climatic changes over a span of years that includes the time of the change. The number of US stations with useful synoptic weather reports decreased slowly throughout the 1980s and rapidly in the mid-1990s, so that the geographical coverage of the US declined from 241 stations in 1981 to only about 27 by the end of 1996. This transition is illustrated in Figure 10, which plots the locations of weather stations reporting visual cloud observations in 1985 and 2005. While Mexico, the Caribbean, Bermuda, Venezuela, remote Eastern Siberia, and Greenland have dutifully continued to report cloud observations, the United States and Canada currently furnish only a handful of reliable weather stations. The United States, Canada, and New Zealand together represent 4% of the Earth's surface, so future global analyses of cloud changes from surface observations will be restricted to the remaining 96% of the globe.

See also: **Clouds and Fog:** Classification of Clouds; Cloud Modeling; Cloud Microphysics; Measurement Techniques *In Situ*; Noctilucent Clouds; Stratus and Stratocumulus. **Numerical Models:** Parameterization of Physical Processes: Clouds. **Radiation Transfer in the Atmosphere:** Cloud-Radiative Processes. **Satellites and Satellite Remote Sensing:** Remote Sensing: Cloud Properties. **Tropospheric Chemistry and Composition:** Cloud Chemistry.

Further Reading

- Climatic Atlas of Clouds over Land and Ocean <http://www.atmos.washington.edu/CloudMap> [accessed 26.03.12].
- Eastman, R., Warren, S.G., Hahn, C.J., 2011. Variations in cloud cover and cloud types over the ocean from surface observations 1954–2008. *Journal of Climate* 24, 5914–5934. Available online: <http://www.atmos.washington.edu/~rmeast/OceanCloudsweb.pdf>.
- Rossow, W.B., Schiffer, R.A., 1999. Advances in understanding clouds from ISCCP. *Bulletin of the American Meteorological Society*, 80, 2261–2287. American Meteorological Society, Boston.
- Surface based cloud climatology webpage: <http://www.atmos.washington.edu/CloudMap/>.
- Warren, S.G., Hahn, C.J., London, J., Chervin, R.M., Jenne, R.L., 1986. Global distribution of total cloud cover and cloud type amounts over land. Technical Note TN-273+STR. National Center for Atmospheric Research, Boulder, CO. Available as pdf online: <http://www.atmos.washington.edu/CloudMap/Atlases/DistLand.pdf>.
- Warren, S.G., Hahn, C.J., London, J., Chervin, R.M., Jenne, R.L., 1988. Global distribution of total cloud cover and cloud type amounts over the ocean. Technical Note TN-317+STR. National Center for Atmospheric Research, Boulder, CO. Available as pdf online: www.atmos.washington.edu/CloudMap/Atlases/DistOcean.pdf.
- Warren, S.G., Eastman, R., Hahn, C.J., 2007. A survey of changes in cloud cover and cloud types over land from surface observations, 1971–1996. *Journal of Climate* 20, 717–738. Available online: http://www.atmos.washington.edu/CloudMap/Publications/WarrenEtal2007_CloudSurvey.pdf.
- World Meteorological Organization, 1987. *International Cloud Atlas*, vols. 1 and 2. WMO, Geneva, Switzerland, p. 212.

Measurement Techniques *In Situ*

D Baumgardner, Universidad Nacional Autónoma de México, Mexico City, D.F., Mexico

J-F Gayet, Université Blaise Pascal, Clermont Ferrand, France

A Korolev, Meteorological Service of Canada, Toronto, ON, Canada

C Twohy, Oregon State University, Corvallis, OR, USA

J Fugal, Max Planck Institute of Chemistry, Mainz, Germany

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

The impact of clouds on weather and climate is determined by their microphysical properties (i.e., the size distribution, water content, optical properties, and shape). Understanding how these properties evolve requires detailed studies with *in situ* measurements using a suite of different sensors. This article gives an overview of the different techniques that are used to determine these cloud characteristics and examples of some of the types of information that can be extracted from these data.

Cloud Properties and Their Impact on the Environment

Climate, weather, and hydrologic cycles are strongly affected by the temporal and spatial properties of clouds. Clouds consist of liquid water or ice hydrometeors whose number concentration, size, density, and shape determine these properties. **Table 1** lists the environmentally important properties of clouds and the hydrometeor characteristics to which they are connected. **Table 2** defines these properties with respect to hydrometeor characteristics and the equations that relate them to the size distributions of cloud particles.

Clouds impact climate primarily by altering the radiative balance in the regions where they form. Hence, cloud particle optical properties (i.e., phase function, extinction coefficient, and effective radius) are the important parameters. Weather is affected by clouds not only because of changes in the radiative balance, but also by precipitation. The formation of precipitation depends on the number concentration, size, density, and shape of cloud particles.

In situ measurements of cloud properties provide detail that cannot be obtained from remote sensors like radars and satellites. This detail is needed to better understand the physical processes that govern the evolution of clouds. *In situ* measurements focus on the evaluation of the hydrometeor

characteristics listed in **Table 2**. Some of these characteristics can be measured directly with currently available instrumentation, while others must be derived from the size distribution that can be measured by some of the sensors discussed in this article.

Measurement Techniques

Table 3 lists the general techniques that are used to measure hydrometeor properties. Hydrometeors are detected using five fundamentally different measurement approaches: impaction, phase change, light scattering from individual particles, light scattering from an ensemble of particles, and hydrometeor imaging. The measurement principles and the more frequently used instruments that utilize these techniques are described in this section.

Hydrometeor Detection by Impaction and Replication

This technique was one of the first approaches for examining the size distribution and concentration of hydrometeors. It has been implemented in several ways, but the basic principle is that a small impaction surface in the shape of a cylinder, disk, tape, or wire is exposed to hydrometeors in a moving air

Table 1 Cloud properties, hydrometeor characteristics, and environmental impact

<i>Cloud property</i>	<i>Hydrometeor characteristic</i>	<i>Environmental impact</i>
Albedo	Number and surface area Phase function and extinction Effective radius	Climate
Lifetime	Number and mass Fall velocity	Climate, weather, and hydrological cycles
Spatial distribution	Number, surface area, and mass Fall velocity	Climate, weather, and hydrological cycles
Precipitation efficiency and rain rate	Number and mass Fall velocity	Climate, weather, and hydrological cycles
Chemical-processing efficiency	Number, surface area, and mass concentration Fall velocity	Climate and hydrological cycles

Table 2 Hydrometeor properties

Number concentration (cm^{-3})	$N = \sum_{i=1}^m n_i$
Surface area concentration ($\mu\text{m}^2 \text{cm}^{-3}$)	$S = \pi \sum_{i=1}^m s_i n_i d_i^2$
Mass concentration (g m^{-3})	$M = \frac{\pi}{6} \sum_{i=1}^m s_i \rho_i n_i d_i^3$
Fall velocity (cm s^{-1})	$V_t = f(d_i, g, \rho_a, C_D)$
Phase function	$P_{\lambda, \theta} = \int_{r_1}^{r_2} F(\eta, \lambda, r, \theta) \pi r^2 n(r) dr$
Extinction (m^{-1})	$\sigma_e = \int_{r_1}^{r_2} Q_e(\eta, \lambda, r) \pi r^2 n(r) dr$
Effective radius (μm)	$R_e = \frac{\sum_{i=1}^m n_i d_i^3}{\sum_{i=1}^m n_i d_i^2}$

m = number of size categories.

n_i = number concentration of hydrometeors in size category i .

d_i = average diameter of size category i .

r_i = average radius of size category i .

s_i = shape factor of the hydrometeor of size category i , to account for asphericity (this factor depends upon the intrinsic property being defined).

ρ_i = density of the hydrometeor in size category i .

ρ_a = air density.

g = gravitational acceleration.

C_D = drag coefficient.

F = angular scattering intensity efficiency.

Q_e = extinction efficiency.

η = hydrometeor refractive index.

λ = wavelength of incident light.

θ = angle of light scattered from the hydrometeor.

stream. The surface is coated with a highly viscous liquid to prevent the particles from escaping the surface after impact. The hydrometeors, or their impressions, are subsequently counted and sized with manual or automatic methods to determine the number concentration and size distribution. The version of this technique that is currently in use is the video ice particle sampler (VIPS) that uses a moving, 8 mm transparent tape that is coated with silicone oil and exposed to the particle-laden air stream. After impaction, the captured droplets or crystals are recorded digitally as the tape moves in front of two video cameras. Another similar device, the Cloudscope, uses a video camera to record hydrometeors as they impact a heated glass window. The size of each hydrometeor is deduced from the size and shape of the impacted particle. The rate at which it evaporates provides information on the water mass.

These two instruments are not commercially available but have been developed for research at several institutes. The National Center for Atmospheric Research and the Desert Research Institute (Reno, NV) use the VIPS and Cloudscope in airborne cloud studies, respectively. A dropsonde and balloon-borne instruments (HYVIS), operating on a principle similar to that of the VIPS, have been used by the Meteorological Research Institute in Japan to sample cirrus and thunderstorm clouds for about the last 20 years.

Hydrometeor Mass Detection by Phase Change

Several instruments measure the water content of hydrometeors by measuring the energy of phase change when they

evaporate. Two of these instruments, the hot-wire liquid water sensor and the Nevzorov probe, derive the water mass from the amount of thermal energy required to evaporate the hydrometeors after impaction. This method utilizes a cylindrical or disk-shaped sensor that is maintained at a constant temperature. Thermal energy is removed from the sensor through convection and by the evaporation of water droplets that strike the heated element. The power required to maintain the sensor at a constant temperature is measured, and the thermal energy removed by convection is calculated directly from heat transfer theory so that the remaining energy loss is due to evaporation of water droplets. The water mass, M , that leads to this energy loss is derived from first principles, that is,

$$M = \frac{P - P_d}{AV(L_v + c_w(T_b - T_a))} \quad [1]$$

where

P = measured power;

P_d = convective heat loss;

A = sensor area;

V = air velocity;

L_v = latent heat of vaporization;

c_w = specific heat of water;

T_b = boiling point of water;

T_a = air temperature.

Sensors that utilize a cylindrical geometry are useful only for water clouds since ice crystals will bounce from the heated element before totally evaporating. This characteristic, however, has been utilized in the Nevzorov probe, shown schematically in Figure 1. This instrument uses two heated sensors, one with a cylindrical geometry and the other in the shape of a concave disk. In water clouds, both sensors respond equally. In mixed-phase or total ice clouds, the concave sensor captures the hydrometeors long enough for them to evaporate. Comparison of the total water content from the concave sensor and the liquid water controller (LWC) from the cylindrical sensor provides a measure of the ice water content.

Another method for deriving water mass by evaporation is to measure water vapor that is formed from evaporation of the hydrometeors. The Counterflow Virtual Impactor (CVI) is one common instrument utilizing this technique. At the CVI inlet tip (Figure 2), cloud droplets or ice crystals larger than some minimum aerodynamic diameter (5–10 μm diameter depending on conditions) are separated from the interstitial aerosol and water vapor and ‘virtually’ impacted into dry nitrogen gas or purified air. This separation is possible via a counterflow stream of gas out the CVI tip, which assures that only larger hydrometeors are sampled. The water vapor and nonvolatile, residual nuclei that remain after droplet evaporation are sampled downstream of the inlet with selected instruments. These may include a hygrometer to determine water content, a condensation nucleus counter, an optical particle counter, or in-line particle filters for various chemical analyses. Since droplets or crystals in a large sampling volume converge into a smaller sample stream within the instrument, concentrations within the CVI are significantly enhanced, which leads to more sensitivity.

Table 3 Measurement techniques

Sensing technique	Hydrometeor properties measured or derived	Instruments in active use
Impaction and replication of hydrometeors on a substrate	N : directly M : size distribution integration V_i : size distribution integration $P_{\lambda,\theta}$: size distribution integration R_e : size distribution integration σ_e : size distribution integration	Video ice hydrometeor sampler (VIPS) ^a HYVIS ^b Cloudscope ^c
Phase change of hydrometeors	N : from CVI only M : directly	Hot-wire liquid water sensor ^d Nevzorov probe ^e Counterflow virtual impactor (CVI) ^{a,d,f}
Single hydrometeor light scattering	N : directly M : size distribution integration V_i : size distribution integration $P_{\lambda,\theta}$: size distribution integration (polar nephelometer measures directly) R_e : size distribution integration σ_e : size distribution integration	Forward scattering spectrometer probe (FSSP-100) ^d Cloud and Aerosol Spectrometer (CAS) ^d CAS with depolarization (CAS-DPOL) ^d Cloud droplet probe (CDP) ^d
Hydrometeor ensemble light scattering	M : direct from PVM $P_{\lambda,\theta}$: partial information from CIN R_e : direct from PVM σ_e : direct from CIN and transmission meter	Particle volume monitor (PVM) ^g Cloud integrating nephelometer (CIN) ^g
Nonintrusive optical imaging	N : directly M : size distribution integration V_i : size distribution integration $P_{\lambda,\theta}$: size distribution integration R_e : size distribution integration σ_e : size distribution integration	Cloud-imaging probe (CIP) ^d Cloud particle imager (CPI) ^h 2D-S probe ^h

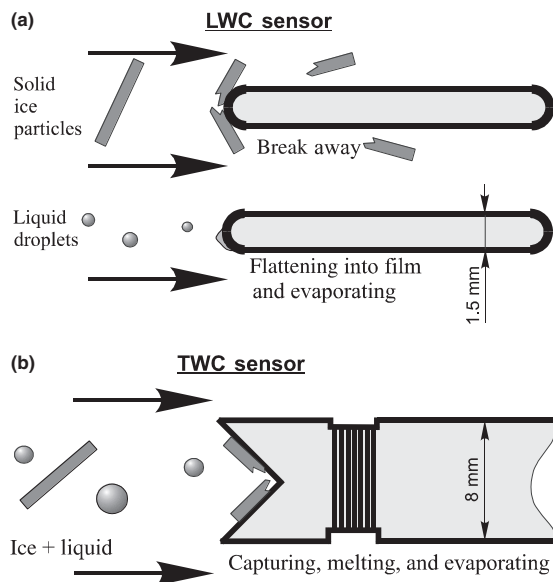
^aNational Center for Atmospheric Research, Boulder, CO, USA.^bMeteorological Research Institute, Japan.^cDesert Research Institute, Reno, NV, USA.^dDroplet Measurement Technologies, Boulder, CO, USA.^eSky Tech Research, Toronto, ONT, Canada.^fBMI, Hayward, CA, USA.^gGerber Scientific Inc., Reston, VA, USA.^hStratton Park Engineering, Boulder, CO, USA.

Figure 1 The Nevzorov instrument uses two sensors to determine the liquid water content (LWC) and total water content (TWC). The LWC sensor (a) responds primarily to water droplets, while the TWC (b) responds to water and ice particles.

Hydrometeor Size Detection by Single Hydrometeor Scattering

Optical particle counters detect the light scattered when a particle passes through a focused laser beam. Instruments that convert the light amplitude into a size, using Mie scattering theory, are called single-particle spectrometers. Currently used instruments for measuring cloud droplet sizes and concentrations are the forward scattering spectrometer probe (FSSP), the cloud droplet probe (CDP), the small ice detector (SID), and the cloud and aerosol spectrometer with depolarization (CAS-DPOL). As individual droplets in the free air stream pass through the laser beam and scatter light, the light is collected through a cone with a solid angle from 4° to 12° and a detector converts the photons into an electrical signal proportional to the optical cross-section of the particle. The CAS-DPOL has two additional sets of optics that collect backward-scattered light from 168° to 176° and separate it into two components: one component is measured with a detector behind a polarizing filter, and the second component is measured by a detector with no filter. Comparison of the two components determines if the hydrometeor was an ice crystal or water droplet. **Figure 3** is a diagram of the CAS-DPOL optical configuration showing the primary features of the light-scattering collection system used in optical spectrometers, the main difference being the

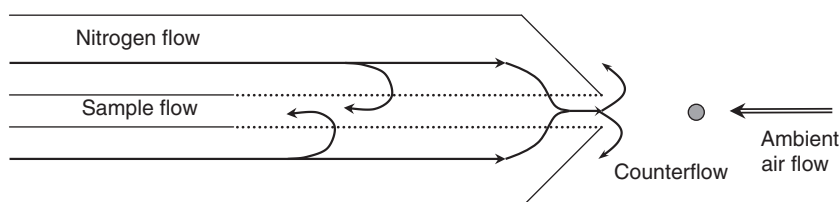


Figure 2 Schematic of the CVI inlet. Nitrogen gas is pumped into the annular region formed by two concentric tubes joined at the tip. The gas flows in opposing directions after passing through the porous center tube. The counterflow out the inlet tip prevents small interstitial aerosol particles from entering the inlet, while cloud droplets or ice particles have enough inertia to pass through the counterflow region and enter the sample flow. Once sampled, droplets or crystals are evaporated in the heated tip region.

additional backscatter optics in the CAS-DPOL. The signal-processing electronics analyze the maximum amplitude associated with each light-scattering event and categorize this signal into a preassigned size bin such that, over a specified period of time, a frequency histogram is accumulated that is called a size distribution (i.e., a method of presenting the number of particles in different size intervals that were found in a particular cloudy volume of air). When the particle number is divided by the volume sampled, this size distribution is expressed as a number concentration per size interval. Mie theory, combined with calibration using spherical particles of known size and refractive index, is used to convert the scattered light signal into a size. The cloud properties are then derived, as shown in Table 2, by integrating over this size distribution. The combination of forward-scattering and backscattering signals from the CAS-DPOL is used to distinguish droplets from ice crystals. Figure 4 illustrates measurements made in a mixed-phase cloud where the separation between droplets and crystals is clearly seen.

The SID was specifically designed to discriminate ice from water based upon the angular distribution of light

scattered in the forward direction. The SID-3, the most recent version, uses a charged coupled detector array to measure the azimuthal distribution of light scattered over a forward-scattering angle by individual cloud particles passing through the laser beam.

The SID-3 can discriminate supercooled liquid drops from small ice particles, based on their scattering pattern, and estimate the size of the ice particles up to approximately 140 μm . Figure 5 illustrates some of the patterns that are measured by the SID-3 when viewing water droplets and ice crystals.

Single-particle scattering is also used by the LaMP polar nephelometer to directly measure the phase function of hydrometeors. In this instrument, shown schematically in Figure 6, the particles intersect a collimated, high-energy laser beam (1 W, $\lambda = 804 \text{ nm}$) at the focal point of a paraboloidal reflector. The scattered light is reflected onto a circular array of 54 detectors. Each detector senses signals corresponding to a range of scattering angles. From these measurements, a phase function can be derived, as shown in Figure 7, where the phase function of ice crystals in a cirrus cloud is compared with theoretical predictions.

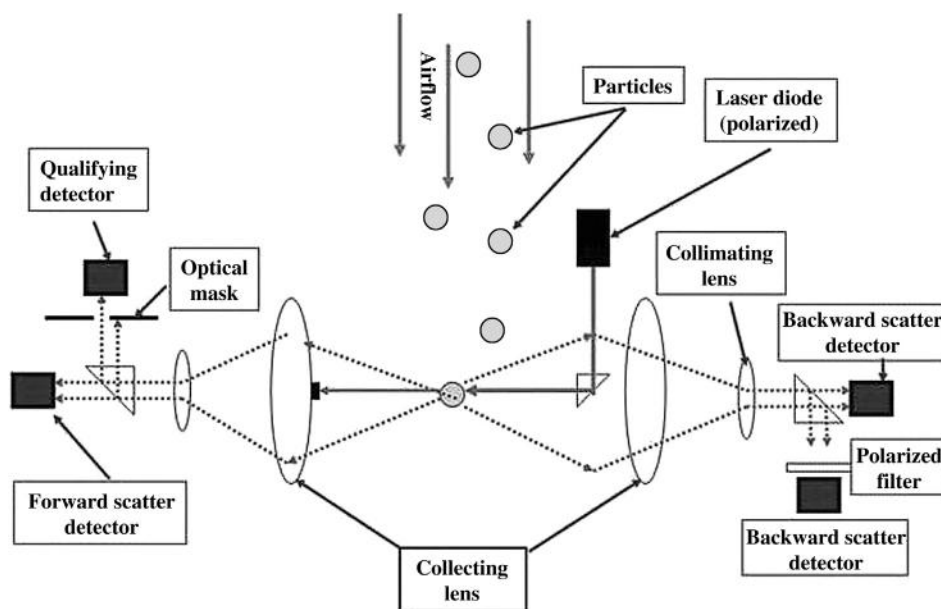


Figure 3 This block diagram shows the general optical layout for components in the CAS. Light that is scattered while particles pass through the CAS laser beam is collected in both the forward and backward directions and measured with photodetectors. The 'masked detector' has aperture that blocks light that is scattered from particles that pass through the laser beam at distances farther from the center of focus than optimum.

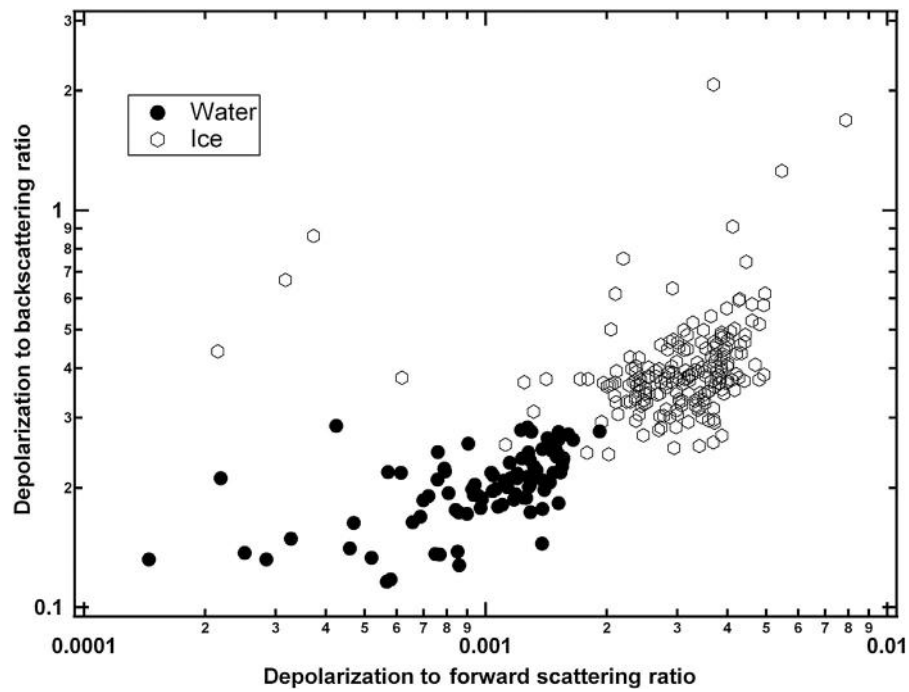


Figure 4 Comparison of the depolarization signal divided by the backscattering nonpolarized signal and the depolarization signal divided by the forward-scattering signal, for individual particle measurements, shows that there are two distinct regimes that separate water droplets from ice crystals. These measurements were made in a mixed-phase cloud.

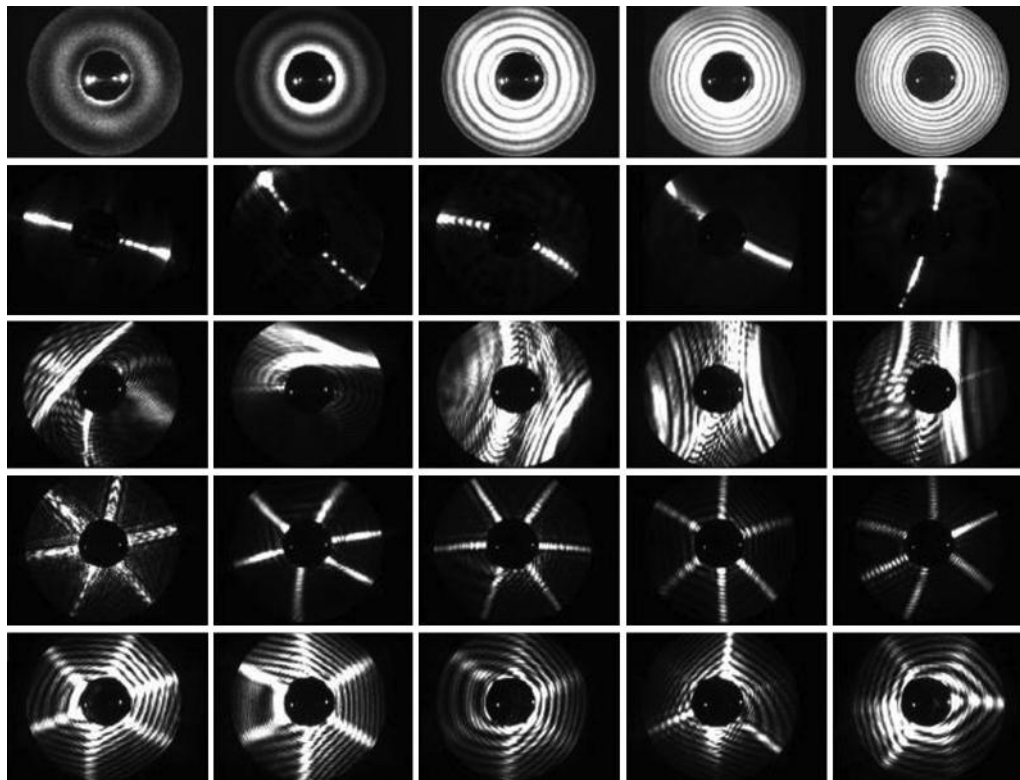


Figure 5 These images are the azimuthal, forward-scattering patterns measured by the SID-3. The top row shows water droplets with different diameters, and the lower rows are ice crystals with various habits and sizes.

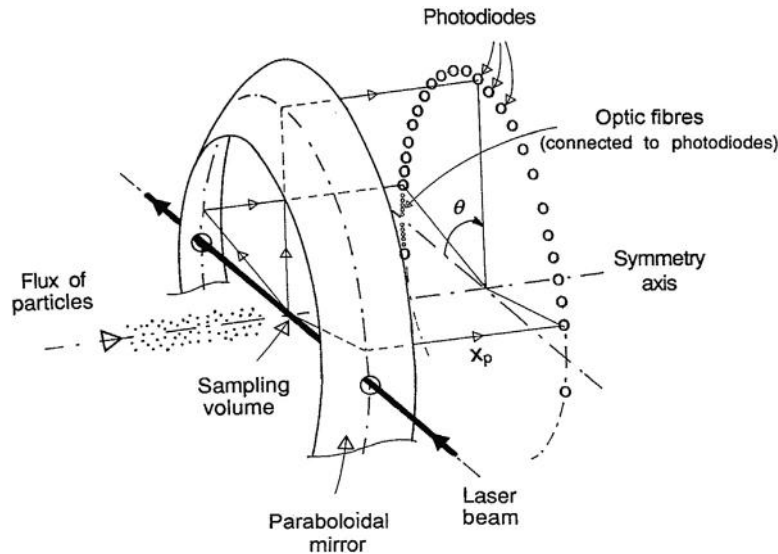


Figure 6 The principal components of the LaMP polar nephelometer are shown in this schematic. The particles intersect a collimated high-energy laser beam at the focal point of the paraboloidal reflector. The scattered light is reflected onto a circular array of 54 detectors.

Hydrometeor Mass and Optical Properties by Ensemble Particle Scattering

Light scattering from an ensemble of hydrometeors is the technique used by several instruments that derive LWC, extinction coefficient, and effective radius from their measurements. Light scattered by particles is composed of three components: reflection, refraction, and diffraction. The FSSP and CAS measure light that is primarily reflected and refracted. In contrast, ensemble-scattering instruments measure mostly the diffracted component of scattered light. The instruments

consist of a light source and one or more detectors that measure transmitted or scattered light from the particle field. The particle volume monitor (PVM), shown schematically in [Figure 8](#), illuminates an ensemble of hydrometeors with a collimated laser and focuses the near-forward scattered light onto two detectors that are masked with variable-transmission filters. The filters have been designed to provide transmission functions that are mathematically derived to approximate inversions of the integral equations that relate particle surface area (PSA) or LWC to the flux of light scattered by the ensemble

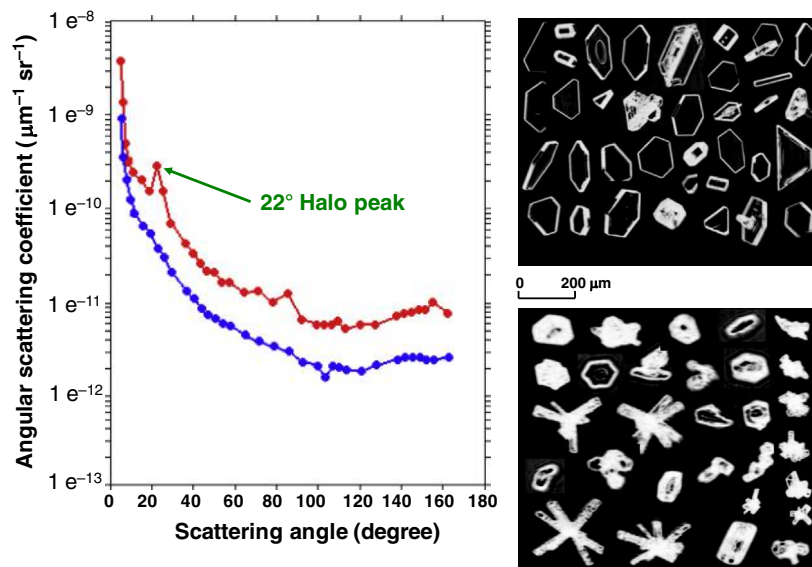


Figure 7 An example of a phase function constructed from measurements in cirrus clouds, with the LaMP polar nephelometer is shown in this figure. The red curve exhibits the 22° halo peak due to pristine hexagonal-plated ice crystals, as exemplified from CPI images (upper-right panel). The smooth feature (blue curve) is a common observation in cirrus clouds with prevalent irregular-shaped ice particles (see CPI images on the bottom-right panel).

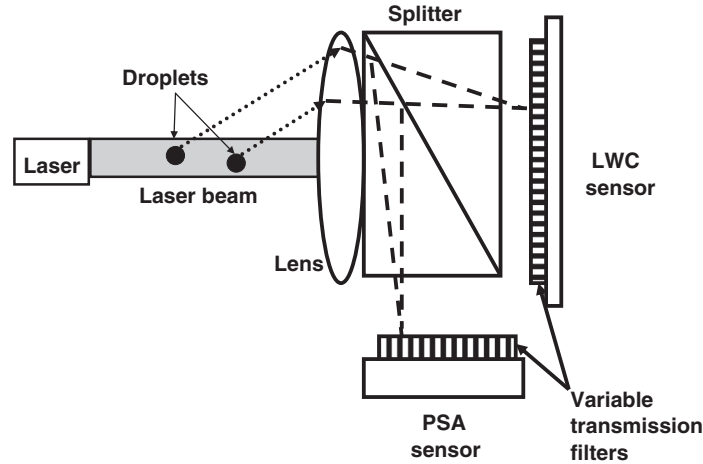


Figure 8 The principal components of the PVM-100 are illustrated in this schematic. Forward-scattered light from an ensemble of particles in the laser beam is measured by the LWC and PSA detectors. These detectors are masked with variable transmissions filters, whose transmission functions are weighted to provide light that is proportional to LWC or PSA, respectively.

of hydrometeors. It has been shown theoretically that the scattered-light flux $F(d)$, measured by the PVM LWC sensor, for an ensemble of hydrometeors of size d , will be proportional to $k_1 d^3$, where k_1 is a scaling constant. This proportionality holds true if the transmission filter consists of annular segments identified by the subscript j , and $F(d) = \sum_{j=1}^N f_j T_j$, where $f_j(d)$ is the flux of diffracted light incident on j , T_j is the transmission of filter j , and N is the total number of annuli. The transmission functions T_j are determined by mathematical inversion of this relationship, and the scaling factor k is determined through calibrations in a well-characterized wet wind tunnel. A similar procedure is followed to measure PSA as a function of $k_2 d^2$, and different sets of annular filters are used on the PSA detector. The ratio of the PVM LWC to PSA gives a measure of the effective radius r_e (Table 2).

The phase function of hydrometeors, as discussed in Section [Hydrometeor Size Detection by Single Hydrometeor Scattering](#),

can be measured for single particles using an array of detectors. An approximation to the phase function at a given wavelength λ is the asymmetry factor g_λ :

$$g_\lambda = \frac{\int \cos \theta P(\lambda, \theta) d(\cos \theta)}{\int P(\lambda, \theta) d(\cos \theta)} \quad [2]$$

The asymmetry factor is the average cosine of the angles for scattered radiation and provides a measure of the relative amount of forward- and backscattered light from an ensemble of hydrometeors. In other words, this factor is >0 when more light is forward scattered than backscattered, 0 for isotropic scattering, and <0 when the majority of light is backscattered.

One instrument that has been specifically designed to measure g_λ is the cloud integrating nephelometer (CIN). The CIN consists of a collimated laser beam that passes through an ensemble of hydrometeors and four detectors that are positioned to measure scattered light in the forward direction

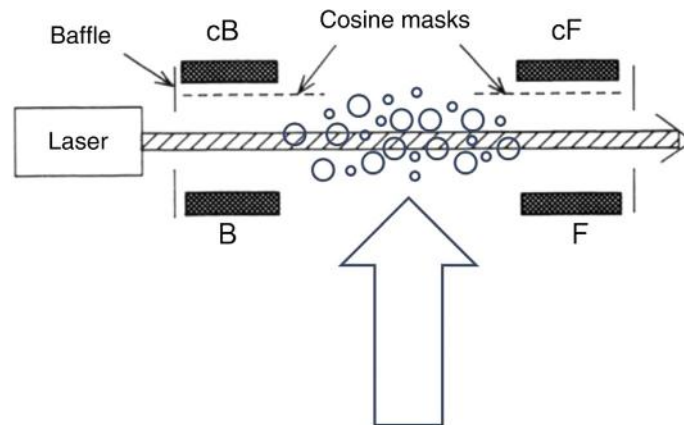


Figure 9 The CIN, whose primary optical components are illustrated in this schematic, measures an approximation to the total scattering coefficient by adding the scattered light measured by detectors B and F. The cosine masks on detectors cB and cF provide an approximation to the asymmetry factor when divided by the scattering coefficient.

(S1 and S3) and in the backward direction (S2 and S4), as shown in Figure 9. The denominator in eqn [2] is proportional to the total scattering coefficient, and the sum of detectors B and F is an approximation to this coefficient. The cB and cF detectors have optical masks that cosine-weight the scattered light collected by these detectors, and their sum is an approximation to the numerator of eqn [2]. After suitable corrections for the angles over which light is not being collected, the ratio of (cB + cF) to (B + F) provides an approximation to the asymmetry factor.

Nonintrusive Optical Imaging

One of the first methods in cloud physics research for optically measuring hydrometeor size was by imaging onto a linear diode array the shadow of a particle that passed through a laser beam (Figure 10). The on-off state of the diodes in the array is recorded at a rate proportional to the velocity of the particles passing through the laser, and the images can be subsequently reconstructed to show the features of the hydrometeors, as shown in Figure 11. These types of instruments, called two-dimensional optical array probes (2D-OAPs), can typically measure in the size ranges of 10–1280 μm (2D-S probe), 12.5–800 μm (cloud-imaging probe, or CIP), 25–800 μm (2D cloud probe, or 2DC), 100–6400 μm (precipitation imaging probe, or PIP), and 200–6400 μm (2D precipitation probe, or 2DP).

This technique has been refined to measure hydrometeors at a higher resolution (measuring ones as small as 2.5 μm) with the cloud particle imager (CPI), by using a pulsed laser and a two-dimensional photodetector array to capture the particle image. In addition, 256 Gy levels are measured in the CPI. Figure 12 shows examples of some of the ice crystals that have been measured with the CPI.

Holography, another technique for viewing ensembles of cloud particles, provides the position in three-dimensional

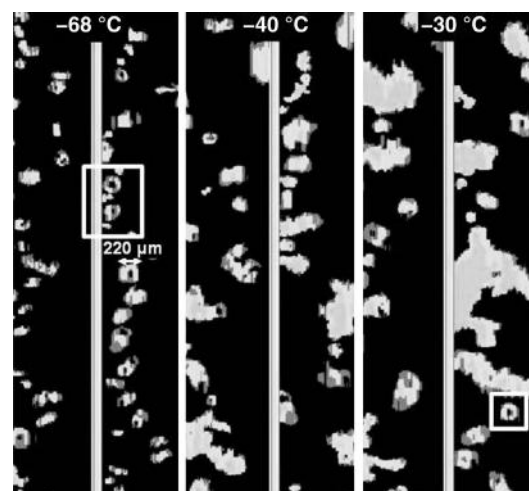


Figure 11 This figure illustrates the reconstructed images recorded with a CIP OAP and shows the structure of ice crystals measured at different temperature levels in clouds.

(3D) space, the shape, and the size of each particle in a dilute collection of cloud droplets and ice particles inside a localized 3D sample volume. Simply stated, an in-line hologram is an interference pattern resulting from the superposition of an incident plane wave (the reference wave) and light scattered by the dilute suspension of illuminated particles. The interference of the reference and object waves are called the real and virtual images, respectively. When the real image is reconstructed, the virtual image reconstructs as a blurry background appearing around the reconstructed 'focused' real image. In practice, images of sampled particles are obtained by numerical reconstruction of the recorded digital holograms. Figure 13 shows several reconstructed images of ice crystals.

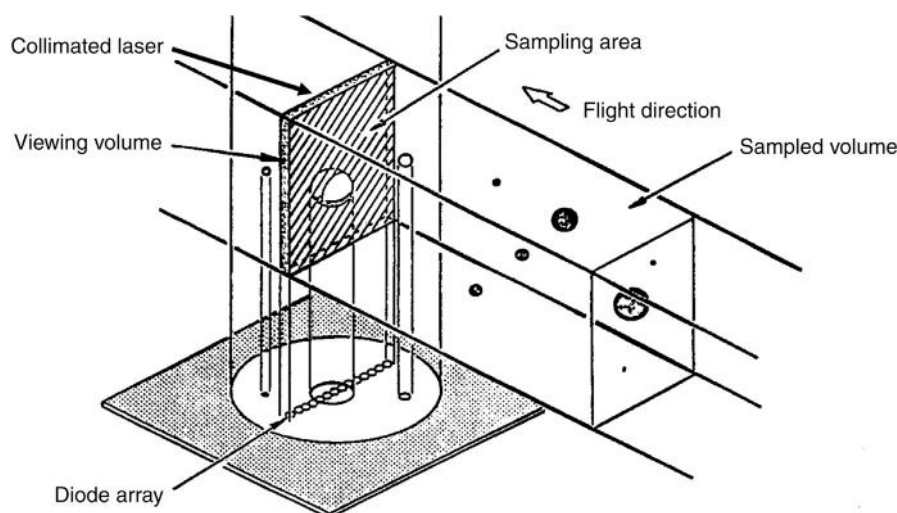


Figure 10 OAPs measure the shadow image of particles that pass through a collimated laser, as shown in this schematic. The laser illuminates a linear array of diodes whose 'on' and 'off' states are monitored by signal-processing electronics to determine when a shadow is present or not. A two-dimensional image is reconstructed by recording the on and off state of these diodes as a function of time.

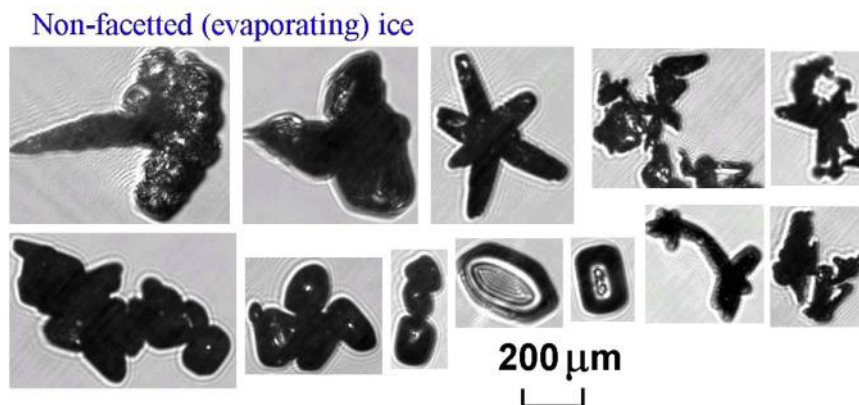


Figure 12 These images are examples of the type of measurements that can be made with the CPI. They are representative of the nonpristine, asymmetric crystals that are more often than not seen in clouds and demonstrate the power of the CPI to resolve the very fine structure of small ice crystals.

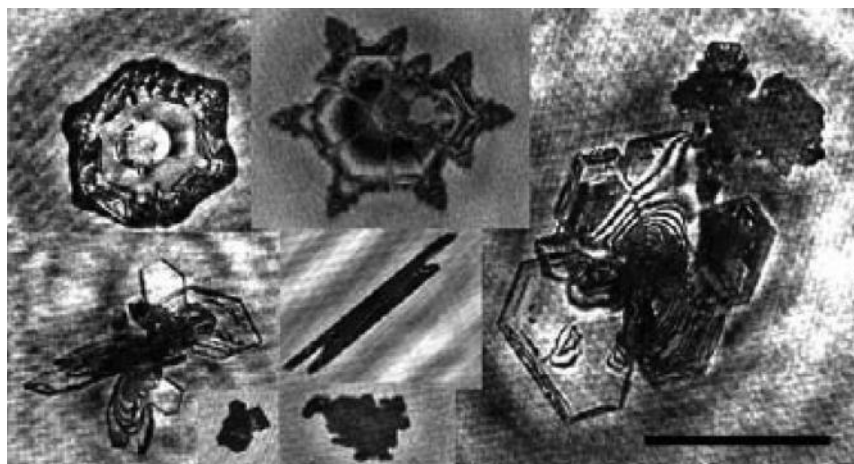


Figure 13 These are holographic images of small ice crystals.

See also: **Aerosols:** Aerosol–Cloud Interactions and Their Radiative Forcing. **Clouds and Fog:** Cloud Modeling; Cloud Microphysics; Contrails; Fog; Stratus and Stratocumulus. **Middle Atmosphere:** Transport Circulation. **Tropospheric Chemistry and Composition:** Cloud Chemistry.

Further Reading

- Baumgardner, D., Brenguier, J.L., Bucholtz, A., Coe, H., DeMott, P., Garrett, T.J., Gayet, J.F., Hermann, M., Heymsfield, A., Korolev, A., Krämer, M., Petzold, A., Strapp, W., Pilewskie, P., Taylor, J., Twohy, C., Wendisch, M., 2011. Airborne instruments to measure atmospheric aerosol particles, clouds and radiation: a cook's tour of mature and emerging technology. *Atmospheric Research* 102, 10–29. <http://dx.doi.org/10.1016/j.atmosres.2011.06.021>.
- Baumgardner, D., Avallone, L., Bansemer, A., Borrmann, S., Brown, P., Bundke, U., Chuang, P.Y., Cziczo, D., Field, P., Gallagher, M., Gayet, J.-F., Heymsfield, A., Korolev, A., Krämer, M., McFarquhar, G., Mertes, S., Möhler, O., Lance, S., Lawson, P., Petters, M., Pratt, K., Roberts, G., Rogers, D., Stetzer, O., Stith, J.,

- Strapp, W., Twohy, C., Wendisch, M., 2012. *In situ*, airborne instrumentation: addressing and solving measurement problems in ice clouds. *Bulletin of the American Meteorological Society* 93, ES29–ES34. <http://dx.doi.org/10.1175/BAMS-D-11-00123.1>.
- Davis, S.M., Avallone, L.M., Kahn, B.H., Meyer, K.G., Baumgardner, D., 2009. Comparison of airborne *in situ* measurements and MODIS retrievals of cirrus cloud optical and microphysical properties during the midlatitude cirrus experiment (MidCIX). *Journal of Geophysical Research* 114, D02203. <http://dx.doi.org/10.1029/2008JD010284>.
- Fugal, J.P., Shaw, R.A., 2009. Cloud particle size distributions measured with an airborne digital in-line holographic instrument. *Atmospheric Measurement Techniques* 2, 259–271.
- Gayet, J.F., Crepel, O., Fournol, J.F., Oshchepkov, S., 1997. A new polar nephelometer for measurements of optical and microphysical cloud properties. Part I: theoretical design. *Annales de Geophysique* 15, 451–459.
- Gerber, H., 1991. Direct measurement of suspended particulate volume concentration and far-infrared extinction coefficient with a laser-diffraction instrument. *Applied Optics* 30, 4824–4831.
- Kaye, P.H., Hirst, E., Greenaway, R.S., Ulanowski, Z., Hesse, E., DeMott, P.J., Saunders, C., Connolly, P., 2008. Classifying atmospheric ice crystals by spatial light scattering. *Optics Letters* 33 (13), 1545.
- King, W.D., Parkin, D.A., Handsworth, R.J., 1978. A hot wire water device having fully calculable response characteristics. *Journal of Applied Meteorology* 17, 1809–1813.

- Knollenberg, R.G., 1981. Techniques for probing cloud microstructure. In: Hobbs, P.V., Deepak, A. (Eds.), *Clouds: Their Formation, Optical Properties, and Effects*. Academic Press, New York.
- Korolev, A.V., Strapp, J.W., Isaac, G.A., Nevzorov, A.V., 1998. The Nevzorov airborne hot wire LWC/TWC probe: principal of operation and performance characteristics. *Journal of Atmospheric and Oceanic Technology* 15, 1495–1510.
- Lance, S., Brock, C.A., Rogers, D., Gordon, J.A., 2010. Water droplet calibration of the cloud droplet probe (CDP) and in-flight performance in liquid, ice and mixed-phase clouds during ARCPAC. *Atmospheric Measurement Techniques* 3, 1683–1706. <http://dx.doi.org/10.5194/amt-3-1683-2010>.
- Twohy, C.H., Schanot, A.J., Cooper, W.A., 1997. Measurement of condensed water content in liquid and ice clouds using an airborne counterflow virtual impactor. *Journal of Atmospheric and Oceanic Technology* 14, 197–202.
- Twohy, C.H., Strapp, J.W., Wendisch, M., 2003. Performance of a counterflow virtual impactor in the NASA Icing Research Tunnel. *Journal of Atmospheric and Oceanic Technology* 20, 781–790.
- Weinstock, E., Smith, J., Sayres, D., Spackman, J., Pittman, J., Allen, N., Demusz, J., Greenberg, M., Rivero, M., Solomon, L., Anderson, J.G., 2006. Measurements of the total water content of cirrus clouds. Part I: instrument details and calibration. *Journal of Atmospheric and Oceanic Technology* 23, 1397–1409.

Fog

PJ Croft, Kean University, Union, NJ, USA

B Ward, Public Works and Natural Resources, Longmont, CO, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by P J Croft, volume 2, pp 777–792, © 2003, Elsevier Ltd.

Synopsis

Fog occurs around the world and leads to many impacts on human and related systems. The formation of fog, its intensity, duration, and coverage are governed by synoptic and mesoscale forces, thermodynamics, and circulations. However, these are modified according to the local physiographic features in a region and vary with atmospheric chemistry and cloud microphysical processes. While the prediction of fog remains difficult due to its significant spatio-temporal variations, advances have been made through numerical modeling. Other uses and impacts of fog have increased with our understanding of the phenomenon and the parameters that determine fog characteristics and behaviors.

Fog as a Phenomenon

The occurrence of fog may lead to a multitude of impacts for a wide variety of activities around the world. These include personal inconvenience and simple annoyance due to delays (e.g., transportation) as well as more serious hazards such as air quality and local visibility. Sometimes these have deadly consequences as well as significant economic effects leading to delays, higher delivery costs, and spoilage of products and goods (e.g., a fog-bound ship near a port of call). While negative consequences are most often associated with fog, several positive benefits exist, including fog harvesting for local agricultural use and water supply. Fog has also been exploited from intelligence and military perspectives and ‘used’ in theatrical and similar productions and performances.

The complexity of fog impacts relates to its character as evident in the atmosphere: while fog may be associated with cleansing of the air it may also enhance contamination of the air by allowing chemicals and aerosols present to be retained. Fog consists of suspended droplets, often of similar size, near or at the ground, some of which may be settling out, scavenged, and/or evaporating, and restricts visibility and persists for a period of time. Fog droplets are characteristically larger than cloud droplets and have diameters near those of drizzle. The droplets are liquid in form and may also consist of supercooled droplets. Fog is often measured according to its restriction to visibility (intensity), its depth (either of the fog layer or the fog ‘top’ relative to the ground surface), and the areal coverage. A great degree of variability in fog’s composition and character is common, even on the smallest of microscales (Table 1).

Although the duration of fog is of importance, there is no standard operational measure or comparative observation. Fog is often labeled by ‘type,’ which is typically based upon its formation process and/or maintenance mechanisms (e.g., radiative, upslope) and on occasion, its location or source (i.e., sea fog, valley fog, etc.). Many types derive from a combination of direct and indirect causes. The origin of the word fog is thought to be primarily from high latitude northern countries and was a term used to describe small/fine rain, sometimes with wind, obscuring visibility. The word is taken from the Greek (nephele) or Latin (nebula) for cloud.

Fog Impacts

World Meteorological Organization Meteorological Terminal Aviation Routine Weather Report (WMO METAR) observational criteria distinguish fog from haze based on visibility considerations, relative humidity, and content. Haze and fog may occur together and have nearly equivalent impacts, but in fog, restriction of visibility predominates. The prevailing conventions define fog by its ‘thickness’ (or ‘intensity’) according to the restriction of horizontal visibility (and sometimes slant range in aviation). Dense, moderate, and light fogs produce visibilities of less than 1 km, 1 to less than 5 km, and 5 to less than 11 km, respectively. These definitions rely on sight distance as a surrogate for measurement of droplet distributions and are exclusive of the occurrence of low stratus clouds. Various designs of transmissometers are used to measure restrictions to visibility; particularly those that limit sight to less than 1 km (Figure 1).

Reductions in visibility are a function of the droplet sizes present and their type, the relative proportion or distribution of

Table 1 Common fog ‘names’ are indicated by country (or language) of origin

<i>Fog name/definition</i>	<i>Origin/language</i>
La niebla	Coastal Chile/Spanish
Le brouillard (or brouiller)	French
Nebel	German
Megla	Slovene
霧	Chinese
안개	Korean
La niebla (a neblina)	Spanish
Tage	Danish
بابضل	Arabic
Dimma	Swedish
Megla	Polish
Mist	Dutch
霧	Japanese
Nevoeiro	Portuguese
Sumutusjärjestelmä	Finnish
Nebbia	Italian
Tåke	Norwegian



Figure 1 Fog may occur in a wide variety of settings with multiple impacts. Reduced or obstructed visibility is typically the most significant impact of fog. The obstruction is dependent upon the drop-size distribution present as well as the concentration of drops and liquid water content. Impacts vary according to application such as runway visibility for aircraft operations. Pictures: <http://www.publicdomainpictures.net/view-image.php?image=7588&picture=docked-boats&large=1#large>; <http://www.publicdomainpictures.net/view-image.php?image=8903&picture=fog&large=1#large>; Schematic images: prepared by Braden Ward.

these, and their concentrations. It is a combination of fog ingredients, processes, and their interaction with characteristics of the local region that modify the physical chemistry (and sometimes the physical behavior of fog) in the local environment, thus leading to impacts. These are acted upon by radiative and advective processes within the planetary boundary layer from the micro to meso to synoptic scale. Together these create an intricate mix of cloud chemistry that is governed by cloud microphysical constraints and thermodynamic pathways. Given the large variation in scales involved, fluxes and interaction between scales (particularly within the roughness and surface layers of the boundary layer), and the ever-present temporal and spatial changes in a region, the occurrence of fog and its intensity and coverage (including its persistence and duration) will vary widely even when a fairly static, uniform, and homogeneous setting appears to exist in the atmosphere. These frequently lead to varying manifestations of fog in spite of seemingly identical conditions across a region or even at a single location.

While fog impacts on transportation (land, sea, and air travel) are commonly known, those that impact rescue operations,

emergency management, and air quality are less often considered with regard to the fog dynamic and this physical chemistry. Historically several important episodes are noteworthy including the Great Smog of 1952 in London, England, and the 1948 Donora Smog in Donora, Pennsylvania, United States – particularly since these occurred in fog-prone regions. This led to a diminished sense of danger or urgency by the local populations and thus resulted in significant loss of life as well as long-term health effects.

During the Great Smog of 1952, the most common heating method used was of low quality coal, which produces high amounts of sulfur dioxide and particulate matter, which can enhance droplet nucleation processes. The accumulation of these pollutants in the air gave London fogs a characteristic color – and led to the use of the phrase ‘pea soup’ to describe the London fog. The Great Smog lasted 4 days and resulted in many deaths and illnesses. The Donora Smog event of 1948 was similar in terms of duration and pollutant-based hazards. In this case two manufacturing plants of steel, wire, and zinc were responsible for emissions that included hazardous fluorine gas. A persistent temperature inversion in the region

restricted mixing and ventilation of the atmosphere and thus sickened thousands of people and resulted in hundreds of deaths during and after the event.

These two events clearly illustrate the difficulty of an emergency response to widespread respiratory illness within a regional population center. This includes difficulty in counting casualties, implementing emergency evacuations, and responding to the sources and cause in an operational and forecast manner. While the resulting knowledge gained from the Donora Smog and from London experiences led to the creation of more stringent environmental and air quality laws, they did not necessarily provide guidelines for preventive actions and mitigation measures that could be considered. In spite of rapid communications and modeling present today, the prospect of fog episodes creating 'bad air quality events' poses a serious threat to many communities globally whether near an industrial region or not.

In addition, little work has been accomplished in better understanding the impacts of fog on a variety of biological populations (i.e., plants, animals, insects, etc.) or those found in marine and estuary settings. These include fog deposition, surface reaction, and chemical ingestion/reaction. Since fog droplets carry chemical matter, or are found attached to various aerosols in the atmosphere (including those that may be radioactive), the surfaces, which they contact will be vulnerable to contamination and related hazards. This suggests a variety of legal issues beyond basic litigation (e.g., as related to motor vehicle accidents) including fog water collection, its safety and distribution. In order to resolve these legal and related issues there is the need for operational observations, predictions, and use of numerical models in real time and even from a hindcast perspective.

Fog Ingredients

Fog formation occurs for a wide variety of combinations of thermodynamic and radiative conditions in the atmosphere most of which are governed by synoptic scale forcing. Local variations in fog are driven primarily by local boundary layer processes and differences in microscale environments. Sufficient moisture and processes of cooling and/or lifting (inclusive of mixing) are essential for the formation of fog. These suggest three basic 'types' or categories of fog: radiational (cooling), advective (cooling and/or lifting), and combinatorial (cooling and lifting, and/or mixing).

The most critical factor for fog to exist and be sustained (as well as to be 'intense' or 'thick') is the presence of sufficient moisture, in terms of total amount and depth, while its horizontal distribution determines coverage. While sufficient moisture may also be achieved by increasing its 'effectiveness,' in other words, making use of the moisture present and realizing it through cooling and/or lifting processes to cause condensation; moisture and its realization is very much a function of meso- and microscale conditions and variations. These in turn are modulated according to microphysical behaviors that are associated with cloud droplet processes.

The presence of moisture (both condensed and uncondensed) provides the setting for the action of physical chemistries and cloud microphysics. These create subtle differences

among fogs based upon their composition and equilibrium states and can be used to distinguish fogs that, while they share synoptic and thermodynamic processes, are very different in terms of droplet size, distributions, and concentrations (e.g., continental vs marine fogs or urban continental vs forest continental fogs). The differences are most often manifest in the drop-size distributions and concentrations. In the case of fog moving from one location to another, it presents an additional complication in terms of fog maintenance and evolution. This is true from a meso- and microscale perspective as well as a physical chemistry point of view.

Fogs are also typically observed to be associated with inherently stable atmospheres. This stability may precede or occur after fog formation and often increases with the advent of fog. Even fog that is associated with strong winds, as is the case with some advective fogs, occurs in relatively stable layers of the boundary layer, which have achieved – and maintained – saturation. Other relevant factors in fog formation which may be considered as secondary in nature include vertical and horizontal distributions of temperature and moisture, orographic effects, sources and sinks of moisture and heat, land use, and surface conditions.

Fog Dynamics

There are several means of cooling an air mass, or parcel of air, that may lead to fog formation. The most obvious and most prevalent (even in the presence of cloud cover) is the diurnal loss of heat by the earth's surface and atmosphere (i.e., radiational cooling). Other means include the cooling of an air mass from below, adiabatic cooling (or mixing), the cooling of an air mass itself due to radiational release, and the evaporational cooling of air due to precipitation through a dry air layer which may induce cooling to saturation and thus result in fog. Depending on the location, time of year, and moisture availability these cooling mechanisms may lead to fog formation with varying persistence and of varying extent and intensity.

Radiational cooling is primarily diurnal in nature and is maximized during the overnight and early morning hours with minimum air and surface temperatures often occurring at or near sunrise. Although the diurnal cooling process occurs year-round, it is favored during both dry and cold seasons of the year when low level moisture may be sufficient, relatively undisturbed, and the cooling period lengthy. The dryness of the atmosphere is most typically observed above the boundary layer and allows significant radiational losses through an open atmospheric window – even in the presence of middle or high-level clouds. Radiational fogs may be brief in duration (e.g., less than 1 h) or may last several hours and are often shallow. The depth and intensity of these fog events is a function of the cooling time, and extent, and amount of moisture available. Therefore these fogs tend to be quite variable spatially across a region. It is not unusual for such fogs to also initiate dew deposition in isolated locations (or riming in the cold season or in colder climates).

The other means of cooling are of varying importance to fog formation and duration. For example, the cooling of an air mass from below is favored in locations and seasons in which the active surface layer is frozen and/or snow covered or when it

experiences a greater albedo (e.g., fallow vs the vegetative growing season). Such fogs may form and persist for hours or days at a time and cover a relatively large area with significant, albeit varying, intensities. The cooling of an air mass itself due to radiational release is typically a slow process and an important factor for persistent fogs that are maintained for extensive periods of time (e.g., sea fogs persisting from hours to days). The extent may be great but the internal variations in coverage and intensity vary largely as a function of the interactions between the air mass and the underlying surface features (i.e., the active surface of the local boundary layer). Evaporational cooling caused by precipitation falling through a layer of dry air may be sufficient to lead to saturation and fog but is typically of short duration and of limited intensity. Such cases more often result in a lowered cloud ceiling that can persist given sustaining synoptic processes (e.g., cold air damming with overrunning precipitation into an initially very dry near-surface layer of air). In the case of a synoptic scale warm front, such fog may form and persist for several days and become quite extensive and intense with minimal local variations.

The second basic means of cooling air to achieve or sustain fog formation, or for realizing the effectiveness of the moisture present in a parcel of air, is through various lifting mechanisms. These include orographic lift, frontal lift, adiabatic ascent, and mixing. In many cases, these processes involve advection and thus give rise to advective fog formation and transport. Although this implies that there are many lifting situations in which fog may form, it is clear that most of these situations involve slow vertical lifting over large horizontal distances or the relatively slow and shallow vertical mixing of two distinct air masses in the boundary layer. In the former case long lasting, extensive, and intense fogs may be expected whereas in the latter short term, shallow, and patchy fogs of varying intensity occur.

Slow vertical lift due to an upslope wind flow, parallel with the elevation gradient, will result in discrete levels of cooling and saturation with increasing distance and transport. Although this process may be slow in the initial formation of fog depending upon the amount of moisture available in the air mass (e.g., from several to nearly 24 h), it is a resilient process that can produce extended events of widespread dense fog (i.e., up to several days). Similarly, frontal lifting may produce conditions that persist for some time depending upon the rapidity of changes in synoptic features. Frontal lift is more commonly warm in nature but may involve cold frontal surfaces, which are of lesser slope than a typical cold front and/or a persistent quasi-stationary front. In both orographic and frontal cases the formation, duration, extent, and intensity of fog events is also a function of the underlying surface and its interaction with the lifted air. For example, the flow of warm and moist air across frozen or snow-covered ground – or simply upslope – increases the depth, intensity, duration, and duration of fog.

This last process is an important aspect and illustrates how two diverse air masses, initially unsaturated, may mix to form a saturated air mass. The use of saturation vapor pressure curves can be made to compare air mass properties as a function of their vapor pressures vs the absolute saturation vapor pressure for various temperatures and pressures. When a cool air mass

with relatively lower vapor pressure combines with a warm air mass of higher vapor pressure, their mixing results in saturation and thus fog. This may be seen by plotting the original vapor pressures at the individual air mass temperatures and connecting the two points with a straight line. When the line crosses the saturation vapor pressure curve, the two mixed air masses will form a saturated air mass. The manner in which these two air masses combine may be through isobaric mixing or weak adiabatic mixing. The process may also be modulated by the type and size distribution of nuclei, droplet activation, and liquid water content, in conjunction with the temperature and moisture regimes.

Lift that involves the adiabatic ascent and mixing of air is greatly dependent upon the existing boundary layer, which evolves during the mixing process to produce fog. Although of limited extent, turbulent mixing through adiabatic ascent can result in fog formation, which is typically of very short duration (i.e., less than a few hours), limited depth (e.g., ground fog), and highly variable in coverage and intensity. Such fogs may occur preceding and following the passage of weak cold fronts with limited pressure and air mass differences, and often following the passage of scattered showers or light rain, and take place in a conditionally stable boundary layer. These fogs tend to be infrequent and of short duration as the dynamics are more likely to lead to low cloud (and ceiling) formation with drizzle (although this is also determined by the nucleation process and varying drop-size distributions). In some cases these fogs may persist and thicken as the frontal boundary decays and/or becomes stationary and ingests local cloud condensation nuclei that may broaden or narrow the drop-size distribution. The processes of adiabatic ascent and mixing also play a role in the formation of Arctic sea smoke and other fogs in which the heat flux is rapid and results from temperature differentials rather than a period of radiational cooling.

Fog Evolution

Based on the preceding discussion it is clear that there are many possible synoptic, dynamic, and microphysical combinations, which may produce fog. It is therefore understandable why so many 'fog types' occur in the literature, have so much internal variability, and are studied around the world. For the same reason, it is clear that these possibilities raise the question of whether fog is readily predicted and whether one type is readily identified over another or whether one type may evolve into another. Since the orographic and frontal lifting processes are typically a gradual cooling process over long distances whereas, the radiational cooling process is gradual over time and specific to a location it is reasonable to consider various combination fogs in the same manner. It is also reasonable to incorporate the effects of cloud microphysics, the vertical and horizontal distribution of temperature and moisture, sources and sinks of moisture and heat, and land use and/or surface conditions with regard to how they modulate fog processes.

For example, given the features described above the longest lasting, most intense, deepest, and potentially most widespread fogs may occur near a coastal region with a moist onshore flow in the vicinity of a warm frontal (or topographic) or quasi-stationary boundary. This would be further enhanced or

avored if the flow of moisture was sustained, the ground frozen or snow covered (and thus the source of cooling maintained), and it was the cool season time of year. The formation and advection of sea fog – typically with a relatively larger size droplet distribution, tend to meet these criteria to varying degrees around the world and create some of the foggiest regions known. Regardless of origin, that is whether the sea fog formed first through radiational cooling or other cooling and lifting processes, it is clear that a variety of factors produces and sustains fog. This is verified by observation of the movement of fog areas and their passage from water to land.

Supporting the underlying causes of formation and growth due to cooling and lifting are the synoptic and mesoscale conditions present in the local atmosphere and their variations over time: diurnally and with dynamic evolution. These variations may occur from timescales of minutes and hours to days and thus pose both difficulties and dilemmas in all ranges of forecasting. The leading synoptic and mesoscale conditions responsible for fog formation, growth, and maintenance also interact with, respond to, and provide feedback into the local physiographic environmental features whether of natural (e.g., a region's changes in land surface cover) or of human origin (e.g., parking lots or garages). These features also complicate the physical chemistry of fog through varying radiative effects and contributions to (or deletions from) condensation nuclei (i.e., type, size distributions, and concentrations).

Synoptic and mesoscale patterns, in conjunction with the local physiographic environment, thus provide key fog mechanisms and ingredients in association with radiative, advective, and mixing processes that lead to the observed frequencies, intensities, durations, and locations of fog worldwide. These are naturally favored by select global circulation features and thus are often found as part of regional and local climate phenomena that describe the prevailing fog distribution, coverage, and occurrence. They are further enhanced – or modulated – according to a region's microphysical contributions. Thus on a global scale certain regions may be identified as prone to fog or not and with regard to fog 'severity' in terms of intensity, duration, and coverage in order to assign a risk factor or assessment. It is however important to understand that even locations that are not fog prone can be susceptible to fog impacts due to high density populations, commerce, and/or industry. Short duration, uncommon, and/or infrequent fogs may produce high impact events with significant repercussions particularly if they coincide with significant activities of a local population.

The foregoing discussion provides a 'family of fogs' in terms of formation, extent, duration, and intensity that may be expressed through conceptual models of fog dynamics. Although generous amounts of research have been accomplished and numerous modeling studies completed to reveal more explicitly the cooling and lifting processes that may produce fog, they are incomplete due to the limited integration of models that provide for improved understanding of the interaction among synoptic-dynamic, thermodynamic, physical chemistry, and microphysics. There is also the need to improve considerations of the interactions and interface between the underlying surface (and the local boundary layer's active surface) over which fog forms and the physical behaviors of fog as related to synoptic-, meso-, and microscale dynamic and thermodynamic processes in relation to local

temperature and moisture distributions. These are critical in the horizontal and vertical dimensions for better understanding and implementation of fog modification and dispersal techniques.

Fog Prediction

The advent of a more scientific and structured observational (or empirical) study of fog followed cloud classification schema and the development of more sophisticated instrumentation for study of the atmosphere, particularly on more localized scales. This phase of exploration led to improved understanding of the observed properties and behaviors of fog based on weather observation data and fog's associated characteristics as a function of dynamic processes in the atmosphere. This linkage afforded the first efforts at fog prediction in order to provide opportunity for the avoidance, mitigation, or prevention of its impacts.

Further development of these methods has included statistical and time series analyses to produce simple conditional climatologies of fog occurrence, intensity, and tendencies; as well as Markov Chain and other analytic techniques to more accurately assess fog occurrence and probability distributions for specific locations or regions (and intensity). In any of these cases, the specificity of the information is limited to the period of record and how representative each location is with regard to its surrounding area. Further specification of fog occurrence is possible when these techniques are linked to the synoptic setting or when statistical or climatic values are mapped across a region in order to infer the frequency of occurrences, intensities, and fog coverage or behavior.

However, while such findings may be applied in a predictive (or diagnostic) and deterministic manner they are limited according to an observational framework and timescale (e.g., hours or days) without full consideration of the process-oriented nature of fog formation, maintenance, evolution, and dissipation. As the observational study of fog has matured, the study of cloud physics and the microphysical processes involved in cloud growth provided distinct thermodynamic information as to the internal mechanisms and factors associated with fog formation (and dissipation) as well as its radiative impacts (as recognized many years before). This much more comprehensive insight of fog as an 'entity' phenomenon to observe and follow was also assisted by the modern development of more sophisticated instrumentation (e.g., transmissometer) to measure cloud (or fog) properties and to monitor subsequent development (or dissipation) for shorter timescales and across shorter distances.

These help to establish the sensing of fog's presence and the impact of fog intensity according to visibility criteria as linked to arbitrary interrogations vs one based on more defined impacts or meaningful thresholds. This provided more distinctive identification of fog and its variations in terms of visibility, or depth/thickness, beyond simple measurements of visual range according to predetermined target ranges. Microphysical studies have also provided sufficient evidence to explain variations in fog behavior based upon its phase composition (i.e., liquid, solid, or mixed) and its subsequent evolution. Fog drop-size distributions have been determined to

range from 1 to 10 μm according to the aerosol (or particulate) matter and the availability of moisture (e.g., mixing ratio of the atmosphere) and condensational processes. Larger sizes are preferred as fog increases during condensation processes while smaller drops dominate in dissipation.

While these also have obvious synoptic signatures as found through observational studies, they are also driven by interactions between fog and the local landscape, including local sources and sinks acting on the micro- and mesoscale. These act to contribute to locally preferred drop-size distributions with specific precipitation or deposition (or terminal velocity) tendencies. An understanding of the dependence and interaction of fog on local features has been aided by air quality and atmospheric chemistry studies (for near-surface locations) as well as the emergent air-surface and air-sea studies (over the last 20 years) that focus on boundary layer exchanges and flux behaviors. These allow a more precise explanation of fog's physical and chemical characteristics (or the statistical family of these), which modulates evolution of fog in time and space – and that create regional and local differences. Due to the physical chemistry associated with fog events, their electrical, radiative, optical, and acoustical properties also necessarily show variation.

Numerical modeling has been used more recently to demonstrate the practicality of predicting fog – and/or its impacts on visibility (or sky condition) – through surrogate parameters in a deterministic manner. For example, threshold values of mixing ratio, relative saturation levels, and cloud liquid water content have been successful in depicting fog occurrence and to a lesser extent fog coverage and intensity. Limitations to operational models have been the inability to properly portray the meso- and microscale environments as well as integrating the planetary boundary layer's characteristics and behaviors with the larger scale environment while accounting for radiative and other transfers and fluxes. Some work has been fruitful in combining these with land cover and land use information (i.e., from a Geographic Information System (GIS) perspective) to provide a more realistic portrayal of fog coverage in a region. The higher resolution examination poses additional

questions with regard to forecast validation and verification. Much work remains in order for these efforts to be reliable and timely and the use of more probabilistic approaches is needed.

In support of the avoidance, mitigation, or prevention of fog hazards, a variety of operational support and decision-making tools and information is available to forecasters as well as the affected and responding communities (e.g., aerodromes, emergency managers, or similar personnel). In some cases these are linked to, or provide data ingest for, numerical or statistical guidance packages. These are particularly effective when tied to GIS databases and decision-support software (or artificial intelligence) and are more commonly used in a military- or disaster-related type of response. The additional deployment of meso- and microscale surface-based observing networks will serve to increase the spatial and temporal acuity of data as related to fog occurrence and its evolution (and study) so that it may be more accurately detected, assessed, and compared with forecasts – particularly those based on digitized data sets (Figure 2).

Satellite and similar remote sensing platforms offer a variety of products (e.g., the 'fog product') that provide a gross estimate of fog occurrence, intensity, and coverage by channel differencing as well as through examination of sounder data to construct vertical and near-surface profiles of temperature and moisture in the atmosphere. While useful, these products only provide information after fog has formed and thus allow tracking its movements and evolution. While other parameter fields are available through satellite imagery, they provide neither adequate nor precise estimates of fog precursors for predictive purposes. Although microwave sensors and ground-based radar, lidar, and profiler platforms may offer additional information and operational support, none are presently suited or dedicated to fog detection or prediction. Additional information on atmospheric chemistry and structure is now available through several remote sensing platforms and these will provide a real-time observation profile of the physical chemistry of the atmosphere as related to fog and other phenomena or conditions in the atmosphere.

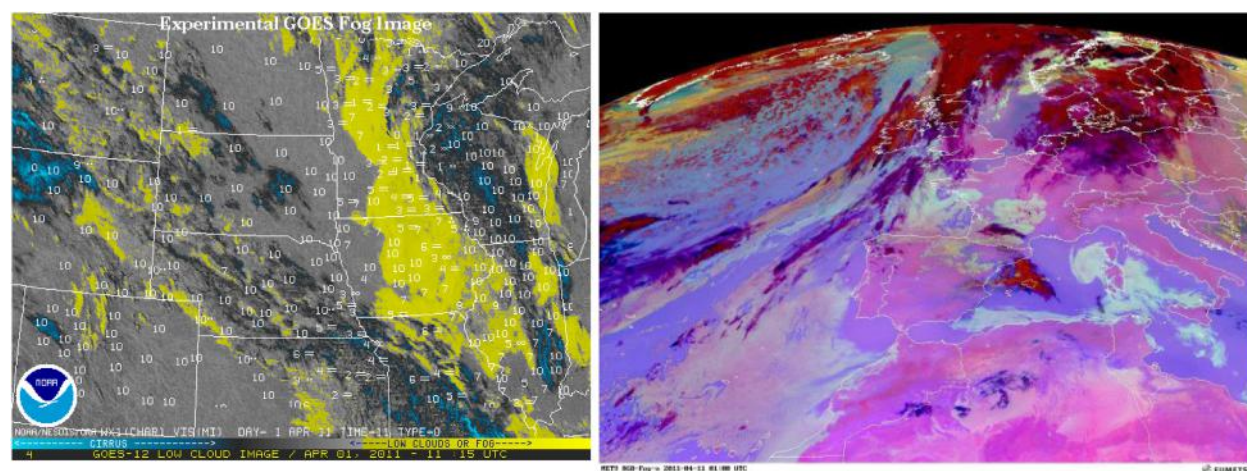


Figure 2 Satellite 'fog product' provides an observational basis for fog coverage, intensity, and duration that may aid operational prediction and nowcasting. Samples are shown from the National Environmental Satellite, Data, and Information Service – NOAA/NWS (USA); and the European Organization for the Exploitation of Meteorological Satellites (United Kingdom).

Satellite, derived remote sensing products, and aerial imagery are essential resources when analyzing fog for development, coverage, and intensity. Indeed, given the often limited spatial coverage of fog, satellite products that can analyze the temperature ranges of cloud tops can effectively accentuate an area of fog within a region of varying cloud heights and type. The higher resolution platforms continue to better specify these parameters in space and time. However, these observational resources cannot be solely relied upon for fog depiction and analysis, as fog is not accurately observed during daylight hours or when specific weather events or synoptic scale features produce upper cloud decks over areas of fog.

Fog Research and Applications

Presently, fog is known to be formed through three basic mechanisms each of which is found to a lesser or greater extent in the creation of each fog type or classification. The basic processes are radiative (cooling), advective (cooling and/or lift), and mixing (thermodynamic). Making use of the historical empirical evidence, statistical evaluations, and the principles of physical chemistry, researchers and applications specialists have used this information to enact two basic types of mitigation or prevention strategies with regard to fog hazards and a third option for use in other cases: (1) fog dispersion, (2) air quality management, and (3) special cases.

In the first case fog dispersion has focused on relatively costly mechanical means (i.e., bulk mixing) as opposed to thermodynamic and physical methods (i.e., heating and/or seeding) that have been applied predominantly at aerodrome locations. For air quality issues, emphasis has been placed on reduced exposure (e.g., avoidance by remaining indoors or masking through the use of personal filters), reduction of contributing sources, or ventilation (similar to bulk mixing) and is often a function of the population affected (i.e., receptors) and the resources required for cost-effective implementation and the likely effectiveness.

The third aspect (special cases) typically offers a mix of the foregoing methods, requires other techniques, or is not well known. These would include situations of ice or acid fog, fog water collection, smog, or fogs occurring with smoke and haze (e.g., forest fires). Related to these anthropogenic issues are the creations of artificial fogs by cooling towers and other industrial processes (through the introduction of moisture or aerosols); or those fogs made for use 'on stage' or for entertainment purposes. While most of these make use of simple ventilation techniques or equipment, these fog environments are often complicated due to the presence of other airborne aerosols which may be more readily deliquesced and thus affect the local population.

Impacts may range from ear, eye, nose, or throat ailments to chronic exposure that may need to be treated on an individual basis. Secondary impacts of fog include wet or dry deposition on plants, buildings, statues, and other surfaces causing damage due to chemical interactions (e.g., scalding). Additional fog techniques, applications, and research are those linked to fog as a water resource and the use of fog as a tactical advantage (military). Water consumption and conservation are

particularly important in less developed nations as well as in dry climates. Fog as a 'weapon' in war has been associated with creating or dispersing fog as well as artificial enhancements – including attempts at physical–chemical changes meant to attenuate radar, communication, and other signals in the battlefield – or to deliver specific aerosols, reactive agents, or gases to a target area.

Other recent scientific fog investigations have focused on the physical chemistry involved in cloud (and thus fog) processes as well as techniques for fog dissipation or dispersal based on the foregoing knowledge of fog's physical, chemical, radiative, and internal properties. Other work has been focusing on the chemical mixture of species present in fog in terms of its impacts on human (or animal/plant) respiration (e.g., acute or chronic; mortality), external contact (e.g., surface of the skin or leaf contact), and ingest (e.g., mucous membranes or eyes; fog water consumption or irrigation). In tandem with prior empirical and statistical study these are assisting in the development of a more comprehensive conceptual model of fog in terms of its characteristics, physical and chemical attributes, and behaviors.

Together these provide for integrated numerical modeling of fog with regard to its formation mechanisms, maintenance, intensity, evolution, and dissipation – particularly as related to fog's physical chemistry and its attendant optical and radiative effects. The ability to quantify these will provide an enhanced ability to model the interactions between atmospheric systems producing fog and the systems impacted by fog (i.e., human and other populations). To best represent and visualize such systems and their interactions current work has focused on artificial intelligence, impact-response models, and the use of GIS systems. It is reasonable to expect applications in catastrophe modeling in select or worst-case scenarios for real-time response and evaluation purposes.

Less obtrusive are connections that have been made with fog in the popular (and even classic) arts and literature. These include music, films, poems, paintings, and other creative works. Fog is often used to convey mystery or represent an enshrouding of truth or information that is to be concealed for often nefarious purposes. In fact, some of the arts dedicated or connected with the fog phenomenon offer clues to climate, climate variability, and climate change. These indicators can be helpful when considering climate modeling studies used to predict variations in frequencies and intensities of atmospheric phenomena such as fog to assess variations in potential or expected impacts for local communities and economies.

Fog has been observed to shape the local cultural manifestations of some societies as well as religious practices around the world. While there has been limited study of the impact of fog with regard to sociology and psychology, there is evidence that atmospheric conditions similar to fog (e.g., persistent cloudiness) do cause and/or relate to the emotional, psychological, and spiritual well-being of people. Fog may also contribute to additional health effects as related to pollen, mold and mildew, as well as disease or virus transmission. There have been suggestions that fog may pose transmission problems for high-tension power lines and interfere with reception of electronic signals.

See also: Boundary Layer (Atmospheric) and Air Pollution: Air Pollution Meteorology. Climate and Climate Change: Climate Prediction: Empirical and Numerical; Overview. Dynamical Meteorology: Coriolis Force. Hydrology, Floods and Droughts: Overview. Inadvertant Weather Modification. Land-Atmosphere Interactions: Overview. Mesoscale Meteorology: Overview. Numerical Models: Methods. Synoptic Meteorology: Forecasting; Weather Maps. Thermodynamics: Moist (Unsaturated) Air; Saturated Adiabatic Processes.

Further Reading

- Achtemeier, G.L., 2008. Effects of moisture released during forest burning on fog formation and implications for visibility. *Journal of Applied Meteorology and Climatology* 47, 1287–1296.
- Agarwal, M., Maze, T.H., Souleyrette, R., 2006. The weather and its impact on urban freeway traffic operations. In: Proc. 85th Annual Meeting of the Transportation Research Board. Transportation Research Board of the National Academies, Washington, DC, 06–1439.
- Air Weather Service, 1979. General Aspects of Fog and Stratus Forecasting. USAF AWS TR 239, 100 pp.
- Anderson, J.R., 1985. Economic impacts. In: Houghton, D.D. (Ed.), *Handbook of Applied Meteorology*. John Wiley and Sons, New York.
- Bankert, R.L., Hadjimichael, M., 2007. Data mining numerical model output for single-station cloud-ceiling forecast algorithms. *Weather Forecasting* 22, 1123–1131.
- Bendix, J., 2002. A satellite-based climatology of fog and low-level stratus in Germany and adjacent areas. *Journal of Atmospheric Research* 64, 3–18.
- Bott, A., Trautmann, T., 2002. PAFOG: a new efficient forecast model of radiation fog and low-level stratiform clouds. *Journal of Atmospheric Research* 64, 191–203.
- Boudala, F.S., Isaac, G.A., Crawford, R.W., Reid, J., 2012. Parameterization of runway visual range as a function of visibility: implications for numerical weather prediction models. *Journal of Atmospheric and Oceanic Technology* 29, 177–191.
- Cho, Y.K., Kim, M.O., Kim, B.C., 2000. Sea fog around the Korean peninsula. *Journal of Applied Meteorology* 39, 2473–2479.
- Cools, M., Moons, E., Wets, G., 2010. Assessing the impact of weather on traffic intensity. *Weather Climate and Society* 2, 60–68.
- Cox, R.E., 2007. Applying fog forecasting techniques using AWIPS and the internet. National Weather Association. *Electronic Journal of Operational Meteorology* FTT1.
- Croft, P.J., Pfost, R., Medlin, J., Johnson, G., 1997. Fog forecasting for the southern region: a conceptual model approach. *Weather Forecasting* 12, 545–556.
- Eagleman, J.R., 1991. *Air Pollution Meteorology*. Trimedia Publishing Company, p. 255.
- Ellrod, G.P., Lindstrom, S., 2006. Performance of satellite fog detection techniques with major, fog-related highway accidents. *Electronic Journal of Operational Meteorology* 7 (3), 1–10.
- Fisher, E.L., Caplan, P., 1963. An experiment in numerical prediction of fog and stratus. *Journal of Atmospheric Sciences* 20, 425–437.
- Fuchs, W., Schickel, K.-P., 1995. Aircraft icing in visual meteorological conditions below low stratus clouds. *Atmospheric Research* 36, 339–345.
- Garriott, E.B., 1904. Forecasts and warnings. *Monthly Weather Review* 32, 547–549.
- Geiger, R., 1965. *The Climate near the Ground*. Harvard University Press, p. 611.
- George, J.J., 1951. Fog. In: Malone, T.F. (Ed.), *Compendium of Meteorology*. Am. Meteorol. Soc., pp. 1179–1189.
- George, J.J., 1960. *Weather Forecasting for Aeronautics*. Academic Press, p. 673.
- Glahn, H.R., Ruth, D.P., 2003. The new digital forecast database of the National Weather Service. *Bulletin of the American Meteorological Society* 84, 195–201.
- Gultepe, I., et al., 2009. The fog remote sensing and modeling field project. *Bulletin of the American Meteorological Society* 90, 341–359.
- Hansen, B., 2007. A fuzzy logic-based analog forecasting system for ceiling and visibility. *Weather Forecasting* 22, 1319–1330.
- Hardwick, W.C., 1973. Monthly fog frequency in the continental United States. *Monthly Weather Review* 101, 763–766.
- Heintzenberg, J., Leck, C., Birmili, W., Wehner, B., Tjernstrom, M., Wiedensohler, A., 2006. Aerosol number-size distributions during clear and fog periods in the summer high Arctic: 1991, 1996 and 2001. *Tellus B* 58, 41–50.
- Heo, K.-Y., Ha, K.-J., 2010. A coupled model study on the formation and dissipation of sea fogs. *Monthly Weather Review* 138, 1186–1205.
- Hilliker, J.L., Akasapu, G., Young, G.S., 2010. Assessing the short-term forecast capability of nonstandardized surface observations using the National Digital Forecast Database (NDFD). *Journal of Applied Meteorology and Climatology* 49, 1397–1411.
- Hoover, R.A., 1950. Forecasting radiation fog at Elkins, W. VA. *Monthly Weather Review* 78, 75–80.
- Houghton, H.G., 1985. *Physical Meteorology*. The MIT Press, p. 442.
- Houze Jr., R.A., 1993. *Cloud Dynamics*. Academic Press, Inc., p. 573.
- Hranac, R., Sterzin, E., Krechmer, D., Rahka, H., Farzaneh, M., 2006. Empirical Studies on Traffic Flow in Inclement Weather. Rep. FHWA-HOP-07–073. Federal Highway Administration, U.S. Department of Transportation, Washington, DC, p. 114.
- Jacob, D.J., Waldman, J.M., Munger, J.W., Hoffmann, M.R., 1984. A field investigation of physical and chemical mechanisms affecting pollutant concentrations in fog droplets. *Tellus B* 36B, 272–285.
- Johnson, J.C., 1954. *Physical Meteorology*. The Massachusetts Institute of Technology and John Wiley & Sons, Inc., p. 393.
- Klemm, O., Wrzesinsky, T., 2007. Fog deposition fluxes of water and ions to a mountainous site in Central Europe. *Tellus B* 59, 705–714.
- Kulshrestha, M.J., Sekar, R., Krishna, D., Hazarika, A.K., Dey, N.C., Rao, P.G., 2005. Deposition fluxes of chemical components of fog water at a rural site in north-east India. *Tellus B* 57, 436–439.
- Lamb, H.H., 1977. *Climatic History and the Future*. Princeton University Press, Princeton.
- Larson, V.E., Schanen, D.P., Wang, M., Ovchinnikov, M., Ghan, Steven, 2012. PDF parameterization of boundary layer clouds in models with horizontal grid spacings from 2 to 16 km. *Monthly Weather Review* 140, 285–306.
- Lee, T.F., Turk, F.J., Richardson, K., 1997. Stratus and fog products using GOES-8-9 3.9- μ m data. *Weather Forecasting* 12, 664–677.
- Leipper, D.F., 1994. Fog on the U.S. west coast: a review. *Bulletin of the American Meteorological Society* 75, 229–240.
- Leipper, D.F., 1995. Fog forecasting objectively in the California coastal area using LIBS. *Weather Forecasting* 10, 741–762.
- Mason, J., 1982. Physics of radiation fog. *Journal of the Meteorological Society of Japan* 60, 486–499.
- Mayer, W.D., Rao, G.V., 1999. Radiation fog prediction using a simple numerical model. *Pure Appl. Geophys.* 155 (1), 57–80.
- NOAA/NWS, January 3, 2005. *National Weather Service Strategic Plan for 2005–2010 – Working Together to Save Lives*, p. 29.
- Nowak, D., Ruffieux, D., Agnew, J.L., Vuilleumier, L., 2008. Detection of fog and low cloud boundaries with ground-based remote sensing systems. *Journal of Atmospheric and Oceanic Technology* 25, 1357–1368.
- Oliver, J.E., 1973. *Climate and Man's Environment: An Introduction to Applied Climatology*. John Wiley and Sons, New York.
- Olivier, J., 2004. Fog harvesting: an alternative source of water on the west coast of South Africa. *GeoJournal* 61, 203–214.
- Pagowski, M., Gultepe, I., King, P., 2004. Analysis and modeling of an extremely dense fog event in southern Ontario. *Journal of Applied Meteorology* 43, 3–16.
- Palmer, A.H., 1917. Fog along the California coast. *Monthly Weather Review* 45, 496–499.
- Peace, R.L., 1969. Heavy-fog regions in the conterminous United States. *Monthly Weather Review* 97, 116–123.
- Pilié, R.J., Mack, E.J., Rogers, C.W., Katz, U., Kocmond, W.C., 1979. The formation of marine fog and the development of fog-stratus systems along the California coast. *Journal of Applied Meteorology* 18, 1275–1286.
- Proctor, F.W., 1904. A new theory of fog formation. *Monthly Weather Review* 32, 406–411.
- Roquelaure, S., Bergot, T., 2008. A local ensemble prediction system for fog and low clouds: construction, Bayesian model averaging calibration, and validation. *Journal of Applied Meteorology and Climatology* 47, 3072–3088.
- Rudack, D.E., Ghirardelli, J.E., 2010. A comparative verification of localized aviation model output statistics program (LAMP) and numerical weather prediction (NWP) model forecasts of ceiling height and visibility. *Weather Forecasting* 25, 1161–1178.
- Schemenauer, R.S., Bridgman, H. (Eds.), July 19–24, 1998. *Proceedings First International Conference on Fog and Fog Collection*. Vancouver, British Columbia, Canada.
- Tardif, R., Rasmussen, R.M., 2007. Event-based climatology and typology of fog in the New York City region. *Journal of Applied Meteorology and Climatology* 46, 1141–1168.
- Tardif, R., Rasmussen, R.M., 2008. Process-oriented analysis of environmental conditions associated with precipitation fog events in the New York City region. *Journal of Applied Meteorology and Climatology* 47, 1681–1703.

- Underwood, S.J., Ellrod, G.P., Kuhnert, A.L., 2004. A multiple-case analysis of nocturnal radiation-fog development in the Central Valley of California utilizing the GOES nighttime fog product. *Journal of Applied Meteorology* 43, 297–311.
- Underwood, S.J., Hansen, C., 2008. Investigating urban clear islands in fog and low stratus clouds in the San Joaquin Valley of California. *Physical Geography* 29, 442–454.
- Unger, E.E., 1923. Dense fog in the tri-cities on November 3, 1922. *Monthly Weather Review* 51, 81–82.
- Ward, B., Croft, P.J., 2008. Use of GIS to examine winter fog occurrences. *National Weather Association Electronic Journal of Operational Meteorology* EJ4.
- Westcott, N.E., Kristovich, D.A.R., 2009. A climatology and case study of continental cold season dense fog associated with low clouds. *Journal of Applied Meteorology and Climatology* 48, 2201–2214.
- Yang, D., Ritchie, H., Desjardins, S., Pearson, G., MacAfee, A., Gultepe, I., 2010. High-resolution GEM-LAM application in marine fog prediction: evaluation and diagnosis. *Weather Forecasting* 25, 727–748.
- Zhang, S.-P., Xie, S.-P., Liu, Q.-Y., Yang, Y.-Q., Wang, X.-G., Ren, Z.-P., 2009. Seasonal variations of Yellow Sea fog: observations and mechanisms*. *Journal of Climate* 22, 6758–6772.
- Zhou, B., Du, J., 2010. Fog prediction from a multimodel mesoscale ensemble prediction system. *Weather Forecasting* 25, 303–322.

Relevant Websites

- <http://www.cco.net/~trufax/fluoride/fog.html>.
- <http://www.fogconference.org/>.
- <http://www.fogquest.org/index.php/home/>.
- <http://meted.ucar.edu> – (COMET Module Radiation Fog – requires registration to view/use).
- <http://outofthefog.ca/>.

Noctilucent Clouds

GE Thomas, University of Colorado, Boulder, CO, USA

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

Noctilucent clouds are the highest clouds on Earth, and are located in the coldest part of planet Earth, the high-latitude summertime mesopause region (80–90 km). Composed of water-ice particles, they are nucleated by meteoric smoke particles. This article discusses brief history since their discovery in 1885, occurrence patterns, the nature of the cold mesopause, methods of active and remote sounding, polar mesospheric summertime radar echoes, theory and modeling of mesospheric ice, and long-term changes of mesospheric ice clouds, possibly related to upper atmospheric climate change.

Introduction

Noctilucent clouds ('night luminous' or NLC) are the highest and coldest clouds in the atmosphere. When viewed from the ground they are referred to as NLC. Viewed from space they are called polar mesospheric clouds (PMC). Occupying a narrow (81–86 km) height zone below the high-latitude mesopause (a temperature minimum vs height, located near 88 km), NLC offer a splendid sight during summer twilights ([Figure 1](#)). They are made visible by scattered sunlight against the dark twilight sky, when the sun lies below the horizon at angles between 6 and 16°. NLC are too faint to be seen during daytime. Visible in both summertime hemispheres, they extend poleward of latitude 50°, generally increasing in their occurrence and brightness toward the pole. Ground observation becomes impossible at the higher summertime latitudes (above approximately 70–75°). Knowledge of NLC properties at the higher latitudes relies upon satellite and light detection and ranging (lidar) measurements. Information regarding these remarkable clouds, existing at the very 'edge of space' has increased enormously since their discovery in 1885, and particularly since the

space age when scientists became aware of the extremely cold conditions in the mesopause region. They are important as sensitive indicators of temperature, humidity, and dynamics, and as possible harbingers of climate change.

Brief History

Mesospheric clouds were first recognized as an upper atmospheric phenomenon in the summer of 1885, when observers throughout Britain, Europe, and Russia reported unusual cloud displays enduring through twilight, and at the higher latitudes throughout the midnight hours. F. Leslie now is given the credit for the first published report, although during that summer, many observers (Jesse, Backhouse, Tserassky, and others) witnessed nocturnal cloud displays throughout northern latitudes. Their great altitude (~83 km) gave rise to speculations that they consisted of cosmic dust or cometary debris. It was also suggested that the great Krakatoa volcanic eruption of 1883 might have injected water vapor into the mesosphere. The 2-year lag between the time of eruption and the NLC appearances of 1885 is known to be consistent with the transport time of air from the upper stratosphere near 50 km (where much of the volcanic material was deposited) to the mesosphere. The Krakatoa event was unique, in that it took place on a small island off the coast of Java, where breaching of the crater rim by seawater caused superheated steam to add to the volcanic explosivity. Its water-rich plume contrasts with all other major eruptions in the modern era whose volatile content consists primarily of sulfates and carbon dioxide. An open question is whether NLC was a pre-existing phenomenon that was only recognized in 1885, or marked the debut of a new type of cloud resulting from increasing anthropogenic activities (see Section on [Long-Term Changes of Mesospheric Clouds](#)).

In 1912, A. Wegener (famous for his advocacy of the continental drift hypothesis) proposed that NLC are composed of water-ice crystals. An obstacle to this idea was the need for either extremely cold air or very high amounts of water vapor. Later on, balloon measurements in the stratosphere showed that the upper atmosphere is very dry, and for a time the ice hypothesis was in disfavor. The water-ice theory received critical support when rocket probes in the early 1960s revealed extremely cold air at high-latitude sites during summertime,



Figure 1 Photograph of an NLC taken from Stockholm, Sweden (59.37° N, 18.06° E), on 16 July 2005. Time: 23:29 UT. Taken with a Canon PowerShot G5. Exposure time: 4 s f/3.5, focal length: 50 mm. Provided by Jacek Stegman of the University of Stockholm, Sweden with permission. The dark clouds near the horizon are ordinary tropospheric clouds, no longer illuminated by direct sunlight.

sufficient for saturation despite the extreme aridity of the mesosphere. It was only in 2001 that satellite occultation spectroscopy confirmed the water-ice composition.

Beginning with the International Geophysical Year in 1957, networks of observers were established at locations in Europe, North America, and the Soviet Union to tabulate cloud frequency and motions and to photograph their structures. While many thousands of detailed observations became available, the interpretation of visual observations was limited until the 1960s by an ignorance of the properties of the cloud environment.

Nevertheless, these early data established the cloud seasonality, and the equatorward boundary of the 'NLC zone' in both hemispheres (which we now know is limited by warmer unsaturated air at lower latitudes). Also, their wavelike structures were among the first evidence of the presence of *gravity waves* in the upper atmosphere. Over the past four decades, ground-based data have established an apparent solar effect with approximately 10- to 11-year periodicity that has an overall inverse relationship to solar activity (conventionally described by the sunspot number). The influence of auroras, which may occur simultaneously with NLC, is of negligible importance, except during rare instances (auroral heating is confined generally to heights above 100 km). Visual and photographic observations continue today, and numerous groups monitor and archive their appearance, motions, and geographical occurrence. However, unveiling the physical nature of NLC awaited new lidar technologies and space-age observations of the properties of the previously inaccessible cloud environment.

Occurrence Patterns

Mesospheric clouds are monsoonal in character, generally existing only between mid-May and mid-August in the Northern Hemisphere, and at similar times of the year (relative to summer solstice) in the Southern Hemisphere. They are seen from all northern landmasses between 50 and 75° with a maximum occurrence rate typically ~40%. Occasionally they are witnessed at midlatitude, the 'record' being at latitude 38° in 1999, in Ignacio, Colorado, USA. In the south, there are few favorable locations, due to the scarcity of land in the NLC zone. The NLC season is about 70 days long, with peak occurrence a few weeks after summer solstice. The factors controlling their growth and decay are discussed in Section [Theory and Modeling of Mesospheric Ice Clouds](#).

The Modern Era: The Discovery of the Mesopause

Several developments led to a scientific understanding of, and renewed interest in mesospheric clouds: (1) realization that they are a possible sensitive proxy of upper atmosphere climate change; (2) discovery of intense radar echoes called polar mesosphere summertime echoes (PMSE); (3) developments in atmospheric lidar technology; and (4) the advent of rocket sounding and satellite remote sensing, revealing their global morphology and the unusual properties of the cloud environment.

A key discovery was that of the cold summertime mesopause (the temperature minimum that routinely occurs near 88 km; see [Figures 2](#) and [3](#)). This behavior is opposite to that occurring in the lower atmosphere, where summer is warmer than winter. This anomaly is a result of a general upwelling of air in the summer hemisphere (and a downwelling in the winter hemisphere) caused by the deposition of wave momentum. Atmospheric waves launched upward by disturbances in the lower atmosphere (e.g., storm fronts) influence the entire mesosphere (50–95 km), but particularly the polar environment of mesospheric clouds. Gravity waves, atmospheric *tides*, and *planetary waves* affect the thermal and dynamical states of the entire mesopause region. In their upward passage through the progressively thinner air, because wave energy tends to be conserved, the wave amplitudes grow with height. Due to the filtering action of the underlying seasonally varying winds, the waves that reach the mesopause also vary seasonally. Above about 80 km,

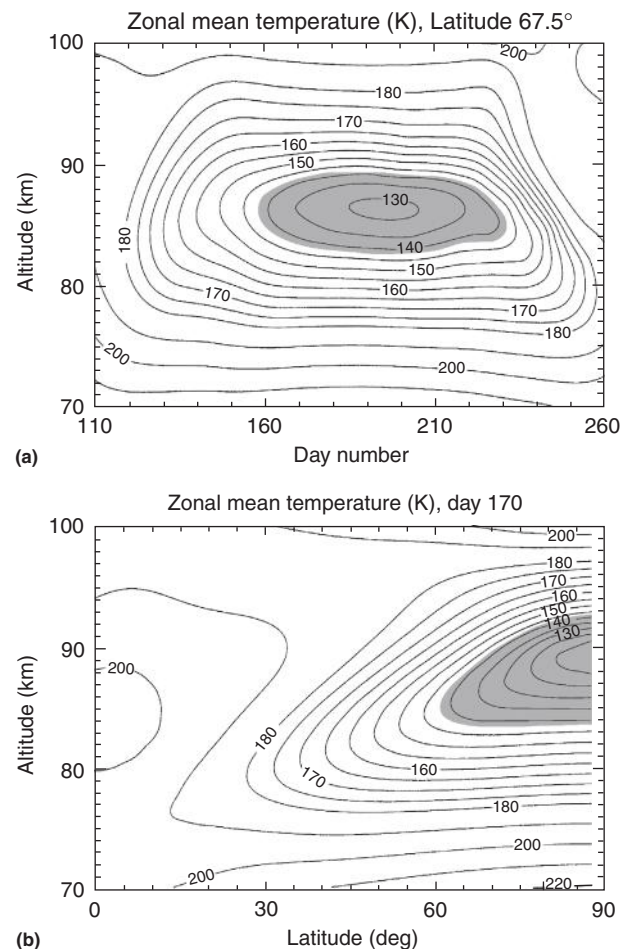


Figure 2 Isotherms in the 70–100-km atmospheric region, derived from a comprehensive 3D model (Thermosphere Ionosphere Mesosphere General Circulation Model) of the upper atmosphere: (a) plotted against day number and height at 67.5° N, and (b) plotted against latitude and height near northern summer solstice (day 170). Shaded areas emphasize saturation regions for water ice, i.e., regions colder than 142 K, where ice particles can exist at water vapor mixing ratios of 2 ppmv.

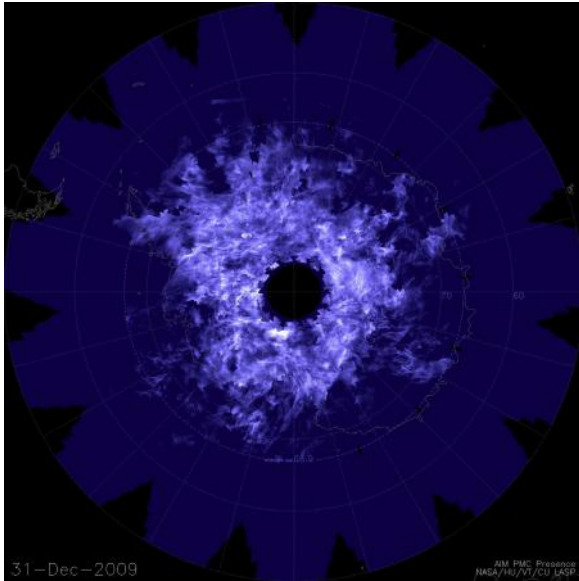


Figure 3 Composite of image strips of polar mesospheric clouds taken by the cloud imaging and particle size suite of UV cameras on the NASA Aeronomy of Ice in the Mesosphere satellite. Fifteen consecutive orbits of data are superimposed (separated by the 90-min orbital period), taken over 24 h on 31 December 2009. Where overlap occurs in the higher latitudes, the brightest cloud of the two strips is shown. Horizontal resolution: 5 km. Note that no data were collected in the dark region centered on the South Pole due to the satellite viewing geometry. Courtesy of the CIPS team at the University of Colorado, Boulder, CO, USA.

they often break down into turbulence, analogous to ocean breakers on a beach. The absorption of these waves causes mesospheric air to be forced equatorward in summer and poleward in winter. This seasonal forcing of the winds causes a chain of events, among them are a lifting/adiabatic cooling of the air in summer, a summer-to-winter circulation, and a sinking/adiabatic warming in winter.

The notion of seasonally and latitudinally varying *saturation* (Figure 2) is basic to the theory of ice formation (see Section [Optical Sensing of Mesospheric Clouds](#)). When air is sufficiently cold and moist, atmospheric ice is stable. When there are sufficient nucleation sites available and the air is *supersaturated*, the ice particles may grow by direct deposition of the surrounding water vapor. These ‘seed particles’ have been long believed to be meteor ‘smoke’ particles (MSPs), which result from condensation of vapors resulting from the atmospheric ablation of incoming interplanetary dust particles. Recent optical measurements from space provide strong support for this long-held hypothesis.

Rocket and Radar Sounding of the Upper Mesosphere

To gain a deeper understanding of the origin and evolution of ice particles, it is necessary to measure the thermal, chemical, electrical, and dynamical properties of their environment. This problem was first attacked through instrumented sounding rockets and continues today. Fortunately, there exist a number

of permanently staffed high-latitude launching sites that allow *in situ* measurements of the upper mesosphere. Rocket probes, radar sounding, and lidar backscattering have made it possible to measure (often with meter-scale resolution) a host of properties: wind, electron, and ion density, ionization state, particle size and density of both ice and mesospheric aerosols, charge state, etc.

It is important to mention the closely related phenomenon of PMSE. The discovery in the early 1980s of very high frequency radar echoes from the high-latitude summertime mesopause region added an extra dimension to NLC studies. Radar does not interact directly with ice particles, but with highly structured variations in the refractive index of the free electron density. The prevailing theory combines neutral air turbulence with reduced electron diffusivity owing to ice particles, which become charged by electron attachment. Subsequently these charged ‘heavy’ particles influence the remaining free electrons by Coulomb interaction. The morphology of PMSE is similar to that of NLC, but the structures are thicker, extending from the lower border of saturation to above the mesopause (Figures 2 and 4). Because of their greater sensitivity to the more numerous small particles, PMSE are often seen when NLC are not detected. Also when they are present, the thin NLC layers (a few kilometers thick) are often detected near the lower PMSE boundary. Only the largest particles in the particle size distribution (equivalent radii, $r > 25\text{--}30\text{ nm}$) appear as NLC layers to lidar backscattering, owing to the strong dependence of light scattering on particle size (as described in Section [Optical Sensing of Mesospheric Clouds](#), highly sensitive near-IR measurements of PMC from space detect ice throughout the saturated region). It should be mentioned that both PMSE and NLC are quite sporadic in occurrence. The morphology of PMSE and ice particles are intimately related but their detailed relationship is not fully understood, and is a subject of continuing research.

Optical Sensing of Mesospheric Clouds

Five basic types of optical measurements are used in the study of mesospheric ice clouds. All but one type relies upon the scattering of solar light: (1) Ground-based NLC sightings are usually visible only near the horizon where the line-of-sight optical paths are longest. Such data are most useful for seasonal, morphological, and long-term trends. Time lapse movies are becoming scientifically valuable in revealing complex dynamical changes not obvious in still pictures; (2) Rocket-borne photometers flown through the region; (3) Active methods which also rely upon the same basic scattering process, but utilizing a series of powerful *laser* pulses (light amplification by stimulated emission of radiation); (4) Measurements from orbit in the visible and ultraviolet (UV) spectral ranges; (5) Methods relying upon precision measurements from space of the attenuation (extinction) of sunlight passing tangentially through the cloud layers.

Lidar Observations

Lidar is an active technique for observing NLC from fixed ground-based sites. Three permanent Northern Hemisphere

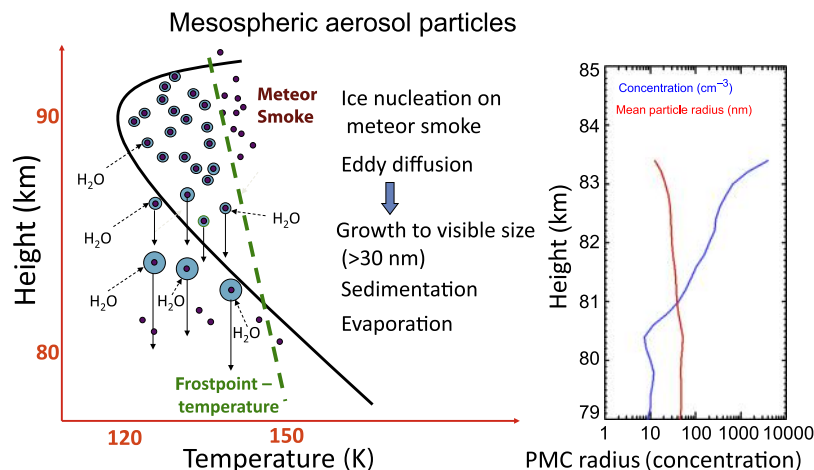


Figure 4 Left: Schematic diagram of the microphysics of mesospheric ice particles. Temperature is shown by the black curve. Meteor smoke particles (MSPs) are shown as purple dots, and ice is shown in blue. Where supersaturation occurs (to the left of the dashed green line showing the frost point temperature where the air is barely saturated), ice forms on the largest MSP. Eddy diffusion carries the small particles downward into a less saturated, but moister environment, where the ice particles may grow to sizes that are optically visible, and constitute a narrow noctilucent cloud (or polar mesospheric cloud) layer near 83 km. Ice particles exit the region within hours by sedimentation and evaporate where the temperature exceeds the frost point. This diagram ignores many other processes such as horizontal transport of ice by winds and strong gravity wave effects (see text). Figure adapted from M. Rapp of the Leibniz Institute of Atmospheric Physics at the Rostock University, Kühlungsborn, Germany, with permission. Right: SOFIE data (see Section [Optical Sensing of Mesospheric Clouds](#)) from 26 July 2007 (orbit 782) for ice particle radius (nm) and concentration (cm^{-3}). For SI units (m^{-3} , multiply blue curve values by 10^6).

lidars, at Sondrestrom and Andoya (Norway), at Fairbanks, Alaska, USA, and in the south, at Davis Station, Antarctica, routinely measure NLC. Of course, lidars can only measure NLC when they are not obscured by tropospheric clouds. The Institute of Atmospheric Physics (IAP) in Germany operates the Rayleigh/Mie/Raman (RMR) lidar at the Arctic Lidar Observatory for Middle Atmosphere Research (ALOMAR) at Andoya (69.3° N). Analysis of three-color ALOMAR lidar backscattered light from the brighter clouds determines NLC mean particle size (30–65 nm range), size dispersion, and ice water content. The long-term data sets have defined the local time dependence of cloud occurrence frequency, which follows a diurnal (24-h) period, at least at this location ([Figure 5](#)). Predictions of the Leibniz-Institute Middle Atmosphere (LIMA) model (see Section [Theory and Modeling of Mesospheric Ice Clouds](#)) suggest that the changes are driven by the atmospheric thermal tide.

Satellite Observations

In 1969, satellite-borne instruments intended to measure airglow revealed solar scattered light from a daytime mesospheric cloud layer extending to near the geographic pole. This discovery, and subsequent satellite experiments, showed that mesospheric ice clouds (now referred to as PMC) increase in brightness and occurrence rate with latitude toward a maximum at the pole. The clouds were found to exist over similar latitude ranges in both northern and southern summers. US Astronauts and Soviet/Russian Cosmonauts are quite familiar with the phenomenon, reporting (and later photographing) PMC from space. Following these early measurements, data have been collected from increasingly

sophisticated satellite instrumentation. A few highlights follow.

A goal of the Odin satellite mission (2001) was to study the cold mesopause region. Recently, analysis of the data from the Odin Optical, Spectroscopic, and Infrared Remote Imaging System (OSIRIS) spectrometer experiment has revealed a remarkable ‘teleconnection’ control of PMC from the *opposite* (winter) stratosphere (15–50 km). This connection occurring over tens of thousands of kilometers is particularly striking in affecting the southern PMC, since the northern stratosphere is more dynamically disturbed in winter than the southern winter. In addition, a second dynamical forcing agent has been found using data from Aeronomy of Ice in the Mesosphere (AIM) Cloud Imaging and Particle Size (CIPS) (see below) and OSIRIS. The seasonal start date of southern PMC appears to be tightly controlled by the timing of the winter–summer stratospheric wind transition (the ‘breakdown of the polar vortex’).

The NASA Aeronomy of Ice In the Mesosphere (AIM, 2007) is the first satellite mission dedicated to the study of PMC. The AIM CIPS experiment measures PMC scattered brightness in the UV, with a 5-km resolution ([Figure 3](#)), PMC spatial distribution over most of the polar region, and the spectrum of gravity waves as they interact with cloud brightness. A surprising result is that cloud structures change nearly completely over consecutive 90-min orbital periods, indicating shorter cloud lifetimes than anticipated from theory. CIPS revealed that large-scale nearly circular ‘voids’ occur in PMC images, whose origin is yet to be determined. Analysis has shown that PMC have similar fractal dimensions as tropospheric clouds, an important clue suggesting that there are common meteorological processes in both types of clouds. Analysis of wave structures in the CIPS large-field images has resolved a wealth and diversity

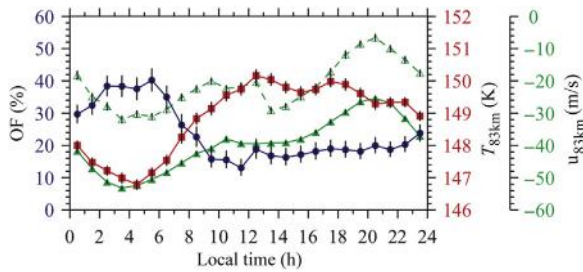


Figure 5 Diurnal variation of noctilucent clouds occurrence determined from Arctic Lidar Observatory for Middle Atmosphere Research Rayleigh/Mie/Raman (RMR)-lidar (using the brighter cloud data) and atmospheric parameters at 83-km altitude during times of lidar measurements: temperature (red squares) and zonal wind (green filled triangles) as calculated by the Leibniz-Institute Middle Atmosphere (LIMA) model, and zonal wind as measured by the medium frequency (MF) radar (green open triangles, dashed line). RMR and LIMA data are integrated from 1997 to 2010, MF data from 1999 to 2010. Vertical bars indicate confidence limits for the occurrence at 95% level and errors of the mean for temperature and wind. Figure by J. Fiedler of the Leibniz Institute of Atmospheric Physics at the Rostock University, Kühlungsborn, Germany, with permission.

of coherent gravity wave activity extending over both the northern and southern polar summer regions with horizontal scales ranging from tens to several thousand kilometers. The AIM Solar Occultation For Ice Experiment (SOFIE) conducts high-precision solar occultation measurements in 16 spectral bands to retrieve vertical profiles of temperature, O_3 , H_2O , CO_2 , CH_4 , NO , and PMC extinction. Due to its unprecedented sensitivity, SOFIE data confirmed the expectation from PMSE (see Section [The Modern Era: The Discovery of the Mesopause](#)) that ice is distributed throughout the saturated region. SOFIE observations have revealed the size, shape, concentration, phase, and temperature of mesospheric ice (see [Figure 4](#)). Ice and water vapor are measured simultaneously in a common volume, revealing in detail the ‘freeze-drying’ process, and rehydration below the cloud region. SOFIE made the first satellite observations of MSP throughout the upper stratosphere to 85 km. Observations in the 330–1037 nm wavelength range have significantly narrowed down the candidates for the MSP composition (carbon, wüstite, magnesiowüstite, olivine, or magnetite). SOFIE measurements also reveal that the ice particles comprising PMCs contain small amounts of smoke (0.02–2% by volume), and that the smoke composition contained in ice is consistent with carbon, wüstite, or magnesiowüstite.

Large rockets affect PMC directly through their exhaust products. The NASA Space Shuttle injects large amounts of water into the lower thermosphere. The water-rich plume is sometimes transported many thousands of kilometers toward the poles, and when the midlatitude launches occur in summer, impulsive increases in PMC brightness are observed several days afterward in satellite PMC data. The mechanism of the fast transport is still not understood, and it remains an open question whether growing amounts of space ‘traffic’ has contributed to the apparent long-term trends of PMC brightness and occurrence (Section [Long-Term Changes of Mesospheric Clouds](#)).

Theory and Modeling of Mesospheric Ice Clouds

The ‘microphysics’ of ice formation and evolution involves the nucleation, growth, transport, and decay of ice particles. The details of the prevailing theory of the nucleation process are sketchy, beginning with the values and seasonal dependence of the incoming meteoroid flux and composition, and continuing with their ablation and subsequent coagulation. The classical ‘liquid droplet’ theory predicts that only the largest MSP sizes (>1.3 nm) have a low enough ‘energy barrier’ that water vapor will deposit on their surfaces more rapidly than they will sublimate. Fortunately, modeling cloud ice mass is quite insensitive to the number of seed nuclei assumed (which is poorly defined in the cloud region). This somewhat magical, but convenient fact is explained as follows. Because nucleation rates are rapid in the upper cold region ($T < 130$ K), the number of particles is very large ($\sim 10^9$ m^{-3}), but their size is limited to tens of nanometers due to the limited water supply. The creation of ice ‘freeze dries’ the mesopause region. Particle sedimentation is too slow to transport the particles downward where clouds form. The cold nucleation region near the mesopause is thus weakly coupled to the cloud region below. These small particles are transported downward mainly through *turbulent eddy diffusion* ([Figure 4](#)), and not because of sedimentation, as previously thought. (The turbulence is thought to be a result of the gravity wave ‘breaking’ process, mentioned in Section [The Modern Era: The Discovery of the Mesopause](#).) A small fraction is transported into the warmer (temperature approximately 140–145 K) and slightly saturated region. The smallest particles reaching this lower region sublimate and contribute to the hydration of the region. The larger particles ($r > 10$ nm) grow rapidly due to the exponential increase of water vapor, and because there are so few of them ($\sim 10^8$ m^{-3}), they grow into optically visible clouds ($r > 25$ –30 nm) without seriously depleting the available water. The total ice mass and also the cloud brightness are limited by the number of ice particles delivered to the cloud region, rather than the amount of available water. The lifetime (a few hours) of the visible cloud is limited either by sedimentation into the warmer air, or by the duration of saturation. The passage of large-amplitude gravity waves through the region somewhat alters this simple picture of an upper region filled with cloud ‘embryos,’ and a lower region occupied by larger visible ice particles. The entry of the cold phase of the wave into the cloud region will lower, and blur, the ‘boundary’ of the two regions, so that nucleation of new particles can occur lower down. The entry of the warm phase will raise the boundary. Modeling of effects of these strong waves produce an overall increase of ice particle number.

Thus the key to understanding many of the observed properties of clouds is the dynamical variability of the medium, on many scales. Gravity waves cause horizontal variability on scales from tens to hundreds of kilometers. These waves ‘stir up’ the medium, modifying the temperature and wind speeds on timescales of tens of minutes to hours. In addition, where they become unstable, the waves dissipate their energy and cause further stirring due to the creation of turbulence, which occurs on scales of tens of meters. Even a few Kelvin of temperature variance can lead to large differences in cloud properties. [Figure 3](#) illustrates that PMC structure occurs on scales down to

at least 5 km, and are quite likely sensitive tracers of temperature variability on similar spatial scales.

Long-Term Changes of Mesospheric Clouds

Long-term trends are of two types: (1) cyclical, with 10- to 11-year periodicity, and (2) secular trends, which are only noticeable after several decades of observation. Identifying secular trends are of great interest, in that they could be caused by anthropogenic or naturally occurring climate change. Two types of data are relevant. The first involves longer duration ground-based results. The second type is from satellites, for which one suite of measurements in particular has the record length necessary to address both cyclical and secular tendencies.

Two sets of NLC observations from the UK and Denmark (latitudes from 51° to 61° N) since 1964, and from near Moscow, Russia (56° N), since 1962 constitute the longest continuous records. The conventional NLC 'climate index,' N , is the total number of nights NLC are spotted within a season, and ranges from as few as four nights per season to as many as 40 or more. The observed cyclic behavior is dubbed the 'quasi-10-year oscillation,' since the derived period (10.4 years) is less than that of the 11-year solar cycle. Within a few percent neither the mean brightness nor N shows significant increases over the last 45 years.

Satellite observations have several important advantages over ground-based data in identifying long-term changes in PMC, covering both hemispheres, all longitudes and most polar latitudes. They are immune to tropospheric cloud cover and provide quantitative photometry. Disadvantages are that their high-inclination orbits are tied to a fixed, or slowly varying solar local time, and analysis of time series are subject to mixing up possible long-term variations in the tidal signature (see Figure 5). Also, most satellite instruments generally have short lifetimes, usually less than 10 years, precluding separation of cyclical and secular trends.

Of greatest value for long-term trends are data from the Solar Backscatter Ultraviolet (SBUV) Spectrometer. Seven nearly identical SBUV instruments have been flown since 1979 in Sun-synchronous orbits, with the objective of monitoring long-term changes in stratospheric ozone. Fortunately, shorter wavelength data also contain signatures of bright PMC. A time series of average PMC average brightness for 1979–2010 is shown in Figure 6. PMC data show both cyclical and secular trends. Similar but smaller trends apply to the southern PMC. Frequency of bright cloud occurrence also shows significant secular trends (with 95% confidence) in both hemispheres. The apparent contradiction between the ground-based and low-latitude SBUV trends is not understood.

Attributing the trends to causal forcings is the task of models. Two different general circulation models including ice formation now reproduce the SBUV trends. The LIMA model

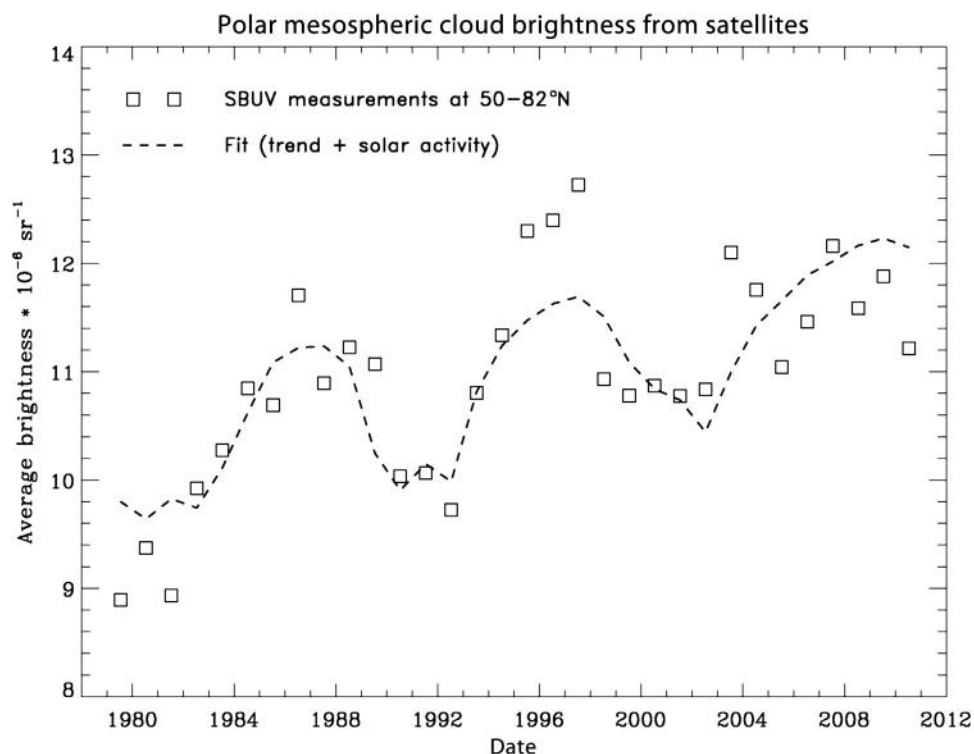


Figure 6 Annually averaged Northern Hemisphere polar mesospheric clouds brightness (squares) Solar Backscatter Ultraviolet data vs year through 2010. Dashed curve: a multiple regression fit of the data, including a constant, a term proportional to solar Lyman alpha irradiance, and a linear secular increase. The latter value is $+3.5 \pm 1.7\%$ per decade (exceeding 95% confidence). The Southern Hemisphere trends are smaller ($+2.3 \pm 1.2\%$ per decade), but still significant. Provided by M. DeLand of Science Systems and Applications, Inc., Lanham, MD, USA, with permission.

was developed by the IAP, and the Whole Atmospheric Community Climate Model (WACCM) by the National Center for Atmospheric Research (NCAR) in Boulder, CO, USA. The versions that include mesospheric ice formation, the IAP LIMA/ICE model and (the latter developed jointly by NCAR and the University of Colorado at Boulder) take different approaches for simulating long-term changes, yet arrive at similar results. Both cyclical and secular trends of modeled PMC match remarkably well the SBUV data set (Figure 6). In both models, the cyclical effect is caused by solar cycle changes in the UV photodissociation rate, reducing water vapor and thus cloud brightness during solar maximum. The IAP group attributes most of the secular trend to stratospheric ‘contraction,’ since their lower boundary is driven by meteorological data, and not by greenhouse gas (GHG) forcing. However, GHG forcing effects are implicitly included in their lower boundary conditions, since the known cooling trends are contained in the stratospheric forcing. In contrast, WACCM-PMC models the entire atmosphere, and so the changes are a result of GHG increases throughout the atmosphere. WACCM-PMC also includes increases in chlorofluorocarbons (CFCs), which destroy ozone and create the ozone hole. The long-term decline in ozone heating results in a cooler PMC region. However, the separate roles of CO₂, CH₄, and CFCs are yet to be sorted out. The attribution issue is yet to be resolved, but the analysis tools are now available to determine the origin of changes revealed in long-term data sets of PMC. As mentioned in Section [Optical Sensing of Mesospheric Clouds](#), the question of whether increased space traffic has contributed to the observed SBUV trend is still open. Much remains to be learned.

See also: **Clouds and Fog:** Cloud Microphysics. **Dynamical Meteorology:** Waves. **Global Change:** Upper Atmospheric Change. **Gravity Waves:** Overview. **Lidar:** Backscatter. **Mesosphere:** Polar Summer Mesopause. **Numerical Models:** General Circulation Models. **Statistical Methods:** Data Analysis: Time Series Analysis. **Stratospheric Chemistry Topics:** Stratospheric Water Vapor. **Thermosphere:** Thermosphere.

Further Reading

- Baumgarten, G., Fiedler, J., Lübken, F.-J., von Cossart, G., 2008. Particle properties and water content of noctilucent clouds and their interannual variation. *Journal of Geophysical Research* 113, 1–13.
- Dalín, P., Kirkwood, S., Andersen, S.H., et al., 2006. Comparison of long-term Moscow and Danish NLC observations: statistical results. *Annals of Geophysics* 24, 2841–2849.
- DeLand, M., Shettle, E.P., Thomas, G.E., Olivero, J.J., 2006. A quarter-century of satellite polar mesospheric cloud observations. *Journal of Atmospheric and Solar-Terrestrial Physics* 68, 9–29.
- Fiedler, J., Baumgarten, G., Berger, U., et al., 2011. NLC and the background atmosphere above ALOMAR. *Atmospheric Chemistry and Physics* 11, 5641–5679.
- Gadsden, M., Schröder, W., 1989. *Noctilucent Clouds*. Springer Verlag, Berlin.
- Hervig, M., Gordley, L.L., Stevens, M.H., et al., 2009. Interpretation of SOFIE PMC measurements: cloud identification and derivation of mass density, particle shape, and particle size. *Journal of Atmospheric and Solar-Terrestrial Physics* 71, 316–330.
- Karlsson, B., Gumbel, J., 2005. Challenges in the limb retrieval of noctilucent cloud properties from Odin/OSIRIS. *Advances in Space Research* 36, 935–942.
- Lübken, F.-J., Berger, U., Baumgarten, G., 2009. Stratospheric and solar cycle effects on long-term variability of mesospheric ice clouds. *Journal of Geophysical Research* 114, D00I06.
- Rapp, M., Thomas, G.E., 2006. Modeling the microphysics of mesospheric ice particles: assessment of current capabilities and basic sensitivities. *Journal of Atmospheric and Solar-Terrestrial Physics* 68, 715–744.
- Russell III, J.M., Bailey, S.M., Gordley, L.L., et al., 2009. The Aeronomy of Ice in the Mesosphere (AIM) mission: overview and early science results. *Journal of Atmospheric and Solar-Terrestrial Physics* 71, 289–299.
- Stevens, M.H., Meier, R.R., Chu, X., DeLand, M.T., Plane, J.M.C., 2005. Antarctic mesospheric clouds formed from space shuttle exhaust. *Geophysical Research Letters* 32, L13810. <http://dx.doi.org/10.1029/2005GL023054>.
- Thomas, G.E., 1991. Mesospheric clouds and the physics of the mesopause region. *Reviews of Geophysics* 29, 553–575.
- Thomas, G.E., Olivero, J.J., Jensen, E.J., Schroder, W., Toon, O.B., 1989. Relation between increasing methane and the presence of ice clouds at the mesopause. *Nature* 338, 490–492.

Stratus and Stratocumulus

R Wood, University of Washington, Seattle, WA, USA

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

Stratus and stratocumulus clouds are two dominant types of low stratiform cloud that together cover 30% of the Earth's surface. Both are thin boundary layer clouds that frequently occur as the sole cloud type under conditions of large-scale subsidence and strong lower-tropospheric static stability, but they can also exist mixed with higher clouds during disturbed periods. Stratocumulus are convective clouds driven largely by the emission of infrared radiation from cloud top and from latent heat release. Stratus, on the other hand, are largely free of convective overturning. Both are typically very thin clouds that can extend horizontally for large distances. Stratocumulus frequently produce light precipitation often in the form of drizzle. Because they strongly reflect solar radiation, stratus and stratocumulus exert an important impact on climate.

Stratus and stratocumulus clouds are the two most common genera of low stratiform cloud by area covered. They often occur under fair weather conditions, and in this way they are distinct from a third low stratiform low cloud type called nimbostratus. However, stratus and stratocumulus also commonly occur during periods of large-scale ascent, under which conditions they coexist with clouds at other levels. In stratocumulus, the cloud layer is comprised of a multitude of individual convective elements giving the layer a lumpy morphology. Stratus, in contrast, is typically more featureless than stratocumulus because of a lack of active convective elements. The layering in both stratus and stratocumulus is commonly maintained by a capping inversion immediately above the cloud top. The temperature inversion can sometimes be strong and is often only a few meters to tens of meters thick. Both stratus and stratocumulus clouds are usually contained in the atmospheric boundary layer.

Stratocumulus is the most common cloud type globally (Warren et al., 1986, 1988), covering approximately one-fifth of Earth's surface in the annual mean (23% of the ocean surface and 12% of the land surface). Stratus clouds cover roughly 10% of the Earth's surface (12% of the ocean surface and 5% of the land). Because stratus and stratocumulus clouds are so common and because they strongly reflect incoming solar radiation, these clouds are important for Earth's radiative balance and therefore climate. In addition, accurate prediction of daytime temperature over many land areas relies on models being able to accurately represent stratus and stratocumulus and their diurnal variability.

Stratocumulus commonly occurs under conditions of large-scale subsidence and strong lower-tropospheric static stability. In the subtropics and tropics, they tend to occur in semi-permanent sheets over the cold eastern ocean basins under the downward branches of the Hadley and Walker circulations (Figure 1(b)). In midlatitudes, stratocumulus sheets tend to be more transient and are typically associated with transient ridges in passing planetary waves, and in winter they commonly form in offshore or polar cold-air outbreaks.

Like stratocumulus, stratus clouds also tend to form under conditions of large-scale subsidence and strong static stability. Whereas stratocumulus dynamics are primarily driven by convective instability caused by cloud top thermal infrared

radiative cooling, stratus tends to be free of convective overturning and so is typically associated with a stable temperature profile in the boundary layer. Stratus may be formed from warm air passing over a cold ocean (warm advection), which can lead to condensation without the need for convective

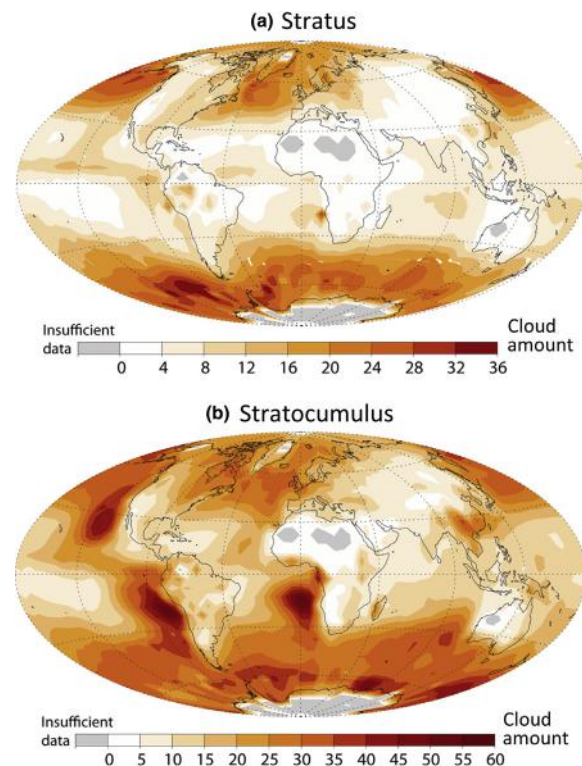


Figure 1 Annual mean cloud amount of (a) stratus and (b) stratocumulus from the volunteer surface observer dataset of Warren, S.G., Hahn, C.J., London, J., Chervin, R.M., Jenne, R.L., 1986. Global Distribution of Total Cloud Cover and Cloud Types over Land. NCAR Tech. Note NCAR/TN-2731STR. National Center for Atmospheric Research, Boulder, CO, 29 pp. + 200 maps and Warren, S.G., Hahn, C.J., London, J., Chervin, R.M., Jenne, R.L., 1988. Global Distribution of Total Cloud Cover and Cloud Types over Ocean. NCAR Tech. Note NCAR/TN-3171STR. National Center for Atmospheric Research, Boulder, CO, 42 pp. + 170 maps. Note the different color scales on the upper and lower panels.

overturning. This can lead to the formation of fog, a cloud genera distinct from stratus and stratocumulus, but wind-driven mixing can elevate the base of the fog layer, leading to stratus. Stratus associated with warm advection is most common over the storm track regions of the Atlantic, Pacific and Southern Oceans, where it can exist 20–30% of the time in both summer and winter months (Figure 1(a)). Over land, stratus is fairly uncommon and is concentrated in regions influenced by extratropical storm tracks. Stratus is also uncommon over the warm tropical and subtropical ocean, but is more common over the cold subtropical oceans (Figure 1(a)). However, in these regions cold advection and strong thermal infrared cloud top cooling under a dry free-troposphere often transforms stratus clouds into stratocumulus by encouraging convective instability.

Structure

Stratocumulus clouds are typically 200–400 m thick and usually occur at the top of the boundary layer below a thermal inversion. It is remarkable that such a thin cloud can extend practically unbroken for a thousand kilometers, but strong negative feedbacks exist to constrain cloud thickness. First, thicker clouds produce drizzle, which can act to reduce condensate and serve as a negative feedback on thickness. Second, and likely the most important in cloud layers without drizzle, the cloud thickness is limited by the entrainment of dry, warm air from the free troposphere, which acts to dry out and thin the cloud. Cloud top entrainment in stratocumulus is driven by the turbulence created by the clouds themselves. Thicker stratocumuli generate more turbulent kinetic energy than thinner ones, so that the strength of cloud top entrainment increases with stratocumulus cloud thickness. The cloud thickness-turbulence-entrainment loop also serves as a strong negative feedback on stratocumulus cloud thickness.

Stratus is often confined to within a few hundred meters of the surface, but can exist at many levels often in conjunction with more vertically extensive cloud systems such as warm fronts and mesoscale convective systems (often as stratus fractus). Stratocumulus top heights are typically 500–2000 m over the Earth, confined mostly by the inversion height of atmospheric boundary layers. Stratocumulus clouds can exist within both well-mixed and intermittently coupled boundary layers. Stratocumulus clouds help to keep the cloud layer relatively well-mixed, but as the boundary layer becomes deeper, the cloud layer frequently becomes disconnected from the surface mixed layer. Coupling of the stratocumulus and surface layers in this case is intermittent and is provided by patches of cumulus clouds that originate at the top of the surface mixed layer and in many cases reach up into the stratocumulus deck. Stratocumulus in coupled, well-mixed boundary layers tends to be more horizontally homogeneous than that in intermittently coupled boundary layers, where stratocumulus tends to exist in patchy regions surrounding the regions of cumulus coupling. Stratus clouds tend to only exist within thermally stratified layers, sometimes as the result of large-scale ascent associated with midlatitude cyclones or orography.

Because stratocumulus are low clouds, most stratocumulus contain liquid condensate, but ice can also be present when the

cloud top temperature is sufficiently supercooled. It is unlikely that completely glaciated stratocumulus clouds can produce sufficient turbulence to maintain the cloud against precipitation losses. Observations of low, stratiform clouds in the Arctic suggest that most contain liquid water even at temperatures well below freezing. In general, liquid water increases upward from cloud base and usually maximizes near cloud top, consistent with what one would expect to occur in a well-mixed layer where water mixing ratio is constant and temperature decreases with height. The turbulent eddies in stratocumulus layers are initiated by cloud top longwave cooling but are strengthened by latent heat release in updrafts and evaporative cooling in downdrafts. The distribution of vertical wind speed is negatively skewed near cloud top, with strong downdrafts and relatively weak updrafts, but the eddies become more symmetric further down in the cloud layer. The strongest vertical winds are usually found near the top of the cloud layer. Typically, updraft speeds in stratocumulus are $0.1\text{--}1\text{ m s}^{-1}$, but are weaker in stratus clouds, where shear-generated turbulence may produce small-scale, weak eddies that can drive a degree of overturning.

The tops of stratocumulus clouds are typically flatter than the bases. This is particularly true over oceans under conditions of large-scale subsidence where time has been allowed to develop a strong thermal inversion atop the clouds that acts to suppress the vertical extent of the cloud tops. Over land, both cloud bases and tops can exhibit comparable variability. Stratus clouds also often have flat tops, and are usually distinguished from stratocumulus by their relatively flat bases.

Stratocumulus clouds exhibit horizontal structure on the scale of the largest turbulent eddies. This is most strikingly seen with the naked eye or in photography of cloud tops and bases, which show clear hummock structures on scales of a few hundred meters. Visible satellite imagery (e.g., Figure 2)

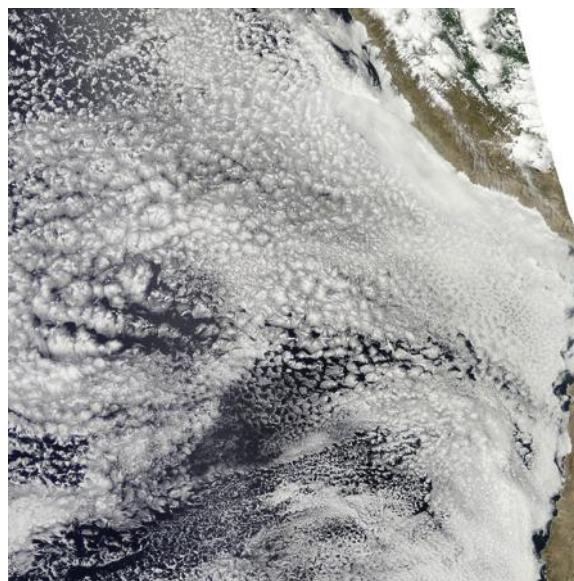


Figure 2 Visible satellite image from the NASA Moderate Resolution Imaging Spectroradiometer to the west of South America showing closed and open cell stratocumulus clouds. A band of stratus clouds is present along the Peruvian coastline. The image scale is approximately 2000 km across. Image from the NASA MODIS Rapidfire archive.

indicates, however, that stratocumulus cloud fields also exhibit a significant degree of structure and organization on scales of tens of kilometers, i.e., on mesoscale-beta and mesoscale-gamma scales. Stratocumulus clouds over oceans tend to organize into one of two common forms: closed and open mesoscale cellular convection (Atkinson and Zhang, 1996). The cells vary in horizontal scale but tend to be 5–50 km in size, with the depth of the stratocumulus-topped boundary layer in which they exist being the primary determinant of cell size (Wood and Hartmann, 2006). Although the theory for such mesoscale organization is not complete, it is clear that diabatic effects and their effects on boundary layer turbulence are the primary drivers leading to an upscaling of turbulent kinetic energy from the large eddy scale to the mesoscale. One example of this is that the evaporation of precipitation below cloud can drive cold pool formation that suppresses small-scale eddy overturning but encourages the formation of mesoscale circulations. Stratus clouds do not generate their own mesoscale variability, but may exhibit such variability if it is imposed by the large-scale flow.

Formation, Maintenance, and Transformation

Stratus and stratocumulus can form from the cooling or moistening of a clear boundary layer. Clear sky longwave cooling drives the clear boundary layer toward condensation by lowering its temperature toward the dew point temperature. Condensation will typically first occur in a thin layer near the top of the boundary layer. This cloud is initially stratus because it likely contains little convective instability. But quickly the cloud layer typically grows sufficiently thick to become strongly emissive in the thermal infrared, so that radiative cooling becomes increasingly concentrated in a thin layer near cloud top. This cooling drives the typical form of ‘top down’ convection commonly found in marine stratocumulus. Over land, surface fluxes are often strong enough to compete with cloud top cooling, and so turbulence in continental stratocumulus tends to be more strongly surface-driven than is the case over oceans. Stratocumulus can also emerge from the transition of another cloud type, most commonly cumulus and stratus.

Stratus clouds, in addition to being formed by radiative cooling of a moist clear layer, can be formed by lifting of moist air masses in regions adjacent to fronts or over orography, and by warm advection of a moist layer over a cold surface. Because thermal infrared absorption from liquid water is so efficient, many stratus clouds are unstable, because, as the cooling strengthens, condensate builds up, longwave cooling further strengthens, and the layer becomes increasingly susceptible to convective overturning (hence stratocumulus formation). In this way, stratus clouds can be seen as transient in nature, the thermodynamics favorable for their existence only existing in fleeting pulses. The resulting stratocumulus sheet tends to be a more robust cloud system. On the other hand, stratocumulus can form as the result of the spreading of shallow cumulus clouds. Although unusual over the ocean, this occurs quite frequently over extratropical land areas. For this to occur, the cumuli need to be strongly capped by a thermal inversion, and surface driving needs to be sufficiently strong.

Stratocumulus cloud layers are maintained by an energetic balance between radiative cooling and surface and cloud top entrainment warming, and by a moisture balance between surface moisture fluxes and drying from cloud top entrainment. Precipitation, when present, can affect both energy and moisture budgets in sometimes complex ways.

Stratocumulus and stratus clouds frequently exhibit strong diurnal variability, primarily as a result of the daytime absorption of solar radiation near cloud top. This serves to weaken the overturning from cloud top longwave cooling during daytime, which weakens the ability of the stratocumulus to maintain a well-mixed boundary layer. The suppression of mixing during daytime can help to decouple the cloud from its surface moisture source, which can result in cloud thinning and also breakup. Observations show that, as a result, stratocumulus clouds are more turbulent, thicker, and contain most condensate at night. Studies show that nighttime stratocumulus can contain 50% more condensate than during the day. The higher nighttime condensate amounts help drive nighttime precipitation maxima. Typically, as stratocumulus breaks up during the day, it tends to be replaced by either fair weather cumulus clouds or by a broken stratocumulus deck with cumulus clouds rising into it. Stratus clouds often dissipate during daytime by direct heating of the cloud layer by solar absorption and subsequent reduction in the relative humidity.

Just as the diurnal breakup of stratocumulus clouds during the day is caused in large part by the transition from a well-mixed to an intermittently coupled boundary layer, so too is the climatological transition from subtropical stratocumulus to tropical trade cumulus over the eastern ocean basins. In the latter case, the transition is driven by increasing surface latent heat flux as the ocean warms below. This drives stronger entrainment of warm, positively buoyant free-tropospheric air, which becomes increasingly difficult to mix down through the entire boundary layer. As a result, the boundary layer transitions from a shallow, well-mixed layer dominated by relatively homogeneous stratocumulus clouds, to a deeper, intermittently coupled boundary layer containing cumulus clouds rising into a layer of stratocumulus. As the transition progresses, the cumulus clouds become increasingly unable to supply sufficient moisture to the stratocumulus layer to offset drying from cloud top entrainment. As a result, the stratocumulus layer thins (Figure 3) and eventually dissipates completely leaving trade cumulus clouds as the sole clouds in the boundary layer.

An alternative fate for well-mixed stratocumulus can occur if sufficient precipitation is produced to generate evaporatively driven cold pools. The subcloud cooling of evaporating drizzle drops, coupled with the dynamic forcing from cold pools undercutting and lifting near-surface air, helps drive locally strengthened regions of ascent that can spawn cumulus clouds that further enhance precipitation formation. The efficient removal of precipitation from the stratocumulus layer can result in thinning and dissipation, with the result being mesoscale cell walls containing thick clouds and cell centers with very thin and broken clouds. This precipitation-driven transition from closed to open cells can be very sharp in space (Figure 2) and time.



Figure 3 Photograph showing thinning sheet of stratocumulus clouds over the southeastern Pacific Ocean with penetrating solar rays. Image taken as part of the NOAA stratus research cruises to the southeastern Pacific.

Microphysical Processes

Warm rain formation in stratocumulus first requires the formation of precipitation drop embryos. These are formed from a small subset of cloud droplets lucky enough to undergo a few coalescence events. Because the probability of coalescence and the speed at which a droplet falls both increase strongly with droplet size, the lucky drops can grow rapidly given a sufficient quantity of cloud droplets to collide with. As they grow, these embryonic drizzle drops begin to fall faster than cloud droplets because a drop's terminal speed increases rapidly with its size. The embryos begin to collect cloud droplets as they fall through the cloud layer (a process termed accretion) and can grow rapidly to sizes of several hundred micrometers. Precipitation formation in stratocumulus is strongly dependent upon the availability of condensate in the column (i.e., cloud thickness) and also upon the ability of the cloud to produce large cloud droplets capable of producing precipitation embryos.

The concentration of cloud droplets in stratocumulus clouds is determined by the availability of soluble aerosol particles (most commonly natural and anthropogenic sulfate aerosols and sea-salt) and by the strength of the updrafts. In stratus, updrafts are particularly weak, whereas convective elements in stratocumulus tend to produce stronger updrafts that results in a higher concentration of cloud droplets being formed for a given aerosol population. The concentration of cloud droplets is particularly important because it strongly impacts the resulting size of droplets for a given amount of condensate. The impacts of cloud droplet size are twofold: first, for a given condensate amount, smaller droplets lead to a greater overall surface area and a higher cloud albedo. Second, because the precipitation process in stratocumulus is caused by the collision and coalescence of cloud

droplets, and because smaller droplets collide less efficiently than larger drops, the production rate of precipitation embryos is faster in a cloud with a low cloud droplet concentration and larger droplets. Thus, observations indicate that the rate at which precipitation falls from stratocumulus is quite strongly dependent upon the cloud droplet concentration. As discussed above, precipitation can exert a significant impact on the dynamics of stratocumulus clouds, and so it is necessary to consider the impact of aerosol particles to understand the degree to which stratocumulus clouds precipitate. Satellite radar data indicate that stratocumulus over remote ocean regions, where the aerosol concentration is particularly low, precipitate more efficiently than those near the coast. Indeed, over land many stratocumulus clouds do not precipitate at all. Regardless of the cloud droplet concentration, stratocumulus clouds thinner than ~ 200 m do not generate much precipitation.

Mixed-phase stratocumulus and stratus, however, can precipitate via the ice phase. Because the growth of ice under water-saturated conditions is quite rapid, collision-coalescence is not necessary for precipitation formation in this case because ice crystals can grow to precipitation size (~ 1 mm) by vapor deposition. However, the availability of liquid water can still influence the precipitation in this case through the riming process: ice crystals collecting cloud droplets as they fall through the liquid cloud. Precipitation formation in mixed-phase stratocumulus and stratus remains poorly understood primarily because the production rate of ice crystals and the heterogeneous nuclei on which they form are not well known.

Precipitation in stratocumulus is horizontally heterogeneous, with a strong tendency for precipitation to be sharply concentrated in regions where the clouds are thickest. Precipitation drops in stratocumulus clouds grow to maximum sizes

of ~ 1 mm, but most are in the range 50–500 μm . As such, stratocumulus produce both drizzle (drops smaller than 500 μm) and, in some circumstances, light rain. Drops smaller than ~ 300 μm usually evaporate below cloud base; larger ones often survive the fall to the surface. The extent to which precipitation evaporates or reaches the surface has an important impact on the moisture and energy budgets of the boundary layer and upon cloud dynamics.

Stratus and Stratocumulus Clouds and Climate Change

Because stratus and stratocumulus together are so plentiful, they exert a strong cooling impact on the climate system by reflecting a large amount of sunlight back to space. It is well understood that relatively small changes in the coverage of these clouds or the brightness of the existing clouds can help to enhance or offset the warming caused by increasing greenhouse gases. Although consensus across models is still relatively poor, current climate modeling suggests that stratus and stratocumulus cloud cover will likely decrease slightly in future, thus enhancing greenhouse warming. However, the fact that stratus and stratocumulus clouds are so thin makes it particularly challenging to represent these clouds in low-resolution climate models.

In addition, the anthropogenic impacts of increasing aerosol concentrations may enhance the albedo of these clouds. These ‘aerosol indirect effects’ are thought to offset

anthropogenic greenhouse warming by as much as 25–50%, but these estimates are highly uncertain due in major part to a poor fundamental understanding of how to represent the myriad processes through which stratus and stratocumulus clouds interact with atmospheric aerosol particles.

See also: [Climate and Climate Change: Overview](#). [Clouds and Fog: Classification of Clouds](#); [Climatology](#); [Cloud Microphysics](#).

Further Reading

- Agee, E.M., 1987. Mesoscale cellular convection over the oceans. *Dynamics of Atmospheres and Oceans* 10, 317–341.
- Atkinson, B.W., Zhang, J.W., 1996. Mesoscale shallow convection in the atmosphere. *Reviews of Geophysics* 34, 403–431.
- Warren, S.G., Hahn, C.J., London, J., Chervin, R.M., Jenne, R.L., 1986. Global Distribution of Total Cloud Cover and Cloud Types over Land. NCAR Tech. Note NCAR/TN-2731STR. National Center for Atmospheric Research, Boulder, CO, 29 pp. + 200 maps.
- Warren, S.G., Hahn, C.J., London, J., Chervin, R.M., Jenne, R.L., 1988. Global Distribution of Total Cloud Cover and Cloud Types over Ocean. NCAR Tech. Note NCAR/TN-3171STR. National Center for Atmospheric Research, Boulder, CO, 42 pp. + 170 maps.
- Wood, R., 2012. Stratocumulus clouds. *Monthly Weather Review* 140, 2373–2423.
- Wood, R., Hartmann, D.L., 2006. Spatial variability of liquid water path in marine low cloud: the importance of mesoscale cellular convection. *Journal of Climate* 19, 1748–1764.

CRYOSPHERE

Contents

Glaciers, Topography, and Climate

Permafrost

Sea Ice

Snow (Surface)

Glaciers, Topography, and Climate

ABG Bush, University of Alberta, Edmonton, AB, Canada

MP Bishop, Texas A&M University, College Station, TX, USA

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

Glaciers are distinct features of Earth's climate system and, combined with sea ice, form the cryosphere. Glaciers currently exist around the globe from the polar regions to tropical high altitudes. Although composed of ice, they are very dynamic and play an extremely important role in shaping the land surface in alpine regions as well as regulating regional climate. Glaciers also play a fundamental role in providing freshwater resources to a significant fraction of the world's population. Monitoring of glaciers is an important component in the effort to determine the effects and ramifications of climate change.

Introduction

Water is arguably the most important component of our climate system. It is responsible for all of our freshwater and agriculture, our oceans and marine ecosystems, and it contributes to our daily weather. Water plays such a primary role on our planet because it can exist in all three phases given our Earth's temperatures and pressures. All substances have a vapor, liquid, and solid phase with transitions between these phases regulated by the thermodynamics of temperature and pressure. Within our climate system, water can naturally cross these thermodynamic thresholds using energy to evaporate water from the oceans or to melt ice, or releasing energy to form precipitation or ice. The energy required to evaporate water or to melt ice comes primarily from solar energy, and the energy released during condensation or ice formation heats the atmosphere, ocean, or a glacier itself depending on where these processes occur.

A glacier is an ice mass that has formed over a long period of time through snow accumulation. Repeated snowfalls compress the snow underneath, increasing its density by squeezing out air pockets. This compaction of snow into firn (a granular state between snow and ice) takes years and, usually over decades, the firn compacts into ice. The geographic

location of a glacier is therefore dependent on both temperature and net snowfall amounts. If it is too warm, the snowfall melts before it can be compacted to ice. In tropical and subtropical regions, these conditions are only satisfied at high elevations and hence glaciers exist at low latitudes only in the high mountains of the Andes, Indonesia, the Himalaya, and Africa. In polar regions these conditions are more easily satisfied, so glaciers are extensive there.

Classification

Geographic Setting

Glaciers are classified according to their size, their topographic surroundings, and their thermal regime. Large, geographically unconstrained glaciers include ice sheets (larger than 50 000 km², of which there are two: the Antarctic and Greenland ice sheets), ice caps (smaller than 50 000 km²; e.g., Iceland), and ice shelves (e.g., Antarctica's Ross and Ronne ice shelves). Smaller glaciers that are typically confined by their topographic setting include mountain glaciers, outlet glaciers, icefields, valley glaciers, piedmont glaciers, cirque glaciers, hanging glaciers, and tidewater glaciers. Glaciers are also classified according to the climate regime in which they exist

(maritime, continental, or transitional, as determined by a combination of temperature and precipitation).

Excluding ice sheets and ice caps, the world's largest glacier is the Lambert glacier in east Antarctica, at approximately 100 km wide, 400 km long, and 2500 m thick. Outside of the polar regions, the largest glacier is the Fedchenko glacier in Tajikistan (77 km long). As part of the western Himalaya, the Karakoram is home to a number of the next largest glaciers, including the Siachen (70 km), the Biafo (63 km), and the Baltoro (62 km long).

Thermal Classification

The thermal regime of a glacier (i.e., its internal temperature structure) has important ramifications for how it moves. A glacier can be polar, in which case its temperature is everywhere below the freezing point (273.15 K), or it can be temperate, meaning that most of the ice is maintained at or near the pressure melting point through a combination of seasonal melting then refreezing of the meltwater as it percolates into the glacier (with the latent heat release during refreezing maintaining a higher temperature). Most glaciers in the Alps, for example, are temperate glaciers whereas high latitude glaciers are polar. There is also the possibility of a combination of these two thermal regimes, called polythermal glaciers, which exist in subpolar–polar Canada and Russia where geothermal heating from below raises the basal ice to the pressure melting point while the surface of the ice is maintained well below the freezing point.

Cumulatively, glaciers cover approximately 10% of the Earth's surface area, they contain about 75% of the world's freshwater and, if they were all to melt then global sea level would rise by approximately 70 m. Conversely, sea level was approximately 120 m below today's level during the peak of the last glacial cycle about 21 000 years ago when much of the water in the oceans today was locked up as terrestrial ice. The glaciers observed today are remnants leftover from that last glaciation which saw 30% of the Earth's surface area glacierized, with ice over 3 km thick covering central Canada. Glaciers are therefore intimately linked to global climate and fluctuate in concern with global temperatures, greenhouse gas concentrations and, going further back in Earth's history, continental configuration as plate tectonics move the continents around the Earth's surface.

Mass Balance

The mass balance of a glacier is an indicator of the health of a glacier and is a measure of the annual mean difference between accumulation and melt. A negative mass balance means the glacier is on average losing mass and, if the climate does not change, it will disappear. A positive mass balance means the glacier is growing in volume (though not necessarily in area). Mass balances of glaciers are not easy to quantify because the underlying topography (and hence ice thickness) is not easily measured. What is more easily measured are the processes occurring at the surface of a glacier, so a 'surface' mass balance, or 'climatic' mass balance is commonly used. These balances are calculated by an energy balance equation that

takes into account net shortwave and longwave fluxes, net accumulation, net melt, and net latent heat fluxes to determine if the surface is accumulating mass or losing mass to melt. If the surface mass balance is negative everywhere, it is likely that the glacier overall has a negative mass balance. Such studies have made use of both ground measurements and satellite observations for Greenland (e.g., Box et al., 2013) and Antarctica (e.g., Shepherd et al., 2012). Recently, the Gravity Recovery and Climate Experiment (GRACE) satellite data have been used to detect gravitational anomalies directly associated with volumetric ice mass gain or loss (see the previous two references). These methods work well for large ice sheets whose fluctuations have a measurable gravitational signal but for the world's extensive network of alpine glaciers, scientists must rely on either aircraft data or *in situ* measurements to obtain surface data (complicated by the fact that many alpine glaciers are not easily accessible and some are located in geopolitically sensitive areas).

Dynamics

Like rivers, ice flows under the constraints of gravity and the underlying topography. The timescale of flow depends on a number of factors including ice thickness and temperature, steepness of topography, atmospheric temperature, and the presence or absence of liquid water underneath the ice. Typical timescales for the flow of very large ice sheets is on the order of centuries to millennia whereas for smaller mountain or valley glaciers the timescales can be as rapid as decadal, annual, or even subannual. Some glaciers exhibit rapid surges through unique combinations of the factors listed above. The fastest observed glacier, the Jakobshavn glacier in Greenland, has varied in speed over the years but values range from 5.7 km per year in 1992 to 15 km per year in 2010 (roughly 41 m per day), a near tripling of speed over this period. This is a significant change considering that this glacier drains approximately 5% of Greenland ice into the ocean, thereby forming approximately 10% of Greenland's icebergs and raising sea level by approximately 4% of the measured twentieth century amount.

From a practical and numerical perspective, the flow of ice is mathematically treated using Glen's flow law, which is an empirically derived relationship between ice stress and the strain rate. The stress is in turn related to the ice density, thickness, and slope. It is greatest at the base of the glacier in contact with the bed, where ice deformation is greatest and the velocity slowest. It is smallest at the glacier surface, where deformation is smaller and velocities are greater.

Like rivers, glaciers flow downhill. Mountain and valley glaciers therefore flow into warmer temperatures at lower elevations (temperature in the free atmosphere decreases with elevation at a rate of between 2 and 9.8 °C km⁻¹). The glacier can therefore flow into climatic conditions not conducive to the presence of ice. Similarly, outlet glaciers flow into oceans or lakes where the ice calves off to form icebergs (e.g., the Perito Moreno glacier, Argentina; Figure 1).

A simplified picture of the evolution of a glacier is of mass accumulation at high elevations, ice dynamically flowing



Figure 1 The active Perito Moreno Glacier, southern Patagonia, Argentina. Photo A.B.G. Bush.

downhill to lower elevations where it melts (or calves off). The elevation at which net accumulation equals net loss is called the equilibrium line altitude (ELA). ELAs are therefore dependent on temperature and precipitation, so change with the regional (and global) climate, with ELAs in general being lower during glacial periods (on the order of hundreds of meters) and higher during warmer interglacial periods. If the ELA is physically higher than the surrounding topography then the existence of a glacier is precluded. The rise of ELAs due to global warming is therefore of concern for those glaciers that are near the top of their topographic environment.

Rheology

The dynamic behavior of ice depends on its rheology. Ice can behave as a brittle substance prone to fracturing if it is subjected to rapid stress, or it can behave as a viscous substance if the applied stress has a relatively slow timescale. The timescale separating these behaviors is the Maxwell time, which is determined by the ratio of the kinematic viscosity of ice (units of pressure–time) to the shear modulus of ice deformation (units of pressure). The Maxwell time for typical glacier viscosities (10^{13} Pa s) and shear moduli (3.5×10^9 Pa) is about 50 min. The viscosity, however, is temperature dependent so the Maxwell time is smaller for warmer glaciers (on the order of a few hours) and longer for colder glaciers (on the order of months or more). This fact plays a role in the numerical modeling of glaciers since the discrete time step chosen in models must be less than the Maxwell time in order to ensure numerical stability.

Fracturing by the former, brittle, behavior results in crevasses when the flow of the glacier is over topographic features or there is convergence/divergence with tributary glaciers or sidewalls (Figure 2). Crevasses can be transverse (across the direction of ice flow, resulting from divergence and extension of ice in the direction of flow); longitudinal (aligned with the direction of flow, resulting from cross-flow divergence usually from underlying topography); or chevron (at an angle with valley walls, caused by friction with topography). Rapid radial spreading of ice unconstrained by topography also produces crevasses.



Figure 2 Crevasse on the Godwin-Austen Glacier, at the base of K2. Photo A.B.G. Bush.

Topography and Glaciers

Climate-glacier dynamics are complicated given the coupling of climate, surface processes, and geological conditions and tectonics. A multitude of feedback mechanisms and forcings govern the response of glaciers to climate change. The topography underlying a glacier plays a critical role in its evolution.

The complexity is highlighted by the fact that the topography is the effective result of system couplings, such that energy and water input, erosion and deposition, and mass influx (uplift) and deformation govern the morphological characteristics of the topography. This in turn defines the magnitude of topographic properties such as altitude, slope angle and azimuth, curvature, relief, and regional topographic structure that partially govern the magnitude of numerous surface processes. Glacier processes are uniquely coupled with mass movement and fluvial processes, thereby resulting in extreme variability in glacier dynamics.

It is well known that orbital dynamics and radiative forcing resulted in glacial expansion in the past. Glaciers have tremendous ability to erode the landscape, as related to ice thickness and meltwater production. As lithospheric mass was removed in the past and the geological system responded via isostatic uplift, glacier dynamics and erosion resulted in the formation of deep U-shaped valleys, polished valley sidewalls, steep local slopes, regional relief production, and a reduction of local slopes at high altitude (Figure 3). Glaciers effectively eroded the landscape and left behind



Figure 3 Valley carved by a tributary glacier on the south side of the Baltoro Glacier, Pakistan. Photo M.P. Bishop.

high-altitude erosion surfaces when they retreated. These surfaces have been identified and mapped in numerous mountain environments.

The changing topographic parameters and structure played an important role in the changing evolution of the landscape. Various local and regional topographic parameters govern the magnitude of surface irradiance, which in turn governs the magnitude of evapotranspiration, ablation, and meltwater production. The magnitude of all three surface irradiance components, direct irradiance, diffuse-skylight irradiance, and adjacent-terrain irradiance, are governed by multiscale topographic parameters that dictate extreme spatiotemporal variability in surface irradiance. These topographic effects on individual glaciers and regional patterns of mass balance are not yet well understood. One of these effects, the adjacent-terrain component, is often responsible for the cross-sectional surface morphology of a glacier, as ablation is higher on the sides of a glacier depending upon ice proximity to adjacent surfaces, valley walls, and the azimuth of solar radiation.

The local and regional topographic structure influences the dominance of climate systems and the magnitude of precipitation, which also governs mass balance. The size and orientation of deep glacial valleys serves as a pathway for weather systems, governed by the global climate. For example, in the Himalaya, glacier valleys serve as a conduit for the seasonal monsoon system to penetrate deeper into the range such that annually, glacier systems may be effected by the westerlies and the monsoon. Furthermore, regional altitude variations and steeper slopes produce an orographic precipitation effect thereby influencing the magnitude of precipitation or snowfall. Such an orographic precipitation effect has been documented in the Karakoram of Himalaya, as a water mass anomaly and advancing glaciers highlight a unique topography-climate-glacier dynamic. This dynamic is also spatially coincident with a large number of recently identified surging glaciers, thought to be the result of mass loading (e.g., Scherler et al., 2011).

The topography also governs ice-flow dynamics and glacier surface characteristics. Glacier erosion is thought to scale with ice thickness and basal velocity, such that the slope angle and ice thickness governs the basal shear stress and

erosion potential. Although many other variables are involved, abrasion and the scouring and removal of rock and sediment that is redistributed along the base of the glacier are important. It becomes part of the bed load, and englacial load, and at the terminus of the glacier, may move upward as part of the englacial load to become part of the supraglacial load on the glacier surface. Rapid glacial incision lowers the valley and coupled with uplift, relief production reduces surface ablation over time, due to cast shadows and increased topographic shielding from solar radiation. Steep slopes and positive mass balance increases ice velocity thereby increasing erosion potential. Over time, erosion and redistribution of sediment may reduce slope angles and erosion potential, but this may be offset by ice thickness and meltwater production governed by surface conditions and climate change.

Glacier surface characteristics strongly control ablation and meltwater production. Most glaciers around the world are heavily debris covered (Figure 4). Depending upon the lithology/minerological composition of the debris and its depth, ablation may be significantly enhanced or reduced. Thin debris layers on a glacier surface enhance ablation and meltwater production, while thick debris layers protect the glacier from losing its mass. Debris depth variations over a glacier surface are highly variable and depend on local topography. Debris-covered glaciers are believed to be relatively insensitive to climate change. However, debris cover can vary significantly along a glacier, with more near the terminus and less at higher elevations. The general trend of debris-cover depth is one of the decreasing debris depth with increasing altitude. The terminus of a glacier usually exhibits the greatest debris depths. Measured debris depths can vary from ~0 to ~5 m. Debris can be highly variable from a lithological and particle size distribution perspective and include car-sized and house-sized boulders.

The adjacent topographic conditions regulate the magnitude of supraglacial debris loads. Snow avalanches, landslides, rock falls, and other mass movement processes dictate the quantity of debris that makes its way onto the glacier surface. As the ice flows down the valley there is a build-up of debris depth, such that it is relatively high near the terminus. In general, this is facilitated with relatively high surface velocities



Figure 4 Debris-covered terminus and outflow of the Baltoro Glacier, Pakistan. Photo M.P. Bishop.

and a systematic decrease toward the terminus. Consequently, as the debris builds up over time, and as the glacier is actively moving downhill, sediment transport is facilitated, and there is a feedback between the cohesive nature of the sediment, the slope of the sediment surface, and the sediment discharge per unit area. So in addition to the ice topographic conditions, the sediment topographic conditions also regulate surface irradiance and ablation.

The aforementioned dynamics are initially controlled by regional topographic conditions related to topographic stress tectonism, and climate. Given the paleoclimatic and glacial chronological conditions that have sculpted the surrounding topography, the three-dimensional topographic stress field strongly influences rock strength and slope-failure potential. Consequently, valley-wall slope failures and landslides contribute to the debris load. Active tectonics and earthquakes also generate material, and there may be considerable variation in supraglacial debris characteristics given variations in the topographic and geological setting. High magnitude glacier erosion and uplift is known to generate relief and therefore can enhance debris production.

Glacier surface topography and debris-load variability governs meltwater production and distribution of water, such that supraglacial runoff and streams are common on many glaciers (Figure 5). Surface water may flow into crevasses and moulins to become part of the englacial conduit network, or make its way to the base of the glacier (Figure 6). Frequently, supraglacial runoff accumulates in local topographic depressions thereby forming a supraglacial lake that can drain englacially or subglacially (Figure 7). Differential ablation or rapidly expanding supraglacial lakes can result in the formation of ice cliffs (Figure 8) that are usually characterized by steep slopes and a thin moisture-laden debris layer that dramatically enhances ablation and meltwater production. Many supraglacial lakes are found adjacent to ice cliffs resulting from rapid melting of ice along the ice-cliff surface. The presence of such numerous and complex surface feedbacks that rapidly change the glacier surface warrants further investigation regarding the sensitivity of debris-covered glaciers, and the role of topography in glacier dynamics.



Figure 5 Supraglacial stream on the Baltoro Glacier. Photo A.B.G. Bush.



Figure 6 Moulin on Lilligo Glacier. Photo A.B.G. Bush.



Figure 7 Supraglacial lake on the Godwin-Austen Glacier, with evidence of either englacial or subglacial drainage. Photo A.B.G. Bush.



Figure 8 Ice cliffs on the Biafo Glacier, Pakistan. Photo A.B.G. Bush.

Glacier–Climate Interactions

In general, and in the absence of debris cover, glaciers have a relatively high albedo compared to their surroundings so they reflect a large fraction of solar radiation. The positive ice-albedo feedback can promote glacier growth (if a glacier grows in area it reflects more solar radiation thereby cooling atmospheric temperatures and promoting further growth) or it can promote glacier demise (if a glacier shrinks in area it reflects less solar radiation thereby warming the atmosphere and promoting further shrinkage). The albedo of a glacier is, however, not straightforward. The presence of surface meltwater, new snow, old snow, rock debris, black soot can change the surface albedo and hence the radiation budget at the surface. Rock debris cover, for example, can protect a glacier from melting if the coverage is sufficiently deep that the ice is no longer in direct contact with the atmosphere. Sparse and scattered debris cover, on the other hand, enhances surface melting because the rocks heat under insolation and melt into the ice. The presence of meltwater ponds on a glacier surface dramatically decreases the albedo and enhances melting during the ablation season.

The fact that 10% of Earth's surface is glacierized means that Earth's globally averaged albedo is higher than it would be if there were no glaciers. Earth's global mean temperature is therefore lower than it would be with no ice. Glaciers not only respond to changes in global climate, they also contribute to that climate through atmosphere–ice interactions.

Global Glaciations

There is geologic evidence for glacial periods as far back as the Precambrian era, approximately 2900 Ma. The multiple Snowball Earth events of the Neoproterozoic period may have created global glaciation (continental and marine; e.g., Hoffman and Schrag, 2002). Less extensive glacial events are known to have occurred during the late Ordovician/early Silurian (~430 Ma), the Carboniferous/Permian transition (~280 Ma), and, more recently, the past 2 million years of the Pleistocene.

When looking at glaciations during such deep geological time periods there are many factors that must be taken into account. One is continental configuration; the presence of land mass near the poles is more likely to produce glaciation. For example, the glacial scarring of rocks that provides evidence for the Ordovician glaciations exists in what is now the Sahara Desert but at that time the African plate was situated near the south pole. Another factor is solar luminosity, which slowly increases with time (e.g., luminosity during the Ordovician was approximately 3% lower than it is today).

The third factor is greenhouse gas concentration in the atmosphere. Changes in carbon dioxide over geologic time-scales (tens of millions of years) is a natural cycle involving volcanism and atmosphere–ocean interactions (e.g., during the Ordovician atmospheric carbon dioxide is believed to have been 10–15 times the preindustrial value). Greenhouse gases play a key role in determining the longwave flux into the surface of a glacier and hence influence the energy balance at the surface. All else being equal, higher greenhouse gas concentrations provide more energy available to melt ice.

There is much more proxy data available to study the recent Pleistocene glaciations, which began during a period of global cooling, and that is simply because glaciations tend to eradicate evidence of previous glaciations through their transformation of the landscape (where proxy data are collected). Earth's orbital parameters of obliquity, eccentricity, and precession have played a key role in modulating glaciations for the past 2 million years. During the early Pleistocene, reconstructions from oxygen isotope data from the seafloor indicate a ~41 000-year cycle of glaciations, indicating that changes in Earth's obliquity modulated glacial cycles. During the midlate Pleistocene, however, glaciations switched to a ~100 000-year cycle, indicating that the eccentricity of Earth's orbit became the dominant factor. Why the timing of glaciations switched periods is still an outstanding question, as is the fact that the 100 000-year eccentricity cycle is radiatively quite weak yet the last nine glacial cycles have been on this period.

Current Climate and Glacier Status

Given the importance of glaciers to climate and to freshwater resources around the globe, monitoring efforts have been underway for decades. The past few decades of global warming have been reflected in the near worldwide retreat of glaciers, particularly the humid-maritime glaciers of Western Canada, Alaska, and Patagonia (e.g., Solomon et al., 2007). In the Arctic, for example, virtually all monitored glaciers lost mass

(one exception being Svalbard) with the greatest loss occurring in the most recent years (Sharp et al., 2012).

Projections for the future fate of glaciers around the world rely not only on *in situ* observations, but also on numerical models that simulate the climate under various greenhouse gas forcing scenarios (e.g., Janes and Bush, 2012; Marshall et al., 2011; Pollock and Bush, 2012). As an example, the Canadian Rockies are projected to lose up to 80–90% of ice volume by the end of the century (Marshall et al., 2011). The Himalaya, dubbed ‘the third pole’ given its number of glaciers and their size, also shows trends of negative mass balance in the central and eastern parts of that range. However, according to observations one exceptional region appears to be the Karakoram, which has a number of stable or even advancing glaciers (Scherler et al., 2011). Nevertheless, modeling projections indicate that by the end of the century, even the glaciers in this anomalous region will only be able to exist at the highest elevations (Janes and Bush, 2012).

There is still much work to be done on glaciers in terms of mass balance, the streamflow that they produce, the erosional characteristics that they exhibit, the biogeochemical nature of subglacial waters, and ultimately their evolving impact on humans and our freshwater resources.

Acknowledgments

ABGB acknowledges funding from the Natural Sciences and Engineering Council (NSERC) of Canada, as well as the Canadian Institute for Advanced Research (Earth System Evolutionary Program).

See also: **Arctic and Antarctic:** Antarctic Climate; Arctic Climate. **Climate and Climate Change:** Overview.

References

- Box, J., Cappelen, J., Chen, C., Decker, D., Fettweis, X., Mote, T., Tedesco, M., van de Wal, R.S.W., Wahr, J., 2013. Greenland ice sheet. In: Arctic Report Card, 2012. http://www.arctic.noaa.gov/reportcard/greenland_ice_sheet.html.
- Hoffman, P.F., Schrag, D.P., 2002. The snowball Earth hypothesis: testing the limits of global change. *Terra Nova* 14, 129–155.
- Janes, T., Bush, A.B.G., 2012. The role of atmospheric dynamics and climate change on the possible fate of glaciers in the Karakoram. *Journal of Climate* 25, 8308–8327.
- Marshall, S.J., White, E.C., Demuth, M.N., Bolch, T., Wheate, R., Menounos, B., Beedle, M.J., Shea, J.M., 2011. Glacier water resources on the eastern slopes of the Canadian Rocky Mountains. *Canadian Water Resources Journal* 36, 109–134.
- Pollock, E.W., Bush, A.B.G., 2012. Climate change in western North America caused by CO₂ rise: a coupled atmosphere-ocean model simulation. *Atmosphere-Ocean* 50 (1), 70–85.
- Scherler, D., Bookhagen, B., Strecker, M.R., 2011. Spatially variable response of Himalayan glaciers to climate change affected by debris cover. *Nature Geoscience* 4, 156–159.
- Sharp, M., Wolken, G., Geai, M.L., Burgess, D., 2012. Mountain glaciers and ice caps (outside Greenland). In: Arctic Report Card, 2012. http://www.arctic.noaa.gov/reportcard/glaciers_ice_caps.html.
- Shepherd, A., et al., 2012. A reconciled estimate of ice-sheet mass balance. *Science* 338 (6111), 1183–1189. <http://dx.doi.org/10.1126/science.1228102>.
- Solomon, S., et al., 2007. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press, Cambridge.

Permafrost

TE Osterkamp, University of Alaska, Fairbanks, AK, USA
CR Burn, Carleton University, Ottawa, ON, Canada

© 2003 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 4, pp 1717–1729, © 2003, Elsevier Ltd.

Introduction

Permafrost, ground that remains at or below 0 °C for two or more years, underlies about a fifth of the land surface of the Earth. Permafrost terrain consists of an active layer at the surface that freezes and thaws each year, underlain by perennially frozen ground. The top of permafrost is at the base of the active layer, and the base of permafrost occurs where the ground temperature rises above 0 °C at depth (**Figure 1**).

The first scientific reports on permafrost were published in the 1830s by the Royal Geographical Society of London. These papers reported the thickness of frozen ground in a well at Yakutsk, Russia, and provided instructions to officers of the Hudson's Bay Company on describing the phenomenon. The first systematic study published in English of perennially frozen ground was prompted by strategic considerations in World War II, when the Alaska Highway was built through northern British Columbia and Yukon to Alaska, with the associated Canol

Pipeline. At that time, the term 'permafrost' was coined by S. W. Müller as a contraction of 'permanent frost'.

Permafrost grows by freezing from its base downward or, when new material is added to the ground surface, from the top upwards, or by a combination of these processes. The first permafrost on Earth likely formed prior to or during the first ice age about 2.3 billion years ago, and its occurrence, distribution, and thicknesses have varied in response to repeated ice ages throughout Earth's history.

There is scientific and geotechnical interest in permafrost principally because it contains ice, is close to its thawing point, and is sensitive to changes in surface conditions caused by human activities and climate. Temperatures in permafrost present retrievable records of past climate, and climatic change commonly leaves an imprint on the stratigraphy of ground ice. Ice-rich permafrost contains ice in excess of the water content at saturation, with the ice masses commonly distributed as lenses, millimeters to centimeters in thickness, within the ground. Massive ground ice with dimensions typically from meters to tens of meters also occurs in permafrost (**Figure 2**).

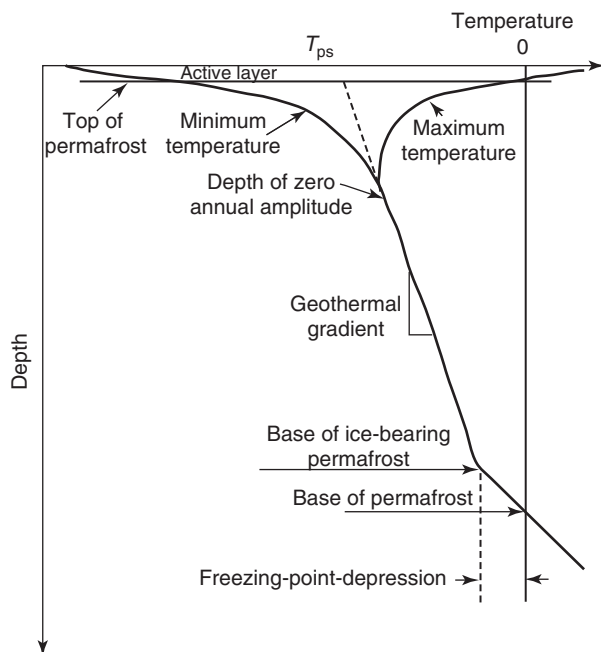


Figure 1 Schematic temperature profiles in permafrost, illustrating the annual maximum and minimum temperatures. Annual mean permafrost surface temperature, T_{ps} , is obtained by extrapolating the common linear portion of the profiles upwards. Soil particle effects, solutes, and hydrostatic pressure decrease the phase equilibrium temperature so that a layer just above the base of the permafrost does not contain ice. The change in slope of the temperature profile at the base of ice-bearing permafrost is caused by the difference in thermal conductivities between the frozen and unfrozen ground.

Occurrence, Distribution, and Thickness

The spatial extent of permafrost generally changes with climate, but there can be considerable regional variation because of snow cover and other factors. Over half of Canada and Russia, most of Alaska, and north-east China are underlain by continental permafrost, while alpine permafrost is found at high elevations in middle and low latitudes (e.g., the summit of Mauna Kea in Hawaii). Some permafrost is found in Scandinavia, but the spatial extent is much less than at corresponding latitudes of North America, because of the warming influence of the Gulf Stream. Permafrost containing water ice and other ices is known to exist on Mars and on other bodies in our solar system.

Permafrost regions are divided into zones with varying spatial extent of perennially frozen ground. In the continuous permafrost zone (**Figure 3**), more than 90% of the ground is underlain by permafrost, and it is usually absent only beneath rivers and lakes that do not freeze to their bottom in winter. At continental scale, the limiting annual mean air isotherm for continuous permafrost is about -8°C . At warmer temperatures, variations due to microclimatic effects lead to a zone of widespread discontinuous permafrost where 50 to 90% of the ground is underlain by permafrost. In the sporadic discontinuous permafrost zone, 10 to 50% of the ground is underlain by permafrost.

Subsea permafrost formed during glacial periods, when lower sea levels exposed large areas of the polar continental shelves. The cold air temperatures and long periods of exposure allowed permafrost to grow there to great depths.



Figure 2 Massive ground ice exposed at the Beaufort Sea coast near Tuktoyaktuk, Northwest Territories. The ice was formed during permafrost aggradation after deglaciation, with water supplied from the underlying sands. Banding indicates variations in the concentration of sediment suspended in the ice. Undulation of the banding indicates displacement of the ice subsequent to formation. Photograph by J. R. Mackay. See Mackay JR and Dallimore SR (1992) Massive ice of the Tuktoyaktuk area, western Arctic coast, Canada. *Canadian Journal of Earth Sciences* 29: 1235–1249.

Thicknesses of several hundreds of meters are known to exist in these shelves and, at present, this permafrost is slowly degrading beneath the relatively warm and salty ocean. However, several tens of thousands of years are required to thaw it completely.

The distribution and thickness of permafrost are controlled by factors that influence its heat balance and heat flow within it. Since pore spaces in permafrost are generally blocked by ice, heat flow is by conduction. While the presence of permafrost is due primarily to climate, considerable modification of the temperature between the atmosphere and permafrost may occur, owing to vegetation, energy exchanges at the snow-ground surface, transfer of heat through the active layer, and local geological and hydrological conditions. In particular, the winter snow cover buffers the ground from frigid air temperatures, and the annual mean ground surface temperature, T_s , is commonly 2 to 4 °C warmer than the annual mean air temperature, T_a . In the summer months, shade from vegetation and a supply of soil moisture for evaporation are two significant site variables that reduce the ground surface temperature. T_a and T_s tend to be similar at windswept sites where there is little snow accumulation. Significant changes in ground temperatures occur across the continental tree line, for within the forest there is less wind at ground level than on the tundra, and hence the snow is deeper, less dense, and a better insulator.

An organic horizon at the ground surface commonly assists the development and persistence of permafrost. Dry moss and peat have low thermal conductivities, reducing heat flow into the ground in summer. However, autumn rain characteristically increases the water content of the moss and peat, increasing their conductivity (particularly when frozen) and facilitating

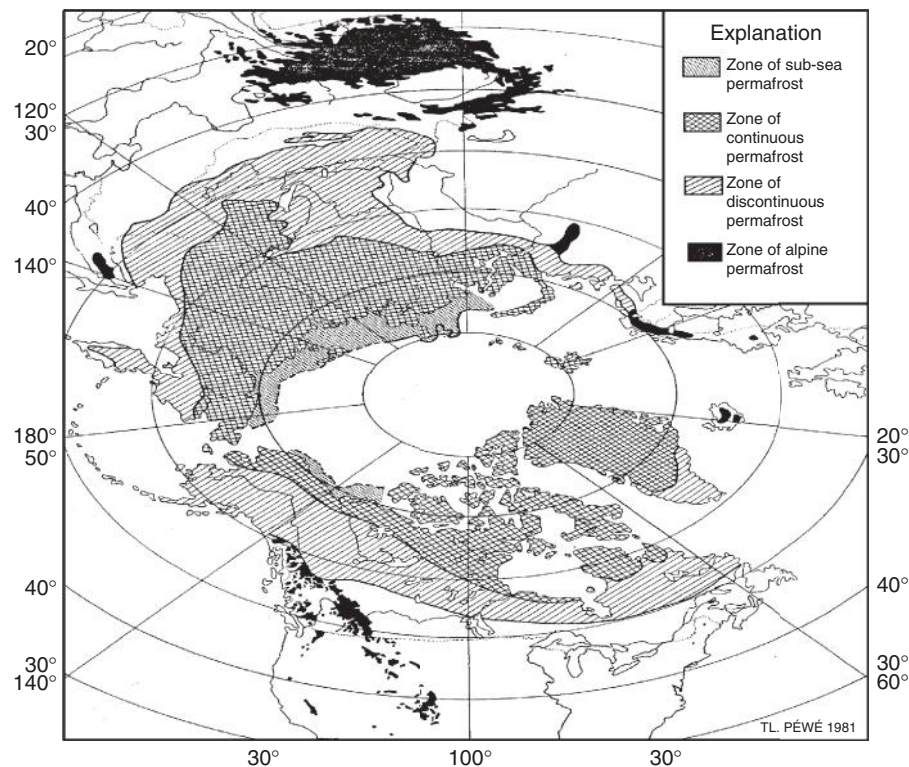


Figure 3 Generalized map of the approximate distribution of permafrost in the north circumpolar region of the Earth. Direct data (probing, drilling, sampling, temperature measurements) are scarce, so the map is somewhat unreliable, especially at local scales in discontinuous permafrost. Subsea permafrost typically exists in the continental shelves of the Arctic Ocean where seabed temperatures remain negative. Adapted from Péwé TL (1983) Alpine permafrost in the contiguous United States: a review. *Arctic and Alpine Research* 15: 145–156, by permission of the Regents of the University of Colorado.

extraction of heat from the ground in winter. These principles indicate why, near the southerly limit for continental permafrost in North America, permafrost generally occurs in peat lands and under thick moss layers.

At regional scale, permafrost distribution may be controlled by topography, in terms both of aspect, modifying the receipt of radiation, and of cold conditions brought by high elevation. However, in the mountain and plateau regions of Alaska and adjacent Canada, strengthening of winter inversions by cold-air drainage into valley bottoms enhances conditions for permafrost there.

The thickness of permafrost is commonly influenced by geothermal heat flow and bedrock stratigraphy. At continental scale, heat flow varies by a factor of four between stable craton, where the flow is low, and tectonically active terranes. This geothermal heat is conducted through the permafrost where the temperature gradient varies with the thermal conductivity of the bedrock. At local and regional scales, the movement of groundwater carries heat that can modify the spatial distribution and thickness of permafrost. Under equilibrium conditions where a constant annual mean permafrost surface temperature, T_{ps} , has existed for a long time, the thickness of homogeneous permafrost, Z_e , is governed by T_{ps} , the thermal conductivity of permafrost, K_p , and the geothermal heat flow, J , where (from Fourier's law)

$$Z_e \approx \frac{K_p}{J} T_{ps} \quad [1]$$

Equation [1] provides reasonable estimates of Z_e when T_{ps} remains near its long-term value or has been close to it for a sufficient period. While T_a is the principal variable governing T_{ps} , the values differ owing to the effects of snow cover, the active layer, and other factors mentioned above.

Properties and Processes

Unfrozen Water and Ice

Seasonally frozen ground and permafrost contain unfrozen water (Figure 4) and ice in equilibrium at temperatures less than 0 °C as a result of the effects of soil particles and solutes. In the absence of solutes, temperature, T , and the soil's specific surface area, S , are the primary determinants of the amount of unfrozen water, θ_u , and are empirically related by

$$\ln \theta_u = 0.2618 + 0.5519 \ln S - 1.449S^{-0.264} \ln |T| \quad [2]$$

where θ_u is in percent. Unfrozen water reduces ground thermal conductivity and distributes latent heat over a range of temperatures, so that temperature changes require freezing or thawing throughout the frozen ground. These effects retard the thermal response of the active layer and permafrost.

Frost Heave

Unfrozen water is also responsible for the development of ice lenses in the ground during freezing, leading to local uplift (heave) of the surface. The water occurs in mobile films on particle surfaces, where it is held in tension. Water flows along the tension and thermal gradients in the films to cooler regions

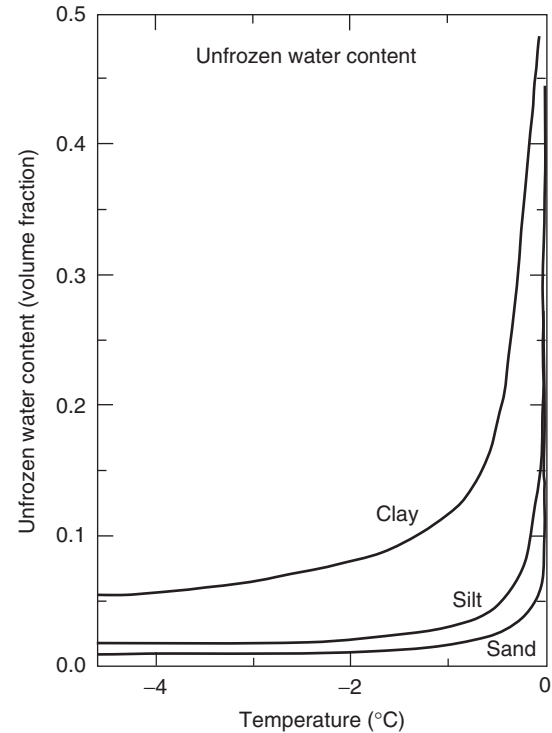


Figure 4 Representative values for the temperature dependence of unfrozen water contents in sand, silt, and clay. Unfrozen water contents typically increase with temperature and finer-grained soil and are small in moss and peat. The presence of solutes increases unfrozen water contents.

of the ground. When the water content in the freezing soil exceeds saturation, the excess volume separates soil particles to form layers, or lenses, of segregated ice. Ice segregation and frost heave characteristically occur in fine-grained soils, where the unfrozen water content is sufficiently large to conduct water into the freezing ground. In coarse-grained soil, the unfrozen water content is small and the permeability is too low to allow water migration during freezing. In saturated, coarse-grained soil, the expansion of water during freezing is accommodated by expelling the excess water into unfrozen ground ahead of the freezing front. When the freezing system is closed, pore-water expulsion may also lead to heave as hydrostatic pressure deforms the overlying frozen ground.

Active Layer

Permafrost is separated from the atmosphere by a boundary layer consisting of the active layer and vegetation in summer with the snow cover added in winter. The active layer transmits heat to and from permafrost, reduces the amplitude of thermal variations at the top of permafrost compared with the ground surface, is the medium through which moisture and gases are exchanged between the permafrost and the atmosphere, and provides water and nutrients for biological processes. In permafrost terrain, the active layer supports plant and animal communities since virtually all biological activity below ground occurs within it.

Permafrost immediately below the active layer is characteristically ice-rich. This ice-rich zone acts as an impermeable barrier to drainage, so much permafrost terrain is wet. The ice-rich zone is the reason permafrost terrain is considered sensitive to disturbance, for deepening of the active layer commonly leads to subsidence (thaw settlement) and accelerated erosion as the ice melts.

In a dry active layer, $T_s \approx T_{ps}$; however, in a wet one, there is an asymmetry in heat flow because the frozen thermal conductivity, K_f , exceeds K_t , the thawed conductivity. This makes T_s warmer than T_{ps} and the difference is the thermal offset (Figure 5)

$$T_{ps} - T_s = \frac{I_t}{P} \left(\frac{K_t}{K_f} - 1 \right) \quad [3]$$

where I_t is the thawing index at the ground surface and P is the period (365 days).

Once the snow melts, the ground surface warms above 0°C and thawing of the active layer begins (Figure 6). In a simplified model, its maximum thickness is

$$X \approx \sqrt{\frac{2K_t I_t}{h}} \quad [4]$$

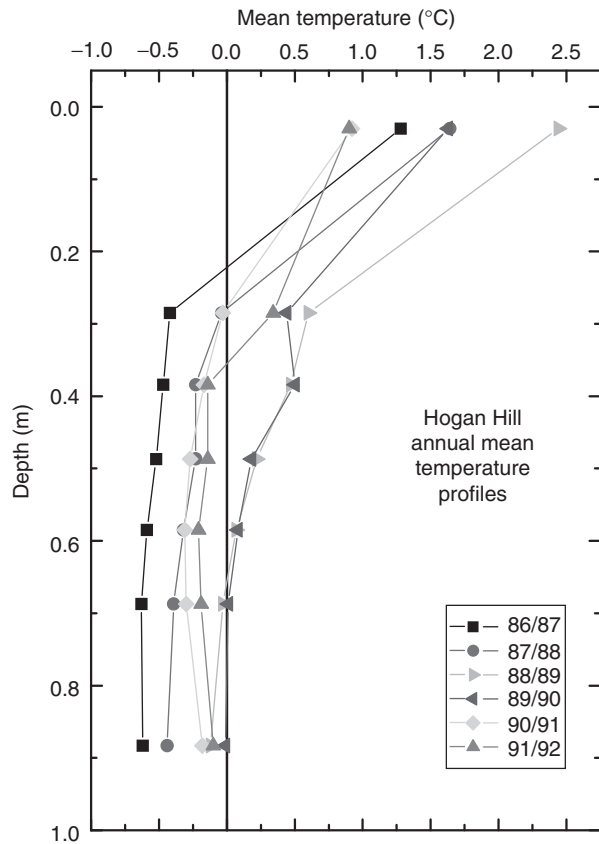


Figure 5 Annual mean temperature profiles in the active layer showing the effects of thermal offset. In the context of climatic change, thermal offset allows permafrost to survive changes that produce annual mean ground surface temperatures warmer than 0°C over multiyear periods as shown at this site. Adapted with permission from Osterkamp TE and Romanovsky VE (1999) Evidence for warming and thawing of discontinuous permafrost in Alaska. *Permafrost and Periglacial Processes* 10: 17–37.

where h is the volumetric latent heat of the ground, which depends on the ice content. The thickest active layers (1 to 2 m or more) develop in dry materials, characteristically bedrock, sand, and gravel. Thin active layers are common in wet organic soils, where there may be considerable amounts of ice at the beginning of summer, and where the thermal conductivity of a dry surface layer is very low. The thinnest active layers, less than 30 cm thick, occur in the High Arctic, owing to short, cool summers.

The active layer typically reaches X and begins freezing upward from the bottom when $T_s > 0^\circ\text{C}$, one or two weeks before freezing downward from the surface. Temperature changes during freezing are retarded by the presence of unfrozen water for a few weeks in cold permafrost and for much of the winter in warm permafrost. The rate of freezing down from the surface depends on the amount and timing of the early winter snow cover and on the thermal conductivity and moisture content of the active layer. Freezing may be slow if a deep snow cover is established early in winter on a wet active layer. Temperatures in the portion of the active layer that remain thawed are constrained near 0°C , a phenomenon termed the ‘zero curtain’, until freezing is complete.

Thermal Regime

A temperature profile in permafrost that is homogeneous, does not contain unfrozen water, and is in equilibrium with a constant long-term T_{ps} , is a straight line described by Fourier’s law (seasonal variations near the surface and freezing point depression near the base cause deviations from a straight line). However, the thermal regime of permafrost is often characteristically different during its formation and growth, after surface temperatures change, and during thawing. Prior to permafrost formation, interannual variability in T_s and in conditions within the seasonally frozen layer can cause the depth of freezing to exceed the depth of thawing during the following summer, resulting in a temporary layer of frozen ground. T_s can be significantly positive when this occurs, owing to thermal offset. A long-term shift in conditions allows such a layer to persist and grow, creating a thickening layer of permafrost with time. When the permafrost freezes from its base downward, the solution of the Stefan problem yields the approximate depth to the bottom of the growing permafrost at time t :

$$Z \approx \sqrt{\frac{2K_p (-T_{ps})t}{h}} \quad [5]$$

In eqn [5], K_p and h depend on ground properties. Equation [5] neglects geothermal heat flow and the effects of freezing point depression near the base of permafrost (Figure 1). It reduces to $Z \approx \text{constant} \cdot \sqrt{t}$, where the constant is typically about 1 to 5 for a wide range of conditions, where t is measured in years and Z in meters.

In principle, permafrost grows until $Z = Z_e$, when it is in equilibrium with T_{ps} . However, in reality, T_{ps} is variable over the long time scales needed to grow even relatively thin permafrost, resulting in permafrost thicknesses that vary with long-term variations in T_{ps} (Figure 7).

A change in T_{ps} (warming or cooling) produces a curvature in the temperature profile that penetrates deeper with time (Figure 8). The magnitude of the change at any depth can be

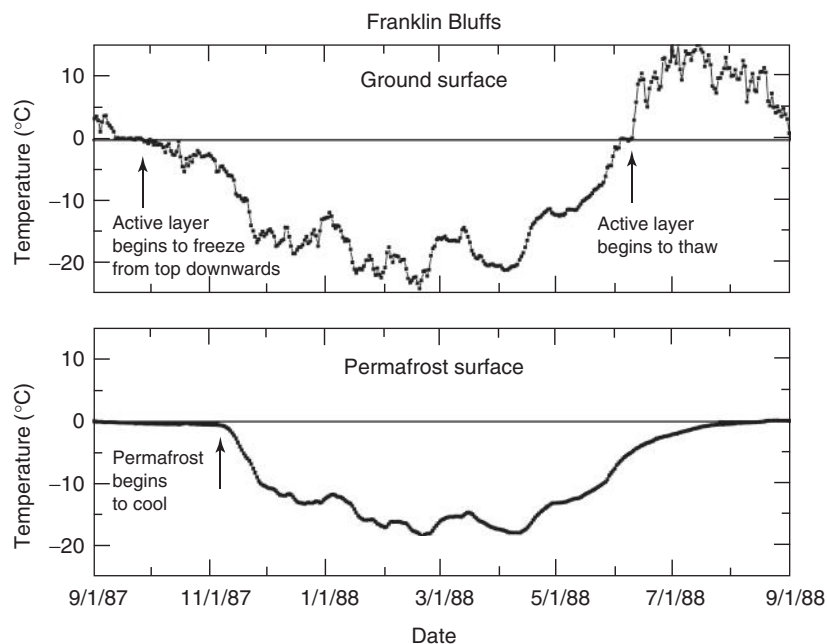


Figure 6 Time series of temperatures at the ground and permafrost surfaces for the annual cycle. Freezing of the active layer from the top downward began about 26 September 1987, and from the bottom upward about ten days earlier. The zero curtain persisted until about 12 November 1987, the date of freeze-up of the active layer. At this time, the lower portion of the active layer and upper portion of the permafrost began to cool. Ground surface temperatures during spring, 1988, remained near 0 °C from 4 to 11 June (the period of snowmelt), when the active layer began to thaw.

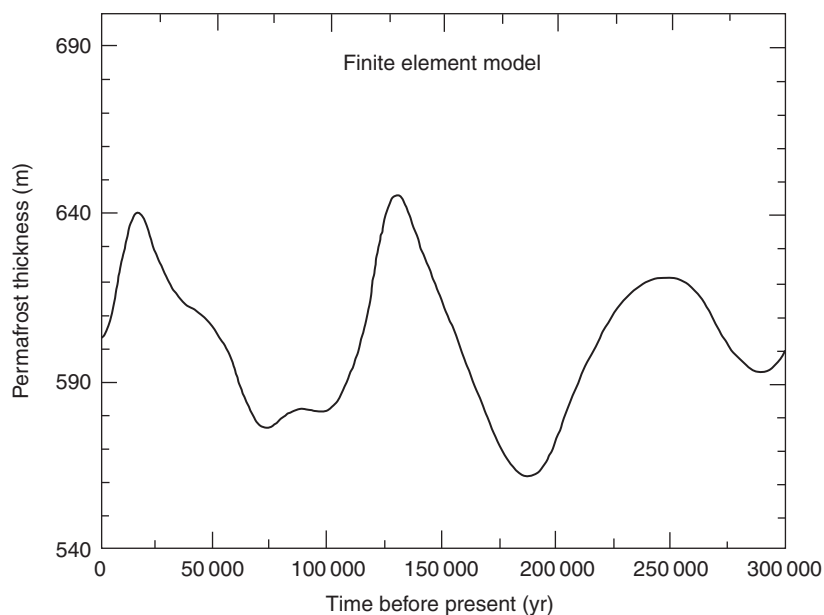


Figure 7 Calculated permafrost thickness variations in response to changes in paleoclimate. Thicknesses varied about 84 m ($562 \text{ m} \leq Z \leq 646 \text{ m}$) with maximum thawing rates of 5 mm yr^{-1} and freezing rates of 2 to 3 mm yr^{-1} . Adapted with permission from Osterkamp TE and Gosink JP (1991) Variations in permafrost thickness in response to changes in paleoclimate. *Journal of Geophysical Research* 96: 4423–4434. Copyright © by the American Geophysical Union.

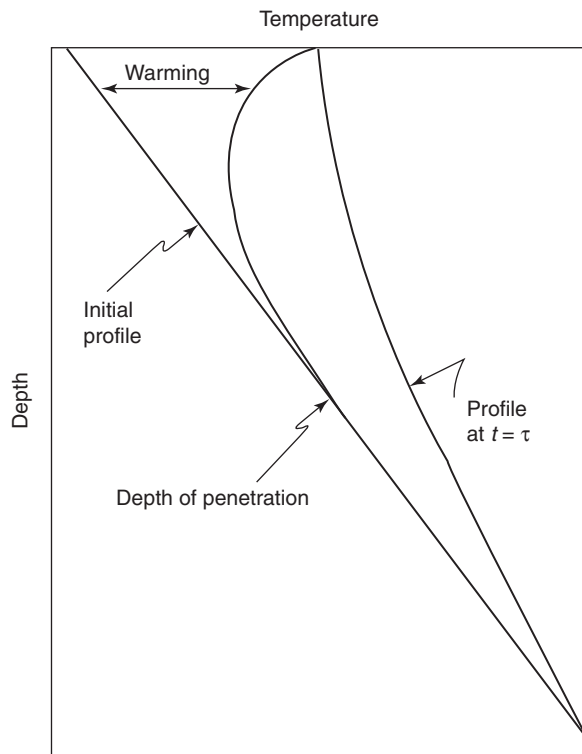


Figure 8 Schematic thermal response of permafrost to warming of its surface temperature. Seasonal variations near the surface are not shown. The magnitude of the warming at any depth can be obtained directly from the measurements and the timing of the warming can be calculated from the depth of penetration of the warming signal. Adapted with permission from Lachenbruch AH, Sass JH, Marshall BV, and Moses TH (1982) Permafrost, heat flow, and the geothermal regime at Prudhoe Bay, Alaska. *Journal of Geophysical Research* 87: 9301–9316. Copyright © by the American Geophysical Union.

obtained from the measurements and the timing can be calculated from the maximum depth of penetration of the thermal signal. For permafrost with thickness Z , the time scale required for the temperature profile to respond to a new surface condition is

$$\tau \approx \frac{Z^2}{4D} \quad [6]$$

where D is the thermal diffusivity of the permafrost. Freezing or thawing at the bottom of the permafrost begins once the thermal disturbance has penetrated there, with calculated rates in thick continuous permafrost that are typically millimeters per year (Figure 7).

When $T_{ps} \approx 0^\circ\text{C}$, interannual variability in T_s and conditions within the active layer may cause the depth of summer thaw to exceed the depth of winter freezing, resulting in a temporary residual thaw layer between seasonally frozen ground and permafrost. T_s can be significantly positive when this occurs, and interannual variability may allow the layer to refreeze and thaw repeatedly. If the conditions that caused the warming persist, the layer may become too thick to refreeze. Then, the permafrost is decoupled from the atmosphere and warms continuously throughout the year. An equation similar to [5] can be obtained for the thickness of the thawed layer

above permafrost. Thawing rates at the top of the permafrost depend on the temperatures at the bottom of the former active layer, K_v and h and can be a few tenths of a meter per year for near-surface permafrost. While the permafrost is thawing, temperatures within it warm very slowly as they approach 0°C (Figure 9) because of the effects of unfrozen water.

Geomorphic Features

There are three principal geomorphic features unique to permafrost terrain: pingos (conical ice-cored hills) and ice-wedge polygons, both associated with permafrost aggradation; and thermokarst terrain, associated with ground thawing. Pingos may form in the unfrozen sandy sediments of drained-lake bottoms that are completely surrounded by permafrost. After drainage, lake sediments freeze primarily from the top downward and from the sides and bottom of the talik (the unfrozen layer) inward. Freezing of the sands results in pore-water expulsion into the enclosed talik, increasing hydrostatic pressure there. The increased pressure lifts the permafrost, and the water freezes in place, to create a core of ice in the mound (Figure 10). Pingos grow until the talik freezes completely. There are 1350 pingos in Canada's western Arctic, and the largest, Ibyuk Pingo, is 50 m high, over 1200 years old, and is

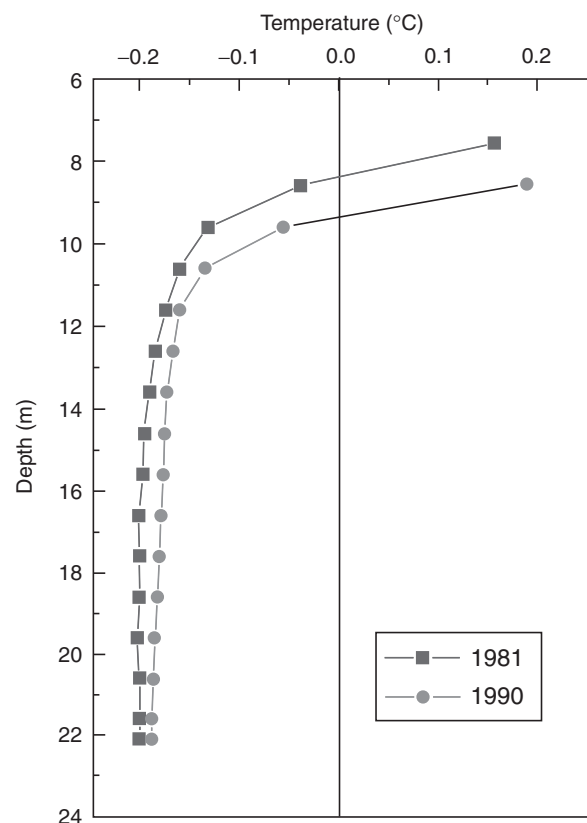


Figure 9 Temperature profiles in thawing permafrost near Fairbanks, Alaska obtained 9 years apart. The presence of unfrozen water retards temperature changes and produces the observed curvature below 8 m depth. Adapted with permission from Osterkamp TE and Romanovsky VE (1999) Evidence for warming and thawing of discontinuous permafrost in Alaska. *Permafrost and Periglacial Processes* 10: 17–37.



Figure 10 Pingo near Wolf Lake, Richards Island, western Arctic coast, Canada, about 20 m high, with the ice core exposed after collapse of the central portion. The vertical extent of exposed ice in the far headwall is about 4 m. Photograph by C. R. Burn. See Mackay JR (1998) Pingo growth and collapse, Tuktoyaktuk Peninsula area, western Arctic coast, Canada: a long-term field study. *Géographie physique et Quaternaire* 52: 271–323.

growing about 2 cm taller annually. Pingos are also found at the base of hill slopes, where the pressure to lift permafrost may be provided by confined groundwater flowing down slope.

Ice-wedge polygons are ubiquitous in continuous permafrost terrain, where cracks occur in the permafrost as a result of thermal contraction in winter. The cracks relieve thermal stress approximately normal to their axes, yielding a polygonal network that may subdivide further as more cracks form. Ice wedges begin to grow in the cracks when spring snowmelt infiltrates them and freezes to form near-vertical sheets of ice. Repetition of the process over many years leads to development of wedges of ice, which commonly extend downwards from the base of the active layer for about 4 m (Figure 11), although some exceed 10 m in depth. Polygons are clearly expressed in lowlands, but movement of the active layer commonly obscures them on hill slopes. Ice-wedge growth deforms the surrounding ground to accommodate the additional volume. Growth of the ice wedges often forces the adjacent ground upwards and laterally, creating a trough above the ice wedge.

Warming or disturbance to permafrost terrain usually leads to deepening of the active layer and thawing of the ice-rich zone at the top of permafrost or of near-surface massive ground ice, causing local subsidence. This thaw settlement produces a pitted relief, called thermokarst terrain. Drainage conditions determine whether standing water will be present or not. When depressions in thermokarst terrain collect water, the disturbance to permafrost is enhanced, leading to growth of thermokarst lakes. ‘Beaded’ streams occur when a series of pools (beads) form along the stream as a result of thawing ice-rich permafrost or large ice masses. Most features of thermokarst terrain range from 1 to 100 m in lateral dimensions, although thermokarst lakes are often larger. Thermokarst depressions up to 100 km² in area and 5 to 20 m in depth form part of the landscape of Siberia, where they are called ‘alasses’.

Thermokarst terrain also includes retrogressive thaw slumps (Figure 12) that commonly develop where ice-rich permafrost is exposed by erosion in riverbanks, lakeshores, and along the coast, or by other processes. These features, with a near-vertical



Figure 11 Ice wedge exposed in a coastal bluff near Tuktoyaktuk, Northwest Territories, Canada, about 4 m tall. The photograph was taken in May 1992, when the winter’s thermal contraction crack was visible. Photograph by C. R. Burn. See Mackay JR (2000) Thermally induced movements in ice-wedge polygons, western Arctic coast: a long-term study. *Géographie physique et Quaternaire* 54: 41–68.



Figure 12 Headwall of a retrogressive thaw slump in ice-rich glaciolacustrine sediments near Mayo, Yukon Territory, Canada. Lenses of segregated ice give the exposure its texture. The ice lenses developed when permafrost formed in glacial lake sediments after drainage of the lake at the end of the last glaciation. A deep active layer, which formed during the warmest climate of the last 10 000 years, left a thaw unconformity in the sediments marked by the abrupt transition between the darker, lower, ice-rich material, and the lighter sediment above. Photograph by C. R. Burn. See Burn CR (2000) The thermal regime of a retrogressive thaw slump near Mayo, Yukon Territory. *Canadian Journal of Earth Sciences* 37: 967–981.

retreating headwall and a low-angled foot slope, are the commonest form of landslide in permafrost terrain. Landslides involving only the thawing active layer also occur, particularly in areas of fine-grained, ice-rich soil, when the active layer and vegetation detach from the underlying frozen material. If ice-rich permafrost or massive ground ice is exposed, the slope failure may develop into a thaw slump.

Impacts of Climatic Warming

The Earth's climate has generally warmed since the mid-to late 1800s. In the Arctic and Sub-Arctic, permafrost has also warmed, particularly in Russia, China, Mongolia, Alaska, and western Canada. In Alaska, T_{ps} for continuous permafrost warmed 2 to 4 °C over the last century followed by a cooling in the early 1980s and then a warming of up to 3 °C since then. In central Alaska and the Yukon Territory of Canada, discontinuous permafrost has warmed since 1970 as a result of changes in air temperatures and snow cover. While it has warmed typically 1 to 2 °C since the late 1980s in Alaska, the ground has been cooling in the Yukon owing to a reduction in snow cover. In the Northwest Territories, warming of permafrost by about 1 °C has been observed since the early 1970s, but in northern Québec there has been cooling. Such regional variations are to be expected as a result of spatial climatic variability. However, the general signal from the Northern Hemisphere indicates permafrost warming, and in some areas permafrost is presently thawing at both top and bottom, the southern boundary of the permafrost is moving northward, and the occurrence of thermokarst terrain is increasing.

Current global circulation models predict that air temperatures in the Arctic will rise 2 to 5 °C during the next half-century,

with the warming in winter larger than in summer and with more precipitation throughout the year. While there is considerable uncertainty in these predictions, warming of T_{ps} by a few degrees would have serious consequences for permafrost regions, especially for discontinuous and sporadic discontinuous permafrost.

In the continuous zone, the active layer, thaw lakes, coastal processes, landscape processes, eolian activity, and vegetation would be sensitive to climatic warming. The effect on the permafrost would be to warm it and, possibly, to change the depth of the active layer. Thawing at the base of the permafrost would start several centuries or more later.

Since most of the discontinuous and sporadic discontinuous permafrost is within a few degrees of thawing, it can be expected to warm, to begin thawing from the top and bottom with an increase in the incidence of thermokarst terrain, and, eventually, to disappear. Although many centuries would be required for the permafrost to disappear, thawing of the warmest permafrost from the top would begin immediately.

Thawing of permafrost as a consequence of human activities and as a result of the climatic warming since the late 1800s serves as a model for what may be expected to occur with additional climatic warming. Where permafrost contains massive ground ice or is ice-rich, extensive differential thaw settlement has occurred, with damage to the natural terrain and to infrastructure. Human-induced thaw settlement is at present responsible for damage to infrastructure on permafrost where the built structures have raised ground temperatures above 0 °C (Figure 13). The magnitude of the thaw settlement is typically 1 to 3 m, but can exceed 5 m (vertical settling) of the ground surface. Repair of the infrastructure is costly and some structures, airports, and roads have been abandoned.



Figure 13 Longitudinal cracks in a road embankment near Fairbanks, Alaska. Snow removal during winter produces a berm along the shoulder and slope of the embankment that warms the underlying ice-rich permafrost, causing it to thaw and settle. Settlement under the shoulder and slope causes the edge of the embankment to tilt outward, putting the top surface of the embankment in tension, which eventually results in cracks. Patches (darker asphalt) in the pavement are a result of two long cracks and the guardrail on the far side of the road has sagged as a result of the thaw settlement. Photograph by T. E. Osterkamp.

Carbon and Trace Gases

Soils in northern regions may contain one-quarter to one-third of the Earth's total soil carbon pool, with much of it stored in frozen peat in near-surface permafrost. When permafrost thaws, the carbon is cycled through terrestrial ecosystems with the gaseous end products (CO_2 and CH_4) emitted to the atmosphere. While the details of this process are not well understood and are subject to environmental constraints, there is evidence that tundra regions may have shifted from being a carbon sink to a carbon source.

Gas Hydrates

Temperature and pressure conditions within and under thick permafrost are favorable for the formation and existence of gas hydrates that are a potentially abundant source of energy. Warming of the permafrost would eventually destabilize these hydrates, producing large quantities of gas (primarily CH_4) that may find its way into the atmosphere. Long time scales (many centuries or millennia) are required for this process. However, subsea permafrost in the continental shelves of the Arctic Ocean that was submerged and warmed by seawater more than a few thousand years ago could be emitting methane at present. Currently, there is too little information to adequately assess this problem.

Ecosystems

In areas of ice-rich permafrost, thaw settlement and development of thermokarst terrain destroys the substrate on which the current ecosystems rest, dramatically changing the nature of the ecosystems. The effects have been observed to include:

1. Destruction of trees and reduction in area of boreal forest ecosystems.
2. Expansion of thaw lakes, grasslands, and wetlands.
3. Destruction of habitat for caribou and terrestrial birds and mammals.
4. Formation of new habitat for aquatic birds and mammals.
5. Coastal and riverbank erosion.
6. Clogging of salmon spawning streams with sediment and debris.
7. Slope instabilities, thaw slumps, landslides, and erosion.
8. Talik development, with increasing depth to water table.
9. Increased methane emissions in wet areas.

Thermokarst has been observed to result in the partial or complete destruction of some ecosystems and their conversion to other types of ecosystems. In one lowland area in central Alaska, permafrost degradation is widespread and rapid, causing large shifts in ecosystems from birch forests to fens and bogs (Figure 14). If current conditions persist, the remaining birch forests will be eliminated by the end of the twenty-first century.

See also: Arctic and Antarctic: Antarctic Climate. Biogeochemical Cycles: Sulfur Cycle. Chemistry of the Atmosphere: Methane. Climate and Climate Change: Energy



Figure 14 An area in the Tanana River valley near Fairbanks, Alaska, showing ponds, floating fens and remnant birch forest underlain by ice-rich permafrost. Thawing is resulting in complete destruction of the trees and forest ecosystem, which is being converted into floating fens with ponds. Standing dead birch trees, some on ground that has settled below the water level, are visible in the center and left half of the picture. Photograph by M. T. Jorgenson. Adapted with permission from Osterkamp TE, Viereck L, Shur Y, *et al.* (2000). Observations of thermokarst in boreal forests in Alaska. *Arctic, Antarctic, and Alpine Research* 32: 303–315 by permission of the Regents of the University of Colorado.

Balance Climate Models. **Cryosphere:** Snow (Surface). **Global Change:** Biospheric Impacts and Feedbacks; Climate Record: Surface Temperature Trends. **Hydrology, Floods and Droughts:** Soil Moisture. **Mountain Meteorology:** Cold Air Damming. **Satellites and Satellite Remote Sensing:** Aerosol Measurements; Precipitation; Temperature Soundings.

Further Reading

- Andersland, O.B., Ladanyi, B., 1994. *An Introduction to Frozen Ground Engineering*. Chapman & Hall, New York.
- Brown, R.J.E., 1970. *Permafrost in Canada: Its Influence on Northern Development*. University of Toronto Press, Toronto.
- Gold, L.W., Lachenbruch, A.H., 1972. Thermal Conditions in Permafrost—A Review of North American Literature. *North American Contribution, Permafrost, 2nd International Conference, 13–18 July 1973, Yakutsk, USSR*, pp. 3–23. National Academy of Sciences, Washington, DC.
- Johnston, G.H. (Ed.), 1981. *Permafrost: Engineering Design and Construction*. Wiley, New York.
- Kudryavtsev, V.A., Garagula, L.S., Kondrat'yeva, K.A., Melamed, V.G., 1974. *Osnovy merzlotnogo prognoza, MSU. [VA Kudryavtsev et al. (1977) Fundamentals of Frost Forecasting in Geological Engineering Investigations. Draft Translation 606. Hanover, NH: US Army CRREL.]*
- Muller, S.W., 1947. *Permafrost or Permanently Frozen Ground and Related Engineering Problems*. Edwards, Ann Arbor, MI.
- Tsytovich, N.A., 1975. *The Mechanics of Frozen Ground*. McGraw-Hill, New York.
- Von Baer, K.E., 1838. On the ground ice or frozen sod of Siberia. *Journal of the Royal Geographical Society* 8, 210–213.
- Washburn, A.L., 1979. *Geocryology: A Survey of Periglacial Processes and Environments*. Arnold, London.
- Williams, P.J., Smith, M.W., 1989. *The Frozen Earth: Fundamentals of Geocryology*. Cambridge University Press, New York.
- Yershov, E.D., 1990. *Obshchaya Geokriologiya. Nedra, Moscow*. [Williams P.J. (Ed.), 1998. *General Geocryology*. Cambridge: Cambridge University Press.]

Sea Ice

MC Serreze and F Fetterer, University of Colorado, Boulder, CO, USA
WF Weeks (Retired), University of Alaska Fairbanks, Fairbanks, AK, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by W. F. Weeks, volume 5, pp 2047–2054, © 2003, Elsevier Ltd.

Synopsis

Sea ice is any form of ice found at sea that originated from the freezing of seawater. It is the most visible feature of the polar seas, with its extent waxing and waning with the seasons. Ice thickness is highly variable, ranging from a thin veneer to tens of meters. While the existence of sea ice reflects the cold conditions inherent to high latitudes, sea ice also strongly modulates the energy budget and climate of the polar seas and beyond, particularly because of its high albedo and through acting as a lid, insulating the underlying ocean from a generally much colder atmosphere. Sea ice extent and thickness are in turn influenced by changing climate conditions. Arctic sea ice extent and thickness as observed at the end of the melt season have declined significantly over the past several decades. In contrast, Antarctic sea ice extent has expanded slightly.

The Seasonal Cycle of Sea Ice

Ice Extent

Sea ice in the Northern Hemisphere attains its maximum seasonal extent in March and its minimum seasonal extent in September. At maximum extent, it can cover more than 15×10^6 km², somewhat less than twice the size of the contiguous United States. It covers essentially all of the Arctic Ocean and extends down the western side of the major ocean basins, paralleling the pattern of cold ocean currents, and reaching the Gulf of St. Lawrence (Atlantic Ocean) and the Sea of Okhotsk (Pacific Ocean). The most southerly site in the Northern Hemisphere where an extensive cover sometimes forms is the Bohai Sea, located off the east coast of China at 40°N. At the seasonal minimum, sea ice is largely confined to the central Arctic Ocean, with extensions into the Canadian Arctic Archipelago and along the east coast of Greenland. Through the 1970s and 1980s, September ice extent averaged around 7.0×10^6 km². As assessed over the period of satellite observations (late 1978 to present), there are downward linear trends in sea ice extent for all months; they are largest for September and smallest during winter. As of this writing, September 2012 held the record for the lowest ice extent in the satellite record, with a monthly average of 3.61×10^6 km². In satellite-derived sea ice maps, extent is generally defined as the region covered by ice with a concentration (fractional ice cover) of at least 15%. **Figure 1** illustrates seasonal contrasts in the coverage of sea ice for the Northern Hemisphere, using 2012 as an example.

The seasonal cycle of ice extent in the Southern Hemisphere is, of course, broadly opposite to that of the Northern Hemisphere, varying between about 3×10^6 km² in March and 20×10^6 km² in September (**Figure 2**). The larger seasonal range in ice extent for the Southern Hemisphere reflects differences in the spatial distributions of land and ocean. The Arctic Ocean is in large part landlocked, with only one major exit for sea ice: Fram Strait, located between Greenland and Svalbard. This constrains the maximum possible winter extent. The Southern Ocean is by comparison essentially unbounded

to the north, which allows unrestricted drift of the ice in that direction and results in the summertime melting of nearly all of the previous season's growth. In sharp contrast to the Arctic, Antarctic sea ice extent as a whole has remained fairly stable over the period of satellite observations and indeed has expanded slightly, with downward trends in some areas offset by positive trends in others. Extent in 2012 was the highest over the period of satellite coverage, acting as a counterpoint to the record low in the Northern Hemisphere. The significance of this record high for Antarctica, and of the record 2012 low for the Arctic, is quite different, however. As assessed over the period of satellite coverage, the steepest Arctic ice extent trend, for September, is $-13.0 \pm 2.9\%$ (relative to the mean over the period 1979–2000). Arctic ice for September 2012 was about 50% below its 1979–2000 average. For March, the trend in extent is $-2.6 \pm 0.6\%$. In contrast, the Antarctic ice extent trend is insignificant for March at 3.2% ($\pm 4.1\%$), while that for September is barely significant at 0.9% ($\pm 0.6\%$). Antarctic sea ice in September 2012, while at a record high, was less than 5% above its 1979–2000 average.

Surface Characteristics

Seasonal changes in ice extent are accompanied by pronounced changes in surface characteristics. As the polar night descends, any ice that survived the summer melt season, as well as new ice that forms in areas of open water, becomes covered with snow. As will be outlined in this article, dynamic forces acting on the ice may result in the development of linear fractures, called leads, as well as areas of ridged ice. In winter, with surface air temperatures usually well below freezing, water within an open lead will quickly freeze. Leads in the Arctic often form networks that extend over hundreds of kilometers, yet the overall Arctic ice concentrations remain high, generally above 95% (i.e., less than 5% open water), except near the ice margin. Refrozen leads in turn become covered with snow. Snow depths by the end of winter range widely depending on the region, the age of the ice, wind drift, and scour, but values of 30–60 cm can be considered typical.

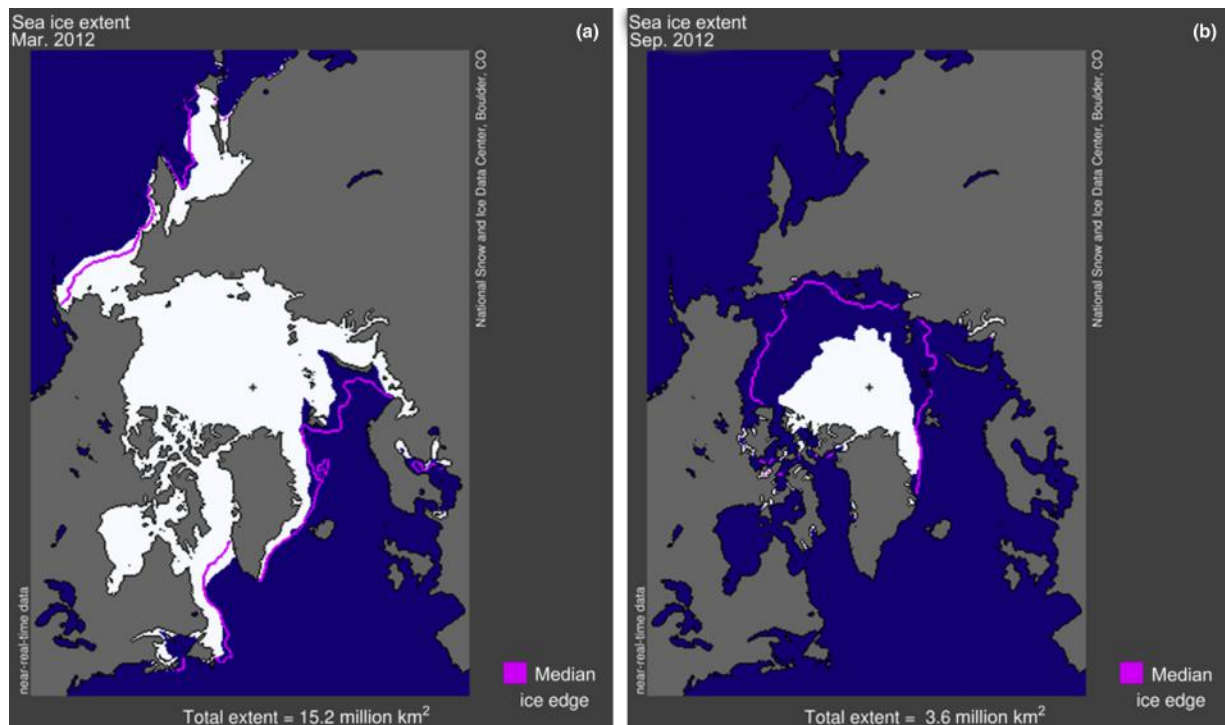


Figure 1 Sea ice extent for the Northern Hemisphere for (a) March 2012 and (b) September 2012. Ice extent is defined as the region with an ice concentration of at least 15%. Arctic ice extent for September 2012 was the lowest over the period of satellite observations. The pink line shows the median ice extent for March and September, over a base period of 1979–2000. Reproduced from the Sea Ice Index, http://nsidc.org/data/seaice_index/; Fetterer, F., Knowles, K., Meier, W., Savoie, M., 2002, updated 2009. Sea Ice Index. National Snow and Ice Data Center, Boulder, Colorado USA. Digital media.

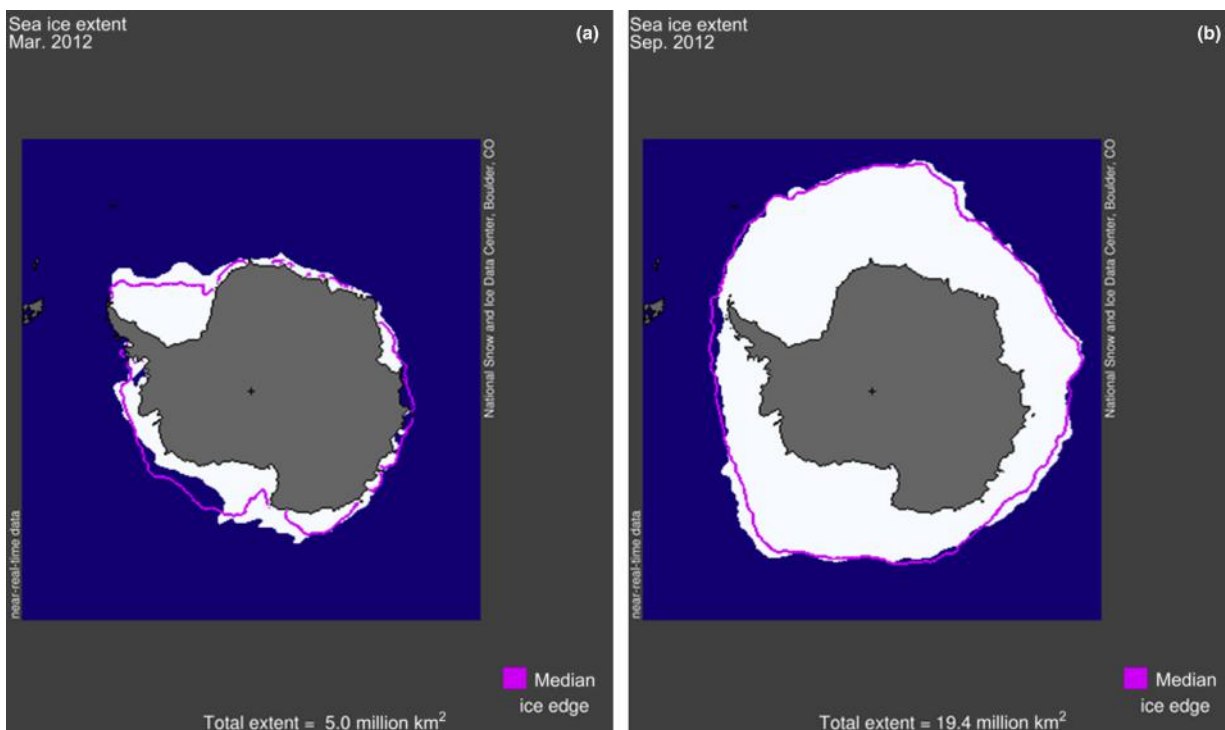


Figure 2 Sea ice extent for the Southern Hemisphere for (a) March 2012 and (b) September 2012. Antarctic ice extent for September 2012 was the highest observed over the period of satellite observations. Reproduced from the Sea Ice Index, http://nsidc.org/data/seaice_index/; Fetterer, F., Knowles, K., Meier, W., Savoie, M., 2002, updated 2009. Sea Ice Index. National Snow and Ice Data Center, Boulder, Colorado USA. Digital media.

With spring, sunlight returns. Because of the overlying snow, the ice-covered part of the ocean may have an albedo (the reflection coefficient for visible-band radiation) as high as 0.85, contrasting sharply with open water, which, except at large solar zenith angles, has an albedo of less than 0.10. As spring advances, it becomes too warm for new ice to form in leads. Snow cover overlying the ice begins to melt, and, as a result, the surface albedo falls. During summer, melt ponds form (Figure 3). As more of the ice surface becomes covered with dark melt ponds, the regional surface albedo further decreases, fostering even more melt. As summer progresses, the open spaces between floes (leads), broadly linear in winter, become more amorphous. Ice floes move more freely and become more rounded, and melt ponds may drain. With the advent of autumn, remaining melt ponds freeze. New ice again starts to form in open-water areas, and the pack takes on its wintertime appearance of linear leads and jagged-edged snow-covered floes. Figure 4 gives examples of how autumn, early melt season, and late melt season ice appears in high-resolution satellite visible-band imagery. The description given here best characterizes the Northern Hemisphere sea ice cover. The Antarctic ice cover is more dynamic and behaves somewhat differently; in particular, the ice tends to be more fractured and summer melt ponds are less extensive.

Ice Formation and Thickness

Initial ice formation occurs at the surface in the form of small platelets and needles termed frazil ice. Frazil crystals are generally less than 3–4 mm. Continued cooling forms a slurry of unconsolidated frazil crystals, termed grease ice. Under calm conditions, frazil crystals freeze together, forming a solid, continuous ice cover of 1–10 cm thickness. However, in the more typical situation, a solid ice cover is inhibited by wind-induced turbulence in the water. Winds and waves advect

frazil ice downwind, where accumulations of up to 1 m thick can form in front of obstacles such as existing ice floes. Once the fraction of ice-covered area exceeds 30–40%, there is sufficient bonding between individual ice crystals to reduce their mobility, initiating the transition to a solid ice cover. Once the initial ice cover is formed, growth occurs mainly by bottom accretion.

Sea ice in the polar ocean has seasonally and regionally varying proportions of first-year (FY) ice, representing the ice growth of a single year, and multiyear (MY) ice, defined as ice that has survived at least one melt season. While, historically, some of the MY ice in the Arctic was over 10 years of age, much of the oldest MY ice is now gone. The majority of ice in the Antarctic is, by contrast, FY ice.

Because ice is a thermal insulator, the thicker it is the slower it grows, other conditions being equal. And because sea ice either ablates or stops growing during the summer, there is a maximum thickness of undeformed FY ice that can form during a specific year. The value depends upon the local climate and oceanographic conditions, reaching slightly over 2 m in the Arctic and as much as almost 3 m at certain protected Antarctic sites. Snow, an excellent insulator, will slow the growth rate of ice. During the winter, the heat flux from areas of open water into the polar atmosphere is significantly greater than the flux through even thin ice, and as much as 200 times greater than the flux through MY ice. This means that even if open-water and thin-ice areas comprise only a few percent of the winter ice pack, as is typical, these areas must still be considered in order to obtain realistic estimates of ocean–atmosphere thermal interactions.

If an FY ice floe survives a summer to become MY ice, the thickness of the ice added during the subsequent winter will be less than the change in thickness of nearby FY ice. This is because the ice growth at the bottom of the MY floe starts later in the season and occurs more slowly. Nevertheless, by the end of the winter, the second-year ice will be thicker than the nearby



Figure 3 Researchers on ponded sea ice. Reproduced from the National Oceanic and Atmospheric Administration/Department of Commerce online photo library (<http://www.photolib.noaa.gov/>).

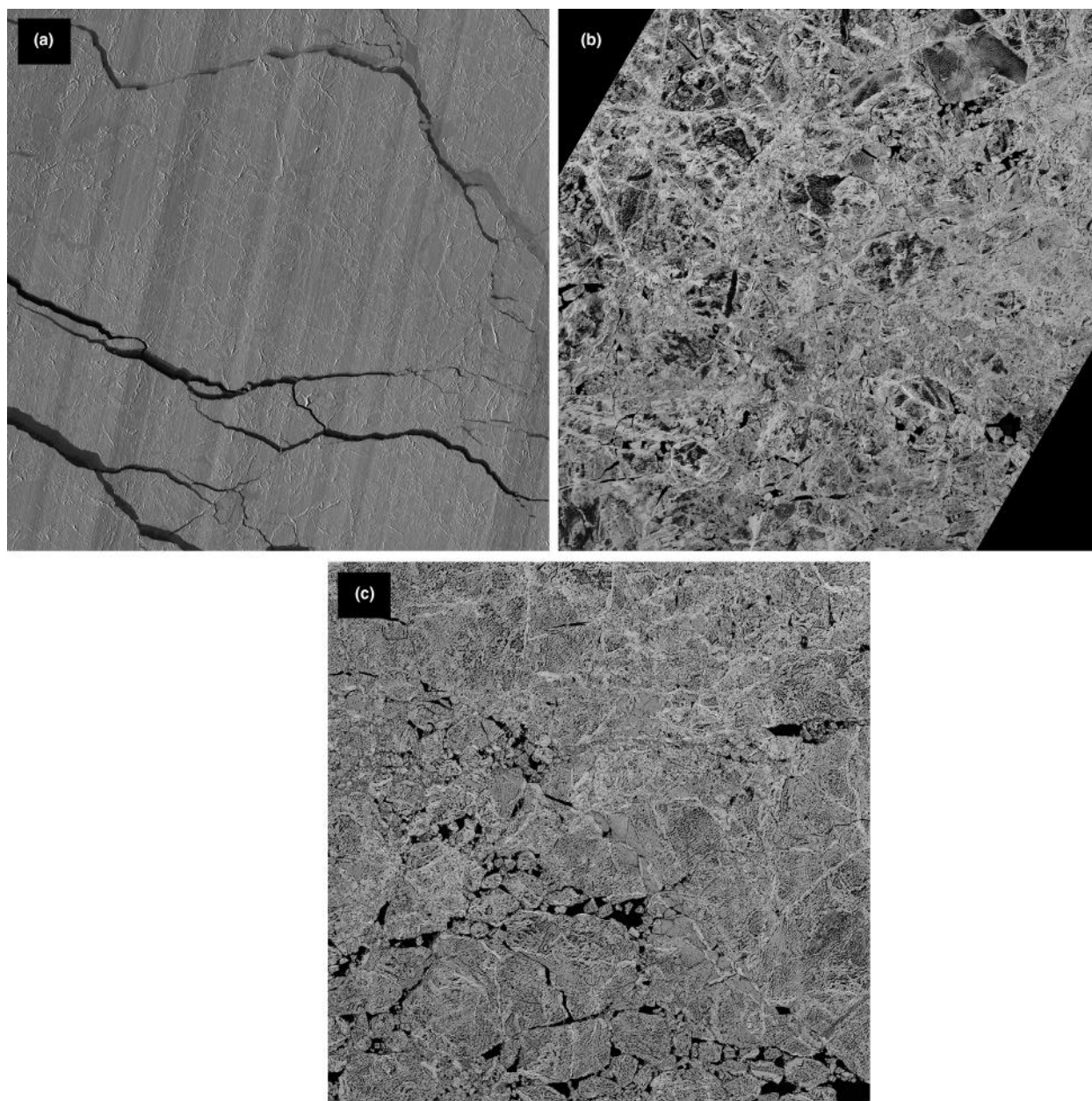


Figure 4 Satellite visible band imagery of the Beaufort Sea. (a) In winter, leads freeze over rapidly; new ice appears smooth and relatively dark. The new ice lightens in appearance as it grows thicker, and it may be dusted by snow (image acquired 10 October 1997). (b) Snow melts soon after the surface air temperature rises above 0 °C. Ponds may cover nearly the entire surface of thinner, younger, less deformed ice, as shown in this image from 22 June 1998. (c) Later in the melt season, ponds generally do not cover as much of the surface. Image from 21 July 1998. All images cover 5 km by 5 km areas. Reproduced from Fetterer, Untersteiner, 2000. SHEBA Reconnaissance Imagery. National Snow and Ice Data Center, Boulder, Colorado, USA. http://nsidc.org/data/docs/noaa/g02180_sheba/index.html.

FY ice. Assuming that the above process is repeated in subsequent years, some ice will be ablated away each summer (largely from the top) and some added each winter (largely on the bottom). As the years pass, the ice melted on top each summer will remain the same (assuming no change in the climate over the ice), while the ice forming on the bottom will become less and less as a result of the increased insulating effect of the thickening overlying ice. Ultimately, a rough equilibrium will be reached, with the winter addition equaling the summer

ablation. At least in the past, steady-state MY ice floes in the Arctic could be layer cakes of 10 or more annual layers with total thicknesses in the range of 3.5–4.5 m. Much of the uncertainty in estimating the equilibrium thickness of such floes comes from uncertainties in the oceanic heat flux. In sheltered fiord sites in the Arctic where the oceanic heat flux is presumed to be near zero, MY fast ice (fast ice is ice that is locked to the shore) with thicknesses of up to roughly 15–20 m has been observed.

Conditions in the Antarctic are rather different. Surface melt rates within the ice pack are small compared with those in the Arctic. The stronger winds and lower humidity encountered over the ice pack also favor evaporation and minimize surface melting. The limited ablation that occurs appears to be controlled by heat transfer processes at the ice–water interface, so that the ice remains relatively cold throughout the summer. However, because most of the Antarctic sea ice is advected rapidly to the north, where it encounters warmer water at the Antarctic convergence and melts rapidly, only small amounts of MY ice remain at the end of summer.

Drift and Deformation

If sea ice was motionless, ice thicknesses would be controlled completely by the thermal characteristics of the lower atmosphere and the upper ocean. The ice cover would presumably have thicknesses and physical properties that would change slowly and continuously from region to region. However, even a casual examination of an area of pack ice reveals striking local lateral variations in ice thicknesses and characteristics. These changes are caused by ice movements produced by the forces exerted on the ice by winds and ocean currents. Such motions are rarely uniform and lead to the building of stresses. If these stresses become large enough, cracks (leads) may form and widen. Leads can vary in width from a few meters to several kilometers and in length from a few hundred meters to several hundred kilometers. As discussed in this article, during much of the year in the polar seas, once a lead forms, it is immediately covered with a thin skin of ice that thickens with time. This is an ever-changing process associated with the movement of weather systems as one lead system becomes inactive and is replaced by another oriented in a different direction.

While the process of lead formation will result in a variety of ice thicknesses, the pack ice thickness distribution observed in the Beaufort Sea in 1976 (Figure 5) reveals a significant amount of ice thicker than the 4.5–5.0 m maximum expected for steady-state MY ice floes. This thicker ice forms by the closing of leads, which commonly results in the piling of broken ice fragments into long, irregular pressure ridges, with a ‘sail’ extending above the water surface paired with a ‘keel’ extending below the surface. There are many small ridges, and large ones are rare. Nevertheless, the large ridges are very impressive: a sail height of 13 m and keel depth of 47 m have been reported in the Arctic (values not from the same ridge). Particularly heavily deformed ice commonly occurs in a roughly 150 km wide band running between the north coast of Greenland and the Canadian Arctic Islands and the south coast of the Beaufort Sea. The limited data available on Antarctic pressure ridges suggest they are generally smaller and less frequent than those in the Arctic Ocean. The general pattern of the ridging is also different in that the long, sinuous ridges seen in the Arctic Ocean are not common. Instead, the deformation can be better described as irregular hummocking accompanied by extensive rafting of one floe over another. Floe sizes are also smaller as the result of the passage of large-amplitude swells through the ice. These are generated by the intense Southern Ocean storms that move to

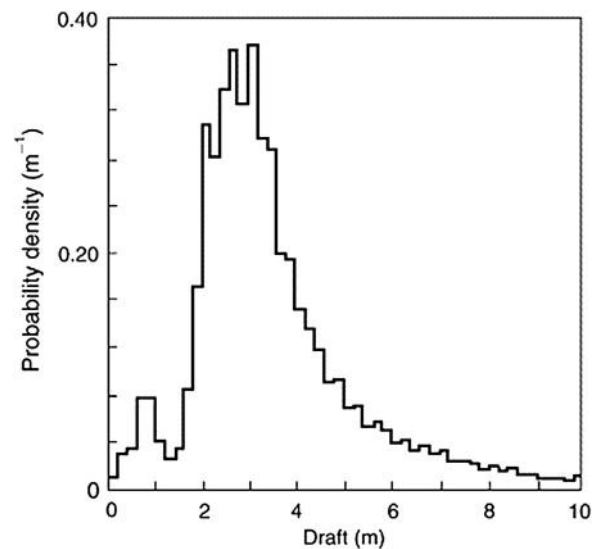


Figure 5 The distribution of sea ice drafts expressed as probability density as determined via the use of upward-looking sonar along a 1400 km track taken in April 1976 in the Beaufort Sea. All ice thicker than ~4 m is believed to be the result of deformation. The peak probabilities in the range between 2.4 and 3.8 m represent the thicknesses of undeformed multiyear ice, while the values less than 1.2 m come from ice that formed more recently in leads.

the north of the ice edge and result in the fracturing of the larger floes, with the large vertical motions facilitating the rafting process.

Pressure ridges are of considerable importance. First, they change the surface roughness at the air–ice and water–ice interfaces, thereby altering the effective surface tractions exerted by winds and currents. Second, they act as plows, forming gouges in the sea floor up to 8 m deep when they ground and are pushed along by the ungrounded pack as it drifts over the shallower (<60 m) regions of the polar continental shelves. Third, as the thickest sea ice masses, they are a major hazard that must be considered in the design of offshore structures. Finally, and most important, the ridging process provides a mechanical procedure for transferring the thinner ice in the leads directly and rapidly into the thickest ice categories. Figure 6 provides a schematic of basic sea ice types, deformation processes, and morphology. Figure 7 illustrates various sea ice formations, including pressure ridges, from photographs taken at the surface and from aircraft.

The large-scale pattern of ice drift in the Arctic has two main features: the Beaufort Gyre, a large clockwise circulation centered in the Beaufort Sea, and the Transpolar Drift Stream, a motion of ice from the Siberian shelves and across the central Arctic Ocean to the Fram Strait. Nearly all of the sea ice that exits the Arctic Ocean does so through the Fram Strait. This large-scale pattern of ice drift broadly reflects the distribution of sea level pressure and hence the surface wind field over the Arctic Ocean (see Figure 2 in Arctic Climate). The time required for the ice to complete one circuit of the Beaufort Gyre averages 5 years, while the transit time for the Transpolar Drift Stream is roughly 3 years. There are many interesting features of the ice drift that occur over shorter time intervals. For instance, in response to the often high frequency of cyclonic storm systems

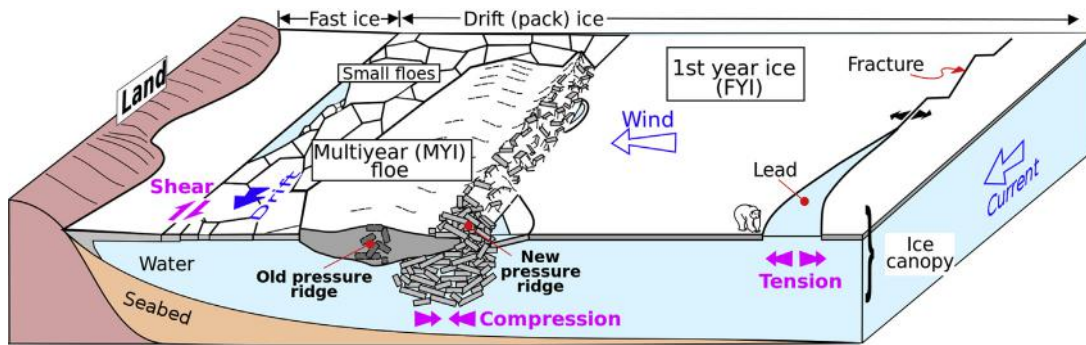


Figure 6 Schematic of basic sea ice types, morphology, and processes. Reproduced from Wikimedia Commons, author Luslier, licensed under the Creative Commons Attribution-Share Alike 3.0 unreported license http://en.wikipedia.org/wiki/Sea_ice.

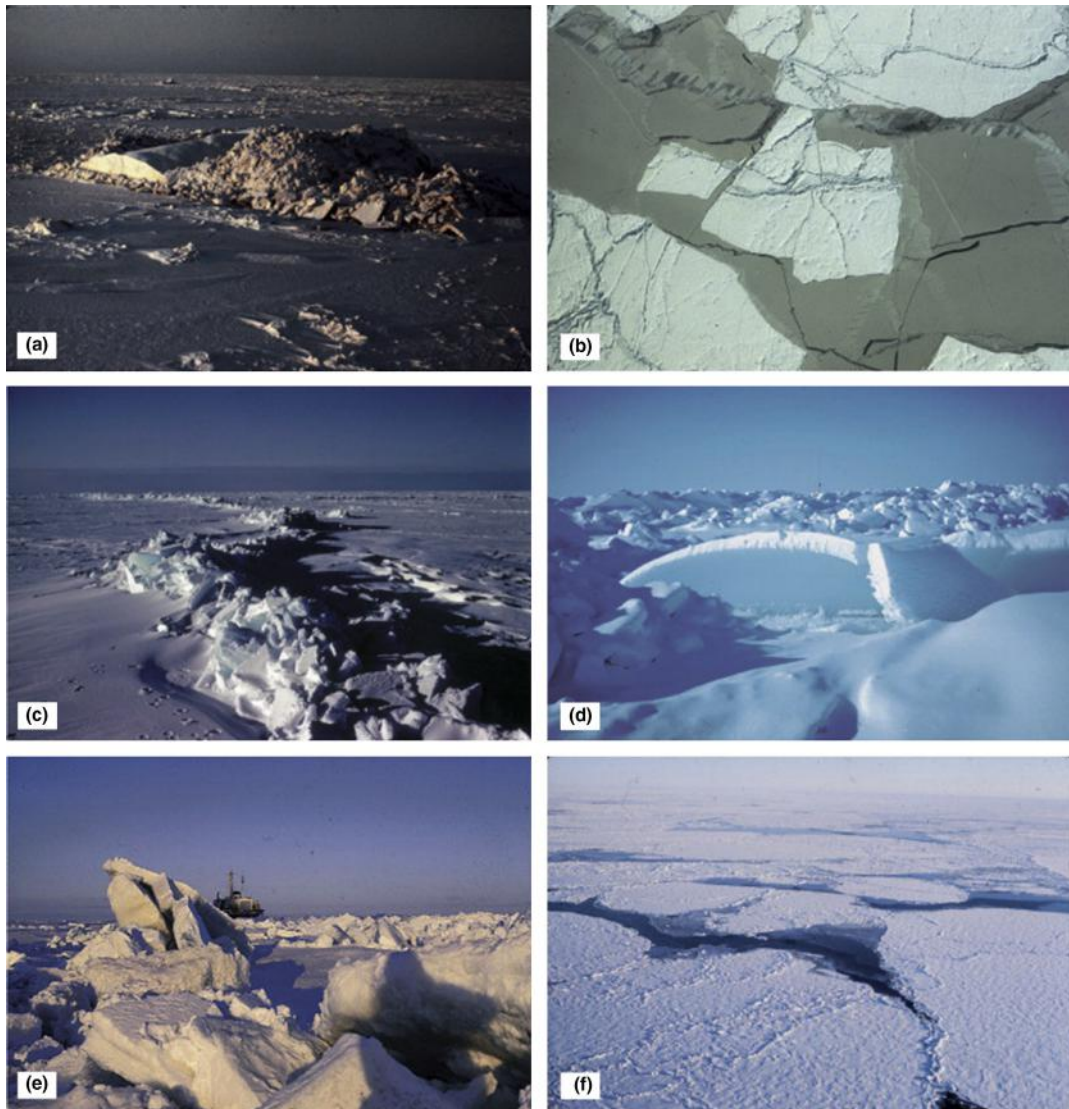


Figure 7 (a) Ice gouging along the coast of the Beaufort Sea. (b) Aerial photograph of an area of pack ice in the Arctic Ocean showing a recently refrozen large lead that has developed in first-year ice. The thinner newly formed ice is probably less than 10 cm thick. (c) A representative pressure ridge in the Arctic Ocean. (d) A rubble field of highly deformed first-year sea ice developed along the Alaskan coast of the Beaufort Sea. The tower in the far distance is located at a small research station on one of the numerous offshore islands located along this coast. (e) Deformed sea ice along the Northwest Passage, Canada. (f) Aerial photograph of pack ice in the Arctic Ocean.

over the central Arctic Ocean in summer and early autumn, the Beaufort Gyre may run backward (counterclockwise) over appreciable periods of time. Typical pack ice velocities range from 0 to 20 cm s⁻¹, although extreme velocities of up to 220 cm s⁻¹ (4.3 knots) have been recorded during storms. During winter, periods of zero ice motion are not rare. During summer, when considerable open water is present in the pack, the ice appears to be in continuous motion. The highest drift velocities are observed near the edge of the pack. Not only are such locations commonly windy, but also the floes are able to move toward the free edge with minimal interfloe interference. Ice drift near the Antarctic continent is generally westerly, becoming easterly further north, but in all cases showing a consistent northerly diverging drift toward the free ice edge.

Geophysical Significance

In winter, sea ice suppresses vertical exchanges of heat (turbulent latent and sensible heat fluxes) between the upper ocean, with its temperature close to the salinity-adjusted freezing point, and a much colder atmosphere (sometimes lower than -40 °C). At the same time, sea ice limits the upward longwave radiation flux from the surface. Snow cover that may be present atop the sea ice cover adds to the insulating effect. In spring and summer, air-sea temperature differences are much smaller than in winter so the insulating effect is not as important. However, because of the high surface albedo of the sea ice cover, much of the solar energy is reflected back to space; even bare ice has a much higher albedo than open water. Where the ocean is largely covered by melting sea ice in summer, the surface has a constant skin temperature. This serves to limit the magnitude of turbulent heat fluxes and the upward longwave radiation flux. All of these effects can be viewed as helping to reinforce the basic temperature gradient between the polar regions and lower latitudes that drives poleward atmospheric energy transports. On a more regional basis, the boundary between sea ice and open water at the surface during winter is characterized by sharp horizontal temperature gradients extending through a considerable depth of the atmosphere. Consequently, shifts in the location of the winter ice margin can alter the formation and tracks of extratropical cyclones and their associated precipitation. Polynyas (semipermanent areas of open water and thin ice at sites where climatically much thicker ice would be anticipated) are another source of moisture to the polar atmosphere, and they can affect precipitation and cloud cover.

Changes in the brine volume of sea ice have geophysical significance as well. When seawater freezes, roughly one-third of the salt in the seawater is initially trapped within the ice in the form of brine inclusions. As a result, initial ice salinities are typically in the range of 10–12 ppt. At low temperatures (<-8.7 °C), solid hydrated salts also form within the ice. The composition of the brine in sea ice is a function of the temperature, with the brine composition becoming more saline as the temperature decreases. Therefore, the brine volume (the volumetric amount of liquid brine in the ice) is determined by the ice temperature and the bulk ice salinity. Not only does temperature vary vertically in the ice floe but also salinity decreases as the ice ages, reaching a bulk value of about

3 ppt in MY ice. Brine volumes are usually lower in the colder upper portions of the ice and higher in the warmer lower portions. They are particularly low in the part of MY ice above sea level, from which the salt has drained almost completely. The upper layers of thick MY ice produce excellent drinking water when melted. Hence, while brine volume is the single most important parameter controlling the thermal, electrical, and mechanical properties of sea ice, these properties exhibit large changes both vertically in the same ice floe and between floes of differing ages and histories. To add complexity to this situation, exactly how the brine is distributed within the floe also affects ice properties such as ice strength and its electro-magnetic characteristics.

The Arctic Ocean is characterized by a low-salinity surface layer up to about 50 m thick that forms as the result of the influx of fresh water from river discharge (particularly from the great rivers of Siberia, the Ob, Yenisei, and Lena; and, on the North American side, the Mackenzie), positive net precipitation over the Arctic Ocean itself (an excess of precipitation over evaporation), and the import of fairly low-salinity waters through the Bering Strait. This stable, low-density surface layer prevents the heat contained in the comparatively warm (temperatures of up to +3 °C) but denser and more saline water beneath the surface layer from affecting the ice cover. As sea ice rejects roughly two-thirds of the salt initially present in the seawater from which the ice forms, the freezing process is equivalent to distillation, producing both a low-salinity component (the ice layer itself) and a high-salinity component (the rejected brine). Both components play important geophysical roles. Over shallow-shelf seas, the rejected brine, which is dense, cold, and rich in CO₂, sinks to the bottom, ultimately feeding the deep-water and bottom-water layers of the world's oceans. Such processes are particularly effective in regions where there are large polynyas. Because this 'salt pump' removes CO₂ from the atmosphere, it has been hypothesized that it is a process contributing to the decrease in CO₂-air ratios observed in ice core samples deposited during times of maximum glacial advance (colder = more sea ice formation = more CO₂ removed from the atmosphere).

Sea ice also has important biological effects. It provides a substrate for a special category of marine life, the ice biota, consisting primarily of diatoms. These form a significant portion of the total primary production and, in turn, support specialized grazers and species at higher trophic levels, including amphipods, copepods, worms, fish, and birds. At the upper end of the food chain, seals and walruses use ice extensively as a platform on which to haul out and give birth to young. Polar bears use the Arctic ice as a platform while hunting. Also important is that in shelf seas such as the Bering and Chukchi, which are well mixed in the winter, the melting of the ice cover in the spring lowers the surface salinity, increasing the stability of the water column. The reduced mixing concentrates phytoplankton in the near-surface photic zone, thereby enhancing the overall intensity of the spring bloom.

Variability, Trends, and the Future

While the most notable aspect of variability in sea ice extent is its large seasonal cycle, extent for all months can vary markedly

from year to year on both a regional basis and for the two hemispheres as a whole. A key driver of variability in ice extent is atmospheric circulation patterns. While changing circulation patterns affect the drift of ice (the dynamic effect), they also affect conditions of temperature, cloud cover, and snowfall that influence ice growth and melt (thermodynamic effects).

As a good example of forcing by atmospheric circulation, it has been shown that September Arctic ice extent following summers with a high frequency of extratropical cyclone activity over the central Arctic Ocean tends to be greater than that following summers with a low frequency of cyclone activity (which tend to feature a persistent Beaufort Sea high). When the prevailing circulation is more cyclonic, the sea ice motion tends to be divergent, spreading the ice over a larger area and increasing ice extent. The stormy cyclonic pattern also tends to bring fairly cool conditions and even snowfall events, inhibiting summer melt. In contrast, an anticyclonic circulation pattern favors ice convergence and relatively warm conditions. These tend to reduce ice extent. Recently, considerable attention has been paid to the importance of what some have called the summer Arctic Dipole Anomaly pattern. In its positive mode, the pattern features above-average sea level pressure centered north of the Beaufort Sea (essentially, a strong Beaufort Sea high) and below-average pressure centered over the Kara Sea. Resulting southerly winds between the pressure anomaly centers at the surface favor transport of ice away from the coasts of Siberia and Alaska toward the North Pole, as well as strong melt in the East Siberian and Chukchi seas. The pattern also favors a strong sea level pressure gradient across the Fram Strait (between Greenland and Svalbard), enhancing wind-driven transport of sea ice out of the Arctic Ocean and into the North Atlantic. The negative phase has a broadly opposing pressure anomaly pattern. The positive pattern was very well developed throughout the summer of 2007 and was a key factor in that year having the second lowest September sea ice in the satellite record (based on data through 2012).

As another example, links have been established between Arctic sea ice conditions and the phase of the Arctic Oscillation (AO) during winter (the AO is also known as the Northern Annular Mode). From the late 1980s through the mid-1990s, the winter AO frequently exhibited its positive phase. This was manifested by reduced winter sea level pressure both over the Arctic Ocean and in the vicinity of the Icelandic Low. The attendant shift in the wind field and hence the ice drift led to the production of more thin FY ice along the Siberian and Alaskan coasts. It also flushed some of the Arctic's store of old, thick MY ice through the Fram Strait and into the North Atlantic. With extensive thin ice in spring, the stage was set for large areas of the pack ice to melt out in summer. This process helps to account for the extreme September sea ice minima that started to be seen in the late 1990s and early 2000s.

Although atmospheric circulation patterns are the main driver of year-to-year variability in ice extent, ocean circulation variability also plays a role. Development of the Weddell Sea polynya in Antarctica is a good example of ocean forcing. This hole in the sea ice cover can occupy an area of over 200 000 km², comparable in size to Great Britain. The polynya appears on an irregular basis (the largest was observed in

1974–76), and it is highly variable in location and shape. While a number of explanations have been offered to explain the development of the feature, it appears that an important role is played by a nearby massive seamount called the Maud Rise. There is evidence from modeling studies that modest variations in the large-scale oceanic flow past the Maud Rise seamount result in the shedding of a horizontal cyclonic eddy from its northeast flank, transmitting a divergent Ekman stress into the sea ice and leading to a crescent-shaped opening in the pack. The opening is then further enhanced by intense heat losses to the atmosphere from the open water, inducing oceanic convection.

Interannual variability in total Northern Hemisphere ice extent is superimposed upon downward linear trends in extent in all months; they are smallest for the winter months and largest for September (*see* Figure 10 in Arctic Climate). Especially interesting is that the general climatic warming widely viewed as driving the basic downward trend in Northern Hemisphere sea ice extent has changed the response of the ice cover to natural climate variability in ways that appear to have enhanced the trend. Because of the general thinning of the ice cover over the past few decades, large areas are more vulnerable to melting out in summer, making unusual circulation patterns such as the strongly developed Arctic Dipole Anomaly seen in 2007 now more effective in forcing large summer ice losses. Furthermore, the thermodynamic thinning process was dynamically enhanced by the shift in the AO from the late 1980s through the mid-1990s. The ongoing decline in Arctic sea ice extent appears to be a dominant driver of what is known as Arctic amplification – the observed enhanced warming at the Arctic surface and lower troposphere relative to the globe as a whole. With more open water in spring and summer, more solar energy is absorbed by the ocean mixed layer compared to past decades; this heat is then released upward in autumn and winter, heating the atmosphere.

By contrast, the Southern Hemisphere sea ice cover has in general expanded slightly since the late 1970s. Based on satellite data through 2010, the overall upward trend for the Southern Hemisphere is dominated by increases in some sectors that are partly countered by negative trends in others. When examined through the annual cycle, there are positive linear trends in Antarctic sea ice cover as a whole for all months, smallest in February and largest in May.

Why has Southern Hemisphere sea ice extent not experienced an overall downward trend like that seen in the Northern Hemisphere? While a number of explanations have been proposed, all of which may contribute, changes in atmospheric circulation appear to play a dominant role. A study of satellite-tracked sea ice motion for the period of 1992–2010 reveals large and statistically significant trends in Antarctic ice drift, which, in most areas, can be linked to local winds. This analysis provides strong evidence that wind-driven changes in ice advection (causing ice to drift more to the north) are a dominant driver of ice concentration trends around much of West Antarctica, whereas wind-driven thermodynamic changes (effects on temperature) dominate elsewhere (Figure 8).

What changes can we expect in our planet's sea ice cover through the twenty-first century? Most hindcast simulations from coupled global climate models that incorporate observed

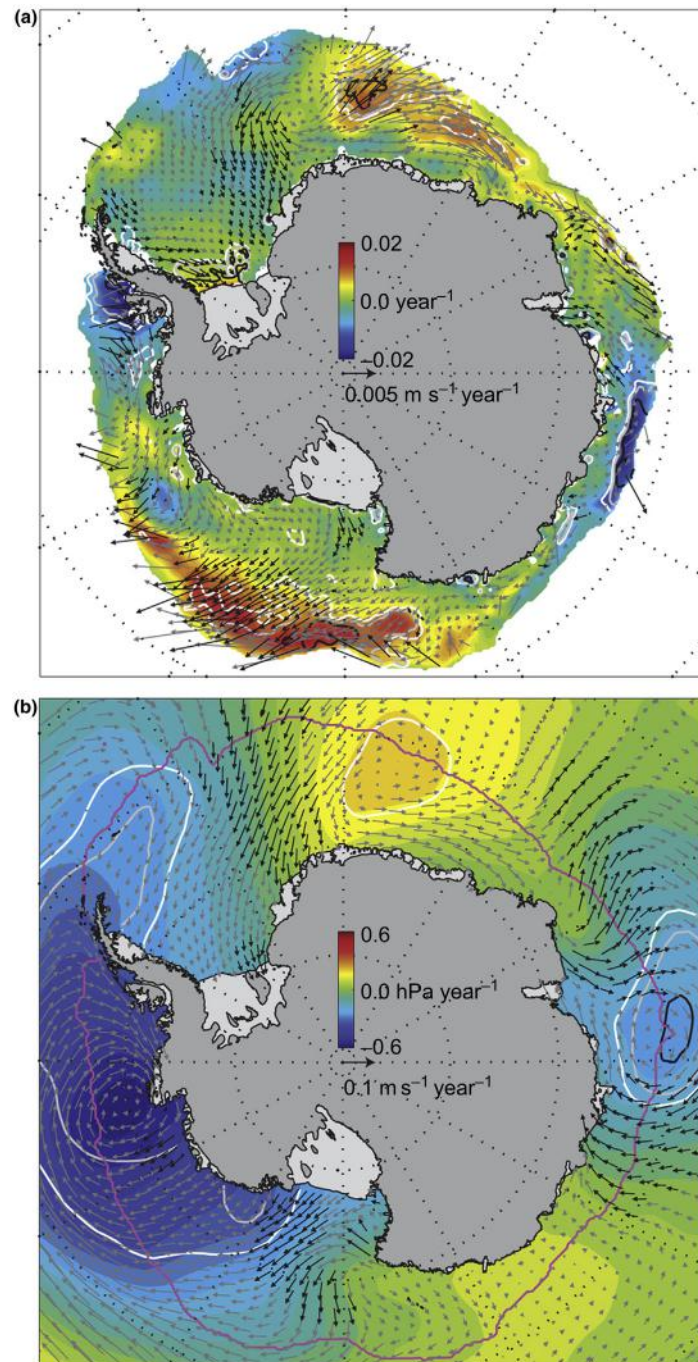


Figure 8 Linkages between wind-driven changes in ice motion and ice concentration changes in Antarctica over the period 1992–2010. (a) Ice motion trend vectors overlaid on ice concentration trends. (b) 10 m wind trend vectors overlaid on trend in sea-level pressure. White, gray, and black contours show underlay field trends significant at 90, 95, and 99%, respectively; black vectors have meridional trends significant at $>90\%$; and magenta contour in (b) shows the extent of concentration trends. Reproduced from Holland, P.R., Kwok, R., 2012. Wind-driven trends in Antarctic sea ice drift. *Nature Geoscience* <http://dx.doi.org/10.1038/NGEO1627>.

climate forcings show declining Northern Hemisphere ice extent for September over the period of observations. While this provides strong evidence for a role of greenhouse gas forcing on the observed downward trend, trends simulated by most models are smaller than observed. This is true even for the newer generation of models contributing to the

Intergovernmental Panel on Climate Change Fifth Assessment Report, which is to be published in 2014. The emerging view is that the Arctic will lose essentially all of its summer sea ice cover by the end of the twenty-first century, perhaps as early as 2030–40. The Arctic amplification of air temperature changes noted in this article will become more pronounced. Many

modeling studies show that this outsize warming may result in changes in patterns of atmospheric circulation and precipitation extending well beyond the Arctic. Another common feature of climate model simulations is a much slower decline in sea ice extent in the Southern Hemisphere.

See also: **Arctic and Antarctic:** Arctic Climate. **Climate and Climate Change:** Climate Feedbacks; Climate Variability: Nonlinear and Random Effects. **Cryosphere:** Snow (Surface). **Global Change:** Biospheric Impacts and Feedbacks; Climate Record: Surface Temperature Trends.

Further Reading

- Alexeev, V.A., Ivanov, V.V., Kwok, R., Smedsrud, L.H., 2013. North Atlantic warming and declining volume of arctic sea ice. *Cryosphere Discussion* 7, 245–265.
- Budikova, D., 2009. Role of Arctic sea ice in global atmospheric circulation: a review. *Global and Planetary Change* 68, 149–163. <http://dx.doi.org/10.1016/j.gloplacha.2009.04.001>.
- Curry, J.A., Schramm, J.L., Ebert, E.E., 1995. On the ice albedo climate feedback mechanism. *Journal of Climate* 8, 240–247.
- Holland, D.M., 2001. Explaining the Weddell Polynya – a large ocean eddy shed at Maud Rise. *Science* 292, 1697–1700. <http://dx.doi.org/10.1126/science.1059322>.
- Holland, P.R., Kwok, R., 2012. Wind-driven trends in Antarctic sea ice drift. *Nature Geoscience*. <http://dx.doi.org/10.1038/NGEO1627>.
- Kwok, R., 2012. Satellite remote sensing of sea-ice thickness and kinematics: a review. *Journal of Glaciology* 56, 1129–1140.
- Ogi, M., Wallace, J.M., 2007. Summer minimum Arctic sea ice extent and the associated summer atmospheric circulation. *Geophysical Research Letters* 34, L12705. <http://dx.doi.org/10.1029/2007GL029897>.
- Parkinson, C., Cavalieri, D., 2012. Antarctic sea ice variability and trends. *Cryosphere* 6, 871–880. <http://dx.doi.org/10.5194/tc-6-871-2012>.
- Persson, P., Ola, G., Fairall, C.W., Andreas, E.L., Guest, P.S., Perovich, D.K., 2002. Measurements near the atmospheric surface flux group tower at SHEBA: near-surface conditions and surface energy budget. *Journal of Geophysical Research* 107 (C10). <http://dx.doi.org/10.1029/2000JC000705>.
- Polyakov, I.V., Timokhov, L.A., Alexeev, V.A., Bacon, S., Dmitrenko, I.A., Fortier, L., et al., 2010. Arctic Ocean warming contributes to reduced polar ice cap. *Journal of Physical Oceanography* 40, 2743–2756. <http://dx.doi.org/10.1175/2010JP04339.1>.
- Stern, H.L., Moritz, R.E., 2002. Sea ice kinematics and surface properties from RADARSAT synthetic aperture radar during the SHEBA drift. *Journal of Geophysical Research* 107 (C10), 8028. <http://dx.doi.org/10.1029/2000JC000472>.
- Stroeve, J.C., Kattsov, V., Barrett, A., Serreze, M., Pavlova, T., Holland, M., Meier, W.N., 2012. Trends in Arctic sea ice extent from CMIP5, CMIP3 and observations. *Geophysical Research Letters* 39, L16502. <http://dx.doi.org/10.1029/2012GL052676>.
- Stroeve, J.C., Serreze, M.C., Holland, M.M., Kay, J.E., Maslanik, J., Barrett, A.P., 2012. The Arctic's rapidly shrinking sea ice cover: a research synthesis. *Climate Change* 110 (3–4), 1005–1027. <http://dx.doi.org/10.1007/s10584-011-0101-1>.
- Untersteiner, N. (Ed.), 1986. *The Geophysics of Sea Ice*. NATO ASI Series, Series B Physics, vol. 146. Plenum Press, New York.
- Wang, J., Zhang, J., Watanabe, E., Ikeda, M., Mizobata, K., Walsh, J.E., Bai, X., Wu, B., 2009. Is the dipole anomaly a major driver to record lows in Arctic summer sea ice extent? *Geophysical Research Letters* 36, L05706. <http://dx.doi.org/10.1029/2008GL036706>.
- Weeks, W.F., Hibler III, W.D., 2010. *On Sea Ice*. University of Alaska Press. p. 664.

Snow (Surface)

M Sturm, US Army Cold Regions Research & Engineering Laboratory-Alaska, Fort Wainwright, AL, USA

© 2003 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 5, pp 2061–2072, © 2003, Elsevier Ltd.

Introduction

Snow blankets more than half of the Northern Hemisphere each winter, remaining in place for periods ranging from less than a month (typically south of 40° N) to more than 8 months of the year (typically north of 60° N). In the Southern Hemisphere, the coverage is less extensive, but still substantial. If the perennial snow covers of the Greenland and Antarctic ice sheets are included, along with the seasonal snow cover that forms on lake and sea ice, then the total percentage of the Earth's surface covered by snow during some period of each year is considerable. This blanket of snow is a complex, layered material that can exhibit a high degree of spatial heterogeneity. Year-to-year variations in coverage and properties can be large and they have a direct and immediate impact on the Earth's climate. In this article, the major types of snow cover are introduced and the layered nature of the snow is discussed. The role of the snow in moderating the exchange of energy and mass with the atmosphere is also described.

Snow Cover and Its Importance

The term 'snow cover' is directly analogous to the term 'formation' when discussing layered sedimentary or metamorphic rocks. Both the sequence and character of the layers, and the lateral variation of each layer (facies changes), contribute to the overall properties of the formation. Similarly, the bulk physical and thermal properties of a snow cover, the properties that are of importance in moderating the exchange of energy and mass between the Earth and the atmosphere, are an aggregate of the properties of the individual layers. For each layer, these properties are the result of the conditions (snowfall, wind, temperature) that prevailed when the layer was deposited, and the post-depositional conditions (temperature, temperature gradients, snow overburden, liquid water percolation, solar radiation) to which the layer was subjected after deposition. Because both deposition and post-deposition conditions vary across the landscape, the layers themselves vary. In order to understand the role of snow cover in atmospheric processes, the layered nature and spatial variability of the material need to be considered.

Much of the impact of snow on climate and atmospheric processes arises because of its high albedo and low thermal conductivity. Snow cover reflects up to 85% of incoming short-wave solar radiation, significantly reducing winter temperatures and retarding melting in the spring. At the same time, snow is an excellent insulator, so it can effectively lower the rate of heat loss from the ground or an underlying ice surface, thereby maintaining higher winter soil temperatures or retarding the rate of sea and lake ice growth. The total winter energy exchange across a snow cover is a complex balance between these two competing processes. Snow cover is also important because it traps aerosols and other atmospheric particulates

like a filter, storing these until the snow melts, then releasing them abruptly. Snow can control, through thermal and physical means, the release of trace gases like CO₂ from subnival plants and soils during the winter, and it functions as a temporary storage reservoir of water, stockpiling winter precipitation then allowing it to run off in a much shorter period of time than it otherwise would have had it not fallen as snow. In some cases in higher latitudes where the snow lasts many months, as much as 80% of the annual river discharge can be from snow melt, and this discharge may occur in a period of less than 2 weeks.

Perennial and Seasonal Snow Covers

Because of their fundamentally different layered structures, it is customary to distinguish between perennial and seasonal snow covers. Seasonal snow covers are deposited in the fall and melt away completely each spring; therefore, they never become very deep. Perennial snow covers form at higher levels on glaciers and ice sheets, where the combined decrease in temperature and increase in snowfall precipitation with altitude is sufficient to allow winter snow accumulation to survive the summer melt. Snowfall of the following winter is deposited on the residual snow of the previous year, forming a sequence of annual layers of snow that can be tens to hundreds of meters thick before compaction at depth converts the snow into glacier ice.

Separate but related climate classification systems for perennial and seasonal snow covers have been suggested and are useful when thinking about both local and global variations in snow cover. For the perennial snow on glaciers and ice sheets, increasing elevation results in a decrease in melting. As a consequence, snow characteristics vary with elevation (**Figure 1**). At the lowest elevation, the melt removes all of the winter snow, and a seasonal rather than perennial snow pack forms each year. Higher, the snow pack survives the summer melt, but percolation of melt water into the snow pack and subsequent refreezing produce extensive icy features like the ice lenses and percolation columns. At the highest elevations, no melting takes place and the dry snow facies is observed. On a steep alpine glacier, the entire sequence is compressed into a distance of tens of kilometers. On ice sheets, sequence may spread over distances of hundreds of kilometers.

For seasonal snow covers, local climate rather than elevation determines the prevailing snow cover characteristics, and this local climate can be represented by three simple binary variables: winter temperature, winter precipitation, and wind. High and low values for each of these variables (**Figure 2**) define eight possible types of seasonal snow covers, most of which have a counterpart in the glacier facies system shown in **Figure 1**. For example, under warmer, wetter winter conditions, a maritime snow cover will develop. This snow cover tends to be deep (>1 m) and warm (near or at freezing temperatures), and

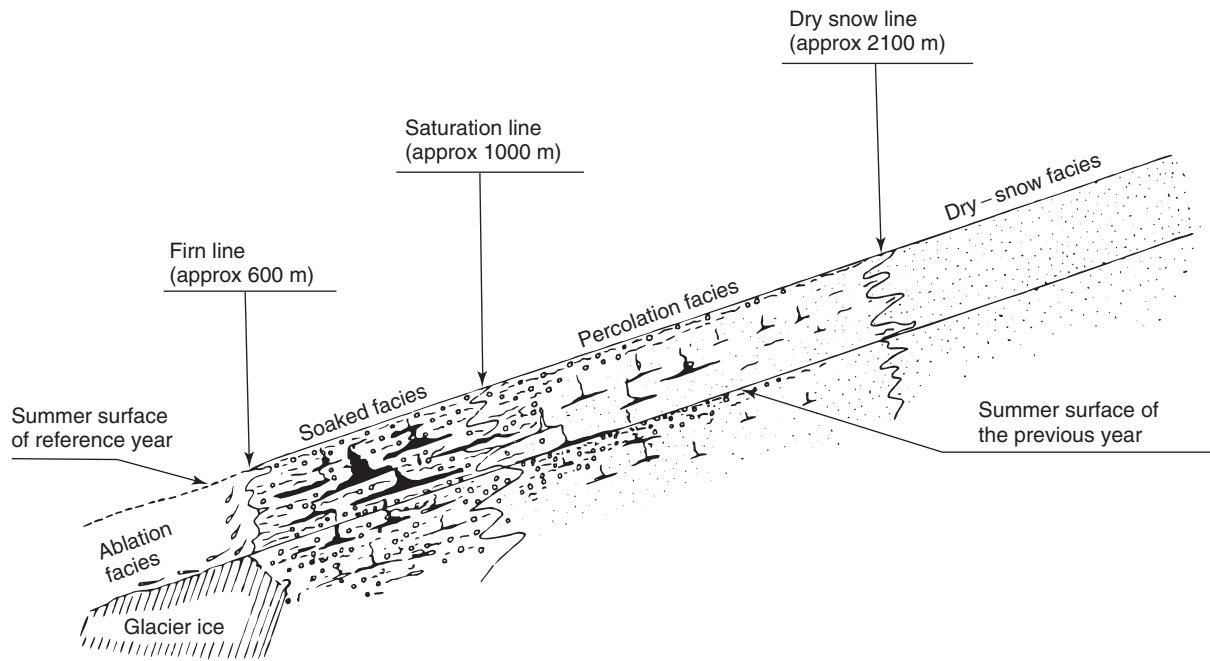


Figure 1 The glacier facies classification of Benson (1962), describing variations in the characteristics of the perennial snow cover found on glaciers and ice sheets. With increasing elevation, there is a decrease in the amount of melting and, as a consequence, a decrease in the amount of icy features in the winter snow pack. At the lowest level, all of the winter snow melts in the summer and the snow cover is essentially seasonal; at the highest level, no melting takes place and the snow has no features in it related to melting. From Benson CS (1962) Stratigraphic studies in the snow and firn of the Greenland Ice Sheet. SIPRE Research Report 70, CRREL.

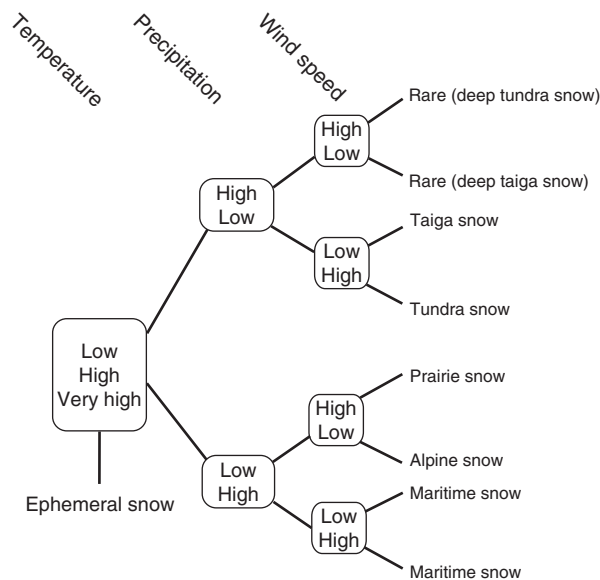


Figure 2 A dichotomous classification of seasonal snow covers based on winter temperature, precipitation, and wind. In [Figure 3](#), a typical snow stratigraphy for each class is shown. Broad similarities in snow characteristics exist between the seasonal snow classes and the glacier facies shown in [Figure 1](#). From Sturm M, Holmgren J, Liston G (1995) A seasonal snow cover classification system for local to global applications. *Journal of Climate* 8: 1261–1283.

exhibits similar icy features to those observed in the percolation facies on glaciers (compare [Figure 3](#) to [Figure 1](#)). Similarly, alpine, tundra, and taiga snow cover classes exhibit features found in the dry snow facies on glaciers. The stratigraphic diagram and key in [Figure 3](#) suggest the main snow cover characteristics associated with each climate class for seasonal snow.

Layer by Layer Development of a Snow Cover

Snow cover builds up layer by layer. The initial characteristics of each layer are determined by how much solid precipitation falls, whether the precipitation is accompanied by wind, and the prevailing temperature at the time of deposition. After deposition, each layer is subjected to mechanical and thermal metamorphic processes that alter the layer characteristics. These vary in intensity and duration depending on when the layer was deposited, its height in the snow pack and the number of overlying layers, the prevailing conditions at the snow surface, and the temperature and temperature gradients in the snow pack as a whole. At any given time, the characteristics of each layer in the snow are a product of its initial deposition and post-depositional metamorphism.

Layer Deposition and Densification

Almost 80 different types of falling snow crystals have been identified. The particular crystals that accumulate at the Earth's surface in a snow storm are determined by the temperature and humidity in the layers of air through which the crystals

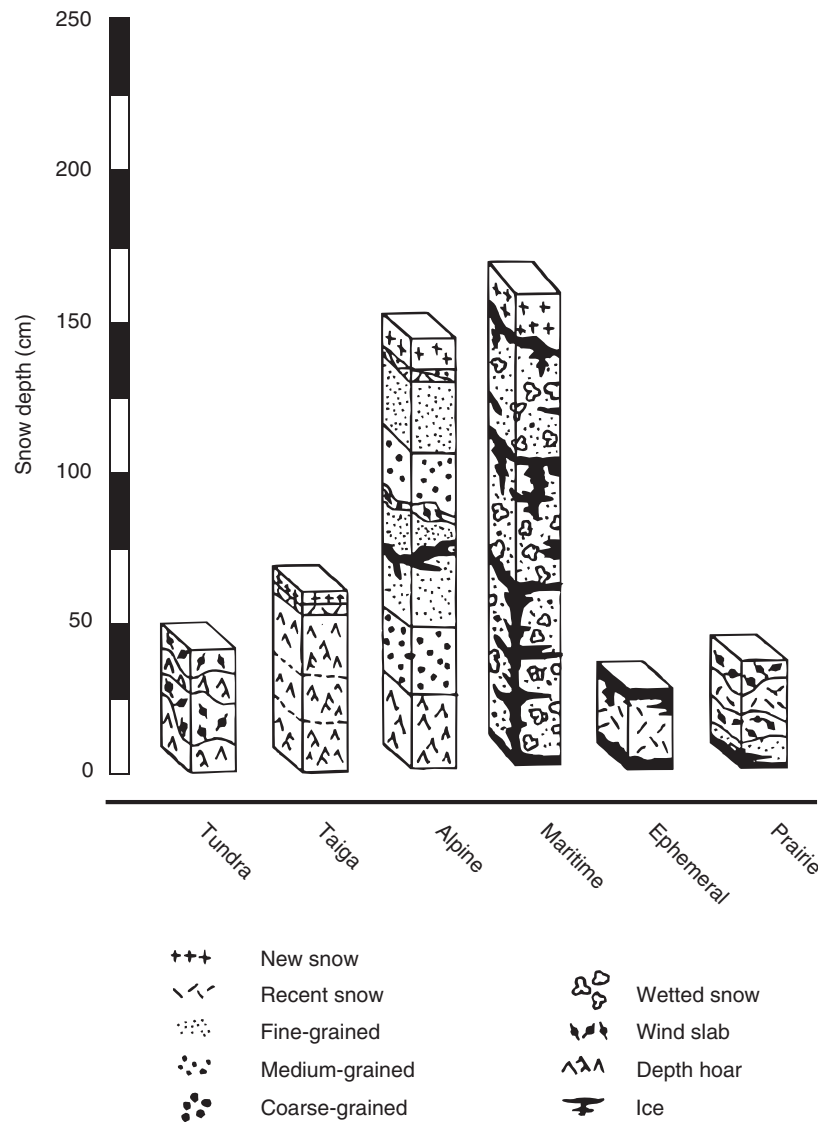


Figure 3 Typical snow stratigraphy for the six seasonal classes listed in Figure 2. From Sturm M, Holmgren J, Liston G (1995) A seasonal snow cover classification system for local to global applications. *Journal of Climate* 8: 1261–1283.

fall and grow. However, crystal form is far less important than the rate of snowfall, the wind speed, and the temperature in determining the initial characteristics of a snow layer. In general, low temperatures, low wind, and low rates of snow fall produce the lowest-density layers of new snow (Table 1).

Once deposited, new snow layers densify rapidly. Initially, much of this densification is a result of thermodynamic

instability. The sharp points and intricate branches of newly fallen snow crystals have high radii of curvature; the water vapor pressure over these highly curved surfaces is greater than elsewhere, so there is a net loss of water molecules from pointed areas to the air spaces in the snow, or to other areas on crystals that have lower degrees of curvature. The crystals rapidly break down and the resulting fragments become more rounded (Figure 4). The breakdown reduces the size of the crystals, increases the number of individual snow grains, and decreases the degree to which the crystals interlock. As a result, the entire snow layer settles.

As additional new layers of snow are added to the snow pack, the overburden load (σ) on buried layers increases. For these layers, compaction due to vertical stresses begins to dominate the snow densification process. The response of the snow to these stresses has been modeled by assuming the snow layer behaves like a viscous fluid (eqn [1]).

Table 1 The density of newly deposited snow

Deposition conditions	Density (g cm^{-3})
No wind, low rate of snowfall, cold	0.02–0.05
Low wind, low rate of snowfall	0.05–0.10
Moderate wind, high rate of snowfall	0.20–0.35
Moderate wind, low rate of snowfall	0.35–0.40
High wind	0.40–0.55

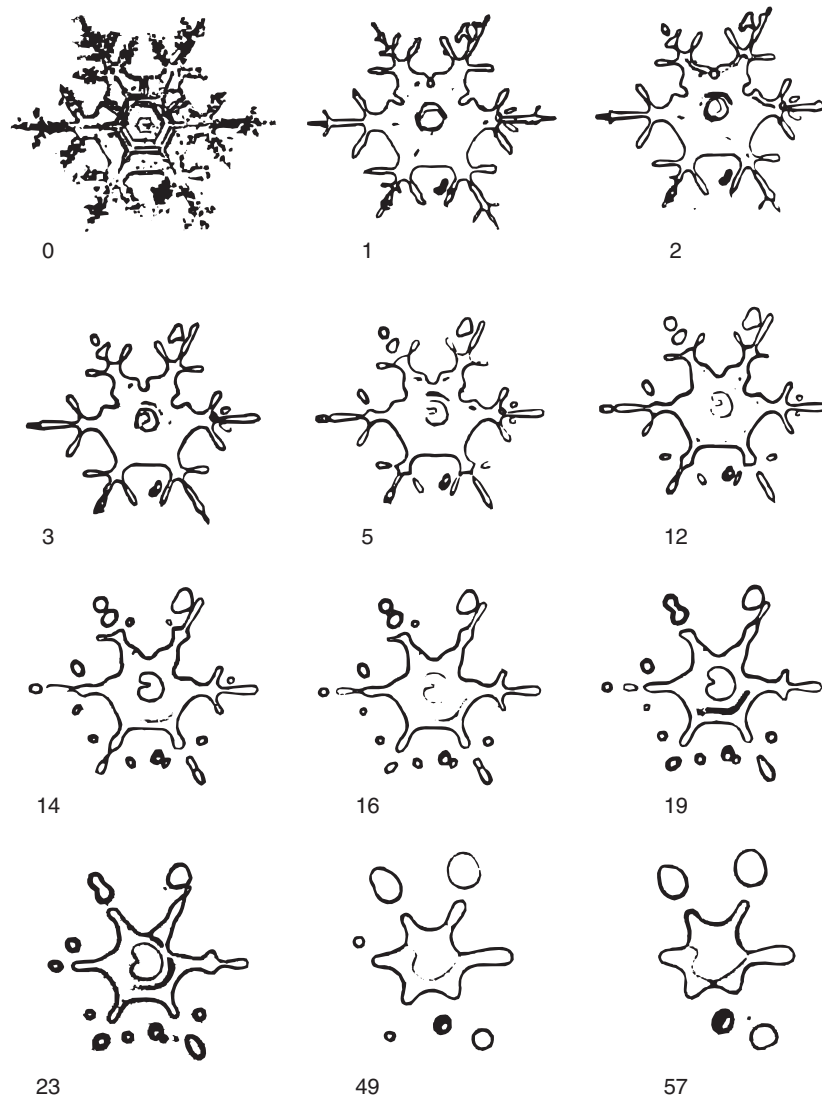


Figure 4 Changes in a snowflake held at a constant temperature of -11.5°C for a total period of 57 days (indicated by small numbers). The snowflake grew in the atmosphere under conditions of supersaturation with respect to water vapor. Once deposited, the sharp points and thin branches were thermodynamically unstable and the snowflake metamorphosed, even in the absence of a temperature gradient or overburden stress. From Bader H, Haefeli R, Bucher E, Neher J, Eckel O, Thams C (1939) *Der Schnee und seine Metamorphose* (Snow and its Metamorphism), US Army SIPRE Translation 14, 1954.

$$-\frac{1}{h} \frac{dh}{dt} = \frac{1}{\rho} \frac{d\rho}{dt} = \frac{\sigma}{\eta_c} \quad [1]$$

In eqn [1], h is the thickness of the layer (m), t is time (s), ρ is the layer density (kg m^{-3}), and η_c is the compactive viscosity. Values of η_c (Pa s) have been determined from observations of the settlement of natural snow layers, from uniaxial strain compressive tests, and from depth–density profiles on glaciers and ice sheets. The combined results show wide scatter, but individual sets of data are usually fitted to the relation in eqn [2], where k is a factor that depends on the type of snow cover (Figure 1 through 3).

$$\eta_c = \eta_0 e^{kp} \quad [2]$$

The effective viscosity term incorporates a number of physical mechanisms including gravity-driven movement of snow grain

centers of mass toward each other, vapor and volume diffusion, and sintering. Not surprisingly, viscosity factor values vary widely depending on the temperature, liquid water content, and grain characteristics of the snow – i.e., the snow cover class (Table 2). Colder, drier, finer-grained layers of snow tend to be more viscous than warmer, wetter, layers with larger grains, and therefore compact more slowly.

In the absence of melting or the introduction of liquid water, snow layers will continue to densify until they reach a limiting density of about 0.6 g cm^{-3} . By this time, the snow grains will have metamorphosed until they have become highly rounded, a shape that minimizes their surface free energy. The rounded grains will be in close contact with each other, and the grain arrangement will approximate that of hexagonal close-packing of ice spheres. Further densification

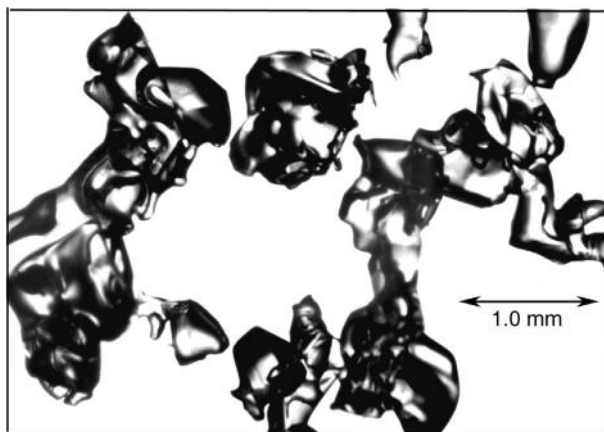
Table 2 Compactive viscosity factors for three classes of snow cover

Snow cover type	k-value ($\text{m}^3 \text{kg}^{-1}$)
Maritime	18–22
Alpine/taiga	35–60
Tundra	>70

will require actual deformation of the individual grains of snow, or the influx and refreezing of melt water in pore spaces. The overburden stresses required to achieve this further deformation are only realized in the deep perennial snow packs found on glaciers and ice sheets.

Snow layers deposited during windy conditions (wind slabs) have much higher initial densities than other new snow layers. The wind tumbles snow crystals as it transports them, breaking the more fragile crystal junctions and pulverizing the crystals in general. The resulting grains are actually crystal fragments, often less than 0.1 mm in length, and these shard-like grains (Figure 5), when they come to rest, pack well and sinter together into a cohesive slablike layer. Initial densities for wind-transported layers of new snow range from 0.35 to 0.6. The upper limit occurs for the same physical reasons as discussed before. Due to their high initial densities and cohesiveness, wind slabs are highly resistant to compaction and often remaining at a fixed density after deposition.

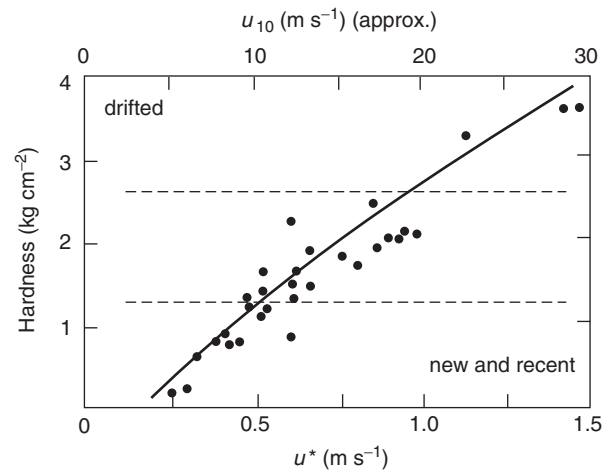
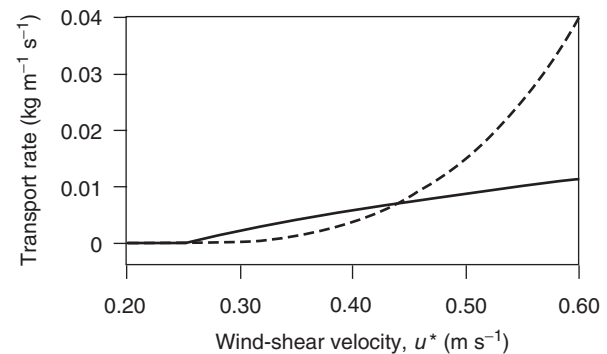
There has been much discussion and experimentation to determine the wind speed necessary to transport snow. The transport takes place through three mechanisms: creep, saltation, and suspension. Creep consists of the rolling movement of grains along the snow surface under the action of the wind. Saltation is the movement of grains along the surface by jumping and ricocheting after impact by other grains. Suspension is the movement of grains in the wind stream at some level above the snow surface. The threshold shear velocity, u^* , at which transport occurs is usually estimated by assuming a logarithmic-shaped wind profile and projecting the 10-m high wind speed (u_{10}) down to the snow surface (u^*). In general the value of u_{10} is between 18 and 30 times the value of u^* . Experimental studies indicate that when u_{10} is greater

**Figure 5** Wind-pulverized snow grains from Arctic Alaska, showing irregular shapes and thick bonds due to rapid sintering after deposition.

than 6 m s^{-1} transport will occur if the snow has fallen recently. If the snow is new and falling while there is wind, transport will occur with wind of 5 or even 4 m s^{-1} . If the snow is aged, was previously transported by the wind, or has undergone some melt-freeze processes, speeds in excess of 30 m s^{-1} may be needed before the snow will start to be transported (Figure 6).

In similar fashion, the flux of snow transported by the wind is a strong function of the wind speed, with increasing speeds producing a marked increase in the total amount transported (Figure 7). For values of u^* between 0.2 and 0.44 m s^{-1} , saltation dominates the transport, but for u^* values in excess of 0.44 m s^{-1} , suspension exceeds saltation in transporting snow.

One other consequence of wind transport of snow is the development of a wide range of drift deposit and erosion

**Figure 6** The critical wind shear velocity (u^*) as a function of snow hardness, which is a good measure of the type of snow. Increasing hardness, common for wind slabs and layers of snow that have undergone melt-freeze, requires considerably higher winds to mobilize these types of snow. u_{10} is the wind speed measured at a standard height of 10 m. From Kind RJ (1981) Snow drifting. In: Gray DM, Male DH (eds). Handbook of Snow, pp. 338–359. Toronto: Pergamon.**Figure 7** Snow transport rates for saltation (solid curve) and suspension (broken curve) as a function of wind shear velocity (u^*). The wind speed at 10 m height is approximately 18–26 times u^* . At $u^* = 0.44$ (10 m height wind speeds of $8\text{--}11 \text{ m s}^{-1}$), suspension begins to transport the majority of the wind-borne flux of snow. From Liston GE, Sturm M (1998) A snow-transport model for complex terrain. Journal of Glaciology 44: 498–516.

features at the snow surface. These features include ripple marks, dunes, barchans, and sastrugi. Surprisingly, little is known about the relationship between these features and the wind speed, the temperature, and the snow conditions necessary for their formation.

The final, and most efficient, method for densifying a layer of snow is through the infiltration of melt or rain water into the snow cover, followed by subsequent refreezing. Water can infiltrate, surround grains as thin films or lie in veins along grain junctions, and refreeze to produce large multiparticle grains. Water can also percolate downward in pipelike structures called percolation columns, or spread out along stratigraphic boundaries (owing to variations in the hydraulic conductivity of the snow). When this water refreezes, ice lenses and layers are created. Frequently, a single infiltration event will produce ice layers at multiple levels in the snow pack. Densities in excess of 0.6 g cm^{-3} , sometimes even as high as 0.9 g cm^{-3} , can result. This mechanism is commonly observed in ephemeral and maritime seasonal snow covers (Figures 2 and 3), and in the percolation facies for perennial snow (Figure 1).

Snow Metamorphism

In addition to compaction and densification, several other metamorphic processes can affect layers of snow. These processes result chiefly in changes in snow grain characteristics and bonding, which in turn affect the thermal conductivity, air permeability, and albedo of the snow. The processes are typically divided into 'wet' and 'dry' categories because different snow grain characteristics are produced depending on whether liquid water is present. Further metamorphic subdivisions are shown in Table 3.

For wet snow metamorphism, the degree to which grains and a snow layer are changed is mainly a function of how much water is present. For low liquid contents ($<5\%$ by weight), the water in the snow exists as thin films and isolated pockets or veins around grains; continuous ice grain and air space pathways still exist through the snow layer. This is called the pendular regime. Under this regime, snow grains will rapidly round, and clusters of grains, looking much like bunches of grapes, will form as a result of the minimization of surface free energy. The clusters themselves are quite strong because the bonds between the spherical grains are still intact and substantial. The wet snow pack will have considerable bearing strength. Spring skiing, which can be excellent, takes advantage of these ball-bearing like grain clusters and the general strength and cohesiveness of this type of wet snow metamorphism. If the temperature of the snow drops and the grain clusters freeze, they will take on the slightly more amorphous shapes of melt-grain clusters (Figure 8), while at the same time the strength of the layer will increase dramatically as all the interstitial water freezes. For higher liquid water contents, snow grains and air spaces become surrounded and isolated by the liquid water

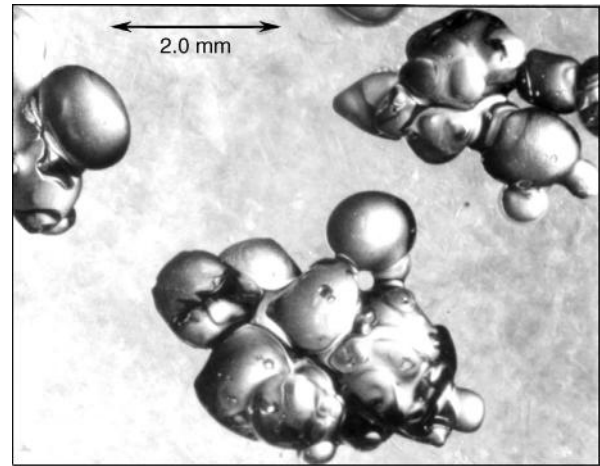


Figure 8 Melt-grain clusters showing the well-rounded grains and the high degree of contact between grains.

present in the layer. This water begins to drain downward under the influence of gravity and is called the funicular regime. Once again, when surrounded by water, the snow grains will round, but now boundaries between grains will not be thermodynamically stable and will melt rapidly, creating a slush. The slush has little or no bearing strength, and can even flow like a fluid under certain conditions. The grains themselves, if surrounded by water at 0°C for long enough (24–36 hours), will metamorphose into oblate spheroids (Figure 9).

In the absence of liquid water, snow will metamorphose in one of two ways depending on the temperature gradient imposed on the snow. Water vapor density over ice is a strong positive function of temperature, so temperature gradients in the snow give rise to water vapor density gradients in the air spaces in the snow and a diffusive flow of vapor from warmer to colder grain surfaces. For convenience the temperature gradient is often defined as the difference between the basal and surface temperatures of the snow cover, divided by the thickness of the snow (Figure 10), but in reality the actual temperature gradient varies continuously with both time and height in the snow. For example, rapid fluctuations in air temperature can produce very large temperature gradients near the snow surface, at least for short periods of time. Experimental work has shown that when the temperature gradient exceeds a magnitude of approximately $0.25^{\circ}\text{C cm}^{-1}$, kinetic crystal growth will occur. If the gradient is lower, equilibrium growth takes place. Not surprisingly, temperature gradients in thick perennial snow covers tend to be lower than those in the thinner seasonal snow covers, particularly thin taiga, tundra, and alpine seasonal classes that can be subjected to very low air temperatures in the winter. As a result, kinetic growth is common in seasonal snow covers but occurs infrequently (often only in autumn) in perennial snow covers.

Table 3 Metamorphic processes that affect the snow cover

<i>Wet snow metamorphism</i>	<i>Dry snow metamorphism</i>	<i>Dry snow metamorphism – older terms</i>
Melt-grain clusters and melt-freeze particles	Equilibrium or rounded growth	Equi-temperature metamorphism (ET)
Slush	Kinetic or faceted growth	Temperature-gradient metamorphism (TG)

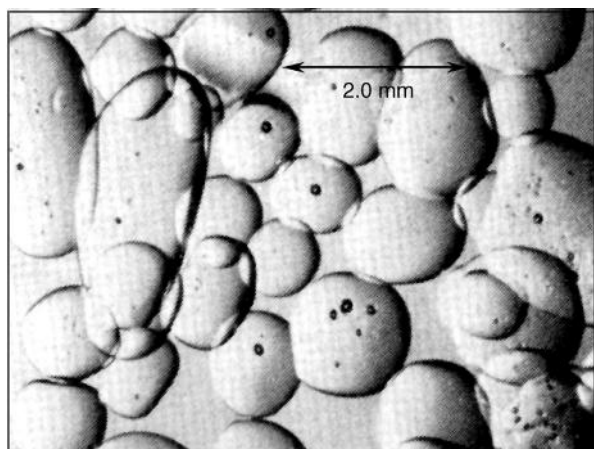


Figure 9 Snow slush, showing the oblate spheroid shape of the grains and the complete lack of bonding. From Colbeck SC (1986) Statistics of coarsening in water-saturated snow. *Acta Metallurgica* 34, 347–352.

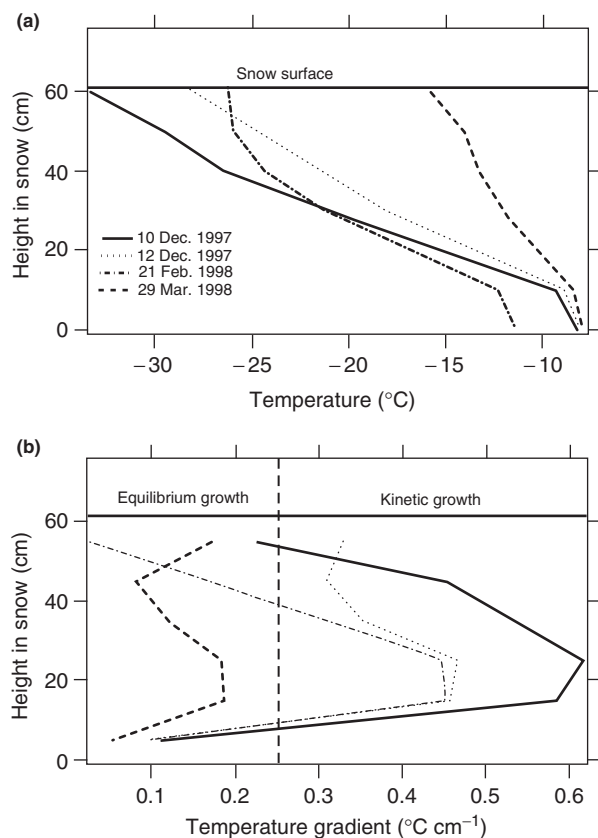


Figure 10 (a) Temperature profiles and (b) computed vertical temperature gradients from the snow cover on the ice of the Beaufort Sea north of Alaska. The temperature profiles are not linear, and as a consequence, the temperature gradients vary in a complex way with height in the snow. Note that at some heights and times the gradient is below the critical magnitude of $0.25^{\circ}\text{C cm}^{-1}$ and kinetic growth will not occur.

Equilibrium crystal growth, also widely known as ‘equi-temperature metamorphism’ (ET-metamorphism) occurs when temperature gradients in the snow pack are less than $0.25^{\circ}\text{C cm}^{-1}$. These low temperature gradients produce weak

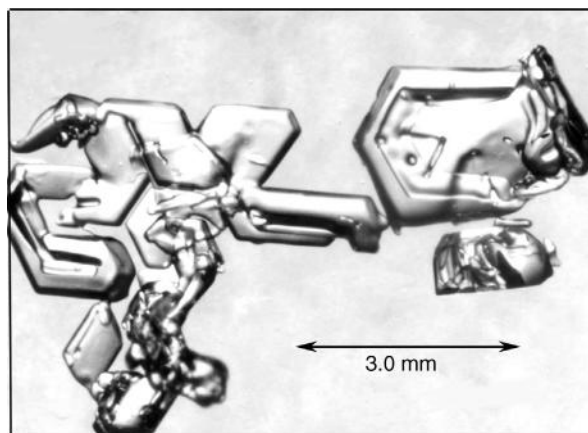


Figure 11 The initial stages of kinetic growth metamorphism. The grains are starting to exhibit distinct faceting.



Figure 12 At-depth hoar cup, shown in typical growth position. The hexagonal pyramidal cup opens downward because the flow of water vapor is upward. Heavy striae can be seen on all crystal facets. This is the late stage of kinetic growth metamorphism.

water vapor density gradients in the snow and low rates of vapor diffusion. The rates are so low that the supply of vapor to a growing crystal, rather than crystal growth dynamics, controls the growth. Rounded, well-bonded grains result.

Kinetic growth, also widely known as ‘temperature-gradient metamorphism’ (TG-metamorphism), produces ornate, faceted

crystals commonly referred to as 'depth hoar.' In this case, temperature gradients imposed on the snow are of a large enough magnitude to produce a flux of water vapor that exceeds the rate at which the crystal can grow. Crystal growth dynamics, rather than vapor supply, control both the growth rate and the crystal form, producing crystals with distinct sharp-edged facets, well-defined interfacial angles, and surface striae (Figure 11 and Figure 12). Unlike the case for equilibrium growth, intergrain bonds are weakened and reduced in number during kinetic growth, producing layers that tend to be brittle and weak. This has two important ramifications: the brittle layers can shear easily and often create failure planes that are responsible for the release of avalanches. Second, the poor bonding creates layers that have low thermal conductivity. In absence of air movement in the snow, these layers provide excellent insulation that contributes to the retention of heat in the ground or ice underlying the snow cover.

Energy and Mass Exchange across a Snow Cover

It is beyond the scope of this article to address in full the mass and energy exchange over a snow cover, but a few points particular to snow are discussed. The reader should also see articles on surface energy balance, albedo, turbulence,

boundary layer meteorology, surface roughness, and solar radiation for more details.

Heat transfer across a snow cover occurs mainly by conduction through the ice network of grains, by conduction across the air-filled pore spaces in the snow, and by diffusion of vapor across the pore spaces. The thermal conductivity of ice is more than 100 times higher than that of air, so the conduction of heat across air spaces is thought to contribute relatively little to the total. The heat transported by vapor diffusion, in contrast, is thought to contribute as much as 40%, particularly at temperatures near freezing when the vapor flux is high. This diffusive vapor transport is envisioned as occurring in a 'hand-to-hand' manner across pore spaces, with vapor diffusion from the warm side of snow grains balanced by vapor condensation on the colder side.

Because the contributions of these three individual mechanisms are difficult if not impossible to separate, in practice they are always lumped together by reporting an 'effective' thermal conductivity for the snow. Both solid body conduction through the ice network and vapor diffusion are driven by the temperature gradients in the snow, suggesting that a simple heat flow equation can be used to model the flux of heat across the snow (eqn [3]).

$$q = k_{\text{eff}} \frac{dT}{dz} \quad [3]$$

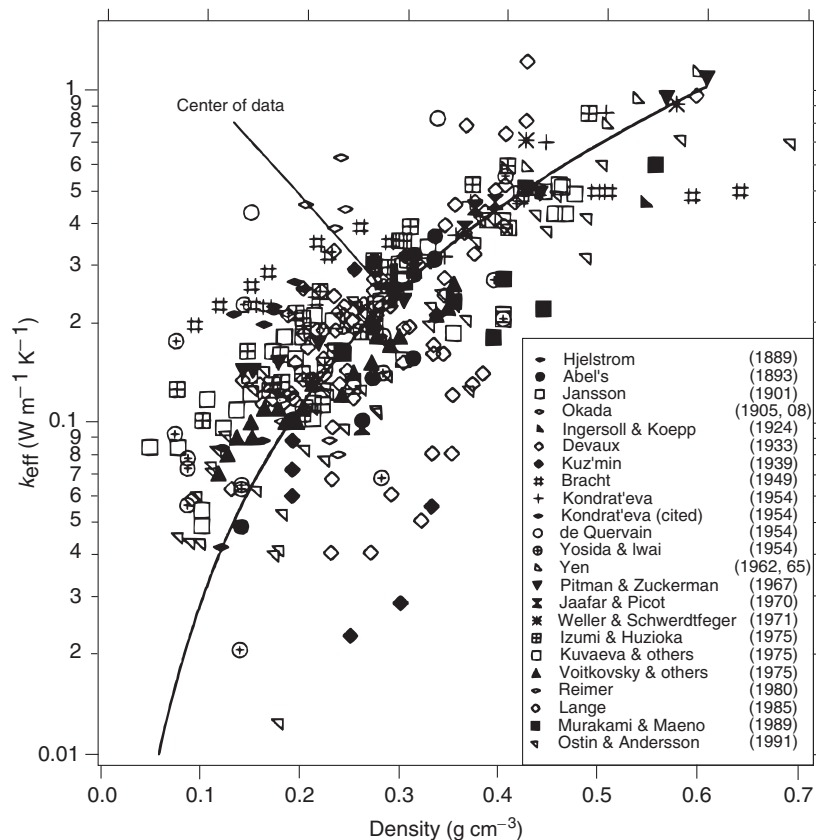


Figure 13 A compilation of most published values of the thermal conductivity of snow. There is nearly an order of magnitude scatter at any given density, and this scatter is real. It arises from differences in snow cover characteristics. From Sturm M, Holmgren J, König M, Morris K (1997) The thermal conductivity of seasonal snow. *Journal of Glaciology* 43: 26–41.

Here q is the vertical heat flow through the snow cover, dT/dz is the temperature gradient across the snow, and k_{eff} is the effective thermal conductivity of the snow. However, the driving temperature gradient in the ice network may be quite different from the gradient across pore spaces that drives vapor diffusion, in which case eqn [3] may be an oversimplification. Be that as it may, it is customary to describe the heat transfer using eqn [3] and assigning an appropriate value for k_{eff} .

Figure 13 shows compilation of most measured values of k_{eff} as a function of density. As the density of the snow increases, so in general does the value of k_{eff} . In many climate models, regression equations relating k_{eff} to density (often using the viscous snow compaction (eqns [1] and [2]) to determine the snow density) are used to set the thermal conductivity of the snow. However, as the figure shows, the scatter in k_{eff} at any given density is large and real. It is the result of differences in the bonding of the snow, and perhaps also due to variations in snow temperature. For a given density, higher temperatures and better bonding between grains lead to higher values of thermal conductivity. Given the scatter, care should be exercised when choosing a value of k_{eff} for modeling. The values should be consistent with the type of snow cover (Figure 1 through 3) as well as a k_{eff} -density relationship. For improved accuracy, a value of k_{eff} for each layer of snow should be determined; then the bulk value for the entire snow cover should be computed using a series-type solution.

Convective heat transfer is also known to operate in snow and complicates the energy exchange across a snow cover. Two types of convection have been reported: buoyancy-driven convection, and convection forced by the wind (wind-pumping). The former has been documented only in a highly permeable snow covers like taiga snow. This snow cover often wholly comprises layers of large, poorly bonded kinetic growth crystals called depth hoar. The layers have extremely high values of air permeability and, owing to low winter air temperatures, are subjected to temperature gradients of high magnitude, both conditions favorable for buoyancy-driven convection. Convective air flow velocities of several millimeters per second have been computed based on observations of temperature fields in the snow, and these air flow speeds are sufficient to increase the heat transfer rate by a factor of 3. The prevalence of buoyancy convection in other types of snow covers may be low, but this has not been shown experimentally.

Forced convection also probably occurs in some snow covers. Theory indicates that pressure differences arising when wind blows across surface irregularities like dunes and sastrugi are most likely to produce a flow of air that can move both heat and mass (in contrast to turbulence or other aspects of the wind over snow). Flow rates are probably on the order of a few millimeters per second and are likely to be confined to near-surface layers of snow. Observations of the mixing depth of aerosols and particulates in snow layers indicate that wind pumping is definitely effective in moving mass, but the magnitude of the effect of wind-pumping on heat transfer has yet to be demonstrated. In addition, it appears that near-surface and surface wind and melt crusts in the snow can effectively eliminate any wind-pumping by reducing the air permeability of the snow creating barriers in the form of impermeable wind on melt crusts that can effectively shut off all air movement.

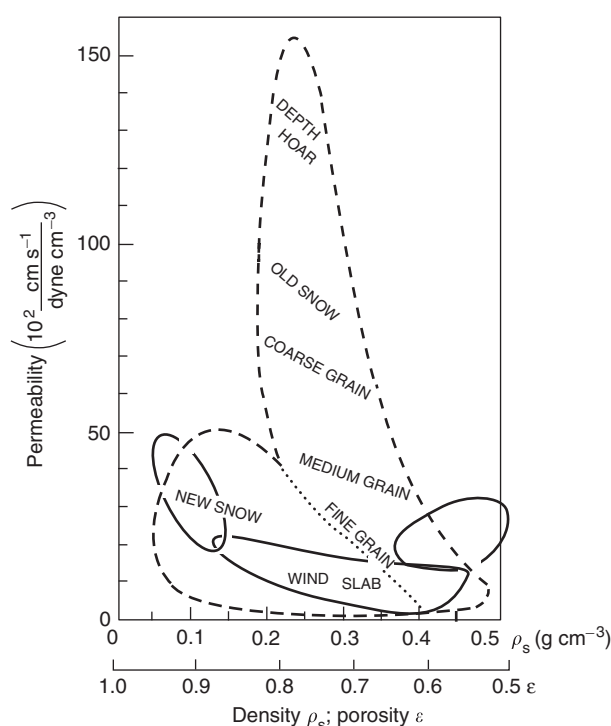


Figure 14 The air permeability of snow. Again, there is a greater variation by snow type than by density. From Shimizu H (1970) Air permeability of deposited snow. Low Temperature Science, Series A, 1–32.

As neither wind-pumping nor buoyancy-driven convection are state properties of the snow, they pose difficulties when one is trying to model heat transfer in snow. Both processes depend on external conditions for their onset and strength, and they can transport anything from zero to several times the conductive heat flux, depending on the snow characteristics, the temperature structure in the snow, and the wind speed and direction.

Water, water vapor, CO_2 , methane, and aerosols and particulates are all transferred across a snow cover and the transfer process for each is complicated. In general, mass transfer is controlled by the air permeability of the snow, the surface topography of the snow cover (for wind-pumping), and the supply rate of particles, gases, or chemicals. As discussed previously, both diffusive and convective transport of air are possible, and the chemicals and gases move with the air. The air permeability of naturally occurring snow (Figure 14) ranges over two orders of magnitude. It is a major control on deposition and transfer rates, which vary widely with chemical species and environmental conditions. For aerosols, when the residence time of the air in the snow is greater than 15 seconds, the filter efficiency of the snow can be almost 100%.

See also: Boundary Layer (Atmospheric) and Air Pollution: Overview. **Climate and Climate Change:** Energy Balance Climate Models. **Land-Atmosphere Interactions:** Canopy

Processes; Overview; Trace Gas Exchange. **Solar System/Sun, Atmospheres, Evolution of Atmospheres:** Solar Terrestrial Interactions: Climate Impact.

Further Reading

- Colbeck, S.C., 1986. Classification of seasonal snow cover crystals. *Water Resources Research* 22 (9), 59S–70S.
- Gray, D.M., Male, D.H., 1981. *Handbook of Snow*. Pergamon Press, Toronto.
- La Chapelle, E.R., 1969. *Field Guide to Snow Crystals*. University of Washington Press, Seattle.
- Magono, C., Lee, C.W., 1966. Meteorological classification of natural snow crystals. *Journal of the Faculty of Science, Hokkaido University* 2 (4), 321–335.
- Seligman, G., 1936. *Snow Structure and Ski Fields*. (Reprinted by the International Glaciological Society, Cambridge, 1980).
- Shimizu, H., 1970. Air permeability of deposited snow. *Low Temperature Science Series A* (22), 1–32.
- Sommerfeld, R.A., 1970. The classification of snow metamorphism. *Journal of Glaciology* 9 (55), 3–17.
- Sturm, M., Holmgren, J., et al., 1997. The thermal conductivity of seasonal snow. *Journal of Glaciology* 43 (143), 26–41.
- Waddington, E.D., Harder, S.L., 1996. The effects of snow ventilation on chemical concentrations. In: Wolff, E.W., Bales, R.C. (Eds.), *NATO ASI Series*, vol. I–43. Springer-Verlag, Berlin, pp. 403–451.
- Warren, S.G., 1982. Optical properties of snow. *Reviews of Geophysics and Space Physics* 20 (1), 67–89.

DATA ASSIMILATION AND PREDICTABILITY

Contents

Data Assimilation

Ensemble-Based Data Assimilation

Ensemble Prediction

Predictability and Chaos

Data Assimilation

AC Lorenc, The Met Office, Bracknell, Berkshire, UK

© 2003 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 6, pp 2542–2546, © 2003, Elsevier Ltd.

Introduction

Modern numerical weather prediction (NWP) models represent the current state of the atmosphere using many millions of numbers. Despite the growing numbers of satellite and *in situ* observations, the state is often poorly determined by the observations alone. Accordingly, a data assimilation process in which the NWP model state is fitted to past and present observations has replaced early computer methods of producing an objective analysis by interpolating the observations. Under certain simplifying assumptions, the optimal assimilation is given by a sequential estimation method known as the Kalman filter, but this cannot be implemented in practice because it requires the manipulation of enormous covariance matrices. A variational approach of fitting the model state to observations is more practical, and has the advantage of being able to use observations nonlinearly related to the model state. Alternative methods using ensembles are being explored, while simpler methods are common for mesoscale systems.

Using Observations in NWP

Modern weather prediction is based on numerical forecast models that represent the atmosphere's variables (wind, temperature, pressure, moisture, cloud, etc.) on a set of levels from the surface to the stratosphere, and a horizontal grid (or equivalent spectral representation) capable of resolving weather systems. For instance, in 2002 the Met Office's global NWP model had 38 levels and a horizontal grid spacing of 60 km; it represents the atmosphere using 10^7 numbers. As available computer power doubles every few years, the size and complexity of models used increases correspondingly.

An illustration of the number of observations available to determine the initial state for a global forecast model is shown in [Table 1](#). In some places some variables are well observed, indeed some observation types are thinned before use, but in many places many variables are poorly determined by the observations. Thanks to advances in remote sensing technology, and in automation, the quality and coverage of observations is continuing to improve, but the requirements for NWP models are increasing as well – some areas and variables will remain relatively poorly observed.

In the days of manual forecasting, before computers, the process of determining the initial conditions from the observations was called weather map analysis. Early computer forecasting systems attempted to replicate this by interpolating between the observations to a regular grid, using a so-called objective analysis program. Because of the sparsity of observations, this approach does not meet the requirements of modern NWP for an accurate and detailed representation of all the atmosphere's variables; extra information must be used. NWP models are based on the physical equations describing the atmosphere's evolution, and hence encapsulate much of our prior knowledge about the atmosphere. Thus we use an NWP model as part of an efficient system to determine the atmospheric state from incomplete observations. This process is called data assimilation. It naturally gives a picture of the atmospheric state in the form needed to start an NWP forecast.

Data assimilation for NWP proceeds sequentially, in cycles. A model state at a particular time summarizes, in organized form, the information gleaned from all earlier observations. The NWP model is integrated forward to the times of the observations in the latest batch. This provides a background state, with which the new information is combined, to create a model state that is a new best estimate.

Table 1 Observations Used in Global Data Assimilation in October 2000

Observation group	Subgroup	Items used	Daily	Percentage used
Ground-based vertical profiles	TEMP	T , V , RH processed to model layer average	1200	97
	PILOT	As TEMP but V only	900	99
	PROFILER	As TEMP but V only (used from February 2001)	300	0 (65)
Satellite-based vertical profiles	TOVS	Radiances directly assimilated with channel selection dependent on surface, instrument, and cloudiness	54 000	11
	ATOVS		700 000	4
Aircraft (manual and automated)	AIREPS	T , V as reported with duplicate checking and blacklist	14 000	21
	ACARS AMDAR ASDAR		67 000	60
Satellite atmospheric motion vectors	GOES 8, 10	High resolution 'BUFR' IR winds	55 000	24
	Meteosat 5, 7	IR, VIS, and WV winds	9200	98
	GMS 5	IR, VIS, and WV winds	5200	93
Satellite-based surface	ERS-2	Wind vector retrievals (ambiguous winds from February 2001)	170 000	0
	SSM/I-13	In-house 1D-VAR wind speed retrieval (no moisture yet)	1 450 000	1
Ground-based surface	Land Synop	Pressure only (processed to model surface)	27 000	80
	Ship Synop	Pressure and wind	6000	90, 95
	Buoy	Pressure	9000	75

T , temperature; V , vector wind; IR, VIS, and WV winds, winds tracked from sequences of infrared, visible, and water vapor imagery.

Typical coverage maps are available at http://www.metoffice.com/research/nwp/observations/data_coverage/index.html. Only a small selection of high-density satellite observations are currently used; about 1.6×10^5 data are presented to the variational analysis.

Following the old manual terminology, this state is often called the analysis. It is used to start the NWP model forecast in the next cycle. If necessary, an initialization method is used to make sure each NWP forecast starts smoothly. It is important to realize that the NWP model is key to the process; the information in the model state from earlier observations is much more important than that in any single batch.

Statistical Estimation Theory

Statistically optimal methods of combining the new observations with the model forecast depend on knowledge of the error probability distribution function of each. The optimal method is simplest when the error distributions are Gaussian, with mean zero. In reality, most observations can suffer from occasional gross errors, due to instrument or communications failures, and some types of observation can suffer from systematic biases. Usually, some quality control and bias correction are done prior to the main assimilation step, in which case the remaining observations may be assumed to have the desired Gaussian distribution. Less theoretically, one can see that a quality control procedure is essential to avoid using the corrupted observations, and that proper calibration of observations will improve their utility. A Gaussian distribution can be completely described by a covariance matrix. Most sources of observational errors are independent, making covariances between them zero; that is, for most observations the error covariance matrix is diagonal.

The errors in the background state are harder to characterize, since during the forecast process errors spread and grow dynamically, leading to significant covariances in space and

time, and between variables. One practical approach, used in so-called statistical interpolation or optimal interpolation algorithms, is to model the background error covariances using structure functions, assuming them stationary, homogeneous, and isotropic. Denoting the model background state by vector $\mathbf{x}^b$, the observations by vector $\mathbf{y}^o$, the linear operator for calculating estimates of the observed values from a model state by $\mathbf{H}$, the observational error matrix by $\mathbf{R}$, and the background error covariance matrix by $\mathbf{B}$, the optimal analysis ($\mathbf{x}^a$) and its error covariance ($\mathbf{A}$) are given by eqns [1] and [2].

$$\mathbf{x}^a = \mathbf{x}^b + \mathbf{B}\mathbf{H}^T(\mathbf{H}\mathbf{B}\mathbf{H}^T + \mathbf{R})^{-1}(\mathbf{y}^o - \mathbf{H}\mathbf{x}^b) \quad [1]$$

$$\mathbf{A} = \mathbf{B} - \mathbf{B}\mathbf{H}^T(\mathbf{H}\mathbf{B}\mathbf{H}^T + \mathbf{R})^{-1}\mathbf{H}\mathbf{B} \quad [2]$$

In principle, under the assumption that errors are small enough to be propagated by the linearization of the NWP model, it would be possible to calculate the background error covariances explicitly from the analysis error covariances of the previous cycle. This is the approach followed in the Kalman filter assimilation method, which for known Gaussian errors and a linear system is the theoretically optimal sequential assimilation method. However, because of the size of NWP models, it is not feasible in practice even to store the full matrices $\mathbf{A}$ and $\mathbf{B}$, let alone to propagate them in time.

Although the basic methods for combining imperfect observations have been understood since Gauss in the early nineteenth century, we are not able to apply even these in a truly optimal way since they require error covariances that we cannot know accurately, and that we could not manipulate on available computers if we did know them. Added complications are the gross errors in observations, which make their error distributions non-Gaussian, and the nonlinear, chaotic

nature of the forecasting equations, which makes background errors non-Gaussian. Theory for handling non-Gaussian probability distribution functions exists, but its application to this size of problem is even less feasible. Practical data assimilation methods for NWP have to use approximate methods, applying the available computer power to those parts where our knowledge justifies it.

Variational Assimilation

Rather than interpolating from observations to a grid, assimilation is best thought of as an inverse method, fitting the model state to scattered data. In variational assimilation, observation operator algorithms are applied to a model state to predict the observations. Any observations can be used, as long as the observation operator to predict their values is a well-behaved, linearizable function of the model values. For instance, it is straightforward to use satellite observed radiances for frequency channels emitted or absorbed by the atmosphere: the observation operator is a radiative transfer calculation of the upward radiances. We represent the observation operator by function H . These predictions are compared with the actual values, and a penalty measuring the misfit is calculated. This is added to a measure of the misfit from the background, to give a total penalty function (eqn [3]).

$$J = \frac{1}{2}(H(\mathbf{x}) - \mathbf{y}^o)^T \mathbf{R}^{-1}(H(\mathbf{x}) - \mathbf{y}^o) + \frac{1}{2}(\mathbf{x} - \mathbf{x}^b)^T \mathbf{B}^{-1}(\mathbf{x} - \mathbf{x}^b) \quad [3]$$

The model state ($\mathbf{x}$) is varied until the minimum of the penalty function is found. Since the terms in the penalty function are weighted using the appropriate error covariance estimates, the variational approach is equivalent to the statistical estimation discussed above (for linear H). However, it can be extended to use nonlinear observation operators, using satellite radiance observations, for instance, and to perform implicit quality control, by allowing for nonnormal distributions of observational errors. To find the minimum in a reasonable number of iterations, a descent algorithm is used requiring the vector of partial derivatives of J with respect to the elements of $\mathbf{x}$ as in eqn [4], where $\mathbf{H}^T$ is the transpose of the linearization of H , often called the adjoint.

$$\left(\frac{\partial J}{\partial \mathbf{x}}\right)^T = \mathbf{H}^T \mathbf{R}^{-1}(H(\mathbf{x}) - \mathbf{y}^o) + \mathbf{B}^{-1}(\mathbf{x} - \mathbf{x}^b) \quad [4]$$

In practice, further steps are needed to make the minimization problem tractable. The background error covariance cannot be explicitly represented as a matrix: it would be too large, and inverting it is even more impracticable. One approach is to express the penalty function in terms of a transformed model state, for which the background term is simpler. The gradient calculation then also requires the adjoint of this transformation, as well as that of the observation operators.

If the observations in a batch are treated as simultaneous, the resulting algorithm is called 3D-Var. The observation operator procedure of calculating model estimates of the observations can be generalized to include a forecast model

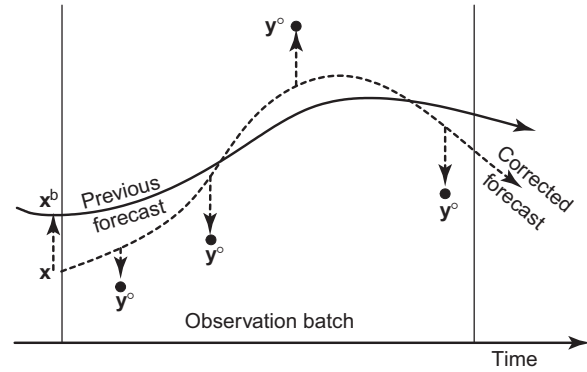


Figure 1 Schematic diagram of 4D-Var. The observed values $\mathbf{y}^o$ are forecast from the model state $\mathbf{x}$ – this process is represented by the dotted lines and by H in eqn [3]. The adjoint of this forecast is used in eqn [4] to determine how to vary $\mathbf{x}$, iterating until the best fit that minimizes J is found.

integration. (If we redefine H to include both the forecast and the observation operator, then the equations above are unaltered.) The adjoint of the forecast model is then needed for the gradient calculation. Each iteration of the minimization's descent algorithm requires a forward integration of the forecast model over a period spanning the observations, and a backward integration of the adjoint model. Since several tens of iterations are generally needed, this 4D-Var algorithm is computationally expensive. 4D-Var is capable of properly assimilating observations distributed in time; for instance, it can use observed tendency information in a dynamically consistent way (Figure 1). Under certain simplifying assumptions, 4D-Var gives the same analysis as a Kalman filter.

Quantifying Uncertainty: Ensembles

A drawback of 4D-Var is that it does not readily provide estimates of the analysis error covariances ($\mathbf{A}$); evaluation using eqn [2] is impracticable. These are required for an algorithm such as the Kalman filter to calculate background error covariances for the next cycle. Without these dynamically varying covariances it is not possible optimally to link the cycles processing batches of observations.

More directly relevant to users of NWP is the desire for an estimate of possible errors in forecasts. The value of forecasts to decision makers is greatly enhanced if the inherent uncertainty can be quantified. This is particularly true of severe weather, which can cause such damage to property and loss of life that precautions may be well advised even if the event is unlikely but possible. Probabilities are a natural way of expressing uncertainty. A range of possible outcomes can be described with associated probabilities, and users can then make informed decisions allowing for their particular costs and risks.

It is impracticable to represent the covariances explicitly as matrices as in the Kalman filter; they are too large to store and manipulate. An alternative approach is to represent the error probability distributions by an ensemble of model states. Because the atmosphere is chaotic, a few patterns of error grow

rapidly, but most do not grow. So it is plausible that a reasonable sample of model states could represent the important components of error covariance matrices, despite having many fewer degrees of freedom. Research is under way into the 'Ensemble Kalman filter' using this approach. Several hundred slightly different analyses are made, and the covariances of forecasts from them are used to approximate any required background error covariances, instead of manipulating and forecasting huge error covariance matrices. The ensemble approach is already well established to forecast uncertainty in medium-range forecasting, but is not yet proven in practice for data assimilation.

Mesoscale Assimilation

Forecasts extending from zero out to about 6 h are based upon a more observations-intensive approach and are referred to as 'nowcasts'. Traditionally, nowcasting has focused on the analysis and extrapolation of observed meteorological fields, with a special emphasis on mesoscale fields of clouds and precipitation derived from satellite and radar. For better handling of developing weather systems, there is increasing emphasis on using NWP for these scales. There are important practical constraints on such a system: it has to assimilate frequent high-resolution data into a high-resolution model, and it has to deliver short-period forecasts promptly, before their utility has expired. To achieve this, usually rather simpler assimilation methods have been employed. One method is to nudge in the observations over a period of time into the

forecast model; for instance, if the observations show rainfall, then the model's temperature variables can be affected by the corresponding latent heating. Correctly tuned, such methods can induce the model state to approach that of the atmosphere.

As computer power increases, it is becoming possible to consider variational methods for assimilating such data. Complicated observation operators, and their adjoints, are needed for predicting precipitation. The variational method tries find a model state that fits the observations to within their estimated error. This potentially can make much better use of cloud and precipitation data.

See also: **Climate and Climate Change:** Climate Prediction: Empirical and Numerical. **Data Assimilation and Predictability:** Ensemble Prediction. **Numerical Models:** Methods; Regional Prediction Models. **Weather Forecasting:** Seasonal and Interannual Weather Prediction; Severe Weather Forecasting.

Further Reading

- Daley, R., 1991. *Atmospheric Data Analysis*. Cambridge University Press, Cambridge.
- Data assimilation in meteorology and oceanography: theory and practice. Special Issue of *Journal of the Meteorological Society of Japan* 75 (1B), 1997.
- Ghil, M., Malanotte-Rizzoli, P., 1991. Data assimilation in meteorology and oceanography. *Advances in Geophysics* 33, 141–266.
- WMO WWRP, 2000. *Proceedings of the 3rd International Symposium on Assimilation of Observations in Meteorology and Oceanography*, Quebec City, June 1999. WMO/TD No. 986.

Ensemble-Based Data Assimilation

Z Meng, Peking University, Beijing, China

F Zhang, Pennsylvania State University, University Park, PA, USA

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

This article introduces the algorithm of ensemble-based data assimilation (EDA) and the main issues in its application to atmospheric sciences. EDA is drawing increasing attentions in data assimilation community mainly due to its flow-dependent background error covariance determined using a short-range ensemble forecast and ease of implementation. Many types of EDA have been applied with different models at different scales in both research and operational or quasi-operational communities. Various aspects involved in EDA are discussed including observations, ensemble initialization, sampling error, covariance inflation and localization, model error, verification, nonlinearity and non-Gaussian errors, intercomparison, and hybrid with variational schemes.

Introduction

The accuracy of weather forecasts is mainly determined by the accuracy of numerical weather prediction (NWP). NWP produces the time evolution of basic meteorological variables including wind, pressure, temperature, moisture, and other hydrometeors based on basic physical laws. Its accuracy depends critically on the qualities of the initial conditions and the forecast model. The initial conditions of an NWP model are usually produced through data assimilation, a procedure that aims to estimate the state and uncertainty of the atmosphere as accurately as possible by combining all available information (including both previous model forecasts and current observations, and their respective uncertainties). Data assimilation methods update the model forecasts through the error covariance of the model forecast between different grid points and variables.

Data assimilation is typically performed through one of the two methods. The first method is variational method, which finds an analysis that has a minimum misfit to the observation and the background forecast (or the first guess). Two main methods in this category are three-dimensional (3DVar, three dimensions in terms of space, Barker et al., 2004) and four-dimensional variational methods (4DVar, three dimensions in terms of space plus one in terms of time, Le Dimet, 1982). Variational methods generally produce their background error covariance using a period of forecast differences between different lead times (Barker et al., 2004; Xiao and Sun, 2007), thus is isotropic and stationary.

The other approach is sequential data assimilation. A theoretically optimum sequential method is the Kalman filter, which assumes Gaussian errors and a linear system. Kalman filter is named as extended Kalman filter for nonlinear models. Since the extended Kalman filter explicitly propagates the background error covariance, it is not practical for large-dimensional systems like the atmosphere or the ocean. Evensen (1994) proposed the ensemble Kalman filter (EnKF) that estimates the background error covariance with a short-term ensemble forecast. The EnKF is drawing increasing attentions in data assimilation community. Since its first application in atmospheric sciences (Houtekamer and

Mitchell, 1998), many types of EnKFs have been applied with different models at different scales (Lorenc, 2003; Houtekamer and Mitchell, 2005; Hamill, 2006; Ehrendorfer, 2007; Meng and Zhang, 2010). The potential of using ensemble-based data assimilation (EDA) is also being explored in several operational meteorological centers. A global- (regional) scale EnKF has been put into operational practice at the Canadian Meteorological Center (Italian Weather Service). There are a few quasi-operational limited area model (LAM)-EnKF systems such as those performed at the University of Washington, the Pennsylvania State University, and the National Center for Atmospheric Research. This article aims to give a brief introduction to the EnKF algorithm and main issues in its application to atmospheric sciences.

The Concept of the EnKF

The EnKF has two steps: an analysis step and a forecast step (Figure 1). The analysis step in the standard EnKF proceeds according to the classical Kalman filter equations with the required sample means and covariances:

$$\mathbf{x}_i^a = \mathbf{x}_i^b + \mathbf{K}(\mathbf{y}_i^o - \mathbf{H}\mathbf{x}_i^b), \quad \text{for } i = 1, \dots, n \quad [1]$$

$$\mathbf{K} = \mathbf{P}^b \mathbf{H}^T (\mathbf{H} \mathbf{P}^b \mathbf{H}^T + \mathbf{R})^{-1} \quad [2]$$

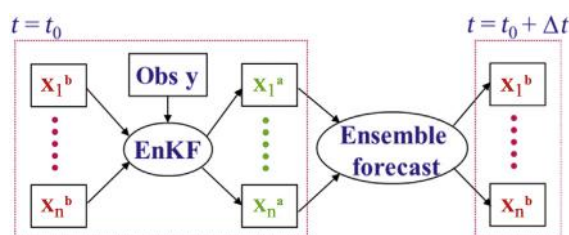


Figure 1 Schematic diagram of the flow chart of the EnKF. The meaning of the various variables is referred to the text. EnKF, ensemble Kalman filter.

$$\mathbf{P}^b = \frac{1}{n-1} \sum_{i=1}^n \left[\left(\mathbf{x}_i^b - \bar{\mathbf{x}}^b \right) \left(\mathbf{x}_i^b - \bar{\mathbf{x}}^b \right)^T \right] \quad [3]$$

$$\mathbf{P}^a = \frac{1}{n-1} \sum_{i=1}^n \left[\left(\mathbf{x}_i^a - \bar{\mathbf{x}}^a \right) \left(\mathbf{x}_i^a - \bar{\mathbf{x}}^a \right)^T \right] \quad [4]$$

where, $\bar{\mathbf{x}}^b$ and $\bar{\mathbf{x}}^a$ are the sample mean values of the forecast and the analysis ensemble. The analysis state vector, $\mathbf{x}_i^a$ for the i th ensemble member is obtained by adding to the background state vector, $\mathbf{x}_i^b$, a weighted difference between observations, $\mathbf{y}^o$, and the background state vector projected to observation space through an observation operator, H . $\mathbf{P}^a$ and $\mathbf{P}^b$ are the analysis and background error covariance respectively, which are approximated in the EnKF by the respective ensemble covariance. $\mathbf{K}$ is the Kalman gain, which takes into account the observation error, $\mathbf{R}$, and the background error covariance, $\mathbf{P}^b$ in determining the extent to which observations are weighted relative to the background. $\mathbf{H}$ and $\mathbf{H}^T$ are the tangent linear and adjoint versions of H , respectively. Tangent linear means the linearization about a time or space-varying trajectory of a nonlinear operator, which gives the evolution of the first-order perturbation of a nonlinear vector. In the EnKF, $\mathbf{P}^b \mathbf{H}^T$ and $\mathbf{H} \mathbf{P}^b \mathbf{H}^T$ are calculated using the sample covariance and thus $\mathbf{H}^T$ and $\mathbf{H}$ are not explicitly needed.

Given the analysis ensemble, the forecast step simply involves forecasting each member forward from time t to $t + \Delta t$ when the next observations are available:

$$\mathbf{x}_i^b(t + \Delta t) = M(\mathbf{x}_i^a(t)), \quad \text{for } i = 1, \dots, n \quad [5]$$

where M is a nonlinear model. The ensemble-based algorithm asymptotically approaches the Kalman filter in the limit of a large ensemble and Gaussian error distributions.

The EnKF has many variants. Based on the method for generating the analysis ensemble, an EnKF can be characterized as stochastic, where the analysis ensemble is obtained with the Kalman gain and randomly perturbed observations, or deterministic, where the analysis ensemble is created by transforming the forecast ensemble without perturbing the observations. Due to the large dimension of the model state vector and the large number of observations needed in the LAM-EnKF scenario, various algorithms of the LAM EnKF have been proposed to improve computational efficiency. The most commonly employed deterministic method has the form of ensemble square root filters (EnSRFs) such as the serial EnSRF, the ensemble transform Kalman filter, and the ensemble adjustment Kalman filter.

Observations

Different observing platforms or different formulations of the same observations may have different impacts on the EnKF performance. For example, radiosonde observations have been found to have the largest impact on regional-scale analyses and forecasts. Due to errors coming from the difference between the real and the model terrain height and uncertainties in the parameterization of boundary layer and land surface physical processes, surface observations have been a big challenge. The analysis and forecast error may benefit from removing the bias and reducing the representativeness error by assimilating only those data whose locations have a terrain height that matches

the model one, and removing the bias of surface wind and temperature before assimilation. However, surface data assimilation may have very limited impact on the levels above the surface and on the time length of the forecast into the integration.

Besides the conventional radiosonde and surface observations, remotely sensed observations are drawing more and more attention. Doppler radars such as Doppler on Wheels, Weather Surveillance Radar 88 Doppler, and airborne radars have sufficient temporal and spatial coverage to fully observe convective clouds. The effectiveness of using the EnKF to assimilate Doppler radar velocities for supercell storms was first demonstrated in [Snyder and Zhang \(2003\)](#) and with real-data in [Dowell et al. \(2004\)](#). An X-band radar network that was deployed by the National Science Foundation Engineering Research Center for Collaborative Adaptive Sensing of the Atmosphere has been helpful in improving the analysis and forecast of convective system, which supports convective-scale warn-on-forecast operations. Airborne Doppler radar radial-velocity observations have also been successfully assimilated in real time for a hurricane prediction model ([Zhang et al., 2011](#)). The EnKF system has demonstrated promising performance, especially on hurricane intensity forecasts, through experiments over 61 applicable The National Oceanic and Atmospheric Administration (NOAA) P-3 airborne Doppler missions during the 2008–10 Atlantic hurricane seasons ([Figure 2](#)). The mean absolute intensity forecast errors initialized with the EnKF analysis of the airborne Doppler velocities at the 24–120-h lead forecast times were 20–40% lower than the National Hurricane Center's official forecasts issued at similar times. Radar reflectivity has been shown to be less effective than Doppler radial velocity. The likely non-Gaussian error distribution, weak cross-correlations between state variables, inherent small-scale variability, and strong dependence of these quantities on the accuracy of model microphysics schemes appear to be the main limiting factors. Nevertheless, the assimilation of differential reflectivity, Z_{dr} , reflectivity difference, Z_{dp} , and specific differential phase, K_{dp} , beyond radar reflectivity and/or radial velocity may improve storm analysis ([Jung et al., 2008](#)). Moreover, the assimilation of even the echo-free radar observations, defined as radar reflectivity below a threshold value of 5 dBZ, sometimes may effectively suppress spurious convection ([Aksoy et al., 2009](#)).

Radar data need to be vigorously preprocessed to have a consistently positive impact. With large volumes of radar observations recorded at a much higher resolution than the forecast model grid spacing, significant data thinning may be necessary. The process of combining multiple observations into one high-accuracy 'super' observation is often referred to as 'superobbing.' This data preprocessing procedure has been implemented operationally on the NOAA hurricane reconnaissance aircraft that allows for more efficient real-time transmission of airborne radar observations to the ground ([Zhang et al., 2011](#)). Assimilating airborne Doppler radar data in storm-relative coordinates under a simultaneity assumption may produce a better kinematic tropical cyclone (TC) structure and a slower error growth ([Aksoy, 2013](#)). This method translates the observations of the whole flight to a common storm center by maintaining their relative positions at the time of their actual sampling, then randomly distributed to different cycles to achieve a more homogenous data coverage among data assimilation cycles.

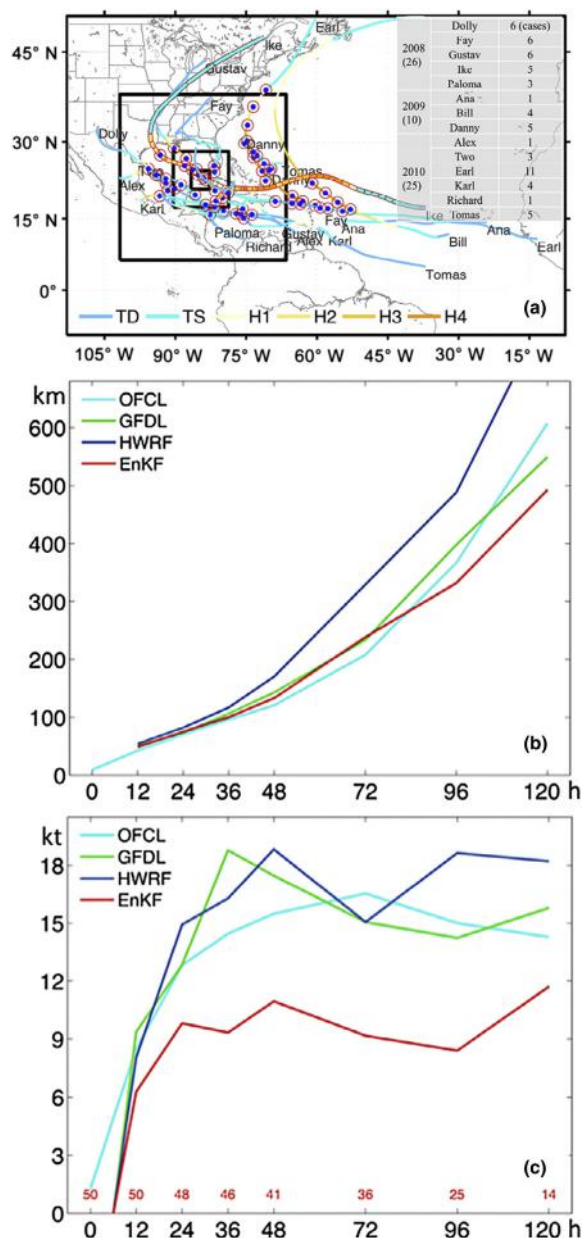


Figure 2 (a) The Weather Research and Forecasting model (WRF) model domain configuration and TC tracks with NOAA airborne Doppler radar mission. The outer domain is fixed for all cases, while the three inner domains are centered at the storm's center at the initial time and movable flowing model vortex center during the forecast. Fourteen storm tracks are colored lines with storm intensity. With blue dots circled with red, 61 missions of NOAA airborne Doppler radar observation are marked. All the storms and case numbers are listed on the right top. Also shown are the mean absolute forecast error averaged over 50 samples homogenized by all 61 airborne Doppler missions during 2008–10 for The National Hurricane Center (NHC) Official forecast ('OFCL,' solid cyan line), The Geophysical Fluid Dynamics Laboratory (GFDL) hurricane model ('GFDL,' green line), The Hurricane Weather Research and Forecasting model ('HWRF,' dark blue line), and the WRF deterministic forecast in 4.5-km resolution initialized with EnKF analysis ('EnKF,' red line) after making the samples homogeneous for all five forecasts for (b) the track position error (km) and (c) the 10-m maximum wind speed error (knots) with simple bias correction

In addition to radars, the assimilation of satellite observations and/or satellite-derived products is also a hot topic. Multivariate specific humidity retrieved from the Atmospheric Infrared Sounder (AIRS), satellite-derived atmospheric motion vectors, and temperature and mixing ratio profiles derived from the AIRS instrument on board the *Aqua* satellite have been shown beneficial for weather analysis and forecast (Jones and Stensrud, 2012). However, direct assimilation of satellite radiance with the EnKF is still in its infancy. Challenging Minisatellite Payload radio occultation refractivity has been found to be beneficial in regions where conventional high-quality observations are sparse. Liu et al. (2012) reported a better performance in assimilating The Advanced Microwave Sounding Unit (AMSU) radiance than without radiance assimilation for the track and intensity forecast of five TC events in 2008 over the Atlantic Ocean with EnKF. A track error reduction of up to 16% was achieved for forecasts beyond 36 h largely due to the improvement in the representation of TC environments. Results also showed more benefit when both radiances and satellite winds than when radiances alone were assimilated. Many issues remain to be explored in satellite radiance assimilation such as observation bias correction.

Besides the aforementioned *in situ* and remotely sensed data, some special synthesized object-oriented observations, such as the vortex position of tropical cyclones, can also be easily assimilated by the EnKF and have been demonstrated to be helpful in improving hurricane forecast ability.

Ensemble Initialization

How to create an initial ensemble that reflects the uncertainty of the initial analysis is an important and critical issue. Global EnKFs typically use random samples from a climatological error distribution, or generate perturbations from a preexisting 3D/4DVar system. The LAM-based EnKF may be initialized from an existing global or larger scale ensemble or by using one of the methods used for the global model if a global ensemble forecast is not readily accessible. The generation of the initial ensemble for convective-scale EnKF systems remains an open question due to the lack of accurate error statistics. For many convective-scale applications, Gaussian noises are added to a horizontally uniform background (sounding) for all state variables (e.g., Snyder and Zhang, 2003).

Compared to the initial condition uncertainties, a proper representation of boundary uncertainties may have a larger impact on the LAM EnKF. A lack of sufficient ensemble spread on the lateral boundaries may propagate inward and lead to filter divergence. Filter divergence means that the ensemble mean deviates further and further away from the truth due to the underestimated variance of the forecast ensemble, which results in more weight being given to the prior (model forecast) than to the observations. Boundary perturbations can be

for all forecasts except for OFCL. The red numbers on the x axis of (c) indicate the sample size for the homogeneous comparison. Adapted from Zhang, F., Weng, Y., Gamache, J.F., Marks, F.D., 2011. Performance of convection permitting hurricane initialization and prediction during 2008–2010 with ensemble data assimilation of inner-core airborne Doppler radar observations. *Geophysical Research Letters* 38, L15810. <http://dx.doi.org/10.1029/2011GL048469>.

generated using the same methods as ensemble initialization described above.

Sampling Error, Covariance Inflation, and Localization

Due to computational constraints, only a limited ensemble size can be afforded in the EnKF, which will result in sampling errors. The number of members sufficient for the EnKF to minimize the impact of sampling errors still remains an open question. Until now, most published EnKF studies use an ensemble size of 30–100. Due to the limited ensemble size, the EnKF generically suffers from a rank-deficiency problem in which only a part of the phase space can be spanned by the ensemble. As a result, the ensemble spread tends to be systematically underestimated.

The underestimation of ensemble spread is commonly treated by covariance inflation through multiplicative or additive scaling, or covariance relaxation. Additive covariance inflation adds a set of ensemble perturbations that can reflect forecast uncertainty to the forecast ensemble, where multiplicative covariance inflation is achieved by multiplying all ensemble perturbations before or after the EnKF analysis with a constant slightly larger than 1. However, multiplicative covariance inflation with a spatially constant inflation factor may cause a model to become unstable due to excessive spread in data-sparse regions. A Bayesian algorithm was proposed using a spatially and temporally varying adaptive inflation factor for each element of the model state vector (Anderson, 2009). Another approach is the covariance relaxation method through relaxation of the final analysis perturbations to the prior forecast (Zhang et al., 2004). The relaxation method only inflates the variance at grid points that are updated by observations, thus avoiding the overinflation deficiency of the conventional inflation method.

Besides the underestimation of ensemble spread, sampling errors in the EnKF may also cause unphysical, distant correlations and subsequently spurious analysis increments. The most commonly used approach for reducing distant spurious correlations is through covariance localization. Covariance localization is often implemented by multiplying a Gaussian shape function that decreases smoothly from one at the observation point to zero at a certain distance from the observation. The impact of an observation is thus confined only within a limited distance. This distance is usually called the radius of influence (ROI). Localization may not only decrease spurious distant correlation, but also reduce the computational cost and alleviate the rank-deficiency problem due to the limited ensemble size. The value of ROI may depend on ensemble size, observation type and density, model error and resolution, as well as the characteristic scales of the underlying dynamic system. For examples, a horizontal ROI of 1000–2000 (60–150) km is often used for standard radiosonde (surface) observations, while a much smaller ROI ranging from several to hundreds of kilometers is often used for radar observations. It was once suggested that the ROI should be two to three times the forecast error scales (Lorenc, 2003). Zhang et al. (2009a) proposed a successive covariance localization (SCL) technique in which a larger ROI is used to assimilate a relatively small subset of observations in the coarser domains, while a smaller ROI is used to assimilate higher density

observations in the inner domains. They found clear advantages of using the SCL method over using single ROIs in the assimilation of dense radar observations for rapidly developing landfalling hurricanes. The vertical ROI is sometimes set to the depth of the atmospheric model for conventional observation assimilation or vertical covariance localization is not performed at all. Besides ROI, the selection of a covariance localization function may also be important. The most widely used covariance localization is the fifth-order correlation function of Gaspari and Cohn (1999). The choice of localization function may become more complicated for observations that have complex spatial, temporal, and physical attributes or with an unknown relation with the state variable.

While accounting for sampling error, covariance localization may cause imbalance when one observation is selected to update the state vector at one grid point, but not selected for a neighboring grid point (Lorenc, 2003). Consequently, covariance localization may produce analyses with weaker flow balance and stronger divergence, which may result in inaccurately balanced background error statistics. Methods used to reduce imbalance in large-scale models such as applying a digital filter (e.g., Lynch and Huang, 1992) or covariance localization performed in the stream function–velocity potential rather than the wind component space (Kepert, 2009).

Model Error

Since the EnKF depends critically on the quality of the first guess and the forecast error covariance estimated from a short-term ensemble forecast, the presence of model error may lead to poor filter performance and even filter divergence. Model error can result from inadequate parameterizations of subgrid-scale physical processes, numerical inaccuracy, truncation error, ill-defined boundary conditions, or other random errors. The presence of model error often results in both a large bias in the ensemble mean and too little spread, which may ultimately cause the ensemble forecast to fail.

There are several ad hoc approaches that have been used to account for model error in the context of the EnKF such as covariance inflation, bias correction, the use of multimodel or multiphysics ensemble, and simultaneous state and parameter estimation. Additive covariance inflation has been shown to be effective in improving the performance of the EnKF. Though most statistical data assimilation methods assume that the model forecast (or first guess) is unbiased, this is rarely the case. Over the past decade, there has been an increasing amount of evidence demonstrating the advantages and effectiveness of using multimodel ensembles (over single-model ensembles) to account for model error in the prediction system (e.g., Krishnamurti et al., 1999; Weisheimer et al., 2009). However, given technical implementation difficulties associated with inherent differences in model numerics, dynamical coordinates, and/or (prognostic) state variables among different forecast models, multimodel ensembles have not been used for the EnKF. Since a considerable part of the model error comes from the parameterization of subgrid-scale physical processes, a more practical approach is to use a variety of physical parameterization schemes available in the same forecast model or a variety of values of some parameters in a particular scheme for different members to account for model uncertainties. This

multiphysics ensemble approach has been shown to be effective in accounting for model error in both global and mesoscale ensemble forecast systems (Meng and Zhang, 2007; Fujita et al., 2007), likely through improved ensemble mean estimates, an increased ensemble spread, and a more effective background error covariance. With the development of more sophisticated physical schemes, how to assign weights of various available schemes in the ensemble may become a critical issue. Besides, stochastic kinetic energy backscatter (Shutts, 2005) has been used as more physically based parameterization of model error (Leutbecher, 2007).

The uncertainty in physical parameterization schemes is closely related to the uncertainty of its parameters. Almost all parameters of subgrid physical parameterization schemes are empirical, due to a lack of direct observations, and could therefore have large and unknown uncertainties. Parameter estimation is a technique in which the EnKF is used to estimate these parameters by treating them as additional model variables and updating them through data assimilation. This technique may help improve the performance of the EnKF via a model error correction. Since there are no direct observations or physical evidence describing the variability of various parameters, the generation of a realistic initial ensemble for the estimated parameter is even more difficult than for standard state variables. Furthermore, not all parameters can be successfully estimated. The performance of a parameter estimation algorithm depends on the correlation between the parameter and model variables and is closely related to the EnKF configuration. Parameter estimation performance using the EnKF is also closely associated with the number of simultaneously estimated parameters. It was found that the estimation of a single parameter is very effective in improving the performance of the EnKF. Increasing the number of estimated parameters inevitably leads to a decline in the improvement from parameter estimation (Aksoy et al., 2006).

Verification

The realism of the analysis and forecast ensembles is usually verified using rank histogram (Anderson, 1996; Hamill and Colucci, 1997). A rank histogram describes the extent to which an ensemble encompasses the verifying data by ranking the verifying data in the sorted ensemble. The reliability of an ensemble can be diagnosed by the shape of its rank histogram. A flat shape implies that the observation can be taken as a random member of the ensemble, and consequently the ensemble is reliable. A U-shape suggests that the ensemble spread is insufficient, while a reversed (upside down) U-shape indicates that the ensemble spread is too large.

Grid point-based root mean square error (RMSE) of different variables integrated over the whole model domain is usually used for verification for the analysis and/or forecast ensemble means of the EnKF. Due to a lack of dense, conventional mesoscale observations, verification of the LAM EnKF can be more difficult than verification of larger scale prediction systems, especially for radar data assimilation. One way is to compare the analysis and/or forecast against radar observations that are not assimilated, but saved specifically for assessment. The equitable threat score can be used to verify the result in

terms of radar reflectivity. An alternative metric that is more pattern-based is the reflectivity correlation coefficient between the observed and simulated reflectivity in observation space. For severe weather systems, such as hurricanes, how to choose an appropriate error metric is still an open question. Since a small displacement of a storm center may result in a substantially large RMSE of wind, performance could be better assessed using feature-based verification.

Nonlinearity and Non-Gaussian Errors

Though the EnKF can be used in nonlinear and non-Gaussian systems, the performance of the EnKF may be affected and ultimately limited by these two characteristics. Variants of algorithms for highly nonlinear and non-Gaussian systems have been proposed, such as the particle filter, the particle Kalman filter, the iterative EnKF and extended Kalman filter, morphing methods, 'Running in place,' and the 'quasi-outer loop.' While these methods have been applied for simple models, many have not been tested on high-dimensional systems, such as full-physics atmospheric models. How to construct proper algorithms to deal with high nonlinearity and non-Gaussianity awaits further efforts.

Intercomparison and Hybrid with Variational Schemes

Despite many of the challenging issues discussed in the previous sections, there are several appealing advantages of the EnKF in comparison to the variational data assimilation techniques. These advantages include: (1) The background error covariance is flow-dependent, which reflects the errors of the day; (2) The model and observation operator can be nonlinear; (3) It provides not only the best estimation of the state, but also the associated flow-dependent uncertainty; therefore, it can be seamlessly coupled with ensemble forecasting; (4) There is no need to code a tangent linear or adjoint model; (5) It is easier to account for model error due to its use of an ensemble forecast; and (6) The ensemble members can be run simultaneously, making it easy to parallelize.

Many studies have shown that the EnKF generally compares favorably with 3DVar (Meng and Zhang, 2008a,b) and 4DVar (Caya et al., 2005). Zhang et al. (2013) found that the advantage of the EnKF over both 3DVar and 4DVar becomes increasingly more after 36-h forecast time for all prognostic variables, while the EnKF moisture forecast field is superior to both 3DVar and 4DVar at all lead times despite fitting less closely to the observations at the analysis time.

Both the EnKF and variational method have their own advantages and disadvantages. The EnKF benefits from its flow-dependent background error covariance but suffers from rank deficiency, while the variational technique has advantages in its analysis algorithm. The variational approach can more efficiently assimilate large numbers of observations and apply physical constraints on the solution when minimizing the cost function, however, it is limited by the use of a static background error covariance matrix. Instead of settling on one particular method, more and more efforts are devoted to the

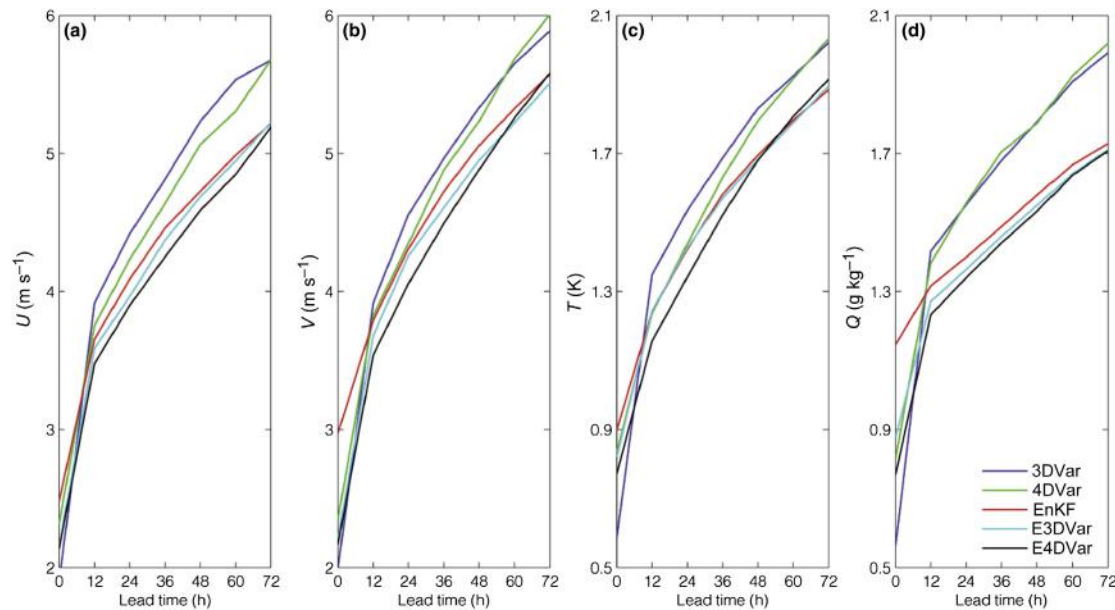


Figure 3 Comparison between WRF-based 3DVar, 4DVar, EnKF, E3DVar, and E4DVar in terms of domain-averaged root mean square errors averaged over all 59 WRF forecasts from June 2003 for each Data Assimilation (DA) experiment at forecast lead times from 0 to 72 h evaluated every 12 h for (a) U (m s^{-1}), (b) V (m s^{-1}), (c) T (K), and (d) Q (g kg^{-1}). 3DVar, three-dimensional variation; 4DVar, four-dimensional variation; EnKF, ensemble Kalman filter. Adapted from Zhang, F., Zhang, M., Poterjoy, J., 2013. E3DVar: coupling an ensemble Kalman filter with three-dimensional variational data assimilation in a limited-area weather prediction model and comparison to E4DVar. *Monthly Weather Review* 141, 900–917. <http://dx.doi.org/10.1175/MWR-D-12-00075.1>.

hybridization of the two approaches. A two-way coupled assimilation scheme benefits from using the state-dependent uncertainty provided by the EnKF, while taking advantage of the variational method from data assimilation algorithm. The EnKF/3DVar hybrid system of The Global Forecasting System (GFS)/The National Centers for Environmental Prediction (NCEP) has been showing consistently better TC track forecast than the operational Gridpoint Statistical Interpolation. A hybrid of EnKF with 4DVar is regarded as one of the most advanced and most promising (as well as most computationally and technically demanding) data assimilation methods in both the research and operational communities. Zhang et al. (2013) demonstrated that EnKF/4DVar produced considerable improvements over EnKF/3DVar at forecast lead times of 1–2 days before they converged to similar results at longer lead times (Figure 3). Both EnKF/3DVar and EnKF/4DVar produced considerably better forecast than stand-alone 3DVar and 4DVar methods that relied solely on static background error covariance.

It is important to bear in mind that a hybrid system will likely inherit issues or complexity from the component systems such as the rank-deficiency problem in the EnKF and the use of an outer loop in 4DVar. More efforts are needed to tackle these issues in the future. Considering the encouraging results obtained by coupling the EnKF with variational methods, several major operational NWP centers have already started testing or implementing such approaches in their global data assimilation systems such as Environment Canada, Météo-France, the European Center for Medium-Range Weather Forecasts, UK Met, and NCEP. Similar efforts for limited area data assimilation systems have not been reported.

See also: Data Assimilation and Predictability: Data Assimilation; Ensemble Prediction; Predictability and Chaos. **Numerical Models:** Methods; Model Physics Parameterization; Parameter Estimation; Regional Prediction Models. **Radar:** Polarimetric Doppler Weather Radar. **Statistical Methods:** Data Analysis: Empirical Orthogonal Functions and Singular Vectors. **Tropical Cyclones and Hurricanes:** Hurricane Predictability; Hurricanes: Observation.

References

- Aksoy, A., 2013. Storm-relative observations in tropical cyclone data assimilation with an ensemble Kalman filter. *Monthly Weather Review* 141, 506–522.
- Aksoy, A., Dowell, D.C., Snyder, C., 2009. A multicase comparative assessment of the ensemble Kalman filter for assimilation of radar observations. Part I: storm-scale analyses. *Monthly Weather Review* 137, 1805–1824.
- Aksoy, A., Zhang, F., Nielsen-Gammon, J.W., 2006. Ensemble-based simultaneous state and parameter estimation with MM5. *Geophysical Research Letters* 33, L12801. <http://dx.doi.org/10.1029/2006GL026186>.
- Anderson, J.L., 1996. A method for producing and evaluating probabilistic forecasts from ensemble model integrations. *Journal of Climate* 9, 1518–1530.
- Anderson, J.L., 2009. Spatially and temporally varying adaptive covariance inflation for ensemble filters. *Tellus* 61A, 72–83.
- Barker, D.M., Huang, W., Guo, Y.-R., Bourgeois, A.J., Xiao, Q.N., 2004. A three-dimensional variational data assimilation system for MM5: implementation and initial results. *Monthly Weather Review* 132, 897–914.
- Caya, A., Sun, J., Snyder, C., 2005. A comparison between the 4DVAR and the ensemble Kalman filter techniques for radar data assimilation. *Monthly Weather Review* 133, 3081–3094.
- Dowell, D.C., Zhang, F., Wicker, L.J., Snyder, C., Crook, N.A., 2004. Wind and temperature retrievals in the 17 May 1981 Arcadia, Oklahoma, supercell: ensemble Kalman filter experiments. *Monthly Weather Review* 132, 1982–2005.

- Ehrendorfer, M., 2007. A review of issues in ensemble-based Kalman filtering. *Meteorologische Zeitschrift* 16, 795–818.
- Evensen, G., 1994. Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *Journal of Geophysical Research* 99, 10143–10162.
- Fujita, T., Stensrud, D.J., Dowell, D.C., 2007. Surface data assimilation using an ensemble Kalman filter approach with initial condition and model physics uncertainties. *Monthly Weather Review* 135, 1846–1868.
- Gaspari, G., Cohn, S.E., 1999. Construction of correlation functions in two and three dimensions. *Quarterly Journal of The Royal Meteorological Society* 125, 723–757.
- Hamill, T.M., 2006. Ensemble-based atmospheric data assimilation. In: Palmer, T., Hagedorn, R. (Eds.), *Predictability of Weather and Climate*. Cambridge University Press, Cambridge, UK, pp. 124–156.
- Hamill, T.M., Colucci, S.J., 1997. Verification of Eta–RSM short range ensemble forecasts. *Monthly Weather Review* 125, 1312–1327.
- Houtekamer, P.L., Mitchell, H.L., 1998. Data assimilation using an ensemble Kalman filter technique. *Monthly Weather Review* 126, 796–811.
- Houtekamer, P.L., Mitchell, H.L., 2005. Ensemble Kalman filtering. *Quarterly Journal of The Royal Meteorological Society* 131, 3269–3289.
- Jones, T.A., Stensrud, D.J., 2012. Assimilating AIRS temperature and mixing ratio profiles using an ensemble Kalman filter approach for convective-scale forecasts. *Weather Forecasting* 27, 541–564.
- Jung, Y., Xue, M., Zhang, G., Straka, J.M., 2008. Assimilation of simulated polarimetric radar data for a convective storm using the ensemble Kalman filter. Part II: impact of polarimetric data on storm analysis. *Monthly Weather Review* 136, 2246–2260.
- Kepert, J.D., 2009. Covariance localization and balance in an ensemble Kalman filter. *Quarterly Journal of The Royal Meteorological Society* 135, 1157–1176.
- Krishnamurti, T.N., Kishtawal, C.M., LaRow, T.E., Bachiochi, D.R., Zhang, Z., Williford, C.E., Gadgil, S., Surendran, S., 1999. Improved weather and seasonal climate forecasts from multimodel superensemble. *Science* 285, 1548–1550.
- Le Dimet, F.X., 1982. A General Formalism of Variational Analysis. CIMMS Rep. 22, Cooperative Institute for Mesoscale Meteorological Studies, 34 pp. [Available from CIMMS, 815 Jenkins St., Norman, OK 73019.]
- Leutbecher, M., Buizza, R., Isaksen, I., 2007. Ensemble forecasting and flow-dependent estimates of initial uncertainty. In: *Proceedings of the ECMWF Workshop on Flow-Dependent Aspects of Data Assimilation*, Reading, ECMWF, United Kingdom, pp. 185–201.
- Liu, Z., Schwartz, C.S., Snyder, C., Ha, S.-Y., 2012. Impact of assimilating AMSU-A radiances on forecasts of 2008 Atlantic tropical cyclones initialized with a limited-area ensemble Kalman filter. *Monthly Weather Review* 140, 4017–4034.
- Lorenc, A.C., 2003. The potential of the ensemble Kalman filter for NWP—a comparison with 4D-Var. *Quarterly Journal of The Royal Meteorological Society* 129, 3183–3203.
- Lynch, P., Huang, X., 1992. Initialization of the HIRLAM model using a digital filter. *Monthly Weather Review* 120, 1019–1034.
- Meng, Z., Zhang, F., 2007. Tests of an ensemble Kalman filter for mesoscale and regional-scale data assimilation. Part II: imperfect model experiments. *Monthly Weather Review* 135, 1403–1423.
- Meng, Z., Zhang, F., 2008a. Tests of an ensemble Kalman filter for mesoscale and regional-scale data assimilation. Part III: comparison with 3DVar in a real-data case study. *Monthly Weather Review* 136, 522–540.
- Meng, Z., Zhang, F., 2008b. Tests of an ensemble Kalman filter for mesoscale and regional-scale data assimilation. Part IV: comparison with 3DVar in a month-long experiment. *Monthly Weather Review* 136, 3671–3682.
- Meng, Z., Zhang, F., 2010. Limited-area ensemble-based data assimilation. *Monthly Weather Review* 139, 2025–2045.
- Shutts, G., 2005. A kinetic energy backscatter algorithm for use in ensemble prediction systems. *Quarterly Journal of The Royal Meteorological Society* 131, 3079–3102.
- Snyder, C., Zhang, F., 2003. Assimilation of simulated Doppler radar observations with an ensemble Kalman filter. *Monthly Weather Review* 131, 1663–1677.
- Weisheimer, A., Doblas-Reyes, F.J., Palmer, T.N., Alessandri, A., Arribas, A., Déqué, M., Keenlyside, N., MacVean, M., Navarra, A., Rogel, P., 2009. ENSEMBLES: a new multi-model ensemble for seasonal-to-annual predictions—skill and progress beyond DEMETER in forecasting tropical Pacific SSTs. *Geophysical Research Letters* 36, L21711. <http://dx.doi.org/10.1029/2009GL040896>.
- Xiao, Q., Sun, J., 2007. Multiple-radar data assimilation and short-range quantitative precipitation forecasting of a squall line observed during IHOP_2002. *Monthly Weather Review* 135, 3381–3404.
- Zhang, F., Snyder, C., Sun, J., 2004. Impacts of initial estimate and observation availability on convective-scale data assimilation with an ensemble Kalman filter. *Monthly Weather Review* 132, 1238–1253.
- Zhang, F., Weng, Y., Gamache, J.F., Marks, F.D., 2011. Performance of convection-permitting hurricane initialization and prediction during 2008–2010 with ensemble data assimilation of inner-core airborne Doppler radar observations. *Geophysical Research Letters* 38, L15810. <http://dx.doi.org/10.1029/2011GL048469>.
- Zhang, F., Weng, Y., Sippel, J.A., Meng, Z., Bishop, C.H., 2009. Cloud-resolving hurricane initialization and prediction through assimilation of Doppler radar observations with an ensemble Kalman filter. *Monthly Weather Review* 137, 2105–2125.
- Zhang, F., Zhang, M., Poterjoy, J., 2013. E3DVar: coupling an ensemble Kalman filter with three-dimensional variational data assimilation in a limited-area weather prediction model and comparison to E4DVar. *Monthly Weather Review* 141, 900–917. <http://dx.doi.org/10.1175/MWR-D-12-00075.1>.

Ensemble Prediction

R Buizza, ECMWF, Reading, UK

© 2003 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 6, pp 2546–2557, © 2003, Elsevier Ltd.

Introduction

The atmosphere is a complex dynamical system with many degrees of freedom. In numerical weather prediction, the state of the atmosphere is described by the spatial distribution of wind, temperature, specific humidity, liquid water content, and surface pressure. The mathematical differential equations used to predict the system's time evolution include Newton's laws of motion and the laws of thermodynamics. Numerical weather prediction models predict the time evolution of the atmospheric state by solving the system equations numerically.

A deterministic forecast is a single integration of the system equations. The practical usefulness of a single deterministic weather forecast is limited by the day-to-day variability in its accuracy. This variability is partly associated with fluctuations in the predictability of the atmospheric flow, with predictable states (i.e., flows characterized by a slow amplification of initial errors) alternating with unpredictable states (i.e., flows characterized by a fast amplification of initial errors).

Ensemble systems are practical tools designed to assess the predictability of the daily atmospheric flow. More generally, they can be used to predict the time evolution of the probability density function (PDF) of forecast states. They can be used, for example, to predict the probability of intense rainfall or cold temperatures over the Euro-Atlantic region (Figure 1).

Ensemble systems should be designed to simulate the effect of all sources of forecast errors. In particular, they should simulate the effect of uncertainties in the knowledge of the initial state of the system and the effects of the approximations made in numerical weather prediction models.

Ensemble systems have been operational since 1992 at the European Centre for Medium-Range Weather Forecasts (ECMWF) in the United Kingdom and at the National Centers for Environmental Prediction (NCEP) in the United States, and since 1995 at the Meteorological Center of Canada (MSC). These three ensemble systems have been designed to estimate the forecast PDF in the short- and medium-forecast range, i.e., for up to 14 days. Beside this operational activity, many international centers, universities, and national and regional meteorological centers have been involved in research and experimental activities in this field. It should also be mentioned that experimental ensemble systems are currently under development and are tested for seasonal time scales.

Despite that fact that most of the examples and of the discussions reported here are based on results obtained during the past years by the operational medium-range ensemble systems, most of the discussions can be applied to seasonal ensemble systems.

Numerical Weather Prediction

In numerical models, the state of the atmosphere is described for a finite number of vertical levels and at a series of grid points

by a set of state variables such as temperature T , velocity v , specific humidity and cloud liquid content q , and surface pressure p . In other words, the state vector y of the system is defined by (T, v, q, p) for all vertical levels and all grid points. The phase space of the system is the N -dimensional space defined by the (T, v, q, p) coordinates.

The system attractor is defined by the set of past, present, and future atmospheric states $y(t)$. For each time t , a unique point on the system attractor identifies the state of the atmosphere. The time evolution of atmospheric states during subsequent times $t_0 < t < t_1$ between the initial time t_0 and the final time t_1 , is represented by the set of points that described the system orbit between t_0 and t_1 (Figure 2).

The mathematical differential equations describing the atmospheric motions include Newton's laws of motion in the form 'acceleration equals force divided by mass', the laws of conservation of mass and energy, and the laws of thermodynamics. These equations, written in terms of the state variables y , can be approximated with a set of algebraic difference equations for the tendencies (i.e., for the variations in time) of the state variables.

These equations include parameterization schemes that simulate the effects of physical processes such as radiative transfer, turbulent mixing, orographic forcing, and moist processes. The parameterization of these physical processes is probably one of the most difficult and uncertain areas of weather modeling.

Schematically, the equations of motion defined in terms of the state variables $y = (T, v, q, p)$ can be written as eqn [1], where A and P identify, respectively, the contribution to the full equation tendency of the nonparameterized and parameterized physical processes.

$$\frac{\partial y}{\partial t} = A(y, t) + P(y, t) \quad [1]$$

The time integration of eqn [1] from the initial time t_0 to the forecast time t (eqn [2]) describes the time evolution of the atmospheric flow from the initial state $y(t_0)$ to the final state. equation (2)

$$y(t) = y(t_0) + \int_{t_0}^t [A(y, \tau) + P(y, \tau)] d\tau \quad [2]$$

The initial state of the system, $y(t_0)$, is defined by observed weather variables. Denote by $\sigma\tau$ the vector of all meteorological observations made in a time interval 2Δ centered at time t_0 , $t_0 - \Delta < \tau < t_0 + \Delta$. The initial state of the system $y(t_0)$ is defined in such a way that, for all times τ with $t_0 - \Delta < \tau < t_0 + \Delta$, $y(\tau)$ is closest to $\sigma(\tau)$. Practically, $y(t_0)$ is computed by minimizing a cost function $J(y(t_0))$ (eqn [3]).

$$J(y(t_0)) = d_B(y(t_0), y_B(t_0)) + \int_{t_0-\Delta}^{t_0+\Delta} d_O(y(\tau), O(\tau)) d\tau \quad [3]$$

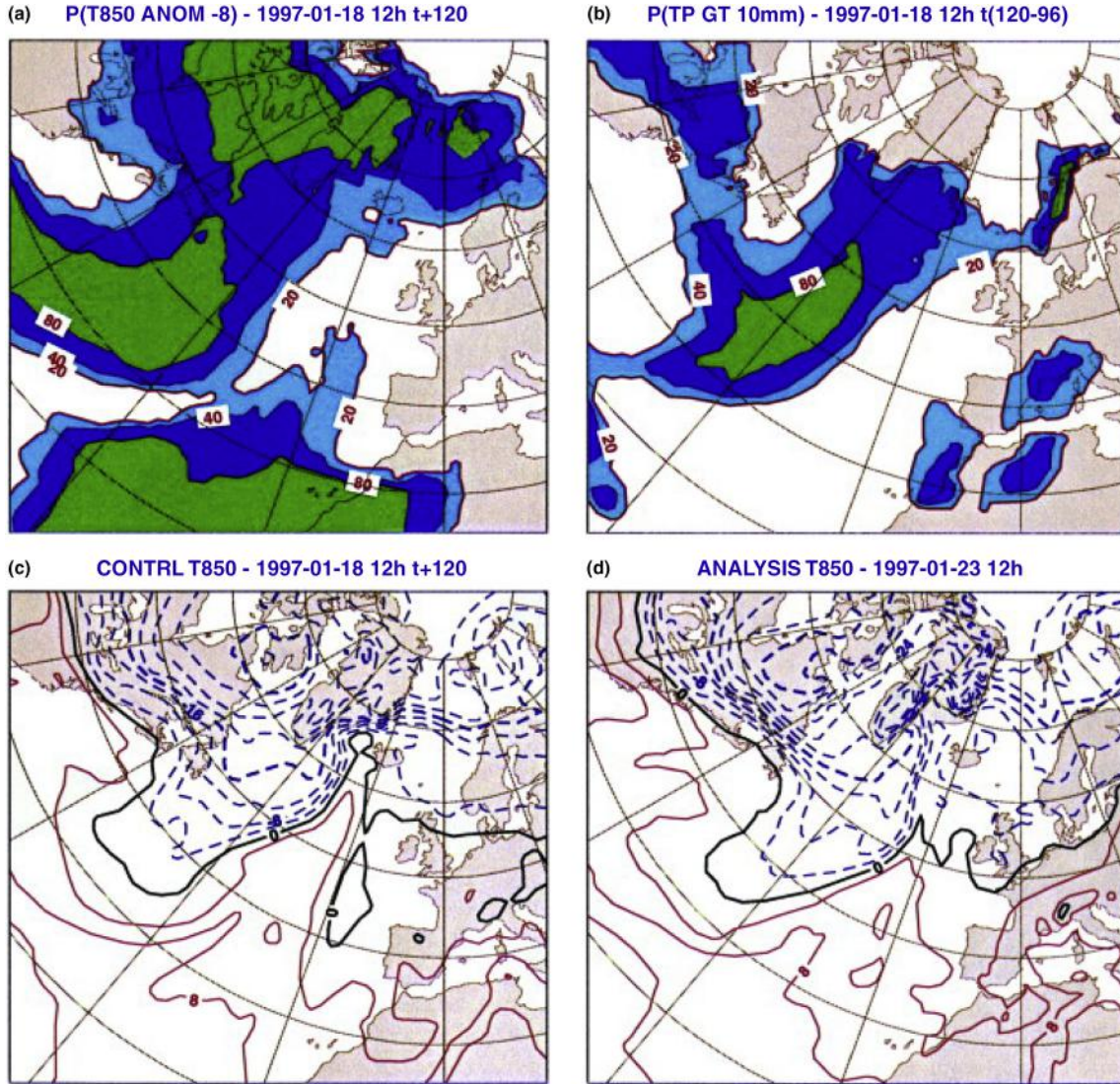


Figure 1 (a) 5-day forecast probability of cold 850 hPa temperatures (850 hPa temperature anomaly with respect to climatology smaller than 8°). (b) 5-day forecast probability of more than 5 mm day^{-1} of precipitation. (c) 5-day 850 hPa temperature forecast given by a single deterministic forecast. (d) Observed temperature field at 850 hPa. Contour isolines for 20%, 40%, and 80% probabilities, with shading for values greater than 20%. Contour interval is 4° for temperatures (dashed blue for negative and solid red for positive values).

The cost function is the sum of the distance $d_B(-, -)$ of $\gamma(t_0)$ from the background field $\gamma_B(t_0)$ and the time-integrated distance $d_O(-, -)$ of $\gamma(\tau)$ and the observation vector $o(\tau)$. (Note that the two distances are different, since one is defined using the covariance matrix of background errors and the other the covariance matrix of observation errors.) The minimum of the cost function can be considered as the best estimate of the true state of the atmosphere. This computational process is referred to as data assimilation. Note that the accuracy of a data assimilation procedure depends on the accuracy of approximations of the atmospheric model used to compute $\gamma(\tau)$ starting from $t_0 - \Delta$ as in eqn [4].

$$\gamma(t) = \gamma(t_0 - \Delta) + \int_{t_0 - \Delta}^t [A(\gamma, \tau) + P(\gamma, \tau)] d\tau \quad [4]$$

Moreover, in order to compare the two vectors γ and o , the atmospheric model is used to transform observed variables into atmospheric state vector variables (i.e., to go from the observations' phase-space to the model variables' phase-space).

Some of the observations, such as those from weather balloons or radiosondes, are taken at specific times at fixed locations. Others, such as those from airplanes, ships, or satellites, are not fixed in space. Generally speaking, there is a great variability in the density of the observation network, with data over oceanic regions, in particular, characterized by very coarse resolution. Observations cannot be used directly to start model integrations, but they must be modified in a dynamically consistent way to obtain a suitable data set.

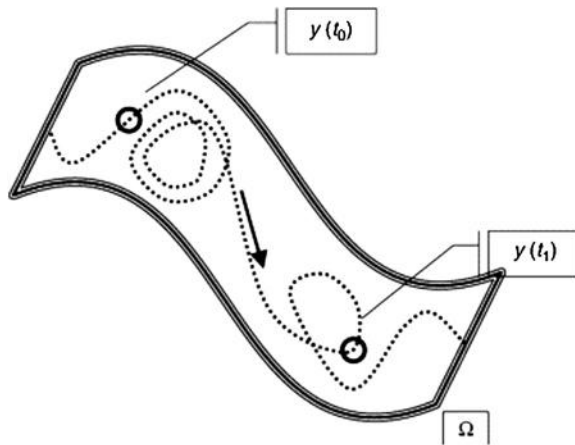


Figure 2 Schematic of a two-dimensional section of the phase space Ω of the atmospheric system. The system attractor is the set of past, present, and future atmospheric states $y(t)$. The time evolution of the system between the initial time t_0 and time t_1 is represented by the dotted line connecting the points $y(t_0)$ and $y(t_1)$ (the orbit of the system between t_0 and t_1).

Sources of Forecast Errors

The facts that at any time t_0 only a limited number of observations (limited with respect to the degrees of freedom of the system) are available and that part of the globe is characterized by a poor coverage introduce uncertainties in the initial conditions. Observational errors, usually in the smallest scales, amplify and through nonlinear interactions spread to the large scales, and eventually affect the skill of the latter. The presence of uncertainties in the initial conditions is one of the sources of forecast errors.

A second source of forecast error is related to the intrinsic approximations made in the numerical models of the atmospheric system. A requirement for skillful predictions is that numerical models are able to accurately simulate at least the effects of the dominant atmospheric phenomena. The facts that the description of some physical processes is only approximate and that numerical models simulate only processes with certain spatial and temporal scales induce forecast errors. (Availability of computer resources is one of the main factors that limits the complexity and the resolution of numerical models and data assimilation procedures, since, to be useful, numerical predictions must be produced in a reasonable amount of time.)

These two sources of forecast errors cause weather forecasts to deteriorate with forecast time. A third source of forecast error that is less important in the short and medium forecast range (say up to 10 days) but that can be very important for longer forecast ranges is related to the system boundary conditions (e.g., soil moisture content, ice coverage, vegetation).

It is worth pointing out that the system's initial conditions will always be known only approximately, since each piece of data is characterized by an error that depends on the instrumental accuracy. In other words, small uncertainties related to the characteristics of the atmospheric observing system will

always characterize the initial conditions. The growth of small initial errors into large forecast errors is due to the chaotic behavior of the atmosphere, which implies that two initial states differing only slightly will depart from one another very rapidly as time progresses.

Chaotic Behavior of the Atmosphere

A dynamical system shows 'chaotic' behavior if orbits exhibit sensitive dependence to initial conditions. An orbit is characterized by sensitive dependence if most other orbits that pass close to it at some point do not remain close to it as time advances. The atmosphere exhibits this behavior.

Figure 3(a) shows a very intense storm that crossed France and Germany during 26 December 1999, and the other three panels of the figure show three 2-day forecasts started from very similar initial conditions at 1200 UTC (coordinated universal time) on 24 December 1999. The differences among the three initial conditions were comparable to estimated analysis errors. After only 2 days of numerical integration, the three forecasts evolved into very different atmospheric situations. In particular, note the different positions of the cyclone forecast over western Germany. The first forecast (**Figure 3(b)**) wrongly positioned the cyclone over Ireland; the second forecast (**Figure 3(c)**) correctly positioned the cyclone over Germany; and the third forecast (**Figure 3(d)**) moved the cyclone too quickly over the Baltic Sea. This is an example of orbits initially close together that rapidly diverge during the time evolution.

Another example of sensitivity to the initial state is shown in **Figure 4**. The figure shows the forecasts for air temperature in London given by 33 different forecasts started from very similar initial conditions for two different dates, 26 June 1995 and 26 June 1994. There are clearly different degrees of divergence among the 33 forecasts during the two cases. All forecasts stay close together up to forecast day 10 for the first case (**Figure 4(a)**), while they all diverge already at forecast day 3 in the second case (**Figure 4(b)**). The level of spread among the different forecasts can be used as a measure of the predictability of the two atmospheric states.

The Ensemble Approach to Numerical Weather Prediction

A complete description of the weather prediction problem can be stated in terms of the time evolution of an appropriate PDF in the atmosphere's phase space (**Figure 5**). The predicted PDF of possible future atmospheric states is a function greater than zero in the phase space regions where the atmospheric state can be, with maximum values identifying the most probable future states.

The problem of the prediction of the PDF can be formulated exactly through the continuity equation for probability (the Liouville equation). Unfortunately, with the current availability of computer power, the Liouville equation that describes the PDF time evolution can only be solved for simple systems characterized by a limited number of degrees of freedom. Furthermore, the initial PDF is not well known. Ensemble prediction based on a finite number of deterministic

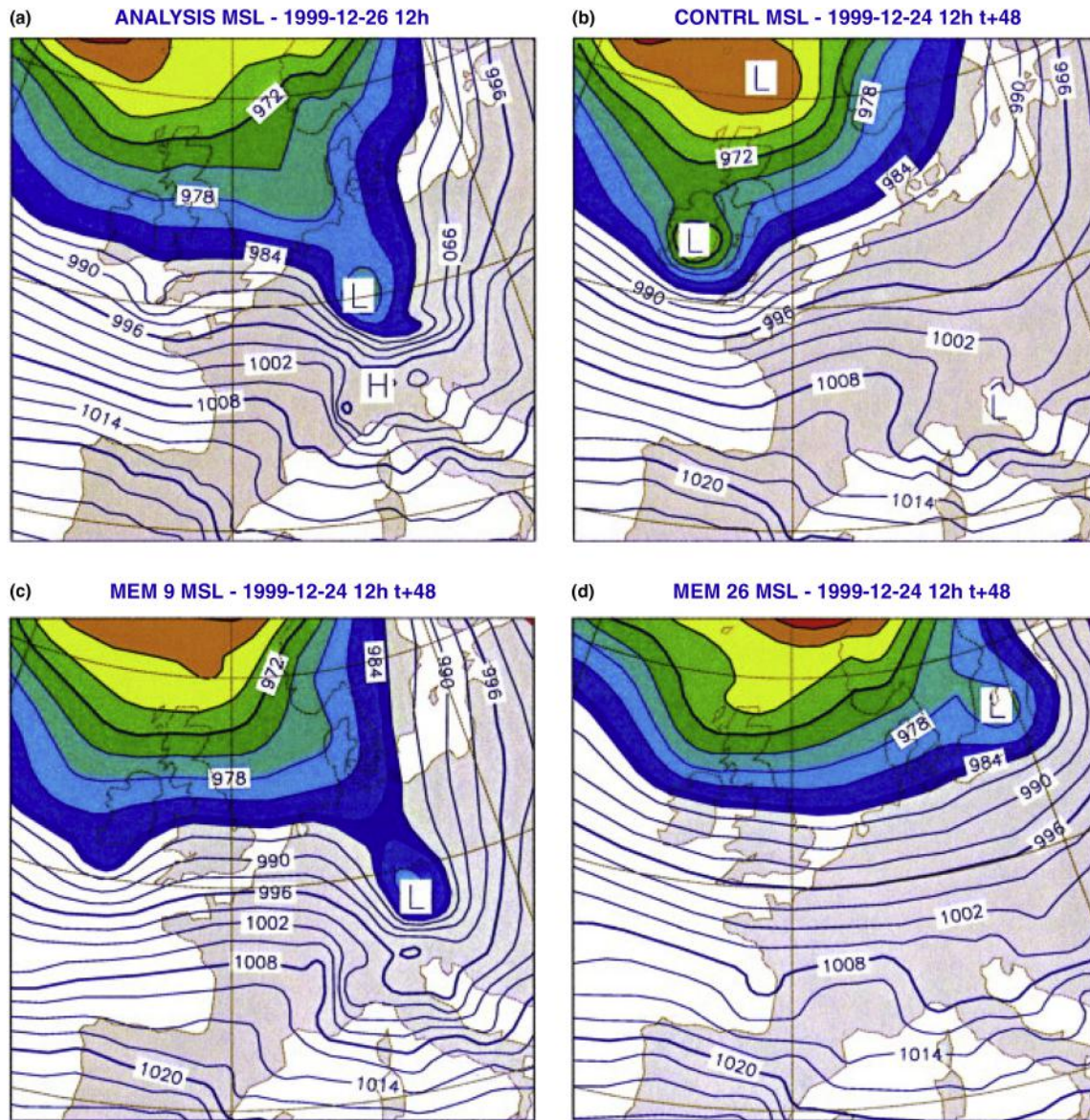


Figure 3 (a) Mean sea-level pressure field at 1200 UTC on 26 December 1999 and (b–d) 2-day mean sea-level pressure forecasts started at 1200 UTC on 24 December 1999 given by three members of the ECMWF ensemble prediction system. Contour interval is 3 hPa, with shading for values lower than 984 hPa.

integrations is a feasible method for predicting the PDF beyond the range of linear error growth. One of the byproducts of ensemble prediction is the possibility of estimating the forecast skill of a deterministic forecast – in other words, to forecast the forecast skill.

Ensemble prediction systems should be designed to simulate all sources of forecast errors, in particular errors due to initial and model uncertainties. The relative importance of these two sources of forecast errors depends on the characteristics (e.g., spatial and temporal scales) of the phenomena under investigation. For large-scale atmospheric patterns, research studies performed with state-of-the-art numerical models have indicated that for short and medium time ranges (say, forecasts for up to 3–5 days) errors are mainly

due to initial uncertainties. Model errors due to the parameterized physical processes start to have a nonnegligible effect after forecast day 3. By contrast, for the prediction of small-scale low-pressure systems and associated precipitation fields, for example, model errors can be as important as initial uncertainties at forecast day 2 or even earlier. For forecast times longer than 10 days (monthly and seasonal prediction), other error sources should be simulated. Examples that should be taken into account include the possible influence of uncertainties in the boundary conditions (e.g., in the soil moisture content or in the ice and vegetation coverage).

At the time of writing (September 2002), global medium-range (i.e. forecasts for up to 15 days) ensemble systems are

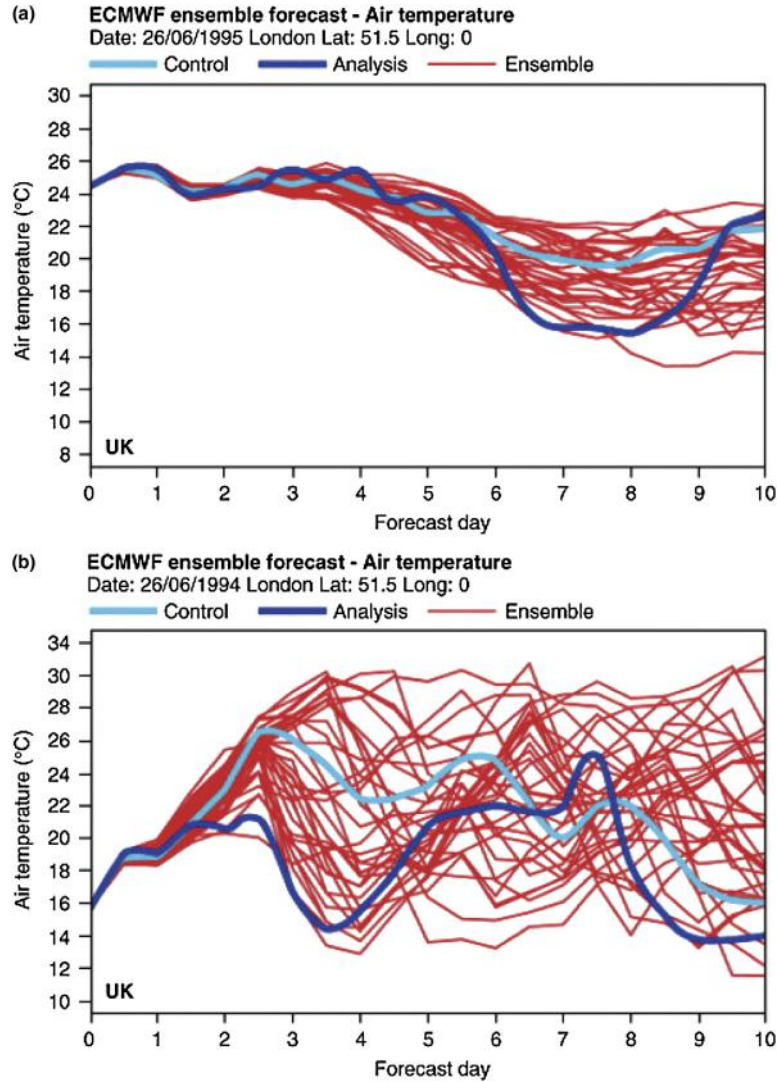


Figure 4 Forecasts given by the ECMWF operational ensemble prediction system (33 members at the time of running) for air temperature in London started from (a) 26 June 1995 and (b) 26 June 1994. Courtesy of Thomas Petroligis, 1995, personal communication.

part of the routine operational suites at the Meteorological Center of Canada (MSC, Dorval, Canada, <http://www.msc-smc.ec.gc.ca>), at the European Centre for Medium Range Weather Forecasts (ECMWF, Reading UK, <http://www.ecmwf.int>), and at the National Centers for Environmental Prediction (NCEP, Washington DC, USA, <http://www.emc.ncep.noaa.gov>). At these three centers global ensemble systems are run daily, and probabilistic forecasts are generated and delivered to their users. Beside this operational activity, many universities and other national and regional meteorological centers have been involved in research and experimental activities in this field.

The three ensemble systems operational at MSC, ECMWF, and NCEP are all based on a finite number of numerical integrations starting from perturbed initial conditions, but they differ in the way the perturbed initial conditions are constructed. The ensemble systems operational at MSC and

ECMWF include different schemes to simulate model errors, while the NCEP system does not simulate them. Furthermore, they are different in their ensemble size, resolution, and forecast length of the numerical integration (Table 1). (It should be stressed that availability of computer power is the main factor affecting configuration parameters such as ensemble size and resolution.)

Despite their differences, schematically each member e_j (with $j = 1, \dots, N_{\text{ens}}$) of any of these operational ensembles is defined by the time integration (eqn [5]) of the perturbed model equations (eqn [6]), starting from perturbed initial conditions $e_j(t_0)$.

$$e_j(t) = e_j(t_0) + \int_{t_0}^t [A(e_j, \tau) + P_j(e_j, \tau)] d\tau \quad [5]$$

$$\frac{\partial e_j}{\partial t} = A(e_j, t) + P_j(e_j, t) \quad [6]$$

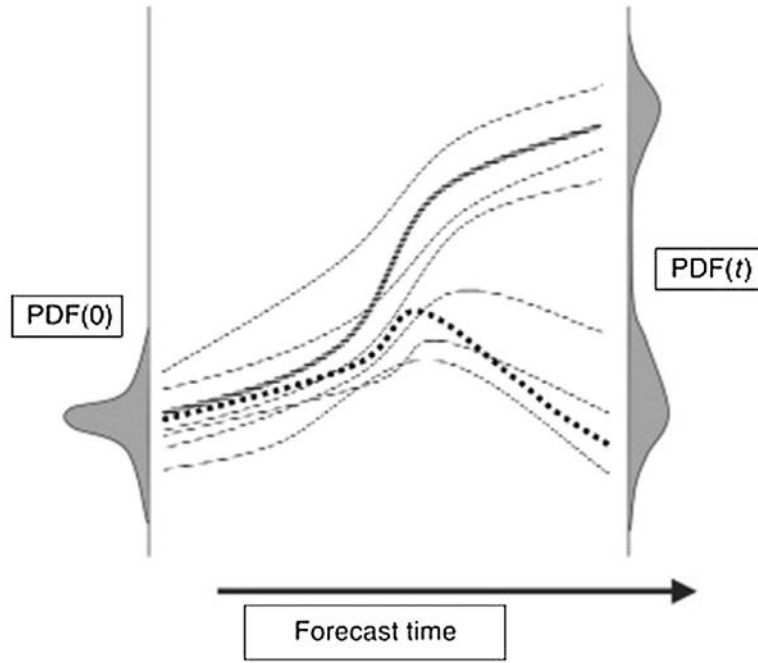


Figure 5 Schematic of ensemble prediction. The initial $PDF(0)$ represents the initial uncertainties. From the best estimate of the initial state, a single deterministic forecast (bold solid curve) is performed. This single deterministic forecast fails to predict correctly the future state (dash curve). An ensemble of perturbed forecasts (thin solid curves) starting from perturbed initial conditions designed to sample the initial uncertainties can be used to estimate the probability of future states $PDF(t)$. In this case, two perturbed forecasts almost correctly predict the future state, and the ensemble system gives a nonzero probability of the future state to reach the observed value.

Table 1 Configuration of the global ensemble systems operational at MSC, ECMWF and NCEP (at the time of writing, September 2002): ensemble size, spectra truncation (TL149 indicates spectral triangular truncation at wavenumber 149 with linear grid), equivalent grid spacing in km, number of vertical levels (note that MSC runs with a different number of vertical levels in each perturbed forecasts), pressure at the top of the model (hPa), forecast length (days), inclusions of the simulation of observation errors, initial uncertainties and model uncertainties (Yes or No)

Ensemble	Size	Resolution (spectral)	Grid (km)	Levels	Top (hPa)	Forecast length	OBS	IC	MOD
MSC	16	TL149	135	23-28-41	10	10d	Y	Y	Y
ECMWF	51	TL255	80	40	10	10d	N	Y	Y
NCEP	24	T126 ($d < 3.5$)– T62 ($3.5 < d$)	120 ($d < 3.5$)– 250 ($3.5 < d$)	28	2.7	15d	N	Y	N

Note that in eqn [6] the term that identifies the contribution to the full equation tendency of the parameterized physical processes, $P_j(y, t)$, is different for each ensemble member. This represents the fact that model errors due to parameterized physical processes are simulated in the ensemble system.

Simulation of Initial Uncertainties

The perturbed initial conditions $e_j(t_0)$ are generated to represent the initial uncertainties. This can be accomplished by following different approaches with the constraint that the number of initial perturbations is limited to a few tens. In probabilistic terms, this is equivalent to saying that the initial-time PDF can be sampled only a small number of times.

At ECMWF, the perturbed initial conditions are defined by adding to the best estimate of the initial state $e_0(t_0)$ (computed

by minimizing $J(-, -)$, see eqn. [2]), an ensemble of initial perturbations $\delta e_j(t_0)$ (eqn [7]).

$$e_j(t_0) = e_0(t_0) + \delta e_j(t_0) \quad [7]$$

The initial perturbations $\delta e_j(t_0)$ have been designed to sample the components of the initial uncertainties with maximum growth during the forecast time (i.e., for times $t > t_0$) and during the data assimilation time, (i.e., for times $t < t_0$). This choice is based on the hypothesis that the components of the initial uncertainties with the fastest growth during the system time evolution and during the data-assimilation time period have the dominant effect on the forecast accuracy. Practically, these ‘growing directions’, called singular vectors, are computed by applying complex techniques based on iterative integrations of forward and adjoint tangent versions of the nonlinear model.

At NCEP, as at ECMWF, initial perturbations are added to the best estimate of the initial state $e_0(t_0)$. But compared to ECMWF, the NCEP initial perturbations are designed to sample only the components of the initial uncertainties that grow during the data-assimilation time interval, i.e., for times $t < t_0$. As a consequence, the two sets of initial perturbations can differ substantially. For example, ECMWF initial perturbations are characterized by smaller scales and by faster growth during the first two days of forecast integration than are the NCEP perturbations.

At MSC, a different approach is followed. First, given the set of observations $o(t)$, an ensemble of perturbed observations $o_j(t)$ is generated, $O_j(t) = O(t) + \delta O_j$. Then, the ensemble of N_{ens} initial perturbations is generated by performing N_{ens} data assimilation cycles (i.e., by minimizing eqn [3] j times, each time using a different set of perturbed observations $o_j(t)$), so that each perturbed initial state $e_j(t_0)$ is the best estimate of the atmospheric state given by the set of perturbed observations $o_j(t)$. It should be mentioned that, since the MSC ensemble also simulates model errors, each data assimilation procedure is performed using a different model version (see next section).

The differences between the three methods seem subtle, but they have a significant impact on the characteristics of the ensemble systems (e.g., on the ensemble dispersion). The debate on what is the best approach is still open.

Simulation of Model Errors

As is the case for the simulation of initial uncertainties, there is not yet agreement within the scientific community about the best way to simulate model errors.

At MSC, different – and on average equally skillful – parameterization schemes are used when numerically integrating each perturbed member. The rationale of this approach is that, despite the fact that these different schemes perform equally on average, they can perform significantly differently on single occasions. Schematically, the MSC approach can be described as follows. Suppose that the following different parameterization schemes are given: three schemes that simulate moist processes (C_1 , C_2 , and C_3), three schemes that simulate turbulent diffusion processes (D_1 , D_2 , and D_3) and two radiation schemes (R_1 and R_2). Using a different combination of the schemes, an ensemble system with 18 perturbed members, each integrated with a different model, can be designed. Schematically, each member of the MSC ensemble is defined by eqn [5] with eqn [8].

$$P_j = C_{k,j} + D_{k,j} + R_{k,j} \quad [8]$$

The full tendency due to the parameterized physical processes is computed by adding the tendencies computed by different combinations of the schemes simulating moist processes, turbulent diffusion, and radiation ($C_{k,j}$ means that the j th ensemble member is integrated using the C_k).

At ECMWF, the random component of model errors due to parameterized physical processes is simulated by stochastically perturbing the tendency due to the physical processes. Schematically, each member of the ECMWF ensemble is defined by eqn [5] with eqn [9], where P is the unperturbed tendency due

to all the parameterized physical processes, and r_j is a vector of random numbers.

$$P_j(e_{j,t}) = [1 + r_j]P(e_{j,t}) \quad [9]$$

At the time of writing, in the ECMWF operational ensemble system the random numbers r_j are uniformly sampled in the interval $-0.5 < r_j < 0.5$.

Operational Applications of Ensemble Prediction

Ensemble prediction products are becoming increasingly popular. Figure 6 and Figure 7 show two examples of weather products designed, respectively, for a forecaster interested only in a single location and for a forecaster interested in a large area. Figure 6 shows the 10-day ensemble prediction of temperature, geopotential height, precipitation, and cloud cover for London given by the ECMWF ensemble prediction system started on 28 March 2000. This shows, for example, that wind speed will decrease during the first three forecast days, that temperature will rise slightly, and that there could be intense precipitation (up to 20 mm/12 h) on 2 April. The fact that the size of the box-and-whiskers increases during the forecast time, especially after 1 April, indicates that the ensemble spread increases, suggesting future less predictable situations.

The second product has been designed for a forecaster interested in assessing the accuracy of an 84 h prediction of mean sea-level pressure (MSLP) and precipitation over north-eastern Spain. Figure 7(a) shows the observed MSLP (more precisely the ECMWF analysis) and the 24 h precipitation forecast given by the ECMWF high-resolution model, which can be considered as a good approximation of the large-scale features of the actual precipitation field. This panel shows that more than 40 mm of rain affected a large area across the Spanish–French border. Figure 7(b) shows the 84 h forecast given by the ECMWF unperturbed (control) forecast. Figure 7(c) shows the 84 h MSLP forecast given by the ECMWF ensemble-mean and the ensemble standard deviation (shading). Figure 7(d) shows the 84 h probability forecast of more than 20 mm day⁻¹ of precipitation given by the ECMWF ensemble system. Figure 7(c) and (d) can be used to assess the possible accuracy of the forecast given by the control forecast (d). Figure 7(c) shows that the ensemble-mean forecast predicts an atmospheric flow (at the surface) similar to that predicted by the control forecast, and that the ensemble standard deviation, which is a measure of the agreement/disagreement among the ensemble members, is not particularly large in the area of interest (i.e., across the Spanish–French border). This can be used as an indication that the ensemble spread in the area of interest is rather small and thus that the situation should be predictable. Figure 7(d) shows that there is a probability of more than 40% that intense precipitation will affect the area of interest, again confirming that there is good agreement among the ensemble members in predicting this phenomenon.

These two figures constitute simple examples of how ensemble prediction can be used to complement single deterministic forecasts with probabilistic information to try to assess the forecast accuracy. It is worth mentioning that ensemble

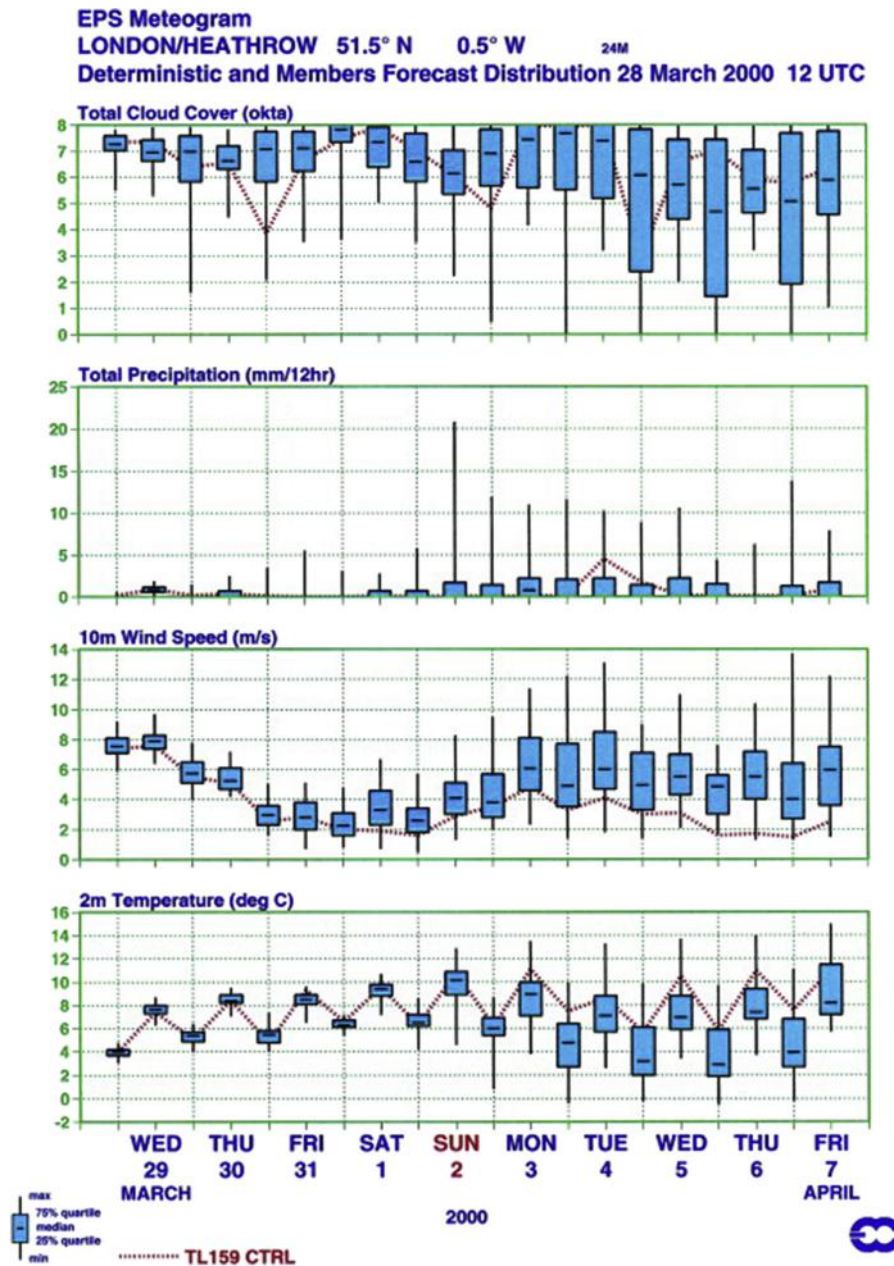


Figure 6 Meteogram for London showing the ECMWF ensemble prediction forecasts for cloud cover, total precipitation, 10-m wind-speed, and 2-m temperature for London/Heathrow. The forecast started at 1200 UTC on 28 March 2000. Box-and-whiskers show the minimum and maximum values, the 25% and the 75% quartiles, and the median of the forecasts given by the 51-member ECMWF ensemble system.

products have started to be used experimentally in business applications such as ship routing and energy-demand predictions.

Validation of Ensemble Systems

The primary purpose of ensemble systems is to estimate the PDF of forecast states. As a consequence, the quantitative evaluation of ensemble systems should be based on the comparison of the forecast PDF with the observed PDF. In

particular, verification measures should be designed to assess the statistical consistency and usefulness of the predicted PDF. These two properties should be assessed by considering the first-order moment of the predicted PDF (the ensemble-mean) and the second-order moment (ensemble standard deviation). Verifications should be performed both at grid-point level (i.e., for single locations) and considering large-scale atmospheric patterns (i.e., regimes characterized by a blocked or a zonal flow over the Euro-Atlantic sector). The skill of an ensemble system should be compared with the skill of some reference

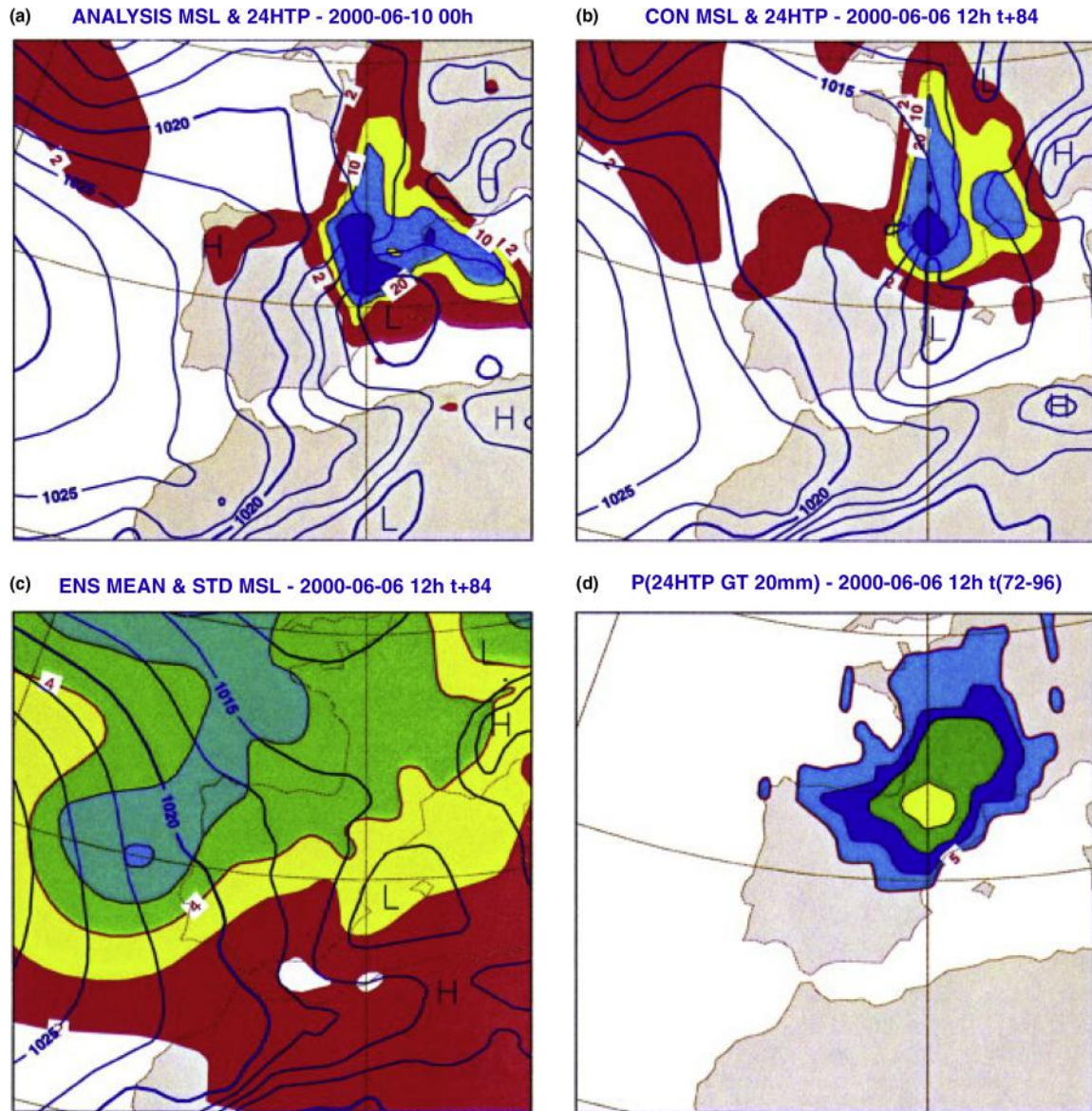


Figure 7 (a) Mean sea-level pressure (MSLP) field observed (i.e., analysis) at 0000 UTC on 10 June 2000 and verification for the 24 h accumulated precipitation (24HTP) between 1200 UTC of 9 and 10 June (this verification field, defined as the 24 h forecast from the ECMWF high-resolution model started on 9 June, can be considered a good approximation of the large-scale features of the observed precipitation field). (b) ECMWF ensemble unperturbed (i.e., control) forecast started on 6 June ($t+84$ h for MSLP and 24HTP predicted between $t+72$ h and $t+96$ h). (c) ECMWF ensemble-mean and standard deviation $t+84$ h forecasts for MSLP. (d) Corresponding ECMWF ensemble probability of more than 20 mm day⁻¹ of precipitation. Contour interval is 2.5 hPa for MSLP and 1 hPa for MSLP standard deviation. Shading for the 24HTP in (a) and (b) is for 2, 10, 20, and 40 mm day⁻¹ and shading for probability in (c) is for 5%, 10%, 20%, and 40%.

systems (benchmarks) such as climatology or simpler PDF-prediction systems.

Numerical weather forecasts are often used by decision-makers to decide whether to take an action to protect against a possible loss. Typically, the decision maker would spend an amount C if an event is predicted to protect himself against a loss L (with $L > C$). The economic value of a forecasting system can be assessed using skill measured by coupling contingency tables and cost-loss decisions. The economic value of a forecast can be defined as a function of the false alarm rate and the probability of detection of the

system. Since this measure is defined for both categorical and probabilistic forecast, it can be used to compare the economic value of a single forecast and of an ensemble forecasting system.

Figure 8 shows, for December 1999 over Europe, the average economic value of the $t + 120$ h $t + 120$ h prediction the event, '24 h accumulated precipitation greater than 10 mm day⁻¹', for three different forecasts, specifically the single deterministic forecasts given by the ECMWF unperturbed (control) forecast and by the ensemble-mean, and the probabilistic prediction generated using the whole ensemble system

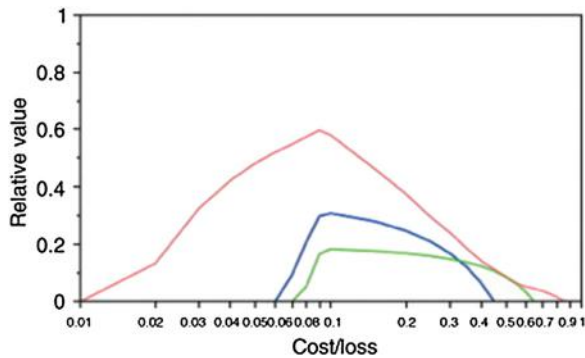


Figure 8 Comparison of the economic value of the ECMWF $t+120$ h forecasts given by the unperturbed (control) forecast (blue line), the ECMWF ensemble-mean (green line), and the ECMWF ensemble probability (red line), average for December 1999 over Europe, for the event '24 h total precipitation larger than 10 mm day⁻¹'.

(the economic value varies from 0, for a forecast as skillful as climatology, to 1 for a perfect forecast).

Figure 8 shows that the economic value for the ensemble probabilistic prediction is definitely greater than the economic value of the two probabilistic predictions, especially for small cost/loss ratios. Considering the deterministic predictions, Figure 8 shows that (for the period and the event considered) the ensemble-mean and the control forecast have about the same economic value for the 5 mm day⁻¹ threshold, while for the 10 mm day⁻¹ threshold the control forecast performs better for small cost/loss ratios but worse for large values.

This result indicates that for December 1999 a decision maker interested in predicting a binary event 'rainfall greater than 5 or 10 mm day⁻¹' over Europe would have had a higher return if by taking decisions (protect/not protect) according to the ensemble forecast than from any single deterministic forecast.

Conclusions

The operational implementation of ensemble prediction systems has changed the approach to numerical weather prediction from deterministic (i.e., based on a single forecast) to probabilistic. Ensemble systems provide a possible way to estimate the probability distribution function of forecast states. They have been developed following the notion that

uncertainties in the initial conditions and in the model formulation are the main sources of forecast errors.

Ensemble prediction systems are particularly useful, if not necessary, for providing early warnings of the risk of extreme weather events. For example, ensemble systems can be used to predict probabilities of intense precipitation events. The economic value of ensemble prediction is higher than the economic value of single deterministic forecasts. Global ensemble systems can be used to provide boundary and initial conditions for higher-resolution, limited-area ensemble prediction systems.

Research activity is in progress in many different areas to further improve the accuracy of the operational ensemble prediction system, and is in hand to improve the representation of initial and model errors, and possibly to use ensemble approaches to data assimilation. Multimodel, multianalysis ensemble systems based on a set of integrations performed with different models and starting from analyses constructed using different data assimilation schemes are under investigation. Work is also under way to further develop ensemble products that can be used more easily in businesses heavily affected by weather (e.g., the shipping industry, the energy sector, safety and protection agencies).

See also: **Climate and Climate Change:** Climate Variability: Seasonal and Interannual Variability. **Data Assimilation and Predictability:** Data Assimilation; Predictability and Chaos. **Mesoscale Meteorology:** Overview. **Numerical Models:** Methods; Regional Prediction Models. **Weather Forecasting:** Operational Meteorology; Seasonal and Interannual Weather Prediction; Severe Weather Forecasting.

Further Reading

- ECMWF Workshop Proceedings, 1992. New Developments in Predictability, 13–15 November 1991. Reading: ECMWF.
- ECMWF Workshop Proceedings, 1999. Workshop on Predictability, 20–22 October 1997. Reading: ECMWF.
- Haltiner, G.J., Williams, R.T., 1979. Numerical Prediction and Dynamic Meteorology. Wiley, Chichester.
- Holton, J.R., 1982. An Introduction to Dynamic Meteorology. Academic Press, London.
- Lorenz, E.N., 1993. The Essence of Chaos. University College London Press, London.
- Richardson, L.F., 1922. Weather Prediction by Numerical Process. Cambridge University Press, Cambridge. (Reprinted, New York: Dover.).
- Wilks, D.S., 1995. Statistical Methods in the Atmospheric Sciences. Academic Press, London.

Predictability and Chaos

LA Smith, London School of Economics, Centre for the Analysis of Time Series, London, UK

© 2003 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 4, pp 1777–1785, © 2003, Elsevier Ltd.

Introduction

Chaos is difficult to quantify. The nonlinear dynamic that gives rise to chaos links forecasting on the shortest time scales of interest with behavior over the longest time scales. In addition, the statistics which quantify chaos forbid an appeal to most of the traditional simplifications employed in statistical estimation. Nevertheless, both the concept of chaos in physical systems and its technical relative in mathematics have had a significant impact on meteorological aims and methods. This is particularly true in forecasting. Chaos implies sensitivity to initial conditions: small uncertainties in the current state of the system will grow exponentially-on-average. Yet, as discussed below, neither this exponential-on-average growth nor the Lyapunov exponents that quantify it reflect macroscopic predictability. The limits chaos places on predictability are much less severe than generally imagined. Predictability is more clearly quantified through traditional statistics, like uncertainty doubling times. These statistics will vary from day to day, depending on the current state of the atmosphere. Maintaining the uncertainty in the initial state within the forecast is a central goal of ensemble forecasting (*see Data Assimilation and Predictability: Ensemble Prediction*). Achieving the ultimate goal of meaningful probability forecasts for meteorological variables would be of great societal and economic value. Fundamental limitations in the realism of models of the atmosphere will limit our ability to make probability (PDF) forecasts, just as uncertainty in the initial condition limits the utility of single forecasts even if the model is perfect.

Initially, it appears that chaotic systems will be unpredictable, and this is true in that it is not possible to make extremely accurate forecasts in the very distant future. Yet chaos *per se* does not imply one cannot sometimes make accurate forecasts well into the medium range. And perhaps just as importantly, with a perfect model one can determine which of these forecasts will be informative and which will not at the time they are made. As we shall see below, both the American and the European weather forecast centers have adopted this strategy operationally, with the aim of quantifying day-to-day variations in the likely range of future meteorological variables. Quantifying this range can be of significant value even without a perfect model. Although accurate probability forecasts are likely to require a perfect model, current operational ensemble prediction systems already provide economically valuable information on the uncertainty of numerical weather prediction (NWP) well into 'week two,' and research programs on seasonal time scales are underway.

Why is perfect foresight of the future state of the atmosphere impossible? First, it should be no surprise that if our knowledge of the present is uncertain then our knowledge of the future will also be uncertain; the question of prediction then turns to how to best quantify the dynamics of uncertainty. Here the full

implications of chaos, or more properly of nonlinearity, mix what are often operationally distinct tasks: observing strategy, data assimilation, state estimation, ensemble formation, forecast evaluation, and model improvement. This complicates the forecasting problem beyond the limits in which classical prediction theory was developed. First the traditional goal of a single best first guess (BFG) forecast with a minimal least-squares error is not a viable aim in this scenario, nor is a least squares solution desired. Indeed (as discussed below) this approach would reject the perfect model which generated the observations! Second, current models are not perfect. The term 'model inadequacy' is used to summarize imperfections in a given model and the entire model class from which that model is a member. Model improvement and the search for a superior model class, along with the investigation of more relevant measures of model skill, are areas of active research.

A Mathematical Framework for Modelling Dynamical Systems

Chaos is a phenomenon found in many nonlinear mathematical models. While one should never forget the distinction between the model and the system being modeled, precise mathematical definitions are more easily made within the perfect model scenario (PMS). Within this useful fiction, one assumes that the model in-hand is itself the physical system of interest. Before moving back to forecasts of the real world, of course, one must recover from this self-deception.

Given a model, an initial condition is simply an assignment of values to all model variables at a particular starting time. Thus the initial condition reflects the current state of the model: it is a vector $\mathbf{x}(t)$ which specifies the value of every variable in the model at time t . For the classical model of a pendulum, the state consists of two variables (the angle and the angular velocity). These two numbers completely define the current state of the model, and so the model is called two-dimensional, or, equivalently, said to have a two-dimensional state space. The famous Lorenz model of 1963 is three-dimensional, as there are three variables: $\mathbf{x} = (x, y, z)$. These low-dimensional systems should be contrasted with the state of an operational NWP model, which may consist of over 10 million variables.

The sequence of states a dynamical system passes through defines the history of the system; this sequence is called a trajectory. For deterministic systems, any single state $\mathbf{x}$ along a trajectory defines all future states of the system. For the classical pendulum, these solutions can be written down analytically; but for all but the simplest nonlinear systems only numerical solutions exist. Thus it is difficult to prove that a given realization of the Lorenz model is chaotic, and even more so modern NWP models.

Chaos

Mathematically, a chaotic system is a deterministic system in which (infinitesimally) small uncertainties in the initial condition will grow, on average, exponentially fast. The average is taken over (infinitely) long periods of time. Of course, finite uncertainties often grow rapidly as well, in which case any uncertainty in the initial condition will limit predictability in terms of a single BFG of the future state, even with a perfect model. But inasmuch as it is defined by the behaviour of (infinitesimally) small uncertainties of (infinitely) long periods of time, chaos *per se* places no limits of the growth of finite uncertainty over a finite period of time. Chaotic systems are often said to display sensitive dependence on initial condition (SDIC), a technical term for systems in which states initially very close together tend to end up very far apart, eventually.

Suppose the true state of the system is $\bar{x}$: what is the behaviour of a near-by solution $\hat{x}$ where $\hat{x} = \bar{x} + \epsilon$? In the pendulum, a small initial ϵ grows slowly, if at all. In a chaotic system, the magnitude of ϵ will grow exponentially-on-average; yet this does not imply that the actual magnitude of ϵ ever grows exponentially in time. Indeed, since ϵ is a distance and $|\epsilon| > 0$ for any given value of t , one can always define a value

$$\lambda = \frac{1}{t} \log[|\epsilon(t)|/|\epsilon(0)|]$$

for any system, chaotic or not! In this case, observing a value of $\lambda > 0$ for finite t does not even suggest exponential growth. Statistics like λ become interesting only when they approach a constant as $t \rightarrow \infty$; by definition, chaos reflects properties only in this limit.

Clearly, chaos includes special cases where magnitude of ϵ is growing uniformly, say doubling every second; but it also allows the more common case where the growth of ϵ is a function of the state x and hence changes with time. In general, the growth will not be uniform in time. In fact in some chaotic systems, including the Lorenz 1963 model, there are regions of the state space in which every ϵ will decrease! Such regions are said to represent ‘return of skill’ as forecasts become more accurate in the least-squares sense as time passes.

For instance, consider the case where, on average, half the time ϵ is constant and the other half of the time it grows by a factor of four. This will yield in the same exponential-on-average growth as doubling every time step, yet there will be times when prediction is easy. Or for a more extreme case, consider where ϵ shrinks by a factor of two 9 times out of 10, but once in 10 grows by a factor of 2^{19} (about half a million); again this is exponential-on-average growth equivalent to doubling every time step. The question, then, is whether or not these variations of predictability can be identified in advance. As discussed below, ensemble forecasts aim to do just that.

The Statistics of Chaos: Lyapunov Exponents and Doubling Times

Given a deterministic system which remains in a bounded region of state space, chaos is defined by a statistic called the Lyapunov exponent. In a one-dimensional system, the Lyapunov exponent reflects the logarithm of the (geometric) average

growth of infinitesimal uncertainties. In the limits $\epsilon \rightarrow 0$ and $t \rightarrow \infty$, the geometric average of $|\epsilon(t+1)|/|\epsilon(t)|$ defines the Lyapunov exponent, usually called λ . Note that nowhere is there the suggestion that $|\epsilon(t)| \approx |\epsilon(0)|e^{\lambda t}$. A system will have as many Lyapunov exponents as the dimension of the state space. The largest Lyapunov exponent is often called the ‘leading’ Lyapunov exponent, and if the leading Lyapunov exponent of a bounded deterministic system is positive, then the system is ‘chaotic.’ Hence the three systems noted above – one in which ϵ doubles every time step, and the other two in which ϵ sometimes grows and sometimes does not – each have the same Lyapunov exponent. Typically, the logarithm is taken with base two, so that if, for example, the uncertainty doubles every second, the Lyapunov exponent is one bit per second, thus the Lyapunov exponent is an average rate.

Also note that for every state x on the attractor, there corresponds a unique direction in state space associated with the leading Lyapunov exponent. If the state space has a dimension greater than one, then estimating Lyapunov exponents involves matrix multiplication along a trajectory. Matrix multiplication does not commute, thus when dealing with statistics like Lyapunov exponents one has to apply multiplicative ergodic theorems; this makes many of their properties appear counterintuitive. Many intuitive methods of statistical estimation fail when applied to chaotic systems. None of this is surprising, since most statistical intuition is developed in the context of more familiar ergodic theorems.

If the sum of all the Lyapunov exponents is negative, then the trajectories will evolve towards a set whose dimension is less than the dimension of the state space; this set is called an ‘attractor.’ An attractor may be something as common as a fixed point, a periodic orbit, or a torus; in such cases the attractor has simple geometry. Alternatively, an attractor may have a strange geometry: it may consist of a fractal set of points in state space, in which case it is called a ‘strange attractor.’ Note that being chaotic reflects a property of the dynamics of the system, while strangeness reflects the geometry of the set on which the system evolves, not the dynamics of the evolution itself. Given a chaotic system with a strange attractor, the choice of initial conditions should reflect the local structure of the attractor, yet this structure is determined by the long-time behavior for the system. In this way, chaos links the longest time scales of the system to the shortest time scales of interest.

The attractor of the Lorenz 1963 model with typical parameter values is shown in [Figure 1](#). It is believed that there are parameter values for which the Lorenz 1963 model has chaotic dynamics on a strange attractor, but, as noted above, such properties are difficult to prove even in this fairly simple system of equations.

In meteorology, the doubling time, τ_2 , provides a more traditional measure of predictability than the Lyapunov exponents. It is also a more relevant measure, although if the growth is not uniform then the time required for an initial uncertainty at x to increase by a factor of four will not be twice the doubling time at x (or, more generally, $\tau_{q^2} \neq 2\tau_q$). In practice, one just has to look at the statistics. An average time is a more relevant measure of predictability than an average rate. Computing an average rate requires stating the duration over which to average *a priori*, while the relevant time scale is itself the quantity of interest. In fact, one can generate a family of chaotic systems

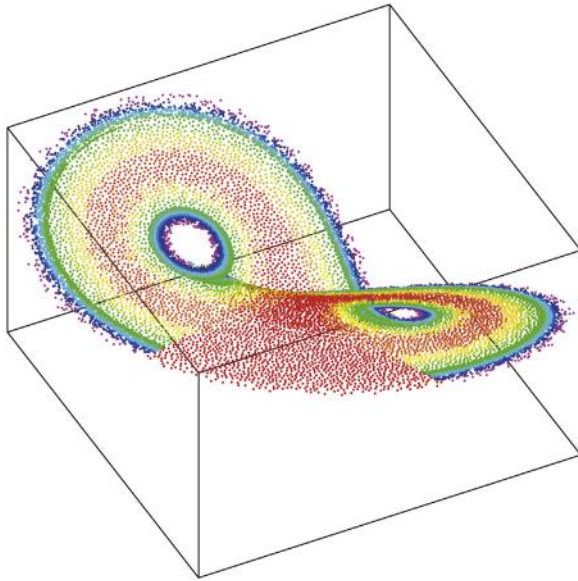


Figure 1 The distribution of uncertainty doubling times on the Lorenz attractor. Points colored red double in less than one Lorenz second. Points colored red have a $\tau_2 \leq 1$ Lorenz time step, orange points $\tau_2 \leq 2$, and so on through yellow, light green, dark green, blue, and purple. The mauve points on the inner and outer edges of the attractor for which $\tau_2 \leq 7$. The density of points with $\tau_2 \leq 5$ has been reduced for clarity. The visible line in the foreground which separates red points from each of the other colors shows the location of points which double just as they enter the region in which all uncertainties decrease, referred to in the text. Adapted with permission from **Figure 1** of Smith LA (1994) Local optimal prediction. *Philosophical Transactions of the Royal Society of London Series A* **348**: 371–381.

each of which has a leading Lyapunov exponent greater than one, yet containing members with an average τ_2 as large as desired! Indeed, estimating a time-like statistic with the inverse of an average rate is a dubious endeavor. To convince yourself of this, consider estimating the average value of a by one over the average of $1/a$, when a is uniformly distributed between zero and one.

In **Figure 1**, points on the Lorenz attractor are colored by the doubling time of an infinitesimal uncertainty aligned initially in the local orientation corresponding to the leading Lyapunov exponent. The coloration is neither uniform nor random. Note the line separating the red points on one side from the band of each color on the other. The origin of this demarcation will be explained in the next section. Red points have a τ_2 of less than one second, for orange points it is less than two, and so on through yellow, light green, dark green, blue, and purple. The mauve points on the inner and outer edges of the attractor for which $\tau_2 > 7$. This is a clear illustration that predictability will vary with initial condition in an organized way! Which, in turn, suggests that predictability will vary in a predictable way: quantifying this in practice is the goal of ensemble forecasting. Yet even within the PMS, one is interested in finite initial uncertainties and forecasts over a finite duration. The accuracy of such forecasts need not reflect the Lyapunov exponents of the system in any way. Thus chaos *per se* places few restrictions on predictability.

True Limits of Predictability

So what are the limits to predictability of a chaotic system? The answer depends on the use to which the forecast is to be put.

Linear prediction theory aims to identify the optimal least-squares predictor: the model which, on average, yields a BFG future state with the smallest (squared) prediction error. This is a coherent approach to Gaussian uncertainties evolved under linear models, but not when applied to nonlinear systems with uncertain initial conditions.

If the initial condition is uncertain, then this uncertainty will evolve nonlinearly. It can be proven that given a series of uncertain observations of a chaotic system, there will always be uncertainty in the current state. This is the case even if a perfect model is in hand and the observations extend into the infinite past. Even then, there will be a set of indistinguishable states, each consistent with the series of observations and with the long-time dynamics of the system. The ideal forecast is then an ensemble forecast, where the members of the ensemble are drawn from the set of indistinguishable states, and each member weighted with its likelihood given the available observations. In the limit of infinitely large ensembles, this forecast can accurately quantify the relative probability of different events and the decay of predictability, correctly reflect the variations in each from day to day.

In practice, ensemble forecasting is a Monte Carlo approach to estimating the probability density function (PDF) of future model states given uncertain initial conditions. An ensemble forecast for the Lorenz 1963 system is shown in **Figure 2**. The vertical axis is time, the horizontal axis is the variable x from the Lorenz system, and each line at constant time represents the probability density function of x at that time.

At $t = 0$ the system is near $x \approx 0$ and the initial ensemble consists of 512 initial states, each of which is both indistinguishable from the true state given the observations and also consistent with the long-term dynamics of the system (that is, 'on the attractor'). This constitutes a perfect ensemble. While only the value of the x component is shown, each member of the ensemble is a complete state of the system, and corresponding figures could be drawn for y and z .

Initially the distribution spreads out as might be expected, while the average value of x increases. At $t \approx 0.4$, however, the volume of the convex hull of the ensemble shrinks, showing a true 'return of skill' as the ensemble enters a region where all uncertainties decrease! Here a BFG forecast at $t \approx 0.4$ is expected to be more accurate than the corresponding forecast at $t \approx 0.2$. This is the origin of the discontinuity in doubling times noted above in **Figure 1**: red points to one side of the line double just before entering the region, while points in the rainbow bands just across the line enter the shrinking region before they double, and must wait a finite time to be advected out of the shrinking region before they might double. Hence the discontinuity.

Returning to **Figure 2**, notice that as distribution returns near $x = 0$ at $t \sim 0.7$, a small fraction of the ensemble members switch to the wing of the attractor with negative values of x , while the majority make another circuit with $x > 0$. Owing to the symmetry of the attractor, there is a somewhat artificial return of skill at $t \sim 1.5$. After this, however, the ensemble members divide more evenly between the two wings of the

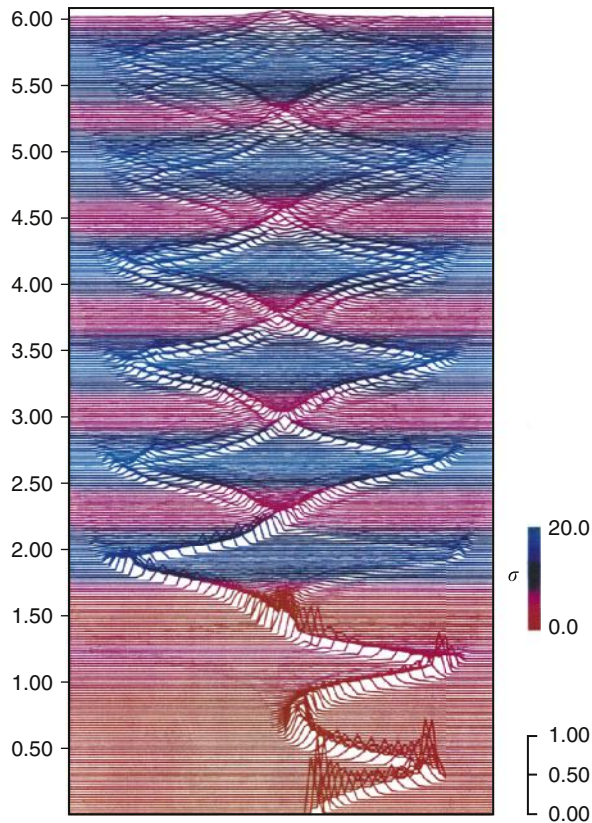


Figure 2 Evolution of a perfect ensemble under a perfect model, showing the probability distribution of one component of the model state, the variable x , in the Lorenz system. Time is denoted along the ordinate, while the abscissa is centered about zero. Each line reflects a particular time: the height reflects the probability density at that value of x , while the color of each line reflects the standard deviation of the ensemble at that time. Adapted with permission from Figure 23 of Smith LA (1997) The maintenance of uncertainty. Proceedings of the International School of Physics “Enrico Fermi”, course CXXXIII, pp. 177–246, Società Italiana di Fisica, Bologna, Italy.

attractor, and the distribution turns blue. The color here reflects the standard deviation of the forecast ensemble.

At this point, the standard deviation of the forecast is greater than that of a set of points taken at random from the attractor (that is, the *climatology*). Many classical measures of predictability would not see the significant information content that the forecast obviously continues to possess; good skill scores should reflect the information content of the ensemble. Even more worrying, the ‘optimal’ least-squares forecast at $t = 2.7$ would be near $x = 0$ where there is zero probability of observing the system. One important point illustrated here is that knowing the mean value exactly is often of much less value than knowing the likely distribution of values even approximately. A second point is that tuning nonlinear models with the aim of a ‘better’ average least-squares error will make the models worse, as it systematically forces model parameters away from more realistic, but heavily penalized, behavior. Such models are expected to be underactive rather than realistic.

The information in the initial ensemble will slowly diffuse away, and whether or not the information in the forecast at any

given time is useful depends on the aims of the user. Eventually, any finite ensemble will itself become indistinguishable from a random draw from the climatology. At this point the forecast is useless, but this time is unrelated to the Lyapunov exponent, or the doubling time, or any other measure of infinitesimal uncertainties.

Accountable Ensemble Forecasts

Corresponding to each probability forecast there is only a single verification; thus no single forecast can be evaluated. Rather, the quality of a (long) series of probability forecasts must be considered. And inasmuch as nonlinearity will mix aspects of data assimilation (see **Data Assimilation and Predictability: Data Assimilation**), ensemble formation and model inadequacy, the ensemble prediction system (EPS) can be evaluated only as a whole. This should not come as a surprise, since in a nonlinear system one expects to lose the benefits of linear superposition.

While the absolute accuracy of the EPS will vary with the level of initial uncertainty, ensemble forecasts under a perfect model using perfect initial ensembles are ‘accountable’: the uncertainty in any forecast variable computed from this ensemble will reflect the true value with an accuracy limited only by the finite number of members in the ensemble. Karl Popper introduced the notion of accountability for BFG forecasts in order to illustrate that a good model should indicate how accurately the initial condition must be measured in order to guarantee the accuracy of a forecast at any fixed lead time. The notion is easily extended to ensemble forecasts, in that an accountable ensemble forecast system should indicate how large an initial ensemble should be in order to reflect events accurately with a given level of probability.

Of course, the detailed shape of each forecast distribution will differ from day to day, **Figure 2** shows the probability distribution for a particular initial condition and set of observations. Yet if the EPS is accountable, then as the number of members in the ensemble is increased, the probability forecast will grow more accurate in a predictable way. For example, every time the ensemble size is doubled, the frequency with which any particular variable in the verification will fall outside the ensemble will be cut in half. [In fact, the ensemble size must be increased from N members to $(2N + 1)$ members, since the probability that the next random draw falls outside the current range is $2/(N + 1)$.]

In practice, ensembles are not drawn from a set of indistinguishable states; there are a number of competing methods now used operationally, and other methods are soon to join them. Current formation schemes include sampling directions of forecast errors of the recent past, or the directions of fastest growth in the near future. Neither approach attempts to sample the initial uncertainty accurately, and thus accurate probability forecasts could not be expected from the raw forecasts, even were the models to be perfect.

Operational ensembles typically consist of between 10 and 100 members, evolved over a duration of two weeks, although seasonal ensembles are a current topic of research. Recalling that operational model-state spaces typically have ten million dimensions gives an indication of just how difficult sampling

the initial uncertainty may prove to be. Despite the technical difficulties, the value of operational ensembles is reflected in [Figure 3](#), a 42-hour forecast for 26 December 1999. The three panels in the top left corner show the low-resolution (control) and high-resolution (T319) BFG forecasts, and the more colourful analysis (1 0 5) which serves as the verification. The color in the analysis reflects the intense winter storm that swept across Europe. The other 50 panels in the figure each show a member of the ensemble forecast at $t = +42$ hours. This collection of ‘postage stamp’ maps is analogous to a single PDF at constant time in [Figure 2](#). Note that about 20% of the ensemble members contain storms, and that even though there is no known way to extract an accountable probability estimate from this operational forecast, there is significantly more information than is provided by the control forecasts. In its present state, this information is already of significant societal and economic value.

Physical Systems and Mathematical Models

Arguably that, no physical system is ever isolated, and perturbations from outside the system imply that no dynamical system can be perfectly modeled as deterministic. What then does one mean by saying that a physical system is chaotic? Lorenz (1993) suggests a physical system should be called chaotic if its behavior would be chaotic were it to be isolated. This, of course, assumes there is perfect mathematical description of the hypothetical isolated system, but it is similar to the manner in which other mathematical terms are interpreted in physics; for example, the definition of periodicity in a physical system. Periodicity is a useful concept in physics, although arguably no physical system is truly periodic. Similarly, chaos may be a useful concept within physics, even if no physical system is truly deterministic.

One property that distinguishes periodic and chaotic systems is that periodic systems eventually return to exactly the same state x in state space. While this never happens in chaotic systems, near returns do occur for all points on the attractor; the longer is the duration of the observations, the closer are the nearest returns. Such systems, like the Lorenz 1963 model, are said to be ‘recurrent.’ What does it mean to say a physical system is recurrent.

At this point one has to leave the perfect-model scenario behind. Observations of a physical system are at best uncertain measurements of variables in the system’s state space (if such a thing exists); in order to use them in the model the observations must be cast into a model-state space. Mathematically, a data assimilation scheme (see [Data Assimilation and Predictability: Data Assimilation](#)) is simply a projection operator which accomplishes this task. Whatever the projection operator may be, the fact that forecasts are made in the model-state space holds deep consequences for attempts to make accountable probability forecasts. Estimates of predictability reflect the limitations of our models, while the underlying physical system is not so constrained.

Once some method of data assimilation is adopted so that the observations of the system can be projected into the model-state space, one can ask if a physical system is likely to be recurrent within a particular model-state space. Will two

similar states be observed during the likely duration of the observations? Over the lifetime of the system? Often the answer is yes.

Many physical systems are also recurrent within the model-state space over the time of a typical experiment. Near recurrence in the model-state space opens up many modeling possibilities, the simplest being to use (local) linear regression (see [Statistical Methods: Data Analysis: Time Series Analysis](#)). It also introduces the possibility that we can learn from past mistakes, improving the model by identifying state-dependent systematic errors. Of course, doing so may increase the dimension of the model-state space to the extent that given the available observations it is no longer recurrent!

Given that the estimated recurrence time of the Earth’s atmosphere is longer than the lifetime of the Solar System (longer, in fact, than the expected lifetime of the Universe!), this remedy is not available to meteorologists modeling the global circulation. Of course, one may be able to exploit recurrence in building parameterizations (see [Boundary Layer \(Atmospheric\) and Air Pollution: Modeling and Parameterization](#)) or in applications on smaller scales over shorter forecast horizons. But chaos is defined for the system (or model) as a whole. There is a clear and useful intuition of what is meant by the concept of ‘approximately periodic,’ which is lacking for the phrase ‘approximately chaotic.’ Further, it is not obvious how to interpret physical systems as chaotic, if they are not expected to exist over the time scales required for chaos to manifest itself in their mathematical analogs.

Loss of Predictability: Model Inadequacy and Shadowing

In addition to uncertainty in the initial condition and uncertainty in parameter values, meteorologists must contend with ‘model inadequacy’: there are aspects of the real physical system that our model is simply unable to mimic. When no model in the available class of models is structurally adequate to duplicate the observed phenomena, it is unclear what is meant by the ‘correct’ initial condition or the ‘true’ parameter values. While the Bayesian agenda provides a principled scheme for handling uncertainty in initial condition and parameter value, no systematic approach is available for handling model inadequacy. Progress here requires having a good idea.

Recall that in the discussion of ‘uncertainty’ in the initial condition above, it was assumed that in addition to a best-guess initial condition, $\hat{x}$, there was also a true initial condition, $\tilde{x}$. The error in the initial condition was defined as the difference between these two points. When the model is imperfect, there is no ‘true’ initial condition (even if the model variables have the same names as the system variables) and the very concept of ‘uncertainty in the initial condition’ has to be reconsidered.

As an example, note that since the current resolution of NWP models is at best tens of kilometers, many different states of the atmosphere (with different futures) will be mapped onto the same state of the model. This is but one example of the projection effects noted above: the model initial conditions are, at best, projections of the true system state into the model-state space. The model cannot, then, be expected to reproduce the evolution of every atmospheric state, simply because there are

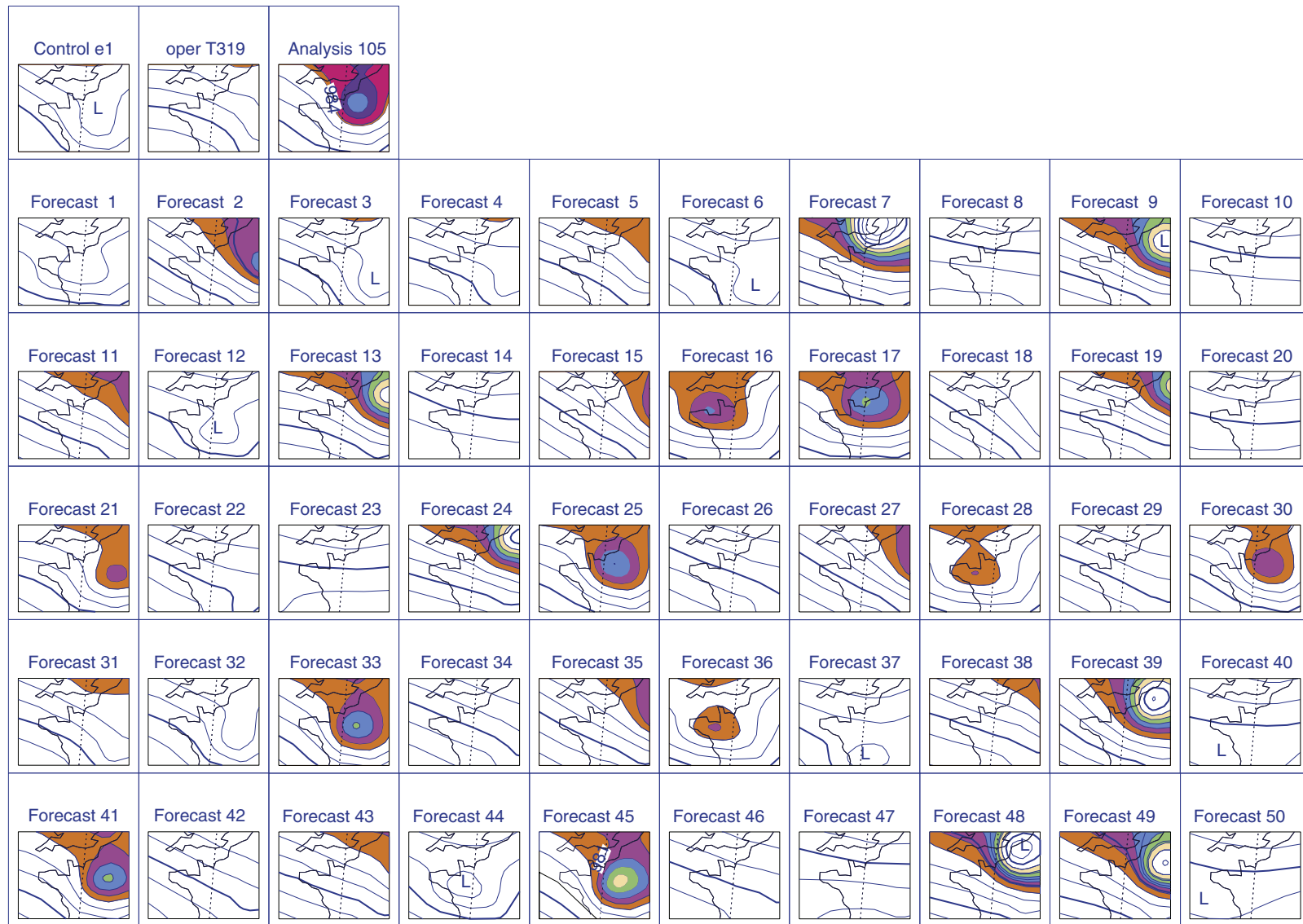


Figure 3 Ensemble forecast for the French storm of 1999. Each 'postage stamp' is a weather map of southern England and France. The three panels in the top left corner show two best-guess forecasts made at different model resolutions and the analysis, which indicates that the verification was rather different from either of these forecasts. Each of the 50 members of the ensemble at the time of the verification is also shown. Reproduced with permission of the European Centre for Medium-Range Weather Forecasts.

more atmospheric states than model states! Of course, the model may have trajectories which shadow the observed atmospheric states, remaining indistinguishable from the trajectory of the atmosphere given the observational uncertainty. How might meteorologists distinguish forecast failures due to limitations in the ensemble formation scheme from those due to model inadequacy? One approach is to look for 'shadowing trajectories' within the historical observations.

Given an imperfect model, there may or may not be a model trajectory that stays close to the series of observed states, no matter which data assimilation method is used to translate the observations into model states. In this context, 'close' must be interpreted in relation to the uncertainty in the observations. A model trajectory that remains near a set of target states is said to 'shadow' the target states. Each analysis will have an associated shadowing time, just as it has an associated value of τ_2 . The distribution of shadowing times reflects the relevance of model inadequacy.

If shadowing trajectories exist, then initial condition(s) which shadow may be cast in the role of 'truth' (that is, the role of $\tilde{x}$) when computing uncertainty in the initial condition, at least for forecasts that are short relative to the duration over which the model can shadow. This suggests that our very definition of 'observational noise' will itself depend on the quality of the model in hand. Indeed, many data assimilation schemes are based on the assumption that long shadowing trajectories exist almost everywhere in state space.

If no shadowing trajectory exists on the time scale of interest, then the model mixes 'uncertainty in the initial condition' and 'model inadequacy' to the extent that the former cannot be unambiguously defined. On these time scales, all model trajectories differ significantly from the observations: the set of indistinguishable states is empty, and there is no optimal method of ensemble formation. Indeed, outside PMS the issue of model improvement is linked to that of forecast usage; there need be no unique best way forward. Nevertheless, current ensemble forecasts are of great value in identifying when the forecasts are sensitive to uncertainties in the initial condition, since any single BFG forecast can be identified, at the time it is issued, as unlikely to be an accurate anticipation of reality. Hence they can be expected to provide useful identification of when the forecast will be unreliable; empirical studies suggest they are also useful in identifying forecasts which are likely to have high skill. In addition, when two members of the same ensemble lead to radically different forecasts in the medium range, determining what distinguishes them at short lead times can suggest valuable observations for improving the forecast.

In addition to ensembles over initial conditions, research is underway aimed at determining how to better include stochastic effects into nonlinear models. Such stochastic effects are commonly referred to as 'dynamical noise' to distinguish

them from observational noise; the latter alters the observations but not the trajectory. A major difficulty here is formulating a relevant state-dependent dynamical noise, as the traditional approaches tend to spread the forecast into unphysical directions. Using multiple models provides one approach, stochastic parameterizations is another.

Current research is also exploring the use of ensembles over distinct models, or even ensembles over trajectories each of which uses a variety of distinct models. Ideally, these models should be independent, so that they share as few common inadequacies as possible. Methods for allocating resources among models, and for the evaluation of the distributions so obtained as forecasts, provide yet other interesting areas of current research. The traditional goal of identifying the 'optimal' least-squares predictor need no longer be a desirable end for any real forecast user. Modern forecast users, in particular industrial users, are quite capable of exploiting probability forecasts.

Since the introduction of the electronic computer, indeed since L. F. Richardson's computations early in the last century, weather prediction has been at the forefront of research into the predictability of nonlinear dynamical systems. One safe forecast is that it will remain there for the foreseeable future.

See also: Boundary Layer (Atmospheric) and Air Pollution: Modeling and Parameterization. Data Assimilation and Predictability: Data Assimilation; Ensemble Prediction. **Statistical Methods:** Data Analysis: Time Series Analysis.

Further Reading

- Lorenz (1993) provides a general introduction to both the history and physics of chaos, while Smith (1998) is an overview of the implications chaos holds for philosophy as well as a general introduction. Introductions to both operational and theoretical ensemble forecasts can be found in Palmer (2000) and Smith, et al. (1999), respectively. A general discussion of the role of model inadequacy in predictability from philosophical, physical, and Bayesian perspectives can be found in Cartwright (1983), Smith (2001), and Kennedy and O'Hagan (2001), respectively and reference thereof.
- Cartwright, N., 1983. *How the Laws of Physics Lie*. Oxford University Press, Oxford.
- Kennedy, M., O'Hagan, A., 2001. Bayesian calibration of computer models. *Journal of the Royal Statistical Society Series B* 63, 425–464.
- Lorenz, E., 1993. *The Essence of Chaos*. University College London Press, London.
- Palmer, T.N., 2000. Predicting uncertainty in forecasts of weather and climate. *Reports on Progress in Physics* 63, 71–116.
- Smith, L.A., 2001. Disentangling uncertainty and error: on the predictability of nonlinear systems. In: Mees, A.I. (Ed.), *Nonlinear Dynamics and Statistics*. Birkhauser, Boston, MA, pp. 31–64.
- Smith, L.A., Ziehmann, C., Fraedrich, K., 1999. Uncertainty dynamics and predictability in chaotic systems. *Quarterly Journal of the Royal Meteorological Society* 125, 2855–2886.
- Smith, P., 1998. *Explaining Chaos*. Cambridge University Press, Cambridge.

DYNAMICAL METEOROLOGY

Contents

Overview

Acoustic Waves

Atmospheric Tides

Balanced Flow

Baroclinic Instability

Coriolis Force

Critical Layers

Hamiltonian Dynamics

Hydraulic Flow

Inertial Instability

Kelvin–Helmholtz Instability

Kelvin Waves

Kinematics

Laboratory Geophysical Fluid Dynamics

Lagrangian Dynamics

Potential Vorticity

Primitive Equations

Quasigeostrophic Theory

Rossby Waves

Solitary Waves

Static Stability

Stationary Waves (Orographic and Thermally Forced)

Symmetric Stability

Vorticity

Wave-CISK

Wave Mean-Flow Interaction

Waves

Overview

JR Holton, University of Washington, Seattle, WA, USA

© 2003 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 2, pp 673–679, © 2003, Elsevier Ltd.

Introduction

Dynamic meteorology is the branch of fluid dynamics concerned with the meteorologically significant motions of the atmosphere. It forms the primary scientific basis for weather and climate prediction, and thus plays a primary role in the atmospheric sciences. Most of the meteorologically important motions studied in dynamic meteorology are profoundly influenced by the facts that the Earth is

a rapidly rotating planet, and that the atmosphere on average has stable density stratification. These facts make the fluid dynamics of the atmosphere very different from traditional engineering fluid dynamics. Planetary rotation places strong constraints on large-scale horizontal motions; stable stratification places strong constraints on vertical motions. These constraints can be understood by considering the fundamental physical laws governing motions of the atmosphere.

The motions of the atmosphere are governed by the laws for conservation of mass, conservation of momentum, and conservation of thermodynamic energy. Application of these laws to motions with horizontal scales of several hundred kilometers or greater leads to simple relations among the horizontal wind, pressure, and temperature distributions. These relations form a set of diagnostic relations essential for understanding the motions that generate weather disturbances. Such motions are generally rotational in character. They can be characterized by a conservable property known as the potential vorticity, which is the fluid dynamical analogue of spin angular momentum in solid mechanics.

The latitudinal gradient of potential vorticity provides the mechanism for generation of global-scale planetary waves, which are primary features of the climate system. Superposed on these global waves are transient cyclones and anticyclones, whose energy is derived primarily from the potential energy associated with the mean Pole-to-Equator temperature gradient. Study of the development and evolution of transient weather disturbances, and of dynamical mechanisms for producing intraseasonal and interannual climate variations, are among the principal areas of study in dynamic meteorology.

The Static Atmosphere

The vertical distribution of pressure, density and temperature is determined by the hydrostatic approximation and the ideal gas law. The hydrostatic approximation, which is derived from the vertical component of the momentum equation (Newton's second law of motion), expresses the balance between the vertical component of the pressure gradient force and the force of gravity. The ideal gas law, or equation of state, expresses the relationship between pressure, density, and temperature in an ideal gas. These two equations may be combined to form the hypsometric equation, which determines the thickness of the layer of temperature T confined between the two pressure surfaces p_2 and p_1 :

$$\Phi_2 - \Phi_1 = g(Z_2 - Z_1) = R \int_{p_2}^{p_1} T \, d \ln p \quad [1]$$

where Φ is the geopotential, Z is geopotential height, T is the absolute temperature, R ($= 287 \text{ J kg}^{-1} \text{ K}^{-1}$) is the gas constant for dry air, and g ($= 9.81 \text{ m s}^{-2}$) is the acceleration due to gravity. The hypsometric equation shows that atmospheric pressure decreases more rapidly in cold air than in warm air. Thus the average altitude of a given upper-level pressure surface (e.g., the 500 hPa surface) decreases towards higher latitudes owing to the decrease of mean temperature with latitude.

Hydrostatic balance requires that pressure decrease monotonically with height in the atmosphere. Pressure may thus be substituted for height as a vertical coordinate; this has the advantages of eliminating explicit reference to the density field in the equations of motion; but it has the disadvantage that pressure varies exponentially with altitude so that equal altitude increments correspond to rapidly decreasing pressure increments as height increases. For this reason, in dynamical meteorology it is often useful to use

log-pressure coordinates in which the independent vertical coordinate z is defined by

$$z = -H \ln (p/p_0) \quad [2]$$

where $H = RT_0/g$ is an atmospheric scale-height, $p_0 = 10^3 \text{ hPa}$ (1000 mbar), and T_0 is a mean temperature. Comparing with eqn [1], it is clear that the log-pressure coordinate corresponds to actual height for an isothermal atmosphere at temperature T_0 . Under most conditions the departure of z from actual altitude is small enough to be neglected.

In the absence of precipitation, changes in temperature following the motion of individual parcels of air are controlled primarily by adiabatic expansion and compression as the air parcels move to lower or higher pressure. The thermodynamic state of such parcels can be characterized by the potential temperature, θ . Potential temperature is the temperature that a parcel of dry air initially at a pressure p and temperature T would acquire if it were moved adiabatically to the reference pressure p_0 . It is defined by the following relation, which can be obtained from the first law of thermodynamics:

$$\theta = T(p_0/p)^{R/c_p} \quad [3]$$

where c_p ($= 1004 \text{ J kg}^{-1} \text{ K}^{-1}$) is the specific heat capacity of dry air at constant pressure. Normally, surfaces of constant potential temperature in the atmosphere are quasi-horizontal with potential temperature increasing with altitude. Air parcels displaced vertically conserving potential temperature are then colder and denser than their surroundings for an upward displacement, and vice versa for a downward displacement. The atmosphere is then said to be statically stable. When diabatic heat sources (such as latent heating and radiation) are neglected, θ remains constant in time for each air parcel; thus potential temperature is conserved following the motion. Air parcels are then constrained to remain on surfaces of constant θ , which are referred to as isentropic surfaces. In a statically stable atmosphere, potential temperature can be used as the independent vertical coordinate. This isentropic coordinate system is useful for analysis of adiabatic motions, since the prediction of atmospheric motions for such conditions is reduced from a three-dimensional problem to a two dimensional problem on each isentropic coordinate surface. Because diabatic temperature changes associated with large-scale weather disturbances in the extratropics are often much smaller than adiabatic changes, isentropic analysis has proved valuable for the study of air motions associated with such disturbances.

The Equations of Motion

It is convenient to express the basic equations of dynamic meteorology in a coordinate system rotating with the Earth, and with the log-pressure altitude defined by eqn [2] as the independent vertical coordinate. The approximate conservation equations for horizontal momentum, mass, and thermodynamic energy are then as follows:

$$\frac{D\mathbf{V}}{Dt} = -f\mathbf{k} \times \mathbf{V} - \nabla\Phi + \mathbf{F}\mathbf{r} \quad [4]$$

$$\nabla \cdot \mathbf{V} + \frac{1}{\rho_0} \frac{\partial}{\partial z} (\rho_0 w) = 0 \quad [5]$$

$$\frac{DT}{Dt} + \frac{wN^2H}{R} = \frac{J}{c_p} \quad [6]$$

Here, t is time, $\mathbf{V}$ is the horizontal velocity, ∇ is the horizontal gradient evaluated at constant log-pressure, Φ is the geopotential, $w \equiv Dz/Dt$, is the vertical velocity, $f = 2\Omega \sin \phi$ is the Coriolis parameter (where $\Omega = 7.292 \times 10^{-5} \text{ s}^{-1}$ is the angular velocity of rotation of the Earth and ϕ is latitude), $\mathbf{k}$ is the vertical unit vector, Fr is the drag due to small-scale turbulent fluctuations, $\rho_0 \equiv \rho_s \exp(-z/H)$ is the basic state density (where ρ_s is density at the surface, H is the scale height defined below eqn [2], J is the diabatic heating rate, and N is the buoyancy frequency whose square is defined in terms of the height dependence of potential temperature as $N^2 = g(\partial \ln \theta / \partial z)$.

In eqns [4] and [6], D/Dt is the rate of change following the horizontal motion of a fluid parcel. This can be related to the rate of change at a given point in space by the expression:

$$\frac{D}{Dt} \equiv \left(\frac{\partial}{\partial t} + \mathbf{V} \cdot \nabla \right) \quad [7]$$

The momentum eqn [4] states that a change in the horizontal velocity following the motion of an air parcel is caused by the net imbalance among three forces: the Coriolis force caused by the rotation of the Earth, the horizontal pressure gradient force (given by the gradient of geopotential on a constant pressure surface) and the force due to turbulent dissipation (important mainly near the surface of the Earth). The equation of mass continuity [5] states that the divergence of the horizontal velocity, which tends to increase or decrease the horizontal cross-section of a fluid parcel, must be balanced by vertical motion, which increases or decreases the depth of the parcel. The thermodynamic energy eqn [6] states that the rate of change of temperature following the horizontal motion is due to the sum of the adiabatic expansion or compression due to vertical motion, plus the net diabatic heating by sources such as latent heat release and solar or thermal radiative heating.

Balanced Flow

When diabatic heating and turbulent dissipation are sufficiently small, eqns [4]–[6], together with the differential form of the hypsometric eqn [1] and suitable initial and boundary conditions, form a closed set for prediction of the meteorological fields $\mathbf{V}$, w , Φ , and T . Solutions of this complicated set of nonlinear partial differential equations can usually only be obtained by numerical methods. There are, however, certain approximate solutions that provide useful information on the relations among these field variables for large-scale atmospheric flows in the extratropical regions.

When turbulent dissipation is small, which is generally true above the lowest kilometer of the atmosphere, large-scale extratropical motions are approximately in geostrophic balance, that is, the horizontal pressure gradient force and the Coriolis force are nearly equal and opposite. The wind velocity for which this balance is exact is referred to as the geostrophic

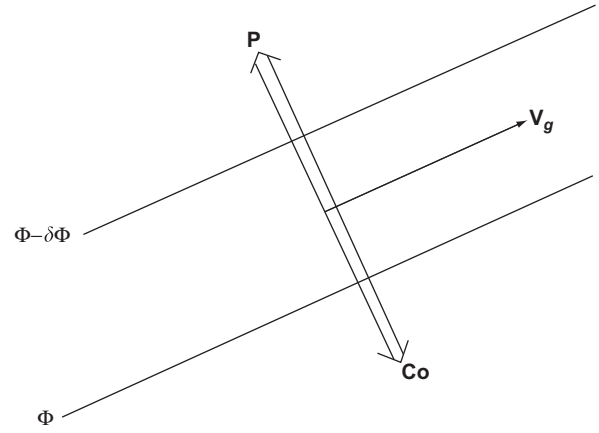


Figure 1 Horizontal plan view of balance of forces for the geostrophic wind. $\mathbf{V}_g$, $\mathbf{P}$ designates the pressure gradient force, $\mathbf{Co}$ designates the Coriolis forces, and Φ and $\Phi - \delta\Phi$ are contours of constant geopotential on an isobaric surface.

wind, $\mathbf{V}_g$. It is derived from eqn [4] with the acceleration and turbulent dissipation terms neglected:

$$f\mathbf{V}_g = \mathbf{k} \times \nabla \Phi \quad [8]$$

This states that the geostrophic wind is parallel to lines of constant geopotential (or constant height) on a pressure surface, with speed proportional to the magnitude of the geopotential gradient on the pressure surface. As illustrated in Figure 1, the Coriolis force acts perpendicular to the wind direction (to the right of the wind in the Northern Hemisphere), while the pressure gradient force is directed opposite to the Coriolis force. Thus, the geostrophic circulation about the center of a low-pressure system in the Northern Hemisphere will be a counterclockwise circulation.

Equation [8] may be combined with eqn [1] to give the thermal wind relation:

$$f[\mathbf{V}_g(p_2) - \mathbf{V}_g(p_1)] = R \int_{p_2}^{p_1} (\mathbf{k} \times \nabla T) d \ln p \quad [9]$$

This relation, which is a consequence of geostrophic and hydrostatic balance, states that the vector difference in the geostrophic wind velocity between two pressure surfaces is proportional to the horizontal gradient of the mean temperature in the layer between the two surfaces. Since large-scale extratropical motions are in hydrostatic balance and are nearly geostrophic, eqn [9] shows that the wind and temperature fields are closely coupled. Because temperature in the extratropical lower atmosphere generally decreases with latitude, the thermal wind relation indicates that the eastward-directed geostrophic wind increases with altitude and that the strongest upper-level winds will occur where that latitudinal temperature gradient is the strongest.

Planetary Boundary Layer

In the lowest kilometer of the atmosphere, momentum transfer by small-scale turbulent eddies becomes an important component in the momentum balance so that the geostrophic approximation is no longer valid. The structure of this boundary

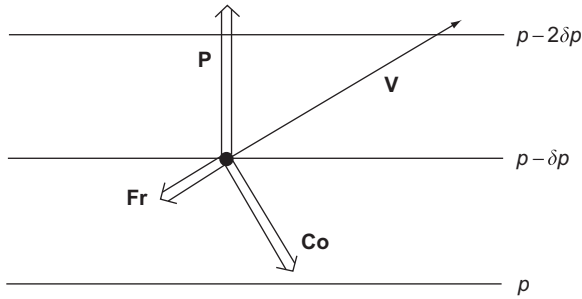


Figure 2 Horizontal plan view of balance of forces in a well-mixed planetary boundary layer. **V** is velocity, **P** designates the pressure gradient force, **Co** designates the Coriolis force, **Fr** designates the turbulent drag force, p is pressure, and δp is a constant pressure interval. Adapted from Holton, J.R., 1992. *Introduction to Dynamic Meteorology*. Academic Press, New York.

layer depends strongly on the vertical stratification in the layer. In regions of strong vertical stability (where potential temperature increases rapidly with height) turbulence is generally weak, except in the lowest few meters above the surface where strong vertical shear of the wind provides a mechanical source for turbulent eddies. Under conditions of strong surface heating, on the other hand, the boundary layer may be convectively unstable (potential temperature decreasing with height) and strong turbulent eddies may extend throughout the lowest kilometer of the atmosphere. Over vast areas of the Earth's surface, however, the boundary layer is often near neutral static stability (potential temperature constant with height). Although accurate representation of the force exerted by turbulent eddies in eqn [4] is a challenging fluid dynamical problem, it is useful as a first approximation to simply assume that turbulent eddies exert a drag on the winds so that $\mathbf{Fr} = -k\mathbf{V}$, where k is a rate coefficient typically taken to be about 10^{-5} s^{-1} .

The force balance in the boundary layer is then a three-way balance among the Coriolis force, the pressure gradient force, and the turbulent drag force. This balance is illustrated schematically in Figure 2. Since the Coriolis force always acts perpendicularly to the wind and the turbulent drag acts in the opposite direction to the wind, a force balance can be achieved only if the wind has a component directed across the isobars towards lower pressure. The component produces net boundary layer inflow into surface low pressure systems, which by transporting mass towards the pressure minimum acts to 'spin down' the circulation.

Vorticity and Potential Vorticity

Because extratropical flow above the boundary layer is approximately in geostrophic balance, and thus tends to be parallel to height contours on pressure surfaces, the flow is characterized by cyclonic and anticyclonic gyres associated with height (or pressure) minima and maxima, respectively. Thus, the large-scale flow is dominated by a rotational flow component, which is conveniently described in terms of the vorticity (defined as the curl of the velocity vector). Although momentum is not conserved for such flows, but is changed in proportion to the small difference between the Coriolis force and the pressure gradient force, there is a dynamical quantity

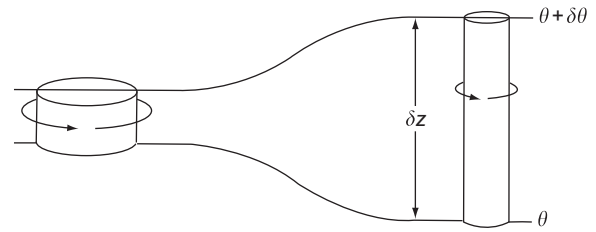


Figure 3 Cylindrical column of air in adiabatic motion from a region of high static stability to a region of low static stability. The column is confined between two potential temperature surfaces, θ and $\theta + \delta\theta$, which are separated by an altitude increment δz . Adapted from Holton, J.R., 1992. *Introduction to Dynamic Meteorology*. Academic Press, New York.

that is conserved following the motion for adiabatic frictionless flows. This quantity, which is referred to as potential vorticity, is a fluid dynamical analogue of spin angular momentum in solid body dynamics. In its simplest form, potential vorticity relates the vertical component of vorticity (or spin) of a fluid column confined between two potential temperature surfaces to the depth of the column (Figure 3).

Potential vorticity can be expressed mathematically as

$$P = (\zeta + f) \frac{1}{\rho_0} \frac{\partial \theta}{\partial z} \quad [10]$$

Here, $\zeta = \mathbf{k} \cdot (\nabla \times \mathbf{V})$ is the vertical component of the relative vorticity due to the horizontal winds and f is again the Coriolis parameter, which is the vertical component of the vorticity owing to the rotation of the Earth. The sum of the local vertical components of the relative vorticity and the Earth's vorticity, $\zeta + f$, is referred to as the absolute vorticity. Equation [10] states that for a column of fluid confined between two potential temperature surfaces separated by a fixed increment of potential temperature $\delta\theta$, the ratio of absolute vorticity to the depth of the column δz remains constant. Thus, for example, a column moving from a region of high static stability (small δz) to a region of low static stability (large δz) will stretch vertically, shrink horizontally and spin faster. Because potential vorticity is conserved following the motion for adiabatic frictionless flow, the evolution of the field of potential vorticity on isentropic surfaces can easily be predicted. The potential vorticity distribution, together with suitable boundary conditions, can in turn be used to deduce the three-dimensional distribution of wind and temperature.

Zonally Symmetric Circulations

In the troposphere, temperature normally decreases from Equator to Pole owing to the latitudinal gradient in solar heating. This heating gradient induces an ageostrophic overturning circulation of air parcels in the meridional (height–latitude) plane. The overturning consists of rising motion in association with convective disturbances in the tropics, poleward drift in the upper troposphere, slow sinking in the extratropics and an equatorward drift in the planetary boundary layer. This zonally symmetric parcel circulation is important for transport of water vapor, momentum, and heat

across latitude circles, but represents only a small deviation from the generally west-to-east directed (zonal) winds.

Because the Equator-to-Pole temperature decrease does not occur uniformly but tends to be concentrated in the subtropics, the zonal winds are also concentrated in the jet stream, a narrow band of strong westerly winds in the upper troposphere that encircles the Earth at an average latitude of about 30° . The association of the jet stream with a strong meridional temperature gradient is a consequence of the thermal wind relationship (eqn [9]), which states that eastward-directed winds must increase rapidly with height where the temperature decreases rapidly in latitude.

The concentration of the meridional temperature gradient (and potential vorticity gradient) in the subtropical jet stream cannot be understood by considering the zonally symmetric circulation forced by solar heating, but rather represents a complex interaction between the symmetric circulation and the disturbances that characterize weather and climate. Thus, the study of atmospheric wave disturbances is one of the primary areas of dynamic meteorology.

Atmospheric Waves

Waves in the atmosphere are motions that can transmit energy and momentum without material transport of air parcels. Most weather disturbances are associated with one or more types of atmospheric wave. Atmospheric waves result from a balance between the inertia of the atmosphere and a restoring force. In acoustic waves, for example, oscillations in the pressure gradient force are balanced by parcel accelerations along the direction of phase propagation. Such waves are longitudinal waves in the sense that the fluid parcel oscillations are parallel to the direction of propagation. Most meteorologically important waves, however, are transverse waves in which the parcel oscillations are perpendicular to the direction of phase propagation. Examples of such waves are buoyancy waves, inertia waves, and Rossby waves.

Buoyancy Waves

The vertical stratification of the atmosphere causes a fluid parcel that is displaced vertically to experience a restoring force due to its buoyancy; the resulting coherent oscillations in the pressure, temperature, and wind fields is called a buoyancy wave, or gravity wave. The structure of an eastward-propagating buoyancy wave, excited from below, is shown in Figure 4. Phase lines, defining the maxima in the perturbation pressure and temperature fields, tilt towards the east with height, and propagate eastwards and downwards in time. However, the flux of energy (measured by the correlation between the pressure and vertical velocity perturbations) is directed upwards, parallel to the lines of constant phase. There is also an upward flux of eastward momentum in this case since positive vertical perturbations are in phase with eastward horizontal velocity perturbations, and vice versa. Thus, buoyancy waves provide an important mechanism for vertical transport of momentum in the atmosphere.

The frequency, ω , of a buoyancy wave is related to the angle, α , of the phase lines to the local vertical by the formula

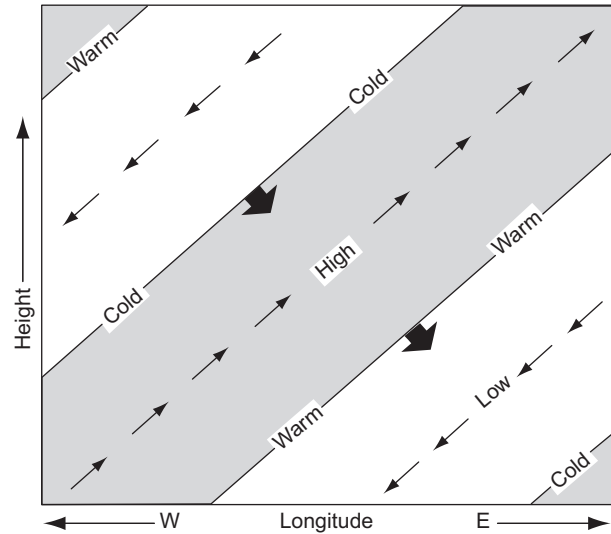


Figure 4 Cross-section showing phases of the pressure, temperature, and velocity perturbations for an internal gravity wave. Thin arrows indicate the perturbation velocity field, blunt solid arrows the phase velocity. Shading shows region of upward motion.

$\omega = N \cos \alpha$, where N is again the buoyancy frequency. Thus, high-frequency buoyancy waves have energy propagation that is closer to the vertical than do low-frequency buoyancy waves. An important special case of buoyancy waves are lee waves formed when air in a stable atmosphere is forced to ascend over a mountain barrier. Oscillations that are induced by the forced uplift may continue several wavelengths downstream of the mountain.

When sufficient moisture is present, condensation may occur in the updraft portions of the waves, leading to regularly spaced bands of cloud in the lee of the mountain. In this case the waves are stationary with respect to the ground but propagate relative to an observer moving with the mean wind.

Rossby Waves

The most important class of large-scale atmospheric waves are called planetary waves, or Rossby waves. These waves are characterized by oscillations in the rotational part of the horizontal wind that are parallel to the horizontal gradient in the potential vorticity. The simplest example of a Rossby wave occurs in a barotropic atmosphere. A barotropic atmosphere is one in which potential temperature is constant on each pressure surface so that $\partial\theta/\partial z$ in eqn [10] is independent of horizontal position. Potential vorticity conservation then reduces to conservation of absolute vorticity following the motion:

$$\frac{D}{Dt}(\zeta + f) = 0 \quad [11]$$

The mechanism of Rossby wave propagation can be understood by considering a tube of fluid parcels that at time t_0 is motionless and lies parallel to a latitude circle. In that case $\zeta(x, t_0) = 0$. If the tube is given a small sinusoidal meridional

displacement $\eta(x, t_0)$ at the initial time, then from eqn [11] at a later time t_1 the relative vorticity will be given by:

$$\zeta(x, t_1) = f_{t_1} - f_{t_0} = -\beta\eta(x, t_1) \quad [12]$$

where $\beta \equiv df/dy$, the rate of increase of the Coriolis parameter with latitude. Since β is positive, it is evident from eqn [12] that the relative vorticity perturbation will be positive for a southward displacement and negative for a northward displacement. Since positive relative vorticity corresponds to a counterclockwise rotation and negative relative vorticity corresponds to a clockwise rotation, the induced perturbation velocities will tend to produce meridional displacement of the tube of fluid parcels to the west of the original displacement, as illustrated in Figure 5. By this mechanism, the originally wavy displacement propagates westwards, perpendicular to the north-south displacement of the fluid parcels.

Quantitative analysis of the relationship of the wave frequency to the horizontal scale reveals that Rossby waves are highly dispersive (i.e., their phase speeds are strongly dependent on the wavelength). In Cartesian coordinates the zonal (eastward) phase speed relative to the ground has the form:

$$c_x = \bar{U} - \beta(L_x^2 + L_y^2)/(4\pi^2) \quad [13]$$

Here $\bar{U}$ is the mean eastward wind, L_x is the zonal wavelength and L_y is the meridional wavelength. Thus, Rossby waves propagate westwards relative to the mean wind with phase speeds that increase rapidly as the wavelength increases. For waves corresponding to typical extratropical storms with wavelengths of a few thousand kilometers, the Rossby wave phase speed is typically less than 10 m s^{-1} , which is less than the mean wind speed. Thus, relative to the ground, such disturbances tend to move eastwards, but more slowly than the average eastward wind speed.

Equation [13] also shows that Rossby waves that are stationary relative to the ground ($c_x = 0$) can exist only when the mean zonal flow is positive (i.e., eastwards). Stationary

Rossby waves are excited in midlatitudes when westerly winds encounter largescale mountain barriers and are displaced meridionally during ascent over the barrier. This process is a primary reason why large-scale stationary cyclonic disturbances are commonly observed to the lee of the Rocky and the Himalayan mountain ranges.

Baroclinic Waves

Barotropic Rossby waves are generated by conversion of the kinetic energy of the zonal flow into kinetic energy of the waves. They do not involve any conversion of potential energy to kinetic energy. Such energy conversion can occur only in the presence of baroclinicity, that is, variations of potential temperature on isobaric surfaces. Baroclinic energy conversion is responsible for the growth and maintenance of most large-scale weather disturbances.

When the latitudinal gradient of the zonal wind in the jet stream is sufficiently strong that the meridional gradient of potential vorticity on a constant potential temperature surface is locally reversed, or when there is a nonvanishing gradient of potential temperature at the surface of the Earth, the equations of motion linearized about a zonally symmetric basic state have solutions in the form of exponentially growing wave disturbances. These baroclinically unstable waves have growth rates, structure, and scales typical of those observed in developing extratropical cyclones. They are quasi-geostrophic in the sense that the geostrophically balanced rotational component of the wind field strongly dominates over the ageostrophic divergent component. The latter is, however, crucial in the energy cycle by which the waves convert potential energy associated with the Pole-to-Equator temperature gradient into disturbance kinetic energy.

Baroclinic instability provides a mode whereby infinitesimal disturbances may be amplified into large-amplitude storms. In many situations, however, it appears that weather disturbances may develop rapidly from preexisting upper-level potential vorticity anomalies in the absence of baroclinic instability. As in baroclinic instability, the growth of storms from upper-level potential vorticity anomalies is associated with conversion of potential energy to kinetic energy in association with the ageostrophic secondary flow induced by adjustments towards thermal wind balance.

Equatorial Waves

In the equatorial region there is a special class of weather-producing waves that combine some of the characteristics of gravity waves and of Rossby waves. Equatorial waves are trapped in latitude, that is, they propagate along the Equator with amplitudes decreasing with latitude. In some circumstances they may also propagate energy and momentum vertically. Two important examples of equatorial waves are the equatorial Kelvin wave and the Rossby-gravity wave. The Kelvin wave has pressure and zonal velocity perturbations symmetric about the Equator, and negligible meridional velocity component. It propagates eastwards, with vertical structure identical to the eastward-propagating buoyancy wave shown in Figure 4, and is an important source of eastward momentum for the equatorial stratosphere.

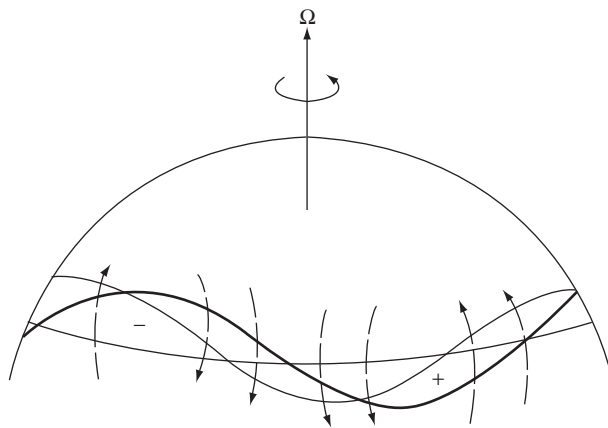


Figure 5 Perturbation vorticity field (+ and -) and induced velocity field (dashed arrows) for a meridionally displaced tube of fluid parcels, showing the mechanism for Rossby wave propagation. Heavy wavy line shows initial perturbation of the tube, light wavy line shows westward displacement of the wave perturbation due to advection by the induced velocity field. From Holton, J.R., 1992. *Introduction to Dynamic Meteorology*. Academic Press, New York.

Rossby-gravity waves are waves that combine characteristics of Rossby waves and gravity waves. They have meridional wind distributions symmetric about the Equator and zonal wind and pressure distributions antisymmetric about the Equator. Away from the Equator the wind and pressure distributions in the Rossby-gravity wave are nearly geostrophic, but near the Equator there are strong departures from geostrophic balance. Rossby-gravity waves propagate westwards relative to the mean flow. They are associated with equatorial weather disturbances, and are also a significant source of westward momentum for the equatorial stratosphere.

Mesoscale Disturbances

If an air parcel is saturated, upward displacement will cause water vapor to condense and release its latent heat of condensation; potential temperature is then no longer conserved, but increases following the parcel motion. If this increase is greater than the potential temperature gradient of the background atmosphere, the atmosphere is said to be conditionally unstable. That is, it is stable with respect to unsaturated parcel displacement but unstable with respect to saturated parcel displacements. The convective storms associated with cumulonimbus clouds can occur only when the atmosphere is conditionally unstable, sufficient moisture is present, and sufficient lifting occurs to bring air parcels to

saturation. The processes that organize convection into mesoscale convective systems are not completely understood. Mesoscale convective systems take a variety of forms. Among these are hurricanes, squall lines, and supercell thunderstorms. In all cases the release of latent heat by convective clouds is the primary energy source, but the character of the largescale environmental flow is generally important for determining the mode of organization for mesoscale systems.

See also: **Dynamical Meteorology:** Baroclinic Instability; Coriolis Force; Rossby Waves; Stationary Waves (Orographic and Thermally Forced); Waves. **Gravity Waves:** Overview.

Mesoscale Meteorology: Waterspouts. **Synoptic Meteorology:** Cyclogenesis.

Further Reading

- Cushman-Roisin, B., 1994. *Introduction to Geophysical Fluid Dynamics*. Prentice-Hall, London.
- Gill, A.E., 1982. *Atmosphere–Ocean Dynamics*. Academic Press, New York.
- Holton, J.R., 1992. *Introduction to Dynamic Meteorology*. Academic Press, New York.
- Pedlosky, J., 1987. *Geophysical Fluid Dynamics*. Springer-Verlag, Berlin.
- Salby, M.L., 1996. *Fundamentals of Atmospheric Physics*. Academic Press, New York.

Acoustic Waves

KE Gilbert, University of Mississippi, University, MS, USA

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

A brief overview of basic acoustic wave propagation is given, followed by a discussion of refraction and diffraction of audible sound in realistic daytime and nighttime atmospheres. Numerical examples show the effect of atmospheric turbulence on sound propagation in an upward-refracting (daytime) unstable atmosphere and a downward-refracting (nighttime) stable boundary layer. Experimental results for pulse propagation in a stable nighttime boundary layer with internal gravity waves are presented and discussed. Finally, remote sensing of the atmosphere using a sodar and a radio acoustic sounding system is briefly discussed. A grayscale plot from a sodar is presented and explained. The plot shows the evolution of the atmospheric boundary layer over a diurnal cycle.

Introduction

This article is concerned with acoustic waves in the atmosphere. Owing to space constraints, the discussion is limited to audible acoustic waves (sound waves). Hence, two important topics – ultrasound (above audible) and infrasound (below audible) – are not discussed. Further, in order to provide a more in-depth discussion of the effect of the atmospheric boundary layer on sound waves, some traditional topics such as ground effects, nonlinear effects, and noise control are omitted. The interested reader should refer to the resources cited under the Further Reading section for information on aspects of acoustic waves not covered here.

At the atomic level, the Earth's atmosphere is a collection of gas molecules, mainly nitrogen and oxygen, bound to the planet by gravity. The microscopic properties of the atmosphere are thus described by the kinetic theory of gases and quantum mechanics. In contrast, at the macroscopic level, the atmosphere can be regarded as a fluid, and, in principle, can be described by the equations of fluid dynamics. Both points of view, molecular and fluid dynamical, are needed to fully understand the generation, propagation, and absorption of the disturbances in the atmosphere that are familiar to us as acoustic waves or 'sound.'

Unlike wave motion on a stretched string or ripples on the surface of water, acoustic waves in the atmosphere have no direct visual representation. Consequently, one must in general rely heavily on a mathematical description. It is useful, nevertheless, to try to connect the mathematical description of sound with an intuitive, physical picture, even if the picture is an approximate representation of reality. Hence, for purposes of visualization, one can schematically represent a planar acoustic wave as shown in Figure 1. In Figure 1(a), regions of compression (positive pressure relative to the ambient background pressure) and regions of rarefaction (negative pressure relative to the ambient background pressure) are indicated schematically by the density of points. Closely spaced points represent a compression, and less closely spaced points represent a rarefaction. The vertical lines in Figure 1(b) indicate regions of constant pressure that are called 'wavefronts.' The maximum pressure regions are indicated by solid vertical lines and the minimum pressure regions by dashed lines. The horizontal line perpendicular to the wavefronts is called an acoustic

'ray.' Acoustic rays are a concise way to indicate the travel paths taken by acoustic wavefronts as they propagate through space. In Figure 1(c), the regions of compression and rarefaction, often called the 'acoustic' pressure, are shown moving to the right with a speed c , which for dry air is 331.6 m s^{-1} at 0°C . For a compact representation of the pressure wave one could, for example, omit the wavefronts and simply show an acoustic ray moving to the right with a speed c .

For sinusoidal pressure variations, a planar acoustic wave can be represented mathematically as $p = p_0 \cos(kx - \omega t + \theta)$, where p_0 is the acoustic pressure amplitude, and the entire argument of the cosine is called the 'phase' of the wave. The angular frequency, ω , is $2\pi f$, where f is the frequency in cycles

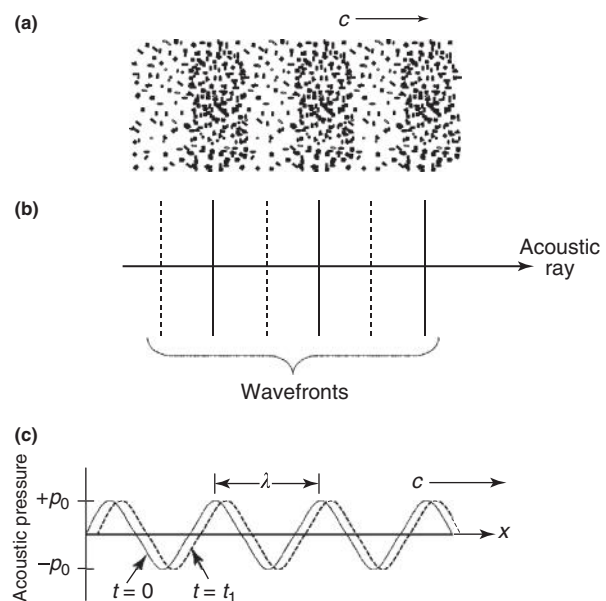


Figure 1 Visualization of a planar acoustic wave moving to the right at speed c . (a) Schematic representation of regions of compression (denser points) and regions of rarefaction (less dense points). (b) Wavefronts (regions of constant pressure); maximum and minimum pressure regions are represented, respectively, by solid and dashed vertical lines. An acoustic ray is drawn perpendicular to the wavefronts. (c) Pressure variation in space at two instants of time for a sinusoidal plane wave of the form $p = p_0 \cos(kx - \omega t + \theta)$.

per second or, more commonly, in hertz (Hz). The wave number k is $2\pi/\lambda$, where λ is the wavelength shown in Figure 1(c). Since $\lambda f = c$, the wave number can also be written as ω/c . The quantity θ is called the 'phase angle' and gives the phase of the wave at $x = 0$ and $t = 0$.

The compressions and rarefactions in an acoustic wave cause variations in density and temperature that also propagate with the wave. For all three quantities – pressure, temperature, and density – the acoustic amplitude is extremely small relative to the ambient background. For example, p_0 might be 0.1 Pa or approximately one millionth of the nominal atmospheric pressure.

The Acoustic Wave Equation

As noted above, acoustic waves in the atmosphere can be viewed as small disturbances on an ambient background fluid, just as water waves are seen as disturbances on a calm surface. For the extremely small pressure perturbations typical of sound, the equations of fluid dynamics can be linearized to arrive at the 'acoustic wave equation,' which is the conventional mathematical description of acoustic pressure waves. In one dimension, the acoustic wave equation is given by eqn [1], where p is the acoustic pressure, x is the distance, and t is the time.

$$\frac{\partial^2}{\partial x^2} p = \frac{1}{c^2} \frac{\partial^2}{\partial t^2} p \quad [1]$$

The general solution to eqn [1] is of the form $p(x, t) = p_R(x - ct) + p_L(x + ct)$, where $p_R(x - ct)$ is a right-going wave and $p_L(x + ct)$ is a left-going wave. The right-going wave, for example, could be a transmitted pulse, and the left-going wave could be an echo. Continuous waves as well as pulses satisfy the wave equation. For example, since $c = \omega/k$, the sinusoidal pressure wave discussed above satisfies the one-dimensional wave equation. Moreover, as indicated in Figure 2, any function of $(x - ct)$ or $(x + ct)$ satisfies eqn [1]. Further, the perturbations in density and temperature associated with an acoustic pressure wave satisfy the same wave equation as the acoustic pressure except that, instead of pressure, the variable is density or temperature, respectively.

The three-dimensional form of eqn [1] is eqn [2], where (x, y, z) are Cartesian coordinates.

$$\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2} \right) p = \frac{1}{c^2} \frac{\partial^2}{\partial t^2} p \quad [2]$$

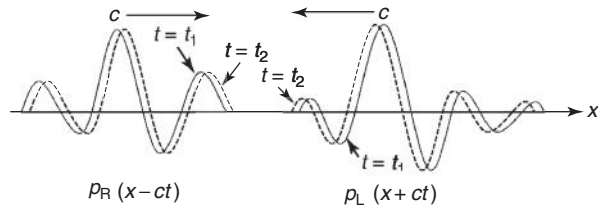


Figure 2 Solutions to the one-dimensional wave equation. The function $p_R(x - ct)$ is a right-going solution and the function $p_L(x + ct)$ is a left-going solution. The complete solution is the superposition of the left- and right-going solutions.

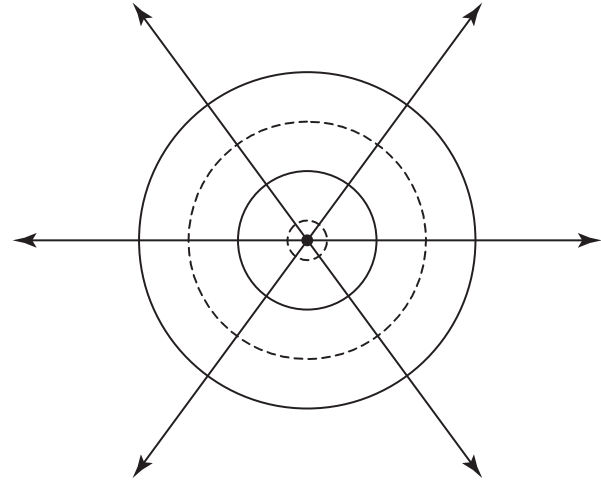


Figure 3 Schematic representation of an out-going spherical wave. The circles are the wavefronts and the straight lines are acoustic rays. The solid and dashed circles denote, respectively, wavefronts for maximum and minimum pressure.

For a symmetrical source, such as a small explosion high above the ground, the three-dimensional wave equation has spherical symmetry and can be written as eqn [3].

$$\frac{\partial^2}{\partial r^2} (rp) = \frac{1}{c^2} \frac{\partial^2}{\partial t^2} (rp) \quad [3]$$

Here $r = \sqrt{x^2 + y^2 + z^2}$. By comparing the form of eqn [3] with that of eqn [1], one can deduce that the general solution of eqn [3] is given by eqn [4], in which r_{ref} is an arbitrary reference distance, generally taken to be 1 m in the meter-kilogram-second (MKS) system.

$$p(x, t) = \left(\frac{r_{\text{ref}}}{r} \right) [p_{\text{OUT}}(r - ct) + p_{\text{IN}}(r + ct)] \quad [4]$$

The quantities $(r_{\text{ref}}/r)p_{\text{OUT}}(r - ct)$ and $(r_{\text{ref}}/r)p_{\text{IN}}(r + ct)$ are out-going and in-going spherical waves, respectively. Note that the spherical wave solution has the same mathematical form as the plane-wave solution except that the amplitude falls off as $1/r$.

For a source far away from boundaries, the acoustic pressure is given by an out-going wave having the same shape in the time domain as the source function. For example, for a time-harmonic source, the acoustic field is a traveling sinusoidal wave of the form $p = p_0(r_{\text{ref}}/r) \cos(kr - \omega t + \theta)$, where p_0 is the pressure amplitude at the reference distance. Pictorially, an out-going spherical wave can be represented as shown in Figure 3, where pressure maxima and minima of the wavefronts are represented, respectively, by solid lines and dashed lines. The radial lines perpendicular to the wavefronts are acoustic rays.

Sound Pressure Levels and Decibels

Acoustic pressure amplitudes encountered in practice typically vary over several orders of magnitude. Consequently, it has become conventional to use a logarithmic scale to describe the amplitudes. For continuous waves, the amplitude of interest is the root-mean-square pressure amplitude, p_{rms} , and is referenced to some standard reference pressure p_{ref} . For pulses, some

'peak' pressure is often chosen. In either case, a logarithmic amplitude measure called the 'sound pressure level' (SPL) is commonly used, defined by eqn [5].

$$\text{SPL} \equiv 10 \log_{10} \left(\frac{p_{\text{rms}}^2}{p_{\text{ref}}^2} \right) = 20 \log_{10} \left(\frac{p_{\text{rms}}}{p_{\text{ref}}} \right) \quad [5]$$

The pressure of interest is denoted here as the rms pressure, but could be any pressure, depending on the application. Although SPL is actually a dimensionless quantity, one refers to the 'units' as decibels (dB), referenced to a reference pressure, p_{ref} . In atmospheric acoustics, the reference level is usually chosen to be 2×10^{-5} Pa or 20 μPa , which is the approximate threshold of hearing. Note that with the above conventions, the SPL for 20 μPa is 0 dB.

The frequency range for audible sound for the human ear is from approximately 20 Hz to approximately 20 kHz. Typical SPLs for sounds in the audible range are given in Table 1.

In addition to being used as a measure for absolute pressure, decibels are also used to describe relative changes in pressure. For example, if a pressure amplitude decreases with distance by a factor of 10, it is conventional to say that, over the distance, the acoustic pressure has decreased by $20 \log_{10}(10) = 20$ dB. If one pressure amplitude were 100 times greater than another, one would say the first pressure was 40 dB greater than the second pressure.

To express in decibels the variation of the rms pressure, $p_{\text{rms}}(r)$, with distance, it is conventional to define the SPL at $r = r_{\text{ref}}$ as the 'source level' (SL) as in eqn [6] and to define the decibel decrease in acoustic pressure with distance as the 'transmission loss' (TL) as in eqn [7].

$$\text{SL} \equiv 20 \log_{10} \left(\frac{p_{\text{rms}}(r_{\text{ref}})}{p_{\text{ref}}} \right) \quad [6]$$

$$\text{TL} \equiv -20 \log_{10} \left(\frac{p_{\text{rms}}(r)}{p_{\text{rms}}(r_{\text{ref}})} \right) \quad [7]$$

In these equations, as noted previously, the reference distance, r_{ref} , is 1 m in the MKS system. (Note that since pressure usually decreases with distance, TL is usually positive.) Using the above definitions for SL and TL, one can write the SPL at a distance r from the source as $\text{SPL} = \text{SL} - \text{TL}$. For example, in MKS units, the rms pressure amplitude for a spherically spreading wave can be written as $p_{\text{rms}}(r) = p_1/r$, where p_1 is the rms pressure at 1 m. Thus, for a spherically spreading wave, the SL is $20 \log_{10}(p_1/p_{\text{ref}})$ and the TL is simply $20 \log_{10}(r)$.

Table 1 Representative list of audible sound pressure levels

Sound	SPL (dB)
Threshold of hearing	0
Leaves rustling	20
Quiet conversation	40
Normal conversation	60
Average street traffic	80
Diesel truck (at 10 m)	90
Jet takeoff (at 10 m)	120
Threshold of pain	140

In general, the TL is not a simple function and must be computed numerically. With numerical computations, it is often useful, for plotting purposes, to subtract the TL due to spherical spreading, that is, to subtract $20 \log_{10}(r)$. Such a convention is equivalent to giving the SPL relative to a spherically spreading wave, and hence is given the name 'relative SPL.' Thus, by definition, the relative SPL for a spherically spreading wave is 0 dB. Expressed as a relative SPL, an SPL above or below that for spherical spreading will be, respectively, greater than or less than zero.

The Speed of Sound in the Atmosphere

To a good approximation, the atmosphere can be treated as an ideal gas, and the acoustic pressure variations in it can be treated as adiabatic; that is, there is no heat flow from the high-pressure (hotter) regions to the low-pressure (cooler) regions. For an ideal gas and adiabatic compression (or rarefaction), the speed of sound is given by $c = \sqrt{\gamma RT}$, where $\gamma = 1.40$ is the ratio of the constant-volume specific heat for air, c_V , to the constant-pressure specific heat, c_P . The quantity $R = 286.69 \text{ J kg}^{-1} \text{ K}^{-1}$ is the gas constant for dry air, and T is the absolute temperature (K). With an ideal gas model, the theoretical value for c at 0°C (273.16 K) is 331.1 m s^{-1} , which is in excellent agreement with the experimental value of 331.6 m s^{-1} given earlier. For values of T not far from 0°C , the square root expression for the speed of sound can be expanded linearly and written approximately as $c = (331 + 0.6T_C) \text{ m s}^{-1}$, where T_C is the temperature in degree Celsius. Thus, for an increase in temperature of 1°C , the speed of sound increases by 0.6 m s^{-1} .

Absorption of Sound in the Atmosphere

In addition to the decrease in pressure amplitude of an acoustic wave due to propagation effects such as 'geometrical' spreading (e.g., spherical spreading), the amplitude is also reduced by atmospheric absorption. A sound wave propagating through 'clean' air (no solid particles) is attenuated owing to two basic mechanisms:

- classical losses due to momentum transfer across a velocity gradient (viscosity) and heat flow across a temperature gradient and
- quantum-mechanical losses due to relaxation processes, mainly relaxation of rotational and vibrational states in nitrogen and oxygen molecules.

For both mechanisms, the effects of absorption can be represented by an absorption coefficient, α , which has units of m^{-1} . The absorption coefficient enters via an exponential, so that the pressure is given by $p = p_0 e^{-\alpha s}$, where p_0 is the unattenuated pressure amplitude and s is the distance the wave has traveled. To indicate the choice of the Napierian base, e , the attenuation coefficient is, by convention, said to have units of nepers m^{-1} . The corresponding attenuation coefficient, a , for decibels (base 10) is $a = 20\alpha \log_{10}(e) = 8.686\alpha$, and, by convention, has units of dB m^{-1} .

Experimental and theoretical studies indicate that the total absorption coefficient can be represented as a sum of

absorption coefficients, with each distinct physical process having its own unique coefficient. That is, the total absorption coefficient α_T can be written as $\alpha_T = \sum_i \alpha_i$, where α_i is the absorption coefficient associated with a particular mechanism. For example, the component of sound absorption due to viscosity is given by eqn [8], in which ω is the angular frequency, c is the speed of sound, ρ_0 is the density of air, and μ is the coefficient of viscosity of air.

$$\alpha_{\text{vis}} = \frac{\omega^2}{2\rho_0 c^3} \frac{4\mu}{3} \quad [8]$$

Since the compressed regions in an acoustic wave are slightly hotter than the ambient temperature, and the expanded regions are slightly cooler, a small amount of heat flows from the compressions to the rarefactions. The conduction of heat converts the organized motion associated with the sound wave into random thermal motion of the gas molecules. Because the heat flow lowers the temperature of the compressions and raises the temperature of the rarefactions, both the pressure maxima and minima are reduced. The reduction manifests itself as a decay of the acoustic wave with distance. The component of absorption due to thermal conduction is given by eqn [9], where κ is the coefficient of thermal conductivity in $\text{J (kg mol)}^{-1} \text{K}^{-1} \text{kg m}^{-1} \text{s}^{-1}$.

$$\alpha_{\text{TH}} = \frac{\omega^2}{2\rho_0 c^3} (\gamma - 1) \frac{\kappa}{\gamma c_V} \quad [9]$$

In addition to the energy loss due to classical mechanisms (viscosity and heat conduction), energy can also be lost via quantum-mechanical ‘relaxation’ processes involving the internal degrees of freedom (rotation and vibration) of oxygen and nitrogen molecules. The transfer of translational energy to and from internal degrees of freedom takes place through an extended sequence of molecular collisions, so there is a time delay associated with the energy transfer. Because of the time delay, relaxation processes cause energy to be lost from the organized translational motion that constitutes the acoustic wave. As a result, just as with the classical mechanisms, the pressure amplitude of the wave decreases as the wave propagates.

For any particular relaxation process, the associated absorption coefficient has the general form of eqn [10].

$$\alpha = \frac{\pi s_v}{c} \frac{f^2/f_r}{1 + (f/f_r)^2} \quad [10]$$

In eqn [10], s_v is the relaxation strength (in nepers), c is the sound speed, f is the frequency, and f_r is called the ‘relaxation’ frequency. The relaxation frequency is the frequency for maximum absorption and is roughly the reciprocal of the characteristic time delay for the transfer between kinetic energy and internal energy of the gas molecules.

For air, there are three important relaxation processes: (1) O_2 vibration, (2) N_2 vibration, and (3) N_2 rotation. The relaxation frequency for N_2 rotational relaxation is very high, so that, below 10 MHz, the denominator in eqn [10] is approximately unity. Thus, N_2 rotational relaxation varies as f^2 and can be combined with the classical absorption coefficient. If one denotes the classical-plus-rotational absorption coefficient (i.e., the coefficient for viscosity, heat flow, and N_2 rotation) as α_C , the absorption coefficient for O_2 vibration as

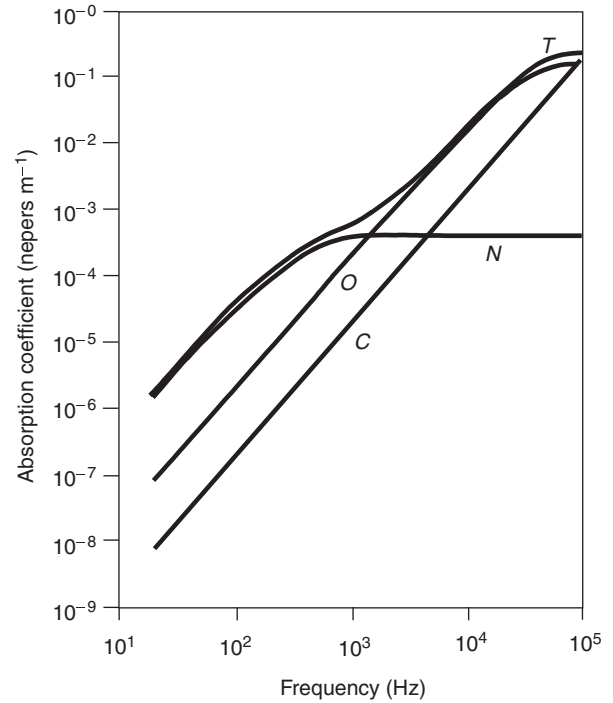


Figure 4 Components and general behavior of the total absorption coefficient for air. The contributions to the total absorption (T) are the classical plus N_2 rotation (C), the O_2 vibration (O), and the N_2 vibration (N). Reproduced with permission from Bass, H.E., 1991. Atmospheric acoustics. In: Trigg, G.L., (Ed.), Encyclopedia of Applied Physics vol. 2, Wiley-VCH Verlag GmbH & Co. KGaA, pp. 145–179.

α_O , and the absorption coefficient for N_2 vibration as α_N , then the total absorption coefficient α_T for air can be written as eqn [11].

$$\alpha_T = \alpha_C + \alpha_O + \alpha_N \quad [11]$$

Figure 4 shows the total absorption coefficient, α_T , together with the components, α_C , α_O , and α_N . Note that below about 10 000 Hz the absorption is dominated by vibrational relaxation. Further, note that, above about 1000 Hz, atmospheric absorption is significant for propagation distances of a kilometer or more, which accounts for the lack of long-range propagation of high-frequency sound.

Refraction of Sound in the Atmosphere

The spherical wavefronts and associated rays shown in Figure 3 represent acoustic waves radiating from a point source in an atmosphere with a constant temperature, and hence a constant sound speed. With a constant sound speed, acoustic ray trajectories are straight lines. In reality, however, the atmospheric temperature is never constant in space or time. Consequently, the speed of sound is not constant but varies spatially and temporally. In a typical daytime situation, the temporally averaged temperature is independent of range but decreases with height (‘lapse’ condition). Thus, on average, the sound speed decreases with height, and sound rays curve upward, as shown in Figure 5. The ray paths shown are for a sound speed that decreases linearly with

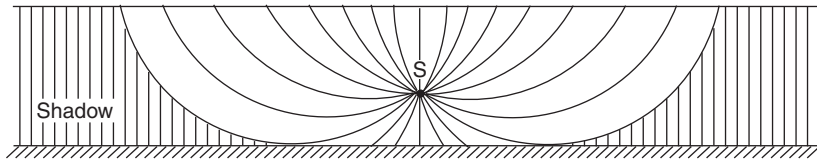


Figure 5 Acoustic rays and shadow zones for an upward-refracting daytime atmosphere. For clarity, rays that are reflected off the ground are not shown. Reproduced with permission from Bass, H.E., 1991. Atmospheric acoustics. In: Trigg, G.L., (Ed.), Encyclopedia of Applied Physics vol. 2. Wiley-VCH Verlag GmbH & Co. KGaA, pp. 145–179.

height. (For the simple case of linear variation, the ray paths are arcs of circles.)

For a general sound speed variation in a stratified atmosphere (i.e., no horizontal variation), ray paths are governed mathematically by Snell's law, which states that the quantity $c(z)/\cos \theta(z)$ is invariant, where at a height z , the quantities $c(z)$ and $\cos \theta(z)$ are, respectively, the sound speed and the cosine of the angle of a ray with respect to horizontal. Thus, Snell's law says that if $c(z)$ decreases with height, $\cos \theta(z)$ must also decrease, so that the ray bends upward. In general, acoustic rays bend toward regions of lower sound speed and away from regions of higher sound speed. The bending of acoustic rays is given the name 'refraction.'

The physical basis for refraction can be understood using the situation in Figure 5. Consider a small section of wavefront associated with a ray that leaves the source nearly parallel to the ground. For a small enough section, the wavefront is nearly planar and nearly vertical. Since the sound speed decreases with height, the lower portion of the wavefront travels faster than the upper portion, causing the wavefront to turn upward. In terms of rays, we would say that the ray is refracted upward due to the decrease in the sound speed with height.

Refraction of acoustic waves is caused by spatially varying wind as well as by spatially varying temperature. The effect of the wind on acoustic waves can be accounted for approximately by defining an 'effective' sound speed, c_e , which is the previously defined 'adiabatic' sound speed, $c_a = \sqrt{\gamma RT}$, plus the component of the vector wind in the direction of propagation. For example, let $\hat{r}$ be a unit vector pointing from the source to a receiver. Then, if the vector wind is denoted as $\mathbf{v} \equiv (v_x, v_y, v_z)$, the effective sound speed is given by $c_e = c_a + \hat{r} \cdot \mathbf{v}$, where $\hat{r} \cdot \mathbf{v}$ is the component of the vector wind in the direction of sound propagation. In general, near the ground, the horizontal wind speed increases with increasing height. For upwind propagation of sound, therefore, the horizontal wind progressively reduces the effective sound speed with increasing height. For downwind propagation, the effect is reversed. That is, with downwind propagation of sound, the

horizontal wind progressively increases the effective sound speed with increasing height. In the daytime, for example, where the temperature and adiabatic sound speed decrease with height, upwind propagation adds to the upward refraction already present. Downwind, if the wind speed gradient is sufficiently large, the horizontal wind can overcome the upward refraction due to the daytime temperature profile and lead to downward refraction. The ray paths for upwind and downwind propagation are illustrated in Figure 6. For propagation directly across the wind, there is little effect due to the wind, but upward refraction persists because of the decreasing temperature with height.

Hereafter in this article, the terms 'sound speed' and 'effective sound speed' are used interchangeably, so that it is always assumed that the effect of wind is included. When the wind is not included, the sound speed is referred to as the 'adiabatic sound speed.'

As indicated in Figures 5 and 6, for upward refraction, there is a region called an acoustic 'shadow,' where no acoustic rays can penetrate. In the shadow region, the acoustic levels are much lower than the SPL one would expect with spherical spreading alone. Because of upward refraction, daytime SPLs for a near-ground source fall off dramatically with horizontal distance as one enters the shadow region, which, for strong upward refraction, can be within 100–200 m of the source.

At night, in contrast to the daytime situation, ground-to-ground propagation is very good. Owing to radiative cooling of the ground, both the near-ground air temperature and the sound speed are lower than at higher altitudes (an 'inversion' condition). As a result, acoustic rays launched near to horizontal (less than about 10° with respect to horizontal) are bent downward, causing sound to be trapped in a 'sound duct' near the ground. Rays launched at steeper angles escape the duct and continue upward (Figure 7). With strong trapping and small ground-bounce loss (e.g., over water) the acoustic field in the near-surface sound duct undergoes approximately cylindrical spreading ($1/\sqrt{r}$) instead of spherical spreading ($1/r$).

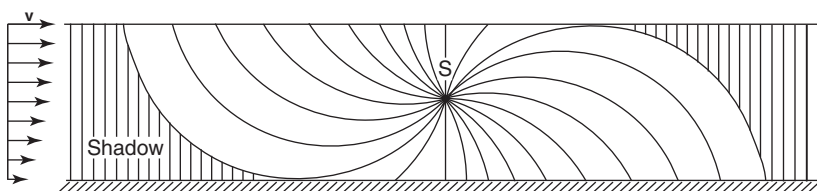


Figure 6 Acoustic rays and shadow zone for an atmosphere that is upward refracting in the upwind direction and downward refracting in the downwind direction. For clarity, rays that are reflected off the ground are not shown. Reproduced with permission from Bass, H.E., 1991. Atmospheric acoustics. In: Trigg, G.L., (Ed.), Encyclopedia of Applied Physics vol. 2. Wiley-VCH Verlag GmbH & Co. KGaA, pp. 145–179.

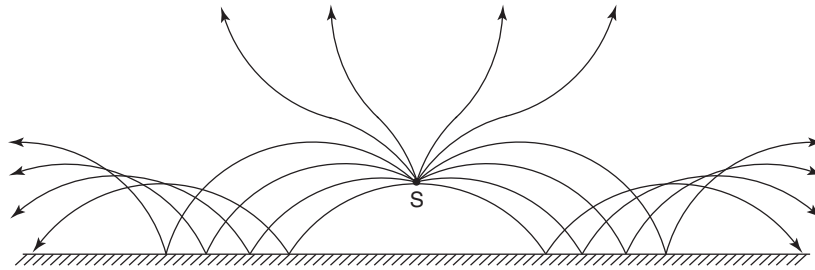


Figure 7 Acoustic rays for a downward-refracting nighttime atmosphere. Rays launched at small angles with respect to the horizontal (less than about 10°) are trapped in a 'sound duct' near the ground and can propagate to ranges of several kilometers. For steeper launch angles, the rays escape the duct and continue upward.

As a consequence of daytime upward refraction and nighttime downward refraction, noise sources that are not generally heard during the day can often be easily heard at long distances (e.g., several kilometers) at night. The long-range propagation of acoustic waves at night makes noise control much more difficult than during the day.

Diffraction of Acoustic Waves in the Atmosphere

As discussed above, one can approximately represent an acoustic field in terms of wavefronts whose propagation directions (i.e., rays) are governed by refraction. Such a representation is useful visually and can be valid computationally when the acoustic wavelengths are much smaller than the smallest sound-speed structure in the atmosphere. The main effect left out in the so-called 'ray theory' of sound propagation

is the wave phenomenon known as 'diffraction.' Diffraction is responsible for the well-known ability of sound to 'bend' around corners and obstacles. In outdoor sound propagation, diffraction fills in gaps in the acoustic field that would be present in a purely ray-based representation. Full-wave solutions to the wave equation (usually numerical) automatically include both diffraction and refraction.

Because of diffraction, every acoustic field has an intrinsic smallest possible scale length that is roughly a quarter of the smallest wavelength present in the field. Owing to the scale limitation, there can be no sharp edges in the acoustic field. For example, instead of the sharp shadow boundary obtained with rays (Figures 5 and 6), a smooth, diffuse boundary is obtained when diffraction is included. Such a situation is illustrated in Figure 8, which shows a numerical solution of the wave equation for a 500 Hz point source in an upward-refracting atmosphere. The color plot in the figure, which is for the relative SPL

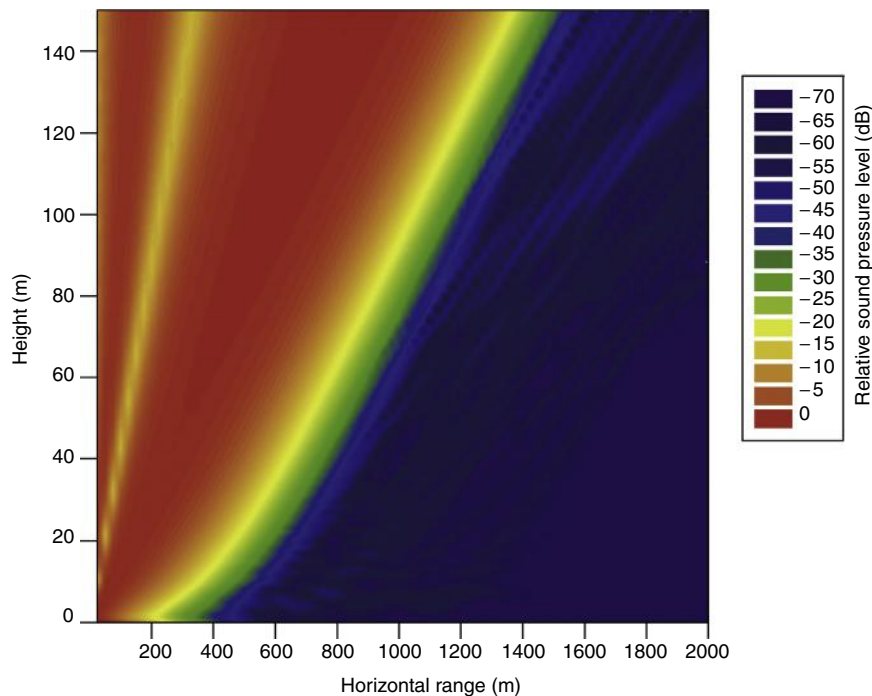


Figure 8 Numerical solution for the sound field in an upward-refracting daytime atmosphere without turbulence. Owing to diffraction, the edge of the shadow boundary is diffuse. The color plot shows the relative SPL as a function of range and height. (Note that the vertical scale is much less than the horizontal scale, so that the actual propagation angles with respect to the horizontal are much smaller than shown.)

as a function of range and height, shows the effects of both diffraction and refraction. The edge of the shadow boundary in [Figure 8](#) would become more diffuse at lower frequencies (longer wavelengths), until finally, at very low frequencies, the shadow boundary would not be discernible at the ranges shown.

When the atmosphere is downward refracting (e.g., at night), the presence of diffraction again causes a ‘blurring’ of the features of the acoustic field, just as with upward refraction. There are no sharp discontinuities in the structure of the acoustic field. Rather, because of the finite wavelengths in the acoustic field, the changes in the field are continuous and smooth, as shown in [Figure 9](#), which is also for a 500 Hz point source. In general, the longer the acoustic wavelengths (i.e., the lower the frequency), the smoother the features of the acoustic field.

In addition to limiting the sharpness of the acoustic field, diffraction is responsible for the scattering of acoustic waves from the complex small-scale structure of the real atmosphere. In a realistic model of the atmosphere, the instantaneous temperature and vector wind fields are not smooth but are highly irregular, containing eddies of all sizes. The eddy sizes of most concern for audible sound are in the region called the ‘inertial subrange,’ which typically begins at a few tens of meters and goes down to a few millimeters. In the inertial subrange, the eddy structure is governed by the well-known Kolmogorov spectrum. As a consequence, at any instant of time, the small-scale spatial structure of the sound speed field, which depends on temperature and vector wind, can also be described by a Kolmogorov spectrum. In the daytime turbulent boundary layer, for example, the sound speed can be approximated as a time-independent mean sound speed, $\bar{c}(z)$, that varies only with height, plus a fluctuating part, $\delta c(x, y, z, t)$, that varies with time, horizontal distance, and height. Hence the total sound speed, $c(x, y, z, t)$, can be represented as in eqn [12].

$$c(x, y, z, t) = \bar{c}(z) + \delta c(x, y, z, t) \quad [12]$$

The quantity $\bar{c}(z)$ approximates the slow, large-scale variations in the sound speed profile and $\delta c(x, y, z, t)$ describes the rapid, smaller-scale fluctuations.

As noted earlier, the quantity $\delta c(x, y, z, t)$ follows the same Kolmogorov statistics as do the temperature and wind fluctuations. At a particular instant of time, an approximate ‘snapshot’ of the sound-speed fluctuation field, $\delta c(x, y, z, t)$, can be synthesized by adding together, with random phase, the wave number components for a Kolmogorov spectrum. The result of such a synthesis is shown in two dimensions in [Figure 10](#). Because audible sound has wavelengths comparable in size to small-scale atmospheric structure, it is scattered in all directions as it propagates through inertial-subrange eddies. As a consequence of diffraction, new wavefronts emanate from every eddy, with the strongest scattering occurring in the near-forward direction. The diffracted acoustic waves that are scattered downward act to fill in the shadow region. An example of this phenomenon is shown in [Figure 11](#), which was computed numerically using realistic representations for $\bar{c}$ and δc . With a realistic model for the sound speed, the predicted mean near-ground levels (–20 to –30 dB relative to spherical spreading) in the shadow region (0.2–2 km) are in good agreement with the measured mean sound pressure levels. The relative SPL for a longer-range interval is shown in [Figure 12](#). It is apparent from [Figure 12](#) that, even with scattering into the shadow region, daytime levels near the ground are very low at ranges beyond a few kilometers.

It can be observed in [Figures 11 and 12](#) that the effects of turbulence are most apparent in the shadow region, where the sound levels would be extremely small in the absence of turbulence. Above the daytime shadow region, in the ‘insonified’ region, the levels are much higher, so that the effect of scattering from turbulence is less dramatic, though the effect increases with increasing distance from the source. Similarly, for nighttime propagation in the near-ground acoustic duct, where the mean

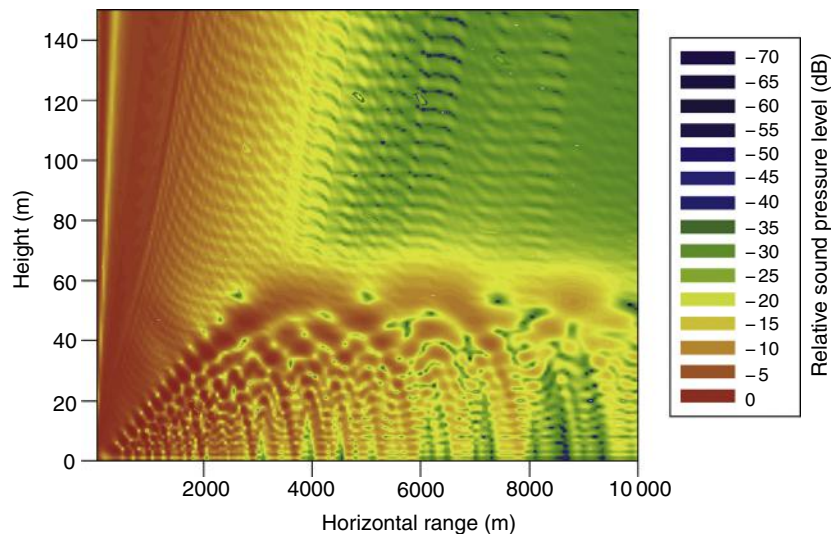


Figure 9 Numerical solution for the sound field in a downward-refracting nighttime atmosphere without turbulence. Note that downward refraction ‘traps’ sound in the near-ground acoustic duct. As in [Figure 8](#), the features of the acoustic field are blurred owing to diffraction. The color plot shows the relative SPL as a function of range and height. (Note that the vertical scale is much less than the horizontal scale, so that the actual propagation angles with respect to the horizontal are much smaller than shown.)

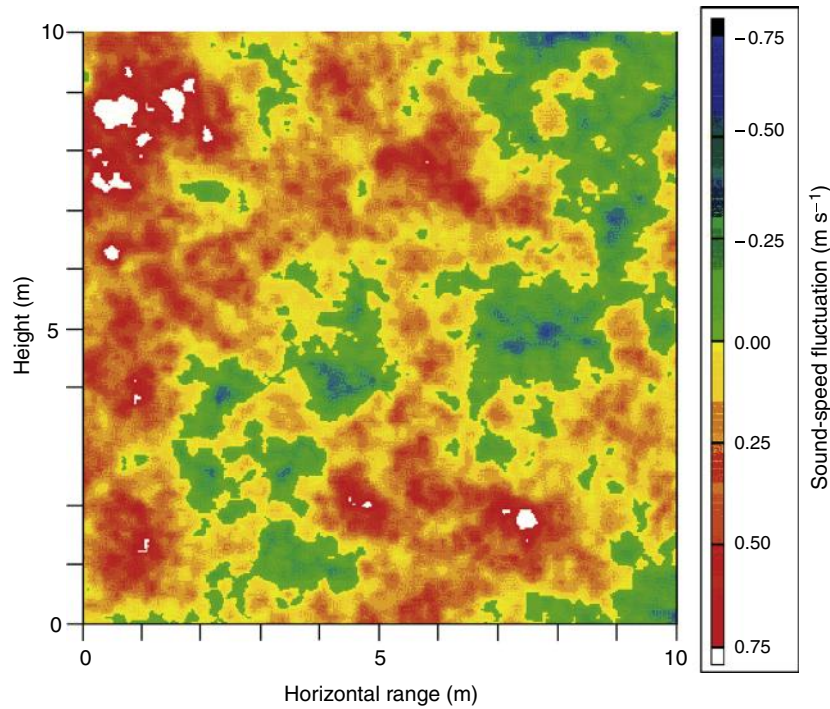


Figure 10 A two-dimensional 'snapshot' of small-scale turbulent fluctuations in a sound-speed field synthesized using a Kolmogorov spectrum and random Fourier components. The fluctuation magnitudes are typical of those created by turbulence in the daytime.

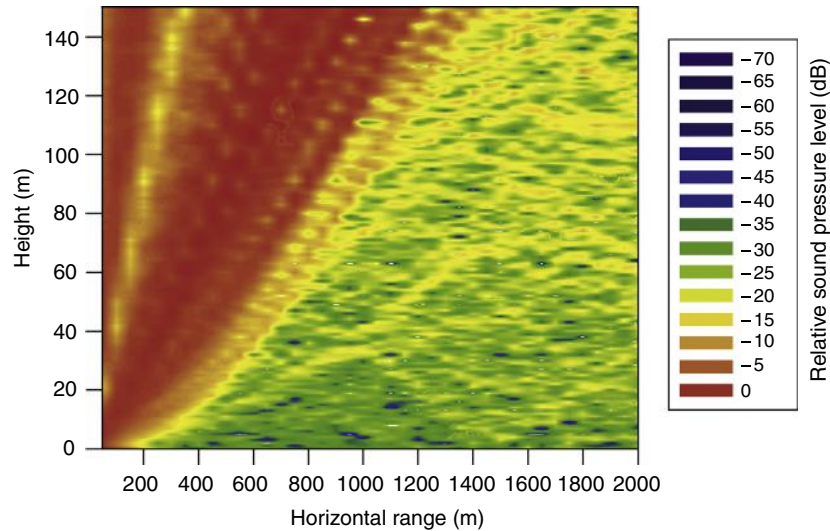


Figure 11 Same as Figure 8 except that the small-scale turbulence shown in Figure 10 is included. Note that the effects of turbulence are most apparent in the shadow region. Scattering of acoustic waves from turbulence 'fills in' the shadow region so the levels there are much higher than the no-turbulence case shown in Figure 8.

levels are high, the effect of scattering from turbulence is not as dramatic as in the shadow region above the duct (Figure 13). Further, the nocturnal boundary layer, being more stable, intrinsically has weaker turbulence than the daytime boundary layer. As a consequence, mean sound levels at night near the ground are not affected by turbulence nearly as much as the near-ground daytime levels. However, in contrast to small-scale turbulence, large-scale nocturnal phenomena such as internal gravity waves (IGW) can have a significant effect on long-range

sound propagation (1–3 km) at night. The effects of such large-scale dynamic features on sound propagation are considered next.

Acoustic Pulse Propagation through IGW in the Stable Nighttime Boundary Layer

IGW in the stable nighttime boundary layer are analogous to gravity waves in the ocean. Like the ocean, the nighttime

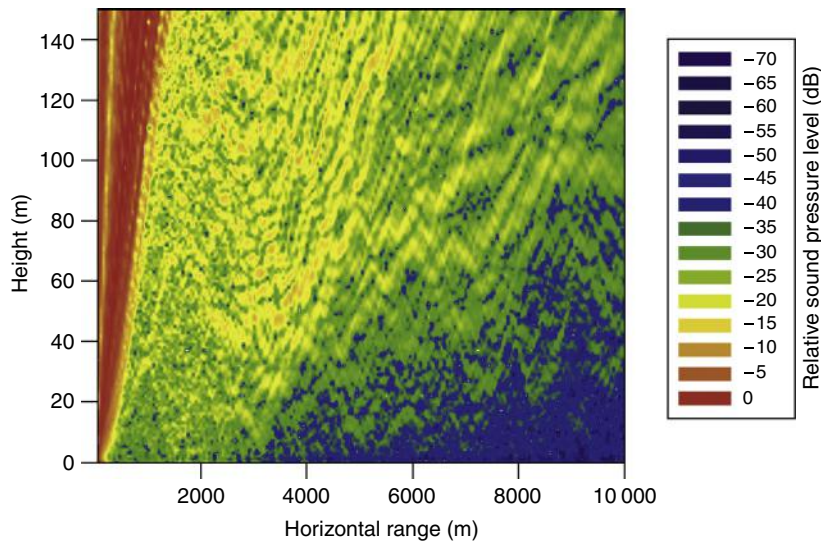


Figure 12 Same as **Figure 11** except that the range extends to 10 km. Note that even with scattering into the shadow region, daytime sound pressure levels near the ground are very low at ranges beyond a few kilometers.

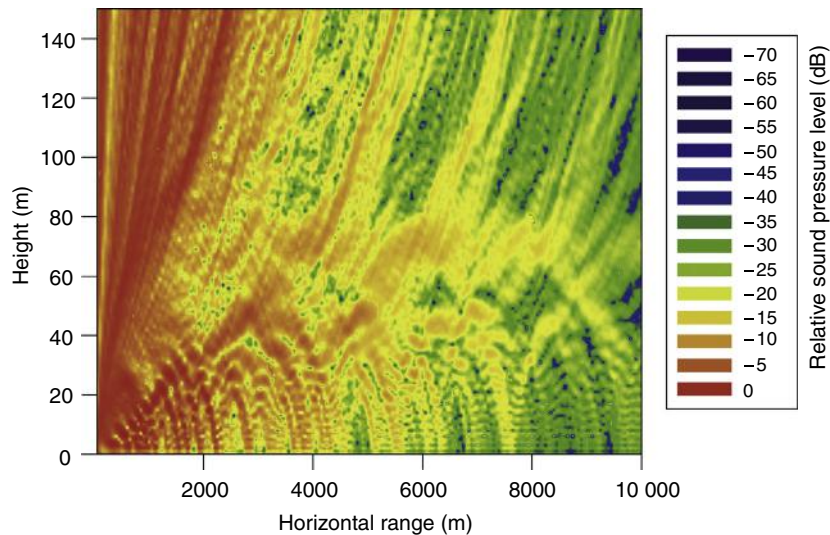


Figure 13 Same as **Figure 9** except that small-scale turbulence typical of a nighttime atmosphere has been included. Since sound levels are relatively high in the near-ground nocturnal acoustic duct, the effects of small-scale turbulence are less evident than during the daytime.

boundary layer is seldom quiescent, but generally has undulations moving through it at wave speeds typically ranging from about 1 to over 10 m s^{-1} . The wave periods of the undulations usually range from a few minutes to over an hour. These wave speeds and periods correspond to wavelengths ranging from less than a kilometer for the short-period waves to over 5 km for the long-period waves. The fluid displacements in the boundary layer caused by the IGW are primarily vertical. Although the displacement amplitude is zero at the ground surface, in the first 100 m it increases approximately linearly with height, becoming as large as 50–100 m near the top of the boundary layer.

The vertical displacements due to the IGW cause a distortion of the effective sound-speed profile, so that it varies in both space and time. Since the IGW displacement increases with

altitude, the sound-speed profile distortion also increases with increasing altitude. The temporal and horizontal spatial scales of the distortion are governed, respectively, by the periods ($\sim 1\text{--}60 \text{ min}$) and wavelengths ($\sim 0.06\text{--}10 \text{ km}$) of the IGW. Hence, compared to turbulence, the distortion of the effective sound speed profile by IGW is large-scale and slowly varying.

The effect of IGW on nighttime sound propagation is most easily seen in acoustic pulse propagation over distances of 1–3 km. Since, at these ranges, scattering by turbulence is relatively weak, the dominant time-dependent effect on pulse propagation results from the large-scale distortion of the effective sound speed profile caused by the IGW. Thus, the IGW mainly refract (i.e., bend) the sound waves rather than scattering them by diffraction. The degree of refraction varies over time as the IGW move through the nighttime boundary layer.

Figure 14 shows a received acoustic pulse from a propane cannon source at distances of 20 m and 1.7 km. The propagation to 1.7 km shown in the figure is for a reference ('mean') state of the acoustic duct that is created by a temperature inversion in the nighttime boundary layer. Note that the propagation of the pulse in the mean acoustic duct has increased ('dispersed') the pulse length from 6 ms (0.006 s) at 20 m to 100 ms (0.100 s) at 1.7 km, an increase of more than a factor of 15. As is discussed below, the length of the pulse received at 1.7 km is not constant but changes over time as the mean state of the acoustic duct is distorted by the IGW. However, before discussing these pulse length fluctuations, it is helpful to first discuss pulse propagation in the mean acoustic duct.

The pulse spreading (dispersion) in Figure 14 results from the acoustic pulse having different wave speeds over different propagation paths in the atmosphere. As indicated in Figure 15, the first arrival is the direct arrival, which travels high in the boundary layer where the sound speed is higher (note the ray paths and sound speed profiles in Figure 17). The later arrivals (single ground reflection, multiple ground reflections, and low-frequency tail) travel progressively lower in the boundary layer where the sound speed is progressively lower. The low-frequency tail, which is confined to the first 10–20 m above the ground, is slowest and arrives last. In general, the lower a ray travels in the boundary layer, the later it arrives.

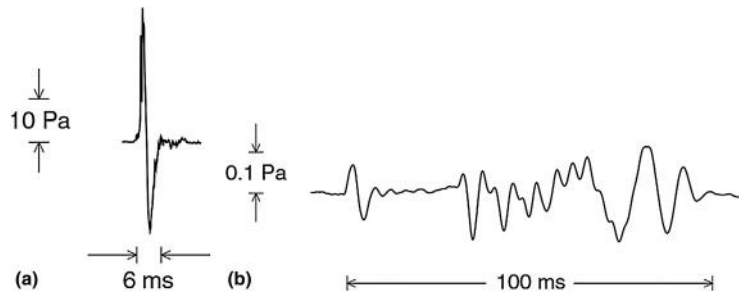


Figure 14 Pressure-versus-time records for propane cannon pulses measured at night at distances of (a) 20 m and (b) 1.7 km. Note that over the 1.7 km propagation path, the pulse length increases from 6 ms (0.006 s) to approximately 100 ms (0.100 s). Reproduced with permission from Waxler, R., Gilbert, K.E., Talmadge, C.L., 2008. A theoretical treatment of the long range propagation of impulse signals under strongly ducted nocturnal conditions. *Journal of the Acoustical Society of America* 124, 2742–2754.

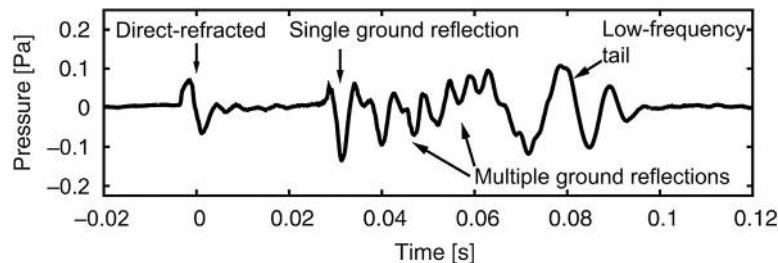


Figure 15 Pressure-versus-time waveform in Figure 14 labeled with the main pulse arrivals, starting with the first arrival and ending with a low-frequency tail (surface wave). As ray paths, the direct arrival travels highest in the boundary layer, and the low-frequency tail travels the lowest. The intermediate arrivals (single and multiple ground reflections), travel at altitudes between that of the direct arrival and the low-frequency tail. Reproduced with permission from Blom, P., Waxler, R., 2012. Impulse propagation in the nocturnal boundary layer: analysis of the geometric component. *Journal of the Acoustical Society of America* 131, 3680–3690.

The dispersed pulse train shown in Figures 14 and 15 is for a single cannon shot propagating through the mean state of the boundary layer. On the successive shots shown in Figure 16, the acoustic duct has been distorted by IGW. In the distorted duct, the overall pulse shape is maintained, but the length of the pulse changes. As the IGW propagate through the boundary layer, the overall waveform of the received pulse expands and contracts (i.e., the pulse length fluctuates) with a time scale governed by the IGW periods. Figure 16(a)–(d) shows acoustic pulse lengths measured over 33 min. Each individual figure in Figure 16(a)–(d) contains 2 min of data, with the pulses separated by 30 s, starting with the first pulse at the bottom of each figure. The time separation between each individual figure is 11 min. Figure 16(a)–(d) shows that, over 33 min, the acoustic pulse length increases from an initial length of approximately 0.08 s (80 ms) to a length of approximately 0.10 s (100 ms), an increase of 0.02 s (20 ms). Note that the pulse length does not change significantly over intervals of minutes but changes measurably over intervals of tens of minutes or more. The repeatability of the pulses over time intervals of minutes shows two things. First, there is little scattering by turbulence, and second, the acoustic pulses are not significantly affected by the short-period IGW. The variation in the pulse length over tens of minutes shows, in contrast to the short-period IGW, that the long-period IGW do significantly affect that the length of the acoustic pulses.

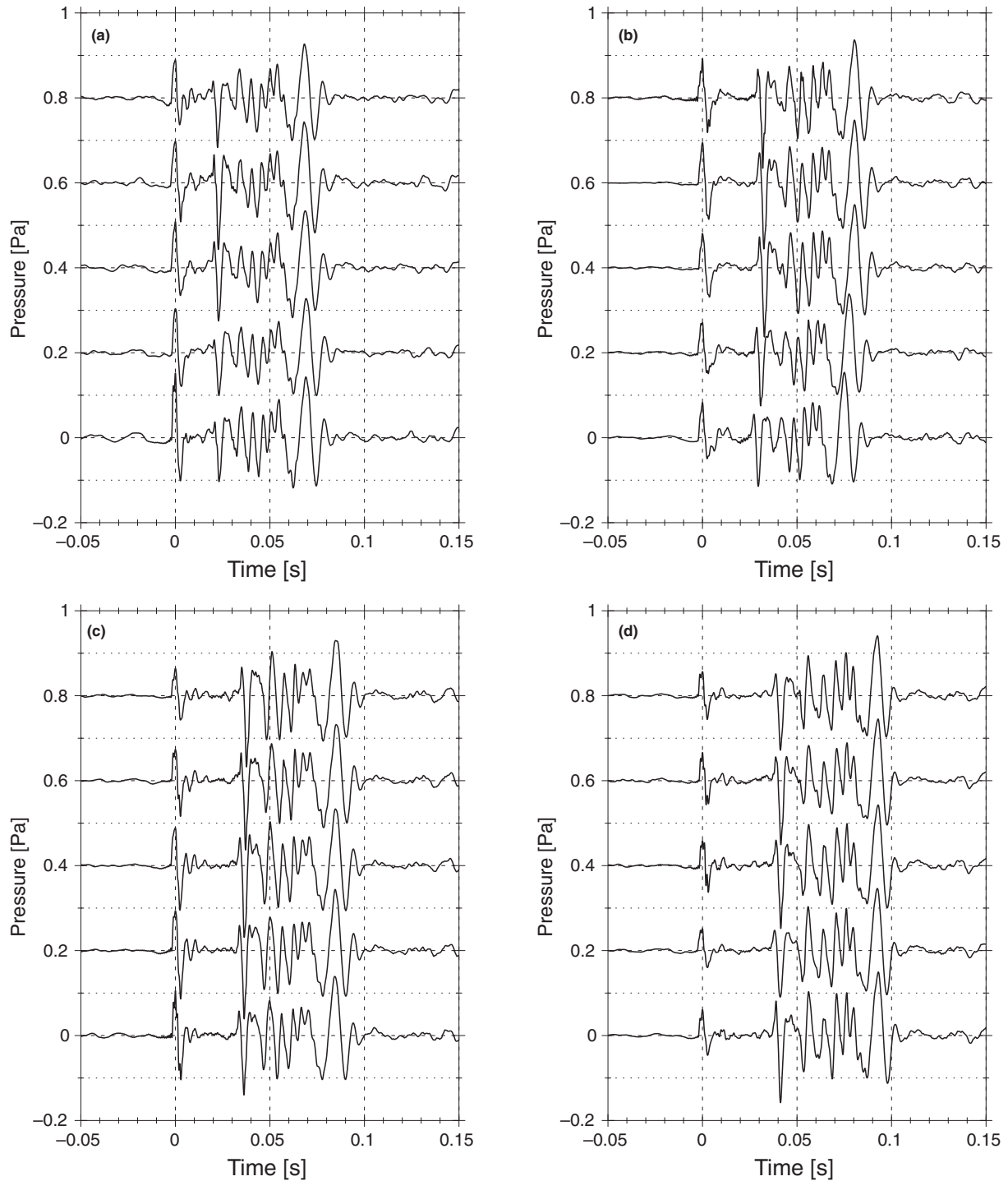


Figure 16 Pressure-versus-time records for propane cannon pulses propagating in the nighttime boundary layer. The pulse timing is shown on the horizontal axis, with the pulse always beginning at $t = 0$. Each vertical division in the figures is 0.1 Pa, and successive pulses are displaced vertically by 0.2 Pa. Each individual figure shows 2 min of data. The successive pulses in each figure are separated by 30 s, starting with the first pulse at the bottom of the figure. The four figures, (a)–(d), are each separated by 11 min, so that **Figure 16(d)** is for cannon shots made 33 min after those in **Figure 16(a)**. Note that over time intervals of minutes, the pulse lengths do not change significantly, but over time intervals of tens of minutes or greater, the pulse lengths change noticeably.

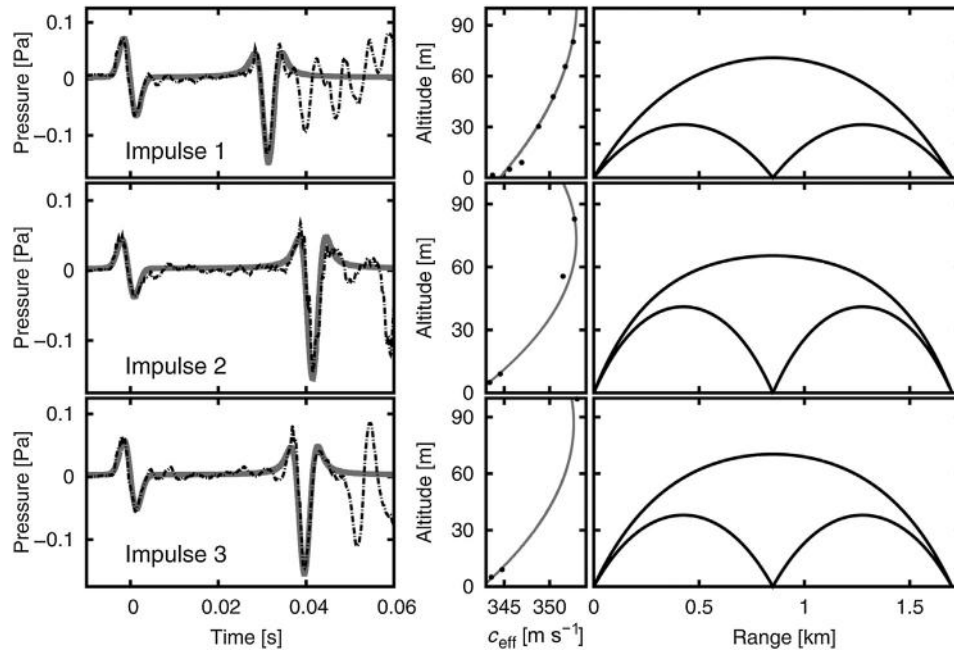


Figure 17 Analysis of the first and second arrivals for three pulses selected from the same experiment as in Figures 14–16. In the left-most panels, which show the measured pulses (dot-dash lines) and least-squares fits (solid lines), Impulse 1 has the smallest time separation between the first and second arrivals, and Impulse 2 has the largest separation. In Impulse 3, the time separation is intermediate between that of Impulse 1 and 2 and corresponds to the approximate mean state of the nighttime boundary layer. The middle panels show the effective sound speed profiles (solid lines) chosen to give the least-squares best fit to the first and second arrivals shown in the left-most panels. The measured effective sound speed profiles are shown as dots. The ray paths are shown in the right-most panels. Reproduced with permission from Blom, P., Waxler, R., 2012. Impulse propagation in the nocturnal boundary layer: analysis of the geometric component. *Journal of the Acoustical Society of America* 131, 3680–3690.

The variation of the pulse length seen in Figure 16 occurs because the IGW distort the upper part of the effective sound speed profile without significantly affecting the lower part. Consequently, the travel time of the high-altitude first arrival varies considerably more than the travel times of later arrivals. The net result is that the pulse stretches and contracts over time, with the largest variation occurring over a time scale of tens of minutes. As noted above, these long-period fluctuations of the pulse length are caused by the long-period IGW.

The above ideas are illustrated quantitatively in the detailed analysis of the first and second arrivals for the three selected pulses shown in Figure 17. The pulses were selected from data taken during the same experiment as in Figure 16. For each pulse, the effective sound speed profile was fitted using a least-squares method so that the predicted travel times matched the measured travel times. (The first arrival always starts at $t = 0$.) The solid line is the least-squares fit for the sound speed profile, and the dots are the directly measured sound speed values. Note that the smallest time separation occurs for Impulse 1, while the largest separation is for Impulse 2. Impulse 3 is intermediate between the two. The corresponding sound speed profile for Impulse 3 can be considered the reference or ‘mean’ effective sound speed profile. Since the sound speed changes are only a few meters per second, the ray paths are essentially the same for all three pulses, but the cumulative effect on travel time over a distance of 1.7 km is significant. As pointed out earlier, the variation with time of the relative travel times is due primarily to the greater distortion of the upper part of the effective sound

speed profile, which in turn mainly affects the travel time of the first arrival.

For Impulse 1, the IGW vertical displacement of the boundary layer is upward, bringing a lower sound speed (from below) upward toward the highest propagation path, thus slowing down the first arrival (highest ray path in Figure 17). Since the first arrival takes longer, its travel time is closer to that of the second arrival (lower ray path in Figure 17). As a result, the time interval between the first two arrivals for Impulse 1 is the least of the three pulses. For pulse 2, the opposite happens. The IGW displacement is downward, bringing a higher sound speed (from above) downward to the highest propagation path, speeding up the first arrival so that it arrives earlier. Thus the time separation between the first and second arrivals is increased. The upward and downward displacements of the effective sound speed profile can be seen in the least-squares fitted sound speed profiles shown in Figure 17. Note that the time interval between the first two arrivals can be directly correlated with the long-period IGW vertical displacements of the atmosphere.

The experiment discussed above (Figures 14–17) was intermittent. That is, it did not continuously track the pulse length variation for successive cannon shots. Figure 18 is for a different experiment where a propane cannon was fired continuously every 30 s for an extended period of time under stable nighttime conditions. The figure shows a time series for the total pulse length variation (relative to a reference pulse length). The propagation is over a horizontal range of approximately 2.5 km to three different receivers (denoted T1, T2, and T3). Note that, as in Figure 16, over time intervals of

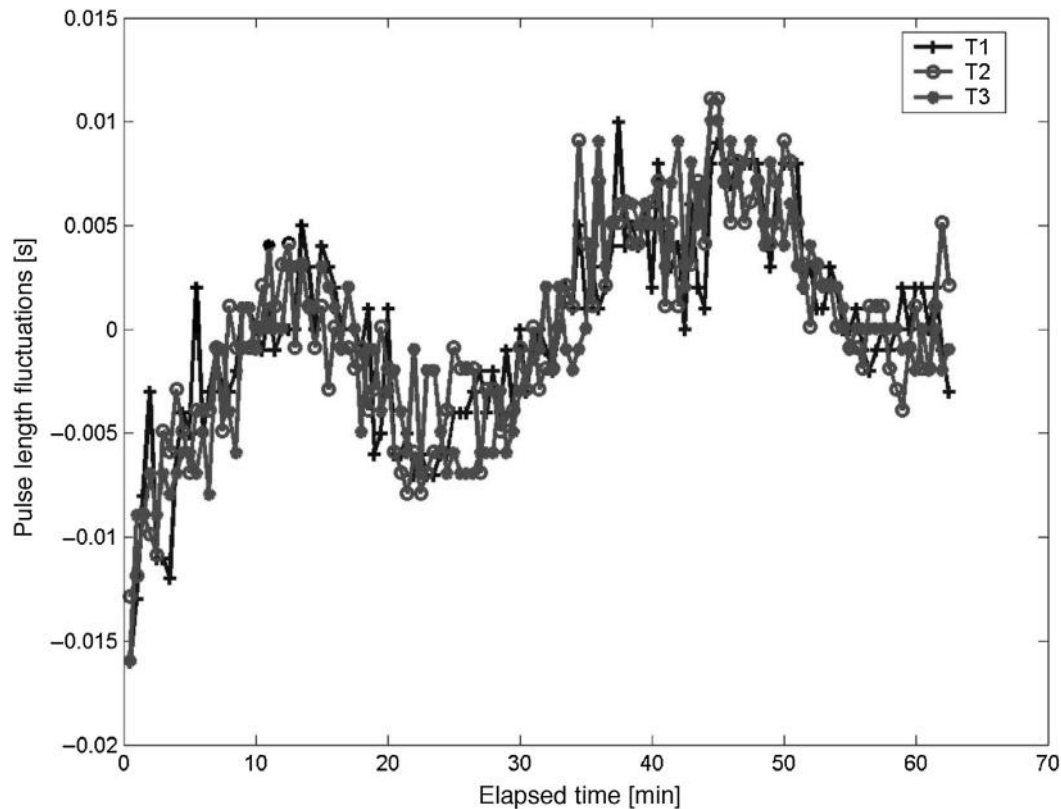


Figure 18 Time series showing the fluctuation of the total acoustic pulse length relative to a reference pulse length as a function of time. The experiment measured pulse propagation in the nighttime boundary layer over a distance of 2.5 km, with propane cannon shots made every 30 s. There were three receivers approximately equidistance from the source. The time series for the three receivers are denoted T1, T2, and T3, respectively. Note that the pulse lengths do not change significantly over time intervals of minutes, but change significantly over time intervals of tens of minutes. Reproduced with permission from Chunchuzov, I., Kulichkov, S., Otrezov, A., Perepelkin, V., 2005. Acoustic pulse propagation through a fluctuating stably stratified atmospheric boundary. *Journal of the Acoustical Society of America* 117, 1868–1879.

minutes, the change in the pulse length is relatively small, but over tens of minutes the pulse length variation is significant. The dominant period is approximately 20–40 min.

The absence of shorter period pulse length fluctuations is partly due to the IGW spectrum and partly due to a ‘filtering’ effect of the experiment itself. The IGW spectrum generally has much larger amplitudes for longer period waves (tens of minutes) than for shorter period waves (minutes). Hence, the long-period distortions of the sound speed profile are much greater than the short-period distortions and generally affect the travel time more. However, even if the IGW spectrum were completely flat, the shorter period variations in the pulse length would be suppressed due to an intrinsic ‘filtering’ effect of the experiment. When the IGW wavelength becomes less than twice the propagation distance, both upward and downward displacements are present along the propagation path. Since upward and downward vertical displacements have opposite effects on the travel time, when both are present, the travel time variations are suppressed, so that the pulse length stays relatively constant. That is, the travel time variations tend to be ‘averaged out’ for IGW wavelengths less than twice the propagation distance (i.e., for shorter period IGW). For a propagation distance of 2.5 km, for example, the response of the acoustic response would begin to be suppressed for IGW wavelengths less than 5 km. The largest suppression is for IGW wavelengths

that ‘fit’ exactly within the propagation distance. If, for instance, the wavelength is exactly the same as the propagation distance, 2.5 km, then there are equal positive and negative fluid displacements along the propagation path. In this case there are approximately equal and opposite changes in the sound speed along all the propagation paths. Since the positive and negative sound speed changes nearly average out for all the propagation paths, the response to an IGW with a 2.5 km wavelength is strongly suppressed.

In reality, the IGW displacement is always a superposition of both short and long wavelengths. However, as noted already, the pulse length responds mainly to the long-wavelength IGW, which have periods of tens of minutes. For these long-wavelength displacements, there is only an upward or downward displacement of the effective sound speed profile along the entire propagation path. Hence, due to both the IGW spectrum and the intrinsic averaging effect, the pulse length variations are dominated by the longer period IGW that have wavelengths greater than twice the propagation distance. These long-period, long-wavelength IGW are the dominant cause of the pulse length variations observed in nighttime pulse propagation experiments.

It is clear from the above discussion that the stretching and compression of an acoustic pulse can be very sensitive to changes in the nocturnal boundary layer. It was seen, for example, that the

time between the first and second arrivals was directly correlated with vertical displacement of the boundary layer along the propagation path of the pulse. Because of the sensitivity of the pulse lengths to IGW, there is the potential for extracting boundary layer information (e.g., the IGW spectrum) from pulse length fluctuations. Such research is a recent and ongoing effort in boundary layer meteorology (see [Further Reading](#) section).

Acoustic Remote Sensing of the Atmosphere

As illustrated above, the sensitivity of acoustic waves to atmospheric wind and temperature variations makes accurate prediction of ground-to-ground sound propagation a challenging problem. Conversely, however, that same sensitivity makes sound a remarkably valuable probe for remotely sensing the complex features of the atmospheric boundary layer. The most widely used acoustic tool for atmospheric sensing is an acoustic pulse-echo probe called a 'sodar,' after the more familiar 'radar,' which is an electromagnetic pulse-echo device. (Note: Sometimes the name 'echosonde' is also used, but that designation is less common than 'sodar'.)

The first sodars, which appeared in the early 1970s, emitted an acoustic pulse in a single vertically pointing beam as shown in [Figure 19](#). The sodar geometry shown, with the acoustic

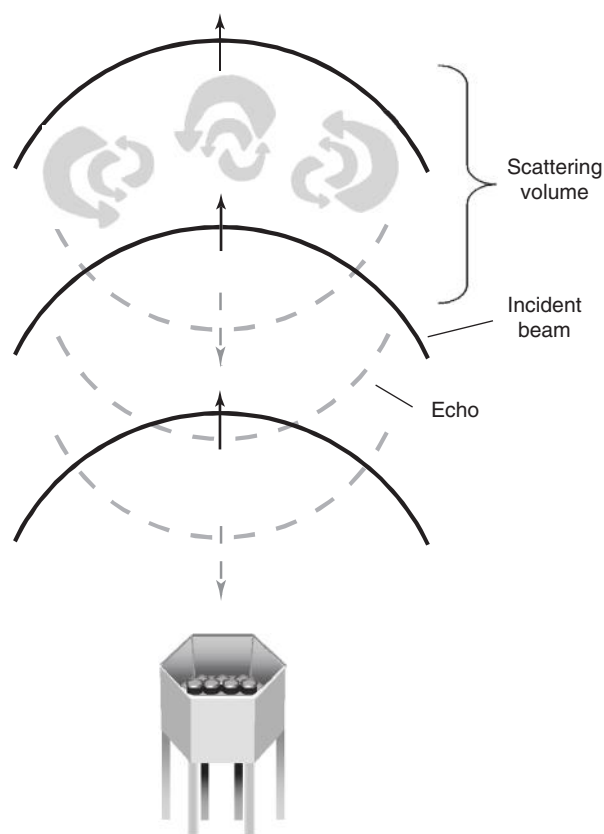


Figure 19 Geometry for a monostatic sodar. An array of transducers projects a short burst of acoustic waves vertically in a beam. Turbulence-generated temperature inhomogeneities scatter sound back toward the transmitting transducers, which act as a directional receiver for the faint echoes received on the ground.

source and receiver collocated, is common and is known as a 'monostatic' sodar. (A less common geometry has the receiver separated horizontally from the transmitter and is called a 'bistatic' sodar.) When the upward-going pulse encounters temperature inhomogeneities produced by turbulence, faint scattered waves are created within the air itself. With monostatic sodar, the part of the acoustic wave scattered back toward the ground, the echo, is detected using the same transducers that produced the probe beam.

Early sodars were used primarily as instruments for detecting turbulence. The time delay between the emitted pulse and its echo determined the height of the turbulence (one-half the time delay times the average speed of sound), while the strength of the echo was a measure of the turbulence intensity. The evolving structure of the atmospheric boundary layer could be 'mapped' by plotting the delay time and echo strength on a vertically moving strip of paper. For example, the horizontal distance on the strip could be proportional to the time delay of the echo, and the darkness could be proportional to the intensity of the echo. With many repeated pulses, the evolution of the boundary layer could be followed visually. With its debut in the 1970s, the sodar immediately provided important insights into the spatial structure and temporal evolution of the atmospheric boundary layer. A typical sodar record is shown in [Figure 20](#). The figure has time moving from left to right and shows the evolution of boundary layer structures over a typical diurnal period. The vertical scale in the figure is 0–500 m, and the thin white vertical streaks are hour markers. Panel (a) shows a typical daytime record of thermal plumes carried through the vertical sodar beam. Panel (b) shows the turbulent boundary layer descending in the late afternoon and evening as solar heating of the ground diminishes. The undulations in the latter part of the record indicate the onset of IGW. Panel (c) shows fully developed IGW activity after midnight.

In addition to visual displays of boundary layer structure and dynamics, modern sodars can provide quantitative measures of wind and temperature. The so-called 'Doppler sodar,' for example, which uses two slant beams in addition to the usual vertically pointing beam, can map vector wind velocity versus height. A typical geometry would have a vertical beam, together with slant beams pointing north and east, respectively, at 70–75° above horizontal (15–20° off vertical). Using the Doppler shift in the echoes (upshift for winds moving toward the receiver and downshift for winds moving away from the receiver), the three vector components of wind velocity (up–down, east–west, and north–south) can be measured as a function of height. Such advances as the Doppler sodar are due, in large part, to the vast increase during the past 30 years in the computing power available with small computers. In addition to providing greatly increased signal processing power, small, powerful computers have also made remote sensing instruments like the Doppler sodar sufficiently 'user friendly' that nonexperts can operate them successfully.

A second important advance in acoustic remote sensing is the 'radio acoustic sounding system,' or RASS, which can provide accurate temperature profiles as a function of height. A RASS uses a single vertically pointing sodar beam together with two radar beams that converge in the air column over the sodar. The radar is used in a bistatic geometry with the transmitter on one side of the sodar and the receiver on

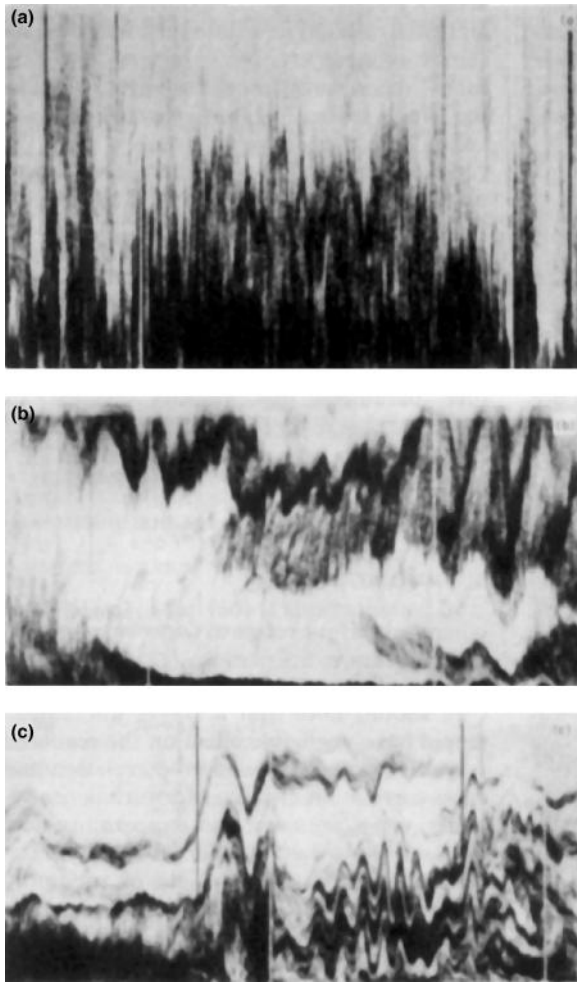


Figure 20 Record from a sodar taken over a diurnal cycle. The vertical scale is 0–500 m, and time increases from left to right. The vertical white lines are hour markers. (a) Unstable daytime boundary layer with thermal plumes generated by solar heating of the ground. (b) Decreasing daytime boundary layer followed by a growing stable nocturnal boundary layer showing evidence of initial gravity wave activity. (c) Stable nocturnal boundary layer after midnight with fully developed internal wave activity. Reproduced with permission from Bass, H.E., 1991. Atmospheric acoustics. In: Trigg, G.L., (Ed.), *Encyclopedia of Applied Physics* vol. 2. Wiley-VCH Verlag GmbH & Co. KGaA, pp. 145–179.

the other side. Using coherent radar backscatter from the upward-going acoustic beam, the RASS measures the speed of the acoustic beam as it propagates upward. After making corrections for the (small) vertical component of wind velocity, one can estimate the adiabatic sound speed (a function of temperature only) as a function of height, which then yields the temperature as a function of height. Extensive comparisons between RASS measurements and *in*

situ measurements have shown that a RASS provides reliable estimates of temperature at heights from a few hundred meters to up to several kilometers.

See also: **Boundary Layer (Atmospheric) and Air Pollution:** Convective Boundary Layer; Diurnal Cycle; Stably Stratified Boundary Layer; Surface Layer. **Gravity Waves:** Buoyancy and Buoyancy Waves: Optical Observations; Buoyancy and Buoyancy Waves: Theory. **Numerical Models:** Parameterization of Physical Processes: Turbulence and Mixing.

Acknowledgments

The author would like to thank Dr C.L. Talmadge for providing the data and the graphics of [Figure 16](#).

Further Reading

- Bass, H.E., 1991. Atmospheric acoustics. In: Trigg, G.L. (Ed.), 1991. *Encyclopedia of Applied Physics*, vol. 2. Wiley-VCH Verlag GmbH & Co. KGaA, New York, pp. 145–179.
- Blom, P., Waxler, R., 2012. Impulse propagation in the nocturnal boundary layer: analysis of the geometric component. *Journal of the Acoustical Society of America* 131, 3680–3690.
- Crocker, M., 1997. Introduction. In: Crocker, M. (Ed.), *Handbook of Acoustics*. Wiley, New York.
- Churchuzov, I., Kulichkov, S., Otrezov, A., Perepelkin, V., 2005. Acoustic pulse propagation through a fluctuating stably stratified atmospheric boundary. *Journal of the Acoustical Society of America* 117, 1868–1879.
- Hetzer, C.H., Gilbert, K.E., Waxler, R., Talmadge, C.L., 2010. Generation of microbaroms by deep-ocean hurricanes. In: Le Pichon, A., Blanc, E., Hauchecorne, A. (Eds.), *Infrasound Monitoring for Atmospheric Studies*. Springer, New York, pp. 249–262.
- Kinsler, L.E., Fry, A.R., Coppens, A.B., Sanders, J.V., 2000. *Fundamentals of Acoustics*. Wiley, New York.
- Morse, P.M., 1981. *Vibration and Sound*. Acoustical Society of America/American Institute of Physics, New York.
- Neff, W.D., Coulter, R.L., 1986. Acoustic remote sensing. In: Lenschow, D.W. (Ed.), *Probing the Atmospheric Boundary Layer*. American Meteorological Society, Boston, MA.
- Ostashev, V.E., 1997. *Acoustics in Moving Inhomogeneous Media*. E&FN Spon, London.
- Pierce, A.D., 1989. *Acoustics: An Introduction to Its Physical Principles and Applications*. Acoustical Society of America, Woodbury, NY.
- Piercy, J.E., Embleton, T.F.W., Sutherland, L.C., 1977. Review of noise propagation in the atmosphere. *Journal of the Acoustical Society of America* 61, 1403–1418.
- Salomons, E.M., 2001. *Computational Atmospheric Acoustics*. Kluwer Academic, Dordrecht.
- Stull, R.B., 1993. *An Introduction to Boundary Layer Meteorology*. Kluwer Academic, Boston, MA.
- Sutherland, L.C., Daigle, G.A., 1997. Atmospheric sound propagation. In: Crocker, M. (Ed.), *Handbook of Acoustics*. Wiley, New York.
- Waxler, R., Gilbert, K.E., Talmadge, C.L., 2008. A theoretical treatment of the long range propagation of impulse signals under strongly ducted nocturnal conditions. *Journal of the Acoustical Society of America* 124, 2742–2754.

Atmospheric Tides

J Oberheide, Clemson University, Clemson, SC, USA

ME Hagan and AD Richmond, National Center for Atmospheric Research, Boulder, CO, USA

JM Forbes, University of Colorado, Boulder, CO, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by M Hagan, J Forbes, A Richmond, volume 1, pp 159–165, © 2003, Elsevier Ltd.

Synopsis

Atmospheric tides are planetary-scale wave motions in the neutral atmosphere with periods defined by Earth's rotation rate. They are among the most striking dynamical features in the mesosphere and thermosphere, and are known to redistribute ionospheric plasma through dynamo processes. This article reviews the salient features of the tides from theoretical and observational perspectives, including the recent discovery that atmospheric tides connect planetary-scale weather patterns with the 'space weather' of the ionosphere-thermosphere.

Introduction

Atmospheric tides are ubiquitous features of the Earth's atmosphere. They are the persistent global oscillations that are observed in all types of atmospheric fields, including wind, temperature, pressure, density, and geopotential height. Tidal oscillations have periods that are some integer fraction of a solar or lunar day. The solar diurnal and semidiurnal tides have 24 and 12 h periods, respectively. The lunar diurnal tidal period is about 24.8 h, while the lunar semidiurnal period is 12.4 h. Scientists often use a shorthand notation to represent solar and lunar tides. S1 and S2 refer respectively to the solar diurnal and semidiurnal tides. Their lunar counterparts are M1 and M2.

Atmospheric tides have been studied for many years, since they are evident in both surface pressure and magnetic observations that date back to the early part of the twentieth century. **Figure 1** illustrates a time series of surface pressure measurements made at Batavia (now known as Jakarta, Indonesia) during the first 5 days of January in 1925. The dominant feature of this time series provides evidence of the solar semidiurnal atmospheric tide. Specifically, there is a 1 to 2-hPa deviation from the average pressure of about 1011 hPa that occurs regularly at 12 h intervals. This semidiurnal variation is modulated by other variations, but the former is such a persistent oscillation that the semidiurnal tide is also the dominant oscillation in monthly, yearly, and even multiyear averages of daily surface pressure measurements made at Batavia.

Atmospheric tides are further characterized by their sources. The Moon's gravity forces the lunar atmospheric tide, while solar

atmospheric tides can be excited in several ways, including the absorption of solar radiation, large-scale latent heat release associated with deep convective clouds in the troposphere, the gravitational pull of the Sun, and as secondary waves due to nonlinear wave-wave interactions. The restoring force that acts on atmospheric tides is gravity, so tides are a special class of buoyancy or gravity waves. Unlike high-frequency gravity waves, tides are affected by the Earth's rotation and sphericity because of their comparatively large periodicities and horizontal scales. Solar atmospheric tides are generally larger than lunar tides and dominate the tidal motions in the middle and upper atmosphere, that is, the stratosphere, mesosphere, and thermosphere. **Movie 1** illustrates the combined diurnal and semidiurnal tidal motions caused by solar atmospheric tides in the lower thermosphere. Temperature and wind speeds can vary by more than 60 K and $>100 \text{ m s}^{-1}$ within a few hours.

The general mathematical expression for a tidal oscillation is given by **eqn [1]**, where A is the magnitude of the variation in some atmospheric field, σ is its frequency, t is universal time, λ is longitude, and $s \geq 0$ is the zonal wavenumber (the number of wave crests that occur along a latitude circle). The $(s\lambda - \sigma t)$ form of **eqn [1]** is chosen so that the sign of σ is indicative of the zonal direction of propagation: $\sigma > 0$ corresponds to eastward propagating waves and $\sigma < 0$ to westward propagating waves. φ is the so-called tidal phase. A crest of the wave occurs when **eqn [2]** is satisfied.

$$A \cos(s\lambda - \sigma t - \varphi) \quad [1]$$

$$\varphi = s\lambda - \sigma t \quad [2]$$

The horizontal phase speed of the tide, c_{ph} , is defined by differentiating **eqn [2]** (**eqns [3] and [4]**).

$$sd\lambda - \sigma dt = 0 \quad [3]$$

$$c_{ph} \equiv \frac{d\lambda}{dt} = \frac{\sigma}{s} \quad [4]$$

For solar tides, the m th harmonic frequency is $\sigma_m = m\sigma_1$, where m is a positive or negative integer and $\sigma_1 = (2\pi/24) \text{ h}^{-1}$. Rewriting the mathematical expression for a tide in terms of LT (hours), $t_L = t + \lambda/\sigma_1$, results in a mathematical expression of the form of **eqn [5]**.

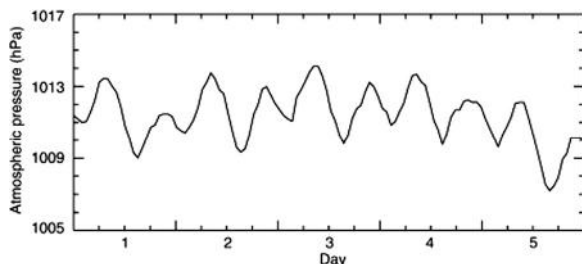


Figure 1 Surface pressure (hPa) at Batavia (Jakarta, Indonesia) against time during the first 5 days of January 1925.

$$A \cos((s + m)\lambda - \sigma_m t_L - \varphi) \quad [5]$$

For the subset of atmospheric tides known as migrating solar tides, $s = -m$ (with $m < 0$) and eqn [5] reduces to eqn [6].

$$A \cos(|m|\sigma_1 t_L - \varphi) \quad [6]$$

Thus, migrating solar tides have the same local time variation at all longitudes. If $m = -1$ and $s = 1$, the tide is diurnal and moves or migrates westward in longitude with the apparent motion of the Sun from the perspective of a ground-based observer. Further, $c_{ph} = -(2\pi/24) \text{ h}^{-1}$. Similarly, if $m = -2$ and $s = 2$, then the wave is a migrating semidiurnal tide. The remaining set of global-scale waves with tidal periods that are not Sun-synchronous are known as nonmigrating tides. Nonmigrating tides may be viewed as waves that propagate to the west more rapidly or slowly than the Sun, or that propagate eastward, or that are standing. Migrating and nonmigrating solar tides are often identified by using a letter/number code that indicates frequency, propagation direction, and zonal wavenumber: DWs or DEs is a westward or eastward propagating diurnal tide, respectively, with positive zonal wavenumber s . For semidiurnal tides, D is replaced by S, and D0, S0 are standing diurnal and semidiurnal tides, respectively. With this nomenclature, the migrating diurnal (semidiurnal) tide is DW1 (SW2) and DE3, for example, is an eastward propagating diurnal tide of zonal wavenumber 3.

All tides contain components that propagate in the vertical direction z . The effects of upward-propagating tidal components are particularly important because these waves grow in amplitude $\sim \exp(z/2H)$ with scale height $H (=k_B T/Mg)$ where k_B is the Boltzmann constant, T is temperature, M is the mean molecular mass, and g is Earth's gravity acceleration) as they conserve energy in an atmosphere whose density decreases with increasing altitude. Thus, tides with insignificant amplitudes in their lower atmospheric regions of excitation often affect the upper atmosphere profoundly because they introduce large atmospheric variations with local time and because they may dissipate and deposit their energy and momentum therein.

Classical Tidal Theory

Classical tidal theory treats the tides as perturbations on a basic state with neither mean-flow nor horizontal temperature gradients in an inviscid atmosphere. It provides a reasonable description of atmospheric tides in the lower and middle atmosphere, including the mesosphere and is therefore quite useful to demonstrate important tidal characteristics. As described in detail by Chapman and Lindzen (1970), the linearized primitive equations (Dynamical Meteorology: Primitive Equations) for wave motions described by eqn [1] and given (s, σ) can be reduced to a single equation for, for example, geopotential. The resulting equation is separable in its latitude and altitude dependence. The latitudinal part is described by Laplace's tidal equation and solved by a complete orthogonal set of eigenfunctions (called Hough modes) and eigenvalues or separation constants (called equivalent depths).

Each Hough mode $\Theta_n^{s,\sigma}$ is a series of associated Legendre polynomials with $|n| \geq s$ being the so-called meridional index

because it provides information on the number of latitudinal nodes and symmetry characteristics. It is quite common to refer to a specific Hough mode as the Θ_n^s mode or simply the (s, n) mode and to provide the frequency information externally. The equivalent depth $h_n^{s,\sigma}$ determines the vertical structure of each Hough mode because it is linked to the vertical wavelength λ_z (eqn [7]),

$$\lambda_{z,n}^{s,\sigma} = \frac{2\pi H}{\sqrt{\left(\kappa H + \frac{dH}{dx}\right) \frac{h_n^{s,\sigma}}{H} - \frac{1}{4}}} \quad [7]$$

with scale height H , $\kappa = (\gamma - 1)/\gamma$ and adiabatic index $\gamma \approx 7/5$, and normalized height $x = z/H$. Equation [7] implies that equivalent depths must be positive but smaller than roughly 8 km for vertical wave propagation. Larger values or negative equivalent depths imply vertically trapped modes. Figure 2 illustrates Hough modes corresponding to the first two propagating and trapped modes of the migrating DW1 and nonmigrating DE3 tides, respectively. Diurnal propagating modes generally maximize at low to middle latitudes and trapped modes at middle to high latitudes.

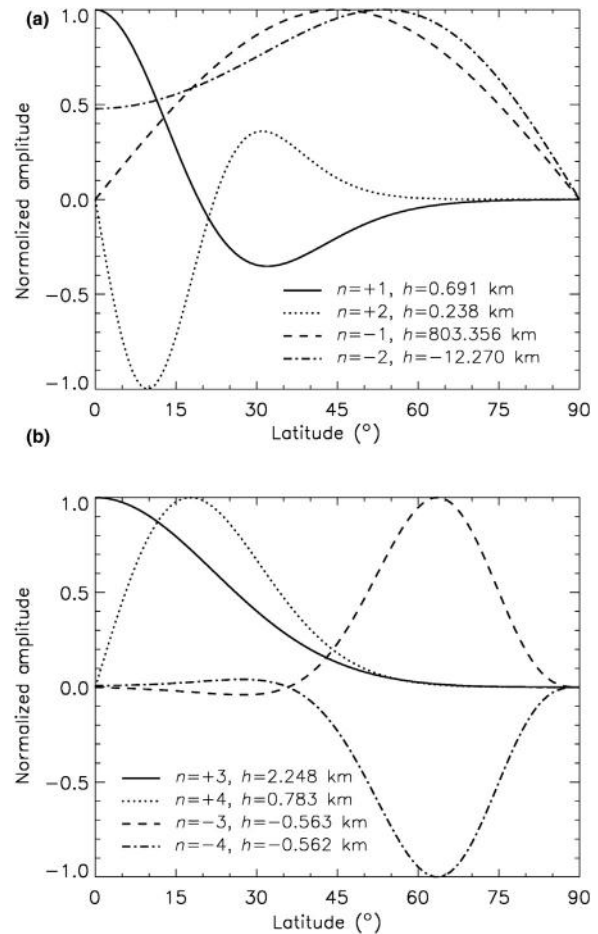


Figure 2 (a) Hough modes for the migrating diurnal solar tide (DW1). The meridional index n is positive for propagating tides and negative for trapped tides. (b) Same as (a) but for the DE3 nonmigrating tide.

Symmetric Hough functions are mirror images about the equator and occur if $n + s$ is even (odd) for positive (negative) values of n . Antisymmetric modes change sign at the equator and occur if $n + s$ is odd (even) for positive (negative) n . Tidal variations δF in temperature, pressure, geopotential, density, and vertical wind as function of normalized height x , latitude ϑ , zonal wavenumber s , and frequency σ are described by eqn [8]

$$\delta F^{s,\sigma}(x, \vartheta) = \sum_n \delta F_n^{s,\sigma}(x) \Theta_n^{s,\sigma}(\vartheta) \quad [8]$$

Zonal (u) and meridional (v) wind variations are described by eqns [9] and [10],

$$\delta u^{s,\sigma}(x, \vartheta) = \sum_n \delta u_n^{s,\sigma}(x) U_n^{s,\sigma}(\vartheta) \quad [9]$$

$$\delta v^{s,\sigma}(x, \vartheta) = \sum_n \delta v_n^{s,\sigma}(x) V_n^{s,\sigma}(\vartheta) \quad [10]$$

with the wind expansion functions shown in Figure 3 (eqns [11] and [12])

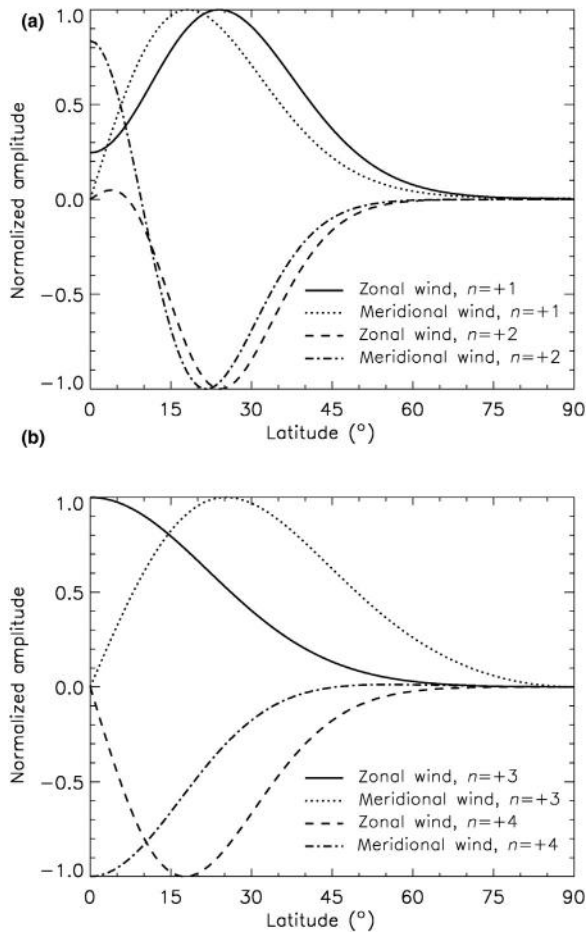


Figure 3 (a) Wind expansion functions for the migrating diurnal solar tide (DW1). The first two propagating modes are shown. (b) Same as (a) but for the DE3 nonmigrating tide.

$$U_n^{s,\sigma}(\vartheta) = \frac{1}{(f^2 - \sin^2 \vartheta)} \left[\frac{s}{\cos \vartheta} + \frac{\sin \vartheta}{f} \frac{d}{d\vartheta} \right] \Theta_n^{s,\sigma}(\vartheta) \quad [11]$$

$$V_n^{s,\sigma}(\vartheta) = \frac{1}{(f^2 - \sin^2 \vartheta)} \left[\frac{s \tan \vartheta}{f} + \frac{d}{d\vartheta} \right] \Theta_n^{s,\sigma}(\vartheta) \quad [12]$$

with $f = \sigma/(2\Omega)$ and Earth's angular velocity Ω . The zonal wind expansion functions $U_n^{s,\sigma}$ have the same symmetry as the corresponding Hough modes. $V_n^{s,\sigma}$ on the other hand are symmetric (antisymmetric) when the corresponding Hough modes are antisymmetric (symmetric).

The classical tidal theory approximates the tidal motions in the lower and middle atmosphere reasonably well, including the mesosphere. Classical methods of computing tides, however, do not work when mean zonal winds or dissipation are included, because the mathematical solutions become inseparable in the latitudinal and vertical coordinates. Hough modes are no longer eigenfunctions of the system and numerical solutions are needed. This is particularly important in the thermosphere where the tides undergo a substantial change in their modal structure when molecular dissipation becomes important. This transition height occurs approximately where the dissipative time scale equals the scale height divided by the vertical group velocity. Amplitudes and phases relax to approximately constant values in the thermosphere and the damping significantly broadens the horizontal structure. It should also be noted that the time constants of eddy and molecular diffusion are proportional to the square of λ_z . Short vertical wavelength tides or modes therefore dissipate more rapidly and cannot propagate into the thermosphere at all.

Migrating Solar Tides

The absorption of radiation by a longitudinally invariant atmosphere is the primary source of migrating solar tides. Owing to the rotation of the Earth, this absorption is periodic in time from the perspective of the ground-based observer. The resultant heating gives rise to migrating tidal oscillations. Solar radiation is absorbed throughout the Earth's atmosphere, thereby exciting migrating solar tides at almost all altitudes. Atomic oxygen, which is the most abundant atmospheric constituent at altitudes about 150 km above the Earth's surface, absorbs the shortest-wavelength solar radiation, known as the extreme ultraviolet. Increasingly longer wavelengths are absorbed as the solar radiation approaches the Earth's surface. Molecular oxygen (O_2) absorbs the far-ultraviolet radiation (100–200 nm) at altitudes near about 100–150 km, and ozone (O_3) absorbs the 200 to 300-nm solar ultraviolet radiation at middle atmospheric altitudes between about 30 and 70 km. Solar infrared radiation may be absorbed by water vapor (H_2O) in the lowest part of the atmosphere.

Even though there is little, if any, tidal forcing due to solar heating in the upper mesosphere (~ 80 – 100 km), measurements of winds and temperatures exhibit strong tidal signatures in this region. Figure 4 illustrates an example of the magnitude of the mean winds and the tidal oscillations over Adelaide, Australia, at these altitudes. The data points represent the eastward winds that were measured with the Buckland Park radar during 2 days in August 1994. The dashed curves

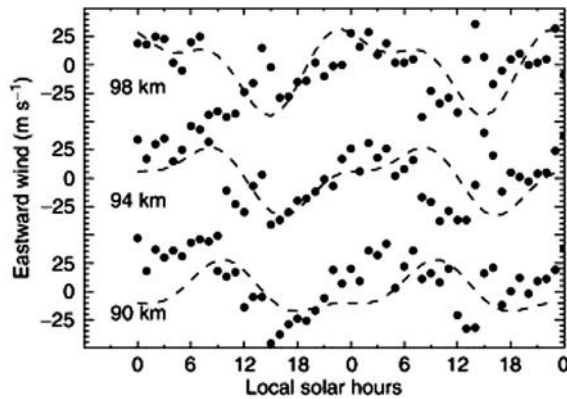


Figure 4 Eastward winds (m s^{-1}) over Adelaide, Australia, against local time (h) on 8–9 August 1994 at 98 km (top), 94 km (middle), and 90 km (bottom). Data points are radar measurements and dashed curves are model predictions that include the migrating diurnal and semidiurnal tidal components. Professor R.A. Vincent provided the Adelaide radar data.

illustrate global-scale wave model (GSWM) tidal predictions for August at the location of Adelaide. While the GSWM differs from the measurements in detail, the model captures the salient features of the observed winds, particularly on 8 August. Differences may be attributable to small-scale waves that are not included in GSWM or to sources of day-to-day tidal variability that are also omitted.

The GSWM predictions include mean winds ($17\text{--}18 \text{ m s}^{-1}$) and both migrating diurnal and semidiurnal components. Notably, the migrating diurnal amplitudes ($20\text{--}30 \text{ m s}^{-1}$) are larger than the mean winds. The GSWM diurnal tidal amplitudes are at least a factor of 2 larger than the semidiurnal amplitudes ($8\text{--}15 \text{ m s}^{-1}$) and the phases of both components shift to earlier times with increasing altitude between 90 and 98 km. This behavior, which is known as downward phase progression, is indicative of upward-propagating wave energy. There are clear signatures of downward phase progression in the GSWM predictions in **Figure 4**. That is, the wind predictions are most westward near 15.00 h at 98 km and there are similar features at progressively later times and progressively lower altitudes. The vertical wavelength of the migrating diurnal tide over Adelaide is much shorter than that of the migrating semidiurnal tide, so the phase of the former progresses far more rapidly than the phase of the latter and their

combined effects result in a pattern of wave maxima and minima that evolves between altitudes.

The migrating diurnal tide below the mesopause (i.e., the region between the mesosphere and thermosphere) originates primarily in the troposphere. Although tropospheric semidiurnal forcing is nonnegligible, there is comparatively more semidiurnal forcing in the middle atmosphere. Thus, the diurnal tidal growth occurs over a deeper altitude region than the semidiurnal growth and it is reasonable to anticipate a diurnal amplitude that is larger than the semidiurnal amplitude in the upper mesosphere. The aggregate characteristics of the mean winds and tides that are illustrated in **Figure 4** support the claim that upward-propagating migrating tides govern the large-scale dynamics of the upper mesosphere.

Migrating tides exhibit somewhat complicated behavior in that the latitudinal structure of the horizontal wind oscillations is dramatically different from the temperature, pressure, or vertical velocity structure. For example, the upward-propagating migrating diurnal tide DW1 is characterized by a primary temperature amplitude maximum over the Equator with secondary maxima near $\pm 30^\circ$. The horizontal wind amplitudes (**Movie 2**) exhibit minima over the Equator and nearly symmetric amplitude peaks at low to middle latitudes ($\pm(20\text{--}30^\circ)$). While ground-based observations provide an important perspective on the local behavior of waves with tidal frequencies, it is impossible to decipher global structures from local structures without conducting correlative analysis of measurements made at multiple locations over a broad range of latitudes. Further, in order to distinguish migrating from nonmigrating tidal components, it is necessary to have a longitudinal distribution of measurements.

The upward-propagating migrating tides dissipate in the lower thermosphere and their contribution to upper thermosphere variability is comparatively small. However, the longitudinally invariant absorption of solar far and extreme ultraviolet radiation efficiently forces a large *in situ* component of the migrating tides that dramatically changes the temperature, density, and wind structure in the upper thermosphere. **Figure 5** illustrates its magnitude based on the output from the NRLMSISE-00 (i.e., the 2000 version of the Naval Research Laboratory Mass Spectrometer Incoherent Scatter Radar Extended) empirical model. Local time mass density variations at 12:00 Universal Time are on the order of a factor of 3 when the tides (diurnal, semidiurnal, terdiurnal) are included (left panel) but almost nonexistent when the

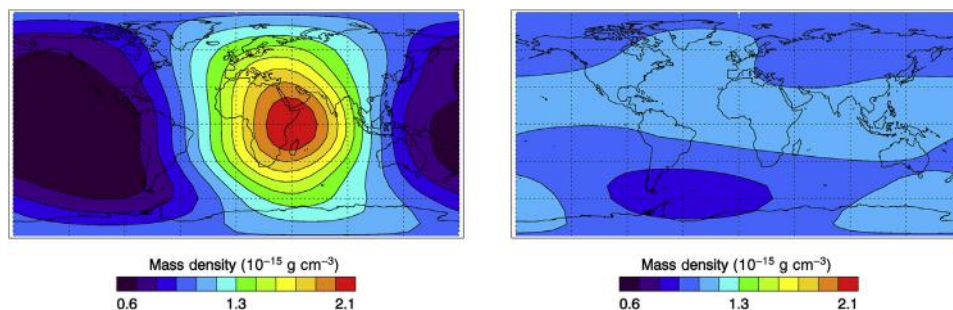


Figure 5 Global distribution of neutral mass density at 12:00 Universal Time and 400 km altitude on 21 September 2010 with (left) and without (right) thermospheric tides. Data shown are from the NRLMSISE-00 model and do not include nonmigrating tides.

tides are excluded (right panel). The largest tidal signal in [Figure 5](#) (left panel) comes from the migrating diurnal tide DW1. It maximizes at the equator (latitude of the subsolar point during equinoxes) and decreases toward higher latitudes along with the noontime solar angle, as expected for an *in situ* forced tide. NRLMSISE-00 does not include nonmigrating tides. Recent diagnostics of upper thermosphere neutral density observations from the CHALLENGING Minisatellite Payload (CHAMP) satellite indicate that nonmigrating tides induce an additional longitudinal variability on the order of 30–50%, mostly due to upward-propagating tides from the troposphere. An accurate description of migrating and nonmigrating diurnal density variations is especially important for predicting low perigee satellite trajectories because atmospheric drag is the dominant error source in operational models.

Nonmigrating Tides

Near the surface of the Earth the strong longitudinal differences in topography, land–sea contrast, and surface interactions produce zonal (i.e., along a latitude circle in the east–west direction) variations in the local time behavior of the atmosphere and thus excite nonmigrating tides. A good example is latent heat release due to condensation in large-scale deep convective systems in the tropical troposphere. [Figure 6](#) illustrates the latitude–longitude distributions of the diurnal (left) and semidiurnal (right) latent heat release amplitudes during the month of September derived from satellite-borne convective rainfall measurements. Both components maximize at low latitudes where the absorption of solar radiation (evaporation) is greatest. Their longitudinal structure reflects the areas of largest deep convective activity in the tropical troposphere: one peak over Africa, followed by two peaks over Indonesia and the western Pacific and a fourth over South and Central America.

Through Fourier analysis the longitude variations depicted in [Figure 6](#) may be decomposed into a series of wave components with different zonal wavenumbers s . Each of the resultant nonmigrating tidal components possesses different vertical propagation characteristics that depend on its sensitivity to the prevailing winds and its vertical wavelength. Good examples are the DW5 and DE3 components. Both are efficiently forced by latent heat release but DW5 dissipates at higher altitudes

due to its short vertical wavelength. DE3 on the other hand has a long vertical wavelength and propagates up into the lower thermosphere where its magnitude rivals the radiatively forced migrating diurnal tide ([Movie 2](#)).

It is difficult to track vertical tidal propagation into the upper atmosphere such as that depicted in [Movie 2](#) because the distribution of ground-based observations is spatially limited. This is particularly true for altitudes between about 30 and 180 km above the Earth's surface, which encompass the region where the upward-propagating waves attain large amplitudes and subsequently dissipate. It is generally impossible to distinguish migrating from nonmigrating tides in the analysis of ground-based remote-sensing measurements from a single site made at these altitudes. During the 1990s remote sensing from the Upper Atmosphere Research Satellite (UARS) considerably ameliorated this problem. Further progress came from the Thermosphere Ionosphere Mesosphere Energetics and Dynamics (TIMED) satellite that was launched in 2001 because it allowed for the first time to observe the global tidal spectrum in various parameters over a range of mesosphere and lower thermosphere altitudes. The data utilized to construct [Movies 1–4](#) consist of wind measurements made by the TIMED Doppler Interferometer (TIDI) and temperature measurements made by the Sounding the Atmosphere using Broadband Emission Radiometry instrument that are extended toward the poles and into the lower atmosphere and upper thermosphere using an empirical tidal model.

[Movies 3 and 4](#) illustrate the latitude–longitude distribution of the diurnal and semidiurnal tides at 100 km for the month of September, averaged over 7 years from 2002 to 2008. The superposition of these movies is shown in [Movie 1](#). The diurnal tide with amplitudes as large as 28 K and 42 m s^{-1} , respectively, mainly consists of the migrating component DW1 and the nonmigrating DE3 tide, with some contributions from the DE2, D0, and DW2 components. The migrating tide is observed as a zonally symmetric oscillation because the movie is animated in LT (compare [eqn \[5\]](#)) and the DE3 as a 4-peaked longitudinal variation for the same reason. The semidiurnal tidal field is more symmetric than the diurnal one with amplitudes of 17 K and 36 m s^{-1} , respectively. Spectral analysis of these particular data reveals that the longitudinal structure is dominated by the migrating tide SW2, with contributions from the SW3, SW1, and SE2 nonmigrating tides.

Complementary numerical modeling studies suggest that the diurnal and semidiurnal nonmigrating tides are

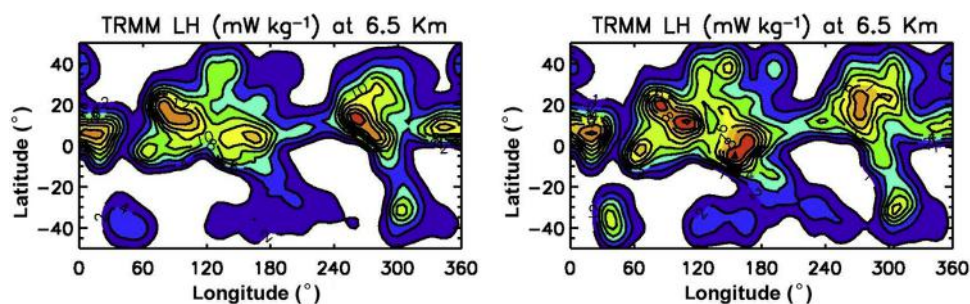


Figure 6 Contours of diurnal (left) and semidiurnal (right) latent heat release amplitudes in the troposphere from Tropical Rainfall Measuring Mission (TRMM) satellite observations during September. Dr X. Zhang provided these figures, which are adapted with permission from Zhang, X., J.M., Forbes, M.E., Hagan, 2010. Longitudinal variation of tides in the MLT region: 2. Relative effects of solar radiative and latent heating. *Journal of Geophysical Research* 115, A06317. doi:10.1029/2009JA014898.

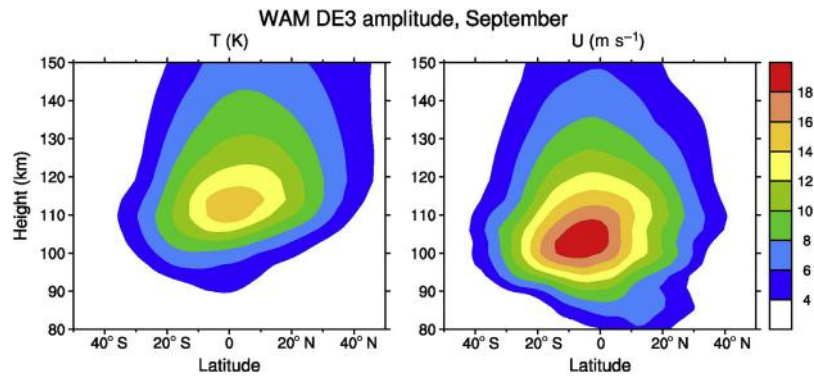


Figure 7 WAM DE3 amplitude of (left) temperature and (right) zonal wind as a function of latitude and height in September. Dr R.A. Akmaev provided the model results, which are adapted with permission from Akmaev, R.A., T.J., Fuller-Rowell, F., Wu, J.M., Forbes, X. Zhang, A.F., Anghel, M.D., Iredell, S., Moorthi, H.-M., Juang, 2008. Tidal variability in the lower thermosphere: comparison of whole atmosphere model (WAM) simulations with observations from TIMED. *Geophysical Research Letters* 35, L03810. doi:10.1029/2007GL032584.

generated in the lower levels of the atmosphere, either by latent heat release (DE3, DE2, SE2) and/or as secondary waves by the nonlinear interaction between stationary planetary waves and the migrating tides (DW2, D0, SW1, SW3). An example from the 'Whole Atmosphere Model' (WAM) is shown in [Figure 7](#). WAM is a general circulation model of the neutral atmosphere built on an existing operational Global Forecast Model used by the U.S. National Weather Service. It includes realistic topography, latent heating associated with tropospheric convection and nonlinear processes, and the model domain extends well into the dissipative thermosphere to a top level of about 600 km. The WAM temperature and zonal wind components associated with the DE3 tide exhibit maxima centered around the equator that can also be discerned in the TIMED satellite measurements shown in [Movie 2](#).

Lunar Tides

Lunar atmospheric tides are only about 5–10% as large as solar tides, but they have clearly detectable effects. The lunar tidal pressure at the ground maximizes at low latitudes, with an average amplitude of about 7 Pa. The corresponding wind amplitude at the Equator is about 0.03 m s^{-1} . The wind amplitude increases with altitude up to about 110 km, where it reaches an amplitude on the order 10 m s^{-1} .

Unlike the solar tides, the lunar atmospheric tides are entirely driven by gravitational forces, as illustrated schematically in [Figure 8](#). Because the lunar gravitational acceleration decreases as the inverse square of the distance from the center of the Moon, this acceleration is not exactly uniform near the Earth, so that atmospheric air parcels at various locations around the Earth experience slightly different lunar accelerations from those of the Earth as a whole. Air parcels in the hemisphere most distant from the Moon are accelerated toward the Moon less strongly than is the Earth, in effect creating a relative acceleration away from the Moon for these air parcels, in the Earth's reference frame. Conversely, air parcels in the moonward hemisphere of the Earth experience a relative acceleration toward the Moon. In each hemisphere, parcels to the west of a line passing through the centers of the Earth and Moon experience an eastward

component of acceleration, while those to the east of this line experience a westward acceleration. As the Earth rotates, during a lunar day (24.8412 h on the average) an air parcel at the Equator successively passes twice through regions of westward and eastward acceleration, comprising two lunar semidiurnal cycles of period 12.4206 h. When the Moon is north or south of the Earth's Equator, an additional diurnal lunar cycle (period 24.8412 h) of acceleration exists at nonequatorial latitudes. There is also a monthly periodicity to the forcing as the Moon cycles between the Northern and Southern Hemispheres of the Earth. Both this cycling and the ellipticity of the Moon's orbit create amplitude and frequency modulation of the lunar semidiurnal and diurnal forcings that can be expressed as combinations of multiple closely spaced periods. The dominant lunar period, representing the average lunar semidiurnal tide, is referred to as the M2 tide, with a 12.4206 h period.

In addition to the direct forcing of lunar gravity on the atmosphere, lunar atmospheric tides are indirectly forced by

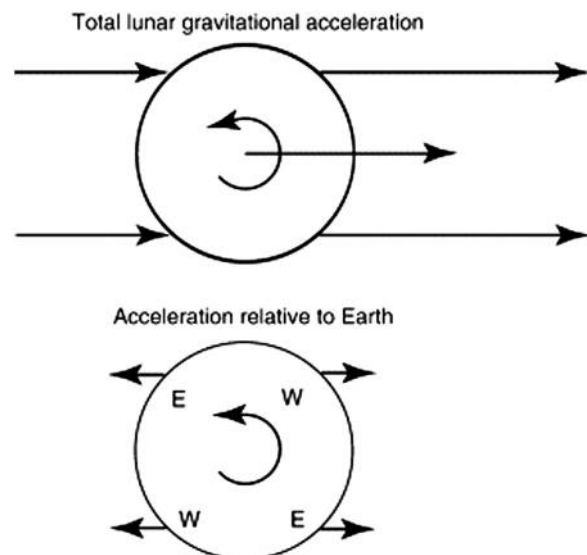


Figure 8 Schematics of lunar tidal forcing in the reference frames of the Moon (top) and the Earth (bottom). The Earth is viewed from above the North Pole and the Moon (not shown) is to the right.

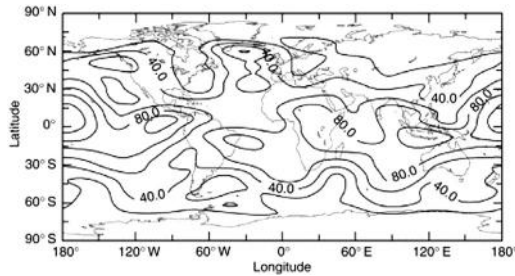


Figure 9 Contours of M2 lunar semidiurnal surface pressure amplitude (dPa) against longitude and latitude for the month of December. Dr F. Vial collaborated with Professor J.M. Forbes to produce the model results.

lunar gravity through deformation of the Earth's surface due to ocean and Earth tides. The vertical velocity associated with this deformation significantly affects the atmospheric tide, and the modulation of terrestrial gravity by the deformation of the Earth's mass distribution also has an effect. These indirect forcing effects can be determined accurately from measurements of ocean and Earth tides, so that the total lunar tidal forcing is well known. This is beneficial for testing theoretical models of tidal propagation and dissipation in the atmosphere. The indirect forcing effects, because they depend on land–sea differences, are a function not only of apparent lunar position but also of geographical location, and generate nonmigrating tides in addition to the primary migrating lunar tides.

It is possible to develop a model of the M2 lunar semidiurnal atmospheric tide that produces results that agree very satisfactorily with the observed tide in the surface pressure. Such a model must account for all direct and indirect lunar forcing effects, and include realistic atmospheric wind and temperature structures. **Figure 9** illustrates prototypical M2 model results and shows how the lunar tidal amplitude varies with latitude and longitude over the Earth for atmospheric conditions representing the month of December. The largest amplitudes in this month are at low latitudes, but they vary from less than 4 Pa (40 dPa) on the east coast of South America to more than 12 Pa over the mid-Pacific. A secondary maximum appears over the northern Atlantic. On the average, the amplitude of the M2 tide at low latitudes is larger around the solstices than at the equinoxes, by roughly 50%. Some of the largest geophysical effects of atmospheric lunar tides appear in the low-latitude ionosphere, as discussed in the next section.

Tides in the Ionosphere

Tides in the ionosphere are spatiotemporal variations in electric fields, currents, and plasma density in the ionosphere primarily due to dynamo effects generated by solar and lunar tidal motions in the neutral background atmosphere. Tidal winds in the low and middle latitude E-region move the partially ionized plasma through the Earth's magnetic field while the electrons with their high gyro frequency/collision frequency ratio remain fixed to the magnetic field lines. An electromotive force is thus created with ensuing electric currents and polarization electric fields.

During the daytime, when the conductivity is large owing to ionizing solar radiation, the electric currents, commonly labeled S_q for 'solar quiet,' flow approximately counter

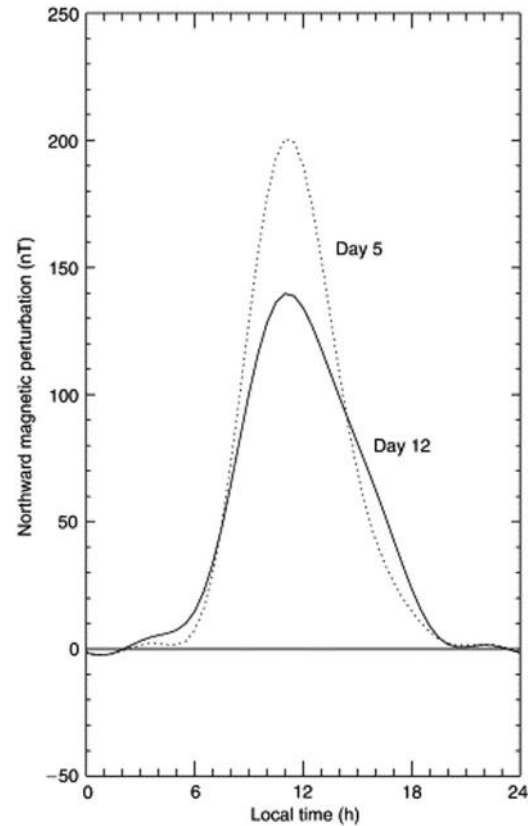


Figure 10 The northward component of magnetic perturbation (in nanoteslas) at Huancayo, Peru, against local time (h) on the 5th and 12th days following new moon reveals evidence of lunar tidal effects on top of the larger solar tidal effects.

clockwise in the Northern Hemisphere and clockwise in the Southern Hemisphere with vortex foci at roughly $\pm 30^\circ$ magnetic latitude. A particularly strong eastward current, called equatorial electrojet, exists along the geomagnetic equator because the effective ionospheric conductivity is unusually high in the lower ionosphere at latitudes where the geomagnetic field is nearly horizontal. The S_q currents produce perturbations in the Earth's magnetic field that are readily measured at the ground. **Figure 10** shows the northward component of the magnetic perturbation at Huancayo, Peru, for two phases of the lunar tide: 5 and 12 days following new moon. These represent average conditions in 1957–58 for the months of November–February, when the lunar tide in the ionosphere is generally largest. The larger solar diurnal and semidiurnal tides produce a northward perturbation that maximizes daily at around 11.00 LT. On day 5, the lunar tide enhances the magnetic perturbation at 11.00 LT, but reduces it in the late afternoon. On day 12 the phase of the lunar tide and its effects are reversed from those on day 5. Clear lunar effects in the low-latitude ionospheric electron density are also found.

The E-region dynamo polarization electric fields are further transmitted along magnetic field lines into the overlying F-region where they drive vertical ($\sim 20 \text{ m s}^{-1}$) and zonal ($\sim 100 \text{ m s}^{-1}$) plasma drifts, which influence many important ionospheric processes. For example, vertical $\mathbf{E} \times \mathbf{B}$ drifts drive the plasma fountain which results in dense bands of plasma

centered near $\pm 15\text{--}20^\circ$ magnetic latitude. This so-called equatorial ionization anomaly can be seen in the top center panel in [Figure 11](#), as the two bluish bands north and south of the magnetic equator. The two ionization crests are predominantly the result of migrating solar tidal winds in the E-region while the apparent longitudinal modulation on top, indicated by brighter colors, is the result of nonmigrating tides excited in the troposphere by latent heat release in deep convective clouds.

The realization that nonmigrating tides due to tropospheric weather impact the F-region ionosphere, and as such couple these atmospheric layers that are 400 km apart, is a relatively new discovery that was made around the year 2005 due to new satellite observations and progress in numerical modeling. [Figure 11](#) summarizes these observations that are all displayed for a constant LT, and sketches the cause-and-effect chain of meteorological impacts on the IT system. The most striking pattern is a 4-peaked ‘wave-4’ longitudinal modulation that is apparent in deep convective cloud occurrence (observed by weather satellites, also compare [Figure 6](#)), E-region zonal winds (observed by TIMED, also compare [Movie 1](#)), thermospheric constituents (observed by the SNOE in equatorial nitric oxide density), electron density (observed by the COSMIC/Formosat-3), ion density (observed by IMAGE spacecraft in the far-

ultraviolet), and neutral mass density (observed by CHAMP). The ‘wave-4’ corresponds to the DE3 nonmigrating tide with some contribution from the SE2, because this is how these components are observed in a LT frame ([eqn \[5\]](#)). [Movie 2](#) and [Figure 6](#) show that the DE3 can achieve zonal wind amplitudes on the order of tens of m s^{-1} in the low latitude E-region. It thus efficiently modulates the E-region dynamo electric fields, resulting in the pronounced ‘wave-4’ F-region plasma density variations. All these observations and corresponding model simulations imply that tropospheric weather is an important contributor to the ‘space weather’ of the geomagnetically quiescent ionosphere, even for solar maximum conditions, and capable to change, for example, electron density by a factor of three within a few thousand kilometers.

Longitudinal variations in the ionospheric plasma due to nonmigrating tides do not always occur as a ‘wave-4’. The ‘wave-4’ dominates from March to October but changes to a 3-peaked ‘wave-3’ from November to February when observed at a fixed LT ([Figure 12](#)). This is mainly due to the seasonal variation of the diurnal nonmigrating tides in the E-region as depicted in [Figure 13](#). The DE3 dominates from March to October but it is exceeded by another component, the DE2 (observed as a ‘wave-3’) during Northern Hemisphere winter.

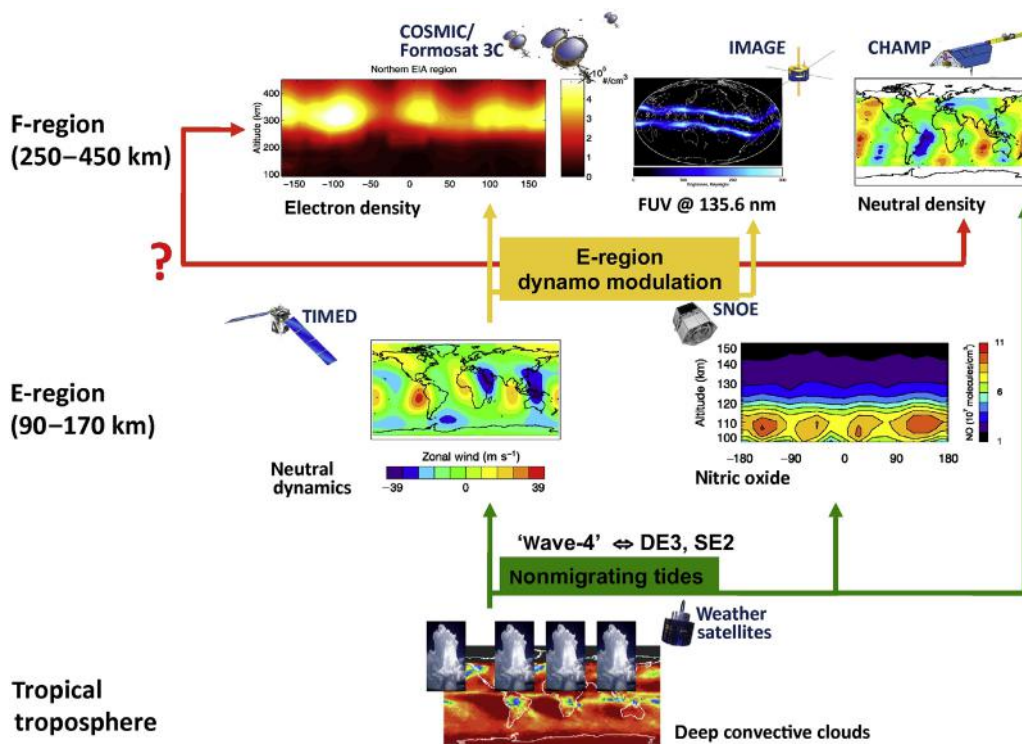


Figure 11 Meteorological impacts on the ionosphere–thermosphere due to nonmigrating tides as observed by different satellites. The ‘wave-4’ like latent heat release pattern in the troposphere is found throughout the ionosphere–thermosphere system in various neutral and ion parameters. Dr C.H. Lin provided the Constellation Observing System for Meteorology, Ionosphere & Climate (COSMIC) figure, which is adapted with permission from Lin, C.H., et al., 2007. Plausible effect of atmospheric tides on the equatorial ionosphere observed by the FORMOSAT-3/COSMIC: three-dimensional electron density structures. *Geophysical Research Letters* 34, L11112. doi:10.1029/2007GL029265. Dr T.J. Immel provided the IMAGE figure, which is adapted with permission from Immel, T.J., et al., 2006. Control of equatorial ionospheric morphology by atmospheric tides. *Geophysical Research Letters* 33, L15108. doi:10.1029/2006GL026161. Dr X. Zhang and Dr S.L. Bruinsma provided the CHAMP neutral density data. The Student Nitric Oxide Explorer (SNOE) nitric oxide figure is adapted with permission from Oberheide, J., Forbes, J.M., 2008. Thermospheric nitric oxide variability induced by nonmigrating tides. *Geophysical Research Letters* 35, L16814. doi:10.1029/2008GL034825. Deep tropical cloud data are from the International Satellite Cloud Climatology Project (ISCCP).

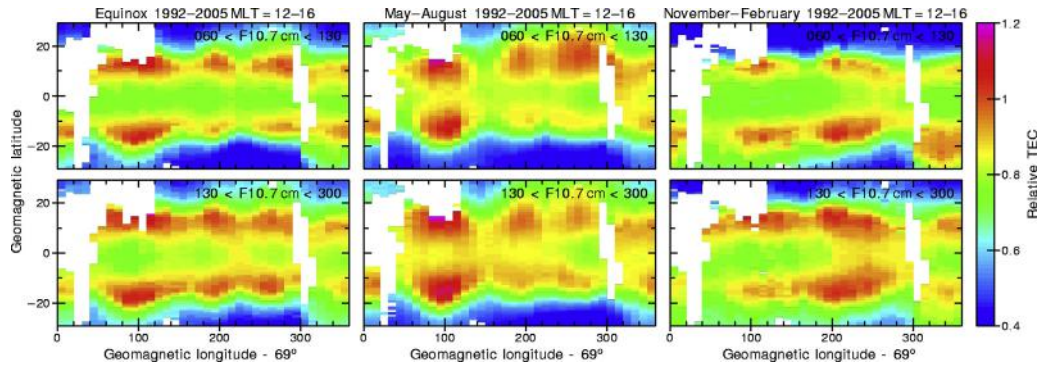


Figure 12 Seasonal variation of normalized total electron content from TOPEX/Poseidon observations for low (upper row) and high (lower row) solar flux conditions. Professor L. Scherliess provided these figures, which are adapted with permission from Scherliess, L., Thompson, D.C., Schunk, R.W., 2008. Longitudinal variability of low-latitude total electron content: tidal influences. *Journal of Geophysical Research* 113, A01311. doi:10.1029/2007JA012480.

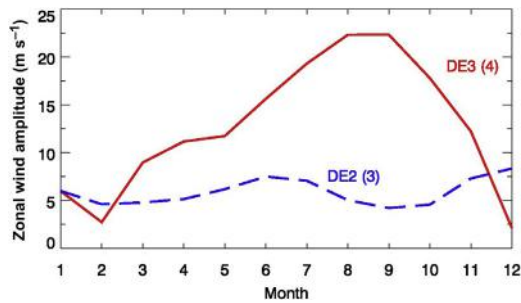


Figure 13 Seasonal variation of the DE2 and DE3 zonal winds at the equator and 105 km as observed by the TIDI/TIMED instrument. The numbers in parentheses indicate the zonal wavenumbers when observed in a LT frame.

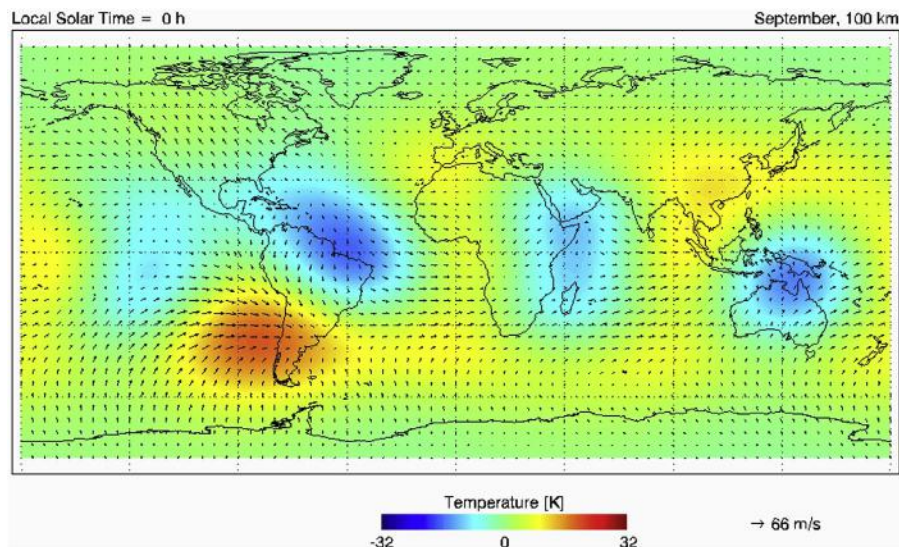
The leading role of E-region dynamo modulation in coupling tidal dynamics into the ionosphere is undisputed but there is growing evidence that other processes may play an important role. While *in situ* F-region ion drag has no measureable effect on the neutral ‘wave-4’ in the upper

thermosphere, models on the other hand suggest that tidal variations in thermospheric $[O]/[N_2]$ and meridional winds at F-region altitudes may add to the observed plasma variations, including effects from semidiurnal nonmigrating tides such as the SE2. Delineating and understanding these processes, including nonlinear wave-wave interactions and secondary wave generation, is one challenge for the time to come as tidal coupling from below constitutes a major energy term for the ionosphere–thermosphere system. It is particularly important since it is a prerequisite for improved space weather predictions including Global Positioning System outages due to ionospheric irregularities or biteouts.

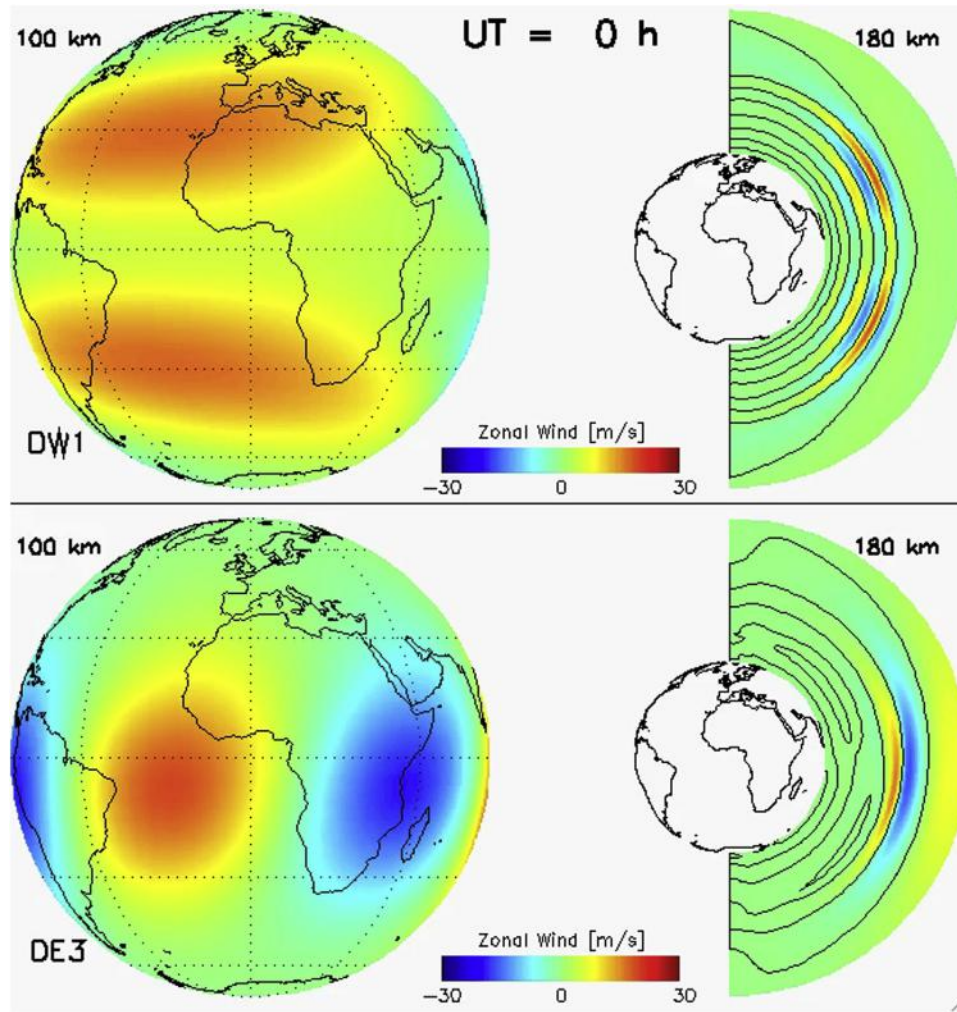
Appendix A: Supplementary Data

Supplementary video related to this article can be found at <http://dx.doi.org/10.1016/B978-0-12-382225-3.00409-6>.

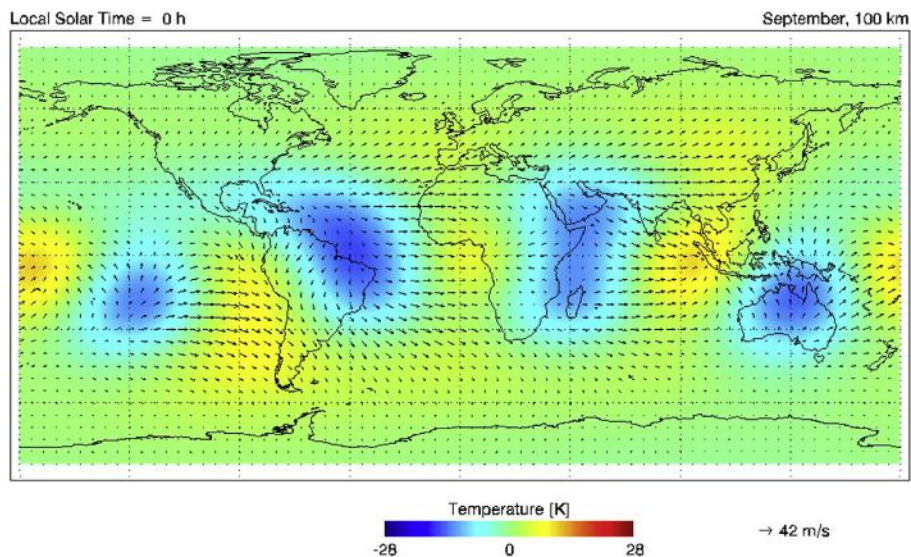
The following are the supplementary data related to this article:



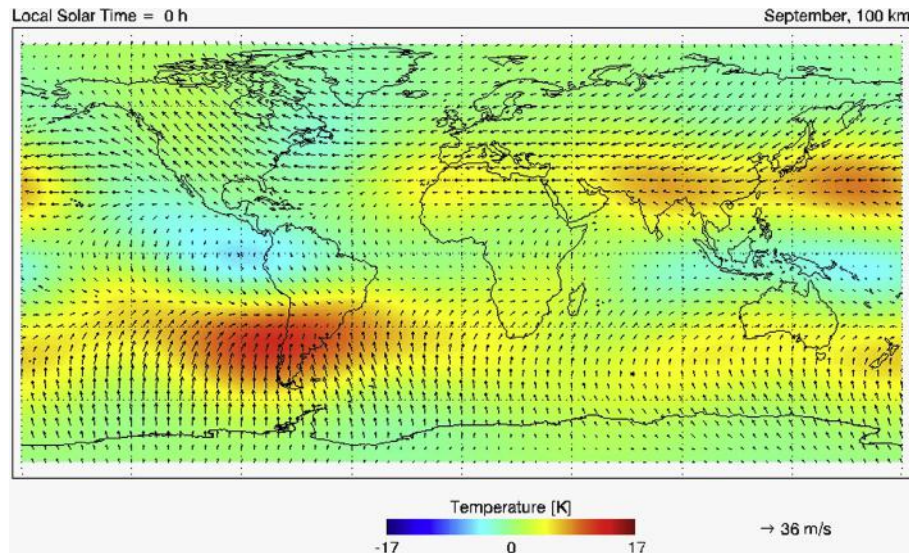
Movie 1 Temperature (color) and wind variations (arrows) at 100 km during September due to diurnal and semidiurnal solar atmospheric tides. The movie is based on TIMED satellite observations and runs in local solar time (LT).



Movie 2 Zonal wind structure of the migrating diurnal tide (DW1, top) and the eastward propagating nonmigrating tide of zonal wavenumber 3 (DE3, bottom). The movie is based on TIMED satellite observations and empirical modeling and runs in universal time. The right column shows the vertical structure of the tides at 90° E longitude. Downward phase progression indicates tidal forcing from below. Thermospheric tides due to far and extreme ultraviolet absorption are not included.



Movie 3 Temperature (color) and wind variations (arrows) at 100 km during September due to diurnal solar atmospheric tides. The movie is based on TIMED satellite observations and runs in LT.



Movie 4 As **Movie 3** but for the semidiurnal tide.

See also: Boundary Layer (Atmospheric) and Air Pollution: Diurnal Cycle. **Gravity Waves:** Buoyancy and Buoyancy Waves: Optical Observations; Buoyancy and Buoyancy Waves: Theory; Overview. **Mesosphere:** Ionosphere. **Numerical Models:** Parameterization of Physical Processes: Gravity Wave Fluxes. **Radar:** Meteor Radar.

Acknowledgment

The authors thank Hanli Liu for comments on an initial draft of the article. The National Center for Atmospheric Research is sponsored by the National Science Foundation.

Further Reading

- Akmaev, R.A., Fuller-Rowell, T.J., Wu, F., Forbes, J.M., Zhang, X., Anghel, A.F., Iredell, M.D., Moorthi, S., Juang, H.-M., 2008. Tidal variability in the lower thermosphere: comparison of whole atmosphere model (WAM) simulations with observations from TIMED. *Geophysical Research Letters* 35, L03810. <http://dx.doi.org/10.1029/2007GL032584>.
- Chapman, S., Bartels, J., 1940. *Geomagnetism*. Clarendon Press, Oxford.
- Chapman, S., Lindzen, R.S., 1970. *Atmospheric Tides*. Reidel, Dordrecht.
- Dai, A., Wang, J., 1999. Diurnal and semidiurnal tides in global surface pressure fields. *Journal of Atmospheric Sciences* 6, 3874–3891.
- England, S.L., Immel, T.J., Huba, J.D., Hagan, M.E., Maute, A., DeMajistre, R., 2010. Modeling of multiple effects of atmospheric tides on the ionosphere: an examination of possible coupling mechanisms responsible for the longitudinal structure of the equatorial ionosphere. *Journal of Geophysical Research* 115, A05308. <http://dx.doi.org/10.1029/2009JA014894>.
- Forbes, J.M., 1995. Tidal and planetary waves. *Geophysics Monographs* 87, 67–87.
- Hagan, M.E., Forbes, J.M., Vial, F., 1995. On modeling migrating solar tides. *Geophysical Research Letters* 22, 893–896.
- Hagan, M.E., Maute, A., Roble, R.G., Richmond, A.D., Immel, T.J., England, S.L., 2007. Connections between deep tropical clouds and the Earth's ionosphere. *Geophysical Research Letters* 34, L20109. <http://dx.doi.org/10.1029/2007GL030142>.
- Haurwitz, B., Cowley, A.D., 1973. The diurnal and semidiurnal oscillations, global distribution and annual variation. *Pure and Applied Geophysics* 102, 193–222.
- Immel, T.J., Sagawa, E., England, S.L., Henderson, S.B., Hagan, M.E., Mende, S.B., Frey, H.U., Swenson, C.M., Paxton, L.J., 2006. Control of equatorial ionospheric morphology by atmospheric tides. *Geophysical Research Letters* 33, L15108. <http://dx.doi.org/10.1029/2006GL026161>.
- Lin, C.H., Wang, W., Hagan, M.E., Hsiao, C.C., Immel, T.J., Hsu, M.L., Liu, J.Y., Paxton, L.J., Fang, T.W., Liu, C.H., 2007. Plausible effect of atmospheric tides on the equatorial ionosphere observed by the FORMOSAT-3/COSMIC: three-dimensional electron density structures. *Geophysical Research Letters* 34, L11112. <http://dx.doi.org/10.1029/2007GL029265>.
- Liu, H., Yamamoto, M., Lüher, H., 2009. Wave-4 pattern of the equatorial mass density anomaly: a thermospheric signature of tropical deep convection. *Geophysical Research Letters* 36, L18104. <http://dx.doi.org/10.1029/2009GL039865>.
- Matsushita, S., 1967. Solar quiet and lunar daily variation fields. In: Matsushita, S., Campbell, W.H. (Eds.), *Physics of Geomagnetic Phenomena*. Academic Press, New York, pp. 301–427.
- Matsushita, S., 1967. Lunar tides in the ionosphere. In: *Handbuch der Physik*. Springer-Verlag, Berlin, pp. 547–602.
- Oberheide, J., Forbes, J.M., 2008. Thermospheric nitric oxide variability induced by nonmigrating tides. *Geophysical Research Letters* 35, L16814. <http://dx.doi.org/10.1029/2008GL034825>.
- Oberheide, J., Forbes, J.M., Zhang, X., Bruinsma, S.L., 2011. Wave-driven variability in the ionosphere–thermosphere–mesosphere system from TIMED observations: what contributes to the “wave-4”? *Journal of Geophysical Research* 116, A01306. <http://dx.doi.org/10.1029/2010JA015911>.
- Picone, J.M., Hedin, A.E., Drob, D.P., Aikin, A.C., 2002. NRLMSISE-00 empirical model of the atmosphere: statistical comparison and scientific issues. *Journal of Geophysical Research* 107 (A12), 1468. <http://dx.doi.org/10.1029/2002JA009430>.
- Richmond, A.D., 1995. The ionospheric wind dynamo: effect of its coupling with different atmospheric regions. In: Johnson, R.M., Killeen, T.L. (Eds.), *The Upper Mesosphere and Lower Thermosphere: A Review of Experiment and Theory*, *Geophysical Monograph* 87. American Geophysical Union.
- Scherliess, L., Thompson, D.C., Schunk, R.W., 2008. Longitudinal variability of low-latitude total electron content: tidal influences. *Journal of Geophysical Research* 113, A01311. <http://dx.doi.org/10.1029/2007JA012480>.
- Vial, F., Forbes, J.M., 1994. Monthly simulations of the lunar semi-diurnal tide. *Journal of Atmospheric and Solar-Terrestrial Physics* 56, 1591–1607.
- Vincent, R.A., Kovalam, S., Fritts, D.C., Isler, J.R., 1998. Long-term MF radar observations of solar tides in the low-latitude mesosphere: interannual variability and comparisons with the GSWM. *Journal of Geophysical Research* 103, 8667–8683.
- Volland, H., 1988. *Atmospheric Tidal and Planetary Waves*. Kluwer Academic, Dordrecht.
- Zhang, X., Forbes, J.M., Hagan, M.E., 2010. Longitudinal variation of tides in the MLT region: 2. Relative effects of solar radiative and latent heating. *Journal of Geophysical Research* 115, A06317. <http://dx.doi.org/10.1029/2009JA014898>.

Balanced Flow

ME McIntyre, University of Cambridge, Cambridge, UK

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

A balanced flow is one in which the three-dimensional velocity field is functionally related to the mass field, presumed hydrostatically related to the pressure field. Such a functional relation between the velocity and mass fields is called a balance relation, or filtering condition. The simplest but least accurate such relation is geostrophic balance. There are more accurate balance relations, of which the most accurate are fully nonlocal. That is, the velocity at a point depends on the mass field throughout the domain. There are ultimate limitations to accuracy, governed by the fuzziness of the slow quasimanifold.

Introduction

The concept of *balanced flow* is the counterpart, in atmosphere–ocean dynamics, to the well-known concept of nearly incompressible flow in classical aerodynamics. In aerodynamics, a key aspect of such flow – long recognized as central to understanding the behavior of subsonic aircraft – is that all the significant dynamical information is contained in the vorticity field. To that extent the flow has, in effect, fewer degrees of freedom than a fully general flow. We may think of it as being elastostatically balanced, in the sense that freely propagating sound waves can be neglected in the dynamics.

In atmosphere–ocean dynamics there is a corresponding statement with vorticity replaced by potential vorticity (PV), understood in a suitably generalized sense; see generalized PV field in the article **Dynamical Meteorology: Potential Vorticity**. For many cases of rotating, stably stratified fluid flow, with parameter values typical of the atmosphere and oceans, all the significant dynamical information is contained in the generalized PV field. One may invert this field at each instant to obtain the mass and velocity fields. The article on Potential Vorticity gives a more precise statement. All such flows may be characterized as balanced.

Again this means that the flow has, in effect, fewer degrees of freedom than a fully general flow. More precisely, balance and invertibility mean that not only sound waves but also freely propagating inertia–gravity waves can be neglected in, or filtered from, the dynamics. Thus balanced flows can be much simpler to understand than fully general flows, thanks to the relatively simple way in which the advective nonlinearity acts on the PV.

Cases of fluid flow describable as balanced come under headings such as Rossby waves, Rossby-wave breaking, vortex dynamics, vortical modes, vortical flow, vortex coherence, vortex resilience, eddy-transport barriers, blocking, cyclogenesis, baroclinic instability and barotropic instability (meaning the wavy shear instabilities), all of which are related to the fundamental Rossby-wave restoring mechanism or quasi-elasticity that exists whenever there are isentropic gradients of PV in the interior of the flow domain, or gradients of potential temperature on an upper or lower boundary. The concept of balanced flow is fundamental, also, to theories of wave–mean interaction and wave–vortex interaction, needed in order to understand, for instance, the gyroscopic pumping that drives global-scale stratospheric circulations and chemical transports

(§6 of **Dynamical Meteorology: Potential Vorticity**). In these theories the mean or vortical flow is usually considered to be balanced, regardless of the wave types involved. Indeed, the concept of balance enters, implicitly or explicitly, into almost any discussion of meteorologically interesting fluid phenomena; and balance versus imbalance is part of the conceptual foundation that underpins data analysis, data assimilation, and weather prediction.

The Elastic Pendulum

Balance has counterparts not only in aerodynamics but also in simple mechanical systems such as the elastic pendulum. This is a massive bob suspended from a pivot by a stiff elastic spring of negligible mass. Such a pendulum has slow, swinging modes of oscillation in which the relatively fast, compressional modes of the bob and spring are hardly excited: they can be neglected in the dynamics if the spring is stiff enough. The slow, swinging modes correspond to balanced flow, and the fast, compressional modes to sound and inertia–gravity waves. One may describe the swinging modes to a crude first approximation by making the spring strictly incompressible, i.e., by making its length strictly constant. There is a hierarchy of more accurate approximations that allow the spring to change its length in a quasi-static or elastostatic way, the spring being longest when the bob moves fastest and shortest when the bob is stationary.

In such a quasi-static description the length of the spring is functionally related to the speed of the bob. The functional relation holds at each instant t , i.e., it holds diagnostically. No derivatives or integrals with respect to t are involved, and values of t do not explicitly enter into the definition of the functional relation. The property of being diagnostic, in this sense, provides us with a useful mathematical and conceptual simplification.

Such approximations and their ultimate limitations can be studied mathematically via techniques ranging all the way from two-timing formalisms (method of multiple scales) and bounded-derivative theory to KAM (Kolmogorov–Arnol’d–Moser) theory and other dynamical-systems techniques; there is an enormous literature.

The error incurred in using the most accurate quasi-static descriptions becomes exponentially small as the fast–slow timescale separation increases. It may even be zero, or in some

circumstances small but inherently nonzero (corresponding to KAM tori breaking into thin chaotic layers, also called fractal layers or stochastic layers).

Balance Relations

In atmosphere–ocean dynamics the defining property of balance is that an analogous functional relation holds – diagnostic in precisely the same sense. The functional relation between bob speed and spring length is replaced by a functional relation between the fluid’s velocity and mass fields. More precisely, a flow is said to be balanced if the three-dimensional velocity field $\mathbf{u}(\mathbf{x}, t)$ satisfies a functional relation of the form $\mathbf{u}(\mathbf{x}, t) = \mathbf{u}^B$ where $\mathbf{u}^B$ depends only on the mass field or mass configuration, i.e., on the spatial distribution of mass throughout the fluid system, presumed to be hydrostatically related to the pressure field. (Knowledge of the mass field then implies knowledge of the pressure, temperature and potential temperature fields, given zero pressure at the top of the atmosphere.)

Such a functional relation $\mathbf{u}(\mathbf{x}, t) = \mathbf{u}^B$ between the velocity and mass fields is called a filtering or balance condition, or balance relation. It supplies just enough information to make the PV field invertible. The property of being diagnostic means that if one knows the mass field at some instant t , but knows nothing about its time dependence, nor the value of t itself, then the balance relation must nevertheless allow one to deduce the complete three-dimensional velocity field $\mathbf{u}$. It must allow the velocity field to be deduced from the mass field and from the mass field alone. (Such a diagnostic relation should not, however, be mistaken for a causal relation. To think that the mass or pressure field causes the velocity field is like thinking that the spring length causes the pendulum’s motion.)

To the extent that a balance relation holds it excludes, or filters, freely propagating sound waves and inertia–gravity waves from the repertoire of possible fluid motions. The system then has too few degrees of freedom to describe such waves. The reduction in degrees of freedom is sometimes expressed by saying that some degrees of freedom are slaved to others, or that the possible states of the dynamical system have been confined to a so-called slow manifold within phase space, having lower dimensionality than the full phase space in which it is embedded. In this language we say that, in particular, the velocity field is slaved to the mass field. A more careful statement would be that in the actual flow the velocity field evolves as if it were slaved to the mass field, to some useful approximation at least. This is like saying that the swinging motion of the pendulum evolves as if the bob speed were slaved to the spring length, to some useful approximation, even though there is no actual mechanical linkage between the two variables.

A standard example of a balance relation is the so-called *geostrophic relation*, which is simple to write and, for typical extratropical parameter values, qualitatively useful though quantitatively not very accurate:

$$\mathbf{u}(\mathbf{x}, t) = \frac{1}{f} \left(-\frac{\partial \Phi(\mathbf{x}, t)}{\partial y}, \frac{\partial \Phi(\mathbf{x}, t)}{\partial x}, 0 \right) \quad [1]$$

Here f is the Coriolis parameter, $\Phi(\mathbf{x}, t)$ is the geopotential height (approximately geometric altitude times gravitational

acceleration), and position $\mathbf{x}$ is specified using pressure altitude along with horizontal position x, y . Thus the horizontal spatial derivatives $\partial/\partial x$ and $\partial/\partial y$ are taken at constant pressure altitude rather than at constant geometric altitude. This qualifies as a balance relation because of the presumption that the hydrostatic relation also holds, as normally assumed when using pressure as the vertical coordinate. Knowing Φ on each constant-pressure (isobaric) surface is then equivalent to knowing the mass field. So eqn [1] is, as required, a diagnostic functional relation between the velocity field and the mass field. The vertical derivative of eqn [1] is the so-called thermal wind equation.

The horizontal coordinates x, y are orthogonal coordinates, and can be taken either as local curvilinear following the Earth’s geometry, or as local Cartesian in a tangent-plane approximation. If we also take $f = \text{constant}$, giving us the so-called f -plane approximation, then eqn [1] asserts not only that $\mathbf{u}$ is slaved to the mass field but also that it is two-dimensionally incompressible or nondivergent, with streamfunction $\Psi = \Phi/f$, so that

$$\mathbf{u}(\mathbf{x}, t) = \left(-\frac{\partial \Psi}{\partial y}, \frac{\partial \Psi}{\partial x}, 0 \right) \quad [2]$$

The geostrophic relation [1] – or relations, plural, if one prefers to think in components rather than vectors – can be motivated as an approximation to the horizontal momentum equation. The accuracy of that approximation depends on smallness of the Rossby number, or, more precisely, on being able to neglect relative particle (Lagrangian) accelerations against Coriolis accelerations, i.e., against f times either side of eqn [1]. The Rossby number Ro , measuring the advective contribution to the relative particle acceleration against the Coriolis acceleration, is usually of the same order as f^{-1} times a typical magnitude of the relative vertical vorticity $\partial v/\partial x - \partial u/\partial y$, the latter being equal to $\nabla_H^2 \Psi$ if eqn [2] holds. Here u and v are the horizontal velocity components corresponding to x and y , and ∇_H^2 is the horizontal Laplacian. Extratropical Rossby numbers have orders of magnitude typically $\sim 10^{-1}$ for synoptic-scale weather systems.

The geostrophic relation [1] was recognized long ago by weather observers as helpful in making sense of synoptic-scale surface pressure patterns, important for instance to ships threatened by cyclonic storms. The history is sometimes discussed under headings such as Buys Ballot’s law and cyclonic development theory. Buys Ballot’s law is a qualitative counterpart of eqn [1] with surface pressure in place of Φ , ‘wind in your back means low on your left’ – low surface pressure, that is, at sea level in the Northern Hemisphere.

Today’s concept of balance recognizes that, like the rigid-pendulum approximation, eqn [1] is merely the first in a hierarchy of more accurate balance relations. Next in the hierarchy is the relation first studied by Bert Bolin and Jule G. Charney in the 1950s, in connection with efforts to develop practical numerical weather prediction. The Bolin–Charney balance relation retains eqn [2] even if f varies with latitude, but redefines Ψ to satisfy

$$\nabla_H \cdot (f \nabla_H \Psi) = \nabla_H^2 \Phi + \nabla_H \cdot (\mathbf{u} \cdot \nabla_H \mathbf{u}) \quad [3]$$

where, as before, ∇_H is horizontally two-dimensional. Equation [3] is an approximation to the divergence equation, the

latter being the result of taking the horizontal divergence of the horizontal momentum equation. The relative particle acceleration is now retained. Its advective part gives rise to the last term of eqn [3] while the remaining, $\partial/\partial t$ part is annihilated when the horizontal divergence is taken, because of eqn [2]. It is only because there are no $\partial/\partial t$ terms that the relation [3], with [2], qualifies as a balance relation. In the special case of an isolated circular vortex on an f -plane, eqn [3] reduces to so-called gradient-wind balance, namely eqn [1] corrected to include the centrifugal force of the relative motion.

Again because of eqn [2], the right-hand side of eqn [3] can be rewritten using a Jacobian in u and v , as $\nabla_H^2 \Phi - 2\partial(u, v)/\partial(x, y)$, or equivalently using a Hessian in Ψ so that eqn [3] becomes

$$\nabla_H \cdot (f \nabla_H \Psi) = \nabla_H^2 \Phi - 2 \left\{ \frac{\partial^2 \Psi}{\partial x^2} \frac{\partial^2 \Psi}{\partial y^2} - \left(\frac{\partial^2 \Psi}{\partial x \partial y} \right)^2 \right\} \quad [4]$$

Regarded as an equation for Ψ when the mass field Φ is given, eqn [4] is of Monge–Ampère type, with an extensive mathematical theory. Iterative methods are needed to solve it numerically because of the nonlinear terms on the right. The problem of finding Ψ becomes ill-posed for certain mass fields Φ , adumbrating, for one thing, that there exist mass fields not even approximately balanceable by any velocity field.

A simple thought experiment to make this last point clear would be to pile up the whole of the Earth’s atmosphere into a narrow column above the North Pole, leaving a vacuum elsewhere. It is obvious that no velocity field $\mathbf{u}$ can be in balance with such a mass field. The free evolution at subsequent times, in any such thought experiment, would start with the column collapsing downward and outward and would involve sound and inertia–gravity waves of enormous amplitude. That is, it would involve gross imbalance as well as, almost certainly, violent wave-breaking and turbulence.

Balance relations are useful in practice only because naturally occurring mass fields, or at least smoothed versions of them are, by contrast, often balanceable to good approximation. In most such cases, eqn [4] with suitable boundary conditions is a well-posed nonlinear elliptic boundary-value problem in the flow domain, the primary exception being flows near the equator, where Rossby numbers are not small and eqn [4] may fail to be elliptic, as can be verified from the theory of Monge–Ampère equations. Again the failure of ellipticity adumbrates a physical reality (though not in a way that is quantitatively precise), namely the fact that balance is liable to break down spontaneously through ‘inertial’ and ‘symmetric’ instabilities near the equator, where f changes sign. There are other varieties of spontaneous imbalance, some only recently clarified. Again these are usually unimportant when Rossby numbers are small.

Balance relations still more accurate than eqn [4] can be defined if one is prepared to deal with more complicated sets of equations. The next relation in the hierarchy – to be referred to here as the generalized Bolin–Charney balance relation – is the first in the hierarchy to yield a nonvanishing vertical component of $\mathbf{u}$. It was implicit in the pioneering work of Charney published in 1962, in a famous paper entitled ‘Integration of the primitive and balance equations’. It starts with eqns [2] and

[4] but then adds to the resulting $\mathbf{u}$ field a horizontally irrotational, divergent correction field governed by another elliptic boundary-value problem in the flow domain, a generalization of the omega equation previously developed by Norman A. Phillips and others. The corrected $\mathbf{u}$ field is an asymptotically consistent improvement on eqn [1], for small Rossby number Ro , in the sense that it is one order more accurate in powers of Ro . The elliptic boundary-value problem is derived by taking $\partial/\partial t$ of eqn [4], then eliminating all the resulting time derivatives using the exact mass conservation and vorticity equations and the inverse Laplacian of the vorticity equation. The vorticity equation expresses $\nabla_H^2 (\partial \Psi / \partial t)$ in terms of diagnostically known, or knowable, quantities such as the corrected $\mathbf{u}$ field; so the inverse Laplacian is needed in order to eliminate $\partial \Psi / \partial t$ from $\partial/\partial t$ of eqn [4].

This process of eliminating all the time derivatives including others such as $\partial \Phi / \partial t$ has to be possible, in principle at least, if the end result is to be a balance relation. By definition, a balance relation may not contain any time derivatives. When the elimination is carried out explicitly, a rather complicated set of integro-differential equations results, containing Green’s function integrals whose details depend on the geometry of the flow domain. It may therefore be notationally and computationally more convenient to work with a set of equations from which $\partial \Psi / \partial t$, $\partial \Phi / \partial t$, etc., have not been eliminated, but have been allowed to remain as unknowns that can, in principle, be eliminated.

Then ‘ $\partial \Psi / \partial t$ ’, in scare-quotes, so to speak, must be regarded not as the rate of change of Ψ but, rather, as an auxiliary variable – better described as a diagnostic estimate of, as distinct from the actual, rate of change. Such a diagnostic estimate must be expected to differ, in general, from the actual rate of change of Ψ , for the reasons explained under *ultimate limitations* below. To avoid confusion over this point a special notation is sometimes used, such as Ψ_1 for a diagnostic estimate of $\partial \Psi / \partial t$ and Ψ_2 for $\partial^2 \Psi / \partial t^2$, and so on.

The general form of the functional dependence defining a balance relation, assuming a balanceable mass field represented by $\Phi(\mathbf{x}, t)$, can be written symbolically as

$$\mathbf{u}(\mathbf{x}, t) = \mathbf{u}^B(\mathbf{x}; \Phi(\cdot, t)) \quad [5]$$

where it is again emphasized that no derivatives or integrals with respect to t may appear. It must be possible, in principle at least, to eliminate them all. Time t enters solely via the second argument $\Phi(\cdot, t)$ of $\mathbf{u}^B$. The notation $\Phi(\cdot, t)$ follows mathematical convention and signifies nonlocal spatial dependence. In other words, the second argument of $\mathbf{u}^B$ is the whole *function*, Φ of $\mathbf{x}$, over the whole flow domain at the given instant t – not merely the *value* of Φ at the single value of $\mathbf{x}$ to which the left-hand side of eqn [5] refers. Such nonlocal functions are sometimes called functionals.

Even the geostrophic relation [1] is enough to illustrate the point, though it involves nothing more than the behavior of Φ in the immediate neighborhood of $\mathbf{x}$ – more precisely, it involves enough about that behavior to permit the evaluation of the two horizontal derivatives. The Bolin–Charney balance relations, generalized or not, are fully nonlocal, as is plain from the occurrence of elliptic partial differential operators like ∇_H^2 and, implicitly or explicitly, the associated Green’s function

integrals. To find $\mathbf{u}$ from Φ one has to solve elliptic partial differential equations in the flow domain, as already emphasized, implying for instance that the value of $\mathbf{u}$ at some position $\mathbf{x}$ will depend on values $\Phi(\mathbf{x}', t)$ at other positions $\mathbf{x}'$ well outside the neighborhood of $\mathbf{x}$.

The generalized Bolin–Charney balance relation is often accurate enough for practical purposes, such as observational data analysis and assimilation, and the initialization of the full dynamics for numerical weather prediction. Of fundamental interest, however, from a theoretical viewpoint, is the fact that the pattern of elimination of time derivatives can be extended systematically to higher derivatives, often resulting in balance relations that are still more accurate.

The ideas involved seem to have been first explored by Karl Hinkelman in the 1960s, in connection with the initialization problem in numerical weather prediction. They were later approached from another direction via the normal modes of the Laplace tidal equations, by Bennert Machenhauer, Ferdinand Baer, Joseph Tribbia and others. The history then went full circle, successively under headings such as nonlinear normal-mode initialization, bounded-derivative method, implicit normal-mode initialization, non-normal-mode initialization, and non-normal-mode filtering, of which the last four represent a rediscovery or further development of Hinkelman's ideas. The ideas were first applied to accurate PV inversion by Warwick A. Norton in the late 1980s. An ingenious numerical approach bypassing the explicit consideration of diagnostic estimates like $\Psi_1, \Psi_2, \dots$ or their normal-mode counterparts was developed by Álvaro Viúdez and David G. Dritschel in 2004.

The Ultimate Limitations

The most accurate balance relations can, in some circumstances, be far more accurate than values of parameters like the Rossby number Ro might ever suggest; and this accuracy extends over a far wider range of parameter values than could reasonably have been expected a priori – with Ro values of order unity, and even greater, in some cases. This astonishing fact – first discovered by Norton through numerical experiments on hemispherical shallow-water flows, for which $Ro = \infty$ at the equator – cannot be deduced by inspection or scaling analysis of the momentum equation or other forms of the equations of motion. It involves great mathematical subtlety, and full understanding has yet to be achieved. Norton's most accurate results used the nonlinear normal-mode technique.

Some insight into the ultimate limitations on accuracy has come from classical aerodynamics. There have been many theoretical, experimental and numerical studies of the weak aerodynamic sound generation or *Lighthill radiation* named after M. James Lighthill's celebrated pioneering work of 1952. This is a simple form of spontaneous imbalance. It is now known that in continuously stratified flows there are further forms of spontaneous imbalance, neither instability-related nor Lighthill-like. Recent work at the cutting edge of this problem can be found in papers by D. Muraki, R. Plougonven, C. Snyder, A. Viúdez, and F. Zhang, appearing in the literature from about 2007.

The spontaneous-imbalance literature gives us a clear answer, in the negative, to a classic question posed in 1980 by Edward N. Lorenz. Could there be an exact balance relation? Could there be unsteady stratified, rotating flows that evolve in such a way that freely propagating sound and inertia–gravity waves are completely absent? More precisely, is there a slow manifold within the full phase space that is an invariant manifold of the full dynamics? Evolution on such a manifold would be such that spontaneous imbalance vanishes exactly.

Lighthill's arguments are enough to show that such a situation is overwhelmingly improbable. Though falling short of rigorous proof, they amount to a very strong heuristic. They show that, whatever else is going on, unsteady vortical flows are practically certain to emit sound and inertia–gravity waves, albeit sometimes very weakly; and practically all the flows of interest are unsteady. This means that spontaneous imbalance is generically nonzero, even though it may often be very close to zero, implying in turn that the so-called 'slow manifold' within the full phase space must be an infinite-dimensional chaotic layer. Though astonishingly thin in places – over a far wider range of parameter values than could reasonably have been expected a priori, as shown by Norton's work, in some cases at least – it is not a manifold, which by definition is infinitesimally thin. Though sometimes astonishingly accurate, the concept of balance is inherently and fundamentally approximate. The layer is sometimes referred to, therefore, as the slow quasimanifold.

(Arguably, a self-contradictory term like 'fuzzy manifold' is best avoided. By its mathematical definition a *manifold* is a perfectly sharp, smooth hypersurface and not at all fuzzy. Thus 'fuzzy manifold' would add yet another item to the list of self-contradictory terms like 'variable solar constant' and 'asymmetric symmetric baroclinic instability' – which of course we inevitably have to live with but, perhaps, need not add to.)

The fact that spontaneous imbalance can take a variety of forms beyond those described by Lighthill's arguments does not change the conclusion that the slow quasimanifold is generically a chaotic layer. Adding to the repertoire of possible imbalance mechanisms can only reinforce that conclusion.

Balanced Models

As already indicated, the swinging modes of the elastic pendulum can be described in a simplified yet in some cases accurate manner by imposing a functional relation between bob speed and spring length, suitably chosen. This reduces the dimensionality of the dynamical system's phase space. Similarly, vortical flows can be described by simplified balanced models or balance models, so called. These are constructed by imposing a balance relation from the start, thereby forcing a true slow manifold into existence. The phase space of the original equations – usually taken as the hydrostatic 'primitive equations', so called – is collapsed into a smaller phase space, though still infinite-dimensional.

The initialization of a balanced model requires only a single scalar field to be specified, such as the mass field, or the PV field in the generalized sense. This scalar field is sometimes called the master field or master variable of the balanced

model, to which all other dependent variables are slaved diagnostically. The model has only one prognostic equation, involving only one true time derivative, the rate of change of the master variable. This rate of change is to be sharply distinguished from the diagnostic estimates of time derivatives that may be hidden inside the definition of the balance relation [5], such as the diagnostic estimates $\Psi_1, \Psi_2, \dots$ already mentioned.

Among those qualifying as balanced models in this standard sense are the models labeled quasigeostrophic theory, semigeostrophic theory, and the Bolin–Charney model, also called ‘the’ balance equations, in isentropic coordinates or in shallow water. In the Bolin–Charney model the master variable can equally well be taken as the mass or as the PV. Both are advected by the same velocity field, a velocity field that satisfies the generalized Bolin–Charney balance relation. Here, as implicitly above, the PV is the exact (Rossby–Ertel) PV and is to be evaluated with the same velocity field.

This property of having a single velocity field is unusual in balanced models. Unlike the Bolin–Charney model, most balanced models have different velocity fields to do different jobs. Semigeostrophic theory is a well-studied example. It has three separate velocity fields. The first advects PV and mass, the second evaluates energy and momentum, and the third evaluates the PV taken as the exact, Rossby–Ertel PV. This last fact is hidden from view in most expositions of the theory. Traditionally the model’s PV is written in terms of the second velocity field, complicating the appearance of the formula for PV and disguising its origin. The three velocity fields differ from each other by fractional amounts $O(Ro)$ where Ro is again the Rossby number.

Semigeostrophic theory has remarkable mathematical properties but is comparable to quasigeostrophic theory in having $O(Ro)$ errors relative to the primitive equations. (Semigeostrophic theory is, however, superior in some respects, such as describing frontogenesis, albeit inferior in others such as describing mesoscale vortices.) The property of having more than one velocity field – for want of a better term we may call it ‘velocity splitting’ – was thought until recently to be a property of all balanced models with the sole exception of the Bolin–Charney model. In all other cases, refining the accuracy was thought to reduce greatly, but not to eliminate, disparities between the velocity fields of a model.

All this was indeed known to be true not only of Norton’s and similar highly accurate balanced models, but true also of another important subclass of balanced models, namely all the Hamiltonian models that can be constructed by Salmon’s method. Semigeostrophic theory is one of these. In the 1980s Rick Salmon showed how to construct balanced from unbalanced models in a systematic way within the Hamiltonian framework. Within that framework one imposes a balance relation as a constraint on the full dynamics, preserving the symplectic geometry of phase space. The constraint is imposed not only on dynamical trajectories but also on functional variations about those trajectories. The resulting balanced models are thus guaranteed to inherit Hamiltonian structure, as well as being accurate to the same formal order in Ro as the imposed balance relation.

A reason for using the Hamiltonian framework is the control it provides over conservation principles. The

framework, properly applied, guarantees that the balanced model will exactly conserve mass, momentum, and energy as well as PV materially. However, there is a fundamental tension between balance relations and conservation principles. A balanced model tries to mimic vortical flows that in reality produce Lighthill radiation or other forms of spontaneous imbalance. The spontaneous imbalance must give rise to wave-induced fluxes of energy and momentum, none of which can be exactly described by the balanced model.

So if one forces a true slow manifold into existence by imposing a balance relation, while insisting that all conservation relations still hold, something has to give way. What gives way, as it turns out, is the concept of a unique velocity field. All balanced models constructed by Salmon’s method exhibit velocity splitting, usually into two separate velocity fields but sometimes, as with semigeostrophic theory, into three. For more detail see the [Further Reading](#) list.

Even if we abandon energy and momentum conservation, there remains a possible tension between balance and local mass conservation. This is because spontaneous imbalance involves local adjustments in the mass field. Until recently, it was thought that this explained why the most accurate non-Hamiltonian balanced models then known still exhibit velocity splitting in one form or another, albeit by tiny amounts. So it was a further surprise when, thanks to recent work by A.R. Mohebalhojeh, a class of highly accurate balanced models was discovered that are entirely free of velocity splitting, yet as far as we know pay no systematic price in terms of accuracy, within shallow-water dynamics at least. Each such model has a unique velocity field, just as does the far-less-accurate Bolin–Charney model. The unique velocity field advects and evaluates the exact PV, advects mass, and evaluates energy and momentum which latter, however, are not conserved. One consequence, though, is that the models possess exact Casimir invariants (*see Dynamical Meteorology: Potential Vorticity*).

These new balanced models have been collectively designated *hyperbalance equations*. In order to write these equations one has to use functional derivatives, which are nonlocal, as well as ordinary partial derivatives. This may perhaps explain why the hyperbalance equations were not discovered long ago.

It is still an open question whether there will prove to be a tradeoff between accuracy and local mass conservation if we go beyond shallow-water dynamics, toward multi-layer models and continuously stratified reality.

Note on Terminology

The reader is warned that the terms geostrophic balance and its shorthand form, geostrophy, are sometimes used in the literature to mean balance more accurate than geostrophic, i.e., more accurate than [eqn \[1\]](#). A common example is the self-contradictory phrase ‘geostrophic adjustment’, which refers to the mutual adjustment of the mass and velocity fields to approach balance or to stay close to balance – and balance, of course, in real fluid flow, nearly always means not geostrophy, [eqn \[1\]](#), but a more accurate balance within the

generic class [5]. The example of a circular vortex adjusting toward ageostrophic, gradient-wind balance while radiating inertia-gravity waves is enough to illustrate the point. As already mentioned, gradient-wind balance is the particular case of Bolin–Charney balance that applies to a steady circular vortex, equivalent to eqn [1] plus a correction term representing relative centrifugal force. Thus by implication we have another piece of self-contradictory terminology, ‘ageostrophic geostrophic adjustment’, unfortunately well established.

It may also be noted that the term adjustment is itself used in two distinct senses that are sometimes confused with each other. The first is Rossby or initial-condition adjustment, the mutual adjustment of the mass and velocity fields toward balance that occurs primarily because a system is started in an unbalanced state. The second is spontaneous adjustment, the continual mutual adjustment of the mass and velocity fields to stay close to balance in unsteady vortical flow, even after initial conditions are forgotten. This second process is far more subtle and sets the ultimate limitations of the balance concept itself, the degree of fuzziness of the slow quasimanifold.

For all the foregoing reasons, some authors are beginning to avoid the term geostrophic adjustment, instead using the terms Rossby adjustment or initial-condition adjustment in the first case, and spontaneous adjustment or spontaneous imbalance in the second.

The term semigeostrophic theory is used here in its standard sense, referring to the balanced model originally introduced by Brian J. Hoskins in 1975. The reader is warned that in Salmon’s papers the same term, semigeostrophic theory, is used in a different, more generic sense.

See also: **Data Assimilation and Predictability:** Data Assimilation. **Dynamical Meteorology:** Coriolis Force; Hamiltonian Dynamics; Inertial Instability; Kelvin Waves; Kelvin–Helmholtz Instability; Lagrangian Dynamics; Potential Vorticity; Primitive Equations; Quasigeostrophic Theory; Symmetric Stability; Wave Mean-Flow Interaction. **Gravity Waves:** Buoyancy and Buoyancy Waves: Theory. **Mountain Meteorology:** Lee Waves and Mountain Waves.

Further Reading

- Hinkelman, K.H., 1969. Primitive equations. WMO No. 297. In: Bykov, V.V. (Ed.), *Lectures on Numerical Methods of Short-Range Weather Forecasting* (Regional Training Seminar of the World Meteorological Organization). Hydrometeoizdat, Leningrad, pp. 306–375 (This paper is of historical as well as fundamental interest. Hinkelman was the first to show how to construct highly accurate balance relations, for the purpose of filtering from initial conditions the so-called meteorological noise represented by sound and inertia-gravity waves. This noise, i.e. imbalance, was known to have been the main reason for a magnificent scientific failure reported in 1922—the failure of Lewis Fry Richardson’s pioneering attempt at a numerical weather forecast using the primitive equations without initialization.).
- Lighthill, M.J., 1952. On sound generated aerodynamically. I. General theory. *Proceedings of the Royal Society of London A* 211, 564–587 (This famous, lucidly-argued and penetrating classic was the first to reveal the surprising properties of acoustic imbalance for non-rotating, nearly incompressible flow in three dimensions.).
- McIntyre, M.E., Norton, W.A., 2000. Potential-vorticity inversion on a hemisphere. *Journal of the Atmospheric Sciences* 57, 1214–1235. Corrigendum 58, 949. (Section 7 describes the only available investigation of a fundamental issue neglected above—how to make (5) Galilean invariant as well as highly accurate. Section 8 points to the possible tension between accuracy and local mass conservation, now surprisingly discounted by A.R. Mohebalhojeh’s recent work on the hyperbalance equations.).
- McIntyre, M.E., 2009. Spontaneous imbalance and hybrid vortex-gravity structures. *Journal of the Atmospheric Sciences* 66, 1315–1326 (This review brings into one convenient place the disparate recent works of Muraki, Plougonven, Snyder, Viúdez, and Zhang that together have clarified when the Lighthill paradigm does and does not apply.).
- Norbury, J., Roulstone, I. (Eds.), 2002. *Large-Scale Atmosphere–Ocean Dynamics. Geometric Methods and Models*, vol. II. University Press, Cambridge (This volume covers many of the deeper mathematical aspects of balanced models, especially Hamiltonian balanced models, including a thorough discussion of Hamiltonian velocity splitting by Roulstone and the author and a wide-ranging, in-depth review of the elastic-pendulum problem by P. Lynch. In particular, Section 5.3 gives a brief but careful discussion of relation between slow manifolds and unbroken KAM tori in the pendulum problem, following the work of O. Bokhove and T.G. Shepherd. In this regard the pendulum problem is rather different from the fluid-dynamical problem.).

Baroclinic Instability

R Grotjahn, University of California, Davis, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

Baroclinic instability refers to a process by which perturbations draw energy from the mean flow potential energy. Conditions in the middle latitude atmosphere are conducive for zonally varying structures (eddies) to grow by this process. The baroclinic energy conversions are proportional to eddy heat fluxes; these eddies also accomplish some of the necessary poleward transport of heat, especially in middle latitudes. Baroclinically unstable solutions arising in simple linear quasigeostrophic formulations have properties similar to observed frontal cyclones. Improving those simplifying assumptions (such as allowing nonlinearity) improve the similarity between simulated and observed properties.

Introduction

Baroclinic instability draws energy from the portion of the potential energy available to be converted (referred to as 'available potential energy' or APE). APE is dependent upon a horizontal gradient of temperature. The conversions of energy are proportional to perturbation heat fluxes in the horizontal and vertical. From thermal wind balance, a horizontal temperature gradient implies the presence of vertical shear. So, baroclinic instability is also an instability of the vertical shear.

Another view of baroclinic instability emphasizes interacting potential vorticity (PV) anomalies. Baroclinic instability is often studied by linearizing the dynamics equations and using eigenvalue or initial value techniques. These alternative views and analysis procedures generally provide complementary means to understand better baroclinic instability.

The atmosphere requires heat fluxes to maintain the observed pattern of net radiation (positive in the tropics, negative poleward of 38° on an annual average). A zonal mean meridional circulation, such as a tropical Hadley cell, can generate these heat fluxes. The poleward moving air in the Hadley cell accelerates while conserving angular momentum. In contrast, lower tropospheric air is slower-moving. Hence, vertical shear builds toward the poleward edge of each Hadley cell. In middle latitudes, baroclinic instability provides a mechanism to explain how the eddies form and evolve whilst including and accomplishing the necessary heat fluxes. Theoretical models of baroclinic instability can simulate various observed properties of midlatitude eddies including the dominant length scales, propagation speed, vertical structure, and energetics.

An Illustrative Model

An illustrative model provides mathematical relations and archetype solutions for the concepts that follow.

Mathematical Formulation

The model uses quasigeostrophic (QG) approximations and nondimensional scaling appropriate for midlatitude frontal

cyclones. PV has contributions from the interior and from temperature gradients at rigid bottom ($z = 0$) and top ($z = Z_T$) boundaries. PV in the QG system (or QG PV) can be written:

$$q = \underbrace{\nabla^2 \psi + f_0 + \beta y}_{\text{RV}} + \underbrace{\gamma \frac{\partial \psi}{\partial z} + \varepsilon \frac{\partial^2 \psi}{\partial z^2}}_{\text{TV}} + \underbrace{\left\{ \varepsilon \frac{\partial \psi}{\partial z} \Big|_{z=0} - \varepsilon \frac{\partial \psi}{\partial z} \Big|_{z=Z_T} \right\}}_{\text{BPV}} \quad [1]$$

where

$$\varepsilon = \frac{f_0^2 L}{g \kappa D}, \quad \kappa = D \frac{\partial \ln \theta_s}{\partial z}, \quad \text{and} \quad \gamma = \frac{\varepsilon \partial \rho}{\rho \partial z} + \frac{\partial \varepsilon}{\partial z} \quad [2]$$

ψ is the horizontal velocity streamfunction, ρ is the density, g is the acceleration of gravity, κ is the static stability from the horizontal mean potential temperature. The coordinates are x eastward, y northward, and z upward. Nondimensional length scales are L in the horizontal and D in the vertical. f_0 is the constant part while β is the meridional derivative (approximated as a constant) of the Coriolis parameter.

An inherent horizontal length scale is the Rossby radius of deformation ($L_R = NHf_0^{-1}$) where N is the Brunt Väisälä frequency ($N^2 = g\kappa D^{-1}$) and $H = RTg^{-1}$ is the scale height (an inherent vertical length scale). Thus, $\varepsilon = (LH)^2 (L_R D)^{-2}$ relates the assumed scales L and H to L_R and D .

QG PV includes three distinct parts: absolute vorticity (AV) that includes relative vorticity (RV), 'thermal' vorticity (TV), and boundary PV (BPV). Positive PV is associated with an interior trough (in geopotential) and/or a warm surface (i.e., boundary) temperature anomaly.

When the vorticity and potential temperature conservation equations are combined, one obtains a time-dependent equation for QG PV conservation:

$$\left[\frac{\partial}{\partial t} + U \frac{\partial}{\partial x} \right] \left(\nabla^2 \Psi + \frac{1}{\rho} \frac{\partial}{\partial z} \left(\rho \varepsilon \frac{\partial \Psi}{\partial z} \right) \right) + \frac{\partial Q}{\partial y} \frac{\partial \Psi}{\partial x} = 0 \quad [3a]$$

with boundary conditions at the bottom and top

$$\left[\frac{\partial}{\partial t} + U \frac{\partial}{\partial x} \right] \left(\frac{\partial \Psi}{\partial z} \right) - \frac{\partial U}{\partial z} \frac{\partial \Psi}{\partial x} = 0 \quad \text{at } z = 0, Z_T \quad [3b]$$

'Basic state' variables are specified: U (independent of x) is zonal wind, and Q is the interior part of the QG PV; meridional and vertical velocities are zero. One can solve eqn [3] as

an initial value problem by specifying an initial streamfunction or PV.

An eigenvalue problem can be formulated from eqn [3]. A common approach assumes time and space dependence in this form:

$$\psi(x, y, z, t) = \text{Re}\{\phi(y, z) \exp(ik(x - ct))\} \quad [4]$$

for the 'perturbation' streamfunction being sought. This solution has zonal wavenumber k and complex phase speed c . The growth rate is given by $k \text{Im}\{c\}$. If U has no meridional variation, then one can assume a wavelike y dependence too: $\exp(iy)$. When meridional wavenumber l equals zonal wavenumber k , the solution is a 'square wave.' Perturbation velocities are defined as $u = -\partial\psi/\partial y$ and $v = \partial\psi/\partial x$.

Additional simplifying approximations are often made. A particularly simple form, commonly labeled the 'Eady model,' was described by E.T. Eady in 1949. The Eady model assumes wavelike meridional structure, $\partial Q/\partial y = 0$, $U = z$, incompressibility ($\rho = \text{constant}$), and $\varepsilon = 1$. Then eqn [3a] is reduced simply to solving $q' = 0$ in the interior where the prime denotes the 'perturbation' sought. The Eady eigenvalue problem can be solved analytically, yielding a pair of 'normal modes' one growing, one decaying for scaled wavenumber $\alpha < \sim 2.4$. The scaled wavenumber:

$$\alpha = \{(k^2 + l^2)\varepsilon^{-1}\}^{1/2} \quad [5]$$

is proportional to absolute wave number and static stability.

Equations for perturbation kinetic energy, K_e and available potential energy, A_e are:

$$\begin{aligned} \frac{\partial A_e}{\partial t} &\equiv \frac{\partial}{\partial t} \iiint \frac{\rho_S}{2} \left\{ \varepsilon \left(\frac{\partial \psi}{\partial z} \right)^2 \right\} dx dy dz \\ &= \underbrace{\iiint \varepsilon \rho_S \frac{\partial U}{\partial z} \frac{\partial \psi}{\partial x} \frac{\partial \psi}{\partial z} dx dy dz}_{(A_z \rightarrow A_e)} - \underbrace{\iiint \rho_S w \frac{\partial \psi}{\partial z} dx dy dz}_{(K_e \rightarrow A_e)} \end{aligned} \quad [6a]$$

$$\begin{aligned} \frac{\partial K_e}{\partial t} &\equiv \frac{\partial}{\partial t} \iiint \frac{\rho_S}{2} \left\{ \left(\frac{\partial \psi}{\partial y} \right)^2 + \left(\frac{\partial \psi}{\partial x} \right)^2 \right\} dx dy dz \\ &= - \underbrace{\iiint \rho_S U \left\{ \frac{\partial \psi}{\partial x} \frac{\partial \psi}{\partial y} \right\}_y dx dy dz}_{(K_z \rightarrow K_e)} + \underbrace{\iiint \rho_S w \frac{\partial \psi}{\partial z} dx dy dz}_{(A_e \rightarrow K_e)} \end{aligned} \quad [6b]$$

The volume integrals are over a closed domain. In the QG system, $\partial\psi/\partial z$ is proportional to potential temperature θ , making the first term on the right hand side of eqn [6a] proportional to a meridional heat flux, while the second term is proportional to a vertical heat flux. The specified vertical shear, $\partial U/\partial z$ is proportional to the available potential energy, A_z of the basic state and is the energy upon which the baroclinic instability mechanism feeds. The first term on the RHS of eqn [6b] is a barotropic energy conversion. The barotropic conversion is

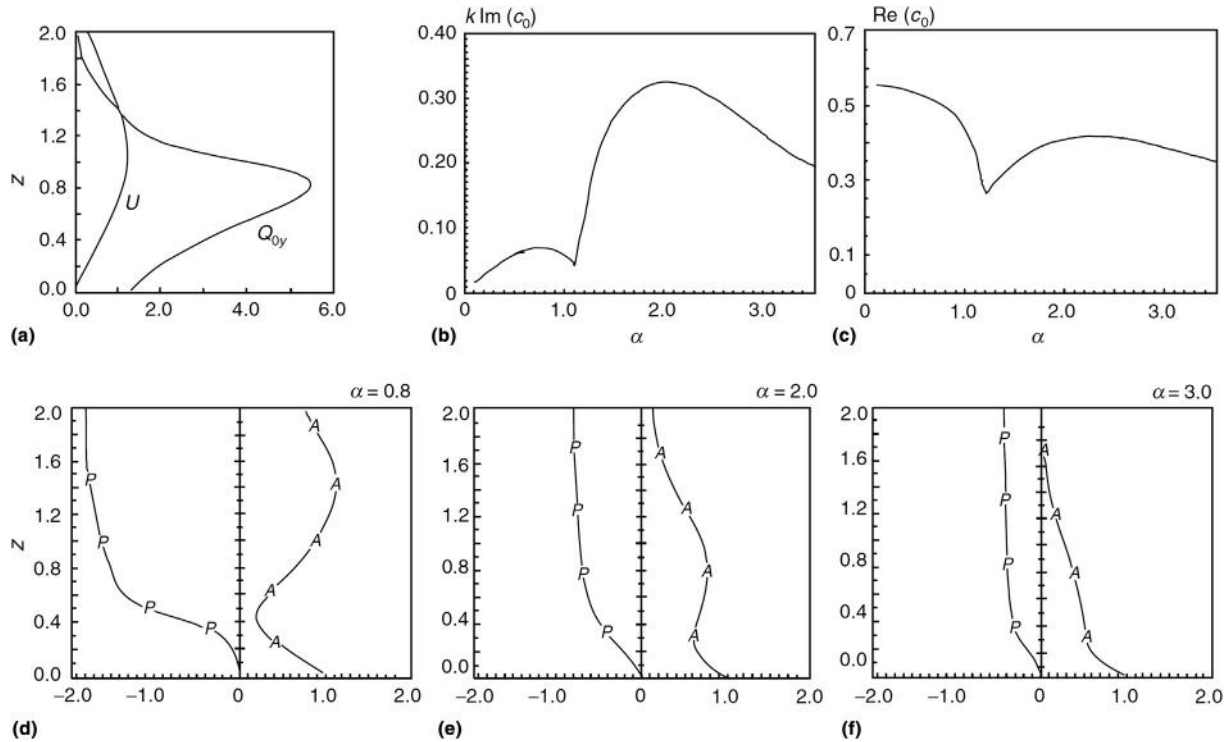


Figure 1 Quasigeostrophic eigenanalysis. (a) Specified zonal wind U , and meridional gradient of interior potential vorticity Q_{0y} versus scaled height. $Z = 1$ is 10 km. (b) Growth rate and (c) phase speed vs absolute wave number α . (d)–(f) Amplitude A , and phase P , for the growing normal mode for $\alpha = 0.8$, $\alpha = 2.0$, and $\alpha = 3.0$, respectively. All three modes tilt westward (upstream) with increasing height. Dimensional wavelengths depend upon scaling assumptions, but reasonable choices imply that $\alpha = 0.8$, $\alpha = 2.0$, and $\alpha = 3.0$ correspond to 11, 4.4, and $\sim 2.9 \times 10^3$ km wavelengths respectively. (Zonal and meridional scales are set equal.) The same scaling implies phase speed of 9 m s^{−1} and doubling time of ~ 1.2 days for $\alpha = 2.0$. Adapted with permission from Grotjahn, R., 1980. *Journal of Atmospheric Science* 37, 2396–2406.

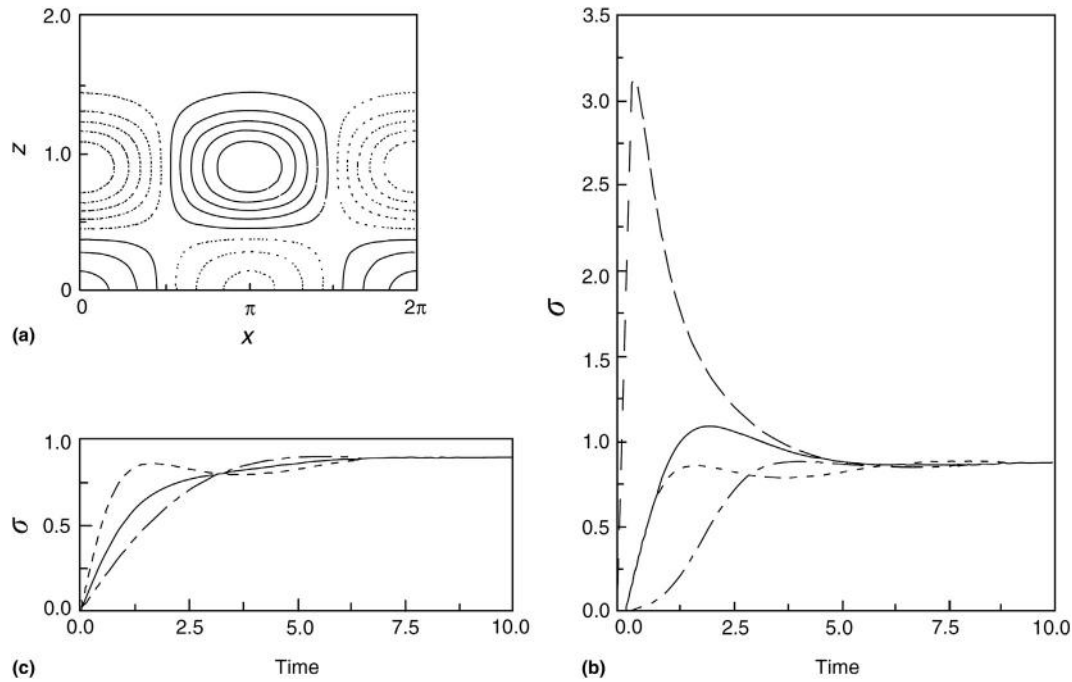


Figure 2 Initial value calculation. (a) Zonal cross section of initial streamfunction, dashed contours used for negative values. (b) Time series of growth rates for domain average potential enstrophy (solid line) and its components: RV^2 (short dashed line), TV^2 (dot-dashed line), and BPV^2 (long dashed line). Growth rates asymptote to the most unstable normal mode rate for this wavenumber ($\alpha = 2.0$). (c) Similar to (b) except for total energy (solid line), kinetic energy (short dashed line), and available potential energy (dot-dashed line).

proportional to the divergence of eddy momentum flux and also draws energy from the mean flow. The second term on the RHS of eqn [6a] and [6b] is the same but with opposite sign indicating a conversion between A_e and K_e .

Example Solutions

This QG eigenmodel of baroclinic instability is applicable to the midlatitudes. In these regions, zonal flow increases with height reaching a maximum near the tropopause. Figure 1(a) is a representative nondimensional profile of U where the tropopause is at nondimensional $z = 1.0$. The growth rate and phase speed spectra along with the (growing normal mode) eigenfunction structures for different k are shown in Figure 1 as well. The growth rate has maximum value at a specific value of α . The vertical structure tends to have relative maxima at the surface and near the tropopause, but it becomes progressively more bottom-trapped for shorter waves. The phase varies such that unstable modes tilt upstream with height, i.e., against the mean flow shear. Other solutions to eqn [3], labeled ‘continuum modes,’ are relevant to ‘nonmodal’ growth.

For shorter waves, the eigenmodes with lower level maximum tend to dominate (when compressibility is included) and the solution decays rapidly away from the boundary. For longer waves, the tropopause level maximum tends to dominate (Figure 1(d)). Eady model normal modes have interior $q' = 0$; from eqn [1]: the Laplacian increases as k^2 requiring a rapid change with height for short waves to make the thermal term comparable to the Laplacian term (this leads to boundary trapping of the solutions). For longer waves, the Laplacian

becomes small and the vertical structure is more evenly spread in the vertical, hence these modes are ‘deeper.’

Typical geopotential patterns observed prior to frontal cyclone development have separate surface and upper troposphere troughs, each equivalent barotropic (vertical trough axis), with the upper level trough more prominent. A crude simulation of that initial state is used to generate solutions shown in Figure 2. Time series of the growth rates of several quantities are tracked over several days. The time series include potential enstrophy ($H = (q')^2$) and total energy ($TE = A_e + K_e$) integrated over the whole domain. The solutions asymptote to the most unstable normal mode growth rate as that eigenmode emerges to dominate the solution. The growth rate has transient peak values that can exceed the asymptotic (normal mode) value.

Classical View

Baroclinic instability draws upon the APE of the environment in which an eddy sits. Since APE is related to a horizontal temperature gradient, and that in turn to the vertical shear, it can be viewed as a type of shear instability. One advantage of doing so is to make comparisons with barotropic instability, which draws energy from the horizontal shear. This view provides a link to eddy fluxes that are observed and necessary for each conversion.

As demonstrated in eqn [6], heat fluxes are necessary to have a baroclinic energy conversion. Horizontal heat fluxes imply that the temperature and mass (here ψ) fields are offset. The offset implies that the trough and ridge axes tilt upstream with elevation.

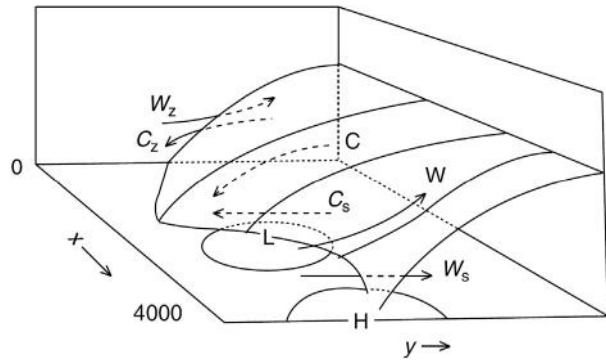


Figure 3 Schematic diagram showing distortion of an isentropic surface by a baroclinically amplifying frontal cyclone. Dotted lines used for objects underneath the three-dimensional isentropic surface. Surface high H, and low L, marked along with two representative contours of surface pressure. Trajectories of representative parcels in the warm air W, and cold air C. Subscript s denotes projection onto the bottom surface while Z denotes projection onto the meridional plane (where $x = 0$). The trajectories do not cross the isentropic surface but distort it. Initially the isentropic surface had negligible variation with x and looked like the current pattern at $x = 0$. The projections W_Z and C_Z seem to cross the initial isentropic surface but in fact are flattening it (which reduces A_z). Rising air is warm while sinking air is cold, which lowers the center of mass converting A_e into K_e .

The QG formulation above is adiabatic, so individual parcels conserve their potential temperature (θ) over time. For unstable modes, the horizontal and vertical eddy heat fluxes must distort the θ field over time as suggested schematically in Figure 3. An isentropic (θ) surface is drawn in three-dimensional perspective; it curves up and over colder areas and dips down over warmer areas. Prior to eddy development, the isentropic surface did not vary in the x direction and had a shape like its intersection with the wall at $x = 0$. The isentropic surface is distorted by flow around the high and low pressure centers and representative cold {C} and warm {W} trajectories are also drawn. When these trajectories are projected onto the $x = 0$ wall they appear to cross the initial zonal mean isentrope and have a slope that is typically half the slope of the mean isentrope. In fact, they are changing the zonal mean of the

isentrope to become more horizontal, thereby reducing the horizontal temperature gradient and thus reducing A_z . In this classical view, A_z is reduced while A_e is increased by increasing the zonal undulations of the isentropic surface. Another aspect is that colder air is sinking while warmer air is rising, a process that lowers the center of mass and thus converts A_e into K_e . To lower the center of mass, the parcel paths must have the vertical component indicated but they must also be less than the slope of the mean isentropes for instability to occur.

The classical view can incorporate latent heat release as follows. The bulk of the precipitation in a developing cyclone forms in the warm air sector of the storm. The release of latent heat further depresses the isentropic surfaces where there is poleward motion implying additional conversion of A_z into A_e and K_e .

Potential Vorticity View

The PV view of instability tracks how two or more PV anomalies interact in a way that causes growth of the PV anomalies. PV is a fundamental conserved quantity for adiabatic motions. The illustrative model is designed around QG PV conservation.

A PV pattern has an associated streamfunction and horizontal wind field. In general, eqn [1] implies that PV emphasizes smaller scale variations than the streamfunction field. Inverting eqn [1] obtains broad patterns of ψ associated with isolated packets of q . For example, PV anomalies in the upper troposphere have corresponding streamfunction extending through the whole troposphere (but somewhat larger amplitude at the level where q has maximum magnitude). The associated winds are displaced from a PV anomaly center by $1/4$ wavelength (~ 1000 km). A similar depiction can be deduced for a PV anomaly associated with a surface temperature gradient.

PV anomalies are created by flow across PV contours. Figure 4 illustrates how two sinusoidal PV anomalies can amplify each other. The PV gradient is reversed between the two levels, increasing with y at upper levels and decreasing with y at the surface. This pattern is consistent with upper tropospheric PV and the surface temperature gradient, respectively. (Recall

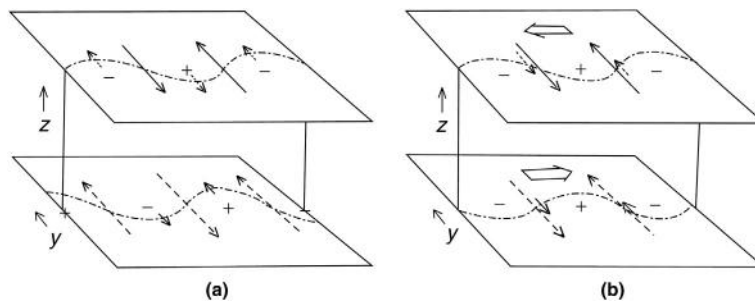


Figure 4 Baroclinic instability from interacting PV anomalies at two levels. A representative PV contour (dot-dashed line) is drawn at each level. Note that the meridional (y direction) PV gradient points opposite directions at the two levels. The offset is (a) $1/4$ wavelength and (b) $1/2$ wavelength. A typical wavelength might be 4000 km. Each anomaly has associated wind component parallel to the PV gradient; dashed arrows are winds from the lower PV anomaly, while solid arrows are from the upper anomaly. The winds created by each anomaly propagate that anomaly. In (a) each PV anomaly has a wind component that amplifies the undulation in the other anomaly (by having a nonzero wind at the center of the other anomaly), thereby causing growth. In (b) each PV anomaly has a wind component that augments the propagation of the other anomaly in the manner indicated by the broad arrows; this causes the anomalies to migrate to a phase offset like diagram (a).

that q is positive for either lower geopotential heights or warmer surface temperature.) The associated winds distort the PV pattern in a way that causes the PV pattern to propagate. However, the meridional wind associated with a PV anomaly is in quadrature with that anomaly so the PV cannot amplify itself. Growth is described simply as advection at the PV extrema that further amplifies the PV pattern. Since the associated winds extend beyond the elevation of the PV anomaly, there can be interaction between the PV anomaly and a second PV anomaly at another level. When the second PV anomaly is offset from the first as in Figure 4(a), the associated winds amplify the first anomaly.

This mechanism also explains how developing cyclones maintain a preferred tilt (i.e., become 'phase locked'). The lower anomaly is shifted horizontally to the right in Figure 4(b) so that upper and lower anomalies are 180° out of phase. The two PV anomalies no longer amplify each other's PV anomalies (shutting off the instability mechanism). Furthermore, the two anomalies reinforce the velocities midway between their positive and negative extremes, thereby enhancing the propagation at each level; but the propagation is directed oppositely at each level, thereby reducing the phase

shift to reestablish the pattern in Figure 4(a). As with the classical view, normal modes are a special case where this phase locking is optimized.

The PV view provides theoretical weight to a classic description of how cyclones develop: an upper level trough (PV anomaly) approaches a low-level baroclinic zone (another PV anomaly), and then growth commences. This paradigm is commonly labeled 'type B' cyclogenesis.

The 'type B' cyclogenesis is illustrated in Figure 5 using a QG nonlinear model. A nearly nondeveloping, nearly coherent upper tropospheric trough is propagating in a flow with vertical shear and is approaching a localized region of warmer surface temperature. The surface warm anomaly is also a positive anomaly of PV in the lower troposphere. The trough has maximum amplitude in the upper troposphere, so there are associated cold anomaly in the troposphere and warm anomaly in the stratosphere. The trough is a region of positive PV. Differential vorticity advection and warm air advection cause rising motion (of cold air) ahead of the trough. Analogously, sinking motion occurs behind the trough. Hence there is positive baroclinic conversion behind and negative ahead of the trough. Integrated over the whole system the net baroclinic

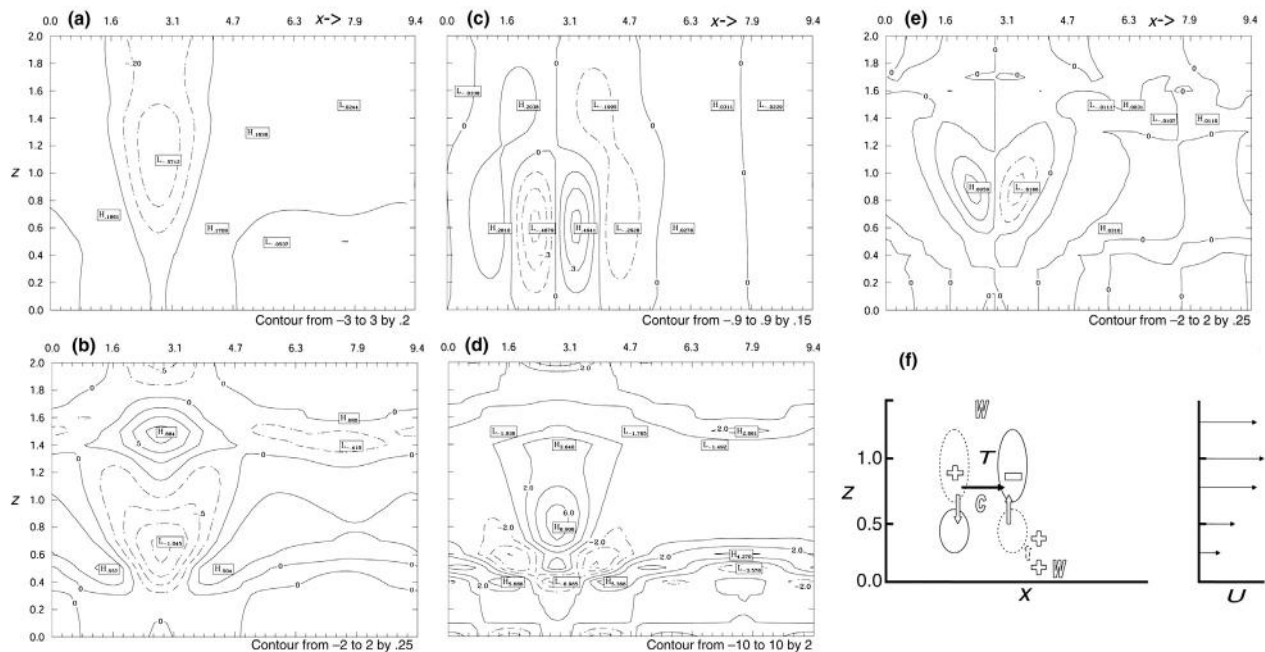


Figure 5 (a)–(e) Zonal cross sections (East–West direction, x versus elevation, z) showing properties across the midpoint of a nearly coherent, nongrowing, upper level trough similar to those observed. z is scaled by 10 km and x by 1000 km. Eddy: (a) streamfunction; (b) temperature; (c) vertical motion, positive upwards; (d) quasigeostrophic potential vorticity; and (e) baroclinic energy conversion shown. The baroclinic energy conversion averaged over the whole upper trough is zero. (f) Schematic zonal cross section of interaction between upper trough encountering a near surface warm anomaly. The upper trough T moves in the direction of the solid arrow. The trough has eddy temperatures indicated by C for colder air, and W for warmer air. Hollow arrows show vertical motions. In the along-flow direction (x) there is upward motion ahead of the trough reaching a maximum near $z \sim 0.6$. Vertical motion is driven both by temperature advection (from $0.6 < z < 1.4$) and differential vorticity advection (from $0.3 < z < 0.7$). The associated divergence is indicated by solid ovals and convergence by dashed ovals. From the quasigeostrophic vorticity equation, these divergence fields oppose vorticity advection by the mean flow at upper levels, and enhance that advection at lower levels. Hence, the trough maintains its vertical tilt in the presence of vertical shear in the zonal mean wind ($U(z)$). The sign of the baroclinic energy conversion, BCE is indicated by open + and - signs. The upper pattern of BCE is similar to panel (e). However, when the upward and poleward motion ahead of the upper trough encounters the warm anomaly, the vertical motion is locally enhanced as is the meridional heat flux. There is net generation of vorticity, net BCE, and the eddy begins to grow. Adapted with permission from Grotjahn, R., 2005. Quarterly Journal of the Royal Meteorological Society 131, 109–124. <http://dx.doi.org/10.1256/qj.03.163>.

conversion is zero. The vertical motion requires divergence and convergence above the peak rising motion. From the QG vorticity equation convergence opposes the positive vorticity advection ahead of the trough at upper levels (reinforces it at lower levels) and in so doing maintains the vertical tilt of the trough even though the advecting wind speed increases with height. When the upper trough begins to encounter the low-level warm anomaly, the warm advection in the poleward flow ahead of the trough is increased. Initially at lower levels, this warmer air rises (instead of the cold air) causing the baroclinic conversion ahead of the upper trough to become less negative and even positive, and the system begins to amplify.

Observations show independent troughs at upper and lower troposphere prior to development with the upper approaching the lower. Neither trough has upstream tilt initially, such tilt emerges only after the two become favorably aligned and growth has commenced.

A necessary condition for instability is that the across-flow mean gradient of PV change sign within the domain. In the illustrative model (Figure 3) and (Figure 4), $\beta > 0$ means that $Q_y > 0$ in the interior, and the surface temperature gradient ($d\theta/dy < 0$) implies that $Q_y < 0$ at the ground. In the Eady model $Q_y = 0$ everywhere in the interior, so the normal mode instability comes from BPV having opposite sign at top and bottom boundaries.

A necessary condition for instability is that a steering level, where $U = \text{Re}\{c\}$, lie within the domain. A supportive kinematic argument is that air parcels remain with the system (rather than blow through it or be left behind) and are more easily mixed laterally. For really long waves, strong retrogressive motion caused by the β -term leads to a different class of unstable eigenmodes for $\alpha < \sim 1.1$ (note cusp in Figure 1(b)) than for larger α .

The eddy meridional flux of potential vorticity is also linked to $\nabla \cdot \mathbf{F}$ in QG theory, where $\mathbf{F}$ is the Eliassen–Palm flux. In addition, due to the strong meridional eddy heat flux present in a baroclinically growing eddy, $\mathbf{F}$ has an upward pointing component. From wave theory, specifically the ‘generalized Eliassen–Palm relation,’ the upward pointing EP flux provides an explanation for upper level amplification of the eddy as it grows nonlinearly.

Normal Modes

Normal modes are physically meaningful eigenfunctions. As in the illustrative model, the equations are linearized about a specified basic state and perturbation solutions are sought. Most commonly, the time and one or more space dependencies are assumed. By assuming a form like eqn [4], unstable solutions grow exponentially. Simple enough models may be solved analytically. More commonly, the eigenvalue problem is solved numerically.

Normal modes are consistent with many observed features:

1. Unstable modes tend to be lined up along the jet axis (if present) in the mean flow.
2. The most unstable wavelength is similar to the observed median size. The normal mode scale can be manipulated by varying the choices made for nondimensional parameters, but is on the order of 4500 km.

3. Solutions tend to develop similar zonal and meridional lengths, the latter responds to the width of the jet providing one natural scale in the model. Other properties (like static stability) also influence the length scales.
4. The vertical structure of the most unstable modes tends to have relative maxima at the surface and upper troposphere.
5. In growing normal modes the temperature lags the mass field in the lower troposphere (typically by 20–50° of phase for the most unstable mode). Two consequences are as follows: First, troughs and ridges in the mass field must be displaced (i.e., tilt) upstream with increasing elevation. There is typically 1/4–1/2 wavelength ($1\text{--}2 \times 10^3$ km) between the trough location at the surface and at tropopause level. Second, the lag allows across-flow heat fluxes down the temperature gradient, as expected from eqn [6a], even for geostrophic winds. In the Eady model the heat flux is uniform with height. Model improvements, most notably compressibility, can emphasize the eddy heat flux in the lower troposphere (where observations find it most prominent).
6. The rate of propagation is $\sim 10\text{--}20$ m s^{−1}: slower than jet stream level winds, but faster than (zonal average) surface winds. A steering level is defined as where the propagation speed of the storm equals the wind component along the storm’s track. The steering level for the most unstable, normal modes is typically between 700 and 500 mb depending on the assumptions made. For shorter waves, the steering level is closer to the surface since these modes move slower. Longer waves respond to competing effects: they have greater upper level amplitude (where U is faster) but greater sensitivity to β (which enhances retrograde motion).
7. The rate of growth is similar to but slower than that of observed cyclones. Observed doubling times are typically 1–2 days at upper levels.
8. Instability is inversely proportional to static stability. For example, the peak growth rate depends on α ($=2.0$ in Figure 1). From eqns [2] and [5], α is proportional to static stability κ . Hence, smaller κ places the most unstable peak at larger k making the growth rate ($k \text{Im}\{c\}$) larger. Kinematically, vertical motion needed in eqn [6] becomes easier for smaller κ .

The fact that normal modes have fixed tilt is not necessarily unrealistic. Observations of the vorticity equation terms support an approximately fixed structure for developing low because the divergence term opposes the horizontal advection at upper levels but reinforces the horizontal advection at low levels. The normal modes (Figure 1) are special structures where the net motion is exactly uniform throughout the depth of the fluid. Tracking observed frontal cyclone troughs over time shows evidence that such storms maintain a roughly fixed tilt during their growth. The vorticity equation also illustrates instability whereupon the divergence term has positive vorticity tendency at a trough where vorticity is a maximum thus amplifying the peak vorticity (and vice-versa for ridges).

In addition to the normal modes, the eigenfunctions include a class of solutions called ‘continuum’ modes. For an adiabatic model, continuum modes have equivalent barotropic structure (no tilt) making them neutral. In the Eady model, continuum modes have zero PV at all levels except at the critical

level, where their amplitude has a 'kink.' Continuum modes play a role in nonmodal growth.

Nonmodal Growth

Nonmodal growth is seen when solving initial value problems like eqn [3]. The formulation can be linear as in eqn [3] or nonlinear. This approach is more general than eigenanalysis since the time dependence is not assumed as it is in eqn [4].

The solution at any time can be decomposed into a combination of eigenfunctions. For an arbitrary initial state, continuum and normal modes are present. These modes move at differing speeds. In a linear formulation the modes operate independently; as modes disperse, positive and negative reinforcement varies. The interference between modes decays algebraically asymptotically. However, for some initial conditions it is possible to have sizable growth over a limited time period.

For the Eady model, analytic solutions can be found which illustrate the process. Using an initial condition with upstream tilt ($\phi \sim \exp(imz)$ in eqn [4] where $m > 0$) yields solutions with normal mode and algebraic parts. The algebraic part has time dependence proportional to $\{(m - kt)^2 + \alpha^2\}^{-1}$ and $\exp\{i(m - kt)z\}$. The amplitude increases while the upstream tilt

becomes more vertical until $t = mk^{-1}$. After that, the wave tilts downstream and decays.

Initial upstream tilt becoming more vertical with time has led to an expectation that RV increases at the expense of TV while interior PV remains conserved. However, exceptions can be found where large nonmodal growth occurs (in H) as upstream tilt *develops* from an initial state with no tilt. The explanation lies in a rough cancellation between RV and TV leaving the BPV evolution to dictate rapid growth in H .

A robust interpretation of nonmodal growth is the progressively more favorable superposition of constituent modes. Continuum modes having mainly upper level amplitude tend to move fast, while modes with mainly lower level amplitude move slowly. Decomposition into eigenmodes of an initial state with upstream tilt finds faster continuum modes located upstream of slower continuum modes. Over time, the modes become more favorably lined up; the tilt becomes more vertical and the total amplitude increases. Figure 6 illustrates the process.

Nonmodal growth can be quite strong in simple models like Eady's. However, most improvements to the model such as adding compressibility, variable Coriolis, and realistic vertical shear of U reduce nonmodal growth. Using more realistic initial states also tends to reduce nonmodal growth (e.g., using a wave packet instead of a wavetrain; using separate untilted upper and lower features instead of connecting them with a tilt).

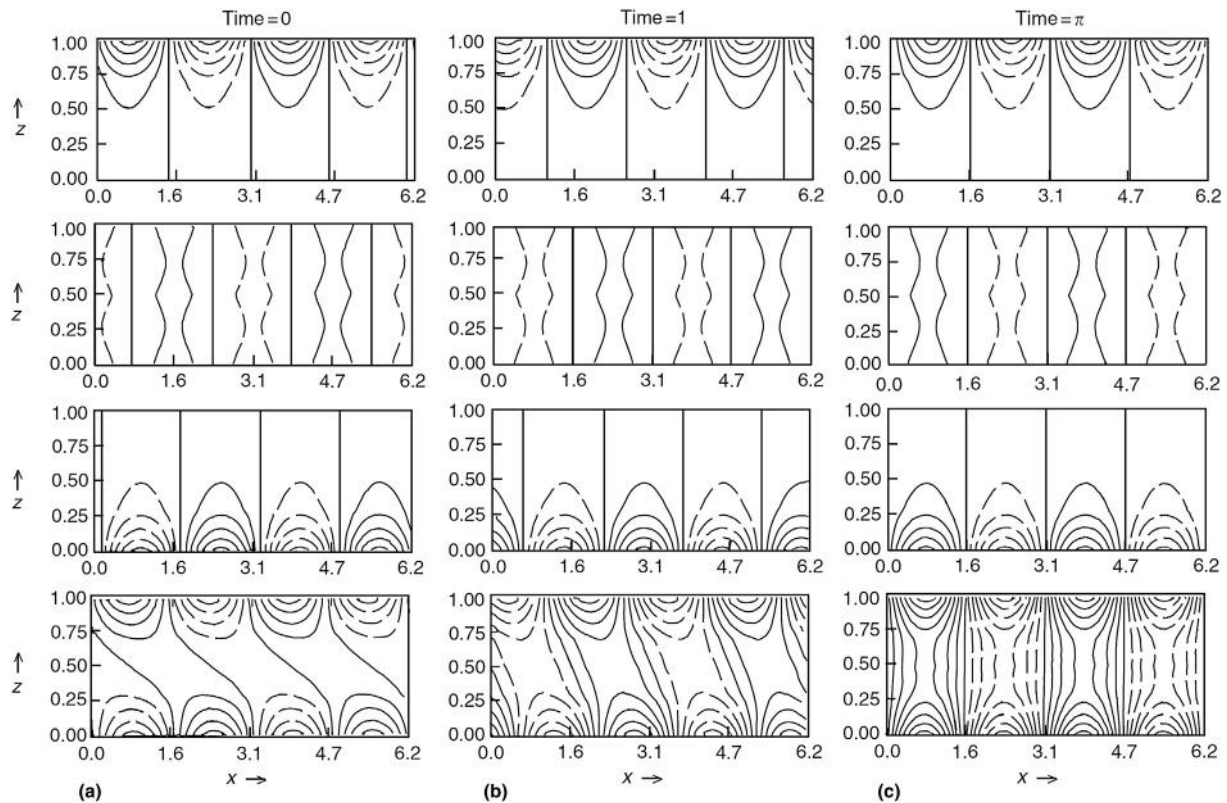


Figure 6 Nonmodal growth as a superposition process. Four initial value linear calculations are shown. The top three rows show three individual neutral continuum modes at three times. The bottom row used the sum of the three modes at the initial time. (a) Initial condition, time = 0, (b) time when energy growth is a maximum in the sum, time = 1, and (c) time when growth rate is zero in the sum, time = π . Adapted with permission from Grotjahn, R., Pedersen, R., Tribbia, J., 1995. *Journal of Atmospheric Sciences* 36, 764–777.

Other Issues

Baroclinic instability has links with barotropic instability. First, each instability draws energy from mean flow shear. Second, barotropic instability has a similar stability criterion (AV gradient changing sign in the domain). Third, there can be interference between the two instabilities. The most unstable baroclinic eigenmode has optimal structure for a flow having the vertical shear alone, but when horizontal shear is added to that flow a different structure is needed otherwise the eddy will be sheared apart. The subsequent structure is unlikely to be as optimal for baroclinic energy conversion. Hence, the baroclinic conversion will usually be reduced, though the barotropic growth mechanism may compensate. Figure 7(c) illustrates such a calculation; in this case adding a purely barotropic flow reduced the growth rate even though the barotropic growth mechanism was activated.

Baroclinically unstable frontal cyclones prefer to develop in certain regions. The preference may arise from local conditions such as lower static stability or locally greater vertical shear. The illustrative model above assumes a wavetrain solution; when more localized development is considered, a variety of issues are raised.

For example, if one uses a single low as the initial condition, the solution typically evolves into a chain of waves as the modal constituents of the initial state disperse. Alternatively, a wave packet initial condition might be used consisting of a 'carrier wave' multiplied by an amplitude envelope. The packet evolution depends upon the mean flow properties and assumptions made in the model. However, for reasonable

choices of parameters, one might find a packet that spreads while propagating downwind. The leading edge of the packet has mainly faster, wider, and deeper modes. The trailing edge has slower, shorter, and shallower waves. It is possible to construct a localized structure, which resists this dispersion by making a judicious combination of eigenmodes having similar phase speed, but different zonal wavenumber. One such example was used when discussing 'type B' cyclogenesis (Figure 5). Figure 8 illustrates another example using neutral continuum modes. When this model is solved as an initial value problem the packet maintains a localized shape for a long time and almost no growth occurs since the normal modes were filtered out and there is very slow phase shifting of the constituent modes. However, when nonlinear advection is allowed, modes interact and soon amplitude is injected into all the eigenmodes including the growing normal modes, which grow rapidly in this example.

Studies of regional development spawned subcategories of baroclinic instability. 'Absolute' instability occurs when the wave packet expands faster than it propagates; the amplitude at a point keeps growing. 'Convective' (in the advection sense) instability occurs when the packet moves faster than it spreads so that growth then decay occurs as the packet moves past a point. 'Global' instability (like the eigensolutions shown here) has growth that is invariant to a Galilean transform. Such is not the case for 'locally' unstable modes. Normal modes for zonally varying basic states look like carrier waves modulated by a spatially fixed amplitude envelope; the envelope locally modifies the growth rate (sometimes called 'temporal' instability); enhancing the global growth locally where the carrier

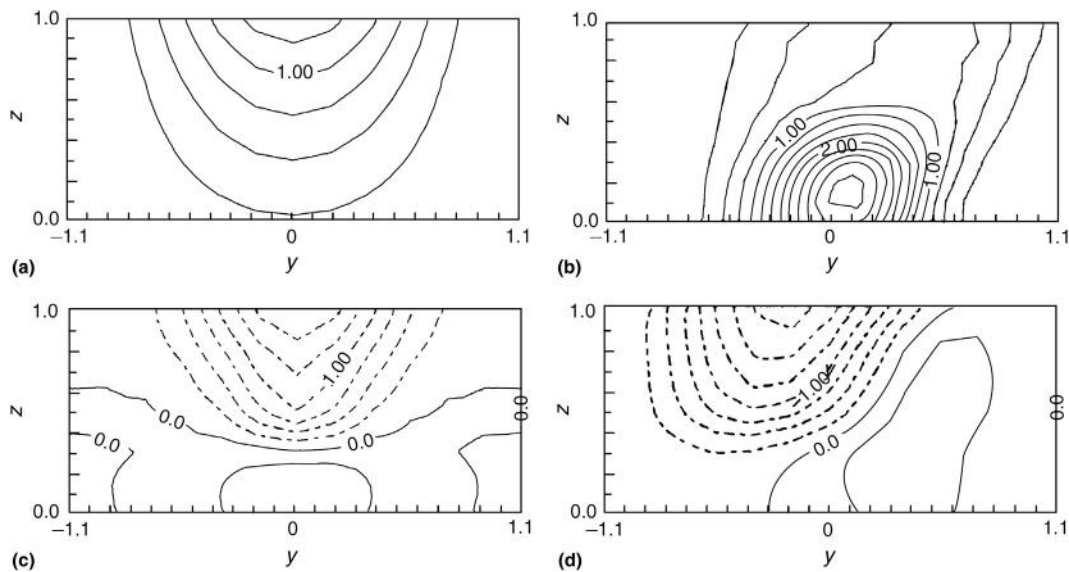


Figure 7 Baroclinic energy conversion ($A_2 \rightarrow A_6$) for four models. (a) Lowest order, square wave solution for an Eady-type model but including compressibility, increasing vertical shear in U , $\beta = 1$. (b) Solution when a surface frontal zone, centered at $y = 0$, is added to the lowest order mean flow U_0 and leading ageostrophic advective effects are included (using geostrophic coordinates). The frontal zone adds wind field: $0.2(2z - z^2) U_1$ where $U_1 = b_1(1 - \tanh^2(\alpha y)) - b_2 - 3b_3 y^2$ to U_0 . The geostrophic coordinate transform causes the asymmetry. (c) Modification to the conversion shown in (a) when barotropically unstable horizontal shear U_1 is added to U_0 . If the total wind is $U = U_0 + \mu U_1$, then the total conversion is (a) + μ (c). The barotropic shear reduces the growth rate. (d) Modification due to all leading order ageostrophic corrections. If those corrections are order μ , then the total conversion is (a) + μ (d). Ageostrophic conversions reduce the conversion and introduce asymmetry. Adapted with permission from Grotjahn, R., 1979. Journal of Atmospheric Sciences 36, 2049–2074.

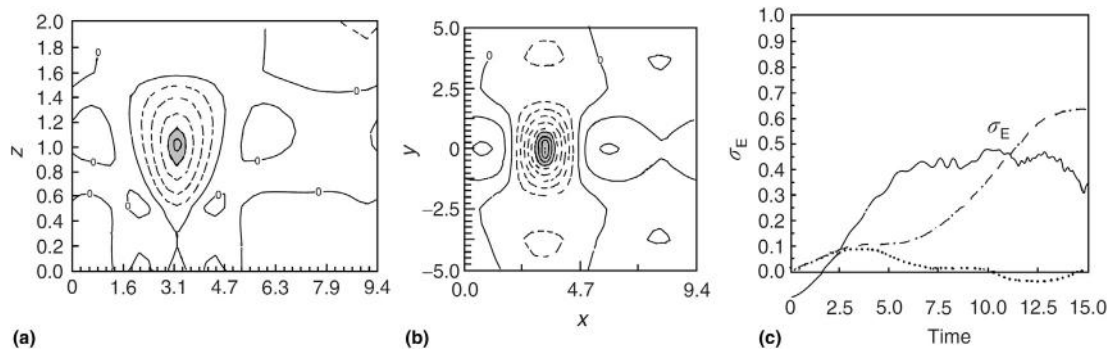


Figure 8 Initial value calculations for a linearly localized initial condition. (a) Zonal cross section showing contours of streamfunction initially. Values < -1.0 are shaded. (b) Horizontal pattern of streamfunction at tropopause level ($z = 1.0$) initially. Initial condition constructed from neutral modes having similar phase speed. Growing or decaying normal modes are excluded. (c) Time series of energy growth rate for three integrations. Linear model (dotted line) showing little growth since the nonmodal mechanism is weak and growing normal modes cannot develop. Also shown are nonlinear calculations for two amplitudes of the initial condition, where the solid line uses three times the initial amplitude of the dot-dashed line. Growing normal modes are activated by nonlinear interaction. Some evidence of nonlinear saturation is seen.

wave propagates from lower to higher amplitude of the envelope. ‘Spatial’ instability allows wave number to be complex while phase speed remains real.

Nonlinear calculations raise other issues related to baroclinic instability. One issue concerns equilibration. The growing wave modifies the mean flow while drawing energy from it. This places a limit upon the cyclone development. In PV theory, this may be where the distortion shown in [Figure 4](#) becomes comparable to the cyclone width. Waves longer than the most unstable wave tend to reach larger amplitude than the linearly most unstable mode. One reason why is that they are deeper and so can potentially tap more APE in the mean flow. Another reason may be the larger scale in both horizontal dimensions provides a longer time for PV contour distortion. Another reason concerns the inversion of a PV anomaly: the streamfunction amplitude is larger for a broader PV anomaly.

‘Life cycle’ studies model cyclones from birth to peak amplitude to decay. These studies typically find baroclinic growth followed by barotropic decay. This cycle fits the observed facts that eddies have a net heat flux and a net momentum convergence. These studies also reveal a characteristic evolution of the eddy structure: upper level amplification compared to the linear eigenmodes. An explanation is that saturation is reached sooner at the critical level and at the surface while upper levels continue to grow. Another was given above regarding the Eliassen–Palm flux F . When averaged over the life cycle, the vertical distribution of the zonal mean eddy heat and momentum fluxes becomes more realistic.

Finally, the atmosphere has higher order processes than the QG system. The biggest impact of ageostrophy is to break symmetries in the solutions. [Figure 7\(d\)](#) shows the leading order ageostrophic effects for a linear model. Ageostrophy causes enhanced eddy development on the poleward side (mainly by negative baroclinic conversion on the equatorward side), builds mean flow meridional shear, and slows down the wave. Ageostrophy also causes contours to be more closely spaced around a low and more widely spaced around a high.

See also: **Dynamical Meteorology:** Balanced Flow; Overview; Potential Vorticity; Quasigeostrophic Theory; Vorticity; Waves. **Synoptic Meteorology:** Cyclogenesis; Extratropical Cyclones.

Further Reading

- Gill, A., 1982. *Atmosphere–Ocean Dynamics*. Academic Press, New York.
- Grotjahn, R., 1993. *Global Atmospheric Circulations: Observations and Theories*. Oxford University Press, New York.
- Holton, J., 2004. *An Introduction to Dynamic Meteorology*, fourth ed. Academic Press, San Diego.
- Hoskins, B., McIntyre, M., Robertson, A., 1985. On the use and significance of isentropic potential vorticity maps. *Quarterly Journal of the Royal Meteorological Society* 111, 877–946.
- Pedlosky, J., 1987. *Geophysical Fluid Dynamics*, second ed. Springer-Verlag, New York.
- Pierrehumbert, R., Swanson, K., 1995. Baroclinic instability. *Annual Reviews in Fluid Mechanics* 27, 419–467.
- Vallis, G., 2006. *Atmospheric and Oceanic Fluid Dynamics: Fundamentals and Large-Scale Circulations*. Cambridge University Press, New York.

Coriolis Force

DW Moore, Pacific Marine Environmental Laboratory, Seattle, WA, USA

© 2003 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 1, pp 570–574, © 2003, Elsevier Ltd.

Introduction

The Coriolis force arises when Newton's equations of motion are written in a rotating coordinate system. It is named for Gaspard-Gustave de Coriolis (1792–1843). His studies of dynamical problems associated with rotating machinery were presented to the Académie des Sciences in Paris in 1831.

All of the terms in the equations of motion related to the Coriolis force were actually included in the tidal equation of PS Laplace, published in 1775 and 1776, and repeated in Book IV of his *Treatise on Celestial Mechanics*, published in 1799.

Newton's equation of motion for a particle is

$$m \frac{d\vec{v}}{dt} = \vec{F} \quad [1]$$

where m is the mass of the particle, $\vec{v} = d\vec{r}/dt$ is the velocity of the particle, $\vec{r}$ is the position vector, and d/dt is the time derivative following the motion of the particle. The quantity $\vec{a} = d\vec{v}/dt$ is the acceleration. This equation holds in an inertial reference frame, fixed with respect to the 'fixed' stars.

But the Earth on which mankind lives rotates at a nearly constant rate about an axis whose direction is more or less fixed in inertial space. The Earth's rotation vector is denoted by $\vec{\Omega}$, so $\Omega = |\vec{\Omega}|$ is the rotation rate. The direction of $\vec{\Omega}$ points from the South Pole to the North. The Earth rotates once each sidereal day, which is about 8.62×10^4 solar seconds. So the Earth's rotation rate is $\Omega = 7.29 \times 10^{-5} \text{ s}^{-1}$.

The motions of the atmosphere and the oceans are generally described in a coordinate system which rotates with the Earth. For example, longitude, latitude, and height above a reference geopotential surface form a convenient rotating system. Because the Earth is rotating, it bulges at the equator and is flattened at the poles. The reference geopotential includes both the gravitational and centrifugal potential, so the reference surface is not quite spherical. The Earth's equatorial radius exceeds the polar radius by about 21 km. Both the shape of the geopotential and the Coriolis force depend on the Earth's rotation rate.

Derivation of the Coriolis Force

Consider two coordinate systems, one of which is inertial, fixed with respect to the fixed stars; and the other rotating with angular velocity $\vec{\Omega}$. A prime is used to denote a quantity as viewed in the rotating frame and use unprimed quantities to denote the inertial frame. Let $\vec{q}$ be a fixed vector in the inertial frame, so that

$$\frac{d\vec{q}}{dt} = 0 \quad [2]$$

Viewed in the rotating frame, $\vec{q}'$ appears to rotate with angular velocity $-\vec{\Omega}$ in the direction opposite to the rotation of the coordinate system. So its apparent velocity in the rotating system is $-\vec{\Omega} \times \vec{q}'$, and $+\vec{\Omega} \times \vec{q}'$ must be added to this apparent velocity to bring it to zero. This leads to the rule

$$\frac{d}{dt}\vec{q} = \frac{d}{dt'}\vec{q}' + \vec{\Omega} \times \vec{q}' \quad [3]$$

where the terms on the left are viewed in the inertial frame and those on the right in the rotating frame. Furthermore, this same result, eqn [3], holds even if the left-hand side is not zero.

Consider a particle with position vector $\vec{r}$ in the inertial frame and $\vec{r}'$ in the rotating frame. The velocity is given by

$$\vec{v} = \frac{d\vec{r}}{dt} = \frac{d\vec{r}'}{dt'} + \vec{\Omega} \times \vec{r}' = \vec{v}' + \vec{\Omega} \times \vec{r}' \quad [4]$$

and the acceleration is

$$\begin{aligned} \vec{a} &= \frac{d\vec{v}}{dt} = \left(\frac{d}{dt'} + \vec{\Omega} \times \right) (\vec{v}' + \vec{\Omega} \times \vec{r}') \\ &= \frac{d\vec{v}'}{dt'} + 2\vec{\Omega} \times \vec{v}' + \vec{\Omega} \times \vec{\Omega} \times \vec{r}' \end{aligned} \quad [5]$$

The first term on the right is the acceleration as seen in the rotating system. The second term is the Coriolis acceleration. The third term is the centripetal acceleration arising from the rotating coordinate frame.

Modification of the Gravitational Potential

Let

$$R^2 = \vec{r}' \cdot \vec{r}' - \frac{1}{\Omega^2} (\vec{\Omega}' \cdot \vec{r}')^2 \quad [6]$$

Then R is the distance from the position $\vec{r}'$ to the axis of rotation, and it is easy to check that

$$\vec{\Omega}' \times \vec{\Omega}' \times \vec{r}' = \nabla' \left(-\frac{\Omega^2 R^2}{2} \right) \quad [7]$$

Since this is the gradient of a scalar, it can be combined with the gravitation potential to define an effective gravitational potential. This then accounts for the nonspherical shape of the rotating Earth and the reduction of the strength of the apparent gravitational force due to the centripetal acceleration.

Motions on a Rotating Earth

We now consider motions as observed on a rotating Earth, and drop the primes used earlier to denote the rotating coordinates.

Consider a local Cartesian coordinate system fixed with respect to the rotating Earth. Let x be positive toward the east, y positive toward the north, and z positive upward, opposite to the direction of the apparent gravitational force. Let $\hat{i}$, $\hat{j}$, and $\hat{k}$ be unit vectors in the directions of x , y , and z , and let u , v , and w be the components of the velocity in those directions. If particle motion is being considered then u , v , and w are the components of the particle velocity. If fluid motion is being considered then they are the components of the velocity of the fluid element at the positions x , y , and z . Let the origin of this system be at latitude θ_0 , which is the angle between the Earth's equatorial plane, and the local vertical direction. Then the Earth's rotation vector $\vec{\Omega}$ written in the local coordinate system is

$$\vec{\Omega} = \Omega \cos \theta_0 \hat{j} + \Omega \sin \theta_0 \hat{k} \quad [8]$$

The Coriolis force $\vec{F}_C = -2\vec{\Omega} \times \vec{u}$ is given by

$$\vec{F}_C = -2\Omega[(-\cos \theta_0 w + \sin \theta_0 v)\hat{i} - \sin \theta_0 u\hat{j} + \cos \theta_0 u\hat{k}] \quad [9]$$

This force is perpendicular to $\vec{u}$ (and to $\vec{\Omega}$) and therefore does no work on the moving particle or fluid. In the equation for the time rate of change of kinetic energy the Coriolis terms vanish identically. This is important to note because in most problems the part of the Coriolis force proportional to the locally horizontal component of $\vec{\Omega}$ can be shown by a scaling argument to be negligible. These are the terms proportional to $2\Omega \cos \theta_0$. When they neglected the $-2\Omega \cos \theta_0 \hat{w}$ term and the $+2\Omega \cos \theta_0 u\hat{k}$ term must both be omitted to avoid introducing a spurious source of kinetic energy. Likewise, if one of these terms is important then the other should be included as well. If these terms are omitted then the only quantity involving the rotation is $f_0 = 2\Omega \sin \theta_0$. This is called the Coriolis parameter. More generally,

$$f = 2\Omega \sin \theta \quad [10]$$

is the Coriolis parameter at latitude θ .

The f -Plane and the Beta Plane

If the problem under consideration involves a physical domain that includes the whole globe or a large portion of it, then spherical coordinates are a natural choice and the full dependence of the Coriolis force on latitude is retained. If the latitudinal extent of the physical domain is small enough that $f = 2\Omega \sin \theta$ is nearly constant, then f is approximated by the constant f_0 and studying the problem 'on an f -plane' is spoken of.

A more common situation which arises in both meteorology and oceanography is one in which the latitudinal extent of the domain is large enough so that the spatial variation of $f = 2\Omega \sin \theta$ needs to be considered, but the full spherical geometry does not. Then the latitude is written as $\theta = \theta_0 + y/R$, where R is the radius of the Earth and $\sin \theta$ is approximated by the first two terms in its Taylor expansion. It is written as

$$f = 2\Omega \sin \theta \approx 2\Omega \sin \theta_0 + \frac{2\Omega}{R} \cos \theta_0 y = f_0 + \beta y \quad [11]$$

The approximation $f = f_0 + \beta y$ is called the beta plane approximation. The parameter $\beta = (2\Omega/R) \cos \theta_0$ was first introduced by Carl Gustav Rossby in 1939.

Examples: The Foucault Pendulum

We now turn to some simple examples to illustrate the effect of the Coriolis force. The examples are the Foucault pendulum, inertial oscillations, Ekman layers, and geostrophic balance. For the Foucault pendulum, let a particle of mass m be suspended by a string of length ℓ . Take the origin at the equilibrium position of the mass, directly below the point of suspension. If the horizontal displacements are small compared to ℓ , then the x and y components of the tension in the string are approximately $-mg(x/\ell)\hat{i}$ and $-mg(y/\ell)\hat{j}$. In the absence of rotation the equations of motion would be $m\ddot{x} = -m(g/\ell)x$ and $m\ddot{y} = -m(g/\ell)y$, where the dots denote a time derivative. If the Coriolis force term is added and divided by m then is obtained

$$\ddot{x} - f\dot{y} + \omega^2 x = 0 \quad [12]$$

and

$$\ddot{y} + f\dot{x} + \omega^2 y = 0 \quad [13]$$

where $\omega^2 = g/\ell$ is the pendulum frequency in the nonrotating system. For positive f , it is seen that the effect of the Coriolis force is to accelerate the particle to the right of its motion. A solution of eqns [12] and [13] with initial conditions $x = x_0$, $y = 0$, $\dot{x} = 0$, and $\dot{y} = 0$ at $t = 0$ is given by

$$x + iy = x_0 \exp\left\{-i\frac{ft}{2}\right\} \left[\cos \omega't + \frac{if}{2\omega'} \sin \omega't\right] \quad [14]$$

where complex notation and $\omega'^2 = \omega^2 + f^2/4$ have been used.

The term in the square bracket describes an elliptical orbit, almost rectilinear if $f' \ll \omega'$, with the pendulum oscillating at frequency ω' . The exponential term indicates that this elliptical orbit slowly rotates clockwise in the Northern Hemisphere, at frequency $f/2$. This rotation is due to the deflection of the orbit by the Coriolis force. The frequency $f/2 = \Omega \sin \theta$. At the North Pole, the period of the exponential terms is 24 h, and can be thought of as the Earth rotating under the swinging pendulum.

Note that the solution given by eqn [14] never passes through the origin. A slight change in initial conditions, with $x = x_0$, $y = 0$, $\dot{x} = 0$, and $\dot{y} = -f/2x_0$ at $t = 0$, produces the solution

$$x + iy = x_0 \exp\left\{-i\frac{ft}{2}\right\} [\cos \omega't] \quad [15]$$

The $[\cos \omega't]$ represents a true rectilinear orbit through the origin and the exponential factor describes the rotation of the plane of the orbit; the nonzero initial value of $\dot{y}$ gives the pendulum just the correct initial velocity to produce this solution.

In the Foucault pendulum problem, the Coriolis force is a relatively small perturbation on the orbit of the pendulum as it goes back and forth once, but over the course of a day it has a substantial cumulative effect. In most atmospheric and oceanic problems, interest is shown on timescales of a day or longer. The Coriolis force then becomes one of the dominant terms in the equations of motion. In the remaining illustrations the Coriolis term is balanced by various other possible important terms.

Geostrophic Balance

If the motions are varying slowly enough with time to be regarded as nearly steady, and turbulent dissipation is small, then the approximate balance in the horizontal momentum equation is

$$+\frac{1}{\rho}\frac{\partial p}{\partial x} = fv \quad [16]$$

and

$$+\frac{1}{\rho}\frac{\partial p}{\partial y} = -fu \quad [17]$$

where p is pressure, ρ the density, and u , v , and f the velocity components and the Coriolis parameter. Equations [16] and [17] show that the pressure gradient on the left balances the Coriolis force on the right. This is called geostrophic balance and is illustrated elsewhere in this encyclopedia (see **Dynamical Meteorology: Overview**, Figure 1). So, if the pressure field is known then the horizontal velocities can be estimated from eqns [16] and [17].

Inertial Oscillations

Various departures from geostrophic balance are possible, and for many of these the Coriolis force still plays a dominant role. Inertial oscillations provide one such example. In the simplest case, horizontal pressure gradients are ignored and horizontal accelerations by the Coriolis force are balanced. The governing equations are

$$\frac{du}{dt} = fv \quad [18]$$

and

$$\frac{dv}{dt} = -fu \quad [19]$$

If the initial condition is $u = U_0$, $v = 0$, then the solution is

$$u = u_0 \cos(ft) \quad [20]$$

$$v = -u_0 \sin(ft) \quad [21]$$

or in complex form

$$u + iv = u_0 \exp\{-ift\} \quad [22]$$

Equation [22] clearly shows that the velocity vector rotates at frequency f , clockwise in the Northern Hemisphere and counterclockwise in the Southern. Such oscillations have been observed in the sea.

One such observation lasting nearly a week was made in the Baltic Sea by Gustafson and Kullenberg. It is described in some detail by Sverdrup, Johnson, and Fleming in their book *The Oceans* (see Figure 1).

Ekman Layers

Another important example is the Ekman layer, where the Coriolis force is balanced by vertical mixing. The original work, published by V Walfrid Ekman in 1905, is called *On the*

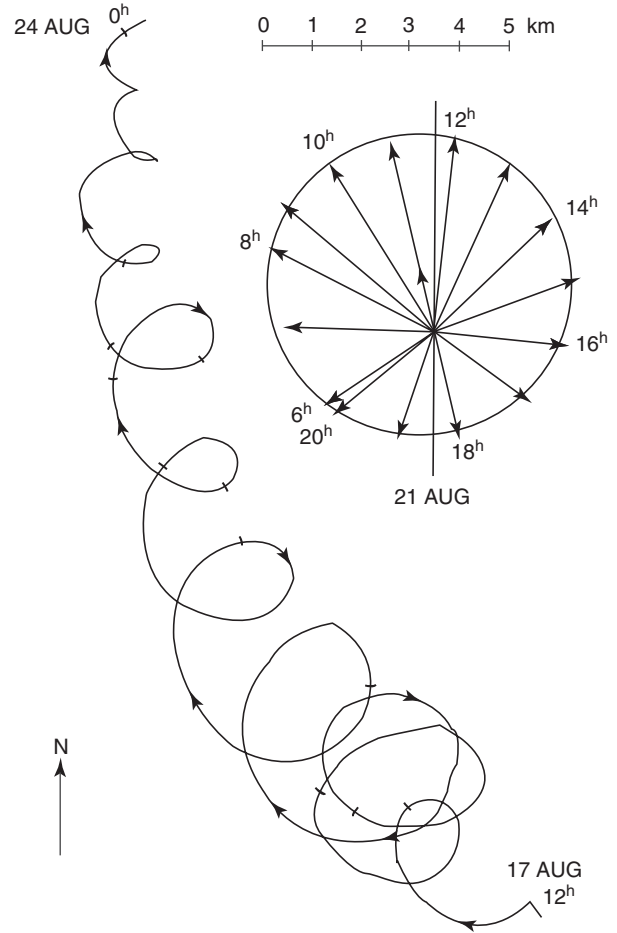


Figure 1 Rotating currents of period one-half pendulum day observed in the Baltic Sea and represented by a progressive vector diagram for the period, 17–24 August 1933, and by a central vector diagram between 06.00 and 20.00 on 21 August (according to Gustafson and Kullenberg). Adapted from Sverdrup, H.U., Johnson, M.W., and Fleming, R.H., 1942. *The Oceans: Their Physics, Chemistry and General Biology*. New York, NY: Prentice-Hall.

Influence of the Earth's Rotation on Ocean Currents. Ekman considered the problem of determining the motion of the upper layer of the ocean due to a wind stress acting on the sea surface. Let z measure distance positive upward from the sea surface, so the domain of interest is $z < 0$. In this model, the only horizontal forces acting on the fluid are the Coriolis force and the divergence of the stress tensor describing the turbulent mixing. The stress is modeled in terms of an eddy viscosity as

$$\tau_z^x = \mu \frac{du}{dz} \quad [23]$$

and

$$\tau_z^y = \mu \frac{dv}{dz} \quad [24]$$

where τ_z^x means the stress in the x direction acting on a surface whose normal is in the z direction. The quantity μ is the

coefficient of eddy viscosity and $\nu = \mu/\rho$ is taken as a constant. The equations of motion in the x and y directions are

$$-fv = \frac{\partial}{\partial z} \left(\nu \frac{\partial u}{\partial z} \right) \quad [25]$$

and

$$fu = \frac{\partial}{\partial z} \left(\nu \frac{\partial v}{\partial z} \right) \quad [26]$$

If a stress of magnitude τ_x^0 acts in the x direction on the sea surface, the boundary conditions are

$$\nu \frac{\partial u}{\partial z} = \frac{\tau_x^0}{\rho} \quad \text{on } z = 0$$

and

$$\nu \frac{\partial v}{\partial z} = 0 \quad \text{on } z = 0 \quad [27]$$

Furthermore, the stresses are assumed to vanish as $z \rightarrow -\infty$. If it is written as

$$q = u + iv \quad [28]$$

then eqns [25] and [26] can be combined in the form

$$\nu q_{zz} = ifq \quad [29]$$

and for $f > 0$ (Northern Hemisphere), a solution which vanishes as $z \rightarrow -\infty$ is

$$q = A \exp \left\{ \frac{(1+i)}{\sqrt{2}} \sqrt{\frac{f}{\nu}} z \right\} \quad [30]$$

Applying the boundary condition given in eqn [27] determines A , and the solution is

$$q = u + iv = \frac{\tau^0}{\rho} \sqrt{\frac{2}{\nu f}} \frac{(1-i)}{2} \exp \left\{ (1+i) \sqrt{\frac{f}{2\nu}} z \right\} \quad [31]$$

Equation [31] describes a flow which at the sea surface $z = 0$ is directed at 45° to the right of the surface stress. The direction

of the flow spirals to the right as z becomes more negative. The net vertically integrated transport can be determined by integrating eqns [25] and [26] and applying the boundary conditions (eqn [27]). The result is

$$U = \int_{-\alpha}^0 u dz = 0 \quad V = \int_{-\alpha}^0 v dz = \frac{-\tau_x^0}{f} \quad [32]$$

If a nonzero stress in the y direction, τ_y^0 , had been specified as well, there would be obtained

$$U = \frac{\tau_y^0}{f} \quad V = -\frac{\tau_x^0}{f} \quad [33]$$

So the net transport in the frictional boundary layer is 90° to the right of the wind stress in the Northern Hemisphere. In the Southern Hemisphere, the transports are 90° to the left of the wind stress, and surface currents 45° to the left.

Since $f = 2\Omega \sin \theta$ vanishes at the equator, there is an equatorial transition from Northern to Southern Hemisphere behavior. All the examples discussed here apply to regions away from the equator. In the tropics, other terms become important in the equations describing the motion.

See also: Dynamical Meteorology: Overview.

Further Reading

- Chandrasekhar, S., 1963. *Ellipsoidal Figures of Equilibrium*. Yale University Press, New Haven, CT [reprinted with minor changes, New York: Dover, 1987].
- Ekman, V.W., 1905. On the influence of the Earth's rotation on the ocean currents. *Arkiv for Matematik, Astronomi, Och Fysik* (Stockholm) 2, 1–53.
- Rossby, C.G., Collaborators, 1939. Relation between variations in the intensity of the zonal circulation of the atmosphere and the displacements of the semipermanent centers of action. *Journal of Marine Research* 2, 38–55.
- Stommel, H.S., Moore, D.W., 1989. *An Introduction to the Coriolis Force*. Columbia University Press, New York, NY.
- Sverdrup, H.U., Johnson, M.W., Fleming, R.H., 1942. *The Oceans: Their Physics, Chemistry and General Biology*. Prentice-Hall, New York, NY.

Critical Layers

P Haynes, University of Cambridge, Cambridge, UK

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

The terms critical line (or critical level) and critical layer, for small-amplitude waves propagating on a shear flow, are defined. At the critical line the simplest linear steady dissipationless theory for waves breaks down and must be resolved by adding other physical processes, such as unsteadiness, nonlinearity, or dissipation in a critical-layer region, which is a finite neighborhood of the critical line. The critical layer for Rossby waves is discussed in detail, highlighting the fact that the behavior within the critical layer affects the waves outside it, in particular determining whether or not the critical layer acts as an absorber, reflector, or overreflector to incident waves. In the case of a nonlinear critical layer, where dissipation can be largely neglected, this effect changes in time, so that the critical layer evolves from an absorbing state to a reflecting state, to an overreflecting state and then oscillates between weak overreflection and weak absorption, tending to a perfectly reflecting state. The behavior seen in the Rossby-wave nonlinear critical layer serves as a paradigm for the more general phenomenon of wave breaking, both with respect to the rearrangement of potential vorticity in the wave-breaking region and to the effect on neighboring wave-propagation regions. The corresponding behavior for internal-gravity waves is also discussed, which is more complicated than for Rossby waves, since there is wave propagation within the critical layer itself, so that the critical layer dynamics is a combination of wave propagation and nonlinear or dissipative behavior.

Introduction

Theoretical models of waves in the atmosphere naturally require consideration of propagation on a background state that is a shear flow. One example is that of Rossby waves (or planetary waves) propagating from the extratropical troposphere into the stratosphere. The background state is here the longitudinally averaged flow, which may include westerly winds increasing in strength with height (e.g., in the winter) or westerly winds at lower levels changing to easterly winds at upper levels (e.g., in the summer). Another example is that of small-scale internal-gravity waves excited by a mountain propagating upward through a large-scale flow that changes strength (and perhaps direction) with height.

Suppose that the background flow (i.e., the flow in the absence of the waves) is in the x -direction with speed U that is a function of a second space coordinate y and that the waves have a well-defined phase speed c in the x -direction. Then a location where $U(y) = c$, i.e., the flow speed matches the phase speed, is a line parallel to the x -axis and at a fixed value of y , called a *critical line*. Where the second space coordinate is height, the equivalent term *critical level* is often used. If the speed U was a function of two space coordinates y and z then the location $U(y, z) = 0$ would define a *critical surface*.

Simple theories for the structure of waves are often based on the assumptions that the waves are steadily propagating, that dissipative or diabatic processes such as friction or radiative transfer may be neglected, and that the waves are small-amplitude, so that terms in the equations of motion that are nonlinear in wave quantities may be neglected. These theories lead to a straightforward differential or partial differential equation that describes the spatial structure of the waves. The importance of the critical line is that it is a location where these differential equations are singular, in other words the solutions imply that some physical quantity becomes infinite.

As in many physical contexts, the appearance of singular behavior in a mathematical model implies that the

simplifications that lead to that model cannot be justified and that some physical process that was neglected must be retained. To remove the critical-line singularity one of the neglected processes mentioned above must therefore be included (however weak such processes might have been estimated to be). The neglected process will be essential only in a small, but finite, region around the critical line and may still be negligible elsewhere. This small but finite region is named the *critical layer*.

The Rossby-Wave Critical Layer

A Simple Model

One of the simplest examples of the critical-line singularity and its resolution in a finite critical layer arises in a two-dimensional model of Rossby-wave propagation on a β -plane (a mathematical device to include the effect of the variation of Coriolis parameter with latitude). Two-dimensional Cartesian coordinates (x, y) may be used, with x measured in the eastward direction and y measured in the northward direction. The corresponding velocity components are taken to be (u, v) . The assumption of incompressibility implies that the velocity components may be expressed in terms of a stream function $\psi(x, y, t)$ (with t time) where $u = -\partial\psi/\partial y$ and $v = \partial\psi/\partial x$.

The governing equation is based on the fact that, in the absence of dissipation, the absolute vorticity, which is the sum of the relative vorticity, $\zeta = (\partial v/\partial x) - (\partial u/\partial y) = \nabla^2\psi$, and the planetary vorticity, βy , is conserved following the fluid motion. β is a constant and in an Earth-like atmosphere is positive. It is convenient to include linear damping of vorticity in the model as a simple representation of a dissipative process. The governing equation then becomes

$$\left(\frac{\partial}{\partial t} + u\frac{\partial}{\partial x} + v\frac{\partial}{\partial y}\right)(\zeta + \beta y) = -\alpha\zeta \quad [1]$$

where α is the damping rate. (Another possibility for a dissipative process would be diffusion of vorticity. Neither linear

damping nor diffusion is likely to be realistic representations of dissipative processes in the real atmosphere, but either serves as a convenient example that captures the basic effect of dissipation in the critical layer.)

It is assumed that in the absence of waves the flow is in the x -direction with speed $U(y)$. Waves are superimposed on this flow giving a contribution $\psi'(x, y, t)$ to the stream function. Then the eqn [1] may be written in terms of ψ' as

$$\begin{aligned} \nabla^2 \psi'_t + U(y) \nabla^2 \psi'_x + [\beta - U''(y)] \psi'_x \\ = -\alpha \nabla^2 \psi' - u' \frac{\partial \nabla^2 \psi'}{\partial x} - v' \frac{\partial \nabla^2 \psi'}{\partial y} \end{aligned} \quad [2]$$

where $u' = -\partial \psi' / \partial y$ and $v' = -\partial \psi' / \partial x$ are the wave velocity components.

If the damping is weak then it is reasonable to neglect the first term on the right-hand side. If the waves are of small-amplitude then it is reasonable to neglect the second and third terms on the right-hand side, which are quadratic in wave quantities. Since the resulting equations are linear and contain no explicit x -dependence it is possible to consider waves with different wavelengths in the x -direction as independent.

Consider one such wave, with x -wavenumber k (i.e., wavelength $2\pi/k$ in the x -direction), assumed to be steadily propagating in the x -direction with phase speed c . It follows that the stream function for this wave may be written in the form $\psi'(x, y, t) = \text{Re}[\hat{\psi}(y)e^{ik(x-ct)}]$, where $\hat{\psi}(y)$ is a (complex) function of y . Substituting into eqn [2] and neglecting terms on the right-hand side gives the ordinary differential equation

$$\hat{\psi}'' + \left(\frac{\beta - U''(y)}{U(y) - c} - k^2 \right) \hat{\psi} = 0 \quad [3]$$

This equation is known as the Rayleigh–Kuo equation and, depending on context, determines the stability of the shear flow $U(y)$ in a problem where c is an eigenvalue, or when c is determined by forcing, which may be included in the problem by an extra term on the right-hand side of eqn [3] or by a boundary condition, describes through the function $\hat{\psi}(y)$ the structure of the forced disturbances. The focus here is on a steady forced problem, where c is given and real.

The appearance of the factor $[U(y) - c]^{-1}$ in part of the coefficient of $\hat{\psi}$ indicates that the equation has a singular point at values of y such that $U(y) = c$, i.e., where the phase speed matches the flow speed. These locations are the critical lines. If $U(y)$ is an increasing or decreasing function of y then there is at most one critical line. If $U(y)$ has a turning point (as would be the case for a jetlike flow, for example) then there may be more than one critical line.

Consider the solution near a critical line at $y = y_c$. The nature of eqn [3] is such that whilst $\hat{\psi}(y)$ (proportional to the velocity in the y -direction) is finite and continuous in the neighborhood of the critical line, $\hat{\psi}''(y)$ (representing part of the vorticity) is proportional to $(y - y_c)^{-1}$ and $\psi'(y)$ (representing the velocity in the x -direction) is proportional to $\log |y - y_c|$, i.e., both are singular. This is clearly unphysical, but what is more problematic, in a way, is that the singularity in $\hat{\psi}'(y)$ implies that there is no unique way to match solutions of eqn [3] across $y = y_c$. In particular the jump in $\hat{\psi}'(y)$, corresponding to the jump in u' , across $y = y_c$ remains unknown. It follows there is no unique solution in the whole flow domain for the

function $\hat{\psi}(y)$. The critical line singularity must therefore be resolved not only to remove the local singular behavior in certain physical quantities but also to determine the structure of the waves over the whole flow.

Absorption–Reflection

To note the implications of the critical layer for the waves elsewhere in the flow it is useful to focus on the following geometry, shown in Figure 1. Assume that the waves are forced at some large positive value of y , with phase speed $c = 0$, i.e., the waves are stationary. Assume also that the flow speed $U(y)$ is positive in $y > 0$ and negative in $y < 0$, so that the waves have a critical line at $y = 0$ and that the curvature term $U''(y)$ is not too large (so that $\beta - U''(y)$ is positive). The eqn [3] predicts that the function $\hat{\psi}(y)$ is oscillatory in $y > 0$, implying that there are propagating waves (as expected from the basic properties of Rossby waves). In $y < 0$, on the other hand, the function $\hat{\psi}(y)$ is exponentially increasing or decreasing with y and physical considerations require that $\hat{\psi}(y)$ decreases as y decreases, representing evanescent waves. One feature of the solution that is naturally of interest is the relative amount of northward- and southward-propagating waves in the region between the wave forcing and the critical layer. This measures the absorption–reflection behavior of the critical layer. If the critical layer acted as an absorber of waves then the region between the forcing and the critical line would contain only waves propagating southward. If it acted as a reflector of waves then there would be some contribution to the solution in this region from waves propagating northwards. The reflection could be partial or perfect. Indeed there could in principle be over-reflection, which would be associated with a greater proportion of northward-propagating than southward-propagating waves, implying that the critical layer was actively emitting waves.

A useful quantitative measure of the wave propagation is the momentum flux, $\overline{u'v'}$, where $\overline{(\cdot)}$ indicates an average in the x -direction (keeping y and t constant). $\overline{u'v'}$ indicates the correlation between the two velocity components and the basic properties of Rossby waves imply $\overline{u'v'} > 0$ for southward-propagating waves and $\overline{u'v'} < 0$ for northward-propagating waves. It follows from eqn [3] that $\overline{u'v'}$ is constant in $y > 0$, except at the critical line at $y = 0$. The evanescence of the waves in $y < 0$ implies that $\overline{u'v'} = 0$ there. However, the value of $\overline{u'v'}$ cannot be determined from eqn [3] alone. Instead the critical line must be resolved into a finite critical layer to allow the jump in $\overline{u'v'}$ across the layer, denoted by $[\overline{u'v'}]_+^+$, and hence the value of $\overline{u'v'}$ in $y > 0$, to be evaluated.

The continuity of $\hat{\psi}(y)$ across the critical line singularity suggests that, when the critical line is resolved into a thin critical layer, ψ' and hence $v' = \partial \psi' / \partial x$ will vary only weakly across the critical layer. In addition, y -derivatives within the critical layer will generally be much larger than x -derivatives (because the layer is thin), implying that $\zeta' \approx -\partial u' / \partial y$. Putting these pieces of information together, it follows that

$$[u']_+^+ = \int \zeta' dy \text{ and } \overline{u'v'} \text{ (in } y > 0) = [\overline{u'v'}]_+^+ = \int \overline{\zeta'v'} dy \quad [4]$$

where $[u']_+^+$ denotes the jump in u' across the critical layer, the integrals are taken across the critical layer, and v' may be taken as constant within the second integral. The first equality is the

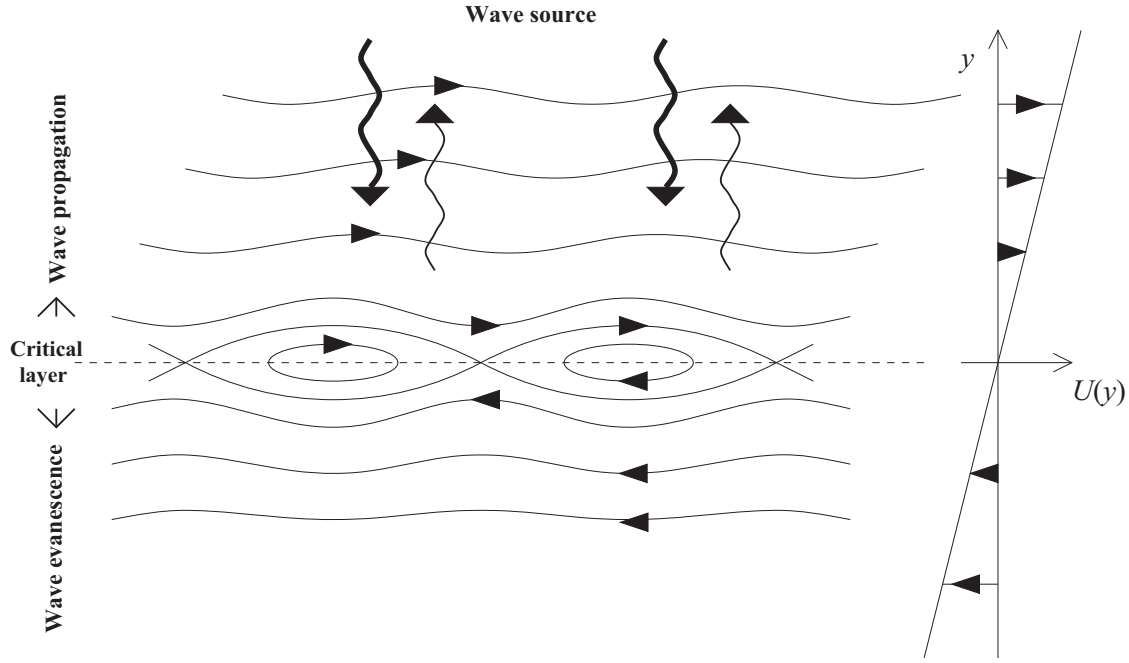


Figure 1 Schematic diagram of Rossby-wave propagation on a shear flow $U(y)$ with a critical line. The flow is positive (i.e., eastward) in $y > 0$ (upper portion of the diagram) and negative (i.e., westward) in $y < 0$ (lower portion of the diagram). The waves are forced, with zero phase speed in x -direction, in $y > 0$ and propagate toward $y = 0$. In $y < 0$ the waves are evanescent (i.e., nonpropagating and decaying as y becomes more negative). The critical line is at $y = 0$, where $U(y) = 0$. In the neighborhood of $y = 0$ the streamlines are closed and form a Kelvin's cat's eye pattern. The width of the closed-streamline region, which increases as the wave amplitude increases, defines the width of the nonlinear critical layer. If dissipation were strong enough then dissipative effects would dominate over a relatively broad region near $y = 0$ and the closed streamlines would essentially be irrelevant to the dynamics. (The critical layer would then be linear and dissipative, rather than nonlinear.) There may be some reflected wave in $y > 0$, but the amount of reflection can be determined only by considering the detailed dynamics of the critical layer.

missing matching condition across the critical layer. The second shows that the critical layer acts as a net absorber of waves when there is (in a y -integrated sense) negative correlation between ζ' and v' in the critical layer, as a perfect reflector when there is zero correlation and as a net emitter (i.e., an over-reflector) when there is positive correlation.

To summarize, the nonuniqueness in the solution of eqn [3] left by the critical-line singularity leaves the absorption-reflection behavior of the critical layer uncertain. Only by determining the correlation between v' and ζ' in the critical layer is it possible to determine the precise absorption-reflection properties.

The Dynamics of the Critical Layer

The dynamical balance in the critical layer depends on the parameters of the problem. Consider in turn the processes that have been neglected in arriving at eqn [3]. Firstly it has been assumed that the waves are steadily propagating (i.e., that their amplitude is not changing with time). It is possible to analyze the nondissipative, linearized equations (eqn [2] with the right-hand side set to zero) without making this assumption and show that the singular behavior predicted by eqn [3] develops with time. For example, both the vorticity ζ' and the x -component of velocity u' are predicted to increase without bound. The time-dependent analysis shows that the terms neglected in going from eqns [2] to [3] inevitably become

important at large times, however small they might have first appeared.

Secondly, consider the dissipative term $-\alpha \nabla^2 \psi'$ in eqn [2]. This may be compared with the advection term $U(y) \partial \nabla^2 \psi' / \partial x$. The relative sizes of these terms, near to $y = 0$, may be estimated as $\alpha / k U'(0) y$ and it follows that the dissipative term cannot be neglected in a region of size $\delta_\alpha = \alpha / [k U'(0)]$. This is (potentially) the thickness of the dissipative critical layer.

Finally, consider the nonlinear term. It turns out that the most important part is $v' \partial \nabla^2 \psi' / \partial y$. If this is to balance $U(y) \partial \nabla^2 \psi' / \partial x \approx U'(0) y \partial \nabla^2 \psi' / \partial x$ in a thin region of thickness δ_{NL} , then $(v' / \delta_{NL} \approx k U'(0) \delta_{NL})$, i.e., $\delta_{NL} \sim [v' / k U'(0)]^{1/2}$. This is (potentially) the thickness of the nonlinear critical layer.

Whether nonlinearity or dissipation is dominant in the critical layer depends on the relative size of $\delta_{NL} / \delta_\alpha = [k v' U'(0)^{1/2} / \alpha]$. If $\delta_{NL} / \delta_\alpha \ll 1$ then the critical-layer dynamics are dominated by dissipation and the critical layer thickness is δ_α . If $\delta_{NL} / \delta_\alpha \gg 1$ then the critical-layer dynamics are dominated by nonlinear processes and the critical layer thickness is δ_{NL} . In the case of Rossby waves in the real atmosphere, wave amplitudes are relatively large and dissipation is relatively weak, so that the nonlinear dynamics are the most relevant.

The fully nonlinear equations state that $\zeta + \beta y$ is conserved following the flow (which is in turn determined by the ζ field). If the critical layer is thin, i.e., δ_{NL} is small, there is a simplification because the flow may be approximated by the superposition of the basic flow $U(y) \approx U'(0) y$ and the y -component

of the disturbance velocity field, which is simply a function of x , because of the continuity of ψ' across the critical layer. This superposition gives a flow whose streamlines form a pattern known as Kelvin's cat's eyes, with closed streamlines near $y = 0$ (see Figure 1). The dynamics of the nonlinear critical layer is therefore that fluid particles are advected around these streamlines, conserving their values of $\zeta + \beta y$. The rearrangement of the $\zeta + \beta y$ field changes the ζ' field, thereby changing $[u']_+^+$ and hence the structure of the waves outside the critical layer. Furthermore there is a corresponding change in the correlation between v' and ζ' , which determines the absorption–reflection properties.

A schematic diagram of the evolution of the vorticity field in a simple nonlinear critical layer is presented in Figure 2 and the absorption–reflection properties deduced. At early times v' and ζ' are anticorrelated and the critical layer acts as an absorber. If there is strong dissipation then the vorticity in the center of the critical layer is essentially frozen in its early time configuration and the critical layer continues to act as an absorber at later times. (Detailed calculation shows that in this early time/dissipative regime, the absorption is effectively perfect.) However if dissipation is weak then the advective rearrangement continues and, after about half a turn-round time for the closed-streamline flow, the ζ' field in the center of the critical layer (which gives the major contribution to the integral) is such that there is no y -integrated correlation with the v' field, i.e., the critical layer acts as a perfect reflector. According to this particular model, the advective rearrangement continues to give a positive correlation between v' and ζ' and hence overreflection and the critical layer subsequently oscillates between a weakly absorbing and weakly overreflecting state, converging to a state of perfect reflection at large times.

The precise details of the evolution depend on the particular flow configuration. However, a general description of absorption–reflection behavior can be formulated by considering $-\overline{u'v'}$ as the flux (in the y -direction) of wave activity (i.e., a quantity that is positive when waves are present and zero when they are not). In the early time absorbing stage wave activity builds up in the critical layer. As the reflecting stage approaches, the rate of buildup decreases to zero and in the overreflecting stage the critical layer reemits some of its wave activity. If there is dissipation then the flux of wave activity into the critical layer may be balanced by local dissipation of wave activity and an absorbing state may persist. However, for the critical layer to continue to act as an absorber without dissipation, then the amount of wave activity in the critical layer must continue to increase. The total amount of wave activity in the critical layer may be shown to depend on the thickness of the region over which the vorticity field has been rearranged (i.e., the thickness of the critical layer). If this thickness is finite then there is an upper bound to the total amount of wave activity that can be stored there and it is therefore not possible to sustain absorption. In such a case the long-time average of the flux of wave activity must approach zero and one can therefore say that the long-time average behavior is perfect reflection. The only way that absorption could be sustained in the long-term would be if the thickness of the critical layer systematically increased in size.

A complementary viewpoint comes from considering $\overline{u'v'}$ as a momentum flux. In the absorbing stage $[\overline{u'v'}]_+^+$ is positive,

implying that there is a negative force exerted on the x -averaged flow in the critical layer. The time-averaged perfect reflection in the case where dissipation is zero translates into no time-averaged x -average force exerted on the flow in the critical layer. (If there was such a force, then the critical line, and hence the critical layer, would move closer and closer to the wave source.) Sustained absorption where there is dissipation translates into a nonzero time-averaged x -average force exerted on the flow in the critical layer, with this force being balanced by forces provided by dissipative processes (i.e., by the linear vorticity damping in the model described above).

The critical layer theory makes clear the nature of the two-way interaction between the wave propagation region outside the critical layer, and the flow in the critical layer itself. The waves outside the critical layer directly determine the flow pattern inside it (because of the continuity of v' across the critical layer). However, inside the critical layer the flow changes the vorticity field and hence the jump in u' across the critical layer, thereby changing the waves outside it. It is important to note that there is no wave propagation within the critical layer itself. The dynamics is simply that of vorticity advection by a simple cat's eye flow whose structure is determined by the waves outside the critical layer. It is not the case that waves can be said to propagate into the critical layer and are reflected by the structure of the flow profile that they encounter there.

Wave Breaking

The behavior seen in the nonlinear critical layer for δ_{NL} small may be interpreted as an example of the breaking of Rossby waves. By 'breaking' it is meant that the material contours or surfaces that would, in wave propagation, be reversibly undulated, are strongly and irreversibly deformed. The most familiar example of wave breaking occurs for surface waves. There the wave dynamics is associated with the undulation of the ocean surface. Waves might be forced in one region (e.g., by a storm), and propagate through large distances. The presence of the waves in this propagation stage is associated with distortion of the ocean surface, but the distortion is weak and reversible. As the waves enter shallow water in a coastal region the distortion of the ocean surface becomes stronger and, ultimately, complex and irreversible and the flow will become three-dimensionally turbulent.

Rossby-wave propagation involves the reversible distortion of contours of potential vorticity (absolute vorticity in the simple two-dimensional context discussed above). In the critical layer region the distortion of these contours is strong and irreversible and the waves may be said to be breaking. Indeed in many cases the flow in the critical layer may be shown to involve a sort of turbulence (quasi-geostrophic or two-dimensional), but this is not essential for the behavior to be described as breaking. As in the surface-wave case, where the breaking may allow the waves to drive systematic long-shore currents, the breaking of Rossby waves allows a systematic force to be exerted by the waves.

Implications for the Atmosphere

In the nonlinear Rossby-wave critical layer described above there is a clear division (described by simplified mathematical

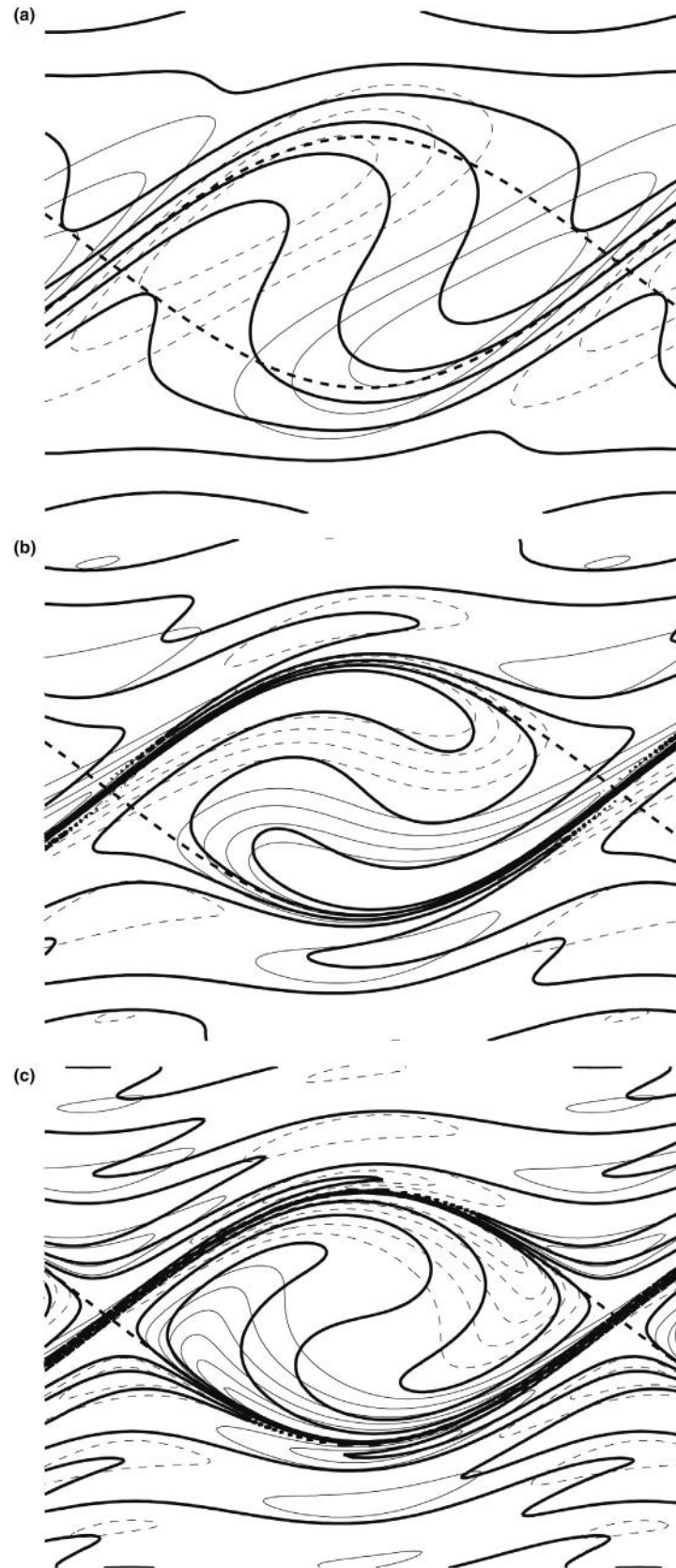


Figure 2 Evolution of vorticity field in the nonlinear critical layer. The panels show an expanded view of the vorticity field in the closed-streamline region indicated in [Figure 1](#). Note that this region may be very thin in the y -direction if the wave amplitude is small. Thick dotted curves are the bounding closed streamlines. Thick solid curves are contours of absolute vorticity $\zeta + \beta y$. Thin curves are contours of wave relative vorticity ζ' , with solid curves indicating positive values and dashed curves indicating negative values. Three panels show (a) absorbing stage, where y -velocity v' is negatively correlated with ζ' , (b) reflecting stage, where correlation between v' and ζ' is close to zero, and (c) overreflecting stage, where v' is positively correlated with ζ' .

equations that may be derived by a formal approximation procedure) between the broad region outside the critical layer where the dynamics is wavelike and the thin critical layer itself where the dynamics is a strong advective rearrangement of the potential vorticity or absolute vorticity field, which might be called wave breaking. In the real atmosphere the wave amplitudes are very large and the formal estimate δ_{NL} of the nonlinear critical layer thickness is generally as large as the other length scales in the problem. The same simplified mathematical equations do not hold precisely. Nonetheless observations and numerical models show clearly that there are regions of wave propagation and regions of wave breaking and that these exist side-by-side. There are at least two important examples. One is in the winter stratosphere, where planetary-scale Rossby waves propagate up from the troposphere, distort and shift the polar-night vortex, and break in what is now called the stratospheric ‘surf zone’ (which covers a large region of the midlatitudes and subtropics). The stirring of potential vorticity in the surf zone leads to weak large-scale gradients in the interior of the surf zone and corresponding strong gradients at its poleward edge, which is the boundary of the polar vortex. These strong potential vorticity gradients act as a kind of wave guide for upward propagation of Rossby waves. A second is in the upper troposphere and lower stratosphere, where synoptic-scale baroclinic eddies lead to a wavelike distortion of the subtropical jet and to wave-breaking regions on the poleward and equatorward sides of the jet. Again the effect on potential vorticity (PV) gradients is to lead to a kind of wave guide for tropospheric Rossby waves.

In these examples nonlinear critical layer theory provides quantitative guidance as to how the different regions interact. For example, it indicates that the wave-breaking regions may be considered to absorb, reflect, or emit wave activity and that the waves may have a systematic effect on the flow in the wave-breaking region.

In the last 10 years or so Rossby-wave critical layer ideas have also been applied to hurricane dynamics. For example, the role of nonaxisymmetric Rossby-wave disturbances to a hurricane in promoting hurricane intensification have been described in terms of the propagation of such disturbances and their subsequent dissipation in a critical layer. In a separate line of argument, the nonlinear critical layer of a subtropical Rossby wave has been interpreted as a preferred location for hurricane vortex development, with one stage of the development being a period of copropagation (in longitude) of the preexisting wave and the growing vortex.

Internal-Gravity Wave Critical Layers

Description

Critical lines and critical layers arise generically in any problem of wave propagation in a fluid. Another example that is particularly important for the atmosphere is that of internal-gravity waves. This has some important differences from the Rossby-wave case.

Consider the propagation of internal-gravity waves on a background state that has stable stratification with buoyancy frequency $N(z)$ and flow in the x -direction with speed $U(z)$, where z is height. Assuming that the flow is incompressible (which is not necessarily defensible for many atmospheric

gravity waves, but the model serves to illustrate important points that continue over to the compressible case) it may be shown that the analogue of eqn [3] is

$$\hat{\psi}'' + \left\{ \frac{N^2(z)}{[U(z) - c]^2} - \frac{U''(z)}{U(z) - c} - k^2 \right\} \hat{\psi} = 0 \quad [5]$$

This equation is known as the Taylor–Goldstein equation and, as is the case for eqn [3] for Rossby waves, depending on the context determines the stability of density-stratified shear flows or the structure of waves propagating on such flows. Again the critical line singularity at $U(z) = c$ is manifested by the inverse powers of $U(z) - c$ appearing in the expression multiplying $\hat{\psi}$. There is an important difference from eqn [3] in that one of the expressions contains the factor of $[U(z) - c]^{-2}$. This means that the behavior of solutions near the critical-line singularity is quite different from the Rossby-wave case. In fact, provided that the local Richardson number $Ri = N^2(z)/U'(z)^2 > 1/4$ (which is precisely the condition required for the flow to be stable) the function $\hat{\psi}(z)$ oscillates rapidly in z near to the critical line and the oscillations become infinitely rapid as the critical line is approached. Indeed there are infinitely many oscillations before the critical line is reached. These oscillations are a manifestation of the rapid shrinking of the vertical wavelength of the wave as the critical line is approached, due to the tilting of the wave by the shear. An analogous shrinking of the wavelength occurs in the Rossby-wave case, but there are only a finite number of oscillations before the critical line is reached – a subtle and important difference between this and the internal-gravity wave case. The reason for the difference is that in the internal-gravity wave case decrease in wavelength gives a stronger decrease in the group velocity (i.e., the propagation velocity for wave packets). Indeed in the Rossby-wave case the idea of group velocity is simply not at all useful in the neighborhood of the critical line, whereas in the internal-gravity wave case it is. (The systematic reduction of the group velocity as the internal-gravity wave approaches a critical line may be recognized as an example of a more general ‘wave-capture’ phenomenon that occurs in a much wider class of background flows, where the deformation by the background flow acts to shrink the wavelength and hence reduce the group velocity.)

For the internal-gravity wave case, as for the Rossby-wave case, it is natural to consider whether the critical layer acts as an absorber, reflector, or overreflector of waves (or indeed as a transmitter, since the different form of the equation means that wave propagation can occur on either side of the critical line) and how this depends on the physical processes acting in the critical layer (e.g., the relative strengths of dissipation and nonlinearity). The arguments just given concerning group velocity, with waves stagnating as they approach the critical line, suggest absorption, at least if dissipation is dominant, since the decrease in group velocity as the critical line is approached means that there is infinite time for the dissipation to act. Indeed the wave will, in practice, dissipate before the critical line is reached (and indeed the thickness of the dissipative critical layer may be defined as the distance to the critical line at which the dissipation occurs).

This heuristic prediction for absorption based on group velocity was refined by the detailed calculations of Booker and

Bretherton in 1967, who considered eqn [5] in the case when the buoyancy frequency N and the vertical shear $U'(z)$ were both constant, so that the only parameter in the (nondissipative) problem is Ri . Their analysis did not require any particular assumption about dissipation, but implicitly assumes that dissipation is present. They showed that a fraction $\exp\left\{-2\pi\left(Ri - \frac{1}{4}\right)^{1/2}\right\}$ of an incident wave would be transmitted across the critical layer, with the remainder absorbed. Thus, according to this linear analysis, the critical layer is a perfect absorber in the limit of large Ri . However, if wave amplitudes are sufficiently large compared to dissipative processes, then nonlinear terms in the equations may become important before dissipation occurs and, again, before the critical line is reached. The distance to the critical line defines the thickness of the nonlinear critical layer. Here the situation is more complicated than in the Rossby-wave case. For example, it is not possible to argue that the velocity component in the z -direction (analogous to v' in the Rossby-wave case) is continuous across the nonlinear critical layer (and therefore independent of z within the critical layer). If governing equations for the nonlinear critical layer are derived, they are essentially the full nonlinear governing dynamical equations, with a slight simplification because the structure is very thin in the z -direction. The critical-layer dynamics is therefore a complex juxtaposition of wave propagation and nonlinearity.

Furthermore if nonlinearity is important it is also almost inevitable that there will be the potential for gravitational instability and therefore, in reality, breakdown of the flow into complex three-dimensional turbulence. For this reason, while there have been some formal asymptotic studies of nonlinear internal-gravity wave critical layers, the behavior that is most likely to be relevant in the real atmosphere is better investigated using full three-dimensional numerical simulation. Evidence from such simulations is that some nonlinear reflection effect is possible, particularly when the local Richardson number is not too large.

Implications for the Atmosphere

Dissipation and breaking of internal-gravity waves as they approach critical lines is potentially an important process in the atmosphere, since it implies the possibility of wave-induced forces. Breaking may also be caused by the decrease of density with height, which leads to a corresponding increase in wave amplitudes. However, there is little doubt that breaking at (or more strictly near) critical lines also plays a major role. For example, the mechanism for the equatorial quasi-biennial oscillation in the stratosphere requires selective filtering, breaking, and dissipation of waves (depending on their horizontal phase speed) by the background flow. Such waves

are believed to arise primarily from convection in the tropical troposphere (on a whole range of different scales). Observations confirm the expected relation between the phase speed of the waves observed at a particular height and the background flow at lower levels, through which they would have propagated.

Critical-line/critical-layer behavior is an important ingredient of gravity-wave parametrizations, which seek to represent the effects (primarily the wave-induced forces) of small-scale gravity waves in global-scale numerical models. Such parametrization is essential for useful simulation of the stratosphere and mesosphere. One very simple parametrization would be that, for a spectrum of upward-propagating gravity waves, each component of the spectrum dissipates at its critical line and therefore gives rise to a force at that location. In practice, some kind of breaking criterion is applied so that waves break before the critical line is reached. Almost all current parametrizations assume the equivalent of critical-layer absorption. If critical-layer reflection had to be taken into account then it would greatly increase the complexity of the parametrization problem.

See also: **Dynamical Meteorology:** Potential Vorticity; Rossby Waves; Wave Mean-Flow Interaction; Waves. **Gravity Waves:** Overview. **Middle Atmosphere:** Planetary Waves; Quasi-Biennial Oscillation.

Further Reading

- Andrews, D.G., Holton, J.R., Leovy, C.B., 1987. *Middle Atmosphere Dynamics*. Academic Press, San Diego, 489 pp.
- Booker, J.R., Bretherton, F.P., 1967. The critical layer for internal gravity waves in a shear flow. *Journal of Fluid Mechanics* 27, 513–539.
- Brunet, G., Haynes, P.H., 1996. Low-latitude reflection of Rossby wavetrains. *Journal of Atmospheric Sciences* 53, 482–496.
- Buehler, O., McIntyre, M.E., 2005. Wave capture and wave-vortex duality. *Journal of Fluid Mechanics* 534, 67–95.
- Dörnbrack, A., 1998. Turbulent mixing by breaking gravity waves. *Journal of Fluid Mechanics* 375, 113–141.
- Dunkerton, T.J., Montgomery, M.T., Wang, Z., 2009. Tropical cyclogenesis in a tropical critical layer: easterly waves. *Atmospheric Chemistry and Physics* 9, 5587–5646.
- Maslowe, S.A., 1986. Critical layers in shear flows. *Annual Review of Fluid Mechanics* 18, 405–432.
- McIntyre, M.E., Norton, W.A., 1990. Dissipative wave-mean interactions and the transport of vorticity or potential vorticity. *Journal of Fluid Mechanics* 212, 403–435. Corrigendum 220, 693.
- McIntyre, M.E., 2000. On global-scale atmospheric circulations. In: Batchelor, G.K., Moffatt, H.K., Worster, M.G. (Eds.), *Perspectives in Fluid Dynamics: A Collective Introduction to Current Research*. University Press, Cambridge, pp. 557–624.
- Staquet, C., Sommeria, J., 2002. Internal gravity waves: from instabilities to turbulence. *Annual Review of Fluid Mechanics* 34, 559–594.

Hamiltonian Dynamics

TG Shepherd, University of Toronto, Toronto, ON, Canada

© 2003 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 3, pp 929–938, © 2003, Elsevier Ltd.

Introduction

Hamiltonian dynamics describes the evolution of conservative physical systems. Originally developed as a generalization of Newtonian mechanics, describing gravitationally driven motion from the simple pendulum to celestial mechanics, it also applies to such diverse areas of physics as quantum mechanics, quantum field theory, statistical mechanics, electromagnetism, and optics — in short, to any physical system for which dissipation is negligible. Dynamical meteorology consists of the fundamental laws of physics, including Newton's second law. For many purposes, diabatic and viscous processes can be neglected and the equations are then conservative. (For example, in idealized modeling studies, dissipation is often only present for numerical reasons and is kept as small as possible.) In such cases dynamical meteorology obeys Hamiltonian dynamics. Even when nonconservative processes are not negligible, it often turns out that separate analysis of the conservative dynamics, which fully describes the nonlinear interactions, is essential for an understanding of the complete system, and the Hamiltonian description can play a useful role in this respect. Energy budgets and momentum transfer by waves are but two examples.

Hamiltonian dynamics is often associated with conservation of energy, but it is in fact much more than that. Hamiltonian dynamical systems possess a mathematical structure that ensures some remarkable properties. Perhaps the most important is the connection between symmetries and conservation laws known as Noether's theorem. Well-known examples are the fact that conservation of energy is linked to symmetry in time, and conservation of momentum to symmetry in space. Less well-known is the fact that material conservation of potential vorticity, so crucial to the theory of dynamical meteorology, is also connected to a symmetry by Noether's theorem, but to a symmetry that is invisible in the Eulerian formulation of the governing equations. It turns out that one can exploit the underlying Hamiltonian structure of a system through the relevant conservation laws even if the explicit form of that structure is not known, which is useful for applications. As is shown in detail below, symmetry-based conservation laws provide a general theory of available potential energy, and show why it is that Rossby waves carry negative zonal momentum, thereby explaining both the maintenance of the westerlies and the stratospheric Brewer–Dobson circulation. Such laws also provide a powerful way of deriving stability criteria.

Dynamical meteorologists use a variety of theoretical models, ranging from the fully compressible equations through the hydrostatic primitive, Boussinesq, and quasi-geostrophic equations to the barotropic equations. With such a zoo of models, it is crucial to know the extent to which theories developed for one model carry over to another.

Hamiltonian dynamics provides this unifying framework. All the models just mentioned are in fact Hamiltonian, and models can be grouped into families according to their Hamiltonian structure. In this way it becomes immediately apparent, for example, that the Charney–Stern stability theorem for baroclinic quasi-geostrophic flow is the counterpart to Rayleigh's inflection-point theorem for barotropic flow, and that an analogous stability theorem will exist for any balanced model having a similar Hamiltonian structure, no matter what the definition is of the potential vorticity. Thus, it is precisely through its abstract character that Hamiltonian dynamics has many powerful applications in theoretical dynamical meteorology.

Canonical and Noncanonical Dynamics

In classical mechanics, canonical Hamiltonian dynamical systems are those described by Hamilton's equation (eqns [1]).

$$\frac{dq_i}{dt} = \frac{\partial \mathcal{H}}{\partial p_i}, \quad \frac{dp_i}{dt} = -\frac{\partial \mathcal{H}}{\partial q_i} \quad (i = 1, \dots, N) \quad [1]$$

$\mathcal{H}(\mathbf{q}, \mathbf{p})$ is the Hamiltonian function, $\mathbf{q} \equiv (q_1, \dots, q_N)$ are the generalized coordinates, and $\mathbf{p} \equiv (p_1, \dots, p_N)$ the generalized momenta. For so-called natural systems with $\mathcal{H} = (|\mathbf{p}|^2/2m) + U(\mathbf{q})$, where m is the mass and U the potential energy, eqns [1] immediately lead to eqn [2], which is Newton's second law for a conservative system.

$$m \frac{d^2 q_i}{dt^2} = -\frac{\partial U}{\partial q_i} \quad (i = 1, \dots, N) \quad [2]$$

Conservation of energy follows directly from eqns [1], for any $\mathcal{H}$, by the chain rule (repeated indices are summed):

$$\frac{d\mathcal{H}}{dt} = \frac{\partial \mathcal{H}}{\partial q_i} \frac{dq_i}{dt} + \frac{\partial \mathcal{H}}{\partial p_i} \frac{dp_i}{dt} = \frac{\partial \mathcal{H}}{\partial q_i} \frac{\partial \mathcal{H}}{\partial p_i} - \frac{\partial \mathcal{H}}{\partial p_i} \frac{\partial \mathcal{H}}{\partial q_i} = 0 \quad [3]$$

Symplectic Formulation

The theory of canonical transformations suggests that there is nothing special about the qs and ps , and Hamilton's equations [1] can be written in the so-called symplectic form, eqn [4].

$$\frac{du_i}{dt} = J_{ij} \frac{\partial \mathcal{H}}{\partial u_j} \quad (i = 1, \dots, 2N) \quad [4]$$

In eqn [4], $\mathbf{u} = (q_1, \dots, q_N, p_1, \dots, p_N)$ and J is given by eqn [5], where I is the $N \times N$ identity matrix.

$$J = \begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix} \quad [5]$$

J has certain mathematical properties, including skew-symmetry. More generally, one can take those properties to be

the definition of Hamiltonian structure, with J not necessarily of the form of eqn [5]. The skew-symmetry of J nevertheless guarantees energy conservation (eqn [6]).

$$\frac{d\mathcal{H}}{dt} = \frac{\partial \mathcal{H}}{\partial u_i} \frac{du_i}{dt} = \frac{\partial \mathcal{H}}{\partial u_i} J_{ij} \frac{\partial \mathcal{H}}{\partial u_j} = 0 \quad [6]$$

There is an important distinction between systems with a nonsingular (or invertible) J , which can always be transformed into the canonical form of eqn [5], and those with a singular (or noninvertible) J . The latter, known as noncanonical systems, possess a special class of invariant functions known as Casimir invariants. These are the solutions of eqn [7] (for canonical systems the solutions are just constants).

$$J_{ij} \frac{\partial \mathcal{C}}{\partial u_j} = 0 \quad (i = 1, \dots, 2N) \quad [7]$$

That they are necessarily conserved in time then follows from the skew-symmetry of J (eqn [8]).

$$\frac{d\mathcal{C}}{dt} = \frac{\partial \mathcal{C}}{\partial u_i} \frac{du_i}{dt} = \frac{\partial \mathcal{C}}{\partial u_i} J_{ij} \frac{\partial \mathcal{H}}{\partial u_j} = -\frac{\partial \mathcal{H}}{\partial u_j} J_{ij} \frac{\partial \mathcal{C}}{\partial u_i} = 0 \quad [8]$$

The best-known example of a noncanonical Hamiltonian system is Euler's equations for rigid-body dynamics. Having an odd number of evolution equations (three in this case), the system is necessarily noncanonical because any skew-symmetric matrix of odd dimension must be singular. There is one Casimir invariant for Euler's equations, the total angular momentum.

Noether's Theorem

For a canonical system, if a particular generalized coordinate q_j does not appear in the Hamiltonian, then the Hamiltonian is invariant under changes in that coordinate; in other words, there is a coordinate symmetry. Translational and rotational symmetries are common examples. Hamilton's equations [1] then immediately imply that the corresponding generalized momentum is conserved: $dp_j/dt = 0$.

This connection between symmetries and conservation laws has a more general and far more powerful form. Given a function $\mathcal{F}(\mathbf{u})$, define $\delta_{\mathcal{F}} u_i = \varepsilon J_{ij}(\partial \mathcal{F} / \partial u_j)$, where ε is an infinitesimal parameter; $\delta_{\mathcal{F}} \mathbf{u}$ is called the infinitesimal variation in $\mathbf{u}$ generated by $\mathcal{F}$. (In the canonical case, $\delta_{\mathcal{F}} \mathbf{u}$ is an infinitesimal canonical transformation.) It then follows that the infinitesimal variation in $\mathcal{H}$ generated by $\mathcal{F}$ is given by eqn [9].

$$\delta_{\mathcal{F}} \mathcal{H} = \frac{\partial \mathcal{H}}{\partial u_i} \delta_{\mathcal{F}} u_i = \varepsilon \frac{\partial \mathcal{H}}{\partial u_i} J_{ij} \frac{\partial \mathcal{F}}{\partial u_j} \quad [9]$$

On the other hand, the time evolution of $\mathcal{F}$ is given by eqn [10].

$$\frac{d\mathcal{F}}{dt} = \frac{\partial \mathcal{F}}{\partial u_i} \frac{du_i}{dt} = \frac{\partial \mathcal{F}}{\partial u_i} J_{ij} \frac{\partial \mathcal{H}}{\partial u_j} \quad [10]$$

Using the skew-symmetry of J , eqns [9] and [10] then imply that $\delta_{\mathcal{F}} \mathcal{H} = 0$ if and only if $d\mathcal{F}/dt = 0$. This connects symmetries and conservation laws: the Hamiltonian is invariant under the variation generated by $\mathcal{F}$ (i.e., that variation represents a symmetry of the Hamiltonian) if and only if $\mathcal{F}$ is a conserved quantity. This result, known as Noether's theorem, is one of the central results of Hamiltonian dynamics

and underpins many of its applications to dynamical meteorology.

Casimir invariants are special because $\delta_{\mathcal{C}} \mathbf{u} = 0$. This suggests that they correspond to invisible symmetries. For example, in rigid-body dynamics the total angular momentum is a conserved quantity in any description of the motion. In the original canonical description it corresponds to the rotational symmetry of the dynamics, but in Euler's equations, where angles have been eliminated, it enters as a Casimir because the underlying physical symmetry is no longer explicit.

Barotropic Dynamics

In what sense are the models of dynamical meteorology Hamiltonian? Consider what is probably the simplest such model, the barotropic vorticity equation (eqn [11]), which describes two-dimensional, nondivergent flow.

$$\frac{\partial \omega}{\partial t} = -\mathbf{v} \cdot \nabla \omega = -\partial(\psi, \omega) \quad [11]$$

Here $\omega(x, y, t) = \hat{\mathbf{z}} \cdot (\nabla \times \mathbf{v}) = \nabla^2 \psi$ is the vorticity, $\hat{\mathbf{z}}$ is the unit vector in the vertical direction, $\mathbf{v}(x, y, t) = \hat{\mathbf{z}} \times \nabla \psi$ is the horizontal velocity, $\psi(x, y, t)$ is the streamfunction, and $\partial(f, g) \equiv f_x g_y - f_y g_x$ is the two-dimensional Jacobian. The candidate Hamiltonian is the conserved energy of this system, which is just the kinetic energy. The obvious dynamical variable is the vorticity. In order to cast eqn [11] in the form of eqn [4], we need to regard every point (x, y) in space as indexing a degree of freedom analogous to the index i ; the sum over i then becomes an integral over space, functions become functionals, and partial derivatives become functional or variational derivatives. Thus we write eqn [12].

$$\begin{aligned} \delta \mathcal{H} &= \delta \iint \frac{1}{2} |\nabla \psi|^2 dx dy \\ &= \iint \nabla \psi \cdot \delta \nabla \psi dx dy \\ &= \iint \{ \nabla \cdot (\psi \delta \nabla \psi) - \psi \delta \omega \} dx dy \end{aligned} \quad [12]$$

Assuming for now that the boundary terms vanish, we identify the variational derivative as $\delta \mathcal{H} / \delta \omega = -\psi$. The need to integrate by parts reflects the fact that the effect of a vorticity perturbation on the kinetic energy density is nonlocal; thus, partial derivatives at fixed points in space make no sense and variational derivatives are essential. Equation [11] can now be cast in Hamiltonian form as eqn [13].

$$\frac{\partial \omega}{\partial t} = J \frac{\delta \mathcal{H}}{\delta \omega} \quad \text{where } J = -\partial(\omega, \cdot) \quad [13]$$

Note that J is now a differential operator rather than a matrix. It is evidently skew-symmetric: $\iint fg dx dy = -\iint gf dx dy$ (under suitable boundary conditions) for arbitrary functions f, g .

Conservation Laws

The form of J in eqn [13] is clearly singular: any function of ω inserted in the argument gives zero. These then represent

Casimir invariants of the system: functionals of the form [14], where $C(\cdot)$ is an arbitrary differentiable function, evidently satisfy $J(\delta \mathcal{C} / \delta \omega) = 0$.

$$\mathcal{C} = \iint C(\omega) dx dy \quad \text{with} \quad \frac{\delta \mathcal{C}}{\delta \omega} = C'(\omega) \quad [14]$$

The fact that such functionals are conserved in time corresponds to the material conservation of vorticity expressed by eqn [11].

To identify the momentum invariants, we need to apply Noether's theorem to the various spatial symmetries. Suppose that the domain is unbounded, with decay conditions at infinity, so that there is symmetry in all directions. The variation in ω corresponding to a translation by δx in the coordinate x is given by $\delta \omega = -(\partial \omega / \partial x) \delta x$. Setting $\varepsilon = \delta x$, we then need to solve for the momentum invariant $\mathcal{M}$ according to eqn [15].

$$-\varepsilon \frac{\partial \omega}{\partial x} = \delta_{\mathcal{M}} \omega = \varepsilon J \frac{\delta \mathcal{M}}{\delta \omega} = -\varepsilon \partial \left(\omega, \frac{\delta \mathcal{M}}{\delta \omega} \right) \quad [15]$$

To within the addition of a Casimir, the solution of eqn [15] is given by $\delta_{\mathcal{M}} \omega = \gamma$. Hence we may choose $\mathcal{M}$ as in eqn [16], where $\mathbf{v} = (u, v)$.

$$\begin{aligned} \mathcal{M} &= \iint \gamma \omega dx dy = \iint \gamma \left(\frac{\partial v}{\partial x} - \frac{\partial u}{\partial y} \right) dx dy \\ &= \iint u dx dy \end{aligned} \quad [16]$$

The first, elementary form of $\mathcal{M}$ given by eqn [16] is known as Kelvin's impulse. It represents the γ 'center-of-mass' of the vorticity distribution, and is in many ways the preferable form because it is local in ω . The final form, however, shows that the invariant $\mathcal{M}$ corresponding to symmetry in x is ultimately just the x -momentum, as expected.

The same argument applied to translation in the coordinate y yields eqn [17].

$$\mathcal{M} = - \iint x \omega dx dy = \iint v dx dy \quad [17]$$

Finally, rotational symmetry leads to eqn [18], where $\mathbf{r} \equiv (x, y)$ and $r = |\mathbf{r}|$, which is the angular momentum about the origin.

$$\mathcal{M} = - \iint \frac{1}{2} r^2 \omega dx dy = \iint \hat{\mathbf{z}} \cdot (\mathbf{r} \times \mathbf{v}) dx dy \quad [18]$$

The discussion has so far neglected any contribution from boundary terms. They are easily included. In the presence of rigid lateral boundaries, for a complete mathematical specification of the problem, eqn [11] must be supplemented with the conditions [19] on each connected portion of the boundary.

$$\mathbf{v} \cdot \hat{\mathbf{n}} = 0, \quad \frac{d}{dt} \oint \mathbf{v} \cdot d\mathbf{s} = 0 \quad [19]$$

Here $\hat{\mathbf{n}}$ is the outward-pointing normal, and $\mathbf{s}$ is the vector arc length along the boundary. The second of eqns [19] represents conservation of circulation, which follows from the underlying momentum equations but must be included as a separate condition in the vorticity formulation of the dynamics. Although the circulation integrals along each

connected portion of the boundary are constants in time, they are independent dynamical variables and are needed to determine $\mathbf{v}$ from ω . The Hamiltonian formulation of eqn [13] may easily be extended to include the circulation integrals in addition to ω as dynamical variables. The Casimir invariants then include functions of these circulation integrals. With regard to the momentum invariants, of course, the rigid boundaries must respect the same symmetries; a zonal channel flow with walls at constant y breaks the translational symmetry in y and the rotational symmetry, leaving only the zonal impulse of eqn [16] as an invariant. The final equality of eqn [16] is then no longer strictly true, but the impulse and momentum differ only by terms involving the circulations along the channel walls, which are Casimirs. Since symmetry-based invariants are only defined to within a Casimir in any case, the impulse and momentum are essentially equivalent.

A simplified model of barotropic dynamics is the point-vortex model, where the vorticity is concentrated in Dirac delta functions. The point-vortex model has been used to study two-dimensional turbulence and certain kinds of atmospheric flow structures. It also turns out to be Hamiltonian, and is in fact a canonical system: the Casimirs are built into the model as parameters through the choice of the point-vortex strengths.

Other Balanced Models

The barotropic vorticity equation has a mathematical structure that is analogous to that of many models of balanced, or potential-vorticity-driven, flow (see **Dynamical Meteorology: Balanced Flow**) and the results derived above extend in an obvious way to such systems. Inclusion of the beta effect means simply a change from ω to the potential vorticity $q = \omega + \beta y$. Since $\delta y = 0$ (recalling that the coordinate y is like an index), $\delta q = \delta \omega$ and eqns [11], [12], [13] and [14] go through unchanged with q in place of ω . However the beta effect breaks translational symmetry in y and rotational symmetry, leaving only the translational symmetry in x represented by the zonal impulse invariant of eqn [16]. Strictly speaking the latter should be written with q in place of ω , but the integrals differ by a constant and so represent the same invariant. Inclusion of topography is no more difficult; one simply includes an additional topographic term $h(x, y)$ in the definition of q . This will generally break all spatial symmetries, leaving only the energy $\mathcal{H}$ and Casimirs $\mathcal{C}$ as invariants. This illustrates a general and important point, namely, that symmetry-based invariants are fragile: a slight change in the conditions of the problem destroys their conservation properties. In contrast, the energy and the Casimirs are robust invariants (robust within the conservative context, of course) that survive such perturbations.

Stratification is most easily introduced in the context of the quasi-geostrophic (QG) model (see **Dynamical Meteorology: Quasigeostrophic Theory**). Layered QG models are completely trivial extensions of the barotropic system: their evolution is determined by the potential vorticity $q_i(x, y, t)$ in each layer i , governed by eqn [11] with q_i in place of ω , together with conservation of circulation along any rigid lateral boundaries that may be present. These are then the dynamical variables. The energy now includes available potential as well as kinetic energy, but, apart from some geometric factors

representing the layer depths, one still recovers $\delta\mathcal{H}/\delta q_i = -\psi_i$ in each layer as well as eqn [13] with q_i in place of ω . The various invariants follow in the obvious way with the spatial integrals summed over the different layers. The same considerations, incidentally, apply to layered non-QG ‘intermediate’ models that still have the form of eqn [11] – namely, nondivergent horizontal advection of the potential vorticity q_i within each layer, with the flow in each layer driven by the potential vorticity in all layers (as described by the particular definition of q_i).

With continuous stratification and with upper and lower boundaries (at $z = 1$ and $z = 0$, say), there is an additional effect. It is well known that the temperature distribution along the upper and lower boundaries is equivalent to potential vorticity (see **Dynamical Meteorology: Baroclinic Instability**), and independent evolution equations for these temperature distributions are required to fully specify the continuously stratified QG system, in addition to the equation for the interior potential vorticity (the latter being eqn [11], with q in place of ω , applied at every value of z ; thus, the advection of q remains purely horizontal). The Eady model is an extreme case where the interior potential vorticity is uniform and the flow is driven entirely by the temperature distributions on the upper and lower boundaries; the dynamical structures driven from each boundary are known as Eady edge waves. Since these temperature distributions also evolve according to eqn [11], with the QG temperature ψ_z in place of ω , it is not surprising that the same kind of Hamiltonian structure also applies to this model. The energy is given by eqn [20].

$$\mathcal{H} = \iiint \frac{\rho_s}{2} \left\{ |\nabla\psi|^2 + \frac{1}{S} \psi_z^2 \right\} dx dy dz \quad [20]$$

In eqn [20], the reference-state density $\rho_s(z)$ and stratification function $S(z) = N^2/f^2$ are both prescribed, with $N(z)$ the buoyancy frequency and f the Coriolis parameter, and where ∇ is still just the horizontal gradient operator. With the potential vorticity given by eqn [21], where f and β are constants, eqn [22] follows.

$$q(x, y, z, t) = \psi_{xx} + \psi_{yy} + \frac{1}{\rho_s} \left(\frac{\rho_s}{S} \psi_z \right)_z + f + \beta\gamma \quad [21]$$

$$\delta\mathcal{H} = \left[\iint \frac{\rho_s}{S} \psi \delta\psi_z dx dy \right]_{z=0}^{z=1} + \iiint \{ \nabla \cdot (\rho_s \psi \delta \nabla \psi) - \rho_s \psi \delta q \} dx dy dz \quad [22]$$

This is like eqn [12], but with an additional term involving the temperature variations $\delta\psi_z$ at the upper and lower boundaries. Including these as independent dynamical variables, in addition to q (and possibly also circulation terms), the governing equations can be cast in the symplectic form of eqn [13]. The Casimirs now involve integrals of arbitrary functions of the temperature on the upper and lower boundaries, in addition to integrals of arbitrary functions of potential vorticity in the interior (eqn [23]).

$$\mathcal{C} = \iiint C(q; z) dx dy dz + \iint C_0(\psi_z) dx dy \Big|_{z=0} + \iint C_1(\psi_z) dx dy \Big|_{z=1} \quad [23]$$

The momentum invariants similarly extend in obvious ways: for example, the zonal impulse invariant is given by eqn [24].

$$\mathcal{M} = \iiint \rho_s \gamma q dx dy dz + \iint \frac{\rho_s}{S} \gamma \psi_z dx dy \Big|_{z=0} - \iint \frac{\rho_s}{S} \gamma \psi_z dx dy \Big|_{z=1} \quad [24]$$

The semi-geostrophic (SG) model (see **Synoptic Meteorology: Frontogenesis**) is widely used in mesoscale dynamics because of its ability to represent realistic frontal structures. It turns out that the SG model can also be cast in the form of eqn [11], and hence in the symplectic form of eqn [13], provided the equations are written in isentropic–geostrophic coordinates. However, in these coordinates rigid boundaries appear to move in time. The SG equations, in contrast to the QG equations, make no geometrical distinction between horizontal and vertical boundaries – this is why they are also useful for the study of coastal dynamics in physical oceanography – and the same kind of independent dynamical degrees of freedom encountered in the QG system on upper and lower boundaries also appear on lateral boundaries. In the special case of channel walls, these degrees of freedom correspond to coastal Kelvin waves and are analogous in some respects to the Eady edge waves represented by both the QG and SG systems. They must be taken into account in the variational calculations, and enter into many of the resulting expressions.

Unbalanced Models

Balanced models are controlled by the advection of potential vorticity (perhaps augmented by the advection of isentropic surfaces on rigid boundaries), so for such models it is natural to seek a Hamiltonian description analogous to eqn [13]. However, models that include a representation of gravity waves or other high-frequency oscillations, called unbalanced models, do not fit within this framework. They necessarily have additional degrees of freedom. For such models, a description in terms of the velocity field is a more natural way to reflect the Hamiltonian structure. For example, the rotating shallow-water equations [25] with $\mathbf{v}(x, y, t) = (u, v)$ the horizontal velocity, $h(x, y, t)$ the fluid depth, g the gravitational acceleration, and with constant f , conserve the energy (eqn [26]).

$$\frac{\partial \mathbf{v}}{\partial t} + (f \hat{\mathbf{z}} + \nabla \times \mathbf{v}) \times \mathbf{v} + \nabla \left(\frac{1}{2} |\mathbf{v}|^2 \right) = -g \nabla h, \quad [25]$$

$$\frac{\partial h}{\partial t} + \nabla \cdot (h \mathbf{v}) = 0$$

$$\mathcal{H} = \iint \frac{1}{2} \{ h |\mathbf{v}|^2 + g h^2 \} dx dy \quad [26]$$

The dynamical variables are $\mathbf{v}$ and h , for which eqns [27] hold.

$$\frac{\delta \mathcal{H}}{\delta \mathbf{v}} = h \mathbf{v}, \quad \frac{\delta \mathcal{H}}{\delta h} = \frac{1}{2} |\mathbf{v}|^2 + g h \quad [27]$$

Note that no integration by parts is necessary in this case; this is characteristic of velocity-based representations of the dynamics. It can easily be verified that eqns [25] may be cast in the symplectic form $(\partial \mathbf{u} / \partial t) = J(\delta \mathcal{H} / \delta \mathbf{u})$ with $\mathbf{u} = (u, v, h)$ with J

given by eqn [28], where $q = (f + \hat{\mathbf{z}} \cdot \nabla \times \mathbf{v})/h$ is the potential vorticity of the shallow-water system.

$$J = \begin{pmatrix} 0 & q & -\partial_x \\ -q & 0 & -\partial_y \\ -\partial_x & -\partial_y & 0 \end{pmatrix} \quad [28]$$

The matrix (28) is evidently skew-symmetric; the signs on the derivative terms are indeed correct, since first-order differential operators are themselves skew-symmetric, as with the J in eqn [13]. The zonal (absolute) momentum invariant is given as expected by eqn [29], for which it is easy to verify that $J(\delta \mathcal{M}/\delta \mathbf{u}) = -\partial \mathbf{u}/\partial x$ in line with Noether's theorem, and the other momentum invariants follow similarly.

$$\mathcal{M} = \iint h(u - fy) \, dx \, dy \quad [29]$$

The Casimirs are given by eqn [30] for arbitrary functions $C(\cdot)$.

$$\mathcal{C} = \iint hC(q) \, dx \, dy \quad [30]$$

Thus, potential vorticity still plays a crucial role in the Hamiltonian description of the dynamics. Special cases of Casimirs are total mass ($C = 1$) and total circulation ($C = q$).

Stratification is easily incorporated. The hydrostatic primitive equations can be cast in Hamiltonian form isomorphic to that of eqn [28] when expressed in isentropic coordinates. Even the fully compressible stratified Euler equations, which form the most general system imaginable for (dry) dynamical meteorology, can be cast in an analogous form, although there are now additional dynamical variables associated with compressibility. The Casimirs are in this case given by eqn [31], where $\rho(x, y, z, t)$ is the density, $\theta(x, y, z, t)$ is the potential temperature, and $q = [(f\hat{\mathbf{z}} + \nabla \times \mathbf{v}) \cdot \nabla \theta]/\rho$ is the Ertel potential vorticity, with $\mathbf{v}$ and ∇ now acting in all three space dimensions.

$$\mathcal{C} = \iiint \rho C(q, \theta) \, dx \, dy \, dz \quad [31]$$

The invariance of the Casimirs is of course evident directly from the dynamical equations (eqn [32]) and reflects the material invariance of q and θ .

$$\begin{aligned} \frac{\partial q}{\partial t} + \mathbf{v} \cdot \nabla q &= 0, & \frac{\partial \theta}{\partial t} + \mathbf{v} \cdot \nabla \theta &= 0, \\ \frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \mathbf{v}) &= 0 \end{aligned} \quad [32]$$

The fully compressible stratified Euler equations are, in fact, a straightforward expression of Newton's second law, without constraints such as hydrostatic balance, provided they are expressed in Lagrangian coordinates (see **Dynamical Meteorology: Lagrangian Dynamics**). In Lagrangian coordinates, the dynamical variables are the positions and momenta of fluid elements, which are natural canonical variables. The thermodynamic fields can be expressed in terms of these variables: ρ can be written in terms of the Jacobian of particle positions (which describes the compression of the fluid), while θ can just be chosen as one of the Lagrangian coordinates. In this way, the fully compressible stratified Euler equations represent a canonical Hamiltonian system. But there are six

dynamical variables in the Lagrangian description, compared with only five in the Eulerian description; in transforming to Eulerian coordinates, a reduction of the phase space takes place. This is where the potential vorticity comes in. In Lagrangian coordinates, the potential vorticity is still materially conserved; but what symmetry does it correspond to? The answer is a particle-relabeling symmetry: if one rearranges fluid elements while preserving the same Eulerian fields, then the dynamics is unchanged. There is just enough freedom to do this, because there is one more Lagrangian than Eulerian variable. Upon reduction to the Eulerian description, this additional degree of freedom disappears, and the particle-relabeling symmetry becomes invisible. That is why potential vorticity conservation then appears in the form of a Casimir invariant.

Disturbance Invariants

Probably the most powerful application of Hamiltonian dynamics to dynamical meteorology arises in the context of studying the properties of disturbances to basic states. In fluid dynamics, the question of how to define the energy of a wave has often been a point of confusion if not contention. For example, in the case of a basic flow, if the wave energy is defined as the energy in the frame of reference moving with the basic flow, then it is positive definite but not conserved. On the other hand, if it is defined as the difference energy relative to the basic-flow energy, then it is conserved but not positive definite. One would like both properties in order to define normal modes, spectra, etc. Another problem, at first sight unrelated, arises with momentum. The momentum of a wave would appear to be zero (the average of a sinusoid is zero), yet waves can certainly transfer momentum; this is what drives the quasi-biennial oscillation in the tropical stratosphere, for example (see **Middle Atmosphere: Quasi-Biennial Oscillation**). How is one to describe this wave momentum?

In canonical Hamiltonian mechanics, the disturbance energy about an equilibrium is always quadratic; from this one assesses stability and defines normal modes. There is no ambiguity. So why are things not equally clear for fluid dynamics? The answer lies in the noncanonical Hamiltonian structure of virtually every fluid dynamical system in the Eulerian representation. If $\mathbf{u} = \mathbf{U}$ is a steady solution of a Hamiltonian system, then eqn [33] holds.

$$J \frac{\delta \mathcal{H}}{\delta \mathbf{u}} \bigg|_{\mathbf{u} = \mathbf{U}} = 0 \quad [33]$$

For a canonical system, the invertibility of J then implies that $\delta \mathcal{H}/\delta \mathbf{u} = 0$ at $\mathbf{u} = \mathbf{U}$. This means that $\mathbf{U}$ is a conditional extremum of $\mathcal{H}$, and $\mathcal{H}[\mathbf{u}] - \mathcal{H}[\mathbf{U}]$ is quadratic in the disturbance. However, for a noncanonical system none of this follows and the disturbance energy is generally linear in the disturbance.

Pseudoenergy

Hamiltonian structure provides the solution to this quandary. Equation [33] is locally the same as the equation defining the Casimirs, which means that $\delta \mathcal{H}/\delta \mathbf{u}$ is locally parallel to $\delta \mathcal{C}/\delta \mathbf{u}$

for some $\mathcal{C}$ (a different $\mathcal{C}$ for each choice of $\mathbf{U}$). In other words, there exists a Casimir $\mathcal{C}$ such that eqn [34] holds.

$$\left. \frac{\delta \mathcal{H}}{\delta \mathbf{u}} \right|_{\mathbf{u}=\mathbf{U}} = \left. \frac{\delta \mathcal{C}}{\delta \mathbf{u}} \right|_{\mathbf{u}=\mathbf{U}} \quad [34]$$

Now, both $\mathcal{H}$ and $\mathcal{C}$ are invariants, and the combined invariant $\mathcal{H} + \mathcal{C}$ satisfies the extremal condition $\delta(\mathcal{H} + \mathcal{C}) = 0$ at $\mathbf{u} = \mathbf{U}$. We have thus constructed what we wanted, namely a disturbance quantity that is conserved and is locally quadratic in the disturbance (eqn [35]).

$$\mathcal{A} = (\mathcal{H} + \mathcal{C})[\mathbf{u}] - (\mathcal{H} + \mathcal{C})[\mathbf{U}] \quad [35]$$

This quantity is known as the pseudoenergy. Provided one has a complete set of Casimirs, eqn [34] can always be solved for a Hamiltonian system and the pseudoenergy can always be constructed according to eqn [35]. This is one of the great attractions of Hamiltonian dynamics: it provides systematic recipes in abstract terms, which can be worked out for any particular application.

A particularly illuminating application is the subject of available potential energy, highly useful in energy budget analyses. We demonstrate the method in the case of the three-dimensional stratified Boussinesq equations. The energy is given by eqn [36].

$$\mathcal{H} = \iint \left\{ \frac{1}{2} \rho_s |\mathbf{v}|^2 + \rho g z \right\} dx dy dz \quad [36]$$

Here ρ_s is the constant reference-state density, and the dynamical variables are $\mathbf{v}$ and ρ , for which eqns [37] hold.

$$\frac{\delta \mathcal{H}}{\delta \mathbf{v}} = \rho_s \mathbf{v}, \quad \frac{\delta \mathcal{H}}{\delta \rho} = g z \quad [37]$$

The term $\rho g z$ in eqn [36] is the gravitational potential energy, and is linear in the dynamical variables. Now consider disturbances to a stably stratified, resting basic state $\mathbf{v} = 0$, $\rho = \rho_0(z)$. Although the Casimirs of this system include functions of the potential vorticity, because the basic state is at rest, $\delta \mathcal{H} / \delta \mathbf{v} = 0$ at $\mathbf{v} = 0$ and this dependence is unnecessary, so we may consider Casimirs of the form of eqn [38].

$$\mathcal{C} = \iiint C(\rho) dx dy dz \quad \text{with} \quad \frac{\delta \mathcal{C}}{\delta \rho} = C'(\rho) \quad [38]$$

Equation [34] then leads to the condition $C'(\rho_0) = -gz$. This is the defining relation for the function $C(\cdot)$. Thus, one has to express gz in terms of the same argument ρ_0 . This can be done by inverting the functional dependence $\rho_0(z)$ to obtain $Z(\rho_0)$, where $Z(\rho_0(z)) = z$. This is always possible provided $\rho_0(z)$ is monotonic, which is the case for a stably stratified basic state. One then has eqn [39].

$$C(\rho) = - \int_{\rho_0}^{\rho} g Z(\tilde{\rho}) d\tilde{\rho} \quad [39]$$

From this the pseudoenergy of eqn [35] takes the form [40].

$$\mathcal{A} = \iiint \left\{ \frac{\rho_s}{2} |\mathbf{v}|^2 + (\rho - \rho_0) g z - \int_{\rho_0}^{\rho} g Z(\tilde{\rho}) d\tilde{\rho} \right\} dx dy dz \quad [40]$$

The first term in the spatial integrand is the kinetic energy and is positive definite; the last two terms can be rewritten as in eqn [41].

$$- \int_0^{\rho - \rho_0} g [Z(\rho_0 + \tilde{\rho}) - Z(\rho_0)] d\tilde{\rho} \quad [41]$$

This is self-evidently positive definite for $d\rho_0/dz < 0$ and has the small-amplitude quadratic approximation [42].

$$- \frac{g(\rho - \rho_0)^2}{2(d\rho_0/dz)} \quad [42]$$

Equation [41] is the exact, finite-amplitude expression for the available potential energy (see **General Circulation of the Atmosphere: Energy Cycle**) of disturbances to a stably stratified, resting basic state $\rho_0(z)$, while eqn [42] is its more familiar small-amplitude counterpart, widely used in the theory of internal gravity waves (see **Gravity Waves: Buoyancy and Buoyancy Waves: Theory**). Similar constructions can be performed to define the available potential energy of any stratified fluid system.

Although the small-amplitude expression of eqn [42] appears to be singular in regions where $d\rho_0/dz = 0$, the finite-amplitude expression of eqn [41] remains perfectly well-defined in such regions.

Pseudomomentum

The same kind of reasoning can be applied for disturbances to zonally symmetric (x -invariant) basic states, assuming that the underlying system possesses the same symmetry. For such states, with $\partial \mathbf{U} / \partial x = 0$, Noether's theorem implies that the zonal impulse or momentum invariant satisfies eqn [43].

$$\left. J \frac{\delta \mathcal{M}}{\delta \mathbf{u}} \right|_{\mathbf{u}=\mathbf{U}} = 0 \quad [43]$$

But just as with eqn [33], there is a Casimir $\mathcal{C}$ such that $\delta(\mathcal{M} + \mathcal{C}) = 0$ at $\mathbf{u} = \mathbf{U}$; with this $\mathcal{C}$, one may immediately construct the invariant [44], which is quadratic to leading order in the disturbance.

$$\mathcal{A} = (\mathcal{M} + \mathcal{C})[\mathbf{u}] - (\mathcal{M} + \mathcal{C})[\mathbf{U}] \quad [44]$$

This quantity is known as the pseudomomentum.

We calculate the pseudomomentum for the case of barotropic flow on the beta-plane. Suppose we are given a monotonic basic state $q_0(y)$. From eqns [14] and [16], with q in place of ω , we have eqn [45].

$$\frac{\delta \mathcal{M}}{\delta q} = y, \quad \frac{\delta \mathcal{C}}{\delta q} = C'(q) \quad [45]$$

The extremal condition $\delta(\mathcal{M} + \mathcal{C}) = 0$ at $q = q_0$ then leads to $C'(q_0) = -y$. This is now isomorphic to the construction of the available potential energy, replacing gz with y and ρ with q . If we define the function $Y(\cdot)$ by $Y(q_0(y)) = y$, then evidently eqn [46] holds.

$$\mathcal{A} = \iint \left\{ - \int_0^{q - q_0} [Y(q_0 + \tilde{q}) - Y(q_0)] d\tilde{q} \right\} dx dy \quad [46]$$

The small-amplitude approximation to the spatial integrand is given by eqn [47].

$$\frac{(q - q_0)^2}{2(dq_0/dy)} \quad [47]$$

Equations [46] and [47] are evidently negative definite for $dq_0/dy > 0$, which is the case when q_0 is dominated by βy . These rather peculiar expressions have no obvious relation to zonal momentum at first sight, but they nevertheless explain why it is that Rossby waves always exert an eastward (positive) force when they leave a source region, and a westward (negative) force when they dissipate and deposit their momentum in a sink region: they carry negative pseudomomentum.

The general nature of the derivation ensures that exactly the same expressions hold for any balanced model having the basic form of eqn [13]. If the basic state q_0 is chosen to be the zonal mean $\bar{q}$, then the zonal mean of eqn [47] becomes eqn [48], where $q' = q - \bar{q}$

$$-\frac{\overline{q'^2}}{2\bar{q}_y} \quad [48]$$

In the case of stratified QG dynamics, the negative of eqn [48] is known as the Eliassen–Palm wave activity (*see Middle Atmosphere: Zonal Mean Climatology*), which has been widely used in dynamical meteorology to assess the effect of Rossby waves on the zonal mean flow. It is such an effective diagnostic precisely because it represents negative pseudomomentum. Moreover, and importantly, its use is not restricted to waves. The exact, finite-amplitude expression of eqn [46] ensures that the concept of pseudomomentum applies to fully nonlinear, even turbulent disturbances.

The robust negative definiteness of the pseudomomentum of balanced disturbances explains a great deal about the general circulation of the atmosphere. Propagation of synoptic-scale Rossby waves away from their source region in the baroclinic storm tracks implies an eastward force in the storm track regions, accounting for the maintenance of the westerlies (*see General Circulation of the Atmosphere: Mean Characteristics*). The westward momentum deposition associated with breaking planetary-scale Rossby waves in the stratosphere drives the poleward Brewer–Dobson circulation (*see Middle Atmosphere: Zonal Mean Climatology*), which is responsible for the observed distribution of ozone and other chemical species in the stratosphere.

Stability Theorems

The pseudoenergy and pseudomomentum are, by their construction, conserved quantities that are quadratic to leading order in the disturbance quantities. In fact, their quadratic approximations are exactly conserved by the linearized dynamics. (The quadratic approximation to the pseudoenergy is the Hamiltonian of the linearized dynamics.) When either of these quantities is sign-definite for a given basic state, it follows that that basic state is stable to normal-mode instabilities. Indeed, in order to reconcile exponentially growing disturbances with conservation of pseudoenergy and

pseudomomentum, the latter quantities must vanish for such disturbances. This fact provides a useful constraint on the structure of normal-mode instabilities, as well as a powerful unifying framework between different models.

This simple framework accounts for virtually every known stability theorem in dynamical meteorology. For resting, stratified basic states in unbalanced models, with pseudoenergy like eqn [40] for the Boussinesq model, the condition of positive definite pseudoenergy is the statement of static stability (*see Dynamical Meteorology: Static Stability*). For basic flows in axisymmetric or symmetric stratified unbalanced models, the same condition is the statement of symmetric stability (*see Dynamical Meteorology: Symmetric Stability*), which reduces to Rayleigh’s centrifugal stability theorem in the special case of axisymmetric homogeneous flow (*see Dynamical Meteorology: Inertial Instability*). These stability theorems are all quite analogous to static stability. A different situation arises for balanced models. There, the pseudoenergy can take either sign depending on the basic flow. The positive-definite and negative-definite cases correspond respectively to Arnold’s first and second stability theorems. (They are analogous to the stability of a rigid body rotating about an axis of symmetry corresponding respectively to a maximum or minimum moment of inertia.) In the special case of a parallel basic flow, Arnold’s first theorem states that the flow is stable if $u_0/(dq_0/dy) < 0$, which is the Fjørtoft–Pedlosky theorem.

With regard to pseudomomentum for balanced models, eqn [46] is sign-definite whenever dq_0/dy is sign-definite. For barotropic flow with $q = \omega$, this corresponds to Rayleigh’s inflection-point theorem; on the beta-plane with $q = \omega + \beta y$, to the Rayleigh–Kuo theorem; and for stratified QG flow with q given either by its multilevel forms q_i or by eqn [21] in the continuously stratified case, to the Charney–Stern theorem. For stratified QG dynamics in the presence of a lower boundary, the second terms of eqns [23] and [24] become relevant and there is an additional contribution to the pseudomomentum involving the temperature distribution on the lower boundary; it is isomorphic to the interior eqns [46], [47] and [48], replacing q with ψ_z . Since the climatological temperature gradient along the Earth’s surface is towards the Equator, the pseudomomentum associated with surface disturbances is generally positive. In this case the Charney–Stern stability criterion is not satisfied for observed flows; on the other hand, normal-mode instabilities are generally required to involve both temperature disturbances on the lower boundary and potential-vorticity disturbances in the interior, in order to create a disturbance with zero total pseudomomentum. The Charney model of baroclinic instability (*see Dynamical Meteorology: Baroclinic Instability*) is the best-known example of this. In the presence of an upper boundary, there is a further contribution to the pseudomomentum, with opposite sign to the lower contribution in accord with eqn [24]. Thus in the Eady model of baroclinic instability, where the potential vorticity is uniform and the interior contribution to the pseudomomentum disappears, the instability can arise from the interaction of disturbances on the upper and lower boundaries that together add up to zero total pseudomomentum.

These statements all concern normal-mode stability. But what can be said about stability goes much further than this.

The existence of finite-amplitude disturbance invariants suggests the possibility of nonlinear, or Liapunov stability: namely, that small disturbances stay small for all time, where small is defined in terms of some disturbance norm. Mathematically, we say that a basic state $\mathbf{U}$ is Liapunov stable to disturbances $\mathbf{u}'$ in a given norm $\|\mathbf{u}'\|$ if for all $\varepsilon > 0$ there exists a $\delta(\varepsilon) > 0$ such that eqns [49] hold.

$$\|\mathbf{u}'(0)\| < \delta \Rightarrow \|\mathbf{u}'(t)\| < \varepsilon \quad \text{for all } t \quad [49]$$

Let us see how this applies to static stability for the Boussinesq model considered earlier. Suppose that the basic state has $d\rho_0/dz < 0$ and that furthermore the basic-state density gradients are bounded according to [50] for some constants c_1, c_2 .

$$0 < c_1 \leq -g \frac{dz}{d\rho_0} = -\frac{g}{d\rho_0/dz} \leq c_2 < \infty \quad [50]$$

Then eqn [41] for the available potential energy is bounded from above and below according to eqn [51].

$$\frac{1}{2} c_1 (\rho - \rho_0)^2 \leq [41] \leq \frac{1}{2} c_2 (\rho - \rho_0)^2 \quad [51]$$

Define the disturbance norm by eqn [52], with $c_1 \leq \lambda \leq c_2$.

$$\begin{aligned} \|(\mathbf{v}, \rho - \rho_0)\|^2 &= \iiint \frac{1}{2} \{ \rho_s |\mathbf{v}|^2 + \lambda (\rho - \rho_0)^2 \} \, dx \, dy \, dz \end{aligned} \quad [52]$$

Then using eqn [51] we obtain the chain [53] of inequalities, valid for any time t , involving the pseudoenergy $\mathcal{A}$ of eqn [40].

$$\begin{aligned} \|(\mathbf{v}, \rho - \rho_0)(t)\|^2 &\leq \frac{\lambda}{c_1} \mathcal{A}(t) = \frac{\lambda}{c_1} \mathcal{A}(0) \\ &\leq \frac{c_2}{c_1} \|(\mathbf{v}, \rho - \rho_0)(0)\|^2 \end{aligned} \quad [53]$$

With the choice $\delta = \sqrt{c_1/c_2} \varepsilon$, eqn [53] establishes Liapunov stability in the norm defined by eqn [52]. Conservation of pseudoenergy is clearly central to the proof.

The finite-amplitude stability of stably stratified flow is not too surprising; it corresponds to physical intuition, and indeed motivates the very concept of available potential energy, which has a long pedigree. What is perhaps more surprising is that exactly the same kinds of constructions can be made for all of the stability theorems mentioned above, and for virtually any model within the same family. They can also be used to obtain rigorous upper bounds on the saturation of normal-mode instabilities, by considering the initial unstable flow (plus infinitesimal disturbance) to be a finite-amplitude disturbance to a stable basic state.

See also: Dynamical Meteorology: Balanced Flow; Baroclinic Instability; Inertial Instability; Lagrangian Dynamics; Potential Vorticity; Quasigeostrophic Theory; Static Stability; Symmetric Stability. **General Circulation of the Atmosphere:** Energy Cycle. **Middle Atmosphere:** Quasi-Biennial Oscillation; Zonal Mean Climatology. **Synoptic Meteorology:** Frontogenesis. **Tropical Meteorology and Climate:** El Niño and the Southern Oscillation: Theory.

Further Reading

- Arnold, V.I., 1989. *Mathematical Methods of Classical Mechanics*, 2nd edn. Springer-Verlag, New York.
- Benjamin, T.B., 1984. Impulse, flow force and variational principles. *IMA Journal of Applied Mathematics* 32, 3–68.
- Holm, D.D., Marsden, J.E., Ratiu, T., Weinstein, A., 1985. Nonlinear stability of fluid and plasma equilibria. *Physics Reports* 123, 1–116.
- Landau, L.D., Lifshitz, E.M., 1976. *Mechanics*, 3rd edn. Pergamon Press, New York.
- Morrison, P.J., 1998. Hamiltonian description of the ideal fluid. *Reviews of Modern Physics*, 467–521.
- Salmon, R., 1988. Hamiltonian fluid mechanics. *Annual Review of Fluid Mechanics* 20, 225–256.
- Salmon, R., 1998. *Lectures on Geophysical Fluid Dynamics*. Oxford University Press, New York.
- Shepherd, T.G., 1990. Symmetries, conservation laws, and Hamiltonian structure in geophysical fluid dynamics. *Advances in Geophysics* 32, 287–338.
- Shepherd, T.G., 1993. A unified theory of available potential energy. *Atmosphere–Ocean* 31, 1–26.

Hydraulic Flow

RB Smith, Yale University, New Haven, CT, USA

© 2003 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 3, pp 966–967, © 2003, Elsevier Ltd.

The study of hydraulic flow is one branch of a broader field of fluid mechanics dealing with the dynamics of density stratified flow under the influence of a gravity field. It has a natural application to the stratified atmosphere and ocean. The field of hydraulics is distinguishable from other studies of stratified flow by its emphasis on layered flow and the use of the hydrostatic or long-wave approximation. Typically, in hydraulic flow formulations, the fluid system is composed of one or more homogeneous fluid layers, separated by sharp interfaces with density discontinuities. This formulation, together with the hydrostatic assumption, insures that the velocity is nearly uniform with height within each layer. In this way, a continuous problem is reduced to a problem with one or more discrete layers; this results in a vast reduction in the number of degrees of freedom. The possibilities for mathematical analysis, numerical computation, and physical conceptualization are greatly enhanced by the simple formulation of hydraulic theory.

Historically, the field of hydraulics arose out of and is still largely involved in, the study of natural river flow and engineering problems related to water flow in channels. Its application to atmosphere and ocean dynamics is more recent. Beginning in the 1950s, a growing number of atmospheric applications have been suggested. On large scales, CG Rossby, G Benton, and NA Phillips developed two-layer mathematical models of the midlatitude atmosphere including the Coriolis force. On smaller scales, following the pioneering work of RR Long and M Tepper, a variety of atmospheric phenomena have been treated with hydraulic models. Cool outflowing air from thunderstorms, sea breeze fronts, and the leading edges of cold fronts all behave like gravity currents. Existing cool layers beneath marine inversions and frontal layers behave hydraulically in mountainous areas, causing barrier jets, gap jets, hydraulic jumps, severe downslope winds, and wake eddies. Cold high terrain can generate layered cold air avalanches and katabatic winds.

In oceanography too, hydraulic theory has found wide application. Basin-to-basin exchange of water masses is limited by hydraulic control at sills and straits. The propagation of tidal currents and tsunamis is controlled by the long-wave speed. Turbidity currents slump into the deep ocean according to gravity current dynamics. Coastally trapped currents obey a modified set of hydraulic equations. Even large-scale wind-driven ocean currents are often modeled as two layers, defined by the thermocline, with wind stress and the Coriolis force playing dominant roles.

The theory of hydraulic flow is based on a few fundamental definitions and concepts. These are reduced gravity, the long-wave speed, Froude number, hydraulic control, conjugate states, the hydraulic jump, and gravity or density current.

Reduced gravity (g') is a measure of the effective magnitude of gravity acting on layers of different density. It is defined as the product of the acceleration of gravity ($g = 9.81 \text{ m s}^{-2}$) times the relative density difference $\Delta\rho/\rho$ between the two superposed layers, thus $g' = (\Delta\rho/\rho)g$. In the compressible atmosphere, the relative density difference is approximately the difference in potential temperature (Θ), so that $g' = (\Delta\Theta/\Theta)g$. For example, if the air above an inversion is 3° warmer than the air below, and the average potential temperature is 300 K, the effective gravity is $g' = (3/300)g = 0.0981 \text{ m s}^{-2}$. The long-wave speed for a single layer is given by $C = (g'H)^{1/2}$, where H is the depth of the layer. If a cool marine layer of air has an effective gravity of $g' = 0.1$ and a depth of 1000 m, long gravity waves will propagate along it at a speed of $C = (0.1 \times 1000)^{1/2} = 10 \text{ m s}^{-1}$.

The Froude number plays a central role in hydraulic theory. It is defined as the ratio of flow speed (u) to long-wave speed (C), i.e., $(Fr) = u/C$. It is related to the ratio of kinetic energy to potential energy in a layer but is more useful as a measure of whether waves can move upstream against the current. Flows are categorized as subcritical, critical, or supercritical according to whether the Froude number is less than, equal to, or greater than unity. In supercritical flow ($(Fr) > 1$), long waves cannot move upstream or even stand steady against the fluid flow. As the long waves are usually the fastest waves in the system, information carried by waves cannot then be felt upstream of a disturbance. The nature of a fluid response to any disturbance is highly sensitive to the Froude number.

Hydraulic control occurs whenever a layered flow is forced to transition from subcritical to supercritical flow by the narrowing of a channel or valley, the rising of a sill or mountain, or the alteration of some other geometric or external parameters (e.g., Coriolis force, coastal slope, etc.). Downstream of the control point, i.e., the point where the Froude number is unity, supercritical flow prevents information from propagating upstream. As a result, the amount of flow through the channel cannot be altered from downstream. Only the upstream conditions and control point characteristics have an influence.

Conjugate states are defined as multiple states of flow, defined by fluid speed and layer depth, with identical mass and momentum flux. They can be computed easily in the hydraulic formalism. If a flow has a conjugate state, it can, in principle, jump spontaneously to its other state, without the loss or gain of mass or momentum. Most commonly, this occurs in a hydraulic jump (i.e., an abrupt thickening of a layer), where the energy may be dissipated by turbulence, but mass and momentum are conserved. Jumps are commonplace events in rivers, and related phenomena have been identified in the atmosphere and ocean.

A so-called gravity current or density current occurs when a new fluid pushes its way over or under an existing denser or less dense fluid, under the influence of gravity. While it resembles a hydraulic jump, a density current is not a sudden thickening of a preexisting layer, but the introduction of a new fluid layer.

The literature is not clear on whether all layered formulations of stratified fluid mechanics should be classified as 'hydraulics'. When friction or Coriolis forces dominate, the term hydraulics is less often used.

See also: **Dynamical Meteorology:** Overview; Solitary Waves; Static Stability. **Mesoscale Meteorology:** Convective Storms: Overview; Density Currents; Gust Fronts; Microbursts; Overview. **Mountain Meteorology:** Downslope Winds; Katabatic Winds; Land and Sea Breezes; Lee Waves and Mountain Waves; Valley Winds.

Further Reading

- Baines, P.G., 1995. *Topographic Effects in Stratified Flows*. Cambridge University Press, Cambridge, UK.
- Hughes, R.L., 1989. The hydraulics of local separation in a coastal current with application the Kuroshio meander. *Journal of Physical Oceanography* 19, 1809–1820.
- Jackson, P.L., Steyn, D.G., 1994. Gap winds in a fjord. 2. Hydraulic analog. *Monthly Weather Review* 122, 2666–2676.
- Pratt, L.J., Lundberg, P.A., 1991. Hydraulics of rotating strait and sill flow. *Annual Review of Fluid Mechanics* 23, 81–106.
- Seitter, K.L., 1987. Numerical study of atmospheric density current motion including the effects of condensation. *Journal of the Atmospheric Sciences* 43, 3068–3076.
- Smith, R.B., 1985. On severe downslope winds. *Journal of the Atmospheric Sciences* 42, 2597–2603.
- Smith, R.B., Smith, D.F., 1995. Pseudoinviscid wake formation by mountains in shallow-water flow with a drifting vortex. *Journal of the Atmospheric Sciences* 52, 436–454.
- Yih, C.S., 1965. *Dynamics of Nonhomogeneous Fluids*. Macmillan, New York, NY.

Inertial Instability

JA Knox, University of Georgia, Athens, GA, USA

© 2003 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 3, pp 1004–1013, © 2003, Elsevier Ltd.

Introduction

Inertial instability is a fundamental, but infrequently documented, hydrodynamic instability characterized by strongly divergent anticyclonic flow. It is the geophysical analog to the centrifugal (or Taylor–Couette) instability in fluid dynamics examined by Rayleigh and Taylor nearly a century ago. Bergen School meteorologists seeking an explanation for cyclogenesis pioneered inertial instability research in the 1930s and 1940s. However, the triumph of baroclinic instability theory and balanced dynamics in the 1950s and 1960s sharply curtailed interest in inertial instability. A revival of research during the past two decades, particularly in middle-atmosphere and mesoscale dynamics, has led to a growing appreciation of the role of inertial instability in geophysical flows. Today, inertial instability arises in a wide range of subjects: the dynamics of mesoscale convection and monsoons, wave generation and breaking in the stratosphere and mesosphere, and the maintenance of jets in planetary atmospheres and equatorial oceanography. Below, we examine what inertial instability looks like physically, how it is represented mathematically, and how it is manifested geophysically. Those new to the subject may wish to begin with the latter.

Physical Description

Centrifugal instability occurs in the Taylor–Couette problem when the angular momentum of a fluid contained between two rotating cylinders (Figure 1(a)) decreases radially outward, violating Rayleigh’s stability criterion. Parcels then rearrange themselves to achieve a stable radial profile of angular momentum.

Inertial instability is the geophysical equivalent of centrifugal instability and occurs when angular momentum decreases as one moves outward from the axis of rotation of the flow. This can be visualized on the global scale (Figure 1(b)) by imagining the roles of the cylinders being played by latitude lines, with the Equator serving as the inner cylinder. The latitude at which angular momentum is nondecreasing outward from the rotation axis forms the outer cylinder.

This analogy between centrifugal and inertial instabilities is correct only if the rotation in the Taylor–Couette apparatus is equated to the total vertical rotation of the flow in the geophysical case. In meteorological terminology, the rotation in Figure 1(b) is the sum of the planetary and relative vertical vorticities. From this viewpoint, inertial instability should be expected where the relative vorticity is opposite in sign to the planetary vorticity, and at least equal in magnitude to it. The threshold for inertial instability – the latitude corresponding to the outer cylinder in Figure 1(b) – is thus the location of zero absolute vorticity. With some restrictions, this is proved mathematically in the next section.

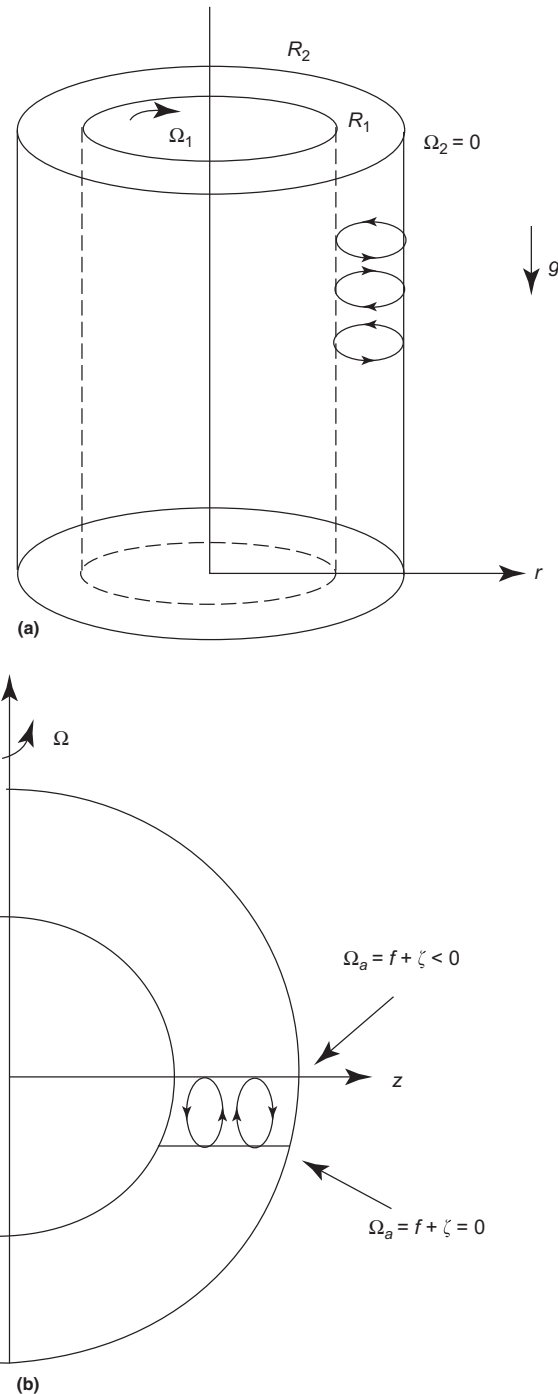


Figure 1 (a) Centrifugal instability in the Taylor–Couette experiment and (b) inertial instability near the equator. Adapted with permission from Hua, B.L., Moore, D.W., Le Gentil, S., 1997. Inertial nonlinear equilibration of equatorial flows. *Journal of Fluid Mechanics* 331: 345–371.

Inertial instability, like centrifugal instability, leads to roll-like vortical motions. The flattened, 'pancake'-like divergent circulations of inertial instability ([Figure 2](#)) attempt to reduce the anticyclonic local rotation via momentum transports so that the total rotation is the same sign throughout the domain.

An inertially unstable flow can exist on a variety of scales. As illustrated in [Figures 1 and 2](#), the flow can be as large as a zonally symmetric latitudinal ring of air around the Earth. It can also occur on scales as small as a mesoscale vortex. When viewed from the perspective of a circular vortex, this instability is triggered when a very strong outward pressure gradient force and the centrifugal force combine to overwhelm the Coriolis force and lead to the breakdown of balance.

In layman's terms, the salient point of inertial instability research is that a planet's rotation sets a local rotational

'speed limit' beyond which violations are corrected via inertial instability. This 'speed limit' can be indirectly inferred from weather charts in a number of ways: for example, the strong tendency for nonnegative absolute vorticity in the Northern Hemisphere, and also the complete absence of intense 'bull's-eye' high-pressure centers analogous to 'bomb' cyclones.

Inertial instability depends on planetary rotation and the horizontal shear and/or curvature of the horizontal wind. In this sense, it shares an affinity with barotropic instability. However, the rapid, strongly divergent motions associated with inertial instability invite analogies with another fundamental, strongly ageostrophic instability: static instability. This analogy is strengthened by the close similarity in the derivation of these two instabilities' criteria, discussed below.

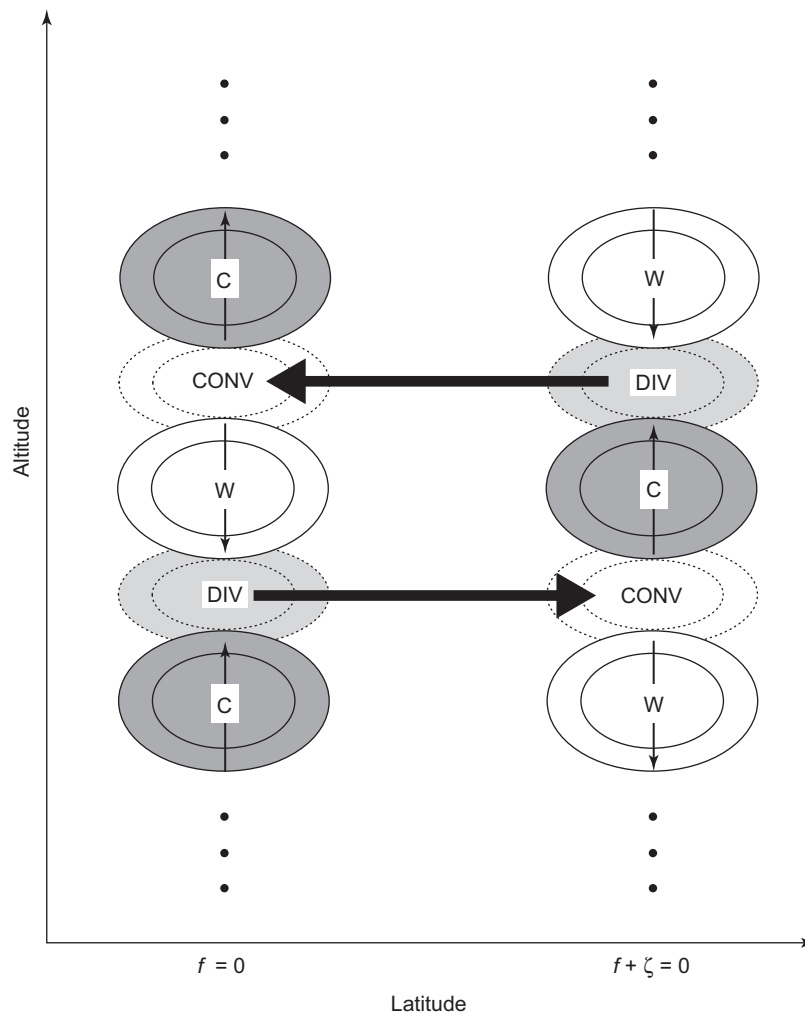


Figure 2 Schematic view of inertially unstable circulations in (y, z) . DIV and CONV refer to regions of divergence and convergence due to the horizontally divergent inertial circulations, shown in heavy bold arrows. Conservation of mass leads to the vertical motions shown in the lighter arrows, which adiabatically create the warm and cold temperature anomalies labeled W and C respectively. Adapted with permission from Dunkerton, T.J., 1981. On the inertial stability of the equatorial middle atmosphere. *J. Atmos. Sci.* 38: 2354–2364, and Hayashi, H., Shiotani, M., Gille, J.C., 1998. Vertically stacked temperature disturbances near the equatorial stratopause as seen in cryogenic limb array etalon spectrometer data. *Journal of Geophysical Research* 103: 19469–19483.

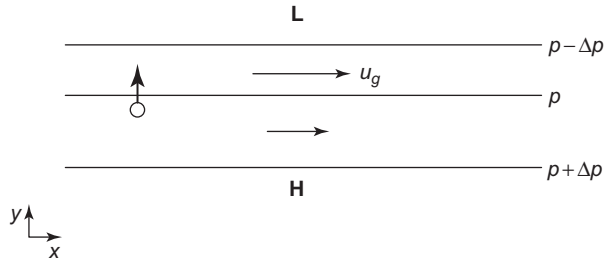


Figure 3 Schematic illustrating the physical situation posed in the mathematical derivation of the inertial instability criterion (eqn [8]), in which a parcel (small circle) in a background flow with horizontal shear is subject to a perturbation (heavy arrow).

Mathematical Criteria for Instability

Inertial instability represents a large departure from geostrophic balance. As a result, inertial instability theory cannot rely on quasi-geostrophic or nonlinear balances to facilitate a generalized theory, making it a far less tractable problem than barotropic or baroclinic instability. In fact, no fully three-dimensional theory for inertial instability exists currently. Instead, the classic derivation relies on the parcel-mode approach, an extremely simplified flow geometry, and ignores nonlinear, frictional, and diabatic effects. Even so, the result is remarkably useful when applied to observed and simulated flows.

We begin by assuming that we have a perturbed parcel embedded in a purely zonal geostrophic flow, as illustrated in **Figure 3**. The horizontal equations of motion describing the parcel are:

$$\frac{D_H u}{Dt} = f v_{ag} \quad [1]$$

$$\frac{D_H v_{ag}}{Dt} = f(u_g - u) \quad [2]$$

In these equations, the horizontal Lagrangian derivative $D_H/Dt = \partial/\partial t + u\partial/\partial x + v\partial/\partial y$ and f is the Coriolis parameter (assumed to be constant for simplicity). In eqn [2], the meridional pressure gradient term has been rewritten in terms of the geostrophic wind.

These two equations can be combined into one equation for the meridional ageostrophic wind in the following manner. Taking the Lagrangian derivative of eqn [2] yields

$$\frac{D_H^2 v_{ag}}{Dt^2} = f \left(\frac{D_H u_g}{Dt} - \frac{D_H u}{Dt} \right) \quad [3]$$

The second term in the parenthesis in eqn [3] can be replaced with the right-hand side of eqn [1], coupling the two equations of motion. The first term in the parenthesis can also be related to the meridional ageostrophic wind via the following approximation:

$$\frac{D_H u_g}{Dt} \approx v \frac{\partial u_g}{\partial y} = -v_{ag} \zeta_g \quad [4]$$

The relation in eqn [4] derives from the steady, zonally uniform mass field in **Figure 3**. It also assumes that the parcel

will not locally change the geostrophic wind but it can advect geostrophic relative vorticity meridionally; thus the analysis is non-quasi-geostrophic but omits the two-way interaction of the geostrophic adjustment problem.

Insertion of eqns [1] and [4] into eqn [3] and rearrangement leads directly to

$$\frac{D_H^2 v_{ag}}{Dt^2} + [f(f + \zeta_g)] v_{ag} = 0 \quad [5]$$

This second-order differential equation bears a close resemblance to the stability equation for static instability, with the meridional ageostrophic wind replacing the vertical displacement and $f(f + \zeta_g)$ replacing the buoyancy frequency. As in the static stability problem, we assume a wave solution and obtain the following cases:

$$f(f + \zeta_g) > 0$$

stable inertial oscillation with period

$$\frac{2\pi}{[f(f + \zeta_g)]^{1/2}} \quad [6]$$

$$f(f + \zeta_g) = 0 \quad \text{inertial neutrality} \quad [7]$$

$$f(f + \zeta_g) < 0$$

$$\text{inertial instability with e-folding time } \frac{1}{[-f(f + \zeta_g)]^{1/2}} \quad [8]$$

Note that the inertial oscillation period in eqn [6] reduces to the usual textbook form if $\zeta_g = 0$ the more general form in eqn [6] has been applied successfully to constant-pressure radiosonde trajectory periodicities. Weak inertial stability is analogous to weak static stability, in which forcing leads to a larger response than in strongly stable conditions.

The criterion in eqn [8] can be interpreted in a variety of ways. In the Northern Hemisphere, it is equivalent to negative geostrophic absolute vorticity. For statically stable conditions, eqn [8] implies negative *potential* vorticity in a geostrophic flow in the Northern Hemisphere. If eqn [8] is calculated on an isentropic surface then it is identical to the criterion for symmetric instability (see **Dynamical Meteorology: Symmetric Stability**). Interestingly, the necessary criterion for barotropic instability is the meridional derivative of eqn [8].

The e-folding times of inertial instability are, by eqn [8], dependent on the latitude and the magnitude of the anomalous absolute vorticity. Observations and modeling studies suggest e-folding times as short as a few hours in the mesoscale midlatitude troposphere and around one day in the equatorial middle atmosphere. These time scales are much longer than for static instability but are usually shorter than for barotropic instability.

Extensions beyond eqn [8] are possible in some cases. For a circular vortex, the criterion in eqn [8] is modified only slightly, with the sum of the shear and curvature vorticities replacing the geostrophic relative vorticity. Extension of the

analysis to the beta plane does not alter the instability criterion, although the condition for stability becomes necessary, no longer sufficient. On the sphere, the metric terms due to the Earth's curvature can alter the criterion, but the change is significant only for high wind speeds and/or high latitudes. The vertical equation of motion may be incorporated into the analysis via the thermal wind law; the resulting instability criterion is eqn [8], with the absolute vorticity replaced with the Ertel potential vorticity. Friction in the form of Rayleigh drag stabilizes the flow linearly, i.e., the more drag there is, the less unstable/more stable the flow is. Extension of the theory to zonally asymmetric flows has been achieved but does not yield an alternative criterion; in general eqn [8] is used locally with good results, except in very strong anticyclones.

Unfortunately, the assumptions in the analysis above preclude any insight into the vertical scale of the instability. The growth rates for inertial instability are greatest at smallest scales; it is hypothesized that eddy diffusion damps these scales out and leads to a preferred intermediate vertical scale. However, the instability itself may be the source of the turbulent eddies that in turn select the vertical scale of the instability. Thus the eddy diffusion hypothesis for vertical scale selection, which depends on externally prescribed approximations such as Rayleigh drag, is incomplete.

Observed Phenomena Related to Inertial Instability

Inertial instability is possible wherever the anticyclonic relative vorticity rivals the Coriolis parameter in magnitude. Therefore, it is not confined to any one region of the planet, although it is

likeliest where the Coriolis parameter is smallest, i.e., the tropics.

Middle Atmosphere

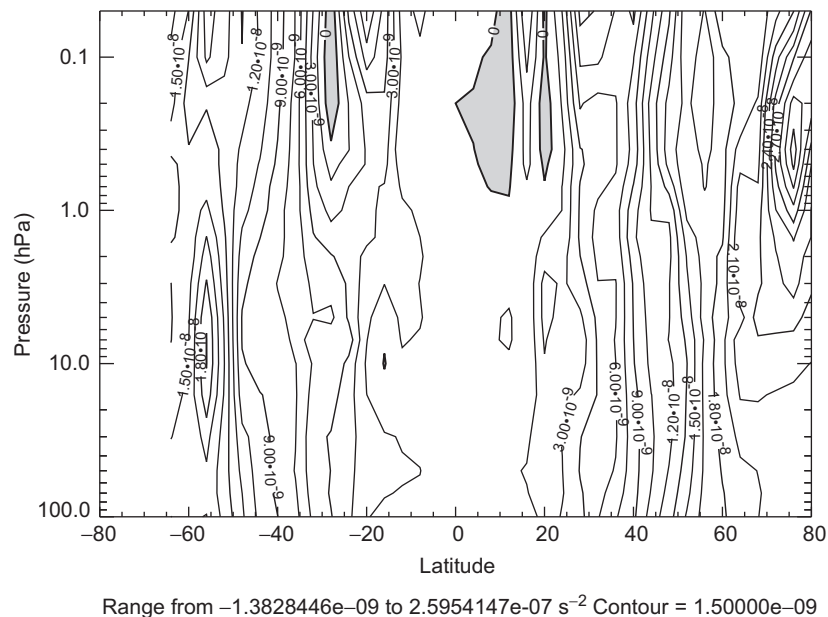
Large-scale inertial instability is observed in the equatorial lower mesosphere, often lasting about a week and occurring within a week or two of the boreal winter solstice. The instability develops in regions of negative absolute vorticity (Figure 4) on the poleward flank of the stratopause semiannual oscillation easterlies; there is also evidence for it around the summer solstice. Its hallmarks are layered, nearly stationary 'pancake structures' in the eddy temperature field (Figure 5) that have a latitudinal width of 10–20 degrees of latitude and a vertical wavelength of roughly 10 km.

Numerical simulations and observations have shown that Rossby wave breaking in the tropics acts as a trigger for the onset of equatorial inertial instability and influences its zonally asymmetric nature in winter. The divergent character of the instability, and some recent observational work, suggest a role for inertial instability in the latitudinal transport of tracers in the tropical middle atmosphere.

Recent research strongly indicates that Rossby wave-triggered inertial instability excites the two-day wave of the equatorial stratopause region. The instability may also play a role in some stratospheric sudden warmings and other high-latitude phenomena.

Upper Troposphere

Inertial instability has been sought for in jet stream analyses since the 1940s. A long-term climatology of inertial instability



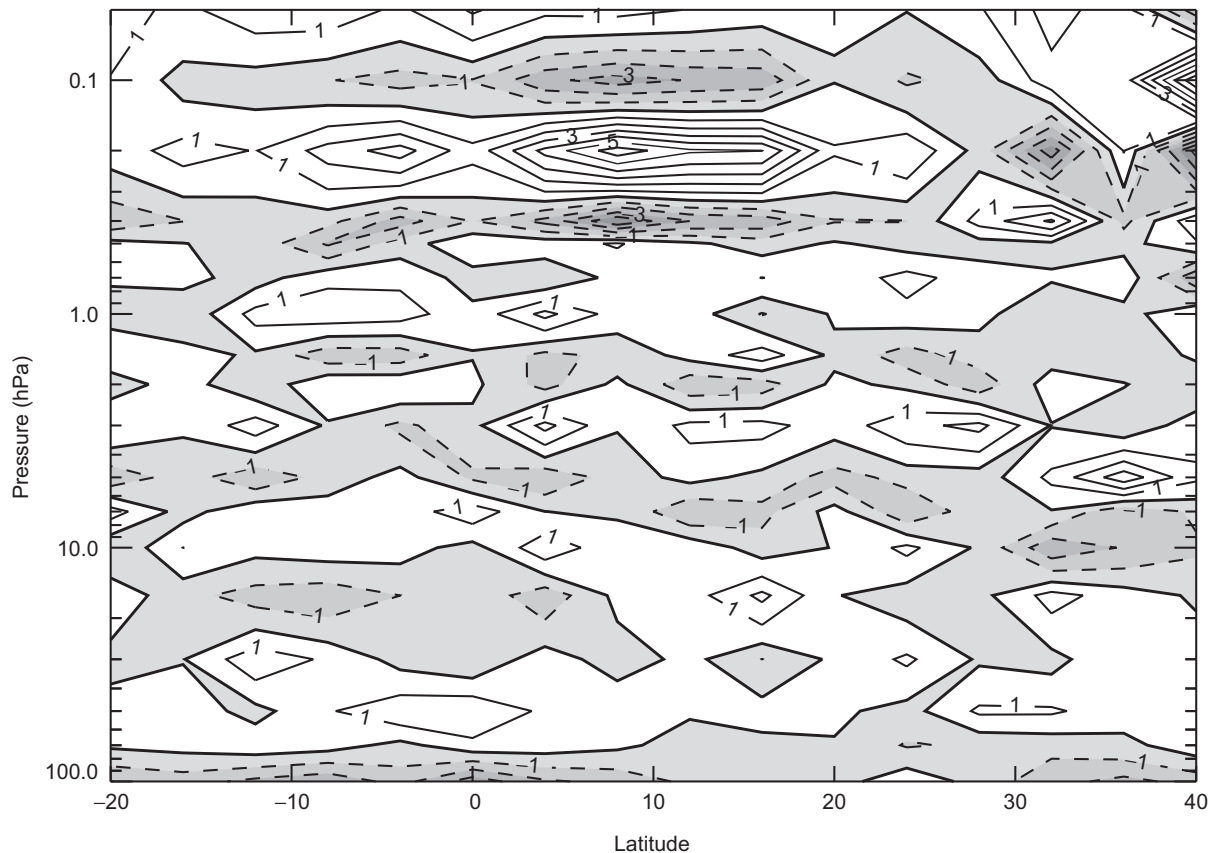


Figure 5 Vertically layered ‘pancake structures’ in the lower mesosphere in LIMS eddy temperatures (contour interval 1 K, negative regions shaded) at the international dateline on 10 December 1978, at the beginning of a week-long episode of inertial instability. Note that the temperature anomalies are strongest just south of 10° N, where the inertial instability is centered in [Figure 4](#), and that anomalies of the opposite sign exist near 30° N, as predicted by the schematic in [Figure 2](#).

based on NCEP geopotential heights ([Figure 6](#)) reveals why this has been a difficult search: the criterion for the instability is achieved in the data-rich midlatitudes only once every few years at most. (The occurrence rate is of course higher for higher-resolution data sets, but is still rare.) However, [Figure 6](#) suggests that inertial instability is a fairly common phenomenon in the subtropical upper troposphere, particularly on the equatorward flank of the East Asian jet. Recent work has shown layered disturbances in low-PV regions near the tropopause just south of Japan.

Inertial instability can enhance the outflow from mesoscale convective systems such as thunderstorms, ‘tropical plumes’, and hurricanes (the Rossby radius of deformation, an estimator of the outflow width, is closely related to eqn [8] and is infinite in the case of inertial instability). However, the evidence for its role in both hurricane outflow and severe thunderstorms is mixed. In [Figure 7](#), the upper-tropospheric flow pattern for the devastating 3 May 1999 Oklahoma City, Oklahoma, tornado outbreak is depicted. Inertially unstable regions are colocated with strong divergence aloft and severe weather at the surface in this case, but not in others. The relationship between inertial instability and convection is still not well understood.

Inertial instability may also have a connection to hazardous weather from an aviation perspective. The gravity wave

radiation expected from strongly anticyclonic regions, such as inertially unstable circulations, may lead to some otherwise unexplained instances of clear-air turbulence.

Lower-to-Middle Troposphere

Inertial instability below the jet stream level is confined to the tropics and a few rare instances in the midlatitudes. The cross-equatorial Asian monsoon circulation appears to accelerate toward the coast of India under the influence of inertial instability. The divergence – convergence couplets of inertial instability appear to determine the location of near-equatorial convection and the mean latitude of the Intertropical Convergence Zone.

Intense extratropical anticyclones may possibly exhibit inertial instability. This is the best explanation for an unforecast pre-dawn elevated convection event near the center of the mid-July 1995 Chicago heat wave high-pressure system. Elongated bands of light precipitation ([Figure 8](#)) coincided closely in time and space with a narrow region of negative potential vorticity ([Figure 9](#)). Very high-resolution model simulations of this event, shown in [Figure 10](#), indicate a checkerboard pattern of vertical motion (and thus divergence and convergence) strongly suggestive of [Figure 2](#). It appears that the

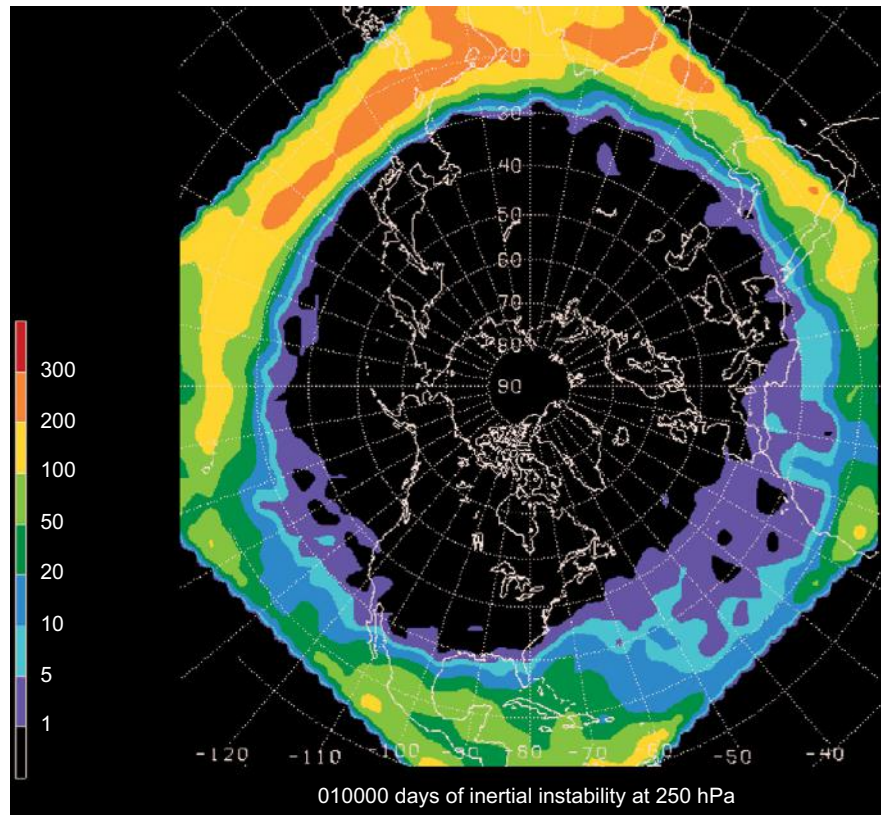


Figure 6 Climatology of inertial instability using National Centers for Environmental Prediction daily geopotential height analyses (horizontal resolution = 381km at 60° N) at 250 mb for December 1966–December 1994 (29 Decembers). Contours indicate the number of analyses in which eqn [8] was satisfied over the 899 days in the study. Figure courtesy Russ Schumacher, Colorado State University, and David Schultz, National Severe Storms Laboratory.

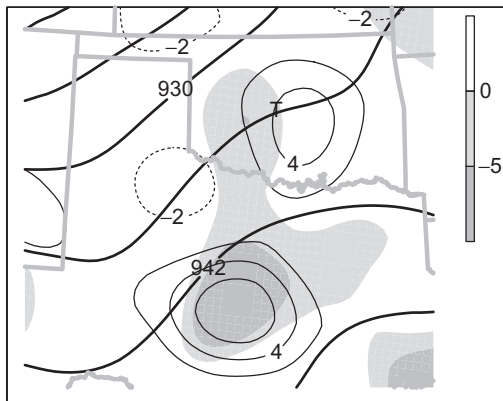


Figure 7 Eta model initial analysis of absolute geostrophic vorticity (shaded, only negative regions shown; scale is $\times 10^5 \text{ s}^{-1}$), geopotential height (heavy contours; in decameters), and horizontal divergence (light contours; scale is $\times 10^5 \text{ s}^{-1}$) at 300 mb over Texas and Oklahoma valid at 0000 UTC on 4 May 1999. The centroid of tornado reports at this time is indicated with a large T. Shaded regions, corresponding to inertial instability via eqn [8], are collocated with severe weather and with high values of divergence. The inertially unstable region over Texas was associated with 29 severe weather reports. Figure courtesy David Schultz, National Severe Storms Laboratory.

rising motions induced by the instability were substantial enough to cause condensation and elevated convection where strong subsidence would normally be expected.

Other Geophysical Fluids

Like the atmosphere, the oceans may also contain inertially unstable flows wherever the anticyclonic current is fast and the effect of the planetary rotation is weak. The layered structure of subthermocline equatorial ocean currents, reminiscent of the middle atmosphere 'pancake structures' and confined to within a degree or two of the Equator, has been linked with inertial instability. Furthermore, anticyclonic ocean eddies, even at high latitudes, can satisfy the criterion in eqn [8]. For example, the overwhelming tendency for 'spiral eddies' on the scales of a few kilometers to rotate cyclonically has been attributed to the limiting effects of inertial instability on small-scale anticyclones.

Farther afield, the atmospheres of Mars and the large gaseous planets are likely venues for inertial instability owing to their strong jet structures at low latitudes. Modeling studies suggest that the very strong shears on the equatorial flanks of the Martian jets should be inertially unstable, even in long-term means.

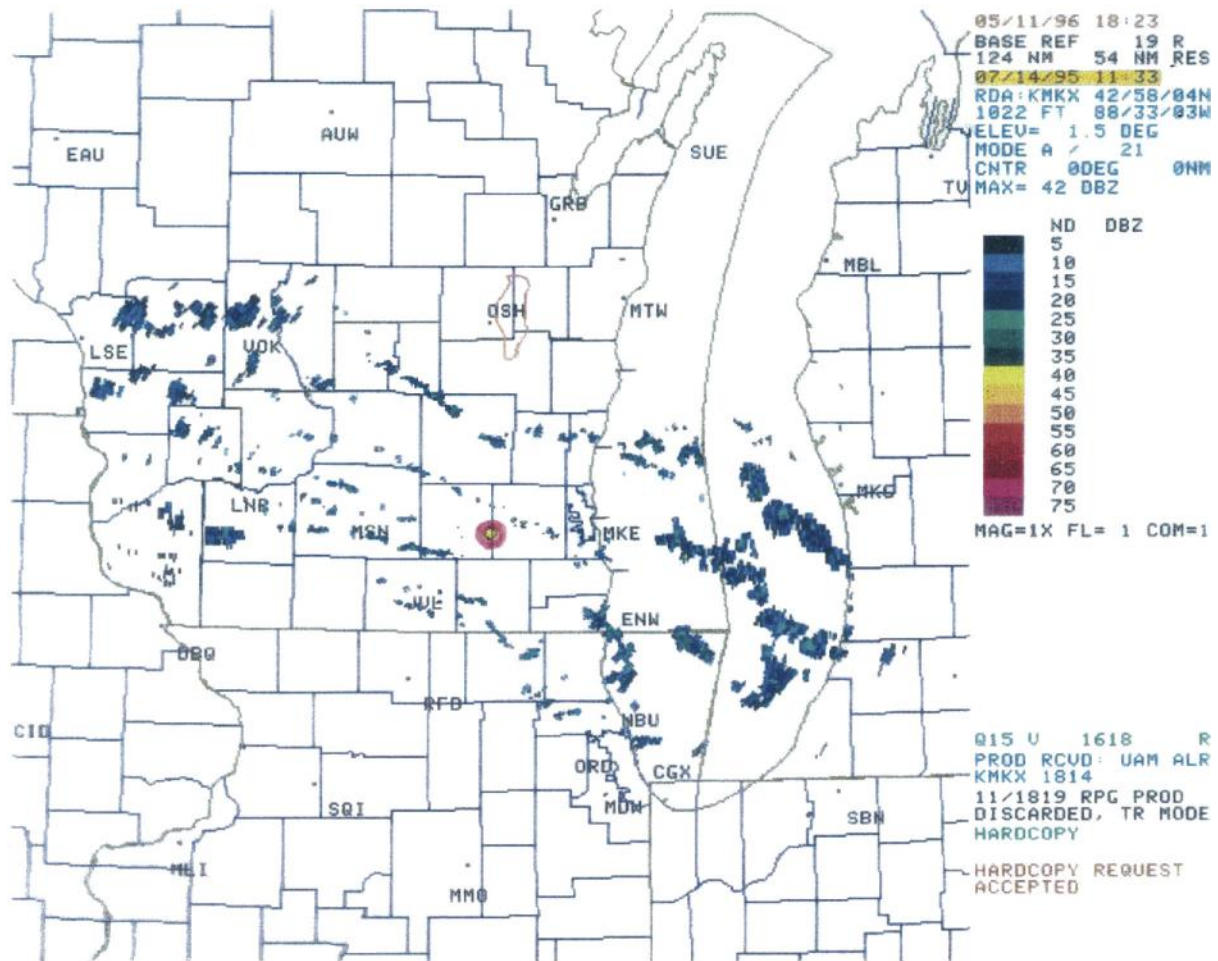


Figure 8 National Weather Service Doppler radar image from Sullivan, Wisconsin (circle at center of image) at 1133 UTC on 14 July 1995. The banded echoes extending east-southeast across central Wisconsin correspond to elevated convection at dawn during a deadly heat wave.

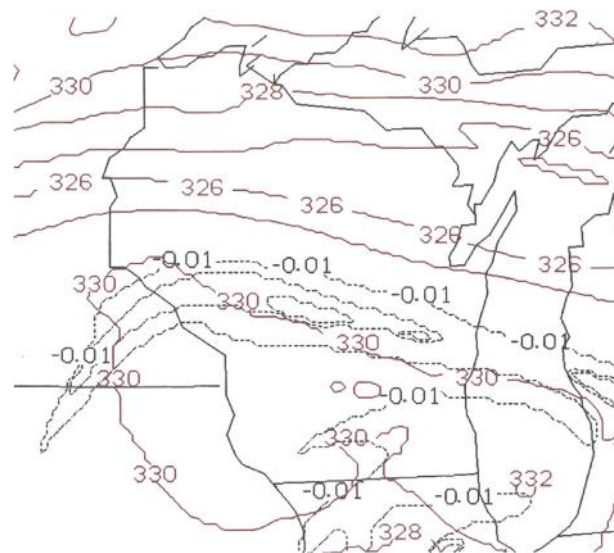


Figure 9 Potential vorticity (.025 PV unit contour interval, only negative values contoured) and equivalent potential temperature (2 K contour interval) at 4 km over the upper Midwest United States, as determined by a 24-hour forecast from the University of Wisconsin nonhydrostatic model (UW-NMS; horizontal resolution 6.67 km, vertical resolution 200–1000 m) valid at 1200 UTC 14 July 1995. Note the region of negative PV extending east-southeastward across central Wisconsin. The equivalent potential temperature field indicates the near-horizontal character of the flow over the region.

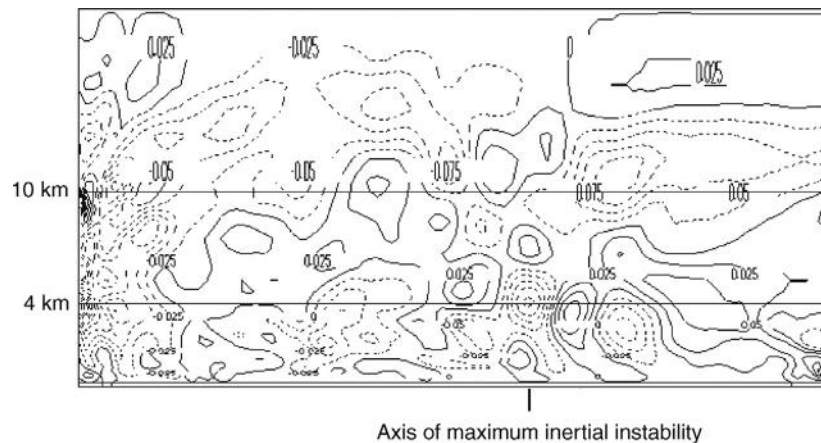


Figure 10 Cross-section of 24-hour forecast vertical motion (contoured every $0.025 \text{ m s}^{-1} = 2.5 \text{ cm s}^{-1}$) from UW-NMS valid 1200 UTC 14 July 1995. The cross-section slices NNW-SSE through the region of maximum inertial instability in Figure 9 (vertical mark below horizontal axis). The checkerboard pattern of vertical motion in the vicinity of the inertially unstable region should be compared with the schematic in Figure 2.

Summary

After several decades of off-and-on attention, inertial instability now seems firmly ensconced in the lexicon of geophysical fluid dynamics. The instability is at its largest and most observable in connection with breaking Rossby waves in the tropical middle atmosphere, but it can happen anywhere anticyclonic shear and/or curvature becomes unusually intense. Its purely horizontal origins and flattened pancake circulations make it orthogonal to static instability and much more difficult to observe. However, as a strongly ageostrophic instability it otherwise shares much in common with static instability. It is perhaps not too much of a stretch to call inertial instability by the nickname ‘horizontal convection’, while keeping in mind that rotation, not density, is at the heart of inertial instability.

Some outstanding research issues involving inertial instability include observation and theoretical explanation of its onset and three-dimensional structure; elucidation of its relationship to wave dynamics, other instabilities, and balanced dynamics; and further investigation of its role in mixing on a wide range of scales, from convection to planetary-scale flows.

See also: **Dynamical Meteorology:** Balanced Flow; Laboratory Geophysical Fluid Dynamics; Rossby Waves; Symmetric Stability; Vorticity; Wave Mean-Flow Interaction. **General Circulation of the Atmosphere:** Angular Momentum of the Atmosphere. **Mesoscale Meteorology:** Mesoscale Convective Systems. **Middle Atmosphere:** Planetary Waves; Semiannual Oscillation; Transport Circulation. **Oceanographic Topics:** General Processes. **Solar System/Sun, Atmospheres, Evolution of Atmospheres:** Planetary Atmospheres: Mars. **Synoptic Meteorology:** Anticyclones. **Tropical Meteorology and Climate:** Intertropical Convergence Zone; Monsoon: Dynamical Theory. **Turbulence and Mixing:** Turbulent Diffusion.

Further Reading

- Andrews, D.G., Holton, J.R., Leovy, C.B., 1987. *Middle Atmosphere Dynamics*. Academic Press, New York.
- Angell, J.K., 1962. The influence of inertial instability upon transsonic trajectories and some forecast implications. *Monthly Weather Review* 90, 245–251.
- Bjerknes, J., 1951. Extratropical cyclones. In: Malone, T.F. (Ed.), *Compendium of Meteorology*. American Meteorological Society, Boston, MA, pp. 577–598.
- Blanchard, D.O., Cotton, W.R., Brown, J.M., 1998. Mesoscale circulation growth under conditions of weak inertial instability. *Monthly Weather Review* 126, 118–140.
- Boyd, J.P., Christidis, Z.D., 1982. Low wavenumber instability on the equatorial beta-plane. *Geophysical Research Letters* 9, 769–772.
- Clark, P.D., Haynes, P.H., 1996. Inertial instability on an asymmetric low-latitude flow. *Quarterly Journal of the Royal Meteorological Society* 122, 151–182.
- Donnelly, R.J., 1991. Taylor-Couette flow: the early days. *Physics Today* 44 (11), 32–39.
- Dunkerton, T.J., 1981. On the inertial stability of the equatorial middle atmosphere. *Journal of Atmospheric Sciences* 38, 2354–2364.
- Dunkerton, T.J., 1993. Inertial instability of nonparallel flow on an equatorial b-plane. *Journal of Atmospheric Sciences* 50, 2744–2758.
- Emanuel, K.A., 1979. Inertial instability and mesoscale convective systems. Part I: Linear theory of inertial instability in rotating viscous fluids. *Journal of Atmospheric Sciences* 36, 2425–2449.
- Hayashi, H., Shiotani, M., Gille, J.C., 1998. Vertically stacked temperature disturbances near the equatorial stratopause as seen in cryogenic limb array etalon spectrometer data. *Journal of Geophysical Research (D16)*, 19469–19483.
- Hayashi, H., Shiotani, M., Gille, J.C., 2002. Horizontal wind disturbances induced by inertial instability in the equatorial middle atmosphere as seen in rocketsonde observations. *Journal of Geophysical Research* <http://dx.doi.org/10.1029/2001JD000922>, 31 July 2002.
- Hitchman, M.H., Leovy, C.B., Gille, J.C., Bailey, P.L., 1987. Quasi-stationary zonally asymmetric circulations in the equatorial lower mesosphere. *Journal of Atmospheric Sciences* 44, 2219–2236.
- Hoskins, B.J., 1974. The role of potential vorticity in symmetric stability and instability. *Quarterly Journal of the Royal Meteorological Society* 100, 480–482.
- Hua, B.L., Moore, D.W., Le Gentil, S., 1997. Inertial nonlinear equilibration of equatorial flows. *Journal of Fluid Mechanics* 331, 345–371.
- Hunt, B.G., 1981. The maintenance of the zonal mean state of the upper atmosphere as represented in a three-dimensional general circulation model extending up to 100 km. *Journal of Atmospheric Sciences* 38, 2172–2186.
- Knox, J.A., 1997. Generalized nonlinear balance criteria and inertial stability. *Journal of Atmospheric Sciences* 54, 967–985.
- Knox, J.A., 1997. Possible mechanisms of clear-air turbulence in strongly anticyclonic flows. *Monthly Weather Review* 125, 1251–1259.
- Mecikalski, J.R., Tripoli, G.J., 1998. Inertial available kinetic energy and the dynamics of tropical plume formation. *Monthly Weather Review* 126, 2200–2216.
- Munk, W., 2001. Spirals on the sea. *Scientia Marina* 65 (Suppl. 2), 193–198.

- Orsolini, Y.J., Limpasuvan, V., Leovy, C.B., 1997. The tropical stratopause in the UKMO assimilated analyses: Evidence for a 2-day wave and inertial circulations. *Quarterly Journal of the Royal Meteorological Society* 123, 1707–1724.
- O'Sullivan, D.J., Hitchman, M.H., 1992. Inertial instability and Rossby wave breaking in a numerical model. *Journal of Atmospheric Sciences* 49, 991–1002.
- Rayleigh Lord, 1916. On the dynamics of revolving fluids. *Proceedings of the Royal Society of London Series A* 93, 148–154.
- Rodwell, M.J., Hoskins, B.J., 1995. A model of the Asian summer monsoon. Part II: Cross-equatorial flow and PV behavior. *Journal of Atmospheric Sciences* 52, 1341–1356.
- Rosier, S.M., Lawrence, B.N., 1999. The January 1992 stratospheric sudden warming: A role for tropical inertial instability? *Quarterly Journal of the Royal Meteorological Society* 125, 2575–2596.
- Sato, K., Dunkerton, T.J., 2002. Layered structure associated with low potential vorticity near the tropopause seen in high-resolution radiosondes over Japan. *Journal of Atmospheric Sciences* 59, 2781–2800.
- Sawyer, J.S., 1949. The significance of dynamic instability in atmospheric motions. *Quarterly Journal of the Royal Meteorological Society* 75, 364–374.
- Smith, A.K., Riese, M., 1999. Cryogenic Infrared Spectrometers and Telescopes for the Atmosphere (CRISTA) observations of tracer transport by inertially unstable circulations. *Journal of Geophysical Research* 104 (D16), 19171–19182.
- Stevens, D.E., 1983. On symmetric stability and instability of zonal mean flows near the equator. *Journal of Atmospheric Sciences* 40, 882–893.
- Stevens, D.E., Ciesielski, P.E., 1986. Inertial instability of horizontally sheared flow away from the equator. *Journal of Atmospheric Sciences* 43, 2845–2856.
- Taylor, G.I., 1923. Stability of a viscous liquid contained between two rotating cylinders. *Philosophical transactions of the Royal Society of London A* 223, 289–343.
- Tomas, R.A., Webster, P.J., 1997. The role of inertial instability in determining the location and strength of near-equatorial convection. *Quarterly Journal of the Royal Meteorological Society* 123, 1445–1482.
- Wilson, R.J., 1997. A general circulation model simulation of the Martian polar warming. *Geophysical Research Letters* 24, 123–126.

Kelvin–Helmholtz Instability

PG Drazin[†], University of Bath, England, UK

© 2003 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 3, pp 1068–1072, © 2003, Elsevier Ltd.

Introduction

Kelvin–Helmholtz instability is the name given, since the 1940s, to an instability of a shear layer in a fluid, which is the mechanism of many phenomena observed in the atmosphere and oceans. It is said that in 1868 the German physiologist and physicist Hermann von Helmholtz first recognized the instability of a shear layer, by writing that “every perfect geometrically sharp edge by which a fluid flows must tear it asunder and establish a surface of separation, however slowly the fluid may move,” although this remark may seem to denote merely recognition of separation of a flow at an edge. However, in 1871 the British physicist William Thomson, Helmholtz’s friend who was later ennobled as Lord Kelvin, posed mathematically, and solved fully, a prototypical problem of linear instability of a horizontal vortex sheet between the uniform moving layers of two fluids of different densities, in an attempt to model the formation of ocean waves by the wind. Later, Helmholtz developed this model and applied it to the formation of billow clouds. At the same time the British physicist Lord Rayleigh was developing the theory of the instability of a shear layer, that is, a parallel flow in which the fluid speed varies across the layer.

These ideas were developed, extended, and applied in the twentieth century. An especially important extension is the instability of a horizontal shear layer in a stratified fluid, that is, a fluid whose density varies with height, because this models more realistically the mechanism of billow clouds, clear air turbulence, and other similar phenomena in the atmosphere.

The visionary British meteorologist Lewis Richardson recognized in the 1920s that atmospheric turbulence could be maintained only if the inertial instability due to shear could overcome the static stability due to heavier air being beneath lighter air. The essence of his argument can be recapitulated in terms of the energetics of the instability of a horizontal shear layer in a stratified fluid as follows. Suppose then that a basic flow of an incompressible inviscid fluid of variable density has velocity $U(z)\mathbf{i}$ and density $\rho(z)$, where $\mathbf{i}$ is a horizontal unit vector and z is the height. The essential mechanism of the instability is the conversion of the available kinetic energy of relative motion of the horizontal layers of the fluid into kinetic energy of a perturbation, overcoming the potential energy needed to raise or lower fluid when $d\rho/dz \leq 0$ everywhere, that is, when light fluid is always above heavier. Thus shear tends to destabilize and buoyancy to stabilize the flow. To quantify these tendencies, suppose that two neighboring fluid particles of equal volume, at heights z and $z + \delta z$, are

somehow interchanged. Then the increment of work δW per unit volume needed to overcome gravity and effect this interchange is

$$\delta W = -g\delta\rho\delta z$$

where g is the acceleration due to gravity and $\delta\rho = (d\rho/dz)\delta z$. In order that the horizontal momentum of the inviscid fluid is conserved in the interchange, the particle initially at height z will plausibly have final velocity intermediate between the velocities of the ambient fluid at its initial and final levels, say $(U + k\delta U)\mathbf{i}$, so that the other particle has final velocity $[U + (1 - k)\delta U]\mathbf{i}$ in order to conserve linear momentum, where $\delta U = (dU/dz)\delta z$ and k is some number between 0 and 1. Then the increment of kinetic energy δT per unit volume released by the interchange is

$$\begin{aligned}\delta T &= \frac{1}{2}\rho U^2 + \frac{1}{2}(\rho + \delta\rho)(U + \delta U)^2 - \frac{1}{2}\rho(U + k\delta U)^2 \\ &\quad - \frac{1}{2}(\rho + \delta\rho)[U + (1 - k)\delta U]^2 \\ &= k(1 - k)\rho(\delta U)^2 + kU\delta U\delta\rho\end{aligned}$$

on neglecting higher-order terms in $\delta\rho$ $(\delta U)^2$. For small increments,

$$\delta T = k(1 - k)\rho(\delta U)^2$$

on neglecting the inertial effects of the variation of density (this is a good approximation for instability in the atmosphere because the buoyancy effects of the variation of density there are almost always much greater). Thus,

$$\delta T = k(1 - k)\rho\left(\frac{dU}{dz}\right)^2(\delta z)^2 \leq \frac{1}{4}\rho\left(\frac{dU}{dz}\right)^2(\delta z)^2$$

with equality holding for $k = \frac{1}{2}$. Now a necessary condition for there being enough energy to effect this interchange, and hence for instability to occur, is that $\delta W \leq \delta T$, and therefore that

$$-g\frac{d\rho}{dz} \leq \frac{1}{4}\rho\left(\frac{dU}{dz}\right)^2$$

somewhere in the field of flow, or that

$$Ri(z) \leq \frac{1}{4}$$

where the local Richardson number is defined as the dimensionless quantity

$$Ri(z) \equiv -\frac{g d\rho/dz}{\rho(dU/dz)^2}$$

This criterion, namely that $Ri(z) \leq \frac{1}{4}$ somewhere in the flow, is a necessary condition for instability of the given basic flow, and

[†] Deceased.

so $Ri(z) > \frac{1}{4}$ everywhere in the flow is a sufficient condition for stability; it is called Richardson's criterion. The above argument for an incompressible inviscid fluid may be adapted for a perfect gas in adiabatic motion, and so for air in the atmosphere, by replacing the density with the potential density, and this leads again to Richardson's criterion but with the Richardson number redefined as

$$Ri(z) \equiv \frac{g}{T} \left(\frac{dT}{dz} + \Gamma \right) / \left(\frac{dU}{dz} \right)^2$$

where T is the absolute temperature and Γ the adiabatic lapse rate. For a perfect gas $\Gamma = g/c_p$, where c_p is the specific heat at constant pressure. In fact, $\Gamma \approx 8 \text{ K km}^{-1}$ in the troposphere.

Theory and Experiments

Now it is needful to go back to Kelvin's problem, and see a few of its details. Suppose that a basic horizontal flow of an incompressible inviscid fluid is given by

$$U(z) = \begin{cases} U_2, & \rho = \rho_2 \text{ for } z > 0 \\ U_1, & \rho = \rho_1 \text{ for } z < 0 \end{cases}$$

This gives a vortex sheet at $z = 0$, as sketched in [Figure 1](#). Kelvin was motivated by the special case with $\rho_2 \ll \rho_1$ to model wind

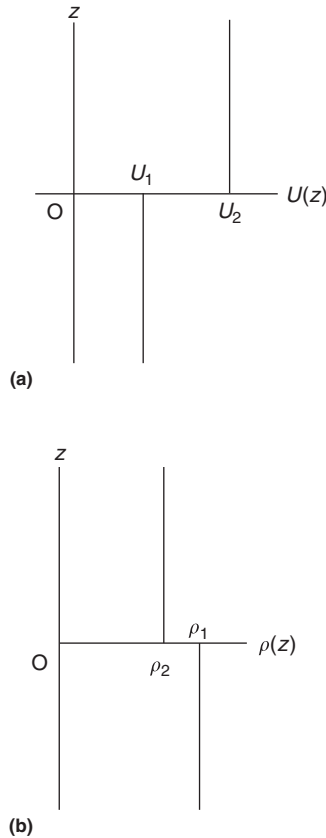


Figure 1 Sketch of (a) velocity and (b) density profiles in Kelvin's basic flow of two horizontal layers of fluid.

blowing on an ocean current. The occurrence of internal gravity waves in the special case with $U_2 = U_1$ may be anticipated. In any event, Kelvin took an irrotational flow coupled to a perturbation of the profile of the vortex sheet and resolved the perturbation into independent normal modes, with its flow quantities proportional to (or rather the real parts of functions proportional to)

$$\exp[i(kx + ly) + st]$$

where k and l are given horizontal wave numbers in the x - and y -directions (so the wavelengths in these directions are $2\pi/k$ and $2\pi/l$, respectively). He deduced, by solving an eigenvalue problem, that

$$s = -ik \frac{\rho_1 U_1 + \rho_2 U_2}{\rho_1 + \rho_2} \pm \left[\frac{k^2 \rho_1 \rho_2 (U_1 - U_2)^2}{(\rho_1 + \rho_2)^2} - \frac{g(k^2 + l^2)^{1/2} (\rho_1 - \rho_2)}{\rho_1 + \rho_2} \right]^{1/2}$$

Expressing $s = \sigma - i\omega$ in real and imaginary parts, ω as the frequency of the mode and σ as its relative growth rate are identified. It follows that if

$$k^2 \rho_1 \rho_2 (U_1 - U_2)^2 > g(k^2 + l^2)^{1/2} (\rho_1 - \rho_2) (\rho_1 + \rho_2)$$

then this mode may grow exponentially with

$$\sigma = \left[\frac{k^2 \rho_1 \rho_2 (U_1 - U_2)^2}{(\rho_1 + \rho_2)^2} - \frac{g(k^2 + l^2)^{1/2} (\rho_1 - \rho_2)}{\rho_1 + \rho_2} \right]^{1/2}$$

and the x -component of its phase velocity is the mass mean basic velocity

$$\frac{\omega}{k} = \frac{\rho_1 U_1 + \rho_2 U_2}{\rho_1 + \rho_2}$$

Now the flow is unstable if any one mode grows exponentially, and so if $\sigma > 0$ for any pair of real values k and l . But the formula above shows that $\sigma > 0$ for some modes with large values of k , provided that $U_2 \neq U_1$ and $\rho_1, \rho_2 > 0$, and therefore that all such two-layer flows are unstable to short waves. However, Kelvin himself showed that surface tension, as well as buoyancy, could in fact stabilize the flow. (In passing, one may note that the above formula for s gives the velocity of internal gravity waves at the interface of two fluids when $U_2 = U_1$.)

In 1931, Sydney Goldstein, Bernhard Haurwitz, and Geoffrey Taylor independently generalized Kelvin's model to deal with basic velocity and density profiles varying smoothly with height. Their results for various shear layers seemed to support Richardson's criterion, but John Miles and Louis Howard confirmed Richardson's criterion mathematically in the 1960s. Typical relative growth rates of a shear layer are found to be $\sigma \approx |U_2 - U_1|/10L$ when the Richardson number is appreciably less than a quarter (note that, by Richardson's criterion, $\sigma = 0$ if $Ri(z) \geq \frac{1}{4}$ everywhere), where L is the thickness of the layer. Taking $|U_2 - U_1| = 10 \text{ m s}^{-1}$ and $L = 100 \text{ m}$, as order of magnitude estimates for billow cloud formation, it is seen that the linear instability breaks up the shear layer with an e-folding time of about $\sigma^{-1} \approx 100 \text{ s}$. Thus Kelvin–Helmholtz instability is a transient process in the atmosphere. The transience presents

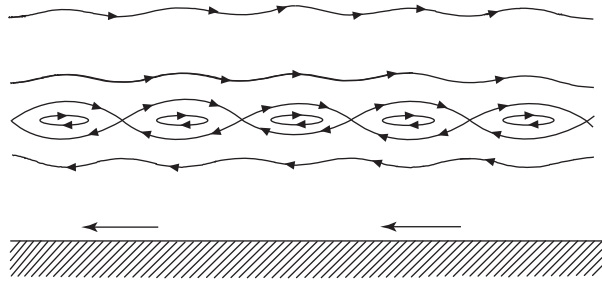
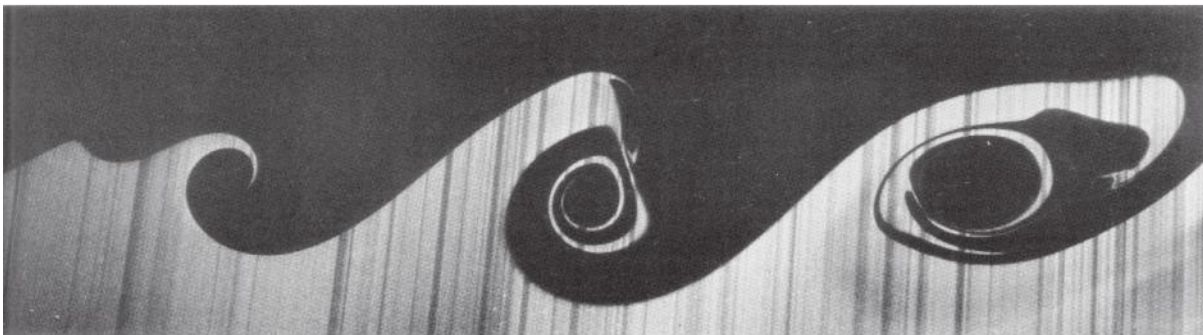
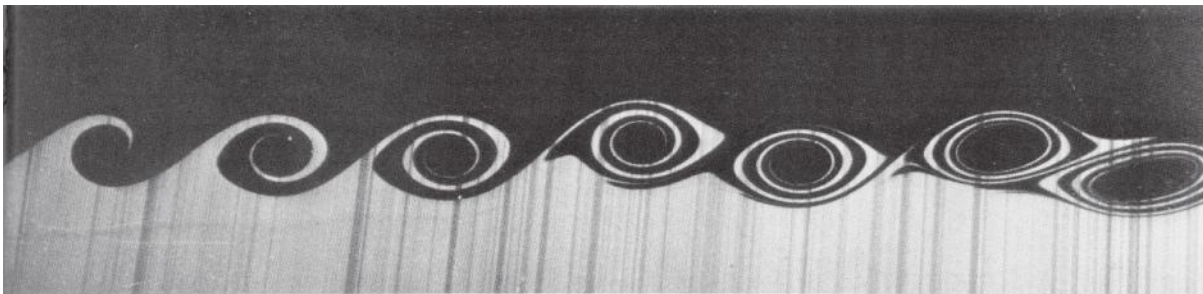


Figure 2 Kelvin's cat's eye pattern. This shows the streamlines near the level where the phase velocity of the waves equals the basic velocity of a smoothly varying profile of a homogeneous inviscid fluid.

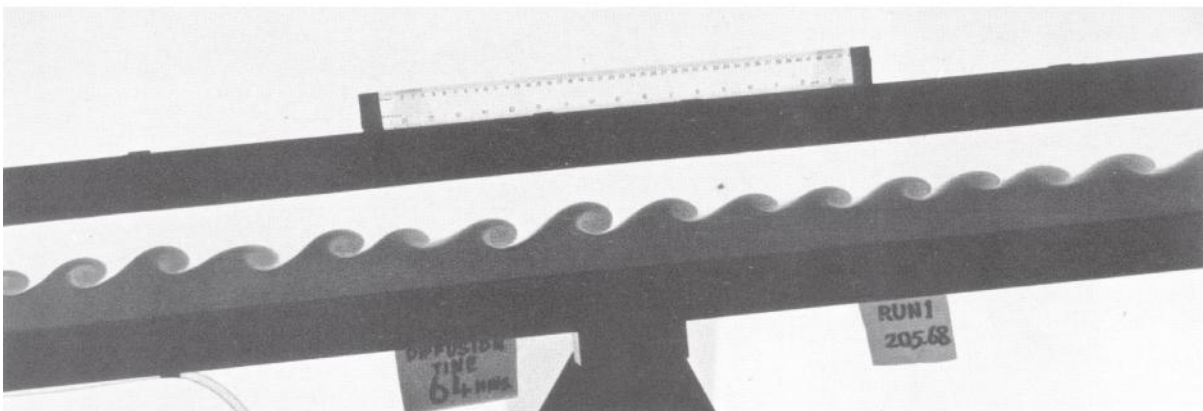
one way to distinguish billow clouds from lee wave clouds, which are forced orographically.

In 1880 Kelvin himself examined the streamlines due to instability of a shear layer of an unstratified fluid, and found what is now called Kelvin's cat's eye pattern, shown in [Figure 2](#). Taylor showed in 1931 that for a stratified fluid the regions of closed streamlines alternate at slightly different levels.

As the sinusoidal waves of the linear instability grow, nonlinearity will moderate the growth. A vortex sheet and a shear layer will then begin to roll up, as shown in [Figure 3](#). Of course, instability in the atmosphere is not so neat as in the careful laboratory experiments of [Figure 3](#).



(a)



(b)

Figure 3 (a) Photograph of instability of a shear layer. The lower stream of water moves leftward faster than the dyed upper stream. Photograph by FA Roberts, PE Dimotakis, and A Roshko. (b) Photograph of instability of a stratified shear layer. The long rectangular tube is filled with water above dyed brine. After the fluid came to rest the tube was suddenly tilted to create the shear layer with downward acceleration of the brine and upward acceleration of the water. The upper stream of water is moving rightward, and the lower stream of brine leftward. Photograph by SA Thorpe. Reproduced from Van Dyke, 1982. *An Album of Fluid Motion*, Parabolic Press, Stanford, CA.

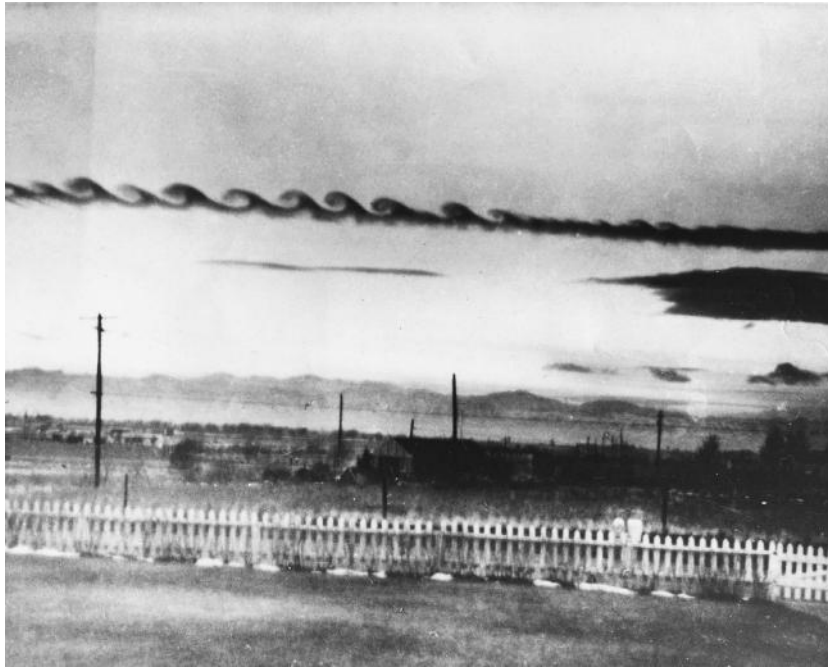


Figure 4 A row of billow clouds photograph by Paul E Branstine. Reproduced from Drazin, P.G., Reid, W.H., 1981. *Hydrodynamic Stability*, Cambridge University Press, Cambridge, UK, p. 22.

Atmospheric Phenomena

Kelvin–Helmholtz instability occurs in the atmosphere as a sporadic, but widespread, phenomenon. It is usually invisible and so can be dangerous as clear air turbulence. But it can be detected by radar or seen by chance as billow clouds when the humidity is such that the rising air in a vortex leads to condensation as cloud and the falling air leads to evaporation (see [Figure 4](#)). When seen at an angle, the lines of billow clouds are often called a ‘mackerel sky,’ this is because of their resemblance to the pattern on the back of a mackerel, the North Atlantic fish.

Kelvin–Helmholtz instability is not only of local significance. It, and the turbulence into which it usually develops, plays an important role in the energy budget of the atmosphere by transferring energy from the larger to smaller scales of motion until it is finally dissipated as heat by viscosity.

Further, it has been conjectured that shear instability in the absence of buoyancy plays a fundamental role in turbulence itself.

See also: **Aviation Meteorology:** Clear Air Turbulence. **Clouds and Fog:** Climatology. **Dynamical Meteorology:** Inertial Instability; Symmetric Stability; Waves. **Mountain Meteorology:** Lee Waves and Mountain Waves. **Gravity Waves:** Buoyancy and Buoyancy Waves: Theory. **Turbulence and Mixing:** Turbulence, Two Dimensional.

Further Reading

- Drazin, P.G., Reid, W.H., 1981. *Hydrodynamic Stability*. Cambridge University Press, Cambridge, UK. §§4.44, Ch. 4.
- Faber, T.E., 1995. *Fluid Dynamics for Physicists*. Cambridge University Press, Cambridge, UK. §§8.11, 8.12.
- Green, J., 1999. *Atmospheric Dynamics*. Cambridge University Press, Cambridge, UK. Ch. 6.
- Van Dyke, M., 1982. *An Album of Fluid Motion*, Photos 145–147. Parabolic Press, Stanford, CA.

Kelvin Waves

B Wang, University of Hawaii, Honolulu, HI, USA

© 2003 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 3, pp 1062–1068, © 2003, Elsevier Ltd.

Introduction

The Kelvin wave is a large-scale wave motion of great practical importance in the Earth's atmosphere and ocean. Discovered by Sir William Thompson (who later became Lord Kelvin) in 1879, the Kelvin wave is a special type of gravity wave that is affected by the Earth's rotation and trapped at the Equator or along lateral vertical boundaries such as coastlines or mountain ranges. The existence of the Kelvin wave relies on (1) gravity and stable stratification for sustaining a gravitational oscillation, (2) significant Coriolis acceleration, and (3) the presence of vertical boundaries or the Equator. An important feature of the Kelvin wave is its unidirectional propagation. The Kelvin wave moves equatorward along a western boundary, poleward along an eastern boundary, and cyclonically around a closed boundary (counterclockwise in the Northern Hemisphere and clockwise in the Southern Hemisphere). The wave amplitude is largest at the boundary and decays exponentially with distance from it. At the Equator, Kelvin waves always propagate eastward, reaching their maximum magnitude at the Equator and decaying exponentially with increasing latitude.

There are two basic types of Kelvin waves: boundary trapped and equatorially trapped. Each type of Kelvin wave may be further subdivided into surface and internal Kelvin waves. Surface, or barotropic, waves penetrate the entire depth of the fluid. Kelvin waves also appear within the stably stratified ocean and atmosphere, and are called internal, or baroclinic, Kelvin waves. Internal Kelvin waves are often found in a layer with large density gradients; the density gradient acts as an interface that allows the existence of internal gravity waves. Examples of such density gradients are the oceanic thermocline (a layer of large vertical temperature gradient separating a shallow layer of warm surface water about 50–200 m deep and a much deeper layer of cold water below) and the lower edge of an atmospheric inversion, a layer in which temperature increases with height. Like gravity waves, Kelvin waves can also propagate vertically in a continuously stratified geophysical fluid.

Atmospheric Kelvin waves play an important role in the adjustment of the tropical atmosphere to convective latent heat release, in the stratospheric quasibiennial oscillation (QBO), and in the generation and maintenance of the Madden–Julian Oscillation (MJO). Oceanic Kelvin waves play a critical role in tidal motion, in the adjustment of the tropical ocean to wind stress forcing, and in generating and sustaining the El Niño Southern Oscillation (ENSO).

Boundary-Trapped Kelvin Waves

Surface Kelvin Waves

The mechanism and properties of the Kelvin wave can be illustrated by considering a horizontally propagating Kelvin

wave in a rotating fluid of uniform finite depth H , where H is small compared to the horizontal extent of the fluid. The fluid has homogeneous density and a free surface, and is confined by a vertical lateral boundary. Such an idealized model is referred to in geophysical fluid dynamics as a shallow water model.

The lateral bounding wall prohibits flow across the boundary, and this absence of transverse motion with respect to the lateral boundary is a defining characteristic of Kelvin waves (Figure 1). Fluid parcels (elements) are constrained to move in a vertical plane parallel to the lateral boundary in the neighborhood of the boundary. Thus, the horizontal longshore (along the boundary) component of the Coriolis force must vanish. Consequently, the wave motion at the lateral boundary, and in any parallel vertical plane, is exactly the same as a hydrostatic gravity wave in a nonrotating system, i.e., the shallow water gravity wave (Figure 1). The wave travels along the boundary with the shallow water gravity wave speed C determined by the square root of the product of the gravitational acceleration (g) and the depth of the fluid, $C = (gH)^{1/2}$. The shape of the wave in the longshore direction is arbitrary and is conserved as the wave travels. This implies that the surface Kelvin wave is nondispersive, and that the wave energy

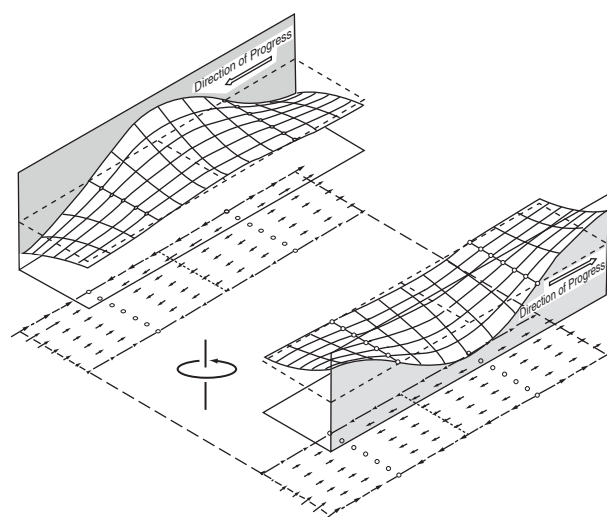


Figure 1 Northern Hemisphere Kelvin waves on opposite sides of a channel that is wide compared to the Rossby radius. In each vertical plane parallel to the coast, the currents (shown by arrows) are entirely within the plane and are exactly the same as those for a long gravity wave in a nonrotating channel. However, the surface elevation varies exponentially with distance from the coast in order to give a geostrophic balance. This means Kelvin waves move with the coast on their right in the Northern Hemisphere and on their left in the Southern Hemisphere. Reproduced from Gill, A.E., 1982. *Atmosphere–Ocean Dynamics*. Academic Press, New York, NY.

is transmitted at the speed of the shallow water gravity wave. Because the Kelvin wave solution in any vertical plane parallel to the lateral boundary is identical to that of the nonrotating case, the energy of a Kelvin wave is partitioned equally between kinetic energy and potential energy.

A fundamental difference between the Kelvin wave and the two-dimensional gravity wave is that the Kelvin wave can propagate in only one direction, rather than in two opposite directions. This is due to the constraints of the Earth's rotation and the presence of the lateral boundary. Rotation modifies the flow by piling up fluid against the lateral boundary, producing an offshore (normal to the boundary) pressure gradient force associated with the surface elevation. Since offshore motion is prohibited by the presence of the boundary, the offshore pressure gradient force must balance the Coriolis force associated with the longshore flow. For this reason, the motion is referred to as semigeostrophic (i.e., geostrophic balance is reached in only one direction). A direct consequence of this geostrophic balance is the exponential decay of the longshore velocity and surface height with distance from the lateral boundary (Figure 1). The Kelvin wave amplitude is significant only within an e-folding distance of the order of the Rossby radius of deformation (R) from the lateral boundary. This important length scale is defined by the ratio of the gravity wave speed, C , over the absolute value of the Coriolis parameter (f). Over this characteristic distance, the tendency of the gravitational force to flatten the free surface is balanced by the tendency of the Coriolis force to deform the surface. This is possible only for a wave traveling in the direction along which the lateral boundary (where the wave has maximum amplitude) is always on the right in the Northern Hemisphere and on the left in the Southern Hemisphere (Figure 1). Therefore, the effects of rotation and the lateral boundary determine the unidirectional propagation and the trapped behavior of the Kelvin wave.

Internal Kelvin Waves

The horizontally propagating internal, or baroclinic, Kelvin wave behaves in the same manner as the surface wave except that the motion varies with depth. Most frequently, it occurs at a stable interface (or a thin layer with large vertical density gradients) separating two relatively homogeneous layers. For internal Kelvin waves, the pressure gradient force normal to the lateral boundary arises from the tilt of the interface and is balanced by the Coriolis force associated with the vertical differential flow parallel to the boundary. The internal Kelvin wave speed depends on the density difference across the interface and is normally much slower than that of surface Kelvin waves. In the ocean, the typical speed for internal coastal Kelvin waves is of the order of 1 m s^{-1} and the Rossby radius of deformation is of the order of 10 km in the midlatitudes. Evidence for coastal Kelvin wave propagation along the eastern boundary of the Pacific has been observed in coastal sea level and temperature records.

In the atmosphere, boundary-trapped Kelvin waves occur primarily in the form of internal waves. They are often found along the edge of a plateau or a mountain range, such as the coast of South Africa, the west coast of California, and the eastern flank of the Tibetan Plateau. These internal Kelvin

waves are created near a stable inversion layer (often located at the top of the boundary layer) against steep topography. The elevated plateau or mountain range rises above the inversion, forming a lateral boundary for the air of the lower layer. Energy is prevented from escaping vertically by the inversion and laterally by the topography. The maximum disturbance intensity of these waves is found near the coast or plateau and decreases exponentially in intensity away from the coast or plateau. The offshore extent of the waves depends on both the thermal structure and topography. When the topography is steep, the depth of the inversion determines the Rossby radius for the atmospheric Kelvin wave. A typical value for atmospheric internal Kelvin waves is on an order of 1000 km.

Equatorially Trapped Kelvin Waves

Matsuno in 1966 showed that the eastward-propagating Kelvin wave is a possible free solution to the perturbation equations of the shallow water model on an equatorial β -plane (β is the meridional gradient of the Coriolis parameter), provided that the meridional velocity vanishes. This type of wave is called an equatorial Kelvin wave, so named because it is extremely similar in character to coastally trapped Kelvin waves, with the Equator serving as a boundary.

Like the coastal Kelvin wave, the propagation of an equatorial Kelvin wave is unidirectional, i.e., eastward only. In each vertical plane parallel to the equatorial vertical plane, the motion of fluid particles is precisely the same as that in a shallow water gravity wave (Figure 2). The Kelvin wave propagates without dispersion at the speed $C = (gH)^{1/2}$, as for nonrotating long gravity waves. Because the Coriolis parameter changes sign at the Equator, eastward flow occurring on both sides of the Equator would induce equatorward Ekman mass transport, piling up fluid at the Equator and generating a meridional pressure gradient force. In this sense, the Equator acts as a lateral wall. The Earth's rotation links the motion in each latitudinal plane, because momentum conservation in the north-south direction requires a geostrophic balance between the eastward velocity and the meridional pressure gradient force. This geostrophic balance results in the perturbation zonal velocity reaching a maximum on the Equator and decaying with increasing distance from the Equator (Figure 2). This is possible only for eastward-traveling waves. Thus, equatorial Kelvin waves are eastward-propagating and have zonal velocity and pressure perturbations that vary with latitude as Gaussian functions centered on the Equator (Figure 2). The e-folding distance for decay with increasing latitude is given by $Rc = (C/2\beta)^{1/2}$, where $C = (gH)^{1/2}$ is the gravity wave speed and β is the meridional gradient of the Coriolis parameter at the Equator. Rc is called the equatorial Rossby radius of deformation, because of its relationship with the decay scale for the case of constant Coriolis parameter. The same analysis can be applied to baroclinic waves in both atmosphere and ocean, with H being interpreted as the equivalent depth. Typical values of the internal gravity wave speed, C , for the tropical atmosphere are $20\text{--}80 \text{ m s}^{-1}$, giving an equatorial Rossby radius of between 6° and 12° of latitude. For baroclinic ocean waves, appropriate values of C are typically in the range of $2\text{--}3 \text{ m s}^{-1}$, so that the equatorial Rossby radius is $200\text{--}250 \text{ km}$.

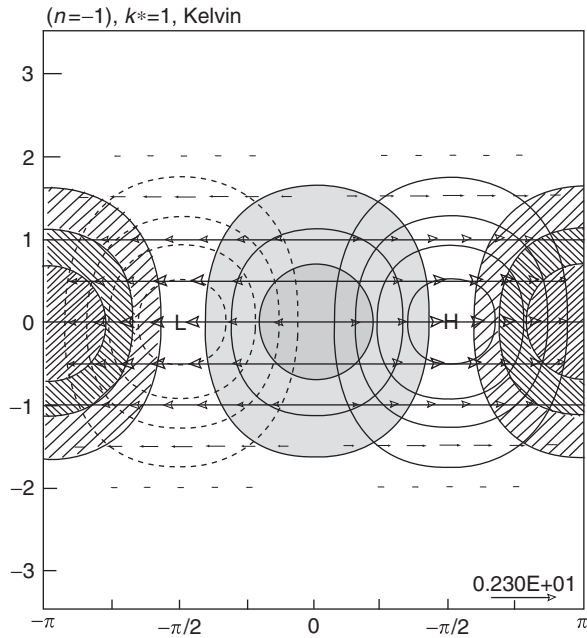


Figure 2 The theoretical equatorially trapped Kelvin wave solution to the linear shallow water equations on an equatorial β -plane for a nondimensional zonal wave number 1, which is constructed from the theory of Matsuno (1966). Hatching is for convergence and shading for divergence, with a .6 unit interval between successive levels of hatching or shading, and with the zero divergence contour omitted. Unshaded contours are for geopotential, with a contour interval of .5 units. Negative contours are dashed and the zero contour is omitted. The largest wind vector is 2.3 units, as marked. Reproduced from Wheeler, M., et al. 2000. Large-scale dynamical fields associated with collectively coupled equatorial waves. *Journal of the Atmospheric Sciences* 57(5), 613–640.

Vertically Propagating Kelvin Waves

In general, the Earth's rotation traps planetary-scale gravity waves in the troposphere unless the frequency of the wave is greater than the Coriolis frequency (about $(1 \text{ day})^{-1}$). For this reason, midlatitude synoptic waves are generally unable to penetrate significantly into the stratosphere. However, near the Equator, the dramatic decrease in the Coriolis parameter allows these longer period waves to propagate vertically. Vertically propagating Kelvin waves have been identified in both the equatorial atmosphere and ocean.

Vertically propagating Kelvin waves can be illustrated by considering a continuously stratified fluid with a constant buoyancy frequency in a semiinfinite vertical domain near a lateral boundary or in the vicinity of the Equator. For simplicity, consider a linear equatorial β -plane model. Solutions can be obtained using the normal mode technique by neglecting meridional perturbations. In a vertical section along the Equator, or in any parallel vertical plane, the motion of vertically propagating equatorial Kelvin waves shares the properties of an ordinary vertically propagating gravity wave. A vertical cross section of the perturbation motion, pressure, and temperature structure for vertically propagating Kelvin waves is shown in Figure 3. The local change in temperature is due to adiabatic warming or cooling, so that the temperature

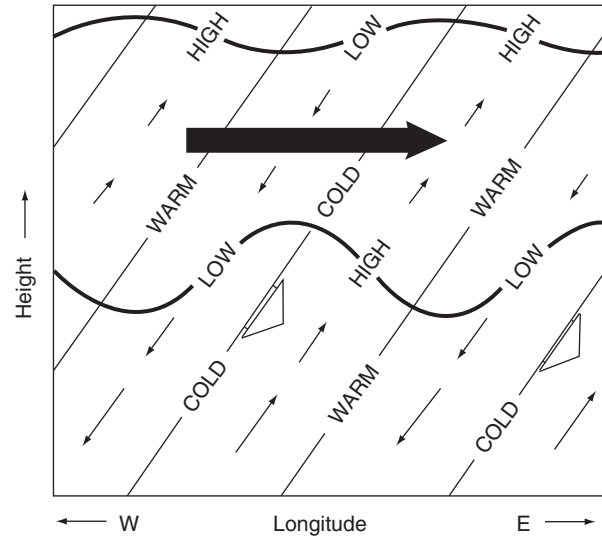


Figure 3 Longitude–height section along the Equator showing pressure, temperature, and wind perturbations for a thermally damped Kelvin wave. Heavy wavy lines indicate material lines; and short blunt arrows show phase propagation. Length of the small thin arrows is proportional to the wave amplitude, which decreases with height owing to damping. The large shaded arrow indicates the net mean flow acceleration owing to the wave stress divergence. Reproduced from Holton, J.R., 1992. *Introduction to Dynamic Meteorology*, third ed. Academic Press, San Diego, CA.

oscillation leads the vertical (zonal) wind and pressure oscillations by a quarter cycle.

Since stratospheric Kelvin waves are forced from below by disturbances in the troposphere, the wave energy propagation must have an upward component. According to theory, Kelvin waves become dispersive in the presence of vertical propagation, and the vertical component of their phase velocity is always opposite to that of their group velocity (the velocity at which the wave energy is transmitted). Thus, the phase velocity must have a downward component. The condition of eastward propagation due to equatorial trapping requires that the vertical wave number be negative. Hence, an eastward and downward-propagating equatorial Kelvin wave has constant phase lines that tilt eastward with height (Figure 3).

Because the variation in zonal wind depends on the pressure gradient force, the highest zonal pressure gradient precedes the largest westerly acceleration by a quarter wavelength, and the zonal wind and pressure waves coincide. This creates an upward flux of wave energy. An individual parcel moves up along the tilted phase line, bringing westerly momentum upward, and moves down, bringing easterly momentum downward. Thus, the Kelvin wave transports westerly momentum upward.

Significance of Kelvin Waves in the Atmosphere and Ocean

Oceanic Kelvin Waves

Kelvin waves are essential in the description of ocean tides. For a deep ocean ($H = 5 \text{ km}$) at 30° N , the Rossby radius for barotropic Kelvin waves is about 3000 km. Continental shelf

regions normally extend about a hundred kilometers seaward; hence, a steep continental slope is practically indistinguishable from a vertical boundary at the scale of the Rossby radius. Thus, a barotropic Kelvin wave extends far from the coast and occupies a substantial fraction of a typical ocean. Much of the energy of tide waves traveling along continents is transmitted in the form of barotropic Kelvin waves with a speed of about 200 m s^{-1} . For instance, along the coast of California more than two-thirds of the semidiurnal and half the diurnal tidal amplitudes can be accounted for by traveling barotropic Kelvin waves. For shallow seas and coastal waters, the Rossby radius is about 200 km. When a Kelvin wave moves through a region in which the fluid depth or the Coriolis parameter varies and the wave energy flux remains constant, the amplitude of the wave varies in proportion to $(f/H)^{1/2}$. Thus, wave amplitude increases when Kelvin waves move into shallow water. In coastal regions, Kelvin waves can also be generated as storm surges are diffracted by vertical boundaries and scattered by irregular coastlines. Variable longshore winds and atmospheric pressure gradients acting on the sea surface are also possible energy sources for oceanic Kelvin waves.

Internal coastal Kelvin waves can be generated by wind-induced, time-dependent coastal upwelling. Coastal upwelling (downwelling) is caused by an Ekman mass flux transported offshore (onshore) and forced by longshore winds. The disturbances can then propagate along the coast as boundary-trapped internal Kelvin waves. Therefore, the amount of upwelling depends not only on local wind forcing but also on the forcing that generated the waves at an earlier time. In a lake, during strong wind forcing, the region of upwelling progresses cyclonically (in the Northern Hemisphere) at the speed of an internal Kelvin wave.

Equatorial Kelvin waves play a critical part in thermocline adjustment. As the equatorial ocean circulation responds to wind stress forcing, equatorial Kelvin waves transmit signals rapidly from the western to the eastern extremities of the ocean basin. Internal Kelvin waves propagating along the thermocline take about 2 months to cross the entire equatorial Pacific. These equatorial Kelvin waves play an essential role in the ENSO cycle. First, they support a positive feedback between the central Pacific zonal wind and eastern Pacific sea surface temperature (SST) anomalies. A westerly wind anomaly excites

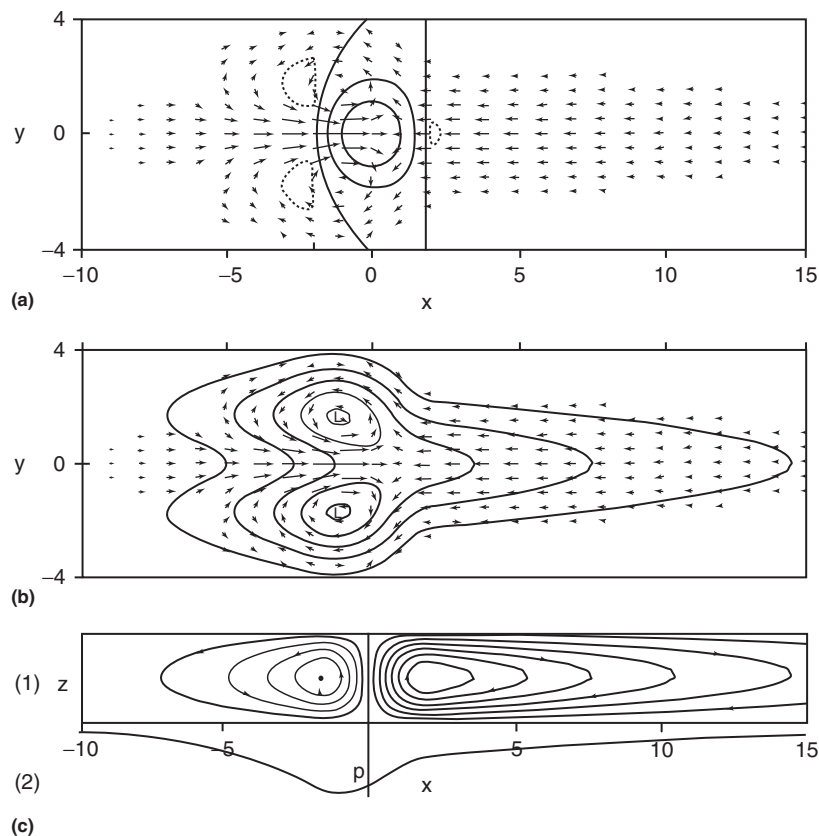


Figure 4 Solution for heating symmetric about the Equator in the region $|x| < 2$ for decay factor $\varepsilon = 0.1$. (a) Contours of vertical velocity w (solid contours are 0, 0.3, 0.6, broken contour is -0.1) superimposed on the velocity field for the lower layer. The field is dominated by the upward motion in the heating region where it has approximately the same shape as the heating function. Elsewhere there is subsidence with the same pattern as the pressure field. (b) Contours of perturbation pressure p (contour interval 0.3) which is everywhere negative. There is a trough at the Equator in the easterly regime to the east of the forcing region. On the other hand, the pressure in the westerlies to the west of the forcing region, though depressed, is high relative to its value off the Equator. Two cyclones are found on the northwest and southwest flanks of the forcing region. (c) The meridionally integrated flow showing: (1) stream function contours and (2) perturbation pressure. Note the rising motion in the heating region (where there is a trough) and subsidence elsewhere. Reproduced from Gill, A.E., 1980. Some simple solutions for heat-induced tropical circulation. *Quarterly Journal of the Royal Meteorological Society* 106, 447–462.

downwelling Kelvin waves, which propagate into the eastern Pacific, suppressing the thermocline and causing the SST to rise; this, in turn, enhances the central Pacific westerly wind anomaly by increasing the eastward pressure gradient force in the atmosphere. This positive feedback provides a development mechanism for ENSO SST warming. Second, the cyclonic wind stress curl associated with the central Pacific westerly wind anomaly can induce upwelling oceanic Rossby waves that propagate westward. These waves are eventually reflected at the western ocean boundary, generating upwelling equatorial Kelvin waves, which propagate into the eastern Pacific and offset the warming by enhancing vertical cold advection. This negative feedback provides a mechanism for turning the coupled system to its opposite (La Niña) phase and sustaining the ENSO cycle. In addition, the atmospheric intraseasonal wind forcing continuously generates equatorial Kelvin waves whose nonlinear rectification to the mean state may also contribute to the eastern Pacific warming.

Atmospheric Equatorial Kelvin Waves

The atmospheric equatorial Kelvin wave is one of the critical wave motions in the response of the tropical atmospheric circulation to a heat source (Figure 4). When an imposed heating centered on the Equator is switched on at some initial time, Kelvin waves carry information rapidly eastward, thereby creating easterly trade winds in that region and forming a Walker cell (rising motion over the heat source region and sinking motion to its east) (Figure 4(c)). Internal equatorial Kelvin waves traveling with typical speeds of $20\text{--}80\text{ m s}^{-1}$ are an effective means by which the equatorial atmosphere becomes homogenized in the zonal direction. The easterly winds are in geostrophic equilibrium, so that there is a trough along the Equator, with the winds along the Equator flowing

directly down the pressure gradient (Figure 4). The westward-propagating Rossby wave regime to the west of the forcing region is about one-third the size of the Kelvin wave regime because Rossby waves travel at one-third of the Kelvin wave speed (Figure 4(a)). The equatorial westerlies between the symmetric Rossby waves provide inflow into the heating region, and are in geostrophic equilibrium, so that a relative ridge appears along the Equator (Figure 4(b)). Meanwhile, the flow converges toward the Equator. If atmospheric damping is taken into account, this simple model can largely explain the steady-state atmospheric response to an imposed heat source.

Kelvin and mixed Rossby-gravity waves are the predominant disturbances in the equatorial stratosphere, and play a critical role in the stratospheric circulation through their vertical transport of energy and momentum. Stratospheric Kelvin waves are excited by oscillations in the large-scale convective heating pattern in the troposphere, and are a source of westerly momentum for the QBO (see **Middle Atmosphere**: Quasi-Biennial Oscillation). The QBO is a zonal wind oscillation in the equatorial stratosphere, and propagates downward with a period of about 24–30 months. Figure 5 shows an example of the zonal wind oscillations caused by the passage of Kelvin waves near the Equator. The descent of the westerly phase of the QBO is shown in the figure. At each level, there is an increase of the zonal wind with time. Superposed on this secular trend is a fluctuating Kelvin wave component with a period of about 12 days and a vertical wavelength of about 10–12 km. As vertically propagating Kelvin waves carry westerly momentum upward, they are damped by radiative cooling, small-scale turbulence, and critical level interaction. As the waves are damped, they lose momentum and accelerate the westerly mean flow. The damping depends on the Doppler-shifted wave frequency. As the Doppler-shifted frequency decreases, the vertical

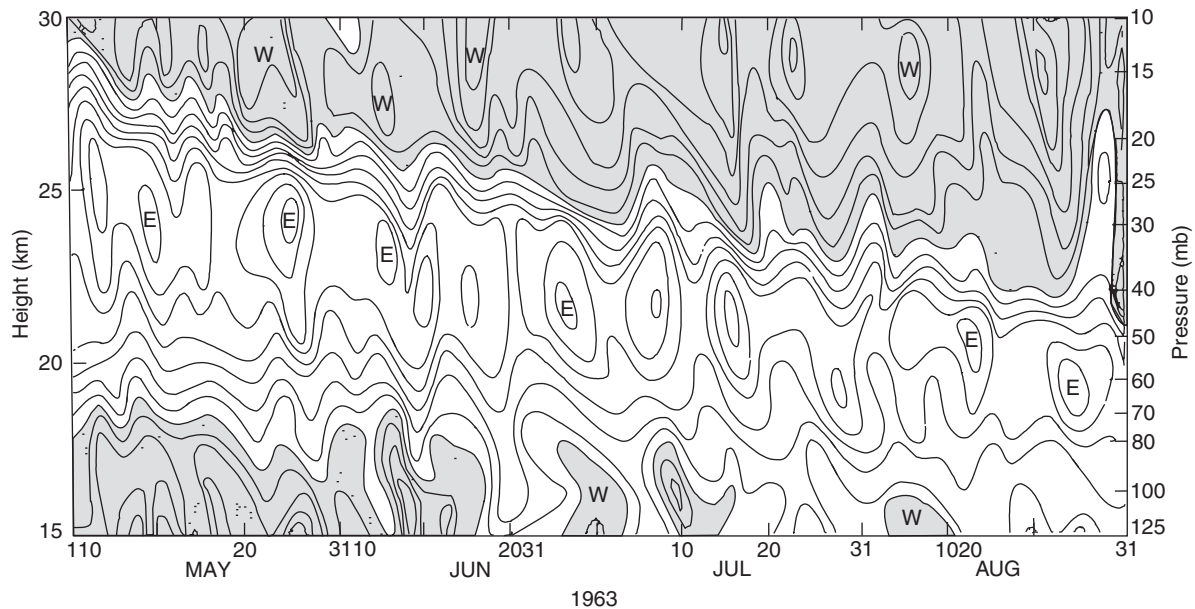


Figure 5 Time–height section of zonal wind at Canton Island (3° S). Isotachs at intervals of 5 m s^{-1} . Westerlies are shaded. Reproduced from Holton, J.R., 1992. Introduction to Dynamic Meteorology, third ed. Academic Press, San Diego, CA. Original courtesy of Wallace, J.M., Kousky, V.E.

component of the group velocity also decreases, and a longer time is available for the wave energy to be damped. Hence, westerly Kelvin waves tend to be damped preferentially in the westerly shear zone where their Doppler-shifted frequencies decrease with height. The associated momentum flux convergence produces westerly acceleration of the mean flow, causing the westerly shear zone to descend. A similar argument is valid for the downward propagation of the easterly phase of the QBO through the action of Rossby-gravity waves.

A peak in the variability of the tropical atmosphere appears in the 30- to 60-day period range, and is known as the MJO. Madden and Julian found that a 30- to 60-day oscillation in zonal winds is in approximate geostrophic balance with varying pressure maxima and minima centered at the Equator. A low-level low-pressure anomaly is accompanied by low-level easterly anomalies. The pressure and zonal wind are out of phase in the upper troposphere. The wave patterns move eastward along the Equator. At the Equator, the meridional winds appear to be insignificant. The amplitude of the oscillation decays with distance away from the Equator. These features are similar to those of internal equatorial Kelvin waves except that the large vertical scale of the MJO implies a faster phase speed than is observed. Arguments involving coupling with equatorial westward-traveling Rossby waves and interaction with the release of latent heat in the

disturbances as well as viscous damping have been invoked to explain the observed slow phase speed.

See also: **Dynamical Meteorology:** Rossby Waves; Waves. **Middle Atmosphere:** Quasi-Biennial Oscillation. **Oceanographic Topics:** Surface/Wind Driven Circulation. **Tropical Cyclones and Hurricanes:** Overview and Theory. **Tropical Meteorology and Climate:** El Niño and the Southern Oscillation: Observation; El Niño and the Southern Oscillation: Theory; Intraseasonal Oscillation (Madden–Julian Oscillation).

Further Reading

- Cushman-Roison, B., 1994. *An Introduction to Geophysical Fluid Dynamics*. Prentice-Hall, Englewood Cliffs, NJ.
- Gill, A.E., 1982. *Atmosphere–Ocean Dynamics*. Academic Press, New York, NY.
- Holton, J.R., 1992. *Introduction to Dynamic Meteorology*, third ed. Academic Press, San Diego, CA.
- LeBlonde, P.H., Lawrence, A.M., 1978. *Waves in the Ocean*. Elsevier, New York, NY.
- Matsuno, T., 1966. Quasi-geostrophic motion in the equatorial area. *Journal of the Meteorological Society of Japan* 44, 25–42.
- Pedlosky, J., 1987. *Geophysical Fluid Dynamics*. Springer-Verlag, Berlin.
- Philander, S.G., 1990. *El Niño, La Niña and the Southern Oscillation*. Academic Press, New York, NY.

Kinematics

DD Houghton, University of Wisconsin-Madison, Madison, WI, USA

© 2003 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 3, pp 1072–1079, © 2003, Elsevier Ltd.

Introduction

Kinematics, when applied to the atmosphere, refers to the description of both air motion and the motion of patterns describing other properties of air, such as moisture content, temperature, and pressure. This description is without regard to forces and other physical processes that cause the motions. Air motion itself is an important causal factor for many of the pattern changes of the other properties of the air. Daily sequences of weather maps showing horizontal wind flow and the movement of patterns in pressure and temperature dramatize the kinematic perspective of the atmosphere.

This article presents an overview of atmospheric kinematics as follows. First, a brief historical perspective is presented. Then, basic characteristics and descriptors for air motion are reviewed. Finally, examples of the wide range of applications of kinematic analysis of the atmosphere are presented.

History

Studies of the kinematics of air motion date back to the advent of observing system networks for wind which first became available in the nineteenth century over parts of Europe and the United States. Kinematics was applied to early studies of the extratropical cyclone. Historical presentations by Kutzbach in 1979 and Eliassen in 1999 document how the analysis and development of theory for these cyclones depended on the kinematics of the surface wind field. In particular, conflicting theories were resolved by noting whether the dominant surface wind flow was primarily convergent, rotational, or confluent – three of the important basic air motion descriptors of atmospheric kinematics to be discussed later.

Weather map depictions of the atmosphere became common in the twentieth century. Kinematics became a basic analysis perspective not only for wind fields but also for describing the motions of weather map patterns, such as the movement of pressure isobars, and weather fronts. The determination of large-scale vertical motion (a quantity that could not be measured directly) was an important application of kinematics. Using a kinematic method, vertical motion was determined from the horizontal divergence in the observed wind field using a relationship from the continuity of mass equation.

By the middle of the twentieth century, kinematic analysis for the atmosphere had become fully developed. Authors such as Saucier in 1955 and Pettersen in 1956 described a wide range of atmospheric kinematic principles and applications.

Basic Air Motion Properties

Motion Characteristics

The advection (transport) of atmospheric properties by air motion is a fundamental means by which the air motion affects the state of the atmosphere. This advection alone can change conditions at a specific point in space as patterns of fluctuation in the atmosphere are moved past this position. For instance, a west wind in an area where the air temperature is warmer to the west than to the east will tend to increase temperatures in the middle area (Figure 1). The mathematical description of this effect can be expressed as

$$\partial T / \partial t = -u \partial T / \partial x \quad [1]$$

where T is the temperature, t the time, u the air motion speed in the west–east direction, x the distance in the west–east direction, and ∂ the partial derivative operator. The right-hand side quantity is generally referred to as thermal advection. The temperature change, $\partial T / \partial t$, is positive as u is positive and $\partial T / \partial x$ is negative for the conditions described above.

The fact that the air motion, itself, varies in space adds an important characteristic to the advection effect. Namely, the differential in advection effects can change the shape and pattern of the quantity of air being moved at the same time it is being displaced. These effects have been examined most thoroughly for the two-dimensional case, i.e., the horizontal wind field and horizontal distribution of air properties. Significant distortions in horizontal patterns can arise with differential advection acting over sufficiently long durations of time.

The differential advection capabilities of the horizontal wind field can change the horizontal distribution patterns of properties such as temperature, pressure, moisture, or the wind field itself. Spatial variation in a velocity field can be described as the summation of four distinct properties of the wind field. These properties and their effects on an individual element, for example a specified region of water vapor, are summarized below and in Figure 2:

1. *uniform translation*: movement with no change in shape, orientation, or horizontal extent;
2. *rotation*: change in orientation without change in shape or horizontal extent;

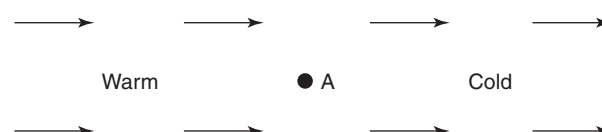


Figure 1 Temperature and wind pattern (arrows) that lead to warming at Point A by the advection (transport) process due to the air motion.

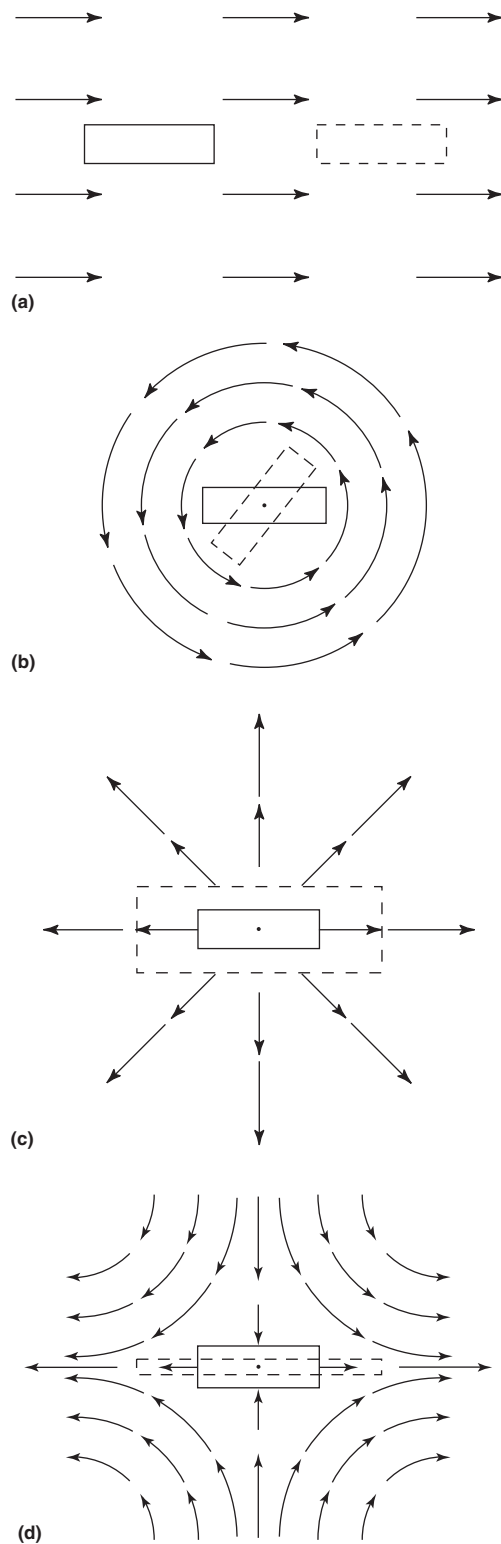


Figure 2 Four basic kinematic properties of air motion and the effect of each, respectively, on an element being advected (transported) by the air. Panels (a–d) show uniform translation, rotation, divergence, and deformation, respectively, for horizontal motion. The solid line enclosed area shows the advected element at the initial time and the dashed-line enclosed area shows the element at a slightly later time.

3. *divergence*: change in horizontal extent or area without change of shape and orientation; and
4. *deformation*: change in shape without rotation or change in horizontal extent.

Mathematical definitions of these four properties for the two-dimensional wind field can be derived using a linear Taylor series expansion of the velocity field with reference to the center position of the specified area of interest. Let u_0 and v_0 be the values of the x component (eastward) and y component (northward) of the horizontal velocity at the center point and $u(x, y)$ and $v(x, y)$ be the velocity component values at distances x and y from the center point. Then, the wind field characteristics relevant for each of the four effects, respectively, are:

1. *uniform translation*: movement with speed u_0 in the x direction and v_0 in the y direction;
2. *rotation*: turning defined by the vertical component of vorticity, $\zeta = \partial v / \partial x - \partial u / \partial y$;
3. *divergence*: expansion defined by horizontal divergence, $\delta = \partial u / \partial x + \partial v / \partial y$; and
4. *deformation*: shape distortion at a rate defined by the horizontal component magnitude.

$$|D| = [(\partial u / \partial x - \partial v / \partial y)^2 + (\partial v / \partial x + \partial u / \partial y)^2]^{1/2}$$

Note that ‘vorticity’ referenced in the rotation definition may be described as ‘relative vorticity’ in atmospheric applications where velocities are measured ‘relative’ to the Earth’s surface which itself is turning. Also note that the deformation expression is more complex than the vorticity and divergence expression in that it depends on the position of the coordinates, i.e., the details of deformation effects depend on the orientation of the deformation field. The deformation has two components described in the definition above by the two quantities that are squared inside the square root function.

The basic differential advection characteristics can be generalized for the full three-dimensional wind field. This results in many more scalar descriptors for the motion properties compared to the two-dimensional case presented above. The two-dimensional (horizontal) perspective is sufficient for many atmospheric applications, especially for large-scale motions, such as those depicted in weather maps, where the flow field is predominantly horizontal. This would not be true for small-scale systems such as turbulence near the Earth’s surface, cumulus clouds, and tornadoes. In these cases, vertical motion fluctuations can be as large as those for the horizontal motions.

Motion Descriptors

There are a number of descriptors for air motion that are frequently used in kinematic analysis of the atmosphere. It is useful to define them and to indicate why they are useful and how they relate to the fundamental definition of motion. Air motion is a vector with a direction and magnitude (speed) that alternatively may be described by the components of motion in the directions of the coordinate system. In the three-dimensional atmosphere, an example would be the u , v , and w components in the eastward (x), northward (y), and vertical (z) directions, respectively.

The four distinct properties of the air motion described above (uniform translation, vorticity, divergence, and deformation) are not only important components for the kinematic description but also serve as basic variables for analyzing the dynamics of the system. For large-scale horizontal motions where the vertical vorticity component is generally larger than the horizontal divergence and deformation components, the governing equations of motion can be approximated by equations for the rate of change of the vertical component of vorticity. Divergence has a direct relationship to vertical motion and is often used in descriptions for large-scale airflow to represent vertical motion conditions. Deformation has more limited use as a separate variable. Its most basic application is in relation to analysis for atmospheric frontogenesis, the situation where horizontal variations (gradients) of temperature, or other atmospheric descriptors, are increasing.

Six other descriptors used to characterize air motions in the atmosphere are discussed below.

1. *Streamline*: line parallel to and following along the flow direction in space at a given time. Streamlines give information only on direction of flow, not the speed of flow. The direction taken by smoke coming out of a smokestack will be in the same direction as the streamlines at the time when it comes out. Often streamlines are shown for two-dimensional space.
2. *Isotach*: contour surface connecting positions with the same wind speed. On a two-dimensional chart, isotachs are contour lines which together with streamlines give a complete description of the two-dimensional flow field.
3. *Trajectory*: line following the course of an air parcel element which is moving with the air motion. In general, the trajectory is three-dimensional. Often trajectories are shown for a nearly horizontal two-dimensional space such as the Earth's surface. Trajectories will follow streamlines if flow conditions are steady state. The track taken by a single puff of smoke emitted from a smokestack defines its trajectory.
4. *Streakline*: line connecting the air parcel elements whose trajectories all emanate from a single position in space. The pattern produced by a smokestack that is continually emitting smoke is one example of a streakline.
5. *Streamfunction*: a function in two-dimensional space (e.g., $\psi(x, y)$) whose spatial gradients define the wind flow for nondivergent flow conditions in that space in the following way: $u = -\partial\psi/\partial y$ and $v = \partial\psi/\partial x$. The contours of constant streamfunction, like the streamlines, are parallel to the wind flow. In addition, the gradients of streamfunction value perpendicular to the lines of constant streamfunction define the wind speed. Thus, in a map showing streamfunction contoured at a regular interval, the distance between adjacent contours will be inversely related to the wind speed (Figure 3). For many quasihorizontal, large-scale atmospheric flows, which have negligible horizontal divergence, the streamfunction is sufficient to define the horizontal wind field. The Laplacian of streamfunction defines the associated vertical component of vorticity, ζ , according to the relationship

$$\zeta = \nabla^2\psi = \partial^2\psi/\partial x^2 + \partial^2\psi/\partial y^2 \quad [2]$$

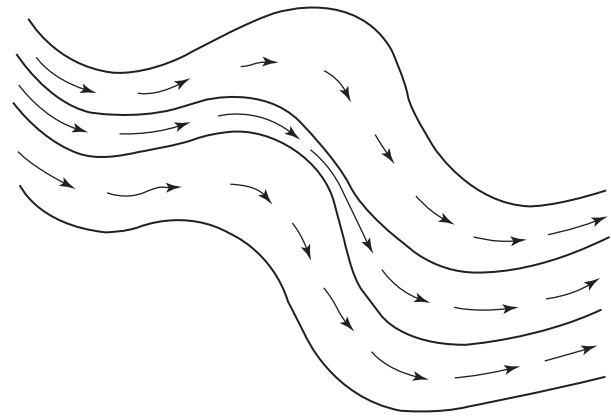


Figure 3 Contours for streamfunction (solid lines) and the associated wind field (arrows) for a two-dimensional and nondivergent flow. The arrows show the direction of flow and have lengths proportional to speed. Streamfunction values of the contours decrease going from the bottom to the top of the chart.

6. *Velocity potential*: a function (e.g., $\chi(x, y, z)$) whose gradients in space define the divergent components of motion. In atmospheric sciences where a two-dimensional formulation is commonly used, the horizontal divergence is defined from $\chi(x, y)$ by the relationships $u_\chi = -\partial\chi/\partial x$ and $v_\chi = -\partial\chi/\partial y$, where u_χ and v_χ refer to the divergent component of velocity in the x and y directions, respectively.
7. *Wind shear*: variations of the wind vector or its components in a given direction. Commonly, this term is used to refer to a lateral wind shear, i.e., a change in speed in a direction perpendicular to the wind direction (Figure 4). Thus for west-to-east wind flow defined by the u component of motion in the x direction, lateral wind shear involves variations in the y direction (north) and z direction (vertical). The presence of wind shear is a factor in many types of flow instability. Wind shears are also representative of both vorticity (rotational) and deformation flow components in the wind.

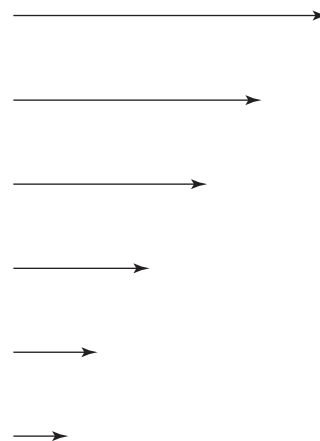


Figure 4 Example of a lateral wind shear field for straight-line wind flow. The arrows show the (uniform) direction of flow and the arrow length is proportional to the speed.

Applications

Kinematic analysis is used in a wide number of applications to describe conditions in the atmosphere. Several of these are briefly described here.

Describing Pressure Field Patterns

As mentioned earlier, kinematic analysis has been applied to the motion of patterns of pressure, temperature, humidity, etc. describing the atmospheric condition, such as seen in weather map sequences. A common depiction is made for pressure distributions, which relate directly to the horizontal wind for large-scale conditions. Pressure at a constant height (often mean sea level) is used to describe the surface weather systems and winds using lines connecting regions with the same pressure (called isobars). At higher levels in the atmosphere, the horizontal pressure variations are commonly represented by variations of the height where a given pressure is found. In this case, lines of constant height on a surface of fixed pressure (called height contours) are used rather than lines of constant pressure (called isobars) on a surface of fixed height to define the horizontal pressure variations. In the discussion below, these terms are used interchangeably.

The motions of the surface pressure isobars and height contours themselves are due to both movement and amplitude changes of weather systems. Such motion is witnessed in the propagation of high- and low-pressure system centers and their associated isobar patterns in a sequence of daily weather map

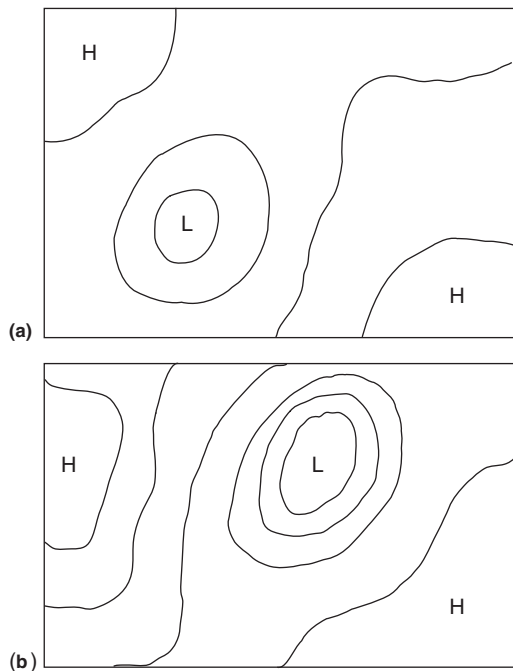


Figure 5 Surface pressure isobars (solid lines) for large-scale weather systems for two consecutive times demonstrating the movement and changes of pressure field patterns, a focus of atmospheric kinematics: (a) earlier time and (b) later time. The H and L denote local regions of highest and lowest pressure, respectively. The contour interval is a fixed value, so changes in the number of isobars go along with changes in pressure magnitudes at the H and L positions.

depictions (Figure 5). These motions relate to the pressure tendencies observed by barometers at fixed locations commonly used to indicate pending weather changes. Such pressure tendencies correspond directly to the motions of isobars and the spacing between isobars (magnitude of horizontal pressure gradient).

For large-scale motions, the isobars and height contours themselves have a well-defined relationship to the wind flow in extratropical latitudes according to the ‘geostrophic law’ approximation. Specifically, the horizontal flow tends to be parallel to the isobars and height contours in the direction where, in the Northern Hemisphere, the lower pressure (or height) is to the left facing in the direction of the flow and the higher pressure (or height) is to the right. (Directions are opposite in the Southern Hemisphere.) In addition, the flow speed is inversely proportional to the horizontal pressure gradient, i.e., the distance between adjacent isobars or height contours. This means that the pressure contour patterns have a direct relationship with flow streamlines so that kinematic properties of the airflow itself, such as vorticity and deformation, are suggested by the pressure contour patterns. Near the Earth’s surface frictional effects have a systematic effect on the large-scale wind flow, causing air to have a flow component across isobars from high to low-pressure. This leads to horizontal wind divergence and associated vertical motion when the isobars are curved as around low- and high-pressure centers (Figure 6).

Considerable specialized and descriptive terminology is used for features and their time changes in the large-scale horizontal pressure field patterns not only by scientists but also by weather presenters on radio and television. Commonly used terms include trough, ridge, digging trough, inverted trough, cutoff low, filling low, building ridge, etc. Some of these terms characterize pressure patterns and their changes which are quite useful for describing dynamic as well as kinematic conditions. For instance, in the situation where isobars or height contours run generally east–west with the lower pressures to the north, a ‘trough’ refers to a region of relative minimum in pressure where isobars for various pressure values are displaced southward compared to positions to the east and west. A ‘tilted trough’ means that the longitude of minimum pressure differs according to latitude, so that a line connecting positions for the location of lowest pressure at each latitude ‘tilts’ eastward or westward with latitude (Figure 7). Such tilts for the associated large-scale motion fields imply systematic correlations involving temperature, momentum, and transport processes. These correlations would be expected to result in extratropical storm system development located in the trough area (in the Northern Hemisphere) if the trough is tilted from southwest to northeast (a ‘positive’ tilt) with the reverse for a tilt from southeast to northwest (a ‘negative’ tilt).

Determining Vertical Motion

Vertical motion is a key factor for many aspects of weather, such as cloud development, precipitation, and development of weather systems. Nevertheless, as mentioned before, few direct measurements of vertical motion are available. However, a number of methods exist for estimating large-scale vertical motion. The kinematic method is based on the principle of conservation of mass which provides a direct relationship

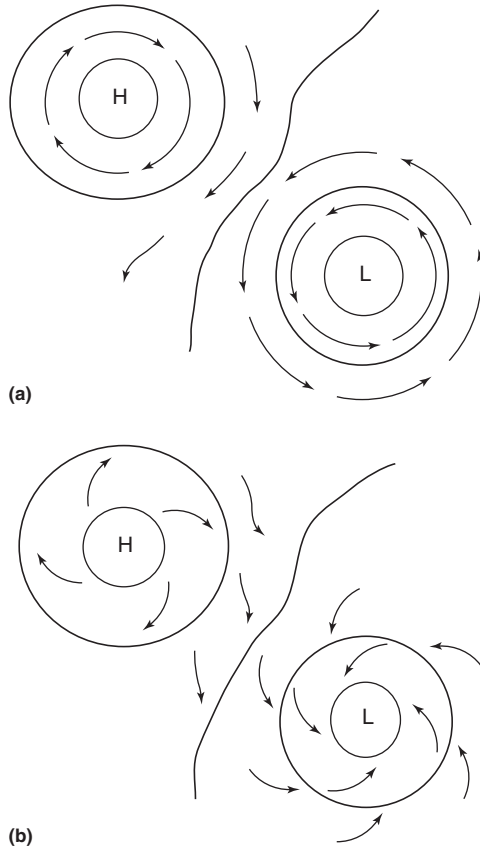


Figure 6 Large-scale horizontal wind field (arrows) in the Northern Hemisphere associated with surface pressure isobars (solid lines): (a) without and (b) with surface friction effects, respectively. The H and L denote local regions of highest and lowest pressure, respectively. Note that in Panel (b) with the curved isobars, the horizontal divergence of the wind is more evident. According to the conservation of mass relationship, vertical motion above the surface would be expected in the low- and high-pressure center areas in Panel (b).

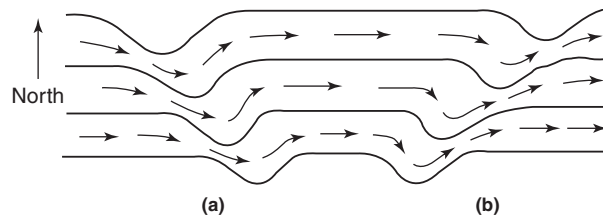


Figure 7 Examples of large-scale pressure field patterns (height contours) and associated motion flow direction (arrows) typical in the upper troposphere (Northern Hemisphere) which demonstrate ‘tilted trough’ terminology. Side A: negative tilt trough (a more westward position of the minimum pressure on a latitude line as one goes further north); and side B: positive tilt trough (a more eastward position of the minimum pressure on a latitude line as one goes further north). Heights (pressures) decrease to the north. The motions associated with the positive tilted trough would be expected to produce development of an associated extratropical cyclone.

between the horizontal divergence field and the vertical motion. Recall that horizontal divergence is one of the basic components of the horizontal motion field as discussed earlier.

The atmosphere is a compressible gas so that variations in density must be included in the relationship. For the incompressible case (e.g., like water), a very simple relationship between horizontal divergence and vertical motion exists:

$$\partial w / \partial z = -[\partial u / \partial x + \partial v / \partial y] \quad [3]$$

where w is the vertical motion, z is the height, the other variables are as defined earlier. Integration of this relationship with respect to height provides magnitudes of vertical motion at all levels as long as it is known at one level. To account for the compressibility aspects important for atmospheric weather systems, a density-weighted formula can be used:

$$\partial(\rho w) / \partial z = -[\partial(\rho u) / \partial x + \partial(\rho v) / \partial y] \quad [4]$$

where ρ is density and the variation of ρ in time is assumed to be negligible. This formulation is less easy to use because of the spatial variations in density. In some applications density variations on the right-hand side of the equation can be neglected.

An equivalent formulation of the continuity of mass relationship using pressure instead of height for the vertical coordinate system is often used for determining large-scale vertical motion. In this coordinate system, the vertical motion is represented by the variable ω representing the pressure change following an element of air. In this case the equation becomes

$$\partial \omega / \partial p = -[\partial u / \partial x + \partial v / \partial y] \quad [5]$$

where the derivatives on the right-hand side are evaluated on constant pressure surfaces.

Frontogenesis

Weather ‘fronts’ relate directly to stormy activity and rapid changes. Fronts commonly refer to regions in the large-scale setting where there are maxima in horizontal variations (gradients) for descriptors such as moisture, temperature, and winds in the atmosphere. Horizontal temperature gradients are commonly used to define the fronts and are shown on weather maps as cold, warm, and stationary fronts. Understanding the movement and changes in such fronts is important for describing, understanding, and predicting rapid changes in atmospheric conditions.

An important factor in the formation of large horizontal gradients is the kinematic deformation in the horizontal wind field. Such an increase in gradients occurs when the horizontal deformation flow in the atmosphere is oriented so that the axis of dilatation (expansion) is parallel to the contours of temperature and the axis of confluence (contraction) is perpendicular to the contours of temperature (Figure 8). This results in the contours of different temperature values being brought closer together, signifying an increase in the horizontal temperature gradient. An examination of the existing horizontal temperature distribution and the horizontal deformation distribution in the wind field or as implied from the pressure field can give considerable insight on how existing temperature gradient regions might change.

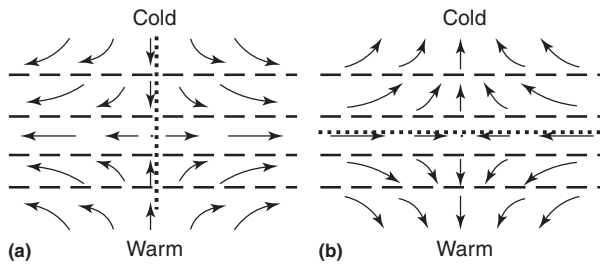


Figure 8 Horizontal flow deformation field (arrows) and temperature contour (dashed-lines) relationships where the temperature variation (gradient) would be expected to increase in time (a) and decrease in time (b) because the deformation field axis of confluence (shown by a dotted line) is perpendicular or parallel, respectively, to the temperature contours moving them closer together and further apart, respectively. Temperatures are colder at the top side of the diagrams as labeled.

Air Pollution

Air pollution arises from the transport of pollutants from source regions often at or near the Earth's surface. Sources may be obvious point sources like a smokestack or more dispersed sources like automobiles. Temperature differences between the source material and the atmosphere near the source region or density differences due to the pollutants themselves may influence how the pollution propagates. However, much of the propagation will be due to transports by the atmospheric motion. Such transports can be separated into small-scale turbulent motion effects which cause a general dispersion of the pollution over time in all directions and larger-scale motions that provide a more systematic propagation of the pollution. Kinematic analysis can provide useful information on where to expect pollution conditions, especially for the larger-scale motions, and also to determine the source regions for pollution conditions.

Flow trajectories define the path of systematic propagation by air motion. Since most pollution cases involve sources that act over sustained periods of time or even continuously, the distinction between trajectory and streamline is important for understanding the overall lateral spread of pollution from a source area. Namely, the wind flow at any one time is not sufficient information. It is the cumulative effect over time measured by the ensemble of trajectories originating in the pollution source area. The availability of computer models for atmospheric motion makes it possible to determine atmospheric trajectories routinely. It is even possible to determine the source regions of trajectories reaching a given region by looking back in time. This allows for definition of the source regions for air pollution observed in a given area.

Current Weather and Short-Range Prediction

Kinematic analysis of the wind flow is important for understanding current weather conditions and making predictions for the near future, i.e., within 12 h. Dynamical weather prediction models are particularly limited in the first 12 h for predicting clouds and precipitation since these variables are poorly represented in the initial conditions, and it takes a while for the prediction model to recreate existing fields in the

forecast. Extrapolation of initial conditions based on motion transports alone determined by kinematic analysis can provide important information on the atmospheric conditions, especially for moisture, clouds, and precipitation.

Kinematic analysis for horizontal wind flow also provides a direct measure of vertical motion through the kinematic relationship between horizontal divergence and vertical motion (discussed before). Such analysis applied on the small-scale (horizontal dimensions of 100 km or less) is used by forecasters to identify regions of upward vertical motions and thus probably more likely for development of convective precipitation systems in the near future, i.e., several hours. Such analysis is also useful for describing heavy rain situations that have occurred.

Measurement of Winds

A variety of methods are used to measure the horizontal wind flow. Some operate from fixed positions such as surface weather station anemometers and weather vanes that give measurements in a fixed (Eulerian) reference system. Others such as weather balloons, clouds tracked from satellites, and raindrops tracked with Doppler radar provide estimates for tracers presumed to be following the wind flow (a Lagrangian system reference). Cloud tracking techniques for this purpose have the additional challenge of determining the height of the cloud and sorting out situations where the clouds do not move with the horizontal wind such as for the stationary wave clouds seen around mountains. Combining such observations into a single description of the wind flow requires taking into account the distinction between trajectories and streamlines and the differing impacts of space and time averaging.

Analysis and Dynamical Process Studies

Published descriptive summaries of atmospheric circulation features include routinely many of the kinematic horizontal flow descriptors discussed here. For example, the publication *Climate Diagnostics Bulletin* published by the US National Weather Service National Centers for Environmental Prediction provides monthly summaries of many global-scale circulation features. Quantities presented include zonal wind flow (the west-to-east ' u ' component of motion), horizontal flow vectors (streamline direction indicators), and isotachs (speed) for both the lower and upper troposphere along with upper tropospheric level streamfunction, divergent wind, and velocity potential fields. The latter two quantities give an indication of large-scale vertical motion in the atmosphere through the kinematic relationships discussed before.

The kinematic descriptors also are used in studies of the dynamical processes in the atmosphere. The kinematic variables include some more complex than the basic kinematic variables described here. As examples, dynamical studies of large-scale weather systems utilize potential vorticity (a combination of rate of temperature decrease with height and vorticity measures) and Q -vectors (diagnostic variables that relate to vertical motion derived from the equations of motion and thermodynamics as well as the continuity of mass relationships). The study of vortices such as dust devils, tornadoes,

and rotating thunderstorms is facilitated by a kinematic variable called helicity. This variable combines the velocity and three-dimensional vorticity (rotational aspects) of the velocity field.

Conclusions

Atmospheric kinematics embodies the description of movement of air and its descriptors such as moisture content, pressure, and temperature with a focus on the air motion itself. The importance of kinematics has been shown by a sampling of its use in a wide range of applications. In the discussion here, the focus has been on the conceptual aspects of kinematics without dwelling on the mathematics. It should be noted that since air motion is a vector, the complete mathematics for the concepts presented here can involve complicated mathematical expressions.

See also: **Aerosols:** Observations and Measurements. **Basic Atmospheric Structure and Concepts:** Beaufort Wind Scale. **Chemistry of the Atmosphere:** Tracers. **Dynamical Meteorology:** Vorticity. **Synoptic Meteorology:** Extratropical Cyclones; Fronts; Weather Maps. **Turbulence and Mixing:** Turbulent Diffusion.

Further Reading

- Bluestein, H.B., 1992. Synoptic-dynamic meteorology in midlatitudes. In: *Principles of Kinematics and Dynamics*, vol. I. Oxford University Press, New York, NY.
- Bluestein, H.B., 1993. Synoptic-dynamic meteorology in midlatitudes. In: *Observations and Theory of Weather Systems*, vol. II. Oxford University Press, New York, NY.
- Climate Prediction Center. Climate Diagnostics Bulletin. National Centers for Environmental Prediction, National Weather Service/National Oceanic and Atmospheric Administration. Camp Springs, MD: US Department of Commerce [Published monthly since 1983].
- Eliassen, A., 1999. Vilhelm Bjerknes's early studies of atmospheric motions and their connection with the cyclone model of the Bergen school. In: Shapiro, M., Grønås, S. (Eds.), *The Life Cycles of Extratropical Cyclones*. American Meteorological Society, Boston, MA, pp. 5–13 [Chapter 1].
- Hess, S.L., 1959. *Introduction to Theoretical Meteorology*. Holt, Rinehart and Winston, New York, NY.
- Kutzbach, G., 1979. *The Thermal Theory of Cyclones: A History of Meteorological Thought in the Nineteenth Century*. American Meteorological Society, Boston, MA.
- Petterssen, S., 1956. Weather analysis and forecasting. In: *Motions and Motion Systems*, vol. I. McGraw-Hill, New York, NY.
- Petterssen, S., 1956. Weather analysis and forecasting. In: *Weather and Weather Systems*, vol. II. McGraw-Hill, New York, NY.
- Saucier, W.J., 1955. *Principles of Meteorological Analysis*. University of Chicago Press, Chicago, IL.

Laboratory Geophysical Fluid Dynamics

RL Pfeffer, Florida State University, Tallahassee, FL, USA

© 2003 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 3, pp 1081–1090, © 2003, Elsevier Ltd.

Introduction

Laboratory experiments represent one of the several approaches employed to gain insight into the processes responsible for natural phenomena. Other approaches include diagnostic studies of observational data, analytic theory, and numerical solutions of the governing equations. Each has strengths and limitations. In concert, they can provide good insight into the mechanisms involved.

Diagnostic studies are useful in the characterization of dynamical properties of observed phenomena such as convection; turbulence; the propagation, growth, and damping of waves; and the time-mean or space-mean circulation. They provide a framework to assess theories and laboratory simulations, which must exhibit essentially the same properties in order to be considered valid. They cannot, however, isolate the roles of controlling parameters (e.g., rotation and differential heating) in bringing about observed phenomena, or provide insight into how things would change if these parameters were altered.

Analytic theory, based on simplified models, is not designed to deal with the full complexity of different phenomena, but rather to determine the roles of the major controlling parameters.

Numerical simulation, using discretized approximations to the full nonlinear governing equations, can treat much more complex systems but is limited in accuracy by the fact that important subgrid-scale phenomena must be parameterized, and by the reality that the solution of the discretized (algebraic) equations departs progressively with time from that of the governing differential equations.

Laboratory experiments, under carefully controlled and reproducible conditions, can be used to verify predictions based on analytic theory and numerical simulation and to discover phenomena that require new theory to explain them. In order to gain insight, it is not desirable to deal with the full complexity of each phenomenon. Moreover, full similarity cannot generally be achieved.

In this article, we select for discussion a few of the many types of experiments that have been employed to understand phenomena in the Earth's atmosphere, oceans, and fluid interior. Examples are chosen that illustrate the interplay of diagnostics, theory, and experiment, and for which regime diagrams have been constructed that map the fluid behavior over a broad range of parameters.

Similarity

In developing a laboratory prototype, one strives to attain similarity in as many ways as possible. An exact model would

be one that had the same geometry as the system to be modeled, but on a smaller scale, and the same governing equations, with the same relative magnitudes of the different terms in these equations. The relative magnitudes are measured by dimensionless numbers, which are ratios of different terms in the governing equations. For example, the ratio of the relative acceleration of a fluid parcel to the Coriolis acceleration in a rotating system leads to the definition of the Rossby number $Ro \equiv |(d\mathbf{V}/dt)|/|2\boldsymbol{\Omega} \times \mathbf{V}|$, where $\mathbf{V}$ is the fluid velocity, t is the time, and $\boldsymbol{\Omega}$ is the angular velocity of the system. If the acceleration is dominated by advection $\mathbf{V} \cdot \nabla \mathbf{V}$, and if V is a characteristic speed of the flow and L a characteristic length scale, the Rossby number can be expressed as $Ro = V/fL$, where f is the Coriolis parameter.

Some dimensionless numbers are arrived at after combining all or several of the governing equations (usually in linearized form) to obtain equations that reveal the relative influences of thermodynamic and mechanical processes. One example is the Rayleigh number $Ra \equiv g\alpha(-\partial T^*/\partial z)H^4/\nu\kappa$, where $-\partial T^*/\partial z$ is the excess of the undisturbed vertical temperature lapse rate over the adiabatic lapse rate in a characteristic depth H of fluid, z is the height, g is the acceleration of gravity, and ν , κ , and α are, respectively, the kinematic viscosity, thermometric conductivity, and coefficient of volume expansion of the fluid due to temperature T . The Rayleigh number measures the ratio of the destabilizing effect of buoyancy to the stabilizing effect of viscous and thermal diffusion.

In geophysical problems, it is usually impossible to construct laboratory analogs with the same geometry as the systems to be modeled. Although some experiments have been done in space vehicles in zero gravity, using electromagnetic forces to simulate gravity normal to a sphere, laboratory experiments on the Earth typically replace the Earth's spherical geometry with cylindrical geometry, in order that gravity be normal to the bottom of the fluid. Moreover, it is not generally possible to achieve the same balances among all terms in the governing equations in the model as in nature. Experimentalists attempt, instead, to model only the balances that are essential to the problem and to explore how the flow changes as the most important parameters are varied.

Vertical Convection between Horizontal Planes (Rayleigh–Bénard Convection)

Convection due to heating at the bottom and cooling at the top of a fluid layer is ubiquitous in the Earth's atmosphere, oceans, and interior. It is also found in the Sun, in other stars, and in the atmospheres and interiors of other planets. The simplest experiment illustrating this phenomenon is accomplished by maintaining the bottom boundary of a liquid at a higher

temperature than the top boundary. In the absence of dissipation, the resulting density increase with height would clearly be unstable, leading to a massive overturning until the more dense fluid is found below the less dense fluid. In the presence of heat conduction and viscosity, however, there is a critical value of the Rayleigh number Ra_c below which convective overturning cannot occur. This value depends on the details of the boundary conditions and the heating arrangement (e.g., constant temperatures or constant heat fluxes at the boundaries), but is always of the order 10^3 . Henri Bénard, who first performed such experiments in 1900, found that, when convection set in, it took the form of polygonal cells (which ultimately became regular hexagons). Later experiments showed that the preferred plan form at instability with most imposed conditions is steady two-dimensional rolls. Today, it is generally accepted that Bénard's cells were produced by surface tension gradients due to temperature variations on the free upper surfaces of very shallow fluids.

At successively larger Rayleigh numbers, the plan form of the convection changes, the details depending on Prandtl number ($Pr = \nu/\kappa$). At low Pr , the transition is from steady rolls to three-dimensional periodic oscillations. At high Pr , there is an intermediate range of Ra in which steady three-dimensional cells are observed. At sufficiently large Ra , the periodic oscillations give way to turbulent convection.

Figure 1 shows patterns and the associated circulation that can arise in cellular convection. Figure 2 is a regime diagram covering a broad range of Ra and Pr . Below Ra_c the vertical heat flux $\mathcal{H}$ is accomplished by heat conduction. With increasing Ra above this transition, $\mathcal{H}$ (accomplished primarily by convection) increases with Ra at fixed Pr , with jumps in the rate of increase at each successive transition. At high Rayleigh numbers, within the turbulence regime, convective plumes tilt with height and transfer momentum between the lower and upper portions of the fluid by means of Reynolds stresses, thereby creating and maintaining a steady large-scale circulation of the fluid. Incorporation of this mechanism into the cloud parameterization in one global numerical prediction model led to the correction of systematic errors in the predicted tropical wind field. At very high Ra , the plumes form clusters that exhibit systematic horizontal drift.

Much theoretical work has been done on the convection problem, beginning with the linear theory of Lord Rayleigh in 1916. The theory establishes that the only parameter controlling the onset of convection is Ra , and that the cell diameters at the onset of convection, which are controlled by the buoyancy and dissipative scales, are about twice the depth of the convecting layer. Laboratory experiments confirm the predicted values of Ra_c and the cell size.

Weakly nonlinear theory has been employed to investigate the stability of each of the possible plan forms allowed by the linear theory. Consistent with the results of laboratory experiments, the solutions show that only two-dimensional rolls are stable to small disturbances just above the critical Rayleigh number Ra_c in the absence of vertical asymmetries. Hexagonal cells become the preferred plan form in the presence of such asymmetries.

It is known from both theory and experiment that if horizontal flow is superimposed, the convection becomes aligned in rolls in the direction of the shear. If rotation is

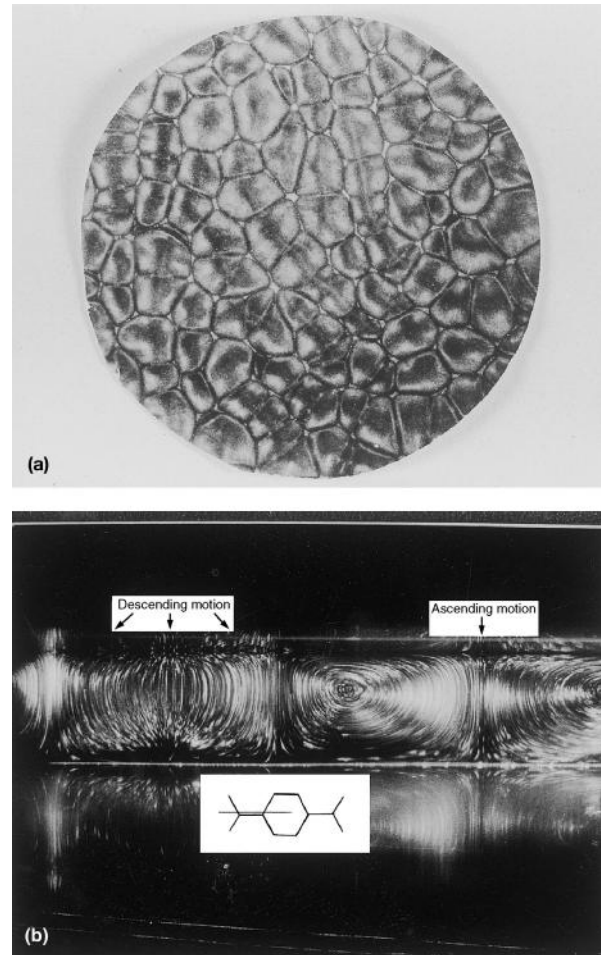


Figure 1 (a) Plan view of convection cells generated by lowering a cylinder containing a high Prandtl number fluid into a larger container of warm fluid. The irregular polygonal plan form is due to the changing mean temperature of the convecting fluid. (b) Vertical cross section depicting broad-scale downward motion along the cell boundaries and upward and outward motions in the interiors of the cells. Experiment by the author, Geophysical Fluid Dynamics Institute, Florida State University.

superimposed, the horizontal dimensions of the convection cells shrink.

In the Earth's atmosphere, cloud patterns give evidence of turbulent convection, clusters, two-dimensional roll convection, and three-dimensional cellular convection under different conditions. Behind maritime cold fronts, convection has been observed in the form of both open and closed cells (Figure 3), the former having upward motion at the cell boundaries and downward motion in the centers, and the latter the reverse circulation. These observations have sparked further experiments and theory that, although highly idealized, seem to have relevance to the observed cloud patterns. Weakly nonlinear theory predicts that three-dimensional open and closed hexagons, rather than rolls, are stable when the mean vertical motion is downward or upward, respectively. Laboratory experiments confirm the existence of these plan forms under the same circumstances. The theory and experiments do not, however, exhibit the aspect ratio observed in the atmosphere,

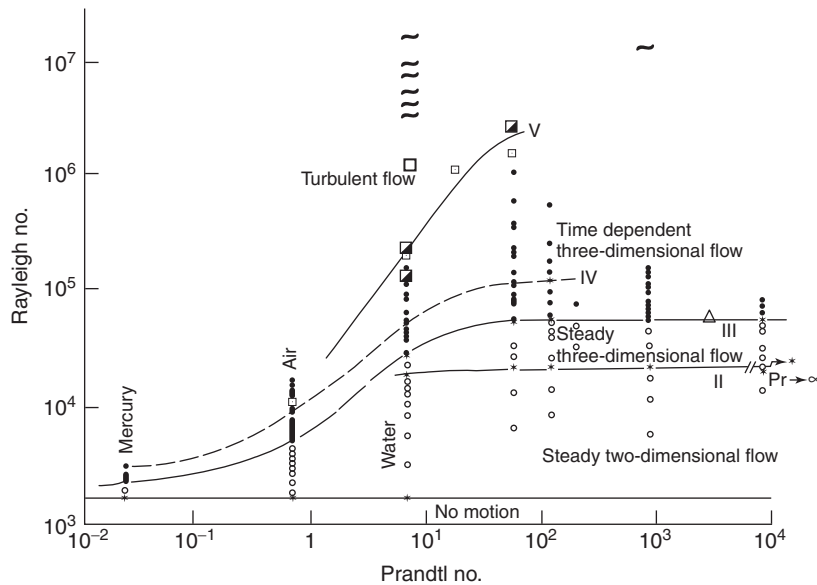


Figure 2 Regime diagram depicting the dependence of plan form and time dependence on Ra and Pr when the top and bottom boundaries are rigid and maintained at different temperatures. Symbols: $\circ$ steady cellular flow; $\bullet$, time-dependent cellular flow; $\blacksquare$, cells and transient bubbles (transitional); $\square$, transient bubbles; $\sim$, large-scale flow with tilted plumes; and $*$, heat flux transitions. The dashed curve represents the transition from one to two hot spots circulating in each convection cell. Reproduced from Krishnamurti, R., Howard, L., 1981. Large-scale flow generation in turbulent convection. *Proceedings of the National Academy of Sciences of the USA* 78: 1981–1985.

where the horizontal scales are from 10 to 50 times the depth of the convecting layer. Clouds are formed by the release of latent heat of condensation, a complication not dealt with in Rayleigh–Bénard convection. More recent experiments

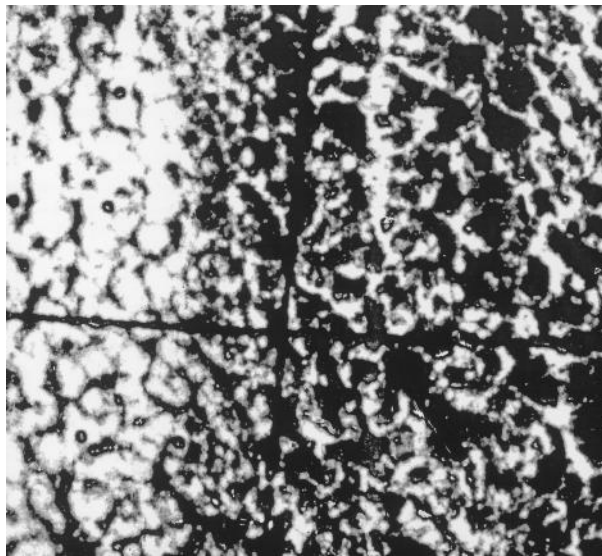


Figure 3 Section of Air Force DAPP satellite photograph (22.30 GMT 4 April 1973) of convective clouds, centered near 30° N and 61° W, showing open cells in a region of large-scale downward motion and closed cells in a region of large-scale upward motion. The region is roughly 500 km on a side. Reproduced from Shaugnessy, J.E., Wann, T.C., 1973. Picture of the month-frontal rope in North Pacific. *Monthly Weather Review* 101: 774–77. Reproduced from Krishnamurti, R., 1975. On cellular cloud patterns, part 3: Applicability of laboratory and mathematical models. *Journal of the Atmospheric Sciences* 32: 1373–1383.

demonstrate that large cell sizes can be realized by simulating the release of latent heat of condensation in clouds that form in an otherwise stably stratified atmosphere.

In the ocean, the water density is controlled by both temperature and salinity, and thermal diffusion is much more rapid than salt diffusion. Convection under these circumstances is controlled by four dimensionless parameters, Ra , Pr (both defined earlier), the ratio of the diffusivities of salt and temperature $\tau = \kappa_s/\kappa (<1)$, and the salinity Rayleigh number $Ra_s = g\gamma(\partial S/\partial z)H^4/\nu\kappa$. Here γ is the coefficient of volume expansion of the fluid due to salt and S is the salt concentration. When $\partial S/\partial z < 0$ and $\partial T/\partial z > 0$, the density gradients due to salt and temperature are both stably stratified, and so no instability is possible. When $\partial S/\partial z > 0$ and $\partial T/\partial z < 0$, both gradients are unstably stratified, and convective instability is observed. When either $\partial S/\partial z > 0$ and $\partial T/\partial z > 0$, or $\partial S/\partial z < 0$ and $\partial T/\partial z < 0$, double diffusive instabilities are possible. Such instabilities occur when the net density stratification is stable, but either the temperature is stably stratified and the salt is unstably stratified, or vice versa.

Theory and experiment confirm that when warm, salty water lies over cold, freshwater, with a net stable density stratification, thin ‘salt fingers’ develop for sufficiently large positive Ra_s (unstable) and small negative Ra (stable), with rising fingers of cold, freshwater and sinking fingers of warm, salty water. For very small negative Ra , the salt fingers become unstable and, below a comparatively thin layer, are replaced by a well-mixed convection layer. If the fluid is sufficiently deep, another salt finger layer forms below the well-mixed layer. Within a range of values of the controlling parameters, reflecting, in particular, small negative Ra and deepwater, the instability takes the form of a number of convecting layers bounded above and below by comparatively thin salt finger

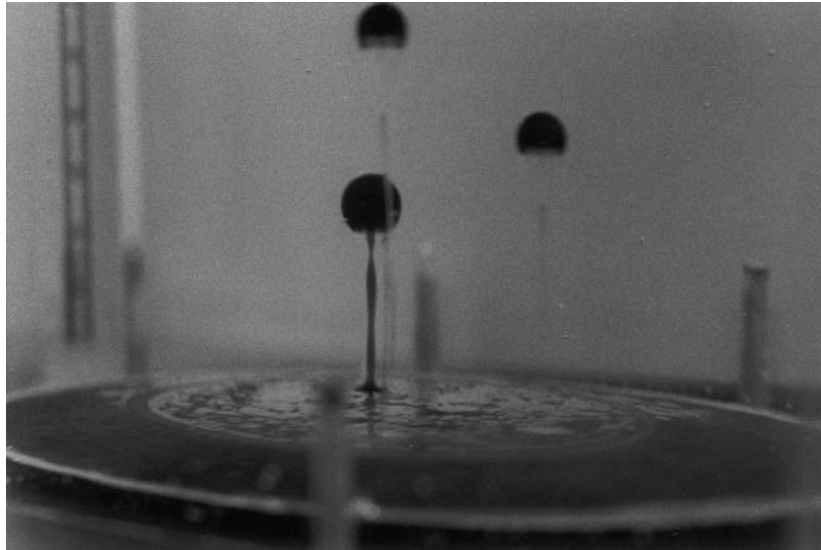


Figure 4 Plume convection produced by maintaining a reservoir of low-density fluid (water) below a more dense, highly viscous fluid (corn syrup). Seen here are two thermals and one ‘starting plume.’ The thermals entrain more dense fluid and decelerate. The plume entrains buoyant fluid from the thin trailing filament leading back to the source and accelerates upward. Courtesy of Loper, D., Geophysical Fluid Dynamics Institute, Florida State University.

layers. Evidence of such phenomena has been found in the ocean (e.g., in the Mediterranean outflow).

When cold, freshwater lies above warm, salty water, with a net stable density stratification, theory and experiment show that oscillatory instability takes place for sufficiently large positive Ra and sufficiently small negative Ra_s . Once again, layering occurs in sufficiently deep fluids, with well-mixed convective layers bounded above and below by stably stratified diffusive layers. Evidence of this phenomenon has been found below an Arctic ice island, as well as in Antarctic lake Vanda and in other lakes, in which warm seawater from old intrusions is found below cold, fresh lake water.

In the Earth’s mantle, convection is shaped by the temperature dependence of the viscosity of the convecting material. The mantle behaves like an extremely viscous, low thermal conductivity fluid. Internal heating, accompanied by cooling at the surface, creates global-scale convection, resulting in the motions of continents. Moreover, the core heats the lowest layer of the mantle, providing buoyancy and lowering the viscosity locally. Evidence suggests that convection from this source, under very high Rayleigh numbers, takes the form of narrow plumes which, upon reaching the surface, form volcanic chain islands (e.g., the Hawaiian islands), among other features. An example of plume convection modeled in the laboratory is shown in Figure 4.

In the Earth’s fluid core, rotation and electromagnetic (Lorentz) forces play important roles, and the four dimensionless parameters needed to study convection under such influences are the Rayleigh number, the Prandtl number, the Taylor number (the square of the ratio of the Coriolis to the viscous force), and the Elsasser number (the ratio of the magnetic body force to the Coriolis force). Rotation renders the motions nearly two-dimensional in planes normal to the angular velocity Ω , which is not parallel to gravity, and tends to make the cells narrower than they would be in its absence. Electromagnetic forces tend to make the cells wider. Buoyancy

is created in part by thermal effects and in part by compositional changes (solidification of iron and nickel, leaving a buoyant residual liquid at the boundary with the inner core). Convection takes place in a geometry of variable depth owing to the alignment of the convection cells within the spherical annulus. Modeling of all these effects in one laboratory experiment would be very complicated and is probably not possible. Separate experiments modeling thermal convection in a self-gravitating sphere or annulus, thermal convection under the simultaneous influence of rotation and a magnetic field, and compositional convection show qualitative agreement with theory and numerical simulation, but much of this work has not approached the parameter range of the core.

Sloping Convection in Rotating, Laterally Heated Fluids

Pioneering experiments by Dave Fultz and Raymond Hide beginning around 1950 demonstrated that it is possible to simulate basic features of the general atmospheric circulation, including the jet stream and the growth and propagation of midlatitude, synoptic-scale ($L \sim 1000$ km) baroclinic waves. In these experiments, a liquid contained in an upright cylinder or cylindrical annulus is rotated about the axis of the cylinder, heated at the outer wall and cooled either at the center or at the inner wall. Differential heating establishes a radial temperature gradient, analogous to the meridional temperature gradient in the atmosphere. Overturnings create a stable vertical density stratification with Brunt–Vaisala (buoyancy) frequency N . Two main dimensionless parameters control the instability and the scale of the baroclinic waves: the Burger number $B \equiv N^2 H^2 / f^2 L^2$ (which measures the competition between stratification and rotation in determining the baroclinic wave scale) and the Taylor number $Ta \equiv |2\Omega \times \mathbf{V}|^2 / \nu \nabla^2 \mathbf{V}|^2$. Here, f is the Coriolis parameter and $N^2 \equiv g\alpha/\partial T \partial z > 0$. The growth rate of the disturbances is proportional to the ratio of the destabilizing

effect of vertical shear to the stabilizing effect of stratification $|\partial V/\partial z|/N$, the square of which is the inverse of the Richardson number. While fluid motions in the cylinder (or 'dishpan') experiments look more like those seen on upper tropospheric weather maps, the more regular wave patterns in the annulus experiments possess the same essential ingredients and are easier to interpret.

Figure 5 shows the streamlines, isotherms, and vertical velocity contours corresponding to a four-wave pattern in an annulus experiment. The relationships among these variables are similar to those in midlatitude developing baroclinic waves and confirm the predictions of analytic theory. Baroclinic growth is achieved by the rising of warm fluid downstream of the troughs and the sinking of cold fluid downstream of the ridges in the stream field. Away from the boundaries, the motions are quasigeostrophic and hydrostatic, implying thermal wind balance and the existence of a jet stream above the strongest temperature gradients.

The slope of the azimuthal mean potential temperature surfaces in the radial plane is such that the coldest fluid is found at the bottom near the cold source and the warmest fluid at the top near the heat source. A parcel stability argument reveals that instability can take place only if fluid parcels are displaced within the angle between the geopotentials and the potential temperature surfaces (called the Eady angle, after the scientist who elucidated this in his 1949 theory of baroclinic instability). According to linear theory, this can take place only at horizontal scales larger than a critical value that depends on

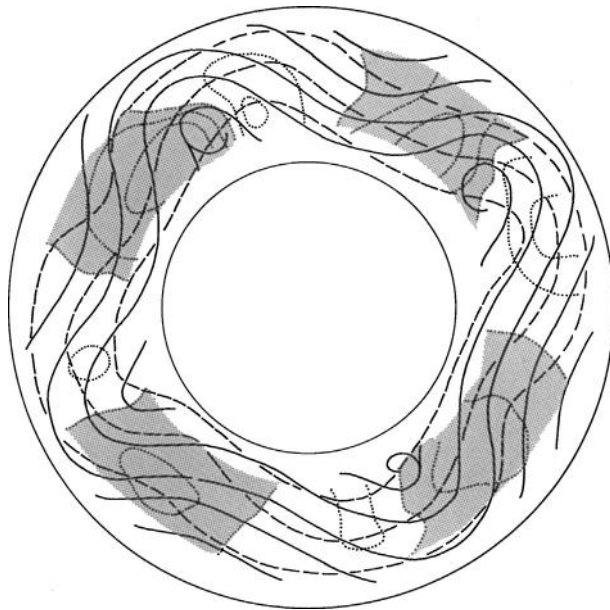


Figure 5 Plan view of streamlines (solid curves), isotherms (dashed curves), and vertical velocity contours (dotted curves with shading for upward motion) at mid-depth in an annulus experiment. The inner and outer cylinders are held at uniformly cold and warm temperatures, respectively. The annulus is rotated counterclockwise. The pattern travels counterclockwise relative to the annulus. A meandering jet stream, following the most closely spaced streamlines, flows counterclockwise through the pattern. Reproduced from Pfeffer, R.L., Buzyna, G., Fowles, W.W., 1974. Synoptic features and energetics of wave amplitude vacillation. *Journal of the Atmospheric Sciences* 31: 622–645.

B . Shorter waves have parcel motions outside the Eady angle. The fastest growing waves are those for which the parcel motions follow the bisector of this angle. The scale of these waves is $L \sim \lambda = NH/f$ (called the Rossby radius of deformation), corresponding to $B \sim 1$. Longer and shorter waves have slower growth rates because their parcel paths have steeper or less steep slopes.

Figure 6 is a sketch of the regime diagram showing the various wave scales and forms of time dependence observed in annulus experiments. The use of the thermal Rossby number $Ro_T \equiv g\alpha H \Delta T / 4\Omega^2 \Delta R$ (in which the annular gap width ΔR is taken as the length scale and the imposed thermal wind as the velocity scale) reflects the fact that the imposed temperature contrast ΔT controls both the Burger and Richardson numbers.

Outside the curve containing the baroclinic wave regime, the motions are axisymmetric, resembling a Hadley cell with rising and sinking motions near the warm and cold walls, respectively, and fluid spiraling inward in the upper layer and outward in the lower layer. The cutoff of wave instability at large Ro_T is due to the fact that the preferred wave scale for instability is larger than that which can fit within the annulus. The cutoffs at small Ta and Ro_T are due to the dominance of dissipation for small wave scales and/or rotation rates. Within the regular wave regime, the observed wave scales correspond to the predicted value λ , becoming smaller as ΔT is decreased and/or Ω is increased. At lower ΔT (i.e., smaller N^2) and higher Ω , where nonlinearity becomes important in the more irregular flow, the observed scales are generally larger than predicted. Two kinds of time dependence are observed in these experiments: amplitude and structural vacillation.

Amplitude vacillation is a nearly periodic growth and decay of waves (Figure 7(a)–(d)). Axisymmetric overturnings near

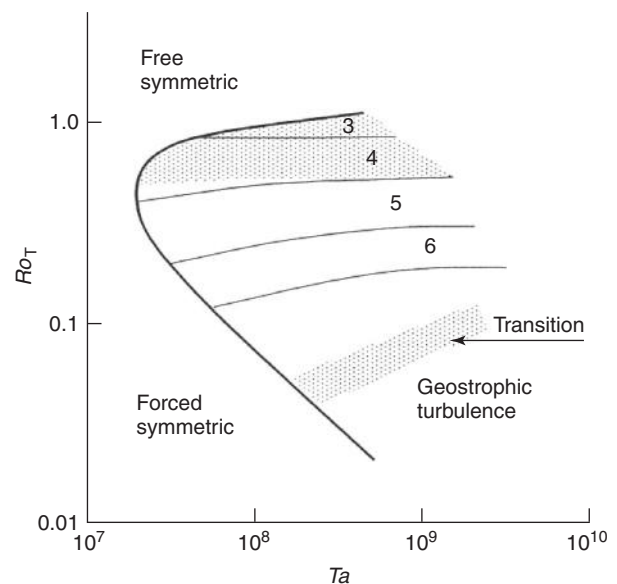


Figure 6 Sketch of the regime diagram for baroclinic flow in an annulus with gap width twice the fluid depth, depicting the most probable wave scales and time dependence as a function of Ro_T and Ta at $Pr = 21$. The upper shaded region shows where amplitude vacillation is most prevalent. The unshaded region immediately below it indicates where structural vacillation is most common. The details differ with aspect ratio and Prandtl number.

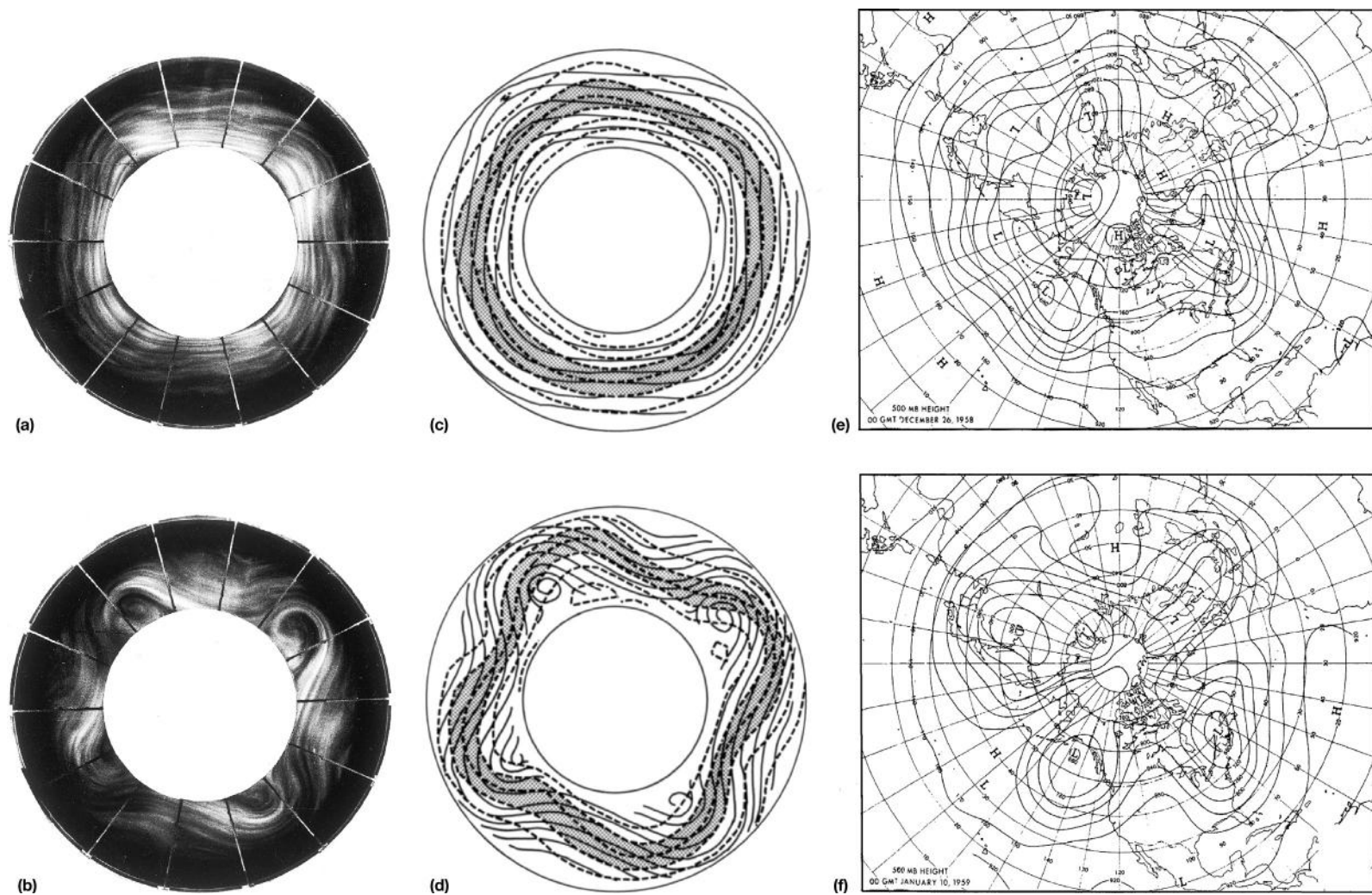


Figure 7 (a,b) Upper-level wave patterns and (c,d) mid-depth temperature fields in a thermally driven rotating laboratory experiment, depicting minimum and maximum wave amplitudes at opposite phases of an amplitude vacillation cycle. (e,f) 500 hPa height charts depicting opposite phases of an atmospheric 'index cycle.' Reproduced from Pfeffer, R.L., Chang, Y., 1967. Two kinds of vacillation in rotating laboratory experiments. *Monthly Weather Review* 95: 75–82.

the side walls set up large temperature gradients across the middle of the annular gap. Baroclinic instability sets in, leading to explosive wave growth. The waves transport heat across the annular gap and induce a thermally indirect Ferrel cell between the thermally direct meridional cells near the walls. The heat transport is so great that it reduces the horizontal temperature gradient below the critical value needed to overcome dissipative effects, and the waves die out. When the waves become sufficiently weak, axisymmetric overturnings reestablish the radial temperature gradient and the cycle begins anew. Although not periodic, striking cases of a similar process of hemispheric baroclinic wave growth and decay in middle latitudes in winter have been observed in the Earth's atmosphere (Figure 7(e) and (f)).

Structural vacillation is characterized by time-dependent changes in the wave shapes with negligible changes in wave amplitude. As Ro_T decreased and Ta is increased, this form of vacillation first appears as a weak, almost periodic, oscillation. In some experiments, it manifests itself as a periodic tilting of the troughs and ridges such that they transport angular momentum first inward and then outward (Figure 8). In others it manifests itself as an energy oscillation in the radial direction. With further decreases of Ro_T and increases of Ta , the oscillations become modulated and intermittent in time, finally giving way to fully developed geostrophic turbulence. There is good evidence to suggest that structural vacillation represents a form of low-order deterministic chaotic behavior, with the order increasing as one moves toward the region of geostrophic turbulence.

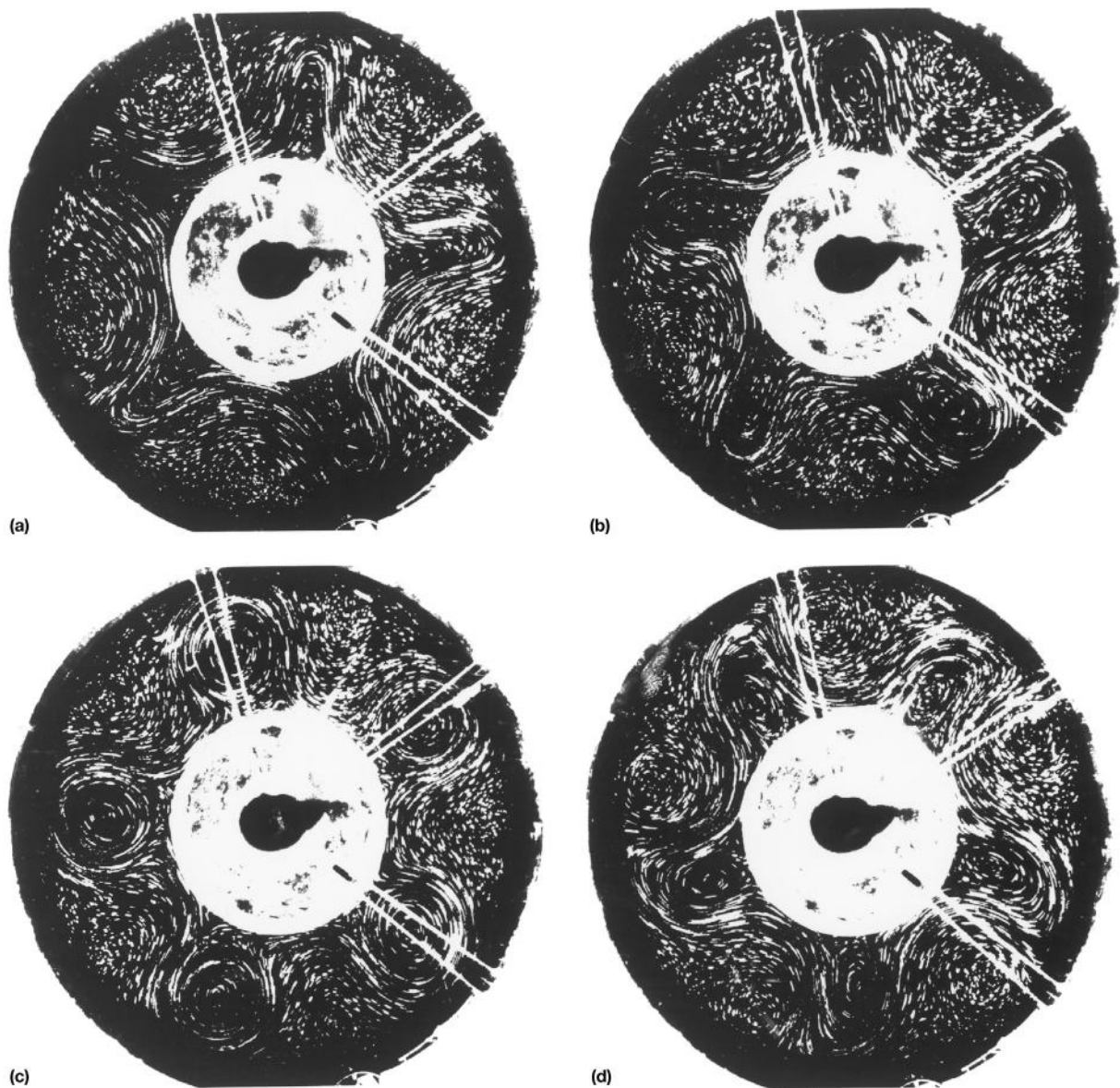


Figure 8 A sequence of flow patterns depicting tilted trough vacillation in a thermally driven rotating annulus of fluid. Courtesy of Professor Fultz, D., University of Chicago.

In addition to the two forms of vacillation, wave dispersion has been observed near the boundaries between wave numbers. In such experiments, two, and sometimes three, adjacent wave numbers are observed, with phase speeds that increase with decreasing scale. It has been speculated that this phenomenon is associated with the slope of the free surface due to rotation, which can act in part like the variation of the Coriolis parameter in the atmosphere, although the wave speed varies linearly with scale rather than quadratically as in the formula for Rossby waves in the atmosphere.

Variations of the annulus experiment have been used to study other influences that affect atmospheric behavior. Among these are bottom topography, land–sea temperature variations, cyclic variations of the imposed temperature contrast (simulating seasonal forcing), and the effect of internal heating in the atmospheres of the major planets on the formation and maintenance of large detached vortices (e.g., Jupiter’s Great Red Spot and related features on Saturn and Neptune). Additional studies of baroclinic fluid behavior have been performed with two homogeneous, incompressible fluids driven by a differentially rotating lid. Here the baroclinicity is concentrated at the interface between the two fluids. Other experiments designed to account for detached vortices have been performed with mechanically driven, rotating barotropic fluids at high Reynolds numbers (the ratio of advection to viscous forces), high Taylor numbers, and low Rossby numbers. Such experiments have also been used to study the properties and stability of the Earth’s stratospheric polar night jet.

One important property of synoptic-scale waves in the atmosphere that has not yet been successfully modeled is the meridional and vertical propagation of such waves, made possible by the presence of a background potential vorticity gradient, the absence of lateral walls, and the unlimited upward extent of the atmosphere. Unlike atmospheric waves, annulus waves are trapped radially and vertically. Innovative experiments attempting to address some of these issues have not yet borne fruit.

Other Experiments

Numerous other geophysical fluid dynamics experiments have been conducted to illustrate fundamental theorems or explain observed phenomena: for example, the Taylor–Proudman theorem (which states that rapid rotation characterized by small Ro and large Ta makes the flow invariant in the direction of Ω); Kelvin–Helmholtz instability (instability generated by a sufficiently large velocity shear across the interface between a lower density fluid and an overlying higher density fluid); axisymmetric inertial instability (the instability of a circular flow when the square of the angular momentum decreases outward and Ta exceeds a critical value: Figure 9); various routes to chaos as a control parameter, such as rotation, are increased; and critical layer absorption (the absorption of neutral linear waves near a surface along which their phase speeds in the direction of the background flow equal the background flow speed). Kelvin–Helmholtz instability is responsible for the formation of billow clouds and wind-driven water waves. Critical

layer absorption plays a crucial role in the quasibiennial oscillation in the equatorial stratosphere.

Other experiments model phenomena such as topographic waves in stratified fluids, penetrative convection in the atmosphere (where tropospheric convection penetrates into the stably stratified stratosphere), and the strong boundary currents found along the western shores of most oceans (e.g., the Gulf Stream, the Kuroshio Current, and the Brazil Current).

Penetrative convection has been modeled by making use of the property that the maximum density of water occurs around 4°C. If the bottom of a container of water is maintained at 0°C, by placing it in contact with a block of ice, and the top is maintained at a temperature well above 4°C, the layer below the 4°C isotherm will have an unstable density stratification, which will lead to convection, and the layer above will have a stable density stratification.

Western boundary currents have been modeled in a rotating cylinder of fluid in which the base is a sloping plane that creates variable depth. It is readily shown that, in an unstratified fluid, the variable depth has the same effect on the flow as the latitudinal variation of the Coriolis parameter has on the

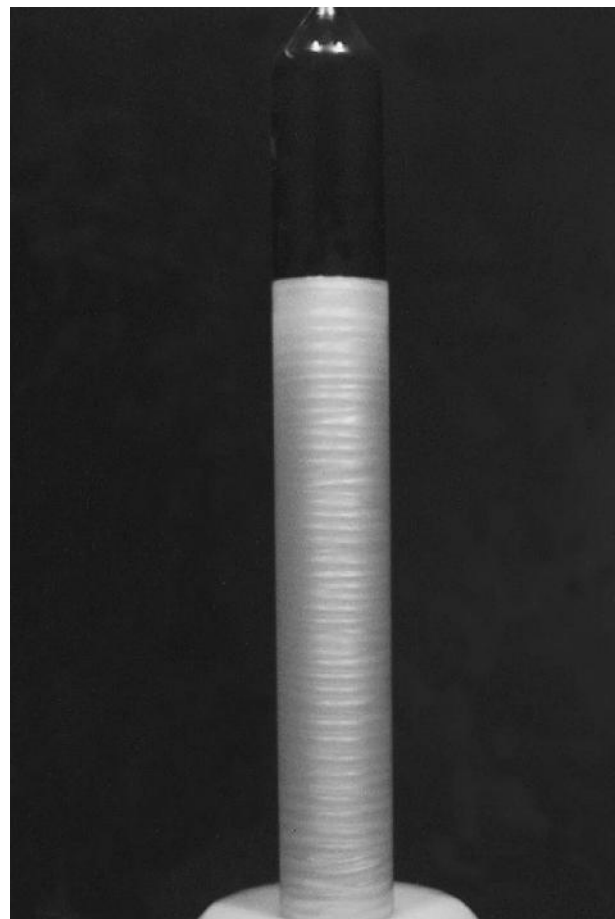


Figure 9 Taylor vortices in a tall cylinder of fluid produced by bringing the fluid into solid rotation and then stopping the rotation. The fluid near the wall slows down while that away from the wall continues to spin, creating an unstable velocity profile. Courtesy of Kung, R., and the author, Geophysical Fluid Dynamics Institute, Florida State University.

atmosphere and the ocean. Experiments and theory confirm that this is the crucial ingredient in the formation of western boundary currents.

Although the present article has by no means covered all types of geophysical fluid dynamics experiments that have been performed, it should give the reader sufficient familiarity with the type of research that is done and the factors that are considered when designing such experiments.

See also: **Data Assimilation and Predictability:** Predictability and Chaos. **Dynamical Meteorology:** Baroclinic Instability; Inertial Instability; Kelvin–Helmholtz Instability; Overview; Quasigeostrophic Theory; Symmetric Stability; Wave-CISK. **Mesoscale Meteorology:** Waterspouts. **Middle Atmosphere:** Quasi-Biennial Oscillation. **Oceanographic Topics:** Surface/Wind Driven Circulation. **Synoptic Meteorology:** Cyclogenesis; Jet Streaks.

Further Reading

- Busse, F.H., Hartung, G., Jaletzky, M., Sommermann, G., 1997. Experiments on thermal convections in rotating systems motivated by planetary problems. In: Navon, I.M., Kalnay, E., Stone, P.H. (Eds.), *Dynamics of Atmospheres and Oceans*. Special issue vol. 27 (3), 161–174.
- Chandrasekhar, S., 1961. *Hydrodynamic and Hydromagnetic Stability*. Oxford University Press, London, UK.
- Cushman-Roisin, B., 1994. *Introduction to Geophysical Fluid Dynamics*. Prentice-Hall, Englewood Cliffs, NJ.
- Drazin, P.G., Reid, W.H., 1981. *Hydrodynamic Stability*. Cambridge University Press, Cambridge, UK.
- Fultz, D., 1962. An experimental view of some atmospheric and oceanic behavioral problems. *Transactions of the New York Academy of Sciences, Series II* 24 (4), 421–446.
- Greenspan, H.P., 1968. *The Theory of Rotating Fluids*. Cambridge University Press, Cambridge, UK.
- Hart, J.E., Adler, B., Leben, R., 1997. Cyclonic/anticyclonic gyre asymmetries: laboratory and intermediate-model experiments. In: Navon, I.M., Kalnay, E., Stone, P.H. (Eds.), *Dynamics of Atmospheres and Oceans*. Special Issue 27 (3), 219–232.
- Hide, R., 1997. On the effects of rotation on fluid motions in cylindrical containers of various shapes and topographic characteristics. In: Navon, I.M., Kalnay, E., Stone, P.H. (Eds.), *Dynamics of Atmospheres and Oceans*. Special issue 27 (3), 243–256.
- Krishnamurti, R., 1997. Convection induced by selective absorption of radiation. A laboratory model of conditional instability. In: Navon, I.M., Kalnay, E., Stone, P.H. (Eds.), *Dynamics of Atmospheres and Oceans*. Special issue 27 (3), 367–382.
- Loper, D.E., 1997. Mantle plumes and their effect on the Earth's surface: a review and synthesis. In: Navon, I.M., Kalnay, E., Stone, P.H. (Eds.), *Dynamics of Atmospheres and Oceans*. Special Issue 27 (3), 35–54.
- Pfeffer, R.L., Kung, R., Applequist, S., Long, C., Buzyna, G., 1998. Progress in characterizing the route to geostrophic turbulence and redesigning thermally-driven rotating annulus experiments. *Theoretical and Computational Fluid Dynamics* 9, 253–257.
- Read, P.L., 1993. Coherent baroclinic waves in a rotating, stably-stratified fluid and transitions to disordered flow. In: Mobbs, S.D., King, J.C. (Eds.), *Waves and Turbulence in Stably Stratified Flows*. Clarendon Press, Oxford, UK.
- Sommeria, J., Meyers, S.D., Swinney, H.L., 1991. Experiments on vortices and Rossby waves in eastward and westward jets. In: Osborne, A.R. (Ed.), *Nonlinear Topics in Ocean Physics*. North-Holland, Amsterdam, pp. 227–269.
- Stern, M.E., 1975. *Ocean Circulation Physics*. Academic Press, New York, NY.
- Stern, M.E., Radko, T., 1997. Maintaining the inshore shear of continental boundary currents. In: Navon, I.M., Kalnay, E., Stone, P.H. (Eds.), *Dynamics of Atmospheres and Oceans*. Special Issue 27 (3), 662–678.
- Swinney, H.L., Gollub, J.P. (Eds.), 1981. *Hydrodynamic Instabilities and the Transition to Turbulence*. Springer-Verlag, Berlin.
- Turner, J.S., 1973. *Buoyancy Effects in Fluids*. Cambridge University Press, Cambridge, UK.

Synopsis

The Eulerian and Lagrangian formulations of fluid mechanics are reviewed, and applications to atmosphere–ocean dynamics are presented. Numerical schemes based on the Lagrangian formulation are discussed.

Introduction

In meteorology and oceanography, we regard our fluid as a continuum – a continuous distribution of mass in space. In so doing, the atomic or molecular nature of the fluid is neglected and this implies that any small volume element is always supposed to be sufficiently large so that it still contains a huge number of molecules. Accordingly, when we speak of an infinitesimal volume element – or a fluid ‘parcel’ – we mean that it is small compared with the volume of the system but still sufficiently large to contain very many molecules.

The adoption of the continuum description is the first step toward specifying how we solve the fundamental problem of the science of kinematics for our fluid; i.e., the specification of suitable methods for describing and analyzing fluid motions. The problem of kinematics, for any given physical system, should be distinguished from that of dynamics, which is concerned with determining the state and/or motion of the system at any instant. It is a common practice to refer to the combined science of kinematics and dynamics as mechanics.

There are two common descriptions of continuum motion (both due originally to Leonhard Euler (1707–83)). In the first method, known as the Lagrangian method, we fix our attention directly on the fluid parcels and study their motion through space. Unless otherwise stated, we shall consider motions in a three-dimensional space. The independent variables are a set of particle labels $\mathbf{a} = (a, b, c)$ and time t . The dependent variables are the coordinates

$$\mathbf{x}(\mathbf{a}, t) = (x(\mathbf{a}, t), y(\mathbf{a}, t), z(\mathbf{a}, t)) \quad [1]$$

of the fluid parcel identified by (a, b, c) . At any given time there is a one-to-one correspondence between the coordinate systems $\mathbf{a}$ and $\mathbf{x}(\mathbf{a}, t)$. We assume that each fluid parcel, which really refers to an imaginary piece of the continuum, is uniquely defined by values of the labels $\mathbf{a}$ for all time. If $\mathbf{x}$ is a suitably differentiable function of its arguments, then the velocity of a parcel is given by $\mathbf{v} = (\partial \mathbf{x} / \partial t)_{\mathbf{a}} = (u, v, w)$ and the acceleration by $(\partial^2 \mathbf{x} / \partial t^2)_{\mathbf{a}}$, where the subscript $\mathbf{a}$ means that the time derivative is evaluated on a parcel (a constant value of $\mathbf{a}$). (In conventional point-particle mechanics, we are accustomed to assigning a position vector, $\mathbf{r}_i$ say, to the i th particle; the continuous label $\mathbf{a}$ in fluid mechanics is the generalization of the discrete label i .) Lagrangian mechanics is the basis for many important conceptual models of fluid flow. In the second method, known as the Eulerian method, we fix our attention on a region of space and study the motion of fluid relative to

that region. That is, the independent variables are the space coordinates $\mathbf{x} = (x, y, z)$ and time t , and the dependent variables are most commonly taken to be the velocity $\mathbf{v}(\mathbf{x}, t)$, the mass density $\rho(\mathbf{x}, t)$, and the pressure $p(\mathbf{x}, t)$. From the density and pressure all other thermodynamic quantities can be determined, provided the equation of state is known.

The time derivative of an arbitrary quantity $A(\mathbf{a}, t) = A(\mathbf{x}(\mathbf{a}, t), t)$ (where we have used the same symbol to denote two different functions) measured in the two descriptions is related by the chain rule

$$\begin{aligned} \left. \frac{\partial A}{\partial t} \right|_{\mathbf{a}} &= \left. \frac{\partial A}{\partial t} \right|_{\mathbf{x}} + \frac{\partial \mathbf{x}}{\partial t} \cdot \frac{\partial A}{\partial \mathbf{x}} \\ &= \left. \frac{\partial A}{\partial t} \right|_{\mathbf{x}} + u \left. \frac{\partial A}{\partial x} \right|_{y, z, t} + v \left. \frac{\partial A}{\partial y} \right|_{x, z, t} + w \left. \frac{\partial A}{\partial z} \right|_{x, y, t} \equiv \frac{DA}{Dt} \end{aligned} \quad [2]$$

which illustrates the relationship between the familiar material or substantial derivative, D/Dt , and $(\partial/\partial t)_{\mathbf{a}}$. The relationship $(\partial/\partial t)_{\mathbf{a}} = D/Dt$ tells us why the material derivative is often referred to as the ‘derivative following a parcel.’

The Lagrangian description of fluid kinematics and dynamics, and related topics, are the subject of this article. The Eulerian description is adopted by most textbooks, but there are many fundamental principles in fluid dynamics that are inherently Lagrangian in nature. To illustrate this point, consider the continuity equation in Eulerian form

$$\left. \frac{\partial \rho}{\partial t} \right|_{\mathbf{x}} + \nabla \cdot (\rho \mathbf{v}) = 0 \quad [3]$$

where, unless otherwise stated, ∇ is the usual gradient operator in three-dimensional space. This familiar equation arises from the requirement that fixed volumes in particle-label space always contain the same mass. That is, if we assign labeling coordinates so that $\mathbf{x}(\mathbf{a}, 0) \equiv \mathbf{a}$ at some initial time $t = 0$, then the mass contained in the small volume $d^3 \mathbf{x}(\mathbf{a}, 0)$ of fluid at $\mathbf{x}(\mathbf{a}, 0)$ is given by $\rho(\mathbf{x}(\mathbf{a}, 0), 0) d^3 \mathbf{x}(\mathbf{a}, 0) = \rho(\mathbf{a}, 0) d^3 \mathbf{a}$. This defines the density, ρ , which we may consider as either a function of Lagrangian labels, $\mathbf{a}$, or Eulerian positions, $\mathbf{x}$. At a later time t , we must have $\rho(\mathbf{x}(\mathbf{a}, t), t) d^3 \mathbf{x}(\mathbf{a}, t) = \rho(\mathbf{a}, 0) d^3 \mathbf{a}$. If we define the specific volume, α , as the Jacobian

$$\alpha \equiv \frac{\partial(\mathbf{x})}{\partial(\mathbf{a})} \quad [4]$$

then the conservation of mass is given by $\rho(\mathbf{x}(\mathbf{a}, t), t) \alpha = \rho(\mathbf{a}, 0)$. For an incompressible fluid, we have $\alpha = 1$ (which corresponds to $\nabla \cdot \mathbf{v} = 0$ in the Eulerian description). It is now possible to

show that $(\partial\alpha/\partial t)_a = \alpha \nabla \cdot \mathbf{v}$, which in turn can be expressed in the form of eqn [3].

Variational Principles and Conservation Laws

In writing down variational principles for fluid systems, one is immediately faced with the choice of adopting either Lagrangian or Eulerian kinematics. Occasionally, it is convenient to change between these representations when performing calculations. In this section we shall review one of the most important features of the Lagrangian description: a connection between an invariant property of this representation and the conservation of a meteorologically significant quantity, potential vorticity.

Hamilton's Principle and Noether's Theorem

We begin by discussing a finite-dimensional problem. Consider a system of n point-particles, with masses $m_i (i = 1, \dots, n)$ and locations $\mathbf{x}_i(t)$, moving under the influence of some potential $V(\mathbf{x}_i)$. Newton's second law of motion for such a system is

$$m_i \frac{d^2 \mathbf{x}_i}{dt^2} + \frac{\partial V}{\partial \mathbf{x}_i} = 0 \quad [5]$$

Hamilton's principle is a variational principle equivalent to Newton's second law, and it states that the action

$$S \equiv \int_{t_0}^{t_1} L dt \quad [6]$$

is stationary, where the Lagrangian $L(\mathbf{x}, \dot{\mathbf{x}}) \equiv T - V$ is the difference between the kinetic energy T and the potential energy V of the system. Hamilton's principle thus states that the variation of the action eqn [6] vanishes for arbitrary, independent variations $\delta \mathbf{x}_i(t)$ that vanish at t_1 and t_0 . Since $\delta \mathbf{x}_i(t_1) = \delta \mathbf{x}_i(t_0) = 0$, we have

$$\int_{t_0}^{t_1} \left[-\frac{d}{dt} \left(\frac{\partial L}{\partial \dot{\mathbf{x}}_i} \right) + \frac{\partial L}{\partial \mathbf{x}_i} \right] \cdot \delta \mathbf{x}_i dt = 0 \quad [7]$$

for all i , where the overdot denotes d/dt . Newton's law follows as a consequence of the arbitrariness of $\delta \mathbf{x}_i(t)$, and the specific form eqn [5] follows from the functional dependence of the potential energy V on $\mathbf{x}_i$ alone, and the usual expression for the kinetic energy

$$T = \frac{1}{2} \sum_{i=1}^n m_i \dot{\mathbf{x}}_i \cdot \dot{\mathbf{x}}_i \quad [8]$$

There is a fundamental connection between certain invariant properties of Lagrangians and the conservation laws for dynamical systems. This connection is embodied in Noether's theorem, which requires considerable expertise in group theory to understand completely. We shall therefore restrict ourselves to a didactic description. Noether's theorem applies to the equations that arise from variational principle like Hamilton's principle. According to *Noether's theorem* (1918), if a variational principle is invariant to a continuous transformation of its dependent and independent variables,

then the equations arising from the variational principle possess a conservation law.

Example. Let us assume that a system has kinetic energy of the form eqn [8], but the potential is independent of x (but still dependent on y and z , say). Therefore the Lagrangian is invariant with respect to small variations in x . The integrand of eqn [7], the so-called Euler–Lagrange equation,

$$\frac{d}{dt} \left(\frac{\partial L}{\partial \dot{x}} \right) - \frac{\partial L}{\partial x} = 0 \quad [9]$$

then yields, assuming independence of x ,

$$\frac{\partial L}{\partial x} = 0 \Rightarrow \frac{d}{dt} \left(\frac{\partial L}{\partial \dot{x}} \right) = 0$$

This is the statement that the x -component of momentum, $m\dot{x}$, is conserved.

Application to Fluid Mechanics: Theorems of Ertel and Kelvin

The fundamental meteorological-cum-oceanographic principle of the conservation of potential vorticity can be traced to the invariance of variational principles for hydrodynamical flows under certain variations of the particle labels, $\delta \mathbf{a}$.

Let $\mathbf{x}(\mathbf{a}, t)$ be the location of the fluid particle identified by labeling coordinates $\mathbf{a} = (a, b, c)$ at time t . The Lagrangian for a perfect fluid is

$$\mathcal{L} = \int d^3 \mathbf{a} \left[\frac{1}{2} \dot{\mathbf{x}}^2 - E(\alpha, S(\mathbf{a})) \right] \quad [10]$$

where the internal energy $E(\alpha, S)$ is a prescribed thermodynamic function of the specific volume, α , and the specific entropy, $S(\mathbf{a})$. The entropy is a Lagrangian conserved quantity. The integration over a measure in particle-label space, $\int d^3 \mathbf{a}$, replaces the discrete sum $\sum_i$, and the overdot is the time derivative $(\partial/\partial t)_a$.

Hamilton's principle states that

$$\delta \int \mathcal{L} dt = 0$$

where δ stands for arbitrary independent variations $\delta \mathbf{x}(\mathbf{a}, t)$, and implies

$$\ddot{\mathbf{x}} + \alpha \nabla p = 0 \quad [11]$$

where $p = -\partial E / \partial \alpha$ is the usual thermodynamic equation relating pressure to internal energy and it may be considered the equation of state. Equation [11] is the momentum balance for a perfect fluid.

The Lagrangian is unaffected by particle-label variations $\delta \mathbf{a}(\mathbf{x}, t)$ that leave the density and entropy unchanged. For the purposes of the following calculation, we shall assume that the particle labels enter only through the specific volume, which is equivalent to the assertion that the fluid is homentropic: $E = E(\alpha)$. (The results can easily be generalized to nonhomentropic fluids where $E = E(\alpha, S)$.) Consider variations $\delta \mathbf{a}(\mathbf{x}, t)$, of eqn [10], that leave the density and entropy unchanged. Then

$$\begin{aligned} \delta \int \mathcal{L} dt &= \iint dtd^3 \mathbf{a} \frac{\partial \mathbf{x}}{\partial t} \cdot \frac{\partial \delta \mathbf{x}}{\partial t} \\ &= - \iint dtd^3 \mathbf{a} \mathbf{A} \cdot \frac{\partial (\delta \mathbf{a})}{\partial t} \end{aligned} \quad [12]$$

where

$$\mathbf{A} = u\nabla_{\mathbf{a}}x + v\nabla_{\mathbf{a}}y + w\nabla_{\mathbf{a}}z \quad [13]$$

$\nabla_{\mathbf{a}}$ is the gradient operator in particle-label space, and we used the chain rule for partial derivatives to go from the second term to the third term in eqn [12]. After integration by parts, noting that $\delta\mathbf{a}(\mathbf{x},t)$ is arbitrary, one can show that

$$\partial/\partial t[(\nabla_{\mathbf{a}} \times \mathbf{A})] = 0 \quad [14]$$

Equation [14] is a general statement of vorticity conservation. Let $\theta(\mathbf{a})$ be any quantity that is conserved on fluid particles. Then

$$\partial/\partial t[(\nabla_{\mathbf{a}} \times \mathbf{A}) \cdot \nabla_{\mathbf{a}}\theta] = 0 \quad [15]$$

Using eqns [4] and [13] we have

$$(\nabla_{\mathbf{a}} \times \mathbf{A}) \cdot \nabla_{\mathbf{a}}\theta = \alpha(\nabla \times \mathbf{v}) \cdot \nabla\theta \quad [16]$$

and, from eqns [15] and [16], the statement

$$\partial/\partial t[\alpha(\nabla \times \mathbf{v}) \cdot \nabla\theta] = 0 \quad [17]$$

expresses the conservation of potential vorticity – *Ertel's theorem*.

Now consider any closed loop in $\mathbf{a}$ -space. By eqn [14], it follows that

$$\partial/\partial t \oint \mathbf{A} \cdot d\mathbf{a} = 0 \quad [18]$$

But $\mathbf{A} \cdot d\mathbf{a} = \mathbf{v} \cdot d\mathbf{x}$ by eqn [13], and the statement

$$\partial/\partial t \oint \mathbf{v} \cdot d\mathbf{x} = 0 \quad [19]$$

is *Kelvin's theorem*.

It should be noted that commuting the time derivative and the contour integral is a nontrivial matter in the Eulerian setting (see Betounes, 1983, and compare with Lamb, 1932).

Reduction of Order

The state of a perfect fluid at a fixed time t corresponds to a point in an infinite-dimensional phase space in which each dimension represents the value of one component of $\mathbf{v}(\mathbf{a})$ or $\mathbf{x}(\mathbf{a})$ at a fixed value of $\mathbf{a}$. The six Lagrangian fields

$$\{\mathbf{v}(\mathbf{a}), \mathbf{x}(\mathbf{a})\}$$

uniquely determine the five Eulerian fields

$$\{\mathbf{v}(\mathbf{x}), \rho(\mathbf{x}), S(\mathbf{x})\}.$$

However, each choice of $\{\mathbf{v}(\mathbf{x}), \rho(\mathbf{x}), S(\mathbf{x})\}$ corresponds to infinitely many choices of $\{\mathbf{v}(\mathbf{a}), \mathbf{x}(\mathbf{a})\}$. Thus the Eulerian description is a *reduced phase space* for the fluid. A group-theoretic picture of the reduction from Lagrangian to Eulerian variables gives a more rigorous description, but this fundamental topic is a subject for further reading.

Numerical Techniques

Lagrangian thinking manifests itself in the design of numerical methods. Here we give a brief review of two schemes: semi-Lagrangian techniques, which are widely used and a novel Lagrangian approach to the semigeostrophic frontogenesis model – the so-called geometric method.

Semi-Lagrangian Schemes

Semi-Lagrangian schemes are important because they offer the promise of allowing longer time steps, with no loss of accuracy, than Eulerian-based advection schemes whose time step is limited by stability criteria. These issues are of practical importance in numerical weather prediction.

To illustrate the essential ideas, let us consider a simple one-dimensional advection equation in Eulerian form

$$\frac{\partial A}{\partial t} + u \frac{\partial A}{\partial x} = 0 \quad [20]$$

where $A(x,t)$ is the transported quantity. The appropriate Lagrangian form of eqn [20] is the pair of equations

$$\frac{dA}{dt} = 0 \quad [21a]$$

$$\frac{dx}{dt} = u \quad [21b]$$

Equation [21a], states that A is conserved along the trajectory, which is in turn given by the eqn [21b]. We seek to integrate eqn [21a] and [21b] by gridpoint techniques. (This is the origin of the term ‘semi’ in semi-Lagrangian – the combination of Lagrangian mechanics with Eulerian gridpoint techniques.) Let Δt denote a time step and let n label the time steps. For each point on the grid we approximate the trajectory that would arrive at that point at time $(n+1)\Delta t$ using the wind u at time n . We refer to the point at which the forecast is made as the arrival point (A) and the point from which the trajectory departed at time $(n-1)\Delta t$ as the departure point (D). The point in the middle of the trajectory at time $n\Delta t$ is called the midpoint (M). The departure point is determined by approximating eqn [21b] with

$$\frac{x_D - x_A}{-2\Delta t} = u(x_M) \quad [22]$$

The midpoint, at which u is needed in eqn [22], is obtained by a similar expression

$$\frac{x_M - x_A}{-\Delta t} = u(x_M) \quad [23]$$

which is implicit since u at x_M cannot be determined until x_M is known. This requires an iterative process in which the next guess of x_M on the left is based on u at x_M from the previous iteration. Once x_M is found, the departure point calculated from eqns [22] and [21a] yields

$$A(x_A)^{n+1} = A(x_D)^{n-1} \quad [24]$$

Typically the departure points and midpoints will not coincide with grid points. The values of A and u at these points must be found by interpolation from neighboring points and the scheme is stable provided the interpolation is based on data points surrounding the departure point (or midpoint). There is essentially no time step restriction and thus, near the poles of a global model where grid lines converge, the departure point may be many grid intervals away from the arrival point without being unstable.

Geometric Method

The geometric method was born out of an attempt to demonstrate the value of a Lagrangian approach to

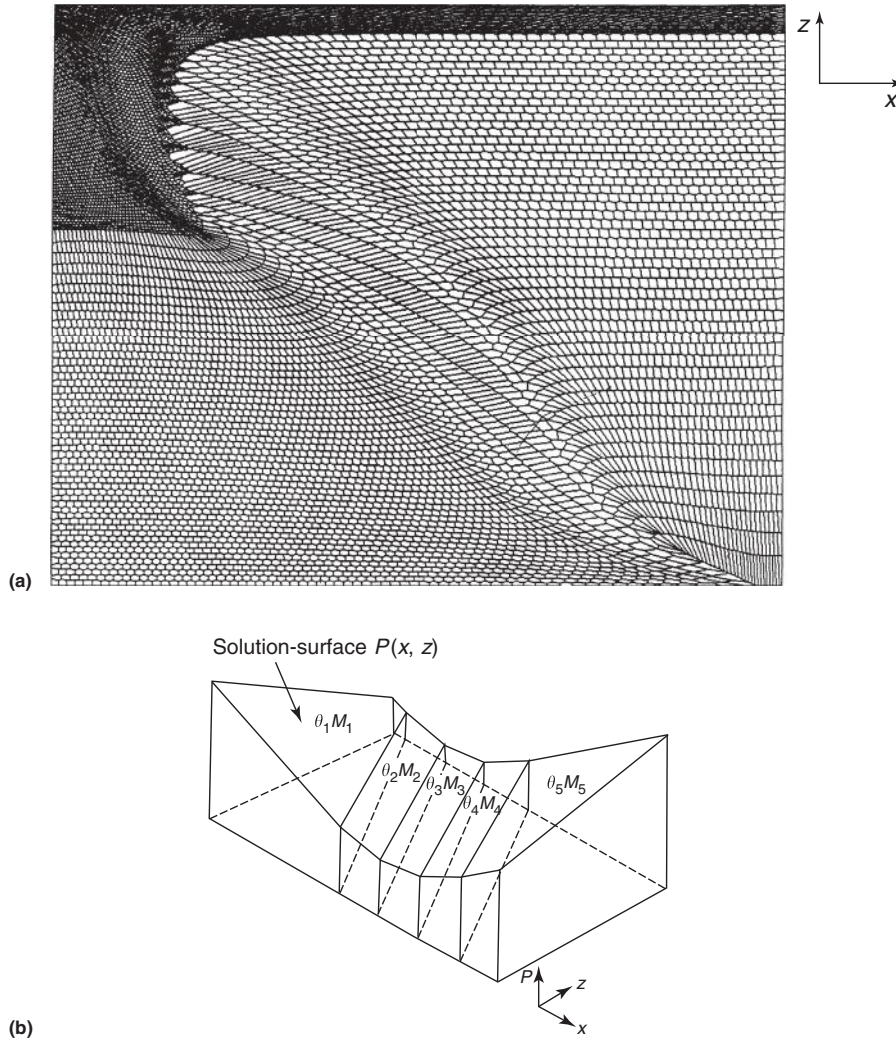


Figure 1 (a) Projection of a polyhedral surface $P(x, z)$ onto the (x, z) -plane. The front is modeled as a discontinuity in the gradients of the faces that make up the piecewise planar surface. (b) Construction of a convex polyhedral surface $P(x, z)$ from faces with given gradients $(\theta_i M_i)$ and areas.

understanding the dynamics of rotating, stratified fluids. It is a novel technique for integrating the semigeostrophic equations that involves dividing the atmospheric domain into elements, each characterized by a certain mass, potential temperature, and absolute momentum. Applications of the geometric method include modeling frontogenesis, embedded convection, sea breezes, and flow over orography. Some of the advantages of the geometric method include the ability to handle a frontal discontinuity, no eddy diffusion requirement and, in principle at least, mountain barrier effects (e.g., drag) can be represented without parameterization. An example of a ‘tropopause fold,’ constructed via the geometric method with 10^5 elements, is illustrated in Figure 1(a).

The geometric method may be summarized as follows. Consider a domain D with coordinates (x, y, z) , taken to represent a region of the Northern Hemisphere (y pointing poleward). Let C be a cross section

$$C = \{(x, z) : -L \leq x \leq L, 0 \leq z \leq H\}$$

(where typical values of L and H may be 1000 and 10 km, respectively). The geometric method discretizes the fluid in C into ‘parcels’ of equal area C_i , so that $C = \sum_i C_i$. To the i th parcel is associated a value of ‘absolute momentum’ $M_i = (v_g + fx)_i$ and a value of potential temperature θ_i . Here, v_g is the y -component of the geostrophic wind, $v_g = (1/f)\partial\phi/\partial x$, $\phi(x, z)$ is the geopotential function, and f is the constant Coriolis parameter. One can show that a convectively and inertially stable arrangement of the fluid parcels can be represented uniquely by a convex function $P(x, z) = \frac{1}{2}f^2x^2 + \phi$ (sometimes referred to as the ‘modified geopotential’ or ‘modified pressure’), such that each element C_i can be associated with a plane whose height ‘above’ the (x, z) -plane at any point corresponds to the value of P at that point. The intersection of these planes defines a piecewise planar solution $P(x, z)$ (Figure 1(b)). The momentum of each parcel, M_i , and potential temperature of each parcel, θ_i , are proportional to the gradients of $P(x, z)$ with respect to x_i and z_i , respectively.

The constraint that P be a convex function is crucial to the solution procedure. It allows us to write the physical solution as the supremum of a family of generating surfaces

$$P(x, z) = \max_i \{xM_i + z\theta_i + S_i\} = \max P_i$$

where P_i denotes the modified pressure plane of the i th element and S_i the point of intersection on the P -axis of a plane in (x, z, P) -space. In a physical problem involving time integration, we begin with a set of values C_i , M_i , θ_i , and S_i . The construction algorithm requires the determination of the set of S_i values such that the element areas are correct – a process that must be carried out iteratively. If the domain boundary is fixed or a known function of time, one may regard S_i as a function of C_i , M_i , and θ_i . In the classical frontogenesis problem where a pure barotropic deformation field is imposed, the governing Lagrangian equations take the form

$$\frac{DM_i}{Dt} = -\eta M_i, \quad \frac{DC_i}{Dt} = -\eta C_i, \quad \frac{D\theta_i}{Dt} = F_i \quad [25]$$

where η is the deformation rate and F_i is a forcing function (which is zero in the frontogenesis problem under consideration). Therefore, at any future time C_i , M_i , and θ_i are known and the solution can be found by determining S_i .

Lagrangian Analysis

The description of weather systems and ocean eddies has always utilized the Lagrangian description of the flow. Examples include synoptic developments in terms of air masses (dating back to the Bergen School in the early twentieth century), the parcel theory of convection, and the description of the dynamics of precipitation systems in terms of conveyor belts. The power of the Lagrangian description in these contexts is the conceptual simplification achieved by burying the nonlinearity of the material derivative D/Dt in the Jacobian of the map between particle labels and Lagrangian positions: $\alpha = \partial(\mathbf{x})/\partial(\mathbf{a})$ (cf. eqn [4]). However, whenever we wish to carry out a theoretical analysis of hydrodynamical flows, the Eulerian framework is the most commonly used vehicle. In this section we give a brief resumé of a Lagrangian method of analysis that, although hitherto not widely used, offers a powerful technique for analyzing, for example, the conceptually simple (yet analytically difficult in the Eulerian setting) problems of transportation along particle trajectories.

Atmospheric Dynamics and Rearrangements

A concept that arises naturally when considering the Lagrangian description of fluid motion is that of the rearrangement of the fluid by the material derivative. The action of the D/Dt operator on an arbitrary quantity A is to advect, or rearrange, it. Consider for example a conserved quantity, such as potential vorticity $q = \rho^{-1}\zeta \cdot \nabla\theta$, that retains its value following a fluid particle. Here, ρ is the fluid density, θ the potential temperature, and ζ the total vorticity. As time advances the fluid particles are permuted or rearranged, but each particle retains its original value of q .

Atmospheric cyclones and anticyclones, and ocean eddies, can be idealized as the stratified, rotating coherent structures

that correspond to circular vortices in ordinary two-dimensional Euler flow. Their interaction and evolution, which play a major role in weather developments and in the behavior of ocean eddies, have been much studied using approximations to Newton's second law of motion. These approximate models seek to describe flows in which there is a dominant balance between the Coriolis, buoyancy, and pressure-gradient forces. Such approximations to Newton's second law are commonly referred to as *balanced models*. Many such models can be described in terms of the Lagrangian conservation of potential vorticity by an equation

$$\frac{Dq}{Dt} = 0$$

together with a so-called invertibility principle

$$E(\mathbf{v}, \theta, p) = q \quad [26]$$

that relates the wind field $\mathbf{v}$, pressure p , and potential temperature θ to the potential vorticity. Typically, E will be an elliptic operator, which may be nonlinear, and certain boundary conditions must be specified. A relationship between the wind, temperature, and pressure fields is known as a balance condition, which is required to define the relationships implicit in eqn [26]. Examples of models that can be formulated in this way are the barotropic vorticity equation, quasi-geostrophic theory, and semigeostrophic theory. The solutions to all these models can be described as rearrangements of the initial potential vorticity distribution, and because the advecting velocity is constrained by the invertibility procedure, progress in understanding features such as existence and uniqueness and the topological properties of solutions can be made because the elliptic operator governs the type of rearrangement that can be reached from the given initial data.

The precise definitions of a rearrangement of functions (both scalar-valued and vector-valued) require concepts from measure theory, and as such have very precise technical definitions. However, the technical nature of the definitions is extremely important, since it gives a firm basis to the mathematical analysis. Bearing this in mind, we will give an intuitive definition. Consider a region, D , spanned by the coordinates (x, y, z) and let $f(\mathbf{x})$ be a function defined on this region (e.g., moisture or potential vorticity). Within the region we have our fluid with Lagrangian coordinates $(\mathbf{a}, t)$ and let us assume we can attach values of the function f to each fluid particle, thereby giving us a function $F(\mathbf{x}, t_0)$, at some reference time t_0 . As the state of the fluid evolves according to a dynamical model, at a later time t_1 we have a new function $G(\mathbf{x}, t_1)$, which we call a rearrangement of F if the two functions satisfy a certain equivalence relationship between the 'sizes' or 'volumes' of the sets on which F and G take values greater than or equal to a datum value, for all real positive datum values. An example of two functions F and G that are rearrangements according to this definition is given in Figure 2.

Possible applications of rearrangements include numerical methods in which the goal is to model the evolution of a quantity such as potential vorticity as accurately as possible. In variational data assimilation it may be useful to work with Lagrangian increments, instead of Eulerian perturbations, in which minimization is carried out over certain classes of 'dynamically accessible' rearrangements. Rearrangements can

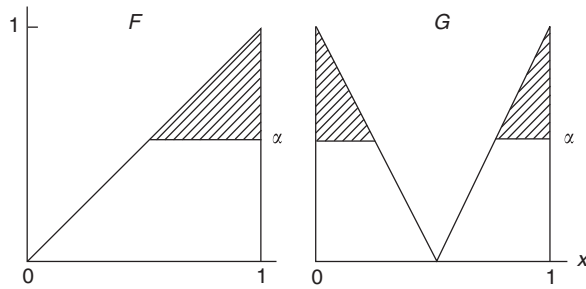


Figure 2 Two rearrangements F and G ; the area of the shaded regions is the same for every value of the datum point α .

be used to compare two functions of spatial variables and as such there may be applications in the future to forecast verification techniques.

A significant achievement of rearrangement theory to date is its application to the study of the stability of steady states. Steady states can be characterized as stationary points of the energy with respect to rearrangement perturbations, which is not possible if Eulerian perturbations are used.

Acknowledgment

The author would like to thank T.J. Bridges for drawing attention to the paper by Betounes.

See also: Chemistry of the Atmosphere: Tracers. **Dynamical Meteorology:** Balanced Flow; Hamiltonian Dynamics; Kinematics; Potential Vorticity; Wave Mean-Flow Interaction. **Numerical Models:** Methods.

Further Reading

- Betounes, D.E., 1983. The kinematical aspect of the fundamental theorem of calculus. *American Journal of Physics* 51, 554–560.
- Hoskins, B.J., 1982. The mathematical theory of frontogenesis. *Annual Review of Fluid Mechanics* 14, 131–151.
- Hoskins, B.J., McIntyre, M.E., Robertson, A.W., 1985. On the use and significance of isentropic potential vorticity maps. *Quarterly Journal of the Royal Meteorological Society* 111, 877–946.
- Lamb, H., 1932. *Hydrodynamics*, sixth ed. Cambridge University Press, Cambridge.
- Marsden, J.E., Ratiu, T., 1994. *Text In Applied Mathematics*. Springer-Verlag, Berlin.
- Norbury, J., Roulstone, I. (Eds.), 2002. *Large-Scale Atmosphere–Ocean Dynamics*: vol. 1 Analytical Methods and Numerical Models; vol. 2 Geometric Methods and Models. Cambridge University Press, Cambridge.
- Salmon, R., 1998. *Lectures on Geophysical Fluid Dynamics*. Oxford University Press, Oxford.
- Shepherd, T.G., 1990. Symmetries, conservation laws and hamiltonian structure in geophysical fluid dynamics. *Advances in Geophysics* 32, 287–338.
- Shutts, G.J., Cullen, M.J.P., Chynoweth, S., 1988. Geometric models of balanced semi-geostrophic flow. *Annals of Geophysics* 6, 493–500.
- Staniforth, A., Côté, J., 1991. Semi-lagrangian integration schemes for atmospheric models: a review. *Monthly Weather Review* 119, 2206–2223.

Potential Vorticity

ME McIntyre, University of Cambridge, Cambridge, UK

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

The significance of the potential vorticity (PV) for atmosphere–ocean dynamics was first explored by Carl-Gustaf Rossby in the 1930s. Reviewed here are its key properties including invertibility, material invariance, and the impermeability theorem – the last two suggesting mixability along stratification surfaces. These properties easily explain the once-mysterious antifriction or ‘negative viscosity’ of strongly nonlinear atmosphere–ocean eddy fields, outside the scope of linear theory and homogeneous turbulence theory. Invertibility implies that eddy fluxes of momentum are intimately related to isentropic eddy fluxes of PV, including those due to strongly nonlinear disturbances, as summarized by the quasigeostrophic Taylor identity.

The Fundamental Definition

The idea of the potential vorticity (PV) as a material invariant central to stratified, rotating fluid dynamics was first introduced and explored by Carl-Gustaf Rossby in the 1930s. Material invariance means constancy on a fluid particle. The PV, a scalar field, will be denoted here by P and can be defined in several ways, as shown shortly. We have

$$DP/Dt = 0 \quad [1]$$

for dissipationless flow, where D/Dt is the material derivative. For such flow we also have material invariance of the potential temperature θ ,

$$D\theta/Dt = 0. \quad [2]$$

Rossby’s idea, as it originally emerged from his papers of 1936, 1938, and 1940, was to introduce a vorticity-like quantity that is related to the vertical component of vorticity in the same way that potential temperature is related to temperature. In his 1938 and 1940 papers he recognized, moreover, that ‘vertical’ can more accurately be replaced by ‘normal to stratification surfaces,’ i.e., in the atmosphere, normal to isentropic or constant- θ surfaces.

Equivalent to this is the idea, clearly emerging on page 252 of the 1938 paper, that P is exactly proportional to the absolute Kelvin circulation C_Γ , eqn [7] below, around an infinitesimally small closed material contour Γ lying on an isentropic surface. The exact material-invariance property [1] is then obvious from Kelvin’s circulation theorem, as generalized by V. Bjerknes, since [2] ensures that the material contour Γ remains on the isentropic surface.

Rossby’s idea is today recognized as having central and far-reaching importance for understanding the dynamical behavior not only of planetary atmospheres and oceans but also of the radiative interiors of solar-type stars. It is especially important for understanding *balanced flow* and thence a vast range of basic dynamical processes, such as Rossby-wave propagation and breaking and its many consequences including, in the Earth’s atmosphere, global-scale teleconnections, antifrictional phenomena such as jet stream self-sharpening, and the genesis of cyclones, anticyclones and storm tracks, answering the child’s age-old question of where the wind comes from.

The relation $P \propto C_\Gamma$ provides the simplest and most fundamental way to define P exactly, not only for continuously

stratified systems but also for single-layer shallow-water or ‘equivalent barotropic’ models and their multilayer extensions. For continuous stratification, today’s standard definition of P chooses the constant of proportionality to be $d\theta$, the potential-temperature increment between a pair of neighboring isentropic surfaces (see Figure 1), divided by the mass of the small material fluid element lying between those surfaces and having perimeter Γ . Mass conservation is assumed throughout this article.

For the single-layer and multilayer models, one needs only to replace the pair of isentropic surfaces by layer boundaries. Then for finite layer thickness the proportionality constant can be chosen as simply the reciprocal of the mass of the material element, or of its volume when the usual incompressible-flow assumption is made. Then from Stokes’ theorem P becomes absolute vorticity divided by layer thickness, the formula first presented in Rossby’s 1936 paper.

For continuous stratification, Rossby derived an approximate formula adequate for use with synoptic-scale observational data. With the foregoing choice of proportionality constant, Rossby’s formula is

$$P \approx g \left\{ \left(\frac{\partial v}{\partial x} - \frac{\partial u}{\partial y} \right)_\theta + f \right\} \left| \frac{\partial \theta}{\partial p} \right| \quad [3]$$

where g is the gravitational acceleration, p is pressure, and f is the Coriolis parameter, a function of latitude. To obtain [3] from the exact relation $P \propto C_\Gamma$ one must assume that the mass and pressure fields are related hydrostatically and that the slopes of isentropic surfaces are small in comparison with unity. In practice these conditions usually hold to more than

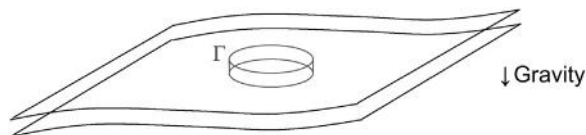


Figure 1 Sketch showing the material mass element defined by a small isentropic contour Γ and a pair of neighboring isentropic (stratification) surfaces with potential temperatures θ and $\theta + d\theta$. The exact PV is the mass-normalized Kelvin circulation around Γ , in the limit of an infinitesimally small element (see text). In a layer model, the two surfaces are taken instead as the layer boundaries.

sufficient accuracy. The horizontal coordinates x, y in [3] are local Cartesian coordinates in a tangent-plane representation, with corresponding horizontal velocity components u, v relative to the Earth. The formula converts to spherical or other coordinates in the same way as the ordinary vertical vorticity.

However, as Rossby pointed out, the quantity within braces is not the ordinary vertical vorticity. The subscript θ is crucial. It signifies that the horizontal differentiations of the horizontal velocity components are to be carried out with θ held constant. That is, one stays on a single isentropic surface, just as one does when calculating C_T . Rossby explains this point very clearly on, for instance, page 253 of his 1938 paper. The resulting quantity, bearing a superficial resemblance to the ordinary vertical vorticity, can more aptly be called Rossby's *isentropic vorticity*. Within the approximations involved in [3], this isentropic vorticity is the same as the component of the vorticity vector normal to the isentropic surface. It can differ substantially from the vertical vorticity.

Such differences are commonplace in balanced flows with strong vertical shear ($\partial u/\partial z, \partial v/\partial z$), where z is geometric altitude or pressure altitude. That is, they are commonplace in balanced flows with high baroclinicity. Examples include tropopause jet streams. Baroclinicity means tilting of isentropic surfaces relative to isobaric surfaces, usually the cross-stream tilting that balances the vertical shear as indicated by the so-called thermal wind equation. A natural measure of baroclinicity is $1/Ri$, where $Ri = N^2/(\partial|\mathbf{u}|/\partial z)^2$, the gradient Richardson number, where $N^2 = g\theta^{-1}\partial\theta/\partial z$, the square of the buoyancy frequency. The shear and cross-stream tilting effects were shown to make substantial contributions to the right-hand side of [3] in, for instance, the 1950s work of R. J. Reed, F. Sanders, and E. F. Danielsen on observational data describing tropopause fronts and jet streams, in which air of stratospheric origin was recognized by its relatively high values of P . Slopes are geometrically small but Ri values low enough for the subscript θ to be important in [3].

Equations [1]–[3] provide a remarkably succinct description of how dissipationless processes affect the component of absolute vorticity normal to an isentropic surface. There are two distinct effects. The first is that the normal component of absolute vorticity increases through vortex stretching if the isentropic surfaces move apart. This is a generalization of angular momentum conservation, i.e., a generalization of the ballerina effect or ice-skater's spin. The second is that the normal component of absolute vorticity is preserved if the isentropic surfaces do nothing but tilt away from the horizontal.

The generalized ballerina effect often contributes to the spin-up of cyclonic vortices, such as the small vortex over the Balkans in Figure 2. The colors mark air with different estimated values of P , on the $\theta = 320$ K isentropic surface at geometric altitudes around 10 km, with the warmest colors marking the highest P values. The vortex over the Balkans has a core of high- P air that has undergone stretching, while moving equatorward out of the stratosphere. The cyclonic, i.e., counterclockwise, rotation of the core relative to the surrounding air shows up as a tendency of the surrounding colored filaments to be wound up into spirals.

The estimated isentropic distribution of P shown in Figure 2 was derived from an initial coarse-grain estimate from operational weather-forecasting analyses together with an assumption that material invariance, [1] with [2], holds to sufficient accuracy over 4 days. A highly accurate tracer advection technique, contour advection, was used. It was first introduced into the atmospheric-science literature by W. A. Norton, R. A. Plumb and D. W. Waugh following work of N. J. Zabusky and D. G. Dritschel. The pattern thus revealed, reminiscent of cream on coffee, illustrates the typical advective effects of the layerwise-two-dimensional flow characteristic of mesoscale and larger-scale flow regimes heavily constrained by stable stratification. Such regimes can often be considered to be balanced flows, whose isentropic distributions of P contain nearly all the

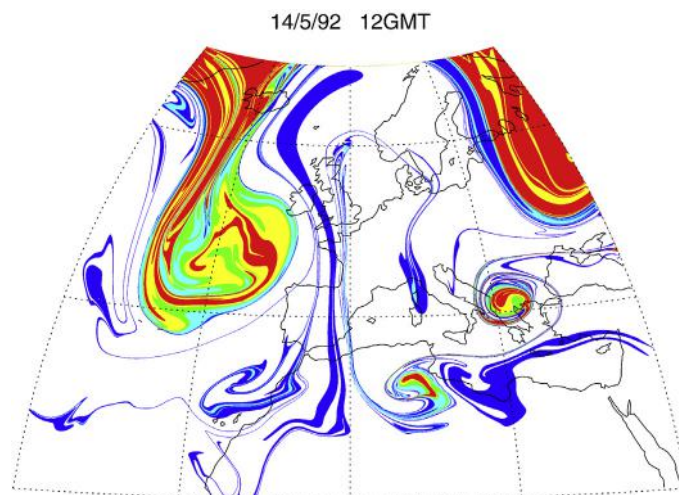


Figure 2 Estimated isentropic distribution of the (Rossby–Ertel) PV on the 320 K isentropic surface on 14 May 1992 at 1200 UT (Greenwich mean time), derived from observations as explained in the text. Over Europe the 320 K surface lies near jetliner cruising altitudes $z \sim 10$ km. The estimate used data from the operational weather-prediction analyses of the European Centre for Medium Range Weather Forecasts (ECMWF). Values from 1 PVU upwards are colored rainbow-wise from dark blue to red, with contour interval 1 PVU, where $1 \text{ PVU} = 10^{-6} \text{ m}^2 \text{ s}^{-1} \text{ K kg}^{-1}$. Courtesy W. A. Norton (personal communication); further details in Appenzeller et al. (1996). Figure 15(b) on p. 1450 of that paper checks that the wind field does, as expected from PV inversion, exhibit the usual tropopause jet structure around the periphery of the large high-PV region on the left. See *PV Mixability and Strong Jets* below.

information about the dynamics. This will be made precise in the section on [PV Inversion](#).

Ertel's Formula

For continuous stratification it is a simple exercise in vector calculus to show, via Stokes' theorem, that Rossby's fundamental relation $P \propto C_\Gamma$ is exactly equivalent to

$$P = \rho^{-1} \zeta^a \cdot \nabla \theta \quad [4]$$

when the constant of proportionality is chosen as before. Here ρ is the mass density, ∇ is the three-dimensional gradient operator, and ζ^a is the absolute vorticity vector, the curl of the three-dimensional velocity field viewed in an inertial frame. In the Earth's rotating frame, ζ^a is the three-dimensional relative vorticity added vectorially to twice the Earth's angular velocity vector $\mathbf{\Omega}$. The formula [4] was first published in 1942 by Hans Ertel, who had visited Rossby at MIT in 1937. The formula has attracted much attention in the mathematical fluid-dynamics community and has been generalized in various ways.

In strongly stratified flows like that of [Figure 2](#) we have $N^2 \gg 4|\mathbf{\Omega}|^2$. Also, the small-slope approximation is valid, making $\nabla \theta$ nearly vertical. In [4], the scalar multiplication by $\nabla \theta$ picks out f , the latitude-dependent vertical component of $2\mathbf{\Omega}$, to good approximation. This is the fundamental reason why f and its latitudinal variation often suffice to capture the main effects of the Earth's rotation $\mathbf{\Omega}$, including the so-called beta effect.

Under the small-slope and hydrostatic approximations, $\rho^{-1}|\nabla \theta|$ is approximately equal to $g|\partial \theta / \partial p|$ in [3]. The contributions to [3] and [4] from $2\mathbf{\Omega}$ therefore agree. It is straightforward to show that the remaining contributions also agree in these circumstances provided that, for consistency with the hydrostatic approximation, the vertical component of velocity is neglected when taking the curl of the relative velocity field to form the relative vorticity.

The small-slope and hydrostatic approximations are usually so good that [3] and [4] give practically indistinguishable results when evaluated from typical meteorological datasets, and from the output of numerical weather-forecasting models. So [3] and [4] are often treated as equivalent for practical purposes, both being called 'exact' when distinguishing them from the much less accurate formulae for the material invariants possessed by certain approximate balanced models, such as quasigeostrophic theory and semigeostrophic theory. Their material invariants are also called potential vorticities but are defined by formulae that differ substantially from [3] and [4], for instance [15]. Unlike [3] and [4] these formulae cannot be considered quantitatively accurate. The PV in its quantitatively accurate sense will be referred to as the *Rossby–Ertel potential vorticity* or simply, for brevity, the PV, whether defined by [3] or [4] or by any other formula accurately equivalent to $P \propto C_\Gamma$.

To check that [4] is accurately, indeed exactly, equivalent to $P \propto C_\Gamma$ and materially invariant for dissipationless flow, we note first that [4] can be rewritten exactly as

$$P = \sigma^{-1} \zeta^a \cdot \mathbf{n} \quad [5]$$

where $\sigma = \rho/|\nabla \theta|$, and $\mathbf{n} = \nabla \theta / |\nabla \theta|$, the upward-directed unit normal to the isentropic surface $\mathcal{S}$, say, on which P is being

evaluated. The scalar field σ , a stratification-related mass density, is a strictly positive quantity. Under the small-slope approximation it is the mass density in isentropic coordinates. With the definition just given, $\sigma d\theta$ is exactly the mass per unit area between neighboring isentropic surfaces, such as those sketched in [Figure 1](#), whose θ values differ by $d\theta$. Thus if dA is the area element of integration on the surface $\mathcal{S}$, then $\sigma dA d\theta$ is exactly the mass element of integration.

For dissipationless flow we have [2] as well as mass conservation, hence

$$\int_{\mathcal{S}(\Gamma)} \sigma dA = \text{constant} \quad [6]$$

where $\mathcal{S}(\Gamma)$ denotes any simply connected portion of $\mathcal{S}$ enclosed by a material contour Γ . Here Γ can, but need not, be small. By definition its Kelvin circulation is

$$C_\Gamma = \oint_{\Gamma} \mathbf{u}^a \cdot d\mathbf{x} = \text{constant} \quad [7]$$

for dissipationless flow, where $\mathbf{u}^a$ is the three-dimensional velocity field in the inertial frame. From Stokes' theorem and [5] we have exactly

$$C_\Gamma = \int_{\mathcal{S}(\Gamma)} \zeta^a \cdot \mathbf{n} dA = \int_{\mathcal{S}(\Gamma)} P \sigma dA \quad [8]$$

and if, as before, we now take Γ to be small – more precisely, if we take the greatest diameter of Γ to be arbitrarily small in comparison with all lengthscales of the flow – then P is simply [8] divided by [6]. This verifies not only the material invariance of P but also the equivalence of [4] and [5] to $P \propto C_\Gamma$ for small Γ , with the choice of proportionality constant made earlier.

For completeness we sketch the alternative derivation given by Ertel, written using the three-dimensional velocity field $\mathbf{u}$ relative to the rotating frame. One takes the scalar product of $\nabla \theta$ with the frictionless three-dimensional vorticity equation, the curl of the nonhydrostatic equation for $D\mathbf{u}/Dt$, and then makes use of $\nabla(D\theta/Dt) = 0$ from [2]. Note that $D/Dt = \partial/\partial t + \mathbf{u} \cdot \nabla$ and that the three-dimensional gradient operator ∇ acts on $\mathbf{u}$ as well as on θ . The baroclinic term in the vorticity equation, proportional to $\nabla p \times \nabla \rho$, is annihilated when the scalar product with $\nabla \theta$ is taken, because the thermodynamics says that θ is a function of p and ρ alone (the standard approximation to this function implying that $\theta \propto T/p^\kappa$, $\kappa \approx 2/7 \approx 0.286$, with temperature $T \propto p/\rho$). The result is a conservation relation in the general sense of the term, in 'flux form,'

$$\frac{\partial}{\partial t}(\rho P) + \nabla \cdot (\rho \mathbf{u} P) = 0 \quad [9]$$

with P defined by [4] or [5]. Putting this together with the corresponding equation

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \mathbf{u}) = 0 \quad [10]$$

expressing mass conservation, we immediately obtain eqn [1] for dissipationless flow.

A corollary of material invariance and mass conservation is the existence of so-called Casimir invariants. They are important in theories that make explicit the Hamiltonian mathematical structure of the dissipationless dynamics, and in associated

theorems on instability and on wave–mean interaction. Note first that we have not only constancy of [8] but also

$$\iint_{\mathcal{S}(\Gamma)} \varphi_1(P) \sigma dA = \text{constant} \quad [11]$$

where $\varphi_1(P)$ is an arbitrary function and Γ is again arbitrary. This is because each mass element has a single value of P and therefore a single value of $\varphi_1(P)$. Extending $\mathcal{S}(\Gamma)$ to span the whole fluid domain and integrating over all surfaces $\mathcal{S}$, with arbitrary θ -weighting, we obtain

$$\iiint \varphi_2(P, \theta) \sigma dA d\theta = \text{constant} \quad [12]$$

with $\varphi_2(P, \theta)$ another arbitrary function, where the integral is taken over the whole fluid domain. These domain integrals [12] are the Casimir invariants. They are exactly constant for any dissipationless flow whatever.

PV Units and the Extratropical Tropopause

Rossby's original choice of proportionality constant differed from today's standard choice. As noted in his 1940 paper, Rossby chose the physical dimensions of P to be the same as those of ordinary vorticity, namely $(\text{time})^{-1}$ drawing on the analogy with potential temperature. (See text between his eqns [11] and [13].) However, the usual practice today is to tolerate the slightly looser analogy and different physical units implied by [3]–[5], for the sake of having simpler formulae. The standard PV unit used today is $10^{-6} \text{ m}^2 \text{ s}^{-1} \text{ K kg}^{-1}$, abbreviated PVU.

By a strange accident, cross-sections of the atmosphere show P values typically around 2 PVU at the extratropical tropopause, and this has proved extremely useful as a way of defining the tropopause outside a tropical band of latitudes, say outside $\pm 20^\circ$ or so. More precisely, the extratropical tropopause is often marked by steep isentropic gradients of P with values ranging from about 1 to 4 PVU. The shape of the 2-PVU contour in Figure 2, dividing dark blue from light blue, gives no more than a slight hint of the complicated three-dimensional shape of the tropopause, where it intersects the 320 K isentropic surface at the instant shown. The instantaneous tropopause is a highly convoluted surface with an overall poleward–downward slope, so that the white areas in Figure 2 are in the troposphere and the main colored areas are in the stratosphere.

Airborne measuring instruments flown along the 320 K surface and crossing from white through dark blue into light blue and warmer-colored areas would see changes in chemical composition characteristic of the transition from tropospheric to stratospheric air. Indeed, such changes have often been observed in association with finer-scale, filamentary structures of the kind seen in the figure, beginning with the pioneering work of D. W. Waugh and R. A. Plumb in the early 1990s using chemical data from NASA's ER-2 aircraft.

The usefulness of the PV as an extratropical tropopause marker is an accident because, for one thing, it depends on the choice of θ as the thermodynamical material invariant that satisfies [2] and appears in the definitions [3]–[5]. There is no fundamental reason for that choice. Everything in the dynamical theory works just as well with other thermodynamical material

invariants such as the specific entropy, or indeed any other smooth, monotonic function of θ . The PV thus redefined is sometimes called a *modified PV*. Isentropic distributions of P like that in Figure 2 remain the same after such modification, apart from changes to the units and to the numerical values assigned to each color. Notice, however, that the normalizing factors for those changes depend on θ and are therefore different on each isentropic surface.

PV Inversion and Generalized PV

Any flow that can be considered balanced whether geostrophically or at higher accuracy (see **Dynamical Meteorology: Balanced Flow**) satisfies what is now called the *invertibility principle* for PV. The principle says that, to an accuracy limited only by the accuracy of the balance relation, one can capture all the dynamical information about the flow by specifying only the following:

1. the mass under each isentropic surface $\mathcal{S}$,
2. the isentropic distributions of P , on all the surfaces $\mathcal{S}$, and
3. the distributions of θ on the lower boundary and on the upper boundary if present.

By implication there exists, then, a nonlocal diagnostic operator, the *PV inversion operator* associated with the given balance relation. Its input is the foregoing information at some instant. Its output is the remaining dynamical information at the same instant including the p , ρ , T , and $\mathbf{u}$ fields. Very often $\mathbf{u}$ is dominated by its horizontal component, the weaker vertical component nevertheless being dynamically significant thanks to its role in the generalized ballerina effect, and in moving and tilting isentropic surfaces.

The idea of PV inversion is implicit in textbook descriptions of, for instance, the Rossby-wave mechanism. The idea is used at the point in the argument where the horizontal component of $\mathbf{u}$ is deduced diagnostically from the disturbance PV field associated with PV-contour undulations. Sometimes the term *induced velocity*, borrowed from aerodynamics, is used. In this context it means the velocity field deduced from the PV field by inversion.

What are PV inversion operators like, qualitatively? A partial answer is that calculating the horizontal component of $\mathbf{u}$ is like calculating the electric field $\mathbf{E}$ induced by a certain electric-charge distribution, and then taking the horizontal component of $\mathbf{E}$ and rotating it counterclockwise through a right angle, for instance from northward to westward. The electric charges correspond to isentropic anomalies in P and boundary anomalies in θ . Thus, for instance, the positive isentropic anomaly in P over the Balkans in Figure 2 corresponds to a positive electric charge, inducing an outward-pointing $\mathbf{E}$ field and hence a cyclonic or counterclockwise velocity field around it. This provides us with a way of saying what the terms vortex, cyclone, and anticyclone really mean. For instance the vortex over the Balkans, an upper-air cyclone, is nothing but a positive isentropic anomaly in P together with its induced velocity field.

Because of the balance relation, these velocity fields are accompanied by p , ρ , and T fields that to a first approximation satisfy the thermal wind equation; for instance the upper-air cyclone has a warm T anomaly above it and a cold T anomaly beneath. Conversely, an upper-air anticyclone has a cold T

anomaly above, a fact crucial to lower-stratospheric polar ozone chemistry. Flow through such a cold anomaly cannot advect the negative PV anomaly beneath, but can give rise to fast cloud formation and accelerated chemical processing.

Similar statements about vortices apply to the distributions of θ at, say, the lower boundary surface. (In practical terms, taking friction into account, this translates to 'just above the planetary boundary layer.')

A surface cyclone or *heat low* is nothing but a positive, i.e., warm, lower boundary anomaly in θ together with its induced velocity field, and conversely for a surface anticyclone.

Severe cyclonic storms in the extratropical atmosphere often arise from the vertical alignment of warm lower boundary anomalies in θ and positive upper-air isentropic anomalies in P like the large cyclonic anomaly seen on the left of [Figure 2](#). Helped by such vertical alignment, the induced velocities can add up to give storm-force winds. Furthermore, the development of such a situation by upper-air positive- P advection along with near-surface warm advection, and poleward upgliding along sloping isentropes, induces large-scale upward motion. Such upward motion is described by any sufficiently accurate PV inversion operator. Alternatively, it can be computed via the so-called omega equation. The large-scale upward motion may trigger latent heat release, creating or intensifying isentropic anomalies in P . Especially in moist air over the extratropical oceans, the upshot can be the sudden explosive marine cyclogenesis feared and respected by sailors: "Three days from land a great tempest arose..."

It hardly needs saying that, whenever the invertibility principle holds to sufficient accuracy, it gives us a vastly simplified conceptual view of the dynamical evolution. The dynamical system is completely specified by a PV inversion operator together with the remarkably simple prognostic equations [\[1\]](#) and [\[2\]](#) or their diabatic, frictional generalizations. Those equations provide us with the simplest way to cope with the bedrock mathematical difficulty of fluid dynamics, the advective nonlinearity.

Since P and θ are scalar fields, keeping track of them using pictures like [Figure 2](#), actual or mental, is a far simpler task than keeping track of the evolving p , ρ , T , and $\mathbf{u}$ fields in three dimensions, including the nonlocal effects mediated by the p field under the constraints imposed by the balance relation. The nonlocal effects are all encapsulated in the PV inversion operator. The foregoing points, implicit in Rossby's work, were articulated with increasing clarity by Jule G. Charney and Aleksandr M. Obukhov in the late 1940s and by Ernst Kleinschmidt in the early 1950s. They allow us to make sense not only of Rossby-wave propagation, cyclogenesis, and anticyclogenesis but also, for instance, of aerodynamical ideas like vortex rollup – the idea that a strong isentropic anomaly in PV can roll 'itself' up into a nearly circular vortex, as in the Balkans example of [Figure 2](#).

In 1966 Francis P. Bretherton pointed out that an even greater conceptual simplification is possible. The single prognostic equation [\[1\]](#) is enough to determine the dissipationless evolution by itself, provided that we consider the PV field $P(\mathbf{x}, t)$ to contain delta-function contributions at the upper and lower boundaries, with strengths determined by the θ distributions at the boundaries. Ignoring frictional boundary-layer phenomena, we may relate this to the idea that isentropic

surfaces $\mathcal{S}$ intersecting the lower boundary, say, can be imagined to continue along the boundary in an infinitesimally thin layer of infinite $|\nabla\theta|$ hence infinite P . In the electrostatic analogy, surface θ distributions correspond to surface charge distributions – electric charge per unit area rather than per unit volume. The PV field with surface θ distributions included may be called the *generalized PV field*, containing all the information in the second and third numbered items above.

Some Illustrations

The idea of PV inversion can be illustrated in a simple way by considering the theoretical limiting case of infinite sound speed and infinite stable stratification. The buoyancy frequency N and gradient Richardson number both tend to infinity. The isentropic surfaces $\mathcal{S}$ become rigid and horizontal – horizontal in the billiard-table sense, with the sum of the gravitational and centrifugal potentials constant. The balance relation degenerates to a statement that the flow on each $\mathcal{S}$ is strictly horizontal and strictly incompressible. Then, in the rotating frame, we have $\mathbf{u} = \hat{\mathbf{z}} \times \nabla_H \psi$ for some streamfunction ψ , where $\hat{\mathbf{z}}$ is a unit vertical vector, and, from [\[5\]](#),

$$P = \sigma^{-1}(f + \nabla_H^2 \psi) \quad [13]$$

with σ now strictly constant. Here ∇_H is the two-dimensional horizontal gradient operator and ∇_H^2 the corresponding Laplacian, so that $\nabla_H^2 \psi$ is the relative vorticity. We may regard [\[13\]](#) as a Poisson equation to be solved for ψ when P is given. Solving it is a well-defined, and well-behaved, operation, given suitable boundary conditions such that the P field on each $\mathcal{S}$ satisfies [\[8\]](#) with Γ taken as the horizontal domain boundary; see also [\[16\]](#). Symbolically, in the rotating frame,

$$\mathbf{u} = \hat{\mathbf{z}} \times \nabla_H \psi \text{ with } \psi = \nabla_H^{-2}(\sigma P - f) \quad [14]$$

expressing PV invertibility in the limiting case. The PV inversion problem now resembles an electrostatics problem in two, rather than three, dimensions. The charge distribution corresponds to σ times the PV anomaly $(P - \sigma^{-1}f)$, with $-\psi$ in the role of the electric potential. In this limiting case, as in general, PV inversion is a diagnostic, nonlocal operation.

Notice that our limiting case is degenerate in another sense as well. The altitude z now enters the problem only as a parameter. There is no derivative $\partial/\partial z$ anywhere in the problem, either in the horizontal Laplacian or in the material derivative $D/Dt = \partial/\partial t + \mathbf{u} \cdot \nabla$ in [\[1\]](#), with $\mathbf{u}$ strictly horizontal. Not only is the flow layerwise-two-dimensional, but the layers are completely decoupled from each other. For the validity of this picture there is, therefore, an implicit restriction on magnitudes of $\partial/\partial z$, i.e., an implicit restriction on the smallness of vertical scales in the limit, with the further implication that the picture cannot be uniformly valid for all time.

More realistically, when N and Ri are large but finite, and when f is finite, $\partial/\partial z$ reappears in the problem and brings back vertical coupling. The flow remains layerwise-two-dimensional in the sense that notional 'PV particles' move along each isentropic surface $\mathcal{S}$ – see *Impermeability Theorem* below – but the surfaces $\mathcal{S}$ themselves are no longer quite horizontal, nor quite rigid. Aside from the vertical advection that moves and tilts the surfaces $\mathcal{S}$, all the vertical coupling comes from the PV

inversion operator. The two-dimensional inverse Laplacian in [14] is replaced by an inverse elliptic operator that qualitatively resembles a three-dimensional inverse Laplacian when a stretched vertical coordinate Nz/f is used; thus the vertical coupling for flows of horizontal scale L is effective over a height scale of the order of the corresponding Rossby deformation height fL/N .

For finite N and Ri there are tradeoffs between accuracy and simplicity. The mathematically simplest though least accurate three-dimensional PV inversion operator is that arising in the standard Charney–Obukhov quasigeostrophic theory, an asymptotic theory whose approximations are valid away from the equator, for large Ri and small Rossby number $Ro \sim Ri^{-1/2}$, where Ro can be defined as f_0^{-1} times a typical relative-vorticity value with f_0 a constant representative value of the Coriolis parameter f . The price paid for the mathematical simplicity includes resorting to a strange double subterfuge in which, first, we retain only the purely horizontal velocity field $\mathbf{u} = \hat{\mathbf{z}} \times \nabla_H \psi$ even though vertical motion is now significant and, second, abandon P , the exact, Rossby–Ertel PV, which is advected by vertical as well as by horizontal velocities, in favor of a so-called *quasigeostrophic potential vorticity*, q , advected by the horizontal velocity only. For background $\rho = \rho_0(z)$ and $N = N_0(z)$ we may define

$$q = f + \nabla_H^2 \psi + \frac{1}{\rho_0} \frac{\partial}{\partial z} \left(\frac{\rho_0 f_0^2}{N_0^2} \frac{\partial \psi}{\partial z} \right) \quad [15]$$

noting the agreement with [13] in the limit $N_0 \rightarrow \infty$, apart from the factor σ^{-1} . Omission of that factor is part of the subterfuge, making vertical advection implicit. The generalized ballerina effect is now hidden inside the last term of [15]. The isobaric anomalies in T and θ , measuring small displacements and tilting of the isentropic surfaces $\mathcal{S}$, are proportional to $\partial \psi / \partial z$. For instance if $\theta_0(z)$ denotes the background potential temperature, so that $N_0^2(z) = g d \ln \theta_0 / dz$, then we have $\theta - \theta_0(z) = g^{-1} \theta_0 f_0 \partial \psi / \partial z$ within the approximations of the theory.

The most efficient way of describing the relation between q and P is to say that $\nabla_H q$, the local horizontal or isobaric (constant- z) gradient of q , is proportional to $(\nabla_H P)_\theta$, i.e., proportional to the corresponding isentropic gradient of P . Isobaric eddy fluxes of q are correspondingly related to isentropic eddy fluxes of P .

From [15] we see that the electrostatic analogy holds, qualitatively, in three dimensions, with stretched vertical coordinate $N_0 z / f_0$. The electric-charge distribution is $q - f$. This can include Bretherton delta functions. If we impose $\partial \psi / \partial z = 0$ at the lower boundary, for instance, when inverting [15] to get ψ from q , then a delta-function contribution to the last term of [15] can accommodate finite $\partial \psi / \partial z$ just above the boundary, hence a nonvanishing θ anomaly there.

Three-dimensional inversions far more accurate than quasigeostrophic are now being used in weather forecasting as well as in research and development. The most accurate possible PV inversion operators are mathematically complicated because accurate balance relations $\mathbf{u} = \mathbf{u}^B$ are mathematically complicated, as discussed in the article on balanced flow. This difficulty can, however, be sidestepped using the forecast-initialization components of today's numerical data-assimilation technology.

The Quasi-westward Ratchet

The single time derivative acting on the generalized PV field in [1] and [2] exposes another fundamental point about the balanced dynamics. This point is well hidden within the equations expressing Newton's laws of motion in terms of the p , ρ , T , and $\mathbf{u}$ fields. The single time derivative shows for instance why all the different types of Rossby waves, including internal and topographic (surface θ) Rossby waves, exhibit one-way phase propagation. The Earth's rotation imposes a handedness or chirality upon the wave dynamics as seen in the rotating frame. In this regard the Rossby-wave mechanism is quite unlike classical wave mechanisms, where the governing equations always contain even numbers of time derivatives, making the propagation time-reversible.

On the global or planetary scale, P has an isentropic gradient whose sign, in a coarse-grain view, is usually set by the sign of the planetary-scale gradient in f . From the Antarctic to the Arctic, f and P go from large negative to large positive values. Planetary-scale Rossby waves feel this gradient. As a result, they exhibit westward, never eastward, phase propagation relative to the mean flow. And in all cases of Rossby waves, planetary-scale or smaller, the sense of the relative phase propagation is *quasi-westward* – meaning as if westward – defined to be such that high or predominantly high generalized PV values are on the right. Thus, for instance, topographic Rossby waves, dependent on a surface gradient in the Bretherton delta function, propagate with warm surface air on the right where 'warm' is measured by θ .

The same chirality accounts for the ratchet-like, one-way character of related processes such as the self-sharpening of jet streams and the irreversible transport of angular momentum due to the dissipation of Rossby waves in the stratosphere, producing a persistent westward or retrograde mean force there, hence the *gyroscopic pumping* – always poleward and never equatorward – that drives the global-scale stratospheric circulations and chemical transports usually discussed under the headings Brewer–Dobson circulation and wave-driven circulation.

(If a zonally symmetric mean force keeps pushing air westward, then Coriolis effects keep turning it poleward – a persistent mechanical pumping action. The best-known example is Ekman pumping, the special case in which the zonal force happens to be frictional, as in classic spindown.)

PV Mixability and Strong Jets

One of the mechanisms involved in the dissipation of Rossby waves is *wave breaking*, the irreversible deformation of otherwise-wavy PV contours. This definition of breaking is motivated by fundamental results in wave–mean interaction theory, namely the so-called nonacceleration theorems, which are corollaries of Kelvin's circulation theorem applied to initially zonal material contours.

Rossby wave breaking gives rise to the irreversible mixing of PV along the isentropic surfaces $\mathcal{S}$. This can happen on a spectacularly large scale in some cases, as in the wintertime stratospheric surf zone commonly observed. Such mixing is a strongly nonlinear phenomenon and, because it tends to be highly inhomogeneous spatially, with surf zones adjacent to

wavy PV contours, it often lies outside the scope of homogeneous turbulence (spectral cascade) theory. The idea of PV mixing does, however, explain the ubiquity of such quintessentially inhomogeneous phenomena as the strong jet streams observed in the atmosphere and oceans. The jet that flows along the poleward border of the stratospheric surf zone is just one example among many.

A strong jet, in the sense at hand, is nothing but a narrow core of concentrated isentropic gradients of P together with its induced velocity fields. The properties of PV inversion operators ensure that these induced velocity fields are always jet-like, flowing quasi-eastward, i.e., flowing with high PV on the left. For instance, in the westernmost part of [Figure 2](#) a strong jet flows southward over the Atlantic, with its core at the edge of the large colored region corresponding to high-PV stratospheric air. The jet continues around the periphery of that region past Spain toward the British Isles. Maximum wind speeds reach values of the order of 50 m s^{-1} in this case.

Once such a jet structure has formed it has a tendency to be self-sustaining or self-sharpening. The concentrated core gradients form a waveguide or duct for Rossby waves whose dispersion properties make them liable to breaking on one or both flanks of the jet, while leaving the core intact. PV mixing adjacent to the core weakens the surrounding PV gradients and strengthens the core's PV gradients, automatically sharpening or reshaping the core and the jet velocity profile. Mixing across the core is strongly inhibited, thanks to the combined effects of the shear and the core's Rossby-wave quasi-elasticity.

The inhibition applies to chemical tracers as well as to PV. Countless observations of chemical tracers verify this, going back to Edwin F. Danielsen's classic 1968 aircraft observations of nuclear bomb-test debris showing distinct isotopic signatures to either side of a strong tropopause jet core. So a strong jet core can be identified with what is sometimes called a PV barrier but more aptly an eddy-transport barrier, recognizing the complementary role of the shear in the jet flanks first noted in the doctoral thesis work of M. N. Jukes. These phenomena clearly have a role in keeping the stratosphere and troposphere chemically distinct and the tropopause sharp.

The idea that the PV is mixable along the isentropic surfaces $\mathcal{S}$ merits closer examination. In using it we are setting up an analogy with chemical mixing. How far can we push that analogy? Despite its evident power to handle some kinds of strongly nonlinear phenomena, including strong-jet formation, the analogy is not always apt because the PV is not a passive tracer. Self-organizing, dynamically active phenomena like vortex rollup, and vortex merging, illustrate that isentropic anomalies in P can, in some situations, transport themselves against mean isentropic gradients of P , contrary to the mixing idea. Furthermore, there are rotational force fields that can systematically widen the range of P values on a surface $\mathcal{S}$. If we think of isentropic anomalies in P as electric-charge anomalies, this is like pair production. Such rotational force fields include those due to dissipating gravity waves.

Nevertheless, the mixing idea seems to work well in situations such as Rossby wave breaking in which a large-scale flow advects smaller-scale PV anomalies, in a manner that becomes increasingly passive-tracer-like as the large-scale strain or deformation fields shrink the advected scales. Once this advective scale-shrinkage takes hold, it goes exponentially fast

on the timescale of the large-scale straining. The passive-tracer-like behavior is possible because PV inversion is relatively insensitive to small-scale PV anomalies.

Scenarios of PV transport along, rather than across, the moving surfaces $\mathcal{S}$ can remain valid even when [eqns \[1\] and \[2\]](#) are replaced by their diabatic and frictional generalizations. More precisely, P can be regarded as the amount per unit mass of a notional chemical substance consisting of charged particles that are permanently trapped on the moving surfaces $\mathcal{S}$. Net charge is conserved: one can have pair production and mutual annihilation, but no net creation or destruction except where a surface $\mathcal{S}$ intersects a boundary. In this picture the surfaces $\mathcal{S}$ are impermeable to the PV particles even when they are permeable to air undergoing diabatic heating or cooling – a behavior very different from that of a real chemical. The corresponding mathematical statement is sometimes called the *impermeability theorem* for PV.

The theorem is simple to prove, along with the conservation of net charge, by repeating the derivation that led to the flux-form conservation [eqn \[9\]](#) but with arbitrary diabatic heating and external forces included. This reveals first that the resulting equation is still of the form $\partial(\rho P)/\partial t + \nabla \cdot (\rho \mathbf{v} P) = 0$, i.e., that it is still a conservation equation in flux form – there are no source and sink terms – and second that the flux itself, the vector field acted on by the three-dimensional divergence operator, naturally takes a form such that it always represents zero transport across moving surfaces $\mathcal{S}$. Thus the surfaces $\mathcal{S}$ behave as if they were impermeable to the charged particles of PV substance.

Of course one can always make the surfaces $\mathcal{S}$ look permeable by adding an identically nondivergent vector field to the flux. But that is arguably a needless complication, for the reasons discussed in the paper by C. S. Bretherton and C. Schär in the Further Reading list.

It is important to remember when using the analogy with chemicals that P is the amount of PV substance or PV charge per unit mass. It is the chemical mixing ratio, so called, to which P is analogous, not the amount per unit volume. Clearly, an inert chemical lacking sources or sinks can be diluted or concentrated. An extreme example is the formation of tropical cyclones, in which, in terms of the foregoing picture, PV charge is advected inwards along the surfaces $\mathcal{S}$ and greatly concentrated near the center of the cyclone. Although such processes cannot create net PV charge, they can and do create strong isentropic anomalies in P , whose inversion may yield hurricane-force winds.

The Inhomogeneity of PV Mixing

Why does PV mixing have such a strong propensity to be inhomogeneous? Part of the answer has already been indicated, namely the self-organizing properties of strong jets as eddy-transport barriers. One can add that the inhomogeneity reflects not only the dispersion properties of jet-guided Rossby waves, but also, arguably, a generic positive-feedback mechanism sometimes called the 'PV Phillips effect.' It can operate at the earlier stages of self-organization. Wherever large-scale isentropic gradients of P are weakened by PV mixing, Rossby-wave quasi-elasticity is weakened, facilitating further mixing. On the borders of such a region, the gradients are strengthened

and mixing is inhibited. If shear and Rossby-waveguide ducting become important at the borders, then mixing is inhibited still further as eddy-transport barriers form.

There is yet another reason to expect PV mixing to be inhomogeneous. It is especially clear in the case of surfaces $\mathcal{S}$ that span the globe and are therefore topologically spherical, as in the stratosphere and upper troposphere (and also in the solar interior). If we extend the surface integrals in eqn [8] to the entire sphere, there is no enclosing contour Γ and we have

$$\iint_{\mathcal{S}} P \sigma \, dA = 0 \quad [16]$$

stating that on each topologically spherical $\mathcal{S}$ there are equal numbers of positively and negatively charged PV particles, regardless of whether the flow is forced, dissipating, or dissipationless. This is consistent with the charge-conservation and impermeability theorems. The integral relation [16] imposes a severe constraint on the possible evolution of the isentropic distributions of PV on each such $\mathcal{S}$, hence on the possible evolution of the flow. That constraint is enough in itself to make uniform or homogeneous mixing highly improbable, as the following argument shows.

Consider a hypothetical situation in which the mixing is uniform, as if the distribution of P on a surface $\mathcal{S}$ were subject to a uniform horizontal diffusivity. Under the constraint [16], in which σ is strictly positive, the perfectly mixed state toward which the distribution of P would then relax can only be a state in which $P = 0$ everywhere on $\mathcal{S}$. But invertibility says that the entire surface $\mathcal{S}$ would then have to be at rest relative to the stars, apart from oscillations representing imbalance such as sound waves and inertia-gravity waves. In a rapidly rotating system like the Earth's atmosphere, with strong Coriolis effects and Rossby numbers typically small, such a state of rest would be overwhelmingly improbable. It would require a redistribution of angular momentum that would not only have an implausibly large magnitude but would also need to take a very special form.

The Taylor Identity

The hypothetical situation just sketched is an implausible extreme case, but it illustrates another fundamental fact. Almost any isentropic redistribution of PV, or other modification to the PV field, will be accompanied by changes in the distribution of angular momentum.

The PV mixing associated with breaking Rossby waves is just one piece of what might be called a wave-turbulence jigsaw in which wave propagation has just as crucial a role as wave breaking, through wave-induced transport of angular momentum such as that giving rise, as already mentioned, to the gyroscopic pumping of the Brewer-Dobson and other global-scale mean circulations. A by-product is that eddy fluxes of momentum often look antifrictional, exhibiting the so-called 'negative viscosity' that was once regarded as a great enigma of atmospheric science, but is now recognized as a natural consequence of the interplay between wave generation, wave propagation, and wave breaking.

The way in which the jigsaw fits together is reflected in a central result from quasigeostrophic theory, which for

historical reasons might be called the Taylor-Charney-Stern-Bretherton-Eady-Green identity. It is traceable back to a seminal 1915 paper by G. I. Taylor that applies to the limiting case [14]. For brevity it will here be called the Taylor identity. It interrelates the eddy fluxes of momentum and PV. The standard form of the identity is for disturbances to a zonal-mean state. Using overbars and primes to denote the zonal mean and fluctuations about it, which can have arbitrary amplitude, we readily find from [15] that

$$\overline{v'q'} = \frac{1}{\rho_0} \left(\frac{\partial F}{\partial y} + \frac{\partial G}{\partial z} \right) \quad [17]$$

where

$$(F, G) = \rho_0 \left(-\overline{u'v'}, \frac{f_0 g}{N_0^2 \theta_0} \overline{v'\theta'} \right) \quad [18]$$

the so-called Eliassen-Palm (EP) flux or effective stress (minus the effective eddy momentum flux). This quantifies the Rossby-wave-induced momentum transport. Here $(u', v') = (-\partial\psi'/\partial y, \partial\psi'/\partial x)$, the eastward and northward components of $\hat{z} \times \nabla_H \psi'$, and $g\theta' = \theta_0 f_0 \partial\psi'/\partial z$. The vertical component of the EP flux is the same as the pressure-fluctuation-induced form stress defined in oceanography (sometimes less aptly called 'form drag'), the mean zonal force per unit area across an undulating stratification surface, whose vertical displacement is $-g\theta'/N_0^2\theta_0$. The Taylor identity has special importance not least because of its validity for strongly nonlinear flows, such as breaking Rossby waves. No small-amplitude assumption is needed.

For instance, in order to create the wintertime stratospheric surf zone, through PV mixing producing downgradient, i.e., negative, $\overline{v'q'}$, there needs to be a convergence of Rossby-wave activity from outside the surf zone, making the right-hand side of [17] negative as well, and reducing the angular momentum of the surf zone. An exquisitely precise illustration of how everything fits together is provided by the Stewartson-Warn-Warn theory of nonlinear Rossby-wave critical layers. These are narrow surf zones and well illustrate the strong inhomogeneity of the wave-turbulence jigsaw and the typical way in which [17] is satisfied.

Further Reading

- Appenzeller, C., Davies, H.C., Norton, W.A., 1996. Fragmentation of stratospheric intrusions. *Journal of Geophysical Research* 101, 1435–1456. (This paper, the source of Figure 2, fills in many of the details of the associated weather systems and their largely advective evolution. There are important cross-checks from satellite water-vapor imagery.)
- Arbogast, P., Maynard, K., Crepin, F., 2008. Ertel potential vorticity inversion using a digital filter initialization method. *Quarterly Journal of the Royal Meteorological Society* 134, 1287–1296. (This work, to which Dr P. Berrisford kindly drew my attention, uses the weather-forecasting technology at Météo France. The inversions produce three-dimensional velocity fields including vertical velocity. At the time of writing, there was also a web page at www.cnrm.meteo.fr entitled "PV inversion as a tool for synoptic forecasting".)
- Bannon, P.R., Schmidli, J., Schär, C., 2003. On potential vorticity flux vectors. *Journal of Atmospheric Sciences* 60, 2917–2921. (Draws on mathematical generalizations of [9] going back to Ertel and Truesdell to include multiply-buoyant fluids like seawater, for which material invariance fails. Caution: the authors use the terms 'source' and 'sink' in their nonchemical, nonconservative, purely causative sense.)

- Bretherton, C.S., Schär, C., 1993. Flux of potential vorticity substance: a simple derivation and a uniqueness property. *Journal of Atmospheric Sciences* 50, 1834–1836. (The choice of flux vector that satisfies the impermeability theorem is shown to be uniquely simple; it is the only choice whose nonadvective term depends linearly on the diabatic heating rates and rotational, non-potential force fields that break material invariance.)
- Bühler, O., 2009. *Waves and Mean Flows* Cambridge University Press, Cambridge. (By far the best account of wave–mean interaction fundamentals, supplying a wealth of telling examples and bringing out, for instance, the connection between non-acceleration theorems and Kelvin’s circulation theorem.)
- Dritschel, D.G., McIntyre, M.E., 2008. Multiple jets as PV staircases: the Phillips effect and the resilience of eddy-transport barriers. *Journal of Atmospheric Sciences* 65, 855–874. (This discussion of strong jets includes a historical introduction noting the seminal contributions of G. I. Taylor, N. A. Phillips, and R. E. Dickinson that were keys to solving the old ‘negative viscosity’ enigma; the PV Phillips effect is named after a different researcher, O. M. Phillips. One historical correction is needed: Charney and Stern appear to have been the first in print with the quasigeostrophic version [17] of the Taylor identity, in 1962.)
- Haynes, P.H., 1989. The effect of barotropic instability on the nonlinear evolution of a Rossby-wave critical layer. *Journal of Fluid Mechanics* 207, 231–266. (The most comprehensive account of the Stewartson–Warn–Warn theory and its further developments, illustrating in precise detail how the wave–turbulence jigsaw works.)
- Haynes, P.H., Anglade, J., 1997. The vertical-scale cascade of atmospheric tracers due to large-scale differential advection. *Journal of Atmospheric Sciences* 54, 1121–1136. (Shows the robustness of exponentially fast advective scale-shrinkage of passive tracers in stratified, rotating, vertically sheared flows.)
- Hoskins, B.J., McIntyre, M.E., Robertson, A.W., 1985. On the use and significance of isentropic potential-vorticity maps. *Quarterly Journal of the Royal Meteorological Society* 111, 877–946. Corrigendum 113, 402–404. (This major review presents typical examples from the real atmosphere and clarifies why it is isentropic – not horizontal or isobaric – maps, distributions, gradients and fluxes of P that are dynamically significant.)
- Korty, R.L., Schneider, T., 2007. A climatology of the tropospheric thermal stratification using saturation potential vorticity. *Journal of Climate* 20, 5977–5991. (Develops a proposal by K. A. Emanuel to characterize afresh the principal atmospheric air masses by means of a saturation PV, in which θ is replaced by the saturation value of the moist equivalent potential temperature θ_e^* ; q.v. also for references to earlier work on varieties of ‘moist PV.’)
- Mestel, L., 2012. *Stellar Magnetism*, second ed. Oxford University Press, Oxford. (See Chapter 8 on ‘late-type stars,’ which include solar-type stars. Recognizing the implications of PV fundamentals for stratified turbulence has led, indirectly but powerfully, to radical advances in understanding the dynamics of the stratified solar interior and the so-called tachocline.)
- Rossby, C.-G., 1936. Dynamics of steady ocean currents in the light of experimental fluid mechanics. *Pap. Phys. Oceanogr. Meteorol.* (Massachusetts Institute of Technology and Woods Hole Oceanographic Institution) 5 (1), 1–43.
- Rossby, C.-G., 1938. On the mutual adjustment of pressure and velocity distributions in certain simple current systems, II. *Journal of Marine Research* 2, 239–263.
- Rossby, C.-G., 1940. Planetary flow patterns in the atmosphere. *Quarterly Journal of the Royal Meteorological Society* 66 (Suppl.), 68–87. (Rossby’s three great pioneering papers seem almost forgotten today. I thank Jule G. Charney, Norman A. Phillips, George W. Platzman, and Roger M. Samelson for dispelling my historical illusions, little by little.)
- Young, W.R., 2012. An exact thickness-weighted average formulation of the Bousinesq equations. *Journal of Physical Oceanography* 42, 692–707. (This is a major advance in the theory of residual circulations and the Taylor identity; see also transformed Eulerian mean. With the help of judiciously chosen averages on stratification surfaces, not necessarily zonal averages, Young finds exact results that, in the case of the Taylor identity, are formally no more complicated than the standard quasigeostrophic Taylor identity, eqn [17])

Primitive Equations

AA White, University of Surrey, Guildford, UK
N Wood, Met Office, Exeter, UK

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by A A White, volume 2, pp 694–702, © 2003, Elsevier Ltd.

Synopsis

The term ‘primitive equations’ has several meanings in atmospheric science. At its most general it means the basic governing equations of atmospheric dynamics. The discussion therefore starts with these equations and goes on to present three approximate forms of them, focusing on the hydrostatic primitive equations, which is most often meant by ‘primitive equations.’ A key aspect of the approximate forms is that they are internally consistent in the sense that they have natural analogues of important conservation properties of the unapproximated equations.

Introduction

Rather confusingly, the term ‘primitive equations’ has several meanings in atmospheric science. At its most general, it denotes any theoretical or numerical formulation in which velocity components are the chosen flow variables. Newton’s Second Law of Motion relates the rate of change of momentum to the net force acting, and applying it to a (small) mass of fluid gives the momentum equation that lies at the heart of fluid dynamics. Since momentum = mass × velocity, the momentum equation is also an equation for the rate of change of the velocity of the mass of fluid. The spatial components of this vector equation are known as ‘primitive equations,’ the word ‘primitive’ being used in the sense of ‘original.’ For the horizontal flow, one alternative to a primitive description is the use of 2D vorticity and divergence as flow variables.

In a general sense, then, the primitive equations are the components of the momentum equation itself, as distinct from derived forms such as vorticity or divergence equations. The term is often extended to include the other equations needed to consistently determine the time evolution of the flow and thermodynamic fields (as required in numerical weather prediction, for example).

In a much more specific (but widespread) usage, the term ‘primitive equations’ is an abbreviation for ‘the hydrostatic primitive equations with the shallow-atmosphere approximation.’ Early computer models of the large-scale atmosphere were based on geostrophically balanced vorticity equations that do not permit gravity waves. The use of the ‘primitive equations’ was a development of the 1960s, and today most numerical weather prediction and climate simulation models are based on them. Sometimes they are integrated not in their velocity component forms, but in derived vorticity and divergence forms, and here a difficulty of terminology arises: one may hear of ‘the vorticity and divergence form of the primitive equations’ – in apparent contradiction to the definition given earlier. In this usage, ‘primitive equations’ alludes to the inclusion of gravity wave motion, the distinction being made from geostrophically balanced vorticity models in which time evolution of the divergence via a forecasting equation is not allowed and gravity waves cannot occur.

In the late 1960s, E.N. Lorenz commented with characteristic humor on the established use of the term ‘primitive

equations’: “Apparently it was thought unlikely that anyone would use the exact equations, which are more primitive than the primitive equations” (Lorenz, 1967). Nowadays, however, such “more primitive” equations are increasingly being used as the foundation of meteorological models. Accordingly, our account begins (next section) with an outline derivation of the spherical polar components of the vector momentum equation that describes motion in a rotating planetary atmosphere. In the third section, three internally consistent approximate forms are considered, of which one is the hydrostatic primitive equations (HPEs) with the shallow-atmosphere approximation. The latter are the “primitive equations” in the parlance discussed by Lorenz; for brevity we shall call them the ‘HPEs.’ The fourth section is devoted to discussion of the HPEs. Various simpler formulations (much used for theoretical exploration and numerical testing) are discussed in the final section; among the most important of these are the shallow-water equations (SWEs).

The Momentum Equation and Its Spherical Polar Components

Newton’s Second Law of Motion is the main physical basis for forecasting flow in the atmosphere and oceans. For an element of fluid, a convenient expression of it is the vector momentum equation:

$$\frac{D\mathbf{u}}{Dt} = -\frac{1}{\rho}\nabla p + \mathbf{F}. \quad [1]$$

Acceleration PGF Other forces

Here, $\mathbf{u}$ is the velocity of the element of fluid, ρ is its density, and p is pressure; per unit mass, $\mathbf{F}$ is the net force acting, except for the pressure gradient force (PGF) which is given by the term $-(1/\rho)\nabla p$, ∇ being the 3D gradient operator of mathematical physics. D/Dt is the material derivative. It gives the rate of change of its operand following the fluid element, and it is related to the local (time) derivative, $\partial/\partial t$, which gives the rate of change of the operand at a fixed point instantaneously coinciding with the element, by the identity

$$\frac{D}{Dt} = \frac{\partial}{\partial t} + (\mathbf{u} \cdot \nabla). \quad [2]$$

To use eqn [1] to forecast $\mathbf{u}$, it is necessary to know $\mathbf{F}$, ρ and p as well as the present field of $\mathbf{u}$. $\mathbf{F}$ includes gravity and friction. The need to know ρ and p draws thermodynamics and mass continuity into the problem, but here we focus on eqn [1], and assume that $\mathbf{F}$, ρ , and p are known. (The equations used to forecast ρ and p will be briefly noted later on.) The spatial components of eqn [1] are primitive equations in the general sense.

Newton's Second Law applies to velocity and its rate of change measured in an inertial (i.e., nonaccelerating) frame. Relative to a rotating frame, such as that of the Earth, a fluid element may appear to be changing its direction of motion whereas relative to an inertial frame it is not. Effects of this type are precisely allowed for by a simple transformation rule. Let D/Dt be the rate of change seen in an inertial frame, and D/Dt_Ω be the rate of change seen in a frame rotating with the Earth's angular velocity Ω ; then for any vector $\mathbf{A}$,

$$\frac{D\mathbf{A}}{Dt} = \frac{D\mathbf{A}}{Dt_\Omega} + \Omega \times \mathbf{A}. \quad [3]$$

A special and important case in which eqn [3] is transparently correct (see Figure 1) is $\mathbf{A} = \mathbf{r}$ = position vector relative to the Earth's center: the velocities $\mathbf{u}$ ($= D\mathbf{r}/Dt$) and $\mathbf{u}_\Omega$ ($= D\mathbf{r}/Dt_\Omega$) measured in the inertial and rotating frames are related by

$$\mathbf{u} = \mathbf{u}_\Omega + \Omega \times \mathbf{r}. \quad [4]$$

By setting $\mathbf{A} = \mathbf{u}$ in eqn [3] and using eqn [4] on the right-hand side, we find that

$$\frac{D\mathbf{u}}{Dt} = \frac{D\mathbf{u}_\Omega}{Dt_\Omega} + 2\Omega \times \mathbf{u}_\Omega + \Omega \times (\Omega \times \mathbf{r}). \quad [5]$$

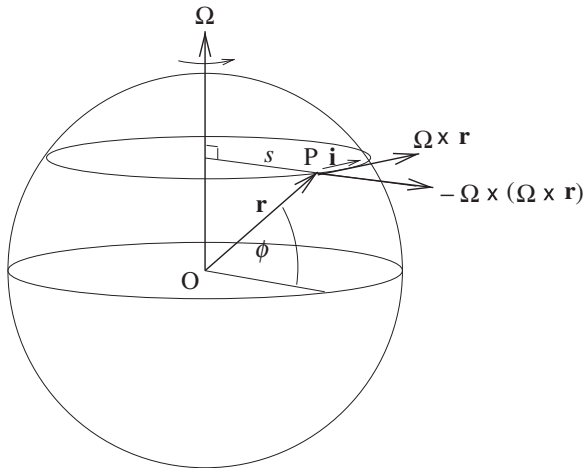


Figure 1 Depicting the Earth's rotation vector Ω , position vector $\mathbf{r}$ of a point P , and two important associated vector products. Point P rotates with the Earth, and $\mathbf{r}$ is defined relative to its center O ; $\mathbf{i}$ is unit vector in the zonal (west $\rightarrow$ east) direction at P , whose latitude is ϕ . The perpendicular distance of P from the rotation axis is $s = |\mathbf{r}| \cos \phi$, so the velocity of P relative to the inertial frame in which O is fixed is $\mathbf{i}|\Omega|s = \mathbf{i}|\Omega||\mathbf{r}| \cos \phi = \Omega \times \mathbf{r}$. Hence the velocity, $\mathbf{u}$, relative to the inertial frame, of a fluid element that is instantaneously at P but moving relative to the Earth with velocity $\mathbf{u}_\Omega$, is simply $\mathbf{u} = \mathbf{u}_\Omega + \Omega \times \mathbf{r}$. The centrifugal force $-\Omega \times (\Omega \times \mathbf{r})$, per unit mass, acts radially outwards in the plane (containing P) perpendicular to Ω . The Coriolis force (not shown) also lies in this plane.

This equation is a great step forward because it allows us to write Newton's Second Law (eqn [1]) in terms of the velocity $\mathbf{u}_\Omega$ and the acceleration $D\mathbf{u}_\Omega/Dt_\Omega$ measured in the rotating frame:

$$\frac{D\mathbf{u}_\Omega}{Dt_\Omega} = \underbrace{-2\Omega \times \mathbf{u}_\Omega}_{\text{Coriolis}} - \underbrace{\Omega \times (\Omega \times \mathbf{r})}_{\text{Centrifugal}} - \underbrace{\frac{1}{\rho} \nabla p}_{\text{PGF}} + \underbrace{\mathbf{F}}_{\text{Other forces}}. \quad [6]$$

The Coriolis and centrifugal forces (per unit mass – a stipulation that will be taken as read from now on) are as indicated beneath eqn [6]. Their presence enables us to work – and think – in terms of quantities measured or defined relative to the rotating Earth, while still respecting Newton's laws of motion.

The Coriolis force, $-2\Omega \times \mathbf{u}_\Omega$ in eqn [6], acts in planes perpendicular to the Earth's rotation axis, and depends on the component of the relative velocity $\mathbf{u}_\Omega$ in such planes, but not on the component of $\mathbf{u}_\Omega$ parallel to the Earth's rotation axis.

The centrifugal force, $-\Omega \times (\Omega \times \mathbf{r})$ in eqn [6], also acts in planes perpendicular to the Earth's rotation axis, but depends on perpendicular distance s (see Figure 1) and not at all on $\mathbf{u}_\Omega$. It acts outwards, and can be expressed as the gradient of a scalar potential:

$$-\Omega \times (\Omega \times \mathbf{r}) = \nabla(\Omega^2 s^2/2). \quad [7]$$

The term $\mathbf{F}$ in eqn [6] consists of Newtonian gravity, which can be represented as the (down) gradient of a scalar potential, say Φ_{Newt} , and the remaining force $\mathbf{F}_{\text{Rem}}$ (mainly friction):

$$\mathbf{F} = -\nabla\Phi_{\text{Newt}} + \mathbf{F}_{\text{Rem}}. \quad [8]$$

It is usual to define apparent (or normal) gravity $-\mathbf{g}\mathbf{k} = -\nabla\Phi_{\text{App}}$ as the vector sum of Newtonian gravity and centrifugal force (as given by eqn [7]). Here, g is the magnitude of apparent gravity, $\mathbf{k}$ is unit vector in the direction of apparent vertical, and Φ_{App} is the apparent gravitational potential:

$$\begin{aligned} -\mathbf{g}\mathbf{k} &= -\nabla\Phi_{\text{Newt}} - \Omega \times (\Omega \times \mathbf{r}) = -\nabla(\Phi_{\text{Newt}} - \Omega^2 s^2/2) \\ &= -\nabla\Phi_{\text{App}}. \end{aligned} \quad [9]$$

g is the quantity measured by a pendulum or a gravity meter, and apparent vertical is the (upward) direction indicated by a plumb line – all observations being made in the frame of the rotating Earth. Equation [6] may now be written as

$$\frac{D\mathbf{u}_\Omega}{Dt_\Omega} = -2\Omega \times \mathbf{u}_\Omega - \mathbf{g}\mathbf{k} - \frac{1}{\rho} \nabla p + \mathbf{F}_{\text{Rem}}. \quad [10]$$

It is illuminating to consider orders of magnitude of some of the terms in eqn [10].

The magnitude, g , of apparent gravity varies slightly with position in the atmosphere, but its order of magnitude value is 10 m s^{-2} . At latitude ϕ (see Figure 2), the centrifugal force contribution has magnitude $\Omega^2 r \cos \phi$; its components parallel and perpendicular to Newtonian gravity are (very nearly) $\Omega^2 r \cos^2 \phi$ and $\Omega^2 r \sin \phi \cos \phi$. Taking $\Omega = 2\pi/1 \text{ day} = 7 \times 10^{-5} \text{ s}^{-1}$ and $r = 6.4 \times 10^6 \text{ m}$ gives $\Omega^2 r = 3.2 \times 10^{-2} \text{ m s}^{-2}$: the contribution of centrifugal force to apparent gravity is at least 300 times smaller than the Newtonian contribution. Similarly, the deflection of the vertical due to centrifugal effects is less than $1/300$ radians in the atmosphere (the maximum value being attained at $\pm 45^\circ$ latitude, where $\sin \phi \cos \phi$ has its extreme values).

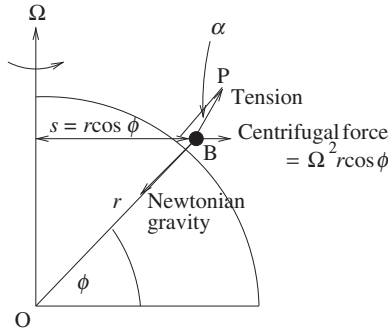


Figure 2 A unit-mass pendulum bob B, at rest relative to the rotating Earth, is in equilibrium under Newtonian gravity, centrifugal force, and the tension in the (short) string by which it is suspended from point P at latitude ϕ . To a good approximation, the angular deflection α of the string PB from the radius OP (O = Earth's center) is given by the representation $g\alpha = \Omega^2 r \cos \phi \sin \phi$ of the balance between the meridional components of tension and centrifugal force.

An upper bound to the magnitude of the Coriolis force is $2\Omega|\mathbf{u}_\Omega|$; taking $|\mathbf{u}_\Omega| \approx 50 \text{ m s}^{-1}$ – a large value for the troposphere and stratosphere – gives $2\Omega|\mathbf{u}_\Omega| \approx 0.7 \times 10^{-2} \text{ m s}^{-2}$. The Coriolis force is thus typically much smaller even than the centrifugal force – but to neglect it would be disastrous. In planes perpendicular to apparent vertical, i.e., in horizontal planes, apparent gravity has no component; the component of the Coriolis force in this plane can be very important, with the major balance of forces being between the pressure gradient and Coriolis forces. This geostrophic balance is characteristic of extratropical weather systems having horizontal scales of 1000 km or more, and is crucial to their spatial configuration and time evolution.

Having derived eqn [10] as the rotating-frame version of the momentum equation, we drop the subscripts Ω : from now on, $\mathbf{u}$ and D/Dt signify the velocity and the rate of change as measured in the rotating system. The form we shall work with is

$$\frac{D\mathbf{u}}{Dt} = -2\boldsymbol{\Omega} \times \mathbf{u} - g\mathbf{k} - \frac{1}{\rho}\nabla p + \mathbf{F}_{\text{Rem.}} \quad [11]$$

Because of centrifugal effects, the surfaces of constant apparent gravitational potential Φ_{App} are essentially nonspherical, but – as we have noted – the contribution of centrifugal force to apparent gravity is small. The usual procedure in meteorological dynamics (Lorenz, 1967; Phillips, 1973; Gill, 1982; Holton, 1992) is (1) to separate the components of eqn [11] perpendicular and parallel to apparent vertical $\mathbf{k}$, but (2) then to treat the surfaces of constant Φ_{App} (to which $\mathbf{k}$ is perpendicular) as if they were spheres. Step (2) is known as the spherical geopotential approximation. We follow this procedure here, referring the reader to van der Toorn and Zimmerman (2008) and White et al. (2008) for further discussion and examination of spheroidal approximations (which are less restrictive and in particular allow the variation of the magnitude of gravity with latitude to be represented; this is not allowed in the spherical case since a spurious source of vorticity would be implied).

In the spherical geopotential approximation it is natural to adopt a spherical polar coordinate system. What are the components of eqn [11] in this curved, non-Cartesian system?

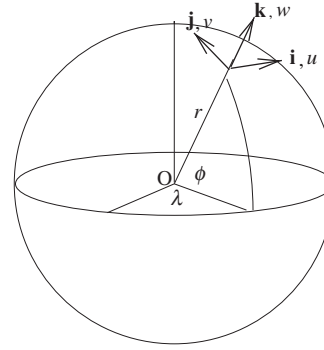


Figure 3 Spherical polar coordinates (λ, ϕ, r) of a point P relative to Earth's center O as origin. The local unit vector triad $\mathbf{i}, \mathbf{j}, \mathbf{k}$ and associated velocity components u, v, w are also shown.

Suppose that λ is longitude, ϕ is latitude, r is distance from the Earth's center, O , and $\mathbf{i}, \mathbf{j}$, and $\mathbf{k}$ are the local triad of unit vectors associated with λ, ϕ , and r ; see Figure 3. If the zonal, meridional, and vertical components of $\mathbf{u}$ are u, v , and w , i.e.,

$$\mathbf{u} = u\mathbf{i} + v\mathbf{j} + w\mathbf{k}, \quad [12]$$

then

$$\frac{D\mathbf{u}}{Dt} = \frac{Du}{Dt}\mathbf{i} + \frac{Dv}{Dt}\mathbf{j} + \frac{Dw}{Dt}\mathbf{k} + u\frac{D\mathbf{i}}{Dt} + v\frac{D\mathbf{j}}{Dt} + w\frac{D\mathbf{k}}{Dt}. \quad [13]$$

The terms involving the material derivatives of $\mathbf{i}, \mathbf{j}$, and $\mathbf{k}$ may be re-expressed by observing that the rotation rate, relative to the Earth, of the unit vector triad as it follows the flow, is

$$\boldsymbol{\Omega}_T = -\frac{v}{r}\mathbf{i} + \frac{u}{r}\mathbf{j} + \frac{u \tan \phi}{r}\mathbf{k}. \quad [14]$$

See Figure 4. Application of a rule similar to eqn [3] gives

$$\frac{D\mathbf{i}}{Dt} = \boldsymbol{\Omega}_T \times \mathbf{i}, \quad \frac{D\mathbf{j}}{Dt} = \boldsymbol{\Omega}_T \times \mathbf{j}, \quad \frac{D\mathbf{k}}{Dt} = \boldsymbol{\Omega}_T \times \mathbf{k}. \quad [15]$$

Hence eqn [13] can be written simply as

$$\frac{D\mathbf{u}}{Dt} = \frac{Du}{Dt}\mathbf{i} + \frac{Dv}{Dt}\mathbf{j} + \frac{Dw}{Dt}\mathbf{k} + \boldsymbol{\Omega}_T \times \mathbf{u}. \quad [16]$$

By noting that the Earth's rotation vector $\boldsymbol{\Omega}$ may be expressed as

$$\boldsymbol{\Omega} = \Omega(\mathbf{j} \cos \phi + \mathbf{k} \sin \phi), \quad [17]$$

and by using eqns [12], [14], and [16], one obtains the zonal ($\mathbf{i}$), meridional ($\mathbf{j}$), and vertical ($\mathbf{k}$) components of eqn [11] as:

$$\begin{aligned} \frac{Du}{Dt} = & -\frac{uv}{r} + \frac{uv}{r} \tan \phi + 2\Omega v \sin \phi - 2\Omega w \cos \phi \\ & - \frac{1}{\rho r \cos \phi} \frac{\partial p}{\partial \lambda} + F_\lambda, \end{aligned} \quad [18]$$

$$\frac{Dv}{Dt} = -\frac{vw}{r} - \frac{u^2}{r} \tan \phi - 2\Omega u \sin \phi - \frac{1}{\rho r} \frac{\partial p}{\partial \phi} + F_\phi, \quad [19]$$

$$\frac{Dw}{Dt} = \frac{u^2}{r} + \frac{v^2}{r} + 2\Omega u \cos \phi - g - \frac{1}{\rho} \frac{\partial p}{\partial r} + F_r. \quad [20]$$

The terms on the left-hand sides of eqns [18]–[20] are the rates of change of the velocity components. The terms quadratic in

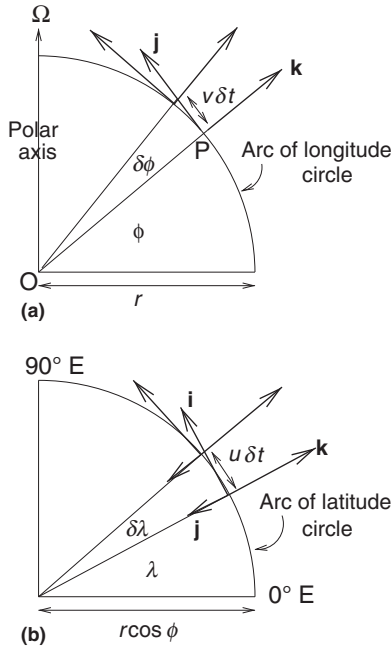


Figure 4 (a) A meridional plane through the instantaneous position P of the unit vector triad $\mathbf{i}, \mathbf{j}, \mathbf{k}$. Displacement $v\delta t$ in time δt due to meridional motion leads to rotation of the unit vector triad through an angle $\delta\phi = \frac{v\delta t}{r}$ about the local zonal axis, and thus to a rotation rate $-\frac{v}{r}\mathbf{i}$. (b) A plane perpendicular to the Earth's rotation axis and passing through the instantaneous position P of the unit vector triad $\mathbf{i}, \mathbf{j}, \mathbf{k}$; note that $\mathbf{i}$ lies in the plane of the diagram, but $\mathbf{j}$ and $\mathbf{k}$ do not. Displacement $u\delta t$ in time δt due to zonal motion corresponds to rotation of the unit vector triad through an angle $\delta\lambda = \frac{u\delta t}{r \cos \phi}$ about the Earth's rotation axis, and thus to a rotation rate $\frac{u}{r \cos \phi}(\mathbf{j} \cos \phi + \mathbf{k} \sin \phi) = \frac{u}{r}\mathbf{j} + \frac{u \tan \phi}{r}\mathbf{k}$ (unit vector along the Earth's rotation axis being $\mathbf{j} \cos \phi + \mathbf{k} \sin \phi$).

the velocity components and having $1/r$ factors are the metric terms. They are of two types: those reflecting the intrinsic uniform curvature of a sphere (the curvature of great circles) and those – involving $\tan \phi$ – reflecting the curvature of the spherical polar coordinate system (with its circles of constant latitude, of which only the Equator is a great circle). The terms involving Ω are the Coriolis terms. They too divide into two sets: those proportional to $\Omega \sin \phi$ and those proportional to $\Omega \cos \phi$; from eqn [17], $\Omega \sin \phi$ and $\Omega \cos \phi$ are, respectively, the vertical and horizontal components of the Earth's rotation vector Ω at latitude ϕ . The vertical component equation, eqn [20], contains the entire contribution of (apparent) gravity. Next come the components of the PGF in the (λ, ϕ, r) system, and finally the components of the remaining force per unit mass: $\mathbf{F}_{\text{Rem}} = F_\lambda \mathbf{i} + F_\phi \mathbf{j} + F_r \mathbf{k}$.

In our spherical polar coordinate system (recall Figure 3), the relation eqn [2] between the material derivative D/Dt and the local time derivative $\partial/\partial t$ becomes specifically

$$\frac{D}{Dt} = \frac{\partial}{\partial t} + \frac{u}{r \cos \phi} \frac{\partial}{\partial \lambda} + \frac{v}{r} \frac{\partial}{\partial \phi} + w \frac{\partial}{\partial r}. \quad [21]$$

Equations [18]–[20] are primitive equations in the general sense discussed in the Introduction. They are subject to neither

the hydrostatic nor the shallow-atmosphere approximations, and are not to be confused with the HPEs discussed in later sections.

Various conservation properties are embodied in or implied by eqns [18]–[20]. Upon noting that $w = Dr/Dt$ and $v = rD\phi/Dt$, the zonal component eqn [18] can be written

$$\frac{D}{Dt} [(u + \Omega r \cos \phi) r \cos \phi] = -\frac{1}{\rho} \frac{\partial p}{\partial \lambda} + F_\lambda r \cos \phi. \quad [22]$$

This is a conservation law for absolute axial angular momentum; the terms on the right represent torques due to the zonal components of the PGF and other forces.

A conservation relation for the sum of kinetic and potential energy per unit mass is obtained by multiplying eqns [18]–[20] by the corresponding components of $\mathbf{u}$, adding the results, and noting eqn [9]:

$$\frac{D}{Dt} \left(\frac{1}{2} \mathbf{u}^2 + \Phi_{\text{App}} \right) = \frac{1}{\rho} \mathbf{u} \cdot \nabla p + \mathbf{u} \cdot \mathbf{F}_{\text{Rem}}. \quad [23]$$

Two further important conservation relations may be established by introducing the thermodynamic and (mass) continuity equations in the forms

$$\rho c_v \frac{DT}{Dt} = -p \nabla \cdot \mathbf{u} + \rho Q \quad \text{and} \quad \frac{D\rho}{Dt} = -\rho \nabla \cdot \mathbf{u}. \quad [24]$$

Here T is temperature, c_v is the specific heat of air at constant volume, Q is the diabatic heating rate per unit mass, and perfect gas behavior is assumed, i.e., $p = \rho RT$, R being the gas constant per unit mass of air. $\nabla \cdot \mathbf{u}$, the divergence of $\mathbf{u}$, is given by

$$\nabla \cdot \mathbf{u} \equiv \frac{1}{r^2 \cos \phi} \left[\frac{\partial u}{\partial \lambda} + \frac{\partial}{\partial \phi} (v \cos \phi) \right] + \frac{1}{r^2} \frac{\partial}{\partial r} (r^2 w). \quad [25]$$

The conservation equation for total energy (i.e., kinetic + potential + internal) is

$$\rho \frac{D}{Dt} \left(\frac{1}{2} \mathbf{u}^2 + \Phi_{\text{App}} + c_v T \right) = -\nabla \cdot (p \mathbf{u}) + \rho (Q + \mathbf{u} \cdot \mathbf{F}). \quad [26]$$

The conservation equation for potential vorticity is

$$\rho \frac{D}{Dt} \left[\frac{1}{\rho} (\mathbf{Z} \cdot \nabla \theta) \right] = \mathbf{Z} \cdot \nabla \left(\frac{D\theta}{Dt} \right) + \nabla \theta \cdot (\nabla \times \mathbf{F}). \quad [27]$$

Potential vorticity is the scalar quantity in the square brackets on the left-hand side of eqn [27], i.e., $(\mathbf{Z} \cdot \nabla \theta)/\rho$. The vector $\mathbf{Z} \equiv 2\Omega + \nabla \times \mathbf{u}$ is the absolute vorticity and θ is the potential temperature defined as

$$\theta \equiv T \left(\frac{p_{\text{ref}}}{p} \right)^{\frac{R}{c_p}}, \quad [28]$$

c_p being the specific heat of air at constant pressure, and p_{ref} a reference pressure (usually taken as 1000 hPa). In the spherical polar system

$$\mathbf{Z} = Z_\lambda \mathbf{i} + Z_\phi \mathbf{j} + Z_r \mathbf{k}, \quad [29]$$

where

$$Z_\lambda = \frac{1}{r} \left[\frac{\partial w}{\partial \phi} - \frac{\partial}{\partial r} (rv) \right], \quad [30]$$

$$Z_\phi = \frac{1}{r} \left[\frac{\partial}{\partial r} (ru) - \frac{1}{\cos \phi} \frac{\partial w}{\partial \lambda} \right] + 2\Omega \cos \phi, \quad [31]$$

$$Z_r = \frac{1}{r \cos \phi} \left[\frac{\partial v}{\partial \lambda} - \frac{\partial}{\partial \phi} (u \cos \phi) \right] + 2\Omega \sin \phi. \quad [32]$$

From eqn [27] may be deduced the conservation of the potential vorticity under conditions in which the right-hand side vanishes, in particular in flow that is frictionless ($F = 0$) and adiabatic ($D\theta/Dt = 0$, which implies $Q = 0$, since $c_p D\theta/Dt = (\theta/T)Q$).

Three Internally Consistent Approximate Models

The conservation laws for energy, angular momentum, and potential vorticity should be respected when the governing equations are being approximated. One way of ensuring this is to observe that the conservation laws stem from the Hamiltonian structure of the governing equations; making approximations within that structure will then deliver individual forecasting equations (such as the components of the momentum equation) that imply consistent approximate forms of the conservation laws. Roulstone and Brice (1995) give details of this approach, which provides a perspective on all other approaches. Here, we shall note without derivation those combinations of approximation within the components of the momentum equation that imply consistent modifications of the conservation laws.

In general, there are two distinct types of approximation: (1) of the dynamics; and (2) of the geometric framework within which the dynamics occurs. In our context, the relevant modification of the dynamics (1) is omission of the term Dw/Dt from eqn [20]; this is an approximation of hydrostatic type – useful when modeling the compressible atmosphere because it only slightly affects relevant timescales and space scales, but prevents the occurrence of computationally demanding acoustic waves having a vertical component of propagation (and minimal meteorological significance). The relevant geometric modification (2) is replacement of everyday Euclidean space, having spherical polar volume element $r^2 \cos \phi d\lambda d\phi dr$ by a (weakly) non-Euclidean, spherical-shell space having volume element $a^2 \cos \phi d\lambda d\phi dr$, where a is a constant representative value of r . This is a shallow-atmosphere approximation. In the hydrostatic and shallow-atmosphere scenario there are three consistent approximations (Staniforth, 2000; White et al., 2005).

- If the shallow-atmosphere approximation is not made, but Dw/Dt and F_r (and only these terms) are omitted from eqn [20], the quasi-hydrostatic equations described by White and Bromley (1995) result.
- If the shallow-atmosphere approximation is made, and both Dw/Dt and F_r are omitted from eqn [20], the result is the HPEs:

$$\frac{Du}{Dt} = \frac{uv}{a} \tan \phi + 2\Omega v \sin \phi - \frac{1}{\rho a \cos \phi} \frac{\partial p}{\partial \lambda} + F_\lambda, \quad [33]$$

$$\frac{Dv}{Dt} = -\frac{u^2}{a} \tan \phi - 2\Omega u \sin \phi - \frac{1}{\rho a} \frac{\partial p}{\partial \phi} + F_\phi, \quad [34]$$

$$g + \frac{1}{\rho} \frac{\partial p}{\partial z} = 0. \quad [35]$$

Here, z = height above mean sea level, and the material derivative is now given by

$$\frac{D}{Dt} = \frac{\partial}{\partial t} + \frac{u}{a \cos \phi} \frac{\partial}{\partial \lambda} + \frac{v}{a} \frac{\partial}{\partial \phi} + w \frac{\partial}{\partial z}, \quad [36]$$

with $u = a \cos \phi D\lambda/Dt$, $v = a D\phi/Dt$, and $w = Dz/Dt$. See Lorenz (1967).

- A third consistent model is obtained if the shallow-atmosphere approximation is made in eqn [20], but Dw/Dt is retained (though with the definition eqn [36] of D/Dt). This is the nonhydrostatic, shallow-atmosphere (NHS) model of Tanguay et al. (1990). Equations [33], [34], and [36] are unchanged but eqn [35] becomes

$$\frac{Dw}{Dt} = -g - \frac{1}{\rho} \frac{\partial p}{\partial z} + F_r. \quad [37]$$

Each of these three approximate models implies conservation properties for energy, angular momentum, and potential vorticity that are natural analogues of those of the unapproximated equations; details for the HPEs are given in the next section.

Comparison of eqns [33]–[35] with eqns [18]–[20] shows that a substantial number of terms have disappeared in the transition to the HPEs. An important feature of both the HPEs and the NHS equations is the absence of the Coriolis terms proportional to $\cos \phi$. The complete Coriolis force may be written (from eqn [17]) as

$$2\Omega \times \mathbf{u} = 2\Omega(\mathbf{k} \times \mathbf{u} \sin \phi + \mathbf{j} \times \mathbf{u} \cos \phi), \quad [38]$$

but the term $2\Omega \mathbf{j} \times \mathbf{u} \cos \phi$ does not survive in the HPEs and the NHS equations. This omission – sometimes referred to as the traditional approximation – amounts to a change in the direction of the Coriolis force as well as in its magnitude. Whereas the true Coriolis force acts perpendicular to the Earth's rotation axis, the shallow-atmosphere Coriolis force acts in local horizontal planes. Further, part of the true Coriolis force that acts in local horizontal planes is omitted in the shallow-atmosphere version; and at the Equator the omission amounts to neglect of the entire horizontal component of the true Coriolis force. If the Coriolis terms proportional to $\cos \phi$ are retained within a shallow-atmosphere framework, then conservation of both angular momentum and potential vorticity are compromised (White et al., 2005). The omission of these Coriolis terms is therefore regarded as part of the shallow-atmosphere approximation.

The metric terms omitted in the HPEs and the NHS model are those not involving $\tan \phi$, i.e., those representing the curvature of the sphere rather than the curvature of the spherical polar coordinate system. In the triad rotation rate Ω_T (see eqn [14]), r is replaced by a and only the vertical component remains. The geometric character of the shallow-atmosphere approximation is to neglect the curvature of great circles; the curvature of small circles (such as latitude circles except the equator) is retained, but slightly underestimated. An equivalent statement is that the divergence of the Earth's radii is neglected. Either statement emphasizes the non-Euclidean character of the shallow-atmosphere approximation; see Zdunkowski and Bott (2003), Chapter 18, for further discussion.

A further consequence of the shallow-atmosphere approximation (that therefore applies to both the HPEs and the NHS model) is that gravity is not permitted to vary with height as otherwise there would be a spurious source of divergence.

The HPEs

The HPEs consist of eqns [33] and [34] (with eqn [36]) as forecasting equations for u and v , eqn [35] as a diagnostic relation between ρ and p , and equations for forecasting ρ and p and diagnosing w . (A diagnostic equation contains no time derivatives.) The appropriate forecasting equations for ρ and p are the continuity and thermodynamic equations written as

$$\frac{D\rho}{Dt} = -\rho \nabla \cdot \mathbf{u}, \quad c_v \frac{Dp}{Dt} = -p c_p \nabla \cdot \mathbf{u} + \rho R Q. \quad [39]$$

Here, D/Dt and $\nabla \cdot \mathbf{u}$ are both shallow-atmosphere forms: eqn [36] gives D/Dt , and

$$\nabla \cdot \mathbf{u} \equiv \frac{1}{a \cos \phi} \left[\frac{\partial u}{\partial \lambda} + \frac{\partial}{\partial \phi} (v \cos \phi) \right] + \frac{\partial w}{\partial z} \quad [40]$$

is recognizable as the shallow-atmosphere version of eqn [25].

The need for diagnostic calculation of w (which contributes to $\nabla \cdot \mathbf{u}$ via eqn [40], and also appears in the material derivative eqn [36]) arises because the prognostic equation eqn [20] (or eqn [37]) for w has been replaced by the diagnostic equation eqn [35] of hydrostatic balance. An expression for w (in terms of u , v , ρ , and p , their derivatives and certain integrals) may be obtained from the condition that the two eqns [39] always forecast values of ρ and p that satisfy eqn [35]. That expression, derived (before 1922) by L.F. Richardson and known as Richardson's equation, is not given here. In physical terms, w adopts the configuration that maintains hydrostatic balance as the fields of u , v , ρ , and p evolve in time; see White (2002).

Calculating w complicates the use of the HPEs in numerical time integration. The compensating advantage (as already noted) is that the omission of Dw/Dt prevents the occurrence of acoustic waves that have a vertical component of propagation. Such acoustic waves are the main mode of motion permitted by the unapproximated equations (as applied to a perfect gas) but not by the HPEs. Lamb waves – horizontally propagating compressibility waves – are permitted by the HPEs, but may or may not be present, depending on the boundary conditions applied in the vertical. Rossby (planetary) waves are present, as are gravity waves modified by rotation; both external and internal types are permitted. A range of special equatorial modes is also present (see Gill, 1982). All of these modes of motion are to some extent modified by the hydrostatic approximation, but only short gravity waves significantly so.

Temperature T may be found at any time from the perfect gas equation $p = \rho RT$, and from it (via eqn [28]) the potential temperature θ . Different forms of the HPEs may be derived by using T , θ or specific volume $\alpha (=1/\rho)$ as forecasting quantities rather than ρ and p .

What are the conditions for applicability of the HPEs? The replacement of r by a requires that the depth of the atmosphere should be small compared to the Earth's mean radius (≈ 6370 km). Since 90% of the mass of the atmosphere lies

below 17 km, this requirement is well satisfied. Other conditions relate to the terms omitted from the original component equations. Scale analysis (White and Bromley, 1995) suggests that the Coriolis terms in $\cos \phi$ are negligible (in comparison with terms retained in the HPEs) if $2\Omega H \cos \phi / U \ll 1$, H being a height scale of the motion and U a typical horizontal flow speed; substituting representative values gives $2\Omega H / U \approx 10^{-1}$. Linearized adiabatic analyses suggest that the $\cos \phi$ Coriolis terms are negligible if $2\Omega \ll N$, where $N \equiv [(g/\theta)d\theta/dz]^{1/2}$ is the buoyancy frequency. Since $2\Omega \sim 10^{-4} \text{ s}^{-1}$ and $N \sim 10^{-2} \text{ s}^{-1}$, this condition, typically, is well satisfied at least above the planetary boundary layer.

More tangibly, the term $-2\Omega w \cos \phi$ in eqn [18] describes the conservation of absolute axial angular momentum in axisymmetric vertical motion, which can lead to changes of zonal flow speed of up to 2 m s^{-1} over the depth of the tropical troposphere; this effect is not described by the HPEs. The omission of the $\cos \phi$ Coriolis terms is the most striking approximation of the HPEs as applied to large-scale motion – particularly diabatically forced motion in the tropics – but there is, as yet, no evidence to suggest that their inclusion makes significant difference in numerical weather prediction and climate simulation. Their inclusion has, however, been found to have a significant effect on the depth of the neutrally stratified, planetary boundary layer as modeled by large-eddy simulations (Mason and Thomson, 1987). The advantages of retaining the $\cos \phi$ Coriolis terms (either by using eqns [18]–[20] or the quasi-hydrostatic model noted in Section Three Internally Consistent Approximate Models) are that justification for their exclusion does not then have to be supplied, that the Coriolis force is accurately and completely represented, and that the shallow-atmosphere approximation is simultaneously abandoned; the geometric framework becomes (once again) Euclidean, and intuition based on everyday experience of 3D space can be safely exercised.

The criterion most often quoted for the neglect of the term Dw/Dt is $H \ll L$, L being a horizontal length scale of the motion; this is well satisfied by many flow phenomena, for example midlatitude weather systems, for which $H/L \sim 10^{-2}$. A more discriminating criterion (which covers cases in which $H/L \sim 1$, but the neglect of Dw/Dt is still a good approximation) is that the Lagrangian timescale τ of the motion should be large compared with that of vertical buoyancy oscillations (which are essentially nonhydrostatic): $\tau \gg N^{-1}$. Thus the HPEs require for their validity that N should be large compared with some threshold value, both as regards the $\cos \phi$ Coriolis terms and Dw/Dt ; they do not remain valid as $N \rightarrow 0$. This situation makes use of the original eqns [18]–[20] attractive and was a motivation for the inclusion of a representation of the $\cos \phi$ Coriolis terms in the numerical simulations of the neutrally stratified planetary boundary layer by Mason and Thomson (1987).

The horizontal component eqns [33] and [34] of the HPEs may be written in various useful alternative forms. A vector form, in terms of the horizontal flow $\mathbf{v} = (u, v, 0)$ is

$$\frac{\partial \mathbf{v}}{\partial t} = -\nabla_z \left(\frac{\mathbf{v}^2}{2} \right) - w \frac{\partial \mathbf{v}}{\partial z} - (\zeta + f) \mathbf{k} \times \mathbf{v} - \frac{1}{\rho} \nabla_z p + \mathbf{F}_h. \quad [41]$$

Here $f \equiv 2\Omega \sin \phi$ (and is customarily called the Coriolis parameter), $\mathbf{F}_h \equiv F_\lambda \mathbf{i} + F_\phi \mathbf{j}$, and the horizontal gradient

operator, ∇_z , and the vertical component of vorticity, ζ , are given by

$$\nabla_z \equiv \left(\frac{1}{a \cos \phi} \frac{\partial}{\partial \lambda}, \frac{1}{a} \frac{\partial}{\partial \phi}, 0 \right), \quad [42]$$

$$\zeta \equiv \frac{1}{a \cos \phi} \left[\frac{\partial v}{\partial \lambda} - \frac{\partial}{\partial \phi} (u \cos \phi) \right]. \quad [43]$$

Forecasting equations for the vorticity, ζ (see eqn [43]), and the horizontal divergence, δ , given by

$$\delta = \nabla_z \cdot \mathbf{v} \equiv \frac{1}{a \cos \phi} \left[\frac{\partial u}{\partial \lambda} + \frac{\partial}{\partial \phi} (v \cos \phi) \right], \quad [44]$$

may be obtained by taking the horizontal curl and divergence of eqn [41]. The resulting equations (not given) are the vorticity-divergence forms of the HPEs. Other useful forms of eqns [33] and [34] involve $u \cos \phi$ and $v \cos \phi$ under the material derivative; these quantities have better transformation properties on the sphere than u and v have.

The HPEs imply good conservation properties in terms of the shallow-atmosphere versions eqn [36] of D/Dt and the ∇ operator (see, e.g., eqn [40]). Axial angular momentum conservation follows from the zonal component eqn [33]:

$$\frac{D}{Dt} [(u + \Omega a \cos \phi) a \cos \phi] = -\frac{1}{\rho} \frac{\partial p}{\partial \lambda} + F_\lambda a \cos \phi \quad [45]$$

(cf eqn [22]). Total energy (kinetic + potential + internal) obeys

$$\rho \frac{D}{Dt} \left(\frac{1}{2} \mathbf{v}^2 + gz + c_v T \right) = -\nabla \cdot (p \mathbf{u}) + \rho (Q + \mathbf{v} \cdot \mathbf{F}), \quad [46]$$

(cf eqn [26]) and the potential vorticity law is

$$\rho \frac{D}{Dt} \left(\frac{\mathbf{Z} \cdot \nabla \theta}{\rho} \right) = \mathbf{Z} \cdot \nabla \left(\frac{D\theta}{Dt} \right) + \rho \nabla \theta \cdot (\nabla \times \mathbf{F}_h), \quad [47]$$

(cf eqn [27]). In eqn [47], $\mathbf{Z}$ is the 3D HPE shallow-atmosphere absolute vorticity vector given by

$$\mathbf{Z} = -\frac{\partial v}{\partial z} \mathbf{i} + \frac{\partial u}{\partial z} \mathbf{j} + (f + \zeta) \mathbf{k}. \quad [48]$$

(cf eqns [29]–[32]). From eqn [47] follows the conservation of the HPE potential vorticity, $(\mathbf{Z} \cdot \nabla \theta)/\rho$, in flow that is frictionless ($\mathbf{F}_h = 0$) and adiabatic ($D\theta/Dt = 0$). Neither the HPE specific kinetic energy $\frac{1}{2} \mathbf{v}^2$ nor the horizontal components of the HPE absolute vorticity $\mathbf{Z}$ involves the vertical velocity w ; $\mathbf{Z}$ also omits the contribution $2\Omega \cos \phi \mathbf{j}$ of the horizontal component of the Earth's rotation to the true absolute vorticity.

The consequences of eqn [35] are extensive and profound. As already noted, acoustic waves having a vertical component of propagation do not occur, and the vertical velocity w must be found diagnostically. In addition, ρ can be eliminated in favor of $\partial p / \partial z$ if desired. More important, pressure becomes an attractive choice of vertical coordinate, a 1:1 correspondence between p and z being assured because ρ (and g) are necessarily positive quantities.

Use of pressure as vertical coordinate simplifies the HPEs when considered purely as a set of partial differential equations (independently of spatial boundary conditions). The horizontal gradient term $(1/\rho) \nabla_z p$ in eqn [41] becomes $\nabla_p (gz)$, where ∇_p is the horizontal gradient operator on pressure

surfaces and z (now a dependent variable) is the height of a pressure surface at any time and horizontal location. The major simplification, however, is that the continuity equation, eqn [39], becomes a simple diagnostic relationship:

$$\nabla_p \cdot \mathbf{v} + \frac{\partial \omega}{\partial p} = 0. \quad [49]$$

Here, $\omega \equiv Dp/Dt$ is the pressure-coordinate 'vertical velocity' and

$$\nabla_p \cdot \mathbf{v} \equiv \frac{1}{a \cos \phi} \left[\frac{\partial}{\partial \lambda} \bigg|_p u + \frac{\partial}{\partial \phi} \bigg|_p (v \cos \phi) \right]. \quad [50]$$

The material derivative eqn [36] is given in 'pressure coordinates' by

$$\frac{D}{Dt} = \frac{\partial}{\partial t} \bigg|_p + \frac{u}{a \cos \phi} \frac{\partial}{\partial \lambda} \bigg|_p + \frac{v}{a} \frac{\partial}{\partial \phi} \bigg|_p + \omega \frac{\partial}{\partial p}, \quad [51]$$

and involves ω rather than w . In eqns [50] and [51], all differentiations with respect to λ , ϕ , and t are taken at constant pressure, and the derivative $\partial/\partial p$ is taken at constant λ , ϕ , and t .

From eqn [49], using appropriate boundary conditions, the 'vertical velocity' ω may be found by integration over p ; the procedure parallels the derivation of Richardson's equation in height coordinates (see above), but the functions involved are fewer and simpler. Against this must be weighed the complication that pressure varies in space and time (even on horizontal surfaces); consequently, before the integration over p to determine ω can be started, computational effort has to be expended in finding the pressure, p_s , at the Earth's surface.

In the pressure-coordinate system, the thermodynamic equation is conveniently rewritten as a forecasting equation for temperature T :

$$\frac{DT}{Dt} = \frac{RT}{\rho c_p} \omega + \frac{Q}{c_p}. \quad [52]$$

In broad outline, the time integration procedure (as applied to appropriately discretized forms of the equations, and assuming required initial values are known) is as follows:

1. forecast u , v , and T using eqns [33], [34], and [52];
2. find surface pressure p_s and then ω from eqn [49];
3. find z from p_s , the terrain height z_s , and the hydrostatic equation, eqn [35], in the form

$$\frac{\partial z}{\partial p} = -\frac{gp}{RT}; \quad [53]$$

4. repeat (1)–(3) until the desired forecast time or simulation length has been attained.

The complications involved in step (2) of this procedure (see earlier remarks) are significantly reduced if, instead of p , a terrain-following coordinate such as $\sigma = p/p_s$ is used; see Figure 5 for orientation, and Haltiner and Williams (1981) for mathematical formulation. As well as p and σ , many other quantities may be used as vertical coordinates; examples are θ and hybrids that behave as p , σ , and θ (say) in different height

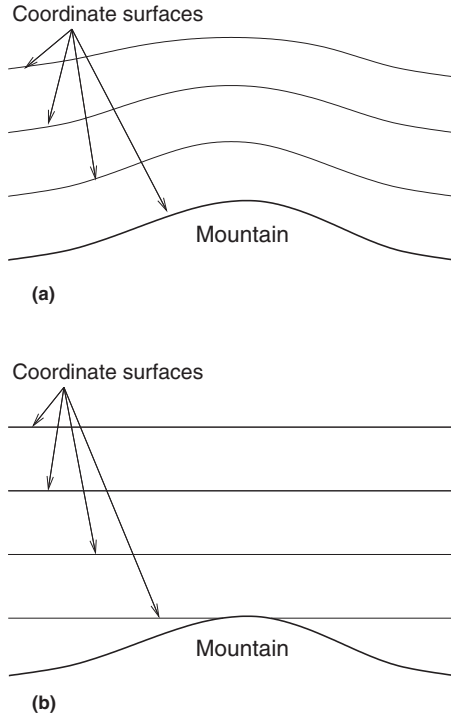


Figure 5 Vertical sections through mountainous terrain showing the behavior of vertical coordinate surfaces of systems which (a) are, (b) are not, terrain-following. Intersection of coordinate surfaces with the Earth's surface, as seen in (b), complicates the construction and operation of numerical models based on non-terrain-following systems. In most respects, the use of terrain-following coordinate systems, as shown in (a), is more convenient.

ranges. See Kasahara (1974), Simmons and Burridge (1981), and Zhu et al. (1992). In all cases, the equations (before discretization) are exact transforms of the HPEs in height coordinates.

Further Simplifications

As is essential for numerical weather prediction and climate simulation, the HPEs treat the 3D structure of the atmosphere's flow and thermodynamic fields. A simpler set of 2D equations, useful as a theoretical and numerical test bed, is the SWEs. These describe the depth (h) variations and depth-independent motion of a shallow layer of incompressible fluid covering a gravitating sphere. The (horizontal) momentum components of the SWEs are

$$\frac{Du}{Dt} = \frac{uv}{a} \tan \phi + fv - \frac{g}{a \cos \phi} \frac{\partial h}{\partial \lambda} + F_\lambda, \quad [54]$$

$$\frac{Dv}{Dt} = -\frac{u^2}{a} \tan \phi - fu - \frac{g}{a \partial \phi} \frac{\partial h}{\partial \phi} + F_\phi, \quad [55]$$

and the (depth-integrated) continuity equation is

$$\frac{Dh}{Dt} = -\frac{h}{a \cos \phi} \left(\frac{\partial u}{\partial \lambda} + \frac{\partial}{\partial \phi} (v \cos \phi) \right). \quad [56]$$

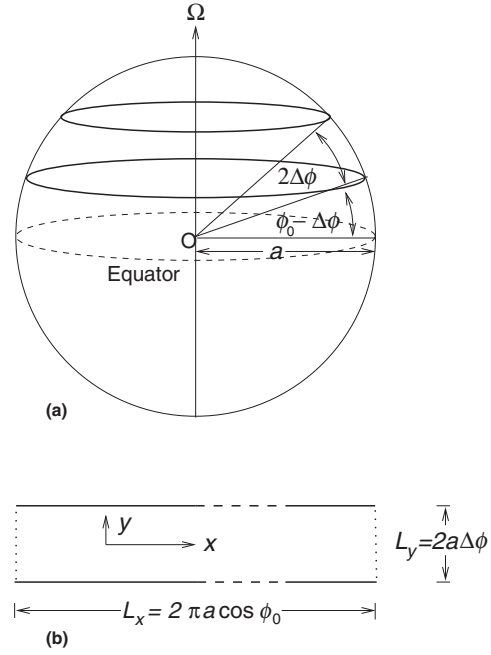


Figure 6 A zone between latitudes $\phi = \phi_0 \pm \Delta\phi$ on a sphere of radius a (a), and a Cartesian 'channel' domain corresponding to it (b). In the channel domain, cyclic conditions in x are applied; boundary conditions in y are more problematic, since no physical boundaries exist at $\phi = \phi_0 \pm \Delta\phi$ in the atmosphere. The channel is usually not a precise mapping of the zone, but is considered as a crude approximation to it. Adopted widths L_y are often somewhat greater than is strictly justifiable, the objective being to obtain theoretical or numerical indications of atmospheric behavior that may then be pursued in more realistic geometrical environments.

In eqns [54]–[56], the material derivative is simply

$$\frac{D}{Dt} = \frac{\partial}{\partial t} + \frac{u}{a \cos \phi} \frac{\partial}{\partial \lambda} + \frac{v}{a} \frac{\partial}{\partial \phi}. \quad [57]$$

Equations [54] and [55] are similar to the horizontal components of the HPEs (especially the p -coordinate forms) except that no vertical variations are allowed; similarly, eqn [56] is comparable to the HPE continuity equation, eqn [39], with vertical variations disallowed.

So far as they are able, the SWEs mimic the rich variety of wave types that are found in the HPEs. They describe depth-independent (barotropic) Rossby waves and surface gravity waves, and have special tropical modes; indeed, from the time of Laplace, studies of the SWEs have helped to elucidate the physics of wave motion in the atmosphere and oceans.

When qualitative rather than quantitative results are sought (as in exploratory theoretical work and numerical testing) the HPEs and the SWEs are conveniently used in simplified forms that neglect the metric terms and replace the spherical polar geometric factors by Cartesian ones. For example, the SWEs may be used in the following forms:

$$\left(\frac{\partial}{\partial t} + u \frac{\partial}{\partial x} + v \frac{\partial}{\partial y} \right) u = fv - g \frac{\partial h}{\partial x} + F_x, \quad [58]$$

$$\left(\frac{\partial}{\partial t} + u \frac{\partial}{\partial x} + v \frac{\partial}{\partial y} \right) v = -fu - g \frac{\partial h}{\partial y} + F_y, \quad [59]$$

$$\left(\frac{\partial}{\partial t} + u \frac{\partial}{\partial x} + v \frac{\partial}{\partial y}\right)h = -h \left(\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y}\right). \quad [60]$$

Here, $x \equiv a\lambda \cos \phi_0$ and $y \equiv a(\phi - \phi_0)$ are zonal and meridional coordinates in a Cartesian plane; as shown in [Figure 6](#), the domain represents a neighborhood of latitude ϕ_0 and the full 2π range of longitude in a ‘channel model.’ The latitude variation of the Coriolis parameter f is neglected in the ‘ f -plane’ approximation. In the ‘ β -plane’ approximation, f is represented as a linear function of y : $f = 2\Omega \sin \phi_0 + 2\Omega(\phi - \phi_0) \cos \phi_0 \equiv f_0 + \beta y$. An ‘equatorial β -plane’ has $\phi_0 = 0$ and hence $f_0 = 0$, $\beta = 2\Omega/a$.

Cartesian forms are also locally valid in a naïve tangent plane approximation for motions whose scale is very small compared with the Earth’s radius. Interestingly, the neglect of curvature in all these Cartesian approximations restores the underlying geometry to Euclidean form; only through the shallow-atmosphere approximation, as represented in spherical geometry in the HPEs (see third section, above) did non-Euclidean effects enter our account.

See also: **Dynamical Meteorology:** Balanced Flow; Coriolis Force; Hamiltonian Dynamics; Kinematics; Overview; Potential Vorticity; Static Stability; Vorticity; Waves. **Gravity Waves:** Buoyancy and Buoyancy Waves: Optical Observations; Buoyancy and Buoyancy Waves: Theory.

Further Reading

Gill, A.E., 1982. *Atmosphere–Ocean Dynamics*. Academic Press, San Diego.
 Haltiner, G.J., Williams, R.T., 1981. *Numerical Prediction and Dynamic Meteorology*. Wiley, New York.

Holton, J.R., 1992. *An Introduction to Dynamic Meteorology*. Academic Press, San Diego.
 Kasahara, A., 1974. Various vertical coordinate systems used for numerical weather prediction. *Monthly Weather Review* 102, 509–522.
 Lorenz, E.N., 1967. *The Nature and Theory of the General Circulation of the Atmosphere*. World Meteorological Organization Technical Note No. 218, TP 115. WMO, Geneva.
 Mason, P.J., Thomson, D.J., 1987. Large-eddy simulations of the neutral-static-stability planetary boundary layer. *Quarterly Journal of the Royal Meteorological Society* 113, 413–443.
 Phillips, N.A., 1973. Principles of large scale numerical weather prediction. In: Morel, P. (Ed.), *Dynamical Meteorology*. Reidel, Dordrecht, pp. 3–96.
 Roulstone, I., Brice, S., 1995. On the Hamiltonian formulation of the quasi-hydrostatic equations. *Quarterly Journal of the Royal Meteorological Society* 121, 927–936.
 Simmons, A.J., Burridge, D.M., 1981. Energy and angular momentum conserving vertical finite differencing schemes and hybrid vertical coordinates. *Monthly Weather Review* 109, 758–766.
 Staniforth, A., 2000. Developing efficient unified nonhydrostatic models. In: Spekat, A. (Ed.), *Proceedings of the Symposium on the 50th Anniversary of NWP*, Potsdam, Germany, 9–10 March 2000. European Meteorological Society.
 Tanguay, M., Robert, A., Laprise, R., 1990. A semi-implicit semi-Lagrangian fully compressible regional forecast model. *Monthly Weather Review* 118, 1970–1980.
 van der Toorn, R., Zimmerman, J.T.F., 2008. On the spherical approximation of the geopotential in geophysical fluid dynamics and the use of a spherical coordinate system. *Geophysical & Astrophysical Fluid Dynamics* 102, 349–371.
 White, A.A., 2002. A view of the equations of meteorological dynamics and various approximations. In: Roulstone, I., Norbury, J. (Eds.), *Large-Scale Atmosphere–Ocean Dynamics*, vol. I. Cambridge University Press, Cambridge, pp. 1–100.
 White, A.A., Bromley, R.A., 1995. Dynamically consistent quasi-hydrostatic equations for global models with a complete representation of the Coriolis force. *Quarterly Journal of the Royal Meteorological Society* 121, 399–418.
 White, A.A., Hoskins, B.J., Roulstone, I., Staniforth, A., 2005. Consistent approximate models of the global atmosphere: shallow, deep, hydrostatic, quasi-hydrostatic and non-hydrostatic. *Quarterly Journal of the Royal Meteorological Society* 131, 2081–2107.
 White, A.A., Staniforth, A., Wood, N., 2008. Spheroidal coordinate systems for modelling global atmospheres. *Quarterly Journal of the Royal Meteorological Society* 134, 261–270.
 Zdunkowski, W., Bott, A., 2003. *Dynamics of the Atmosphere: A Course in Theoretical Meteorology*. Cambridge University Press, Cambridge.
 Zhu, Z., Thuburn, J., Hoskins, B.J., Haynes, P.H., 1992. A vertical finite-difference scheme based on a hybrid σ – θ – p coordinate. *Monthly Weather Review* 120, 851–862.

Quasigeostrophic Theory

HC Davies and H Wernli, Institute for Atmospheric and Climate Science, Swiss Federal Institute of Technology (ETH), Zürich, Switzerland

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

Synoptic- and larger-scale atmospheric flow in the extratropics is almost geostrophic in character, indicating an approximate balance between the Coriolis and pressure gradient forces. The concept of quasigeostrophy (QG) encapsulates elegantly and simply the dynamics whereby the flow retains this character, although both the maintenance of geostrophy and the existence of other key flow features require a nongeostrophic flow component.

Consideration is given to the QG concept's observational foundation, its mathematical formulation, and the interpretation of its subtle physics. In addition, a cursory case study analysis is provided to demonstrate the essence and nature of QG dynamics and the utility of the QG perspective.

Introduction

The concept of quasigeostrophic flow is a cornerstone for the study of synoptic and larger-scale atmospheric flow in the extratropics. It is founded upon two seemingly contradictory deductions regarding the nature of such flows: the flow is observed to be almost geostrophic, and yet both the maintenance of geostrophy and the existence of other key features of synoptic weather systems require a nongeostrophic flow component.

This paradox is resolved by the concept of quasigeostrophy (QG). It is encapsulated in two simple and elegant mathematical formulations that prescribe, respectively, the evolution of the primary 'geostrophic' flow and the form of an accompanying secondary 'ageostrophic' flow that is forced by the primary field. It is the net of these two flow components that constitutes QG flow.

Here, we successively illustrate the concept's observational foundation, tabulate its mathematical formulation, and interpret its subtle physics. Finally, a cursory case study analysis is provided to demonstrate the essence and nature of QG dynamics and the utility of the QG perspective.

Foundation

Governing Equations

For synoptic-scale flow of the free atmosphere away from the Earth's surface, the conservation equations for horizontal momentum, mass, and thermodynamic energy can be written in the following approximate form:

$$D_H \mathbf{V} / Dt + w \partial \mathbf{V} / \partial z = -f(\mathbf{k} \wedge \mathbf{V}) - \nabla_H(p^* / \rho_0) \quad [1]$$

$$\nabla_H \cdot \mathbf{V} + \mathcal{L}w = 0 \quad [2]$$

$$D_H(g\theta^* / \Theta) / Dt + N^2 w = H \quad [3]$$

The corresponding approximate form for the hydrostatic relationship is given by

$$\partial(p^* / \rho_0) / \partial z = (g\theta^* / \Theta) \quad [4]$$

Here, t stands for time, and $\mathbf{k}$ is the vertical unit vector. The flow variables $(\mathbf{V}, w)$ represent, respectively, the horizontal vector velocity and the vertical velocity. The thermodynamic variables $(p, \rho, \text{ and } \theta)$ denote, respectively, the pressure, density, and

potential temperature. For the latter three variables, the zero subscript refers to a time-mean and horizontally averaged measure of the atmosphere's vertical structure, whereas the starred superscript signifies the deviation away from that mean state. In addition, ∇_H is the horizontal gradient operator, $f = 2\Omega \sin \varphi$ is the Coriolis parameter (with Ω denoting the Earth's angular velocity and φ the latitude), N is the mean state buoyancy frequency such that $N^2 = g d(\ln \theta_0) / dz$, Θ is a reference potential temperature, and H is the net diabatic heating rate. Also, $\mathcal{L}$ is a differential operator, such that $\mathcal{L}A = (1/\rho_0) \partial(\rho_0 A) / \partial z$, and in the incompressible limit this reduces to $\mathcal{L}A = \partial A / \partial z$. Finally, D_H / Dt is the rate of change following the horizontal motion of an air parcel, and it is related to the local rate of change at a fixed point by the following expression:

$$D_H / Dt = \{ \partial / \partial t + \mathbf{V} \cdot \nabla_H \} \quad [5]$$

The momentum equation (eqn [1]) states that a change in the horizontal velocity of an air parcel results essentially from an imbalance between two forces: the Coriolis force due to the Earth's rotation and the horizontal pressure gradient force. The approximated equation for mass continuity (eqn [2]) states that expansion in the horizontal cross-section of an air parcel induced by divergence of the horizontal wind field is compensated by a vertical contraction accomplished by the vertical motion. The thermodynamic equation (eqn [3]) states that a change in the potential temperature of an air parcel following the horizontal motion can be induced in two ways: a vertical transport of the potential temperature, and a diabatic contribution that is the net of contributions related primarily to latent heat release, radiative transfer, and turbulent diffusion.

Character of Geostrophic Flow

For synoptic- and larger-scale atmospheric flow, the dominant terms in the horizontal momentum equation (eqn [1]) are usually the Coriolis term and the pressure gradient. If an exact balance prevailed between these two forces, then it would follow that

$$f(\mathbf{k} \wedge \mathbf{V}) = -\nabla_H(p^* / \rho_0)$$

In such a state of so-called geostrophic balance, the flow would be aligned perpendicular to the horizontal pressure gradient (i.e., along the contours of constant pressure), and its

strength would be proportional to the pressure gradient. Alternatively, for a given pressure field, this equation can be used to define a 'geostrophic flow,' and the difference between it and the observed flow would provide a measure of the departure from geostrophy.

Assuming for the moment a single value ($f = f_0$) for the Coriolis parameter, then the equation for the geostrophic flow field ($\mathbf{V} = \mathbf{V}_G$) and the hydrostatic relationship eqn [4] can be written in the following compact form:

$$\mathbf{V}_G = \mathbf{k} \wedge \nabla_H \psi, \quad \text{and} \quad g\theta^*/\Theta = f_0 \partial \psi / \partial z \quad [6]$$

where $\psi = (p^*/f_0\rho_0)$ is referred to as the geostrophic stream function. A striking inference is that the spatial distribution of ψ (or, equivalently, p^*) defines both the geostrophic velocity field and the potential temperature field.

There are other notable features to geostrophic flow. The geostrophic vorticity ζ_G and geostrophic divergence D_G are given, respectively, by

$$\zeta_G = (\nabla_H \wedge \mathbf{V}_G) = \nabla_H^2 \psi \quad [7]$$

$$D_G = (\nabla_H \cdot \mathbf{V}_G) = 0 \quad [8]$$

Also, from eqns [6] and [7], the vertical variations of the geostrophic flow and of the vorticity are given by

$$\partial \mathbf{V}_G / \partial z = (g/f_0\Theta)(\mathbf{k} \wedge \nabla_H \theta^*) \quad [9]$$

$$\partial \zeta_G / \partial z = (g/f_0\Theta) \nabla_H^2 \theta^* \quad [10]$$

A simple interpretation can be given to eqns [7] and [10] by noting that, in the neighborhood of a region where the scalar variable A has a local extremum,

$$\nabla_H^2 A \propto -A$$

Thus, eqn [7] indicates that the geostrophic vorticity is positive in a localized region of low pressure (and vice versa), and eqn [10] that the vertical gradient of vorticity is positive in a localized region of colder air (and vice versa). Equation [8] equates the horizontal divergence of the geostrophic flow to zero, and thereby emphasizes the purely horizontal character of geostrophic flow. Equation [9] is a thermal wind relationship indicating that the vertical variation of the geostrophic wind is aligned such that the potential temperature is relatively cold (warm) on its left (right) flank.

Nature of Synoptic-Scale Flow

Figures 1 and 2 indicate the extent to which the atmosphere's synoptic-scale flow both conforms to the geostrophic interrelationships (i.e., eqns [6–10]) and is consistent with the aforementioned physical interpretations. The displayed synoptic charts are for a portion of the Atlantic sector and are valid at 1200 UTC 31 December 2006. At this time, an incipient low-pressure system (i.e., an extratropical cyclone) was developing on a surface baroclinic zone located southeast of Newfoundland, while aloft a deepening trough was approaching the low.

Figure 1(a) and 1(c) shows the geopotential height and velocity fields on the 300 and 700 hPa levels, respectively. The displayed patterns are in broad accord with the geostrophic relationship (eqn [6]), with the flow aligned approximately along the iso-height contours and with its strength roughly

inversely proportional to the local contour spacing. However, significant departures from geostrophy (white contours) are evident. These regions tend to be smaller in spatial scale and irregular in structure, but there are some coherent features, for instance those associated with strong flow curvature located within and at the base of the major troughs at the upper and lower levels.

Figure 1(b) shows the wind difference between the 700 and 300 hPa levels plus the isentropes at the intermediate level of 500 hPa. In this case, the patterns are in broad qualitative accord with the thermal wind relationship (eqn [9]), with the wind difference field aligned approximately along the isentropes.

Figure 2(a) and 2(c) displays the relative vorticity and the divergence fields on, respectively, the 300 and 700 hPa surfaces. First, consider the vorticity patterns. These patterns, viewed in conjunction with the height fields displayed in Figure 1(a) and 1(c), are broadly consistent with the inverse relationship between the vorticity and pressure fields (eqn [7]). In particular at the 300 hPa level, there is an enhanced band of vorticity linked to the trough in the height field, and at the 700 hPa level, there is a maximum of vorticity in the vicinity of the cyclone.

Again, comparison of the vorticity patterns in Figure 2(a) with the contemporaneous potential temperature field at 500 hPa (Figure 1(b)) indicates that, in qualitative accord with eqn [10], the increase of the vorticity with height beneath the aforementioned strong band of vorticity at the upper level corresponds to a relatively colder region in the midtroposphere.

Comparison of the vorticity and divergence fields in Figure 2(a) and 2(c) suggests that, in accord with eqns [7] and [8], the amplitude of the divergence fields is indeed significantly less than that of the counterpart vorticity fields. Nevertheless, the divergence fields are decidedly nonzero, and this is indicative of a nongeostrophic flow component. Again, the divergence fields tend to be smaller in scale and fragmented in structure. There is a distinct and coherent region of divergence at the upper level ahead of the trough over the Atlantic Ocean. Likewise, at lower levels, there are patches of strong convergence in the neighborhood of the Low, and moreover the aforementioned region of divergence aloft surmounts a fragmented band of convergence at a lower level that is aligned ahead of the cold front. This reversal in sign with height connotes ascent between the 700 and 300 hPa levels (see eqn [2]), and it is thereby consistent with the patterns of the mid-tropospheric vertical velocity field and surface rainfall displayed in Figure 2(b). In assessing this qualitative agreement between vertical velocity and rainfall, note that the former quantity is instantaneous, whereas the latter is integrated over the preceding 6 h.

Thus, inspection of the synoptic-scale fields confirms the flow's predominantly 'geostrophic' character, but it also points to the existence of more fragmented but nevertheless coherent 'nongeostrophic' flow features. This duality is important on two counts. First, a synoptic weather system characterized by a purely geostrophic flow is not compatible with the occurrence of cloud and rain because these processes generally require a nonzero vertical velocity field. Indeed, the precipitation patterns displayed in Figure 2(b) correspond well with the vertical velocity and divergence fields. Second, there is an enigma associated with the sustainment of the geostrophic nature of the atmospheric fields. Equations [1] and [3] are

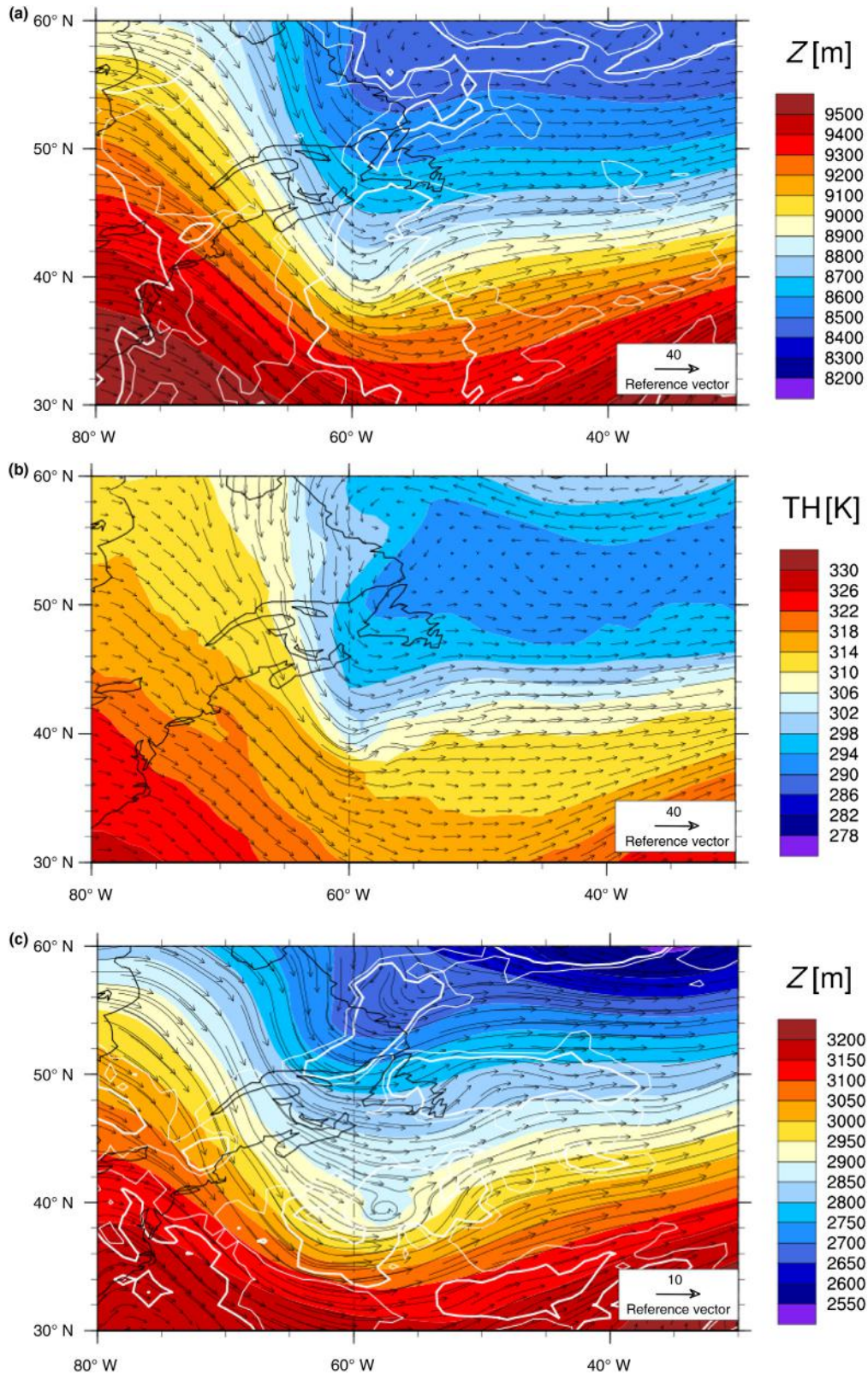


Figure 1 Synoptic charts for a portion of the North Atlantic valid at 12 UTC 31 December 2006, at which time a low-pressure system was developing south of Newfoundland. The upper and lower panels are for, respectively, the 300 and 700 hPa pressure surfaces (i.e., altitudes of ~8 and 3 km), and they indicate the flow pattern (see the representative wind vector in units of m s^{-1}) and the geopotential height field (a proxy for the pressure distribution at an appropriate altitude), plus the departure from geostrophy (thin and bold white contours corresponding to 20 and 30% deviations from geostrophy). The middle panel depicts the potential temperature field at the 500 hPa level and the thermal wind field vectors corresponding to $\{\mathbf{v}(300) - \mathbf{v}(700)\}$.

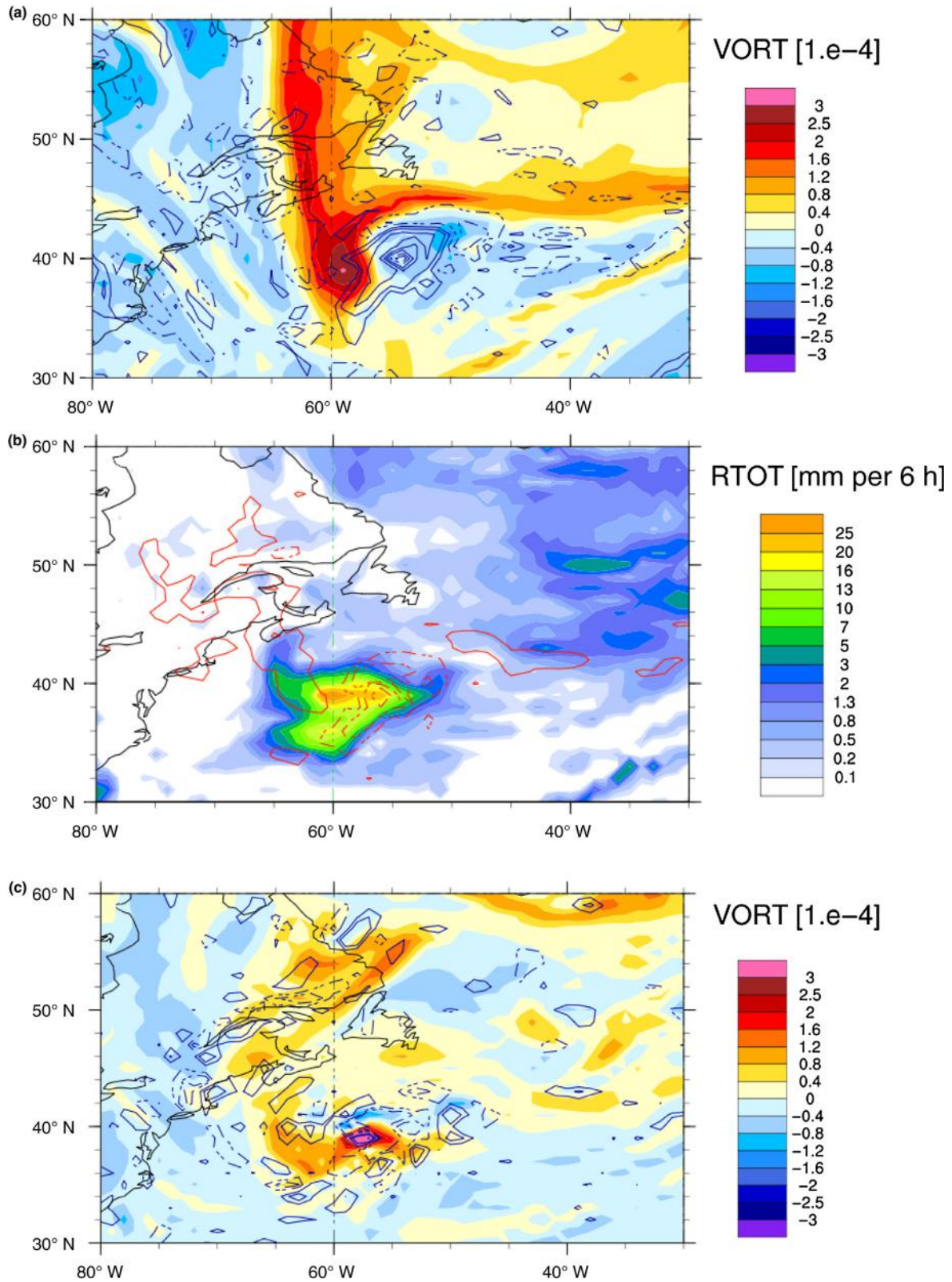


Figure 2 Counterpart of Figure 1, but now displaying in the upper and lower panels the relative vorticity fields and the horizontal divergence fields (contour intervals in units of 10^{-5} s^{-1}). In the upper panel, solid (dashed) contours denote divergence (convergence), with the converse applying for the lower panel. The middle panel depicts the precipitation accumulated during the previous 6 h and the vertical velocity field on the 500 hPa surface (contours for $\pm 0.3, 1$, and 2 Pa s^{-1} , with dashed lines for negative values indicating ascending motion).

independent prediction equations for the horizontal velocity $\mathbf{V}$ and the potential temperature θ , and hence they do not necessarily ensure that geostrophy is maintained in time. These two issues serve to prompt, and indeed are central to, the concept of QG.

Mathematical Formulation

Flow Subdivision

The concept of QG enables the dynamics of synoptic-scale flow to be viewed from a simple and self-consistent perspective that condenses to an elegant mathematical formulation.

In the QG framework, the three-dimensional atmospheric flow field ($\mathbf{U}$) is conceived to comprise two components: a primary, purely horizontal geostrophic component $\mathbf{V}_G$, defined by eqn [6]; and a secondary, three-dimensional ageostrophic component ($\mathbf{V}_{AG}, w$):

$$\mathbf{U} = (\mathbf{V}_G, 0) + (\mathbf{V}_{AG}, w) \quad [11]$$

The ageostrophic component is deemed to be the additional velocity field required to maintain the flow in an almost-geostrophic (i.e., QG) state (discussed further in this article).

The concept itself is formulated here in the idealized setting of flow taking place on and above a plane that is tangential to the Earth's surface at a specified latitude ($\varphi = \varphi_0$) located in the extratropics. On this surface, commonly referred to as the β -plane, the Coriolis parameter (f) is taken to vary linearly in the y -direction (i.e., poleward in the Northern Hemisphere) and is represented by

$$f = f_0 + \beta y \quad [12]$$

where $(f_0, \beta) = (2\Omega \sin \varphi_0, 2\Omega \cos \varphi_0/a)$, with a denoting the Earth's radius.

A systematic introduction of the foregoing flow subdivision into eqns [1–4] in a manner consistent with the primacy of the geostrophic flow component results in the following quasi-geostrophic set of equations:

$$D_G \mathbf{V}_G / Dt = \mathbf{G} \quad [13]$$

$$\text{with } \mathbf{G} = \boxed{-f_0(\mathbf{k} \wedge \mathbf{V}_{AG})} - \beta \gamma(\mathbf{k} \wedge \mathbf{V}_G)$$

$$\boxed{\nabla_H \cdot \mathbf{V}_{AG} + \mathcal{L}w = 0} \quad [14]$$

$$D_G b / Dt + \boxed{N^2 w} = H \quad [15]$$

Here, D_G/Dt refers to the rate of change following the horizontal geostrophic motion, $\mathbf{G}$ is a reduced form for the sum of the Coriolis and pressure gradient forces of eqn [1], and the buoyancy (b) equates to $b = (g\theta^*/\Theta)$. For later use, we note that the boxed terms in eqns [13–15] include contributions from the ageostrophic flow.

In the next two sub-sections, eqns [6, 13–15] are recast into two quasi-independent subsets for, respectively, the geostrophic and the ageostrophic flow.

Prognostic Geostrophic Equations

A set of equations governing the evolution of the geostrophic flow component is derivable from eqns [6, 13–15]. There is

a prognostic equation for the so-called QG potential vorticity, q , that takes the following form:

$$D_G \{q + \beta y\} / Dt = f_0 \mathcal{L}(H/N^2) \quad [16]$$

with q given by

$$q = [\nabla_H^2 \psi + f_0^2 \mathcal{L}\{(1/N^2) \partial \psi / \partial z\}] \quad [17]$$

Likewise, there is a prognostic equation for the surface potential temperature of the following form:

$$D_G(b)/Dt_{z=0} = H_{z=0} \quad [18]$$

And, furthermore, $b (= f_0 \partial \psi / \partial z)$ is merely a function of ψ , the geostrophic stream function.

Thus, these equations prescribe the space–time evolution of q in the interior (eqn [16]) and b at the surface (eqn [18]). They have an initial-boundary value character in that

1. at an initial time, their solution requires the specification of ψ to permit the forward integration of eqns [16] and [18] to yield the subsequent distribution of q in the interior and $\partial \psi / \partial z$ at the surface; and
2. the boundary specification of $\partial \psi / \partial z$ at the surface is required at all subsequent times to permit the inversion of the elliptic eqn [17] to obtain the spatial distribution of ψ .

Note, moreover, that this set of equations is self-contained provided H is known externally or is specifiable in terms of the geostrophic state.

It follows that the evolution of synoptic-scale atmospheric flow is captured to the first order by a remarkably simple and mathematically elegant simplification of the far more complex governing equations.

Diagnostic Ageostrophic Equation

Equations [13–15] can be manipulated to show that the vertical component, w , of the ageostrophic flow satisfies the following elliptic equation:

$$N^2 \nabla_H^2 w + f_0^2 \partial \{\mathcal{L}w\} / \partial z = F + f_0 \beta \partial \mathbf{V}_G / \partial z + \nabla_H^2 H \quad [19]$$

with F , a known function (discussed later in this article) of the geostrophic stream function.

In effect, w is a diagnostic variable whose spatial distribution at any specified time can be determined by inverting eqn [19] given the contemporaneous distribution of the geostrophic stream function ψ and H in the interior and applying the bottom boundary condition $w_{z=0} = 0$. This diagnostic equation is customarily referred to as the ‘omega equation.’

Physical Interpretation

Relationship between the Geostrophic and Ageostrophic Flow

The physical interpretation of QG flow hinges upon the nature of the relationship between the geostrophic and ageostrophic flow components. The subtlety of this relationship can be exemplified in two ways.

First, in their raw form, the QG eqns [13–15] point to a direct influence of the ageostrophic flow (the boxed ageostrophic terms) upon the evolution of the geostrophic velocity and the potential temperature fields. However, this seemingly ostensible link is deceptive. In particular, the form of the

prognostic relationships (eqns [16–18]) demonstrates that the evolution of the QG potential vorticity, q , depends upon only the geostrophic flow itself and is independent of the ageostrophic flow.

Second, deriving separately the rate of change of the two terms in the thermal wind relationship (eqn [9]) yields

$$D_G\{\partial \mathbf{V}_G/\partial z\}/Dt = -f_0 \mathbf{R} - \{\mathbf{k} \wedge (f_0 \partial \mathbf{V}_{AG}/\partial z + \beta \gamma \partial \mathbf{V}_G/\partial z)\} \quad [20]$$

$$D_G\{\mathbf{k} \wedge \nabla_H b\}/Dt = +f_0 \mathbf{R} - \{\mathbf{k} \wedge \nabla_H (N^2 w - H)\} \quad [21]$$

with $\mathbf{R} = [\partial \mathbf{V}_G/\partial z \cdot \nabla_H] \mathbf{V}_G$

so that $\mathbf{R}$ denotes the rate of change of the geostrophic flow along the thermal wind vector.

It follows that the rate of change of the vertical shear of the geostrophic wind (eqn [20]) and the baroclinicity (eqn [21]) are influenced by both the geostrophic flow itself and the ageostrophic flow. However, the influence of the geostrophic $\mathbf{R}$ term upon the shear and baroclinicity terms is equal and opposite. In effect, the geostrophic component acts to destroy the thermal wind balance, and hence the ageostrophic flow must serve to restore that balance. Two stark inferences are that (1) maintenance of the primary flow's geostrophy necessitates the existence of the secondary ageostrophic flow; and (2) the ageostrophic component is determined by, and slaved to, the geostrophic component. In effect, the diagnostic omega equation (eqn [19]) prescribes the ageostrophic component of the flow that must prevail for a specified geostrophic field.

In summary, the foregoing considerations indicate that (1) the geostrophic flow is independent of the ageostrophic flow; and (2) the ageostrophic flow is enslaved to the geostrophic flow. The ramifications of these two deductions are explored in the next two sub-sections.

Evolution of the Geostrophic Flow

The essence of QG flow dynamics is encapsulated by eqns [16–18]. This set of equations forms the basis for the QG version of the so-called Potential Vorticity (PV) perspective. This PV perspective of synoptic and larger-scale flow is physically insightful and cartographically attractive.

Dynamics

Equation [16] states that $\{q + \beta \gamma\}$ is conserved as it is advected with the horizontal geostrophic flow in the absence of diabatic heating ($H = 0$). The physics of this conservation is subtle and simple. The subtlety results from a compensating effect that influences both the rate of change of the vorticity (ζ_G) and a measure (Π) of the stratification, with Π defined such that $q = (\zeta_G + \Pi)$.

It can be shown from eqns [13–15] that

$$D_G \zeta_G / Dt = -\beta v_G - f_0 (\nabla_H \cdot \mathbf{V}_{AG}) \quad [22]$$

$$D_G \Pi / Dt = f_0 \mathcal{L}\{H/N^2\} + f_0 (\nabla_H \cdot \mathbf{V}_{AG}) \quad [23]$$

where $\Pi = f_0 \mathcal{L}\{(1/N^2)g\theta^*/\Theta\}$

Thus, the evolution of both ζ_G and Π is influenced by the ageostrophic horizontal divergence (the term involving $\nabla_H \cdot \mathbf{V}_{AG}$) and the accompanying vertical contraction. For example, horizontal convergence (and thus vertical stretching)

spins up the fluid and thereby increases ζ_G (see eqn [22]) while concomitantly reducing the vertical gradient of θ^* and thereby the stratification measure Π (see eqn [23]). However, the net contribution of this ageostrophic effect upon the development of q is zero.

The ramification for the physics is that the entire flow evolution is condensed to two conservation relationships: pseudo air parcels moving adiabatically with the geostrophic flow in the interior conserve their value of $\{q + \beta \gamma\}$ (eqn [16]), and pseudo air parcels moving adiabatically with the geostrophic flow along the bottom boundary conserve their value of θ^* (eqn [18]).

Thus, the flow evolution, although intrinsically nonlinear, results directly from the prevailing geostrophic flow field modifying, via geostrophic advection, the spatial distribution of q in the interior and θ^* at the surface. Indeed, the instantaneous development can be qualitatively inferred from inspection of the relative configuration of the prevailing QG potential vorticity, q , and the geostrophic flow field itself.

In addition, it can be shown that a localized positive (negative) anomaly of q in the interior is associated with enhanced (reduced) values of vorticity and reduced (enhanced) pressure, and that these signatures decay with distance from the anomaly. Also, the stratification is increased in the vicinity of the anomaly and reduced both above and below. Similar signatures are associated with warm (cold) anomalies of θ^* on the bottom boundary. These features are illustrated schematically in Figure 3.

Flow development and cartography

The ramifications for charting the flow development emerge from recognizing that

- the geostrophic flow field can be viewed as the net of two components, $\psi = \psi_I + \psi_B$, associated respectively with q in the interior and θ^* at the bottom boundary; and
- the interior q and the boundary θ^* fields can each be decomposed into, say, n arbitrary subelements (e.g., q_i and θ_i^*), such that

$$q = \sum_{i=1,n} (q_i), \text{ and } \theta^* = \sum_{i=1,n} (\theta_i^*) \quad [24]$$

It follows, from the linearity of eqn [17], that (ψ_I, ψ_B) satisfy the following relationships:

$$\nabla_H^2 \psi_I + f_0^2 \mathcal{L}\{(1/N^2)\partial \psi_I/\partial z\} = q$$

with $f_0 \partial \psi_I/\partial z = 0$ on $z = 0$

and

$$\nabla_H^2 \psi_B + f_0^2 \mathcal{L}\{(1/N^2)\partial \psi_B/\partial z\} = 0$$

with $f_0 \partial \psi_B/\partial z = (g\theta^*/\Theta)$ on $z = 0$

Analogous relationships also apply for each q and θ^* subelement.

For the cartography of the flow, the foregoing multiple decomposition is particularly apt provided the synoptic-scale flow does resemble numerous distinctive interior q and surface θ^* substructures. In this circumstance, an appropriate and insightful depiction of the flow would be to map the various subelements, and then the flow evolution could be viewed as the interaction between these subelements plus their intrinsic self-development.

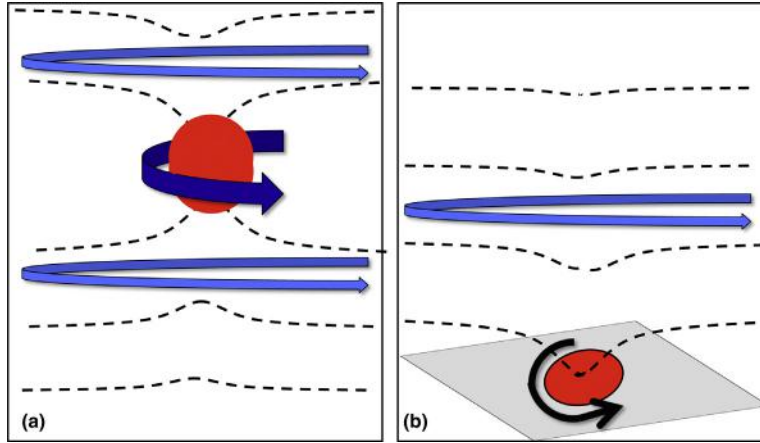


Figure 3 A schematic of the flow and thermal pattern associated with a localized positive anomaly of quasigeostrophic potential vorticity (left panel) and a local positive surface thermal anomaly (right panel). The arrows indicate the circumferential flow around the anomalies, with the strength and scale indicated, respectively, by the thickness and lateral extent of the arrows. The dashed lines represent the isentropes.

An illustration

One paradigm for cyclogenesis is the approach aloft of a major trough toward and over a strong low-level baroclinic zone. A schematic illustration of this paradigm when viewed from the QG PV perspective is shown in Figure 4.

At the initial instance (Figure 4(a)), an isolated positive upper-level q -anomaly aloft approaches a low-level baroclinic zone that is aligned east–west. The velocity signature of the upper-level anomaly at the surface serves to distort the baroclinic zone into a wave shape, creating positive and negative anomalies in the surface thermal pattern (Figure 4(b)). Thereafter (Figure 4(c)), the upper-level anomaly's continued distortion of the wave increases the amplitude of the surface thermal anomalies, and in addition the far-field signature of these surface anomalies at upper levels can advect the q -anomaly overhead of the evolving surface front. The latter movement enhances both the strength of the interlevel interaction and the amplitude of the incipient frontal cyclone.

Diagnosis of the Ageostrophic Flow

The interpretive challenge

For simplicity, we consider here, and in the remainder of this section, that the flow is on an f -plane ($\beta = 0$), and is adiabatic and incompressible. In this limit, the omega equation (eqn [19]) reduces to the simpler form:

$$N^2 \nabla_H^2 w + f_o^2 \partial^2 w / \partial z^2 = F \quad [25a]$$

or, more succinctly,

$$\mathcal{M}^2 w = F \quad [25b]$$

where $\mathcal{M}^2$ is the corresponding three-dimensional elliptic operator, and F is, as noted earlier in this article, a known function of the geostrophic stream function.

The numerical solution of eqn [19] for a specified F distribution would yield the full three-dimensional distribution of the vertical velocity field (w) forced by the prevailing geostrophic flow. Here, the focus is on using the omega equation directly to qualitatively infer from inspection of synoptic charts the regions that are favorable for ascent and descent.

The procedure adopted to address this interpretative challenge is to first recognize that

$$(\mathcal{M}^2 w) \alpha - (w)$$

so that a localized region of large negative ' F forcing' favors ascent. Then, to diagnose regions of major ascent and descent, the locations of regions of large (negative or positive) F remain to be inferred from inspection of synoptic charts.

The F forcing

The expression for F can be, and has been, cast in markedly different yet mathematically equivalent forms. Consider the following four formulations.

The conventional formulation:

$$F = f_o \partial [(\mathbf{V}_G \cdot \nabla_H) \zeta_G] / \partial z - \nabla_H^2 [(\mathbf{V}_G \cdot \nabla_H) b] \quad [26a]$$

Refined-Sutcliffe formulation:

$$F = 2f_o \{ [(\partial \mathbf{V}_G / \partial z) \cdot \nabla_H] \zeta_G - D^2 (\partial \lambda / \partial z) \} \quad [26b]$$

'Q vector' formulation:

$$F = 2(\nabla_H \cdot \mathbf{Q}) \quad [26c]$$

'PV perspective' formulation:

$$F = f_o \partial [(\mathbf{V}_G \cdot \nabla_H) q] / \partial z - \mathcal{M}^2 \{ (\mathbf{V}_G \cdot \nabla_H) b \} / N^2 \quad [26d]$$

Here, D^2 is the total deformation and λ the angle of the deformation field's dilatation axis, and $\mathbf{Q}$ is a horizontal vector that for interpretative purposes can be written in the following form:

$$\mathbf{Q} = -|\nabla_H \theta^*| \{ \mathbf{k} \wedge \partial \mathbf{V}_G / \partial s \}$$

where s is a local coordinate aligned along an isentrope such that cold air is located to its left.

Versions [26a] and [26b] arise from direct manipulation of the vorticity and thermal eqns [22] and [23], the $\mathbf{Q}$ vector version from manipulation of eqns [20] and [21], and the PV version from operating directly on eqns [15] and [16]. The relative merit of these formulations depends in part upon their mathematical uniqueness, their potential ability to provide physical insight, and the ease of their usage to diagnose vertical motion.

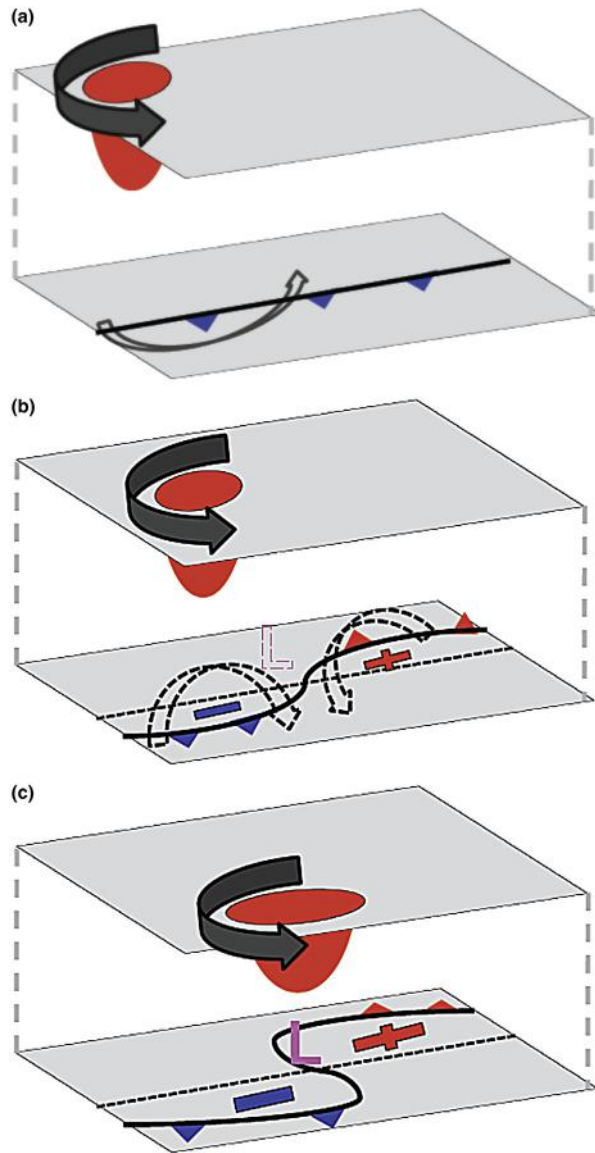


Figure 4 A schematic illustrating the nature, location, and movement of salient PV features associated with one paradigm for cyclogenesis. A positive PV anomaly at the upper level (red feature) is depicted approaching and interacting with a surface front initially aligned east–west. The flow associated with the PV anomaly is shown as a filled arrow aloft and an unfilled arrow at the surface, and the dashed arrows indicate the flow associated with the evolving surface warm (+) and cold (–) thermal anomalies.

Versions [26a] and [26d] suffer from similar defects: (1) they are not Galilean invariant because, for each version, the relative contributions of their two forcing terms differ in the presence of an additional uniform translational velocity; and (2) for each version, the two terms of the forcing contain equal and opposite contributions. Both these defects militate against attaching a separate physical significance to the two terms. Notwithstanding, version [26d] does provide an elegant interpretation and linkage of the forcing to the PV perspective, and it has been used to shed light on the distinction between mere pattern translation versus flow development.

The remaining two schemes, the refined-Sutcliffe formulation and the $\mathbf{Q}$ vector formulation (eqn [26b] and [26c]), are Galilean invariant and are not subject to the aforementioned cancellation effect. Also, they can be interpreted physically in terms, respectively, of offsetting vortex stretching by thermal adjustment and of maintaining the thermal wind balance. Also, in terms of practical utility, these schemes have been deployed to assess the forcing at a given level using a chart displaying the geopotential and thermal patterns at that level. The next subsection illustrates the application of these two schemes to infer regions of major ascent and descent.

Illustrations

Consider again the paradigm introduced earlier (see Section An Illustration) of cyclogenesis following the approach aloft of a major trough toward and over a surface baroclinic zone. Figure 5 illustrates schematically the nature of the F forcing for, respectively, the refined-Sutcliffe and the $\mathbf{Q}$ vector formulations on a level in the lower troposphere (e.g., 700 hPa) where the vertical velocity field is expected to be significant.

For the refined Sutcliffe formulation, it has been customary to neglect the second of the two forcing terms, and then to qualitatively assess the forcing due to the remaining term:

$$[(\partial \mathbf{V}_G / \partial z) \cdot \nabla_H] \zeta_G$$

This term (the Sutcliffe development term) is proportional to the product of the strength of the thermal wind (the in situ baroclinicity) and the variation of the vorticity along the thermal wind. Hence, for the specified flow setting (Figure 5(a)), the vorticity along the baroclinic zone increases ahead of the trough and decreases behind it, and this in turn demarks the major descent and ascent regions.

For the $\mathbf{Q}$ vector formulation, the method is to exploit the divergence form, $(\nabla_H \cdot \mathbf{Q})$, of the forcing by noting that (1)

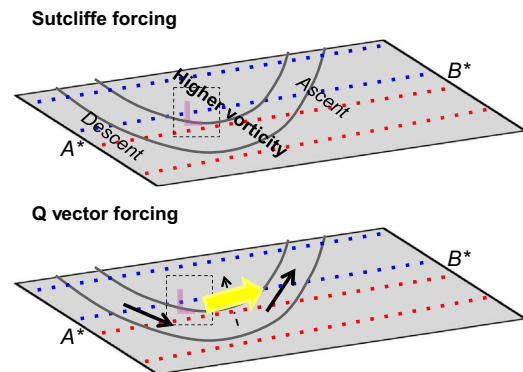


Figure 5 A schematic of the forcing term for the omega equation as represented by the Sutcliffe (upper panel) and the $\mathbf{Q}$ vector (lower panel) formulations of the F forcing. The dotted colored lines represent the isentropes of a strong baroclinic zone on the 700 hPa level, the gray continuous lines represent the geopotential height field of a trough, and L is the location of a surface Low. For the Sutcliffe formulation, consideration of the vorticity change along the thermal wind line A^*B^* shows first an increase and then a decrease. For the $\mathbf{Q}$ vector formulation, consideration of the vector wind change along the same line shows a change at the base of the trough indicated by the dotted arrow, and hence a $\mathbf{Q}$ vector (yellow arrow) pointing along the thermal wind.

a region of maximum Q convergence (i.e., negative $\nabla_H \cdot \mathbf{Q}$) favors ascent; and (2) in the vicinity of such a region, the Q vector will tend to point toward the region of convergence and away from a divergent region. The formula for the Q vector

$$\mathbf{Q} = -|\nabla_H \theta^*| \{ \mathbf{k} \wedge \partial \mathbf{V}_G / \partial s \}$$

indicates that it will be large in a region characterized by both strong baroclinicity and a marked change in the strength and/or direction of $\mathbf{V}_G$ along the accompanying isentrope. In effect, the Q vector will be aligned 90° to the right (in the Northern Hemisphere) of the vector wind change.

For the specified flow setting, the largest change of the vector wind along the baroclinic zone is around the base of the trough and is indicated by the dotted arrow in [Figure 5\(b\)](#). Hence, the resulting Q vector (yellow arrow) points down the thermal wind, signifying ascent and descent (respectively) in the regions ahead of and behind the vector. In effect, both formulations agree on the pattern of the vertical motion.

Another diagnostic follows from noting that a causal chain links ascent to cyclogenesis. Ascent in the lower troposphere connotes flow convergence ([eqn \[2\]](#)), and hence vorticity enhancement of an air-parcel advecting downstream with the geostrophic wind ([eqn \[22\]](#)). Further, that enhancement is accompanied by a concomitant decrease in the ambient pressure ([eqn \[7\]](#)), and thereby a tendency for cyclogenesis downstream of the prevailing region of ascent. Likewise, descent is linked to anticyclogenesis. On this basis, the pattern inferred for the vertical motion in [Figure 5](#) would imply that the pre-existing incipient low-pressure system would move along the thermal wind (the so-called steering effect).

Cursory Case Study

Here, the appropriateness and the utility of the QG PV perspective and the interpretative diagnostic schemes discussed in Section Physical Interpretation are examined for one particular realized event of cyclogenesis. The event occurred over the Atlantic from 31 December 2006 to 1 January 2007. The synoptic setting at 12 UTC on 31 December 2006 was discussed in Section Mathematical Formulation and is shown in [Figures 1](#) and [2](#).

The PV Perspective

[Figure 6](#) shows the flow field and a proxy for the net QG PV (i.e., $q + \beta y$) on the near-tropopause 315 K isentropic surface (upper panels), and the geopotential and potential temperature fields on the 1000 hPa surface (lower panels). The displays are for 12 UTC 31 December and 00 UTC 1 January.

At the upper level, the initial potential vorticity distribution has a narrow and elongated band of sharp q gradient aligned along the tropopause-level jet with an accompanying short-wavelength equatorward protrusion of the jet at 60° W. The latter constitutes a positive q anomaly, and in the subsequent 12 h this feature is seen to dilate and extend further south-eastward to form a strong and distinctive filament.

At the surface, a band of strong baroclinicity extends eastward from the Carolinas at 12 UTC. An incipient Low is located

on this band beneath the southeastern extremity of the PV protrusion aloft. In the subsequent 12 h, the surface pattern exhibits a classical development with the thermal field undergoing a strong distortion to form a cold and warm frontal palette separated by a warm sector. The attendant Low develops rapidly, and its eastward movement shadows that of the q anomaly aloft.

This development resembles the cyclogenesis paradigm introduced in Section Cursory Case Study. [Figure 6](#) lends credence to the assertion that the atmosphere has distinctive PV subelements, pinpoints clearly the subelements involved in the development, and illustrates and confirms that insightful qualitative inferences can be drawn regarding the interaction between the subelements and the subsequent short-term movement of the anomalies.

Q Vector Diagnosis

[Figure 7](#) is the counterpart of [Figure 6](#), but now the displayed fields are for the 700 hPa surface and show the potential temperature and wind field (upper panels) and the Q vector pattern and vertical motion (lower panels).

The patterns in the upper panels allow one to quickly and qualitatively infer the location of large Q vector contributions. At the earlier time, there is a significant flow change up- and downstream of the cusp in the thermal pattern near (39° N, 55° W), and this is consistent with strong eastward-pointing Q vectors in this region. At the later time, both the new location of the cusp and the warm frontal zone coincide with strong flow changes with dominant Q vectors aligned, respectively, north-eastward and southeastward.

In the lower panels, the quantitatively computed Q vectors confirm the foregoing qualitative inferences, but their pattern also indicates that the F forcing has a more spatially refined substructure that is itself confirmed by the displayed pattern of the vertical velocity field.

Furthermore, note that at the earlier time ([Figure 7\(c\)](#)), the Q vector pattern is indicative of ascent in the neighborhood of the surface Low and along the contiguous cold and warm fronts. This signature, taken in conjunction with the south-westerly geostrophic flow above the Low at 700 hPa ([Figure 7\(a\)](#)), correctly foreshadows the subsequent movement of the Low to the northeast (cf. [Figure 6\(c\)](#) and [6\(d\)](#)). At the later time, the Q vector pattern ([Figure 7\(b\)](#)) has a similar overall structure but is stronger and more coherent. Again, consistent with this signature and the 700 hPa flow above the surface Low, the latter moves northeastward and continues to develop rapidly (not shown).

The foregoing results underline the practical utility of the omega equation for indicating qualitatively and readily the regions of major ascent, and concomitantly pointing to the regions favorable for cyclogenesis.

Limitations and Refinements

The QG set of equations provides a first-order and remarkably insightful representation of synoptic and larger-scale atmospheric flow. The emphasis here has been to understand the fundamental dynamics of QG flow and then to

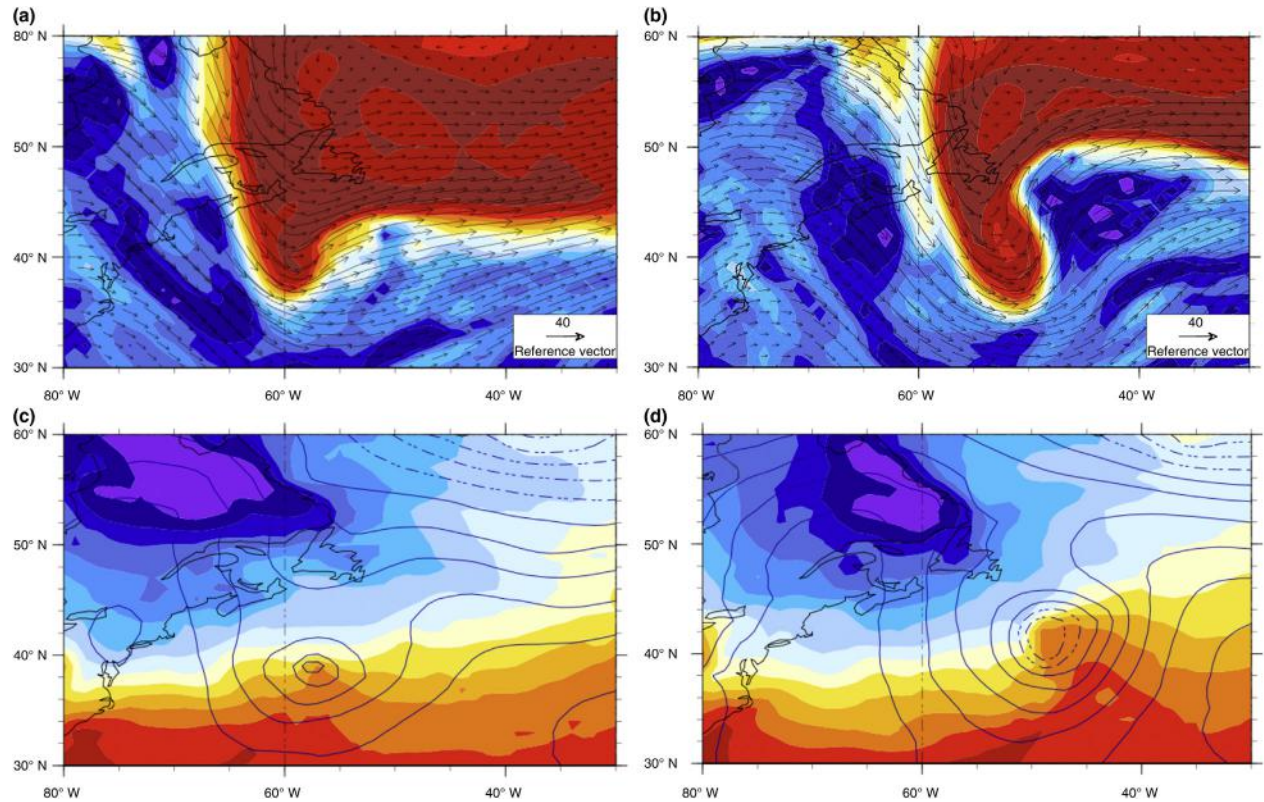


Figure 6 Charts for 12 UTC 31 December 2006 (left panels) and 00 UTC 1 January 2007 (right panels). Panels (a) and (b) display the flow field and the natural logarithm of PV, i.e., $\ln PV$ (a proxy for the quasigeostrophic PV) on the 315 K isentropic surface. Panels (c) and (d) display the height field (blue contours every 50 m) and potential temperature distribution on the 1000 hPa surface for the corresponding times.

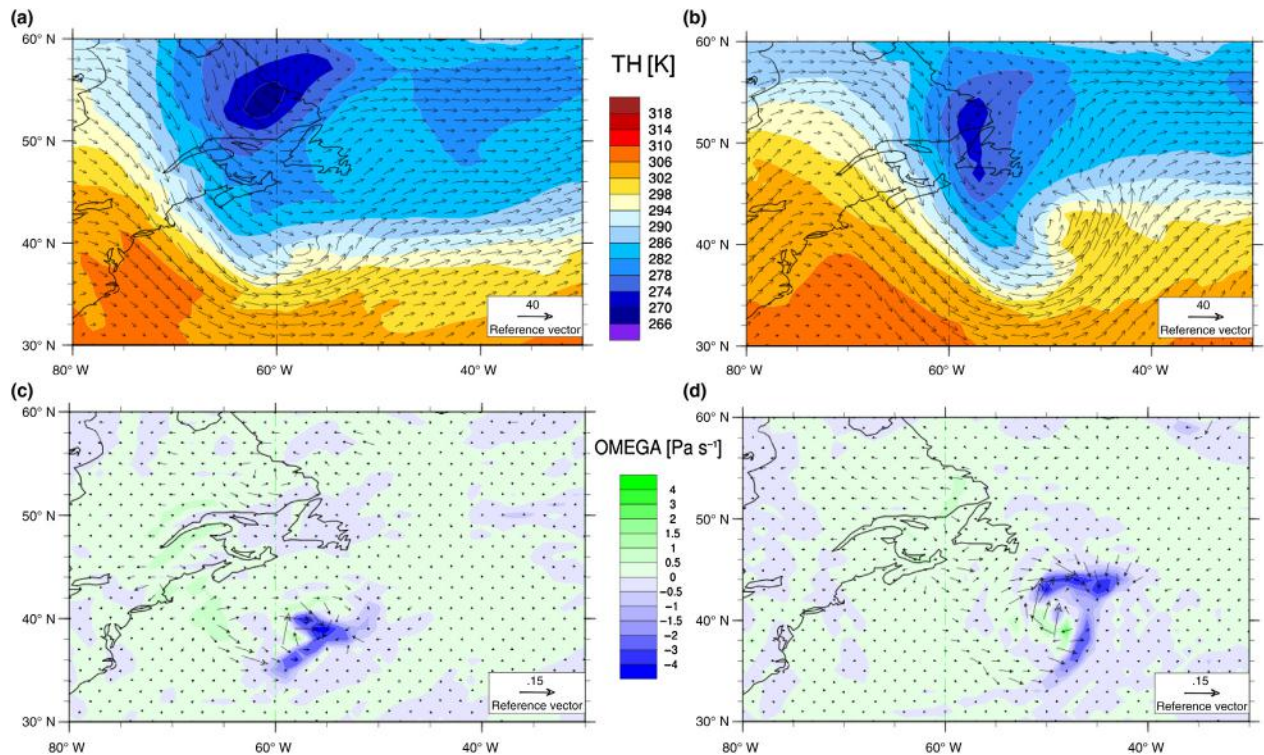


Figure 7 Panels A and B show the flow field and potential temperature pattern on the 700 hPa surface at, respectively, 12 UTC 31 December 2006 and 00 UTC 1 January 2007. Panels C and D display both the vertical motion field (negative values indicate rising motion) and the $\mathbf{Q}$ vectors on the 700 hPa surface for the corresponding times. Note that here the $\mathbf{Q}$ vectors have been computed with the full horizontal wind rather than the geostrophic wind.

illustrate its utility for understanding day-to-day weather development. However, the range of application of the QG set is much broader. For example, the set provides the platform for the theoretical study of large-scale planetary (Rossby) waves and their propagation, barotropic and baroclinic instability, the perturbing effect of orography upon the large-scale flow, and the large-scale dynamics of the stratosphere.

Nevertheless, QG is only a first-order representation of synoptic-scale flow, and it is not strictly valid in flow regimes with strong ageostrophic flow such as usually prevails for large-scale tropical convective systems, in the finer-scale features of extratropical weather systems, and in the vicinity of major mountain ranges. Also intrinsic to the QG representation is the exclusion of buoyancy and acoustic-wave phenomena.

Extensions of QG theory have been developed that are predicated upon the existence of some higher form of dynamical balance beyond that of geostrophy (e.g., the semi-geostrophic and geostrophic momentum set of equations).

See also: **Dynamical Meteorology:** Balanced Flow; Baroclinic Instability; Coriolis Force; Potential Vorticity; Rossby Waves;

Vorticity. Middle Atmosphere: Planetary Waves. **Synoptic Meteorology:** Cyclogenesis; Extratropical Cyclones; Frontogenesis; Fronts; Weather Maps.

Further Reading

For an overview of synoptic-scale atmospheric flow:

Newton, C., Holopainen, E.O., 1990. *Extratropical Cyclones: The Erik Palmén Memorial Volume*. American Meteorological Society, Boston.

Shapiro, M., Gronas, S., 1999. *The Life Cycle of Extratropical Cyclones*. American Meteorological Society, Boston.

For a discussion of geostrophy and potential vorticity in relation to synoptic-scale flow: Eliassen, A., 1984. Geostrophy. *Quarterly Journal of the Royal Meteorological Society* 110, 1–12.

Hoskins, B.J., 1997. A potential vorticity view of synoptic development. *Meteorological Applications* 4, 325–334.

For a consideration of various formulations of the quasigeostrophic omega equation: Durran, D.R., Snellman, L.W., 1987. The diagnosis of synoptic-scale vertical motion in an operational environment. *Weather and Forecasting* 2, 17–31.

Pedder, M.A., 1997. The omega equation: Q–G interpretations of simple circulation features. *Meteorological Applications* 4, 335–344.

Rossby Waves

PB Rhines, University of Washington, Seattle, WA, USA

© 2003 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 5, pp 1923–1939, © 2003, Elsevier Ltd.

Introduction

Large-scale undulations in the westerly winds are related to an ideal form of motion known as the ‘Rossby’ or ‘planetary’ wave. These waves owe their existence to the rotation and spherical shape of the Earth. Weather patterns and the general circulation are mostly much wider than the depth of the atmosphere: viewed from the side, a weather system is 100 to 1000 times thinner (vertically) than its width. This extreme thinness, beyond reminding us of the fragile nature of the atmosphere, causes horizontal winds to be stronger than vertical winds in such weather systems. Stable layering of the air, with its great variations in density, reinforce this inequality. Waves then become possible, which are dominated by nearly horizontal wind patterns, in many ways unlike the familiar waves on the sea or those of sound or light.

In their most general form, Rossby waves have an important bearing on what we call ‘weather’ and on the form of the general circulation of the Earth’s atmosphere and its oceans, and on the atmospheres of other planets. Indeed, the form of the general circulation is in part shaped by Rossby waves. In their most general form, Rossby waves occur widely in fluid flows of many kinds (e.g., in hurricanes). To understand these waves completely requires a challenging amount of mathematics and physics, but many of their properties can nevertheless be appreciated through ideas, observations, and experiments that could be found in a high-school physics course. Because the language of science can be artificially complex, we provide translations of some unfamiliar terms inside brackets {}. Some mathematical equations are included, but these can be skipped by those unfamiliar with them.

Horizontal Propagation

If a westerly wind {one that blows eastward} passes over a mountain range, undulations develop downstream, meandering north and south with regularity. This is somewhat like a brook flowing rapidly over a stony bottom, where small ripples and longer gravity waves are seen on the surface. The uneven bed of the fluid disturbs the flow, sending out waves. Waves trying to ‘stem the current’ and propagate upstream are particularly strong: these waves are held fixed in space. They ‘resonate’ with the rocks and build up in amplitude (yet, more subtly, their energy does propagate away from the stony source, upstream for ripples and downstream for gravity waves). Such a pattern, calculated for the much larger-scale flow appropriate to Rossby waves, is seen in Figure 1. The waves here are excited by flow of a single layer of fluid over an isolated mountain; the ‘mountain’ is idealized in the form of a circular cylinder of finite height. For a westerly wind speed of 30 m s^{-1} , at latitude 40° on the Earth, the wavelength {the distance from one northward undulation to the next} is about 8200 km. Rossby

waves are primarily horizontal motion, rather than vertical. Unlike waves at the surface of the sea, but somewhat like longer gravity waves on a stream, they involve the entire fluid, and help to shape its circulation. If the mean winds were reversed in Figure 1, they would flow over the mountain with very little disturbance: Rossby waves cannot then be generated.

In large-scale atmospheric flow, isobars {lines along which pressure is constant} nearly coincide with streamlines {lines along which the wind is directed}; this is why useful weather maps show the atmospheric pressure. Notice at once that the meandering flow in Figure 1 has a cyclonic low-pressure center just downstream of the mountain, with relatively high pressure upstream. This indicates that the wind is pressing eastward on the mountainous solid Earth, and in response the Earth is pressing westward on the atmosphere {with a horizontal force found by adding up everywhere the pressure multiplied by the slope of the topography}. It seems that in generating a Rossby wave, the mountain is also exerting, over time, a westward force on the atmosphere. In fact, this ‘wave-drag’ is a key part of the angular momentum balance of the atmospheric circulation. With the passing of the seasons, fluctuations in this balance cause the speed of rotation of the Earth to fluctuate, and the length of the day to change by about 1.5 ms.

The observed northern hemisphere flow in the upper troposphere is shown as an average over the winter season (Figure 2a). The deviation from circular streamlines, which would correspond to strictly east–west winds, is the ‘stationary’ or ‘standing’ wave field of the atmospheric general circulation (see *Dynamical Meteorology: Stationary Waves (Orographic and Thermally Forced)*). The flow is westerly throughout most of the region, which does not extend south into the trade winds. Where the streamlines squeeze close together the winds

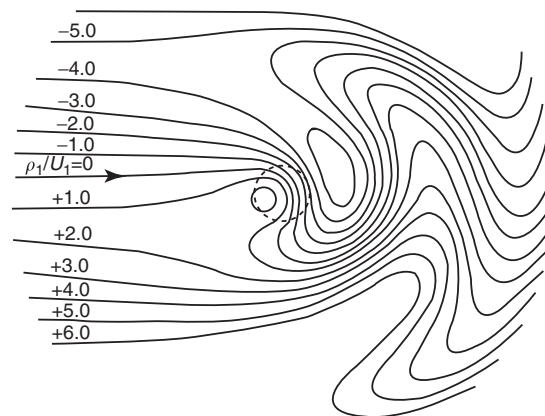


Figure 1 Rossby waves in westerly flow over an idealized mountain (a circular cylinder). From McCartney M, *Journal of Fluid Mechanics*, 68: 71–95, 1975. Shown are the streamlines, stationary pattern along which the wind is directed.

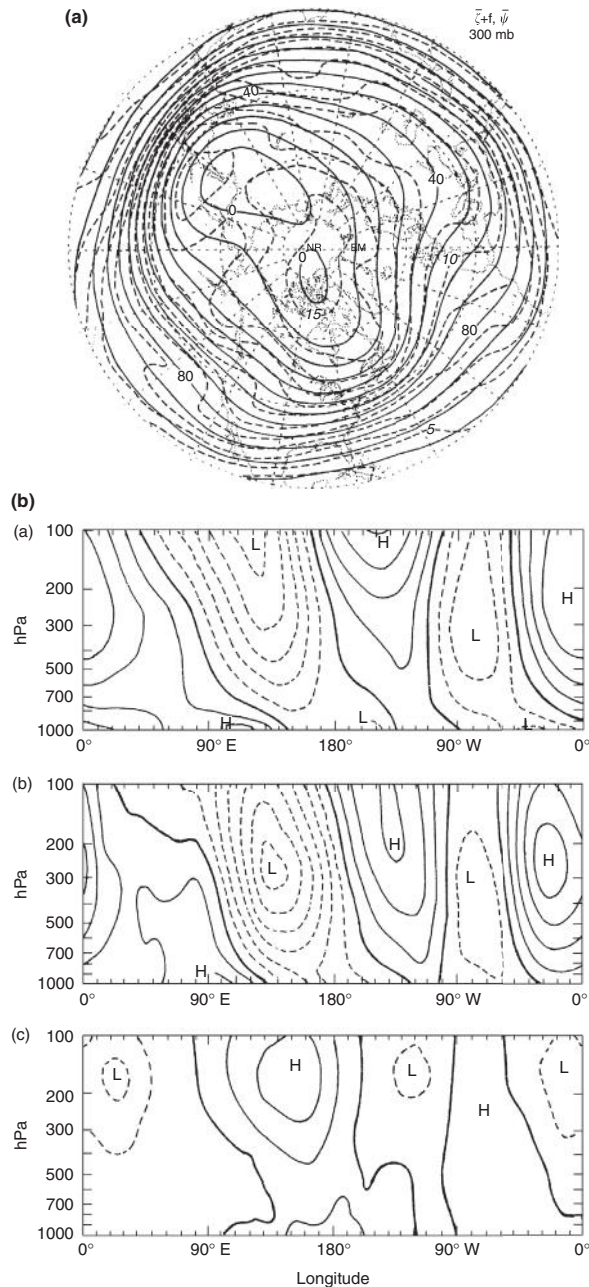


Figure 2 (a) Streamlines (solid) and absolute vorticity, $\xi + f$, averaged at the 300 hPa level in the upper troposphere. Figure from Lau NC, *Journal of the Atmospheric Sciences*, 36: 996–1016, 1979. The deviation of the streamlines from circles is a wave pattern analogous to Figure 1. If this were simply barotropic Rossby waves, the dashed and full lines would coincide. Continents are shown with dotted curves; North America is at the bottom of the figure. (b) Cross sections of the geopotential height of the atmosphere at 601° N (a), 451° N (b), and 251° N (c). This is the flow averaged in time during the winter season. Here the vertical axis is pressure-level and the horizontal axis is longitude. To a good approximation this can be viewed as a plot of pressure with respect to actual altitude and longitude. Figure from Lau NC, *Journal of the Atmospheric Sciences*, 36: 996–1016, 1979. The flow in (a) cuts through this figure at the 3000 hPa level.

are stronger: this occurs downwind of the Himalayan plateau at the western side of the Pacific Ocean, and downwind of the Rocky Mountains, at the western side of the Atlantic Ocean. Of course at any given moment the winds will not look like this figure, for transient waves and eddies, some of them associated with the jet stream, are as strong as the mean winds. Between two and three waves fill a latitude circle, in this pattern.

Vertical Propagation

Complementing this purely horizontal view of the circulation, a vertical cross section of the atmosphere, running east and west, shows that in a broad-brush sense the wind patterns vary very little with altitude, though they grow in strength as one moves up to the top of the troposphere (where the jet streams are strongest). This means that a ‘barotropic’ model, {such as one with a single layer of uniform-density fluid, as in Figure 1}, should have some validity. Notice, as in the horizontal map of the same wind field (Figure 2(a)), and the theoretical calculation behind Figure 1, that major low-pressure regions lie in the lee of the Himalayan Plateau in Asia, near 90° E and in the lee of the Rocky Mountains of North America at 90° W longitude. Because a geostrophic {large-scale} wind moves with high pressure to its right (in the Northern Hemisphere, conversely in the Southern Hemisphere), there are northerly winds in the lee of the mountain ranges, as in Figure 1.

A more subtle feature, however, shows that the winds and temperature are organized in a pattern that tilts to the west as one moves upward through the atmosphere (Figure 2(b)). This tilt is a signature of several important things. It signifies upward propagation of energy and of easterly momentum. The easterly {westward} pressure force exerted by the solid Earth is transmitted upward through the atmosphere by ‘wavy’ layers of air that exert pressure forces similar to those on the solid mountain slopes. These forces alter the winds aloft. Upward propagation of a given amount of energy into regions of thinner (less dense) atmosphere leads to increasing velocity with height (because the kinetic energy is the product of air density and squared velocity). For this reason, Rossby waves that reach the stratosphere can be accompanied by strong winds, and can drive strong changes in the mean westerly winds.

The tilted pattern also indicates that north–south wind velocities, v , and temperature, T , have a systematic correlation. Their product, vT , is the poleward {toward the North or South Pole depending on which hemisphere one is in} transport of heat. Recall (see **Radiation Transfer in the Atmosphere: Radiation, Solar**) that most of the significant motions of the atmosphere and ocean are driven by solar radiation, heating the tropics, with the heat radiating back to space most strongly near the poles. Rossby waves, or motions with some of their properties, contribute to the north–south ‘heat pipe’ in this giant heat engine. This discussion fills out the fully three-dimensional picture of the two-dimensional problem shown in Figure 1.

Because Rossby waves exist throughout the depth of the atmosphere, they are not ‘superficial oscillations,’ but rather an expression of forces acting on the entire circulation. An isolated cyclone, if it is large enough in size and weak enough in wind velocity, will ‘burst apart’ into Rossby waves, forming new,

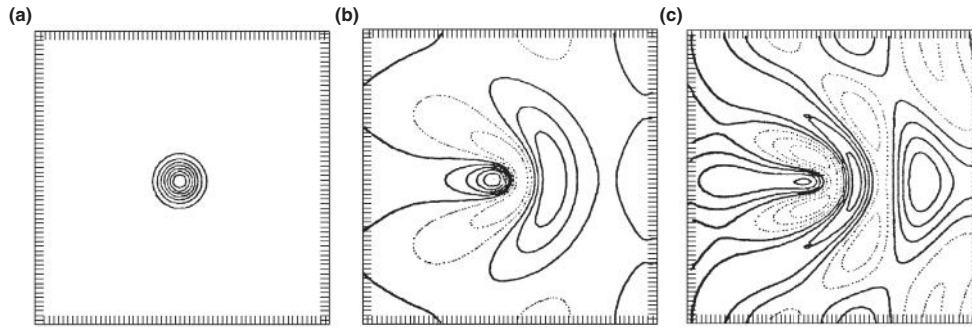


Figure 3 The breakup of an initial weak vortex (a). Shown are streamlines or constant-pressure curves. The ‘banana’ shaped patterns are high-pressure and low-pressure cells associated with Rossby waves. This is a single layer of fluid without a mean westerly wind, using a β -plane approximation to the rotating Earth. This numerical solution is periodic east and west, with waves entering the domain from the east after exiting in the west. From Rhines PB, *The dynamics of unsteady currents*. In ED Galdberg (ed.), *The Sea*, Vol. VI, pp. 189–318, New York: Wiley, 1977.

elongated cyclones and anticyclones that gradually fill the fluid with motion (Figure 3). In this experiment, the fluid itself oscillates gently, moving only slightly compared with the movement of the wind patterns. In the figure we have no mean westerly wind, so that most of the fluid is motionless. Stronger, more realistic cyclones do not radiate waves so efficiently, but the forces involved in the waves are still at work, for example, nudging hurricanes out of the tropics toward the poles.

Examples: Blocking Patterns in the South-East Pacific and Teleconnections over North America

Observations of winds at the 300 hPa level in the Pacific (Figure 4) show a structure that illustrates the nature of Rossby wave propagation. Cumulus convection in the western tropical Pacific provides a large-scale pattern of divergent winds aloft. A train of cyclones and anticyclones appears south of Australia and veers south-eastward toward Chile, where it creates a lasting circulation cell strong enough to be called a ‘blocking’ pattern. The figure shows correlations of the winds with a time-series that expresses this blocking: the winds themselves form a similar pattern. The waves arrive quickly (the speed of individual cyclones and anticyclones being about 5° of longitude per day, which appears to be slower than the south-eastward development of the pattern as a whole). They decay slowly, influencing a vast region of the South Pacific and reaching into the South Atlantic. In a simpler fluid than the atmosphere, convection in the tropics would stimulate a more local response: here, waves cause a ‘teleconnection’ round half of the circumference of the Earth.

When the close-in views of theoretically solved Rossby waves (Figures 1 and 3) are expanded to the whole sphere, the waves tend to propagate along great-circle paths. While theory predicts their structure in the simplest circumstances, computer models, which solve the mathematical equations approximately, must be employed if realistic mean wind patterns and land topography are included. The modeled Rossby-wave field generated by a similar pattern of equatorial heating by the ocean, Figure 5, has two branches propagating south-eastward and north-eastward from the western Pacific. The wave train crossing North America is similar to the ‘PNA pattern’ that is

associated with ENSO events (more will be said about this pattern below).

Some Specific, and Rather Technical, Results

What features of the atmosphere are explained in some way by Rossby waves? Begun a century ago as an exploration of weak oscillations of the atmosphere and oceans (e.g., the tides raised by Moon and Sun), the theory of Rossby waves now provides insight into the very heart of atmospheric (and oceanic) circulation dynamics. These waves are related to the meandering north and south of the westerly winds and, less directly, to the synoptic eddies that shape our weather. Rossby waves contribute to understanding the global pattern of these westerly winds, the enhancement of cyclonic disturbances in the lee of major mountain chains, the location and shape of storm tracks in the western Atlantic and western Pacific, some forms of blocking and stagnation of air masses, the propagation of energy in long waves upward to the stratosphere, the transport of east–west momentum with these waves, and the attendant deceleration and ‘sudden warming’ of the wintertime vortex that sits above the North Pole.

Along the Equator, oceanic heat and water vapor drive cumulus towers that heat the larger-scale atmosphere. The winds converge below and diverge above such a heat source, and air pulled into the pattern creates a pattern of circulation extending both east and west from the heat source. Rossby waves propagating westward from the region of forcing control the shape of this pattern to the west, while Kelvin waves describe the movement east of the heating. At a yet larger scale, the atmosphere signals the onset of El Niño in the tropical Pacific by sending a train of waves across North America: in the simplest idealization these are Rossby waves (meanwhile, in the sea below, Rossby waves move westward along the Equator to reinforce the recurrence of El Niño). In the lower troposphere in summer great anticyclones fill the North Atlantic and Pacific Oceans, and these are established by monsoon forcing (warming of the land surface) yet organized and shaped by westward propagation of low-frequency Rossby waves. More distant relatives of the Rossby wave account for the basic instability of the primary, east–west

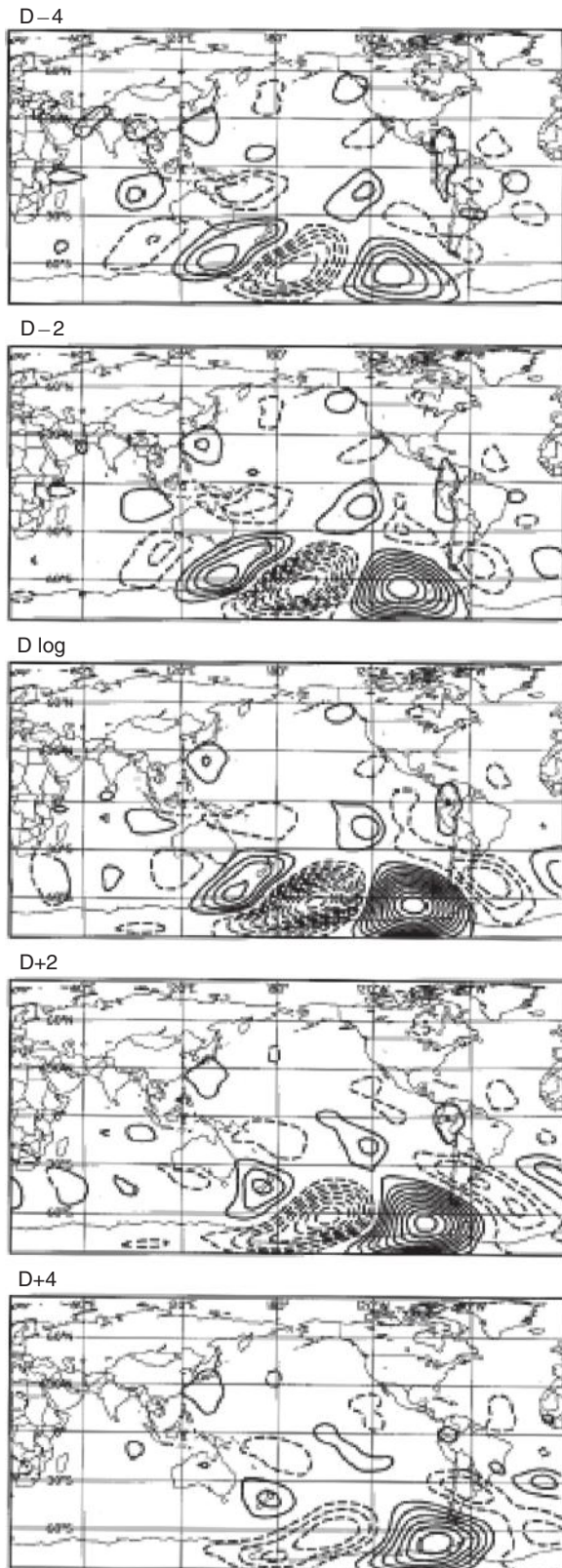


Figure 4 A map showing the correlation between the north-south winds at 300 hPa level with the time series of south-east Pacific blocking. The panel labeled D-4 represents the correlation pattern 4 days before maximum blocking, and the sequence proceeds in time to

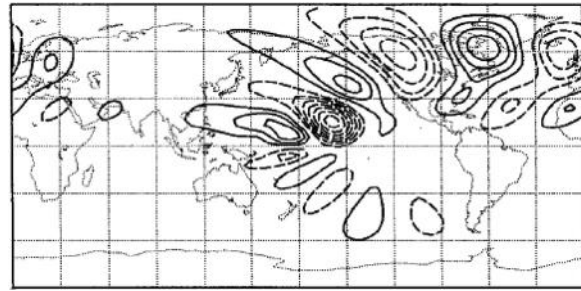


Figure 5 Computer models of the atmospheric circulation play a key role, not just in weather forecasting, but in developing ideas and testing theories of the underlying dynamics. Here, trains of Rossby waves propagate in both hemispheres from a source of cumulus-convective heating in the western equatorial Pacific. The model takes the observed, fully three-dimensional structure of the circulation and calculates the change in the winds arising from tropical heating by the warm ocean. The waves all emanate from the western equatorial Pacific, moving eastward along (very approximately) great-circle paths. Plotted contours show the north-south wind (not including that of the time-averaged winds) in the upper troposphere; contour interval 0.5 meters per second. From Jin F and Hoskins BJ, *Journal of the Atmospheric Sciences*, 52: 307-319, 1995.

atmospheric circulation: baroclinic instabilities, which are the model of cyclonic storm development, tapping the potential energy of the atmosphere, and barotropic instabilities, which tap the kinetic energy of the mean atmospheric flow. In the stratosphere, very large-scale Rossby waves describe the undulations of the vortex sitting over the winter pole. They are excited by upward propagation of Rossby wave energy from the intense winter circulation below. The restoring force that gives us Rossby waves also inhibits mixing of fluid north and south, and in this way makes possible the ozone hole.

Before being completely carried away by the potency of this idea, however, we have to warn that Rossby waves are in competition with other forms of flow, particularly with turbulent, large-amplitude winds that are not waves at all. At the scale of the larger weather systems, the 'flow' dynamics and the 'wave' dynamics are nearly equal in importance.

Basic Principles

The conservation principle for potential vorticity (*see Dynamical Meteorology: Potential Vorticity*) helps to simplify the notion of Rossby waves, and also unifies them with the 'flow' dynamics just mentioned. Potential vorticity (PV) combines the dynamical effects of the spin of the Earth about its axis with the much smaller-scale spin of elements

4 days after peak blocking. The apparent wave train propagates from Australia south-eastward across the Pacific. 'Blocking' here means a period of at least 5 days when the 500 hPa pressure is at least 0.5 standard deviation above the norm. The source of the wave train is thought to be the cumulus convection in the western equatorial zone, and hence there is a strong correlation of warm El Niño periods with blocking patterns at higher latitude (SOI index-blocking index correlation reaches -0.8 during the past 15 years). From Renwick JA and Revell MJ, *Monthly Weather Review*, 127: 2233-2247, 1999.

of the fluid about their centers. The spin of the Earth is, in effect inherited by the fluid atmosphere, and concentrated into small, spinning storms. PV also incorporates effects of the sloping isentropic surfaces, and the shape of the atmosphere's lower boundary. PV is thus a combination of small-scale fluid properties and large-scale environmental properties. Using the time derivative following the motion of the fluid (and with PV denoted by q), eqn [1] describes conservation of q following an ideal fluid element as it moves.

$$\frac{Dq}{Dt} = 0 \text{ where } q = \frac{f + \zeta}{h} \quad [1]$$

We have neglected friction, heat sources, and the effects of small, unobserved, turbulence here. PV is related to, yet more general than, the conservation of angular momentum encountered in the physics of spinning bodies. Here f is known as the Coriolis frequency. It is equal to twice the vertical component of the Earth's rotation vector (eqn [2], where Ω is the Earth's rotation rate (in radians per second) and ϕ is the latitude).

$$f = 2\Omega \sin \phi \quad [2]$$

Rather like a Foucault pendulum, the horizontal flow of the atmosphere picks out the vertical component of the rotation vector. ζ is known as relative vorticity (also resolving just the vertical component), which is twice the average rate of spin of small fluid elements about their centers; finally, h is the thickness of the layer of fluid, measured along the local vertical. As the fluid moves, it may trade off its small-scale spin (ζ) for its large-scale 'planetary spin' (f) by moving north or south: this is the essence of the Rossby wave. The thickness h may also be involved in the trade-off, helping or hindering it. A key quantity suggested by this description is the rate of change of the planetary spin, f , with latitude. This is known as β , defined as in eqn [3],

$$\beta = \frac{2\Omega \cos(\text{latitude})}{a} \quad [3]$$

where a is the Earth's radius, 6380 km (± 3 km). In SI units (meters $\times$ seconds) $^{-1}$, $\beta = 2.28 \times 10^{-11}$ multiplied by the cosine of the latitude. The 'beta-effect' is the name given this systematic gradient {variation} of PV provided by the spherical shape of the planet.

For a single layer of homogeneous fluid like ordinary water, h is the full depth of the layer: in this case we can have 'topographic Rossby waves' due to a slope of the solid bottom (variation of h), instead of the spherical shape of the Earth (variation of f). For a fluid layered with significant density variation (like the ocean or the atmosphere), the conservation of PV can be applied to a small fluid element, with h being the vertical thickness of the density layer. PV thus has a dual nature: it gives a conserved quantity at each point in the fluid and yet it also has a vertically averaged sense of being conserved, for an entire layer of fluid {this known as 'barotropic' PV}. Remarkably, many aspects of atmospheric Rossby waves can be largely understood in terms of the latter, simpler, barotropic PV. A map of PV throughout a fluid can be mathematically 'inverted' to give much of the velocity and density field (though a part of the flow for which PV = 0 is 'invisible' to PV analysis), and a map showing curves of constant PV for the time-averaged state of the atmosphere (with mean winds and mean temperatures) describes the

restoring effect for the waves. Indeed, fluid can flow rather freely along such curves of constant PV (known as 'geostrophic contours' or just mean-PV contours). The persistent variation of f with latitude tells us that east-west winds are favored on a rotating planet, and north-south winds may often lead to waves.

It is exceptionally handy that the fundamental dynamical quantity (PV) for the atmosphere is nearly unchanging, like the concentration of a trace chemical following the circulation of the air. This adds great intuitive resource because we can actually see tracers move, distort and mix, and we can 'see' PV in models and observations behaving in many of the same ways. Rossby waves are the shimmering of the mean PV contours of the atmosphere.

Barotropic Rossby Waves

For the case when the layer thickness, h , is effectively constant, we have $\zeta + f$ as the active part of PV. An ideal situation would be a single layer of incompressible fluid (like water) of constant depth. It happens that this idealization is immensely powerful, providing approximate wave solutions for the more complex environment of the stratified atmosphere.

Carl-Gustav Rossby, working at MIT in 1939, introduced the useful approximation for middle latitudes, known as the 'beta-plane.' It approximates the spherical Earth locally by a plane tangent to it, allowing simpler mathematics using Cartesian coordinates to replace the full spherical coordinates. Far from the tropics, the Coriolis frequency can be approximated as in eqn [4] where γ is the north-south position, measured from some mean latitude γ_0 .

$$f = 2\Omega \sin(\text{latitude}) \approx f_0 + \beta\gamma \quad [4]$$

Rossby Waves in an Atmosphere at Rest

Let us now use these ideas to construct a basic Rossby wave for an atmosphere otherwise at rest (without the usual east-west mean winds). Newton's second law, the conservation of momentum, gives us equations in both horizontal directions, x (eastward) and y (northward), for the corresponding velocity components u and v . If we set up the wave with purely north-south motion, $u = 0$, the momentum equations express an east-west force balance between the pressure gradient and the Coriolis force (known as geostrophic balance; eqn [5]) and a north-south force balance between acceleration per unit mass and pressure gradient (eqn [6])

$$-fv = -\frac{1}{\rho} \frac{\partial p}{\partial x} \quad [5]$$

$$\frac{\partial v}{\partial t} = -\frac{1}{\rho} \frac{\partial p}{\partial y} \quad [6]$$

Eliminating the pressure, p , between these two equations gives a wave equation for v (eqn [7]).

$$\frac{\partial^2 v}{\partial x \partial t} + \beta v = 0 \quad [7]$$

In eqn [7], where $\beta = \partial f / \partial y$, approximated as a constant in the equation. Assuming a wave of the form $v = A \cos(kx - \omega t)$, we substitute into the wave equation to find eqn [8].

$$\omega = -\frac{\beta}{k} \quad [8]$$

This key relation between the wavenumber k {which is 2π divided by the wavelength} and the frequency ω tells us that longer waves have higher frequency, and their propagation speed, c (the phase speed) is westward, with magnitude β/k^2 . In more familiar wave systems, for example, nondispersive sound waves, light waves, or waves on a vibrating string, the frequency varies directly with wavenumber and the propagation speed c is a constant. The character of dispersive waves is that a localized disturbance breaks into sine-wave components of gradually varying length (as with a pebble thrown into a pond). By contrast, nondispersive waves like sound and radio waves preserve the properties of isolated pulses, making possible communication to a great distance.

The Restoring Force for Rossby Waves

The force balance above shows that the pressure gradient is the restoring force for the waves, and that this north–south gradient arises indirectly from the east–west force balance. Large-scale winds, slowly varying in time (relative to a day) are nearly geostrophic (see **Dynamical Meteorology: Quasigeostrophic Theory**), with pressure gradient balancing the horizontal Coriolis force (which is at right angles to the wind). Because the Coriolis frequency and hence the Coriolis force on the air parcels increase with latitude, so too do the north–south pressure variations, and these provide the needed restoring force, driving the acceleration of the wind.

The principle of conservation of PV gives a clearer description of the workings of the Rossby wave. An air mass that moves northward in a standing wave pattern, conserving the sum $\zeta + f$, will have to develop negative spin or vorticity, ζ , as it encounters smaller values of f found at high latitude. This anticyclonic spin matches the northward velocity, west of the parcel, and the southward velocity to its east (Figure 6),

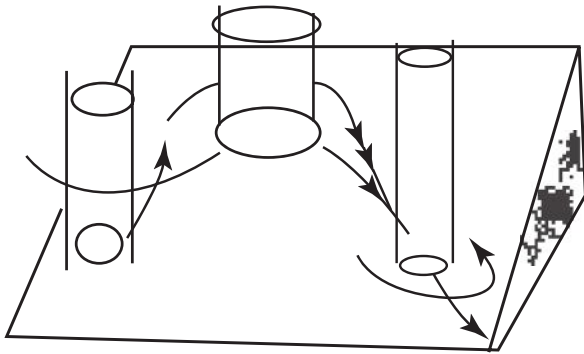


Figure 6 Showing how a parcel of air moved northward in a standing wave, conserving its potential vorticity $\zeta + f$, develops negative (anticyclonic) spin ζ , which affects fluid to the east and west, reinforcing the north–south winds.

enforcing the basic wave pattern. In downwind regions where the wave has not yet penetrated, this spin will extend the pattern downwind at a rate twice the mean westerly wind speed.

When the wave is oriented in an arbitrary direction, it has two wavenumbers or, more succinctly, a wave-vector $\mathbf{k}$ with components k (east–west) and l (north–south). The corresponding equation (for a fluid without mean east–west winds) is now (written in terms of the stream function, ψ) given by eqn [9].

$$\frac{\partial}{\partial t} \nabla^2 \psi + \beta \frac{\partial \psi}{\partial x} = 0 \quad [9]$$

The quantity ψ is a close approximation to the pressure or geopotential height field, as well as giving the horizontal velocities ($u = -\partial \psi / \partial y$, $v = \partial \psi / \partial x$), and $\nabla^2 \psi$ is the horizontal Laplacian, $\partial^2 \psi / \partial x^2 + \partial^2 \psi / \partial y^2$. The equation is a form of conservation of PV, rewritten as $PV = \nabla^2 \psi + \beta y$. The frequency relation, found by substituting a wave of the form $\psi = \exp(ikx + il y - i\omega t)$ in the equation, is now given by eqn [10], where α is the direction of $\mathbf{k}$ with respect to east.

$$\omega = \frac{-\beta k}{k^2 + l^2} = \frac{-\beta \cos \alpha}{|\mathbf{k}|} \quad [10]$$

This relationship is plotted in Figure 7. A key property of dispersive waves is the velocity of energy propagation, known as the group velocity. This vector has magnitude (5) equal to the westward component of phase speed, and it points in a direction 2α . This vector has magnitude (eqn [11])

$$\frac{\beta}{|\mathbf{k}|^2} \quad [11]$$

The group velocity is perpendicular (pointing inward) to the circles of constant frequency in Figure 7. A remarkable

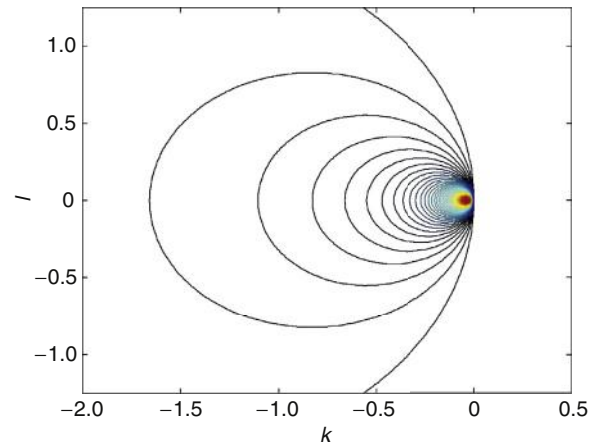


Figure 7 Frequency–wavenumber diagram for Rossby waves in a fluid without mean east–west winds. Here we have wavenumbers k and l in the east and north directions, respectively, as horizontal and vertical axes. Curves of constant frequency, are plotted, with frequency increasing toward the origin. The waves reverse the normal property of nondispersive waves, having higher frequency for longer wavelength. Energy propagation for a given wave is directed perpendicular (and inward) to these curves. Thus, for example, a wavevector pointing from the origin upward/leftward (northwestward) has wave crests and winds directed north–east–south–west, and energy propagation southward.

property of Rossby waves is that their wave crests always move westward relative to the air, with westward speed also given by expression [11], even though their energy can propagate in any direction. If a steadily oscillating force is exerted at a point in the fluid, it will radiate Rossby waves in all directions (Figure 8). The theory gives the solution as a form of Bessel function {known as the Hankel function of the second kind}, multiplied by a westward-traveling sine wave, written explicitly as $\psi = \exp(i\beta x/2\omega - i\omega t)H_0^{(2)}(\beta r/2\omega)$. The wave crests form parabolas on the horizontal plane that sweep westward while collapsing toward the westward part of the x -axis.

Effect of Mean Zonal Winds

An east–west wind, U , Doppler-shifts the waves; if we restrict our interest to standing or stationary waves, we replace time derivatives with x -derivatives, or equivalently replace ω by $-kU$ in the frequency relation. Instead of eqn [10], this gives eqn [12].

$$kU = \frac{k\beta}{k^2 + l^2} \quad [12]$$

And instead of the wave equation [9], we have eqn [13].

$$\nabla^2 \psi + \frac{\beta}{U} \psi = 0 \quad [13]$$

Evidently, a westerly wind generates Rossby waves, all of the same wavelength (eqn [14]).

$$\frac{2\pi}{\sqrt{k^2 + l^2}} = 2\pi \sqrt{\frac{U}{\beta}} \quad [14]$$

We see from this formula that faster winds make longer waves, and in regions of easterly wind ($U < 0$) there are no simple waves at all. The group velocity is found by adding the intrinsic group velocity of the Rossby wave to the mean wind velocity. Using relation [11], the intrinsic group velocity has

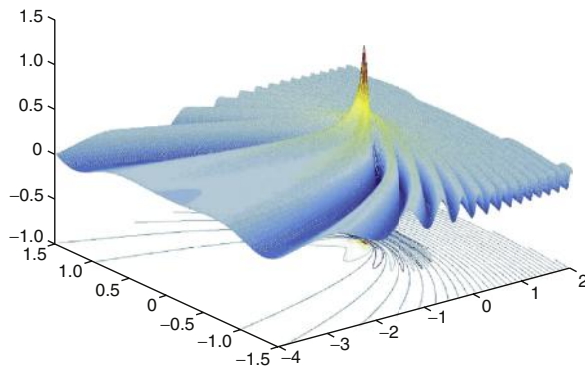


Figure 8 Rossby waves generated in a fluid, initially at rest, by an oscillating force where the field has its highest peak. This is a perspective view of the pressure, or streamfunction, seen from the south-west. Short Rossby waves have eastward energy velocity, and hence appear to the east, while long waves with nearly east–west velocity extend west of the forcing. On the plane beneath, the contour plot of the same field shows the parabolic shape of the wave-crests. With time they sweep westward and ‘collapse’ on the latitude line emanating westward from the forcing.

magnitude that just equals U . Taking account of the direction of energy propagation, the waves’ energy propagates with velocity $2U(1 + \cos \theta)$, in a direction θ measured with respect to east. They fill an ever-expanding circle downstream of the source of waves.

In the more general situation of propagating (rather than stationary) waves on a westerly wind, the two above analyses combine to give an east–west wavespeed ω/k which is $U - (\beta/|k|^2)$. This is Carl-Gustav Rossby’s ‘trough formula.’

Forcing and β -Plumes

There is one important signal missing from the above discussion. Waves with crests (and winds) running nearly east and west have $k \ll l$ and have significant energy velocity even though their intrinsic frequency is small. At very low frequency, keeping the wavelength constant, the direction of energy propagation is due westward, and is sufficiently fast to overcome a westerly wind. This produces what can be called ‘ β -plumes,’ which are nearly steady cells of circulation reaching westward from their point of generation. The above expression for the energy velocity, $\beta/|k|^2$, tells us that for north–south length scale L , the circulation plume will propagate westward at speed βL^2 relative to the mean wind. In the case of flow over a mountain range, this plume can reduce the flow upwind of the mountains, expressing a blocking of the wind by the mountain. With a little friction added, the β plume can become a steady, closed circulation. A solution for the streamlines due to a point-source of PV, Figure 9,

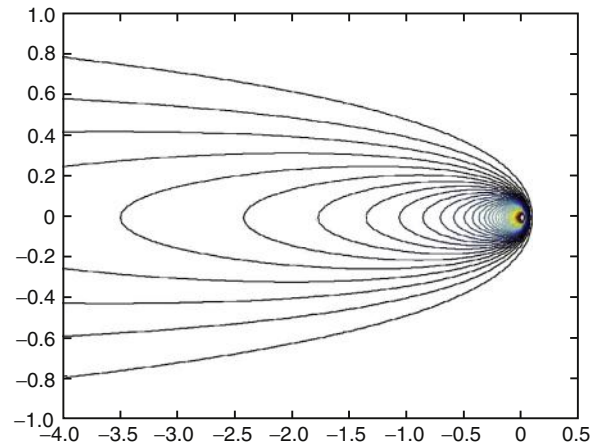


Figure 9 The β -plume is a circulation generated by a small region of forcing (e.g., heating of the atmosphere with convective clouds). Here, in the middle latitudes, on a β -plane the circulation reaches west of the forcing forming an elongated gyre of streamlines. Rossby waves of very low frequency are active in setting up this western extension of the circulation. The forcing at the origin could be a small region of heating by cumulus convection (driving divergent flow in the upper troposphere, and hence anticyclonic circulation and convergent winds, with cyclonic circulation in the lower troposphere), or a mechanical ‘twisting’ force concentrated at the origin. This is a plot of the theoretical result, which is a Bessel function with imaginary argument multiplied by an exponential.

complements the oscillating Rossby wave point source in [Figure 8](#). Here the forcing would produce a circular vortex in absence of the β effect (and indeed, near the forcing the streamlines are circular). But Rossby wave propagation makes the vortex lop-sided, extending far to the west.

The beta effect is particularly strong in the tropics and, together with gravity wave and Kelvin wave dynamics, helps to shape circulations there. An idealized steady heat source at the Equator, [Figure 10](#), causes a low-level convergence of the winds, which rise into the heating region. This is for an atmosphere initially at rest. Yet this convergence forms a double cell of circulation west of the heating region, which is again a β -plume. The winds are drawn in from east of the heating in another 'plume' that is shaped by Kelvin wave dynamics (see **Dynamical Meteorology: Kelvin Waves**). Motions that involve strong horizontal temperature variations and corresponding vertical velocity variations (through the thermal-wind balance) are termed baroclinic (see **Dynamical Meteorology: Baroclinic Instability**), whereas winds with little vertical variation in the pattern of the velocity are termed equivalent-barotropic. The winds produced by heating here are quite baroclinic, yet with a significant equivalent-barotropic contribution. When the observed mean atmospheric winds are added to this model, the very different pattern of nearly stationary Rossby waves

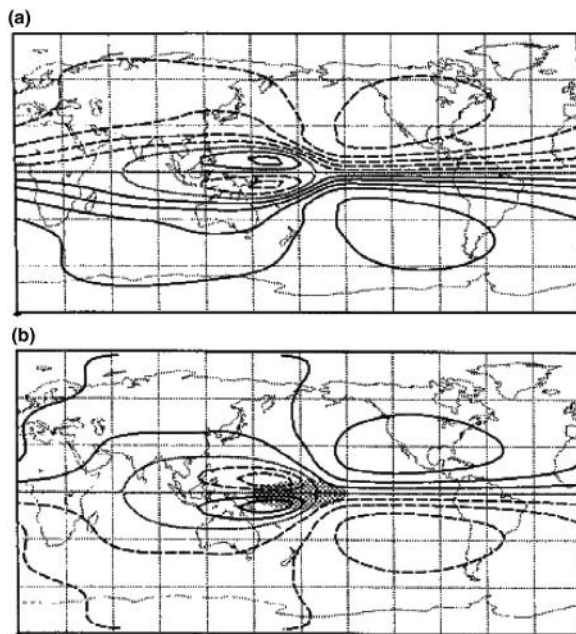


Figure 10 On the Equator, the pressure response of an atmosphere without mean winds to heating by cumulus convection (shaded region) involves both Rossby and Kelvin waves. Here the β -plume draws air in from the west to feed the low-level convergence, while drawing in from a 'Kelvin plume' to the east. Panel (a) shows the upper troposphere, where the updrafts diverge outward, and panel (b) shows the lower atmosphere, where convergent winds feed the updraft. Because the sign of f changes across the Equator, we have a double-celled pattern rather than the single circulation cell in [Figure 8](#) (cyclonic wind cells below, anticyclonic above in the β -plume west of the forcing). From Jin F and Hoskins BJ, *Journal of the Atmospheric Sciences*, 52: 307–319, 1995.

appears downwind of the forcing, [Figure 5](#): yet we can still see the β -plume upwind.

Horizontal Propagation: Refraction, Waveguides, and Instability

Rossby waves follow propagation pathways ('rays') in an approximate sense, and these rays are bent by variations in large-scale PV from the time-averaged winds, the thermal structure, and topography of the solid Earth. Preferred paths ('waveguides' or 'ducts') of the waves are created in this way, for example, in the core of the westerly winds, along the Equator, and in the upper polar troposphere.

We have seen some of the topographically induced standing wave patterns of the atmosphere that at least in part are attributable to Rossby wave dynamics. Consider now the kind of perturbation {alteration, often weak or slight} to the mean standing-wave circulation arising from an additional source of waves, as in [Figure 5](#). For example, during El Niño events in the Pacific, the extraordinarily warm sea-surface temperature can excite waves in the atmosphere passing above. Yet it is found that the wave pattern generated is sensitive to the location (east and west) of the forcing region; this would not be the case for a simple Rossby wave problem. We must generalize the restoring effect for Rossby waves to include PV gradients in the mean winds themselves. In eqn [13], β is replaced by $dq/dy = \beta - \partial^2 U / \partial y^2$ for this barotropic model. Now consider what this does. The curvature of the $U(y)$ profile (which is the gradient of relative vorticity) is subtracted from β . For an easterly jet, the $q(y)$ profile now has a 'flat spot' with small PV gradient cut like a plateau in the ' β hillside.' A westerly jet gains a concentrated gradient at the core of the jet. The concentration of vorticity in the mean flow augments the β -effect for the case of the westerly jet: it is the simple sum of the planet-scale PV and the vorticity of the time-averaged winds that counts. The ray paths describing propagation of groups of short Rossby waves will bend toward from regions of large β^* , as defined in eqn [15].

$$\beta^* = \left[\frac{dq/dy}{U} \right]^{1/2} \quad [15]$$

In this way waves will be deflected away from an easterly jet and trapped inside a westerly jet, which thus acts as a waveguide.

Three panels in [Figure 11](#) show the observed mean winter westerly wind, the barotropic PV gradient, dq/dy , and the effective restoring term β^* . Using similar, but three-dimensional, fields from observations, linear Rossby waves were generated by a stationary source of vorticity {a twisting force in a small part of the fluid} as an exploratory computer experiment ([Figure 12](#)). The size of the forced region is about 30° of latitude. The waves indeed follow the westerly Northern Hemisphere jet. Preferred propagation into and out of the tropics occurs ([Figure 11](#)) where the zonal winds are weak or westerly, rather than the more usual easterly winds. Lines along which $U = 0$ lead to infinite values of β^* , and these 'critical lines' tend to reflect Rossby waves, after a certain amount absorption of their wave activity and momentum. Computer models are sensitive to the way such regions are handled, and to the levels of frictional damping assumed

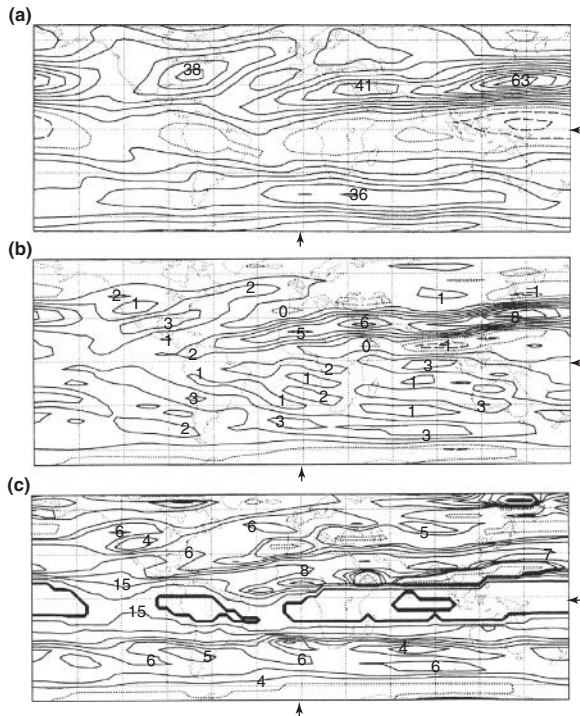


Figure 11 (a) Mean westerly wind speed at 300 hPa in Northern Hemisphere winter. (b) Mean north-south barotropic potential vorticity gradient, $\beta - \partial^2 U / \partial y^2$. Note the large values in the mean westerly jets. (c) Effective restoring force for Rossby waves, β^* . Heavy line in (c) is a concentration of high values of β^* at critical lines. Zero lines are dotted, negative are dashed. The Greenwich meridian is marked with an arrow. From Hoskins BJ and Ambrizzi T, *Journal of the Atmospheric Sciences*, 50: 1661–1671, 1993.

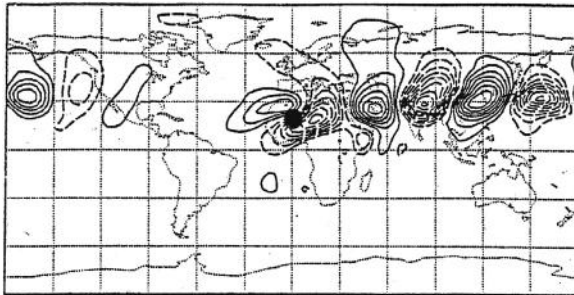


Figure 12 Meridional wind anomaly at day 10 for circular heat source at 20° N, 0° E, with December–January mean flow. This is a three-dimensional model calculation, but with a wave environment dominated by the barotropic PV (Figure 11). From Ambrizzi T and Hoskins BJ, *Quarterly Journal of the Royal Meteorological Society*, 123: 919–929, 1997.

in their formulation, leading to lingering uncertainty about many features of the circulation.

While the idea of Rossby waves having preferred waveguides is attractive, a slightly different interpretation of this experiment is that the jetstream itself is prone to meandering. It has a strongly concentrated PV gradient, and when disturbed it develops intense oscillations which have both stable and unstable components. This is not quite the same thing as

a Rossby waveguide, and perhaps better describes the ‘capture’ of Rossby wave energy by the underlying circulation.

This idea has been developed extensively, and it is found that the time-averaged winter winds can actively contribute to the wave field, exhibiting barotropic instability that can resemble a simple train of Rossby waves. Thus, in the PNA pattern, describing the atmospheric response to a warm tropical Pacific Ocean energy can be added to the wave train, transferred from the large-scale circulation, en route to North America. The weakly unstable modes are not easy to sort out because of the more rapidly growing baroclinic instabilities of the system.

Energetic Eddies and Breaking Rossby Waves

The atmosphere is made more complex, however, by the great strength of the winds and the large temperature range. With winds as strong as those typically observed, nonlinear effects (neglected in the simplest theory of Rossby waves) are strong, and lead to a large-scale form of turbulence (‘geostrophic turbulence’). We have just described how Rossby waves interact with the general circulation. This goes both ways: the waves induce new and important arteries of general circulation and the circulation, through something like linear instability, generates waves and eddies.

A view of the great energy of the atmosphere is readily seen in daily satellite images. The potential vorticity field at the 320 K isentropic surface on 14 May 1992 (Figure 13) cuts through the tropopause, showing the highly convoluted path of the jet stream. Obviously this is not a ‘small perturbation to a westerly wind.’

General circulation aside, the eddies interact among themselves in ways that are not subject to the propagation rules of waves. Geostrophic turbulence {the strong interaction among eddies that are of the scale of storms and weather, and larger} obeys none of the rules of classical turbulence observed in nature: energy cascades {flows, moves} predominantly to large horizontal scale and into barotropic eddies {tall motions with flow that is similar over a wide range of altitude}. The merging of two eddies of the same sign {both either cyclonic or anti-cyclonic} is a part of this cascade, and can lead to concentration of the flow into a few, sparse, intense eddies. The ‘life-cycle’ of intensifying cyclonic storms in the atmosphere is a manifestation of this cascade toward tall eddies with reduced vertical shear and hence reduced potential energy, followed by horizontal propagation: barotropic Rossby waves propagate rapidly, and the turbulence cascade feeds energy into them. Typically the growing storm develops cut-off {closed patterns of} temperature and PV fields, whereupon it is more an eddy than a wave.

More generally, the interaction of the transient and stationary waves, as calculated from observations, is an expression of the complex mix of Rossby waves and eddies (stationary and traveling) found in the general circulation.

The chaotic development of geostrophic turbulence has similarities to the pairing of cyclones and their north-south movement, which reorganizes the larger-scale circulation. Purely barotropic turbulence also coexists with Rossby waves, and can feed energy into them: either through the jostling of

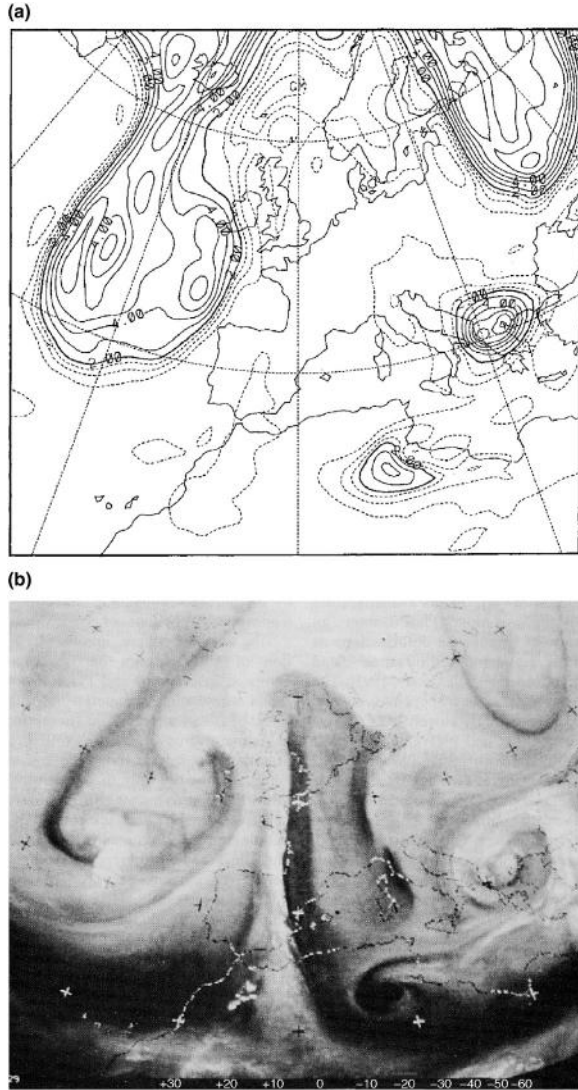


Figure 13 (a) Potential vorticity (PV) over Europe on the 3201 K surface at 1200 UT on 14 May 1992, from ECMWF analysis. The tropopause, dividing troposphere and stratosphere, cuts through the figure along the 0.5 to 1.5 contours. (b) Meteosat image of water vapor (5.7 to 7.1 mm wavelength radiation) at the same time as in (a). From Appenzeller C and Davies HC (1992) *Nature* 359: 570–572.

a pair of adjacent eddies or through distortion of an eddy by the larger-scale circulation.

When the fluid motion is very energetic, PV mixing occurs, and this directly forces changes in the large-scale atmospheric circulation by shifting momentum about. In some circumstances this can be described as a ‘breaking’ Rossby wave, in which lines of constant PV fold over sideways (still with nearly horizontal motion) and curl up.

More about Vertical Propagation

As we have seen, the defining properties of Earth’s atmosphere and ocean (and, indeed, the atmospheres of other planets) arise from the underlying rotation of the planet and its form

as an irregular sphere; to these we now add the extreme thinness and the nearly horizontal layering of air and water, with gradation from large density below to small density above. This density varies significantly over one scale height of the atmosphere, defined as $H_s = RT/g$ (R being the gas constant, T the temperature, and g the gravitational acceleration); it is typically 8 km. The layered stratification allows vertical variation of the horizontal wind, according to the thermal wind relation (see **Dynamical Meteorology: Overview**). This introduces a new class of *baroclinic* Rossby wave which propagates in all three dimensions. We now have a basis for understanding **Figure 2(b)**.

The extreme gravitational stability of these fluid layers is expressed in the buoyancy frequency (eqn [16]).

$$N = \left(\frac{g}{\theta} \frac{\partial \theta}{\partial z} \right)^{1/2} \quad [16]$$

This is the frequency of bobbing of a small air parcel, after it is given an upward or downward impulse (see **Gravity Waves: Buoyancy and Buoyancy Waves: Theory**). In the troposphere, the period $2\pi/N$ is typically 10–20 min; and in the lower stratosphere it is about 5 min. Here we use a local Cartesian coordinate system, (x, y, z) (east, north, up).

The form of the potential vorticity eqn [1] including vertical propagation is given in eqn [17].

$$\frac{D}{Dt} \left(\nabla^2 \psi + \rho^{-1} \frac{\partial}{\partial Z} \frac{\rho f^2}{N^2} \frac{\partial \psi}{\partial Z} \right) + \frac{dq}{dy} \frac{\partial \psi}{\partial x} = 0 \quad [17]$$

Here $Z = H_s \ln(p_0/p)$ is a slightly modified vertical coordinate, proportional to the log of the pressure. For waves with simple sine-wave variation east and west, and oscillating like a sine wave in time, we find a Helmholtz equation related to eqn [9] but now with respect to y and z (north and vertical). For stationary waves with mean westerly wind $U(y, z)$, and simplifying by considering N^2 to be constant, we let $\psi = \exp(ikx) \exp(Z/2H) \Phi(y, z)$ giving eqn [18a], with ν^2 and λ given in eqn [18b].

$$\frac{\partial^2 \Phi}{\partial y^2} + \frac{f^2}{N^2} \frac{\partial^2 \Phi}{\partial Z^2} + \nu^2 \Phi = 0 \quad [18a]$$

$$\nu^2 = \left[-\lambda^2 - k^2 + U^{-1} \frac{dq}{dy} \right], \quad \lambda = \frac{f}{2NH} \quad [18b]$$

Here ν^2 is the index of refraction for the waves; simple refraction occurs, in which rays bend toward regions of large index of refraction, and their paths can be sketched on a plot of ν^2 . Solutions of this equation are wavelike where ν^2 is greater than zero, and waves cannot penetrate regions where ν^2 is negative; only a small tail of the disturbances reach inside such regions. In particular, the waves can penetrate only where U falls in the range given in eqn [19].

$$0 < U < \frac{dq/dy}{k^2 + \lambda^2} \quad [19]$$

Jule Charney and Philip Drazin showed, in 1961, how in this analysis the zonal wind must be westerly and yet not too strong for stationary waves to propagate upward toward the stratosphere. The signature of this upward propagation is a three-dimensional wave-vector $\mathbf{k}$ that points downward and to the west, which means that the wave-crests are sloped

upward to the west, just as in Figure 2(b). This happens in a window of time in midwinter. It is remarkable that the very dispersive Rossby wave should, in the case of a single wave in x and t , be governed by a simple, single index of refraction (as if it were nondispersive). Tracing of the rays along which the waves propagate in the y – z plane (northward, upward) can be done immediately and easily (see model simulation in Figure 14). Errors due to east–west variations in the large-scale ‘mean’ atmosphere can be quite important and tracing of rays in all three dimensions would seem important, yet the very long wavelength east and west invalidates simple ray-tracing theory.

The upward propagation of Rossby waves into the stratosphere is particularly energetic in the Northern Hemisphere in winter, and the forces they exert act to decelerate the strong winds of the polar vortex. The important waves are very long, with just one or two wavelengths around a latitude circle. The dynamics of breaking Rossby waves and potential vorticity mixing by the turbulence that follows are a crucial part of the dynamics of the stratosphere (see **Middle Atmosphere: Planetary Waves**). Confinement of air over the wintertime pole by the strong potential vorticity gradient makes possible the chemically induced ozone hole.

Rossby Wave Forcing of the General Circulation

Transport of momentum and energy by Rossby waves and by geostrophic turbulence is a key to the ‘shaping’ of the general circulation. For example, the north–south circulation, averaged east and west, of the stratosphere, known as the Brewer–Dobson circulation, is related to strong radiative forcing, but is largely enabled by Rossby waves transporting easterly momentum upward from the troposphere. The lower atmosphere thus ‘pushes on the stratosphere above,’ and on our rapidly rotating planet this leads to both a deceleration of the polar vortex and flow at right angles and poleward. The Rossby wave force that does this has been diagnosed from observations (Figure 15).

The transmission of energy and momentum in Rossby waves has an elegant theory, with deep relations to classical physics. Potential vorticity, being conserved following air parcels (in an ideal sense, ignoring dissipation and external forcing), provides remarkable connections between the ‘stirring’ of fluid (as seen by marking it with colored smoke) and forces and momentum. These ideas can be discovered by the interested reader and he or she will be rewarded.

Concluding Remarks

There are many more manifestations of Rossby waves in the atmosphere than are described here, including ultralong modes with rapid easterly propagation, modes related to tidal forcing, and particularly the large derivative class of potential-vorticity waves involved in baroclinic growth of storms. Waves, instability, wave-induced acceleration of the mean circulation, life cycles of geostrophic turbulence, and potential vorticity mixing fill out the dynamics of the atmosphere, and Rossby waves are just the first step toward their understanding.

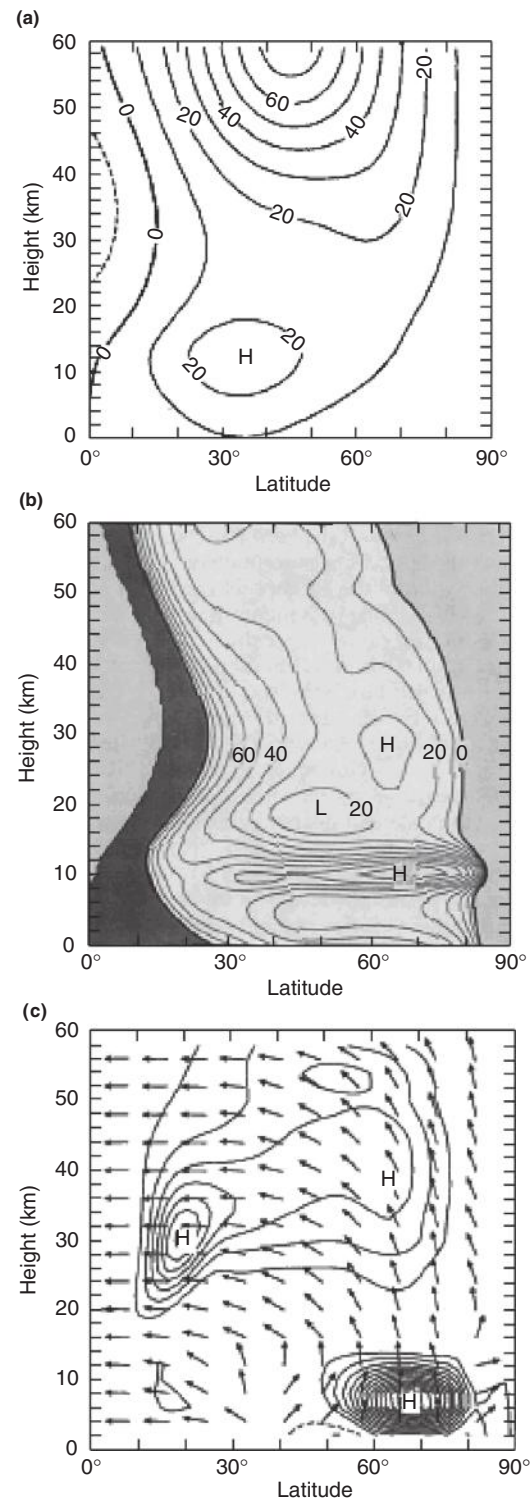


Figure 14 (a) Mean east–west wind used in the model. (b) Refractive index for steady zonal wavenumber 2. Dark band has high positive values. (c) Propagation paths for idealized Rossby waves, refracting toward the Equator (toward large refractive index, or slow wavespeed) with height. Contours show the convergence of the Eliassen–Palm flux, which exerts a westward force on the general circulation; it is particularly strong in the tropics where U goes to zero. Numerical model experiments from Chen P and Robinson WA, *Journal of the Atmospheric Sciences*, 49: 2533–2545, 1993.

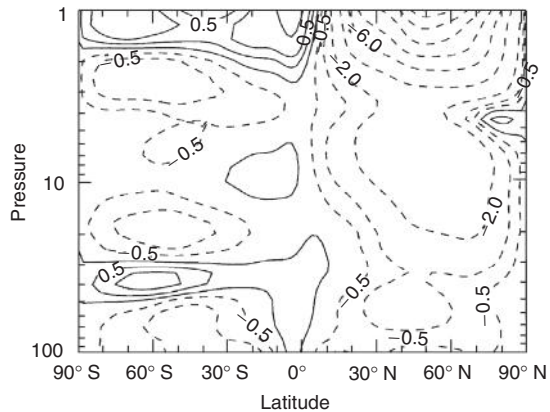


Figure 15 Wave-induced zonal force per unit mass, G , in units of meters per second, per day, for January 1993. Derived indirectly from the zonal momentum balance using radiatively derived meridional circulation and observed zonal mean wind. The dashed curves indicate the negative values expected from upward flux of east–west momentum transported by long Rossby waves. After Rosenlof K and Holton JR (1993) *Journal of Geophysical Research*, 10: 10465–10479. © 1993 American Geophysical Union.

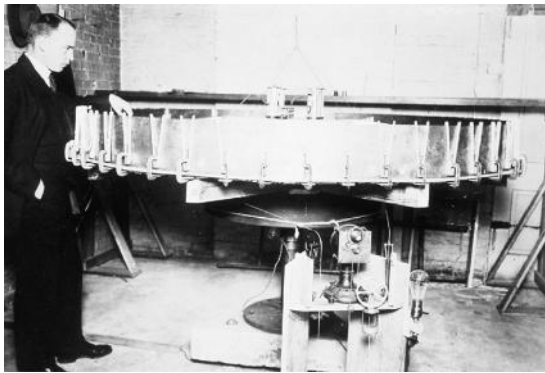


Figure 16 Carl-Gustav Rossby in 1926 or 1927, with a rotating platform designed to simulate the Earth's rotation and produce waves and 'weather.' NOAA historic photo archive, <http://www.photolib.noaa.gov/historic>.

Finally, we must remark that ideas in science should be tested when possible by physical experiments. On a rotating platform rather like Rossby's (Figure 16), we can easily produce a 'polar β -plane' in which the steady paraboloidal free surface of water in a cylinder provides a potential vorticity gradient very much like that in the middle and high latitudes of Earth. By oscillating a small glass cylinder up and down, squashing vortex lines, we generate an energetic Rossby wave (Figure 17), most visible toward east of the wavemaker. The short waves seen there have phase propagation toward their energy source, rather like the theoretically derived waves in Figure 3. The waves also transport momentum, driving easterly flow at most latitudes (as predicted by theory), and a westerly jet at the latitude of the forcing. They coexist with turbulent eddies. Furthermore, the orange polar cap region, despite strong wave activity, does not mix with the lower latitudes. This is the 'ozone hole effect,' by which potential vorticity gradients inhibit mixing just as they promote Rossby waves.

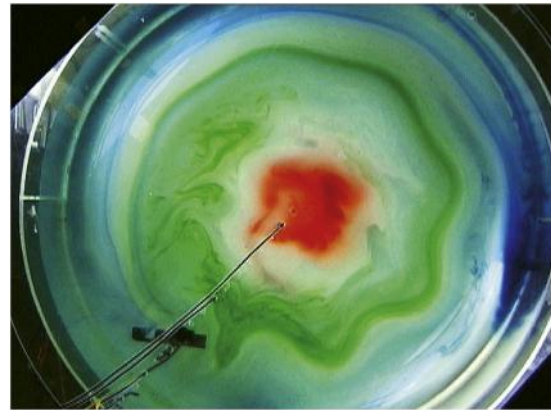
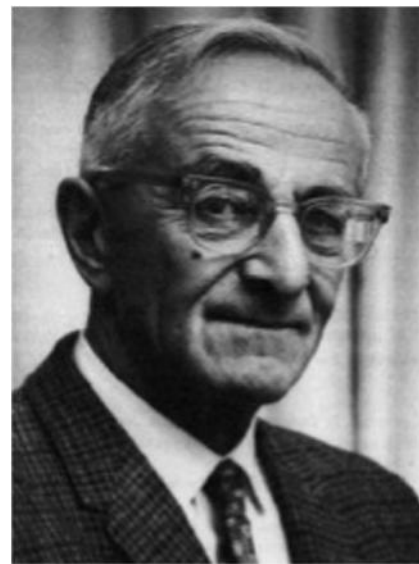


Figure 17 Rossby waves in the laboratory, as if viewed by a satellite above the North Pole. The wave source is at the lower left, an oscillating body. There is no preexisting circulation, but the waves induce easterly flow at most latitudes, and westerly flow at the latitudes near the forcing (as seen in the dye drawn into circles). At the North Pole, the red dye remains unmixed by the strong wave activity. Courtesy of the Geophysical Fluid Dynamics Laboratory, University of Washington.



Bernard Haurwitz

Figure 18 Bernard Haurwitz, one of the pioneers of Rossby wave theory.

Historical Notes

Large-scale waves on a rotating sphere were predicted with mathematical theory at the end of the nineteenth century by Hough and Margules. Bernard Haurwitz (Figure 18), after leaving Germany in the early 1930s, derived the essential properties of these modes in a 1937 paper. His status as 'enemy alien' in the United States in 1942 did not prevent the Army Air Corps from asking him to direct a research program on weather

forecasting. Carl-Gustav Rossby (see [Figure 16](#)) developed an elegantly simple approximation known as the β -plane, with which the derivation of the waves is greatly simplified. As so often happens in science, the full force of the earlier theory did not become apparent until long after its discovery. Rossby's influential first paper on the subject appeared in the *Journal of Marine Research* in 1939 (reminding us that these waves exist in both ocean and atmosphere). By emphasizing the simple propagation formula, Rossby successfully brought the ideas to bear on observations of the circulation at MIT in important papers in 1939, and through the war years. Before the era of computer simulation of the atmosphere, Rossby waves provided a foothold of dynamical theory in aid of weather

forecasting. Much later, in the last quarter of the twentieth century, Rossby wave dynamics has filled out like a powerful floodlight our understanding of the dark corners of atmospheric dynamics.

See also: **Dynamical Meteorology:** Baroclinic Instability; Kelvin Waves; Quasigeostrophic Theory. **Gravity Waves:** Buoyancy and Buoyancy Waves: Theory. **Radiation Transfer in the Atmosphere:** Radiation, Solar. **Tropical Meteorology and Climate:** El Niño and the Southern Oscillation: Theory.

Solitary Waves

JP Boyd, University of Michigan, Ann Arbor, MI, USA

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

Solitary waves are steadily translating disturbances in which nonlinearity and dispersion balance to create a disturbance of permanent form. The alternative name 'soliton' was coined because solitary waves are long-lived and elementary particle-like. Other coherent structures such as vortices or vortex pairs may exhibit a similar nonlinear persistence. Many distinct species of solitons are observed in both the ocean and the atmosphere. Solitary wave misconceptions are also reviewed.

Definition

In the narrowest sense, a solitary wave is a single, isolated wave crest, which propagates steadily without either steepening or widening. However, the concept has been broadened by the discovery of many new species of similar phenomena. Also, nonpropagating 'coherent structures,' especially vortices, have much in common with solitary waves. Thus, 'solitary wave' is no longer a phenomenon but a theme. The theme is nonlinear self-preservation of a crest or a vortex in the face of opposing, disruptive forces.

Dispersion, Frontogenesis, and the Bell Soliton

The left curve in Figure 1 is a schematic of the simplest solitary wave. It is called a 'bell soliton' because its shape resembles a church bell.

Waves are said to be 'dispersive' if the propagation speed c of a sinusoidal wave varies with the wavelength λ . It is possible to superimpose many sine waves of different wavelengths to make a bell shape, which is centered where all the crests are in phase. However, the bell shape rapidly disperses into an ever-widening patch of ever-shrinking ripples as illustrated in the upper right of Figure 1. In a marathon race, the runners are elbow to elbow at the start, but disperse into an ever-widening pack with the fastest runners in front and the slowest runners falling farther and farther

behind at the rear. Wave dispersion is the same as track-and-field dispersion: When waves travel at different speeds, the disturbance must spread over time unless some other mechanism intervenes.

One such mechanism is advective steepening. If the fluid velocity is proportional to height, then an initial bell shape will evolve a leading edge front (Figure 1). As the fast-moving tip overtakes the lower, slower fluid, the trailing (left) edge is stretched while the leading edge steepens ('frontogenesis').

In a solitary wave, dispersion and nonlinear steepening *exactly balance* so as to create a wave, which propagates *without change of shape*.

History of Solitary Waves

Solitary waves were discovered by the naval architect John Scott Russell in 1834. When a canal barge hit an underwater obstruction and stopped suddenly, Russell expected that the bow wave would dissolve into lots of little ripples through dispersion. Instead, a smooth, bell-shaped crest perhaps half a meter tall, independent of the cross-channel direction, emerged from the froth. On horseback, he followed the unchanging, steadily propagating crest for a couple of kilometers until he lost it 'in the windings of the canal.'

Russell later made solitary waves in a long, narrow water tank. Dropping a square block into the water produced a localized wave disturbance, which speedily organized itself into one or more solitary waves followed by a few small dispersing ripples.

Forty years later, Rayleigh and Boussinesq showed that

$$A(x, t) \approx \text{constant } B^2 \text{sech}^2(B[x - ct]) \quad [1]$$

where the phase speed c is proportional to B^2 , where B is a positive parameter that simultaneously controls the width, speed, and amplitude of the solitary wave.

Korteweg and deVries (1895) assumed that the wave was independent of the cross-channel direction, as approximately true both on the canal and in Russell's water tank, and also that the horizontal current was depth-independent ('shallow water approximation'). Thus, the only nontrivial spatial coordinate is down the channel. The surface height is then proportional to the solution $A(x, t)$ of the Korteweg–deVries (KdV) equation:

$$A_t + c_0 A_x + \mu A_{xxx} + \nu A A_x = 0 \quad [\text{KdV Eq.}] \quad [2]$$

where the subscripts denote partial differentiation with respect to the subscripted coordinate, and where the coefficients

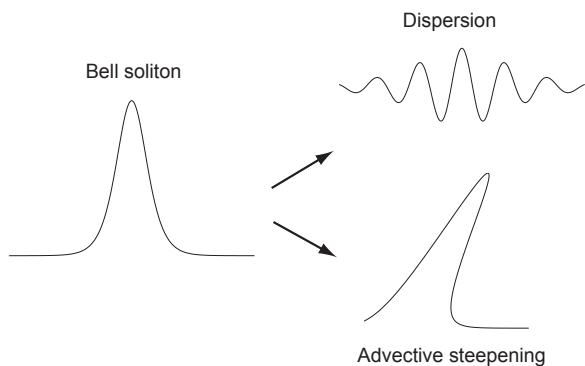


Figure 1 A bell-shaped crest (left) will dissolve into little ripples under pure wave dispersion; it will steepen and eventually break if advective steepening is unopposed by dispersion. In a solitary wave, dispersion and steepening exactly balance so that a bell-shaped curve propagates steadily without change of shape.

depend on the water depth, channel width, and the gravitational constant.

Today, it is known that the KdV equation is *generic* in the sense that it can be rederived for weakly nonlinear long waves in a wide variety of physics and engineering contexts, not just water waves. For 70 years after Korteweg and deVries, however, solitary waves were only a curiosity, mentioned in textbooks out of duty rather than conviction.

Zabusky and Kruskal in 1965 numerically integrated the KdV equation on a spatially periodic domain and discovered that a large amplitude initial sine wave dissolved into a sequence of solitary waves. These peaks collided elastically (i.e., without loss of energy from the colliding pair), as robust as if they were elementary particles. Zabusky and Kruskal therefore coined the name ‘soliton’ which is today used as a synonym for solitary wave.

A couple of years later, Gardner, Green, Kruskal, and Miura discovered the ‘inverse-scattering’ method. Although the KdV equation is *nonlinear*, the inverse-scattering method is an exact algorithm, which solves the general initial value problem for arbitrary time by solving a sequence of *linear* subproblems.

The KdV equation has special solutions, which consist of N solitary waves of different sizes where N is arbitrary. Tall solitons overrun shorter solitary waves, collide elastically, and then all rematerialize in their precollision amplitudes and widths.

The general KdV solution has two parts: a finite number of solitary waves plus a dispersing wavetrain. Except for rather special conditions, such as an initial height which is nonpositive, at least one solitary wave is *always* generated, even from wildly nonsoliton initial conditions, such as the bow wave of a canal barge. The flow spontaneously either steepens or disperses so as to *evolve* an exact balance between nonlinearity and dispersion in the solitary waves even when these two competing mechanisms are wildly imbalanced initially.

The following 10 years were the Golden Age of Solitons. Envelope solitary waves, kink solitons, and other new species were discovered, each solving an inverse-scattering-solvable generic partial differential equation in (usually) one space dimension. For a time, it seemed that the inverse-scattering method was the algorithm for everything. Then the thermonuclear bomb known as Chaos theory detonated, and it became clear that inverse scattering fails for most physical systems including the three-dimensional hydrodynamic equations.

The last quarter of the twentieth century has been the Golden Age of the Generalized Solitary Wave. Many species of coherent structures *almost* satisfy the classical definition of a solitary wave. Further, nonlinear steepening can preserve vortices and other moving structures even in the absence of wave propagation. Monopoles, modons, and weakly nonlocal solitons, described below, are important quasisolitons.

Solitary Vortices: The Vortex in a Strain Field

Nonlinear self-preservation is not limited to waves. Blobs of fluid are teased out into long, stringy filaments by the ‘strain’ or ‘deformation’ field created by distant vortices, pulled apart like

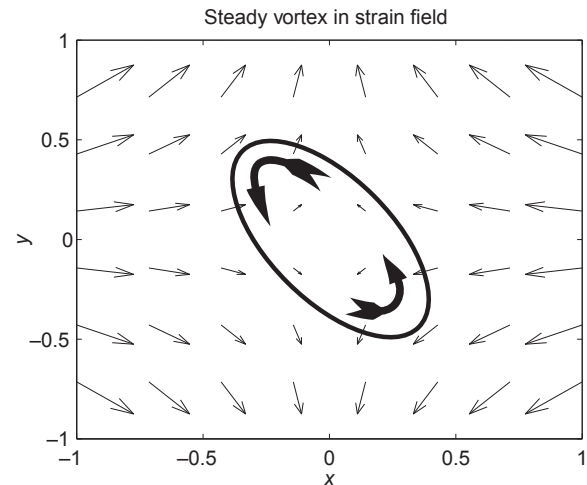


Figure 2 Thin arrows: a pure straining field. A patch of vorticity-free fluid would be irreversibly contracted toward the x -axis and stretched along the y -axis. If the fluid is a sufficiently strong vortex, the patch will be deformed into an ellipse (heavy curve) with its long axis oriented at 45° to the contraction and dilation axes of the strain field. The sense of rotation within the ellipse of uniform vorticity is shown by the heavy double-ended arrows.

taffy candy. However, if the blob is a sufficiently strong vortex, its own self-interaction will preserve it. A patch of uniform vorticity will distort into an elliptically shaped vortex with its long axis at angle of $\pi/4$ to the axis of strain, as shown in 1971 by Saffman and Moore (Figure 2).

The vortex in a strain field is not in any sense a wave. Nevertheless, a vortex is a coherent, isolated structure, which is preserved by nonlinearity in the face of disruptive mechanisms.

On a rotating earth, a large vortex will drift westward because of the beta-effect – a wavelike behavior. It is then impossible to speak of ‘solitary waves’ and ‘isolated vortices’ as separate species. The coherent structure is both wave and vortex.

A History of Isolated Vortices

Smoke rings were discovered by casual observation long before there was any science of fluids. Any smoker can make a ring merely by blowing out a mouthful of smoke. The smoke is trapped in a torus of fluid, which propagates away from the mouth under its own self-interaction. The propagating torus is a vortex ring, rotating about its narrow diameter.

In the 1860s, Tait showed that a trailing smoke ring can overtake and pass through another, then slow to be passed in its turn, as in the child’s game of leapfrog. This robust survival of leapfrogging rings is reminiscent of the durability of KdV solitons under collisions that was discovered by Zabusky and Kruskal a century later. Anticipating their analogy of coherent fluid structures with elementary particles, Lord Kelvin was inspired by Tait’s experiments to create a theory that atoms were vortex rings, and molecules were interlocking vortex rings.

Around 1900, Chaplygin and Lamb independently discovered analytic solutions for a pair of contra-rotating vortices, now usually called ‘modons’ or ‘Chaplygin–Lamb dipoles.’ Three quarters of a century later, Stern and Larichev and Reznik generalized these solutions to incorporate the beta-effect. Vortex pairs form spontaneously through random near-collisions of one-signed vortices in turbulence, through injection of river and estuary flows into the oceans, and a variety of other mechanisms.

Boyd generalized modons to vortex pairs that straddle the equator in the early 1980s. For small amplitude, these modons are well-described by the KdV equation and are classical ‘bell’ solitons in longitude with the usual structure of linear equatorial waves in latitude; most of the propagation is wavy, due to the Rossby beta-effect. As the amplitude increases, equatorial modons develop pockets of recirculating fluid, just as in Chaplygin and Lamb’s solutions, and the westward propagation is more and more due to the mutual interaction of the two vortices.

One can no more say that an equatorial modon is either a wave or a vortex pair than one can assert that the color purple is either red or blue.

Periodic Generalizations of Solitary Waves: Cnoidal and Polycnoidal Waves

The adjective ‘solitary’ is as misleading as ‘wave.’ Korteweg and deVries showed that the KdV equation has an exact elliptic function solution that they dubbed the ‘cnoidal wave.’ This is spatially periodic with an arbitrary period. In the limit of small amplitude for fixed period, the cnoidal wave is an ordinary cosine function. The large amplitude cnoidal wave has narrow peaks, which are well approximated by the sech^2 shape of the solitary wave (Figure 3). The soliton is just a limiting case of the cnoidal wave.

Eighty years later, Toda proved that the cnoidal wave is the *exact* sum of an infinite number of copies of the solitary wave

where the copies are evenly spaced, one centered on each spatial period. This sum-of-solitons relationship is true even in the small amplitude regime (foreground of Figure 3) where $A(x,0)$ is also well approximated by the cosine function. This nonlinear superposition principle has since been extended to many other wave equations.

Similarly, the KdV equation has exact analytical N -soliton solutions on an unbounded spatial interval, which have been extended to spatially periodic exact solutions. These generalizations, which are ratios of multidimensional theta functions, are called ‘ N -polycnoidal’ waves where the cnoidal wave is the special case $N = 1$. Polycnoidal waves depend on N independent phase speeds. It can be proved that the general solution to the KdV equation with periodic boundary conditions can be approximated to arbitrary accuracy by a polycnoidal wave of sufficiently large N and appropriate phase speed and amplitude parameters.

Thus, solitary waves need not be solitary. This is true for solitary waves in general and not merely for KdV solutions. Because solitons usually decay *exponentially* with distance from the core of the structure, a pair of solitons can be rather close and yet have a negligible dynamic interaction.

Weakly Nonlocal Solitary Waves

If the phase speed c of a coherent structure is multivalued in the sense that there are infinitesimal amplitude waves of some wave number k that have the same phase speed as the structure, then the solitary wave will not decay to zero at large distances from its center, but will instead radiate waves of wave number k . In many cases, the amplitude of the radiation is exponentially weak so that the structure behaves very much like a classical solitary wave. Such structures are called ‘weakly nonlocal’ solitary waves (Figure 4).

Ironically, water waves, the prototype of solitons, are weakly nonlocal. The solitary wave radiates capillary waves, but these were too small for J.S. Russell to observe.

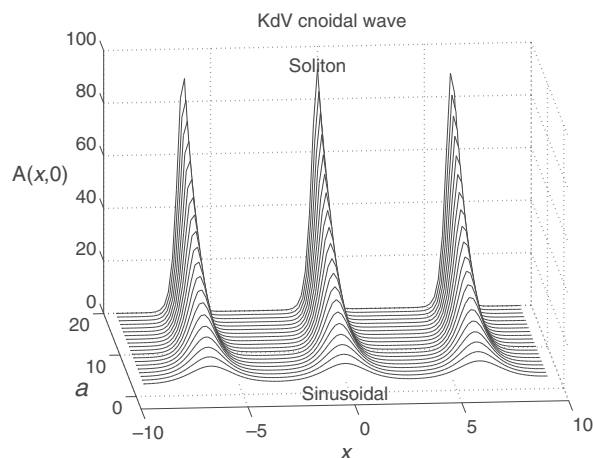


Figure 3 The KdV cnoidal wave $A(x,t)$ as a function of x (for fixed time $t = 0$) and amplitude a . For small amplitude (foreground), the cnoidal wave is sinusoidal. As the amplitude increases, the peaks become taller, narrower, and more soliton-like.

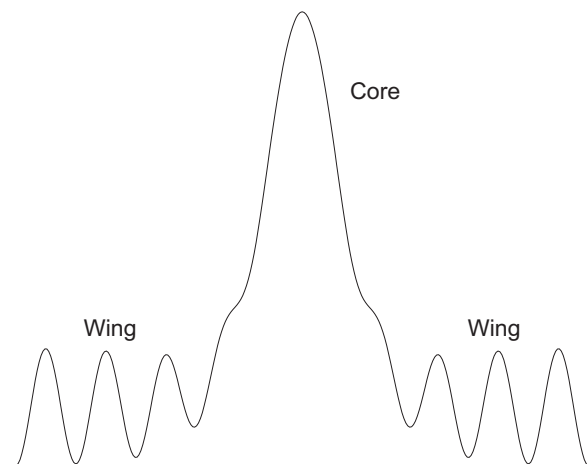


Figure 4 The ‘core’ of a nonlocal solitary wave is similar to a bell soliton, but the wave decays to small amplitude sinusoidal ‘wings’ instead of to zero.

Weakly nonlocal solitary waves are found in all branches of physics and seem to be just as common as classical, decay-to-zero solitons. Baroclinic vortices and solitons, for example, are usually nonlocal through weak radiation in the barotropic vertical mode.

A Bestiary of Solitary Waves and Coherent Vortices

Figure 5 shows the diversity of solitary waves and coherent structures. The six species illustrated are only a set of interesting creatures from a much larger zoo.

Bell solitons, such as those that solve the KdV equation, have been described above.

An ‘envelope solitary wave’ is the product of a sinusoidal ‘carrier wave’ with a slowly varying amplitude factor called the ‘envelope,’ which is dashed in the figure. The *envelope* solves the nonlinear Schrodinger equation and propagates at approximately the group velocity of infinitesimal waves of the wavelength of the ‘carrier wave.’

‘Breathers’ are solitary waves whose amplitude oscillates in time. The breather may be either stationary or propagating, but the period and amplitude of the ‘breathing’ oscillations never changes. The sine–Gordon, self-induced transparency, and ϕ^4 field theory equations also have breathers.

‘Kinks,’ also known in some contexts as ‘traveling shocks,’ occur in both inviscid models (such as the sine–Gordon equation) and viscous equations, such as Burgers’ equation and the Kuramoto–Sivashinsky equation. Viscous shocks seem a paradox since mechanical energy is being damped and yet the shocks, like solitary waves, are independent of time except for a steady propagation. The plateaus, extending indefinitely away from the shock, act as limitless reservoirs of energy to

sustain the shock. Real kinks do not extend indefinitely, but are consistent local approximations to coherent structures of finite width.

Vortices, whether monopoles or modons, are not always identified as solitary waves. If the diameter of the vortex is sufficiently small compared to the radius of the earth, then wave effects may be only a small correction to vortex dynamics. However, vortices often exhibit the same robustness and nonlinear self-preservation as KdV solitons.

Monopole vortices have vorticity which is everywhere of the same sign except perhaps for an annular ring surrounding the core. Modons are pairs of contra-rotating vortices as described earlier. These have a strong nonlinear translation, indicated by the hollow arrow in the figure, which is augmented or opposed by westward Rossby wave propagation.

Solitons and Coherent Vortices in the Ocean

In the Andaman Sea, tidal flow triggers regular trains of internal gravity solitons. These are visible in satellite photographs as long parallel streaks and are well modeled by the KdV equation.

When the Gulf Stream separates from the coast at Cape Hatteras, it develops unstable, amplifying meanders that eventually roll up into Gulf Stream rings. Most ‘cold core’ rings perish in a few months by reabsorption, but the few that drift south of the Gulf Stream live a couple of years in the Caribbean. This is an order of magnitude longer than the lifetime of a small amplitude Rossby wave of the same initial size (roughly 200 km in diameter). Similar coherent, long-lived eddies split from the Agulhas Current off South Africa.

The high evaporation of the Mediterranean Sea creates dense, salty water that flows out through the Straits of Gibraltar into the Atlantic Ocean. As it sinks to a depth of 1000 m, the anomalously hot and salty water rolls up into anticyclonic vortices called ‘Meddies.’ These spinning lens-shaped masses, perhaps 60 km in diameter and a kilometer thick, have lifetimes of half a dozen years or more.

Smaller coherent vortices, both monopoles and dipoles, are very common. Dipoles with long stems of vorticity are called ‘mushroom vortices’ from their shape. These are easily made in the laboratory merely by injecting a jet of fluid into a rotating tank. River outflows and melting at the edge of the ice pack are prolific generators of such vortex pairs, a few kilometers in diameter.

Solitary waves and isolated vortices and vortex pairs seem to be very important components of ocean dynamics. There is room here to catalog only a small subset of the rather wide range of observed species.

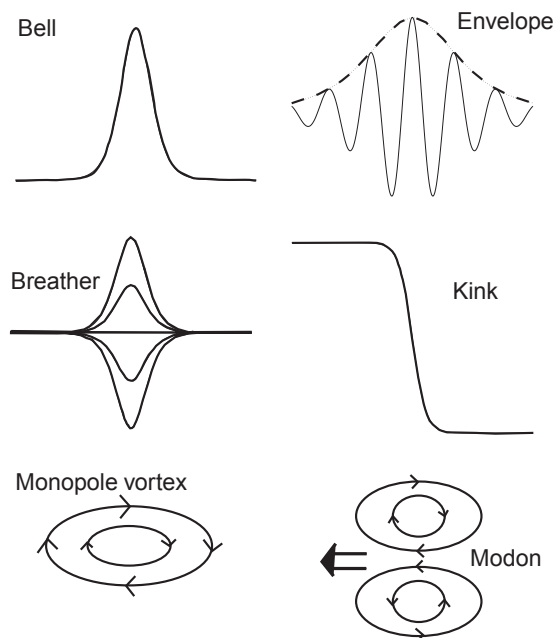


Figure 5 A selection of soliton species. All are snapshots at a single time except for the breather, which shows the oscillation at intervals of one-quarter of the temporal period.

Why Atmospheric Solitons Are Vertically Trapped

When a wave propagates upward into thinner and thinner air with weak or negligible dissipation, its amplitude u grows so that the energy flux remains constant even as the mass density decreases exponentially with altitude. A steady balance between nonlinearity and dispersion is impossible because the nonlinearity is *steadily increasing with height*. In contrast, the dispersion depends only upon the wavelengths of the

sinusoidal waves that comprise the wave pulse and thus does *not change* with height.

However, some waves are reflected by wind shear or static stability variations at some level, thus being trapped below the reflection height. Only such ‘vertically trapped’ waves can form solitons.

The ocean is a bounded fluid of almost constant density, so the difficulties of propagation to space and vertically increasing nonlinearity do not arise. This is one important reason why solitary waves are more readily observed in the ocean than the atmosphere.

Atmospheric Solitary Waves

New species of atmospheric solitary waves and new applications of previously studied types are inevitable. The following four examples are representatives of the diversity of ‘soliton thinking.’

Internal Gravity Waves: The Morning Glory

A low-level temperature inversion can create a layer of very stable air in the lowest kilometer or two of the atmosphere. Internal gravity waves are vertically trapped, and then can be reshaped by nonlinearity into a sequence of solitary waves. This mechanism operates all over the world. In particular such gravity solitons have been detected by Doppler radar and surface networks in Oklahoma. On the shores of the Gulf of Carpentaria in northern Australia, conditions are especially favorable to generate such soliton trains, and to further make the soliton crests visible through condensation. These trains of roll clouds are known as the Morning Glory (Figure 6).

The waveguide is leaky, so these solitary waves are ‘weakly nonlocal.’ Indeed, the upward leakage is so strong that recent articles have argued that convective forcing may be as important as soliton dynamics in sustaining the crests as they roll in from the Gulf.

Great Red Spot of Jupiter

The Great Red Spot (GRS) is an anticyclonic, eye-shaped vortex embedded in a shear zone between alternating east–west jets at about 20° S latitude on Jupiter (Figure 7). It has been spinning for at least three centuries with only minor fluctuations in amplitude and appearance. It is an isolated vortex in the sense that it is the only large feature in the shear zone. However, it cannibalizes smaller eddies that appear on the edges of the zone, and this may help to sustain the GRS against losses to viscosity and radiative damping.

A KdV theory has produced plausible agreement with observations; the eddy is both vortex and Rossby wave. Numerical models by G. Williams and P. Marcus offer a vivid explanation of GRS genesis. Generically, shear instabilities roll up into a string of vortices. Such chains of same-sign vortices are unstable to vortex mergers, and eventually a single large quasisteady vortex emerges as the end product of the instability. But Jupiter is banded with many alternating jets; why is there a strong vortex in only one of these, and only in the Southern Hemisphere?

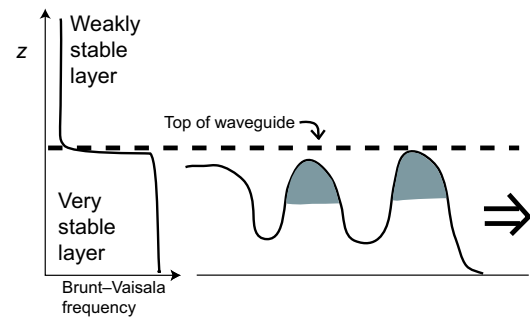


Figure 6 Schematic of the Australian Morning Glory. Left: the vertical profile of Brunt–Väisälä frequency after convection has created a very stable, well-mixed surface layer. When sea breezes collide at night over Cape Yorke Peninsula, this excites a gravity wave disturbance, which is trapped in the bottom layer by reflection from the interlayer boundary (dashed line) where the stability changes abruptly. The disturbance spontaneously organizes into an undular bore as it propagates over the Gulf of Carpentaria. Each peak is a solitary wave, and its updrafts cause condensation (shaded). The roll clouds may extend for over 100 km perpendicular to the direction of propagation, which is indicated by the large arrow.

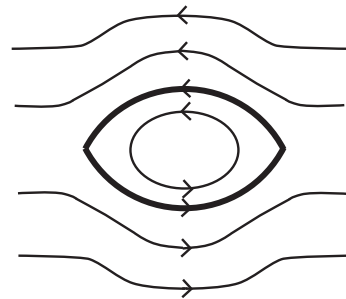


Figure 7 Schematic streamlines and velocity arrows of the GRS of Jupiter.

Hurricanes

Atlantic hurricanes originate through the roll-up of the shear instability of the East African jet. However, unlike classical solitary waves, hurricanes are strongly forced and damped; they perish rapidly after making landfall, strangled by the cessation of moist convection. Hurricanes are certainly coherent vortices, but the notion of ‘hurricane-as-soliton’ has generated little enthusiasm.

The reason is pragmatic. There are many similarities between hurricanes and other roll-up vortices such as the GRS of Jupiter. However, convective forcing and damping and small-scale embedded structures such as the hot towers lining the eyewall are so important to hurricane dynamics that the hurricane-as-soliton paradigm has been ineffective in forecasting or understanding hurricanes.

Atmospheric Blocking

Atmospheric ‘blocking’ is the formation of a quasistationary vortex or vortex pair over mountains which is sufficiently strong to block the usual midlatitude storm track, forcing weather systems to detour around the block. There are many

conflicting theories of blocking. However, the block is certainly a finite amplitude, quasistationary coherent structure that propagates westward against the prevailing westerlies so as to remain fixed above the mountain range. Much theoretical work has explored the idea that mountain chains are able to excite coherent blocks because the forcing is resonant: the forced solutions are very strong because they are close to unforced, finite amplitude vortex pairs (modons). Because the forcing is weak, the modon paradigm is much more useful for blocks than for strongly forced-and-damped vortices-like hurricanes.

Misconceptions

1. Solitary waves are necessarily waves.
Vortices and traveling shocks display the same nonlinear self-preservation as KdV solitons, and may move under a mixture of advection and Rossby wave dynamics.
2. Solitary waves are solitary; periodic waves have nothing to do with solitons.
Wave crests and coherent vortices may be very close *geographically* and yet have almost no *dynamic* interaction. Chains of crests that appear wavy or sinusoidal may in fact be weakly interacting solitary waves.
3. Solitary waves are small amplitude only.
This misconception was created by the derivation of the KdV and other simplified wave equations, which usually employ expansions in powers of the amplitude. However, numerical solutions show that the solitary waves do not magically cease to exist above a tiny limiting amplitude. Instead, solitons persist as a continuous family of solutions to such large amplitude that the soliton contains entrained fluid that is trapped within the structure as it propagates. 'Small amplitude' is a restriction of the mathematics, not physics.
4. Solitary waves are one-dimensional.
The KdV equation has only a single space coordinate. However, KdV theories often multiply the KdV solution, $A(x,t)$, by a function $Y(y)$ which is spatially confined because of Coriolis refraction, as true of equatorial solitary waves, or shear trapping, as in the GRS of Jupiter.
5. Solitary waves are unforced and inviscid.
Traveling shocks of Burgers' and the Kuramoto–Sivashinsky equations are solutions to *viscous* differential equations. Furthermore, weakly forced and damped nonlinear structures may be accurately approximated by unforced, undamped solitons. However, the soliton paradigm is not very useful when the forcing dominates the flow, as true of hurricanes.

The 'Leonardo–Kolmogorov Duality'

Leonardo da Vinci, who sketched turbulent streams and scribbled notes on what he called *turbolenza* in 1500, seems to have

known that turbulence could only be described (or painted!) as a mixture of coherent and random motion. Science progresses through a willful blindness to some aspects of a phenomenon to think deeply about others. (In a language of Papua New Guinea, this is 'mokita,' which means 'things we all know but agree not to talk about.')

Kolmogorov in 1941 made the first great breakthrough in turbulence by willfully ignoring the coherent structures, and pretending that turbulence is purely random.

The Voyager photographs of Jupiter showed instead what Frisch has called the 'Leonardo–Kolmogorov duality.' The Jovian atmosphere is neither completely coherent and predictable nor completely random. Instead, the 'Leonardian' GRS, which is a solitary vortex, coexists with a seething Kolmogorovian sea of billowing, fluctuating, random-appearing turbulence. The mystery of this soliton/random duality challenges our understanding today as it challenged Leonardo's pen five centuries ago.

See also: Dynamical Meteorology: Kelvin Waves; Rossby Waves. Gravity Waves: Overview.

Further Reading

- Ball, P., 1999. *The Self-Made Tapestry: Pattern Formation in Nature*. Oxford University Press, New York.
- Boyd, J.P., 1989. New directions in solitons and nonlinear periodic waves: polycnoidal waves, imbricated solitons, weakly non-local solitary waves and numerical boundary value algorithms. In: Wu, T.-Y., Hutchinson, J.W. (Eds.), *Advances in Applied Mechanics*, vol. 27. Academic Press, New York, pp. 1–82.
- Boyd, J.P., 1998. *Weakly Nonlocal Solitary Waves and Beyond-All-Orders Asymptotics: Generalized Solitons and Hyperasymptotic Perturbation Theory*. Mathematics and Its Applications, vol. 442. Kluwer, Amsterdam, 680 pp.
- Boyd, J.P., 1999. The devil's invention: asymptotics, superasymptotics and hyperasymptotics. *Acta Applicandae* 56, 1–98.
- Boyd, J.P., Haupt, S.E., 1991. Polycnoidal waves: spatially periodic generalizations of multiple solitary waves. In: Osborne, A.R. (Ed.), *Nonlinear Topics of Ocean Physics: Fermi Summer School, Course LIX*. North-Holland, Amsterdam, pp. 827–856.
- Ding, R.Q., Feng, G.L., Liu, S.D., Liu, S.K., Huang, S.X., Fu, Z.T., 2007. Nonlinear atmospheric and climate dynamics in China (2003–2006): a review. *Advances in Atmospheric Science* 24, 1077–1085.
- Grimshaw, R.H.J. (Ed.), 2007. *Solitary Waves in Fluids*. WIT Press, Southampton.
- Johnson, R.S., 1997. *A Modern Introduction to the Mathematical Theory of Water Waves*. Cambridge University Press, Cambridge.
- Lugt, H.J., 1983. *Vortex Flow in Nature and Technology*. John Wiley, New York.
- Nezlin, M.V., Snezhkin, E.N., 1993. *Rossby Vortices, Spiral Structures, Solitons*. Springer-Verlag, New York.
- Nihoul, J.C.J., Jamart, B.M. (Eds.), 1989. *International Liège Colloquium on Ocean Hydrodynamics*, no. 20 in *Liège Colloquium on Ocean Hydrodynamics*. Elsevier, Amsterdam.
- Osborne, A., 2010. *Nonlinear Ocean Waves and the Inverse Scattering Transform*. Academic Press, New York.
- Remoissenet, M., 1991. *Waves Called Solitons: Concepts and Experiments*, third ed. Springer-Verlag, New York.
- Soomere, T., 2007. Nonlinear components of ship wake waves. *Applied Mechanics Reviews* 60, 120–138.
- Van Dyke, M., 1982. *An Album of Fluid Motion*, second ed. Parabolic Press, Stanford.

Static Stability

JA Young, University of Wisconsin, Madison, WI, USA

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

Static stability is a property of the vertical thermal stratification of the atmosphere which acts through gravitational buoyancy forces to suppress vertical motions. Although it is defined locally, it influences all scales of motion and helps to control the weather systems and climate of the Earth. Small-scale turbulence, mesoscale motions, mountain influences, and large-scale atmospheric wind systems are all influenced by static stability.

In addition, regions of reduced static stability may become unstable in humid air masses, when the release of latent heat of condensation for ascending air creates positive buoyancy. Thus, the distribution of static stability influences the production of strong moist convection which is essential for severe weather events as well as the tropical driving of the global climate.

Introduction

Static stability measures the gravitational resistance of an atmosphere to vertical displacements. It results from fundamental buoyant adjustments, and so it is determined by the vertical stratification of density or potential temperature. It influences the dynamics of many kinds of atmospheric motions, which in turn are responsible for determining its variations.

Static stability is represented commonly by the square of the buoyancy frequency N , which plays a role in theories for flow instabilities, wave propagation, and forced motions. As summarized below, these theories apply to a wide range of spatial scales, from small-scale turbulence to convection, mesoscale motions, and large-scale circulations for which the ratio of N to the Coriolis frequency f is paramount. Finally, the stability concepts generalized to include condensation of water vapor are shown to have predictive value for prediction of severe weather organized by large-scale weather patterns.

Basic Buoyant Stability and Instability

The role of density fluctuations in a gravity field is best in the vertical component of the equations of motion. In an absolute sense, the gravity and pressure gradient forces are usually in a state of hydrostatic balance to within 1%. However, the slight imbalances account for vertical accelerations dw/dt which are often driven by buoyancy:

$$dw/dt = -\rho_0^{-1}[dp'/dz] + B \quad [1]$$

Here, w is the vertical velocity dz/dt , t is the time, $\rho_0(z)$ is the density of a static 'environmental' reference state, and a prime indicates deviation from that reference state. B is the buoyancy force per unit mass, given by $B = -\rho'/\rho_0 g$. For many buoyant motions, B is an upper bound on vertical accelerations dw/dt since the pressure gradient term tends to oppose B . The most useful approximate form for B is

$$B = (\theta'_v/\theta_{v0})g \quad [2]$$

where θ_v is the potential temperature augmented by a small (at most, a few °C) amount proportional to water vapor, reflecting the contribution of humidity fluctuations to buoyancy.

For a dry adiabatic vertical displacement δz , a parcel conserves θ_v so that $\theta'_v = -(\partial\theta_{v0}/\partial z)\delta z$. For a stable system, the squared frequency of oscillation is commonly equal to the restoring force per displacement or $B/\delta z$ in this case. Thus,

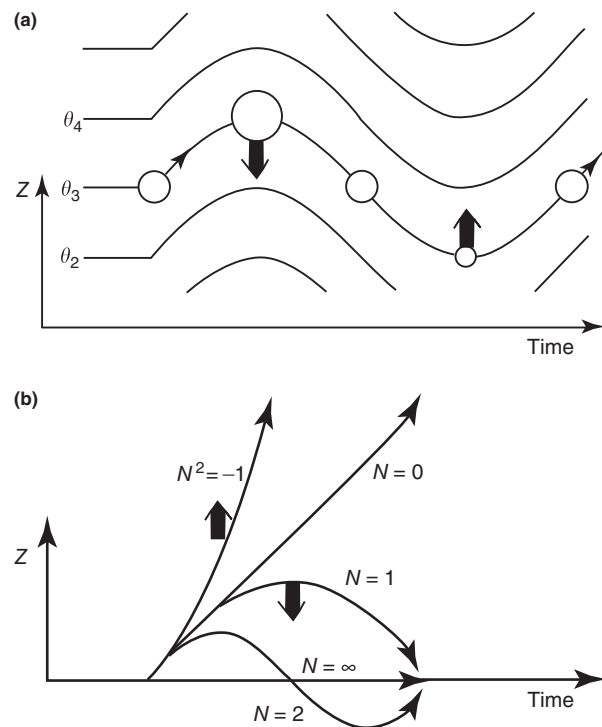


Figure 1 Simple buoyancy motions and varying environmental static stability. (a) Stable oscillation for $N = 1$. Isentropic surfaces are shown; increasing labels indicate warmer θ . Impulsive force creates initial vertical motion W (thin arrow), adiabatic displacements of θ surfaces, changes in air parcel volume (circles), and buoyancy force (vertical arrows). (b) Parcel motions for five stability conditions. Moderate stability: $N = 1$, shown in (a). Stronger stability: $N = 2$ stable oscillation has shorter period, smaller vertical displacements. Extreme stability: $N = \infty$ has no vertical displacement. Neutral stability: $N = 0$ has displacements growing linearly, with no restoring force. Unstable conditions: $N^2 = -1$ has buoyancy forces creating amplifying vertical parcel displacements.

if pressure effects are ignored in eqn [1], the simple buoyancy frequency N is given by

$$N^2 = (g/\theta_{v0})\partial\theta_{v0}/\partial z \quad [3]$$

N , also known as the Brunt–Vaisalla frequency, is determined by the vertical gradient of θ_{v0} or equivalently by the difference between virtual temperature lapse rate $-\partial T_v/\partial z$ and the dry adiabatic rate $\Gamma_d = g/c_p$. Unless conditions are super-adiabatic, θ_{v0} increases upward, corresponding to static stability. In this case, N^2 is positive and eqns [1]–[3] imply

$$d^2w/dt^2 + N^2w = 0 \quad [4]$$

It follows that the solution is a simple oscillation $w(t) = W \cos(Nt + \varepsilon)$, where W is the maximum vertical velocity amplitude and ε is a phase constant. W is a function of the initial motion $w(0)$ and buoyancy $B(0)$. $1/2 W^2$ is equal to the total energy, which is a constant as kinetic and potential energies oscillate out of phase between zero and their maximum values. The period is $2\pi/N$, typically about 10 min in the troposphere. Figure 1(a) shows the vertical oscillation, and its driving by buoyancy, which is a quarter cycle ahead of the parcel displacement δz . The buoyancy oscillation is analogous to that of a spring, so N^2 is equivalent to the ‘stiffness’ of the atmosphere when it is subjected to vertical displacements. The stiffness increases with the closeness of θ surfaces. Figure 1(b) shows that a larger stability produces a faster oscillation and inhibits the maximum vertical displacements W/N . The associated downward buoyancy forces have a maximum magnitude proportional to WN . For a limited upward displacement, the force and potential temperature deviation is proportional to $-N^2\delta z$. This corresponds to a potential temperature deficit which is a ‘lifted index’ LI of parcel stability, which has generalized use in cumulus convection.

For smaller values of static stability, the restoring buoyancy forces are weaker and the oscillations are slower. Neutral stability occurs when $\partial\theta_{v0}/\partial z$ is zero (dry adiabatic conditions); a displaced parcel with no initial buoyancy remains that way, so there is no vertical acceleration. ‘Absolute instability’ occurs when it is further reduced to a negative value (superadiabatic lapse rate). In this case, $N^2 = -|N^2|$ is negative, and the solutions to eqn [4] are exponential in time (Figure 1(b)). The growing mode ($\exp(|N|t)$) corresponds to a cooperative relation between buoyancy and motion (e.g., warm air rising) and may be thought of as the initial stage of convection. (A decaying mode ($\exp(-|N|t)$) corresponds to a mismatch of B and w (e.g., cold air rising) and so it is of no long-term consequence.) For convective motions, the increase of vertical kinetic energy is equal to the buoyancy work $\int B\delta z$, known as the convective available potential energy (CAPE) along the parcel’s vertical path. The LI for such a parcel is negative, indicating instability. (In vertically confined convective systems, a growing mode requires that thermal and viscous dissipation must be overcome, so a critical value of $|N^2|$ must be exceeded, as expressed in a critical ‘Rayleigh number’ necessary for convection.) For many applications, the distinction between θ and θ_v is of secondary importance, as is assumed in the remaining discussion.

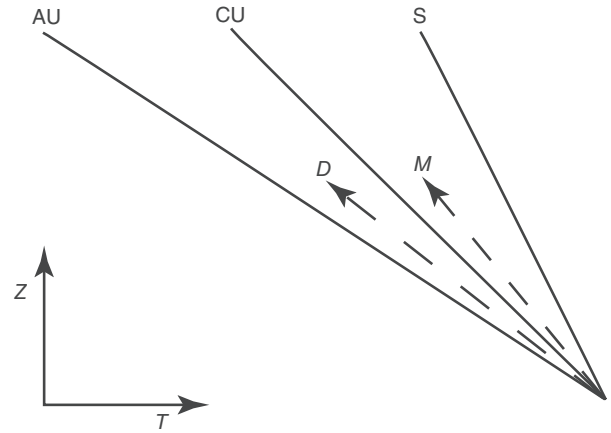


Figure 2 Vertical temperature profiles (solid) for three categories of static stability. Temperature changes for dry and moist adiabatic parcel displacements are dashed. AU: absolutely unstable; CU: conditionally unstable (for saturated parcels); and S: absolutely stable.

Moist Instability

In a humid atmosphere, phase changes in the water content may cause instability even when N^2 is positive. In this case, a parcel conserves its equivalent potential temperature θ_e rather than θ . θ_e exceeds θ by a temperature-dependent amount depending on humidity. As an example, conservation of θ_e is consistent with upward motions leading to saturation, the release of latent heat of condensation, and the diabatic increase of θ . These ‘moist’ diabatic processes reduce the effective static stability for cloud systems. For example, the stability of a cloud layer to internal displacements depends most strongly upon $\partial\theta_e/\partial z$, with negative values corresponding typically to instability. (This criterion is used to describe the ‘potential instability,’ an often-misused concept that describes the stability of an unsaturated layer which is lifted hypothetically until it becomes a cloud layer.)

The most important example of moist processes affecting stability occurs when rising, saturated parcels in cumulus clouds penetrate a dry ‘environmental’ layer. In this case, ‘conditional instability’ may occur even when $\partial\theta_{v0}/\partial z$ is positive and the ‘dry dynamics’ of the environment are stable. This instability criterion may be expressed as $\partial\theta_{es}/\partial z < 0$, where θ_{es} is the saturation equivalent potential temperature, a known function of T and pressure p . This criterion is met if the virtual temperature lapse rate $-\partial T_v/\partial z$ exceeds the smaller moist adiabatic rate Γ_m . The result is that the unstable combination of positive buoyancy with a rising parcel occurs if a saturated parcel moves upward through a layer of air where N^2 is insufficiently positive. Figure 2 illustrates the three fundamental types of stability for an atmosphere.

The growth of cumulus clouds is overestimated by this simple parcel reasoning because updrafts require compensating subsidence of the environment. The resulting adiabatic warming decreases the relative buoyancy of the cloud. A simple ‘slice’ theory shows that the effective stability of the system is then increased for finite-sized clouds; it can be represented as a combination of the moist and dry static stabilities. Additional stabilizing influences are turbulent mixing of momentum and thermodynamic quantities between the cloud and the environment, and the effects of pressure adjustments.

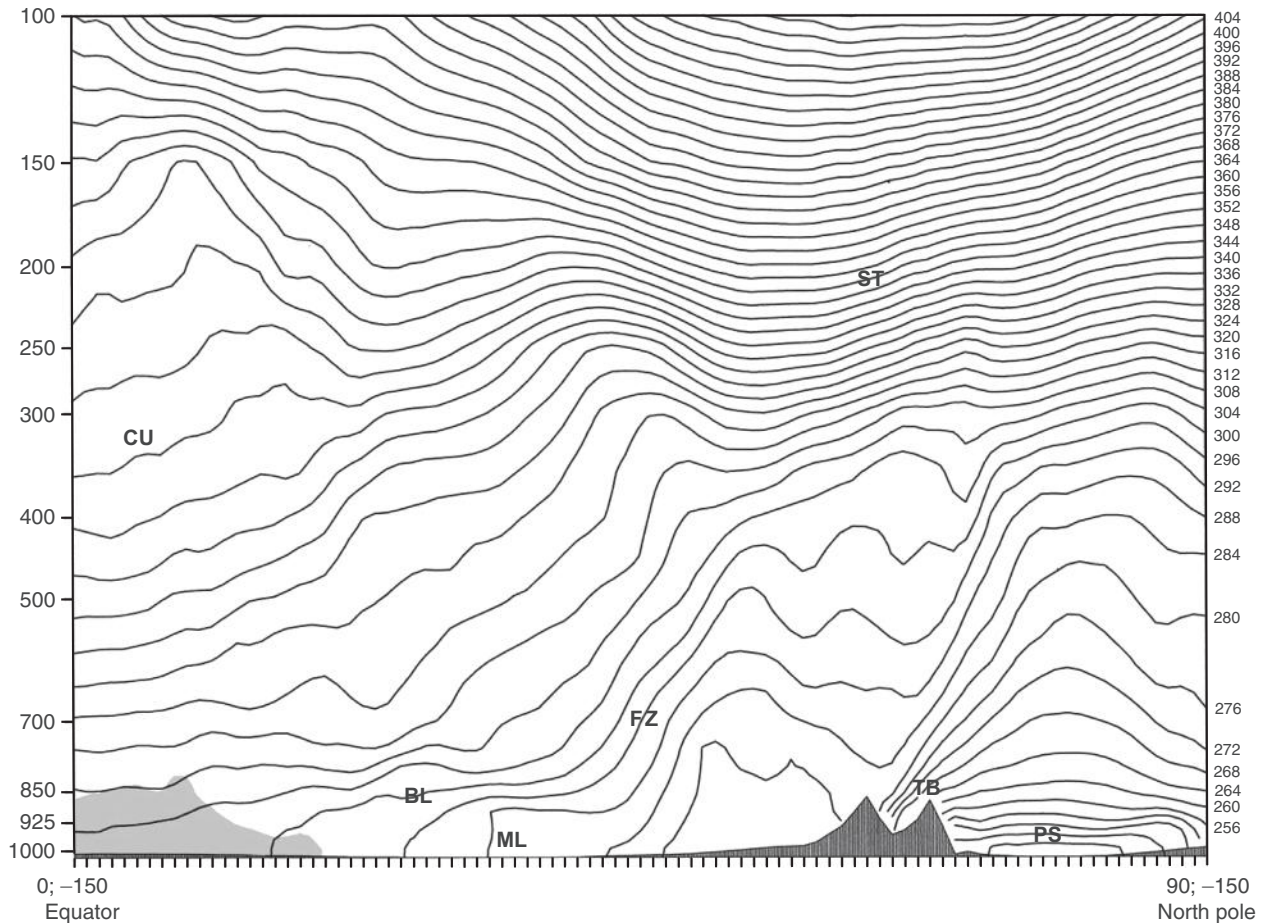


Figure 3 Vertical cross section of θ from Equator to pole. Static stability is indicated by vertical closeness of θ surfaces. Left scale is pressure in hectopascal. Dark shading: Earth's topography. Light shading: boundary layer air with moisture mixing ratio exceeding 12 g kg^{-1} . Strong stability cases: ST – stratosphere, PS – polar surface, BL – boundary layer top, FZ – frontal zone, and TB – topographic blocking by mountains. Weak stability: CU – conditionally unstable tropical troposphere and ML – convectively mixed boundary layer. GFS analysis, National Centers for Environmental Prediction (NCEP: www.ncep.noaa.gov). Provided by Unidata (www.unidata.ucar.edu).

Climatology of Static Stability

In the simplest terms, the dry and moist static stability indices depend upon vertical profiles of potential temperature and to a lesser extent on the profile of water vapor. Figure 3 shows some typical features in a vertical cross section. Strong static stability (N^2) regions are associated with isentropic surfaces that are closely spaced in the vertical, a symptom of the vertical 'stiffness.' Weak stability regions have greater spacing, and the limit of zero stability may correspond to a vertical orientation of the isentropic surface. Regions of moist unstable motions are possible where there is a conditionally unstable temperature profile and sufficient moisture supply (e.g., the tropical boundary layer).

The distribution of static stability $\partial\theta/\partial z$ can be explained first by considering the processes that change the spacing Δz of potential temperature surfaces. From the first law of thermodynamics, it is easily shown that local changes of stability are caused by (1) advection of stability from upwind, (2) (vertically) differential temperature advection, and (3) differential diabatic heating. The differential diabatic term (3) explains many basic stability features in the atmosphere. The term is proportional to dJ/dz , where J is the diabatic heating rate per

unit mass; negative J connotes cooling. This term increases the stability where J increases with height, and decreases it where J decreases with height.

Examples of diabatic influence on static stability are seen in Figure 3. The strongest stability is seen in the stratosphere, where stability is maintained by the radiative heating increase due to absorption of solar ultraviolet radiation by ozone. The tropospheric static stability is several times smaller, due especially to downward longwave radiation. Near the Earth's surface, strong stability at high latitudes is created by longwave radiative cooling, whereas weaker stability at other latitudes is driven by sensible heat from the surface. The sensible heating is concentrated in the atmospheric boundary layer, which often resembles a 'convective mixed layer' of low stability, especially over land. In the tropics, the troposphere is moist at low levels, conditionally unstable, and deep; heavy cumulus convection is prevalent and its latent heating is the essential driving of the tropical climate system.

Some of the smaller scale features in the figure are affected by adiabatic circulation processes. Term 2 includes the effect of vertical wind shear in a baroclinic region; near the Earth's surface, warm (cold) advection situations are associated commonly with stabilization (destabilization) of the lower

atmosphere by this process. This term also explains the development of strong static stability by subsidence at the top of the atmospheric boundary layer and in frontal zones.

There are seasonal and diurnal variations in stability that cannot be represented in the snapshot (Figure 3). These variations are caused by those of solar radiative forcing of the Earth's surface, which results in variations of sensible and latent heating. Broadly speaking, the static stability fields tend to shift poleward in the summer season and Equatorward in the winter season. The destabilization of the lower atmosphere is a maximum over land on summer days, while it is a maximum over the midlatitude oceans in winter.

Static Stability and Circulation Dynamics

Static stability influences the motions of the atmosphere on a range of scales and may permit waves to connect distant regions. Simple vertical buoyancy concepts are not sufficient for understanding these effects. In reality, one must also consider the coupling to horizontal winds and the ways in which pressure links the motion of different air parcels. The spatial distributions of static stability and wind determine the outcomes, which range from flow instability to various kinds of wave propagation in the horizontal and vertical.

Small-Scale Turbulence

Turbulence in the atmosphere may be caused by convection or by wind shear, and static stability is influential in each case. Ignoring moist dynamics, convection requires $\partial\theta_v/\partial z$ to be negative, which occurs most commonly when the air is in contact with a warmer Earth's surface, such as a sunny day over dry land. In such cases, N^2 is strongly negative in the surface layer (roughly the lowest 50 m), reflecting a superadiabatic lapse rate of virtual temperature. Static stability is then near neutral ($N^2 = 0$) in a deeper 'mixed layer' up to the boundary layer top. Thus, neutral boundary layers are symptoms of surface-induced convection.

Positive static stability inhibits turbulence induced by wind shear. The production of shear turbulence may be understood by imagining a layer of concentrated wind shear which, when perturbed by vertical displacements, creates a pressure feedback that amplifies the displacements of the layer. The result is mixing of fast and slow air parcels by a growing pattern of Kelvin–Helmholtz instability (KHI) motions. Obviously, the vertical restoring forces of a statically stable atmosphere will oppose the vertical components of such KHI displacements. The competition between shear instability and stable stratification is best measured by the Richardson number:

$$Ri = N^2 / SH^2 \quad [5]$$

where SH is most generally the magnitude of the vector wind shear $\partial\mathbf{V}/\partial z$. Ri is the squared ratio of the stable buoyancy oscillation frequency N to the maximum shear-induced growth rate SH . Theory and observation show that when $Ri > 1/4$, shear growth is eliminated: static stability wins, and perturbations are stable oscillations as in Figure 1. On the other hand, when static stability is reduced so that $Ri < 1/4$, the shear instability is not suppressed totally, and the perturbations may grow into turbulence.

In the free atmosphere, intense frontal zones are associated commonly with 'clear air turbulence,' despite the zones having a maximum static stability. This is because they are sloping regions of strong gradients, and Ri is reduced more effectively by the strong shear as the vertical width of the zone becomes small. The mixing by this turbulence is thought to modify the mesoscale structure of the static stability and shear near jets.

Very near the Earth's surface, strong shear is created by frictional drag, but the turbulence is limited by the surface and static stability. In such surface boundary layers, the intensity of shear turbulence is the greatest beneath the height L , the Monin–Obukhov length. L varies inversely with the stable air–surface temperature difference and static stability near the ground. Higher in the boundary layer, the turbulent fluxes are often represented by eddy mixing coefficients which are a decreasing function of Ri (and hence static stability).

Mesoscale Motions

Static stability and its spatial variations may produce complex mesoscale motions. Since wind speeds and the frequencies of weather systems are strongly subsonic, it follows that the pressure fields are in a state of 'anelastic' balance with the temperature and velocity patterns. The simplest balance involving buoyancy B is described by the three-dimensional p.d.e. $\nabla^2 p = \partial B / \partial z$, where ∇^2 is the elliptic Laplacian operator in three spatial dimensions. The buoyancy gradient term 'forces' a smooth pressure response which decreases inversely with distance. For a vertically oriented pattern of B , the pressure response is negligible, and simple buoyancy forces dominate the motion. However, a pattern of B tilted toward the horizontal produces a pressure gradient force that opposes B . Thus, static stability may be associated with motions that may or may not be in hydrostatic balance, depending on the distribution of buoyancy in the vertical plane.

The simplest tool for understanding these motions is the theory of buoyancy waves (see Atmospheric Waves). For patterns of motion and temperature with phase fronts tilted at an angle α from the vertical, the free oscillation has a frequency $\omega = N \cos \alpha$. We see that N is actually an upper limit on the frequency, corresponding to the vertical orientation for a simple buoyancy oscillation. Such motions are nonhydrostatic. Much slower oscillations occur when the wave patterns are tilted toward the horizontal, a result of the 'braking' effect of the pressure field on the buoyant parcel. These motions are nearly hydrostatic, and the waves may propagate with a nondispersive phase speed obeying

$$c_g^2 = N^2 / m^2 \quad [6]$$

where m is the vertical wave number. Strong static stability corresponds to fast horizontal wave speeds.

There are dramatic consequences of the simple frequency dispersion relation. For example, the energy of the waves is transmitted along the sloping wave front at a group speed:

$$c_g = N \sin \alpha / K \quad [7]$$

where K is the two-dimensional wave number (inverse scale) of the wave pattern. We see that the energy propagation rate

increases with static stability and with angle α from the vertical. It follows that the response to a confined impulse will rapidly spread low-frequency energy horizontally, while higher frequencies will be found immediately above and below the region. Imposed frequencies greater than N are 'evanescent': such energy cannot be propagated away from the forcing. Interestingly, the orthogonal relation between phase and group velocity vectors implies that downward phase propagation is associated with upward energy propagation.

These properties have implications for a variety of meso-scale responses of a stable atmosphere to surface heating or mountains. For example, steady airflow U over a mountain complex may be envisioned in terms of periodic forcing. The above theory for low frequencies predicts that (1) a wide mountain may cause upwind 'blocking' of low-level air with high static stability and (2) motions over the mountain are nearly in hydrostatic balance. The theory for higher frequencies suggests that very narrow mountains do not disturb the flow far above the mountain, but an intermediate mountain width yields a complex pattern of vertically propagating wave patterns extending upward and downwind of the mountain. In order for energy to propagate upward, the wave fronts must tilt upwind with increasing altitude and the waves transport wind momentum down into the mountain. An example is shown in Figure 4.

Static stability and wind variations influence the vertical fluxes of mesoscale wave energy and momentum which may

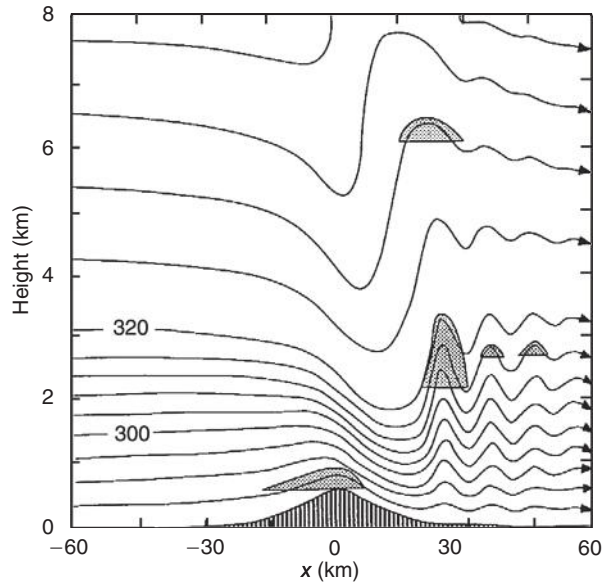


Figure 4 Streamlines and θ surfaces for flow over an isolated ridge. Upwind conditions have high static stability below 3 km, so $P(z)$ decreases upward. Wind speeds vary along streamlines in proportion to closeness of streamlines. Proceeding from the left, note the slowing of air on the upwind side, strong downslope wind, vertically tilted flow pattern, downwind jump, and lee waves trapped in the stable layer. Shading denotes possible clouds due to lifting of moist layers. Reproduced with permission from Houze, R.A. Jr., 1993. *Cloud Dynamics*. Academic Press, San Diego (Figure 12.9). Courtesy of Durran, D.R., 1986. Another look at downslope windstorms, Part 1. *Journal of the Atmospheric Sciences* 43, 2527–2543.

link the upper atmosphere with the surface. For example, the vertical structure of the steady response with horizontal wave number k is governed by a propagation coefficient:

$$P(z) = [N^2/U^2 - k^2] \quad [8]$$

The wave profile 'propagates' vertically only when P is positive, or when static stability makes the Scorer parameter N^2/U^2 sufficiently large. The vertical wave number is then $P^{1/2}$. Variations in stability or wind will cause $P(z)$ to vary, which corresponds to wave refraction in the vertical plane. Two categories of phenomena result, depending upon whether $P(z)$ decreases or increases with height.

If stability decreases with height, then $P(z)$ may become negative, and the wave may be reflected downward. Since the rigid Earth is also a reflecting surface for the wave vertical motion, the mountain-induced wave energy may become trapped in this layer. In this case, intense downslope winds and resonant 'lee' waves are possible. Other wave mechanisms, such as wave absorption at a critical layer where $U = 0$, depend more strongly on the wind profile.

In the other extreme, weak static stability in the boundary layer causes $P(z)$ to increase with height above the surface. A common idealization is a mixed layer ($N^2 = 0$) capped at height H by a sharp inversion of strength $\Delta\theta_v$. In this case, horizontal scales larger than H are hydrostatic and move with speeds of 'shallow water' gravity waves obeying

$$c_G^2 = [g'H] \quad [9]$$

We see that $g' = g(\Delta\theta_v/\theta_v)$, the 'reduced gravity' parameter for the inversion, plays an analogous role to static stability for these hydrostatic motions. An example of this kind of motion is the propagation of a gust front, the leading edge of thunderstorm outflow in the boundary layer. Another example is where this kind of air layer is forced to flow over a mountain at speed U ; the inversion stability appears inversely in the Froude number $F = U^2/(g'H)$. This number represents a competition between the flow inertia and the inversion stability or equivalently between advection by U and gravity wave propagation c_G . Values exceeding $O(1)$ may be associated with blocking on the upwind side of mountains and strong downslope winds and hydraulic jumps on the downwind side. A similar definition of F is useful for a stably stratified atmosphere encountering a mountain of height H^* : $F = [U/NH^*]^2$.

Large-Scale Circulations

Large-scale circulations are those of large horizontal dimension, associated with low frequencies and hydrostatic balance. For such motions, static stability and the rotation of the Earth are important. Coriolis effects limit horizontal parcel motions in a fashion somewhat analogous to the buoyancy oscillation. The natural frequency of this 'inertia oscillation' is simply the Coriolis parameter f , which is about 100 times smaller than N . Thus, large-scale dynamics is ruled by the two fundamental frequencies of geophysical fluid dynamics: N and f . The most important large-scale flow variable is the combination known as the potential vorticity:

$$q = (f + \epsilon)N^2 \quad [10]$$

which is proportional to both the absolute vorticity of the winds and the static stability. Two frequency classes of large-scale waves are possible. The higher frequency class is inertio-gravity waves that obey

$$\omega^2 = f^2 + N^2(k^2/m^2) \quad [11]$$

Static stability is seen to increase the minimum frequency f . These motions are never in a state of geostrophic balance, so they play an important role in the transient adjustments to thermal and mechanical forcing of the atmosphere. Vertical propagation of wave energy occurs only when frequency ω exceeds f . For example, diurnal atmospheric tides propagate vertically only Equatorward of 30° latitude.

Horizontal energy propagation is highly dispersive as a result of the Coriolis term: the largest scales propagate energy very slowly, while the smallest scales do so at the fast gravity wave speed c_G . The separation between large and small horizontal scales occurs at

$$\lambda = c_G/f = (N/f)m \quad [12]$$

known as the Rossby deformation radius.

The deformation radius is the natural horizontal scale for large-scale atmospheric dynamics. From eqn [12], it is the distance traveled by a gravity wave in the time (f^{-1}) required for Coriolis forces to deflect the velocity. It represents the spatial scale for adjustment of wind and pressure to geostrophic balance. This scale of adjustment increases with the static stability parameter N , and it decreases with rotation f .

The lowest frequency class of large-scale dynamics is that of quasi-geostrophic (QG) dynamics for which ' $\omega \ll f$ ' (see **Dynamical Meteorology: Quasigeostrophic Theory**). These motions are always near a state of geostrophic and hydrostatic balance, and are influenced strongly by static stability and the Earth's rotation. The QG form of the potential vorticity corresponding to eqn [10] has a variable part proportional to

$$q^* = \left\{ \bar{N}^2 \nabla^2 p' + f^2 \partial^2 p' / \partial z^2 \right\} \quad [13]$$

The response of p' to thermal or vorticity forcing is determined by eqn [13], which is a three-dimensional Laplacian in coordinates that are stretched vertically according to N/f . It follows that point forcing yields an elliptically shaped response, with the major axis lying in the direction of least resistance. For example, large static stability of the stratosphere yields responses that are stretched horizontally and compressed vertically. For a given vertical scale, this property implies a horizontal influence distance equal to the deformation radius. For a given horizontal scale L , it implies a vertical influence distance called the Rossby depth, given by $H_R = (f/N)L$, so that increasing the stability decreases the vertical coupling distance H_R .

Similar considerations may be applied to the QG 'omega equation' to distinguish the total response to various patterns of thermal and vorticity forcing, illustrating the crucial importance of static stability on large-scale dynamics through the ratio N/f . There are obvious global implications, since f is small at low latitudes. Two major regimes of large-scale atmospheric circulation are the result. For example, QG instability theory indicates that baroclinic wave and cyclone growth are possible only at mid-high latitudes. Hence the tropics are less variable, except in concentrated areas of moist

convection (such as tropical cyclones) where conditionally unstable air lowers the effective ratio N/f . Similar arguments account for the difference among the atmospheric circulations of other planets.

Static Stability Variability: Phenomena and Prediction

The previous static stability concepts have application to many variable phenomena in the atmosphere, particularly near the earth's surface and in combination with other variables such as water vapor.

Temporal fluctuations of static stability occur on time scales from hours to years. For example, the atmospheric boundary layer over land responds strongly to the diurnal solar cycle as it experiences surface heating during the day and cooling by longwave radiation at night. The static stability consequently is a minimum in the lower daytime atmosphere and is a strong maximum in a thin (e.g., 200 m) nighttime boundary layer. As a consequence, low-level sources of natural gases or man-made pollutants are trapped near the ground during the night, a result of suppressed turbulence and weaker winds. These products are then vigorously dispersed to higher altitudes during the afternoon, typically by convection into a deeper boundary layer (e.g., 1 km) or even the free troposphere. Spatial differences in diurnal boundary layer heating/cooling are found near coastlines and sloping terrain: (1) The daytime 'sea breeze' along coastlines is a flow of relatively cool, stable air inland from over the water. (2) The nighttime 'katabatic wind' is a downhill flow of cool, stable air which can accumulate in stagnant pools in valleys. (3) Over flat land, strong convective rains storms often produce 'downbursts' of cool air which spread horizontally as a stable 'pool,' following a leading edge 'gust front' which can produce damaging winds.

The annual cycle of temperature and stability is strongest in middle and high latitudes, where the solar energy flux has its most extreme seasonal changes, and where 'synoptic' weather systems create variability of temperature and stability over periods of days. These traveling systems result from the instability of the intense baroclinic zones separating polar air from subtropical air. These typically consist of a time progression of warm, weakly stable air masses, low-pressure cyclonic winds, cold, stable air masses, and high-pressure anticyclonic winds which are characteristic of the Norwegian synoptic weather model. (Figure 5 illustrates the spatial differences of static stability for a springtime weather pattern over the US when the synoptic systems organized large areas of convective storms.) The surges of polar air in wintertime weather systems amount to a downward 'slumping' of cold air domes as they propagate toward the equator, increasing the low-level static stability following the passage of cold fronts. The slumping process is a manifestation of the conversion of potential energy to kinetic energy of winds in large-scale weather systems and an increase in overall tropospheric static stability.

Synoptic wind patterns typically create patterns of moist convective stability due to the destabilizing advection processes discussed earlier, as well as the transport of water vapor from source regions. Since all of these processes are represented in numerical weather and climate models, a byproduct of the

systems and climate of the Earth. In the Earth's atmosphere, radiation and surface energy fluxes act to create three main categories of static stability.

1. *Strong stability*: The stratosphere is the most extensive example. Strong stability there encourages the vertical propagation of forced planetary waves through westerly winds regions, but it suppresses the growth of synoptic-scale circulations and convection.
2. *Weak static stability*: The troposphere is the atmosphere's dominant region of lesser, more variable static stability. As a result, instabilities may produce weather systems on a range of scales. For example, baroclinic wave circulations create variable weather in middle and high latitudes, and conditional instability may be realized as moist convection. Moist convection may be organized on the global scale (e.g., Hadley and Walker circulations), the synoptic scale (e.g., tropical cyclones), or the mesoscale (deep cumulus convection and severe weather). The static stability for dry processes may be strong enough to allow mesoscale mountain influences on the upper atmospheric wind or to suppress small-scale shear instability which would otherwise produce clear air turbulence.
3. *Static instability*: The energy balance of the Earth system requires that the Earth's surface provides energy to the atmospheric boundary layer. This is often associated with static instability, dry convective motions, and sensible heating. Neutrally stable conditions are also very common, in which case turbulence transports latent energy away from the surface, enhancing the possibility of subsequent conditional instability.

In summary, the three regimes of static stability account for much of the variety of weather and climate. Ultimately, the various kinds of circulations feedback on the static stability field itself, leading to increased complexity of its space-time variability.

Finally, looking ahead to the time scales of years, the prospect of a warmed climate is consistent with a changed

energy budget that requires the surface to release more energy upward into the boundary layer. The energy increases will be in the form of both sensible (thermal) energy and latent energy of water vapor. These changes will inevitably tend to reduce the static stability, either in the frequency of stable conditions, or in its intensity. These changes are expected to affect many weather systems, including precipitation events and storms.

See also: Dynamical Meteorology: Vorticity. Thermodynamics: Moist (Unsaturated) Air; Saturated Adiabatic Processes.

Further Reading

- Andrews, D.G., Holton, J.R., Leovy, C.B., 1987. *Middle Atmosphere Dynamics*. Academic Press, Orlando, FL.
- Chapman, S., Lindzen, R.S., 1970. *Atmospheric Tides. Thermal and Gravitational*. Reidel, Dordrecht.
- Durrán, D.R., 1986. Another look at downslope windstorms, Part 1. *Journal of the Atmospheric Sciences* 43, 2527–2543.
- Durrán, D.R., 1990. Mountain waves and downslope winds. In: Blumen, W. (Ed.), *Atmospheric Processes over Complex Terrain*. American Meteorological Society, Boston, pp. 59–82.
- Emanuel, K.A., 1994. *Atmospheric Convection*. Oxford University Press, New York.
- Gill, A.E., 1982. *Atmosphere–Ocean Dynamics*. Academic Press, New York.
- Holton, J.R., 1992. *An Introduction to Dynamic Meteorology*, third ed. Academic Press, New York.
- Houze Jr., R.A., 1993. *Cloud Dynamics*. Academic Press, San Diego.
- Irbane, J.V., Godson, W.L., 1981. *Atmospheric Thermodynamics*, second ed. Reidel, Dordrecht.
- Pedlosky, J., 1987. *Geophysical Fluid Dynamics*, second ed. Springer Verlag, New York.
- Scorer, R.S., 1978. *Environmental Aerodynamics*. Ellis Horwood, Chichester, UK.
- Sorbian, Z., 1989. *Structure of the Atmospheric Boundary Layer*. Prentice-Hall, Englewood Cliffs, NJ.
- Stull, R.B., 1988. *An Introduction to Boundary Layer Meteorology*. Kluwer, Boston.
- Tritton, D.J., 1996. *Physical Fluid Dynamics*, second ed. Oxford University Press, New York.
- Turner, J.S., 1973. *Buoyancy Effects in Fluids*. Cambridge University Press, London.
- Yih, C.S., 1965. *Dynamics of Non-homogeneous Fluids*. Macmillan, London.

Stationary Waves (Orographic and Thermally Forced)

S Nigam, University of Maryland, College Park, MD, USA

E DeWeaver, University of Wisconsin, Madison, WI, USA

© 2003 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 4, pp 2121–2137, © 2003, Elsevier Ltd.

Introduction

The term stationary waves refers to the zonally asymmetric features of the time-averaged atmospheric circulation. They are also referred to as standing eddies, where standing refers to the time averaging over a month to season, and eddy is a generic term for zonally asymmetric patterns. The zonal asymmetries of the seasonal circulation are particularly interesting because they occur despite the longitudinally uniform incidence of solar radiation on our planet. Stationary waves must arise, ultimately, due to asymmetries at the Earth's surface – mountains, continent–ocean contrasts, and sea surface temperature asymmetries. Understanding precisely how the stationary waves are generated and maintained is a fundamental problem in climate dynamics.

Stationary waves have a strong effect on the climate through their persistent northerly and southerly surface winds, which blow cold and warm air. Advection of moisture by the stationary wave flow contributes to hydroclimate variations over the continents. Beyond their direct advective impact, stationary waves control the location of stormtracks – the preferred paths of synoptic weather systems in the midlatitudes, and the zone of tropical–extratropical interaction in the subtropics. Stationary waves are important also on longer time scales, since interannual climate variability projects substantially on the zonally asymmetric component of the flow. Finally, stationary waves contribute significantly to the maintenance of the complementary zonally symmetric circulation, in both climatological and anomalous states; the contribution is through quadratic fluxes of meridional momentum and heat. Stationary waves are thus a fundamental feature of the general circulation of the troposphere.

Observed Structure

Stationary waves are stronger in the Northern Hemisphere because of greater orography and continentality. Wave amplitudes in the Northern Hemisphere are largest during winter, modest during the transition seasons of spring and autumn, and weakest during summer. The Southern Hemisphere stationary waves and their seasonal variation are substantially smaller in comparison.

Northern Hemisphere Winter Structure

Because of the geostrophic balance condition, stationary waves in the upper-level flow can be conveniently displayed using the height of the 300 hPa pressure surface. The geostrophic wind blows along the height contours, with lower heights to the left in the Northern Hemisphere, and with a speed proportional to

the gradient of the height field. The height of the 300 hPa surface varies considerably with latitude and longitude (**Figure 1(a)**), with the mean height being close to 9 km. The polar vortex is clearly recognizable in this projection. The vortex is due to insolation and planetary rotation, both zonally symmetric inputs, but the vortex has notable departures from symmetry: troughs over northern Canada and western Siberia, and ridges over the eastern Atlantic and Pacific. (The zonally asymmetric component of the field which highlights the troughs and ridges is shown later, in **Figure 2(a)**). The regions where the height contours are close together correspond to strong westerly (coming from the west) jets: the Asian–Pacific and North American jets.

Stationary waves at the Earth's surface can be identified using the sea-level pressure field, which in elevated areas is the surface pressure reduced to sea level. The lightly shaded regions in **Figure 1(b)** are surface lows, and the dark regions are highs. Lows are found over both ocean basins, the Aleutian Low in the Pacific and the Icelandic Low in the Atlantic. The Aleutian Low is centered off the tip of the Aleutian Islands chain, and the counterclockwise flow around the low brings southerly marine air to coastal Canada and Alaska, lessening the severity of the winter season. To the south of the Icelandic Low is a high-pressure center known as the Azores High. Strong onshore surface flow occurs between the Icelandic Low and the Azores High, again lessening the severity of coastal winters in Europe. Much higher surface pressure can be found over central Asia in a center called the Siberian High. Between the Siberian High and the Aleutian Low is a region of strong northerly flow, which brings down colder air and lowers near-surface temperature along the east coast of Asia. The winter sea-level pressure field can be broadly characterized as being high over the continents and low over the comparatively warmer northern oceans. Since sea-level pressure is related to column temperature, vertical coherence of the continent–ocean temperature contrast in the lower troposphere is key, as discussed later.

The stationary wave pattern changes considerably between the surface and 300 hPa, and these changes are highlighted in **Figure 2**. The top panels show eddy heights at the 300 and 850 hPa levels, revealing the troughs and ridges. These features are displaced westward with increasing height, i.e., westward tilted, assuming that the same features are being tracked at the two levels. The low-level trough over the Pacific is positioned 15–20° westward of the Aleutian Low, and gives way to a trough centered on the east Asian coast at 300 hPa, which is associated with the Asian–Pacific jet. The 850 hPa trough over the North Atlantic is likewise shifted relative to the Icelandic Low, and migrates further westward towards Hudson Bay at upper levels; it brings cold Arctic air into the central and eastern United States and Canada. The low-level feature over Eurasia (**Figure 2(b)**) is more definitely linked to the surface Siberian

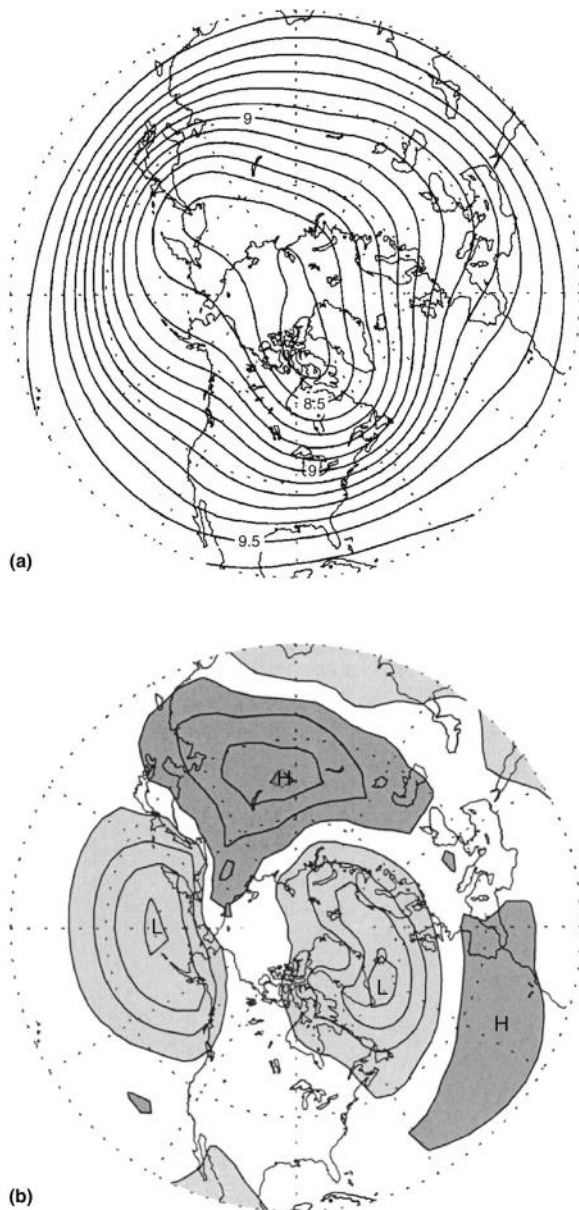


Figure 1 (a) Average height of the 300 hPa pressure surface in northern winter months (December, January, February, and March: DJFM). The average is over 20 winter seasons (December 1979 through March 1999), and is computed from the reanalysis fields produced by the US National Center for Environmental Prediction (NCEP). The contour interval is 100 m. (b) Average sea-level pressure (SLP) for the same months and years, with a contour interval of 5.0 hPa. Sea-level pressure data come from Trenberth's analysis, which is archived at NCAR. Dark (light) shading represents values above 1015 hPa (below 1010 hPa). The letters 'L' and 'H' designate the prominent centers of action: the Aleutian Low, Siberian High, Icelandic Low, and the Azores High. Map domain begins at 20° N.

High, but there is no corresponding feature of significance present at the upper level – in contrast with the vertically coherent structure of the Azores High.

The vertical structure of stationary waves is plotted in Figure 2(c) and (d), which are cross-sections of the eddy

height field at 40° N and 60° N. The shading in these panels depicts the eddy temperature field. (In hydrostatic balance, this is the vertical derivative of the height field in $\log(p)$ coordinates.) These plots allow for the tracking of features. The northern section shows the pronounced westward tilt of the east Asian trough, the Rocky Mountain ridge, and the Azores High. The tilt is a consequence of meridional temperature advection by the associated geostrophic wind, which induces cooling to the west (east) of the low (high). Interestingly, connection with the prominent surface features is not strong, except in case of the highs. The Aleutian and Icelandic Lows, in particular, are quite shallow ($p \geq 800$ hPa), exhibiting little connectivity to the westward displaced upper-level troughs. The southern section (40° N) nicely reveals the limited vertical extent of the Siberian High, in contrast with the deep structure of the Azores High.

Comparison of the two cross-sections indicates a striking difference in vertical variation of the eddy heights, particularly in the Eastern Hemisphere. In the northern section, the wave amplitude keeps growing with height up until the tropopause, and even beyond. The structure is indicative of upward propagation of stationary wave energy into the polar lower stratosphere. (Note that the wave's phase is stationary, so the phase velocity is zero, but its group velocity – the velocity of energy propagation – is not zero.) In contrast, the 40° N structure is indicative of trapping of wave energy within the troposphere. The eddy height at the 10 hPa level, displayed using shaded contours in Figure 3, reveals the presence of a large-amplitude stationary wave at an altitude of nearly 30 km. The zonal wavelength of this pattern is evidently close to the circumference of the latitude circle, i.e., the largest possible. Both observations and theory (see **Dynamical Meteorology: Rossby Waves**) suggest that disturbances of such large wavelengths can propagate into the stratosphere. The wave pattern in Figure 3 moves the center of the polar vortex away from the geographical pole and reduces the strength of the vortex.

Equatorial westerly duct

An important circulation feature in the deep tropics during northern winter is the presence of strong upper-level westerlies ($\sim 10 \text{ m s}^{-1}$) over the Pacific and Atlantic longitudes. This is notable because the equatorial belt is otherwise occupied by easterly winds. Zonal winds at 200 hPa are shown in Figure 4(a), with the easterly region shaded. A vertical section at the equator (Figure 4(b)) shows westerly zones to be confined to the near-tropopause region (100–300 hPa), with maximum values ($\sim 15 \text{ m s}^{-1}$) at 200 hPa. The origin of equatorial westerly zones is not well understood, but their absence in northern summers and El Niño winters suggests that their occurrence is linked to the absence of strong convection in the central equatorial Pacific and Atlantic longitudes.

Rossby wave propagation theory (see **Dynamical Meteorology: Rossby Waves**) suggests that tropical easterlies are an effective dynamical barrier, shielding the equatorial zone from the influence of midlatitude perturbations. Openings in this barrier, or westerly ducts, thus provide a conduit for equatorward penetration of midlatitude waves during northern winter – a timely opening, since the midlatitude stationary and transient wave activity is most vigorous in winter. Interaction between midlatitudes and the equatorial zone can impact convection and

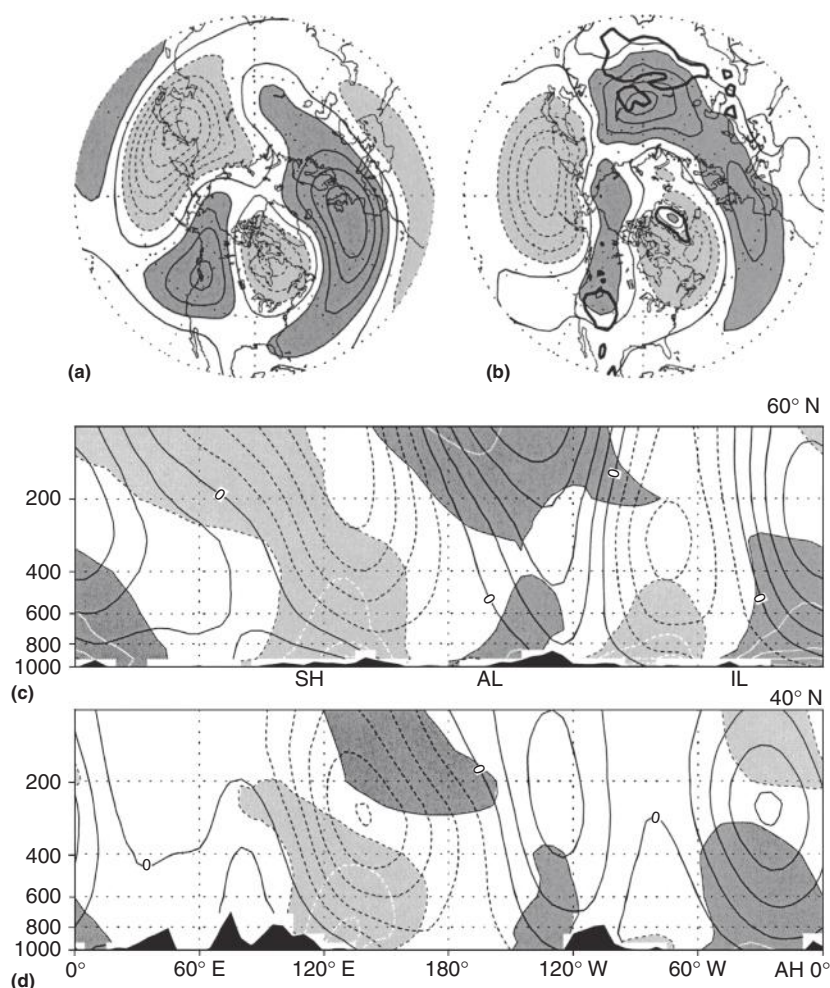


Figure 2 (a) Eddy height at 300 hPa during northern winter (DJFM), or height of the 300 hPa pressure surface after subtracting the zonal average. The contour interval is 50 m, with dark (light) shading for positive (negative) values in excess of 50 m. (b) Eddy height at 850 hPa for the same period, with a contour interval of 25 m and dark (light) shading for positive (negative) values in excess of 25 m. The thick contours in b enclose regions where the surface pressure is less than 850 hPa. In these regions, the pressure surface is interpolated below ground. (c and d) 1000–100 hPa zonal–vertical cross-sections of eddy height and temperature at (c) 60° N and (d) 40° N. Contour interval for eddy height is 50 m, and dashed contours represent negative values. Eddy temperature is plotted in 3 K contours with dark (light) shading for positive (negative) values in excess of 3 K, and zero contours suppressed. ‘SH’, ‘AL’, ‘IL’, and ‘AH’ are the surface lows and highs of Figure 1(b).

water vapor distribution in the tropics and subtropics. Lateral mixing from extratropical intrusions can also influence tracer transports.

Northern Hemisphere Summer Structure

The northern polar vortex is much weaker in summer than in winter. The summer vortex is shown at the 150 hPa level in Figure 5(a), and is evidently quite symmetric. It also lacks the tight meridional gradients that characterized the winter vortex. A somewhat higher level was chosen for displaying the summer pattern in order to capture fully the divergent monsoonal flow and accompanying rotational circulations over the warmer landmasses. The upper-level asymmetries include the very prominent anticyclone over Tibet, and troughs over the subtropical ocean basins which are easier to appreciate in the eddy height plots, shown later.

The summertime sea-level pressure (Figure 5(b)) has almost a reversed winter structure. Two subtropical anticyclones of comparable strength are present in the ocean basins, underneath the upper-level troughs. They are referred to as the Pacific High and the Bermuda High. The Bermuda High is the summer equivalent of the Azores High, which expands while the Icelandic Low retreats northward during the transition from winter to summer. The Pacific sector undergoes a similar winter to summer transition. The subtropical anticyclones constitute the descending branch of the regional Hadley cells which are driven by deep convection in the tropics. Descending motions induced to the north-west of subtropical monsoonal heating may also contribute to anticyclone development.

Over the continents, sea-level pressure is low during summer. A large region of low sea-level pressure is present over Asia beneath the Tibetan anticyclone (which is actually

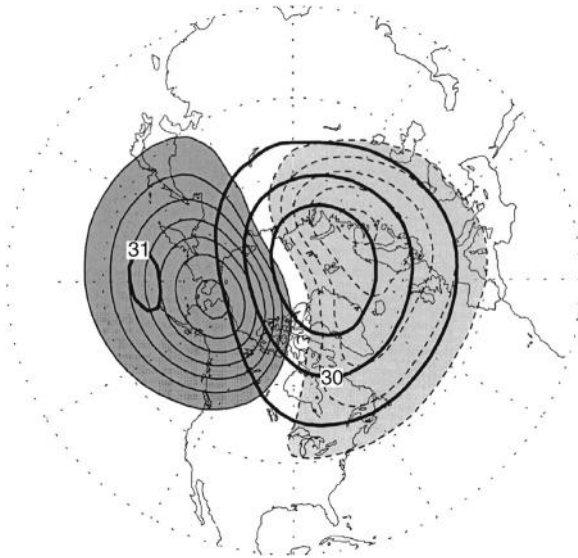


Figure 3 Average height of the 10 hPa pressure surface in northern winter (DJFM). Thick contours show the height field in 500 m increments. The eddy height is plotted at 100 m intervals, with dark (light) shading for positive (negative) values in excess of 100 m. The zero contour for eddy height is suppressed.

centered over northern India). The continental-scale anticyclone is an integral element of the Asian monsoon circulation, being the rotational response to deep heating.

The cross-section of eddy height at 30° N (Figure 5(c)) shows the internal baroclinic structure that is typically produced

by deep heating in the tropics. The Tibetan anticyclone reaches maximum amplitude at 150 hPa, the level displayed in Figure 5(a). Over the oceans, the structure is also baroclinic, but the Pacific and Bermuda Highs are evidently shallow features – although not as shallow as their winter counterparts in Figure 2(c). The strong positive temperature centered over the Tibetan plateau is caused by latent heat release in the monsoon rains. On the other hand, negative temperatures over the Pacific and Bermuda Highs are produced, in part, from the long-wave radiative cooling to space.

The eddy height fields during summer are displayed in Figure 6. The Tibetan anticyclone is the prominent feature at upper levels. Baroclinic structure is evident in the Northern Hemisphere, with upper-level troughs positioned over the subtropical highs. Also evident at the upper level is a weak ridge over North America that is associated with the local monsoon system, which includes the Mexican monsoon. The western edge of the Bermuda High produces low-level southerly flow, which brings in significant amounts of moisture from the Gulf of Mexico into the US Great Plains. A notable low-level feature in the Southern Hemisphere is the Mascarene High centered south of Madagascar, which generates strong easterlies along its northern flank (recall that the flow around a Southern Hemisphere High is counterclockwise). After turning northward along the African coast and crossing the Equator, this flow evolves into the south-westerly monsoon flow over the Arabian Sea.

In the Asian monsoon circulation, equatorial and cross-equatorial flows play an important role, and these cannot be appreciated in the height field, since the geostrophic relationship breaks down at the Equator. The summer circulation

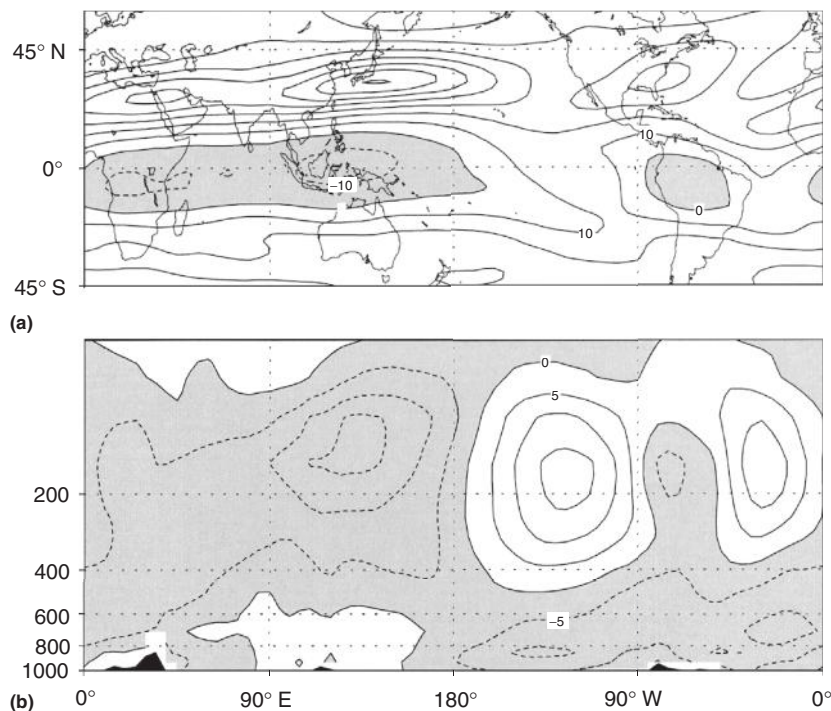


Figure 4 (a) The 200 hPa zonal wind in northern winter (DJFM), contoured in 10 m s^{-1} intervals. Regions where the zonal wind blows from the east are shaded. (b) Zonal-vertical cross-section of zonal wind at the Equator, contoured in 5 m s^{-1} increments, with shading for easterly regions. Top level in (b) is 50 hPa.

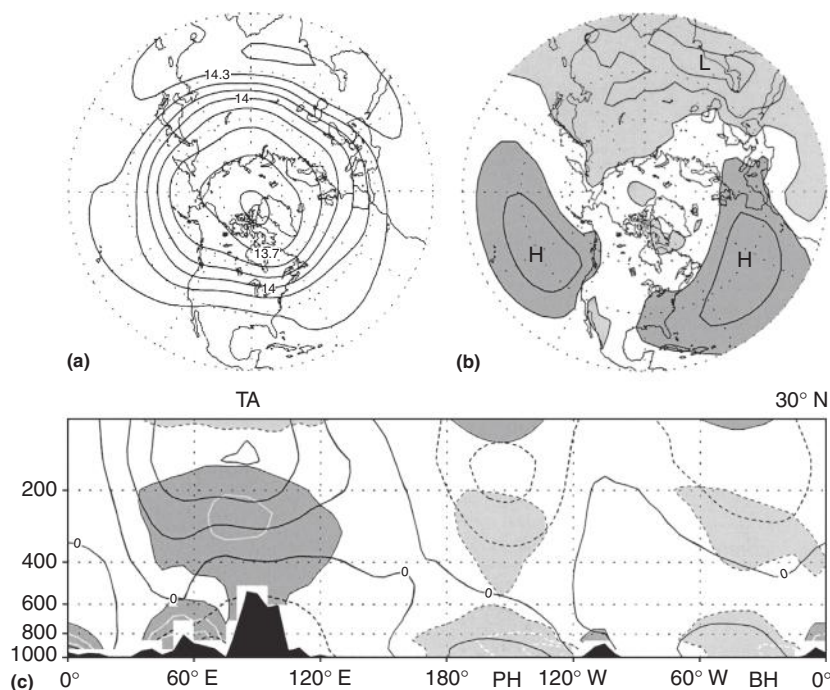


Figure 5 (a) Average height of the 150 hPa pressure surface in northern summer months (June, July, and August: JJA). The average is over 20 summer seasons (June 1980 through August 1999), and is computed from NCEP reanalysis. The contour interval is 100 m. (b) Average sea-level pressure for the same months, plotted as in Figure 1. The letters 'L' and 'H' designate the prominent centers of action: the Bermuda High, the Pacific High, and the broad region of low pressure associated with the Asian monsoon. Map domain begins at 15° N. (c) The 1000–100 hPa zonal-vertical cross-section of eddy height and temperature at 30° N, with contours and shading as in Figure 2(d). 'TA' gives the location of the Tibetan anticyclone, which is enclosed by the 14.3 km contour near the top of panel (a).

figures are thus complemented with a vector-wind plot at 850 hPa (Figure 6(c)); only the zonally asymmetric components of winds are plotted. Strong cross-equatorial flow occurs along the east coast of Africa, bringing moisture to the Asian continent. Easterly flow is found all along the Equator, particularly along the southern flank of the Pacific and Bermuda Highs.

Southern Hemisphere Stationary Waves

The Southern Hemisphere has much less land than the Northern Hemisphere, resulting in weaker asymmetries at its lower boundary. A more zonally symmetric circulation, with smaller-amplitude stationary waves, is thus expected. Due to the larger fraction of ocean, the seasonal cycle will also be muted. The seasonal change in surface temperature, for example, will be smaller than in the Northern Hemisphere.

The southern vortex is shown during the December–March (southern summer) and June–August (southern winter) periods in Figure 7. As before, the winter vortex is shown at 300 hPa and the summer one at the higher 150 hPa level. Thick lines mark the height contours while the shaded region shows the corresponding eddy height patterns. Note that in the Southern Hemisphere the flow around a low is clockwise rather than counterclockwise. The southern vortex is considerably more symmetric than the northern one. Eddy heights are thus smaller, and contoured at 25 m in both summer and winter (Figure 7). The summer and winter patterns are both dominated by the wave number 1 component in the high latitudes

so that opposite along a latitude circle have opposite polarities. The wave component exhibits similar phase and amplitude structure in the two seasons, indicating a significant role of Antarctic orography in its forcing. The subtropics shows greater seasonality, with a ridge over northern Australia in summer; this upper-level feature is linked to the Australian monsoon outflow.

The extent of zonal asymmetries at the surface is examined using sea-level pressure which is contoured with a 2.5 hPa interval as opposed to 5.0 hPa in the Northern Hemisphere. The summer distribution (Figure 7(c)) is much like the one in the Northern Hemisphere (Figure 5(b)), with high-pressure cells occupying the midlatitude ocean basins. In summer, the subtropical highs are interrupted by continental heat lows, caused by the warmer land temperatures. The winter sea-level pressure (Figure 7(d)) is more zonally symmetric, unlike the Northern Hemisphere where asymmetries are most pronounced during winter (cf. Figure 1 and Figure 2). A prominent feature of the southern winter pattern is the Mascarene High extending from Africa to Australia, which generates strong south-easterly flow along its northern flank. Its linkage with south-westerly flow over the northern Indian Ocean and Asian summer monsoon can be seen in Figure 6(c).

The vertical structure of Southern Hemisphere stationary waves is shown in Figure 8 at 30° S in summer and 60° S in winter – the latitude of the subtropical highs and the polar wavenumber 1 pattern, respectively. Contour intervals in Figure 8(a) are 10 m for height and 1.5 K for temperature, as

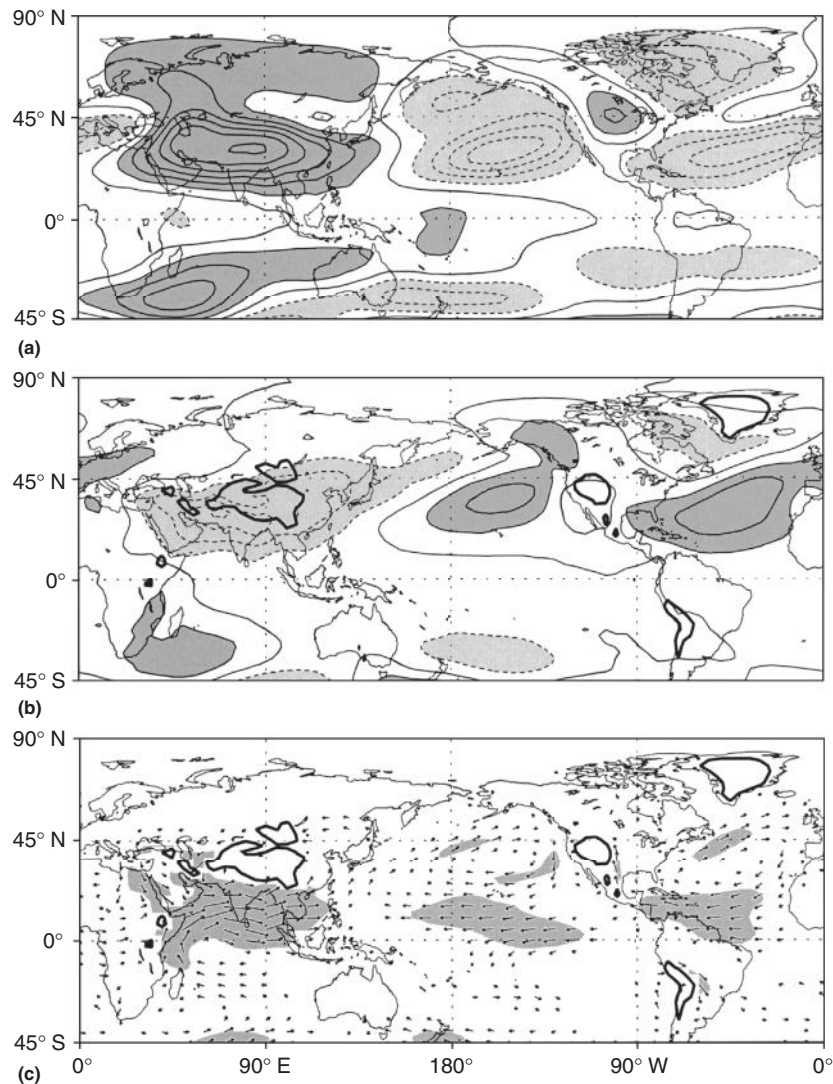


Figure 6 Eddy height at (a) 150 hPa and (b) 850 hPa during northern summer (JJA), with a contour interval of 25 m and dark (light) shading for positive (negative) values in excess of 25 m. (c) Eddy wind vectors at 850 hPa. Regions where the eddy wind speed is in excess of 5 m s^{-1} are shaded, and the longest arrow represents a wind speed of 18 m s^{-1} . Eddy winds with speeds below 2 m s^{-1} are suppressed. As in (b), the thick closed contours in (c) surround mountainous regions where the surface pressure is less than 850 hPa.

opposed to 50 m and 3.0 K in northern summer (Figure 5(c)). As in the Northern Hemisphere, the subtropical highs have a baroclinic structure with upper-level troughs superimposed on surface highs. The heat lows over Australia and southern Africa are quite shallow and intense: this is typical of arid regions where rainfall and mid-tropospheric latent heating do not occur in response to the surface heat low. The winter height and temperature structures (Figure 8(b)) are plotted using 25 m and 1.5 K intervals, as opposed to 50 m and 3.0 K in northern winter (Figure 2(c, d)), due to their relative weakness. The southern winter pattern evidently changes little with height. There is much less westward tilt in comparison with the northern winter structure (Figure 2(c)), indicating less upward propagation of wave energy. Although westerlies are necessary for upward propagation, theoretical considerations suggest that propagation is hindered by the presence of excessive westerlies

(westerlies exceeding the Rossby critical velocity), the southern winter vortex is substantially stronger than its northern counterpart (cf. Figure 1(a) and Figure 7(b)); note the larger contouring interval in the latter figure).

Transience in the Atmosphere

The above review of stationary wave structure does not convey the extent to which these waves are representative of the instantaneous circulation. For example, how stationary (or transient) is the upper-level circulation during northern winter? Can the stationary waves be ‘seen’ on synoptic weather charts? The degree to which these charts depart from the climatological pattern is a measure of the strength of the transient flow. Transient activity is estimated in northern winter in Figure 9

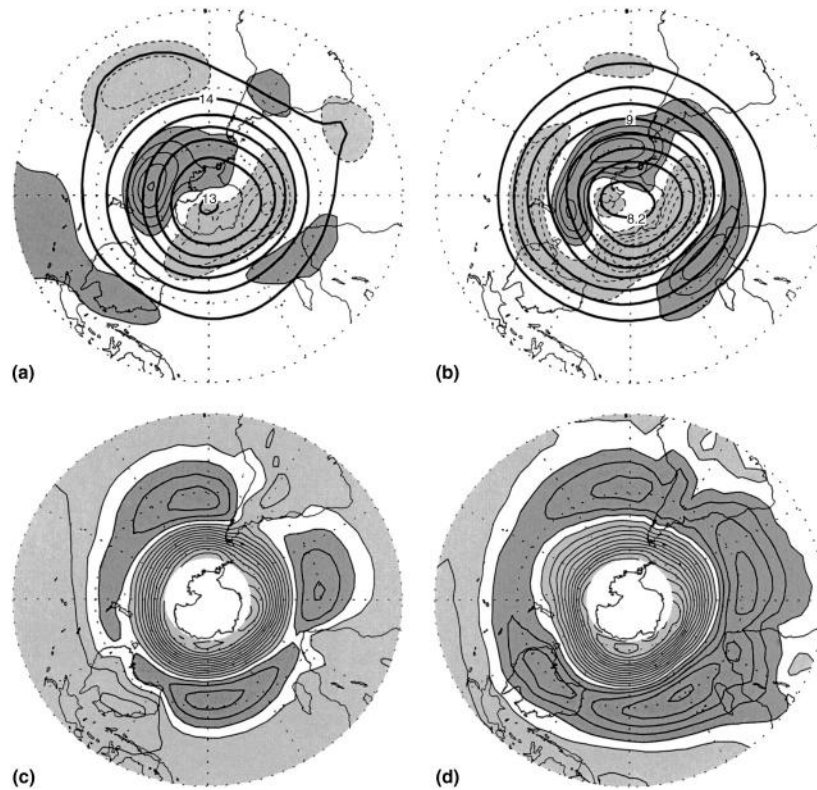


Figure 7 (a) Height of the 150 hPa pressure surface in the Southern Hemisphere during DJFM months (southern summer). (b) Height of the 300 hPa surface during JJA months (southern winter). Thick solid contours show the total height field in 200 m increments, while thin contours represent the eddy height. The contour interval for eddy height is 25 m, with dark (light) shading for positive (negative) values in excess of 25 m. The zero contour for eddy height is suppressed. (c, d) Sea-level pressure for (c) DJFM and (d) JJA months, in 2.5 hPa increments, with dark (light) shading for values above 1015 hPa (below 1012.5 hPa). Map domain is from the Equator to the South Pole. Sea-level pressure values over Antarctica are unreliable and hence suppressed.

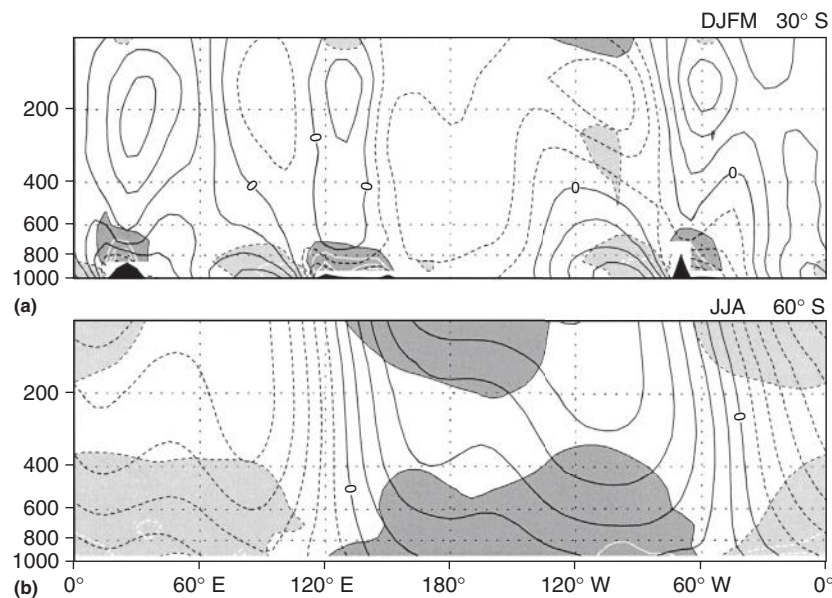


Figure 8 Zonal-vertical cross-sections of eddy height and temperature in the Southern Hemisphere at (a) 30° S in DJFM and (b) 60° S in JJA months. In (a), the contour interval for eddy height is 10 m, with dashed contours for negative values. The contour interval for eddy temperature in (a) is 1.5 K, with dark (light) shading for positive (negative) values in excess of 1.5 K, and zero contours suppressed. In (b), contour intervals for eddy height and temperature are 25 m and 1.5 K, respectively, with plotting conventions as in panel (a).

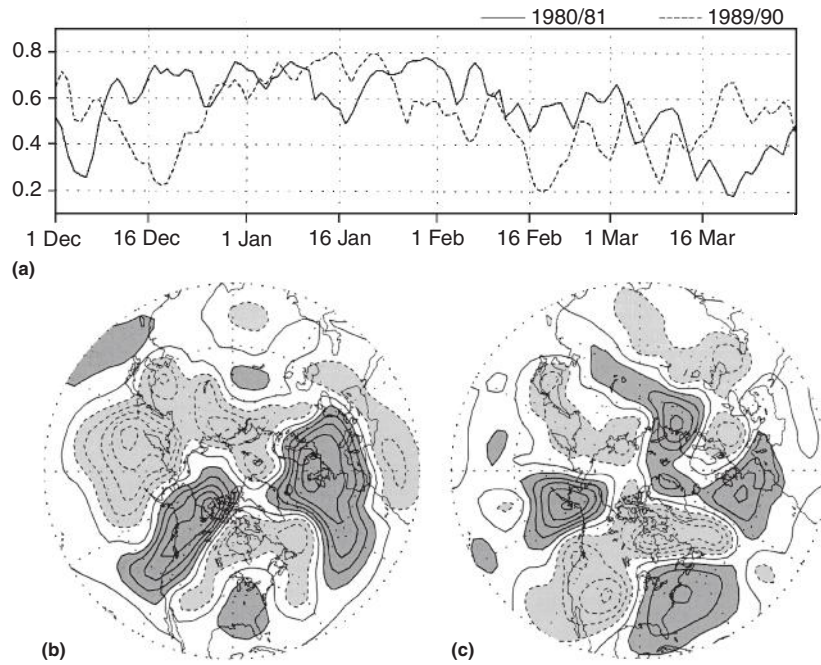


Figure 9 (a) Spatial correlation between the instantaneous (00UTC) and climatological 300 hPa eddy heights during 1980/81 (solid curve) and 1989/90 (dashed curve) winters. Correlations are area weighted and include height data from 20° N to the north pole. (b) 300 hPa eddy height on 1 February 1981, when the spatial correlation was 0.74. (c) Eddy height on 16 February 1990, when the correlation was 0.29. In (b) and (c), the contour interval is 100 m (twice the interval in Figure 2(a), with dark (light) shading for positive (negative) values in excess of 100 m.

because it is expected to be strongest in this season. The greater vertical shear of the thermally balanced Asian–Pacific and Atlantic jets in winter makes them prone to hydrodynamic instability, which in the context of geostrophic flows is called baroclinic instability. Baroclinic instability produces transient disturbances on subweekly time scales.

The extent to which the stationary wave structure is representative of instantaneous flow is depicted in Figure 9 by projecting the daily, instantaneous (00UTC), 300 hPa circulation on the climatological wave pattern (Figure 2(a)) during the winters of 1980/81 and 1989/90. Correlation – a measure of the structural similarity of the two maps (without regard to amplitude) – is plotted on the y-axis. The correlation ranges from 0.2 to 0.8 in these winters, indicating that the climatological pattern accounts for up to 65% of the spatial variance. High correlation is however achieved only on a few days in each winter. More typically, the correlation is 0.5 and 0.6 between. Interestingly, the correlation drops and recovers over a 2–3 week period, 1–2 times each winter, revealing the establishment time scale of the climatological pattern. Dynamical analysis of such episodes, especially of the recovery phase, can shed light on the establishment mechanisms of stationary waves.

The question of whether the climatological wave pattern can be ‘seen’ on synoptic charts is addressed in Figure 9(b) and (c), which show the instantaneous (00UTC) wave pattern on two days: 1 February 1981, when the spatial correlation is high (0.74; Figure 9(b)), and 16 February 1990, when the correlation is low (0.29; Figure 9(c)). The climatological pattern (Figure 2(a)) can be clearly recognized in the former plot, but not in the latter. Even when structurally similar, the patterns

can evidently have very different wave amplitudes; the contour interval is 50 m in Figure 2(a) but 100 m in Figure 9(b).

Forcing of Stationary Waves

Stationary waves are generated, ultimately, by the zonal asymmetries at the Earth’s surface: orography, continent–ocean contrasts, and sea surface temperature gradients. Through dynamic and thermodynamic interactions with the zonal-mean flow, and subsequent mutual interactions, surface inhomogeneities produce zonally asymmetric circulation and precipitation features at upper levels. Comprehensive numerical models of the atmosphere, which include coupling between physical and dynamical processes, are able to realistically model the observed stationary waves. In a sense, the often posed question – on relative contribution of orography and other processes in forcing of stationary waves – has been addressed by such prognostic general circulation models (GCMs). Comparison of GCM simulations obtained with and without orography provide insight. In these assessments, the change in the heating distribution is attributed to orographic forcing, whose circulation impact is found to be comparable to that of all other processes put together.

Historically, answers to the above question were sought in a framework where ‘orographic forcing’ was used more restrictively – to refer to the dynamical forcing of flow from mechanical diversion. In such analysis, the entire heating distribution, regardless of its origin (e.g., from condensation in adiabatically cooled upslope flow), was regarded as an independent forcing. This framework was adopted, perhaps, because mechanical

diversion of flow by an orographic barrier is conceptually easier to model. Such studies lead to the rapid advancement of stationary wave theory, including construction of potential vorticity conserving models for the responses to orograph, meridional and vertical wave propagation analysis, and understanding of troposphere–stratosphere interaction.

Theoretical Considerations

Large-scale atmospheric motions in the extratropics are approximately hydrostatic and quasi-geostrophic (QG) in character. The hydrostatic approximation recognizes the operative balance between the horizontally varying pressure and density perturbations, while the QG approximation acknowledges the near-balance between the Coriolis force and the horizontal pressure gradient. QG flow is thus dominated by the rotational component. Its evolution, however, is determined, in part, by the comparatively weaker divergent flow component, as described by the vorticity equation

$$\frac{\partial \zeta}{\partial t} + \mathbf{V}_h \cdot \nabla \zeta + \beta v = -(f + \zeta)(\nabla \cdot \mathbf{V}_h) \quad [1]$$

Here, ζ is the QG relative vorticity ($= \hat{k} \cdot (\nabla \times \mathbf{V}_h) = \nabla^2 \psi$), ψ is the geostrophic streamfunction, f is the Coriolis parameter with $\partial f / \partial y = \beta$, and $\mathbf{V}_h$ is the horizontal QG flow. The right-hand term is the product of absolute vorticity ($f + \zeta$) and horizontal convergence, and is often called the ‘stretching term’ because convergent flow leads to stretching of vortex tubes.

Due to compressibility of air, evolution of the thermodynamic state is conveniently described using potential temperature, $\theta = T(p_0/p)^{R/C_p}$, which is conserved in adiabatic motion; p_0 is the reference pressure (1000 hPa). Potential temperature thus changes only in response to diabatic heating $\dot{Q}$, as follows:

$$\frac{\partial \theta}{\partial t} + \mathbf{V}_h \cdot \nabla \theta + w \frac{\partial \theta}{\partial z} = \frac{\dot{Q}}{C_p} (\theta_0/T_0) \quad [2]$$

where $\theta_0 = T_0(p_0/p)^{R/C_p}$, with $T_0 = T(p_0)$.

Since stationary waves refer to the zonally varying component of the flow (here onward denoted by prime), their dynamics can be described, to first order, by linearizing eqns [1] and [2] about the zonal-mean circulation, $\bar{U}(y, z)$ and $\bar{\theta}(y, z)$. The linearized equations are valid for small-amplitude perturbations:

$$\zeta'_t + \bar{U} \zeta'_x + v'(\beta - \bar{U}_{yy}) \approx -f(\nabla \cdot \mathbf{V}'_h) \quad [3]$$

$$\theta'_t + \bar{U} \theta'_x + v' \bar{\theta}_y + w' \bar{\theta}_z \approx (\dot{Q}' \theta_0 / T_0 C_p) \quad [4]$$

For convenience, subscripts are used to denote the partial derivatives.

Orographic Forcing and Response

The forcing and propagation of stationary waves can be discussed using eqns [3] and [4]. In contrast with diabatic heating, which is explicitly present as right-hand forcing in the thermodynamic equation, the mechanical forcing by orography (h') is implicitly present through its kinematic impact on vertical velocity at the lower boundary (w_s). In the presence of the zonal-mean circulation, the linearized vertical velocity, w'_s , equals $\bar{U} h'_x$.

In simplified treatments of orographic interaction, the geophysical fluid is additionally considered to be homogeneous and incompressible ($\nabla \cdot \mathbf{V}_h = -\partial w / \partial z$), so that response is determinable using the vorticity equation alone – this ‘shallow water’ approximation is indeed reasonable for the interaction of oceanic flows with underwater topography, but somewhat limited in capturing aspects of the atmospheric interaction. In shallow water theory, density (or temperature) is constant, and the horizontal flow, including horizontal divergence, is height independent. Assuming that a rigid lid is placed at the top of the fluid ($z = H$), so that vertical velocity vanishes there, one obtains $\partial w' / \partial z \approx -(\bar{U} h'_x) / H$. The forced waves are then modeled by eqn [5], where $\bar{U}$ has been additionally assumed to be latitude independent, and perturbation vorticity is dissipated (e.g., by Ekman spin-down) on an ε^{-1} time scale:

$$[\partial / \partial t + \bar{U} \partial / \partial x + \varepsilon] (\psi'_{xx} + \psi'_{yy}) + \beta \psi'_x \approx -(f/H) \bar{U} h'_x \quad [5]$$

To understand the forced response, consider an arbitrary Fourier component of the geostrophic streamfunction: $\psi'(x, y, t) = \text{Real}\{\hat{\psi}_{k,l} e^{i(kx+ly-\omega t)}\}$, where the hat denotes the complex amplitude corresponding to zonal and meridional wavenumbers, k and l , and associated frequency ω . For such a perturbation, eqn [5] yields the solution

$$\hat{\psi} = \frac{f \hat{h}}{H[(k^2 + l^2)(\bar{U} - c) / \bar{U} - \beta / \bar{U} - i\varepsilon(k^2 + l^2) / (\bar{U} k)]} \quad [6]$$

For stationary waves ($\omega = 0$), the zonal phase speed, $c (= \omega/k)$, vanishes. This simplifies the first term in the denominator to $(k^2 + l^2)$. In the presence of dissipation, the orography and streamfunction are not in phase, since the denominator in eqn [6] is complex. The possibility of resonance is also indicated in the inviscid case ($\varepsilon = 0$), when $(k^2 + l^2) = \beta / \bar{U}$. Dissipation however limits the wave amplitude at resonance, with $\hat{\psi} \propto i \hat{h}$. The streamfunction and orography are 90° phase-shifted (or in quadrature) in this case, with the trough in the flow being a quarter wavelength downstream of the mountain ridge. When forcing is on larger scales, $(k^2 + l^2) < \beta / \bar{U}$, planetary vorticity advection dominates zonal advection of relative vorticity in balancing the orographically induced vorticity on upslopes and downslopes. Both the real and imaginary parts of the denominator in eqn [6] are negative in this case, which puts the trough within a quarter wavelength downstream of the ridge.

It is interesting that troughs in the observed 300 hPa stationary wave pattern (cf. Figure 2(a)) are also downstream of the orographic features, but the extent to which these are forced by local orography remains somewhat uncertain, as discussed later. Also, the assumed Fourier representation of the streamfunction implies the presence of meridional boundaries which confine wave energy to a midlatitude channel – a set-up conducive for resonance. The sinusoidal zonal structure is also unrealistic, since orographic features are generally localized. In nature, the wave energy propagates away zonally, meridionally, and vertically from the localized forcing region, thus calling into question the validity of the solution [6].

The pedagogically useful shallow-water model of orographic interaction is limited for other reasons as well. The tropospheric flow cannot be assumed to be homogeneous and incompressible, since cooling (heating) from adiabatic

expansion (compression) during ascent (descent) is important in the thermodynamic budget. Moreover, a flow configuration which satisfies the vorticity eqn [3] can be unbalanced from the viewpoint of this budget. For example, when planetary vorticity advection dominates the left-hand side in balancing the upslope divergent flow in eqn [3], the thermodynamic budget [4] is unbalanced in the absence of condensation heating (precipitation), since both adiabatic ascent and equatorward flow lead to colder temperatures. The generation of orographic response must thus be understood by considering the vorticity and thermodynamic equations together, so that any implications that one may have for the other are fully accounted for. Such considerations lead to the development of the QG potential vorticity equation.

QG Potential Vorticity Equation

The prediction equation for QG flow that does not explicitly reference the divergent flow component is called the QG potential vorticity equation. Although it can be derived quite generally, the focus here is on its simplified linearized version, when the zonal-mean flow $\bar{U}$ is independent of latitude. The equation is derived by eliminating the divergent flow from eqns [3] and [4]. In the $z^* = (RT_0/g) \ln(p_0/p)$ coordinate, which reduces to geometric height in an isothermal atmosphere, the QG potential vorticity equation is

$$q'_t + \bar{U}q'_x + v'\bar{q}_y = \frac{R}{Hc_p\rho_0} \frac{\partial}{\partial z^*} \left(\frac{\rho_0 \dot{Q}'}{N^2} \right) \quad [7]$$

where

$$q'(x, y, z) = \psi'_{xx} + \psi'_{yy} + \frac{f^2}{N^2\rho_0} \frac{\partial}{\partial z^*} \left(\rho_0 \frac{\partial \psi'}{\partial z^*} \right)$$

and

$$\bar{q}_y(z) = \beta - \frac{f^2}{N^2\rho_0} \frac{\partial}{\partial z^*} \left(\rho_0 \frac{\partial \bar{U}}{\partial z^*} \right) \quad [8]$$

Since divergent flow is not referenced by this equation, it is of some interest to examine the manifestation of orographic forcing in this analysis framework. Not surprisingly, this forcing enters as a lower boundary condition, but in the thermodynamic equation [4]. This is because of the direct reference to vertical velocity in eqn [4], in contrast with the vorticity equation [3] which refers only to its vertical gradient. With $w'_s = \bar{U}h'_x$, the boundary condition conveying orographic forcing is

$$\begin{aligned} \theta'_t + \bar{U}\theta'_x + v'\bar{\theta}_y \\ = -\bar{U}h'_x\bar{\theta}_z + (\dot{Q}'\theta_0/T_0c_p) \text{ at the surface} \end{aligned} \quad [9]$$

Assume for purposes of this discussion that diabatic heating vanishes at the surface, so that only adiabatic cooling (warming) is occurring on the upslope (downslope). In steady flows, the heating can be balanced by zonal eddy advection and/or meridional advection of mean temperature. If upslope cooling is compensated by the latter, the upslope flow will be poleward, and a high-pressure center will be positioned over the mountain ridge near the surface. The response at upper levels depends upon the zonal scale of mountains: large wavelengths will propagate into the lower stratosphere, and phase lines will tilt westward with increasing height, all as depicted in the

(Figure 10(a) schematic. It is interesting that although thermal advection and vertical wave propagation are absent in the shallow water model, the horizontal structure of the long-wavelength solution (in the presence of damping, eqn [6] is not too different from that indicated at upper levels in Figure 10(a).

Heating Response

The stationary wave response to heating can be qualitatively understood from the thermodynamic equation [4]. In the deep tropics, horizontal variations of geopotential (and temperature) are much smaller since it is difficult to maintain them in the presence of the weak Coriolis force. Consequently, horizontal temperature advection is ineffective in balancing diabatic heating in eqn [4]. Away from the surface, heating is

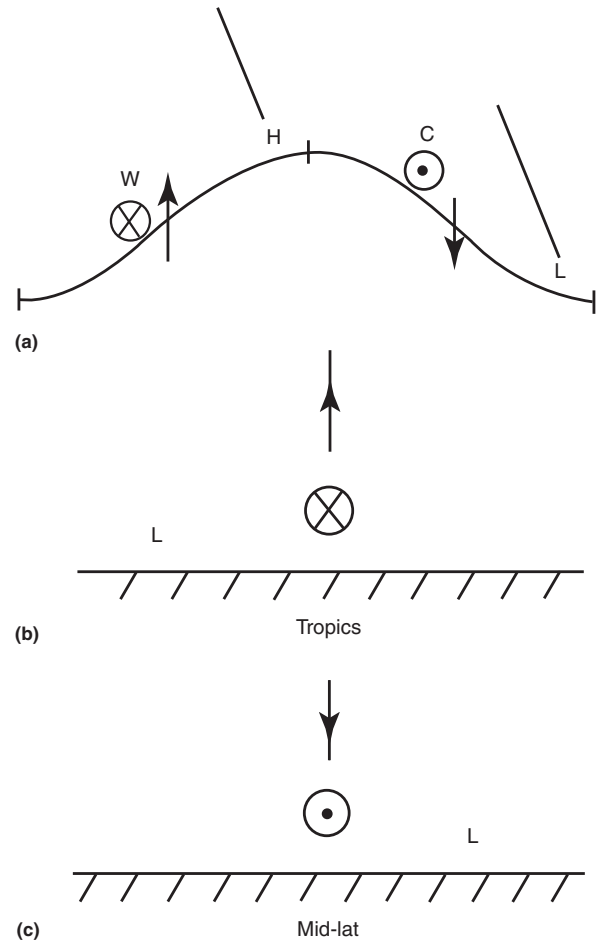


Figure 10 Schematic depiction of the longitude–height response forced by (a) westerly flow over midlatitude orography, (b) tropical heating, and (c) midlatitude heating, all taken from Hoskins and Karoly (1981). The orographic response is shown for the long-wavelength case, and is determined from both dynamic and thermodynamic (i.e., quasi-geostrophic potential vorticity) considerations. The arrows depict vertical motion, and circled crosses and dots denote poleward and equatorward flow, respectively. H and L denote the pressure ridge and trough, with the lines showing the vertical tilt of the pressure wave. W and C are the warmest and coldest air, respectively.

thus balanced, almost entirely, by adiabatic cooling, with the vertical profile of w' closely following that of heating. A substantial portion of heating in the tropics results from deep convection, which produces strongest heating in the mid-to-upper troposphere, as shown later in Figure 11(c). Such heating distribution leads to convergence (divergence) in the lower (upper) troposphere, which results in vortex stretching (squashing). The rotational response to the induced vorticity depends on the horizontal forcing scale: if the scale is large, the stretching is offset by poleward advection of planetary vorticity, which is tantamount to the surface low being positioned westward of the heat source, as schematically illustrated in Figure 10(b).

In the midlatitudes, heating does not extend as deeply into the troposphere as in the tropics. Heating in the Pacific and Atlantic stormtracks, for example, is confined mostly to the lower troposphere, as shown later in Figure 11(b). Midlatitude heating is offset to a large extent by horizontal temperature advection; larger temperature gradients are sustainable in midlatitudes due to the greater Coriolis force. Large-scale heating in midlatitudes is balanced, mostly, by cold advection from the north; the near-surface low is thus positioned eastward of the heating. Interestingly, vertical motion in the vicinity of midlatitude heating is determined by vorticity balance considerations – a complete reversal of the tropical

situation: cold advection from the north brings with it higher vorticity air as well, and this induced vorticity advection must be offset if a steady state is to be maintained. The compensation is accomplished by vortex compression, which has implications for the temperature field.

Wave Propagation

The qualitative arguments discussed above are helpful in understanding the nature of response in the forcing region. The stationary wave response is however not confined to the forcing region alone, since Rossby waves propagate zonally, meridionally, and vertically, carrying the disturbance (energy) into the far field (unforced region). The energy propagation, or group velocity, characteristics depend both on the perturbation scale and structure of the basic state. Some zonal-mean zonal wind configurations encourage Rossby wave propagation, while others impede it. Basic state flow can thus profoundly impact wave propagation into the tropics and the stratosphere.

Theoretical analysis helps to focus on the basic state attributes that are influential, e.g., the direction and curvature of the zonal-mean zonal wind. A useful quantity in wave propagation analysis is the refractive index which seizes on these and other relevant attributes. A display of refractive index variations is often helpful, since it conveys, to first order, the wave

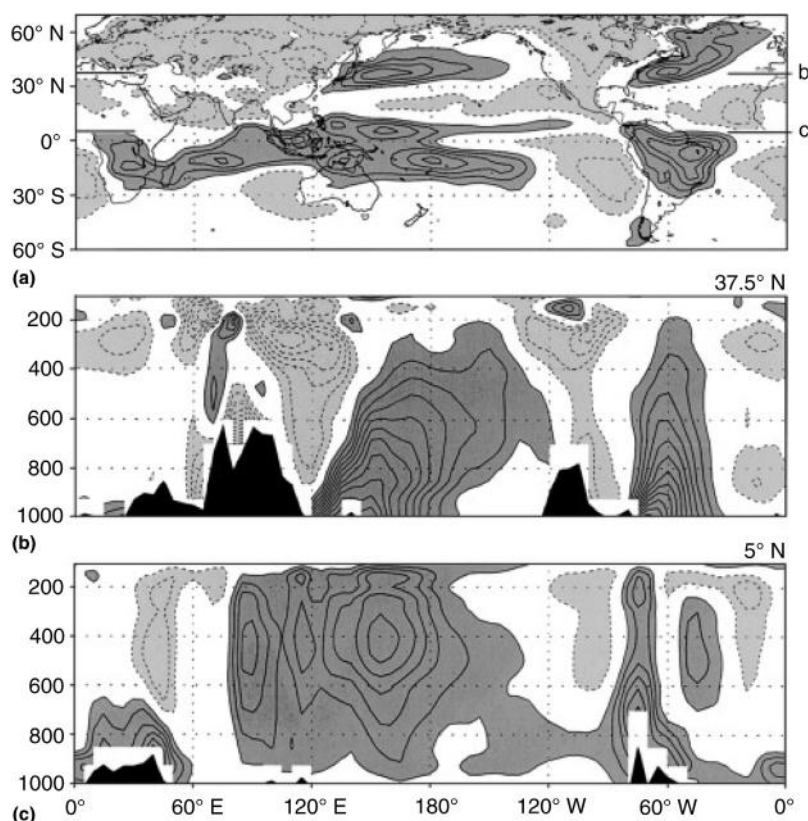


Figure 11 (a) Mass-weighted vertical average of diabatic heating, calculated as a residual from the thermodynamic equation. The winter season (DJFM) diagnosis is obtained from NCEP reanalysis fields 20 winter seasons (1979/80–1998/99). The contour interval is 0.5 K day^{-1} , with dark (light) shading for positive (negative) values in excess of 0.5 K day^{-1} , and zero contours suppressed. (b, c) Zonal-vertical (1000–100 hPa) cross-section of diabatic heating at (b) 37.5° N and (c) 5° N , with contours and shading as in panel (a). The latitudes of the cross-sections in (b) and (c) are marked with thick lines at the edges of panel (a). A 9-point smoother is applied to the heating field before plotting.

propagation pathways, as waves are generally refracted towards higher refractive index regions. Such analysis suggests that midlatitude stationary waves are refracted towards the Equator, drawn there by the large index values resulting from diminishing westerly winds. The tropical easterlies, in contrast, present an effective dynamical barrier to equatorward propagation of midlatitude stationary waves. In the vertical, waves with large horizontal scales alone can propagate upward, but only when the upper-level westerlies are not too strong.

Diabatic Heating in Northern Winter

Diabatic heating plays a prominent role in the forcing of stationary waves. In stationary wave theory, it is an explicit forcing in the QG potential vorticity equation [7], and even orographic forcing in this theoretical framework manifests as surface heating [9].

In nature, heating resulting from the change of phase of water substance, turbulent eddy diffusion, and short-wave and long-wave radiative fluxes is referred to as diabatic heating. (Note that the temperature of air parcels can change even without any diabatic heating, from adiabatic compression or expansion.) In contrast with Earth's orography, whose highly accurate measurements are widely known, the three-dimensional structure of diabatic heating is only beginning to be described. The main reason why the heating distribution has remained uncertain is that, unlike other quantities, heating is not directly measured. It is, instead, estimated, usually as a residual in the thermodynamic budget. Since the heating estimate is only as good as the quality of atmospheric data from which it is diagnosed, the quality of atmospheric analysis is critical for the diagnosis. Fortunately, data coverage and quality and analysis methods have all improved in the last two decades, and are reflected in the modern reanalysis data sets. Heating diagnosis from one such data set, the US National Centers for Environmental Prediction (NCEP) reanalysis, is shown in **Figure 11**.

The mass-weighted vertical average of diabatic heating,

$$\frac{1}{(p_s - 100)} \int_{100}^{p_s} \dot{Q} c_p^{-1} dp$$

is shown during northern winter in units of K day^{-1} ; here, p_s is the surface pressure, c_p is the specific heat of air at constant pressure, and the integration is from the surface to 100 hPa. Key features in **Figure 10(a)** include the heating centers in the extratropical Pacific and Atlantic basins, which effectively define the two midlatitude stormtracks. The northern continents, in contrast, constitute the cooling regions. In the tropics, heating is strong over the South Pacific convergence zone and the Amazon basin. A narrow zone of heating is also present in the Pacific just northward of the Equator; this intertropical convergence zone (ITCZ) is much stronger during northern summer when it is positioned a few degrees farther northward and fully extended across the Pacific basin.

Diabatic heating has a complicated vertical structure which changes with latitude and season. The changes with latitude are shown in **Figure 10(b)** and (c), which depict height–longitude cross-sections through the midlatitude stormtracks (37.5° N

and the ITCZ (5° N). The stormtrack heating is evidently strongest near the surface, with peak values close to 6 K day^{-1} , and diminishes rapidly with height. Latent heating due to precipitation in baroclinically unstable synoptic-scale disturbances is the primary contributor to stormtrack heating. Diabatic cooling, on the other hand, is comparatively weaker, and focused more near the tropopause in this estimation, for reasons that are not clear.

The vertical structure of ITCZ heating is strikingly different. Although the entire column is being heated, heating is generally strongest in the mid-to-upper troposphere. For example, over the tropical Pacific warm pool – the site of persistent deep convection – heating is strongest ($\sim 3 \text{ K day}^{-1}$) at 400 hPa. In contrast, heating over land (e.g., equatorial Africa) is strongest near the surface due to sensible heating. The heating structure over Central America is also similar, except that elevated surface heating there has produced some deep convection as well.

Interaction with Transients

The climatological stationary waves coexist with vigorous atmospheric motions occurring on a variety of time scales (**Figure 9**), and there are strong interactions between these transient motions and the stationary waves. Transient motions are the instantaneous departures of the flow from its climatological state, and the time mean of transient motion thus vanishes, by definition. However, fluxes of heat and vorticity by transients do not vanish in general. For example, the contribution of transients to the advection terms in eqn [2] can be written as

$$\begin{aligned} V''_h \cdot \nabla \theta'' + w'' \partial \theta'' / \partial z \\ \approx \nabla \cdot (V''_h \theta'') + \partial (\omega'' \theta'') / \partial p \end{aligned} \quad [10]$$

where the double prime denotes the transient component and ω is the vertical velocity in p -coordinates. The right-hand side of eqn [10] is the heat-flux divergence from transient motions. In synoptic systems, northward (southward) transient motions are typically accompanied by positive (negative) temperature fluctuations, so that heat flux diverges to the south of a storm-track and converges to the north. The heat-flux divergence acts as heat sources and sinks for the time-mean flow, and the stationary waves respond to this thermal forcing just as they respond to diabatic heating. Likewise, the convergence of transient vorticity flux ($-\nabla \cdot (V''_h \zeta'') - \partial (\omega'' \zeta'') / \partial p$) provides sources and sinks of vorticity for the stationary waves.

The net effect of transient thermal and vorticity fluxes on stationary waves is not easy to characterize. However, it is clear that transient forcing is strong enough in northern winter to exert a powerful influence on stationary waves. The 700 hPa heat-flux convergence by perturbations lasting less than 1 month is superimposed in **Figure 12(a)** on the local winter eddy temperature pattern. The two fields evidently oppose each other. For example, transient heat fluxes diverge from the warmer regions over the Atlantic and the west coast of North America and converge in the colder regions above northeastern Canada. Thus, throughout most of the northern midlatitudes, transient thermal fluxes have a damping effect on the lower tropospheric stationary eddy temperature pattern. Furthermore, the forcing by transient thermal fluxes is on the

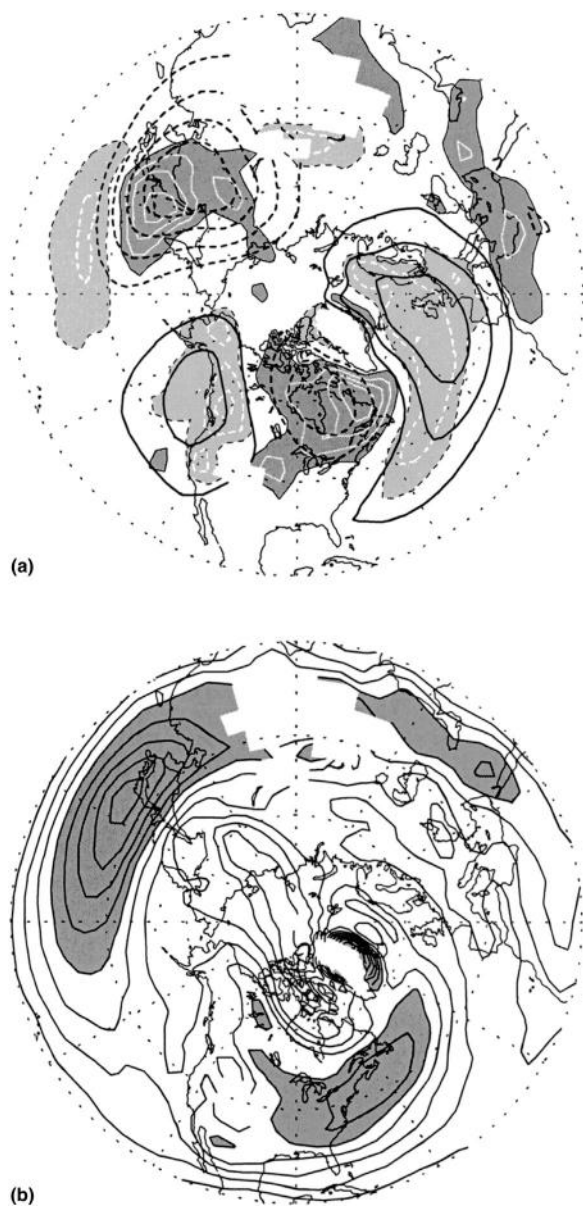


Figure 12 (a) The 700 hPa eddy temperature (thick contours) and heat-flux convergence by transient motions (shading and white contours) in northern winter months (DJFM). The contour interval for temperature is 2 K, and dashed lines represent negative values. The contour interval for heat-flux convergence is 0.5 K day⁻¹, with dark (light) shading for positive (negative) values in excess of 0.5 K day⁻¹. Zero contours for temperature and heat-flux convergence are suppressed, and regions where the surface pressure is less than 700 hPa are masked out. (b) Eady growth parameter in northern winter (DJFM), calculated from the 700 hPa temperature gradients. The contour interval is 0.1 day⁻¹, with dark shading for values in excess of 0.6 day⁻¹.

order of 0.5 K day⁻¹, while the eddy temperatures are about 4 K. In the absence of other processes, it would take little more than a week for the thermal fluxes to reduce dramatically the 700 hPa eddy temperature field. Such a reduction also implies a substantial weakening of the upper-level geopotential pattern, since temperature is the vertical derivative of geopotential in hydrostatic balance.

While transients are an important influence on stationary waves, the stationary waves can be equally influential for the transients. One way in which stationary waves organize transients is by creating localized regions of strong cyclogenesis. Synoptic systems tend to develop in regions of strong lower tropospheric temperature gradients, and such regions are present off the coasts of Asia and North America in Figure 12(a). The ability of the local temperature gradient to enhance the growth of synoptic systems can be measured by the Eady growth parameter (ϵ):

$$\epsilon = 0.31 \frac{|\nabla T|}{|T(T_0 \partial \ln \theta / g \partial z)|^{1/2}}$$

Its plot in Figure 12(b) shows large values in the same regions where stormtrack heating occurs in Figure 11(a). Comparison of these panels suggests that stationary waves play an important role in determining the locations of stormtracks. In addition to this effect on growth rate, stationary waves can also have a mechanical effect on stormtracks, by changing the steering winds that determine storm paths and by exchanging mechanical energy with the storms.

The Eady growth rate is applicable to synoptic transients, which grow by extracting energy from the thermal gradients (or vertical shear) of the climatological state. Synoptic systems, in fact, account for less than half of the transient forcing in the vorticity and thermodynamic equations for the climatological stationary waves. Furthermore, the slower transients (those with time scales between, say, 10 days and 1 month) are quite distinct from synoptic transients. They do not generally travel along concentrated stormtracks, nor do they typically grow by extracting energy from the climatological temperature gradients. Although these transients are certainly influenced by the climatological stationary waves, the nature of this influence is rather complex and cannot be easily summarized.

Modeling of Northern Winter Stationary Waves

Modeling of orographically forced stationary waves dates back to the seminal paper of Charney and Eliassen in 1949, in which linear shallow water theory was applied to the longitudinal distribution of orography at 45° N. The earlier discussion here, including the development of eqn [5] and its solution [6], largely follows the analysis reported in that paper. Charney and Eliassen found the midlatitude mountains to be rather influential, accounting for almost all of the observed signal in their analysis.

Since that time, diabatic heating due to continent–ocean contrasts, and transient fluxes of heat and momentum have also been advocated as important mechanisms for the generation of stationary waves. In the intervening period, the atmosphere has been more closely observed, both spatially and temporally, and there has been a tremendous increase in computational power for modeling studies. A reassessment of the relative roles of various forcing mechanisms is thus in order. In effect, more complete versions of the dynamical and thermodynamical equations can now be solved globally at high resolution, and verified against the extensive record of upper-air observations that have been compiled since.

It is still advantageous to linearize the system of equations, at least initially, since this allows the influence of individual forcing terms to be examined separately. Of course, the forcing terms can have strong mutual interactions. For example, heating can cause eddy flow which then impinges on mountains, and the subsequent orographically induced uplift can generate convection and lead to further heating. However, linear diagnostic models serve a valuable purpose in assessing the relative importance of the different forcing terms in different regions. They also indicate the degree to which linear perturbation theory applies to stationary waves.

A linear simulation of northern winter stationary waves is presented in Figure 13. The linear model uses $\sigma = p/p_s$ as the vertical coordinate (here p_s is the surface pressure) so that mountains do not intersect the model levels. There are 15 levels in the vertical, ranging from 1000 to 25 hPa, and the horizontal resolution is 7.5° in the zonal direction and about 2.5° in the meridional direction. As in eqns [3] and [4], the model is linearized about the zonal-mean climatological state. The momentum and thermal dissipation in the model is roughly equivalent to a 5-day damping time scale in eqns [3] and [4]. The forcing consists of three-dimensional diabatic heating (Figure 11), orographic height, and transient fluxes of heat and momentum.

The 300 hPa response obtained with all forcings is shown in Figure 13(a), and plotted using the convention used in Figure 2(a), with which it should be compared. The linear model can simulate the ridge over the Atlantic, the low off the east Asian coast, and the trough over Canada. A notable flaw in the simulation is the weakness of the ridge over the Rockies. The solution shows that linear perturbation theory can explain many, though not all, aspects of the observed stationary wave pattern.

The model response when forced separately by diabatic heating, mountains, and transients is shown in panels (b)–(d), respectively; note that contour interval in these three panels is half that in panel (a). All three forcings contribute significantly to the total pattern. The heating response includes jets in the Asian-Pacific and Atlantic sectors. Heating is evidently important in establishing the ridge over the Atlantic and northern Europe, and contributes significantly to the trough over Canada as well.

The response to mountains in panel (c) shows troughs downstream of the Himalayan-Tibetan complex and the Rockies, as suggested by the shallow water solution [6] and also QG potential vorticity considerations (cf. Figure 10(a)), in each case for long wavelengths. Thus, orography contributes to the forcing of the jets as well. The high amplitudes directly over

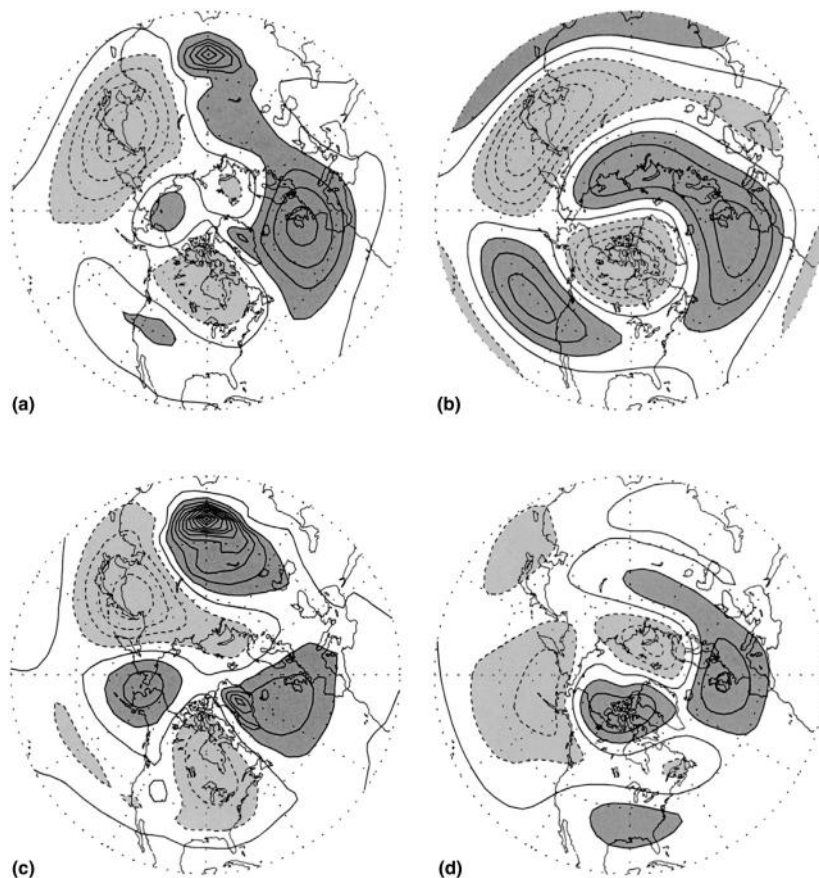


Figure 13 (a) The 300-hPa height response of a linear stationary wave model forced by heating, mountains, and transient fluxes of heat and momentum. The contour interval is 50 m, with dark (light) shading for positive (negative) values in excess of 50 m. (b–d) Response of the model when forced separately by (b) heating, (c) mountains, and (d) transient fluxes. The contour interval in (b–d) is 25 m.

Greenland and Tibet are a consequence of the linearization of the hydrostatic equation in σ -coordinates. Examination of geopotential heights gives a somewhat misleading impression that waves generated by mountains propagate primarily in the zonal direction. Examination of the modeled streamfunction (a more suitable variable for describing the rotational response in the tropics; not shown), however, reveals considerable equatorward propagation of the forced waves.

The forcing by submonthly transients (panel d) produces a somewhat intricate pattern with no clear relationship to the synoptic stormtracks. Transients are apparently responsible for a large part of the response over eastern Atlantic and northern Europe.

Studies of stationary wave dynamics have traditionally focused on the question of the relative importance of the various forcing terms in generating the observed pattern. Yet recent simulations such as the one in [Figure 13](#) show clearly that the northern winter stationary waves do not constitute a simple linear response to a single form of forcing. Furthermore, linearized equations, such as eqns [3] and [4], neglect the advection of eddy heat and vorticity by the stationary waves themselves, and also the effect of eddy winds impinging on the mountains. These terms play an important role in generating some features of the stationary wave pattern, such as the ridge over the Rockies. Future examinations of stationary wave dynamics will have to assess not only

the relative importance of various forcing terms but their mutual interactions, and the nonlinear interactions of the stationary waves themselves.

See also: **Climate and Climate Change:** Climate Variability: Seasonal and Interannual Variability. **Dynamical Meteorology:** Coriolis Force; Overview. **Stratosphere/Troposphere Exchange and Structure:** Global Aspects. **Synoptic Meteorology:** Cyclogenesis.

Further Reading

- Gill, A.E., 1982. *Atmosphere–Ocean Dynamics*. Academic Press, Orlando.
- Grotjahn, R., 1993. *Global Atmospheric Circulations: Observations and Theories*. Oxford University Press, New York.
- Holton, J.R., 1992. *An Introduction to Dynamic Meteorology*. Academic Press, New York.
- Hoskins, B.J., Karoly, D.J., 1981. The steady linear response of a spherical atmosphere to thermal and orographic forcing. *Journal of the Atmospheric Sciences* 38, 1179–1196.
- Hoskins, B.J., Pearce, R., 1983. *Large-Scale Dynamical Processes in the Atmosphere*. Academic Press, London.
- James, I.N., 1994. *Introduction to Circulating Atmospheres*. Cambridge University Press, Cambridge.
- Saltzman, B., Manabe, S. (Eds.), 1985. *Advances in Geophysics*, vol. 28. *Issues in Atmospheric and Oceanic Modeling*. Academic Press, Orlando.

Symmetric Stability

HB Bluestein, University of Oklahoma, Norman, OK, USA

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

The necessary conditions for symmetric and conditional symmetric instability in a frictionless atmosphere are considered on the basis of parcel theory. The relationship between symmetric and conditional symmetric stability, and gravitational and inertial instability is noted. The possible relationship between some mesoscale precipitation bands in extratropical systems and in tropical cyclones, and conditional symmetric instability, is discussed.

Introduction

Symmetric stability is a state of the atmosphere in which an inviscid, dry air parcel displaced from its equilibrium position with respect to some axis along/about which the flow has no variations, i.e., along/about an axis of symmetry, experiences a restoring force that makes it oscillate about its original position. For axially symmetric displacements in a ring about an axially symmetric vortex the wave motions are called *inertial* or *centrifugal* waves. Centrifugal oscillations are like buoyancy waves with the horizontal centrifugal (inertial) force playing the role of buoyancy (gravity).

Similar oscillations can also occur in a statically stable, rotating atmosphere when the thermal-wind shear vector is unidirectional and does not vary along the direction it is oriented. Parcels in the form of a tube are displaced in a vertical plane normal to the thermal-wind vector. In this case the axis of symmetry is the axis along which the thermal wind is directed.

If potential temperature increases with height and the geostrophic absolute vorticity is anticyclonic, the atmosphere is *inertially unstable*; if the potential temperature decreases with height and the geostrophic absolute vorticity is cyclonic, the atmosphere is *gravitationally* (or *statically*) *unstable*. If the geostrophic absolute vorticity is cyclonic and potential temperature increases with height the atmosphere is both *inertially stable* and *gravitationally stable*; however, if infinitesimal displacements in the plane normal to the vertical shear are accompanied by forces that move the air parcel farther away from its equilibrium position, the atmosphere is *symmetrically unstable*. Since the atmosphere is baroclinic, owing to the thermal wind, this instability is a special case of baroclinic instability for a flow in which there is no temperature-gradient component along the axis of symmetry.

When tubes of moist, unsaturated air are lifted in a symmetrically stable atmosphere to a level at which condensation occurs (and water and ice loading are not significant or are neglected) and thence to a level at which the atmosphere is symmetrically unstable with respect to saturated processes (i.e., when vertical trajectories follow surfaces of constant equivalent or wet-bulb potential temperature instead of surfaces of potential temperature), then the atmosphere is in a state of *conditional symmetric instability* (CSI). CSI is analogous to *conditional instability* for air parcels lifted vertically. Since CSI involves forces that are both horizontally and vertically directed, the process by which the instability is released is also referred to as *slantwise convection*. When a layer of moist air that

is initially symmetrically stable is lifted to saturation and the vertical displacement of air itself creates the conditions for slantwise convection, then the process is referred to as *potential symmetric instability* (PSI), which is analogous to potential instability for upright convection. At saturation, CSI and PSI are equivalent.

The importance of CSI is that it is thought to be responsible for the formation of some mesoscale bands of precipitation that are oriented along the thermal wind. Since the thermal wind is oriented along the elongated zone of strong temperature gradient associated with fronts and is quasi-two-dimensional, CSI may be triggered in response to slantwise, ageostrophic, frontal circulations initiated by confluence/diffuence acting on a cross-frontal temperature gradient. It is also thought that CSI may be responsible for eyewall rainbands in some tropical cyclones.

The Parcel Theory of Symmetric Instability in an Inviscid, Dry Atmosphere

The analysis of symmetric stability is simplified by using a parcel approach analogous to that used in the parcel theory of upright convection. Consider a Cartesian coordinate system in which there is a temperature gradient in the y - p plane and that $\partial/\partial x$ of all variables is 0 (this choice of an axis of symmetry is arbitrary; sometimes the y axis is chosen to be the axis of symmetry). For simplicity the dynamics are described for the Northern Hemisphere. Consider the quantity:

$$m = u - fy, \quad [1]$$

where u is the x -component of the wind and f is the Coriolis parameter. In inviscid flow m , the *absolute momentum* or *pseudo-angular momentum* is conserved; it is attributed to an infinitesimal tube of air extending through some point (y, p) infinitely off in both the $+x$ and $-x$ directions.

The inviscid momentum equation in the y direction, with height as the vertical coordinate, is

$$\frac{Dv}{Dt} = -fu - \frac{1}{\rho} \frac{\partial p}{\partial y} = -f(m - m_g), \quad [2]$$

where v is the y -component of the wind, ρ is the density, p is the pressure, D/Dt is the total (material) derivative, and the geostrophic absolute momentum:

$$m_g = u_g - fy, \quad [3]$$

where u_g is the geostrophic component of the wind in the x direction. Therefore the net force in the y direction on a tube is proportional to the difference between the m of the tube, which is conserved, and the m_g of the environment into which the tube is displaced. It is assumed for simplicity that the tube does not mix with its environment. The original value of m of the tube is just its geostrophic value at its equilibrium point in the y - p plane. Therefore there will be a net force in the y direction on the tube if it moves into an environment where m_g is different from that of its equilibrium, starting location. Vertical gradients of m_g are associated with thermal-wind shear in the x direction (i.e., with temperature gradients in the y direction); gradients of m_g in the y direction are associated with geostrophic absolute vorticity.

The inviscid vertical equation of motion is

$$\frac{Dw}{Dt} = \frac{g}{\theta}(\theta' - \theta), \quad [4]$$

where w is the vertical velocity, g is the acceleration of gravity, θ is the potential temperature of the environment, and θ' is the potential temperature of the tube. If the flow is adiabatic and there is no diffusion of heat, θ' is conserved following the motion of the tube. It is assumed that the environment is not disturbed by the tube's motion so that there is no vertical perturbation pressure-gradient force. Therefore there will be a net force in the vertical on the tube if it moves into an environment where θ is different from that of its equilibrium, starting location.

Whether or not there is a restoring force on the tube that brings it back to its equilibrium point about which it undergoes a stable oscillation (symmetric stability) or whether it continues to move in the direction of its displacement (symmetric instability) depends on how the surfaces of m_g and θ are oriented and what the direction of displacement is with respect to the surfaces (Figure 1). Symmetric instability is possible (panel b of Figure 1) if the slope of the θ surfaces is greater than the slope of the m_g surfaces and if the tube is displaced infinitesimally along a plane whose slope is intermediate between that of the θ surfaces and that of the m_g surfaces (i.e., along paths a or c , but not along paths b or d), and if $\partial\theta/\partial z > 0$ and $\partial m_g/\partial y < 0$. If $\partial\theta/\partial z < 0$ (panel (d) of Figure 1) or if $\partial m_g/\partial y > 0$ (panel (e) of Figure 1), then the atmosphere is statically unstable or inertially unstable, respectively, and not symmetrically unstable. Panel (a) in Figure 1 depicts neutral stability and panel (c) in Figure 1 depicts absolute stability.

The thermal-wind relation in terms of potential temperature is to a good approximation:

$$\frac{\partial u_g}{\partial z} = -\frac{g}{f} \frac{1}{\theta} \frac{\partial \theta}{\partial y}. \quad [5]$$

The slope of a surface of constant θ is therefore

$$\left(\frac{dz}{dy}\right)_\theta = \frac{f(\partial u_g/\partial z)}{g(\partial \theta/\partial y)}, \quad [6]$$

and the slope of a surface of constant m_g is

$$\left(\frac{dz}{dy}\right)_{m_g} = \frac{f - \partial u_g/\partial y}{\partial u_g/\partial z}. \quad [7]$$

It follows that the necessary condition for symmetric instability is

$$Ri = \frac{(g/\theta)(\partial\theta/\partial z)}{(\partial u_g/\partial z)^2} < \frac{f}{(\zeta_g + f)}, \quad [8]$$

where Ri is the Richardson number for the geostrophic wind and ζ_g is the geostrophic vorticity, which for symmetric flow ($\partial/\partial x = 0$) is $-\partial u_g/\partial y$. In typical synoptic-scale flow in mid-latitudes the geostrophic vorticity is an order of magnitude smaller than f ; then the necessary condition for symmetric instability is that $Ri < 1$. In the vicinity of fronts where geostrophic vorticity is much larger, Ri must be smaller.

Ertel's potential vorticity for an atmosphere in geostrophic and hydrostatic balance is

$$Z = \frac{Cp}{\rho} f \frac{g}{\theta} \frac{\partial \theta}{\partial z} \left(\frac{\zeta_g + f}{f} - \frac{1}{Ri} \right). \quad [9]$$

It follows from eqn [8] that an equivalent necessary condition for symmetric instability is that Ertel's potential vorticity for the geostrophic wind is negative (anticyclonic in either hemisphere). Since

$$(\zeta_g + f)_\theta = (\zeta_g + f)_z - \frac{f}{Ri}, \quad [10]$$

where $(\zeta_g + f)_\theta$ is the geostrophic absolute vorticity evaluated on an isentropic surface and $(\zeta_g + f)_z = f - \partial u_g/\partial y$ is the geostrophic absolute vorticity evaluated on a surface of constant height, then negative (anticyclonic in either hemisphere) isentropic geostrophic absolute vorticity is also an equivalent necessary condition for symmetric instability. Thus, symmetric instability is favored on the anticyclonic-shear side of jets and jet streaks or near sharply curved ridges of high pressure.

It can also be shown that the necessary condition for symmetric stability is equivalent to the ellipticity condition for the Sawyer–Eliassen equation, which describes the vertical circulation about a front forced by geostrophic confluence/diffuence and differential diabatic heating and whose dynamics are governed by the geostrophic-momentum approximation. Since the Sawyer–Eliassen equation is a second-order, constant-coefficient, partial differential equation, the condition of ellipticity is necessary for it to have unique solutions. Thus, balanced frontal circulations are possible only if the atmosphere is symmetrically stable. However, if friction is included in the equations of motion, it turns out that the ellipticity condition can be met even when Ertel's potential vorticity is negative.

The Parcel Theory of Slantwise Convection in an Inviscid, Moist Atmosphere

The analysis of symmetric instability in a moist atmosphere is complicated by latent heat release, evaporation and melting-related cooling, and by water and ice loading. The governing momentum equation remains as eqn [2]. The governing vertical equation of motion, on the other hand, is different from eqn [4] since it must account for latent heat release, and if there is a condensate, for water and ice loading also. Surfaces of constant entropy that account for latent heat release and for condensate loading replace potential temperature in eqn [4]. If both the environment and the tube are unsaturated and there is

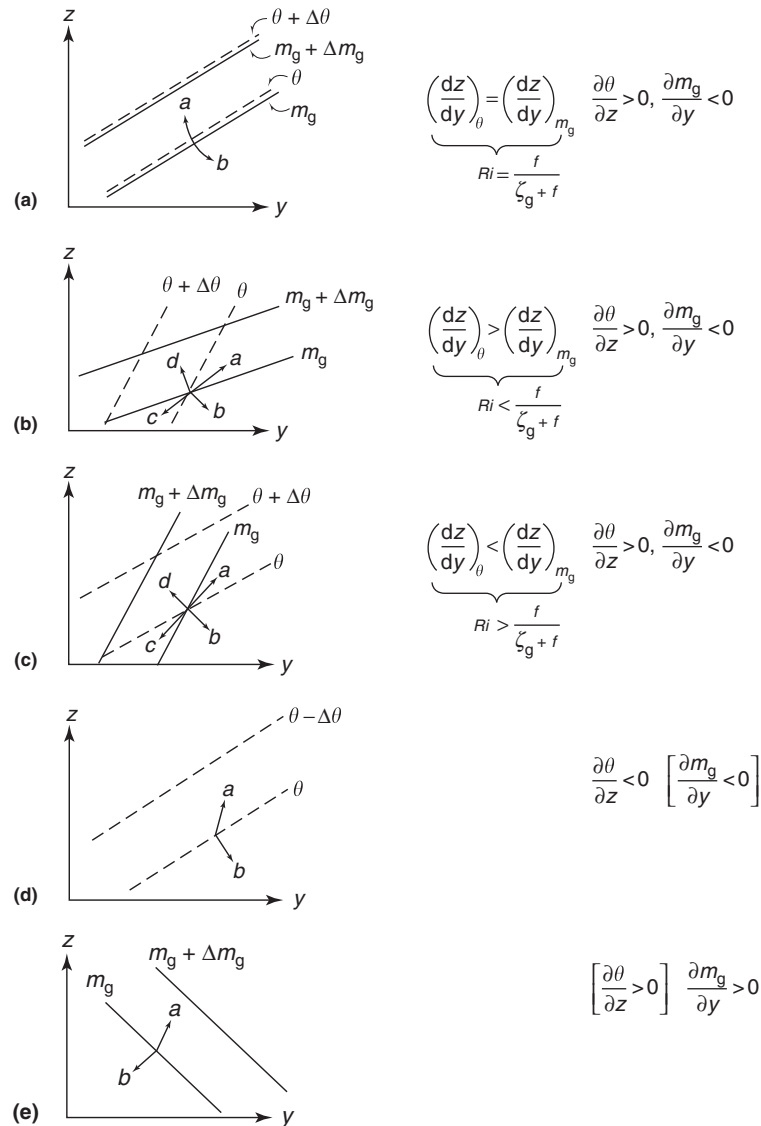


Figure 1 Stability criteria for symmetric flow ($\partial/\partial x = 0$) in terms of the slope of the m_g (solid lines) and potential-temperature θ (dashed lines) surfaces. Displacements in the directions a , b , c , and d are referred to in the text. (a) Neutral stability for displacements along the θ and m_g surfaces; otherwise stable for infinitesimal displacements; (b) symmetric instability: unstable for infinitesimal slantwise displacements intermediate in slope between that of θ surfaces and m_g surfaces; otherwise stable; (c) absolute stability: stable for any infinitesimal slantwise displacement; (d) gravitational instability: unstable for any infinitesimal slantwise displacement not along a θ surface; and (e) inertial instability: unstable for any infinitesimal slantwise displacement not along an m_g surface. Adapted from Bluestein, H.B., 1993. *Synoptic-Dynamic Meteorology in Midlatitudes. Observations and Theory of Weather Systems*, vol. II. By permission of Oxford University Press, New York.

no condensate, then virtual potential temperature (θ_v) may be used in place of potential temperature. If the tube is saturated and the environment is unsaturated, and if condensate is ignored, then potential temperature of the tube may be replaced by equivalent virtual potential temperature (θ_{ev}); if both the tube and the environment are saturated, and if condensate is ignored, then the potential temperature of both the tube and environment may be replaced by equivalent virtual potential temperature. (Wet-bulb virtual potential temperature (θ_{wv}) may be used instead of equivalent virtual potential temperature.)

For the purpose of illustration consider an atmosphere that is unsaturated and has no condensate, but is moist.

Suppose that the distribution of θ_v , θ_{ev} , and m_g is as shown in **Figure 2**. Since the slope of the surfaces of constant θ_v are not steeper than the surfaces of constant m_g at low levels, the atmosphere there is symmetrically stable or even neutral with respect to unsaturated displacements. However, there are regions aloft where the slope of surfaces of constant θ_{ev} is steeper than the surfaces of m_g . In these regions, if condensate loading is ignored, the atmosphere is symmetrically unstable with respect to infinitesimal saturated displacements. Since the stability depends upon whether or not a tube is saturated or unsaturated, the symmetric instability condition is *conditional*.

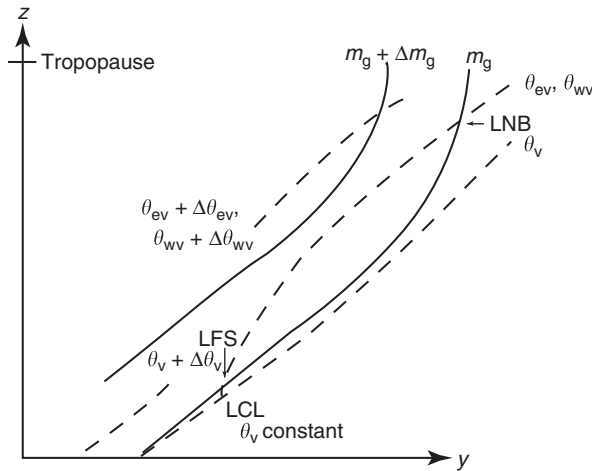


Figure 2 Idealized example of a vertical cross section in the Northern Hemisphere, normal to the thermal-wind shear vector, showing surfaces of constant m_g (solid lines), and constant θ_v , θ_{ev} , and θ_{wv} . In this example θ_v , θ_{ev} , and θ_{wv} increase with height (gravitational and conditional stability), m_g decreases with increasing y (inertial stability), and θ_v decreases with y (baroclinic atmosphere, with colder air at larger values of y). Lifting condensation level (LCL); level of free slantwise convection (LFS); level of neutral buoyancy (LNB). Below the LCL the slope of the θ_v surface is less than that of the m_g surface. Note that the slope of θ_{ev} and θ_{wv} surfaces is greater than the slope of θ_v surfaces because the lapse rate of a θ_v surface is greater than that of a θ_{ev} or θ_{wv} surface, and θ_v , θ_{ev} , and θ_{wv} decrease with y . Adapted from Bluestein, H.B., 1993. *Synoptic-Dynamic Meteorology in Midlatitudes. Observations and Theory of Weather Systems, vol. II*. By permission of Oxford University Press, New York.

Suppose an unsaturated tube at low levels is lifted a *finite* distance along a surface of constant θ_v (e.g., by the ascending branch of a frontal circulation or more slowly as a result of quasigeostrophic forcing) until it reaches its lifting condensation level and that condensate is ignored: If lifted any further, it follows a surface of constant θ_{ev} . Thus far the tube is neutrally buoyant. Owing to the inclusion of the effects of latent heat, the surfaces of constant θ_{ev} have different slopes than that of the surfaces of constant θ_v . The m of the tube is greater than that in its environment everywhere to the right of the original m_g surface; therefore according to eqn [2] the tube is symmetrically stable because it feels a restoring force that has a component to the left; if the tube were not forced any further, it would become negatively buoyant and move back down and to the left toward its original equilibrium position. The tube is symmetrically stable even though it is saturated and the slope of the surfaces of constant θ_{ev} is greater than the slope of the surfaces of constant m_g because the tube has undergone a finite displacement rather than an infinitesimal displacement.

If the tube is lifted further, however, so that eventually it crosses to the left of the original m_g surface, and it is displaced upward and to the right at a slope intermediate between that of the θ_{ev} and m_g surfaces, then according to eqns [4] and [2] it would continue to accelerate upward and to the right if it were not forced any more. The level at which it would first realize symmetric instability is called the *level of free slantwise convection*, in analogy with the level of free convection for upright convection.

Eventually the tube will reach a level at which it crosses back to the right side of the original m_g surface where the slope of the θ_{ev} surfaces is now less than that of the m_g surfaces. Above this level, the *level of neutral buoyancy* (LNB) for slantwise convection, the atmosphere is symmetrically stable. The LNB tends to be near the tropopause where θ surfaces are more horizontally oriented owing to the strong static stability [6] in the lower stratosphere and where m_g surfaces are more vertically oriented owing to the lack of horizontal temperature gradient and thermal-wind shear [7] at the tropopause.

The amount of kinetic energy it takes to lift a tube to its level of free slantwise is called the *slantwise convective inhibition* (SCIN). The potential energy in the environment that is converted into kinetic energy of the tube while it is symmetrically unstable is called the *slantwise convective available potential energy* (SCAPE). SCIN and SCAPE are the analogs to the convective inhibition and convective available potential energy (CAPE) in the parcel theory of upright convection. The SCAPE is equivalent to CAPE computed along a surface of constant m_g . In the geostrophic coordinate system used in semigeostrophic theory, m_g surfaces are parallel to the geostrophic coordinate, which is directed opposite in direction to the horizontal temperature gradient. Thus SCAPE is CAPE computed in geostrophic coordinates. In the limit of vanishing baroclinicity, m_g surfaces become vertically oriented (see eqn [7] when $\partial u_g / \partial z \rightarrow 0$) and SCAPE is identical to CAPE.

If the atmosphere is saturated, then an equivalent necessary condition for CSI is that Ertel's potential vorticity for saturated moist processes is negative (anticyclonic in either hemisphere). If the effects of condensate are accounted for, the criteria for CSI are more complicated. The loading depends on what phase of water substance is present; condensate that falls out does not follow air parcel motion and may evaporate into unsaturated air.

In order for precipitation bands to form as a result of slantwise moist convection, there must be an adequate supply of water vapor, strong enough lift to release CSI, and the necessary conditions for CSI must be satisfied.

The slantwise ascent of symmetrically unstable tubes must be compensated for by slantwise-descending air. The regions of slantwise-descending air modify the environment so as to make the tubes less symmetrically unstable, just as compensating subsidence around buoyant clouds (according to the *slice method* in the theory of upright convection) warms the environment and lessens the buoyancy in the clouds (i.e., the CAPE is diminished). The narrower the slantwise-ascending branch and the wider the slantwise-descending branch, the less is the slantwise acceleration (i.e., the less is the SCAPE). The most unstable configuration is one of thin, relatively rapidly slantwise-ascending layers of saturated air sandwiched in between thick, less-rapidly slantwise-descending layers of unsaturated air that are being cooled evaporatively from precipitation falling out from above. If the most unstable configuration is the one most likely to occur, then CSI precipitation bands should be relatively narrow and widely spaced.

The horizontal scale of CSI precipitation bands estimated from the horizontal extent of a sloping m_g surface [7] is U_g/f , where U_g is the change in geostrophic wind in the layer of CSI. For typical values of U_g in midlatitudes, the horizontal scale of CSI precipitation bands is on the order of 100 km, which is mesoscale.

Observations of Conditional Symmetric Instability and Precipitation Bands

Slantwise convection may be triggered suddenly after a gradual buildup in SCAPE by synoptic-scale or mesoscale processes, or it may be in statistical equilibrium so that the SCAPE is nearly zero and constant. Since the latter is the frequently observed case, it is not easy to use SCAPE as a forecast tool because its absence does not preclude the possibility that CSI is in fact being released; the existing observational network cannot resolve the space and time scales of the production of CSI and its release when the atmosphere is in statistical equilibrium.

Bands of precipitation ahead of warm fronts may be due to CSI. The bands are important in that large amounts of precipitation may accumulate in mesoscale regions, which makes the forecasting of floods and heavy snow difficult; synoptic-scale dynamics, on the other hand, can only explain how lower amounts of precipitation accumulate over broader regions. The slanted convection in the eyewall of some rapidly intensifying tropical cyclones might also be a result of CSI. In this case, the temperature gradient is directed toward the center of the tropical cyclone.

While lightning activity is common when conditional instability is released, it is thought it can also occur when CSI is released, even though updrafts in slantwise convection are usually much weaker. Lightning is commonly observed in the trailing precipitation region of mesoscale convective systems and sometimes in wintertime convection and in the eyewall of deepening tropical cyclones, where the likelihood of conditional instability is small, but there can be CSI.

When Ertel's potential vorticity for moist processes is very small, i.e., when the atmosphere is nearly neutral with respect to CSI, the atmosphere's response to frontogenetical forcing is enhanced. Since frontal secondary circulations and the secondary circulations in tropical cyclones themselves can produce precipitation bands, it is therefore not always easy to distinguish between bands forced by the secondary circulations and the bands forced by CSI.

The state of the atmosphere sometimes evolves so that the necessary conditions for gravitational, inertial, and symmetric instability/CSI appear and disappear. When small-scale

gravitational instability weakens and is followed by mesoscale, symmetric instability, there is said to be 'upscale development.' There is some evidence that the trailing, 'stratiform' precipitation region in some mesoscale convective systems may represent upscale development, as air ascends at the leading convective line in an environment of conditional instability, and then leans over into a trailing region of CSI.

When mesoscale frontal ascent in a symmetrically unstable atmosphere leads to latent heat release and is followed by small-scale, elevated gravitational instability, there is said to be 'downscale development.' In any event, it has been argued that there is a continuum between the gravitational and symmetric instabilities, so that a distinction between the two is somewhat arbitrary.

See also: Dynamical Meteorology: Inertial Instability; Potential Vorticity; Quasigeostrophic Theory. **Mesoscale Meteorology:** Cloud and Precipitation Bands. **Synoptic Meteorology:** Frontogenesis.

Further Reading

- Bluestein, H.B., 1993. Synoptic-Dynamic Meteorology in Midlatitudes. In: *Observations and Theory of Weather Systems*, vol. II. Oxford University Press, New York.
- Emanuel, K.A., 1994. *Atmospheric Convection*. Oxford University Press, New York.
- Holton, J.R., 1992. *An Introduction to Dynamic Meteorology*. Academic Press, New York.
- Lilly, D.K., 1986. Instabilities. In: Ray, P.S. (Ed.), *Mesoscale Meteorology and Forecasting*. American Meteorological Society, Boston, pp. 259–271.
- Schultz, D.M., Schumacher, P.N., 1999. The use and misuse of conditional symmetric instability. *Monthly Weather Review*, 2709–2732. American Meteorological Society, Boston.
- Schumacher, R.S., Schultz, D.M., Knox, J.A., 2010. Convective snowbands downstream of the Rocky Mountains in an environment with conditional, dry symmetric, and inertial instabilities. *Monthly Weather Review*, 4416–4438. American Meteorological Society, Boston.
- Thorpe, A.J., 1999. Dynamics of mesoscale structure associated with extratropical cyclones. In: Shapiro, M.A., Gronas, S. (Eds.), *The Life Cycles of Extratropical Cyclones*. American Meteorological Society, Boston, pp. 285–296.
- Xu, Q., 1989. Frontal circulations in the presence of small viscous moist symmetric instability and weak forcing. *Quarterly Journal of the Royal Meteorological Society*, 1325–1353. Royal Meteorological Society, Reading, UK.

Vorticity

JR Holton, University of Washington, Seattle, WA, USA

© 2003 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 6, pp 2500–2504, © 2003, Elsevier Ltd.

Introduction

The spin of a solid body is characterized by the angular velocity about its axis of rotation. This angular velocity is related in a simple manner to the spin angular momentum, which is conserved in the absence of torques, thus providing a powerful constraint on the motion. Owing to the fact that, unlike a solid body, a fluid provides little resistance to shearing deformation, angular velocity as defined for a solid body is not a suitable measure of rotation in a fluid. In the presence of velocity shear, straight-line flow can have strong local rotation, while curved flow may have zero local rotation (Figure 1). Vorticity is a vector field which, by providing a local measure of the instantaneous rotation of a fluid parcel, plays a role in fluid dynamics analogous to angular velocity in solid body

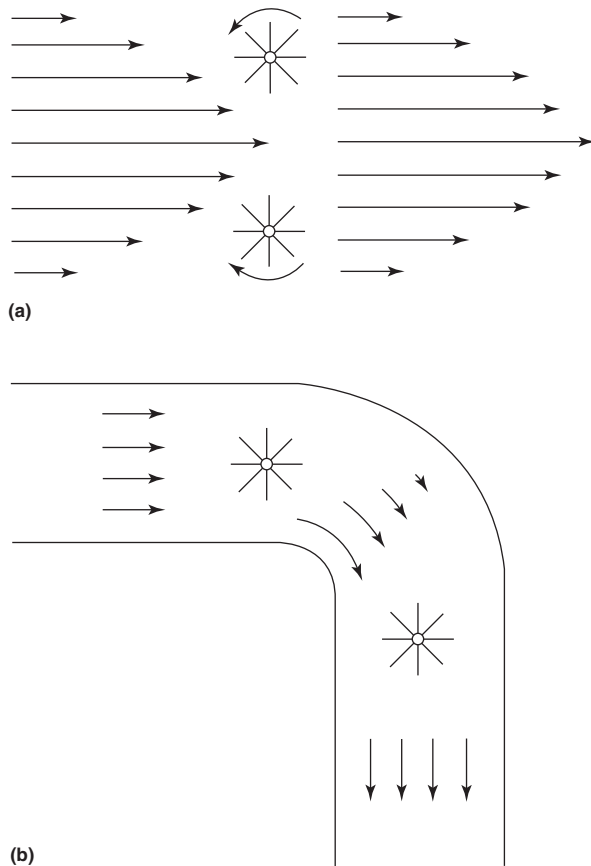


Figure 1 Behavior of a paddle wheel spinning about a vertical axis under two types of horizontal flow: (a) linear jet with positive and negative shear vorticity and (b) curved flow with zero net vorticity. Reproduced from Holton, J.R., 1992. *An Introduction to Dynamic Meteorology*, third ed. Academic Press, Orlando, FL.

mechanics. A related quantity called potential vorticity (*see Dynamical Meteorology: Overview*) is a fluid analog to the spin angular momentum of a solid body.

In most meteorological systems, horizontal wind velocities greatly exceed the vertical velocity, and the vertical component of vorticity (which depends on the horizontal velocity field) is of primary interest. Vorticity has the property that it is advected by the winds, but at the same time influences the wind distribution. Temporal and spatial variations of vorticity play key roles in the diagnosis of the development and evolution of weather systems, and of the general circulation of the atmosphere.

The Nature of Vorticity

Vorticity, ω , is defined as the curl of the fluid velocity. It is thus a vector field, and has a direction and amplitude at every point in space. Lines parallel to the vorticity field are referred to as vortex lines, while the volume defined by all of the vortex lines passing through a closed loop is referred to as a vortex tube. In meteorology, it is necessary to distinguish between the absolute vorticity ω_a , which is the curl of the absolute velocity observed in a coordinate frame fixed in space ($\mathbf{U}_a$), and the relative vorticity ω , which is the curl of the relative velocity observed in a coordinate system rotating with the Earth ($\mathbf{U}$, the wind velocity):

$$\omega_a \equiv \nabla \times \mathbf{U}_a, \quad \omega \equiv \nabla \times \mathbf{U} \quad [1]$$

In Cartesian coordinates rotating with the Earth, with x, y , and z directed eastward, northward, and upward, respectively, and with the velocity components designated by $\mathbf{U} = (u, v, w)$, the components of vorticity are

$$\omega = \left(\frac{\partial w}{\partial y} - \frac{\partial v}{\partial z}, \frac{\partial u}{\partial z} - \frac{\partial w}{\partial x}, \frac{\partial v}{\partial x} - \frac{\partial u}{\partial y} \right) \quad [2]$$

Large-scale dynamic meteorology is concerned with the vertical components of absolute and relative vorticity, which are designated by η and ζ , respectively:

$$\eta \equiv \mathbf{k} \cdot (\nabla \times \mathbf{U}_a), \quad \zeta \equiv \mathbf{k} \cdot (\nabla \times \mathbf{U})$$

That vorticity is the natural measure of local fluid rotation that can be demonstrated by considering the special case of horizontal motion with no z dependence: $\mathbf{U} = (u, v)$. In that case, the relative vorticity has only a vertical component:

$$\zeta = \frac{\partial v}{\partial x} - \frac{\partial u}{\partial y} \quad [3]$$

Figure 2 shows two short line segments, radiating from point A at (x_0, y_0) of lengths δx and δy , which are perpendicular to each other at time $t = 0$. The velocity at point A is (u_0, v_0) .

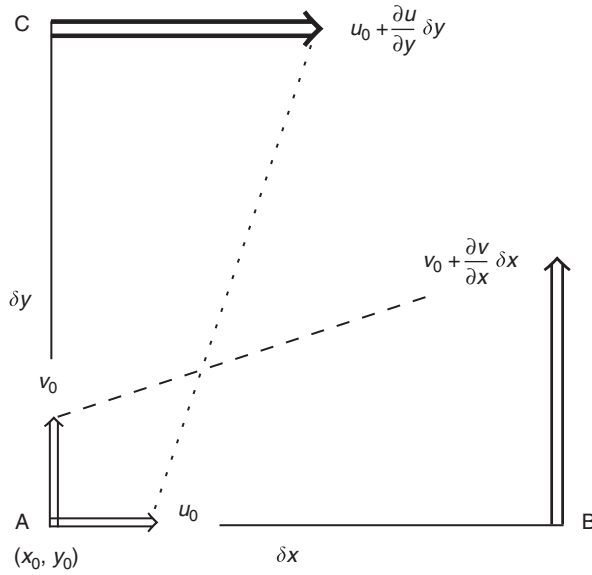


Figure 2 Relation of vorticity in two-dimensional flow to the instantaneous rotation of two orthogonal line segments.

The velocity normal to the line AB at the point $x_0 + \delta x$ is $v_0 + (\partial v / \partial x) \delta x$, while the velocity normal to the line AC at the point $y_0 + \delta y$ is $u_0 + (\partial u / \partial y) \delta y$. The instantaneous counterclockwise angular velocity of line AB is thus $[(v_0 + (\partial v / \partial x) \delta x) - v_0] / \delta x = \partial v / \partial x$, while for line AC the instantaneous counterclockwise angular velocity is $[-(u_0 + (\partial u / \partial y) \delta y) + u_0] / \delta y = -\partial u / \partial y$. The average of the angular velocity in the two orthogonal directions is thus $\zeta / 2 = \frac{1}{2}(\partial v / \partial x - \partial u / \partial y)$, that is, one-half of the vorticity.

The difference between absolute and relative vorticity is the planetary vorticity, which is just the local vertical component of the vorticity of the Earth owing to its rotation; $2\Omega \sin \phi \equiv f$, where ϕ is the latitude, and f is referred to as the Coriolis parameter. Thus, $\eta = \zeta + f$, or, in Cartesian coordinates,

$$\zeta = \frac{\partial v}{\partial x} - \frac{\partial u}{\partial y}, \quad \eta = \frac{\partial v}{\partial x} - \frac{\partial u}{\partial y} + f$$

The Relation between Vorticity, Circulation, and Potential Vorticity

Vorticity is related to another quantity, referred to as the fluid circulation. Circulation, which is a scalar integral quantity, is a macroscopic measure of rotation for a finite area. Vorticity, on the other hand, is a vector field that gives a microscopic measure of the rotation at every point in the fluid.

The circulation, C , is defined as the line integral about the contour, $\mathbf{l}$, of the component of the velocity vector that is locally tangent to the contour (Figure 3):

$$C \equiv \oint \mathbf{U} \cdot d\mathbf{l} \quad [4]$$

By convention, the circulation is evaluated by counterclockwise integration around the contour. Thus, $C > 0$ corresponds to a cyclonic circulation in the Northern Hemisphere. According to Kelvin's circulation theorem, the absolute

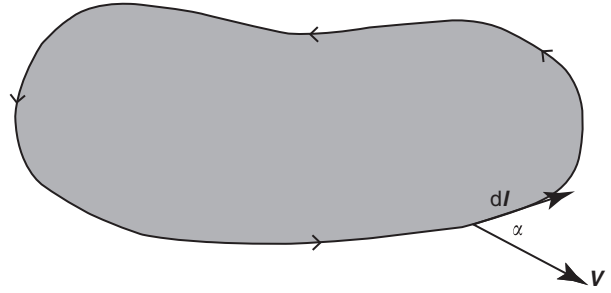


Figure 3 Circulation about an arbitrary closed contour.

circulation (computed using the velocity with respect to coordinates fixed in space) is conserved following the fluid motion for loops of fluid parcels on isobaric or isentropic surfaces. This conservation law is somewhat analogous to conservation of angular momentum in solid body mechanics.

Circulation is related to vorticity via Stokes's theorem, which states that

$$C \equiv \oint \mathbf{U} \cdot d\mathbf{l} = \iint_A (\nabla \times \mathbf{U}) \cdot \mathbf{n} dA \quad [5]$$

where A is the area enclosed by the contour, and the unit normal $\mathbf{n}$ is defined by the counterclockwise sense of the line integration using the right-hand thumb rule. Equation [5] shows that the vertical components of absolute and relative vorticity are related to the absolute and relative circulations by

$$\eta = \lim_{\delta A \rightarrow 0} \frac{C_a}{\delta A}, \quad \zeta = \lim_{\delta A \rightarrow 0} \frac{C}{\delta A} \quad [6]$$

where C_a and C are the absolute and relative circulations, respectively, given by line integrals over closed loops in the horizontal plane.

For an infinitesimal parcel of air, the Kelvin circulation theorem and eqn [6] imply that

$$\eta \delta A = (\zeta + f) \delta A = \text{Const} \quad [7]$$

If the atmosphere is approximated as a barotropic layer of depth h , in which horizontal velocity is independent of height and the density is assumed to be constant then eqn [7] leads to a simple and powerful conservation law. The arbitrary area in eqn [7] can be eliminated by observing that the mass of a column of fluid of depth h , cross-sectional area δA , and density ρ is $M = \rho h \delta A$. Since mass is conserved, this relation can be combined with eqn [7] to show that

$$\left(\frac{\eta}{h} \right) = \left(\frac{(\zeta + f)}{h} \right) = \text{Const} \quad [8]$$

Thus, following the horizontal motion of a cylinder of fluid, an increase in depth h must lead to a corresponding increase in the absolute vorticity. If latitude remains constant this implies an increase in relative vorticity. On the other hand, a decrease in depth h requires a reduction in the absolute vorticity. The conserved quantity in eqn [8] is referred to as the barotropic potential vorticity. This relation shows clearly how vertical stretching concentrates vorticity, while vertical shrinking dilutes vorticity.

Vorticity Equation

Equation [8] states that the barotropic potential vorticity of a fluid cylinder is conserved following the motion. This implies that

$$\frac{D_h}{Dt} \left(\frac{\zeta + f}{h} \right) = 0 \quad [9]$$

where the derivative is taken following the fluid motion.

If the flow is purely horizontal ($w = 0$), as must be the case for a homogeneous incompressible fluid of constant depth then eqn [9] can be simplified to obtain the barotropic vorticity equation

$$\frac{D_h(\zeta + f)}{Dt} = 0 \quad [10]$$

which states that absolute vorticity is conserved following the horizontal motion. More generally, absolute vorticity is conserved for any fluid layer in which the divergence of the horizontal wind vanishes, without the requirement that the flow be barotropic. For horizontal motion that is nondivergent ($\partial u/\partial x + \partial v/\partial y = 0$), the flow field can be completely represented by a stream function $\psi(x, y)$ defined so that lines of constant C are parallel to the velocity vector at every point, and the velocity $\mathbf{V} = \mathbf{k} \times \nabla\psi$ is proportional to the gradient of ψ . The velocity components are thus given by $u = -\partial\psi/\partial y$, $v = \partial\psi/\partial x$, and the vorticity is $\zeta = \partial v/\partial x - \partial u/\partial y = \partial^2\psi/\partial x^2 + \partial^2\psi/\partial y^2 \equiv \nabla^2\psi$. Thus, the velocity field and the vorticity can both be represented in terms of the variation of the single scalar field $\psi(x, y)$, and the barotropic vorticity equation can be written as a prognostic equation for vorticity in the form

$$\frac{\partial}{\partial t} \nabla^2\psi = -\mathbf{V} \cdot \nabla(\nabla^2\psi + f) \quad [11]$$

where $\mathbf{V} = \mathbf{k} \times \nabla\psi$ is the nondivergent horizontal wind. This equation states that the local tendency of relative vorticity is given by the advection of absolute vorticity. Equation [11] can be solved numerically to predict the evolution of the stream function, and hence of the vorticity and wind fields. Since the flow in the mid-troposphere is often nearly nondivergent on the synoptic scale, eqn [11] provides a surprisingly good model for short-term forecasts of the synoptic scale 500 hPa flow field.

Vorticity Patterns in the Atmosphere

Regions of positive ζ tend to develop in association with cyclonic storms in the Northern Hemisphere and on the poleward sides of westerly jet streams, while regions of negative ζ are associated with anticyclones and the equatorward side of westerly jet streams. (The signs are reversed in the Southern Hemisphere.) Thus, the distribution of relative vorticity is an excellent diagnostic for weather analysis. As discussed in the previous section, absolute vorticity tends to be conserved following the motion at mid-tropospheric levels; this conservation property is the basis for the simplest dynamical forecast scheme given by eqn [11].

The relationship between relative vorticity and the relative circulation C discussed above, and the roles of flow curvature and shear in contributing to vorticity can be clearly seen by

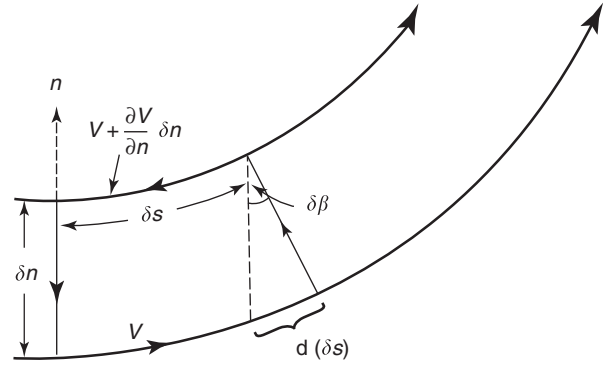


Figure 4 Circulation for an infinitesimal loop bounded by two streamlines and two lines orthogonal to the wind direction. Reproduced from Holton, J.R., 1992. *An Introduction to Dynamic Meteorology*, third ed. Academic Press, Orlando, FL.

considering the circulation about the closed contour shown in Figure 4. If the circulation is computed for the infinitesimal contour shown in Figure 4 (noting that there is no contribution along the line segments AC and BD, because these are perpendicular to the velocity), the result is

$$\delta C = V[\delta s + d(\delta s)] - \left(V + \frac{\partial V}{\partial n} \delta n \right) \delta s$$

But from Figure 4, it is seen that $d(\delta s) = \delta\beta \delta n$, where $\delta\beta$ is the angular change in the wind direction in the distance δs . Hence,

$$\delta C = \left(-\frac{\partial V}{\partial n} + V \frac{\delta\beta}{\delta s} \right) \delta n \delta s$$

or in the limit $\delta n, \delta s \rightarrow 0$

$$\zeta = \lim_{\delta n, \delta s \rightarrow 0} \frac{\delta C}{(\delta n \delta s)} = -\frac{\partial V}{\partial n} + \frac{V}{R_s} \quad [12]$$

where $R_s = (\delta\beta/\delta s)^{-1}$ is the radius of curvature of the streamlines. The net vertical vorticity component is thus the result of the sum of two parts: (1) the rate of change of wind speed normal to the direction of flow $-\partial V/\partial n$, called the shear vorticity; and (2) the turning of the wind along a streamline V/R_s , called the curvature vorticity. The contributions of shear and curvature vorticity to the pattern of vorticity on the 500 hPa isobaric surface in the atmosphere can be seen in the vorticity map of Figure 5.

Integral Constraints on Vorticity

When the influence of the divergence and convergence associated with the large-scale vertical motion is taken into account, the vorticity equation can be rewritten as

$$\frac{\partial \zeta}{\partial t} = -\nabla \cdot [(\zeta + f)\mathbf{V}] \quad [13]$$

Equation [13] expresses the remarkable fact that if frictional and diabatic effects are neglected then vorticity can be changed only by the divergence or convergence of the horizontal flux of vorticity given in brackets on the right-hand side. The vorticity cannot be changed by vertical transfer of vorticity across a horizontal surface. Integration of eqn [13] over the globe

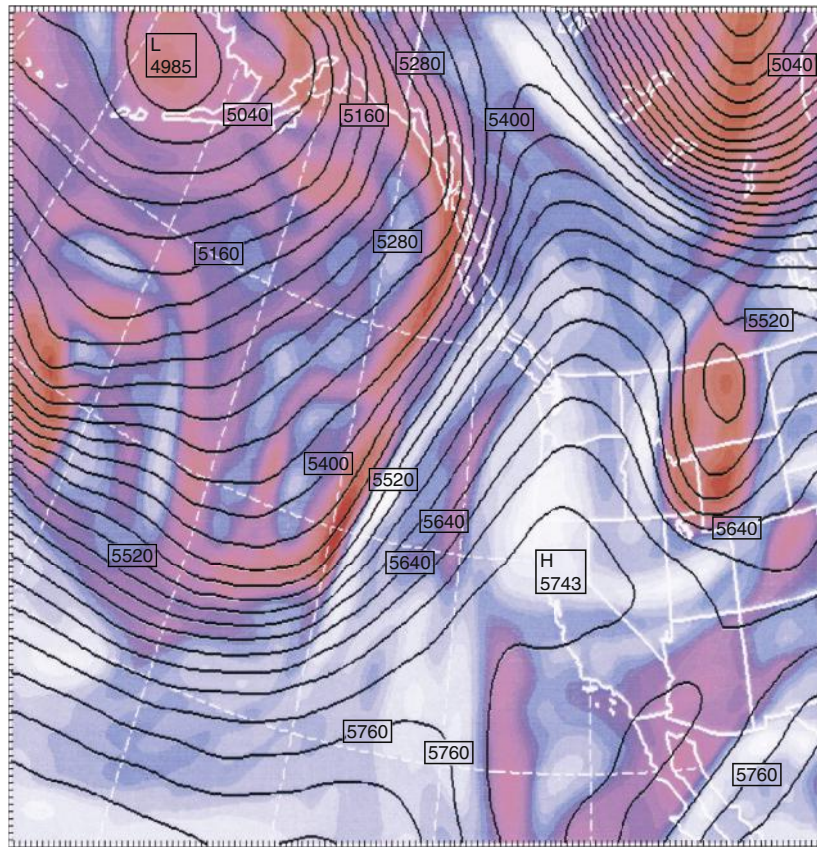


Figure 5 Vorticity (color) and geopotential height (solid contours, meters) for the 500 hPa surface in the Eastern North Pacific sector on 4 February 2002. The height contours approximate a stream function for the large-scale flow, so are approximately parallel to the winds, and where the lines are close together the wind is strong. Regions of large absolute vorticity (red) occur in conjunction with troughs in the geopotential height (e.g., over Canada), in closed cyclonic circulations (e.g., over Montana), and along the cyclonic shear side of the jet stream (e.g., the South West (SW) to North East (NE) flow over the Pacific). Negative relative vorticity (light blue) occurs in ridges and on the equatorward side of the jet stream. Courtesy of Professor Clifford Mass, Atmospheric Sciences Department, University of Washington.

shows that for a horizontal surface that does not intersect the surface of the Earth the global average of ζ is constant. Furthermore, the integration of ζ over the sphere shows that the global average ζ is exactly zero. Vorticity on such a surface is neither created nor destroyed; it is merely concentrated or diluted by the convergence or divergence of horizontal fluxes. This last result holds even when friction and diabatic effects are included.

Further Reading

Acheson, D.J., 1990. *Elementary Fluid Mechanics*. Clarendon Press, Oxford.
 Gill, A.E., 1982. *Atmosphere-Ocean Dynamics*. Academic Press, New York.
 Holton, J.R., 1992. *Introduction to Dynamic Meteorology*. Academic Press, New York.
 Salby, M.L., 1996. *Fundamentals of Atmospheric Physics*. Academic Press, New York.

See also: **Dynamical Meteorology:** Overview; Rossby Waves; Waves.

Introduction

In 1964, Charney and Eliassen introduced the term CISK (conditional instability of the second kind) to describe a positive feedback between deep moist convection and a large-scale circulation. They hypothesized that hurricane intensification was such a process, in which surface friction helps to induce low-level convergence into a vortex, resulting in deep convection and latent heating that amplify the vortex. In the late 1960s, Yamasaki and Hayashi first considered the feedbacks between deep convection and large-scale equatorial waves. In their models, convection could intensify (destabilize) the wave in some cases through purely inviscid processes not dependent on feedbacks with surface drag or surface thermodynamic fluxes. In 1974, Lindzen termed this destabilization wave-CISK, to distinguish it from Charney and Eliassen's frictional CISK, and put forth perhaps the most expansive view of the role of wave-CISK, implicating it in the development of tropical circulations from squall lines to the Madden-Julian Oscillation.

Wave-CISK has proved a somewhat slippery hypothesis to test, and has largely fallen from favor among specialists in convective dynamics. The predicted instabilities are very sensitive to the representation of cumulus convection. For simple models the fastest-growing instabilities have very short wavelengths, and are not clearly separable from conventional conditional instability of individual cumulus clouds. However, wave-CISK is a mode of instability permitted by many convective parameterizations, including some used in climate models, so it can be a useful concept in interpreting model output even if physically dubious.

The class of convective parameterizations that tend to lead to wave-CISK instabilities are those that diagnose convective mass flux based on column-integrated horizontal moisture convergence. The Arakawa-Schubert scheme, a typical 'quasi-equilibrium' closure in which the convective mass flux is chosen so as to regulate the local convectively available potential energy, does not support wave-CISK, as shown by Stark. Neither does the Betts-Miller scheme, a typical moist adjustment convective parameterization, as shown by Neelin and Yu, although Hayashi and Golder showed that CISK could be excited if the convective adjustment turns on and off frequently. In recent years, moisture-convergence-based convective parameterizations have been criticized for using a nonlocal measure (moisture convergence) to regulate a local thermodynamic process (convection), allowing unrealistic soundings to develop. Furthermore, the development of wave-CISK is strongly influenced by the vertical profile of convective heating perturbation selected by the parameterization. 'Top-heavy' heating perturbations concentrated in the upper troposphere are most favorable for wave-CISK instabilities.

A Simple Wave-CISK Model

In its simplest form, wave-CISK can be phrased in terms of a nonrotating inviscid gravity wave interacting with a simple parameterization of moist convection. Wave-induced perturbations in surface heat flux and radiative fluxes are neglected. Classically, a linear stability analysis of a small-amplitude wave is used to assess the convective feedback.

We consider the mathematical structure of an extremely simple wave-CISK model based on small-amplitude (linear) two-dimensional inviscid hydrostatic motions of a nonrotating atmosphere. The pressure velocity ω is assumed to be zero at the mean surface pressure $p = p_s$. At the tropopause pressure p_t , a similar boundary condition or a boundary condition that allows upward-propagating gravity waves to radiate out of the domain may be applied.

The mean convection is assumed to maintain the same profile of water vapor mixing ratio $q(p)$ everywhere. This is a very strong simplifying assumption on the column moisture budget, and one of the weakest links in simple wave-CISK models. In addition, most such models have not explicitly attempted to maintain a consistency between the mean temperature and moisture profiles that ensures that boundary layer air will be conditionally unstable, and hence able to convect, but will not penetrate much above the tropopause. Such consistency requires a basic state in which the moist static energy is similar at the tropopause to that in the boundary layer, which relates the assumed $q(p_s)$ to the assumed troposphere-mean static stability.

The perturbation convective heating in any column is assumed to be caused by the conversion of the converged moisture into rainfall; the resulting latent heating is redistributed through the column by turbulent convection according to a fixed vertical heating profile $\eta(p)$, which is normalized to have a mean value of unity averaged over the depth of the atmosphere. The thermodynamic equation for geopotential perturbation $\phi(x, p, t)$ is eqn [1].

$$\frac{\partial}{\partial t} \frac{\partial \phi}{\partial p} + \sigma \omega = \frac{LR}{pC_p(p_b - p_t)} \eta(p)M \quad [1]$$

In eqn [1], x is the horizontal coordinate, t is time, σ is the static stability parameter, $M(x, t)$ is the perturbation column-integrated horizontal moisture convergence, R and C_p are the gas constant and isobaric specific heat of air, and L is the latent heat of vaporization for water vapor. Using the continuity equation, the moisture convergence can be written as eqn [2].

$$M = \int_{p_t}^{p_s} q(p) \frac{\partial \omega}{\partial p} \frac{dp}{g} \quad [2]$$

Applying the horizontal momentum and continuity equations results in a linear, separable equation for $\omega(x, p, t)$ (eqn [3]).

$$\frac{\partial^2}{\partial t^2} \frac{\partial^2 \omega}{\partial p^2} + \sigma \frac{\partial^2 \omega}{\partial x^2} = \frac{LR}{pC_p(p_b - p_t)} \eta(p) \frac{\partial^2 M}{\partial x^2} \quad [3]$$

With a rigid-lid boundary condition, $\omega(p_s) = \omega(p_t) = 0$. A radiation boundary condition would relate $\omega(p_t)$ to its vertical derivative. In either case, this equation admits normal modes of the form $\omega = \hat{\omega}(p) \exp(ik[x - ct])$. In the case of constant σ , Cho and Pendlebury showed how the complex phase speed c can be calculated by a transcendental equation most easily derived by Fourier-analyzing $\hat{\omega}(p)$ and the normalized heating profile $\eta(p)$. Unstable modes ($\text{Im}(c) > 0$) exist for a variety of heating profiles. Use of a radiation boundary condition instead of a rigid lid at the tropopause usually has a minor effect on the unstable modes.

Raymond has given a simple physical description of wave-CISK by regarding the convective heating as a series of pulses, each of which acts as a vertically distributed source of gravity waves. He then showed that for a wave moving at a particular phase speed, the gravity wave generated by the pulse heating would generate low-level horizontal convergence in phase with the preexisting wave, causing the wave to amplify.

An important feature of the eigenvalue problem is that the horizontal wavenumber k cancels out of the equation. This implies that if an instability is present, shorter wavelengths will grow fastest (no short-wave cutoff), since the growth rate is proportional to $\text{Im}(kc)$. This suggests that wave-CISK might manifest itself at the shortest scale at which the model assumptions are still plausible. Since the model assumes an ensemble convective heating response, this scale would have to be somewhat larger than the spacing between convective clouds or cloud groups. However, on such scales other mechanisms such as cold-pool dynamics seems to play a more important role in organizing convection into mesoscale systems.

The short-wavelength behavior is related to the difficulty of clearly separating wave-CISK in this model from conventional conditional instability of individual cumuli. For a given temperature profile, moistening the near-surface air will render the profile more conditionally unstable to individual cumuli. Our simple CISK model exhibits a distorted version of this same mechanism, rendering dubious the notion of CISK as a distinct instability of cumulus cloud ensembles. To see this, we can integrate eqn [2] by parts as in eqn [4].

$$M = - \int_{p_t}^{p_s} \omega \frac{\partial q}{\partial p} \frac{dp}{g} \quad [4]$$

This expresses the parameterized moisture convergence, and hence the cumulus-induced heating rate, as proportional to the vertical velocity weighted by dq/dp . This heating counteracts the adiabatic cooling associated with rising motion, reducing the effective static stability of the lower and mid-troposphere. If the near-surface air is sufficiently humid, this effective static stability can become negative, promoting short-wavelength CISK instabilities. Within the framework of our CISK model, this condition of negative effective static stability plays the same role as conventional conditional instability does for growth of individual cumuli.

The stability analysis is more involved for convective parameterizations that do not employ a moisture convergence closure, and only a few such studies have been published. The complication is that the right-hand side of eqn [3], which

involves the horizontal Laplacian of the heating, is not usually expressible purely in terms of ω . Although a similar eigenvalue problem for c can often still be formulated, it usually must be solved numerically by vertical discretization and may now have a short-wave cutoff.

Elaborations on the Basic Model

Many elaborations on the above model have been proposed. In 1970, Hayashi extended a similar model to continuously stratified motions on an equatorial beta-plane using separation of variables in the meridional direction. In particular, the equatorial Kelvin wave has a similar zonal structure and growth rate to a nonrotating gravity wave. This has led to many theories that rationalize the tropical Madden-Julian (intraseasonal) oscillation as a wave-CISK mode. More sophisticated general circulation model simulations using moisture-convergence-based convective parameterizations, starting with Hayashi and Sumi in 1986, have also frequently produced intraseasonal oscillations that have been interpreted as wave-CISK.

In 1979, Davies obtained a short-wavelength cutoff by assuming a short delay between the moisture convergence and the convection. With a 30–60-minute delay, the fastest growing wavelength is a few hundred kilometers, corresponding to a typical size of a mesoscale convective system. In 1987, Lau and Peng considered ‘positive-only heating’, in which only upward motion (creating moisture convergence) produces perturbation heat release. This can be thought of as a simple way to represent waves sufficiently strong to suppress all convection in their subsiding branches, but still weak enough to be approximated by linear dynamics. Such models produce unstable modes with a propagating narrow band of ascent surrounded by a broad subsidence region. This could be considered as a parameterized representation of the circulation around a single intense cumulonimbus.

Additional physical feedbacks have been considered. In 1990 Wang and Rui considered frictional wave-CISK, the impact of surface friction and convective heating on an equatorial wave, and found that the surface drag could stimulate a pattern of convective heating that helps destabilize an equatorial Kelvin wave. Mak considered the feedback of cumulus convection (represented via eqn [2]) with an Eady model of mid-latitude baroclinic instability. He showed that the most unstable Eady mode becomes shorter and intensifies more rapidly when moderate cumulus-driven latent heating is included; these effects can be interpreted in part as consequences of a reduced effective static stability. With sufficiently strong cumulus heating, his theory also predicted new classes of boundary-trapped CISK modes. Emanuel found a baroclinic wave-CISK mode varying perpendicularly to the wind shear, somewhat akin to symmetric instability, which can be excited in a broader set of conditions than classical wave-CISK modes. These theories await decisive testing against observations and more sophisticated numerical models.

Current Status of Wave-CISK

Neither observations nor current cloud-resolving numerical model simulations clearly show classical wave-CISK-like

modes. Furthermore, the theoretical models that predict wave-CISK are based on dubious parameterizations of cumulus convection. However, there are many intriguing indications that moist convection may in fact help destabilize some large-scale waves through mechanisms not considered in classical wave-CISK. Some of these include: (i) the effect of wave-associated surface wind perturbations on the surface fluxes and boundary layer structure (wind-induced surface heat exchange, or WISHE); (ii) radiative feedbacks on the wave associated with convectively produced anvils or moisture redistribution; and (iii) feedbacks between the convection and the humidity profile in the convecting column. The importance of these convective/large-scale feedbacks in producing transient variability in the tropics and parts of the mid-latitudes on all time scales remains an active and very important topic of research.

See also: **Dynamical Meteorology:** Baroclinic Instability; Hamiltonian Dynamics; Inertial Instability; Kelvin–Helmholtz Instability; Lagrangian Dynamics; Quasigeostrophic Theory; Rossby Waves; Symmetric Stability; Vorticity. **Electricity in the Atmosphere:** Sprites.

Further Reading

- Charney, J.G., Eliassen, A., 1964. On the growth of the hurricane depression. *Journal of the Atmospheric Sciences* 21, 68–75.
- Cho, H.-R., Pendlebury, D., 1997. Wave CISK of equatorial waves and the vertical distribution of cumulus heating. *Journal of the Atmospheric Sciences* 54, 2429–2440.
- Davies, H.C., 1979. Phase-lagged wave-CISK. *Quarterly Journal of the Royal Meteorological Society* 105, 325–353.
- Emanuel, K.A., 1982. Inertial instability and mesoscale convective system. Part II: Symmetric CISK in a baroclinic flow. *Journal of the Atmospheric Sciences* 39, 1080–1092.
- Hayashi, Y., 1970. A theory of large-scale equatorial waves generated by condensation heat and accelerating the zonal wind. *Journal of the Meteorological Society of Japan* 48, 140–160.
- Hayashi, Y., Golder, D.G., 1997. United mechanisms for the generation of low- and high-frequency tropical waves. Part I: Control experiments with moist convective adjustment. *Journal of the Atmospheric Sciences* 54, 1262–1276.
- Hayashi, Y.-Y., Sumi, A., 1986. The 30–40 day oscillation simulated in an ‘aqua-planet’ model. *Journal of the Meteorological Society of Japan* 64, 451–467.
- Lau, K.-M., Peng, L., 1987. Origin of low-frequency (intraseasonal) oscillations in the tropical atmosphere. Part I: Basic theory. *Journal of the Atmospheric Sciences* 44, 950–972.
- Lindzen, R.S., 1974. Wave-CISK in the tropics. *Journal of the Atmospheric Sciences* 31, 156–179.
- Mak, M., 1994. Cyclogenesis in a conditionally unstable moist baroclinic atmosphere. *Tellus* 46A, 14–33.
- Neelin, J.D., Yu J.-Y., 1993. Modes of tropical variability under convective adjustment and the Madden–Julian oscillation. Part I: Analytical theory. *Journal of the Atmospheric Sciences* 51, 1876–1894.
- Raymond, D.J., 1983. Wave-CISK in mass-flux form. *Journal of the Atmospheric Sciences* 40, 2561–2572.
- Stark, T.E., 1976. Wave-CISK and cumulus parameterization. *Journal of the Atmospheric Sciences* 33, 2383–2391.
- Yamasaki, M., 1969. Large-scale disturbances in the conditionally unstable atmosphere in low latitudes. *Papers in Meteorology and Geophysics* 20, 289–336.
- Wang, B., Rui, H., 1990. Dynamics of the coupled moist Kelvin–Rossby wave on an equatorial beta plane. *Journal of the Atmospheric Sciences* 47, 397–413.

Wave Mean-Flow Interaction

M Jukes, University of Oxford, Oxford, UK

© 2003 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 6, pp 2526–2532, © 2003, Elsevier Ltd.

Introduction

The complexity of atmospheric flow encourages the use of simplified conceptual models to study specific processes. A conceptual model is here interpreted as a model that includes a broad range of physical features but is still sufficiently simple to permit a complete description of its properties. The theory of wave mean-flow interaction encompasses a family of such conceptual models. The flows considered generally have nontrivial dependence on all three spatial coordinates and also on time.

Some degree of simplification is achieved by assuming that the flow can be split into a wave and a mean component, such that the mean has large amplitude and is slowly evolving, while the wave has smaller amplitude and may evolve on a faster time scale and a shorter length scale.

The theory is based on the assumption the mean flow depends only on a reduced number of coordinates. It is further assumed that the mean-flow variation with respect to one of the remaining coordinates is weak. The mean could be taken over time or space, but the discussion below will concentrate on an example for which a spatial mean is used and the system has multiple scales of variation in time.

Superimposed on the mean flow is a wave field. The waves depend on all four coordinates but are often assumed to have small amplitude.

The concept of multiple scales is important here. Atmospheric dynamics is often analysed by assuming that every component of the flow varies according to the same set of representative scales. In the theory of wave mean-flow interaction that assumption is relaxed slightly by allowing different scales for different components of the flow. The two components are a mean flow and a wave field. These interact through the nonlinearity of the governing equations. The time scale for the evolution of the mean flow is much longer than that which characterizes the wave propagation. In many applications there is also a faster time characterizing the period of individual waves.

Zonal Averaging and Multiple Time Scales

The theory of wave mean-flow interaction can be applied to a wide range of flows. Here the general principles will be illustrated through a discussion of a single example, described later. This section introduces the relevant equations and averaging procedure. The complete set of equations for this problem consists of the thermodynamic equation, the horizontal momentum equations, mass conservation, and the hydrostatic equation. The hydrostatic approximation has the effect of reducing the system from five to three predictive equations by filtering out sound waves. This approximation is valid when the typical velocities of the flow are small compared

to the phase speed of sound waves. These equations may be written as [1], [2], [3], [4] and [5].

$$\text{Thermodynamic} \quad \theta_t + u\theta_x + v\theta_y + w\theta_z = Q \quad [1]$$

$$\begin{aligned} \text{x - Momentum} \\ u_t + uu_x + vv_y + ww_z - fv = \Phi_x + F \end{aligned} \quad [2]$$

$$\begin{aligned} \text{y - Momentum} \\ v_t + uv_x + vv_y + ww_z + fu = \Phi_y + G \end{aligned} \quad [3]$$

$$\text{Mass} \quad u_x + v_y + \rho_s^{-1}(w\rho_s)_z = 0 \quad [4]$$

$$\text{Hydrostatic} \quad \theta = \frac{\theta_{\text{ref}}}{g} \Phi_z \quad [5]$$

In these equations, z is the log pressure vertical coordinate (see **Dynamical Meteorology: Overview**) and $\rho_s(Z) = \rho_s(0) \exp(-z/H)$ is the corresponding pseudodensity (H is a reference scale height). (u, v, w) are the three components of the wind, θ is potential temperature, and Φ is the geopotential height. Q represents diabatic heating and F and G represent diffusion terms. In numerical or conceptual models F and G can also include the effects of small-scale turbulent mixing that is not explicitly represented.

We now introduce a small parameter ε that will represent the amplitude of the waves. The associated changes to the mean flow will occur on a time scale proportional to ε^{-2} . To represent this mathematically, we introduce a fast time $\tau = t$ and a slow time $T = \varepsilon^{-2}t$. The variables u, v, w, θ , and Φ are then all expanded in powers of ε as follows in eqn [6a].

$$\begin{aligned} u(x, y, z, t) = & u^{(0)}(y, z, T) \\ & + \varepsilon u^{(1)}(x, y, z, \tau, T) \\ & + \varepsilon^2 u^{(2)}(x, y, z, \tau, T) + \dots \end{aligned} \quad [6a]$$

Note that the leading-order term is independent both of the zonal coordinate x and of the fast time τ . It follows from the definition of the fast and slow times that temporal derivatives can be expanded as in eqn [6b].

$$\frac{\partial}{\partial t} = \frac{\partial}{\partial \tau} + \varepsilon^2 \frac{\partial}{\partial T} \quad [6b]$$

With the introduction of two time variables to replace a single one, the problem is rendered underspecified. That is, we have increased the number of variables that need to be determined without having a corresponding increase in the number of equations we can use to determine these variables. This property will be turned to advantage later on, when it is used to impose conditions that facilitate the clean separation of different parts of the problem along physically meaningful lines.

The leading order equations are [7], [8], [9], [10] and [11].

$$v^{(0)}\theta_y^{(0)} + w^{(0)}\theta_z^{(0)} = Q^{(0)} \quad [7]$$

$$v^{(0)} \left[u_y^{(0)} - f \right] + w^{(0)} u_z^{(0)} = F^{(0)} \quad [8]$$

$$v^{(0)} v_y^{(0)} + w^{(0)} v_z^{(0)} + f u^{(0)} = \Phi_y^{(0)} + G^{(0)} \quad [9]$$

$$v_y^{(0)} + \rho_s^{-1} \left(w^{(0)} \rho_s \right)_z = 0 \quad [10]$$

$$\theta^{(0)} = \frac{\theta_{\text{ref}}}{g} \Phi_z^{(0)} \quad [11]$$

It will be assumed that the leading-order solution is unforced, so that $Q^{(0)}$, $F^{(0)}$, and $G^{(0)}$ all vanish. The first and second equations imply, in general, $v^{(0)}, w^{(0)} = 0$. Equation [10] is then satisfied automatically. The remaining two equations then show that the zonal flow is in hydrostatic and geostrophic balance. These two can be combined to form the thermal wind eqn [12].

$$\theta_y^{(0)} = g^{-1} \theta_{\text{ref}} f u_z^{(0)} \quad [12]$$

The wave equations, at order ε , are [13], [14], [15], [16] and [17].

$$\theta_t^{(1)} + u^{(0)} \theta_x^{(0)} + v^{(1)} \theta_y^{(0)} + w^{(1)} \theta_z^{(0)} = Q^{(1)} \quad [13]$$

$$u_t^{(1)} + u^{(0)} u_x^{(0)} + v^{(1)} u_y^{(0)} + w^{(1)} u_z^{(0)} - f v^{(1)} = \Phi_x^{(1)} + F^{(1)} \quad [14]$$

$$v_t^{(1)} + u^{(0)} v_x^{(1)} + f u^{(1)} = \Phi_y^{(1)} + G^{(1)} \quad [15]$$

$$u_x^{(1)} + v_y^{(1)} + \frac{1}{\rho_s} \left(w^{(1)} \rho_s \right)_z = 0 \quad [16]$$

$$\theta^{(1)} = \frac{\theta_{\text{ref}}}{g} \Phi_z^{(1)} \quad [17]$$

These five equations describe a wide variety of wave propagation mechanisms, including gravity waves, Rossby waves, inertial waves, Kelvin waves and mixed Rossby gravity waves.

The equations are invariant with respect to the x coordinate and it is known that the underlying physics can be expressed in a Hamiltonian formulation. For Hamiltonian systems there is a general theorem, due to Emmy Noether, stating that there is an invariant related to every symmetry. The invariant that is associated with the x -symmetry of the underlying physics is momentum. The momentum-like invariant that is associated with the x -symmetry of the perturbation equations is known as the pseudomomentum and takes the form shown in eqn [18a].

$$A = -\rho_s \frac{\overline{u_z^{(1)} \theta^{(1)}}}{\overline{\theta_z^{(0)}}} + \left(\frac{\rho_s}{\overline{\theta_z^{(0)}}} \right)^2 \left(\frac{\overline{u_z^{(0)}}}{\rho_s} \right) \frac{\overline{\theta^{(1)^2}}}{2} + \frac{1}{2} \left(\frac{\rho_s}{\overline{\theta_z^{(0)}}} \right)^2 \frac{\overline{(q^{(1)} \theta_z^{(0)} - \theta^{(1)} q_z^{(0)})}}{q_y^{(0)} \theta_z^{(0)} - \theta_y^{(0)} q_z^{(0)}} \quad [18a]$$

In eqn [18], q is Ertel's potential vorticity, given by eqn [18b].

$$q = \frac{(f + v_x - u_y) \theta_z + u_z \theta_y - v_z \theta_x}{\rho_s} \quad [18b]$$

The bracketed group in the numerator of the last term on the right-hand side of [18a] is related to the isentropic anomaly of q . The isentropic anomaly is of greater dynamical significance than the isobaric anomaly $q^{(1)}$ because q is materially conserved on isentropes in adiabatic, inviscid flow. A is also known as the 'wave activity'.

The expression for wave activity [18a] is complicated, but this complexity is compensated for by a relatively simple evolution equation (eqn [19], where $\nabla = (\partial/\partial y, \partial/\partial z)$).

$$A_t + \nabla \cdot \mathbf{F}^{\text{ep}} = D. \quad [19]$$

The right-hand side, D , is a dissipative term depending linearly on $Q^{(1)}$, $F^{(1)}$, and $G^{(1)}$. The conservative wave dynamics are described by the Eliassen-Palm flux vector $\mathbf{F}^{\text{ep}}$ whose components are given by eqns [20] and [21].

$$F_y^{\text{ep}} = \rho_s \left(\overline{-u^{(1)} v^{(1)}} + \frac{\overline{\theta^{(1)} v^{(1)}}}{\overline{\theta_z^{(0)}}} u_z^{(0)} \right) \quad [20]$$

$$F_z^{\text{ep}} = -\rho_s \left(\overline{u^{(1)} w^{(1)}} + \frac{\overline{\theta^{(1)} v^{(1)}}}{\overline{\theta_z^{(0)}}} \left[u_y^{(0)} - f \right] \right) \quad [21]$$

The evolution equation for the wave activity [19] follows from the linear wave equations [13], [14], [15], [16] and [17] without any further approximation. The unwieldy definition of A is compensated for by the simple structure of the evolution equation. While it is easy enough to verify that A has the properties listed, it would not be easy to find such a quantity without the guidance that comes from Noether's theorem. The existence of such an exact form is important in guaranteeing the validity of the approximations made to the wave equations below. This would not be a problem if we were only concerned with a local approximation, but the problem here requires an approximation to be valid over many vertical wavelengths and many wave periods.

A further useful property of the wave action is that, under circumstances in which a group velocity c_g is well defined, the flux of wave activity is given by $F^{\text{ep}} = c_g A$.

Mean-Flow Evolution

The equation for the slow evolution of the mean flow is obtained from the order ε^2 terms in the expansion of the governing equations. The zonal-mean zonal-momentum and thermodynamic equations take following forms [22] and [23].

$$u_T^{(0)} - \overline{v^{(2)}} \left[f - u_y^{(0)} \right] + \overline{w^{(2)}} u_z^{(0)} = - \left(\overline{u^{(1)} v^{(1)}} \right)_y - \rho_s^{-1} \left(\rho_s \overline{u^{(1)} w^{(1)}} \right)_z \quad [22]$$

$$\theta_T^{(0)} + \overline{v^{(2)}} \theta_y^{(0)} + \overline{w^{(2)}} \theta_z^{(0)} = - \left(\overline{\theta^{(1)} v^{(1)}} \right)_y - \rho_s^{-1} \left(\rho_s \overline{\theta^{(1)} w^{(1)}} \right)_z + \overline{Q^{(2)}} \quad [23]$$

It was mentioned earlier that the introduction of two time variables made the problem underspecified. The associated freedom has been used here to impose the conditions $\overline{u^{(2)}} = 0$ and $\overline{\theta^{(2)}} = 0$, so that all changes in the mean zonal flow and

potential temperature are included in the leading-order terms $\overline{u^{(0)}}$ and $\overline{\theta^{(0)}}$. Equations [22] and [23] show that the mean flow is forced by the convergence of the Reynolds' stress and the mean potential temperature is changed by the convergence of the heat flux. The system can respond to forcing through two physically distinct mechanisms: the tendency (represented by $u_T^{(0)}$ and $\theta_T^{(0)}$) and the mean meridional circulation (represented by $\overline{v^{(2)}}$, $\overline{w^{(2)}}$).

The analysis in the previous section showed that the eddy fluxes that occur in [22] and [23] as forcings of the mean flow also occur in [20] and [21] representing the flux of wave activity. A transformation to what are known as the transformed Eulerian mean equations clarifies the link between mean-flow forcing and the wave propagation. The aim of the transformation is to reorganize the equations so that all the forcing terms can be collected into a single expression. This achieved by defining the residual circulation as in eqns [24] and [25].

$$\overline{v^*} = \overline{v^{(2)}} - \frac{1}{\rho_s} \frac{\partial}{\partial z} \left(\frac{\rho_s \overline{v^{(1)} \theta^{(1)}}}{\theta_z^{(0)}} \right) \quad [24]$$

$$\overline{w^*} = \overline{w^{(2)}} + \frac{\partial}{\partial y} \left(\frac{\overline{v^{(1)} \theta^{(1)}}}{\theta_z^{(0)}} \right) \quad [25]$$

Modified means are defined for the potential temperature [26] and heating rate [27].

$$\overline{\theta^*} = \overline{\theta^{(0)}} + \varepsilon^2 \frac{\partial}{\partial z} \left(\frac{\overline{\theta^{(1)2}}}{\theta_z^{(0)}} \right) \quad [26]$$

$$\overline{Q^*} = \overline{Q^{(2)}} + \frac{\partial}{\partial z} \left(\frac{\overline{\theta^{(1)} Q^{(1)}}}{\theta_z^{(0)}} \right) \quad [27]$$

The resulting transformed Eulerian mean equations are [28a] and [28b] where terms that are fourth-order in wave amplitude have been neglected. The transformed Eulerian mean equations [28] gain

$$u_T^{(0)} - \overline{v^*} [f - u_y^{(0)}] + \overline{w^*} u_z^{(0)} = \nabla \cdot \mathbf{F}^{\text{ep}} \quad [28a]$$

$$\overline{\theta_T^*} + \overline{v^*} \overline{\theta_y^*} + \overline{w^*} \overline{\theta_z^*} = \overline{Q^*} \quad [28b]$$

their significance not only from the fact that the eddy forcing terms have been collected into a single term but also from the fact that this single term, $\nabla \cdot \mathbf{F}^{\text{ep}}$, is the same as the term describing the evolution of the pseudomomentum in [19]. There is thus a clear link between the wave forcing of the mean flow and the wave propagation. The solutions of eqns [28a] and [28b] are constrained by the thermal wind equations [12] for the zonal flow and by mass conservation in the residual circulation. This in turn means that the eddy forcing and the thermal forcings ($\nabla \cdot \mathbf{F}^{\text{ep}}$ and $\overline{Q^*}$ respectively) cannot, in general, be independently specified. When considering wave-driven circulations, the thermal term may be considered as a relaxation term, pulling the atmosphere back toward a thermal equilibrium on a time scale T_Q , say. In other contexts

it may make more sense to consider the thermal term as forcing, rather than responding.

The system still has two modes of response: zonal acceleration or meridional circulation. The relative magnitude of these two modes can be analyzed by eliminating the meridional circulation from [28a] and [28b] using the equation of mass conservation ($\overline{v_y^*} + \rho_s^{-1} (\rho_s \overline{w^*})_z = 0$), and expressing the tendencies in terms of a potential ($u_T^{(0)}, \theta_T^{(0)} = (\chi_y, -\theta_{\text{ref}} f g^{-1} \chi_z)$) (which is possible because of the thermal wind equation). An elliptical equation for χ can then be derived. Analysis of this equation reveals that the nature of the response depends on the nondimensional parameter $C_F = C_T / (1+B-1)$, where $C_T = T_F / T_Q$ is the ratio of times scales. (T_F is the time scale of the forcing) and $B = L_F^2 f^2 / (H_F^2 N^2)$ is the Burgers number (N is the Brunt-Väisälä frequency; L_F, H_F are the horizontal and vertical length scales of the eddy forcing). If C_F is large, the response will tend to be in the meridional circulation; when C_F is small, the response will be mainly in the zonal acceleration. Small C_F may result from either a small ratio of time scales or a small Burger number. If, for example, we take parameter values representative of large-scale seasonally varying forcing ($L_F = 10^6$ m, $H_F = 8 \times 10^3$ m, $T_F = 8 \times 10^6$ s, $T_Q = 8 \times 10^5$ s, and $N^2 = 4 \times 10^{-4}$ s⁻¹) then taking a tropical value of the Coriolis parameter ($f = 10^{-5}$ s⁻¹) gives $C_F \approx 0.04$, whereas taking a mid-latitude value ($f = 10^{-4}$ s⁻¹) gives $C_F \approx 4$.

Thus, for eddy forcing on seasonal times scale the mid-latitude response should be dominated by the meridional circulation, whereas the tropical response will be dominated by zonal accelerations. The next section analyses the latter scenario in more detail.

The Quasi-Biennial Oscillation

As the title 'wave mean-flow interaction' implies, it is important not only to understand how the waves modify the mean flow but also to quantify how the mean flow affects the waves. The latter component of the interaction is illustrated beautifully by the cyclic variation in the zonal winds of the tropical stratosphere known as the quasi-biennial oscillation (QBO; see **Middle Atmosphere: Quasi-Biennial Oscillation**).

The QBO is a wave-forced oscillation with an irregular period slightly longer than 2 years. The atmospheric phenomenon is described in more detail in a separate article. The QBO is forced by a variety of different waves. Here the focus will be on a simple conceptual model forced by a single class of waves that are thought to play an important role. To further simplify the exposition, the meridional structure will be neglected and the Coriolis parameter will be taken as zero. These simplifications eliminate the Rossby, Rossby-gravity, inertial, and Kelvin waves from the system, leaving only gravity waves.

In this case there is a very good separation of scales. The mean flow varies with a period of slightly over 2 years, whereas the time scale for the propagation of waves through the stratosphere is a matter of hours or days, depending on the type of waves in question.

Since the wave equations include a z -dependent coefficient, $u^{(0)}$, they do not have simple sinusoidal solutions. However, if the vertical scale of the variation is large compared to the

vertical wavelength of the waves, this variation can be neglected to leading order. This approximation can be made rigorous using the Liouville–Green method, but a detailed analysis of the approximations made will be omitted here to preserve the focus on wave mean-flow interaction. To deal with this rigorously, it is necessary to introduce a scale separation between the wavelength of the eddies and the length scale of variations in the mean flow.

If the vertical variation of $u^{(0)}$ is neglected these equations have solutions of the form [29], where $\gamma = 1/(2H)$ and $[\tilde{\theta}, \tilde{u}, \tilde{w}, \tilde{\Phi}]$ are complex constants.

$$\begin{aligned} [\theta^{(1)}, u^{(1)}, w^{(1)}, \Phi^{(1)}] &= \text{Re}\{[\tilde{\theta}, \tilde{u}, \tilde{w}, \tilde{\Phi}] \\ &\times \exp(i(kx + mz - \omega t) + \gamma z)\} \end{aligned} \quad [29]$$

If [29] is substituted into the linear eqns [13], [14], [15], [16] and [17], it is found that these constants must satisfy the matrix equation [30].

$$\begin{pmatrix} i(\omega - u^{(0)}k) & 0 & \theta_z^{(0)} & 0 \\ 0 & i(\omega - u^{(0)}k) & 0 & -ik \\ 0 & ik & im - \gamma & 0 \\ 1 & 0 & 0 & (im + \gamma)\theta_0 g^{-1} \end{pmatrix} \begin{pmatrix} \tilde{\theta} \\ \tilde{u} \\ \tilde{w} \\ \tilde{\Phi} \end{pmatrix} = 0 \quad [30]$$

Equation [30] only has nonzero solutions if the determinant of the matrix vanishes. This condition leads to the dispersion relation [31] relating the frequency ω to the wave vector (k, m) .

$$\omega = ku^{(0)} \pm \frac{kN}{\sqrt{m^2 + \gamma^2}} \quad [31]$$

The corresponding group velocity is given by eqn [32].

$$c_{gz} = \frac{\partial \omega}{\partial m} = \frac{kmN}{(m^2 + \gamma^2)^{3/2}} \quad [32]$$

If such a plane wave propagates into a region of different $u^{(0)}$, then [31] must continue to be fulfilled. The frequency and zonal wavenumber are fixed by the boundary conditions, so the only means that the wave has of maintaining [31] is by adjusting the vertical wavenumber m . Rearranging [31] gives eqn [33].

$$m^2 = \left(\frac{N}{c - u^{(0)}} \right)^2 - \gamma^2 \quad [33]$$

As $c - u^{(0)}$ decreases, m^2 must increase. This will have two effects: The speed of vertical propagation will decrease and the rate at which the wave is dissipated by thermal damping will increase.

With the neglect of the meridional structure, the wave activity eqns [18a], [19], [20] and [21] are greatly simplified:

$$A_t - \left(\overline{\rho_s w^{(1)} u^{(1)}} \right)_z = D \quad [34]$$

Equation [34] shows that steady ($A_t = 0$) conservative ($D = 0$) waves have vanishing Reynolds stress. In other words,

there is no wave mean-flow interaction. Nonconservative processes could include radiation and nonlinear processes associated with wave breaking. In keeping with the policy of taking the simplest representation necessary to illustrate wave mean-flow interaction, these nonconservative processes will be represented by a simple linear damping on the wave activity.

The wave field is most easily represented as a superposition of weakly interacting waves, each identified by a wavenumber and frequency. The orthogonality of the different waves then implies that the Reynolds stress and the wave activity can both be expressed as the sum of the contributions from individual waves. Let $\tilde{A}(k, \omega; z)$ be the wave activity associated with wavenumber k and frequency ω , and suppose that the corresponding linear damping rate is $\alpha(k, \omega)$. For steady conditions (that is, with no variation on the short time scale) and neglecting order ε^4 term associated with slow time scale variations, the wave activity equation becomes [35].

$$\frac{\partial}{\partial z} \left[\overline{\rho_s w^{(1)} u^{(1)}} \right] = \sum_{k, \omega} \alpha \tilde{A}(k, \omega; z) \quad [35]$$

As mentioned above, the z dependence of A is assumed to take place on a scale much larger than the vertical wavelength.

With $u_z^{(0)} = 0$ and $q^{(0)} = q^{(1)} = 0$, the contribution to the pseudomomentum from a single wave component is given by eqn [36], where the $\dagger$ indicates a complex conjugate.

$$\tilde{A}(k, \omega; z) = -\text{Re}[(im + \gamma)\tilde{u}\tilde{\theta}^\dagger] \frac{\rho_s}{\theta_z^{(0)}} \quad [36]$$

The Reynolds stress can also be expressed as a sum of different wavenumber and frequency contributions, and it can be shown that eqn [37] holds.

$$\begin{aligned} \overline{w^{(1)} u^{(1)}} \rho_s &= \sum_{k, \omega} \text{Re}[\tilde{u}\tilde{w}^\dagger] \rho_s \\ &= -\sum_{k, \omega} c_{gz} \tilde{A}(k, \omega; z) \end{aligned} \quad [37]$$

Hence, given the independence of the propagation of different wave components, eqn [35] becomes eqn [38].

$$\frac{\partial}{\partial z} [c_{gz} \tilde{A}(k, \omega; z)] = -\alpha \tilde{A}(k, \omega; z) \quad [38]$$

Integrating with respect to z gives the vertical structure of the wave activity (eqn [39]).

$$\begin{aligned} \tilde{A}(k, \omega; z) &= \frac{c_{gz}(0) \tilde{A}(k, \omega; 0)}{c_{gz}(z)} \\ &\times \exp\left(-\int_0^z \frac{\alpha}{c_{gz}(z')} dz'\right) \end{aligned} \quad [39]$$

The corresponding slow variation in the mean flow is given by eqn [40].

$$\rho_s u_T^{(0)} = -\sum_{k, \omega} \alpha \tilde{A}(k, \omega; z) \quad [40]$$

In general, the dissipation rate α will depend both on the wavenumber and the amplitude of the waves. Figure 1 shows sketches of five different types of behavior, depending both on the ratio of the dissipation time scale to the vertical propagation time scale and on the type of dissipation. In I and II the dissipation rate is constant. The key factor then is whether the

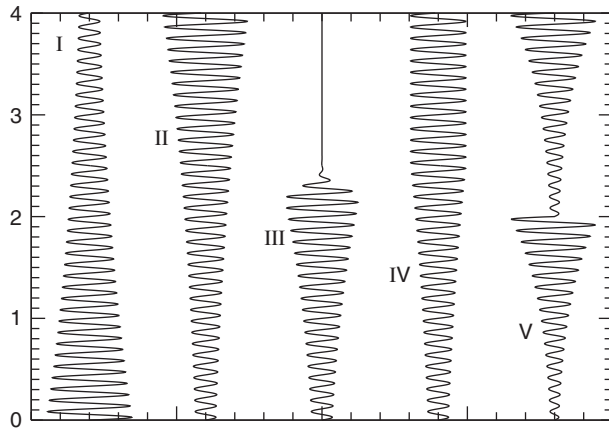


Figure 1 Different paradigms of dissipating vertically propagating waves. I, overdispersed; II, underdispersed; III, critical level; IV, continuous breaking; and V, catastrophic breaking.

growth of the wave due to the density effect dominates over the dissipation. For $2\alpha H > c_{gz}$ the wave decays with height (I), otherwise it grows (II). Another possibility is that the wave encounters a critical line (where the phase speed matches the mean zonal wind, $u^{(0)}$). In this case α is peaked near the critical line and there is no propagation above the critical line (III). Finally, the growth in amplitude with height can lead to wave breaking. This can either be continuous, where the value of α is such as to keep the amplitude constant with height (IV), or catastrophic, in which episodes of wave breaking drastically reduce the wave amplitude (V).

In the equatorial stratosphere, Rossby-gravity waves generally fit into category I, while Kelvin waves and gravity waves are more likely to be in III–V.

In order to focus on wave mean-flow interaction, the simplest form of dissipation will be assumed here: $\alpha = \text{constant}$. It will further be assumed that the criticality parameter $2\alpha H c_{gz}^{-1}$ is close to unity, so that the wave amplitude does not vary greatly throughout the domain. This situation lies between cases I and II in Figure 1.

Figure 2 shows the acceleration that will result from an initially sinusoidal vertical structure in $u^{(0)}$ (the heavy solid line) when the wave field consists of two vertically propagating waves with equal and opposite zonal phase speeds, $c = \pm c_0$. The dot-dashed lines show how the wave activity would decay with height in the absence of any mean flow. The thin solid lines show the actual decay with height. There is a relative enhancement of the $c > 0$ wave in the region $u^{(0)} > 0$ and of the $c < 0$ wave in the region $u^{(0)} < 0$. Taking the sum of these, and rescaling, gives the heavy dashed line. The forcing of the mean flow is proportional to this line. There is a clear tendency to accelerate the mean flow and also to displace the flow downward.

Summary

The theory of wave mean-flow interaction rests on a division of the flow into a large-amplitude, slowly evolving mean and a small-amplitude, rapidly evolving wave. Despite their small amplitude, the waves are able to force significant changes to the

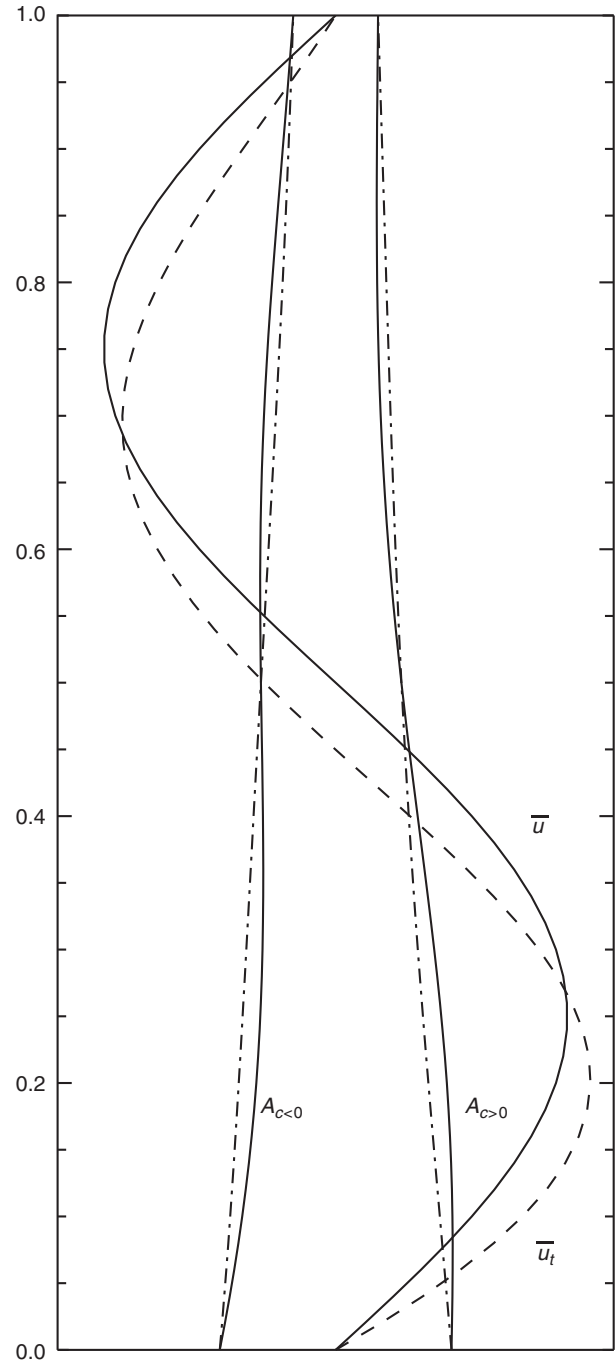


Figure 2 Idealized forcing generated by gravity waves propagating through a sinusoidally varying basic state (see discussion in text).

mean flow on the longer time scale relevant for mean-flow evolution. The time scale separation was used to derive one set of equations describing the wave propagation and another set describing the mean flow changes forced by the waves.

The pseudomomentum is introduced to describe the wave evolution over time and space scales much greater than the typical wave scales. The pseudomomentum gives a phase-independent measure of the wave amplitude without introducing any approximations. This allows subtle effects arising

from the slight, mean flow-induced, variations in wave propagation and absorption properties to be described accurately and concisely.

It was then shown how a transformation of the mean flow evolution equations allows the eddy forcing terms they contain to be related to the evolution of the pseudomomentum.

The way in which the mean-flow modification by the waves and the modulation of wave propagation by the mean flow can feedback on each other was illustrated with a simple model of the quasi-biennial oscillation, forced by gravity waves.

The amplification of zonal flow anomalies due to gravity wave forcing can arise because a positive zonal flow anomaly slows down the vertical propagation of those gravity waves with a positive phase velocity. The slower vertical propagation leads to a stronger interaction within a given height range. In the case of gravity waves, this interaction tends to drive the zonal flow toward the gravity wave phase speed. The end result is that the positive flow anomaly will be reinforced through a stronger interaction with gravity waves having a positive phase velocity.

The downward propagation of the zonal flow anomalies depends on the details of the selective wave absorption. If there is strong absorption, the amplitude of the positive phase speed waves will decrease with height through the region of positive winds. This means that a given flow anomaly will be reinforced more strongly below its maximum than above. The displacement of the forcing relative to the anomaly causes the anomaly to drift downward. The fact that waves of a positive phase velocity are being preferentially absorbed in a positive flow anomaly means that higher up there will be a deficit relative to the negative phase velocity waves. This implies that there will be a net negative forcing at higher levels, tending to generate negative flow anomalies above positive ones.

The negative anomalies will then drift downward, so that a repeating pattern of alternating and descending mean flow anomalies is generated.

Wave mean-flow interaction is important in many other areas of atmospheric dynamics, such as sudden warmings in the winter polar stratosphere, in the maintenance of the jet streams in the extratropical upper troposphere, and in the formation of blocks at the end of the mid-latitude storm tracks. The theory makes it possible to analyze these complex flows in terms of the propagation of small-amplitude waves and their feedback on the mean flow. The wave activity formalism allows the wave propagation to be described in terms of the group velocity and provides a direct link between that propagation and induced changes in the mean flow.

See also: **Dynamical Meteorology:** Overview; **Waves.** **Gravity Waves:** Overview. **Middle Atmosphere:** Quasi-Biennial Oscillation; Stratospheric Sudden Warmings. **Tropical Meteorology and Climate:** Equatorial Waves.

Further Reading

- Andrews, D.G., 1987. On the interpretation of the Eliassen–Palm flux divergence. *Quarterly Journal of the Royal Meteorological Society* 113, 323–338.
- Andrews, D.G., McIntyre, M.E., 1976. Planetary waves in horizontal and vertical shear: the generalized Eliassen–Palm relation and the mean zonal acceleration. *Journal of the Atmospheric Sciences* 33, 2031–2048.
- Andrews, D.G., Holton, J.R., Leovy, C.B., 1987. *Middle Atmosphere Dynamics*. In: *International Geophysics Series*, 40. Academic Press, London.
- Hinch, E.J., 1991. *Perturbation Methods*. Cambridge University Press, Cambridge.
- Olver, F.W.J., 1974. *Introduction to Asymptotics and Special Functions*. Academic Press, London.
- Pedlosky, J., 1987. *Geophysical Fluid Dynamics*. Springer-Verlag, Berlin.

Waves

JR Holton, University of Washington, Seattle, WA, USA

© 2003 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 2, pp 703–711, © 2003, Elsevier Ltd.

Introduction

Waves in the atmosphere are oscillatory motions that result from a balance between the inertia of the atmosphere and a restoring force. Atmospheric waves are important because they can transmit energy and momentum without material transport of air parcels. Most weather disturbances are associated with one or more types of atmospheric wave. Owing to wave propagation, disturbances in one region can influence the weather in a remote region on time scales much shorter than the transit time for air parcels. In order to predict the weather or to simulate the climate it is necessary to properly account for the influence of wave propagation. Examples of important atmospheric waves are buoyancy waves, inertia-gravity waves, and Rossby waves. In buoyancy waves the vertical stability of the atmosphere acts as a restoring force; in inertia-gravity waves the Coriolis force and buoyancy force both act as restoring forces; and in Rossby waves a fluid analog of spin angular momentum (potential vorticity) acts as the restoring force.

Phase Velocity and Group Velocity

The simplest atmospheric waves are sinusoidal oscillations that satisfy linearized forms of the dynamical equations. The dynamical equations are linearized by assuming that all dependent field variables can be divided into two parts, a basic state portion, which is usually assumed to be an east–west flow independent of time and longitude, and a disturbance portion, which represents an oscillatory deviation of the field from its basic state and is assumed to be sufficiently small so that terms quadratic in the disturbance variables can be neglected. Thus, for example, if $\bar{u}$ designates a time- and longitude-averaged zonal velocity, u' is the deviation from that average, and x and t designate the zonal coordinate and time, respectively, then the complete zonal velocity field is $u(x, t) = \bar{u} + u'(x, t)$ and the inertial acceleration $u\partial u/\partial x$ can be written as eqn [1].

$$u \frac{\partial u}{\partial x} = (\bar{u} + u') \frac{\partial}{\partial x} (\bar{u} + u') \approx \bar{u} \frac{\partial u'}{\partial x} \quad [1]$$

When the dynamical equations are linearized in this fashion, and the basic state flow is sufficiently simple, it is often possible to find analytic solutions in the form of sinusoidal waves. For example, a sinusoidal wave propagating in the zonal direction can be represented by eqn [2], where $|C|$ is the amplitude and ϕ , the argument of the complex exponential, is the phase.

$$f(x, t) = \text{Re}[C \exp(i\phi)] \quad [2]$$

In a propagating wave, phase depends on time and space. Thus, for a one-dimensional wave propagating in the x

direction, the wave phase can be represented by $\phi(x, t) = kx - \nu t$. Here the wave number, k , is defined as 2π divided by the wavelength, L_x , and frequency, ν , is 2π divided by the wave period. For zonally propagating waves, the phase is constant for an observer moving at the zonal phase speed $c_x = \nu/k$. This may be verified by observing that if phase is to remain constant following the motion, eqn [3] must be satisfied.

$$\frac{D\phi}{Dt} = \frac{D}{Dt} (kx - \nu t) = k \frac{Dx}{Dt} - \nu = 0 \quad [3]$$

where D/Dt designates differentiation following the motion. Thus, $(Dx/Dt)\phi=\text{const} = c_x = \nu/k$. By convention it is assumed that $k > 0$; then for $\nu > 0$, $c_x > 0$, so that phase propagates in the positive direction, while for $\nu < 0$ phase propagates in the negative direction.

In some waves (for example, acoustic waves) parcel oscillations are parallel to the direction of phase propagation. However, most meteorologically important waves are transverse waves in which the parcel oscillations are perpendicular to the direction of phase propagation.

For propagating waves, frequency generally depends on the wave number of the perturbation as well as the physical properties of the basic state. Thus, since $c_x = \nu/k$, the phase speed also depends on the wave number except in the special case where $\nu \propto k$. For waves in which the phase speed varies with k , the various sinusoidal components of a disturbance originating at a given location are at a later time found in different places. Such waves are referred to as dispersive, and the formula that relates ν and k is called a dispersion relationship. Some types of waves, such as acoustic waves, have phase speeds that are independent of the wave number. In such nondispersive waves, a spatially localized disturbance consisting of a number of Fourier wave components (a wave group) will preserve its shape as it propagates in space at the phase speed of the wave. For dispersive waves, however, the shape of a wave group will not remain constant as the group propagates. Since the individual Fourier components of a wave group may either reinforce or cancel each other, depending on the relative phases of the components, the energy of the group will be concentrated in limited regions.

When waves are dispersive, the speed of the wave group is generally different from the average phase speed of the individual Fourier components. Hence, as shown in Figure 1, individual wave components may move through the group as the group propagates in space. Furthermore, the group generally broadens in the course of time, that is, the energy is dispersed. A familiar example is the wake of a ship, in which individual wave crests are observed to move twice as fast as the wave group.

An expression for the group velocity, which is the velocity at which the observable disturbance (and hence the energy) propagates, can be derived by considering the superposition of

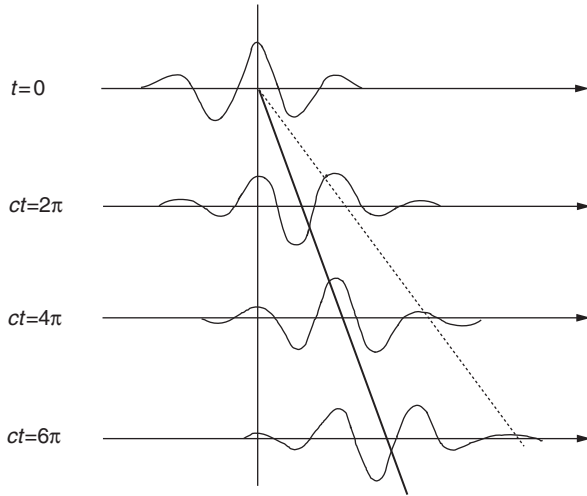


Figure 1 Propagation of a wave group. Heavy line shows group speed; dashed line shows phase speed. Reproduced from Holton, J.R., 1992. *Introduction to Dynamic Meteorology*, 3rd edn. San Diego, CA: Academic Press.

two horizontally propagating waves of equal amplitude but slightly different wavelengths with wave numbers $k \pm \delta k$ and frequencies $\nu \pm \delta \nu$, respectively. The total disturbance is thus given by eqn [4], where it is understood that only the real part of the right-hand side has physical meaning.

$$\psi(x, t) = \exp\{i[(k + \delta k)x - (\nu + \delta \nu)t]\} + \exp\{i[(k - \delta k)x - (\nu - \delta \nu)t]\} \quad [4]$$

Rearranging terms gives eqn [5].

$$\begin{aligned} \psi(x, t) &= \left[e^{i(\delta kx - \delta \nu t)} + e^{-i(\delta kx - \delta \nu t)} \right] e^{i(kx - \nu t)} \\ &= 2\cos(\delta kx - \delta \nu t) e^{i(kx - \nu t)} \end{aligned} \quad [5]$$

The disturbance is the product of a high frequency carrier wave of wavelength $2\pi/k$ whose phase speed, ν/k , is the average for the two Fourier components, and a low frequency envelope of wavelength $2\pi/\delta k$ that travels at the speed $\delta \nu/\delta k$. Thus, in the limit as $\delta k \rightarrow 0$, the horizontal velocity of the envelope, or group velocity, c_{gx} , is simply as given in eqn [6].

$$c_{gx} = \frac{\partial \nu}{\partial k} \quad [6]$$

This result applies generally to arbitrary wave disturbances provided that the wavelength of the wave group, $2\pi/\delta k$, is large compared to the wavelength of a characteristic wave component, $2\pi/k$.

Wave Classification

Wave activity may propagate horizontally and vertically over large distances from wave sources to regions where transience, nonlinear wave breaking, or dissipation causes interactions with the mean flow. Waves can thus provide a strong nonlocal influence on the momentum and heat budgets and on atmospheric transport. The nature of the major types of wave

motions in the atmosphere can be classified according to several wave dualities.

- **Energy source.** Waves can be divided into free modes and forced modes. Free modes are the normal (i.e., resonant) modes of oscillation of the atmosphere. These can be excited by small random forcing, and are limited in amplitude by dissipation. Free planetary waves (so-called Rossby modes) of global extent and periods of about 2, 5, and 16 days have been detected in the atmosphere, and are especially notable in the stratosphere. Most waves important for weather and climate, however, are forced modes. Examples of forced modes are Rossby waves excited by flow over topography or by land-ocean heating contrasts, and gravity waves generated by convection.
- **Horizontal structure.** Waves can be divided into global modes, which can propagate meridionally as well as zonally, and equatorial modes, which are trapped in an equatorial waveguide and propagate zonally along the Equator. Rossby waves are the most important global modes. Kelvin waves (a special type of gravity wave) and Rossby-gravity waves are the most important equatorial modes.
- **Vertical structure.** Waves can be divided into external and internal modes. External modes are vertically trapped; the energy density of such modes decays exponentially in the vertical. Internal modes are vertically propagating (phase surfaces tilt with height); they can transfer momentum and energy vertically over many scale heights. Under suitable conditions, forced Rossby waves, Kelvin waves, Rossby-gravity waves, and gravity waves can all propagate vertically and transfer energy and momentum between the lower atmosphere and the upper atmosphere.

Typical vertical distributions of forced Rossby and gravity wave amplitudes in the extratropics are shown schematically in **Figure 2**. Gravity wave amplitudes increase exponentially with height at all seasons. Forced Rossby waves of planetary scale propagate to high altitudes during the winter season when the mean zonal winds are westerly in the middle atmosphere, but are trapped during the summer when the mean zonal winds are easterly. Synoptic-scale Rossby waves are trapped in the troposphere at all seasons.

Internal Gravity Waves

Basic Properties

Internal gravity waves are vertically propagating waves associated with the buoyancy restoring force in stably stratified fluids. Parcel oscillations in these waves are parallel to the phase lines, and hence perpendicular to the direction of phase propagation as indicated in **Figure 3**. Thus, for example, the phase surface corresponding to the maximum negative temperature perturbation occurs one-quarter cycle after the maximum upward motion, hence coinciding with maximum upward displacement (and adiabatic cooling) of fluid parcels.

The physical mechanism for internal gravity wave propagation can be understood by considering a fluid parcel displaced from its equilibrium altitude. If a parcel is displaced at

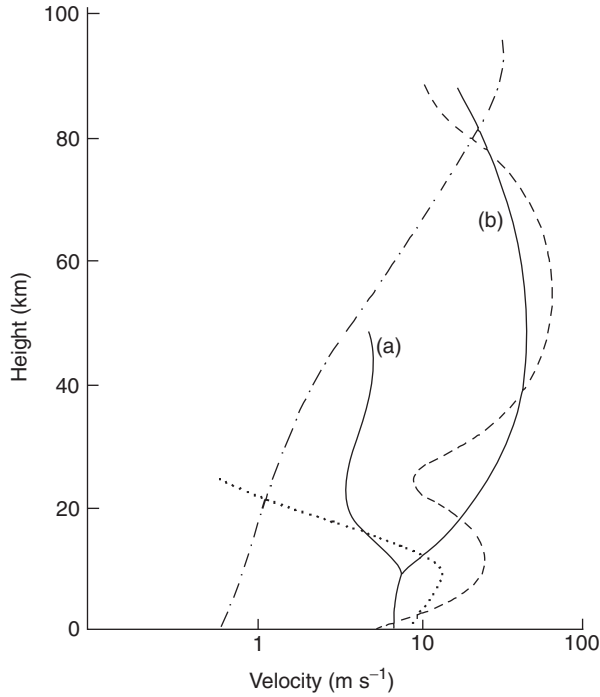


Figure 2 Vertical profiles of characteristic horizontal velocity amplitudes for various types of atmospheric waves and the mean zonal flow. Solid lines: planetary waves (a) summer, (b) winter. Dashed line: zonal mean. Dotted line: synoptic scale. Dotted-dashed line: gravity waves.

a distance δs along a line tilted at an angle α to the vertical, as shown in Figure 4, it will undergo a vertical displacement $\delta z = \delta s \cos \alpha$. For such a parcel, the buoyancy force per unit mass, which is directed in the vertical, is just $-N^2 \delta z$, where

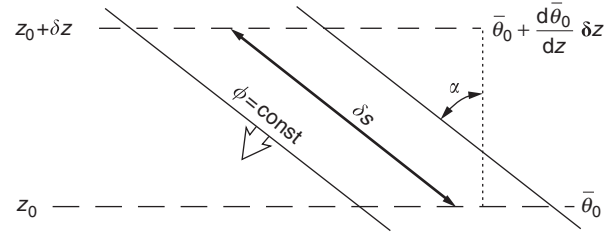


Figure 4 Parcel oscillation path (heavy arrow) for pure gravity waves with phase line tilted at an angle to the vertical. Reproduced from Holton, J.R., 1992. Introduction to Dynamic Meteorology, 3rd edn. San Diego, CA: Academic Press.

$N^2 = -g d \ln \bar{\rho} / dz$ is the square of the buoyancy frequency (with $\bar{\rho}$ the equilibrium density). The component of the buoyancy force perpendicular to the parcel path is balanced exactly by the wave's pressure gradient force, which is itself perpendicular to the parcel path as shown in Figure 3. The parcel acceleration is thus given by the component of the buoyancy force parallel to the parcel path (eqn [7]).

$$-N^2 (\delta s \cos \alpha) \cos \alpha = -(N \cos \alpha)^2 \delta s \quad [7]$$

The momentum equation for the parcel oscillation is then given by eqn [8].

$$\frac{d^2(\delta s)}{dt^2} = -(N \cos \alpha)^2 \delta s \quad [8]$$

This has the general solution $\delta s = \exp[\pm i(N \cos \alpha)t]$. Thus, the parcel executes a simple harmonic oscillation at the frequency $\nu = N \cos \alpha$. This frequency depends only on the static stability (measured by the buoyancy frequency N) and the angle of the phase lines to the vertical.

Dispersion Relationship

In the absence of rotation, the parameters governing two-dimensional buoyancy waves approximately satisfy the dispersion relation of eqn [9].

$$(\nu - \bar{u}k)^2 \equiv \hat{\nu}^2 = N^2 k^2 (m^2 + k^2)^{-1} \quad [9]$$

Here $k = 2\pi/L_x$ and $m = 2\pi/L_z$ are the zonal and vertical wave numbers, respectively, L_x and L_z are the zonal and vertical wavelengths, respectively, and $\hat{\nu} = \nu - \bar{u}k$. The parameter $\hat{\nu}$, which is the frequency relative to the mean flow, is referred to as the Doppler shifted frequency or intrinsic frequency. For most observed gravity waves, eqn [10] applies.

$$\frac{\hat{\nu}}{N} \approx \pm \frac{k}{m} \approx \pm \cos \alpha \quad [10]$$

α is again the angle of phase lines to the vertical. Thus the ratio of the intrinsic wave frequency to the buoyancy frequency determines the slope of the phase lines. By convention, the zonal wave number k is always positive, so if the frequency ν is positive, the zonal phase speed $c_x = \nu/k$ is also positive. For waves in which phase lines slope upward in the positive x direction (as in Figure 2) the vertical wave number m must be negative. This can be easily verified by noting that along lines of constant phase eqn [11] applies, so that for t constant and

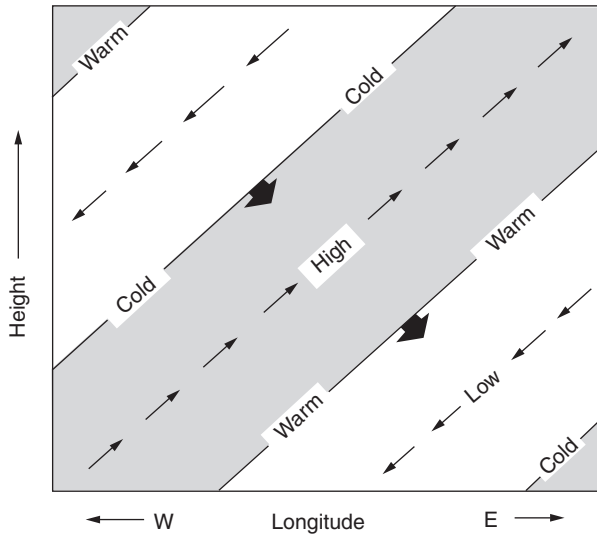


Figure 3 Schematic cross section showing phase relationships among the geopotential, temperature, and velocity perturbations for an internal gravity wave. Thin arrows indicate the perturbation velocity field; and blunt solid arrows show the phase velocity. Shading shows regions of upward motion.

k positive, m must be negative if ϕ is to be constant for x and z both increasing.

$$\phi = kx + mz - vt = \text{constant} \quad [11]$$

In this case the negative root must then be taken in eqn [10], since frequency is positive.

Energy propagates with the group velocity $c_g \equiv (c_{gx}, c_{gz})$, which for the dispersion relation eqn [10] can be expressed as eqn [12].

$$(c_{gx}, c_{gz}) = \left(\frac{\partial v}{\partial k}, \frac{\partial v}{\partial m} \right) = (\bar{u}, 0) \pm \frac{N}{m^2} (m, -k) \quad [12]$$

For the eastward-propagating wave of Figure 3 the negative root applies and m is negative, so that the group velocity relative to the mean flow $\bar{u}$ is eastward and upward. Since $|c_{gz}/(c_{gx} - \bar{u})| = |k/m|$, the group velocity vector relative to the mean flow is parallel to lines of constant phase. Thus, energy propagates eastward and upward relative to the mean flow as phase propagates eastward and downward phase and energy propagate in opposite directions in the vertical!

Most gravity waves have sources at the surface or in the troposphere and propagate energy upward into the stratosphere. They thus have downward phase propagation as in Figure 3. Gravity waves of periods greater than a few hours are influenced by the Coriolis effect, which causes parcel trajectories to be elliptical rather than linear. For a detailed discussion of these waves, see **Gravity Waves: Buoyancy and Buoyancy Waves: Theory**.

Rossby (Planetary) Waves

The wave type that is of most importance for large-scale meteorological processes in the troposphere and the stratosphere is the Rossby wave or planetary wave. In its most general sense the Rossby wave owes its existence to the latitudinal gradient of potential vorticity. This gradient causes fluid parcels that are displaced in latitude to develop anomalous potential vorticity, which induces velocity anomalies that in turn produce further anomalies in potential vorticity. For an illustration of the mechanism of the Rossby wave, see **Dynamical Meteorology: Overview**.

Barotropic Rossby Waves

The simplest Rossby waves occur in the case in which motions are nondivergent and barotropic (i.e., independent of depth). Such conditions are a reasonable first approximation for the flow at the 500 hPa level (mid-troposphere). The dispersion relationship for barotropic Rossby waves may be derived formally by finding wave-type solutions of the linearized barotropic vorticity equation. The barotropic vorticity equation states that the vertical component of absolute vorticity, $\zeta + f$, is conserved following the horizontal motion. Here ζ is the relative vorticity owing to the winds and f is the planetary vorticity (or Coriolis parameter) due to the rotation of the Earth. The analysis is facilitated by the use of the mid-latitude β -plane approximation. In this approximation, the variation of the planetary vorticity with latitude is approximated by expanding the latitudinal dependence of the Coriolis parameter, $f = 2\Omega \sin \phi$ (where Ω is the angular velocity of rotation of the Earth and ϕ

is latitude) in a Taylor series about a reference latitude ϕ_0 and retaining only the first two terms of the series to yield $f = f_0 + \beta y$, where f_0 and β are, respectively, the values of f and df/dy evaluated at the reference latitude ϕ_0 . With this approximation, eqn [13] gives the barotropic vorticity equation.

$$\left(\frac{\partial}{\partial t} + u \frac{\partial}{\partial x} + v \frac{\partial}{\partial y} \right) \zeta + \beta v = 0 \quad [13]$$

The flow is assumed to consist of a constant basic state zonal velocity plus a small horizontal perturbation. Since the horizontal velocity in the barotropic model is nondivergent, the perturbation velocity can be represented in terms of a perturbation stream function (eqn [14]).

$$u' = -\frac{\partial \psi'}{\partial y} \quad v = v' = \frac{\partial \psi'}{\partial x} \quad [14]$$

$$\zeta = \zeta' = \left(\frac{\partial v'}{\partial x} - \frac{\partial u'}{\partial y} \right) = \nabla^2 \psi'$$

The linearized form of eqn [13] is then given by eqn [15].

$$\frac{\partial}{\partial t} + \bar{u} \frac{\partial}{\partial x} \nabla^2 \psi' + \beta \frac{\partial \psi'}{\partial x} = 0 \quad [15]$$

A solution of eqn [15] can be obtained in the form of eqn [16].

$$\psi' = \text{Re} \{ \Psi \exp[i(kx + ly - vt)] \} \quad [16]$$

Here k and l are wave numbers in the zonal and meridional directions, respectively. Substituting eqn [16] into eqn [15] and solving for $c_x = v/k$ yields eqn [17].

$$c_x = \bar{u} - \frac{\beta}{k^2 + l^2} \quad [17]$$

According to eqn [17] the zonal phase propagation of Rossby waves is always westward relative to the mean zonal flow and depends inversely on the square of the horizontal wave number. Therefore, Rossby waves are dispersive. This result is consistent with the fact that the advection of planetary vorticity, which tends to make disturbances propagate westward, increasingly dominates over relative vorticity advection as the wavelength of a disturbance increases. For a typical mid-latitude synoptic-scale disturbance, with similar meridional and zonal scales ($l \approx k$) and zonal wavelength of order 6000 km, the Rossby wave phase speed relative to the zonal flow calculated from eqn [10] is approximately -8 m s^{-1} . Since the mean zonal wind near the 500 hPa level in mid-latitudes is generally greater than 8 m s^{-1} , synoptic-scale Rossby waves usually move eastward, but at a phase speed relative to the ground that is somewhat less than the mean zonal wind speed.

For longer wavelengths, the westward Rossby wave phase speed may be large enough to balance the eastward advection by the mean zonal wind so that the resulting disturbance is stationary relative to the surface of the earth. From eqn [17] the barotropic Rossby wave solution is stationary relative to the ground ($c_x = 0$) when eqn [18] holds.

$$(k^2 + l^2) = \frac{\beta}{\bar{u}} \quad [18]$$

The group velocity for barotropic Rossby waves is given by eqn [19].

$$(c_{gx}, c_{gy}) = \left(\frac{\partial v}{\partial k}, \frac{\partial v}{\partial l} \right) = \left[\bar{u} + \frac{\beta(k^2 - l^2)}{(k^2 + l^2)^2}, \frac{2\beta kl}{(k^2 + l^2)^2} \right] \quad [19]$$

Unlike the Rossby wave phase speed, which is always westward relative to the mean flow, the zonal component of group velocity may be either eastward or westward relative to the mean flow, depending on the ratio of the zonal and meridional wave numbers. Eastward (i.e., downstream) propagation is believed to be important for the development and evolution of tropospheric weather disturbances in the westerlies of the extratropical regions. For stationary modes, the phase velocity relative to the mean flow is westward, but the zonal component of group velocity is always eastward, as can be seen by substituting $\bar{u} = \beta/(k^2 + l^2)$ into eqn (19). In that case Rossby wave dispersion produces downstream propagation of energy.

Stationary Rossby modes (i.e., modes with $c_x = 0$) are strongly excited by flow over major mountain ranges. Such waves can be illustrated by the barotropic vorticity equation if a term is included to account for dilution and concentration of vorticity due to the divergence and convergence associated with flow forced over the mountains. For typical mid-latitude zonal winds the stationary Rossby wave pattern excited by flow over

the Rockies and Himalayas (Figure 5) features lee-side troughs in agreement with observations.

Vertically Propagating Rossby Waves

Vertically propagating Rossby waves are forced modes generated by flow over continental-scale topography, by continent-ocean heating contrasts, and by nonlinear interactions among transient tropospheric wave disturbances. Such motions are approximately governed by the conservation following the geostrophic wind of a dynamical field called quasigeostrophic potential vorticity.

The linearized quasigeostrophic potential vorticity equation for a basic state with constant mean zonal wind $\bar{u}$ can be expressed as in eqn [20], where q' is given by eqn [21].

$$\frac{\partial}{\partial t} + \bar{u} \frac{\partial q'}{\partial x} + v' \beta = 0 \quad [20]$$

$$q' = \nabla^2 \psi' \frac{f_0^2}{\rho_0} \frac{\partial}{\partial z} \left(\frac{\rho_0}{N^2} \frac{\partial \psi'}{\partial z} \right) \quad [21]$$

Here ρ_0 , the reference density, decreases with height exponentially at a constant scale height H , so that $\rho_0 = \rho_s \exp(-z/H)$ and ρ_s the density at sea level. Equation [20] states that the rate of change of q following the mean flow $\bar{u}$ is given by the advection of the planetary vorticity by the disturbance meridional wind component. Equation [20] has solutions in the form of [22], where k , l , and m are the zonal, meridional, and vertical wave numbers, respectively; and v is the frequency of the waves.

$$\psi' = A \exp(i(kx + ly + mz - vt) + \frac{z}{2H}) \quad [22]$$

Note that the factor $\exp(z/2H)$, which is proportional to $\rho_0^{-1/2}$, is introduced to account for the fact that in this simple model the perturbation stream function for vertically propagating waves grows in height as the inverse square root of density. Substituting eqn [22] into eqn [20] and solving for m^2 yields eqn [23] where again $c_x = v/k$ is the zonal phase speed relative to the ground.

$$m^2 = \frac{N^2}{f_0^2} \left[\frac{\beta}{\bar{u} - c_x} - (k^2 + l^2) \right] - \frac{1}{4H^2} \quad [23]$$

For known c_x , k , and l , eqn [23] gives the vertical structure of the waves. For forced wave solutions the wave frequency is specified as that of the forcing. Thus, for topographically forced solutions the zonal phase speed relative to the ground is zero ($c_x = 0$), and $m^2 > 0$ only if eqn [24] is satisfied.

$$0 < \bar{u} < u_c, \text{ where } u_c = \beta \left[k^2 + l^2 + \frac{f_0^2}{4N^2 H^2} \right]^{-1} \quad [24]$$

For $m^2 > 0$ waves can propagate vertically, while for $m^2 \leq 0$ the waves are trapped. Thus, for stationary waves, vertical propagation exists only in the presence of mean westerly winds that are less than a critical value, u_c . This result is referred to as the Charney–Drazin criterion. For example, for a wave with large meridional scale, $l \approx 2\pi/(12\,000 \text{ km})$ and planetary wave number 1 ($k = a^{-1}$), this theory gives $u_c \approx 60 \text{ m s}^{-1}$. More accurate calculations with spherical geometry and realistic winds give $u_c \approx 100 \text{ m s}^{-1}$. It is clear from eqn [24] that u_c decreases rapidly as k and l increase, i.e., the critical value is

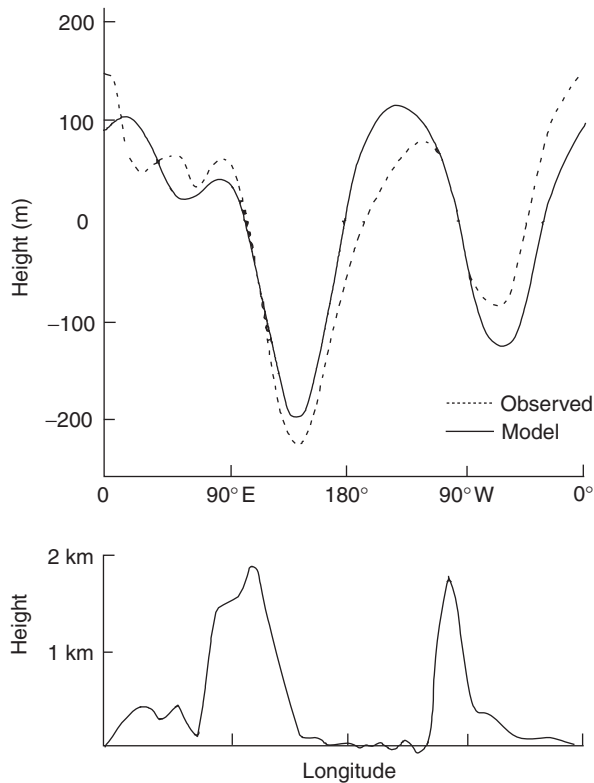


Figure 5 Upper panel shows the longitudinal variation of the stationary barotropic Rossby wave disturbance geopotential height field (solid line) compared with the observed 500 hPa height disturbance at 45° N in January (dashed line). The lower panel shows the topography profile used in the computation. Reproduced from Held, I.M., 1983. In: Hoskins, B.J., Pearce, R., (eds.) *Large-Scale Dynamical Processes in the Atmosphere*. New York, NY: Academic Press.

determined largely by horizontal wave scales. In practice only zonal wave numbers 1 and 2 propagate significantly into the extratropical stratosphere, and this happens only in the winter hemisphere where $\bar{u} > 0$. Therefore, the Charney–Drazin criterion provides an approximate explanation for the absence of stationary planetary waves in the summer stratosphere (where winds are easterly) and the dominance of waves of zonal wave numbers 1 and 2 in the winter stratosphere.

Equatorial Waves

The change of sign of the Coriolis parameter at the Equator leads to a special class of large-scale atmospheric waves that are trapped laterally in the equatorial region but propagate zonally and vertically. Such waves have important consequences for the circulation of the equatorial middle atmosphere. Equatorial waves are most easily analyzed by utilizing an equatorial β -plane approximation, in which the Coriolis parameter is assumed to vary linearly with distance from the equator so that $f = \beta y$, where $\beta \equiv (df/dy)_{\phi=0} = 2\Omega/a$, with a the radius of the Earth.

Solutions for zonally propagating wave disturbances on an equatorial β -plane in a basic state at rest can be expressed in the form of eqn [25].

$$v' = \text{Re} \left\{ \Psi(y) \exp \left[i(kx + mz - \nu t) + \frac{z}{2H} \right] \right\} \quad [25]$$

Again it is assumed that $k > 0$, so that $\nu > 0$ implies eastward propagation relative to the ground at phase speed $c_x = \nu/k$.

The Meridional Structure

The meridional structure in eqn [25] can be determined by letting λ be given by eqn [26] and defining the transformed variable $\xi = (\beta|m|/N)^{1/2}y$.

$$\lambda^2 = \frac{N}{\beta|m|} \left(\frac{m^2 \nu^2}{N^2} - k^2 - \frac{k\beta}{\nu} \right) \quad [26]$$

It can then be shown that Ψ satisfies eqn [27], which is just the Schrodinger equation for a quantum harmonic oscillator.

$$\frac{d^2 \Psi}{d\xi^2} + (\lambda^2 - \xi^2) \Psi = 0 \quad [27]$$

This equation is analogous to the governing equation for a simple harmonic oscillator, eqn [28], where I^2 is a positive constant.

$$\frac{d^2 \Psi}{dy^2} + I^2 \Psi = 0 \quad [28]$$

The equatorial case given by eqn [27] differs from the simple harmonic oscillator in that oscillating solutions exist only for $\xi^2 < \lambda^2$ (with $\lambda^2 > 0$), while exponential behavior occurs otherwise. Solutions of eqn [27] that satisfies the condition of equatorial trapping ($\Psi \rightarrow 0$ as $\xi \rightarrow \pm\infty$) exist if and only if $\lambda^2 = 2n + 1$, where $n = 0, 1, 2, \dots$. These solutions are the Hermite functions defined by eqn [29].

$$\Psi_n = e^{-\xi^2/2} H_n(\xi) \quad [29]$$

$H_n(\xi)$ is the Hermite polynomial of order n , so that successive values are given by eqn [30].

$$H_0(\xi) = 1 \quad H_1(\xi) = 2\xi \quad H_2(\xi) = 4\xi^2 - 2 \quad [30]$$

Substitution of eqn [29] into eqn [27] then yields the dispersion relation (eqn [31]).

$$\frac{N}{\beta|m|} \left(\frac{m^2 \nu^2}{N^2} - k^2 - \frac{k\beta}{\nu} \right) = 2n + 1 \quad [31]$$

Rossby-Gravity Waves

The lowest meridional mode solution of eqn [31], the $n = 0$ mode, is called a Rossby-gravity wave. It has a meridional velocity perturbation, v' , that is symmetric about the Equator. For this mode eqn [32] applies, which determines the vertical wave number for specified zonal wave number and frequency.

$$|m| = \frac{N}{\nu^2} (\beta + \nu k) \quad [32]$$

Letting $s = ka$, where s is the number of wavelengths around a latitude circle ($s = 0, 1, 2, \dots$), and recalling that $\beta = 2\Omega/a$, eqn [32] implies that with $\nu < 0$ (corresponding to the observed westward propagation of these modes) a solution exists only for $|\hat{\nu}/\Omega| < 2/s$. For $s = 4$, $|\hat{\nu}/\Omega| < 1/2$, so that the period in this case must be greater than 2 days.

In terms of the original y coordinate, the solution for the meridional velocity perturbation in the Rossby-gravity mode can be expressed as eqn [33].

$$\tilde{v} = V_0 \exp \left[-\frac{(1 + kv\beta^{-1})\beta^2 y^2}{2\nu^2} \right] \quad [33]$$

The corresponding solutions for the zonal wind and geopotential perturbations are given by eqn [34].

$$\begin{aligned} \begin{Bmatrix} \tilde{u} \\ \tilde{\Phi} \end{Bmatrix} &= V_0 \begin{Bmatrix} -i\beta y(1 + kv\beta^{-1})\nu^{-1} \\ +i\nu y \end{Bmatrix} \\ &\times \exp \left[-\frac{(1 + kv\beta^{-1})\beta^2 y^2}{2\nu^2} \right] \end{aligned} \quad [34]$$

For observed Rossby-gravity waves in the equatorial stratosphere, $\nu \sim 2\pi/(4\text{days})$, $s = 4$, and $L_z = 2\pi/m \sim 6\text{--}8$ km. The horizontal and vertical structures of the Rossby-gravity wave are shown in Figures 6 and 7, respectively.

Kelvin Waves

There is an important type of equatorial wave in which the meridional velocity perturbation vanishes ($v' = 0$). For this

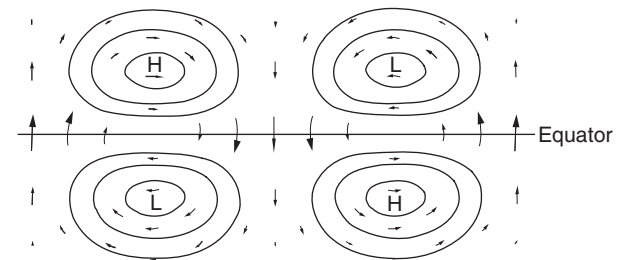


Figure 6 Latitude-longitude section showing horizontal velocity and geopotential perturbations associated with an equatorial Rossby-gravity wave.

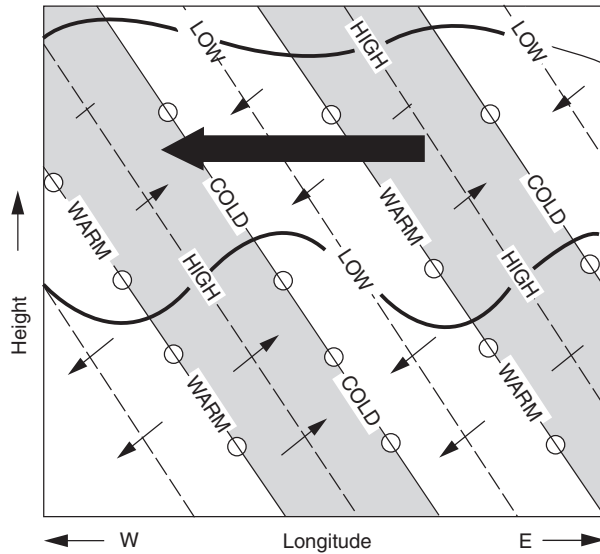


Figure 7 Longitude–height section along the equator showing geopotential, temperature, and wind perturbations for a thermally damped Rossby-gravity wave. Heavy wavy lines indicate material lines. Areas of high pressure are shaded. Small thin arrows indicate zonal and vertical velocity perturbations, with length proportional to wave amplitude, which decrease with height owing to damping. The large black arrow indicates the net wave-induced force owing to wave damping. Reproduced from Holton, J.R., 1992. *An Introduction to Dynamic Meteorology*, 3rd edn. San Diego: Academic Press.

mode it turns out that eqn [35] applies, which is the same as the dispersion equation for an eastward-propagating internal gravity wave.

$$v = \pm \frac{kN}{m} \quad [35]$$

The zonal wind disturbance for this mode has a Gaussian distribution about the Equator (eqn [36]).

$$\tilde{u} = u_0 \exp \left[\frac{-\beta y^2 k^2}{2v} \right] \quad [36]$$

Temperature and pressure are similarly symmetric about the Equator. Note that this solution requires that $v/k = c_x > 0$, otherwise it would grow without bound away from the Equator, which explains why only eastward-propagating Kelvin waves are allowed.

For observed Kelvin waves in the equatorial stratosphere, eqn [37] is satisfied.

$$\frac{v}{k} \sim 30 \text{ m s}^{-1} \quad \text{so} \quad L_z = \frac{2\pi}{|m|} \sim 12 \text{ km} \quad [37]$$

The horizontal and vertical structures for the Kelvin mode are shown in **Figures 8 and 9**, respectively. The vertical structure is the same as that for a two-dimensional internal gravity wave. The meridional width for observed Kelvin waves is of order Y_L given by eqn [38].

$$Y_L = \left| \frac{2v}{\beta k} \right|^{1/2} = \left| \frac{2c_x}{\beta} \right|^{1/2} \sim 2000 \text{ km} \quad [38]$$

Kelvin waves are a significant source of westerly momentum for driving the westerly phase of the equatorial stratospheric quasi-biennial oscillation (QBO).

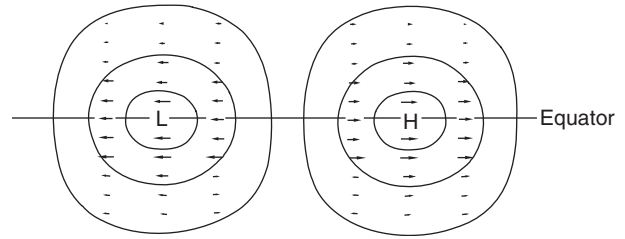


Figure 8 Latitude–longitude section showing horizontal velocity and geopotential perturbations associated with an equatorial Kelvin wave.

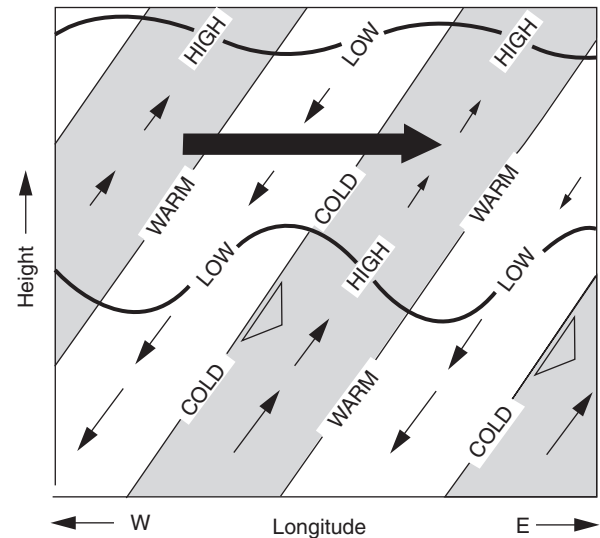


Figure 9 Longitude–height section along the equator showing geopotential, temperature, and wind perturbations for a thermally damped Kelvin wave. Heavy wavy lines indicate material lines; and short blunt arrows show phase propagation. Areas of high pressure are shaded. Small thin arrows indicate zonal and vertical velocity perturbations, with length proportional to wave amplitude, which decreases with height owing to damping. The large black arrow indicates the net wave-induced force owing to wave damping. Reproduced from Holton, J.R., 1992. *An Introduction to Dynamic Meteorology*, 3rd edn. San Diego, CA: Academic Press.

See also: Dynamical Meteorology: Baroclinic Instability; Kelvin Waves; Overview; Rossby Waves. **Gravity Waves:** Buoyancy and Buoyancy Waves: Theory; Overview. **Mesoscale Meteorology:** Waterspouts. **Middle Atmosphere:** Quasi-Biennial Oscillation. **Mountain Meteorology:** Lee Waves and Mountain Waves. **Tropical Meteorology and Climate:** Equatorial Waves.

Further Reading

Andrews, D.G., Holton, J.R., Leovy, C.B., 1987. *Middle Atmospheric Dynamics*. Academic Press, Orlando, FL.
 Gill, A.E., 1982. *Atmosphere–Ocean Dynamics*. Academic Press, New York, NY.
 Holton, J.R., 1992. *Introduction to Dynamic Meteorology*, 3rd ed. Academic Press, San Diego, CA.
 Lighthill, J., 1978. *Waves in Fluids*. Cambridge University Press, Cambridge, UK.

Second Edition

ENCYCLOPEDIA OF ATMOSPHERIC SCIENCES

Edited by
Gerald R. North **Editor-in-Chief**
John Pyle and Fuqing Zhang



6 VOLUME SET

ENCYCLOPEDIA OF **ATMOSPHERIC SCIENCES**

SECOND EDITION

VOLUME 3

This page intentionally left blank

ENCYCLOPEDIA OF **ATMOSPHERIC SCIENCES**

SECOND EDITION

EDITOR-IN-CHIEF

GERALD R NORTH

Texas A&M University, College Station, TX, USA

EDITORS

JOHN PYLE

Cambridge University, Cambridge, UK

FUQING ZHANG

Pennsylvania State University, University Park, PA, USA

VOLUME 3



Amsterdam • Boston • Heidelberg • London • New York • Oxford
Paris • San Diego • San Francisco • Singapore • Sydney • Tokyo
Academic Press is an imprint of Elsevier



Academic Press is an imprint of Elsevier
32 Jamestown Road, London NW1 7BY, UK
225 Wyman Street, Waltham, MA 02451, USA
525 B Street, Suite 1800, San Diego, CA 92101-4495, USA

Copyright © 2015 Elsevier Ltd. unless otherwise stated. All rights reserved.

No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means electronic, mechanical, photocopying, recording or otherwise without the prior written permission of the publisher

Permissions may be sought directly from Elsevier's Science & Technology Rights Department in Oxford, UK: phone (+44) (0) 1865 843830; fax (+44) (0) 1865 853333; email: permissions@elsevier.com. Alternatively you can submit your request online by visiting the Elsevier web site at <http://elsevier.com/locate/permissions>, and selecting *Obtaining permission to use Elsevier material*

Notice

No responsibility is assumed by the publisher for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions or ideas contained in the material herein, Because of rapid advances in the medical sciences, in particular, independent verification of diagnoses and drug dosages should be made

British Library Cataloguing in Publication Data

A catalog record for this book is available from the British Library

Library of Congress Catalog Number:

A catalog record for this book is available from the Library of Congress

ISBN (print): 978-0-12-382225-3

For information on all Elsevier publications
visit our website at store.elsevier.com

Printed and bound in the United Kingdom

15 16 17 18 19 10 9 8 7 6 5 4 3 2 1

		Working together to grow libraries in developing countries
www.elsevier.com • www.bookaid.org		

<i>Acquisitions Editor:</i> Simon Holt <i>Project Manager:</i> Michael Nicholls <i>Associate Project Manager:</i> Marise Willis <i>Designer:</i> Matthew Limbert

DEDICATION

This second edition of the *Encyclopedia of Atmospheric Sciences* is dedicated to the memory of James Holton who was editor-in-chief of the first edition. He was a great researcher and colleague inspiring an entire generation of atmospheric scientists.

This page intentionally left blank

CONTENTS

<i>List of Contributors</i>	xxvii
<i>Preface to the First Edition</i>	xxxix
<i>Preface to the Second Edition</i>	xli
<i>Editor Biographies</i>	xlili
<i>Guide to Using the Encyclopedia</i>	xlvi

VOLUME 1

BASIC ATMOSPHERIC STRUCTURE AND CONCEPTS **1**

Beaufort Wind Scale	1
<i>L Hasse</i>	
Wind Chill	7
<i>M Bluestein</i>	
Standard Atmosphere	12
<i>W W Vaughan</i>	

AEROSOLS **17**

Aerosol–Cloud Interactions and Their Radiative Forcing	17
<i>U Lohmann</i>	
Aerosol Physics and Chemistry	23
<i>M Kalberer</i>	
Climatology of Stratospheric Aerosols	32
<i>L W Thomason and J-P Vernier</i>	
Climatology of Tropospheric Aerosols	40
<i>N Bellouin and J Haywood</i>	
Dust	48
<i>I N Sokolik</i>	
Observations and Measurements	53
<i>P H McMurry</i>	
Role in Radiative Transfer	66
<i>G A Ban-Weiss, and W D Collins</i>	

Role in Climate Change <i>N Bellouin</i>	76
Soot <i>P Chylek, S G Jennings, and R Pinnick</i>	86
Agricultural Meteorology and Climatology <i>E S Takle</i>	92
ARCTIC AND ANTARCTIC	98
Antarctic Climate <i>J Turner</i>	98
Arctic Climate <i>M C Serreze</i>	107
Arctic Haze <i>L M Russell and G E Shaw</i>	116
AIR SEA INTERACTIONS	
Freshwater Flux <i>J Schulz</i>	122
Momentum, Heat, and Vapor Fluxes <i>P K Taylor</i>	129
Sea Surface Temperature <i>W J Emery</i>	136
Surface Waves <i>A Benilov</i>	144
AVIATION METEOROLOGY	153
Aircraft Emissions <i>R R Friedl</i>	153
Aircraft Icing <i>M K Politovich</i>	160
Aviation Weather Hazards <i>A J Bedard, Jr</i>	166
Clear Air Turbulence <i>G P Ellrod (Retired), J A Knox, P F Lester, and L J Ehernberger (Retired)</i>	177
BIOGEOCHEMICAL CYCLES	187
Sulfur Cycle <i>P Brimblecombe</i>	187
Bromine <i>R von Glasow and C Hughes</i>	194
Heavy Metals <i>T D Jickells and A R Baker</i>	201

Iodine	205
<i>L J Carpenter</i>	
BOUNDARY LAYER (ATMOSPHERIC) AND AIR POLLUTION	220
Overview	220
<i>P J Mason and D J Thomson</i>	
Air Pollution Meteorology	227
<i>X-M Hu</i>	
Coherent Structures	237
<i>F T M Nieuwstadt and J C R Hunt</i>	
Complex Terrain	242
<i>J J Finnigan</i>	
Convective Boundary Layer	250
<i>M A LeMone</i>	
Microclimate	258
<i>M W Rotach and P Calanca</i>	
Modeling and Parameterization	265
<i>A A M Holtslag</i>	
Observational Techniques <i>In Situ</i>	274
<i>E F Bradley</i>	
Observational Techniques: Remote	284
<i>W M Angevine and C J Senff</i>	
Ocean Mixed Layer	290
<i>L Kantha and C A Clayson</i>	
Stably Stratified Boundary Layer	299
<i>L Mahrt</i>	
Surface Layer	305
<i>G L Geernaert</i>	
Urban Heat Islands	310
<i>J C Luvall, D A Quattrochi, D L Rickman, and M G Estes, Jr</i>	
Diurnal Cycle	319
<i>A Betts</i>	
CHEMISTRY OF THE ATMOSPHERE	324
Chemical Kinetics	324
<i>R P Wayne</i>	
Ion Chemistry	333
<i>J L Fox</i>	
Isotopes, Stable	348
<i>C A M Brenninkmeijer</i>	
Laboratory Kinetics	356
<i>D J Donaldson and S N Wren</i>	

Methane <i>E Dlugokencky, and S Houweling</i>	363
Observations for Chemistry (<i>In Situ</i>): Ozone Sonde <i>H G J Smit</i>	372
Observations for Chemistry (<i>In Situ</i>): Particles <i>T Deshler</i>	379
Observations for Chemistry (<i>In Situ</i>): Water Vapor Sonde <i>J B Smith</i>	387
Observations for Chemistry (Remote Sensing): IR/FIR (Satellite, Balloon and Ground) <i>H Fischer and F Hase</i>	401
Observations for Chemistry (Remote Sensing): Lidar <i>G Vaughan</i>	411
Observations for Chemistry (Remote Sensing): Microwave <i>J Waters</i>	418
Principles of Chemical Change <i>R P Wayne</i>	429
Radioactivity: Cosmogenic Radionuclides <i>D Lal</i>	437
Volcanoes: Composition of Emissions <i>M T Coffey and J W Hannigan</i>	446
Tracers <i>K A Boering</i>	450

VOLUME 2

CLIMATE AND CLIMATE CHANGE	1
Overview <i>D L Hartmann</i>	1
Carbon Dioxide <i>C L Sabine and R A Feely</i>	10
Climate Feedbacks <i>A E Dessler and M D Zelinka</i>	18
Climate Prediction: Empirical and Numerical <i>S Hastenrath</i>	26
Climate Variability: Decadal to Centennial Variability <i>D G Martinson</i>	33
Climate Variability: Nonlinear and Random Effects <i>M Ghil</i>	38
Climate Variability: North Atlantic and Arctic Oscillation <i>J W Hurrell</i>	47
Climate Variability: Seasonal and Interannual Variability <i>D S Gutzler</i>	61

Energy Balance Climate Models <i>G R North and K-Y Kim</i>	69
Global Impacts of the Madden–Julian Oscillation <i>C Zhang</i>	73
Greenhouse Effect <i>G R North</i>	80
History of Scientific Work on Climate Change <i>S Weart</i>	87
Intergovernmental Panel on Climate Change <i>K E Trenberth</i>	90
Nuclear Winter <i>A Robock</i>	95
Radiative–Convective Equilibrium Climate Models <i>N O Renno and X Huang</i>	102
Volcanoes: Role in Climate <i>A Robock</i>	105
CLOUDS AND FOG	112
Cloud Modeling <i>W-K Tao and M Moncrieff</i>	112
Contrails <i>P Minnis</i>	121
Cloud Microphysics <i>D Lamb</i>	133
Classification of Clouds <i>A L Rangno (Retiree)</i>	141
Climatology <i>S Warren, R Eastman, and C J Hahn</i>	161
Measurement Techniques <i>In situ</i> <i>D Baumgardner, J-F Gayet, A Korolev, C Twohy, and J Fugal</i>	170
Fog <i>P J Croft and B Ward</i>	180
Noctilucent Clouds <i>G E Thomas</i>	189
Stratus and Stratocumulus <i>R Wood</i>	196
CRYOSPHERE	201
Glaciers, Topography, and Climate <i>A B G Bush and M P Bishop</i>	201
Permafrost <i>T E Osterkamp and C R Burn</i>	208

Sea Ice	217
<i>M C Serreze, F Fetterer, and W F Weeks (Retired)</i>	
Snow (Surface)	227
<i>M Sturm</i>	
DATA ASSIMILATION AND PREDICTABILITY	237
Data Assimilation	237
<i>A C Lorenc</i>	
Ensemble-Based Data Assimilation	241
<i>Z Meng and F Zhang</i>	
Ensemble Prediction	248
<i>R Buizza</i>	
Predictability and Chaos	258
<i>L A Smith</i>	
DYNAMICAL METEOROLOGY	265
Overview	265
<i>J R Holton</i>	
Acoustic Waves	272
<i>K E Gilbert</i>	
Atmospheric Tides	287
<i>J Oberheide, M E Hagan, A D Richmond, and J M Forbes</i>	
Balanced Flow	298
<i>M E McIntyre</i>	
Baroclinic Instability	304
<i>R Grotjahn</i>	
Coriolis Force	313
<i>D W Moore</i>	
Critical Layers	317
<i>P Haynes</i>	
Hamiltonian Dynamics	324
<i>T G Shepherd</i>	
Hydraulic Flow	332
<i>R B Smith</i>	
Inertial Instability	334
<i>J A Knox</i>	
Kelvin–Helmholtz Instability	343
<i>P G Drazin</i>	
Kelvin Waves	347
<i>B Wang</i>	
Kinematics	353
<i>D D Houghton</i>	

Laboratory Geophysical Fluid Dynamics <i>R L Pfeffer</i>	360
Lagrangian Dynamics <i>I Roulstone</i>	369
Potential Vorticity <i>M E McIntyre</i>	375
Primitive Equations <i>A A White and N Wood</i>	384
Quasigeostrophic Theory <i>H C Davies and H Wernli</i>	393
Rossby Waves <i>P B Rhines</i>	404
Solitary Waves <i>J P Boyd</i>	417
Static Stability <i>J A Young</i>	423
Stationary Waves (Orographic and Thermally Forced) <i>S Nigam and E DeWeaver</i>	431
Symmetric Stability <i>H B Bluestein</i>	446
Vorticity <i>J R Holton</i>	451
Wave-CISK <i>C S Bretherton</i>	455
Wave Mean-Flow Interaction <i>M Juckes</i>	458
Waves <i>J R Holton</i>	464

VOLUME 3

ELECTRICITY IN THE ATMOSPHERE	1
Global Electrical Circuit <i>E R Williams</i>	1
Ions in the Atmosphere <i>K L Aplin and R G Harrison</i>	9
Lightning <i>M B Baker</i>	14
Sprites <i>W A Lyons</i>	20
Forensic Meteorology <i>L E Branscome</i>	28

GENERAL CIRCULATION OF THE ATMOSPHERE	33
Overview	33
<i>J M Wallace, D W J Thompson, and P Beresford</i>	
Angular Momentum of the Atmosphere	43
<i>D A Salstein</i>	
Energy Cycle	51
<i>R Grotjahn</i>	
Weather Regimes and Multiple Equilibria	65
<i>F Molteni</i>	
Mean Characteristics	73
<i>R Grotjahn</i>	
Teleconnections	90
<i>S Nigam and S Baxter</i>	
 GLOBAL CHANGE	 110
Climate Record: Surface Temperature Trends	110
<i>P D Jones</i>	
Sea Level Change	121
<i>R S Nerem</i>	
Upper Atmospheric Change	128
<i>R G Roble</i>	
Biospheric Impacts and Feedbacks	132
<i>B A Hungate and G W Koch</i>	
 GRAVITY WAVES	 141
Overview	141
<i>D C Fritts</i>	
Buoyancy and Buoyancy Waves: Optical Observations	153
<i>M J Taylor and W R Pendleton, Jr</i>	
Buoyancy and Buoyancy Waves: Theory	160
<i>T J Dunkerton</i>	
Gravity Waves Excited by Jets and Fronts	164
<i>R Plougonven and F Zhang</i>	
Convectively Generated Gravity Waves	171
<i>T P Lane</i>	
 HYDROLOGY, FLOODS AND DROUGHTS	 180
Overview	180
<i>R C Bales</i>	
Deserts and Desertification	185
<i>V P Tchakerian</i>	
Drought	193
<i>S Quiring</i>	

Flooding	201
<i>C A Doswell III</i>	
Groundwater and Surface Water	209
<i>S Ge and S M Gorelick</i>	
Modeling and Prediction	217
<i>Z Yu</i>	
Palmer Drought Severity Index	224
<i>L Nkemdirim</i>	
Soil Moisture	232
<i>A Robock</i>	
LAND-ATMOSPHERE INTERACTIONS	240
Overview	240
<i>R E Dickinson</i>	
Canopy Processes	244
<i>P D Blanken</i>	
Trace Gas Exchange	256
<i>J N Cape and D Fowler</i>	
LIDAR	262
Atmospheric Sounding Introduction	262
<i>P S Argall and R Sica</i>	
Backscatter	270
<i>C M R Platt and R L Collins</i>	
Differential Absorption Lidar	277
<i>S Ismail and E V Browell</i>	
Doppler	289
<i>R M Hardesty</i>	
Raman	296
<i>D N Whiteman</i>	
Resonance	305
<i>C S Gardner and R L Collins</i>	
Magnetosphere	309
<i>G K Parks</i>	
MESOSCALE METEOROLOGY	316
Overview	316
<i>D J Parker</i>	
Cloud and Precipitation Bands	323
<i>R M Rauber and M Ramamurthy</i>	
Gust Fronts	331
<i>R Rotunno</i>	

Hail and Hailstorms	334
<i>C Knight, N Knight, and H E Brooks</i>	
Mesoscale Convective Systems	339
<i>A Laing</i>	
Microbursts	335
<i>R M Wakimoto</i>	
Severe Storms	361
<i>C A Doswell III</i>	
Waterspouts	369
<i>J H Golden</i>	
Bow Echoes and Derecho	384
<i>M L Weisman</i>	
Density Currents	395
<i>P G Baines</i>	
Convective Storms: Overview	401
<i>M L Weisman</i>	
MESOSPHERE	411
Atomic Species in the Mesopause Region	411
<i>M G Mlynczak and L A Hunt</i>	
Ionosphere	422
<i>M C Kelley</i>	
Metal Layers	430
<i>J M C Plane</i>	
Polar Summer Mesopause	436
<i>R H Varney and M C Kelley</i>	
VOLUME 4	
MIDDLE ATMOSPHERE	1
Planetary Waves	1
<i>A K Smith and J Perlwitz</i>	
Polar Vortex	12
<i>M R Schoeberl and P A Newman</i>	
Quasi-Biennial Oscillation	18
<i>T J Dunkerton, J A Anstey, and L J Gray</i>	
Semiannual Oscillation	26
<i>K Hamilton</i>	
Stratospheric Sudden Warmings	30
<i>A O'Neill, A J Charlton-Perez, and L M Polvani</i>	
Transport Circulation	41
<i>S E Strahan</i>	

Zonal Mean Climatology <i>P Braesicke</i>	50
MOUNTAIN METEOROLOGY	57
Overview <i>R B Smith</i>	57
Cold Air Damming <i>B A Colle</i>	62
Downslope Winds <i>D R Durran</i>	69
Katabatic Winds <i>T R Parish</i>	75
Land and Sea Breezes <i>R A Pielke, Sr</i>	80
Lee Vortices <i>C C Epifanio</i>	84
Lee Waves and Mountain Waves <i>D R Durran</i>	95
Orographic Effects: Lee Cyclogenesis <i>C Schär</i>	103
Valley Winds <i>D Zardi</i>	114
NUMERICAL MODELS	135
Chemistry Models <i>M P Chipperfield and S R Arnold</i>	135
Coupled Ocean-Atmosphere Models: Physical Processes <i>M Zhang</i>	144
General Circulation Models <i>C R Mechoso and A Arakawa</i>	153
Methods <i>J Thuburn</i>	161
Model Physics Parameterization <i>D J Stensrud, M C Coniglio, K H Knopfmeier, and A J Clark</i>	167
Parameter Estimation <i>A Aksoy</i>	181
Parameterization of Physical Processes: Clouds <i>R Forbes, C Jakob, and M Miller</i>	187
Parameterization of Physical Processes: Gravity Wave Fluxes <i>M J Alexander</i>	194
Parameterization of Physical Processes: Turbulence and Mixing <i>A Beljaars</i>	200

Spectral Models <i>F Baer</i>	212
Mesoscale Atmospheric Modeling <i>R A Pielke, Sr</i>	219
Cloud-System Resolving Modeling and Aerosols <i>W-K Tao and T Matsui</i>	222
Large-Eddy Simulation <i>C-H Moeng and P P Sullivan</i>	232
Regional Prediction Models <i>B W Golding</i>	241
Convective Storm Modeling <i>M D Parker</i>	246
OBSERVATIONS PLATFORMS	255
Balloons <i>J-P Pommereau</i>	255
Buoys <i>J M Hemsley</i>	264
Kites <i>B B Balsley</i>	268
Radiosondes <i>W F Dabberdt and H Turtiainen</i>	273
Rockets <i>M F Larsen</i>	285
OCEANOGRAPHIC TOPICS	290
General Processes <i>N C Wells</i>	290
Surface/Wind Driven Circulation <i>R X Huang</i>	301
Thermohaline Circulation <i>R X Huang</i>	315
Water Types and Water Masses <i>W J Emery</i>	329
OPTICS, ATMOSPHERIC	338
Optical Remote Sensing Instruments <i>G G Shepherd</i>	338
Airglow Instrumentation <i>M Conde</i>	346

OZONE DEPLETION AND RELATED TOPICS	353
Long-Term Ozone Changes <i>N R P Harris</i>	353
Ozone as a UV Filter <i>J E Frederick</i>	359
Ozone Depletion Potentials <i>D J Wuebbles</i>	364
Photochemistry of Ozone <i>G K Moortgat and A R Ravishankara</i>	370
Stratospheric Ozone Recovery <i>D J Hofmann and R Müller</i>	380
Surface Ozone Effects on Vegetation <i>M Ashmore</i>	389
Surface Ozone (Human Health) <i>M Lippmann</i>	397
PALEOCLIMATOLOGY	404
Ice Cores <i>E J Steig</i>	404
Varves <i>R Gilbert</i>	411
RADAR	415
Cloud Radar <i>T Uttal</i>	415
Incoherent Scatter Radar <i>M P Sulzer</i>	422
Mesosphere—Stratosphere—Troposphere and Stratosphere—Troposphere Radars and Wind Profilers <i>G Vaughan and D Hooper</i>	429
Meteor Radar <i>N J Mitchell</i>	438
Polarimetric Doppler Weather Radar <i>R J Doviak and R D Palmer</i>	444
Precipitation Radar <i>S E Yuter</i>	455
Synthetic Aperture Radar (Land Surface Applications) <i>R K Vincent</i>	470
VOLUME 5	
RADIATION TRANSFER IN THE ATMOSPHERE	1
Radiation, Solar <i>Q Fu</i>	1

Absorption and Thermal Emission <i>R M Goody and X Huang</i>	5
Cloud-Radiative Processes <i>Q Fu</i>	13
Non-local Thermodynamic Equilibrium <i>M López-Puertas and B Funke</i>	16
Scattering <i>M Mishchenko, L Travis, and A Lacis</i>	27
Ultraviolet Radiation <i>K Stamnes</i>	37
Ultraviolet, Surface <i>R McKenzie and S Madronich</i>	45
SATELLITES AND SATELLITE REMOTE SENSING	51
Aerosol Measurements <i>R A Kahn</i>	51
Earth's Radiation Budget <i>N G Loeb and B A Wielicki</i>	67
GPS Meteorology <i>S S Leroy</i>	77
Measuring Ozone from Space — TOMS and SBUV <i>R D McPeters and R S Stolarski</i>	87
Orbits <i>S Q Kidder</i>	95
Precipitation <i>G Liu</i>	107
Remote Sensing: Cloud Properties <i>P Yang and B A Baum</i>	116
Research <i>M D King</i>	128
Surface Wind and Stress <i>W T Liu</i>	138
Temperature Soundings <i>A Dudhia</i>	145
Water Vapor <i>J E Harries</i>	157
SOLAR SYSTEM/SUN, ATMOSPHERES, EVOLUTION OF ATMOSPHERES	163
Evolution of Earth's Atmosphere <i>Y L Yung, M L Wong, and E J Gaidos</i>	163
Planetary Atmospheres: Mars <i>R M Haberle</i>	168

Planetary Atmospheres: Venus <i>P J Gierasch and Y L Yung</i>	178
Solar Terrestrial Interactions: Climate Impact <i>J D Haigh</i>	183
Solar Winds <i>S T Suess and B T Tsurutani</i>	189
Meteors <i>P Jenniskens</i>	195
STATISTICAL METHODS	201
Data Analysis: Empirical Orthogonal Functions and Singular Vectors <i>C S Bretherton</i>	201
Data Analysis: Time Series Analysis <i>G R North</i>	205
STRATOSPHERIC CHEMISTRY TOPICS	211
Overview <i>J A Pyle</i>	211
Halogens <i>D Toohey</i>	215
Halogen Sources, Anthropogenic <i>A McCulloch and P M Midgley</i>	221
Halogen Sources, Natural (Methyl Bromide and Related Gases) <i>S Yvon-Lewis and J H Butler</i>	228
HO _x <i>T F Hanisco</i>	233
Hydrogen Budget <i>J E Harries</i>	238
Reactive Nitrogen (NO _x and NO _y) <i>Y Kondo</i>	242
Stratospheric Water Vapor <i>K H Rosenlof</i>	250
STRATOSPHERE/TROPOSPHERE EXCHANGE AND STRUCTURE	257
Global Aspects <i>J R Holton</i>	257
Local Processes <i>J F Lamarque and P Hess</i>	262
Tropopause <i>M Dameris</i>	269

SYNOPTIC METEOROLOGY	273
Anticyclones <i>S J Colucci</i>	273
Forecasting <i>D Mansfield</i>	280
Weather Maps <i>R Reynolds</i>	289
Cyclogenesis <i>G J Hakim</i>	299
Extratropical Cyclones <i>A Joly</i>	304
Fronts <i>D M (David) Schultz and W Blumen</i>	337
Fronts in the Lower Stratosphere <i>A L Lang</i>	344
Frontogenesis <i>D M (David) Schultz</i>	353
Jet Streaks <i>P Cunningham and D Keyser</i>	359
Lake-Effect Storms <i>P J Sousounis</i>	370
Polar Lows <i>I A Renfrew</i>	379
Thermal Low <i>R H Johnson</i>	386
 THERMODYNAMICS	 391
Humidity Variables <i>J A Curry</i>	391
Moist (Unsaturated) Air <i>J A Curry</i>	394
Saturated Adiabatic Processes <i>J A Curry</i>	398
 Thermosphere <i>S C Solomon and R G Roble</i>	 402
 VOLUME 6	
TROPICAL CYCLONES AND HURRICANES	1
Overview and Theory <i>R A Tomas and P J Webster</i>	1

Hurricane Dynamics <i>Y Wang</i>	8
Hurricane Predictability <i>J A Sippel</i>	30
Hurricanes: Observation <i>F D Marks</i>	35
Tropical Cyclogenesis <i>Z Wang</i>	57
Tropical Cyclones and Climate Change <i>T R Knutson</i>	65
Tropical Cyclones in the Western North Pacific <i>J C L Chan</i>	77
Tropical Cyclones: Secondary Eyewall Formation <i>C-C Wu and Y-H Huang</i>	85
TROPICAL METEOROLOGY AND CLIMATE	91
El Niño and the Southern Oscillation: Observation <i>N Nicholls</i>	91
El Niño and the Southern Oscillation: Theory <i>P Chang and S E Zebiak</i>	97
Equatorial Waves <i>M C Wheeler and H Nguyen</i>	102
Hadley Circulation <i>J Lu and G A Vecchi</i>	113
Intertropical Convergence Zone <i>D E Waliser and X Jiang</i>	121
Intraseasonal Oscillation (Madden–Julian Oscillation) <i>R A Madden</i>	132
Madden–Julian Oscillation: Skeleton and Conceptual Models <i>A J Majda and S N Stechmann</i>	137
Monsoon: Overview <i>J Slingo</i>	146
Monsoon: Dynamical Theory <i>P J Webster and J Fasullo</i>	151
Monsoon: ENSO–Monsoon Interactions <i>K-M Lau</i>	165
Tropical Climates <i>S Hastenrath</i>	170
Walker Circulation <i>K-M Lau and S Yang</i>	177

TROPOSPHERIC CHEMISTRY AND COMPOSITION	182
Aerosols/Particles <i>J H Seinfeld</i>	182
Aliphatic Hydrocarbons <i>J Rudolph and O Stein</i>	188
Aromatic Hydrocarbons <i>I Barnes</i>	204
Biogenic Hydrocarbons <i>A Guenther</i>	214
Cloud Chemistry <i>P Herckes and J L Collett, Jr</i>	218
H ₂ <i>U Schmidt and T Wetter</i>	226
Hydroxyl Radical <i>K C Clemitshaw</i>	232
Mercury <i>J Munthe and J Sommar</i>	239
Oxidizing Capacity <i>D H Ehhalt, F Rohrer, and A Wahner</i>	243
Peroxyacetyl Nitrate <i>H B Singh</i>	251
Sulfur Chemistry, Organic <i>I Barnes</i>	255
Volatile Organic Compounds Overview: Anthropogenic <i>R G Derwent</i>	265
TURBULENCE AND MIXING	268
Overview <i>P Haynes</i>	268
Turbulence, Two Dimensional <i>P Bartello</i>	273
Turbulent Diffusion <i>A Venkatram and S Du</i>	277
WEATHER FORECASTING	287
Marine Meteorology <i>L Xie and B Liu</i>	287
Operational Meteorology <i>D R Novak</i>	293
Seasonal and Interannual Weather Prediction <i>J P Li and R Q Ding</i>	303
Severe Weather Forecasting <i>D J Stensrud, H E Brooks, and S J Weiss</i>	313

Wildfire Weather <i>J Coen</i>	323
Inadvertant Weather Modification <i>S A Changnon</i>	332
<i>Appendix 1: Physical Constants</i>	337
<i>Appendix 2: Units and their SI Equivalents</i>	339
<i>Appendix 3: Periodic Table of the Elements</i>	340
<i>Appendix 4: The Geologic Time Scale</i>	341
<i>Index</i>	343

This page intentionally left blank

LIST OF CONTRIBUTORS

A. Aksoy
University of Miami, Miami, FL, USA; and NOAA Hurricane Research Division, Miami, FL, USA

M.J. Alexander
NorthWest Research Associates (NWRA), Boulder, CO, USA

W.M. Angevine
CIRES, University of Colorado; and NOAA Earth System Research Laboratory, Boulder, CO, USA

J.A. Anstey
University of Oxford, Oxford, UK

K.L. Aplin
University of Oxford, Oxford, UK

A. Arakawa
University of California, Los Angeles, CA, USA

P.S. Argall
The University of Western Ontario, London, ON, Canada

S.R. Arnold
University of Leeds, Leeds, UK

M. Ashmore
University of York, York, UK

F. Baer
University of Maryland, College Park, MD, USA

P.G. Baines
University of Melbourne, Melbourne, VIC, Australia

A.R. Baker
University of East Anglia, Norwich, UK

M.B. Baker
University of Washington, Seattle, WA, USA

R.C. Bales
University of Arizona, Tucson, AZ, USA

B.B. Balsley
University of Colorado, Boulder, CO, USA

G.A. Ban-Weiss
Lawrence Berkeley National Laboratory, Berkeley, CA, USA; and University of Southern California, Los Angeles, CA, USA

I. Barnes
University of Wuppertal, Wuppertal, Germany

P. Bartello
McGill University, Montréal, QC, Canada

B.A. Baum
University of Wisconsin–Madison, Madison, WI, USA

D. Baumgardner
Universidad Nacional Autónoma de México, Mexico City, D.F., Mexico

S. Baxter
University of Maryland, College Park, MD, USA

A.J. Bedard, Jr.
National Oceanic and Atmospheric Administration, Boulder, CO, USA

A. Beljaars
European Centre for Medium-Range Weather Forecasts, Reading, England

N. Bellouin
University of Reading, Reading, UK

A. Benilov
Acute Solutions, Highlands, NJ, USA

P. Beresford
European Centre for Medium-Range Weather Forecasts, Reading, UK

A. Betts
Atmospheric Research, Pittsford, VT, USA

M.P. Bishop
Texas A&M University, College Station, TX, USA

P.D. Blanken
University of Colorado at Boulder, Boulder, CO, USA

H.B. Bluestein
University of Oklahoma, Norman, OK, USA

M. Bluestein
Indiana University – Purdue University, Indianapolis, IN, USA

W. Blumen[†]
University of Colorado Boulder, Boulder, CO, USA

K.A. Boering
University of California – Berkeley, Berkeley, CA, USA

J.P. Boyd
University of Michigan, Ann Arbor, MI, USA

E.F. Bradley
CSIRO Land and Water, Canberra, ACT, Australia

P. Braesicke
Karlsruhe Institute of Technology, Karlsruhe, Germany

L.E. Branscome
Climatological Consulting Corporation, FL, USA

C.A.M. Brenninkmeijer
Max Planck Institute for Chemistry, Mainz, Germany

C.S. Bretherton
University of Washington, Seattle, WA, USA

P. Brimblecombe
University of East Anglia, Norwich, UK

H.E. Brooks
National Oceanic and Atmospheric Administration, Norman, OK, USA

E.V. Browell
STARS II Affiliate, NASA Langley Research Center, Hampton, VA, USA

R. Buizza
ECMWF, Reading, UK

C.R. Burn
Carleton University, Ottawa, ON, Canada

A.B.G. Bush
University of Alberta, Edmonton, AB, Canada

J.H. Butler
NOAA Earth System Research Laboratory, Boulder, CO, USA

P. Calanca
Agroscope Reckenholz-Taenikon, Zurich, Switzerland

J.N. Cape
Edinburgh Research Station, Midlothian, UK

L.J. Carpenter
University of York, York, UK

J.C.L. Chan
City University of Hong Kong, Hong Kong

P. Chang
Texas A&M University, College Station, TX, USA

S.A. Changnon
University of Illinois, IL, USA

A.J. Charlton-Perez
University of Reading, Earley Gate, Reading, UK

M.P. Chipperfield
University of Leeds, Leeds, UK

P. Chylek
Dalhousie University, NS, Canada

A.J. Clark
University of Oklahoma and National Oceanic and Atmospheric Administration, Norman, OK, USA

C.A. Clayson
Woods Hole Oceanographic Institution, Woods Hole, MA, USA

K.C. Clemitshaw
Imperial College of Science, Technology, and Medicine, Ascot, UK

J. Coen
National Center for Atmospheric Research, Boulder, CO, USA

M.T. Coffey
National Center for Atmospheric Research, Boulder, CO, USA

B.A. Colle
Stony Brook University – SUNY, Stony Brook, NY, USA

J.L. Collett, Jr.
Colorado State University, Fort Collins, CO, USA

R.L. Collins
University of Alaska Fairbanks, Fairbanks, AK, USA

W.D. Collins
Lawrence Berkeley National Laboratory, Berkeley, CA, USA

S.J. Colucci
Cornell University, Ithaca, NY, USA

M. Conde
University of Alaska Fairbanks, Fairbanks, AK, USA

M.C. Coniglio
National Oceanic and Atmospheric Administration, Norman, OK, USA

[†] Deceased.

- P.J. Croft
Kean University, Union, NJ, USA
- P. Cunningham
Florida State University, Tallahassee, FL, USA
- J.A. Curry
Georgia Institute of Technology, Atlanta, GA, USA
- W.F. Dabberdt
Vaisala Company, Boulder, CO, USA
- M. Dameris
Deutsches Zentrum für Luft- und Raumfahrt, Institut für Physik der Atmosphäre, Oberpfaffenhofen, Wessling, Germany
- H.C. Davies
Institute for Atmospheric and Climate Science, Swiss Federal Institute of Technology (ETH), Zürich, Switzerland
- R.G. Derwent
rdscientific, Newbury, UK
- T. Deshler
University of Wyoming, Laramie, WY, USA
- A.E. Dessler
Texas A&M University, College Station, TX, USA
- E. DeWeaver
University of Wisconsin, Madison, WI, USA
- R.E. Dickinson
University of Texas at Austin, Austin, TX, USA
- R.Q. Ding
Institute of Atmospheric Physics, Chinese Academy of Sciences, Beijing, China
- E. Dlugokencky
NOAA Earth System Research Laboratory, Boulder, CO, USA
- D.J. Donaldson
University of Toronto, Toronto, ON, Canada
- C.A. Doswell, III
University of Oklahoma, Norman, OK, USA
- R.J. Doviak
National Severe Storms Laboratory, Norman, OK, USA
- P.G. Drazin[†]
University of Bath, England, UK
- S. Du
California Air Resources Board, Sacramento, CA, USA
- A. Dudhia
University of Oxford, Oxford, UK
- T.J. Dunkerton
Northwest Research Associates, Bellevue, WA, USA
- D.R. Durran
University of Washington, Seattle, WA, USA
- R. Eastman
University of Washington, Seattle, WA, USA
- L.J. Ehernberger
National Aeronautics and Space Administration, Dryden Flight Research Center, Edwards, CA, USA
- D.H. Ehhalt
Forschungszentrum Jülich, Jülich, Germany
- G.P. Ellrod
National Oceanographic and Atmospheric Administration/National Environmental Satellite, Data, and Information Service, Granby, CT, USA
- W.J. Emery
University of Colorado, Boulder, CO, USA
- C.C. Epifanio
Texas A&M University, College Station, TX, USA
- M.G. Estes
Universities Space Research Association, Huntsville, AL, USA
- J. Fasullo
University of Colorado – Boulder, Boulder, CO, USA
- R.A. Feely
NOAA Pacific Marine Environmental Laboratory, Seattle, WA, USA
- F. Fetterer
University of Colorado, Boulder, CO, USA
- J.J. Finnigan
CSIRO Atmospheric Research, Black Mountain, ACT, Australia
- H. Fischer
Institute of Meteorology and Climate Research (IMK), Karlsruhe Institute of Technology, Karlsruhe, Germany
- J.M. Forbes
University of Colorado, Boulder, CO, USA
- R. Forbes
European Centre for Medium-Range Weather Forecasts, Reading, UK
- D. Fowler
Edinburgh Research Station, Midlothian, UK

[†] Deceased.

J.L. Fox
Wright State University, Dayton, OH, USA

J.E. Frederick
The University of Chicago, Chicago, IL, USA

R.R. Friedl
California Institute of Technology, Pasadena, CA, USA

D.C. Fritts
GATS Inc., Boulder, CO, USA

Q. Fu
University of Washington, Seattle, WA, USA

J. Fugal
Max Planck Institute of Chemistry, Mainz, Germany

B. Funke
Instituto de Astrofísica de Andalucía, CSIC, Granada, Spain

E.J. Gaidos
University of Hawaii at Manoa, Honolulu, HI, USA

C.S. Gardner
University of Illinois at Urbana-Champaign, Urbana, IL, USA

J.-F. Gayet
Université Blaise Pascal, Clermont Ferrand, France

S. Ge
University of Colorado, Boulder, CO, USA

G.L. Geernaert
US Department of Energy, Washington, DC, USA

M. Ghil
Ecole Normale Supérieure, Paris, France; and University of California, Los Angeles, CA, USA

P.J. Gierasch
Cornell University, Ithaca, NY, USA

K.E. Gilbert
University of Mississippi, University, MS, USA

R. Gilbert
Queen's University, Kingston, ON, Canada

J.H. Golden
Forecast Systems Laboratory, NOAA, Boulder, CO, USA

B.W. Golding
Met Office, Exeter, UK

R.M. Goody
Harvard University (Emeritus), Cambridge, MA, USA

S.M. Gorelick
Stanford University, Stanford, CA, USA

L.J. Gray
University of Oxford, Oxford, UK

R. Grotjahn
University of California, Davis, CA, USA

A. Guenther
Pacific Northwest National Laboratory, Richland, WA, USA

D.S. Gutzler
University of New Mexico, Albuquerque, NM, USA

R.M. Haberle
NASA/Ames Research Center, Moffett Field, Mountain View, CA, USA

M.E. Hagan
National Center for Atmospheric Research, Boulder, CO, USA

C.J. Hahn
University of Arizona, Tucson, AZ, USA

J.D. Haigh
Blackett Laboratory, Imperial College London, London, UK

G.J. Hakim
University of Washington, Seattle, WA, USA

K. Hamilton
University of Hawaii, Honolulu, HI, USA

T.F. Hanisco
Harvard University, Cambridge, MA, USA

J.W. Hannigan
National Center for Atmospheric Research, Boulder, CO, USA

R.M. Hardesty
NOAA Environmental Technology Laboratory, Boulder, CO, USA

J.E. Harries
Imperial College of Science, Technology and Medicine, Blackett Laboratory, London, UK

N.R.P. Harris
University of Cambridge, Cambridge, UK

R.G. Harrison
The University of Reading, Reading, UK

D.L. Hartmann
University of Washington, Seattle, WA, USA

F. Hase
Institute of Meteorology and Climate Research (IMK), Karlsruhe Institute of Technology, Karlsruhe, Germany

- L. Hasse
Universität Kiel, Kiel, Germany
- S. Hastenrath
University of Wisconsin, Madison, WI, USA
- P. Haynes
University of Cambridge, Cambridge, UK
- J. Haywood
Met Office, Exeter, UK
- J.M. Hemsley
National Data Buoy Center, Stennis Space Center, MS, USA
- P. Herckes
Arizona State University, Tempe, AZ, USA
- P. Hess
National Center for Atmospheric Research, Boulder, CO, USA
- D.J. Hofmann[†]
NOAA Climate Monitoring and Diagnostics Laboratory, Boulder, CO, USA
- J.R. Holton
University of Washington, Seattle, WA, USA
- A.A.M. Holtslag
Wageningen University, Wageningen, The Netherlands
- D. Hooper
Science & Technology Facilities Council (STFC), Didcot, UK
- D.D. Houghton
University of Wisconsin-Madison, Madison, WI, USA
- S. Houweling
SRON Netherlands Institute for Space Research, Utrecht, The Netherlands
- X.-M. Hu
University of Oklahoma, Norman, OK, USA
- R.X. Huang
Woods Hole Oceanographic Institution, Woods Hole, MA, USA
- X. Huang
University of Michigan, Ann Arbor, MI, USA
- Y.-H. Huang
National Taiwan University, Taipei, Taiwan
- C. Hughes
University of York, York, UK
- B.A. Hungate
Northern Arizona University, Flagstaff, AZ, USA
- J.C.R. Hunt
University College London, London, UK
- L.A. Hunt
Science Systems and Applications Incorporated, Hampton, VA, USA
- J.W. Hurrell
National Center for Atmospheric Research, Boulder, CO, USA
- S. Ismail
Science Directorate, NASA Langley Research Center, Hampton, VA, USA
- C. Jakob
Monash University, VIC, Australia
- S.G. Jennings
National University of Ireland, Galway, Ireland
- P. Jenniskens
SETI Institute, Moffett Field, CA, USA
- X. Jiang
University of California, Los Angeles, CA, USA
- T.D. Jickells
University of East Anglia, Norwich, UK
- R.H. Johnson
Colorado State University, Fort Collins, CO, USA
- A. Joly
Centre National de Recherches Météorologiques – Groupe d'étude de l'Atmosphère Météorologique, Météo-France and CNRS, Toulouse, France
- P.D. Jones
Climatic Research Unit, University of East Anglia, Norwich, UK
- M. Jukes
University of Oxford, Oxford, UK
- R.A. Kahn
NASA Goddard Space Flight Center, Greenbelt, MD, USA
- M. Kalberer
University of Cambridge, Cambridge, UK
- L. Kantha
University of Colorado, Boulder, CO, USA
- M.C. Kelley
Cornell University, Ithaca, NY, USA

[†] Deceased.

D. Keyser
*University at Albany, State University of New York,
Albany, NY, USA*

S.Q. Kidder
*Colorado State University, Fort Collins,
CO, USA*

K.-Y. Kim
Seoul National University, Seoul, Korea

M.D. King
University of Colorado, Boulder, CO, USA

C. Knight
*National Center for Atmospheric Research, Boulder,
CO, USA*

N. Knight
*National Center for Atmospheric Research, Boulder,
CO, USA*

K.H. Knopfmeier
*University of Oklahoma; and National Oceanic and
Atmospheric Administration, Norman, OK, USA*

J.A. Knox
University of Georgia, Athens, GA, USA

T.R. Knutson
*NOAA Geophysical Fluid Dynamics Laboratory,
Princeton, NJ, USA*

G.W. Koch
Northern Arizona University, Flagstaff, AZ, USA

Y. Kondo
The University of Tokyo, Tokyo, Japan

A. Korolev
*Meteorological Service of Canada, Toronto, ON,
Canada*

A. Lacis
*Goddard Institute for Space Studies, New York,
NY, USA*

A. Laing
*National Center for Atmospheric Research, Boulder,
CO, USA*

D. Lal
*Scripps Institution of Oceanography, La Jolla,
CA, USA*

J.F. Lamarque
*National Center for Atmospheric Research, Boulder,
CO, USA*

D. Lamb
*The Pennsylvania State University, University Park,
PA, USA*

T.P. Lane
*The University of Melbourne, Melbourne, VIC,
Australia*

A.L. Lang
*University of Albany – State University of New York,
Albany, NY, USA*

M.F. Larsen
Clemson University, Clemson, SC, USA

K.-M. Lau
*NASA/Goddard Space Flight Center, Greenbelt, MD,
USA*

M.A. LeMone
*National Center for Atmospheric Research, Boulder,
CO, USA*

S.S. Leroy
*Harvard School of Engineering and Applied Sciences,
Cambridge, MA, USA*

P.F. Lester
San Jose State University, San Jose, CA, USA

J.P. Li
*Institute of Atmospheric Physics, Chinese Academy of
Sciences, Beijing, China*

M. Lippmann
New York University, Tuxedo, NY, USA

B. Liu
North Carolina State University, Raleigh, NC, USA

G. Liu
Florida State University, Tallahassee, FL, USA

W.T. Liu
*California Institute of Technology, Pasadena,
CA, USA*

N.G. Loeb
NASA Langley Research Center, Hampton, VA, USA

U. Lohmann
ETH Zurich, Zürich, Switzerland

M. López-Puertas
*Instituto de Astrofísica de Andalucía, CSIC, Granada,
Spain*

A.C. Lorenc
The Met Office, Bracknell, Berkshire, UK

J. Lu
*Pacific Northwest National Laboratory, Richland, WA,
USA*

J.C. Luvall
*National Space Science and Technology Center,
Marshall Space Flight Center, Huntsville, AL, USA*

- W.A. Lyons
FMA Research Inc., Fort Collins, CO, USA
- R.A. Madden
National Center for Atmospheric Research, Boulder, CO, USA
- S. Madronich
National Center for Atmospheric Research, Boulder, CO, USA
- L. Mahrt
Oregon State University, Corvallis, OR, USA
- A.J. Majda
New York University, New York, NY, USA
- D. Mansfield
National Meteorological Center, Bracknell, UK
- F.D. Marks
Hurricane Research Division, Miami, FL, USA
- D.G. Martinson
Columbia University, Palisades, NY, USA
- P.J. Mason
Met Office, Bracknell, UK
- T. Matsui
NASA/Goddard Space Flight Center, Greenbelt, MD, USA; and University of Maryland, College Park, MD, USA
- A. McCulloch
University of Bristol, Bristol, UK
- M.E. McIntyre
University of Cambridge, Cambridge, UK
- R. McKenzie
National Institute for Water and Atmospheric Research, Lauder, Central Otago, New Zealand
- P.H. McMurtry
University of Minnesota, Minneapolis, MN, USA
- R.D. McPeters
NASA Goddard Space Flight Center, Greenbelt, MD, USA
- C.R. Mechoso
University of California, Los Angeles, CA, USA
- Z. Meng
Peking University, Beijing, China
- P.M. Midgley
M & D Consulting, Leinfelden Metternich, Germany
- M. Miller
European Centre for Medium-Range Weather Forecasts, Reading, UK
- P. Minnis
Science Directorate, NASA Langley Research Center, Hampton, VA, USA
- M. Mishchenko
Goddard Institute for Space Studies, New York, NY, USA
- N.J. Mitchell
The University of Bath, Bath, UK
- M.G. Mlynarczyk
NASA Langley Research Center, Hampton, VA, USA
- C.-H. Moeng
National Center for Atmospheric Research, Boulder, CO, USA
- F. Molteni
Abdus Salam International Centre for Theoretical Physics, Trieste, Italy
- M. Moncrieff
National Center for Atmospheric Research, Boulder, CO, USA
- D.W. Moore
Pacific Marine Environmental Laboratory, Seattle, WA, USA
- G.K. Moortgat
Max-Planck-Institute for Chemistry, Mainz, Germany
- R. Müller
Institute for Energy and Climate Research (IEK-7), Forschungszentrum Jülich, Jülich, Germany
- J. Munthe
IVL Swedish Environmental Research Institute, Göteborg, Sweden
- R.S. Nerem
University of Colorado, Boulder, CO, USA
- P.A. Newman
NASA Goddard, Space Flight Center, Greenbelt, MD, USA
- H. Nguyen
Centre for Australian Weather and Climate Research, Bureau of Meteorology, Melbourne, VIC, Australia
- N. Nicholls
Bureau of Meteorology Research Centre, Melbourne, VIC, Australia
- F.T.M. Nieuwstadt
Delft University of Technology, Delft, The Netherlands
- S. Nigam
University of Maryland, College Park, MD, USA

L. Nkemdirim
University of Calgary, Calgary, AB, Canada

G.R. North
Texas A&M University, College Station, TX, USA

D.R. Novak
Weather Prediction Center, College Park, MD, USA

A. O'Neill
University of Reading, Earley Gate, Reading, UK

J. Oberheide
Clemson University, Clemson, SC, USA

T.E. Osterkamp
University of Alaska, Fairbanks, AK, USA

R.D. Palmer
University of Oklahoma, Oklahoma, OK, USA

T.R. Parish
University of Wyoming, Laramie, WY, USA

D.J. Parker
University of Leeds, Leeds, UK

M.D. Parker
North Carolina State University, Raleigh, NC, USA

G.K. Parks
University of Washington, Seattle, WA, USA

W.R. Pendleton
Utah State University, Logan, UT, USA

J. Perlwitz
University of Colorado, Boulder, CO, USA

R.L. Pfeffer
Florida State University, Tallahassee, FL, USA

R.A. Pielke, Sr.
University of Colorado at Boulder, CO, USA

R. Pinnick
US Army Research Laboratory, Adelphi, MD, USA

J.M.C. Plane
University of Leeds, Leeds, UK

C.M.R. Platt
Colorado State University, Fort Collins, CO, USA

R. Plougonven
Ecole Polytechnique, Palaiseau, France

M.K. Politovich
National Center for Atmospheric Research, Boulder, CO, USA

L.M. Polvani
Columbia University, New York, NY, USA

J.-P. Pommereau
LATMOS, CNRS, Guyancourt, France

J.A. Pyle
University of Cambridge, Cambridge, UK

D.A. Quattrochi
National Space Science and Technology Center, Marshall Space Flight Center, Huntsville, AL, USA

S. Quiring
Texas A&M University, College Station, TX, USA

M. Ramamurthy
University Corporation for Atmospheric Research, Boulder, CO, USA

A.L. Rangno (Retiree)
University of Washington, Seattle, WA, USA

R.M. Rauber
University of Illinois at Urbana-Champaign, Urbana, IL, USA

A.R. Ravishankara
Colorado State University, Fort Collins, CO, USA

I.A. Renfrew
University of East Anglia, Norwich, UK

N.O. Renno
University of Michigan, Ann Arbor, MI, USA

R. Reynolds
University of Reading, Reading, UK

P.B. Rhines
University of Washington, Seattle, WA, USA

A.D. Richmond
National Center for Atmospheric Research, Boulder, CO, USA

D.L. Rickman
National Space Science and Technology Center, Marshall Space Flight Center, Huntsville, AL, USA

R.G. Roble
National Center for Atmospheric Research, Boulder, CO, USA

A. Robock
Rutgers University, New Brunswick, NJ, USA

F. Rohrer
Forschungszentrum Jülich, Jülich, Germany

K.H. Rosenlof
Earth System Research Laboratory, National Oceanic and Atmospheric Administration, Boulder, CO, USA

- M.W. Rotach
University of Innsbruck, Innsbruck, Austria
- R. Rotunno
National Center for Atmospheric Research, Boulder, CO, USA
- I. Roulstone
University of Surrey, Guildford, UK
- J. Rudolph
York University, Toronto, ON, Canada
- L.M. Russell
Scripps Institution of Oceanography, University of California, San Diego, La Jolla, CA, USA
- C.L. Sabine
NOAA Pacific Marine Environmental Laboratory, Seattle, WA, USA
- D.A. Salstein
Atmospheric and Environmental Research, Inc., Lexington, MA, USA
- C. Schär
Atmospheric and Climatic Science ETH, Zürich, Switzerland
- U. Schmidt
Johann Wolfgang Goethe-University, Frankfurt am Main, Institut für Meteorologie und Geophysik, Frankfurt am Main, Germany
- M.R. Schoeberl
Science and Technology Corporation, Lanham, MD, USA
- D.M. (David) Schultz
University of Manchester, Manchester, UK
- J. Schulz
Meteorological Institute, University of Bonn, Bonn, Germany
- J.H. Seinfeld
California Institute of Technology, Pasadena, CA, USA
- C.J. Senff
CIRES, University of Colorado; and NOAA Earth System Research Laboratory, Boulder, CO, USA
- M.C. Serreze
University of Colorado, Boulder, CO, USA
- G.E. Shaw
Geophysical Institute, University of Alaska, Fairbanks, AK, USA
- G.G. Shepherd
York University, Toronto, ON, Canada
- T.G. Shepherd
University of Toronto, Toronto, ON, Canada
- R. Sica
The University of Western Ontario, London, ON, Canada
- H.B. Singh
NASA Ames Research Center, Mountain View, CA, USA
- J.A. Sippel
National Aeronautics and Space Administration (NASA), Greenbelt, MD, USA
- J. Slingo
University of Reading, Reading, UK
- H.G.J. Smit
Research Centre Jülich, Jülich, Germany
- A.K. Smith
National Center for Atmospheric Research, Boulder, CO, USA
- J.B. Smith
Harvard University, Cambridge, MA, USA
- L.A. Smith
London School of Economics, Centre for the Analysis of Time Series, London, UK
- R.B. Smith
Yale University, New Haven, CT, USA
- I.N. Sokolik
Georgia Institute of Technology, Atlanta, GA, USA
- S.C. Solomon
National Center for Atmospheric Research, Boulder, CO, USA
- J. Sommar
Göteborg University, Göteborg, Sweden
- P.J. Sousounis
AIR Worldwide, Boston, MA, USA
- K. Stamnes
Stevens Institute of Technology, Hoboken, NJ, USA
- S.N. Stechmann
University of Wisconsin–Madison, Madison, WI, USA
- E.J. Steig
University of Washington, Seattle, WA, USA
- O. Stein
IEK 8: Troposphere, Research Center Juelich, Juelich, Germany

D.J. Stensrud
*National Oceanic and Atmospheric Administration,
Norman, OK, USA*

R.S. Stolarski
Johns Hopkins University, Baltimore, MD, USA

S.E. Strahan
*Universities Space Research Association, NASA Goddard
Space Flight Center, Greenbelt, MD, USA*

M. Sturm
*US Army Cold Regions Research & Engineering
Laboratory-Alaska, Fort Wainwright, AL, USA*

S.T. Suess
*NASA Marshall Space Flight Center, Huntsville, AL,
USA*

P.P. Sullivan
*National Center for Atmospheric Research, Boulder,
CO, USA*

M.P. Sulzer
Arecibo Observatory, Arecibo, PR, USA

E.S. Takle
Iowa State University, Ames, IA, USA

W.-K. Tao
*NASA/Goddard Space Flight Center, Greenbelt, MD,
USA*

M.J. Taylor
Utah State University, Logan, UT, USA

P.K. Taylor
*Southampton Oceanography Centre, Southampton,
UK*

V.P. Tchakerian
Texas A&M University, College Station, TX, USA

G.E. Thomas
University of Colorado, Boulder, CO, USA

L.W. Thomason
NASA Langley Research Center, Hampton, VA, USA

D.W.J. Thompson
Colorado State University, Fort Collins, CO, USA

D.J. Thomson
Met Office, Bracknell, UK

J. Thuburn
University of Exeter, Exeter, UK

R.A. Tomas
University of Colorado – Boulder, Boulder, CO, USA

D. Toohey
University of Colorado Boulder, Boulder, CO, USA

L. Travis
*Goddard Institute for Space Studies, New York,
NY, USA*

K.E. Trenberth
*National Center for Atmospheric Research, Boulder,
CO, USA*

B.T. Tsurutani
Jet Propulsion Laboratory, Pasadena, CA, USA

J. Turner
British Antarctic Survey, Cambridge, UK

H. Turtiainen
Vaisala Company, Helsinki, Finland

C. Twohy
Oregon State University, Corvallis, OR, USA

T. Uttal
NOAA, Boulder, CO, USA

R.H. Varney
Cornell University, Ithaca, NY, USA

G. Vaughan
University of Manchester, Manchester, UK

W.W. Vaughan
*University of Alabama in Huntsville, Huntsville,
AL, USA*

G.A. Vecchi
GFDL/NOAA, Princeton, NJ, USA

A. Venkatram
*University of California – Riverside, Riverside,
CA, USA*

J.-P. Vernier
*Science Systems and Applications, Inc., Hampton,
VA, USA*

R.K. Vincent
*Bowling Green State University, Bowling Green,
OH, USA*

R. von Glasow
University of East Anglia, Norwich, UK

A. Wahner
Forschungszentrum Jülich, Jülich, Germany

R.M. Wakimoto
*National Center for Atmospheric Research, Boulder,
CO, USA*

D.E. Waliser
California Institute of Technology, Pasadena, CA, USA

J.M. Wallace
University of Washington, Seattle, WA, USA

- B. Wang
University of Hawaii, Honolulu, HI, USA
- Y. Wang
University of Hawaii at Manoa, Honolulu, HI, USA
- Z. Wang
University of Illinois at Urbana-Champaign, Urbana, IL, USA
- B. Ward
Public Works and Natural Resources, Longmont, CO, USA
- S. Warren
University of Washington, Seattle, WA, USA
- J. Waters
California Institute of Technology, Jet Propulsion Laboratory, Pasadena, CA, USA
- R.P. Wayne
University of Oxford, Oxford, UK
- S. Weart
Center for History of Physics, American Institute of Physics, College Park, MD, USA
- P.J. Webster
Georgia Institute of Technology, Atlanta, GA, USA
- P.J. Webster
University of Colorado – Boulder, Boulder, CO, USA
- W.F. Weeks
University of Alaska Fairbanks, Fairbanks, AK, USA
- M.L. Weisman
National Center for Atmospheric Research, Boulder, CO, USA
- S.J. Weiss
National Oceanic and Atmospheric Administration, Norman, OK, USA
- N.C. Wells
University of Southampton, Southampton, UK
- H. Wernli
Institute for Atmospheric and Climate Science, Swiss Federal Institute of Technology (ETH), Zürich, Switzerland
- T. Wetter
Johann Wolfgang Goethe-University, Frankfurt am Main, Institut für Meteorologie und Geophysik, Frankfurt am Main, Germany
- M.C. Wheeler
Centre for Australian Weather and Climate Research, Bureau of Meteorology, Melbourne, VIC, Australia
- A.A. White
University of Surrey, Guildford, UK
- D.N. Whiteman
NASA Goddard Space Flight Center, Greenbelt, MD, USA
- B.A. Wielicki
NASA Langley Research Center, Hampton, VA, USA
- E.R. Williams
Massachusetts Institute of Technology, Cambridge, MA, USA
- M.L. Wong
California Institute of Technology, Pasadena, CA, USA
- N. Wood
Met Office, Exeter, UK
- R. Wood
University of Washington, Seattle, WA, USA
- S.N. Wren
University of Toronto, Toronto, ON, Canada
- C.-C. Wu
National Taiwan University, Taipei, Taiwan
- D.J. Wuebbles
University of Illinois, Urbana, IL, USA
- L. Xie
North Carolina State University, Raleigh, NC, USA
- P. Yang
Texas A&M University, College Station, TX, USA
- S. Yang
NOAA/NWS/NCEP, Climate Prediction Center, Camp Springs, MD, USA
- J.A. Young
University of Wisconsin, Madison, WI, USA
- Z. Yu
College of Hydrology and Water Resources, Hohai University, Nanjing, China; and University of Nevada Las Vegas, Las Vegas, NV, USA
- Y.L. Yung
California Institute of Technology, Pasadena, CA, USA

S.E. Yuter
*North Carolina State University, Raleigh,
NC, USA*

S. Yvon-Lewis
*Texas A&M University, College Station,
TX, USA*

D. Zardi
University of Trento, Trento, Italy

S.E. Zebiak
*International Research Institute for Climate Prediction,
Palisades, NY, USA*

M.D. Zelinka
*Lawrence Livermore National Laboratory, Livermore,
CA, USA*

C. Zhang
University of Miami, Miami, FL, USA

F. Zhang
*Pennsylvania State University, University Park,
PA, USA*

M. Zhang
*Stony Brook University, Stony Brook,
NY, USA*

PREFACE TO THE FIRST EDITION

A half century ago the American Meteorological Society published the *Compendium of Meteorology*, which in a single volume of 1334 pages summarized the state of understanding of the atmosphere at that time. A perusal of the contents of that volume indicates that although a broad range of topics was covered, the vast bulk of the volume was devoted to traditional meteorological topics such as atmospheric dynamics, cloud physics, and weather forecasting. Barely 4 percent of the volume was devoted to articles related to atmospheric chemistry or air pollution and, of course, none of the volume was devoted to techniques such as satellites and remote sensing.

As Sir John Mason aptly notes in his foreword to the present work, the atmospheric sciences have expanded in scope enormously over the past 50 years. Topics such as atmospheric chemistry and global climate change, of only marginal interest 50 years ago, are now central disciplines within the atmospheric sciences. Increasingly, developing areas within the atmospheric sciences require students, teachers, and researchers to familiarize themselves with areas far outside their own specialties. This work is intended to satisfy the need for a convenient and accessible references source covering all aspects of atmospheric sciences. It is written at a level that allows undergraduate science and engineering students to understand the material, while providing active researchers with the latest information in the field.

More than 400 scientists, from academia, government, and industry have contributed to the 330 articles in this work. We are very grateful to these authors for their success in providing concise and authoritative summaries of complex subjects. As editors, we have benefited from the chance to learn from these articles, and we believe that all students and active scientists who want to increase their knowledge of the atmosphere will benefit enormously from access to this work.

We are also grateful to the 31 members of the Editorial Advisory Board who have guided us in our coverage of the very broad range of topics represented in this encyclopedia. Their willingness to suggest topics and authors, and to carefully review draft articles has contributed significantly to our success.

The production of this multivolume encyclopedia would not have been possible without the dedicated work of the staff of the Major Reference Works group at Academic Press. We are especially grateful to the Major Reference Work Development Manager, Colin McNeil, who has worked closely with us during the entire process. Finally, we appreciate the liberal use of color figures in the printed encyclopedia.

James R Holton, Judith A Curry, and John Pyle

This page intentionally left blank

PREFACE TO THE SECOND EDITION

Since the publication of the first edition of the *Encyclopedia of Atmospheric Sciences*, significant advances in research have been achieved all across the broad and expanding spectrum of the field and related disciplines. In particular, climate science with primary input from the atmospheric research emerges as a new field and integrator of interlocking peripheral disciplines over the last decade. These events have demanded the solicitation of new and updated articles for the 2003 edition. Some articles from the earlier publication were judged to be of such a fundamental and enduring nature that they did not require modification. But huge amounts of new information from Earth-orbiting satellite observatories have brought much new insight to the field. In addition there are new findings in many areas such as the latest simulations of meteorological and climatic processes of interest as well as simulations and observations of the composition and interaction of the field's chemical constituents. While interest in the ozone hole and its ramifications may have reached a plateau, ever more understanding of the stratosphere and its role in climate change emerges. The study of past climates provides new means of testing climate models and theories. In weather prediction we see new progress on how data are to be better assimilated for much improved initialization of the forecast model leading to the promise of more accurate predictions of severe weather and tropical cyclones over longer lead times. These are just a few of the new features of the second edition.

The editors of the second edition are greatly indebted to our predecessors in the first edition. They set the outline of topics and solicited the original authors, while establishing a high standard for the content of this publication. In many cases we decided to reprint those articles or request only minor updates. Nevertheless, many articles in this edition are entirely original, based on which we also made significant reorganization of the content. We are proud of our product and hope it provides the same assistance to students, researchers, and practitioners throughout the science and engineering communities.

Editors of the second edition
Gerald R North
Fuqing Zhang
John Pyle

This page intentionally left blank

EDITOR BIOGRAPHIES



Gerald R North received his PhD in theoretical physics from the University of Wisconsin in 1966. After postdoctoral research at the University of Pennsylvania he became a faculty member in physics at the University of Missouri–St. Louis. He shifted his research focus to climate science research during his sabbatical year at the National Center for Atmospheric Research, where he won the Outstanding Paper Award in 1975. He moved to NASA Goddard Space Flight Center (GSFC) in 1978 where he was awarded the NASA Medal for Research Excellence. During his stay at GSFC, he was the proposer and first study scientist for the Tropical Rainfall Measuring Mission, which was launched in 1997 and is still orbiting in 2014.

He moved to Texas A&M University in 1986 as a university distinguished professor of atmospheric sciences where he served as department head from 1995 to 2003. He has served as editor-in-chief of the *Reviews of Geophysics* and is recognized as one of the most cited authors in geosciences (Web of Science). He has chaired and/or served on a number of national committees and is a Fellow of the American Geophysical Union, American Meteorological Society (AMS) and the American Association for the Advancement of Science, and winner of the Jule Charney Award for Research (AMS).

He has published about 150 refereed papers not including many book chapters and reviews. His books include *Paleoclimatology*, co-authored with Thomas Crowley, and *An Introduction to Atmospheric Thermodynamics* co-authored with Tatiana Erikhimova.

North's interests are focused on the use of mathematical and statistical tools to solve climate problems over a wide range of issues including: analytical solutions of simplified energy balance climate models, use of random field techniques in representing and interpreting climate data and model simulations, detection of deterministic signals in climate change, statistical analysis satellite remote sensing for mission planning and analysis of data, paleoclimate problems using simplified climate models.



John Pyle obtained a BSc in Physics at Durham University before moving to Oxford where he completed a DPhil in Atmospheric Physics, helping to develop a numerical model for stratospheric ozone studies. After a short period at the Rutherford Appleton Laboratory he moved to a lectureship at Cambridge University in 1985. In 2000 he was appointed professor of atmospheric science and since 2007 has been the 1920 professor of physical chemistry. He is a Professorial Fellow at St Catharine's College. He has been a codirector of Natural Environment Research Council's National Centre for Atmospheric Science, where he is currently Chief Scientist.

His research focuses on the numerical modelling of atmospheric chemistry. Problems involving the interaction between chemistry and climate have been addressed; these range from stratospheric ozone depletion to the changing tropospheric oxidizing capacity and have included the environmental impact of aviation, land use change, biofuel technologies, and the hydrogen economy. He has studied palaeochemistry problems as well as the projected atmospheric composition changes during the current century. He has published more than 250 peer reviewed papers.

He played a major role in building an EU stratospheric research programme in the 1990s, coordinating several major field campaigns. He has contributed to all the WMO/UNEP assessments on stratospheric ozone since the early 1980s and is now one of the four international cochairmen

of the Scientific Assessment Panel, responsible for these assessments. He was a convening lead author in the IPCC Special report "Safeguarding the ozone layer and the global climate system," published in 2006.

He was elected Fellow of the Royal Society in 2004 and an American Geophysical Union Fellow in 2011. He was awarded the Cambridge ScD degree in 2012. Other honours and awards include membership of Academia Europaea (1993), Royal Society of Chemistry (Interdisciplinary award, 1991, and John Jeyes lectureship, 2008), and the Royal Meteorological Society Adrian Gill Prize, in 2004.



Fuqing Zhang is a professor with tenure in the Department of Meteorology at the Pennsylvania State University, with a joint appointment in the Department of Statistics, along with an endowed position as the E Willard & Ruby S Miller Faculty Fellow at the College of Earth and Mineral Sciences at the Pennsylvania State University. His research interests include atmospheric dynamics and predictability, data assimilation, ensemble forecasting, tropical cyclones, gravity waves, mountain plains and sea-breeze circulations, warm-season convection, and regional-scale climate. He earned his BS and MS in meteorology from Nanjing University, China in 1991 and 1994, respectively, and his PhD in atmospheric science in 2000 from North Carolina State University.

He spent seven years as an assistant and then associate professor at Texas A&M University before coming to Penn State University as a full professor in 2008. In 2000, he spent a year and a half as a postdoctoral fellow at the National Center for Atmospheric Research. He also held various visiting scholarship appointments at various academic and research institutions including the National Center for Atmospheric Research in Boulder, Colorado; the Navy Research Laboratory in Monterey, California; NOAA/AOML Hurricane Research Division, Miami, Florida; Peking University and Nanjing University, China; the Chinese State Key Laboratory of Severe Weather in Beijing, China; and Laboratoire de Meteorologie Dynamique, École Normale Supérieure in Paris, France.

He has authored/co-authored about 130 peer reviewed journal publications and has given more than 160 keynote speeches or invited talks at various institutions and meetings. He has served as principal investigator/co-principal investigator for 30 federal or state-sponsored research grants. He has chaired/cochaired more than 10 scientific meetings or workshops. He also served on various review or advisory panels for numerous organizations that include National Science Foundation, Office of Naval Research, NASA, NOAA, and National Academies. He has also served as editor of several professional journals including *Monthly Weather Review*, *Science China*, *Atmospheric Science Letter*, *Acta Meteorologica Sinica*, and *Computing in Science & Engineering*.

He has also received numerous awards for his research and service. Notably, in 2007 he received the **Outstanding Publication Award** from the National Center for Atmospheric Research. In 2009, was the sole recipient of the American Meteorological Society's 2009 **Clarence Leroy Meisinger Award** "for outstanding contributions to mesoscale dynamics, predictability, and ensemble data assimilation." Most recently, he received the 2014 American Meteorological Society's **Banner Miller Award** "for valuable insights into incorporating real-time airborne Doppler radar measurements via ensemble data assimilation, leading to improvements in forecasts of tropical cyclone track and intensity."

GUIDE TO USING THE ENCYCLOPEDIA

Structure of the Encyclopedia

The material in the encyclopedia is not arranged by ordinary alphabetical order, but by alphabetical order according to 49 principal topic areas taken to allow all papers belonging to each principal topic to appear together in the same volume. Within each principal subject, article headings are also arranged alphabetically, except where logic dictates otherwise. For example, overview articles appear at the beginning of a section.

There are four features that help you find the topic in which you are interested:

- i. the contents list
- ii. cross-references to other relevant articles within each article
- iii. a full subject index
- iv. contributors

i. Contents List

The contents list, which appears at the front of each volume, lists the entries in the order that they appear in the encyclopedia. It includes both the volume number and the page number of each entry.

ii. Cross-references

All of the entries in the encyclopedia have been cross-referenced. The cross-references, which appear at the end of an article as a See also list, serve four different functions:

- i. To draw the reader's attention to related material in other entries

- ii. To indicate material that broadens and extends the scope of the article
- iii. To indicate material that covers a topic in more depth
- iv. To direct readers to other articles by the same author(s)

Example

The following list of cross-references appears at the end of the article.

See also: **Biogeochemical Cycles:** Biogeochemistry of Iodine. **Stratospheric Chemistry Topics:** HO_x; Halogen Sources, Natural (Methyl Bromide and Related Gases); Halogens; Hydrogen Budget; Overview; Reactive Nitrogen (NO_x and NO_y).

iii. Index

The index includes page numbers for quick reference to the information you are looking for. The index entries differentiate between references to a whole article, a part of an article, and a table or figure.

iv. Contributors

At the start of each volume there is list of the authors who contributed to that volume.

This page intentionally left blank

ELECTRICITY IN THE ATMOSPHERE

Contents

Global Electrical Circuit

Ions in the Atmosphere

Lightning

Sprites

Global Electrical Circuit

ER Williams, Massachusetts Institute of Technology, Cambridge, MA, USA

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

The Earth–ionosphere cavity, composed of a thin insulator (the atmosphere) confined between two spherical conductors (the Earth and the ionosphere), provides the natural framework for both the DC and AC global circuits. The DC global circuit involves a nominal voltage drop of 250 kV between Earth and ionosphere. The AC global circuit is an electromagnetic wave phenomenon known as Schumann resonances. The historical discovery of these phenomena and their salient features is reviewed. The source currents for the DC and AC global circuits are different, but cloud-to-ground lightning is a source common to both. Accordingly, the variations in amplitude of the two global circuits are expected to be correlated but not identical. The value of monitoring the global circuit for climate change applications is emphasized, but robust measurement schemes need to be implemented.

Introduction

We live in a thin layer of air glued to the Earth's surface by gravity. This gaseous atmosphere is composed largely of neutral molecules of oxygen and nitrogen and as a consequence is an electrical insulator. The atmosphere is bathed in radiation – ultraviolet (UV) radiation from the Sun and cosmic radiation from deep space. This radiation ionizes the atmosphere and makes it a good electrical conductor at upper levels where the radiation is more energetic. The Earth beneath the atmosphere is abundant in liquid water. All water contains ions in solution, and the ions provide conduction. Three-quarters of the Earth's surface is covered with conductive seawater. Liquid water is also present virtually everywhere on land, permeates the cracks and joints within the Earth's crust, and makes landmasses electrical conductors too. The thin layer of insulating air sandwiched between these two conductors forms the medium for the global electrical circuit. For the so-called 'DC' global circuit, this medium is a giant spherical capacitor. For the 'AC' global circuit, otherwise called the Schumann resonances, the medium is an electromagnetic waveguide.

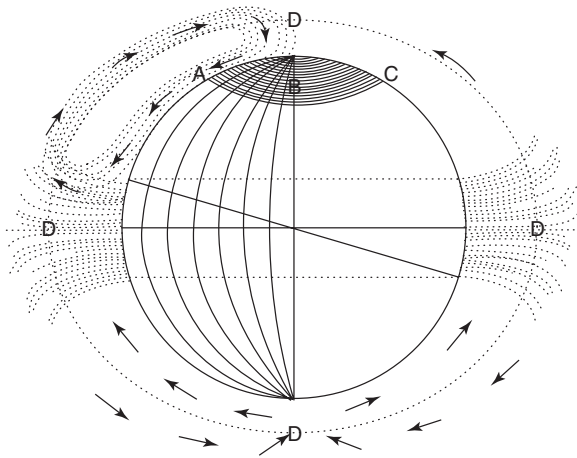
Historical Development

The development of ideas on the DC global circuit received great impetus from three giants of research in atmospheric

electricity: Benjamin Franklin, William Thomson (Lord Kelvin), and C.T.R. Wilson. Their three contributions, which dominated each of three successive centuries, are discussed in turn.

Franklin can be credited with the first enunciation of a global flow of moist, electrified air. His concept is best discussed around his own picture (**Figure 1**), as presented to the Royal Academy of Sciences in Paris in 1779. Warm, moist air ascends in the tropics and descends in the polar regions. This cloudy air was believed to deliver electricity to the cold polar ice cap on snow, where it would accumulate until breakdown of the rarefied upper atmosphere occurred in the form of the aurora. While this explanation for the aurora is now known to be flawed, the postulated role for the tropics and the picture of the large-scale circulation of the atmosphere were clearly prescient.

Lord Kelvin, 100 years later developed potential theory, a mathematical tool needed for theoretically underpinning the global circuit. Remarkably, more than 40 years before the conductive ionosphere was postulated by A. Kennelly and O. Heaviside in 1902, Kelvin advanced the spherical capacitor picture for the global circuit. His expectation for an outer conductor was based on his knowledge that rarefied air of the upper atmosphere was a poor insulator in comparison with air at the Earth's surface. He also advocated organized measurements of the Earth's electric field, and this suggestion undoubtedly motivated subsequent electrical observations from the research vessels *Carnegie* and *Maude* by the Carnegie



*The Arrows represent the general Currents of the Air
ABC, the great Ice & Snow in the Polar Regions.
D.D.D.D. the Medium Height of the Atmosphere.
The Representation is made only for one Quarter and one
Meridian of the Globe; but is to be understood the same
for all the rest.*

Figure 1 Benjamin Franklin's picture of airflow and transport of electricity from equatorial to polar regions. Reproduced with permission from Silverman, S., 1970. Franklin's theory of the aurora. *Journal of the Franklin Institute* 290, 177.

Institution. Kelvin undertook his own surface measurements of potential gradient, verified that the Earth carried a negative charge, and concluded that the global circuit peaked in winter, a result now believed to be dominated by local effects. He also prophesied the use of electrical measurements for purposes of weather prediction.

There can be no doubt that electric indications, when sufficiently studied, will be found important additions to our means for prognosticating the weather and the speaker hoped soon to see the atmospheric electrometer generally adopted as a useful and convenient weather glass.

Measurements by Wilson of the field changes associated with lightning in thunderclouds led him to conclude that the polarity of thunderclouds was systematically positive in upper levels and negative at lower levels. Wilson was also engaged with measurements of the currents flowing to Earth during periods of fair weather. The observation of the transatlantic propagation of radio waves in 1903 by G. Marconi verified the presence of the conductive ionosphere. This collective information led Wilson in 1920 to formulate his famous hypothesis for the global electrical circuit: thunderstorms are batteries and drive current upward to the conductive ionosphere where it spreads out to return to Earth in fair weather regions, as illustrated in Figure 2. Wilson's idea led F. J. W. Whipple to compare the universal time (UT) variation of electric field over the oceans, now referred to as the 'Carnegie curve,' with the UT diurnal variation of thunder areas on a global basis, as shown in Figure 3. Three major tropical continental zones are activated sequentially by the surface heating associated with the passage of the Sun. The similarity in phase between these two curves has long stood as key

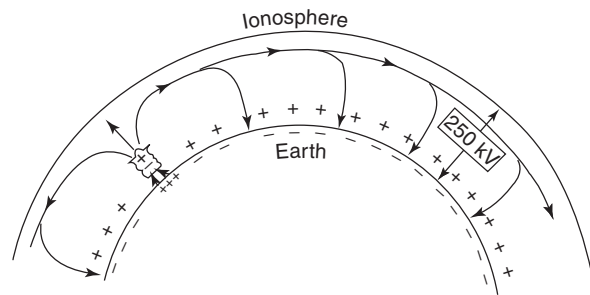


Figure 2 Simple illustration of the operation of the DC global circuit with electrified clouds as generators and a return current to earth in fair weather regions.

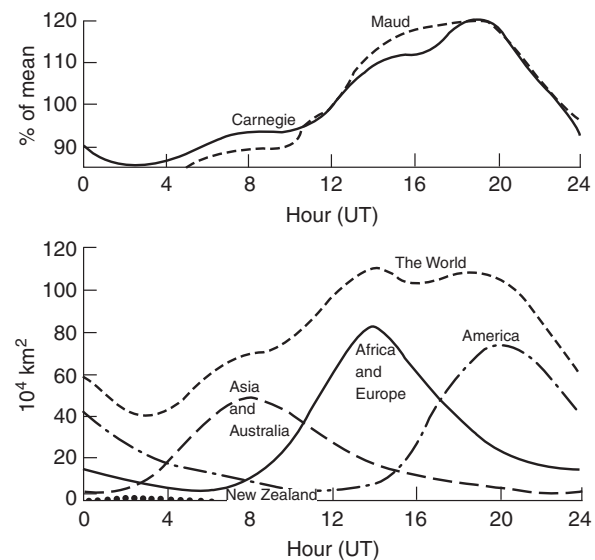


Figure 3 Comparisons of the UT variation of electric field over the ocean (the Carnegie curve) and the UT variation of thunder areas world-wide. Adapted from Whipple, J.N., 1929. On the association of the diurnal variation of electric potential gradient in fine weather with the distribution of thunderstorms over the globe. *Quarterly Journal of the Royal Meteorological Society* 55, 1–17.

substantiating evidence for the Wilson's global circuit hypothesis. Additional support came in 1950 when O. Gish and G. Wait measured upward currents over thunderstorms from an airplane. More recent measurements have shown upward currents over showerclouds.

Wilson's student T.W. Wormell later extended the surface measurements of current and formulated a statistical charge 'balance sheet' for the global circuit. These results showed that point discharge current dominated over the lightning current in modulating the negative charge transfer to the Earth's surface by electrified storms.

The first coordinated measurements of the global circuit were made by R. Muhleisen, one set from Germany and another from a ship in the Atlantic Ocean. These balloon measurements integrated the vertical electric field in the

atmosphere to provide the so-called ionospheric potential of the global circuit. The simultaneous soundings at two locations agreed to within 5% in three-quarters of the measurements, providing considerable support for the worldwide nature of the global circuit response. Extensive measurements of the DC circuit were also carried out by R. Markson using instrumented aircraft and with balloons from stations in Massachusetts and Australia with similar correlated results, and with diurnal variations that closely follow the classical Carnegie curve.

The first suggestion that the spherical capacitor of the DC global circuit also served as an electromagnetic waveguide appeared when in 1952 W. Schumann postulated the existence of electromagnetic resonances maintained by global lightning activity. Partial experimental verification of the resonances was obtained by Schumann's student H. Koenig in Munich in the late 1950s. M. Balser and C. Wagner of the MIT Lincoln Laboratory verified the multimodal resonances with the first spectral measurements in 1960. Transient excitations of the Schumann resonances by single extraordinarily energetic flashes were reported in the early 1970s by D.L. Jones and his colleagues.

Structure and Operation of the DC Global Circuit

The conductivity structure of the giant spherical capacitor and electromagnetic waveguide is established by an interplay between ionization from the Sun and from deep space, and the exponential decline of air density with altitude above the Earth's surface. The principal radiation components in this context are the galactic cosmic radiation and the UV and X-ray photons from the Sun. The cosmic radiation is the most energetic and serves to dominate the ionization and electrical conductivity at altitudes from 0 to 50 km. The cosmic radiation is also largely isotropic and so the conductivity structure of the lower atmosphere is spherically symmetrical. The UV radiation from the Sun is largely removed in causing ionization at higher altitudes – in the 'D' and 'E' regions of the ionosphere – and since this radiation is strongly anisotropic, the conductivity structure is endowed with a modest asymmetry between the daytime and the nighttime hemispheres. As a result, the electromagnetic waveguide departs slightly from spherical symmetry.

In treating the DC global circuit, we are concerned primarily with the spherically symmetric lowest region of the atmosphere, where small ions are the dominant charge carrier. Over the oceans, where the air is relatively free of aerosol particles, the altitude (z) dependence of the electrical conductivity can be approximated by the exponential function

$$\sigma(z) = \sigma_0 \exp(z/z_0) \quad [1]$$

where $\sigma_0 = 5 \times 10^{-14} \text{ S m}^{-1}$, and the e-folding scale height $z_0 = 5.0 \text{ km}$. The minimum atmospheric conductivity is found at the Earth's surface and contrasts markedly with conductivity values for the oceans and typical surface crustal material. This information establishes quantitatively the medium of the global circuit – a spherical capacitor with variable resistivity among good conductors.

Current sources for the DC global circuit are in principle any mechanism that separates positive from negative charge vertically between the conductive Earth and the upper atmosphere.

Such mechanisms include the vertical transport of electric space charge in the planetary boundary layer, the gravity-driven descent of selectively charged aerosol particles in the atmosphere, and the separation of electric charge by microphysical processes involving ice in showerclouds and thunderclouds. The Wilson hypothesis postulates the special subset of thunderstorms as the dominant current sources, or batteries, for the global circuit, though much evidence has appeared that showerclouds without lightning are also major contributors. The Wilson hypothesis establishes a firm link between the polarity of charge on the Earth and the polarity of thunderclouds, but does not account for the polarity of either entity. Today, it seems likely that this polarity is determined by some fundamental property of ice, still to be identified.

The behavior of current sources in a medium whose conductivity structure is exponential in the vertical has been examined mathematically. The solution for a point source in an exponential medium infinite in the vertical is for all current to flow upward toward higher conductivity. When horizontal conductor boundaries and the associated image sources are added, some current flows locally to Earth with the remaining flowing upward toward larger conductivity. Consistently with these theoretical calculations, observations over thunderstorms do show systematic upward current flow to higher altitudes, in line with the Wilson hypothesis. In a closed global circuit, the integrated upward current from globally distributed sources, approximately 1 kA, must ultimately return to Earth, as illustrated in Figure 2. On account of the spherical uniformity of conductivity structure, the current flow will also be spherically uniform in the resistive lower atmosphere. According to Ohm's law, this uniform integrated current density, J_0 , will set up a vertical electric field given by

$$E(z) = \frac{J_0}{\sigma(z)} = \frac{J_0}{\sigma_0 \exp(-z/z_0)} \quad [2]$$

giving an electric field at the Earth's surface of 40 V m^{-1} which declines exponentially with increasing altitude. The total voltage drop across the giant spherical capacitor will then be given as the line integral of this vertical electric field (where one-dimensional coordinates are used rather than spherical coordinates, as the vertical extent of the Earth-ionosphere cavity is small in comparison with the radius of the Earth, R_E):

$$V_1 = \int E(z) dz = \frac{J_0 z_0}{\sigma_0} \quad [3]$$

While the limits of field integration extend from the Earth's surface to infinity, the 5-km scale height for conductivity and electric field indicate that in three scale heights the field is reduced to 5% of its surface value and hence more than 95% of the voltage drop is achieved at 40 km altitude. The integrated potential difference between the Earth and the upper atmosphere is called 'ionospheric potential.'

The numerical evaluation of V_1 in the simplified treatment of the global circuit includes the experimentally determined current density in fair weather, J_0 ($2 \times 10^{-12} \text{ A m}^{-2}$), the conductivity scale height z_0 (5 km), and the (extrapolated) conductivity of air at the surface ($5 \times 10^{-14} \text{ S m}^{-1}$). The resulting V_1 value from eqn [3] is 200 kV, which is somewhat less than the value of actual measurements (mean value 250 kV). The primary reason for this discrepancy is the neglect

of the aerosol-laden planetary boundary layer, the shallow (~ 1000 m) zone near the surface in which the conductivity departs from the simple exponential form given earlier in eqn [1]. The local complications of the planetary boundary layer have been the major obstacle to the measurement of the global circuit from the Earth's surface. Measurements at sea, where the boundary layer is substantially less polluted, have shown records similar to the Carnegie curve on individual days, but even at sea problems can arise.

The steady-state distribution of electric space charge within the resistive lower atmosphere of the spherical capacitor may be determined from Poisson's equation,

$$\rho(z) = \epsilon_0 \frac{dE}{dz} = -\frac{\epsilon_0 J_0}{\sigma_0 z_0} \exp\left(\frac{-z}{z_0}\right) \quad [4]$$

where ϵ_0 is the permittivity of free space. According to these simple electrostatic predictions, positive space charge is distributed throughout the atmosphere, with maximum value in the lowest part of the atmosphere, as indicated schematically in Figure 2. An equal and opposite negative charge resides on the Earth's surface, with charge density per unit area given by

$$S = \epsilon_0 E(z = 0) = \frac{\epsilon_0 J_0}{\sigma_0} \quad [5]$$

and with total negative charge on the Earth

$$Q_{\text{tot}} = 4\pi R_E^2 S = \frac{4\pi R_E^2 \epsilon_0 J_0}{\sigma_0} = 2 \times 10^5 \text{ C} \quad [6]$$

Note that this charge is large in comparison with the charge transferred by single lightning discharges (~ 10 C), which as a consequence have no discernible effect on the steady-state V_1 values or the surface electric field. The capacitance of the giant spherical capacitor is, by definition,

$$C = \frac{Q_{\text{tot}}}{V_1} \quad [7]$$

whose value is approaching 1 F. The integrated resistance of the global circuit can also be computed analytically, again ignoring the contribution of the resistive and highly variable planetary boundary layer, which is most prevalent over land areas. This global resistance is most easily obtained directly from Ohm's law as

$$R = \frac{V_1}{I_{\text{tot}}} = \frac{V_1}{4\pi R_E^2 J_0} = 200 \Omega \quad [8]$$

The time required for the voltage across the giant capacitor to decay if all sources suddenly ceased is the RC relaxation time:

$$RC = (200 \Omega)(1 \text{ farad}) = 200 \text{ s} \quad [9]$$

Using typical parameters, this gives a relaxation time of 200 s or about 3 min. This time is somewhat longer if the effect of the resistive planetary boundary layer is included.

Schumann Resonances

The AC global circuit shares the same Earth-ionosphere medium as the DC global circuit but is a substantially richer and more complex phenomenon. Electromagnetic standing

waves within the Earth-ionosphere cavity, excited by lightning flashes worldwide, are known as Schumann resonances. The resonant frequencies are determined by the circumference of the natural waveguide and by the speed of light within the waveguide. The fundamental resonant frequency is 8 Hz.

The effective ionospheric height for the AC global circuit is systematically greater than for the DC circuit. The propagation of electromagnetic waves requires the displacement current to exceed conduction currents; the latter currents damp the waves. A meaningful estimate of waveguide height is found by equating the displacement current and the conduction current

$$J_D = J_C \quad [10]$$

$$\omega \epsilon_0 E = \sigma E \quad [11]$$

$$\omega \epsilon_0 = \sigma(z) \quad [12]$$

The height dependence of electrical conductivity is given by eqn [1], and so for a frequency of 8 Hz this condition can be solved for a height, which is approximately 50 km. This value is approximately 10 conductivity scale heights and hence is significantly greater than the height needed to achieve the ionospheric potential [3] of the DC global circuit.

The application of Maxwell's equations to a thin spherical waveguide (assumed uniform for simplicity of treatment) leads to the normal mode equations for the electric and magnetic fields resulting from a single vertical lightning discharge with frequency-dependent current moment $I dS(\omega)$, where I is the current in amperes and dS is the vertical extent of the current-carrying channel. The role of the ionosphere is treated with the complex eigenvalue ν , which is frequency dependent. The parameter a is the radius of the Earth and h is the height of the waveguide. The P_ν^0 and P_ν^1 are complex Legendre polynomials and their derivatives, respectively. The distance between the lightning source and an observer along a great circle path is given by θ .

$$E(\omega) = \frac{I(\omega) dS \nu(\nu+1) P_\nu^0(-\cos \theta)}{4R_E^2 \epsilon_0 \omega h \sin(\pi \nu)} \quad [13]$$

measured in $\text{V m}^{-1} \text{ Hz}^{-1}$

$$H(\omega) = \frac{I(\omega) dS P_\nu^1(-\cos \theta)}{4R_E h \sin(\pi \nu)} \quad [14]$$

measured in $\text{A m}^{-1} \text{ Hz}^{-1}$

These equations have been successfully merged with measured electromagnetic fields to construct global maps of extraordinarily energetic lightning flashes, whose fields dominate all other lightning on the planet for periods of several hundred milliseconds. Such maps strongly resemble the maps produced by satellite observations in Figure 4 in showing strong continental dominance.

The majority of lightning flashes are not sufficiently energetic to stand out above all other events, and the electromagnetic fields of this far larger population of ordinary flashes superimpose to produce the quasi-steady 'background' resonances. The phase information for the integrated activity is lost and the observations are normally documented as power spectra of electric and magnetic fields. Sample spectra, integrated for a complete 24-h period for three field components

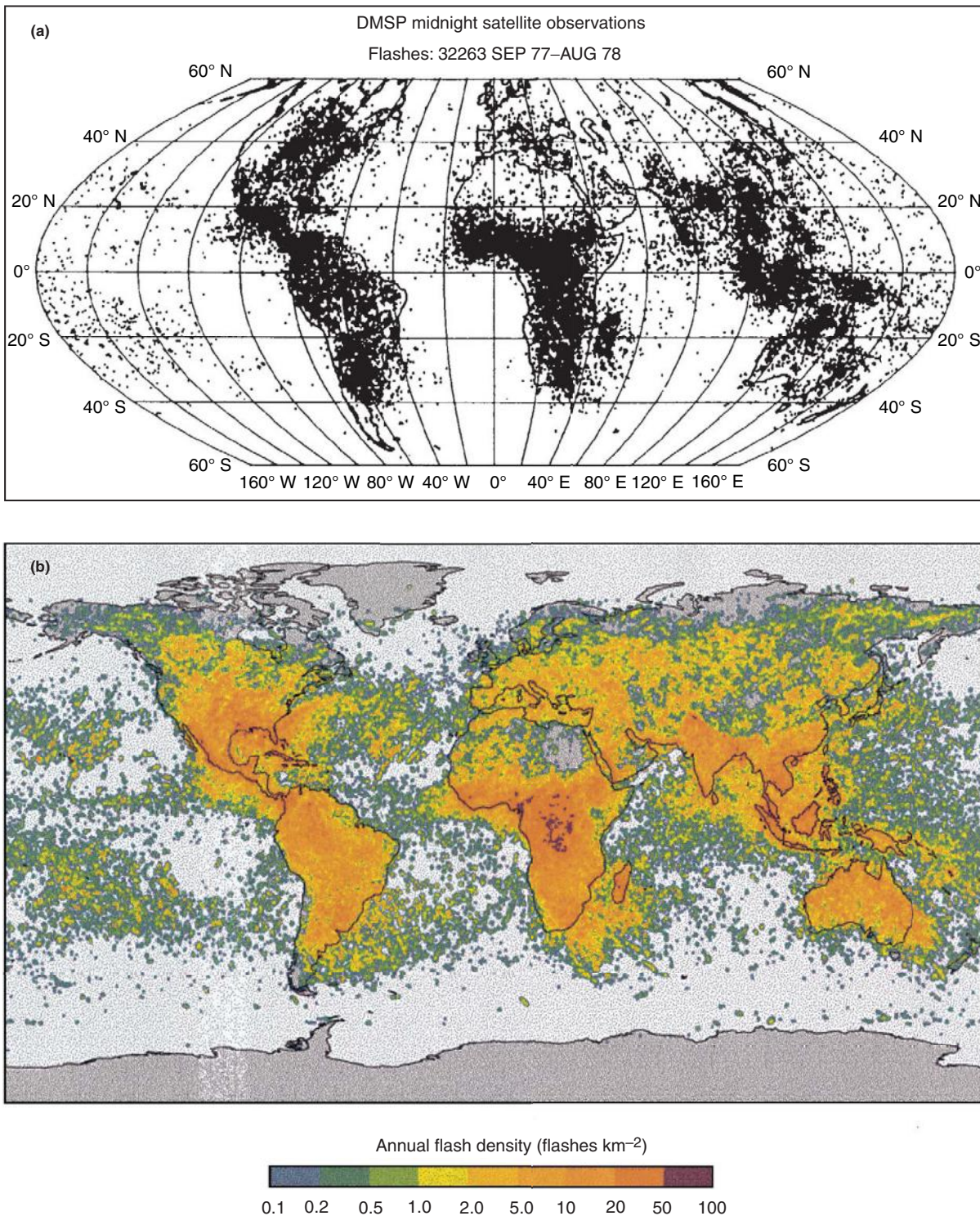


Figure 4 Schumann resonance power spectra in the vertical electric (top panel) and two magnetic field components (middle and bottom panel), as measured in West Greenwich, Rhode Island, on 1 January 2000. The fundamental mode near 8 Hz is dominant in all spectra. Four to five higher order modes are also discernible.

(vertical electric, north–south magnetic, and east–west magnetic), are shown in **Figure 5**, where well-defined peaks at 8 Hz for the fundamental mode are seen as well as higher resonant frequencies at 14, 20, 26, and 32 Hz. The widths of the various spectral peaks are indicative of the quality factor Q of the Earth–ionosphere cavity. One definition of Q is

$$Q = \frac{f_0}{\Delta f} \quad [15]$$

where f_0 is the center frequency and Δf is the frequency interval between half-power points of the resonance curve. Dimensionless Q values are typically in the range of 3–8, indicating that the Schumann resonances are not sharply tuned owing to the lossy nature of the ionosphere.

The background Schumann resonances are more difficult to interpret for lightning source properties (location and intensity), because the sources are less localized than for single events and because lightning near the receiver can dominate the signal and mask more distant activity. Nevertheless, because of the overwhelming dominance of lightning in continental zones (**Figure 4**), even the sources for the background resonances exhibit considerable localization. A modified form of eqns [13] and [14] has been used to infer the

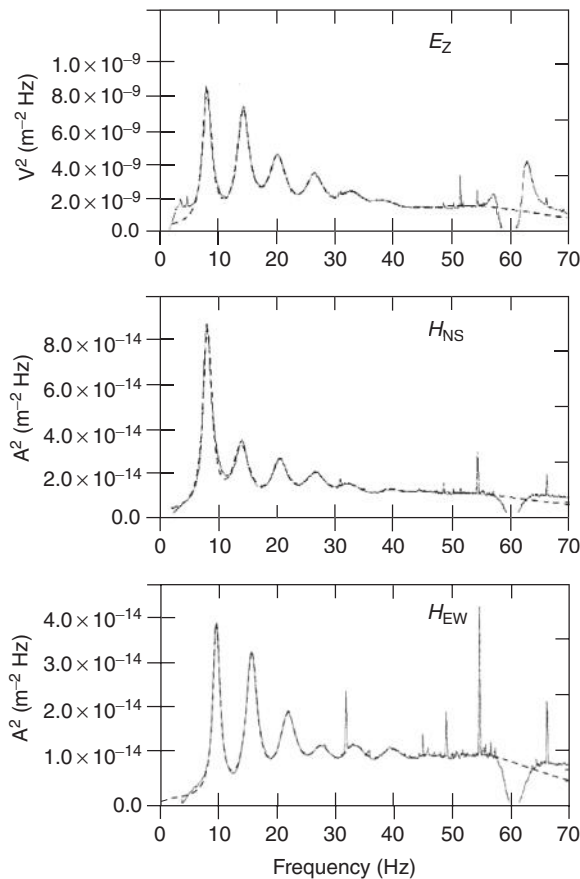


Figure 5 Global distribution of lightning based on (a) optical observations with the DMSP satellite at local midnight (Orville and Henderson, 1986), and (b) optical observations during both day and night with the Optical Transient Detector. Courtesy NASA Marshall Space Flight Center.

global vertical dipole moment change squared per unit time based on observations of the electric and magnetic power spectra. These measurements are broadly consistent with earlier estimates of 100 s^{-1} for the global lightning flash rate.

Comparative Response of the DC and AC Global Circuits

The first three modes of the electric field of the global circuit are illustrated within the Earth–ionosphere cavity in **Figure 6**. Mathematically speaking, the DC global circuit is the zeroth-order mode of Schumann resonances, the AC global circuit. On this basis, one might expect the ionospheric potential of the DC circuit and the global charge moment squared per unit time of the AC global circuit to be highly correlated on all timescales. While high-quality comparisons of simultaneous behavior are scarce, the available evidence does not support this prediction. A likely reason for different behavior lies with the sources, which drive the two global circuits, and the different meteorological origins of these sources. **Table 1** summarizes the principal source currents for the DC and AC global circuits. Charge transfer by point discharge current, by falling precipitation, and by lightning are all deemed important for the DC global circuit, whereas only lightning is believed to contribute substantially to Schumann resonances.

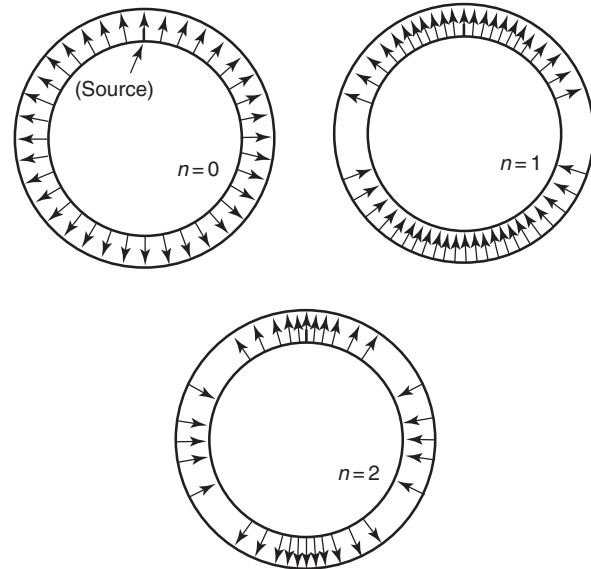


Figure 6 Illustrations of the behavior of the vertical electric field for the first three modes of the Earth–ionosphere cavity. The zeroth-order mode is the DC global circuit. Adapted from Wait, J.R., 1996. *Electromagnetic Waves in Stratified Media*, IEEE Press, Piscataway, NJ.

Table 1 Dominant current sources for the global circuit

DC circuit	AC circuit
Point discharge current	Lightning current
Precipitation current	
Lightning current	
Boundary layer convection	

Historically, the behavior of the diurnal variation of the DC global circuit shown in [Figure 3](#) is viewed as the cornerstone of atmospheric electricity and a critical test of the Wilson hypothesis that thunderstorms sustain the source currents. According to the simplified earlier development of ideas, a diurnal variation in integrated current J_0 will manifest itself as a diurnal variation in V_1 and in the surface electric field $E(z=0)$ on a global basis. As noted earlier, the phase agreement between the thunder day variations and the Carnegie curve is quite good, but the amplitude variation of the proxy lightning record is more than twice that of the Carnegie curve. Three major tropical storm regions are represented in the thunder day data in [Figure 3](#) as separate peaks but such clean distinction is not apparent in the Carnegie curve. These results strongly suggest that lightning is not the main contributor to the DC global circuit. Consistent with this conclusion is the experimental evidence that the local diurnal variation of point discharge current, rainfall, and rain current are all substantially smaller than the local diurnal variation of lightning. Since point discharge currents, rain currents, and boundary layer convective charge transport are present in weather systems other than thunderstorms, it would appear that a larger collection of weather elements than thunderstorms alone is needed to account for the modulations of the DC global circuit.

Further evidence for diverging behavior between the DC and AC global circuits comes from quantitative comparisons of the contributions from each of the three major tropical zones with the diurnal variation in UT ([Figure 3](#)). Africa is frequently the dominant lightning producer both in the classical thunder day analysis and in numerous observations of lightning from satellites in space ([Figure 4](#)). By contrast, the Carnegie curve peaks at 19.00 UT when the Americas are most active, suggesting that the Americas are the dominant tropical contributor to the DC global circuit.

The annual variation of the global circuit is more thoroughly documented for the AC component, because of numerous available observations of global lightning, and shows a maximum in Northern Hemisphere summer. The comparisons also indicate a similar result to that on the diurnal timescale: the annual variation of lightning (nearly a factor of two) is substantially larger than the annual variation of surface electric field and ionospheric potential.

Applications of the Global Circuit

A Monitor for Global Change

The verification and physical understanding of global change has become a topic of enduring interest worldwide. The connection between a currently increasing CO_2 concentration and an increasing mean global temperature is controversial and still unresolved. The global circuit provides a unique natural framework for monitoring global integrals: the collective current from electrified convection and the total lightning activity. The expectation that global circuit integrals are responsive to temperature is based on well-recognized local behavior. For example, lightning is more likely in the hot afternoon than during cool nights. At midlatitude locations, lightning is more frequent in the hot summer than in the cold winter. In the tropics, lightning is more frequent during warm

equinox than during cooler solstice. At all latitudes, lightning is more common over warmer land than over cooler ocean. In general, in the present climate, one finds greater cloud electrification and higher lightning flash rates in regions characterized by stronger conditional instability, larger CAPE (convective available potential energy), and larger updraft speeds. In the tropics, such regions are characterized by large wet-bulb potential temperature, a thermodynamic quantity representing both temperature and moisture.

With the aim of understanding global circuit response to temperature change, a general strategy in recent years has been to examine global circuit behavior on timescales for which the temperature variations and the underlying forcings are reasonably well established. Substantial correlations between global circuit behavior and the underlying global meteorology (which is often temperature related) have now been identified on the diurnal, the 5-day, the 25-day, the 30- to 60-day Madden-Julian oscillation, the semiannual, the annual, and the interannual (El Niño Southern Oscillation) timescales. The global circuit sensitivity to surface air temperature on these various timescales is generally in the range of 10–100% per $^{\circ}\text{C}$. For many of these studies, the determination of a quantitative sensitivity to temperature has been hindered by the poor sampling of temperature in regions dominated by deep convective activity (i.e., tropical South America and Africa).

The response of the global circuit to a long-term increase in temperature associated with global warming remains an open question. Measurements of the global circuit have been underway for only 50 years and very sporadically. The Schumann resonance records are substantially shorter. These records do not show significant upward trends. Results from a general circulation model point to an increase in lightning, but the key quantities for lightning in the models, like cloud buoyancy and convective scale updraft strength, are poorly resolved. If CAPE is invariant on the long timescale, the only change to be expected is the increase in moisture with temperature following the fundamental Clausius-Clapeyron relation. This dependence is approximately 6% per Kelvin and is substantially smaller than the global circuit sensitivity to temperature change documented on shorter timescales. Further work and longer records are required to discern a meaningful long-term change.

The DC global circuit also affords an opportunity to investigate changes in boundary layer pollution on a global basis. The strategy would be to distinguish variations in V_1 caused by source current from variations in V_1 caused by global circuit resistance. The latter quantity is influenced by the resistance of the planetary boundary layer, which is in turn influenced by the aerosol population that removes small ions and reduces the local electrical conductivity. Simultaneous measurements of the ionospheric potential V_1 and numerous single-station measurements of the air-earth current density would suffice to compute a global electrical resistance value (recall [eqn \[8\]](#)) whose changes could then be monitored. This experiment has not yet been performed, but the requisite observational methods have been established and verified.

Mesospheric Discharge Phenomena

Optical phenomena in the mesosphere high above large thunderstorms (mesoscale convective systems) have been

firmly documented within the last decade, and named sprites, elves, and trolls. All of these phenomena appear to be accompanied by a high-amplitude ‘ringing’ of the Schumann resonances by single-energetic cloud-to-ground lightnings. The application of the normal mode eqns [13] and [14] to these isolated transients enables the location and vertical charge moments to be extracted on the basis of ELF (extremely low frequency) global circuit measurements. Such observations appear to verify Wilson’s speculations in the 1920s that a sufficiently large dipole moment change by lightning in the lower atmosphere can cause a field increase in the upper atmosphere large enough for dielectric breakdown and ensuing optical emission. Global maps of event locations can be prepared on the basis of ELF measurements from a single location. Such maps are proxy maps for sprites.

Curiously, the great majority of flashes that cause upper atmospheric discharges are ground flashes with positive polarity, the opposite of the far more common negative ground flash that recharges the Earth negatively. Studies of the giant positive discharges with Schumann resonance methods have shown that they exhibit larger current and transfer greater amounts of charge than their negative counterparts. It is not known whether this asymmetry is related primarily to local meteorology and differences in the nature of the local charge reservoirs for the two lightning types, or whether a contribution arises from the fact that negative flashes are charging the DC global circuit and positive flashes are discharging it.

Diagnostic for the D Region of the Ionosphere

The upper conductive boundaries of both the DC and AC global circuits lie in the lowermost regions of the ionosphere, where the electron density is some four orders of magnitude smaller than its daytime maximum near 300-km altitude. Even large-aperture radars are insufficiently sensitive to detect and monitor the lower D region. Measurements *in situ* by rockets are limited in space and time. Schumann resonance methods may provide a useful global diagnostic for ionization in this region. The parameters of greatest interest are the Schumann resonance frequencies, determined by phase velocities of the resonant waves, and Q values, both of which depend on conductivity profiles that change with modulations in ionizing radiation. Recent results have shown systematic increases in the resonant modal frequencies over the recent solar cycle that are attributable primarily to ionization increases in the 80–90 km altitude range.

Conclusion

Renewed interest in the global electrical circuit within the last decade has spurred new methods for measurement, the initiation of dedicated monitoring programs, and a greatly improved understanding of the global circuit’s relationship with meteorology and ionospheric physics on many timescales.

The DC global circuit is difficult to measure on a continuous basis, but possesses a global invariant, ionospheric potential, that is well defined and quantitatively accessible. The Schumann resonances are relatively easy to measure on a continuous basis and are insensitive to the local variations in the planetary boundary layer, but the global invariant is more difficult to extract from single-station records. Coordinated measurements with both DC and AC aspects are most likely to bear the greatest fruit.

See also: Chemistry of the Atmosphere: Chemical Kinetics; Principles of Chemical Change. **Electricity in the Atmosphere:** Sprites. **Middle Atmosphere:** Quasi-Biennial Oscillation; Semiannual Oscillation. **Tropical Meteorology and Climate:** Intertropical Convergence Zone.

Further Reading

- Adlerman, E.J., Williams, E.R., 1996. Seasonal variations of the global electrical circuit. *Journal of Geophysical Research* 101, 29679–29688.
- Bering, E.A., July 1997. The global circuit: global thermometer, weather by-product, or climate modulator. *Reviews in Geophysics (Suppl.)*, 845–862.
- Boccippio, D., Williams, E., Heckman, S., Lyons, W., Baker, I., Boldi, R., 1995. Sprites, ELF transients and positive ground strokes. *Science* 269, 1088–1096.
- Holzer, R.E., Saxon, D.S., 1952. Distribution of electrical conduction currents in the vicinity of thunderstorms. *Journal of Geophysical Research* 57, 207–216.
- Markson, R., 1985. Aircraft measurements of the atmospheric electrical global circuit during the period 1971–1984. *Journal of Geophysical Research* 90, 5967–5977.
- Muhleisen, R., 1977. The global circuit and its parameters. In: Dolezalek, H., Reiter, R. (Eds.), *Electrical Processes in Atmospheres*. Steinkopf, Darmstadt, pp. 467–476.
- Polk, C., 1982. Schumann resonances. In: Volland, H. (Ed.), *Handbook of Atmospherics*. CRC Press, Boca Raton, FL, pp. 112–178.
- Price, C., Rind, D., 1993. Modeling global lightning distributions in a general circulation model. *Monthly Weather Review* 122, 1930–1939.
- Satori, G., Zieger, B., 1996. Spectral characteristics of Schumann resonances observed in Central Europe. *Journal of Geophysical Research* 101, 29663–29669.
- Wait, J.R., 1996. *Electromagnetic Waves in Stratified Media*. IEEE Press, Piscataway, NJ.
- Whipple, F.J.W., 1929. On the association of the diurnal variation of electric potential gradient in fine weather with the distribution of thunderstorms over the globe. *Quarterly Journal of the Royal Meteorological Society* 55, 1–17.
- Williams, E.R., 1992. The Schumann resonance: a global tropical thermometer. *Science* 256, 1184.
- Williams, E.R., Heckman, S.J., 1993. The local diurnal variation of cloud electrification and the global diurnal variation of negative charge on the earth. *Journal of Geophysical Research* 98, 5221.
- Williams, R., 1999. Global circuit response to temperature on distinct timescales: a status report. In: Hayakawa, M. (Ed.), *Atmospheric and Ionospheric Phenomena Associated with Earthquakes*. Tena, Tokyo, pp. 939–949.
- Wilson, C.T.R., 1920. Investigations on lightning discharges and the electric field of thunderstorms. *Philosophical Transactions of the Royal meteorological Society, Series A* 221, 73–115.

Relevant Website

<http://fallmeeting.agu.org/2012/events/franklin-lecture-ae31a-lightning-and-climate-video-on-demand/> – Williams, E.R., 2012. Franklin Lecture.

Ions in the Atmosphere

KL Aplin, University of Oxford, Oxford, UK

RG Harrison, The University of Reading, Reading, UK

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

Earth's atmosphere, like all planetary atmospheres, is constantly ionized by cosmic rays. Close to the surface, natural radioactivity is the dominant source of ionization. This article describes the generation and physical characteristics of atmospheric cluster-ions, including their size, composition, transport, and electrical properties, and gives an overview of relevant measurement technologies. The role of cluster-ions in weather and climate through modulating atmospheric particle formation and a range of other electrical effects is summarized. Nomenclature and categorization of atmospheric cluster-ions, and the properties of ions in the indoor environment are also briefly discussed.

Introduction to Atmospheric Ions

In Earth's atmosphere, an ion is a cluster of molecules carrying an overall charge, known as a *molecular cluster ion*. Such cluster ions, with dimensions of approximately 1 nm, have usually been referred to as *small ions*, and their motion in air constitutes a small electric current. *Large ions* (or Langevin ions), by comparison, are physically larger (tens to hundreds of nanometers) and consequently electrically less mobile. Usage of the term 'ion' to represent these molecular clusters originates from the early history of atmospheric electricity, which spans the discovery of the electron and the elucidation of the structure of matter. The distinction between large and small ions originates from distinguishing ions that could be accelerated by atmospheric electric fields (and therefore directly contribute to the conductivity of air), and those (the large ions) which were insufficiently electrically mobile to contribute to electrical conduction in air.

Particle Spectrum

Modern measurements have identified ions as part of the atmospheric aerosol spectrum. Aerosol particles span a wide range of sizes in the atmosphere, with dimensions from a few nanometers to tens of microns, and the smaller particles are more numerous. Large ions are effectively submicron aerosol particles, which have acquired electrical charges (Table 1). Many of the smaller aerosol particles are capable of acting as nuclei for trace gas condensation, depending on the supersaturation, but an important small fraction can act as cloud condensation nuclei (CCN).

Table 1 Typical properties of atmospheric particles and droplets

Particle	Radius (μm)	Mobility ($10^{-6} \text{ m}^2 \text{ V}^{-1} \text{ s}^{-1}$)
Small ions	0.005–0.008	0.5–100
Large ions	0.008–0.03	0.05–0.5
Aerosol	0.003–30	
Cloud droplets	2–30	
Rain drops	30–3000	

Ion Formation

Small ions form by the interaction between energetic particles, from natural radioactivity and cosmic rays, with air molecules. The atomic ions and free electrons formed by this radiolysis are highly polarizing, and clustering ligands rapidly attach to them by hydrogen bonding. Large ions are formed by collisions between preexisting aerosol and small ions, which lead to the small ion becoming attached to the larger particle. The contribution of the small ion's mass to the combined particle is insignificant, but the attachment process transfers charge to the larger particle. Further collisions between small and large ions can occur, generating multiply charged large ions, or neutral particles if the combining ions have opposite polarities. Interactions between small and large ions occur from kinetics, and electrical forces only become significant for highly charged large ions. Collisions between aerosol and small ions serve to remove small ions: in air containing large quantities of aerosol the small ion content becomes correspondingly small. The removal of ions by aerosol is the basis for one type of domestic smoke alarm. This also means that, although the conductivity of air is principally due to its small ion content, the conductivity is inversely influenced by large ions and aerosol through their removal of small ions.

Chemical Composition

Atmospheric small ions consist of clusters of molecules collected around a singly charged ion. Clustering occurs within microseconds of ionization, and the ions have a lifetime of the order of a hundred seconds. Positive clusters take the chemical form $\text{I}^+(\text{X})_n$, where I^+ is a typical atmospheric species, for example, H_3O^+ , NO^+ , or NO_2^+ , and X is usually an atmospheric trace gas with high proton affinity, such as ammonia, pyridine, or water vapor. The rate of attachment and the ion species created depend on both the abundance and proton affinity of trace gases present. Similarly, negative ions can be represented as $\text{I}^-(\text{Y})_n$, where the ion could be O_2^- , CO_4^- , NH_3^- , or NO_2^- and the ligand is an atmospheric electronegative species such as SO_2 , HNO_3 , or water vapor; n is typically between 4 and 10. The chemical difference between the species in the positive and negative ions leads to some physical asymmetries in the ion

properties, so the negative ions tend to have fewer molecules clustered around them, and as a result are observed to be about 20% more mobile than the positive ions.

Electric Charge

Small ions are singly charged, and respond to electric fields. The *electrical mobility* μ describes the speed that a molecular ion will attain in an electric field. It is the ratio of the magnitude of the electric field to the ion's *drift velocity*. (The electrostatic forces acting on the particle are assumed to balance the drag forces, so there is no net acceleration.) Small ions have a mobility of around $10^{-4} \text{ m}^2 \text{ V}^{-1} \text{ s}^{-1}$, and are more strongly influenced by electric fields than larger aerosol particles. The current flowing through the air is almost entirely due to these small ions, and the total *conductivity* σ can be written as

$$\sigma = e(n_+\mu_+ + n_-\mu_-) \quad [1]$$

where there are, respectively, number concentrations n_+ and n_- of positive and negative ions, with mean mobilities μ_+ and μ_- and e is the magnitude of the electronic charge.

Ion Transport

With typical vertical atmospheric electrical fields in fair weather of 100 V m^{-1} , small ions will typically migrate under electrical forces at about 1 cm s^{-1} or less. However, the electrical migration is able to occur in regions where there is little dynamical motion, in meteorological terms, and consequently clean air will always have fresh ions introduced into it either by electrical migration or *in situ* ionization. In contrast, large ions and atmospheric aerosol have negligible electrical acceleration, and their primary motion is due to advective transport arising from dynamical processes. The charge transported may therefore be ionic, or particulate (large ions), and the partitioning of the atmospheric *space charge* (the net amount of charge per unit volume) determines whether air motion or electric fields dominate.

Ions in the Global Atmospheric Electric Circuit

The terrestrial atmospheric electrical circuit exchanges charge between the planetary surface and the upper atmosphere, and is maintained by a combination of charge separation in thunderstorms and vertical ionic conduction currents in the larger fair weather regions. The conduction currents can only flow because there are charged molecular clusters (ions) present, which are sufficiently mobile to be accelerated by the small electric fields present in the atmosphere. Consequently, ions have an important part to play in maintaining the global atmospheric electric circuit. The charge-generating effect of thunderstorms and shower clouds leads to a small conduction current throughout fair weather regions of current density about 2 pA m^{-2} . The *conduction current density* J is defined by

$$J = -\sigma E \quad [2]$$

where σ is the conductivity and E the vertical electric field. If integrated across the surface of the whole planet, the ionic conduction current totals about 2000 A, [Figure 1](#).

Ion Balance Equation

The time variation of atmospheric ion concentrations can be expressed as a simplified equation, the ion balance equation, in which rates of ion formation and removal are considered separately.

Source Term

Formation of ions is, as mentioned above, essentially due to radiolysis of air molecules. There are three principal sources near the continental surface: radon isotopes, cosmic rays, and terrestrial gamma radiation, and the partitioning between the sources varies vertically. Close to the surface, ionization from turbulent transport of radon and other radioactive isotopes is important, together with gamma radiation emitted by isotopes below the surface. Cosmic rays are always present, comprising about 20% of the ionization at the surface. The cosmic ray contribution to ionization increases with increasing height in the atmosphere and dominates above the planetary boundary layer, as depicted in [Figure 1](#). The rate of production of ion pairs per unit volume is usually written as q , and this quantity shows considerable temporal and spatial variability because of geological variations, or, over the oceans, solar activity. The increasing fractional contribution of cosmic ray ionization with height has already been mentioned, but the temporal variations in q at the surface are also substantial due to the effects of turbulence and the transient production of ions. q has a frequently quoted surface value of $10 \text{ ion pairs cm}^{-3} \text{ s}^{-1}$, which is a long-term mean. On timescales of seconds, there can be significant departures from this value. In the middle and upper troposphere, the cosmic contribution to q dominates.

Bipolar Ion Equations

The ion balance equations describing production, recombination, and attachment are

$$\frac{dn_+}{dt} = q - \alpha n_+ n_- - n_+ \beta_+ Z \quad [3]$$

$$\frac{dn_-}{dt} = q - \alpha n_+ n_- - n_- \beta_- Z \quad [4]$$

in which the rate of change of ion concentration is determined by an ion source term q , and the two loss terms described above. Ion pairs are produced at a rate q per unit volume. If there are number concentrations n_+ and n_- of positive and negative ions, respectively, then the rate at which they recombine is proportional to $n_+ n_-$, with a constant of proportionality α , also known as the *ion-ion recombination coefficient* and typically assumed to be $1.6 \times 10^{-6} \text{ cm}^3 \text{ s}^{-1}$. Recombination is the principal loss mechanism of ions in clean, aerosol-free air. In more polluted air, the rate of ion loss by attachment is given by $n\beta Z$, where β_+ and β_- are the appropriate ion-aerosol attachment coefficients and Z the aerosol number concentration. In these equations, a considerable simplification has been made in assuming that it is possible to regard the aerosol as *monodisperse* (i.e., having a single particle size) so a single value of Z is appropriate. The attachment coefficient varies with

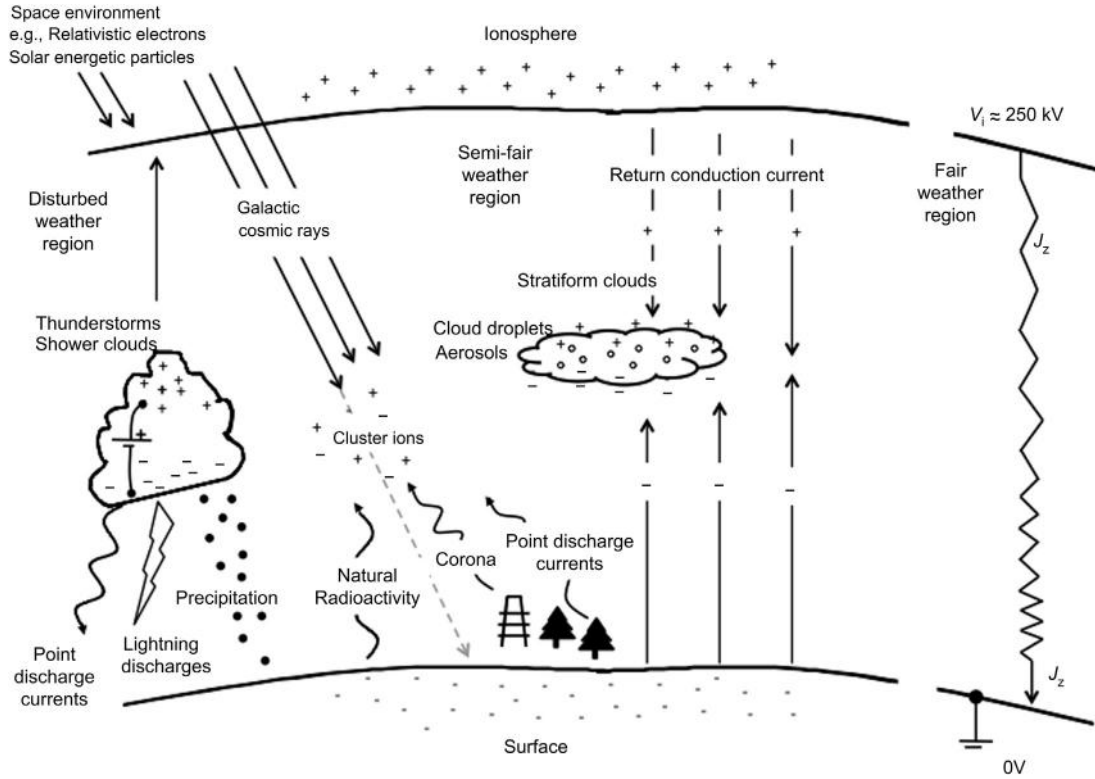


Figure 1 The role of ions in the atmospheric electrical system. Major sources of atmospheric ions are cosmic rays, and, in the boundary layer, radon isotopes diffusing from the surface, with a small contribution from point discharge currents and anthropogenic corona ions. The ions formed may become attached to atmospheric aerosol particles, or to water droplets in clouds. The remaining ions are accelerated by atmospheric electric fields, originating from charge separation occurring in thunderstorms, to form the conduction current with density J_z . Diagram kindly provided by Dr Nicoll, K.A. (Department of Meteorology, University of Reading).

aerosol radius, therefore βZ is in fact more accurately determined for atmospheric particles as an integral evaluated across the aerosol size distribution. There are many theoretical approaches to the problem of calculating β , which depend on the aerosol particle size considered, but β is typically $5 \times 10^{-5} \text{ cm}^3 \text{ s}^{-1}$ for aerosol particles with the radii most prevalent by number ($0.2 \text{ } \mu\text{m}$) in lower regions of the atmosphere.

Time-Dependent Solution

The ion balance equation can be simplified by neglecting the ion sign (i.e., $n_+ \approx n_- = n$), so that the ion-aerosol equation can be written as

$$\frac{dn}{dt} = q - \alpha n^2 - n\beta Z \quad [5]$$

Integrating this equation gives the ion concentration n as a function of time t . This time-dependent solution is

$$n(t) = \frac{[(\beta^2 Z^2 + 4\alpha q)^{1/2} - \beta Z]}{2\alpha} \left[\frac{(1 - e^{-(\beta^2 Z^2 + 4\alpha q)^{1/2} t})}{(1 + e^{-(\beta^2 Z^2 + 4\alpha q)^{1/2} t})} \right] \quad [6]$$

showing that firstly, if the ion-pair production rate q is uniform and the removal rates are also steady, the ion concentration tends to a steady value for large values of t . Secondly, the equation can be simplified according to the situations in which attachment or recombination dominates as the removal mechanisms, according to whether αn^2 or $n\beta Z$ is the bigger term. In moderately polluted air, these two terms are roughly comparable, and therefore all the terms in eqn [6] have to be evaluated.

In clean air, when ion loss can be assumed to be solely by recombination, eqn [6] reduces to

$$n(t) = \left[\frac{q}{\alpha} \right]^{1/2} \left[\frac{(1 - e^{-2(\alpha q)^{1/2} t})}{(1 + e^{-2(\alpha q)^{1/2} t})} \right] \quad [7]$$

and the concentration after a long time has elapsed is given by $n_\infty = (q/\alpha)^{1/2}$. Inserting typical atmospheric values of $q \approx 10 \text{ ion pairs cm}^{-3} \text{ s}^{-1}$ and $\alpha = 1.6 \times 10^{-6} \text{ cm}^3 \text{ s}^{-1}$ gives $n_\infty = 2500 \text{ ion pairs cm}^{-3}$. Typical values of small ion concentrations observed in mountain air are about 500 ions cm^{-3} of each sign, suggesting that attachment processes are almost always a further factor acting to modulate natural ion concentrations. There is one other factor which influences measured ion concentrations: adjacent to the surface, when the atmosphere is stable, the planet's negative charge causes an

increase of positive ions and a depletion of negative ions. This is known as the *electrode effect*.

Ions and Health

Indoor atmospheres are frequently deficient in small ions, and asymmetries in ion removal due to charged surfaces on furnishings lead to asymmetries in the remaining ion concentrations. It has been suggested that this may be the cause of some health problems, but this work appears inconclusive. Ionizers are widely used domestically to produce additional ions (generally negative ions, by corona), to replenish concentrations in indoor environments, but it is difficult for this not to affect the charging of aerosols and therefore their removal. Both positive and negative ions are readily produced in atmospheric air, therefore if there is any conceivable support for the anecdotal effect on health it is likely to be a result of the asymmetry in their concentrations. However, there is some experimental evidence that negative ions have bactericidal effects.

Ion Measurement

Atmospheric ions can be measured by exploitation of their electrical properties, for example, their migration toward an opposite charge in an electric field. The simplest way of counting ions is to apply an electric field, for example, by blowing air between two metallic plates or in a conducting cylinder with a central electrode. If one of the plates is grounded, and the electric field sufficiently strong to ensure all the unipolar ions are collected on the other electrode, then a current can be measured which is proportional to the concentration of air ions. As small ions are known to be singly charged, then the number of ions can easily be calculated.

Gerdien Condenser

If only a fixed fraction of the ions contributes to the ion current, then the electrical conductivity of the air can be measured. Gerdien, of Göttingen, first developed a method to measure air conductivity in 1905, which uses a cylindrical condenser, with a voltage applied between two cylindrical electrodes (Figure 2).

Ions are sucked into the cylinder with a fan, and those of the same sign as the polarizing voltage are repelled by the outer electrode, and move in the electric field to meet the inner electrode, where they cause a small current. In Gerdien's original configuration, the cylindrical capacitor was charged up to a fixed voltage, and the subsequent rate of decay depends on the concentration of air ions. If the time constant of the decay is τ , then the conductivity is given by

$$\sigma = \frac{\epsilon_0}{\tau} \quad [8]$$

An alternative technique, which was introduced later, was to earth the central electrode via a sensitive ammeter and measure the ion current. Typical currents in the large instruments of the early twentieth century were about 10^{-12} A, but in modern, miniaturized Gerdien condensers, sensitive electrometry can detect currents down to 10^{-15} A, giving resolution of about 20 ions cm^{-3} . Unipolar conductivity $\sigma_{\pm}$ is given by

$$\sigma_{\pm} = \frac{i\epsilon_0}{CV_{\pm}} \quad [9]$$

where i is the current, C is the capacitance, and V is the voltage across the electrodes. This relationship is valid for measuring conductivity, as long as the output current is proportional to the bias voltage, which indicates that a fixed fraction of air ions are collected by the central electrode. If the current is not linearly related to the voltage, then all the atmospheric ions are collected, and the instrument is operating as an ion counter. The method of current measurement described above is commonly deployed due to its simplicity and high time resolution, but the alternative voltage decay technique remains useful for calibration as it only requires relative voltage measurements rather than sensitive current measurements.

Mobility Spectrometry

In an aspirated cylindrical capacitor, the fraction of ions contributing to the measurement can be varied by changing the magnitude of the electrical field or the dimensions of the capacitor. This affects the maximum mobility (or minimum size) of ion that can pass through the instrument without being drawn to the central electrode and contributing to the current. A *critical mobility* μ_c , the minimum mobility of ion contributing to the conduction current, can be defined. If L is the length of the tube, a and b the radii of the outer and inner

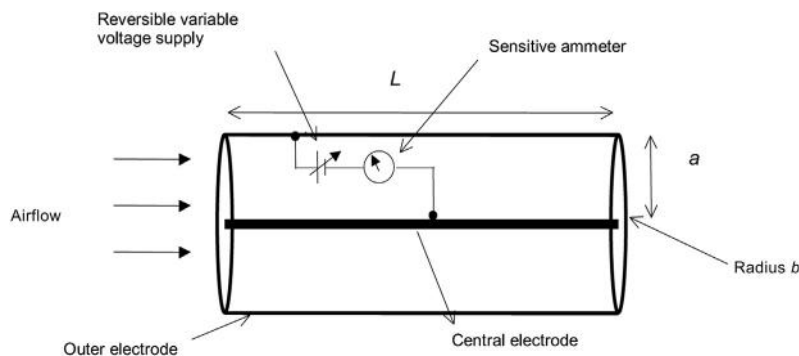


Figure 2 Schematic of a Gerdien condenser for measuring atmospheric ions.

electrodes, u the air speed, and V the voltage across the tube, then μ_c is given by

$$\mu_c = \frac{(a^2 - b^2) \ln\left(\frac{a}{b}\right) u}{2VL} \quad [10]$$

Varying the voltage across the electrodes, or the flow rate through the instrument, therefore provides a way to study the abundance of ions of different sizes. If current measurements are made as different voltages are swept across the capacitor, ion concentration n can be estimated from the conductivity, and the critical mobility using the equation below.

$$n = \frac{\sigma}{e\mu_c} \quad [11]$$

Ion *mobility spectra*, which are important for observing the behavior of ions and their evolution, can be resolved with equipment based on this relatively simple principle.

Modern Implementations of Ion Spectrometry

The approach of using a cylindrical capacitor to measure ions is by no means new. However, present day electronics, data processing and materials technology make these sensitive measurements more straightforward than in the past, and there are now many types of mobility spectrometers and analyzers available. Modern instruments employ sophisticated engineering to ensure smooth airflow, reducing random collisional currents, and complex algorithms for mobility retrieval, but they are almost all based on the simple principles described above. Some mass spectra of ambient atmospheric ions have been measured, but this method is complicated by the need for a vacuum inside the mass spectrometer. Mobility-based techniques are usually considered most appropriate for obtaining detailed measurements of atmospheric ion composition.

Ions and Climate

Direct Condensation onto Ions

It has long been assumed that atmospheric ions have little relevance to the climate system, as they have no direct role in cloud formation. It is possible to demonstrate theoretically that condensation on ions cannot occur unless the air is highly supersaturated with water vapor, at levels vastly greater than the few percent observed in atmospheric air. The cloud chamber invented by C.T.R. Wilson is able to visualize the tracks of subatomic particles by the clouds forming on the ions produced only because the supersaturation it generates is typically 400%. Such supersaturation levels are not found in the atmosphere, therefore cloud formation by the direct nucleation of water clouds on ions is not expected to occur.

Proposed Mechanisms for Electrical Influences on Climate

Recent work has identified a greater possible significance for small ions within the climate system, because they are able to assist the formation of ultrafine aerosol particles, some of which may subsequently be able to act as CCN. Numerous modeling studies have indicated the theoretical likelihood of the growth of ions from particles. Laboratory experiments have shown that particles can grow from ions in air, in the presence of a range of radioactive sources and, recently, in beams of subatomic particles that more realistically simulate cosmic ray ionization. There are now many sets of atmospheric observations of small ions growing into intermediate ions and onto small aerosol particles, via a range of charge-enhanced nucleation pathways, referred to as 'ion-induced' or 'ion-mediated' nucleation. Nucleation enhanced by atmospheric ions has been observed both in the boundary layer and in the free troposphere, but it only makes a small contribution to the total background ultrafine aerosol particle concentration. Ions appear to contribute most to aerosol formation in clean locations with low background aerosol concentrations, such as the free troposphere, but it is not known if these new aerosol particles grow to sizes that influence solar or terrestrial radiation. At the moment there are no observations of CCN growing from ions in the atmosphere, and modeling studies indicate that the radiative effect from ion nucleation is negligible. However, there is evidence for a host of additional small effects that could link atmospheric ionization to climate. These include the modification of droplet charges (and therefore potentially their activation) at the base of clouds, charge-related scavenging and coagulation, and direct absorption of infrared radiation by the bending and stretching of hydrogen bonds inside atmospheric cluster ions.

See also: Chemistry of the Atmosphere: Radioactivity: Cosmogenic Radionuclides. **Electricity in the Atmosphere:** Lightning; Sprites.

Further Reading

- Aplin, K.L., 2013. *Electrifying Atmospheres: Charging, Ionisation and Lightning in the Solar System and Beyond*. Springer.
- Chalmers, J.A., 1967. *Atmospheric Electricity*, second ed. Pergamon.
- Harrison, R.G., Carslaw, K.S., 2003. Ion-aerosol-cloud processes in the lower atmosphere. *Reviews in Geophysics* 41 (3), 1012.
- McGorman, D.R., Rust, W.D., 1998. *The Electrical Nature of Storms*. Oxford University Press.
- Pruppacher, H.R., Klett, J.D., 1997. *Microphysics of Clouds and Precipitation*, second ed. Kluwer.

Lightning

MB Baker, University of Washington, Seattle, WA, USA

© 2003 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 3, pp 1216–1223, © 2003, Elsevier Ltd.

Introduction

This article outlines the current understanding of lightning from tropospheric clouds. It begins with a brief history of research in this area and a summary of the tools used in the present-day studies. It then presents a short description of a lightning flash and discusses the major physical processes that lead to the production of lightning. This discussion provides the basis for the next topics: the links between lightning and other meteorological parameters, and the climatology of lightning.

Background

History of Lightning Research

Lightning is arguably the most dramatic naturally occurring atmospheric phenomenon, but the first understanding of its electrical nature came with Benjamin Franklin's famous experiments in the 1750s. These were followed by rapid advances in the theory of electromagnetic phenomena, with practical advances in lightning protection devices. In the 1920s C.T.R. Wilson suggested the role of thunderstorms in the atmospheric electrical system. In the 1950s and 1960s, as the field of cloud physics came into its own, the microphysical processes involved in thunderstorm electrification were widely studied, if not explained. Improvements in remote sensing of the relevant variables have further advanced our understanding of lightning as an atmospheric phenomenon, and lightning observations now provide information on cloud and atmospheric processes that complement traditional meteorological measurements.

Data and Models

Our understanding of lightning is derived from data from many sources. *In-situ* field mills, current meters, and induction rings mounted on meteorological balloons and/or aircraft yield small-scale information on in-cloud electric fields and charge distributions. Surface sensors located beneath storms and/or on mobile laboratories record electric field changes and currents associated with lightning along with colocated cloud physical variables. Sophisticated instruments measuring various components of the radiation associated with lightning can pinpoint the in-cloud locations of lightning strokes, and follow the lightning channel trajectories. Finally, lightning can be detected and its location mapped from great distances by optical detectors placed on space-based platforms and by detection at the surface of low-frequency electromagnetic signals propagating in the global waveguide.

These data can be used in numerical thunderstorm models of varying complexity to simulate the evolution of cloud and electrical properties over a cloud life cycle. These provide useful

means for testing new ideas, although at present incorporation of electrical phenomena in models requires semiempirical parameterizations.

Lightning: the Process

Lightning is an electrical discharge; i.e., a rapid redistribution of electrical charge within thunderclouds or between a cloud and its surroundings. IC, or intra-cloud lightning, is a discharge between two points in the same cloud; C–G, or cloud-to-ground lightning, transfers charge between a point in cloud and a point on the ground below. (Inter-cloud lightning, connecting points in different clouds, will not be treated here.)

A lightning flash consists of several components. In C–G flashes (which are those most widely and closely studied) a low-current (~ 1 kA), low-luminosity channel is initiated in the cloud. The channel is a fully ionized plasma tube several centimeters in diameter. Joule heating by electrical currents in the plasma raises the channel temperature to above 20 000 K. The heated gas expands, giving rise to shock waves producing the sound called thunder. The current carried by the channel discharges tens of millions of volts. The high temperatures inside the channel favor certain chemical reactions of atmospheric significance; in particular, the production of NO_x , treated elsewhere in this encyclopedia (*see Stratospheric Chemistry Topics: Reactive Nitrogen* (NO_x and NO_y)).

This first channel propagates downward in discrete steps 2–50 m in length separated by pauses of ~ 50 μs , and is commonly referred to as a stepped leader. This is followed by a return stroke (i.e., a return current) from ground to cloud. There may be several leader–return-stroke pairs but the time between strokes is usually too short for our eyes to resolve and thus what we call a lightning flash usually consists of several individual strokes. The return stroke is the brightest phase of the flash with typical peak currents ~ 40 kA; the average total current is several amps. **Figure 1** shows a sequence of strokes recorded on film.

The entire discharge lasts about 0.1 s. C–G flashes typically lower 20–30 °C to ground but values in excess of 200 °C have been recorded.

The high temperature inside the channel increases the pressure inside it and gives rise to a shock wave that expands into the surrounding air. Air resistance damps the wave and the pressure of the disturbance decreases rapidly, so that the far field effect is that of sound waves, or thunder.

Thunderstorms and Lightning

To understand the electrification of thunderstorms we review the electrical context in which they develop.

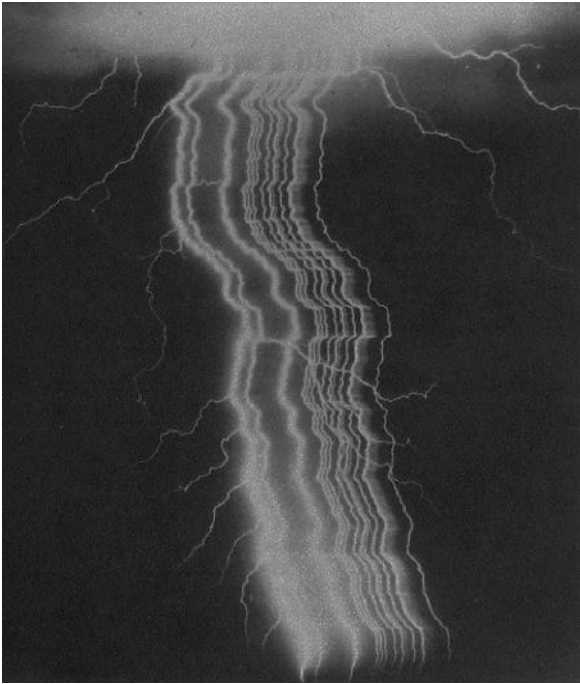


Figure 1 Lightning flash photographed with a streak camera. The camera moves during the lightning flash, enabling the resolution of a number of individual strokes. Reproduced with permission from Uman et al. (1987) *The Lightning Discharge*. International Geophysics Series, Vol. 39. San Diego, CA: Academic Press.

The Global Circuit

The surface of the Earth behaves like a negatively charged conductor. This conductor and the electrosphere (a region of net positive charge whose base is around 65 km altitude) constitute the 'plates' of a concentric spherical condenser. There is a fairly constant, spatially uniform fair-weather electric field in the atmosphere between the plates: $E_{\text{fair}} \approx 100 \text{ V m}^{-1}$ at the Earth's surface. This field is associated with a fair weather current of positive charge to the ground: $J_{\text{fair}} \approx 2 \times 10^{-12} \text{ A m}^{-2}$. Without 'batteries' the current would discharge the field in a few minutes. The major batteries, or generators, are thunderstorms, which deposit predominantly negative charge on the Earth's surface through lightning and precipitation. Thus lightning at any point perturbs the local electric charge distribution and the electromagnetic waves travelling in the global waveguide.

Thunderstorm Electrification

The electrical conductivity inside a cloud is much lower than that in the free air, since ions become attached to the hydrometeors. Thus insertion of a completely passive cloud into the fair-weather field results in the attachment of ions from the environment on hydrometeors at the cloud edges. As these move, the electric field distribution in the cloud changes. Fields up to $100\text{--}1000 \text{ V m}^{-1}$ are often found in "nonelectrified" clouds. In thunderstorms, on the other hand, internal charging mechanisms produce electric fields of up to 100 kV m^{-1} .

The information we have on in-cloud charge distributions is very limited, and we know only some general features, shown schematically in an idealized isolated convective thunderstorm in **Figure 2**. In simple storms like this one, a vertical dipolar or

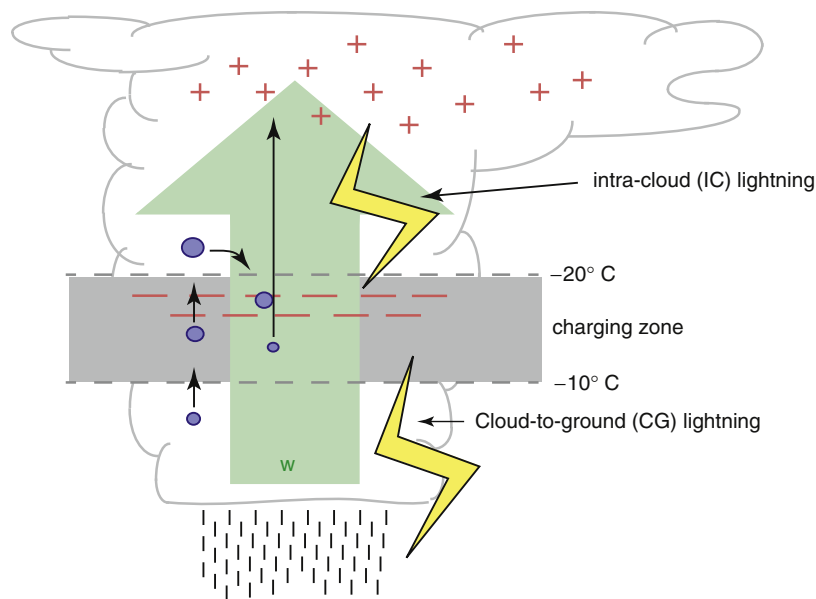


Figure 2 Schematic picture of an idealized thunderstorm, showing the ascent of hydrometeors (round circles) in the updraft (velocity w) and sedimentation of the larger particles, which collide with smaller ones below. Rebounding collisions between hail particles and small ice crystals result in charge transfer between them. Gravitational separation of the lighter (positively charged) ice particles and the negatively charged graupel (i.e., soft hail) particles leads to the electric charge distribution shown. This precipitation-based charging is thought to dominate early thunderstorm electrification. Reproduced with permission from Schroeder et al. (2000) *How Does Lightning Initiate and What Controls Lightning Frequency?* PhD thesis, University of Washington.

tripolar charge distribution develops as the cloud grows, with typical charges $Q \approx 10\text{--}100\text{C}$ in each center at maximum, carried by the cloud water and ice particles. The positive charge is distributed in a diffuse region aloft; typically, the charge density in this region is several C km^{-3} . The negative charge center can be more compact, and is always located in the part of a cloud (called the 'charging zone') in which the temperature lies between -10°C and about -20 to -25°C . This is the only region inside clouds in which vapor-grown ice, supercooled water, and hail particles coexist. There are often other important pockets of charge, including a secondary positive charge center; these become larger and more complex as the cloud ages.

The fact that the negative charge center is confined in temperature suggests that microphysical interactions control the electrical charge distributions in this type of storm, although dynamic processes (transport in updrafts and downdrafts and turbulent motions) are also important. The major microphysical charging mechanism in convective storms appears to be charge separation between colliding ice particles.

Laboratory studies show that electric charge is separated during rebounding collisions between ice particles, so that the rebounding particles carry equal and opposite charge. The sign and magnitude of the charge are reproducible functions of temperature, particle types and sizes, and the cloud water content and distribution. At temperatures above a 'reversal' threshold, around -15°C , the hail generally receives positive charge, and at lower temperatures it receives negative charge during a collision. Typically the charge separated is around 10^{-14}C per collision when one of the partners is hail; far more charge is separated when both are hail and less when both are vapor-grown ice crystals. In general, the large (precipitating) hail particles receive net negative charge as they fall through the cloud of smaller ice crystals. Subsequently, gravitational separation of the large, negatively charged precipitation particles and the lighter, positively charged ice crystals constitutes a 'generator current' $J_{\text{gen}} \approx 0.1 \text{ A km}^{-2}$.

The collision-based process is the most important, but not the only one responsible for creating the observed in-cloud charge distributions. Other processes become more important as electrification proceeds; for example, cloud particles in the existing electric field become charged inductively; melting appears to be associated with charging in stratiform clouds; ions become attached to aerosol particles and cloud particles near cloud edges; finally, motions of charged particles redistribute them throughout the cloud system.

This inhomogeneous electric charge distribution produced by these processes is associated with an electric field. The field grows as the cloud develops and eventually conditions are met in which lightning is triggered. The first flash is typically initiated just above the lower, negative charge center and the first lightning is usually IC. The lightning flash transfers charge between the upper, positive charge center and points below. The net charge density aloft is thereby decreased and positive charge builds up somewhat below the negative charge center. Continued IC lightning as the convective cell reaches full vigor continues this charge redistribution process. At some point the electric field between the lower charge centers and ground becomes strong enough to trigger C-G flashes.

This classical picture does not describe all thunderstorms. Recent measurements (for example, those made during the

Severe Thunderstorm Electrification and Precipitation Study (STEPS) project in 2000) reveal that the charge structure in clouds is often inverted, with negative charge aloft. The exact mechanisms for this inversion are not known. Moreover, in complex cloud systems containing large stratiform regions and/or multicellular structures the charge distribution can be highly three-dimensional and complex.

Most C-G flashes from convective storms carry negative charge to ground, indicating that the in-cloud point of origin of the flash is associated with a negative charge pocket, but in midlatitude storms around 10% deposit positive charge to ground. Positive flashes tend to be associated with low precipitation regions in mesoscale storm systems, and they are often very intense. In general, reversed lightning polarity results from (1) wind shear, displacing the upper positive charge so that it overlies cloudfree air and is electrically coupled to ground; (2) inverted charge structure in cloud, possibly due to charging at temperatures higher than the reversal temperature; and (3) absence of large concentrated pockets of negative charge.

Microphysical Processes Leading to Lightning Production

In order to make further progress in understanding lightning generation we must solve two outstanding puzzles; namely, (1) what is the mechanism for the charge transfer between colliding ice particles, and (2) how is lightning initiated? In this section we briefly discuss our current understanding of these processes and the hypotheses now being tested to explain them.

Charge Transfer via Ice-Ice Collisions

The mechanism by which the rebounding ice particles become charged remains somewhat unclear. Both sign and magnitude of the charge depend on the rate of growth of the ice particle from the vapor. The charge transfer appears to be due to the fact that positive water ions in ice have much higher mobilities than do negative ions. Ions are created at high rates near the ice surface, and the positive ions migrate inside, so that the growth creates net negative charge at the ice-vapor interface. During a collision, material from the two particles is mixed and net negative charge is transferred to the particle with the slower growth rate. The various characteristics of the hail – i.e., its smoothness or roughness, the density of grain boundaries and dislocations, and the chemical composition of the frozen water – all modify the charge transfer.

Lightning Initiation

Lightning initiation is thought to involve the following stages. (1) The in-cloud electric fields intensify via microphysical and dynamic processes. (2) Free electrons in some region of the field begin to accelerate, creating more electrons via ionization, in so-called electron avalanches. If these continue they result in propagating corona streamers, or small-scale currents. (3) Heating associated with the propagation of high electric current through and beyond the region of very high local fields produces the hot, completely ionized lightning channel, or leader.

This process is known as ‘dielectric breakdown’. In the laboratory, dielectric breakdown can occur only if the ambient electric field reaches a (pressure-dependent) threshold value, denoted $E_{th}(p)$. At surface pressure (1000 hPa) $E_{th} \approx 2600 \text{ kV m}^{-1}$. This value, and the physics of breakdown in the laboratory context, are well understood. However, measured electric fields inside thunderclouds never reach the threshold value and are usually more than an order of magnitude less than this value.

Two classes of hypotheses are typically invoked to resolve this puzzle. While both involve acceleration of an initial electron and subsequent production of electron avalanches, everything else about the two mechanisms is different. The first is the ‘conventional breakdown’ hypothesis, which operates on very small spatial scales. It is based on the fact that in the vicinity of conducting hydrometeors the electric field magnitude can reach the breakdown threshold value. According to this lightning initiation hypothesis, free electrons near cloud ice and water particles are accelerated by the local field and can produce local streamers.

The second class of triggering mechanism hypotheses depends on the fact that in-cloud electric fields are fairly large, even if not of breakdown magnitude, over several kilometers. An electron moving in such a field is accelerated by the field and decelerated by its interactions with neutral molecules. If the electric field magnitude is sufficiently great, it can compensate for the deceleration. The balancing electric field is called the ‘breakeven’ field; $E_{be}(p) [\text{kV m}^{-1}] \approx 200p [\text{atm}]$, much smaller than the breakdown threshold field.

The ‘runaway breakdown’ hypothesis of lightning initiation suggests that if E_{be} is exceeded, even if slightly, over sufficient distance, then high-energy electrons (resulting from cosmic ray showers or atmospheric radioactivity, or short-lived, small-scale intense electric fields) travelling in the field can create sufficient daughter electrons by ionization to initiate a leader, or ionized channel. The more the field exceeds the breakeven value the shorter the distance needed to create the leader. Measured electric fields in clouds are generally much smaller than the breakeven field, and just reach it or barely exceed it immediately prior to a lightning stroke, lending credence to the runaway breakdown hypothesis.

Neither the conventional nor the runaway mechanisms are completely satisfactory; neither can explain all the observations. Unfortunately, balloon measurements cannot solve the lightning initiation puzzle because they are Lagrangian in nature, they sample very small volumes of cloud, and their spatial and temporal resolution is too coarse. Therefore considerable effort will have to be devoted to interpretation of complementary measurements in order to clarify the nature of the lightning trigger(s).

Lightning and Cloud Properties

Although much remains to be learned about lightning generation, we can examine some basic relationships linking lightning frequency $F [\text{s}^{-1}]$ to other important parameters. A simple dimensional argument illustrates the important links. F is a function of the charge generation rate, $\dot{\rho}_Q$, the depth H of the electrically active region and the threshold field for producing

lightning (which we assume here is a function of E_{be}). The charge-generating current density $J_{gen} \propto \dot{\rho}_Q H$.

Therefore

$$F \propto \frac{\dot{\rho}_Q H}{E_{be}} \quad [1]$$

Flash Rate and Radar Reflectivity in Charging Zone

For simplicity we neglect all contributions to $\dot{\rho}_Q$ except that due to collisions between hail particles and vapor grown crystals. Consider a situation in which $N_{hail} [\text{m}^{-3}]$ hail particles of diameter D_{hail} collide with N_{ice} vapor-grown ice particles of diameter d_{ice} . The laboratory charging results can be parameterized in terms of these variables, leading to an expression of the form

$$\dot{\rho}_Q \propto N_{hail} D_{hail}^6 N_{ice} d_{ice}^3 \quad [2]$$

$$\approx Z M_{ice} \quad [3]$$

where Z is the radar backscatter and M_{ice} the mass of vapor-grown ice in the region of the collisions. Thus, we can write

$$F \propto \frac{Z M_{ice} H}{E_{be}} \quad [4]$$

Therefore we expect the measured lightning frequency to increase with radar reflectivity (i.e., the large hail concentration) and with the mass of vapor grown ice. Both of these predictions have been borne out in preliminary analyses of lightning frequency vs. radar reflectivity and 85 GHz signal from thunderstorms. As a rule of thumb, a minimum radar reflectivity of around $Z = 40 \text{ dBZ}$ for temperatures about -7°C seems to be required for rapid electrification.

Flash Rate and Updraft Velocity

We can rewrite eqn [4] to relate F to updraft velocity in the charging zone. The fall velocity of the hail particles is very roughly proportional to the hail particle diameter. Therefore, if we assume that most of the collisions take place at the balance point, where the fall velocity of the hail particles is equal to the updraft velocity w , then

$$\dot{\rho}_Q \propto w^6 M_{ice} \quad [5]$$

Thus

$$F \propto \frac{w^6 M_{ice} H}{E_{th}} \quad [6]$$

Since updraft velocity in the mixed-phase zone is of importance in weather prediction, the strong dependence of lightning on this factor may provide a useful remote indicator of storm strength. More careful studies show that as a general rule the very existence of lightning implies vertical velocities of at least $7\text{--}8 \text{ m s}^{-1}$ in the charging zone. Updraft velocities are often tied to buoyancy, or CAPE (convective available potential energy), so that it is not surprising that observations tie F to CAPE in regions where sounding shape is not highly variable, such as over oceans; however, this relationship is regional over land and attempts to identify a single CAPE ‘threshold’ with the onset of lightning have not been very fruitful.

Flash Rate and Vertical Water Fluxes

It has long been hoped that, because of the role of hail in generating lightning, reliable and useful relationships could be found by which lightning frequency could be used as a surrogate for precipitation at the ground. Unfortunately, although it is possible to derive dimensional relationships between flash rate and in-cloud hail concentrations (see eqn [3]), universal quantitative relationships of sufficient accuracy for use in prediction are far from our reach. Many high precipitation storms produce no lightning at all (particularly if most of the precipitation comes from warm cloud processes), and the complex processes involved in producing precipitation (particularly precipitation at the ground, where predictions are most needed) are so variable on small spatial and temporal scales that precipitation cannot be accurately predicted from simple macroscopic measures like lightning frequency. Estimates of precipitation 'yield' γ for example, range from $10^7 \leq \gamma \leq 10^{10} [\text{kgH}_2\text{O flash}^{-1}]$ in one given location.

Since lightning frequency increases with increasing rate of supply of water to the charging zone and upward flux of small ice particles, lightning can be used in certain predictive schemes. Remote sensing of lightning has been successfully incorporated into predictions of hurricane landfall time and location, and more such applications are likely to follow. On larger scales, flash rate can also be considered a surrogate for the rate of water lofting to the upper troposphere in convective towers. A simple argument suggests that globally around $M_{\text{ice}} \approx 10^6 [\text{kg ice flash}^{-1}]$ is lofted to the upper troposphere. At present, attempts to make more quantitative estimates of regional condensate lofting from lightning observations have proved frustrating because the relationship is highly variable, but global lightning activity as inferred from ELF (extremely low-frequency) waves in the global circuit appears to be strongly correlated with tropical water lofting.

Lightning and Atmospheric Aerosols

Forest fires and urban pollution have been shown to produce anomalies in C–G lightning flash rate, lightning intensity (peak current, and/or radiance) and in the sign of the charge brought to ground by lightning. These intriguing observations might be

due to modifications of the hydrometeor populations by variations in the distributions of aerosol particles, to modification of the soundings by anomalous surface heating, to chemical modifications of the charge transfer process or to some combination of these.

Lightning Climatology

The advent of routine lightning monitoring programs from satellites and important advances in ground sensing capability have allowed great advances in our understanding of the distribution and frequency of lightning. Here we present an overview of lightning climatology.

The global, diurnal average C–G flashrate is $\bar{F}_{\text{C-G}} \approx 12\text{--}16 \text{ s}^{-1}$; $F_{\text{max}} \approx 55 \text{ s}^{-1}$ in NH summer, over land. As shown in Figure 3, about 70% of all lightning occurs between 30°S and 30°N , echoing the distribution of convective activity. In the tropics there is no seasonal cycle in the lightning frequency. The seasonal cycle in midlatitude lightning flash rate (with more lightning in the summer months) is due mostly to differences in the number of storms (i.e., presumably, in CAPE), rather than in differences in the flash rate per storm.

F is maximum at 1500–1800 local time, with much smaller amplitude variation over the oceans than over land. The diurnal cycle at any fixed location is due in part to a cycle in flash rate per storm and in part to the diurnal cycle in the number of storms.

Lightning production requires both high vertical velocity and high upward water flux in the charging or mixed phase zone. Water fluxes in this region are relatively weak in many oceanic storms, which may provide a clue to the large land–ocean differences in F indicated in Figure 3: $F_{\text{storm}}(\text{on land}) \approx 2 \times F_{\text{storm}}(\text{over ocean})$, but $F_{\text{total}}(\text{land}) \approx 10 \times F_{\text{total}}(\text{ocean})$. (Probably the difference is enhanced by large continental storms.)

The electric field distribution in and below a thunderstorm determines whether the lightning produced is IC or C–G. Globally, the ratio IC/C–G ranges from about 1 to 5, increasing with increasing latitude. This latitudinal variation probably occurs because the base of the charging zone is higher and the depth of the charging zone larger in the tropics than in midlatitudes.

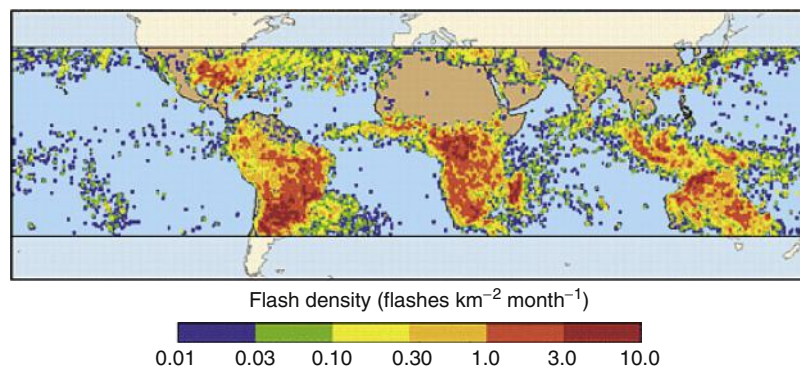


Figure 3 Lightning incidence in the months of December 1997, January 1998 and February 1998, as measured by Lightning Imaging Sensor, in tropical orbit aboard the Tropical Rainfall Measuring Mission (TRMM) satellite. Note preponderance of lightning over land and surrounding land masses. Reproduced with permission from Christian HJ, Blakeslee RJ, Goodman SG, *et al.* (1999) The lightning imaging sensor. In: Proceedings of the 11th International Conference on Atmospheric Electricity, pp. 746–749. Guntersville, Alabama, 7–11 June 1999.

Summary

Lightning is an electrical discharge that arises in the high electric fields inside a thunderstorm. While neither the field generation mechanisms nor the details of the lightning initiation are well understood, we have enough information to begin to relate lightning to the atmospheric conditions in which it originates. Lightning frequency, polarity, intensity, and the spatial distribution of lightning flashes in principle carry information about spatial distributions of updraft velocities and hydrometeors, particularly in the mixed-phase regions of storms. Lightning can be sensed at great distance and it yields a measure of the state of a storm that is independent of those measures derived from meteorological instruments. Thus, as our understanding of the links of lightning to other atmospheric phenomena improves, we can design ways to utilize lightning data to improve storm prediction. Moreover, distributions of lightning can provide long-term information on the atmospheric portion of the hydrological cycle. Incorporation of lightning-related variables into the suite of measures used routinely to diagnose surface temperature and atmospheric stability may eventually aid in monitoring global and regional climate change.

Acknowledgements

This article was prepared with the help of V. Schroeder and R. Solomon, and was supported by NASA NAG1819.

See also: **Aviation Meteorology:** Aviation Weather Hazards.

Electricity in the Atmosphere: Global Electrical Circuit; Ions in the Atmosphere; Sprites. **Radar:** Cloud Radar. **Stratospheric**

Chemistry Topics: Reactive Nitrogen (NO_x and NO_y).

Further Reading

- Christian H.J., Blakeslee R.J., Goodman S.J., et al. (1999) The lightning imaging sensor. In: Proceedings of the 11th International Conference on Atmospheric Electricity, pp. 746–749. Gunterville, Alabama, 7–11 June 1999.
- Houze, R.A., 1993. *Cloud Dynamics*. Academic Press, San Diego, CA.
- Latham, J., 1981. The electrification of thunderstorms. *Quarterly Journal of the Royal Meteorological Society* 107, 277–298.
- MacGorman, D.R., Rust, W.D., 1998. *The Electrical Nature of Storms*. Oxford University Press, Oxford.
- Schroeder, V., 2000. *How Does Lightning Initiate and What Controls Lightning Frequency?*. PhD thesis University of Washington.
- Uman, M., 1987. The Lightning Discharge. vol. 39. In the *International Geophysics Series*. Academic Press, San Diego, CA.

Sprites

WA Lyons, FMA Research Inc., Fort Collins, CO, USA

© 2003 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 2, pp 738–747, © 2003, Elsevier Ltd.

Introduction

Science often advances at a deliberate and cautious pace. Over a hundred years passed before the persistent reports of luminous events in the stratosphere and mesosphere associated with tropospheric lightning were accepted by the scientific community. Since 1886, dozens of eyewitness accounts, mostly published in obscure meteorological publications, have been found alongside articles describing meteorological esoterica such as half-meter-wide snowflakes and toads falling during rain showers. The phenomena were variously described using terms such as ‘cloud-to-space lightning’ and ‘rocket lightning’. A typical description might read ‘In its most typical form it consists of flames appearing to shoot up from the top of the cloud or, if the cloud is out of sight, the flames seem to rise from the horizon’. Such reports were largely ignored by the nascent atmospheric electricity community, even when they were posted by a Nobel Prize-winning physicist. As early as 1925, C.T.R. Wilson proposed possible mechanisms to explain such phenomena. In 1956 Wilson commented

It is quite possible that a discharge between the top of the cloud and the ionosphere is a normal accompaniment of a lightning discharge to earth ... a diffuse discharge between the top of the cloud and the upper atmosphere ... many years ago I observed what appeared to be discharges of this kind from a thundercloud below the horizon. They were diffuse, fan-shaped flashes, green in color ... extending up into a clear sky.

During the last three decades, several compendia of similar subjective reports from credible witnesses worldwide were prepared by Otha H. Vaughan (NASA Marshall) and the late Bernard Vonnegut (The State University of New York – Albany). The events were widely dispersed geographically from equatorial regions to above 50° latitude. About 75% of the observations were made over land. The eyewitness descriptions shared one common characteristic: they were perceived as highly atypical of ‘normal’ lightning. The reaction of the atmospheric science community could be summarized as indifference at best. Then, as so often happens in science, serendipity intervened.

Hard Evidence Emerges

The air of mystery began to dissipate at 0414 UTC on 6 July 1989. Scientists from the University of Minnesota, led by Professor John R. Winckler, were testing a low-light CCD video camera system (LLTV) for a forthcoming rocket flight at an observatory in central Minnesota. The resulting tape contained, quite by accident, two fields of video that provided the first hard evidence for what are now called sprites. The twin pillars of light were assumed to originate with a thunderstorm system some 250 km

to north along the Canadian border. The storm system, while not especially severe, did contain a larger than average number of positive-polarity cloud-to-ground (+CG) lightning flashes. From this singular observation emanated a decade of fruitful research into the electrodynamics of the middle atmosphere.

In the early 1990s, NASA scientists searched tapes from the Space Shuttle’s LLTV camera archives and confirmed at least 17 apparent sprites above storm clouds occurring worldwide. The orbital perspective suggested a relationship between sprites and tropospheric lightning. The specific lightning flashes associated with the observed sprites were often among the brightest in the region. The sprites occurred within milliseconds of the brightest cloud illumination, and were apparently triggered by especially energetic discharges within the storm cell. The parent convective systems, while often larger than their neighbors, had otherwise unexceptional flash rates. [Figure 1](#) portrays the global distribution of sprite reports. By late 1999 and early 2000 the first sprites (and elves) had been detected above European convective storms.

By 1993, NASA’s Shuttle Safety Office had developed concerns that this newly discovered ‘cloud-to-space lightning’ might be fairly common and thus pose a potential threat to Space Shuttle missions, especially during launch or recovery. Based upon the available evidence, the author’s hunt for these elusive events was directed above the stratiform regions of large mesoscale convective systems (MCSs), known to generate relatively few but often very energetic lightning discharges. On the night of 7 July 1993, an LLTV was deployed for the first time at the Yucca Ridge Field Station (YRFS), on high terrain about 20 km east of Fort Collins, Colorado, USA. Exploiting an uninterrupted view of the skies above the High Plains to the east, the LLTV was trained above a large nocturnal MCS in Kansas, some 400 km distant. Once again, good fortune intervened as 248 sprites were imaged over the next four hours. Analyses revealed that almost all the sprites were associated with +CG flashes, and assumed an amazing variety of shapes ([Figure 2](#)). Within 24 h, in a totally independent research effort, sprites were imaged by a University of Alaska team on board the NASA DC8 aircraft over Iowa. The following summer, the University of Alaska’s flights provided the first color videos detailing the red sprite body with bluish, downward-extending tendrils. The same series of flights documented the truly bizarre blue jets.

By 1994 it had become apparent that there was a rapidly developing problem with the nomenclature being used to describe the various findings in the scientific literature. The name ‘sprite’ was elected so as to avoid employing a term that might presume more about the physics of the phenomena than our knowledge warranted. Sprite replaced terms as ‘cloud-to-space’ lightning and ‘cloud-to-ionosphere discharge’ and similar appellations that were initially in use. The intentionally fanciful names given to phenomena were selected in the same spirit. Today a host of phenomena have been named: sprites, blue jets, blue starters, elves, sprite halos, and trolls, with

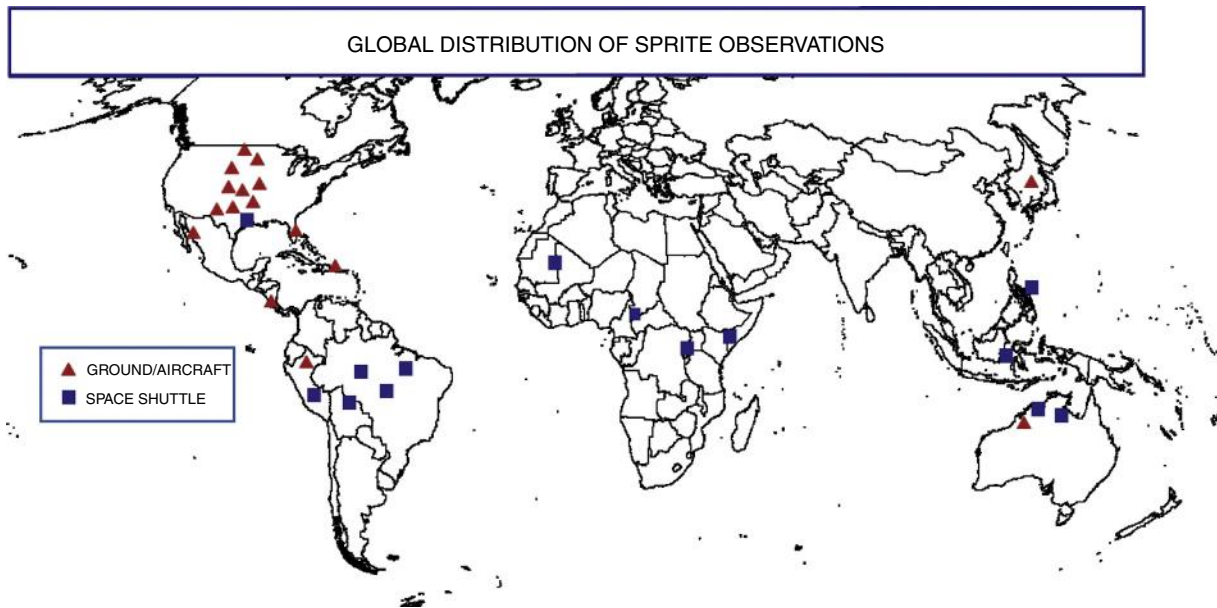


Figure 1 Global distribution of reports of sprites and elves (as of 2000). The US events are shown schematically as hundreds of storms have been monitored in that region.

perhaps others remaining to be discovered. Collectively they have been termed Transient Luminous Events (TLEs).

A Description of TLEs

Since the first sprite observations in 1989, the scientific community's misperception of the middle atmosphere above thunderstorms as 'uninteresting' has completely changed. Much has been learned about the morphology of TLEs in recent years. Sprites can extend vertically between 95 km and less than 30 km. While telescopic investigations reveal that individual tendril elements may be of the order of 10 m across, the total illuminated volume can exceed 10^4 km^3 . Sprites are almost always preceded by +CG flashes, with time lags of less than

1 ms to over 100 ms. To date, there are only two documented cases of sprites associated with negative-polarity CGs. The sprite parent +CG peak currents range widely, from under 10 kA to over 100 kA, though on average the sprite +CG peak current is 50% higher than other +CGs in the same storm. High-speed video images (1000 fps) suggest that many sprites usually initiate around 70–75 km from a small point, and first extend downward and then upward development at speeds of around 10^7 m s^{-1} . Sprite luminosity on typical LLTV videos can endure for tens of milliseconds. Photometry suggests, however, that the brightest elements usually persist for a few milliseconds, though occasionally small, bright 'hot spots' linger for tens of milliseconds. By 1995, sprite spectral measurements by Russell Armstrong of Mission Research Corporation confirmed the presence of the N_2 first positive emission lines. In 1996, photometry provided clear evidence of ionization in some sprites associated with blue emissions within the tendrils and sometimes the sprite body. Peak brightness within sprites is on the order of 1000 kR. In seven years of observations at Yucca Ridge, sprites were typically associated with larger storms ($>10^4 \text{ km}^2$ radar echo), especially those exhibiting substantial regions of stratiform precipitation. The TLE-generating phase of High Plains storms averages about 3 h. The probability of optical detection of TLEs from the ground in Colorado is highest between 0400 and 0700 UTC. It is suspected that sprite activity is maximum around local midnight for many storms around the world. The TLE counts observed from single storm systems has ranged from 1 to 776, with 48 as an average count. Sustained rates as high as once every 12 s have been noted, but more typical intervals are on the order of 2–5 min.

Like snowflakes, no two sprites are alike. Developing a taxonomy to describe sprite shapes has proven to be a challenge, though several names have come into common use. The vertically oriented, narrow (order 1 km) columnar sprites are often referred to as c-sprites (Figure 3). Sprites that have



Figure 2 A low-light television image of a sprite at a range of about 600 km from the camera. The sprite extends from about 90 km downward to about 40 km (The bright spot at the lower left is Jupiter).



Figure 3 An example of thin columnar sprites (c-sprites) with downward-extending tendrils. (The bright spot is Jupiter.)

pronounced downward-extending tendrils and upward-flaring streamers have been referred to as ‘carrots’ (Figure 4). Some amorphous and apparently structureless events have been termed ‘smudges’. Rapid-fire successions of sprites appearing to propagate above a cloud top have been termed ‘dancers’ (Figure 5).

Can sprites be detected with the naked eye? The answer is a qualified yes. Most sprites do not exceed the threshold of detection of the dark-adapted human eye, but some do. Naked-eye observations require a dark (usually rural) location, no moon, very clean air (such as visibilities typical of the western United States), and a dark-adapted eye (5 min or more). A storm located 100–300 km distant is ideal if it contains a large stratiform precipitation area with +CGs. The observer should stare at the region located some 3–5 times the height of the storm cloud (Figure 6). It is best to shield the eye from the lightning flashing within the parent storm. Often sprites are best seen out of the corner of the eye. The event is so transient often observers cannot be sure of what they may have seen. The perceived color may not always appear ‘salmon red’ to



Figure 4 A close-range (around 175 km) view of a ‘carrot’ type sprite with both downward-extending thin tendrils and bright upward branches.

any given individual. Given the human eye’s limitations in discerning color at very low light levels, some report seeing sprites in their natural color but others see them as white or even green (as did C.T.R. Wilson). Figure 7 is an LLTV sprite imaged rendered to what are believed to be close to the sprite’s true colors.

While most TLE discoveries came as a surprise, one was predicted in advance from theoretical arguments. In the early 1990s, Stanford University researchers proposed that the electromagnetic pulse (EMP) from CG flashes could induce a transient glow at the lower ledge of the ionosphere between 80 and 100 km altitude (Figure 8). Evidence for this was first noted in 1994 using LLTV at Yucca Ridge. Elves were confirmed the following year by photometric arrays deployed at Yucca Ridge by Tohoku University. Elves, as they are now called, are believed to be expanding quasi-toroidal structures that attain a width of up to several hundred kilometers. (The singular is elve, rather than elf, in order to avoid confusion with ELF (extremely low frequency) radio waves which are used intensively in TLE studies.) Photometric measurements suggest that an elve’s intrinsic color is red owing to strong N_2 first positive emissions. While relatively bright (1000 kR), their duration is $<500 \mu s$. Elves follow very high peak current (often >100 kA) CGs, most of which are positive in polarity, by about $\sim 300 \mu s$. Stanford University researchers, using sensitive photometric arrays, documented the outward and downward expansion of the elve’s disk. They also suggest many more dim elves occur than are detected with conventional LLTVs. These fainter elves have been suggested to be more evenly distributed between positive- and negative-polarity CGs.

It has been determined recently that some sprites are preceded by a diffuse disk-shaped glow that lasts about a millisecond and superficially resembles an elve. However, these structures, now called ‘sprite halos’, are less than 100 km wide, and propagate downward from about 85 to 70 km altitude. Sprite elements sometimes emerge from the lower portion of the sprite halo’s concave disk.

Blue jets are rarely observed from ground-based observatories, in part because of atmospheric scattering of the shorter wavelengths. LLTV video from aircraft missions revealed blue jets emerging from the tops of electrically active thunderstorms. The jets propagate upward at speeds of $\sim 100 \text{ km s}^{-1}$, reaching terminal altitudes around 40 km. Their estimated brightness is on the order of 1000 kR. Blue jets do not appear associated with specific CG flashes. Curiously, however, CG lightning activity appears to cease for several seconds within a 15 km radius after each occurrence of a blue jet. Some blue jets appear not to extend very far above the clouds, only propagating as bright channels for a few kilometers above the storm tops. These nascent blue jets have been termed blue starters. During the 2000 observational campaign at Yucca Ridge, the first blue starters ever imaged from the ground were noted. They were accompanied by very bright, short-lived (~ 20 ms) ‘dots’ of light at the top of MCS anvil clouds. There have been anecdotal associations of blue jets with hail-producing storms. A NASA ER2 pilot flying over the Dominican Republic above Hurricane Georges in 1998 described seeing luminous structures that resembled blue jets.

The troll is the most recent addition to the TLE family. In LLTV videos, trolls superficially resemble blue jets, yet they are

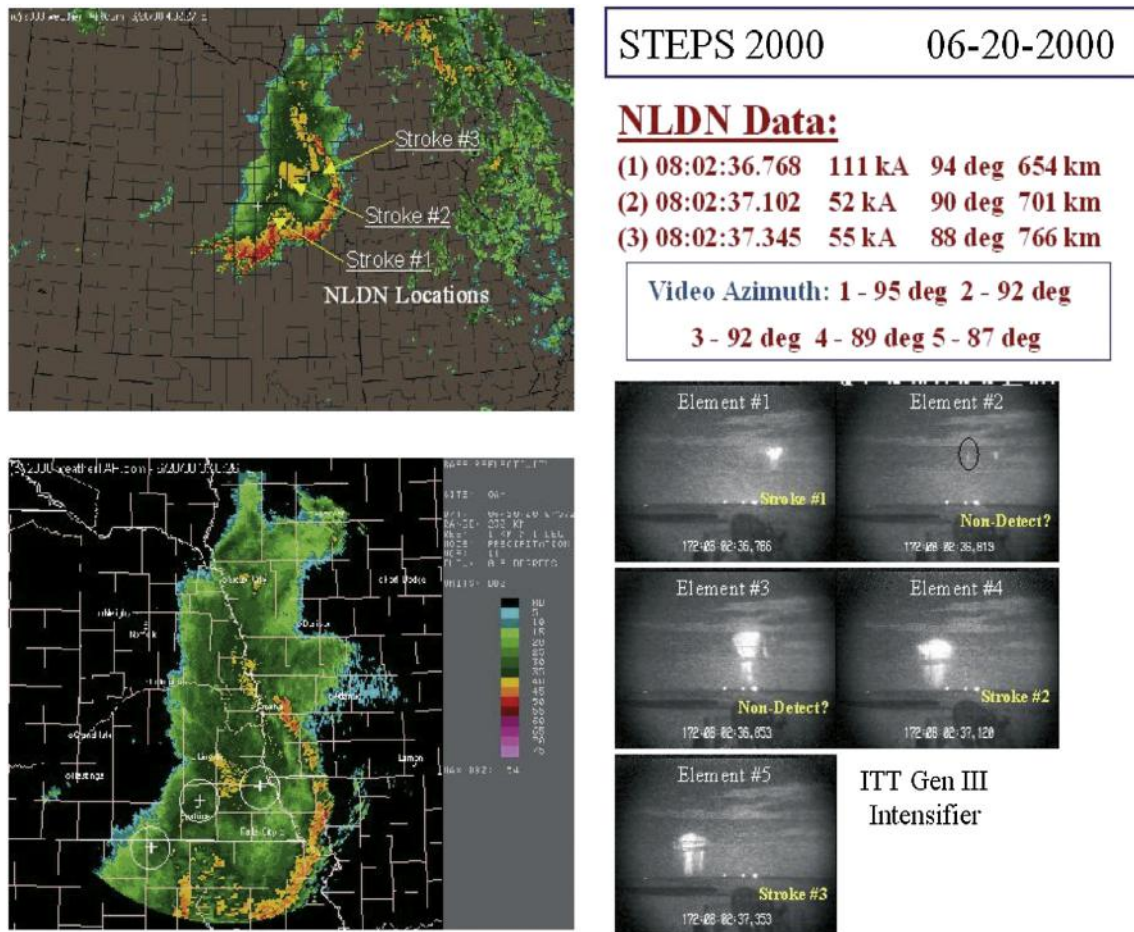


Figure 5 Example of a 'dancing' sprite in which five separate sprites appear in succession moving north-eastward above the trailing stratiform region of a mesoscale convective system in eastern Nebraska. Three +CGs were detected by the NLDN that were associated with given sprites. It is believed the three +CGs are part of a continuous discharge characterized by a >100 km long horizontally propagating 'spider' lightning discharge. The sprite images were obtained with an off-the-shelf ITT Night Vision systems mated with a DV camcorder.

clearly dominated by red emissions. Moreover, they occur after an especially bright sprite in which tendrils have extended downward to near cloud tops. The trolls exhibit a luminous head leading a faint trail moving upward initially around 150 km s^{-1} , then gradually decelerating and disappearing by 50 km. It is still not known whether the preceding sprite tendrils actually extend to the physical cloud tops or whether the trolls emerge from the storm cloud per se.

The Lightning Source Term for TLEs

Worldwide, a variety of storm types have been associated with TLEs. These include the larger mid-latitude MCS, tornadic squall lines, tropical deep convection, tropical cyclones, and winter snow squalls over the Sea of Japan. It appears, however, that the central United States may be home to some of the most prolific TLE producers, even though only a minority of High Plains thunderstorms produce TLEs. Some convective regimes, such as supercells, have yet to be observed producing many TLEs, and the few that have been observed are mostly confined to the stratiform precipitation region that may develop during

the late mature and decaying stages. Furthermore, the vast majority of +CGs, even many with peak currents above 50 kA, produce neither sprites nor elves that are detectable using standard LLTV systems. While large peak current +CGs populate both MCSs and supercells, only certain +CGs possess characteristics that generate sprites or elves.

Monitoring of ELF radio emissions in the Schumann resonance bands (8–120 Hz) has provided a clue to what differentiates the TLE parent CG from 'normal' flashes. The background Schumann resonance signal is produced from the multitude of lightning flashes occurring worldwide. It is generally a slowly varying signal, but periodically brief amplitude spikes, called Q-bursts, are noted. Their origin was a matter of conjecture for several decades. In 1994, visual sprite observations at Yucca Ridge were coordinated in real time with ELF transients (Q-bursts) detected at a Rhode Island receiver station operated by Earle Williams of the Massachusetts Institute of Technology. This experiment, repeated many times since, clearly demonstrated that Q-bursts are companions to the +CG flashes generating both sprites and elves. ELF measurements have shown that sprite parent +CGs are associated with exceptionally large charge moments (300 to >2000 C km).

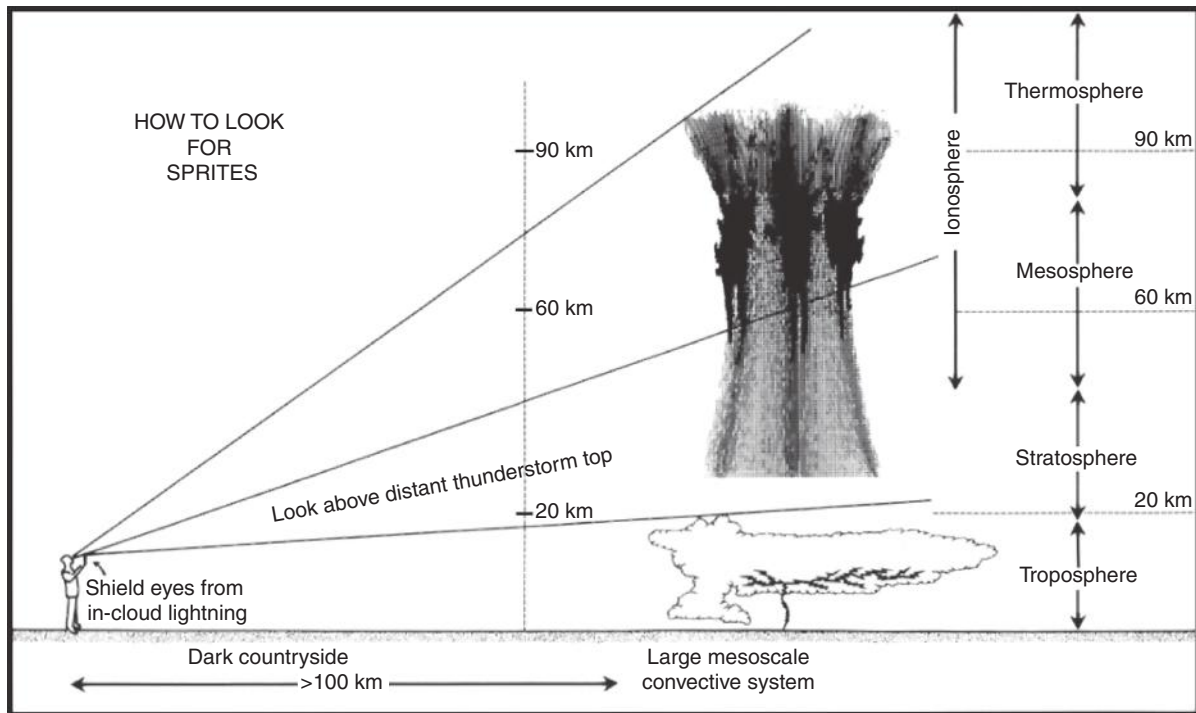


Figure 6 The geometry of sprite watching.

The sprite +CG ELF waveform spectral color is 'red', that is, peaked toward the fundamental Schumann resonance mode at 8 Hz. Lightning charge transfers of hundreds of coulombs may be required for consistency with theories for sprite optical intensity and to account for the ELF Q-burst intensity. Lightning that causes elves has a much flatter ('white') ELF spectrum and, though associated with the very highest peak current +CGs, exhibits much smaller charge moments (<300 C km).

Recent studies of High Plains MCSs confirm that their electrical and lightning characteristics are radically different from the textbook 'dipole' thunderstorm model, derived largely from studies of rather small convective storms. Several horizontal laminae of positive charge are found, one often near the

0°C layer, and these structures persist for several hours over spatial scales of ~ 100 km. With positive charge densities of $1\text{--}3$ nC m $^{-3}$, even relatively shallow layers (of order 500 m) covering 10^4 to 10^5 km 2 can contain thousands of coulombs. Some 75 years ago, C.T.R. Wilson postulated that large charge transfers and particularly large charge moments from CG lightning appear to be a necessary condition for conventional breakdown that could produce middle-atmospheric optical emissions. Sprites occur most readily above MCS stratiform precipitation regions with radar echoes larger than $\sim 10^4$ km 2 . It is not uncommon to observe rapid-fire sequences of sprites



Figure 7 A low-light television image of a sprite at a range of about 400 km from the camera. The glow on the horizon is from the parent lightning stroke. The sprite has been colorized to show what is believed to be close to the true color.

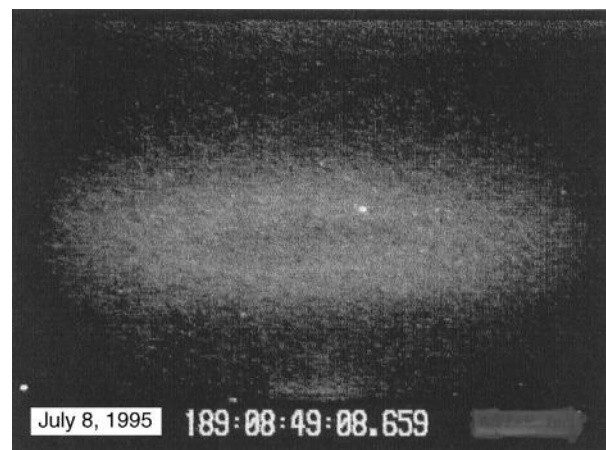


Figure 8 Low-light television image of a large elve above a distant thunderstorm. Note the lower intensity near the center, indicative of the expanding toroidal-shaped structure of the elve.

Table 1 Current ideas on TLE storm/lightning parameters in selected storm types

	<i>Core of supercells</i>	<i>MCS stratiform region</i>	<i>'Ordinary' convection</i>
+CG peak currents	> ~40 kA	> ~60 kA	~30 kA
Storm dimension	10–20 km	10–500 km	<100 km
Spider discharges	Few/small	Many/large	Some
Continuing current	Short if any	Longest	Some
+CG Channel height	10–15 km?	5 km?	10 km?
Sprites occur?	No (except at end) ^a	Many	Rare?
Elves occur?	No (except at end) ^a	Many	Rare?
Blue jets occur?	Yes?	Rare?	Rare?

^aSome supercells may generate a few sprites during their final phase when/if extensive stratiform develops.

propagating above storm tops, apparently in synchrony with a large underlying horizontal lightning discharge. One such 'dancer' included a succession of eight individual sprites within 700 ms along a 200 km long corridor. This suggests a propagation speed of the underlying 'forcing function' of $\sim 3 \times 10^5$ m s⁻¹. This is consistent with the propagation speed of 'spider' lightning – vast horizontal dendritic channels tapping extensive charge pools once a +CG channel with a long continuing current becomes established.

It is suspected that only the larger MCS, which contain large stratiform precipitation regions, give rise to the +CGs associated with the spider lightning networks able to lower the necessary charge to ground. The majority of sprite parent +CGs are concentrated in the trailing MCS stratiform regions. The radar reflectivities associated with the parent +CGs are relatively modest, 30–40 dBZ or less. Only a small subregion of trailing stratiform area produces sprite and elves. It would appear that this portion of the MCS possesses, for several hours, the requisite dynamical and microphysical processes favorable for the unique electrical discharges that drive TLEs. [Table 1](#) and [Table 2](#) summarizes the relationships between lightning and the major TLE types.

The Causes of Transient Luminous Events

TLEs have captured the interest of many theoreticians. Several basic mechanisms have been postulated to explain the observed luminous structures. These include sprite excitation by a quasi-electrostatic (QE) mechanism. Sprite production by runaway electrons in the strong electric field above storms has been suggested. The formation of elves from lightning electromagnetic pulses is now generally accepted. More than one mechanism may be operating, but on different temporal and spatial scales, which in turn produce the bewildering variety of TLE shapes and sizes. Absent from almost all theoretical modeling efforts are specific data on key parameters characterizing lightning flashes that actually produce TLEs. Many modelers refer to standard reference texts, which, in turn, tend to compile data taken in storm types and locations that are not representative of the nocturnal High Plains. Specifically, many invoke the conventional view that the positive charge reservoir for the lightning is found in the upper portion of the cloud at

Table 2 Current speculations about characteristics of TLEs and their parent lightning (none for blue jets)

	<i>Sprite</i>	<i>Elve</i>	<i>Blue jet</i>
<i>Color of emission</i>	Red top/blue base	Red?	Deep blue
Polarity of parent CG	Positive	Positive (mostly)	None
+CG Peak current	> ~50 kA	> ~100 kA	None
Charge transferred	Largest	Large	N/A
Charge moment	Largest	Large	N/A
Parent CG location	Stratiform area	Stratiform area?	N/A
Parent CG vertical channel height	5 km?	10 km?	N/A
D I/DT value	?	Large?	N/A
Total flash duration	Very long	Short?	N/A
Spider involved	Yes?	No?	N/A
Continuing current duration	Very long?	Short if any?	N/A
VLF/ELF slow tail	Distinct	Possible	N/A
ELF spectral color	Red	White	None
VLF audio character	Low frequency	Higher frequency	None
Duration of TLE	1–150 ms	0.5 ms	100–200 ms
Altitude range of TLE	25–95 km	75–105 km	Cloud–40 km
Onset delay after CG	1–100 ms	0.325 ms	N/A
Brightness	50–1000 kR	1000 kR	1000 kR
Horizontal size of emission	100 m–100 km	100–400 km	~2 km

altitudes of ~10 km. The positive dipole (or tripole) storm model has been found wanting in many midcontinental storms. A survey was made of the range of lightning parameters used in over a dozen theoretical modeling studies. While the proposed heights of the vertical +CG channel ranges from 4 to 20 km, there is a clear preference for 10 km and above. The amount of charge lowered varies over three orders of magnitude, as does the time scale over which the charge transfer occurs. Only a few papers consider the possible role of horizontal components of the parent discharge. The charge moment (in C km) – not the peak current as measured by the National Lightning Detection Network (NLDN) – is the key parameter in the basic QE conventional breakdown mechanism first proposed by Wilson. The key physics of the problem appears to involve the altitude and magnitude of the removed charge and the time scale on which this occurs, parameters about which little agreement exists. Many theorists note that even with an assumed tall +CG channel (~10 km), extremely large (~100 C) charge transfers, typically ten times larger than in 'conventional' lightning, are required to trigger sprites. Some models yield a thousandfold enhancement in optical intensity at 75 km for a doubling of lightning charge removal altitude from 5 to 10 km. The use of shorter channels to ground, say 5 km, would imply truly large charge transfers. Evidence is accumulating that indeed such may be the case. Recent estimates of charge moments associated with sprite parent lightning +CGs suggest values are typically at least 300–600 C km, and often much larger.

While the various models simulate optical emissions bearing some (though in many cases rather minimal) resemblance to the observations, the wide ranges in the lightning

source term parameters used by modelers do not appear physically realistic. If, in fact, such a range of lightning characteristics could produce sprites, why does only a very small subset of +CGs (<1:20 even in active storms) actually produce observable TLEs (with current sensors)? It appears that most models have made assumptions about the lightning in order to produce something resembling a TLE, rather than starting with hard physical constraints on the source term. The reason, of course, is that there is very little data on the actual CGs that generate specific TLE occurrences.

Shortly after the confirmation in 1993 of sprites as a frequently occurring phenomenon, speculation began that meteors might serve as a trigger for sprites in the intense electric fields above large MCS. This notion was quickly dismissed during several summers of observations of meteor showers coincident with large MCS systems over the United States, which found no sprite–meteor pairings. However, in 1998, a small meteor descending below 100 km altitude appeared to have triggered a sprite above an MCS. No coincident CG lightning flash was observed and the sprite appeared to originate close to the end of the luminous meteor trail. Even more curiously, a jetlike feature appeared to move about 10 km back up the ionized trail left by the meteor. This (so far) unique event illustrates how complex and interrelated middle-atmospheric electrical discharge phenomena may be.

Experimental Measurements

Since the late 1990s more and more complex measurement programs have begun to be undertaken to study TLEs. One of the more important gaps in our knowledge concerns the electric environment above large thunderstorms during TLE episodes. Consequently, the 1999 NASA Sprite Balloon Campaign, directed by E.A. Bering of the University of Houston, conducted several high-altitude balloon flights. The last mission flew out of Ottumwa, Iowa, on the night of 20–21 August 1999. The balloon payload floated at 32 km and drifted westward at ~30 knots. Ground-based LLTV observations were made from three sites (in Wyoming, South Dakota, and Colorado). All three stations had clear skies. There were two small TLE-producing storms, one in eastern South Dakota and one in central Kansas. Of 67 TLEs seen by at least one station or the balloon, 5 were seen by two or more stations. The balloon data at a typical range of 300 km show that the sprite is accompanied by a positive vertical electric field pulse of $\sim 0.2 \text{ V m}^{-1}$. Curiously, no perturbation in any component of the electromagnetic fields was observed during the several milliseconds between the lightning flash and a sprite. Also, two very bright elves were detected by ground and balloon optical sensors. The NLDN, however, failed to compute any associated CG. Analysis of the raw network sensor data did reveal that the elve parent lightning events were each observed by over 75% of the NLDN's sensors nationwide. This powerful sferic was so complex as to prevent classification by the NLDN algorithms. Preliminary results from this flight have indicated that more data are required before we can understand the complex physics involved.

Until the summer of 2000, no lightning flash known to produce a sprite or elve had ever been well characterized. To

simulate complex TLEs, modelers require information on the total charge removed and its waveform, the continuing current characteristics, the rate and altitude from which charge was removed, and the geometry of the vertical and, especially, horizontal lightning channels. Between 22 May and 16 July 2000, a major field observation effort called the Severe Thunderstorm Electrification and Precipitation Study (STEPS) was undertaken. The STEPS domain, located 100–400 km east-south-east of Yucca Ridge, was ideally situated for acquiring a wide array of optical measurements of TLEs above the storm concurrently with the lightning discharges below. With the inclusion of the 3D lightning mapping array (LMA) from New Mexico Tech, analyses of STEPS data will facilitate determination of the unique lightning discharges that give rise to TLEs. STEPS may confirm whether massive 'spider' discharges are a necessary condition for sprites.

Scheduled programs during the early 2000s include a joint US–Israeli sprite-observing mission on the Space Shuttle. Plans are also under way for Taiwan to launch a sprite-monitoring satellite.

Why Study TLEs?

Apart from their intrinsic scientific interest, there may be some rather practical reasons to explore TLEs in more depth. It has been suggested that there may be significant production of nitrogen oxides (NO_x) in the middle atmosphere by sprites. This becomes even more interesting in the light of recent observations that regional smoke palls from biomass burning radically enhance the percentage of +CGs within storms, and thus increase sprite counts (and middle atmosphere NO_x production?). Currently, no global chemical model accounts for any potential effects of TLEs.

Once a better estimate of NO_x production per sprite is obtained, it will be necessary to know the global frequency and distribution of sprites. It has been demonstrated that several Schumann resonance monitoring sites working in tandem are capable of obtaining a worldwide TLE census.

There is growing interest in determining the sources of unusual infrasound emissions detected above sprite-capable MCSs as determined by NOAA's Environmental Technology Laboratory near Boulder, Colorado. TLEs thus produce optical, radiofrequency, and acoustic emissions that have the potential of mimicking or masking signatures from clandestine nuclear tests. Such findings may have important implications for global monitoring efforts supporting the Comprehensive Test Ban Treaty.

TLEs may contribute in ways not yet understood to the maintenance of the global electrical circuit. To quantify the impacts of TLEs, we require information on the global frequency (now roughly estimated between 1 and 10 per minute) and their geographic distribution. Recent findings may clarify some of the broader issues concerning any potential TLE hazards to aerospace safety above 15 km.

See also: Electricity in the Atmosphere: Global Electrical Circuit; Ions in the Atmosphere; Lightning.

Further Reading

- Corliss, W.R., 1983. *Handbook of Unusual Natural Phenomena*. Anchor Books/Doubleday, Garden City, NY.
- Franz, R.C., Nemzek, R.J., Winckler, J.R., 1990. Television image of a large upward electrical discharge above a thunderstorm system. *Science* 249, 48–51.
- Huang, E., Williams, E., Boldi, R., et al., 1999. Criteria for sprites and elves based on Schumann resonance observations. *Journal of Geophysical Research* 104, 16943–16964.
- Journal of Atmospheric and Solar-Terrestrial Physics: the May–June, 1998 issue was dedicated to TLEs and provides a valuable source of references.
- Lyons, W.A., 1996. Sprite observations above the U.S. High Plains in relation to their parent thunderstorm systems. *Journal of Geophysical Research* 101, 29641–29652.
- MacGorman, D.R., Rust, W.D., 1998. *The Electrical Nature of Storms*. Oxford University Press, New York.
- Rowland, H.L., 1998. Theories and simulations of elves, sprites and blue jets. *Journal of Atmospheric and Solar-Terrestrial Physics* 60, 831–844.
- Uman, M.L., 1987. *The Lightning Discharge*. Academic Press, New York.

Forensic Meteorology

LE Branscome, Climatological Consulting Corporation, FL, USA

© 2003 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 2, pp 793–798, © 2003, Elsevier Ltd.

Introduction

Forensic meteorology is the study of past weather events for the primary purpose of assisting a court of law in adjudicating disputes in which weather may have been a factor. The task of a forensic meteorologist is to diligently investigate the weather conditions pertinent to the litigation and objectively present to the court the most accurate description possible of the relevant meteorological events. Expert testimony by a meteorologist often provides crucial information in various types of civil and criminal proceedings.

Civil litigation is often initiated when a personal injury or financial loss occurs, or when a contractual dispute cannot be resolved. In many cases the weather is a contributing factor, or even the primary cause of the loss or dispute. For example, weather frequently contributes to maritime cargo losses, aircraft and highway accidents, property damage, construction delays, air pollution emergencies, slip and fall injuries, and structural failures. The party sustaining the loss or injury may seek compensation through the legal system from another party who is alleged to be responsible for the incident. A judge, jury, or other adjudicative body must determine the facts of the case and make a decision in favor of one of the parties in the litigation, either the plaintiff or the defendant. If the weather was a factor in the incident or dispute on which the lawsuit is based, then expert testimony about the weather conditions relevant to the case must be presented to the adjudicative body in a courtroom or other judicial setting. Meteorological analysis and testimony are sometimes necessary simply to rule out weather as a contributing factor. Together with other facts, testimony, opinions, exhibits, and evidence, the adjudicative body considers the relevant weather evidence and opinions in making its determination.

In addition to civil litigation, testimony about the weather is occasionally required in criminal proceedings. For example, the ability of an eyewitness to identify an alleged assailant in low natural lighting is occasionally an issue in criminal trials. In such cases a meteorologist may be called to testify about the sky conditions, phase of the moon, or sunrise and sunset times. Other criminal proceedings have depended on meteorological analysis and testimony as it relates to the decomposition of the bodies of murder victims, the transport of odors from illegal drugs, the death of children trapped inside motor vehicles exposed to intense sunlight, and various types of criminal negligence.

In a broader sense, the work of a forensic meteorologist is not necessarily limited to civil litigation and criminal prosecutions. For example, a windstorm that causes property damage may result in a claims dispute between an insured party and its insurer. A meteorologist may be hired by one of the parties in the dispute to ascertain the peak wind speed during the storm. The conclusions of the meteorologist may be sufficient to allow the client to make a determination about the

claim and settle the dispute with the opposing party, thus avoiding the effort and expense of a lawsuit. Similarly, a forensic meteorologist may be hired to provide information and expert opinion about past weather events to parties involved in administrative or regulatory disputes with government agencies.

The Practice of Forensic Meteorology

General Procedures and Considerations

Technical experts are often retained by attorneys representing the parties involved in a lawsuit. The immediate purpose of retaining a meteorologist is to obtain information and expert opinion about the weather conditions so that the attorney can evaluate, with assistance from the expert, the importance of weather to the case. The attorney will consider how the weather facts and opinions may support or refute theories for the causation of the loss that led to the litigation. If the meteorological information and opinions are relevant and helpful, the weather expert may be asked by the attorney to prepare a formal report, evaluate the reports and opinions of the meteorologist (if any) working for the opposing legal counsel, assist the attorney in evaluating the merits and demerits of the case from a technical perspective, develop exhibits for presentation at trial, and provide expert testimony and opinions at deposition and trial.

The expert investigating and testifying about the weather should be a meteorologist with sufficient knowledge, skill, experience, and education to offer an opinion about the particular weather conditions related to the litigation. Although meteorologists are not usually granted professional licenses like engineers or architects, the American Meteorological Society, for example, has a certification program for consultants, identified as Certified Consulting Meteorologists, that involves extensive testing and board review. Many attorneys find that the certification assists them in identifying qualified experts. The forensic meteorologist must also have a high degree of integrity and composure under pressure, so that the expert testimony that he or she gives in the courtroom is trustworthy, unbiased and professional.

The admissibility of expert opinions in the federal courts of the United States is governed by the 'Daubert test' in which the judge assesses whether the reasoning or methodology underlying the opinion is scientifically valid and the offered testimony is relevant to the case. Many non-federal courts still adhere to the earlier 'Frye test' in which expert opinions must be based on principles and techniques that are 'generally accepted' as reliable within the relevant scientific community.

Before accepting work as an expert, the forensic meteorologist should inquire about the nature of the case and the parties and attorneys involved, in order to avoid possible conflicts of interest. A conflict may arise if the expert has had prior exposure

to the facts and legal arguments of the case or if the opposing attorney is a regular client of the expert, particularly for cases of a similar nature. In some instances the expert should decline to accept the work if the case involves a special area of meteorology in which the expert has limited knowledge or training. For example, an expert in air pollution meteorology may refer the client to an aviation weather expert if the case involves an aircraft accident in a severe thunderstorm.

After accepting the case, the weather expert should obtain all the information that the attorney possesses that might be relevant to the meteorological investigation. This information will include, at a minimum, the date, time, location, and nature of the event that led to the litigation. It may also include reports and statements describing the event that were prepared by investigative agencies, law enforcement officers, or eyewitnesses. For example, the National Transportation Safety Board of the United States produces investigative reports that provide information about the particulars and contributing factors in aircraft accidents. The testimony of eyewitnesses regarding the weather conditions can sometimes provide descriptive details about a weather event that would not be evident in official meteorological data. The expert may also find it useful to visit the site of the incident to assess the importance of exposure, surrounding terrain, and site orientation to the analysis and interpretation of the meteorological events.

The Value of Meteorological Data and Analysis

Although the attorney retaining the expert may have a substantial amount of information about the event, a forensic meteorologist cannot offer a reliable opinion until weather data are obtained and reviewed. The nature and quality of the data needed to formulate expert opinions depends on the nature of the case and the relative importance of weather to the arguments and theories in the litigation. In the case of a slip and fall injury, it may only be necessary to obtain hourly rainfall data from a nearby weather station. In the case of a fatal aircraft accident, a wide variety of data may be needed to formulate a clear and complete description of the weather conditions at the time and location of the accident. The data may include surface weather observations, upper-air weather charts, satellite and radar images, lightning strike data, aviation weather forecasts, and pilot reports of weather conditions aloft.

Forensic meteorological investigations often depend on the analysis of severe weather phenomena that have significant variations over small temporal and spatial scales. For example, a severe thunderstorm can create strong winds and large hail that cause isolated property damage in a few seconds to minutes. The ability to reconstruct past weather for forensic purposes has been greatly enhanced by recent advances in observational techniques that target mesoscale weather phenomena. For example, the National Weather Service of the United States operates a network of Doppler weather radar stations that provide nearly complete and continuous coverage of the country. Similar radar networks are found in other countries. The radar monitoring of severe weather and the archiving of the radar data have allowed forensic meteorologists to develop more accurate and detailed descriptions of severe weather events.

Similarly, the weather is monitored continually from space by geostationary and polar-orbiting satellites. Improvements in the resolution of the on-board sensors and the addition of more observing channels at infrared wavelengths allows forensic meteorologists to investigate past weather events in greater detail. For example, combinations of data from different infrared channels permits the identification of potential aviation hazards such as fog or airframe icing. Other satellites are capable of measuring ocean wave heights and estimating surface wind speeds at sea, which is valuable information for the investigation of cargo and ship losses during intense storms.

Other advances in meteorology are related to improvements in computer models of the atmosphere. The ability to computationally simulate or analyze mesoscale weather phenomena has greatly improved. The global re-analysis of historical weather data with computer models that are also used in operational weather forecasts has provided another new tool for the forensic investigator.

Acquisition of Weather Data

The meteorological data used in forensic work are commonly stored in government archives and can be accessed by the general public for usually modest fees. The primary sources of data in the United States are the National Climatic Data Center, which is part of the National Oceanic and Atmospheric Administration, and Regional Climate Centers and State Climatologists. Similar government archives are operated by the weather services of other countries. In addition to national government sources, a number of local government agencies, universities, and private companies archive meteorological data and operate weather observing stations in special monitoring networks. While anyone involved in litigation can obtain weather data, a forensic meteorologist is usually more efficient in selecting and obtaining the relevant data and organizing it into a coherent source of information for the analysis of weather events.

The collection, archiving, and retrieval of meteorological data had undergone considerable change in recent years, affecting the practice of forensic meteorology. In the past most meteorological data were recorded and archived on paper or film. For example, wind data were recorded on instrument recorder charts, radar scope images on film, and surface observations on handwritten forms. The data and analysis techniques were often limited in quantity and complexity. Forensic meteorologists usually ordered, and in many cases continue to order, paper copies of data and images that bear a fixed certification and seal of authenticity from the government agency providing the data. The physical seal and certification of the data copy are still regularly used as a means of authenticating the copy for admission into evidence at trial. Production and delivery of the hard-copy data and images by the data supplier often resulted in waiting times of a few weeks to months. Analysis of the data by a forensic meteorologist was labor-intensive, particularly for cases in which years of data had to be reviewed to determine normal climatological conditions, since the data had to be extracted from the certified paper forms.

The recording, storage, and delivery of data on paper, microfiche, and film is rapidly diminishing. Low-cost computer storage of data, data-intensive observational systems such as

Doppler weather radar and multichannel satellite sensors, complex computer models of the atmosphere and ocean, digital surface weather sensors, computer networks, and the Internet are radically changing the work of the forensic meteorologist. Many government and private data sets are now maintained on computer servers on the Internet and are easily accessible at low or no cost to the end user. Even large data sets, such as satellite and radar data, can be extracted from computer tapes at government archives and delivered to the end-user by file transfers over the Internet. Many data providers are also providing large data sets on CD-ROM and DVD. Large quantities of data can thus be retrieved and analyzed on the computer workstations of forensic meteorologists. Waiting periods for data delivery have generally been reduced from weeks and months to minutes and days. While the time required to acquire and analyze the data on a per-unit basis has been substantially reduced, the quantity and variety of data available for forensic work has increased dramatically, so that the total amount of effort in a typical investigation has not diminished. Even so, improved access to data often allows experts to provide their clients with preliminary analysis and opinions in a time frame shorter than previously possible.

The physical certification of digitized data is often impractical, if not impossible. The authenticity of digitized meteorological data as admissible trial evidence is usually not problematic, provided the testifying expert can demonstrate that the data was obtained from reliable sources and is customarily used by meteorologists in their work.

Reports and Testimony

After the weather data has been obtained and analyzed, a report is made to the client either verbally or in writing. A formal report that includes expert opinions is sometimes required by the client or by the judicial rules governing the litigation. The report becomes a basic reference document for future testimony by the expert.

Before a case goes to trial, a discovery period occurs when the attorneys in the litigation request copies of the documents and information used or prepared by the other side's technical experts. The information includes the data gathered and any written reports prepared by the weather expert. Furthermore, the expert is questioned in a deposition by the opposing legal counsel to discover the expert's opinions and the foundational basis for those opinions. The discovery and evaluation of the weather data and expert opinions sometimes encourage a settlement between the opposing parties, particularly if the weather was the primary cause of the loss or dispute that initiated the lawsuit.

As the trial approaches, the expert prepares exhibits that display the data and information relevant to the issues of the case. The exhibits are often directly derived from diagrams, tables, and images found in the expert's report. The availability of digitized data has led to improvements in the presentation of meteorological information in the courtroom. Radar and satellite digital images and computer simulations of weather events can easily be annotated for trial presentation and stored on a CD-ROM. Using a high-resolution projector attached to a laptop computer, the testifying expert can display the images in the courtroom using animation sequences, stop-action

frames, zoom views, and various enhancements. Presentations of this kind allow the jury or judge to arrive at a better understanding of the weather conditions associated with the event that led to the litigation. In fact, many lay people who serve on juries see animated satellite and radar images on the weather segments of television news shows and are somewhat familiar with basic meteorological concepts related to these observations, prior to seeing such data in the courtroom.

The expert should assist the attorney in preparing for the expert's trial testimony. The attorney needs to have a thorough understanding of the expert's opinions, along with the basis for those opinions, so there are no surprises in the courtroom. The testimony and opinions of the expert, together with the trial exhibits, are presented during the direct examination by the attorney who retained the expert. The expert has the responsibility of clearly and simply explaining the weather elements of the case to the judge or jury. Cross-examination of the expert by the opposing attorney is standard practice. If some points of the testimony need to be clarified following the cross-examination, the attorney who retained the expert has an opportunity to ask additional questions in redirect examination. Once the expert's courtroom testimony is finished, the expert's participation in the litigation usually ends, unless appeals in the matter require a re-hearing of the expert's opinions.

Examples of Meteorological Investigations

A tragic aviation accident, in which weather played a critical role, was the crash of a commercial aircraft, American Eagle flight 4184, near Roselawn, Indiana, on 31 October 1994. The aircraft was in a holding pattern at 3000 m altitude for about 30 minutes while waiting for clearance to land at Chicago's O'Hare airport. Shortly after it was released by air traffic controllers from its holding pattern, it descended toward 2500 m, at which time the pilots could no longer control the aircraft and it crashed into a field, killing all 68 people on board.

Airframe icing can significantly degrade the performance of an aircraft. Based on the findings of detailed meteorological investigations of the Roselawn accident, the cause for the loss of control is strongly suspected to have been the accumulation of a ridge of ice behind the leading edge of the wings. Because the wings of the aircraft operating as Flight 4184 were located at the top of the fuselage, the pilots were unable to see the formation of ice on top of the wings. Relatives of the deceased passengers sued the airline partly over the alleged inaction of the pilots with respect to the hazardous icing conditions. Attorneys in the litigation retained aviation weather experts and much of the meteorological investigation focused on the nature, timing, and severity of the icing conditions. The microphysics of supercooled drizzle drops was an important aspect of the investigation, since the atmospheric conditions were indicative of the presence of such drops in the area of the holding pattern. These drops can be particularly hazardous to aircraft because they can flow over and freeze behind the icing-prevention devices on the leading edge of the wings and create ice formations that seriously disrupt air flow over the wings. Data and images from the National Weather Service Doppler radar near Chicago (see example in [Figure 1](#)) were carefully studied to determine when the aircraft most likely encountered

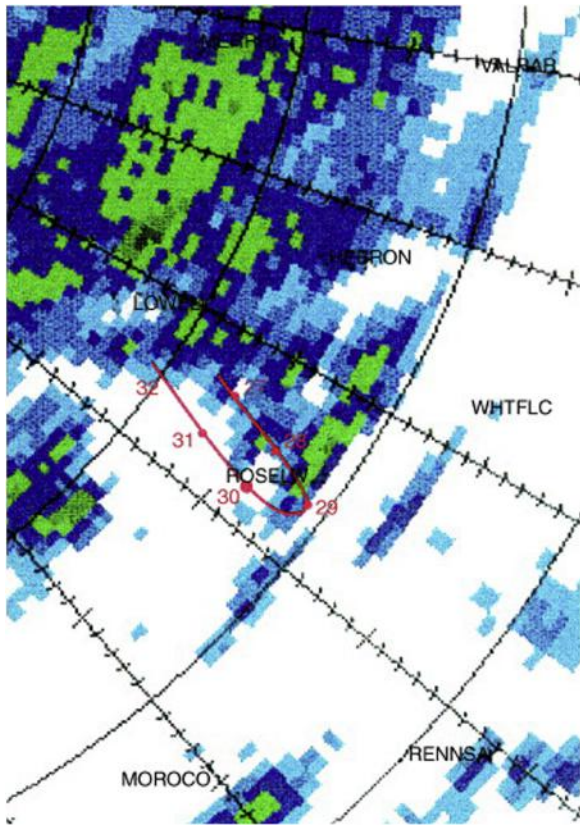


Figure 1 Track of American Eagle Flight 4184 during part of its holding pattern near Chicago, Illinois, on 31 October 1994, plotted on a National Weather Service Weather Surveillance Radar reflectivity image. Time associated with each aircraft position is in minutes after 21:00 Universal Coordinated Time. The aircraft occasionally encountered areas of light precipitation, most likely supercooled drizzle drops, during its hold.

supercooled drizzle drops during the holding pattern. The meteorological analysis was partly intended to determine whether any early visible indications of icing were present that might have given the crew an opportunity to respond and take evasive action. A substantial amount of time and effort was spent on the preparation of trial exhibits related to the weather testimony, but the litigation was settled shortly before trial.

Commercial and private aircraft are also subject to hazardous weather associated with thunderstorms. The occurrence of microbursts (intense small-scale downbursts in the decaying stages of a thunderstorm) are of particular concern on approach or departure from an airport. Large changes in wind speed and direction across a microburst can create sudden and unexpected changes in lift. The ability of the pilots to recognize and avoid the hazardous conditions associated with a microburst are usually at the center of the litigation in such cases. Microbursts were the primary cause of several major accidents, such as the Delta Air Lines crash at Dallas, Texas, on 2 August 1985 and the US Airways crash at Charlotte, North Carolina, on 2 July 1994. Litigation related to these crashes relied not only on detailed analysis of the meteorological data but on eyewitness statements from pilots of other aircraft and air traffic controllers regarding the weather.

Severe turbulence can cause passenger injuries on commercial flights and such incidents often lead to lawsuits against the airlines operating the flights. Typically the plaintiff will allege that the pilots should have anticipated and avoided the turbulence. Aircraft operation in and around thunderstorms increases the likelihood of a severe turbulence encounter. Avoidance of thunderstorms is made possible through direct visual observation, on-board weather radar, and pre-flight and in-flight weather briefings. Other atmospheric phenomena, such as breaking mountain waves and strong vertical or horizontal shear in jet streams, can also cause severe turbulence, but such conditions are usually more difficult to anticipate and avoid than convective turbulence.

The examination of satellite, radar, and upper-air data is usually a critical component in meteorological investigations of turbulence incidents. Two passengers who allegedly suffered turbulence-related injuries during a flight across the southwestern United States in March 1996 sued the airline. The airline hired an aviation meteorologist who concluded that there was no meteorological evidence for the turbulence encounter. However, the meteorologist working for the plaintiff obtained satellite photographs that showed the existence of strong mountain waves at the time and location of the incident, consistent with the plaintiffs' allegation. The case was settled shortly after the report of the plaintiff's expert was submitted in the litigation.

Sometimes unofficial weather measurements are at least as valuable as official records. The construction of a natural gas pipeline in South America was interrupted by heavy rain and destructive flooding in April 1998. The construction company was insured against losses of this kind and filed a large claim for extensive damage and delay associated with the event. The insurer was uncertain about paying the claim since rainfall totals from official weather stations in the region were not sufficiently large to have caused the flood damage claimed by the construction company. The insured party produced an unofficial measurement of 325 mm of rain in one day from a hotel located near the project. The rainfall reading at the hotel was many times larger than any of the official readings. In order to resolve the validity of the claim, the insurer hired a meteorologist to investigate the weather conditions, as well as forensic civil engineers who inspected the physical damage. An examination of weather satellite data demonstrated that an isolated severe thunderstorm with exceptionally high cloud top heights (Figure 2) did occur along the segment of the pipeline where the most severe damage was found. A range of probable rainfall amounts was inferred from satellite data and was generally consistent with the high rainfall reading at the hotel. Although the heavy rainfall did not cover as much of the pipeline project as claimed by the construction company, the insurer concluded that at least part of the claim was valid.

In another example, the owner of a Florida resort hotel claimed that extensive water damage to the interior of the building was the result of roof leaks during a very heavy rainfall event. The hotel owner filed a lawsuit against a construction company that had repaired the roof shortly before the alleged damage. Attorneys for the construction company retained a forensic meteorologist to determine whether a heavy rainfall event actually occurred at the property during the period in which the water damage occurred. Rain gauges in the area did not confirm the

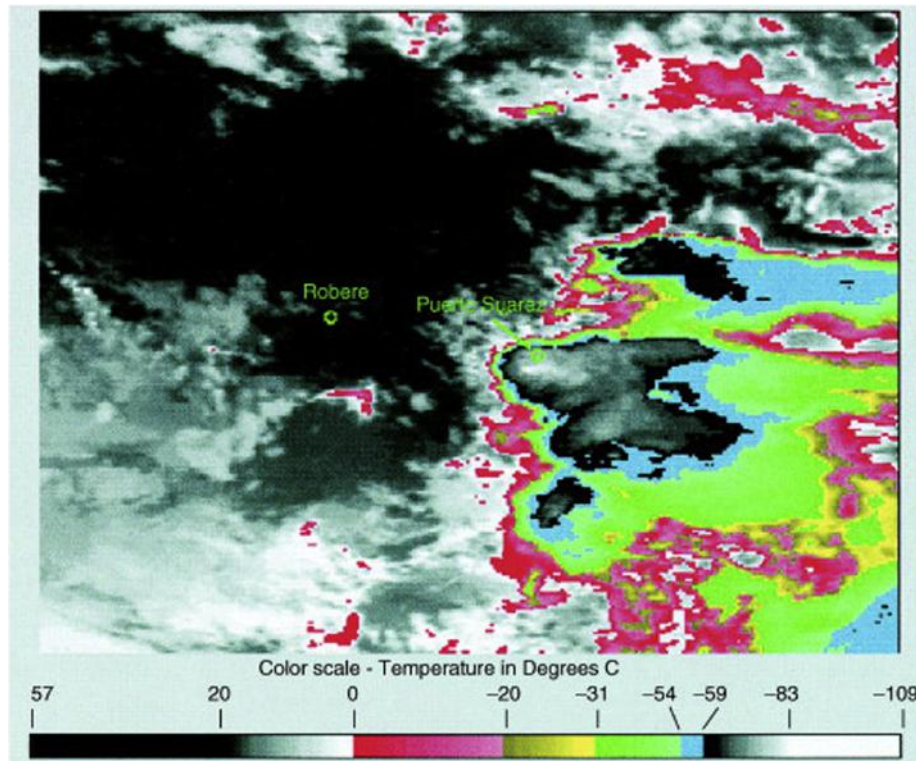


Figure 2 Color-enhanced infrared satellite image from Geostationary Operational Environmental Satellite 8 on 8 April 1998 at 11:45 Universal Coordinated Time near Puerto Suarez, Bolivia. Lower temperatures indicate higher cloud tops. Note the very cold temperatures for the heavy thunderstorm near the center of the image. Heavy rain from this thunderstorm caused flooding and extensive damage to a pipeline construction project.

occurrence of a rain event of the magnitude alleged by representatives of the hotel. The amounts were far below the quantity of water associated with the interior damage. Attorneys for the hotel asserted that since isolated heavy rainfall often occurs in Florida, the rain gauges simply missed the 'deluge' at the hotel. Examination of data from the National Weather Service Doppler radar station that monitors the region demonstrated conclusively that the rainfall amounts at the hotel were similar to the amounts measured at the rain gauges. The case reached trial and the meteorologist testified about the rainfall and radar analysis. The jury found in favor of the defendant, i.e., the construction company, concluding that it was impossible that rain could have caused the water damage to the interior of the hotel.

The Outlook for Forensic Meteorology

The density and value of physical property is steadily increasing in regions especially susceptible to weather damage, such as coastal areas. While significant improvements are being made in aviation safety procedures, particularly related to the distribution and analysis of weather information for pilots and air traffic controllers, the frequency of air travel is also expected to grow rapidly. Furthermore, judicial systems are generally increasing the monetary value assigned to human life and health. As a result, the amount of litigation related to weather is likely to expand, increasing the need for forensic meteorological services.

With the expectation of additional enhancements in the quality and quantity of data from remote sensors and computer

models of the atmosphere and ocean, the work of the forensic meteorologist will become more complex and require continual upgrades in professional skills and knowledge. On the other hand, advances in observational and analytical techniques will likely also reduce uncertainties regarding past weather conditions and, therefore, enhance the value of meteorological research and testimony with respect to litigation.

See also: **Air Sea Interactions:** Surface Waves. **Aviation Meteorology:** Aviation Weather Hazards; Clear Air Turbulence. **Clouds and Fog:** Fog. **Hydrology, Floods and Droughts:** Flooding. **Mesoscale Meteorology:** Gust Fronts; Microbursts. **Mountain Meteorology:** Lee Waves and Mountain Waves. **Radar:** Mesosphere–Stratosphere–Troposphere and Stratosphere–Troposphere Radars and Wind Profilers; Precipitation Radar. **Synoptic Meteorology:** Weather Maps. **Weather Forecasting:** Severe Weather Forecasting.

Further Reading

- Bradley, M.D., 1983. *The Scientist and Engineer in Court*, Water Resources Monograph Series, 8. American Geophysical Union, Washington DC.
- Bronstein, D.A., 1993. *Law for the Expert Witness*. Lewis Publishers, Boca Raton, FL.
- Falconer, P.D., Haggard, W.H., 1990. Forensic meteorology. In: Wecht, C.H. (Ed.), *Forensic Sciences*, Ch. 35B. Matthew Bender, New York.
- National Climatic Data Center, 1999. *Weather records in litigation*. Environmental Information Summaries, vol. C-1. National Oceanic and Atmospheric Administration, National Climatic Data Center, Asheville, NC.

GENERAL CIRCULATION OF THE ATMOSPHERE

Contents

Overview

Angular Momentum of the Atmosphere

Energy Cycle

Weather Regimes and Multiple Equilibria

Mean Characteristics

Teleconnections

Overview

JM Wallace, University of Washington, Seattle, WA, USA

DWJ Thompson, Colorado State University, Fort Collins, CO, USA

P Beresford, European Centre for Medium-Range Weather Forecasts, Reading, UK

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

The atmospheric general circulation connotes the motion of the atmosphere, as viewed from a global, long-term perspective. This term is usually applied to the seasonally-varying, climatological-mean circulation, including the statistical properties of the variability of the circulation on day-to-day, week-to-week, and month-to-month time scales, but it also encompasses year-to-year variability and long-term variations associated with climate change. A unifying concept in studies of the general circulation are the so-called balance requirements for the conservation of angular momentum, total energy, and the mass of water vapor and other trace substances. Also of interest are the processes that contribute to the generation and dissipation of the kinetic energy, the exchange of kinetic energy between the zonally averaged circulation and the longitudinally varying 'eddy', the exchange of air between troposphere and stratosphere, and the long range transport of trace substances.

The atmospheric general circulation encompasses the planetary-scale wind systems that shape the Earth's climate. Features of interest include the belts of midlatitude westerlies and subtropical trade winds at the Earth's surface, the jet streams aloft, and the storm tracks. Most of these features are apparent in the figures discussed in this article. They are summarized in the schematic in [Figure 1](#).

The general circulation can be partitioned into zonally symmetric and eddy components, where 'zonally symmetric' denotes longitudinally (*zonally*) averaged (i.e., averaged around latitude circles) and 'eddy' denotes departures from the zonal average. By construction, the zonally symmetric flow is a function of latitude and height only. Certain properties of the eddies, such as the root mean squared amplitude of the eddies in various fields, can also be zonally averaged and displayed in the form of two-dimensional (latitude vs height) cross sections. Understanding the climatology of such zonally averaged fields is one of the principal goals of general circulation research.

[Figure 2](#) shows the zonally symmetric component of the zonal wind and temperature fields for (1) December through

February (DJF) and (2) June through August (JJA). At the Earth's surface, westward (i.e., east to west) 'trade winds' prevail equatorward of 30° latitude, while eastward wind prevails at higher latitudes. The eastward winds strengthen with height and reach a peak at the 10 km level just poleward of 30° latitude in both hemispheres. These zonal wind maxima commonly referred to as the 'tropospheric jet streams' lie along the boundary between troposphere and stratosphere. Dry, ozone-rich stratospheric air is found above and poleward of the tropospheric jet streams; moist, ozone-poor tropospheric air is found below and equatorward of them. Tropospheric and stratospheric air can also be distinguished in [Figure 2](#) by the marked difference in lapse rate, which determines the vertical spacing between the isotherms.

The jet streams are strongest during winter, when the north-south temperature gradient in the troposphere is strongest. The even stronger so-called polar night jet is observed in the middle and upper stratosphere of the winter hemisphere, separating the cold, dark polar cap region from the less cold sunlit portion of the hemisphere. The Southern Hemisphere

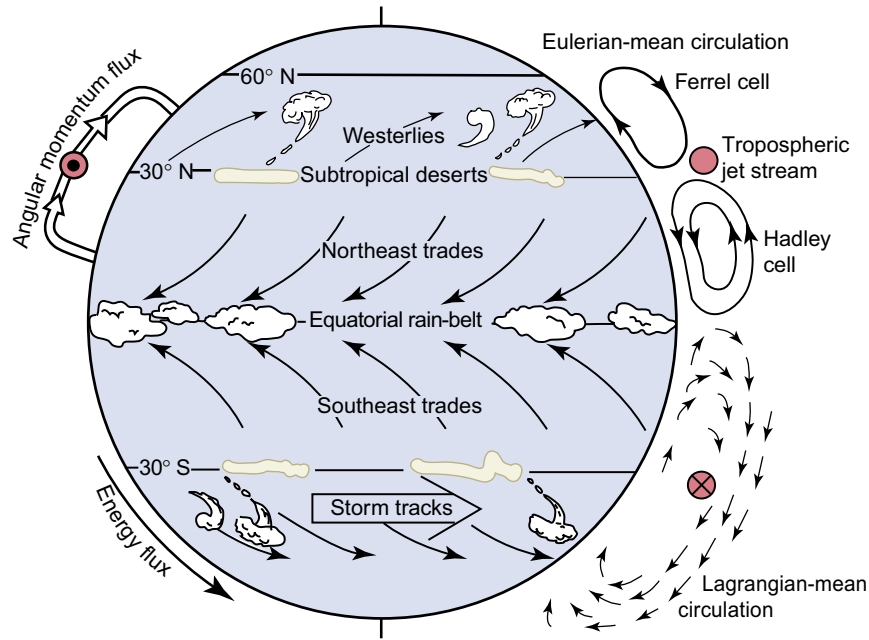


Figure 1 Schematic showing some of the essential features of the zonally averaged general circulation as they would appear on an idealized ocean covered but otherwise 'Earthlike' planet under equinox conditions. The red arrows pointing into the page at right and out of the page at left represent the eastward tropospheric jet streams. © Elsevier.

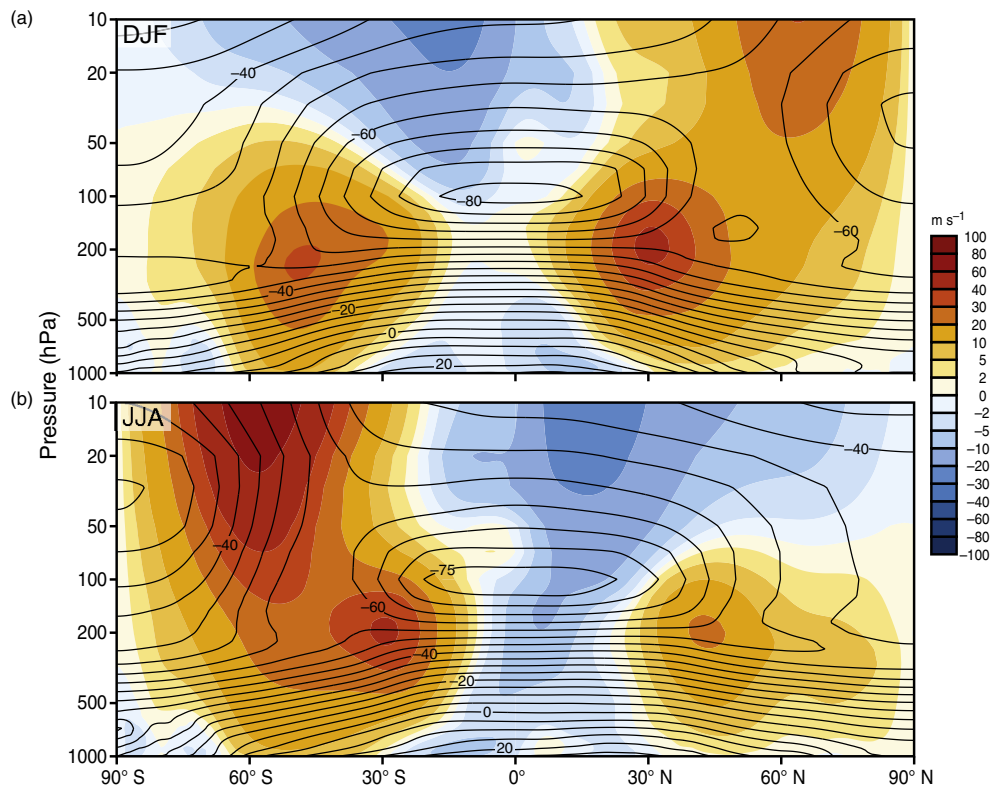


Figure 2 Pole-to-pole cross sections showing zonally averaged and time-averaged temperature (dashed contours, 5 °C interval) and zonal wind (shading) for the seasons (a) December through February (DJF) and (b) June through August (JJA). Based on the European Centre Reanalysis (ERA)-Interim Reanalyses. © Elsevier.

polar night jet is much stronger than its Northern Hemisphere counterpart.

Figure 3 shows the zonally averaged mean meridional (i.e., north/south) circulations for the same seasons as Figure 2 and for the annual mean. The trade wind belt is characterized by equatorward flow at the Earth's surface; the belt of surface eastward winds at higher latitudes is characterized by poleward flow. The low-level meridional flows constitute the lower branches of closed circulation cells extending through the depth of the troposphere. By far the strongest of these cells is the tropical 'Hadley cell,' whose rising branch delineates the belt of heaviest tropical rainfall and whose sinking branch is closely associated with the desert regions. Weaker 'Ferrel cells' in which air circulates in the opposite sense to the Hadley cell, are discernible in midlatitudes.

The rising branch of the Hadley cell is located at tropical latitudes in the summer hemisphere, while the sinking branch is located at cooler subtropical latitudes in the winter hemisphere. Circulations such as the Hadley cell are said to be 'thermally direct' or 'thermally driven,' since they are characterized by the rising of warmer air and the sinking of cooler air, as in convective cells. In contrast, circulations like the Ferrel cell are said to be 'thermally indirect' or 'thermally damped.' They are marked instead by the rising of air in cooler regions and the sinking of air in warmer regions.

In the tropics, the mean meridional circulation is strongly seasonally dependent and straddles the equator during the solstitial seasons DJF and JJA. However, when the data are averaged over the whole year, a different pattern emerges with rising motion near the equator, poleward flow in the upper

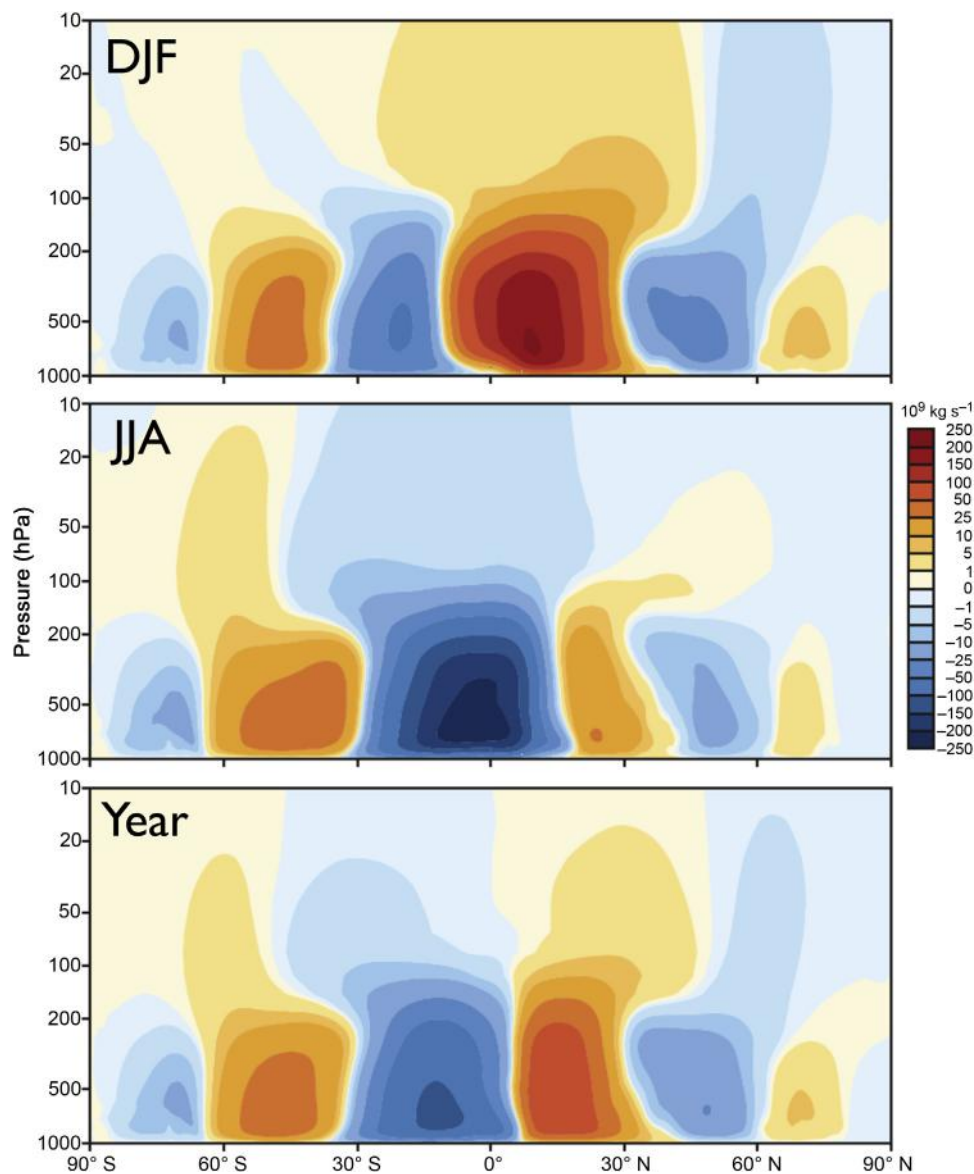


Figure 3 Pole-to-pole cross sections showing zonally averaged and time-averaged mean meridional circulations for the seasons December through February (DJF) and June through August (JJA), and for the calendar year (Year). Based on the ERA-Interim Reanalyses. © Elsevier.

troposphere, sinking motion in the subtropics and equatorward flow in the trade wind belts just above the Earth's surface.

The circulation cells pictured in Figure 3 represent time averages and zonal averages of wind measurements taken at fixed points in space. The circulation patterns derived from such measurements are said to be 'Eulerian mean circulations.' Composites of the trajectories of numerous individual air parcels, averaged over longitude and over an extended period of time give a different impression of the mean meridional circulations. As viewed from this so-called Lagrangian perspective, the Hadley circulations extend into high latitudes and the midlatitude Ferrel cells virtually disappear, as indicated on the lower right quadrant of Figure 1.

Balance Requirements

The general circulation must satisfy a number of so-called balance requirements relating to fundamental conservation laws. The balance requirements can be posed as a series of statements that for any specified region (e.g., the tropics or the Arctic polar cap region), the imports, exports, sources, and sinks of any given conserved quantity must sum to zero when averaged over a sufficiently long time. For the atmosphere as a whole, in the absence of long-term trends, sources must equal sinks.

Kinetic Energy

Frictional dissipation observed within the planetary boundary layer and within patches of turbulence within the free atmosphere is continually depleting the kinetic energy of planetary-scale wind systems. Half the energy would be gone within a matter of days were there not some mechanism continually operating to restore it. The source of this kinetic energy is the 'available potential energy (APE)' inherent in the distribution of atmospheric mass. The size of the APE reservoir depends upon the height of the atmosphere's center of mass relative to mean sea-level. APE is released whenever the center of mass is lowered through the sinking of colder, denser air and the rising of warmer, less dense air in thermally direct circulations like the Hadley cell. Such circulations also act to flatten out the potential temperature surfaces, thereby weakening existing horizontal temperature gradients on pressure surfaces.

The kinetic energy released by such large scale, thermally direct circulations is imparted not to the vertical component of the motion, which is so small as to be inconsequential with respect to kinetic energy, but directly to the horizontal component of the flow, which is pushed across the isobars from higher to lower pressure by the horizontal pressure gradient force. Such cross-isobar flow toward lower pressure is particularly strong close to the Earth's surface, where the dissipation of kinetic energy is most intense. For example, the trade winds in the lower branch of the Hadley cell (Figures 1 and 3) exhibit a component directed down the pressure gradient, out of the subtropical high-pressure belt and into the belt of low pressure that coincides with the rising branch of the Hadley cell in equatorial latitudes. The poleward flow in the upper branch of the Hadley cell is also down the pressure gradient, as evidenced by the existence of eastward winds at

that level, which implies (from the geostrophic wind equation) that pressure decreases with latitude.

Thermally direct circulations like the Hadley cell that are characterized by the rising of warmer, lighter air and the prevalence of cross-isobar horizontal flow toward lower pressure, thus release APE and convert it to the kinetic energy of the horizontal flow. Thermally indirect circulations like the Ferrel cell can be viewed as 'mechanically forced,' and cycle energy in the opposite direction. In the absence of diabatic heating and friction, the sum of the available potential and kinetic energy is conserved: for example, for a thermally direct circulation, the APE released is equal to the kinetic energy generated.

For the atmosphere as a whole, thermally direct circulations are prevalent. Since thermally direct circulations are continually depleting the atmosphere's reservoir of APE reservoir, something must be operating to restore it. Heating of the atmosphere by radiative transfer and by the release of the latent heat of condensation of water vapor in clouds acts to restore the APE in two ways: (1) by warming the atmosphere in the tropics (where the sum of the condensation heating and the absorption of incoming solar radiation exceeds outgoing infrared radiation) and cooling it at higher latitudes, where the reverse is true; and (2) by heating the air in the lower and middle troposphere, where most of the condensation heating takes place, and cooling it at higher levels, where infrared cooling to space prevails.

Mechanism (1) acts to maintain the equator-to-pole temperature contrast on pressure surfaces. Mechanism (2) acts to expand the air in the lower troposphere and compress the air in the upper troposphere, thereby lifting the air at intermediate levels, which acts to maintain the height of the atmosphere's center of mass against the lowering produced by thermally direct circulations. Hence, the maintenance of the atmospheric general circulation requires both horizontal and vertical heating gradients.

The above can be summarized in terms of a 'kinetic energy cycle' as depicted in Figure 4, with generation of APE (G) by diabatic heating, conversion of APE to kinetic energy (C) by thermally direct circulations, and the dissipation of kinetic energy (D). In the long-term mean, for the atmosphere as a whole, the rate of dissipation of kinetic energy is sufficient to deplete the global reservoir of kinetic energy in only about a week. If APE as well as kinetic energy is taken into account, the timescale for depleting (or recharging) the system is on the order of a month.

Angular Momentum

The angular momentum of an air parcel is given by $(\Omega R \cos \varphi + u)R \cos \varphi \, dm$, where Ω is the angular velocity of

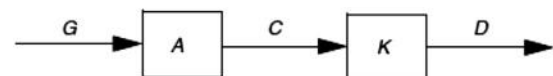


Figure 4 Kinetic energy cycle showing the reservoirs of available potential energy (A), and kinetic energy (K), the generation of available potential energy by horizontal gradients of diabatic heating (G), the conversion of available potential energy to kinetic energy in thermally direct circulations (C), and the frictional dissipation (D). © Elsevier.

the Earth's rotation, R is the radius of the Earth, ϕ is the latitude, and dm is the mass of the air parcel. Apart from small tidal interactions with the Moon, the total angular momentum of the atmosphere plus oceans plus solid Earth is conserved. Despite its enormous mass, the angular momentum of the ocean is very small, owing to the inhibition of circumpolar currents by the continents. Hence, whenever the atmosphere gains angular momentum it tends to be at the expense of the solid Earth, and vice versa. A strong correlation between length of day and atmospheric angular momentum is, in fact, observed on timescales ranging from days up to a few years. On longer timescales, slow motions within the Earth's molten core also affect the length of day.

Eastward winds circulate around the Earth's axis in the same sense as the Earth's rotation. Hence, air parcels in the atmosphere's eastward wind belts rotate more rapidly than the solid Earth does, and air parcels in westward wind belts rotate more slowly. It follows that the frictional drag that is acting to slow the tropical trade winds has the effect of transferring angular momentum from the solid Earth to the atmosphere. In a similar manner, frictional drag on the surface eastward flow that prevails poleward of 30° latitude transfers angular momentum from the atmosphere back to the solid Earth.

The torques (force times distance from the Earth's axis) that the atmosphere exerts upon the solid Earth by virtue of the atmospheric pressure difference between the east and west slopes of large, north-south oriented mountain ranges like the Rockies and Andes also transfer angular momentum between the atmosphere and solid Earth. This effect tends to be of the same sign as the frictional torques on the surface winds. Hence, there exists a balance requirement for a poleward transport of angular momentum within the atmosphere. The transport must be largest near 30° latitude, which marks the transition between the tropical trade winds and the eastward surface flow at midlatitudes, as depicted in Figure 2.

Angular momentum can be transported poleward across a latitude circle by either (1) a systematic poleward flux of atmospheric mass, or (2) 'exchange processes' in which poleward-moving air parcels carry more angular momentum (i.e., have a stronger eastward wind component) than equatorward-moving parcels. The net mass flux in the Earth's atmosphere is very small, as evidenced by the fact that mean

surface pressure over the region poleward of 30° latitude is not systematically changing with time. Additionally, there are no appreciable internal sources and sinks of atmospheric mass other than water vapor, which accounts for too small a fraction of the mass of the atmosphere to yield an appreciable transport. Hence, exchange processes must be responsible for the poleward transport of angular momentum.

The exchange processes, in turn, may be divided into two kinds: (1) those involving the Eulerian mean meridional circulations pictured in Figure 3 (and depicted in the upper right quadrant of Figure 1); and (2) those involving the eddies, as depicted in the upper left quadrant of Figure 1. In the annual average, 30° latitude coincides with the boundary between the Hadley and Ferrel cells. Hence, the mean meridional motions cannot contribute to the required poleward transport across that latitude: the eddies must play a key role in fluxing angular momentum from tropical to extratropical latitudes.

The distribution of the northward flux of eastward momentum by the eddies is shown in Figure 5. As inferred above, the eddies do, in fact, exhibit a systematic poleward transport of angular momentum across 30° . The flux is just enough to satisfy the balance requirement inferred from the frictional torques on the surface winds and the mountain torques. To accomplish the required transport, the horizontal flow in the eddies must exhibit a north-south tilt, as pictured in Figure 6. Such a tilt is widely observed in the atmospheric eddies.

It is evident from Figure 5 that most of the poleward transport of angular momentum takes place around the jet stream (~ 10 km) level, where the amplitude of the eddies is largest and the eastward tilt with increasing latitude is most pronounced. The Hadley cell is instrumental in transporting the angular momentum acquired by the trade winds upward to the jet stream level. The air in its rising (equatorial) branch contains much more angular momentum per unit mass than the air in its sinking (subtropical) branch due to its larger distance from the Earth's axis of rotation. The vertical exchange of air parcels of equal mass, but containing differing amounts of angular momentum per unit mass results in a net upward transport of angular momentum in the Hadley cell. In a similar manner, the Ferrel cell is instrumental in transporting eastward angular momentum downward from the jet stream level to the

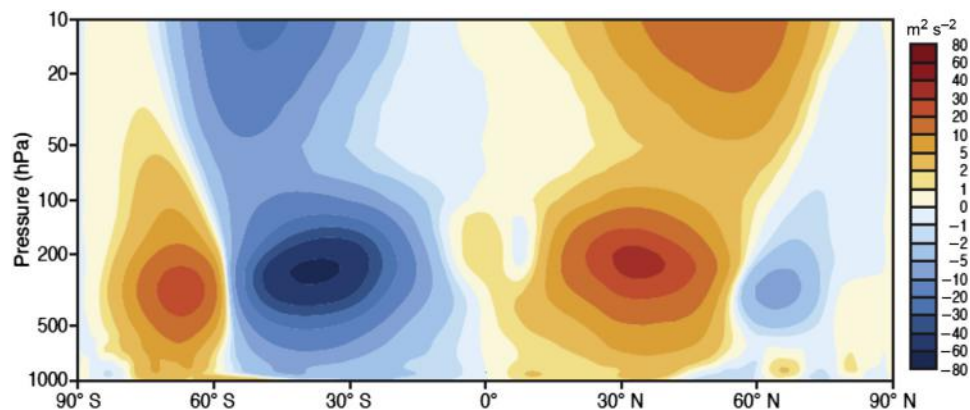


Figure 5 Annual mean northward flux of eastward momentum due to the horizontal motions in eddies, expressed in units of $\text{m}^2 \text{s}^{-2}$ (i.e., per unit mass and per unit length of the latitude circle). Based on the ERA-Interim Reanalyses. © Elsevier.

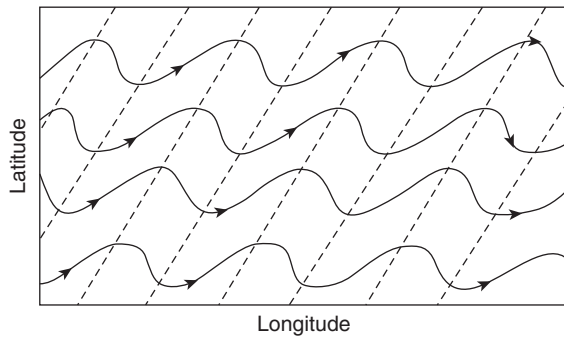


Figure 6 Idealized sketch showing the relationship between the tilt of the eddies and the poleward transport of eastward momentum. In order for poleward-moving air to have a stronger eastward wind component than the equatorward return flow, the eddies must have a southwest–northeast tilt, as shown. © Elsevier.

surface where it is lost to the solid Earth via frictional and mountain torques. The resulting upward/poleward/downward transport of angular momentum is depicted in the upper left quadrant of [Figure 1](#).

Total Energy

Averaged over the year, the net radiation at the top of the atmosphere is downward at low latitudes and upward at high latitudes. These local imbalances reflect the differing meridional profiles of insolation (incoming solar radiation) and outgoing terrestrial radiation pictured in [Figure 7](#). The former slopes steeply downward toward the poles, reflecting the strong latitudinal dependence of solar zenith angle. The latter mirrors the more gently sloping profile of the infrared radiation emitted to space from clear air, cloud tops, and the underlying surface. Since the equator-to-pole temperature gradient is not observed to be increasing systematically with time, these local imbalances must be fully compensated by the poleward transport of energy from the low latitude ‘surplus’ region to

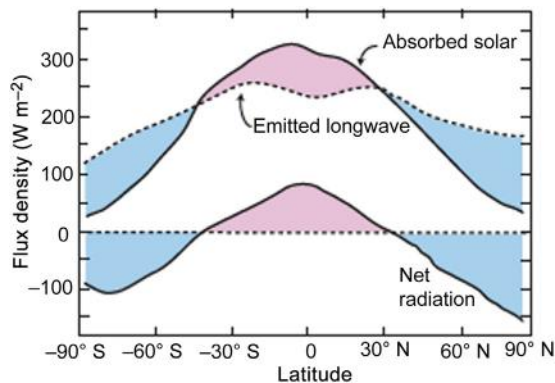


Figure 7 Annual mean net incoming solar radiation and outgoing terrestrial radiation as a function of latitude, expressed in units of W m^{-2} . Distance on the latitude scale is proportional to area on the Earth’s surface. From Wallace, J.M., Hobbs, P.V., 2006. *Atmospheric Science: An Introductory Survey*, second ed. Elsevier, 483pp.

the high latitude ‘deficit’ region, as depicted in the lower left quadrant of [Figure 1](#). The transport must be largest in midlatitudes where the curves in [Figure 7](#) cross one another.

The atmosphere and oceans both contribute to the poleward transport of energy. In both media, the transport involves exchange processes, with poleward-moving fluid parcels carrying larger amounts of total energy than equatorward-moving parcels by virtue of their having recently been heated while residing at low latitudes. In the ocean, both the shallow, wind-driven gyre circulations and the deeper thermohaline circulation contribute to the transport of energy: the gyres by transporting warm water poleward in the western boundary currents and cooler water equatorward on the eastern side of the oceans; the thermohaline circulation by transporting relatively warm surface water poleward and cold bottom water equatorward, primarily in the North Atlantic.

In the atmosphere, much of the total required transport is accomplished by eddies, in which poleward-moving air parcels carry with them greater amounts of sensible and latent heat than do equatorward-moving parcels. The thermally indirect Ferrel circulation does not transport energy poleward across middle latitudes: when the gravitational potential energy of air parcels is taken into account as well as the latent and sensible heat, the equatorward-moving air parcels in the upper branch of the Ferrel cell carry with them more energy per unit mass than the warmer, more moist poleward-moving parcels in its lower branch.

Eastward-moving baroclinic waves, such as that depicted in [Figure 8](#), are responsible for most of the poleward eddy transport of sensible and latent heat across middle latitudes. These waves derive their energy from the prevailing north–south temperature gradient (*see Dynamical Meteorology: Baroclinic Instability*). At the Earth’s surface, they are marked by intensifying cyclones (gyres that circulate in the same sense as the Earth’s rotation) attended by sharp frontal boundaries that separate warm, poleward-moving air masses on the cyclones’ eastern flanks from cold, equatorward-moving air masses on their western flanks ([Figure 8](#)). The meridional displacements of warm and cold air masses in the waves serve to sharpen the east–west temperature gradients. Meanwhile, the rising of the poleward-flowing warm air masses, in combination with the sinking of the equatorward-flowing cold air masses, serves to release APE, providing the kinetic energy needed to amplify the waves. Baroclinic waves tilt westward with increasing height: the wave troughs in the pressure field at the 5 km level overlie the cold air masses at the Earth’s surface.

In the Northern Hemisphere wintertime, planetary-scale stationary waves (sometimes also referred to in the general circulation literature as ‘standing eddies’) forced by the large thermal contrasts between the cold continents and the warmer oceans and by the blocking effect of the Rockies and the Himalayas also make an appreciable contribution to the poleward transport of sensible heat (*see Dynamical Meteorology: Stationary Waves (Orographic and Thermally Forced)*). The Icelandic and Aleutian lows are prominent stationary wave features in the wintertime mean sea-level pressure pattern. Poleward flow to the east of these low-level cyclones carries warm air northward, keeping western Europe and coastal southeast Alaska relatively warm compared to other regions at the same latitude, and the

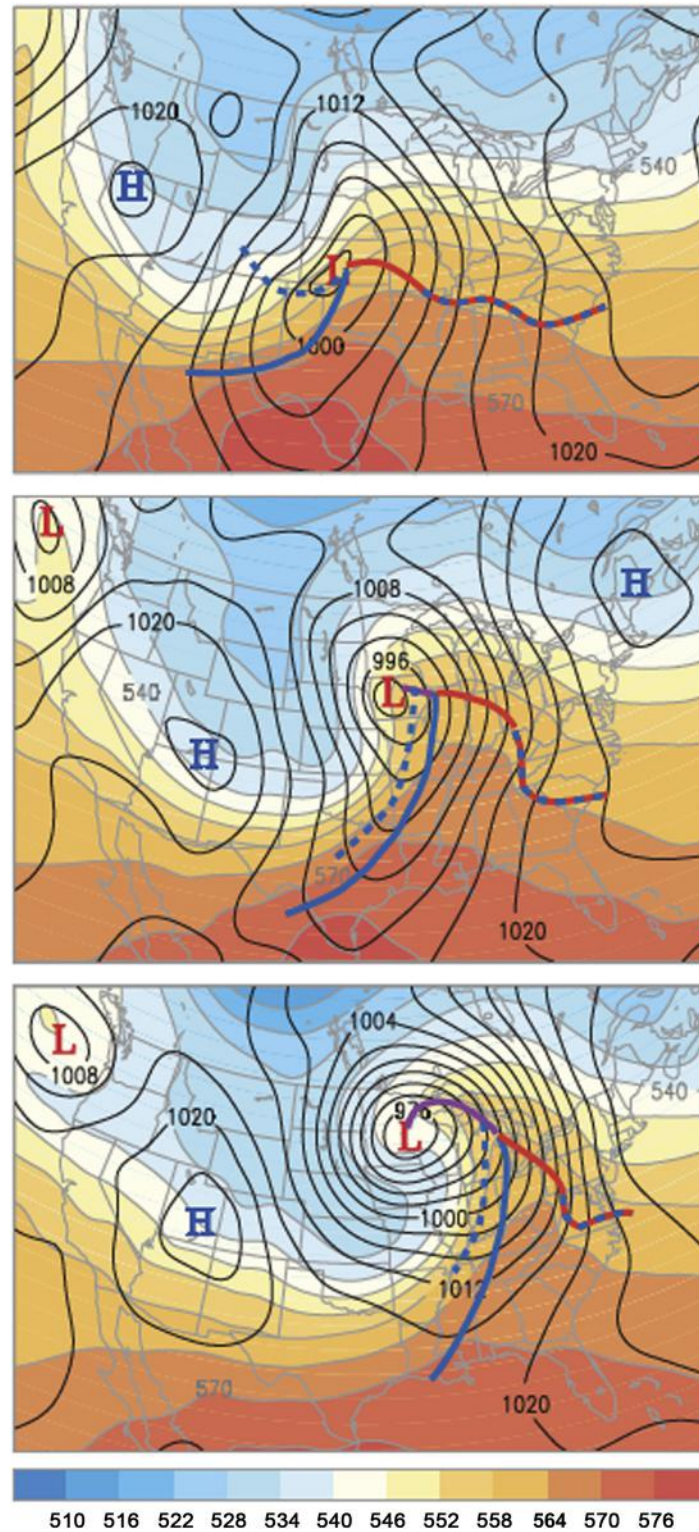


Figure 8 Synoptic charts at 12 h intervals in an amplifying baroclinic wave passing over the Central United States 10 November 1998. Sea-level pressure (contours at 4-hPa intervals) and 1000- to 500-hPa thickness (colored shading: contour interval 60 m; labels in dkm). Surface frontal positions are overlaid. Courtesy of Jennifer Adams, COLA/IGES. From Wallace, J.M., Hobbs, P.V., 2006. *Atmospheric Science: An Introductory Survey*, second ed. Elsevier, 483pp.

equatorward flow of cold air to the west of them contributes to the coldness of Siberia and eastern Canada. The Northern Hemisphere wintertime stationary waves tilt westward with height and they extend upward into the stratosphere. The poleward heat transport by the eddies peaks in midlatitudes and in the lower troposphere, as shown in Figure 9 and in the lower left quadrant of Figure 1.

Water Vapor

The mass of water vapor in the atmosphere is not changing appreciably on timescales comparable to the residence time of individual water vapor molecules in the atmosphere, which is only about a week. Hence, for the globe as a whole, a balance between evaporation and precipitation is required on timescales of a week or longer. However, on a regional basis there can be large imbalances between evaporation and precipitation due to the import or export of atmospheric water vapor by large-scale wind systems. Precipitation exceeds evaporation by a factor of two or more in the equatorial rain belts and in the midlatitude storm tracks and an imbalance in the opposite sense prevails in the subtropical deserts and cloud-free

maritime anticyclones. These features are highlighted in Figure 10, which also indicates the directions of the required atmospheric water vapor transports. The eddies are responsible for most of the poleward transport of water vapor from the subtropical dry zones into the midlatitude storm tracks, while the steadier trade winds that constitute the lower branch of the Hadley cell are responsible for most of the equatorward transport into the tropical rain belts, as depicted in Figure 1.

Regional imbalances between evaporation and precipitation also have implications for the oceanic thermohaline circulation. Excess precipitation freshens the surface waters, decreasing their density and thereby rendering them more stably stratified. The impact of the atmospheric hydrological cycle upon the oceanic circulation is particularly important in the regions of bottom water formation over the North Atlantic Ocean and the Weddell Sea.

The remarkable dryness of the lower stratosphere is a consequence of the Lagrangian mean, so-called *Brewer–Dobson* circulation, in which air parcels enter the stratosphere through the very cold tropical ‘tropopause’ (Figure 2), move poleward and eventually reenter the troposphere at higher latitudes, often in discrete intrusions in the vicinity of the jet stream. The

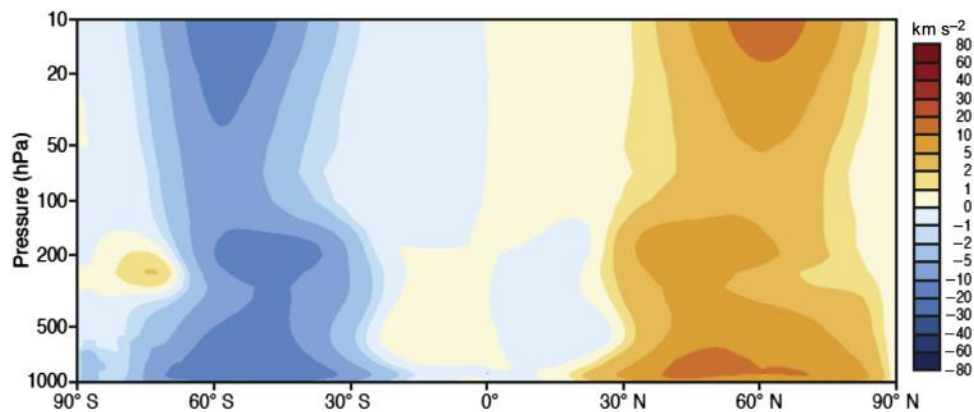


Figure 9 Annual mean northward heat flux due to the horizontal motions in eddies, expressed in units of Km s^{-1} (i.e., per unit mass and per unit length of the latitude circle). To convert temperature flux to sensible heat flux, multiply by specific heat at constant pressure $1004 \text{ J kg}^{-1} \text{ K}^{-1}$. Based on the ERA-Interim Reanalyses. © Elsevier.

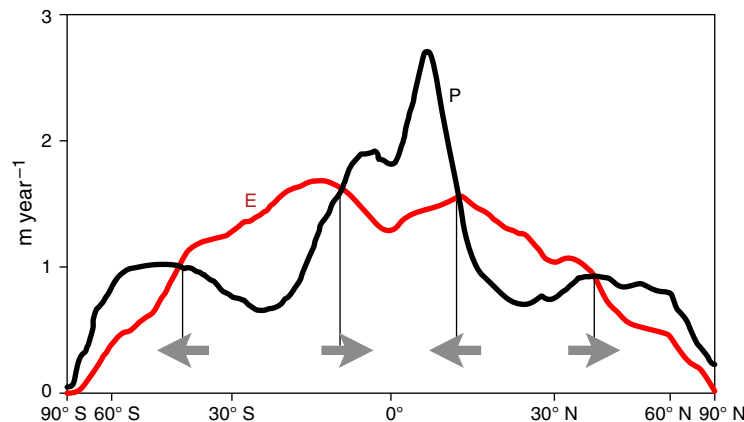


Figure 10 Annual mean precipitation (P) and evaporation (E) as a function of latitude, expressed in units of meters per year. Distance on the latitude scale is proportional to area on the Earth's surface. Based on ERA-Interim Reanalyses. Horizontal arrows indicate the directions of the required atmospheric water vapor transports. © Elsevier.

same circulation carries ozone poleward and downward from its photochemical source region in low latitudes around the 30 km level to its principal reservoir in the 10–20 km layer in high latitudes.

What Determines the Character of the General Circulation?

To understand why the observed planetary-scale wind systems have the particular geographical and seasonal dependence and amplitude that they do, scientists rely upon experience gained from working with numerical models based on the conservation laws that pose the balance requirements considered in the previous section. When such a model atmosphere is ‘turned on,’ starting from a state of rest (i.e., stable stratification, a horizontally uniform temperature distribution, flat pressure surfaces, and no wind) the tropics warm and the polar regions cool in response to the imposed distribution of radiative heating, which is designed to mimic that in the real atmosphere. As the tropical atmosphere warms, the thermal expansion of air causes pressure surfaces in the upper troposphere to

bulge upward relative to the cooling air at the higher latitudes, as depicted in [Figure 11\(a\)](#). The downward sloping of the pressure surfaces from equator to pole gives rise to a poleward flow at the upper levels as depicted in the figure. The poleward mass flux causes mass to accumulate and sea-level pressure to rise at high latitudes, driving a compensating, equatorward low-level flow. Hence, the initial response to the heating gradient is the development of a thermally direct circulation reminiscent of the Hadley cell, but extending all the way from equator to pole, as shown in [Figure 11\(b\)](#). The conservation of angular momentum imparts a westward component to the equatorward flow in the lower branch of the cell and an eastward component to the poleward flow in the upper branch. The vertical wind shear between the low-level westward flow and the upper-level eastward flow increases in proportion to the strengthening north–south temperature gradient, in accordance with thermal wind balance ([Figure 11\(c\)](#)).

For a few weeks of simulated time this equator-to-pole thermally direct circulation persists and the upper-level eastward winds and the meridional temperature gradient continue to strengthen. Up to this point, the circulation is zonally symmetric. But when the temperature gradient reaches a critical

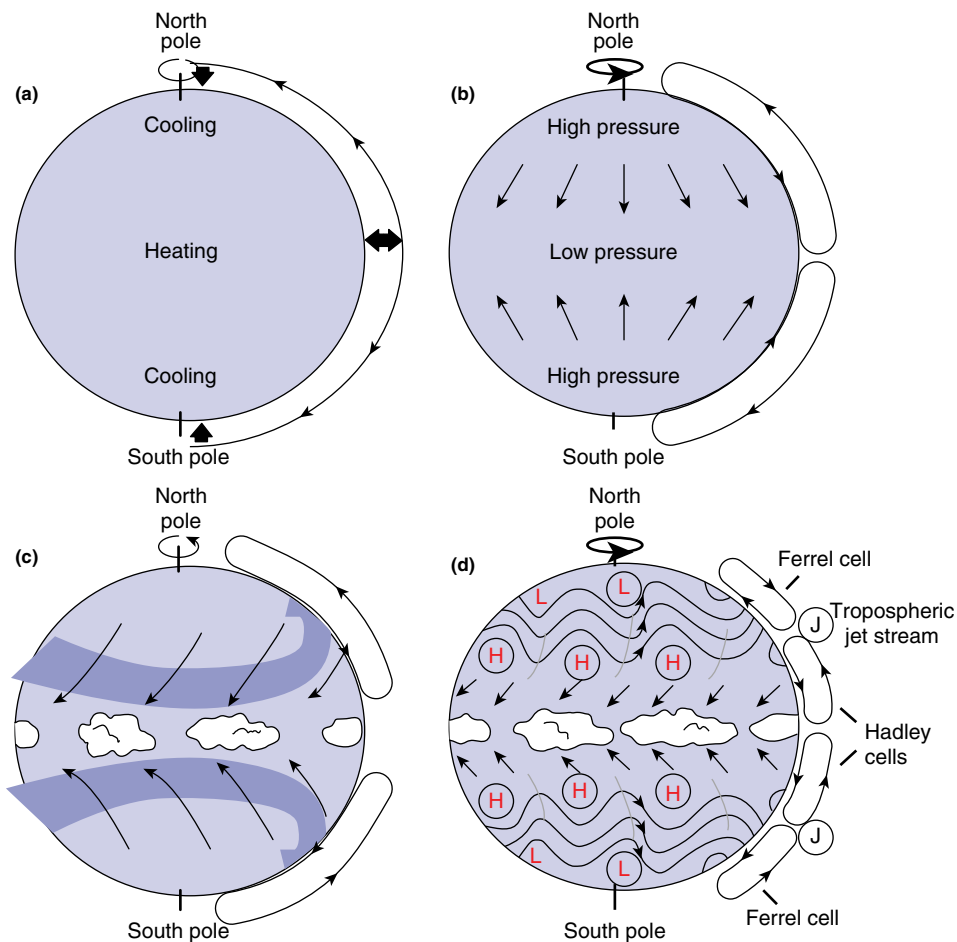


Figure 11 Schematic depiction of the general circulation as it develops from a state of rest on an idealized ocean covered but otherwise ‘Earthlike’ planet under equinox conditions. See text for further explanation. From Wallace, J.M., Hobbs, P.V., 2006. *Atmospheric Science: An Introductory Survey*, second ed. Elsevier, 483pp.

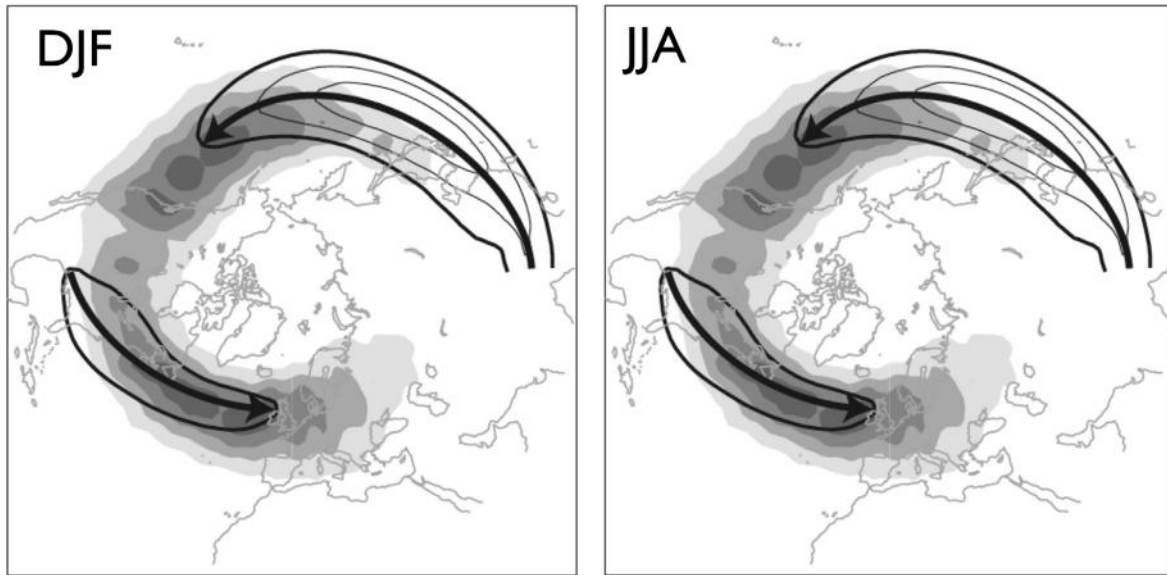


Figure 12 250 hPa zonal-mean zonal wind (contoured) and 850 hPa transient eddy heat flux (shading) for the seasons December through February (DJF) and June through August (JJA) based on the ERA-Interim Reanalyses. © Elsevier.

threshold value for *baroclinic instability*, the simulated circulation undergoes a fundamental reorganization. Successive generations of baroclinic waves spontaneously develop, amplify, and decay in the manner shown in [Figure 8](#). As they do so, they modify the general circulation as indicated in [Figure 11\(d\)](#). As the waves amplify, they produce large poleward eddy heat fluxes that oppose the further buildup of the meridional temperature gradient across midlatitudes. As they disperse upward toward the jet stream (10 km) level and thence equatorward into the tropics, they are attended by a poleward flux of angular momentum from the tropics into midlatitudes. The wave-induced heat and momentum fluxes give rise to the observed thermally indirect midlatitude Ferrel cell ([Figures 1 and 3](#)), which maintains the surface westerlies against frictional dissipation. Were it not for the baroclinic waves, there would be no Ferrel cell, and the broad belt of surface westerlies poleward of 30° would not exist.

Beyond the Zonally Averaged General Circulation

The zonally averaged diagnostics considered in the foregoing sections deal with only the broad outlines of the general circulation: they leave many important zonally varying features unaccounted for, particularly in the Northern Hemisphere where the orography and land/sea thermal contrast are most pronounced. For example, it is evident from [Figure 12](#) that the eastward wind maximum near 30° N in [Figure 2\(a\)](#) is a composite made up of intense wintertime jets over Japan and the eastern United States, as contrasted against more diffuse westerlies over other parts of the hemisphere. In a similar manner, the eddy flux cross sections presented in [Figures 5 and 9](#) tend to be dominated in the Northern Hemisphere by well-defined ‘storm tracks’ over the oceans downstream of the continents. Understanding the zonally varying structure of the general circulation requires consideration of more complex,

three-dimensional balance requirements and numerical simulations that incorporate careful treatment of land–sea thermal contrasts and mountains.

Nor can the statistics that describe the general circulation necessarily be regarded as perfectly reproducible, year after year. For example, they are discernibly different during contrasting years of the El Niño/Southern Oscillation cycle, particularly over the Pacific sector during the months of December through March. They may also occur in response to changes in the distribution of radiative heating brought about by anthropogenic emissions of radiatively active trace gases and aerosols.

Acknowledgments

We would like to thank the Metview and Magics development teams at ECMWF for their support in producing the figures for ERA-Interim Reanalysis.

See also: **Dynamical Meteorology:** Stationary Waves (Orographic and Thermally Forced). **Stratosphere/Troposphere Exchange and Structure:** Global Aspects; Local Processes. **Tropical Meteorology and Climate:** Hadley Circulation. **Turbulence and Mixing:** Overview; Turbulence, Two Dimensional.

Further Reading

- Holton, J., Hakim, G.J., 2012. The general circulation. In: *An Introduction to Dynamic Meteorology*, fifth edition. Academic Press, San Diego, CA. pp. 325–375.
- Vallis, G.K., 2006. Large-scale atmospheric circulation. In: *Atmospheric and Oceanic Fluid Dynamics: Fundamentals and Large-Scale Circulation*. Cambridge University Press, Cambridge, UK. pp. 449–580.
- Wallace, J., Hobbs, P., 2006. Climate dynamics. In: *Atmospheric Science*, second edition. Academic Press, San Diego, CA. pp. 419–465.

Angular Momentum of the Atmosphere

DA Salstein, Atmospheric and Environmental Research, Inc., Lexington, MA, USA

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

Variability in the angular momentum of the atmosphere, related to wind and mass, mirror climate and weather fluctuations. Important signals derive from the strength of the subtropical upper-level jets. Notable variability occurs on diurnal, sub-seasonal, annual, semiannual, interannual, and decadal/interdecadal scales, including from the quasi-biennial oscillation and the El Niño Southern Oscillation. Torques against the surface transfer angular momentum by pressure against mountains or by frictional action, related to changes in Earth's rotation, namely length of the day changes pole motion. Observations and models provide evidence for secular increases in angular momentum in both the recent and upcoming centuries.

Introduction

Angular momentum is a property of mass in motion about a given axis, which in a closed domain is conserved. In the context of the atmosphere, angular momentum is a useful parameter for studying dynamics on different temporal and spatial scales. When the reference axis is identified with that of the Earth's figure, which we may call the principal axis, the resulting globally integrated axial angular momentum value, moreover, may be treated as a fundamental index of atmospheric circulation. As such, this parameter mirrors many aspects of the signature of climate and weather. Furthermore, how angular momentum is exchanged across the atmosphere's lower boundary, by means of the interactive torques with the oceans and solid Earth below, is important to quantify so that one can understand how the Earth acts as a system. Small but measurable changes in the Earth's rotation rate, moreover, are a consequence of the exchanges of angular momentum between the solid Earth and its fluid envelope; this aspect of the variability is of importance to the study of Earth's physics and to the monitoring of reference frames for satellite orbits and navigation. The relevance of atmospheric angular momentum fluctuations to geodesy and geophysics has been recognized by the formal organization of the Special Bureau for the Atmosphere of the International Earth Rotation and Reference Systems Service to supply such atmospheric data to geoscientists.

The angular momentum of a parcel of air in the plane perpendicular to an axis is given as its mass multiplied by the length of the radius arm to the reference axis, multiplied by the component of the velocity of the parcel in that plane, normal to the radius arm. The angular momentum of the global atmosphere about such an axis is the sum of that of all its air parcels, which may be calculated by integration over the volume of the atmosphere. Because the atmosphere is a fluid, variations in its angular momentum relate to changes in both motion terms (relative to the Earth), as well as to changes in its mass distribution.

As a conservative property, angular momentum in a closed system has constant total but can be redistributed within that system. For example, the atmosphere transfers angular momentum poleward in both hemispheres principally by means of transient eddies. Poleward transfers of mass however

mean decreased axial angular momentum because the mass moves to a smaller radius arm from the Earth's axis of rotation. If there are no interactions with the solid Earth, a Coriolis torque will ensue, acting to increase the zonal winds, and keep the angular momentum constant. The transport of angular momentum is also accomplished vertically, carrying angular momentum as part of the Hadley and other mean meridional circulations.

The atmosphere, however, is far from a closed system in this respect, and a stream function/divergence analysis of zonal mean angular momentum reveals sources and sinks at the atmosphere's lower boundary. Indeed, there are exports and imports of angular momentum across the atmosphere's lower interface by means of torques. But the whole Earth, including its fluid components, functions essentially as a closed system with respect to the angular momentum budget (but for the influence of certain well-known tides, principally involving the moon).

If we consider atmospheric angular momentum about the fundamental axis, the relative angular momentum is largely dependent on the westerly component of the wind, with the portion related to mass changes rotating with the Earth very small. Because of the variations in the axial atmospheric angular momentum quantity, the angular momentum in the other components of the Earth's system must change in compensation. Indeed, observations using a number of space-geodetic techniques have demonstrated that the Earth's rotation rate changes perceptibly on many timescales. Such a change is most conveniently expressed in terms of variations in the length of day (l.o.d.), which are very nearly proportional to those in atmospheric angular momentum.

Besides the principal axis, angular momentum may be calculated as well about pairs of other axes in the plane perpendicular to the principal axis, intersecting the equator. In the components in these equatorial axes, the term related to the mass of the atmosphere dominates. These changes of atmospheric angular momentum lead to motions of the Earth's pole about the mean rotation axis – the wobble of the Earth. Such polar motions have also been measured by several space-geodetic techniques.

Changes in the angular momentum of a body must be produced by an imposed torque. In the case of the atmosphere, such interactions occur across its lower interface, with the solid

Earth and the ocean below. These torques are related principally to two mechanisms. In one, surface winds transfer angular momentum by tangential stresses across that surface, yielding a so-called friction torque on the Earth. A second mechanism comes as a result of the existence of surface pressure variability near areas of high topography. Such 'mountain' torques result from the variability of the normal pressure-gradient forces that push harder on one side of the mountain than they do on the other.

In the following, we will examine the distribution of angular momentum in the atmosphere, principally the axial component, and then we discuss its variability on a number of timescales. We also expand upon the relationship to the corresponding motions of the Earth.

Axial Angular Momentum in Regions

The dominant relative angular momentum about Earth's axis depends on the strength of zonal winds, which tend to be persistent features of the atmospheric circulation. In **Figure 1** we present the latitude–pressure distribution of the long-term zonal mean zonal wind based on fields from the reanalysis dataset from the U.S. National Centers for Environmental Prediction and National Center for Atmospheric Research (NCEP–NCAR); we used a 30-year period at the end of the twentieth century to form a climatology of the mean zonal wind. A similar signature exists in the two hemispheres. Mean easterly winds are found in the tropical regions, with a broader latitudinal extent at the surface than higher in the troposphere. The winds are westerly over most of the extratropics. Increases in the strength of these winds with height in the atmosphere lead to very strong westerlies in the upper troposphere, at levels near 200 hPa; above this jet level, they tend to decrease again. Regional maps of such jets at this level reveal that the strongest values are located over the eastern North American and Asian

continents. The strong winds at these regions contribute heavily to the relative component of the axial angular momentum of the atmosphere.

Angular momentum can be computed in zonal belts from values of the zonal wind, so that a profile of angular momentum (**Figure 2**) reveals the general distribution with latitude. The seasonality of the angular momentum can be noted as the substantial difference between the December–January–February periods and the June–July–August periods. It is clear that middle latitude belts have their largest values during their winter, in both the Southern and Northern Hemispheres, but the annual cycle is larger in the northern than in the southern hemispheric belts. Most of the atmosphere has westerly relative angular momentum, indicating that in these regions the atmosphere superrotates with respect to the underlying planet.

Global Atmospheric Angular Momentum

From series of the four-times daily zonal winds given based on the NCEP–NCAR reanalyses, global values of the relative atmospheric angular momentum are calculated by integration over the volume of the atmosphere. In the resulting series of atmospheric angular momentum values, shown in **Figure 3**, a host of interesting signals on a number of important time-scales emerges. For example, it is clear that a strong annual cycle exists, whose phase yields a peak around January, during the period of the strongest Northern Hemisphere jets in the boreal winter. The angular momentum signal in each hemisphere peaks in its respective winter months, but the annual signal in the Northern Hemisphere is of stronger amplitude than that of the Southern Hemisphere, due to the greater continentality of the Northern Hemisphere. As a result, the phase of the global signal is that of the Northern Hemisphere, though the amplitude is reduced. Remarkably, a factor of two exists between the

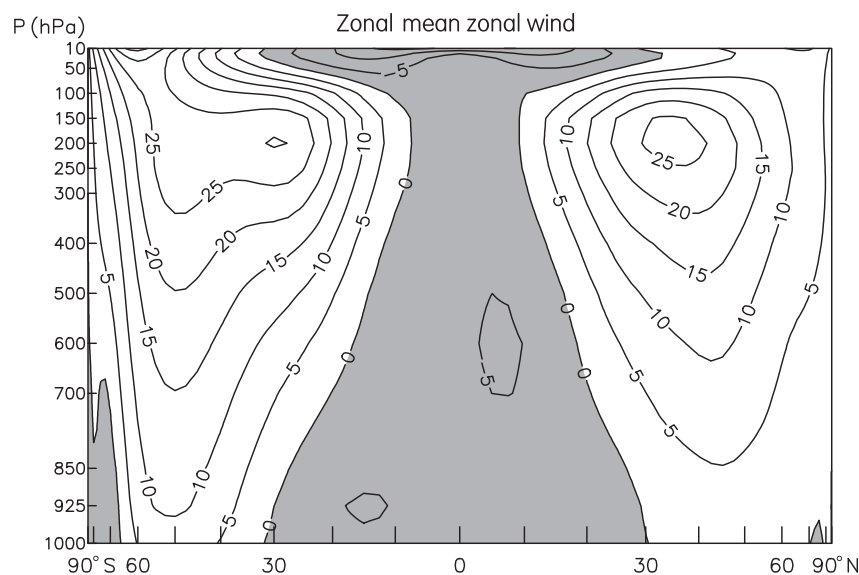


Figure 1 Latitude–pressure cross-section of zonal-mean zonal winds, from which relative atmospheric angular momentum is derived. Based on 30 years of the NCEP–NCAR analysis system. Units are meter per second.

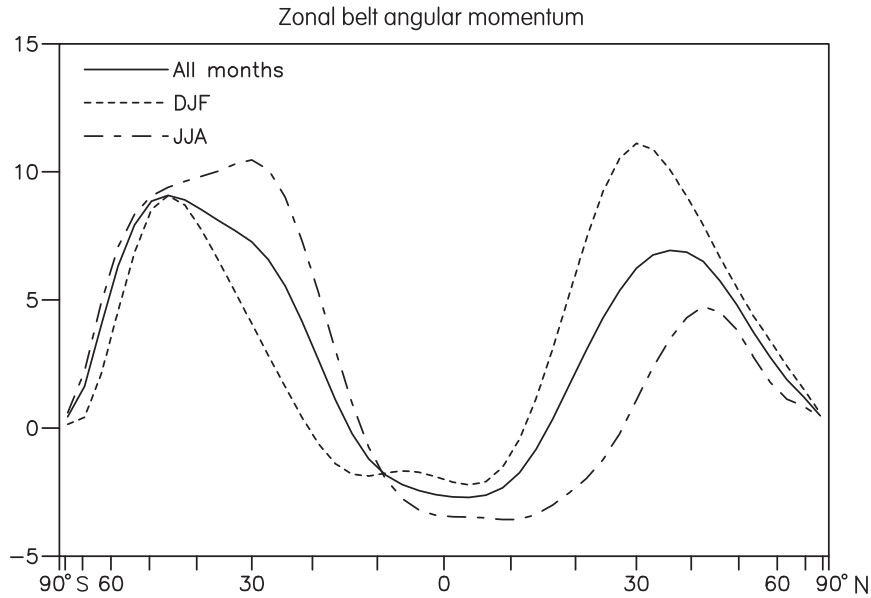


Figure 2 Long-term mean angular momentum distribution in a set of 46 equal area belts spanning a 30-year time period. Shown are values for all months, for December–January–February and June–July–August months. Units are $10^{24} \text{ kg m}^2 \text{ s}^{-1}$.

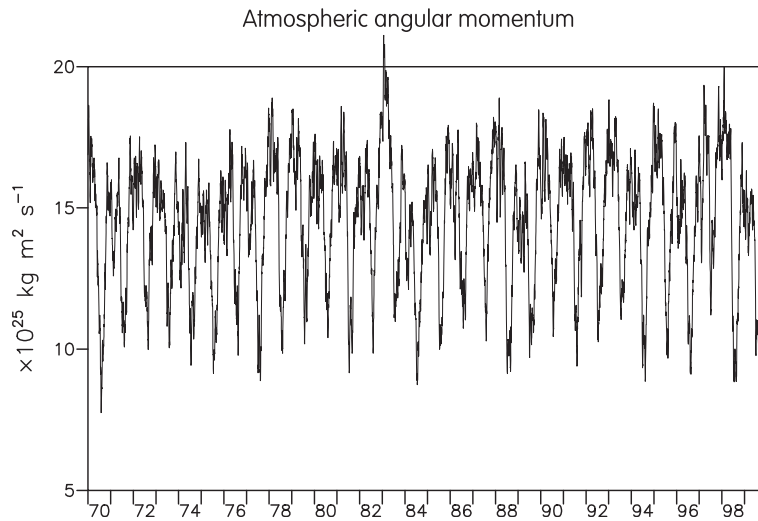


Figure 3 Series over 30 years, of globally integrated atmospheric angular momentum between the 1000 and 10 hPa levels, based on the NCEP–NCAR reanalyses. Units are $10^{25} \text{ kg m}^2 \text{ s}^{-1}$.

relative angular momentum values in Northern Hemisphere winter and summer.

In the figure there is also evidence of a superimposed semiannual signal, which can be noted as a combination of a dip during the middle of the northern winter, and a sharp plunge in the middle of the southern winter. This semiannual signal arises largely from the corresponding wind signal in the stratosphere. Such an overall signature is derived from annual patterns at different latitudes that peak 6 months out of phase. These varying patterns can be noted in the time–latitude diagram of angular momentum in the stratosphere in [Figure 4](#).

On interannual timescales, we find two prominent signals in the global signature in [Figure 3](#), one on scales slightly longer than 2 years, and a second on timescales closer to 4 years.

The shorter of the two relates to the so-called ‘quasi-biennial oscillation,’ a result of the reversal of the zonal winds in the tropical stratosphere. The second of the two has a signature that attains a maximum around the time of peaks in the El Niño Southern Oscillation (ENSO) over the tropical Pacific Ocean, and minima at the time of the La Niña events.

Quasi-biennial Oscillation in Atmospheric Angular Momentum and Stratospheric Winds

The distribution of winds in the stratosphere is such that westerly winds predominate in middle latitudes and easterly winds are found in the tropics. In alternate years, approximately,

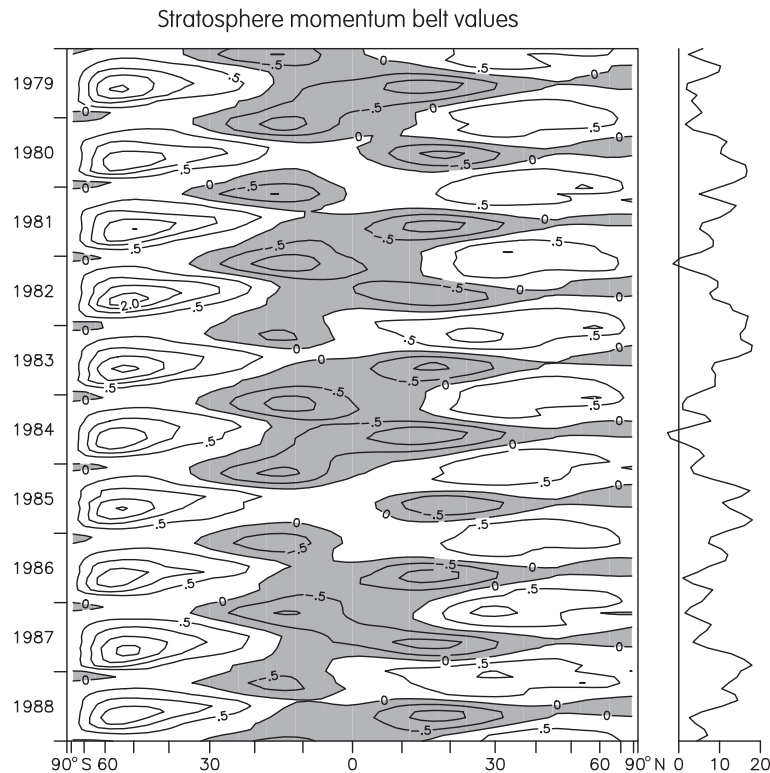


Figure 4 Time–latitude diagram of angular momentum of the stratosphere in 46 equal area belts, between the 100 and 10 hPa pressure levels. The resolution of data here is monthly. Units are $10^{24} \text{ kg m}^2 \text{ s}^{-1}$.

however, the tropical easterly winds tend to diminish substantially or even reverse their direction to become westerly. Such an alternating signal, though noted first over the western Pacific, has been observed at other longitudes, and it is very well captured by a zonal average. It can be observed in [Figure 4](#) in the 100–10 hPa layer in the stratosphere in the belts surrounding the equator. Because angular momentum is calculated with weights related to the distance to the rotation axis, the contributions from the zonal winds at the lowest latitudes, furthest from the rotation axis, are most important here. The vertical curve on the right hand side of [Figure 4](#), the sum over all the belt values in the stratosphere, clearly reflects an alternation every other year in the global time series of angular momentum in the stratosphere.

ENSO Influence on Angular Momentum

The influence of the ENSO produces a clear signature in the evolution of angular momentum. The origin of the strong peaks in the global relative atmospheric angular momentum may be noted in the time–latitude diagram, such as [Figure 5](#), which, to emphasize the ENSO timescales, filters out signals longer than 4 years and is confined to the region below 100 hPa. During periods of El Niño, the tropospheric zonal winds have westerly anomalies, equivalent to weakened easterlies or to westerlies in part of the tropics initially, and then anomalously strong westerlies more poleward on the order of 0.5–1 year later. During the peak of the westerly anomaly

period, especially, the globally integrated atmospheric angular momentum is notably strong. Two such strong values during this period were during the 1982–83 and the 1997–98 El Niño events. Subsequent events after 1998 have also been noted in records of atmospheric angular momentum. During these episodes, the global signal in relative atmospheric angular momentum was exceptionally high. However, the record value in January 1983 came about as a result of the superposition of the El Niño signal with that of the normally strong seasonal signal during northern winter. With the cooling of the waters in the Pacific, the La Niña ushers in a different circulation from that of the El Niño, and anomalously easterly winds create a negative anomaly in atmospheric angular momentum. The transition can be quite abrupt, as occurred during May 1998, a month that featured a reversal of the sign of angular momentum anomaly across a very wide meridional band from the middle latitude southern to the middle latitudes of the northern hemispheres.

Torques across Atmosphere’s Lower Boundary

The angular momentum of the atmosphere may fluctuate quite rapidly, and so it is apparent that means must exist to accomplish this change at the atmosphere’s lower boundary. Two principal torque mechanisms to effect the angular momentum transfer have been identified. In one, the atmosphere sets up a pressure-gradient force on opposite sides of mountainous topography, and when considered at a distance from the axis,

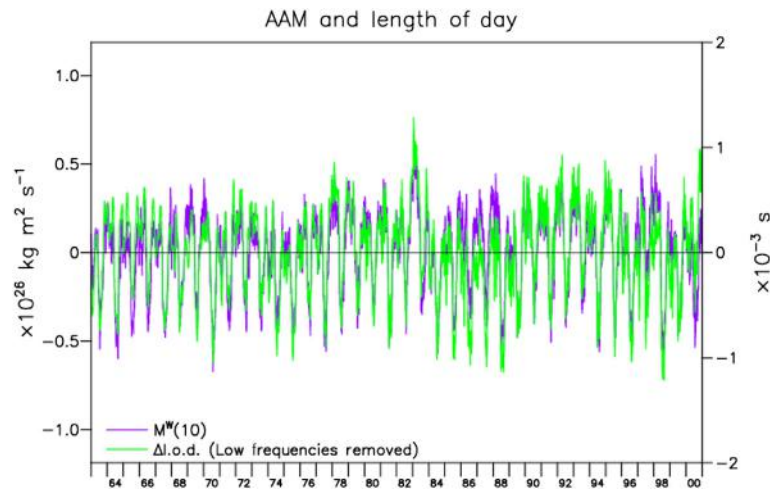


Figure 6 A comparison of atmospheric angular momentum (AAM) and length of day series, given on equivalent scales. Frequencies with timescales of longer than 4 years have been filtered out. AAM is in units of $10^{26} \text{ kg m}^2 \text{ s}^{-1}$, and length of day in 10^{-3} s .

year 2000. Here, however, data with timescales greater than around 4 years are filtered out, because the Earth's rotation signals at these scales are driven by interactions with the core of the Earth. The very good agreement between the two series is remarkable because of the extremely different data types from which they are derived. The seasonal and intraseasonal fluctuations occur quite closely in both series. Indeed, coherence between the two is very strong on scales down to 7 days, and has some significance on scales as short as daily. Interestingly, intraseasonal fluctuations in l.o.d. were discovered independently from those of the atmospheric Madden–Julian oscillation, related to fluctuations across the Pacific Ocean and observed in both the tropics and extratropics. Differences in angular momentum between the Earth and atmosphere point to either errors in the data sets or to the role in the exchange of angular momentum of a third component such as the oceans.

Models and Historical Series of Atmospheric Angular Momentum

The atmosphere has been simulated by a large number of models that are driven solely by the temperature of the underlying ocean. Based on these models, atmospheric angular momentum has been calculated and used, moreover, as a parameter for model validation to determine the success of model simulations. Aside from observations of angular momentum, independent measurements of l.o.d. have been used to examine the results of models. Lengthy runs of models are possible because sea surface temperatures are available since the latter part of the nineteenth century; these models are unlike atmospheric analyses, whose dependence on upper air winds, are confined to the period since the middle of the twentieth century. Such runs indicate that an increase in such values since that time appear to have occurred, possibly related to the relative increase in El Niño activity. Increases in the short-term variability of atmospheric angular momentum moreover appear to have taken place.

Models can be run in a prognostic mode as well, to determine, for example, the effect of an increase in greenhouse gases on the angular momentum of the Earth. Such effects may include changes in the annual signature, and a possible decrease in angular momentum could be related to the warming of the higher latitudes, which could induce a reduction in temperature gradient and the strength of zonal winds. Use of a coupled atmosphere–ocean models help resolve changes that would be needed for the prediction of angular momentum trends.

Also, because of the close relationship on timescales from days to a few years, between l.o.d. and atmospheric angular momentum, earlier records of l.o.d. have been used as a proxy for the global variations in the atmosphere. A record of l.o.d. back to the dawn of the telescope era in the seventeenth century has been examined, though it is of insufficient accuracy for atmospheric purposes until the end of the nineteenth or the beginning of the twentieth century. Signals relating to changes in variability of the atmosphere during certain decades (like the 1920s and the 1940s, which had high and low interannual variability, respectively, in l.o.d.), and dominant interannual timescales (3.4 and 2.1 years) have been determined from such a proxy record.

Projections into the Future of Atmospheric Angular Momentum

Coupled models of the atmosphere–ocean are used currently to simulate the present state of the atmosphere and to extend the expected state into the future. The Coupled Model Intercomparison Project has organized the running of a number of models throughout the world in different scenarios of the future involving the build-up of greenhouse gases. The results have been used by the Intergovernmental Panel on Climate Change for studying possible effects of greenhouse warming. Possible scenarios include 'business as usual,' extending the rapid growth of CO_2 , as well as more modified scenarios

indicating versions of less rapid growth of greenhouse gases. In general the expectations from these studies are that angular momentum of the atmosphere may increase by some 10% due to changes in the vigor of the zonal winds, particularly in the upper troposphere of both hemispheres in the middle and high latitudes. Three such estimates using the Goddard Institute for Space Studies (GISS) coupled model from the Third Coupled Model Intercomparison Project (CMIP3) archives, out to the year 2100 are given by the increasingly higher curves in **Figure 7** for the increasing amounts of CO_2 that may be put according to these scenarios. Note that results also show the amount of relative angular momentum already increasing from the late 1800s to the present, according to the models and the observations used to construct the angular momentum numbers. Given that the angular momentum of the solid Earth generally decreases with the increasing atmospheric portion, a resulting change in the l.o.d. can extend it by an amount on the order of 0.2 ms.

Atmospheric Angular Momentum on Very Short Timescales

On the other side of the time spectrum from the climate timescales that we just considered here are very rapid changes in the atmospheric angular momentum on the order of several days or less. In fact recent capabilities of models and atmospheric data assimilation systems allow for the calculation of hourly signals. Interestingly, at these timescales, power in the amplitudes of atmospheric angular momentum fluctuations fall quite short of those of the Earth's angular momentum, that is, the Earth's rotation signal, so at these timescales other factors may be in play in the angular momentum balance. Such responses are related to the tides of the atmosphere. Particular peaks in the spectrum of atmospheric angular momentum, shown in **Figure 8**, occur at the diurnal time scale and its next three harmonics (strong semidiurnal, peaks at 8 and 6 h).

Atmospheric Angular Momentum in the Equatorial Plane and Polar Motion

Besides its rotation about the principal axis discussed for most of this article, the other two components of the atmospheric angular momentum vector, namely those in the equatorial plane, can be determined. Though not of clear fundamental

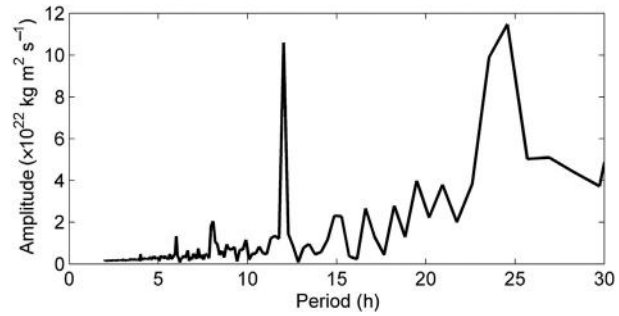


Figure 8 High-frequency spectrum of atmospheric angular momentum based on European Center for Medium Range Weather Forecasts model and analyses during a special observing period in August 2008. The period is given in hours. Note the strong broad annual diurnal period, and also the sharp semidiurnal as well as other peaks in the spectrum. Courtesy of Schindelegger, M., Boehm, J., Salstein, D., Schuh, H., 2011. High-resolution atmospheric angular momentum functions related to Earth rotation parameters during CONT08. *J. Geodesy* 86 (7), 425–433. <http://dx.doi.org/10.1007/s00190-011-0458-y>.

interest to atmospheric studies, this component of angular momentum is related importantly to certain motions of the Earth known as Earth wobble, or polar motion. Related fluctuations of angular momentum in these components are stronger in the so-called matter (surface pressure) term than in the motion (wind) term. Thus, pressure variability over certain regions like the northern Pacific and Atlantic (Aleutian and Icelandic lows, respectively), the southern oceans, and over Eurasia have been determined to be important to fluctuations of equatorial angular momentum (**Figure 9**).

When atmospheric pressure fluctuations over the oceans are observed and carefully accounted for, it has been noted that those on timescales of several days and longer influence the distribution of the ocean mass below. This effect, the so-called inverted barometer, acts so that a high atmospheric pressure will depress the surface below, moving ocean mass away from that region; the opposite action occurs with a relatively low atmospheric pressure. Such an inverted barometer relationship has the affect of dramatically reducing the mass component of the effective angular momentum signal of the atmosphere over the oceans.

For the continental regions remaining, the mass fluctuations over Eurasia, predominantly, and North America, secondarily, appear to be the biggest regional atmospheric influences exciting polar motions on subseasonal and other timescales

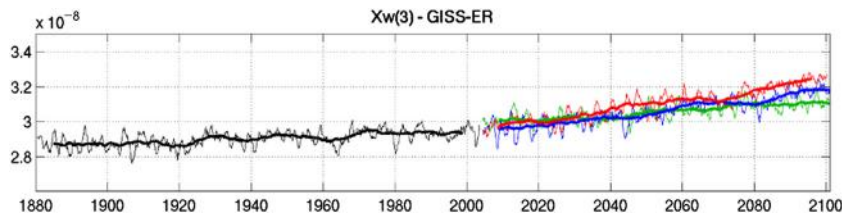


Figure 7 Excitations of length of day, proportional to atmospheric angular momentum from the NASA Goddard Institute for Space Studies (GISS) model contributed to the Third Coupled Model Intercomparison Project (CMIP3). Values prior to the year 2000 were based on observations and the underlying model. Those since 2000 are based on three different scenarios of greenhouse gas emissions with the green, blue, and red curves in increasing order of emissions due to economic growth and energy policy.

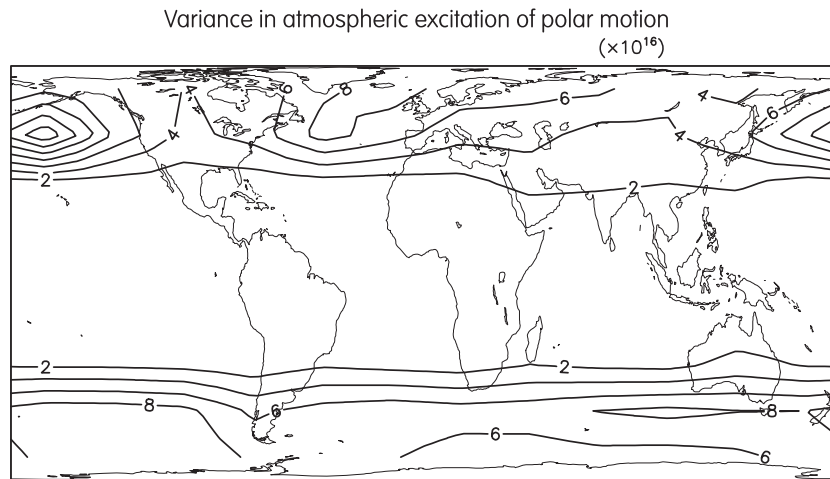


Figure 9 Variability of (nondimensional) excitation of polar motion, based upon surface pressure analyses for a 30-year period.

which have been monitored using space-geodetic methods. On longer timescales, the ocean plays a role of similar importance to that of the atmosphere. Such more dominant ones for polar motion are the annual term and that of the natural modal response of the Earth's polar motion, with period near 430 days, known as the Chandler wobble. At the other end of the frequency spectrum, a motion of the Earth on near daily timescales, known as nutation, is also driven partially by atmospheric angular momentum forcing.

See also: **Air Sea Interactions:** Momentum, Heat and Vapor Fluxes. **Climate and Climate Change:** Climate Variability: Seasonal and Interannual Variability. **Land-Atmosphere Interactions:** Overview. **Middle Atmosphere:** Quasi-Biennial Oscillation; Semiannual Oscillation. **Mountain Meteorology:** Overview. **Tropical Meteorology and Climate:** El Niño and the Southern Oscillation: Observation.

Further Reading

Barnes, R.T.H., Hide, R., White, A.A., Wilson, C.A., 1983. Atmospheric angular momentum fluctuations, length-of-day changes and polar motion. *Proceedings of the Royal Society of London A* 387, 31–73.

Grotjahn, R., 1993. *Global Atmospheric Circulations, Observations and Theories*. Oxford University Press, New York.

Hide, R., Dickey, J.O., 1991. Earth's variable rotation. *Science* 253, 629–637.

Lambeck, K., 1980. *The Earth's Variable Rotation: Geophysical Causes and Consequences*. Cambridge University Press, Cambridge.

Lorenz, E.N., 1967. *The Nature and Theory of the General Circulation of the Atmosphere*. World Meteorological Organization, Geneva.

Lott, F., de Viron, O., Viterbo, P., Vial, F., 2008. Axial atmospheric angular momentum budget at diurnal and subdiurnal periodicities. *Journal of Atmospheric Science* 65, 156–171.

Nastula, J., Salstein, D.A., 1999. Regional atmospheric momentum contributions to polar motion. *Journal of Geographical Research* 104, 7347–7358.

Newton, C.W., 1971. Global angular momentum balance: Earth torques and atmospheric fluxes. *Journal of Atmospheric Science* 28, 1329–1341.

Oort, A.H., 1989. Earth's angular momentum cycle in the atmosphere–earth–solid earth system. *Bulletin of the American Meteorological Society* 70, 1231–1242.

Peixoto, J.P., Oort, A.H., 1992. *Physics of Climate*. American Institute of Physics, New York (Chapter 11).

Rosen, R.D., 1993. The axial momentum balance of Earth and its fluid envelope. *Survey Geophysics* 14, 1–29.

Salstein, D.A., Kann, D.M., Miller, A.J., Rosen, R.D., 1993. The sub-bureau for atmospheric angular momentum of the International Earth Rotation Service: a meteorological data center with geodetic applications. *Bulletin of the American Meteorological Society* 74, 67–81.

Salstein, D.A., Kolaczek, B., Gambis, D. (Eds.), 1999. *The Impact of El Niño and Other Low-Frequency Signals on Earth Rotation and Global Earth System Parameters*, IERS Technical Note 26. Central Bureau of the IERS.

Schindelegger, M., Boehm, J., Salstein, D., Schuh, H., 2011. High-resolution atmospheric angular momentum functions related to Earth rotation parameters during CONT08. *Journal of Geodesy* 86 (7), 425–433. <http://dx.doi.org/10.1007/s00190-011-0458-y>.

Weickmann, K.M., Kiladis, G.N., Sardeshmukh, P.D., 1997. The dynamics of intraseasonal atmospheric angular momentum oscillations. *Journal of Atmospheric Science* 54, 1445–1461.

Energy Cycle

R Grotjahn University of California, Davis, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

The energy cycle provides a physically meaningful system to understand several constraints upon and properties of the general circulation. Energy is conserved and it can be tracked even as it changes from one form to another. Energy properties can be analyzed to deduce the net circulations as well as the rates at which the circulations are created, maintained, or destroyed.

Introduction

The total energy per unit mass (TE) is defined as:

$$\underbrace{C_p T + gZ + Lq}_{\text{MSE}} + \underbrace{\frac{1}{2}(u^2 + v^2 + w^2)}_{\text{KE}} = \text{TE} \quad [1]$$

where C_p is the specific heat at constant pressure, T the temperature, g the acceleration of gravity, Z the geopotential height, L the latent heat of vaporization or sublimation, q the specific humidity, and (u, v, w) the eastward, northward, and upward wind components, respectively. The terms above are internal energy at constant pressure, gravitational energy, latent energy from phase changes of water, and kinetic energy (KE) per unit mass, respectively. Together the first two terms define dry static energy (DSE) while including the third defines moist static energy (MSE).

The distributions of MSE and DSE in summer and winter seasons are shown in [Figure 1](#). The DSE monotonically increases with height because the large-scale atmosphere is statically stable. DSE decreases toward the poles due to cooler temperatures. The addition of moisture, with most moisture mass in the lower troposphere, causes the vertical derivative of MSE to reverse sign (implying convective potential instability) in the lower tropical troposphere and causes the contours to be nearly vertical in the midlatitude troposphere (the latter indicative of a nearly moist adiabatic lapse rate). In the horizontal, higher values of MSE favor the tropical and summer continents as well as the intertropical and other convergence zones (indicative of low-level moisture convergence). In middle latitudes, lower values of MSE occur on the colder east sides of the continents.

Potential energy (PE) is related to DSE. PE is useful for global energy balance. A tiny fraction of the PE, called the available potential energy (APE), is usable to drive the KE. APE is defined as the difference between the PE and the minimum PE that could be achieved by an adiabatic rearrangement of mass. APE is used to understand the links between PE and KE. Sometimes latent heating is included directly in the APE, usually it is treated as a separate diabatic process.

Solar radiant energy does not reach the Earth equally everywhere. On an annual average, the tropics receive and

absorb far more solar energy than the polar regions. This distribution of absorbed energy creates an uneven distribution of temperature. Temperature, pressure, and density are related, so the PE has an uneven distribution too. The existence of APE is essentially due to the horizontal variations in density and temperature. A portion of the APE is converted into motions (KE) as the atmosphere tries to remove these density and temperature variations. The motions redistribute some mass, but mainly the atmosphere transports heat. The atmospheric circulation becomes a complex balance between radiant energy input and output that create APE, which is needed to generate the KE of circulations that in turn strive to create a state of no APE.

APE and KE are defined in formal mathematical ways. The mathematics show interactions from which physical mechanisms (like baroclinic instability) can be identified. The energy equations describe the following chain of events: radiation creates APE; some APE is converted into motions that redistribute the heat energy; and KE in turn is lost by conversion back to APE and by friction. The forms of energy and the *net* conversions between them can be represented via a 'box' diagram. However, the box diagram does not show the energy cycle in an intuitive sense. To make the physical mechanisms clear, energy must be examined regionally and one phenomenon at a time.

Conceptual Models

Two-Fluid Model

A fluid flow analog of the pendulum illustrates forms and conversions of energy. Imagine a tank holding two immiscible fluids of different densities, separated by a vertical barrier ([Figure 2\(a\)](#)). The initial state has the highest center of mass and thus greatest gravitational PE. If the barrier is suddenly removed, the fluids begin to move. The motion accelerates until the point in time where the greatest amount of the denser fluid underlies the greatest amount of the less dense fluid ([Figure 2\(b\)](#)). The center of mass is now lowest as is the gravitational PE. Ignoring friction, mixing, and turbulent effects, KE is maximized at this point. As time proceeds further, the fluids overshoot this state and KE begins converting back to PE ([Figure 2\(c\)](#)).

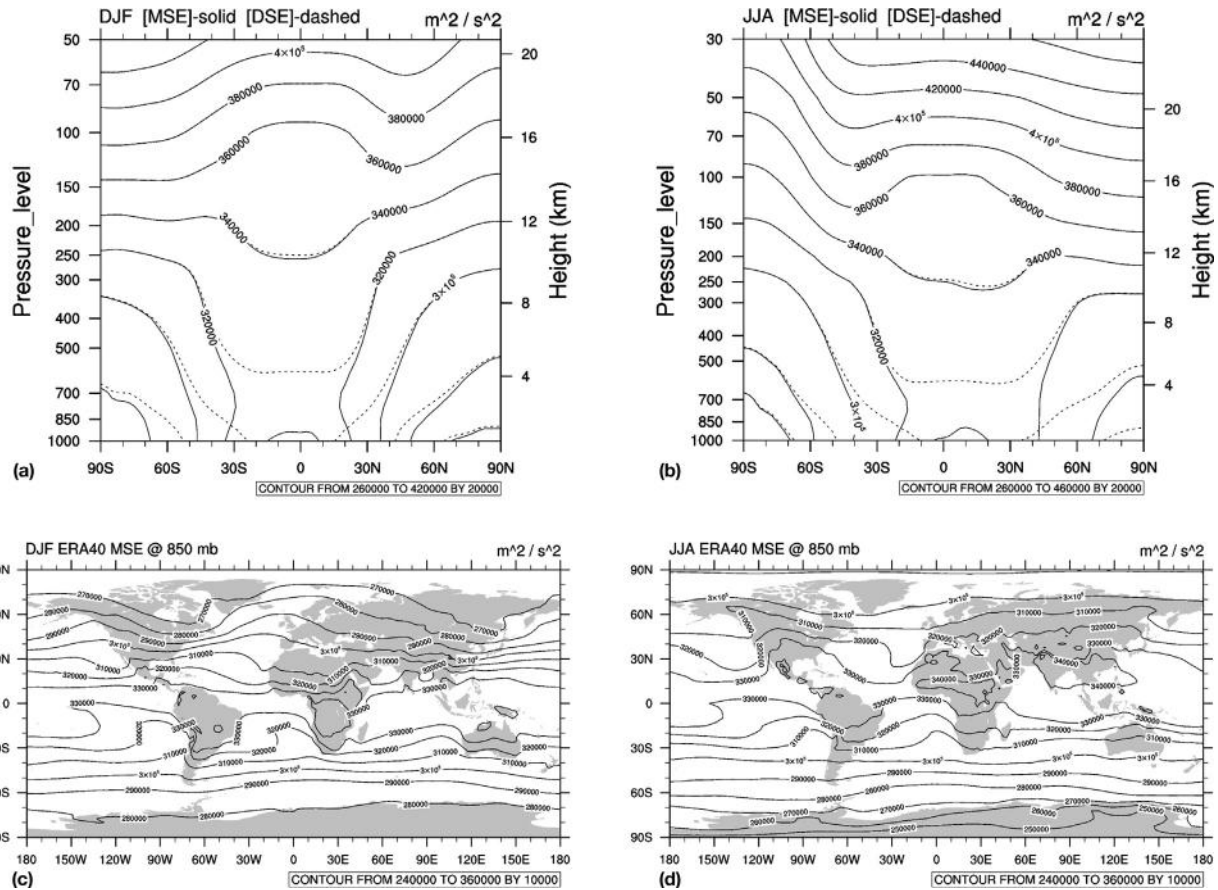


Figure 1 Distributions of DSE and MSE during December–February (left column) and June–August (right column) extreme seasons. In (a) and (b), solid lines are MSE while dashed lines are DSE, which differ mainly in the lower and tropical troposphere. In the deep tropics, the vertical gradient of MSE reverses sign from lower to upper troposphere. In (c) and (d), the longitudinal distribution of MSE is illustrated using the pattern at 850 hPa. ERA-40 reanalysis data from 1979 to 2001 used were provided by the European Centre for Medium Range Weather Forecasts.

APE is defined as the difference between the current PE and the minimum PE. The state with lowest PE is the ‘reference state’ having zero APE. The reference state definition is somewhat arbitrary. Another mechanism could possibly occur at some later time to lower further the minimum PE, for example, a net temperature decrease. However, the size of the conversions, generation, and destruction are less arbitrary. APE is intended to represent PE available for driving motions, so the reference state is usually defined by adiabatically rearranging

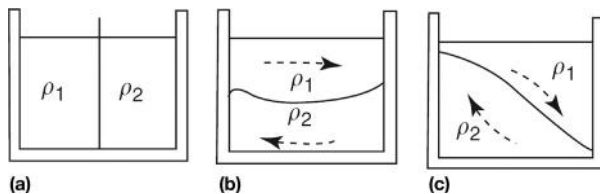


Figure 2 Schematic model of fluid motion illustrating APE and KE concepts. The tank holds two immiscible fluids with density $\rho_1 < \rho_2$. (a) Initial state; (b) state with maximum KE but minimum PE reached during the first oscillation; and (c) state where KE is being converted back to PE.

atmospheric properties so as to reach a state of minimum PE. This model reveals that density differences across a fixed elevation in the tank are proportional to the APE.

This model relates to the atmosphere as follows:

1. Temperature differences create the density differences. The less dense fluid represents the tropics; the denser fluid represents polar regions.
2. The reference state has minimum center of mass when the air ‘layers’ are flat. However, flat fields of pressure and temperature imply no geostrophic winds and remove a driver for ageostrophic winds.
3. Horizontal density differences (manifest as sloping fluid layers) have APE but also produce horizontal pressure gradients that accelerate the air. On the rotating Earth, geostrophic winds and thus KE are present, too. So, reservoirs, sources, and sinks of APE are not independent of KE.

Carnot Cycle

A Carnot cycle analysis can estimate KE generation from thermodynamic changes that an air parcel undergoes

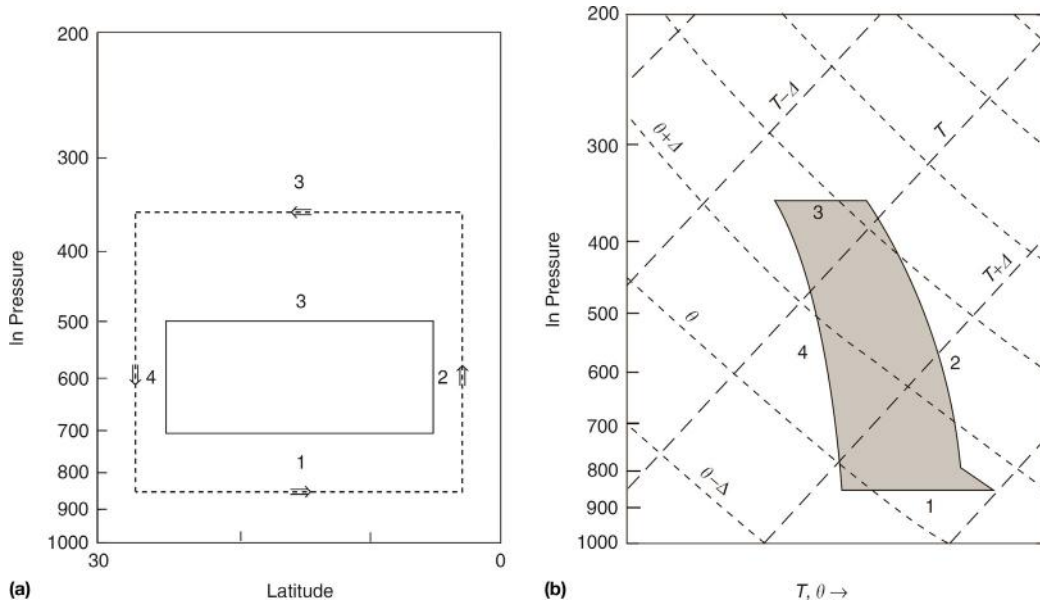


Figure 3 Interpretation of the Hadley circulation as a Carnot cycle. (a) Meridional cross section showing the idealized circulation. The dashed line shows an average path followed by the parcels with numeric labels for each leg. (b) Skew T-lnP plot of the thermodynamic changes along each of the four legs drawn in part (a). The shaded area is proportional to the energy converted from PE to KE.

while completing an atmospheric circuit. The Hadley cell is a conceptual model for the zonal mean tropical circulation. Air in the lower troposphere moves equatorward while gaining heat and moisture from surface fluxes. Near the equator, rapid ascent within thunderstorms releases and advects much latent heat energy. Reaching the upper troposphere, air moves poleward, cools radiatively, and sinks, completing a circuit.

KE generation can be estimated by plotting the thermodynamic properties of air parcels on a skew T-lnP chart. A unit area anywhere on the chart corresponds to a specific amount of energy exchange. Figure 3 shows a possible circuit around an annual mean Hadley cell. The amount of APE converted into KE by a kilogram of air while it completes the plotted circuit is $E \sim 1.4 \times 10^3 \text{ J kg}^{-1}$.

The rate of energy release per unit horizontal area, R , by all the air in motion can be compared with the rate per unit area of energy absorbed from the Sun.

$$R = ME t^{-1} A^{-1} \quad [2]$$

where M is the mass in motion, t is the time to complete the circuit, and A is the area of the Hadley cell. For the schematic circulation indicated in Figure 3(a), $M \sim 10^{18} \text{ kg}$, $A \sim 10^{14} \text{ m}^2$, and $t \sim 3 \times 10^6 \text{ s}$. The total rate of energy released by the Hadley cell is $ME t^{-1} \sim 5 \times 10^{14} \text{ J s}^{-1}$. However, the rate per unit area is only $R \sim 3.6 \text{ W m}^{-2}$. The absorbed solar radiation in the tropics is 100 times larger than R , making the atmosphere an 'inefficient' heat engine. (Efficiency of the Carnot cycle is often measured in a way dependent on the temperature difference during the cycle, but this estimate is related to energy input.)

The model has the following properties:

1. Warmer air is rising and cooler air is sinking, so the center of mass is lowered and KE is created; the circuit is counterclockwise and the circulation is 'thermally direct'. In contrast, the Ferrel cell is a clockwise circuit that reduces KE to increase PE.

2. A steady state is reached if the frictional losses balance the KE generation.
3. The rate of KE generation depends on the area enclosed by the circuit divided by the time to complete the circuit. That time depends on the speed of parcels, a point reinforced later when the APE to KE conversion term is considered.
4. The amount of energy converted is proportional to a circuit integral of temperature, so it increases as the temperature difference increases between the warm and cold stages of the cycle. During winter the meridional temperature gradient is stronger than summer, and so is the Hadley cell.
5. In winter, the air motion of the Hadley cell is five to seven times stronger than the Ferrel cell, but larger temperature differences occur along the Ferrel circuit. So, the net energy conversions are similar (see Energy Box Diagram section below).
6. Large energies are involved, but only a tiny fraction of the solar radiation actually drives the observed motions.
7. The path followed by air parcels was specified, not predicted.

Available Potential Energy

DSE and therefore PE combine both gravitational and internal energies. To the extent that hydrostatic balance and ideal gas law are valid and working at constant volume, then:

$$\begin{aligned} \text{PE} &= \int_0^\infty \rho g Z dz + \int_0^\infty \rho C_V T dz = \int_0^{P_s} Z dP + \int_0^\infty \rho C_V T dz \\ &= \int_0^\infty \rho (R + C_V) T dz = \int_0^\infty \rho C_P T dz \end{aligned} \quad [3]$$

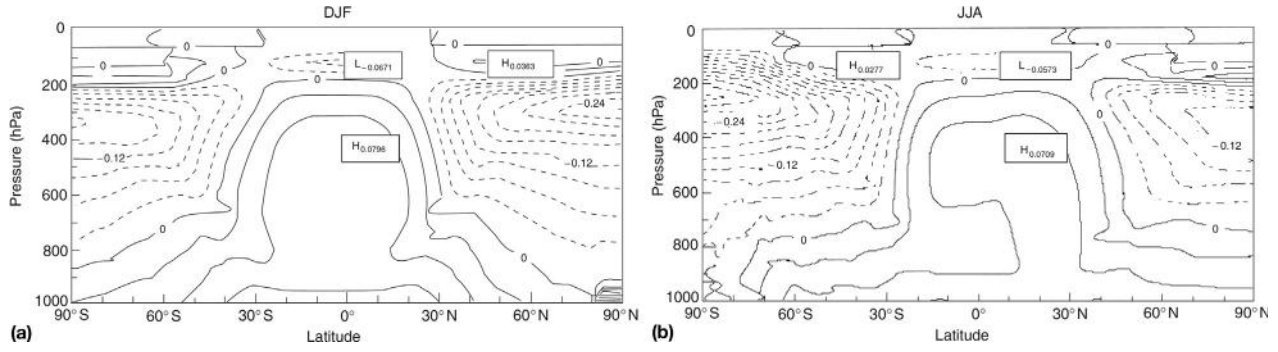


Figure 4 Zonal mean efficiency factor $[\varepsilon]$ for (a) December–February and (b) June–August. $[\varepsilon]$ is estimated from zonal mean 1979–99 National Center for Environmental Prediction/National Center for Atmospheric Research (NCEP/NCAR) reanalysis data. The contour interval is 0.03.

Hence, gravitational and internal energies are bound together in an R/C_p ratio. APE is approximated by temperature variations on a pressure surface.

$$\text{APE} = \int \varepsilon C_p T dM \approx \frac{1}{2} \kappa C_p P_{00}^{-\kappa} \times \int P^{\kappa-1} \{\theta - \bar{\theta}\}^2 \left(\frac{\partial \theta}{\partial P} \right)^{-1} dM \quad [4]$$

$$\varepsilon = 1 - \left(\frac{P_r}{P} \right)^\kappa \quad [5]$$

M is the mass while other variables are potential temperature θ , $P_{00} = 10^5$ Pa, specific heat at constant pressure C_p , $\kappa = RC_p^{-1}$, ideal gas constant R , and ‘efficiency factor’ ε . $P_r(\theta)$ is the reference pressure: the average pressure on a potential temperature surface θ . APE is zero when $P = P_r$ everywhere in the domain. The overbar denotes the horizontal average on an isobaric surface. APE has the following properties:

1. For an integral over the depth of the atmosphere, APE differs from PE by the factor ε .
2. Observed PE is about a thousand times greater than estimates of global average APE.
3. Hemispheric PE is greater in summer since the air is generally warmer than in winter.
4. Hemispheric APE is greater in winter when the meridional temperature gradient is stronger making the $\{\}$ term larger than in summer. The larger the atmosphere departs from the reference state mean, the larger the magnitude of ε becomes.
5. Diabatic heating or cooling can create APE if it magnifies the departures from the reference state, but the same heating or cooling can destroy APE if it reduces the departures. In simplistic terms, APE is generated by ‘heating where it is hot or cooling where it is cold’.
6. $\varepsilon > 0$ in ‘hot’ regions and $\varepsilon < 0$ in ‘cold’ regions. From **Figure 4**, ε has a positive maximum in the tropical middle troposphere and negative minimums in high latitudes. In middle latitudes, the sign varies with longitude: $\varepsilon > 0$ over oceans during winter or warm sectors of frontal cyclones while $\varepsilon < 0$ over continents in winter or behind cold fronts.

Kinetic Energy

KE is primarily contained in horizontal winds:

$$\text{KE} = \frac{1}{2} (u^2 + v^2) dM \quad [6]$$

KE has the following properties:

1. The distribution of zonal mean KE (**Figure 5**) has maximums at upper levels near the subtropical jets.
2. KE is related to atmospheric momentum and torque. Momentum fluxes by the Hadley cells and by midlatitude eddies maintain the KE maximum near the subtropical jet. Slowing down surface easterlies in the Hadley circulation imparts westerly momentum that is transported to higher latitudes by the upper branch of the Hadley Cell. Eddy momentum flux convergence is another source.

Energy Generation and Conversion

Energy Equations

To understand how energy evolves one needs formulae for APE and KE tendencies in a limited domain. The domain may be a unit area in the meridional plane (useful for calculating zonal means) or enclosing a single phenomenon to the exclusion of others (e.g., a single frontal cyclone). Tendency equations for APE and KE in a mass M between two isobaric surfaces are

$$\begin{aligned} \frac{\partial \text{APE}}{\partial t} = & \underbrace{\int (\varepsilon q) dM}_a + \underbrace{\int (\varepsilon \omega \alpha) dM}_b - \underbrace{\int \varepsilon (\nabla_P \cdot (V_P C_p T)) dM}_c \\ & + \underbrace{\int \left(C_p T \frac{\partial \varepsilon}{\partial t} \right) dM}_d \end{aligned} \quad [7]$$

$$\frac{\partial \text{KE}}{\partial t} = - \underbrace{\int (V_P \cdot F) dM}_a - \underbrace{\int (V_P \cdot \nabla_P \Phi) dM}_b - \underbrace{\int (\nabla_P \cdot (V_P \text{KE})) dM}_c \quad [8]$$

where q contains all diabatic heating and F is the frictional force. Also, ω is the pressure coordinate (‘vertical’) velocity, α is

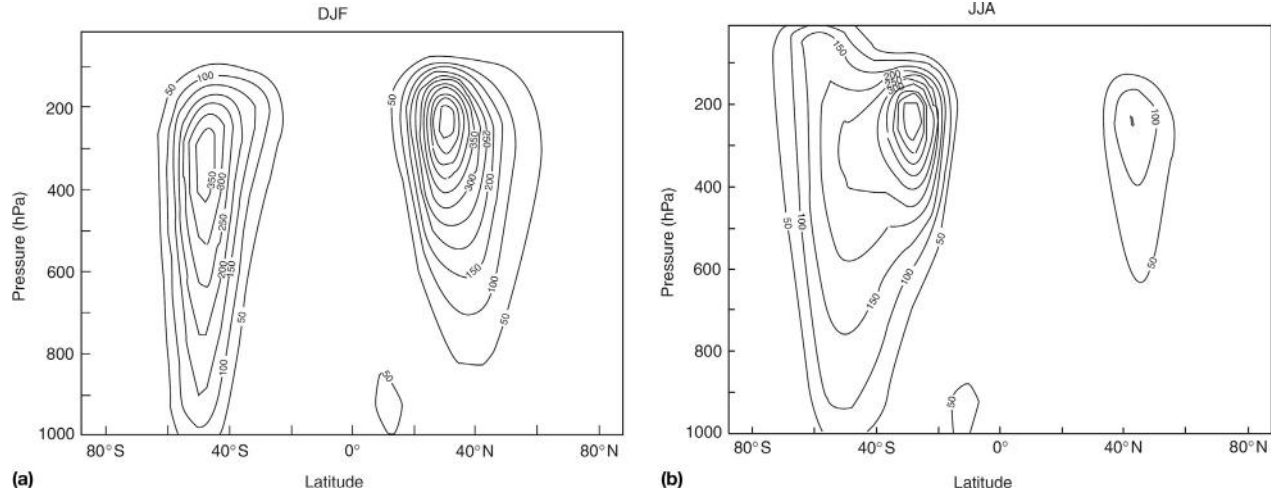


Figure 5 Zonal mean KE density for (a) December–February and (b) June–August using 1979–99 NCEP/NCAR reanalysis data. The contour interval is $50 \text{ kg s}^{-2} \text{ m}^{-1}$.

the specific volume, Φ is the geopotential, and P subscript indicates evaluation on isobaric surfaces.

The terms are ordered to match similar processes. Term 'a' in each equation has diabatic source/sink mechanisms. Term 'b' is similar but has opposite sign in the APE and KE equations and so represents a (baroclinic) conversion between these two forms of energy. Term 'c' is divergence of PE or KE flux; it is a conversion between the APE or KE inside and that external to the domain; and baroclinic or barotropic conversions, respectively, appear in this term.

Diabatic Sources and Sinks of Energy

There are five categories of diabatic processes: solar and terrestrial radiation, latent and sensible surface heat flux, and friction. **Figure 6** illustrates how diabatic heating is distributed on the seasonal and zonal means.

1. Solar radiation absorbed. Over much of the atmosphere the emission exceeds the absorption (net radiative cooling

(NRC) in **Figure 6**) but in the tropical and summer portions of the stratosphere there is net solar absorption (NRH).

2. Terrestrial radiation emitted. NRC (**Figure 6**) is very strong in the winter high latitudes and subtropics through the depth of the atmosphere. The net cooling is strong in the Southern Hemisphere subtropics in summer. Since emission is stronger from lower (warmer) cloud tops, the net cooling increases toward lower subtropical elevations. While the terrestrial emission ($q < 0$) is greater in the tropics, suggesting destruction of APE, the bulk of the net cooling in the diabatic heating field is in the winter hemisphere. APE is generated because the stronger net emission in high latitudes is from cloud tops where ε is strongly negative. In the less cloudy subtropical latitudes emission mainly occurs where ε has smaller magnitude.
3. Latent heat release. Some solar radiation absorbed by the Earth's surface evaporates water. Evaporation introduces water vapor into the atmosphere, primarily in the subtropics. The latent heat is released some distance away where

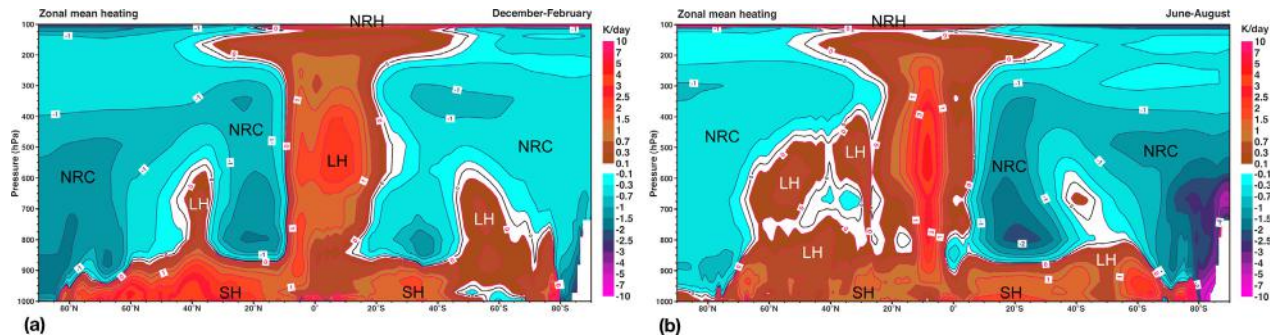


Figure 6 Zonal mean heating calculated as a residual in the temperature conservation equation. Areas near zero have no shading. Areas where shading darkens as values become larger in magnitude are negative values (cooling, 'blue' if in color). Areas where shading becomes lighter as values become larger are positive (heating, 'warm colors' if in color). Labels indicate the processes that tend to be larger in given regions. NRC indicates net radiative cooling, NRH means net radiative heating, LH indicates latent heating, SH indicates surface sensible heat flux mixed upward in the boundary layer. Units are Kelvin per day. Reproduced from Källberg, P., Berrisford, P., Hoskins, B., et al. 2005. European Centre for Medium Range Weather Forecasts (ECMWF) ERA-40 Publication 19: The ERA-40 Atlas, 191 pp.

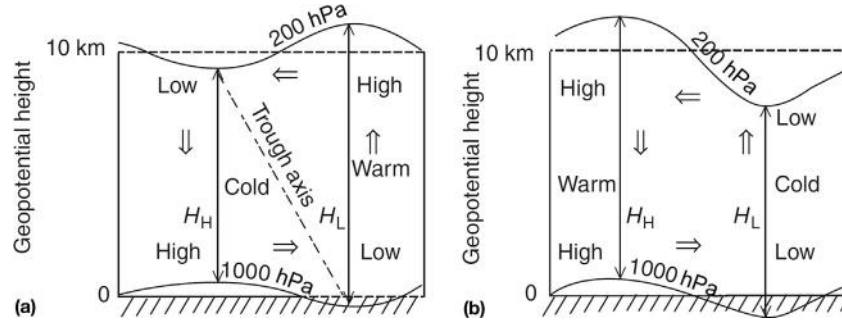


Figure 7 Highly idealized schematic illustration of a midlatitude frontal cyclone during (a) baroclinic growth and (b) baroclinic decay. The 1000 and 200 hPa surfaces are marked. Areas of relatively warmer and colder air at an isobaric level are noted. Double-shafted arrows indicate the relevant part of the ageostrophic circulation. The dot-dashed line in (a) shows the trough axis through the troposphere.

condensation occurs. In the tropics, much latent heat is released in the middle troposphere (mainly in the inter-tropical convergence zone (ICZ)) where $\varepsilon > 0$, hence creating APE. Moisture from the subtropics as well as evaporation over warm waters of an oceanic western boundary current (WBC; e.g., the Gulf Stream) is transported poleward and feeds precipitation in the midlatitude storm tracks. So midlatitude secondary maxima of latent heating occur in the lower and middle troposphere. The midlatitude latent heating is more apparent in summer because the opposing NRC is much stronger in winter. The Asian summer monsoon is seen as a heating maximum near 30° N in Figure 6(b).

4. Surface sensible heat flux. Solar radiation absorbed by the Earth's surface also heats up the ground and that is mixed into the air by turbulent processes. The transfer of sensible heat is large in the subtropics and also near midlatitude WBCs. Generally, the sensible heating is where $\varepsilon > 0$, thereby creating APE. Sensible heating is input into the cold air sector of the extratropical cyclone, thus destroying the temperature contrast between the warm and cold sectors, and so indicating APE destruction. However, sensible heating in the cold sector does lower the static stability, which allows vertical motions to proceed more freely encouraging baroclinic conversion.
5. Friction is only important for KE and it always destroys KE.

Baroclinic Conversions

Terms 'b' in eqns [7] and [8] are similar though opposite in sign, thereby indicating conversion between APE and KE in the limited volume energy equations. Using the hydrostatic and continuity equations:

$$\vec{\nabla}_P \cdot \nabla_P \Phi = \omega \alpha + \nabla_3 \cdot (\vec{\nabla}_3 \Phi) \quad [9]$$

For a closed system, there is no mass divergence and the pressure work term ($\nabla_3 \cdot (\vec{\nabla}_3 \Phi)$) vanishes when integrated over the domain. A large fraction of the wind is described by the geostrophic wind and since the geostrophic wind is proportional to the cross product of Φ , the geostrophic wind does not contribute to the left-hand side of eqn [9]. Hence, ageostrophic motions across geopotential height contours contribute to the left-hand side. For an open system, $\vec{V}_P \cdot \nabla_P \Phi \ll \nabla_3 \cdot (\vec{\nabla}_3 \Phi)$ and so there must be partial cancellation with the $\omega \alpha$ term to have the right-hand side also small. $\vec{V}_3 \Phi$ is a flux of geopotential that

can transport (gravitational) PE into or out of a region. Since $\varepsilon \ll 1$, $\varepsilon \omega \alpha$ has comparable magnitude as $\vec{V}_P \cdot \nabla_P \Phi$.

Two approximate forms of this conversion aid interpretation of the process. For a closed system illustrated in Figure 7, the conversion depends on the difference in thickness above the surface high (H_H) versus above the surface low (H_L):

$$-S g |\omega| (H_H - H_L) \sim \int \vec{V}_P \cdot \nabla_P \Phi \, dM \quad [10]$$

S is the horizontal area for each half of the domain and $|\omega|$ is the magnitude of the mean ω . Since thickness is proportional to the mean temperature of the layer of air, the sign and magnitude of the conversion depend on vertical motion in relatively warmer and colder regions. In Figure 7(a), warm air over the surface low rises while cold air sinks, which means $H_L > H_H$ resulting in cyclogenesis (APE $\rightarrow$ KE). In Figure 7(b), warm air overlies the surface high, so it has greater thickness than the air over the low resulting in cyclolysis. The magnitude of the conversion during the developing stage exceeds that during the decay stage implying net generation of KE.

A simple description of the baroclinic conversion is 'warm air rising or cold air sinking converts APE into KE'. The relationship is seen in the quasigeostrophic system, where

$$-\omega \alpha \approx g \mu w \theta \quad [11]$$

and $\mu \ll 1$ is a nondimensionalizing constant. The Carnot cycle model illustrates this mechanism as does the A_Z to K_Z conversion found in the tropical Hadley cells. (A_Z is the APE constructed from zonal mean quantities and K_Z is the KE constructed from zonal mean winds.)

For the Hadley cell, the integrand of term b in eqn [8] can be approximated by $-[v] \frac{\partial [\Phi]}{\partial y}$ if one neglects zonally varying phenomena like the Asian Monsoon and the Walker Cell. There is no zonal component because the geopotential term is averaged around a latitude circle and identically zero (if no topography is intercepted, i.e., if there is no mountain torque). The low-level flow in the Hadley cell is down the gradient of geopotential (from higher to lower Φ), so the term (including the minus sign) is positive and thereby generating zonal mean KE. For the upper level generally poleward return flow, both the meridional wind and the Φ gradient reverse sign and the KE are being created. Consistently, the rising ($\omega < 0$) branch of the

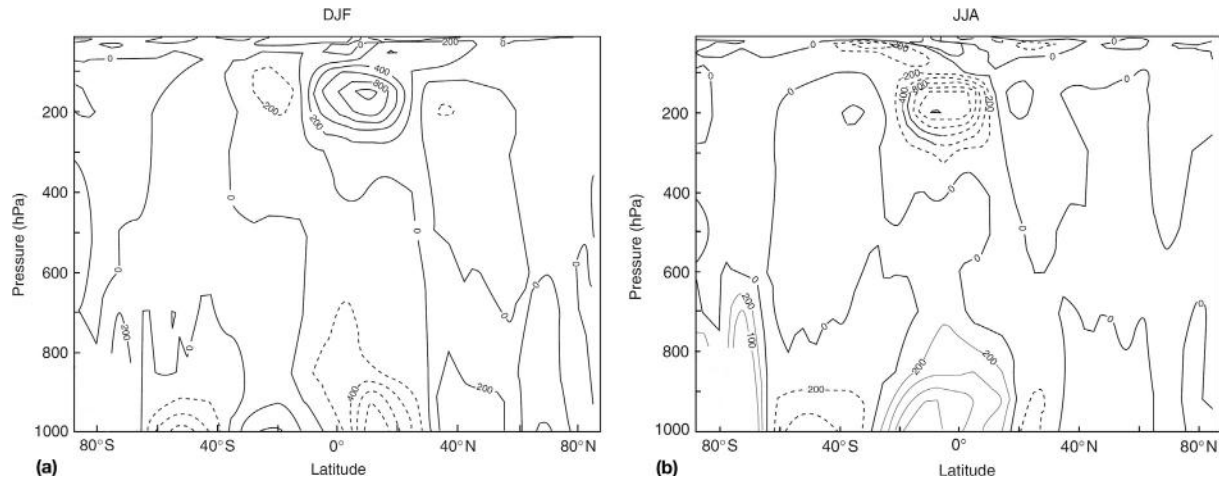


Figure 8 Zonal mean meridional flux of potential temperature [$v\theta$] for (a) December–February and (b) June–August using 1995–99 NCEP/NCAR reanalysis data. Hadley cell fluxes are apparent in the tropics, while eddy fluxes predominate in midlatitudes.

Hadley cell occurs in the warmer air ($\varepsilon > 0$), so term b in eqn [7] is negative and APE is being lost. Assuming that half the KE conversion takes place in a 300 hPa deep lower layer, moving at 3.5 m s^{-1} , with a geopotential height gradient of 10^{-4} , the full KE conversion in eqn [8] is 2.1 W m^{-2} . The results are consistent with the Carnot cycle calculation and indicate that the KE conversion is small relative to the solar absorption.

In middle latitudes zonally varying phenomena (eddies) dominate the circulation. To drive midlatitude eddies, two conversions are incorporated into the ‘baroclinic’ conversion label: A_Z to A_E and A_E to K_E . In the quasigeostrophic context, the latter is proportional to a vertical eddy heat flux while the former is proportional to a horizontal eddy heat flux. From thermal wind balance, the environment has a meridional temperature gradient and westerly vertical shear. Eddy horizontal heat fluxes are directed down the temperature gradient (toward the pole) if the eddy tilts against the vertical shear. The upstream tilt and vertical heat flux are both visible in the schematic diagram (Figure 7(a)) of cyclogenesis.

Figure 8 shows the observed zonal mean heat flux by all motions. In the tropics the heat flux follows the Hadley cell: lower-level equatorward heat flux and upper-level poleward heat flux. Because MSE is larger at the higher elevation (Figure 1) each Hadley cell has a net poleward heat flux. At higher latitudes heat is advected poleward, mainly in the lower troposphere. The midlatitude eddies have sizable poleward heat flux in the lower troposphere and somewhat less poleward flux in the upper troposphere. The mean meridional (Ferrel) cell circulation opposes the upper troposphere eddy flux and reinforces the lower troposphere eddy flux making the net $[v\theta]$ flux larger in the lower troposphere.

Intra-KE (Barotropic) Conversion

The barotropic conversion rearranges KE. A commonly shown redistribution is between zonal mean and eddy KE. In the KE tendency eqn [8] this conversion (term c) appears as a boundary flux that is zero for a closed domain. The term originates from the horizontal advection terms in the original

component momentum equations and when one derives the conversion between zonal mean and eddy KE then the term is related to eddy momentum flux convergence as well as the mean flow horizontal shear. Zonal mean KE is generated from eddy KE where eddy momentum fluxes are up the gradient of the mean flow.

The eddy momentum fluxes, meridional cells, and jets are linked in the barotropic mechanism. Meridional momentum transport (Figure 9) tends to be largest in the upper troposphere. One reason for this is that midlatitude eddy momentum transport becomes stronger as frontal cyclones reach a mature stage. Momentum transport by the Hadley cell is up the gradient of the zonal mean zonal wind toward the subtropical jet stream; the poleward moving air contains large angular momentum that causes air parcels to accelerate relative to the Earth’s surface. The meridional cells have little meridional motion at the subtropical jet, but the eddies carry momentum further poleward. The Ferrel cell momentum flux opposes the flux by the eddies in the upper troposphere.

Complexity arises from several sources.

1. The eddies have preferred regions of genesis and decay and their momentum fluxes vary greatly between these regions.
2. Mature lows migrate to the cold side of the jet and thus deflect the jet stream equatorward.
3. Eddy momentum fluxes build vertical shear while the eddy heat fluxes reduce the temperature gradient, a combination that destroys thermal wind balance. One consequence is the formation of a secondary circulation, appearing as the ‘Ferrel cell’ on a zonal average, which partially opposes the eddy heat and momentum fluxes. So, the jet streams are equatorward of where the eddy momentum flux has greatest convergence. This Ferrel cell brings westerly momentum downward.
4. The momentum flux convergence in Figure 9 is largest in the winter hemisphere between 30 and 40° N in the upper troposphere. This location of momentum convergence thus seems to be poleward of the subtropical jet. The jet and momentum convergence positions match better if one

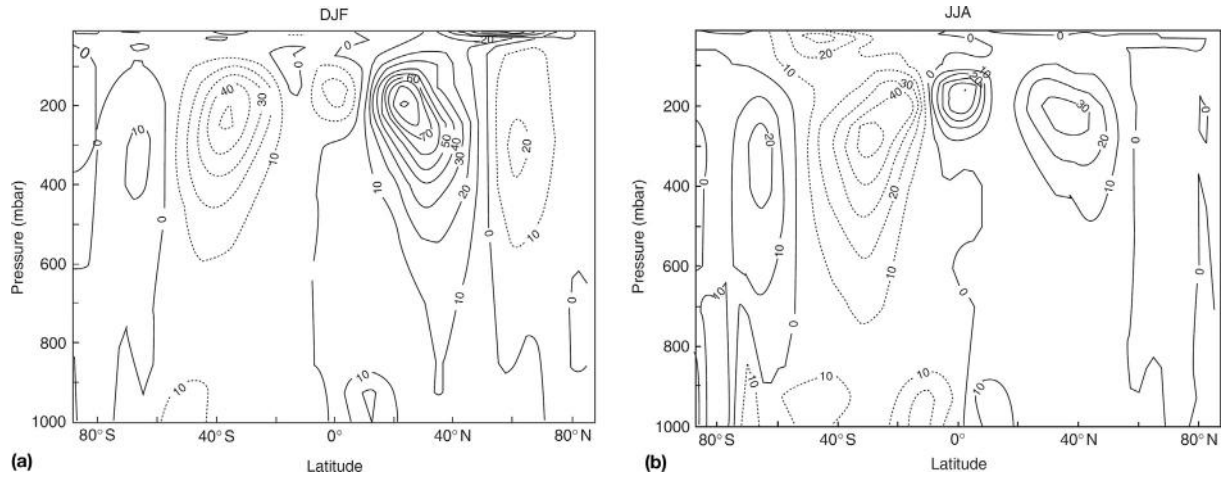


Figure 9 Zonal mean meridional flux of zonal wind $[vu]$ for (a) December–February and (b) June–August using 1995–99 NCEP/NCAR reanalysis data. Hadley cell fluxes are apparent in the tropics, while eddy fluxes predominate in midlatitudes.

accounts for the spherical geometry. The zonal component of zonal mean KE tendency is given by eqn [12]:

$$\frac{\partial}{\partial t} \int \left(\frac{[u]^2}{2} \right) dm = - \underbrace{\int [u] \left(\frac{1}{r \cos^2 \varphi} \frac{\partial \{ \cos \varphi \}^2 [uv]}{\partial \varphi} + \frac{\partial [u\omega]}{\partial P} \right) dm}_c + \int f [u] [v] dm - \int [u] [F] dm \quad [12]$$

where r is the radius of the Earth, φ is the latitude, $[\]$ indicates a zonal average, f is the Coriolis parameter, and dm is an increment of atmospheric mass. The subtropical jets have large contribution to the zonal component of KE: $[u]^2$ and the momentum flux $[uv]$ are modified by $\cos(\varphi)$ and $[u]$ factors that shift the larger values of the integrand (term c) to lower latitudes closer to the subtropical jet location.

The eddies need horizontal tilts to accomplish the observed momentum fluxes. **Figure 10** illustrates the southwest to northeast tilts for poleward transport by Northern Hemisphere eddies.

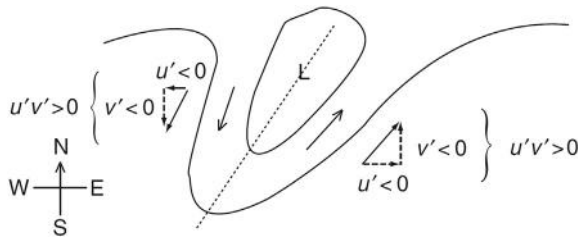


Figure 10 Schematic illustration of how a horizontal tilt of the trough axis (dotted line) leads to a net meridional transport of eddy zonal momentum. Primes denote winds with the zonal average removed. In this case the zonal average eddy momentum flux is northward. In contrast, a low that is symmetric about a north–south axis has $u'v'$ contributions on the east and west sides that cancel in the zonal mean.

Observed Energetics

Global Energy Balance

Ignoring seasonal heating and climatic change heating, there should be a balance between the solar energy absorbed by the Earth and that radiated away to space. The actual energy budget between the Earth and space depends on a variety of factors such as cloud cover, atmospheric composition, and surface properties. Estimates of energy fluxes for the Earth's surface and atmosphere on global and annual averages are presented in **Figure 11**.

Some limitations of this summary depiction are listed below:

1. The balance shown is global, so net radiation is zero. Net radiation (solar minus terrestrial) is positive from 38° N to 38° S and negative elsewhere.
2. Heat fluxes sustain the net radiation pattern; those motions are not included.
3. Vertical fluxes of heat and radiation are not shown, only the net transfer for the whole atmosphere.

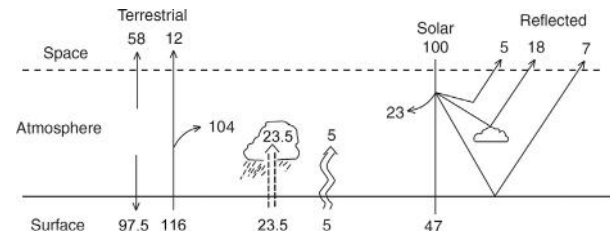


Figure 11 Global average energy balance expressed as percentages of the solar radiation striking the top of the atmosphere. Estimates of the solar constant vary, but the 100 units in the figure correspond to $\sim 341 \text{ W m}^{-2}$. Right side: solar radiation processes showing amounts reflected and absorbed. Middle: surface sensible heat flux (wavy arrow) and surface latent heat flux (dashed arrow). Left side: terrestrial radiation processes showing emission, transmission, and absorption. Data in the figure are based primarily on Trenberth, K., Fasullo, J., Kiehl, J., 2009. Earth's global energy budget. *Bulletin of the American Meteorological Society* 90: 311–323.

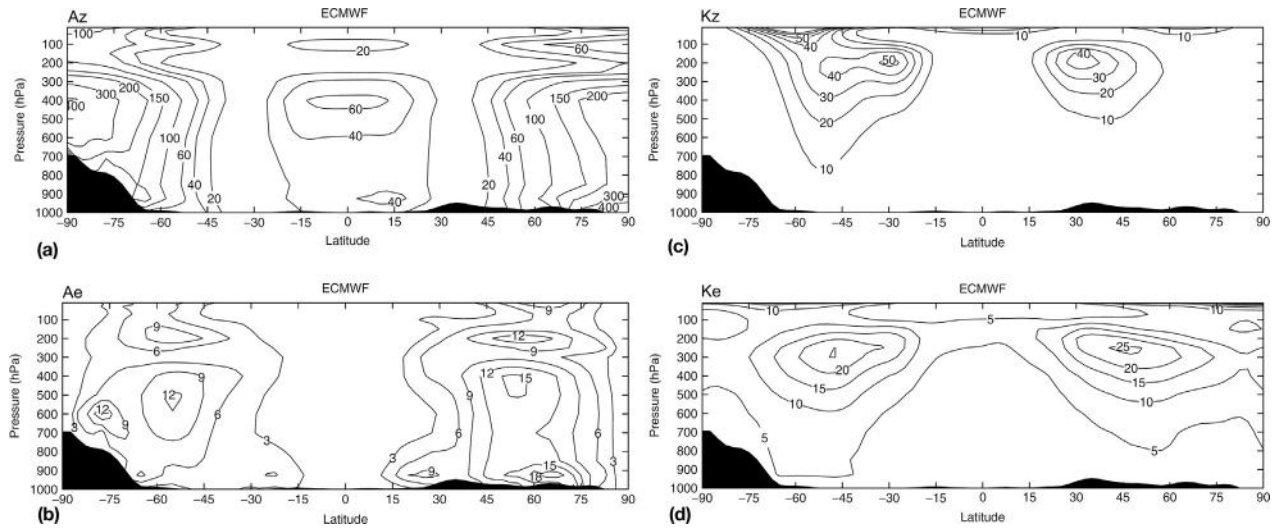


Figure 12 Zonal and annual average APE (a) of the zonal mean state, A_z , and (b) of the zonally varying state, A_e . Zonal and annual average KE (c) of the zonal mean flow, K_z , and (d) of the zonally varying flow, K_e . The units are $10^5 \text{ J m}^{-2} \text{ bar}^{-1}$. Reproduced from Marques, C., Rocha, A., Corte-Real, J., Castanheira, J., Ferreira, J., Melo-Gonçalves, P., 2009. Global atmospheric energetics from NCEP–Reanalysis 2 and ECMWF–ERA40 Reanalysis. *International Journal of Climatology*, 29: 159–174.

Figure 11 shows the following:

1. At the Earth's orbit, the longwave radiation from the Earth greatly exceeds the longwave radiation from the Sun. So, solar (shortwave) radiation can be treated separately from terrestrial (longwave) energy.
2. The (shortwave) albedo is greatly affected by clouds, which also strongly affect terrestrial emission.
3. More solar radiation is absorbed by the ground (47%) than by the air (23%).
4. The solar radiation reaching the ground evaporates water, is emitted as longwave radiation, or creates a sensible heat flux, in that order of (net) magnitude.
5. The net surface emission must balance the input (18.5 units) not lost by surface fluxes. However, the actual surface longwave emission (116 units) exceeds the shortwave input (47 units) because of downward radiation from the atmosphere, this means the surface temperatures are larger than expected from just the shortwave absorption, an increase known as the 'greenhouse' effect.
6. The average global annual energy absorbed is $\sim 235 \text{ W m}^{-2}$. Multiplied by the Earth's surface area, this is about 121 PW ($1.2 \times 10^{17} \text{ W}$).

The Energy Box Diagram

Since the tropical circulations are dominated by the zonal average Hadley circulation and the midlatitude weather is dominated by the zonally varying frontal cyclones, it is logical to partition the energy into zonal average and zonally varying (eddy) parts. The total APE can be partitioned into zonal average (A_z) and eddy (A_e) parts. Similarly, the total KE is partitioned into zonal average (K_z) and eddy (K_e) parts. These four categories of energy are shown in Figure 12. Zonal average of the APE in the zonal mean fields (Figure 12(a)) is large in the tropics and increases from a subtropical minimum to largest values at

the poles. The zonal average of APE from eddy structures is largest in the middle latitudes (Figure 12(b)). The zonal mean KE of the zonal mean flows (Figure 12(c)) has largest values in the middle latitudes where the major jet streams are found, including the dual jet in the Southern Hemisphere during local winter. The zonal mean KE from eddy motions (Figure 12(d)) is largest in the middle latitudes, with largest values in the upper troposphere. Compared with K_z the distribution of K_e is broader in the vertical and centered at higher latitudes. Comparison with Figure 5 indicates that most of the tropopause level KE is in the zonal mean flow while in the middle and lower troposphere the KE is composed of approximately equal contributions from zonal mean and eddy flows.

The sources and sinks of energy discussed above can be summarized for the atmosphere using the 'box' diagram. Each box discussed here is a combination of zonal or eddy quantities that contribute to KE or APE. Other forms are possible, such as transient versus time mean or subdividing the eddies into different wave number groups. Arrows indicate input and output from each box as identified in the energy equations. Energy is converted back and forth between various forms, but only net changes are shown. The Hadley circulation is contained in the zonal mean quantities at the top row of the box diagram. Midlatitude frontal cyclones are included in the bottom row. The box diagram has some limitations:

1. The diagram only refers to global mean properties. For example, A_z to A_e is largest mainly in middle latitudes.
2. Only net changes are shown, whereas large regional variations occur. A_z to $K_z > 0$ for the Hadley cells but < 0 for Ferrel cells. Some studies show the global value of this conversion as negative.
3. Some conversions are simultaneous, as mentioned above, especially the A_z to A_e to K_e route by which frontal cyclones intensify. The small size of the A_z to K_z conversion hides a redundant geostrophic and hypsometric link between A_z and K_z .

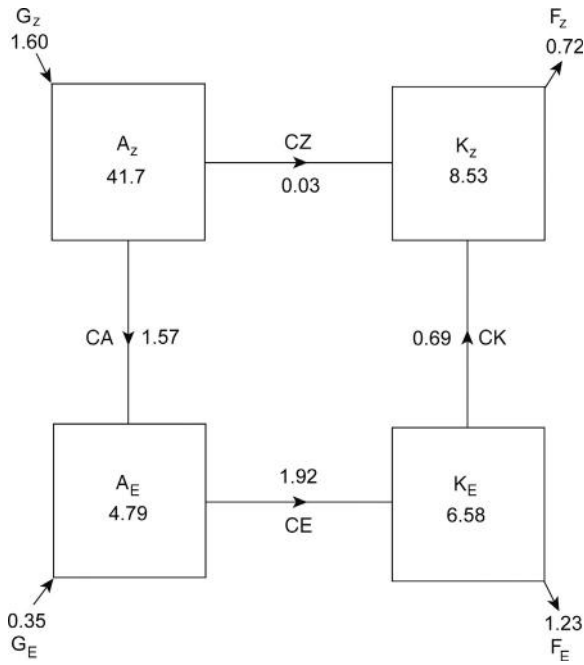


Figure 13 Energy ‘box’ diagram showing reservoirs of zonal mean APE (A_Z) and KE (K_Z); eddy APE (A_E) and KE (K_E) in the same orientation as Figure 12. Also shown are net energy conversions (CA, CE, CK, CZ); and net diabatic generation or destruction. Global and annual average values indicated. The reservoirs are in units of 10^5 J m^{-2} and the conversions and diabatic processes have units of watts per square meter. The conversions are given letter labels for future reference. The numbers for A_Z and A_E are somewhat arbitrary, but consistent with each other. The values are redrawn from Marques et al. (2009) and based on ERA-40 data from 1979 to 2001. Marques, C., Rocha, A., Corte-Real, J., Castanheira, J., Ferreira, J., Melo-Gonçalves, P., 2009. Global atmospheric energetics from NCEP–Reanalysis 2 and ECMWF–ERA40 Reanalysis. *International Journal of Climatology* 29: 159–174.

4. Some conversions are hard to measure directly and are sometimes approximated (e.g., A_Z to K_Z) or deduced as a residual (e.g., G_E and A_E to K_E).
5. The box diagram gives arbitrary if not misleading results for certain situations of wave–mean flow interaction.

Observed reservoirs, conversions, sources, and sinks of energy are depicted in Figure 13. Diabatic processes create much A_Z which in turn drives K_Z of zonal mean circulations such as the Hadley cells and the zonal mean midlatitude westerlies. Some A_Z is converted into A_E which becomes K_E especially in midlatitude frontal cyclones. Convergence of eddy momentum is a net conversion of K_E to K_Z . Since friction removes K_E there must be net A_E to K_E conversion. Depending on the estimated strength of eddy frictional loss: F_E , eddy APE generation: G_E may be negative.

The box diagram for global energy shows the following:

1. While the total value of the APE is somewhat arbitrary, the relative sizes of similar energy ‘reservoirs’ are worth noting. For example, K_E is almost half of the total KE. A_E is about a tenth of the total APE.

2. The flow of energy consists primarily of G_Z to A_Z to A_E to K_E to F_E . However, the small conversion CZ can be misleading since there is some arbitrariness in how the conversions are defined as will be discussed below.
3. Different phenomena follow different paths. The A_Z to K_Z relates to the Hadley cell in terms of Figure 3. Baroclinic instability is A_Z to A_E to K_E and obviously a primary track in the diagram. Barotropic instability is K_Z to K_E and it is clearly negative (eddies feed KE into the zonal mean flow).

Transformed Eulerian Mean (TEM) and Stratospheric Energy Box Diagram

An alternative energy analysis can be constructed in a TEM framework based on an ‘eddy-induced circulation’ that results (in large part) from eddy heat fluxes. The purpose of the residual circulation is to simplify the net effect of the eddies on the mean flow in part by including their contribution to a zonal mean meridional circulation. The residual circulation (*subscripts) might be defined as

$$[v_*] = [v] - \frac{1}{p} \frac{\partial}{\partial z} \left(p \frac{[v' \theta']}{\partial [\theta]} \right) \quad [w_*] = [w] + \frac{\partial}{\partial y} \left(p \frac{[v' \theta']}{\partial [\theta]} \right) \quad [13]$$

The TEM formalism is applied more commonly to understand the stratospheric general circulation than the tropospheric. One reason is that the divergence of Eliassen–Palm flux term in the zonal momentum equation has a lot of cancelation with the flux of planetary angular momentum by the TEM residual circulation. For stratospheric planetary waves, the residual circulation induced by eddy heat flux convergence exceeds that from the diabatic heating, but that does not hold for the troposphere, in part due to the lower boundary.

If there is nonacceleration, then there is no net gain of energy in the different ‘energy boxes.’ But in some non-acceleration situations the ‘conventional’ formulation of the previous section has transfers of energy, though they are such that energy input to a ‘reservoir’ equals energy output. The transfers of energy could be viewed as misleading, since there is no net generation or destruction despite the large conversions with adjacent reservoirs. Under the same nonacceleration assumption, the energy conversions are all zero in the TEM formulation. While that might seem superior, problems arise when the nonacceleration condition is not met such as for growing baroclinic waves.

The definitions of the energy conversions differ between the TEM and conventional formulations. The TEM formulation does not allow the CA conversion, instead, energy is exchanged between waves and zonal mean through a CK conversion. However, in the TEM formulation the conversion between zonal and eddy KE is related to the Eliassen–Palm flux, which in turn is a vector that depends on eddy momentum and eddy heat fluxes.

When considering the stratosphere alone, a new source or sink of KE arises from the work done by deformation of the tropopause. These boundary fluxes are substantial and a major driver for planetary-scale stratospheric circulations. In a conventional energy cycle depiction (like Figure 13) the energy conversions would appear large, but largely go in a loop: K_Z to A_Z

to A_E to K_E and back to K_Z in the stratosphere during winter. So, much of the large conventional energy conversions do not result in corresponding large net gain or loss of energy in the individual components. The same situation in the TEM framework shows mechanical forcing of K_Z , most of which is transferred to K_E and then to A_E , where it is lost by diabatic cooling. The TEM framework is more clear than the conventional view, though it might be more physically reasonable to have a direct forcing of the stratospheric K_E by the tropospheric longest waves instead of going through K_Z first. During summer, in the stratosphere the eddies transport heat poleward making the polar region warmer than it would be from local radiative balance (geometry makes the incident solar radiation larger at the pole than at the adjacent middle latitudes), so the presence of the eddies results in polar radiative net cooling instead of net warming and the energy flow is from eddies to zonal mean.

Column Average Energetics

Vertically integrating the combination of the temperature conservation and KE equations yields

$$\begin{aligned} \frac{1}{g} \frac{\partial}{\partial t} \int_0^{P_s} (C_p T + \Phi_s + KE) dP \\ = -\frac{1}{g} \nabla_H \cdot \int_0^{P_s} (C_p T + \Phi + KE) V_H dP + Q_h + Q_f \end{aligned} \quad [14]$$

Subscript H indicates the horizontal components, KE is defined in eqn [1], and Φ_s is the surface geopotential. Q_h contains the net diabatic heating terms from the temperature conservation equation: net downward radiation at the top of the atmosphere, surface heat flux, and precipitation rate, all expressed as a heating rate. Q_f is heating due to friction (which is very small, $\sim 2 \text{ W m}^{-2}$). Vertically integrating the moisture conservation equation yields

$$\frac{L}{g} \frac{\partial}{\partial t} \int_0^{P_s} q dP = -\frac{L}{g} \nabla_H \cdot \int_0^{P_s} q V_H dP - Q_w \quad [15]$$

where Q_w is the difference in precipitation minus surface evaporation rates in the atmosphere, expressed as a heating rate. Combining the two previous equations gives an equation for a type of TE that includes internal, potential, kinetic, and latent energy:

$$\begin{aligned} \frac{1}{g} \frac{\partial}{\partial t} \int_0^{P_s} (C_p T + \Phi_s + KE + Lq) dP \\ = -\frac{1}{g} \nabla_H \cdot \int_0^{P_s} (C_p T + \Phi + KE + Lq) V_H dP + Q_h + Q_f - Q_w \end{aligned} \quad [16]$$

Hence, vertical and horizontal advections of MSE + KE are balancing the diabatic processes. A similar balance, between radiative cooling and adiabatic warming of sinking air, was exploited for the Carnot cycle model of the Hadley cell discussed above. Ignoring the tiny Q_f , eqn [14] shows that Q_h is proportional to a divergence of the flux of the combination DSE + KE. Similarly, from eqn [16] $Q_h - Q_w$ is proportional to the divergence of the TE. Hence, those fluxes can be deduced from the Q_h and $Q_h - Q_w$ fields and reveal large-scale flows of energy within the atmosphere.

Figure 14 shows the distribution of Q_h and Q_w for the extreme seasons; the total diabatic energy sources and sinks ($Q_h - Q_w$) are subdivided into quasistationary and transient contributions along with the corresponding energy fluxes. Figure 14 shows the following:

1. Q_h has strong sources in regions with heavier precipitation such as the ICZ, south Pacific convergence zone (SPCZ), and winter midlatitude frontal cyclone storm tracks. Strong surface sensible heat flux over WBCs also amplifies the Q_h .
2. Q_h is a strong sink: over winter continents due to intense radiative cooling and the eastern subtropical oceans due to negative net radiation.
3. Q_w has maxima where there is net precipitation (ICZ, SPCZ, and midlatitude winter storm tracks) and minima where there is greater net evaporation (subtropical oceans). Evaporation exceeds precipitation more strongly during winter in the subtropics. The moisture fluxes are out of the subtropics toward the ICZ, SPCZ, and midlatitude storm tracks.
4. The slowly varying diabatic sources of TE, $Q_h - Q_w$ are shown in Figure 14(c) and 14(g). The sources of TE are areas of strong evaporation over the subtropical and tropical oceans. Precipitation does not enter into eqn [16] because the flux (and hence conservation) of MSE means that latent heating is compensated by adiabatic cooling of the rising air, such as in the ICZ. Strong surface heat fluxes (and some evaporation) occur over the WBCs during winter. (During summer there is a sink over the WBCs.) During winter there is a sink of TE at higher latitudes due to the strong negative net radiation, with some enhancement of the sink over the continents.
5. The flow of TE for monthly mean and longer features is from the tropics, subtropics, and WBCs toward the higher latitudes in the vertical mean. However, for the Hadley cell, it has to be transported first equatorward, then upward, and then poleward, including phase changes of the water.
6. The faster varying diabatic sources of TE: $Q'_h - Q'_w$ (Figure 14(d) and 14(h)) are generally along the equatorial side of the midlatitude storm tracks with greater intensity during winter months. The sink tends to be immediately poleward of the storm tracks, a pattern reinforced by the TE flux.
7. The flux of TE is greatest where the $Q_h - Q_w$ gradient is greatest. The peak is in the middle latitudes ($\sim 40^\circ$ latitude) with peak value of $\sim 5 \text{ PW}$. The ocean transports heat as well, though peak values amount to $\sim 1/4$ (Northern Hemisphere) to $\sim 1/10$ (Southern Hemisphere) of the transport by the atmosphere.

Summarizing the Energy Cycle

Energy in the general circulation follows paths from sources to sinks. Energy is mathematically cast in several useful forms and it is linked to several global constraints. To provide an overview of these diverse properties, Table 1 and Figure 15 summarize how energy in kinetic and potential forms flows in the general circulation. Heat and APE are similar concepts and are compared in the table and figure. Similarly, momentum and KE may be considered together. The table and figure make clear that heat and momentum have similar circuits.

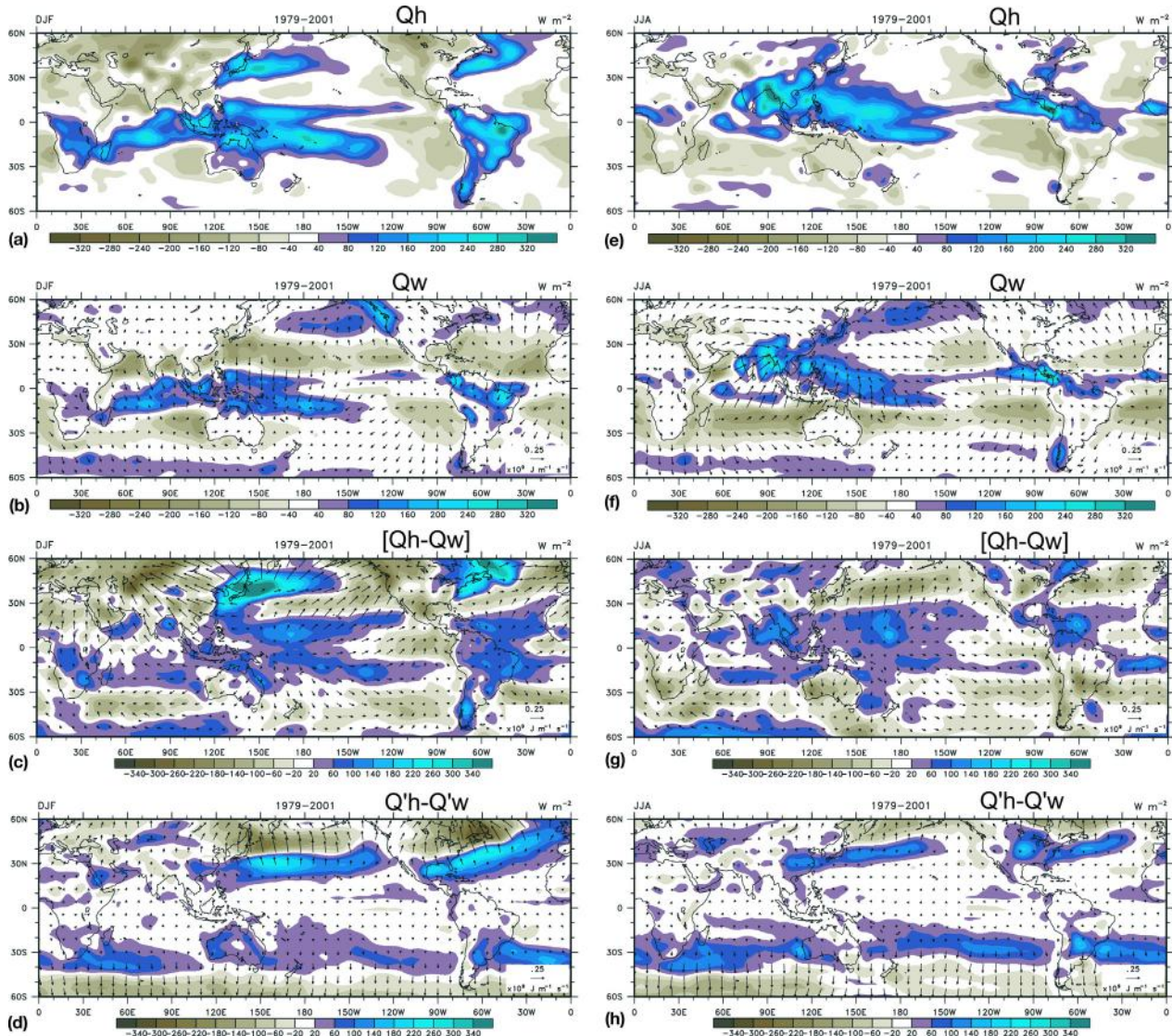


Figure 14 Vertically integrated components of the energy flow during December–February (left column) and June–August (right column) using data from 1979–2001. (a) and (e) are net diabatic heating. (b) and (f) are moisture sources and sinks (contoured) and the implied net transport (vectors). (c) and (g) are TE change (contours) and net energy fluxes (vectors) due to ‘quasistationary’ processes varying on monthly mean or longer timescales. (d) and (h) are similar to (c) and (g) but for ‘transient’ processes varying faster than monthly mean timescales. The sources and sinks are all expressed as energy fluxes with units of watts per square meter. The vector scales and units are indicated with a reference vector in the lower right corner of individual panels. Reproduced from Trenberth, K., Stepaniak, D., 2004, The flow of energy through the earth’s climate system. *Quarterly Journal of the Royal Meteorological Society*, 130: 2677–2701.

Table 1 Stages in the APE (heat) and KE (momentum) energy cycle

APE or heat flow	KE or momentum flow
1. Solar and terrestrial radiation create excess heating in the tropics and a deficit poleward of 38°. 2. Result of item (1) is a poleward heat flux. 3. In midlatitudes eddies are the main mechanism for heat transport. 4. The CA and CE conversions (see Figure 13) show that horizontal and vertical heat fluxes create eddy energy (baroclinic process). Latent heat release may also contribute. 5. Net radiation being positive causes the heat flux to increase with latitude in the subtropics. (More and more heat must be transported poleward to maintain quasisteady PE.) 6. The eddy heat flux is down the T gradient. 7. The heat flux is maximum where net radiation is zero. 8. Poleward of the heat flux maximum there is cooling by net radiation.	1. Westerly momentum is introduced in the tropics and is removed by friction in midlatitudes. 2. Result of item (1) is a poleward westerly momentum flux. 3. In midlatitudes eddies are the main mechanism for momentum transport. 4. Eddy momentum fluxes also provide sources and sinks of eddy KE from the CK conversion (see Figure 13). Since global average CK is positive, eddies must lose KE to the mean flow in the net (barotropic process). 5. The flux of zonal mean KE keeps increasing with latitude in the subtropics where CK is positive. (More and more momentum must be transported poleward to maintain quasisteady KE.) 6. Eddy momentum flux is up the gradient of $[u]\cos^{-1}\phi$ at many latitudes. 7. The flux of (KE) reaches a maximum where CK equals zero. 8. Poleward of the angular velocity, $[u]\cos^{-1}\phi$, maximum, eddies remove energy from (KE) in the net.

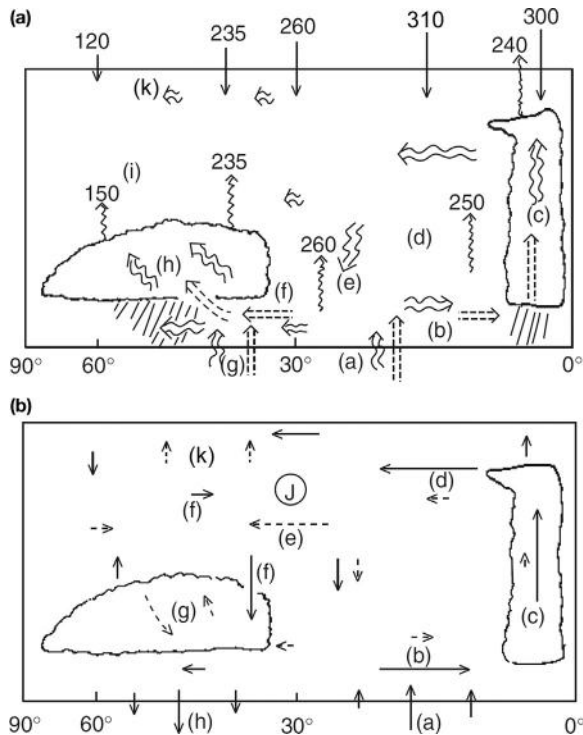


Figure 15 (a) Schematic meridional cross sections of the atmospheric cycles of (a) APE and heat and (b) KE and momentum. Arrow lengths are intended to suggest relative magnitudes. In panel (a) single-shafted arrows depict radiation (straight for absorbed solar; wavy for terrestrial emitted to space) with representative numerical values given in watts per square meter. Double-shafted arrows depict heat fluxes (dashed for latent; wavy for sensible heat). Letters correspond to distinct parts of the cycle. Solar radiation (a) is absorbed in the tropics and subtropics, then that heat energy is transported (b) equatorward. Latent heat (c) is converted into sensible heat in the ICZ then transported poleward (d) in the upper troposphere. Sinking in the Hadley cell (e) brings heat downward. Eddy fluxes (f) extract heat from the subtropics and are augmented by surface fluxes (g) at WBCs. Heat is mixed upward vertically (h) in frontal cyclones, whereupon a net loss occurs to space (i). A weak poleward flux (k) results from the eddy-induced stratospheric circulation. In panel (b), momentum transport in the troposphere is separated into mean meridional cells (solid arrows) and eddies (dashed arrows). Letters correspond to these parts of the cycle: Slowing down surface easterlies imparts westerly momentum (a) into the tropical boundary layer. That momentum is transported equatorward (b) then upward (c) in the ICZ convection. The upper Hadley cell transports the momentum (d) poleward. Eddies further transport the momentum (e) poleward, out of the subtropics, while the Ferrel cell (f) both opposes the eddy flux and mixes some momentum downward. Frontal cyclones also have a net downward mixing (g) of westerly momentum (e.g., behind cold fronts) that is lost by friction at the surface (h). The transport creates westerly momentum convergence in the subtropics and midlatitudes forming a subtropical jet (J) and a KE maximum there, (k) and other arrows in the stratosphere are meant to indicate mechanical forcing by the midlatitude ultralong waves and by deep tropical convection.

Westerly momentum is defined positive, so frictional slowing of easterlies in the tropical boundary layer is a *source* of westerly momentum but a *sink* of KE. The low-level flow in the Hadley cells gains westerly momentum and transports it equatorward. Surface fluxes of heat and water provide a diabatic source of warmth and moisture to air parcels as they approach the ICZ.

In the ICZ, westerly momentum is transported to the upper troposphere. Latent heat is released where the efficiency factor is large and positive, leading to strong diabatic generation of APE. While the precipitation does not contribute directly to the TE generation, it does reinforce the equatorward pressure gradient that drives the upper level Hadley cell motion, establishes the westerlies, and creates other indirect impacts.

The upper branch of each Hadley cell transports westerly momentum poleward. Conservation of angular momentum builds the velocity relative to the Earth's surface, i.e., creating KE and providing one mechanism to create the subtropical jets. The poleward extent of the motion arises from interaction between the (poleward) pressure gradient force and the Coriolis force (at a right angle to the motion); these two forces would accelerate parcels along a trochoidal path. Because the MSE of the poleward moving air is much greater than the MSE of the equatorward moving air below, the Hadley cell has a net poleward transport of heat.

The subtropical oceans are where much energy is input into the atmosphere through evaporation. The subtropics are also a transition between the convection-dominated tropical circulations and the frontal cyclone-dominated midlatitude circulations. Negative net radiation encourages parcels in the Hadley cell to sink, bringing down some westerly momentum as well as high potential temperatures. The frontal cyclones mix momentum and heat vertically. The vertical fluxes of heat by each cyclone and the vertical shear of the jet streams both are fundamental parts of the baroclinic instability mechanism.

In middle latitudes, air in the cyclones' warm sector has a poleward component of motion, while air in the cold sector moves equatorward: in both sectors *eddy* heat transport is poleward. Much precipitation accompanies the frontal cyclones with the bulk of it occurring in the warm sector, so there may be diabatic generation of eddy APE but loss of zonal APE. Further poleward, radiative cooling, especially from cloud tops, generates APE. Mature frontal cyclones develop momentum fluxes that converge upper level westerly momentum, while their heat fluxes reduce the meridional temperature gradient on a zonal average. To maintain thermal wind balance a secondary circulation forms which also transports momentum (equatorward at upper levels). The westerly momentum mixed to the surface (for example, in subsiding air of cold sectors) is removed by friction in the boundary layer, becoming a sink of westerly momentum and of KE.

In the stratosphere, there is input of mechanical (kinetic) energy by tropical convection and by midlatitude long waves as well as heating as ozone absorbs radiation. The eddy-induced circulation transports some heat poleward though angular momentum is mixed vertically and horizontally.

See also: **Climate and Climate Change:** Energy Balance Climate Models. **Dynamical Meteorology:** Balanced Flow; Baroclinic Instability; Overview; Quasigeostrophic Theory. **General Circulation of the Atmosphere:** Mean Characteristics. **Land-Atmosphere Interactions:** Overview. **Middle Atmosphere:** Transport Circulation. **Numerical Models:** General Circulation Models. **Stratosphere/Troposphere Exchange**

and Structure: Global Aspects. **Thermodynamics:** Moist (Unsaturated) Air. **Tropical Meteorology and Climate:** Hadley Circulation.

Further Reading

Dutton, J., Johnson, D.R., 1967. The theory of available potential energy and a variational approach to atmospheric energetics. In: *Advances in Geophysics*, Vol. 12. Academic Press, New York, NY. 333–436.

Grotjahn, R., 1993. *Global Atmospheric Circulations: Observations and Theories*. Oxford University Press, New York, NY.

James, I., 1994. *Introduction to Circulating Atmospheres*. Cambridge University Press, Cambridge, UK.

Karoly, D.J., Vincent, D.G., 1998. *Meteorology of the Southern Hemisphere*. American Meteorological Society, Boston, MA.

Peixoto, J.P., Oort, A.H., 1992. *Physics of Climate*. American Institute of Physics, New York, NY.

Trenberth, K.E., Stepaniak, D.P., 2004. The flow of energy through the Earth's climate system. *Quarterly Journal of the Royal Meteorological Society* 130, 2677–2701.

Weather Regimes and Multiple Equilibria

F Molteni, Abdus Salam International Centre for Theoretical Physics, Trieste, Italy

© 2003 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 6, pp 2577–2586, © 2003, Elsevier Ltd.

Introduction

The atmosphere can be regarded as a dynamical system with an infinite number of degrees of freedom. If one considers the whole spectrum of atmospheric phenomena, covering different spatial and temporal scales, the number of states that can be assumed by the atmospheric variables is indeed infinitely large. However, when interest is restricted to the large-scale features of the atmospheric circulation (especially in the extratropics during the cold season), many statistical analyses of the multidecadal record of observed data suggest the existence of preferred circulation patterns that seem to be particularly recurrent and/or persistent.

Examples of these recurrent flow types are presented in **Figure 1**: the monthly mean geopotential height at 500 hPa in January of four different years is shown, superimposed on the corresponding anomalies (i.e., deviations from climatology) over the northern extratropical regions. It is evident that the two circulation anomalies shown in each row of **Figure 1** are quite similar to each other, while there are strong differences between the anomalies in the top row and those in the bottom row. Indeed, these two sets of patterns are characterized by nearly opposite anomalies over the north-east Pacific and most parts of North America and Eurasia.

Several explanations have been proposed for the existence of preferred circulation patterns. Some of them are based on

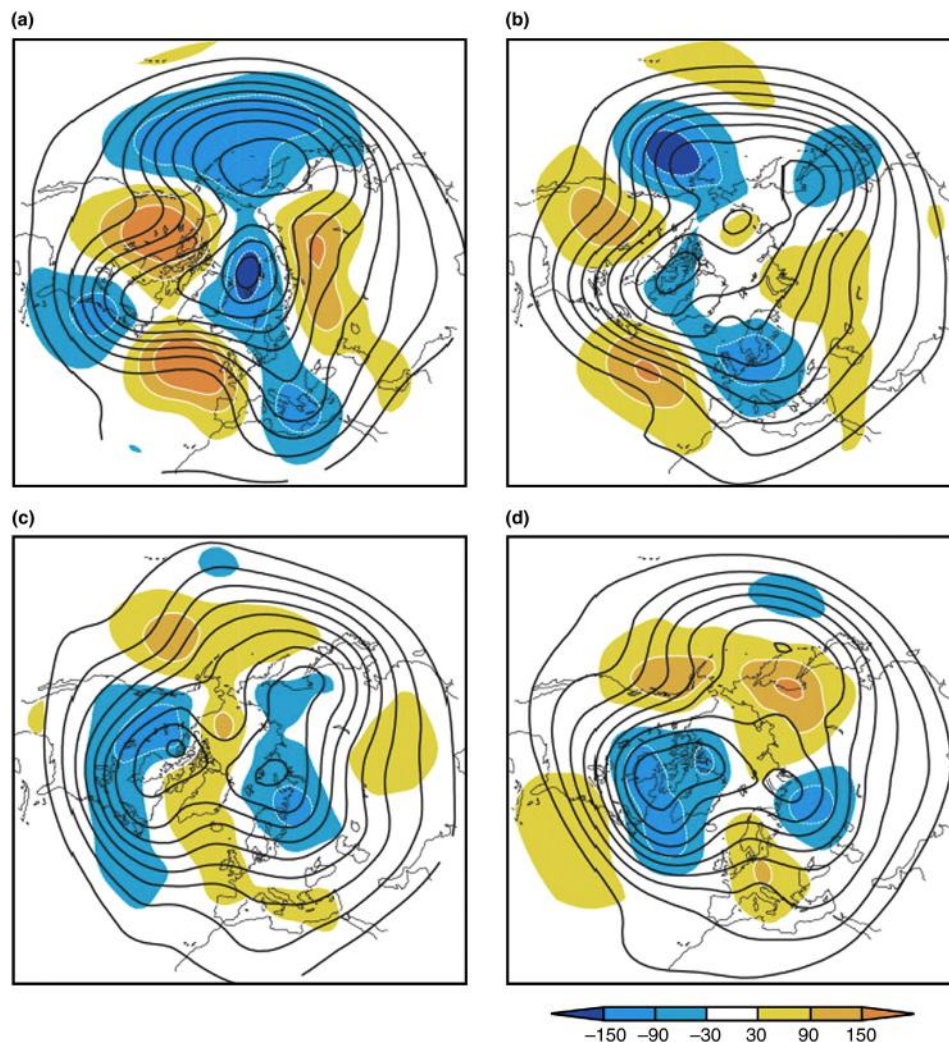


Figure 1 Monthly mean geopotential height at 500 hPa (black solid contours) and its deviation from the climatological mean (shaded) in (a) January 1981, (b) January 1986, (c) January 1982, and (d) January 1991. Contour interval 100 m for full fields, 60 m for anomalies.

linear theories of the interactions between the observed time-mean state of the atmosphere and large-scale circulation anomalies. Because of the zonal asymmetry of the climatological basic state, some large-scale patterns may be much more efficient than others in amplifying at the expense of the energy of the time-mean state, or may generate a much stronger response to anomalies in the distribution of diabatic heating.

Theories of a different kind are based on the nonlinear nature of the equations that govern the atmospheric motion. Because of their nonlinearity, these equations may possess a number of different stationary states. The so-called multiple equilibria of a number of highly simplified atmospheric models have therefore been proposed as an explanation of the preferred persistent anomalies of the atmospheric circulation.

In reality, no observed atmospheric state is exactly stationary: even though large-scale features may remain relatively constant, unstable synoptic disturbances always generate variability of the day-to-day weather. On the other hand, the intensity and geographical distribution of the weather events is strongly affected by the characteristics of the large-scale flow. Therefore, rather than looking at the equilibria of the instantaneous flow, it may be more appropriate to investigate the quasi-stationary states produced by the interactions of large-scale flow anomalies with the associated distribution of synoptic-scale disturbances. These statistical-dynamical equilibria, which involve interactions between phenomena with different spatial and temporal scales, are usually referred to as 'weather regimes' or 'flow regimes'.

The concept of multiple equilibria has been used to explain nonlinear phenomena in many branches of physical climatology and geophysical fluid dynamics. In oceanography, for example, it is widely accepted as an explanation of some aspects of the long-term variability of the thermohaline circulation. In atmospheric sciences, theories of multiple equilibria have always been more controversial, partly because of the highly simplified nature of the models proposed, partly because of the difficulty of finding observational evidence to support them. The concept of weather regimes, on the other hand, has been more widely accepted, since such regimes have been simulated in a number of intermediate-complexity models that produce a fairly realistic description of the atmospheric flow.

The summary of theoretical and observational studies on flow regimes given in the following sections will be focused on the northern extratropical circulation, following the majority of investigations on this issue. It should be mentioned that the existence of weather regimes in the Southern Hemisphere and in the tropical circulation has also been advocated. Whether a regimelike behavior may explain some aspects of the variability of the Asian summer monsoon is currently a much-debated issue. On the basis of results from simplified numerical models of the monsoon circulation, it has been suggested that the interannual variability arising from anomalies in tropical sea surface temperatures may be represented (particularly in the Indian region) by a modification of the frequency of two regimes corresponding to the so-called active phases and break phases of the monsoon. So far, the observational evidence supporting this interpretation of monsoon variability is scarce. However, the relationship between the statistical properties of regimes and variations in atmospheric forcing is

receiving attention also for the extratropical regions, since it may be a key to understanding the regional patterns of atmospheric response to the perturbations of the Earth's radiative balance arising from anthropogenic emissions of greenhouse gases.

Dynamical Models with Multiple Equilibria and Flow Regimes

The first (and simplest) atmospheric model that has been used to investigate the existence of multiple equilibria is represented by the barotropic vorticity equation forced by bottom topography. If ψ is the streamfunction at the equivalent barotropic level, and $\mathbf{V}_\psi$ given by eqn [1] is the associated nondivergent wind, the model can be written as eqn [2].

$$\mathbf{V}_\psi = \left(-\frac{\partial\psi}{\partial y}, \frac{\partial\psi}{\partial x} \right) \quad [1]$$

$$\begin{aligned} \frac{\partial \nabla^2 \psi}{\partial t} + \mathbf{V}_\psi \cdot \nabla \left[\nabla^2 \psi + f \left(1 + \frac{H}{H_0} \right) \right] \\ = -k \nabla^2 (\psi - \psi^*) \end{aligned} \quad [2]$$

where f is the Coriolis parameter, H/H_0 is the topographic height scaled by a suitable reference value, k is a dissipation coefficient, and ψ^* is an equilibrium state in the absence of topography.

A minimal version of this model (either in a so-called beta-channel or in a hemispheric domain) can be obtained by expressing ψ as a linear combination of just three orthogonal functions as in eqn [3].

$$\psi = u\Psi_0(y) + w_1\Psi_1(x, y) + w_2\Psi_2(x, y) \quad [3]$$

In eqn [3], Ψ_0 represents the zonal-mean flow; Ψ_1 and Ψ_2 represent the two components (in quadrature along the x -direction) of a planetary-scale Rossby wave that has near-zero phase speed when the zonal mean flow is close to its time-mean value. If the three functions Ψ_0 , Ψ_1 , and Ψ_2 are normalized in such a way as to have the same kinetic energy, the flow is relaxed toward a zonal-mean state $\psi^* = u^*\Psi_0$, and the topographic height is assumed to project only on Ψ_1 , then the model represented by eqn [2] is reduced (by a suitable scaling) to the three-variable dynamical system [4] and [6].

$$\dot{u} = -hw_2 - k(u - u^*) \quad [4]$$

$$\dot{w}_1 = (u - u_0)w_2 - kw_1 \quad [5]$$

$$\dot{w}_2 = -(u - u_0)w_1 + hu - kw_2 \quad [6]$$

In these equations, u_0 represents the value of u at which the planetary wave (Ψ_1, Ψ_2) is stationary, and h is a coefficient proportional to the topographic height. The nonlinear terms in eqns [5] and [6] arise from the absolute vorticity advection, while the terms proportional to h in eqn [4] and [6] represent the orographic drag on the zonal-mean flow and the orographic forcing of planetary waves, respectively.

For a given value of h , the system [4] and [6] possesses only one (stable) stationary solution when u^* is sufficiently small. Above a bifurcation value for u^* , three stationary solutions exist, one of which is unstable. Of the two stable equilibria, that

with the stronger zonal wind has a planetary wave pattern with a ridge approximately centered over the orographic maxima (i.e., $u > u_0$ and $w_1 > 0$). In the second stable state (with $u < u_0$ and $w_1 < 0$) the planetary wave pattern has a stronger amplitude, and is out of phase with respect to the topography. Examples of these stationary solutions are illustrated in Figure 2; the equilibrium with larger wave amplitude was originally interpreted as the counterpart of the atmospheric regimes characterized by blocking highs over the eastern sides of the oceans.

Because of the existence of stable steady states, the simple barotropic model above cannot simulate transitions from one regime to another. Trajectories started in different parts of the model phase-space will converge to either of the two ‘fixed points’ represented by the stable steady states. In the real extratropical flow, transitions from one regime to the other are due to a combination of barotropic and baroclinic instability, which can only be reproduced by multilevel models with wave components of different wavenumber.

A number of diagnostic studies, however, have revealed that high-frequency transients do not simply act as quasi-stochastic perturbations causing transitions between regimes. Instead, they interact with the large-scale flow in such a way as to modify the dynamical equilibria of the system. For example, in the case of blocking highs, it has been shown that the nonlinear feedback of transient eddies onto the large-scale anomaly tends to counterbalance the dissipative processes in the lower troposphere, and to oppose the downstream advection of the blocking pattern by the zonal-mean flow in the upper troposphere.

These processes may be simulated by a multilevel quasi-geostrophic model, described (at individual pressure levels) by an equation of the form [7].

$$\frac{\partial q}{\partial t} + \mathbf{V}_\psi \cdot \nabla q = -\mathbf{D}(\psi - \psi^*) \quad [7]$$

In eqn [7], q is the quasi-geostrophic potential vorticity (PV) given by eqn [8].

$$q = \nabla^2 \psi + f + f_0^2 \frac{\partial}{\partial p} \left(\frac{1}{\sigma} \frac{\partial \psi}{\partial p} \right) \quad [8]$$

σ is the appropriate static stability parameter, and $\mathbf{D}$ is a linear, multilevel operator representing dissipative processes that relax the model streamfunction toward ψ^* . A forcing term for planetary waves can be introduced in this model either by defining a zonally asymmetric ψ^* (for example, to simulate the thermal contrast between oceanic and continental regions during the Northern Hemisphere winter) or by introducing orographic forcing through an additional term in the definition of PV at the lowest model level (analogous to the orographic term in eqn [2]).

To introduce the concept of weather regimes, it is convenient to decompose the streamfunction field into a low-frequency component and a high-frequency component by introducing a time scale τ , equal to a typical lifetime of persistent large-scale anomalies. For the extratropical flow, τ is of the order of 10 days. If the overbar represents a (running) mean over time τ , the low-frequency component can be defined as $\bar{\psi}$, and the high-frequency part as $\psi' = \psi - \bar{\psi}$.

The time-averaged version of eqn [7] is given in eqn [9], where the last term on the right-hand side represents the nonlinear feedback of the high-frequency transients onto the low-frequency component.

$$\frac{\partial \bar{q}}{\partial t} + \bar{\mathbf{V}}_\psi \cdot \nabla \bar{q} = -\mathbf{D}(\bar{\psi} - \psi^*) - \overline{\mathbf{V}_{\psi'} \cdot \nabla q'} \quad [9]$$

This quantity should be regarded as a statistical–dynamical variable; its expected value (for any given $\bar{\psi}$) will be indicated as in eqn [10].

$$F'(\bar{\psi}) = -\langle \mathbf{V}_{\psi'} \cdot \nabla q' \rangle \quad [10]$$

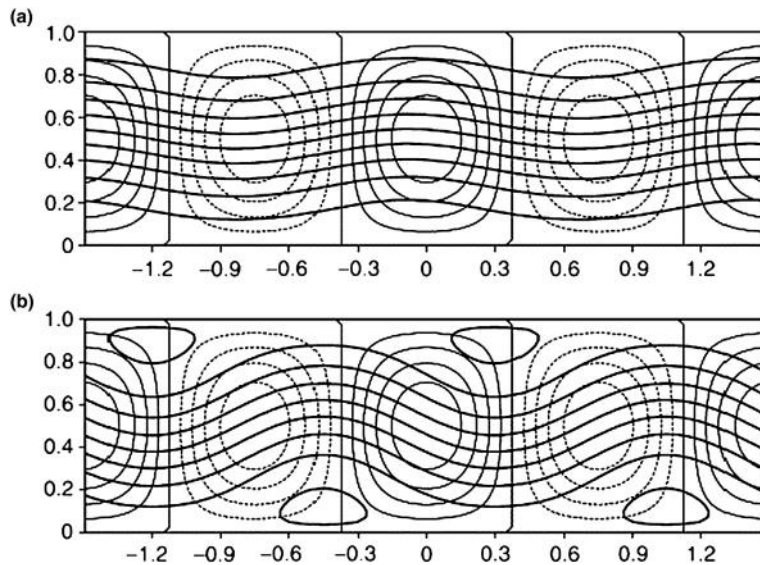


Figure 2 Streamfunction (thick solid contours) of the two stable stationary states of eqns [4] and [6], with $u_0 = 10$, $u^* = 12$, $h = 0.3$, $k = 0.6$. (a) State with $u > u_0$, (b) state with $u < u_0$. The topographic height is shown in thin contours (solid for positive values, dashed for negative values). Units are arbitrary.

Weather regimes can be defined as the solution of the statistical-dynamical equation [11].

$$\bar{\nabla}\psi \cdot \nabla \bar{q} = -D(\bar{\psi} - \psi^*) + F'(\bar{\psi}) \quad [11]$$

That is the weather regimes are those low-frequency patterns for which a balance between large-scale PV advection, dissipation, and feedback from high-frequency transients is achieved. As mentioned above, the high-frequency term is mainly balanced by large-scale advection in the upper troposphere and by dissipation in the lower troposphere.

Except for very idealized models, it is very difficult to write an analytical formulation of the term F' as a function of $\bar{\psi}$. Therefore, in studies performed with intermediate-complexity models, a semiempirical approach has often been used, in which first a very long integration of the original model equations is performed, then individual realizations of the high-frequency feedback are computed (each one corresponding to a $\bar{\psi}$ field in the model-generated time-series), and finally F' is defined as a weighted average over a neighborhood of $\bar{\psi}$ in a suitable subspace of the model phase-space. This methodology can also be adapted to the diagnostic analysis of observed data.

Another approach, applicable to the analysis of both modeled and observed data, is based on the search for multiple maxima (i.e., modes) in the probability density function (PDF) of the variables describing the low-frequency flow. If flow regimes do correspond to quasi-stationary states, then the large-scale flow patterns corresponding to them should be more persistent than other patterns, and this should be reflected in a higher probability density in phase-space.

An example of regimes of a three-level quasi-geostrophic model, identified by multiple maxima in the PDF of large-scale parameters, is given in Figure 3 and Figure 4. The model used for this integration has about 500 degrees of freedom for each level, covers a global domain, and represents the planetary wave forcing due to (real) topography and to vorticity sources generated by tropical convection. In models of this kind, it is difficult to select a priori the variables that are expected to display a multimodal behavior. Therefore, multivariate statistical analysis of the modeled variability is often used as a preliminary step to identify the most relevant parameters. Figure 3 shows the bidimensional PDF of the first two principal components (PCs) of streamfunction anomalies at 500 hPa, computed from 5-day means over the Pacific–North American region. The spatial patterns associated with these PCs (called empirical orthogonal functions, EOFs) are shown in Figure 4(a) and 4(b) and represent planetary wave anomalies with maximum amplitude in the north-eastern Pacific and over the western coast of North America. The PDF is bimodal, with one maximum along the PC-1 axis, and a second one along the PC-2 axis. Therefore, the two EOFs shown in Figure 4 and Figure 4 are also representative (in this particular case) of the anomalies associated with the two regimes of the model. In Figure 4 and Figure 4, the full 500 hPa streamfunctions corresponding to these regimes are shown. In the first regime, a predominantly zonal flow covers most of the North Pacific, and a ridge of moderate amplitude is centered on the western edge of the North American continent. In the second regime, the ridge has a stronger amplitude, its axis is shifted about 30° upstream, and an anticyclonic circulation prevails over the north-east Pacific.

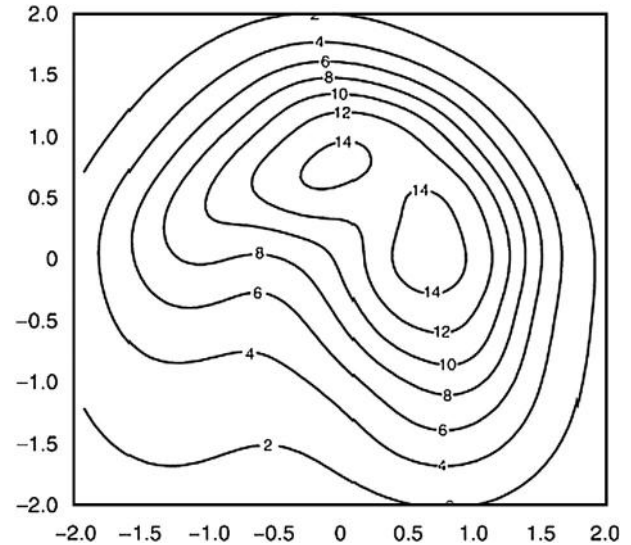


Figure 3 Probability density function (PDF) in the plane spanned by the first two principal components of 500 hPa streamfunction in the Pacific–North American region, from an integration of a three-level quasi-geostrophic model representing topographic forcing and tropical vorticity sources (PC-1, x-axis; PC-2, y-axis). Adapted from Molteni F and Corti S (1998) Long-term fluctuations of the statistical properties of low-frequency variability: Dynamical origin and predictability. Quarterly Journal of the Royal Meteorological Society 124: 495–526.

Are these regimes the counterparts of the multiple equilibria of highly truncated barotropic or baroclinic models, such as the three-variable system described by eqns [4] and [6]? The answer is not straightforward. Even when one considers the minimal two-level model in which interactions between planetary waves and baroclinically unstable synoptic-scale waves are represented, one does not find a one-to-one correspondence between stationary solutions and regimes. There are different reasons for this. First, in models with chaotic attractors, quasi-stationary regimes (if any) may only occur in the neighborhood of weakly unstable steady states, while no regimes are associated with strongly unstable equilibria. Second, baroclinically unstable eddies grow at the expense of the available potential energy (APE) of the time-mean flow, and therefore the mean effect of the term F' in eqn [11] is equivalent to a sink of APE. Since a reduced energy source will be available for direct energy conversions from the time-mean flow to large-scale anomalies, the planetary waves associated with flow regimes may have a smaller amplitude than those corresponding to the stationary states of the instantaneous flow for the same forcing parameters.

On the other hand, the flow-dependent component of F' may act as a kinetic energy source for the large-scale anomalies, maintaining them against dissipation. This is especially the case for dipolar structures resembling atmospheric blocking patterns. In fact, multiple regimes associated with the alternation of blocked and zonal flows, occurring downstream of a region with enhanced westerlies, have been simulated in quasi-geostrophic models possessing only one stationary solution for the instantaneous flow.

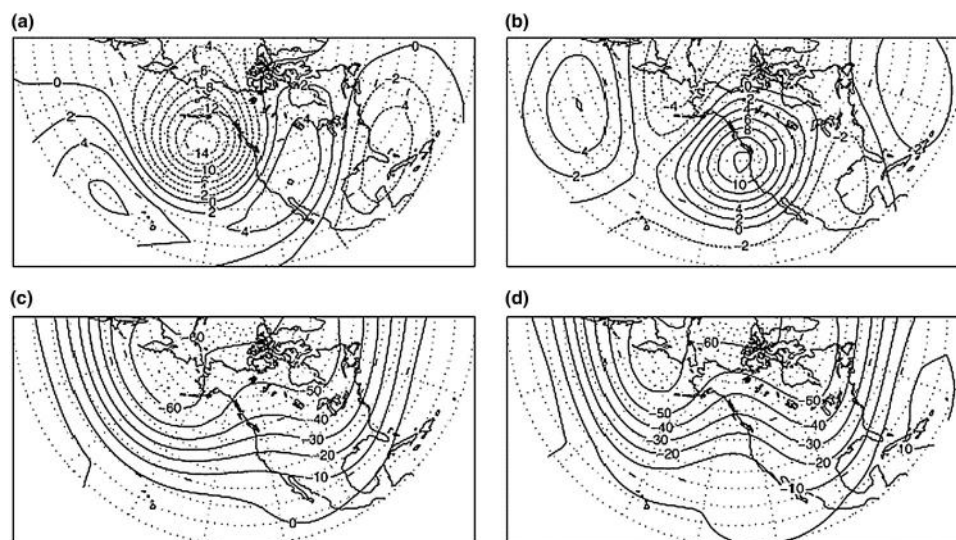


Figure 4 Empirical orthogonal functions (EOFs) (a, b) of 500 hPa streamfunction in the Pacific–North American region, and streamfunction patterns (c, d) corresponding to the two regimes (PDF maxima) in Figure 3, from an integration of a three-level quasi-geostrophic model representing topographic forcing and tropical vorticity sources. (a) EOF-1; (b) EOF-2; (c) streamfunction for regime with positive PC-1; (d) streamfunction for regime with positive PC-2. Units: $10^6 \text{ m}^2 \text{ s}^{-1}$. Adapted from Molteni F and Corti S (1998) Long-term fluctuations of the statistical properties of low-frequency variability: Dynamical origin and predictability. *Quarterly Journal of the Royal Meteorological Society* 124: 495–526.

Regimes in the Extratropical Atmosphere

A number of dynamical models of the atmospheric circulation that possess multiple steady states and/or regimes have been described in the scientific literature. At least for the extratropical flow in the Northern Hemisphere winter, a reasonably good understanding of the dynamical mechanisms that lead to the formation of flow regimes has been established. However, finding observational evidence of regime-like behavior in the real atmosphere has been a much more difficult task, often leading to controversial results.

In the early 1980s, attempts to detect a ‘signature’ of flow regimes in the atmospheric flow were guided by the results of theoretical studies on topographically forced multiple equilibria. Emphasis was therefore put on the analysis of indices of the intensity of zonal wind and of planetary wave amplitude in northern mid-latitudes. Although no evidence of multiple regimes was obtained from the analysis of zonal-mean wind (as justified by later theoretical and modeling studies), an index of the combined amplitude of planetary waves with zonal wavenumbers 2 to 4 showed a bimodal behavior when computed from a multiyear record of analyses of 500 hPa geopotential height. These findings, extended from a few years to about four decades of data in subsequent analyses, have been subject to a lot of critical scrutiny because of the sensitivity of the detected bimodality to the specific definition of the planetary wave index (namely, to the extent of the latitudinal band in which waves are analyzed and to the time-filtering procedure applied to the time series). Despite these uncertainties, the discovery of bimodality in the planetary wave index was quite a fundamental step in advancing the nonlinear theories of flow regimes to the role of reputable candidates for the explanation of extratropical low-frequency variability.

Part of the problems affecting the significance of the planetary wave index were related to the difficulty of condensing all information about the regime structure into a single index, defined a priori on the basis of theoretical considerations. Alternative approaches to the analysis of large-scale flow patterns were attempted in which multivariate statistical techniques were used to identify a suitable low-dimensional space in which the search for regimes could be performed effectively. The simplest of these approaches consisted in performing a PC analysis of observed planetary wave fields, and looking for bimodality in the combined amplitude of the leading PCs. In later studies, the leading PCs of anomalies of 500 hPa height were used to search for local maxima in multidimensional PDFs, or as input to cluster analysis techniques. In all cases, the goal was to reveal the ‘coarse grain’ structure of the attractor for the extratropical low-frequency circulation.

If on the one hand the search of regimes in multidimensional spaces removed some of the constraints associated with the definition of a unidimensional index, on the other it introduced greater difficulties in assessing the statistical significance of the results. Also, the arbitrariness in the choice of the domain is still present in multidimensional analyses. While some studies were concerned with regimes of the whole extratropical circulation (either in the Northern or in the Southern Hemisphere), it was argued that separate analyses of the regimes in the North Atlantic and North Pacific regions would provide more meaningful results because of the regional nature of the interactions between large-scale anomalies and high-frequency transients in the Atlantic and Pacific stormtracks.

Differences in methodologies, space–time domains, and dimensionality of the analyzed subspaces have produced regime classifications that obviously are not entirely compatible with one another. Still, consistency can be found in at least a subset of the regimes identified by different research groups.

From the analyses of the whole northern extratropical flow, a widespread consensus is found on the existence of two regimes that are well represented by the flow patterns in the top row and the bottom row, respectively, of Figure 1. When related to the results of linear teleconnection studies, these two regimes are characterized by opposite phases of the Pacific–North American teleconnection pattern. A third regime found in most Northern Hemisphere studies, and in analyses focused on the Atlantic sector, is associated with the negative phase of the North Atlantic Oscillation pattern and a general reduction in the strength of high-latitude westerlies. More examples of regimes defined in the Euro-Atlantic region will be given below.

Interannual and Interdecadal Variability of Regimes

One fundamental difficulty arising in the interpretation of observational analyses of flow regimes is due to the non-stationarity of the energy sources for the atmospheric flow. In theoretical and simplified numerical models of the large-scale flow, the statistical properties of the circulation depend on the amplitude and spatial distribution of the forcing terms acting as energy sources for the model. Usually, multiple regimes occur only in a certain range of the forcing parameters

and, within that range, regime properties such as the position in phase space and the frequency of occurrence depend on the parameter values.

In the real atmosphere, apart from the periodic fluctuations due to the seasonal cycle, the energy sources for the atmospheric flow are modified by interannual and interdecadal variations in the properties of the surface boundary (e.g., sea surface temperature, snow, and sea ice distribution) arising from the long-term natural variability of the whole climate system. In addition, the planetary radiative balance is being altered, in a slow but continuous way, by the changes in the concentration of radiatively active constituents produced by human activities. Therefore, even when the yearly cycle is accounted for by computing anomalies with respect to a seasonally varying climate, the multidecadal sample of observed anomalies in any given season is not homogeneous as far as the forcing terms are concerned.

The effect of these forcing variations on the regime properties are very difficult to distinguish from the natural fluctuations due to the chaotic nature of the atmospheric dynamics. Based on the theoretical behavior of nonlinear dynamical systems, one may hypothesize two different kinds of response. In one scenario, multiple regimes are well defined and the forcing variations are small compared to the range of forcing

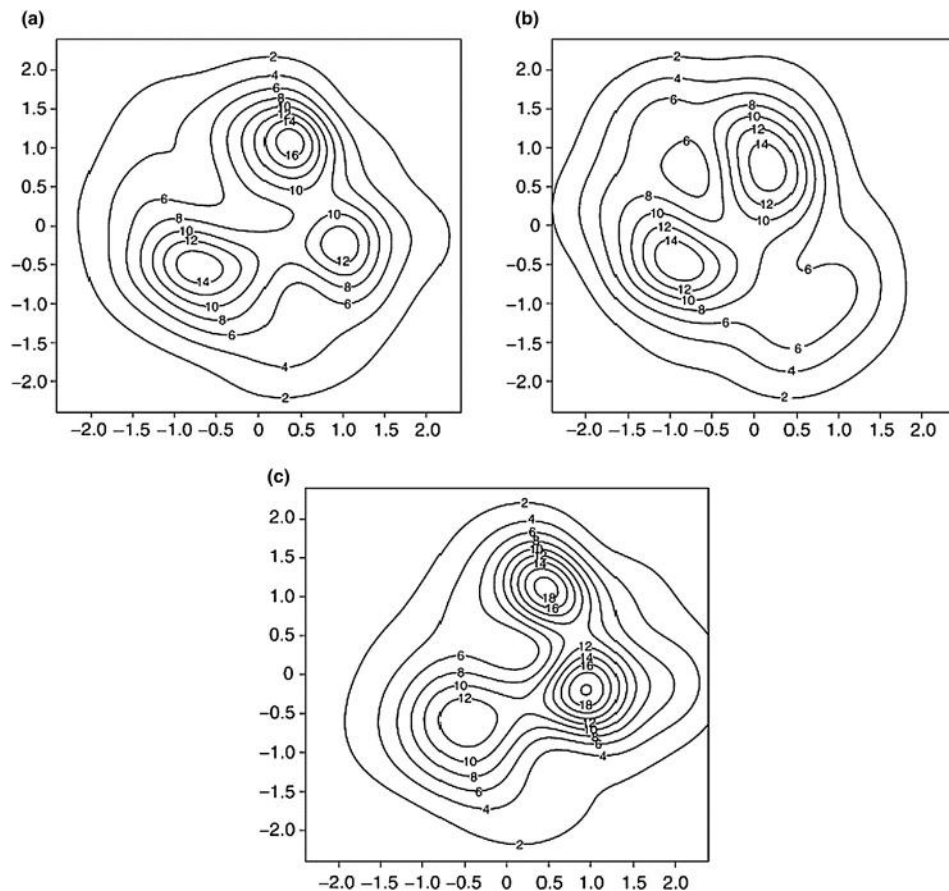


Figure 5 Probability density function (PDF) in the plane spanned by the first two principal components of wintertime (Dec–Mar) monthly mean 500 hPa geopotential height in the Euro-Atlantic region, computed from the reanalyses of the US National Centers for Environmental Prediction. (a) Data from 50 winters, 1948/49 to 1997/98; (b) data from 25 winters, 1948/49 to 1972/73; (c) data from 25 winters, 1973/74 to 1997/98.

parameters in which those regime exist. In this case, the circulation patterns associated to the regimes will not be strongly modified, but the stability of the regimes may be altered in such a way that the frequency of occurrence of one (or more) regimes will be increased, while other regimes will become less populated. Alternatively, if the system is close to a bifurcation point, or the forcing terms are changed substantially, then the patterns or even the number of regimes may be modified.

Which scenario is better suited to describe the observed atmospheric behavior? The answer may depend on the region considered. [Figure 5](#) shows the bidimensional PDF of the first two principal components of wintertime monthly mean anomalies of 500 hPa geopotential height in the Euro-Atlantic sector of the Northern Hemisphere (20° – 90° N, 90° W– 60° E). The first PC, corresponding to the x -axis in the figure, is an index of the North Atlantic Oscillation (NAO). [Figure 5\(a\)](#)

shows the PDF for the whole 50-winter record available, while [Figure 5\(b\)](#) and (c) show the PDF computed from the first half and the second half of the record, respectively.

The 50-winter PDF has three well-defined modes (i.e., local maxima), which are associated with the anomalies shown in [Figure 6](#). The regime corresponding to positive PC-2 (y -axis) shows a positive anomaly covering the central part of the North Atlantic, corresponding to an enhanced planetary wave ridge in this region. The other two modes, which occur at small negative values of PC-2, represent opposite phases of the NAO. In the negative phase, an anticyclonic anomaly develops over Greenland, which strongly reduces the westerly flow in the Atlantic poleward of 50° N. In the positive phase, the high-latitude westerlies are strengthened, and an anticyclonic anomaly develops over central Europe. These regimes have been found in a number of observational studies of Atlantic

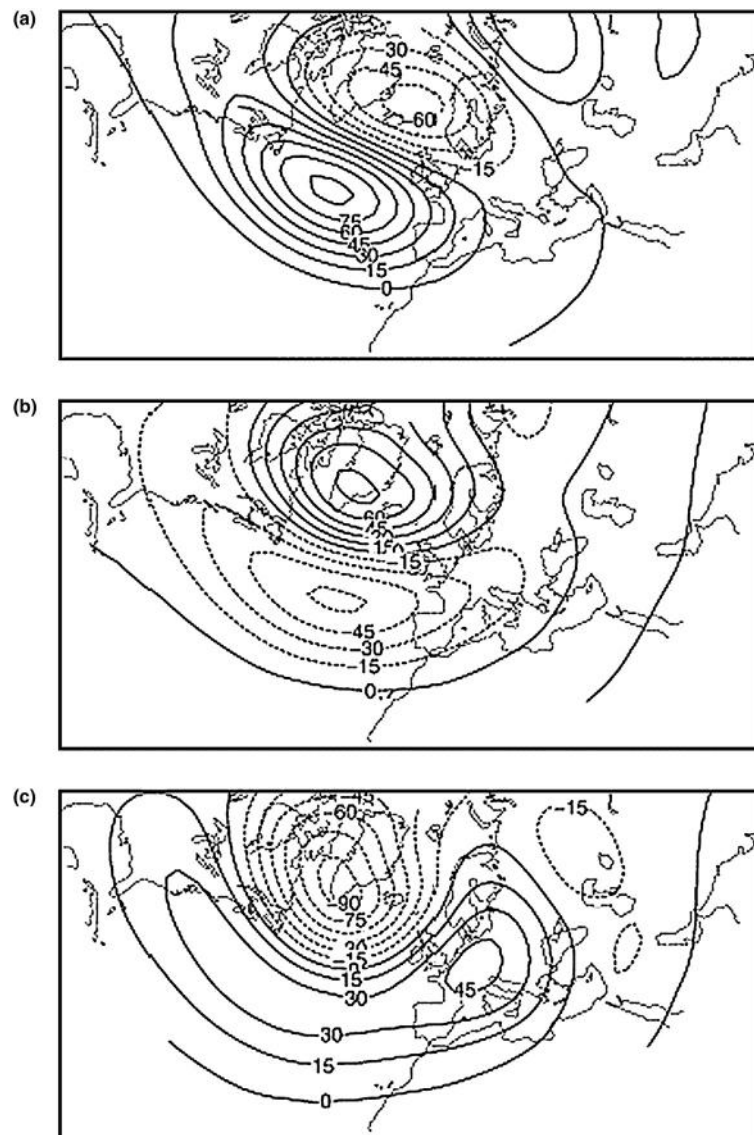


Figure 6 Anomalies of 500 hPa geopotential height (m) corresponding to the three regimes (PDF maxima) in [Figure 5](#). (a) Regime with positive PC-1 and positive PC-2 (Atlantic ridge); (b) regime with negative PC-1 and negative PC-2 (Greenland anticyclone); (c) regime with positive PC-1 and negative PC-2 (zonal flow with positive NAO phase).

variability (although more detailed analyses reveal at least one more regime, associated with a blocking high over Europe) and also correspond quite well to the Euro-Atlantic portion of the most reproducible hemispheric regimes.

If one looks at the PDFs for the two halves of the records, the first two modes may be clearly identified in both periods, although with different amplitude. Conversely, the mode corresponding to positive NAO is only evident in the second half of the record. This behavior is consistent with a widely documented trend of the NAO index in the last few decades. This trend is not simply reflected in a uniform shift of the PDF in the PC plane: although a partial shift is detectable, the most relevant effect is the appearance of the third mode corresponding to positive NAO. Overall, the regime change in the Atlantic region seems to fall into an intermediate category between the two theoretical scenarios outlined above.

One of the main differences between the properties of low-frequency variability in the Atlantic and Pacific sectors arises from the different importance of tropical–extratropical interactions in the two regions. While the interannual variability associated with tropical sea surface temperature anomalies accounts for a small fraction of the total variability in the Atlantic, the extratropical Pacific is strongly influenced by the El Niño Southern Oscillation (ENSO) phenomenon. Since the east–west asymmetry in the distribution of diabatic heating in the tropical Pacific is an important source of forcing for the planetary waves in the region, the changes in convective activity induced by ENSO cannot be regarded as weak anomalies in the forcing patterns. A number of observational and modeling studies have found that the intensity of low-frequency variability in the North Pacific region is enhanced in the cold ENSO phase (La Niña) and reduced during warm (El Niño) events. In fact, there is little evidence of multiple regimes in the Pacific during strong El Niño events.

See also: **Climate and Climate Change:** Climate Variability: North Atlantic and Arctic Oscillation; Climate Variability: Seasonal and Interannual Variability. **Dynamical Meteorology:** Balanced Flow; Quasigeostrophic Theory; Rossby Waves;

Stationary Waves (Orographic and Thermally Forced); Wave Mean-Flow Interaction. **General Circulation of the Atmosphere:** Overview; Teleconnections. **Global Change:** Upper Atmospheric Change. **Statistical Methods:** Data Analysis: Empirical Orthogonal Functions and Singular Vectors. **Tropical Meteorology and Climate:** El Niño and the Southern Oscillation: Observation.

Further Reading

- Benzi, R., Saltzman, B., Wiin-Nielsen, A.C. (Eds.), 1986. *Anomalous Atmospheric Flows and Blocking*. Academic Press, Orlando.
- Charney, J.G., DeVore, J.G., 1979. Multiple flow equilibria and blocking. *Journal of the Atmospheric Sciences* 36, 1205–1216.
- Corti, S., Molteni, F., Palmer, T.N., 1999. Signature of recent climate change in frequencies of natural atmospheric circulation regimes. *Nature* 398, 799–802.
- Ghil, M., 1987. Dynamics, statistics and predictability of planetary flow regimes. In: Nicolis, C., Nicolis, G. (Eds.), *Irreversible Phenomena and Dynamical Systems Analysis in Geophysics*. Reidel, Dordrecht, pp. 241–283.
- Hansen, A.R., Sutera, A., 1986. On the probability density distribution of large-scale atmospheric wave amplitude. *Journal of the Atmospheric Sciences* 43, 3250–3265.
- Held, M., 1983. Stationary and quasi-stationary eddies in the extratropical troposphere: Theory. In: Hoskins, B.J., Pierce, R.P. (Eds.), *Large-Scale Dynamical Processes in the Atmosphere*. Academic Press, London, pp. 127–168.
- Reinhold, B., Pierrehumbert, R.T., 1982. Dynamics of weather regimes: Quasi-stationary waves and blocking. *Monthly Weather Review* 110, 1105–1145.
- Vautard, R., Legras, B., 1988. On the source of midlatitude low-frequency variability. 2. Nonlinear equilibration of weather regimes. *Journal of the Atmospheric Sciences* 45, 2845–2867.
- Wallace, J.M., 1996. Observed climatic variability: Spatial structure. In: Anderson, D.L.T., Willebrand, J. (Eds.), *Decadal Climate Variability: Dynamics and Predictability*. Springer-Verlag, Berlin, pp. 31–81.
- Wilks, D.S., 1995. *Statistical Methods in the Atmospheric Sciences: An Introduction*. Academic Press, London.

Mean Characteristics

R Grotjahn, University of California, Davis, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

The large-scale average conditions of the atmosphere are described. Emphasis is placed on the primary observed variables: radiation, temperature, pressure, wind, clouds, and precipitation rate. Zonal and time average fields as well as time average fields at representative levels are shown.

Introduction

The atmosphere of the Earth has a diverse range of motions. The general circulation refers to the larger scale motions having horizontal length scales greater than 1000 km and persisting for a season or longer. In addition, this subject includes all processes necessary to explain sufficiently, or maintain directly, the large-scale circulation.

The general circulation is a broader subject than simply the large-scale movement of air. Understanding the circulation requires examination of other atmospheric quantities. The large-scale circulations are created by imbalances in the radiation fields that lead to temperature gradients that the atmosphere tries to eliminate. The circulations that develop are limited by various physical constraints such as radiative energy balance, mass balance, and angular momentum balance. So, while the primary scope of this article is to display the average properties of circulation, it is also necessary to consider related variables that are directly observed. The related variables are connected by constraints and the underlying laws of dynamics and thermodynamics. Other articles on the general circulation discuss how the circulation is maintained and how it can be simulated.

The general circulation of the atmosphere has complex structure in all dimensions as it evolves over the seasons. However, the Earth rotates fast enough relative to the radiative response time of much of the atmosphere so that many properties of the atmosphere have a strong zonal average component. To make a discussion of this subject manageable, the properties of a variable are first shown when longitudinal averages are taken in addition to seasonal time averages. Longitude averages are commonly labeled 'zonal means'. Zonal averages do miss important phenomena that can be seen in time averages. Time means reveal longitudinal variations that one must consider in order to understand the properties as well as the maintenance of the zonal mean state. The three-dimensional structure of time averages is illustrated by showing horizontal maps of time-averaged fields at a few representative levels.

The general circulation undergoes seasonal change. In many fields, the seasonal change is much less in the Southern Hemisphere than in the Northern Hemisphere. The difference arises because the middle latitudes in the Southern Hemisphere have a much higher fraction of ocean coverage than in the Northern Hemisphere. Land and ocean have different thermodynamic properties: heating and cooling are more readily mixed through a greater amount of mass in the ocean than on the land; the albedo of land can change drastically with season unlike ice-free ocean, and water has a higher heat capacity

than soil. These factors magnify seasonal change over large landmasses.

The difference in middle latitude land coverage has another implication. There are more major mountain ranges in the Northern Hemisphere. The mountains, along with differences that arise between land and sea areas, lead to more prominent planetary waves in the Northern Hemisphere. Alternatively, many fields in the Southern Hemisphere tend to be more zonally uniform.

Radiation and Temperature

Radiation

A discussion of the general circulation starts with radiation, since the distribution of radiation absorbed and emitted is the ultimate driving mechanism for the circulation.

The Earth (including its atmosphere, solid, and ocean parts) absorbs radiation emitted by the Sun. While the Earth rotates about an axis that is tilted with respect to the Sun, the equator is perpendicular to the Sun's rays on an annual average. Simple geometry (see [Figure 1](#)) shows that the solar radiation reaching a unit horizontal area on the Earth diminishes from equator to pole. The amount of solar radiation absorbed is influenced by the reflectivity of the Earth's surface, the amount of cloud cover, and the path length through the atmosphere. On an annual average, the amount of radiation absorbed decreases greatly from the equator toward each pole ([Figure 1](#)). The rate that absorption decreases with latitude is greater in polar than in tropical regions.

To balance the radiant energy absorbed, the Earth emits energy back to space. Like the absorbed solar radiation, terrestrial emission also decreases from the equator toward each pole on an annual average. The terrestrial emission is governed by the temperature and the radiative properties of the emitter. Emission comes from the surface of the Earth as well as from the atmosphere. For equivalent radiative substances, more radiant energy is emitted from objects having higher temperature. So, the terrestrial emission shown in [Figure 1](#) is consistent with the tropics being warmer than the polar regions. A key fact is that the emission does not change with latitude nearly as fast as does the absorption. Consequently, most latitudes are not in radiative equilibrium and there is a net radiation surplus in the tropics (more solar radiant energy is absorbed than terrestrial radiation is emitted). At latitudes higher than 38°, there is a net radiation deficit (more emission than absorption locally). In order to sustain the surplus and deficit over time, a poleward energy transport is required, and that has implications for the temperature field. The terrestrial

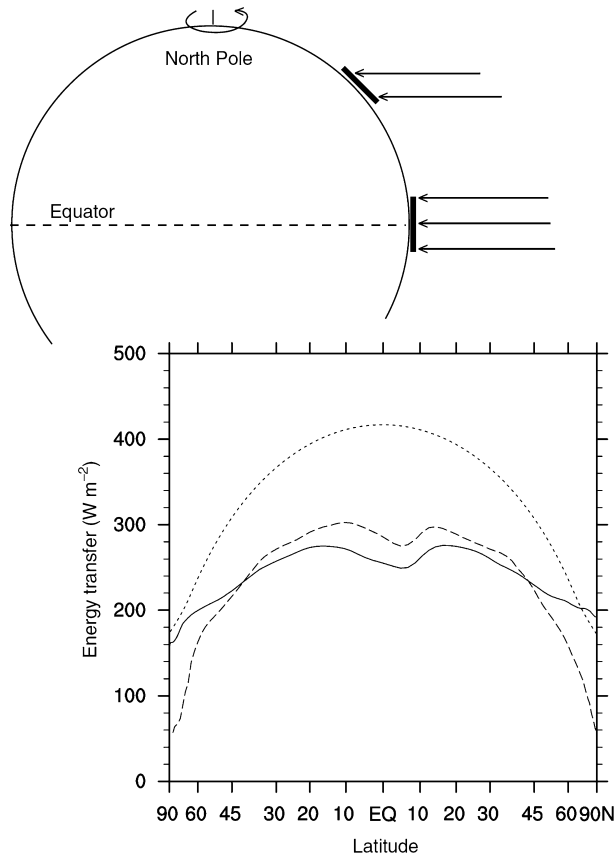


Figure 1 Schematic diagram of solar radiation reaching two horizontal areas on the Earth at an equinox. Inset: annual average latitudinal distribution of incoming solar (dotted line), absorbed solar (short dashed line), and terrestrial emission (solid line) in watts per square meter. A stretched latitude is used based on the area within each latitude band. Data from National Center for Environmental Prediction (NCEP)/Department of Energy (DOE) AMIP Reanalysis II four times daily mean data averaged from January 1979 through December 2009.

emission means that temperatures are lower in the tropics and warmer in the polar regions than they would be if there was no heat transport. While some transport occurs by the oceans, the remainder occurs in the atmosphere and to transport thermal energy the atmosphere must have circulation. The atmosphere may transport the heat by direct means (labeled sensible heat transport) or by transporting water vapor (labeled latent heat transport). The latter process gains or releases heat during a phase change of the water. So, the distribution of radiation implies links to temperature, velocity, and moisture fields.

The solar absorption and terrestrial emission shown in **Figure 1** are plotted on a stretched latitudinal scale. The stretching is proportional to the horizontal area of each latitude band. Consequently, the annual average energy balance for the Earth as a whole is also seen in the figure because the area under the dashed line equals the area under the solid line. More is said about the energy distribution and flow elsewhere in the encyclopedia (see **General Circulation of the Atmosphere: Energy Cycle**).

While **Figure 1** shows the zonal average distribution of absorbed and emitted radiations, the longitudinal distribution of these two quantities is shown in **Figure 2**. The annual

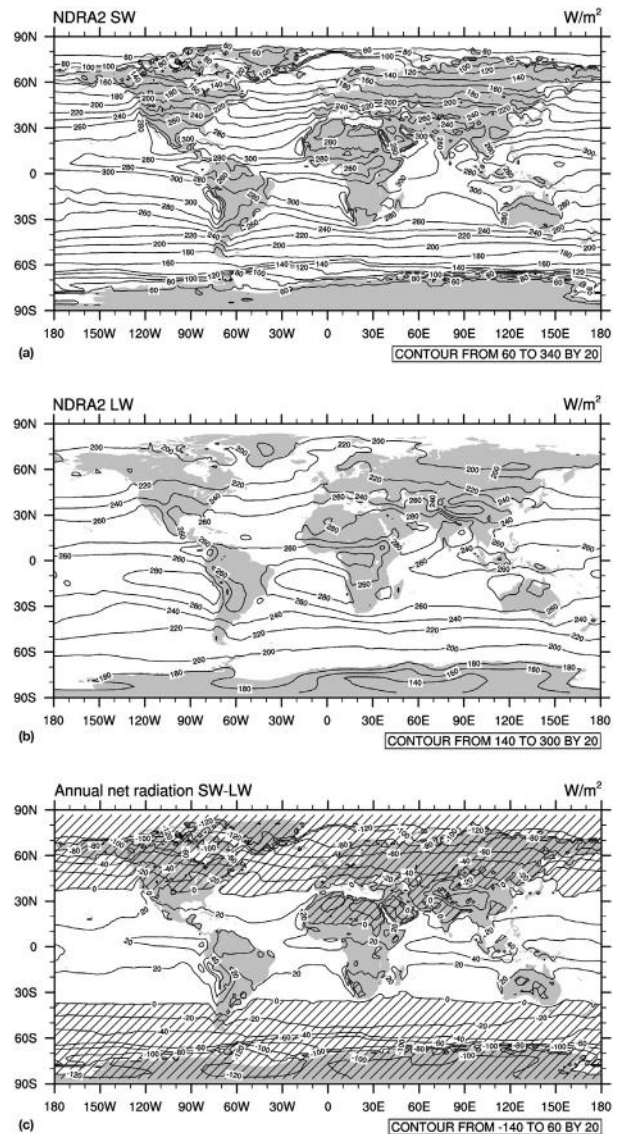


Figure 2 Annual radiative balance for the Earth. (a) Time average incoming, downward, solar, and absorbed radiation estimated at the top of the atmosphere. (b) Time average upwelling, longwave radiation from the Earth. (c) Net radiation, the solar minus terrestrial radiation difference. NCEP/DOE AMIP Reanalysis II four times daily mean data averaged from January 1979 through December 2009. Contour interval is 20 W m^{-2} .

average solar radiation absorbed (**Figure 2(a)**) has a strong zonal mean component over the oceans. The principal deviations from the zonal average are over ice-covered regions (Greenland) and desert areas (Sahara and Arabia). Both areas have high annual average reflectance due to a bright surface. Variation across each ocean basin occurs with greater cloudiness along the storm tracks (discussed below) such that those tracks avoid the eastern sides and allow more solar absorption on the east side. The dips in the zonal average radiant energies near 10° N latitude (**Figure 1**) are a consequence of persistent high-elevation cloudiness associated with the intertropical convergence zone (ICZ). The ICZ is seen in **Figure 2(a)** as most of the lower values in the tropics that occur over the

major landmasses and along a narrower band across the oceans. The annual average terrestrial radiant flux at the top of the atmosphere (**Figure 2(b)**) also has a strong zonal mean component over oceans. Longitudinal variations occur over the eastern subtropical oceans, deserts, and higher topography. The ICZ is visible as lower values of terrestrial radiant fluxes in the tropics because the emission is primarily from the cold tops of high clouds. The same clouds strongly reflect solar radiation causing a corresponding solar radiation pattern. The net radiation (**Figure 2(c)**) has a strong latitudinal trend as deduced from **Figure 1**. The primary zonal variations are over the subtropical deserts. These areas are hot (high thermal emission) but reflective (reduced solar absorption) leading to negative net radiation. For thermal balance, energy must be exported from areas of positive to areas of negative net radiation; so, the subtropical deserts maintain their high heat in part because heat energy converges into these deserts. Also, the dry desert surface means the solar energy absorbed mainly heats the surface and overlying air but evaporates little water.

Temperature

In the troposphere and lower stratosphere, absorption and emission of radiation alter the air temperature at a rate of a few degrees per day. This change of air temperature is generally small compared to the difference between the equatorial regions and the polar regions. Hence, the radiant energy absorption and emission do not create large temperature differences between the daylight and night sides of the Earth. Therefore, much about the atmosphere's thermal structure is seen in a zonal mean.

Seasonal averages for winter and summer temperatures are displayed in **Figure 3**. The following properties are evident in the figure:

1. Temperature decreases with increasing height in the troposphere and winter stratosphere. The rate of change with height (equaling minus the lapse rate) is greater in the troposphere and notably less in the stratosphere. The lapse rate is less than the lapse rate for neutral stability; so, the atmosphere is statically stable on the large scale. Temperature increases with height in the tropical lower stratosphere.
2. The tropopause marks the boundary between the stratosphere and troposphere. The zonal mean tropopause is not level, but ranges from 6–8 km (~ 500 – 350 hPa) in polar regions to 16–18 km (~ 100 – 80 hPa) in the tropics.
3. In the troposphere, zonal mean temperature along an isobaric surface decreases toward each pole. The rate of decrease (the meridional gradient) is small in the tropics. This gradient is larger in the middle latitudes (30 – 60° latitude). An exception is the sharp gradient near the Antarctic coast. The gradient is stronger during winter. The temperature change from equator to pole is also greater at the surface than in the middle troposphere.
4. The coldest temperatures are in the winter stratospheric polar region and near the equatorial tropopause.
5. In much of the lower stratosphere, the zonal mean temperature gradient is reversed from the troposphere. The

reversal is evident in much of the tropics and middle latitudes. The main exceptions are the higher latitudes and part of the Southern Hemisphere middle latitudes during winter.

The longitudinal distribution of the temperature illustrated by the 850 hPa level temperatures is shown in **Figure 4**. As anticipated from the terrestrial radiation data, the field has notable zonal symmetry over the oceans. The colder temperatures in middle latitudes tend to be over the eastern sides of the continents during winter. Large ice-covered regions have colder temperatures. During summer, the hottest temperatures are located over the subtropical deserts.

Mass Fields

Geopotential Height Fields

The temperature field is related to the mass fields in several ways. For example, the hypsometric equation demonstrates that the spacing between isobaric surfaces is proportional to the mean virtual temperature of the air between these surfaces. In the troposphere, the vertical distance between any two isobaric surfaces (the 'thickness') is larger in the tropics than outside the tropics. Since the variation of sea level pressure (SLP) over the globe is rather small, this implies that the altitude of an upper air constant pressure surface (500 hPa, e.g.) on average increases from pole to equator.

The zonal and time mean geopotential heights of the 500 and 1000 hPa geopotential height surfaces (Z_{500} and Z_{1000} , respectively) are shown in **Figure 5** for the two extreme seasons. The Z_{500} surface is representative of geopotential height surfaces in much of the troposphere and lower stratosphere. At 500 hPa, higher heights are found in the tropical regions with lowest values near the poles. The gradient between equator and pole is strongest in the middle latitudes and during winter. The Z_{1000} pattern is representative of the mass field near the surface. The zonal average Z_{1000} has lower values near the equator, with higher values in the subtropics (more so in local winter). The lower values in middle latitudes are associated with the midlatitude storm tracks.

Midlatitude Planetary Waves and Storm Tracks

The mass and temperature fields have significant longitudinal variation. Away from the surface, the time mean pattern is characterized by long waves. The time mean Z_{500} (**Figure 6**) has patterns typical of much of the troposphere. Prominent troughs are seen near the midlatitude east coasts of Asia and North America. The temperature field beneath (e.g., **Figure 4**) also has colder values (thermal troughs) near these regions. A weaker geopotential trough occurs over the Mediterranean and Northwest Africa. At the base of each trough the height contours are more closely spaced; from geostrophic wind balance one expects the wind speeds to be relatively stronger there. Ridges are found in between, most prominently in northwestern North America and Europe. In summer, the North American trough remains visible due to the cold air over the Baffin and Greenland ice sheets.

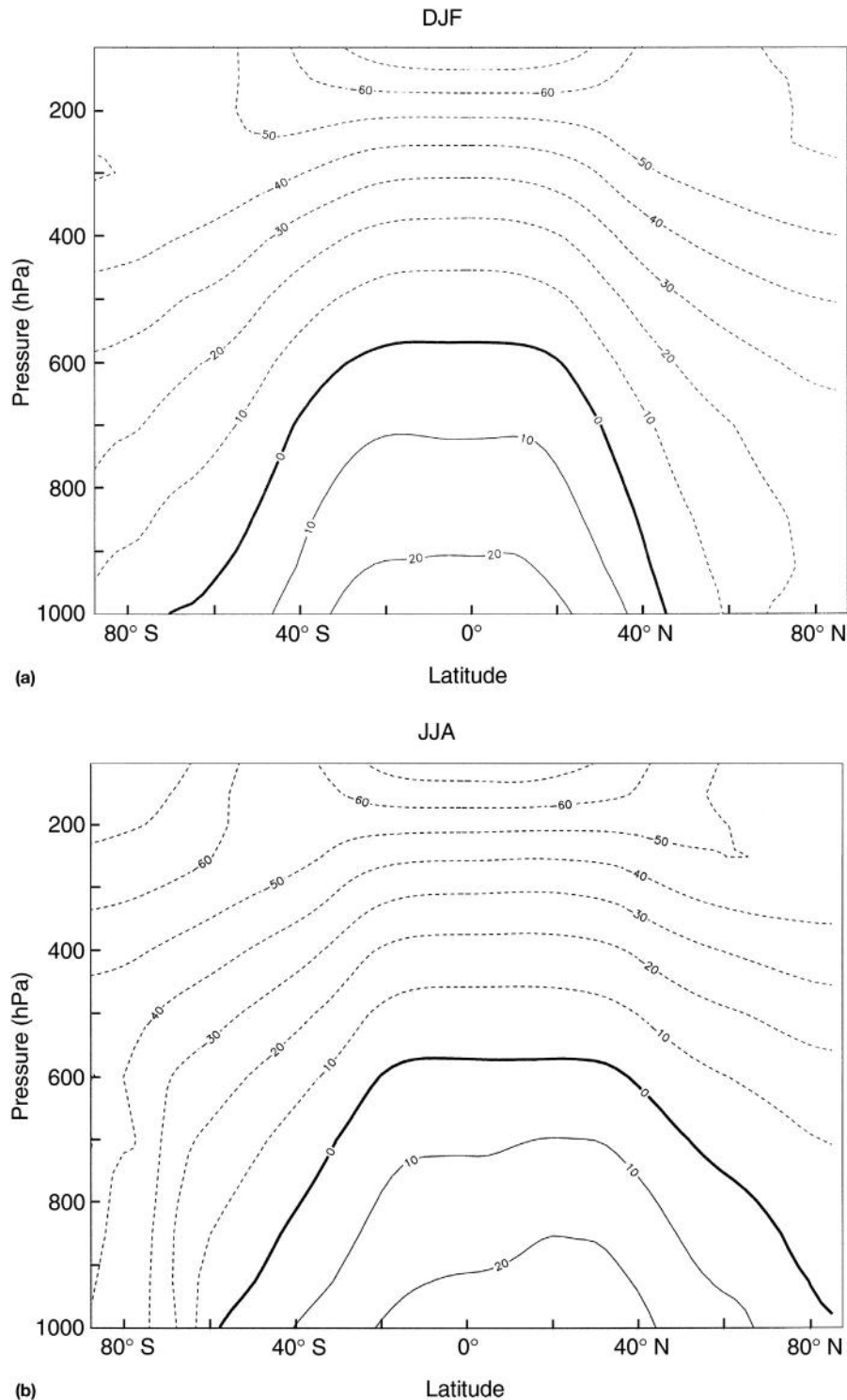


Figure 3 Zonal mean temperature during (a) December–February and (b) June–August. Units are degrees Celsius. The data are from the NCEP/National Center for Atmospheric Research (NCAR) reanalysis period 1979–99. Contour interval is 10 °C.

Consistent with the zonal average distribution (Figure 5) the time average gradient is strongest in middle latitudes. While the gradient has similar strength during winter, during summer the time mean gradient is stronger in the Southern Hemisphere.

Most of the weather in the middle latitudes is created by the traveling frontal cyclones, otherwise known as extratropical cyclones. These cyclones interact with the planetary wave pattern in various ways. These cyclones prefer to form, propagate, and decay in specific regions. Generally speaking, frontal

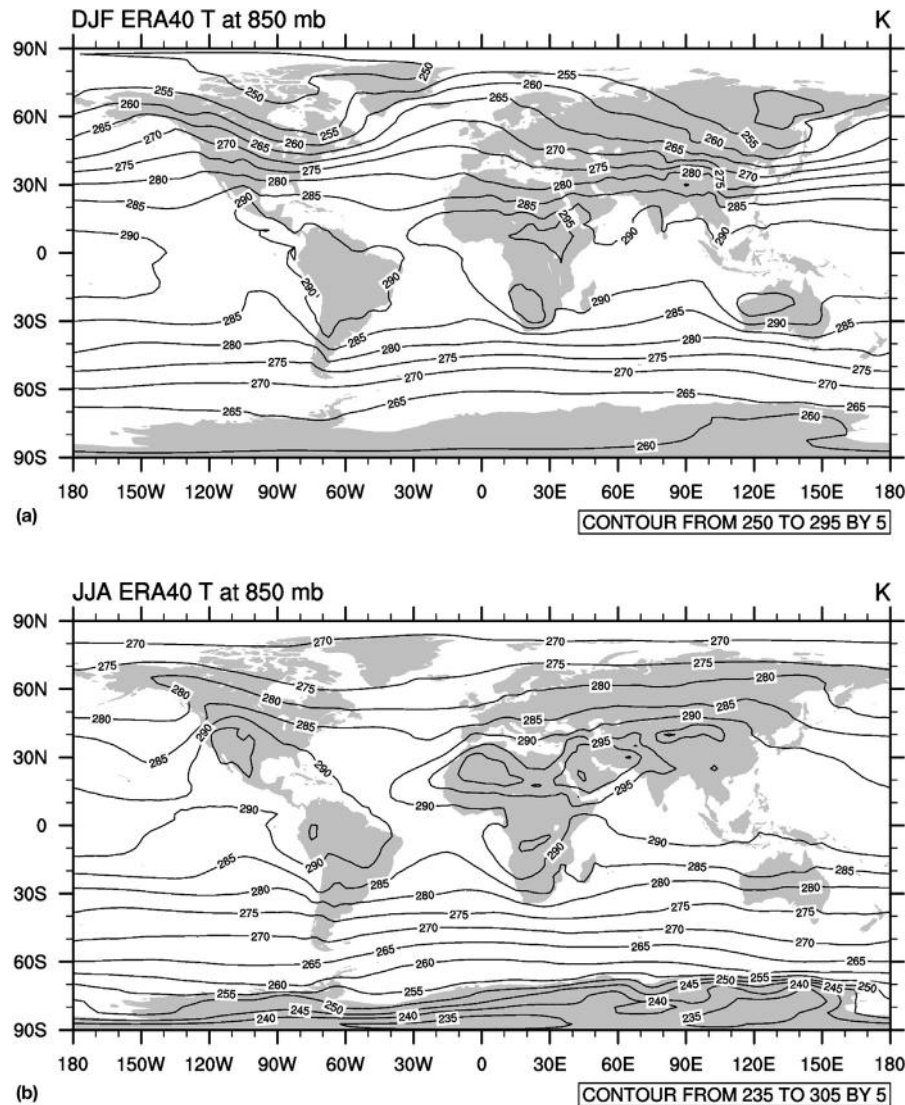


Figure 4 Temperature at 850 hPa in the two extreme seasons (a) December–February and (b) June–August. Units are degrees Kelvin. The data are from the European Centre for Medium-Range Weather Forecasts ERA-40 reanalysis monthly averaged data from 1979 to 2002. Contour interval is 5 K.

cyclogenesis is favored in three types of regions: near the east coasts of continents, on the lee side of major mountain ranges, and where large-scale surface temperature gradients are strong. Most of these regions coincide with the longwave trough locations. A region of stronger sea surface temperature gradient lies south of Africa and across much of the Southern Indian Ocean. Further south, the near surface temperature gradient is intensified at the edge of Antarctic sea ice. The frontal cyclones generally progress eastward and poleward as they evolve. Figure 6 also shows various archetypal tracks followed by many frontal cyclones. Individual tracks of cyclones vary, but the thicker dashed arrows in Figure 6 indicate the more common paths. In the Northern Hemisphere, cyclones often merge with or supplant the ‘semipermanent’ Aleutian and Icelandic lows that are visible in the SLP field discussed next. The storm tracks show up in the precipitation fields (shown later) as bands of heavier precipitation in the middle latitudes across the oceans. Precipitation also is enhanced where

westerlies encounter mountain ranges of North America and Western Europe.

Sea Level Pressure and Subtropical Highs

The SLP pattern (Figure 7) differs from the mass field pattern at mid and upper troposphere levels (e.g., the Z_{500} pattern of Figure 6). An obvious difference is that the field is noisy over major topographic features because it is based on extrapolating the pressure to sea level. Another difference is the meridional height gradient (created by the meridional temperature gradient through a depth of the atmosphere), which is not so evident at the surface. So, the SLP pattern is more cellular. Another difference is that colder air aloft can result in relatively higher surface pressure near the ground (due to higher average density of the air above compared to an adjacent region). The converse is also true for lower pressure where the air aloft is warmer than the surroundings (so-called

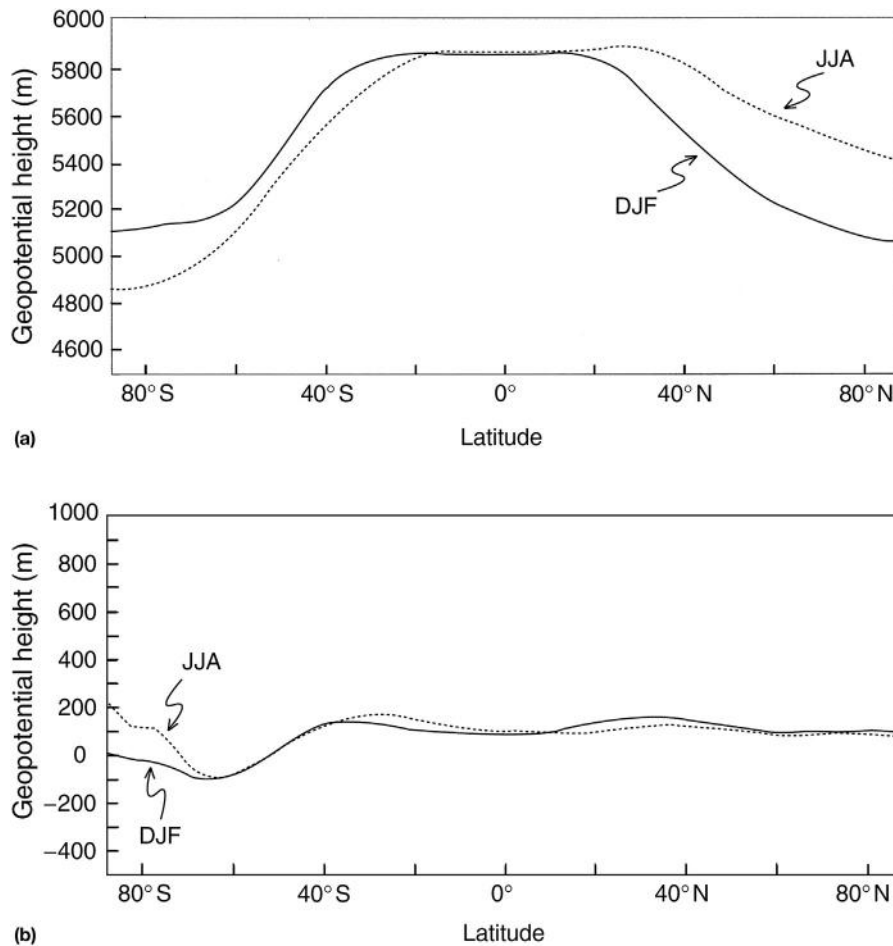


Figure 5 Zonal mean geopotential heights (a) at 500 hPa and (b) at 1000 hPa. Solid line is for December–February seasonal average and dotted line is for June–August. These NCEP/NCAR reanalysis data are from 1979 to 1999.

‘thermal lows’). This relationship is the opposite to that described above for upper air geopotential height because the hypsometric equation relates to the thickness of a layer, not the bottom of a layer.

At the surface one finds prominent high-pressure cells in the subtropics. These cells generally prefer the eastern sides of the ocean basins where thermal and other forcings build higher pressure. For example, equatorward winds on the east side of the subtropical high can drive upwelling of colder subsurface ocean water and the resultant cooler surface is observed to lead strengthening of the high (a result also consistent with potential vorticity reasoning). The colder temperatures foster low stratus clouds, which in turn create a net radiative cooling of the air and additional forcing for the high. The areal extent of the highs is influenced by other factors such as the midlatitude storm tracks and tropical convection. In summer, these factors allow much greater expansion of the subtropical highs in the Northern Hemisphere than in the Southern Hemisphere. A three-way balance between the pressure gradient, Coriolis, and turbulent drag forces implies divergence at a surface high. Surface divergence is consistent with the sinking, and eastern sides of the subtropical highs can be viewed as regions where the

sinking branch of a Hadley cell is enhanced. The sinking is apparent as areas where upper-level divergent winds converge (shown below). The tropical convection, mainly equatorward and to the west of the subtropical high, feeds the circulation supporting the high; this link is visible in the divergent winds.

Wind Fields

Zonal Velocity

The zonal wind is directed positive when blowing toward the east. Outside the equatorial region, the mass and wind fields are in approximate geostrophic balance. The meridional gradient seen in the tropospheric geopotential height fields implies a westerly wind which is stronger in middle latitudes. At the surface, comparatively weak winds are expected.

The thermal wind relation states that vertical *shear* of the zonal wind is proportional to the meridional gradient of temperature and inversely proportional to the Coriolis parameter. The temperature gradient in much of the troposphere is directed equatorward, an orientation implying westerly wind shear. Since the tropospheric temperature gradient is

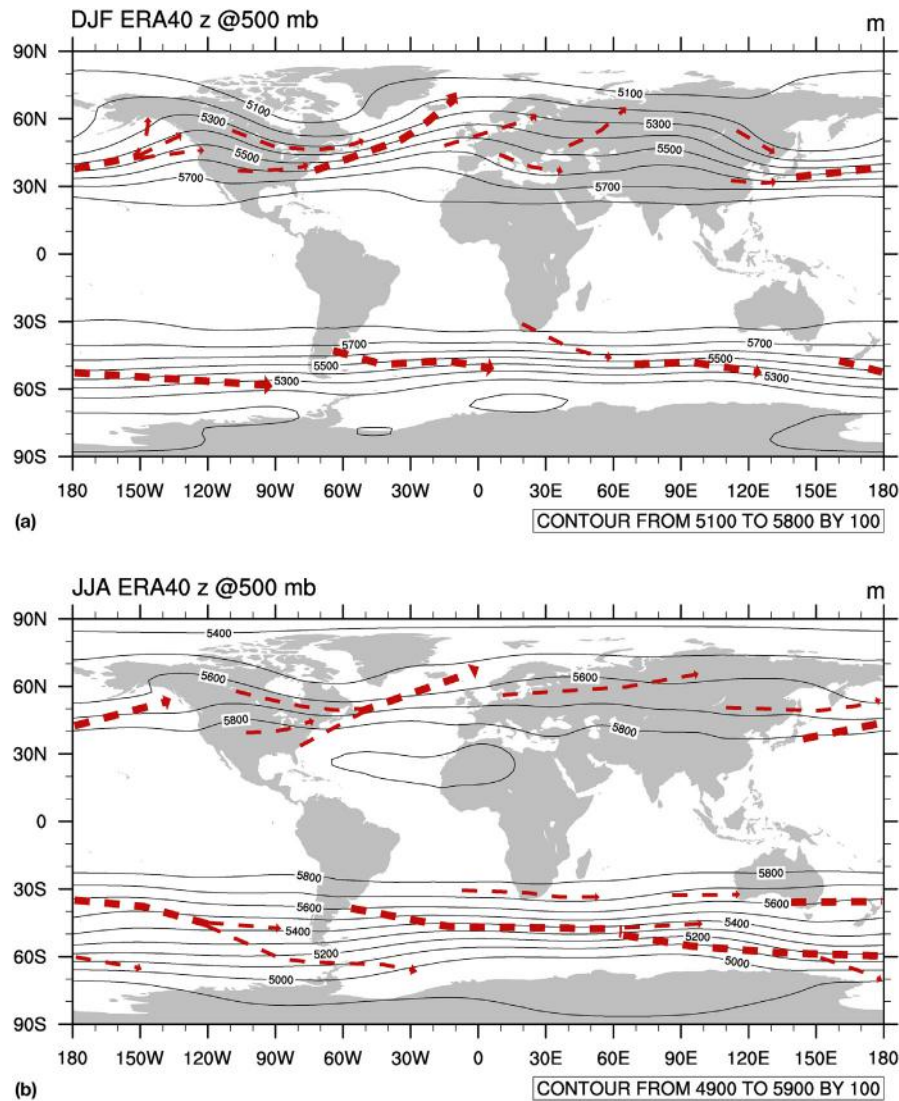


Figure 6 Geopotential height of the 500 hPa surface (contours) in the two extreme seasons (a) December–February and (b) June–August. Units are in meters. The data are from the European Centre for Medium-Range Weather Forecasts ERA-40 reanalysis averaged data from 1979 to 2000. Also plotted are general indicators of the frontal cyclone storm tracks (dashed arrows) where a wider arrow indicates more frontal cyclone passages than for a thinner arrow. Frontal cyclone storm tracks shown merge information from these sources:

van Loon (1966) On the annual temperature range over the Southern Oceans. *Geographical Review* 56: 497–515.
 Whitaker and Horn (1984) Northern Hemisphere extratropical cyclone activity for four mid-season months. *Journal of Climatology* 4: 297–310.
 Simmonds and Murray (1999) Southern extratropical cyclone behavior in ECMWF analyses during the FROST special observing periods. *Weather and Forecasting* 14: 878–891.
 Hoskins and Hodges (2002) New perspectives on the Northern Hemisphere winter storm tracks. *Journal of the Atmospheric Sciences* 59: 1041–1061.
 Hoskins and Hodges (2005) A new perspective on Southern Hemisphere storm tracks. *Journal of Climate* 18: 4108–4129.
 Dos Santos Mesquita (2008) Climatological properties of summertime extra-tropical storm tracks in the Northern Hemisphere. *Tellus* 60A: 557–569.

stronger in middle latitudes, one expects the westerly shear to be stronger there as well. The temperature gradient reverses in the tropical and middle latitudes of the lower stratosphere implying easterly shear. Therefore, one anticipates westerly winds to increase with height in the troposphere and to decrease above, in the lower stratosphere. In short, the stronger westerly winds tend to occur at tropopause level.

Another constraint on the zonal wind is angular momentum balance. If the winds at the surface were everywhere

westerly (say) then those winds would apply a net torque upon the surface of the Earth. A net westerly torque would speed up the rotation of the Earth and the days would be getting shorter. However, the angular momentum of the Earth is essentially constant. So, one expects that areas of easterly winds are balanced by areas of westerly winds at the surface.

The zonal mean of the zonal wind is shown in [Figure 8](#) for winter and summer. Areas that are shaded indicate easterly winds. The following properties are evident in the figure.

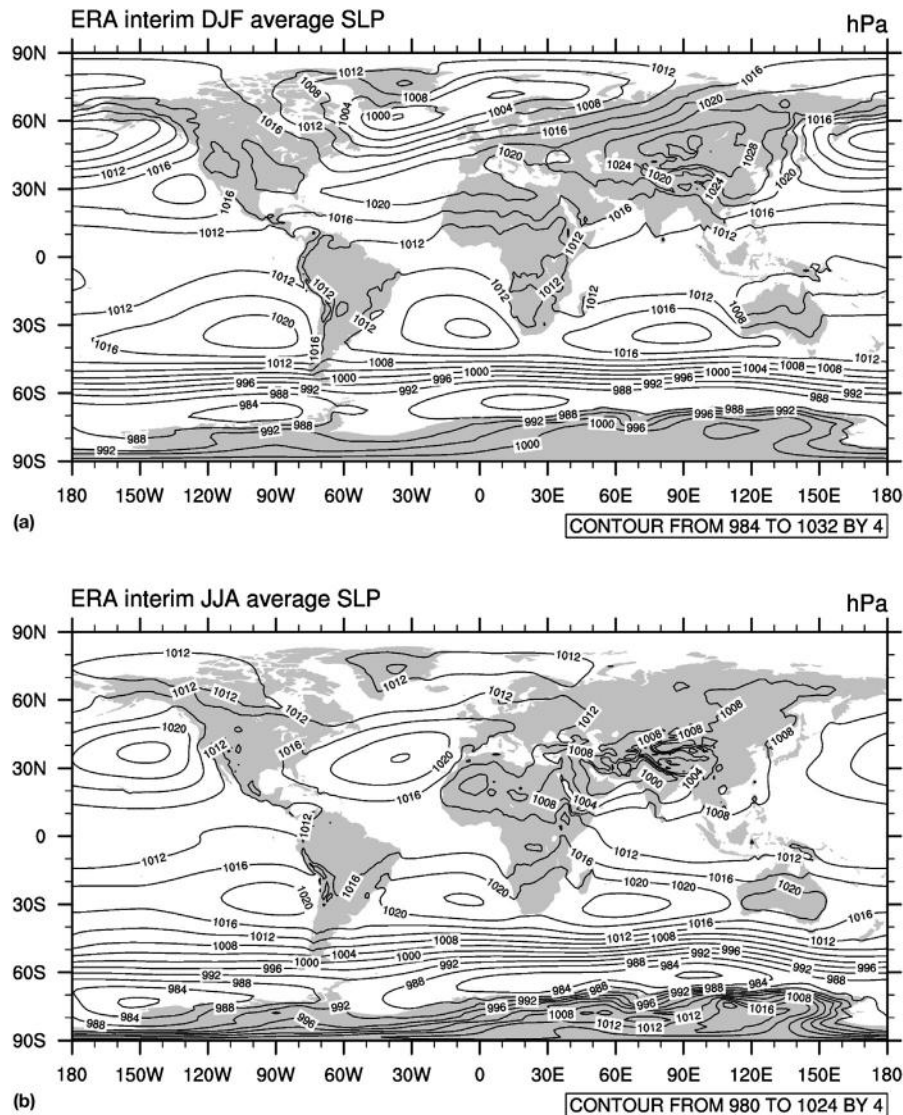


Figure 7 SLP contours in the two extreme seasons (a) December–February and (b) June–August. Units are in hectoPascals. Contours use 4 hPa interval. The data are from the European Centre for Medium-Range Weather Forecasts ERA-Interim Reanalysis averaged data from 1989 to 2010.

1. The surface winds are generally easterly in the tropics, which cover approximately half of the surface area of the globe. The surface winds are primarily westerly in the middle latitudes.
2. The winds generally gain a westerly component with elevation. For the tropical troposphere, the easterly wind decreases with increasing elevation. For much of the middle latitudes, the westerly wind increases with height until the tropopause. Above, the vertical shear reverses and the westerlies decrease with increasing height. The principal exceptions are the high latitudes in winter. These properties, including the exceptions, are anticipated from the temperature gradients shown above (Figure 3).
3. In the high latitude, winter stratosphere, westerly winds increase with height. It is difficult to see with the vertical coordinate chosen, but these westerly winds are associated with the polar night jet. This jet reaches maximum speed in the upper stratosphere (10–30 hPa level).
4. The subtropical jets are prominent maxima at the midlatitude tropopause. These jets are stronger and migrate to a lower latitude in winter. In the Southern Hemisphere during winter, strong winds of the polar night jet extend into the troposphere creating the impression that there are two tropospheric jets on upper-level isobaric charts. During northern summer an easterly jet is visible at tropopause level just north of the equator.

Meridional Circulations

The meridional wind at a specific location can often be comparable in magnitude to the zonal wind. However, if the meridional wind is geostrophic, then a zonal average of the meridional wind is the integral (with respect to longitude) of

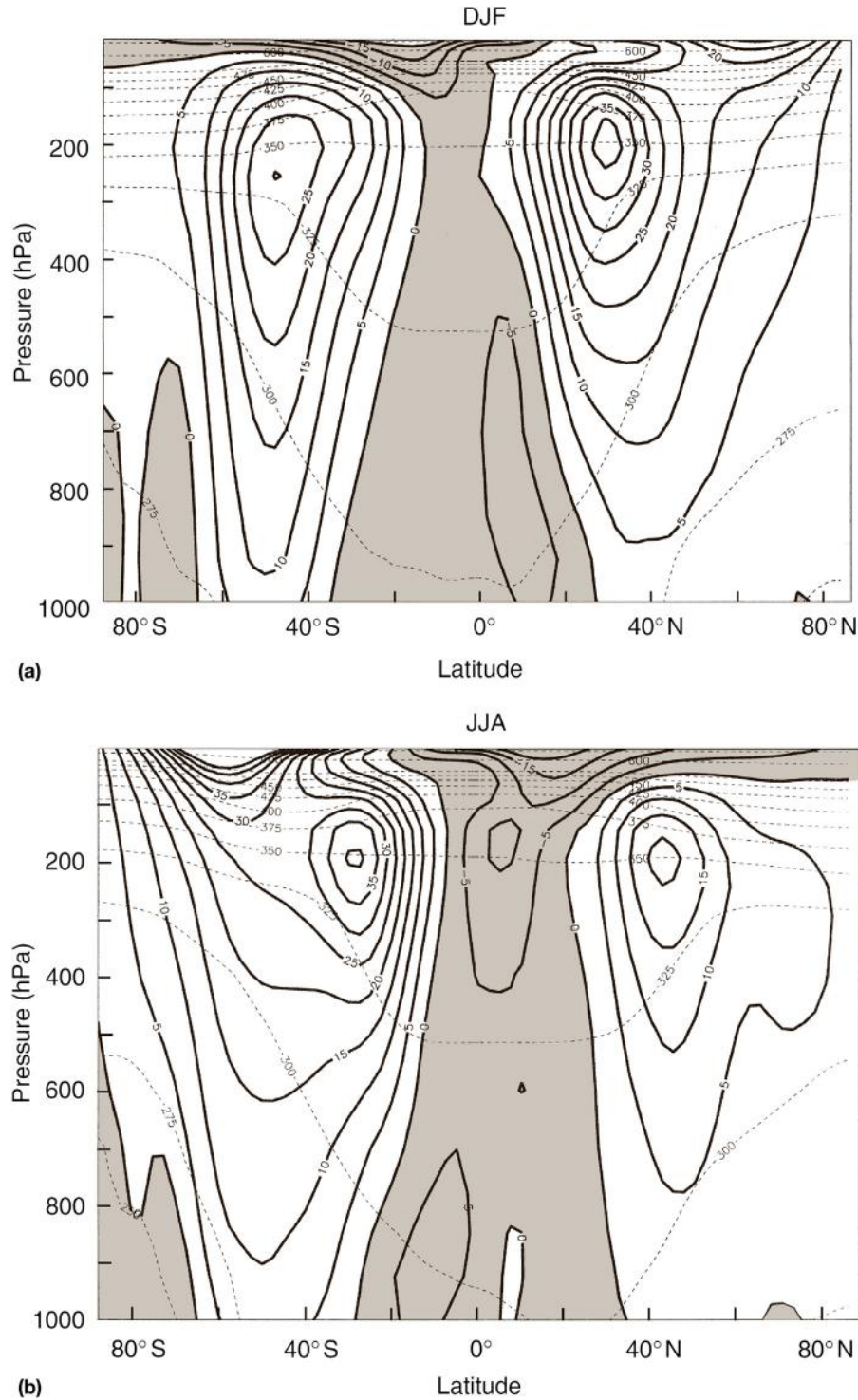


Figure 8 Zonal mean zonal wind (solid contours; in m s^{-1}) with potential temperature (dashed contours; in Kelvin) for seasonal averages (a) December–February and (b) June–August. Areas of easterly winds are shaded. NCEP/NCAR reanalysis data from 1979 to 1999.

a longitudinal derivative. Unless a mountain intercepts the integral path, this integral must be zero because the integral completes a circuit. Since the total wind is nearly in geostrophic balance outside the tropics, the zonal mean meridional wind is very small outside the tropics.

Since the observed lapse rate is less than the dry adiabatic lapse rate, it follows that vertical motions are resisted. One consequence is that vertical motions are far smaller in magnitude than are horizontal motions on this large scale. Large-scale vertical motion cannot be directly measured since it is smaller

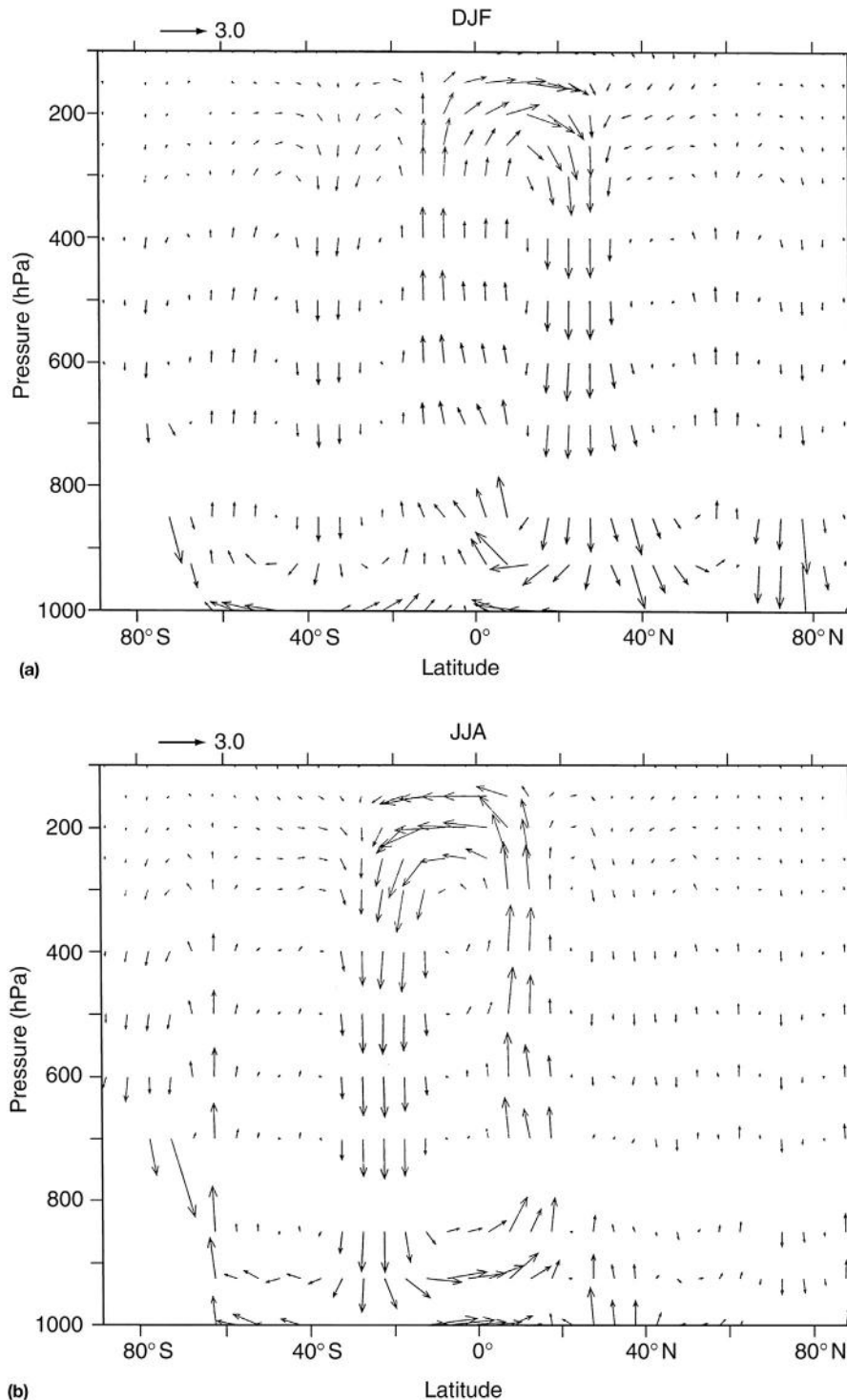


Figure 9 Zonal mean meridional circulations for (a) December–February and (b) June–August. Some of the large vectors for $p > 700$ hPa over Antarctica may not be meaningful. Vectors based on NCEP/NCAR reanalysis data from 1979 to 1999.

than the errors of the observing systems. Vertical velocity can be estimated indirectly from quantities that are measured with some confidence. One procedure is to merge measured meridional winds in the tropics with meridional winds elsewhere estimated from an equation for angular momentum

balance and then deduce a stream function in the meridional plane. An alternative procedure is to input observed fields into a primitive equation general circulation model and let the model deduce the vertical motion. The latter procedure obtained the motions shown in Figure 9.

The zonal mean motions in the meridional plane have the following properties:

1. The motion is organized into distinct patterns commonly referred to as meridional cells. The most prominent cells occur in the tropics and are often named the 'Hadley' cells. In the middle latitudes of each hemisphere is found a much weaker circulation usually labeled the 'Ferrel' cell.
2. The meridional cells are much stronger during winter both in terms of the areal extent they occupy and the vigor of the circulation. The winter Hadley cell has significant flow across the equator.
3. In the Hadley cell, air circulates in an intuitive sense: rising motion occurs where the mean temperature through the depth of the atmosphere is warmer, sinking motion where it is cooler. The temperature distribution over the globe is statically stable in the dry sense. However, in the lower tropical troposphere the air is very moist and nearly neutral with respect to a pseudoadiabatic lapse rate. The upward motion of the Hadley cell is driven by latent heat release as water vapor is converted into precipitation. So, one expects precipitation to be maximum in the tropics. In order to overcome the moist static stability and entrainment as air parcels rise into the upper tropical troposphere, the rising motion is embedded within thunderstorms. These thunderstorms occupy a very small fraction (about 0.5%) of the tropical surface area. Air parcels in the poleward moving branch of the Hadley cell cool radiatively, as their potential temperature decreases these parcels sink.
4. In the Ferrel cell, air appears to rise where temperatures are cooler and sink where they are warmer. This Eulerian mean motion should not be confused with the actual paths of air parcels. In middle latitudes parcel motions are strongly influenced by baroclinic waves. The zonally varying (eddy) flow associated with baroclinic waves has comparable meridional and zonal components. When an average is taken around a latitude circle at constant pressure, the northward and southward motions nearly cancel and the resulting mean has the sense given by the Ferrel cell. Such a mean creates a misleading picture of the motions of air parcels. However, if an average is taken along constant entropy surfaces, the resulting zonal mean in isentropic coordinates shows a meridional circulation having rising in the subtropics and sinking at high latitudes, i.e., the same sense as the Hadley circulation. The pattern in isentropic coordinates follows from the mass distribution of baroclinic eddies: in an upper isentropic layer slightly more mass is heading poleward than equatorward, vice versa for a lower layer. Therefore, in isentropic coordinates the motion is more intuitive: rising motion at lower latitudes and sinking motion at higher latitudes. Even though the Ferrel cell does not reflect the actual motion of parcels, the Ferrel cell does depict components of the motion necessary to counteract eddy fluxes of heat and momentum that would otherwise destroy thermal wind balance. (Eddy heat fluxes reduce the meridional temperature gradient while eddy momentum fluxes increase the westerly vertical shear. Planetary angular momentum and vertical adiabatic heat transports by the Ferrel cell can counteract these eddy fluxes.)
5. The radiative energy distribution requires a meridional circulation to transport heat poleward. The meridional motions of the Hadley cell transport the same mass poleward as southward (ignoring the mass due to moisture). The Hadley cell has a net heat transport because the upper-level air has higher moist static energy than does the lower-level air. In the case of the Ferrel cell, the frontal systems have strong heat fluxes that can be deduced from westward tilts with height of the trough and ridge axes of these waves.

Divergent Tropical Circulations

The Hadley cell upward motion is part of the divergent winds of the tropics. Diverging arrows in [Figure 10](#) imply rising motion below. Following the divergent wind vectors appears to connect areas of preferred rising with the sinking above the eastern sides of several subtropical highs. Some divergent winds are poleward and thus consistent with the Hadley circulation at 200 hPa. The strongest areas of divergence in [Figure 10](#) overlie Southeast Asia, Indonesia, and Amazonia. In northern summer, the Southeast Asian Monsoon is prominent. The figure also shows prominent east–west motion in the Pacific that is generally referred to as the 'Walker' circulation. The Walker circulation apparently connects rising motion (and thus precipitation) over the far Western Pacific with enhanced sinking over the east side of the subtropical high in the Eastern Pacific. The rising air is fed by low-level convergence that results from ageostrophic motions that occur for relatively low surface pressure ([Figure 7\(b\)](#)). Low-pass filtered observations show a correlation between heavier precipitation in Indonesia, lower pressure there, and stronger Pacific subtropical highs.

Subtropical Jet Streams

Zonal variations of the subtropical jet streams are linked to several phenomena discussed above. The time mean wind speed is plotted in [Figure 11](#). Consistent with geostrophic balance, the winds are stronger where the horizontal gradient of time mean geopotential height is stronger. That gradient was stronger at the bases of longwave troughs in geopotential (and temperature) near the east coasts of Asia, North America, and Australia (during local winter).

The Northern Hemisphere subtropical jet streams have a larger relative maxima near the east coast of Asia than the corresponding part of North America. These maxima are much stronger during winter (December–February (DJF)). The downstream end of each of these maxima is further poleward than the upstream end. Consequently, two jets occur over both the Eastern Pacific and the Eastern Atlantic, one velocity maximum south of the other. This combination of velocity maxima contributes to sinking above the east side of the SLP subtropical highs.

In the Southern Hemisphere the jet streams are more zonally oriented. In summer there is a tendency for stronger flow south of Africa. In winter (June–August (JJA)) the stronger winds occur east of Australia. As anticipated from the zonal mean ([Figure 8](#)) there is a secondary maximum at a higher

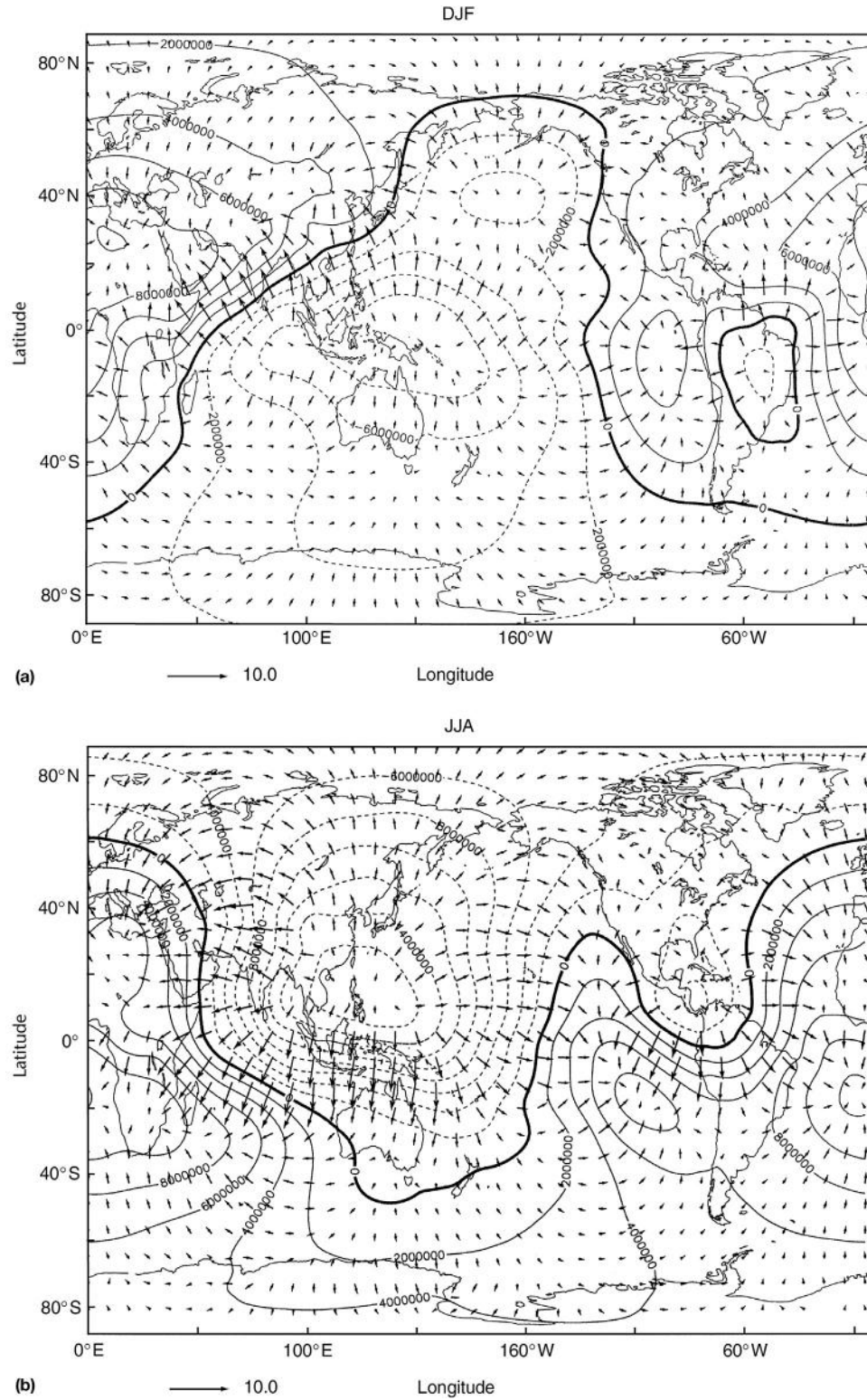


Figure 10 Time mean divergent wind (arrows) and velocity potential (contours) at 200 mbar during (a) December–February and (b) June–August. The longest arrow on the map is approximately 7 m s^{-1} and the contour interval is $2 \times 10^6 \text{ m}^2 \text{ s}^{-1}$. NCEP/NCAR reanalysis data from 1979 to 1999.

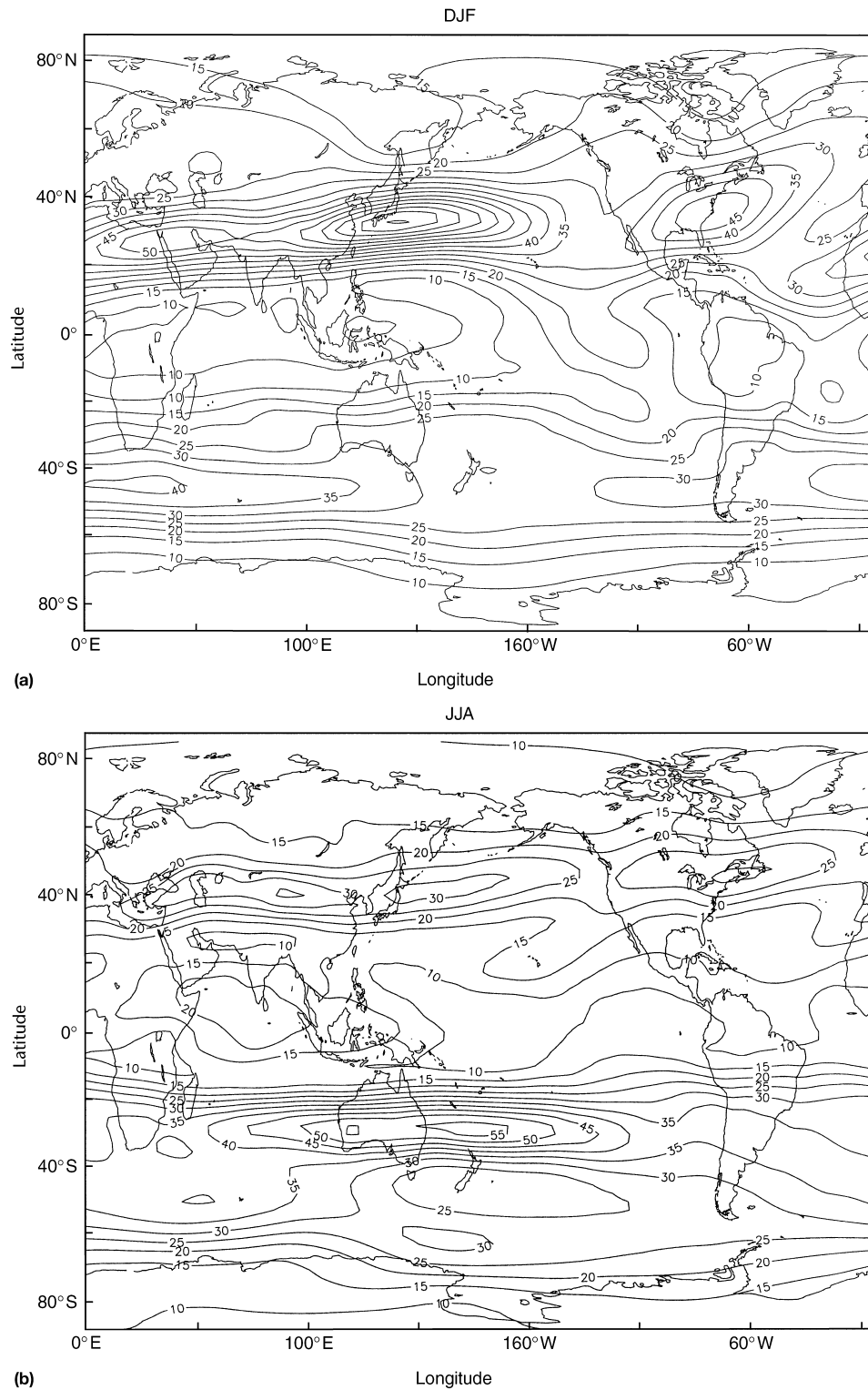


Figure 11 Time mean horizontal wind at 200 hPa for (a) December–February and (b) June–August. Contour interval is 5 m s^{-1} . NCEP/NCAR reanalysis data from 1979 to 1999.

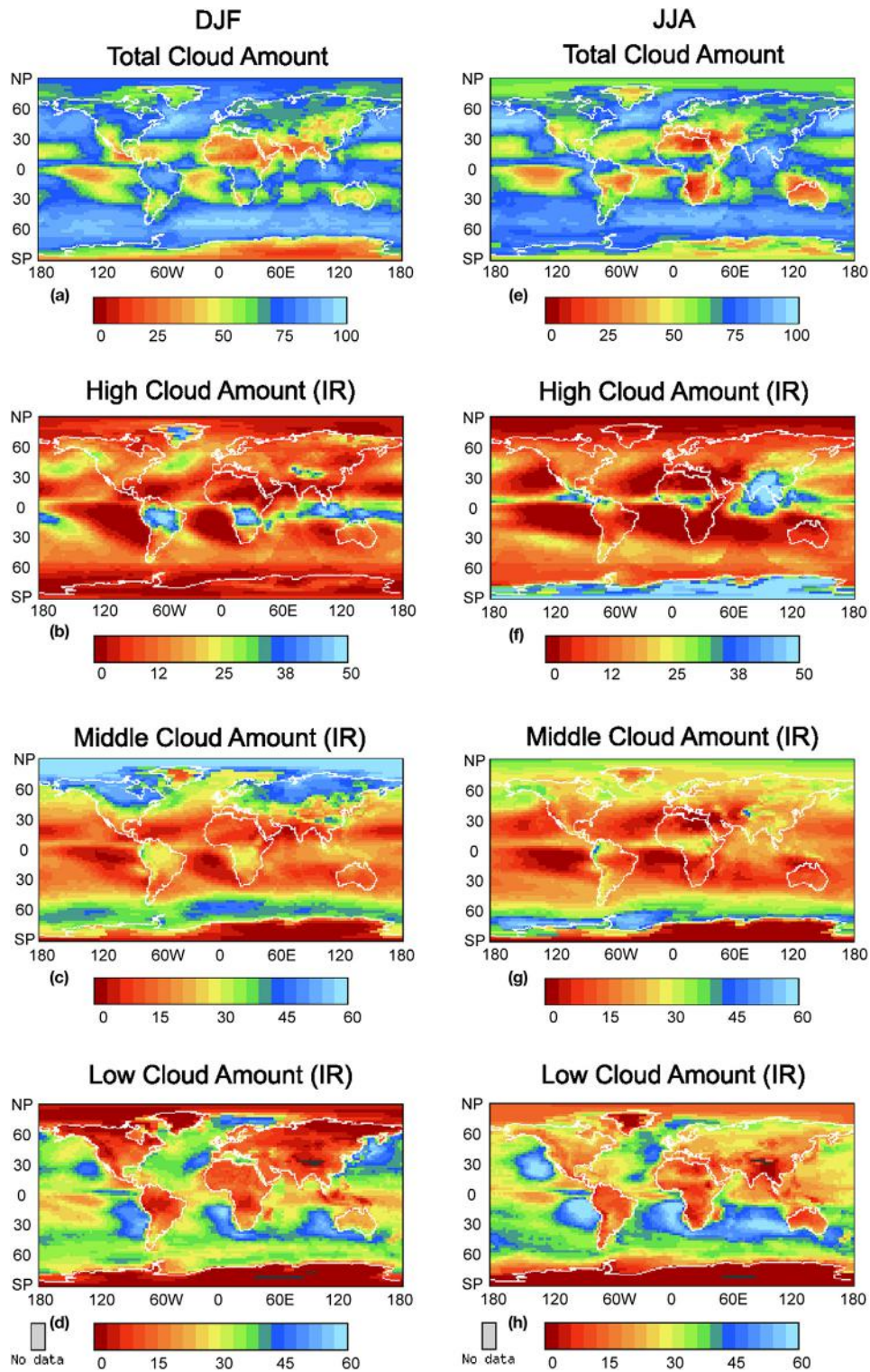


Figure 12 Cloud amount during December–February (left column) and June–August (right column). (a) and (e) Total amount at all levels. Estimates of average cloud amounts assigned to three general elevation ranges are also shown. (b) and (f) High clouds in the range of 50–440 hPa. (c) and (g) Middle tropospheric clouds in the range of 440–680 hPa. (d) and (h) Low-level clouds below 680 hPa pressure surface elevation. Cloud amounts based on satellite estimates of the fraction of 5 km wide pixels report cloud in each 280 km wide region. Geostationary and polar orbiting satellite data used, leading to some artificial boundary effects (e.g., Indian Ocean). Data are provided by the International Satellite Cloud Climatology Project (ISCCP) D2 monthly means combined from the July 1983 through June 2008 time period. If reproduced in grayscale, dark gray form few clouds, lightest gray for a middle amount of clouds for each indicated range, and light inside darker gray for higher amounts in the range.

Data maintained by the ISCCP research group at the NASA Goddard Institute for Space Studies, New York, NY, on January, 2005. Reproduced from Rossow, W.B., Schiffer, R.A., 1999. Advances in understanding clouds from ISCCP. *Bulletin of the American Meteorological Society* 80: 2261–2288.

latitude (in the Southern Indian Ocean and south of New Zealand). The more southerly maxima are a downward expression of the stratospheric polar night jet.

The divergent winds of the Hadley cell advect planetary angular momentum poleward and thereby strengthen the subtropical jets. As shown in Figure 10, the tropical circulation is not uniform with longitude but rising motion is enhanced in certain regions. The poleward motion consequently strengthens the subtropical jet in favored locations. Near the east coast of Asia, divergent winds blowing northward from the Indonesia region during northern winter build higher pressure to the southeast of the Asian longwave trough. The higher pressure there amplifies the height gradient on the southeast side of that trough making the jet stronger. The divergent flow southward from the Indonesia region contributes to the stronger jet over Australia by similar reasoning.

Moisture

Clouds

Figure 12 shows the distribution of clouds in terms of cloud amount. Cloud amount is a measure of the fractional area of the sky covered by clouds. Figure 12 shows the total cloud and the cloud amounts in three ranges based on an estimate of the pressure at the top of each cloud. Clouds at one location often occur at multiple levels, however, satellites observe the topmost layer. Since satellites observe the highest clouds, the seasonal means of the middle and low clouds shown in Figure 12 have been augmented by other observations of humidity beneath higher clouds over some land areas.

Areas of rising motion deduced from the velocity potential (Figure 10) are collocated with some, but not all, of the cloudy regions. The broad regions of rising motion deduced from Figure 10 correspond with generally smaller cloudy regions with some details in the cloud structure related to the topography. Generally, the match between Figures 10 and 12 is better for deep clouds (areas having higher cloud amounts at all levels). Some areas of sinking deduced from the velocity potential over the eastern ocean basins are also near persistent clouds, but these are low clouds.

The following properties are evident in the seasonal mean cloud amounts:

1. Deep clouds are found along the ICZ. The clouds tend to occupy a narrower band over the oceans (e.g., the Eastern Pacific) and are broader over the large tropical landmasses (e.g., Southeast Asia). The impact of the ICZ cloudiness on radiation was noted in Figure 2.
2. Cloud amounts are low over subtropical deserts but high over the adjacent east sides of the subtropical oceans. These clouds are confined to low elevation and tend to be more prominent during local spring (Southern Hemisphere) or summer (Northern Hemisphere).
3. Cloud amounts are high along the midlatitude storm tracks (Figure 6). In the Northern Hemisphere, the high and middle level cloudinesses along the storm tracks are notably greater during winter. The corresponding seasonal change in the Southern Hemisphere is less obvious.

Precipitation

Precipitation is linked to the general circulation in several ways. Figure 13 shows the zonal mean distribution of precipitation rate. The largest precipitation rate occurs in the tropics along the ICZ and is associated with the rising branch of the Hadley cells. The peak value migrates to about 10–15° latitude in the summer hemisphere (on a zonal average); a migration consistent with the dominance of the winter hemisphere Hadley cell (Figure 9). In some seasons, two equatorial maxima are found whose explanation becomes clear when time mean fields are consulted. Subtropical sinking motion seen in the meridional cells suppresses the precipitation rate. Secondary maxima occur in middle latitudes that are superficially linked to the rising branch of the Ferrel cells, but are more properly associated with the extratropical cyclones. While the cloudiness (Figure 12) is high in high latitudes precipitation is light, the colder air does not contain as much water vapor mass as in the tropics.

The zonal mean Hadley circulation implies transport of moisture from the subtropics toward the ICZ. The vigorous upward motion near the equator is driven by precipitation formed within thunderstorms that cover only a small fraction of the tropics at any one time. The time mean precipitation, Figure 14, shows tropical precipitation matching the larger cloud masses (especially high cloud) shown in Figure 12. Therefore, precipitation rate is higher over tropical land areas and regions of warmest sea surface temperature (i.e., Western Pacific). During southern summer, precipitation is greatly enhanced over the Southern Indian Ocean and tropical landmasses, while over the Eastern Pacific the ICZ remains north of the equator resulting in the double maximum seen in the zonal average during DJF (Figure 13). The South Pacific Convergence Zone is also prominent; part forms the ICZ of the Western Pacific and part is the northwest–southeast oriented line of higher precipitation across the South Pacific.

Higher time mean precipitation rates (Figure 14) are found along the midlatitude storm tracks (Figure 6) with larger

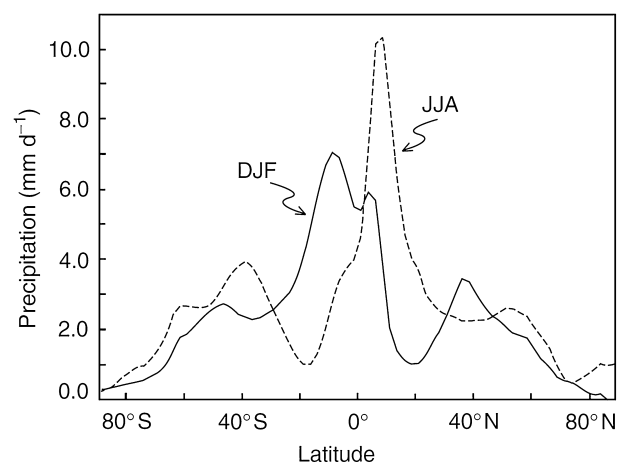


Figure 13 Zonal mean precipitation rate in millimeters per day for December–February (solid line) and June–August (dotted line). Climate Prediction Center Merged Analysis of Precipitation (CMAP) 1979–99 data used. Data provided by the NOAA/OAR/ESRL PSD, Boulder, Colorado, USA, from their Web site at <http://www.esrl.noaa.gov/psd/>.

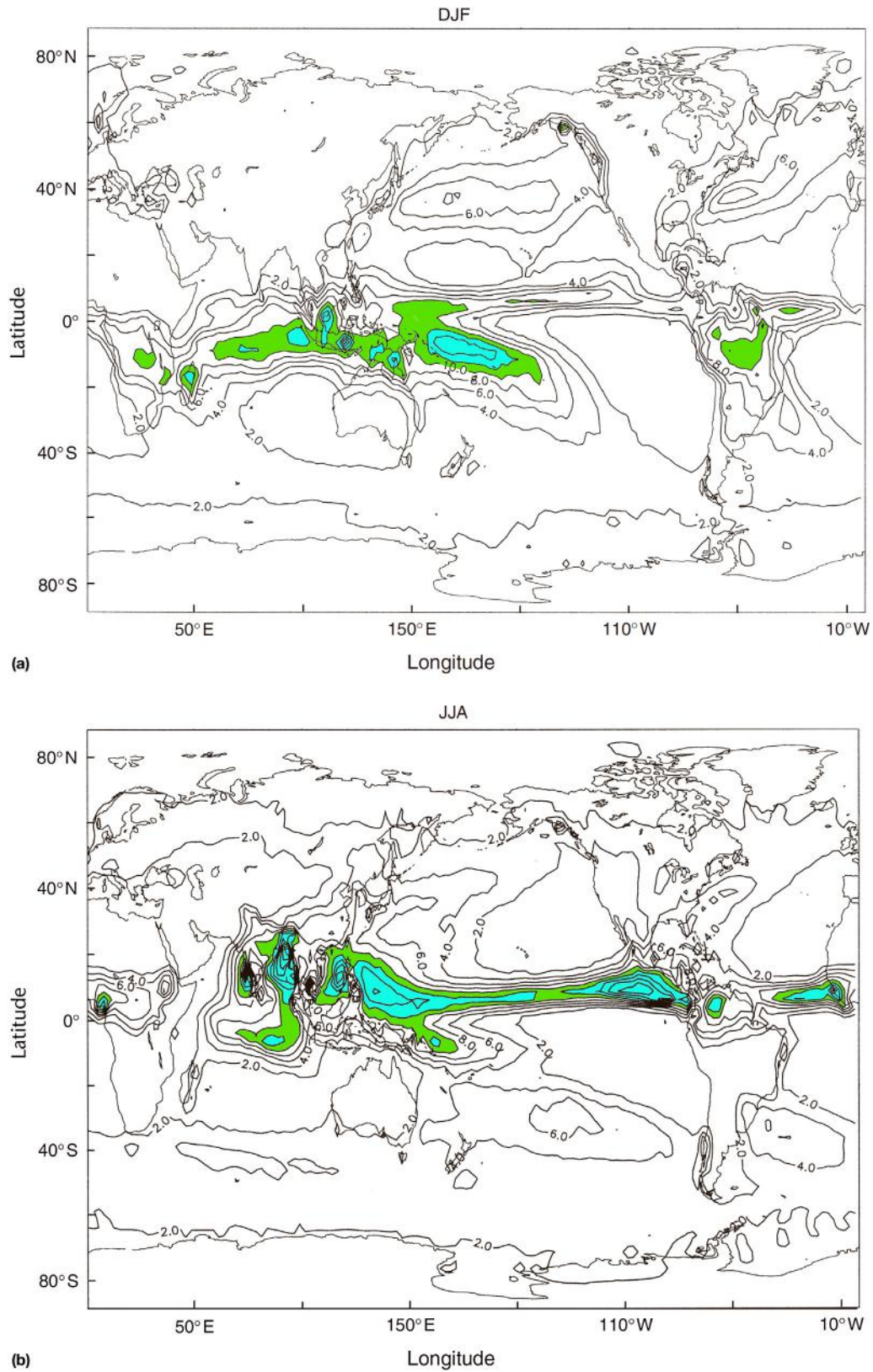


Figure 14 Seasonal time mean precipitation rate for (a) December–February and (b) June–August. Contour interval is 2 mm day^{-1} with areas exceeding 10 mm day^{-1} shaded. CMAP data from 1979 to 1999.

amounts near the start of the tracks and where the flow encounters topography (west coasts of the Americas).

Broad Summary

The observed time and zonal mean properties of the large-scale primary atmospheric variables were briefly described. Above the planetary boundary layer, these variables tend to have stronger meridional than longitudinal variation, thereby validating the use of zonal average depictions. The zonal variations are related to surface properties (land versus sea, ice versus vegetation, surface temperature gradient) and major topographic features. Seasonal changes were described; the most general comment is that seasonal variation is less in the Southern Hemisphere compared to the Northern Hemisphere.

See also: **Climate and Climate Change:** Climate Variability: Seasonal and Interannual Variability. **Clouds and Fog:** Climatology. **Dynamical Meteorology:** Balanced Flow; Overview; Quasigeostrophic Theory. **General Circulation of the Atmosphere:** Angular Momentum of the Atmosphere; Energy Cycle. **Global Change:** Upper Atmospheric Change. **Middle Atmosphere:** Planetary Waves; Polar Vortex; Transport Circulation; Zonal Mean Climatology. **Numerical Models:** General Circulation Models. **Satellites and Satellite Remote**

Sensing: Earth's Radiation Budget. **Solar System/Sun, Atmospheres, Evolution of Atmospheres:** Planetary Atmospheres: Mars; Planetary Atmospheres: Venus. **Synoptic Meteorology:** Anticyclones; Weather Maps. **Thermodynamics:** Moist (Unsaturated) Air. **Tropical Meteorology and Climate:** Hadley Circulation; Monsoon: Overview; Tropical Climates; Walker Circulation.

Further Reading

- Grotjahn, R., 1993. *Global Atmospheric Circulations: Observations and Theories*. Oxford University Press, New York, NY.
- Grotjahn, R., 2008. Different data, different general circulations? A comparison of selected fields in NCEP/DOE AMIP-II and ECMWF ERA-40 reanalyses. *Dynamics of Atmospheres and Oceans* 44, 108–142.
- James, I.N., 1994. *Introduction to Circulating Atmospheres*. Cambridge University Press, Cambridge, UK.
- Johnson, D.R., 1989. The forcing and maintenance of global monsoonal circulations: An isentropic analysis. In: *Advances in Geophysics*, vol. 31. Academic Press, San Diego, CA, pp. 43–316.
- Karoly, D.J., Vincent, D.G., 1998. *Meteorology of the Southern Hemisphere*. American Meteorological Society, Boston, MA.
- Peixoto, J.P., Oort, A.H., 1992. *Physics of Climate*. American Institute of Physics, New York, NY.
- Schneider, T., Sobel, A.H., 2007. *The Global Circulation of the Atmosphere*. Princeton University Press, Princeton, NJ.

Teleconnections

S Nigam and S Baxter, University of Maryland, College Park, MD, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by S Nigam, volume 6, pp 2243–2269, © 2003, Elsevier Ltd.

Synopsis

Teleconnections refer to the climate variability links between non-contiguous geographic regions. Teleconnection patterns are extracted from analysis of the sea-level/tropospheric pressure variations on monthly (and weekly) timescales. The methods used in extraction of teleconnection patterns are discussed and applied to recent period data. Teleconnections are especially well-developed in Northern winter, when they strongly influence subseasonal variability, notably, in surface temperature and precipitation; the stratospheric and SST links are also noted. The patterns in boreal summer are shown, as are Southern Hemisphere teleconnections. Aspects of tropical–extratropical teleconnections are discussed, along with the relationship between annular modes and teleconnections. Finally, a teleconnection analysis with pentad (5-day averaged) data is presented to initiate discussion of teleconnection evolution.

Introduction

The term teleconnection is often used in atmospheric sciences to describe the climate links between geographically separated regions. The remote region need not exhibit fluctuations of the same sign in order to be ‘teleconnected.’ In fact, the interesting teleconnections often involve contemporaneous variations of opposite signs. Climate analysis is facilitated by the construction of a teleconnection map, which describes the linkage between a region of interest (a base point) and all other points in the domain that are farther than the decorrelation length scale of the variable. Teleconnection maps thus provide information about the structure of recurrent climate variability, especially its correlation-at-a-distance features. The maps are useful because climate variability is often manifest with such structure: for example, winter variations in temperature and rainfall over southern Europe and the Iberian Peninsula are frequently opposite to those over northwestern Europe and Scandinavia.

Teleconnection maps were first constructed for meteorological parameters measured at the Earth’s surface, such as the atmospheric pressure. The selection of the base point is a critical first step, and was, historically, guided by the investigator’s insights and interests. Today, base points can be selected more objectively, and the robustness of teleconnection maps can be ascertained by independent analyses. The statistical correlation of the fluctuations provides a measure of the teleconnection strength. The structure and strength of the teleconnection patterns change with season, altitude, choice of variable, and even temporal averaging of data. There are interesting differences between the hemispheres too, in part, due to the presence of extensive continents and mountainous zones in the Northern Hemisphere (NH). Teleconnectivity in the NH winter circulation has been extensively analyzed as teleconnections account for a significant portion of the winter variance in this region.

Climate teleconnections are present in observations that have been averaged in time over a period that is long enough to suppress the day-to-day weather fluctuations, but short enough to retain the seasonal-to-interannual component of climate variability. Although teleconnection patterns often evolve on submonthly timescales (~weeks), their spatial patterns are characterized in monthly and seasonal data. Monthly averages

are typically used since a month is longer than the period of most large-scale synoptic waves in the troposphere. Climate teleconnections thus highlight the ‘standing’ component of low-frequency variability – one with geographically fixed nodes and antinodes. The connectivity of remote regions manifest in the teleconnection map indicates the potential significance of remote forcing in the generation of regional climate anomalies. A teleconnection map based on contemporaneous correlations however cannot, by itself, discriminate between the forcing and response regions.

Although the structure of the prominent teleconnection patterns has been known for some time, the reasons for their origin are not yet well understood. For example, the mechanisms that excite and sustain the North Atlantic Oscillation (NAO) – one of the notable and earliest discovered patterns – are still being investigated. In the context of such investigations, it has been questioned if the teleconnection patterns that are typically regional (e.g., NAO), robustly portray the spatial structure of variability from the viewpoint of elucidation of the underlying dynamical processes.

Such concerns are relevant since the canonical teleconnection patterns typically represent the mature phase of variability. The mature-phase pattern, however, need not resemble the nascent-phase structure, which may be more revealing of the excitation mechanism. Identification of the evolution process from analysis of the mature-phase structure is thus difficult. The spatial imprint of variability captured by a teleconnection pattern can also be ineffective in revealing the underlying mechanisms if the region in question is the locus of two temporally independent physical and/or dynamical processes. While cautionary, these remarks do not call for a radically new analysis paradigm. Instead, they point to the need for more comprehensive analysis of variability, particularly, in the spatiotemporal domain, to facilitate insights into the evolutionary process. While in-depth analysis of this kind is beyond the scope of this article, it is important to let the reader know that contemporary teleconnection research is rooted in spatiotemporal analyses.

Additionally, while teleconnectivity has historically been identified from assessment of the correlation between some defined base point and distant points in the domain, not all

teleconnection patterns are uncovered from such analysis. Examples on subseasonal to seasonal timescales include the Madden–Julian Oscillation (MJO) and the annular modes, both of which are briefly discussed in subsequent sections. An example on interannual timescales is the El Niño–Southern Oscillation (ENSO), which impacts climate around the globe.

Analysis Method

Teleconnection patterns can be extracted from correlation analysis and from the calculation of principal components (PCs), among other techniques. Both methods have been widely used in climate research, and each offers some advantages.

Correlation

Correlation analysis is the more straightforward of the two methods. Consider a meteorological field such as geopotential height that denotes the height of an isobaric surface in the atmosphere. Geopotential height, ϕ , is a function of longitude and latitude, and assume that it is defined at M grid points; ϕ_i represents height at the i th longitude–latitude grid point. Geopotential height is also a function of time, and its monthly mean value is assumed to be available for several winters (N winter months). Interest in the variability of climate leads to the consideration of departures of monthly mean heights from their calendar month climatologies, with prime (') denoting the departures; $\phi'_{i,k}$ representing the height departure at the i th grid point in the k th month. The correlation in height departures at two grid points, i and j , is denoted by H_{ij} , and defined as

$$H_{ij} = \frac{\sum_{k=1}^N \phi'_{i,k} \phi'_{j,k}}{\left(\sum_{k=1}^N \phi'^2_{i,k} \right)^{1/2} \left(\sum_{k=1}^N \phi'^2_{j,k} \right)^{1/2}}$$

Knowledge of the correlation matrix, $\mathbf{H}$, can be used to construct a teleconnectivity map ($\mathbf{T}$) that objectively identifies the base points associated with various teleconnection patterns. The map is constructed by associating the magnitude of the strongest negative correlation between a grid point and all others with that grid point, i.e., $T_i = |\text{most negative member in the } i\text{th row of the correlation matrix } \mathbf{H}|$. The local maxima in this map ($\mathbf{T}$) identify the potential base points. Linkage between the neighboring base points is assessed by examining the sites of their strongest negative correlations. A cluster of linked base points constitutes the core of the teleconnection pattern, and three prominent patterns are identifiable using this technique.

Teleconnection Map

Climate teleconnections were first investigated in the sea-level pressure (SLP) field. SLP is however an ill-defined quantity over land, particularly near the mountains; it can moreover be influenced by local meteorological processes. Teleconnectivity is thus better analyzed in upper-air data, which became available since the mid to late 1940s. The preferred analysis variable in recent decades has been the geopotential height – a measured quantity, whose horizontal and vertical gradients are proportional to the wind and temperature, respectively. Northern winter is the preferred season for teleconnection

analysis in view of the pronounced climatological stationary waves (from impressive mountains and continents in the NH) that impart a significant ‘standing’ component to the interannual fluctuations.

Figure 1(a) shows the teleconnectivity in the 500-mb geopotential height field during Northern winter. The map is constructed from correlation analysis of December, January, and February (DJF) height anomalies from 0 to 90° N for the 1979–2008 period; the data are from NOAA’s Climate Forecast System Reanalysis (CFSR). This recent reanalysis takes advantage of the latest advances in data assimilation and incorporates important data from the satellite era, and thus only extends back to 1979. This figure is comparable to **Figure 7(b)** in the pioneering study by [Wallace and Gutzler \(1981\)](#). It identifies three major teleconnection patterns in the NH: the NAO, the North Pacific Oscillation/West Pacific pattern (NPO/WP), and the Pacific–North America (PNA) pattern. The NAO and NPO/WP each consist of two base points over the Atlantic and Pacific basins, respectively. The PNA, as identified by [Wallace and Gutzler \(1981\)](#), consists of four base points: the subtropical Pacific near the dateline, the Aleutians, interior northwestern North America, and the southeastern United States. Note, however, that the fourth base point over southeastern United States is missing in **Figure 1(a)**.

This brings to light an issue regarding the stability of teleconnection patterns. Changing the domain and time period of the analysis can introduce some differences: Besides the missing PNA center, there are other differences between the seminal analysis of Wallace and Gutzler and **Figure 1(a)**. The NAO-related pattern exhibits a noticeable eastward shift, as does the third PNA center. Note, data set differences are unlikely to be the origin as the large-scale rotational flow is similarly represented in atmospheric data sets. While notable, the teleconnectivity differences are viewed as modest considering the nonoverlapping periods of the two analyses. This assessment is supported by **Figure 1(b)**, which shows the teleconnectivity map from a similar analysis of longer period data (1949–2009), one encompassing the previous nonoverlapping periods. While the NAO centers and the third PNA center here are more in accord with the Wallace and Gutzler analysis, the fourth PNA center remains indiscernible. Interestingly, the NPO/WP centers of action are also not identifiable in the longer period analysis. Even so, the overall structure of the teleconnectivity map is similar in **Figure 1(a)**.

The correlation analysis reveals the base points of a teleconnection pattern. An index constructed using the magnitude of fluctuations at the base points is used to track the pattern amplitude. For example, the PNA index is defined as

$$\alpha_{\text{PNA}}(k) = \frac{1}{3} \left(\frac{\phi'_{A,k}}{S_A} - \frac{\phi'_{B,k}}{S_B} + \frac{\phi'_{C,k}}{S_C} \right)$$

where A, B, and C are the marked base points of the PNA pattern (**Figure 1(a)**). S_i in the above definition denotes the standard deviation of height anomalies at the i th base point. Likewise an index can be constructed for the NAO- and NPO/WP-related patterns, using two base points in each case.

Correlation analysis is a physically intuitive and objective method for identifying climate teleconnections but the obtained patterns may not be independent, especially, if spatial

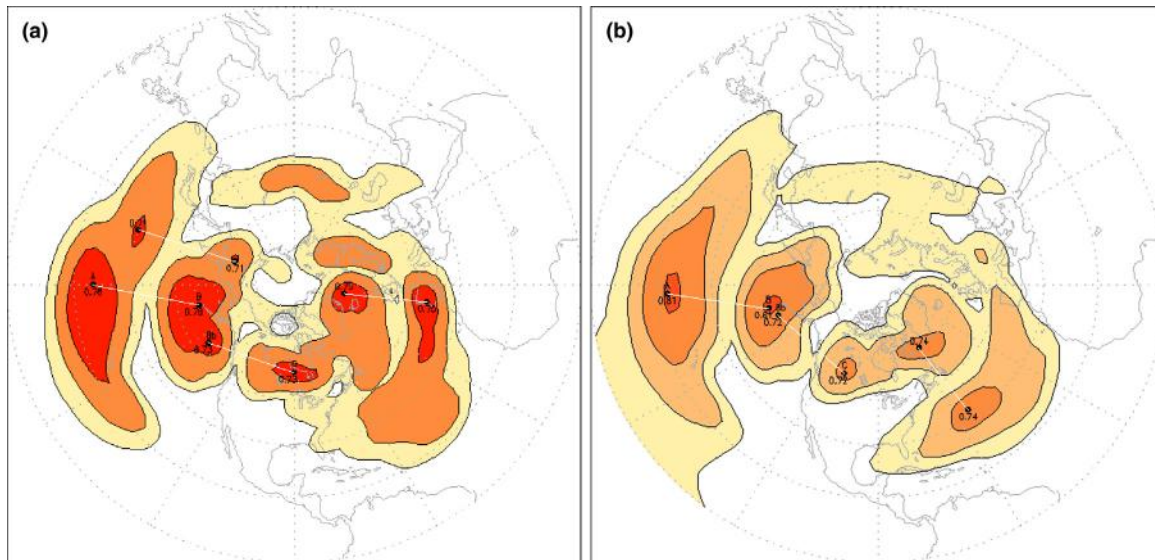


Figure 1 (a) 500-mb geopotential height teleconnectivity map for Northern Hemisphere winter (DJF), comparable to seminal **Figure 7(b)** in Wallace, J.M., Gutzler, D.S., 1981. Teleconnections in the geopotential height field during the northern hemisphere winter. *Mon. Weather Rev.* 109, 784–812. Contour/shading starts at a correlation of 0.5 and increases at 0.1 intervals. Points A, B, Bb, and C refer to the centers of actions of the PNA pattern. Points B and Bb are generally associated with the Aleutian Low center of action. (b) 500-mb geopotential height teleconnectivity map for Northern Hemisphere winter (DJF), comparable to seminal **Figure 7(b)** in Wallace, J.M., Gutzler, D.S., 1981. Teleconnections in the geopotential height field during the northern hemisphere winter. *Mon. Weather Rev.* 109, 784–812. **Figure 1**, except for a prolonged analysis period from 1949 to 2009. Contour/shading starts at a correlation of 0.5 and increases at 0.1 intervals. Points A, B, Bb, and C refer to the centers of actions of the PNA pattern. Points B and Bb are generally associated with the Aleutian Low center of action. DJF, December, January, February.

structures overlap. This, for instance, is the case in the Pacific, where three prominent patterns – PNA, WP/NPO, and the ENSO response (shown later) – overlap to varying extent in the extratropical sector. Not surprisingly, the correlation method proved unsuccessful in capturing the ENSO response in the extratropics as a teleconnection pattern. Could this be a consequence of focusing on analysis of midtropospheric variability? The 500-hPa level – a level of near-zero divergence – is, perhaps, not the level of choice for identifying tropical–extratropical interactions instigated by deep convective heating in the Tropics. Horizontal divergence associated with deep convection, such as during El Niño winters, is usually strongest in the tropical upper troposphere (~ 200 hPa). The associated midlatitude response, on the other hand, is quasi-geostrophic in character, and hence approximately nondivergent. Tropical–extratropical interactions are thus best diagnosed at a pressure level that captures the divergent outflow in the Tropics and that is near a nondivergent level in the extratropics. The 200-hPa level meets these criteria to a large extent.

Repeating the correlation analysis of height fluctuations at 200 hPa did not yield any new information on teleconnectivity. In particular, no new base points were identified and the ENSO response in the extratropics remained unidentified as before.

Principal Component Analysis

Principal component analysis (PCA) is an elegant and widely used method for determining the structure of recurrent variability. A common name for it is empirical orthogonal function (EOF) analysis. This method also analyzes the structure of the correlation matrix – the covariance matrix is preferred, though –

but focuses on regions that account for a substantial portion of the temporal variance rather than just those which exhibit strong negative correlations with distant points in the domain. In contrast with the previous method, the technique yields both the spatial patterns of recurrent variability and the extent to which these are present, or projected, in the observed anomaly record. The projection, or amplitude, is called the PC while the spatial pattern is referred to as the loading vector (LV) in the technical literature.

Recurrent variability patterns identified from PCA are spatially and temporally independent, or orthogonal. Such relationship among patterns is often helpful in investigating the origin and governing mechanisms of variability but can be relaxed in one of the dimensions – space or time – if the orthogonality constraints prove restrictive. For example, it is conceivable for two temporally independent variability patterns to have overlapping spatial structure. Imposition of temporal and spatial orthogonality constraints in this case may not lead to a physically meaningful analysis.

The EOF analysis results in a series of spatial patterns, or LVs, each one explaining successively smaller portion of the total temporal variance. One can view these recurrent patterns as teleconnection patterns. The pattern amplitude is given by the time-dependent PC. The anomaly record can be reconstructed by summing the product of the spatial patterns (LVs) and their PCs.

Rotated Principal Component Analysis

PCA provides an efficient and unique characterization of recurrent variability in terms of a small number of uncorrelated spatial patterns. The patterns are chosen so that each one

successively explains the maximal residual variance in the anomaly data set. For instance, the leading PC is obtained by requiring that it maximizes the sum of the squared temporal correlation between itself and the anomaly time series at all spatial points in the domain. The resulting PCs are temporally orthogonal while the LVs are spatially orthogonal. Spatial orthogonality can however be restrictive and, in many cases, undesirable, as discussed earlier: Although the leading LV is not directly impacted, subsequent LVs are often constrained to have predictable geometric relationships vis-a-vis the leading pattern; domain geometry, thus, becomes an influencing factor, itself.

For these and other reasons, a variant of PCA, called the rotated principal component analysis (RPCA), has become popular since it yields patterns that are no longer constrained to be spatially orthogonal; domain geometry is thus much less influential; rotated PCs continue to be temporally orthogonal, though. The linear transformation (or solid rotation) of PCs that is widely used in meteorology is called the 'varimax' rotation. It is determined by the requirement that the variance of the squared correlations between each rotated PC and the original time series be maximized. Focusing on the variance, rather than sum (as in unrotated analysis), of the squared correlations increases spatial discrimination, and facilitates interpretation of the obtained patterns.

Typically, only a subset of the leading PCs is rotated. Although several criteria exist to guide the choice of this subset, the sensitivity of results to the rotated number offers good practical guidance. In most meteorological applications, 8–10 of the leading PCs are rotated; in most applications reported here, the number is eight.

A more technical discussion of RPCA can be found in Barnston and Livezey where RPCA is advanced as a preeminent method for defining and monitoring of teleconnection patterns.

Empirical Orthogonal Teleconnection

An alternate method for identifying the teleconnection patterns is empirical orthogonal teleconnection (EOT) analysis, developed by Van den Dool et al. (2000) EOT analysis is conducted by finding a base point in the domain that explains the most variance at all the other points combined, using linear regressions. This leading spatial pattern is defined from the linear regression coefficients of this base point, while its time series is simply the raw time series at the base point. Subsequent patterns are obtained by repeating this procedure on a reduced data set, from which the first pattern has been linearly removed.

The analysis results in a series of spatial patterns, or EOTs, that are not constrained to be spatially orthogonal, but whose time series are independent. In this way, it is similar to RPCA. Other than the inherent simplicity, EOT analysis holds another advantage over RPCA: In RPCA, only a small subset of PCs are subject to rotation; no such truncation is required in EOT analysis.

Northern Winter Teleconnections from RPCA

The leading patterns of recurrent height variability are extracted from RPCA and shown in Figure 2. The DJF anomalies during 1979–2008 winters were analyzed in the 30° S–90° N domain. Analysis was conducted at the 200-hPa level in order to also

capture the tropical–extratropical interactions (e.g., ENSO related) that are prominently manifest in the upper troposphere, for reasons stated earlier. Height anomalies were multiplied by the square root of the cosine of latitude to achieve grid-area parity, which prevents polar regions with many more points on a regular latitude–longitude grid, from unduly influencing the analysis. The covariance, rather than correlation, matrix was analyzed so that regions with large variance can exert greater control on the analysis outcome. The eight leading PCs were rotated using the varimax criterion. The PCs are normalized and dimensionless and the spatial patterns (LVs) show nothing but the regression of the PCs onto the 200-mb height field.

The leading pattern explains ~15% of the monthly variance in the domain, with two centers of action in the Atlantic sector. Note the spatial similarity to the Atlantic Basin teleconnectivity structure in Figure 1, sans its earlier noted eastward shift. This leading pattern is clearly the NAO. Its structure is consistent with its historical characterization as a north–south dipole in SLP that represents out-of-phase fluctuations of the Icelandic Low and Azores High. For example, Hurrell's monthly NAO index is constructed from the difference of normalized SLP anomalies at Ponta Delgada, Azores and Stykkisholmur, Iceland; the NAO pattern depicted in Figure 2 is thus in its positive phase, which strengthens the regional winter circulation features. The identification of NAO as the leading pattern in the 200hPa analysis is important both for explanation of regional variance and for showing NAO variability to be temporally orthogonal to the other NH variability patterns. The NAO pattern is manifest on timescales ranging from subseasonal to decadal, as seen in Figure 3. The NAO pattern is closely linked with meridional excursions of the Atlantic jet, and related storm track displacements; northward in the positive phase of the pattern.

The second, third, and fourth leading patterns all have centers of action in the Pacific sector, attesting to RPCA's skill in separating overlapping variability structures. The second leading pattern represents higher geopotential heights in all sectors of the northern subtropics, but especially over the central/eastern Pacific. Higher upper-level heights are typically associated with a warmer air column underneath, since the atmosphere is in hydrostatic balance. The pattern of variability is thus associated with a warming of the Tropics, such as during El Niño winters. The presence of a subtropical ridge to the southeast of the Hawaiian Islands is indicative of linkage with ENSO, for deep convection in this tropical Pacific sector and the related divergent outflow are linked with the development of an upper-level anticyclonic circulation in the subtropics. The subtropical ridge also reflects the southeastward extension of the Asian-Pacific jet in El Niño winters. The ENSO pattern is also characterized by an east–west dipole near North America with a negative anomaly along and off the west coast and a positive anomaly near the south shore of Hudson Bay.

The PC associated with this spatial pattern exhibits inter-annual variability and is correlated ($r = 0.6$) with the Niño 3.4 index, which is the average sea surface temperature (SST) anomaly in the central-eastern equatorial Pacific (5° S–5° N, 170–120° W) and operationally used to define the state of ENSO. While this pattern, which explains ~13% of the monthly winter variance, is characterized as the ENSO response, the related PC exhibits some long-term trend as well.

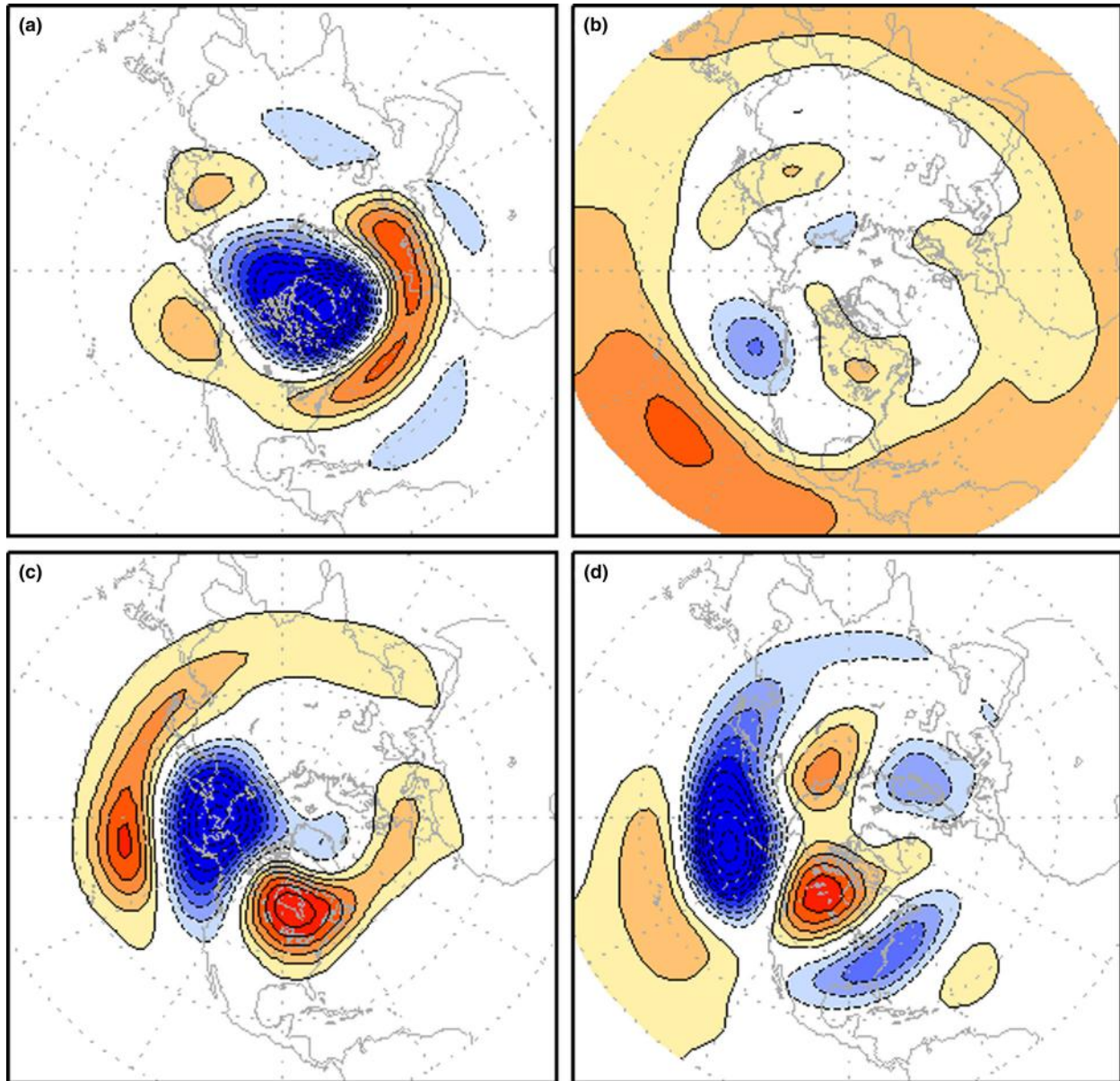


Figure 2 The four leading teleconnection patterns from a rotated EOF analysis conducted on monthly 200-mb wintertime height anomalies: (a) NAO, (b) ENSO related, (c) NPO/WP, and (d) PNA. The domain of the analysis is 30°S – 90°N . Contour/shading interval is 10 m; the zero contour is suppressed. Percentage of explained variance: (a) 14.6, (b) 12.5, (c) 11.4, (d) 11.1. NAO, North Atlantic Oscillation; ENSO, El Niño–Southern Oscillation; NPO/WP, North Pacific Oscillation/West Pacific; PNA, Pacific–North America.

It is remarkable that RPCA of upper-tropospheric heights can identify the ENSO-related height pattern without any reference to the underlying SST variability. Note that teleconnection analysis of the 200-hPa height variability in the same period was unsuccessful in this regard.

The third and fourth leading patterns of Northern winter height variability each explain $\sim 11\%$ of the monthly variance. The third leading pattern in **Figure 2** consists of a north–south dipole in the North Pacific with a downstream height anomaly over eastern North America centered over Hudson Bay. This pattern is referred to as the NPO/WP pattern as the related SLP pattern (shown later) closely resembles the NPO, a recurrent subseasonal variability pattern in SLP, first identified by Walker

and Bliss in 1932. The ‘West Pacific’ part of the name is from Wallace and Gutzler’s teleconnection analysis of 500-hPa heights. The NPO/WP pattern can be viewed as the NAO analog in the Pacific Basin: both patterns describe subseasonal height fluctuations in the jet-exit regions of the climatological jets streams. The fluctuations have a meridional dipole structure with a stronger polar center in both cases, and linked with meridional displacements of the jets and related storm tracks. A multivariate description of the NPO/WP pattern and its similarity with the NAO can be found in Linkin and Nigam.

The fourth leading pattern is the PNA, defined by the coherent arcing pattern of height fluctuations of alternating signs, beginning with the center over Hawaii and followed by

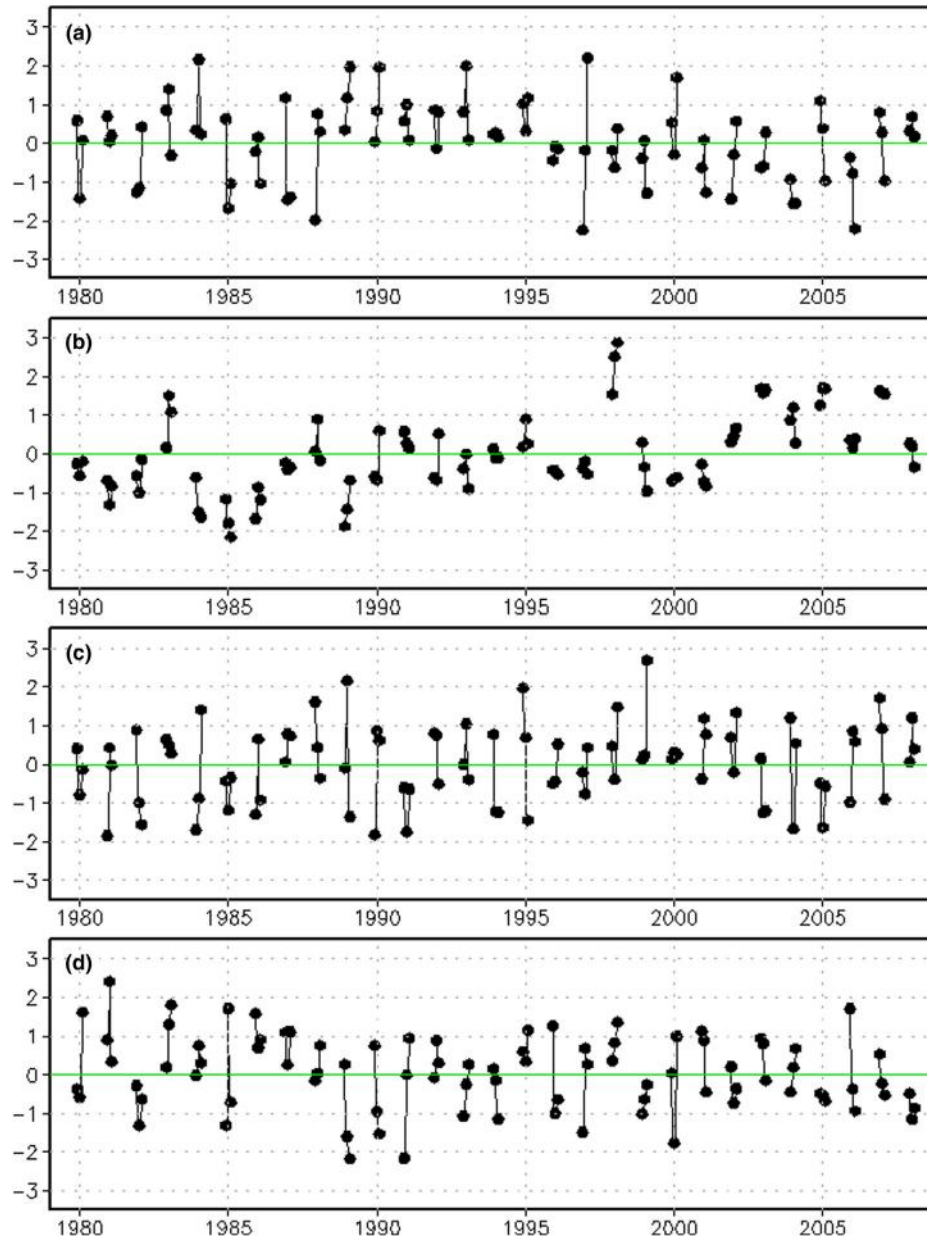


Figure 3 The principal component time series for each pattern is shown. (a) NAO, (b) ENSO related, (c) NPO/WP, (d) PNA. Note that there are only values for the DJF season. NAO, North Atlantic Oscillation; ENSO, El Niño–Southern Oscillation; NPO/WP, North Pacific Oscillation/West Pacific; PNA, Pacific–North America.

centers over the tip of the Aleutians, northwestern North America, and Southeast United States. In the positive phase (depicted), the pattern consists of positive height anomalies over northwestern North America. In the Pacific sector, the height anomalies represent eastward displacement and meridional focusing of the Asian-Pacific jet. It is interesting to note the position of the subtropical ridge in the PNA- and ENSO-related patterns, as this is helpful in pattern identification: The ridge is over the Hawaiian Islands in the former but southeastward of Islands in the latter. The variability timescales of the patterns are also distinct: subseasonal and interannual, respectively (Figure 3). The separation of the PNA- and ENSO-related patterns is important as PNA was

once viewed as the extratropical response of ENSO. The PNA and NPO/WP patterns are also distinct, notwithstanding the broadly similar structure over the North Pacific–North American region; closer inspection reveals the two patterns to be in spatial quadrature in this sector; for example, the NPO/WP trough tracks the North American coastline, which is a nodal line in the PNA pattern.

Surface Temperature and Precipitation Impacts

Teleconnection patterns are of considerable interest because of their impact on surface hydroclimate. Significant climate anomalies on subseasonal-to-interannual timescales are often

attributable to a dominant phase of one or more teleconnection patterns. Figure 4 shows the SLP, surface temperature, and precipitation footprints of the NAO. The temperature and precipitation data sets used in this analysis are from the Climate Research Unit (CRU) of the University of East Anglia, namely the TS 3.1 climate database. The SLP pattern bears striking resemblance to the 200-hPa height pattern (Figure 2(a)), indicating an equivalent barotropic vertical structure. The strongest impacts of the NAO emerge over Europe and North Africa. The NAO results in nearly all of Europe experiencing above-normal temperatures, while cooler temperatures prevail from Saharan Africa east-northeastward to the eastern Mediterranean, Middle East, and the Arabian Peninsula. The temperature pattern at the surface results largely from advection of climatological temperature by the NAO circulation (i.e., $-V_{NAO} \cdot \text{Grad}(T_{CLIM})$); for example, cooler temperatures over Africa arise from the cold advection by

northeasterlies tracking the eastern flank of the high SLP cell. Correlation coefficients are shown in Figure 4 to indicate the significance of the impact, although they conceal the fact that temperature anomalies over Europe are larger than those over Africa. The precipitation pattern likewise shows a wet-dry dipole with northern Europe experiencing above-normal, and the Iberian Peninsula and southern Europe below-normal precipitation. This impact is not obvious from the geopotential and SLP patterns unless the impact of NAO on the jet stream, and thus storm tracks, is considered. In the depicted positive phase, NAO results in enhanced zonal wind north of the Azores center and reduced zonal wind to the south. The northward displaced jet stream and storm track account, in part, for the observed precipitation dipole.

The NAO influences the surface climate of North America as well. A north-south dipole in temperature is again observed: cold in northeastern North America, warm to the south; both

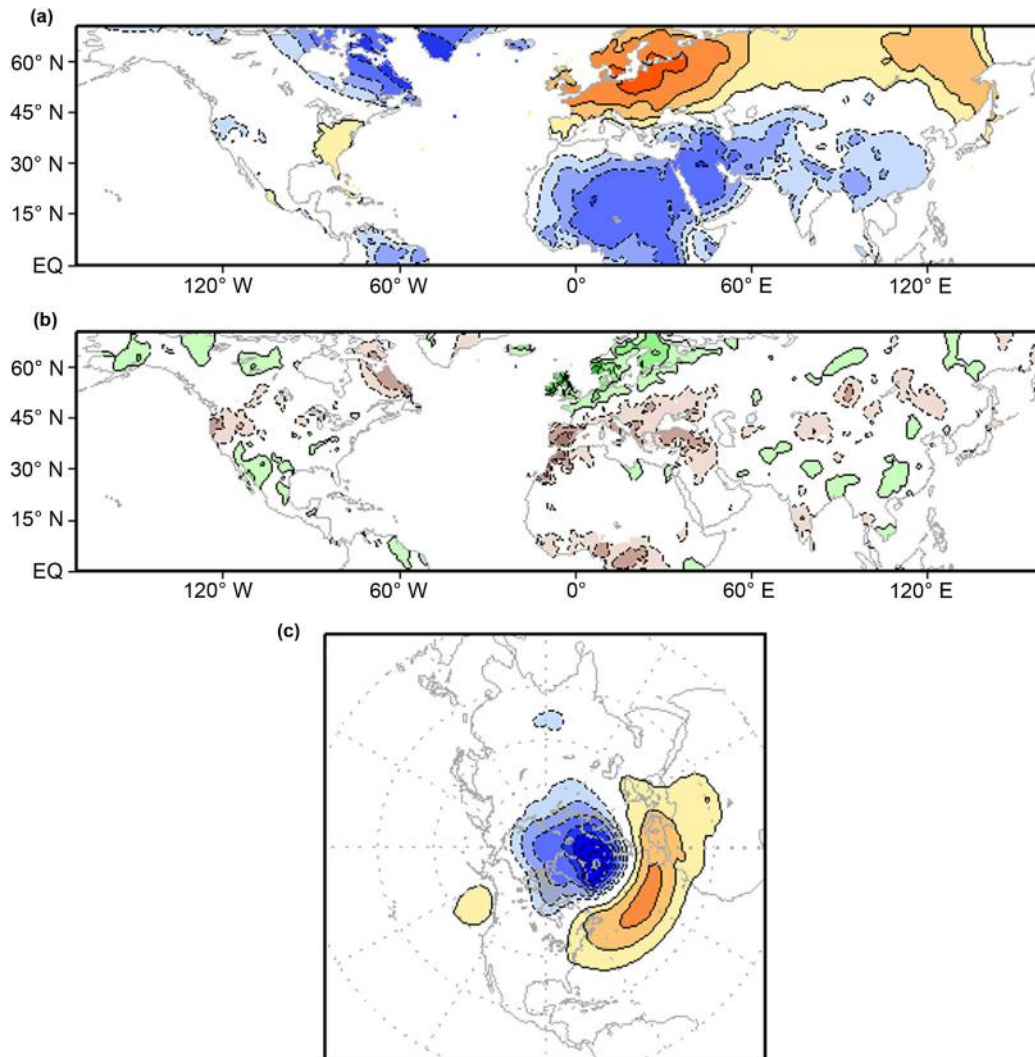


Figure 4 (a) Temperature, (b) precipitation, and (c) SLP footprints of the wintertime NAO pattern seen in Figure 2. Correlation between NAO principal component and CRU temperature and precipitation is contoured at 0.15 intervals; the zero contour is suppressed. For SLP the regression coefficient between the principal component and the SLP field from the CFSR is contoured at 1-hPa intervals, and the zero contour is suppressed. NAO, North Atlantic Oscillation; CFSR, Climate Forecast System Reanalysis.

consistent with the regional SLP signal. The precipitation impacts are less coherent and significant than in Europe, except along the west coast of the Labrador Sea and Davis Strait; the deficit here resulting from off-shore moisture advection by the NAO circulation. Interestingly, there is a significant dry anomaly on the central US West Coast that is associated with a weak ridge over the Gulf of Alaska ([Figure 2\(a\)](#)). This impact feature is revisited when annular modes are discussed.

The surface signatures of the ENSO-related pattern are shown in [Figure 5](#). The SLP signature is notably weak and focused in the Gulf of Alaska. Unlike NAO where regional thermal advection shaped surface temperature, the significant ENSO-related warming of the global northern Tropics (at both lower and upper levels) results from the propagation of atmospheric waves instigated by ENSO-related diabatic heating changes over the tropical Pacific. There is also a tendency for warmer winters across much of North America, centered on the United States–Canadian border, along with northern Europe and East Asia. Precipitation impacts are also notable in the

deep Tropics, with drying in eastern South America and increased precipitation over eastern Africa. The ENSO-related pattern results in a southward shift of precipitation across North America with increased precipitation along the US West Coast and southern tier states. Equally noteworthy are the positive precipitation anomalies over the northern Indian subcontinent, including the Himalayas. It is important to note that [Figure 5](#) does not fully capture the canonical ENSO impact on winter hydroclimate (e.g., as defined from correlations of the Niño 3.4 SST index) but comes close in most regions.

The NPO/WP pattern results in significant modulation of the North American surface climate ([Figure 6](#)). The SLP pattern is robust and oriented much as the NAO's, reflecting similar variability mechanisms. The NPO/WP circulation in the lower troposphere leads to impressive temperature correlations (exceeding 0.5) over central-eastern North America from anomalous thermal advection and related meteorological feedbacks; see Linkin and Nigam for more information. Intrusion of the sub-Arctic air into the continental interior in

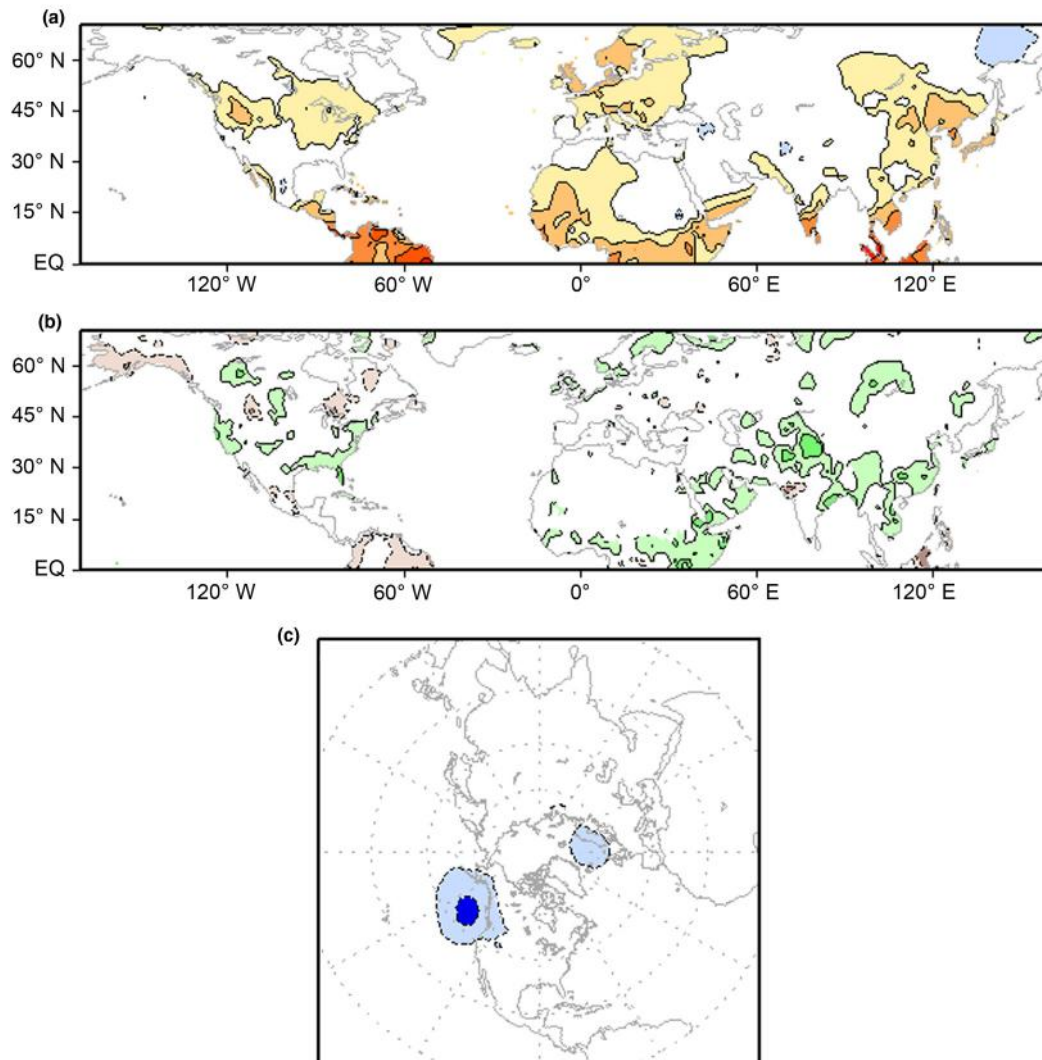


Figure 5 Same as in [Figure 3](#) except for ENSO pattern (a) temperature, (b) precipitation, (c) sea-level pressure. ENSO, El Niño–Southern Oscillation.

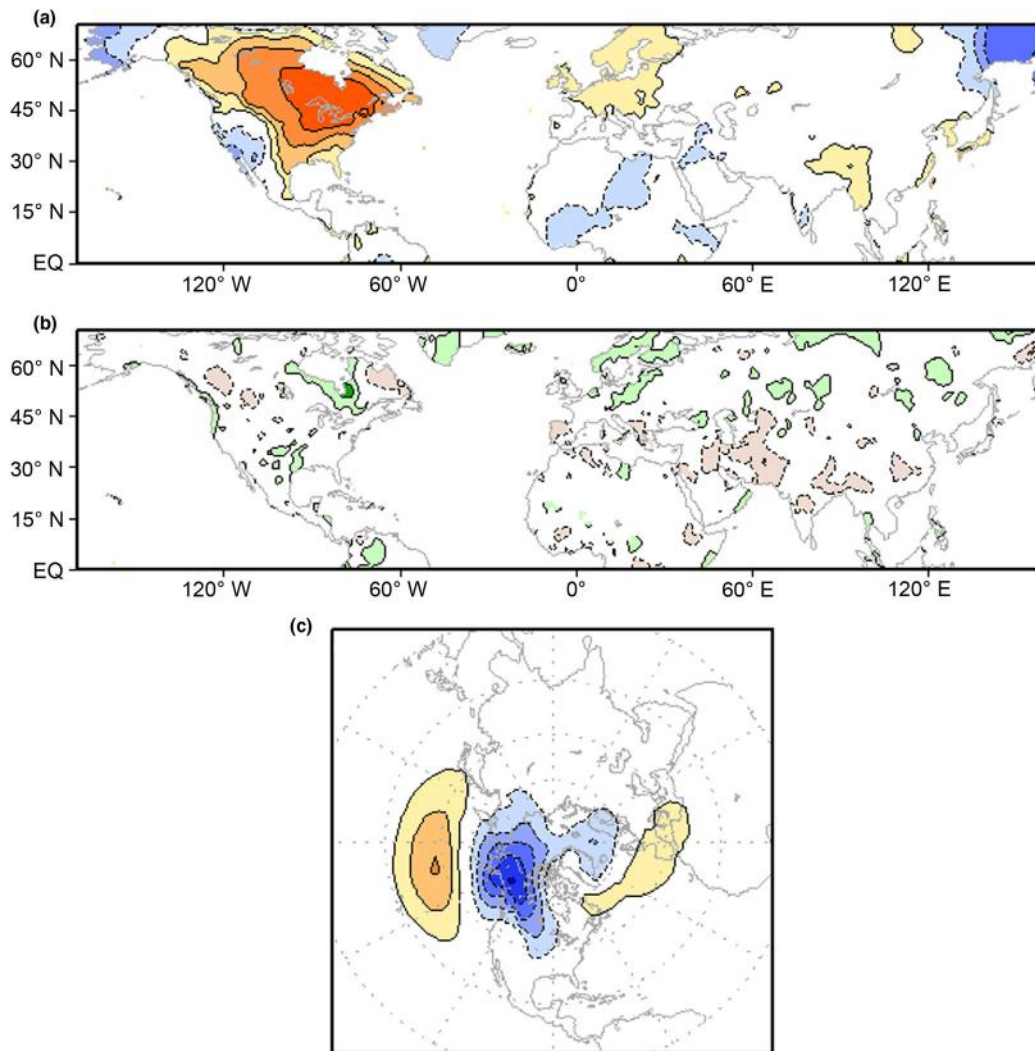


Figure 6 Same as in [Figure 3](#) except for the WP/NPO pattern (a) temperature, (b) precipitation, (c) sea-level pressure. WP/NPO, West Pacific/North Pacific Oscillation.

the negative NPO/WP phase (weakened Aleutian Low) leads to below-normal temperatures, especially to the east of the Rockies, for example. The high correlations suggest that between one-quarter and one-half of the monthly temperature variance in the core influence region is explainable by NPO/WP variability. The precipitation impact is weaker and the spatial pattern less coherent, but potentially important in the Pacific Northwest and southern Great Plains. Like the NAO, the NPO results in meridional perturbation of the jet stream in its exit region; since this occurs over the wide Pacific Basin, the related storm track modification is not manifest in analysis of continental precipitation.

The PNA pattern's influence on North American surface hydroclimate ([Figure 7](#)) is, perhaps, most impressive. The SLP footprint in the Bering Sea and Gulf of Alaska is ~ 6 hPa, larger than of the other Pacific Basin patterns, including ENSO. The temperature impact is characterized by a strong northwest–southeast dipole across North America. Not surprisingly, the impact is focused on Northwest

Canada and the Southeast United States, the third and fourth centers of action of the pattern, respectively; again, with anomalous thermal advection defining the large-scale impact structure, if not its amplitude. At first glance, the precipitation impact appears similar to the opposite phase of the ENSO-related pattern ([Figure 5\(b\)](#)). Closer inspection, however, reveals a subtle shift in the greatest precipitation correlation regions. The PNA response is defined by large correlations over interior eastern North America, centered on the Tennessee and Ohio Valleys and extending across the Great Lakes into Northern Canada. A similarly signed signal is evident in the Pacific Northwest and Alaska as well. The positive-phase PNA precipitation response ([Figure 7\(b\)](#)) thus consists of only deficits across the United States. The deficit over the eastern United States results from the southward displacement of the Atlantic jet and storm track that are linked to the circulation anomalies associated with the third and fourth centers of the PNA height pattern.

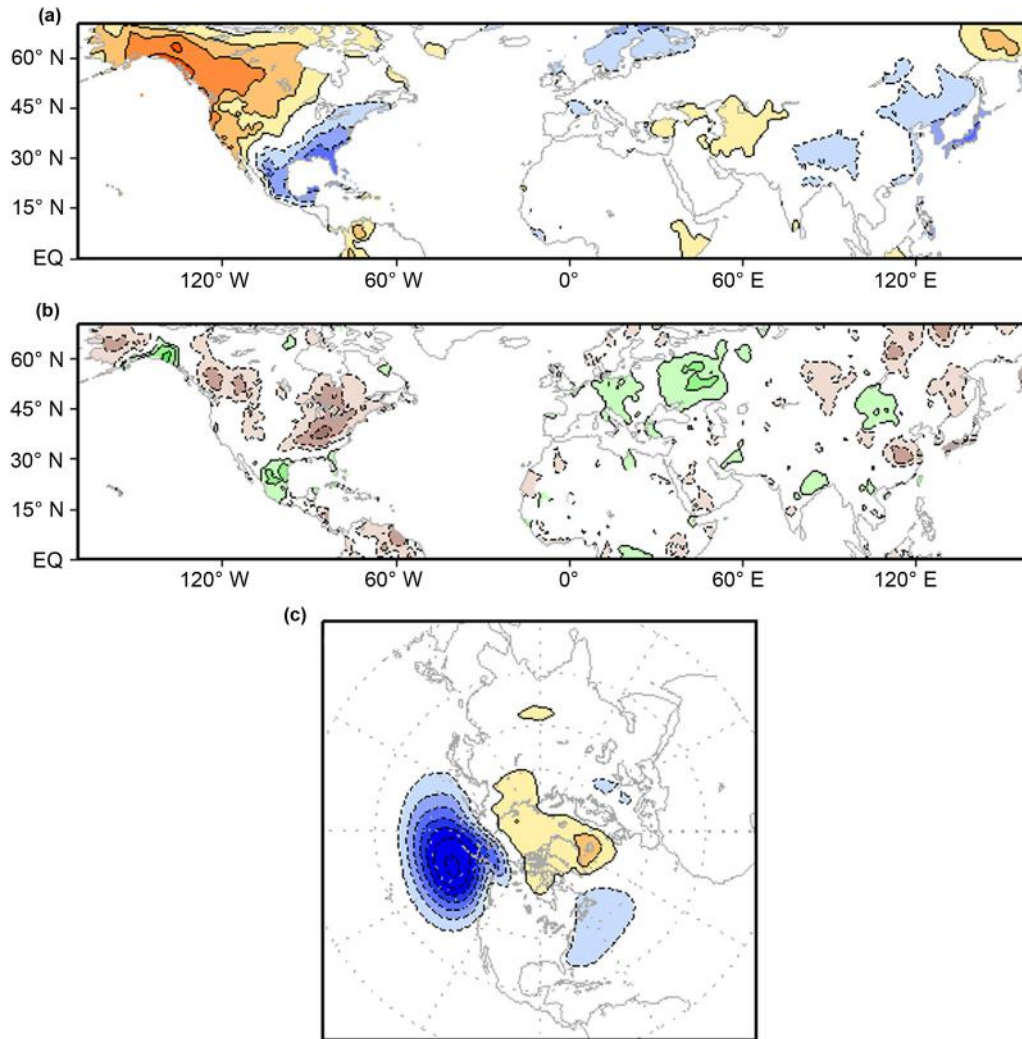


Figure 7 Same as in [Figure 3](#), except for the PNA pattern (a) temperature, (b) precipitation, (c) sea-level pressure. PNA, Pacific–North America.

Stratospheric Signature

Teleconnection patterns are obtained from the analysis of connectivity of tropospheric climate variability. Understanding the mechanisms generating the remote response, i.e., teleconnectivity, in some cases warrants analysis of the related stratospheric links since stratospheric circulation can be an influence pathway between distant tropospheric regions. The enabling role of the stratosphere in generating aspects of tropospheric teleconnectivity is getting increasing attention in observational and modeling analyses. Two of the four tropospheric teleconnection patterns – NAO and NPO/WP – exhibit impressive stratospheric links ([Figure 8](#)). Not surprisingly, both involve meridional displacement of the subtropical jets, especially in the jet-exit sector. Such change in the tropospheric jet structure (a poleward shift in the positive NAO and NPO/WP phase) modulates the refractive index for the meridional-vertical propagation of planetary waves; a poleward jet shift, for instance, reduces the orographically forced stationary wave response in both the troposphere and stratosphere in dynamical models. Jet perturbations influence not only wave

propagation but also wave forcing; the latter by modulating flow over orography and regions of strong climatological vorticity gradients, for instance. The dynamical and thermodynamical mechanisms linking the troposphere and stratosphere are of great interest in modern climate studies.

[Figure 8](#) shows the regressions of the NAO and NPO/WP PCs on geopotential heights at 100 hPa (~16 km), 50 hPa (~20 km), and 10 hPa (~30 km); the NAO ones are in the left column. The NAO footprint in the stratosphere is significantly annular, i.e., anomalies of one sign in the polar region surrounded by opposite-signed anomalies in the midlatitudes. The tropospheric structure is less annular, prompting some to view NAO as the regional expression of the annular mode of SLP; the annular modes are discussed in a later section. The NAO stratospheric structure can be broadly characterized as consisting of out-of-phase variations of the polar vortex and surrounding regions. The NPO/WP one, on the other hand, reflects displacement of this vortex into North America. This difference between displacement of the vortex and the breakdown or splitting of the vortex is analogous to the two distinct

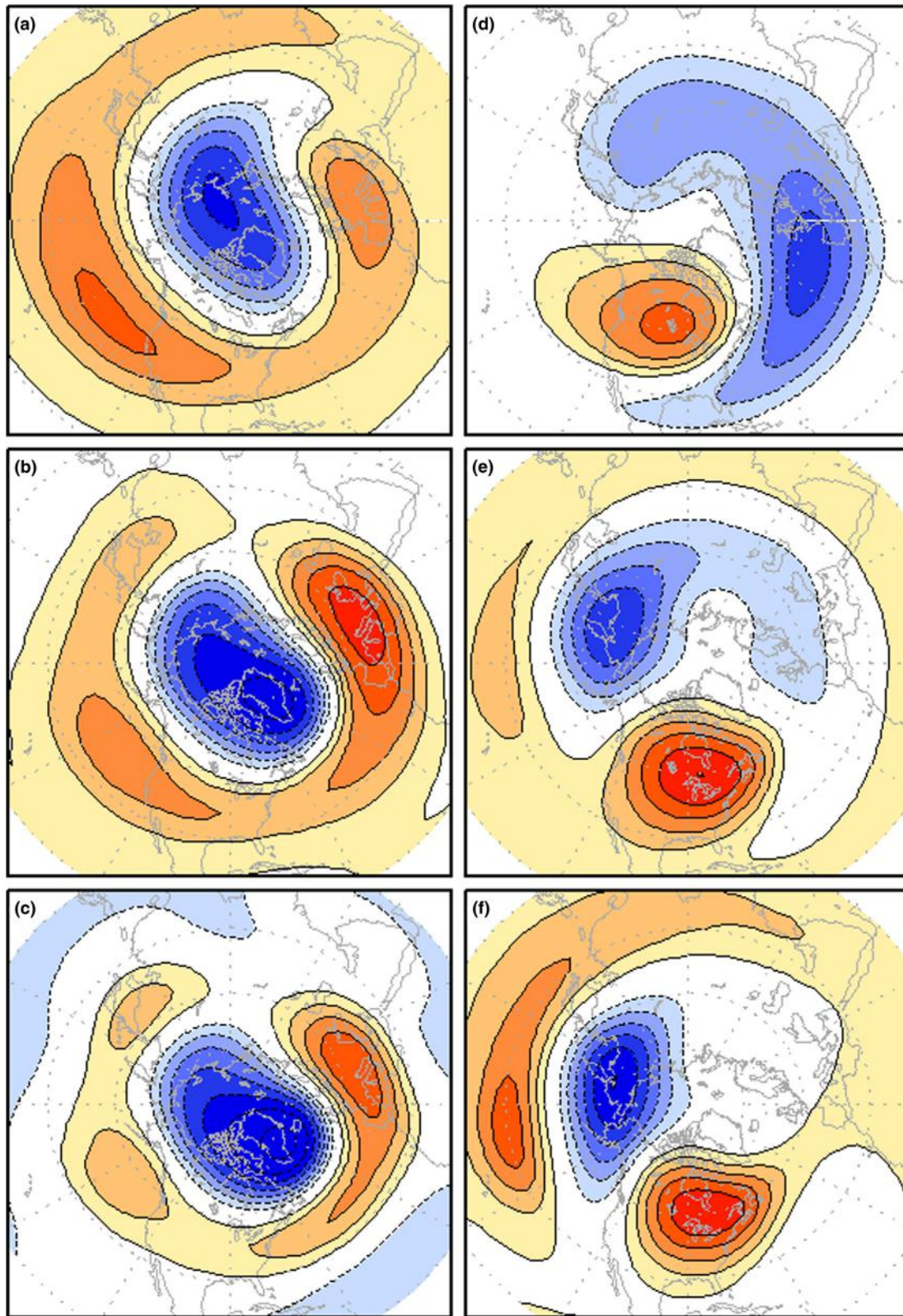


Figure 8 Stratospheric footprints of the two leading teleconnection patterns with substantial stratospheric signatures: (a)–(c) show the regression of the NAO PC onto 10-mb, 50-mb, and 100-mb normalized geopotential height anomalies, respectively. The contour interval is 0.2 standard deviations. The zero contour is suppressed in all cases. Panels (d)–(f) show the same except for the WP/NPO pattern. NAO, North Atlantic Oscillation; PC, principal component; WP/NPO, West Pacific/North Pacific Oscillation.

types of sudden warming events that occur in the stratosphere. In the case of a split vortex or complete breakdown of the vortex, an equivalent barotropic structure is observed as in the case of the NAO stratospheric signature. The NPO-related vortex displacement tends to be more baroclinic in nature, at least in the stratosphere, as evidenced by the westward tilt with height.

SST Anomalies

The teleconnections described above represent standing modes of atmospheric circulation variability on subseasonal-to-interannual timescales. As such, it is of interest to inquire

into the nature of related SST anomalies, which in the Pacific Basin can provide additional discrimination between the overlapping height variability patterns. Figure 9 shows the SST correlation of the four PCs. The NAO SST footprint reveals a characteristic tripole feature in the Atlantic Basin, with cooler temperatures underneath strengthened westerlies in the North Atlantic. SST anomalies associated with the NAO also appear in the extratropical Pacific that are only slightly weaker than in the Atlantic. It is noteworthy that NAO variability is not correlated with tropical Pacific SSTs, at least contemporaneously during Northern winter months. Herein lies a great dilemma that serves as the basis for a significant portion of contemporary

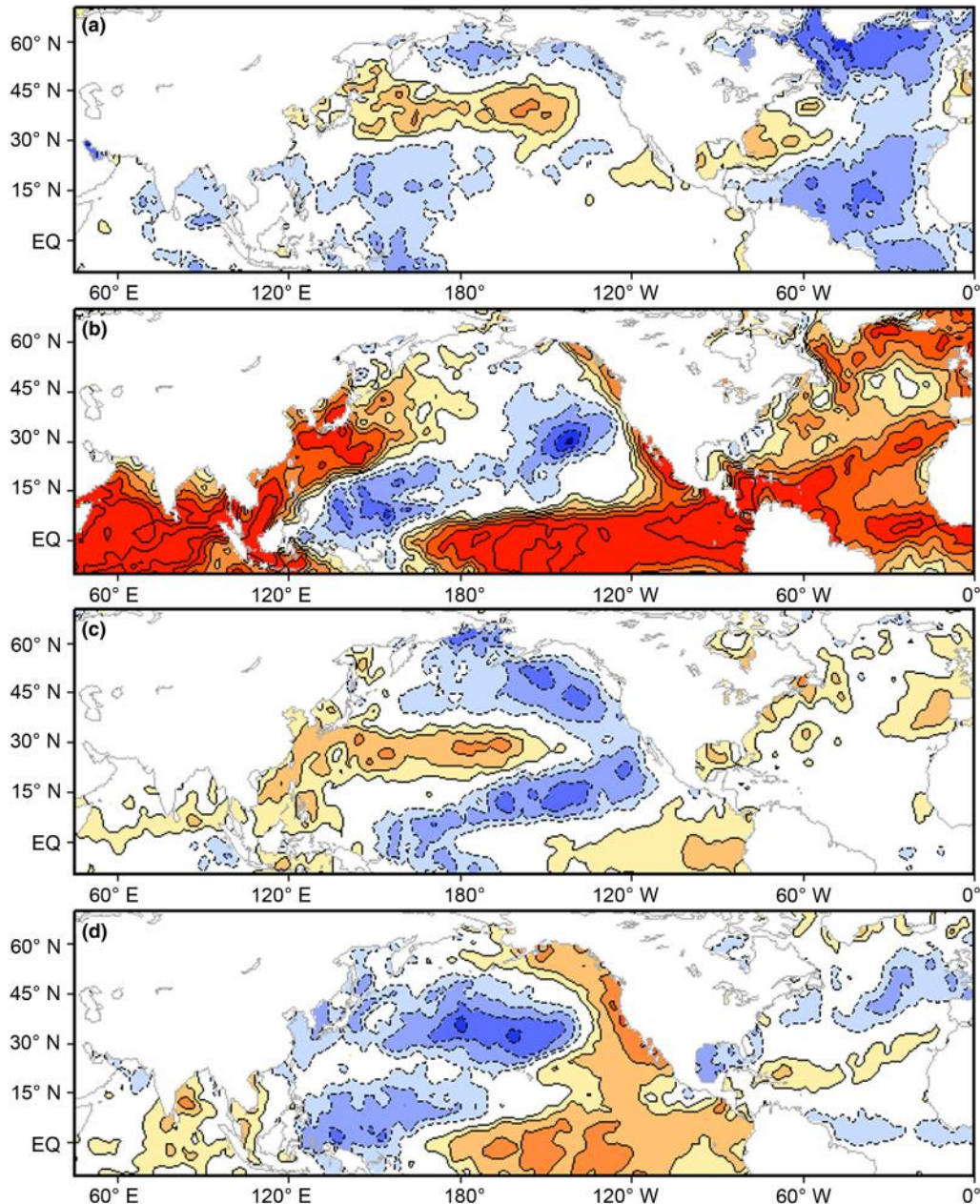


Figure 9 Correlation between the leading wintertime teleconnection patterns: (a) NAO, (b) ENSO related, (c) NPO/WP, (d) PNA, and SSA anomalies. Contour/shading interval is 0.1 m while the zero contour is suppressed. NAO, North Atlantic Oscillation; ENSO, El Niño-Southern Oscillation; NPO/WP, North Pacific Oscillation/West Pacific; PNA, Pacific-North America; SSA, sea surface temperature.

subseasonal-to-interannual climate research: ENSO, which is slowly varying and more predictable on seasonal timescales, explains relatively little hydroclimate variance in the Northern extratropics. The NAO, by contrast, explains more surface climate variance but is substantially less predictable.

Not surprisingly, the SST signature of the ENSO-related height pattern consists of large tropical correlations that resemble El Niño SSTs in the equatorial Pacific. Notable correlations are also present in the western Indian Ocean and the tropical and extratropical Atlantic. These include the well-documented ENSO links to warming of the adjacent basins, as well as capturing of more than just ENSO variability by this PC; including, possibly, the secular trend and multidecadal variability. The NPO/WP's SST correlations are not unlike NAO's Atlantic ones: a weak, tripolar, horseshoe-like structure in the extratropical basin; unlike NAO, however, it exhibits weak links with other basins.

The PNA SST correlations are broadly similar to the NPO's – an SST tongue extending from Japan into the midlatitude Pacific, surrounded by opposite-signed SSTs along the North American coast from the Aleutians to Baja Peninsula – but somewhat stronger, especially in the deep Tropics. A helpful distinction is the placement of the coastal anomalies: the off-coastal maximum in NPO/WP vis-à-vis the coastal maximum in PNA. The SST correlations of all three Pacific height patterns likely contain elements of the ENSO and Pacific decadal SST variability patterns, in view of the potential aliasing of SST variability due to the short analysis period (1979–2008) and the atmospheric-only nature of the PCA. An RPCA analysis of the combined variability of upper-tropospheric geopotential height and SST in a longer record minimizes such aliasing; see [Nigam \(2003\)](#) for more information.

The PNA SST correlations extend into the equatorial Pacific ([Figure 9\(d\)](#)), as noted above. This pattern was once thought to be the extratropical atmosphere's response to El Niño because of such correlations, despite the dominant intra-seasonal timescales of its PC (or teleconnection index) and the modest correlation ($0.3 \sim 0.4$) of the latter with the Niño 3.4 SST index. The potential aliasing of SST variability, noted above, led to PNA's spurious linkage with tropical SSTs: The PNA PC from the combined analysis of 200-hPa geopotential height and SST variations is correlated with the Niño 3.4 SST index at only 0.07.

The contemporaneous SST correlations of the recurrent 200-hPa height variability patterns raise the interesting question of cause and effect: Are the height patterns forced by the underlying SST patterns, or vice versa? The answer to this question is, unfortunately, complex as it is dependent on both the basin region and the variability pattern. Generally speaking, SST variations in the deep Tropics (e.g., ENSO related) are influential on the atmosphere on monthly timescales whereas variations in the middle–high latitude basins are the response. Consider the NAO: In the displayed phase (positive), the SLP pattern ([Figure 4\(c\)](#)) strengthens both the Icelandic Low and Azores High along with a northward displacement of the jet stream and storm track. This will result in anomalous westerly wind stress on the ocean northward of the climatological jet position, leading to increased vertical mixing and latent heat flux, and thus cooling of the surface

layer. The opposite arguments apply in the region off the US coast, where SSTs are warmer. The strengthened Azores High is, likewise, associated with stronger trade winds in the tropical Atlantic, leading to more upwelling and mixing off the African coast, and thus colder SSTs. As noted earlier, the influence direction is dependent on the basin region and variability timescales. The extratropical oceans can be influential on longer timescales and even on subseasonal ones through their preconditioning role. Clearly, this is an active research area in climate science requiring the use of advanced modeling techniques.

Northern Summer Teleconnections from RPCA

The Northern summer stationary waves are considerably weaker than the winter ones because of the weaker and northward displaced jet (and storm track). The summer regime thus consists of monsoonal circulations and only modest stationary Rossby wave propagation. For these reasons, the winter anomalies have been the target of most teleconnection analyses. The structure of recurrent height variability in Northern summer is however still of interest. A rotated EOF analysis (RPCA) of June, July, and August (JJA) 200-hPa geopotential height anomalies is conducted in a fashion similar to the wintertime analysis except for the analysis domain. The domain used in the summer analysis is restricted to $20\text{--}90^\circ\text{N}$. This change is made to emphasize the extratropical modes; inclusion of the Tropics leads to a noisier analysis.

The summer results are displayed in [Figure 10](#). The leading patterns are, perhaps, relatable to the winter patterns. The first one is labeled NPO/WP due to the presence of a north–south dipole over the North Pacific near the date-line. The second one is referred as the NAO on account of the dipole over the northeastern Atlantic. An ENSO-related pattern or a PNA look-alike was not identified. The third leading pattern exhibits higher geopotential heights in the global Tropics and a coherent ridge over the western sub-Arctic. The first four modes explain less variance than in winter, indicating the more disorganized structure of summertime variability. Notice that even the patterns resembling NPO/WP and NAO are more regionally confined, consistent with brief remarks on reduced Rossby wave activity in Northern summer; see [Folland et al. \(2009\)](#) for more discussion of the summer NAO, the most studied summer teleconnection.

Southern Winter Teleconnections from RPCA

The very different distribution of continents and mountains in the Southern Hemisphere (SH) results in muted seasonality and stationary waves. The SH winter climatology is dominated by a strong polar vortex, and thus more zonally symmetric flow. The lack of continents and large mountain ranges lead to reduced orographic forcing and land–sea contrast, both resulting in diminished stationary Rossby wave activity. An RPCA analysis of the SH winter height variability is briefly discussed here. It is identical to the Northern winter

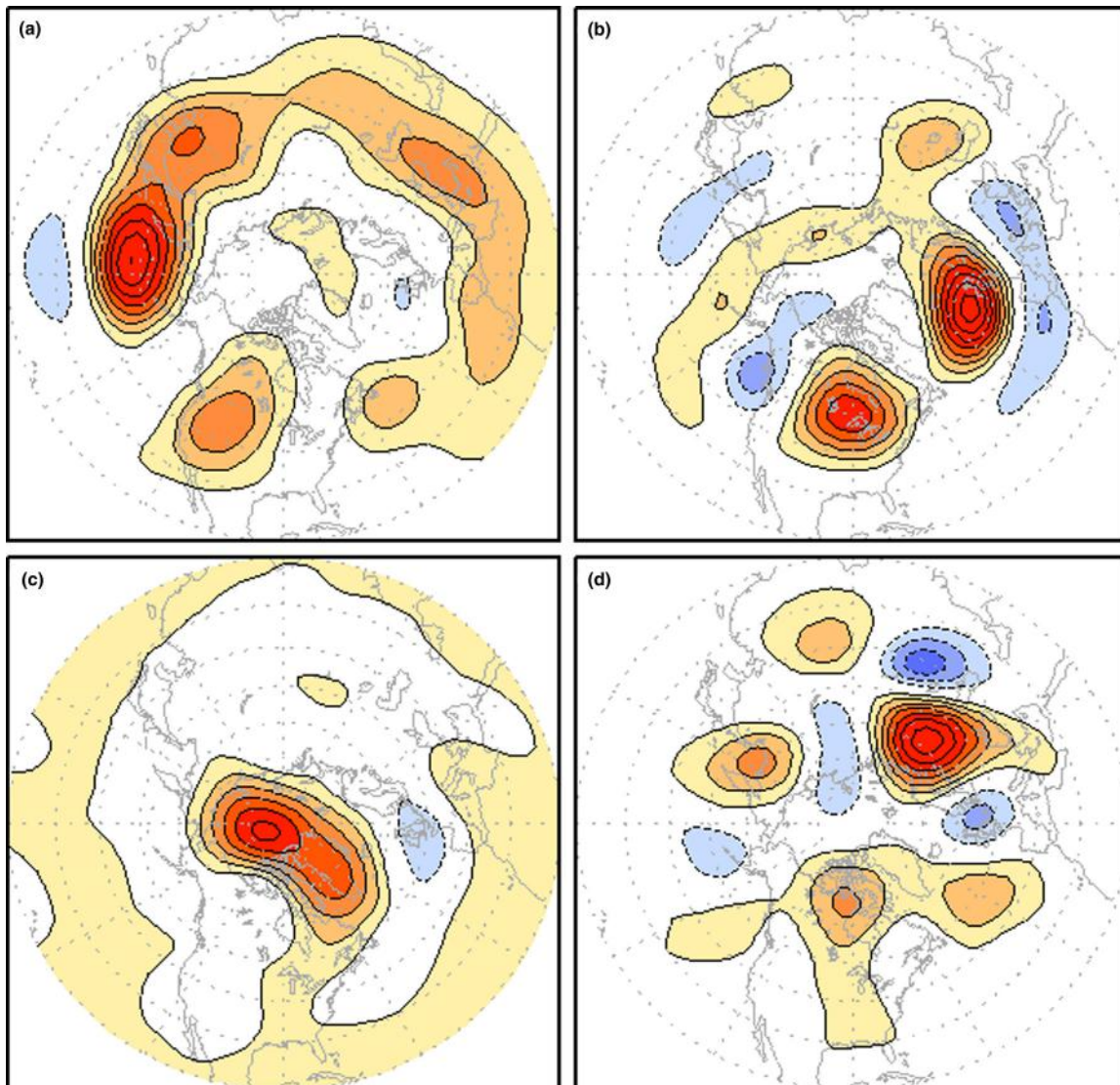


Figure 10 Four leading summertime (JJA) teleconnection patterns that emerge from a rotated EOF analysis of 200-mb geopotential height. The domain of the analysis is 20–90° N. Only the leading two modes, (a) and (b), are clearly identifiable as the NPO/WP and NAO, respectively. Contour/shading interval is 8 m; the zero contour is suppressed. Percentage of explained variance: (a) 11.9, (b) 8.5, (c) 8.0, (d) 7.4. JJA, June, July, August; EOF, empirical orthogonal function; NPO/WP, North Pacific Oscillation/West Pacific; NAO, North Atlantic Oscillation.

analysis: the domain includes the Tropics (30° N–90° S), and the eight leading patterns are rotated using the varimax criterion.

The four leading patterns are shown in [Figure 11](#). The first one is analogous to the NAO, but slightly more annular in nature. The larger amount of explained variance (~16.4%) relative to the other patterns indicates that SH variability is dominated by fluctuations of the polar vortex; not surprising, given the zonal nature of the SH circulation. Patterns 3–4 are more classic teleconnections: The third pattern consists of a strong center in the Bellingshausen Sea in the far South Pacific, surrounded by three weaker centers, while the fourth one is dominated by a wave-3 pattern around the pole. Patterns 1, 3, and 4 were identified in some form by [Mo and White \(1985\)](#), the first analysis of SH

winter variability; see this study for more details on SH teleconnections.

For reference, [Figures 12–14](#) show the surface temperature and precipitation footprints of select winter patterns of SH height variability; in the interest of space, only a few prominent features are noted below. Because of its polar domain, the first pattern has weak temperature and precipitation footprint over the inhabited continents. Pattern 2 ([Figure 12](#)) is linked with wide-spread warming of the Tropics, and more precipitation over Amazonia and southern South Africa, and a precipitation dipole over Australia (dry to the east, wet in the north); some of these signals are well documented in the literature as El Niño's SH winter impact. Pattern 3's precipitation impact is weak but it is linked to a prominent north–south temperature dipole over South America ([Figure 13](#)).

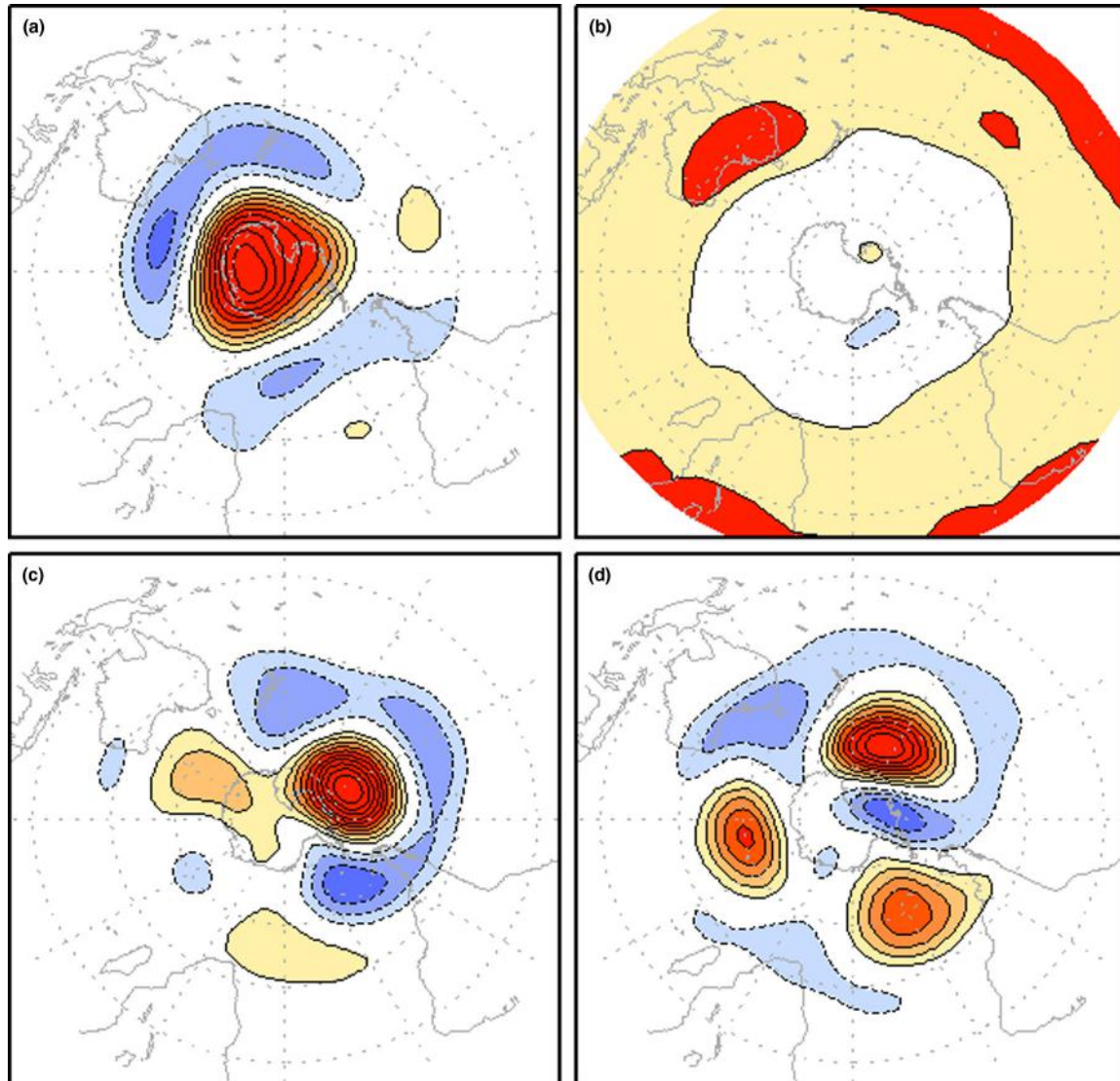


Figure 11 The leading patterns that emerge from a rotated EOF analysis of 200-mb height in SH wintertime (JJA). The domain used is 30° N–90° S. The contour interval is 10 m while the zero contour is suppressed. Panels (a)–(d) represent patterns 1–4. Percentage of explained variance: (a) 16.4, (b) 11.9, (c) 11.5, (d) 10.4. JJA, June, July, August; EOF, empirical orthogonal function; SH, Southern Hemisphere.

Although not expected from the height anomaly structure, mode 4 exerts substantial impact over Australia (Figure 14), with wet anomalies everywhere except the southeastern region.

Annular Modes

Closely related to teleconnections are the annular modes, which have received considerable attention in recent years. While teleconnections are the climate links between geographically separated regions, the annular modes represent the seesaw (i.e., out-of-phase links) between the poles and the surrounding midlatitudes; annular modes can be viewed as the teleconnections of the polar region. The annular mode in the NH and SH is referred as the Arctic Oscillation and

Antarctic Oscillation (AO and AAO), respectively. The annular pattern is obtained from EOF analysis of the monthly 1000-hPa geopotential height anomalies (700 hPa for SH) poleward of 20° for the entire year. Key differences from the previous RPCA analysis are the lack of rotation of the PCs and the search for recurrent patterns among all calendar month (and not just winter) anomalies. Note, unrotated analyses often generate more spatially extended patterns in order to maximize the explanation of variance, as seen in Figure 15, which also shows the temperature and precipitation impacts of the annular patterns. Note, ‘annular pattern’ rather than ‘annular mode’ is, perhaps, more reasonable terminology as only spatial (and not spatiotemporal) variability is analyzed for their extraction.

The annular modes bear close resemblance to the leading mode extracted from RPCA of 200-hPa heights in each

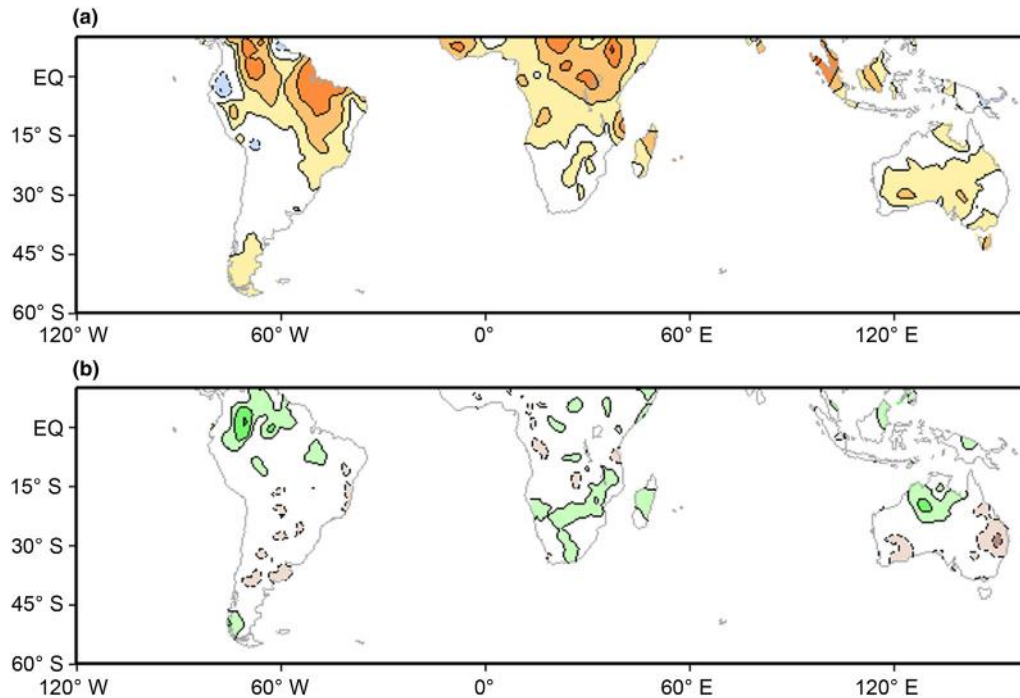


Figure 12 Same as in [Figure 3](#), except for the SH pattern 2. This is likely related to ENSO. ENSO, El Niño-Southern Oscillation; SH, Southern Hemisphere.

hemisphere. The AO has a stronger, coherent center of action in the Aleutians in addition to the essential elements of the NAO pattern. Not surprisingly, AO's year-round temperature and precipitation impacts are very similar to the winter impacts of the NAO (cf [Figure 4](#)). The close similarity of AO and NAO has led many to question whether the NAO is merely a region manifestation of the larger-scale annular mode (AO). It is noteworthy that neither the AO nor AAO are

strikingly annular; especially the AO. The AO footprint in the lower stratosphere (not shown) however appears substantially more annular; as does the NAO footprint (cf [Figure 8\(c\)](#)).

Subseasonal Tropical Variability – MJO

The ENSO-related pattern reveals one way in which the Tropics can influence the extratropics: An interannual shift in the preferred area of tropical convection/precipitation leads to anomalous divergent motions, which generate a Rossby wave source that, in turn, alters midlatitude storm tracks and extratropical climate. Zonal variations in tropical convection occur on timescales other than interannual ones as well: The MJO – representing convection variability in the tropical Indian Ocean and Western Pacific basins on 30- to 50-day timescales – can, like ENSO, influence the climate in the extratropics, albeit on subseasonal timescales.

When the MJO is active, convection is enhanced over the eastern Indian Ocean at the expense of the West Pacific. This generates an anomalous divergent circulation that leads to retraction of the East Asian jet. The jet retraction and related modulation of storm tracks influence the circulation and hydroclimate over the North Pacific and North American regions; for example, weaker Aleutian Low and higher than normal heights over eastern North America. These MJO influences can implicitly influence the NAO and PNA structure, but only modestly as these teleconnection patterns are robustly manifest in the MJO-filtered data as well.

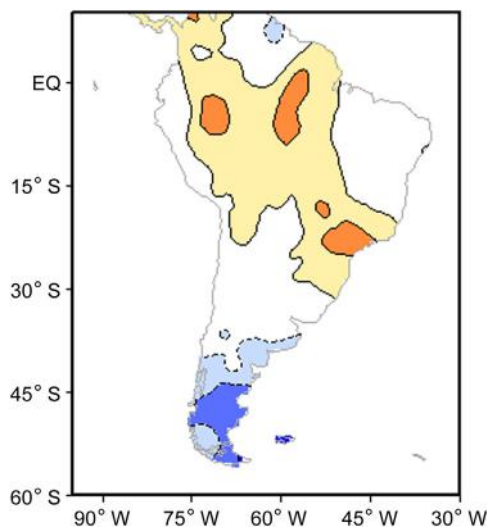


Figure 13 Similar to [Figure 12](#), except only showing the South American temperature footprint for pattern 3.

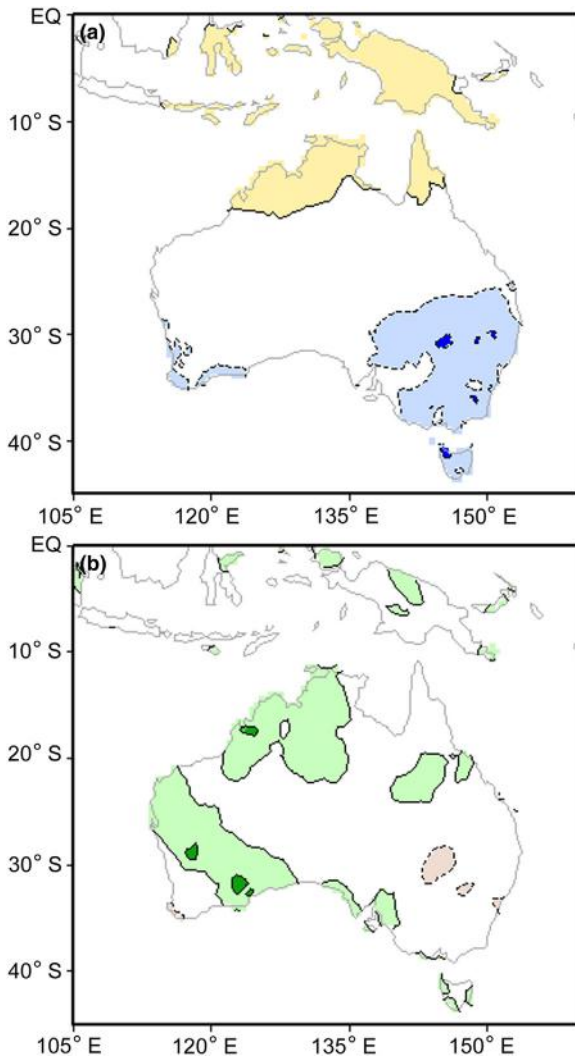


Figure 14 Similar to [Figure 12](#), except the temperature (a) and precipitation (b) footprints are shown only over Australia for pattern 4.

Subseasonal Teleconnection Analysis

Climate teleconnections, discussed above, were obtained from the analysis of monthly averaged anomalies, using techniques keyed to an optimal accounting of the temporal variance. These attributes steered the analysis toward extraction of the mature phase of the leading teleconnection patterns. Monthly averaging filtered the large-scale synoptic waves in the troposphere and facilitated focus on the ‘standing’ component of low-frequency variability. A month is not too long a period in context of large-scale ocean–atmosphere interaction, but it is an extended averaging period in context of the dynamical and thermodynamical processes generating climate teleconnections. These processes evolve rapidly – on pentad-to-weekly timescales – and their resolution is critical for the description of the nascent-to-mature and decay phases of the teleconnection patterns, and for advancing understanding of the excitation mechanisms. Insights into the evolution of teleconnection

patterns can advance the subseasonal prediction of the climate system.

In the following, the submonthly evolution of PNA variability is briefly discussed to illustrate an interesting line of contemporary teleconnection research. The PNA pattern is chosen as its evolution was described earlier from a simpler (and less accurate) analysis ([Nigam, 2003](#)), facilitating assessment of the new analysis strategy. The CFSR, used earlier at monthly resolution, is analyzed at pentad (5-day average) resolution in this section. The analysis period covers 30 extended winter (November–March) seasons from 1979 to 2008, and the 200-hPa geopotential height anomalies are analyzed, as before. To emphasize pattern development and decay, a technique called *extended* EOF analysis is used. Extended EOF analysis is a powerful spatiotemporal analysis technique, but one that can be easily implemented using the standard EOF analysis infrastructure; by substituting a contiguous *series* (e.g., from $T - 2$ to $T + 2$ time steps) of anomaly patterns in place of the contemporaneous one (i.e., $T = 0$). The leading variability mode in this case is not a single spatial pattern but the most recurrent temporal series of spatial patterns, or the most recurrent spatiotemporal pattern. In this illustrative example, a 5-pentad window is chosen ($T - 2$ to $T + 2$ pentads). The seven leading modes are rotated using the varimax criterion to allow for more spatial discrimination.

The leading mode from this spatiotemporal analysis represents the development and decay of the NAO (not shown). The fourth leading mode is closely related to the PNA; a 5-pentad evolution is shown in [Figure 16](#). An interesting finding here is that the development and decay of the PNA is associated with westward progression of height anomalies across the high latitudes. At $T = -3$, there is a notable height center in the NAO region of the North Atlantic. It is thus not surprising that there is a small but significant lag correlation between modes one and four ($r = -0.35$), with the negative NAO *leading* the positive-phase PNA pattern. The retrogression of the northern height anomalies is consistent with the advection of planetary vorticity, but cannot adequately explain the development of the robust anomaly center south of the Aleutians.

To help set forth a plausible basis for the NAO–PNA link, the impact of the NAO on the meridional wind is assessed; meridional winds highlight the shorter zonal scales of the midlatitude quasi-geostrophic flow, and thus the zonal propagation of stationary Rossby waves. [Figure 17\(a\)](#) shows contemporaneous regressions of the PC associated with the positive NAO pattern onto the 200-hPa meridional wind, along with the climatological winter position of the East Asian jet. The NAO apparently forces a wave pattern in the subtropics between East Africa and southern Asia that culminates in the North Pacific. [Figure 17\(b\)](#) shows the 2-pentad lag regression of the same NAO PC onto the 200-hPa zonal winds. The result is a subtle retraction of the East Asian jet, consistent with the negative-phase PNA pattern.

The analysis presented above serves to illustrate the nature of contemporary research in climate teleconnections rather than definitive findings on PNA evolution; durable and corroborated findings on teleconnection evolution have yet to

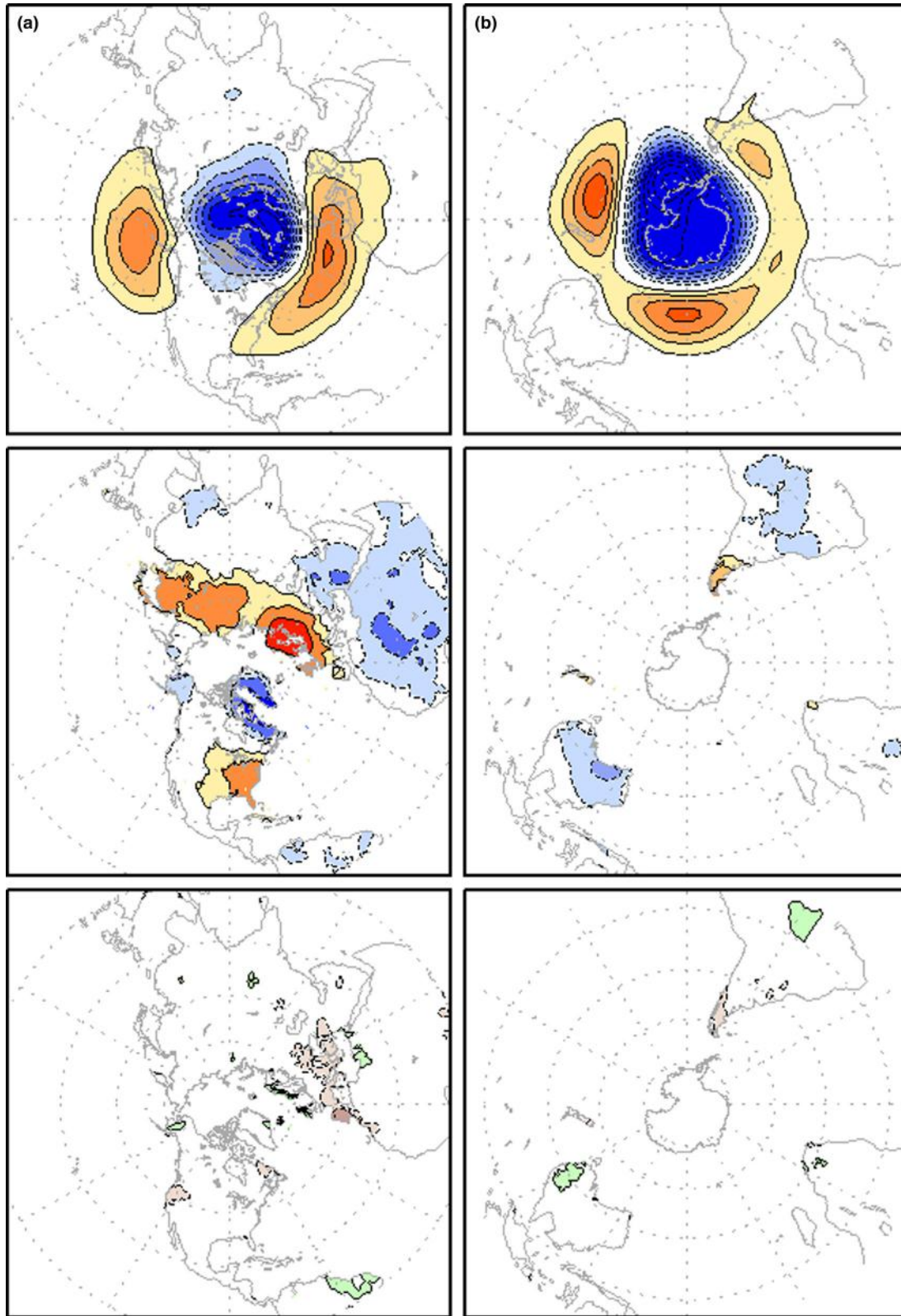


Figure 15 The two annular modes, (a) the Arctic Oscillation and (b) the Antarctic Oscillation (AAO), are shown along with their temperature and precipitation footprints. Annular modes are calculated as the leading EOF of 1000-mb geopotential height (700-mb geopotential height for AAO). Note that the entire year is used to calculate the pattern. Contour interval for the height patterns is 5 m. Hydroclimate footprints are shown as correlations between the principal components and CRU temperature and precipitation, contoured at 0.15 intervals. For all panels, the zero contour is suppressed. EOF, empirical orthogonal function.

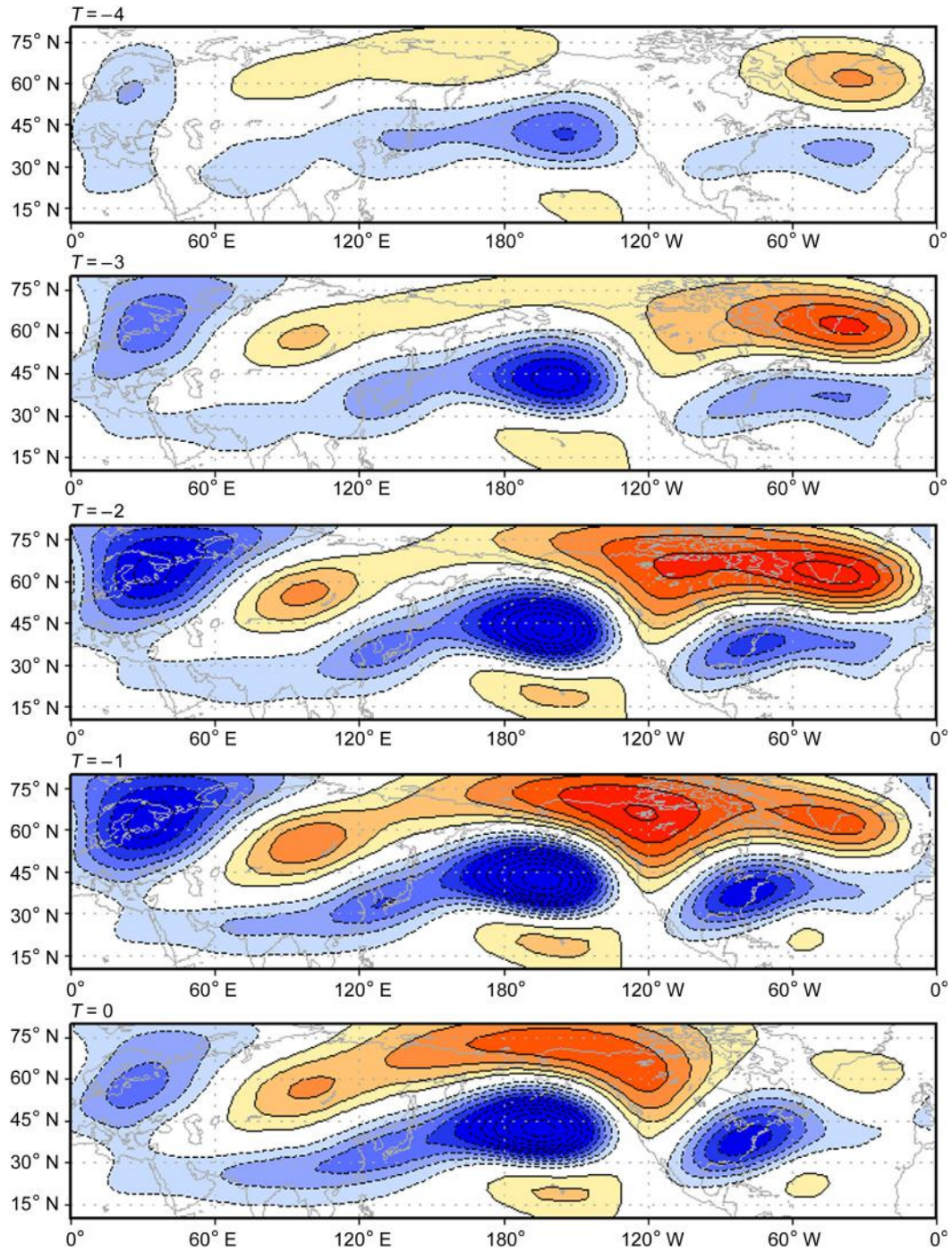


Figure 16 The fourth leading mode from a rotated, extended EOF analysis on extended winter 200-hPa height at pentad resolution. This mode closely resembles the PNA. The figure is presented as the lead/lag regression ($T = -4$ to $T = 0$) of the PC onto the height field. The contour interval is 10 m and the zero contour is suppressed. EOF, empirical orthogonal function; PNA, Pacific–North America; PC, principal component.

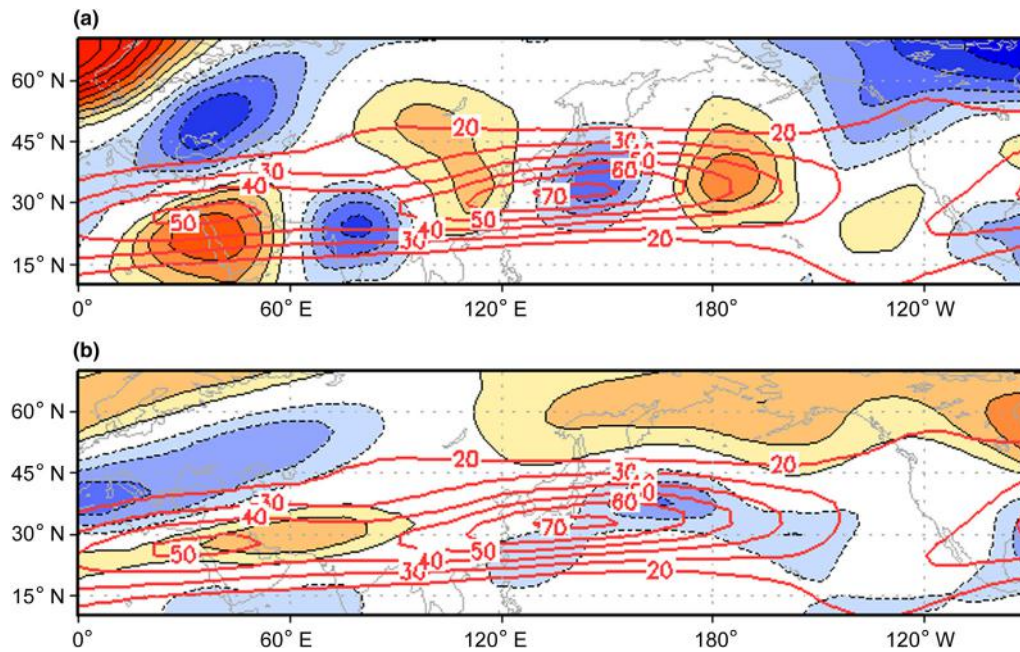


Figure 17 (a) The contemporaneous regression of the leading PC (NAO) onto 200-hPa meridional wind is contoured/shaded at 0.5 m s^{-1} intervals. The zero contour is suppressed. Bold red contours show the winter climatology of the 200-hPa zonal wind at 10 m s^{-1} intervals. (b) The 2-pentad lag regression of the leading PC (NAO) onto the 200-hPa zonal wind field is contoured/shaded at 1 m s^{-1} intervals with the zero contour suppressed. As in (a), bold red contours represent the winter jet climatology. NAO, North Atlantic Oscillation; PC, principal component.

emerge. The recent focus on submonthly climate prediction has however spurred climate scientists to investigate the development and decay of the leading winter teleconnection patterns.

See also: **Climate and Climate Change:** Climate Variability: Seasonal and Interannual Variability. **Dynamical Meteorology:** Rossby Waves; Stationary Waves (Orographic and Thermally Forced). **Stratosphere/Troposphere Exchange and Structure:** Global Aspects. **Tropical Meteorology and Climate:** Madden–Julian Oscillation: Skeleton and Conceptual Models.

- Folland, C.K., Knight, J., Linderholm, H.W., Fereday, D., Ineson, S., Hurrell, J.W., 2009. The summer North Atlantic Oscillation: past, present, and future. *Journal of Climate* 22, 1082–1103.
- Mo, K.C., White, G.H., 1985. Teleconnections in the southern hemisphere. *Monthly Weather Review* 113, 22–37.
- Nigam, S., 2003. Teleconnections. In: Holton, J.R., Pyle, J.A., Curry, J.A. (Eds.), *Encyclopedia of Atmospheric Sciences*. Academic Press, Elsevier Science, London, pp. 2243–2269.
- van den Dool, H.M., Saha, S., Johansson, Å., 2000. Empirical orthogonal teleconnections. *Journal of Climate* 13, 1421–1435.
- Wallace, J.M., Gutzler, D.S., 1981. Teleconnections in the geopotential height field during the northern hemisphere winter. *Monthly Weather Review* 109, 784–812.

Further Reading

Barnston, A.G., Livezey, R.E., 1987. Classification, seasonality and persistence of low-frequency atmospheric circulation patterns. *Monthly Weather Review* 115, 1083–1126.

GLOBAL CHANGE

Contents

Climate Record: Surface Temperature Trends

Sea Level Change

Upper Atmospheric Change

Biospheric Impacts and Feedbacks

Climate Record: Surface Temperature Trends

PD Jones, Climatic Research Unit, University of East Anglia, Norwich, UK

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

Surface temperatures have risen by about 0.6 °C during the twentieth century. This article addresses the quality of the basic temperature data over the terrestrial and marine domains. The warming during the century has not occurred in a linear fashion but in two periods, from about 1920 to 1945 and since about 1975. Spatial patterns of the change over the twentieth century indicate many regions showing statistically significant warming, but not all. Changes in temperature are also assessed in greater detail: showing that the warming is occurring more by reductions in cold extremes compared to increases in warm extremes and occurring more at night than during the daytime. The twentieth century warming is finally placed in a longer context by considering millennial-length paleoclimatic information from many diverse proxies. The latest evidence shows that the twentieth century has been both the warmest of the millennium and the warming rate during it has been unprecedented. The 1990s (1991–2000) was the warmest decade of the twentieth century and 1998 the warmest year. The first decade of the twenty-first century (2001–10) was warmer again, 0.20 °C above the 1990s, with 2010 almost as warm as 1998. The 10 warmest years are all the years from 2001 to 2010, but with 2008 replaced by 1998.

Quality of Temperature Data

Any assessment of trends or changes in temperature requires that all the observations have been taken in a consistent manner. Climatologists refer to this property as homogeneity. Time series of temperature are homogeneous if the variations exhibited are due solely to the vagaries of the weather and climate. Numerous nonclimatic factors influence the basic data. Without some form of adjustment, erroneous conclusions can be drawn regarding the course of 'true' temperature change. The factors vary depending upon the data source and are briefly considered in the next two subsections for the terrestrial and marine components of the Earth's surface.

Land

It is extremely rare if observational protocols and the environment around an observing location have remained exactly the same during the station's history. Changes are likely to have occurred with the instruments, their exposure and measurement techniques, in the location of stations and the height of the instruments, in the number and times of observations per day, and the methods used to calculate daily and monthly

averages. The two most important factors with respect to long-term consistency are changes brought about by the introduction of Stevenson screens and changes in the environment around some of the stations across the world. Both factors are difficult to deal with using conventional approaches of station homogeneity assessment, as neighboring sites are all likely to be similarly affected.

The commonly used louvered screen developed by Stevenson in the 1870s is now the standard around the world, although different countries use variants of a similar design. Prior to this, most thermometers were positioned on poleward-facing walls (i.e., out of direct sunlight), but this poses problems in high-latitude regions in the summer. The issue of the different exposures before screens has recently begun to be addressed by rebuilding the early exposures (using nineteenth century information) and taking modern parallel observations. Comparison of these measurements confirms expectations and indicates that prescreen temperatures are about 0.5 °C too warm during the summer months from May to September. Winter temperatures are barely affected by the change. Only the longest series are affected, but countries or regions that introduced screens later (e.g., Australia in the 1910s) have yet to fully allow for the effects in the prescreen parts of their series.

The second factor is that the location may have been a small town in the nineteenth century, but now could be a city of several million. Development around the site (urbanization) leads to relative warming of city sites compared to, still rural, neighbors. On certain days, particularly with anticyclonic weather, cities can be warmer than rural surroundings by up to 10 °C. For monthly averages, this reduces to up to 2 °C, more so for inland continental, compared to coastal, locations. Cities that have grown rapidly over the twentieth century tend to be more affected, compared to European locations where development has taken place over many centuries. Assessment of the urbanization influence suggests that the overall influence (at hemispheric scales) is small (up to 0.02 °C per decade).

Additional factors influencing homogeneity are that most stations have moved at least once during their lifetime. Also, of importance is the time observations that are made each day. Even today, there is no accepted standard, and countries are allowed to choose whatever times suit them. English-speaking countries have tended to use the average of the daily maximum and minimum readings each day to measure daily and monthly averages. Some countries have switched to this method mainly because of its ease, while others retain their national standards (averages of measurements made at fixed hours, between 3 and 24 times per day). Changes to sites or to the methods used to calculate monthly averages influence the time series, often in an abrupt manner (temperatures changing to a new level by up to 2 °C in extreme cases). Ideally, when new sites or observation protocols are adopted, parallel measurements are recommended, enabling corrections to be calculated. These corrections are referred to as homogeneity adjustments. Sadly, although clearly recognized as being necessary, few countries carry out sufficient overlapping measurements. The most common problems relate to location moves, particularly to airports in the 1940s and 1950s.

Recently, many countries have switched from mercury-in-glass thermometers to electrical resistance thermistors, to reduce manpower, automating measurements. The sum total of all these problems can be disentangled, if adequate station history information is available, but it is generally a tedious process locating all the necessary information. In some countries, it is just not available in sufficient detail. Site moves and changes to observation protocols are less important than the widespread changes to screens and changes in the environment around a station as the effects can be of both signs and occur at irregular points in time. The overall effect on large-scale averages is only important if many sites are affected by changes occurring at the same time (such as the introduction of screens and urbanization issues). Homogeneity adjustments are necessary and vital for local scales, but are relatively unimportant at the hemispheric and global scales.

Several groups in the UK, USA, Russia, and Japan have extensively analyzed the basic surface temperature data (between two and seven thousand stations), adjusting the data for the abrupt changes and removing urban-affected stations, and have reached similar conclusions about the course of temperature change over the instrumental period since 1850. It is highly unlikely that every problem has been corrected for, but the different techniques used, give confidence that large-scale changes over the last 160 years are both real and well-documented. The agreement on large-scale trends with marine

and lower tropospheric temperature estimates (see later discussion of satellite records) confirms this confidence.

Marine

Terrestrial parts of the world constitute only 30% of the Earth's surface, so for a global picture it is vital to include the oceans. Historical temperature data over marine regions are largely derived from *in situ* measurements of sea surface temperature (SST) and marine air temperature (MAT) taken by ships and buoys. To be of use, each measurement must be associated with a location. Up to 15% of marine data are thought to be mislocated (ships located on the land!) and these values must be discarded. It is obviously harder to reject data still located over the ocean, but all analyses of the raw data also attempt to remove or correct these problems. Plotting ship tracks has helped to considerably reduce the numbers of mislocated reports.

Marine data are also beset with homogeneity problems, but they are distinctly different from the terrestrial realm. For MAT data, the average height of ships' decks above the ocean has increased during the twentieth century, but more importantly, daytime measurements are influenced by the solar heating of the ship, restricting use, at present, to only the nighttime MAT (NMAT). For SST data, the changes in sampling method from uninsulated canvas buckets (generally prior to the early 1940s) to engine intake measurements (early 1940s onward) cause an artificial rise in SST values of 0.3–0.7 °C. Recently, the greater preponderance of buoy (as opposed to ship) SST data has led to a further need for adjustments. Data from buoys were rapidly used when they became available, as coverage was dramatically improved. Now with over 20 years of overlap in measurements it is becoming apparent that the absolute temperatures from buoys are slightly cooler (0.1–0.2 °C) than those taken by ships. Estimates of SST from satellites are also beginning to be widely used, but these can be offset from *in situ* measurements by several degrees.

In the combination of marine data with land-based surface temperatures, SST data are preferred to NMAT, because they are generally more reliable, principally, as there are at least twice as many observations, daytime MAT values having been contaminated by the ships' infrastructure. Additionally, the much stronger day-to-day correlation of SST compared to MAT means that averages of a few SST values are much more reliable than comparable averages of MAT data. Absolute values of SST and land air temperatures may differ by up to 10 °C near some coastlines, so the two cannot be directly combined. Instead, anomalies are used (departures or differences from average) assuming that anomalies of SST and MAT agree on climatological (monthly and greater) timescales. Correction of the SST data for the change from canvas buckets is achieved using a physical-empirical model to estimate the degree of seawater cooling that occurs in buckets of varying designs. The cooling depends on the ambient weather conditions, but this can be approximated by climatological averages. Corrections are greatest in regions with the largest air-sea temperature differences (i.e., winters compared to summers) and the technique minimizes residual seasonal cycles in pre-World War II (pre-WWII) SST values compared to post-1945 values.

Since both the marine and the land components are independent, the two records can be used to assess each other after they have been separately corrected. The components have been shown to agree with several groups on both hemispheric scales using island and coastal data.

Aggregation of the Basic Data

Both the land and the marine data are irregularly located over the Earth's surface. To overcome the greater density of data in some regions, it is necessary to interpolate the data generally to some form of regular latitude/longitude grid.

Land

Differing station elevations and national practices with regard to the calculation of monthly mean temperatures means that interpolation to a regular grid is much more easily achieved by converting all the monthly data to anomalies from a common reference period (often referred to as a climatology). The period with best available data is 1961–90. The simplest interpolation scheme is the average of all stations that are located within each $5 \times 5^\circ$ grid box. More complex interpolation methods yield essentially the same results on all spatial scales. A potential drawback of gridding schemes is that the variance of grid box time series is affected by changing numbers of stations within each grid box through time, although it is possible to allow for this.

Marine

For SST, the aggregation is approached in a somewhat different manner. The changing location of each observation means that it is necessary, by interpolation, to derive the 1961–90 climatology for each 1×1 square of the world's oceans for each 5-day period (pentad). SST anomaly values with respect to this climatology are then averaged together for each month for each 5×5 grid box, the same as used for the land component.

Combination into One Dataset

Combination of the two components occurs in the simplest manner. Anomaly values are taken from each component. They are combined using weights determined by the errors of estimates of the land and the marine part. Around coastal areas this gives greater weight to the marine part.

Hemispherical Global Time Series

With the basic data now in 5° latitude/longitude grid boxes, calculation of large-scale averages is relatively simple but must take into account the different sizes of grid boxes in tropical, compared to polar, latitudes. This is simply achieved by weighting each grid box by the cosine of its central latitude value.

Figure 1 shows annual hemispheric and global time series for the 1850–2010 period using the HadCRUT3 dataset. The series derived by the other groups are very similar. Table 1

gives monthly linear trend values, estimated by least squares, for the three domains calculated over the 161-year period and for some other subperiods (1901–2010, 1920–44, and 1975–2010). For the 1901–2010 period, global average surface temperature has risen by 0.83°C , a value that is statistically significant at the 99.9% level. All the monthly values also exhibit a significant warming. The choice of periods such as 1925–44 and since 1975 is determined by looking at Figure 1, so an element of subjectivity could have influenced the choice. Analysis of all possible periods longer than 10 years has been considered in one study. Since 1945, all periods longer than 22 years indicate warming but are only statistically significant for periods ending after about 1990. There is also a segment of significant warming for periods ending in the mid-1940s. All periods longer than 82 years all produce positive trends.

While both hemispheres show similar degrees of warming, it is also apparent that many warm and cool years, relative to the underlying trend, are in common. Many anomalous warm years are coincident because they relate to El Niño years in the eastern equatorial Pacific. El Niño events cause somewhat predictable patterns of temperature and precipitation patterns over the world, with more regions experiencing warmer than cooler conditions. The opposite phase of an El Niño is termed a La Niña event. Anomalous patterns of temperature and precipitation also occur here, which to a first order are opposite to those during an El Niño event. Polar regions and much of northern Eurasia are largely unaffected by such an influence. A commonly used measure of the El Niño or La Niña state of the atmosphere is the normalized pressure difference between Tahiti and Darwin, Australia (referred to as the Southern Oscillation index (SOI)). Figure 2 is a scatter plot of the residual annual global temperature averages (the difference between each annual value and the smoothed curve) and the SOI for the 12-month average from July of the previous year through to June of the present year. The SOI explains about 30% of the variance of these residual temperatures. This analysis indicates that the next warm year will likely occur when the next El Niño occurs.

A few cool years can be related to the climatic effects of explosive volcanic eruptions, which are large enough to put considerable amounts of dust into the stratosphere. Once there, the dust forms a veil over the Earth, reducing solar radiation and cooling the surface, particularly land areas. Surface cooling of about $0.2\text{--}0.3^\circ\text{C}$ followed the eruption of Mt Pinatubo in the Philippines in June 1991, mainly in the northern summer months of 1992 and 1993. Volcanic eruptions, which only affect the troposphere (e.g., Mt St Helens in 1980), have little climatic effect as their ejecta are quickly dispersed by rain-making processes. The next major explosive volcanic eruption in the tropics is likely to cool global temperatures for 2–3 years following the eruption.

Accuracy of the Hemispheric and Global Series

The series in Figure 1 are subject to three sources of error: reductions in coverage earlier in the record; errors associated with the necessary bias adjustments in the basic data (discussed earlier for marine and land data, principally the switch to engine intake measurements in the early 1940s over the oceans

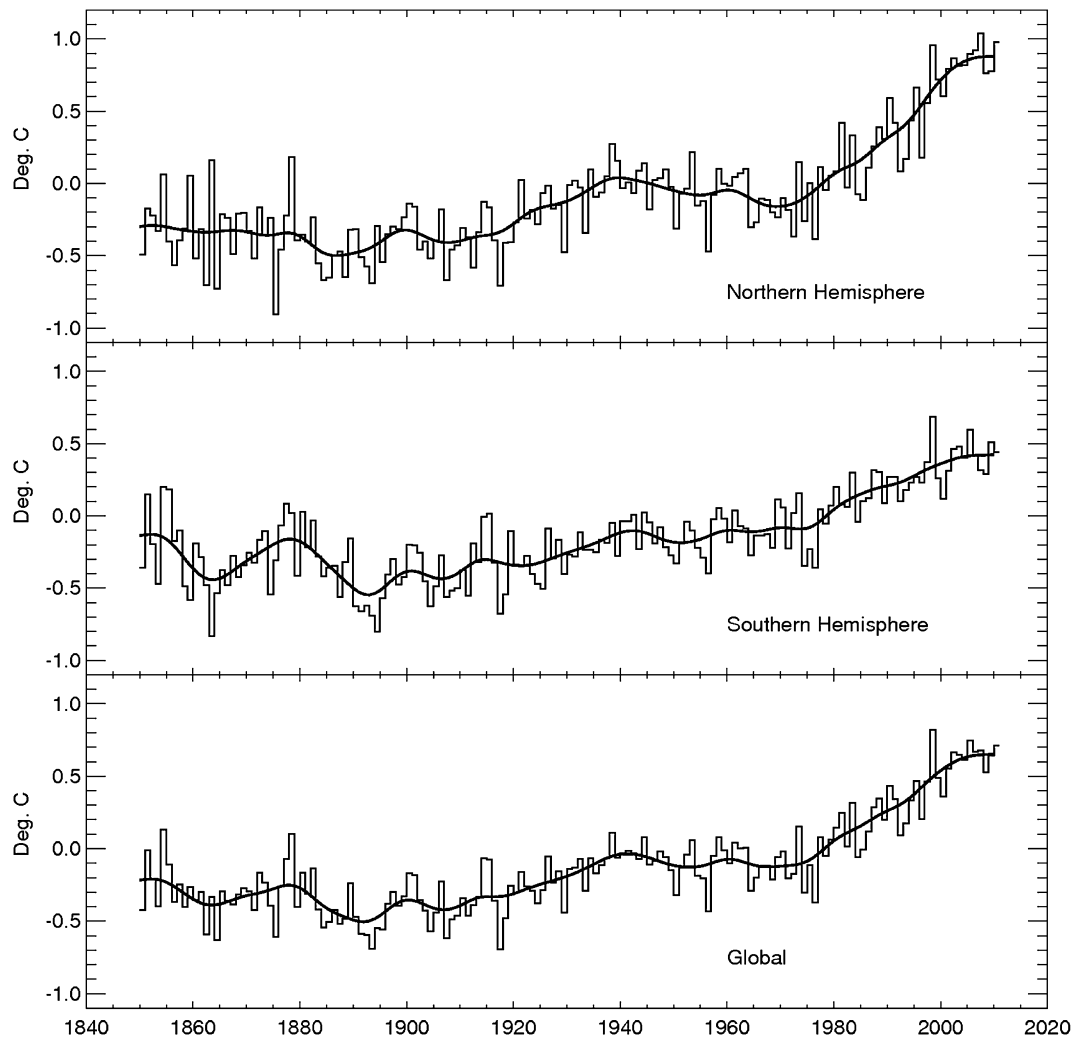


Figure 1 Hemispheric and global temperature averages on the annual timescale (1850–2010 relative to 1961–90). The smooth curves highlight variations on 20-year timescales.

Table 1 Temperature change (°C) explained by the linear trend over four periods: 1850–2010, 1901–2010, 1920–44, and 1975–2010

	1850–2010			1901–2010			1920–44			1975–2010		
	NH	SH	Globe	NH	SH	Globe	NH	SH	Globe	NH	SH	Globe
Jan.	0.77	0.69	0.73	0.81	0.80	0.80	0.26	0.55	0.40	0.65	0.44	0.55
Feb.	0.88	0.66	0.77	0.99	0.80	0.89	0.63	0.35	0.49	0.79	0.45	0.62
Mar.	0.97	0.71	0.84	1.03	0.84	0.94	0.25	0.35	0.30	0.82	0.50	0.66
Apr.	0.75	0.72	0.74	0.93	0.80	0.87	0.51	0.36	0.43	0.75	0.51	0.63
May	0.64	0.77	0.71	0.84	0.85	0.84	0.40	0.44	0.42	0.73	0.42	0.58
Jun.	0.50	0.82	0.66	0.82	0.84	0.83	0.38	0.51	0.44	0.80	0.47	0.64
Jul.	0.50	0.73	0.61	0.77	0.86	0.82	0.47	0.62	0.55	0.88	0.45	0.66
Aug.	0.58	0.71	0.65	0.78	0.83	0.80	0.49	0.48	0.48	0.91	0.46	0.69
Sep.	0.64	0.71	0.67	0.71	0.81	0.76	0.52	0.36	0.44	0.79	0.41	0.60
Oct.	0.78	0.71	0.74	0.75	0.83	0.79	0.62	0.44	0.53	0.91	0.44	0.68
Nov.	0.95	0.60	0.78	0.81	0.80	0.80	0.45	0.25	0.35	0.85	0.35	0.60
Dec.	0.84	0.65	0.74	0.91	0.78	0.84	0.50	0.44	0.47	0.67	0.35	0.51
Year	0.73	0.71	0.72	0.84	0.82	0.83	0.46	0.43	0.44	0.80	0.44	0.62

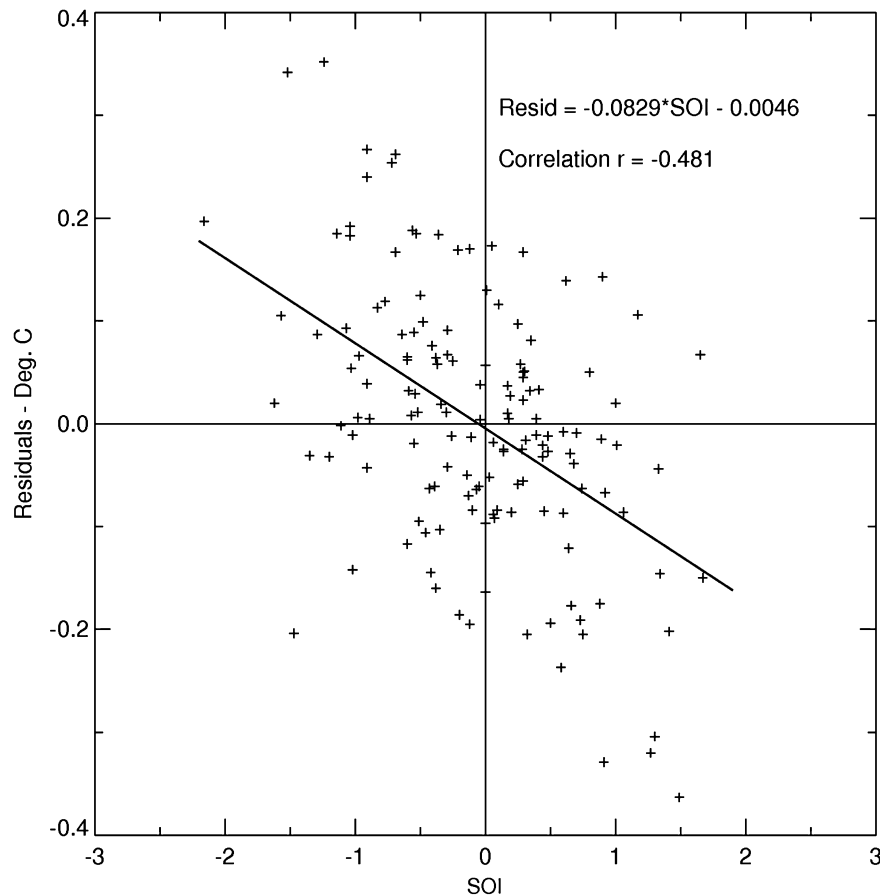


Figure 2 The inverse relationship between the SOI and the residual global temperature (annual value minus the smoother curve seen in [Figure 1](#)). The SOI is the normalized sea level pressure difference between Tahiti and Darwin. Residual temperatures are for the calendar year, while the SOI is the average of July of the previous year to June of the current year. Also shown is the regression line (based on annual values for the years 1867–2010) and the correlation between the two series. The regression line slope of ~ -0.8 means that a change of one SOI unit affects global temperature by $\sim 0.08^\circ\text{C}$. Very strong El Niño events can be up to 0.16°C warmer and very strong La Niña years up to 0.16°C cooler.

and the introduction of screens in the nineteenth century and effects of urbanization around some sites for the land); and homogeneity adjustments to land records for site and procedural changes. These are all taken into account in the error estimates and annual hemispheric averages are accurate to within $\pm 0.05^\circ\text{C}$ (one standard error). Errors in the mid-nineteenth century were roughly twice modern values.

Analyses of the Temperature Record

The surface record has been extensively analyzed, principally over the past 35 years. The series in [Figure 1](#) has become one of the foremost series in major international reviews of the climate change issue, most recently by the Intergovernmental Panel on Climate Change (IPCC). Here several diverse aspects of the record are analyzed:

- trends in areas affected by monthly extremes;
- trends in maximum and minimum temperatures;
- daily extremes of temperature in two long European series; and
- the last 150 years in the context of the last 1000 years.

[Figure 1](#) clearly shows recent warming since the late 1970s. Five different groups monitor surface temperatures and all show similar courses of change over the last 160 years and similar rates of change to those given in [Table 1](#). Completely independent estimates of temperature change have been developed for the lower part of the troposphere, first from weather balloons (called radiosondes) from the 1940s and more recently by satellite estimates from microwave sounding units aboard polar-orbiting satellites (since the late 1970s). [Figure 3](#) shows a comparison at the global scale, of the two principal groups who collate the satellite data, for the period from 1979 to 2010. The surface temperature indicates a warming of 0.16°C per decade, in exact agreement with one of the satellite records. The difference between the two satellite records relates to the adjustments to their records that must be applied to derive a consistent record over the 32 years. [Figure 3](#) also illustrates the influence of the SOI on global temperatures with the major El Niño event of 1997/1998 and the slightly lesser ones during the 1980s and during 2009/2010. Variability from month to month is markedly greater for the lower troposphere than at the surface, particularly during the larger El Niño and La Niña events.

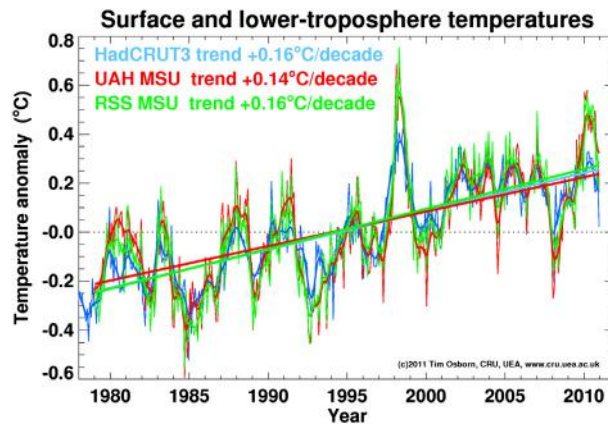


Figure 3 Comparison of monthly surface and lower tropospheric temperatures for the period 1979–2010. All series have been re-zeroed to the period 1979–2009. The trends given are based on all the years shown from 1979 to 2010.

Trends in Areas Affected by Monthly Extremes

The various groups that produce global and hemispheric temperature series also produce spatial temperature maps showing regions that were cooler or warmer than the 1961–90 base period. Similarly, it is possible to produce maps of temperature trends over specified periods. Interpreting these maps is often difficult as they tend to be dominated by the local level of variability. For example, the largest anomalies for a given month or year tend to occur in polar regions. Warming in Siberia may be large, but it is barely significant because of the large year-to-year variability. Trends in some tropical oceanic areas indicate highly statistically significant warming, but it may only be about 0.3–0.7 °C. To enable easier intercomparison of trends and extremes, from a local impact point of view, removing the effects of year-to-year variability highlights where significant changes are occurring. An appropriate transformation is the gamma distribution. Each grid box time series, on a monthly basis, is transformed from one in anomalies with respect to 1961–90 to percentiles based on the same period. Percentiles can be easily related to return periods (e.g., the 5th/95th percentiles are equivalent to the one in the 20-year return period). Using a normal distribution (i.e., simply dividing the grid box anomaly series by the standard deviation calculated over the 1961–90 period) works almost as well as the gamma distribution, but the latter is better in many regions of the world as monthly temperatures are often significantly negatively skewed.

Figure 4 compares the anomaly and percentile method for displaying annual temperatures for 2010. The zero anomaly and the 50th percentile contour are essentially the same in both plots. The percentile map, however, indicates extremely warm annual temperatures over many tropical and oceanic regions that might not warrant a second glance in anomaly form. 2010 shows 34% of the world's surface with data above the 90th percentile and 3% below the 10th percentile. How unusual is this, compared to other years? Figure 5 shows the percentage of the world's surface with data with temperatures greater than the 90th percentile (in red) and less than the 10th percentile (in blue) since 1900. An increase in the percentage of the analyzed area with warm extremes is evident (the largest area being 35%

in the warmest year (1998)), but by far the greatest change is a reduction in the percentage of the analyzed area with cold extremes. Some caution should be exercised while interpreting these results because of the large changes in coverage, particularly before 1951. The implicit assumption being made is that the average of the unsampled regions is the same as the average of the sampled regions. Coverage changes since 1951 are minimal, though, and even analyzing only those regions with data for the 1900–20 period produces similar series to those seen in Figure 5. The implications of these series are that before the mid-1970s most of the warming in this century was more apparent through less cold annual averages than excessively warm ones. Over the last 35 years, regions experiencing very warm annual anomalies have begun to increase dramatically.

Trends in Maximum and Minimum Temperatures

Up to now, all the surface temperature analyses have been based on monthly mean temperatures. This situation has arisen due to the widespread availability of this variable. As mentioned earlier, English-speaking countries have tended to measure daily and monthly means using maximum and minimum temperatures. Recently, extensive datasets of monthly mean maximum and minimum temperatures have become available since the 1950s. These enable recent warming patterns to be assessed for both day (maximum) and night (minimum) temperatures. The difference between day and night (the diurnal temperature range (DTR)) should prove a useful variable when considering what the causes of changes might be due to.

Homogeneity of the series poses more severe problems than for mean temperatures, as the various factors discussed earlier generally cause differential effects in the maximum and minimum series and station history information is even more important to decide upon adjustments. Analyses are restricted to the period 1950–2004 because of data availability issues in many regions of the world. Combining all available land regions, 'global' minimum averages warmed by 1.12 °C over the 55 years, while maximums warmed by only 0.78 °C. The DTR decreased by 0.34 °C. Most of the differences in warming rates occurred over the period from 1950 to about 1980. Over the last 25 years, trends have been similar in the two series.

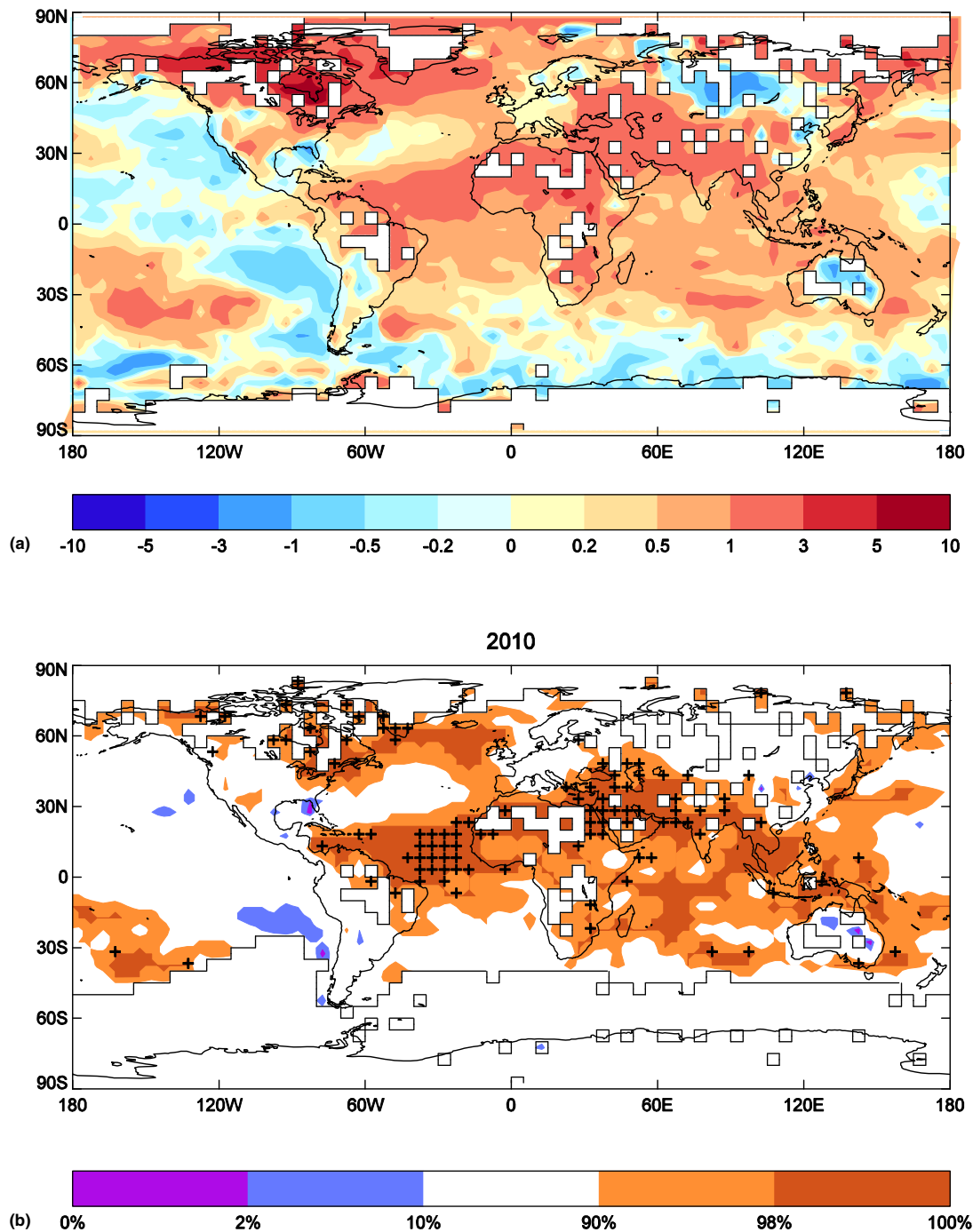


Figure 4 Surface temperatures for 2010, relative to the 1961–90 average (a) as anomalies and (b) as percentiles. The percentiles were defined by fitting gamma distributions to the 1961–90 annual deviations relative to the 1961–90 base period for all 5×5 grid boxes with at least 21 years of annual data in this period.

Urbanization influences have been shown to have the same signature (warmer nights compared to days), so these studies have restricted analyses to nonurban stations. In most regions, however, these differential trends can be clearly related to increases in cloudiness which will raise nighttime, compared to daytime, temperatures. Longer records, back to the turn of the twentieth century, are available in a few limited regions. Analyses over the USA and southern Canada, for example,

show little change over the first half of the twentieth century, so the drop in DTR over the period 1950–80 is the main feature of the record.

Daily Temperature Extremes in Long European Series

The last two sections have considered extremes on a monthly basis, but public perception of climate change is often

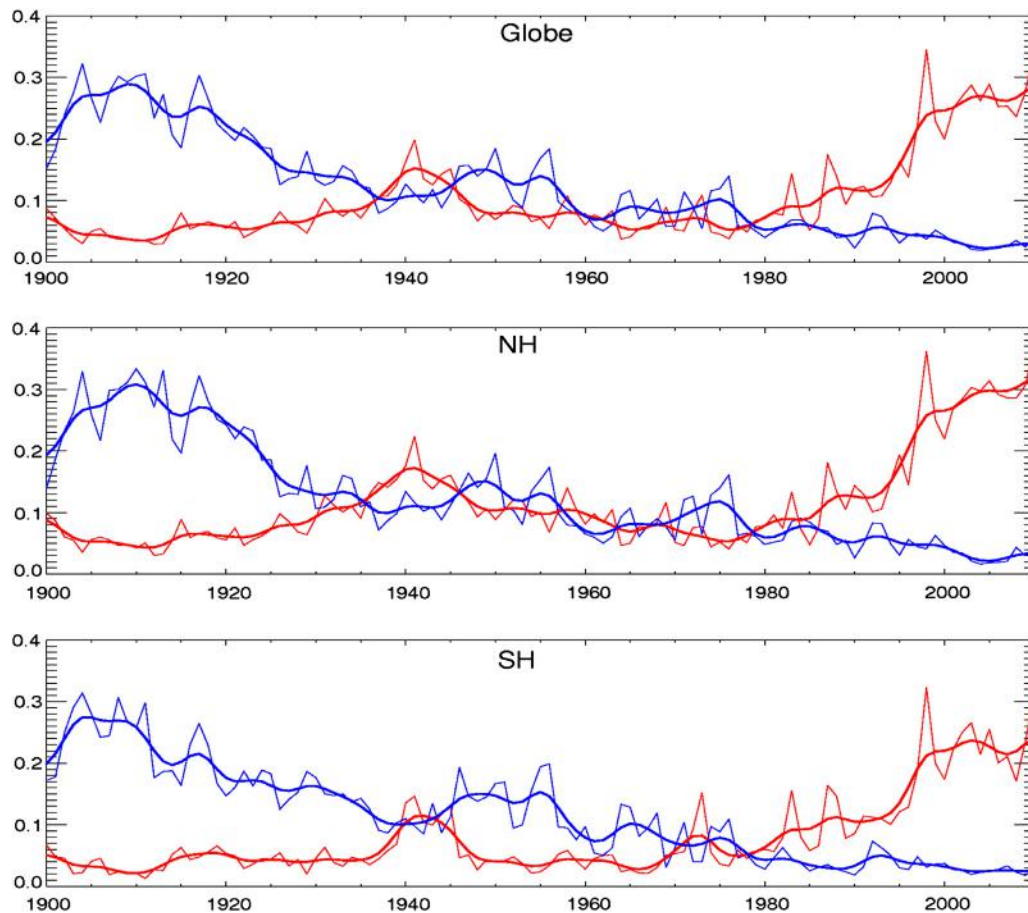


Figure 5 Percentage of the monitored area of the globe having monthly surface temperatures above the 90th percentile (in red) and below the 10th percentile (in blue). Annual averages from these separate monthly analyses were averaged to get the annual values. The percentiles were defined by fitting gamma distributions to each month based on the 1961–90 period, using all 5×5 boxes with at least 21 years of data. The smooth curves highlight variations on decadal timescales.

influenced by daily extremes or runs of warm/cold days (e.g., heat waves or cold surges). In the context of the global warming issue, daily data are relatively unimportant as detection and attribution of human influences are primarily concerned with underlying trends on decadal timescales. In the public and political worlds though, changing frequencies of daily extremes are that how much of the global warming debate is perceived.

Daily temperature series present even greater problems for climatologists with respect to homogeneity than monthly data. Site and observation time changes are particularly important and in some cases it may not be possible to fully correct for all the problems. Few long daily temperature series, therefore, are totally homogeneous. Furthermore, the availability of long series in some parts of the world is often restricted to the last 50 years because earlier data have not been digitized (particularly in some developing countries of the world). Changes in the frequency of extremes may be occurring, but without long series it is difficult to judge whether recent changes are really unprecedented. In Europe, however, several 200+ year series have recently been developed which will be ideal for analysis.

The public perception of extremes is clearly cold winter and hot summer days, but in different regions it is necessary to define somewhat arbitrarily what is meant by cold and hot.

A cold day threshold of 0°C clearly has important consequences but what is hot in northern Europe clearly differs from what would be regarded as hot in southern Europe. Also, considering only absolute extremes ignores changes that might be taking place in the transition seasons. A better and universally applicable means of defining extremes is to let the data define the thresholds and to allow these to change from place to place and during the year. A number of groups over the recent decade have developed different sets of indices of extremes. These have recently been combined into a large set by the Expert Team on Climate Change, Detection and Indices (ETCCDI, <http://www.clivar.org/organization/etccdi/etccdi.php>). Here, a relatively simple example is presented, which is applied to three long European series.

The first step is an analysis to define the annual cycle of temperature on a daily basis, based on a common period such as 1961–90. Some smoothing of this cycle is necessary as 30 years is a relatively short period for definition. 1961–90 is chosen for compatibility with the other analyses in this section. Variability of a single day's temperatures from the annual cycle shows greater variability in Europe during winter compared to summer. Also most station data series throughout the year, but particularly in winter, tend to be negatively skewed, so

a normal distribution would be inappropriate as this would give a bias to the cold day count. Instead, it is necessary to fit a gamma distribution to the daily anomalies for each day of the year, again using the thirty 1961–90 days for each day. Now it is a simple matter to count the number of days above the 90th/95th (warm/very warm) and below the 10th/5th (cold/very cold) percentiles in a calendar year or in a season.

Figure 6 shows counts of warm/cold days for two of the long European series (Central England and Stockholm). Although there are differences between the stations in the timings of change, the overall picture is of an increase in warm days in the second half of the twentieth century, but the largest trend is a reduction in the number of cold days. Recent increases in warm days at the sites with longer records have only just exceeded similar counts in some decades of the eighteenth century. Cold day counts, in contrast, are clearly lower than at any period in the long records. The analysis method is insensitive to the choice of base period, another choice producing similar trends but centered around a different base.

All the last three sections consider extremes in different ways, but all show similar conclusions. Until the recent 35 years, the warming of the twentieth century is mostly manifest, not by increases in warm extremes, but by reduction in cold extremes. Cold extremes often pass by unnoticed by the majority, except in sectors and seasons where they have important effects.

The Last 150 Years in the Context of the Last 1000 Years

Global average surface temperature has clearly risen over the last 160 years (Figure 1), but what significance does this have when compared to changes over longer periods of time? The last millennium is the period for which most is known about the preinstrumental past, particularly spatially, but it must be remembered that such knowledge is considerably poorer than since 1850. The millennium, particularly the last 500 years, is also the most important when considering attribution of

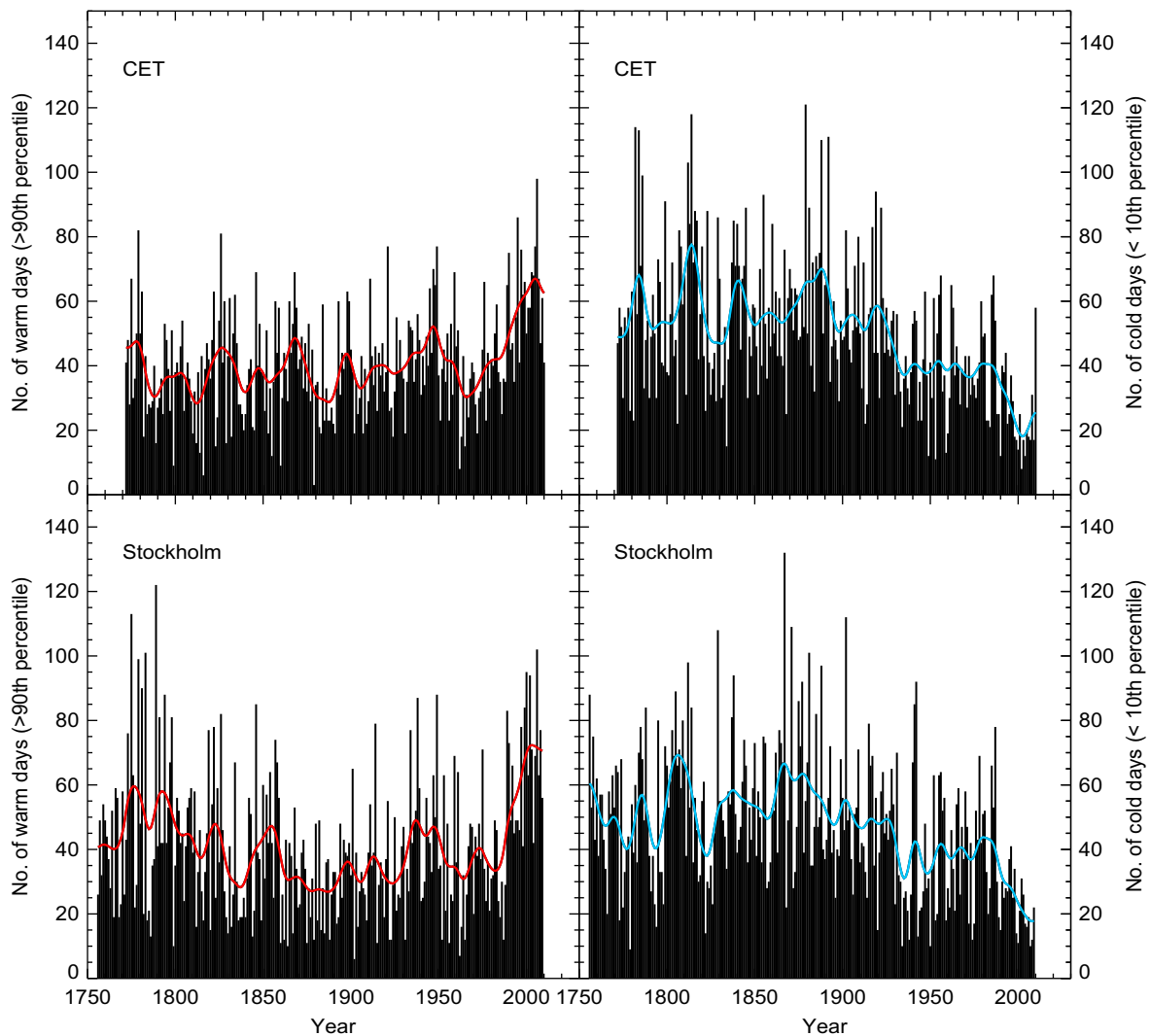


Figure 6 Numbers of cold days (<10th percentile) and warm days (>90th percentile) for four European locations with long daily records (Stockholm and Central England). The smooth curves highlight variations on decadal timescales (red for warm days and blue for cold days).

recent changes to human influences. Earlier millennia are also important, but they are known to have experienced longer timescale changes in solar irradiance caused by orbital changes (the Milankovitch effect), giving, for example, higher irradiance in summer to northern high latitudes around 9000 years ago. Such differences in insolation mean that comparisons to today are not fair. Summers 9000 years ago in the higher latitudes of the Northern Hemisphere (NH) were warmer, but they experienced an 8% greater amount of solar insolation compared to today.

Information about the past millennium comes from a variety of high-frequency and low-frequency proxy sources. High-frequency sources, giving information on the annual timescale include early instrumental records (back to the late seventeenth century in Europe), written historical documents (mainly Europe and the Far East), tree ring densities and widths (mid-to-high latitudes of both hemispheres), ice cores (both polar ice caps and also high-elevation tropical and smaller polar latitude ice caps), corals (tropical), and some highly

resolved lake and marine sediments. Low-frequency (decadal to century timescale change) evidence comes from boreholes, glacial advances/retreats and peat, lake, and marine cores. Uncertainties in all proxy information are considerable, both because evidence is restricted to where these written and natural archives survive, and more importantly, all proxy records are only imperfect records of past temperature changes.

The last two decades have seen a dramatic improvement in both the availability of past evidence and also in information from diverse regions and sources. Figure 7 compares several different reconstructions of NH temperature change for most of the last millennium. The reconstructions are of different seasons, so based on the instrumental record they would be expected to differ somewhat. None of the series are strictly independent of each other, as they contain some common sources, but each has made different assumptions in their averaging. The most striking feature of the multiproxy averages is the warming over the twentieth century, both for its magnitude and duration. Agreement with the instrumental record

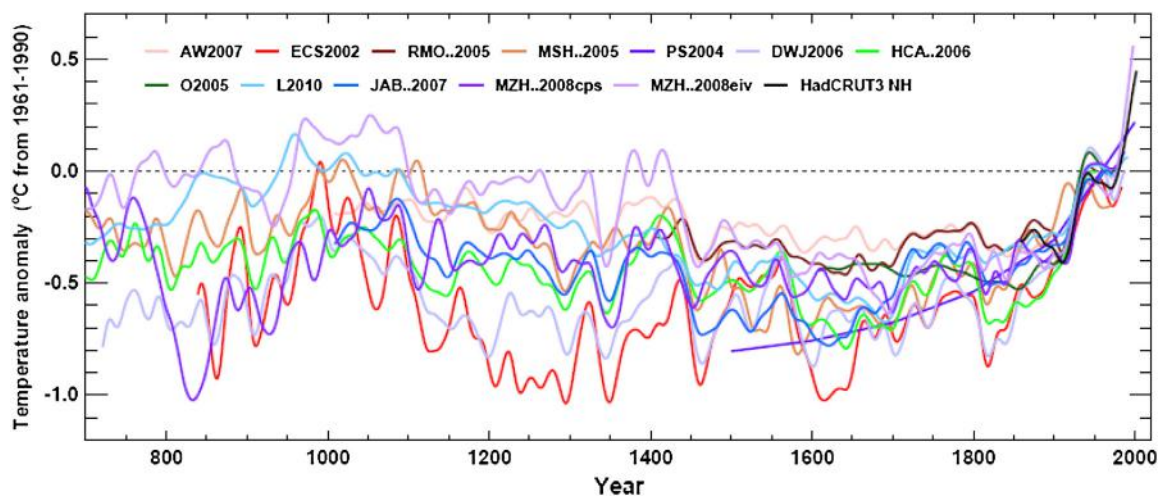


Figure 7 Reconstructions of NH temperatures from several different combinations of proxy data (multiproxy averages). All the series have been smoothed with a 40-year Gaussian filter and all are plotted as departures from the 1961–90 average. The different reconstructions are shown by the colored lines, the black being the instrumental record for April–September for the NH. All the series have been assembled recently and represent cutting edge research in paleoclimatology.

Reproduced from Esper, J., Cook, E.R., Schweingruber, F.H., 2002. Low-frequency signals in long tree-ring chronologies for reconstructing past temperature variability. *Science* 295: 2250–2253.

Reproduced from Rutherford, S., et al., 2005. Proxy-based Northern Hemisphere surface temperature reconstructions: Sensitivity to method, predictor network, target season, and target domain. *Journal of Climate* 18: 2308–2329.

Reproduced from Moberg, A., et al., 2005. Highly variable Northern Hemisphere temperatures reconstructed from low- and high-resolution proxy data. *Nature* 433: 613–617.

Reproduced from Pollack, H.N., Smerdon, J.E., 2004. Borehole climate reconstructions: Spatial structure and hemispheric averages. *Journal of Geophysics Research* 109: D11106, doi: 10.1029/2003JD004163.

Reproduced from D'Arrigo, R., Wilson, R., Jacoby, G., 2006. On the long-term context for late twentieth century warming. *Journal of Geophysics Research* 111: 12, doi: 10.1029/2005JD006352.

Reproduced from Hegerl, G.C., Crowley, T.J., Hyde, W.T., Frame, D.J., 2006. Climate sensitivity constrained by temperature reconstructions over the past seven centuries. *Nature* 440: 1029–1032.

Reproduced from Oerlemans, J., 2005. Extracting a climate signal from 169 glacier records. *Science* 308: 675–677.

Reproduced from Mann, M.E., et al., 2008. Proxy-based reconstructions of hemispheric and global surface temperature variations over the past two millennia. *Proceedings of the National Academy of Sciences of the United States of America* 105, doi: 10.1073/pnas.0805721105.

Reproduced from Ljungqvist, F.C., 2010. A new reconstruction of temperature variability in the extra-tropical Northern Hemisphere during the last two millennia. *Geografiska Annaler: Series A, Physical Geography* 92: 339–351, doi: 10.1111/j.1468-0459.2010.00399.x.

Reproduced from Ammann, C.M., Wahl, E.R., 2007. The importance of the geophysical context in statistical evaluations of climate reconstruction procedures. *Climatic Change* 85: 71–88, doi: 10.1007/s10584-007-9276-x.

Reproduced from Juckes, M.N., et al., 2007. Millennial temperature reconstruction intercomparison and evaluation. *Climate of the Past* 3: 591–609.

should be assumed, as all the components of the series have, to some extent, been calibrated against instrumental data, either locally or as a whole. The twentieth century was the warmest of the millennium and the warming rate during it has been unprecedented.

The number of series averaged together in each reconstruction depends on the particular study. Some use as many as possible, while others are restricted to specific proxy types (e.g., trees) or limited so that all proxies extend back a thousand years. Even the most spatially extensive studies only incorporate in number less than 5% of the number of instrumental sites. At the temporal scale plotted in Figure 7, however, the number of contributing proxy series is adequate to reconstruct large-scale temperatures, much more so for the NH compared to the Southern Hemisphere (SH), so Figure 7 only considers the NH. If standard errors are assigned to these series, as was the case for the instrumental period, errors would be considerably larger even for the 50-year timescale plotted.

Earlier studies, using considerably fewer proxy datasets, have considered the past millennium and two periods, the Little Ice Age (variously defined as AD 1450–1850) and the Medieval Warm Epoch (less well-defined in the literature, but AD 900–1200 encompasses most earlier works) are often discussed. To some extent, these two periods have become accepted wisdom but the various curves in Figure 7 indicate only partial support. Spatial analysis of the proxy data shows that no century-scale periods in the millennium were universally colder or warmer everywhere, with considerable variability being present. The latter is to be expected even by studying the instrumental period since 1850. Just as the early 1940s were warm in many parts of the world Europe was cold, the early seventeenth century was cool in many regions, but was relatively mild in Iceland. In many respects, therefore, paleoclimatology is in the process of reassessing the evidence for these past periods and further changes are in prospect as more evidence becomes available.

The various series in Figure 7 differ in some respects with regard to the coldest and warmest periods of the millennium, but they have all analyzed orders of magnitude more data than available in the early 1970s. The cooler centuries of the millennium were the sixteenth to the nineteenth, the seventeenth being the coldest in Europe, and the nineteenth coldest in North America. These regions are still the best studied and it will be vital in future to extend the knowledge to other areas, particularly in the SH. At present, for every one long SH proxy reconstruction there are at least 10 in the NH. Just as with the instrumental record it is important to gain as much evidence from as many regions as possible, if how global and hemispheric temperatures have varied over this long time are to be fully understood. Contrasts in the timing of changes between regions and particularly between the hemispheres must be recognized if the causes of the changes are to be fully

understood. A more complete understanding of the causes of the changes will allow to determine how much climate can change naturally, enabling to better distinguish the degree of human influence on surface temperature during the twentieth century.

See also: **Climate and Climate Change:** Climate Variability: Decadal to Centennial Variability; Climate Variability: Nonlinear and Random Effects; Climate Variability: North Atlantic and Arctic Oscillation; Climate Variability: Seasonal and Interannual Variability. **Global Change:** Biospheric Impacts and Feedbacks; Sea Level Change; Upper Atmospheric Change. **Ozone Depletion and Related Topics:** Long-Term Ozone Changes. **Paleoclimatology:** Ice Cores; Varves. **Statistical Methods:** Data Analysis: Empirical Orthogonal Functions and Singular Vectors; Data Analysis: Time Series Analysis. **Tropical Meteorology and Climate:** El Niño and the Southern Oscillation: Observation.

Further Reading

- Böhm, R., Jones, P.D., Hiebl, J., Frank, D., Brunetti, M., Maugeri, M., 2010. The early instrumental warm-bias: A solution for long Central European temperature series, 1760–2007. *Climatic Change* 101, 41–67.
- Hansen, J., Ruedy, R., Sato, M., Lo, K., 2010. Global surface temperature change. *Reviews of Geophysics* 48, RG4004. doi:10.1029/2010RG000345.
- Jones, P.D., New, M., Parker, D.E., Martin, S., Rigor, I.G., 1999. Surface air temperature and its changes over the past 150 years. *Reviews of Geophysics* 37, 173–199.
- Jones, P.D., Wigley, T.M.L., 2010. Estimation of global temperature trends: What's important and what isn't. *Climatic Change* 100, 59–69.
- Jones, P.D., Briffa, K.R., Osborn, T.J., et al., 2009. High-resolution paleoclimatology of the last millennium: A review of current status and future prospects. *The Holocene* 19, 3–49.
- Karl, T.R., Hassol, S.J., Miller, C.D., Murray, W.L. (Eds.), 2006. *Temperature Trends in the Lower Atmosphere: Steps for Understanding and Reconciling Differences. A Report by the U.S. Climate Change Science Program and the Subcommittee on Global Change Research*. National Oceanic and Atmospheric Administration, National Climatic Data Center, Asheville, NC, p. 164.
- Liebmann, B., Dole, R.M., Jones, C., Bladé, I., Allured, D., 2010. Influence of choice of time period on global surface temperature trend estimates. *Bulletin of the American Meteorological Society* 91, 1485–1491.
- Mann, M.E., Zhang, Z., Hughes, M.K., et al., 2008. Proxy-based reconstructions of hemispheric and global surface temperature variations over the past two millennia. *Proceedings of the National Academy of Sciences of the United States of America* 105, 13252–13257.
- Parker, D.E., 2010. Urban heat island effects on estimates of observed climate change. *Wiley Interdisciplinary Reviews: Climate Change* 1, 123–133.
- Thompson, D.W.J., Kennedy, J.J., Wallace, J.M., Jones, P.D., 2008. A large discontinuity in the mid-twentieth century in observed global-mean surface temperature. *Nature* 453, 646–649.
- Trewin, B., 2010. Exposure, instrumentation, and observing practice effects on land temperature measurements. *WIREs Climate Change*, 490–506. DOI 10.1002/wcc.46.
- Vose, R.S., Easterling, D.R., Gleason, B., 2005. Maximum and minimum temperature trends for the globe: An update through 2004. *Geophysical Research Letters* 32, L23822. doi:10.1029/2005GL024379.

Sea Level Change

RS Nerem, University of Colorado, Boulder, CO, USA

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

Sea level rise is expected to become a serious problem in the future as the Earth warms. Changes in global average sea level are an excellent indicator of how the Earth is responding to climate change because it responds to the melting of land ice and increases in ocean heat content. Satellite altimeter and gravity measurements have revolutionized how we monitor and understand sea level change. Currently, sea level is rising at a rate of $\sim 3 \text{ mm year}^{-1}$ with roughly equal contributions from thermal expansion, the melting of ice in Greenland and Antarctica, and the melting of mountain glaciers.

Introduction

Sea level rise is likely to be a serious socioeconomic problem in the future as inundation impacts coastal populations and the threat from storms increases. But it is also an excellent indicator of changes in the water reservoirs on the Earth as they respond to climate change. As such, sea level change has emerged as one of a handful of sensitive climate indicators that reveals how the Earth's water reservoirs are changing. Sea level can change due to a variety of factors including the ocean tides, changes in ocean circulation, changes in atmospheric pressure, and a variety of other phenomena, but long-term sea level change is mainly influenced by changes in the heat content of the oceans (thermal expansion) and the exchange of water between the oceans and the continents (Cazenave and Nerem, 2004). Variations in global mean sea level (GMSL) are often used to characterize the climate aspects of sea level change, because much of the natural variability of sea level averages out in the global average. Changes in land water storage, mainly due to changes in precipitation patterns, are the leading cause of interannual GMSL variations, while the melting of land ice (and subsequent run off into the ocean) is the dominant cause of long-term GMSL change, along with thermal expansion. Thus, as the Earth warms, sea level change is a critical climate variable that reflects the amount of heat absorbed by the oceans and land-based ice.

While GMSL may be a critical climate variable, regional sea level change can be quite different from the global mean, and thus is the critical variable for assessing the socioeconomic impacts of the change. In addition, the regional pattern of sea level change may offer clues as to the cause of the changes, which is important for predicting the magnitude and regional variations of future sea level change. For sea level science, accurate global observations of sea level, ocean heat content, and land ice changes are the key for better understanding how the Earth system is responding to climate change and how much sea level might rise in the future. Fortunately, such an observation system has become a reality over the last few decades.

Observations of Sea Level Change

Prior to the satellite era, tide gauge and salt marsh measurements were the primary methods for determining changes in sea level (Woodworth et al., 2011). Tide gauge

measurements have been used to reconstruct GMSL back to 1700 (Figure 1(c)), although only a few gauges extend to the 1700s. Salt marsh measurements have generally corroborated the tide gauge record (Figure 1(b)). While many of our best long-term sea level records come from tide gauge measurements, they are limited by poor spatial distribution and are influenced by vertical land motion at the tide gauge site. Because tide gauges measure sea level change relative to the land, the data are often referred to as 'relative' sea level change measurements. The vertical land motion is important for assessing regional sea level impacts, but it confuses the issue when one is trying to detect true climate signals. Nevertheless, tide gauges provide one of the few direct measurements of sea level change prior to the 1990s, and so attempts are often made to correct the data for the effects of vertical land motion – either through modeling (such as models of glacial isostatic adjustment (GIA)) or through measurements (e.g., using a Global Navigation Satellite System (GNSS) receiver to measure the vertical land motion directly). There have been many different analyses of the tide gauge sea level record (Figure 1(b)). Averaged over the last century, tide gauges generally show that sea level has been rising at an average rate of $\sim 1.8 \text{ mm year}^{-1}$. However, the tide gauge record has provided little insight into the spatial variability of sea level change because of its limited geographic distribution.

The precision satellite altimeter era began in the early 1990s. The launches of TOPEX/Poseidon (1992), Jason-1 (2001), and Jason-2 (2008) into identical orbits with special attention paid to calibration and validation of the measurements have led to a 20-year climate data record of sea level change with a 10-day temporal resolution. The record of GMSL from these missions (Figures 1(d), and 2) has provided a fundamental indicator for how climate change is affecting the Earth system. The average rate of sea level rise over 1993–2013 is $3.2 \pm 0.4 \text{ mm year}^{-1}$. The error is mostly driven by errors in the calibration of the instruments, which is accomplished by locally comparing tide gauge sea level measurements to the altimeter measurements (Mitchum, 2000; Nerem et al., 2010). GMSL also shows significant interannual variability, much of which has been shown to be related to El Niño – Southern Oscillation (ENSO) processes (Boening et al., 2012; Fasullo et al., 2013; Nerem et al., 2010, 1999) through changes in land/ocean precipitation and other factors.

The satellite measurements also allow one to determine the geographic variation of the 20-year sea level rates (Figure 3). This

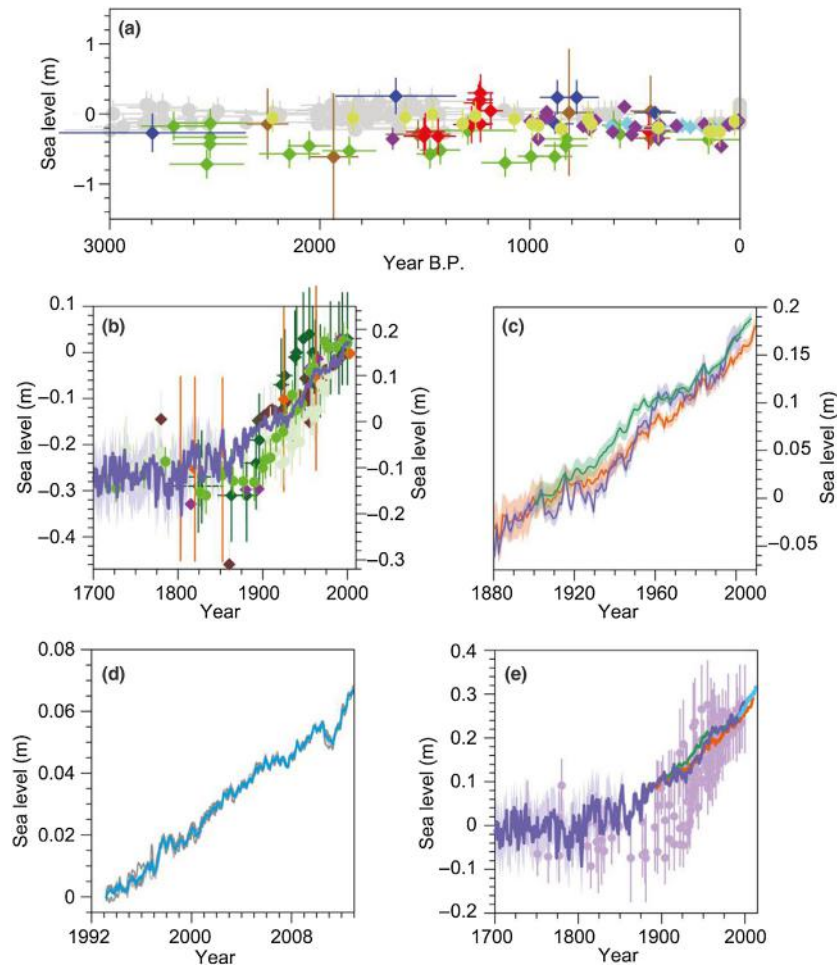


Figure 1 (a) Paleo sea level data from salt marshes since 1700 from Northern and Southern Hemisphere sites compared to sea level reconstruction from tide gauges (blue time series with uncertainty). The effects of GIA have been removed from these records. Green and light green = North Carolina, orange = Iceland, purple = New Zealand, dark green = Tasmania, brown = Nova Scotia. (b) Yearly average GMSL reconstructed from tide gauges (1900–2010) by three different approaches (orange, blue, and green). (c) Altimetry data sets from five groups with mean of the five shown as bright blue line. (d) Comparison of the paleo data from salt marshes (purple symbols, from panel a), with tide gauge and altimetry data sets (same line colors as in panels b and c). Figure 13.3 from Intergovernmental Panel on Climate Change (IPCC) AR5 WG1.

shows considerable geographic variation ($\pm 10 \text{ mm year}^{-1}$) relative to the global mean rate of 3 mm year^{-1} . Much of this is related to the Pacific decadal oscillation (PDO), and will presumably average out as longer time series become available (Hamlington et al., 2013). The current regional rates of sea level change are dominated by steric sea level changes (Woodworth et al., 2011); however, as the loss of land ice accelerates, it is eventually expected to dominate the patterns of regional sea level change. These patterns will be driven by the magnitude and location of the ice mass loss, creating ‘ice sheet fingerprints’ (Figure 4) in the geographic patterns of regional sea level change (Tamisiea and Mitrovica, 2011). Because loss/gain of land ice can change the Earth’s gravity field and the loading on the solid Earth, regional sea level is changed due to changes in the geoid and vertical land motion.

Satellite altimeter measurements have been invaluable for measuring changes in total sea level, but they offer only limited insight into the causes of that change. Specifically, a better understanding is desired of how much is due to

thermal expansion versus the melting of land ice. The launch of the Gravity Recovery and Climate Experiment (GRACE) in 2002 has revolutionized our understanding of the contribution of land ice and water storage variations to sea level change. GRACE provides monthly maps of the Earth’s time-varying gravity field with a spatial resolution of $\sim 300 \text{ km}$ and an accuracy of 1 cm water equivalent. GRACE has allowed the measurement of ice mass loss since 2002 over Greenland (0.6 mm year^{-1}), Antarctica (0.3 mm year^{-1}) (Velicogna and Wahr, 2013), and from mountain glaciers (0.4 mm year^{-1}) (Jacob et al., 2012). In addition, GRACE has allowed the determination of changes in global mean ocean mass (1.8 mm year^{-1}) (Chambers et al., 2004; Chambers et al., 2010) (Figure 2).

While satellite altimeter and gravity measurements have revolutionized the field of sea level science, improved *in situ* measurements of ocean heat content have also taken a leap forward. Our knowledge of changes in ocean heat content has improved significantly in the last decade with the deployment of the Argo network of profiling floats (Leuliette and Willis,

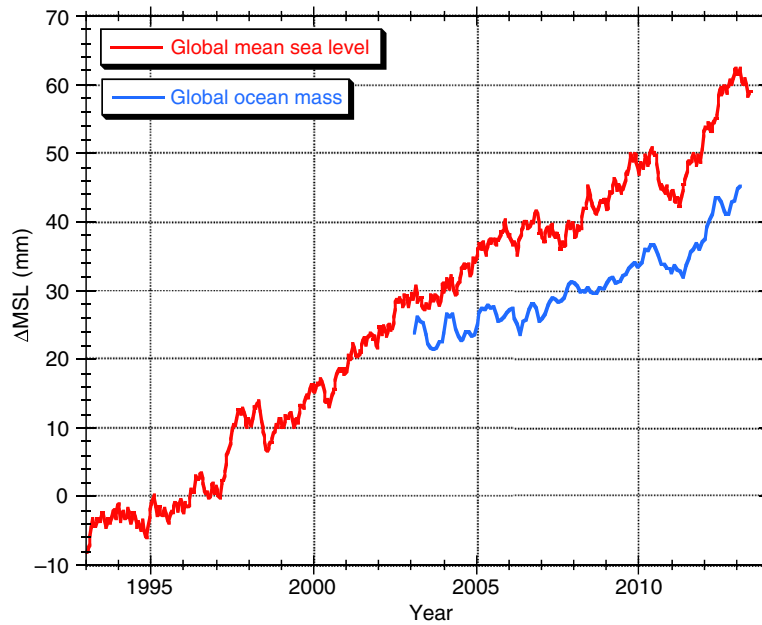


Figure 2 Global mean sea level variations from TOPEX/Poseidon, Jason-1, and Jason-2 altimeter measurements. Seasonal variations have been removed, an inverted barometer correction has been applied, glacial isostatic adjustment is removed, and 2-month smoothing is applied. Reproduced with permission from <http://sealevel.colorado.edu>. Global mean ocean mass variations computed from the GRACE satellite gravity mission (<http://xena.marine.usf.edu/~chambers/SatLab/Home.html>). Seasonal variations and glacial isostatic adjustment have been removed, and 2-month smoothing has been applied.

2011). Prior to Argo, shipboard hydrographic measurements (XBTs, etc.) are used. These measurements suggest that the contribution of thermal expansion in the upper ocean to GMSL over the last two decades is $\sim 1 \text{ mm year}^{-1}$. Much less is known about the heat absorbed by the deep ocean, and a major goal of future research is to better determine that contribution to sea level change.

Causes of the Observed Sea Level Change

Changes in GMSL are primarily driven by changes in ocean heat content (thermal expansion), land water storage, and the

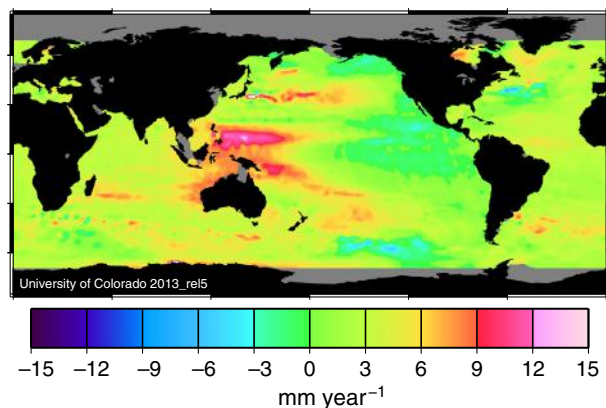


Figure 3 Regional variations in the rate of sea level change over 1993–2013. Reproduced with permission from <http://sealevel.colorado.edu>.

melting of land-based ice. Interannual changes in GMSL have been shown to be primarily driven by changes in land water storage, which are in turn driven by changes in land/ocean precipitation (Boening et al., 2012; Fasullo et al., 2013; Nerem et al., 2010, 1999). However, there is little evidence for long-term GMSL trends related to land water storage, except for those due to the building of artificial water storage reservoirs on land, which reduces sea level rise (e.g., Chao et al., 2008).

Together, satellite altimetry, GRACE data, and Argo data can be used to ‘close’ the sea level budget over the time period when all three observation systems were active. Roughly, ocean mass (GRACE) plus steric changes (Argo) should equal total sea level change (altimetry). This has in fact been demonstrated within the errors of these systems (Leuliette and Willis, 2011). Over the GRACE time period, Greenland and Antarctica are estimated to have contributed 0.6 and 0.3 mm year^{-1} , respectively, to GMSL change (Shepherd et al., 2012; Velicogna and Wahr, 2013), with the remaining mountain glaciers contributing 0.4 mm year^{-1} (Jacob et al., 2012). Decadal variability related to the PDO could have contributed an additional 0.5 mm year^{-1} over this time period (Hamlington et al., 2013). Global ocean mass (which reflects the total land/ocean mass exchange) is increasing at a rate of 1.8 mm year^{-1} since 2002 (Chambers et al., 2004; Chambers et al., 2010). While the advent of GRACE and Argo measurement systems has been a huge advance for sea level science, their use for improving our understanding of sea level change is limited until longer time series are available from these systems.

Although the contributions vary somewhat depending on the time period considered, during the altimeter era (1992 to

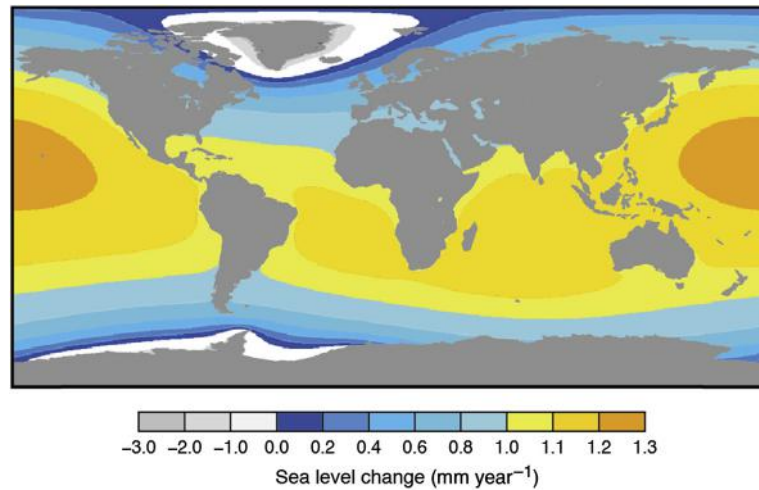


Figure 4 Model output showing relative sea level change due to uniform melting of the Greenland Ice Sheet and the West Antarctic Ice Sheet at rates of 0.5 mm year^{-1} each (giving a global mean value for sea level rise of 1 mm year^{-1}). The modeled sea level changes are less than the global mean value in areas near the melting ice but enhanced further afield. IPCC AR5 FAQ13.1, Figure 2, Adapted from Milne, G.A., Gehrels, W.R., Hughes, C.W., Tamisiea, M.E., 2009. Identifying the causes of sea-level change. *Nature Geoscience* 2, 471–478.

present) the rate of global sea level rise (3.2 mm year^{-1}) can roughly be ascribed to (1) thermal expansion ($\sim 1 \text{ mm year}^{-1}$), (2) Greenland and Antarctica ($\sim 0.9 \text{ mm year}^{-1}$), (3) mountain glaciers (0.8 mm year^{-1}), and (4) decadal variations that appear as trends over the 20-year record ($\sim 0.5 \text{ mm year}^{-1}$) with much smaller contributions from halosteric changes, hydrologic variations in land water storage, and the construction of artificial reservoirs.

Projections of Future Sea Level Change

Assessing the potential socioeconomic impacts of sea level change depends on being able to accurately project future sea level change – both the global average and its regional variation – in addition to many other factors (Tebaldi et al., 2012). This is generally accomplished with either ‘process-based models,’ which use theoretical concepts and computational methods to simulate the behavior of the Earth systems affecting sea level or ‘semiempirical models,’ which use historical observations of sea level and temperature (or radiative forcing) to project sea level change based on projections of future changes in temperature (or radiative forcing) from global climate models. These models have mainly been used to project changes in GMSL, although projections of regional sea level change can also be done.

The IPCC Fifth Assessment Report (AR5) thoroughly evaluated these two techniques for projecting future sea level (Church et al., 2013) based on different scenarios for changing greenhouse gases and radiative forcing called representative concentration pathways (RCPs). For the projected sea level rise from 2000 to 2100, the process-based models project a rise in GMSL of 53–98 cm for RCP8.5 with a rate of change of $8\text{--}16 \text{ mm year}^{-1}$ during 2081–2100 (Figure 5). RCP2.6 ranges from 28 to 61 cm by 2100 and the other RCP scenarios fall in between these to end members. The two projection techniques tend to give somewhat different results, with semiempirical

models generally projecting higher sea levels than process-based models (Figure 6). The IPCC AR5 assessed both techniques, and evaluated the process-based models as ‘medium confidence’ and the semiempirical models as ‘low confidence.’

These techniques for projecting GMSL change can be extended to project regional sea level change as well. Figure 7 shows projections of regional sea level change for different RCP scenarios from the IPCC AR5 (Church et al., 2013). These were computed by combining projections of thermal expansion from climate models and process-based models of ice mass loss, including the effects of ice sheet fingerprints shown in Figure 4 (Church et al., 2011). The projected regional variations from the global mean (Figure 5) can be significant, especially near the ice sheets, where there is very little sea level rise due to the effects of the fingerprints.

While the year 2100 is often a horizon that is used to evaluate future sea level rise, sea level rise will continue well past 2100 even after greenhouse gas emissions have ceased, because of inertia in the global climate system. Levermann et al. (2013) evaluated the potential for sea level rise over the next several millennia and found that thermal expansion and the melting of Antarctica respond roughly linearly with GMSL contributions of 0.4 and $1.2 \text{ m } ^\circ\text{C}^{-1}$ of warming, respectively. Mountain glaciers and Greenland were found to behave non-linearly, but in a compensating way, with Greenland increasing its contribution after the mountain glacier contribution has decreased (because they have effectively melted out). The total sea level commitment over the next 2000 years was found to be $2.3 \text{ m } ^\circ\text{C}^{-1}$ of warming.

Discussion

Global sea level change is a sensitive indicator of changes in the climate system, because on long timescales it responds mainly to heat being absorbed by the oceans and the ice. Observations of sea level change have shown that sea level is rising in response to

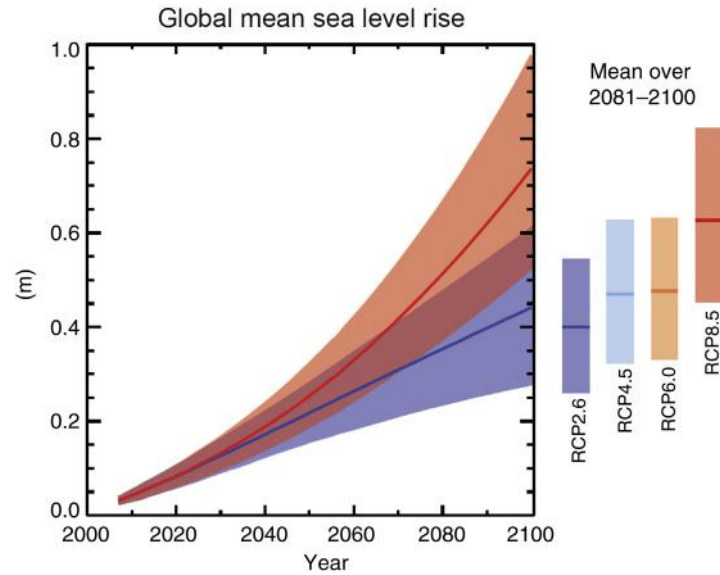


Figure 5 Global sea level rise projections from the IPCC AR5 for different RCP scenarios. RCP8.5 ranges from 0.53 to 0.98 m by 2100 with a rate of change of 8–16 mm year⁻¹ during 2081–2100. RCP2.6 ranges from 0.28 to 0.61 m by 2100. Figure 8 of the SPM for IPCC WG1 AR5.

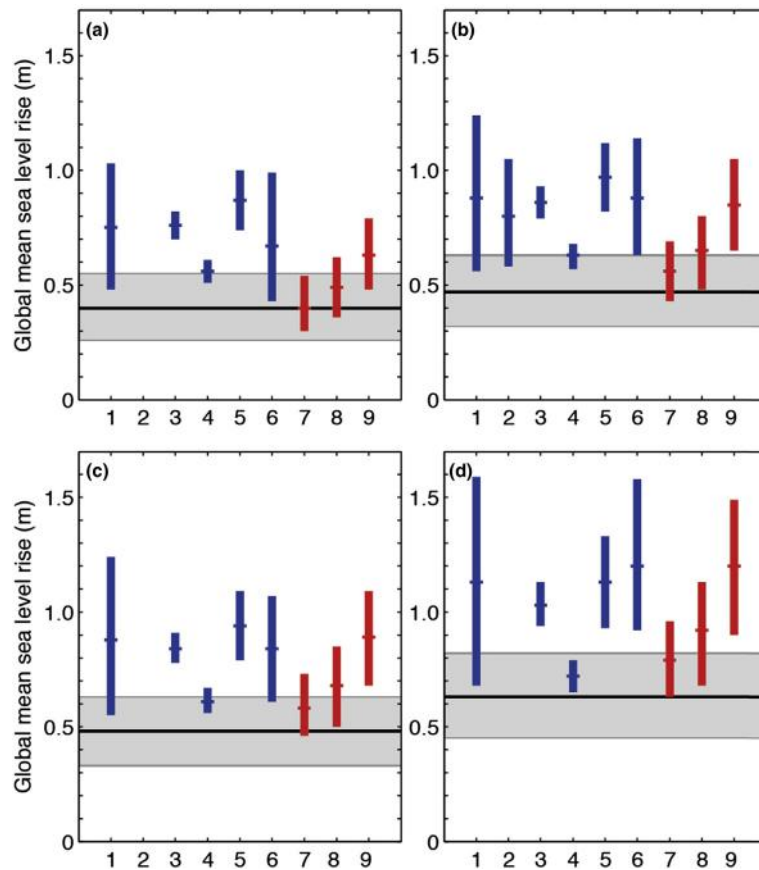


Figure 6 Projections (5–95%) of global mean sea level rise (m) from semiempirical models for 2081–2100 relative to 1986–2005 for (a) RCP2.6, (b) RCP4.5, (c) RCP6.0, and (d) RCP8.5. Blue bars are results from the models using RCP temperature projections; red bars are using RCP radiative forcing. The numbers on the horizontal axis refer to different literature sources. Also shown for comparison is the median (thick black line) and likely range (horizontal gray bar) from the process-based projections. Figure 13.12 from IPCC AR5 WG1.

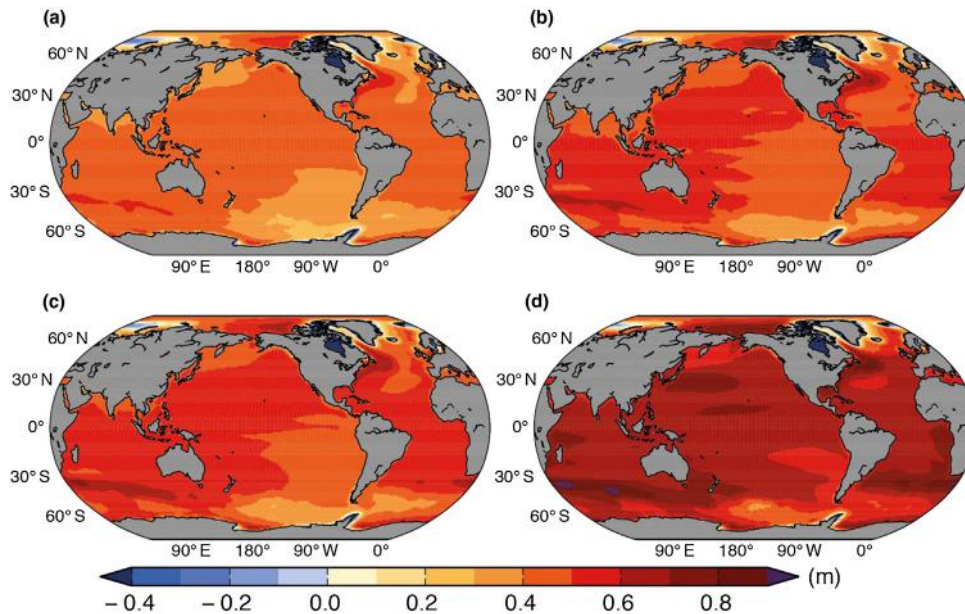


Figure 7 Ensemble mean net regional sea level change (m) evaluated from 21 CMIP5 models for the RCP scenarios (a) 2.6, (b) 4.5, (c) 6.0, and (d) 8.5 between 1986–2005 and 2081–2100. Each map includes effects of atmospheric loading, plus land ice, GIA, and terrestrial water sources. Figure 13.20 from IPCC WG1 AR5.

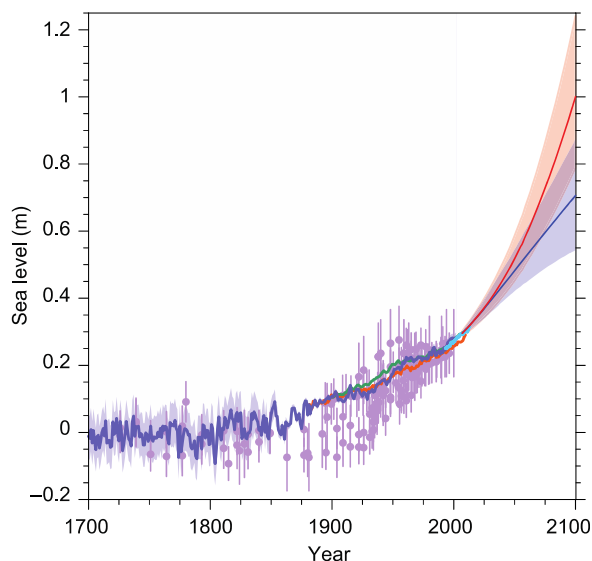


Figure 8 Compilation of paleo sea level data, tide gauge data, altimeter data, and central estimates and likely ranges for projections of global mean sea level rise for RCP2.6 (blue) and RCP8.5 (red) scenarios, all relative to preindustrial values. Figures 13.27 and 13.3 from IPCC AR5 WG1.

the warming of the planet (Figure 8) and that rate is increasing in time. Currently, the rate of GMSL rise is 3.2 mm year^{-1} with the main contributions coming from thermal expansion, the melting of Greenland and Antarctica, and mountain glaciers. However, this rate is expected to accelerate in the coming decades. The IPCC estimates GMSL could rise between 28 and 98 cm this century (Figure 8) depending on how greenhouse gases change in the future. The uncertainties in these projections

are more about how long it will take sea level to rise – the magnitude of the rise is much better understood.

References

- Boening, C., Landerer, F.W., Nerem, R.S., Willis, J.K., Fasullo, J., 2012. The 2012 La Nina: so strong, the oceans fell. *Geophysical Research Letters* 39, L19602. <http://dx.doi.org/10.1029/2012GL053055>.
- Cazenave, A., Nerem, R.S., 2004. Present-day sea level change: observations and causes. *Reviews in Geophysics* 42, RG3001. <http://dx.doi.org/10.1029/2003RG000139>.
- Chambers, D.P., Wahr, J., Nerem, R.S., 2004. Preliminary observations of global ocean mass variations with GRACE. *Geophysical Research Letters* 31, L13310. <http://dx.doi.org/10.1029/2004GL020461>.
- Chambers, D.P., Wahr, J., Tamisiea, M.E., Nerem, R.S., 2010. Ocean mass from GRACE and glacial isostatic adjustment. *Journal of Geophysical Research-Solid Earth* 115. <http://dx.doi.org/10.1029/2010JB007530>.
- Chao, B.F., Wu, Y.H., Li, Y.S., 2008. Impact of artificial reservoir water impoundment on global sea level. *Science* 320 (5873), 212–214. <http://dx.doi.org/10.1126/Science.1154580>.
- Church, J.A., et al., 2013. Sea level change. In: *Climate Change 2013: The Physical Science Basis. Working Group I Contribution to the IPCC Fifth Assessment Report (AR5)*, edited.
- Church, J.A., Gregory, J.M., White, N.J., Platten, S.M., Mitrovica, J.X., 2011. Understanding and projecting sea level change. *Oceanography* 24 (2), 130–143.
- Fasullo, J.T., Boening, C., Landerer, F.W., Nerem, R.S., 2013. Australia's unique influence on global sea level in 2010–2011. *Geophysical Research Letters* 40 (16), 4368–4373. <http://dx.doi.org/10.1002/grl.50834>.
- Hamlington, B.D., Leben, R.R., Strassburg, M.W., Nerem, R.S., Kim, K.Y., 2013. Contribution of the Pacific decadal oscillation to global mean sea level. *Geophysical Research Letters* 40. <http://dx.doi.org/10.1002/grl.50950>.
- Jacob, T., Wahr, J., Pfeffer, W.T., Swenson, S., 2012. Recent contributions of glaciers and ice caps to sea level rise. *Nature* 482 (7386), 514–518. <http://dx.doi.org/10.1038/Nature10847>.
- IPCC, 2013. Summary for policymakers. In: *Stocker, T.F., Qin, D., Plattner, G.-K., Tignor, M., Allen, S.K., Boschung, J., Nauels, A., Xia, Y., Bex, V., Midgley, P.M. (Eds.), Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, Cambridge, UK and New York.

- Leuliette, E.W., Willis, J.K., 2011. Balancing the sea level budget. *Oceanography* 24 (2), 122–129.
- Levermann, A., Clark, P.U., Marzeion, B., Milne, G.A., Pollard, D., Radic, V., Robinson, A., 2013. The multimillennial sea-level commitment of global warming. *Proceedings of the National Academy of Sciences U.S.A.* 110 (34), 13745–13750. <http://dx.doi.org/10.1073/Pnas.1219414110>.
- Milne, G.A., Gehrels, W.R., Hughes, C.W., Tamisiea, M.E., 2009. Identifying the causes of sea-level change. *Nature Geoscience* 2, 471–478.
- Mitchum, G.T., 2000. An improved calibration of satellite altimetric heights using tide gauge sea levels with adjustment for land motion. *Marine Geodesy* 23, 145–166.
- Nerem, R.S., Chambers, D.P., Choe, C., Mitchum, G.T., 2010. Estimating mean sea level change from the TOPEX and Jason altimeter missions. *Marine Geodesy* 33, 435–446. <http://dx.doi.org/10.1080/01490419.2010.491031>.
- Nerem, R.S., Chambers, D.P., Leuliette, E.W., Mitchum, G.T., Giese, B.S., 1999. Variations in global mean sea level associated with the 1997–1998 ENSO event: implications for measuring long term sea level change. *Geophysical Research Letters* 26 (19), 3005–3008.
- Shepherd, A., et al., 2012. A reconciled estimate of ice-sheet mass balance. *Science* 338 (6111), 1183–1189. <http://dx.doi.org/10.1126/Science.1228102>.
- Tamisiea, M.E., Mitrovica, J.X., 2011. The moving boundaries of sea level change: understanding the origins of geographic variability. *Oceanography* 24 (2), 24–39. <http://dx.doi.org/10.5670/oceanog.2011.25>.
- Tebaldi, C., Strauss, B.H., Zervas, C.E., 2012. Modelling sea level rise impacts on storm surges along US coasts. *Environmental Research Letters* 7 (1), 014032. <http://dx.doi.org/10.1088/1748-9326/7/1/014032>.
- Velicogna, I., Wahr, J., 2013. Time-variable gravity observations of ice sheet mass balance: precision and limitations of the GRACE satellite data. *Geophysical Research Letters* 40 (12), 3055–3063. <http://dx.doi.org/10.1002/Grl.50527>.
- Woodworth, P.L., Gehrels, W.R., Nerem, R.S., 2011. Nineteenth and twentieth century changes in sea level. *Oceanography* 24 (2), 80–93.

Upper Atmospheric Change

RG Roble, National Center for Atmospheric Research, Boulder, CO, USA

© 2003 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 2, pp 910–915, © 2003 Elsevier Ltd.

Introduction

The composition of the Earth's atmosphere has changed considerably over geological time but very little is known about how the upper atmosphere and solar–terrestrial interactions responded to these changes. Indeed, it is only since the dawn of the space age that information on the structure of the upper atmosphere and its response to solar–terrestrial influences have been measured. Since the space age began in the 1950s, this does not leave a long record for determining the sensitivity of the upper atmosphere to changes driven by recently observed variations in composition. However, certain measurements and model calculations have made it increasingly clear that releases of trace gases from human activity have a potential for causing a change in the present-day climate of the Earth. Most discussions of the projected change during the past decade have dealt with changes in the Earth's troposphere and stratosphere. There have been many papers written on the subject and several assessment panels have issued reports which indicate that the troposphere will warm and the stratosphere will cool in response to a doubling of the present-day composition of several trace gases, such as carbon dioxide and methane, toward the end of the twenty-first century. Yet not much is known about how the upper atmosphere will respond to these influences. Therefore, the following is based primarily on theoretical considerations and model simulations that await verification by adequate trend measurements of the upper atmosphere properties. The few available observations, however, have suggested that the mesosphere and thermosphere/ionosphere are experiencing a cooling trend over the past few decades.

Trace Gas Influences on the Troposphere and Stratosphere

The amounts of certain trace gases in the atmosphere, such as carbon dioxide (CO_2), methane (CH_4), nitrous oxide (N_2O), chlorofluorocarbons (CFCs), and tropospheric ozone (O_3), have been increasing during the past few decades. These gases are mainly transparent to incoming short-wave solar radiation but they absorb and emit long-wave IR radiation and thus are able to influence the Earth's climate. These gases are generally referred to as 'greenhouse gases.' Various projections, based on past and current trends of the rate of emission of these gases, suggest that the present-day concentration of several of these species will double by the end of the twenty-first century.

Most studies dealing with the consequences of the increasing greenhouse gases have been concerned with the troposphere and the stratosphere. These studies suggest that the global mean temperature of the troposphere will increase by 1–5 K and the stratosphere will cool by about 10–20 K in response to a doubling of present day CO_2 concentrations and various increases in other greenhouse gases. The predictions are

based primarily on radiative considerations and there is still considerable uncertainty as to the magnitude of the global response because of uncertainties in cloud feedback, aerosols, and dynamics.

Numerical model studies of the stratospheric response have shown that the combined effects of projected trace gas increases to the end of the twenty-first century will result in major changes to both the ozone and temperature distributions. **Figure 1** shows the results of one such model calculation. For a doubling of the CO_2 concentration alone, the calculated stratopause temperature decreases by 10–15 K. But when chemical feedback is included, the concentration of ozone will increase because certain temperature-dependent chemical rate coefficients are slower at lower temperature. This results in less ozone destruction. Higher ozone concentrations will absorb more solar radiation and thus partially offset the cold temperatures calculated by CO_2 cooling alone thus producing a temperature healing. When the chlorofluorocarbon concentration increases, say from 2.0 to 6.6 ppbv, the ozone in the upper stratosphere decreases about 30–50% and the temperature decreases by 10–15 K. The combined effect of both the CO_2 doubling and the CFC increases is shown in **Figure 1** where the stratopause temperature is seen to decrease by 15–25 K and ozone in the lower mesosphere to increase by 10% due to the healing effect but to decrease by 10–30% in the upper stratosphere because of

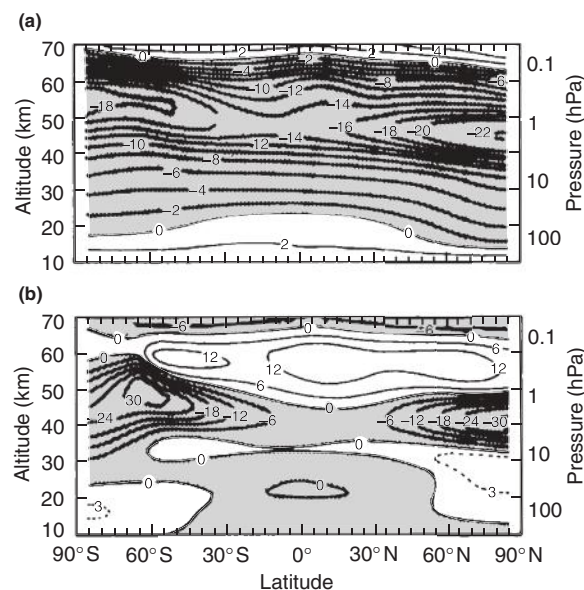


Figure 1 Predicted response of the middle atmosphere in June to a doubling of CO_2 and an increase of chlorofluorocarbons from 2.0 to 6.6 ppbv. The contours are temperature difference in degrees K in (a) and percent change in ozone in (b), both from present-day conditions. Reproduced from Brasseur and Hitchman (1998). Stratospheric response to trace gas perturbations: changes in ozone and temperature distribution. *Science* 240: 634–637.

catalytic destruction of ozone caused by CFCs. These results suggest that significant stratospheric change should occur as a result of increased trace gases from present-day levels. CO₂ increases lead to lower temperatures in the middle and upper stratosphere because of increased emission in the infrared radiation to space. In addition, ozone depletions in the middle and upper stratosphere above about 30 km lead to less absorption of ultraviolet radiation, and thereby to less solar heating and enhanced cooling. As a result of the combined effect, the stratosphere is projected to cool significantly.

Trace Gas Influences on the Mesosphere and Thermosphere

There have been a few studies to investigate how the atmosphere above the stratopause might respond to the global change expected to occur in the troposphere and stratosphere. These studies suggest that global change effects are not confined entirely to the lower atmosphere but that significant changes can extend further into the upper atmosphere, affecting the mesosphere, thermosphere, and ionosphere. For a doubling of present-day concentrations of CO₂ and CH₄ mixing ratios, which is expected to occur by the end of the twenty-first century, the global mean temperature of the mesosphere is projected to cool by 10 K and that of the thermosphere to cool by 50–100 K from present-day conditions. The predicted sensitivity of the mesosphere and thermosphere to variations of CO₂ and CH₄ is shown in Figure 2, where temperature departures from the present-day globally averaged temperature structure are shown for a doubling of present-day CO₂ concentrations, which should occur by the end of the twenty-first century, and the halving of present-day CO₂ concentrations that occurred at the end of the last ice age about 18 000 years ago. These predictions indicate a slow response to trace gas increases, yet the few available observation during the past decade indicate that the mesosphere and thermosphere are cooling much more rapidly than suggested by the trace gas predictions, indicating that other factors are involved in the upper atmosphere that are not accounted for in present-day model simulations.

The atmosphere expands as it is heated and contracts as it cools. Therefore, the troposphere should expand slightly as it is warmed by a few degrees, but the upper atmosphere should contract much more because of the greater cooling. As a result, the density at a given height in the upper atmosphere should decrease. The sensitivity of the upper atmosphere density to a halving and doubling of the CO₂ concentrations from present-day concentrations is also shown in Figure 2. The calculations suggest that at a given height in the upper thermosphere, densities were 40–50% larger at the end of the last ice age and should be 40–50% smaller by the end of the twenty-first century for similar solar and auroral forcings. The decrease shown could occur at the end of the twenty-first century and so the density variations could be occurring slowly at the present time. However, recent observations of the decay in the orbits of satellites indicate that the decadal trends in atmospheric drag are occurring faster than model predictions that include only CO₂ doublings. There appears to be missing physics or chemistry in current models if the

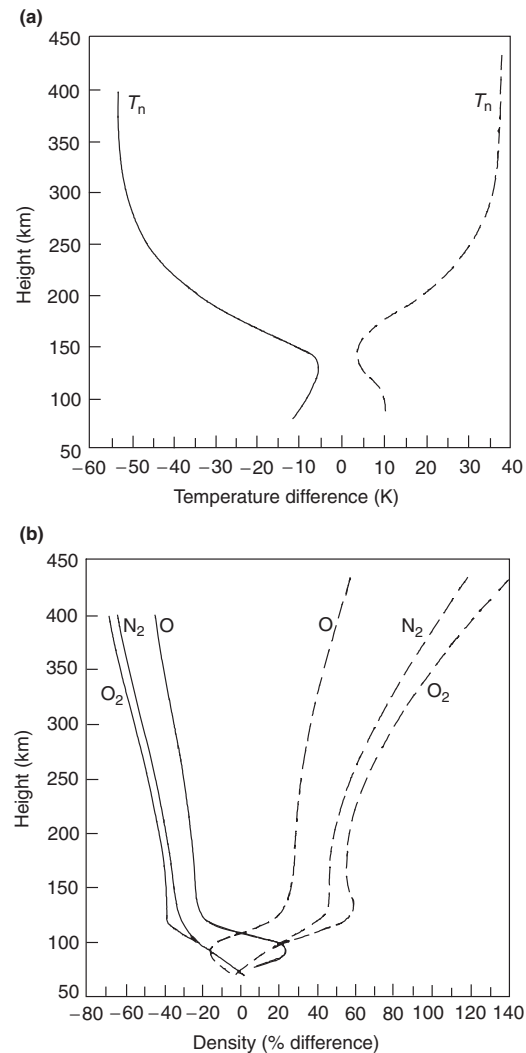


Figure 2 (a) Calculated neutral gas temperature (T_n) difference profiles for doubling of CO₂ and CH₄ concentrations (solid line) and a halving of the CO₂ and CH₄ concentrations (dashed line) from present-day concentrations. (b) The compositional response of the major neutral gas constituents of the upper atmosphere. Solid lines are from the doubling of the trace gas concentrations and dashed lines are for halving the concentrations from present-day values. O₂, N₂, and O are molecular oxygen, molecular nitrogen, and atomic oxygen respectively. Reproduced from Roble and Dickinson (1989). How will changes in carbon dioxide and methane modify the mean structure of the mesosphere and thermosphere. *Geophysical Research Letters* 16: 1441–1444.

observations that suggest the current observed rate of atmospheric cooling are correct.

In addition to the changes in the major constituents of the mesosphere and thermosphere, there are significant variations that occur in the concentrations of minor neutral gas constituents such as O₃, H₂O, OH, HO₂, CO, NO, NO₂, N(⁴S), and others. The minor species change because many of the chemical rate constants for these species are highly temperature-sensitive and turbulent diffusion and advective transport, which determine their vertical distributions, depend on the vertical temperature structure. The results show that for a doubling of

CO₂ and CH₄, NO, and O₃ could decrease in the thermosphere and H₂O, N(⁴S), CO, and H could increase. A doubling of the present-day methane concentration could also increase the H₂O and with colder mesopause temperatures could result in an increase in noctilucent clouds. An increase in tropospheric methane of about 1 ppmv could cause an increase in exospheric hydrogen by about 30% with a corresponding increase in the total density in the hydrogen-dominated part of the exosphere.

A change in these species and temperature structure could also have an effect on the ion composition of the D-, E-, and F-ionospheric regions. Even though the atmospheric models that predict these global mean changes are quite sophisticated, there are still many uncertainties concerning the chemistry, reaction rates, and various atmospheric processes, such as turbulence and atmospheric transport by gravity and planetary waves, that suggest caution in evaluating the magnitude of the predicted changes in chemical structure that respond to atmospheric cooling.

In addition to the changes in thermal and compositional structure on a global mean basis, there are also variations in the latitudinal distribution of heat and momentum sources that alter the mean meridional and global circulation patterns from present-day conditions.

Predicted Changes in Ionospheric Structure

The predictions of change in the neutral upper atmosphere could also cause changes in the ionosphere. The thermospheric cooling and associated composition changes could lower the E-layer and F2-layer peaks in the ionosphere by about 2 km and 20 km, respectively. However, the predicted changes in the E-layer and F2-layer electron densities should be small. Changes in the 'maximum usable frequency' for radio propagation should also be small for the trace gas doubling scenario. These changes are consistent with a lowering of the ionosphere as the neutral atmosphere contracts from 'greenhouse cooling.'

There is an important need to continue model development using coupled models of the Earth's atmosphere, ionosphere, and magnetosphere to evaluate the entire atmospheric readjustment to global change induced by human activity.

Do Observations Support the Predictions of Global Change?

In addition to the modeling studies mentioned above, there have been various investigations that suggest that a change in the composition and temperature structure of the upper atmosphere has been taking place in the past few decades. The overall trend in upper atmosphere response is not clear at the present time but a brief summary of some of the observational studies is given below:

- Radiosonde and satellite observations indicate a decadal cooling trend of the global, annual-mean lower stratosphere since about 1980. Over the period 1979 to 1994, its amplitude is approximately 0.6 K/decade in the 15–20 km altitude range.
- The vertical profile of the annual-mean stratospheric temperature change observed in the Northern Hemisphere for the 1979–94 period consists of an approximately 0.75 K/decade cooling of the 15–35 km region, a slight reduction in the cooling at about 35 km, and increased cooling with height above 35 km (approximately 2 K/decade at 50 km).
- In the upper stratosphere and mesosphere, rocket observations made at various Russian and other stations indicate a few degrees of cooling in the 30–40 km region, increasing to 20 K cooling in the mesosphere near 60–70 km. These changes are not only occurring at high latitudes but penetrate to low and equatorial latitudes.
- In the middle and upper atmosphere, both the greenhouse gas increases and observed ozone changes contribute to the cooling. Atmospheric models are currently being used to investigate the observed decadal trends.
- Lidar observations of the height of the atmospheric sodium layer near 90 km suggest that the height of the sodium layer has decreased during the past decade, indicating that the mesosphere temperature has cooled over the same period.
- Observations have suggested that the trend in increasing atmospheric methane has the potential for increasing the water vapor concentration in the middle atmosphere and contributes to an increase in noctilucent cloud formation in the high-latitude summer mesopause region near 82 km. It is well known that water vapor is an important product of methane oxidation in the stratosphere and thus the increasing methane concentration may be responsible for an increase in the number of sightings of noctilucent clouds in recent years. An observed secular increase over the last few decades may, in part, be caused by increasing concentrations of water vapor and colder temperatures in the mesopause region.
- Lidar measurements over the south of France have indicated a secular change in temperature over the 33–87 km altitude range. A –4 K/decade temperature trend in the 60–70 km altitude region between the years 1978 and 1989 has been observed. Ground-based lidar data and satellite temperature data have found a change in mesospheric temperatures with a trend of –0.10 to –0.14 K per year during the 1980–90 period. These measurements also indicated a solar cycle variability that must be considered when evaluating secular trends.
- Observations have shown that the solar cycle variation of stratospheric ozone over the 1979–99 period is about 5% near 40 km altitude. The observed trend in ozone over this period is monotonically decreasing with a 6–8% per decade change.
- Predicted changes in the atmospheric vertical structure that should occur as a result of a doubling of trace gas concentrations could have a significant effect on the gravitational and thermal tides that are excited in the atmosphere. In particular, the effects in the middle atmosphere may be significant.
- An increase in tropospheric methane of 1 ppmv could cause an increase in exospheric hydrogen number density of 30%. Since atomic hydrogen is the dominant species in the exosphere, this suggests that the total exospheric density is sensitive to global change. The increase in exospheric

hydrogen could have important consequences on the plasma density of the magnetosphere, plasma exchange between the magnetosphere and ionosphere, and perhaps solar–terrestrial plasma coupling processes.

- Perhaps the largest uncertainty in the energetics of the upper mesosphere and lower thermosphere is the magnitude of the rate coefficient for deactivation of the bending mode of carbon dioxide by atomic oxygen. There has been considerable uncertainty in the rate of this coefficient, ranging from 2×10^{-13} in early investigations to a value of $6 \times 10^{-12} \text{ cm}^3 \text{ s}^{-1}$ in more recent analyses of satellite data. This coefficient influences the predictions of upper atmospheric response to global change.
- Recent laboratory and satellite data have indicated that certain crucial rate coefficients important for the energy balance of the upper atmosphere have changed considerably, suggesting that the overall aeronomy of the upper atmosphere is still not sufficiently known. The photodissociation coefficient for nitric oxide and the O–NO vibration deactivation rate coefficient have changed recently, which, along with the uncertainty of the CO_2 –O rate coefficient, makes accurate predictions of upper atmosphere global change difficult at least until these aeronautical issues are resolved.

- Will present-day solar–terrestrial couplings be altered by changed upper atmosphere and ionosphere structures?
- Will new upper atmosphere chemical and physical processes become important as a result of an altered atmospheric structure (e.g., heterogeneous chemistry was found to be important for understanding changes in the Antarctic ozone hole)?
- Will an altered upper atmosphere structure have important feedback on the troposphere and weather systems?
- Will present-day technological systems have to be redesigned to adapt to a changed upper atmosphere environment?

These are only a few of the many questions that emerge from the realization that the present-day upper atmospheric structure may be changing.

See also: Chemistry of the Atmosphere: Principles of Chemical Change. **Climate and Climate Change:** Climate Variability: Decadal to Centennial Variability; Overview. **Clouds and Fog:** Noctilucent Clouds. **Mesosphere:** Ionosphere; Metal Layers; Polar Summer Mesopause. **Ozone Depletion and Related Topics:** Ozone Depletion Potentials. **Solar System/Sun, Atmospheres, Evolution of Atmospheres:** Solar Terrestrial Interactions: Climate Impact.

Future Studies

The studies described have been motivated by the realization that global change is not confined only to the lower atmosphere but that the consequences of man's activities may be more global and extend into the upper atmosphere and even affect plasma processes that couple the solar wind into the Earth's upper atmosphere. Indeed, the application of current, well-tested models of the upper atmosphere suggests that significant changes to the present-day structure of the atmosphere above about 50 km will occur by the middle of the twenty-first century. The studies conducted thus far are only suggestive of possible changes and there is an important need for further investigations and long-term measurements to detect natural and anthropogenic trends into the twenty-first century.

Studies of the possible effect of global change on the upper atmosphere are in their infancy and a number of important questions emerge:

- Will gravity wave, tidal, and planetary wave propagation characteristics change in response to an altered atmospheric structure?
- How will energy, mass and momentum exchange rates between atmospheric regions change in response to increased CO_2 cooling and a contracted upper atmosphere?

Further Reading

- Akmaev, R.A., Fomichev, V.I., 1998. Cooling of the mesosphere and lower thermosphere due to doubling of CO_2 . *Annales Geophysicae* 16, 1501–1512.
- Brasseur, G., Hitchman, M.H., 1988. Stratospheric response to trace gas perturbations: changes in ozone and temperature distribution. *Science* 240, 634–637.
- Brasseur, G.P., Orlando, J.J., Tyndall, G.S., 1999. *Atmospheric Chemistry and Global Change*. Oxford University Press, New York.
- Bruhl, C., Crutzen, P.J., 1988. Scenarios of possible changes in atmospheric temperatures and ozone concentrations due to man's activities, estimated with a one-dimensional coupled photochemical climate model. *Climate Dynamics* 2, 173–203.
- Ramanathan, V., 1988. The greenhouse theory of climate change: a test by an inadvertent global experiment. *Science* 240, 293–299.
- Rishbeth, H., Roble, R.G., 1992. Cooling of the upper atmosphere by enhanced greenhouse gases: modeling the thermospheric and ionospheric effects. *Planetary and Space Science* 40, 1011–1026.
- Roble, R.G., 1993. 'Greenhouse cooling' of the upper atmosphere. *EOS* 74, 92–93.
- Roble, R.G., Dickinson, R.E., 1989. How will changes in carbon dioxide and methane modify the mean structure of the mesosphere and thermosphere. *Geophysical Research Letters* 16, 1441–1444.
- Thomas, G.E., 1996. Global change in the mesosphere–lower thermosphere region: has it already arrived? *Journal of Atmospheric and Terrestrial Physics* 58 (14), 1629–1656.
- World Meteorological Organization, 1998. *Scientific Assessment of Ozone Depletion: 1998, Global Ozone Research and Monitoring Project, Report No. 44*.
- Yung, Y.L., DeMore, W.B., 1999. *Photochemistry of Planetary Atmospheres*. Oxford University Press, Oxford.

Biospheric Impacts and Feedbacks

BA Hungate and GW Koch, Northern Arizona University, Flagstaff, AZ, USA

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

The terrestrial biosphere interacts with the atmosphere through exchanges of trace gases and of energy. These interactions are bidirectional, with atmospheric conditions modifying ecological controls of trace gas and energy exchange, and the very exchanges feeding back to affect atmospheric chemistry and climate. This article addresses the role of biosphere–atmosphere material and energy exchange in atmospheric processes, the biological processes involved, and how biotic responses to ongoing and future global changes could influence the future chemical and radiative balance of the atmosphere.

Introduction

The terrestrial biosphere interacts with the atmosphere through exchanges of trace gases and of energy. These interactions are

bidirectional, with atmospheric conditions modifying ecological controls of trace gas and energy exchange, and the very exchanges feeding back to affect atmospheric chemistry and climate (Figure 1). Here, we review the role of

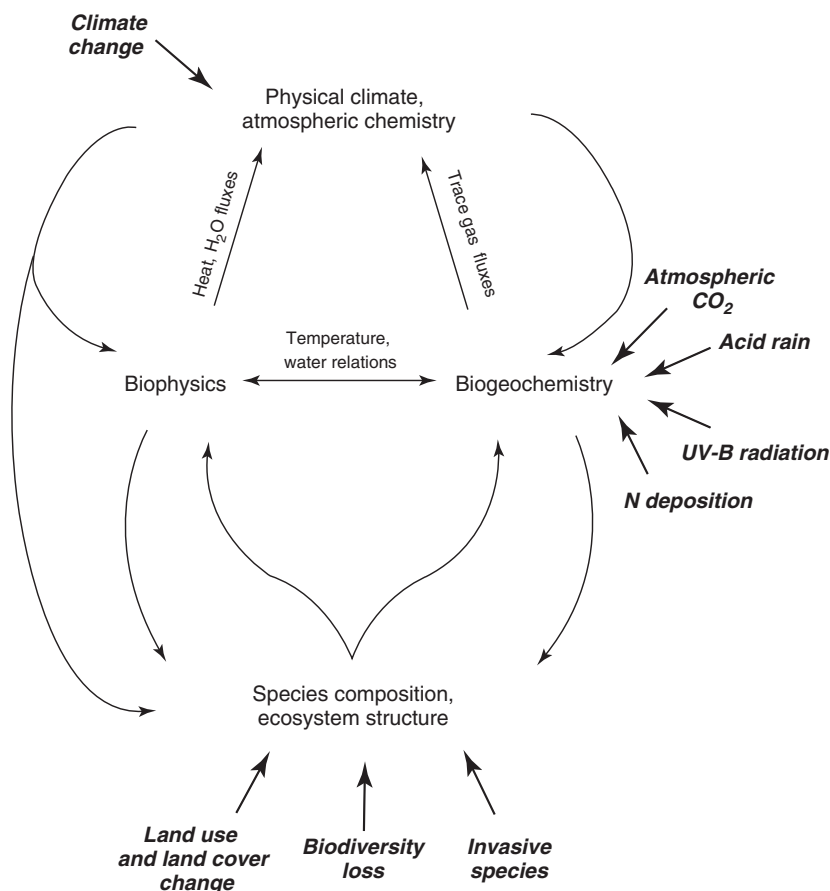


Figure 1 Interactions between the terrestrial biosphere and the atmospheric composition and climate (plain type, thin arrows), and impacts of global environmental change on the coupled biosphere–atmosphere system (bold italics, thick arrows). Climate influences biophysical and biogeochemical processes, creating feedback effects to both atmospheric composition and climate through fluxes of heat, water, and trace gases. Global changes such as N deposition, acid rain, enhanced UV-B radiation, and rising CO₂ directly affect ecosystem biogeochemistry, feeding back to the atmosphere through altered trace gas fluxes, altering ecosystem structure by favoring some species over others, and affecting biophysically mediated exchanges of energy by altering physiological processes (e.g., stomatal conductance). Other global changes affect ecosystem structure directly, with, for example, invading species introducing new biological entities with unique contributions to biogeochemical and biophysical properties. Modified from Sellers, P.J., et al., 1991. Charting the boreal forest's role in global change. EOS Transactions American Geophysical Union 72 (4), 33–40.

biosphere–atmosphere material and energy exchange in atmospheric processes, the biological processes involved, and how biotic responses to ongoing and future global changes could influence the future chemical and radiative balance of the atmosphere. Global environmental change is already altering biological production and consumption of trace gases and energy partitioning between the biosphere and the atmosphere. However, the magnitude and direction of these effects are not well constrained, and this uncertainty contributes to the difficulty in predicting future global climate. Global changes likely to substantially alter biosphere–atmosphere trace gas and energy exchange include climate change, rising atmospheric CO₂, acid rain and associated nitrogen deposition, species invasions, losses of biodiversity, and land use and land cover change (Figure 1).

Trace gases in the atmosphere are arbitrarily defined as gas species with concentrations equal to or less than that of carbon dioxide (388.5 ppmV, average concentration for 2010). Through biogeochemical transformations in terrestrial ecosystems, plants and soil microorganisms produce and consume many trace gases, often at rates high enough to affect atmospheric concentrations. Climate is a key determinant of biogeochemical transformations, and thereby directly affects rates of trace gas exchange. Because species differ in their physiological capacities to produce and consume trace gases, the suite of organisms comprising a biological community and the number of organisms present in an ecosystem both strongly influence biogeochemical cycling and associated trace gas fluxes. In turn, biogeochemical transformations include processes that regulate nutrient availability to plants, thereby influencing species composition and ecosystem structure and productivity. Energy exchanges between ecosystems and the atmosphere include fluxes of sensible heat and latent heat (i.e., evaporation of water), exchanges that are directly mediated by the biophysical properties of ecosystems. Like biogeochemical transformations, biophysically mediated energy exchange is sensitive to climate and to the physiological properties of organisms, mostly plants. Thus, changes in climate and in the structure of terrestrial ecosystems alter patterns of biosphere–atmosphere energy exchange.

Trace Gases

Carbon Dioxide

The concentration of carbon dioxide (CO₂) in the atmosphere is increasing by 0.5% per year (1980–2008), a rate unprecedented for the last 400 000 years. A heat-trapping or greenhouse gas, CO₂, is estimated to be responsible for 60% of current global warming and that which is predicted to occur over the next several hundred years. While the current increase in CO₂ concentrations is clearly driven by human activities (fossil fuel burning, deforestation, and cement manufacture), two biological processes strongly influence the concentration of CO₂ in the atmosphere. Photosynthesis by plants (including algae) converts atmospheric CO₂ to organic carbon, and this process is the primary entry point of carbon into the biosphere. Respiration by plants, animals, and microorganisms oxidizes organic carbon, returning it to the atmosphere as CO₂. Photosynthetic and respiratory rates

of CO₂ exchange between the terrestrial biosphere and the atmosphere amount to roughly 120 Gt C year⁻¹, compared to about 9.5 Gt C year⁻¹ added through the combined anthropogenic effects of fossil fuel burning, land use change, and cement manufacture. Thus, predicting the future trajectory of atmospheric CO₂ concentrations requires understanding how global changes might alter the balance between CO₂ release to the atmosphere versus CO₂ uptake and storage in ecosystems.

Spatial and seasonal changes in atmospheric CO₂ concentration reflect the balance of photosynthesis and respiration in the terrestrial biosphere. For example, photosynthesis is a larger flux than respiration during the growing season, and this imbalance is reflected in lower atmospheric concentrations of CO₂ during the spring and summer. This pattern is most pronounced above the temperate zone in the northern hemisphere (Figure 2); less so in the southern hemisphere, where landmasses are smaller; and essentially absent in tropical regions, where the activities of photosynthesis and respiration tend to be synchronous throughout the year. Oceanic exchanges of CO₂, while large, do not show strong seasonality. These seasonal oscillations in atmospheric CO₂ concentration reflect the metabolism of the terrestrial biosphere and underscore the importance of biological control over atmospheric CO₂ concentrations.

Global change can alter these biological controls, affecting fluxes of photosynthesis and respiration and thus the net exchange of carbon between the biosphere and atmosphere. Several lines of evidence suggest that such changes are already underway. The amplitude of the seasonal oscillations in atmospheric CO₂ has increased over the past 40 years, particularly above the northern hemisphere, indicating greater interannual variation in activity of the terrestrial biosphere in this region. Inverse modeling of atmospheric CO₂ concentration, stable carbon isotope composition (¹³C/¹²C), and a number of land-based measurements of biosphere–atmosphere CO₂ exchange indicate a net sink for atmospheric CO₂ in the northern hemisphere. While land use change in the tropics currently amounts to a substantial source of CO₂ to the atmosphere, on the order of 1.1 Gt C year⁻¹, this is partly offset by afforestation of abandoned agricultural lands in the northern hemisphere and net uptake of CO₂ by forests. Experimental additions of either CO₂ or N often stimulate ecosystem carbon uptake, at least in the short term, suggesting that globally pervasive N deposition and increasing atmospheric CO₂ may already be contributing to the terrestrial CO₂ sink. For 2009, this terrestrial carbon sink is estimated to have been 2.4 Gt C year⁻¹. However, it is unclear whether this terrestrial sink will persist. For example, because photosynthesis is less sensitive to temperature than are respiration and microbial decomposition of plant residues, future stores of organic C in soils could decline by up to 11 Gt C globally for every 1 °C of warming. Particularly at high latitudes, the combination of warming and drying could stimulate CO₂ release from tundra and peat bogs, ecosystems that contain most of the world's soil carbon. Furthermore, while experimental CO₂ and N fertilization increases ecosystem CO₂ uptake in the short term, photosynthesis shows a saturating response to rising CO₂, and excessive N deposition often decreases soil fertility and plant production. Because of the

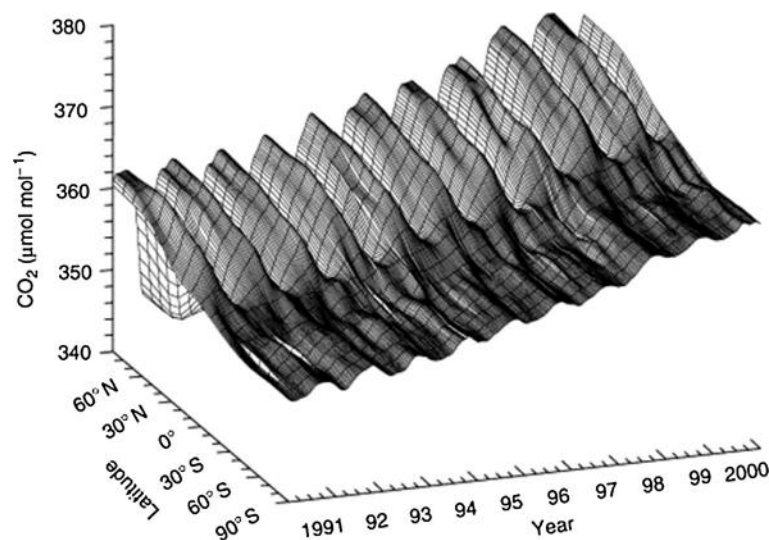


Figure 2 Three-dimensional representation of atmospheric carbon dioxide in the marine boundary layer. The surface represents data that are smoothed in time and latitude. Reprinted with permission of the principal investigators, Pieter Tans and Thomas Conway, of the National Oceanic and Atmospheric Administration (NOAA) Climate Monitoring and Diagnostics Laboratory (CMDL) Carbon Cycle Greenhouse Gases, Boulder, CO.

slow turnover time of soil organic carbon, increased carbon uptake observed in experiments tends to overestimate the long-term potential for ecosystem carbon storage.

Methane

Methane (CH_4) is also a heat-trapping gas, contributing roughly 15% to the radiative forcing associated with global warming. As with CO_2 , two biological processes are important in the production and consumption of atmospheric CH_4 : methanogenesis and methanotrophy. Methanogenesis is the largest source of CH_4 to the atmosphere and is conducted by strictly anaerobic microorganisms of the domain Archaea. Methanogens thrive in many habitats where oxygen concentrations are low, including wetlands, sediments in aquatic ecosystems, and the guts of termites and ruminant mammals. Methane produced by soil microorganisms can be released by plant leaves, and a small amount of methane is produced by UV degradation of plant pectins. With intensification and extensification of livestock farming and rice cultivation, humans have increased methanogenic activity and associated CH_4 emissions to the atmosphere. Fossil fuel mining and biomass burning are additional important anthropogenic sources of CH_4 to the atmosphere. While natural sources add around $160 \text{ Tg CH}_4 \text{ year}^{-1}$ to the atmosphere, anthropogenic sources are larger, totaling nearly $375 \text{ Tg CH}_4 \text{ year}^{-1}$. Most CH_4 is destroyed in the atmosphere through interactions with OH radicals or is lost to outer space. However, methanotrophs – literally ‘methane eaters’ – are an additional important global CH_4 sink, consuming as much CH_4 as accumulates in the atmosphere each year (40 Tg) and constituting approximately 10% of the global CH_4 sink. Methanotrophs are bacteria that oxidize CH_4 to generate energy and fix CO_2 into organic compounds. Methanotrophs are widespread in nature, and are quantitatively important in the global CH_4 cycle in well-aerated surface soils of terrestrial ecosystems. Many temperate forests,

for example, are net CH_4 sinks, consuming CH_4 from the atmosphere.

Global change has the potential to increase or reduce CH_4 production in soils. For example, global changes that increase plant production, such as elevated CO_2 , often stimulate CH_4 efflux from wetlands, because increased plant production in wetlands enhances substrate availability for anaerobic decomposition, stimulating CH_4 production. In experiments, elevated CO_2 often reduces transpiration (water loss) from wetland plants, raising the water table in wetlands and enhancing anaerobic conditions, again potentially stimulating CH_4 efflux. In the arctic, thawing of permafrost may already be stimulating CH_4 emissions by increasing methanogenic activity in previously frozen soils. By contrast, if warming causes widespread drying of wetlands and tundra, anaerobic activity could decline and rates of methane oxidation increase.

Globally, methane oxidation in soils is a smaller flux than methane production, and changes in methane oxidation are likely to have less influence on atmospheric concentrations than are changes in methane production. Some evidence points to a diminished capacity of terrestrial ecosystems to consume atmospheric CH_4 with increasing atmospheric CO_2 concentrations, but the magnitude of this effect appears to be negligible at the global scale. By increasing soil water content (through reduced transpiration) and thereby reducing CH_4 diffusion into soil, elevated CO_2 has been shown to reduce rates of CH_4 oxidation by forest soil. In another case, CH_4 consumption rates declined in CO_2 -treated forest plots even without a change in soil water content or any other obvious mechanism, raising the possibility of reduced efficiency in CH_4 oxidation due to changes in the methanotrophic bacteria community. Reduced CH_4 oxidation has also been observed in grasslands exposed to elevated CO_2 . Thus, reduced CH_4 uptake may be a general consequence of rising atmospheric CO_2 .

Land use change is likely to reduce the terrestrial CH_4 sink: converting forests, woodlands, and savannahs to cultivated or

grazed lands reduces CH_4 uptake, sometimes even causing a shift from net methanotrophy to net methanogenesis. Continued expansion of agriculture and livestock husbandry is likely to further reduce the soil CH_4 sink. Acid deposition also decreases CH_4 oxidation, due to physiological responses of methanotrophs to low soil pH, and also, in the case of nitrogen deposition, to inhibitory effects caused by increased NH_4^+ availability in soil. Ammonium-oxidizing bacteria (nitrifiers; see the “Oxides of Nitrogen” section) have the capacity to oxidize methane, because of structural similarities in the ammonia- and methane-binding enzymes. For this reason, it is not always apparent which groups of organisms are responsible for methane oxidation in soils. Compared to methanotrophs, however, nitrifiers have a lower affinity for CH_4 and therefore oxidize CH_4 more slowly, so displacement of methanotrophs by nitrifiers in response to N deposition may exacerbate inhibition of CH_4 oxidation by ammonium. The inhibitory effects of NH_4^+ on methane oxidation may also explain the reduced capacity for CH_4 oxidation following land use change, as NH_4^+ availability often increases after land clearing or because of direct application of nitrogenous fertilizers.

Oxides of Nitrogen

Nitrous oxide (N_2O) is a potent greenhouse gas, with 300 times the warming potential of CO_2 on a molar basis. N_2O concentration in the atmosphere is increasing at a rate of 0.3% per year, and it is responsible for around 5% of the radiative forcing associated with global warming. N_2O also plays a critical role in stratospheric ozone depletion. Nitric oxide (NO) and nitrogen dioxide (NO_2), together referred to as ‘ NO_x ’ are important reactive gases, influencing tropospheric concentrations of O_3 , OH, HNO_3 , and CH_4 . While their effects on atmospheric processes fundamentally differ, N_2O and NO_x are produced and consumed by the same groups of organisms, and so these gases are treated together here.

Nitrification and denitrification are the two major biological sources of N_2O and NO_x . Nitrifying bacteria use ammonia (NH_3) as an energy source, oxidizing it to nitrate (NO_3^-). Like methanotrophs, which oxidize CH_4 , nitrifiers use their substrate (NH_3) to fix carbon in a manner analogous to a plant’s use of light in photosynthesis. Incomplete oxidation of NH_3 results in gaseous losses of nitrogen dioxide (NO_2), nitric oxide (NO), and N_2O ; emissions of these oxides of nitrogen to the atmosphere can account for as much as 5% of the total nitrification flux. Denitrifying bacteria respire NO_3^- , using it as a terminal electron acceptor just like animals (and other aerobic organisms) use O_2 . In the process, denitrifiers reduce NO_3^- to N_2 gas, with several intermediates, including NO_2 , NO, and N_2O . The final step in the denitrification chain is reduction of N_2O to N_2 , a step that is often incomplete, such that gaseous N_2O escapes from the reaction site. NO_x , too, can be emitted during the earlier steps. Complete denitrification can also consume both N_2O and NO_x and other reports suggest that NO_x consumption by other (nondenitrifying) soil microorganisms may serve a detoxifying function, or may have no clear physiological role (co-metabolism). However, it is not clear that net N_2O or NO_x consumption by soils has any major influence on the global budget of these trace gases. Plants also

play a role in atmospheric NO_x : both uptake and emission of NO_2 by plant leaves are known to occur.

Gaseous losses of NO_x and N_2O during nitrification and denitrification represent small proportions of the total flux of N through these pathways. Nevertheless, the quantities produced are appreciable for the global fluxes of these important trace gases. Globally, soils amount to 65% of total atmospheric sources of N_2O and 40% of NO_x . Tropical (N_2O), wetland (N_2O), arid (NO_x), and cultivated (both) systems are the major soil contributors of these important trace gases. Rates of N_2O production by soils are likely to increase with climate warming, because, like most biological reactions, N_2O production is temperature sensitive and has been found to increase in response to experimental warming in a number of cases. A lowering of the water table in tundra and peatlands associated with warming increases N_2O emissions, likely because the slightly more oxidizing conditions favor incomplete reduction of NO_3^- during denitrification, increasing the amount of N_2O lost in the reaction chain. Conversion of tropical forests to pasture, one of the more prevalent land use changes in the tropics, usually decreases N_2O losses from soils. However, pasture reclamation, which often involves fertilizer application, can cause N_2O fluxes to return to or even exceed levels typical of native forests. Additionally, conversion of native ecosystems to agriculture often dramatically increases N_2O losses from soils. Thus, ongoing global changes are likely to increase rates of N_2O accumulation in the atmosphere.

Emissions of NO_x to the atmosphere are also increasing, particularly from cultivated (and fertilized) soils. As with N_2O , land clearing, cultivation, and fertilization in agriculture increase soil emissions of NO_x , and global increases in these land use changes are increasing the soil source of these reactive gases. Rising atmospheric CO_2 has been shown to reduce NO_x fluxes associated with nitrification by increasing the demand for NH_3 by nonnitrifying microorganisms, reducing substrate availability for nitrifiers. By contrast, warming and N deposition are both likely to increase NO_x fluxes from terrestrial ecosystems.

Other Trace Gases

Carbon monoxide (CO) affects atmospheric chemistry by contributing to tropospheric ozone formation and interfering with methane destruction in the stratosphere. Up to 3% of net primary productivity can be lost as direct CO emission by plants or as losses of volatile organic hydrocarbons (VOCs) subsequently oxidized to CO. Soils are both an important source and sink of CO to the atmosphere. A number of bacteria are capable of oxidizing CO to CO_2 , including some methanotrophs (also important in CH_4 uptake) and nitrifiers (important in N_2O production). Based on selective inhibitor experiments, even eukaryotes may contribute to CO consumption in some forest soils. CO oxidation is sensitive to the water and organic matter content of soils, but is less sensitive to temperature compared to many other ecological processes. While most soils are net CO sinks, some, particularly in arid regions, are net CO producers. The mechanism for CO production is unknown, but it is apparently abiotic, as sterilization often converts a soil from a net CO sink to a source.

Compared to other trace gases, the biology of CO consumption is not very well understood.

Plants emit a number of VOCs to the atmosphere, including isoprenoids (isoprene), terpenes, and oxygenated compounds. These compounds influence carbon monoxide, ozone, and aerosol formation in the troposphere. About half of plant VOC emissions to the atmosphere occur as isoprene. Effects of global change on VOC emissions have not been characterized as well as for other trace gases; nevertheless, changes in VOC emissions are likely to occur. Because plant species naturally differ in rates of VOC production, land use and other global changes that result in shifts in species composition are likely to profoundly alter VOC emissions to the atmosphere. For example, in general, native tropical forest trees produce more isoprene than either crops or pasture grasses, such that land clearing for agriculture or grazing likely reduces VOC emissions. Secondary forests, however, contain many species with high rates of VOC emissions, and net emissions from secondary forests may exceed those from primary forests in some cases. For this reason, land use change could cause VOC emissions to the atmosphere to increase or decline. Isoprene and terpene emissions increase exponentially with temperature, and are thus likely to increase in response to global warming. Global changes that increase plant production (CO_2 and N deposition) are also likely to increase VOC production. In concert with emissions of NO_x from combustion and fertilizer application, increased VOC emissions could increase concentrations of tropospheric ozone and carbon monoxide.

Carbonyl sulfide (COS) is the most abundant sulfur species in the atmosphere. COS is primarily produced abiotically in the oceans; anaerobic production by soil microorganisms occurs, particularly in salt marshes, but is a small source globally. In the stratosphere, COS is oxidized by photolysis to sulfate aerosols, forming an aerosol layer important in cooling the planet, and also in enhancing depletion of stratospheric ozone through chemical interactions with nitrogen and chlorine species. Biological uptake of COS by soils has been

documented, but its significance on a global scale is not well constrained. Lichens are also capable of COS uptake, likely representing a small but nontrivial sink. Uptake by terrestrial vegetation is the major global sink for COS (Table 1). Based on enzyme inhibition studies, COS uptake is likely mediated by carbonic anhydrase, an enzyme that occurs in plants, lichens, and bacteria. COS enters higher plants through stomata, and there is some evidence that reduced stomatal aperture impedes COS uptake by vegetation. Thus, global changes that reduce stomatal conductance (e.g., increasing atmospheric CO_2) may reduce the terrestrial COS sink, possibly constituting a negative feedback to global warming. By contrast, COS uptake in lichens apparently increases with temperature, a possible positive feedback to global warming.

Water and Energy Exchange

Terrestrial ecosystems interact with the atmosphere through exchanges of energy, moisture, and momentum at all temporal and spatial scales. These interactions are functions of key land surface parameters, which derive from the structure and physiology of terrestrial ecosystems. The bidirectional nature of ecosystem–atmosphere interactions is clearly evident. For example, the gas exchange of leaves responds to variation in light, temperature, vapor pressure, and CO_2 partial pressure, and, in turn, affects the water, energy, and trace gas exchange with the atmosphere. Over long periods and broad spatial scales, ecosystem–atmosphere interactions are yet more evident. Water availability and temperature range are the major determinants of the structure and function of terrestrial ecosystems over decades to centuries, and the paleoecological and paleoclimatological records indicate that glacial–interglacial cycles involve coupled changes in the distribution of terrestrial ecosystems, surface albedo, biogeochemistry, and climate. It is understood that the Earth’s climate and ecological systems have interacted over geologic time. Current changes in

Table 1 Trace gases influencing atmospheric chemistry and climate that are produced and consumed by the terrestrial biosphere

<i>Gas</i>	<i>Producing organisms</i>	<i>Terrestrial production</i>	<i>Consuming organisms</i>	<i>Terrestrial consumption</i>	<i>Atmospheric consequences</i>
CO_2	All organisms	40%	Plants and autotrophic bacteria	40%	Radiative forcing
COS	Sulfur-reducing microorganisms in salt marshes and soils	21%	Plants and soil microorganisms	72%	Aerosol
Nonmethane VOCs	Plants	80%	Soil microorganisms	Negligible	Reacts with oxidized pollutants
CH_4	Methanogens	70%	Methanotrophs (nitrifiers)	10%	Radiative forcing
N_2O	Denitrifiers and nitrifiers	65%	Denitrifiers	Unknown	Radiative forcing and O_3 removal (stratosphere)
NO and NO_2	Denitrifiers, nitrifiers, and plants	40%	Denitrifiers and plants	Unknown	Photochemical smog
CO	Plants	5%	Soil microorganisms and plants	15%	O_3 formation (troposphere) and CH_4 destruction (stratosphere)

Shown are the gas species of interest, the organisms responsible for producing and consuming them, a rough estimate of the percentage of total annual production or consumption mediated by the terrestrial biosphere (e.g., of all global processes that consume COS each year, plants are estimated to consume 70%), and a brief description of the role of the gas species in atmospheric chemistry and climate.

climate, atmospheric composition, and land cover arising from human activities will likely continue to affect and be affected by ecosystem–atmosphere interactions.

In this section, we first explain the biophysical and physiological basis for water and energy interactions between terrestrial ecosystems and the atmosphere. We then examine some of the major ways in which global change affects ecosystem–atmosphere interactions. Much of our current understanding derives from atmospheric general circulation models (AGCMs), which have developed rapidly in recent years to include realistic parameterizations of the land surface, including the physiological responses of vegetation to multiple environmental factors. Although much uncertainty exists in the magnitude, and in some cases direction, of ecosystem–atmosphere interactions, modeling and large-scale experimental studies clearly indicate that the potential feedbacks are large.

The Biophysics and Physiology of Ecosystem–Atmosphere Interactions

The presence of vegetation affects a number of key land surface characteristics that are important determinants of the surface energy balance and the water and energy exchange with the atmosphere.

Albedo

Vegetation affects the reflectivity, or albedo, of the Earth's surface, the land surface parameter having the largest influence on the surface radiation budget. Albedo is the integrated reflectance over the solar spectrum (0.0–4.0 μm) and is lower for vegetation than bare soil (Table 2) because leaves absorb strongly in the visible wavelengths (0.4–0.7 μm) useful in photosynthesis and moderately from 0.7 to 4.0 μm . In contrast, soils have lower average absorption across the solar spectrum. Although leaves of different plant species vary somewhat in reflectivity, albedo is primarily sensitive to the leaf area index (LAI), the average amount of leaf area per ground area ($\text{m}^2 \text{m}^{-2}$). Values of LAI range from 0 for extreme deserts to one in arid regions and up to 5–7 or more for forests. The vegetation influence on albedo also varies temporally. For example, forest albedo generally increases in winter when deciduous trees are leafless or evergreens are snow covered, cropland and grassland albedo varies seasonally as the plant canopy develops and then senesces, and extreme events including droughts, pest outbreaks, and severe storms can reduce vegetation cover and increase albedo. Over longer time scales, the changing distribution of land surface albedo is closely tied to movement of vegetation zones in response to climate variation.

Table 2 Reflectivity (albedo) of various surfaces

Surface	Albedo
Forests	0.05–0.18
Grassland	0.22–0.28
Crops	0.15–0.26
Snow	0.75–0.95
Wet soil	0.09 ± 0.04
Dry soil	0.19 ± 0.06
Water	0.05 to >0.20

Sensible heat, latent heat, and evapotranspiration

Ecosystems have a major influence on how the net radiation received at the surface is balanced by losses of sensible and latent heat, which in turn have profound effects on weather and climate. Sensible heat flux from land warms the overlying air and planetary boundary layer. Latent heat is the energy required to evaporate water from soils and plants, and acts to cool the surface. Evaporated water can be transported above the planetary boundary layer, where it may release heat during condensation to form clouds (which affect radiation balance) or precipitation, often at considerable distance from the site of surface evaporation. The presence of vegetation generally increases latent heat flux relative to sensible heat flux. Of the 111 000 km^3 of precipitation on land each year, about 71 000 km^3 returns to the atmosphere by evaporation from soils and transpiration from plants (together considered 'evapotranspiration,' ET), the balance reentering the oceans as river flow. Globally, latent heat releases from the surface about three-quarters, and sensible heat one-quarter, of the annual average net radiation received by the surface.

The ratio of sensible to latent heat flux densities is known as the Bowen ratio (β). Average β varies inversely with water availability on land, ranging from about 0.2 for tropical rain forests to 0.4–0.8 in temperate forests and grasslands, 2–6 for semiarid regions, and 10 for deserts. The Bowen ratio is not static, however; distinct daily, seasonal, and annual trends arise from variation in LAI, soil and plant surface moisture availability, the surface–atmosphere vapor pressure deficit (VPD), and the resistance of the plant canopy to transpiration. Transpiration is the diffusion of water vapor along the concentration gradient from the saturated interior surfaces of leaves to the surrounding air via microscopic pores known as stomata

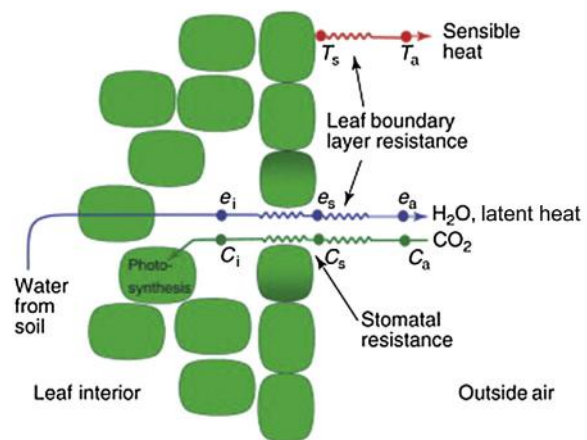


Figure 3 Schematic representation of exchange of water vapor (e), carbon dioxide (C), and heat by leaves. Gas flux rates are determined by the leaf-to-air concentration gradient and the combined stomatal and boundary layer resistance to diffusion. Light-dependent photosynthetic assimilation of CO_2 causes daytime $C_i < C_a$ and a net uptake of CO_2 by the leaf, while in the absence of light, leaf respiration causes $C_i > C_a$ and the net flux is from the leaf. Stomatal resistance is under physiological control and responds to light, temperature, humidity, and CO_2 concentration. Subscripts i, s, and a refer to interior, surface, and ambient, respectively. After Sellers, P.J., Dickinson, R.E., Randall, D.A., et al. 1997. Modeling the exchanges of energy, water, and carbon between continents and the atmosphere. *Science* 275, 502–509.

(Figure 3), and it is the plant process that most directly impacts ecosystem–atmosphere exchange of water and latent heat. Transpiration can enhance total surface evaporation because plants can extract water from below the soil surface, which itself develops high resistance with mild drying, and because plant canopies can present multiple layers ($LAI > 1$) for evaporation. An example of the importance of transpiration to water balance is apparent from figures for the Amazon basin estimating that up to 50% of regional precipitation arises from water transpired from plants elsewhere in the basin.

Stomata act as variable-aperture valves, the resistance (or conductance) of which responds to environmental factors including light, temperature, soil moisture, VPD, and CO_2 concentration. Canopy conductance is the stomatal conductance averaged over the total canopy leaf area. Because photosynthetic uptake of CO_2 by land plants shares the same stomatal diffusion pathway as transpiration, stomatal conductance simultaneously influences water, energy, and carbon fluxes. Theoretical studies suggest that stomatal conductance sometimes varies so as to optimize the efficiency of carbon gain relative to water loss. Conditions tending to enhance water stress to plants, particularly low soil moisture and high VPD, typically lead to partial stomatal closure, which then causes a decrease in transpiration and latent heat flux and an increase in surface temperature and sensible heat flux (i.e., β increases). The potential for ecosystem–atmosphere feedbacks mediated by stomatal physiology can be appreciated from the fact that stomatal conductance, through its control on transpiration and latent heat, can influence the conditions (e.g., VPD) to which it also responds. This is relevant in the context of global change because several global change factors (altered temperature, moisture availability, and CO_2) affect stomatal conductance and, thus, potentially can feedback to either amplify or moderate these changes.

Aerodynamic resistance

A third means by which ecosystems affect surface–atmosphere interactions is via their influence on the aerodynamic resistance of the land surface. Aerodynamic resistance varies inversely with wind speed and the log of the surface roughness length, the latter being about 10% of the vegetation height. Thus, for a given wind speed, taller vegetation such as forests has a lower aerodynamic resistance than shorter vegetation (e.g., grasslands), and promotes greater turbulent transfer of sensible and latent heat away from the surface. The vertical structure of the daytime atmospheric boundary layer is also highly dependent on partitioning of sensible and latent heat; a deeper boundary layer develops when sensible heat flux is large and convection increases.

Terrestrial ecosystems have several additional influences on land surface properties and processes. Vegetation shields soil from radiation inputs, reducing the magnitude of the components of the soil energy budget. Plant canopies also intercept precipitation, decreasing moisture reaching the soil and cooling the canopy surface when intercepted rain reevaporates. Finally, the distribution of roots in the soil profile affects the water available for ET. Studies of variation in the natural abundance of stable isotopic forms of water ($H_2^{18}O$ and $^2H^1HO$) in precipitation, groundwater, soils, plants, and the atmosphere are beginning to reveal the different sources of

water that is evaporated from soils or transpired by plants, and the importance of these atmospheric inputs to local and regional vapor pressure and precipitation.

Terrestrial Ecosystem Effects on Climate

The land surface parameters described in this article underlie influences of ecosystems on climate that are evident from observational and modeling studies. Sparsely vegetated urban areas feel warmer, and are warmer, than nearby forests or grasslands. In the north central United States, the rate of increase of mean surface temperature during spring slows abruptly as deciduous trees leaf out and increased latent heat flux imparts a cooling effect that counteracts the increasing insolation. The aridity of deserts is reinforced by their lack of vegetation, and, conversely, water transpired by tropical rain forests returns to these regions as precipitation. A variety of AGCMs have explored vegetation controls on climate for major regions of the terrestrial biosphere. Forest clearing for agriculture in the eastern and central United States during the 1800s is estimated to have resulted in cooler summers and autumns, consistent with the instrumental record. The cooling was due to changes in vegetation characteristics, in this case primarily reduced leaf area, which increased albedo, and reduced net radiation and sensible heating of the atmosphere. AGCM simulations that replaced Amazon forests with grassland estimated substantial reductions (300 mm year^{-1}) in ET and precipitation and increases in surface temperatures of $3\text{--}5^\circ\text{C}$. Changes in ET and precipitation arose from reduced LAI and decreased roughness length, which reduced turbulent transfer of moisture above the planetary boundary layer. Surface temperatures increased because latent heat flux decreased more (12 W m^{-2}) than did absorption of solar radiation (6 W m^{-2}), the latter due to increased albedo. Reduced ET following rainforest clearing is also predicted to decrease cloud cover, lessening the effect of increased surface albedo by increasing incident solar radiation, but also allowing more outgoing longwave radiation to be lost by the warmed land surface.

An AGCM scenario of replacing all boreal forests with bare ground or tundra (vegetation of short stature, with low leaf area vegetation) showed that boreal forests have a strong warming influence during winter and summer relative to the alternative cover situations. The warming results from the lower albedo of forests than snow, which is masked by the overlying trees. Under the deforestation scenario, land surfaces at high latitudes were up to 12°C colder in April, and remained as much as 5°C colder in July, when the albedo effect was small. These terrestrial changes were linked to sea surface conditions; the colder winter air temperature caused by deforestation reduced sea surface temperature (SST) in Arctic regions, inducing a thermal lag that inhibited warming of land surfaces in the summer. Lower SSTs increased the extent of sea ice, increasing ocean albedo and reinforcing the cooling effect of deforestation. These simulated effects of boreal deforestation were not limited to the boreal region itself; at latitudes as low as 30°N , simulated air temperatures were up to 3°C cooler throughout the year in the deforestation scenario compared to current, control conditions. These examples of ecosystem impacts on climate illustrate the

importance of including land cover change in models of future global climate.

Global Change and Ecosystem–Atmosphere Interactions

Global change, including changes in climate, atmospheric composition, and land use, will alter ecosystem–atmosphere interactions by changing the abundance, distribution, and functioning of different terrestrial ecosystems. For clarity, we discuss ecosystem responses and feedbacks to each of these major global changes separately, but recognize that these interactions are likely to be highly interdependent in the real world.

Climate change

Global climate change models project an increase in globally averaged temperature of 2.4–6.4 °C for the period 2090–2100 compared to 1980–99. The broad range of likely temperature increase is primarily due to uncertainty in projections of future greenhouse gas emissions by human activities. Temperature increases for terrestrial ecosystems should be greater than the global average because the land surface is not as thermally buffered as the oceans. High-latitude regions are very likely to warm considerably more than equatorial and midlatitudes (a phenomenon known as ‘polar amplification’) because warming will reduce the extent of snow and ice cover, decreasing albedo and increasing net radiation. Winter warming may exceed summer warming in high latitudes, while where soil moisture is low, summer warming may exceed winter warming because of low latent heat loss. Warming will drive greater ocean evaporation, and globally averaged precipitation is expected to increase by about 10%. Precipitation is likely to increase in high latitudes in winter as major storm tracks shift northward. At midlatitudes, precipitation is likely to increase in winter, with rain increasing relative to snow, but likely to decrease in summer in continental interiors. Aridity is likely to increase in currently arid regions because although increased surface temperatures will drive more evaporation, precipitation will decrease because the water-holding capacity of air increases nonlinearly with temperature according to the Clausius–Clapeyron relationship, and so air vapor pressure will tend to be farther from saturation. Most AGCM projections also forecast increased frequency of extreme weather events, including severe storms and droughts.

The primary approach to understanding climate-driven changes in vegetation distribution is based on current correlations between natural vegetation and climate, the so-called ‘equilibrium’ approach. The assumption is that, after climate changes, vegetation will reequilibrate according to the same relationship that currently exists. Thus, for example, as high latitudes warm, boreal forests will migrate northward and arctic tundra will be compressed into a narrower latitudinal band. Similarly, vegetation zones should shift upward in mountainous regions to track increases in temperature, a pattern of change that would reduce the extent and biodiversity of the uppermost, alpine tundra ecosystems. Zones of optimal grain production in North America may shift northward into Canada. In some continental regions, predictions are confounded by large uncertainties in precipitation

changes and the relative influence of increased temperature and altered precipitation in increasing or decreasing moisture stress to vegetation.

Although simple, an equilibrium approach is potentially misleading. Vegetation is unlikely to migrate rapidly enough to track the predicted pace of anthropogenic climate change, meaning there will be transient periods in which vegetation and climate are mismatched with respect to the equilibrium models. In addition, other determinants of future vegetation distribution, including herbivory, fire, severe storms, and human activities, can strongly affect vegetation composition, independent of the climate–vegetation correlation. In Australia, for example, fire frequency is expected to increase with climate change because of an increased probability of the high temperatures associated with ignition of bush fires. In the western United States, the frequency and severity of wildfires have increased as runoff from mountain snow occurs earlier, lengthening the potential fire season. Fire frequency and intensity may also be affected by changes in chemical properties of plant tissues produced in elevated CO₂. Outbreaks of insect herbivores are increasing in temperate and boreal forests, in part because milder winters increase insect survival, but their implications for local, regional, or global climate are yet to be explored. Given these complexities, coupled vegetation–AGCM models are increasingly including much transient, and in some cases nonlinear, behavior of vegetation distribution, behavior that can feedback to alter the pattern and pace of climate change.

Rising atmospheric CO₂

The stomatal response to CO₂ is critical to determining overall vegetation feedback to climate because it impacts latent and sensible heat exchange and return of water vapor to the atmosphere in transpiration. Experimental studies with a large number of plant species indicate that stomatal conductance is reduced by about 25% with a doubling of current atmospheric CO₂; there is considerable variation among species, with coniferous trees typically responding less than deciduous trees and herbaceous species typically showing the largest responses. The stomatal response to CO₂ can change over time as plant physiology adjusts to altered ecosystem biogeochemistry and the availability of essential resources, notably soil nitrogen. Whereas early AGCM studies addressed only the radiative effects of increased CO₂, current models specifically examine the role of vegetation physiology by allowing stomatal conductance and photosynthesis to respond to climatic conditions and atmospheric CO₂. Comparisons of AGCM simulations for the radiation-only (physical effects of CO₂) case to situations where vegetation physiology also responded in a realistic manner to increased CO₂ clearly illustrate the importance of vegetation feedbacks to the climate system. For tropical latitudes, the fully adjusted physiological response (reduced stomatal conductance) accounts for one-third of the nearly 3 °C increase in surface temperature. Global CO₂ assimilation by vegetation increased by 11–36% depending on the degree of physiological adjustment of levels of photosynthetic enzymes. Lesser increases in CO₂ assimilation and the larger increases in surface temperature come about when leaf nitrogen concentration, which is correlated with levels of photosynthetic

enzymes, decreases in elevated CO₂ and stomatal conductance declines strongly, consistent with maintenance of high photosynthetic water use efficiency. Thus, ecosystem biogeochemical responses, particularly components of the nitrogen cycle, can influence the response of water and energy exchanges to elevated CO₂.

Land cover and land use change

Ongoing human-caused changes in land use and land cover will undoubtedly impact biosphere–atmosphere interactions, water and energy exchanges being no exception. These impacts arise from the influence of key land surface parameters (roughness length, albedo, and LAI) on climate, as described in this article. Predictions of patterns of future land cover change carry substantial uncertainty, because land use depends on many different social and economic factors. It is safe to assume, however, that in tropical regions the trend toward increased conversion of forests to grasslands and croplands will continue. In the developed regions of the temperate zone, there has been a net increase in forested area in some regions as increased agricultural efficiency has allowed croplands to be abandoned and to convert back to forests. Increased rates of timber extraction in some regions such as Siberia may introduce climatic effects qualitatively similar to those estimated in simulations of extreme boreal deforestation. Retrospective studies of past, well-quantified patterns of conversion of natural vegetation to various agricultural conditions (pastures, crops, and associated irrigation) indicate that land cover change has been regionally important, but counteracting influences in tropical and temperate regions have probably caused little change in global temperature. In general, tropical forest conversion has warmed those regions, while temperate forest conversion has produced a cooling. This balance will likely not continue, however, because tropical deforestation has accelerated and temperate deforestation may be decelerating.

Summary and Conclusions

Together, the atmosphere and terrestrial biosphere form a coupled, interactive system, exchanging materials and energy in ways that critically regulate both biological and atmospheric processes (Figure 1). Biological production and consumption of greenhouse gases and aerosols influence the radiative balance of the atmosphere, while production and consumption of reactive gases affect atmospheric chemistry, including the formation and destruction of important pollutants. Similarly, plants mediate the partitioning of solar radiation incident on the terrestrial surface, partitioning that

affects regional and even global climate. Primarily mediated by plants and microorganisms, biosphere–atmosphere exchanges vary over the globe because of organisms' physiological and ecological responses to the environment. Ongoing, human-caused global changes are already altering biosphere–atmosphere exchanges, and these impacts are large enough that they could substantially shape the responses of the coupled biosphere–atmosphere system to future global changes. To better understand the consequences of anthropogenic global change, it will be critical to incorporate these complex interactions in AGCMs.

Acknowledgments

We thank Jeff Amthor, Paul Dijkstra, James Holton, Manuel Lerdau, and Oleg Menyailo for comments on an earlier version of this article, and Pieter Tans for permission to reproduce Figure 2 here. We thank the editors for inviting us to write this article.

See also: **Aerosols:** Climatology of Tropospheric Aerosols. **Boundary Layer (Atmospheric) and Air Pollution:** Overview. **Climate and Climate Change:** Carbon Dioxide; Energy Balance Climate Models; Overview. **Global Change:** Climate Record: Surface Temperature Trends. **Hydrology, Floods and Droughts:** Overview. **Land-Atmosphere Interactions:** Canopy Processes; Overview; Trace Gas Exchange.

Further Reading

- Bonan, G.B., 2008. Forests and climate change: forcings, feedbacks, and the climate benefits of forests. *Science* 320 (5882), 1444–1449. <http://dx.doi.org/10.1126/science.1155121>.
- IPCC, 2007. In: Solomon, S., Qin, D., Manning, M., Chen, Z., Marquis, M., Averyt, K.B., Tignor, M., Miller, H.L. (Eds.), *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.
- Mooney, H.A., Canadell, J.G., 2002. The Earth system: biological and ecological dimensions of global environmental change. In: Munn, T.A. (Ed.), 2002. *Encyclopedia of Global Change*, Vol. 2. John Wiley & Sons, Ltd, Chichester.
- Sellers, P.J., Dickinson, R.E., Randall, D.A., et al., 1997. Modeling the exchanges of energy, water, and carbon between continents and the atmosphere. *Science* 275, 502–509.
- Singh, B.K., Bardgett, R.D., Smith, P., Reay, D.S., 2010. Microorganisms and climate change: terrestrial feedbacks and mitigation options. *Nature Reviews Microbiology* 8, 779–790.

GRAVITY WAVES

Contents

Overview

Buoyancy and Buoyancy Waves: Optical Observations

Buoyancy and Buoyancy Waves: Theory

Gravity Waves Excited by Jets and Fronts

Convectively Generated Gravity Waves

Overview

DC Fritts, GATS Inc., Boulder, CO, USA

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

This article begins with a brief overview of the theory describing internal gravity wave structure, propagation, and influences in the atmosphere. Of particular relevance are the gravity wave characteristics that allow them to attain high altitudes and account for significant fluxes of energy and momentum. The dominant gravity wave sources are reviewed and the propagation of the resulting gravity waves through representative mean wind profiles is discussed. Also addressed are the dynamics that contribute to gravity wave spectral evolution, dissipation, turbulence generation, energy and momentum deposition, and the constraints imposed by these dynamics on the character of the gravity wave wave number spectra with increasing altitude. We conclude with a brief overview of the primary gravity wave influences on the mean circulation and structure of the atmosphere.

Introduction

The middle atmosphere refers to regions of the atmosphere extending from the tropopause (e.g., the top of the troposphere at $\sim 10\text{--}16$ km) to the homopause (an altitude of ~ 110 km, below which the atmosphere remains relatively well mixed). It includes (1) the stratosphere, extending from the tropopause to the stratopause at ~ 50 km in which mean temperature increases with altitude, (2) the mesosphere, extending from the stratopause to the mesopause (which varies in altitude between ~ 87 and 100 km) in which mean temperature decreases with altitude, and (3) the lower thermosphere above the mesopause in which mean temperature again increases with altitude. Radiative processes are largely responsible for the increasing temperatures with altitude in the stratosphere and thermosphere. In the mesosphere, however, dynamical mixing and transport processes contribute to the decrease of mean temperature with altitude. The mean state of the middle atmosphere is stably stratified and in approximate hydrostatic balance at all altitudes, and the corresponding mean density decreases approximately exponentially with increasing altitude. Stable stratification enables vertical wave propagation and decreasing mean

density causes wave amplitudes to increase dramatically with increasing altitude. Those waves for which gravity and stable stratification are the dominant influences are referred to as gravity waves.

Gravity waves play key roles in middle atmosphere dynamics for several reasons. They are ubiquitous throughout the atmosphere, arising in response to a variety of sources in the troposphere and middle atmosphere. Because they increase in amplitude as they propagate vertically, they dominate the motion spectrum at small and intermediate spatial and temporal scales (from a few to hundreds or thousands of kilometers and from minutes to tens of hours). More importantly for the middle atmosphere, they propagate readily over large distances and depths, account for substantial fluxes of energy and momentum, and experience filtering and dissipation due to interactions with the environments through which they propagate. The flux divergences arising from gravity wave dissipation, in particular, result in significant forcing of the large-scale circulation and the thermal and constituent structures of the middle atmosphere.

The goal here is to provide a brief survey of the theory describing gravity wave propagation and effects, the significant sources of gravity waves in the lower and middle atmosphere,

the factors influencing their evolution with increasing altitude, and their various roles in middle atmosphere dynamics.

Theory

Gravity waves are a subset of more general atmospheric motions that are described by equations expressing conservation of momentum, mass, and energy for a fluid system. These equations, known as the Navier–Stokes equations, are nonlinear, time dependent, dissipative, and have no general analytic solutions. We can gain significant insights into the structure and behavior of gravity waves, however, by seeking approximate solutions for linear waves in idealized environments.

Perturbation Equations

Making approximations that eliminate both acoustic and planetary wave solutions, the linear, inviscid equations describing gravity wave perturbations in a hydrostatic mean state with mean horizontal velocity $(\bar{u}, \bar{v})$ may be written as

$$du'/dt + w'\partial\bar{u}/\partial z + fu' + (1/\bar{\rho})\partial p'/\partial x = 0 \quad [1]$$

$$dv'/dt + w'\partial\bar{v}/\partial z + fv + (1/\bar{\rho})\partial p'/\partial y = 0 \quad [2]$$

$$dw'/dt + (1/\bar{\rho})\partial p'/\partial z - g\theta'/\bar{\theta} = 0 \quad [3]$$

$$d\theta'/dt + w'\partial\bar{\theta}/\partial z = 0 \quad [4]$$

$$\partial u'/\partial x + \partial v'/\partial y + (1/\bar{\rho})\partial(\bar{\rho}w')/\partial z = 0 \quad [5]$$

In these equations, u' , v' , w' , p' , and θ' are the perturbation zonal, meridional, and vertical velocities, pressure, and potential temperature; $\bar{\rho}$ and $\bar{\theta}$ are the mean density and potential temperature, respectively; $f = 2\Omega \sin \phi$ is the Coriolis parameter, with Ω the rotational frequency of the Earth and ϕ the latitude; and $d/dt = \partial/\partial t + \bar{u}\partial/\partial x + \bar{v}\partial/\partial y$ is an advective derivative.

Potential temperature is a conserved quantity for adiabatic (energy conserving) motions and is equal to the temperature of an air parcel restored adiabatically to a reference surface pressure. The static stability of the mean state (a measure of the degree of stratification of the atmosphere) may be expressed in terms of mean temperature or potential temperature as eqn [6], where g is the gravitational acceleration, $\bar{T}$ is the temperature, and C_p is the specific heat of air at constant pressure.

$$N^2 = (g/\bar{\theta})\partial\bar{\theta}/\partial z = (g/\bar{T})(g/C_p + \partial\bar{T}/\partial z) \quad [6]$$

The quantity N is the buoyancy frequency (see **Gravity Waves: Buoyancy and Buoyancy Waves: Optical Observations; Buoyancy and Buoyancy Waves: Theory**), which is the natural frequency of oscillation for an air parcel displaced vertically from its equilibrium altitude. We see from eqn [6] that N is larger (smaller) where temperature increases (decreases) with altitude. Thus, the stratosphere and lower thermosphere are more stratified than (and have higher static stability than) the troposphere and mesosphere.

To a good approximation, the atmospheric density varies as shown in eqn [7], where $H = R\bar{T}/g$ and R is the gas constant.

$$\bar{\rho}(z) = \rho_0 e^{-z/H} \quad [7]$$

For typical temperatures in the middle atmosphere, $H \approx 7$ km, so that mean density decreases by $\sim 10^6$ from the tropopause to the mesopause (~ 90 km).

In order to provide the simplest insights into gravity wave structure, we assume, for now, that the mean shear terms in eqns [1] and [2] may be neglected and that N^2 is constant, and seek solutions for u' having the form of eqn [8], with similar forms for other perturbation quantities.

$$u' = u_0 e^{z/2H} e^{i(kx + ly + mz - \omega t)} \quad [8]$$

These perturbations comprise solutions with wave number components (k, l, m) and a ground-based frequency ω . The exponential growth term in the perturbation results from the mean density decrease with altitude and implies that even small-amplitude waves may achieve appreciable amplitudes and effects at greater altitudes. Substitution into eqns [1]–[5] yields a set of algebraic equations for $(u', v', w', \theta', p')$ that can be combined to form a dispersion relation given by eqn [9].

$$m^2 = (k^2 + l^2)(N^2 - \omega_i^2)/(\omega_i^2 - f^2) - 1/4H^2 \quad [9]$$

Here, $\omega_i = \omega - k\bar{u} - l\bar{v}$ is the intrinsic frequency, or frequency of the gravity wave relative to the mean motion $(\bar{u}, \bar{v})$. The dispersion relation shown in eqn [9] relates the wave intrinsic frequency to the wave spatial structure (k, l, m) , mean atmospheric stability N , Coriolis parameter f , and density scale height H .

For illustration, we further assume that gravity wave propagation is in the zonal direction such that $l = 0$ and that the vertical wavelength, $\lambda_z = 2\pi/m \ll 4\pi H$, such that the last term in eqn [9] is negligible. Then the polarization relations that relate the various perturbation quantities can be expressed as in eqns [10]–[13].

$$v' = -ifu'/\omega_i \quad [10]$$

$$w' = -ku'/m \quad [11]$$

$$\theta'/\bar{\theta} = -iN^2 w'/\omega_i g = ikN^2 u'/m\omega_i g \quad [12]$$

$$p'/\bar{p} = -(N^2 - \omega_i^2)w'/m\omega_i = k(N^2 - \omega_i^2)u'/m^2\omega_i \quad [13]$$

The dispersion relation, eqn [9], reveals that gravity waves propagate vertically (with real m) only for $f < \omega_i < N$. For $\omega_i > N$, or more generally for $m^2 < 0$ owing to the last term in eqn [9], m is imaginary and gravity wave motions decay exponentially with altitude. Within the range where vertical propagation occurs, however, a wide range of wave motions is possible. For gravity waves with $\omega_i \sim f$ (called inertia-gravity waves), horizontal motions are ellipses with a major to minor axis ratio of ω_i/f , vertical motions are suppressed, $m^2 \gg k^2 + l^2$ (i.e., horizontal wavelengths are much larger than vertical wavelengths), pressure perturbations are negligible, and the dispersion relation takes the form of eqn [14], where $k_h = (k^2 + l^2)^{1/2}$ is the horizontal wave number in the direction of propagation.

$$m^2 = k_h^2 N^2 / (\omega_i^2 - f^2) \quad \text{or} \quad \omega_i^2 = k_h^2 N^2 / m^2 + f^2 \quad [14]$$

At the opposite extreme, with $\omega_i \sim N$, either large k_h (small horizontal wavelength, $\lambda_h = 2\pi/k_h$) or mean atmospheric density variations (the $1/4H^2$ term in eqn [9]) can result in gravity waves having large vertical wavelengths, $\lambda_z = 2\pi/m$, and large vertical motions, and the appropriate dispersion relation is given by eqn [15].

$$m^2 = k_h^2 N^2 / \omega_i^2 - k_h^2 - 1/4H^2 \quad [15]$$

For gravity waves having intermediate intrinsic frequencies, $f^2 \ll \omega_i^2 \ll N^2$ (describing a large fraction of relevant gravity waves), transverse motions are negligible ($v' \sim 0$), motions are largely in a vertical plane, vertical motions are larger than for inertia-gravity waves, and gravity waves easily propagate vertically. In this case, the dispersion relation (eqn [16]) is especially simple and insightful.

$$m^2 = k_h^2 N^2 / \omega_i^2 \quad [16]$$

This yields an intrinsic frequency given by eqn [17], where ϕ is the angle of the gravity wave phase surfaces from the horizontal.

$$\omega_i = \pm k_h N / m = N \sin \phi \quad [17]$$

We now return to the influences of mean wind shear that were neglected above to allow a simple and insightful derivation of the dispersion and polarization relations. Influences of mean wind shear need to be included in more general assessments of gravity wave behavior because the large majority of gravity waves arising in the lower atmosphere and influencing the middle atmosphere have phase velocities within the range of large-scale winds (including mean, tidal, and planetary wave winds) in the lower and middle atmosphere. Writing the component of the mean wind $(\bar{u}, \bar{v})$ in the direction of gravity wave propagation as $\bar{u}_h(z)$, the ground-based and intrinsic frequencies may be written as $\omega = k_h c$ and $\omega_i = k_h (c - \bar{u}_h(z)) = k_h c_i(z)$, respectively (where c_i is the intrinsic phase speed along the direction of gravity wave propagation). For intermediate-frequency gravity waves, eqn [16] may then be written as eqn [18].

$$m(z) = \pm N / (c - \bar{u}_h(z)) = \pm N / c_i(z) \quad \text{or} \quad \lambda_z(z) = 2\pi c_i(z) / N \quad [18]$$

Thus, m and λ_z are now functions of z that are simply related to the atmospheric stability, N , and the variable intrinsic gravity wave phase speed, c_i , due to mean wind variations in altitude.

Wind shear and curvature also have implications for gravity wave structure and propagation that require a more general dispersion relation for mid- and high-frequency gravity waves accounting for variable $\bar{u}_h(z)$ and displayed as eqn [19].

$$m^2 = k_h^2 N^2 / \omega_i^2 - k_h \bar{u}_z / \omega_i H + k_h \bar{u}_{zz} / \omega_i - k_h^2 - 1/4H^2 \quad [19]$$

This dispersion relation differs from that in eqn [15] by inclusion of shear and curvature terms, where $\bar{u}_z(z)$ and $\bar{u}_{zz}(z)$ are the first and second derivatives of $\bar{u}_h(z)$ in the vertical, and by allowing all quantities except k_h to now vary in z . Variable mean wind in altitude along the gravity wave propagation direction allows two limiting cases of special interest because of

their impacts on gravity wave structure and propagation. One occurs where ω_i approaches zero, termed a critical level with $c_i(z) = 0$. The behavior of the solution in the vicinity of a critical level at $z = z_0$ in a linear mean wind shear, $\bar{u}_h(z) = U_z(z - z_0)$, is obtained as the lowest order term in a Frobenius series expansion and has the form given by eqn [20].

$$w'(z) \sim \delta z^{1/2} e^{\pm i\mu \delta z} \quad [20]$$

Here, $\delta z = z - z_0$, $\mu = (\text{Ri} - 1/4)^{1/2}$, and $\text{Ri} = N^2 / U_z^2 > 1/4$ for stability. Equations [11] and [20] imply $u'_h(z) \sim \delta z^{-1/2} e^{\pm i\mu \delta z}$ such that the gravity wave vertical flux of horizontal momentum (see below), $\bar{\rho} u'_h w' / 2$ is constant with altitude for conservative wave propagation.

The steady-state critical-level gravity wave structure inferred above is singular and implies strong gravity wave attenuation for realistic propagation environments having $\text{Ri} \gg 1/4$. It also implies strong potential roles for viscosity, nonlinearity, and transient effects, all of which have been found to contribute to gravity wave instability and/or dissipation in such regions. Indeed, these various processes typically play significant roles in constraining gravity wave amplitudes before a critical level is approached, causing a critical level to be a very effective barrier to continued vertical propagation of the incident gravity wave.

A second barrier to upward gravity wave propagation occurs where increasing c_i (or ω_i) and/or decreasing $N(z)$ with increasing altitude in eqn [19] results in $m^2 = 0$, termed a turning level, above which the gravity wave is evanescent ($m^2 < 0$), resulting in vertical reflection. More generally, variable mean winds and atmospheric stability can yield a layer within which a gravity wave has $m^2 > 0$ and propagating character, but above and below which $m^2 < 0$ and the gravity wave is evanescent. Such a layer is termed a duct and the gravity wave is termed a ducted gravity wave. It remains effectively trapped within the duct until the environment changes, or the gravity wave is dissipated, tunnels through an evanescent layer to another region of vertical propagation, or transfers energy to other motions via nonlinear interactions that may or may not be ducted. A final effect of mean winds on gravity wave propagation seen in eqn [19] is the new curvature term, $k_h \bar{u}_{zz} / \omega_i$, which, for $c_i > 0$ or $c > \bar{u}_h(z)$, causes a reduction in m^2 in the duct with $\bar{u}_{zz}(z) < 0$, where $k_h^2 N^2 / \omega_i^2$ is maximum because of small $c_i > 0$.

Ray paths in the plane of gravity wave propagation depicting the various gravity wave responses to a mean shear layer, including critical level absorption, reflection at a turning level, and refraction to larger and smaller intrinsic phase speeds (or frequencies) in an environment with constant $N(z)$ are illustrated in Figure 1. For illustration, we assume $|\omega_i| = |k_h N / m| = N |\sin \phi|$ is constant for all cases. Thus, all gravity waves have the same initial frequency, but varying intrinsic frequency, wavelength ratio, and vertical wavelength after encountering the shear layer. In Figure 1, ray path origins denote initial phase speeds relative to the mean wind. Gravity waves propagating to the right (along the mean wind) experience decreasing $|\omega_i|$ and phase slopes, those propagating to the left experience increasing $|\omega_i|$ and phase slopes. Smaller initial phase speeds are more susceptible to refraction. Dashed ray paths denote gravity waves that are prevented from propagating to higher altitudes due to either a critical level or a turning level, occurring where $\omega_i = 0$ or $m^2 = 0$ ($\omega_i \sim N$), respectively.

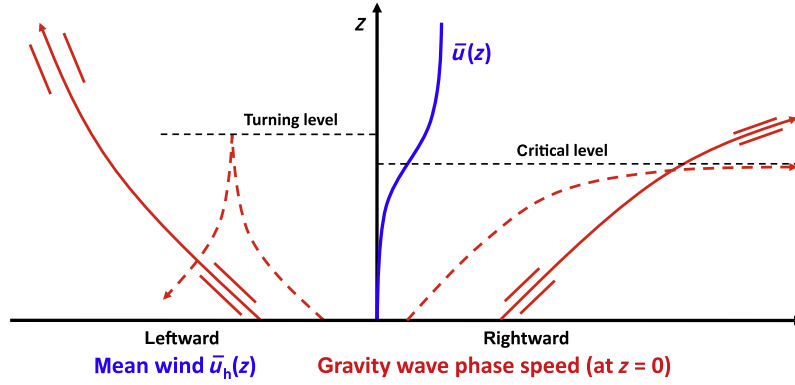


Figure 1 Ray paths for gravity waves having common initial intrinsic frequencies $\omega_i = N/1.4$, but different phase speeds and directions of propagation, encountering a mean wind shear. Solid lines denote gravity waves and phase structures that are refracted to higher (left) and lower (right) intrinsic frequencies. Dashed lines denote gravity waves that encounter a turning level (left) or critical level (right) due to more severe refraction accompanying smaller initial phase speeds.

The ray paths shown in **Figure 1** lead naturally to a discussion of gravity wave group velocity. This is the velocity at which a gravity wave packet propagates and transports energy and momentum. The group velocity components for medium and high-frequency gravity waves in the plane of wave propagation derived from [eqn \[19\]](#), assuming mean wind shear and curvature are small, are shown in [eqn \[21\]](#).

$$(c_{gh}, c_{gz}) = (\partial\omega/\partial k_h, \partial\omega/\partial m) = (\bar{u}_h, 0) + [k_h(N^2 - \omega_i^2), -m\omega_i^2] / [\omega_i(k_h^2 + m^2 + 1/4H^2)] \quad [21]$$

[Equation \[21\]](#) illustrates the two limiting cases for gravity wave vertical propagation nicely. For both a critical level and a turning level, [eqn \[21\]](#) implies (c_{gh}, c_{gz}) approaches $(\bar{u}_h, 0)$ and the gravity wave ceases to propagate vertically. For intermediate-frequency gravity waves for which $f^2 \ll \omega_i^2 \ll N^2$, [eqn \[21\]](#) reduces to [eqn \[22\]](#) below.

$$(c_{gh}, c_{gz}) = (\bar{u}_h, 0) + (\omega_i/k_h, -\omega_i/m) \quad [22]$$

Thus, for most gravity waves of importance in the middle atmosphere, the horizontal group velocity is the same as the horizontal phase speed, but the opposite progression in the vertical. Except for cases for which $\bar{u}_h(z)/c < -1$, downward phase progression implies upward gravity wave propagation and energy and momentum transport. This also implies that for upward gravity wave propagation, having c_i and $\omega_i > 0$, $c_{gz} = -\omega_i/m > 0$ and $m < 0$.

Low-frequency gravity waves are also described by the dispersion relation in [eqn \[9\]](#) for intrinsic frequencies $\omega_i^2 \ll N^2$ and $m^2 \gg 1/4H^2$ yielding the approximate relation in [eqn \[23\]](#).

$$\omega_i^2 = N^2 k_h^2 / m^2 + f^2 \quad [23]$$

The group velocity for low-frequency gravity waves is nearly horizontal, with the ratio of components $|c_{gz}/c_{gh}| = |k_h/m| = (\omega_i^2 - f^2)^{1/2}/N$. As a result, these gravity waves can propagate very large distances horizontally. While their vertical propagation is very slow, they nevertheless often comprise a large fraction of the total gravity wave kinetic and

potential energy and contribute to the environments through which intermediate- and high-frequency gravity waves propagate.

Gravity wave structure, refraction by mean winds, and amplitude growth with altitude noted above imply significant effects where amplitudes become large. Exponential amplitude growth with altitude enables even very small-amplitude gravity waves to achieve large amplitudes at high altitudes. The polarization relations imply that gravity wave propagation provides energy and momentum transport due to the fluxes arising from the gravity wave velocity and pressure fluctuations, the effects of which also grow exponentially with altitude. Of these fluxes, the vertical transport of horizontal momentum plays the dominant role, effectively transporting horizontal momentum in the direction of gravity wave propagation from the gravity wave source to the region in which the wave is dissipated. Other correlations arise accompanying mean density variations, compressibility, and dissipation, but these will not be addressed here. Finally, amplitude growth with increasing altitude ultimately yields sufficiently large gravity wave amplitudes, and momentum fluxes, to initiate a spectrum of instabilities. These various instabilities extract energy from the gravity wave, reducing its amplitude and momentum flux, thus causing a momentum flux divergence that acts as a local body force on the mean flow.

Considering first the effects of amplitude growth, a conservative gravity wave will experience an amplitude growth of $\sim e^{z/2H}$ according to [eqn \[8\]](#). In reality, however, dissipation and interactions occur at all altitudes and affect all gravity waves, hence the observed growth in gravity wave variances is less than half the conservative rate, with a ‘variance’ scale height $H_E \approx 2.3H$ throughout the lower and middle atmosphere. A growing gravity wave ultimately reaches an amplitude above which the atmosphere becomes locally ‘convectively unstable,’ having an inverted potential temperature gradient due to relative advection along slanted gravity wave phase lines. This condition may be written in terms of the gravity wave variables as in [eqn \[24\]](#).

$$(\partial\theta'/\partial z)_{\min} < -\partial\bar{\theta}/\partial z \quad \text{or} \quad |u'_h|_{\max} > c_i \quad [24]$$

In principle, gravity waves can succumb to instabilities at significantly smaller amplitudes, as all gravity waves at all amplitudes are unstable. Equation [24] nevertheless provides a useful guide for the likely occurrence of instability because instabilities typically progress much more rapidly at larger amplitudes. The consequence of instability is a reduction of the gravity wave amplitude to a value often much smaller (by 30% or more) than the nondimensional amplitude defined by the ratio of terms in eqn [24], e.g., $a = |u'_h|_{\max}/c_i > 1$, where $a = 1$ corresponds to incipient overturning and will be referred to hereafter as the 'breaking' amplitude near which instabilities become more likely. Equation [24] and the linear gravity wave structure near a critical level reinforce the expectation that critical level approach also yields gravity wave instability and dissipation. Given the impacts of instability on gravity wave amplitudes, propagation, and the environments in which they propagate, the mean state responses, and the character and effects of instabilities will be addressed in greater detail in the following sections.

Mean State Equations

The role of gravity waves in forcing the mean circulation and thermal structure of the middle atmosphere can be understood by examining the zonally averaged momentum and continuity equations. Assuming for simplicity a steady mean flow with negligible vertical shear, the appropriate equations are eqns [25] and [26] with a mean meridional and vertical (or residual) circulation, $(\bar{v}^*, \bar{w}^*)$, satisfying eqn [27].

$$-f\bar{v}^* = -(1/\bar{\rho})\partial F_x/\partial z = D_{F_x} \quad [25]$$

$$f\bar{u} + (1/\bar{\rho})\partial \bar{p}/\partial y = -(1/\bar{\rho})\partial F_y/\partial z = D_{F_y} \quad [26]$$

$$\partial \bar{v}^*/\partial y + (1/\bar{\rho})\partial (\bar{\rho} \bar{w}^*)/\partial z = 0 \quad [27]$$

Here, the asterisks denote the transformed Eulerian mean circulation, which expresses the net effects of wave forcing, and (F_x, F_y) is given by eqn [28] for a gravity wave having arbitrary propagation direction and intrinsic frequency ω_i .

$$(F_x, F_y) = \bar{\rho}(\overline{u'w'}, \overline{v'w'}) (1 - f^2/\omega_i^2) \quad [28]$$

For a gravity wave propagating upward and eastward relative to the mean flow (i.e., $c_{gz} > 0$, $c_{pz} < 0$, $k > 0$, and $c_i > 0$), we see from eqns [22] and [11] that $m < 0$ and $\overline{u'w'} > 0$. The terms (F_x, F_y) represent the mean vertical fluxes of horizontal momentum by gravity waves in the zonal and meridional directions. Momentum fluxes are nondivergent for conservative wave motions but are generally divergent when wave dissipation occurs. In the absence of wave dissipation, the solution to eqns [25] and [27] is a geostrophically balanced zonal mean flow with $(\bar{v}^*, \bar{w}^*) = 0$ and $\bar{u} = -(1/\bar{\rho}f)\partial \bar{p}/\partial y$. Momentum flux divergences arising from wave instability or filtering processes, on the other hand, lead to significant departures from a geostrophic mean flow and to a wave-driven residual circulation and thermal structure exhibiting large departures from radiative equilibrium.

Gravity Wave Sources

Two major gravity wave sources in the troposphere are believed to be topography and convection. Airflow over mountains leads to vertical displacements dictated by the terrain height and horizontal scale at zero phase speed relative to the ground. For horizontal wavelengths of ~ 10 km or longer, the vertical wavelength is determined by eqn [16] with $c = 0$: $\lambda_z \sim 2\pi \bar{u}_h(z)$. Mountain waves are prevented from reaching high altitudes by mean winds that cause critical levels near $\bar{u}_h(z) = 0$, such as occurs for nearly zonal propagation at middle and high latitudes in summer, or at low latitudes due to wind reversals in the stratosphere accompanying the quasi-biennial oscillation. In winter, however, sustained eastward winds extending into the mesosphere provide a propagation channel for the larger mountain wave horizontal wavelengths having largely zonal propagation and $m^2 > 0$. In such cases, these waves can reach the mesosphere or lower thermosphere, depending on the mean wind and tidal environment at those altitudes. Two examples of mountain waves seen by the Atmospheric Infrared Sounder (AIRS) on the Aqua satellite at ~ 41 km and one seen in the OH airglow at ~ 87 km are shown in the upper two and lower left panels of Figure 2.

Convection excites gravity waves by vertical motions within clouds, through buoyancy driven by latent heating, or by airflow over clouds in a sheared environment. The scales and frequencies of gravity waves that arise are dictated by the geometry and timescale of the convection. Shallow, broad convection excites primarily lower frequency gravity waves at smaller vertical scales; deep, narrow convection excites higher frequency gravity waves having larger vertical scales and vertical group velocities. Additionally, cloud complexes may act collectively as gravity wave sources on much larger spatial scales. Convection thus excites gravity waves spanning a wide range of scales and phase speeds. Deep and fast convective gravity waves (those having large vertical scales and relatively small horizontal scales, approximately tens to hundreds of kilometers) have sufficiently large horizontal phase speeds to often penetrate well into the mesosphere and thermosphere. An example of concentric gravity waves seen in OH airglow at ~ 87 km excited by deep convection and propagating through a weak mean wind environment at lower altitudes is shown in the lower right panel of Figure 2.

At greater altitudes, jet streams excite gravity waves through a process known as spontaneous imbalance through which an unbalanced flow achieves a new balanced state and conserves energy via gravity wave emission. The initial unbalanced flows are typically relatively shallow and broad (approximately 10 km deep and hundreds of kilometers across the jet stream), resulting in gravity waves having large horizontal scales, near-inertial frequencies, and small vertical group velocities. Such waves comprise the most energetic part of the gravity wave frequency spectrum in the stratosphere and mesosphere, but often propagate several thousand kilometers horizontally before reaching these higher altitudes. Unstable jet stream shear flows, often enhanced by inertia-gravity waves, also excite gravity waves through a process called envelope radiation, in which gravity wave scales are imposed by the event scale rather than by the detailed dynamics of shear instability itself.

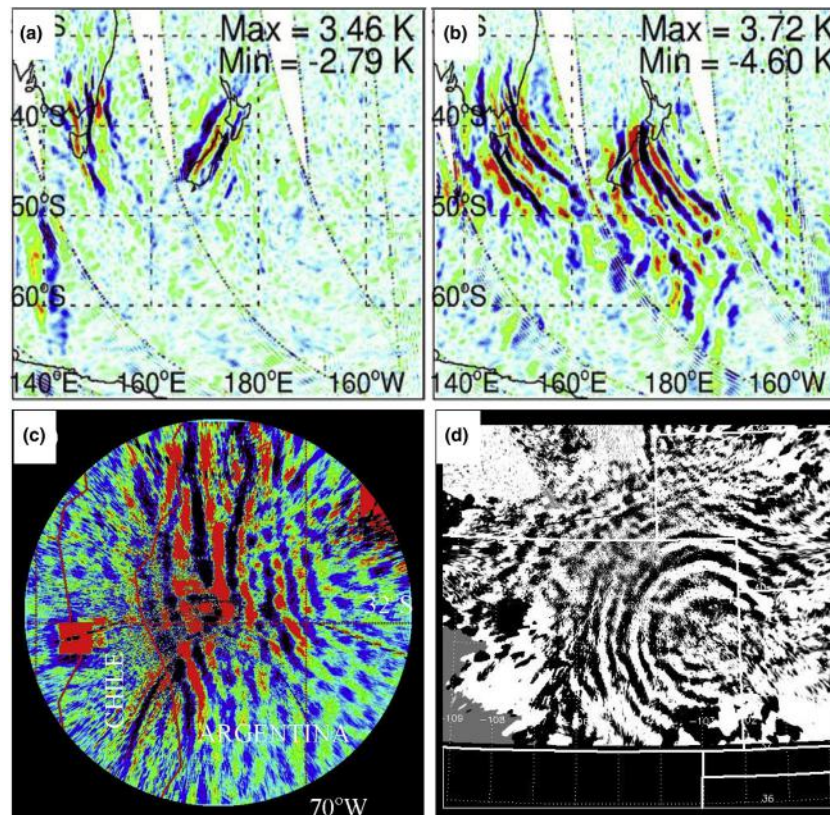


Figure 2 Apparent mountain waves seen in AIRS temperatures at 2 hPa (~ 41 km) over New Zealand and Tasmania on 6–7 July 2011 (a and b; courtesy of Eckermann, S.D. (unpublished)) and in the OH airglow layer (~ 87 km) over the Andes centered at $\sim 32^\circ$ S on 3 July 2009 (c; courtesy of Smith, S., Baumgardner, J., Mendillo, M., 2009. Evidence of mesospheric gravity-waves generated by orographic forcing in the troposphere. *J. Geophys. Res.* 36, L08807. doi:10.1029/2008GL036936.). Shown in (d; courtesy of Yue, J.) are convective gravity waves due to deep convection over Northeast Colorado seen in OH airglow (~ 87 km) at a time of light mean winds.

Additional gravity wave sources at higher altitudes accompany gravity wave amplitude growth with altitude. As noted above, amplitude growth enhances the growth rates for wave-wave interactions and local shear- and buoyancy-driven instabilities. Both processes lead directly to excitation of additional gravity waves at scales defined by the specific wave-wave interactions or local instability dynamics. Local body forcing due to momentum flux divergence arising from localized instabilities and dissipation also generate what are called secondary gravity waves. Secondary gravity waves have scales imposed by the geometry of the forcing, allowing them to often have significantly larger scales, and higher propagation altitudes, than the primary gravity wave. Evidence for these additional gravity wave sources has been provided largely by numerical studies to date.

Spectra and Evolution with Altitude

Near major sources in the troposphere, the spectrum of gravity waves is determined largely by the source characteristics. But as altitude increases, the spectrum is shaped increasingly by other factors. Contributions from various sources necessarily broaden the spectrum, with topography, convection, jet streams, wind shears, and other sources at higher altitudes

typically contributing very different spatial scales, horizontal phase speeds, and intrinsic frequencies. Thereafter, variations in mean winds cause intrinsic frequencies and corresponding vertical wavelengths to also vary considerably with altitude. An illustration of gravity wave propagation for a mountain wave (with $c = 0$) and two convective gravity waves having equal and opposite zonal phase speeds in an idealized midlatitude summer zonal wind profile is provided in the left panel of Figure 3. For the convective gravity wave having $c > \bar{u}(z)$ on the right, increasing c_i above the eastward wind maximum at lower altitudes increases λ_z , accelerates vertical propagation, and enables the gravity wave to penetrate to much higher altitudes. Those gravity waves with $c = 0$ and $c < 0$, however, encounter critical levels where $c = \bar{u}(z)$ and are dissipated by viscosity or instability processes below this altitude. Hence, gravity wave propagation in mean wind shear redistributes wave energy in vertical wave number and removes those components of the spectrum that encounter critical levels within this altitude range. Additional spectral transfers accompany pervasive wave-wave interactions that act throughout the atmosphere.

The approximate constraints placed on gravity wave amplitudes described by eqn [24] imply what is termed ‘saturation’ of the gravity wave spectrum, which imposes an approximately universal spectral character where these dynamics occur. This

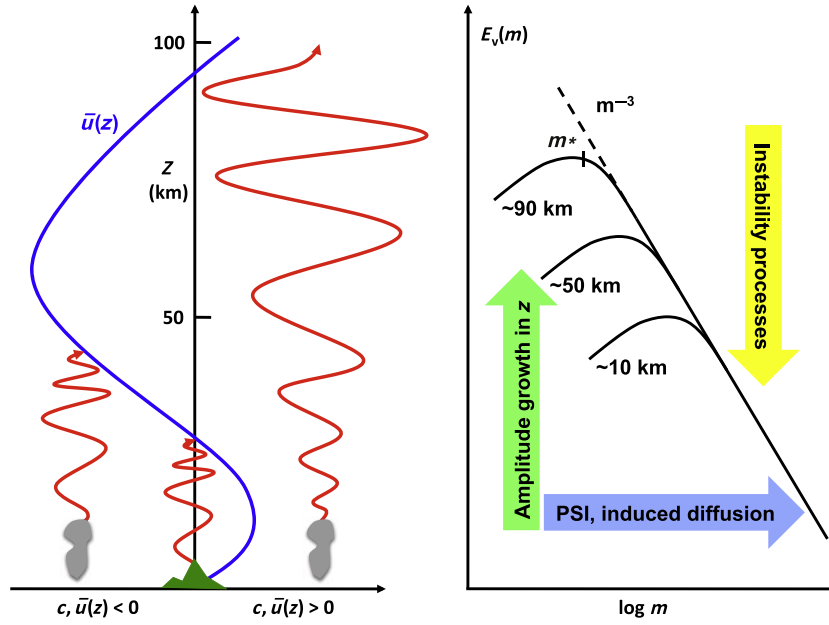


Figure 3 Gravity wave vertical wavelength dependence (left panel) on $c_i = c - \bar{u}_h(z)$ for a mountain wave having $c = 0$ and two convective gravity waves with westward ($c < 0$, left) and eastward ($c > 0$, right) phase speeds in a summer midlatitude zonal wind. Evolution of the vertical wave number spectrum of horizontal velocity variance $E_v(m)$ with altitude (right panel). The slope at large m is due to the constraint on wave amplitudes imposed by gravity wave saturation processes, allowing variance increases only at smaller m and causing a reduction in the characteristic vertical wavelength, m^* , with altitude. Induced diffusion and PSI in variable wind environments account largely for the transfer energy to smaller vertical scales.

‘saturated’ spectrum has a form and amplitude at large m given approximately by $E_v(m) \approx N^2/6m^s$, with $s \sim 2.5$ – 3 suggested by multiple observations. Instabilities and dissipation occur preferentially at smaller vertical wavelengths because these gravity waves are more easily Doppler shifted toward smaller $|c_i|$ by changing mean winds. Gravity waves at larger vertical wavelengths tend to be excited at much smaller initial amplitudes, have larger phase speeds so are less influenced by mean winds, grow approximately conservatively with altitude, and typically propagate to significantly higher altitudes before they attain sufficiently large amplitudes to participate in saturation processes. Wave–wave interactions, studied extensively in both oceanic and atmospheric contexts, likewise act to shape the gravity wave wave number and frequency spectra.

These various dynamics lead to a vertical wave number spectrum of horizontal velocity having an approximate form and amplitude given by $E_v(m) \approx N^2\mu/6(1 + \mu^4)$, where $\mu = m/m^*$ and m^* is a characteristic vertical wave number that decreases with altitude. Approximately conservative gravity wave propagation at small m yields a corresponding growth of total gravity wave kinetic energy (per unit mass) with a scale height $H_E \sim 2.3H$, as noted above. This approximate spectral form, and its evolution with altitude (for which there is now considerable observational and theoretical evidence), is illustrated in the right panel of Figure 3. The growth of horizontal kinetic energy with altitude occurs largely at $m < m^*$ and equates to an increase of ~ 100 between the tropopause and the mesopause and a corresponding increase of m^* by ~ 10 over this same altitude range. Characteristic vertical wavelengths are typically ~ 1 – 3 km in the lower stratosphere and

~ 10 – 30 km near the mesopause, with a similar increase in the magnitude of horizontal velocities (i.e., from ~ 3 m s $^{-1}$ at lower altitudes to ~ 30 m s $^{-1}$ near the mesopause). Despite the growth with altitude of gravity wave vertical wavelengths and amplitudes, the kinetic energy (per unit volume) decreases by $\sim 10^3$ between the tropopause and mesopause, implying that the large majority of gravity wave energy does not propagate to high altitudes, but rather contributes to instabilities, turbulence, mixing, and transport throughout the lower and middle atmosphere.

Also observed are ground-based and intrinsic frequency spectra, and airborne horizontal wave number spectra, of horizontal winds (and temperatures) that vary as $E_u(\omega) \sim \omega^p$ and $E_u(k_h) \sim k_h^p$ where $p \sim -5/3$, with corresponding frequency spectra of vertical velocities having the form $E_w(\omega) \sim \omega^{p+2}$ (except near f and N) under light wind conditions for which Doppler shifting effects are small. These frequency spectra imply that the major contributions to gravity wave momentum flux occur for the highest frequency gravity waves, with periods $T_{GW} < 1$ h contributing the dominant momentum fluxes at most altitudes.

Instability Dynamics

As noted above, there is a broad spectrum of instabilities that cause energy exchanges among various modes and/or gravity wave dissipation, amplitude reductions, mixing and transport, and energy and momentum flux divergence. Resonant and off-resonant wave–wave interactions transfer energy to gravity

waves having different frequencies and wave numbers than the parent gravity wave(s), but typically without explicit dissipation. Local instabilities also exhibit many forms, depending on gravity wave frequency, amplitude, spatial localization, and environment. Collectively, these interactions and instabilities are responsible for the approximately universal character of the gravity wave spectrum in the atmosphere, and its analog in the oceans. Indeed, approximate descriptions of these various dynamics underlie all parameterization schemes intended to describe gravity wave propagation, spectral evolution, energy and momentum transport and deposition, and their influences throughout the atmosphere.

Resonant wave-wave interactions comprise triad interactions satisfying conditions given by eqn [29].

$$\mathbf{k} = \mathbf{k}_1 + \mathbf{k}_2 \quad \text{and} \quad \omega_i = \omega_{i1} + \omega_{i2} \quad [29]$$

Here, $\mathbf{k}$ and ω_i denote the primary gravity wave's wave number vector and intrinsic frequency, and subscripts denote these quantities for secondary gravity waves arising from these interactions. There are three primary classes of interactions having different character. These include (1) elastic scattering, (2) induced diffusion, and (3) parametric subharmonic instability (PSI). Each class has unique characteristics. Elastic scattering refers to the backscatter of an upward propagating gravity wave into a downward propagating gravity wave (or vice versa) of comparable vertical wave number by a low-frequency gravity wave (or mean motion) having twice the vertical wave number (i.e., Bragg scattering). Induced diffusion amounts to a transfer of energy from one gravity wave to another that has an almost identical wave number via interaction with a low-frequency gravity wave (or mean motion) having a much smaller vertical wave number (i.e., analogous to refraction of a gravity wave packet in a large-scale shear flow). Finally, PSI refers most often to the exchange of energy between a primary gravity wave of intermediate frequency and two secondary gravity waves having approximately half the primary gravity wave frequency and nearly opposite (and more nearly vertical) wave numbers. Of these interactions, PSI plays the major role at low frequencies and high wave numbers, and induced diffusion dominates energy transfers at high frequencies and high wave numbers. Wave-wave interactions are weaker (hence gravity wave propagation is more nearly linear) at low frequencies and low wave numbers, $m < m^*$, where gravity wave amplitudes are below their saturated values and grow strongly with increasing altitude.

Higher-order, off-resonant interactions can also occur, particularly in cases in which gravity waves having significant amplitudes and different characters are superposed. In such cases, mutual deformation of the velocity fields leads to rapid energy transfers and additional gravity wave excitation. Collectively, these various wave-wave interactions cause the gravity wave spectrum to depart increasingly from a collection of independent, discrete, linear gravity waves as altitude and gravity wave amplitudes increase.

Of more direct relevance in terms of gravity wave effects in the middle atmosphere are the instabilities that lead to turbulence, gravity wave dissipation, and energy and momentum flux divergence. Important insights into the character of these instabilities can be obtained by employing various theories that describe possible instability character in

general sheared flows for which there is not a set of linear, orthogonal eigenvectors. Optimal perturbation theory (or singular vector analysis) enables the identification of both transient and 'global' perturbation characters and growth rates in general flows. Floquet theory similarly allows a perturbation analysis of spatially and temporally variable flows that typically corresponds to the global optimal perturbations that most often dictate finite-amplitude responses. Analytical and numerical efforts have also assessed instabilities accompanying wave-mean flow interactions, while direct numerical simulations (DNSs) offer an ability to evaluate instability dynamics and the resulting turbulence structure and morphology for idealized and more complex flows. These studies reveal multiple instabilities that depend to varying degrees on the characteristics of the underlying gravity waves and their environments. Here, we review those instability types that appear to contribute most significantly to gravity wave dissipation for various wave frequencies.

At near-inertial frequencies, gravity wave character is strongly influenced by rotation (see eqns [10] and [14]), they evolve slowly due to small vertical group velocities, and they support instabilities expected in approximately plane parallel shear flows. These include the familiar Kelvin-Helmholtz (KH) shear instabilities at Richardson numbers, $Ri = N^2(d\bar{u}_h/dz)^2 < 1/4$, and to an instability comprising counter-rotating rolls aligned along the shear when $N^2 < 0$. The latter instability closely resembles Langmuir circulations in a sheared ocean mixed layer, roll vortices in a sheared, convective atmospheric boundary layer, and the secondary instabilities that account for the transition to turbulence that arise in the highly sheared outer portions of KH billows at small Ri and sufficiently high Reynolds number, Re . A KH billow and secondary instabilities are illustrated in Figure 4 during the transition to turbulence for a shear flow having uniform N , $\bar{u}_h(z) = U_0 \tanh(z/h)$, where U_0 and h are the initial half velocity difference and the half shear depth, an initial $Ri = 0.05$, and $Re = U_0 h/\nu = 10\,000$.

Intermediate-frequency gravity waves exhibit wave 'breaking,' which can occur for waves that are convectively unstable or stable ($Ri \sim 1$ or greater), and Kelvin-Helmholtz instability (KHI), especially in the presence of mean wind shear or superposed gravity waves. Wave breaking can display a range of behaviors and orientations, depending on the gravity wave amplitude, intrinsic frequency, and Re . An example of wave breaking in a high-resolution DNS resolving the turbulence dynamics for a gravity wave having $a = 0.9$, $\omega_i = N/3.2$, a minimum flow $Ri \approx 3.8$, and $Re = \lambda_z^2/T_b\nu = 10\,000$, where $T_b = 2\pi/N$, is illustrated in Figure 5 with three-dimensional (3-D) images of λ_2 (a measure of flow rotation defined in terms of the rotation and strain tensors) during initial instability and following the transition to turbulence. Despite the high initial Ri , this gravity wave is highly unstable and exhibits a rapid transition to strong turbulence. Similar dynamics occur for a wide range of frequencies and spatial scales throughout the lower and middle atmosphere and can extend into the lower thermosphere for sufficiently large scales, amplitudes, and Re .

High-frequency gravity waves likewise exhibit a range of instabilities. Wave breaking at large amplitudes is similar to that for intermediate-frequency gravity waves. These gravity waves also enable instabilities accompanying evolution of the mean

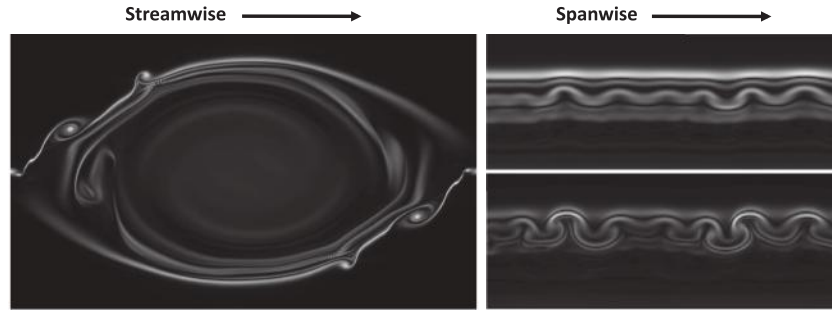


Figure 4 KH billow structure depicted with vorticity magnitude for $Ri = 0.05$ and $Re = 10\,000$ at maximum billow amplitude. A streamwise-vertical cross section (left) shows vorticity maxima accompanying secondary instability structures having spanwise (smaller-scale KHI at the edges) and streamwise (billow interior) alignments. Streamwise-aligned instabilities in a spanwise-vertical plane at the top of the primary KH billow at the same time, and $\sim 0.1 T_b$ later, are shown at right. Fritts, D.C., Wan, K., Franke, P., Lund, T., 2012. Computation of clear-air radar backscatter from numerical simulations of turbulence, III: off-zenith measurements and biases throughout the lifecycle of a Kelvin-Helmholtz instability. *Journal of Geophysical Research* 117, D17101. doi:10.1029/2011JD017179.

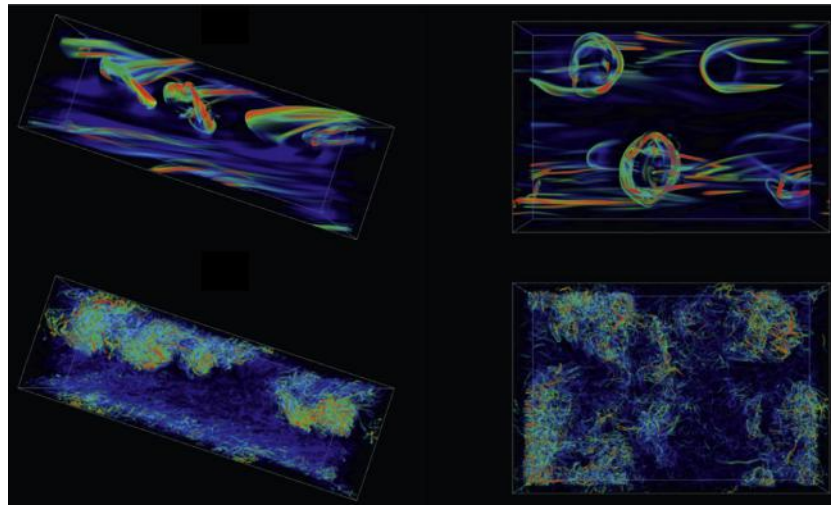


Figure 5 3-D vorticity structures during wave breaking for a gravity wave having $a = 0.9$, $\omega_i = N/3.2$, a minimum initial $Ri \approx 4$, and $Re = 10\,000$ separated by ~ 1.5 wave periods. Left views of the computational domain tilted along the gravity wave phase are from the side. Right views are from below. Strong vorticity is shown with yellow and red. The strongest turbulence accompanies plunging motions within the vortex rings seen forming at upper right. Fritts, D.C., Wang, L., Werne, J., Lund, T., Wan, K., 2009. Gravity wave instability dynamics at high Reynolds numbers, 2: turbulence evolution, structure, and anisotropy. *Journal of Atmospheric Sciences* 66, 1149–1171. doi:10.1175/2008JAS2727.1.

flow, termed ‘self-acceleration’ instability. This happens when the mean flow in which the wave resides is accelerated beyond the horizontal group velocity and is a consequence of the induced mean motion increasing as $\Delta \bar{u} \sim 1/\bar{\rho}$ because of its dependence on gravity wave momentum flux, $\bar{\rho} u_h' w'(z)$, whereas gravity wave velocities vary as $u_h' \sim 1/\bar{\rho}^{1/2}$. A DNS of such a self-acceleration event is shown in **Figure 6** with w' fields for a gravity wave packet localized in z , spanning 5 gravity wave periods, and having $\lambda_h = 20$ km and $\omega_i = N/1.4$. Instability accompanies the phase structure becoming vertical (fifth panel), which results from accelerations along the gravity wave propagation direction where amplitudes increase with time and decelerations where amplitudes decrease with time. Self-acceleration instability is 2-D, though 3-D instabilities also occur thereafter. The gravity wave amplitude decays and momentum transport ceases accompanying instability, resulting in local forcing that acts as a source of secondary gravity wave generation.

The instability dynamics reviewed above almost always occur in environments representing a superposition of gravity waves having a range of frequencies, spatial scales, and orientations. Such environments can significantly enhance the potential for instability and turbulence through interactions among finite-amplitude motions that readily yield flow deformations and energy transfers among modes (e.g., the $\mathbf{u} \cdot \nabla \mathbf{u}$ term in the Navier–Stokes equations contributing the advective terms in eqns [1]–[3]). Numerical studies of such complex dynamics are now becoming possible with nested mesoscale, high-resolution DNS, and large-eddy simulation models. Such simulations promise to significantly advance our understanding of these various dynamics throughout the atmosphere. An example of the turbulence dynamics arising from a simple superposition of a single gravity wave (with $a = 0.5$ and $\omega_i = N/10$) and fine-structure mean shears is shown for illustration in the top panel of **Figure 7**. This streamwise-vertical cross section

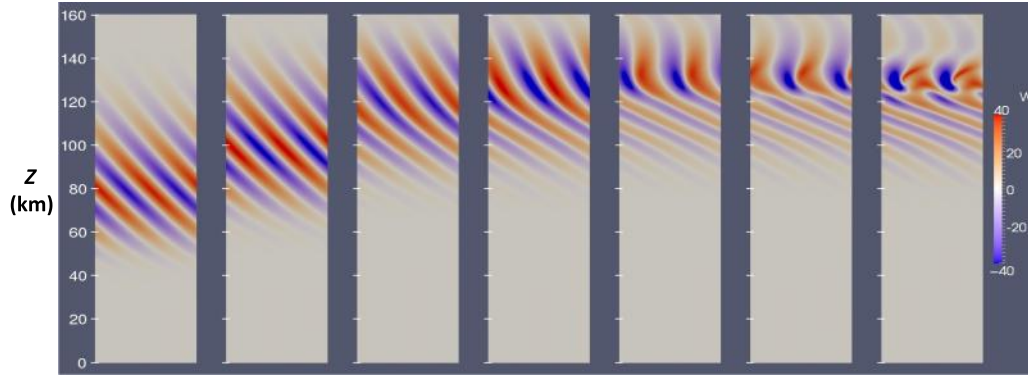


Figure 6 Streamwise-vertical cross sections of vertical velocity exhibiting gravity wave propagation and self-acceleration spanning ~ 5 wave periods. Maximum velocities increase from $\sim 2.5 \text{ m s}^{-1}$ initially to $\sim 60 \text{ m s}^{-1}$ in frame 5 and decrease as self-acceleration instability arises and the gravity wave dissipates. Fritts, D.C., Lund, T., 2011. Gravity wave influences in the thermosphere and ionosphere: Observations and recent modeling. In: Abdu, M., Pancheva, D. (Eds.), *Aeronomy of the Earth's Atmosphere and Ionosphere*. Springer, pp. 109–130.

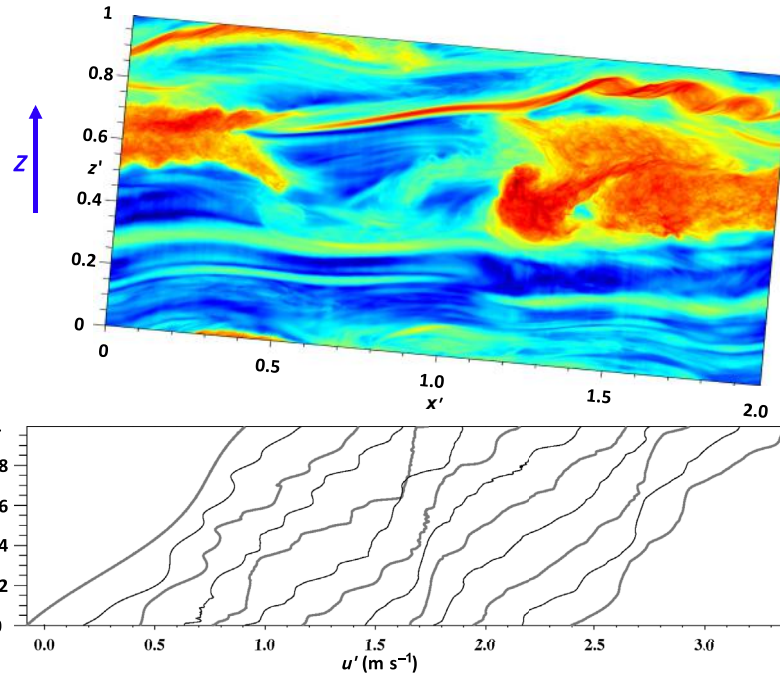


Figure 7 Gravity wave instabilities (top panel) arising due to a superposition of a gravity wave having $a = 0.5$ and $\omega_i = N/10$ and fine-structure (FS) mean shears having a larger vertical wave number ($m_{\text{FS}} \sim 5 m_{\text{GW}}$) and maximum initial shear $d\bar{u}_h/dz = 2N$. KHI and wave breaking are seen occurring simultaneously at different altitudes due to the evolving gravity wave-FS superposition. Vertical profiles of potential temperature from $t = 0 - 24 T_b$ at $2 T_b$ intervals at the center of the computational domain. The image at the top is at $t = 11.5 T_b$ just before the seventh profile.

of the energy dissipation rate, ε , reveals a highly layered response and simultaneous wave breaking and KHI at different altitudes. The lower panel shows vertical profiles of potential temperature at two T_b intervals throughout the DNS that reveal the creation of ‘sheet and layer’ fine structure that closely resembles these structures often seen in high-resolution measurements in the atmosphere and oceans. These results suggest that multiscale interactions may play prominent roles extending from the atmospheric boundary layer into the thermosphere.

Middle Atmosphere Effects

Without wave forcing of the mean circulation of the middle atmosphere, the mean motion at middle and high altitudes would be in geostrophic balance, mean winds would increase continuously with altitude, there would be no wave-driven residual circulation, i.e., $(\bar{v}^*, \bar{w}^*) = 0$, the thermal structure would be in radiative equilibrium, and the winter polar mesopause would be cold and the summer polar mesopause would be warm. In reality, however,

gravity wave dissipation and momentum flux divergence in the mesosphere lead to body forces that decelerate the zonal mean winds, reverse the vertical shears in the mesosphere, and induce a wave-driven residual circulation with rising motions in the summer hemisphere, subsidence in the winter hemisphere, and strong meridional flow from the summer to the winter hemisphere near the mesopause. Rising and descending motions lead to adiabatic cooling and heating of the summer and winter mesosphere, causing a reversal of the meridional gradient of temperature imposed by solar radiation at lower altitudes and the coldest temperatures in Earth's atmosphere at the polar summer mesopause. A similar, but weaker, residual circulation occurs at lower altitudes, to which planetary waves

also make a significant contribution. Gravity wave contributions are due largely to mountain waves in winter, with both mountain and nonstationary waves contributing in summer. The impact of wave forcing, denoted (D_{F_x}, D_{F_y}) in eqns [25] and [26], effects on the mean wind and temperature profiles, and the character of the residual circulation at lower and higher altitudes, are depicted schematically in Figure 8. (For further discussion of the quasi-biennial oscillation (QBO) and semiannual oscillation (SAO), see **General Circulation of the Atmosphere: Overview.**)

At equatorial latitudes, geostrophic motions do not occur, and gravity wave filtering, dissipation, and momentum flux divergence contribute instead to forcing of two equatorial oscillations of the zonal mean wind at different altitudes. For

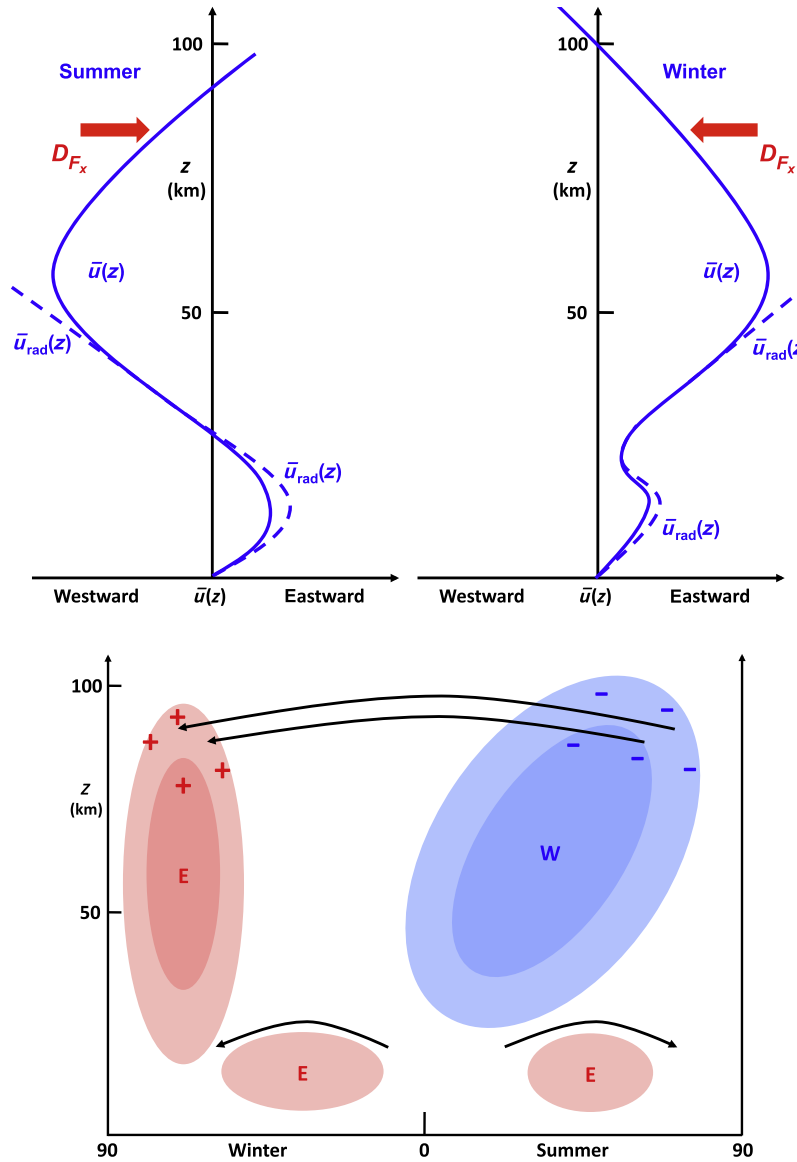


Figure 8 Schematic of winter and summer zonal wind profiles (top panel) and the wave-driven residual circulation (bottom panel). Solid (dashed) lines at the top show the observed (radiative equilibrium) profiles and red arrows indicate the mean body force accompanying gravity wave momentum deposition. Red (blue) regions at the bottom show eastward (westward) zonal jets, black arrows indicate the wave-driven circulation, and signs indicate heating and cooling due to induced vertical motions.

both the QBO and the SAO, gravity waves contribute to the forcing of each phase of the oscillation due to filtering and momentum deposition much as described above. In each case, gravity wave filtering in the lower phase of the oscillation imposes anisotropy on the wave spectrum that contributes to forcing of the opposite phase of the oscillation at greater altitudes. (For further discussion of the QBO and SAO, *see* **Middle Atmosphere**: Quasi-Biennial Oscillation; Semiannual Oscillation.)

Turbulence arising from gravity wave dissipation contributes both heating and transport through direct energy dissipation and mixing of local gradients of heat, momentum, and constituents. Turbulent mixing drives the thermal structure toward an adiabatic lapse rate ($\partial\bar{T}/\partial z = -g/C_p$), thus contributing to the mean thermal structure of the mesosphere. Because of the nature of gravity wave dissipation, however, mixing of heat and constituents is believed to be much less efficient than mixing of momentum. Turbulent mixing processes likely also play roles in gravity wave interactions with, and modulation of, larger scale tidal and planetary wave motions, given their significant responses to variations in local winds and static stability, such as so-called mesospheric inversion layers to which these large-scale dynamics frequently contribute. Relatively little is known about these dynamics at present.

See also: **Aviation Meteorology**: Clear Air Turbulence. **Dynamical Meteorology**: Kelvin–Helmholtz Instability;

Overview; Static Stability; Wave Mean-Flow Interaction.

General Circulation of the Atmosphere: Overview. **Gravity Waves**: Buoyancy and Buoyancy Waves: Optical Observations; Buoyancy and Buoyancy Waves: Theory. **Mesosphere**: Polar Summer Mesopause. **Middle Atmosphere**: Quasi-Biennial Oscillation; Semiannual Oscillation. **Turbulence and Mixing**: Overview; Turbulent Diffusion.

Further Reading

- Andrews, D.G., Holton, J.R., Leovy, C.B., 1987. *Middle Atmosphere Dynamics*. Academic Press, Orlando, FL.
- Bühler, O., 2009. *Waves and Mean Flows*. Cambridge University Press, Cambridge.
- Fritts, D.C., Alexander, M.J., 2003. Gravity wave dynamics and effects in the middle atmosphere. *Rev. Geophys.* 41, 1003. <http://dx.doi.org/10.1029/2001RG000106> also see corrigendum to gravity wave dynamics and effects in the middle atmosphere published in *Reviews of Geophysics*, 41, 1/1003, 2003(2012).
- Fritts, D.C., Lund, T.L., 2011. Gravity wave influences in the thermosphere and ionosphere: observations and recent modeling. In: Abdu, M., Pancheva, D. (Eds.), *Aeronomy of the Earth's Atmosphere and Ionosphere*. Springer, pp. 109–130.
- Gossard, E.E., Hooke, W.H., 1975. *Waves in the Atmosphere*. Elsevier, Amsterdam.
- Lighthill, J., 1993. *Waves in Fluids*. Cambridge University Press, Cambridge.
- Sutherland, B.R., 2010. *Internal Gravity Waves*. Cambridge University Press, Cambridge.

Buoyancy and Buoyancy Waves: Optical Observations

MJ Taylor and WR Pendleton, Jr., Utah State University, Logan, UT, USA

© 2003 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 1, pp 323–330, © 2003, Elsevier Ltd.

Introduction

During the past four decades, a wide variety of remote and *in situ* observational techniques have revealed a rich spectrum of wave activity in the upper atmosphere, with horizontal wavelengths ranging from a few kilometers (km) to global scales and temporal periodicities ranging from a few minutes to about 2 weeks. In terms of decreasing spatial-temporal scales, the observed spectrum of freely propagating atmospheric waves is subdivided into planetary, inertia-gravity, and buoyancy (or acoustic-gravity) waves. Although physically similar to inertia-gravity waves, the buoyancy waves are of sufficiently small temporal and spatial scales that the rotation and curvature of the Earth play minor roles in determining their properties. Optical observations of these mesoscale (~ 10 – 1000 km horizontal wavelengths) waves are the primary focus of this article.

A complete spatial-temporal description of the global distribution of buoyancy wave sources has yet to be made, but a growing body of evidence indicates copious sources in the lower atmosphere. This is not particularly surprising since any perturbation of a stably stratified, low-dissipation region with frequency components between the high (Brunt-Väisälä) and low (Coriolis) frequency limits will generate buoyancy waves. The propagation of such waves in a realistic atmosphere will be strongly impacted by the background winds and temperature structure, producing such effects as refraction, reflection (complete or partial), and ducting (partial or complete). Hence, propagation of tropospherically generated buoyancy waves to the mesosphere and lower thermosphere (MLT) region (altitude range ~ 80 – 100 km) is not ensured, but the mesoscale waves observed in this region frequently exhibit downward phase progression, which is indicative of upward transport of energy from lower-atmospheric sources. In the upper mesosphere and lower thermosphere, processes involving viscosity and thermal conduction attenuate and dissipate energy for waves in the saturation region of growth. In [Figure 1](#), a simplified sketch captures much of the basic physics embodied in the modern concept of a realistic atmosphere permeated by packets of buoyancy waves.

It is now widely accepted that buoyancy waves with periods less than ~ 1 h achieve major importance in the MLT region because of the relatively large quantities of momentum (and to a lesser extent energy) that they transport to this region. In this connection, growth of many of the tropospherically generated buoyancy waves between their sources and the mesopause region is expected to increase their energy (per unit mass) by a factor of ~ 100 . A number of comprehensive experimental and theoretical evaluations of gravity-wave forcing have resulted in the conclusion that this process has a profound impact on the large-scale circulation in the MLT, as well as on the thermal and minor-constituent structures of the region. In addition, such forcing is expected to impact variability in the

MLT on many spatial and temporal scales as a result of various types of interactions such as wave-tide and wave-mean flow.

Optical Observations

Optical observations of buoyancy waves in the terrestrial upper atmosphere have been greatly facilitated by ‘nature’s gifts’ of several vertically distinct airglow layers, alkali metal layers, and noctilucent clouds (NLC). Some of the earliest evidence for atmospheric buoyancy waves was provided by wavelike deformations frequently observed during NLC displays. These are very tenuous mesospheric ice clouds (altitude ~ 82 km) that form during the summer months at high latitudes (typically $>55^\circ$) when the mesopause cools to the lowest temperature on Earth (110 – 150 K). As the clouds are tenuous, they can only be seen from the ground by the scattering of sunlight during the hours of twilight, when the observer and the atmosphere below the cloud layer are in darkness while the clouds themselves remain illuminated. (This condition occurs for solar depression angles between $\sim 6^\circ$ and 16° .) Optimum locations for observing mesospheric clouds now and over the past 100 years are Scandinavia/northern Europe, central Asia, and Canada in the Northern Hemisphere and the southern tip of South America in the Southern Hemisphere, where they are readily seen during the prolonged hours of twilight. An excellent example illustrating the wave forms frequently observed

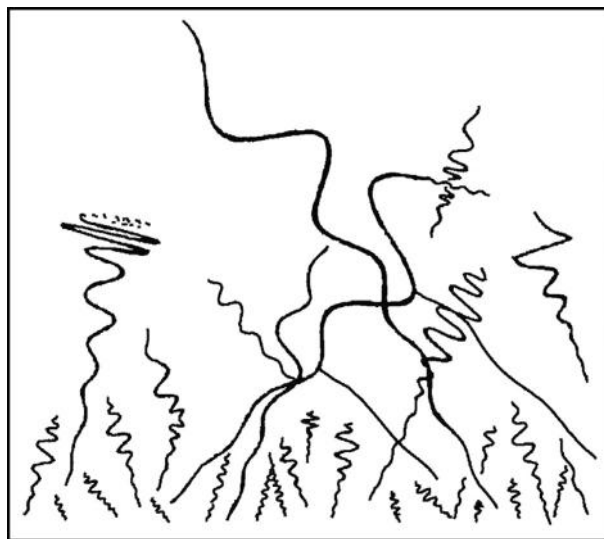


Figure 1 Surrealistic representation of several key physical processes associated with the propagation of internal atmospheric gravity waves. Hines, C.O., et al., *The Upper Atmosphere in Motion*, Geophysical Monograph 18, American Geophysical Union, Part 2, Atmospheric Gravity Waves, 194, Figure 1.



Figure 2 High-latitude summer twilight photograph of noctilucent clouds at ~ 82 km altitude showing band-type buoyancy waves and many smaller-scale billows. Copyright of P. Parviainen, Finland, e-mail: pekka.parviainen@ppl.inet.fi.

in NLC is shown in **Figure 2**. The photograph clearly shows three large-scale NLC 'bands' with many smaller-scale, near-perpendicular, 'billow' waves. The average horizontal wavelengths associated with such deformations lie in the range 5–10 km (billows) and ~ 20 –50 km (bands), the latter of which are typical of the lower buoyancy range. It is interesting to note that current theories for the formation of NLC draw heavily on gravity-wave forcing from below to provide the requisite extremely cold near-mesopause environment during the summer months for successful ice nucleation and growth.

It is convenient to distinguish between *active* and *passive* optical methods for studying buoyancy waves. The active methods are usually associated with Rayleigh and resonant-scatter lidars that probe the vertical structure of temperature and/or density fields by studying the return signal from short, very intense, gated pulses of monochromatic light scattered by the upper-atmospheric medium. In contrast, most passive methods (with the exception of NLC) utilize the wave-induced spatial-temporal modulation of the airglow emissions to detect and study certain characteristics of the waves. The passive optical methods include various types of narrowband imaging, photometry/radiometry, and spectrometry/interferometry. Clearly, the lidar techniques yield direct information on vertical structure and motion, including the sense of vertical phase progression. In contrast, many of the passive techniques yield direct information on horizontal structure and motion and limited direct information on vertical structure when two or more layer observations are employed.

Lidar Measurements

With the development of powerful lidar systems in the 1980s, a new era of active remote sensing of the middle and upper atmosphere was born. Studies of buoyancy wave dynamics using lidars have emphasized Rayleigh scatter methods for investigating structure and wave motions in the stratosphere and lower mesosphere (~ 25 –70 km) and resonant scatter (e.g. using sodium) for studying waves in the MLT region (~ 80 –105 km). The low-altitude limit (~ 25 km) for the

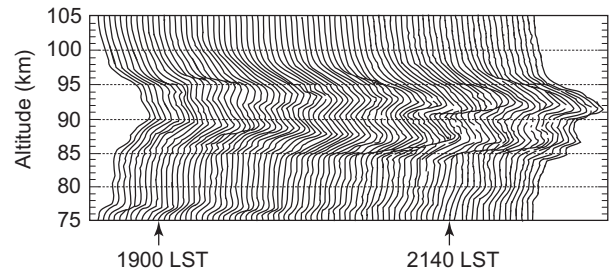


Figure 3 Sequence of Na lidar profiles (3-min spacing) illustrating the downward phase progression of a coherent gravity wave of 6.9-km vertical wavelength and observed period of ~ 5.8 -hr. Collins R.L., et al., 1996. Gravity wave activity in the upper mesosphere over Urbana, Illinois: lidar observations and analysis of gravity wave propagation models. *Journal of Atmospheric Terrestrial-Solar Physics* 58, 1905–1926, Figure 2.

stratospheric measurements is determined principally by signal distortion associated with Mie scattering from aerosols and particulates, whereas the upper-altitude limit (~ 70 km) results from limitations imposed by signal photon noise. In practice, each limit is determined by the characteristics of the particular lidar system (primarily its power-aperture product) and by the information retrieval methods used in the data analysis.

Lidar studies of wave dynamics in the mesopause region have utilized the alkali metal layers (e.g., Na and K) that are created primarily by the ablation of meteors in this region of the atmosphere. As Na is much more abundant and has a large back-scatter cross-section, it is the most well suited for buoyancy wave studies. In all cases, the lower- and upper-altitude limits of the measurements are determined by signal photon noise and/or information retrieval uncertainties. It is usually assumed that the Na atoms act as a passive tracer of wave dynamics under typical measurement conditions. However, extensive modeling studies of the chemistry of the Na layer and its response to wave forcing suggest that this basic assumption may be invalid for altitudes < 85 km. In addition, this assumption can be compromised under the special conditions that apply when sporadic layers form within the Na layer. This said, lidar studies represent one of the most powerful and advanced tools for sounding the atmosphere.

Current lidar studies of buoyancy waves provide a direct measure of vertical structure in Na number density induced by well-developed quasi-monochromatic events. Vertical wavelengths in the measurement range, typically 1 to ~ 20 km (limited by layer thickness), in the mesopause region and ~ 3 –30 km in the stratopause region are common. Quasi-monochromatic waves with observed periods in the range of several minutes (close to the local Brunt-Väisälä value) to several hours (usually limited by the data-record length) have been measured. However, in practice lidars have proven to be most sensitive to a class of waves with short vertical wavelengths (< 10 km) which exhibit relatively long observed periods and hence slow vertical phase progression. These waves tend to follow the so-called diffuse-damping limit for wave growth. **Figure 3** shows a sequence of Na layer profiles (~ 3 min spacing), tracing the downward progression of a well-developed long-period (5.8 h) buoyancy wave observed in the Na layer. This typical observation yielded a mean vertical

Table 1 Layer characteristics (full-width at half-maximum), except the Na layer which is of full width

Emission/layer	Peak altitude (km)	Width (km)
OI (557.7 nm)	96	6–10
O ₂ (0,1)	94	8–10
Na (589.2 nm)	90	~10
NIR OH	87	8–10
Na layer	92	80–105
NLC	82	2

wavelength of 6.9 km and a vertical phase velocity of -0.33 m s^{-1} , where the minus sign signifies downward phase progression, as illustrated in the figure. An inferred horizontal wavelength of about 450 km follows from these directly measured parameters. More sophisticated wind/temperature lidar systems have recently been developed that are capable of measuring temperature perturbations induced by long-period waves and tides (which are forced rather than freely propagating waves) and studying the background wind field through which the waves progress using thermal broadening and Doppler shifting of the resonance line. Together with the wave measurements these data can be used to estimate the sensible heat flux and vertical flux of horizontal momentum transported into the MLT region.

Image Measurements

Over the past 25 years the capability of imaging instrumentation for remote sensing faint structures in the upper-atmospheric nightglow emissions has evolved considerably. Early photographic observations demonstrating the existence of wavelike motions (akin to those seen in NLC) have been superseded by low-light, intensified TV cameras and more recently by solid state CCD imaging systems that now provide an exceptional capacity for quantitative studies of mesospheric wave motions.

Images of the naturally occurring nightglow emissions afford an excellent method for investigating the horizontal morphology and dynamics of short-period (typically $<1 \text{ h}$) buoyancy waves. To date, most imaging studies have utilized the bright near infrared (NIR) hydroxyl (OH) Meinel band emissions that originate from a well-defined layer (typical half-width $\sim 8 \text{ km}$) centered at $\sim 87 \text{ km}$. However, there are a growing number of observations of the NIR O₂ (0,1) atmospheric band emission at $\sim 865 \text{ nm}$ and the visible wavelength OI (557.7 nm) and Na (589.2 nm) line emissions. Although these emissions are considerably weaker than the broadband OH emission, they each exhibit well-defined nighttime profiles at different, but closely spaced altitudes in the MLT region. **Table 1** lists the properties of the nightglow emission layers together with the metal layers that are most frequently used in resonant lidar studies and the NLC layer characteristics. As the waves propagate and dissipate within the MLT region they induce significant modulations in the line-of-sight brightness (and rotational temperature) of these emission layers which is detected as 'structure'. Measurements of two (or more) nightglow emissions therefore provide an important additional

method for investigating the vertical as well as horizontal propagation of short-period buoyancy waves.

In particular, all-sky image data (180° field of view) yield unique information on the occurrence frequency, two-dimensional horizontal spatial characteristics, and the prevailing directionality of small-scale waves over an exceptionally large geographic area ($\sim 400 \text{ km}$ radius) corresponding to $\sim 500\,000 \text{ km}^2$ at MLT heights with high temporal and spatial resolution. These cameras are most sensitive to relatively fast-moving waves exhibiting vertical wavelengths somewhat greater than the layer thickness (i.e., $> 8 \text{ km}$) and horizontal wavelengths (λ_h) $\sim 5\text{--}200 \text{ km}$ (i.e., significantly less than the maximum field of view). In general, lidar and image measurements therefore sample different (but overlapping) regions of the buoyancy wave spectrum.

Optical measurements of the airglow emissions can be made at any latitude and season providing a global, all year round capability. Such studies, once limited to the realm of NLC, have revealed a wealth of small-scale wave activity at equatorial, mid- and high-latitudes from many sites around the world, and it is not uncommon to observe several different wave patterns during the course of a night suggesting copious sources. **Figure 4** illustrates a variety of wave patterns that are most commonly observed in the MLT emissions. These observations have mainly been made from mid- and low-latitude mountain sites, and distinct spatial and temporal properties have emerged which suggest the existence of two dominant types of short-period waves, termed 'bands' and 'ripples' (akin to those seen in NLC). **Figure 4(a)** illustrates the most prominent quasi-monochromatic pattern which usually appears as an extensive, coherent series of waves that exhibit horizontal wavelengths of a few to several tens of kilometers and horizontal phase speeds up to 100 m s^{-1} . Band displays are persistent, usually lasting for a few to several hours, and are spatially extensive, often occupying an area of sky much larger than the instantaneous all-sky field of view. In this image the bands appear as a series of curved waves (like the segments of an orange) due to the format of the all-sky lens. However, when they are mapped into geographic coordinates they often (but certainly not always) appear as an extensive series of quasi-planar waves. In this example the wave pattern appeared linear and exhibited a horizontal wavelength of $38 \pm 2 \text{ km}$, and an observed phase speed of $34 \pm 3 \text{ m s}^{-1}$ indicating an observed period of $19 \pm 2 \text{ min}$. The histogram plots of **Figure 5** illustrate the range of values typically observed for bands.

In contrast, the second type of wave motion (termed ripples) is quite distinct from the bands, exhibiting much smaller spatial and temporal scales. Ripples usually occur in localized wave packets that occupy much smaller regions of the sky, typically $< 5 \times 10^3 \text{ km}^2$. They have relatively short lifetimes (a few minutes to $\sim 45 \text{ min}$) and almost always exhibit periods close to the local Brunt-Väisälä period ($\sim 5 \text{ min}$). Two ripple patterns are also evident in **Figure 4(a)** superimposed on the band-type wave pattern. The similarity between the wave forms seen in this image and the narrow angle NLC of **Figure 2** is striking.

The importance of multilayer measurements is illustrated in **Figure 4(b)**, which shows wave structure imaged in the Na emission layer at approximately the same time as **4(a)**. (Note that the bright lines are due to an Na lidar beam probing two

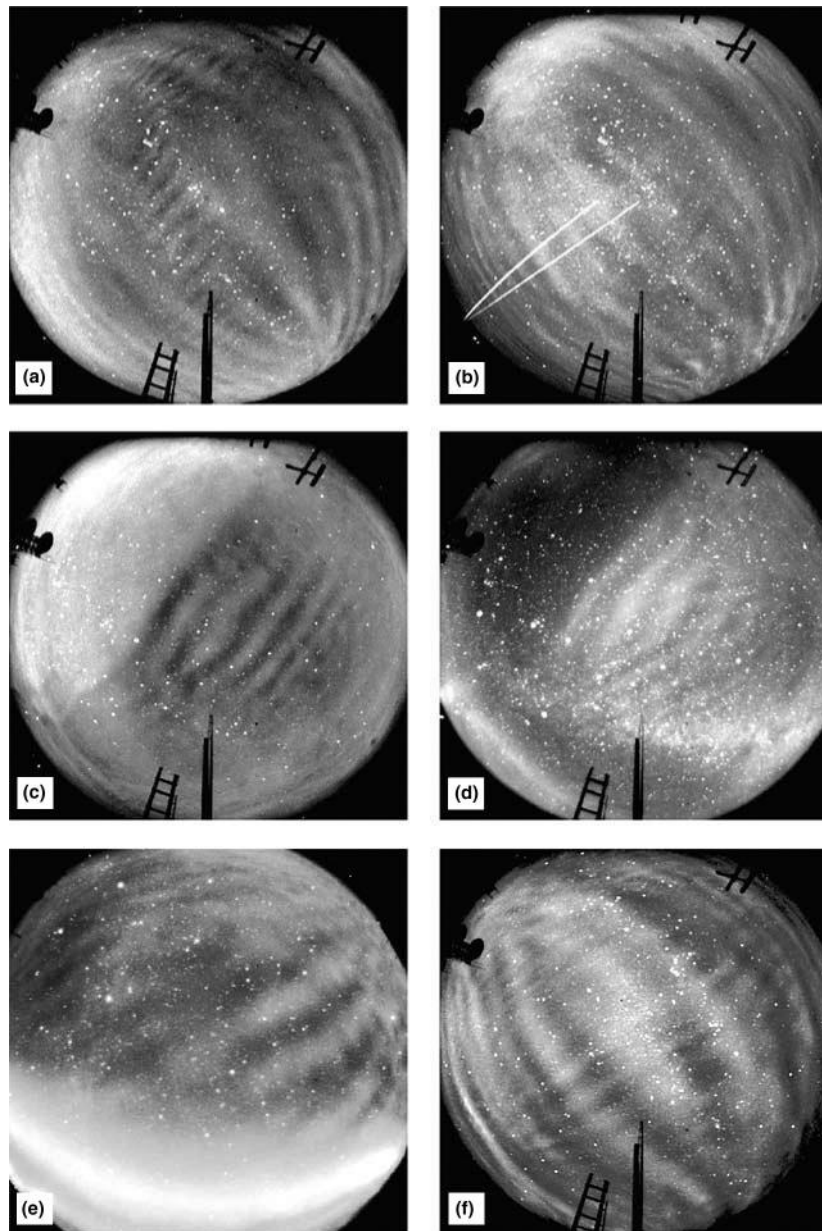


Figure 4 Example all-sky (180°) images illustrating the dominant spatial signatures of short-period buoyancy waves observed frequently in the MLT nightglow emissions. Adapted from Taylor, M.J., et al., 1995. All-sky measurements of short-period waves imaged in the OI(557.7 nm), Na(589.2 nm) and the near infrared OH and O₂(0,1) nightglow emissions during the ALOHA-93 campaign. *Geophysical Research Letters* 22, 2833–2836, Figure 1.

regions of sky during the 120 s image exposure.) The same band pattern is clearly evident in this emission (and in the OH and O₂ emissions) indicating that this wave motion was coherent and extended vertically throughout the MLT region whereas the ripples are absent revealing their limited horizontal and vertical extent.

A likely source of ripples and billows is the chance combination of wind and long-period wave motions (including tides) creating localized regions of strong wind shear which then generate small-scale waves *in situ* through the Kelvin–Helmholtz instability. Alternatively, the waves may be generated by a three-dimensional convective-type instability which

predicts that the waves should form near-orthogonal to the perturbing wave. In each case, the wave patterns will be short lived and spatially localized as evident from the image data. In contrast, the band-type waves have been shown to be due to freely propagating or ducted buoyancy waves most probably of tropospheric origin.

Figure 4(c), (d) depicts different class of band-type motion termed a ‘frontal event’. Unlike most band patterns this type of wave is characterized by a sharp leading edge followed by a discrete number of (typically <10) trailing wave crests, similar in morphology to a bore on a river. This type of wave motion is much less common and is thought to be the

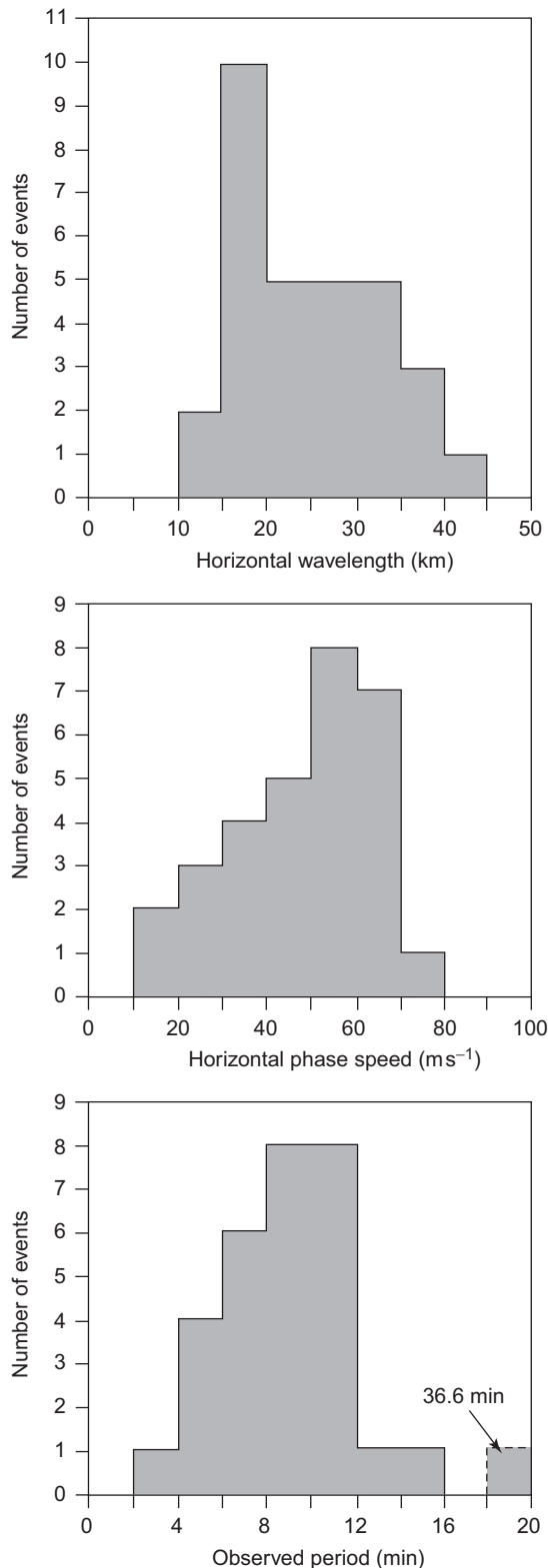


Figure 5 Histogram plots illustrating the typical ranges of horizontal wavelength, phase speed and observed period associated with short-period gravity wave patterns. Adapted from Taylor, M.J. et al., 1997. Image measurements of short-period gravity waves at equatorial latitudes. *Journal of Geophysical Research* 102, 26283–26299, Figures 4, 5 and 6.

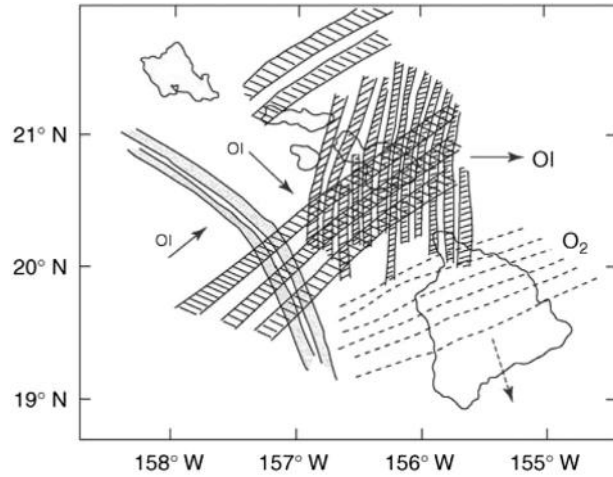


Figure 6 Map illustrating a relatively complex wave display imaged in the OI(557.7 nm) and O₂(0,1) nightglow emissions and consisting of two quasi-linear wave patterns and one curved wave train. Taylor, M.J., et al., 1995. All-sky measurements of short-period waves imaged in the OI(557.7 nm), Na(589.2 nm) and the near infrared OH and O₂(0,1) nightglow emissions during the ALOHA-93 campaign. *Geophysical Research Letters* 22, 2833–2836, Figure 2.

signature of a trapped (or ducted) wave propagating near-horizontally at mesospheric heights. In this case, the effect of the bore intrusion on the airglow layers was to simultaneously lower the height of the NIR OH (image d) and raise the OI (557.7 nm) layer (image c) which is manifested as a reversal in contrast to the structuring.

Figure 4(e) shows another example of band-type wave structure ($\lambda_h = 35.5 \pm 1.0$ km) but this time imaged in the high-latitude ($\sim 65^\circ$ N) OH emission from central Alaska. The bright arc in the lower part of the image is due to auroral precipitation. Joule heating and other forcing associated with solar-induced magnetic storms are a known source of large-scale buoyancy waves that are often detected in the high-altitude ionosphere as traveling ionospheric disturbances (TIDs). However, in this case the bands were moving towards the auroral zone suggesting other, tropospheric-type sources.

These examples have been chosen for their clarity to illustrate the types of short-period wave motions that exist at MLT heights. However, as one would expect from the quasi-random ensemble of sources and waves propagating into the upper atmosphere (as depicted in the sketch of Figure 1) there are oftentimes many more waves evident, resulting in a complex, time-varying airglow pattern. Such a situation is shown in the OI (557.7 nm) image of Figure 4(f). In this case two bandtype motions progressing in almost orthogonal directions are evident. A map illustrating the geographical location and orientations of an even more complex wave display observed over the Hawaiian Islands is shown in Figure 6. The display consists of two quasilinear wave patterns and one curved wave train. The temporal evolution of these events reveals the true/exceptionally dynamic nature of the MLT region.

Figure 7(a), (b) shows two narrow angle images depicting exciting new evidence of wave breakdown creating turbulent structures. The two images show a portion of a well-developed

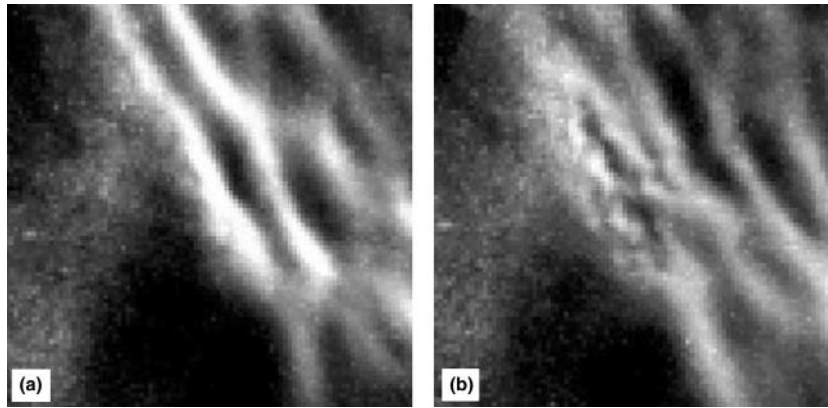


Figure 7 (a, b) Narrow-angle images illustrating evidence of wave breaking in the OH emission leading to turbulent structures. Adapted from Yamada, Y., et al., 2001. Breaking of small-scale gravity wave and transition to turbulence observed in OH airglow, *Geophysical Research Letters* 28, 2153–2156, Figure 1.

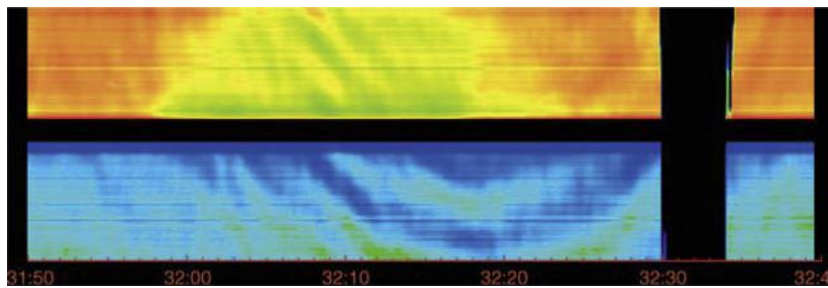


Figure 8 MSX satellite observations of upper stratospheric CO₂ emissions illustrating a near circular buoyancy wave pattern likely produced by a ‘point-like’ thunderstorm located at its center of curvature. Dewan, E.M., et al., 1998. MSX satellite observations of thunderstorm-generated gravity waves in mid-wave infrared images of upper stratosphere, *Geophysical Research Letters* 25, 2809–2812, Figure 2(b).

band pattern ($\lambda_h = 27$ km) which appears to increase in contrast, as the wave becomes nonlinear (image a) and then breaks into a number of much smaller scale turbulent features (image b). Such observations are currently rare, yet it is expected that the breakdown of these waves to smaller scale sizes eventually resulting in turbulence and the associated transfer of energy and momentum into the background medium is a major driver of the MLT region dynamics.

Ground-based, airborne, and, most recently, satellite-borne image measurements have been employed to study buoyancy waves. The advantage of satellite-based measurements is their ability to study waves and their potential sources over remote, inaccessible areas on a global scale. This new capability is illustrated in Figure 8, which shows data from the Midcourse Space Experiment (MSX) providing the first observations from space of gravity waves generated by a thunderstorm over Indonesia. In this case, the data are nadir pointing from the midwave infrared CO₂ emission at 4.3 μ m in the upper stratosphere (altitude 40 km) and show circular wave fronts ($\lambda_h = 25$ km) launched by a ‘pointlike’ thunderstorm at their center. Until recently most satellite measurements have focused on much larger horizontal scale size ($\lambda_h \sim$ a few thousand kilometers), longer period waves that can be readily discerned in limb airglow measurements. However, this new nadir capability is poised to revolutionize our understanding of wave generation and propagation in the middle atmosphere on a global scale.

Interferometry, Photometry, and Spectrometry Measurements

Interferometers, photometers (radiometers), and spectrometers continue to play major roles in quantifying the line-of-sight-integrated intensity and temperature responses of the airglow

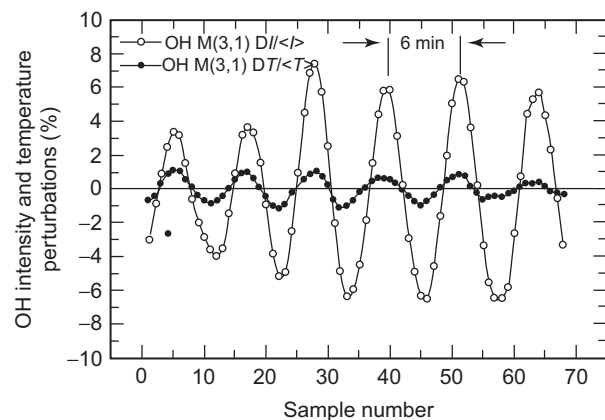


Figure 9 Ground-based measurements of OH Meinel (3,1) band emission illustrating the intensity and rotational temperature signatures induced by the passage of a short-period (~ 6 -mm) buoyancy wave. Courtesy Utah State University.

layers to buoyancy waves. For quasi-monochromatic events, the ‘layer-averaged’ amplitude of the perturbing wave can, in principle, be obtained directly from the observed temperature variation. However, in practice, an accurate determination of the wave amplitude requires a knowledge of the background temperature structure, the distribution of emitters in the layer, and the vertical wavelength of the wave. This information set is seldom, if ever, available. Furthermore, the column emission rate (CER) response is less directly related to the amplitude of the induced perturbation. As a result, ratio of the fractional perturbations in CER and temperature (which is independent of wave amplitude) has become the key parameter in comparing experiment with theory.

In their simplest operating mode (static, single field), these instruments continue to provide much useful information for testing the predictions of sophisticated chemical–dynamical models for the layer response. For example, [Figure 9](#) shows the response of the OH Meinel airglow layer to short-period ($\cong 5.8$ min) buoyancy waves near the Brunt–Väisälä limit as measured by a high-throughput Michelson interferometer. A train of six oscillations clearly illustrates that the high-frequency wave-induced fractional temperature perturbations are nearly an order of magnitude smaller than the fractional CER perturbations. As the induced perturbation amplitudes

for such highfrequency waves are expected to be small (typically $\leq 1\%$), precise measurements pose a significant experimental challenge.

In their more sophisticated multifield or imaging modes, these instruments have provided useful information on the horizontal spatial scales and motions of the waves, in addition to defining the integrated layer responses to perturbations induced by the waves. Furthermore, as with imaging studies, measurements using two (or more) vertically distinct airglow emissions (e.g., OH and O₂ (0,1) bands) provides a means for assessing horizontal and vertical scales of motion and the amplitude of wave growth.

See also: **Chemistry of the Atmosphere:** Observations for Chemistry (Remote Sensing): Lidar. **Clouds and Fog:** Noctilucent Clouds. **Gravity Waves:** Buoyancy and Buoyancy Waves: Theory; Overview. **Lidar:** Atmospheric Sounding Introduction. **Optics, Atmospheric:** Airglow Instrumentation; Optical Remote Sensing Instruments.

Buoyancy and Buoyancy Waves: Theory

TJ Dunkerton, Northwest Research Associates, Bellevue, WA, USA

© 2003 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 1, pp 330–334, © 2003, Elsevier Ltd.

Introduction

Under the influence of gravity, a fluid of variable density settles into a configuration in which surfaces of constant density are approximately horizontal and density decreases with height. Such a fluid is said to be stably stratified. Stratification is an important property of geophysical flows, affecting motions over a wide range of spatial and temporal scales. The buoyancy of a fluid parcel (or of a solid object) suspended within a fluid describes the tendency of the parcel or object to move upward or downward in response to gravitational and pressure-gradient forces. Buoyancy might be regarded as an effective ‘force’, but it actually represents a combination of these two fundamental forces. Coherent oscillations resulting from the restoring force of buoyancy are known as buoyancy waves, or gravity waves; these waves play an important role in atmospheric dynamics.

Buoyancy

The vertical acceleration of a parcel of fluid is governed by

$$\frac{dw}{dt} + g + \frac{1}{\rho} \frac{\partial p}{\partial z} + \cdots = 0 \quad [1]$$

where w is the vertical component of velocity, d/dt denotes the material derivative following the parcel, g is the acceleration of gravity, ρ is density and p is pressure. Linearizing the second and third terms of eqn [1] about a basic state in hydrostatic balance, such that

$$p = \bar{p} + p' \quad [2a]$$

$$\rho = \bar{\rho} + \rho' \quad [2b]$$

and

$$\frac{\partial \bar{p}}{\partial z} = -\bar{\rho}g \quad [3]$$

the equation may be written

$$\frac{dw}{dt} + g \frac{\rho'}{\bar{\rho}} + \frac{1}{\bar{\rho}} \frac{\partial p'}{\partial z} + \cdots = 0 \quad [4]$$

The second term describes the buoyancy of a parcel:

$$b' \equiv -g \frac{\rho'}{\bar{\rho}} \quad [5]$$

implying a tendency for lighter parcels to accelerate upward, and vice versa, under the influence of gravity. The buoyancy variable in eqn [4] acts as an effective ‘force’ in the vertical direction.

In a nondiffusive incompressible fluid, density is conserved following the motion:

$$\frac{d\rho}{dt} = 0 \quad [6]$$

which may be linearized about a state of rest and combined with the definition of buoyancy variable to give

$$\frac{\partial b'}{\partial t} + w'N^2 = 0 \quad [7]$$

where

$$N^2 \equiv -\frac{g}{\bar{\rho}} \frac{\partial \bar{\rho}}{\partial z} \quad [8]$$

is the Brunt–Väisälä frequency squared, a quantitative measure of stratification. N^2 is positive in a stably stratified fluid. In this case, N is the frequency of oscillation of a fluid parcel about its equilibrium level of neutral buoyancy, as can be seen, for example, by combining eqn [7] with the fully linearized version of eqn [4] and neglecting the vertical pressure gradient force.

In an ideal gas, potential temperature θ is conserved following the motion (apart from diabatic effects) and may be used as a surrogate for density in the definition of buoyancy. This approximation is valid if the speed of sound is large and the effect of the pressure perturbation on density is thereby minimized. The corresponding definition of N^2 is

$$N^2 \equiv \frac{g}{\bar{\theta}} \frac{\partial \bar{\theta}}{\partial z} \quad [9]$$

so that potential temperature increases with height in a stably stratified atmosphere. In this case, once again, N is the frequency of oscillation of an air parcel about its equilibrium level of neutral buoyancy.

Unlike solid objects that are shape-preserving, fluid parcels are stretched and deformed by the motion; buoyant plumes entrain or detrain fluid while mixing with the environment. The concept of buoyancy, as applied to individual parcels, is therefore of limited use except in special situations where the integrity of parcels is maintained. Moreover, individual parcels are rarely able to move about without affecting neighboring ones. Buoyancy remains important, nonetheless, because it influences the collective motion of parcels coupled together by the fundamental forces. Stratification affects virtually all classes of atmospheric motion. It is particularly important for buoyancy waves – more commonly known as gravity waves – which arise in a stably stratified flow as a result of the restoring force of buoyancy.

Buoyancy Waves

Gravity waves exist in two forms: internal and external. The most common type of external wave propagates horizontally along the interface between two fluids of different density, with amplitude diminishing away from the interface. Internal waves, on the other hand, propagate horizontally and vertically in a fluid of continuously varying density or potential

temperature. Internal waves are more important in the atmosphere. While propagating vertically, gravity waves transport horizontal momentum vertically, often over many density scale heights. To conserve wave energy, the amplitude of a gravity wave generally grows with altitude because of the density decrease, so that a small perturbation (perhaps undetectable in the lower atmosphere) becomes large in the middle and upper atmosphere. Momentum is irreversibly deposited in a region where the waves attain large amplitude and become locally unstable (e.g., at a critical level (see later)), breaking via convective or shear instability. Vertical mixing of heat and constituents also can be attributed to gravity-wave breaking.

Gravity waves assume the form of plane waves, consisting of alternating slabs of fluid sliding upward or downward at an angle, as illustrated in Figure 1. They also appear with surfaces of constant phase radiating diagonally from a two-dimensional source (e.g., a vibrating horizontal rod) or as conical phase surfaces in three dimensions emanating from a point source. Waves originating from a point source appear as concentric rings in a horizontal cross-section, and have been observed in the stratosphere above isolated thunderstorms.

Figure 1 depicts a plane wave with phase propagation downward and to the right. For this wave the group propagation is upward and to the right, orthogonal to the direction of phase propagation as is generally the case with gravity waves. A downward propagating phase is commonly observed in the middle atmosphere, suggesting upward group propagation and a tropospheric source. An upward propagating phase is sometimes observed in the troposphere, suggesting downward group propagation and an upper tropospheric source. In other cases, upward group propagation is observed throughout the

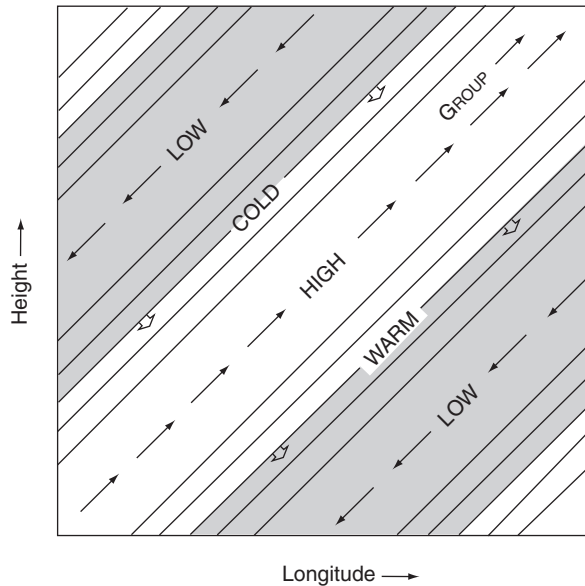


Figure 1 Structure of an internal gravity wave with phase propagation downward and to the right. Contours indicate zonal velocity, vertical velocity, or pressure, with negative values shaded. Rising and sinking motions produce adiabatic cooling and warming, respectively. The resulting temperature anomalies are in quadrature with the pressure anomalies, leading the pressure anomalies by a quarter-cycle.

troposphere, indicating a source near the surface. Vertical transport of horizontal momentum is in the direction of vertical group propagation, as implied by the positive correlation of horizontal and vertical velocity components in Figure 1.

In a simple mean flow without shear, or slowly varying in time and space, the intrinsic frequency of a vertically propagating gravity wave must lie in the range

$$|f| < |\hat{\omega}| < N \quad [10]$$

where f is the local Coriolis frequency. The intrinsic frequency is

$$\hat{\omega} = \omega - \mathbf{k} \cdot \bar{\mathbf{u}} \quad [11]$$

where ω is the wave frequency relative to the coordinate system, $\mathbf{k}$ is the wavevector, and $\bar{\mathbf{u}}$ is the mean flow, assumed horizontal. The special situation $\hat{\omega} = 0$ (relevant if $f = 0$) is illustrated in Figure 2, and is referred to as the Hines critical circle, named after its inventor Colin Hines, an atmospheric scientist of the latter half of the 20th century who has written extensively on the role of gravity waves in the middle atmosphere. This circle traces out a locus of points for which the component of mean flow velocity in the direction of horizontal wave propagation equals the horizontal phase speed of the wave, $c = \omega/k$, where k is the horizontal wavenumber. Equivalently, for any phase velocity vector extending from the origin to a point on the circle, the mean flow speed $|\bar{\mathbf{u}}|$ matches the apparent speed of wave crests in the direction of the mean flow.

When the wavevector and mean-flow vector are parallel, the wave is said to have a critical level at the altitude where mean-flow speed and horizontal phase speed are equal. Waves with

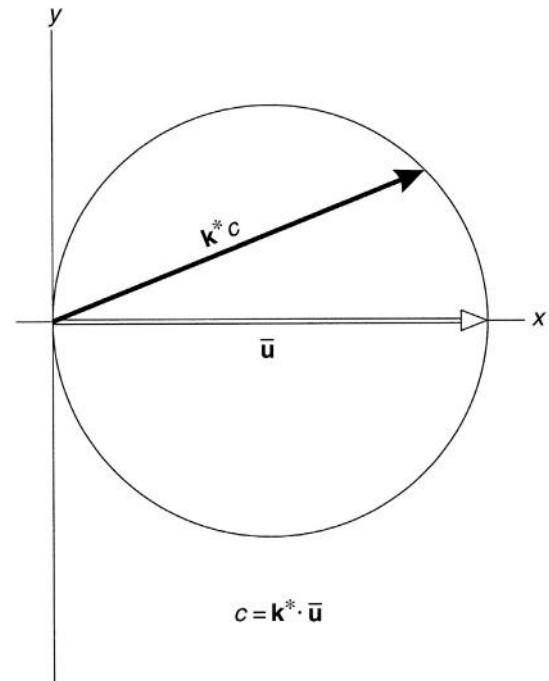


Figure 2 Hines's critical circle for plane waves with horizontal component of wavevector, $\mathbf{k}^*$, normalized to unit length in a mean flow $\bar{\mathbf{u}}$, assumed horizontal. Along the circle, the mean flow speed projected onto the direction of wave propagation is equal to the wave's horizontal phase speed, c .

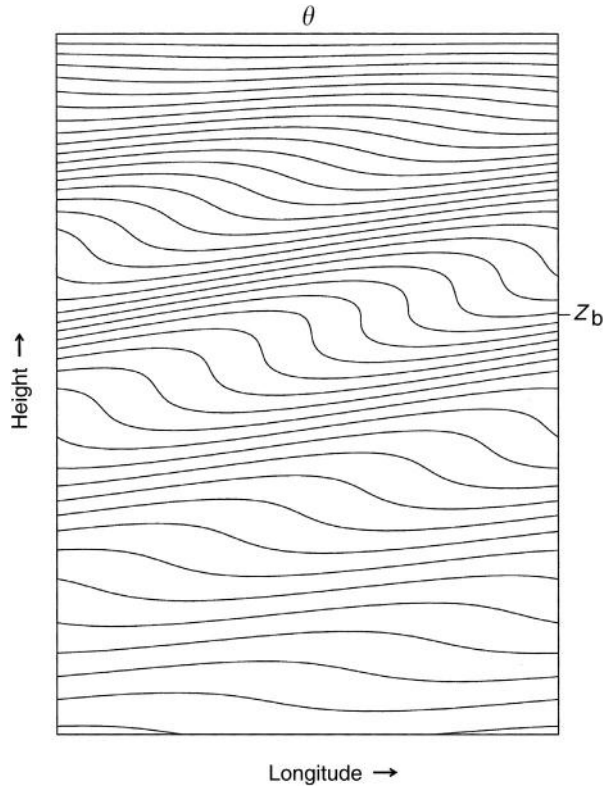


Figure 3 Potential temperature surfaces displaced by an upward propagating internal gravity-wave packet experiencing exponential growth with height in a compressible atmosphere. The wave is about to overturn and break in the upper part of the domain, beginning at the breaking level, z_b .

upward group propagation, approaching a critical level from below, grow to large amplitude, overturn, and break in a critical layer of finite depth lying below the critical level. In a compressible atmosphere where density decreases with height, waves generally grow in amplitude with height, with or without shear, and eventually break. This situation is illustrated in [Figure 3](#), showing potential temperature surfaces distorted to such an extent that some of these surfaces are about to overturn and create conditions for local static instability. Parcels within a convectively unstable region experience exponential growth of displacement amplitude with time, quickly restoring a locally stable configuration. The process of wave breaking leads to turbulence generation, momentum deposition, mixing, and diminution of wave amplitude, and in some cases, excitation of waves not present in the original spectrum.

Waves with intrinsic frequency near f are known as inertia-gravity waves. The effect of positive f is illustrated in [Figure 4](#). A vertical profile of horizontal velocity would look something like the hodograph in [Figure 5](#), with phase increasing downward ($m < 0$). Horizontal wind components obtained from a balloon ascending rapidly through this wave would exhibit clockwise rotation with height, indicating upward group propagation. Velocity data displayed in this format are informative. The eccentricity of the ellipse (a/b) indicates the ratio

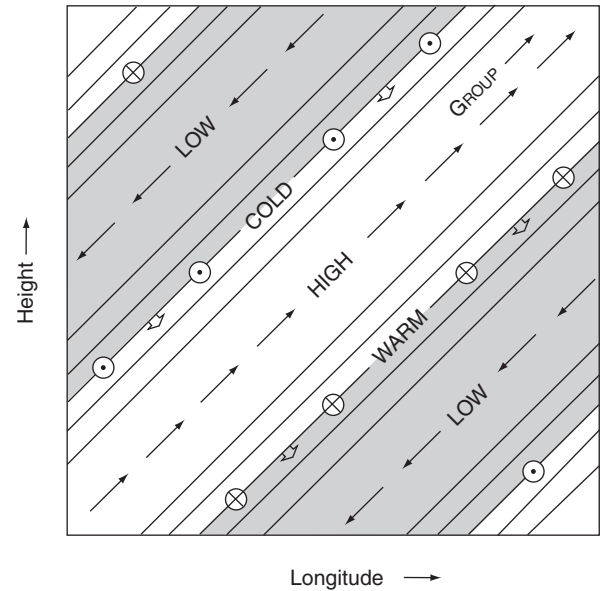


Figure 4 Structure of an inertia-gravity wave in the Northern Hemisphere, where the Coriolis parameter f is positive. Features resemble those of the internal gravity wave shown in [Figure 1](#), but with a meridional component of velocity included. Eastward and westward zonal wind anomalies generate southward and northward meridional wind anomalies, respectively, in quadrature with the zonal wind anomalies, and positively correlated with the temperature anomalies.

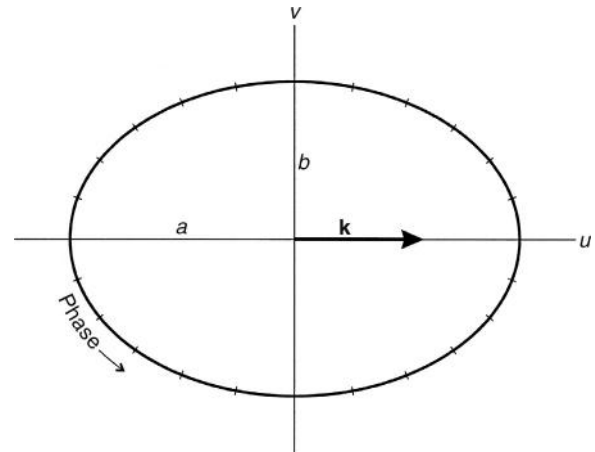


Figure 5 Hodograph of horizontal wind components for the inertia-gravity wave of [Figure 4](#). The horizontal component of the wavevector points in the positive x -direction, along the major axis of the ellipse. The vertical wavenumber m is negative, so that phase decreases with height.

of wave intrinsic frequency to local Coriolis frequency, while the orientation of the major axis indicates the direction of horizontal phase propagation, to within 180° . This directional ambiguity can be resolved if simultaneous temperature data are available, as implied by [Figure 4](#). In addition to a vertical flux of momentum, the inertia-gravity wave contributes to a horizontal flux of heat transverse to the direction of horizontal phase propagation.

Inertia-gravity wave parameters are related through a dispersion relation

$$(\hat{\omega}^2 - f^2)m^2 = (N^2 - \hat{\omega}^2)k^2 \quad [12]$$

where k and m are horizontal and vertical wavenumbers, respectively. This formula is valid for waves on an f -plane, in a constant or slowly varying mean flow, under the incompressible approximation; it is also approximately valid for atmospheric waves with vertical wavelength $\lambda_z \ll 4\pi H$ where $H = -(1/\rho) \times (d\rho/dz)$ is the density scale height where ρ is the horizontally averaged density. Most waves of interest satisfy this inequality. It can be seen from the dispersion relation that when the intrinsic frequency approaches f , the vertical wavenumber increases to infinity, i.e., the vertical wavelength approaches zero. The presence of f^2 in eqn [12] implies a larger vertical wavenumber than would otherwise be the case without the Coriolis force. One implication of this result is that inertia-gravity waves may break down via shear instability rather than via convective instability as wave amplitude increases.

According to eqn [12], waves with large intrinsic frequency have relatively steep surfaces of constant phase, while waves with small intrinsic frequency have relatively flat surfaces of constant phase. That phase surfaces become vertical in the limit $\hat{\omega} \rightarrow N$ is intuitively reasonable since N is the frequency of oscillation for vertical displacements in a stably stratified fluid. In this limit, the vertical group velocity and vertical wavenumber go to zero, and the waves are reflected from the surface where $\hat{\omega} = N$. In the opposite limit $\hat{\omega} \rightarrow f$, the vertical group velocity also approaches zero, but the vertical wavenumber approaches infinity, and the waves are likely to break and be absorbed.

Gravity waves are sometimes trapped in vertical waveguides or 'ducts' and are able to traverse a large horizontal distance. Long-period waves have relatively small vertical group velocity and are also able to propagate over a range of latitudes, so that the value of f seen by the waves changes slowly with time. Near the equator, inertia-gravity waves may become trapped within an equatorial waveguide, forming a modal structure, or equatorially trapped inertia-gravity wave. These modes constitute the 'fast manifold' of waves on an equatorial beta plane. The intrinsic period of equatorial inertia-gravity waves can be longer than a day, in contrast to mid-latitude gravity waves which are typically restricted to a range of intrinsic periods from several minutes to several hours.

There are several sources of gravity waves in the atmosphere, including flow over topography, convection, shear instability, and geostrophic adjustment. Orographic excitation is favored when surface winds are strong and directed across undulations of the topography. Convection generates gravity waves through pulsations in forcing or through a 'topographic' effect as the convective plume impinges on a stably stratified shear layer. Organized clusters of convection also generate gravity waves, but on a much larger scale than that of individual clouds. Sometimes the clusters themselves are organized by inertia-gravity waves to form a coupled moist dynamical system. This behavior is observed in the tropics,

over open ocean. Shear instability generates larger-scale gravity waves through envelope radiation. Geostrophic adjustment has long been regarded as a source of gravity waves radiating away from a state of initial imbalance. A similar process, the spontaneous emission of gravity waves, is now recognized for its role in evolving large-scale flows, such as baroclinic instabilities, in which regions of imbalance are generated through formation of frontal zones and regions of strong ageostrophic acceleration.

Gravity waves have several important effects on the general circulation of the atmosphere. They can accelerate the flow spontaneously, giving rise to such phenomena as the quasibiennial and semiannual oscillations of the tropical middle atmosphere. Gravity waves also decelerate the flow, reducing the speed of the winter polar night jet, summer mesospheric jet, and the topside of the tropospheric jet stream. These waves stimulate weather events in the troposphere, such as precipitation bands and new convective elements. Convection therefore is not only a source of gravity waves, but is also triggered by gravity waves generated by distant convection.

Simulation of the atmospheric general circulation requires either a proper parametrization of gravity-wave effects, or explicit simulation of the waves. Because gravity waves span a wide range of horizontal scales, it is impractical to simulate the entire spectrum. In most models, some combination of explicit modeling and parametrization is utilized.

See also: Arctic and Antarctic: Antarctic Climate. **Dynamical Meteorology:** Kelvin Waves; Kelvin–Helmholtz Instability; Symmetric Stability; Wave Mean-Flow Interaction; Waves. **Gravity Waves:** Buoyancy and Buoyancy Waves: Optical Observations; Overview. **Mesoscale Meteorology:** Density Currents. **Mountain Meteorology:** Lee Waves and Mountain Waves; Overview; Valley Winds.

Further Reading

- Andrews, D.G., Holton, J.R., Leovy, C.B., 1987. *Middle Atmosphere Dynamics*. Academic Press, Orlando, FL.
- Baines, P.G., 1995. *Topographic Effects in Stratified Flows*. Cambridge University Press, Cambridge.
- Gill, A.E., 1982. *Atmosphere–Ocean Dynamics*. Academic Press, Orlando, FL.
- Gossard, E.E., Hooke, W.H., 1975. *Waves in the Atmosphere*. Elsevier Scientific Publishing Co, New York.
- Hines, C.O., et al., 1974. *The Upper Atmosphere in Motion*. Heffernan Press, Worcester.
- LeBlond, P.H., Mysak, L.A., 1978. *Waves in the Ocean*. Elsevier Scientific Publishing Co, New York.
- Lighthill, J., 1978. *Waves in Fluids*. Cambridge University Press, Cambridge.
- Turner, J.S., 1973. *Buoyancy Effects in Fluids*. Cambridge University Press, Cambridge.
- Wallace, J.M., Hobbs, P.V., 1977. *Atmospheric Science: an Introductory Survey*. Academic Press, New York.

Gravity Waves Excited by Jets and Fronts

R Plougonven, Ecole Polytechnique, Palaiseau, France

F Zhang, Pennsylvania State University, University Park, PA, USA

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

Jets and fronts are known to be significant sources of gravity waves. This has been shown by numerous studies with real-data observations and/or numerical models. Case studies have highlighted the jet exit region as particularly favorable to low-frequency gravity waves. They have also identified wave ducting as a key maintenance mechanism for waves propagating over long distances near the ground. A number of observational studies also hypothesized that the primary mechanism for wave generation in the jet exit region is geostrophic adjustment, which has been recently generalized to the concept of spontaneous balance adjustment. However, the precise mechanisms responsible for the generation of these waves remain an active area of research. Further research remains necessary to draw a more complete picture of waves in the vicinity of jets and fronts, assessing the relevance of theoretical results on one hand, and the importance of other factors that contribute to the gravity waves, in particular moist processes and topography.

Introduction

In midlatitudes, jet/front systems have been known to be sources of gravity waves for a long time. More precisely, jets and fronts are associated with motions on fairly small scales (a few to a few hundred kilometers) and part of these have clearly been identified as gravity waves. Several essential elements leading to the generation of gravity waves from jets and fronts have been understood, yet the complexity of the flow is such that no simple diagnostic is readily available for a quantitative prediction of waves generated near jets and fronts. In such a complex, three-dimensional and time-dependent environment, the propagation of the waves needs to be considered alongside with their generation. The combination leads to the characteristics and amplitude of the waves that are found.

Such mesoscale waves are important for tropospheric dynamics and convection in the vicinity of fronts, for small-scale motions (mixing and turbulence) in the tropopause layer near upper-level jets, and for the energy and momentum transport that affects the general circulation of the atmosphere. The description of the latter, in climate models, depends in part on parameterizations of gravity waves. The description of nonorographic sources at midlatitudes remains unsatisfactory to date, and calls for a better understanding and knowledge of waves emanating from jets and fronts.

Below, we first review evidence from observations showing that jets and fronts are significant sources of gravity waves. Next, the elements of understanding coming from theoretical studies are presented. Open issues and current challenges are then discussed.

Observational Evidence

Gravity waves in the vicinity of jets and fronts have been identified in observations for several decades, both in case studies and in more statistical investigations. The focus has expanded, over the years, from tropospheric waves, typically detected in networks of surface barographs, to upper-tropospheric and stratospheric waves, detectable from a variety of different

measurements (radar, radiosondes, superpressure balloons, and satellites).

Early case studies have identified a configuration of the flow regularly present in events of intense gravity waves in the troposphere. In a midlatitude baroclinic wave, it consists of the region downstream of an upper-level jet, cold-air side of the surface front and close to the ridge of geopotential (Uccellini and Koch, 1987, see Figure 1, adapted from Koch and O'Handley, 1997). The relevance of this paradigm has been confirmed numerous times in subsequent case studies. For example, in the lower troposphere, Bosart et al. (1998) have documented an intense gravity wave observed in a midlatitude cyclone on the eastern coast of the United States (Figure 2). As the wave propagated northward along the coast, ducting in the lower layers and amplification of the wave played major roles, with the pressure falls reaching 13 hPa over a period of 30 min. Such large-amplitude waves in

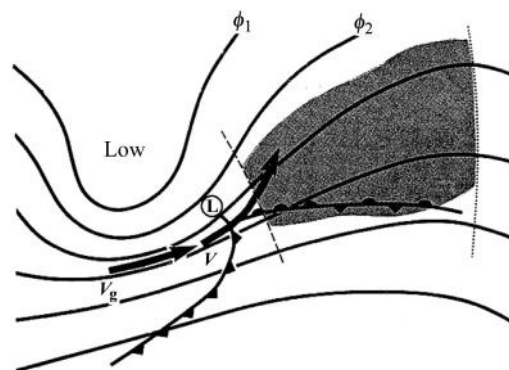


Figure 1 Schematic of the flow configuration conducive to the presence of large amplitude gravity waves, depicting the pressure at the surface (black lines), the surface fronts (conventional symbols), and the tropospheric jet (thick arrows). The region where large amplitude waves are commonly observed is shaded. Adapted from Uccellini, L., Koch, S., 1987. The synoptic setting and possible energy sources for mesoscale wave disturbances. *Mon. Wea. Rev.* 115, 721–729 by Koch, S., O'Handley, C., 1997. Operational forecasting and detection of mesoscale gravity waves. *Weather Forecasting* 12, 253–281.

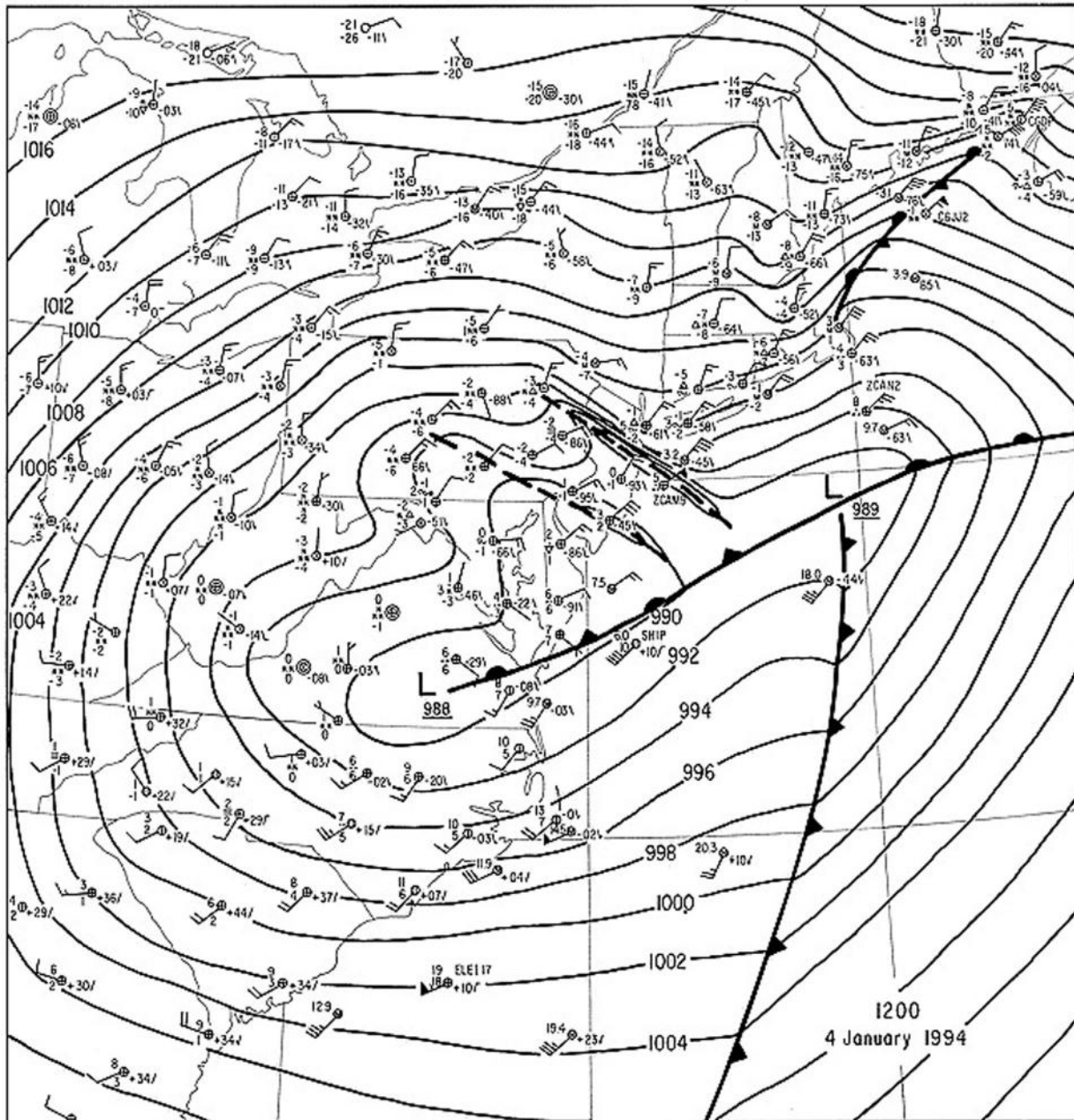


Figure 2 Large-amplitude gravity wave with a clear signature in the surface pressure (black lines), on the cold-air side of an occluded front. From the case study of Bosart, L., Bracken, W., Seimon, A., 1998. A study of cyclone mesoscale structure with emphasis on a large-amplitude inertia-gravity wave. *Monthly Weather Review* 126, 1497–1527.

the lower layers can have impacts on the organization and enhancement of moist convection but are challenging to be accurately simulated in convection-permitting models (Zhang et al., 2001). At upper levels, such waves emanating from jets have a potential to create regions of clear-air turbulence by modulating a background shear. In regions where the gravity wave enhances the shear, the latter can become sufficiently strong to become unstable (Lane et al., 2004).

Several other studies also identified the exit region of the upper-level jet as the source for stratospheric gravity waves (Guest et al., 2000; Hertzog et al., 2001; Wu and Zhang, 2004)

using sounding or satellite measurements in combination with numerical models and/or ray-tracing techniques. Other regions of the flow have also been highlighted as significant from observational case studies. First, jet exit regions upstream of a trough also appear favorable for enhanced gravity wave activity (Plougonven et al., 2003). In several studies of waves from upper-level jets, gravity waves have been found above and below the jet, propagating upward and downward, respectively. This is clear evidence that, in these cases, the jet is the source. These waves typically have intrinsic frequencies close to the Coriolis parameter (often called inertia-gravity waves). In

contrast to this, the vicinity of a cold front has also been highlighted as a source, in particular when significant wind impinges on the front (e.g., [Ralph et al., 1999](#)).

Complementary to case studies, observational evidence showing gravity waves from jet/front systems also comes from statistical studies, i.e., studies that examine small-scale motions in a large number of observations and relate those to likely sources (e.g., [Fritts and Nastrom, 1992](#); [Koppel et al., 2000](#); [Guest et al., 2000](#)). Many different types of observations can be used to identify gravity waves, from *in situ* measurements by high-resolution radiosondes, to global remote-sensing observations from satellites. Each type of observation will see only a part of the gravity wave spectrum. For example, high-resolution profiles of horizontal wind from radiosondes or Doppler radars are well suited to identify low-frequency gravity waves which have a clear signature in the hodograph (curve obtained from plotting u as abscissa and v as vertical coordinate), and waves with vertical wavelengths rather smaller than 10 km. Satellite observations from limb sounders, on the other hand, will capture deep gravity waves ([Alexander et al., 2010](#)). Presently, there is an ongoing effort to combine the knowledge acquired from the different datasets available, complemented by outcome of numerical models, in order to obtain a global estimation of the gravity wave field.

Global descriptions of the gravity wave field obtained from satellite observations typically highlight several large-scale regions with enhanced gravity wave activities: regions of orography, the Tropics, and the winter midlatitudes associated, respectively, with mountain waves, convective waves, and waves from jets and fronts. Often, maps will tend to highlight the hotspot regions of orography on the pass of large scale jet streams where the largest wave amplitudes are often found. Yet the area covered by storm tracks where waves from jets and fronts are found can be larger (e.g., [Wu and Zhang, 2004](#)). A quantitative comparison of the momentum fluxes integrated over a latitude band due to either source has been carried out using measurements from long-duration superpressure balloons flying in the lower stratosphere. Recent campaigns around Antarctica in austral spring have showed that, integrated over latitude circles, the contribution to momentum fluxes from gravity waves over the oceans was at least comparable to that from orographic waves ([Hertzog et al., 2008](#)).

In summary, observations have provided compelling evidence regarding the importance of jets and fronts as sources of gravity waves, and have identified several configurations particularly favorable to gravity waves, in particular jet exit regions. Mechanisms of generation, however, have not been clearly identified from observations alone. Starting in the 1990s, case studies have frequently included numerical simulations with a mesoscale meteorological model as complementary to observations (e.g., [Powers and Reed, 1993](#); [Zhang et al., 2001](#)). The more complete description of the flow that is provided by simulations allows a further investigation of generation mechanisms. Numerical simulations have highlighted the complexity of the flow, which is three-dimensional and time-evolving. Indeed, for a given wave event, there are generally several processes occurring, each of which may influence the emission and maintenance of waves (emission from the jet and/or front, interaction with topography and/or moist processes leading to amplification of the waves, ducting,

etc.). In consequence, there can be several interpretations to a given gravity wave event, even with the full description of the event available from a full three-dimensional numerical simulation. Different mechanisms have been described theoretically that contribute to generating and modulating the waves. This is currently a very active area of research, as it is difficult to isolate and describe quantitatively these source and maintenance mechanisms, for fundamental reasons as outlined below. A more extensive review of this subject can be found in [Plougonven and Zhang \(2014\)](#).

Theoretical Understanding

One would wish for a simple idealized model of gravity wave emission from jets and fronts that would capture the essential physics, as the linear study of constant flow over orography conveys the essential physics to start understanding orographic waves. Unfortunately, in contrast to orographic waves, it is not an external element which introduces short (intrinsic) timescales in the flow and hence forces gravity waves. It is the internal dynamics which produces these short timescales, making the problem much more difficult. Theory has nonetheless described several mechanisms that explain the propagation or the generation of some of the waves present in the vicinity of jets and fronts. Below we successively describe wave ducting, then paths that have been followed to illustrate how balanced motions can lead to the emission of gravity waves, and finally summarize recent studies which quantitatively explain low-frequency gravity waves that are found near jet exit regions.

Some elements of understanding have been obtained for quite a long time regarding the behavior of gravity waves near jets and fronts. Namely, in many early studies of waves detected near the surface, propagating away from a surface front, it was shown that the environment was such as to guide and maintain waves near the surface. These are called ducted waves ([Eom, 1975](#)). Ducting occurs when a stable layer is present near the ground, capped by a layer which efficiently reflects waves (e.g., of low stability, possibly beneath a critical level) ([Lindzen and Tung, 1976](#)). The presence of such a duct has been highlighted in many case studies (e.g., [Powers and Reed, 1993](#); [Zhang et al., 2001](#)). This emphasizes an important point: the background environment in which the waves propagate is very important, and it can help select some of the characteristics of the waves present.

The mechanisms for generation of the waves, on the other hand, have long remained elusive; quite often geostrophic adjustment has been referred to as the generation mechanism. Indeed, as a baroclinic wave grows in amplitude, one can diagnose a region of increasing imbalance (using diagnostics such as Lagrangian Rossby numbers or the residual of the nonlinear balance equation). Part of this imbalance takes the form of mesoscale gravity waves. Case studies have repeatedly shown the relevance of identifying regions of imbalance, which can be identified even in a coarse-grained description of the flow ([Koch and Dorian, 1988](#); [Bosart et al., 1998](#); [Zhang et al., 2001](#)).

Classical geostrophic adjustment theories typically describe the evolution of either small perturbations to a fluid at rest, or

of perturbations to a flow that has a symmetry (zonally symmetric flow or axisymmetric flow) (Rossby, 1938; Blumen, 1972). In both cases, balanced motions and gravity waves decouple, so that the geostrophic adjustment problem is well-posed and analytical treatment becomes possible. Geostrophic adjustment constitutes a fundamental building block of our understanding of motions in a stratified, rotating fluid. It illustrates how, in a first approximation, motions decompose into slowly evolving motions near geostrophic balance and fast gravity waves. However, to understand how gravity waves are excited by jets and fronts, one needs to understand more precisely in what circumstances these motions couple, and how some imbalance energy may continuously be produced from the balanced motions and to generate gravity waves. One also needs to consider higher-order balance than geostrophy. To generalize the classical geostrophic adjustment theory, Zhang (2004) introduced the concept of (spontaneous) balance adjustment under which the large-scale synoptic background flow such as baroclinic jet-front system continuously produces flow imbalance (away from nonlinear balance that is more precise than geostrophy) which continuously forces gravity waves. A framework to describe spontaneous balance adjustment has been proposed (Plougonven and Zhang, 2007) and its successful use to describe waves emitted in dipoles will be described further below, at the end of the present section.

An important nondimensional number which indicates the relevance of balanced dynamics is the Rossby number. Consider a flow with L a typical length scale and U a typical velocity. The Rossby number is the ratio of the inertial timescale, $1/f$, where f is the Coriolis parameter, and the advective timescale, L/U :

$$Ro = U/fL \quad [1]$$

When $Ro \ll 1$, this indicates that advection occurs on timescales much longer than the fastest of the gravity waves (inertia-gravity waves with periods close to the inertial period). The dynamics is then well captured by approximations which describe the part of the flow that is tied to the advection of potential vorticity, i.e., the balanced or vortical part of the flow. Now, jets and fronts are predominantly balanced motions: they are rather well described by assuming the flow to be near geostrophic and hydrostatic balance. Such assumptions lead to approximations, such as quasigeostrophy, which simplify the flow by explicitly filtering gravity wave motions. Higher-order approximation, such as *semigeostrophy*, has been shown to describe essential features of frontogenesis. In other terms, the theoretical models available to understand fronts exclude gravity waves by construction.

Therefore, determining what gravity waves are generated from jets and fronts requires to understand how the evolution of flows that are balanced initially leads to emission of gravity waves. This *spontaneous emission* (or spontaneous adjustment emission, SAE) is quite distinct from classical geostrophic adjustment problems, in which the imbalance is prescribed in the initial conditions. Quantifying spontaneous emission sheds light on the limitations of balanced models, but this is an arduous problem which has motivated numerous studies over the past three decades.

Three pathways for spontaneous emission have been identified and quantified analytically and are briefly discussed here: Lighthill radiation, unbalanced instabilities and transient

generation in shear. Lighthill radiation is the emission of large scale waves by small-scale balanced motions which have frequencies high enough to project onto the gravity wave spectrum (Ford et al., 2000), i.e., for which $Ro > 1$. The study of this mechanism builds on the pioneering work of James Lighthill (1924–98) who had described the generation of acoustic waves from turbulent motions. Although this mechanism applies to waves generated from intense vortices in the atmosphere, the scale separation described does not transpose to waves emanating from jets and fronts. The other two mechanisms, unbalanced instabilities and transient generation in shear, couple balanced motions and gravity waves in flow configurations with $Ro \ll 1$. An unbalanced instability involves both balanced and unbalanced motions, and is therefore absent from all balanced models, whatever their accuracy. Unbalanced instabilities have been calculated for a wide range of flows, and have recently been observed in the laboratory, but they generally have weak growth rates (see Plougonven and Zhang, 2014 and refs. therein). Transient generation in shear is the emission of waves from sheared anomalies of potential vorticity. This emphasizes that in a nontrivial background flow, the decomposition into balanced motions and gravity waves is no longer well-defined, in contrast to small perturbations in a fluid at rest. For both mechanisms, the main point to retain is that shear allows to couple ‘slow’ and ‘fast’ motions because of Doppler shifting. The waves have been quantified analytically and are exponentially small in Ro , i.e., their amplitude typically scales as

$$\exp(-\alpha/Ro) \quad [2]$$

with a prefactor that may include an algebraic dependence in Ro , and where α is a constant. In practice, such a dependence implies that the waves seem absent for small Ro , and appear to ‘turn on’ past a finite value of Ro . Such results have been obtained using sophisticated asymptotic techniques (see Vanneste, 2013 and refs. therein) and constitute valuable milestones. However, the mechanisms described above do not relate straightforwardly to cases of inertia-gravity waves in the vicinity of jets and fronts. Flow configurations closer to those observed have been obtained by investigating flows of intermediate complexity using idealized numerical simulations.

Much understanding of baroclinic waves and frontogenesis has come from idealized simulations of these processes. Likewise, idealized simulations of frontogenesis and of baroclinic life cycles have provided essential elements of understanding for gravity waves generated from jets and fronts (Snyder et al., 1993). The first simulations of baroclinic life cycles with a hemispheric model (O’Sullivan and Dunkerton, 1995) confirm the presence of inertia-gravity waves emitted from the upper-level jet in simulations initiated from an initially balanced jet. However, the model grid resolution they used raised concern of possible numerical artifacts.

With increasing computing power and the development of mesoscale meteorological models, it became possible to run an idealized simulation with nested domains allowing a resolution fine enough to be confident that the waves were well resolved (Figure 3). The waves found in these simulations differed from the previous ones, had shorter scales and larger amplitudes. Both upper-level jets and surface fronts were shown to contribute, in different ways. Gravity waves from

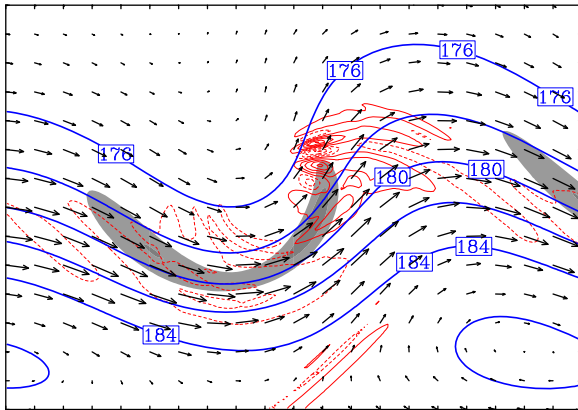


Figure 3 The 13-km pressure (thick blue line, every 2 hPa), horizontal divergence (thin red line; solid, positive; dashed, negative; every $5 \times 10^{-6} \text{ s}^{-1}$), and wind vectors (maximum of 25 ms^{-1}) simulated from the triple-nested mesoscale model MM5 with horizontal (vertical) resolutions of 10 km (360 m). The wind speed at 8 km (near the maximum jet strength level) greater than 45 ms^{-1} is shaded in gray (every 5 ms^{-1}). The distance between tick marks is 300 km. Adapted from Figure 12d of Zhang, F., 2004. Generation of mesoscale gravity waves in upper-tropospheric jet-front systems. *Journal of Atmospheric Science* 61(4), 440–457.

these high-resolution multiply nested mesoscale model simulations as shown in Figure 3 (Zhang, 2004) are to a large extent consistent with the conceptual model of Uccellini and Koch (1987; Figure 1) that are derived from observational mesoscale gravity wave case studies.

In many cases, both from observations and from idealized case studies, jet exit regions have been emphasized as a favored locus for the presence of intense inertia-gravity waves. This suggested that this region was particularly relevant for generation. It turned out to be insightful to consider also the propagation of the waves. Indeed, jet exit regions are characterized by strong strain and important vertical shears. Simple considerations (a WKB approach) on the propagation of a wave packet

in a background flow made of a pure deformation field and vertical shear was shown to lead to waves with a specific intrinsic frequency and with phase lines oriented along the extensional axis of the deformation field. This is consistent with the orientation and intrinsic frequencies of waves simulated and observed in such regions. This phenomenon has been called ‘wave-capture’, and has been emphasized as providing a route for the background flow to lead a gravity wave to dissipation (as the wavelength contracts) (Bühler and McIntyre, 2005). Whereas the implications for wave-mean flow interactions remain an area of investigation, the relevance of these ‘propagation effects’ have been verified and illustrated in a number of idealized studies (Plougonven and Snyder, 2005, 2007; Lin and Zhang, 2008; Wang et al., 2010).

Albeit idealized, baroclinic life cycles remain complex (three-dimensional, time-evolving flow). More detailed and quantitative understanding of the wave generation process has been obtained over recent years by focusing on a yet simpler flow: vortex dipoles in a continuously stratified fluid. This constitutes a simple model of jet streaks. It is truly three-dimensional, retains a local maximum of wind speed, and hence a region of deceleration where streamlines are diffluent. Yet it greatly simplifies the flow relative to baroclinic life cycles, because it propagates nearly steadily. More precisely: the simulated dipoles were not exact steadily propagating solutions to the model equations, but their evolution had timescales (tens of days) much longer than those of gravity waves, making it possible to consider that a dipole is stationary in the appropriate comoving frame.

Numerical experiments have been carried out by several groups with very different numerical models and configurations (Snyder et al., 2007; Viudez, 2008; Wang et al., 2009). A robust phenomenology emerged from these simulations, with a wave packet found in the front of the dipole, in the jet exit region, with phaselines, intrinsic frequencies and sensitivity to resolution that are consistent with expectations from wave-capture (Figure 4). In initial times of the different simulations, an adjustment takes place, but the vertical velocity then settles down to a pattern that is essentially stationary in the frame moving with the dipole. For small Rossby number ($Ro < 0.05$),

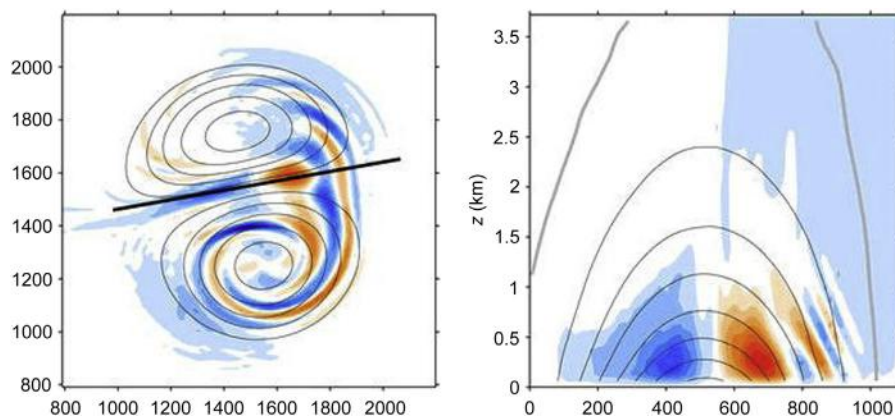


Figure 4 Idealized simulation of a surface dipole, with a conspicuous wave packet emitted in the jet exit region. The left panel shows vertical velocity at altitude $z = 62.5 \text{ m}$ (colors), and potential temperature (contours). The right panel shows a vertical cross section along the axis of the dipole, with vertical velocity (colors) and horizontal velocity in the plane of the cross section (contours). All axis labels are in kilometers. Adapted from Figure 9 of Snyder, C., Muraki, D.J., Plougonven, R. and Zhang, F. 2007. Inertia-gravity waves generated within a dipole vortex. *Journal of Atmospheric Science* 64, 4417–4431. © American Meteorological Society. Used with permission.

no easily detectable wave signal is found, whereas for the finite Rossby numbers that can be attained ($0.1 < Ro < 0.2$) the waves may dominate the vertical velocity field (Snyder et al., 2009; Wang et al., 2009).

The generation of the waves in these dipole experiments is explained by decomposing the flow into a balanced approximation of the dipole and a correction (Snyder et al., 2009; Wang and Zhang, 2010). The correction can be considered a solution of equations that are linearized about the balanced dipole. Now the latter is not an exact solution, so that when injected into the full equations it yields nonzero residual tendencies. These provide a forcing for the correction. In the center of the dipole, advection is strong, leading to short intrinsic timescales. In this region the waves are generated, and their propagation in the jet exit influences the characteristics consistent with wave capture (McIntyre, 2009). It is essential in the above that the equations for the corrections are linearized about the dipole flow. This background flow is three-dimensional and strongly constrains the waves that emerge. Intuitions built on models of wave emission in a fluid at rest are not applicable here.

The above discussion has not touched on another pathway for energy to be transferred from balanced motions to gravity waves: shear instability is expected to occur in regions where frontogenesis leads to very intense shears. It has been suspected for nearly four decades that this could constitute a source of gravity waves (e.g., Jones, 1968; Mastrantonio et al., 1976). However, it has been difficult to quantify this emission. Linear stability analysis shows that, in general, the most unstable modes due to shear will be of Kelvin–Helmholtz type, occurring on scales too short (intrinsic frequencies too high) to connect directly to gravity waves. Part of the energy of the unstable motions may project onto gravity waves, through fluctuations of the envelope of the unstable region, or by the geostrophic or balanced adjustment in this region once it has been partially and rapidly mixed by the instability. It is necessary to describe a wide range of scales to quantify this emission, from the large scale of the baroclinic disturbances (a thousand kilometers) to the short scales (a few kilometers) of the instability. This hinders the precise estimation of waves from this source, which remains an open issue.

Open Questions and Challenges

One remarkable outcome from observations and numerical modeling has been the robustness of the paradigm put forward by Uccellini and Koch (1987), and the dynamical understanding obtained since. The convergence of different approaches (observational case studies, idealized simulations) and the recurrence of this configuration in numerous studies are indications of the robustness of this result. Theory has highlighted propagation effects, either through wave-ducting of Lindzen and Tung (1976) or in ‘wave-capture’ of Bühler and McIntyre (2005), as mechanisms for maintaining and enhancing IGW in certain regions of the flow, the large-scale environment determining certain of the wave characteristics. The emission mechanism has been explained as the linear response to the differences between the balanced and the full tendencies. The key point is that the dynamics are linearized on the background of an approximation of the dipole.

This constitutes an encouraging, positive result, but several issues remain to be explored: it is necessary to go back to real flows with this enhanced theoretical understanding, to confirm that the characteristics of the waves are indeed consistent with those expected from propagation effects in the shear and strain present. The amplitudes of waves in idealized experiments have repeatedly been found to be weaker than in observations, and this constitutes a second issue. More importantly, in both idealized simulations and real flows, other sources of waves need to be assessed, and in particular sources that lead to waves with shorter scales and higher intrinsic frequencies, which may carry significant contributions to momentum fluxes toward the stratosphere (Fritts and Alexander, 2003). Simulations and observations have provided indications that such waves are present, but their shorter scale makes them more difficult to identify and quantify, let alone to predict.

In coming years, a major challenge will be to bring together descriptions of gravity waves from different observation sets and from high-resolutions modeling (Alexander et al., 2010). Tying gravity waves quantitatively to diagnostics of the tropospheric flow will be an important goal. Several possible routes for such a relation have been suggested, but a systematic, quantitative assessment of their relevance remains to be undertaken. Furthermore, theoretical studies so far have concentrated on dry dynamics, yet case studies have often emphasized moist processes, at least as amplifying waves. Recent studies showed that moisture may lead to significantly enhanced gravity wave activities in the vicinity of jets and fronts (Waite and Snyder, 2012). In the idealized simulations of moist baroclinic waves, Wei and Zhang (2014) found that not only moist convection can be triggered by dry gravity waves from jet streaks, it can significantly enhance the parent gravity waves by as much as orders of magnitude, along with additions of high-amplitude and high-frequency smaller scale gravity waves. The interactions of jet-front systems with convection and/or topography in the generation of gravity waves remain to be thoroughly explored. Such progress in the quantitative understanding of waves near fronts will be an important step for understanding the impact of gravity waves to the general circulation, and for the improvement of parameterizations of gravity wave effects in climate models.

See also: Aviation Meteorology: Clear Air Turbulence.

Dynamical Meteorology: Balanced Flow; Overview; Wave Mean-Flow Interaction; Wave-CISK; Waves. **Gravity Waves:** Buoyancy and Buoyancy Waves: Optical Observations; Buoyancy and Buoyancy Waves: Theory; Convectively Generated Gravity Waves; Overview. **Synoptic Meteorology:** Frontogenesis; Jet Streaks.

References

- Alexander, M., et al., 2010. Recent developments in gravity-wave effects in climate models and the global distribution of gravity-wave momentum ux from observations and models. *Quarterly Journal of the Royal Meteorological Society* 136, 1103–1124.
- Blumen, W., 1972. Geostrophic adjustment. *Reviews in Geophysics and Space Physics* 10 (2), 485–528.
- Bosart, L., Bracken, W., Seimon, A., 1998. A study of cyclone mesoscale structure with emphasis on a large-amplitude inertia-gravity wave. *Monthly Weather Review* 126, 1497–1527.

- Bühler, O., McIntyre, M., 2005. Wave capture and wave-vortex duality. *Journal of Fluid Mechanics* 534, 67–95.
- Eom, J., 1975. Analysis of the internal gravity wave occurrence of 19 April 1970 in the midwest. *Monthly Weather Review* 103, 217–226.
- Ford, R., McIntyre, M.E., Norton, W.A., 2000. Balance and the slow quasimanifold: some explicit results. *Journal of Atmospheric Science* 57, 1236–1254.
- Fritts, D., Alexander, M., 2003. Gravity wave dynamics and effects in the middle atmosphere. *Reviews in Geophysics* 41 (1), 1003.
- Fritts, D., Nastrom, G., 1992. Sources of mesoscale variability of gravity waves. Part II: Frontal, convective, and jet stream excitation. *Journal of Atmospheric Science* 49 (2), 111–127.
- Guest, F., Reeder, M., Marks, C., Karoly, D., 2000. Inertia-gravity waves observed in the lower stratosphere over Macquarie Island. *Journal of Atmospheric Science* 57, 737–752.
- Hertzog, A., Souprayen, C., Hauchecorne, A., 2001. Observation and backward trajectory of an inertia-gravity wave in the lower stratosphere. *Annals of Geophysics* 19, 1141–1155.
- Hertzog, A., Boccara, G., Vincent, R., Vial, F., Coquerez, P., 2008. Estimation of gravity-wave momentum fluxes and phase speeds from long-duration stratospheric balloon flights. 2. Results from the Vorcore campaign in Antarctica. *Journal of Atmospheric Science* 65, 3056–3070.
- Jones, W.L., 1968. Reflexion and stability of waves in stably stratified fluids with shear flow: a numerical study. *Journal of Fluid Mechanics* 34, 609–624.
- Koch, S.E., Dorian, P.B., 1988. A mesoscale gravity wave event observed during CCOPE. Part III: Wave environment and possible source mechanisms. *Monthly Weather Review* 116, 2570–2591.
- Koch, S.E., O'Handley, C., 1997. Operational forecasting and detection of mesoscale gravity waves. *Weather Forecasting* 12, 253–281.
- Koppel, L., Bosart, L., Keyser, D., 2000. A 25-yr climatology of large-amplitude hourly surface pressure changes over the conterminous United States. *Monthly Weather Review* 128 (1), 51–68.
- Lane, T., Doyle, J., Plougonven, R., Sharman, R., Shapiro, M., 2004. Numerical modeling of gravity waves and shearing instabilities above an observed jet. *Journal of Atmospheric Science* 61, 2692–2706.
- Lin, Y., Zhang, F., 2008. Tracking gravity waves in baroclinic jet-front systems. *Journal of Atmospheric Science* 65, 2402–2415.
- Lindzen, R., Tung, K.-K., 1976. Banded convective activity and ducted gravity waves. *Monthly Weather Review* 104, 1602–1617.
- Mastrantonio, G., Einaudi, F., Fua, D., Lalas, D.P., 1976. Generation of gravity waves by jet streams in the atmosphere. *Journal of Atmospheric Science* 33, 1730–1738.
- McIntyre, M., 2009. Spontaneous imbalance and hybrid vortex-gravity wave structures. *Journal of Atmospheric Science* 66, 1315–1326.
- O'Sullivan, D., Dunkerton, T., 1995. Generation of inertia-gravity waves in a simulated life cycle of baroclinic instability. *Journal of Atmospheric Science* 52 (21), 3695–3716.
- Plougonven, R., Snyder, C., 2005. Gravity waves excited by jets: propagation versus generation. *Geophysical Research Letters* 32 (L18892). <http://dx.doi.org/10.1029/2005GL023730>.
- Plougonven, R., Snyder, C., 2007. Inertia-gravity waves spontaneously generated by jets and fronts. Part I: Different baroclinic life cycles. *Journal of Atmospheric Science* 64, 2502–2520.
- Plougonven, R., Zhang, F., 2007. On the forcing of inertia-gravity waves by synoptic-scale flows. *Journal of Atmospheric Science* 64, 1737–1742.
- Plougonven, R., Zhang, F., 2014. Internal gravity waves from atmospheric jets and fronts. *Reviews in Geophysics* 52 (1), 33–76. <http://dx.doi.org/10.1002/2012RG000419>.
- Plougonven, R., Teitelbaum, H., Zeitlin, V., 2003. Inertia-gravity wave generation by the tropospheric mid-latitude jet as given by the FASTEX radio soundings. *Journal of Geophysical Research* 108 (D21), 4686. <http://dx.doi.org/10.1029/2003JD003535>.
- Powers, J., Reed, R., 1993. Numerical simulation of the large-amplitude mesoscale gravity wave event of 15 December 1987 in the Central United States. *Monthly Weather Review* 121, 2285–2308.
- Ralph, F.M., Neiman, P.J., Keller, T.L., 1999. Deep-tropospheric gravity waves created by leeside cold fronts. *Journal of Atmospheric Science* 56, 2986–3009. [http://dx.doi.org/10.1175/1520-0469\(1999\)056<2986:DTGWCB>2.0.CO;2](http://dx.doi.org/10.1175/1520-0469(1999)056<2986:DTGWCB>2.0.CO;2).
- Rossby, C., 1938. On the mutual adjustment of pressure and velocity distributions in certain simple current systems II. *Journal of Marine Research* 1, 239–263.
- Snyder, C., Skamarock, W., Rotunno, R., 1993. Frontal dynamics near and following frontal collapse. *Journal of Atmospheric Science* 50 (18), 3194–3211.
- Snyder, C., Muraki, D., Plougonven, R., Zhang, F., 2007. Inertia-gravity waves generated within a dipole vortex. *Journal of Atmospheric Science* 64, 4417–4431.
- Snyder, C., Plougonven, R., Muraki, D., 2009. Forced linear inertia-gravity waves on a basic-state dipole vortex. *Journal of Atmospheric Science* 66 (11), 3464–3478.
- Uccellini, L., Koch, S., 1987. The synoptic setting and possible energy sources for mesoscale wave disturbances. *Monthly Weather Review* 115, 721–729.
- Vanneste, J., 2013. Balance and spontaneous wave generation in geophysical flows. *Annual Reviews of Fluid Mechanics* 45, 147–172.
- Viudez, A., 2008. The stationary frontal wave packet spontaneously generated in mesoscale dipoles. *Journal of Physical Oceanography* 38, 243–256.
- Wang, S., Zhang, F., 2010. Source of gravity waves within a vortex-dipole jet revealed by a linear model. *Journal of Atmospheric Science* 67, 1438–1455.
- Waite, M.L., Snyder, C., 2012. Mesoscale energy spectra of moist baroclinic waves. *Journal of Atmospheric Science* 70, 1242–1256.
- Wang, S., Zhang, F., Snyder, C., 2009. Generation and propagation of inertia-gravity waves from vortex dipoles and jets. *Journal of Atmospheric Science* 66, 1294–1314.
- Wang, S., Zhang, F., Epifanio, C., 2010. Forced gravity wave response near the jet exit region in a linear model. *Quarterly Journal of the Royal Meteorological Society* 136, 1773–1787.
- Wei, J., Zhang, F., 2014. Mesoscale gravity waves in moist baroclinic jet-front systems. *Journal of Atmospheric Science* 71, 929–952.
- Wu, D.L., Zhang, F., 2004. A study of mesoscale gravity waves over the North Atlantic with satellite observations and a mesoscale model. *Journal of Geophysical Research* 109, D22104. <http://dx.doi.org/10.1029/2004JD005090>.
- Zhang, F., 2004. Generation of mesoscale gravity waves in upper-tropospheric jet-front systems. *Journal of Atmospheric Science* 61 (4), 440–457.
- Zhang, F., Koch, S., Davis, C., Kaplan, M., 2001. Wavelet analysis and the governing dynamics of a large amplitude mesoscale gravity wave event along the east coast of the United States. *Quarterly Journal of the Royal Meteorological Society* 127, 2209–2245.

Convectively Generated Gravity Waves

TP Lane, The University of Melbourne, Melbourne, VIC, Australia

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

Atmospheric convection is an important source of gravity waves. The gravity waves are caused by vertical displacements of stable air associated with convective elements, along with the diabatic heating and cooling within moist convection. Gravity waves can be generated by shallow or deep convection and the waves normally propagate vertically and horizontally away from their source. While in the troposphere, convective gravity waves can affect the stability and modify further convective development. In the troposphere, stratosphere, and mesosphere the dissipation of convective gravity waves exerts a tendency on the mean flow and thereby contributes to the momentum budget of those layers.

Underlying Theory and Generation Mechanisms

Dry thermals or plumes represent the simplest model of a convective cloud. Consider an isolated warm perturbation (i.e., a thermal) in a stratified environment. The thermal will have positive buoyancy and be accelerated upward. Once the thermal reaches the altitude where its temperature is equal to that of the environment, known as the equilibrium level, its buoyancy will be zero but it will continue to rise due to its existing upward momentum. As the thermal penetrates above the equilibrium level, the thermal will have negative buoyancy and be forced to return to the equilibrium level. Hence an oscillation of the thermal ensues, which also causes an oscillation of the stable air above and below the thermal's boundary. This oscillation generates gravity waves that radiate away from the thermal. See the model simulation in Figure 1.

Here, the thermal is analogous to a simple harmonic oscillator with the restoring force provided by buoyancy associated with the stable stratification. However, the oscillation is damped by entrainment processes that dilute the buoyancy and by the flux of momentum and energy away from the thermal caused by the propagating waves. Thermals in neutral environments do not generate gravity waves, except if they impinge upon a stable interface like an inversion at the top of the convective boundary layer.

The pattern of waves generated by the thermal shown in Figure 1 is similar to the classical St Andrew's cross pattern obtained in laboratory experiments. That pattern is obtained in the laboratory using a cylindrical solid body oscillating in a stratified fluid. Consider the two-dimensional linear equations of motion for an inviscid Boussinesq fluid, which are:

$$\frac{\partial u}{\partial t} + U \frac{\partial u}{\partial x} = -\frac{1}{\rho_0} \frac{\partial p}{\partial x}, \quad [1a]$$

$$\frac{\partial w}{\partial t} + U \frac{\partial w}{\partial x} = -\frac{1}{\rho_0} \frac{\partial p}{\partial z} + b, \quad [1b]$$

$$\frac{\partial b}{\partial t} + U \frac{\partial b}{\partial x} + w N^2 = 0, \quad [1c]$$

$$\frac{\partial u}{\partial x} + \frac{\partial w}{\partial z} = 0, \quad [1d]$$

where u is the horizontal velocity perturbation from the mean horizontal velocity U , w is the vertical velocity, p is the perturbation pressure, ρ_0 is the reference density, b is the buoyancy, N is the Brunt-Väisälä frequency, x and z are the horizontal and vertical coordinates, respectively, and t is time. Here $N^2 = (g/\theta_0)(d\theta/dz)$, where g is the gravitational acceleration, θ is the potential temperature, and θ_0 is its reference value; the buoyancy is defined as $b = g\theta'/\theta_0$, where θ' is the perturbation potential temperature. The St Andrew's cross wave response can be easily reproduced in numerical experiments with the solid body replaced by an isolated oscillating source of momentum or heat applied to the right hand sides of eqns [1a–b] or eqn [1c], respectively.

For example, Figure 2 shows results from two numerical experiments that apply a periodic source of localized heating and cooling to eqn [1c] (and solve eqns [1a–d]). This diabatic forcing has the form $Q = Q_0 q(x, z) \cos(\Omega t)$, where Ω is the source frequency and for this example q is a Gaussian function in the horizontal and a half sinusoid in the vertical (see Figure 2 for the outline of the forcing). In these experiments the background wind, U , is zero. The diabatic forcing is a simplified representation of a moist convective cloud; waves generated by real clouds will be reflected at the ground; and wave propagation characteristics will change with vertical variations in wind and stability. As with real clouds, gravity waves are generated that propagate in both horizontal directions away from the source.

When the amplitude of the oscillation is small (i.e., nonlinear effects can be neglected), the wave frequency, ω , is equal to the source frequency, Ω . The frequency of the source and the Brunt-Väisälä frequency of the environment determine the angle of the wave phase lines from the vertical, α . Specifically the dispersion relation for two-dimensional linear gravity waves in the absence of background flow is

$$\frac{\omega^2}{N^2} = \frac{k^2}{m^2 + k^2} = \cos^2(\alpha), \quad [2]$$

where k is the horizontal wave number, m is the vertical wave number, and the horizontal and vertical wavelengths are $\lambda_x = 2\pi/k$ and $\lambda_z = 2\pi/m$, respectively. Thus, for all else equal, a higher frequency source generates waves with phase lines that are oriented closer to the vertical than a lower frequency source. This property is illustrated in Figure 2. When $|\Omega| > N$ the

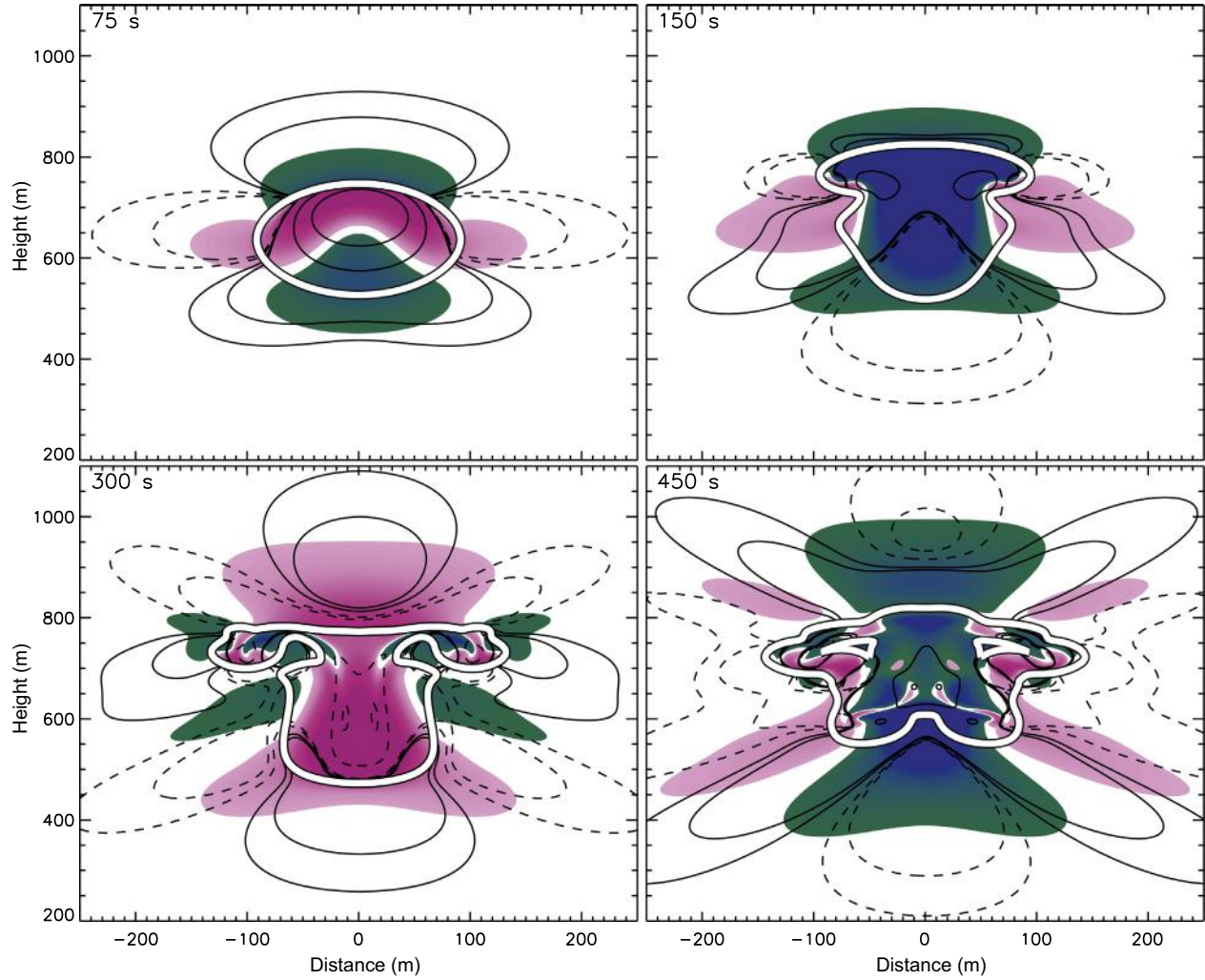


Figure 1 Numerical simulation of a thermal evolving in a stably stratified environment (with $N = 0.02 \text{ s}^{-1}$). The black-edged white line is the thermal boundary. Colored shading represents buoyancy with positive values pink and negative values green and blue. Vertical velocity is contoured black (at intervals of $\pm 0.025, \pm 0.05, \pm 1, \pm 2 \text{ m s}^{-1}$) with negative values dashed.

generated waves are evanescent, which means that their amplitude decays exponentially with height (and satisfy $m^2 < 0$).

Linear theory of gravity waves also shows that in steady background flows that only vary in the vertical (i.e., flow with no horizontal variations in mean wind speed or stability), the horizontal phase speed $c = \omega/k$, the horizontal wavelength, and the wave frequency are all conserved following a wave packet. In such a framework, gravity waves respond to vertical variations in stability and horizontal wind through changes in the vertical wave number. The influence of changes in horizontal wind on gravity wave properties can be incorporated into eqn [2] as

$$\frac{(\omega - Uk)^2}{N^2} = \frac{k^2}{m^2 + k^2} = \cos^2(\alpha), \quad [3]$$

where terms involving wind shear and wind curvature have been neglected for simplicity. The 'intrinsic frequency', $\omega - Uk$, is the wave frequency in a frame of reference moving with the

wind. Unlike the actual wave frequency (ω), the intrinsic frequency is not conserved and therefore changes in the wind will modify the vertical wavelength and the angle of the phase lines from the vertical. If, for example, U increased above a convective cloud or thermal, it can be shown that the phase lines of the waves propagating into the wind shear vector would be tilted toward the vertical and those propagating with the wind shear vector would be tilted to the horizontal. If this change of wind is sufficiently large, such that $|\omega - Uk| > N$, the waves propagating into the wind shear vector would become evanescent. Alternatively, a wave critical level would occur for those waves propagating with the wind shear vector if $\omega - Uk = 0$, and the waves would be dissipated in some way (e.g., wave breaking).

The examples provided in Figure 2 created gravity waves by applying an external source of diabatic heating and cooling. Diabatic forcing in moist convection is derived from phase changes of water: condensation causes heating and evaporation causes cooling. However, dry convection (i.e., thermals)

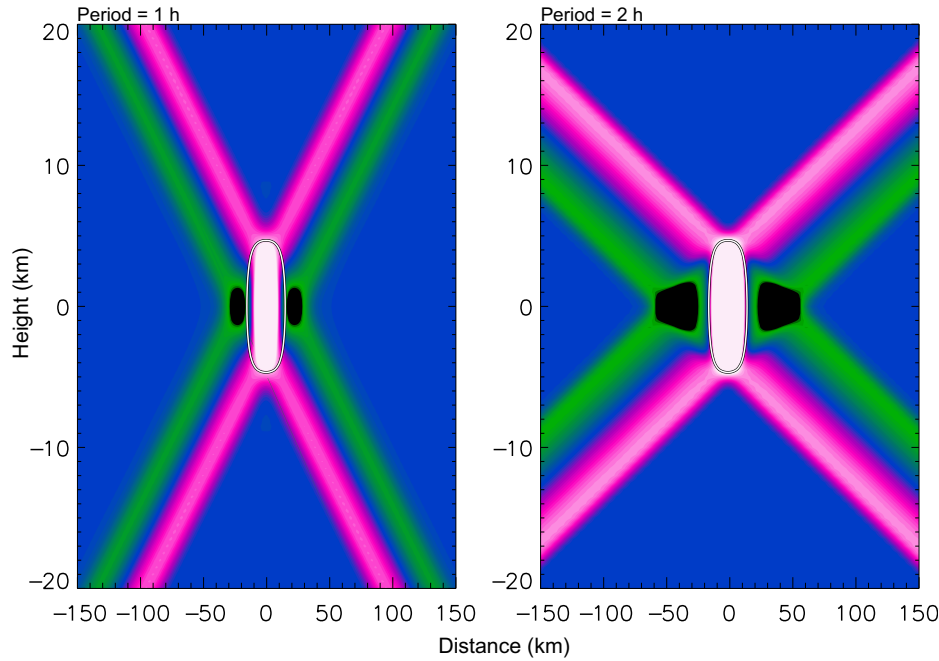


Figure 2 Vertical velocity patterns generated by localized thermal forcing in a stratified environment with $N = 0.01 \text{ s}^{-1}$. The forcing periods ($2\pi/\Omega$) are 1 h (left) and 2 h (right), and the solutions are shown at time (t) that satisfies $\cos(\Omega t) = 1$. Pink through white colors are upward motion and green through black colors are downward motion. The region of forcing is outlined with the black-edged white line.

generates gravity waves via adiabatic processes, which means that there is no diabatic forcing on the right-hand side of eqn [1c] associated with phase changes of water. Instead, the wave generation is associated with internal body forces caused by the oscillating thermals. These body forces can be represented mathematically as forcing on the right hand side of eqns [1a–c] that is sourced from nonlinear processes within and around the oscillating thermals.

Moist convective clouds generate gravity waves via a combination of diabatic forcing and nonlinear body forces. The heating and cooling associated with condensation and evaporation explains part of the wave spectrum. The typical life cycle of a mesoscale convective system (MCS) is characterized by mean heating and cooling that is transient with timescales greater than about an hour. Accordingly, the transient heating and cooling explains waves with periods greater than about an hour, with horizontal wavelengths defined, in part, by the horizontal extent of the system. The flow shown in Figure 2 is directly analogous to this generation process.

Convective updrafts embedded within MCSs behave similarly to thermals (e.g., Figure 1) and can be represented by nonlinear body forces within the clouds. Like thermals, convective updrafts overshoot their equilibrium level and generate waves with frequencies that are close to the Brunt–Väisälä frequency; thus, these waves can have periods as small as 10 min, i.e., shorter than the typical convective life cycle of about an hour or more. The horizontal wavelength of these waves is related to the width of the convective updrafts, which can be smaller than the overall size of the MCS that the updrafts are embedded within. Deep convective clouds that span the troposphere contain convective updrafts that

overshoot the tropopause, which can increase the amplitude and frequency of the generated waves considerably.

The generation of updrafts and downdrafts within moist convection is, however, ultimately related to diabatic processes and it could be argued that the diabatic processes are responsible for all of the wave generation. However, for deep clouds the overshoot and oscillation of convective updrafts occurs in the upper troposphere or lower stratosphere, where latent heating is weaker (due to the smaller mass of condensed water) and a large part of the wave generation occurs due to the vertical displacement of stable air above the cloud, where the flow is adiabatic (i.e., internal body forces). Although there is no definitive separation of scales, it is instructive to consider the wave generation as a combination of two separate processes: (1) diabatic forcing from phase changes of water that generates waves with spatial and temporal scales similar to the size and life cycle of convective systems and (2) internal nonlinear body forces associated with oscillating updrafts that generate waves with horizontal wavelengths closer to the size of individual convective updrafts and wave frequencies that approach the Brunt–Väisälä frequency.

Tropospheric Waves

Some of the earliest observations of convective gravity waves were made by glider pilots above boundary layer thermals. Glider pilots exploit the ascending motion within the waves to achieve enhanced lift, and their experiences provided motivation for a number of early studies. Like the examples discussed so far, the thermals penetrate and exert a body force on the inversion at the top of the mixed layer, and the response of the

vertically displaced stable air generates gravity waves. The horizontal wavelengths of the waves are directly related to the size of the thermals, and in conditions without wind shear the spectrum of waves is rich, contains a range of scales, and the spectrum of phase speeds is close to symmetric relative to the mean flow. However, it is common for there to be wind shear at the top of the mixed layer, which can create a coherent wave pattern in the troposphere that appears close to monochromatic.

Changes in wind speed and stability within and above the boundary layer can shape the spectrum due to vertical variations in wave propagation. As described earlier, variations in the wind and stability affect the vertical wavelength of gravity waves, which corresponds to changes in the wave's vertical group velocity. As waves propagate vertically, vertical variations in the environment can lead to wave evanescence ($|\omega - Uk| > N$), which more readily affects the highest-frequency waves and/or those with short horizontal wavelengths (i.e., large k). The waves can also encounter a wave critical level and be dissipated at altitudes where the background wind becomes equal to the phase speed of the waves ($c = \omega/k$). Critical level interactions first affect those waves with slow phase speeds relative to the mean flow that propagate in the same direction as the wind shear vector. The combined

effect of evanescence and critical level interactions is often termed 'wave filtering.'

An example of waves above a convective boundary layer is presented in Figure 3. This idealized case features a mixed layer about 1.5–2 km deep, a background wind that is zero in the mixed layer and equal to 5 m s^{-1} in the free troposphere, and a shear layer that extends between approximately 2 and 3 km altitude. This example illustrates how the shear layer filters the waves: between 2 and 3 km altitude, shorter wavelength waves are present with amplitudes that decay with height. Above the shear layer the wave field is coherent and being dominated by waves with horizontal wavelengths equal to approximately 5 km. The process of wave filtering also reflects some wave energy downward, and those reflected waves help determine the horizontal scale of the thermals in the mixed layer, i.e., the convection and the waves it generates interact with one another via a feedback process. This resonant process is thought to explain some occurrences of regularly spaced boundary layer clouds and shallow cloud bands.

Deep convection can also be influenced by the gravity waves in the troposphere, with populations of deep clouds and the waves they generate interacting via a complicated feedback process. The lower-frequency waves generated by diabatic forcing within the clouds probably make the most important

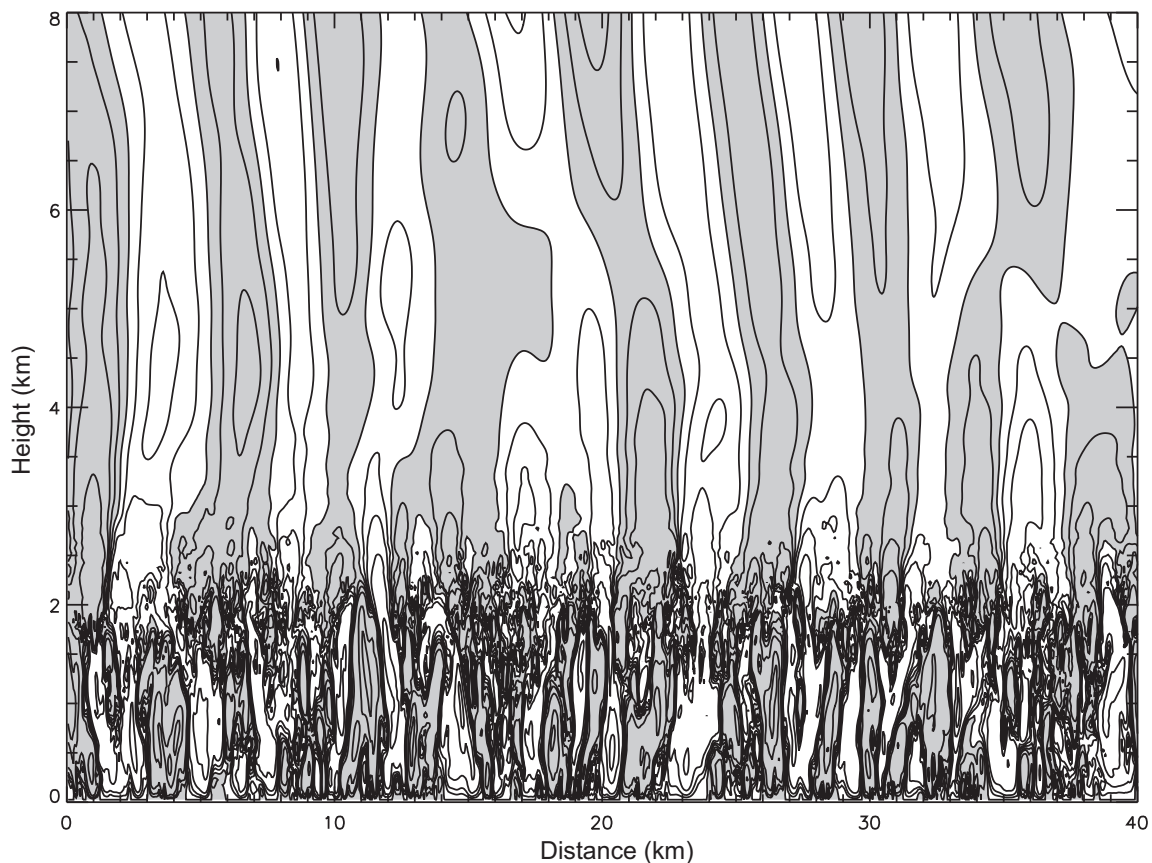


Figure 3 Numerical simulation of gravity waves generated by a (dry) convective boundary layer. Lines are contours of vertical velocity (at 0, ± 0.25 , ± 0.5 , ± 1 , ± 2 , ..., m s^{-1} intervals) and gray shading denotes upward motion. Adapted from Lane, T.P., Clark, T.L., 2002. Gravity waves generated by the dry convective boundary layer: two-dimensional scale selection and boundary layer feedback. *Quarterly Journal of the Royal Meteorological Society* 128, 1543–1570.

contribution to these wave–convection interactions. Precipitating deep convection normally generates a set of gravity waves with different vertical wavelengths that are related to the depth of the convective systems and the vertical distribution of latent heating and cooling. The longest vertical wavelength will be twice the depth of the convection and is related to the depth of the diabatic heating profile. The next longest wavelength is approximately equal to the depth of the convection and related to upper-level heating and low-level evaporative cooling in the system (i.e., the so-called ‘stratiform heating’ profile). When the convection extends to the tropopause, the vertical structure of these waves also corresponds to harmonics of the tropospheric depth (Z_T). These waves are commonly referred to as the $n = 1$ and $n = 2$ waves, with corresponding vertical wavelengths equal to $2Z_T$ and Z_T , respectively. Shorter vertical wavelength waves (and higher-order tropospheric harmonics) are also generated by the complexities in the vertical structure of the clouds and the range of cloud depths in cloud populations.

The $n = 1$ and $n = 2$ waves generated by deep convection have relatively long periods and therefore propagate at an angle close to the horizontal. Long-lived convection generates $n = 1$ and $n = 2$ waves with long periods and appears like horizontally propagating bores; shorter-lived convective systems generate $n = 1$ and $n = 2$ waves with periods of a few hours. The change in stability at the tropopause also acts to partially reflect the waves, and coupled with the relatively long wave periods, can cause $n = 1$ and $n = 2$ waves to extend significant horizontal distances from the convective systems while remaining in the troposphere. These waves also propagate quickly: in the hydrostatic limit ($k \ll m$) the horizontal phase speed relative to the mean flow is equal to N/m and for typical values of the stability and tropopause height the speed is approximately 30–50 and 15–25 m s^{-1} , for $n = 1$ and $n = 2$ waves, respectively. See Figure 4 for a schematic.

The deep tropospheric gravity waves generated by convection communicate the diabatic heating and cooling within the clouds to their environment. In doing so, these waves can either stabilize or destabilize their environment, which can suppress or promote further convection. The first phase of the $n = 1$ wave normally features descending motion that extends over

the entire tropospheric depth. The subsidence stabilizes the atmosphere and suppresses convection in the immediate vicinity. Following the onset of precipitation, the $n = 2$ wave is generated and its first phase is normally ascent in the lower half of the troposphere and descent in the upper half. This low-level ascent destabilizes the atmosphere, either creating a region of preferred convective development around the convective system or initiating new convection when convective inhibition is sufficiently low.

The property of the $n = 1$ and $n = 2$ waves in modifying the environment of convective systems has been used to explain aspects of cloud organization and the upscale growth of convective systems. Convective systems in the tropics are known to be ‘gregarious,’ and the destabilizing effects of gravity waves explain the rapid growth in area of convective activity around mesoscale convective complexes in the tropics. A theory known as wave-conditional instability of the second kind (CISK) attempts to explain the propagation of convective systems as a result of gravity wave-induced convergence. However, in many cases the application of this theory is challenged because the gravity wave speeds and system propagation speeds are different. Nonetheless, other studies have demonstrated that in the tropics gravity waves can organize the cloud population through creating a sequence of convective initiation events, or by modulating the cloud population and determining the spacing of convective elements.

Although many studies have focused on tropical regions, where the convective inhibition is normally low and wave effects are maximized, vertical variations in the background shear and stability in the midlatitudes can lead to ducting of waves, which amplifies their role in convective initiation. Studies have shown success in explaining the propagation of some midlatitude convective systems using a combination of wave ducting and wave-CISK (called ‘ducted wave-CISK’). Forward protruding storm anvils can also define a local wave duct ahead of convective systems; waves in that duct have been shown to cause discrete system propagation. In many ways, convective systems and the tropospheric waves they generate are intertwined and inseparable; and gravity waves are an important part of system morphology and longevity.

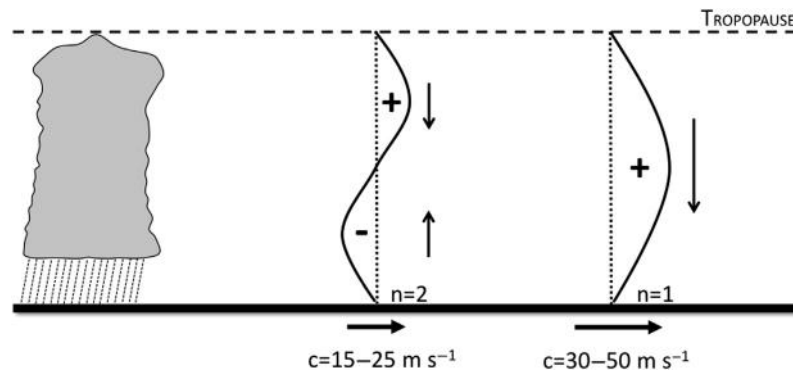


Figure 4 Schematic of the vertical structure and horizontal propagation of the $n = 1$ and $n = 2$ tropospheric waves from a precipitating convective system. Curves represent the vertical structure and sign of the temperature perturbations associated with each wave, and vertical arrows represent the corresponding vertical motion associated with the first phase of each wave. Horizontal arrows represent the horizontal phase speed, c , with the range of typical values also shown.

Stratospheric Waves

The stratosphere is normally about twice as stable as the troposphere (in terms of its Brunt–Väisälä frequency, N) and permits the vertical propagation of the high-frequency waves generated in the troposphere by convection. These highest-frequency waves have phase lines that are oriented closer to the vertical than lower-frequency waves and the vertical component of their group velocity can be similar in magnitude to the horizontal component. Consequently, these waves often occur in the stratospheric air directly above convective systems. Nonetheless, as the waves propagate horizontally as well, the total horizontal area affected by high-frequency gravity waves eventually exceeds the horizontal area of the convective system that generated them.

Figure 5 illustrates the pattern of waves in the stratosphere above a numerically simulated isolated midlatitude convective system. (The tropopause is at approximately 12 km altitude in this case). The wave fronts are approximately circular, with the center of the fronts corresponding to the convective system itself. As is typical, the horizontal wavelength of the highest-frequency waves is smaller than the width of the convective system and corresponds to the size of individual convective updrafts. Gravity waves with horizontal wavelengths as small as 5 km have been observed directly above active convection by aircraft, consistent with those shown here. Organized MCSs comprises multiple regions of active convection and often

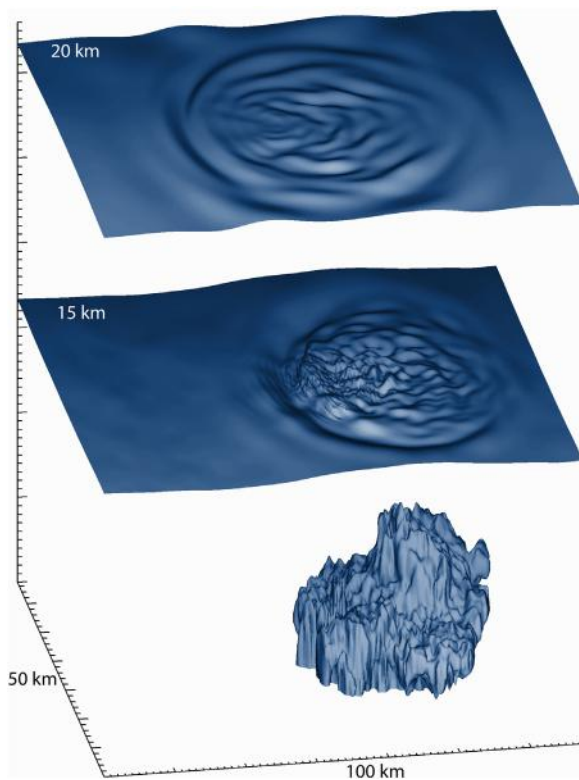


Figure 5 Results from a model simulation of an isolated thunderstorm. The portion of the cloud that penetrates above 10 km altitude is shown by the lower surface. Isentropic surfaces at approximately 15 and 20 km altitudes are also shown.

show separate sets of circular wave fronts; each set of wave fronts is centered on a different region of active convection within the system. The result can be complex interference patterns in the wave field in the stratosphere, with additional complexity arising from source intermittency as convective regions become active and decay with time.

Convective clouds generate a spectrum of waves and lower-frequency waves that are not directly attributable to overshooting updrafts that also reach the stratosphere. For example, $n = 1$ and $n = 2$ waves are only partially reflected at the tropopause and propagate into the stratosphere as well. Typical wave periods above isolated convection can range from 10 min to a few hours, horizontal wavelengths range from ~ 5 to >50 km, and there is a corresponding range of horizontal phase speeds. For example, Figure 6 shows an example spectrum of the vertical flux of horizontal momentum (ρuw , where ρ is the density and u and w are the perturbation horizontal and vertical velocities, respectively) at 20 km altitude above a numerically simulated midlatitude squall line. For upward propagating waves the sign of the momentum flux is equal to the sign of the phase speed relative to the mean flow, and the fluxes in this figure are consistent with upward propagating waves in an environment moving at approximately 10 m s^{-1} . In this example, peaks in the spectrum exist at phase speeds between approximately 10 and 30 m s^{-1} in magnitude relative to the 10 m s^{-1} background flow.

Although the examples presented earlier in Figures 1 and 2 showed symmetric wave patterns, the waves shown in Figures 5 and 6 show that the waves and their spectrum are different for the different propagation directions. This asymmetry occurs for two main reasons: wave generation and wave propagation.

Long-lived convective systems, like squall lines, usually feature convective updrafts that are tilted in the vertical. Studies have shown that although tilted updrafts generate gravity waves propagating in all horizontal directions, the largest amplitude waves generated are normally those that propagate in the direction of the updraft tilt. The tilted structures project onto that part of the wave spectrum more efficiently than other parts, generating an asymmetric spectrum of gravity waves.

The spectrum of waves at a given altitude in the stratosphere is also shaped by changes in wind shear and stability between that altitude and the convective wave source, i.e., wave filtering. Filtering by wind shear can create significant asymmetries in the stratospheric wave spectrum, which is closer to symmetric in the absence of shear. Moreover, as illustrated by Figure 5, the shortest gravity waves are often filtered in the upper troposphere and lower stratosphere, leaving only the longer wavelength waves further aloft.

The motion of convective systems, viz. advection and propagation, complicates interpretation of the dynamics underlying the wave generation. Analyzing the wave spectrum in a frame of reference moving with the convective source (as opposed to the ground-based reference frame, e.g., Figure 6) is sometimes illustrative. However, choosing this reference frame is complicated by the fact that individual convective cells often propagate at a different velocity to the MCSs they are embedded within. Some studies have shown, however, that when the convective updrafts move with the mean wind at their equilibrium level, the peaks in the spectrum occur at similar phase speeds in all directions relative to that reference frame.

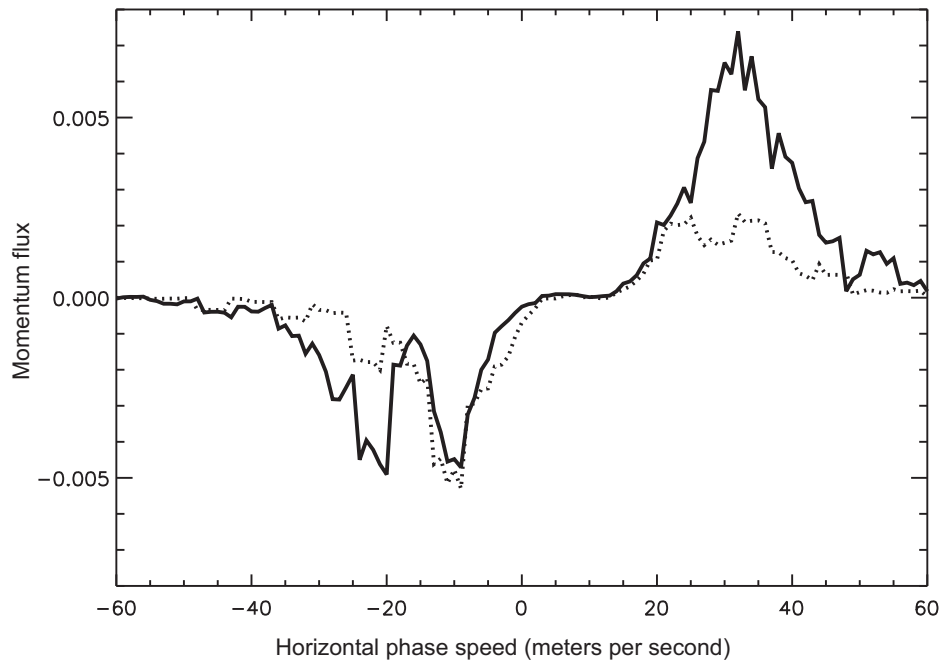


Figure 6 Spectrum of the vertical flux of horizontal momentum (in units of N m^{-2} per 1 m s^{-1} bin) vs horizontal phase speed above a numerically simulated squall line. Results from two simulations are shown that use horizontal grid spacing of 2 km (solid) and 125 m (dotted).

Sheared flows with an ensemble of cloud top heights can broaden the spectrum considerably because the ensemble of wave sources moves at different speeds.

Organized convective systems, viz. systems larger than a single convective cell that persist for multiple convective overturn times, generate gravity waves with horizontal wavelengths beyond the scale of individual clouds. As convective clouds become organized, the spectrum of gravity waves broadens to encompass longer wavelengths and longer periods that are related to the spatial scale of the organized system and its lifetime. For example, regions of organized convective activity in the tropics can contain a rich spectrum of multiscale convective systems and may include individual clouds, MCSs, mesoscale convective complexes, clusters, and superclusters. Each of these systems generates gravity waves leading to a spectrum that contains notable contributions from horizontal wavelengths of 10 km through to scales beyond 1000 km. Waves with these longer wavelengths can be influenced by the Earth's rotation and are then classified as 'inertia-gravity waves.' The intrinsic period of inertia-gravity waves can be tens of hours. (The minimum intrinsic frequency of inertia-gravity waves is the local inertial frequency). Accordingly, the phase lines of inertia-gravity waves are oriented close to the horizontal and can propagate significant horizontal distances from their source. Indeed, inertia-gravity waves observed in the lower stratosphere have been traced back to convection more than 1000 km away.

Tropical cyclones and hurricanes are also notable sources of gravity waves. Patterns of vertical velocity above these systems show a combination of short-scale gravity waves from individual thunderstorms and long-scale (~ 100 km horizontal wavelength) waves that are associated with the evolving eye wall and spiral rainbands. Numerical simulations have found

coherent spiral wave fronts above the convectively active cyclone rainbands. Notable variations in wave amplitude from these separate sources have also been linked to different stages of the cyclone intensification and decay cycle. The rotation of the wind around cyclones affects the wave propagation as well; the subsequent wave filtering, combined with different sectors of the storm featuring more active convection than others, creates considerable spatial variability in the gravity wave activity above cyclones.

Approaches

Numerical models have been utilized to study convective gravity waves, their generation mechanisms and to characterize the wave spectra. The advantage of numerical models is that they provide complete and consistent four-dimensional datasets that can be used to relate aspects of the waves to properties of the convection. Idealized approaches that systematically vary properties of the convection and/or the environment have also made valuable progress toward understanding and parameterizing these waves. However, models have their deficiencies, which pose challenges for studying this problem. Ideally, models should contain high enough resolution that convective structures are well resolved. Models that explicitly resolve moist processes, i.e., convection-permitting or cloud-resolving models, are preferable over models that parameterize convection. However, for studies of large areas and inertia-gravity waves such resolutions are not always achievable. Moreover, even at relatively short model grid spacings of about 1 km, convective updraft structures are not properly resolved, which makes the simulated convection and the waves it generates sensitive to horizontal resolution. For example, Figure 6 shows

a comparison of the wave spectrum from similar model simulations with 2 km and 125 m grid spacing; as the grid spacing is reduced the width of convective updrafts and the horizontal wavelength of the high-frequency waves also reduce. These sensitivities lead to changes in the spectrum and the shorter wavelengths are more likely to be filtered. Finally, the vertical resolution is also important and needs to resolve the vertical structure of the waves. This is particularly important in the stratosphere, where the vertical wavelength is usually smaller than in the troposphere due to the increased stability.

Observations are crucially important for studying convective gravity waves because they provide a true representation of the gravity waves in the atmosphere. Numerous platforms are used to study gravity waves including research aircraft, radiosondes, and remote sensing using radar and satellites.

Radiosondes, in particular, have been used extensively to describe the characteristics of gravity waves and their spatial and temporal variability. As a radiosonde rises, profiles of its three velocity components are determined from the change in position of the sensor with time; the horizontal position of the sensor is normally determined using Global Positioning System and the vertical position is determined from the measured temperature and pressure using the hypsometric equation. An example profile is shown in Figure 7, with the solid lines representing the horizontal atmospheric velocities (left and center) and the balloon's ascent rate (right). Gravity wave velocity perturbations are normally extracted from such a profile by defining them as deviations from smoothly varying profiles that are each defined using polynomials fitted to each profile; example background profiles are shown as dashed lines in Figure 7.

As radiosondes ascend they are advected horizontally as well, making the profiles shown in Figure 7 not true vertical profiles of the atmosphere. However, for waves with horizontal

wavelengths much longer than the horizontal motion of the radiosonde (i.e., inertia-gravity waves), the radiosonde profile can be assumed to approximate a vertical profile and vertical wavelength information can be determined through spectral or wavelet analysis. Moreover, different parts of the wave spectrum affect different variables more strongly than others; the horizontal velocity is more strongly influenced by inertia-gravity waves (i.e., low-frequency waves) and the vertical velocity is more strongly influenced by high-frequency waves. These properties, along with wave theory, are commonly exploited to determine the character of the inertia-gravity waves measured by radiosondes. The ascent rate perturbations provide a representation of the amplitude of the higher-frequency waves and newer techniques have been developed to use the ascent rate and temperature perturbations to determine the wave frequency as well. However, because the higher-frequency gravity waves usually have horizontal wavelengths shorter than the horizontal distance traveled by the radiosonde, the ascent rate profiles cannot be interpreted to represent the vertical wavelength of the gravity waves. Information about the horizontal wavelength and phase speed of the waves is combined in these profiles and is inseparable without additional assumptions or measurements.

Remote sensing applications like radar and satellite have also been used for studying convectively generated gravity waves. Vertically pointing radars have had success in measuring the frequency, vertical wavelength, and amplitude of gravity waves for many parts of the wave spectrum. Such platforms, however, have difficulty in inferring the horizontal wavelength of the gravity waves and can suffer from gaps in coverage due to variations in the atmospheric properties and finite radar range. Satellites offer much promise for determining statistics of gravity waves generated by convection through radiance

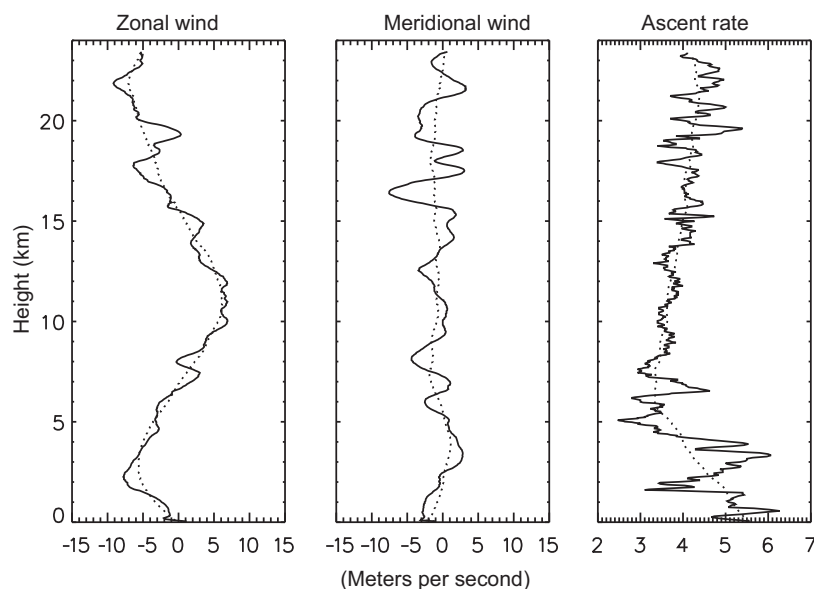


Figure 7 Gravity waves observed using radiosonde data (at 0300 UTC 27 November 1995, near Darwin Australia). Solid lines show the observed horizontal wind components (left and center) and the radiosonde ascent rate (right). The dotted lines represent background profiles used to define the wave perturbations. Adapted from Lane, T.P., Reeder, M.J., Guest, F.M., 2003. Convectively generated gravity waves observed from radiosonde data taken during MCTEX. *Quarterly Journal of the Royal Meteorological Society* 129, 1731–1740.

measurements that are used to infer temperature perturbations. Satellites have demonstrated their effectiveness in identifying large-amplitude wave events and their spatial structures, as well as the global statistics of wave activity. Like all observational platforms, however, satellites cannot measure the entire wave spectrum, and are restricted by the footprint of the sensor and the vertical structure of the weighting functions assigned to specific sensing channels.

Although observations do not suffer from the same approximations and sensitivities that models do, each measurement platform can usually only observe a part of the wave spectrum. Thus, approaches with multiple platforms or combined observational/modeling studies are normally required to fully characterize the wave properties. In addition, purely theoretical approaches using mathematical theory have been responsible for many advances in the understanding of gravity waves, their propagation, and generation.

Impacts

Gravity waves generated by convection have important impacts on the surrounding atmosphere. As described above, convective gravity waves in the troposphere can influence the morphology and longevity of convective systems and thereby play key roles in the dynamics of moist and dry convection.

The upper troposphere/lower stratosphere is an important region for convective gravity waves. The strong change in static stability at the tropopause contributes to the generation of waves by overshooting deep convection. The wind shear is often strong near the tropopause, and therefore the upper troposphere/lower stratosphere is normally a region of significant wave filtering. In particular, gravity wave breaking through critical level interactions near the tropopause generates turbulence and mixing as part of the wave dissipation process. The turbulence has been shown to transport water vapor and other constituents across isentropic surfaces, contributing to stratosphere–troposphere exchange. The turbulence associated with breaking convective gravity waves is also an important aviation hazard and is responsible for many commercial aircraft turbulence encounters.

Momentum transport associated with convective gravity waves also plays an important role in the dynamics of the middle atmosphere. For example, convective gravity waves contribute to the formation of the stratospheric quasi-biennial oscillation (QBO); it has been estimated that up to 80% of the forcing of the QBO originates from convectively generated waves (including convectively generated inertia-gravity waves). The importance of these waves, and the fact that most of the spectrum is unresolvable in global climate models, has motivated the development of parameterizations of the convective gravity wave source. These parameterizations have been informed by theoretical, observational, and numerical modeling studies of convective wave generation. The most common approach is to link the wave momentum flux to properties of the diabatic heating in the troposphere using

linear theory. Nonetheless, parameterization of these processes remains challenging because tropospheric convection is also parameterized, and the complicated nonlinear processes responsible for convective wave generation and dissipation are still not fully understood.

See also: **Aviation Meteorology:** Aviation Weather Hazards; Clear Air Turbulence. **Boundary Layer (Atmospheric) and Air Pollution:** Coherent Structures; Convective Boundary Layer. **Dynamical Meteorology:** Wave-CISK. **Gravity Waves:** Buoyancy and Buoyancy Waves: Theory; Overview. **Mesoscale Meteorology:** Cloud and Precipitation Bands; Convective Storms: Overview; Mesoscale Convective Systems. **Middle Atmosphere:** Quasi-Biennial Oscillation. **Numerical Models:** Convective Storm Modeling; Parameterization of Physical Processes: Gravity Wave Fluxes. **Observations Platforms:** Radiosondes. **Stratosphere/Troposphere Exchange and Structure:** Local Processes; Tropopause.

Further Reading

- Ern, M., Preusse, P., 2012. Gravity wave momentum flux spectra observed from satellite in the summertime subtropics: implications for global modeling. *Geophysical Research Letters* 39, L15810. <http://dx.doi.org/10.1029/2012GL052659>.
- Fritts, D.C., Alexander, M.J., 2003. Gravity-wave dynamics and effects in the middle atmosphere. *Reviews in Geophysics* 41, 1003. <http://dx.doi.org/10.1029/2001RG000106>.
- Karoly, D.J., Roff, G.L., Reeder, M.J., 1996. Gravity wave activity associated with tropical convection detected in TOGA COARE sounding data. *Geophysical Research Letters* 23, 261–264. <http://dx.doi.org/10.1029/96GL00023>.
- Kim, S.-Y., Chun, H.-Y., 2010. Stratospheric gravity waves generated by Typhoon Saomai (2006): numerical modeling in a moving frame following the typhoon. *Journal of Atmospheric Sciences* 67, 3617–3636.
- Kim, Y.-J., Eckermann, S.D., Chun, H.-Y., 2003. An overview of past, present and future of gravity-wave drag parametrization for numerical climate and weather prediction models. *Atmosphere-Ocean* 41, 65–98.
- Kuetner, J.P., Hildebrand, P.A., Clark, T.L., 1987. Convection waves: observations of gravity wave systems over convectively active boundary layers. *Quarterly Journal of the Royal Meteorological Society* 113, 445–467.
- Lane, T.P., 2008. The vortical response to penetrative convection and the associated gravity-wave generation. *Atmospheric Science Letters* 9, 103–110.
- Lane, T.P., Clark, T.L., 2002. Gravity waves generated by the dry convective boundary layer: two-dimensional scale selection and boundary layer feedback. *Quarterly Journal of the Royal Meteorological Society* 128, 1543–1570.
- Lane, T.P., Moncrieff, M.W., 2008. Stratospheric gravity waves generated by multi-scale tropical convection. *Journal of Atmospheric Sciences* 65, 2598–2614.
- Lane, T.P., Sharman, R.D., 2006. Gravity wave breaking, secondary wave generation, and mixing above deep convection in a three-dimensional cloud model. *Geophysical Research Letters* 33, L23813. <http://dx.doi.org/10.1029/2006GL027988>.
- Lane, T.P., Zhang, F., 2011. Coupling between gravity waves and tropical convection at mesoscales. *Journal of Atmospheric Sciences* 68, 2582–2598.
- Mapes, B.E., 1993. Gregarious tropical convection. *Journal of Atmospheric Sciences* 50, 2026–2037.
- Mowbray, D.E., Rarity, B.S.H., 1967. A theoretical investigation of the phase configuration of internal waves of small amplitude in a density stratified liquid. *Journal of Fluid Mechanics* 28, 1–16.
- Nappo, C.J., 2002. *An Introduction to Atmospheric Gravity Waves*. Academic Press, p. 276.
- Sutherland, B.R., 2010. *Internal Gravity Waves*. Cambridge University Press, p. 377.

HYDROLOGY, FLOODS AND DROUGHTS

Contents

Overview

Deserts and Desertification

Drought

Flooding

Groundwater and Surface Water

Modeling and Prediction

Palmer Drought Severity Index

Soil Moisture

Overview

RC Bales, University of Arizona, Tucson, AZ, USA

© 2003 Elsevier Ltd. All rights reserved.

This article is a reproduced from the previous edition, volume 3, pp 968–973, © 2003, Elsevier Ltd.

Introduction

Hydrology is the science that encompasses the study of water on the Earth's surface and beneath the surface of the Earth, the occurrence and movement of water, the physical and chemical properties of water, and its relationship with the living and material components of the environment. Ultimately, many hydrologic questions involve the transport of solutes, nutrients, energy, sediment, or contaminants, as well as the fluxes of water itself.

As a science, hydrology has both basic and applied aspects. The first relates to questions about the Earth system, and specifically about the role of water in natural processes, particularly as related to the Earth's biosphere. The second relates to using scientific knowledge to provide a sound basis for wise usage of water resources. The development of hydrologic science in recent years is based on both of these aspects, which are equally important and intimately linked.

Water is central to most natural processes. Water weathers, then transports sediment and solutes to lakes and oceans, thereby shaping the landscape. The land-based part of the Earth's water cycle is important for transporting carbon from the continents to the ocean. The high capacity of water for storing thermal energy and the large amount of heat required to change between solid, liquid, and vapor forms of water strongly influence the global energy balance. The distribution of atmospheric water and its regulation by oceanic and land-surface processes make it a central aspect of climate. Water vapor is the most important greenhouse gas. In short, life depends on water.

Hydrologic science has an important place in the field of water resources, especially freshwater resources, which are the subject of intense concern and study. In arid and semi-arid regions, the fair allocation and wise use of water are significant societal challenges, affecting relations between nations, states, cities, and individual users. As a global resource, water appears abundant. Locally and regionally it is often taken for granted. However, the twentieth century has witnessed a tremendous growth in the use of water, resulting in limits on both its availability, due to human exploitation, and its quality, owing to contamination.

Water Cycle

A fundamental concept of hydrology is the hydrologic cycle, which can be described at many different scales of space and time. At the global scale, the hydrologic cycle is the endless recirculatory process linking water in the atmosphere, on the continents, and in the oceans. We can think of this recirculatory process in terms of reservoirs or compartments that store water (e.g., oceans, atmosphere, glaciers, ice sheets, ground water) and the movement of water between them. Movement of water from one compartment to another can occur in any of the three phases. For example, the movement of water between the land surface and the atmosphere occurs in the vapor phase (evaporation and condensation), liquid phase (rain), and solid phase (snowfall).

Solar energy and gravity are the main forces driving the hydrologic cycle. The dynamic processes of water vapor

formation and transport of vapor and liquid in the atmosphere are driven largely by solar energy. Precipitation and the flow of water on and beneath the Earth's surface are driven primarily by gravity. Within partially dry soil, gravitational pressure and capillary forces are responsible for the movement of water.

The hydrologic cycle can be considered to start anywhere, but it is often convenient to consider atmospheric water first. The dominant hydrologic process involving atmospheric water is precipitation of water to the land surface. Condensation of water from the atmosphere to the land surface (e.g., dew, frost) and fog deposition can be important locally, in the absence of precipitation. Precipitation that reaches the land surface as snow or ice can be retained on vegetation and subsequently evaporate or fall to the ground, accumulate in seasonal snow-packs, and later melt or evaporate or accumulate in glaciers, ice caps, or ice sheets. Water is also lost from glaciers and ice caps by evaporation and melting; pieces of ice can also break off from the edges of glaciers and ice sheets (this is known as calving) and be returned directly to the ocean, in the form of icebergs. A portion of rain or snowfall can be retained temporarily on vegetation surfaces, and subsequently evaporate or fall to the ground. Rain or snowmelt can also collect in surface depressions, enter into the soil (infiltration), or flow as runoff over the land surface into small rivulets and ultimately into larger streams, lakes, and rivers. Water that infiltrates into the soil can also follow different paths. Some returns to the atmosphere by evaporating from the soil or being transpired by plants (transpiration), the sum of which is termed evapotranspiration. The remaining water continues to move downward through the soil and recharges the saturated portion of the subsurface, becoming groundwater. Ground water discharges into streams and rivers, or directly to the ocean. Water evaporates from the surface of the oceans and thereby replenishes the water in the atmosphere.

Much effort in hydrology goes toward estimating the amounts of water in the various storage compartments and the magnitudes of the various flows to and from these compartments at local, regional, and global scales. Nearly 97% of all water on the Earth is stored in the oceans, while only about 0.001% is stored in the atmosphere. Fresh water accounts for about 2.5% of the total storage, 70% of which is contained in the two polar ice sheets and 30% is found in ground water. Only about 0.4% is found in glaciers and ice caps. The fresh water in lakes, streams, rivers, and marshes represents only 0.26% of all fresh water and 0.008% of all water on Earth. That is, if the Earth's ocean were represented as a 16 l (4 gallon) bucket, the fresh water fraction would be equal to a little over 1 ml ($\frac{1}{4}$ teaspoonful).

Another useful concept for enumerating reservoirs and the flows of water through them is residence time, which is a measure of how long, on average, a molecule of water spends in that reservoir before moving on to another reservoir of the hydrologic cycle. For a system at steady state, i.e. with inflow and outflow the same, residence time is equal to the size of the reservoir (e.g., in m^3) divided by the flow through the system (e.g., in $\text{m}^3 \text{ yr}^{-1}$). Water in the oceans has a residence time approaching 3000 years, less than half of the residence time for ice sheets, while in the atmosphere it has a residence time of only 0.02 years or about 8 days; the residence time in rivers is 0.05 years or about 17 days.

Most of hydrology deals with scales smaller than global; however, the same concepts of fluxes and reservoirs apply. A catchment is an area in which water falling on or flowing across the land surface drains into a particular stream or river and flows ultimately through a single point or outlet. Thus a catchment is defined relative to a specific location and includes all of the land area that drains to that point; i.e., it can be considered to catch the water that flows past that point. Clearly, then, any number of catchments can be defined for a particular river (corresponding to any location along the river). Some special locations of interest for defining catchments are points corresponding to dams, stream gauges, cities, and a river's mouth. Ground water reservoirs (aquifers), on the other hand, are defined by subsurface geologic structure. Aquifers, or water-bearing formations, are bounded by material of low permeability, i.e., material with a very small ability to transmit water.

Precipitation

Precipitation is the deposition of liquid water droplets and ice particles that have formed in the atmosphere and grown to a size sufficient to fall to the Earth's surface by gravity. Precipitation is classified according to the phase it is in when it reaches the ground, i.e., solid (snow, sleet, and hail) or liquid (rain and freezing rain). Other deposition processes (e.g., direct deposition of dew and fog), though generally small, can however be important in terms of chemical fluxes (e.g., acidic fog).

Most of the precipitation falling over North America originates from the bordering oceans, even in the interior of the continent. However, over the Amazon basin, a significant fraction of the precipitation is derived from within-basin evapotranspiration. In other continental basins, local evapotranspiration does have some influence on local precipitation, but most of the precipitated water must be transported significant distances across the continents from the oceans.

Average precipitation over the continents is extremely variable geographically, reflecting the influence of a number of important physiographic factors. First, precipitation increases with elevation owing to orographic cooling. Second, precipitation is typically higher on windward than on leeward sides of mountain ranges. Third, precipitation tends to drop off as air masses move further inland, away from the ocean. Fourth, the temperature differences between adjacent land and ocean influence moisture transport. Fifth, prevailing wind direction has local to regional effects.

Global average precipitation is about 1000 mm yr^{-1} . In the continental US, average annual precipitation ranges from about 40 mm yr^{-1} at Death Valley, California (in the Mojave Desert), to over 3000 mm yr^{-1} in parts of the Pacific Northwest. In the Atacama Desert of northern Chile rainfall is infrequent, averaging under 1 mm yr^{-1} .

Rainfall and snowfall are measured at a point by collectors of very simple construction. Essentially, any receptacle with a reasonable opening can serve to estimate the precipitation per unit area. In the US, the standard gauge has a 20 cm diameter opening. Wind is probably the single most important factor in rain gauge accuracy. Updrafts resulting from air moving up and around the instrument reduce the catch, which has led to the

development of elaborate shields to mitigate wind effects and detailed correction procedures.

In a few areas, radar is also used for precipitation measurement, although rain gauges are generally required for calibration. Radar emits electromagnetic energy in narrow bands, which upon hitting precipitation is partially absorbed, scattered, and reflected. Some of the reflected energy returns to the transmitter, and the attenuation of the signal indicates the intensity of rainfall. In principle, the same method could be used for snowfall.

Snow accumulation at a point is more accurately measured, either in situ by snow pillows equipped with pressure transducers or manually. Snow accumulation is reported in terms of equivalent water depth rather than snow depth, the proportionality factor being the density. Snow accumulation can also be measured from aircraft in terms of the extinction of gamma or other types of radiation. Satellites are used for estimating snowpack depth and areal coverage. A combination of visible and infrared signals are used to extract snow-covered area. Over flat terrain, passive microwave signals indicate snow water equivalent. Because much snow falls in heterogeneous terrain and is thus not uniformly distributed, accurate measurements of snow water equivalent remain a major challenge. Snow measurement networks in the western US consist of index sites, for which correlations with streamflow have been developed. They do a relatively poor job of providing direct estimates of snow volume.

Snowmelt plays a major role in the hydrology of midlatitudes, as many rivers originate in mountains where snow dominates the precipitation. In contrast to rainfall, snowfall has a delayed effect on river flow and hydrology. Accumulation occurring during winter months becomes all-important during spring runoff, which provides much of the streamflow, ground water recharge, and soil moisture for wide areas of the Earth. Melt waters can also cause serious floods, particularly when compounded with spring rainfall.

Catchment

Much effort in hydrology focuses on water balances at the scale of a catchment, be it a 1 km² headwater basin or a 100 000 km² regional basin. Hydrologists often think of a catchment as functioning like a group of reservoirs that store and release water. Precipitation and snowmelt are the inputs to the catchment reservoir, and evapotranspiration and runoff are the outputs. Integrating the various processes that affect runoff involves quantitative partitioning of the different fluxes and reservoirs into downstream fluxes and reservoirs. For example, a fraction of precipitation is intercepted by vegetation, with the remainder falling to the ground. Intercepted rain and snow then either evaporates or eventually falls to the ground. Snow on the ground partitions between snowmelt and sublimation. Rainfall or snowmelt reaching the ground either infiltrates or runs off, depending on the infiltration capacity of the soil. A portion may flow downstream in the subsurface and reemerge at a lower point in the catchment. Much infiltrating water may eventually be transpired by plants or evaporated, though some will travel downward sufficiently far to enter, or recharge, a ground water aquifer. Of

all these quantities, surface runoff in streams and rivers, or discharge, is most often measured.

Discharge is estimated by continuously measuring the stage, or height of a stream at a point using a pressure transducer or a mechanical float. This point measurement is then related to the volumetric discharge through a calibration, or rating, curve, which is developed by measuring the volumetric discharge at the gauged point for many different stages. A time series graph of discharge is commonly referred to as a hydrograph.

A streamflow hydrograph is often defined in terms of two components, quickflow and baseflow. Separation of a hydrograph into two components suggests that water is being routed through two different storage reservoirs. During and after rainfall and snowmelt events, water moves through the catchment into the stream channel and the discharge increases (quickflow). The resulting peak in the hydrograph is generally defined as a flood, regardless of whether the river actually leaves its banks and causes damage. Background discharge between floods (baseflow) is supplied by inflow of ground water, which may lag the occurrence of precipitation by days, weeks, or even years.

Water flowing across the ground surface is termed overland flow. For this to occur, water must accumulate at the surface rather than infiltrate into the soil. This happens for three main reasons: (1) the catchment surface may be nearly impermeable owing to the presence of exposed bedrock or pavement; (2) the instantaneous rate of infiltration through the pervious surface may be exceeded by the instantaneous rate of rainfall (or snowmelt) onto the catchment surface, causing ponding of water at the surface, and (3) the catchment soil upon which the rainfall is precipitated may be saturated to the soil surface, causing ponding because the precipitated water cannot infiltrate into an already saturated soil. Overland flow in catchments is one of the most rapid paths that water can follow to the stream channel.

Water that has infiltrated the soil surface continues to be influenced by gravity, so that it percolates downward through the soil profile. In general, the ability of the soils and rocks of a catchment to conduct water (hydraulic conductivity) decreases with depth; water percolating downward has thus been observed to back up, causing local areas of saturation in the soil. In these instances, water may move laterally toward a stream by a process known as shallow subsurface stormflow. Some of the water in subsurface stormflow moves at a relatively slow pace through the soil and contributes to the baseflow of streams, particularly during wetter winter and spring periods. Subsurface stormflow also may occur quite rapidly along preferred flow pathways or macropores (e.g., soil cracks, old animal burrows, and decayed root channels).

Vadose

In most areas, the water table is some distance below the ground surface. Between the ground surface and the water table is a region in which the pore spaces of the rock or soil may be partly filled with air and partly with water. This region is referred to as the unsaturated zone or vadose zone, and water in this zone is referred to as soil moisture. Hydrologists want to be able to describe the amount of water in the unsaturated zone and

fluxes through the zone for two main reasons. First, ground water recharge occurs through this zone. Second, most terrestrial plants extract water from the vadose zone. Plants wilt when soils become too dry because the tension forces holding the water in the soil are too great to allow the plants access to the water. Related to the water balance of plants is the practice of irrigation in agriculture, which accounts for about two-thirds of global water use. Understanding the movement of soil water, its uptake by plants, and its loss through evapotranspiration and recharge to the groundwater system is essential in this regard.

Hydrologists have traditionally recognized three divisions within the unsaturated zone: the capillary fringe, the intermediate belt, and the belt of soil water. The capillary fringe is a zone in which the pressure is less than atmospheric, overlying the zone of saturation and containing capillary interstices, some or all of which are filled with water that is continuous with the water in the zone of saturation but is held above that zone by capillarity forces acting against gravity. That is, the capillary fringe is a saturated zone above the water table where water is affected by capillary forces. Above that is a zone of soil water from which water is discharged to the atmosphere by the action of plants or by evaporation. For the most part, plants extract water from a portion of the soil near the surface (the 'root zone'). Depending on the depth of the vadose zone and the plant, roots may lie only in the upper part of the vadose zone, or extend into the water table. For example, most grasses have roots extending only a few centimeters to tens of centimeters, whereas some trees in semi-arid regions have roots that extend through vadose zones that are tens of meters thick and reach the water table. The volumetric moisture content (volume of water per bulk volume of soil sample) in the capillary fringe is the saturation value. In other words, the pores are completely filled with water. As water drains or is withdrawn by plants, soil moisture content decreases from saturation to a fairly constant value, termed the field capacity.

Rates of removal of water from the unsaturated zone by evapotranspiration are controlled by a number of factors, including the wetness of the soil itself. If a vegetated surface is supplied with plenty of water (e.g., a well-watered lawn), evapotranspiration will be controlled by atmospheric conditions, e.g. solar radiation, wind speed, and humidity. That is, evapotranspiration will proceed at the maximum rate (potential evapotranspiration). As a soil dries, evapotranspiration will proceed at the potential rate for some time, but ultimately the rate will drop. As water is pulled from the soil near a plant root, the moisture content in the soil surrounding the root decreases. In order to maintain a steady flow of water to the plant root, the plant must exert ever greater suction (ever greater negative capillary pressure heads). At some point, the plant cannot sustain this battle with a drying soil and the transpiration rate falls below the potential rate.

Most plants have openings (stomata) on their leaves to allow them to take up carbon dioxide from the atmosphere. When the stomata are open, plants transpire water. Unlike evaporation, transpiration is not controlled solely by physical conditions because plants regulate the rate at which water is released in transpiration in a manner that varies by plant type and ecological conditioning. Of the water taken up by plant roots, most is transpired through the stomata. A few percent is concerted to biomass through photosynthesis. Hence, to first

order, the water taken up by the roots is converted to vapor and lost to the atmosphere. The degree of stomatal restriction varies considerably across plant species and even throughout the year for a given species.

Ground Water

The largest accessible fresh water reservoir for human exploitation is ground water, or water that is present in the fractures and interstitial spaces in subsurface geologic materials. In contrast to the vadose zone, in ground water the void spaces are completely filled with water. An aquifer is a saturated geological formation that contains and transmits significant quantities of water under normal field conditions. 'Significant' is a vague term, but the implication is that aquifers are formations that can be used for water supply. Many aquifers are unconsolidated materials, mainly gravel and sand. Examples of this type of aquifer include those in coastal plains and intermontane valleys. Limestones, partially cemented sandstones and conglomerates, and permeable volcanic and igneous rocks are also important as aquifers.

An aquitard is a formation of relatively low permeability, and includes both formations that contain water but do not transmit significant quantities (e.g., clays and shales) and formations that neither contain nor transmit significant quantities of water (e.g., unfractured crystalline rocks).

Aquifers are classified according to hydraulic conditions as well as type of material. Ground water by definition refers to water in the saturated zone of the subsurface; one type of aquifer is an unconfined or water table aquifer. Deeper in the soil profile, saturated conditions prevail (the saturated zone). The water table is defined as a surface of zero (gauge) pressure within the subsurface, and separates the saturated and unsaturated zones. Water will flow into an excavation or well up to this level; the water table is equivalent to a free surface. An aquifer with the water table as the bounding surface at its top is an unconfined aquifer.

The second type of aquifer is a confined or artesian one. This is found when permeable material (the aquifer) is overlain by relatively impermeable material. The water in a confined aquifer is under pressure and, in a well penetrating the aquifer, will rise above the top of the aquifer. The height to which water rises in a well defines the piezometric surface, or pressure of water in the aquifer.

In areas where the water table is sufficiently close to the ground surface, ground water levels are influenced directly by transpiration. During the day, when transpiration is high, water movement is upward from the water table and the level declines. At night, transpiration is reduced, groundwater flows laterally from locations relatively unaffected by direct transpiration effects, and the water table recovers.

Recharge to aquifers can occur from direct infiltration of rainfall or snowmelt past the root zone and to the water table. Water can also seep from surface water bodies, such as rivers, ponds and lakes, into the ground. Artificial recharge (recharge induced by activities of people as opposed to that which occurs naturally) can be implemented by introducing water into recharge wells or by routing water into infiltration basins in permeable material.

The specific yield is a hydrologic parameter that determines the response of the water table to changes in inputs and outputs. In the case of an increase in evapotranspiration, the change in water table level may be fairly uniform over a given area, although variations will occur due to the lateral movement of groundwater and spatial variations in evaporation rate and vegetation. Pumping a well has a different level in the pumping well, or in observation wells nearby, is referred to as drawdown. The amount of this drawdown will decrease as one moves away from the pumping well, and the pattern produced is called a cone of depression because of its characteristic shape.

Integration

The need for an integrated understanding of water cycles and linked chemical cycles is critical for a number of reasons. For example, a quantitative understanding of how ground water recharge changes with precipitation in a variable or changed climate requires an understanding of how catchment processes respond to translate precipitation into recharge. Understanding how evapotranspiration and ecosystem functioning respond to change in land use depends on how water and nutrient fluxes

change. Hydrologic science is at the center of many pressing issues in other natural sciences.

See also: **Climate and Climate Change:** Climate Variability: Seasonal and Interannual Variability. **Dynamical Meteorology:** Hydraulic Flow. **Hydrology, Floods and Droughts:** Modeling and Prediction; Soil Moisture. **Land-Atmosphere Interactions:** Overview.

Further Reading

- Bras, R.L., 1990. *Hydrology – An Introduction to Hydrologic Science*. Addison-Wesley, Reading MA.
- Fetter, C.W., 1988. *Applied Hydrogeology*. Macmillan, New York.
- Freeze, R.A., Cherry, J.A. (Eds.), 1979. *Groundwater*. Prentice-Hall, Englewood Cliffs, NJ.
- Hornberger, G.M., Raffensperger, J.P., Wiberg, P.L., Eshleman, K.N. (Eds.), 1988. *Elements of Physical Hydrology*. Johns Hopkins University Press, Baltimore, MD.
- Maidment, D.R., 1992. *Handbook of Hydrology*. McGraw-Hill, New York.
- National Research Council, 1991. *Opportunities in Hydrologic Sciences*. National Academy Press, Washington, DC.

Deserts and Desertification

VP Tchakerian, Texas A&M University, College Station, TX, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by G Wang, G S Jenkins, volume 2, pp 633–640, © 2003, Elsevier Ltd.

Synopsis

Deserts cover about 35% of the land surface area of the world and are typically located between and astride the Tropic of Cancer (35° N) and the Tropic of Capricorn (35° S). About 20% of the world's population resides in this geographic region. The location of this global arid zone is primarily the result of the semi-permanent high-pressure cells that dominate this region along with such other factors as rain-shadow effects, continentality or remoteness from moisture sources, upwelling of cold currents that suppress the development of precipitation, and the nature and direction of the prevailing winds. Most deserts exhibit a combination of the above factors. Desertification refers to land degradation in the global arid zone owing to a series of complex climatic, biophysical, and anthropogenic factors and became a major global topic during the severe drought of the Sahel region in northern Africa in the 1970s. Desertification has been erroneously represented as the irreversible march of the desert and is now believed to result primarily from the degradation of arid ecosystems largely because of human induced factors along with natural climatic oscillations and drought cycles.

Introduction

Deserts (arid lands/drylands) constitute about 35% of the land areas of the world and are typically characterized by rainfall scarcity, higher temperatures and evapotranspiration, lower humidity, and a general paucity of vegetation cover. A unique combination of atmospheric, geologic, and geomorphic conditions is responsible for the formation of deserts primarily between the Tropics of Cancer and Capricorn (Figure 1). Compared to humid lands, the relative importance of desert atmospheric and geomorphic processes and the magnitude and frequency of their operation is rather distinctive. Arid lands

comprise the most widespread terrestrial biome on Earth and are home to over 20% of the world's people. In the twentieth century, a combination of natural and anthropogenic factors gave rise to the concept of desertification – first as a rather simplified 'march of the desert' into bordering semiarid regions, and then recently as a more complex phenomenon that includes both natural and anthropogenic causes, the latter primarily the consequence of increased population numbers in the semiarid regions of the world. The distinctive natural environment in arid lands, coupled with the growing human populations in drylands, is one of the primary reasons behind the recent upsurge in the global study of deserts.

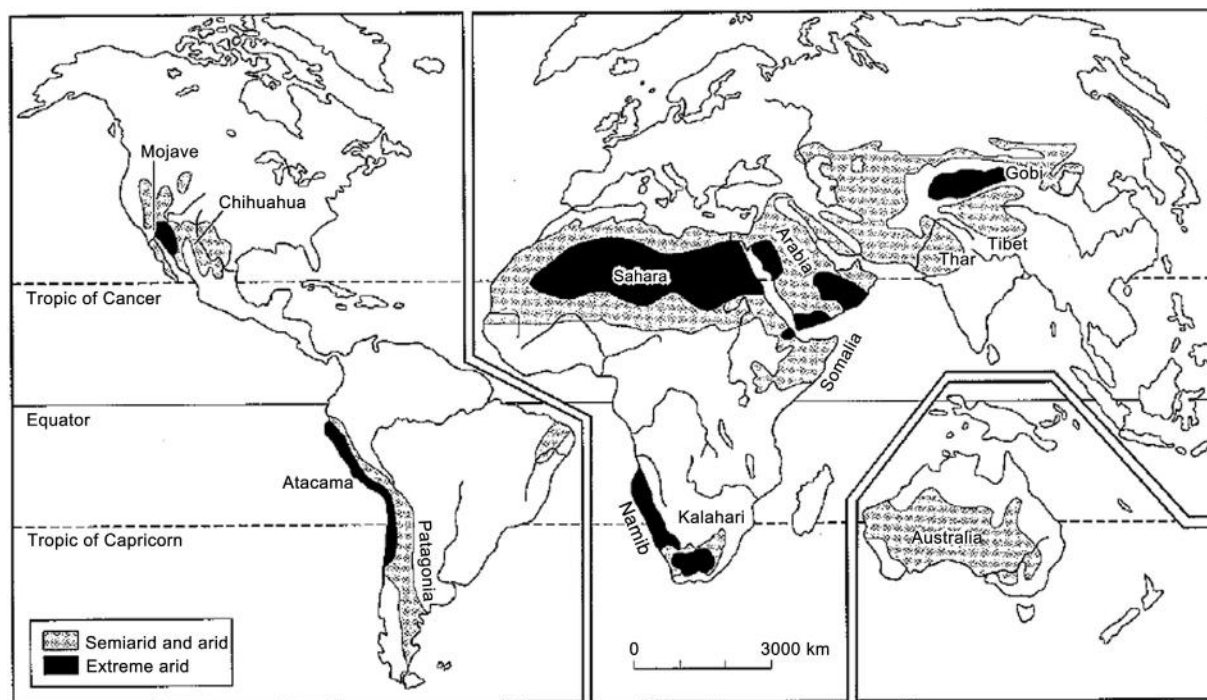


Figure 1 Global distribution of deserts. Reproduced from Goudie, A.S., 2002. *Great Warm Deserts of the World*. Oxford University Press, Oxford.

Deserts of the World

Drylands of the world can be either classified as arid, based on climate or desert, based on surface characteristics (landforms, vegetation, etc.). For the purpose of this review, we will use the climatic classification adopted by UNESCO in 1979. Deserts thus can be classified as (1) semiarid (precipitation less than 500 mm); (2) arid (precipitation less than 250 mm); and (3) hyperarid (precipitation less than 25 mm). Approximately 16% of the global arid zone is semiarid, 15% arid, and 4% hyperarid. The division between the three is somewhat arbitrary and based on limited climatic data owing to the fact that there are large variations in annual precipitation regimes and that natural and anthropogenic activities have shifted those boundaries. Africa contains the greatest proportion of the global arid zone at about 37%, while Australia is the most arid continent, with about 75% of the land area being arid or semiarid. Contrary to public opinion, sand dunes and sand seas (ergs) are not the dominant landform type in deserts, although dunes cover about 40% of the surface area of Australia (most are stabilized, relict dunes), constituting 38% of the world's dune fields. A recent landform map of North Africa (mostly the Sahara Desert) produced through moderate resolution imaging spectroradiometer, indicated that the two most dominant landform types were stone pavements (hamada, serir, reg, desert pavement) at about 25% cover, followed by sand seas (ergs) at about 20%. On the other hand, the North American arid zone includes extensive areas of alluvial fans, mountains, desert flats, playas, and arroyos, while sand dunes and sheets constitute less than 5% of the arid zone.

There are a number of reasons for the global arid zone. The majority of the world's deserts are subtropical in distribution, covering about 20% of the Earth's land surface and located between the Tropic of Cancer (23.5° N) and the Tropic of Capricorn (23.5° S). The strong subsidence of air in these regions is largely the result of the descending branch of the Hadley cell, which causes air at the surface to be hot and dry. These are the locations for some of the most famous deserts in the world such as the Sahara Desert, the Rub'al Khali in the Arabian Peninsula, the Thar desert of India and Pakistan, the Kalahari Desert in southern Africa, the Sonoran and Chihuahuan Deserts of North America and the Australian deserts – Simpson, Gibson, Great Sandy, Tanami, and the Great Victoria (Figure 1). These deserts typically exhibit very high insolation values, very low humidity, very high evapotranspiration rates and extreme spatially and temporally variable precipitation regimes, with occasional heavy, short-lived thunderstorms, largely associated with the seasonal movements of the intertropical convergence zone (ITCZ). Another type of desert is formed when moisture is prevented from reaching continental interior locations because of either the distance from water bodies or the presence of mountain ranges. Examples of continental interior deserts include most of the midlatitude Asian deserts such as the Taklimakan and the Gobi Desert in China, the KaraKum deserts in Kazakhstan and Uzbekistan, and parts of the Great Basin and the Colorado Plateau in the United States. For example, moist air masses that originate around Scotland (wet) can move over Poland (moist) and will most likely be dry by the time they cross the Volga River on their way to central Asia. Other deserts owe their

existence because of their location on the lee side of major topographic barriers such as the Mojave Desert being on the rain shadow of the Sierra Nevada Mountains and the Transverse Ranges of southern California. Parts of the Great Basin and the Colorado Plateau in the United States are also considered rain-shadow deserts as well as the Patagonian Desert in Argentina. The north–south orientation of the Great Dividing Range in eastern Australia also contributes to the lee side aridity of the Simpson Desert, as easterly trade winds are prevented from bringing their moisture past the Great Divide. Deserts are also formed on the western coastline regions of continents, where the upwelling of cold, ocean currents, suppresses any precipitation potential. Other climatic conditions include low sea-surface evaporation, high atmospheric humidity, low annual temperature ranges, and extremely low rainfall amounts. Warm air as it moves over these cool waters is chilled/condensed (only a thin layer is affected) forming mostly fog and dew, which for many years is the only source of moisture for plants. Examples include the Namib Desert, located mostly in Namibia (Benguela Current), Baja Deserts in Mexico (California Current), the Atacama Desert of Chile and Peru (Humboldt or Peru Current), and smaller arid zones off the coasts of Mauritania, Somalia, and NW Australia. The Atacama Desert of northern Chile is considered 'the driest place in the world,' with less than 10 mm of annual precipitation, and runs roughly about 4000 km from north to south. In addition, winds in this region typically blow parallel to the coast and thus inhibit the eastern movement of moist air from the Pacific to the Atacama region. In summary, the general causes of aridity and hence the presence of deserts in the world can be summarized as (1) atmospheric stability (subtropical highs/Hadley cell circulation); (2) rain-shadow effects; (3) upwelling cold currents off west coasts; (4) prevailing winds parallel to coasts; and (5) remoteness from moisture sources (continentality). It should be noted that most deserts are arid because of a combination of the above five factors.

Desert Hydroclimatology

Clear skies, lack of cloud cover, and low water vapor content are responsible for the high persistent temperatures that characterize most desert environments (variations discussed under microclimates) with maximum temperatures commonly between 45 and 50 °C. Subtropical deserts tend to experience hot summers and cool winters, while midlatitude deserts tend to have hot summers but very cold winters. Annual temperature fluctuations tend to be highest in midlatitude deserts, while diurnal temperature changes can be extreme in all desert environments, except the cool, coastal deserts (Figure 2). During the daytime, the incoming solar radiation heats up surfaces very rapidly causing the temperature (sensible heat) to rise. During the night, when terrestrial long-wave radiation dominates the surface energy budget, the surfaces cool very rapidly owing to the fact that clear, dry, and cloudless skies cannot trap the outgoing terrestrial radiation, and thus the very large diurnal temperature fluctuations. A >30 °C range is very typical in subtropical deserts and a 50 °C diurnal range has been reported from dark, basaltic rocks in the central Sahara Desert. The breakdown of rocks as a result of volumetric

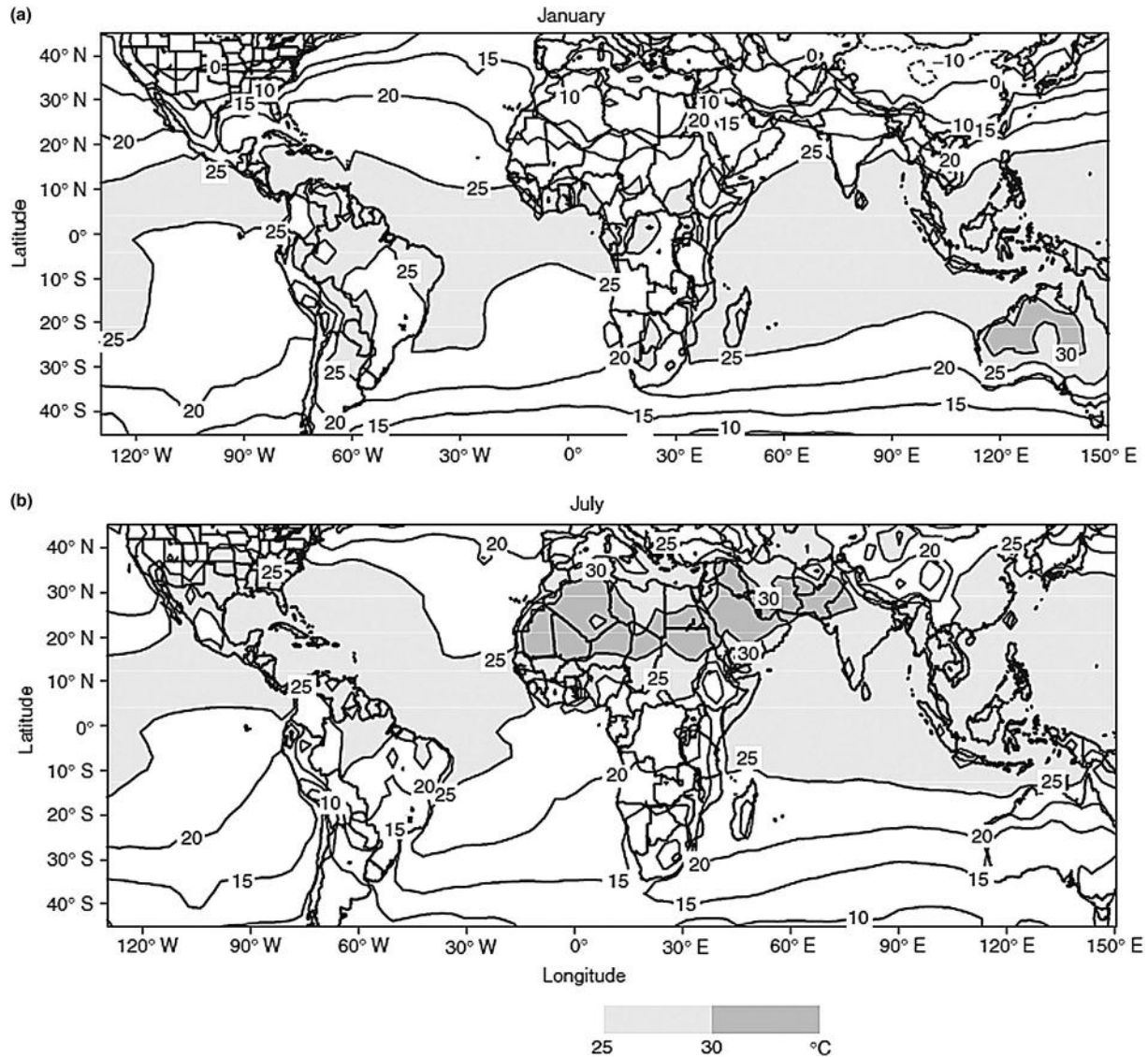


Figure 2 (a) January and (b) July long-term observed temperature. Temperatures greater than 25 °C are shaded. The darkest shading represents temperatures greater than 30 °C.

changes from thermal expansion and contraction has been proposed as one of the physical weathering processes in hyperarid deserts (most likely it is the individual minerals that are found in rocks that will respond to these thermal changes, leading to granular disaggregation. Unequal thermal expansion of the dominant minerals is the controlling factor).

Deserts typically receive less than 1 mm per day rainfall when averaged on an annual basis (Figure 3). Precipitation in deserts typically occurs in short durations but at high intensities, with low overall amounts, at irregular intervals, often with a strong seasonal bias and usually with a very large interannual variability. Enhanced precipitation owing to orographic (mountain) effects is especially prominent and has a strong impact on the spatial distribution of flora and fauna. Storms typically form as discrete convective cells and are unlikely to affect the entire drainage network within a desert – hence storms have low frequencies and high magnitudes and are

typically discontinuous in space and time. The rain falls on ground with a sparse or nonexistent vegetation cover, which is irregular in its distribution and especially adapted to collect rainfall. Interception rates are low and highly variable and rapid direct evaporation of excess surface water is characteristic. Evaporation rates from exposed surfaces are high in subtropical deserts, particularly during the summer months. Additionally, infiltration is largely controlled by the bare surface characteristics, which range from sands and alluvium to organic crusts and from stone (desert) pavements to duricrusts (a product of processes acting within the zone of weathering to cause the accumulation of iron and aluminum oxides, silica, calcium carbonate, or less commonly gypsum), such as calcrete (caliche/calcium carbonate).

Most runoff in deserts occurs as overland flow (Hortonian overland flow) with vegetation type and densities as well as surface and subsurface (soil) characteristics controlling the

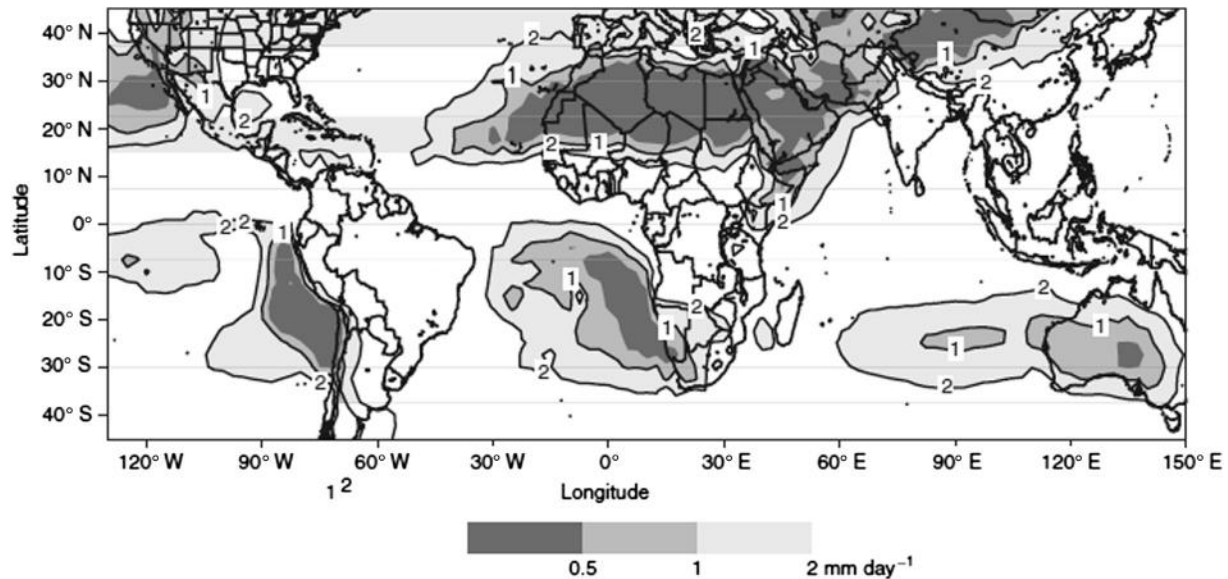


Figure 3 Annual precipitation rates for areas that receive less than 2 mm day^{-1} of rain. The darkest shaded regions receive less than 0.5 mm day^{-1} of rain.

intensity and duration of overland flow. Most overland flow is ephemeral, lasting only hours or days. Desert streams exhibit flashy hydrographs with short recessional limbs because of the predominance of overland flow, and seepage and transmission losses into the underlying alluvial stream beds are common, thus the importance of groundwater in providing moisture to plants and people. Although water paucity is the norm in deserts, there are water supplies that when used wisely, can sustain people and ecosystems in the global arid zone. Some of these sources are as follows: (1) perennial and ephemeral rivers, (2) wadis and arroyos, (3) groundwater from shallow (alluvial fans) and deep aquifers (mostly fossil aquifers), (4) lakes and playas, (5) fog, dew, and snow, (6) desalination plants, and (7) dams and reservoirs. The hydrological conditions in deserts leads to less weathering and leaching and hence soils tend to be shallow with coarse textures, high aeolian content, and retain many soluble substances (such as carbonates and salts). Soil pedogenesis (formation) takes a long time and paleosols (ancient soils) are rather common in many global deserts and can be used as proxy data for past climates.

Deserts also exhibit certain geomorphic processes and landforms that are rather unique or operate more favorably than in other environments. Some of these attributes include a rather open and exposed landform assemblages in part because of the limited vegetation cover with high and frequent changes in relative relief, many superimposed landforms and sediments from previous geologic periods – hence a veritable laboratory for studying past climates (such as lake cores and sediments), and the greater efficacy of wind as a geomorphological agent of erosion and deposition (hence the vast expanses of sand dunes and sand seas), and dust entrainment and transport.

Desert Microclimates

Microclimates are significant within deserts because they offer less arid conditions for plants, animals, and humans. Some

examples include (1) modification of relative humidity – mostly by nocturnal radiation and the shade effects or mixing with cooler air masses. The drops in temperatures increase the relative humidity of the air and the chances of moisture condensation as either fogs or dew (Namib Desert – up to 150 mm of moisture has been calculated). This process is vital for the survival of rich desert ecosystems such as in the Namib, Atacama, and Baja California deserts; (2) reduction of temperature extremes – any shade-giving object produces direct reduction of air temperatures in the arid lands because of the importance of the direct radiation component in the cloud-free atmosphere; and (3) reduction of wind speeds – shelter from wind movement reduces the amount of moisture loss from evapotranspiration – vegetation is sometimes used as a wind break; however, its competition for soil moisture may reduce crop yields close to the barrier (economic factors to be considered).

The single most important microclimate is provided by mountains, which offer the maximum modifications with respect to overall climatic variables. These include the reduction of air temperatures with altitude, shade effects (plants can grow which themselves provide shade – pinyon–juniper trees), increased chances of precipitation (orographic effects), as well as air drainage among the basins and the ranges (peaks and valleys), which can ameliorate diurnal temperatures and humidity.

Desertification

Desertification refers to land degradation in the global arid zone resulting primarily from various anthropogenic (human land use) and biophysical factors (climatic variations). In 1978 at the first UN Conference on Desertification (UNCOD) the following definition was proposed “Desertification is the diminution or destruction of the biological potential of the land which can lead ultimately to desert-like

conditions.” Desertification ultimately reduces the sustainability of arid lands whereby agriculturally productive lands become barren and thus prone to wind and water erosion and other forms of land degradation. For example, UNCOD estimates that moderate desertification can lead to a 10–25% drop in agricultural productivity. Contrary to public opinion, desertification does not refer to the expansion of deserts – although the margins of deserts are known to oscillate north and south owing to natural perturbations in climate and the resulting response of ecosystems. Desertification has occurred because desert ecosystems, which cover over one third of the world’s land area, are extremely vulnerable to overexploitation and inappropriate land use, compounded by poverty, political instability, deforestation, overgrazing, and wasteful irrigation practices. According to UNCOD, over 250 million people are directly affected by desertification. In addition, some 1000 million (or 1 billion) people in over 100 countries are at risk. These people include many of the world’s poorest, most marginalized, and politically weak citizens.

Another way for looking at desertification is to analyze the ‘5Ds’ specifically and these include drylands, drought, desiccation, degradation, and desertification. ‘Drylands’ are the world’s arid lands (semiarid, arid, and hyperarid) that are inherently prone to natural perturbations throughout geologic time; ‘drought’ is a short-term (a few years) and natural decline in precipitation and desert ecosystems and economic systems (people) adapt to those changes and eventually there is full recovery during moister times; ‘desiccation’ is drought conditions that lasts over an extended period of time (such as decadal), and has an adverse impact on both natural and cultural ecosystems with some systems never recovering or needing many years to reestablish (such as certain plant species or transborder migration of peoples); ‘degradation’ is the end result of drought and desiccation with the land losing its agricultural productivity leading to water and wind erosion, salinization on one hand, and the loss of the natural vegetation as a result of overgrazing, firewood collection, and groundwater

removal on the other; and ‘desertification’ would then be the ultimate end of this cycle whereby desert conditions overwhelm the whole ecosystem. Although the system above seems like a positive feedback scenario, land degradation and desertification are more likely to operate on a negative feedback method, whereby eventually the system will revert back to its original position albeit having crossed a number of thresholds and experienced a few positive feedbacks during its cyclical journey.

Desertification and the Sahel

It was the severe droughts beginning in the late 1960s in the Sahel (the areas immediately to the south of the Sahara Desert) and its subsequent socioeconomic consequences that enabled desertification to be firmly established within the global community as one of the most consequential environmental events of the late twentieth century (Figure 4). We will use the Sahel as a case study since it contains all the variables necessary for understanding global desertification. Three interrelated factors contributed to the environmental disaster of the Sahel:

1. Climatic – the Sahel is a transitional geographic region between the dry, desert climates of the Sahara and the moist and humid savanna environments to the south, with significant variability in the spatial and temporal distribution of precipitation. A short rainy season associated with the northward movement of the ITCZ is separated by long stretches of dry and dusty weather. Any prolonged interruptions in the arrival of the rainy season can lead to drought and desiccation. This transitional climatic zone is thus highly susceptible to both long-term and short-term climatic oscillations, and has undergone major climatic swings since the Last Glacial Maximum at about 18 ka. During the last Ice Age (Wisconsin Age in North America), the Sahel experienced an intense arid period characterized by major dune building episodes and dust deposition, followed by a much cooler and wetter period from 11 to 8 ka (the period of major rock art in the Sahara with petroglyphs

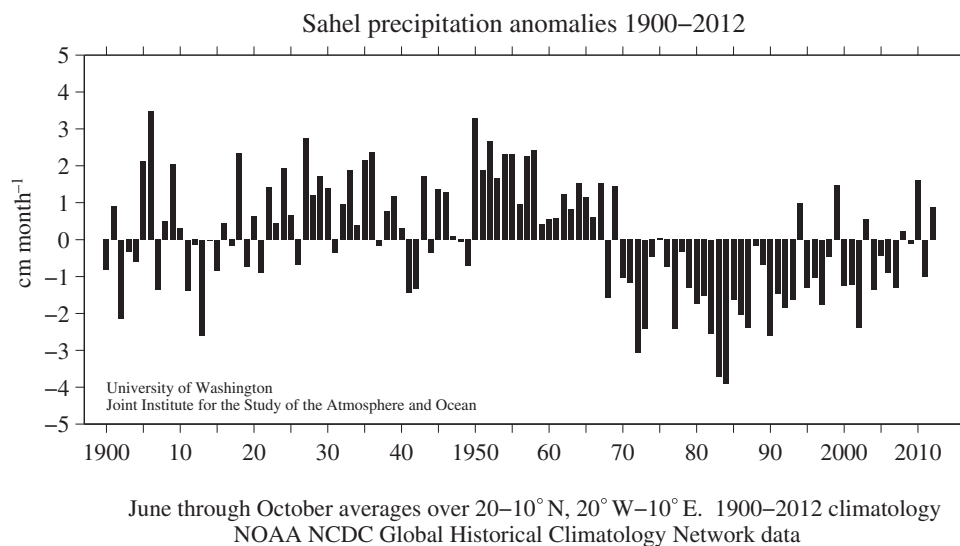


Figure 4 Sahel precipitation anomalies 1900–2012. Reproduced from <http://jisao.washington.edu/data/sahel/>.

of crocodiles and giraffes painted in rock shelters and caves). This period was then followed by a brief arid episode and then another wetter than present period from 7 to 5 ka. Since 5 ka, aridity has slowly returned to the Sahara–Sahel, as the subtropical high-pressure cells assumed their current position, with today's hyperarid central Saharan core region well established by 2 ka. Desert margins, such as the Sahel, are thus very prone to extreme climatic variability, with many decades of good wet years followed by short to decadal length drought periods. The natural vegetation is rather resilient and adapted to these chronic environmental stresses and will typically recover after brief or even decadal perturbations in rainfall regime. However, twentieth century changes in land use and land cover from increased population pressures on the landscape have altered the natural ecosystem cycle (more below). Rainfall in the Sahel appears to be strongly influenced by the combined effects of North Atlantic Oscillation (NAO) and the El Niño–Southern Oscillation (ENSO). The interannual variability in the position of the northern boundary of the Sahel (southern boundary of the Sahara), as represented by the 200 mm isohyet, can be explained in large measure by changes in NAO and ENSO. Various studies have indicated that up to 75% of the interannual variation in the extent of the Sahara Desert is accounted for by the combined effects of NAO and ENSO. The drier years in the Sahel tend to be associated

with warm sea surface temperatures in the southern oceans and Indian Ocean, and anomalously cold sea surface temperatures to the west of the continent. Two recent studies have advanced our understanding of the physical factors controlling long-term, persistent drought in the Sahel. One study has shown that the Sahel drought of the last 40 years was likely initiated by a change in worldwide ocean temperatures (multidecadal variations in sea-surface temperatures (SSTs)), which reduced the strength of the African monsoon, shifted the ITCZ, and was exacerbated by land–atmosphere feedbacks through natural vegetation and land cover change (Figure 5). Land use changes by humans may have also played an important role. Another study, using laminated sediment cores from a lake, found that the recent Sahelian droughts of the 1960s and 1970s, are characteristic of the monsoon and are linked to natural variations in Atlantic SSTs and, furthermore, these droughts have occurred many times during the past three millennia, and although the most recent multidecadal drought of the 1970s had widespread ecological, political, and socioeconomic impacts, the climatic oscillations of the past 3000 years are capable of much more severe and longer drought cycles. In summary, large-scale atmospheric circulation changes brought on by multidecadal oscillations in global sea-surface temperatures, seem to be the primary biophysical mechanism (in addition to the human

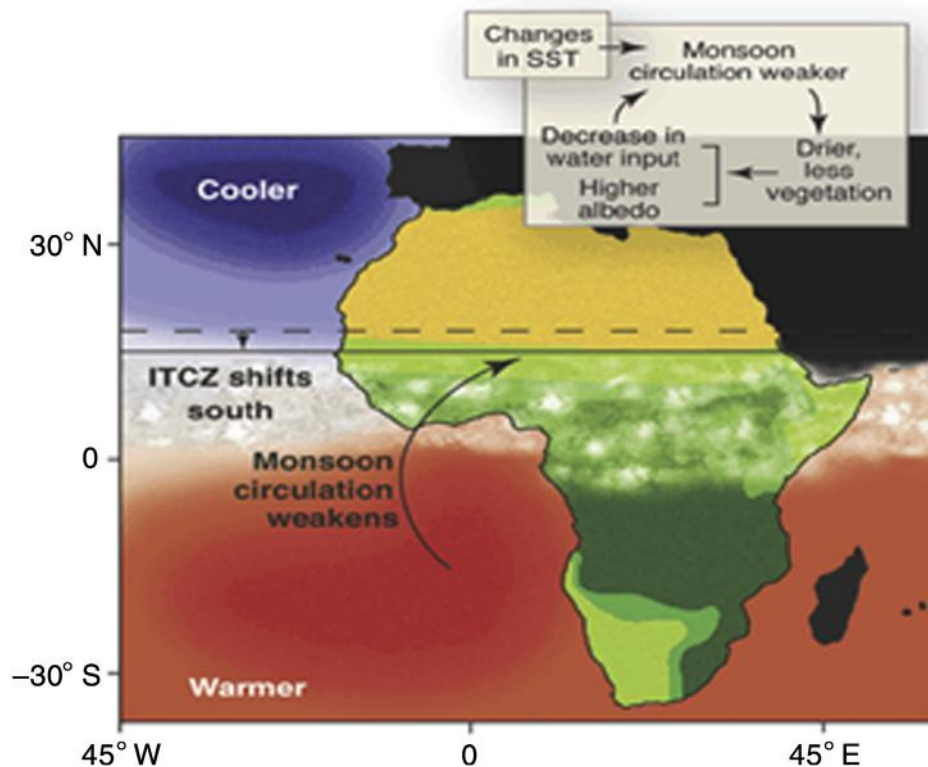


Figure 5 Complex feedbacks. The recent Sahel drought was likely initiated by a change in worldwide ocean temperatures, which reduced the strength of the African monsoon, and was exacerbated by land–atmosphere feedbacks through natural vegetation and land cover change. Land use changes by humans may have also played an important role. SST, sea-surface temperature; ITCZ, intertropical convergence zone. Reproduced from Zeng, N., Meyerson, J., 2003. Drought in the Sahel. *Science* 302 (5647), 999–1000.

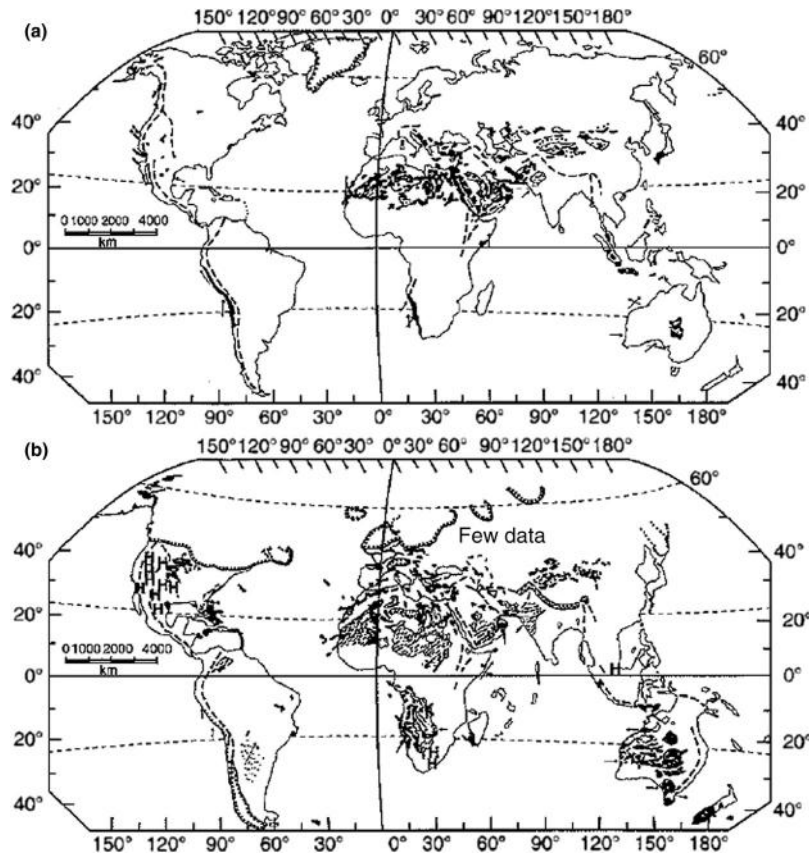


Figure 6 Global distribution of sand seas for (a) the present and (b) the Last Glacial Maximum (LGM), 18 ka (after Sarnthein, 1978). H denotes humid conditions. Reproduced from Tchakerian, V.P., 2009. Paleoclimatic interpretations from desert dunes and sediments. In: Parsons, A.J., Abrahams, A.D. (Eds.), *Geomorphology of Desert Environments*. Springer-Verlag, New York, pp. 757–772.

component discussed later) that contributes to desertification, and that their length, severity, and origin can be traced back at least 3000 years.

2. Geomorphic – underlying the natural surface of the Sahel is a mantle of relict dune sands (ergs) from the Pleistocene presently stabilized and anchored by the vegetation (Figure 6). The removal of the vegetation cover by either anthropogenic or natural causes will lead to a significant increase in aeolian activity and in sand/dust storms. Owing to the nature of the aeolian sand, the topsoil is thin, with shallow root systems, and the water table very low and highly susceptible to increases in water use as well as to pedogenic carbonate formation (which can affect water quality and translocation of nutrients). Aeolian activity is further enhanced in this region by the presence of many winds such as the trade winds, Harmattan, and Khamsin, and other convective mesoscale systems. The Harmattan, is a dry and dusty Sahelian tradewind that sometimes extends all the way to the ITCZ. It blows south from Sahara into the Gulf of Guinea from November to March. Over the Sahara, it picks up fine dust particles some of which might end up in the Caribbean and even North America. Since sediment transport varies with the cube of the wind speed, a slight increase in wind speed will result in a threefold increase in sand and dust transport. Reactivation of some of the

stabilized Sahelian dormant/relict dune systems has been the result of both increased population pressure in the region (stabilized dunes provide a richer plant cover for grazing and firewood gathering, and are also easier to cultivate), and to long-term decadal drought owing to the atmospheric perturbations from global sea-surface temperature changes.

3. Anthropogenic – factors such as overgrazing and conversion of woodland to agriculture are among features that have been proposed as human-induced causes for desertification. Persistent, decadal drought, and desiccation leads to increased pressure on land use and land cover, compounded by a concomitant increase in population (from high birth rates, immigration, and population movements because of conflict). The demand for water and energy also soar putting additional pressure on people and governments. The human-induced stresses lead to such manifestations of desertification as (1) soil erosion and salinization (and/or waterlogging) as a result of the population exceeding the environmental thresholds of agricultural sustainability owing to overgrazing, overcultivation, deforestation, etc.; (2) wind erosion leading to increased frequencies of dust and sand storms (including loss of topsoil); (3) drawdown in groundwater, wells, and diminished use of dams and other irrigation schemes; (4) introduction of exotic species;

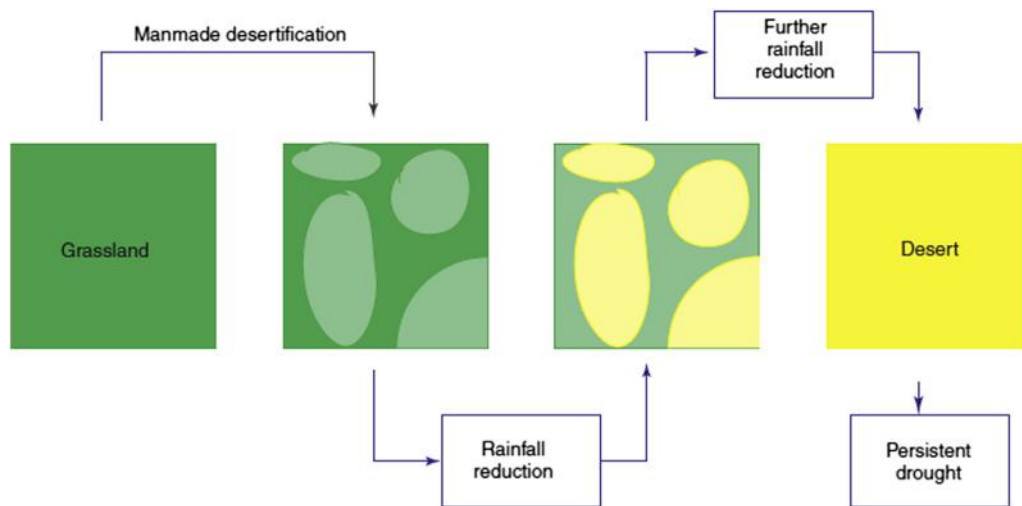


Figure 7 Schematic mechanism for the enhancement of desertification through biosphere-atmosphere interactions.

(5) war and civil unrest with all its socioeconomic ramifications – the Sahelian population is doubling every 20 years. All the above ultimately changes the ecosystem with a marked loss in biodiversity and agricultural productivity leading to a prolonged hiatus whereby a return to the original state may take decades or longer to attain.

The environmental effects accompanying desertification are many and widespread and include such factors as (1) increase in surface albedo (less sunlight is absorbed), (2) decrease of evaporation and transpiration, (3) reduction in the moisture supply to the atmosphere (less water vapor available for condensation), (4) decrease of soil moisture, (5) decrease in precipitation (compounded by increases in albedo and reduced moisture in the atmosphere), which ultimately leads to less favorable conditions for plant growth, and (6) surface temperature changes tend to be highly seasonal and closely related to the hydrologic cycle. On one hand, an increase in surface albedo will result in a decrease of surface net radiation, hence cooling the surface, while on the other hand, a decrease in latent heat as a result of lowered evapotranspiration can lead to the warming of the surface – thus an increase or decrease in surface temperatures will depend largely whether it is the dry or the wet season in the Sahel. Increased dust production from denuded dry soils and dry dune sands as well as from biomass burning (mineral aerosols) can have a number of effects on the environment including influencing the atmospheric radiative transfer directly by scattering and absorbing solar radiation, and indirectly by modifying the optical property and lifetime of clouds. Biomass burning from firewood, charcoal, and animal

dung releases significant amount of greenhouse gases as well as black carbon (soot), which could affect both atmospheric phenomena and human health. Atmosphere-biosphere interactions can also lead to enhancement or a positive feedback loop and become self-perpetuating until a certain threshold is reached. Drought-induced ecosystem degradation will then reinforce the initial anthropogenically driven changes and thus ultimately changing the regional climate, leading to a self-degradation of the land surface as well as persistent drought (Figure 7). Only a return to a decadal or longer above average hydrologic conditions accompanied with changes in land use can jolt (threshold) the system to its former stage.

Further Reading

- Giannini, A., Saravanan, R., Chang, P., 2003. Oceanic forcing of Sahel rainfall on interannual to interdecadal time scales. *Science* 302, 1027–1031.
- Goudie, A.S., 2002. *Great Warm Deserts of the World*. Oxford University Press, Oxford.
- Lait, J.E., 2008. *Deserts and Desert Environments*. Wiley-Blackwell Publishers.
- Shanahan, T.M., Overpeck, J.T., Anchukaitis, K.J., Beck, J.W., Cole, J.E., Dettman, D.L., Peck, J.A., Scholz, C.A., King, J.W., 2009. Atlantic forcing of persistent drought in West Africa. *Science* 324, 377–380.
- Tchakerian, V.P., 1999. Dune palaeoenvironments. In: Goudie, A.S., Livingstone, I., Stokes, S. (Eds.), *Aeolian Environments, Sediments and Landforms*. John Wiley & Sons, New York, pp. 261–292.
- Tchakerian, V.P., 2009. Palaeoclimatic interpretations from desert dunes and sediments. In: Parsons, A.J., Abrahams, A.D. (Eds.), *Geomorphology of Desert Environments*. Springer-Verlag, New York, pp. 757–772.
- Thomas, D.S.G., Middleton, N.J., 1994. *Desertification: Exploding the Myth*. John Wiley & Sons, Chichester.

Drought

S Quiring, Texas A&M University, College Station, TX, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by D A Wilhite, volume 2, pp 650–658, © 2003, Elsevier Ltd.

Synopsis

Drought is more than just a precipitation deficit. It occurs as a result of the complex interplay between natural precipitation deficiencies and the demands of human and environmental water use. It is the most far-reaching climate-related disaster each year, causing hardship to millions of people. Impacts ripple through the economy and may linger for years after the termination of the drought episode. Societal vulnerability to drought is escalating in both developing and developed countries, and at a significant rate. Future climate changes may lead to more frequent and severe droughts.

Introduction

Drought results from a deficiency of precipitation from expected or 'normal.' However, drought is more than just a precipitation deficit. Drought occurs as a result of the complex interplay between natural precipitation deficiencies on varying time and space scales and the demands of human and environmental water use, which may be exacerbated by inefficiencies in water distribution, planning, and management. Drought is a natural feature of the global climate system that occurs in virtually all regions of the world. Drought affects various sectors of society and the natural environment in many ways. It is the most far-reaching climate-related disaster each year, causing hardship to millions of people. According to the National Climatic Data Center, losses from drought in the United States exceeded \$210 billion (United States) during the period 1980–2011. Droughts accounted for roughly 24% of all losses from major weather events (i.e., floods, hurricanes, and severe storms). Droughts and hurricanes are the most costly of natural hazards in the United States. For example, the 2012 drought in the United States affected more than half the country and caused \$30 billion in damages. It was the largest drought, in terms of spatial extent, in the United States since the 1930s. While drought rarely causes fatalities in the developed nations like the United States, it can lead to famine and political instability in other regions of the world. Impacts of drought appear to be increasing in both developing and developed countries, which in many cases is an indication of unsustainable development. Drought is considered to be one of the most complex and least understood of all natural hazards, affecting more people than any other hazard. Numerous definitions of drought exist, reflecting regional variations in climatic characteristics and sector-specific impacts. Simply stated, drought originates from a deficiency of precipitation that results in water shortage for some activity or group.

The Concept of Drought

Drought differs from other natural hazards (such as floods, tropical cyclones, and earthquakes) in several ways. First, drought is a slow-onset, creeping natural hazard. Its effects often accumulate slowly over a considerable period of time and

may linger for years after precipitation returns and the drought ends. Therefore, the onset and end of drought are difficult to identify. This means that scientists and policy makers continue to debate the basis (i.e., the criteria) for declaring an end to a drought. Second, the absence of a precise and universally accepted definition of drought adds to the confusion about whether or not a drought exists and, if it does, its degree of severity. Realistically, definitions of drought must be region-specific and application-specific. This is one reason why there are so many definitions of drought and so many different metrics/indices for measuring it. Third, drought impacts are nonstructural and spread over a larger geographical area than damages that result from other natural hazards. Therefore, quantifying the impacts of drought events and providing disaster relief are far more difficult tasks than they are for other natural hazards. Emergency managers, for example, are more accustomed to dealing with impacts that are structural and localized, responding to these events by restoring communication and transportation channels, providing emergency medical supplies, ensuring safe drinking water, and so forth. These characteristics have hindered our understanding of drought and the development of appropriate strategies to reduce vulnerability and improve response and recovery.

Many people consider drought to be largely a natural or physical event. Figure 1 illustrates that, in reality, drought has both natural and social components. The risk associated with drought for any region is a product of both the region's exposure to the event (i.e., probability of occurrence at various severity levels) and the vulnerability of society to the event. The natural event (i.e., meteorological drought) is a result of the occurrence

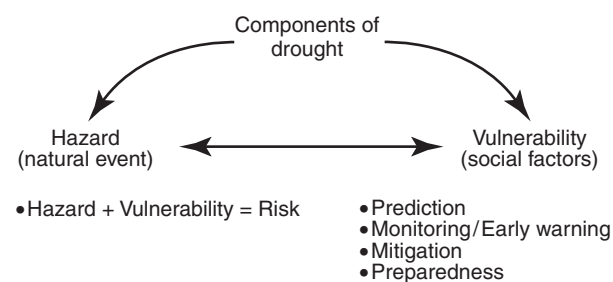


Figure 1 Components of drought. National Drought Mitigation Center.

of persistent large-scale disruptions in the global circulation pattern of the atmosphere. Exposure to drought varies spatially and there is little, if anything, that we can do to alter drought occurrence. Vulnerability, on the other hand, is determined by social factors such as population changes, population shifts (regional and rural to urban), demographic characteristics, technology, policy, and social behavior. These factors change over time and thus vulnerability is likely to increase or decrease in response to these changes. Subsequent droughts in the same region will have different effects, even if they are identical in intensity, duration, and spatial characteristics, because societal characteristics will have changed. However, much can be done to lessen societal vulnerability to drought.

Defining Drought

Drought is the consequence of a natural reduction in the amount of precipitation received over an extended period of time, although other climatic factors (such as high temperatures, high winds, and low relative humidity) are often associated with drought and can significantly aggravate its severity. Drought is also related to the timing (principal season of occurrence, delays in the start of the rainy season, occurrence of rains in relation to principal crop growth stages) and the effectiveness of the rains (rainfall intensity, number of rainfall events). Thus, each drought event is unique in its climatic characteristics, spatial extent, and impacts. The area affected by drought is rarely static during the course of an event. As drought emerges and intensifies, its core area shifts and its spatial extent expands and contracts throughout the duration of the event. Because drought affects so many economic and social sectors, numerous definitions have been developed. In addition, because drought occurs with

varying frequency in nearly all regions of the world, all types of economic systems, and developed and developing countries alike, the approaches taken to define it also reflect regional and ideological differences. Impacts also differ spatially and temporally, depending on the societal context of drought. A universal definition of drought is an unrealistic expectation.

There are two main types of drought definitions: *operational* and *conceptual*. Operational definitions are impact- or application-specific that are used in an operational mode by decision makers to identify the beginning, end, and degree of severity of a drought. Conceptual definitions are formulated in general terms and they are used to help explain what a drought is. The large number of conceptual drought definitions that exist illustrates the challenge of accurately describing drought and the lack of agreement in regards to drought definitions. Conceptual definitions of drought are commonly classified into four categories: meteorological, agricultural, hydrological, and socio-economic drought. [Figure 2](#) explains the relationship between these various types of drought and the duration of the event. The impacts associated with drought usually take weeks to months to develop, but this period can vary considerably, depending on the timing of the initiation of the precipitation deficiency.

Meteorological (or climatological) drought is expressed solely on the basis of the degree of dryness (often in comparison to some normal or average amount) and the duration of the dry period. Thus, intensity and duration are the key characteristics of these definitions. Meteorological drought definitions must be considered as region-specific since the atmospheric conditions that result in deficiencies of precipitation are dependent on the climate regime. Most meteorological drought definitions relate actual precipitation departures to average amounts on monthly, seasonal, water year, or annual time scales.

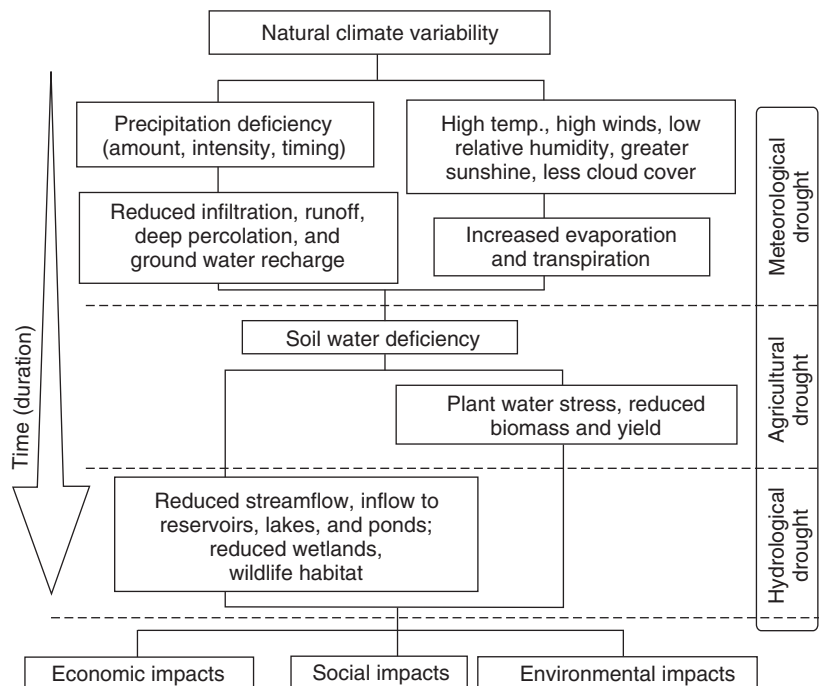


Figure 2 Relationship between types of drought. National Drought Mitigation Center.

Agriculture is usually the first economic sector to be affected by drought because soil moisture supplies are often quickly depleted, especially if the period of moisture deficiency is associated with high temperatures and windy conditions. Agricultural drought links various characteristics of meteorological drought to agricultural impacts, focusing on precipitation shortages, differences between actual and potential evapotranspiration, and soil water deficits. A plant's demand for water is dependent on prevailing weather conditions, biological characteristics of the specific plant, its stage of growth, and the physical and biological properties of the soil. A definition of agricultural drought should account for the variable susceptibility of crops at different stages of crop development. For example, deficient subsoil moisture in an early growth stage will have little impact on final crop yield if topsoil moisture is sufficient to meet early growth requirements. However, if the deficiency of subsoil moisture continues, a substantial yield loss may result.

Hydrological droughts are associated with the effects of prolonged periods of precipitation shortfall on surface or subsurface water supply (i.e., streamflow, reservoir and lake levels, groundwater). Hydrological droughts usually lag the occurrence of meteorological and agricultural droughts. More time elapses before precipitation deficiencies are detected in other components of the hydrological system (e.g., reservoirs, groundwater). Also, water in hydrological storage systems (e.g., reservoirs, rivers) is often used for multiple and competing purposes (power generation, flood control, irrigation, recreation), further complicating the sequence and quantification of impacts. Competition for water in these storage systems escalates during drought, and conflicts among water users increase significantly.

Finally, socioeconomic drought associates the supply and demand of some economic good or service with elements of meteorological, hydrological, and agricultural drought. Socioeconomic drought is associated directly with the supply of some commodity or economic good (e.g., water, hay, hydroelectric power) that is the result of precipitation shortages. Increases in population can substantially alter the demand for these economic goods over time. This concept of drought supports the strong symbiosis that exists between drought and its impacts and human activities. Thus, the incidence of drought could increase because of a change in the frequency of meteorological drought, a change in societal vulnerability to water shortages, or both. For example, poor land-use practices such as overgrazing can decrease animal carrying capacity and increase soil erosion, which exacerbates the impacts of and vulnerability to future droughts.

Operational Definitions of Drought

Developing operational definitions of drought is also challenging because drought has no definitive onset/end, is slow to evolve, and is regionally relative. Operational drought definitions are used to trigger water conservation measures and determine whether (and how much) drought assistance will be provided to affected regions. Failure to adequately define and monitor drought can have a significant negative impact, particularly if it fails to trigger a response (e.g., limiting water use) when one is sorely needed, or if it triggers a response when

one is not required. Operational definitions of drought typically utilize drought indices or metrics like percentage of normal precipitation, Palmer Drought Severity Index, or the Standardized Precipitation Index. However, most drought indices were not specifically designed with this purpose in mind. There is much less discussion about operational drought definitions in the scientific literature, regarding and how they should be developed and used to monitor drought and to trigger drought response. Most of the state drought plans in the United States employ a single set of operational drought definitions for the entire state and these definitions have usually been developed using subjective criteria. It is more appropriate to use objective and standardized approaches to develop operational drought definitions.

Drought Characteristics and Severity

Droughts differ from one another in three essential characteristics: intensity, duration, and spatial coverage. Intensity refers to the degree of the precipitation shortfall and/or the severity of impacts associated with the shortfall. It is generally measured by the departure of some climatic index from normal and is closely linked to duration in the determination of the impact. The simplest index in widespread use is the percentage of normal precipitation, but there are numerous more complex and more effective indices available. It is generally recommended that several indices be used to monitor drought onset and development because each index has its particular strengths and weaknesses. One of the principal difficulties with this (or any) index is the choice of the threshold below which the deficiency of precipitation must fall (e.g., 75% of normal) to define the onset of drought and trigger various mitigation actions or response programs. Thresholds are usually chosen arbitrarily. In reality, they should be linked to the impact. Another distinguishing feature of drought is its duration. Droughts usually require a minimum of weeks to months to become established but then can continue for months or years. The magnitude of drought impacts is closely related to the timing of the onset of the precipitation shortage, its intensity, and the duration of the event. For example, a dry period that begins in the late fall and continues through the winter months in the midwestern United States will likely have negligible impacts. However, if dry conditions persist into the spring and early summer months, agricultural and urban demands for water supplies increase dramatically. Pasture growth will be reduced, affecting livestock producers through supply shortages and increasing prices. Diminished topsoil and subsoil moisture will also affect seed emergence and early growth development of grain crops such as corn and soybeans, eventually affecting yield and crop production, if dry conditions persist throughout the summer months. Urban water supplies are often reduced, forcing water suppliers to impose voluntary or mandatory water conservation measures. As drought conditions extend over more than one growing season, impacts magnify substantially as a result of declining surface and subsurface water supplies and an expanding circle of impacts.

Droughts also differ in terms of their spatial characteristics. The areas affected by severe drought evolve gradually, and regions of maximum intensity shift from season to season. In

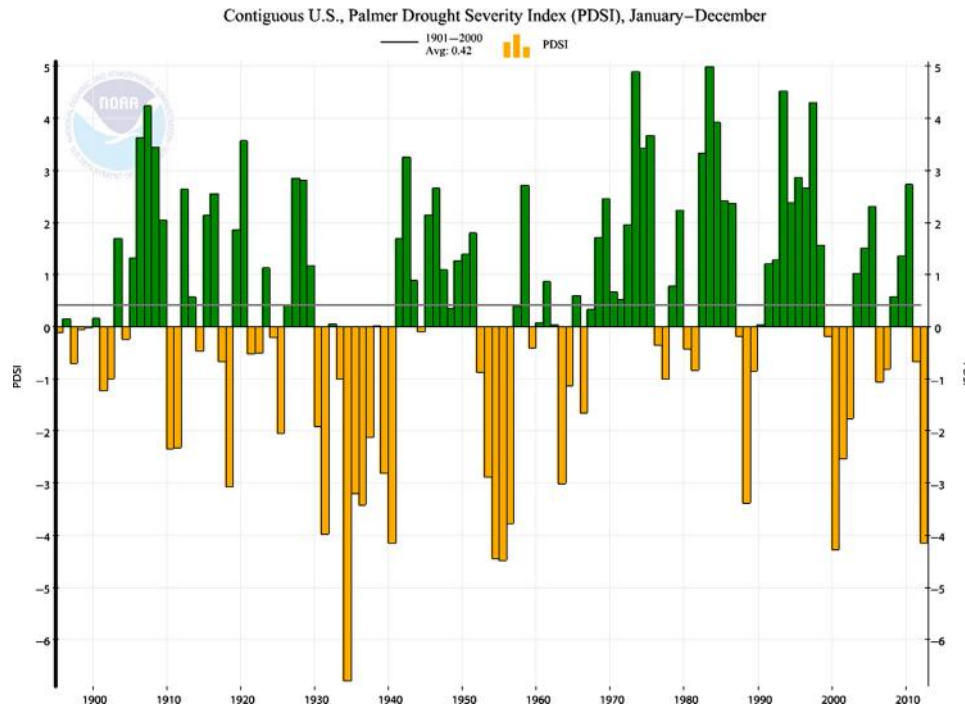


Figure 3 Percentage area of the United States in severe and extreme drought, 1895–2013. National Drought Mitigation Center, based on data provided by the National Climatic Data Center, National Oceanic and Atmospheric Administration, US Department of Commerce. <http://www.ncdc.noaa.gov>.

larger countries, such as Brazil, China, India, the United States, or Australia, drought would rarely, if ever, affect the entire country. During the severe drought of the 1930s in the United States, for example, the area affected by severe drought never exceeded 65% of the country (Figure 3). By contrast, drought affected more than 95% of the Great Plains region in 1934. On the other hand, it is indeed rare for drought not to exist in a portion of the United States each year. Thus, the governments of larger countries are more accustomed to dealing with water shortages and have established an infrastructure to respond, albeit reactively. For smaller countries, it is more likely that the entire country might be affected since droughts are usually regional phenomena. From a planning perspective, the spatial characteristics of drought have serious implications. Nations should know the probability that drought may simultaneously affect all or several major crop-producing regions within their borders and develop contingencies for the occurrences of such an event. Likewise, it is important for governments to know the chances of a regional drought simultaneously affecting agricultural productivity in their country as well as adjacent or nearby nations on whom they are dependent for food supplies. In some instances, a nation's primary drought mitigation strategy may be to import food from nearby nations, ignoring the likelihood that a drought may have significant regional impacts on food supplies. Likewise, the occurrence of drought worldwide or in the principal grain-exporting nations, such as occurred during the El-Niño Southern Oscillation (ENSO) event of 1982–83, may significantly alter a developing country's access to food from donor governments.

The Impacts of Drought

The impacts of drought are diverse and often ripple through the economy. Thus, impacts are referred to as direct or indirect. Conceptually speaking, the more removed the impact from the cause, the more complex the link to the cause is. In other words, a loss of yield resulting from drought is a direct or first-order impact of drought. However, the consequences of that impact (e.g., loss of income, farm foreclosures, government relief programs) are secondary or tertiary impacts. Because of the number of affected groups and sectors associated with drought, its spatial extent, and the difficulties connected with quantifying environmental damages and personal hardships, the precise determination of the financial costs of drought is an arduous task. It has been estimated that the average annual impacts of drought in the United States are \$6–8 billion. These figures may be misleading because drought years often occur in clusters, such as the 1930s, 1950s, 1987–89, and 1999–2002 (Figure 3). Impacts during each of these years were much above the annual average. The 1988 drought caused an estimated \$40 billion in damages. The impacts of drought can be classified into three principal areas: economic, environmental, and social. Table 1 presents a simplified illustration of the impacts associated with each of these areas. Economic impacts range from direct losses in the broad agricultural and agriculturally related sectors, including forestry and fishing, to losses in recreation, transportation, banking, and energy sectors. Other economic impacts would include added unemployment and loss of revenue to local, state, and federal government. Environmental losses are the result of damages to plant and animal

Table 1 Classification of drought-related impacts (costs and losses)

<i>Problem sectors</i>	<i>Impacts</i>
Economic	Loss from crop production Annual and perennial crop losses; damage to crop quality Reduced productivity of cropland (wind erosion, etc.) Insect infestation Plant disease Wildlife damage to crops Loss from dairy and livestock production Reduced productivity of range land Forced reduction of foundation stock Closure/limitation of public lands to grazing High cost/unavailability of water for livestock High cost/unavailability of feed for livestock High livestock mortality rates Increased predation Range fires Loss from timber production Forest fires Tree disease Insect infestation Impaired productivity of forest land Loss from fishery production Damage to fish habitat Loss of young fish due to decreased flows Loss of national economic growth, retardation of economic development Income loss for farmers and others directly affected Loss of farmers through bankruptcy Loss to recreational and tourism industry Loss to manufacturers and sellers of recreational equipment Increased energy demand and reduced supply because of drought-related power curtailments Costs to energy industry and consumers associated with substituting more expensive fuels (oil) for hydroelectric power Loss to industries directly dependent on agricultural production (machinery and fertilizer manufacturers, food processors, etc.) Decline in food production/disrupted food supply Increase in food prices Increased importation of food (higher costs) Disruption of water supplies Unemployment from drought-related production declines Strain on financial institutions (foreclosures, greater credit risks, capital shortfalls, etc.) Revenue losses to federal, state, and local governments (from reduced tax base) Deterrence of capital investment, expansion Dislocation of businesses Revenues to water supply firms Revenue shortfalls Windfall profits Loss from impaired navigability of streams, rivers, and canals Cost of water transport or transfer Cost of new or supplemental water resource development
Environmental	Damage to animal species Reduction and degradation of fish and wildlife habitat Lack of feed and drinking water Disease Increased vulnerability to predation (e.g., from species concentration near water) Loss of biodiversity Wind and water erosion of soils Reservoir and lake drawdown Damage to plant species Water quality effects (e.g., salt concentration, increased water temperatures, pH, dissolved oxygen) Air quality effects (dust, pollutants) Visual and landscape quality (dust, vegetative cover, etc.) Increased fire hazard Estuarine impacts; changes in salinity levels, reduced flushing

(Continued)

Table 1 Classification of drought-related impacts (costs and losses)—cont'd

<i>Problem sectors</i>	<i>Impacts</i>
Social	Increased groundwater depletion (mining), land subsidence Loss of wetlands Loss of cultural sites Insect infestation Food shortages (decreased nutritional level, malnutrition, famine) Loss of human life (e.g., food shortages, heat) Public safety from forest and range fires Conflicts among water users, public policy conflicts Increased anxiety Loss of aesthetic values Health-related low flow problems (e.g., diminished sewage flows, increased pollutant concentrations, etc.) Recognition of institutional constraints on water use Inequity in the distribution of drought impacts/relief Decreased quality of life in rural areas Increased poverty Reduced quality of life, changes in lifestyle Social unrest, civil strife Population migration (rural to urban areas) Reevaluation of social values Increased data/information needs, coordination of dissemination activities Loss of confidence in government officials Recreational impacts

National Drought Mitigation Center.

species, wildlife habitat, and air and water quality; damage from forest and range fires; degradation of landscape quality; and soil erosion. Although these losses are difficult to quantify, growing public awareness and concern for environmental quality has forced public officials to focus greater attention on these effects. Social impacts mainly involve public safety, health, conflicts among water users, and inequities in the distribution of impacts and disaster relief programs. As with all natural hazards, the economic impacts of drought are highly variable within and between economic sectors and geographic regions, producing a complex assortment of winners and losers with the occurrence of each disaster. For example, decreases in agricultural production result in enormous negative financial impacts on farmers in drought-affected areas, at times leading to foreclosure. This decreased production also leads to higher prices of grains, vegetables, and fruit. These price increases have a negative impact on all consumers as food prices increase. However, farmers outside the drought-affected area with normal or above-normal production or those with significant grain in storage reap the benefits of these higher prices. Similar examples of winners and losers can be given for other economic sectors as well.

Drought Response and Preparedness

With the occurrence of any natural disaster come appeals for disaster assistance from the affected area. For decades, governments have typically responded to drought by providing emergency, short-term, and long-term assistance to distressed areas. Emergency and short-term assistance programs are often reactive, a kind of 'band-aid' approach to more serious land and water management problems. Actions of this type have long been criticized by scientists and

government officials, as well as by recipients of relief, as inefficient and ineffective. Long-term assistance programs are far fewer in number, but they are proactive. They attempt to lessen a region's vulnerability to drought through improved management and planning. Governmental response to drought includes a wide range of potential actions to deal with the impacts of water shortages on people and various economic sectors. In the United States and other developed countries, agencies of the federal government typically respond by making massive amounts of relief available to the affected areas. Most of this relief is in the form of short-term emergency measures to agricultural producers and few, if any, of these assistance measures in recent years have been aimed at reducing future vulnerability. In developing countries, emergency response is often provided by donor governments, nongovernmental organizations, and international organizations. The assistance that is provided typically is in the form of food aid, health services, access to potable water supplies, and transportation services. Because of the unique character of drought, governments have been less inclined to invest resources to develop well-conceived mitigation programs and contingency plans. This reactive approach to natural disasters is commonly referred to as crisis management. Research has demonstrated that reaction to crisis often results in the implementation of hastily prepared assessment and response procedures that lead to ineffective, poorly coordinated, and untimely response. An alternative approach is to initiate planning between periods of drought, thus developing a more coordinated and proactive response that would more effectively address those persons, areas, and sectors mostly at risk. Also, the limited resources available to government to mitigate the effects of drought could be allocated in a more beneficial manner.

Drought Policy and Planning

Drought planning is defined as actions taken by individual citizens, industry, government, and others in advance of drought for the purpose of mitigating some of the impacts and conflicts associated with its occurrence. Because drought is a normal part of climate variability for virtually all regions, it is important to develop plans to deal with these extended periods of water shortage in a timely, systematic manner as they evolve. This planning process needs to occur at various levels of government and to be integrated between levels of government. The purpose of a drought plan is to reduce the impacts of drought by identifying the principal sectors, groups, or regions mostly at risk and developing mitigation actions and programs that can reduce these risks in advance of future drought events. Generally, drought plans have three basic components: monitoring and early warning; risk and impact assessment; and response and mitigation. Plans will also improve coordination within agencies of government and between levels of government. In the United States, there has been a remarkable growth in the number of states with drought plans. In 1982, only three states had drought plans. By late 2000, 30 states have drought plans and 6 additional states are in various stages of plan development. Today, only three states do have state drought plans in place or under development (Figure 4). Planning methodologies are available to facilitate the planning process (see [Further Reading](#)). This trend demonstrates an increased concern about the potential impacts of extended water shortages and the complexity of those impacts. Drought plans are at the foundation of improved drought management, but only if they emphasize risk assessment and mitigation programs and actions. Many of the drought plans that currently exist in the United States

still emphasize emergency response or crisis management, but in recent years there has been increased focus on state drought plans that are based on mitigation.

Other developed and developing countries are also making substantial progress in drought preparedness through the application of appropriate risk management techniques. Australia developed a national drought policy in the early 1990s and this policy has undergone revision as experience with the policy has increased. The Australian policy is focused on the agricultural sector and seeks to improve the self-reliance or drought-coping capacity of farmers and to promote the sustainable use of natural resources. The government has invested resources in improving longer-range forecasts and other decision support systems to aid farmers in farm management. South Africa also developed a national drought policy that emphasizes the same principles as that implemented by Australia. Many other countries in South America, Europe, Africa, and Asia are following this trend toward improving drought management through the application of the principles of risk management. Before developing a preparedness plan, government officials should first define what they hope to achieve by that plan. Thus, a drought policy statement should be prepared in advance of a plan. The objectives of drought policy should encourage or provide incentives for agricultural producers, municipalities, and other water-dependent sectors or groups to adopt appropriate and efficient management practices that help to alleviate the effects of drought. Past relief measures have usually discouraged the adoption of appropriate management techniques. Assistance should also be provided in an equitable, consistent, and predictable manner to all without regard to economic circumstances, industry, or geographic region. An objective should also seek to protect the natural and agricultural resource base. Degradation of natural resources can result in spiraling economic, environmental, and social costs.

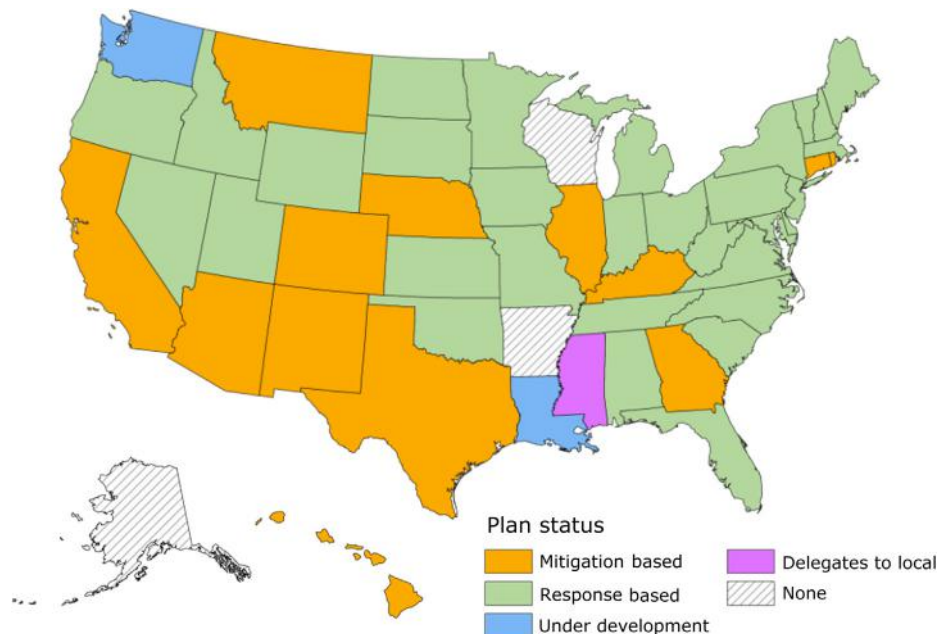


Figure 4 Status of drought planning, United States. National Drought Mitigation Center. <http://drought.unl.edu/>.

Future Droughts

The impacts of future droughts will be determined not only by the frequency and intensity of drought events, but also by the number of people at risk, their degree of risk, and the resiliency of natural systems. As demand for water and other shared natural resources increases as a result of population growth and migration to drought-prone areas, urbanization, environmental degradation, government policies, land use changes, technology and other factors, future droughts can be expected to produce increasingly negative impacts even without any drought frequency or intensity changes. The 2007 Fourth Assessment Report of the Intergovernmental Panel on Climate Change states that “drought-affected areas are projected to increase in extent, with the potential for adverse impacts on multiple sectors, e.g., agriculture, water supply, energy production and health.” In addition, the United States Global Change Research Program assessment states that droughts are likely to become more common and more intense as regional and seasonal precipitation patterns change, and rainfall becomes more concentrated into heavy events.

Therefore, projected changes in climate, population growth, and increased water demands suggest that future droughts will be more severe and have greater impacts than those experienced during the past century.

Summary

Drought is an insidious natural hazard that is a normal part of the climate in virtually all regions. It should not be viewed as only a physical phenomenon. Rather, drought is the result of an interplay between a natural event and the demand placed on water supply by human-use systems. Drought should be considered relative to some long-term average condition of balance between precipitation and evapotranspiration. Many definitions of drought exist; it is unrealistic to expect a universal definition to be derived. The three characteristics that differentiate one drought from another are intensity, duration, and spatial extent. The impacts of drought are diverse and generally classified as economic, social, and environmental. Impacts ripple through the economy and may

linger for years after the termination of the drought episode. Because of the large number of groups and economic sectors affected by drought, the nonstructural nature of its impacts, its spatial extent, and the difficulties in quantifying environmental damages and personal hardships, the precise calculation of the financial costs of drought is difficult. It appears that societal vulnerability to drought is escalating in both developing and developed countries, and at a significant rate. It also appears that future climate changes may lead to more frequent and severe droughts. Therefore, it is imperative that increased emphasis be placed on mitigation, preparedness, and prediction and early warning if society is to reduce the economic and environmental damages associated with drought and its personal hardships.

See also: **Climate and Climate Change:** Climate Variability: Seasonal and Interannual Variability; Overview. **Global Change:** Climate Record: Surface Temperature Trends; Upper Atmospheric Change. **Hydrology, Floods and Droughts:** Groundwater and Surface Water; Palmer Drought Severity Index; Soil Moisture. **Tropical Meteorology and Climate:** El Niño and the Southern Oscillation: Observation; El Niño and the Southern Oscillation: Theory. **Weather Forecasting:** Wildfire Weather.

Further Reading

- Dai, A.G., 2011. Drought under global warming: a review. *Wiley Interdisciplinary Reviews Climate Change* 2, 45–65.
- Glantz, M.H. (Ed.), 1994. *Drought Follows the Plow: Cultivating Marginal Areas*. Cambridge University Press, Cambridge.
- Quiring, S.M., 2009. Monitoring drought: an evaluation of meteorological drought indices. *Geography Compass* 3, 64–88.
- Quiring, S.M., 2009. Developing objective operational definitions for monitoring drought. *Journal of Applied Meteorology and Climatology* 48, 1217–1229.
- Vogt, J.V., Somma, F. (Eds.), 2000. *Drought and Drought Mitigation in Europe*. Advances in Natural and Technological Hazards Research. Kluwer Academic, Dordrecht.
- Wilhite, D.A. (Ed.), 2000. *Drought: A Global Assessment*. Hazards and Disasters: A Series of Definitive Major Works, vols. 1 and 2. Routledge, London.

Flooding

CA Doswell III, University of Oklahoma, Norman, OK, USA

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

Flooding occurs when water encroaches on normally dry land. Floods can arise directly from atmospheric precipitation or from events that cause precipitation that had fallen earlier to inundate dry land. Flooding is a hazard that occurs worldwide and can have enormous societal impacts. As people encroach on flood-prone regions and urbanization occurs, they become increasingly vulnerable to the dangers of flooding.

Introduction

Flooding is the weather-related hazard that is arguably the most widespread around the globe. It can occur virtually anywhere. A flood is defined as water overflowing onto land which is typically dry. Flooding is often thought of as a result of heavy rainfall, but floods can arise in a number of ways that are not directly related to ongoing weather events. Thus, a complete description of flooding must include processes that may have little or nothing to do with meteorological events. Nevertheless, it is clear that in some ultimate sense, the water involved in flooding has fallen as precipitation at *some* time, perhaps long ago. The origins of flooding, therefore, ultimately lie in atmospheric processes creating precipitation, no matter what specific event causes the flooding.

Floods produce damage through the immense power of moving water and through the deposition of dirt and debris when floodwaters finally recede. People who have not experienced a flood may have little or no appreciation for the dangers of moving water. The energy of that moving water goes up as the square of its speed; when the speed doubles, the energy associated with it increases by a factor of four. Flooding is often coupled to water moving faster than normal, in part because of the weight of an increased amount of water upstream, leading to an increase in the pressure gradient that drives the flow. In most cases, the damage potential of the flood is magnified by the debris the waters carry, which includes trees, vehicles, boulders, buildings, etc. When the water moves fast enough, it can sweep away all before it, leaving behind scenes of terrible destruction (Figure 1).

The effect of the water itself can be devastating on structures and on the objects within them: books, furniture, photographs, electronic equipment, etc. can be damaged simply by being immersed in water, even if they are not directly damaged by the water movement. Moreover, floodwaters typically contain suspended silt, potentially toxic microorganisms, and dissolved chemicals. This means that floods usually compromise drinking water supplies, resulting in short-term shortages of potable water, with the additional long-term costs in restoring drinking water service to the residents of a flooded area. The mud and debris left behind when floodwaters recede can be costly to clean up and represent a health hazard, as well, especially when there are decomposing bodies in the debris. In some situations, floods drive wild animals (including invertebrates of all sorts) from their normal habitats to human

habitations near and within the flooded areas, which can create various problems, especially when those animals are venomous or aggressive.

Although flooding has some large negative impacts on humans, it is also part of the natural processes shaping the Earth. Floodplains along rivers and streams are among the most fertile regions known through deposition of nutrients and relatively high soil moisture content. Most of the so-called cradles of civilization are within floodplains for this very reason (e.g., the Nile River, the Tigris–Euphrates River, among others). Hence, humans have been affected by flooding both positively and negatively since before historical times, whenever they find themselves in the path of these natural events.



Figure 1 Damage resulting from the 1977 Johnstown, Pennsylvania, flash flood event. © The Johnstown Tribune-Democrat, used by permission.

Floods as a Direct Result of Precipitation

When the waters of a flood arise directly from precipitation, atmospheric processes can be identified which are responsible for the event. That is, rainfalls occur that are usually well beyond the average values for the affected area. It is only when those rainfalls exceed the average that land which is usually dry can be affected; that is, a flood occurs. Thus, the rainfall amounts needed for floods cannot be defined in absolute terms. A precipitation event that causes a flood in one location might be well within the bounds of what is typical for another location. Generally speaking, the threshold for flood-producing rainfalls increases as the annual average rainfall for a region increases.

Flash Floods

Flash floods are defined as those flood events where the rise in water is either during or within a few hours of the rainfall that produces the rise. Therefore, flash floods occur within small catchments, where the response time of the drainage basin is short. In part owing to the rapidly rising, fast-moving waters of a flash flood, the damage from them can be devastating (Figure 1). Many hydrological factors have relevance to the occurrence of a flash flood: terrain gradients, soil type, vegetative cover, human habitation, antecedent rainfall, etc. In steep, rocky terrain or within heavily urbanized regions, even a relatively small amount of rainfall can trigger flash flooding. These hydrological factors determine the response of the catchment to the precipitation event. Thus, a flash flood is clearly the result of the concatenation of both meteorological and hydrological circumstances.

Most flash floods associated with rainfall are produced by thunderstorms; that is, deep, moist convection. A single thunderstorm cell is unlikely to produce enough rainfall to cause a flash flood, so the typical flash flood is the result of several thunderstorms moving successively over the same area, known as 'training' thunderstorms (Figure 2), because it resembles the passage of cars in a train. A succession of thunderstorms results when new thunderstorms pass repeatedly over the same place while the overall system of thunderstorms is very nearly stationary. The infamous Johnstown, Pennsylvania flash flood of 19–20 July 1977 was produced by such a system. Thunderstorms forming in northwestern Pennsylvania moved south-eastward, only to be replaced by newly formed thunderstorms, a process that went on for several hours. The result was torrential rainfall concentrated near Johnstown, with amounts exceeding 400 mm. The ensuing flash flood was responsible for 77 fatalities and \$550 million (in 1999 dollars) in damage.

Occasionally, flash floods are created in conditions that are not favorable for thunderstorms but which still produce heavy rainfalls. This can occur when moist air is forced upward over mountains by the wind flow, called *orographic* precipitation. Whenever the air forced upward is very moist, the rainfall can be quite heavy. The steep, rocky terrain also promotes rapid runoff of the rainwater. Flooding along the West Coast of the United States or in the European Alps is often of this type; that is, not involving thunderstorms.

A characteristic of flash floods is the localized nature of the heaviest rainfall. As shown in Figure 3, the most intense

rainfall is typically confined to a relatively small area. When large amounts of this localized precipitation fall within a small drainage basin, flash floods can occur. Sometimes, the location where flash flood damage occurs may actually receive little or no rainfall; that is, the rainfall that causes the problem falls upstream of areas susceptible to damage from the flash flood. This separation between the rainfall and the flood can cause confusion because it may not even be raining in an area for which flash flood warnings are issued. Another factor in the impact of flash floods is that the precipitation causing the event often falls during the night, when it can be difficult to give warnings to sleeping residents. The central part of the United States is well known for its heavy thunderstorm-produced rains during nighttime hours. Worldwide, thunderstorms are most common during the day, but on the central plains of the United States (and in a few other places around the world), the unique geography of the region favors nocturnal thunderstorms. This setting promotes a strong poleward flow of air near the surface from the Gulf of Mexico, called a *low-level jet stream*, during the warm months of the year. Moisture carried by the low-level jet stream helps to maintain thunderstorm systems that often begin during daytime hours on the higher terrain to the east of the Rocky Mountains. Because of the low-level jet stream, such storms can persist well into the nighttime hours, often forming clusters of thunderstorms known as *mesoscale convective systems* (Figure 4).

It is the rapidity of the event that makes flash floods so damaging and dangerous. Flash floods involve rapidly rising, fast-moving waters that can do immense damage; the suddenness of the onset of the flood can result in people being caught unawares and unprepared. Most fatalities result from drowning, with perhaps some traumatic injuries from being carried along in the debris-laden waters and being swept into stationary objects. The potential for loss of human life with flash floods is high. Debris carried in flash floods can form temporary 'debris dams' that typically fail as waters back up behind them. Failure of these debris dams then results in a 'wall of water' surging downstream. Debris dam failure events can happen repeatedly during the course of the flash flood. Not all flash floods are characterized by a 'wall of water' but all of them (by definition) involve rapidly rising floodwaters.

Flash flooding is more likely in cities than in rural areas surrounding a city, because urbanized areas promote runoff of rainfall, rather than permitting most of the rain to be absorbed into the ground. It takes much less rainfall in a city to create a flash flood situation than in a rural area of comparable size.

Flash floods continue to be a major contributor to loss of life, in spite of improved precipitation forecasting. Some noteworthy examples include events in the Big Thompson Canyon in Colorado (1976 – 144 fatalities) and near the town of Biescas in the Spanish Pyrenees (1996 – 86 fatalities).

Tropical cyclones often create devastating flash floods as a result of torrential rainfalls. In late October of 1998, Hurricane Mitch caused more than 9000 fatalities (the exact number is not known), mostly in Nicaragua and Honduras, in Central America, from flash floods and landslides associated with its rainfall. It was the worst weather disaster in terms of

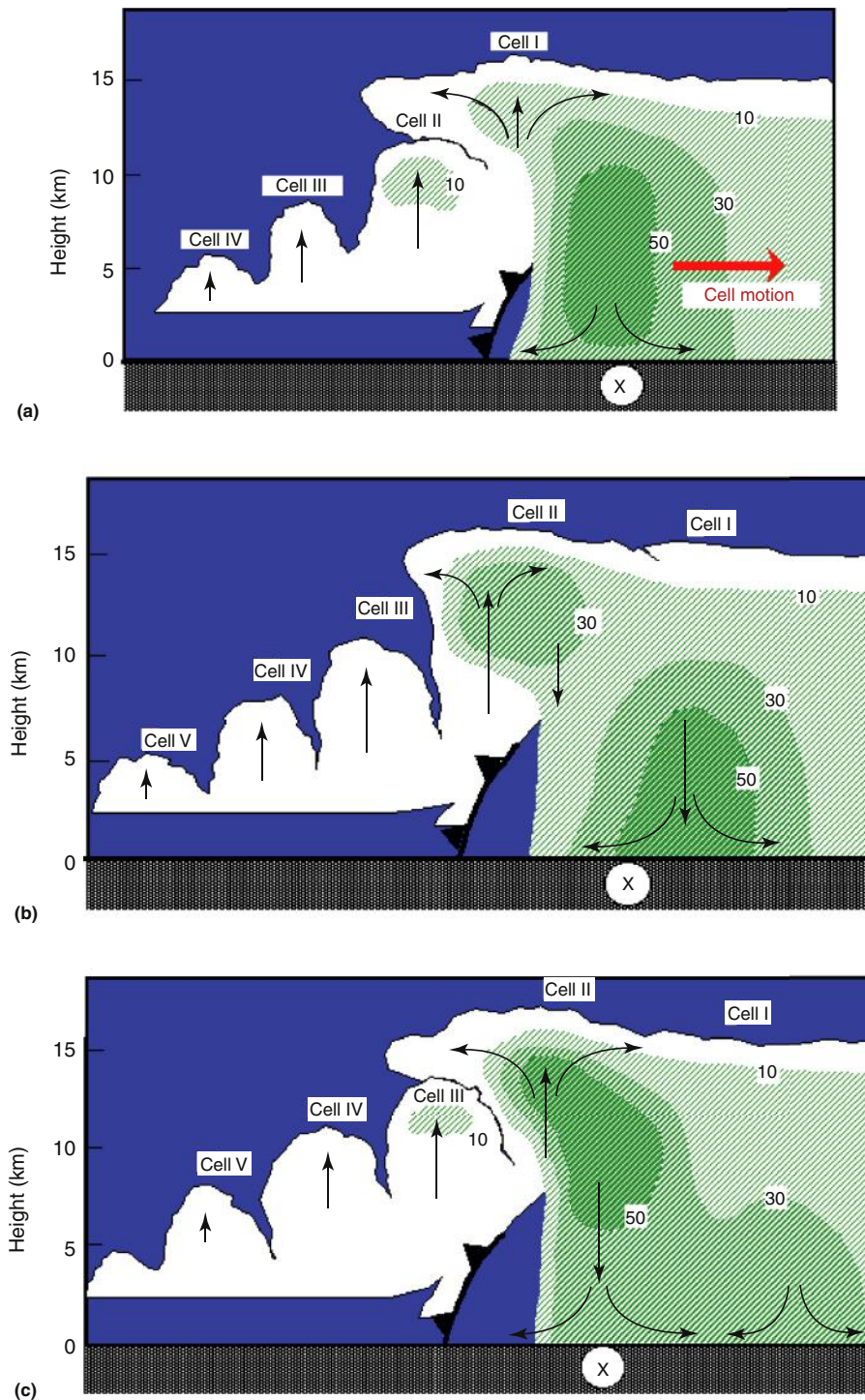


Figure 2 Schematic of the 'training' effect. (a) At this time, there are four numbered thunderstorm cells in various stages of development. Cell I is mature, with both updrafts and downdrafts, and heavy rain is about to commence at point 'X.' Cells II, III, and IV are still developing, and have only updrafts. Cell II has precipitation forming aloft. The hatched contours are radar reflectivity, in units of dBz, which is related to the rainfall rate. (b) About 15 min later, Cell I's updraft is dissipated, and it is now dominated by downdraft. Heavy rain continues at 'X' while Cell II is maturing and developing a downdraft. Cells III, IV, and now V are still immature. (c) About 15 more minutes have elapsed. Cell I's rainfall is continuing but it is now nearly dissipated, while Cell II is entering late maturity. It is still raining at 'X' but now the rainfall is from Cell II, and heavy rain from Cell II is descending from aloft. Now Cell III is developing its first precipitation aloft. Cells IV and V are still immature. Adapted from Figure 7 in Doswell, C.A. III, Brooks, H.E., Maddox, R.A., 1996. Flash flood forecasting: an ingredients-based methodology. *Weather Forecasting* 11, 560–581.

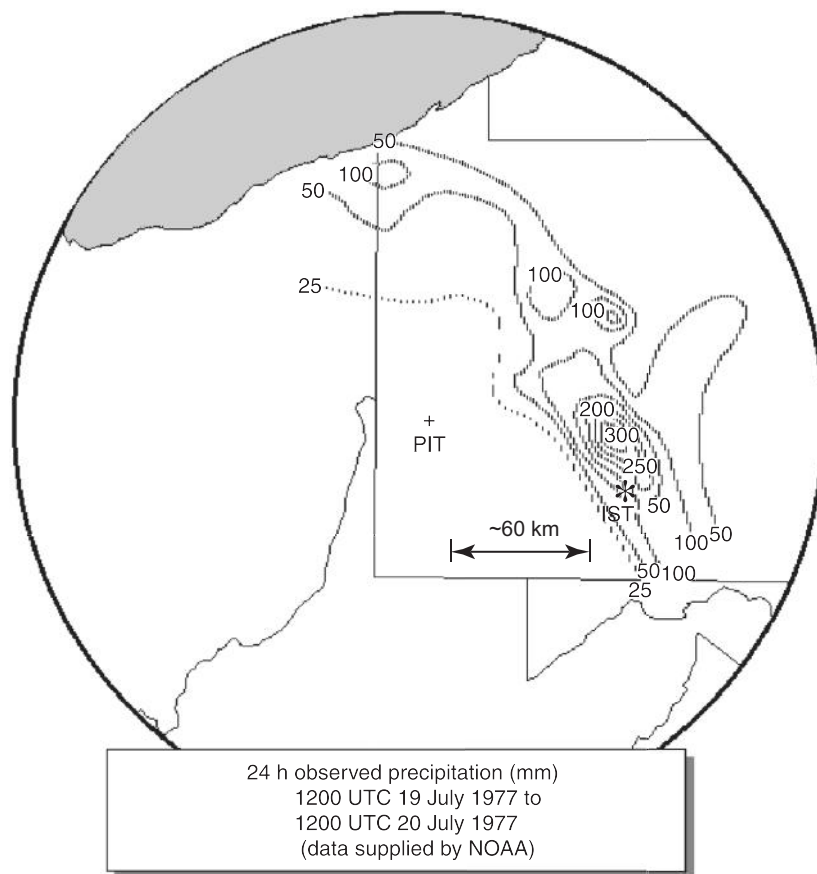


Figure 3 Observed total precipitation (mm) during the Johnstown, Pennsylvania (JST, located by an asterisk), flash flood event. For reference, Pittsburgh, Pennsylvania (PIT, located by the plus sign), is also shown. Adapted from Figure 14(a) in Hoxit, L.R., Maddox, R.A., Chappell, C.F., Zuckerberg, F.L., Mogil, H.M., Jones, I., Greene, D.R., Saffie, R.E., Scofield, R.A., 1987. Meteorological Analysis of the Johnstown, Pennsylvania, Flash Flood, 19–20 July 1977. NOAA Tech. Rept. ERL 401-APCL 43, NTIS Accession No. PB297412.

casualties in the Western Hemisphere during the twentieth century.

River Floods

River floods, in contrast to flash floods, typically unfold over days, or even months. This is because they occur in large basins involving 'main stem' rivers like the Missouri, or the Nile and are usually the result of many individual rainfall episodes spread out over many days. In fact, within a river flood event, numerous localized flash flood events can occur. Again, hydrological factors often contribute to a river flood, but river floods are not so sensitive to them as are flash floods. River floods are usually the result of a stagnant synoptic-scale weather pattern, whereas individual thunderstorm systems can cause flash floods. Localized heavy rainfall events occur many times during a period of days or even months, each contributing its share of rainfall to the tributaries, which then discharge into the main stem of a river. Saturated soils from repeated heavy rainfall promote runoff. The river rises gradually in response to all the input rainfall. The river flood potential of a situation can be increased by concurrent snow-melt and other factors besides rainfall.

The major flooding event during June and July of 1993 was the result of a weather pattern (Figure 5(a)) that produced a storm track across the upper Midwestern United States. Abnormally low heights of the pressure surfaces (associated with cool temperatures) over the Northern Plains produced a pattern in which traveling weather disturbances intensified in the Midwest after crossing the Rocky Mountains. This pattern aloft also produced an anomalously strong poleward flow of low-level moisture from the Gulf of Mexico into the Midwest. Mesoscale convective systems developed almost every evening during the early summer, typically persisting through the night. These passed repeatedly over the nearly the same areas, resulting in widespread significant rainfalls (Figure 5(b)) for the period over the Lower Missouri and Upper Mississippi basins. In addition to these factors, considerable rainfall had fallen over the region during the previous several months, providing a hydrological setting that favored runoff of the precipitation. This event produced disastrous flooding that persisted for many weeks.

Owing to the long timescale of the rising waters, river floods pose a lower risk of fatalities; people have more time to take proper actions. Of course, some casualties result from waiting

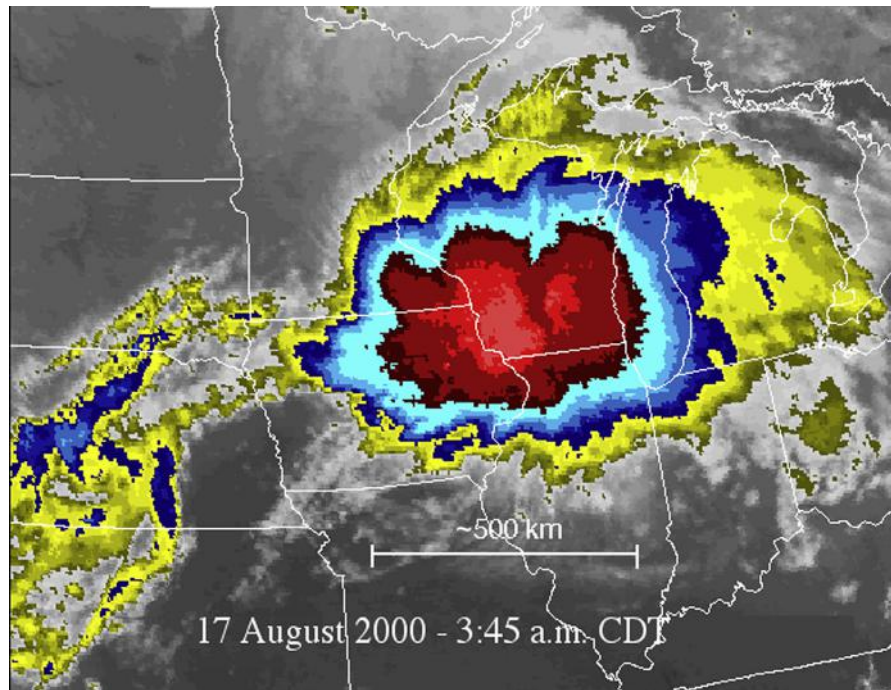


Figure 4 False-color enhanced infrared satellite image of a mesoscale convective system, with the light red colors indicating the coldest (therefore the highest) clouds. Note that this image is at 0345, local time, which corresponds to 0845 UTC. Adapted from NOAA image.

until it has become too late to respond to the threat. Levee and dam failures, as well as intentional rapid release of impounded waters to prevent the catastrophic failure of the flood-control structures, can produce rapidly rising water situations embedded within a river flood, and these also can contribute to loss of life.

Because of the large scale of river floods, the damage figures may be enormous; easily into billions of dollars. Crop losses are a major factor in the costs of river floods, whenever large tracts of prime agricultural land along floodplains are inundated. Levees are often used to protect populated areas, so the failure of those levees can generate major property losses. The damage and dislocations along the Upper Mississippi and Lower Missouri basins during the summer floods of 1993, during which several levees were breached, illustrate the huge impact such events can have.

Floods Arising from Nonprecipitation Events

Apart from floods resulting directly from rainfall, there are many ways that precipitation can cause floods, perhaps long after it has fallen. When flowing water is impounded by the construction of dams, there is some risk that the dams will fail. For example, Johnstown, Pennsylvania was inundated by a dam failure during a rainfall event in 1889. Such rapid releases of stored water can be cataclysmic, manifesting themselves as an enormous 'wall of water' choked with debris.

Floods also can arise through the melting of snowfall. In situations where the preceding winter's snowpack is deep, a sudden change to warm temperatures in the spring can result

in abnormally rapid melting and runoff of the snowmelt. The devastating flood created in Grand Forks, North Dakota in April of 1997, is an example. Occasionally, warm rain falls directly onto the melting snow, exacerbating such situations by speeding the melting process and adding more liquid water.

Deposits of snow and ice on volcanic peaks can melt rapidly during eruptions. The resulting runoff, often turned into a thick slurry by the inclusion of volcanic ash, roars down the mountainside and is called a *lahar*. A tragic example occurred with the Nevado del Ruiz volcano in Colombia on 13 November 1985, which killed more than 23 000 people, mostly in the town of Armero. Another occurred in Iceland in 1996 on the Vatnajökull glacier, with no fatalities owing to its remote location. Lahars can continue occasionally for years after an eruption, when heavy rains fall onto ash deposited by the volcano.

During the winter and late spring, ice can build up on rivers in cold climates. The breakup of the ice can create *ice dams* on the river. These ice dams cause the waters to back up, sometimes flooding the land upstream of the ice dam. Then, the breakup of the ice dam can result in a flash flood wave that surges downstream of the ice dam's position.

Other flood situations can develop along the shores of the world's oceans and even with large freshwater lakes. *Tsunamis*, such as the tragic event on 26 December 2004 caused by a major earthquake in the Indian Ocean near Indonesia, which also can be initiated by landslides (both above and below the water), flood coastal areas with multiple surges of water that can penetrate deep inland in flat coastal zones. The effects of tsunamis can be experienced far from their source, as

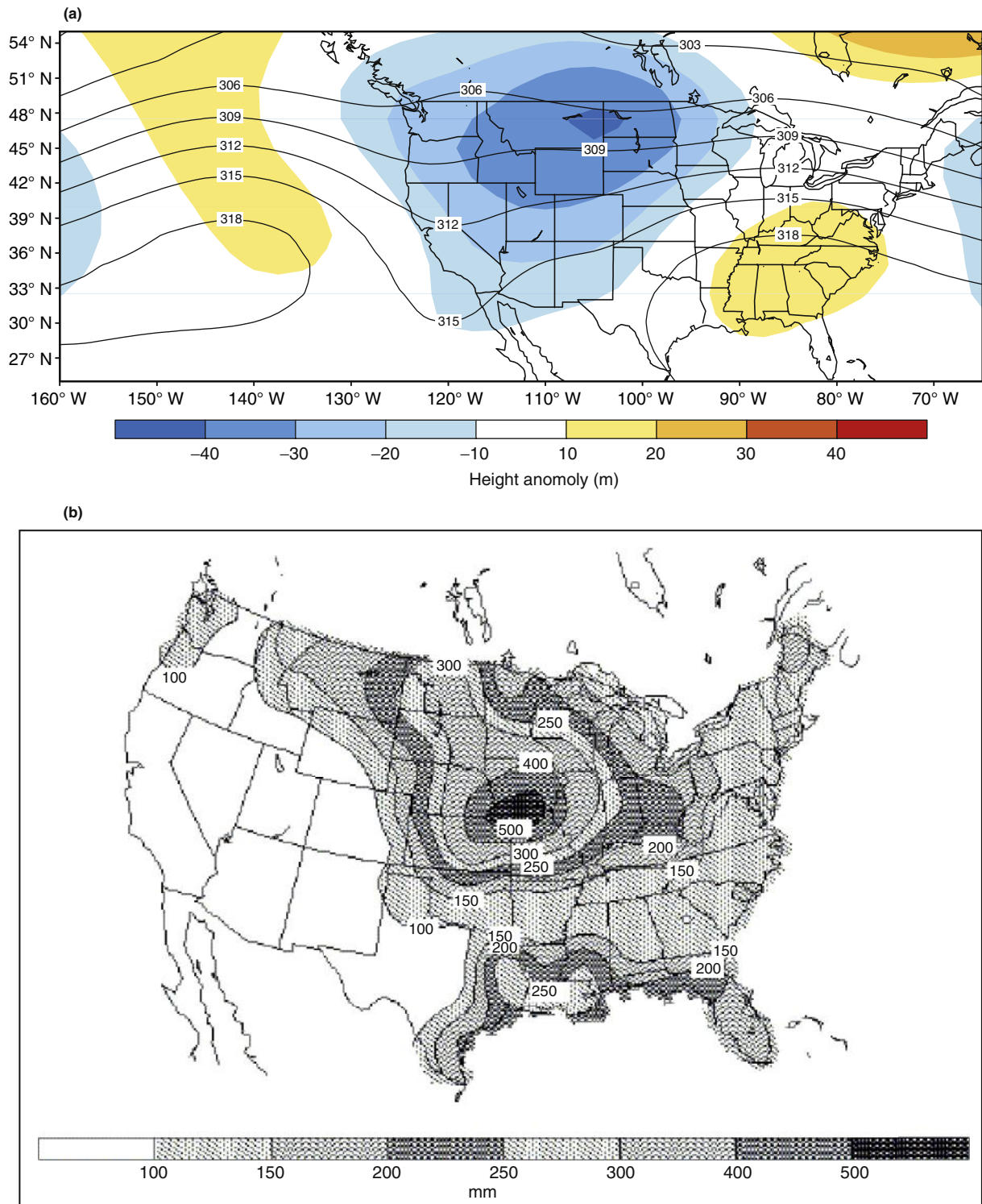


Figure 5 (a) Map of the 700-hPa heights (thin lines, in dam) and height anomalies (shading, in m) for June/July 1993, and (b) observed precipitation for the same period (shading, in mm). Based on data supplied by Janowiak, J., of the NOAA Climate Prediction Center.

illustrated in Figure 6. Storms of all sorts, including tropical cyclones, can drive the waters before the winds into *storm surges* that inundate coastal zones when the storms are near the land. Large lakes, such as the Great Lakes in the North

Central United States can experience flooding due to *seiches*, which are surges of water (usually wavelike) within enclosed bodies of water. Seiches can be caused by earthquakes or by atmospheric processes.

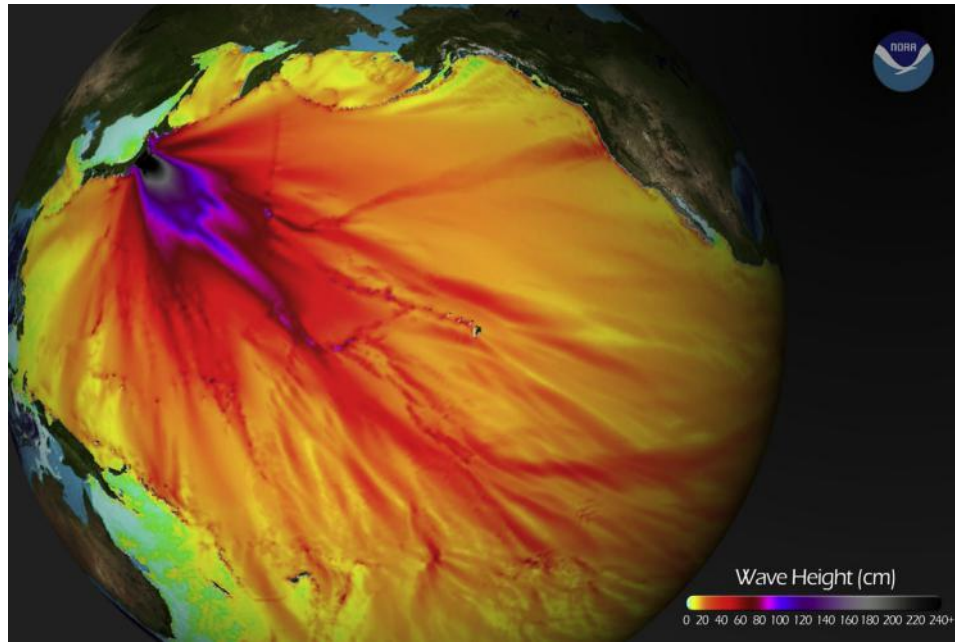


Figure 6 Estimated wave heights derived from a computer model of the Pacific Ocean basin associated with the tsunami caused by a powerful earthquake under the ocean near Japan on 11 March 2011. Adapted from NOAA image.

Societal Impacts and Their Mitigation

The results of floods on society worldwide are substantial. Flooding is responsible for many drowning fatalities in tropical cyclones, either from storm surges or from freshwater rain-induced flash floods. Flash floods and river floods typically produce more fatalities on the average than either tornadoes or hurricanes in the United States. In many parts of the world, flood fatalities are associated with the most significant weather-related disasters. Flood damage cost in the United States is now on the order of several billion dollars annually and the numbers continue to rise.

Many people live and play in flood-prone areas; for example, within floodplains of rivers and their tributaries, as well as along coastlines that are vulnerable to storm-caused flooding from tsunamis, tropical cyclones, and nontropical storms. Development of flood-prone areas for habitation and recreation has been increasing, with a corresponding increase in the risks to life and property. In 1993 and again in 2011, Mississippi River floods provided a grim reminder of the risks of building permanent structures within floodplains, even when flood-control measures have been taken.

In the case of flash floods, it is difficult to take measures to protect property, owing to the rapidity with which the event happens. However, prevention of flash flood casualties is possible, provided warnings can be issued and acted upon properly in a timely fashion. Considerable attention has been paid to increasing public awareness of the dangers of driving into rapidly rising floodwaters, for instance, as a result of recent experiences with flash floods. Unfortunately, situations can still arise where warnings are not issued in time. People living and engaging in recreational activities in places prone to flash floods need to be alert during heavy rainfalls and be

prepared to seek safety even when they do not receive timely warnings.

For river floods and other relatively slow-developing situations (such as rising snowmelt or ice action events), it may be possible to reduce the property damage as well by removing the contents of structures. Obviously, any structures (and their contents) built in flood-prone areas are permanently at risk; the only way to guarantee them not being affected by floods is to move them out of those areas. Prevention of fatalities in river flood events is a matter of heeding the warnings of danger and getting residents out of the danger areas before the number of options is reduced by the rising waters and by the breaching of levees or other flood-prevention structures.

Forecasting the details of flooding events is an important part of mitigation. Knowing precisely when and where a flood will occur would no doubt be helpful, but it is also important to be able to anticipate the magnitude of the flood. An example of this is the tragedy of the 1997 Grand Forks, North Dakota case, where the river level was only a few feet higher than that forecast. Those few feet, however, had a large impact, because the flood-control operations were based on the lower forecast value. When the river rose above that level, the flood-control measures failed catastrophically. In reality, such a forecast can never be a precise statement; uncertainty is implicitly a part of every forecast, a point that deserves greater emphasis in the future.

Flooding, by its very nature, is usually a result of both meteorological and hydrologic processes; the character of a flood is determined both by the detailed behavior of the precipitation and by the nature of situation in which the event is likely to occur (soil conditions, amount of antecedent rainfall, etc.). It is not likely that precisely detailed forecasts of flooding events will ever be possible, although it certainly is

well within our capability to anticipate the *possibility* of most flood events. The challenge for reducing the social impacts of floods is how best to make use of the *uncertain* meteorological and hydrological forecasts that are within practical means. The requirement is to make effective use of whatever forecasting capability we have, even as we seek to improve that capability.

Effects of Human Activities on Flooding

In addition to the risks to lives and property that people take by moving into flood-prone areas, development for human use often involves clearing land of its native vegetation and altering the characteristics of the ground cover. Vegetation works together with the soil to store rainfall, so when that vegetation is cleared, rainfall runoff can increase substantially. Rather than being absorbed by the soil and its natural vegetation, in areas where that vegetation has been cleared (either for construction or for agriculture), heavy rainfall is more likely to run off and pour into streams and rivers, increasing the potential threat from flash floods and river floods. Construction of roads and buildings also acts to increase runoff, replacing soil with impervious materials. Such construction increases dramatically the fraction of the rainfall that runs off, regardless of antecedent rainfall. Human-caused fires also can produce at least temporary increases in the runoff potential in the headwater regions of streams and rivers. It is evident that human activities are increasing the potential for floods around the world.

Again recalling the Mississippi River floods of 1993 as an example, the issue of flood control through levees and other structures was dramatically recalled to public attention. The value of structural methods for flood control (levees, flood-control dams, breakwaters, etc.) remains controversial, but the 1993 floods made it apparent that structures such as levees *can* be breached during *major* flooding episodes, even though they may be able to contain lesser events. Structural failures create rapidly rising waters (flash floods) artificially within a river flood event, increasing the hazards to human life as well as destroying property. The decision about when and where to take structural approaches will continue to be a challenge.

Finally, the use of flood-prone areas for human activities puts lives and property at risk, although the major flood events may be separated by many years. The relatively long time between events can lead to complacency and subsequent

disasters. The choices associated with land use are a continuing challenge, now and in the future. When humans live and play in ways that put them in the path of potential floodwaters, major societal impacts are inevitable.

See also: **Air Sea Interactions:** Momentum, Heat, and Vapor Fluxes. **Data Assimilation and Predictability:** Predictability and Chaos. **Hydrology, Floods and Droughts:** Palmer Drought Severity Index; Soil Moisture. **Mesoscale Meteorology:** Cloud and Precipitation Bands; Convective Storms: Overview; Mesoscale Convective Systems; Severe Storms. **Radar:** Precipitation Radar. **Satellites and Satellite Remote Sensing:** Precipitation. **Tropical Cyclones and Hurricanes:** Hurricanes: Observation. **Weather Forecasting:** Severe Weather Forecasting.

Further Reading

- Agnone, J.C. (Ed.), 1995. Raging Forces: Earth in Upheaval. National Geographic Society, Washington, DC.
- Anthes, R.A., 1985. Tropical cyclones – their evolution, structure, and effects. American Meteorological Society, Boston, MA.
- Barry, J.M., 1997. Rising Tide: The Great Mississippi Flood of 1927 and How It Changed America. Simon and Schuster, New York.
- Brutsaert, W., 2005. Hydrology: An Introduction. Cambridge University Press, New York, 605 pp.
- Cluckie, I.D., Collier, C.G. (Eds.), 1991. Hydrological Applications of Weather Radar. Ellis Horwood, New York.
- Doswell III, C.A. (Ed.), 2001. Severe Convective Storms. American Meteorological Society, Boston, MA.
- Fradin, J., Fradin, D., 2005. Tsunami: Witness to Disaster. National Geographic Society, Washington, DC.
- Hill, C.E. (Ed.), 1986. Nature on the Rampage: Our Violent Earth. National Geographic Society, Washington, DC.
- Lorenz, E.N., 1993. The Essence of Chaos. University of Washington Press, Seattle, Washington, DC.
- Ludlam, F., 1980. Clouds and Storms. Pennsylvania State University Press, University Park, PA.
- Markowski, P., Richardson, Y.P., 2010. Mesoscale Meteorology in Midlatitudes. Wiley-Blackwell.
- Sarewitz, D., Pielke Jr., R.A., Byerly, R. (Eds.), 2000. Prediction: Decision-Making and the Future of Nature. Island Press, Covelo, CA, 405 pp.
- Schuster, R.L., Lynn, M., 2001. Socioeconomic and Environmental Impacts of Landslides in the Western Hemisphere. U.S. Geological Survey.

Groundwater and Surface Water

S Ge, University of Colorado, Boulder, CO, USA

SM Gorelick, Stanford University, Stanford, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

Groundwater and surface water are two important components of the hydrologic system. One objective of hydrologic studies is to understand the spatial and temporal variations of water storage and movement. This article first presents an overview of the global distribution of water, the hydrologic cycle, and the water balance concept. It then describes the physical principles governing the movement of groundwater and surface water by focusing on the hydraulic gradient as the driving force for flow: Darcy's law for groundwater and Manning's equation for streams. Finally, applied aspects of hydrology are discussed as they relate to water contamination, land subsidence, hydrothermal fluids, and hydroseismicity.

Introduction

Water is one of the most precious and indispensable natural resources. As important components of the hydrologic system on Earth, groundwater and surface water impact numerous aspects of Earth processes and many facets of daily life. Water at the Earth's surface directly interacts with the atmosphere, and water in the subsurface continuously redistributes geothermal energy and dissolved minerals in the Earth's crust at a variety of temporal and spatial scales. Hydrology encompasses the study of the occurrence and movement of water both at the land surface and in the subsurface. Although the term groundwater usually refers to the water that occurs beneath the water table in saturated soils and rocks, study of soil moisture movement in the unsaturated zone above the water table is well within the realm of groundwater studies.

Focusing on the physical dynamics of groundwater and surface water, this article first presents a brief overview of water as a resource. The discussion is then devoted to the main concepts and governing principles applied to physical processes controlling the movement of groundwater and surface water. Finally, the applied aspects of hydrology as they relate to water contamination, land subsidence, and geological processes are briefly reviewed.

Water as a Resource

The presence of abundant liquid water makes the Earth a unique planet in the solar system. This abundance has been challenged throughout human history as numerous local and regional conflicts over water resources have erupted. Evidence of early attempts to harness water for human purpose has been documented by archaeologists. For example, clever water usage for irrigation can be traced as far back as 4500 years ago in the Middle East. A remarkable water-collecting tunnel system dating from around 500 BC in Egypt has been unearthed. As world population grows, the local demand for water is expected to grow, particularly in arid developing countries. Shown in [Figure 1](#) are the historical and anticipated

global water extractions from 1900 to 2025 and how water use is partitioned among agricultural, domestic, and industrial uses. Increasing trends are observed in all three sectors and in the total extraction (bar graphs in [Figure 1](#)). The partitioning of water (pie charts in [Figure 1](#)) indicates an increased share in domestic use from 14% in 1900 to forecasted 24% in 2025 but a decreased share in agricultural use from 80% in 1900 to forecasted 60% in 2025. The increased domestic share and the decreased agricultural share may suggest food shortage problems in the future if crop production water use efficiency cannot keep up with global population growth and increased demand for food. Total freshwater withdrawal in the United States from 1950 to 2005 also followed a generally increasing trend ([Figure 2](#)). The decrease in the 1980s was primarily due to increased irrigation efficiency and reduction in water consumption by the thermoelectric power industry as a result of improved power plant technologies and efficiencies. The continual increase in domestic water use has compensated for the decrease in industrial water use.

Global Water Distribution

[Figure 3](#) shows the water distribution in the Earth system. Of all water on the Earth, 97.33% of it is stored in the ocean and is too salty to be directly used for human consumption. Ice caps and glaciers, the next largest water reservoir, hold 2.12% of the global water and account for 79% of the total freshwater. Groundwater in the upper 800 m of the subsurface holds 0.31% of the water on the Earth. Only the portion in the upper few hundred meters of the Earth's crust is economically accessible and fresh enough for human consumption. The salinity of groundwater increases with depth and often becomes too high to be useful as a resource below 1 or 2 km. Surface waters including lakes and streams hold 0.158% of the global water, and they have served as the main water resource for people, primarily owing to their easy accessibility. The water in the atmosphere, 0.083% of total water, is much more than that occurring in all of the streams (0.003%) plus all soil moisture (0.002%) in the world.

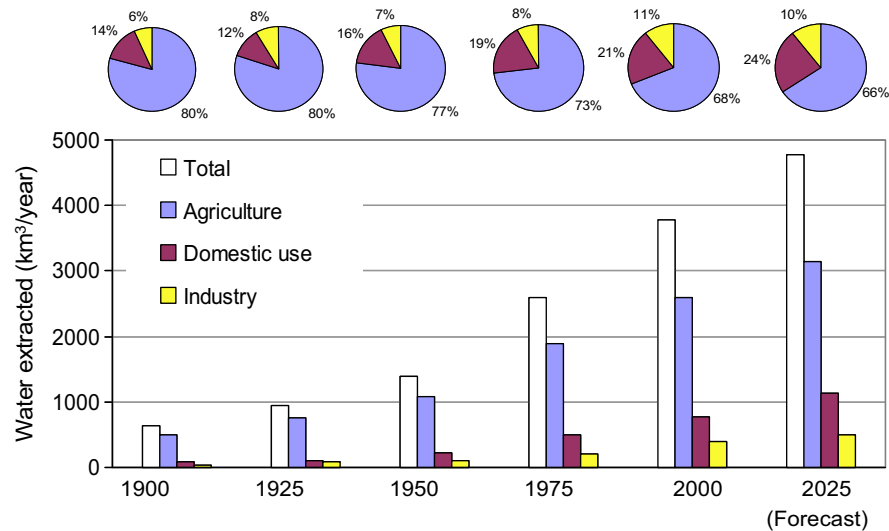


Figure 1 Global water extractions from 1900 to 2025 (bar charts) and the partition of water by different sectors (pie charts). Reproduced from UNEP, 2008. *Vital Water Graphics – An Overview of the State of the World's Fresh and Marine Waters*, 2nd edn. Nairobi, Kenya: UNEP, ISBN: 92-807-2236-0.

Hydrologic Cycle

Powered by solar energy, the hydrologic cycle is the endless movement of water from one part to another part in the Earth system (Figure 4). Water evaporates into the atmosphere from bare land surfaces and from open waters such as oceans and lakes. Plants also lose water to the atmosphere through transpiration. Evaporation and transpiration are collectively known as evapotranspiration. Ice and snow can lose water through sublimation, the process of changing solid water into vapor without first melting to liquid. Water falls back to the Earth's surface as precipitation in the form of snow and rain. Upon reaching the surface, water flows overland as runoff to streams or infiltrates into the subsurface. Some of that infiltrated water becomes recharge to the groundwater system. Groundwater moves

through geologic media, discharging to springs and surface waters. Much of the migrating groundwater and surface waters eventually make their way to the oceans. The rates of water flow in the hydrologic cycle vary spatially and temporally. Streamflow is relatively rapid, with residence times of days to months. This is in contrast to the residence time of groundwater, the time it remains in the subsurface since recharge, which varies from months in sediments at shallow depths to tens of thousands of years in rocks several kilometers deep in the Earth's crust.

Water Budget Balance

One of the primary objectives in studying groundwater and surface water is to understand the spatial and temporal variations of water storage and movement. The basic principle

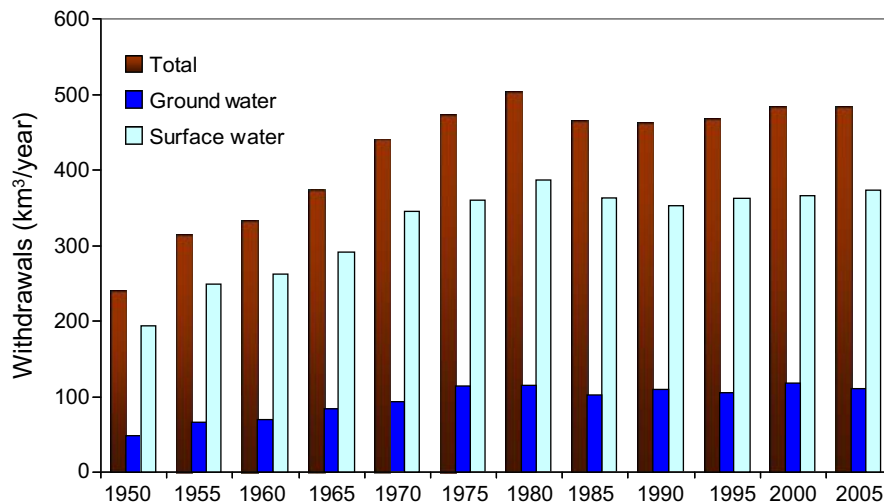


Figure 2 Freshwater withdrawals in the United States from 1950 to 2005. Reproduced from Kenny, J.F., Barber, N.L., Hutson, S.S., Linsey, K.S., Lovelace, J.K., Maupin, M.A., 2009. *Estimated Use of Water in the United States in 2005*. U.S. Geological Survey Circular 1344.

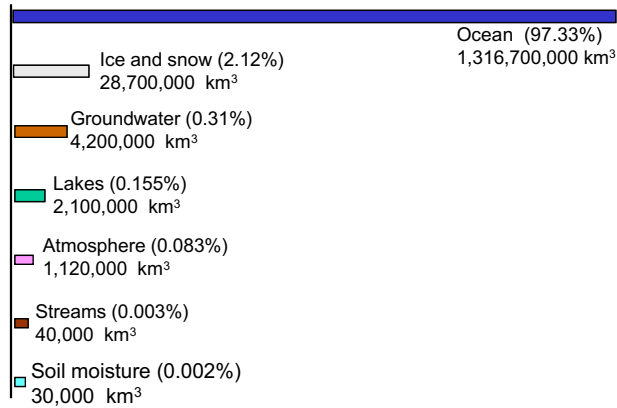


Figure 3 Global water distribution. Bar lengths are not to scale. Reproduced from Fierro, P., 2007. *The Water Encyclopedia*, Hydrologic Data and Internet Resources, 3rd edn. Boca Raton, FL: CRC Press. doi: 10.1201/9781420012583.ch4.

governing these variations is conservation of mass or water balance. This principle requires that the amount of water entering a control volume during a specified time period minus the amount of water exiting equals the change in storage in that volume. The principle of a water balance is often applied over a watershed, also referred to as a drainage basin, which is a geographic region where all the water flows to a common destination or an outlet. Watershed delineation is primarily based on topography where topographic divides, or ridges, form the boundaries of a watershed. Smaller watersheds or subbasins can be nested within a larger watershed. It should be noted that groundwater can flow across the boundaries of watersheds, under the topographic divides of the basin. Considering steady-state flows averaged over a long period of time, the water balance equation is:

$$P + Q_{in} + G_{in} = ET + Q_{out} + G_{out} \quad [1]$$

where P is the precipitation, Q_{in} and Q_{out} are stream inflow and outflow, respectively, G_{in} and G_{out} are groundwater inflow and outflow, respectively, and ET includes evaporation and transpiration. All quantities have the dimension ($L \, t^{-1}$).

It is a common misconception that the amount of groundwater that can be safely extracted from the subsurface at steady state is equal to some fraction of preextraction groundwater recharge to that system. However, as suggested by the water balance shown in eqn [1], any extraction of groundwater, adding to the right side of the equation, must be balanced by increased inflows on the left side or decreased outflows on the right side of eqn [1]. In general, any extraction of groundwater during development will reduce natural groundwater discharge to springs and surface waters over the long term, which may be undesirable to people and harmful to the environment.

Physical Hydrological Processes

Surface Water Dynamics

Streams, lakes, and wetlands are the surface waters of primary interest in hydrologic studies. The following discussion focuses on streams. A stream is a body of water flowing down slope along a more or less confined course. A river is a stream with a significant amount of flow. A stream with no tributaries is designated as a first-order stream, the confluence of two first-order streams is the beginning of a second-order stream; the confluence of two second-order streams is the beginning of a third-order stream, and this pattern can continue to form higher-order streams. The branching patterns of stream orders have been studied using a fractal approach that provides a mathematical framework for treatment of similar geometric characteristics over a range of scales. Streams play vital geologic roles in incising valleys, transporting dissolved load and sediments to the sea, and reshaping the landscape over time. Stream processes are affected by a variety of factors such as the steepness of the stream, the cross-sectional area of the stream, water velocity, and sediment load. The dimensionless Reynolds

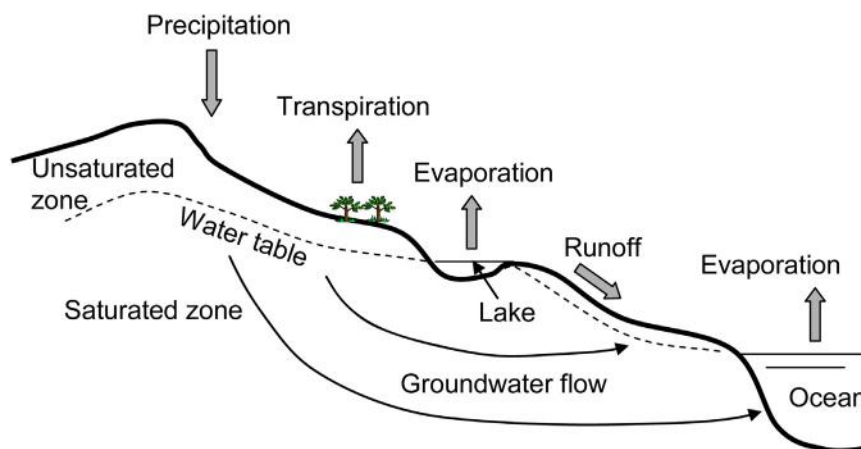


Figure 4 The hydrologic cycle: Water evaporates from open waters at the Earth's surface and transpires from vegetated lands. Upon reaching the land surface, precipitation infiltrates the soil to replenish groundwater as recharge or is removed by evapotranspiration. The remainder flows overland as runoff to open waters. Groundwater flows through the subsurface and returns to surface waters and oceans. The water table is the boundary between the unsaturated zone above and the saturated zone below. This particular diagram is most representative of humid regions where the water table is near the ground surface and streams and lakes are surface manifestations of the water table.

number, (Re), is a convenient parameter describing the state of flow as laminar or turbulent. It is defined as $(Re) = \rho vY/\mu$, where ρ is the density of water ($M L^{-3}$), v is the average flow velocity ($L t^{-1}$), Y is the average flow depth (L), and μ is the dynamic viscosity of water ($M L^{-1} t^{-1}$). In streams, laminar flow occurs when (Re) is less than 500. Turbulent flow occurs when (Re) is greater than 2000 and circulating eddies form in turbulent regions. Transitional flow lies between laminar and turbulent flow regimes. Actual streamflow is seldom laminar, but when the degree of turbulence is small flow is often considered to be in the laminar range. The Manning equation is one of the most commonly used equations for computing the average flow velocity in a stream channel:

$$v = n^{-1} R^{\frac{2}{3}} S^{\frac{1}{2}} \quad [2]$$

where n is the Manning roughness coefficient, R is the hydraulic radius (L) defined as the ratio of the flow channel's cross-sectional area to its perimeter, and S is the channel slope.

Water Table

In near-surface sediments and soils, water is in the form of moisture in the unsaturated zone where void spaces are partially filled with water. At depth, void spaces are completely filled with water, which forms the saturated zone. The boundary between the unsaturated and the saturated zone marks the water table (Figure 4), although a transitional tension-saturated region called the capillary fringe occurs at the interface. The configuration of the water table in humid regions is generally a subdued replica of the topography, near the surface in low lands but deeper at high elevations. The depth of the water table varies in both space and time. In humid regions, the water table can be at or near the surface and streams and lakes are surface expressions of the water table. In drier semiarid and arid regions, the water table can be hundreds of meters below the land surface. Under natural conditions, the water table can rise in wet seasons as excess precipitation percolates through the unsaturated zone to the water table and drop in dry seasons when water is lost through evaporation and transpiration. The most reliable way to locate the water table is to drill wells. Because lateral groundwater flow follows the slope of the water table, knowledge of the water table configuration gives useful basic information about the direction of groundwater flow.

Porosity, Permeability, and Hydraulic Conductivity

The primary factors controlling groundwater occurrence and movement in the subsurface are the hydraulic properties of the geologic material and the hydrologic driving force. The most important material properties are porosity and permeability. Porosity, defined as the percentage of pore space in a unit volume, may vary from 0–5% for tight igneous and metamorphic rocks to 25–50% for sands or fractured rocks. Clay can have a porosity as high as 90%, but few of the pores are interconnected. Porosity is a measure of geologic materials' capacity for storing water. A related quantity is known as specific yield, which is used to describe the volume of water that is partially drained by gravity from a porous material as the water table declines.

Permeability is a measure of geologic materials' ability to transmit water and reflects how well pores are interconnected. In the simplest case, it is defined as follows: $k = Cd^2$, where k is the permeability (L^2), C is a proportionality constant related to grain size, sediment sorting, and packing arrangement, and d is the particle diameter for which 10% of the grains by weight are finer (L). Depending on the type of fluid flowing through a system, the ease of flow differs. One can imagine that a fluid that is sticky like honey would flow much slower than would clean water in the same medium under the same conditions. Therefore, it is necessary to consider not only the medium but also the fluid properties. The hydraulic conductivity is introduced and is defined as: $K = k\rho g/\mu$, where K is the hydraulic conductivity ($L t^{-1}$), ρ is the fluid density ($M L^{-3}$), g is the gravitational acceleration ($L t^{-2}$), and μ is the dynamic viscosity of the fluid ($M L^{-1} t^{-1}$). Values of hydraulic conductivity vary over many orders of magnitude from $10^{-13} m s^{-1}$ for tight rocks to over $1 m s^{-1}$ for sands and gravels.

The most reliable means of obtaining hydraulic conductivity values are field-scale pumping tests using sets of observation wells. When water is withdrawn from or injected into a well, the rate of water level decline and recovery in the well and in adjacent observation wells can be measured by a pressure transducer and recorded by a data logger. The rate of water level decline during pumping, and recovery after pumping is stopped, is used to compute the hydraulic conductivity of the material in the vicinity of the pumping and observation wells. Large-scale tests with many observation wells are more reliable than single well tests, known as slug tests. This is largely because local material properties near the slug test well are not necessarily representative of the regional hydraulic conditions. Laboratory tests on core samples are commonly conducted to obtain hydraulic conductivity values, but they are typically valid at the centimeter scale and not necessarily representative of larger scales because hydraulic properties typically vary over small distances in natural groundwater systems. Computer modeling is also commonly employed as an indirect means of inferring hydraulic conductivity at different scales based on changes in groundwater levels due to regional addition or removal of groundwater over time and space.

Aquifers and Aquitards

Aquifer and aquitard are terms used to characterize hydrogeologic systems. A geologic unit that is highly permeable and can store and transmit a significant amount of groundwater is called an aquifer. When an aquifer is bounded by the water table on the top, the aquifer is called an unconfined aquifer. When an aquifer is confined between two much less permeable units, it is called a confined aquifer. Water pressure in confined aquifers is usually higher than pressure in unconfined aquifers. Thus, when a well is drilled into a confined aquifer, the water level in the well will rise to above the top of the aquifer, and may even rise above the ground surface, which creates artesian flow. An aquitard, also known as a confining bed, is a much less permeable geologic unit. Because no naturally occurring porous material is completely impermeable, aquifers and aquitards are identified to distinguish their relative degree of high and low permeability, respectively. In general, gravel,

sandy materials, limestone, or highly fractured rocks make good aquifers, whereas clay-rich, poorly sorted sediments, and unfractured rocks often form aquitards. The term aquiclude has been used for describing an impermeable unit, but this term has become obsolete.

Water in an Unsaturated Zone

Near the land surface at shallow depths, soils and sediments are often partially saturated, which creates a region known as an unsaturated or vadose zone. Saturation is defined as the fraction of pores that contain water, and it varies from 0 to 1, representing dry and fully saturated conditions, respectively. The water in an unsaturated zone clings onto solid particle surfaces and is sustained by capillary forces. Pore pressures in the unsaturated zone are conventionally expressed as negative values, reflecting the use of atmospheric pressure as the zero reference pressure. The pore pressure distribution and the rate of moisture movement vary spatially, depending on soil type, and temporally in response to short- and long-term climate conditions.

Infiltration is an important process in the unsaturated zone, which involves downward movement of moisture under wet conditions. The infiltration rate over a small area can be measured using a ring infiltrometer. A commonly used type of infiltrometer consists of two rings with a smaller inner ring nested inside a larger outer ring. The infiltrometer can be pushed several centimeters into the subsurface and both rings filled with water. The rate of water seepage into the ground from inside the inner ring is used to estimate the infiltration rate. The diameter of the rings of a portable infiltrometer varies from a few centimeters to 1 or 1.5 m. It is important to note that not all infiltrated water makes its way to the water table as recharge because some is temporarily stored in the near-surface sediments and subject to evapotranspiration.

In contrast to infiltration, evaporation and transpiration draw moisture upward. Evaporation not only causes water loss from surface waters, such as lakes and rivers, but also from near-surface soils and sediments. Water evaporates as a vapor diffusion process that is largely controlled by the energy exchange between radiation, or sensible heat from the atmosphere or ground, and the heat energy change in the evaporating body. A direct method to determine the evaporation rate is known as the pan evaporation approach. It involves exposing a cylindrical pan of water to the atmosphere in clearings where precipitation also can be monitored. The standard U.S. National Weather Service Class-A pan is 1.22 m in diameter and 25.4 cm deep. Transpiration is a process where water is lost to the atmosphere through the vascular systems of plants. The transpiration process works by absorption of water by plant roots, translocation of liquid through plant vascular systems, and transpiration into the atmosphere through stomata or openings in leaf surfaces. Although transpiration is considered a diffusion process, water is first pulled through the plant by a potential energy gradient, before diffusing into the air in response to a vapor pressure difference. There are a variety of methods to estimate transpiration. One is based on direct measurement using a lysimeter, a tank of soil, water, and plants in which the water loss due to transpiration can be assessed.

Another general approach is based on a surface energy balance that relies on a set of measurements to compute the latent heat flux, which is proportional to transpiration.

Groundwater in a Saturated Zone

Below the water table, the movement of groundwater occurs as slow-moving seepage through the pore spaces in sediments and rocks or as relatively fast flow through rock fractures and dissolution channels. Groundwater velocities are generally much slower than streamflow rates, and may be as low as 1 mm day^{-1} . Under natural conditions, a groundwater velocity on the order of centimeters per day would be considered typical and 1 m day^{-1} or more would be considered high. The low velocity of groundwater has important implications for geological processes like metamorphisms and ore formation as well as contaminant movement because it leads to long residence times. The extent of groundwater flow systems varies from local meter-scale hill slopes to multikilometer-scale regional basins. Groundwater velocities are typically faster in shallow, local flow regimes and slower in deeper, regional flow systems. As a result, the residence time of groundwater varies significantly, ranging from days in shallow, small systems to tens of thousands of years when considering deep flow in large sedimentary basins. Quantitative descriptions of groundwater flow require a depiction of the hydraulic head field and application of Darcy's law, which are discussed in the following subsections.

Hydraulic Head

Hydraulic head is one of the key variables in describing a groundwater system. It represents the mechanical energy per unit weight of fluid in the system. Hydraulic head, h , is defined as: $h = h_p + h_z$, where h_p is the pressure head and h_z is the elevation head. All three quantities have the dimension (L). The pressure head represents the energy due to pore fluid pressure, and the elevation head represents the gravitational potential energy arising from elevation. Because groundwater velocities are so slow, kinetic energy is typically negligible. Water flows from high to low hydraulic heads along a hydraulic gradient (discussed presently). *In situ* measurement of hydraulic head is accomplished by measuring the water level elevations in wells. First, the depth to groundwater is measured with the aid of a manual tape, electric sounding instrument, pressure transducer, or similar devices. Next, water level elevations are obtained by subtracting the measured depth to water from the land surface elevation. In regions where spatial hydraulic head differences are slight, typically in areas of gentle topography, water level measurements require accuracies of millimeters.

Darcy's Law

The basic equation governing groundwater movement is Darcy's law. In 1856, Henry Darcy, a French engineer in Dijon, France, performed an experiment involving water flow through a cylindrical sand column. The experimental data led to an empirical relationship between water flow and the experiment setup parameters. This relationship later became well-known as

Darcy's law in groundwater studies. Darcy's law is a simple gradient-flux relation and is as common in groundwater studies as analogous laws are in other fields such as Fick's law describing solute flux, Ohm's law describing electrical current, and Fourier's law describing heat conduction. In a one-dimensional system, Darcy's law is expressed as:

$$Q = -K \frac{dh}{dx} A \quad [3]$$

where Q is the volumetric flow rate ($L^3 t^{-1}$), K is the hydraulic conductivity ($L t^{-1}$), h is the hydraulic head (L), x is the coordinate (L), A is the cross-sectional area of flow (L^2), and dh/dx is the hydraulic gradient. It is clear from eqn [3] that the hydraulic gradient is the primary driving force for groundwater flow. The negative sign in the equation denotes that groundwater flow is in the direction of decreasing hydraulic head, down the hydraulic gradient. In describing groundwater flow, groundwater velocity, V_{gw} , is commonly computed. It is defined by the following equation:

$$V_{gw} = \frac{Q}{(A)(n_{eff})} = \frac{K}{n_{eff}} \frac{dh}{dx} \quad [4]$$

where n_{eff} is the effective porosity, or pore space through which water flows unimpeded by stagnant, dead end zones. V_{gw} is the average rate of a parcel of water occupying a volume of many pores and grains.

It should be noted that groundwater flow can be driven by other spatial differences in potentials besides a hydraulic gradient. For example, groundwater flow also can be driven by gradients in temperature, solute concentration, or both. In the absence of these other gradients, Darcy's law is a simple and sound formula to compute groundwater flow.

Wells

The primary need for wells is to withdraw water from the subsurface. Wells also serve as a window to the subsurface in the study of groundwater. Some wells are used for monitoring hydraulic heads (water levels) and also for sampling water for chemical analyses. When a well is pumped, a cone of depression forms around the well as the water level declines; the drawdown in the depression cone is greatest near the pumping well and less severe with distance from the well. The rate of water level decline is typically quite different depending on whether the pumping occurs in a confined or in an unconfined aquifer. Under the same pumping rate in different aquifers consisting of similar geologic materials, initially, a larger water level decline is expected in a confined aquifer and a smaller decline in an unconfined aquifer. For similar declines due to pumping, the volume of water obtained from confined aquifer storage is much less than that obtained from unconfined aquifer storage. When pumping from a confined aquifer, the porous material is not drained and does not desaturate. Rather, water comes from storage by compaction of the aquifer material accompanying slight rearrangement of grains plus expansion of water when pore pressure is reduced due to pumping. When pumping from an unconfined aquifer, water comes from storage by partial draining of saturated pores as the upper portion of the aquifer is converted into an unsaturated zone.

Surface Water and Groundwater Interaction

Groundwater interacts with lakes, streams, and wetlands across the landscapes from mountains to plains to shores of bays and oceans. In managing water resources, conjunctive use of surface water and groundwater has increasingly become a common practice, particularly in arid and semiarid regions. The basic concept of conjunctive use is to utilize surface water while storing excess water in aquifers under wet climate conditions when streamflow is high, and to withdraw water from the aquifers under drier conditions when demand is high but streamflow is low. The success of a conjunctive use project depends heavily on the dynamics of the interaction between the surface water and the groundwater. Streams can either gain water from or lose water to aquifers. Many streams do both in different reaches of the stream and at different times in the same reach. The rate and direction of flow in or out of the stream can also vary as the elevation of the water table relative to the stream surface fluctuates. Pumping of groundwater near streams can decrease the quantity of flow feeding a stream as baseflow, and even change a gaining stream into a losing stream. Moreover, interactions between groundwater and surface water affect water quality. When the groundwater in shallow aquifers is contaminated (for example, from agricultural practices of applying fertilizer and pesticides), the shallow aquifers can contaminate surface water as the groundwater flows toward a stream. The opposite can occur when a stream is heavily contaminated (for example, from mine waste drainage in mountainous regions). Mixing of groundwater and surface water affects other natural environments such as wetlands and aquatic environments when acidity, temperature, nutrients, and dissolved oxygen are altered by mixing. Streams, lakes, and wetlands may become acidic as they receive atmospheric deposition of chemicals, such as sulfate and nitrate. Acidic precipitation directly affects the well-being of aquatic ecosystems. In some cases, significant groundwater flow into a stream may help neutralize and reduce the stream acidity to tolerable levels for aquatic organisms.

Applied Aspects of Hydrology

The scientific aim of hydrology is to seek understanding of the mechanisms of water movement in the Earth system and the roles that water plays in natural processes. The applied aspect of hydrology, on the other hand, relates to using scientific knowledge to guide safe and sustainable use of water resources, investigating the impact and consequences of improper water use, and evaluating means to protect water resources. Although only two areas are discussed below, applied hydrology contributes to society and the environment far beyond what is included in these subsections.

Water Contamination

Water contamination has increasingly become a concern in modern times. Application of pesticides and fertilizer is common agricultural practice and has resulted in aerial contamination of water as excess irrigation water percolates through soils and carries the chemicals into groundwater and

directly or indirectly into surface waters. Contamination sources that occur over large areas are known as nonpoint sources; sources confined to small areas are called point sources. Multiple closely spaced point sources can form a nonpoint source. Landfills are major point sources of water contamination. Aging and leaking liners around landfills allow leachate, a mixture of water and dissolved chemical, to leak into groundwater. Wastes from mines, industrial disposal, nuclear reactor and weapon facilities, petroleum spills, and leaking underground storage tanks have all contributed to the contamination of groundwater and surface waters. Understanding contaminant transport is important for predicting future behavior of contaminated waters and designing effective remediation procedures.

Three major mechanisms control the transport of contaminants in water: advection, spreading, and chemical reactions resulting in retardation, decay, or degradation. Advection is the process of transporting contaminants by moving water and is often the dominant mechanism once the contaminants make their way into highly permeable aquifers. Spreading involves diffusing, dispersing, and mixing of the dissolved contaminants with clean waters. Spreading is due in part to a suite of mixing processes and dispersal mechanisms accompanying solute migration along relative preferential flow paths at multiple spatial scales. Spreading, mixing, and attenuation also occur due to the slow movement of solutes into and out of relatively stagnant zones, such as dead end pores and local immobile regions of low permeability. Advection and spreading are the physical processes of contaminant transport, while retardation, degradation, and decay result from chemical reactions occurring as the contaminated water migrates through and interacts with geologic media. Retardation, degradation, and decay are simple terms used here to express a collection of many complex chemical and biological processes. Chemical reactions can slow down the rate of migration of contaminants and even degrade or sequester certain compounds. These reactions may include sorption, precipitation, oxidation, reduction, ion exchange, and biological activity.

Land Subsidence

Extraction of groundwater plays a direct role in land subsidence. Uneven subsidence of the historic Tower of Pisa in Italy has created a tourist attraction. Subsidence has been a problem in many cities and towns as building foundations and road surfaces become cracked and tilted as the ground subsides. When a large amount of water is withdrawn from the subsurface, void spaces in rocks, and sediments collapse, which leads to compaction and subsidence. The Santa Clara Valley and nearby San Joaquin Valley in Northern California have experienced regional subsidence of several meters due to excessive groundwater pumping in order to sustain productive agriculture and water use in rapidly urbanizing areas. As Las Vegas Valley in Nevada has turned into a fast-growing metropolitan area, groundwater is being rapidly depleted and this area has also suffered problems with land subsidence and ground fissures due to horizontal movement. Subsidence may also occur from drainage of soils that are rich in organic carbon as microbial decomposition converts organic carbon to

carbon dioxide gas and water. Subsidence at the rate of 20–80 mm year⁻¹ has been observed as a result of the decomposition of the remains of shallow water sedges and reeds in California's Sacramento–San Joaquin Delta and in Florida's Everglades. More catastrophic subsidence takes place with the formation of sinkholes associated with localized collapse of subsurface cavities. Often triggered by a decline in the groundwater level, sinkholes typically form in areas underlain by carbonates (e.g., limestone and dolomites) and evaporites consisting of easily dissolved minerals such as salt, gypsum, and anhydrite.

Role of Groundwater in Geologic Processes

Water exists in pore spaces in sediments and rocks to a depth of more than 10 km. The degree of pore water decreases with depth in response to a general decrease of porosity with depth. Groundwater plays an essential role in mineral dissolution and precipitation, and impacts metamorphic processes by altering rocks' mineral compositions over geologic time. As water flows through deep sections of the crust or passes through thermally active regions, such as in the vicinity of a cooling magma, the heated waters become hydrothermal. Hot springs emerge at the locations of hydrothermal water discharge. Groundwater carries dissolved minerals and transports them to ore-forming locations. Petroleum forms, as natural gas or crude oil, at significant depths of burial of marine organism and sediments. Groundwater then transports the petroleum to shallower locations, a process known as secondary migration in petroleum system studies.

The mechanical interaction between groundwater and rock deformation has been thought to contribute to triggering earthquakes. As pore pressures in faults and the surrounding area increase, faults become lubricated and weakened, setting the stage for an earthquake. The best-known example is the documented earthquakes between 1962 and 1972 in the Denver area when liquid waste was injected underground into fractured granites a few kilometers deep at the Rocky Mountain Arsenal. The time and frequency of the earthquakes were correlated strongly with the time and volumetric rate of waste injection. Interest in fluid-induced seismicity is on the rise as fluid injection activities associated with carbon dioxide sequestration in geologic formations and hydrothermal exploration are expected to intensify in coming years.

See also: **Hydrology, Floods and Droughts: Overview; Soil Moisture; Modeling and Prediction.**

Further Reading

- Alley, W.M., Reilly, T.E., Franke, O.L., 1999. Sustainability of Ground-Water Resources. U.S. Geological Survey Circular 1186.
- Black, M., King, J., 2009. *The Atlas of Water*. University of California Press, Berkeley, CA.
- Bredehoeft, J.D., Papadopoulos, S.S., Cooper, H.H., 1982. Groundwater: The water-budget myth. In: *Scientific Basis of Water-Resource Management, Studies in Geophysics*. National Academy Press, Washington, DC, pp. 51–57.

- Dingman, S.L., 2002. *Physical Hydrology*. Prentice Hall, Upper Saddle River, NJ.
- Domenico, P.A., Schwartz, F.W., 2008. *Physical and Chemical Hydrogeology*. John Wiley & Sons, Inc., New York, NY.
- Fetter, C.W., 1999. *Contaminant Hydrogeology*. Prentice Hall, Upper Saddle River, NJ.
- Fetter, C.W., 2001. *Applied Hydrogeology*. Prentice Hall, Upper Saddle River, NJ.
- Fierro, P., 2007. In: *The Water Encyclopedia, Hydrologic Data and Internet Resources*, third ed. CRC Press, Boca Raton, FL DOI: 10.1201/9781420012583.ch4.
- Freeze, A.R., Cherry, J., 1979. *Groundwater*. Prentice-Hall, Englewood Cliff, NJ.
- Gleeson, T., Marklund, L., Smith, L., Manning, A., 2011. Classifying the water table at regional and continental scales. *Geophysical Research Letters* 38, L05401. doi:10.1029/2010GL046427.
- Gleick, P.H., 2001. Making every drop count. *Scientific American* February, 41.
- Haggerty, R., Gorelick, S.M., 1995. Multiple-rate mass-transfer for modeling diffusion and surface-reactions in media with pore-scale heterogeneity. *Water Resources Research* 31 (10), 2383–2400.
- Hornberger, G.M., Raffensperger, J.P., Wiberger, P.L., Eshleman, K.N., 1998. *Elements of Physical Hydrology*. The Johns Hopkins University Press.
- Kenny, J.F., Barber, N.L., Hutson, S.S., Linsey, K.S., Lovelace, J.K., Maupin M.A., 2009. *Estimated Use of Water in the United States in 2005*. U.S. Geological Survey Circular 1344.
- Rodríguez-Iturbe, I., Rinaldo, A., 2001. *Fractal River Basins*. Cambridge University Press, Cambridge, UK.
- The World's Water Series*, <http://www.worldwater.org/books.html>
- Tindall, J.A., Kunkel, J.R., 1999. *Unsaturated Zone Hydrology*. Prentice Hall, Englewood Cliffs, NJ.
- UNEP., 2008. *Vital Water Graphics – An Overview of the State of the World's Fresh and Marine Waters*, second ed. UNEP, Nairobi, Kenya. ISBN: 92-807-2236-0.
- Vorosmarty, C.J., Green, P., Salisbury, J., Lammers, R.B., 2000. Global water resources: vulnerability from climate change and population growth. *Science* 289, 284–288.
- Winter, T., 1998. *Groundwater and Surface Water, a Single Resource*. U.S. Geological Survey Circular 1139.

Modeling and Prediction

Z Yu, College of Hydrology and Water Resources, Hohai University, Nanjing, China; and University of Nevada Las Vegas, Las Vegas, NV, USA

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

This article describes the general approach for hydrologic modeling and prediction. After the hydrologic system is conceptualized, mathematical model can be set up and numerical model can be developed. Then data such as precipitation, land use–land cover, digital elevation, and soil texture need to be prepared for the hydrologic simulation. Before the model can be used for the prediction, it needs to be calibrated for assessing how well the simulation can reproduce the observed hydrologic variables.

Introduction

Hydrologic models have become an indispensable tool to study hydrologic processes and the impact of natural variability and anthropogenic factors on the hydrologic system. Mathematical models, which are governed by the laws for conservations of mass and momentum, are used to describe the temporal and spatial variation on various physical processes in the hydrologic system along with information concerning climate, land use–land cover, and hydrology. Modeling the hydrologic response to various natural and human-induced changes has the potential to contribute to the understanding of these physical processes such as flow and solute transport in the surface and subsurface and the atmosphere–land surface interaction.

Two types of hydrologic models have been used in most applications: lumped conceptual models and physically based models. A lumped model is generally applied in a single point or a region for the simulation of various hydrologic processes. The parameters used in the lumped model represent spatially averaged characteristics in a hydrologic system and are often unable to be directly compared with field measurements. In general, lumped models use simple bookkeeping procedures to quantify physical processes by simulating the temporal variation of various physical processes in a hydrologic system. The advantage of these models over physically based models is that the conceptual parameterization in the models is simple and computation is efficient. With the availability of spatially distributed digital and remotely sensed data sets such as precipitation, elevation, vegetation, etc., many distributed lumped models have been developed in recent years. These kinds of models have been widely used in climate, meteorological, and hydrologic studies to simulate hydrologic processes.

Many physically based distributed-parameter models have been developed to facilitate various hydrologic and climatic applications over recent years, especially for simulating spatiotemporal variation in hydrologic processes under climate changes such as global warming. These models represent hydrologic processes in a physically rigorous manner because they use process-based partial differential equations (PDEs) to describe the spatial variability of hydrologic processes over time. Within this framework provided by these equations, it is possible to estimate governing parameters through field studies. One disadvantage of such models is that the representation of physical processes in these models is often too

crude and the scales of measurement for many hydrologic parameters are incompatible with the scales used in the models. Studies indicated that the uniform effective parameters (e.g., saturated hydraulic conductivity) cannot represent the ensemble heterogeneity, resulting in a significant underestimation of hydrologic variables (e.g., streamflow). These physically based distributed models can be referred to as conceptual lumped models in some sense. The physically based models are more complex and require more computing time for solving PDEs numerically and considerable effort to master all their intricacies, such as model calibration. For simulating the hydrologic response (e.g., soil moisture content, groundwater level) to climate forcing (e.g., storms and human-induced global warming), these models currently offer no advantage over the traditional conceptual lumped water-balance models.

Physically based hydrologic models are an important evolutionary step in representing hydrologic processes and spatially distributed data. At the present time, the ability to represent spatially varying processes is outstripping our ability to collect various data sets. The need for the research on the better representation of physical processes in space and time is evident given the increasing availability of digital products (e.g., distributions of elevation, soil, vegetation) and remotely sensed data (e.g., soil moisture, vegetation), along with new remote sensing technologies for measuring temporal and spatial variability in precipitation.

Research on data assimilation and analysis, subgrid-scale variability in precipitation and hydraulic parameters, subgrid-scale model calibration, prediction uncertainty analysis, and inclusion of fine-resolution surface and subsurface hydrology in various hydrologic and climatic models are being conducted in various research institutes to facilitate an interactive link between the hydrosphere and the atmosphere. Simulation and prediction of various physical processes in the hydrologic system are among the principal areas of study in the current fields of hydrology and climate.

General Approach of Hydrologic Modeling

Climate variability has substantial impact on hydrologic, biologic, and ecologic systems such as water availability and quality, floods, and droughts. Such climate variability has been intensified due to the human-induced climate changes and

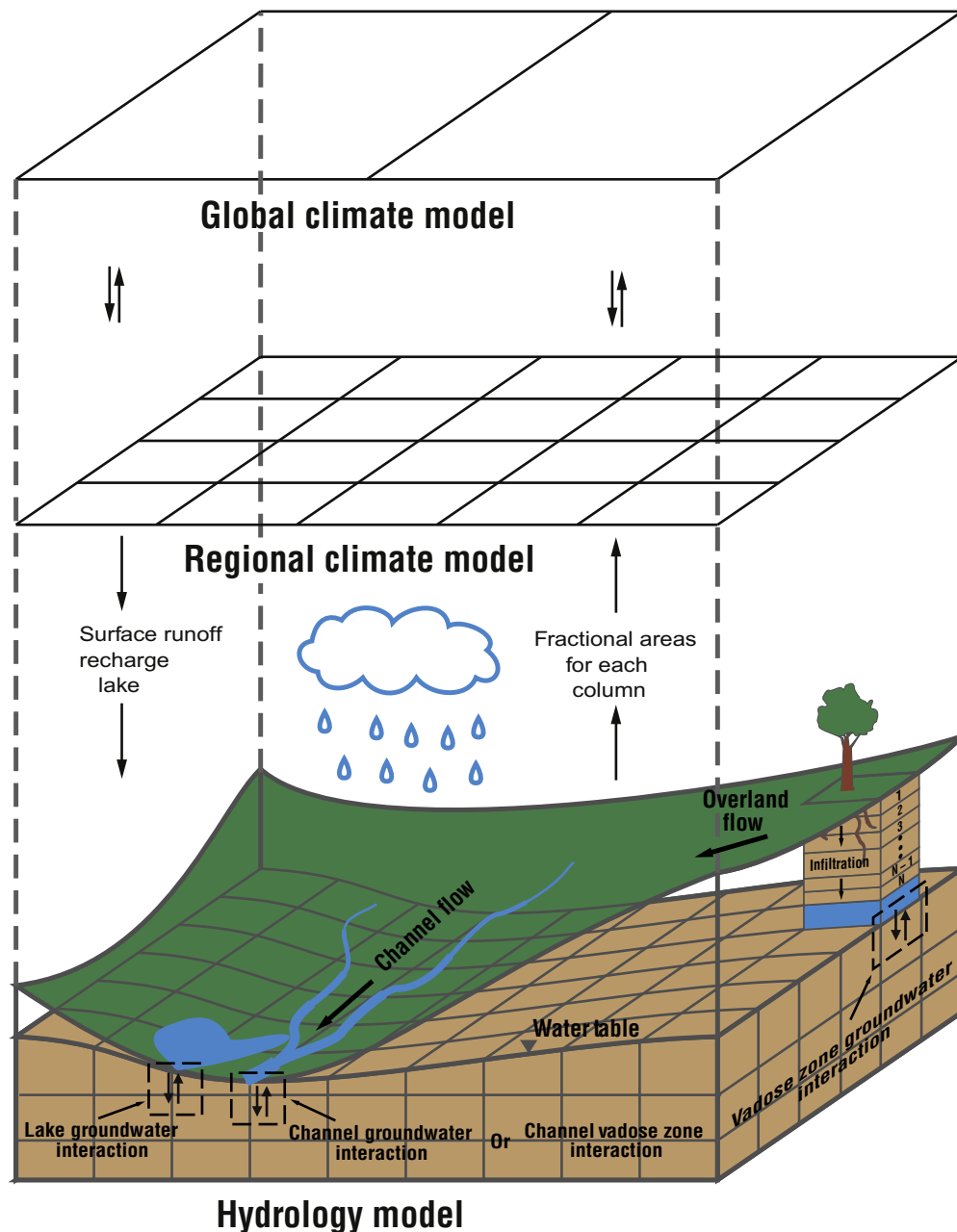


Figure 1 Frame structure of coupled climate and hydrology models. Source: Yu, Z., Pollard, D., and Cheng, L., 2006. On continental-scale hydrologic simulations with a coupled hydrologic model. *Journal of Hydrology* 331, 110–124.

results in many extreme flood and drought events in recent years. The effects of climate variability on economic vitality and the quality of life (e.g., 1988 North American drought and 1993 Mississippi flood in the USA) indicate that future climate change will be of considerable global, national, and regional importance. The current emphasis on the climate impact analysis and hydrologic modeling requires improved specifications of the interactions between the atmosphere and the hydrosphere. The difficulty in reconciling the discordant scales in climate change and hydrologic research hampers the understanding of how regional hydrologic systems respond to climate forcing. Part of the difficulty can be attributed to

insufficient progress on the theoretical side, related to deficiencies in our abilities to couple atmospheric and hydrologic processes that occur over spatial scales ranging from microscopic to global (specifically ranging from 10^{-4} to 10^6 m) and over time scales ranging from 10^{-4} to 10^7 s. Various parameterization schemes have been used to reconcile the discordant scales of hydrologic and atmospheric models. The approach involves numerical modeling to bridge the gap in scale between global circulation models (GCMs) and hydrologic models by embedding a regional climate model within a GCM so that the natural variability at the regional scale can be represented within the system (Figure 1).

The nested climate models used for studying the effect of climate change on the hydrologic system indicate that changes in the hydrologic system induced by a climatic change could lead to a further change in climate, ecosystems, agriculture, and hydrology. Many comprehensive hydrologic models such as the Biosphere–Atmosphere Transfer Scheme have been developed to integrate hydrology into the climate system by coupling the physical processes from different systems. Such an integrated hydrologic model system, including components from soil hydrology, surface water hydrology, and groundwater hydrology, is necessary to understand how perturbed climate conditions can modify regional hydrologic systems and, in turn, how the modified hydrologic conditions can modify the climatic system.

Traditionally most hydrologic models have been developed to simulate individual hydrologic processes such as rainfall–runoff partitioning, soil moisture flow, overland flow, and groundwater flow while some comprehensive models are used to simulate all the processes within the hydrologic cycle. In general, any use of hydrologic models for hydrologic and climate studies follows five main steps (Figure 2). After the purpose of hydrologic modeling is well-defined, the conceptual model then can be constructed to describe the hydrologic system and its components that need to be simulated. In the second step, the various mathematical models with various PDEs can be used to describe individual processes within the hydrologic system and computer codes in various computer languages (e.g., C, Fortran) can be developed to solve these PDEs in the mathematical models. In general, the numerical solution of these PDEs must be verified by comparing to known or analytic solutions. The second step can also be completed by selecting existed computer codes that have been verified in previous applications (e.g., MODFLOW). The third step, data preparation, is to compile and analyze available

meteorological and hydrologic data for the hydrologic simulation. Additional data can be obtained through laboratory and field works, if necessary. After all required data being integrated into the models, the models should be calibrated to available field-observed data (i.e., streamflow, groundwater level). So the calibrated models can then be used to predict the hydrologic response to various external forcings (e.g., future water usage, land use–land cover change, and climate change).

System Conceptualization, Mathematical Model Setup, and Model Development

There are many surface and subsurface hydrologic processes in the hydrologic system (Figure 1). Many conceptual and mathematical models have been developed to describe these physical processes in time and space and various numerical schemes (i.e., finite-difference and finite-element methods) are available to solve various PDEs for the simulation of hydrologic processes. The following sections describe some of the major flow processes such as soil water flow, overland flow, and groundwater flow while other processes such as sediment and contaminant transport will not be described here.

When the precipitation reaches the ground, it partitions into components of surface runoff and infiltration. After the water enters the soil column, the soil water flow in the unsaturated zone can be described by the following form of Richards' equation:

$$-\frac{\partial q(z, t)}{\partial z} = \frac{\partial}{\partial z} \left[K(z, t) \frac{\partial \psi(z, t)}{\partial \theta} \frac{\partial \theta(z, t)}{\partial z} \right] + \frac{\partial K(z, t)}{\partial \theta} \frac{\partial \theta(z, t)}{\partial z} \quad (1)$$

where q is the vertical moisture flux, θ is the volumetric water content, z is depth, t is time, K is the unsaturated hydraulic conductivity, and ψ is the hydraulic potential. The Richards' equation can be numerically solved using the Crank–Nicholson numerical scheme and a finite-difference scheme of forward in time and backward in space or other numerical methods. Various schemes are available for relating K and ψ to the volumetric water content because K at a given time is a function of constant saturated hydraulic conductivity and soil moisture status. Infiltration and evaporation can be treated as either a source or a sink, respectively, in Richards' equation or the upper boundary condition of the soil profile. The calculation of evaporation on the bare soil and evapotranspiration on the vegetation canopy can be completed with various methods (e.g., the Penman–Monteith method).

The overland flow will form soon after the rainfall–runoff partitioning and can be formulated as a kinematic wave with a flow-direction algorithm to account for the overland flow delay and storage in each grid cell. Mathematically, one-dimensional overland flow may be described as

$$\frac{\partial d}{\partial t} + \frac{\partial q}{\partial x} = i_e, \quad q = \alpha d^m \quad (2)$$

where $d = d_{(x,t)}$ is the depth of overland flow, $q = q_{(x,t)}$ is the rate of overland flow, i_e is the excess rainfall rate, α is conveyance, and m is surface roughness. The overland flow is discretized into a large number of smaller segments to solve the kinematic-wave flow and the equation can be solved numerically for

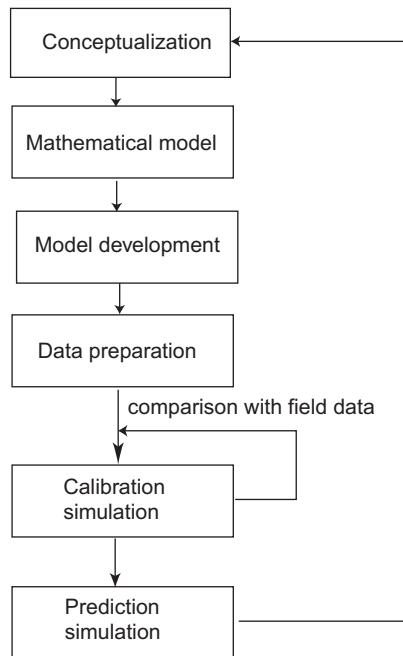


Figure 2 Flow chart of hydrologic model application.

the one-dimensional flow. In general, a conservative numerical scheme is implemented in the model to maintain the numerical stability in some extreme cases.

The overland flow quickly reaches nearby streams to form channel flow. The channel flow can be simulated hydraulically and hydrologically. The Muskingum–Cunge method or other channel routing schemes can be used for channel-flow routing through the channel networks that are derived from digital elevation models (DEMs). For instance, the extended finite-difference form of the governing equation can be written in the Muskingum channel routing method:

$$K \frac{d}{dt} [\varepsilon Q_x + (1 - \varepsilon) Q_{x+1}] = Q_x - Q_{x+1} \quad (3)$$

where Q_x is the inflow to a given stream reach, Q_{x+1} is the outflow from a given stream reach, K is a travel time parameter, and ε is a weighting factor. The routing equations are produced using a finite-difference form of the above equation. The outflow of a stream reach is calculated at each time step and routed to downstream reach along DEM-derived channel networks, eventually to the basin outlet.

The following second-order PDE is used to describe groundwater flow in an aquifer:

$$\frac{\partial}{\partial x} \left(T \frac{\partial h}{\partial x} \right) + \frac{\partial}{\partial y} \left(T \frac{\partial h}{\partial y} \right) = S \frac{\partial h}{\partial t} + Q_{\text{net}} \quad (4)$$

where T is transmissivity, h is hydraulic head, S is storativity, t is time, and Q_{net} is net groundwater withdrawal rate. Q_{net} includes groundwater recharge from infiltration, evaporation of shallow groundwater, withdrawal of groundwater from wells, and other sources or sinks, such as induced infiltration of groundwater from the stream network.

In general, the simulation domain for a basin is discretized into a set of rectangular cells with dimensions Δz , Δx , and Δy . Δz is the vertical thickness of the aquifer, and Δx and Δy are the cell dimensions in the x - and y -coordinate directions (Figure 2). A finite-difference or finite-element scheme can be used to discretize the hydrologic system for the numerical solution of the flow domain. Many iterative methods such as the method in MODFLOW (developed by U.S. Geological Survey (USGS)) are available for obtaining solutions for the above PDE.

The interaction between a stream system and groundwater can be simulated with Darcy's law. The calculated inflow/outflow is a function of the hydraulic head difference between the groundwater and the river stage at a given stream reach. When the groundwater level is lower than the streambed, water flows from the stream to the groundwater. When the groundwater level is lower than the streambed, the stream will dry out and eventually there will be no channel groundwater interaction.

After the implementation of numerical schemes for solving various PDEs in the hydrologic system, the developed codes for hydrologic models need to be tested against to known and analytic solutions before they can be used for modeling applications. For the specific applications, the simulation domain must be set up with defined boundary conditions. So the given forcings (i.e., meteorological data and water usage) can be used to drive the simulation of various hydrologic processes.

Data Compilation and Preparation for Hydrologic Modeling

Data compilation and preparation is a major fundamental part of hydrologic modeling. The work involves compiling various available field and digital data sets and processing and integrating these data sets into models for hydrologic simulations. Sources of data can be from various government agencies and institutes. With the availability of new technology, ground-based and remotely sensed data sets as well as the collected data through traditional field methods have been used for hydrologic modeling. Some data are required to be reprojected or reformatted for the specific models.

For topography-based hydrologic models, DEMs are routinely used to describe the land surface topography that contains crucial information for the surface water flow and the interaction of surface water and groundwater. After DEMs are conditioned for removing spurious sinks and peaks, DEMs can be used to determine the flow direction of surface water at a grid cell by comparing the elevation at a grid cell to elevations of neighboring grid cells within the simulation domain assuming that the water always flows from high level to low level. With DEMs, a flow accumulation data set can be obtained in which the contributing area that drains to a grid cell is determined for each grid cell. The DEMs and derived data sets of flow direction and flow accumulation can then be processed to delineate drainage and stream networks, overland flow paths, watershed and subwatershed boundaries, and other features. Derived stream networks facilitate the channel–shallow groundwater interaction while the derived overland flow paths provide accurate prediction of the rainfall–runoff partitioning and soil moisture status over overland flow paths.

Simulations of various hydrologic processes require spatially distributed information on land use–land cover, soil type, and hydraulic properties besides the topography. Many digital soil data sets are available from various sources. For instance, the Natural Resources Conservation Service, U.S. Department of Agriculture, compiled the State Soil Geographic (STATSGO) soils data. In general, these digital data sets are stored in a Geographical Information System format. The hydraulic parameters (i.e., the average saturated hydraulic conductivity and the average capillary suction) data sets can be derived by assigning a value for each grid cell based on various parameterization schemes of hydraulic variables and the soil texture from STATSGO soil data. Then, these derived data sets of various hydraulic parameters can be integrated into models for the hydrologic simulation.

With the development of remote sensing and satellite technology, these data sets hold great potential for practical application to regional ecology, hydrology, and planning. The satellite data can be used to derive a variety of surface parameters, such as radiant surface temperature and vegetation fraction; these variables can be used for both spatial and temporal comparisons and for regional hydrologic modeling of important hydrologic processes. The remotely sensed data such as vegetation classification can be obtained from the AVHRR and Landsat TM images. Spatial distributed data of meteorological variables (e.g., precipitation, temperature, wind speed) are available from NEXRAD, radar, and satellite images at a time resolution from minutes to days. With the improved

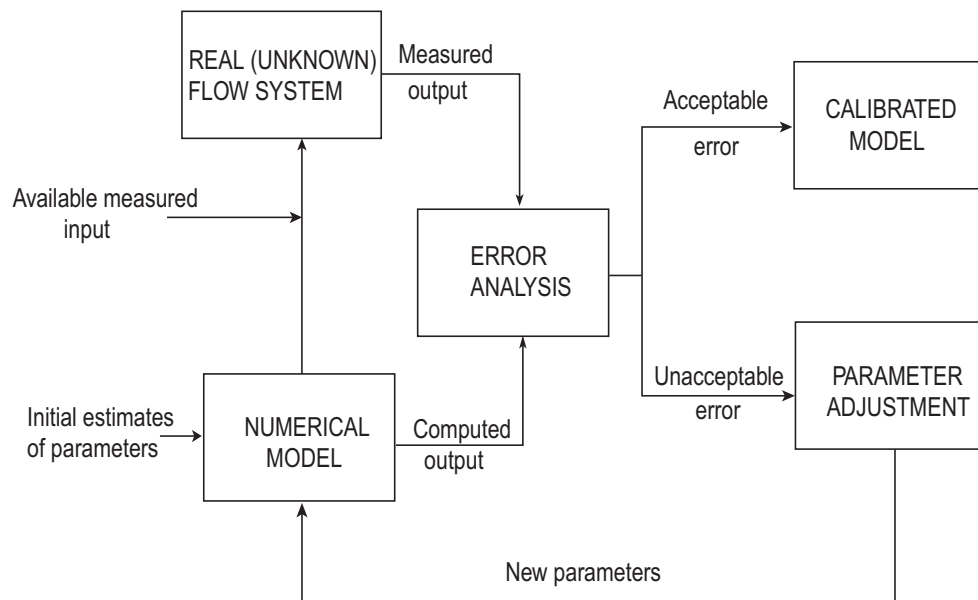


Figure 3 Procedure of model calibration. Source: Freeze, R.A., and Cherry, J.A., 1979. Groundwater. Prentice-Hall, pp. 604.

resolution on these data sets, the more accurate simulation of various hydrologic processes is expected in both temporal and spatial scales.

Different types of field-observed data are available for the hydrologic simulation and model calibration that will be discussed in the next section. There are networks of rain gages over the world. At outlets of many river systems, gage stations are available to measure daily streamflow rates and other hydrologic variables. Groundwater levels are also available from many wells and many other data such as evapotranspiration and moisture content are obtainable from local specific regions. For instance, USGS provides infrastructure for monitoring stations and collecting measured data of streamflow and groundwater level while the National Oceanic and Atmospheric Administration maintains the collection of meteorological data.

Model Calibration

Complexity in the hydrologic modeling over large space and long times has prompted a significant need for model calibration or parameter optimization. Model calibration is a demonstration that the model is capable of reproducing field-observed values of various hydrologic variables (e.g., streamflow, soil moisture, and well-observed groundwater level) (Figure 3). Prediction of various hydrologic variables based on an uncalibrated flow model are sterile and undefensible. Generally, the goodness-of-fit between simulated and measured variables is not satisfactory based on the initial values of hydrologic and hydraulic parameters used in the model. The goodness-of-fit can be improved by the adjustment or optimization of these parameter values until the difference between simulated and measured variables is satisfactory during this model practice. The adjustment process most commonly is based on trial-and-error changes in a parameter while other parameter values are held constant. Some

numerical models are now equipped with a semi-automated or automated procedure to optimize one or multiple parameters. The range of adjustment to values of hydrologic and hydraulic parameters must be constrained by plausible site-specific field data such as streamflow, water levels, hydraulic conductivity, and so on. The difficulty in achieving a good calibration is that boundary conditions and values of hydrologic and hydraulic parameters are always known with uncertainty.

Goodness-of-fit calibration can be evaluated through visual comparison and statistical measures. Visual comparison includes scatterplot of simulated versus measured variables, simulated and field-based temporal and spatial distribution, and spatial distribution of residuals. Statistical measures consist of mean error, absolute mean error, and root mean-squared error, between simulated and observed variables.

The trial-and-error calibration involves with manually adjusting parameters to match the simulated results (e.g., streamflow and groundwater level hydrographs) with field-observed historical records. The fitness between the simulation and observation can be evaluated visually and statistically after each simulation with changed parameters. Knowledgeable modelers are able to conduct acceptable model calibration with this approach. On the other hand, the calibration process could be a very frustrating and time-consuming practice for many untrained modelers and sometimes even for experienced modelers when many parameters are changed simultaneously. The disadvantage for such approach is the low efficiency and the lack of standard measures in evaluating the performance of model calibration.

While the trial-and-error calibration is still the most common method for calibrating models of all kinds, many automatic or semi-automatic direct or indirect techniques are available to solve an inverse problem in various hydrologic systems. Automatic direct or indirect model calibrations require setting up objective functions, developing optimization algorithms, setting up termination criteria, and of course,

using observed data or derived variables as calibration targets. The direct approaches use an inverse operator to solve the problems directly, while the indirect approaches use an iterative scheme. For the indirect approaches, various sampling procedure schemes (e.g., Monte Carlo and Latin Hypercube samplings) are available to sample data that cover the range of adjustable variables systematically. In effect, the simulation problem is solved many times over with the coverage of the potential solution space. Ultimately, the process finds the optimum solution, providing a set of parameters that yield the best fit between the observed data and simulated results.

For any model calibration, objective functions need to be set up for variables (e.g., streamflow, groundwater level) that will be optimized during the calibration process. The least-squares method, a common method in the model calibration, is equivalent to the minimization of the sum of squares of the residuals (also called objective function or fitness measure). The general form for a steady state is given by

$$\text{Obj} = \sum_{i=1}^n w_i [s_i^o - s_i^c]^2 \quad (5)$$

where w_i is a weighted factor, s_i^o is the observed data, s_i^c is a simulated result, i is the index for location points, and n is the number of location points.

The weighted factor represents the percentage of importance allocated to a variable at particular locations. If the sum of all weighted factors is 1.0, the above equation becomes the simple least-squares function. It is feasible to zone spatially for both steady state and transient state, although execution times will increase significantly, especially for the transient state, as the degrees of freedom in parameter values increase. The objective function can be constructed for hydrologic variables that can be used as calibration targets (e.g., streamflow and groundwater level). Objective functions, based on a weighted least-square error criterion, can then be used to drive the calibration process. Different combinations of various changes in different optimizing parameters are selected for each simulation run. The objective is to minimize the difference between measured data and simulated results. The objective function for a transient simulation can also be formed as

$$\text{Obj} = \sum_{j=1}^m \sum_{i=1}^n w_{i,j} [s_{i,j}^o - s_{i,j}^c]^2 \quad (6)$$

where m is the number of time steps, j is the index for time steps, and other variables are as defined previously. The transient calibration requires much more computational time than the steady-state calibration due to the additional time dimension. There are other alternatives to the least-squares method such as the maximum-likelihood approach for objective functions.

The direct approaches use an inverse operator to solve the problems directly. Let h^o be the observed hydrologic variables and h^c be the simulated hydrologic variables. Then, according to Taylor's series expansion, the equation can be written as

$$\sum_{j=1}^m \frac{\partial h_i^c}{\partial p_j} \delta p_j = h_i^o(p + \delta p) - h_i^c(p) \quad i = 1, \dots, n \quad (7)$$

In a matrix form, the above equation is written as

$$\mathbf{J} \delta \mathbf{p} = \mathbf{r} \quad (8)$$

where $\mathbf{J}$ is a rectangular Jacobian matrix, $\mathbf{p}$ is the parameter vector to be inverted, $\delta \mathbf{p}$ is the parameter correction vector, and $\mathbf{r}$ is the residual vector between the measured and simulated hydraulic heads. In order to solve the above equation, the transpose of matrix $\mathbf{J}$ multiplies both sides of the equation. The result is a system of equations with m unknowns. A problem occurs because the solution of matrix system is unstable. Marquardt's approach, which involves adding terms to the diagonals of the matrix $\mathbf{J}^T \mathbf{J}$, can overcome this problem. The Levenberg-Marquardt method or Marquardt method (iterative procedures) can be used to solve this system.

After the model calibration, the sensitivity analysis is needed to investigate the uncertainty in the calibrated model caused by the uncertainty in the estimation of various parameters. The sensitivity analysis can be conducted by evaluating the hydrologic flow system in response to various parameter disturbances based on the calibrated hydraulic parameters. The analysis involves perturbing the values of the hydraulic parameters with respect to the best estimates and examining how the simulated hydrologic variables change. The common approach is to evaluate one parameter at a time; all other parameters are kept constant while a given parameter is being evaluated. Then various parameters are varied simultaneously to examine the behavior of the hydrologic flow system under different situations. By doing this, the nonuniqueness of solution of the parameter estimations for the hydrologic flow system can be evaluated and other issues along with the model uncertainty can be assessed.

Model Prediction

With the set of optimized hydraulic parameters from the model calibration, the calibrated model can be used to conduct predictive simulations. The predictive simulations can be driven by external forcings (e.g., climate change scenarios) to derive the system response to future events. Some climate and hydrology problems require the hydrologic modeling to predict the system response for as many as thousands of years. For instance, the risk assessment of contaminant transport for storing low-level nuclear wastes in arid regions requires predicting the water flow in the unsaturated and saturated zones for a time scale of 10 000 years.

How good predictions are depends on how well the model is calibrated assumed mathematical and numerical models are properly established and how creditable future external forcings are, which are used to drive predictive simulations. Because of the nonlinearity of natural hydrologic systems and simplified model structures or assumptions as compared to the actual complex natural physical processes, the calibration process cannot guarantee that the global minimum can be found and so the calibrated model could produce unrealistic results beyond the confident period. In general, the calibrated model should not be used to predict the system in the future longer than twice the period of model calibration with available field-observed

data. In some cases, due to the lack of available long-period historical data for model calibration and the nature of some applications that require an extended period prediction, there will be always uncertainty in predictive simulations. So users must conduct a detailed assessment of the uncertainty on any future prediction of system response as well as the uncertainty associated with model calibration and future external forcings.

The hydrologic cycle plays an important role in the climate system. The two-way interaction between the hydrologic and climate is a crucial element in our current and future hydrology and climate modeling to understand how perturbed climate conditions can modify regional hydrologic systems and, in turn, how the modified hydrologic conditions can influence the climatic system. Data assimilation and analysis of remotely sensed and satellite-derived information should be integrated into hydrologic models to better represent subgrid-scale spatial and temporal variability in meteorological and hydrologic parameters. The model calibration and subgrid-scale model simulation of various hydrologic components. In addition, prediction of uncertainty analysis and inclusion of fine-resolution surface and subsurface hydrology are some important topics in such an interactive link.

Further Reading

- Anderson, M.P., Woessner, W.W., 1992. *Applied Groundwater Modeling*. Academic Press, San Diego, CA, pp. 381.
- Beven, K.J., Moore, I.D., 1993. *Terrain Analysis and Distributed Modeling in Hydrology*. John Wiley & Sons, New York, NY, pp. 249.
- Freeze, R.A., Cherry, J.A., 1979. *Groundwater*. Prentice-Hall, Upper Saddle River, NJ, pp. 604.
- Maidment, D.R., 1993. *Handbook of Hydrology*. McGraw-Hill, New York, NY.
- Neuman, S.P., 1973. Calibration of distributed parameter groundwater flow models viewed as a multiple-objective decision process uncertainty. *Water Resources Research* 9, 1006–1021.
- Singh, V.P., 1995. *Computer Models of Watershed Hydrology*. Water Resource Publications, Littleton, CO, pp. 1130.
- Yang, C., Lin, Z., Yu, Z., Hao, Z., Liu, S., 2010. Analysis and simulation of human activity impact on streamflow in the Huaihe river basin with a large-scale hydrologic model. *Journal of Hydrometeorology* 11, 810–821.
- Yu, Z., Schwartz, F.W., 1999. Automated calibration applied to constrained groundwater flow modeling. *Hydrological Processes* 13, 191–209.
- Yu, Z., Lakhtakia, M.N., Yarnal, B., Johnson, D.L., White, R.A., Miller, D.A., Barron, E.J., Duffy, C., Schwartz, F.W., 1999. Simulation of the hydrologic response to atmospheric forcing in large river basins: linking a mesoscale meteorological model and a hydrologic model system. *Journal of Hydrology* 218, 72–91.
- Yu, Z., Barron, E.J., Schwartz, F.W., 2000. Retrospective simulation of a storm event: a first step in coupled climate/hydrologic modeling. *Geophysical Research Letters* 27, 2561–2565.
- Yu, Z., Pollard, D., Cheng, L., 2006. On continental-scale hydrologic simulations with a coupled hydrologic model. *Journal of Hydrology* 331, 110–124.

Palmer Drought Severity Index

L Nkemdirim, University of Calgary, Calgary, AB, Canada

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

Drought damages the physical environment, which could compromise the socioeconomic well-being of affected communities. The degree of damage and compromise typically depends on the severity of the event. Wayne Palmer (1965) designed and calibrated an index that is widely used to measure drought severity (Palmer Drought Severity Index – PDSI) both in the present and the past. It is also utilized in a forecast mode. The index is based on the difference between the amount of water required to support the existing water requirement at a place and time and the actual water available at that place and time. Droughts occur when the difference is negative, while wet conditions prevail when positive. The Index of drought (wetness) is determined from a cumulative plot of that difference over monthly time steps. Palmer's model is a two-part system. The first part is the Hydrological Index, which thoroughly accounts for the exchange of water between the atmosphere and the ground. The second part is a meteorological modification that times the onset and end of a drought (wet) event.

Introduction

The Palmer drought severity index (PDSI) is widely used to classify and manage droughts and wet periods. Details of its development and rationale are provided in this article. Some of the criticisms directed at the Index are discussed briefly in the Section [Criticism of the Palmer Index](#). However, this text is not a critical review of PDSI. The objective is to help a potential user decide its adequacy for a specific application. Droughts are the main focus, but wet events, which are scaled by the Index, are treated when relevant.

Drought: Definition, Significance, and Cause

The general sense of a drought is a prolonged absence of sufficient precipitation during a period when it is normally expected. Some disciplines have particular outcome-dependent definitions. Meteorologists defined droughts by the time interval between large precipitation events. In agriculture, drought is said to occur if the supply of moisture from precipitation or soil storage is inadequate for optimum crop growth. A long rainless period may not be so classified if crop yield is unaffected. When streamflow is adequate for established uses under a given water management system, a significant precipitation-free period may not be considered a drought by the water engineer. However, if the duration of precipitation deficiency over a large geographical area is months rather than weeks, and precipitation amounts are well below normal, the senses of drought conveyed by various discipline-specific definitions converge.

Droughts degrade the physical environment. They dry out soils, damage vegetation, and reduce streamflow. Impacts on socioeconomic health can be far reaching. They include threats to food security, industrial production, and population stability. Drought-driven environmental refugees are found in every continent. The African Sahel has been particularly vulnerable to mass migration due to persistent droughts. In Canada and the United States, the droughts of the 1930s significantly restructured the socioeconomic landscape in the West and Midwest.

Droughts are caused by anomalies in atmospheric circulation, which produce air subsidence over a stricken area. They may be intensified and/or prolonged by land surface processes

and the energy–water balance they drive. An index tuned to capture this combination of atmospheric and land surface forcing could approach general applicability.

Drought Index

A drought index is a numerical scale used to describe the severity of a dry (wet) period. There are many drought indices in current usage, many of them discipline specific. [Palmer \(1965\)](#) distinguished himself by constructing one that

- is conceptually simple for qualitative interpretation;
- crosses disciplinary boundaries with sufficient flexibility to recommend its use, in modified form if necessary, in many sectors;
- allows comparison of severity to be made over both time and space;
- ensures that values assigned to severity reflect variations in both duration and intensity such that impacts of long-lasting but mild events can be distinguished from those caused by short but very intense ones;
- contains a mechanism for determining the onset and end of an event;
- enables differentiation between aridity and drought and between drought and a naturally occurring dry season;
- is calculable from easily obtainable data; and
- possesses a forecast potential.

Drought Parameters

A drought (wet period) is normally described by three parameters, namely, duration, magnitude, and severity. In order to isolate and relate these parameters, a truncation level Y_0 is established to separate drought from wet events ([Figure 1](#)). Y_0 may be constant, as in [Figure 1](#), or may vary through time and space to reflect different water requirement thresholds. The function $Y_{(t)}$, which quantifies relative dryness (wetness) at time t , can be construed as a drought index. Index values with $Y_{(t)} < Y_0$ represent drought, while those with $Y_{(t)} > Y_0$ indicate wet conditions. Duration is the time span between successive crossings of the Y_0 line by Y_t (e.g., $t_2 - t_1$). Magnitude is the mean deviation of moisture conditions at $Y_{(t)}$

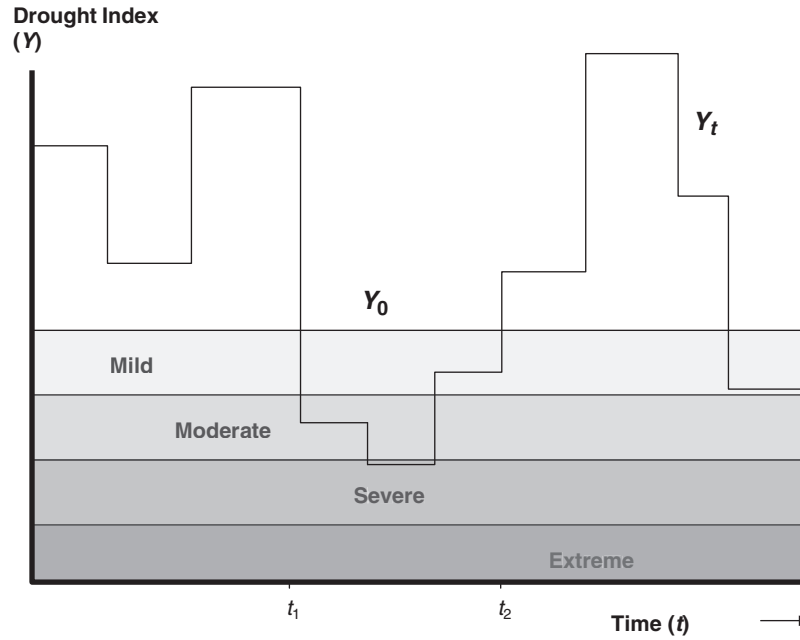


Figure 1 Conceptual model of a drought index.

from Y_0 over duration $t_2 - t_1$, and severity is the cumulative deviation of those conditions over the duration. The dimension of t_j (the averaging period) should be small enough to provide enough samples for meaningful statistical analysis but sufficiently long so that events are statistically independent of one another. Additionally, the duration Δt_i must not trivialize the incidence of an event. Each of the two broad index categories, namely, $Y_t < Y_0$ and $Y_t > Y_0$, may be subdivided into classes to reflect increasing severity of dryness or wetness, and decisions must be made on the location of subdivision boundaries that are physically meaningful. Figure 1 shows severity division for droughts.

Palmer Drought Severity Index

The main features of the PDSI are listed in Section [Drought Index](#). The present section explains how they were achieved. At a later stage, a distinction is made between Palmer's hydrological index and his meteorological index. Until that point is reached, the term 'PDSI' is utilized in a generic sense.

Drought as Defined by Palmer

Palmer defined a drought as a significant reduction of available moisture below that required for the near-normal operation of the established economy of a region. The qualification 'established economy' is important because it differentiates a drought from normal states of aridity such as deserts and semiarid regions as well as dry seasons. The specification of 'region' underscores spatial variations in water culture, which permits different degrees of dryness for drought to exist in different areas. For established economies, Palmer chose croppable land areas in the US Midwest.

PDSI is a meteorological concept grounded in detailed hydrological accounting. Hydrological accounting, also known as water balance, is a systematic numerical budget of moisture inflow, outflow, and storage. Based on Palmer's definition, drought occurs when water supply is significantly less than potential supply (water needed for normal operation of the established economy). Palmer determined potential supply as the amount of precipitation that is climatically appropriate for existing conditions (CAFEC) at a given place and time. The difference between the actual precipitation and the CAFEC precipitation is the deficit (negative) or surplus (positive). If $P_{i,j}$ is the actual precipitation in month j ($1, 2 \dots 12$) in year i ($1, 2 \dots n$), and $\hat{P}_{i,j}$ is the CAFEC precipitation equivalent, then the difference:

$$d_{i,j} = P_{i,j} - \hat{P}_{i,j} \quad [1]$$

is the deficit (surplus) for the time and place. It is the basic variable used for constructing the index. Palmer used the water balance equation to determine $\hat{P}_{i,j}$.

Water Balance Equation

The water balance equation for a natural surface is represented as:

$$P = ET + R + RO \quad [2]$$

where P is precipitation, ET is actual evapotranspiration, R is moisture storage in the soil (recharge), and RO is surface runoff. Actual evapotranspiration is the combined moisture loss to the atmosphere from vegetation and soil. In climatology, eqn [1] assumes the following sequence in the destination of precipitation: (1) loss to the atmosphere up to full atmospheric demand; (2) storage in the soil (soil water recharge) until filled to capacity; and (3) surface runoff – water transported over land – after sequence steps (1) and (2) are completed.

The term potential evapotranspiration (PE), also called atmospheric demand, is the maximum amount of water lost to the atmosphere through evapotranspiration under the existing atmospheric condition. PE is achieved only when water from precipitation, soil storage, or both is enough to enable it. When water from the two sources is less than PE, a deficit occurs.

Palmer's use of the water balance equation

Palmer used monthly data for mean surface temperature and precipitation to solve the water balance equation in monthly time steps. The Thornthwaite method (Thornthwaite, 1948) was used to obtain potential evapotranspiration from mean monthly temperature. All precipitation was treated as rain. Snowfall was cumulated over the winter and its rainfall equivalent added to the soil in the spring. Palmer made several assumptions concerning soil moisture storage and its usage. The soil was divided into two layers. The top layer (Layer 1), roughly equivalent to plow depth, was assumed to contain 25 mm of available water (AW) at field capacity. Field capacity is the maximum volume of water the soil holds when fully drained of excess water. AW is the amount of moisture held in the soil between field capacity and wilting point (the minimum plant-extractable soil moisture content). Palmer assumed that the equivalent capacity for the second (underlying) layer (Layer 2) is site specific. All moisture is removed from Layer 1 before any is taken from 2. Withdrawal from the underlying layer is subject to increasing resistance as the

moisture level is drawn down. Palmer assumed that the rate of withdrawal is a function of potential evapotranspiration and of the ratio of the water content in that layer to the combined available water holding/moisture capacity (AWC) of both layers. The water balance over a crop at a specified time step is set up as follows. Input is precipitation. Crop water demand is potential evapotranspiration. If precipitation exceeds PE, actual evapotranspiration $ET = PE$ (Figure 2). The difference $P - PE$ is stored in the soil (recharge) up to field capacity. Runoff occurs after full soil recharge. If precipitation is insufficient to satisfy PE, water is drawn from the soil layers as specified here until PE is achieved. When that happens, $ET = PE$. However, if after withdrawal, water is still insufficient to satisfy PE, the actual evapotranspiration (ET) is the sum of precipitation and soil withdrawal and ET is less than PE. The difference $PE - ET$ is the deficit (Figure 2). A surplus (runoff) occurs when precipitation exceeds the sum of potential evapotranspiration and amount of moisture required to bring the soil to field capacity. Equation [2] is reformulated to accommodate the exchanges discussed here:

$$ET = P - R - RO + L \quad [3]$$

where L is loss from the soil (water withdrawal from storage).

Palmer's challenge was to:

1. adapt data derived from eqn [3] for use within his CAFEC environment;

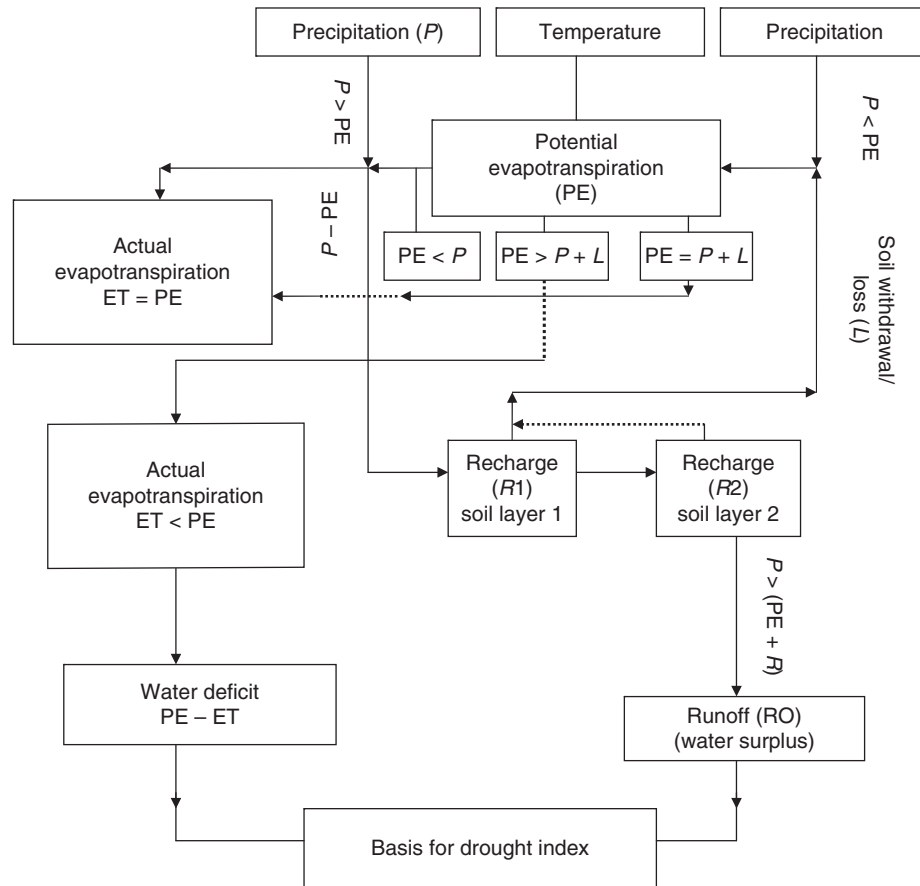


Figure 2 A water balance flow chart.

2. establish a numerical scale for the assessment of drought (wet period) severity based on the water balance (this water balance-based scale is the Palmer hydrological index (PHDI));
3. determine the onset and end of a drought (wet period); and
4. refine the PHDI to improve its meteorological sensitivity. The refined product, which includes a mechanism for determining the onset and end of a drought (wet period), is the PDSI.

The steps taken to meet the challenge and the supporting arguments are covered in Sections [Determination of CAFEC Values](#), [Calculating the Palmer Hydrological Drought Index](#), [Ending a Drought \(Wet Period\)](#), and [Palmer's Meteorological Drought Severity Index](#).

Determination of CAFEC Values

Palmer utilized several variables to calculate CAFEC precipitation $\hat{P}$ for a region. In addition to PE, ET, R , RO, and L , he introduced three new ones, namely, potential recharge (PR), potential loss (PL), and potential runoff (PRO). The term 'potential' refers to the expected normal for the region. The physical meaning of the additional terms is given in [Table 1](#).

Palmer used the variables to derive four coefficients: α is the coefficient of evapotranspiration, also referred to as the coefficient of climatically appropriate moisture efficiency; β the coefficient of recharge; γ the runoff coefficient; and δ the coefficient of loss.

$$\alpha_j = \overline{ET_j}/\overline{PE_j} \quad \text{if } \overline{PE} > 0, \alpha_j = 1 \text{ otherwise} \quad [4]$$

$$\beta_j = \overline{R_j}/\overline{PR_j} \quad \text{if } \overline{PR_j} > 0, \beta = 0 \text{ otherwise} \quad [5]$$

$$\gamma_j = \overline{RO_j}/\overline{PRO_j} \quad \text{if } \overline{PRO_j} > 0, \gamma_j = 0 \text{ otherwise} \quad [6]$$

$$\delta_j = \overline{L_j}/\overline{PL_j} \quad \text{if } \overline{PL} > 0, \delta_j = 0 \text{ otherwise} \quad [7]$$

In eqns [4]–[7], an overbar denotes monthly means obtained over the years of data.

Based on these calculations, Palmer reformulated eqn [3] to obtain the CAFEC precipitation for a place:

$$\hat{P}_{ij} = \alpha_j PE_{ij} + \beta_j PR_{ij} + \gamma_j PRO_{ij} - \delta_j PL_{ij} \quad [8]$$

CAFEC precipitation for a given month may differ from the long-term mean. If the preceding months have been hot and dry, the stored moisture will be nearly depleted, and CAFEC

precipitation, as well as water demand, will be high. The difference between actual precipitation and CAFEC precipitation d_{ij} will be negative with a high absolute value, as the deficit with respect to demand may have been much greater than the deficit with respect to the long-term average precipitation. Similarly, above normal precipitation in a month when the soil moisture is at capacity and PE is low will show as a greater surplus than if the soil moisture had been close to depletion and PE was high.

Palmer further recognized that two regions with the same deficit might not suffer identical economic consequences if they had different moisture supply bases. The region with the lower supply base will be more stressed. To discriminate with respect to differences in impact over the two regions, Palmer calculated k_j , the ratio of the mean demand to the mean supply for a given place and month, which he used to adjust the d_{ij} to reflect the disparity arising from differences in supply. Thus:

$$k_j = (\overline{PE_j} + \overline{R_j})/(\overline{P_j} + \overline{L_j}) \quad [9]$$

The moisture anomaly index z_{ij} was formulated to provide the desired spatial weighting for d_{ij} :

$$z_{ij} = d_{ij} k_j \quad [10]$$

Calculating the Palmer Hydrological Drought Index

Generally, droughts become more severe the longer they last. Thus, cumulative plots of z at a place not only capture the magnitude of an event but also reflect its duration. Based on similar cumulative plots over two locations, Palmer determined that drought severity could be adequately represented by four distinct classes (mild, moderate, severe, and extreme) to which he assigned values based on his cumulative plots ([Table 2](#)). The four classes of drought were separated from the opposite classes of wetness by normal conditions. Palmer reconciled the effect of intense short-period deficits against those produced by less intense but longer-lasting ones by considering maximum rates of accumulation of z at different stations. He concluded that the accumulation of z at a rate of -12.0 for 1 month or -85.0 over 60 months would constitute extreme drought. He used this argument to derive a quantitative expression for drought severity X_g :

$$X_g = \sum_{j=1}^g \frac{z_{ij}}{0.309j + 2.691} \quad [11]$$

Table 1 Partial list of variables for computation of Palmer's CAFEC precipitation

Symbol	Description	Physical meaning
PR	Potential recharge	Amount of water needed to bring the soil to field capacity
PRO	Potential runoff	Total amount of soil moisture storage available
PL	Potential loss	Amount of water that could be lost from the soil if no precipitation occurred during the month

Table 2 Palmer drought severity index for moisture conditions

Index value	Moisture condition
Less than -4	Extreme drought
-3.0 to -4.0	Severe drought
-2.0 to -3.0	Moderate drought
-1.0 to -2.0	Mild drought
-1.0 to 1.0	Normal condition
1.0 – 2.0	Mild wet spell
2.0 – 3.0	Moderate wet spell
3.0 – 4.0	Severe wet spell
Greater than 4.0	Extreme wet spell

where g is the target (current) month and $j = 1, 2, \dots$ (eqn [11]) accumulates weighted values of z beginning with the first month ($j = 1$) of drought (wet period) to the target month ($j = g$).

Recognizing that a single very dry or very wet month in a long series of months with opposite moisture conditions may unrealistically reflect in the severity value, Palmer revised eqn [11] to include each month in the index in an incremental basis. The revised equation:

$$X_g = 0.897X_{g-1} + Z_g/3 \quad [12]$$

accomplishes two things. First, the duration factor is included implicitly through the sequentially arranged values of X . Second, the severity index acquires a memory through the recursive term X_{g-1} , the previous month's severity.

Based on eqn [12], the values of X_g will range from greater than +4 (very much wetter than normal) to -4.0 (very much drier than normal, or extreme drought) with values close to zero constituting near-normal conditions.

Reevaluating the Spatial Weighting Factor

Palmer's formulas were created with data from Kansas and Iowa. When applied to other climatic regions, the original expression for spatial weighting produced unrealistic values. This discrepancy led Palmer to modify his approach to spatial weighting. He assumed that for a given area, the driest 12-month period on record represents extreme drought. He next determined that the accumulated $z_{i,j}$ value, which would yield $X = -4.0$ from eqn [12], was -25.60. A new value of the 12-month mean weighting factor $\bar{K}$ could then be determined by dividing the 12-month sum of $d_{i,j}$ for the driest 12 months on record by -25.60. By experimenting with several stations outside Iowa and Kansas, he found that values of $\bar{K}$ based on long-term records varied from 1.06 in humid areas to 1.73 in drier ones. From this, he concluded that $\bar{K}$ and its monthly equivalent K_j depend on average water supply conditions, as did the k values. This led Palmer to further modify his approach to regional weighting by adding average runoff to the moisture demand term. He discovered that with this inclusion, $\bar{K}$ varied inversely with $\bar{D}$, the mean of the absolute values of $d_{i,j}$. Plotting all his experimental values on a graph, he concluded that a new monthly weighting value

$$K_j = 1.5 \log_{10}\{(S_j + 2.8)/D_j\} + 0.50 \quad [13]$$

where

$$D_j = \sum_{i=1}^n (|d_{ij}|)/n \quad [14]$$

and

$$S_j = (\overline{PE_j} + \overline{R_j} + \overline{RO_j})/(\overline{P_j} + \overline{L_j}) \quad [15]$$

improved the spatial adjustment factor k_j . The final form of the spatial weighting factor became

$$K = 17.67K' / \left(\sum_{j=1}^{12} \overline{DK'} \right) \quad [16]$$

Following the modifications, a new set of values $Z_{i,j}$ was calculated to define moisture anomalies normalized with respect to both time and space:

$$Z_{ij} = K_j d_{ij} \quad [17]$$

which replaced the $z_{i,j}$ in eqn [10]. The X_g calculated with the new $Z_{i,j}$ was Palmer's first approximation to his drought severity index and is often considered the PDSI by many researchers and used as such. However, at this point the index could more properly be considered a hydrological drought index (PDHI) since it consists basically of the systematic moisture budget referred to in the Section [Drought as Defined by Palmer](#).

Ending a Drought (Wet Period)

A wet period or drought is considered to have ended when X_g approaches zero. At that point, moisture demand is satisfied. But it may not endure. A return to normal weather following readjustments of large-scale or regional-scale atmospheric circulation patterns will have to persist much longer in order for this index to truly reflect a return to normal conditions. Therefore, in meteorological terms, the criteria for ending a drought (wet period) are more stringent than for achieving X_g close to 0. Consequently, another method is needed to establish the beginning and ending of dry and wet periods. Also, requiring the severity index to drop all the way to zero or near zero could result in an unjustified perception that a mild drought persisted over many years if followed by slightly dry or even normal weather. However, if the drought index was required to reduce to only -1.0, a single normal month could lead to a premature end of a mild drought that could still evolve into a severer one.

Palmer conceded that drought would most reasonably have ended with the index somewhere between -1.0 and 0.0, and chose -0.5. Thus, as the index reaches the 'near-normal' category and the drought or wet spell is considered terminated, Palmer then asked the question of how much moisture would be required to reduce the index value to -0.5. Allowing X_g to equal -0.5, Palmer solved for this value, which he assigned the symbol Z_e ,

$$Z_e = -2.6619 X_{g-1} - 1.50 \quad [18]$$

for a dry spell, and

$$Z_e = -2.6691 X_{g-1} + 1.50 \quad [19]$$

for a wet one.

Also, there is some smallest value of Z , which could occur month after month and produce X_g equal to -0.5. In such cases, using eqn [11], ΔX would be zero, $X_g - 1$ would be -0.5, and Z_g would be -0.15. This value indicates that a drought may end even if the weather is consistently slightly drier than normal. Therefore, once a drought has definitely begun ($X_g \leq -1.0$), any value of Z greater than -0.15 should end it. Palmer could then answer the question that he posed by calculating U_w , or effective wetness, which would apply *only* once $X_g \leq 1.0$.

$$U_w = Z_g + 0.15 \quad [20]$$

Assume that for month g , $X_g \leq -1.0$. If, during that month, the amount of wetness required to end the drought is greater than the effective wetness ($Z_e > U_w$), the drought severity index will be lower than for month $g-1$, but the drought will not have ended. However, the amount of wetness required to end the drought during the following month, $g+1$, will have been diminished by the accumulated wetness such that a new value of Z_e must be calculated.

To enable that calculation, Palmer redefined Z_e . If $x_j \leq -1.0$ for all months h from h^* , the first month of the drought through g , the current drought month, then for month g ,

$$Z_{e_g} = Z_{e_{(g-h)}} + \sum_{h=0}^{h=h^*} (U_{(g-h)} - U_g) \quad [21]$$

U_g is U_w in month g and $h^* \leq h \leq g$. The drought duration is $(g - h^* + 1)$ Z_e computed for month g during any dry spell then represents that amount of wetness needed to end the drought in that month. The value U_g for the month g is the effective wetness, the amount that actually occurred. By keeping track of accumulated values of U as well as the current Z_e , Palmer was able to determine a percentage probability Pe_g that a drought has ended in month g . Pe_g is not a true stochastic probability but rather a ratio of the amount of moisture received in month g to that required to end the drought in that month, expressed as a percentage. Pe_g is formulated as follows:

$$Pe_g = \left(100 \sum_{h=0}^{h=h^*} U_{g-h} \right) / \left(Z_e + \sum_{h=1}^{h=h^*} U_{g-h} \right) \quad [22]$$

where $h^* \leq h < g$.

This probability may then be used as a basis for defining drought ending times, particularly when the physical sense of the index is ambiguous.

Palmer's Meteorological Drought Severity Index

Once a method for defining the ending of droughts and wet spells was established, Palmer further refined the index to make it more sensitive to meteorological conditions. The final product with the added meteorological sensitivity is the true PDSI. He accomplished this by keeping track of all the previous dry and wet spells using three different indices: $X1_{ij}$, the severity index of a wet spell that is becoming established; $X2_{ij}$, the severity index for a dry spell that is becoming established; and $X3_{ij}$, the severity index for any wet or dry spell that has become established. The three indices are calculated as follows:

$$X1_g = 0.897X1_{g-1} + Z_g/3 \quad \text{if } X1_g > 0; \text{ otherwise } X1_g = 0 \quad [23]$$

which gives positive and nonzero values for wet spells, and

$$X2_g = 0.897X2_{g-1} + Z_g/3 \quad \text{if } X2_g < 0; \text{ otherwise } X2_g = 0 \quad [24]$$

which gives negative and nonzero values for any dry spell.

If $Pe_g < 100$, then

$$X3_g = 0.897X3_{g-1} + Z_g/3 \quad [25]$$

Otherwise, $X3_g = 0$, which gives nonzero positive or negative values only when the probability of ending the current wet or dry spell is less than 100%.

Equations [23]–[25] follow definitions of the three indices. However, assigning a value to $X3$ for the beginning of an established wet or dry spell requires changing from one index to another, as follows:

$$\text{If } X3_g = 0 \text{ and } X1_g > 1 \text{ then } X3_g = X1_g \quad [26]$$

Equation [26] indicates that a wet spell has become established for the first time. Similarly:

$$\text{If } X3_g = 0 \text{ and } X2_g < -1 \text{ then } X3_g = X2_g \quad [27]$$

Equation [27] identifies the start of a dry spell.

To assign a value to the actual meteorological index (i.e., the PDSI_g for month g), one of the three indices defined in eqns [23]–[25] must be chosen. In some cases, the procedure is straightforward, but in others, confusion may arise. Palmer used examples from a long-running data set from western Kansas to demonstrate whether the assigned value represents the beginning of an established wet spell or a dry spell. When Pe drops to zero, the PDSI value for that month is the $X3$ value. When Pe is 100%, the current spell has ended, and the PDSI value will have the value of $X1$ or $X2$, whichever is opposite in sign from the spell that just ended. Problems arise when Pe is between 0 and 100 and when both $X1$ and $X2$ have absolute values less than 1.0, indicating that no serious wet or dry spell is underway. Palmer's solution is to continue assigning values to both the $X1$ and $X2$ indices according to the criteria established in this article, but delay assigning a value to PDSI until $X1$ and $X2$ become zero at some future month. If $X2$ drops to zero first, then the $X1$ value is given to the unassigned cases, and a mild wet spell is considered to have occurred. Palmer himself considered this method to be a compromise. However, to calculate the PDSI through Palmer's method during times when both $X1$ and $X2$ are indicating low absolute values requires backtracking and, very occasionally, subjective decisions. While this does not present major problems when working with fewer than 10 data sets listed in tabular format, difficulties encountered when electronically processing scores of sets have led to a number of modifications. For example, the *Weekly Weather and Crop Bulletin* assigns values to PDSI_g based on Pe_g , $X1_g$, $X2_g$, and $X3_g$ as follows:

$$1. \quad \text{If } 0 < Pe_g \leq 50, \text{ then } PDSI_g = X3_g \quad [28]$$

This constraint covers the case when $X3$ is nonzero, and a wet or dry spell has been established, but the probability of the spell ending, while not zero as yet, is low. Had this constraint been applied to Palmer's western Kansas example, one 8-month dry spell would have lasted 10 months:

$$2. \quad \text{If } 50 < Pe_g < 100, \text{ then } PDSI_g = X1_g \text{ or } X2_g \quad [29]$$

whichever results in an index having the opposite sign from $X3_g$. This change in sign would indicate a switch from a dry to a wet spell or from a wet to a dry spell.

The drought-ending criterion is as follows: once the probability of ending the wet or dry spell exceeds 50%, the value of $X1$ or $X2$, which would indicate a change of weather conditions, is used as the PDSI value. Had this been used in Palmer's

example, one 4-month near-normal wet spell would have been a 3-month near-normal wet spell followed by a single near-normal dry spell.

$$3. \text{ If } X_{3g} = 0, \text{ then } PDSI_g = X_{1g} \text{ or } X_{2g} \quad [30]$$

whichever has the larger value. Through this criterion, the case of $P_e = 100\%$ is covered, as X_{3g} would in that case be zero. Also, since the constraint of $X_{3g} = 0$ means that the absolute values of X_{1g} and X_{2g} are both less than 1.0, the case in which Palmer required backtracking and occasional arbitrary decisions is resolved. Had this criterion been applied to Palmer's example, the results would be unchanged.

Criticism of the Palmer Index

The PDSI has been criticized for various reasons. Alley (1984) and Karl and Knight (1985) listed the following issues, which are deemed drawbacks:

1. The values quantifying the intensity of drought and signaling the beginning and end of a drought or wet spell were arbitrarily selected based on Palmer's study of central Iowa and western Kansas and have little scientific meaning.
2. Because the Index is sensitive to the available moisture capacity (AWC) of a soil type, applying it to a climate division may be too general.
3. The two soil layers within the water balance computations are simplified and may not be accurately representative of a location.
4. Snowfall, snow cover, and frozen ground are not recognized in the Index. Since all precipitation is treated as rain, the timing of PDSI or PHDI values may be inaccurate in the winter and spring months in regions where snow and frost occur.
5. The natural lag between when precipitation falls and the resulting runoff is not considered. In addition, no runoff is allowed to take place in the model until the water capacity of the surface and subsurface soil layers is full. This could lead to an underestimation of runoff.
6. Potential evapotranspiration is estimated using the Thornthwaite method. While this technique has wide acceptance, it is still only an approximation.

Whereas these comments and criticisms have earned respect within a large section of the scientific community, there is also widespread support for Palmer's approach and methodology. What follows reflects the views of the latter group.

Palmer may have 'arbitrarily' tuned his model to obtain agreement with observed data from Kansas and Iowa, but model tuning is widely used in science to enhance robustness. The accusation of arbitrariness appears extreme given his meticulous argumentation and inclusion of data from several climatically diverse regions outside the two states. Successful results achieved with the model in many world regions have given it added credibility. As well, work aimed at overhauling Palmer's criteria for the onset and end of droughts and wet periods has produced very modest improvements (see Section [Palmer's Meteorological Drought Severity Index](#)). The issue concerning snow and AWC would be serious if the index

was constructed as a short-term moisture status indicator. However, because Palmer targeted the longer term (months to years), time lag between input and output (snow, soil moisture change, and runoff) is not a significant factor in Palmer's drought scaling and timing (Weber, 1996). Even though more theoretically based models normally yield better potential evapotranspiration data, van der Schrier et al. (2006) show that PDSI estimates computed with the Penman-Montieth and Thornthwaite methods are very similar. Dai et al. (2004) reported significant agreement between observed soil moisture and Palmer model-based equivalents in the United States, Mongolia, China, and the former Soviet Union.

Current Applications

The PDSI is widely used as an environmental moisture indicator. Historically, the PDSI was used to classify, map, and interpret past drought events. While it still serves those purposes, today it is increasingly used in a forecast mode. For example, Drought Monitor, an organization comprising federal and state agencies, universities, and the private sector, provides monthly forecasts of moisture conditions in the United States based on the Index. The US National Oceanic and Atmospheric Administration publishes weekly Palmer maps. PDSI is also used to reconstruct past droughts at geological time scale based on temperature and precipitation derived with dendrochronology. Such reconstructed drought patterns have been used to explain the emergence and collapse of civilization, including the demise of the Bronze Age. Historical patterns of soil moisture derived through the Index help inform the contribution of soil moisture to global warming.

Conclusion

Palmer's drought index fulfills most of the criteria specified for an ideal index. Perhaps his most important contribution to moisture scaling is giving quantitative expression to differing thresholds for drought across regions as well as land use and economic practices. While the assumptions may appear too numerous and simplistic to some scientists, several experiments designed to test them have either confirmed the robustness of Palmer's approach or suggested modifications that are not significant enough to alter the fundamentals of his methods. But changes are being made and will continue to be made in how data are acquired and processed for PDSI. Remote sensing enables data acquisition from large geographical areas at high frequency. Several software packages are available for computing the index for various time and spatial scales. Used within a Geographical Information System environment, the PDSI can facilitate spatial and temporal analyses of drought pattern anywhere. Although the PDSI has been used primarily to study past drought events, its use as a forecasting tool for future events is expanding. The flexibility offered by the CAFEC concept makes it an ideal tool for constructing water budget scenarios for climate change impacts.

See also: Agricultural Meteorology and Climatology. **Hydrology, Floods and Droughts:** Deserts and Desertification; Drought; Overview.

Further Reading

- Alley, W.M., 1984. The Palmer drought severity index: limitations and applications. *Journal of Applied Meteorology* 23, 1100–1109.
- Alley, W.M., 1985. The Palmer drought severity index as a measure of hydrological drought. *Water Resource Bulletin* 21, 105–114.
- Dai, A., Trenberth, K.E., Qian, T., 2004. A global dataset of Palmer severity drought index for 1870–2002: relationship with soil moisture and effects on surface warming. *Journal of Hydrometeorology* 5, 1117–1130.
- Karl, T.R., 1983. Some spatial characteristics of drought duration in the United States. *Journal of Climate and Applied Meteorology* 22, 1356–1366.
- Karl, T.R., Knight, R.W., 1985. Atlas of Monthly Palmer Hydrological Drought Indices (1931–1983) for the Contiguous United States. Historical Climatology Series 3–7. National Climatic Data Center, Asheville, North Carolina.
- Nkemdirim, L.C., Weber, L., 1999. Comparison between the droughts of the 1930s and the 1980s in the southern prairies of Canada. *Journal of Climate* 12, 2434–2450.
- Palmer, W.C., 1965. Meteorological Drought. Research Paper No. 45. US Department of Commerce Weather Bureau, Washington, DC.
- Palmer, W.C., 1968. Keeping track of crop moisture conditions nationwide: the new moisture index. *Weatherwise* 21, 156–161.
- Thornthwaite, C.W., 1948. An approach toward a rational classification of climate. *Geographical Review* 38, 55–94.
- van der Schrier, G., Briffa, K.R., Jones, P.D., 2006. A Global Dataset of the Palmer Drought Index – Sensitivity to the Potential Evapotranspiration. Koninklijk Nederlands Meteorologisch Instituut, p. 13.
- Weber, L., 1996. Drought Dynamics in the Southern Canadian Parries. PhD dissertation. University of Calgary, Calgary, Alberta, Canada. p. 272.

Soil Moisture

A Robock, Rutgers University, New Brunswick, NJ, USA

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

Soil moisture is an important variable in the climate system. Understanding and predicting variations of surface temperature, precipitation, drought, flood, and the impacts of future climate change depend critically on knowledge of soil moisture variations. There are few long-term observations of soil moisture. Remote sensing of soil moisture is in its infancy and can observe only the wetness of the top centimeter or two of soil. The production of a global soil moisture data set will require an excellent land surface model and data assimilation system, validated with observations, that calculates soil moisture based on soil moisture observations and observations of precipitation, temperature, insolation, and other weather variables.

Introduction

Soil moisture is the amount of water in the active layer of the soil, typically the top 1–2 m. Soil moisture is extremely important because it is the main source of water for agriculture and natural vegetation. Near-surface soil moisture also controls the partitioning of available energy at the surface into sensible and latent heat exchange with the atmosphere, thus linking the water and energy balances through the moisture and temperature states of the soil. Soil moisture is the source of water that evaporates and transpires from the soil and vegetation into the atmosphere, thus affecting the distribution of clouds and precipitation. Surface temperature is controlled by soil moisture, as a wetter surface will be cooler, with more of the available energy going into evapotranspiration (evaporation and transpiration) rather than heating the surface. Soil moisture also affects runoff, determining how much precipitation or snowmelt goes immediately into rivers and streams, or in extreme cases into flooding. A deficit of soil moisture is often connected to 'drought,' and soil moisture interactions with the atmosphere may be important in maintaining droughts. Soil moisture, along with snow cover, is also the most important component of meteorological memory for the climate system over the land. Thus, the soil moisture state is an important predictor of monthly to seasonal climate variations.

Soil moisture is measured in a variety of ways, from the simple gravimetric technique to more complex electronic instruments. For the top 1 m soil layers, the temporal scale of soil moisture variation in the Northern Hemisphere midlatitudes is 1.5–2 months and the spatial scale is about 500 km. Despite the importance of soil moisture, parameterization schemes used in weather forecast and climate simulation models, in general, do not capture the observed soil moisture variations when forced with either model-generated or observed meteorology. During the twenty-first century, a significant vulnerability for human society may be summer drying in the subtropics and midlatitudes. Observations of summer soil moisture variations for the past several decades, however, show that for the stations with the longest recorded summer, soil moisture in the top 1 m has increased while temperatures have risen. In Eastern Europe, where these observations were taken, increasing tropospheric pollution following World War II produced solar dimming, reducing the

evaporative demand at the surface. This effect dominated any trends in temperature or precipitation.

Definition of Soil Moisture

The term 'soil moisture' refers to the amount of water in the upper layer of soil that interacts with the atmosphere. This varies depending on soil type and vegetation, but is typically about the top 1 m. Soil moisture can be expressed in different units. The most common are as plant-available volumetric soil moisture (W) or as total volumetric soil moisture (W_T), expressed as the depth of a column of water contained in a given depth of soil, or as the volumetric percentage of water in a given soil depth. A fraction, typically less than half, of soil consists of pores that can be filled with air or water. This fraction is called the *porosity* (P). If this fraction were completely filled with water, the soil would contain its *total water-holding capacity* (W_o), and the water table would be at the surface. For any layer of depth (thickness) D , $W_o = PD$. If the soil were saturated, so that $W_T = W_o$, and then gravitational drainage were allowed to occur until it was negligible, the amount of water left in the soil is called the *field capacity* (W_f). If vegetation then extracted as much water as possible until it wilted, the remaining soil moisture is called the *wilting level* (W^*), and this amount of water is unavailable to plants. The plant-available soil moisture is $W = W_T - W^*$.

Measurement of Soil Moisture

There are many different techniques to measure soil moisture. The choice of a particular method depends on the application and the resources available. Here, the principal techniques are briefly described.

Gravimetric

The gravimetric method, also called the thermostat-weight technique, has been in use for a long time. Soil samples are taken using coring devices or augers at required depths and locations. Typically (in the Russian method), 10 cm segments reaching a depth of 1 or 1.5 m are extracted and a smaller

sample is removed from each segment. The sample is weighed, oven-dried, and weighed again. The difference in mass gives the total soil moisture in the sample, which is converted to volumetric units using the density of the soil. The wilting level is then subtracted, giving plant-available soil moisture, expressed as depth of liquid water.

Under the leadership of the State Hydrological Institute, the Russians established a system of soil moisture observations throughout the former Soviet Union. The network consisted of several thousand stations using the gravimetric technique. Typically, samples are taken at four different locations at a station each 10 days. After 10 days, the cores are replaced in the holes and samples are taken from other holes. In this way, after some time the effects of sampling disappear and the same location can be reused. In very cold regions, observations were taken either once a month or not at all. Other countries that were influenced by the Soviet Union also set up soil moisture observation programs, and extensive data sets now also exist for China and Mongolia.

The gravimetric method is low-tech and simple, making it an excellent technique for long homogeneous climatological records. As it is labor-intensive and somewhat destructive to its site, new electronic methods are being introduced that are indirect and require calibration and theoretical assumptions. With suitable parallel measurements, the new methods can produce useful long records, and they are briefly described in this article. However, all the electronic methods suffer from some problems and require extensive calibration with gravimetric measurements that span the time scales of the seasonal cycle and interannual variations. Furthermore, hysteresis effects may give ambiguous readings, and some techniques may be less accurate in certain soils and for extremely dry or wet soil.

Neutron Probe

The neutron probe is relatively easy to use, accurate, and capable of measurements in real time. A probe with a fast neutron source is placed on the surface or lowered in an access tube (transparent to the neutrons), and the back-scattered slow neutrons are measured. The back-scattered flux of slow neutrons is proportional to the density of hydrogen atoms. Water is the major source of hydrogen atoms that changes with time; therefore, the neutron probe provides a good measure of soil water content. Calibration of slow neutron counts with gravimetric samples of soil moisture content and bulk densities yields a relationship to estimate the volumetric soil moisture content. Since radioactive scattering occurs over a spherical domain, a neutron probe samples a volume of soil rather than a point. As this volume of influence depends on soil moisture content, there are differences in the soil volumes sampled in dry and wet soils. These differences are generally small as compared to the total volume sampled by the probe, but they influence the depth resolution of the probe. The probe's relatively large volume of influence makes observations at shallow depths prone to errors, as adjoining air is also sampled. Disadvantages of neutron probes include that they are also labor intensive, the need for precautions associated with handling radioactive material, and the relatively high costs. An extensive data set of soil moisture observations from Illinois

starting in 1981 was made with neutron probes, with observations made 2–3 times per month year round.

Heat Dissipation Sensors

These sensors make point measurement of soil moisture tension by measuring temperature changes in response to a heat pulse. A small ceramic block with an embedded sensor is briefly heated. The rate of heating is affected by the ability of the ceramic block to dissipate the heat. This is related to its soil moisture content, which has equilibrated with the surrounding soil. The measured heating rate must be calibrated, and soil moisture tension related to volumetric water content, by gravimetric observations for each location or with theoretical relationships. These sensors are relatively inexpensive and can produce measurements every 30 min. An extensive network in Oklahoma and Kansas now uses these sensors. Once installed and calibrated, a very labor-intensive effort, an advantage of these sensors is that they are automatic and can take observations of both soil temperature and soil moisture with high temporal resolution.

Other In Situ Sensors

Other in situ soil moisture measurement techniques include the tensiometer (a bulb of porous ceramic material is placed inside the soil and connected to a water-filled tube, which is used to measure soil moisture tension after allowing the system to equilibrate), the gypsum block (small cylindrical gypsum blocks embedded with electrodes are buried at required depths in the soil and measure the electrical resistance, which is related to the water content), time domain reflectometry (TDR; based on monitoring changes in the dielectric properties of the soil at microwave frequencies), frequency domain reflectometry (FDR; similar to TDR, except that it derives soil moisture content based on changes in the frequency of signals due to the dielectric properties of the soil), and gamma densitometry (based on the relatively greater gamma radiation attenuation factor of water compared to other soil components). Each technique has limitations and advantages.

Remote Sensing of Soil Moisture

If radiance or gamma rays from soil moisture can be observed above the soil and related to the soil moisture, using gravimetric observations for calibration, then remote sensing can be used over a much larger region than in situ observations. Satellites have the potential to produce global soil moisture observations, but such observations are restricted to the top few centimeters of the surface. Reflections of global positioning system (GPS) signals using GPS sensors mounted near the ground or measuring cosmic rays from a sensor mounted near the ground have the potential for measuring a deeper layer (up to 50 cm) over a small area near the sensor.

Satellite remote sensing of soil moisture

Satellite remote sensing of soil moisture works on the principle that microwave emissions from the soil depend on the soil moisture because of the large difference in dielectric constant between dry and wet soil. However, the emission also depends

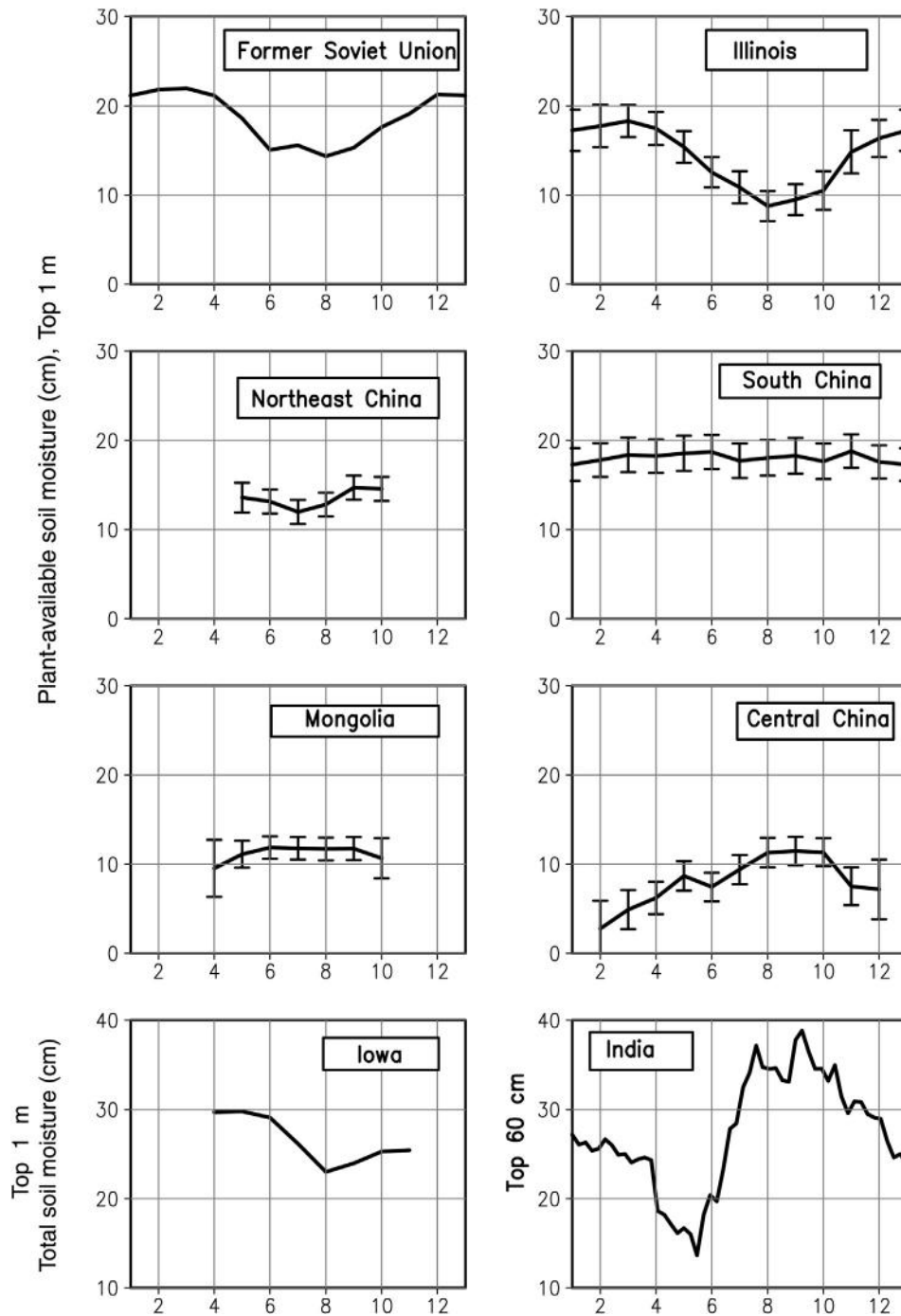


Figure 1 Average seasonal cycles of soil moisture, averaged for the entire length of each data set (Figure 2). Units are centimeters of water in the top 1 m, except the top 60 cm for India. The shape depends on the seasonal cycles of precipitation and evapotranspiration. For those regions where optimal averaging was used, whiskers indicate the estimated one standard deviation error range associated with the spatial distribution of the stations. Reproduced from Robock, A., Vinnikov, K.Y., Srinivasan, G., et al., 2000. The global soil moisture data bank. *Bulletin of the American Meteorological Society* 81, 1281–1299. Copyright 2000 American Meteorological Society.

on surface roughness, surface temperature, and vegetation, so ancillary information is needed, and remote sensing under a dense canopy is not possible. The longer the wavelength of the radiation, the deeper the potential signal. Typically, C-band (frequency ≈ 6 GHz, wavelength ≈ 5 cm) and L-band (frequency ≈ 1.3 GHz, wavelength ≈ 23 cm) radiometers have

been used on satellites. L-band observations have the potential to sample more of the top few centimeters of soil but require larger antennas on the satellites. Passive observations have footprints several tens of kilometers across and repeat cycles of about 3 days, while active observations (radar) have smaller footprints but sample the ground much less frequently. The

most advanced satellite now measuring soil moisture is the Soil Moisture Ocean Salinity (SMOS) satellite that was launched by the European Space Agency in late 2009 and makes passive L-band measurements. NASA plans to launch the Soil Moisture Active & Passive (SMAP) satellite in late 2014, which will use both active and passive L-band measurements. By combining observations from different satellites and validating them with in situ observations, it is possible to monitor flood and drought regions and get useful information on the wetness of the soil surface. Visible and infrared radiation can also be used for indirect satellite soil moisture monitoring, but cannot make measurements when cloudy. Microwave remote sensing offers the most promise for future global data sets but will have to be combined with validated land surface models using data assimilation to produce soil moisture fields for the entire active soil layer.

GPS soil moisture observations

GPS satellites transmit L-band signals that are used for navigation. By comparing the direct microwave signal received at a GPS receiver mounted on a pole in an open field with the signal reflected from the ground, the soil moisture in the top several tens of centimeters can be measured. Since satellites pass over a particular location in different directions, soil

moisture can be observed in different directions from the receiver, and a 1000 m² region around the receiver can be sampled. This technique is currently being tested and has the potential to measure soil moisture anywhere that GPS receivers are installed near the soil for other purposes, such as geodetic observations. GPS receivers can also be installed on airplanes to measure large areas with the same technique.

The COsmic-ray Soil Moisture Observing System (COSMOS)

The intensity of low-energy cosmic-ray neutrons above the ground is inversely correlated with soil water content. By measuring these neutrons, it is possible to measure the average soil moisture to depths of 50 cm over a region of about 350 000 m². This system is now being tested and has the possibility of measuring soil moisture over large regions more easily than with in situ sensors.

Soil Moisture Variations

Scales of Temporal and Spatial Variations of Soil Moisture

It is well known that the complex topography of natural landscapes, with spatially variable vegetation and soil types, and gravitational drainage and infiltration of water after heavy

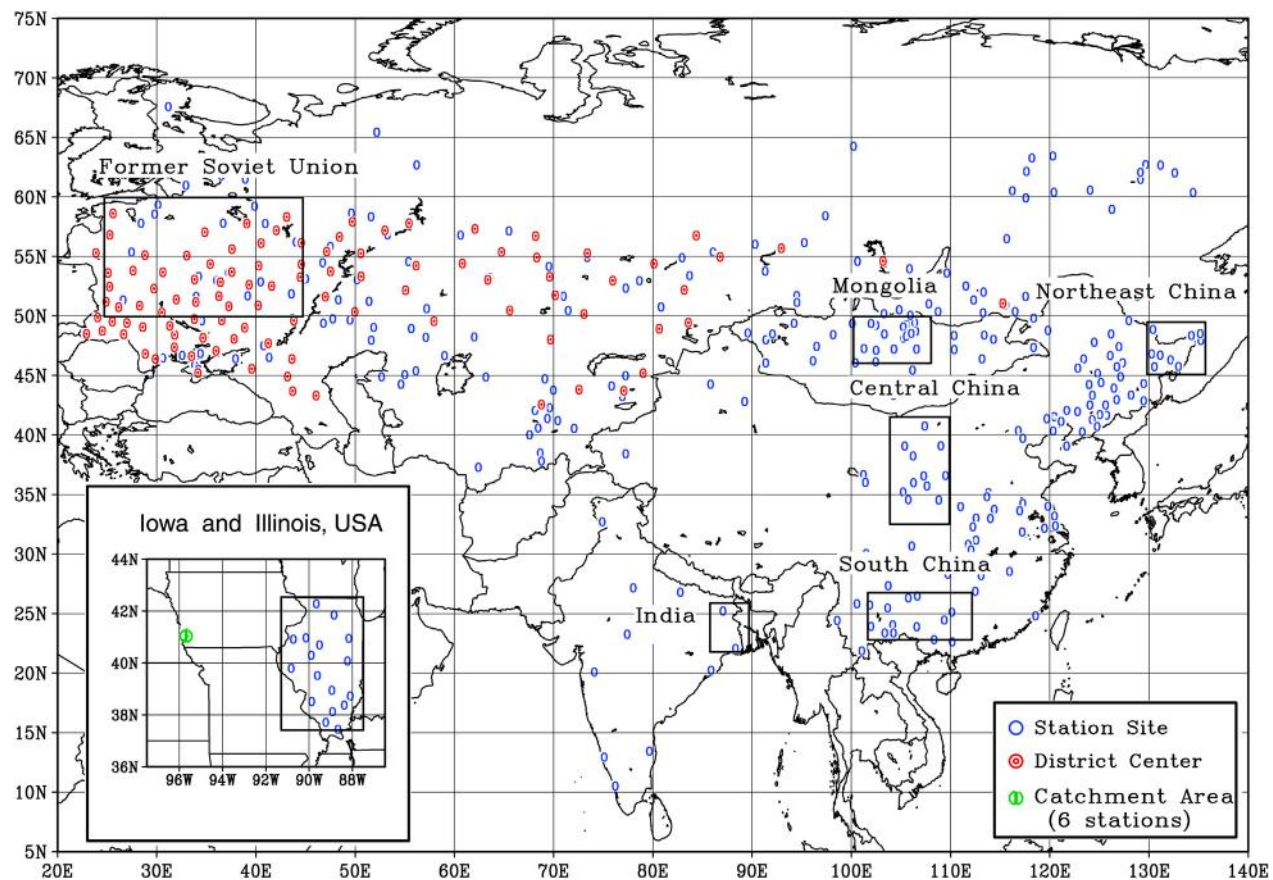


Figure 2 Map of the distribution of the stations in the Global Soil Moisture Data Bank, and location of the regions used for sample seasonal cycle plots (Figure 1). Part of the Russian data represents averages for administrative districts, rather than individual stations, and is indicated as circled dots. Reproduced from Robock, A., Vinnikov, K.Y., Srinivasan, G., et al., 2000. The global soil moisture data bank. *Bulletin of the American Meteorological Society* 81, 1281–1299. Copyright 2000 American Meteorological Society.

rains, is responsible for very small-scale spatial (tens of meters) and temporal (up to a few days) variability in the soil moisture field. This scale of soil moisture variation is of intense interest to hydrologists, who connect it to variations of the water budget within catchments. When studying climate change, however, a larger scale is of interest, driven by fields of precipitation and evapotranspiration. In addition to this small-scale component of soil moisture variability, analysis of spatial fields and time series of soil moisture observations in the Northern Hemisphere midlatitudes also finds a long-term (about 1–4 months) and large-scale (about 400–800 km) signal related to atmospheric forcing. The meteorological component of soil moisture field variability has been found in observations and later received theoretical explanation. For example, the temporal scale is equal to the field capacity divided by the potential evaporation. The spatial structure of actual soil moisture obviously depends on the distribution

of topography, soils, and vegetation on all scales, but the scale of soil moisture variations can be related to the scale of atmospheric forcing.

Climatology

Figure 1 shows the mean seasonal cycle from eight different regions where actual in situ soil moisture observations exist (Figure 2). For each region, optimal averaging was used to produce one value representative of the area. Optimal averaging takes into account the spatial scale of soil moisture, and it gives less weight to repeated information and more weight to independent information. This is particularly important when there is missing information at particular times producing a changing distribution of stations. It also produces an estimate of the error of the estimated average, which is also shown in Figure 1.

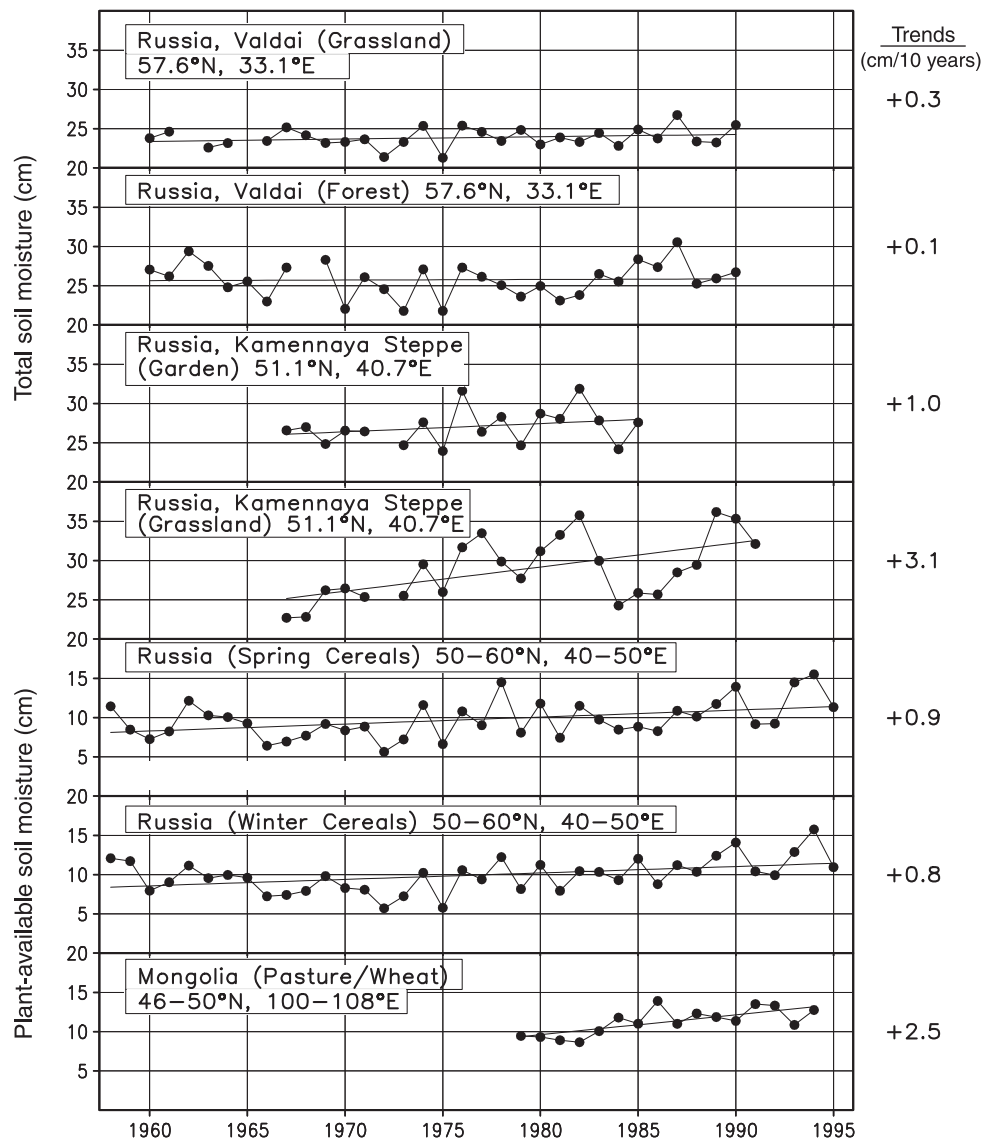


Figure 3 Trends of summer soil moisture from stations or regions with the longest records. Reproduced from Robock, A., Vinnikov, K.Y., Srinivasan, G., et al., 2000. The global soil moisture data bank. *Bulletin of the American Meteorological Society* 81, 1281–1299. Copyright 2000 American Meteorological Society.

Western Russia, Illinois, and Iowa show a typical midlatitude seasonal cycle in a climate with precipitation distributed uniformly throughout the year. Soil moisture is high in the winter, and when the snow melts in the spring, some of the water infiltrates into the soil producing the peak soil moisture for the year. At the end of the summer, when evapotranspiration exceeds precipitation, soil moisture falls to the annual minimum value. In the autumn, evapotranspiration falls and soil moisture increases until the winter. When the ground is covered by snow, there is less change in soil moisture. Vapor exchange and infiltration through cracks and wormholes are still important in frozen soils.

The seasonal cycle in eastern Asia, on the other hand, shows a different character, as there is a large summer precipitation maximum associated with the summer monsoon. In Mongolia, during the half-year for which we have data, soil moisture stays almost constant. The increased evapotranspiration is almost exactly matched by the increased precipitation. In India, where evapotranspiration is much larger, soil moisture falls from the end of the rainy season and reaches a minimum just before the onset of the summer monsoon precipitation.

Trends

Figure 3 shows the trends of summer (June, July, and August) soil moisture from the longest soil moisture time series available 10 years ago. Although all the stations or regions had an upward trend in temperature during the period (not shown), they all showed an upward trend in soil moisture. The longest record of observed soil moisture is from the Ukraine (Figure 4). Although there is no trend in precipitation or temperature, the summer soil moisture shows a strong upward trend (Figure 5). This is because, as mentioned in this article, rising tropospheric

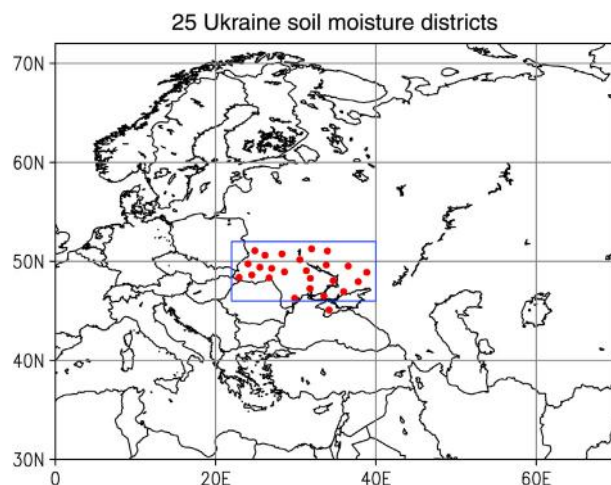


Figure 4 Location of 25 soil moisture districts with 45 years of soil moisture observations, for the period 1958–2002. Each district has on average six soil moisture stations. Also shown is the 22–40° E, 46–52° N box used for averaging the observations and model simulations shown in Figure 5. Reproduced from Robock, A., Li, H., 2006. Solar dimming and CO₂ effects on soil moisture trends. *Geophysical Research Letters* 33, L20708. <http://dx.doi.org/10.1029/2006GL027585>. Copyright 2006 American Meteorological Society.

pollution following World War II led to solar dimming and less evaporative demand at the surface, which dominated any trends in temperature or precipitation.

Virtually, all general circulation model (GCM) simulations of the future climate show summer drying of the midlatitudes. This summer desiccation is caused by evapotranspiration going up faster than precipitation in a greenhouse gas-warmed world. Observations, however, do not show this effect yet. Better models (discussed in Modeling of Soil Moisture) and longer observational time series will be needed to decide whether this prediction is valid.

Modeling of Soil Moisture

For many purposes, including weather forecasting, accurate climate modeling, seasonal prediction, and water resource management, it is crucial to have accurate land surface models. The soil moisture component that is driven by atmospheric forcing may be modeled using routine meteorological observations at regular meteorological stations. Small-scale variability of the soil moisture field is unpredictable and appears as a stochastic process in this context. However, land surface schemes, whether driven by GCMs or by observations, currently do an imperfect job of simulating the mean seasonal cycle and interannual variation of soil moisture, when compared to observations.

Soil moisture (W) variations in the active soil layer can be calculated with the following equation:

$$\frac{dW}{dt} = P + M + WT - E - R \quad [1]$$

where P is liquid precipitation, M is snowmelt, WT is water table contributions, E is evapotranspiration, and R is runoff. Sources of water for the soil are rain, melting snow, and possible subsurface contributions from a rising water table. Soil moisture is lost by evaporation from bare soil, transpiration through plants, and runoff either from the surface or through subsurface drainage. The dependence of evapotranspiration on soil moisture is a function of soil moisture, incoming energy, and rooting depth. When the soil is close to saturation or very dry, when the incoming energy is low (such as in high latitudes), or when plants have a shallow rooting depth, soil moisture has only a small influence on evapotranspiration. Models of soil moisture changes differ in their treatment of evapotranspiration, runoff, snow, frozen soil, and water table. Some models also explicitly contain a parameter that gives a temperature threshold to separate incoming liquid precipitation from solid precipitation.

The first soil moisture schemes in GCMs were based on the Manabe bucket model with 15 cm field capacity, implicit vegetation, and runoff only with bucket overflow. More recently, models have been developed with explicit consideration of vegetation that account for canopy interception of a portion of the precipitation, and explicit stomatal resistance to transpiration. The latest models also incorporate fluxes of CO₂ and other gases as part of the plant and soil biology and physics. In the past, it was convenient to refer to bucket-type models with implicit vegetation as simplistic, and models with explicit consideration of vegetation as sophisticated, complex, and presumably more accurate. But these more

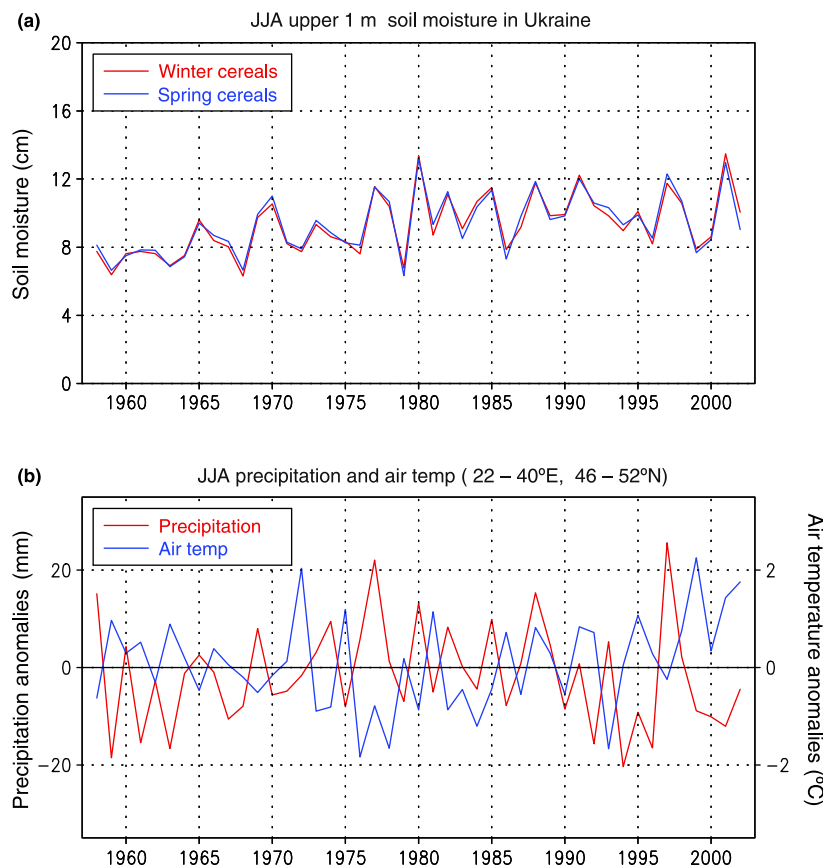


Figure 5 (a) Summer (June, July, and August) plant-available soil moisture in the top 1 m of soil, averaged separately for fields with winter cereals and spring cereals for the region 22–40° E, 46–52° N in the Ukraine (Figure 4). (b) Summer precipitation and temperature anomalies (with respect to the mean for 1971–2000), averaged for the same stations. Reproduced from Robock, A., Li, H., 2006. Solar dimming and CO₂ effects on soil moisture trends. *Geophysical Research Letters* 33, L20708. doi:10.1029/2006GL027585. Copyright 2006 American Meteorological Society.

complex models have so far not shown themselves to be more accurate than the simple ones in climate model applications. Runoff can also be modeled in a more complex manner, including partitioning a fraction of heavy precipitation into immediate runoff and subsoil drainage. Different models have different field capacities and rooting depths, and divide the soil into different numbers of vertical layers for temperature and soil moisture computations.

Weather forecast models now incorporate land surface schemes to predict latent and sensible heat fluxes from the Earth's surface, as well as the radiative balance. In the process, soil moisture and other hydrological quantities, such as runoff, are calculated. Archives of climate model-produced soil moisture, output from GCM runs, are also now becoming available. The soil moisture results from different types of simulations will be different, depending on the way the calculation was done. Some of these come from reanalyses of the global climate system, but the soil moisture calculations in these reanalyses so far do not conserve water, adding or subtracting moisture to make up for atmospheric model deficiencies. Therefore, it is important to not confuse actual observations with model calculations. If a soil moisture 'data set' is to be used to initialize a land surface scheme, then the 'data set' must be calculated with the same land surface scheme, even if the land surface scheme is known to produce a bias in soil moisture.

Otherwise, there will be spin-up problems with the model trying to adjust from the initial values to its own climatology.

The Global Soil Moisture Data Bank has been incorporated into the new International Soil Moisture Network, a data hosting facility for global in situ soil moisture measurements. A continually increasing collection of soil moisture observations may be accessed for free at <http://ismn.geo.tuwien.ac.at/>.

See also: **Climate and Climate Change:** Climate Feedbacks; Climate Prediction: Empirical and Numerical. **Cryosphere:** Snow (Surface). **Global Change:** Biospheric Impacts and Feedbacks. **Hydrology, Floods and Droughts:** Groundwater and Surface Water; Palmer Drought Severity Index. **Satellites and Satellite Remote Sensing:** Precipitation; Water Vapor. **Weather Forecasting:** Seasonal and Interannual Weather Prediction.

Further Reading

- Budyko, M.I., 1956. *Balance of the Earth's Surface*. Gidrometeoizdat, Leningrad, 255 pp. (in Russian).
- Dorigo, W.A., Wagner, W., Hohensinn, R., et al., 2011. The international soil moisture network: A data hosting facility for global in situ soil moisture measurements. *Hydrology and Earth System Sciences* 15, 1675–1698. <http://dx.doi.org/10.5194/hess-15-1675-2011>.

- Delworth, T., Manabe, S., 1993. Climate variability and land surface processes. *Advances in Water Resources* 16, 3–20.
- Larson, K.M., Small, E.E., Gutmann, E., et al., 2008. Using GPS multipath to measure soil moisture fluctuations: initial results. *Earth and Environmental Sciences* 12, 173–177. <http://dx.doi.org/10.1007/s10291-007-0076-6>.
- Manabe, S., 1969. Climate and the ocean circulation, 1. The atmospheric circulation and the hydrology of the earth's surface. *Monthly Weather Reviews* 97, 739–774.
- Manabe, S., Wetherald, R.T., Stouffer, R.J., 1981. Summer dryness due to increase of atmospheric CO₂ concentration. *Climate Change* 3, 347–384.
- Robock, A., Vinnikov, K.Y., Srinivasan, G., et al., 2000. The global soil moisture data bank. *Bulletin of the American Meteorological Society* 81, 1281–1299.
- Robock, A., Li, H., 2006. Solar dimming and CO₂ effects on soil moisture trends. *Geophysical Research Letters* 33, L20708. <http://dx.doi.org/10.1029/2006GL027585>.
- Robock, A., Mu, M., Vinnikov, K., et al., 2005. Forty five years of observed soil moisture in the Ukraine: no summer desiccation (yet). *Geophysical Research Letters* 32, L03401. <http://dx.doi.org/10.1029/2004GL021914>.
- Sellers, P.J., 1992. Biophysical models of land surface processes. In: Trenberth, K.E. (Ed.), *Climate System Modeling*. Cambridge University Press, London, pp. 451–490. (Chapter 14).
- Vinnikov, K.Ya., Yeserkepova, I.B., 1991. Soil moisture: empirical data and model results. *Journal of Climate* 4, 66–79.
- Vinnikov, K.Y., Robock, A., Speranskaya, N.A., Schlosser, C.A., 1996. Scales of temporal and spatial variability of midlatitude soil moisture. *Journal of Geophysical Research* 101, 7163–7174.
- Zreda, M., Desilets, D., Ferré, T.P.A., et al., 2008. Measuring soil moisture content non-invasively at intermediate spatial scale using cosmic-ray neutrons. *Geophysical Research Letters* 35, L21402. <http://dx.doi.org/10.1029/2008GL035655>.

LAND-ATMOSPHERE INTERACTIONS

Contents

Overview

Canopy Processes

Trace Gas Exchange

Overview

RE Dickinson, University of Texas at Austin, Austin, TX, USA

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

Land through its vegetation and soil exchanges with the atmosphere energy, water vapor, carbon dioxide, and other important trace gases. Absorption of solar energy is a major determinant of these exchanges. Both models and data are needed to describe these processes.

Introduction

Land is closely coupled to the atmosphere and the latter to oceans. This system is driven by the absorption of solar radiation, much of it is absorbed by the underlying land and ocean surfaces. Land has much lower capacity to store the thermal energy than the oceans, but rather raises its temperature until it adjusts to lose as much energy as it has received. The consequent fluxes are an important determinant of the day-to-day weather, e.g., clouds and thunderstorms, which are important factors determining climates over land. Properties of the overlying atmosphere are strongly controlled by the influences of underlying surfaces that it has previously been in contact with, so the climate at a particular location acts to average influences from elsewhere, i.e., the climate of near-coastal regions resembles that of adjacent ocean areas, whereas that of continental interiors depends on the absorption of solar radiation by the surfaces it has come in contact with.

The clouds that form when atmospheric water vapor condenses into water droplets substantially reduce the available solar radiation, but their cooling effect is compensated to some extent by their increase of downward thermal radiation. When clouds are able to form drops or crystals that grow large enough to fall, the resulting precipitation supplies water to the soil, and hence plants for their growth. The surplus that is not so used fills streams, rivers, ground water, and reservoirs, and is either captured for human requirements or delivered at the mouths of rivers to the ocean. Incident solar radiation provides the energy required to drive photosynthesis and hence assimilates carbon, and in doing so leads to plant transpiration. The

transpiration of plants provides much of the evaporative cooling of the land surface by the atmosphere in moist regions.

The winds, temperature, and moisture of the overlying atmosphere also strongly affect the land surface, and in turn are affected by the land surface. Describing these influences requires knowledge of the mechanisms by which the land surface balances the energy and water it receives.

How Does the Land Surface Absorb Solar Radiation?

Solar radiation absorbed by any surface is given by the product of how much solar energy is incident on that surface, depending on the direction of the Sun and the orientation of the surface relative to this direction, and the fraction of that incident radiation that is absorbed.

The amount of solar radiation absorbed at the surface determines in large part the climate over land and climate in general. On average for the Earth, it is in nighttime half the time, and during daytime, on average, the angle formed between a vertical line from the surface of the Earth and the Sun is 60° . Consequently, the daytime average sunlight received at the top of the atmosphere is half of that which a point receives when the Sun is directly overhead, and the day–night average is a quarter of that received from an overhead Sun. Solar radiation at the surface is additionally affected by greater removal of radiation by the atmosphere when the Sun is closer to the horizon. These geometric factors reduce the solar radiation received in high latitudes and in winter, and enhance the solar radiation received in the tropics and

in summer, hence largely determining the seasonal and geographic variations of climate.

Quantitative details of climate are also highly dependent on how the atmosphere responds to the seasonal and geographical variations of solar energy delivery. Clouds are the primary mechanism by which the atmosphere alters the reception of solar radiation by the land surface. In addition, solar radiation is attenuated by other smaller atmospheric particles, referred to as aerosols, and atmospheric gaseous absorbers, such as water vapor and ozone. The total net radiation that heats the land surface includes that of downward thermal emission from the atmosphere, also depending on clouds and water vapor, and is reduced by the upward emission of thermal radiation. This 'Earth' radiation depends on surface temperature but less so than the fluxes of energy from surface evaporation and dry sensible heat. Hence, in discussing the atmospheric radiative fluxes to the surface in the next section, we use the term 'total net radiation.'

The fraction of solar radiation that is reflected by a surface is referred to as its albedo. Because land consists of surfaces oriented in all directions, its albedo is not simply determined by the reflection from a single flat surface but also by how much of reflected light is absorbed by other surfaces. Leaves are so arranged that a plant canopy reflects less than half as much solar radiation as do individual leaves. The albedo of land surfaces, and especially those vegetated depends on the wavelength of the solar radiation. The solar radiation that plants use is essentially at the same wavelengths as those of human vision that is 'visible light.' Approximately half of solar radiation occurs at longer wavelengths than visible. These longer wavelengths have much higher albedos since they are not used for photosynthesis and otherwise may lead to overheating.

What Determines Turbulent Fluxes of Water and Heat from the Land Surface?

Besides its absorption of radiation, the land surface and its interaction with the atmosphere depends on other factors. Over the course of a day, some of the daytime heating can conduct downward into the soil and then be released again at night. Averaged over day and night, the net absorption of radiation is largely balanced by fluxes of energy that the land surface delivers to the atmosphere. The turbulent motions of air near the land surface determine these fluxes. Convection and mechanical mixing in turn determine the intensity of the turbulence. The latter is determined by the strength of surface winds and by the roughness of the surface. Positive net land surface heating normally occurs only during the day and leads to convection. It usually has small negative values at night.

The energy carried from the surface by water vapor corresponds to the energy that was required to evaporate the water from its liquid state at the land surface or equivalently to the energy that will be released when this water vapor is converted back to liquid form through the formation of clouds and precipitation. Dry atmospheric energy is transported upward by relatively warm air rising and cold air sinking, providing the intense daytime convection that generates the boundary layer turbulence needed to remove the energy supplied to the surface by daytime radiation. This flux of dry energy is proportional to

the difference between land surface temperatures (i.e., that of leaves and soil surface) and that of the overlying air. The flux of water vapor is likewise proportional to the difference between water vapor concentrations at the surface and that of the overlying air. Where materials at the surface are supplied with water, e.g., inside leaves and in moist soil, the consequent water vapor concentration depends only on temperature. How the land surface responds to a given amount of net radiation depends on how it divides the removal of this energy between evaporation and dry atmospheric energy flux. The ratio of dry atmospheric energy flux to energy carried by water vapor is called the 'Bowen ratio.'

The Bowen ratio depends on surface temperatures, the relative humidity of the overlying air, and various resistances to removal of water from the soil. In the extreme, if the surface is very dry so that there is very little or no water available to be moved into the atmosphere, then the Bowen ratio becomes very large. Conversely, the smallest Bowen ratios occur in warm areas when there are no limitations on the removal of water from the surface but the atmospheric relative humidity is low. This situation can only happen locally with conditions different from those of the surrounding land surface, since the consequent flux of energy will exceed that provided by net radiation and so depends on air energy carried to that point from elsewhere. Such conditions are referred to as the 'oasis' effect.

Temperature alone determines the Bowen ratio when surface air is at 100% relative humidity. Evapotranspiration can still occur provided the surface is heated to raise the concentration of water in the near-surface soil and inside plant leaves to values of atmospheric humidity larger than that of the overlying air. At larger temperatures, these concentrations increase substantially and the Bowen ratio is less.

In sum, the Bowen ratio of moist surfaces will be greater for colder temperatures and higher relative humidity. It is further increased by various surface resistances that affect only the movement of water vapor or affect it more than the movement of dry heat, e.g., when the surface soil is dry, water vapor must diffuse upward from deeper soil layers, and the rate at which this diffusion occurs may limit transport of water into the air.

Precipitation and hence soil moisture are normally accompanied by growth of plants. Hence, the fluxes of water vapor from the land surface to the atmosphere largely occur through extraction of soil water by roots and transport through plants and their leaves. This transport is called transpiration. Normally, with adequate soil moisture, plant leaves are the main obstacle to this movement of water and hence act as an important control on the Bowen ratio. A leaf normally has to contain water to maintain its structure, and if the roots cannot supply water as fast as it loses it, it wilts. Hence, its water loss can be a threat to its survival. The leaves lose water primarily through tiny holes called 'stomates.' Why do leaves have this stomatal loss mechanism for water and what determines the extent by which the stomates lose water? Leaves may benefit from the resulting evaporative cooling in warm regions where they may otherwise be threatened by thermal damage from high temperatures. However, the primary reason is an even more basic element of plant's requirements for growth and survival. The photosynthesis of plants uses light from the Sun to convert carbon dioxide to the carbohydrates and proteins needed for plant growth (some of which eventually become

our own food). This can only work if the plants receive, besides light, an adequate supply of carbon dioxide. The stomates pass carbon dioxide from atmosphere to the inside of leaves where it can reach the chloroplasts and be used. This function of the stomates, however, is in a somewhat delicate balance, controlled by the stomatal opening, with how much water is lost. If the net stomatal passageway to the atmosphere is too large, the leaves may lose more water than the soil can provide, whereas if it is too small, the leaf may be starved of carbon dioxide, at least relative to the light energy available to convert it to carbohydrate.

Plants and related aspects of the land surface exchange not only carbon dioxide but also other important gases with the atmosphere, as addressed in the next section.

What Determines the Fluxes of Carbon Dioxide and Other Gases to and from the Land Surface?

As concluded in the previous section, the flux of carbon dioxide into plants is closely linked to the transpiration by vegetation and hence to much of the flux of water from the land into the atmosphere. However, if this removal of carbon dioxide were simply one-way or unbalanced in its direction of removal, the atmosphere would become exhausted and no longer provide an adequate supply of this essential plant nutrient. This may not be such a far-fetched possibility, because the amount of carbon from carbon dioxide currently contained in land plants is as large as that held in the atmosphere, including the atmosphere over the oceans; and that stored in the soil, originating from plants, is considerably larger. Large amounts of carbon dioxide have been converted to fossil fuels at earlier times in the Earth's history. Indeed, currently accessible coal deposits have been estimated to hold over 10 times the carbon currently contained in the atmosphere, and these deposits may have formed when the atmosphere held several times as much carbon dioxide as it does now.

Even more carbon is held in the oceans than by land, mostly dissolved as bicarbonates and carbonates. This carbon in seawater eventually makes its way into limestone sediments, which through movements of the Earth's crust are returned to land and the atmosphere. Fortunately, the natural exchanges of carbon dioxide between land, atmosphere, and ocean are normally very close to being in balance with the land, in particular, acting to move as much carbon dioxide back into the atmosphere as it removes. Currently, land appears to take up more carbon dioxide than it gives back to the atmosphere in response to the extra carbon dioxide being supplied by the human conversions of fossil fuels back into carbon dioxide and consequent increases of atmospheric carbon dioxide. Hence the net removal of this excess carbon dioxide by land must be regarded as a very valuable ecosystem service, reducing the consequent climate change. Although we know much about how the land exchanges carbon with the atmosphere, our current understanding of the details of the net removal into land is sufficiently poor that we cannot accurately determine under what conditions this removal could cease or possibly reverse.

How does carbon dioxide removed from the atmosphere by plants return back to the atmosphere? The carbohydrates that the plants produce from photosynthesis are largely eaten. At

the cellular level the mitochondria present in all eukaryotic cells metabolize (that is oxidize) these carbohydrates, generating energy and carbon dioxide. This energy supplies the needs of the cell and so that of more complete organisms, and the carbon dioxide is returned to the atmosphere. In our bodies, this carbon dioxide is then exhaled from our lungs. Only about 1% of the carbohydrate energy from plants is metabolized by humans. What happens to the rest?

The plants themselves use about half of their carbohydrate energy stores both to convert the carbohydrates to more complex molecules such as proteins and fat compounds and to continuously repair complex molecules such as enzymes that tend to 'wear out.' Additional energy is needed by roots to facilitate their acquisition of soil nutrients either by directly powering the movement of soil ions or indirectly by feeding other soil organisms such as mycorrhizae that facilitate the movement of insoluble soil nutrients to the roots.

The carbon that the plants do not use themselves, and that is not harvested by humans or other animals, is delivered to the land surface and soils as dead plant materials. This dead plant material feeds many small to microscopic organisms, at the bottom of the food chain mostly bacteria and fungi, the ultimate decomposers of plant materials that are responsible for return of carbon dioxide to the atmosphere. All these biological processes are strongly controlled by the physical environment, the overlying atmosphere, the water provided by precipitation, and the supply of energy and warmth by the Sun.

Some of the carbon compounds contained by plants are oxidized directly by high-temperature combustion: that is, by fires. Human use of plant carbon for energy, human land management practices, and natural or accidentally started fires release comparable amounts of carbon dioxide, in total perhaps as much as 20% of that captured by plants on average. Natural fires are themselves a land process that strongly interacts with the atmosphere.

Besides carbon dioxide, numerous other important carbon compounds are exchanged between the land surface and the atmosphere. Especially important for climate is methane, which is emitted in places where there is very little oxygen, such as swamps and rice paddies. In addition, complex organic compounds are given off in copious amounts by leaves and contribute to enhancing regional levels of ozone and photochemical smog. Forest fires are an especially important mechanism for the supply of other complex organic materials to the atmosphere, including important aerosols.

Other elements important for living cells that are exchanged between land and the atmosphere include nitrogen and sulfur. The molecular nitrogen in the atmosphere is very inert to chemical change, but enough nitrogen must be supplied to the land surface in the form of nitrate or ammonium compounds to maintain this element in living cells. Ammonium compounds are generated naturally by nitrogen-fixing organisms that live freely in the soil or are attached to the root systems of some plants, such as the legume family. Much of the nitrogen supplied by soil to plants comes from the fast recycling of the nitrogen residing in plants, e.g., the leaves and roots that die and return carbon to the soil also return, nitrogen that is released by microbial processes along with the carbon.

Humans now add large amounts of ammonium and nitrate to the soils directly as fertilizers or indirectly as wet and dry

deposition of these compounds from atmospheric pollution. Although some sulfur moves from the land into the atmosphere, much more is initially put into the atmosphere by the combustion of fossil fuels and this is then deposited back to the land.

Land processes supply ammonia and oxides of nitrogen to the atmosphere, which are major elements in determining aerosols and atmospheric chemistry. Most of the ammonia comes from areas where it has been concentrated by human management practices, such as from feedlots and heavily fertilized fields. Bacteria convert soil ammonium to nitrate ions and other bacteria convert some of this to nitrous oxide that escapes to the atmosphere.

What Data Are Needed about the Geographical Variations of Land Properties to Determine Their Interaction with the Atmosphere?

Models of the Earth system that describe the interactions of the land with the atmosphere require not only the best efforts to solve equations for the processes reviewed above but also knowledge of the geographically varying properties of the land surface that determine these interactions. Many of these properties are generated by the models themselves, e.g., those involving the absorption of solar radiation, temperatures, and aspects of the hydrological cycle such as precipitation and soil moisture. However, some necessary features are more accurately observed than modeled, or are best treated by a combination of modeling and observations.

Some particularly important parameters that need to be constrained by observations are the land albedos and aspects of vegetation that control these properties of the soil and other aspects of the land surface that determine the movement of water and how much ends up in streams and rivers or escapes to the atmosphere. The latter include the roughness of the vegetation for producing turbulence in the air that flows over it,

the degree of resistance the leaves produce to the water movement through them, and the depth into the soil to which the roots are able to mine water.

These required properties can be estimated by mapping of the world's land surfaces into plant functional types or major biomes, overlapping approaches where biomes consist largely of single plant functional types. For example, the tropical evergreen biome consists largely of the tropical evergreen tree plant functional type. On the other hand, the savanna class biome consists of a mixture of trees and grasses. For either of these approaches or more detailed land cover classifications and for determination of albedos, global mapping requires the use of satellite imagery of the land surface. Such measurements are not as simple as determining the presence or absence of clouds but have become possible with satellite instruments developed over the last decade or two.

See also: **Boundary Layer (Atmospheric) and Air Pollution:** Overview; Stably Stratified Boundary Layer; Surface Layer.
Global Change: Biospheric Impacts and Feedbacks.
Land-Atmosphere Interactions: Canopy Processes; Trace Gas Exchange.

Further Reading

- Dickinson, R.E., 1983. Land surface processes and climate – Surface albedos and energy balance. In: Saltzman, B. (Ed.), *Theory of Climate, Advances in Geophysics*, vol. 25. Academic Press, New York, pp. 305–353.
- Dickinson, R.E., 2011. In: Henderson Sellers, A., McGuffie, K. (Eds.), *Interaction between Future Climate and Terrestrial Carbon and Nitrogen, Future of the World's Climate*. Elsevier Inc, Oxford, UK.
- Wang, K.C., Dickinson, R.E., 2012. A review of global terrestrial evapotranspiration: observation, modeling, climatology, and climatic variability. *Reviews in Geophysics* 50. <http://dx.doi.org/10.1029/2011RG000373>. RG2005, 54 pp.

Canopy Processes

PD Blanken, University of Colorado at Boulder, Boulder, CO, USA

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

This article summarizes the processes that connect vegetation canopies to the atmosphere. After a review of canopy structure and leaf arrangement, photosynthesis is discussed, including a review of the various biochemical pathways that vascular plants use to sequester carbon from and release water vapor to the atmosphere. Atmospheric controls on the connection of vascular plants to the atmosphere through stomata at both the leaf and canopy levels are discussed. Other topics discussed include light penetration through canopies, canopy productivity, wind and turbulence, and the remote sensing of canopy properties.

Introduction

The arrangement of leaves and supporting structures (branches and stems), integrated with height between the ground and the top of vegetation, forms our definition of a vegetation canopy. There is almost infinite variety in the morphology of canopies in the world's vegetation. Canopies may have several, well-defined layers, such as in rainforests, or may have one well-defined overstory, with or without a vegetated ground cover (the understory; [Figure 1](#)). A typical way to quantify the canopy is in terms of canopy closure (or gaps, where sky is visible) or the leaf area index (L ; half of the total leaf area per unit ground area, where the other half is the underside). The latter takes into account dense, multilayered canopies where L can exceed one. In general, L varies zonally across the globe, varying roughly with precipitation and temperature, and reaching a maximum near the equator and a minimum near the poles.

Superimposed on the world's spatial variability in canopy structure, there is often a great deal of temporal variability for

a given canopy. In the long term, the canopy changes through the ecological processes of succession, as a disturbance such as fire destroys the canopy and L generally increases over perhaps hundreds of years as different species colonize the area. Annually, L can vary from a maximum during the summer to a minimum during the winter as deciduous canopies seasonally lose their leaves, effectively shedding most of their canopy. On a short-term basis (hours), some species can change leaf orientation in response to water stress (e.g., soybeans), track the sun (e.g., alfalfa, cotton, and soybeans), or shed leaves in response to water stress (e.g., cotton). Even within a single species at any given time, leaf structure can vary from being thick near the top of a light-rich canopy to being thin within a shaded canopy. Oak leaves, for example, grown in the upper canopy tend to be smaller, more deeply lobed, and inclined at steeper angles than those grown beneath in the shade.

Regardless of this spatial and temporal variability, it can be argued that the motivation driving all atmospheric-related canopy processes is to achieve an optimum L that maximizes



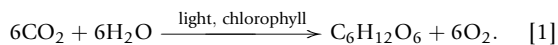
Figure 1 Contrasting understory growth and canopy structure in a mixed hardwood deciduous forest in Michigan (left), and a subalpine coniferous forest in Colorado (right).

light absorption for photosynthesis while minimizing transpiration water loss and balancing nutrient demands with supplies. The quantity of light absorbed as it passes through a canopy drives the process of photosynthesis, which is connected to the transpiration water loss through individual leaf stomata (sing. stoma). A measure of the openness of leaf stomata at the canopy scale, canopy conductance, responds to the air temperature, humidity, wind speed (turbulence), and CO₂ concentrations, all of which to various degrees are influenced by the canopy. As a result, a canopy often creates its own microclimate.

A description of canopy processes requires integration or aggregation of processes across spatial scales ranging from the cellular to canopy scales. Such integration remains an issue, as an accurate representation of canopy processes requires the ability to 'scale up' from leaf- to canopy-level measurements, and 'scale down' back to the leaf from the canopy. In this article, canopy processes will be described for vascular plants following a scaling-up approach, covering the following topics: photosynthesis, canopy conductance, light penetration, canopy productivity, wind and turbulence, and remote sensing. The responses of plant canopy processes to changing environmental conditions such as atmospheric CO₂ concentrations, solar radiation, wind, and temperature will be discussed.

Photosynthesis

Leaf surfaces are remarkably well adapted for intercepting light. The conversion of light into carbohydrates and starches occurs through the process of photosynthesis:



Respiration is the reverse of eqn [1], with energy produced instead of required. Joseph Priestley in 1771–72 discovered that a candle would quickly extinguish when placed with a plant in a sealed chamber, yet could be relit 27 days later. Similarly, a mouse could survive in a sealed chamber only when kept with a plant. Thus, Priestley concluded that plants were somehow 'restoring' the air (oxygen had just been discovered). Subsequent experiments by others showed that only shortwave (solar) radiation between wavelengths of 400 and 700 nm (photosynthetically active radiation (PAR)), and especially between 680 and 700 nm, is absorbed by the plant for photosynthesis and required for the production of sugars and starches.

The way of CO₂ entrance into the leaf, and subsequent loss of water vapor, is through the stoma (Figure 2). Liquid water inside the leaf is required for the carbon to dissolve into the plant, and nitrogen is required to fix or sequester the carbon into a usable form. Stomata are small openings (typical length 15 μm) that are typically on the underside of leaves and are flanked by two guard cells. These guard cells regulate the aperture of the stomata by changing their turgor (internal water pressure). The plant actively regulates the stomata aperture (often expressed as stomatal conductance, or the reciprocal, stomatal resistance) by changing the guard cell turgor through changing the potassium ion concentration inside the guard cells. If, for example, the tension of water inside the water-conducting vessel in the plant (the xylem) became too great,

to prevent cavitation inside the xylem (irreparably damaging it), potassium would flow out of the guard cells. In response to the developing water potential gradient, water would flow out of the guard cells and into the adjacent epidermal cells. The turgor in the guard cells would decrease, and the stomata would close, thus reducing water loss (also carbon gain). The signal to transfer potassium in or out of the guard cells is thought to be controlled by the hormone abscisic acid (ABA), which is produced at the root hairs and carried through the xylem to the guard cells in the leaf.

Stomata provide a direct connection for gas exchange between the canopy and atmosphere, and the plant can regulate CO₂ uptake and transpiration water loss on short time scales (seconds) through changes in stomatal conductance, but also on long time scales (years) through changes in the stomatal density (number of stomata per area of leaf). The environmental clues that trigger changes in stomatal density are linked to the variables given in eqn [1], namely light quantity and atmospheric CO₂ concentrations. For example, plants grown under increased PAR levels tend to have a higher stomatal density compared to plants grown under normal PAR levels, thus taking advantage of the increased energy available to drive photosynthesis. Conversely, plants grown under elevated atmospheric CO₂ concentrations tend to have a decrease in stomatal density, and this response is also influenced by relative humidity (and therefore temperature), as one may have guessed by inspection of eqn [1]. So both the aperture of existing stomata (short term) and the development of stomata (long term) show dynamic responses to environmental clues expressed in the basic variables of the photosynthesis equation.

Biochemical Pathways

As eqn [1] shows, water is required for photosynthesis as it provides the solution in which carbon can dissolve. In addition to modifying canopy and leaf morphology to optimize light absorption, plants have evolved a number of strategies to maximize photosynthesis when water is the limiting resource. Biochemically, this includes supplementing the C₃ photosynthetic carbon reduction (PCR) cycle with the C₄ photosynthetic carbon assimilation (PCA) or crassulacean acid metabolism (CAM) pathways.

All photosynthetic eukaryotes reduce CO₂ to carbohydrates using the PCR (Calvin) cycle. In this cycle, atmospheric CO₂ and water are combined with a five-carbon acceptor molecule (ribulose 1,5-bisphosphate) to generate two molecules of a three-carbon intermediate, phosphoglycerate. This carboxylation reaction is catalyzed by the enzyme ribulose bisphosphate carboxylase oxygenase (RuBisCO), the most abundant enzyme in the world (estimated 10⁷ tons). Using the photochemically derived energy adenosine triphosphate (ATP), phosphoglycerate is then reduced to form carbohydrates (sucrose and starch). Next, the CO₂ acceptor molecule is regenerated, again using ATP. At a thermodynamic efficiency of about 90%, six turns of this cycle are required to produce the equivalent of one glucose molecule from six carbon atoms.

The enzyme RuBisCO discriminates poorly between CO₂ and O₂, and it will oxygenize ribulose 1,5-bisphosphate and release CO₂ if combined with O₂. This process, known as photorespiration (the C₂ photorespiratory carbon oxidation

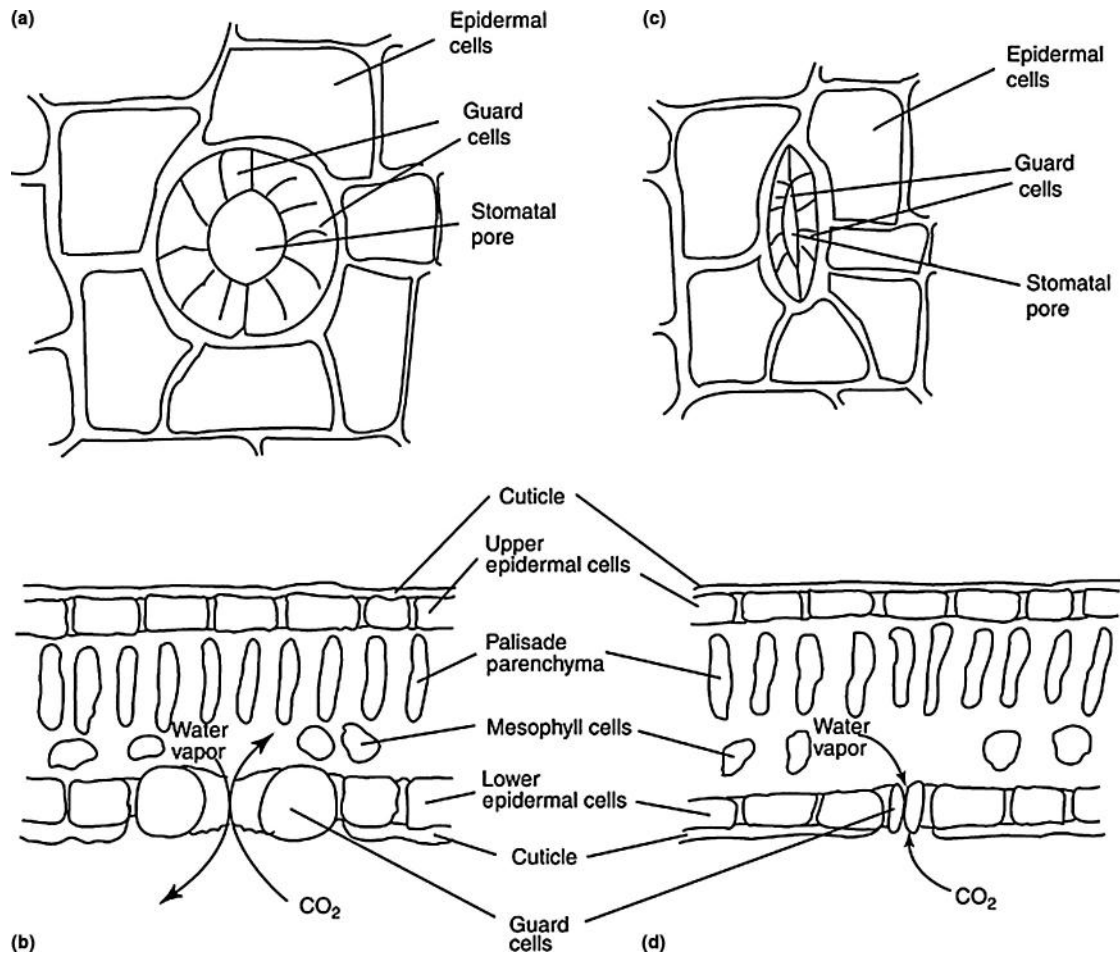


Figure 2 Plan (a and c) and cross-sectional (b and d) views of an open (a and b) and closed stoma (c and d). The transfer of water vapor out of the leaf, and of CO₂ into the leaf, is regulated by the aperture of the stomatal pore; this is quantified as the leaf stomatal conductance or its reciprocal, the leaf stomatal resistance.

(PCO) cycle), is diametrically opposed to photosynthesis and results in a loss of up to 50% of the CO₂ gained by photosynthesis. To recover some of the photorespired CO₂, species found in somewhat dry and high-energy (light) tropical and subtropical climates (e.g., tropical grasses, sugarcane, corn, and sorghum) have a particularly well-developed ability to form the C₄ acids malate and/or aspartate. These acids are broken down near the site of carboxylation (the mesophyll) to regenerate an additional CO₂ acceptor molecule. Hence, C₄ species typically have high photosynthesis rates and a low stomatal conductance (high water use efficiency, as discussed further in this article), yet they require high temperatures and ample light, and are therefore seldom found in cool or shady locations such as those often found beneath a canopy.

In especially arid environments, CAM species such as cacti have the ability to reduce water loss by closing their stomata during the day and opening them at night. This high water use efficiency is achieved by obtaining CO₂ at night and fixing it in the form of the acid malate, which is stored in vacuoles. During the day, carboxylation of malate releases CO₂, which cannot escape back to the atmosphere since the stomata are closed, and is reduced to carbohydrates via the C₃ PCR cycle.

Each of these biochemical pathways for vascular plant photosynthesis is affected by atmospheric CO₂ concentrations. For example, the C₃ pathway is favored under higher atmospheric CO₂ concentrations, and the C₄ pathway is favored under lower atmospheric CO₂ concentrations. This is because there is no real advantage for a plant to spend energy to increase CO₂ concentrations in the bundle sheath cells if ambient concentrations are already high. Recently, experiments such as FACE (Free-Air CO₂ Enrichment) have been conducted to explore how contemporary increases in atmospheric CO₂ concentrations will affect these photosynthesis biochemical pathways, and therefore the distribution and abundance of terrestrial vegetation with elevated CO₂. Key findings from this 15-year study across several different canopy types included the following: trees (C₃ species) responded more than other vegetation groups (C₄ species) with an increase in carbon uptake (photosynthetic stimulation), increased *L*, an increase in dry matter production, and a decrease in stomatal conductance under elevated atmospheric CO₂ concentrations (200 ppm above current concentrations). Experiments such as FACE and other similar studies on plants and ecosystems under natural conditions show that C₃ species, especially trees, are

much more responsive to elevated CO_2 concentrations than C_4 species. Thus, we could expect long-term ecosystem shifts in composition and distribution due to changes in canopy processes caused by changing atmospheric CO_2 concentrations.

Water Use Efficiency

The success of these various processes of maximizing carbon uptake while minimizing water loss is expressed by the water use efficiency (WUE):

$$\text{WUE} = \frac{\text{moles of } \text{CO}_2 \text{ fixed}}{\text{moles of } \text{H}_2\text{O} \text{ transpired}} \quad [2]$$

This ratio captures what is often referred to as the photosynthesis–transpiration dilemma; how to maximize carbon intake while minimizing water loss. In regions where water is nonlimiting, canopies usually maximize L to maximize light interception without regard for water loss. Many freshwater wetland species, and saltwater species such as mangroves, however, do actively regulate water loss (transpiration) thorough decreasing stomatal conductance to prevalent xylem capitation in times of high atmospheric water demand (freshwater wetland species) or to prevent salt from entering the plant (mangroves; Figure 3). Ironically, these species growing in standing water show xylem water pressure potentials as low as those of desert species. In regions where water is a limiting factor (sometimes seasonally or even diurnally), various strategies in addition to the various biochemical pathways just discussed may be used to minimize transpirational water loss.

Leaf and canopy morphology are often modified to minimize water loss. New leaves and stems grown under water deficits tend to be smaller than those grown under nonstressed

conditions. Existing leaves often senesce and fall off during periods of water stress. These leaf area limitations and adjustments are usually the first responses of plants to slow, long-term dehydration. The structure of the canopy and of individual leaves (e.g., edge shapes and leaf hairs) can also affect water loss. For example, a canopy that is aerodynamically rough (e.g., forests) tends to enhance water and heat loss by having a low aerodynamic resistance (see the ‘Wind and Turbulence’ section), whereas aerodynamically smooth canopies (e.g., crops) tend to have a large aerodynamic resistance that suppresses water loss and heat exchange. Prolonged drought can also lead to an expansion of the root system into deeper soils, where water may be more plentiful than at the surface.

In response to short-term water stress, or after the plant has reached its maximum leaf area, stomatal closure can effectively reduce transpiration water loss. Stomatal closure occurs either passively when rapid water loss from the guard cells cannot be replenished by water from adjacent epidermal cells, or metabolically when solute transport from the guard cells results in water loss, decreased turgor, and hence closure. It is thought that ABA delivered to the leaf from the roots, in addition to playing a role in leaf abscission, also plays a key role in initiating stomatal closure, as it is at the roots where drying is first detected.

As discussed in this article, the quantity of CO_2 sequestered through photosynthesis depends on several factors and variables, including the ambient CO_2 concentration, quantity and quality of light (PAR), temperature and relative humidity, stomatal conductance and density (and therefore L), and biochemical pathway through which photosynthesis occurs. With elevated ambient CO_2 concentrations (and all other variables held similar), WUE generally increases (improves) for C_3 species. This is primarily due to the decrease in stomatal



Figure 3 A mangrove island remarkably well adapted to saline water in the Gulf Coast of Florida.

conductance and associated decrease in transpiration, allowing for increased soil water content and increased growth. What needs to be remembered is that other climatological changes associated with changes in atmospheric CO₂ concentrations (e.g., cloud cover hence PAR, precipitation, temperature, humidity, soils, and nutrient concentrations such as nitrogen) will almost certainly also change, hence complicating this general finding.

Canopy Conductance

The transpirational water loss from a leaf is controlled by not only processes internal to the leaf but also external processes, and considerable feedbacks occur between the two. The aggregation of leaves into a canopy alters not only the light regime but also the temperature, wind, and concentration of gases such as water vapor and CO₂. There are also concerns with scaling between the leaf and canopy, that is, does a canopy behave like a 'big leaf' so that processes observed at the leaf level can be simply multiplied to the canopy level as a function of leaf area?

An Ohm's law electrical analogy ($I = V/r$ where V is voltage, I is current, and r is resistance) has been successfully applied to describe the transfer of heat, water vapor, and CO₂ between the leaf and atmosphere, and between the canopy and atmosphere (Figure 4), and hence, it forms a basis for predicting either the leaf stomatal (g_L) or canopy (g_C) conductance. Note that conductance is the reciprocal of resistance (i.e., $g = 1/r$). Simply stated, the flux or exchange per unit area over a given time (i.e., I) is equal to the difference in concentration (potential difference) between the leaf or canopy and the atmosphere (i.e., V), divided by the resistance to this transfer (i.e., r):

$$\text{flux} = \frac{\text{potential difference}}{\text{resistance}}. \quad [3]$$

A popular approach for calculating g_C from generally available canopy-level meteorological measurements is by solving the Penman–Monteith combination equation for g_C :

$$r_C = \frac{1}{g_C} = \frac{r_a[S(R_n - G) - \lambda E(S + \gamma)] + \rho c_p D}{\gamma \lambda E}, \quad [4]$$

where r_C and r_a are the canopy and aerodynamic resistances, respectively; S is the slope of the saturation vapor pressure

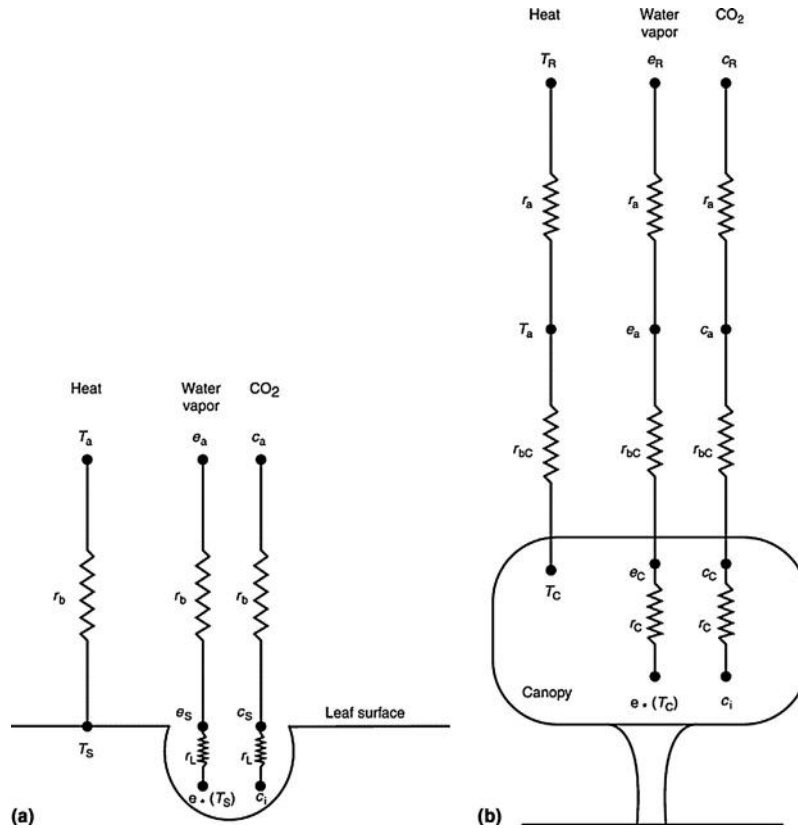


Figure 4 Schematic of transfer of heat, water vapor, and CO₂ between a leaf (a) or canopy (b) and the atmosphere. At the leaf level, heat transfer depends on the difference between the leaf surface (T_S) and air temperature (T_a) divided by the leaf's boundary layer resistance (r_b), created by the transfer across the layer of still air adjacent to the leaf surface. The transfer of water vapor from the leaf is described by the difference in the saturation vapor pressure calculated at T_S ($e^*(T_S)$) and the vapor pressure at the leaf surface (e_S) divided by the leaf stomatal resistance (r_L), and the difference between e_S and the atmospheric vapor pressure (e_a) divided by r_b . Similarly, CO₂ transfer into the leaf depends on the difference in CO₂ concentration in the atmosphere (c_a) and the leaf surface (c_S) divided by r_b , and the difference between c_S and the leaf's internal CO₂ concentration (c_i) divided by r_L . At the canopy level, leaf-level values must be replaced with canopy-level values (subscript C), and r_b must be replaced with a canopy boundary layer resistance, r_{bC} . In addition, the aerodynamic resistance (r_a) between the atmospheric values and those measured at some reference height above the canopy (subscript R) must be included.

versus temperature curve; R_n is the net radiation; G and λE are the soil and latent heat fluxes, respectively; γ is the psychrometric constant; ρ is the air density; c_p is the specific heat of dry air; and D is the saturation deficit. This energy balance approach works well in an analytical sense when λE has already been measured, for example by eddy covariance (Figure 5) and when soil water evaporation is negligible. Soil water evaporation is difficult to measure, and it varies with both time and space. New measurement techniques such as stable oxygen isotopes, however, offer a promising means to partition λE into transpiration and soil water evaporation. When measurements of λE are not available, the use of eqn [4] in a predictive fashion requires that g_C be modeled, which is not an easy task given the complexity of a plant's response to the environment and vice versa.

At both the leaf and canopy levels, conductance has been found to respond to several variables, for example PAR, temperature (leaf or air), humidity (relative or the saturation deficit), and the CO_2 concentration, or $[\text{CO}_2]$ (leaf or air). The response of individual species to these or other variables varies greatly, but examples of general responses in the absence of any other environmental stresses are shown in Figure 6.



Figure 5 Eddy covariance instruments used to directly measure the fluxes of heat, water vapor, CO_2 , and momentum. A sonic anemometer measures the vertical, horizontal, and lateral wind velocity components by measuring the speed of sound between pairs of transducers. An open-path gas analyzer measures the densities of water vapor and CO_2 in a volume of air by measuring the attenuation of infrared radiation between a source and detector. The covariance between fluctuations in simultaneous high-frequency measurements of the vertical wind speed, air temperature, water vapor, CO_2 , and horizontal wind speed are used to calculate each of the fluxes.

The approach taken by researchers to develop species-specific relationships has been either to make measurements on individual leaves in a laboratory under controlled conditions, or to make measurements on individual leaves in the natural environment under either controlled conditions or a wide variety of conditions. Individual leaf measurements are obtained by placing a leaf into a cuvette (Figure 7), and then using either multiple regression (an additive model; see eqn [5], where a through d are regression coefficients, and x_1 through x_3 are the various independent variables influencing g_L) or a boundary-line analysis to determine the stomatal response to the measured environmental variables. The latter analysis involves measuring g_L under as many ambient conditions as possible (and thus requires large amounts of data), then fitting and defining curves to the upper data points (e.g., $f(x_1)$), where points below these curves represent times when g_L was limited by some other variable. The data are then standardized by the maximum observed g_L ($g_{L\max}$), to form a multiplicative model (eqn [6]). Once the relationships between g_L and the environment are quantified, scaling from leaf to canopy is then accomplished with knowledge of the leaf area index (eqn [7]). If a single-species canopy has several layers, then g_L and L should be measured for each individual layer, multiplied, and then summed. If there are several species in the canopy, then g_L and L for each species should be measured, multiplied, and then summed to scale up to the canopy.

$$g_L = a + bx_1 + cx_2 + dx_3 \dots \quad [5]$$

$$g_L = g_{L\max} [f(x_1)f(x_2)f(x_3) \dots] \quad [6]$$

$$g_C = g_L L \quad [7]$$

It has also been recognized, however, that in addition to g_L responding to relative humidity at the leaf surface (h_s), g_L in some species also responds to the $[\text{CO}_2]$ by varying g_L to maintain a constant $[\text{CO}_2]$ at the leaf surface (c_s). These two driving variables, h_s and c_s , are combined in an empirical model widely recognized as the Collatz model, which is based on the Ball-Berry-Woodrow index:

$$g_L = m \frac{A}{c_s} h_s + b, \quad [8]$$

or, at the canopy scale:

$$g_C = m \frac{A_C}{c_C} h_C + bL, \quad [9]$$

where m and b are empirically derived coefficients based on cuvette gas exchange studies, A is the net carbon assimilation rate, and the subscript C refers to canopy-level values. In eqns [8] and [9], conductance is now a function of the net assimilation rate and vice versa, hence eqns [8] or [9] must be solved iteratively using a series of equations that describes both conductance and photosynthesis.

Equation [9] still remains the common means of calculating canopy conductance in climate models, such as the Community Land Model (CLM4). While the basic form of eqn [9] remains unchanged, improvements in model accuracy have been achieved when the physiologically based parameters are changed based on improved measurements under a greater range of ambient conditions for more canopy types.

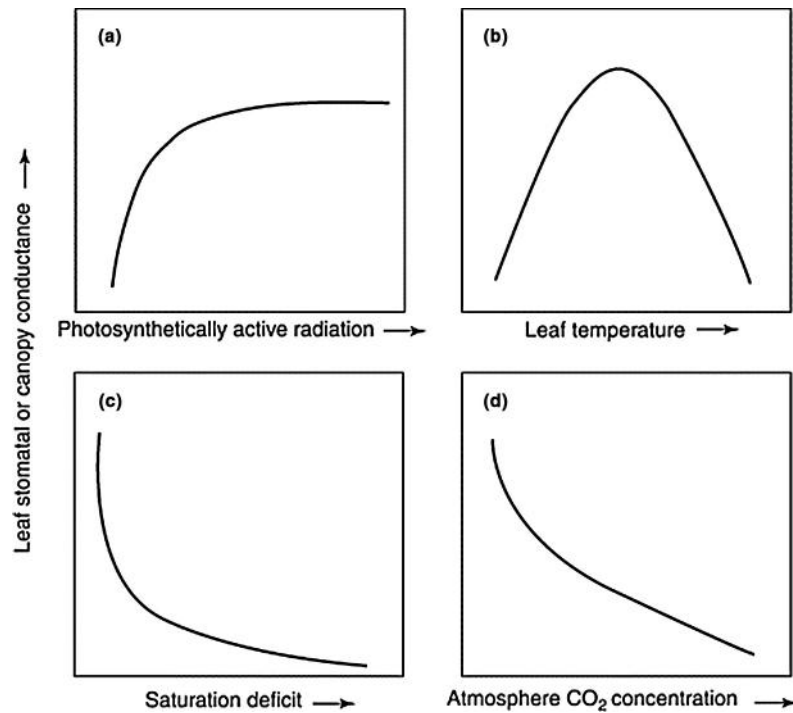


Figure 6 Examples of the response of leaf stomatal or canopy conductance to photosynthetically active radiation (a), leaf temperature (b), saturation deficit (c), and atmospheric CO₂ concentrations (d) when all other environmental stresses are absent.



Figure 7 Using a steady-state porometer to measure the leaf stomatal conductance of *Carex aquatilis* in the Hudson Bay Lowland near Churchill, Manitoba, Canada.

Light Penetration

A canopy develops because of competition for sunlight, which is required for photosynthesis and hence growth. Trees, through their intricate structure of leaves supported by stems and branches, are remarkably well adapted for light interception, and hence, they outgrow nonstemmed species. Whereas an individual leaf typically absorbs roughly 50% of the incident shortwave radiation, a canopy typically absorbs roughly 80%. The canopy's advantage stems from the absorption by leaves lower in the canopy of the scattered and reflected light created as light passes through the upper canopy.

To quantify the fraction of the top of the canopy (incident) radiation (f) that penetrates the canopy to a depth z , a version of Bouguer's or Beer's law is often used:

$$f_z = e^{-KL_T}, \quad [10]$$

where K is the extinction coefficient and L_T is the cumulative leaf area index between the top of the canopy and z . The fraction of the incident beam radiation intercepted by the canopy is $1 - f_z$. The extinction coefficient K varies with the solar zenith angle (thus K varies both diurnally and seasonally), the optical properties of the leaves, and the canopy's leaf angle distribution; thus, it can vary largely with the canopy architecture. Equation [10] can also be used to estimate the attenuation of other streams of radiation with canopy depth (e.g., net all-wave radiation or PAR) as long as the appropriate radiation-specific K is used.

When L is measured, either optically with commercially available instruments or by leaf harvesting or litter collection, eqn [10] allows for profiles of light penetration to be calculated (Figure 8). This in turn allows canopy processes

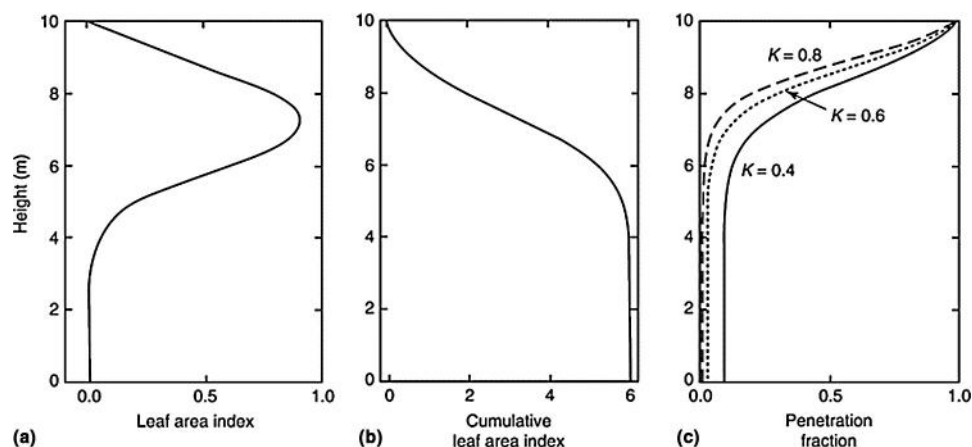


Figure 8 Example of the profiles of the leaf area index (a), the cumulative leaf area index (b), and the fraction of radiation at the top of the canopy that penetrates to a given height calculated using eqn [10] (c). The extinction coefficient K determines the rate at which the radiation is attenuated through the canopy.

that are driven by light absorption to be calculated at multiple levels within a canopy. Use of eqn [10] assumes a homogeneous canopy with randomly distributed leaves that absorb all of the radiation. Realistically, leaves are not randomly distributed in a canopy but are often grouped or ‘clumped’ together around whorls or branches. To account for this, eqn [10] can be modified by multiplying L_T by a clumping factor, Ω . Various additional modifications to eqn [10] exist to account for light scattering and reflection, multispecies canopies, and spatially heterogeneous canopies. Nonetheless, inherent in all of these modifications is the assumption that radiation decreases exponentially through the canopy as a function of the cumulative leaf area index and an extinction coefficient.

The leaf area index and the structure and arrangement of leaves are influenced by the quality of the light being received at the leaf level. Changes in not only the wavelength of the incident radiation but also the direction of that radiation (e.g., as a direct beam from the Sun or diffused after scattering) have been shown to affect the rate of photosynthesis and hence plant growth and productivity. For example, many species, especially coniferous ones with their small, cylindrically leaves and canopy structure, have higher photosynthetic rates (greater carbon uptake) and light use efficiencies under cloudy conditions when the portion of diffuse-beam PAR radiation is high, since the leaves (canopy) can absorb radiation from multiple angles due to their cylindrical shape. In contrast, broad-leaf species (e.g., most deciduous species) have higher photosynthetic rates under clear-sky conditions since their leaves are better positioned to intercept PAR radiation at lower solar zenith angles. Recent global surface and satellite-based measurements of surface solar radiation have shown periods of decreased surface solar radiation (global dimming; ~1960–90) and subsequent periods of increased surface solar radiation (global brightening; ~1990–present). These changes will affect plant productivity, for example by increasing carbon uptake during periods of increased cloud cover due to the increase in the fraction of diffuse-beam relative to direct-beam solar radiation.

Canopy Productivity

For a canopy, we define the net primary productivity (NPP) as the difference between gross primary productivity (photosynthesis, or GPP) and respiration (R):

$$\text{NPP} = \text{GPP} - \text{R}. \quad [11]$$

In other words, the carbon gained by the canopy is the difference between the amount of carbon that is gained by photosynthesis minus losses of carbon by respiration. If water and nutrients are available, eqns [1] and [11] show that the NPP of a canopy is limited only by the absorbed PAR and by respiratory carbon loss, the latter largely a function of temperature. Since NPP, or canopy growth, is a function of these two different processes, changes in the environment that affect the balance between GPP and R will affect canopy growth.

The main variables affecting GPP have already been discussed, namely, light absorption, leaf area, and canopy architecture. Other variables, including water and nutrient availability, also affect GPP. For example, the growth of most plants is limited by the relatively low atmospheric CO_2 concentration, and an increase in atmospheric CO_2 concentration would result in increased photosynthesis, biomass, and water use efficiency in many species, as discussed in this article. Recently, FACE has been investigating the effects of increasing atmospheric CO_2 concentration (370 versus $550 \mu\text{mol mol}^{-1}$) under open-air conditions on several natural ecosystems. With CO_2 enrichment, results have shown increased photosynthesis rates by up to 75%, increased canopy temperature due to reduced transpiration by partial stomata closure, and increased success and dominance of exotic, invasive species.

Any increase in GPP due to an increase in atmospheric CO_2 concentrations, however, may be limited by the availability of several other nutrients, especially nitrogen. Despite an abundance of nitrogen in the atmosphere (~78% in the homosphere), it is not in a form that plants can use due to the stable triple covalent bond between the two nitrogen atoms, which is difficult to break. Only roughly 10% of the nitrogen in a usable

form is made available to plants in precipitation by breaking of the bonds via lightning. The remaining 90% is made available through the process of biological nitrogen fixation. As relatively few plant species are themselves capable of converting or fixing nitrogen into a usable form, they rely on microorganisms to do this. The cycling of nitrogen in its various forms between the atmosphere, plants, and soil is the key to productivity in many of the world's canopies.

Especially in forests, NPP is the difference between two large numbers; respiration is just as important as or more important than GPP. Several large-scale experiments such as the Boreal Ecosystem-Atmosphere Study (BOREAS), EUROFLUX (the Long-term Dioxide and Water Vapor Fluxes of European Forests and Interactions with the Climate System project), and AmeriFlux were conducted to study the long-term carbon balance across several forests around the world. Some results show that, especially at higher latitudes, it is the ecosystem respiration that determines the net ecosystem carbon exchange, not photosynthesis. Moreover, the winter season is important for the annual NPP since R , although small, is large annually when cumulated over the longer winter season. As short-term experiments have shown that microbial respiration increases with warming, this suggests that with global warming, canopies will become less effective in removing atmospheric CO_2 since any carbon gains would be offset by the increased respiratory carbon loss. Over long periods of time (decades), however, recent studies have shown no control on respiration rates by temperature. If this were correct, global warming would result in an increase in atmospheric carbon sequestration by plant canopies, as the respiratory carbon loss would not increase.

Wind and Turbulence

The final section in our discussion of canopy processes is on the effects of wind. Wind cannot only affect the shape and size of vegetation (Figure 9) but also influence the turbulent dispersion of heat and gases such as water vapor and CO_2 . It also influences light penetration (through the movement of leaves creating sun flecks); the removal of intercepted precipitation; the transport of particulates such as pollen, spores, and pollutants; and canopy stability (wind sway or wind throw).

Air flowing over a leaf does not behave as if the leaf were a flat, smooth surface. Serrations at the edges of leaves as well as leaf ribs generate turbulence over the leaf. Turbulence serves to increase the exchange of heat, water vapor, and CO_2 , since exchange in turbulent air is much more efficient than exchange in still air. Over a canopy, air flow is reduced both by skin friction, or the transfer of momentum across the horizontal canopy boundary layer, and by form drag, or the force exerted when air strikes the canopy in a direction other than horizontal. The combination of these two forces is expressed by the total shearing stress, τ :

$$\tau = K_M \rho \frac{du}{dz}, \quad [12]$$

where K_M is the eddy viscosity, and du/dz is the change in horizontal wind speed u with height z . Since leaves and canopies are flexible, τ changes with wind speed as the canopy changes form, often becoming more streamlined as u increases.



Figure 9 Effects of persistent high wind speeds on the shape and structure of alpine vegetation in Colorado. A common alpine vegetation feature known as a krumholtz forms with a dense mat of branches near the ground, with a few stems with branches prevalent on the leeward side.

Wind can also affect photosynthesis, especially in understory species and leaves. A thick, multilayered canopy can effectively reduce u within the canopy, sometimes resulting in leaf temperatures and relative humidity increasing, which, as shown in eqn [8], acts to increase leaf stomatal conductance. Sun flecks, patches of full sunlight that pass through gaps in the canopy, can increase in frequency, duration, and size as wind moves the canopy. Exposure to sun flecks has been found to lower the photosynthetic rate in light-intolerant understory species or increase the photosynthetic rate in shade-tolerant understory species.

Canopies can influence the wind around them, in addition to being affected by the wind itself. Depending on the canopy properties (height, density, and leaf area index), extensive canopies alter the wind profile by creating a layer known as the roughness boundary layer, a layer of turbulent flow at the top of the canopy that influences the exchange of heat and mass (carbon and water vapor) from the vegetation. Within many forests, higher wind speeds are observed within the trunk space due to decreased canopy structure (fewer leaves and branches) and topography-driven cold-air drainage flows. In boreal or alpine regions, vegetation often promotes the retention of the beneath-canopy snowpack due to the canopy's shading of the snowpack, providing the benefit of insulating the soil (roots) from subzero temperatures and temperature fluctuations, and providing soil moisture during spring snow melt. In the lee of canopies and even isolated vegetation patches such as alpine krumholtz, the decrease in wind speed is often sufficient to deposit wind-blown snow and create snowdrifts (Figure 10). If properly designed,

vegetation can be used to create effective and economical living snow fences.

Precipitation Interception

Canopies are effective at intercepting precipitation. Water external to the leaf is free from direct physiological controls, thus it is either evaporated (rainfall) or sublimated (snowfall); both processes are influenced by wind. Depending on the canopy leaf area and architecture, roughly 10–55% of the total annual precipitation can be intercepted and subsequently lost through evaporation. Generally, the ability to intercept precipitation increases with leaf area, and the ability to evaporate this precipitation also increases due to a decrease in the aerodynamic resistance with increasing leaf area. Increasing wind speed increases the evaporation rate of intercepted rainfall by decreasing the aerodynamic resistance or by mechanically shedding water from leaves, resulting in throughfall. Especially in coastal and mountainous environments, leaves can act as condensation nuclei as fog collides and coalesces on individual leaves.

Similarly, as the canopy's L increases, the potential to intercept snow increases, and therefore, the snow water equivalent at the forest floor decreases. Sublimation losses of intercepted snow are especially significant in boreal coniferous stands (e.g., up to 40% of the annual precipitation) where the leaves are present during the long winter snowfall season. The bridging of snow across leaves and branches can effectively increase the interception area, and strong winds can effectively remove large volumes of snow off the canopy.



Figure 10 An example of vegetation's ability to alter its microclimate. Just above the alpine tree line in Colorado, the decrease in wind speed behind the vegetation created a snowdrift that insulated leaf tissue and roots from temperature extremes and provides valuable soil moisture.

Particulate Transport

The transportation of particulates such as pollen, spores, and pollutants is affected by wind, especially within the canopy where origination or deposition of many particulates occurs. Wind decreases exponentially with height within most canopies, with an attenuation or extinction coefficient generally a function of the cumulative L . In canopies with a dense overstory and open space beneath, however, the wind speed and direction can be unrelated to that above the canopy due to within-canopy horizontal pressure differences. In such canopies, a secondary maxima in u is often observed within the stem space.

In canopies, both the take-off and dispersion or flight of particles must be considered. To dislodge a particle from a surface, the drag force (a function of u) and lift force produce a turning moment (created by a difference in static pressure between the top and bottom of the particle) that must exceed the turning moment due to the force of gravity. Hence, a threshold u at the particle level is required to entrain the particle into the atmosphere, yet the u at this level is usually very low with nonturbulent flow. Therefore, many species have developed active methods to project or assist the entrainment of particles into the moving air stream.

Once the particle is in the air stream, its path is usually chaotic and random, and it has been successfully described for small particulates by diffusion theory borrowed from air pollution studies, Lagrangian simulations of particle trajectories, or, more recently, large-eddy simulations. The particle will land when the forces of gravity and drag exceed the lift force.

Structural Failure

The structural failure of canopies is of great economical significance for agricultural canopies, whether crops or forests. The shaking of plants as a result of wind has been proven to inhibit plant growth, as has the exposure of plants to particulates such as ice crystals or sand. The force required to catastrophically overturn a plant can be calculated from the forces acting on the plant, where the lateral force required to exceed the turning moment is a function of u and the drag coefficient for the canopy. The average threshold u required for catastrophic plant structural failure, however, fails to predict when stem failure will occur because canopies behave as aeroelastic structures. Such structures fail when the gust frequency, with mean wind speeds that may be below the theoretically required threshold average wind speed, coincides with the natural oscillation frequency of the canopy. Additional factors such as soil texture, structure, and moisture should also be considered, as mechanical failure of the soil could occur if, for example, the soil became saturated.

Remote Sensing of Canopy Properties

Given the sparseness of ground-based point measurements of canopy processes, and the need for large-scale and long-term observations, there have been many recent developments in the use of satellite observations to quantify the biophysical properties of plant canopies. For example, data from the NOAA Advanced Very High Resolution Radiometer

(AVHRR) can be used to calculate several biophysical properties at a resolution of 1.1×1.1 km. Of the several remote-sensing vegetation indices that exist, one using AVHRR data is the Normalized Difference Vegetation Index (NDVI):

$$\text{NDVI} = \frac{(\rho_{\text{NIR}} - \rho_{\text{RED}})}{(\rho_{\text{NIR}} + \rho_{\text{RED}})}, \quad [13]$$

where ρ_{NIR} ($0.7\text{--}1.1\ \mu\text{m}$) and ρ_{RED} ($0.6\text{--}0.7\ \mu\text{m}$) are the reflectances in the near-infrared and red wave bands, respectively, and can be used to estimate L and the fraction of PAR absorbed by a canopy (FPAR). Complementary to NDVI, the Moderate Resolution Imaging Spectroradiometer (MODIS) can be used to calculate the Enhanced Vegetation Index (EVI), which has greater sensitivity to canopy structure:

$$\text{EVI} = \frac{2.5(\rho_{\text{NIR}} - \rho_{\text{RED}})}{(L_C + \rho_{\text{NIR}} + C_1\rho_{\text{RED}} - C_2\rho_{\text{BLUE}})}, \quad [14]$$

where L_C (typically 1) is the canopy background adjustment factor, C_1 and C_2 are aerosol resistance weights (typically 6 and 7.5, respectively), and ρ_{BLUE} is the reflectance in the blue wave band ($0.45\text{--}0.52\ \mu\text{m}$). Hyperspectral remote-sensing instruments (capable of resolving hundreds of spectral bands), such as the Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) and Hyperion, can be used to not only identify vegetation cover but also identify the vegetation type (species). Various passive microwave sensors and radar can be used to detect changes in soil moisture, especially in agricultural crops when the canopy cover is still sparse. New satellite-based measurements such as those provided by the Gravity Recovery and Climate Experiment (GRACE) have been used to help estimate canopy processes such as evapotranspiration in combination with measured precipitation and runoff measurements.

Whereas L and FPAR can be used as measures of the capacity of the canopy to intercept PAR, the actual photosynthetic rate can be determined from the absorbed PAR (APAR). This can be derived from knowledge of APAR below the top of the canopy (APAR_{SFC}), the amount of PAR reflected from the top of the canopy (A_{PAR}), and FPAR:

$$\text{APAR} = \frac{\text{APAR}_{\text{SFC}}}{1 - A_{\text{PAR}}} \text{FPAR}. \quad [15]$$

The ability to remotely sense APAR is important as APAR inherently drives all of the canopy processes discussed in this article.

Despite this tremendous potential for remotely sensing these canopy properties, complex algorithms are required to correct for atmospheric effects such as clouds, aerosol and ozone absorption, sun angle, and bidirectional reflectance. Thus, ground-based point measurements are still required to ensure that the remotely sensed canopy properties are indeed accurate.

See also: Agricultural Meteorology and Climatology. **Boundary Layer (Atmospheric) and Air Pollution:** Microclimate; Observational Techniques *In Situ*; Observational Techniques: Remote. **Land-Atmosphere Interactions:** Trace Gas Exchange.

Further Reading

- Campbell, G.S., Norman, J.M., 1998. Introduction to Environmental Biophysics, second ed. Springer-Verlag, New York.
- Grace, J., Ford, E.D., Jarvis, P.G. (Eds.), 1981. Plants and Their Atmospheric Environment. Blackwell Scientific Publications, Oxford.
- Kaimal, J.C., Finnigan, J.J., 1994. Atmospheric Boundary Layer Flows: Their Structure and Measurement. Oxford University Press, New York.
- Kirkham, M.B., 2011. Elevated Carbon Dioxide: Impacts on Soil and Plant Water Relations. CRC Press Taylor and Francis Group, Boca Raton.
- Jones, H.G., 1992. Plants and Microclimate. Cambridge University Press, Cambridge.
- Marshall, B., Roberts, J.A. (Eds.), 2000. Leaf Development and Canopy Growth. Sheffield Academic Press, Sheffield.
- Monteith, J.L. (Ed.), 1975. Vegetation and the Atmosphere. Principles, vol. 1. Academic Press, London.
- Monteith, J.L., Unsworth, M.H., 2008. Principles of Environmental Physics, third ed. Academic Press, Oxford.

Trace Gas Exchange

JN Cape and D Fowler, Edinburgh Research Station, Midlothian, UK

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

The theory and background to the exchange of trace gases at the Earth's surface are presented as an introduction to the experimental methods used for measuring trace gas exchange, and the roles of weather and surface vegetation in controlling the process. Measurement systems over a range of spatial scales, including small chambers, micrometeorological techniques, and aircraft studies, are described, and the use of measurement data in deposition models is briefly introduced.

Introduction

The exchange of trace gases between the Earth's surface and the atmosphere is a fundamental process of life. Primary productivity, as photosynthesis, involves uptake of carbon dioxide and release of oxygen and water vapor. Many trace gases are also involved in biogeochemical cycling of nutrient elements, in particular, sulfur and nitrogen. Gases such as sulfur dioxide (SO_2), hydrogen sulfide (H_2S), carbonyl sulfide (COS), dimethylsulfide ($(\text{CH}_3)_2\text{S}$), nitrous oxide (N_2O), nitric oxide (NO), nitrogen dioxide (NO_2), and ammonia (NH_3) are emitted from and absorbed at the Earth's surface, by plants or soil, as part of natural processes. Industrialization has led to large artificial emissions, especially from combustion sources, of many of these trace gases, which are then returned to vegetation, soil, and water surfaces with important implications for effects including acidification and eutrophication. It has therefore become important to be able to estimate the exchange of trace gases between the surface and the atmosphere. The deposition of pollutant gases to vegetation and soils (*dry deposition*) contributes to the nutrient budgets of ecosystems, and may exceed thresholds above which irreversible changes occur. Direct gaseous deposition forms part of 'Critical Load' calculations, which are used in Europe as a means of setting strategies for pollution control. Trace gas exchange is also measured to assess responses to land-use change and pollutant stress. For example, increased deposition of N (either as fertilizer, or as atmospherically deposited pollution) can increase soil emissions of nitrous oxide (N_2O), which is an important 'greenhouse' gas. Methods of measuring trace gas exchange have also been developed to investigate the natural biogenic emissions of other radiatively active gases, such as methane (CH_4), so that forecasts can be made of the effects of changing land use or climate on uptake by, and emission from, soils. Vegetation also emits a wide range of gases, including volatile organic compounds (VOCs), which contribute to the formation of tropospheric ozone (O_3). Knowledge of biogenic VOC emission rates, and the factors that control them, is necessary for modeling the production of ozone on regional and global scales. Deposition at the ground, and uptake by vegetation, is one of the major sinks for tropospheric ozone, so this process also must be understood before atmospheric chemistry and transport models can be developed.

Transfer Processes

The transfer of gases between the Earth's surface and the atmosphere can occur in both directions, and is controlled both by atmospheric processes and by the state of the ground surface. Sources and sinks of trace gases at the ground may be in the soil itself, or both in and on the vegetation. The magnitude of the trace gas flux depends on both the physical and physiological state of the vegetation – that is, is controlled by the structure of the vegetation as it affects air turbulence, and by the biological and chemical processes occurring inside and on the surface of leaves.

Some of the trace gases, such as ammonia and nitric oxide, are deposited on and emitted from natural surfaces and the exchange is therefore bidirectional, as is the case for water vapor and carbon dioxide. Other trace gas fluxes are unidirectional, for example, the very reactive gases nitric acid and hydrogen chloride. Fluxes of trace gases to or from the surface are usually expressed as mass fluxes per unit ground area (μg or $\text{ng m}^{-2} \text{s}^{-1}$). These are often referenced to an air concentration (for downward fluxes) at a fixed height above the ground, expressed as a mass per unit volume ($\mu\text{g m}^{-3}$), in part to normalize the flux for the ambient concentration. The ratio of flux to concentration in these units is called the deposition velocity (v_d), because it has units of velocity (m s^{-1}).

Since the late 1970s, methods for measuring trace gas concentrations over short time periods (several minutes or less) have been developed, which have permitted trace gas fluxes to be measured over a range of different land surfaces by measuring vertical concentration gradients. More recently, the development of rapid response gas sensors, which can provide precise concentration measurements at frequencies of tens of Hertz, has permitted the direct measurement of trace gas fluxes using eddy covariance methods. Both of these techniques integrate over a 'footprint' of land surface upwind of a fixed measurement point, and several conditions must be satisfied for the micrometeorological theory used to determine fluxes to be valid. For example, the terrain should be uniform, and there should be very small horizontal gradients in concentration, or small changes in concentration with time, to avoid advection and storage errors in the measured flux. On a smaller scale (typically of the order of $0.1\text{--}10 \text{ m}^2$), trace gas fluxes can be measured using boxes placed on the ground, in which changes in trace gas concentrations are measured either statically or

dynamically. On a larger scale, trace gas fluxes have been calculated on scales of around 10^2 km^2 by measuring changes in trace gas concentration with time below a well-defined inversion layer when vertical mixing is suppressed. At much larger scales, fluxes have been inferred from measurements in the well-mixed boundary layer upwind and downwind of a region typically 10^4 km^2 in area.

Pathways of Gas Transfer

The transfer of gases between the atmosphere and the surface, whether bare soil or vegetation, may be thought of in terms of three consecutive pathways. Each of these pathways imposes a resistance on the overall transfer process, and the network of resistances is a convenient way of describing the different contributions of the pathways to the overall process, by analogy with an electrical circuit (Figure 1). In this model, resistance has units of time per unit length (s m^{-1}); the reciprocal (units: m s^{-1}) is referred to as a conductance. The flux is then driven by the difference in trace gas concentration between a reference height (z) in the atmosphere and the concentration at the sink (or source) at the surface, analogous to a potential difference applied across the combined resistances of the transfer pathways.

The first pathway is the transfer of gases by turbulence within the atmosphere, and depends upon the surface roughness and wind speed. The associated aerodynamic resistance (r_a) decreases with increasing wind speed, and is small for rough vegetation (e.g., forests; typically $2\text{--}10 \text{ s m}^{-1}$) and larger for short vegetation (e.g., short grass; typically $30\text{--}300 \text{ s m}^{-1}$). It is implicitly a function of height above the canopy (z), and is

generally calculated between a reference height ($z - d$) and the notional height ($z_0 + d$), where d is the displacement height associated with aerodynamically rough plant canopies, and z_0 is the roughness length of the surface. The second part of the transfer process involves the molecular diffusion of gases across the laminar boundary layer that forms at surfaces. The thickness of this layer is a function of the geometry of the leaf or surface element, and also depends on the wind speed. Although several formulations of this term (r_b) exist, it is generally small relative to the third, or surface resistance (r_c), which for most of the trace gases is the controlling step in the transfer pathway. When considering the flux of gases to or from a vegetation canopy, the surface resistance term includes several parallel pathways that depend upon the characteristics of the ultimate absorbing (or emitting) surface. Different sinks (and sources) are present in a canopy, and the surface resistance is often split into a stomatal term and a nonstomatal term. The nonstomatal term may be further decomposed into parallel pathways to understorey vegetation or bare soil, dry canopy surfaces (e.g., leaf cuticles), and wet surfaces. These parallel resistances are shown in Figure 1. Biological or chemical processes occurring at the surface control the magnitude of the surface resistance. Stomatal resistance increases as stomata close, either at night or in response to a large vapor pressure deficit. Stomatal resistance to trace gas transfer can be estimated from measurements of water vapor exchange through stomata, making allowance for the relative diffusion rates of the trace gas molecules. Alternatively, stomatal resistance may be modeled based on parameterizations of stomatal opening and closure based on light intensity and vapor pressure deficit. The nonstomatal terms are difficult to model, particularly where the gas exchange process at the surface depends upon solubility or

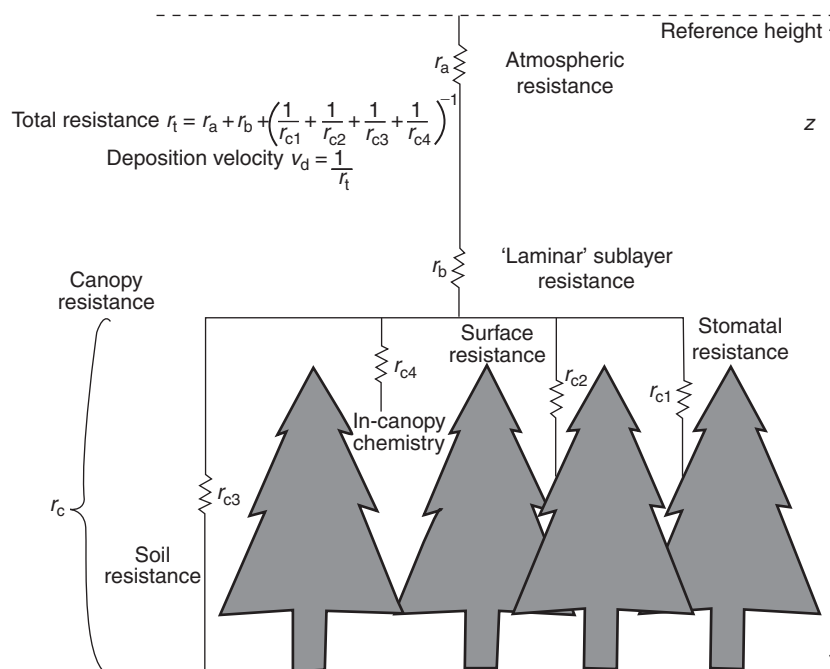


Figure 1 Schematic diagram showing the transfer pathways for a gas between the atmosphere and the Earth's surface, and the associated transfer resistances, which determine the overall transfer rate or deposition velocity. The deposition velocity is referenced to a height (z) above the ground.

reaction in water. The nonstomatal resistance is often obtained as the residual between the (measured) surface resistance and an estimated or modeled stomatal resistance.

Micrometeorological Theory and Methods

Measurements of trace gas exchange to the land surface over distances of 10–1000 m may be made by three techniques, which operate on different timescales. The particular method employed may be determined by the availability of an appropriate sensor for measuring the trace gas concentration and the characteristics of the surface over which measurements are to be made.

The Aerodynamic Gradient Method

The transport of gases to and from the surface is regarded as occurring by the same turbulent processes that transport momentum and heat. The flux is then determined from measurements of the vertical gradient above the surface of wind speed, temperature, and gas concentration. Formally, the flux of trace gas s (F_s) is proportional to the vertical concentration gradient: $F_s = -\rho K_s \partial s / \partial z$, where K_s is the transfer coefficient (eddy diffusivity) for gas s , ρ is the air density, and $\partial s / \partial z$ is the vertical gradient in concentration (expressed as a mixing ratio in dry air). By convention, fluxes toward the surface ($\partial s / \partial z < 0$) have a negative sign. Similarly, the flux for momentum may be defined in terms of the vertical gradient in wind speed, and the flux for heat in terms of the vertical gradient in potential temperature, with transfer coefficients K_m and K_h , respectively. The transfer coefficients are functions of the measurement height (z). In atmospheric conditions of neutral stability, the eddy diffusivities for heat, momentum, and trace gases are equal, so that the trace gas flux may be calculated from measurements of the momentum flux obtained from the vertical gradients of wind speed. This technique can use relatively slow analytical methods for measuring concentration gradients, on the assumption that atmospheric turbulence is relatively constant over periods up to about 30 min, but requires a capability to measure concentration differences with height of a few percent of the mean. In stable or unstable conditions, corrections must be applied to correct for differences between eddy diffusivities for heat and momentum transfer. Semiempirical relationships have been developed based on the measured gradients in wind speed and potential temperature, which permit the estimation of trace gas fluxes under a wide range of conditions. The development of conditional time-averaged gradient (CoTAG) methods, in which average vertical gradients are measured over long periods but only in conditions of neutral stability (by turning on and off the gas sampling pumps), has meant that average fluxes can be measured over different vegetation types on a weekly or monthly basis, and these average fluxes provide parameters for use in deposition models.

The Eddy Covariance Method

Trace gases are transported to and from the surface in turbulent eddies of air. The method can be envisaged as estimating the net flux of a gas by the simultaneous measurement of the air

concentration and vertical wind speed in different upward and downward moving eddies as they move through a hypothetical plane above the surface. Formally, the flux (F_s) is given by the mean (denoted by a horizontal bar) of the product of vertical wind speed (w) and air concentration (χ_s):

$$F_s = \overline{w \cdot \chi_s} = \overline{w} \cdot \overline{\chi_s} + \overline{w' \cdot \chi_s'}$$

where w and χ_s are the mean vertical wind speed and air concentration, and w' and χ_s' are the instantaneous deviations from the means. In order to capture the full spectrum of eddy sizes responsible for trace gas transport, the wind speeds and air concentrations must be measured at a frequency of 5–20 Hz, depending on the scale of surface roughness and the height of the reference plane above the surface. The analytical method must also be able to resolve small differences in gas concentration at that frequency. Care must be taken to ensure that sensible and latent heat fluxes, which cause vertical gradients in air density, do not introduce large uncertainties into the measurements. This method is able to resolve rapid changes in surface characteristics, and is applicable for use in aircraft, where the horizontal resolution is determined by the frequency of measurements.

There have been many variants of the eddy covariance technique, mainly to overcome the limitations in the ability of the available instrumentation to sample the trace gas concentration at a sufficiently high frequency. A technique known as disjunct eddy covariance takes a series of spot samples of air at discrete intervals and collects the high frequency turbulence data for each of the air samples. In this way the air samples may be analyzed for their composition (e.g., for VOCs), and the flux calculated from the statistics of the sampling regime.

The Eddy Accumulation Method

This is a variant of the eddy covariance method, but does not require rapid gas analysis. Instead, the air sampling system is controlled by the vertical wind speed sensor (usually a sonic anemometer) to direct the sample air stream through one of three sample inlet lines, where gas can be accumulated (in a bag, or trapped on an absorbent material) prior to analysis. The three inlet lines correspond to upward moving eddies, downward moving eddies, and slack air. Very rapid (>1 Hz) valve switching is required, and the sampling rate should be proportional to the measured vertical wind speed. In practice, this latter requirement is technically difficult to achieve, and the method of relaxed eddy accumulation is more often employed. In this case, the sampling rate remains constant, and the vertical flux is estimated as the product of the standard deviation of vertical wind speed, the difference in average concentration in upward and downward moving eddies, and a scaling factor that depends on the nature of the surface. The scaling factor is determined by reference to the flux of water vapor obtained by eddy covariance or gradient methods, or may be estimated empirically for a given surface.

Constraints and Uncertainties in Measurements

The above methods are based on the assumption that the flux to or from the surface is sufficiently close to the value measured at some reference height above the surface that

simple corrections may be applied to derive the actual surface flux. In practice, several processes may upset this 'constant flux' assumption:

1. Chemical reaction between the measurement height and the ground

Trace gas concentrations may be depleted or enhanced below the measurement height, so that there are additional sinks/sources besides transfer at the surface. Examples include the loss of ozone (O_3) by reaction with nitric oxide (NO), which may have originated from microbial processes in the soil, the production of nitrogen dioxide (NO_2) from the same reaction, or the reaction of terpenes and other biogenic hydrocarbons with O_3 and hydroxyl (OH) radicals. The uncertainty is greatest for fast chemical reactions that occur on timescales similar to the transport processes. In such cases, actual surface fluxes can usually only be calculated by independent measurement of all the factors influencing the reaction rates, and detailed modeling of the coupled transport/chemistry system. Simplifications can be introduced if constraints on the system can be measured independently (e.g., in the case of $NO + O_3$, by measuring the NO flux from the soil surface directly).

2. Changes in mean concentration with time

Changes in air concentration with time lead to 'storage' errors, where a component of the vertical transport relates to the restoration of equilibrium across the reference plane as concentrations above or below change through factors not related to the surface processes. Large uncertainties can arise when there are marked changes in mixing layer height, for example, just after sunrise.

3. Horizontal gradients in concentration

Uncertainties are caused by advection of air horizontally below the measurement height when there are horizontal gradients in air concentration, and may be important for major pollutant gases close to sources. For example, an error of around 40% is introduced for a reference height of 1 m by a horizontal gradient of $10 \mu\text{g m}^{-3} \text{ km}^{-1}$ where the vertical flux is $50 \text{ ng m}^{-2} \text{ s}^{-1}$. Such gradients may be found close to urban areas or large point sources, in regions where flux measurements may be important in determining the rates of pollutant deposition and their effects.

Large-Scale Measurements

The Nocturnal Box Method

When a stable temperature inversion forms close to the Earth's surface, the air closest to the ground is effectively isolated from the atmosphere above. Provided that the boundary layer is well mixed, the emission or deposition of gases at the surface results in a gradual increase or decrease in concentration under the inversion. If the depth of the surface layer is known, or can be estimated, the change in concentration with time can be used to infer the rate of exchange of a gas at the surface. Where conditions are stable for several hours, the change in concentration measured at a fixed point follows a first-order increase or decrease from which the flux may be calculated directly. Such measurements integrate over a large surface area. This

method is not applicable to gases where chemical production or loss in the surface air would cause temporal changes in concentration, but has been used to estimate rates of O_3 deposition in rural air (where reaction with NO is negligible), and the deposition of the nonreactive gases CO and H_2 .

The Daytime Boundary Layer Budget Method

In a well-mixed daytime boundary layer with a well-defined capping inversion, and in the absence of deep convection, concentration measurements made within the boundary layer can be taken as representative of the whole mixing depth, typically of the order of 1 km. If there is a steady horizontal wind over a period of hours then air can be sampled (from an aircraft) along a transect upwind of a region, and subsequently measured along a transect downwind of the region. In this context a region can be several hundred kilometers across. The flux to or from the surface is then calculated from a mass balance of gas entering and leaving the hypothetical box formed by the capped boundary layer. In practice, this method is only applicable to gases that react very slowly in the atmosphere, but has been applied to the estimation of total fluxes of methane (CH_4) and nitrous oxide (N_2O) from the United Kingdom.

Small-Scale Measurement Methods

Gas exchange fluxes can be measured on spatial scales of $0.1\text{--}10 \text{ m}^2$ using enclosure methods. At small spatial scales there are problems with surface heterogeneity, and such techniques are often used to elucidate functional relationships between fluxes and environmental conditions such as temperature or surface wetness, while recognizing that the absolute fluxes measured may vary greatly over distances of $10\text{--}100 \text{ m}$.

Static Boxes

This technique is well suited to unreactive gases such as CH_4 and N_2O , where interactions with the box materials are not important. The rate of change of gas concentration within a box sealed to the surface is measured with time, and the flux determined from the volume–surface area ratio and the rate of change in concentration. The method is well suited to sampling fluxes across wide-ranging conditions of soil and short vegetation, but using such data to estimate field-scale fluxes generates large uncertainties unless the surface is homogeneous. The environment inside the closed box will in general be very different from that outside, with turbulent mixing greatly suppressed (unless an internal fan is used), and large changes in concentrations of water vapor and carbon dioxide if vegetation is enclosed.

Dynamic Boxes

As in the static box method, a box is sealed to the surface, and the gas concentration is measured in air pumped into and out of the enclosed volume. The surface flux is then calculated from the volume–surface area ratio, the airflow rate through the box, and the difference between inlet and outlet concentrations.

Although still subject to the same problems of sampling the underlying surface as for static boxes, the finite residence time means that fluxes of chemically reactive gases can be measured in this way, particularly if the box is constructed of chemically inert materials. However, such methods inevitably alter the conditions above the surface, if only because the surface is isolated from atmospheric turbulence. Internal mixing fans may be used to provide turbulence, but in general these will not reproduce the air movements in the absence of the box. This does not invalidate the use of such methods for process-based studies, but introduces unquantified uncertainties into the measured fluxes and their relationship to the actual surface flux outside the box.

Control of Trace Gas Fluxes by Surface Processes

Stomatal vs Nonstomatal Pathways

The transfer resistance between the atmosphere and the surface has been described in terms of atmospheric factors, which control r_a and r_b , and surface factors, r_c . In the simplest case, with highly reactive gases such as nitric acid vapor (HNO_3) or hydrogen chloride (HCl), the surface resistance term is small relative to the atmospheric transfer resistances. Under these conditions, trace gases are deposited to vegetation at rates approaching those for momentum, with deposition velocities up to several centimeters per second, depending on wind speed and surface roughness. At the other extreme, the nonstomatal surface resistance is very large, so that transfer of the gas only occurs when stomata are open. This type of behavior is observed for nitrogen dioxide (NO_2). More generally, emission or deposition occurs through both stomatal and nonstomatal pathways, with total surface resistances varying diurnally and seasonally. Although the contribution of the stomatal pathway can be estimated fairly precisely from simultaneous measurements of water vapor fluxes, elucidation of the nonstomatal pathways is more difficult. The nonstomatal resistance at the surface is usually obtained by difference between the momentum flux above a vegetation canopy ($r_c = 0$), the overall gas flux and the estimated or measured stomatal flux. The residual nonstomatal flux comprises several different potential sink or source processes, as illustrated in Figure 1. For example, although deposition of O_3 follows a diurnal cycle that coincides with stomatal opening, the nighttime flux (when stomata are closed) can be a large fraction of the daytime flux. Integrated over a whole year, the nonstomatal flux may even be greater than the flux to the inside of the leaf. However, the precise mechanisms regulating reactions on cuticular surfaces remain uncertain. The focus of interest in O_3 deposition is on the stomatal term, which leads to physiological damage to mesophyll tissue. For emission, similar patterns may be observed, with emissions from vegetation of some VOCs following stomatal opening, while emissions of other VOCs appear to be relatively invariant through day and night, and more responsive to temperature.

Chemical Reactions at the Surface

Additional complications arise for gases that dissolve in, or react with, surface water. Water-soluble gases such as sulfur

dioxide (SO_2) are rapidly absorbed through stomata, and if external leaf surfaces are dry the flux is determined by stomatal opening. However, in the presence of surface moisture, a parallel transfer pathway becomes possible, and the overall flux may be dominated by nonstomatal deposition. The surface moisture is not an infinite sink for SO_2 , and in practice the surface flux will be determined by factors such as the acidity of the surface water and the rate of oxidation to sulfate. A dynamic equilibrium may be established between the atmosphere and the solution of the trace gas on the surface, so that the net flux approaches zero. If the surface water subsequently evaporates, dissolved SO_2 may be released back into the atmosphere, producing an upward flux. If sufficient ammonia (NH_3) is also present, rapid uptake of this highly soluble gas into the surface water maintains the conditions of low acidity that favor SO_2 uptake, and the surface resistance can approach zero. More generally, a detailed understanding of all the chemical processes, including surface ion exchange and oxidation pathways, is needed before the overall transfer of SO_2 can be successfully modeled.

Bidirectional Exchange

The surface exchange of NH_3 is possibly the most complex of all the trace gases. The internal intercellular fluids of leaves contain ammonium ions, which are in dynamic equilibrium with gaseous NH_3 . The overall direction of the surface exchange therefore depends upon the relative concentrations of the gas inside and outside the leaf. If the surface exchange were purely stomatal, then the process could be relatively simply modeled from knowledge of the internal ammonium concentration and the air concentration of the gas. However, NH_3 also interacts with external leaf surfaces, which can provide a sink for both atmospheric NH_3 and NH_3 emitted from stomata when air concentrations are very small. Moreover, the difference between internal and external NH_3 concentrations may vary through a plant canopy, so that NH_3 emitted close to the ground is reabsorbed before it can escape from the vegetation canopy. In this situation even the simple resistance model of Figure 2 is not appropriate, and terms equivalent to capacitors in the electrical analog must be introduced. Much progress has been made in recent years, however, and measured hourly fluxes (both upward and downward) have been successfully modeled over periods of days for several vegetation types.

Bidirectional trace gas fluxes are not restricted to NH_3 . Microbial processes in soil can switch from consumption to production of trace gases as soil water status changes. Depending on soil type, anaerobic conditions lead to emission of N_2O , CO , and CH_4 , while in drier soils CO and CH_4 are consumed, and NO is released rather than N_2O . The switch from one type of behavior to the other can be seen in the same soil, as the soil water status changes. At the field scale, local differences in soil chemistry and structure can mean that some patches are emitting a gas which is being absorbed only a few meters away. Perhaps the best illustration of this type of behavior is in blanket peat bogs, where CH_4 is consistently emitted from wet pools and removed (by oxidation) at neighboring drier tussocks. At the scale of the landscape, the net flux is determined by the relative proportion of wet and dry areas.

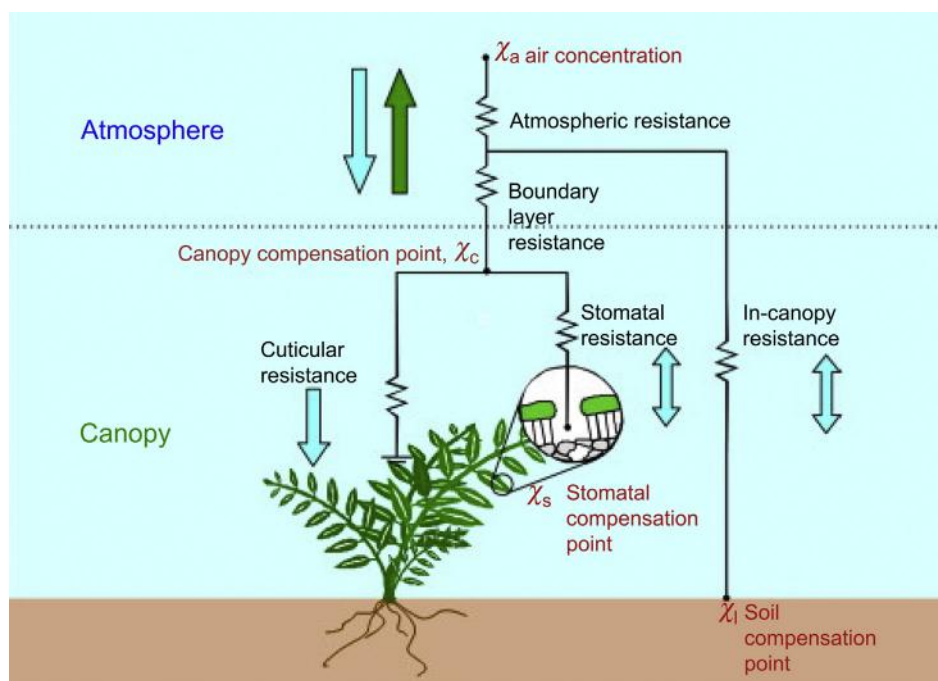


Figure 2 Schematic representation of bidirectional exchange using a resistance analogy (cf Figure 1). This formulation permits both upward and downward fluxes to be modeled by comparing the gas concentration above the surface (χ_a) with the canopy compensation point concentration (χ_c). Wichink Kruit, R.J., van Pul, W.A.J., Sauter, F.J., van den Broek, M., Nemitz, E., Sutton, M.A., Krol, M., Holtslag, A.A.M., March 2010. Modeling the surface-atmosphere exchange of ammonia. *Atmospheric Environment* 44 (7), 945–957.

Model Parameterization for Mapping Trace Gas Fluxes

One of the main objectives of measuring trace gas exchange is to be able to quantify fluxes at the regional, national, or global scale. The complexity of the processes involved, for all but a few gases, means that simplifications must be introduced into models designed to estimate large-scale fluxes and budgets. Initially, the modeling of surface exchange as part of schemes to represent long-range transport and fate of trace gases used single parameters for the deposition velocity of each gas. More complex models might vary the deposition velocity by night and day. As understanding improved, models were modified to calculate explicitly the atmospheric resistance terms, as a function of mean wind speeds and vegetation type, using climatological maps of wind speed and vegetation classifications based on survey or remote sensing. The most recent models include factors to account for surface wetness, and may model stomatal opening from parameterizations based on a range of different vegetation types and responses, with explicit dependence on temperature and light levels. The introduction of detailed surface chemical processes (e.g., the dependence of SO_2 deposition rates on the availability of NH_3) is still beyond the scope of regional and national modeling, not least because the required data on air

concentrations are not available at the spatial scales necessary. The evaluation of the uncertainties in such models, arising from the inevitable use of spatially and temporally averaged data (for meteorological as well as chemical parameters) is an active area of current research.

See also: **Climate and Climate Change:** Carbon Dioxide. **Satellites and Satellite Remote Sensing:** Water Vapor.

Further Reading

- Aubinet, M., Vesala, T., Papale, D. (Eds.), 2012. *Eddy Covariance: A Practical Guide to Measurement and Data Analysis*. Springer, Dordrecht, 442 pp.
- Burba, G.G., Anderson, D.J., 2010. *A Brief Practical Guide to Eddy Covariance Flux Measurements: Principles and Workflow Examples for Scientific and Industrial Applications*. LI-COR Biosciences, Lincoln, USA, 211 pp. http://www.licor.com/env/applications/eddy_covariance/book.html.
- Fowler, D., et al., 2009. Atmospheric composition change: ecosystems-atmosphere interactions. *Atmospheric Environment* 43, 5193–5267.
- Gasche, R., Papen, H., Rennenberg, H. (Eds.), 2002. *Trace Gas Exchange in Forest Ecosystems*. Kluwer, Dordrecht.
- Lee, X., Massman, W., Law, B. (Eds.), 2004. *Handbook of Micrometeorology*. Kluwer, Dordrecht.
- Monteith, J.L., Unsworth, M.H., 2008. *Principles of Environmental Physics*, third ed. Academic Press, London.

LIDAR

Contents

Atmospheric Sounding Introduction

Backscatter

Differential Absorption Lidar

Doppler

Raman

Resonance

Atmospheric Sounding Introduction

PS Argall and R Sica, The University of Western Ontario, London, ON, Canada

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

Light detection and ranging (Lidar) is a remote sensing technique that is used for measuring atmosphere parameters, such as temperature, composition, and wind. Lidar, like radar, operates by transmitting a beam of electromagnetic radiation and subsequently detecting any radiation scattered back to the instrument. The scattered radiation is analyzed in order to determine some property, or properties of the medium through which the radiation traveled. The range of atmospheric parameters that can be measured with lidar include: temperature, wind velocity, atomic and molecular species concentration, and aerosol and cloud properties.

Introduction

Light detection and ranging (Lidar) is a remote sensing technique that is used predominately for measuring atmosphere parameters, such as temperature, composition, and wind. Lidar operates on the same principle as radar; in fact, it is sometimes called laser radar. Both these techniques operate by transmitting a beam of electromagnetic radiation and subsequently detecting any radiation scattered back to the instrument. The scattered radiation is analyzed in order to determine some property, or properties of the medium through which the radiation traveled. Lidar and radar differ in the wavelength of the radiation utilized. Radar uses wavelengths longer than about 1 cm, in the radio-band, while lidar uses light in the ultraviolet, visible, and infrared, which in modern lidar systems, is generated by lasers. The different wavelengths used by radar and lidar leads to the very different forms the actual instruments take.

The range of atmospheric parameters that can be measured with lidar include: temperature, wind velocity, atomic and molecular species concentration, and aerosol and cloud properties.

In addition to its atmospheric applications, lidar is also used in ocean research and military applications including the

detection of chemical and biological agent and remote identification and tracking of vehicles. Lidar equipped binoculars are used by hunters and golfers as they provide accurate range measurements.

Evolution

The principle of lidar was first proposed in 1930. The original proposal suggested the measurement of atmospheric density profiles by the detection of scattering from a beam of light projected into the atmosphere. This proposed scheme suggested an antiaircraft searchlight as the source of the beam and a distant large telescope for the receiver. In this configuration, now known as bistatic, the range of the scattering can be determined by geometry. In the bistatic configuration, shown in [Figure 1](#), the receiver field of view is scanned along the transmitted beam in order to obtain an altitude profile of the scattered light.

The first results obtained using this principle were reported in the late 1930s when photographic recordings of light scattered from a searchlight beam were made.

Typically, modern lidar systems are monostatic in configuration with the transmitter and receiver colocated. Monostatic

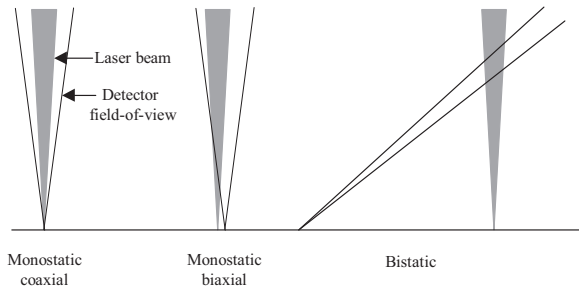


Figure 1 Three possible alignment arrangement of a lidar's transmitted beam and receiver field of view.

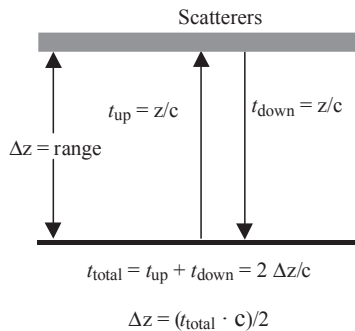


Figure 2 Schematic illustrating the process of ranging based on timing the returned signal.

systems can be subdivided into two categories; coaxial systems where the laser beam is transmitted coaxially with the receiver's field of view, and biaxial systems, where the transmitter and receiver are located adjacent to each other. Monostatic lidar systems use pulsed light sources, thereby enabling the range at which scattering occurs to be determined from the round trip time of the scattered light (**Figure 2**).

By the early 1950s refinements in technique and improved instrumentation, including electrical recording of the intensity of the backscattered light, allowed the measurement of atmospheric density profiles up to around 67 km altitude. These measured density profiles were then used to derive temperature

profiles using the Rayleigh lidar technique, which is described later.

The invention of the laser in 1960 and the giant pulse, or Q-switched, laser in 1962 provided a powerful new light source for lidar. The first use of a laser in a lidar system was reported in the early 1960s and since then developments in lidar have been closely linked to advances in laser technology.

Instrument Basics

Lidar hardware can be conveniently divided into three subsystems: the transmitter, the receiver, and the detection and recording systems. **Figure 3** is a block diagram of a generic lidar system, which shows how these subsystems work together to form a complete lidar.

Transmitter

The transmitter generates light pulses with the required properties and directs them into the atmosphere. Pulsed lasers, with their inherently low divergence, narrow spectral width, and short intense pulses are ideal as the light sources for lidar systems.

In addition to a laser, the transmitter of a lidar often includes a beam expander whose purpose is to reduce the divergence of the beam being transmitted into the atmosphere. This allows a reduction in the background measured by the lidar. At night, the background is due to light from the Moon, stars, airglow, and artificial lights. During the day background is predominately due to the Sun. Background can enter the lidar receiver either directly or after scattering in the atmosphere. A reduction in the divergence of the transmitted beam allows the field-of-field of the receiver to be reduced, resulting in a lower background.

The narrow spectral width of the laser has been used to advantage in a variety of ways in lidar systems. This narrow spectral width allows the spectral filtering of light by the lidar receiver. A band-pass filter tuned to the laser wavelength selectively transmits photons backscattered from the laser

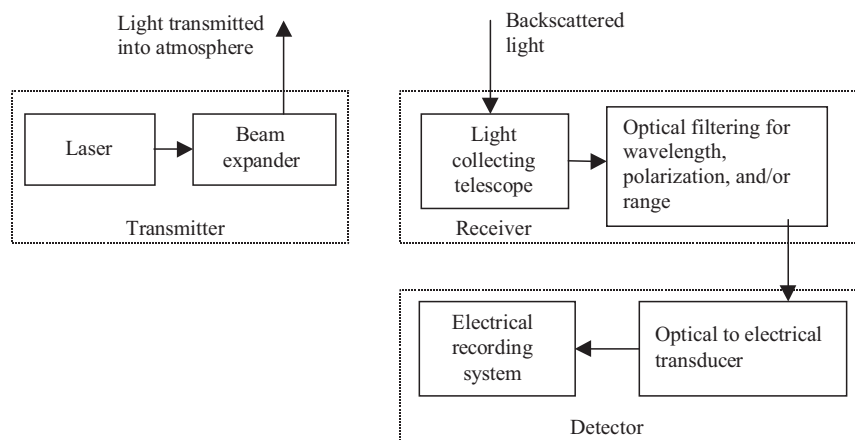


Figure 3 Schematic of a generic lidar.

beam, while rejecting photons at other wavelengths, thereby enabling a reduction in the background by several orders of magnitude. The pulse properties of pulsed lasers allow ranging to be achieved by timing the backscattered signal thus allowing the simpler monostatic configuration.

The major influence on the type of laser used in a lidar is the parameters the lidar is being designed to measure. Some measurements require a very specific wavelength and/or tunability, i.e., resonance-fluorescence and differential absorption lidar (DIAL). These types of lidars can require complex laser systems to produce the required wavelengths, while other simpler lidars, such as Rayleigh, Raman, and aerosol lidars, can operate over a wide wavelength range. Although it may be possible to specify the exact performance characteristics of the laser required of a particular lidar measurement, these characteristics often need to be compromised in order to select from the types of lasers available.

Receiver

The receiver of a lidar collects and processes the scattered laser light before directing it onto the detector. The first optical component, the primary optic, in the receiver usually has a large diameter enabling it to collect a large amount of the scattered laser light.

Lidar systems typically utilize primary optics with diameters ranging from about ten centimeters up to a few meters in diameter. Optics at the smaller end of this scale are used in lidar systems that are designed to work at close range, a few hundred meters, and may be lenses or mirrors. Optics at the larger end of this range are used in systems designed to probe the middle and upper atmosphere and are typically mirrors.

After collection by the primary optic, light is usually processed in some way before being directed to the detector system. Processing can be based on wavelength, polarization, and/or range depending on the purpose for which the lidar has been designed.

As described previously, the simplest form of processing based on wavelength is the use of a narrow-band interference filter to reduce the background. Much more sophisticated spectral filtering schemes are employed in Doppler and high-spectral-resolution lidar systems.

Signal separation based on polarization is a technique that is often used in the study of atmospheric aerosols. Information on aerosol properties can be obtained from the degree to which light scattered from a polarized laser beam is depolarized.

Processing of the backscattered light based on range can be performed in order to protect the detector from the intense near-field returns of high-power lidar systems. This protection is achieved by using a fast shutter that closes the optical path to the detector while the laser is firing and for a short time afterward. The shutter opens again in time to allow transmission of light backscattered from the range being studied.

Detection and Recording

The signal detection and recording section of a lidar takes light from the receiver and produces a permanent record of the

measured intensity as a function of altitude. In the first lidar systems, the detection and recording system comprised a camera and photographic film. Today detection and recording is achieved electronically. The detector is a device that converts light into an electrical signal and the recorder is an electronic device, often part involving a microcomputer, which processes and records this electrical signal.

Photomultiplier tubes (PMTs) are devices used as detectors for incoherent lidar systems working in the visible and ultraviolet. PMTs convert an incident photon into an electrical current pulse large enough to be detected by sensitive electronics. Other devices that are less commonly used as detectors in lidar systems include multianode PMTs, microchannel-plates, and avalanche photodiodes.

There are two ways in which the output of a PMT can be recorded electronically, the pulses can be counted individually (photon counting) or the average current due to the pulses can be measured and recorded (analogue recording). Which of these methods is the most appropriate depends on the rate at which the PMT produces output pulses, which is proportional to the intensity of the light incident on the PMT. If the average time between PMT output pulses is much less than the average pulse width, then individual pulses can be easily identified and photon counting is the more appropriate recording method. However, if the average time between PMT output pulses is close to, or greater than, the average pulse width, then it becomes impossible to distinguish overlapping pulses and so analogue recording becomes the more appropriate method. Some commercially available detection systems are able to simultaneously measure both analog and digital signals from a single PMT, increasing the dynamic range of the PMT by 100 compared to a digital-only counting system.

Coherent Detection

There is a class of lidar systems that determine wind speed by measuring the Doppler shift of backscattered light. There are two ways in which these measurements can be achieved; they are referred to as incoherent and coherent detection. Incoherent systems measure the wavelength of the transmitted and received light independently, using a spectrometer and determine the Doppler shift of the backscattered light from these two measurements. Coherent detection systems use a local oscillator, a narrow-band cw laser, to set the frequency of the transmitted pulses. The backscattered light is then mixed with light from the local oscillator on a photomixer. This arrangement results in the output of the photomixer being a radio-frequency (RF) signal whose frequency is the difference of the frequencies of the local oscillator and the backscattered light. Standard RF techniques are then used to measure and record this RF signal. The measured RF signal is used to determine the Doppler shift of the backscattered light and thus the wind speed.

The Lidar Equation

The lidar equation is used to determine the number of photons detected by a lidar system. The lidar equation takes into

account both instrumental parameters and geophysical variables. The general form of the lidar equation includes all forms of scattering and it can be used to calculate the signal strength for any lidar.

The number of photons detected as pulses at the photomultiplier output, per laser pulse is

$$A \int_{\Delta\lambda} PS(\lambda) \tau_t(\lambda) \tau_r(\lambda) Q(\lambda) \int_{R_1}^{R_2} \xi(r) \tau_a(r, \lambda)^2 \frac{1}{r^2} \times \sum_i \frac{d\sigma_i}{d\Omega}(\lambda) N_i(r) dr d\lambda \quad [1]$$

where,

A is the area of the telescope.

$PS(\lambda)$ is the convolution of $P(\lambda)$ and $S(\lambda)$, where $P(\lambda)$ is the number of photons emitted by the laser in a single laser pulse. $S(\lambda)$ is a function that takes into account any wavelength shift during scattering, including Doppler and Raman shifts. $\Delta\lambda$ is the wavelength range for which $PS(\lambda)$ is nonzero.

$\tau_t(\lambda)$ and $\tau_r(\lambda)$ are the optical transmission coefficients of the transmitter and receiver optics, respectively.

$Q(\lambda)$ is the quantum efficiency of the photomultiplier.

r is the range and R_1 and R_2 are the minimum and maximum ranges for a range bin.

$\xi(r)$ is the overlap factor that takes into account the intensity distribution across the laser beam and the physical overlap of the transmitted laser beam and the field of view of the receiver optics.

$\tau_a(r, \lambda)$ is the optical transmission of the atmosphere along the laser path.

$\frac{d\sigma_i}{d\Omega}(\lambda)$ is the backscatter cross section, for scattering of type i .

$N_i(r)$ is the number density of scattering centers, which cause scattering of type i .

The general form of the lidar equation, as expressed in eqn [1], can usually be greatly simplified when applied to a particular lidar system.

Rayleigh Lidar

Rayleigh lidar is the name given to the class of lidar systems that measure the intensity of light backscatter by molecules from altitudes between about 30 and 100 km. The intensity profiles measured by Rayleigh lidars are used to calculate relative density profiles, which are in turn used to calculate absolute temperature profiles. The terms Rayleigh scattering and molecular scattering are often used interchangeably, as are the terms Mie scattering and aerosol scattering. Rayleigh theory describes the scattering of light by molecules, small compared to the wavelength, and Mie theory describes scattering by aerosols, not small compared to the wavelength, so there is a strong connection between these two pairs of terms.

Rayleigh theory, named after its founder, Lord Rayleigh, describes the scattering of light by objects that are small compared to the wavelength of the incident radiation. Rayleigh scatter explains the color, intensity distribution, and polarization of the blue sky in terms of scattering by atmospheric

molecules. For objects with dimensions greater than about 0.003 times the incident wavelength, the more general Mie theory must be used to calculate scattering effects. Strictly speaking what is commonly referred to as Rayleigh scattering is in fact the sum of pure elastic scattering (the Cabannes line) plus the rotationally excited Raman lines which occur near the Cabannes line. Lidars exist which measure only the pure Cabannes line and the rotational lines (for aerosol properties and temperature respectively).

The Rayleigh backscatter ($\theta = \pi$) cross section for the atmosphere below 90 km can be expressed as

$$\frac{d\sigma_{\text{m}}(\theta = \pi)}{d\Omega} = \frac{C}{\lambda^4} \text{m}^2 \text{sr}^{-1} \quad [2]$$

where the value of C is between about 4.75×10^{-57} and 5.00×10^{-57} , depending on the value used for the index of refraction of air. Above 90 km altitude, the concentration of atomic oxygen becomes significant causing the refractive index of air to change, resulting in eqn [2] becoming less accurate with increasing altitude. The Rayleigh backscatter cross section eqn [2] can be used in conjunction with the lidar eqn [1] to determine the intensity of the backscatter that can be expected for a particular Rayleigh lidar system.

The Rayleigh lidar technique relies on the measured signal being proportional to the atmospheric density; this is not the case in any region that contains aerosols. From the surface to the top of the stratospheric aerosol layer about 25–30 km, the atmosphere contains a significant concentration of aerosols, thus the Rayleigh technique cannot be directly applied to this region. However, the atmosphere above this altitude contains very few aerosols allowing the application of the Rayleigh technique.

The principle of operation of a Rayleigh lidar system is quite simple. A pulse of laser light is fired up into the atmosphere, and any photons that are backscattered and collected by the receiving system are counted as a function of range. The lidar eqn [1] can be directly applied to a Rayleigh lidar system to calculate the expected signal strength. For Rayleigh lidar, a number of simplifications can be made to eqn [1] allowing it to be expressed as,

$$\text{Signal} = K \frac{1}{R^2} N_a(R) \delta R \quad [3]$$

where K is a constant that includes all constant terms from eqn [1], R is the range, δR is the length of a range bin, and $N_a(R)$ is the number density of air. Equation [3] shows that after correction for range, a Rayleigh lidar signal will be proportional to the atmospheric number density profile.

Due to the uncertainty in atmospheric transmission and instrumental parameters it is not possible to determine the value of the constant K in eqn [3] with sufficient precision to enable the determination of an absolute density profile. The measured relative density profile can be scaled to a model density profile to obtain a density profile that is well scaled.

Until recently the only method available to compute temperature from the relative density profile was prescribed by Hauchecorne and Chanin (1980). In this scheme, the relative density profile is integrated, using the hydrostatic equation, to

determine a relative pressure profile. This integration requires an initial or seed pressure, usually chosen from a model atmosphere, to initiate the integration at the maximum altitude of the density profile. The pressure profile calculated in this way has the same ratio to the actual pressure as the relative density profile has to the actual density, i.e., their scaling factors are the same. An absolute temperature profile can be calculated by applying the ideal gas law to the relative density and pressure profiles. The application of the ideal gas law divides the relative pressure by the relative density so that their scaling factors, which are the same, cancel out resulting in an absolute temperature profile.

The selection of the seed for the pressure integration introduces a systematic uncertainty into the temperatures retrieved at the greatest heights. The magnitude of this error is proportional to the difference between the actual pressure and the seed pressure used. As the actual pressure is not known, the resulting error in temperature is unknown. However, the magnitude of this error reduces as the calculation of temperature proceeds downward. **Figure 4(a)** shows differences in temperature profiles retrieved using seed pressures, which differ by ± 1 , 5, and 10% from some initial value. Users of this technique are well advised to ignore temperatures from at least the uppermost 10 km of the retrieval (and more safely 15 km), since the uncertainties introduced by the seed pressure estimate are not easily quantified.

Khanna et al. (2012) have developed a temperature retrieval, which relies on an inversion approach to obtaining

temperature. In this case the temperature profile is varied to find a 'best' match between the measured photocount profile and a photocount profile calculated using the lidar equation. In this method, the temperature is not required to be constant within a retrieval layer and the density profile can be integrated from the bottom to the top of the measurements. These improvements greatly decrease the systematic uncertainty in the seed pressure (or seed temperature). **Figure 4(b)** is in the same format as **Figure 4(a)**, but the temperature profiles are calculated using the inversion approach. The differences in temperature retrieved using an inversion approach introduces a much smaller uncertainty than the traditional method, and useful temperature can be retrieved over the entire measurement range significantly extending the upper altitude of retrieved temperature profiles.

Above about 90 km changes in composition of the atmosphere causes the Rayleigh backscatter cross section and the mean molecular mass to change with altitude. These changes lead to errors in the temperatures derived using the Rayleigh lidar technique. For the current generation of Rayleigh lidar systems there are other sources of error, statistical fluctuations, and seeding error, which are generally larger than those errors due to composition changes above 90 km. However, more powerful Rayleigh lidar systems may ultimately be limited in their maximum altitude extent by composition changes as has been shown by Argall (2007).

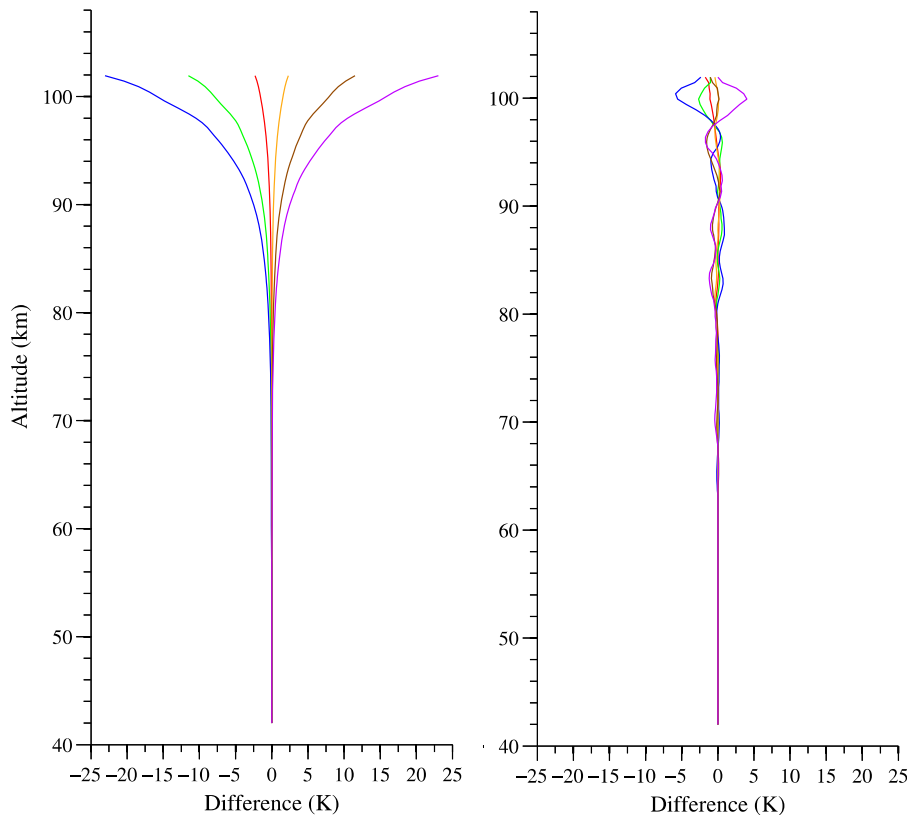


Figure 4 Effect of seed pressure choice on the retrieved temperature profile calculated from Rayleigh-scatter lidar measurements: left, the method of Hauchorne and Chanin and right, the inversion approach of Khanna et al. Left: as the initial seed pressure is varied from -10 to $+10\%$ the retrieved temperatures above 85 km are uncertain to a degree which exceeds the statistical uncertainty of the measurement. Right: using an inversion approach the retrieved temperatures are much less sensitive to the initial pressure guess and the retrieved temperature profile is useful over the entire range.

While even the most technically advanced, ground-based, middle-atmosphere lidar systems need clear skies to operate, the addition of Fabry–Perot etalons in the receiver allows daytime measurements. This daytime capability is technically complex and has been implemented on only very few Rayleigh lidar systems.

Doppler Effects

The motion of air molecules has components due to both random thermal motions and wind. When light is scattered by a molecule it suffers a change in frequency due to the Doppler effect. The magnitude and direction of the Doppler shift is determined by the component of the molecules velocity along the direction of the lidar beam.

The random thermal motions of air cause backscattered laser light to be spectrally broadened. Using Maxwell's velocity distribution function and the Doppler equation, it can be shown that the broadening function is Gaussian and has a temperature-dependent width.

Wind, the average motion of air molecules, causes back-scattered laser light to suffer a frequency shift while maintaining its shape. The frequency shift is directly proportional to the component of the wind velocity in the direction of scattering, the radial wind velocity. [Figure 5](#) shows how the spectrum of a narrow-bandwidth laser is modified due to scattering by atmospheric molecules.

Middle atmospheric winds can be determined by measuring the spectrum of backscattered light, however, Rayleigh–Doppler temperature measurements are quite difficult as the signal-to-noise requirements are much greater than those for wind velocity measurement using this technique.

Aerosol Lidar

The theory of scattering that was developed by Mie early this century is a general solution to the scattering of electromagnetic radiation by a homogeneous sphere. This early work has been extended to cover numerous other geometries and so provides a useful approximation for scattering from atmospheric aerosols.

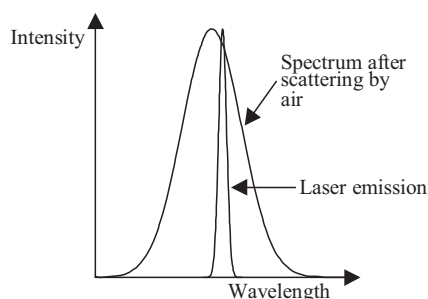


Figure 5 Doppler shift effects on Rayleigh scattering a narrow-line-width laser from atmospheric molecules. The broadening is due to thermal motion and the shift is due to wind. The intensities of the two spectra are not to scale.

The influence of clouds and aerosols on the atmospheric energy budget is complex as they scatter and absorb both incoming solar and outgoing terrestrial radiation. Since the early 1960s a large number of lidar systems have been operated at various stations around the world to study aerosols and clouds in the troposphere and lower stratosphere. Aerosols and clouds are easily detected by elastic backscatter lidar, however, instruments using multiple wavelength transmitters and receivers and polarization techniques provide significantly more information on their properties.

In September of 1994, NASA flew a space shuttle mission, STS-64, which included the LITE instrument, the first successful space-based lidar. LITE was used to measure tropospheric and stratospheric aerosols, clouds, and surface reflectance on a global scale. The next step in space-based lidar was the ICESat mission, which obtained measurements of aerosol and cloud properties from 2003 to 2010. Measurements of backscatter from 532 to 1064 nm were obtained (Schutz et al., 2005).

In June of 2006, the Cloud-Aerosol Lidar and Infrared Pathfinder Satellite Observation (CALIPSO) payload was launched. CALIPSO is still fully functional as of December 2012 and is routinely acquiring backscatter measurements at both 532 and 1064 nm. Routine data products include aerosol optical depth, backscatter, and extinction in addition to cloud height, thickness, optical depth, extinction backscatter, and properties of water and ice particles (Winker et al., 2007).

Similar measurements of cloud and aerosol properties to those on Earth were also made on Mars in 2008 as part of the NASA Phoenix mission. This highly successful mission found evidence for water on Mars. Whiteway et al. (2009) used the Phoenix aerosol lidar to measure fall streaks from clouds on Mars similar to fall streaks associated with precipitating ice particles on Earth.

Lidar systems can utilize the backscatter from aerosols to measure wind velocity. Light backscattered from aerosols undergoes the same Doppler shift due to wind as light scattered back from molecules. However, the spectral broadening of the light backscattered from aerosols is much narrower than that backscattered from molecules due to the difference in the masses of the two types of scatterers. The high signal level offered by scattering from aerosols in the lower atmosphere allows the use of coherent detection for the determination of wind velocity. Steerable lidars using this technique are capable of making high-resolution wind field maps.

Differential Absorption Lidar

The DIAL technique is used for measuring the concentration of trace species in the atmosphere. The DIAL technique relies on sharp variations in optical transmission near an absorption line of an atmospheric constituent. A DIAL transmits two closely spaced wavelengths, one coinciding with an absorption line of the constituent of interest and the other in the wing of this absorption line. During the transmission of the two wavelengths through the atmosphere the emission that is tuned to the absorption line will suffer greater attenuation than the emission in the wing of the absorption line. The intensities of the two wavelengths that are backscattered to the DIAL

instrument can then be used to determine the optical attenuation due to the constituent and thus the concentration of that constituent.

The DIAL technique has proven to be useful in providing tropospheric measurements with good time and spatial resolution for a number of trace species including NO, H₂O, O₃, SO₂, and CH₄, as well as stratospheric ozone measurements. The Network for Detection of Atmospheric Climate Change (NDACC), an international program, includes a global network of DIAL ozone lidars. DIAL also allows mapping and wide-area monitoring of industrial effluents and pollution.

Raman Lidar

When monochromatic light, or light of sufficiently narrow spectral width, is scattered by a molecular gas or liquid the spectrum of the scattered light can be observed to contain lines at wavelengths different from that of the incident radiation. This effect was first observed by Raman and it is due to the interaction of the radiation with the quantized vibrational and rotational energy levels of the molecule. Raman scattering involves a transfer of energy between the scattered light and the molecule; it is therefore an inelastic process. As the energy levels for each type of molecule are unique and so the Raman spectrum is unique and provides a method of sensing a particular molecular species.

The term Raman lidar is generally used to refer to a lidar that utilizes light scattered by molecules that undergo a change in their vibrational quantum number. Measurement of the intensity of the scattered Raman light allows the calculation of the abundance of the molecular species. The selection of vibrational Raman lines can be achieved with high-quality narrow-band interference filters. However, it is necessary to ensure that the blocking of such a filter is sufficiently high, so that elastic backscatter from molecules and aerosols is effectively attenuated. Due to the small cross sections for Raman scattering, Raman lidar is limited to molecules with a relatively high abundance, such as water vapor and molecular nitrogen, and due to its relative simplicity, is used in preference to DIAL.

Raman lidar is predominately used for the measurement of atmospheric water vapor and temperature. Raman molecular nitrogen profiles can be used to determine atmospheric temperature profiles, using the Rayleigh technique described above, even in regions containing aerosols. Elastic scattering from aerosols can be effectively separated from the Raman nitrogen backscatter by spectral filtering. The Raman nitrogen signal is therefore approximately proportional to the number density profile, although a correction for the optical attenuation of the atmosphere, due to both aerosols and molecules, must be made.

The pure rotational Raman spectrum (PRRS), which is due to scattering involving a change in the rotational quantum state only, is difficult to measure as the spectral shift of the lines is quite small so that separation of the elastic and rotational Raman lines is technically challenging. The separation of lines in the PRRS of N₂ is about 16 cm⁻¹ while the first vibrational transition causes a shift of about 2331 cm⁻¹. The shape of the PRRS is temperature dependent allowing pure rotational Raman lidar to make atmospheric temperature measurements.

Resonance-Fluorescence Lidar

The constant ablation of meteors in the Earth's upper atmosphere leads to the existence of extended layers of alkali metals at altitudes around 90 km. These metals have low abundance but very high resonant-scattering cross sections. Resonant-scattering occurs when the energy of an incident photon is equal to the energy of an allowed transition within an atom. In this elastic process, the atom absorbs a photon and instantly emits another photon at the same frequency. As resonant-scattering involves an atomic transition between allowed energy levels, the probability of this process occurring is much greater than that for Rayleigh scattering, leading to the much higher scattering cross sections. The resonant-scattering cross section for sodium at 589 nm is about 10¹⁵ times larger than the cross section for Rayleigh scattering by air at the same wavelength. As each species of alkali metal has a unique absorption and hence, resonant-scatter and fluorescence spectrum, these may be used to identify and measure the concentration of each individual species. Although most commonly applied to sodium, resonance-fluorescence lidar has been applied to calcium (Ca and Ca⁺), potassium, lithium, and iron.

Sodium lidar systems are used to measure abundance profiles of sodium between 85 and 105 km with time resolution of tens of seconds and altitude resolution of a few hundred meters. Density perturbations due to wave motions are present in the sodium density profiles, enabling the determination of wave parameters in this dynamically active region of the atmosphere to be determined. Spectral resolution of resonance-fluorescence scattering from sodium allows the determination of the temperature and wind. This technique, narrow-band resonance-fluorescence lidar, allows accurate, high resolution, temperature, and wind measurements in the mesopause region. These systems are currently being refined to use all solid-state lasers to generate the required sodium wavelengths and to make more robust systems capable for being used in harsh environments such as encountered on orbital platforms.

Iron Boltzmann lidar uses alexandrite lasers operating at 370 nm to probe the atmospheric iron layer, which exists between about 80 and 100 km. As alexandrite lasers are solid-state lasers and typically very reliable, this technique is particularly well suited to remote operation. Another advantage of this technique is that the wavelength used is in a spectral region where little solar radiation reaches the ground; this allows daytime operation without the need for complex spectral filtering apparatus to reduce the background light levels.

Lidars measuring stimulated emission from atmospheric potassium have also been successfully used to measure atmospheric dynamics and temperatures in a manner similar to sodium (see the review by Chu and Papan, 2005).

See also: **Aerosols:** Aerosol–Cloud Interactions and Their Radiative Forcing; Climatology of Stratospheric Aerosols; Climatology of Tropospheric Aerosols; Observations and Measurements; Role in Climate Change; Role in Radiative Transfer. **Aviation Meteorology:** Clear Air Turbulence.

Biogeochemical Cycles: Iodine. **Boundary Layer (Atmospheric) and Air Pollution:** Observational Techniques *In Situ*; Observational Techniques: Remote. **Chemistry of the Atmosphere:** Observations for Chemistry (Remote Sensing): Lidar. **Clouds and Fog:** Cloud Modeling; Measurement Techniques *In Situ*; Noctilucent Clouds; Stratus and Stratocumulus. **Dynamical Meteorology:** Acoustic Waves. **Electricity in the Atmosphere:** Ions in the Atmosphere. **Global Change:** Climate Record: Surface Temperature Trends; Sea Level Change; Upper Atmospheric Change. **Gravity Waves:** Overview. **Lidar:** Backscatter; Differential Absorption Lidar; Doppler; Raman; Resonance. **Mesosphere:** Atomic Species in the Mesopause Region; Metal Layers; Polar Summer Mesopause. **Middle Atmosphere:** Planetary Waves; Polar Vortex; Stratospheric Sudden Warmings; Zonal Mean Climatology. **Numerical Models:** Parameterization of Physical Processes: Clouds; Parameterization of Physical Processes: Gravity Wave Fluxes; Parameterization of Physical Processes: Turbulence and Mixing. **Optics, Atmospheric:** Airglow Instrumentation; Optical Remote Sensing Instruments. **Ozone Depletion and Related Topics:** Long-Term Ozone Changes. **Radar:** Cloud Radar; Mesosphere–Stratosphere–Troposphere and Stratosphere–Troposphere Radars and Wind Profilers; Meteor Radar; Synthetic Aperture Radar (Land Surface Applications). **Radiation Transfer in the Atmosphere:** Scattering. **Satellites and Satellite Remote Sensing:** Aerosol Measurements; Measuring Ozone from Space: TOMS and SBUV; Remote Sensing: Cloud Properties; Temperature Soundings; Water Vapor. **Solar System/Sun, Atmospheres, Evolution of Atmospheres:** Meteors; Planetary Atmospheres: Mars. **Stratospheric Chemistry Topics:** Stratospheric Water Vapor. **Tropospheric Chemistry and Composition:** Cloud Chemistry.

Further Reading

- Argall, P.S., 2007. Upper altitude limit for Rayleigh lidar. *Annals of Geophysics (Germany)* 25 (1), 19–25.
- Chu, X., Papen, G., 2005. Resonance fluorescence lidar for measurements of the middle and upper atmosphere. In: Fukuchi, T., Fujii, T. (Eds.), *Optical Science and Engineering*. CRC Press, Boca Raton, pp. 179–432.
- Frehlich, R., 1996. Coherent Doppler lidar measurements of winds. In: Consortini, A. (Ed.), *Trends in Optics: Research, Development and Applications*. Academic Press, London, England, pp. 351–370.
- Grant, W.B., 1995. Lidar for atmospheric and hydrospheric studies. In: Duarte, J.F. (Ed.), *Tunable Laser Applications*. Marcel Dekker, New York, pp. 213–305.
- Hauchecorne, A., Chanin, M., 1980. Density and temperature profiles obtained by lidar between 35 and 70 km. *Geophysical Research Letters* 7, 565–568.
- Khanna, J., Bando, J., Sica, R.J., McElroy, C.T., 2012. New technique for retrieval of atmospheric temperature profiles from Rayleigh-scatter lidar measurements using nonlinear inversion. *Applied Optics* 51 (33), 7945–7952. <http://dx.doi.org/10.1364/AO.51.007945>.
- Killinger, D.M., Mooradian, A. (Eds.), 1983. *Optical and Laser Remote Sensing*. Springer-Verlag, New York.
- Measures, R.M., 1984. *Laser Remote Sensing: Fundamentals and Applications*. John Wiley & Sons, New York.
- Schreiber, U., Werner, C. (Eds.), 1999. *Laser Radar Ranging and Atmospheric Lidar Techniques II (Europto Series)*. Society of Photo-optical Instrumentation Engineers.
- Schutz, B.E., Zwally, H.J., Shuman, C.A., Hancock, D., DiMarzio, J.P., 2005. Overview of the ICESat Mission. *Geophysical Research Letters* 32, L21S01. <http://dx.doi.org/10.1029/2005GL024009>.
- Sedlacek, A.J., Fischer, K.W. (Eds.), 1999. *Application of Lidar to Current Atmospheric Topics III (Proceedings of Spie, 3757)*. Society of Photo-optical Instrumentation Engineers.
- Singh, U.N., 1997. In: Rastogi, R.K. (Ed.), *Optical Measurement Techniques and Application*. Artech House, Norwood, MA, pp. 369–396.
- Thomas, L., 1995. In: Clark, R.J.H., Hester, R.E. (Eds.), *Spectroscopy in Environmental Science*. John Wiley & Sons, Chichester, England, pp. 1–47.
- Weitkamp, C., 1996. Lidar measurements: atmospheric constituents, clouds, and ground reflectance. In: Raschke, E. (Ed.), *Radiation and Water in the Climate System*. Springer-Verlag, Berlin, Germany, pp. 217–247.
- Whiteway, J.A., et al., 2009. *Science* 325, 68. <http://dx.doi.org/10.1126/science.1172344>.
- Winker, D.M., Hunt, W.H., McGill, M.J., 2007. Initial performance assessment of CALIOP. *Geophysical Research Letters* 34, L19803. <http://dx.doi.org/10.1029/2007GL030135>.

Backscatter

CMR Platt, Colorado State University, Fort Collins, CO, USA

RL Collins, University of Alaska Fairbanks, Fairbanks, AK, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by CMR Platt, volume 3, pp 1176–1183, © 2003, Elsevier Ltd.

Synopsis

This article summarizes the principles and practices of the backscatter lidar techniques used to measure clouds and aerosols in the atmosphere. Scattering phase functions of molecules and particles are presented and polarization techniques are reviewed. The lidar equation is presented with a simple solution that highlights the importance of the backscatter phase function and the use of the extinction-to-backscatter ratio. The technique is illustrated with examples of lidar measurements of stratospheric aerosols, cirrus clouds, and in midlevel ice, water, and mixed-phase clouds. Finally contemporary global measurements of clouds and aerosols by the Cloud Aerosol Lidar with Orthogonal Polarization (CALIOP) lidar aboard the Cloud-Aerosol Lidar and Infrared Pathfinder Satellite Observations (CALIPSO) are described.

Introduction

This article presents several aspects of lidar backscatter. After a general introduction, various definitions pertaining to lidar backscatter are described. This is followed by an explanation of the scattering and backscattering phase functions, including presentation of typical scattering phase functions of molecules, water drops, and clouds. The polarized nature of the radiation and its treatment by a scattering matrix are then described briefly. The lidar equation is presented, together with a simple solution in order to point out the importance of the backscatter phase function and its relation to the extinction-to-backscatter ratio and its use in solving the equation. Examples of extinction-to-backscatter ratios of various atmospheric constituents are presented. Several examples of profiles of measured atmospheric backscatter are described, including stratospheric aerosols, cirrus clouds, and depolarizing effects in midlevel ice, water, and mixed-phase clouds. The article does not cover inelastic backscatter such as Raman scattering and fluorescence.

While current backscatter lidars employ lasers, the first backscatter predates the invention of the laser and employed searchlights. Lidar is used to detect and profile certain constituents in the atmosphere, such as molecules, aerosols, and clouds. The backscatter from such entities is important in lidar because most lidar (laser radar) systems are monostatic, that is, there is a telescope receiver placed close to, or coaxial with, a laser pulse transmitter. Pulses of light sent into the atmosphere are scattered in all directions by molecules, aerosols, and clouds, and a small amount scattered into the back direction is returned to the receiver. The time taken for the laser pulse to return gives the range of the atmospheric volume being studied, and the amplitude of the return is proportional to the volume density of the atmospheric particles or molecules. The total amount scattered by a particle is dependent on the diameter of the particle and its size compared with the wavelength of light. For example, particles that are small compared with the wavelength scatter less than if the scatter were determined solely by the particle's geometric cross section. The ratio of the scattering cross section to the geometric cross section is termed the scattering efficiency and used to quantify the total scattering. The amount scattered in any direction forms a pattern that is described by the single scattering phase

function $P(\theta)$, where θ is the scattering angle that is the angle between the scattered light and the forward direction. The geometry is shown in Figure 1. Backscatter corresponds to $\theta = 180^\circ$, while forward scatter corresponds to $\theta = 0^\circ$.

Backscatter Cross Section, Efficiency, and Coefficient

The scattering cross section $C_{sc}(\lambda, r)$ (m^2) of an atmospheric particle (molecule, aerosol, water drop, or ice crystal) determines how much radiation is scattered in all directions by the particle. Here λ is wavelength and r is the particle dimension. Consider a uniform light beam of irradiance I (W m^{-2}) incident on a particle of area of geometric cross section A (m^2). The particle scatters an amount of power W_{sc} (W) into all directions. The scattering cross section is defined as

$$C_{sc}(\lambda, r) = \frac{W_{sc}}{I}. \quad [1]$$

The scattering efficiency, $Q_{sc}(\lambda, r)$, is defined as the ratio of the scattering cross section to the geometric cross section,

$$Q_{sc}(\lambda, r) = \frac{C_{sc}(\lambda, r)}{A}. \quad [2]$$

Because of the nature of electromagnetic scattering, $Q_{sc}(\lambda, r)$ can approach a value of 2 for nonabsorbing particles that are large compared with the wavelength λ . This is because diffraction occurs around and outside the edges of the particle, causing the scattering cross section to be about $2A$ for large particles. For particles much smaller than the wavelength, the scattering is less efficient and the cross section is much less than $1A$.

The backscatter cross section $C_{\Pi}(\lambda, r)$ ($\text{m}^2 \text{sr}^{-1}$) determines the amount of power scattered into the backscattering direction per unit solid angle. The backscatter efficiency $Q_{\Pi}(\lambda, r)$ is again the ratio of the scattering cross section to the geometric cross section. The backscatter phase function, as well as the total scattering efficiency, determines the backscatter efficiency, as shown under 'scattering phase function' below.

When lidar is observing the atmosphere, it is observing the molecular atmosphere into which are mixed aerosols and clouds in different proportions and varying in both time and space. Of course, aerosols and clouds – particularly clouds – tend to form

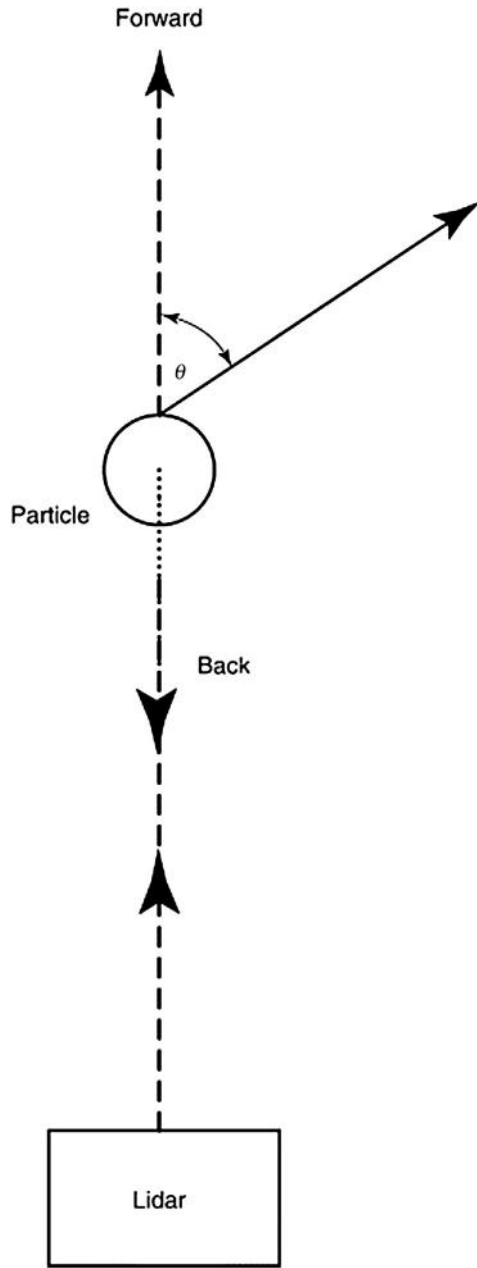


Figure 1 Schematic of backscattering at an atmospheric particle.

in well-defined layers, whereas gas molecules are distributed throughout the atmosphere, the molecular density depending on the local pressure and temperature. Aerosols and clouds generally contain a range of particle sizes with different number densities, defining a smooth particle number size distribution. The backscatter efficiency also varies with particle size. Lidar measures the returns within a volume of the atmosphere and the volume backscatter coefficient $\beta(\lambda)$ ($\text{m}^{-1} \text{sr}^{-1}$) is defined as

$$\beta(\lambda) = \int_0^\infty Q_\pi(\lambda, r) n(r) \pi r^2 dr, \quad [3]$$

where $n(r)$ is the number of particles of dimension r , and in the case illustrated above the particles are spheres (a water cloud, for example) of radius r . For ice crystals, such as hexagons, the

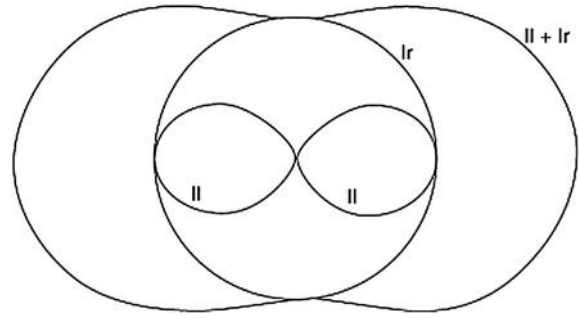


Figure 2 The scattering phase function for scattering from a molecule – Rayleigh scattering. Labels are described in the text. Adapted with permission from Twomey, S., 1977. *Atmospheric Aerosols*. Developments in Atmospheric Science, vol. 7. Elsevier, Amsterdam.

dimensions and area depend on the crystal type or habit and orientation. Irregular particles are often modeled as equivalent spheres. If the ice crystals are tumbling then an effective dimension has to be determined.

Scattering Phase Function

The phase function $P(\theta)$ defines the intensity of the radiation scattered by a particle at an angle θ to the forward direction. The total scattered radiation normalized to 4π is then

$$\int_0^{180} P(\theta) \cos \theta \sin \theta d\theta = 4\pi. \quad [4]$$

The value of the backscatter phase function, together with the scattering efficiency, will determine the backscatter efficiency. This can be either larger or smaller than the isotropic backscatter efficiency of 0.0796, depending on whether the scattering phase function is peaked or depressed near the backscattering direction.

We now consider some phase functions for a range of atmospheric constituents and the corresponding backscatter phase function and efficiency. The single scattering phase function for an atmospheric molecule is shown in Figure 2. As noted above, the phase function denotes the fraction of the total radiation that is scattered into a given angle to the incident direction. Note that the scattering differs according to the polarization of the radiation. For scattered radiation in a plane containing the direction of polarization and the vector of the incident radiation, denoted by l , the phase function falls to 0 at a scattering angle of 90° . For the radiation perpendicular to the plane containing the incident vector, denoted by r , scattering is independent of direction. The scattering direction in Figure 2 is depicted in a polar diagram where incident radiation is entering from the left of the diagram. The normalized scattering phase function for unpolarized light is given by

$$P(\theta) = \frac{3}{16\pi} (1 + \cos^2 \theta). \quad [5]$$

The backscatter phase function $P(\pi)$ is then $6/16\pi$ (0.119 sr^{-1}).

For particles of molecular size, or of dimensions small compared with the wavelength, Rayleigh scattering theory is appropriate. For particle dimensions approaching, or larger

than, the wavelength of the incident radiation, Mie scattering theory is appropriate. Figure 3 shows the polar scattering diagram for a water sphere, as might be found in a water cloud. The diagram shows a succession of high and low values of scatter, caused by complex interference effects between penetrating and surface waves on the sphere. The quantity r is the radius of the sphere and x is the size parameter, $2\pi r/\lambda$. The backscatter phase function, which is not normalized, is seen to be small compared with the forward scattering. This is a characteristic of scattering on large particles, with some exceptions, as we will see. The phase function depicted here is the sum of the two directions of polarization (see Polarization Effects section).

Figure 4 shows typical scattering phase function for a hexagonal ice crystal that has large dimensions relative to the wavelength. The presentation is different, with scattering angle on the x -axis and phase function on the y -axis. The scattering peak at 22° that gives rise to the halo is clearly visible. Again one can see that the backscatter phase function is quite low, and much lower than that from a molecule. Of course, the cross sections of ice clouds are much greater per unit volume than for molecules, so that backscatter intensity from ice clouds tends to be greater, although this is not always the case. The crystal phase function is calculated either from a complex geometrical theory or by geometrical ray tracing.

What is relevant here to lidar backscatter is the large differences in the backscatter phase function between the various atmospheric components in Figures 2–4. These differences can be exploited in lidar systems as the phase function varies with size parameter. Lidar systems with multiple wavelengths can be used to estimate the size of particles based on the relative differences in backscatter at each wavelength.

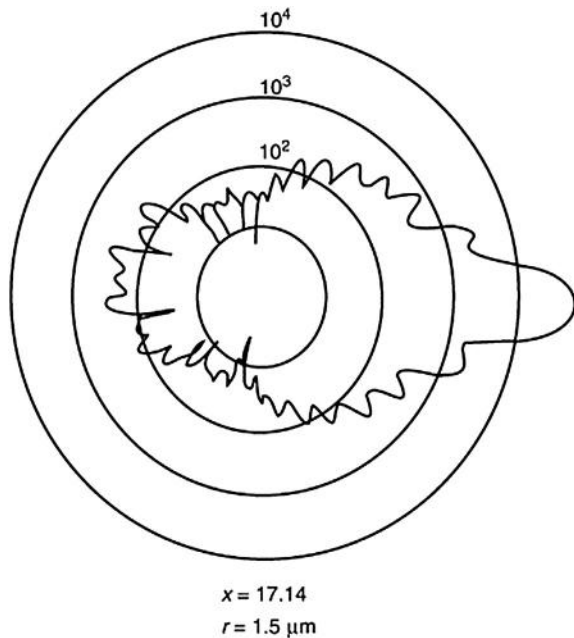


Figure 3 Scattering phase function for a water drop. The size parameter and water droplet size are shown below. Adapted with permission from Twomey, S., 1977. *Atmospheric Aerosols*. Developments in Atmospheric Science, vol. 7. Elsevier, Amsterdam.

The backscatter-to-extinction ratio k (sr^{-1}) is

$$k = \frac{\beta}{\sigma} = \frac{Q_\pi}{Q_{sc}} = P(\pi), \quad [6]$$

where β ($\text{m}^{-1} \text{sr}^{-1}$) is the backscatter coefficient, σ (m^{-1}) is the volume extinction coefficient, Q_{sc} is the effective scattering efficiency, and Q_π is the backscatter efficiency of the volume of particles. The backscatter-to-extinction ratio is numerically equal to the normalized backscatter phase function. The value of k is needed to solve the backscatter lidar equation for volume attenuation of the lidar beam and is therefore a crucial quantity to lidar backscatter measurements. A popular quantity is S , defined as the lidar ratio, which is simply the reciprocal of the backscatter-to-extinction ratio and therefore is measured in steradians. Typical ranges of values of S and k are shown in Table 1. Note that aerosol populations also possess a range of values of these quantities.

Polarization Effects

Thus far we have considered the total radiation backscattered by a particle without consideration of whether the light is polarized or not. In fact, pulsed lasers used in lidar are usually linearly polarized to a high degree. The backscattered light can have significantly different polarization based on the scattering properties of the particles and their orientation. A complete analysis of the polarization effects can be conducted using Stokes parameters. However, nearly all lidar systems employ linear depolarization effects where the backscattered light is split into two channels. In one channel, the light passes through a polarizer parallel to the polarization of the transmitted light and in the other channel, it passes through a polarizer perpendicular to the polarization of the transmitted light. The degree of depolarization is given by the amount of signal in the perpendicular channel and used to discriminate between different constituents. Molecules and water spheres have only very weak depolarization due to their electronic properties. Ice crystals and aerosols, however, are found to depolarize by various amounts.

Scattering of radiation is conveniently described by a scattering matrix, whose components can be measured by optical instruments. Here, we consider a backscattering matrix for nonspherical particles as a general example (Sassen, 2000):

$$F(180) = \text{diag}[F_{11}(180), F_{22}(180), F_{33}(180), F_{44}(180)], \quad [7]$$

This matrix is simplified as it applies to particles that are randomly oriented in space and which contain a plane of symmetry, such as typical hexagonal crystals. The medium is called macroscopically isotropic and symmetric. If there is horizontal orientation of ice crystals, which can occur, then more scattering elements in the matrix need to be considered. The present treatment represents the present state of progress in the field.

In terms of the above matrix a depolarization ratio Δ , defined as the ratio of the perpendicular to parallel components, is given by

$$\Delta = \frac{F_{11}(180) - F_{22}(180)}{F_{11}(180) + F_{22}(180)}. \quad [8]$$

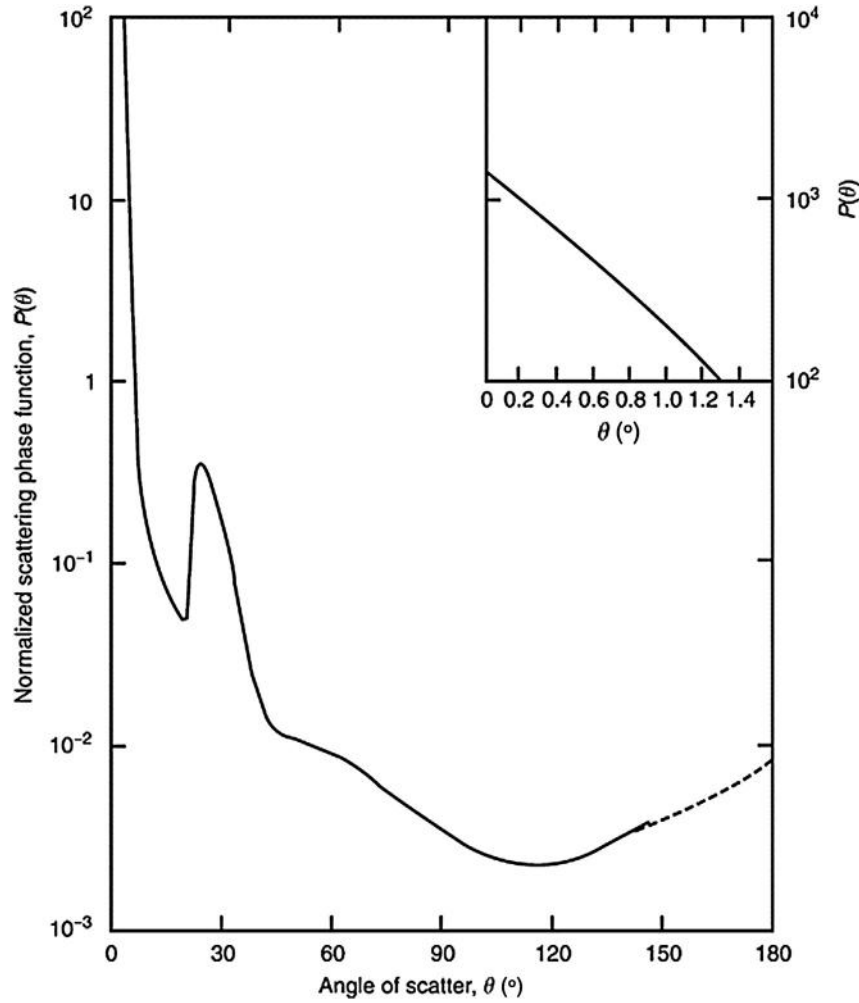


Figure 4 Scattering phase function for an ice crystal; schematic representation from a number of experimental phase functions. Adapted with permission from Platt, C.M.R., 1981. Remote Sounding of High Clouds. III: Monte Carlo Calculations of Multiple-Scattered Lidar Returns. *Journal of Atmospheric Sciences* 38, 156–167.

Table 1 Range of values of backscatter parameters k and S for various atmospheric scatterers

Atmospheric constituent	k (sr^{-1})	S (sr)
Molecular	0.119	8.4
Water cloud	0.047–0.050	20.0–21.3
Ice cloud theoretical	0.008–0.072	14–125
Ice cloud experimental	0.011–0.036	28–91
Aerosol	0.015–0.05	20–60

Now, for spheres, $F_{11}(180) = F_{22}(180)$, so that $\Delta = 0$ is a well-known result for this case, although there is some residual depolarization.

Hexagonal crystals depolarize because the laser radiation undergoes several internal reflections before ending up in the back direction, so that (except for specular reflections at perpendicular surfaces) some rotation of the plane of the polarization vector occurs. Measured values of depolarization ratio are quite large in cirrus clouds, which allows some distinction in cloud phase using a depolarization lidar.

The use of such a lidar in such observations is thus very desirable.

The Lidar Equation

Lidar is used to measure the properties of atmospheric layers through their backscattering and depolarization properties. As the layers scatter the laser beam, then they also attenuate the beam during passage through that layer. Thus we must understand the retrieval of the backscatter coefficient of a particular layer. Measured backscatter must be corrected for this attenuation. It is apparent then as to how the extinction-to-backscatter ratio S of a volume of particles is important to the retrieval process. The lidar equation can be written in the first place in terms of the power $P(r)$ measured from a particular range r :

$$P(r) = \frac{EAc}{r^2} (\beta_{pa} + \beta_{pe}) \exp \left\{ -2 \int_{r_1}^{r_2} \sigma(r) dr \right\}, \quad [9]$$

where E is the pulse energy, A the telescope area, and c the velocity of light. The backscatter coefficient β is the sum of

the two polarization components in the parenthesis and $\sigma(r)$ is the volume extinction coefficient of the layer within ranges r_1 and r_2 . The power is calibrated against that from a molecular layer at a different range. The lidar (eqn [9]) can then be written:

$$\beta'(r) = \beta(r)T_{12}^2, \quad [10]$$

where

$$\begin{aligned} T_{12}^2 &= \exp - 2 \int_{r_1}^{r_2} \sigma(r) dr \\ &= \exp - 2S \int_{r_1}^{r_2} \beta(r) dr, \end{aligned} \quad [11]$$

and we consider $\beta(r)$ as the sum of the two polarization components.

The solution for $\beta(r)$ in eqns [10] and [11] is then

$$\beta(r) = \frac{\beta'(r)}{1 - 2S \int_{r_1}^{r_2} \beta'(r) dr}. \quad [12]$$

The crucial importance of the quantity S in the solution is clear.

The solution is generally more complex, as there is the molecular atmosphere to consider. The above is an example of how a volume backscatter coefficient can be recovered if we know the extinction-to-backscatter ratio S .

Table 1 shows the values of k and S for various atmospheric constituents from both experimental and theoretical data.

Backscatter lidar has also been used at infrared wavelengths where the absorption by both water and ice can be quite strong. In that case, the scattering efficiency and backscatter efficiency will be correspondingly less for transparent particles. The same is true for some aerosols, such as those composed of soot where strong absorption of visible radiation can occur. Values of S will be correspondingly larger than for the transparent-particle case.

Measurements of Atmospheric Backscatter

Early lidar observations were made of the stratospheric aerosol, which responds to large volcanic eruptions. Such eruptions send a cloud of enhanced aerosol mass around the globe that is easily observed by lidar. Series of observations now exist covering 30 years that show how peak backscatter in the Junge layer has varied over the years and how it responds to volcanic activity. An example is shown in Figure 5. The total backscatter is a mixture of aerosols and molecules. A model of molecular backscatter from local meteorological information is shown, and the enhanced backscatter in the stratosphere is indicated clearly. The aerosol backscatter amplitude waxes and wanes in strength over the years as various volcanic clouds appear and then dissipate.

The atmospheric boundary layer also contains copious amounts of aerosol as a capping temperature inversion impedes upward movement out of the layer. Such aerosol layers can be dense in large urban areas, in regions of desert dust, and from forest fires, as examples.

Cirrus was also an early target of investigation because of the semitransparent nature of such clouds. The Sun's disk is often seen hazily through even quite deep layers of cirrus. Lidar pulses of radiation can often penetrate through such layers with sufficient photons returning to the receiver to be detectable.

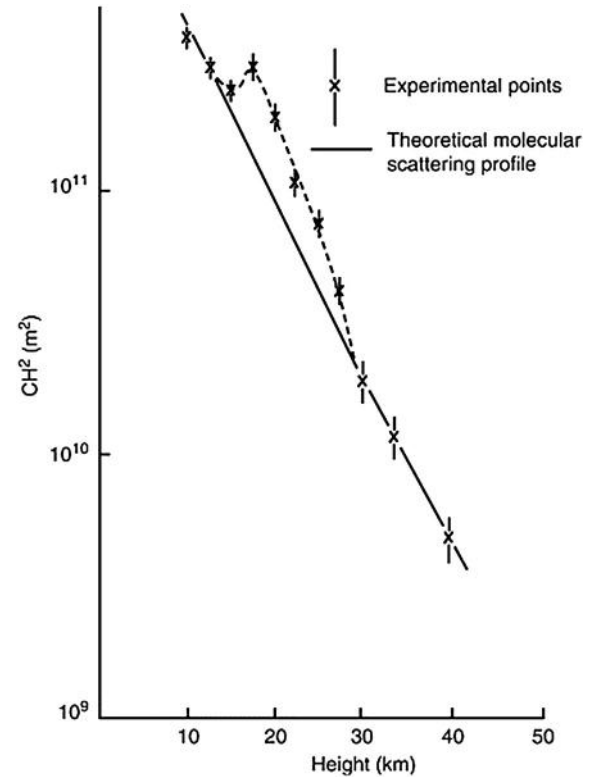


Figure 5 Typical backscatter profile (in units of range squared) for the stratospheric aerosol layer. The backscatter for a model molecular atmosphere is also shown. Adapted with permission from Kent, G.S., Clemesha, B.R., Wright, R.W., February 1967. High altitude atmospheric scattering of light from a laser beam, *Journal of Atmospheric and Solar-Terrestrial Physics* 29 (2), 169–181.

Figure 6 shows a typical return from cirrus, the depth being rather typical for such clouds. Attenuation is fairly weak for this cloud. The figure also reveals typical structure with variations in backscatter with altitude. The Mt. Pinatubo volcanic cloud was strong at the time this profile was taken and shows through the cirrus. When several profiles taken at successive time intervals are investigated, cirrus ice crystals falling out and being swept sideways by the wind are revealed. There is often also an indication of cirrus cloud base becoming progressively lower with time.

Examples of linear depolarization ratio are shown in Figure 7. Here, layers of midlevel cloud exhibit very variable characteristics. This is because layers of ice crystals, water drops, and mixed-phase cloud can exist separately in the atmosphere. The lidar backscatter is shown as the full line and the depolarization ratio Δ as the broken line. The bottom layer just above 4 km has a value of Δ that is typical of ice clouds. The next layer has intense backscatter but very low values of depolarization. This is an interesting case that is fairly common at these altitudes and temperatures. Between atmospheric temperatures of -10 and -20 °C, ice crystals are often hexagonal plates that float horizontally through the air. They thus present a large area of cross section but also mirrorlike surfaces, which do not depolarize in the backward direction. If the lidar is tilted a few degrees off the horizontal, such returns disappear rapidly, indicating the close angle to horizontal at which the crystals fall. Such

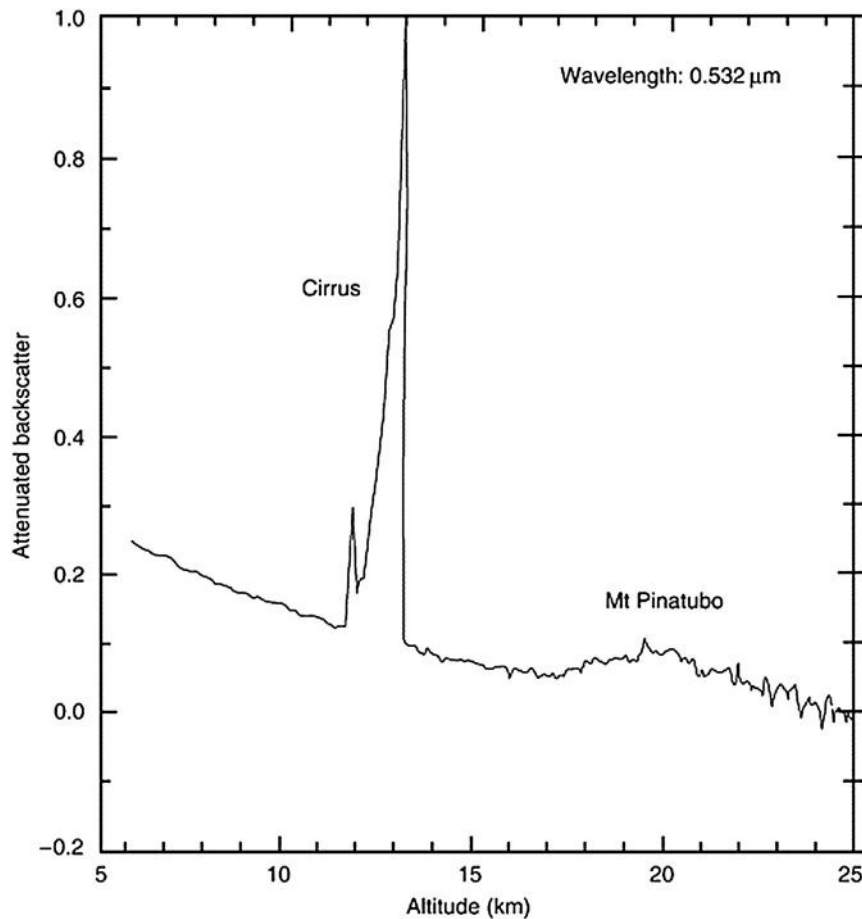


Figure 6 Typical backscatter profile from a cirrus cloud, showing the strong backscatter and atmospheric attenuation. Adapted with permission from Platt, C.M.R., Young, S.A., Manson, P.J., Patterson, G.R., Marsden, S.C., Austin, R.T., Churnside, J.H., 1998. The Optical Properties of Equatorial Cirrus from Observations in the ARM Pilot Radiation Observation Experiment. *Journal of Atmospheric Sciences* 55, 1977–1996.

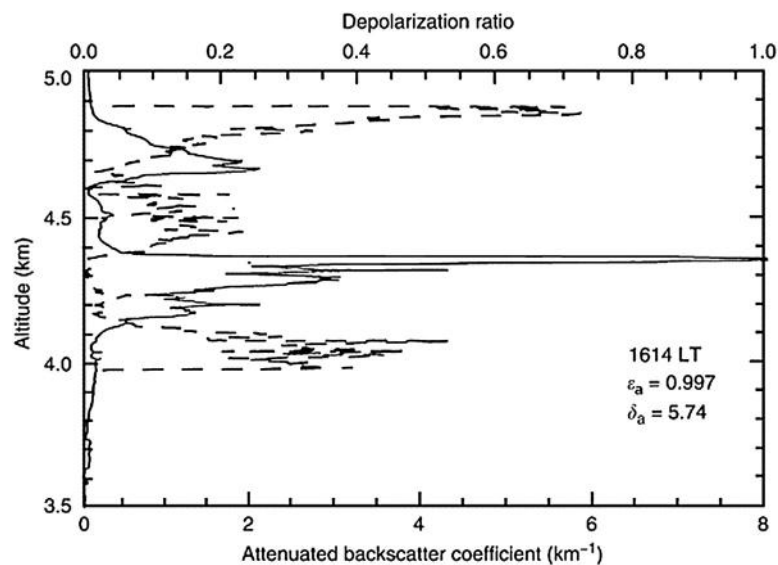


Figure 7 Examples of returns from a mixed-phase cloud, showing various patterns of backscatter (full line) and depolarization ratio (broken line). Adapted with permission from Young, S.A., Platt, C.M.R., Austin, R.T., Patterson, G.R., 2000. Optical Properties and Phase of Some Midlatitude, Midlevel Clouds in ECLIPS. *Journal of Applied Meteorology* 39, 135–153.

unusual and dramatic returns from the atmosphere show the strength of lidar backscatter in picking out details in cloud layers. The layer at about 4.5 km is probably composed of ice, with some supercooled water drops possibly also present. The top layer is interesting because there appears to be strong attenuation and a pattern of depolarization ratio that commences at the cloud base with a low value but increases steadily upwards. This happens to be the characteristic of a layer of water drops. Because these are relatively small and numerous, attenuation is quite strong. The depolarization pattern is a consequence of strong multiple scattering in the beam, where the scattered photons near the back direction, or even at larger angles, can be scattered several times more, thus finding their way back into the telescope receiver beam. This process can, in a manner equivalent to crystal internal reflections, rotate the plane of polarization to give depolarization in the back direction.

Global measurements of aerosols and clouds have entered a new phase with the launch of the CALIPSO satellite mission in 2006. The CALIPSO satellite carries CALIOP, a nadir viewing two-wavelength, polarization sensitive lidar. The CALIOP wavelength and polarization capabilities allow discrimination between aerosol and cloud particles just as they have been employed in earlier ground-based lidar systems. However, being satellite-borne, CALIOP measurements allow scientists to explore upper- and middle-level clouds that are often shielded from surface-based observers by low-level clouds. Furthermore, the CALIOP measurements are made as part of a suite of instruments aboard CALIPSO and companion satellites that includes imagers, radiometers, and radars. Meteorological studies seek to understand the roles of clouds and aerosols as well as the interactions between them. The CALIOP measurements have mapped aerosol layers above clouds where the heating effects of aerosols are enhanced by the underlying clouds. These aerosol layers were previously unobserved and are now being incorporated into assessments of aerosol radiative forcing. CALIOP measurements have also provided a more accurate detection of clouds than earlier imaging satellites, as the lidar detects both upper- and lower-level clouds. The lidar measurements have revealed significantly higher occurrence of stratus decks over the Pacific and Atlantic Oceans than previously reported. Furthermore, the lidar has the sensitivity to detect thin cirrus clouds that have been previously unreported. These thin cirrus clouds have significant impact on the radiative budget of the tropical upper troposphere and are understood to influence troposphere–stratosphere exchange.

The above examples show the power of lidar backscatter to distinguish various layers of aerosols and clouds in the

atmosphere and understand their meteorological impacts. Lidar backscatter has established itself as a key remote-sensing technique in understanding the role of aerosols and clouds in the Earth's climate system.

See also: **Aerosols:** Aerosol Physics and Chemistry; Aerosol–Cloud Interactions and Their Radiative Forcing; Climatology of Tropospheric Aerosols; Observations and Measurements; Role in Radiative Transfer. **Chemistry of the Atmosphere:** Observations for Chemistry (Remote Sensing): Lidar. **Clouds and Fog:** Cloud Microphysics. **Lidar:** Atmospheric Sounding Introduction; Differential Absorption Lidar; Doppler; Raman; Resonance. **Optics, Atmospheric:** Airglow Instrumentation; Optical Remote Sensing Instruments. **Radiation Transfer in the Atmosphere:** Cloud–Radiative Processes.

Further Reading

- Bohren, C.F., Huffman, D.R., 1983. *Absorption and Scattering of Light by Small Particles*. Wiley, New York.
- Deirmendjian, D., 1969. *Electromagnetic Scattering on Spherical Polydispersions*. Elsevier, New York.
- Gimmestad, G.G., 2008. Reexamination of depolarization in lidar measurements. *Applied Optics* 47, 3795–3802. <http://dx.doi.org/10.1364/AO.47.003795>.
- Grant, W.B., Browell, E.V., Menzies, R.T., Sassen, K., She, C.Y. (Eds.), 1997. *Selected Papers on Laser Applications in Remote Sensing*. SPIE Milestone Series, vol. 141. SPIE, Bellingham, WA, p. 662.
- Lynch, D.E., Sassen, K., Starr, D.O., Stephens, G.L. (Eds.), 2002. *Cirrus*. Oxford University Press, Oxford.
- Measures, R.M., 1984. *Laser Remote Sensing. Fundamentals and Applications*. Wiley, New York.
- Mishchenko, M.I., Hovenier, J.W., Travis, L.D. (Eds.), 2000. *Light Scattering by Nonspherical Particles. Theory, Measurements, and Applications*. Academic Press, London.
- Sassen, K., 2000. The lidar backscatter depolarization technique for cloud and aerosol research. In: Mishchenko, M.I., Hovenier, J.W., Travis, L.D. (Eds.), *Light Scattering by Nonspherical Particles: Theory, Measurements, and Geophysical Applications* chap. 14. Academic, San Diego, CA, pp. 393–416.
- Sassen, K., 2005. Meteorology: dusty ice clouds over Alaska. *Nature* 434, 456. <http://dx.doi.org/10.1038/434456a>.
- Twomey, S., 1977. *Atmospheric Aerosols*. In: *Developments in Atmospheric Science*, vol. 7. Elsevier, Amsterdam.
- Van de Hulst, H.C., 1957. *Light Scattering by Small Particles*. Wiley, New York.
- Weikamp, C. (Ed.), 2005. *Lidar: Range-Resolved Optical Remote Sensing of the Atmosphere*. Springer Series in Optical Sciences, vol. 102, pp. 307–323.
- Winker, C.D.M., Pelon, J., Coakley Jr., J.A., et al., 2010. The CALIPSO mission: a global 3D view of aerosols and clouds. *Bulletin of the American Meteorological Society* 91, 1211–1229. <http://dx.doi.org/10.1175/2010BAMS3009.1>.

Differential Absorption Lidar

S Ismail, Science Directorate, NASA Langley Research Center, Hampton, VA, USA

EV Browell, STARS II Affiliate, NASA Langley Research Center, Hampton, VA, USA

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

Differential absorption lidar (DIAL) is a laser remote-sensing technique that is used for range-resolved (profile) measurements of atmospheric gas concentrations. Technological advancements in DIAL systems have greatly increased the measurement capabilities of ground-based and airborne DIAL systems for measurements of tropospheric and stratospheric O_3 and tropospheric H_2O , along with pollution measurements of many other gases in range-resolved and column measurements. The DIAL technique has also been proposed for measurements of H_2O and O_3 from space, and a variation of the DIAL technique called integrated path DIAL (IPDA) is being explored for global, high-precision column measurements of CO_2 and CH_4 . All of these topics are discussed in this article.

Introduction

Differential absorption lidar (DIAL) is a laser remote-sensing technique that is used for range-resolved (profile) measurements of atmospheric gas concentrations. This technique was first applied in 1966 for remote measurements of water vapor (H_2O), and since then it has been used to measure other naturally occurring atmospheric gases such as ozone (O_3) and many pollutant gases, such as sulfur dioxide (SO_2), nitrogen dioxide (NO_2), ammonia (NH_3), mercury (Hg), carbon monoxide (CO), carbon dioxide (CO_2), and hydrocarbons. While the initial DIAL technique development focused on H_2O , the main thrust of the DIAL applications in the 1970s and early 1980s was on pollution monitoring. The first airborne measurements with DIAL were aimed at studying tropospheric O_3 in large-scale pollution studies over the East Coast of the United States in 1980. Subsequently, airborne DIAL measurements of H_2O were demonstrated in 1982. Technological advancements in airborne DIAL systems have greatly increased the measurement capabilities of ground-based and airborne DIAL systems for measurements of tropospheric and stratospheric O_3 and tropospheric H_2O , along with pollution measurements of many other gases in range-resolved and column measurements. Even the possibility of DIAL profile measurements of carbon dioxide (CO_2) and methane (CH_4) are being investigated. The DIAL technique has also been proposed for measurements of H_2O and O_3 from space, and a variation of the DIAL technique, called integrated path differential absorption lidar (IPDA), is being explored for global, high-precision column measurements of CO_2 and CH_4 . All of these topics are discussed in this article.

DIAL Technique

DIAL is a remote-sensing technique that uses two lidar returns to determine the distribution of a selected gas along the direction of the lidar beams. A simplified version of the DIAL concept is shown in Figure 1. The molecules and aerosols in the atmosphere provide the backscattering media for the laser light via Rayleigh and Mie scattering, respectively. Two laser wavelengths are employed, one tuned to a strong absorption feature

of the gas of interest, generally called the 'on' wavelength (λ_{on}), and the other tuned to a nearby wavelength with weak absorption by the gas, generally called the 'off' wavelength (λ_{off}). The value of the average gas concentration N_A in the range interval from R_1 to R_2 can be determined from the ratio of the backscattered lidar signals at λ_{on} and λ_{off} , as shown in Figure 1. In that equation, $\Delta\sigma = \sigma_{on} - \sigma_{off}$ is the difference between the absorption cross-sections at the on and off wavelengths, and $Pr_{on}(R_1)$ and $Pr_{off}(R_2)$ are the signal powers received from range R at the on and off wavelengths, respectively. This equation is essentially an application of the familiar Beer-Lambert law for an absorbing medium. The λ_{off} lidar return also provides important information on the molecular and aerosol-scattering properties of the atmosphere, and this contributes greatly to the science interpretation of the gas profile measurement.

Some of the key considerations for being able to use the DIAL technique for measuring range-resolved gas concentrations include:

1. having a tunable, pulsed laser source that can generate sufficient pulse energies at the DIAL wavelengths on and near a suitable absorption feature of a gas of interest, with the λ_{on} and σ_{on} optimized so that the integrated absorption by the gas at the maximum measurement range has a one-way optical depth of about one. This ensures an optimum compromise between having a large $\Delta\sigma$ and having enough signal from the most distant range to be able to make a DIAL measurement;
2. keeping the DIAL wavelengths as close together as possible in order to minimize measurement errors that result from differences in the atmospheric scattering and attenuation at these two wavelengths. When the laser wavelength separation is unavoidably large, a correction to account for wavelength differences in atmospheric scattering and attenuation must be applied;
3. selecting a DIAL wavelength region where the atmospheric scattering is sufficient to provide adequate backscattered signals from the atmosphere. Molecular scattering drops off as λ^{-4} and is very weak for wavelengths longer than about $1\ \mu m$. Thus, for longer wavelengths (typically longer than $1\ \mu m$), aerosols, clouds, or some sort of surface has to provide the backscattered signal. For species that are

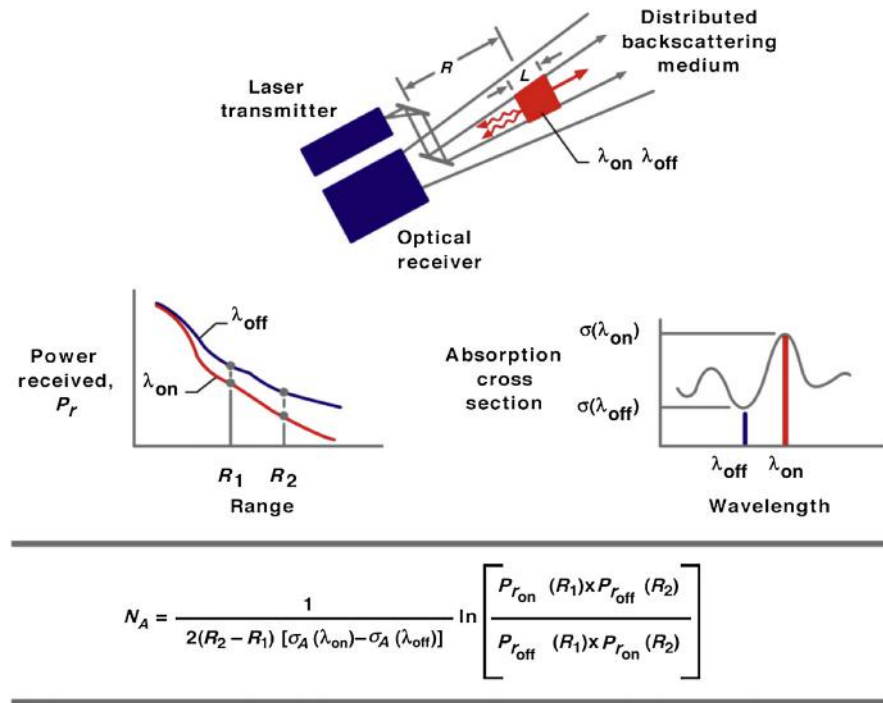


Figure 1 The differential absorption lidar (DIAL) concept.

- measured in the thermal infrared spectral region (3–12 μm), gas profile measurements are generally constrained to the planetary boundary layer (mixing layer), although long-path measurements can be made using topographic targets;
4. minimizing unknown absorption interference from other gas species;
 5. using lasers with short pulse lengths compared to the desired measurement range resolution. For example, a 100 ns laser pulse length has a 15 m folded scattering length; and
 6. using detectors with the necessary sensitivity and low noise level for making the DIAL measurements.

Three types of lasers are being used in the development of new DIAL systems in the ultraviolet (UV) to infrared (IR) spectral region: rare-earth solid-state lasers that have a limited (<100 nm) tuning range, custom-designed optical parametric oscillators, and fiber lasers. In some instances, the laser frequency is doubled or tripled to achieve wavelengths in the UV and visible (VIS) spectral regions. For narrow-band DIAL systems operating on single absorption lines in absorption bands in the IR, the spectral position and width are controlled by locking the laser frequency of a seed laser with respect to an absorption feature and by injection seeding the laser transmitter. DIAL detectors of choice include photomultiplier tubes (PMTs) in the UV to mid-VIS region, avalanche photodiodes (Si:APDs) in the mid-VIS to about 1 μm region, and InGaAs:APD in the 1.0–1.7 μm region. New mercury-cadmium-telluride (HgCdTe) detectors are being developed for use in the 1–4.5 μm region. Detector noise corrupts measurements in long-wavelength (>2 μm) regions, and in these cases, heterodyne detection methods are employed to overcome detector noise.

Note that the DIAL technique generally performs better in the nadir direction than in the zenith because molecular density and, in general, aerosol scattering decrease with increasing altitude, so that the backscattered lidar signal falls off more rapidly in the zenith than in the nadir. Thus, the DIAL technique is well suited for operation from airborne platforms, especially for relatively well-mixed gases such as O_3 , H_2O , and CO_2 . Zenith measurements from aircraft platforms at wavelengths greater than 1 μm are in general more difficult because of rapid reduction in molecular and aerosol backscattering at higher altitudes. Airborne DIAL systems have demonstrated measurement accuracies better than 5% in the troposphere for H_2O and in the troposphere and lower stratosphere for O_3 . To achieve these accuracies, data from these DIAL systems are generally averaged over 200–500 m in vertical distances and over 10–70 km in horizontal distances. Airborne DIAL systems have the ability to measure gas concentrations over long ranges (~1000 km) and map their distribution over regional scales for application to studies of atmospheric processes. Aerosol-scattering profile information is simultaneously obtained from these DIAL systems with much higher spatial resolutions (~30 m in the vertical and <200 m in the horizontal). Atmospheric CO_2 measurements for climate studies require an accuracy of <0.3%. Rapid progress has been made recently to demonstrate this capability from airborne lidar systems.

In designing and operating DIAL systems, one must be aware of a number of atmospheric and instrumental effects that can cause bias in the measurements, including scattering and extinction differences between on and off wavelengths, sensitivity of absorption and scattering to atmospheric temperature and pressure, interference due to other absorbing species, laser spectral characteristics, background radiation, detection system

noise, and so on. A well-designed and optimally operated DIAL system reduces these errors to a manageable, if not insignificant, level. In some cases, these biases are minimized or removed during data processing.

Use of Topographic Targets

When the DIAL measurement does not have sufficient atmospheric backscatter or does not have range resolution for a gas profile measurement, or there is a requirement for a very high-precision measurement, such as for CO_2 and O_2 , a topographic target may be employed to provide the backscattered laser radiation. This DIAL measurement results in a long-path or column measurement of the gas, which is known as an IPDA measurement. Low-pulse-energy high-repetition-rate laser systems and intensity-modulated (IM) continuous-wave (CW) laser systems often need to use the IPDA method. There are several issues that need to be addressed when using topographic targets, including (1) the need for accurate range information that can be achieved by proper design of the pulsed or pulse-encoded laser systems and during their associated data retrieval methods; (2) rapid variations in the height of topographic targets that can be minimized either by simultaneous measurements of on and off signals from these targets or by randomizing these effects by collecting a large number of samples; (3) the introduction of bias in IPDA measurements due to very sharp spectral changes in the reflection features of the materials contained in the topographic targets, which can be reduced by selecting closely spaced on and off wavelengths; and (4) interferences in IPDA measurement by intervening clouds and aerosols that need to be compensated for, particularly when using IM-CW measurement techniques. Another factor to consider is that unless the target is moving, such as from an aircraft or satellite, or is being scanned, the measurement accuracy will not necessarily increase rapidly with the number of pulses averaged.

Application Areas

The primary applications of the DIAL technique have been in the areas of O_3 and H_2O measurements. These gases are of great importance in such areas as atmospheric chemistry, radiation, health, and weather, and they are discussed in detail in this article. A number of other gases have been or can be studied using the DIAL technique. Although H_2O was the first gas measured with the DIAL technique (via the temperature-tuned ruby laser lidar system in 1966), NO_2 was the first pollutant gas measured using this technique in the early 1970s. NO_2 is the only gas of interest with a strong absorption band in the VIS spectral region, with features in the blue region. However, since NO_2 occurs in low concentrations in most situations, appearing at high concentrations only in major pollutant plumes, there has not been significant activity in measuring it using DIAL. SO_2 has been measured using dye lasers operating in the UV spectral region near 300 nm, where it has a strongly modulated absorption spectrum. The interest in SO_2 was primarily related to emissions from power plants, but there has also been some interest in SO_2 emitted from volcanoes. Hg is another gas of interest. Hg has a strong absorption line near 254.3 nm, and it is emitted from

a number of sources, including chlorine-alkali plants, geothermal fields, volcanoes, cinnabar mining areas, and coal-burning power plants. Hydrocarbons have seen some interest as well. The absorption bands vary from near 300 nm for benzene to 3.4 μm for methane and 10.5 μm for ethylene. Other hydrocarbons can also be measured in the 3.2–3.7 μm and 9.3–10.7 μm spectral regions. Hydrocarbons have various urban and industrial sources, and it can often be cost-effective to use laser systems to detect hydrocarbon gas leaks in industrial plants so they can be eliminated. NH_3 is commonly found near feedlots and other agricultural operations, and it has absorption lines in the 9.3–10.7 μm spectral region, which can be accessed by CO_2 laser transmitters operating in that region.

There is also increasing interest in the measurement of CO_2 due to its importance in global climate change and the carbon cycle. The sources and sinks of atmospheric CO_2 are not well understood on a global scale, and there is a need to map the large-scale sources and sinks of CO_2 , preferably from space. There are candidate absorption lines in the 1.6 and 2 μm spectral regions, and fiber-based and solid-state laser technology is being developed in these regions for possible space-based IPDA systems. CO_2 is difficult to measure in part because it is a long-lived gas and has a relatively high 'background' concentration, which does not vary much around the Earth. As a result, the measurement accuracy and precision must be extremely high for meaningful measurements. Also, it is important to measure the atmospheric mixing ratio of CO_2 , and thus it will be important to measure atmospheric number density, via an IPDA O_2 column measurement, to convert CO_2 concentrations to CO_2 mixing ratios.

Airborne DIAL Systems and Applications

Determining variations of O_3 and H_2O over large geographic regions is important to our understanding of a broad range of atmospheric processes. For example, measurements of O_3 and H_2O distributions can lead to an improved understanding of the relative role of transport versus photochemistry in the tropospheric O_3 budget. O_3 and H_2O are important radiatively and contribute to the radiation budget and climate change. H_2O is influential in many different meteorological processes and in the transport of energy on large scales. Better knowledge of upper tropospheric and lower stratospheric water vapor is needed to understand the dynamics and chemical processes in this region and for climate studies. For a better understanding of the atmosphere, it is important to study the spatial and temporal variations of these gases over many regions of the Earth and ultimately be able to make measurements of them from space. Lidar systems have the potential to provide the high vertical resolution needed in these studies. DIAL measurements of O_3 and H_2O were initially demonstrated from the ground. Ground-based O_3 DIAL systems are being used to monitor air quality and to study stratospheric changes, and ground-based water vapor DIAL systems are being used in meteorological studies. Airborne DIAL systems evolved from ground-based DIAL systems and are used in many field experiments to study a number of atmospheric chemical, dynamical, transport, and meteorological processes.

Airborne DIAL systems are also considered precursors to the development of space-based DIAL systems for making global-scale measurements.

Airborne O₃ Measurements

The first airborne DIAL system was flown for O₃ and aerosol investigations, and it was performed in conjunction with a large-scale pollution field experiment conducted over the East Coast of the United States in the summer of 1980. This initial system evolved into the advanced UV DIAL system that has been used in many airborne field experiments over the last two decades. A schematic of the UV DIAL system is shown in [Figure 2](#). This system uses four frequency-doubled Nd:YAG lasers to pump sequentially two dye lasers that are frequency-doubled into the UV to produce on- and off-line wavelength pairs of 289 and 300 nm or 301 and 310 nm for DIAL O₃ measurements in the troposphere or stratosphere, respectively. The parameters of the O₃ DIAL system are given in [Table 1](#). The residual 1064 and 590 nm beams from the frequency-doubling processes of the Nd:YAG and dye lasers, respectively, are also transmitted for aerosol and cloud measurements. This system has a demonstrated absolute accuracy for O₃ measurements of better than 5% or 2 ppbv (parts per billion by volume), whichever is larger, and a measurement precision of 2% or 1 ppbv with a vertical resolution of 300 m and an averaging time of 5 min (a horizontal resolution of about 70 km at typical aircraft ground speeds). An example of the O₃ measurements made with this system is shown in [Figure 3](#). This figure shows many different aspects of O₃ processes that occur from the tropics to high latitudes, including photochemical loss and production, vertical and horizontal transport, and stratosphere–troposphere exchange.

The NASA Langley airborne UV DIAL systems have made significant contributions to the understanding of O₃ in both the troposphere and stratosphere. One of the key features of these measurements has been the characterization of the air masses using simultaneous O₃ and aerosol properties. This

characterization has led to the identification of photochemical, meteorological, dynamical, and transport processes in the atmosphere. These DIAL systems have been used in 26 international and 3 national field experiments during the past 30 years; during these field experiments, measurements were made over, or near, all of the oceans and continents of the world. A list of field experiments and base locations for these airborne field experiments is given in [Table 2](#). Data derived from these field experiments are publicly available from NASA's Langley Research Center.

Airborne H₂O Measurements

The first DIAL measurements of H₂O with a continuously tunable laser were demonstrated in the late 1970s. In an initial step toward the development of a space-based H₂O DIAL system, the first airborne H₂O DIAL system was developed and demonstrated in 1982. This system was based on Nd:YAG-pumped dye laser technology, and it was used in the first airborne H₂O DIAL investigation of the marine boundary layer over the Gulf Stream. This laser was later replaced with a flashlamp-pumped solid-state alexandrite laser, which had high spectral purity (>99% of laser energy contained within a narrow 1 pm spectral region), and this system was used to make accurate H₂O profile measurements across the lower troposphere under a variety of atmospheric conditions. A third H₂O DIAL system called Lidar Atmospheric Sensing Experiment (LASE) was developed as a prototype for a space-based H₂O DIAL system. This system was designed to operate autonomously from a high-altitude ER-2 aircraft, and it uses a Ti:sapphire laser and 1–3 different H₂O absorption cross-sections to make H₂O measurements across the entire troposphere. While the LASE system was initially designed for operation from NASA's high-altitude ER-2 aircraft, it was also modified to fly on conventional medium-altitude P-3 and DC-8 aircraft.

The Ti:sapphire laser in LASE is pumped by a double-pulsed, frequency-doubled Nd:YAG to produce laser pulses

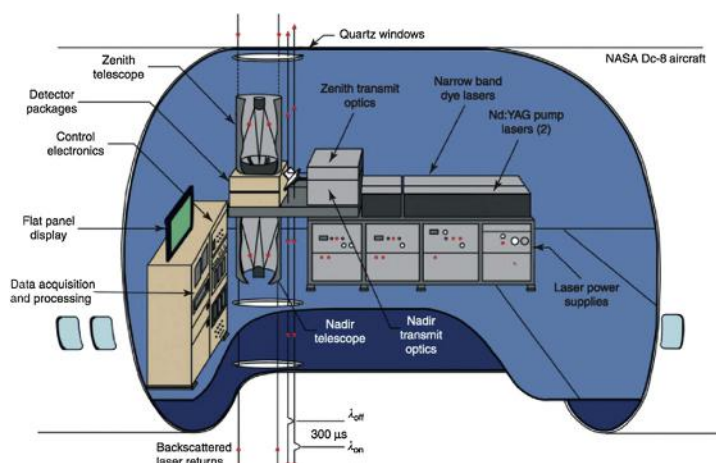
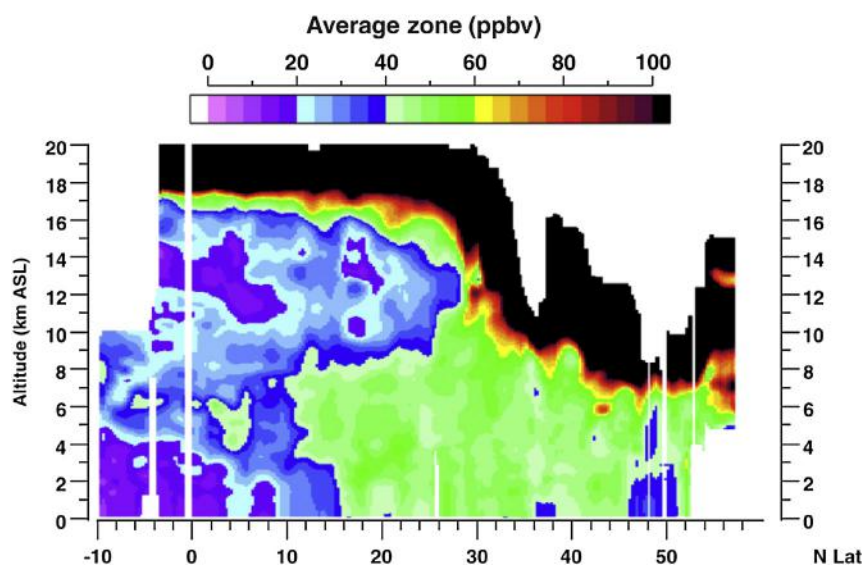


Figure 2 Configuration of the NASA Langley airborne UV DIAL system is shown on the left panel. Four beams are simultaneously transmitted in the nadir and zenith directions for measurements of O₃ profiles with DIAL wavelengths of 289 and 300 nm and for aerosol and cloud profiles at 600 and 1064 nm. The right panel shows a photograph of the DIAL system as configured on the DC-8 aircraft.

Table 1 Characteristics of an airborne DIAL system

Transmitter	
Pump lasers (4)	Big Sky CFR 800
Pulse separation, μ s	300
Pulse energy at 532, mJ	400
Pulse length, ns	7
Repetition rate, Hz (pulse pairs interleaved)	40
Transmitted laser energy at 1064 nm, mJ	200
Dye lasers (2 lasers)	Continuum ND-6000
Dye laser output energy, fundamental, each laser, mJ	105
Doubled fundamental (UV laser energy), each laser, mJ	25 on and off
UV laser line width, pm	<40
Transmitted UV laser energy, each direction, mJ	13 on and off
Transmitted visible energy, each direction, mJ	70

Receiver	Wavelength regions		
	289–300 nm	578–600 nm	1064 nm
Efficiency to detector, %	31	40	31
Detector quantum efficiency, %	21 (PMT)	8 (PMT)	40 (APD)
Total receiver efficiency, ^a %	6.5	3.2	12.4
Receiver field-of-view (selectable), mrad	<1.5	≤ 1.5	≤ 1.5
Physical parameters			
Total weight (lbs) (includes laser structure, data rack, and support equipment)	2521		
Dimensions of lasers and laser support structure (L $\times$ W $\times$ H) (telescope included in length but not height)	78" $\times$ 40" $\times$ 43"		
Total system power requirements (kW)	10		

^aIncludes filter transmission for daytime operation.**Figure 3** Average latitudinal O₃ distribution measured by the airborne UV DIAL system over the western Pacific Ocean during the 1994 Western Pacific Exploratory Mission (PEM-West B).

in the 815 nm absorption band of H₂O. The wavelength of the Ti:sapphire laser is controlled by injection seeding with a diode laser that is frequency-locked to an H₂O line using an absorption cell. Each pulse pair consists of an on-line and off-line wavelength for the H₂O DIAL measurements. Parameters of the LASE system are shown in Table 3. Operation of LASE

on the side of the H₂O absorption line permits the use of the same H₂O line for different absorption cross-sections, and as a result, the DIAL measurements of H₂O can cover four orders of magnitude of dynamic range. The accuracy of LASE H₂O profile measurements was determined to be better than 6% or 0.01 g kg⁻¹, whichever is larger, over the full dynamic range of

Table 2 Airborne lidar investigations of O₃ and aerosols during major field experiments

Year	Lidar experiment	Base of operation
1980	Persistent Elevated Pollution Episode (PEPE)—US Environmental Protection Agency (EPA)	Wallops, Virginia
1981	North American Plume Study-1	San Juan, Puerto Rico
1981	Cloud Transport Study	Wallops, Virginia
1982	North American Plume Study – II	Bermuda
1984	Tropopause Fold Experiment (TFE)	Las Vegas, Nevada
1984	Atlantic Boundary Layer Experiment (ABLE-1)	Barbados
1985	Amazon Boundary Layer Experiment – Dry Season (ABLE-2A)	Manaus, Brazil
1987	Amazon Boundary Layer Experiment – Wet Season (ABLE-2B)	Manaus, Brazil
1987	Airborne Antarctic Ozone Experiment (AAOE)	Punta Arenas, Chile
1988	Arctic Boundary Layer Experiment (ABLE-3A)	Barrow and Bethel, Alaska
1989	Airborne Arctic Stratospheric Expedition (AASE-1)	Stavanger, Norway
1990	Arctic Boundary Layer Experiment (ABLE-3B)	Hudson Bay, Canada
1991	Pacific Exploratory Mission – West A (PEM-West-A)	Guam, Hong Kong, and Tokyo
1992	Airborne Arctic Stratospheric Expedition (AASE-11)	Stavanger, Norway
1992	Transport and Atmospheric Chemistry Near Equator (TRACE-A)	Brazil, Namibia, and South Africa
1994	Pacific Exploratory Mission – West B (PEM-West-B)	Guam, Hong Kong, and Tokyo
1995–96	Tropical/Vortex Ozone Transport Experiment (TOTE/VOTE)	Alaska, Hawaii, and Iceland
1996	Pacific Exploratory Mission – Tropics (PEM-Tropics-A)	Fiji, New Zealand, and Tahiti
1997	SASS Ozone and NO _x Experiment (SONEX)	North Atlantic Flight Corridor
1999	Pacific Exploratory Mission – Tropics (PEM-Tropics-B)	Fiji, Tahiti, and Easter Island
1999–2000	SAGE III Ozone Loss and Validation Experiment (SOLVE)	Kiruna, Sweden
2000	Tropospheric Ozone Production about the Spring Equinox (TOPSE)	Churchill, Canada, and Greenland
2001	Transport and Chemical Evolution over the Pacific (TRACE-P)	Guam, Hong Kong, Tokyo, and Hawaii
2002	SAGE III Ozone Loss and Validation Experiment-11 (SOLVE-11)	Kiruna, Sweden
2004	Intercontinental Chemical Transport Experiment – Phase A (INTEX-A)	Mid-America, Missouri, and Pease, NH
2005	Polar Aura Validation Experiment (PAVE)	Pease, NH
2006	Intercontinental Chemical Transport Experiment – Phase B (INTEX-B)	Houston, TX, Honolulu, HI, and Anchorage, AK
2007	Tropical Composition, Cloud, and Climate Coupling (TC4) Field Experiment	San Jose, Costa Rica
2008	Arctic Research of the Composition of the Troposphere from Aircraft and Satellites (ARCTAS)	Fairbanks, AK, and Cold Lake, AB

H₂O concentrations in the troposphere. LASE measurements of water vapor, aerosol, and cloud distributions have been used in many atmospheric-process and hurricane model forecast studies. An example of LASE measurements of water vapor in the vicinity of Hurricane Erin on 10 September 2001 is shown in Figure 4. LASE operated onboard the DC-8

aircraft at an altitude of 9 km, and the flight track of the DC-8 is shown in the left panel in Figure 4. Elevated and highly variable water vapor distributions, shown in the right panel of Figure 4, were observed up to and above the aircraft. These data and data from other NASA hurricane field experiments have been used at Florida State University to conduct hurricane model forecast studies. These studies have shown that inclusion of LASE water vapor measurements improved model forecasts of hurricane track and intensity. LASE has participated in 12 major airborne field experiments since 1995. A list of field experiments and their base locations is given in Table 4. During the last decade, the German Aerospace Center (DLR) has developed an airborne DIAL system operating in a strong H₂O band near 940 nm. This system has been flown on a Falcon 20 aircraft to measure upper tropospheric and lower stratospheric moisture profiles. The 940 nm band has much stronger H₂O absorption lines than the 820 nm band and is more suitable for moisture profiling in the upper troposphere and lower stratosphere, where the concentrations of H₂O are very low and variable. Measurements of water vapor in the upper troposphere and lower stratosphere are needed in climate studies. An example of upper tropospheric and lower stratospheric water vapor measurements by the DLR DIAL system is shown in Figure 5. This example shows variations of water vapor mixing ratios

Table 3 Parameters of the LASE H₂O DIAL system

Transmitter	
Laser: double-pulsed Ti:sapphire developed at NASA Langley	
Wavelength region (nm)	813–818
Pulse energy (mJ)	100–150
Pulse-pair repetition frequency (Hz)	5
Line width (pm)	<0.25
Stability (pm)	<0.35
Spectral purity (%)	>99
Beam divergence (mrad)	<0.4
Pulse width (ns)	35
Spectral separation of H ₂ O DIAL λs (pm)	≤70
Time separation of H ₂ O DIAL λs (μs)	300
Receiver	
Area (m ²)	0.11
Optical efficiency to detector (%)	50 (night); 35 (day)

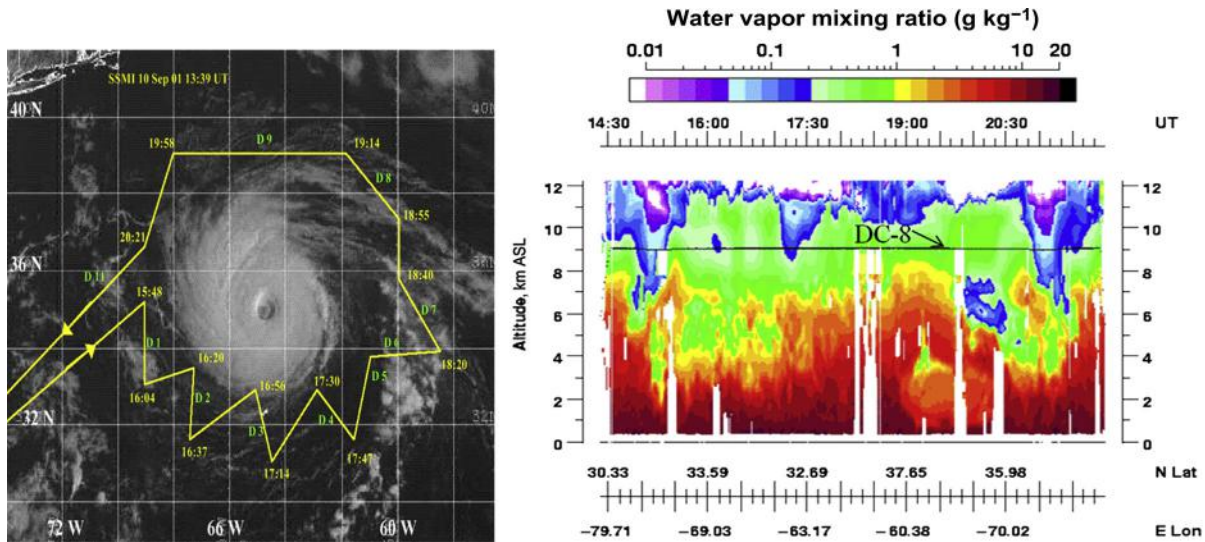


Figure 4 Flight track of the DC-8 overlaid on a satellite image showing distributions of clouds associated with Hurricane Erin on 10 September 2010 over the Atlantic Ocean. The right panel shows the distribution water vapor mixing ratios measured by LASE along the flight track. The mass mixing ratio unit g kg^{-1} represents the mass of water in grams divided by the mass of air in kilograms in an enclosure. The black line shows the altitude of the DC-8, and blank regions indicate an absence of data.

Table 4 LASE investigation of H_2O , aerosol, and clouds during major airborne field experiments

Year	LASE field experiments	Aircraft	Base of operation
1995	LASE Validation Experiment	ER-2	Wallops Island, VA
1996	Tropospheric Aerosol Radiative Forcing Observation Experiment (TARFOX)	ER-2	Wallops Island, VA
1997	Southern Great Plains Experiment (SGP97)	P-3	Oklahoma City, OK
1998	Convection and Moisture Experiment-3 (CAMEX-3)	DC-8	Cocoa Beach, FL
1999	GTE Pacific Exploratory Mission in Tropics-B (PEM Tropics-B)	DC-8	Hawaii, Fiji, and Tahiti
1999–2000	SAGE III Ozone Loss and Validation Experiment (SOLVE)	DC-8	Kiruna, Sweden
2000	ARM/FIRE Water Vapor Experiment (AFWEX)	DC-8	Oklahoma City, OK
2001	Convection and Moisture Experiment-4 (CAMEX-4)	DC-8	Jacksonville NAS, FL
2002	International H_2O Project (IHOP)	DC-8	Oklahoma City, OK
2006	NASA African Monsoon Multidisciplinary Analyses (NAMMA)	DC-8	Cape Verdi, Africa
2007	Tropical Composition, Cloud, and Climate Coupling (TC4)	DC-8	San Juan, Costa Rica
2010	Geneses and Rapid Intensification Processes (GRIP)	DC-8	Fort Lauderdale, FL, and St. Croix, VI

across a tropopause fold region where there is a potential for exchange of air between the stratosphere and troposphere.

Airborne Column CO_2 Measurements

The first high-precision CO_2 column measurements with the IPDA method were demonstrated from an aircraft in 2008. These measurements were made by a multifrequency, single-beam laser absorption spectrometer (LAS) operating in the vicinity of a CO_2 absorption line in the $1.57 \mu\text{m}$ region. The LAS system uses an IM-CW laser transmitter with synchronous detection to produce high signal-to-noise ratio (SNR) measurements of CO_2 columns between the aircraft and the surface or tops of clouds. In addition, a pseudorandom-noise (PRN) transmitter and receiver are part of the overall system

to measure the range from aircraft to the backscattering surface. This multifrequency fiber laser lidar (MFLL) was built by ITT Exelis Geospatial Systems as a prototype of a laser system to be flown in a future space mission for global CO_2 measurements.

The MFLL transmitter produces three simultaneous laser wavelengths with different IM frequencies (one on the CO_2 absorption line and two off the line on opposite sides) with a combined average power of 5 W. Since all LAS beams are also transmitted simultaneously, they have 100% spatial and temporal overlap, eliminating sensitivity to regions of highly varying surface reflectance and minimizing effects of atmospheric turbulence by making them common mode terms. The PRN altimeter also has an average power of 5 W. The reflected signals from the surface or cloud tops for both the LAS

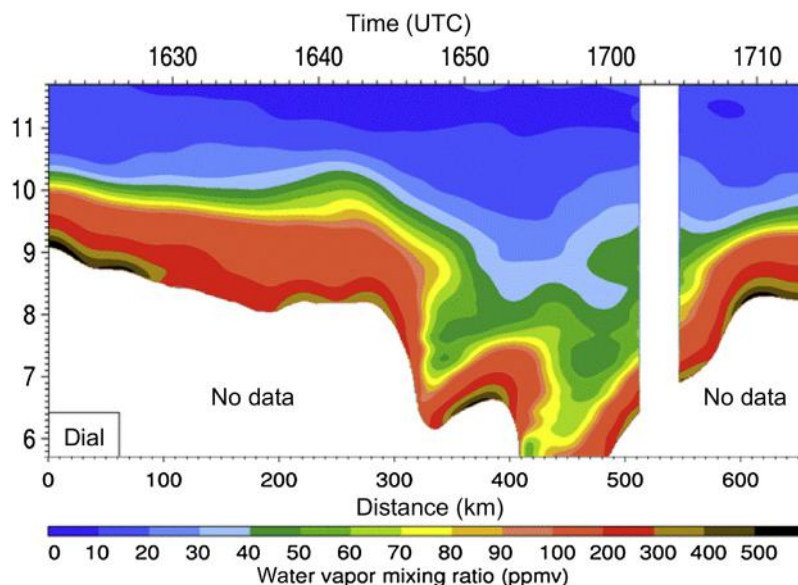


Figure 5 Cross-section of water vapor mixing ratios measured by the DLR DIAL system onboard the DLR Falcon aircraft on 12 November 1998 across a tropopause fold region. Tropopause is the boundary between the troposphere and stratosphere, and it was generally located near the blue-green border in this figure. Reproduced from Ehret, G., Hoinka, K., Stein, J., Fix, A., Kiemle, C., Poberaj, G., 1999. Low stratospheric water vapor measured by an airborne DIAL. *Journal of Geophysical Research* 104, D24. <http://dx.doi.org/10.1029/1999JD900959>.

and PRN altimeter are collected by the same telescope. The LAS and PRN altimeter returns are optically separated and detected by separate low-noise detectors. The energy-normalized LAS return signals are ratioed to yield the differential transmission of the transmitted wavelengths, and the natural log of the differential transmission is directly proportional to the CO₂ column amount. The PRN altimeter return is analyzed to determine the range to the surface or cloud, and provides information on the relative intervening aerosol-scattering profile and surface reflectance.

The MFL system has been extensively evaluated in many flight test campaigns conducted both during the day and at night, over a wide range of surface conditions (land and water), and in clear and scattered-cloud conditions. An example of the MFL measurements of surface reflectance and CO₂ columns (shown in terms of equivalent CO₂ mixing ratios) is shown in [Figure 6](#). *In situ* CO₂ profiles were obtained by a UC-12 (red) and Cessna (blue) aircraft for comparison with the remote MFL CO₂ measurements. The 1 s average surface reflectance variations obtained from the off-line backscatter returns are shown along the flight leg. The reflectance varied by a factor of 2.5 over very short distances as the laser beam backscattered from different surface types. This would create large errors in the IPDA CO₂ measurement unless the beams are transmitted simultaneously and the reflectance divides out in the IPDA analysis. The SNR for the MFL CO₂ measurement was found to be 760 for 1 s average (0.6 ppmv) and over 2000 for 10 s average (0.2 ppmv), which both exceed the CO₂ column measurement precision identified by the National Research Council (NRC) for the active sensing of CO₂ emissions over night, days, and seasons (ASCENDS) mission.

Other airborne lidar systems for IPDA CO₂ column measurements are under development and flight testing, and they include narrow and broadband-pulsed lidar systems

operating in the 1.57 and 2.06 μm regions and a heterodyne CW LAS system operating in the 2.06 μm region.

Recent DIAL System Studies and Developments

The idea of putting DIAL systems in space has been discussed since the late 1970s, and the initial emphasis was on the measurement of H₂O and O₃. This possibility became more likely with the successful deployment of the first cloud and aerosol lidar system on the space shuttle, the Lidar In-space Technology Experiment (LITE), in 1994 and the Cloud-Aerosol Lidar and Infrared Pathfinder Satellite Observation (CALIPSO) mission launched in 2006. The rationale for putting DIAL systems in space is to provide high-vertical-resolution measurements of gases such as O₃ and H₂O (along with aerosol and cloud measurements) across the troposphere, and of O₃ and aerosols across the stratosphere. Achieving total column CO₂ at high spatial resolution from space using laser remote sensing was considered to be of high priority for climate studies by the 2007 NRC decadal survey study. This study identified the ASCENDS space mission as being capable of providing global-scale CO₂ mixing ratio measurements by day and night, over all latitudes and seasons. Several approaches are being investigated for development of a CO₂ and O₂ space-based lidar system to achieve this capability.

Global O₃ Measurements

A number of key global atmospheric science issues can be addressed by a space-based O₃ DIAL system, including photochemical O₃ production, destruction, and transport in the troposphere; stratosphere-troposphere exchange; and

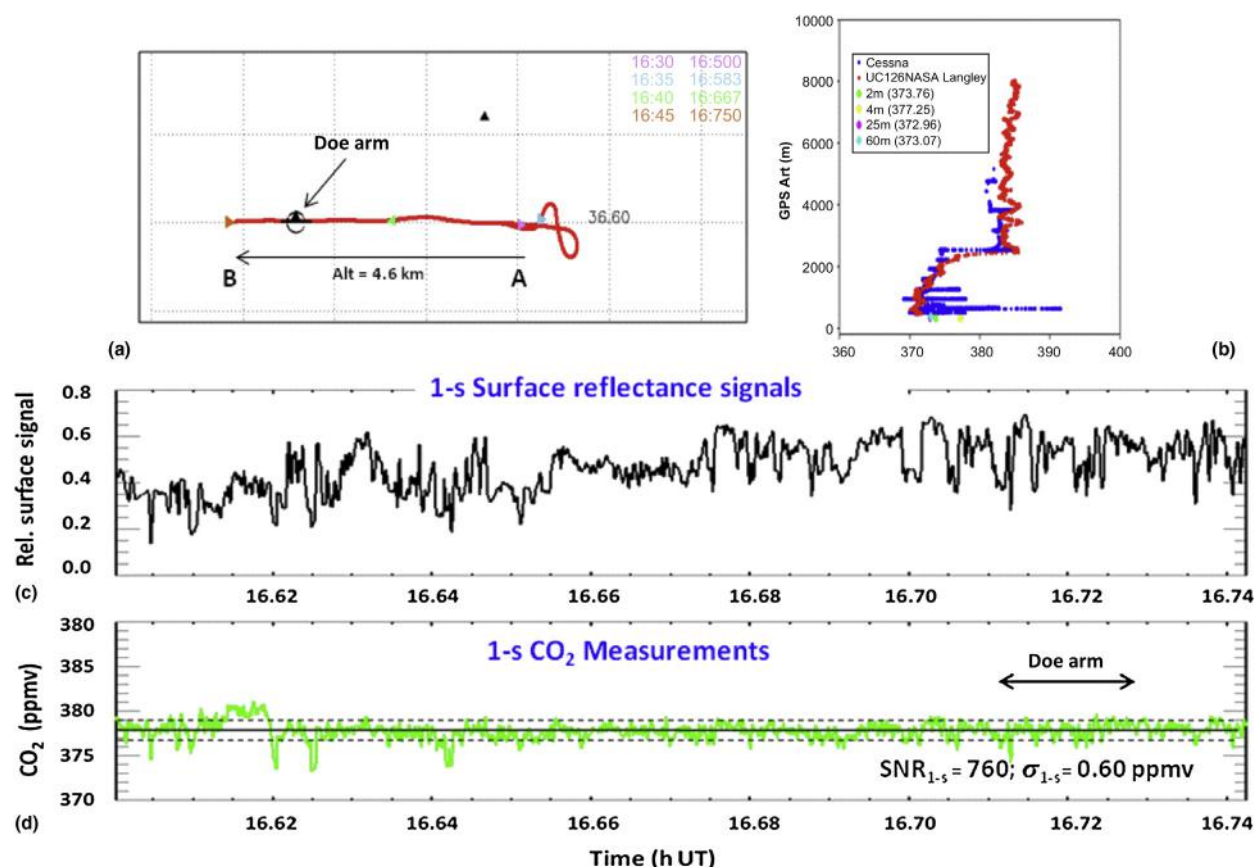


Figure 6 MFL measurements made on 31 July 2009 over the US Department of Energy's (DOE) Atmospheric Radiation Measurement (ARM) site in Lamont, Oklahoma. Flight track and altitude (top left), *in situ* CO₂ profiles (top right), relative surface backscattering (off-line wavelength) (middle), and MFL CO₂ column measurements with 1 s averaging along the flight track (bottom) are shown.

stratospheric O₃ depletion and dynamics. High-vertical-resolution (2–3 km) tropospheric O₃ measurements from space are needed for most of these studies, and this capability is not available from current or planned passive remote-sensing satellites.

Figure 3 shows an example of the type of latitudinal O₃ cross-section that could be provided in one pass of a space-based O₃ DIAL system. Even with the DIAL system optimized for tropospheric O₃ measurements (see the system description in Table 1), it would also provide simultaneous high-resolution stratospheric measurements of O₃ (1 km vertical, 100 km horizontal) and aerosols (100 m vertical, 10 km horizontal). In addition, these DIAL measurements will be useful in assisting in the interpretation of passive, remote-sensing measurements and in helping to improve their data-processing algorithms.

Figure 7 shows a simulation for expected measurement uncertainties as a function of altitude for an O₃ DIAL system at an orbital altitude of 400 km for the lidar parameters given in Table 5 using the US Standard Midlatitude O₃ model. In the middle of the stratosphere, where O₃ number densities are highest, the measurement uncertainty is a minimum of about 2%. In the troposphere, it increases to about 10% for altitudes below 6 km, which is similar to the performance of the airborne UV DIAL system.

Measurement of O₃ from space with the DIAL technique requires the development of high-power (>5 W) tunable lasers in the UV and large collection area (>3 m²) receivers. Current technology advancements have led to the development of low-pulse-energy (~5 mJ), moderate-power (~1 W), tunable solid-state lasers. As an interim step to space, these lasers have been incorporated into a compact (72 × 57 × 49 cm) DIAL system called the Global Ozone Lidar Demonstrator (GOLD) for autonomous operation on a Global Hawk uninhabited aerial system (UAS) aircraft. The packaging of the GOLD system is illustrated in Figure 8. With a nominal vertical measurement resolution of about 300 m and a horizontal averaging distance of 40 km, an O₃ measurement accuracy of better than 10% is expected. GOLD is planned to serve as a demonstrator of a space-based DIAL system and for use in field experiments to acquire long-range long-duration measurements from Global Hawk and for validation of O₃ measurements from the Thermal Emission Spectrometer (TES) and other passive satellite instruments.

Global H₂O Measurements

H₂O and O₃ are important to the formation of OH in the troposphere, and OH is at the center of most of the chemical reactions in the lower atmosphere. In addition to influencing

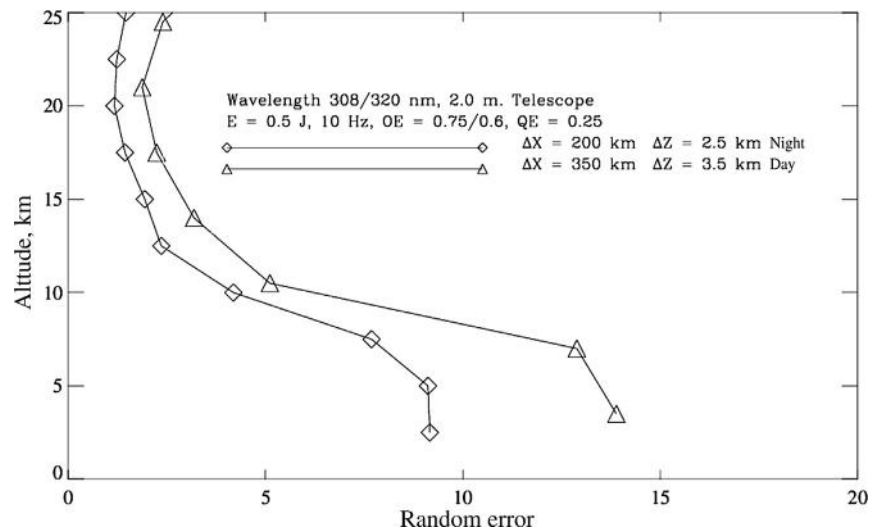


Figure 7 Simulations of expected space-based O_3 DIAL performance using the parameters given in Table 4 and nighttime and daytime background conditions.

Table 5 Example system parameters and performance goals for space-based O_3 and H_2O DIAL and CO_2 IPDA measurements from low Earth orbit (<400 km)

Parameters and goals	O_3 DIAL	H_2O DIAL	CO_2 IPDA
Laser transmitter			
Wavelength (λ): $\lambda_{on}/\lambda_{off}$ or λ region and $\Delta\lambda = \lambda_{on} - \lambda_{off}$ (nm)	308/320	940 region $\Delta\lambda < 70$ pm	1571/2050 region $\Delta\lambda < 100$ pm
Pulse energy, each λ (mJ)	500	200	Various
Pulse-pair repetition frequency (Hz)	10	20 (3 λ_{on} s)	Various
Pulse width (ns)	20	200	Various
Laser average power, each λ (W)	5	4	10
Line width (pm)	50	0.1	<0.06
Spectral purity (%)	>99	>99	>99.9
Beam divergence (mrad)	<0.15	<0.2	<0.2
Receiver			
Area (m^2)	3.0	3.0	>1.7
Field-of-view (mrad)	0.3	0.3	0.3
Filter bandwidth (nm)	0.5	0.04	1
Overall optical efficiency (%)	40	40	>60
Detector quantum efficiency (QE) (%)	31	60	Various

the production of OH, H_2O is an excellent tracer of vertical and horizontal transport of air masses in the troposphere, and it can be used as a tracer of stratosphere–troposphere exchange. Increased aerosol sizes due to high relative humidity can also affect heterogeneous chemical processes in the boundary layer and in cloud layers. Thus, knowledge of H_2O distributions can be used in several different ways to understand better the chemical and transport processes that influence the composition of the global troposphere.

The combination of active and passive measurements can provide significant benefits for H_2O , temperature, aerosol, and cloud information. High-vertical resolution H_2O (≤ 1 km), aerosol (≤ 100 m), and cloud top (≤ 50 m) measurements from the lidar along the satellite ground track can be combined with the horizontally contiguous data from nadir passive sounders like a Fourier transform spectrometer

(FTS) to generate a more complete high-resolution H_2O , aerosol, and cloud field for use in the various studies indicated in this article. The Combined Active and Passive Environmental Sounder (CAPES) concept of the combination of DIAL and FTS measurements is shown in Figure 9. The high-vertical H_2O distribution information provided by the DIAL system will assist in improvements in the retrieval of temperature (T) profiles and H_2O distributions from FTS data. In addition, the combination of T from the FTS and the improved H_2O field can be used to derive a more accurate 3D relative humidity field. Simultaneous aerosol and cloud profiling by the DIAL system will further enhance the strong synergism with aerosol and cloud mapping by other active and passive instruments like those in the A-train orbit for missions addressing atmospheric chemistry, radiation, hydrology, natural hazards, and meteorology.

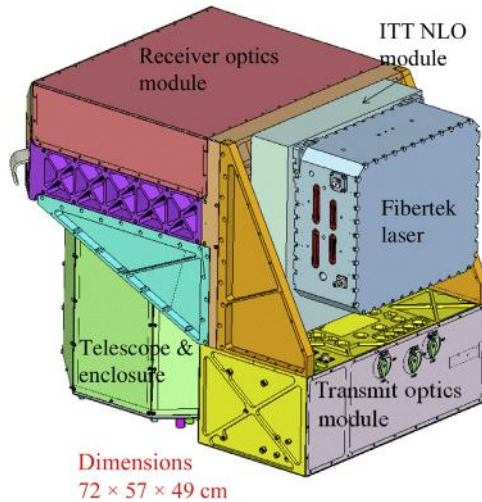


Figure 8 Schematic diagram of the packaging of the GOLD system for installation on the Global Hawk aircraft. Figure courtesy of Hair, J., NASA Langley.

The technology for a space-based H₂O DIAL system is developing incrementally in the areas of high-efficiency, high-energy, high-spectral-purity, long-life lasers with tunability in the 940 nm region; low-weight, large-area, high-throughput, high-background-rejection receivers; and high-quantum-efficiency, low-noise, photon-counting detectors. Figure 10 shows simulations of random errors for a space-based 940 nm H₂O DIAL system at 400 km using the parameters given in

Table 5. The simulation uses several lines in the 940 nm region that are chosen for their different absorption cross-sections and low temperature dependence. In the altitude range, where each cross-section is most sensitive, the random error is less than 5%. Thus, using two or three interleaved lines, the entire altitude range from near the surface to above 8 km can be covered to better than 10% accuracy. The European Space Agency (ESA) has conducted a system study for the development of a H₂O DIAL system from space called Water Vapor Lidar Experiment in Space (WALES). Achieving high pulse energy and high-power lasers for space-based H₂O DIAL applications remains the critical element in the development of H₂O DIAL systems for space.

Global CO₂ Measurements

The US National Research Council's report, entitled *Earth Science and Applications from Space: National Imperatives for the Next Decade and Beyond*, identified the need for a near-term space mission of ASCENDS. The primary objective of the ASCENDS mission is to make CO₂ column measurements across the troposphere during the day and night over all latitudes and all seasons, and in the presence of scattered clouds. These measurements would be used to reduce significantly the uncertainties in global estimates of CO₂ sources and sinks, to provide an increased understanding of the connection between climate and CO₂ exchange, to improve climate models, and to close the carbon budget for improved forecasting and policy decisions. To meet the science objectives of understanding carbon cycle processes, a CO₂ mixing ratio (XCO₂) measurement precision of about 1 ppmv over an atmospheric column

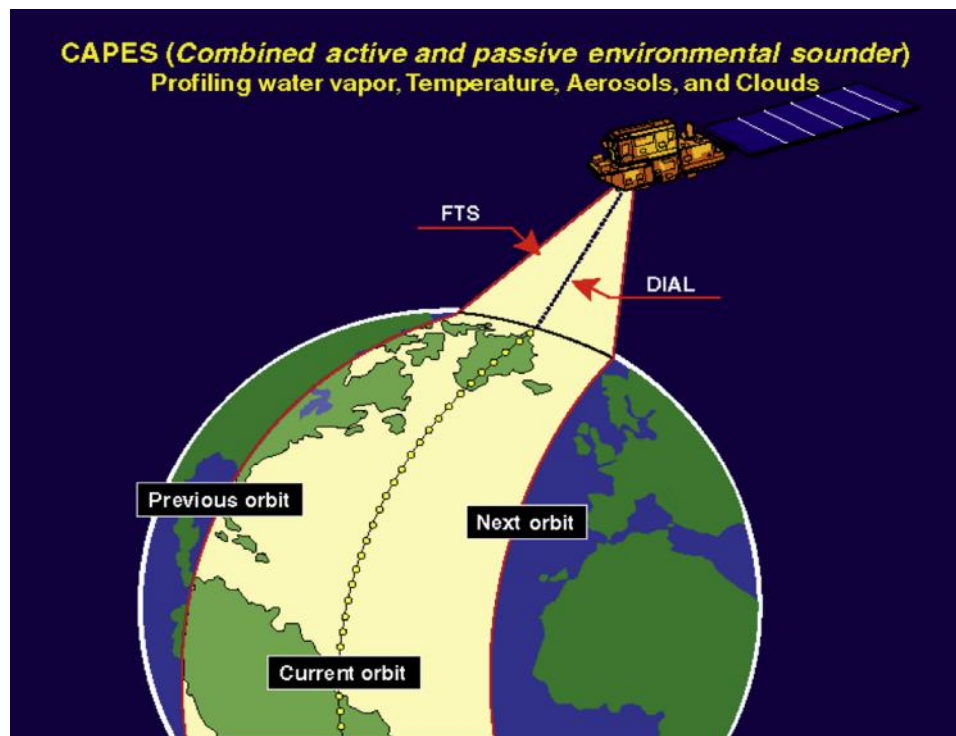


Figure 9 The concept of the combination of space-based DIAL and FTS measurements is illustrated in this figure. High-vertical resolution from DIAL is combined with large spatial coverage provided by FTS to enhance the quality of measurements.

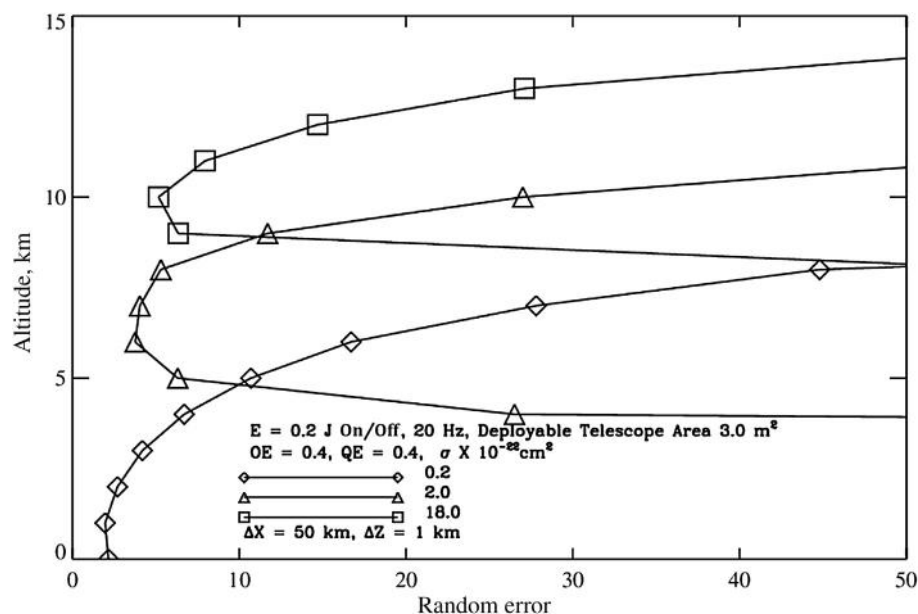


Figure 10 Simulations of expected space-based H₂O DIAL performance using the parameters in Table 5 and nighttime background conditions. Any altitude region of the troposphere can be probed by the choice of an appropriate H₂O absorption cross-section using the electronically tunable laser.

weighted across the troposphere is needed. This represents a combined CO₂ and O₂ column measurement uncertainty of about 0.3% for XCO₂. This is a very challenging requirement, and in addition to the airborne lidar systems mentioned in this article for CO₂ column measurements, there are two lidar systems being flight-tested for the O₂ column measurements: an IM-CW O₂ lidar system operating in the 1.26 μm region, which is being incorporated into the previously described MFL CO₂ lidar system, and a pulsed O₂ lidar system operating in the 0.765 μm region, which is being combined with the pulsed 1.57 μm CO₂ lidar system. These O₂ lidars operate using surface and cloud backscattering in an IPDA measurement of the O₂ column, which is then divided into the CO₂ column to derive XCO₂. Other techniques to obtain high-precision (≤ 1 hPa) estimates of the dry atmospheric density from meteorological analyses are also under evaluation. While airborne DIAL and IPDA systems continue to make significant contributions to our understanding of O₃, H₂O, CO₂, aerosol, and cloud distributions in investigations of atmospheric chemistry, dynamics, meteorology, and climate processes, they cannot provide the global and temporal coverage that a space-based DIAL and IPDA system can. Significant progress has been made in identifying the system requirements and enabling technologies needed for space-based DIAL and IPDA systems, and progress is being made in their development (see Table 5). Within the next decade, it is expected that DIAL and IPDA systems will be deployed in space for global measurements of CO₂ and CH₄, which use IPDA in the 1.65 μm region, and possibly H₂O and O₃.

Further Reading

- Ansmann, A., Neuber, R., Rairoux, P., Wandinger, U. (Eds.), 1997. *Advances in Atmospheric Remote Sensing with Lidar*. Springer Verlag, Berlin.
- Browell, E., 1989. Differential absorption lidar sensing of ozone. *Proceedings of the IEEE* 77 (3), 419–432.
- Browell, E., Ismail, S., Grant, W.B., 1998. Differential absorption lidar (DIAL) measurements from air and space. *Applied Physics B* 67, 399–410.
- Byer, R.L., Gustafson, E.K., Trebino, R. (Eds.), 1984. *Tunable Solid State Lasers for Remote Sensing*. Springer-Verlag, Berlin.
- Dabas, A., Loth, C., Pelon, J. (Eds.), 2001. *Advances in Laser Remote Sensing – Selected Papers Presented at the 20th International Laser Radar Conference*. Ecole Polytechnique, Palaiseau, France.
- Ehret, G., Hoinka, K., Stein, J., Fix, A., Kiemle, C., Poberaj, G., 1999. Low stratospheric water vapor measured by an airborne DIAL. *Journal of Geophysical Research* 104, D24. <http://dx.doi.org/10.1029/1999JD900959>.
- ESA, 2001. *WALEX – Water Vapour Lidar Experiment in Space: The Five Candidate Earth Explorer Core Missions*. SP-1257(2). ESA Publications Division, Noordwijk, The Netherlands.
- Fujii, T., Fukuchi, T. (Eds.), 2005. *Laser Remote Sensing*. Taylor and Francis, Boca Raton, FL.
- Grant, W.B., Browell, E.V., Menzies, R.T., Sassen, K., She, C.-Y. (Eds.), 1997. *Laser Applications in Remote Sensing*, SPIE Milestones Series. SPIE Optical Engineering Press, Bellingham, WA, p. 690, 86 papers.
- Kaminen, R., Krishnamurti, T.N., Ferrare, R.A., Ismail, S., Browell, E.V., 2003. Impact of high resolution water vapor cross-section data on hurricane forecasting. *Geophysical Research Letters* 30. <http://dx.doi.org/10.1029/2002GL016741>.
- Kovalev, V.A., Eichinger, W.E., 2004. *Elastic Lidar*. John Wiley and Sons, Hoboken, NJ.
- Measures, R.M., 1984. *Laser Remote Sensing – Fundamentals and Applications*. John Wiley & Sons, New York.
- National Research Council, 2007. *Earth Science and Applications from Space: A Community Assessment and Strategy for the Future*. The National Academies Press, National Academy of Sciences, Washington, DC.

See also: **Chemistry of the Atmosphere:** Observations for Chemistry (*In Situ*): Water Vapor Sondes. **Optics, Atmospheric:** Optical Remote Sensing Instruments.

Doppler

RM Hardesty, NOAA Environmental Technology Laboratory, Boulder, CO, USA

© 2003 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 3, pp 1194–1202, © 2003, Elsevier Ltd.

Introduction

The Doppler effect, in which radiation scattered or emitted from a moving object is shifted in frequency owing to the movement of the object, can be utilized by specially designed lidar systems to measure atmospheric motions, including winds. Atmospheric Doppler lidars operate by irradiating a volume of atmosphere with a pulse of very narrowband, laser-produced radiation, then detecting the change in frequency of the radiation backscattered from atmospheric aerosol particles or molecules present in the volume. The technique is directly analogous to that employed in Doppler weather radars, except that lidar wavelengths are three to four orders of magnitude shorter, which leads to some important differences between radar and lidar characteristics. Because radiation at wavelengths in the optical region of the spectrum is efficiently scattered by aerosols and molecules, Doppler lidars do not require the presence of hydrometeors or insects to obtain useful results. Optical radiation can be tightly focused, virtually eliminating ground clutter and enabling lidar probing of volumes to within a few meters of terrain or structures. However, because optical radiation is severely attenuated by cloud water droplets and fog, Doppler lidars do not typically probe into or beyond most atmospheric clouds – the one exception being tenuous ice clouds such as cirrus, which often are characterized by low optical extinction and high backscatter, making them excellent lidar targets.

The characteristics of Doppler lidar make the technique well suited to making detailed measurement of wind flows for a wide variety of applications. Lidar beams can easily be scanned to characterize motions within very confined three-dimensional spaces such as shallow atmospheric boundary layers, narrow canyons, and turbulent structures. Doppler lidar has also been proposed as a satellite-based technique for obtaining global measurements of atmospheric wind fields. By scanning a lidar beam from an orbiting satellite and analyzing backscattered returns from clouds, aerosols, and molecules, a satellite-based instrument could provide important wind information for numerical forecast models.

Basic Principles

The Atmospheric Backscattered Signal

As laser light propagates through the atmosphere, a portion of the incident energy is scattered by the atmospheric molecules and constituents through which the light passes. In lidar applications, the light backscattered (scattered directly back toward the source) is collected and analyzed. The backscattering properties of an atmospheric particle depend on the refractive index, shape, and size of the scattering particle. If the particle is much smaller than the wavelength of the incident radiation, as is the case for laser radiation scattered by

atmospheric molecules, the scattering process is characterized as Rayleigh scattering. In the Rayleigh scattering regime, the energy backscattered by a particle increases proportionally with the inverse of the fourth power of the wavelength. Consequently, Doppler lidar systems designed for molecular scatter can take advantage of significantly increased backscattered signal intensity by operating at short wavelengths. Molecular Doppler lidar systems typically operate in the visible or ultra-violet spectral regions.

Aerosol particles, the other component of the atmosphere that scatters laser light, result in Mie scattering, which is the more generalized scattering case that applies when the diameter of the scatterers is not orders of magnitude smaller than the incident wavelength. Aerosol particles include dust, soot, smoke, and pollen, as well as liquid water and ice. Although in Mie scattering the relationship between the power backscattered by an ensemble of aerosol particles and the incident wavelength is not simply characterized, most studies have shown that Mie backscatter in the atmosphere increases roughly with the first or second power of the inverse of the incident wavelength. In a high aerosol environment, such as in the vicinity of urban areas, an abundance of large particles often results in a roughly linear variation between the inverse of the incident wavelength and the backscattered energy. In more pristine environments, such as the free troposphere, the inverse wavelength/backscatter relationship can approach or exceed a square-law relationship.

The primary objective in Doppler lidar is to measure the Doppler frequency shift of the scattered radiation produced by the movements of the scattering particles. Figure 1 shows

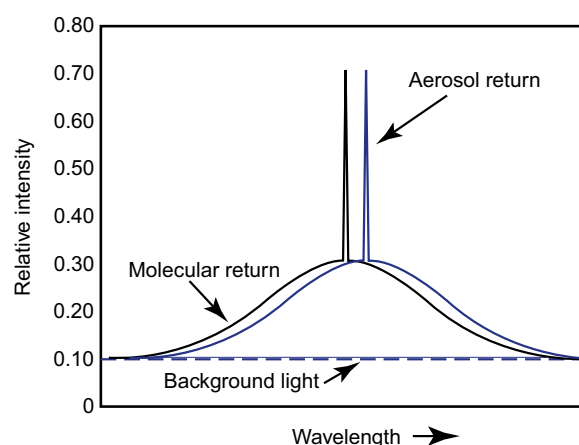


Figure 1 Spectrum of frequencies in a backscattered lidar return, showing unshifted (black) and Doppler shifted (blue) spectra, where the Doppler shift is due to the wind-induced motion of the scatterers. The ratio of aerosol to molecular signal intensity increases with increasing wavelength.

a typical spectrum of the radiation collected at the lidar receiver for a volume of atmosphere irradiated by a monochromatic laser pulse. Molecular scattering produces the broadband distribution in [Figure 1](#), where the broadening results from the Doppler shifts of the radiation backscattered from molecules moving at their thermal velocities. The width of the molecular velocity distribution in the atmosphere ranges from about 320 to 350 m s⁻¹, scaling as the square root of the temperature. In the center of the spectrum is a much narrower peak resulting from scattering of the light by aerosol particles. Since the thermal velocity of the much larger aerosol particles is very low, the width of the distribution of the aerosol return is determined by the range of velocities of particles moved about by small-scale turbulence within the scattering volume. This is typically on the order of a few meters per second. Also shown in [Figure 1](#) is an additional broadband distribution due to scattered solar radiation collected at the receiver. If the laser source is not monochromatic, the backscattered signal spectrum is additionally broadened, with the resulting spectrum being the convolution of the spectrum shown in [Figure 1](#) with the spectrum of the laser pulse. As seen in the figure, the entire spectrum is Doppler-shifted in frequency, relative to the frequency of the laser pulse. The object of a Doppler lidar system is to measure this Doppler shift, given by $\delta f = 2v_{rad}/\lambda$, where v_{rad} is the component of the mean velocity of the particles in the direction of propagation of the lidar pulse and λ is the laser wavelength.

The relative intensity of the scattered aerosol and molecular returns shown in [Figure 1](#) varies considerably as a function of laser wavelength and atmospheric turbidity. For lidar applications, molecular scatter is very strong at ultraviolet lidar wavelengths, decreasing as the fourth power of the wavelength until it is mostly negligible at wavelengths beyond about 1 μ m. Because most aerosol scattering is Mie scatter, which has a much weaker dependence on the laser wavelength, the ratio of molecular to aerosol scattered radiation decreases for longer wavelengths.

Components of a Doppler Lidar

Doppler lidar systems can be designed primarily to measure winds from aerosol-scattered radiation, or from molecule-scattered radiation, or from both. The type of system places specific requirements on the primary components that comprise a Doppler lidar system. A Doppler lidar is typically made up of a laser transmitter to produce pulses of energy that irradiate the atmospheric volume of interest; a receiver that collects the backscattered energy and estimates the backscattered energy and Doppler shift of the return; and a beam-pointing mechanism that directs the transmitter and receiver together in various directions to probe different atmospheric volumes and measure different components of the wind. Whether the primary scatterers are molecules or aerosol particles, in a Doppler system the system design criteria are driven by a fundamental relationship between the error in the estimate of mean frequency shift δf_1 , the bandwidth of the return f_2 , and the number of incident backscattered photons detected N , as in [eqn \[1\]](#).

$$\delta f_1 \propto \frac{f_2}{N^{1/2}} \quad [1]$$

Thus, the precision of the velocity estimate is always enhanced by an increase in the number of detected signal photons and/or a decrease of the bandwidth of the backscattered signal. From the equation, it is obvious that a significantly greater number of photons is required to achieve the same precision in a Doppler measurement from a molecular backscattered signal, characterized by higher bandwidth f_2 , compared to the number required in an aerosol Doppler instrument. The improved measurement precision gained by a narrow-bandwidth return also implies that the transmitter in a Doppler lidar system should be designed for narrow spectral width (typically on the order of the Doppler shift to be measured), and, as in most lidar systems, maximum transmitted energy. The need for a narrowband laser pulse has a strong effect on the laser performance requirements, and usually results in a transmitter that is more complex than the transmitter in a simple aerosol backscatter lidar. Injection seeding, in which a small amount of power from a separate laser is injected into the laser transmitter optical cavity to 'seed' the laser into operating at a single frequency, is used to produce the required narrow-bandwidth laser pulses in Doppler lidar instruments.

The lidar receiver gathers the backscattered photons and extracts the wind velocity as a function of the range to the scattering volume by analyzing the return as a function of time. This requires a telescope, to gather and focus the scattered radiation, and a system element that analyzes the scattered radiation to compute the Doppler shift. The frequency analysis function in a Doppler lidar receiver is carried out using one of two fundamentally different techniques: coherent detection (also known as heterodyne detection) or direct detection (alternatively labeled incoherent detection). The techniques differ fundamentally. In a heterodyne receiver, local oscillator laser radiation is mixed with the backscattered radiation at an optical detector, and the detector output signal is digitized and spectrally processed. A direct detection lidar employs an interferometer to optically analyze the backscattered radiation.

Coherent Doppler Lidar

Description

Coherent or heterodyne lidar is implemented by optically mixing the backscattered laser light with radiation from a continuous-wave, local oscillator (LO) laser whose wavelength is precisely controlled to be equal to, or at a known displacement from, that of the laser transmitter ([Figure 2](#)). The mixing process at the face of an optical detector generates an electrical signal with a frequency equal to the difference frequency between the backscattered signal and the LO laser signal. This signal is typically digitized and processed as a function of time, using digital signal processing techniques, to obtain an estimate of the range-dependent mean frequency shift of the backscattered signal, from which the radial wind component can be derived. A characteristic of coherent lidar is that single pulse returns, even with high signal levels, are characterized by random fluctuations in the backscattered field and the resulting frequency spectrum. Consequently, averaging of multiple pulses usually is employed to increase the accuracy of the mean frequency estimate.

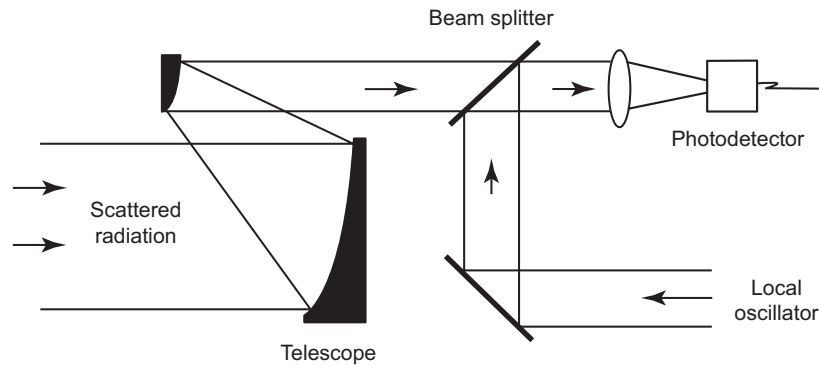


Figure 2 Schematic of coherent detection of backscattered radiation.

The optical mixing process that is fundamental to coherent Doppler lidar provides both benefits and design challenges. Receiver efficiencies (measured as equivalent fraction of incident photons converted to electrons) are quite high, meaning that wind velocities can be estimated from weak backscattered signals. Also, because signal processing is performed on a time-series derived from the mixed signal, receiver bandwidths can be made extremely narrow, effectively eliminating the broadband solar background light as a major source of noise. Thus, unlike that of many lidars, coherent Doppler lidar performance is not degraded under daylight conditions. However, the added photon noise from the local oscillator radiation results in a system noise threshold below which very weak signals cannot practically be extracted, even with substantial multiple-pulse averaging. Because efficient mixing requires phase coherence of the backscattered signal field across the detector, coherent lidar performance at longer ranges can be degraded by strong optical turbulence along the path of the laser pulse. Turbulence effects are more pronounced at shorter wavelengths.

Coherent Doppler lidar systems generally operate in the eye-safe, infrared portion of the spectrum at wavelengths longer than $1.5\ \mu\text{m}$. Currently, the two most common system wavelengths for coherent lidar wind systems are in atmospheric window spectral regions around 2.0 and $10.6\ \mu\text{m}$. Early coherent Doppler lidar measurements, beginning in the 1970s, employed CO_2 laser transmitters and local oscillators operating at wavelengths near $10.6\ \mu\text{m}$. Pulsed CO_2 laser transmitters with as much as $10\ \text{J}$ of energy have since been demonstrated, and systems with $1\ \text{J}$ lasers are still routinely used for atmospheric probing to ranges extending to $30\ \text{km}$ or more. In the late 1980s, solid-state laser transmitters operating near $2\ \mu\text{m}$ wavelengths were introduced into coherent lidar wind-measuring systems. The compact size and potential reliability advantages of solid-state transmitters, in which the transmitter laser is optically pumped by an array of laser diodes, provide advantages over older CO_2 laser technology. Also, because the range resolution obtainable for a given measurement accuracy scales with wavelength, $2\ \mu\text{m}$ instruments are characterized by better range resolution than their $10\ \mu\text{m}$ counterparts, and hence have enhanced ability to probe small-scale features. However, although development of high-energy laser sources operating in the $2\ \mu\text{m}$ region is currently an active research area, lasers with pulse energies greater than

several tens of millijoules have yet to be incorporated into laser systems.

Applications of Coherent Doppler Lidar

Coherent lidars have been employed to measure winds for a variety of applications, and from an assortment of platforms, such as ships and aircraft. Since these lidars operate in the infrared, where aerosol scattering dominates molecular scattering, they require aerosol particles to be present at some level to obtain usable returns. Although clouds also provide excellent lidar targets, most of the more useful applications of coherent lidars have been associated with probing the atmospheric boundary layer or lower troposphere, where aerosol content is highest. One of the special applications of lidar probing is measurement of wind structure and evolution in complex terrain such as mountains and valleys. Over the past 20 years, results of Doppler lidar studies have been used, for example, to characterize the intensity and structure of damaging downslope windstorms in the lee of mountain ranges, advection of pollution by drainage flows in valleys, and formation of mountain leeward turbulence as a potential hazard to landing aircraft. An example of the capability of Doppler lidar to probe complex terrain is shown in Figure 3, in which the advance of a cold front northward during a southerly foehn wind event in the Austrian Alps is clearly indicated. Such measurements are studied to improve understanding of the mechanisms associated with the onset and evolution of severe wind events.

Because coherent Doppler lidars are also well matched to applications associated with probing small-scale, turbulent phenomena, they can be applied to improving aviation safety. Doppler lidars have been deployed in the vicinity of airports to detect and track wing tip vortices generated by arriving or departing aircraft. In the future, a network of ground-based lidars to provide information on vortex location and advection speed could decrease congestion at major airports by increasing landing and takeoff capacity. Also, compact lidar systems deployed on research aircraft have detected wave structures ahead of the aircraft associated with potentially hazardous clear-air turbulence. Doppler lidar installed on the commercial aircraft fleet might be able to look ahead and provide a warning to ensure that passengers have their seat belts fastened before the turbulence is encountered.

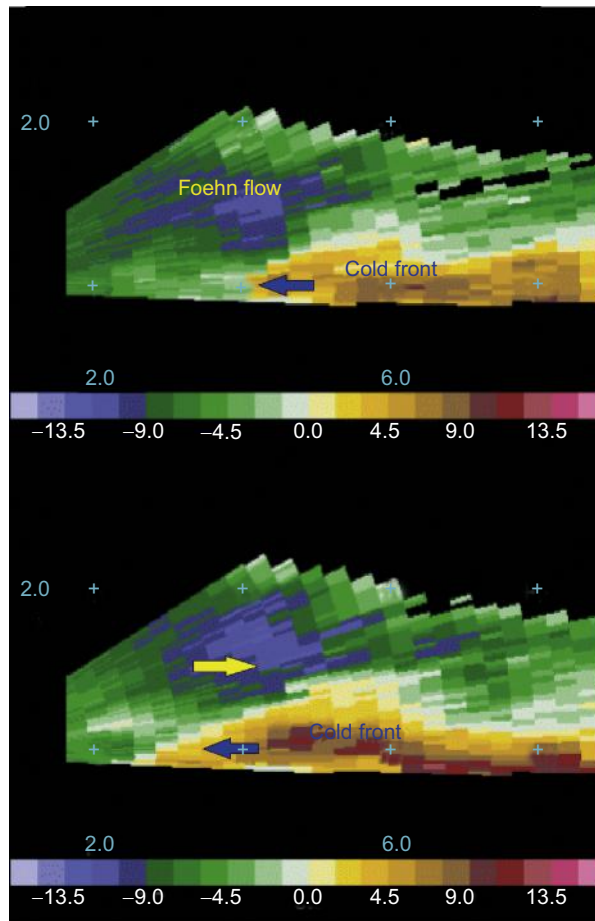


Figure 3 Successive lidar vertical plane scans of radial wind speed, showing the progression of a cold front beneath a strong foehn flow in the Austrian Alps. Colors represent radial velocity in m s^{-1} ; distances are in km from the lidar. Courtesy L. Darby, NOAA.

The high resolution obtainable in a scanning lidar can produce informative images of the details of small-scale phenomena. **Figure 4** shows an image of a shallow wave associated with a low-level nocturnal stable layer structure just 50 m above the surface obtained using a scanning $2.02 \mu\text{m}$ Doppler lidar. Similar images taken during different evenings as part of the same experiment illustrated markedly different characteristics, such as much more turbulence along the interface. Such observations enable researchers to improve models and better understand the conditions associated with vertical turbulent transport and mixing of atmospheric constituents such as pollutants.

Direct Detection Doppler Lidar

Description

Direct-detection, or incoherent, Doppler lidar has received significant attention in recent years as an alternative to coherent lidar for atmospheric wind measurements. In contrast to coherent lidar, in which an electrical signal is processed to estimate Doppler shift, an optical interferometer, usually

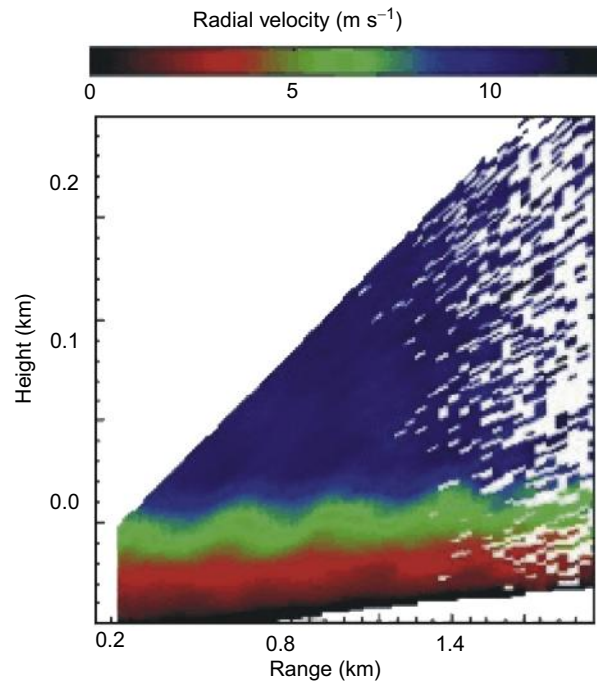


Figure 4 Wave structure within a nocturnal stable layer at a height of 50 m. Note the stretching of the vertical scale. Courtesy R. Banta, NOAA.

a Fabry–Perot etalon, serves as the principal element in a direct-detection lidar receiver for determining the frequency shift of the backscattered radiation.

One implementation of a direct-detection Doppler lidar receiver is the ‘fringe imaging’ technique. In this design, the interferometer acts as a spectrum analyzer. The backscatter radiation is directed through a Fabry–Perot interferometer, which produces a ring pattern in the focal plane (**Figure 5**). The spectral content information of the incident radiation is contained in the radial distribution of the light. Each ring corresponds to an order of the interferometer and is equivalent to a representation of the backscattered signal frequency spectrum. As the mean frequency of the backscattered radiation changes, the rings move inward or outward from the center. To extract the spectrum of the backscattered light, one or more of the rings are imaged onto a multielement detector, and the resulting pattern is analyzed.

An alternative direct detection receiver configuration employs two interferometers as bandpass filters, with the center wavelengths of respective filters set above and below the laser transmitter wavelength, as shown in **Figure 6**. The incoming radiation is split between the two interferometers, and the wavelength shift is computed by comparing the radiation transmitted by each interferometer. This method, sometimes called the ‘double edge’ technique, has been used to measure winds to heights well into the stratosphere. The major challenge associated with this implementation of a direct detection receiver is optimizing the instrument when both aerosol-scattered and molecular-scattered radiations are present, since in general the change in transmission as a function of velocity is different for the aerosol and molecular signals.

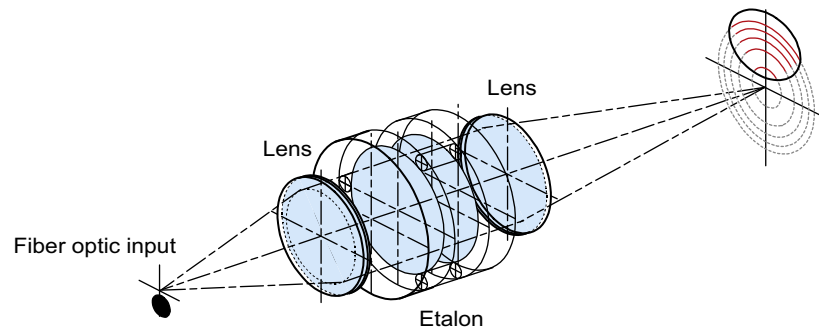


Figure 5 Schematic of Fabry–Perot etalon in a direct-detection, fringe-imaging lidar receiver. Courtesy P. Hays, Michigan Aerospace Corporation.

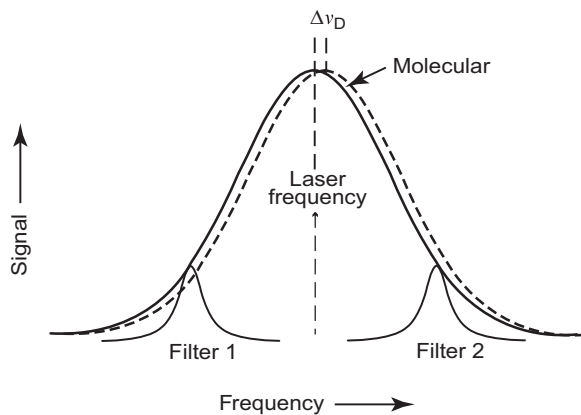


Figure 6 Spectrum of molecular lidar return, showing placement of bandpass filters for a dual-channel (double-edge) direct detection receiver. Courtesy B. Gentry, NASA.

For both types of direct-detection receiver configurations described above, a portion of the radiation incident on the interferometer is reflected out of the system, reducing the overall efficiency of the receiver. Recently, designs that incorporate fiber optics to collect a portion of the reflected radiation and recycle it back into the etalon have been proposed as a way to reduce receiver losses.

Doppler Measurements Based on Molecular Scattering

One of the primary advantages of direct-detection Doppler lidar is the capability for measurements based on scatter from atmospheric molecules. Measurement of Doppler shifts from molecular scattered radiation is challenging because of the large Doppler-broadened bandwidth of the return. Typically, one wants to measure a mean wind velocity with a precision that is much less than the velocity standard deviation of 320 m s^{-1} . This requires a large number of photons. As an example, more than 10^5 photons must be detected to achieve a measurement precision of a few m s^{-1} from molecular-scattered radiation, which typically requires some combination of multiple pulse averaging, powerful lasers, and large receivers for the lidar system.

This need to collect and detect large numbers of photons also means that molecular Doppler measurements are made at

short wavelengths where scatter is strongest. Molecular-scatter wind measurements have been demonstrated in the visible spectral region at 532 nm wavelength and at 355 nm in the near ultraviolet. The ultraviolet region has the dual advantages of enhanced molecular scatter and less-restrictive laser eye-safety restrictions. Figure 7 shows the time series of a wind profile measured throughout the troposphere using a molecular-scatter, ground-based Doppler fringe-imaging lidar operating at 532 nm. The figure also shows measurements from a second receiver channel in which the interferometer design was optimized for the more narrowband aerosol return. For this measurement, returns from 500 pulses, each with 400 mJ of energy, were collected by a 0.5 m receiver aperture and processed.

Heterodyne and Direct-Detection Doppler Trade-Offs

Lively debates within the lidar community have taken place over the past decade regarding the relative merits of heterodyne versus direct-detection Doppler lidars. To a large extent, the instruments are complementary. Generally, heterodyne instruments are quite sensitive for measurements when significant aerosols are present. Processing techniques have been developed that can produce accurate wind measurements rates using only a few lidar pulses, such that several wind observations per second can be obtained. This inherent sensitivity has enabled numerous applications in which a lidar beam has been scanned rapidly over a large volume to obtain time-varying, three-dimensional wind measurements. Heterodyne lidars also operate in the eye-safe infrared portion of the spectrum, which is highly advantageous for general atmospheric research and field studies.

The primary advantage of direct-detection instruments is their demonstrated capability to measure winds from molecular-backscattered returns in the middle and upper troposphere. In pristine air, direct detection offers the only method for long-range wind measurements, even though significant averaging may be required. Direct-detection lidars are also not degraded by atmospheric refractive turbulence. From an engineering perspective, optical system specifications for direct detection systems tend to be somewhat less demanding. For optimum performance, a heterodyne lidar requires a very pure laser pulse and a diffraction-limited receiver field of view that is matched to and precisely aligned

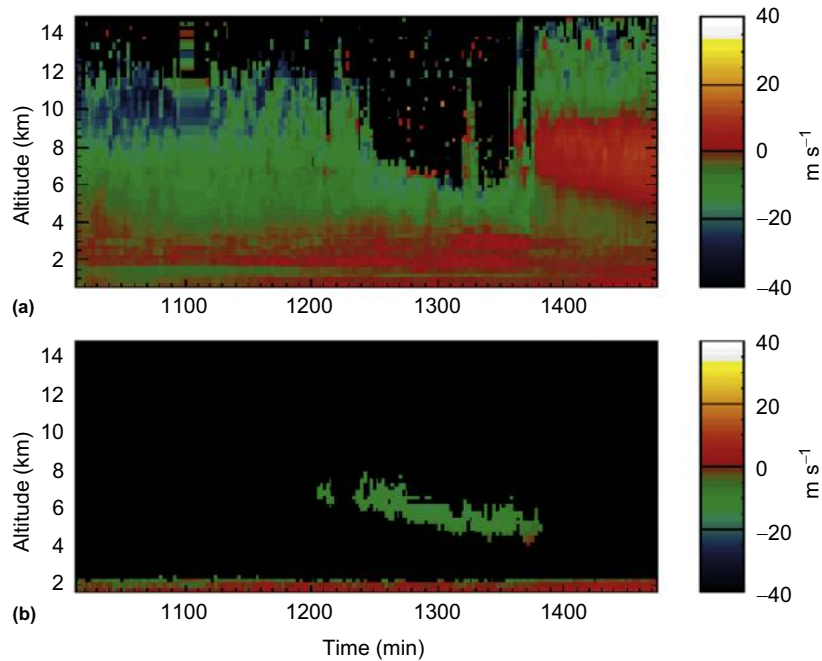


Figure 7 Time-series (in minutes after local midnight) of radial wind speed profiles measured by a fringe-imaging direct detection lidar molecular channel (a) and aerosol channel (b) at Bartlett, NH. Change in wind speed and blocking of the return by clouds are clearly seen. Courtesy P. Hays, Michigan Aerospace Corp.

with the transmitted beam. The requirements for direct detection are more relaxed, such that a somewhat wider bandwidth transmitter and a receiver field of view several times diffraction-limited is sufficient. In direct-detection lidars, the field of view is typically constrained by the need to limit background light for daytime operation. The major design challenge for direct detection instruments is associated with maintaining stability of the Fabry-Perot etalon over a range of temperatures and in high-vibration environments.

Global Wind Measurements

A satellite-based Doppler lidar has frequently been proposed as a way of measuring wind fields over most of the Earth. At present, wind profiles are not well measured from orbiting platforms. Measurement of winds is especially important over regions of the Earth that are not currently well sampled, such as over Northern Hemisphere oceans, as well as over most of the tropics and the Southern Hemisphere. Wind profile information is currently obtained from radiosondes and by tracking cloud and water vapor inhomogeneities using satellite images. Doppler lidar wind measurements could greatly augment the current data set by providing wind estimates throughout the troposphere under clear conditions, and strongly height-resolved observations down to cloud tops when cloud decks are present. Observing system simulation experiments conducted in recent years indicate that satellite-based lidar global wind measurements could lead to a significant improvement in long-term forecast skill if the wind fields can be observed with sufficient accuracy and spatial resolution.

Figure 8 shows an artist's concept of a shuttle-based wind lidar mission. In an operational mission, a satellite carrying the lidar would be placed in a nearly polar orbit. The pulsed laser beam is scanned conically from nadir to obtain different components of the wind velocity. The scanning can be either



Figure 8 Concept for a demonstration wind lidar mission from a shuttle orbiter. In an operational mission, the lidar would be deployed on a polar-orbiting satellite for global coverage. Courtesy M. Kavaya, NASA.

continuous or 'stop and stare.' After sufficient returns are averaged to produce an acceptable estimate, the radial component of the velocity is computed, which would most likely be assimilated directly into numerical forecast models.

Doppler lidar measurement of winds from space is theoretically feasible but technologically difficult. Depending on the orbital height, the scattering volume is anywhere from 450 to about 850 km from the satellite, which challenges the sensitivity of current system types. Because weight and power consumption are critical parameters for space-borne systems, telescope diameter and laser power cannot easily be increased to obtain the necessary sensitivity. Similarly, the ability to average returns from multiple pulses is also limited. The satellite moves at about 7 km s^{-1} ; therefore, in order to obtain measurements over a horizontal distance of 300 km (the resolution of the radiosonde network), only about 45 s are available to make observations from the multiple look angles needed for a useful measurement. It should be also noted that, as a result of the high orbital velocity of the satellite, the precision and knowledge of beam pointing is extremely critical for measurements from satellites. For a lidar operating with a nadir angle of 45° , an error in the pointing knowledge of just 1 mrad results in an error of about 5 m s^{-1} introduced into the measured radial component of the wind by the satellite motion.

Despite the challenge of employing a Doppler lidar for satellite-based wind measurements, efforts are continuing to develop the appropriate technology and to assess the impact of the observations. The European Space Agency is planning a Doppler wind lidar demonstration mission for the late 2000s that would incorporate a nonscanning, direct-detection instrument. Doppler lidar technology research and simulation experiments aimed at satellite-based wind sensing continue at several research centers within the United States, Europe, and Japan.

See also: Chemistry of the Atmosphere: Observations for Chemistry (Remote Sensing): Lidar. **Lidar:** Atmospheric Sounding Introduction; Backscatter; Differential Absorption Lidar; Raman; Resonance. **Radar:** Polarimetric Doppler Weather Radar.

Further Reading

- Baker, W.E., Emmitt, G.D., Robertson, F., et al., 1998. Lidar-measured wind from space: a key component for weather and climate prediction. *Bulletin of the American Meteorological Society* 76, 869–888.
- Gentry, B., Chen, H., Li, S.X., 2000. Wind measurements with a 355 nm molecular Doppler lidar. *Optics Letters* 25, 1231–1233.
- Grund, C.J., Banta, R.M., George, J.L., et al., 2001. High-resolution Doppler lidar for boundary-layer and cloud research. *Journal of Atmospheric and Oceanic Technology* 18, 376–393.
- Huffaker, R.M., Hardesty, R.M., 1996. Remote sensing of atmospheric wind velocities using solid state and CO_2 coherent laser systems. *Proceedings of the IEEE* 84, 181–204.
- Menzies, R.T., Hardesty, R.M., 1989. Coherent Doppler lidar for measurements of wind fields. *Proceedings of the IEEE* 77, 449–462.
- Post, M.J., Cupp, R.E., 1990. Optimizing a pulsed Doppler lidar. *Applied Optics* 29, 4145–4158.
- Rees, D., McDermid, I.S., 1990. Doppler lidar atmospheric wind sensor: reevaluation of a 355 nm incoherent Doppler lidar. *Applied Optics* 29, 4133–4144.
- Rothermel, J., Cutten, D.R., Hardesty, R.M., 1998. The multi-center airborne coherent lidar atmospheric wind sensor. *Bulletin of the American Meteorological Society* 79, 581–599.
- Skinner, W.R., Hays, P.B., 1994. Incoherent Doppler lidar for measurement of atmospheric winds. *Proceedings of the SPIE* 2216, 383–394.
- Souprayen, C., Garnier, A., Hertzog, A., 1999. Rayleigh–Mie Doppler wind lidar for atmospheric measurements II: Mie scattering effect, theory, and calibration. *Applied Optics* 38, 2422–2431.
- Souprayen, C., Garnier, A., Hertzog, A., Hauchecorne, A., Porteneuve, J., 1999. Rayleigh–Mie Doppler wind lidar for atmospheric measurements. I: instrumental setup, validation, first climatological results. *Applied Optics* 38, 2410–2421.

Raman

DN Whiteman, NASA Goddard Space Flight Center, Greenbelt, MD, USA

© 2003 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 3, pp 1202–1212, © 2003, Elsevier Ltd.

Introduction

Optical techniques for remote sensing of the atmosphere have existed for many decades. In the 1930s, Hulburt used mechanically chopped searchlights as an optical source and a telescope as the receiver to measure atmospheric signals to 28 km. In the 1950s, Elterman used the same searchlight technique to derive atmospheric temperature to altitudes in excess of 60 km. Theodore Maiman's invention of the laser in 1960 created a new optical source, enabling the development of lidar and revolutionizing the field of remote sensing.

Lidar stands for *light detection and ranging*, and is used to describe the use of laser radiation to make remote measurements of the atmosphere, ocean, or hard target where the backscattered signal is collected by use of a telescope. Some significant early achievements are summarized in Table 1. Following these early efforts, researchers soon broadened the range of phenomena used for laser remote sensing by using the Raman effect to study molecular species remotely. Raman lidar technology can be used to provide scientifically meaningful measurements of a wide variety of atmospheric quantities including temperature, ozone and aerosol size. The presentation here will focus on the use of Raman lidar in the measurements of water vapor, aerosols, and cirrus clouds.

Raman Scattering

In 1921, Chandrasekhara V. Raman and his students began investigating the light scattering properties of various substances. Their experiments were carried out at the University of Calcutta in India, where Raman held an endowed chair in physics. These investigations led to the discovery of a very weak type of secondary light that is generated at wavelengths shifted from the incident wavelength. Initially they used filtered sunlight as the source and the shifted wavelengths

were observed by eye. Raman soon realized that they were observing a completely new scattering phenomenon of fundamental importance. They found that the frequency shifts, their relative intensities, and the state of polarization were independent of the exciting radiation. The frequency shifts observed were attributed to the frequencies of oscillation of the atomic bonds in a molecule. Raman was awarded the Nobel Prize for his work in 1930.

A quantum mechanical description of the Raman scattering process (Stokes shift) involves a transition from an initial state of energy E_i to an intermediate state, known as a virtual state, prior to the transition to the final state E_f . This is shown in the Figure 1 where both Stokes and anti-Stokes Raman scattering are shown. After a Stokes scattering event, the system is left in a higher energy state so that the scattered photon is of longer wavelength (lower energy) than the incident photon. In anti-Stokes scattering, the final energy state is lower than the initial state.

The Raman Lidar Technique

The Raman lidar technique entails measuring the rotational or vibrational-rotational component of Raman scattering from the atmosphere. This approach has proven to be a highly versatile one permitting a wide variety of atmospheric studies to be performed. Water vapor, nitrogen, and oxygen are molecules of interest in the normal atmosphere that exhibit convenient vibrational Raman shifts. These shifts are approximately 3657, 2331, and 1556 cm^{-1} , respectively. A spectrum of Raman scattering (Stokes shift) from the normal atmosphere is shown in Figure 2.

The Standard Raman Lidar Equation

The standard single scattering Raman lidar equation can be expressed as in eqn [1].

Table 1 Early developments in the history of lidar

Event	Date	Researchers	Comments
Laser invented	1960	Maiman	
First laser remote sensing measurements – echos from the moon	1962	Smullin and Fiocco	
Invention of the Q-switch	1962	McClung and Hellwarth	Permitted much higher power levels making atmospheric measurements feasible
First atmospheric laser remote sensing measurements	1963	Fiocco and Smullin	First atmospheric lidar measurements closely followed the invention of the Q-switch
First differential absorption lidar (DIAL) measurements of water vapor	1966	Schotland	DIAL measurements of atmospheric water vapor preceded Raman lidar measurements of water vapor
First measurement of the atmospheric sodium layer using resonance lidar	1968	Bowman <i>et al.</i>	

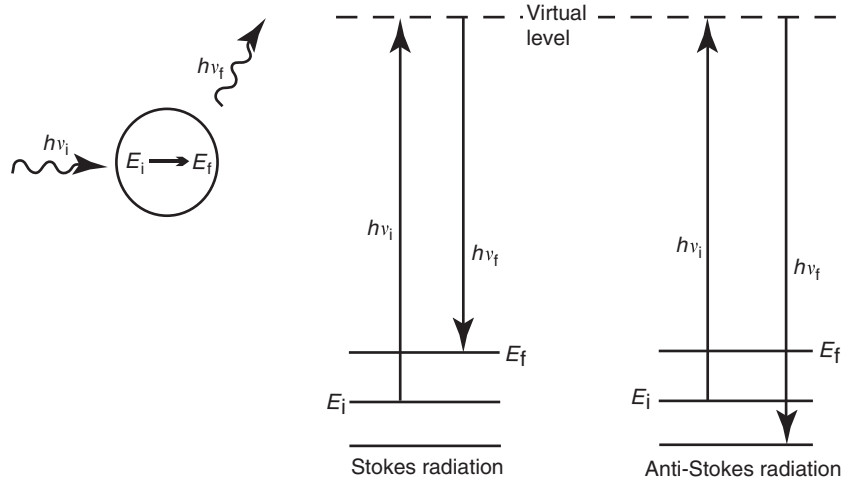


Figure 1 Schematic diagrams of both Stokes and anti-Stokes Raman scattering. Stokes scattering is the predominant phenomenon observed in the atmosphere. This results in a scattered photon of lower energy and thus longer wavelength than the incident. Anti-Stokes scattering results in a higher energy or shorter wavelength photon being scattered.

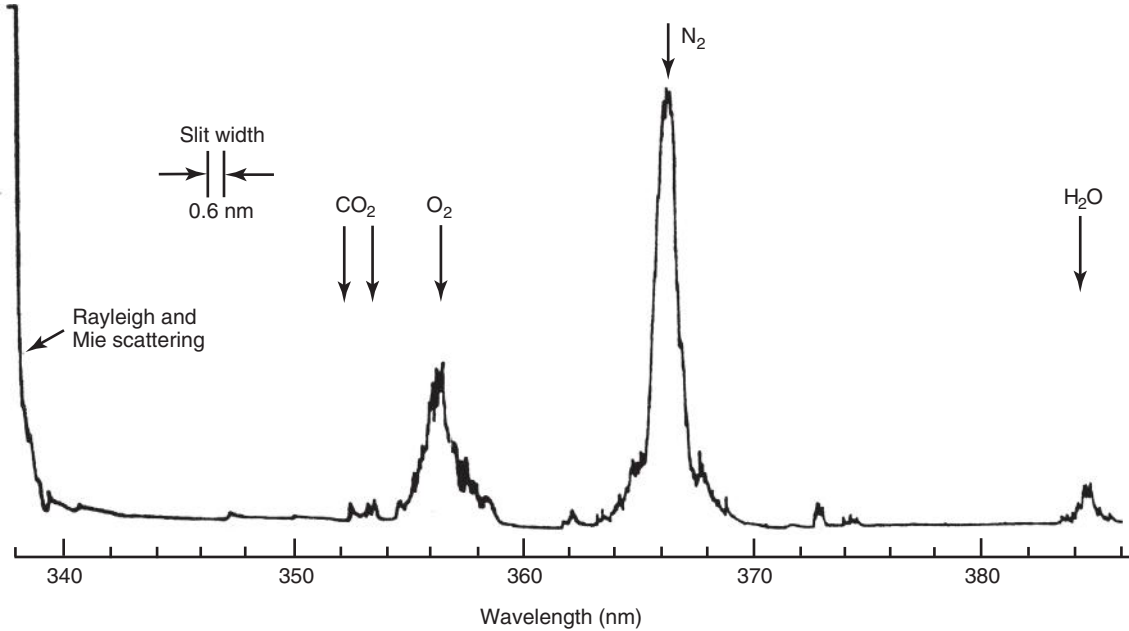


Figure 2 Measured spectrum of Raman-shifted and unshifted backscatters from the normal atmosphere. Redrawn with permission from Ianaba and Kobayashi (1972).

$$P(\lambda_{X,r}) = \frac{O_X(r)P_0(\lambda_L)N_X(r)[d\sigma_X(\lambda_L, \pi)/d\Omega]A\xi(\lambda_X)}{r^2} \times \exp\left\{-\int_0^r [\alpha(\lambda_L, r') + \alpha(\lambda_X, r')]dr'\right\} \quad [1]$$

Here $P(\lambda_{X,r})$ is the background subtracted power received at the Raman-shifted wavelength appropriate for molecular species X as a function of range, r . $O_X(r)$ is the channel overlap function, and $P_0(\lambda_L)$ is the output power of the laser at the laser wavelength, λ_L . $N_X(r)$ is the number density of molecules and $d\sigma_X(\lambda_L, \pi)/d\Omega$ is the Raman differential backscatter cross-section at the laser wavelength. $\xi(\lambda_X)$ is the

total lidar receiver optical efficiency at the wavelength of the Raman species of interest and includes factors such as the reflectivity of the telescope, the transmission of any conditioning optics, the transmission of any filters, and the quantum efficiency of the detector. A is the receiver telescope area. The exponential factor gives the two-way atmospheric transmission, where $\alpha(\lambda_X, r)$ is the total extinction coefficient due to scattering and absorption by molecules and aerosols at the Raman-shifted wavelength as a function of range along the path of the laser beam. In this context, the term 'aerosols' may be used to describe any nonmolecular atmospheric constituent such as dust,

water droplets, ice crystals, etc. This equation uses the simplifying assumption that the Raman scattered signals are monochromatic.

The Raman effect has been used to great effect in laboratory studies of various materials. Since the advent of the laser, it has also been used for atmospheric studies. Table 2 presents a chronological list of various milestones in the history of Raman lidar measurement capability. As an illustration of the Raman lidar technique for measuring water vapor, aerosols, and clouds, the measurement capability of a single Raman lidar system will now be illustrated.

The NASA/GSFC Scanning Raman Lidar

The NASA/GSFC scanning Raman lidar (SRL) is housed in a single mobile trailer and contains a 0.76 m telescope, two lasers (XeF excimer (351 nm) and a tripled Nd:YAG (355 nm)), various wavelength-selection optics and combined photon counting and analog data acquisition electronics. The trailer also provides laboratory space for new experiment development as well as workspace for up to three analysts. The system

is designed to measure the Raman scattering from water vapor, liquid water, nitrogen, and oxygen. In addition, the Rayleigh–Mie signal is measured as well. A large mirror enables full aperture scanning within a single plane perpendicular to the trailer axis. The scanning capability is used to assess horizontal variability of various constituents as well as to improve the retrieval of scientific quantities in the lowest part of the atmosphere. The data are typically acquired with 1 min temporal resolution and 7.5 m vertical resolution. Some of the basic measurement capabilities of this system will now be illustrated.

Measurements

Water vapor

The water vapor mixing ratio is defined as the ratio of the mass of water vapor to the mass of dry air in a given volume. The mixing ratio is conserved in atmospheric processes that do not involve condensation or evaporation, and thus serves well as a tracer of the movement of air parcels in the atmosphere. It can be calculated from a ratio of Raman lidar signals for water vapor and nitrogen using eqn [1] above.

Table 2 History of atmospheric measurements using Raman Lidar

<i>Event</i>	<i>Measurement date</i>	<i>Publication date</i>	<i>Researchers</i>	<i>Comments</i>
Earliest claimed measurement of Raman scattering (nitrogen) from the atmosphere	1966	1968	Cooney	
Measurement of Raman scattering from oxygen and nitrogen	1967	1967	Leonard	Definitive evidence of measurements provided by photographs of oscilloscope traces
Water vapor measurements:				
Cooney	1969	1970	Cooney	
Melfi	1969	1969	Melfi	
Mobile Raman lidar capable of remote measurement of pollutants in smoke stacks and automobile exhaust	1969, 1972	1969, 1972	Inaba and Kobayasi	Clear atmosphere measurements presented showing Rayleigh/Mie and Raman N ₂ , O ₂ , CO ₂ , H ₂ O. Polluted samples show various hydrocarbons along with liquid water.
Atmospheric temperature using vibrational–rotational Raman scattering	1971?	1971	Strauch <i>et al.</i>	Correlated temperature variations measured at the top of a 30-m tower and Raman backscatter
Atmospheric temperature using pure rotational Raman scattering	1973?	1973	Salzman and Cooney	Separated the anti-Stokes component of pure rotational scattering into two bands with differing temperature sensitivities
Remote measurement of water temperature	1976	1977	Leonard	Sea water temperature quantified at a depth of 1 m
Evolution of boundary layer water vapor	1977	1979	Pourney <i>et al.</i>	
Daytime water vapor measurements	1979?	1980	Renaut <i>et al.</i>	
Evolution of tropospheric water vapor	1985	1985	Melfi and Whiteman	First meteorologically meaningful measurements of tropospheric water vapor evolution
Aerosol extinction, optical depth	1990?	1990	Ansmann	
Stratospheric ozone in the presence of aerosols	1993?	1993	McGee <i>et al.</i>	Raman extension used to the differential absorption technique to permit quantification of ozone in the presence of stratospheric aerosols due to the eruption of Mount Pinatubo in 1991
Automated measurements of water vapor and aerosols	1998	1998	Goldsmith <i>et al.</i>	
Multi-wavelength determination of aerosol size and refractive index	1997	2000	Müller <i>et al.</i>	

During April 1994, the SRL participated in the first intensive field campaign sponsored by the US Department of Energy's Atmospheric Radiation Measurements (ARM) program at their northern Oklahoma Cloud and Radiation Testbed site. The SRL was operated for approximately 9 h on the night of 21 April during this field campaign. Figure 3 shows the SRL water vapor mixing ratio measurements (1 min summation of lidar data) compared with a radiosonde measurement made at 0535 UTC. The agreement is excellent. The figure indicates that the atmosphere was relatively well mixed up to an altitude of approximately 2 km. Above this altitude there was abrupt drying. Cirrus clouds were present during the entire evening as well.

A color image of water vapor evolution can be created by combining the vertical lidar profiles into a time–height image. The result is shown in Figure 4. This image reveals that the water vapor field showed variations in the regions below 2 km throughout the night. Furthermore, the dry slot located just above this mixed layer generally rose in altitude during the night from approximately 2.2 km at 0230 UTC to 2.7 km at 1100 UTC. Above the dry slot, the atmosphere was quite stable throughout the night.

Aerosols

The corresponding aerosol scattering ratio image from the night of 21 April 1994 is shown in Figure 5. The aerosol scattering ratio is used to quantify the ratio of aerosol to molecular scattering. It is defined as the ratio of the volume backscatter coefficients for total (molecular+aerosol) scattering to pure molecular scattering. The vertical scale has been expanded to 15 km in altitude, allowing the evolution of aerosols at all levels of the troposphere to be studied. In the aerosol field, the dry slot that was revealed in the water vapor field is seen as a decrease in the scattering ratio indicative of reduced aerosol loading. A comparison of the water vapor and aerosol images would tend to indicate that below a mixing ratio value of approximately 6.5 g kg^{-1} the aerosol scattering ratio was less than approximately 1.05. This creates the apparent increase in the dry slot at 0800 UTC as revealed in the aerosol scattering ratio image. Cirrus clouds formed in the upper troposphere on this night as well. The evolution of cirrus cloud structure can be studied over the course of the night using Raman lidar imagery as shown here. The bases of the cirrus clouds were as low as 9 km while the tops

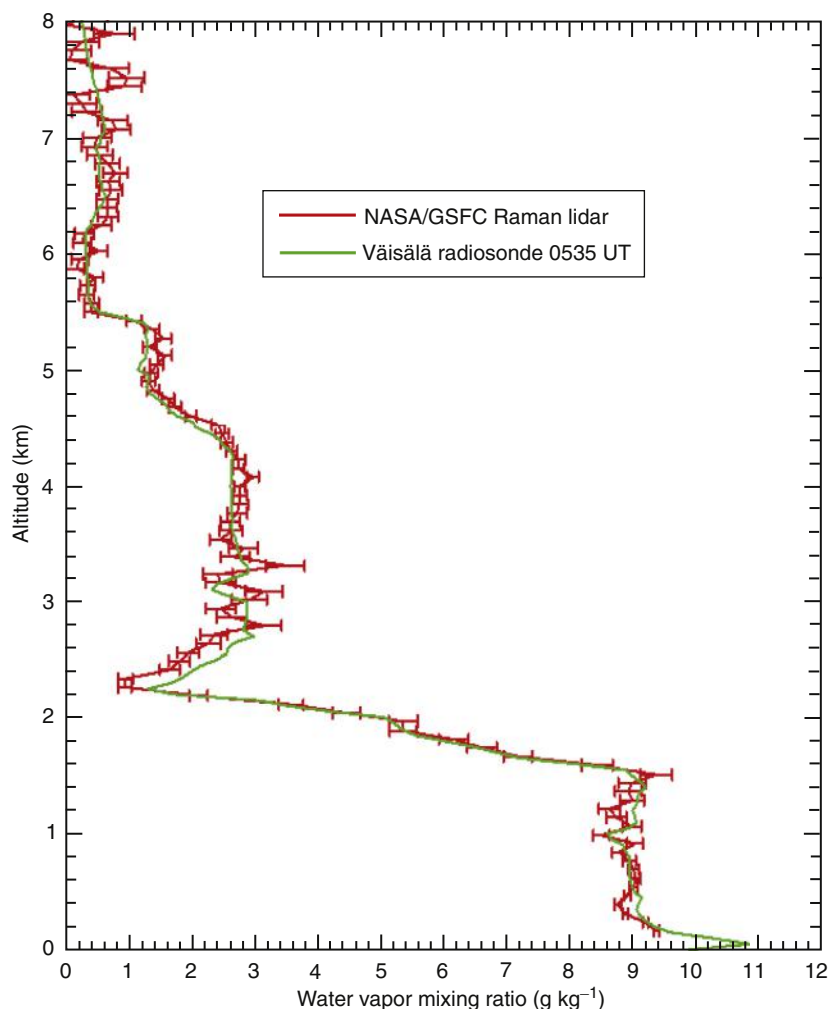


Figure 3 Comparison of Raman lidar measurement of water vapor mixing ratio and radiosonde. Measurements occurred 21 April 1994 at the DOE ARM CART site in northern Oklahoma. A 1-min summation of lidar data was used.

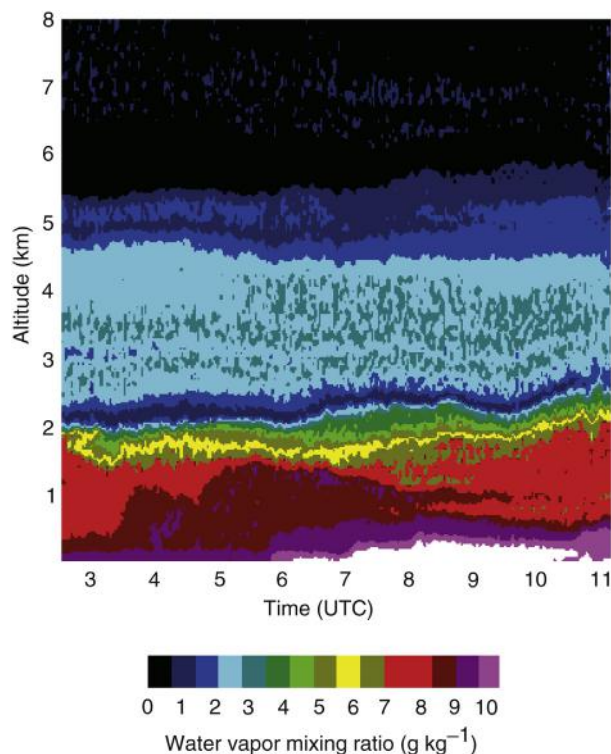


Figure 4 Color image of water vapor mixing ratio showing the evolution of water vapor in the troposphere. A thin, persistent dry slot rose in altitude during the night.

rose to approximately 12.5 km. As will be shown later the cirrus cloud optical depth can also be quantified using Raman lidar.

By analyzing the slope of the Raman nitrogen or oxygen signal, the extinction due to aerosols can be quantified. Aerosol optical depth can be obtained by integrating aerosol extinction with range. The Raman lidar is able to simultaneously quantify aerosol extinction and aerosol backscattering in the same atmospheric volume. When these quantities are studied along with the relative humidity (calculated from the Raman lidar water vapor mixing ratio using a temperature profile obtained from radiosonde), changes in aerosol optical properties due to hygroscopic growth of the aerosol particles can be studied. For example, **Figure 6** shows four simultaneously acquired images of water vapor mixing ratio, relative humidity, aerosol backscattering coefficient, and aerosol extinction.

The ratio of extinction to backscattering can be used to quantify aerosol growth versus relative humidity. A relative humidity value of 70% is generally taken to be the level above which aerosols can begin swelling. This can be revealed by the ratio of extinction to backscattering which will tend to increase as the aerosol grows in size.

Daytime Water Vapor

The Scanning Raman Lidar was deployed to the panhandle of Oklahoma in the United States for the first International H₂O Project held in May–June, 2002. Prior to the deployment, the SRL received technology upgrades to improve measurements of water vapor during the daytime where high solar

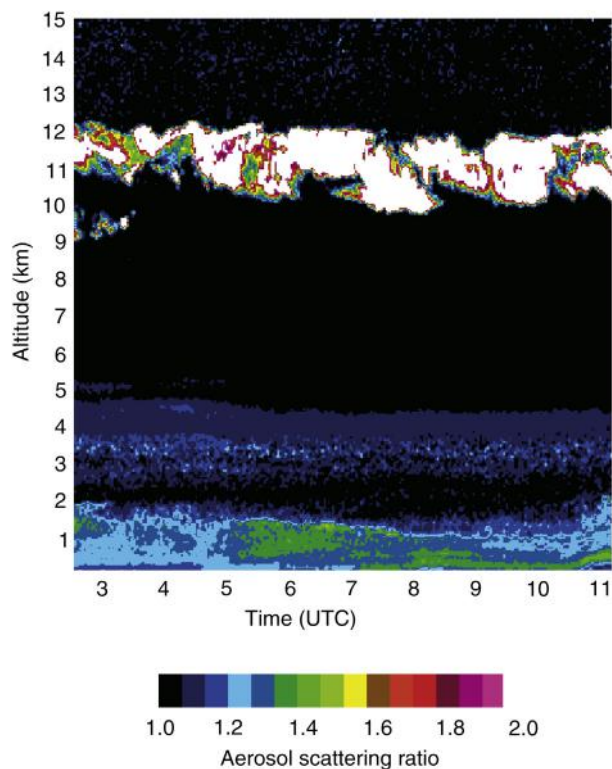


Figure 5 Color image of aerosol scattering ratio depicting the evolution of aerosol structure throughout the troposphere. Cirrus clouds were present during the night between the altitudes of 9 and 12.5 km. The dry region in the water vapor image is seen again here as a region of low aerosol scattering.

background makes measurement of the weak Raman signals more difficult than at night. In **Figure 7** is shown the water vapor image acquired on May 22, 2002 during a dryline event near the lidar site. This image is displayed with 3 min temporal and between 100 and 300 m vertical resolution and reveals convective plume structures in the boundary layer, which is seen to increase in height as the day progresses. Approximately 1 hour before sunset, which occurred at 0200 UT (shown as 26 in the image), the dryline became better defined resulting in a significant drying above ~1.5 km and increased moisture below.

Upper Tropospheric Water Vapor

Small errors in the quantification of water vapor at high altitudes can create large errors in radiative transfer calculations. Therefore ground-based monitoring of upper tropospheric water vapor is important in climate change studies. The same hardware upgrades to the SRL that permitted the improved daytime measurements of water vapor shown in **Figure 7** also permitted very high-quality upper tropospheric retrievals of water vapor during the nighttime. In **Figure 8** is shown a comparison of a 30-min summation of SRL water vapor mixing ratio data and a radiosonde measurement on the night of 5 December 2000. Excellent agreement is seen between the two sensors on this particular occasion, although radiosondes

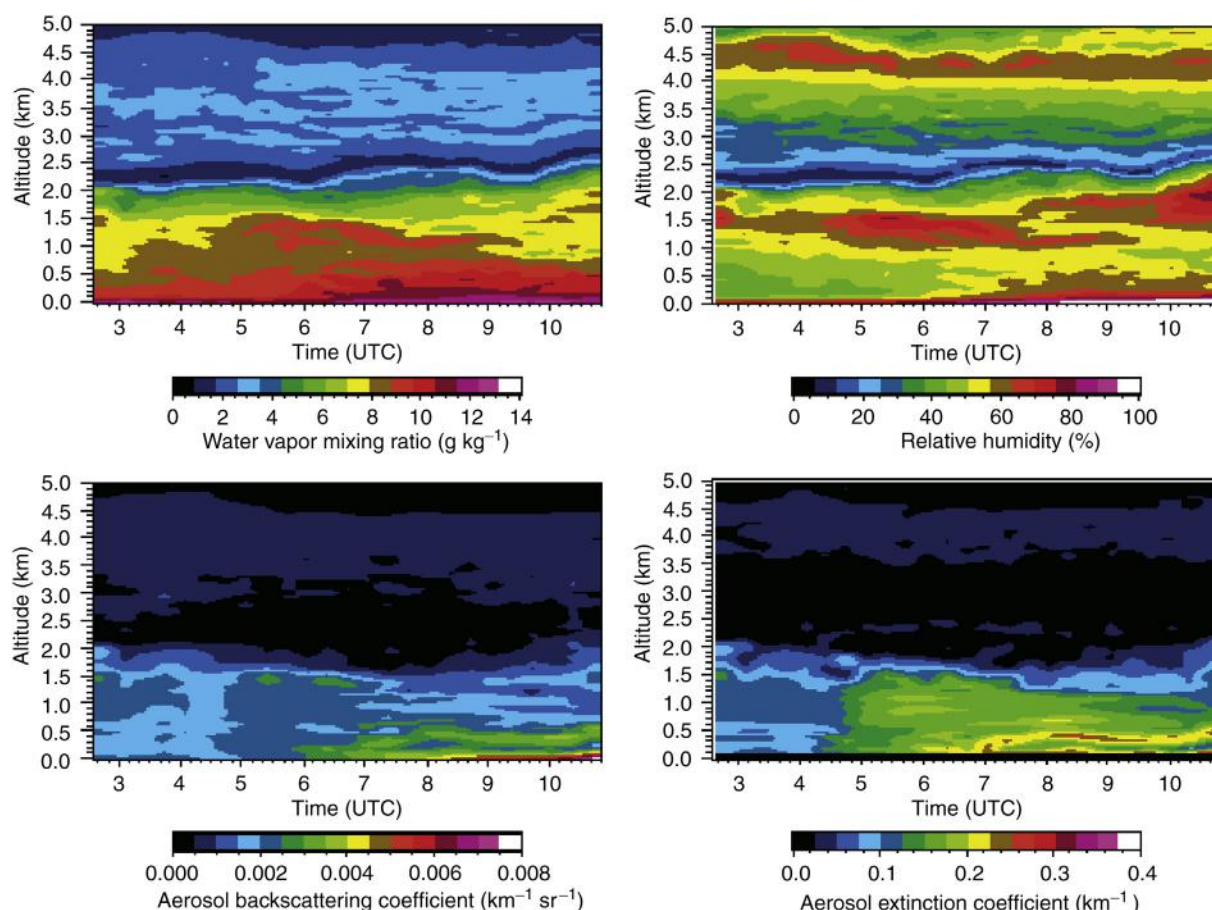


Figure 6 Four-panel display of the water vapor mixing ratio, relative humidity, aerosol backscattering ratio, and aerosol extinction. The bottom 5 km of the atmosphere is displayed so that changes in boundary aerosol properties can be revealed. Simultaneous studies of these parameters can reveal information about the growth of aerosols as a function of relative humidity.

can have difficulty quantifying the very small amounts of water vapor present in the cold upper troposphere.

Cirrus Cloud Optical Depth Measurements and Influence on Satellite Radiances

The Scanning Raman Lidar was deployed to Andros Island, Bahamas during August and September 1998 for the third Convection and Moisture Experiment (CAMEX-3). CAMEX-3 was a multi-instrument field campaign designed to improve hurricane tracking and intensification forecasting. During the period of August 21–25, hurricane Bonnie passed near Andros Island and influenced both the water vapor and the cirrus cloud environment. The SRL measurements of cirrus cloud acquired on August 23 indicated that the cirrus optical depths ranged from a high of ~ 0.7 to a low of ~ 0.005 . This range of optical depths provided a convenient dataset to test the sensitivity of satellite retrievals of precipitable water vapor to the presence of cirrus clouds.

In [Figure 9](#) are shown the SRL measurements of total atmospheric precipitable water and cirrus cloud optical depth from August 23 along with the corresponding GOES-8 (Geostationary Observational Environmental Satellite)

precipitable water retrievals. The lidar measurements indicate that the total precipitable water (TPW) changed relatively little during the measurement period. All significant variation in the retrieved TPW from GOES is attributed to the influence of cirrus. There is a strong correlation between cirrus cloud optical depth and the GOES-derived TPW. Furthermore, the GOES and lidar TPW measurements converge at the end of the data record where optical depths are very low as expected.

It is clear from [Figure 9](#) that increases in cirrus optical depth elevate the retrieved TPW. These results indicate that satellite radiances are noticeably affected for cirrus optical depths above approximately 0.005. Undetected cirrus cloud will create a consistent high bias in GOES satellite retrievals of TPW. Using the cirrus cloud detection criteria of the most recent International Satellite Cloud Climatology Project analysis indicates this bias will be up to 20% for cirrus clouds measured over water and 40% for cirrus clouds measured over land. It is apparent from the results presented here that satellite retrieval algorithms need to be able to detect the presence of cirrus clouds with IR optical depths as small as 0.005 in order to avoid significant influences on satellite radiances and thus potential errors in retrievals.

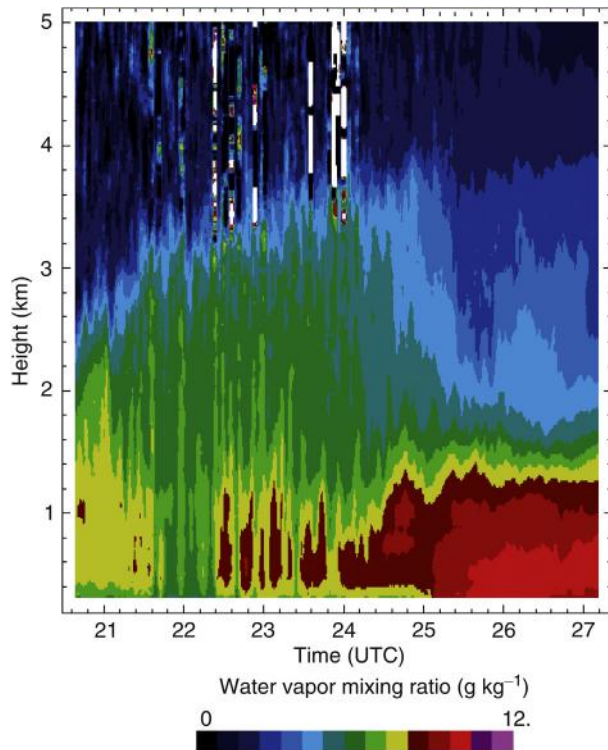


Figure 7 Water vapor mixing ratio measurements acquired on May 22, 2002 during the passage of a dryline near the SRL location in western Oklahoma. The boundary layer depth is seen to increase until ~2400 UT at which time the dryline influence became more noticeable. Sunset on this day was ~0200 UT (shown as 26 UT). The white stripes at for example 24 UT and 3.2 km are due to attenuation of the laser beam by clouds.

Future Developments: Airborne Raman Water Vapor Lidar

Despite the great success of Raman lidar technology for both daytime and nighttime measurements of water vapor from ground-based platforms, Raman lidar measurements from aircraft are very limited. No airborne Raman lidar system has been constructed to measure water vapor during both daytime and nighttime. To address this deficiency, the performance of an airborne Raman water vapor lidar was studied as a part of the NASA Instrument Incubator Program. A numerical Raman lidar model was constructed and used to study the anticipated measurements of this new system for water vapor conditions ranging from subtropical to arctic. The results demonstrate that a significant performance increase is realized by operating a Raman lidar looking downward from an aircraft compared to that same system looking upward from the ground. The lidar system simulated was based on a 15 W tripled Nd:YAG laser and a 0.6 m telescope. The optical field of view of the telescope is 0.25 mrad and the spectral widths of the water vapor and nitrogen interference filters are both 0.3 nm. In Figure 10 are shown the simulated profiles of water vapor mixing ratio measured by this airborne Raman lidar under both nighttime (A) and daytime (B) conditions for subtropical conditions during August (the radiosonde profile used was from Andros

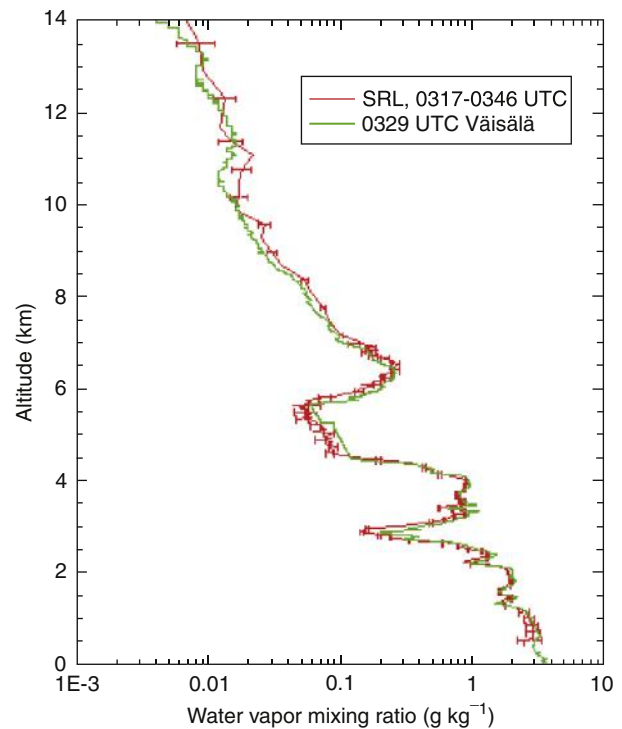


Figure 8 Nighttime profile of SRL water vapor mixing ratio and radiosonde measurement plotted to 14 km using a 30-min summation of lidar data. Upper tropospheric water vapor is very important in atmospheric radiation studies. The Raman lidar can provide long-term ground-based monitoring of this important quantity.

Island, Bahamas on 22 August 1998). The nighttime simulations used a 10 s summation while during the daytime the summation was 3 min. The daytime simulations assume a 0° solar zenith angle over water. The vertical smoothing applied to the profiles is as follows: For the nighttime case,

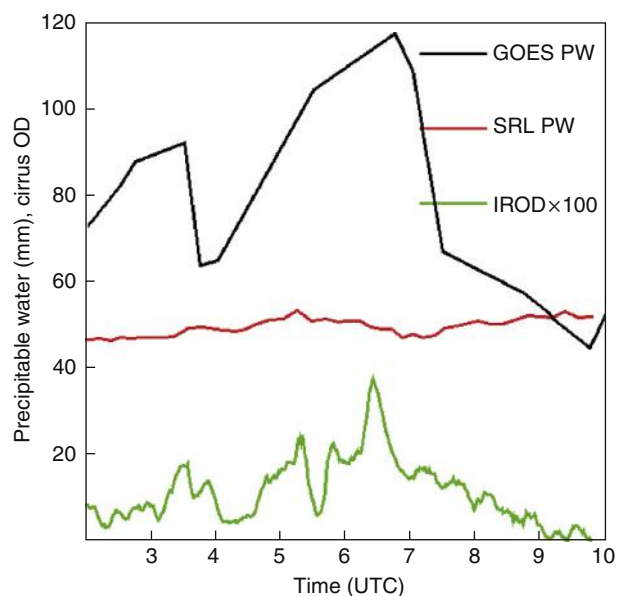


Figure 9 Cirrus optical depth-line profile.

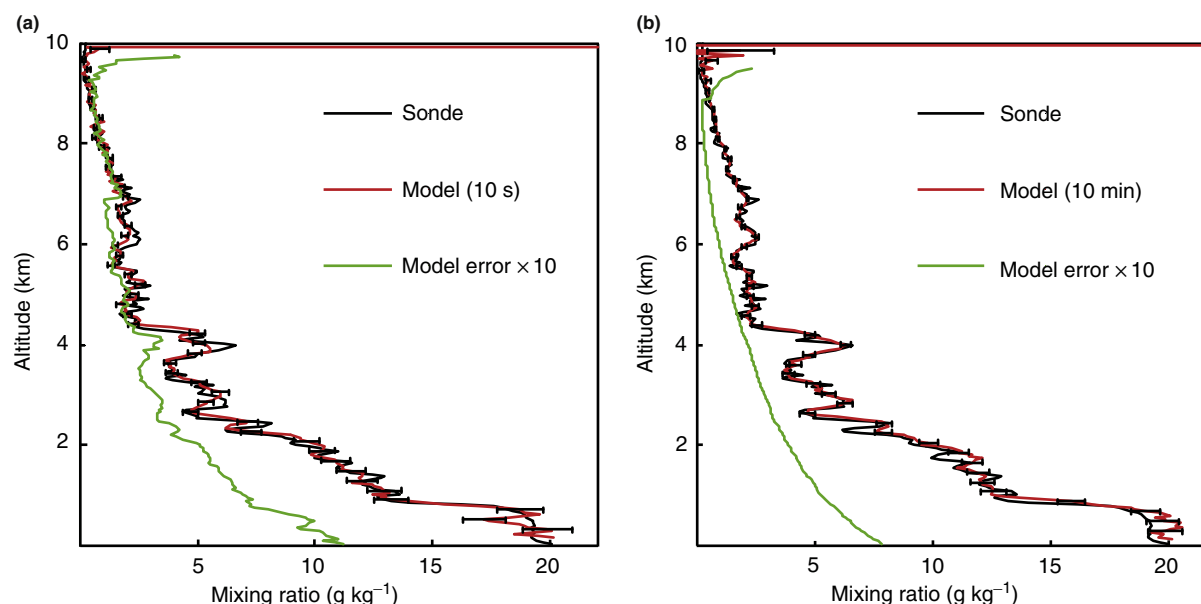


Figure 10 Simulations of the nighttime (a) and daytime (b) water vapor measurement performance of an airborne Raman lidar. A subtropical water vapor profile was assumed and the daytime simulations used a 0° solar zenith angle over a water surface. These simulations reveal that an airborne Raman lidar can provide high-quality water vapor measurements using as little as 10 s of integration time during the nighttime and typically 3 min during the daytime.

between the altitudes of 0–4 km, a 200 m smoothing is used, between 4 and 7 km, the smoothing is 120 m, and above 7 km, the smoothing is 40 m. The random error in the retrieval is less than 10% up to 9 km and closer to 5% in the very moist region near the surface. For the daytime case, the data are smoothed to 200 m for all altitudes. The modeled error is generally less than 5% except in the region between 5 and 6 km, where it is closer to 7%. In the lowest 2 km of the profile, the error is 3–4%. These profiles demonstrate that an airborne Raman lidar of the specifications modeled here would be capable of very high-quality profiles of water vapor under both daytime and nighttime conditions. Such a system would indicate a significant advance in airborne remote sensing since the additional Raman measurements that have been described here could be made simultaneously with the water vapor mixing ratio.

Summary and Conclusions

Raman Lidar technology has been used to measure a broad range of atmospheric phenomena including water vapor, cloud liquid water, water temperature, atmospheric density and temperature, aerosol backscattering and extinction, stratospheric ozone, and pollutants. Systems have been constructed with the ability to measure many of these parameters simultaneously. This ability to make numerous simultaneous lidar measurements is quite unique to the Raman technique due to the spectral shifting of the return signals. This has permitted the construction of systems such as the NASA/GSFC Scanning Raman lidar, which is capable of precise tropospheric measurements of water vapor along with measurements of aerosols and cirrus clouds. Liquid water

measurements are also possible simultaneously with those of water vapor and aerosols.

Raman lidar has a further attractive feature in its relative simplicity compared to competing techniques such as differential absorption lidar. This has permitted an automated Raman lidar to be developed by the Department of Energy. This system provides continuous automated measurements of water vapor and aerosols at the northern Oklahoma Cloud and Radiation Testbed Site. With improvements in technology, we can expect the future to bring several advances in Raman lidar technology and corresponding advances in atmospheric measurements. These include: (1) higher performance systems for ground-based research; (2) lower cost automated systems for precise characterization of the atmospheric state over extended periods; and (3) airborne systems combining water vapor, liquid water, aerosol scattering, extinction, and depolarization. Such systems and new analysis techniques should permit Raman lidar to remain one of the most powerful remote sensing techniques being applied to atmospheric studies.

Acknowledgements

Support from NASA and the Department of Energy in the development and scientific application of the Scanning Raman Lidar is gratefully acknowledged. In addition I would like to express my appreciation to all members of the Raman lidar group at NASA/GSFC for their efforts over the years.

See also: Lidar: Doppler; Resonance.

Further Reading

- Anderson, A. (Ed.), 1971. *The Raman Effect*. Marcel Dekker, Inc, New York.
- Ansmann, A., Riebesell, M., Weitkamp, C., 1990. Measurement of atmospheric aerosol extinction profiles with a Raman lidar. *Optics Letters* 15 (13), 746–748.
- Bowman, M.R., Gibson, A.J., Sandford, M.C.W., 1969. Atmospheric sodium measured by a tuned laser radar. *Nature* 22, 456.
- Cooney, J.A., 1968. Measurements on the Raman component of laser atmospheric backscatter. *Applied Physics Letters* 12, 40.
- Cooney, J.A., 1970. Measurement of the ratio of the vibrational Raman cross-section for H₂O vapor to the nitrogen vibrational–rotational Raman band. *Spectroscopy Letters* 3, 305–300.
- Derr, V.E., Little, C.G., 1970. A comparison of remote sensing of the clear atmosphere by optical, radio and acoustic radar techniques. *Applied Optics* 9, 1976–1992.
- Goldsmith, Blair, Bisson, Turner, 1998. Turnkey Raman lidar for profiling atmospheric water vapor, clouds, and aerosols. *Applied Optics* 37, 4979–4990.
- Leonard, D.A., 1967. Observation of Raman scattering from the atmosphere using a pulsed nitrogen ultraviolet laser. *Nature* 216, 142–143.
- Leonard, D.A., Caputo, B., Johnson, R.L., Hoge, F.E., 1977. Experimental remote sensing of subsurface temperature in natural ocean water. *Geophysical Research Letters* 4, 279–281.
- Little, R.G., Derr, V.E., Cupp, R.E., 1972. Atmospheric water vapor measurement by Raman lidar. *Remote Sensing of Environment* 2, 101–108.
- McGee, T.J., Gross, M., Ferrare, R., et al., 1993. Raman DIAL measurements of stratospheric ozone in the presence of volcanic aerosols. *Geophysical Research Letters* 20, 955–958.
- Melfi, S.H., Lawrence Jr., J.D., McCormick, M.P., 1969. Observation of Raman scattering by water vapor in the atmosphere. *Applied Physics Letters* 15, 295.
- Melfi, S.H., 1972. Remote measurement of the atmosphere using Raman scattering. *Applied Optics* 11, 1605–1610.
- Müller, D., Wagner, F., Wandinger, U., et al., 2000. Microphysical particle parameters from extinction and backscatter lidar data by inversion with regularization: experiment. *Applied Optics* 39, 1879–1892.
- Pourney, J.C., Renaut, D., Orszag, A., 1979. Raman-lidar humidity sounding of the atmospheric boundary-layer. *Applied Optics* 13, 1141–1148.
- Renaut, D., Pourney, J.C., Capitini, R., 1980. Daytime Raman-lidar measurements of water vapor. *Optics Letters* 5, 233–235.
- Schotland, R.M., 1966. Some observations of the vertical profile of water vapor by means of a laser optical radar. *Fourth Symposium on Remote Sensing of the Environment*. April.
- Strauch, R.G., Derr, V.E., Cupp, R.E., 1971. Atmospheric temperature measurement using Raman backscatter. *Applied Optics* 10, 2665–2669.
- Whiteman, D.N., Melfi, S.H., Ferrare, R.A., 1992. Raman Lidar System for measurement of water vapor and aerosols in the Earth's atmosphere. *Applied Optics* 31, 3068–3082.
- Whiteman, D.N., Melfi, S.H., 1999. Cloud liquid water, mean droplet radius and number density measurements using a Raman lidar. *Journal of Geophysical Research* 104, 31411–31419.
- Whiteman, D.N., Evans, K.D., Demoz, B., et al., 2001. Raman lidar measurements of water vapor and cirrus clouds during the passage of hurricane Bonnie. *Journal of Geophysical Research* 106, 5211–5225.
- Whiteman, D.N., Schwemmer, G., Berkoff, T., et al., 2001. Performance modeling of an airborne Raman water vapor lidar. *Applied Optics* 40, 375–390.

Resonance

CS Gardner, University of Illinois at Urbana-Champaign, Urbana, IL, USA

RL Collins, University of Alaska Fairbanks, Fairbanks, AK, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by CS Gardner, volume 3, pp 1212–1216, © 2003, Elsevier Ltd.

Synopsis

This article summarizes the principles and practices of the resonance lidar techniques used to measure constituents, temperatures, and winds in the middle and upper atmosphere (~70–150 km). The lidar equation is presented with a resonance solution that highlights the robustness of the technique due to the complete quantum mechanical description of scattering by atoms and molecules. The technique is illustrated with contemporary examples of high-resolution (~100 m, 1 min) measurements of mesospheric metals, temperatures, and winds. These measurements allow estimation of the constituent, heat, and momentum fluxes that are crucial for understanding nonlinear coupling and transport.

Lidar technologies, the optical counterpart to radar, have been especially important for studying the middle atmosphere from the middle stratosphere (from about 30 km altitude) to the lower thermosphere (to about 110 km altitude). This region is inaccessible to *in situ* probing from aircraft, balloons, and satellites. Only rocket probes and remote-sensing instruments can be used to measure the atmospheric composition and structure at these high altitudes. Lidars typically employ high-energy pulsed lasers, large optical telescopes, and range-gated photon counting detectors to derive atmospheric profiles. There are two primary types of lidar systems that are being used to probe the middle atmosphere. Rayleigh/aerosol systems employ backscattering from air molecules and ice particles to measure temperature, wind, and aerosol profiles up to about 80 km altitude. Resonance fluorescence lidars employ resonant backscattering from the atomic metal layers (i.e., Na, Fe, K, Ca, Ca⁺, and Li) and from the hydroxyl (OH) layer. These layers are usually found in the upper mesosphere and lower thermosphere in the 70–110 km altitude range, though recent observations have revealed events where Fe extends to 155 km. The metal layers are the product of meteor ablation while the OH layer is formed by chemical processes.

Because of its relatively high abundance, large resonant backscatter cross-section, and visible wavelength, Na has been the most widely studied species. In fact, the first metal resonance lidar measurements were made in 1968 just 2 years after the invention of the tunable dye laser. Existing systems based upon Na provide the highest resolution and most accurate wind and temperature measurements. However, rugged K and Fe systems have also been developed to make upper-atmosphere temperature measurements at remote sites and from research aircraft. Although OH airglow imagers are widely employed to study mesopause region dynamics, the OH lidar technique has been demonstrated only recently and few data have yet been collected using this new instrument.

System Design

Figure 1 is a block diagram of a typical resonance fluorescence lidar system. The pulsed laser is tuned to a resonant absorption line of the metallic species being probed. A small fraction of the output beam is directed to wavelength and energy monitors

and to a pulse detector, which triggers the data acquisition system. The laser pulse propagates into the atmosphere, where it is Rayleigh-scattered by air molecules at lower altitudes and then resonantly scattered by the metal atoms in the mesopause region. For resonance fluorescence scattering, individual metal atoms absorb photons in the laser beam and are excited to a higher energy state. Because the wavelength of the photons corresponds to an absorption line of the species, the probability of absorption is many orders of magnitude higher than the probability of Rayleigh scattering. In the low-density mesopause region, where the mean time between collisions is long compared to the fluorescence lifetime of the excited species, the excited atoms return to the ground state by spontaneously emitting a photon. In most cases, the emission wavelength is nearly the same as the excitation wavelength. Those photons that are backscattered into the receiving telescope and focused onto the photomultiplier tube are counted by the data acquisition system and then the data are recorded for later processing and analysis.

The detected photons are counted in small sequential subintervals of a few tenths of a microsecond duration, which correspond to scattering layers of a few tens of meters thickness. As in radar, the measured time delay between firing the laser and detecting the backscattered photons is used to compute the range to the scattering layer. To improve the signal level and minimize the effects of photon noise, the signal counts from several 100 laser pulses, transmitted over a period of several tens of seconds, can be integrated. This integration can be performed by the data acquisition system and/or in post-acquisition data processing to achieve a desired statistical significance in the measurement. The integrated data are then processed to yield profiles of the species concentration. If a narrowband laser is tuned to several different frequencies within the absorption line spectrum, it is also possible to infer the temperature and line-of-sight velocity of the species as well as its concentration.

Plotted in Figure 2 is an Na lidar signal profile obtained at the Starfire Optical Range near Albuquerque, NM. This lidar signal profile is a representative of resonance lidar signals from Na and other metals. The profile illustrates the molecular (Rayleigh) scattering between 30 and 60 km and the resonant scattering from Na between 80 and 105 km. The thin dense layers near 93 and 99 km altitude are meteor ablation trails that

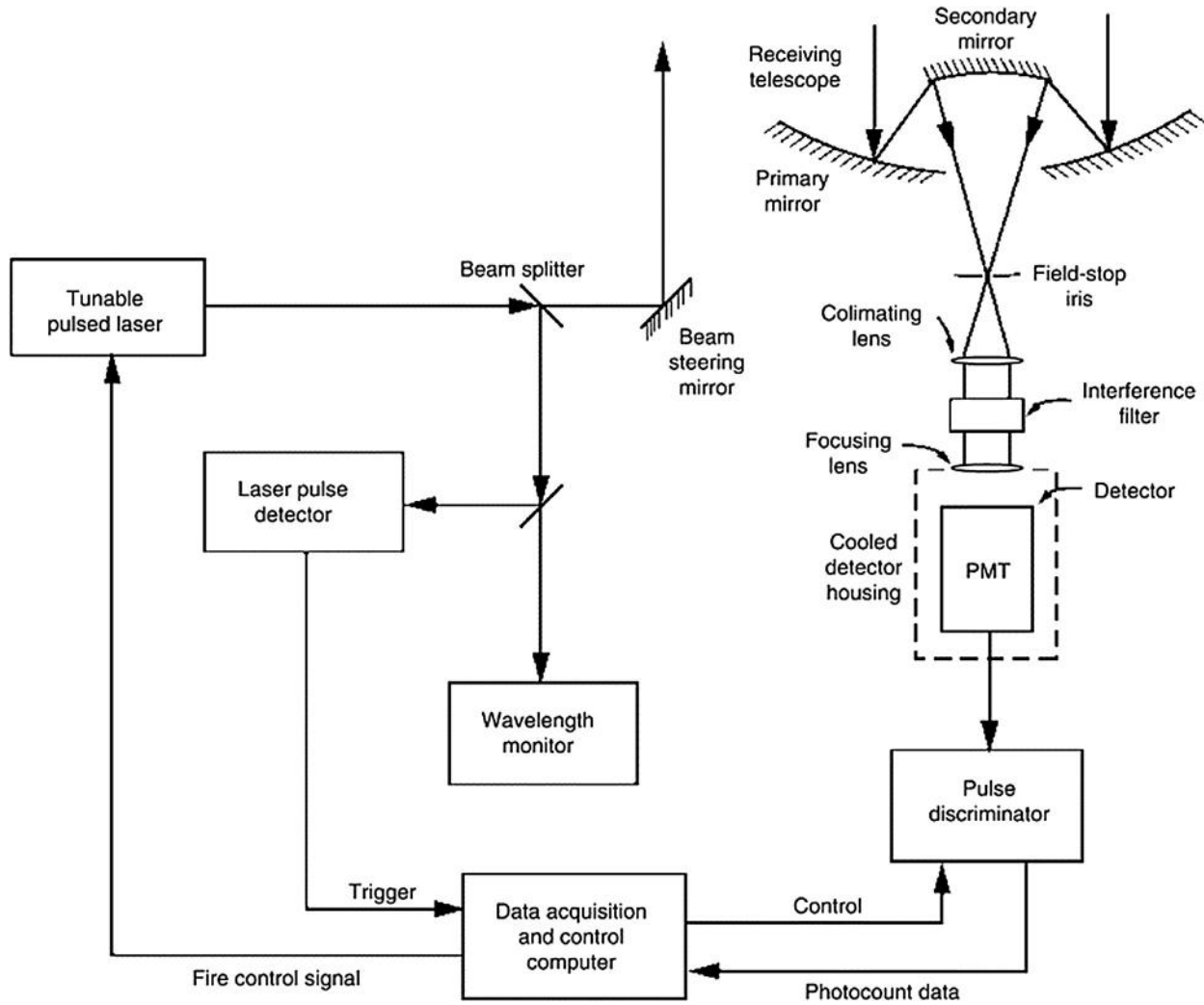


Figure 1 Block diagram of a typical resonance fluorescence lidar system.

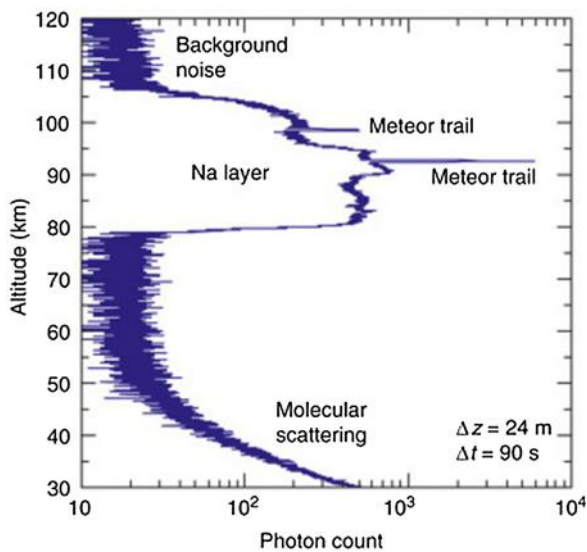


Figure 2 Na lidar photon count profile obtained at the Starfire Optical Range, NM on 17 November 1998 during the Leonid meteor shower.

drifted through the field-of-view of the lidar system during the 90 s observation period. The molecular scattered signal is proportional to atmospheric density, while the Na signal is proportional to the Na density.

Data Processing

For a zenith pointing lidar, the signal from a metallic scattering layer of thickness Δz located at an altitude z can be expressed in terms of the detected photon count as given by eqn [1].

$$N_m(\lambda, z) = \left(T_a^2 E^2(\lambda, z) \frac{P_L \Delta t}{hc/\lambda} \right) \times (\sigma_m(\lambda, T, v_R, \sigma_L) \rho_m(z) \Delta z) \times \left(\eta \frac{A_R}{4\pi z^2} \right) \quad [1]$$

In eqn [1], T_a^2 is the round-trip transmittance of the lower atmosphere; $E^2(\lambda, z)$ is the round-trip extinction associated with absorption by the metallic species; P_L is the average power of the laser; Δt is the accumulation period; h is Planck's constant; c is the vacuum speed of light; $\sigma_m(\lambda, T, \sigma_L)$ is

the effective absorption (or backscatter) cross-section of the resonance line, which is a function of laser wavelength λ , temperature T , radial velocity v_R , and laser linewidth σ_L ; $\rho_m(z)$ is the metallic species density; η is the overall optical efficiency of the receiving system including the detector; and A_R is the area of the receiving telescope. The first term in brackets on the right-hand side of eqn [1] is the effective number of transmitted photons; the second term is the probability that a transmitted photon is scattered by the metal layer of thickness Δz located at altitude z ; and the third term is the probability that a scattered photon is collected by the telescope and counted by the detector and data acquisition system. Because accuracy and useful resolution are related to signal strength, the most accurate, highest-resolution measurements are obtained with large power-aperture product ($P_L A_R$) lidars utilizing powerful lasers and large-diameter receiving telescopes. Typically, the laser average power levels are a few watts and the telescope diameters range from a few tens of centimeters to a few meters. Furthermore by taking the ratio of the resonance signal in the upper mesosphere and lower thermosphere (~ 70 – 110 km) to the Rayleigh signal in the stratosphere (~ 30 km), it is possible to remove the effects the

Table 1 Absorption cross-sections of several mesospheric metals

Metallic species	Nominal peak density (cm^{-3})	Resonance line wavelength (nm)	Peak absorption cross-section (10^{-16} m^2)
Ca	300	422.67	39
Ca ⁺	100	393.37	14
Fe	10 000	371.99	0.94
K	50	769.90	13
Li	3	670.78	7.3
Na	5000	589.00	15

atmospheric transmittance, optical efficiency, and laser power from the calculation of concentration, temperature, and wind.

A key strength of resonance lidar as a remote-sensing technique is that the scattering is described by the quantum mechanical interaction of light with atoms and molecules and unlike aerosol and cloud lidars is not dependent on particle shape, size, or orientation. The absorption cross-section can be calculated using the fundamental principles of quantum spectroscopy. The peak cross-sections for several mesospheric metals are listed in Table 1 along with the wavelength of the

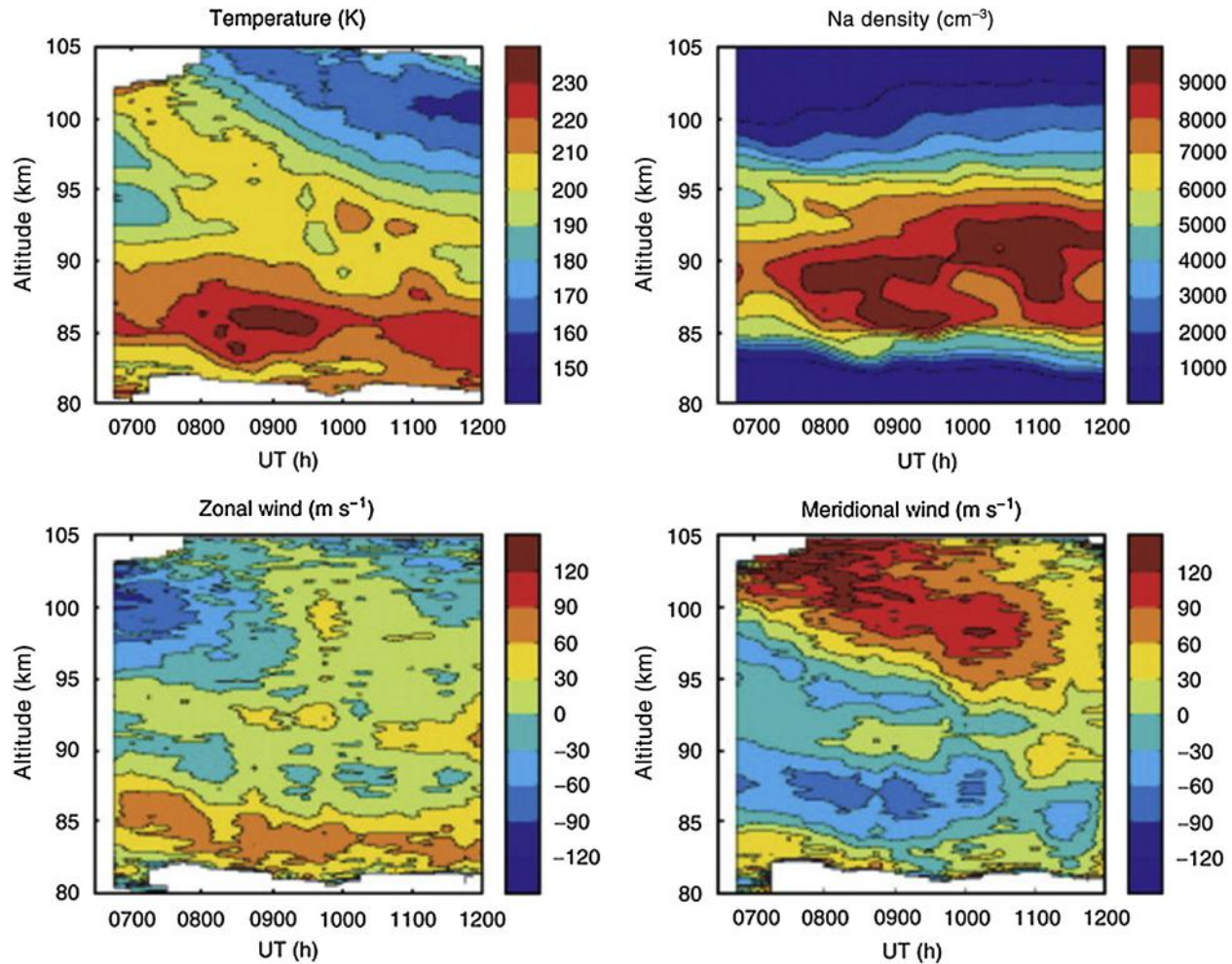


Figure 3 Profiles of temperature, Na density, zonal wind, and meridional wind obtained with an Na wind–temperature lidar using the 3.5-m steerable telescope at the Starfire Optical Range, NM 2 November 2000.

absorption line. These cross-sections were computed by assuming that $T = 200$ K and $\nu_R = 0$, that the laser linewidth is small compared to the thermally broadened width of the absorption line, and that the laser is tuned to the peak of the absorption line. The nominal peak densities of these species usually occur between 90 and 95 km. The nominal densities are also listed in Table 1. The probability of resonance fluorescence scattering is given by the second term on the right-hand side of eqn [1]. The probability of scattering from a layer of Na 100 m thick with a nominal density of 5000 cm^{-3} is $7.5 \cdot 10^{-4}$. By comparison, the probability of Rayleigh scattering from air molecules in a 100-m-thick layer at 90 km is $3 \cdot 10^{-9}$. This simple example illustrates the enormous advantage of resonance fluorescence lidars for making observations in the upper mesosphere.

Measurement Capabilities

Na resonance fluorescence lidars were the first to be developed to measure wind and temperature profiles. These parameters are derived from the backscattered Na signal by tuning a narrowband laser over the Na D₂ fluorescence spectrum near 589 nm to measure the width of the fluorescence spectrum, which is related to temperature, and to measure the center frequency, which is related to Doppler shift associated with the radial winds. Na density, temperature, and radial wind velocity can be determined by measuring the backscattered signal at as few as three different frequencies within the D₂ spectrum. To measure all three wind components it is necessary to scan the lidar sequentially among several zenith and off-zenith directions using a steerable telescope. While simple in concept, Na wind-temperature lidars employ sensitive ring dye lasers, pulsed dye amplifiers, and complex frequency-locking techniques to achieve the required tens of kilohertz frequency accuracy necessary for the wind and temperature observations. Accuracies of ± 1 K and $\pm 1 \text{ m s}^{-1}$ (radial wind) at resolutions of ~ 100 m and ~ 1 min are readily achievable with current laser and telescope technologies. Plotted in Figure 3 are the temperature, Na density, zonal wind, and meridional wind profiles measured with a steerable Na lidar at the Starfire Optical Range, NM. Narrowband K and Fe lidars have also been used to measure temperature profiles using robust solid-state ring lasers. More recently, a rugged broadband Fe lidar was used to measure the first middle and upper-atmosphere temperature profiles over the North and South Poles using the Boltzmann technique.

Resonance lidar technologies are now making crucial contributions to the studies of the chemistry and dynamics of the middle atmosphere through measurements of constituents, winds, and temperatures. The high-resolution of resonance

lidar measurements not just allows measurements of these atmospheric state variables but also allow measurements of the constituent, heat, and momentum fluxes that are crucial for understanding nonlinear coupling and transport in the atmosphere. Resonance lidar systems are now part of instrument suites that make it possible to profile winds and temperatures throughout the atmosphere. These measurements are making fundamental contributions to our knowledge of this region and to the impact of global climate changes. During the 40 years since the invention of the laser, lidar technologies have firmly established themselves as one of the key tools for probing the Earth's atmosphere.

See also: Lidar: Atmospheric Sounding Introduction; Backscatter; Differential Absorption Lidar; Doppler; Raman.

Further Reading

- Abo, M., 2005. Resonance scattering lidar. In: Weitkamp, C. (Ed.), *Lidar: Range-Resolved Optical Remote Sensing of the Atmosphere*, Springer Series in Optical Sciences, vol. 102. Springer, New York, NY, pp. 307–323.
- Bills, R.E., Gardner, C.S., She, C.Y., 1991. Narrowband lidar technique for Na temperature and Doppler wind observations of the upper atmosphere. *Optical Engineering* 30 (1), 13–21.
- Brinksma, E.J., Meijer, Y., McDermid, S., et al., 1998. First lidar observations of mesospheric hydroxyl. *Geophysical Research Letters* 23 (1), 51–54.
- Chu, X., Papen, G.C., 2005. Resonance fluorescence lidar for measurements of the middle and upper atmosphere. In: Fujii, T., Fukuchi, T. (Eds.), *Laser Remote Sensing*. Taylor and Francis, Boca Raton, FL, pp. 179–432.
- Chu, X., Yu, Z., Gardner, C.S., Chen, C., Fong, W., 2011. Lidar observations of neutral Fe layers and fast gravity waves in the thermosphere (110–155 km) at McMurdo (77.8° S, 166.7° E), Antarctica. *Geophysical Research Letters* 38. <http://dx.doi.org/10.1029/2011GL050016>.
- Gardner, C.S., 1989. Sodium resonance fluorescence lidar applications in atmospheric science and astronomy. *Proceedings of the IEEE* 77 (3), 408–418.
- Gardner, C.S., Papen, G.C., Chu, X., Pan, W., 2001. First lidar observations of middle atmosphere temperatures, Fe densities, and polar mesospheric clouds over the North and South Poles. *Geophysical Research Letters* 28 (7), 1199–1202.
- Grant, W.B., Browell, E.V., Menzies, R.T., Sassen, K., She, C.Y. (Eds.), 1997. *Selected Papers on Laser Applications in Remote Sensing*. SPIE Milestone Series, vol. 141. SPIE, Bellingham, WA, p. 662.
- Lautenbach, J., Höffner, J., 2004. Scanning iron temperature lidar for mesopause temperature observation. *Applied Optics* 43 (23), 4559–4563.
- Pfommer, T., Hickson, P., She, C.Y., 2009. A large-aperture sodium fluorescence lidar with very high resolution for mesopause dynamics and adaptive optics studies. *Geophysical Research Letters* 36. <http://dx.doi.org/10.1029/2009GL038802>.
- She, C.Y., Yu, J., Latifi, J., Bills, R., 1992. High-spectral resolution fluorescence light detection and ranging for mesospheric sodium temperature measurements. *Applied Optics* 31 (12), 2095–2106.
- Su, L., Collins, R.L., Krueger, D.A., She, C.Y., 2008. Statistical analysis of Doppler wind-temperature lidar measurements of vertical heat flux. *Journal of Atmospheric and Oceanic Technology* 25, 401–413.
- von Zahn, U., Höffner, J., 1996. Mesopause temperature profiling by potassium lidar. *Geophysical Research Letters* 23, 141–144.

Magnetosphere

GK Parks, University of Washington, Seattle, WA, USA

© 2003 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 3, pp 1229–1237, © 2003, Elsevier Ltd.

Introduction

Magnetospheres are new magnetic structures discovered during the space age by satellite-borne instruments that made possible physical measurements in distant regions previously not accessible. The first magnetosphere discovered was Earth's. Soon afterward, another discovery showed that space is not empty as once thought but is filled with ionized gases emanating from the Sun, stars, and other celestial bodies with high temperatures. It then became evident that magnetospheres are ubiquitous in space. What is a magnetosphere, how is it formed, and what are some of the important internal dynamics?

Let us first describe the environment in which magnetospheres are found. In our solar system, for example, the Sun's coronal atmosphere is hot, $\sim 10^6$ K, and dynamic, so it expands into space. The expanding solar coronal atmosphere is called solar wind and consists mostly of hydrogen ($\sim 95\%$ H^+) and helium ($\sim 5\%$ He^{2+}) ions and an equal number of electrons. Matter in the ionized state is called plasma and much of known matter in the Universe exists as plasmas. Since ionized matter is a good electrical conductor and magnetic fields decay slowly in conductors, it was immediately verified that the solar wind carries with it solar magnetic fields into space. Space is therefore permeated with magnetized plasma.

The solar wind is different from winds in the lower atmosphere because it is always blowing. All of the planets immersed in the solar coronal atmosphere are interacting with

it all the time. The electromagnetic (EM) interaction induces large-scale currents and forms magnetic cavities around magnetized planets. These cavities are called magnetospheres. Except for Mars and Venus, which do not have intrinsic magnetic fields, the planets in our solar system all have magnetospheres. This article will focus on planetary magnetospheres and emphasize features that are associated with Earth's magnetosphere, which has all of the elements to characterize a planetary magnetosphere (Figure 1).

The lower boundary of a planetary magnetosphere begins from that part of the atmosphere where ionized constituents play an important role in the dynamics of the upper atmosphere. For Earth, this boundary is located at ~ 100 km where the ionosphere begins (ionospheres are formed by the Sun's ultraviolet radiation). The ionosphere is therefore part of the magnetosphere. The outer boundary of the magnetosphere is called the magnetopause, and it separates the domains of the planetary magnetic field and the solar wind that blows outside it. Its location is determined by the pressure balance between the solar wind and the planetary magnetic field. On an average day Earth's magnetopause at local noon (subsolar point) crosses the equatorial plane at $\sim 10 R_E$ (average Earth radius, ~ 6367 km), and at $\sim 20 R_E$ in the dawn and dusk sectors. In the antisunward direction, the magnetosphere has a magnetic tail. The geomagnetic tail extends beyond $100 R_E$.

As the Sun's coronal atmosphere expands into space the wind speed increases, and near Earth's position it is ~ 400 $km s^{-1}$. This is faster than the speed of Alfvén waves in the

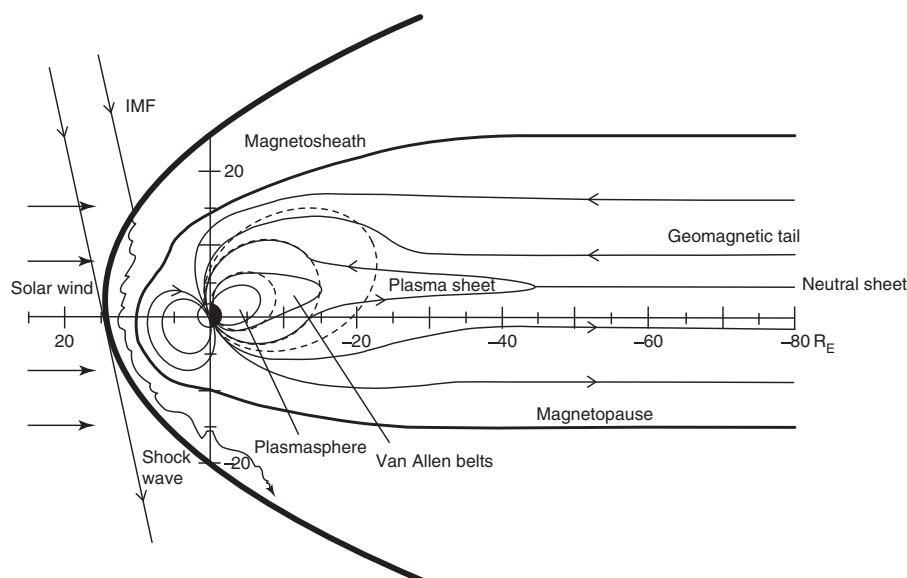


Figure 1 A sketch of Earth's magnetosphere in the noon–midnight plane. The dashed lines are the original dipole field. The solid lines are magnetic fields modified by external currents. IMF stands for interplanetary magnetic field, which is of solar origin. Major features of the magnetosphere are shown. (R_E earth radius.)

solar wind medium. Alfvén waves, named after the Swedish Nobel laureate Hannes Alfvén, are transverse magnetohydrodynamic waves travelling in the direction of the ambient magnetic field. As with objects that travel faster than the sound speed in the terrestrial atmosphere, a shock wave forms in front of Earth's magnetosphere. The Alfvén wave steepens nonlinearly and a shock forms as the magnetosphere plows through the super-Alfvénic solar wind. The Alfvén Mach number M_A is about 8, which makes the Earth's shock wave a strong shock. The shock wave is detached and separated from the magnetopause by $\sim 1 R_E$.

The region behind the shock wave is called *magnetosheath* and it extends to the *magnetopause*. The magnetosheath is a turbulent region permeated by large amplitude waves and hot particles that have been created in the shock formation. As in ordinary shocks, the solar wind stream energy is converted to the thermal energy. The solar wind speed in the downstream region just behind the shock is much reduced, $\sim 50 \text{ km s}^{-1}$. The solar wind speed picks up again further downstream.

Inside the magnetosphere, Van Allen radiation belts, named after their discoverer, James Van Allen, are found. These are divided into inner (ionosphere to $\sim 4\text{--}5 R_E$) and outer ($\sim 4\text{--}5 R_E$ to the magnetopause) radiation belts. The inner radiation belt energetic particles come from neutrons produced by cosmic rays that bombard the planet's atmosphere. Neutrons are unstable and have short lifetimes and they decay in flight into protons, electrons, and antineutrinos. The charged particles are captured by the planetary magnetic field. The primary source of Earth's energetic population in the inner radiation belt comes from these cosmic ray albedo neutron decay (CRAND) particles.

The source of the outer radiation belt particles is tied to solar wind and auroral disturbances which are dynamic. The outer magnetosphere is sometimes quiet, sometimes stormy, like the weather in the lower atmosphere of Earth. But unlike terrestrial weather, 'space weather' is driven by electrical forces powered by the disturbed solar wind connected to solar storms that produce flares and coronal mass ejections (CME). Spectacular auroral displays and intense radio emissions that occur in the polar regions of the planet are manifestations of space storms. The dancing lights of aurora are atmospheric emissions excited by precipitating energetic electrons that bombard the Earth's outer atmosphere. The radio emissions are generated by the unstable auroral particles.

Particles with millions of electron volt (MeV) energies are frequently produced during large space storms. These penetrating particles can impact on mankind as they can disrupt communication, impair satellite instrumentation and even cause damage to spacecraft. Another havoc is that currents of several million amps flow in the ionosphere during these storms. These ionospheric currents induce strong currents on the ground and have caused power outages in cities located in the auroral zone. A new practical goal of magnetospheric research is to learn to forecast space weather so as to forewarn when disruptive storms will occur and to predict which storms accelerate particles to MeV energies.

Particles in space rarely collide, because the density is so low and the mean free path so long. For example, in the outer magnetosphere there are ten thousand or so particles per cubic meter and in the solar wind a few million. The mean free path is

of the order of an astronomical unit, $1.5 \times 10^{11} \text{ m}$. This unique feature of the space system makes it different from other particle systems where collisions are prevalent.

Many space phenomena are driven by 'collisionless' processes that involve collective interactions through the long-range electromotive force. Very little is known about these collective processes. Even though it is known that the disturbed solar wind fuels auroral and magnetic storms, the actual mechanism of how the solar wind mass, momentum, and energy are transported across the magnetopause is not yet understood. The Earth's shock wave is a collisionless shock and, like ordinary shocks, it dissipates energy. But the collisionless dissipation mechanism is very different from the classical dissipation mechanism where viscosity is produced by the colliding particles. How viscosity is produced in a collisionless process is a fundamental problem yet to be solved.

Basic Equations

Understanding magnetospheres requires knowledge of how EM fields interact with charged particles and how large-scale currents are generated. The fundamental equations that describe the physics of magnetospheres are the Maxwell equations of electrodynamics,

$$\nabla \cdot \mathbf{B} = 0 \quad [1]$$

$$\nabla \times \mathbf{H} = \mathbf{J} + \frac{\partial \mathbf{D}}{\partial t} \quad [2]$$

$$\nabla \cdot \mathbf{D} = \rho \quad [3]$$

$$\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t} \quad [4]$$

$\mathbf{B}$ is magnetic induction and is related to the magnetic field intensity $\mathbf{H}$ by the constitutive relationship $\mathbf{B} = \mu_0 \mathbf{H}$, where μ_0 is the magnetic permeability of free space and equal to $4\pi \times 10^{-7} \text{ Hm}^{-1}$, $\mathbf{D}$ is the electric displacement vector related to the electric field $\mathbf{E}$ by the constitutive relationship $\mathbf{D} = \epsilon_0 \mathbf{E}$ where ϵ_0 is the dielectric constant of free space equal to $8.85 \times 10^{-12} \text{ F m}^{-1}$, ρ is the charge density and $\mathbf{J}$ is the current density. The equations for ρ and $\mathbf{J}$ are given by

$$\rho = \sum_{\Delta V} \frac{q_k}{\Delta V} \quad [5]$$

$$\mathbf{J} = \sum_{\Delta V} \frac{q_k \mathbf{v}_k}{\Delta V} \quad [6]$$

where q_k and $\mathbf{v}_k$ are charge and velocity of the k th particle. The summation is carried over a suitably chosen small volume ΔV .

The velocity of the k th particle is obtained from the Lorentz equation of motion,

$$m_k \frac{d\mathbf{v}_k}{dt} = q(\mathbf{E} + \mathbf{v}_k \times \mathbf{B}) \quad [7]$$

There are as many equations of motions as there are particles and they are coupled through the electromagnetic fields. Equations [1], [2], [3], [4], [5], [6] and [7] plus the constitutive relationships define a system of charged particles and EM fields. The physics of magnetosphere studies the science of

large-scale electromagnetic dynamics with an extremely large number of particles, or alternatively it can be viewed as a branch of statistical physics of charged particles driven by electromagnetic forces.

Steady-State Magnetosphere

For an understanding of how magnetospheres are formed, and of the physics of the interaction that induces external currents strong enough to deform and modify the planetary dipole field and excite the dynamic activity inside the magnetosphere, eqns [1], [2], [3], [4], [5], [6] and [7] need to be solved self-consistently. But this is not yet possible. Even though the density in space is very small, a magnetosphere occupies a large volume and particles are interacting through the long-range electromagnetic field.

Our present knowledge of the magnetosphere is based primarily on the synthesis of various pieces of magnetospheric elements that have been studied. A quiet-time picture of the magnetosphere based on time-independent formulation is given below. This simple model is a reasonable starting point for describing the more real magnetospheres.

Magnetic Field

The equations that govern the time independent magnetic field are

$$\nabla \cdot \mathbf{B} = 0 \quad [8a]$$

$$\nabla \times \mathbf{H} = \mathbf{J} \quad [8b]$$

The solution of these equations is

$$\mathbf{B}(\mathbf{r}) = \frac{\mu_0}{4\pi} \int \frac{\mathbf{J}(\mathbf{r}') \times (\mathbf{r} - \mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|^3} d^3 r' \quad [9]$$

where use was made of $\mathbf{B} = \mu_0 \mathbf{H}$. Equation [9] is Biot-Savart's law and it states that given the current density $\mathbf{J}$ at $\mathbf{r}$, we can obtain the magnetic field $\mathbf{B}$ everywhere, or, given $\mathbf{B}$, we can invert the equation and obtain the source current $\mathbf{J}$ responsible for the magnetic field.

Equation [9] is difficult to use for the magnetosphere as a whole, because the particles contributing to the source term $\mathbf{J}$ are not completely understood and they have not been measured for all regions. Also, spacecraft-borne magnetometers have measured $\mathbf{B}$ over a large region of the magnetosphere but the magnetosphere is dynamic and single-point measurements made at different times are not easily related.

Information on $\mathbf{J}$ can be obtained if $\nabla \times \mathbf{B}$ can be measured. This requires simultaneous measurements with identical instruments from multiple spacecraft that can measure gradients of the magnetic field. This is one of the primary goals of the Cluster mission to be launched in the summer of 2000 by the European Space Agency (ESA). It is anticipated that Cluster experiments will definitely improve our knowledge of magnetospheric currents.

The Dipole Field

We begin with the dynamo current interior to the solid Earth, which is the source of the geomagnetic field. The dipole field is

a good approximation for describing the relatively stable magnetic field in the lower region of the magnetosphere. The dipole configuration also provides a standard of reference for many other planets and celestial bodies.

Let the dipole moment of the planet be given by $\mathbf{M}$. In regions outside the planet, there are no currents, hence $\mathbf{J} = 0$. We can then let $\mathbf{B} = -\nabla \Psi$, where Ψ is the magnetic scalar potential of a dipole given by

$$\Psi = \frac{\mu_0}{4\pi} \mathbf{M} \cdot \nabla \frac{1}{r} \quad [10]$$

Assume that Earth has a centered dipole moment $\mathbf{M}$ (Figure 2). In spherical coordinate system with $\mathbf{M} = -M\hat{z}$ (—for Earth), the three components of the dipole magnetic field are

$$\begin{aligned} B_r &= -\frac{\mu_0}{2} \frac{M}{\pi} \frac{\sin \lambda}{r^3} \\ B_\lambda &= -\frac{\mu_0}{4} \frac{M}{\pi} \frac{\cos \lambda}{r^3} \\ B_\phi &= 0 \end{aligned} \quad [11]$$

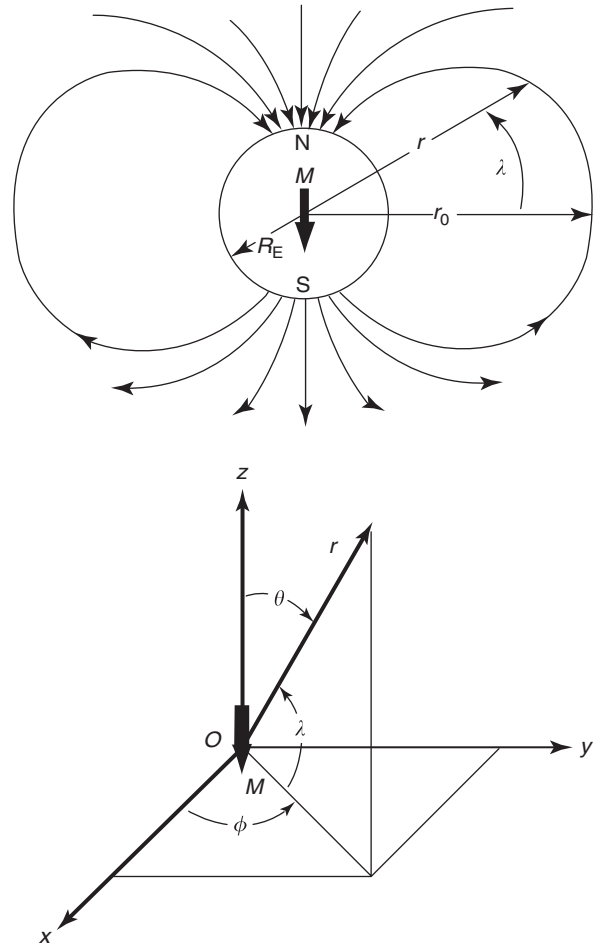


Figure 2 The top diagram shows the contours of the planetary dipole magnetic field. The magnetic moment for Earth points from north to south. The bottom diagram shows the Earth-centered coordinate system in which the dipole field is defined.

where for Earth $M = 8 \times 10^{22} \text{ A m}^2$, λ is the latitude ($\lambda = 0$ at the Equator) and r is the radial distance. $B_\phi = 0$ because the centered dipole is symmetric about ϕ . The magnitude of the field at (r, λ) is obtained from [11] and is given by

$$B(r, \lambda) = \frac{\mu_0}{4\pi} \frac{M}{r^3} (1 + 3 \sin^2 \lambda)^{1/2} \quad [12]$$

which shows r^{-3} dependence of the dipole field strength. A dipole has the strongest field at the pole ($\lambda = \pi/2$) and the smallest on the Equator ($\lambda = 0$). Another useful relation is the equation of the locus of a dipole field given by

$$r = r_0 \cos^2 \lambda \text{ and } \phi = \phi_0 \quad [13]$$

Here r_0 and ϕ_0 are distance and longitude of the line of force at $\lambda = 0$.

The dipole approximation is fairly good until about 4–5 R_E from the Earth. Further out, the magnetic field begins to depart as the other current sources become important.

Magnetopause

Consider now the dayside boundary of the magnetosphere. For simplicity, consider the boundary to be a plane and assume the solar wind consists of electrons and protons. The planetary magnetic field is in vacuum (no particles in the magnetosphere) and the solar wind is not magnetized (no interplanetary magnetic field). The solar wind particles that impinge on this boundary are deflected by the Lorentz force. This creates a boundary current running from east to west and modifies the planetary field.

In Figure 3, the total magnetic field B_t produced by the current just inside the boundary can be estimated as

$$B_t = B_d + B_s \quad [14]$$

where B_d is the dipole field of the planet at that point and B_s is the surface magnetic field produced by the magnetopause current. Just outside the boundary,

$$B_t = B_d - B_s = 0 \quad [15]$$

Therefore,

$$B_t = 2B_d \quad [16]$$

This result states that the current at the magnetopause will produce a magnetic field whose intensity is twice the value of the undistorted dipole field at that point. It is based on an ideal model which stipulates the planetary field to totally vanish inside the solar wind.

The magnetopause boundary separates the two domains completely and there is no normal component of the magnetic field in this model. This magnetosphere is essentially the model proposed by S. Chapman and V. C. A. Ferraro in 1931 to explain terrestrial magnetic storms. This magnetosphere is 'closed' and particles cannot enter the magnetosphere. If the boundary supports a normal component of the magnetic field, it can lead to an 'open' magnetosphere which allows the particles to enter the magnetosphere from the solar wind (see 'Dynamic Magnetosphere' below).

If the boundary is curved then the same procedure can be used, but now currents that give curved field configurations need to be taken into account. Another feature not incorporated

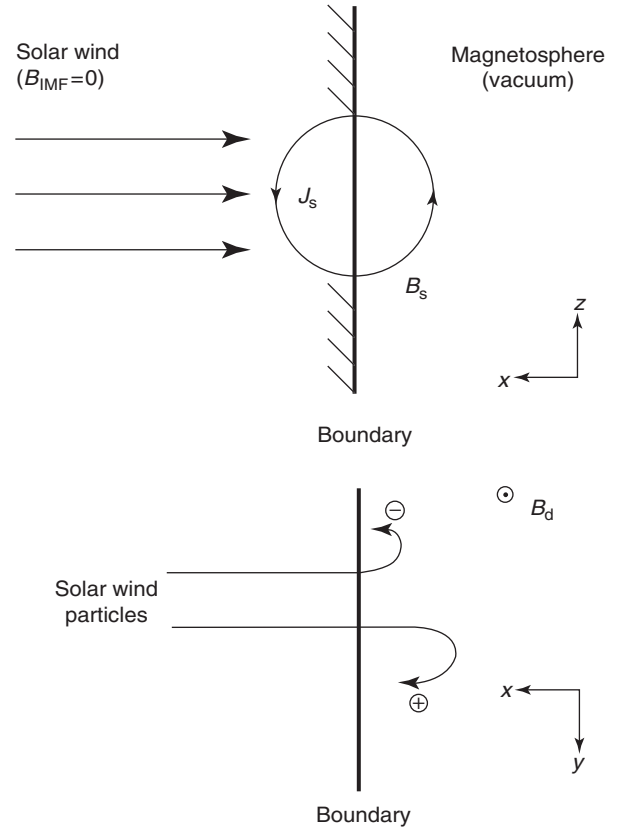


Figure 3 The top sketch shows a boundary in the noon–midnight plane that separates the solar wind and the magnetosphere. The current at the boundary is out of the page. The bottom sketch shows how the current is set up by the turning around of the solar wind particles due to the Lorentz force.

in this simple pedagogic model is that the real magnetopause includes a boundary layer. How boundary layers can be produced in collisionless plasmas is still not understood.

We can estimate where the outer boundary is located. Let the solar wind be specularly reflected at the boundary. The transfer of momentum per particle per collision is $2mV_{sw} \cos \xi$, where m is the mass of H^+ , V_{sw} the velocity of the solar wind, and ξ the angle of incidence. The number of particles striking a unit area of the boundary per second is $NV_{sw} \cos \xi$, where N is the number density of the solar wind. Thus, the total energy density of the particles perpendicular to the surface is $2NmV_{sw}^2 \cos^2 \xi$. At the magnetopause boundary, we require a balance between the solar wind and planetary magnetic energy densities. Hence,

$$2NmV_{sw}^2 \cos^2 \xi = B^2/2\mu_0 \quad [17]$$

Typical values are $V_{sw} = 400 \text{ km s}^{-1}$ and $N = 5 \times 10^6 \text{ m}^{-3}$. Equation [17] then yields $B \approx 70 \text{ nT}$ for the subsolar point, $\xi = 0$. From equation [16], we see that in the absence of the solar wind, the undistorted dipole field has a value approximately $\sim 35 \text{ nT}$. The surface field of Earth is $\sim 0.32 \times 10^{-4} \text{ T}$. Scaling this as $1/r^3$, we deduce the location for the magnetopause to be $\sim 9.2 R_E$. Observations show the subsolar magnetopause position varies typically between $\sim 9 R_E$ and $11 R_E$.

Geomagnetic Tail

The solar wind imparts momentum on the planetary magnetic field and creates a current such that when superposed on the dipole field, the resulting field on the antisunward direction has a tail-like geometry (Figure 1). The equation $\nabla \times \mathbf{H} = \mathbf{J}$ shows that a line or sheet current in the x -direction with the current flowing from dawn to dusk in the noon–midnight plane is needed. Several functional models have been proposed to account for the tail geometry. One model of this current is given by

$$J_y = \frac{B_0}{\mu_0 L} \sec h^2 \frac{z}{L} \quad [18]$$

where B_0 is the magnitude of the magnetic field at the outer boundary and L the half-thickness in the z -direction. This current is uniform in the y -direction but has a z -dependence. The magnetic field deduced from eqn [18] yields

$$B_x = B_0 \tanh \frac{z}{L} \quad [19]$$

The magnetic field is directed along the x -direction and increases with z . The magnetic field vanishes at $z = 0$. This is the magnetic field free line (neutral line) that results from the fields above and below that point in opposite directions. While functional models provide a useful picture of how the tail might be formed, the structure of the real tail current is very complicated. Although it is attributed to the solar-wind–geomagnetic-field interaction, how the current is produced and maintained is still unknown.

Electric Field

Understanding the origin of electric fields in magnetospheres is important because they can change particle energies and also alter their trajectories. For example, in the presence of an electric field perpendicular to the direction of the magnetic field, the particles drift across the magnetic field (further discussion below).

The two Maxwell equations governing the electric field for static magnetospheres are

$$\nabla \cdot \mathbf{D} = \rho \quad [20a]$$

$$\nabla \times \mathbf{E} = 0 \quad [20b]$$

The electric field can be defined in terms of the scalar potential ϕ , $\mathbf{E} = -\nabla\phi$, which inserted in eqns [20a] and [20b] yields Poisson's equation,

$$\nabla^2 \phi = \rho \quad [21]$$

whose solution is

$$\phi(\mathbf{r}) = \frac{1}{4\pi\epsilon_0} \int \frac{\rho(\mathbf{r}') d^3\mathbf{r}'}{|\mathbf{r} - \mathbf{r}'|} \quad [22]$$

where $\mathbf{r}'$ is the location of the charge density and the potential is evaluated at $\mathbf{r}$. This is the Coulomb potential that results from the charge density ρ . Electric field is obtained by taking the gradient of ϕ . Although magnetospheres are populated by charged particles, free charges ρ are not maintained (charges do not accumulate in good conductors). Thus, eqn [22] is not useful when considering large-scale electric fields in space.

Large-scale electric fields in magnetospheres are produced mainly by inductive effects. Michael Faraday in 1831 showed that an electromotive force (e.m.f.) is induced when magnetic flux changes in time or in space. In the magnetosphere, the inductive field comes from motions of a plasma across a magnetic field or rotation of a magnetized planet through a plasma medium. These motions induce a motional e.m.f., which is the primary source of large-scale electric fields in space.

Motional Electric Fields

Electric and magnetic fields are measured on moving platforms through a plasma medium which itself may be in motion. The relationship of electric and magnetic fields in the rest and moving frames of references is given by the Lorentz transformation. For linear motions and nonrelativistic case ($V/c \ll 1$), the relations are

$$\mathbf{E}' \approx \mathbf{E} + \mathbf{V} \times \mathbf{B} \quad [23a]$$

$$\mathbf{B}' \approx \mathbf{B} \quad [23b]$$

Here the prime (') denotes a moving frame and $\mathbf{V}$ is the velocity of that frame relative to the rest frame (unprimed) and c is the speed of light. Magnetic fields measured in the two frames are nearly equal, but electric fields are different. Thus, it is necessary to specify the coordinate frame in which the measurements are made.

An important result of the Lorentz transformation is that if the velocity of the moving frame is given by

$$\mathbf{V} = \frac{\mathbf{E} \times \mathbf{B}}{B^2} \quad [24]$$

then $\mathbf{E}' = 0$ and no work will be done on the particles in that moving frame. This has important application for plasma systems in motion. Consider, for example, the motion of an arbitrarily shaped system of charged particles in the magnetosphere. Lorentz's result states that if the motion of this plasma is given by eqn [24], then $\mathbf{E}' = 0$. Thus, the electric field vanishes in the moving frame which is also the plasma frame of reference. Equation [23a] then states

$$\mathbf{E} = -\mathbf{V} \times \mathbf{B} \quad [25]$$

The electric field in the rest frame is related to the vector product of the velocity of the plasma system and the magnetic field.

Equations [24] and [25] are equivalent expressions. These equations state that if there is an electric field, the plasma will move. Equivalently, if the plasma is moving, there is an electric field. In a manner analogous to the convective motion in fluids which arises to equalize the nonuniform temperature, the plasma motion arises to transform away the electric field (charges) in the moving frame (plasma frame) because free charges cannot be maintained in good conductors. In that sense, the plasma motion in eqn [24] is referred to as a convective motion.

If the coordinate frame is spinning, this frame is not an inertial frame and the Lorentz transformation theory is not valid. General relativity effects must be taken into account. We state only the results here. The relationships of EM fields are given by

$$\mathbf{E}'' = \mathbf{E} + \mathbf{V} \times \mathbf{B} \quad [26a]$$

$$\mathbf{B}'' = \mathbf{B} \quad [26b]$$

where the double prime (") denotes the spinning frame of reference, $\mathbf{V} = \boldsymbol{\omega} \times \mathbf{r}$ and $\boldsymbol{\omega}$ is the angular frequency vector. Equations [23a], [23b], [26a] and [26b] have the same form, except that eqns [26a] and [26b] are exact to all orders in V/c , whereas eqns [23a] and [23b] are only approximate.

Electric Field at the Magnetopause

In closed magnetospheres, Maxwell equations require the tangential component of the electric field to vanish at the boundary, in addition to the requirement of the absence of a magnetic normal component. The presence of a tangential electric field will allow particles to drift across the boundary which is not permitted in closed models. This contrasts the open model which invokes a magnetic field merging process (see 'Dynamic Magnetosphere' below). In closed models, particles can cross the boundary, for example, by diffusive mechanisms.

Electric Field in the Plasma Sheet

The geomagnetic tail is populated by plasmas of the solar wind and the ionosphere. When these plasmas move, we can apply the above results to study the behavior of the electric field. Although the motions are generally turbulent, we will assume the motion is laminar. Consider an observer on the equatorial plane at rest observing the plasma moving toward the Earth. Equations [23a] and [23b] can be used to estimate the magnitude and direction of the electric field. For example, in the noon-midnight meridional plane, $\mathbf{V}$ is earthward and $\mathbf{B}$ is upward, and thus the electric field $\mathbf{E}$ points from dawn to dusk. The dawn-dusk electric field measured during quiet solar wind and geomagnetic field is $\sim 0.3 \times 10^{-3} \text{ V m}^{-1}$. The origin of this large-scale dawn-dusk field is thought to be the solar wind, but the mechanism for establishing this field across the tail of the magnetosphere is not known.

Electric Field in the Plasmasphere

The Earth's rotation is important for particles in the lower region of the magnetosphere. This region, dominated by the dipole field and populated mostly by ionospheric plasmas, is called plasmasphere. The plasmasphere starts from the ionosphere and extends out to about 4–5 R_E in the equatorial plane. It thus includes the inner radiation belt. The boundary of the plasmasphere is called plasmopause. Outside this boundary is the outer radiation belt.

The rotation of the planet induces an electric field in the plasmasphere. Consider an observer at rest at position $\mathbf{r}$ on the equatorial plane. Noting the plasma is rotating with the planetary angular frequency they apply eqns [26a] and [26b] to calculate the induced electric field. Since the electric field in the rotating frame (plasma frame) vanishes, they use

$$\mathbf{E} = -(\boldsymbol{\omega} \times \mathbf{r}) \times \mathbf{B} \quad [27]$$

to find the direction and magnitude of $\mathbf{E}$. The result shows the electric field is radially outward. Induced electric field due to

the rotation of the planet is also known as the corotational electric field. The Earth's corotational electric field is $\sim 0.5 \times 10^{-3} \text{ V m}^{-1}$ at ionospheric heights.

Van Allen Radiation Belts

The motion of a charged particle is governed by the Lorentz equation of motion given in eqn [7]. We can study the behavior of a single particle using this equation, neglecting the presence of other particles. This test particle approach will give us a sense of how a charged particle moves around in magnetospheres and how Van Allen radiation belts are formed. Since we are ignoring the collective effects of interacting particles, information on the dynamic behavior of magnetospheric particles in aurorae is not revealed.

Consider first the motion of a particle in a time-independent magnetic field and assume there is no electric field ($\mathbf{E} = 0$). The equation of motion is then $m d\mathbf{v}/dt = q\mathbf{v} \times \mathbf{B}$. In inhomogeneous magnetic fields, $\mathbf{B} = \mathbf{B}(\mathbf{r})$, this differential equation yields three types of motion. The first is the cyclotron motion around the magnetic field. The cyclotron frequency of this circular motion is $\omega_c = qB/m$ and the cyclotron radius is $r_c = mv_{\perp}/qB$. Here $v_{\perp}$ is the magnitude of the particle velocity perpendicular ($\perp$) to the magnetic field direction. In terms of the pitch α of the particle, the angle between the particle velocity $\mathbf{v}$ and the magnetic field $\mathbf{B}$, $v_{\perp} = v \sin \alpha$.

The second type of motion comes from particles with $\alpha \neq \pi/2$. These particles can move along the direction of the magnetic with $v_{\parallel} = v \cos \alpha$. Here $\parallel$ denotes parallel to $\mathbf{B}$. In an inhomogeneous magnetic field such as the dipole field, the field becomes stronger as the particle gets closer to the planet. The particle orbit winds tighter as it approaches the stronger field region and it also encounters a force that pushes the particle back in the direction from which it arrives. This results in 'mirroring' of the particles, which bounce back and forth between the Northern and Southern Hemispheres.

The third type of motion comes from particles travelling on magnetic fields that are curved and not uniform in the radial direction. Particles travelling on curved magnetic fields experience a centrifugal force. Particles in fields that are not uniform experience a continually changing cyclotron radius. Both of these effects result in particle drifts in the azimuthal direction, westward for positively charged particles and eastward for negatively charged particles. The curvature and gradient drift velocities are energy-dependent, with higher energy particles drifting faster.

In the presence of electric fields, the motions described above must now be augmented with the effects of electric fields. Consider a time-independent electric field given by $\mathbf{E} = \mathbf{E}_{\parallel} + \mathbf{E}_{\perp}$, where the $\parallel$ and $\perp$ are directions relative to the direction of the magnetic field. The effect of $\mathbf{E}_{\parallel}$ is to accelerate or decelerate particles travelling along the magnetic field. $\mathbf{E}_{\perp}$ gives rise to a drift in the azimuthal direction given by eqn [24]. A peculiar feature here is that the drift is the same for $+$ and $-$ particles and also independent of the mass, charge, and energy of the particles.

In summary, the total motion of particles in magnetospheres consists of a superposition of the cyclotron motion, bounce motion and the drift motion. In the absence of $\mathbf{E}_{\parallel}$ these

are energy-conserving motions. In the outer radiation belt at synchronous altitudes ($\sim 6.6R_E$ geocentric) for example, the electron cyclotron frequency is ~ 1 kHz, the bounce period for 40 keV electrons is ~ 1 s, and the drift period is ~ 2 hours. For the ions, the cyclotron frequency is smaller by the ratio of electron to ion mass, the bounce frequency by the square root of the mass ratio, and the drift times are same for the same energies.

The lifetime of the magnetospheric particles is determined by how close they approach the planetary atmosphere. If they approach close to the planet where the atmospheric densities are sufficiently high, so that they collide with them, then these particles will be lost into the atmosphere. These are called precipitated particles and they are the source of atmospheric emissions responsible for aurorae at high magnetic latitudes. If the particles bounce at sufficiently high altitudes where collisions are infrequent then these particles can persist for a long time and drift around the magnetosphere many times. These particles are trapped particles and they form the Van Allen radiation belts.

Dynamic Magnetosphere

Contrary to the static magnetosphere described above, the real magnetosphere is time-dependent and very dynamic. As an example, we describe phenomenologically what happens when the solar wind is moderately disturbed. The magnetopause boundary moves in and out in response to the solar wind variations. The boundary is no longer smooth but is modulated with surface waves, reminiscent of atmospheric and ocean waves in stormy weather. These magnetopause waves could be excited by the Kelvin–Helmholtz instability mechanism owing to the presence of the large solar wind velocity shear across the boundary. Inside the magnetosphere, particles are injected from the plasma sheet into the outer radiation belt, and trapped Van Allen particle intensity and energy increase by orders of magnitude. This happens when the tail current abruptly disrupts and the geomagnetic field returns for a short time to the dipole shape. The trapped particle intensity increases considerably during magnetic storms and the particles form a ‘ring current’ at $\sim 4R_E$ strong enough to affect the terrestrial magnetic field measured on the ground. In the polar ionosphere, brilliant and wild auroral displays luminate the night sky accompanied by the roaring of natural electromagnetic radio waves that are emitted over millihertz to megahertz frequencies. The geomagnetic tail wags and flaps like a wind sock on a windy day. An observer in the interplanetary space sees the magnetosphere soaring through the heliosphere with bright flickering lights, resembling a comet.

Unable to solve eqns [1], [2], [3], [4], [5], [6] and [7] in a self-consistent way, space researchers have thus far obtained only approximate solutions. For this reason, the picture of the magnetosphere is incomplete. The approach most commonly used has been to treat the collisionless space plasmas as a magnetohydrodynamic (MHD) fluid and then solve the

mass, momentum, and energy conservation equations using the variables of density, velocity, and temperature. MHD theory has provided insight into the ways the solar wind flows and how magnetospheres respond to the solar wind. An important paradigm of the MHD theory is that interplanetary and Earth’s dipole magnetic fields can ‘merge’, which opens the magnetosphere so the solar wind can enter. This innovative concept was introduced by James W. Dungey nearly 40 years ago. The physics of merging requires a kinetic treatment which has not yet been solved.

A limited number of dynamical processes have been studied from the kinetic point of view, assuming that an ensemble of collisionless particles can be defined in terms of the distribution function. One then solves the Boltzmann transport equation coupled to the EM equations. This approach has been most fruitful in the study of the microphysics of wave–particle interactions and instabilities. Progress has been made towards understanding the microphysics of auroras observed at ionospheric heights, and much has been learned about the structure of aurorae, how particles are accelerated in the ionosphere and how certain types of radio emissions are excited. The full kinetic formulation has not however been applied to problems of large-scale global spatial structure and dynamics, and this will remain at the forefront of magnetospheric studies.

Among the important dynamic problems to be studied in magnetospheres include how the solar wind gets into the magnetosphere, how the particles are accelerated in the magnetosphere, how electric fields are set up in the geomagnetic tail, how the magnetosphere and the auroral ionosphere are coupled, how global currents can be generated during auroral storms to reconfigure the entire magnetosphere and how some of the solar wind energy is captured to produce the global aurora.

See also: Mesosphere: Ionosphere. Solar System/Sun, Atmospheres, Evolution of Atmospheres: Solar Winds.

Further Reading

- Alfven, H., Fälthammar, C.G., 1963. *Cosmical Electrodynamics, Fundamental Principles*, second ed. Oxford University Press, Oxford.
- Baumjohann, W., Treumann, R.A., 1996. *Basic Space Plasma Physics*. Imperial College Press, London.
- Hargreaves, J.K., 1992. *The Solar-Terrestrial Environment*. Cambridge University Press, Cambridge.
- Hess, W.N., 1968. *The Radiation Belt and Magnetosphere*. Blaisdell, Waltham, MA.
- Kamide, Y., 1988. *Electrodynamic Processes in the Earth’s Ionosphere and Magnetosphere*. Kyoto Sangyo University Press, Kyoto.
- Kivelson, M.G., Russell, C.T. (Eds.), 1995. *An Introduction to Space Physics*. Cambridge University Press, Cambridge.
- Parker, E.N., Kennel, C.F., Lanzerotti, L.J. (Eds.), 1979. *Solar System Plasma Physics*. North Holland, Amsterdam.
- Parks, G.K., 1991. *Physics of Space Plasmas, An Introduction*. Addison-Wesley, Reading, MA.
- Uberoi, C., 2000. *Earth’s Proximal Space*. Universities Press, Hyderabad.

MESOSCALE METEOROLOGY

Contents

Overview

Cloud and Precipitation Bands

Gust Fronts

Hail and Hailstorms

Mesoscale Convective Systems

Microbursts

Severe Storms

Waterspouts

Bow Echoes and Derecho

Density Currents

Convective Storms: Overview

Overview

DJ Parker, University of Leeds, Leeds, UK

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

Mesoscale meteorology occurs in the range of horizontal scales between larger, synoptic-scale and smaller boundary-layer scales. The mesoscale includes many of the weather systems with the highest societal impact, including cumulonimbus storms, intense cyclones and locally generated winds. Weather systems on the mesoscale also control many transport processes in the global atmosphere, causing rapid movement of air from the boundary layer to the upper troposphere. This article reviews the physical processes which control weather systems on the mesoscale, and their mathematical representation, then makes a broad survey of some important mesoscale phenomena.

Introduction

The mesoscale is an intermediate regime between the larger, synoptic scale and the smaller scales of boundary layers, turbulence and micrometeorology. The fact that the mesoscale is an intermediate regime is reflected in the physical processes, which influence mesoscale systems. While larger, synoptic-scale systems can be regarded as being close to geostrophic balance, and small-scale systems are variously independent of Coriolis force, compressibility or nonhydrostatic effects, at the mesoscale we cannot immediately neglect any components of the equations of motion, except for effects associated with the Earth's curvature. In a sense, the mesoscale could be defined as the regime in which all the physical processes represented in the basic dynamic and thermodynamic equations of the atmosphere may be important. By necessity, this article refers in only superficial detail to a wide range of phenomena and processes, each of which is discussed in more detail in separate articles.

Since it spans a range of phenomena and processes, the mesoscale is often subdivided into the rather unmemorable categories of,

- meso- α : 200–2000 km, e.g., secondary frontal cyclones,
- meso- β : 20–200 km, e.g., mesoscale convective systems (MCSs), and
- meso- γ : 2–20 km, e.g., cumulonimbus cells.

Since these categories are in order of roughly increasing Rossby number (defined as $Ro = U/fL$, with U and L the velocity- and lengthscales and f the Coriolis parameter), they correspond to a decreasing importance of planetary rotation and geostrophically balanced dynamics.

The relegation of mesoscale meteorology to a transition zone between larger and smaller scales of the atmosphere is to belie the fact that the mesoscale exhibits a majority of the most interesting, observable weather systems. Many of these systems are of the highest societal impact. These include cumulonimbus storm cells and cumulonimbus complexes, a variety of

frontal structures, and various coherent flows relating to topography or coastal regions. These systems are fascinating as remarkably coherent and repeatable structures in a nonlinear and complex dynamical regime of the atmosphere. Mesoscale systems are also important practically, being responsible for producing damaging storms (such as squall lines, supercells, mesoscale convective complexes, secondary frontal cyclones, polar lows, and downslope windstorms), complex rainfall distributions, and significant inhomogeneity in local weather conditions. In addition, cumulonimbus convection and active fronts are the two families of systems responsible for rapid vertical transport of trace gases from the boundary layer (where most natural and anthropogenic sources exist) to the upper troposphere and lower stratosphere. Such weather systems are poorly resolved or parametrized in global-scale models, and they are critical to global-scale chemical transport.

Increasing computer power over recent decades has led to a rapid increase in the possibilities for numerical modeling of mesoscale phenomena. In the study of cumulonimbus convection, for example, simulations of evolving cloud systems have been able to explore the sensitivity of such storm systems to environmental parameters (such as background winds and thermodynamics), with useful benefit to forecasting. As remarked above, these kinds of simulations tend to contain all the components of the dynamic and thermodynamic equations, since they relate to a scale where nothing can be wholly neglected, and are often termed ‘full physics’ simulations. Indeed, simulations of MCSs are performed using ‘large eddy’ simulations incorporating explicit cloud microphysics as well as the influence of the Coriolis acceleration.

Until recently, the resolution of global-scale models has meant that representation of mesoscale systems has by necessity involved subgrid parametrization. Currently, however, the horizontal resolution of global operational models is approaching the mesoscale, and we are left with a delicate balance between what is resolved and what is parametrized: if a model has resolution of 25 km, and the coherent convective structures are of a similar scale, the convective structures are neither well-resolved nor well-parametrized.

Despite the increasing possibilities of mesoscale numerical modeling, the observed coherence of mesoscale weather systems, in the sense that the same coherent structures are observed on many occasions, points to the importance of idealized and intuitive interpretations of the atmospheric dynamics, and this remains an active area of research. The simple model of the density current, for example, is very efficient at obtaining robust estimates of the propagation characteristics of small-scale fronts, is intuitively easy to understand, and requires some considerable effort to better with a numerical model.

Mathematical Description of Different Mesoscale Regimes

The only universal simplification appropriate to all mesoscale systems is to neglect effects of the Earth’s curvature and consider an f -plane version of the ‘primitive equations’ (which represent a shallow layer on the quasi-spherical Earth).

The resulting equations of motion can be cast in different forms, depending on the choice of vertical coordinate: here they will be stated in physical height coordinates. The vector momentum equation is,

$$\frac{D\mathbf{u}}{Dt} + f\mathbf{k} \times \mathbf{u} = -g\mathbf{k} - \frac{1}{\rho}\nabla p' + \frac{\partial \tau_{ij}}{\partial x_j}, \quad [1]$$

where $\mathbf{u}$ is the vector velocity, $\mathbf{k}$ is the upward unit vector, g is the acceleration due to gravity, ρ is the density, p' the perturbation pressure, and τ_{ij} is the Reynolds stress tensor, with the summation convention assumed for this last term. The system also requires a thermodynamic equation,

$$\frac{D\theta}{Dt} = S, \quad [2]$$

where θ is the potential temperature and S is a thermodynamic source term (say due to radiative flux convergence or phase changes of water and including a term according to the divergence of the turbulent flux of potential temperature) and conservation of mass, or continuity,

$$\frac{D\rho}{Dt} + \rho\nabla \cdot \mathbf{u} = 0. \quad [3]$$

Turbulence Closure

The momentum and thermodynamic equations involve turbulent fluxes on the right-hand side, which need to be obtained by some form of turbulence closure. Although a great deal of effort can be expended on dealing with the turbulent terms, there are no schemes which are entirely satisfactory. The most reliable approach for numerical modeling is regarded to be large eddy simulation, in which, crudely speaking, a spatial resolution is chosen to resolve the anisotropic scales of eddies, with the assumption that smaller (isotropic) scales can be dealt with effectively by a relatively simple turbulent scheme. In practice, the resolutions required by this method, on the order of meters, mean that computational demands are extremely high when mesoscale systems on the scales of 10 km or more are studied.

For some flows, such as small-amplitude buoyancy waves, the turbulence terms will be very small, but most mesoscale systems involve turbulent mixing to an important degree.

Simplifications to the Continuity Equation

In most instances the continuity equation may be simplified using the assumption of subsonic flow, to the anelastic form,

$$\nabla \cdot (\rho_r \mathbf{u}) = 0, \quad [4]$$

in which $\rho_r(z)$ is a reference density which depends only on height, z . If, as in the case of a sea breeze current, the vertical scale of the motion is small relative to the scale height,

$$H = \frac{R\bar{T}}{g} \sim 8 \text{ km}, \quad [5]$$

where R is the gas constant for air and $\bar{T}$ is a mean temperature for the atmosphere, it is possible to use the incompressible form,

$$\nabla \cdot \mathbf{u} = 0. \quad [6]$$

There are few instances where one of these approximations will not be valid, but it should be recalled that deep waves in the atmosphere have speeds, which scale approximately with the vertical wavelength. In a quiescent atmosphere of constant static stability, the maximum group and phase speeds of buoyancy waves are

$$c = \frac{N}{m}, \quad [7]$$

where N is the Brunt–Väisälä frequency and m is the vertical wave number: for a wave with a single maximum in the troposphere this can be of the order of 50 m s^{-1} , while for a wave with a first maximum at the tropopause it can be 100 m s^{-1} . Further refinements to the continuity equation exist, and have been proposed to take into account such very deep motions.

Hydrostatic Balance

It is often possible to assume a state of hydrostatic balance, provided the vertical scale of the motion is significantly smaller than the horizontal scale, which tends to imply the meso- α regime. This is, for example, applicable to most scales of frontal dynamics, but not to the circulations in a cumulonimbus storm. For inertia-gravity waves, the applicability of hydrostatic balance depends on the aspect ratio of the waves, being suitable when the ratio of horizontal to vertical wave numbers, k/m , is small.

The Boussinesq Approximation

This approximation amounts to neglecting density variations in inertia terms but retaining them from the point of view of buoyancy, which is valid when the vertical scale of the motion is much less than the scale height, H . A basic state density field $\rho_r(z)$, dependent only on height, is then incorporated into a modified pressure function. This is very convenient in that it retains the simplicity of uniform, incompressible dynamics, while accommodating buoyancy variations. There are several versions of the Boussinesq approximation: typically it leads to momentum equations of the form,

$$\frac{D\mathbf{u}}{Dt} + f\mathbf{k} \times \mathbf{u} = g'\mathbf{k} - \nabla\phi + \frac{\partial\tau_{ij}}{\partial x_j}, \quad [8]$$

where $\phi = p'/\rho_r$ is the modified pressure and g' is the buoyancy.

Semigeostrophic Theory

The semigeostrophic (SG) model is a refinement of quasi-geostrophic (QG) dynamics, to allow for short lengthscales in the region of a synoptic front. This approach, although based on balanced dynamics, is thought to work effectively in quasi-two dimensional frontal zones down to scales of tens of kilometers, and is therefore one of the few areas where analytical, mathematical theory has been able to bridge the mesoscale regime. Broadly, SG frontal dynamics involves an along-front component of the wind, v_g , in geostrophic balance with the cross-frontal pressure structure, but also accommodates advection by the ageostrophic cross-frontal and vertical winds (a feature which is absent from QG dynamics). This is made

analytically more tractable by transformation to a geostrophic momentum coordinate,

$$X = x + \frac{v_g}{f}, \quad [9]$$

where x is the cross-front direction. The transformation between coordinate systems can develop a singularity in finite time, and mathematically it is this which leads to the formation of frontal singularities in the model; physically, it is the process of advection by the ageostrophic winds, which leads to the enhanced frontogenesis.

SG models are based on thermal wind balance, which states that the vertical gradient of the geostrophic wind is proportional to the horizontal gradient of the temperature. Using this balance SG models can conveniently be solved by inversion of a conserved (in the absence of diabatic processes) potential vorticity function. This makes SG potential vorticity a powerful quantity with the properties that it is conserved moving with the flow, and that the temperature and wind patterns can be retrieved from it. Weather patterns can be identified in terms of their coherent potential vorticity patterns, which then move with the weather system, and define the winds and temperatures in this system. This has major advantages relative to (for example) studying weather systems in terms of their surface pressure patterns, since the surface pressure is not conserved, changing as the weather system moves.

The inversion procedure for SG potential vorticity fails when the potential vorticity becomes negative, a condition, which implies 'slantwise instability' of the atmosphere.

Higher order balance models have also been applied to mesoscale phenomena such as organized convective systems. However, the prevalence of latent heating and turbulent processes, as well as the difficulty of applying a balance condition at high Rossby number, has meant that the use of potential vorticity concepts is less common to mesoscale systems than it is in synoptic analysis.

A Tour of Mesoscale Phenomena

Buoyancy Waves

Often termed 'gravity waves' these patterns occur as a universal response of the stably stratified atmosphere to mesoscale perturbations. They propagate with components in the vertical and horizontal, and are modified strongly by the vertical structure of both stratification and horizontal wind. For shallow waves, for which k/m becomes small, the waves are increasingly influenced by the Coriolis terms, and are known as inertia-gravity waves. Two characteristic regimes of solution are upward-propagating waves, which carry energy and momentum upwards away from the wave source, and trapped waves, where a wavelike region is bounded by an evanescent layer, leading to ducting of the waves in the horizontal. Although simple, linear buoyancy wave solutions for a quiescent, Boussinesq atmosphere are relatively easy to obtain, for real profiles it seems that detailed consideration of the full vertical profile is necessary in order to diagnose the correct buoyancy wave characteristics.

In regions of orography, vertically propagating waves tilt backwards against the mean flow (Figure 1(a)), and the

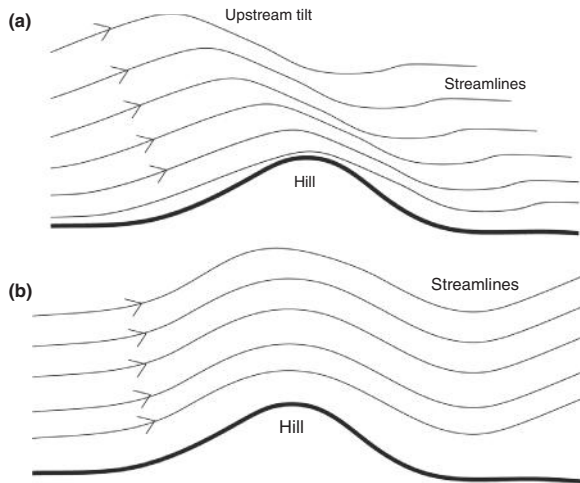


Figure 1 Schematics of streamlines over a hill in the cases of (a) upward wave propagation and (b) wave trapping. In (a) the wave tilt with height contributes to a wind speed maximum in the lee.

resulting shift in the streamlines contributes to downslope acceleration of the wind. In conditions of trapping (**Figure 1(b)**), lee waves can propagate significant distances downstream and can often be seen in low-level cloud patterns (**Figure 2**). Such lee waves give rise to local regions of low-level convergence, which may occasionally initiate moist convection.

Deep convection forces waves propagating both upward into the stratosphere and horizontally in the troposphere: deep modes in the troposphere propagate at speeds upward of 50 m s^{-1} , and evidence of wave-trains forced by convection has been observed in the small surface pressure changes measured by arrays of microbarographs.

The horizontally propagating waves act to modify the environment of the convection, through adiabatic warming and cooling, and are one mechanism whereby the convection communicates its thermodynamic forcing into its environment. The primary mode of response to the convective heat

source is downward motion, warming the environment: heating and strong ascent within a convective cloud causes subsidence in its environment, and this region of subsidence spreads outwards with gravity wave propagation. Within the subsidence region the air warms through adiabatic descent, meaning that the heating within the cloud (due to cloud condensation) has caused heating of the environment through subsidence. Although the heating profile of a cumulonimbus storm is dominated by midlevel warming, the vertical structure of the heating (including forcing due to downdraft cooling) forces modes of different vertical structure. As in the case of orographic waves, there is an important possibility of wave ducting, if waves are trapped in the troposphere, in which case a strong response can occur at large distances.

Generation of buoyancy waves by orography and by convection and the subsequent momentum flux convergence when the waves break are thought to be significant processes in the global circulation. However, neither wave source, whether orographic or convective, is ideally represented in global models. Diagnosis of the relevant wave regime is not easy, and waves may propagate long vertical or horizontal distances, equivalent to many model gridlengths, before breaking.

Cumulonimbus Convection

Cumulonimbus storms are the most dramatic of cloud features, and are a vital component in the atmospheric circulation. Individual cumulonimbus cells occur on horizontal scales of around 10 km and commonly extend to the tropopause, over a timescale of an hour or so (**Figure 3**). However, it is common for cumulonimbus systems to self-organize into an MCS with a significantly longer life cycle: squall lines in West Africa can persist for 48 h and propagate for thousands of kilometers across the continent.

Cumulonimbus storms involve rapid and active microphysical transitions, among water vapor, cloud drops, rain, and many forms of ice. The microphysical changes occur on time-scales of a few minutes, yet the storm systems can persist for many hours and organize into a significant degree of geostrophic balance, so these systems can truly encompass the



Figure 2 Linear bands of cumulus, representing low-level 'gravity wave clouds,' observed at West Lutton in Yorkshire at 1937 UTC on 1 June 2011. On this day, similar clouds were observed over large parts of the UK.



Figure 3 A typical cumulonimbus storm, observed near Varberg, Sweden, 1336 UTC on 6 August 2005.

whole mesoscale atmospheric regime. In terms of mathematical representation, the strong up- and downdrafts in these storms, combined with intense turbulence and their great depth, mean that none of the mathematical approximations outlined above can comfortably be applied (although the anelastic approximation may be used if the fast buoyancy wave response is not critical to the solution). Cumulonimbus systems also involve other distinct mesoscale flows as components, or responses to the forcing. The buoyancy wave response has been discussed above; other component flows include the cold pool and its gust front. Within a cumulonimbus storm, evaporation in the precipitation-driven downdraft causes the cold pool of air at the surface, and this propagates like a density current, of around 1 km depth, into the environmental boundary layer. Forced lifting at the gust front acts to trigger new convective cells (Figure 4). Although the gust front spreads in all directions outward from its source, the lifting is optimized in a given direction in relation to the ambient wind, and this in turn feeds back on the morphology of the cumulonimbus system. For instance, when regeneration of new convective cells occurs on one side of a storm, due to favorable interaction with the ambient winds on that side, this leads to development of the convective system in that direction. In some cases, where there is a balance between the ambient flow and the cold pool, stationary storms can develop, and these can lead to long periods of intense rainfall on a fixed location, with the possibility of flash flooding.

The interaction between boundary-layer shear and differential lifting at the gust front is also thought to lead to precursor vortices for tornado development. In this case, the horizontal vorticity in the boundary layer, due to the change in strength and direction of the wind with height, is tilted into the vertical by interaction with a gust front, and then the vertical vorticity is very much intensified by vortex stretching in the vertical.

When two gust fronts from adjacent storm cells collide, they produce enhanced lifting at the region of collision, and this can be another favored location for storms to be generated.

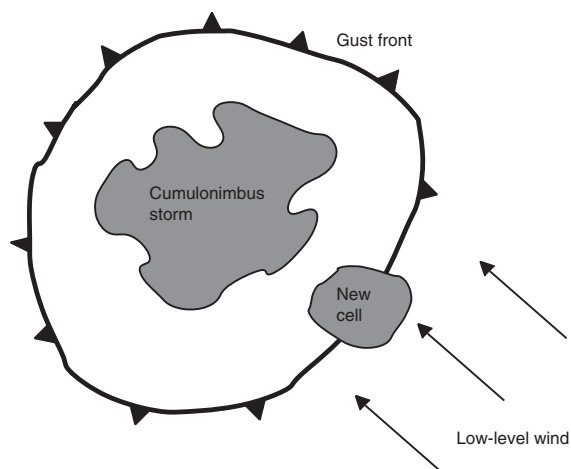


Figure 4 Plan view of an idealized cumulonimbus system. A new cumulonimbus cell is often formed at the gust front, in a direction determined by the low-level wind and shear structure.

The regeneration of cumulonimbus storms at the gust front is an intriguing way in which the most turbulent, nonlinear, and irreducible of atmospheric flows can be understood in terms of relatively simple, coherent components.

In some cases, extremely intense downdrafts lead to downbursts and microbursts, which produce extremely intense gusts at the surface, and are a severe hazard for aircraft. Such events are hard to forecast and can result from convective clouds which, to the eye, appear relatively benign.

Sea Breezes and Other Thermally Generated Winds

The sea breeze is the best example of a flow generated by the relatively rapid generation of a baroclinic zone (a horizontal temperature or density gradient) at low levels of the atmosphere. Warming of the land surface after sunrise is rapid, while the sea surface temperature remains almost constant, so the boundary-layer inland becomes relatively warm over the space of a few hours. This leads to a 'baroclinic overturning' (Figure 5(a)) in which the warm air tends to rise and the cool air to subside, with corresponding horizontal flows to balance mass. As this circulation develops, the development of a flow resembling a density current is observed, as a cold sea breeze front pushes inland. The front may be perceived in cloud formed due to the forced ascent, or in visibility changes. Continued heating of the land surface means that there is likely to be relatively strong convective turbulence in the boundary-layer inland, which tends to dissipate the sea breeze front and render it a more diffuse baroclinic zone. However, as the surface heating

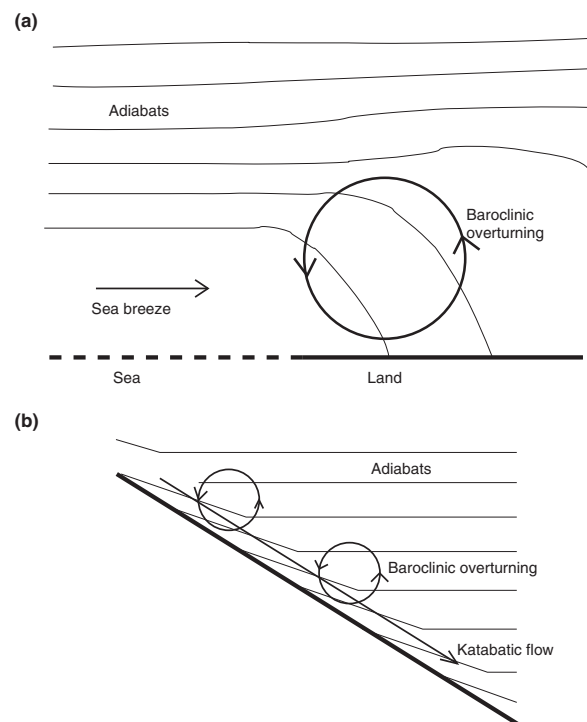


Figure 5 Baroclinic tendencies in horizontal vorticity lead to thermally driven flows; (a) the sea breeze and (b) a katabatic, downslope wind.

diminishes in the evening, the convective turbulence decays and the front may intensify and propagate further inland. Like a cold pool from a convective storm, the sea breeze front is also sensitive to the ambient winds. In strong background, winds a sea breeze will not be observed. In lighter winds, an intense sea breeze front tends to occur in conditions of light offshore flow.

A density current in the atmosphere is in a state of balance between the pressure gradient force due to the density change across the front, and drag on the current due to turbulent stresses (principally Kelvin–Helmholtz instability at the head). However, the sea breeze is also influenced by the Coriolis acceleration, over a timescale $1/f$, and can be expected to turn with the Coriolis acceleration as the day progresses. Sea breezes are quite sensitive to the large-scale flow, and will not develop if the ambient winds are strong.

After sunset it is possible for a land breeze to develop, as the land surface cools more rapidly than the sea surface. However, the ensuing surface inversion, which suppresses turbulence, does not develop as deeply as the daytime convective boundary layer, so the land breeze tends to be less active and less intense than the sea breeze.

Under conditions of light ambient winds, comparable circulations can be observed at boundaries between land surface types. These circulations are generally weaker, less coherent and harder to observe than the sea breeze, because the contrasts in surface heating between different land surface types is weaker than the land–sea contrast. However, coherent sea breeze-like circulations occur over cities (related to the ‘urban heat island’), at forest–crop boundaries, and in semiarid regions at contrasts in soil moisture. There is increasing evidence, notably from the Sahel zone of West Africa, that these circulations forced by land surface contrasts in vegetation or soil moisture influence the generation of cumulonimbus rainfall. This is an important way in which the land surface patterns exert some control on rainfall, and therefore lead to feedbacks between the slow evolution of the land surface and the rapid mesoscale processes causing rainfall. Similarly, other results from the Amazon basin have indicated that mesoscale patches of deforestation may result in increased rainfall over the deforested areas, with a consequent promotion of forest regeneration in those areas.

On sloping terrain, the diurnal cycle of surface heating leads to baroclinicity relative to the background air, and tends to cause upslope, or anabatic, flow in the daytime and downslope, katabatic, flows at night (**Figure 5(b)**). Again, these flow regimes only develop under conditions of light ambient wind, but when they do occur they may dominate the local meteorology.

Thermally generated winds can be very important to the local meteorology and climate of a specific location, influencing the daily temperatures, cloud cover, winds, and rainfall on scales of a few to hundreds of kilometers. Therefore, these mesoscale features need to be accounted for in making assessments of the weather and climate of locations where surface contrasts are strong, such as coastal regions or large cities. Accurate representation of all of these thermally generated flows in numerical models requires close attention to the surface and boundary-layer conditions,

as well as suitable model resolution to represent the mesoscale dynamics.

Synoptic Fronts

The fronts which form within synoptic cyclones exhibit lengthscales far shorter than those of the parent system, and are controlled by a spectrum of dynamical processes, so that synoptic fronts are really mesoscale features. It has been known since the early twentieth century that synoptic fronts exist primarily as a balance between the tendency for thermal overturning due to the frontal pressure gradients (with warm air tending to rise) and the Coriolis force associated with along-front geostrophic winds. Therefore, although fronts are associated with temperature contrasts primarily, they are always also associated with changes in the winds, as well as humidity and clouds changes.

The genesis of synoptic fronts is well represented in primitive equation simulations, but the genesis is best understood in SG theory, which predicts the collapse of a synoptic thermal gradient (on the scales of hundreds of kilometers, for example) to a mathematical singularity in finite time. In the atmosphere, a singularity in winds and temperatures does not occur, because the flow moves to a low Richardson number (predicted by the SG theory), at which point turbulent effects lead to mixing of the temperature and wind contrasts. Subsequent to this point, it is assumed that a real front is in a state of balance between Coriolis force, thermally derived pressure gradient forces and turbulent stresses, and observational studies have largely confirmed this state.

In addition to the interplay between balanced and turbulent forces at the front, most synoptic fronts are characterized by significant cloud and precipitation features, which can have a first-order influence on the dynamics. Cold fronts, in particular, tend to trigger cumulonimbus convection, which in turn generates strong downdrafts and cold pool/gust front structures at the surface. A number of observations of the low-level structure of active cold fronts have shown the details of the surface front to resemble a density current quite closely. Sometimes, a frontal zone is composed of a number of rainbands, aligned almost parallel to the front, whose nature can resemble that of a squall line. Rainbands have been attributed to a number of processes, such as conditional symmetric instability, which occurs when the (moist equivalent) potential vorticity becomes negative. Synoptic fronts, then, can span the range of mesoscale phenomena: they are formed out of the larger, synoptic-scale flow, and can exhibit cumulonimbus convection, rainbands, and density current structure.

One of the important features of the SG model of a synoptic front is that the system remains a continuous feature – a zone of smooth changes in the atmospheric properties – up to the time at which a mathematical singularity occurs. Since these fronts form as components of a baroclinically unstable wave (for instance, fronts appear in an SG version of the theoretical Eady wave), the front may propagate as a wavelike phenomenon, meaning that the fronts are propagating as the steep part of a wave moving through the atmosphere. For this reason, there is an important possibility of material transport through the frontal zone. This behavior is observed and has important consequences for chemical transport: a front is not a barrier to

airflow, and air may flow from one side of a front to the other (generally with a strong along-front motion also). This, in conjunction with the rapid vertical motion due to the cross-frontal circulations and the frontal cloud features, means that synoptic fronts are important agents of transport in the atmosphere.

Mesoscale Cyclones

Vortices occur on all scales in the atmosphere, and those on the mesoscale are particularly important in terms of observable weather. As ever in this regime, mesoscale vortices are formed and influenced by the whole spectrum of physical and thermodynamic processes.

Secondary cyclones commonly form on the synoptic fronts of parent cyclones. These secondaries can develop by baroclinic instability of the parent front (feeding on the frontal temperature contrasts and potential energy), but may also have a significant or even dominant component of barotropic instability, arising from the horizontal shears at the front and feeding off the frontal kinetic energy. Such systems may develop explosively and cause significant damage. The horizontal scales of these secondary cyclones range from the synoptic scales right down to scales of tens of kilometers or less.

In addition to the formation of cyclones through dynamical energy conversions, cloud processes can contribute significantly to the energetics of cyclogenesis. Equivalent static stability (defined in terms of a moist equivalent potential temperature) is generally lower than dry stability, so the growth rate of baroclinic waves in a saturated atmosphere is faster than the dry growth rate, and the lengthscale of the resulting cyclone is shorter.

Tropical cyclones, including hurricanes and typhoons, also lie within the mesoscale regime. These extreme weather systems have been the subject of a great deal of research and forecasting effort over many years, yet remain difficult to predict. Their dynamics are dependent on many processes in the atmosphere, including balance, moist convection and radiative fluxes, as well as complex ocean–atmosphere interactions. Polar lows occur in high latitudes, usually associated with rapid release of latent heat where cold air moves from the ice sheets over a relatively warm ocean, but also may derive energy from baroclinic conversions.

MCSs, which appear through self-organization of cumulonimbus convective storms, develop mesoscale cyclones in their lower levels over a period of a few hours. On the mesoscale, the strong ascent, which develops in cumulonimbus storms can lead to mesocyclones through tilting, stretching and baroclinic generation of vorticity.

It appears that mesoscale vortices can occur on all scales, and are responsible for a large number of the most severe windstorms we encounter, including tornadoes, hurricanes, and secondary midlatitude storms. In many cases, there have as yet been too few observations of such vortices to categorize them properly, and the balance between dynamic and thermodynamic energy sources for cyclogenesis across the spectrum of

scales is still not well understood: this remains a very active area of research.

Conclusions

In terms of societal impact, mesoscale meteorology encompasses some of the most important weather systems observed. Cumulonimbus storms, surface-forced mesoscale circulations, synoptic fronts, and mesoscale cyclones all produce wind, rain, and thermodynamic patterns of significant influence on society. However, by definition of this scale as one lying between the large and small scales of the atmosphere, it is generally not possible to make many simplifications in our descriptions of the physics and mathematics of these systems. This article has attempted a survey of some of the most important mesoscale phenomena, and tried to link these to the underlying physics and mathematics describing systems in different parts of the mesoscale regime. Each of these phenomena is described in more detail in other articles.

See also: **Boundary Layer (Atmospheric) and Air Pollution:** Modeling and Parameterization; Overview. **Climate and Climate Change:** Climate Feedbacks. **Clouds and Fog:** Cloud Modeling. **Dynamical Meteorology:** Baroclinic Instability; Coriolis Force; Inertial Instability; Overview; Potential Vorticity; Symmetric Stability; Wave-CISK. **Gravity Waves:** Buoyancy and Buoyancy Waves: Optical Observations; Buoyancy and Buoyancy Waves: Theory. **Mesoscale Meteorology:** Cloud and Precipitation Bands; Density Currents; Gust Fronts; Mesoscale Convective Systems; Microbursts. **Mountain Meteorology:** Cold Air Damming; Downslope Winds; Katabatic Winds; Lee Waves and Mountain Waves; Valley Winds. **Numerical Models:** Convective Storm Modeling; Large-Eddy Simulation; Mesoscale Atmospheric Modeling. **Synoptic Meteorology:** Cyclogenesis; Fronts; Polar Lows.

Further Reading

- Atkinson, B.W., 1981. *Meso-Scale Atmospheric Circulations*. Academic Press, London, 495 pp.
- Bader, M.J., Forbes, G.S., Grant, J.R., Lilley, R.B.E., Waters, A.J., 1995. *Images in Weather Forecasting: A Practical Guide for Interpreting Satellite and Radar Imagery*. Cambridge University Press, Cambridge, 499 pp.
- Carlson, T.N., 1991. *Mid-Latitude Weather Systems*. Harper Collins, London, 507 pp.
- Cotton, W.R., Anthes, R.A., 1992. *Storm and Cloud Dynamics*. Harcourt, New York, NY.
- Emanuel, K.A., 1994. *Atmospheric Convection*. Oxford University Press, New York, 580 pp.
- Gill, A.E., 1982. *Atmosphere–Ocean Dynamics*. Academic Press, New York, 662 pp.
- Holton, J.R., 1979. *An Introduction to Dynamic Meteorology*. Academic Press, San Diego, CA.
- Markowski, P., Richardson, Y., 2010. *Mesoscale Meteorology in Midlatitudes*. Wiley-Blackwell, Oxford, UK.
- Pielke, R.A., Pearce, R.P. (Eds.), 1994. *Mesoscale Modeling of the Atmosphere*. American Meteorological Society, Boston, MA, 167 pp.
- Ray, P.S. (Ed.), 1986. *Mesoscale Meteorology and Forecasting*. American Meteorological Society, Boston, MA.

Cloud and Precipitation Bands

RM Rauber, University of Illinois at Urbana-Champaign, Urbana, IL, USA

M Ramamurthy, University Corporation for Atmospheric Research, Boulder, CO, USA

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

The term 'precipitation band' describes an area of precipitation that originates from updrafts that are either nonconvective or weakly convective, and is sufficiently elongated that an orientation can be assigned. Precipitation bands are common within extratropical and tropical cyclones, and also develop along topographical and geographical features, such as mountains, islands, coastlines, and lake shores. The taxonomy used to classify precipitation bands derives from radar studies of precipitation structures in various areas of the world. Precipitation bands commonly form as the result of frontogenesis, deformation flow, moist symmetric instability, boundary-layer convergence, gravity waves, topographic effects, and seeder-feeder processes.

Introduction

When viewed from satellites, large weather systems such as extratropical and tropical cyclones are often composed of clouds that organize on the mesoscale in linear features called bands. From a radar perspective, precipitation falling from these clouds also organizes along lines. Meteorologists use the term 'squall line' to describe a line where the precipitation originates from thunderstorms produced by strong convective updrafts. The term 'precipitation band' is used to describe an area of precipitation that typically originates from updrafts that are either nonconvective or weakly convective, and is sufficiently elongated that an orientation can be assigned. Precipitation bands are called 'rainbands' and 'snowbands' depending on the type of precipitation they produce. When precipitation echoes appear on radar in several long lines or bands, the weather system is said to exhibit banded structure. If bands of heavier precipitation appear within a larger field of widespread precipitation, they are often called 'embedded bands.'

The structure, intensity, and orientation of precipitation bands in extratropical cyclones are primarily related to forcing occurring along frontal zones. In tropical cyclones, rainbands outside the eyewall are related to forcing associated with the relative motion of the vortex through its environment and to inertia-gravity waves. Precipitation bands also exist far from cyclones. For example, precipitation bands can form and sometimes align along topographical features, such as mountains and islands, or can be forced by variations in surface heating, such as along coastlines or over the U.S. Great Lakes in winter. Precipitation bands organize on the mesoscale, and range in width from about 5 to 250 km, and in length from tens to over a thousand kilometers. Precipitation bands have time-scales that range from less than an hour to more than a day. These scales are usually determined by scales of the forcing mechanisms that trigger and maintain them.

Band Classification

Extratropical Cyclones

The taxonomy used to classify precipitation bands in extratropical cyclones arose primarily from studies of cyclones in coastal locations of the United States, the British Isles, and

Japan. This classification, shown in Figure 1, consists of six major band groups and is based on the physical relationship of the bands to frontal boundaries. The narrow cold frontal rainband, typically about 5 km wide, is aligned with the surface position of the cold front. On satellite images, the band often appears as a narrow ropelike cloud that can extend hundreds or even thousands of kilometers along the surface cold front. Updrafts associated with forced convection can exceed 5 m s^{-1} within the band, and locally heavy rain can occur as the band passes. From a radar perspective, the narrow cold frontal rainband often consists of small ellipsoidal cores of heavier rainfall oriented about $30\text{--}35^\circ$ to the cold front, separated by gaps of lighter rainfall. These regularly spaced precipitation cores and gaps are believed to develop as a result of horizontal shearing

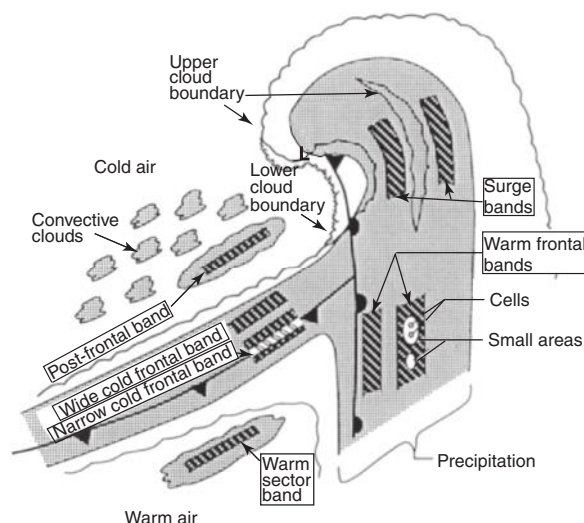


Figure 1 Idealization of the cloud and precipitation pattern associated with a mature extratropical cyclone. From Houze Jr., R.A., 1993. Cloud Dynamics. International Geophysical Series, vol. 53. Academic Press, p. 573, originally adapted from Matejka, T.J., Houze Jr., R.A., Hobbs, P.V., 1980. Microphysics and dynamics of clouds associated with mesoscale rainbands in extratropical cyclones. Quart. J. Royal Met. Soc., 106, 29–56 and Houze Jr., R.A., 1981. Structures of atmospheric precipitation systems – a global survey. Radio Sci., 16, 671–689; reprinted with permission from Academic Press, American Geophysical Union and the Royal Meteorological Society.

instability associated with strong wind shear across the frontal interface. Occasionally, the updrafts in the precipitation cores have been shown to produce weak tornadoes. Wide cold frontal rainbands develop behind the surface cold front. These bands, which form within the warm air rising along and over the frontal surface, are about 50 km wide and are parallel to the front. The updrafts in these bands are generally much weaker than the narrow cold frontal rainband, on the order of 0.5 m s^{-1} or less. Warm frontal rainbands have similar physical characteristics as wide cold frontal rainbands, except that they develop north of the surface warm front, as warm air from the south overruns and rises over the cold air. They are typically parallel to the surface warm front, are often embedded in lighter precipitation, and have updrafts with magnitudes similar to wide cold frontal rainbands.

Mature oceanic and continental cyclones often appear as well-defined comma-shaped cloud patterns on satellite pictures. These comma-shaped cloud patterns are characterized by a sharp boundary aloft between cool, dry air subsiding from the upper troposphere and moist air rising from the lower troposphere. This boundary, which appears on infrared satellite images as a sharp line demarking high and shallow or sometimes no clouds (e.g., [Figure 2](#)), typically occurs ahead (east) of the surface cold front and has been described in meteorological literature as an overrunning upper-level cold front, a prefrontal cold surge, a cold front aloft, or as a split front. The dry air behind this boundary is called the dry slot. The passage of the upper-level front is characterized by a sharp drop in equivalent or wet-bulb potential temperature, which is always associated with a sharp reduction in relative humidity and sometimes with a reduction in temperature. The organization of precipitation ahead of the upper-level front (at the leading edge of the dry slot) depends on the stability and vertical wind shear in the moist air ahead of the front. In oceanic and cold season continental cyclones, a rainband called the prefrontal cold surge rainband or cold front aloft rainband, typically develops along or slightly ahead of this boundary ([Figure 2\(a\)](#)). This rainband can extend hundreds of kilometers. In warm season continental cyclones, the moist air ahead of the front is often unstable and the vertical shear may be significant. In these cases, a squall line or a line of supercell thunderstorms may be triggered ahead of the upper-level cold front ([Figure 2\(b\)](#)). Convection, organized in bands, can also develop behind the upper-level cold front, where enhanced radiation (due to the dry, cloud free air aloft within the dry slot), along with moist low-level surface air can create extreme thermodynamic instability.

A wide band of precipitation typically wraps around the northwest quadrant of the low-pressure center of mature cyclones ([Figure 3](#)). In continental winter cyclones, this band often produces moderate to heavy snowfall, and is responsible for major winter storms and blizzard conditions. Smaller bands of heavier snowfall, often aligned along the midtropospheric thermal wind, are sometimes embedded within the broader wrap-around precipitation region. As a cyclone decays, the band typically elongates and narrows under the influence of background deformation flow.

All of the bands described above occur within the general envelope of clouds that compose the cyclone's comma-cloud pattern. With the exception of the narrow cold frontal

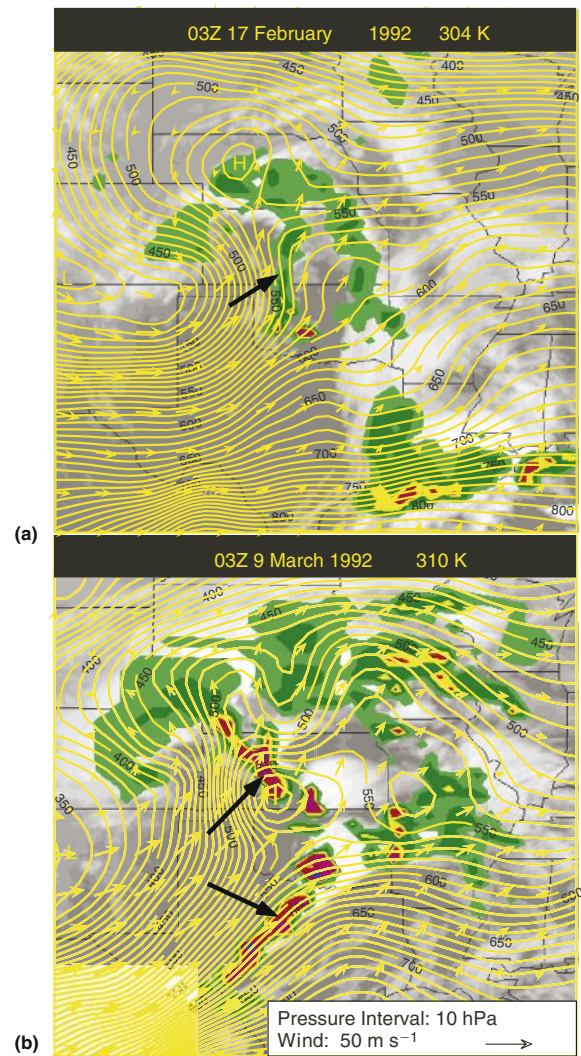


Figure 2 Isentropic analyses of pressure and wind on the (a) 304 K surface and (b) 310 K surface overlaid on an infrared satellite image and radar echoes. Areas of cold (warm) advection are indicated as cross-isobaric flow toward high (low) pressure. On each panel, a precipitation band appears at the leading edge of the cold air advection marking the position of the upper-level cold front (see arrows).

rainband, these bands are general enhancements of stratiform precipitation that characterizes the comma-cloud pattern. Bands also occur in the cyclone's warm sector, ahead of the primary cyclone cloud pattern. When they occur, these bands are called warm sector rainbands. Bands well behind the surface cold front are called post-cold frontal rainbands.

Tropical Cyclones

Concentric and spiral bands of clouds and precipitation are among the most striking features of a mature tropical cyclone. The classification of precipitation bands in tropical cyclones, shown in [Figure 4](#), has been developed primarily from studies of hurricanes over the North Atlantic Ocean. In some tropical cyclones, particularly intense hurricanes, precipitation inside

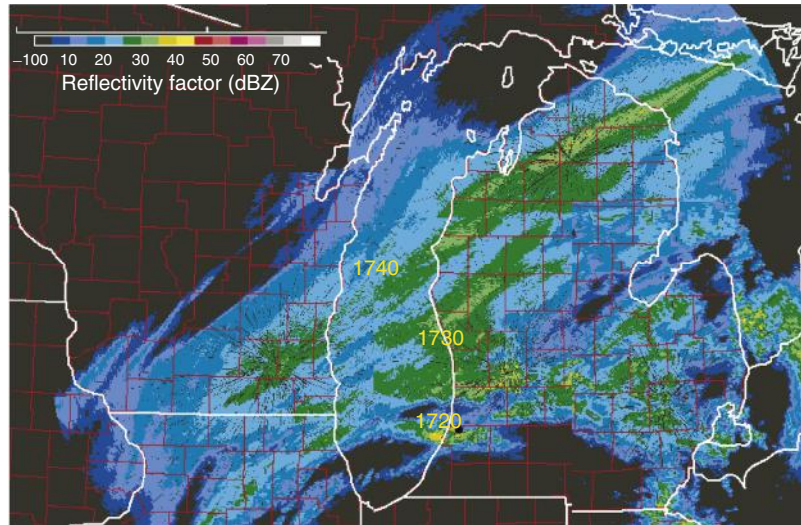


Figure 3 Radar image of a precipitation band in the northwest quadrant of a cyclone on 8 January 1998. This band produced heavy snow across the state of Michigan.

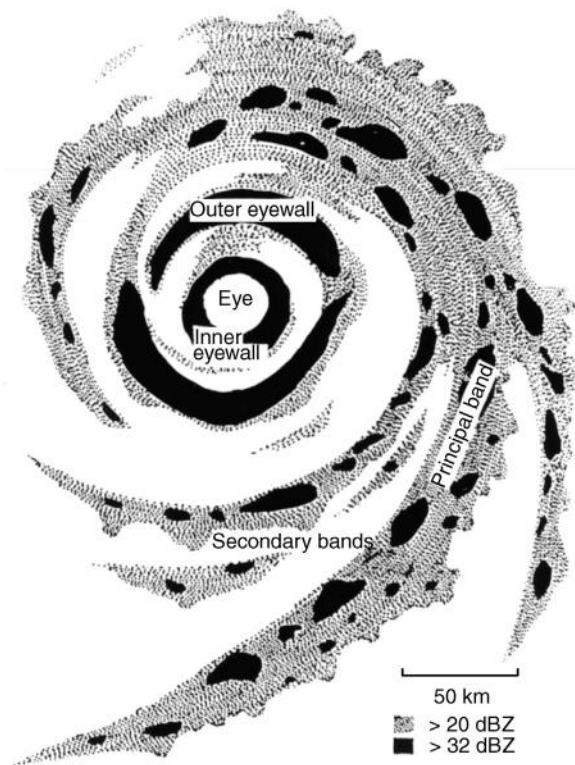


Figure 4 Schematic illustration of the radar reflectivity in a Northern Hemisphere tropical cyclone with a double eyewall and outer bands. From Willoughby, H.E., 1988. The dynamics of the tropical hurricane core. *Aus. Meteor. Mag.*, 36, 183–191. Commonwealth of Australia copyright, reproduced with permission.

a radius of about 100 km from the vortex center tends to be axisymmetric, with convection organized in one or more concentric rings. These rings, called the outer and inner eyewalls, contract toward the center of the vortex as a hurricane

intensifies. With time, the outer ring replaces the inner ring, leading to cyclic changes in central pressure. Precipitation organizes in spiral rainbands outside the radius of the outer eyewall in stronger hurricanes, and sometimes into the eyewall radius in weaker tropical cyclones. Unlike the concentric eyewall rainbands, precipitation in the outer spiral bands can be either stratiform or convective in nature. A large rainband, called the principal band, often extends from near the radius of the outer eyewall to the outer radius of the hurricane cloud pattern. This band normally occurs on the east side of a cyclone. The principal band typically has cellular convection along its axis, a clear region along its concave side, and stratiform precipitation falling from an anvil-like feature on its outer side. Other bands, called secondary bands, are found on either side of the principal band. These typically contain shallow convection and vary significantly in intensity. Sometimes short rainbands, called connecting bands, are present between the principal and secondary bands. A distinct connecting band often joins the principal band to the eyewall. This band either contains stratiform precipitation or weak convection. These bands together are termed the stationary band complex, since they move slowly, if at all, relative to the vortex. Tropical storms also contain smaller-scale banded precipitation features that propagate outward relative to the vortex.

Mechanisms Leading to the Formation of Banded Features

Extratropical Cyclones

Several mechanisms have been advanced to explain the existence of banded precipitation regions within extratropical cyclones. These include frontal lifting and/or frontogenetic circulations, moist symmetric instability (MSI), boundary layer convergence, gravity waves, topographic effects, and seeder-feeder processes. While it is difficult to generalize which of these mechanisms will explain the formation of a specific

precipitation band, there is considerable evidence that the aforementioned mechanisms are among the most important ones.

Frontogenesis

Banded features in extratropical cyclones are associated with vertical circulations organized along lines. Frontogenesis, a primary forcing mechanism for these vertical motions, is the process by which the thermal and moisture gradients between air masses are concentrated into narrow zones called fronts. During frontogenesis, cloud and precipitation producing transverse vertical circulations are generated, with ascending motion, concentrated into a quasi-linear narrow zone parallel to a front, on the warm side of the front. Frontogenetic circulations have been studied using quasi-geostrophic and semi-geostrophic theory, two of a hierarchy of approximations to the fundamental governing equations. The most realistic of these theories uses the semi-geostrophic system of equations, with inclusion of latent heat release to simulate the effects of condensation. From a quasi-geostrophic perspective, the intensification of the cross-front thermal gradient by background geostrophic deformation, leading to warm and cold advection on opposite sides of the front, destroys the thermal wind equilibrium. The atmosphere tries to restore the thermal wind equilibrium by generating a secondary circulation that produces vertical and ageostrophic motions that counter the effects of frontogenesis. During this process, adiabatic cooling is generated by the rising motion on the warm side, and adiabatic warming is produced by the sinking motion on the cold side of the front, partially negating the effects of frontogenesis. Semi-geostrophic theory takes into account the characteristically different length and velocity scales in the along- and cross-front directions by including the effects of thermal advection by the ageostrophic flow, which are ignored by quasi-geostrophic theory. In addition, the inclusion of moisture in the semi-geostrophic equations accounts for effects of latent heating associated with condensation on the warm side of the front. The resulting semi-geostrophic frontal circulations include a more intense, narrow sloped updraft on the warm side of the front (Figure 5), a frontal surface that tilts in the vertical toward the cold air, and a stronger thermal gradient near the surface than in quasi-geostrophic theory. The orientation, scale, and vertical motions associated with certain rainbands found in the vicinity of fronts qualitatively match these theoretical predictions.

Boundary layer convergence

The most prominent example of rainbands associated with boundary layer convergence is the narrow cold frontal rainband. Under the influence of deformation flow and the accompanying frontogenesis, the thermal gradient at a cold front sometimes collapses to a near-discontinuity at the surface. When this collapse occurs, the air behind the cold front takes on characteristics of a density current, and strong localized convergence occurs at the front. A sharp narrow updraft develops in the warm air ahead of the front as the warm air rises over the advancing density current. The resulting rainband, the narrow cold frontal rainband discussed above, often contains updrafts exceeding several meters per second and produces locally heavy rain. Studies of the narrow cold frontal rainband

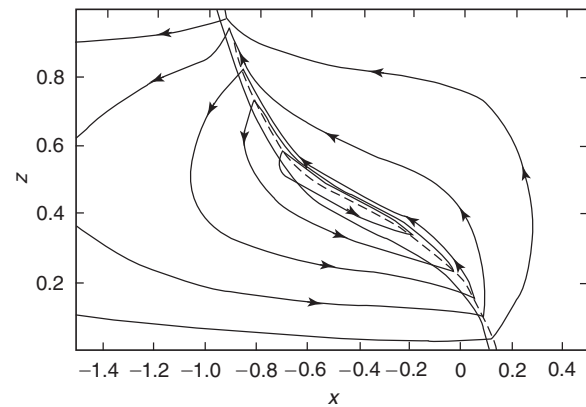


Figure 5 Ageostrophic stream function about a front obtained by solving the semi-geostrophic Sawyer–Eliassen equation for a case of deformation forcing under moist conditions. Height and distance (x) are expressed in nondimensional units. From Emanuel, K.A., 1985. Frontal circulations in the presence of small, moist symmetric stability. *Journal of Atmospheric Sciences* 42, 1062–1071; reprinted with permission from the American Meteorological Society.

in oceanic regions have shown that the warm air lifted by the density current is often stable or very slightly unstable, indicating that the updrafts are forced rather than due to free convection. The depth of the updrafts is limited to the depth of the forced ascent, usually no more than 3–5 km. The updrafts in the narrow cold frontal rainband are often complicated by small vortices that form along the wind-shift line associated with the advancing density current. Shearing instability is believed to be the principal mechanism for the formation of such vortices. The updrafts then organize in discrete precipitation cores, separated by gaps. Narrow cold frontal rainbands are not as common in continental cyclones as they are in oceanic cyclones, probably because the air ahead of cold fronts is more likely to be potentially unstable and, when lifted, produce squall lines or other deep convective phenomena.

Ducted gravity waves

Gravity waves are buoyancy oscillations in a stably stratified fluid where the restoring force is gravity. Gravity waves are ubiquitous in the atmosphere, but most are not large enough in amplitude to trigger precipitation bands. This is true because atmospheric gravity waves are highly dispersive, propagating vertically and rapidly losing their energy to the upper atmosphere. As a result, gravity waves normally cannot travel very far from their source before they no longer have sufficient energy to lift air to the lifting condensation level and trigger clouds and precipitation. A particular class of gravity waves, termed *ducted gravity waves* because their energy is trapped in an atmospheric duct, has been associated with precipitation bands. Ducted gravity waves require very specific conditions that include (1) a sufficiently deep low-level stable layer that can accommodate a quarter vertical wavelength of the wave, (2) a deep moist neutral layer above the stable layer, and (3) a critical level above the stable layer, where the mean wind speed in the direction of propagation equals the wave phase speed. These conditions most commonly occur north of warm fronts associated with cyclones, but have been observed in other

environments. When these conditions exist, theory predicts that wave energy will be trapped within the duct, thus allowing the wave to propagate long distances. If the wave amplitude is sufficiently large, the vertical motion induced by the wave can be sufficient to trigger a precipitation band. For rainbands to occur, the atmosphere above the stable layer must be near saturation, a common condition over warm frontal surfaces. The structure of a propagating ducted gravity wave is shown in **Figure 6**. The location of greatest rising motion, where a precipitation band would be triggered, is a quarter of a wavelength upstream from the location of lowest pressure. Observations of precipitation bands in conditions where ducting is favorable often show this relationship in surface precipitation and pressure traces, suggesting that these bands may be forced by ducted gravity waves.

Moist symmetric instability

MSI is a two-dimensional, semigeostrophic mesoscale instability in which both gravitational and inertial body forces determine the stability of a displaced air parcel. The term symmetric refers to a basic state and resulting circulations that do not vary in a particular horizontal direction, e.g., along a baroclinic zone. In the Northern Hemisphere, the condition for inviscid, inertial instability is $\partial m_g / \partial x < 0$, where $m_g = v_g + fx$ is the geostrophic absolute momentum, v_g is the geostrophic wind in the direction perpendicular to the temperature gradient, f is the Coriolis parameter, and x is the cross-front distance, increasing toward warmer air. The condition for conditional instability is that $d\theta_{ev}^* / dz < 0$ at a level where $\Gamma_m < -dT_v / dz < \Gamma_d$. Here θ_{ev}^* is the saturation virtual equivalent potential temperature, T_v is the virtual temperature, and Γ_m and Γ_d are, respectively, the moist and dry adiabatic lapse rates. In an atmosphere where the lapse rate of θ_{ev}^* is negative when evaluated along a surface of constant m_g , a saturated air parcel can be inertially stable to horizontal displacements ($\partial m_g / \partial x > 0$) and gravitationally stable to vertical displacements ($d\theta_{ev}^* / dz > 0$), but unstable with respect to slantwise displacements. The release of this instability, termed *conditional symmetric instability*, has been suggested as a mechanism for the development of banded structure in frontal precipitation. In situations where the atmosphere is unsaturated,

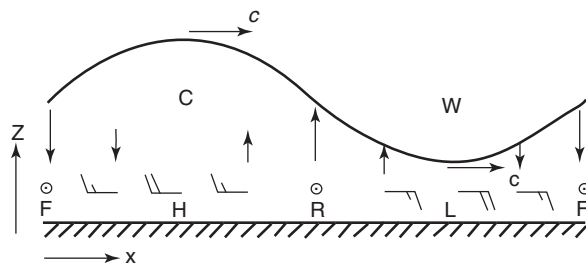


Figure 6 Idealized vertical cross section of a linear plane gravity wave with no base current, propagating to the right at speed c . The heavy sinusoidal line is a representative isentropic surface or a temperature inversion. Surface pressure extrema are labeled H and L, while cold and warm temperature anomalies are denoted by K and W, respectively. From Bosart, L.F., Sanders, F., 1986. Mesoscale structure in the megapolitan snowstorm of 11-12 February 1983. Part III: a large amplitude gravity wave. *J. Atmos. Sci.*, 43, 924-939; reprinted with permission from the American Meteorological Society.

but may be brought to saturation by lifting (for example, during frontogenesis), the potential for slantwise instability to occur can be evaluated by identifying regions where the gradient of the virtual equivalent potential temperature along an m_g surface, $d\theta_{ev}^* / dz|_{m_g} \leq 0$. This condition is identical to the equivalent potential vorticity being negative. Potential and conditional symmetric instability are both types of MSI. Bands associated with MSI are aligned along the thermal wind, move with the environmental flow, and have spacing that is related to both the depth of the unstable layer and the slope of the moist isentropes. **Figure 7** shows a schematic cross section through an environment susceptible to MSI. The slantwise ascent within the bands occurs between the slope of the m_g and θ_{ev}^* surfaces. Observations in regions of banded precipitation in frontal systems have often found that $d\theta_{ev}^* / dz|_{m_g} \approx 0$, suggesting that the atmosphere may have undergone symmetric overturning, reaching a state of moist slantwise neutrality.

Seeder-feeder processes

The seeder-feeder mechanism is a combined dynamical-microphysical mechanism for organizing ice crystal plumes into banded structures. The 'seeder' component consists of elevated convection, called generating cells, near the cloud top region that creates ice particle plumes (**Figure 8**). The 'feeder' component results from broad scale ascent and moisture convergence in the lower part of the cloud, typically associated with frontogenesis. Deformation associated with frontogenesis may organize the hydrometeor plumes to produce linear banded features. Together these provide conditions for the ice particles to nucleate, grow to become snowflakes or raindrops, and organize into linear bands.

Tropical Cyclones

Tropical cyclones can be considered dynamically to consist of two distinct regions, an inner core where accelerations due to

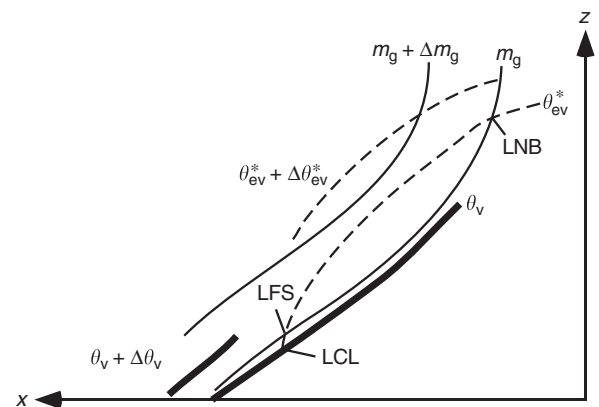


Figure 7 Idealized example of a vertical cross section in the Northern Hemisphere, normal to the thermal wind vector, showing surfaces of constant m_g (solid lines), saturation virtual equivalent potential temperature, θ_{ev}^* . In this example, θ_{ev}^* increases with height (conditional stability), m_g increases with increasing x (inertial stability). The lifting condensation level, level of free slantwise convection, and level of neutral buoyancy are denoted by LCL, LFS, and LNB, respectively.

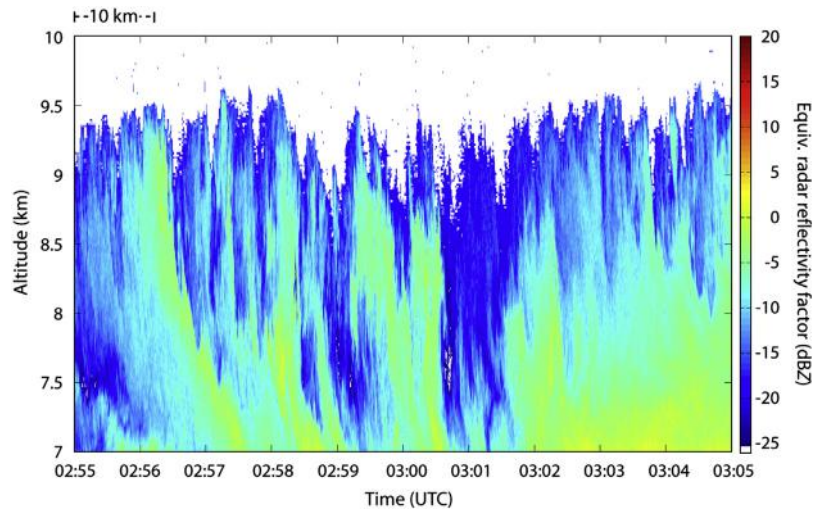


Figure 8 Radar data from the comma-head region of an extratropical cyclone showing precipitation streamers from 'seeder' generating cells near cloud top merging and increasing in intensity as they descend lower into the cloud into the 'feeder' zone.

relative rotation dominate over the Coriolis force, and an outer region where the Coriolis force is comparable to rotation in influencing storm dynamics. As a result, in the inner gyre the air trajectories form closed paths, whereas in the outer region they do not. The boundary between these regimes appears to be the boundary between bands that appear as concentric rings, such as the eyewall, and outer bands, composed of a stationary band complex and moving convective spiral bands. The location of the boundary between these two regimes varies with hurricane strength, so the stationary band complex tends to occur nearer the center of weaker hurricanes and on the peripheries of stronger hurricanes.

Analyses of hurricane structure obtained from aircraft penetrations suggest that the principal band lies along the flow streamline separating two distinct regions of the storm. On the concave, or inner side of the principal band, moist high θ_e air lies within the closed vortex circulation of the hurricane and orbits the center several times during its residence time within the vortex (see Figure 9). On the convex outside of the principal band, somewhat drier lower θ_e environmental air remains in the vortex circulation for less time than that required to orbit the center once. The principal band develops as a result of the relative motion of the vortex through its environment. In the environment of low-level easterly winds, the relative motion of the vortex leads to a region of concentrated low-level convergence along the streamline where environmental flow encounters the rotating flow within the vortex. Connecting bands generally have stratiform characteristics, although they may contain shallow convection. A stratiform band will result if a plume of ice particles originating from the eyewall falls through the 0°C level after spiraling outward at upper levels of the storm. The connecting bands in hurricanes generally sharply cross low-level streamlines, suggesting that they arise from this process.

Propagating bands also exist in tropical storms. Some bands propagate outward from the core vortex. These bands are thought to be associated with vortex Rossby waves or mixed

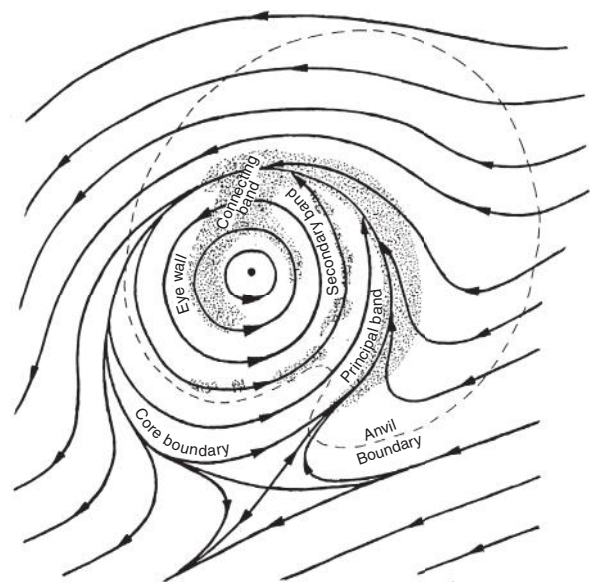


Figure 9 Schematic representation of the stationary band complex of a tropical cyclone, the bands that compose the complex, and the flow in which the complex is embedded. From Willoughby, H.E., Marks Jr., F.D., Feinberg, R.J., 1984. Stationary and moving convective bands in hurricanes. *Journal of Atmospheric Sciences* 41, 3189–3211; reproduced with permission from the American Meteorological Society.

Rossby waves and gravity waves. Some radiating gravity waves are believed to be generated during oscillations of the storm track, although stationary tropical cyclones are known to be capable of radiating gravity waves and rainbands. Track oscillations are thought to be excited both when tropical storms encounter land and by normal-mode oscillations of the hurricane vortex. There are also inward propagating rainbands, which likely form due to low-level convergence. Inward

propagating rainbands sometimes result in secondary eyewall formation as an outer rainband contracts into a closed circular band surrounding the primary eyewall and eventually contracts and replaces the inner eyewall as the inner eyewall contracts and dissipates.

Other Precipitation Bands

Although precipitation bands most commonly appear in extratropical and tropical cyclones, they also develop outside these circulations, primarily forced by variations in topography and surface conditions. Flow over or around topographic features, such as a mountain range or an island, can create elongated zones of convergence and vertical motion that create precipitation bands. For example, the Island of Hawaii, which consists of two volcanic mountains exceeding 4000 m elevation, lies in the northeasterly trade wind regime of the North Pacific Ocean. As the trade winds encounter Hawaii, they are unable to flow over the island because of an inversion at the top of the trade wind layer and must, instead, flow around the island. At night, cool drainage flow originating in the higher elevations of the island often flows offshore on the east (upwind) side of the island and meets the incoming trade winds. The boundary between the drainage flow and the trade winds leads to a persistent zone of convergence, which forces vertical motion, clouds, and precipitation. The deformation of the trade wind flow around the island organizes the precipitation into rainbands that align approximately parallel the shoreline, as shown in Figure 10.

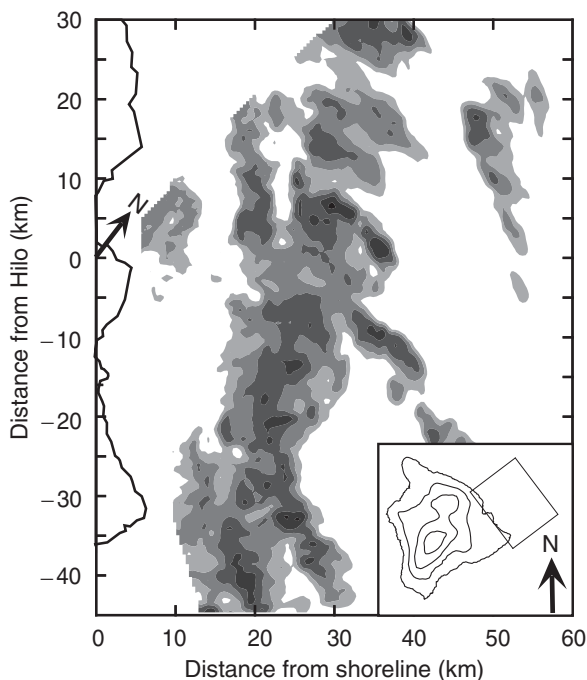


Figure 10 Radar reflectivity on 3 August 1990, measured by a radar located along the northeast shore of the island of Hawaii, showing a rainband located just offshore of the island.

Variations in surface properties can also lead to the formation of precipitation bands. In fall and winter months, when cold, arctic air masses move across warmer water bodies such as the Great Lakes, lake-effect snowstorms may occur. Under suitable conditions of lake–air temperature difference, wind speed and direction, extent of fetch over the lake, and environmental stability upstream of the lake, strong organized convective precipitation bands may develop. Specifically, three morphological types of precipitation bands have been identified to occur over the Great Lakes in wintertime. They are wind parallel bands, shoreline bands, and midlake bands. For example, when air over the lake is very cold, and winds are either weak or parallel to the long axis of the lake, a long precipitation band will sometimes develop near the center of a lake approximately parallel to the lake's shores (Figure 11). This band develops as a result of a land breeze circulation between the lake and its shorelines. Air residing over the lake, heated by the warm lake surface, becomes unstable and rises. Cooler air flows inward from both shorelines to replace the rising air. The air from shore, in turn, is heated as it flows over the lake. A circulation develops with air flowing inward over the lake, rising in a narrow zone at the lake center, and returning shoreward aloft while slowly descending. On the other hand, shoreline bands, which are usually weaker than midlake bands, develop along or close to the shoreline of a lake, when winds have a cross-lake component. Such lake-effect snowbands have been shown to result in locally heavy snowstorms in the lee of the lakes.

Other local topographic and geographic features, such as mountains, shorelines, islands, and bays, frequently induce vertical motions organized along lines. The precipitation bands that form often have characteristics unique to the specific region, but can have a significant local impact on rain or snowfall.

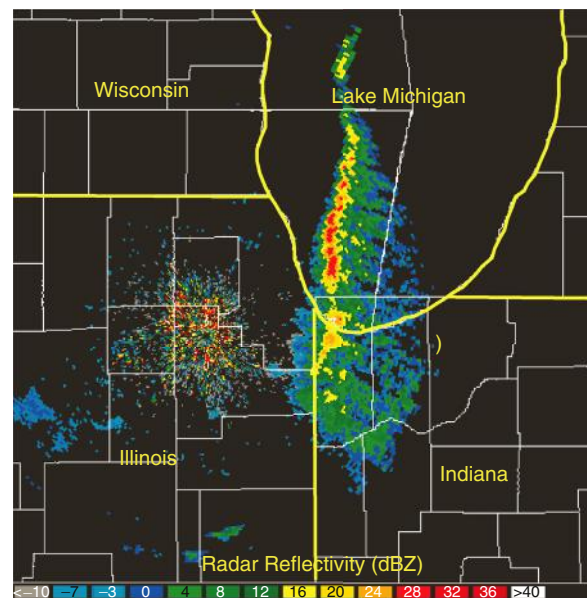


Figure 11 Radar reflectivity on 7 March 1996, measured by the WSR-88D radar located in Romeoville, Illinois. A lake-effect snowband extends down the center of Lake Michigan and onto the southern shoreline.

See also: **Dynamical Meteorology:** Inertial Instability; Symmetric Stability; Wave-CISK. **Gravity Waves:** Overview. **Mesoscale Meteorology:** Convective Storms: Overview; Mesoscale Convective Systems. **Mountain Meteorology:** Land and Sea Breezes. **Numerical Models:** Mesoscale Atmospheric Modeling; Parameterization of Physical Processes: Clouds. **Radar:** Cloud Radar; Incoherent Scatter Radar; Mesosphere–Stratosphere–Troposphere and Stratosphere–Troposphere Radars and Wind Profilers; Meteor Radar; Polarimetric Doppler Weather Radar; Precipitation Radar; Synthetic Aperture Radar (Land Surface Applications). **Satellites and Satellite Remote Sensing:** Precipitation. **Synoptic Meteorology:** Extratropical Cyclones; Frontogenesis; Fronts; Lake-Effect Storms. **Tropical Cyclones and Hurricanes:** Hurricanes: Observation.

Further Reading

- Atkinson, B.W., 1981. Meso-scale Atmospheric Circulations. Academic Press, p. 495.
- Bluestein, H.B., 1993. Synoptic-Dynamic Meteorology. In: Observation and Theory of Weather Systems, vol. II. Oxford University Press, p. 594.
- Bosart, L.F., Sanders, F., 1986. Mesoscale structure in the megalopolitan snowstorm of 11–12 February 1983. Part III: a large amplitude gravity wave. *Journal of Atmospheric Sciences* 43, 924–939.
- Cotton, W.R., Anthes, R.A., 1989. Storm and Cloud Dynamics. In: *International Geophysical Series*, vol. 44. Academic Press, San Diego, p. 883.
- Emanuel, K.A., 1985. Frontal circulations in the presence of small, moist symmetric stability. *Journal of Atmospheric Sciences* 42, 1062–1071.
- Houze Jr, R.A., 1981. Structures of atmospheric precipitation systems – a global survey. *Radio Science* 16, 671–689.
- Houze Jr, R.A., 1993. Cloud Dynamics. In: *International Geophysical Series*, vol. 53. Academic Press, San Diego, New York, Boston, p. 573.
- Kelly, R.D., 1986. Mesoscale frequencies and seasonal snowfalls for different types of Lake Michigan snowstorms. *Journal of Climatology and Applied Meteorology* 25, 308–312.
- Lin, Y.-L., 2007. *Mesoscale Dynamics*. Cambridge University Press, 625 pp.
- Markowski, P., Richardson, Y., 2010. *Mesoscale Meteorology in Midlatitudes*. Wiley-Blackwell, Oxford, UK, p. 407.
- Matejka, T.J., Houze Jr, R.A., Hobbs, P.V., 1980. Microphysics and dynamics of clouds associated with mesoscale rainbands in extratropical cyclones. *Quarterly Journal of the Royal Meteorological Society* 106, 29–56.
- Ray, P. (Ed.), 1986. *Mesoscale Meteorology and Forecasting*. American Meteorological Society, p. 793.
- Willoughby, H.E., Marks Jr, F.D., Feinberg, R.J., 1984. Stationary and moving convective bands in hurricanes. *Journal of Atmospheric Sciences* 41, 3189–3211.
- Willoughby, H.E., 1988. The dynamics of the tropical hurricane core. *Australian Meteorological Magazine* 36, 183–191.

Gust Fronts

R Rotunno, National Center for Atmospheric Research, Boulder, CO, USA

© 2003 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 2, pp 915–918, © 2003, Elsevier Ltd.

Introduction

Evaporation of rain within a thunderstorm produces a ground-based pool of cold air that spreads under its own weight and thereby generates gusty surface winds. The leading edge of the spreading cold pool is therefore known as the ‘gust front’. The gust front is an example of a ‘density current’ (see **Mesoscale Meteorology**: Density Currents), a phenomenon that occurs in a great variety of geophysical and engineering applications. As a characteristic feature of thunderstorms, the gust front plays a role in other aspects of thunderstorms, such as storm cell redevelopment (see **Mesoscale Meteorology**: Severe Storms).

The basic phenomenon of cold air spreading beneath a thunderstorm has been recognized in writing since the early nineteenth century. However, the term ‘gust front’ started to appear in the literature only in the early 1960s. Before that time the phenomenon was variously referred to as the pressure jump line, squall front, micro cold front, or outflow boundary, among other designations; the latter two are still used frequently.

Physics

Evaporation of Rain

Choosing the simplest case to illustrate the basic physics, imagine a vertically erect thunderstorm with rain falling through it (**Figure 1**); since the air beneath the thunderstorm is subsaturated, rain falling into it may evaporate and thereby cool the subcloud layer. Considering an isobaric process in which the rain evaporates until the subcloud layer is saturated, the first law of thermodynamics gives [1].

$$T_f - T_i = -\frac{L}{c_p} [q_{vs}(T_f) - q_{vi}] \quad [1]$$

Here T_i and q_{vi} are the initial temperature and water vapor mixing ratio (mass of water vapor per unit mass of dry air) of the subcloud layer, respectively, T_f is final temperature at saturation, and q_{vs} is the saturation mixing ratio; the latent heat of vaporization $L = 2.5 \times 10^6 \text{ J kg}^{-1}$ and $c_p = 1006 \text{ J kg}^{-1} \text{ K}^{-1}$ is the heat capacity for dry air at constant pressure (the effect of water vapor on the heat capacity has been ignored). The saturation mixing ratio is related to the pressure p and saturation vapor pressure $e_v(T)$ through eqn [2].

$$q_{vs} = \frac{R_d}{R_v} \frac{e_v}{p - e_v} \approx \varepsilon \frac{e_v}{p} \quad [2]$$

Here $\varepsilon \equiv R_d/R_v$, and $R_d = 287 \text{ J kg}^{-1} \text{ K}^{-1}$ and $R_v = 462 \text{ J kg}^{-1} \text{ K}^{-1}$ are the gas constants for dry air and water vapor, respectively. To obtain an explicit formula for T_f , one can expand q_{vs} in a Taylor series as in eqn [3].

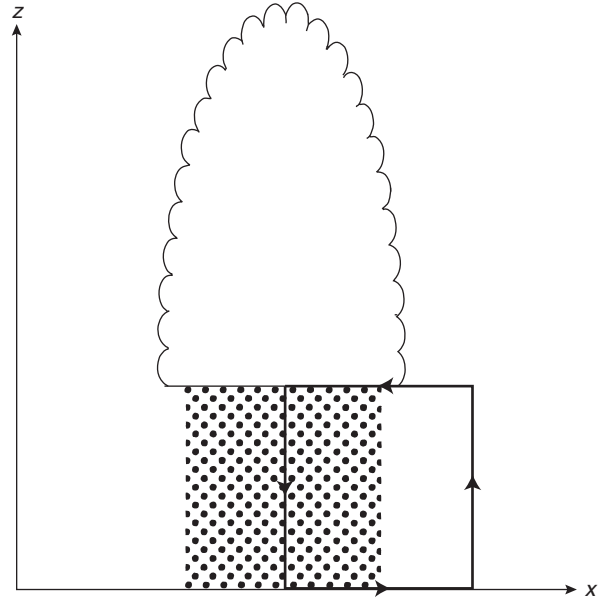


Figure 1 Schematic diagram of a rain shaft below a thunderstorm. Evaporative cooling and weight of rain produce negative buoyancy and hence negative circulation C on the circuit shown (adopting the convention that the line element $\mathbf{l}$ on the circuit points in the clockwise direction).

$$q_{vs}(T) \approx q_{vs}(T_i) + \left. \frac{dq_{vs}}{dT} \right|_{T=T_i} (T - T_i) \quad [3]$$

Using the second part of eqn [2] to calculate the derivative in eqn [3], and then substituting for de_v/dT from the Clausius–Clapeyron equation, yields eqn [4] where RH_i is the initial subcloud relative humidity.

$$T_f - T_i = -\frac{(L/c_p)q_{vs}(T_i)(1 - RH_i)}{1 + \frac{L}{c_p} \frac{Lq_{vs}(T_i)}{R_v T_i^2}} \quad [4]$$

With typical subcloud values of $p = 900 \text{ hPa}$, $T_i = 293 \text{ K}$, and $RH_i = 0.7$, eqn [4] gives $T_f - T_i \approx -3.5 \text{ K}$. While this is not untypical of observed values, $T_f - T_i \sim -10 \text{ K}$ can be obtained in situations more complex than that depicted in **Figure 1**. For example, if drier mid-level air is entrained into the thunderstorm, RH_i in eqn [4] can be much smaller, and so $T_f - T_i$ is much lower. In any case, the basic idea expressed in eqn [4] is the same: evaporation of rain produces colder temperatures at low altitude in a thunderstorm.

Generation of Motion

Fluid motion is governed by Newton’s second law, expressed as in eqn [5], where u is the velocity, p the pressure, and ρ the

density of the fluid mixture of dry air and water substance; $-\rho g$ is the external body force (per unit volume) due to the Earth's gravity, and d/dt is the substantial derivative.

$$\frac{du}{dt} = -\frac{1}{\rho}\nabla p - g \quad [5]$$

Since we are dealing with a fluid mixture of compressible gas and liquid, we begin our development by recasting the first term on the right-hand side of eqn [5] as eqn [6].

$$-\frac{1}{\rho}\nabla p = -c_p\theta\frac{1+\varepsilon^{-1}q_v}{1+q_v+q_l}\nabla\pi \approx -c_p\theta[1+0.61q_v-q_l]\nabla\pi \quad [6]$$

Here $\pi = (p/p_{00})^{R_d/c_p}$ (p_{00} is a reference pressure), the potential temperature $\theta = T\pi^{-2}$, and q_l is the mixing ratio of liquid water; the perfect gas law $p = \rho_d R_d T + \rho_v R_v T$ and the fact that $\rho = \rho_d + \rho_v + \rho_l$ have been used. The second part of eqn [6] is an approximation based on the typical smallness of the water substance mixing ratios.

For clarity of exposition, imagine that the subcloud air shown in Figure 1 is initially motionless, although rain has fallen into it; the gust front originates with the circulation $C(= \oint \mathbf{u} \cdot d\mathbf{l})$ created around the circuit shown. From eqns [5] and [6] one can derive eqn [7] for C

$$\frac{dC}{dt} = \oint -c_p\theta[1+0.61q_v-q_l]d\pi \quad [7]$$

Considering as a reference a rain-free ($q_{l0} = 0$) atmosphere with constant $\theta_{v0} (= \theta_0 + 0.61q_{v0})$ in hydrostatic balance ($d\pi_0/dz = -g/c_p\theta_{v0}$), and expressing the dependent variables as the reference state value plus a small deviation, eqn [7] may be approximated as eqn [8].

$$\frac{dC}{dt} = \oint g \left[\frac{\theta'}{\theta_0} + 0.61q'_v - q'_l \right] dz \quad [8]$$

The prime denotes the deviation from the reference state, and terms involving products of water substance mixing ratios have been neglected. The integrand of eqn [8] is the buoyancy and is composed of three terms. The first term is the thermal buoyancy, which in the case under discussion is negative since $\theta' \approx T'\pi_0^{-1}$ and, as previously demonstrated, evaporative cooling produces $T' < 0$. The second term represents the contribution to buoyancy by the presence of water vapor which, being lighter than air, increases the buoyancy when $q'_v > 0$ (as it is in the present case). (Negative thermal buoyancy produced by evaporation of rain is numerically much greater than the offsetting effect on positive buoyancy of the accompanying increase in water vapor.) The third term represents the downward force that liquid water exerts on the air in which it resides. Although the negative buoyancy associated with rain-water loading can be significant (e.g., the negative buoyancy associated with $q_l = 10 \text{ g kg}^{-1}$ is approximately equivalent to that associated with $\theta' \approx -3 \text{ K}$), in most cases thermal buoyancy dominates.

The foregoing is simply a precise way of saying that cold, rainy air sinks and spreads out along the ground. Understanding exactly how it does so requires a model.

Models of Motion

The simplest model is to consider the initial volume of cold air as a rectangle (with unit breadth into the page) that retains its

rectangular shape as it spreads. The model predicts that at late time the front speed decreases as $t^{-1/3}$, while the height decreases as $t^{-2/3}$. However, observations (Figure 2) indicate that outflows from thunderstorms move with roughly constant speed and height.

A more sophisticated model can be obtained by considering the atmosphere as composed of two fluid layers of different potential temperatures, and using the hydrostatic approximation throughout (see Mesoscale Meteorology: Density Currents). With these assumptions, and restricting attention to motion in the x - z plane, the full x -momentum eqn [5] may be reduced to the so-called shallow water momentum equation (eqn [9]).

$$\partial_t u + u\partial_x u + g'\partial_x h = 0 \quad [9]$$

The equation expressing conservation of mass becomes eqn [10], where h and u are, respectively, the depth and horizontal speed of the cold air with potential temperature θ_0 spreading into an environment with $\theta = \theta_1$; $g' \equiv g\theta_0^{-1}(\theta_1 - \theta_0)$ is the reduced gravity.

$$\partial_t h + u\partial_x h + h\partial_x u = 0 \quad [10]$$

Using the method of characteristics, the solution of eqns [9] and [10] for the so-called 'dam-break' problem was obtained in the nineteenth century and is depicted in Figure 3(a). These solutions are characterized by a wave of depression moving into the initial reservoir and a constant speed outflow with depth falling to zero. Observations and laboratory tests of the model predictions failed to show the predicted time-varying parabolic shape near the leading edge, but rather a zone of nearly constant height falling abruptly to zero, as shown in Figure 2 and depicted in Figure 3(b). Recognizing this deficiency, researchers in the early 1960s solved eqns [9] and [10] with the condition of eqn [11] instead of $h \rightarrow 0$ at the front.

$$u_f = k\sqrt{g'h_f} \quad [11]$$

The constant k was left to be determined empirically. For $k \geq 1$, eqn [11] produces solutions exhibiting a zone of constant state behind the leading edge, in which u_f and h_f are given by eqn [12], where h_0 is the initial reservoir height (Figure 3(b)).

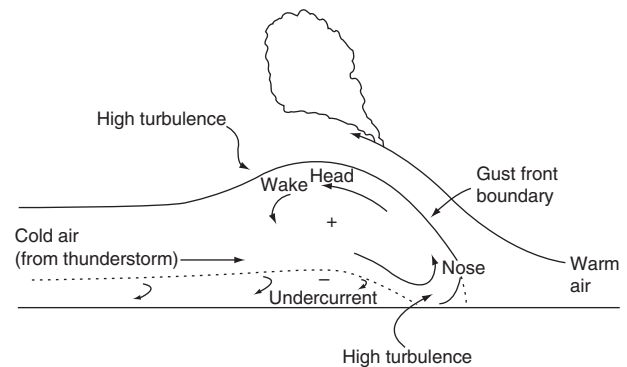


Figure 2 Schematic diagram of a gust front based on observations. Reproduced from Simpson (1997) Gravity Currents in the Environment and the Laboratory, 2nd edn. Cambridge, UK: Cambridge University Press.

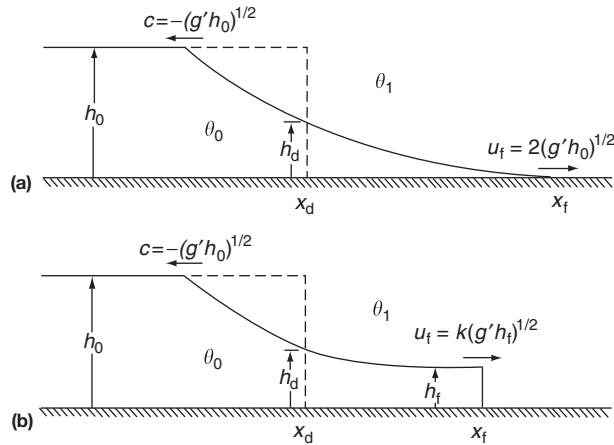


Figure 3 Models of the gust front based on the ‘dam-break’ problem with (a) $h \rightarrow 0$ or (b) $u_f = k\sqrt{g'h_0}$ at $x = x_f$. The dashed line indicates the original position of the dense rain-cooled air; x_d and h_0 indicate, respectively, the position of the edge (i.e., the ‘dam’) and the height of the rain-cooled air; and h_d is the height of the outflow at the dam site. Reproduced from Klemp J, Rotunno R, and Skamarock WC (1994) On the dynamics of gravity currents in a channel. *Journal of Fluid Mechanics* 269: 169–198.

$$u_f = \frac{2}{2+k} \sqrt{g'h_0} \quad [12]$$

$$h_f = h_0 \left(\frac{2k}{2+k} \right)^2$$

Although these solutions have the observed features of constant (u_f , h_f) behind the front, one has no way of determining k from eqns [9] and [10]. To make such a determination, one has to return to the full x -momentum eqn [5].

The reason why the simple solution shown in Figure 3(a) is not realized in a two-layer fluid is that Kelvin–Helmholtz instability produces a nontrivial coupling between the cold pool and the environment into which it spreads. Although the details of the instability and turbulence at the interface are hopelessly complicated, one can deduce the relation between u_f and h_f from a simple control volume analysis. Taking as an observational fact that a dense fluid of height $H (= h_f)$ moves steadily at speed $U (= u_f)$ into an environment of less dense fluid in a channel of depth d (Figure 4), one can deduce eqn [13] from the vertically integrated x -momentum equation along with mass conservation.

$$\frac{U}{\sqrt{g'H}} = \sqrt{\frac{2(1-\alpha)(1-0.5\alpha)}{1+\alpha}} = k \quad [13]$$

$$\alpha = \frac{H}{d}$$

(The restriction of the flow to the finite depth d is an attempt to account for the effect of a stable layer located above the generally constant θ air in the subcloud layer.) The steady flow depicted in Figure 4, and described by eqn [13], is generally known as a ‘density current’ (or ‘gravity current’), and occurs in a wide range of geophysical and engineering problems (see *Mesoscale Meteorology: Density Currents*). With the parameter k determined by eqn [13], eqn [12] now expresses the motion in terms of the external parameters of the problem.

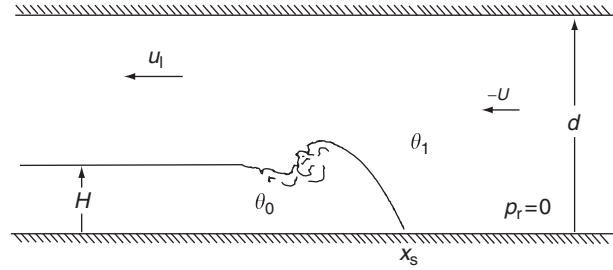


Figure 4 Moving at the constant density current speed U , a control volume analysis of the mass and momentum equations can be done to derive eqn [13], which relates U to H , d and $g' \equiv g\theta_0^{-1}(\theta_1 - \theta_0)$. Reproduced from Klemp J, Rotunno R, and Skamarock WC (1994) On the dynamics of gravity currents in a channel. *Journal of Fluid Mechanics* 269: 169–198.

The connection described here between the production of buoyancy, circulation, and cold outflow constitutes the minimal model for understanding the origin and nature of gust fronts. Neglected, but potentially important for quantitative prediction, are the effects of surface friction, environmental stratification, and time dependence of the buoyancy source.

Influence of the Gust Front

In addition to bringing relief from the heat of the day, the cold outflow from a thunderstorm displaces air in its path upward (Figure 2) and so may regenerate a new thunderstorm cell. Experience and models have shown that the regenerated cells occur in a preferred compass direction if the prevailing wind increases with height. So, for example, if the prevailing westerly wind increases with height, the gust front from a thunderstorm produces a new cell on its east side; a collection of such cells in close proximity, all regenerating cells on their east side, will soon give the system of cells the appearance of a line running north–south. The latter is termed a squall line. Since the squall line by its nature lives much longer than its constituent cells, these systems are the producers of copious rain, and frequently, severe weather such as flash floods, tornadoes and other high wind phenomena.

See also: *Dynamical Meteorology:* Hydraulic Flow; Kelvin–Helmholtz Instability. *Mesoscale Meteorology:* Density Currents; Severe Storms. *Thermodynamics:* Saturated Adiabatic Processes.

Further Reading

- Emanuel, K.A., 1994. *Atmospheric Convection*. Oxford University Press, Oxford, UK.
- Houze, R.A., 1993. *Cloud Dynamics*. Academic Press, New York, NY.
- Klemp, J., Rotunno, R., Skamarock, W.C., 1994. On the dynamics of gravity currents in a channel. *Journal of Fluid Mechanics* 135, 169–198.
- Simpson, J.E., 1997. *Gravity Currents in the Environment and the Laboratory*, 2nd edn. Cambridge University Press, Cambridge, UK.
- Stoker, J.J., 1957. *Water Waves*. Interscience, New York, NY.

Hail and Hailstorms

C Knight and N Knight, National Center for Atmospheric Research, Boulder, CO, USA

HE Brooks, National Oceanic and Atmospheric Administration, Norman, OK, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by C Knight, N Knight, volume 3, pp 924–929, © 2003, Elsevier Ltd.

Synopsis

Hail forms as ice particles collect and freeze supercooled water drops within clouds, at temperatures below freezing. The primary variables that control growth rate are the content of supercooled liquid water within clouds and the terminal velocities of the hailstones, and the ultimate hailstone sizes at the ground depend upon the updraft velocities they encounter while growing and, again, their terminal fall speeds. They form primarily in the more severe thunderstorms, which are often quite turbulent, and consequently they encounter complicated successions of growth environments. They often have several growth layers that reflect changes of the liquid water content and temperature in their environment. Small hailstones usually are approximately spherical but may be soft or slushy; whereas, large hail usually is not spherical (the word 'diameter' is misleading when used to mean longest dimension) but usually has a density near that of solid ice. As a phenomenon of severe storms, large hail correlates well with tornadoes and hailstorms generally produce lightning, though lightning often occurs without hail. Hail suppression attempts by cloud seeding are carried out in several countries, and remote detection of hail by radar is recently operational in the United States.

Introduction

Hailstones are balls of ice that typically fall from cumulonimbus clouds. By convention, they must be greater than 5 mm in diameter but their composition, size, and shape are variable. The largest hailstones can have the longest dimensions of 15 cm or more.

Hailstone amounts are also highly variable, but generally the largest hailstones and the heaviest hailfalls are from the most severe storms; that is, storms with the strongest updrafts, tallest cloud tops, and largest size. Thus hail is correlated with tornadoes, and also with lightning, though many storms produce lightning with no hail at the ground. Hail is not as well correlated with flooding, which often results from long-lasting and slow-moving precipitation systems without strong updrafts that do not produce hail.

Hailstones include various amounts of air bubbles, often in layers that indicate growth stages, but when larger than about 2 cm in diameter their densities are usually within 5% of solid ice, 0.91 g cm^{-3} . However, hail may be slushy, containing significant amounts of liquid water, and, especially at small sizes, the air content may be sufficient that the hail is soft. Soft hail is distinguished from graupel (accumulations of rime on snow particles or small frozen water drops) only by size, and since nearly all hail falls through a thick layer of air above the freezing point before reaching the ground, soft hail is often slushy, from partial melting, when it falls. Much rainfall from cumulonimbus clouds in temperate climates is melted graupel and small hail.

Fundamental Concepts of Hail Formation

Hail forms by the accretion of water droplets onto ice particles falling through supercooled cloud. The basic elements

needed for understanding the principles of hail formation are as follows.

The Updraft and Its Consequences

Humid air rising in the cores of cumulus clouds cools as it rises. The cooling causes the condensation and growth of water droplets, forming visible cloud. The droplets supercool when the temperature falls below the freezing point. These water droplets are very small, and rise in the updraft almost as fast as the air rises, typically several tens of meters per second in hailstorms, because their terminal fall speeds are less than 10 cm s^{-1} . An ice particle big enough to have a higher terminal fall speed, within such an updraft, collides with and collects the supercooled droplets, which freeze upon impact and stick to it. This accretion is the basic mechanism of hailstone growth. The initiating particle may be a snow crystal, a snowflake, or a frozen water drop. The other main role of the updraft in hail formation is to be strong enough and long lasting enough to hold the hailstones aloft, within supercooled cloud above the freezing level, long enough to grow to their final sizes. If they are to reach the ground as hail, they must then be big enough not to melt on the way down.

Terminal velocities of hailstones are described by

$$V_T = \{(4g\rho_i D/3C_D\rho_a)\}^{1/2} \quad [1]$$

where C_D is the drag coefficient, a dimensionless quantity that expresses how the drag force (the air resistance) relates to the velocity and the fluid properties. The densities of ice and air are indicated by ρ , g is the acceleration due to gravity, and D is the diameter of the hailstone. Numerical values are plotted in [Figure 1](#), for different values of C_D and, with $C_D = 0.55$, for both sea level and 500 mb pressure. Pressure of 500 mb corresponds very roughly to -10°C and 6 km above sea level, with considerable variability depending upon local conditions.

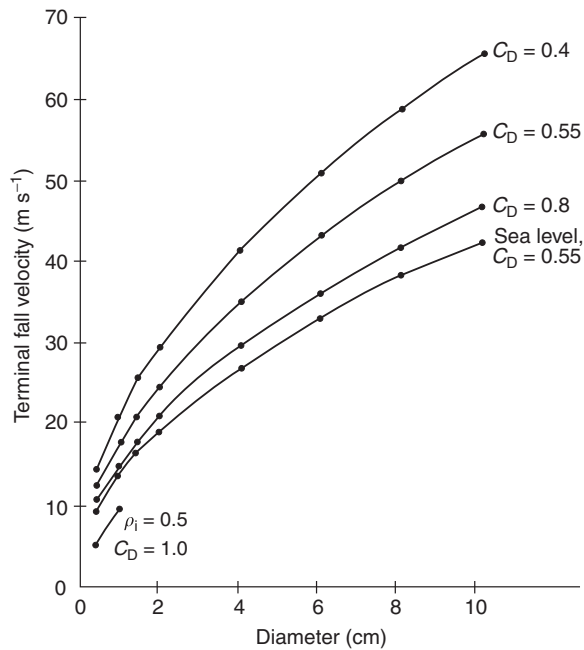


Figure 1 Terminal fall velocities of spherical hailstones plotted against diameter, for the drag coefficients (C_D) indicated, calculated for a pressure of 500 mb (50 kPa, about 6 km above sea level) and a hailstone density of 0.9 g cm^{-3} , except where otherwise indicated. From Knight, C.A., Knight, N.C., 2001. Hailstorms. In: Charles A. Doswell III. (Ed.), *Severe Convective Storms*. American Meteorological Society, Boston, pp. 223–254, by permission of AMS.

The large range of values for C_D comes about because of the variability of hailstone shape, which influences fall speed considerably. Hailstone ‘diameter’ is often used, deceptively, to refer to the longest dimension. In the context of [Figure 1](#), the more appropriate meaning is the equivalent spherical diameter. A hailstone growing within an updraft may be ascending or descending, depending upon whether its terminal fall speed is less or greater than the updraft velocity.

The Thermodynamics of Hailstone Growth

Water remains in the liquid state when cooled below 0°C unless it contains a particle of a solid material, that is, an ice nucleus with the property of initiating ice. The content of ice nuclei in the atmosphere is variable, but as a rough average there might be one that acts at -20°C per liter of air and only one per many cubic meters active at -5°C . Thus clouds that are not too much colder than 0°C are often composed almost entirely of supercooled water droplets, since the droplet populations in clouds vary between about 100 and 1000 cm^{-3} . Hailstones grow within such clouds.

The droplets freeze when they collide with ice, releasing latent heat, about 80 cal g^{-1} . However, since the specific heat of water is 1 cal g^{-1} , a drop supercooled to -10°C is warmed to 0°C when only about one-eighth of it is frozen. If the remaining seven-eighths is to freeze, the rest of the latent heat must be absorbed by the environment, which in this case is the surrounding air, at -10°C . Growing hailstones are thus warmer than the surrounding air, and in extreme cases their

temperature may be 0°C and they may grow as a mixture of ice and water, called spongy hail.

Hailstone layering consists of shells with varying air bubble contents, which arise because of air temperature variations (changes in growth altitude) or cloud water content variations, which cause changes in the hailstone temperature.

Hail Falling Behavior, Shape, and Growth

Since hail grows by accreting and freezing water droplets, it grows mostly on its underside where the collisions occur. Small hailstones, falling at only $10\text{--}15 \text{ m s}^{-1}$, usually maintain a single falling orientation. This often leads to a roughly conical shape, with the growth ‘center’ on top and consecutive broadening layers beneath that constitute most of the bulk. When hailstones grow larger, the higher fall speed creates more turbulence and they tumble. Often their growth shapes are then flattened, most of the collisions occurring around a perimeter because of a rapid and symmetrical, but complicated, tumbling motion. [Figure 2](#) shows two sections through such a hailstone, illustrating the conical center that grew into a rather thick, somewhat elongated, disk. The tumbling has important consequences for shape and also for terminal fall speed and heat exchange, and it may aid in shedding any unfrozen water that otherwise might accumulate.

Hailstones larger than a few centimeters in diameter are often lumpy. When the hail grows wet with liquid water, lumps may form like icicles, from water flowing over the surface. Hailstones of this kind are seen in [Figure 3](#). The lumpy shapes often suggest aggregation of smaller hail, but sections through lumpy hail never have shown this to be the case. The lumpiness makes the tumbling more chaotic and the shapes more

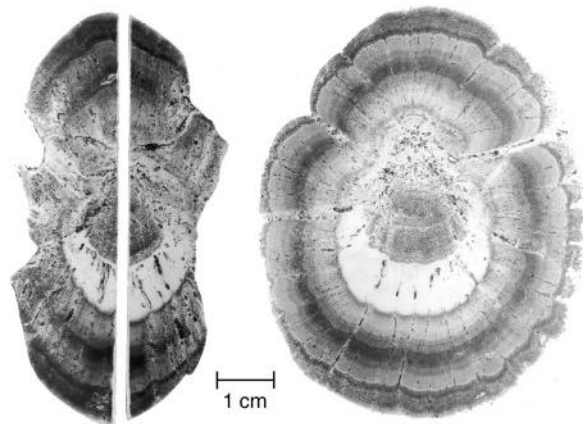


Figure 2 Two slices through a large, oblate hailstone: the first, at right, perpendicular to the short axis of the stone and the second, at left, parallel to the short and the long axes. The photos were taken with a bright background, so clear ice appears white and ice with small air bubbles is darker. The hailstone grew in a constant falling orientation and developed a conical shape up to 2–3 cm in length. After that the growth was fastest around the perimeter of the flattened hailstone, due to a rapid, complicated but symmetrical, tumbling motion. Note the growth layers (shown by differing air bubble content) that signify changes in the growth environment.

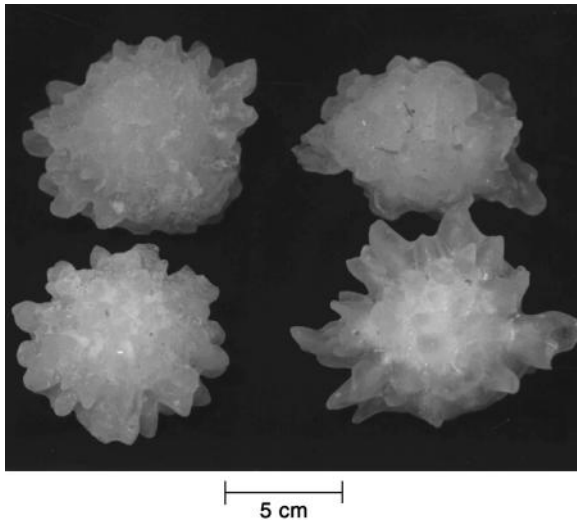


Figure 3 Four hailstones greater than 5 cm in diameter that show prominent projections. These form as the hail grows while wet, with liquid water flowing over the hailstone surface, and grow in a manner similar to the formation of icicles.

complex, and also influences the fall speed by increasing the drag coefficient.

The rate of hailstone growth is expressed as

$$dD/dt = V_T WE/2\rho_i \quad [2]$$

where D is the diameter, W the cloud water content in mass per unit volume, and E the collection efficiency. $E = 1$ for the case of simple sweep-out, the collection of all water droplets in the volume of air traversed by the hailstone. In reality E is usually between roughly 0.3 and 0.8 depending upon the sizes of both the droplets and the hailstone, because some droplets in the path of the hailstone are carried around it by the airflow. **Figure 4** shows the typical range of values of growth rates as a function of hail size, for $WE = 2.5 \text{ g m}^{-3}$ and $\rho_i = 0.9 \text{ g cm}^{-3}$.

Hailstorms

The necessary and sufficient conditions for hail formation in a thunderstorm are easy to state in general terms. The storm needs to provide a strong updraft containing supercooled water droplets, as the environment for hailstone growth, and into this environment must come hailstone ‘embryos.’ These are particles of ice with terminal fall speeds of a few meters per second that can grow in a few tens of minutes into hailstones. Their growth must be fairly rapid because the growth environment may disappear as the updraft weakens (the updrafts in most thunderstorms are intermittent), or the supercooled water content of the updraft may become depleted, either by evaporation as the updraft air mixes with drier surrounding air or by conversion to many small ice particles. If the ascent rate of the potential hail embryo is too great it may be elevated out of the growth region before it grows big enough to fall as hail, or if the ascent rate is too slow it may fall out before it reaches large enough size.

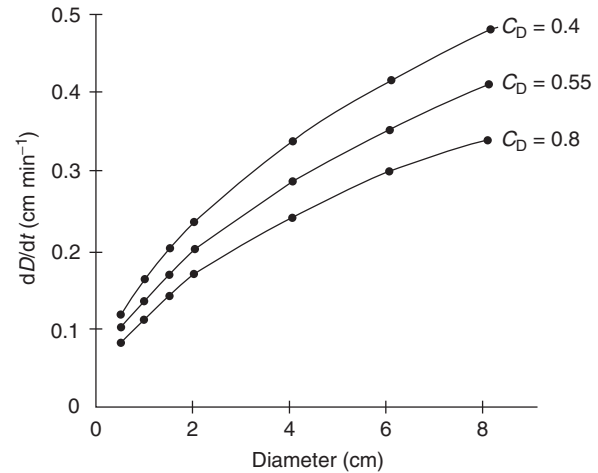


Figure 4 Hailstone growth velocities, in cm min^{-1} , as a function of diameter calculated for the three top curves in **Figure 1**, assuming an effective cloud liquid water content WE of 2.5 g m^{-3} and no shedding of liquid water. Cloud water contents vary from zero to perhaps twice this value, and growth rates vary accordingly (see eqn [2]). From Knight, C.A., Knight, N.C., 2001. Hailstorms. In: Charles A. Doswell III. (Ed.), Severe Convective Storms. American Meteorological Society, Boston, pp. 223–254, by permission of AMS.

The formation of hail, then, depends upon a critical interrelation among the three-dimensional airflow within storms, the fate of the supercooled droplets that occur within the updrafts, and the trajectories of the potential hail embryos. These factors work together to determine how much of the water vapor that condenses reevaporates in the middle and upper atmosphere, how much of it falls as rain, and how much as hail. The percentage of it that falls as hail is difficult to measure, but is usually very small.

Hail in Supercell and Multicell Storms: Steady and Time-Dependent Concepts

There are at least two specific hypotheses for hail formation that refer to two idealized storm types, the supercell storm and the multicell storm. The supercell is a type of storm that can last a long time and travel considerable distances, often producing tornadoes and often producing long swaths of large, damaging hailstones. The diagrams in **Figure 5** represent a typical case in the central United States, a map view looking down at the storm and a cross-sectional view, both indicating possible growth trajectories. The humid air entering the storm comes in a vigorous flow at the surface from the south or southeast, rises up through the middle of the storm, and exits to the east because the flow in the upper levels is from the west, that carries the southern part of the outflow in the upper levels over the top of the inflowing air at the surface. The updraft in the central part of the storm can be $30\text{--}50 \text{ m s}^{-1}$, so small ice particles that grow within the updraft do not have time to grow very large and attain much fall speed before being ejected into the outflow at the top of the storm, producing the extensive anvils characteristic of supercells. However, in the outflow there is little vertical air motion, and the idea is that some of these particles fall back into the inflow at low levels. They may

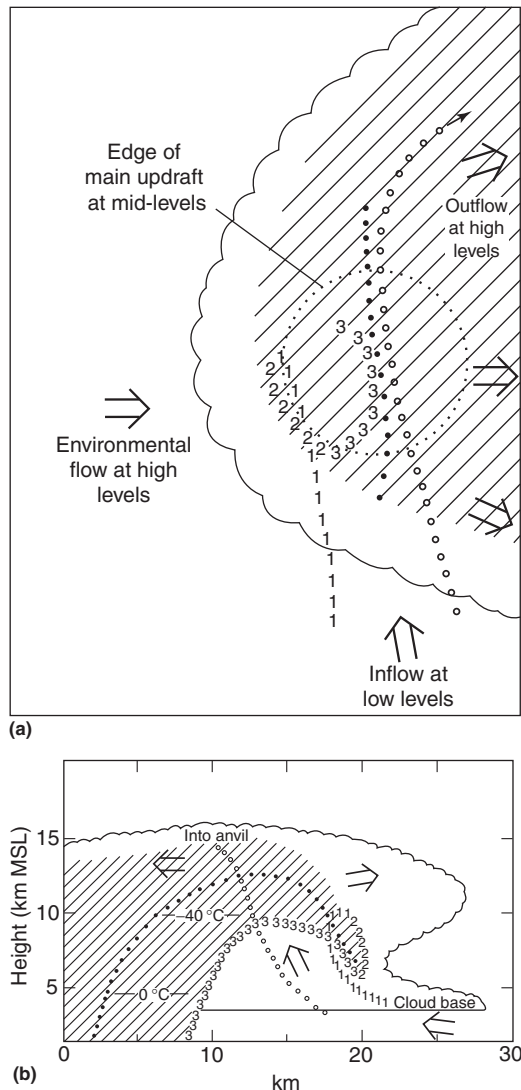


Figure 5 Schematic of airflow and some hydrometeor trajectories deduced in a supercell storm. (a) Plan view, looking down at the storm, showing low-level inflow entering from the south (the bottom of the drawing), the main updraft within the storm shown by the dotted circle, upper-level environmental airflow from the west and storm outflow to the east. The area within the cloud containing precipitation-sized hydrometeors is hatched. (b) Cross section south to north parallel to the inflow, with the upper-level flow away from the reader. The trajectory indicated by open circles in both views represents direct growth in the main updraft: condensation, freezing, and some riming leading to small ice particles that travel more or less with the air and out into the upper-level anvil. The trajectory indicated by the solid circles indicates particles that start in weaker upward flow on the south side of the updraft, grow to larger sizes and higher fall velocities and fall as small hail or rain to the north of the location of the main updraft. The trajectory indicated by the numbers – 1, 2, 3 – represent large hail formation: (1) slow growth rising at the west side of the updraft, (2) descent back into the inflow while traveling in the environmental flow around the south side of the updraft, and (3) final growth as hail within the strong updraft and fallout at its northern edge. Adapted from Browning, K.A., 1977. In: Foote, G.B., Knight, C.A. (Eds.), *Hail: A Review of Hail Science and Hail Suppression*. American Meteorological Society, Boston, pp. 1–47, by permission of AMS.

become hail embryos, getting a chance to grow further while rising a second time within the updraft. Falling faster with respect to the air, they ascend more slowly and have time to grow into full-fledged hailstones.

This is a particularly simple and organized recycling of precipitation particles to form hail embryos in a storm: simplified because even most supercells are not especially in steady state, but evolve in various ways.

Multicell storms, on the other hand, are composed of individual convective cells that grow and decay in proximity to one another, sometimes in a systematic way. The individual cells may last 45–60 min, as their updrafts increase up to a maximum and then die out. This can be enough time to produce significant hail, and if the initial updraft is not too strong, embryos may have time to grow and attain a fall speed of several meters per second without being elevated too high in the cloud. Now as the updraft strengthens the embryos may already be there in the right locations and ready to form hailstones.

Multicell hailstorms typically produce spotty, discontinuous hailfalls, which may be organized within a larger hailstorm if the multicell is an organized one or may be distributed irregularly.

Hail Suppression

Hail suppression by seeding clouds with artificial ice nuclei has been practiced in many parts of the world for several decades. It is still widespread, although it is controversial and there has been no definitive demonstration of positive effects. The main idea is that, furnishing ice nuclei might increase the number of potential hail embryos in the clouds, thereby depleting the supercooled water earlier and reducing the hailstone size. If the size is reduced enough, most or all of the hail may then melt before reaching the ground. Ideally, the hail would be suppressed and the rainfall increased. There has been a great deal of discussion about optimizing the seeding materials and the timing and location of seeding, but the knowledge of the natural evolution of the ice contents within hailstorms is still rudimentary, so it may be many years before a consensus is reached on the prospects of hail suppression.

Hail Climatology

Hailfall at the ground is a small-scale phenomenon generally affecting areas of one to a few tens of square kilometers, though much larger hailfalls have been documented. It is highly variable and poorly resolved by routine weather observation networks. In many countries, most reports come from target-of-opportunity observations from members of the general public or public safety officials. In some limited areas, hailpad networks provide measurements of hail size, but in most parts of the world, size is estimated. Many small hailfalls are missed altogether. Thus, data on hail climatology are statistical in nature and much of the information derives from insurance records. Recently, estimates based on the environmental conditions in which hail occurs have been

developed, but they are based on data from North America (see Figure 9 in **Weather Forecasting: Severe Weather Forecasting**).

Consideration of the environments favorable for hail formation leads to understanding the preferred locations of hailfall. Exceptionally high humidity at low levels along with exceptionally cold temperatures aloft are the ingredients for the instability that leads to thunderstorm updrafts strong enough to support large hail. Strong changes of the horizontal wind with height (as depicted in **Figure 5(a)**) also contribute to updraft velocity by influencing the vertical pressure gradient. Both of these influences are common to the east of high mountain ranges in middle latitudes. The predominant westerlies aloft, over warm, moist winds from nearer the equator near the ground, are the reason why thunderstorms with large hail are generally found east of the Rocky Mountains in North America, the Andes in South America, and the Himalayas in Asia. Although the regions east of those three large mountain ranges are where the largest hail (>5 cm diameter) most commonly falls, areas near the Pyrenees and Alps experience smaller hail where both the instability and the wind shear are usually less pronounced. With the high-value and sensitive crops in those regions, such as wine grapes, crop losses from small hail can be very important.

Tropical thunderstorms are less likely to produce hail than their midlatitude counterparts for two primary reasons. First, there is less atmospheric instability, so that updrafts may not be as strong, and second, the height of the freezing level is far above the ground. As a result, hail that forms tends to be small and melt on the way down, arriving at the surface as rain. Hail is rare at high latitudes because conditions are generally too cold and dry for thunderstorms.

Some global climate simulations predict an increase in the frequency of severe, hail-producing thunderstorms in the future. However, observations from China have indicated that the frequency of hail occurrence has decreased since the mid-1980s. Using radiosonde observations and a simple one-dimensional model of hail growth, this decrease has been associated with an observed rise in the freezing level as a result of increasing surface temperatures, indicating that the melting of hail on the way to the ground may have dominated any potential for increased creation of hail. Hailpad studies from France and Italy have indicated a slight tendency for a shift in hail size distributions toward larger sizes since the mid-1980s, consistent with the production of larger hail aloft, but with preferential melting of the small, slower-falling hail below the freezing level.

Hail Detection by Radar

Radar methods for remote sensing of hail within clouds have recently come into use for research but not yet for the operational radar networks. These rely on varying the polarization of the transmitted pulses and sensing different polarizations of the backscattered radiation. Horizontal and vertical linear polarizations are often used, and sometimes circular polarization. Intense radar echoes may be caused either by heavy rain or by hail, and until recently the interpretation has been ambiguous. However, big raindrops are consistently flattened, with larger horizontal than vertical dimensions, and this produces a substantially stronger echo with a horizontally than a vertically polarized radar. Hail gives a more nearly equal radar echo strength at the two polarizations because small hail is not flattened and larger hail tumbles so that its elongations are about equally distributed in space. Thus the difference, termed the differential reflectivity, is used as a hail signal when the radar echo itself is intense.

Widespread application of this and other advanced radar techniques will greatly increase the knowledge of hail production as related to storm structure and behavior, especially if techniques are developed to provide information about hail size and amount within storms. Radar sensing of hail, if developed to be sufficiently quantitative and used routinely, would also contribute greatly to hail climatology and hail research.

See also: **Clouds and Fog:** Cloud Microphysics. **Cryosphere:** Snow (Surface). Inadvertant Weather Modification. **Mesoscale Meteorology:** Severe Storms. **Radar:** Precipitation Radar. **Weather Forecasting:** Severe Weather Forecasting.

Further Reading

- Hail: a review of hail science and hail suppression. No. 38. In: Foote, G.B., Knight, C.A. (Eds.), *Meteorological Monographs*, vol. 16. American Meteorological Society, Boston.
- Rogers, R.R., Yau, M.K., 1989. *A short course in cloud physics*, third ed. Pergamon Press, Oxford, NY, p. 293.
- Severe convective storms. No. 50. In: Doswell III, C.C. (Ed.), *Meteorological Monographs*, vol. 28. American Meteorological Society, Boston.

Mesoscale Convective Systems

A Laing, National Center for Atmospheric Research, Boulder, CO, USA

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

Mesoscale convective systems are organized, multicellular thunderstorm systems that have a strong impact on heavy rainfall, severe weather, and vertical transport in the atmosphere. The significant types in terms of those impacts range from bow echoes at smaller scales, which cause damaging surface winds, to mesoscale convective complexes with mesoscale convective vortices at larger scales, which can produce persistent, widespread heavy rainfall and flash flooding. This article describes their structure, life cycle, large-scale environments, global distribution, impact on heat and moisture transport, chemical transport, and electricity.

Introduction

A mesoscale convective system (MCS) is an organized, multicellular thunderstorm system marked by an extensive mid- to upper-level stratiform cloud shield. Tropical and warm-season convections in midlatitudes are most often organized into mesoscale systems with length scales of 100–1000 km and typical lifetimes of 6–12 h. The definition encompasses a continuous spectrum of phenomena including midlatitude squall lines, bow echoes, mesoscale convective complexes (MCCs), tropical squall and nonsquall clusters, small multicellular storms, and even some tropical cyclones. They typically begin as cumulonimbus along a low-level convergence boundary then merge and organize upscale to form a single cloud system with large contiguous areas of rain (on the order of 100 km). A midtropospheric mesoscale convective vortex (MCV) sometimes forms in the stratiform region, can focus new convection, and may contribute to tropical cyclone genesis. Their mode of organization, strength, and longevity depend on variables such as the vertical shear, the strength of the low-level convergence, the convective available potential energy (CAPE, Appendix), and the distribution of moisture.

These systems have garnered much attention mainly because of their strong impact on heavy rainfall, severe weather, and vertical transport. The significant MCSs in terms of those impacts range from bow echoes at smaller scales, which cause damaging surface winds, to MCCs with MCVs at larger scales, which can produce persistent, widespread heavy rainfall and flash flooding. This article describes, in turn, their structure, life cycle, large-scale environment, global distribution, impact on heating and moisture transport, chemical transport, and electricity.

Structure and Life Cycle

The structure and life cycle of MCSs strongly depends on the environmental wind shear. How the system organizes and evolves depends on the balance between the horizontal vorticity of the cold pool generated by convection and horizontal vorticity associated with the low- to midtropospheric shear. The Coriolis force also influences the structure of long-lasting MCSs, particularly in the midlatitudes. MCSs become

more asymmetric with stronger Coriolis force and develop larger areas of rain on their poleward side.

Midlatitude Squall Lines (Linear MCS)

The term 'squall line' applies to any line of rapidly moving thunderstorms including large multicellular storms organized along a line, i.e., a linear MCS with length scales of 1000 km along the line and 100 km normal to the line. Squall lines are the most common type of midlatitude MCS and severe lines are most frequent in the spring, producing twice as many tornadoes, high wind, and hail as more circular MCSs.

Spring squall lines have four primary modes of genesis that all evolve into a quasi-linear MCS. A radar study of 55 major rain events over six springs in Oklahoma showed that most MCSs have a contiguous region of deep convective cells and an associated region of stratiform rain. The precipitation pattern comprises a continuous spectrum of mesoscale structures. Initial organization is into convective lines followed by stratiform rain regions with slantwise front-to-rear ascent and rear-to-front descent (Figure 1). Two-thirds of those systems were classified as asymmetric, a categorization more applicable to an MCS stage than type. Typical surface features are presquall mesolow (due to subsidence warming in the mid-to-upper troposphere), mesohigh (due to heavy precipitation and convective-scale downdrafts), and wake low (due to subsidence warming and as a surface manifestation of the descending rear-inflow jet (RIJ)). Another study of 88 linear MCSs over the US Great Plains found that 60% had similar leading line/trailing stratiform structure while those with stratiform rain *ahead* of or *parallel* to the convective line each accounted for about 20% (Figure 2).

Mesoscale Convective Complex

The MCC is a large, quasi-circular, meso- α -scale (length scale 250–2500 km) convective system that meets size, shape, and duration criteria based on satellite infrared brightness temperature (Table 1). Their association with devastating flash floods has been recognized since the late 1970s. Note that, with the exception of size, no significant differences in characteristics exist between MCCs and slightly smaller quasi-circular MCSs. The MCC definition succeeds in identifying the largest,

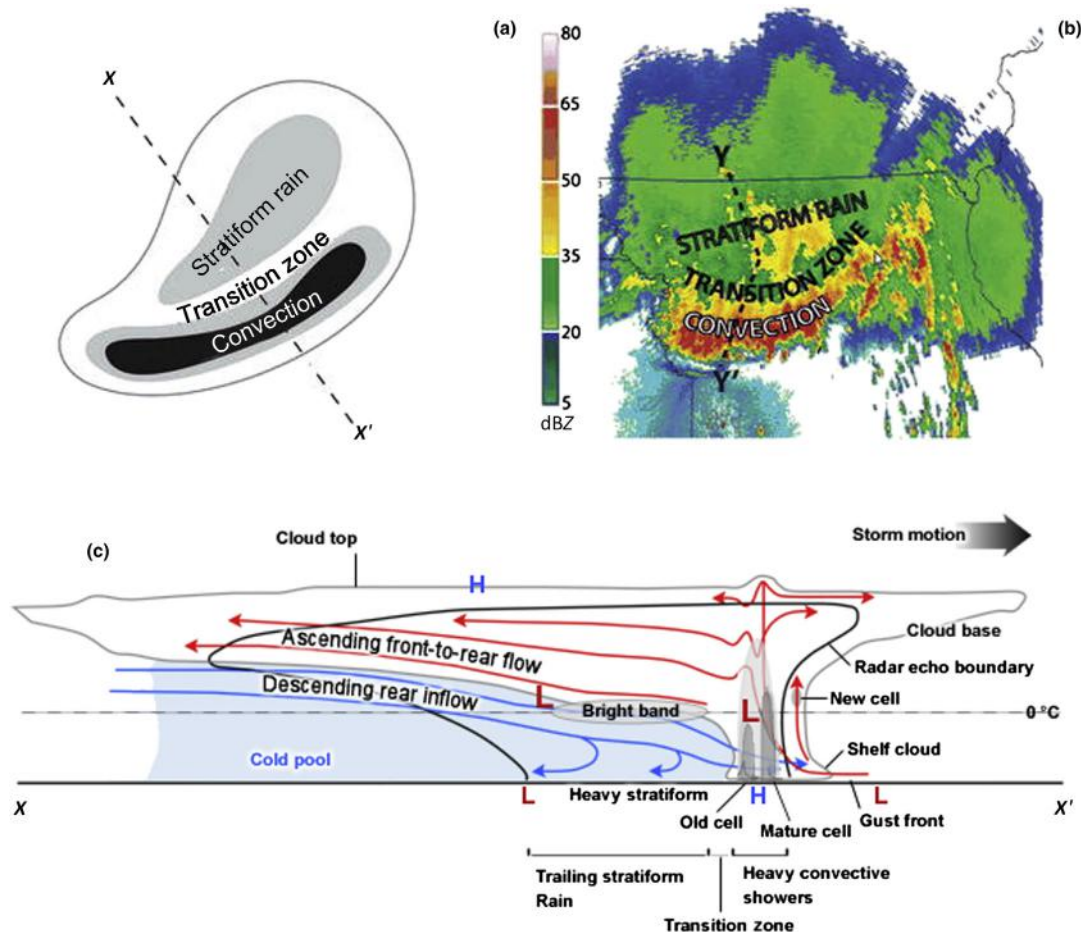


Figure 1 (a) Schematic plan view of radar reflectivity of an MCS with trailing stratiform rain. (b) Radar reflectivity of an MCS. Adapted from Smith, A.M., McFarquhar, G.M., Rauber, R.M., Grim, J.A., Timlin, M.S., Jewett, B.F., Jorgensen, D.P., 2009. Microphysical and thermodynamic structure and evolution of the trailing stratiform regions of mesoscale convective systems during BAMEX. Part I: Observations. *Monthly Weather Review* 137, 1165–1185. (c) Schematic cross section of the relative airflow through (a). Flow is perpendicular to the convective line and to the motion of the line. H and L denote the centers of positive and negative pressure perturbations. Adapted from Houze, R.A., Biggerstaff, M., Rutledge, S., Smull, B., 1989. Interpretation of Doppler weather radar displays of midlatitude mesoscale convective systems. *Bulletin of the American Meteorological Society* 70, 608–619. © The American Meteorological Society.

long-lived convective systems that are important for hydro-meteorological purposes.

All populations of MCCs (Table 2) exhibit a nocturnal life cycle. First thunderstorms usually develop in the mid- to late afternoon; a single organized system initiates by early nighttime; reaches its maximum extent after midnight; and dissipates shortly after daybreak (Figure 3).

MCCs have mean cloud shield areas of 350 000 km² and persist for an average of 11 h. Oceanic systems are, on average, slightly larger and longer lasting but less intense in terms of cloud top temperature and radar reflectivity. Not surprisingly, size and duration are positively correlated, with summer correlation values nearly twice the spring and fall. It appears that a meso- β -scale convective cycle that occurs early in the growth phase distinguishes the long-lived MCCs from other shorter-lived MCSs. Strong growth within the first 3 h is a strong predictor of the maximum extent and duration of MCCs.

In contrast to the uniform cloud top signature, MCC precipitation is highly varied. In the 1985 Oklahoma-Kansas Preliminary Regional Experiment for Stormscale Operational

Research Meteorology Central-Phase (PRE-STORM), 75% of MCCs matched the MCC criteria and displayed asymmetric precipitation patterns. Although the internal structure of some MCCs resembles linear MCSs, the surface weather can be quite different. Squall line passage is noted for more precipitous temperature, dew point, and wind changes than MCCs, which have longer periods of steady rain and rain showers.

Peak precipitation rates (over 25 mm h⁻¹) occur in the 2–6 h period after initiation and favor the equatorward flank, contributing to the reflectivity and precipitation pattern depicted in Figure 1. The growth stage also produces hail, high winds, and tornadoes. As the cloud system matures, the overall rain rate decreases, becomes more stratiform, and the rain area expands. The stratiform rainfall, 30–40% of the system total, is associated with mesoscale ascent inside clouds extending upward from the melting level. After 10–12 h, rainfall ends and deep convection ceases. A 10-year study of extreme warm season rainfall found that 22% of extreme rainfall in the Central United States occurred in MCCs, although less than 9% of all precipitation observations occurred in MCCs. Thus, MCCs

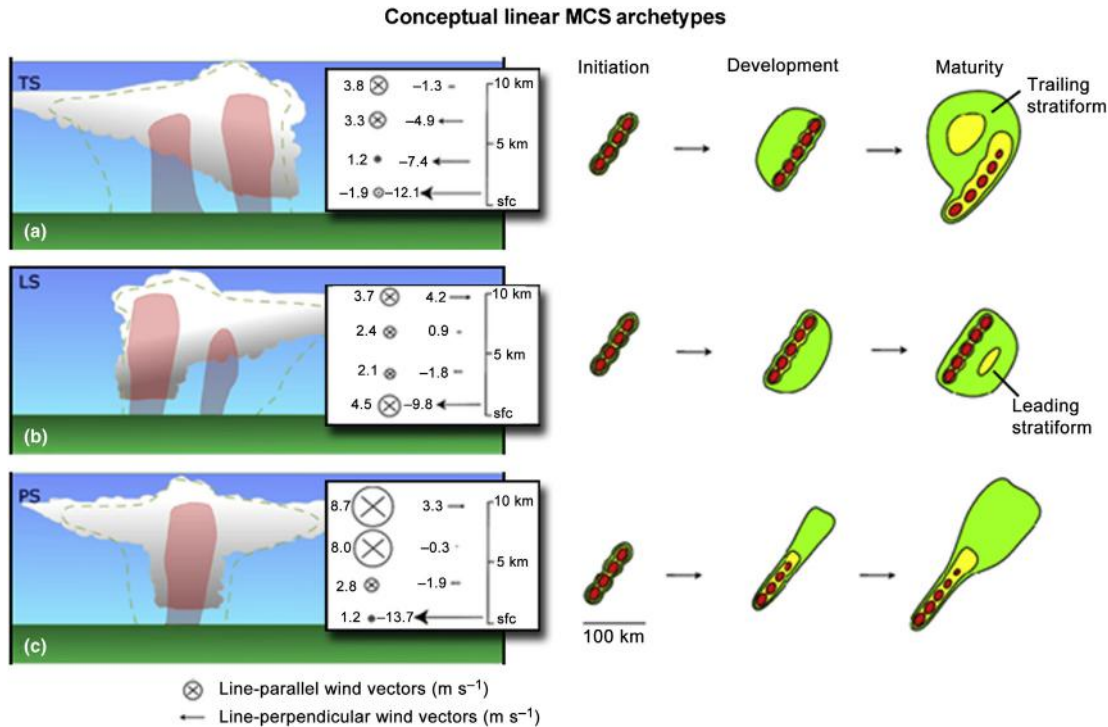


Figure 2 Schematic reflectivity drawing of idealized life cycles for three linear MCS archetypes adapted from NOAA/Parker, M.D., Johnson, R.H., 2000. Organizational modes of midlatitude mesoscale convective systems. *Monthly Weather Review* 128, 3413–3436: (a) Leading line with trailing stratiform precipitation. (b) Convective line with leading stratiform precipitation. (c) Convective line with parallel stratiform precipitation. Approximate time interval between phases: for trailing stratiform (TS), 3–4 h; for leading stratiform (LS), 2–3 h; for parallel stratiform (PS), 2–3 h. Levels of shading correspond roughly to 20, 40, and 50 dBZ.

Table 1 Mesoscale convective complex definition

Size	(A) Cloud shield with IR temperature $\leq -32^\circ\text{C}$ must have an area $\geq 100\,000\text{ km}^2$ (B) Cloud shield with IR temperature $\leq -54^\circ\text{C}$ must have an area $\geq 50\,000\text{ km}^2$
Initiate	Size definitions A and B are first satisfied
Duration	Size definitions A and B must be met for a period $\geq 6\text{ h}$
Maximum extent	Continuous cold cloud shield (IR temperature $\leq -32^\circ\text{C}$) reaches maximum size
Shape	Eccentricity (minor axis/major axis) ≥ 0.7 , a time of maximum extent
Termination	Size definitions A and B no longer satisfied

From Maddox, R.A., 1980. Mesoscale convective complexes. *Bulletin of the American Meteorological Society* 61, 1374–1387.

Table 2 Climatology of tropical squall lines

Region	Average propagating speed (m s^{-1})	Lifetime (h)	Average displacement (km)	Length (km)	Width (km)
West Africa	14.8	39.7	2100	750	433
East Atlantic	14.6	9.7			
Amazon coastal squall line	16	16	2000	1400	170
Venezuela	13	3.5	150	98	29

Adapted with permission from Cohen J., Silva, D.M., Nobre, C., 1995. Environmental conditions associated with Amazonian squall lines: a case study. *Monthly Weather Review* 123, 3163–3174.

produce extreme rainfall at a far greater rate than that produced by other sources.

MCCs and smaller circular MCSs, beneficially, contribute 30–70% of the growing season precipitation in the US Central Plains. Other locations with large MCC populations, like southeastern China, Sahelian Africa, and subtropical South America, rely on deep convection for a significant portion of their summer season rainfall. For example, during 1998–2007 the MCC contribution to rainfall in subtropical South America was 11–20% for all months and 30–50% for the midsummer. In southeastern Africa, MCCs contribute 20–24% of summertime rainfall. However, there is considerable interannual and spatial variability, with MCCs accounting for up to 40% in some warm seasons.

Bow Echo

A line of strong convective cells or a single large convective cell can evolve into an MCS known as a bow echo, because of its characteristic bow shape on radar displays. The typical evolution of a bow echo is shown in Figure 4, a pattern observed in single cells or squall lines of hundreds of kilometers horizontal scale. Bow echoes cause wide and long swaths of severe, straight-line winds called *derechos*. Derechos are defined as a family of downburst clusters with winds exceeding 33 m s^{-1} (64 kts) across an area whose major axis is at least 400 km. Large MCSs can produce multiple bow echoes and derechos.

Prominent features of bow echoes include a strong RIJ, which, if it descends to the surface, is likely to produce

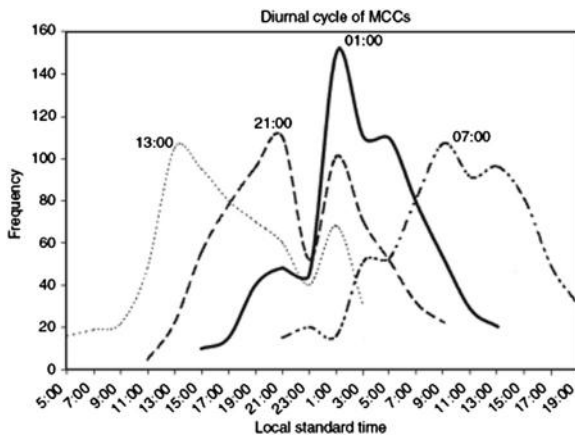


Figure 3 Life cycle of the global population of MCCs (----- first thunderstorm; --- initialization; — maximum extent; and --- termination).

damaging winds, and a rear-inflow notch, which is often evident just before and during the initial formation of the bow. Vortices, known as bookend vortices, form on the ends of the convective line. In longer-lived bow echoes, the Coriolis force

promotes stronger cyclonic shear behind the poleward end of the bow and weak anticyclonic shear equatorward of the bow (Figure 4).

Tropical Squall and Nonsquall Clusters

Tropical squall cluster or squall line is identified by a line of vigorous convective cells extending for hundreds to thousands of kilometers along its major axis. The typical tropical squall line, in prevailing easterly flow, has a rearward tilting updraft, and expanding stratiform rain area as the system matures (Figure 5).

Typical squall clusters move rapidly westward as a continuous disturbance or family of systems that can propagate for several days while undergoing cycles of regeneration. Some West African squall clusters could be classified as MCCs from their satellite signature while Amazon squall lines tend to elongate over time. Table 2 summarizes the climatology of various squall line populations.

Squall line passage is noted for a distinct roll cloud followed by a sudden wind squall. A heavy downpour then sets in, often producing 30 mm of rainfall in about 30 min. Vigorous convection is then replaced by stratiform rainfall (about 40% of system total) continuing for several hours.

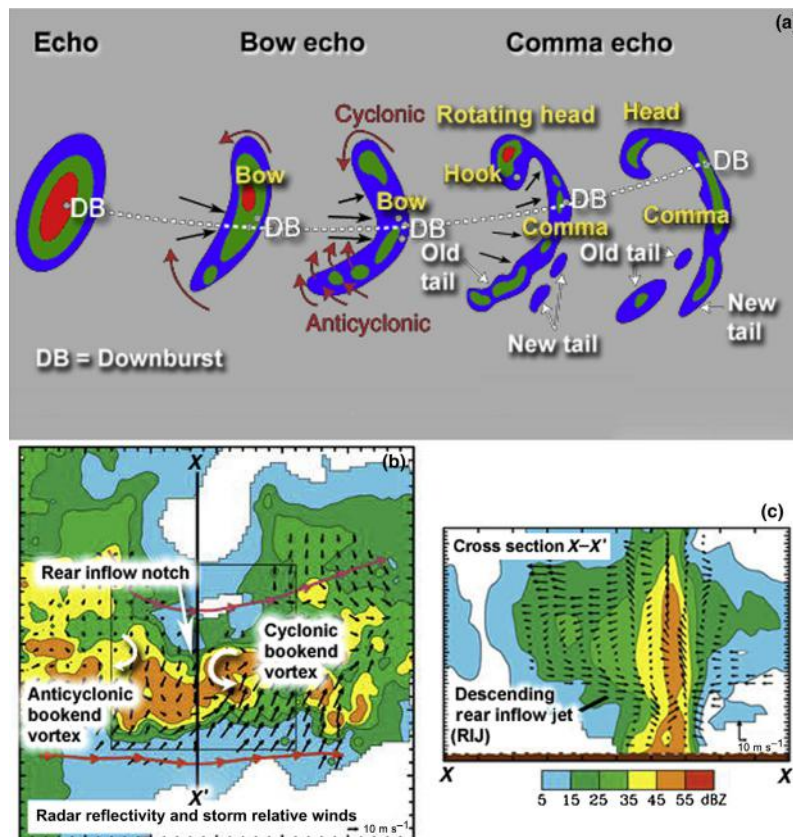


Figure 4 (a) Typical morphology of radar echoes associated with bow echoes that produce strong and extensive downbursts. Adapted from Fujita, T.T., 1978. Manual of downburst identification for project NIMROD. Satellite and mesometeorology Research Paper 156, Department of Geophysical Sciences, University of Chicago, 104 pp). Areas of cyclonic and anticyclonic rotation are capable of producing tornadoes. (b) Radar reflectivity (dBZ) and system-relative winds showing vortices and rear-inflow notch. (c) Cross section from X-X' in (b). Adapted from Grim, J.A., Rauber, R.M., McFarquhar, G.M., Jewett, B.F., Jorgensen, D.P., 2009. Development and forcing of the rear inflow jet in a rapidly developing and decaying squall line during BAMEX. Monthly Weather Review 137, 1206–1229.

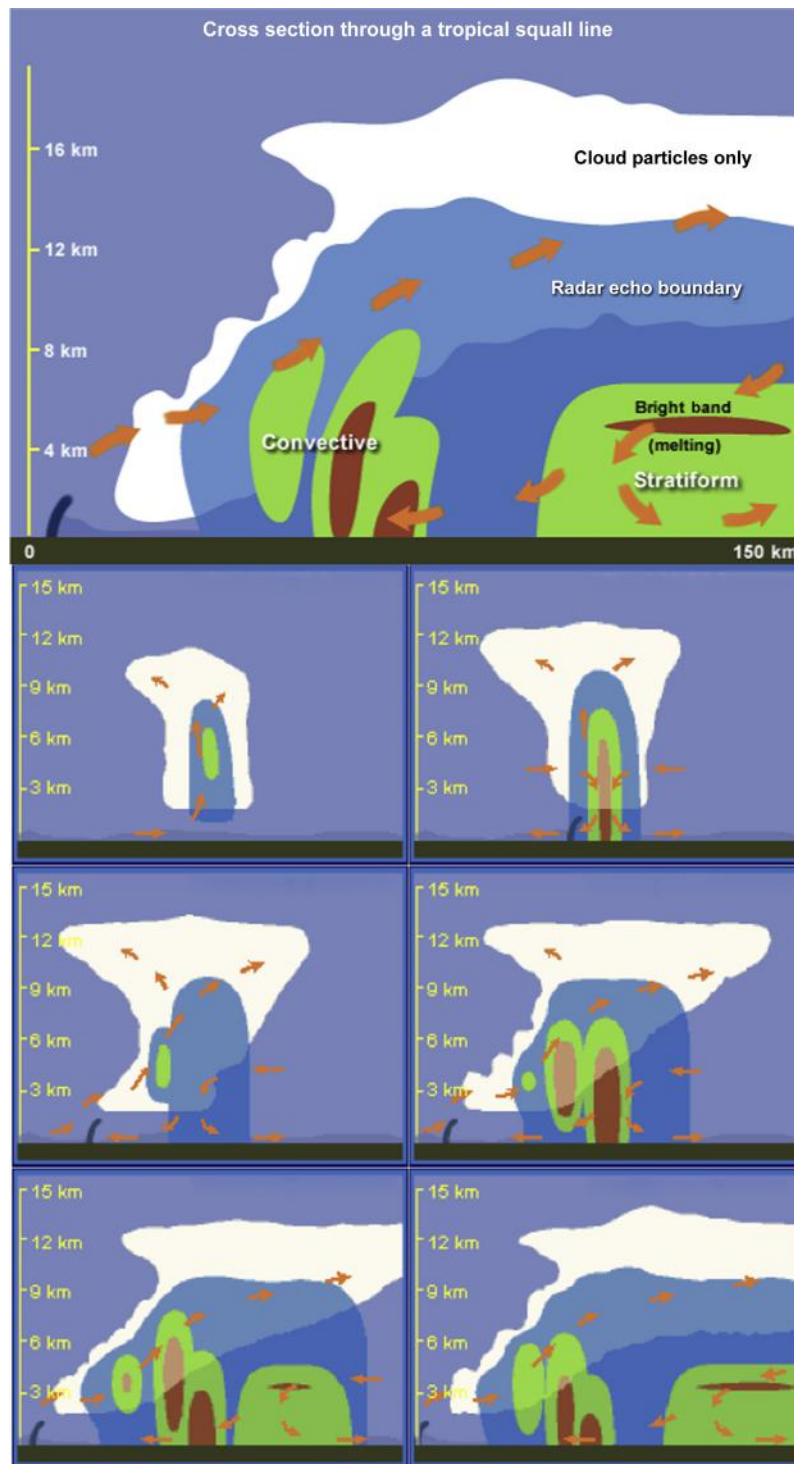


Figure 5 Schematic diagram of the kinematic structure and radar echoes of a tropical squall line with trailing stratiform rain from initial convection to maturity; viewed as a cross section perpendicular to the convective line and its motion. © The COMET Program. https://www.meted.ucar.edu/about_legal.php.

Beneath the melting level, evaporative cooling of precipitation drives mesoscale downdrafts in the wake of the squall (Figure 5). Some convection in the tropical Pacific and tropical North Africa have a 2-day cycle, with daytime surface heat initiating a strong MCS that matures at night, then

downdrafts from the MCS stabilizing the boundary layer the next day.

Eastward-moving tropical squall lines, under westerly steering winds, occur in the monsoon regions of Asia, Australia, the western tropical Pacific, and, occasionally, in

tropical South America. For large MCSs or superclusters, observed by Doppler radar over the western tropical Pacific, updrafts consist of layers of potentially unstable air that nearly always ascend slantwise over an apparent downdraft cold pool. Inflow layers of 0.5–4.5 km in depth form convective cells as soon as the rising layer becomes saturated but retain their layered structure, phenomena now known as moist absolutely unstable layers.

Tropical nonsquall clusters occur more frequently than squall clusters and possess more varied precipitation structures although their precipitation life cycle is similar to that of squall lines and MCCs.

MCS Propagation

It has been suggested that MCSs propagate via mechanisms such as (1) new convection forming along the cold pool boundary; (2) discrete propagation, in which the MCS propagation vector is the cold pool vector (approximated by the mean cloud layer wind) plus a low-level jet (LLJ) vector; (3) waves generated by the MCS itself, such as gravity waves, force new cells in the near MCS environment; (4) advection by large-scale waves to which the MCS becomes phase-locked; and (5) formation of a midlevel mesoscale vortex and new convection with a different velocity than the initial MCS.

Large-Scale Environments

The Role of Topography

Global climatologies have shown that most MCSs form in the lee of elevated terrain, which act as heat sources in response to diurnal heating. Large-scale, mountain–plains solenoidal circulations help to organize MCSs during summertime in the midlatitudes and certain regions of tropical continents. In a typical diurnal cycle, convection initiates over elevated or sloping terrain then propagates coherently downstream in moderate vertical shear while growing and merging into mesoscale systems.

Midlatitude Squall Lines

Squall lines are observed frequently in the warm sector of a midlatitude cyclone, usually about 100–300 km in advance of the cold front. Divergence ahead of an upper-level trough induces low-level convergence. The advection of warm, moist air by an LLJ, sometimes accompanied by cold advection in the upper-level jet stream, provides a source of instability and promotes long-lived systems (Figure 6).

Over the US southern Plains, low–midtropospheric warm, dry air from the Mexican plateau meets the warm, humid air from the Gulf of Mexico along a moisture boundary known as the ‘dry line.’ Dry-air intrusion produces a capping inversion that leads to great instability (CAPE) and more intense squall lines. This scenario is common elsewhere, such as east of the subtropical Andes, the South African escarpment, and southeast Australia. Squall lines are also triggered along coastlines, due to differential surface heating, and wind shear lines.

Increasing mid- to upper-level shear marks midlatitude squall line environments. Large quantities of dry air are entrained into the rear of the system, cooled, moistened, and sink, thereby strengthening the downdraft. The resulting cold pool and gust front enhances the system updraft and leads to greater intensity and longevity. The cold pool acts as a gravity current because it is denser than its environment and new convection forms by uplift along its boundary. The organization of the mature system is strongly governed by the vertical wind shear and the strength of the cold pool. Other predictors of intensity and longevity are the strength of the midlevel winds, the low-level storm-relative flow, and shear from the surface to midtroposphere. Midlatitude squall lines can be maintained with moderate midlevel shear, which is generally destructive to tropical squall lines.

Mesoscale Convective Complexes

MCCs initiate within prominent baroclinic zones (like stationary fronts) characterized by locally large values of lower-tropospheric vertical wind shear and CAPE. Typically, an LLJ of air with low static stability, high equivalent potential

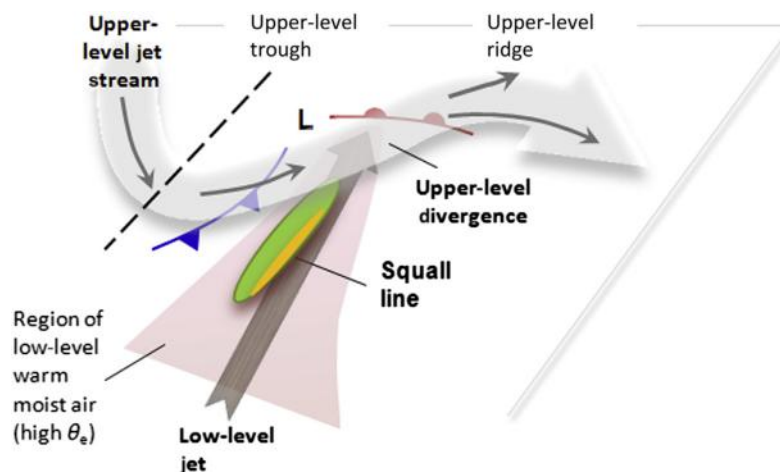


Figure 6 Schematic diagram of the large-scale features associated with midlatitude prefrontal squall lines.

temperature (θ_e), oriented nearly perpendicular to the baroclinic zone, intrudes into the genesis region, and is forced to ascend over a relatively shallow, surface-based layer of relatively cool air (Figure 7). Pronounced warm advection accompanied by strong lower-tropospheric veering overlays the surface-based cool layer. Local maxima in absolute humidity and static instability mark the preferred MCC formation region. Low-level convergence, upper-level divergence, and an approaching midlevel vorticity maximum related to a weak shortwave trough are also characteristics of the mean genesis

environment. In most instances, an upper-level jet is present in the near vicinity.

The thermodynamic structure and flow of the mean MCC environments suggest that low-level slantwise ascent along sloping potential temperature surfaces in regions of lower-tropospheric warm advection is an inherent part of most MCC environments. The low-level warm advection coupled with differential cyclonic vorticity advection forces upward motion throughout the troposphere, with maximum values near midtroposphere.

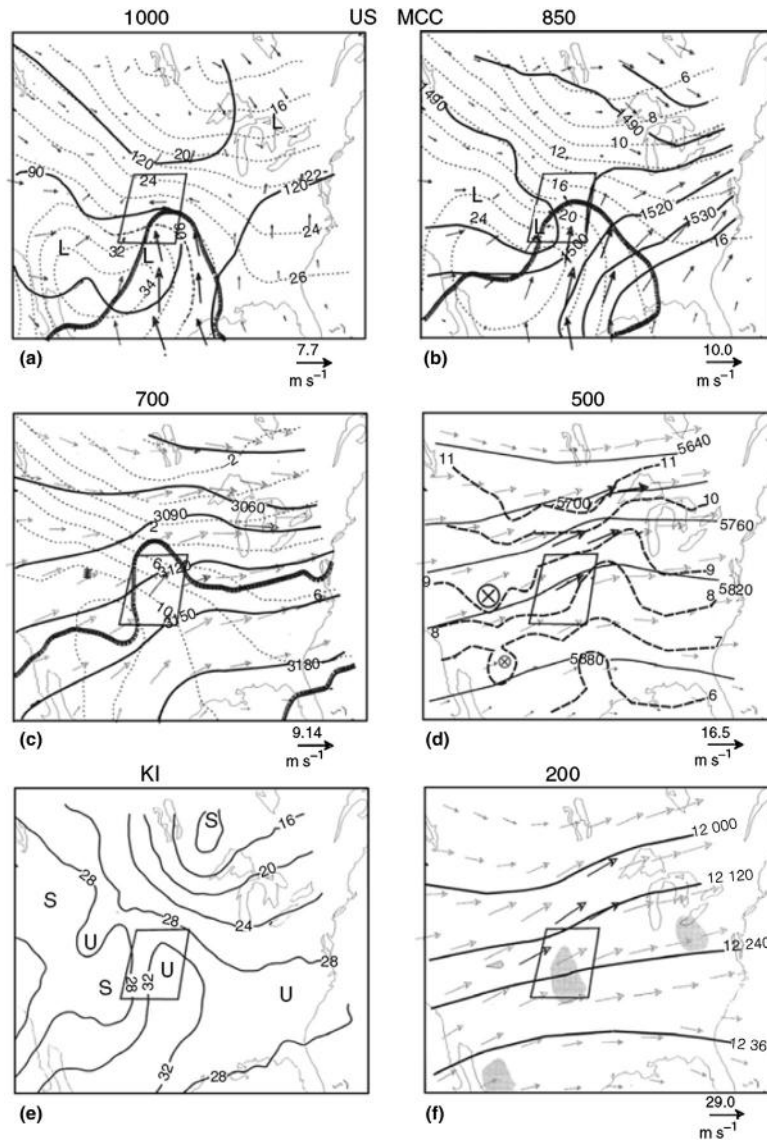


Figure 7 Analyses of the mean synoptic environment of 12 US MCCs: temperature, geopotential height, q_e , and wind vectors for (a) 1000 hPa and (b) 850 hPa; geopotential height, mixing ratio, and wind vectors for (c) 700 hPa; geopotential height, vorticity, and wind vectors for (d) 500 hPa; (e) K-index (KI) (Appendix); and geopotential height, divergence, and wind vectors for (f) 200 hPa. The quadrilateral in the center indicates the approximate genesis region (GR). Velocity vectors (m s^{-1}) are plotted at every other grid point. Heights (m) and KI values are solid contours, isotherms ($^{\circ}\text{C}$) are dashed, and hachured areas are (a) $q_e > 350 \text{ K}$, (b) $q_e > 338 \text{ K}$, and (c) mixing ratio $> 5 \text{ g kg}^{-1}$. The letters U and S identify the relatively unstable and stable areas, respectively, in (e). Bold velocity vectors, in (a) and (b), highlight the low-level inflow of high- q_e air and 850-hPa warm advection, respectively. Bold vectors also mark wind speed maxima in (d) and (f). The 500-hPa absolute vorticity ($\times 10^{-5} \text{ s}^{-1}$) is shown by dashed lines in (d). Shaded areas in (f) are divergence $> 4 \times 10^{-6} \text{ s}^{-1}$. Reproduced from Laing, A.G., Fritsch, J.M., 2000. The large-scale environments of the global population of mesoscale convective complexes. *Monthly Weather Review* 128, 2756–2776.

Although most MCCs form in conjunction with frontal lifting, a significant number (about 40% in the United States) develop on the warm side of the baroclinic zone (Figure 8). A favorable combination of a convectively generated cold pool and substantial low-level vertical wind shear can produce long-lived convective weather systems with a deep layer of slantwise ascent and a large stratiform cloud region *without large-scale forcing*. Figure 9 shows the conceptual models of MCC configurations. Figure 9(a) is similar to the squall line arrangement, except that, because of an LLJ, the ambient shear in the cold pool is opposite to the squall line case. Therefore, the cold-pool-induced circulation always has the same sign as the ambient shear. This device is also likely in tropical MCC populations that have weak baroclinic forcing. Nevertheless, it is unlikely that this mechanism would dictate the organization of most MCCs, for which convection originates from an elevated layer of high- θ_e air overrunning a quasi-stationary surface-based frontal zone (Figure 9(b)). It is difficult for moist downdrafts to penetrate to the surface and form a mesoscale cold pool. However, if a cold pool were to develop, because of the LLJ the ambient shear would have the same sign as any convectively derived cold pool. That configuration would also produce slantwise ascent and a large stratiform cloud layer.

Mature MCCs contain three distinct circulations: a large, cold, and shallow anticyclone in the vicinity of the tropopause, a boundary layer cold mesohigh, and, in the midlevels, strong latent heating, convergence, and a mesoscale updraft. MCCs moisten a deep tropospheric layer, and amplify the midlevel

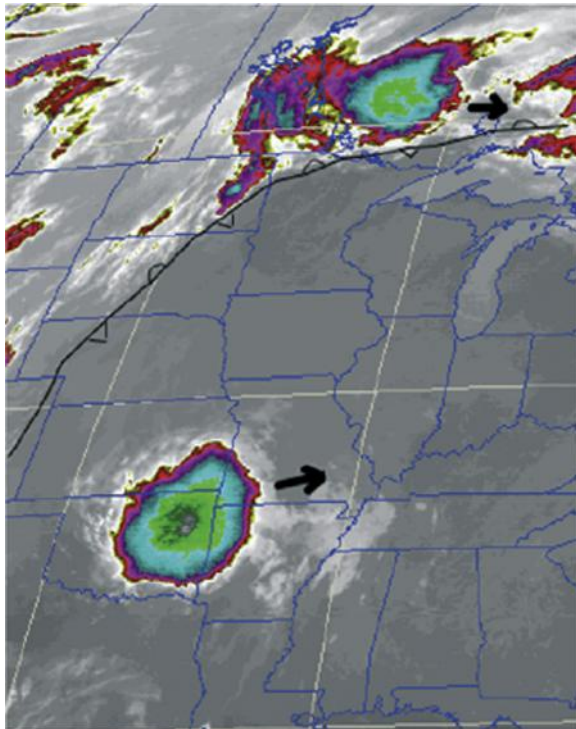


Figure 8 Examples of two types of North American MCCs: one in Canada on the cool side of the stationary front (dominant type) and another in the US Southern Plains on the warm side of the front at 1200 UTC, 6 May 2000. Arrows show the direction of movement of the MCCs.

shortwave trough as it adjusts to latent heating. The LLJ strengthens, veers, and helps to maintain high moisture and maximum conditional instability in the southwestern flank of the system (a region susceptible to floods). An anticyclonically curved upper-level jet streak develops to the north of the MCC. Near maximum mean upward motion, upper-tropospheric divergence, and vorticity occur throughout the mature stage and is maintained through the decay stage.

Once outside the region of conditionally unstable air and low-level warm advection, the system decays leaving a significantly modified environment extending beyond the MCC boundary, including an amplified midlevel shortwave trough and high upper-tropospheric heights and wind perturbations.

LLJ and Movement of MCCs

The presence of LLJs in MCC environments around the globe implies that the jet helps to organize and maintain mesoscale convection. Detailed investigations of MCCs in the United States, South America, and China showed that pronounced convergence occurs at the apex of the jet and that vertical motion from the jet dramatically increases the moist convective instability (e.g., Figure 6). Furthermore, *the temporal evolution (evening into early morning) and the orientation of the LLJ closely parallel the nocturnal life cycle and movement, respectively, of MCSs*.

MCC motion vector is approximated by summing the mean from 700- to 500-hPa flow parallel to thickness isopleths, a proxy for propagation of the meso- β -scale elements (MBEs), and the LLJ vector. Studies of US MCCs found that, as the system mature the MBEs move about 45° to the right of the mean flow, partly due to the MCC modifying the mean environmental wind. Extreme right movers account for about 25% of US MCCs and they produce derechos. MCCs become quasi-stationary or build backward if the moist, unstable inflow is toward the rear, then new cells form upstream (to the west in the midlatitudes).

Bow Echoes

Bow echoes and derechos most commonly occur in an environment with strong, deep layer shear in combination with high CAPE, steep midlevel lapse rates, and a strong cold pool. Derecho-producing environments can be distinguished from nonderecho environments by the strength of shear and the system-relative flow in the lowest 3 km. About half of derechos in the United States during 1980–2001 had shear magnitudes greater than 20 m s^{-1} . *Progressive* derechos are prevalent when the bow is oriented with a large angle to the mean wind direction and moves for hundreds of kilometer along a quasi-stationary front while *serial* derechos occur when multiple bow echoes are embedded in a long squall line and move along the direction of the mean steering wind.

Tropical Squall Line/Clusters

The dominant large-scale forcing of squall lines in the western Pacific and tropical Atlantic is convergence in the equatorial trough, monsoon trough, and intertropical convergence zone. Systems observed during the Global Atmospheric Research Program Atlantic Tropical Experiment (GATE) formed several

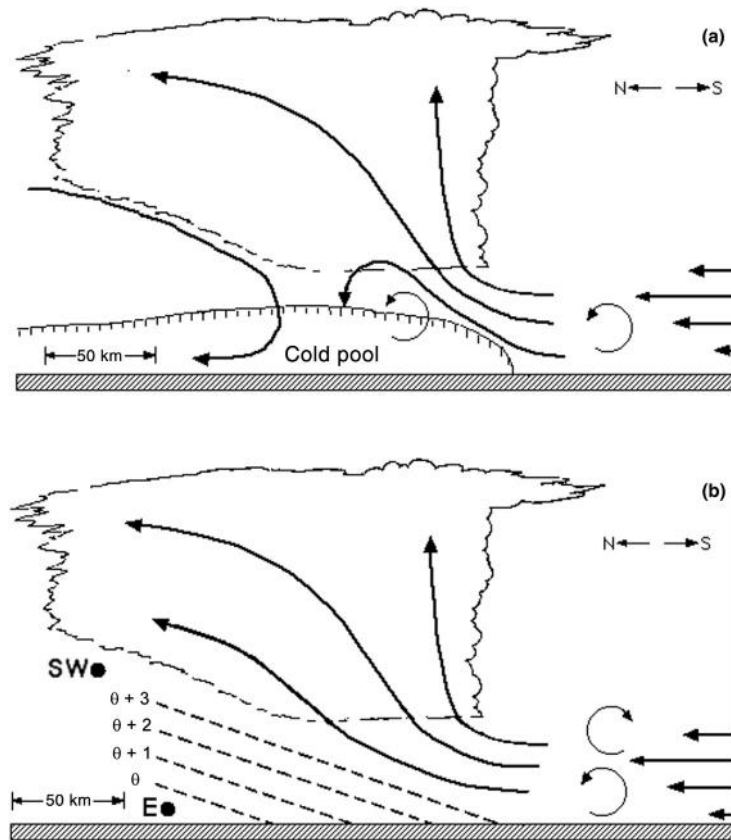


Figure 9 Storm evolution in contrasting environments: thin circling arrows indicate circulations affiliated with ambient wind shear and/or a moist-downdraft-generated cold pool. (a) Cold pool forms in the warm sector. The circulation induced by the cold pool combines with the circulation associated with the shear to tilt the system downstream and downshear (northward). (If the shear-induced circulation is of the opposite sign, the cold-pool circulation overwhelms the shear and tilts upshear, northward, over the cold pool.) (b) LLJ overruns a frontal zone. Negative buoyancy of moist downdrafts is insufficient to penetrate to the base of the synoptic-scale cold air layer, so a surface-based mesoscale cold pool does not develop. The letters W and E indicate southwesterly and easterly winds, respectively. The edge of the cloud system is scalloped. Reproduced from Laing, A.G., Fritsch, J.M., 2000. The large-scale environments of the global population of mesoscale convective complexes. *Monthly Weather Review* 128, 2756–2776.

hours after large-scale convergence was established and moisture flux increased in the low-middle troposphere. Postsquall environments are characterized by lower-tropospheric warming and drying.

While most tropical MCS generation is correlated with topography, vertical shear, high CAPE, and diurnal heating, some squall lines occur with African easterly waves. Squall lines associated with easterly waves most often form ahead of or within the easterly wave trough, move at about twice the speed of the wave, and tend to die just behind the ridge. However, in the northern Sahel, MCSs form within and behind the trough, in the region of maximum convergence of moisture from southerly monsoon winds. Amazon squall lines have added forcing from the sea breeze front.

Tropical squall line development is aided by strong low-level wind shear. For GATE squall lines, the critical threshold was 13 m s^{-1} in the 950- to 650-hPa layer, mostly perpendicular to the leading line. Squall lines, in turn, transport momentum *up-gradient*, increasing the vertical wind shear.

Tropical MCS environments include topographic effects and gravity waves. For example, mesoscale convection over northwest South America is initiated by elevated heating over the

Andes, which induces sea breeze inflow. Convection then moves westward over the tropical Pacific because of a gravity wave response to the elevated heating. During the day, a warm anomaly over the ocean caps convection while after sunset, cooling over the elevated terrain causes a nearby gravity wave response of upward motion in the lower troposphere. Thus, allowing convection to develop over the ocean at night. Similar mechanisms operate over the northern South China Sea except with extra forcing from an onshore LLJ. Another theory is suggested for nocturnal, offshore mesoscale convection downstream of tropical mountains. The mountain blocks the sea breeze current, thereby producing relatively cold, stagnant air at the base of the mountain, which then produces a stronger land-breeze density current and triggers convection far offshore.

Tropical Nonsquall Clusters

The environment of nonsquall clusters is similar to squall clusters in terms of conditional instability and formation relative to the easterly wave trough. However, large, longer-lived nonsquall clusters tend to move at less than the speed of the easterly wave and dissipate once they fall behind the

trough axis. GATE clusters are formed where mean 950- to 650-hPa environmental shear was about 6 m s^{-1} and mostly parallel to the lines. A distinctive feature of nonsquall cluster environments is the absence of strong shear between 700 and 250 hPa. Winter monsoon clusters west of Borneo are initiated by the low-level convergence of offshore breezes, northeasterly monsoon wind surges, and synoptic-scale Borneo vortices, which form when northeasterly monsoon winds curve to the east and north near the equator. MCS formation over the South China Sea and Borneo is suppressed on days without a vortex and enhanced further downstream.

Composite nonsquall cluster environments have maximum low-level convergence during the growing stage of the system life cycle. The maximum vertical motion is close to 300 hPa and occurs during the mature phase in contrast to midlatitude MCCs whose maximum vertical motion occurs during the mature-to-dissipating phase.

Basically, environments with weak shear in the middle levels and deep layer moisture tend to produce more

MCCs and nonsquall clusters while strongly sheared environments with dry-air inflow at middle levels favor more linear MCS.

Impact of Tropical Waves and the Madden-Julian Oscillation

Tropical mesoscale convection is sometimes coupled to equatorially trapped waves such as Rossby-gravity waves and Kelvin waves. The Madden-Julian Oscillation (MJO), the dominant mode of intraseasonal variability in the tropics, also modulates MCS activity. The propagation of tropical MCSs is affected by variations in the large-scale wind velocity in the lower troposphere and near the tropopause, as occurs with convectively coupled equatorial waves and the MJO. For example, during the wet phase of convectively coupled Kelvin waves, MCSs are larger and have higher cold cloud tops. Daily initiation and westward propagation of MCSs continue within the wave envelope while the dry phase has few, weak systems (Figure 10).

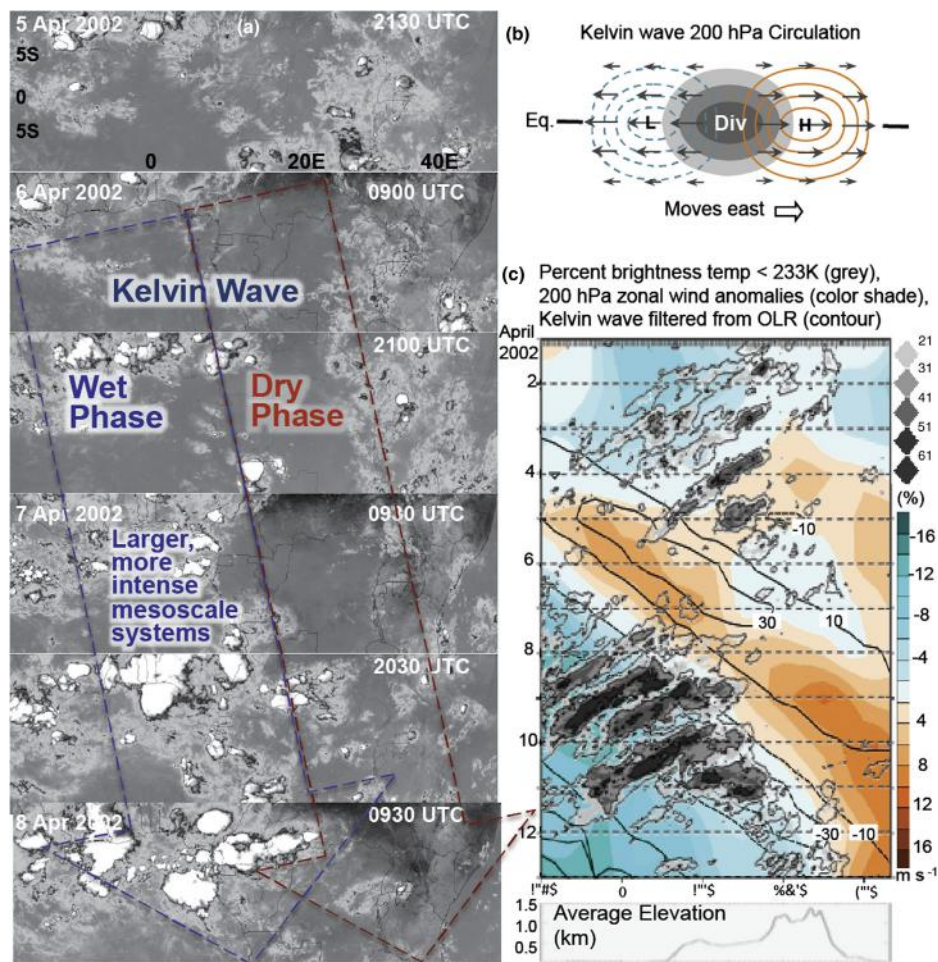


Figure 10 (a) Enhanced IR images for 5–8 April 2002. (b) Theoretical equatorially trapped Kelvin wave circulation at 200 hPa. (c) Hovmöller of brightness temperature $<233 \text{ K}$ frequency (solid contours 10%, gray shade 21%) overlaid on 200-hPa zonal wind anomaly (orange shades are easterly, turquoise shades are westerly) and Kelvin waves filtered from daily outgoing long-wave radiation (OLR) anomaly (negative is dashed black, positive is solid black at 20 W m^{-2} intervals). Average elevation (m) is below the Hovmöller diagram in (c). Averages are taken between 7.5° S and 7.5° N . Part (c) is reproduced from Laing, A.G., Carbone, R.E., Levizzani, V., 2011. Cycles and propagation of deep convection over equatorial Africa. *Monthly Weather Review* 139, 2832–2853. © The American Meteorological Society.

Global Distribution

Mesoscale Convective Complex

MCCs occur frequently (300–400 annually, Table 3) but in certain preferred locations (Figure 11). Most MCCs: (1) occur overland; (2)

develop in belts of easterlies and westerlies; and (3) tend to form in the lee of high terrain (association with LLJs). It is intriguing to note that MCCs are rare in some locations where deep moist convection including squall lines is common, e.g., the Amazon River basin or the Southeastern United States (Figure 10).

Table 3 MCC populations included in the global data set

MCC data sets			
Region/period of study	Geographic domain	Avg no. per season (high–low annual range)	Reference
South, Central America, May 1981–Apr 1983	40° S–30° N, 120–28° W	96 (47)	Velasco and Fritsch (1987)
Western Pacific Region, 1983, 1985	50° S–50° N, 90° E–170° W	82 (14)	Miller (1990)
United States, 1986–87	25–50° N, 130–70° W	51 (14)	Augustine and Howard (1991)
Africa, 1986–87	40° S–35° N, 45° W–45° E	97 (7)	Laing (1992)
Indian subcontinent, Apr–Dec 1988	0–50° N, 45–110° E	49	Laing (1992)
Europe, 1986–87	35–55° N, 20° W–45° E	6 (1)	Laing and Fritsch (1997)
Total		381 (83)	

Adapted from Laing, A.G., Fritsch, J.M., 1997. The global population of mesoscale convective complexes. Quarterly Journal of the Royal Meteorological Society 123, 389–405.

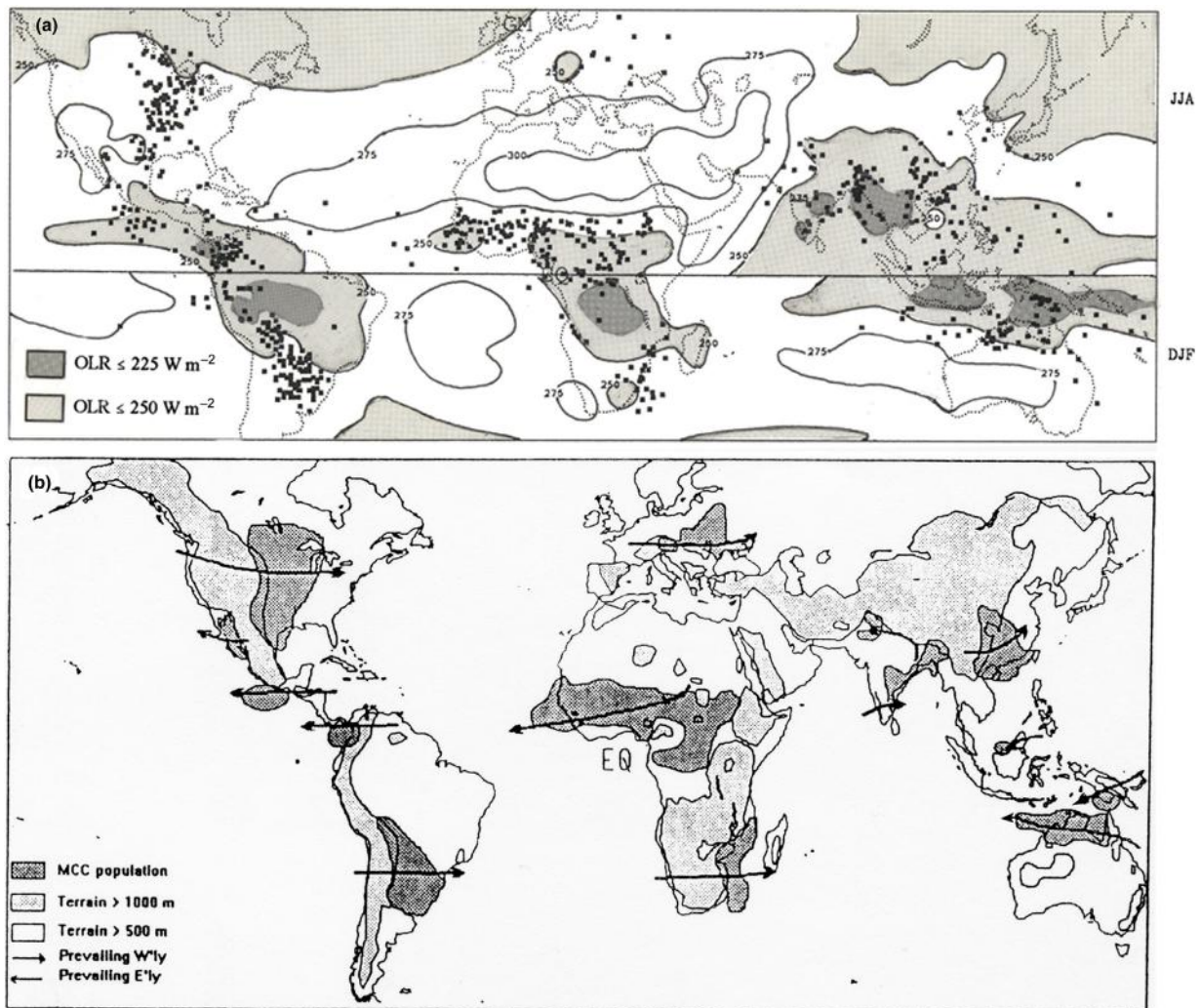


Figure 11 (a) Global distribution of MCCs (dots) and regions of widespread frequent deep convection as inferred by OLR minima (shading). (b) Relationship among MCC locations, elevated terrain, and prevailing midlevel flow. From Laing, A.G., Fritsch, J.M., 1997. The global population of mesoscale convective complexes. Quarterly Journal of the Royal Meteorological Society 123, 389–405. © The Royal Meteorological Society.

Table 4 SSM/I-derived ice scattering definition for MCS and intense MCS based on polarization corrected temperature (PCT)

MCS	Minimum area	2000 km ²
	Contour PCT	250 K
	Minimum enclosed PCT	≤225 K
Intense MCS	Minimum area	2000 km ²
	Contour PCT	200 K
	Minimum enclosed PCT	≤175 K

Adapted from Mohr, K.I., Zipser, E.J., 1996. Mesoscale convective systems defined by their 85-GHz ice scattering signature: size and intensity comparison over tropical oceans and continents. *Monthly Weather Review* 124, 2417–2437.

Midlatitude MCC activity is most frequent from June to August in the Northern Hemisphere and November to February in the Southern Hemisphere. The latitudinal progression of all MCS activities is related to the seasonal migration of large-scale circulation patterns and the migration varies substantially from region to region. During summer in northern continents, when the poleward shift in the jet stream is pronounced, MCC activity shifts strongly poleward. In the ocean-dominated hemisphere, where jet streams are quasi-stationary, poleward migration is not as evident.

Tropical MCS

Tropical MCS climatologies were developed using Special Sensor Microwave Imager (SSM/I) and the Tropical Rainfall Measurement Mission (TRMM) satellite measurements. Systems were classified according to their 85-GHz ice scattering signature to distinguish mesoscale regions of convectively produced large ice particles aloft (Table 4), the maximum heights of the 30-dBZ contour, and 6-km reflectivities (Table 5). MCSs are most prevalent over tropical South America, tropical Africa, and the oceanic warm pool (similar locations to Figure 11). They are 60% more frequent over continents at sunset than at sunrise and 35% more frequent over the oceans at sunrise than at sunset. Except over subtropical oceans, MCSs are larger at sunrise than at sunset. The total population forms a continuous, approximately log normal, distribution with the frequency inversely proportional to the area and intensity. No significant

relationship was found between size and intensity. Intense MCSs, which accounted for only 4% of the SSM/I-derived database, were mainly continental, colder than oceanic systems, and more occurred in the subtropics than in the tropics, a pattern confirmed by TRMM measurements. The reverse is true for rain area as oceanic intense MCSs contained a larger area of rain than continental intense MCS. The most intense systems, those with reflectivity of 20 dBZ reaching 14 km or higher, are most frequent over the African Congo, a significant fraction are also found in subtropical South American, northern India/Bangladesh, areas also favored for MCC development.

Heating and Moist Convective Transport

The propensity of convective cells to organize into mesoscale systems has important implications for the large-scale circulation and heat budget because convection has a greater impact on the large-scale when organized into mesoscale systems than as individual thunderstorms.

In examining the transport of heat and moisture, it helps to consider an ensemble of cumulus clouds embedded in a large-scale circulation. The heat and moisture budget can be expressed as:

$$Q_1 - Q_R = Q_{1c} + Q_{1m}$$

$$Q_2 = Q_{2c} + Q_{2m}$$

where Q_R is the net radiative heating, Q_1 and Q_2 are the residuals from the budgets of the 'resolvable' motion called the apparent heat source and moisture sink, respectively. When the budgets are divided into cumulus (Q_{1c} , Q_{2c}) and mesoscale (Q_{1m} , Q_{2m}) stratiform components, the stratiform exerts great influence on the overall vertical heating profile with cooling below the melting level and warming above (Figure 12(a) and 12(b)). Midlatitude MCSs generally have a sharper peak in heating and at a higher level than tropical MCSs (Figure 12(c)). Heating characteristics vary even among tropical MCSs. Maximum heating in Western Pacific MCSs occurs in the mid-to upper levels during the mature-to-decaying stages while GATE systems experience maximum heating earlier and in the lower troposphere.

Table 5 Criteria used to classify three precipitation feature (PF) types, based on the TRMM Microwave Imager (TMI) and Precipitation Radar (PR) data

Category name	Data	Criteria
PF with an MCS	PR near-surface reflectivity and nearest neighbor TMI 85-GHz PCT (within the PR swath)	● 108 or more contiguous data bins (area ≥2000 km²) with PCT ≤250 K with associated PR rain area ● 10 or more data bins (area 185 km²) with PCT ≤225 K
PF with ice scattering and without an MCS	PR near-surface reflectivity and nearest neighbor TMI 85-GHz PCT (within the PR swath)	● 4 or more contiguous data bins (area ≥75 km²) with a PR near-surface reflectivity ≥20 dBZ or a PCT ≤250 K ● At least one data bin with PCT ≤250 K ● Does not meet the PF with an MCS criteria
PF without ice scattering (250 K)	PR near-surface reflectivity and nearest neighbor TMI 85-GHz PCT (within the PR swath)	● 4 or more contiguous data bins (area ≤75 km²) with a PR near-surface reflectivity ≥20 dBZ ● No data bins contain PCTs ≤250 K

Adapted from Nesbitt, S.W., Zipser, E.J., Cecil, D.J., 2000. A census of precipitation features in the Tropics using TRMM: radar, ice scattering, and lightning observations. *Journal of Climate* 13, 4087–4106.

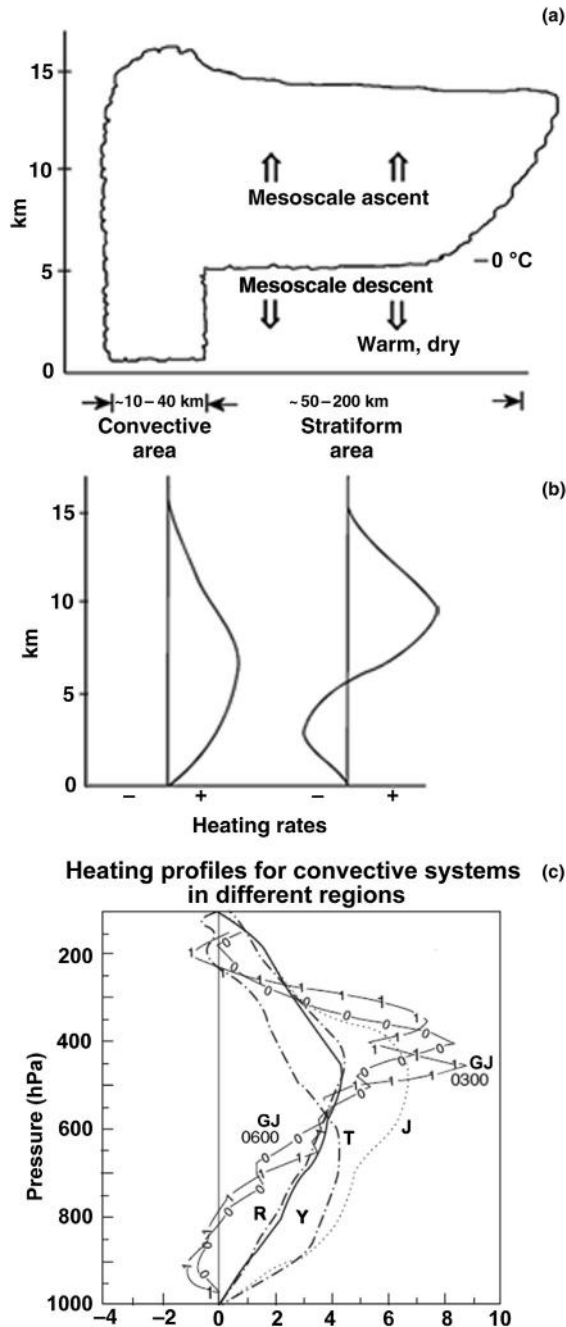


Figure 12 (a) Idealized mature stage of MCS illustrating convective and stratiform precipitation areas along with (b) associated heating profiles. Reproduced from Johnson, R.H., 1986. Implications of lower tropospheric warming and drying in tropical mesoscale convective systems for the problem of cumulus parameterization. *J. Meteorol. Soc. Japan* 64, 721–726. (c) Comparisons of Q_1 profiles normalized by rainfall rates for averages taken over an entire squall line for 11 June at 0300 (curve GJ0300) and 0600 (curve GJ0600) UTC. Other curves are from tropical or subtropical cases by Yanai, M., Esbensen, S., Chu, J.H., 1973. Determination of bulk properties of tropical cloud clusters from large-scale heat and moisture budgets. *J. Atmos. Sci.* 30, 611–627 (curve Y), Reed and Recker (1971, curve R), Thompson et al. (1979, curve T), and Johnson (1976, curve J). Reproduced with permission from Gallus, W.A., Johnson, R.H., 1991. Heat and moisture budgets of an intense midlatitude squall line. *Journal of Atmospheric Sciences* 48, 122–146. © The American Meteorological Society.

Dynamical Adjustment

The atmospheric response to latent heat depends on the horizontal scale of the heating relative to the Rossby radius of deformation, L_R , where

$$L_R = \frac{C_N}{(\zeta + f)^{\frac{1}{2}}(2VR^{-1} + f)^{\frac{1}{2}}}$$

ζ is the vertical component of relative vorticity, f is the Coriolis parameter, C_N is the phase speed of an inertial gravity wave, and V is the tangential component of the wind at the radius of curvature, R . If the horizontal scale of heating is less than L_R , then most of the energy released by heating will propagate away from the disturbance as gravity waves. If the scale of the latent heating is close to or exceeds L_R , then energy maintains nearly geostrophically balanced flow. For midlatitude systems, L_R is about 300 km, close to MCC length scale, explaining their longevity and ability to influence the large-scale circulation.

Diabatic heating in MCCs forces upper-level jet streaks downstream by enhancing ageostrophic flow in the jet entrance. Dynamical adjustment to the 1993 US MCSs created an upper-level wind maximum to the northeast of the heaviest precipitation, upper-level divergence over the region of heaviest rainfall, and an anomalously strong southerly low-level flow into the Upper Midwest. Thus, a continental-scale adjustment favored more convection and flooding.

Mesoscale Convective Vortices

Since the late 1970s, researchers have observed spiral bands of midlevel clouds that remain after the decay of MCSs, appearing like a tropical cyclone on land (e.g., Figure 13). The cyclonic circulation is centered on a warm-core, MCV that sometimes develops in the stratiform region of MCSs (Figure 14).

The development can be examined in terms of potential vorticity (PV) theory, which dictates that mass-integrated PV remains constant between two isentropic surfaces regardless of changes in mass transport or diabatic heating. Therefore, when diabatic heating occurs in moist convection, evacuation of mass across isentropes will lead to an increase in PV, i.e., the spin of the fluid adjusts to changes in the depth of the rotating column. *Heating in the positive PV anomaly can persist for days, regenerating convection and amplifying the vortex.* Under certain conditions, series of MCSs form in roughly the same location and follow similar tracks, thereby aggravating their flood potential, e.g., during the 1993 US Midwest Floods.

Tropical Cyclone Effects

It has been theorized that when successive cycles of convection occur over a warm water surface, they can play a crucial role in the transition of a loosely organized cluster of convection into a tropical depression, provided the large-scale environment is conducive. Tropical cyclones sometimes form when MCSs move over water, probably due to midlevel vortices extending downward.

Most tropical cyclone circulations are synoptic scale but their precipitation structure and inner core dynamics are predominantly mesoscale. Radar and satellite microwave

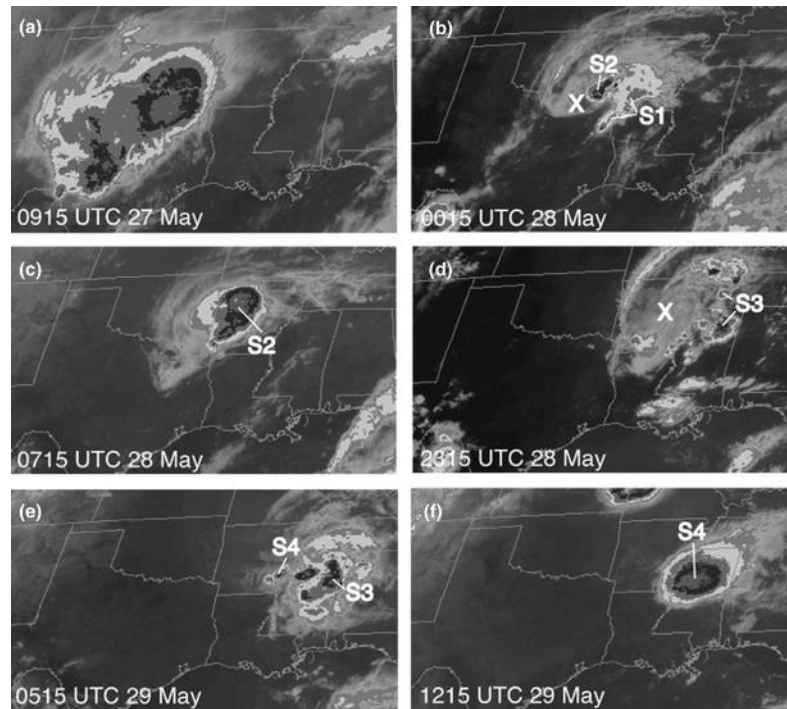


Figure 13 Infrared satellite imagery illustrating the deep convection associated with an MCV at (a) 0915 UTC 27 May, (b) 0015 UTC 28 May, (c) 0715 UTC 28 May, (d) 2315 UTC 28 May, (e) 0515 UTC 29 May, and (f) 1215 UTC 29 May 1998. 'X' marks the center of the MCV; other annotations mark secondary convection that formed around the MCV. Reproduced with permission from Trier, S.B., Davis, C.A. Tuttle, J.D., 2000. Long-lived mesoconvective vortices and their environment. Part I: Observations from the central United States during the 1998 warm season. *Mon. Wea. Rev.* 128, 3376–3395. © The American Meteorological Society.

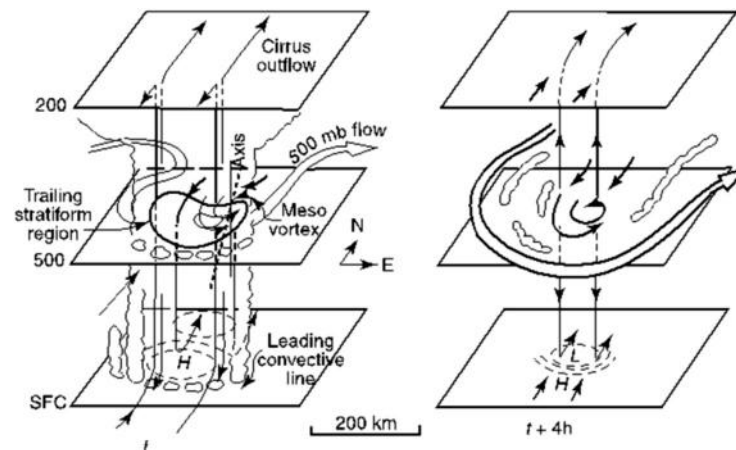


Figure 14 Idealized cloud and system-relative flow structure during decaying squall line stage (t) and fully developed mesovortex stage ($t + 4h$). Solid arrows denote storm-relative flow and open arrows denote large-scale 500-hPa flow. The dotted line indicates approximate vertical extent and tilt (toward the northeast) of the mesovortex core. Reproduced with permission from Johnson, R.H., Bartels, D.L., 1992. Circulations associated with a mature-to-decaying midlatitude mesoscale convective system. Part II: Upper-level features. *Monthly Weather Review* 120, 1301–1320. © The American Meteorological Society.

measurements suggest that mesoscale convective bursts can lead to rapid changes in cyclone intensity and structure through latent heat feedback mechanisms that enhance convergence and vorticity. Tropical Cyclone Oliver (1993) is a well-documented case of tropical cyclone genesis from MCSs.

Radiative Balance

One of the most important aspects of deep convection is its influence on the tropospheric radiation budget. The radiation balance depends on the distribution and characteristics of hydrometeors and other cloud particles. A tropical MCS can

cause net cooling from solar extinction, outweighing warming inside the system. Differential radiative heating can alter the evolution and mass circulation of individual MCSs through feedback among latent heat release, convective updrafts, and precipitation.

The time of day that MCSs occur can have varying radiative effects. Even with constant total cloud fraction, the sign of the radiation balance is sensitive to the diurnal distribution of deep convective cloud systems. For example, warming due to increased nocturnal deep convective cloud fraction exceeded the shortwave cooling over North America (July 1985 and 1986 comparison).

The diurnal radiative cycle modulates tropical oceanic convection, which attains peak intensity near sunrise. One hypothesis for the overnight growth holds that horizontal divergence in the radiation field near MCSs can lead to mass circulation that enhances cloud development. Another theory suggests that long-wave cooling at the cloud top and warming at cloud base lead to destabilization and increased convection.

Chemical Transport

Mesoscale convection is important for the transport and chemical cycling of trace gases. The convective updraft displaces the tropopause upward leading to strong vertical mixing. In one US MCC, an anvil formed *within the stratosphere* resulting in a layer of tropospheric air approximately one kilometer thick lying above an equally thick layer of stratospheric air. This displacement, along with a broken tropopause around the edges, injected stratospheric ozone into the troposphere and ozone-poor tropospheric air into the stratosphere. During the Stratosphere–Troposphere Analyses of Regional Transport (2008), convective injections were observed up to 5 km above the stratospheric intrusion base.

In the Amazon Boundary Layer Experiment, ozone concentrations in the mid- to upper levels of squall lines were 3–4 times above background values. MCCs often occur in series, disturbing atmospheric thermal and trace gas fields over a large area, making them likely important local, natural sources of tropospheric ozone in MCC populated regions.

Atmospheric Electrical Effects

Since the early 1990s interest has grown in the electrical properties of MCSs, particularly their association with upper atmospheric transient luminous events known as sprites, blue jets, and elves. These features are observed above the stratiform precipitation region of large MCSs that support horizontally extensive and layered regions of positive charge generation (in contrast to convective elements that are predominantly negatively charged). Over the US Great Plains, a necessary but not sufficient condition for sprites is an MCS with stratiform precipitation area over 20 000 km². Sprites are observed in other MCS-favored zones but indications are that sprites are less frequent on a per storm basis in tropical systems. These discoveries provide more incentives for continued investigation of MCSs.

Appendix

Glossary of Thermodynamic Parameters

Convective Available Potential Energy (J Kg⁻¹)

CAPE is the vertically integrated positive buoyancy of an adiabatically rising parcel, an excellent measure of latent instability. Increasing values of CAPE generally lead to progressively, more vigorous convection. CAPE depends on the parcel chosen for lifting and on whether the initial lift is mechanically forced or from surface heating. CAPE is represented on a thermodynamic diagram (Skew-*T*) as the positive area between the moist adiabatic curve and the sounding temperature curve, from the level of free convection to the equilibrium level.

$$\text{CAPE} = g \int_{z_{\text{LFC}}}^{z_{\text{EL}}} \left(\frac{T_p - T_e}{T_e} \right) dz$$

K-Index (KI, °C)

The K-Index has been shown to be particularly useful for identifying convective and heavy rain-producing environments because it accounts for the differences in moisture not simply temperature. George's (1960) results indicate that thunderstorm probability ranges from near zero when $K < 20$ to widespread activity when $K > 35$.

$$\text{KI} = (T_{850} - T_{500}) + Td_{850} - Tdd_{700},$$

where Td is the dewpoint temperature and Tdd is the dewpoint depression.

See also: Dynamical Meteorology: Potential Vorticity. **Electricity in the Atmosphere:** Sprites. **Mesoscale Meteorology:** Bow Echoes and Derecho; Convective Storms: Overview; Gust Fronts. **Radar:** Polarimetric Doppler Weather Radar; Precipitation Radar. **Radiation Transfer in the Atmosphere:** Cloud-Radiative Processes. **Satellites and Satellite Remote Sensing:** Precipitation; Remote Sensing: Cloud Properties. **Stratosphere/Troposphere Exchange and Structure:** Local Processes. **Synoptic Meteorology:** Forecasting; Weather Maps. **Thermodynamics:** Moist (Unsaturated) Air. **Tropical Cyclones and Hurricanes:** Overview and Theory.

Further Reading

- Carbone, R.A., Tuttle, J.D., Ahijevych, D.A., Trier, S.B., 2002. Inferences of predictability associated with warm season precipitation episodes. *Journal of Atmospheric Sciences* 59, 2033–2056.
- Cohen, J., Silva, D.M., Nobre, C., 1995. Environmental conditions associated with Amazonian squall lines: a case study. *Monthly Weather Review* 123, 3163–3174.
- Cotton, W.R., Bryan, G., van den Heever, S., 2010. *Storm and Cloud Dynamics*, second ed. Academic Press, San Diego. 820 pp.
- Durkee, J.D., Mote, T.L., Shepherd, J.M., 2009. The contribution of mesoscale convective complexes to rainfall across subtropical South America. *Journal of Climate* 22, 4590–4605.
- Fritsch, J.M., Kane, R.J., Chelius, C.H., 1986. The contribution of mesoscale convective weather systems to the warm season precipitation in the United States. *Journal of Climatology and Applied Meteorology* 25, 1333–1345.
- Fritsch, J.M., Forbes, G., 2001. Mesoscale convective systems. *Meteorological Monographs* 28, 323–358.

- Fujita, T.T., 1978. Manual of downburst identification for project NIMROD. Satellite and mesometeorology Research Paper 156, Department of Geophysical Sciences. University of Chicago, 104 pp.
- Gallus, W.A., Johnson, R.H., 1991. Heat and moisture budgets of an intense midlatitude squall line. *Journal of Atmospheric Sciences* 48, 122–146.
- Grim, J.A., Rauber, R.M., McFarquhar, G.M., Jewett, B.F., Jorgensen, D.P., 2009. Development and Forcing of the Rear Inflow Jet in a Rapidly Developing and Decaying Squall Line during BAMEX. *Monthly Weather Review* 137, 1206–1229.
- Houze, R.A., Biggerstaff, M., Rutledge, S., Smull, B., 1989. Interpretation of Doppler weather radar displays of midlatitude mesoscale convective systems. *Bulletin of the American Meteorological Society* 70, 608–619.
- Houze, R.A., Smull, B.F., Dodge, P., 1990. Mesoscale organization of springtime rainstorms in Oklahoma. *Monthly Weather Review* 118, 613–654.
- Houze Jr., R.A., 2004. Mesoscale convective systems. *Reviews in Geophysics* 42, RG4003. <http://dx.doi.org/10.1029/2004RG000150>.
- Johnson, R.H., 1986. Implications of lower tropospheric warming and drying in tropical mesoscale convective systems for the problem of cumulus parameterization. *Journal of the Meteorological Society Japan* 64, 721–726.
- Johnson, R.H., Bartels, D.L., 1992. Circulations associated with a mature-to-decaying midlatitude mesoscale convective system. Part II: Upper-level features. *Monthly Weather Review* 120, 1301–1320.
- Laing, A.G., Fritsch, J.M., 1997. The global population of mesoscale convective complexes. *Quarterly Journal of the Royal Meteorological Society* 123, 389–405.
- Laing, A.G., Fritsch, J.M., 2000. The large-scale environment of the global population of mesoscale convective complexes. *Monthly Weather Review* 128, 2756–2776.
- Laing, A.G., Carbone, R.E., Levizzani, V., 2011. Cycles and propagation of deep convection over equatorial Africa. *Monthly Weather Review* 139, 2832–2853.
- Liu, C., Zipser, E.J., Cecil, D.J., Nesbitt, S.W., Sherwood, S., 2008. A cloud and precipitation feature database from nine years of TRMM observations. *Journal of Applied Meteorology and Climatology* 47, 2712–2728.
- Maddox, R.A., 1980. Mesoscale convective complexes. *Bulletin of the American Meteorological Society* 61, 1374–1387.
- Mohr, K.I., Zipser, E.J., 1996. Mesoscale convective systems defined by their 85-GHz ice scattering signature: size and intensity comparison over tropical oceans and continents. *Monthly Weather Review* 124, 2417–2437.
- Nesbitt, S.W., Zipser, E.J., Cecil, D.J., 2000. A census of precipitation features in the Tropics using TRMM: Radar, ice scattering, and lightning observations. *Journal of Climate* 13, 4087–4106.
- Parker, M.D., Johnson, R.H., 2000. Organizational modes of midlatitude mesoscale convective systems. *Monthly Weather Review* 128, 3413–3436.
- Ray, P.S. (Ed.), 1986. *Mesoscale Meteorology and Forecasting*. American Meteorological Society, Boston.
- Riehl, H., 1979. *Climate and Weather in the Tropics*. Academic Press, San Diego.
- Smith, A.M., McFarquhar, G.M., Rauber, R.M., Grim, J.A., Timlin, M.S., Jewett, B.F., Jorgensen, D.P., 2009. Microphysical and thermodynamic structure and evolution of the trailing stratiform regions of mesoscale convective systems during BAMEX. Part I: Observations. *Monthly Weather Review* 137, 1165–1185.
- Trier, S.B., Davis, C.A., Tuttle, J.D., 2000. Long-lived mesoconvective vortices and their environment. Part I: Observations from the central United States during the 1998 warm season. *Monthly Weather Review* 128, 3376–3395.
- Yanai, M., Esbensen, S., Chu, J.H., 1973. Determination of bulk properties of tropical cloud clusters from large-scale heat and moisture budgets. *Journal of Atmospheric Sciences* 30, 611–627.
- Zipser, E.J., 1977. Mesoscale and convective-scale downdraughts as distinct components of squall-line circulation. *Monthly Weather Review* 105, 1568–1589.

Synopsis

The fundamentals of downdrafts with an emphasis on those events, referred to as microbursts, that produce strong outflow winds near the surface are discussed.

Introduction

The atmosphere undergoes a dramatic overturning when a convective storm forms. Warm, moist air in the boundary layer is transported aloft, while relatively cool and dry air is brought down to the lowest levels. The primary mechanism that transports the latter air mass to the boundary layer is the downdraft. The thermodynamic conditions of the downdraft are familiar to the general public once it reaches the surface by the respite from hot, humid conditions that are prevalent in the ambient air that is convectively unstable. The leading edge of this outflow is referred to as the gust front.

Many convective downdrafts do not produce strong winds at the surface; however, in some instances the downward velocities and subsequent outflow winds can result in severe damage. Indeed, these strong winds can result in considerable crop, tree, and structural damage and have been identified as a causal factor in a number of aircraft accidents. These intense wind events are called downbursts and are defined as an area of strong winds produced by a downdraft over an area from <1 to 10 km in horizontal dimensions. Downbursts can be further subdivided into macrobursts and microbursts with the following definitions.

- *Microburst.* Small downburst, less than 4 km in outflow diameter at the ground, with peak winds lasting only 2–5 min. They may induce dangerous tailwind and downflow wind shears which can reduce aircraft performance.
- *Macroburst.* Large downburst, with 4 km or larger outflow diameter at the ground; damaging wind lasts 5–20 min. An intense macroburst causes tornado-force damage up to EF3 intensity (on the Enhanced Fujita scale of damage intensity). (The EF-scale rates the damage intensity of tornadoes and ranges from EF1 to EF5. A rating of EF5 has been associated with the most intense tornadoes.)

It is the former phenomenon that has garnered the most interest, owing to its small temporal and spatial scales. Indeed, during the period 1974–85, microburst winds were a factor in at least 11 civil transport accidents and incidents in the United States, involving over 400 fatalities and 145 injuries. Extensive research has been focused on documenting the reasons why this small subset of downdrafts produces damaging winds.

Fundamentals of the Downdraft

In order to understand the microburst phenomena, it is instructive to describe the fundamental characteristics of the downdraft

and how it differs from the updraft. A naive view of the downdraft is that it is an upside-down version of the updraft; however, this is far from an accurate description. While the updraft is typically slightly supersaturated, the downdraft is often appreciably subsaturated, owing to the inability for condensate cooling by evaporation, melting, or sublimation to offset completely the warming from adiabatic compression. For a number of years this was, perhaps, one of the greatest misconceptions of the thermodynamic characteristics within the downdraft. It was often assumed that downdraft air parcels descended moist adiabatically (i.e., assumed saturated descent). The wet-bulb temperature was also believed to be representative of the temperature within a descending air parcel. It is now known that the saturated wet adiabat can only be approached by the downdraft if it is weak (this allows sufficient time for latent cooling to offset adiabatic compression during descent), the mean drop size is small (see discussion below), or the rainfall is heavy.

It is more important to document the microphysical details of the liquid or solid condensate within the downdraft than within the updraft. For example, for a prescribed water content, small raindrops are more conducive to producing stronger downdrafts than large drops for two reasons: (1) there is greater surface area exposed to the environment; and (2) smaller drops have greater curvature which results in a larger equilibrium vapor pressure and, hence, lower relative humidity. Both effects increase the potential for evaporation, resulting in latent cooling. Finally, individual parcel vertical excursions tend to be less than 4 km for downdrafts and often greater than 10 km (i.e., the depth of the troposphere) for updrafts. The primary reason for this difference in length scale is that the positive buoyancy in the updraft is much greater than the negative buoyancy in the downdraft.

The forcing mechanisms of the microburst can be understood by considering the inviscid vertical momentum equation

$$\frac{d\bar{w}}{dt} = -\frac{1}{\rho} \frac{\partial \bar{p}}{\partial z} + g \left(\frac{\theta'_v}{\theta_{v0}} - \frac{c_v}{c_p} \frac{p'}{p_0} - (r_c + r_r + r_i) \right) \quad [1]$$

(1) (2) (3) (4)

where $\bar{w}$ is the mean vertical velocity, p the pressure, θ'_v the virtual potential temperature, c_p the specific heat at constant pressure, c_v the specific heat at constant volume, r_c the mixing ratio of cloud water, r_r the mixing ratio of rainwater, and r_i the mixing ratio of ice water.

The primes denote departures from a basic state (subscript 0), which varies only in height.

Term 1 is the vertical gradient of perturbation pressure. Term 2 represents thermal buoyancy accounted for in parcel theory and term 3 is the perturbation pressure buoyancy. Condensate

loading of cloud, rain, and ice water is represented by term 4. In addition, entrainment (or mixing) of environmental air with cloudy air or precipitation has been shown to be an important factor in downdraft dynamics.

The vertical gradient of perturbation pressure is generally small for most downdrafts; however, its effect becomes significant in intense cumulonimbi and mesoscale convective systems associated with strong vertical wind shear. Strong downdrafts can develop owing to rapid pressure falls generated in response to an intensifying low-level mesocyclone within a supercell thunderstorm.

The effect of thermal buoyancy is the best-known term in the vertical momentum equation. Relatively cold (hot) air promotes negative (positive) vertical velocities. The importance of using virtual potential temperature in the vertical momentum equation has been shown. Increasing the environmental relative humidity at low levels increases the virtual temperature difference and will lead to stronger downdrafts even if the ambient air temperature remains the same. The pressure buoyancy term suggests that a downdraft will be produced if an air parcel has a higher perturbation pressure than its surroundings. This effect is ignored in classical parcel theory, which assumes that the environmental and parcel pressures are equal. In general, this effect is relatively small in comparison to thermal buoyancy and the pressure gradient effects.

Some of the early hypotheses on downdrafts suggested that the air was initially dragged downward by the weight of precipitation particles (term 4 in eqn [1]), and then cooled by evaporation. This has been reinforced by numerous studies that have examined the forcing mechanisms of downdrafts. While precipitation drag can be important in initiating a downdraft, the critical role of latent cooling produced via evaporation can be shown by considering a case where the water mixing ratio is evaporated completely. A water content of 1 g kg^{-1} is approximately equivalent to a temperature deficit of 0.30°C in the buoyancy term in eqn [1]. If this water mixing ratio evaporates then the resulting temperature deficit is given by

$$\theta'_v = L \frac{r_r}{c_p} \approx 2.5 \text{ K} \quad [2]$$

where L is the latent heat of evaporation.

Thus, the temperature deficit has increased by a factor of 8.3 or nearly an order of magnitude by evaporating the water. Accordingly, the evaporation of raindrops rather than water loading would be more effective in accelerating a downdraft.

The effect of entrainment is complex. At higher levels, the entrainment of dry air promotes downdrafts by evaporation or sublimation of cloudy air and precipitation, especially when the entrained region corresponds to the level of minimum equivalent or wet-bulb potential temperature. The effect of entrainment is different at lower levels once a downdraft forms. If the downdraft is driven primarily by negative thermal buoyancy then its intensity will be determined by the virtual temperature difference between the descending air parcel and the environment as shown in eqn [1]. Mixing of environmental air at this stage will deplete the negative buoyancy by decreasing this difference and will lead to reduced downdraft speeds.

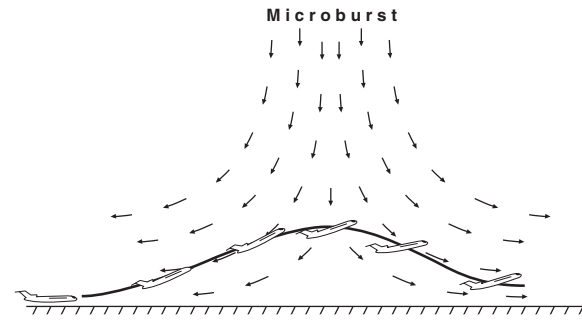


Figure 1 Schematic diagram illustrating the impact of a microburst on aircraft performance during takeoff. The airplane first encounters a headwind and first experiences added lift. This is followed in short succession by a decreasing headwind component, a downdraft, and finally a strong tailwind which may lead to an impact with the ground. Composite drawing based on numerous studies of aircraft accidents. Reproduced from Fujita, T.T., 1985. *The Downburst – Microburst and Macroburst*. SMRP Research Paper No. 210, NTIS PB-148880, report of Projects NIMROD and JAWS. University of Chicago Printing Department, Chicago, IL.

The Microburst

The discovery of the microburst can be traced back to a series of aircraft accidents and the existence of small-scale divergent damage patterns in crops and forests discovered after the passage of a thunderstorm. Both events were shown to be a result of a transient but intense downdraft accompanied by strong outflow. The typical scenario for a microburst-related aircraft accident is shown in Figure 1. When an aircraft flies through a microburst during takeoff it first encounters a headwind component from the microburst outflow. This headwind increases lift by increasing the relative airflow over the wing. The plane may then pitch up, and the pilot may attempt to compensate by leveling off. But only a matter of seconds later the plane encounters a decreasing headwind, downdraft (within the center of the microburst), and then a strong tailwind. The plane has lost lift and could find itself flying too low and with insufficient air speed to avoid a crash. A similar scenario could be created when an aircraft is on approach for a landing at an airport.

The conceptual model of the descending microburst is shown in Figure 2. The microburst is characterized by a shaft of strong downward velocity. It is also associated with strong divergent flow at its center when it reaches the ground, and is

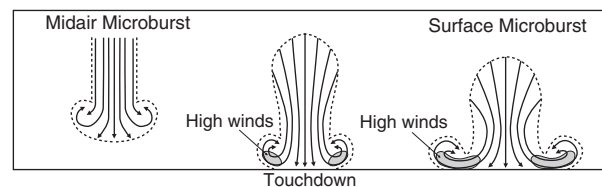


Figure 2 Conceptual model of a microburst hypothesized to explain ground-damage patterns. Three stages of development are shown. The regions of high wind within the rotor are indicated. Reproduced from Fujita, T.T., 1985. *The Downburst – Microburst and Macroburst*. SMRP Research Paper No. 210, NTIS PB-148880, report of Projects NIMROD and JAWS. University of Chicago Printing Department, Chicago, IL.

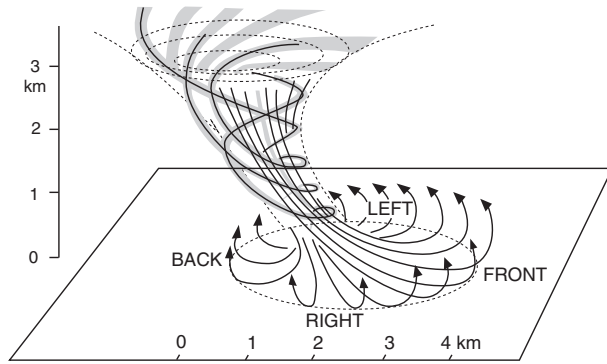


Figure 3 Three-dimensional visualization of a microburst. Reproduced from Fujita, T.T., 1985. *The Downburst – Microburst and Macrobust*. SMRP Research Paper No. 210, NTIS PB-148880, report of Projects NIMROD and JAWS. University of Chicago Printing Department, Chicago, IL.

followed by an accelerating outburst of strong winds in an overturning rotor propagating away from the center of the microburst. The highest wind speeds are usually associated with these rotors with the peak speeds occurring in the lower portion of the ring vortex, where outflow speeds are augmented by the circulation of the ring. The mechanism for intensifying the outflow winds is the stretching of the vortex as the ring expands. Horizontal pressure gradients can also contribute to high-outflow wind speeds, as discussed below. The vertical gradient of perturbation pressure (refer to eqn [1]) appears to play a relatively minor role in most microbursts. The exception appears to be the microburst associated with the supercell storm. Microburst damage has been frequently documented very near tornado damage tracks. Minimum pressures within the low-level mesocyclones are believed to enhance the downward velocities in these downdrafts.

The microburst in three dimensions is shown schematically in Figure 3. A key feature in the figure is that a number of microbursts are associated with small-scale circulations aloft. In fact, it is common for many microbursts to rotate, with the strongest ones associated with stronger rotation. The magnitude of the vertical vorticity within these circulations can be comparable to the vorticity present in the parent mesocyclone associated with a tornado.

Results from numerous observational and numerical studies suggest that microburst winds are associated with a continuum of rain rates that range from heavy precipitation from thunderstorms to virga shafts from either altocumuli, or clouds that have been referred to as shallow, high-based cumulonimbi. New definitions were created to account for the following two extreme microphysical situations.

- *Dry/low-reflectivity microburst.* A microburst associated with <0.25 mm of rain or a radar echo <35 dBZ in intensity.
- *Wet/high-reflectivity microburst.* A microburst associated with 0.25 mm of rain or a radar echo >35 dBZ in intensity.

Low-Reflectivity or Dry Microburst

Numerical simulations have shown the sensitivity of downdraft intensity as a function of drop size, rain intensity, and subcloud lapse rate. One of their conclusions is that when the

environmental lapse rate is approximately equal to the dry adiabatic lapse rate then the rates of evaporation place little restriction on the downdraft magnitude, and even in light precipitation there may be strong downdrafts generated. In the absence of pressure effects and in light rain situations (i.e., dry microbursts), it is often convenient to view the maintenance of the downdraft as the competing forces of cooling due to evaporation and sublimation versus dry adiabatic warming due to compression. When the subcloud lapse rate is dry adiabatic, any cooling by condensate results in a negative temperature perturbation that will maintain the downdraft. This effect increases with the depth of dry adiabatic lapse rate. Compressional warming in a descending parcel can counteract this cooling when the subcloud lapse rate is less than dry adiabatic. Condensate loading plays a relatively minor role in producing dry microbursts because only light precipitation is present.

These results have been confirmed by observations of virga shafts from weakly precipitating cloud systems producing low-reflectivity microbursts. These microbursts are particularly hazardous to aircraft because the parent cloud and pendant virga shafts appear innocuous. The weak echoes and low precipitation rates often result in little or no temperature change at the surface. In some cases, the air temperature has been shown to rise although the virtual temperature typically falls. Figure 4 is a plot of microburst occurrence as a function of radar reflectivity and subcloud lapse rate. Each point on the figure represents a microburst event identified by Doppler radar. Microbursts can be seen to occur most frequently at lapse rates of temperature exceeding $\sim 8.5 \text{ K km}^{-1}$. Practically no microbursts occurred for lapse rates less than $\sim 8.0 \text{ K km}^{-1}$. The axial profiles for cooling owing to phase change of condensate for the dry microburst is shown in Figure 5. The dry microburst is associated with small cooling, but it occurs through a deep column. Thus the integrated negative buoyancy is large. The prevalence of dry microburst over the high plains of the United States is attributed largely to the deep, dry adiabatic layer characteristic of this geographic region.

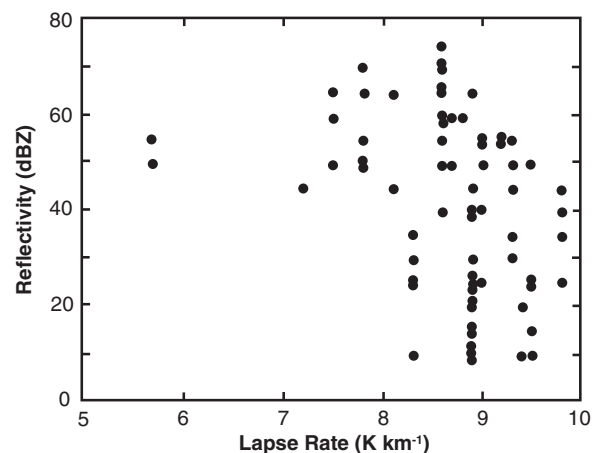


Figure 4 Plot of microburst occurrence as a function of radar reflectivity and environmental lapse rate. Reproduced from the American Meteorological Society from Srivastava, R.C., 1985. A simple model of evaporatively driven downdraft: Application to microburst downdraft. *Journal of the Atmospheric Sciences* 42, 1004–1023.

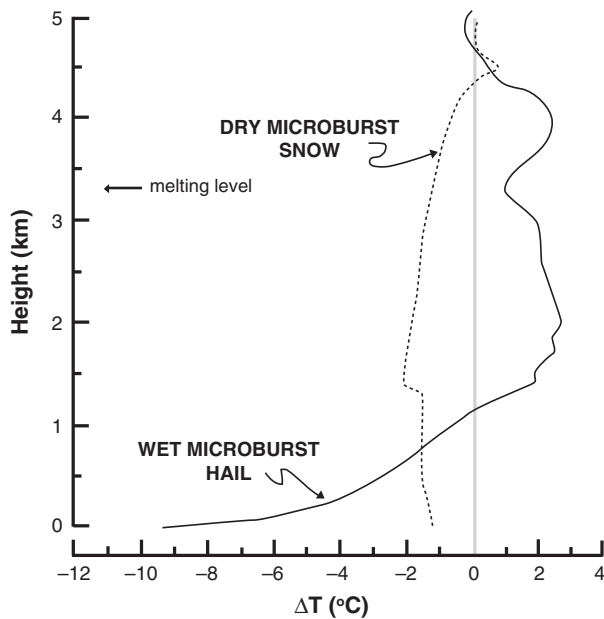


Figure 5 Comparison of the axial profiles for cooling owing to phase changes of condensate for the dry microburst with snow and the wet microburst with hail. Location of the melting level based on the environmental sounding is shown. Reproduced from the American Meteorological Society from Proctor, F.H., 1989. Numerical simulations of an isolated microburst. Part II: Sensitivity experiments. *Journal of the Atmospheric Sciences* 46, 2143–2165.

The sensitivity of the dry microburst to the microphysics of the condensate has also been shown. Numerical studies initiated with different precipitation particles reveal that snowflakes produce downdraft speeds nearly twice as strong as those generated by hail. The sublimation process associated with snowflakes is important for three apparent reasons: (1) the numerous low-density snow particles readily sublimate, with much of the snow content depleted before melting into rain; (2) the latent heat of sublimation is greater than the latent heat of either evaporation or melting; and (3) the cooling from sublimation takes place at a relatively high altitude within the deep adiabatic layer, allowing the downdraft to accelerate through a deep column.

High-Reflectivity or Wet Microburst

In more stable lapse rates, the downdraft tends to be warmer than the environment. In such a case, the drag of the condensed water may become important in accelerating the downdraft, especially at high levels. Accordingly, higher rainwater mixing ratios (i.e., radar reflectivities) are required for microbursts to form. This effect is well illustrated in Figure 4. The few microbursts that occurred for lapse rates less than $\sim 8.0 \text{ K km}^{-1}$ were associated with radar reflectivities in excess of 45 dBZ, suggesting that higher water contents are needed to produce strong downdrafts. This often manifests itself in radar reflectivity images as a prominent descending precipitation core. It has been shown that melting also plays an important role in producing a wet microburst. Narrow shafts of hail, coincident

with the microburst downdraft, embedded within a heavy shower of large raindrops have been noted.

The comparison of wet and dry microburst simulations is shown in Figure 5. The dry and wet microbursts were driven by snow and hail, respectively. Note that the temperature deviations for the two cases are different. The dry microburst is associated with smaller cooling, but it occurs through a deep column. The temperature deviation for the wet microburst is largest at the ground but diminishes rapidly with height until it becomes warmer than the ambient air above 1 km. The warming aloft for the hail case indicates the importance of precipitation loading in the early stages of the wet microburst which overcomes the positive buoyancy. It is hypothesized that hail produces stronger microbursts in more stable environments, since the downdraft is maintained only at lower elevations where it is less likely to be depleted of negative buoyancy because of compressional heating. Note that strong negative buoyancy is confined to the lowest levels of the wet microburst in Figure 5. The dominance of negative thermal buoyancy near the ground combined with the decreasing effect of precipitation loading can result in a displacement between the location of the strongest downdraft velocities and the maximum in radar reflectivity.

An important observation was revealed in the simulations shown in Figure 5. Although the downdraft associated with the dry microburst was much deeper and almost twice as intense as that of the wet microburst, both produced identical outflow speeds. This points to the immense difficulty in estimating peak outflow speeds based on the expected downdraft velocities. The cold air for the wet microburst is situated primarily at low levels in Figure 5. Although this cooling near the ground is not able to translate into strong enhancement of the downdraft, it may strengthen the outflow speeds through enhanced horizontal pressure gradient forces produced by the formation of a mesohigh. Dry microbursts are typically accompanied by small, cold pools (i.e., weak mesohighs). Thus, strong outflow winds for the dry microburst are possible only when the downdraft speeds are also intense. This comparison of these two cases emphasizes the nonlinear relationship between vertical velocities and outflow speeds.

Forecasting and Detection

Although intense microbursts are often associated with the supercell storm, most microbursts develop within environments characterized by weak vertical wind shear. The thermodynamic profile determined by nearby upper-air soundings is particularly useful for identifying when strong, convectively induced winds are likely to form. The two types of profiles that have been associated with strong outflow are the 'inverted V' profile conducive to dry or low-reflectivity microbursts and the weakly capped wet or high-reflectivity microburst. The former is shown in Figure 6. Its main characteristics are the deep, dry adiabatic, subcloud layer from near the surface to the mid-levels, a dry lower layer, and a moist midtropospheric layer. The convective instability is small or marginal, therefore the convection is often weak. The latter characteristic is consistent with numerous observations of dry microbursts associated with virga shafts pendant from innocuous clouds. Forecasting schemes are usually based on the 1200 UTC sounding and expectations for solar heating and maximum surface temperatures later in the day.

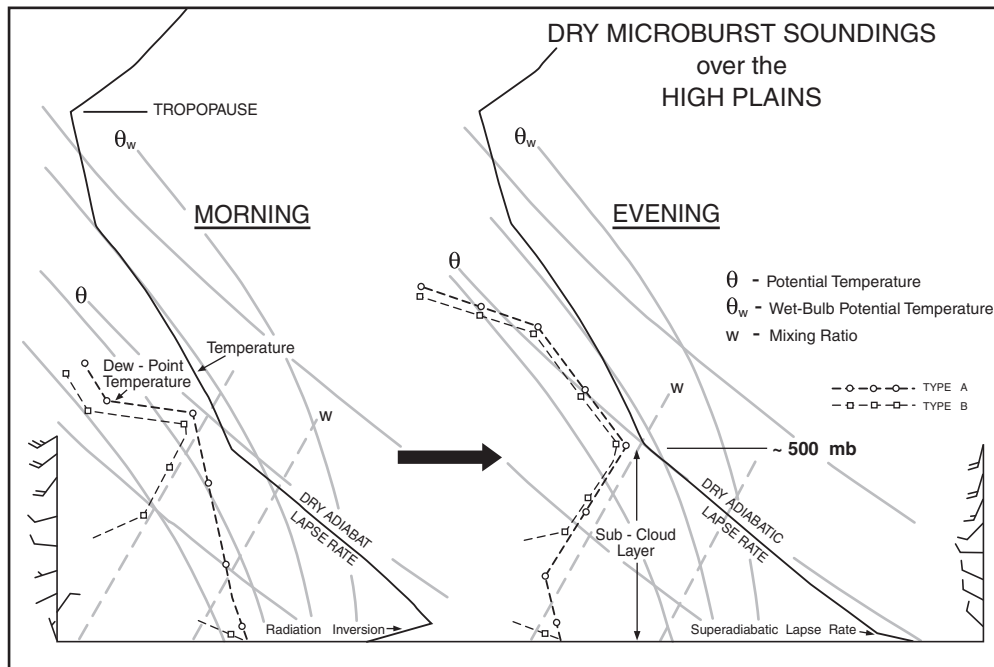


Figure 6 Schematic of the characteristics of the thermodynamic profile of the morning and evening soundings favorable for dry microburst activity over the high plains. Reproduced from the American Meteorological Society from Wakimoto, R.M., 1985. Forecasting dry microburst activity over the high plains. *Monthly Weather Review* 113, 1131–1143.

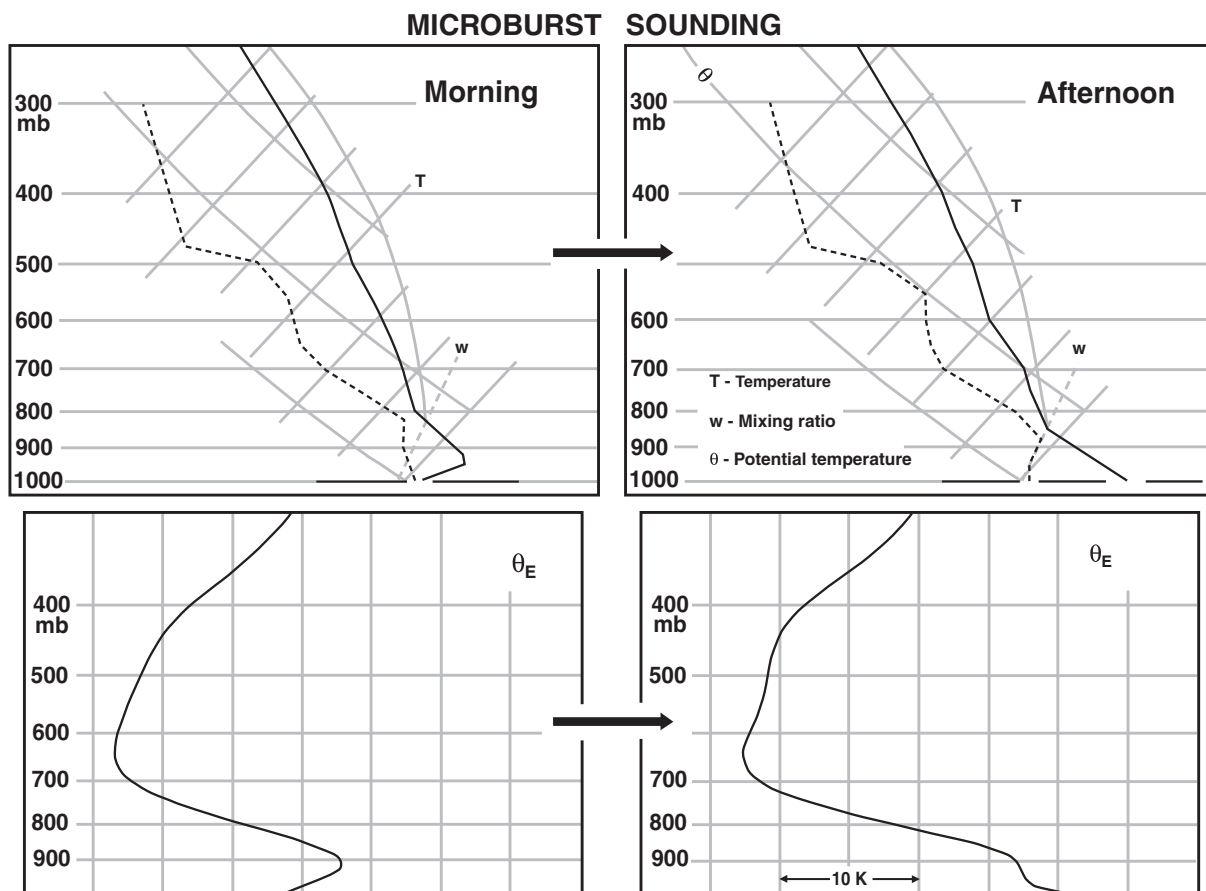


Figure 7 Thermodynamic model summarizing the environment conducive for wet microburst occurrence. Reproduced from the American Meteorological Society from Atkins, N.T., Wakimoto, R.M., 1991. Wet microburst activity over the southeastern United States. *Weather and Forecasting* 6, 470–482.

Wet microburst profiles typically display high moisture values through a deep, surface-based layer, with the top of the moist layer sometimes extending beyond 4–5 km AGL. Relative humidities above the moist layer are low. The dry adiabatic subcloud layer may only be 1.5 km deep and the capping inversion is weak. It has been suggested that the vertical profile of equivalent potential temperature can be useful in identifying environmental conditions capable of supporting wet microbursts (Figure 7). The difference in the equivalent potential temperature between the surface and midlevels equal to or greater than 20 °C appears to be a characteristic profile during wet microburst events. In addition to the morning sounding, predicted moisture and thermal advection patterns must be taken into consideration.

Microburst detection in real time has focused on the use of single-Doppler radar. A number of radar signatures have been identified that occur typically 2–6 min prior to the initial surface outflow. Descending reflectivity cores, increasing radial convergence within the cloud (in response to an accelerating downdraft), intense small-scale circulations, and weak echo reflectivity notches (in response to entrainment of low equivalent potential temperature air) have all been found to be important microburst precursors.

Basic research on microbursts has been quickly and successfully transferred into the operational community. Many of the airports around the United States have now implemented systems that provide timely warning to air traffic controllers and pilots of impending wind shear events associated with microbursts.

See also: **Mesoscale Meteorology:** Gust Fronts; Mesoscale Convective Systems. **Numerical Models:** Convective Storm Modeling.

Further Reading

- Fujita, T.T., 1981. Tornadoes and downbursts in the context of generalized planetary scales. *Journal of the Atmospheric Sciences* 38, 1511–1534.
- Fujita, T.T., 1985. The Downburst–Microburst and Macrobust SMRP Research Paper No. 210, NTIS PB-148880, report of Projects NIMROD and JAWS. University of Chicago Printing Department, Chicago, IL.
- Proctor, F.H., 1989. Numerical simulations of an isolated microburst. Part II: Sensitivity experiments. *Journal of the Atmospheric Sciences* 46, 2143–2165.
- Srivastava, R.C., 1985. A simple model of evaporatively driven downdraft: application to microburst downdraft. *Journal of Atmospheric Sciences* 42, 1004–1023.

Severe Storms

CA Doswell III, University of Oklahoma, Norman, OK, USA

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

The concept of a severe storm is defined, because there are many different types of weather that can be called a 'storm' and the severity of a storm is measured by the intensity of the weather produced. It is generally the case that as the scale of a storm decreases, its intensity increases, but it affects a smaller area and does not last as long as a larger scale storm. Storms roughly the size of a continent (several thousand kilometers across) are called 'synoptic-scale' storms. When the size of a storm is roughly the size of a few hundred kilometers across, it is called a 'mesoscale' storm. Severe thunderstorm systems can be considered mesoscale, but the individual storms are a few tens of kilometers across. Severe storms in the tropics form a special class of their own. Property and casualty mitigation from severe storm events is discussed.

Introduction

The word 'storm' implies a disturbance of some sort in the weather, but many different types of weather can result in an event called a 'storm.' Thus, it is possible to have windstorms, dust storms (which also are windstorms), hailstorms, thunderstorms, winter storms, tropical storms, and so on. Generally speaking, events called 'storms' are associated with cyclones; undisturbed weather is usually found with anticyclones.

Similarly, the meaning of severity needs to be considered. The *intensity* of the event in question is often the basis for deciding on the severity of that particular storm. However, if storm intensity is to be our basis for categorizing a storm as severe, then we have to decide what *measure* we are going to use for intensity. This also usually implies an arbitrary threshold for deciding the issue of severity. That is, weather events of a given type are going to be called severe when some measure of that event's intensity meets or exceeds a threshold, which can be more or less arbitrary. A hailstorm might be severe when the hailstone diameters reach 2.5 cm or larger, a winter snowstorm might be called severe when the snowfall rate equals or exceeds 5 cm h⁻¹. On the other hand, some storms of *any* intensity might be considered severe. A tornado is a 'storm' embedded within a thunderstorm; tornado of any intensity is considered a severe storm.

The difficulty with *arbitrary* definitions is that they imply a change in character whenever the threshold criterion is met. That is, if a hailstorm produces hailstones of 2.4 cm in diameter, a threshold of 2.5 cm means that such a storm is not severe. However, from most anyone's viewpoint, is it reasonable to try to distinguish between a storm producing 2.4 cm diameter hailstones from one producing 2.5 cm hailstones? In the majority of cases within the science of meteorology, there is no obvious way to distinguish events with this sort of precision. A small *quantitative* change in some intensity measurement is not necessarily associated with a *qualitative* change in the character of a storm. It is near the threshold (wherever that threshold is chosen) that it becomes challenging to analyze and predict storm 'severity.' This will be elaborated on in dealing with the specific events described below. However, the

challenge of *defining* severity should be kept in mind in the following discussion, as we consider various types of severe storms.

Severe Midlatitude Storms

The Tropics are defined formally as lying equatorward of 23.26° latitude in the Northern and Southern Hemispheres: the 'Tropics' of Cancer and Capricorn, respectively. Poleward of these latitudes and equatorward of 60° latitude lie the so-called midlatitudes. There are important distinctions between the weather of midlatitudes and that of the Tropics. Notably, in midlatitudes, the Coriolis force is a dominant part of the meteorology, whereas in the Tropics, its impact on the large-scale weather is of lesser importance, with some notable exceptions, including tropical cyclones.

Synoptic-Scale Storms

Cyclones in midlatitudes that are thousands of kilometers in horizontal extent are known as *synoptic-scale* systems. These are the familiar rotating weather systems (Figure 1) shown routinely in newspapers and on television. Such storms serve an important function in the global circulation, helping to carry warm air poleward from the tropics and cold air from the polar regions equatorward. This process keeps the imbalance of solar heating from creating an extremely strong temperature contrast between the poles and the equator. In association with these synoptic-scale cyclones, intense temperature contrasts can develop (Figure 2), called *fronts*, which are the leading edges of the cold air flowing equatorward and warm air flowing poleward. These midlatitude cyclones are part of the normal progression of weather systems, typically bringing clouds and precipitation with them.

In some situations, where the hemispheric weather patterns have become slow moving, these cyclones can result in prolonged periods of heat or cold in some regions. When extreme temperatures (hot or cold) are reached, these can be hazardous to humans for a variety of reasons, but would not generally be considered 'storms' despite that.

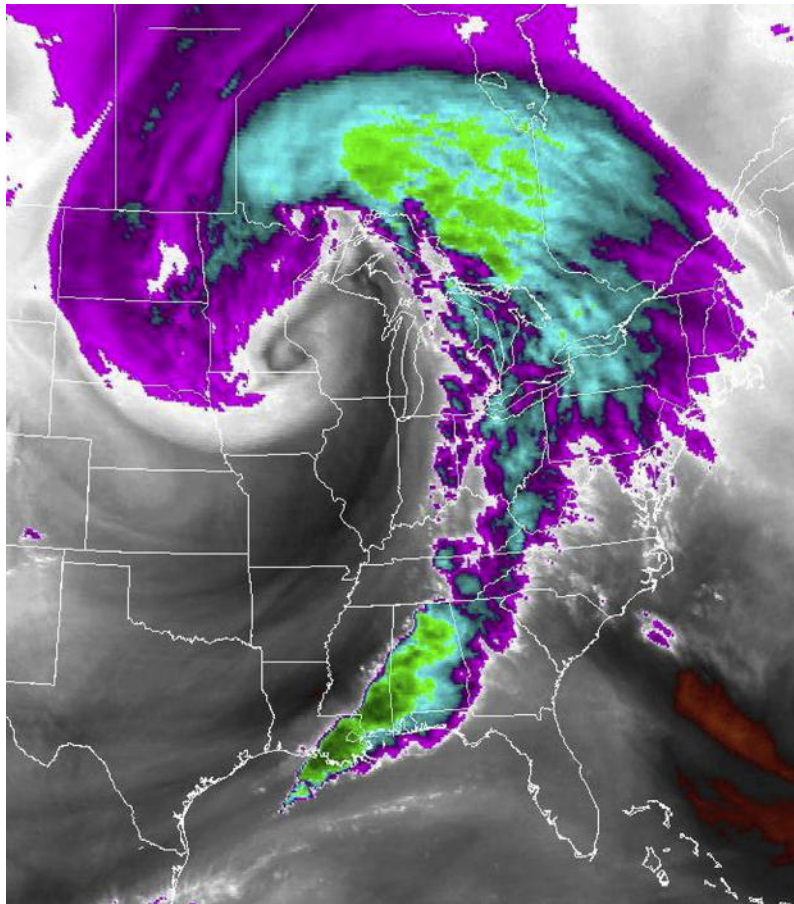


Figure 1 False color-enhanced satellite image of a synoptic-scale cyclone on the afternoon of 10 November 1998, showing the center of the cyclone near the spiral of clouds in southeastern Minnesota. This cyclone was producing severe thunderstorms in and near the Gulf of Mexico, as well as snow and high winds on the northern plains, in North and South Dakota. NOAA image.

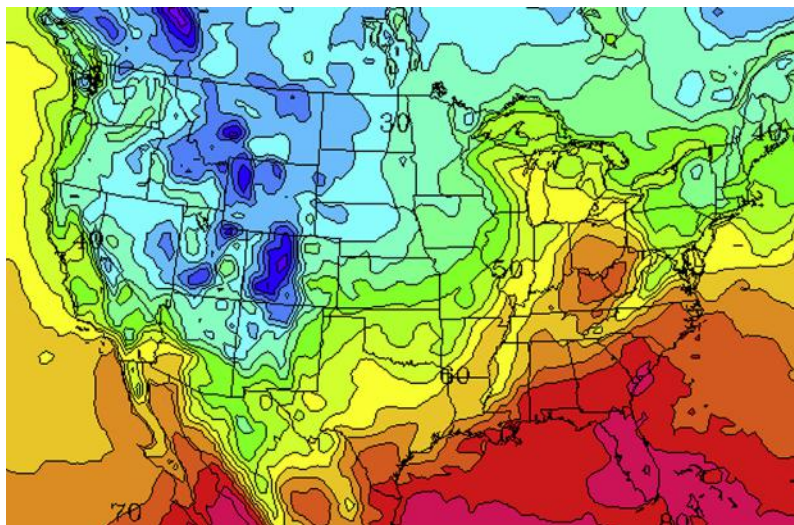


Figure 2 Map of surface temperatures at about the same time as **Figure 1**, showing the strong contrast in temperatures along the cold front, with subfreezing temperatures in North and South Dakota at the same time that quite warm temperatures are present over the Gulf of Mexico. Many subsynoptic-scale features can also be seen in mountainous regions; for example, in the Appalachian and Rocky mountains. Image courtesy of National Center for Atmospheric Research.

Synoptic-scale cyclones sometimes become particularly intense, and the pressures at their cores can become quite low, in comparison to the average. In such cases, during the process of intensification, pressures can fall quite rapidly as the result of the dynamic processes operating to cause the intensification. Cyclones with fast-falling pressures are sometimes called 'bombs' and, whereas they can be considered storms in their own right, these synoptic-scale cyclones may be responsible for several different types of stormy weather within them.

Rapidly falling pressures create strong winds over a wide region. Winds and heavy precipitation resulting from synoptic-scale cyclones can produce considerable damage and associated casualties; recent examples occurred in France during December 1999 and along the east coast of the United States in January 2000 and March 1993 (the so-called Superstorm of 1993). Another well-publicized example hit the United Kingdom in October 1987. Damaging winds can extend over many hundreds of kilometers and last in any one place for a full day or more. The result of such widespread damaging wind can be overwhelming to emergency services, and power outages alone can last for days in some places simply because of the sheer size of the affected area. At sea, strong winds from intense synoptic-scale cyclones also produce large waves that represent hazards to ships of all sorts. The winds from intense cyclones at sea can cause serious damage, including beach erosion, when they affect coastal areas.

In addition, intense synoptic-scale cyclones can produce a full spectrum of hazardous precipitation. Such storms can occur at any time of the year but are most common from autumn through spring, and so the cyclones are capable of paralyzing snowstorms, ice storms, heavy rainstorms, and even severe thunderstorms. Accumulations of ice and snow during winter storms of this type are potentially hazardous to ships and aircraft.

Depending on the circumstances, two or more of these different severe weather types could be happening at the same time, in different places. A given location might experience all of them in the course of a single day during the passage of a synoptic-scale cyclone. In other situations, only one form of severe weather occurs within such a cyclone.

Synoptic-scale cyclones are important in creating the conditions for the development of smaller scale storms. It is a general principle in meteorology that as the size of a weather system decreases, the maximum intensity of the weather it can create increases. Although synoptic-scale systems certainly can produce widespread damage, it is usually not of the most extreme intensity. However, the conditions within such storms can result in smaller concentrations of severe weather that become even more potentially hazardous.

Mesoscale Storms

Mesoscale weather is in the range of hundreds of kilometers, whereas synoptic-scale weather happens on scales of several thousand kilometers. Synoptic-scale weather processes go on essentially all the time (although the really intense events are generally rare), whereas mesoscale storms are intermittent. That is, they arise only occasionally in any given location and then only when the conditions for their formation are produced by the processes operating on the synoptic scale. There are two

general classes of mesoscale storm systems: those that arise from interactions between the atmosphere and the underlying surface, and those that occur even in regions of relatively uniform conditions at the surface.

Those systems that depend on the underlying surface cover a wide range of phenomena. There are many atmospheric circulations, like land-sea breezes that are more or less routine, driven by the underlying topographic conditions and occurring on most days of the year; in the case of the land-sea breeze, it is the temperature contrast between the land and the sea that drives the flow. During the day, the land is warmer and air tends to rise over land, to be replaced by cooler air flowing in from the sea. At night, the opposite happens. Of course, most of these circulations would not be considered 'storms' in the sense that we have been using. However, such processes as land-sea breezes can be influential in the *development* of stormy weather, often in the form of thunderstorms that are initiated along them.

Occasionally, the circumstances produced by the synoptic-scale flow as it interacts with the surface result in stormy conditions. A common example is when the airflow interacts with complex terrain, producing localized windstorms. There are examples of these mesoscale windstorms around the world, often given colorful names. Mesoscale windstorms such as the Chinook (in Alaska), the Foehn (in the European Alps), the Tramontana (in the western Mediterranean), the Bora (in the Adriatic), and so forth have been recognized as important weather events for centuries. Windstorms in complex terrain arise in different circumstances; they are not all driven by the same mechanism. Some are simply cases where cool, stable air is being funneled through gaps in the terrain (e.g., the Tramontana); others develop when strong winds aloft are brought down to the surface by processes induced by airflow over the mountains (as in Boulder, Colorado). The situation creating the windstorms is created by the synoptic-scale flow, but the strongest winds are confined to a mesoscale area.

Another class of mesoscale storms can arise when cold air flows over relatively warm waters. Certain storms of this sort, called 'polar lows,' apparently arise through processes not unlike those of tropical cyclones, drawing energy from the ocean to develop their intense circulations. They occur when outbreaks of very cold polar air flows over relatively warm waters. Given their mesoscale size, they often are characterized by intense pressure gradients, leading to the occurrence of strong windstorms. Their size means the weather they bring may only last for part of a day, but during the passage of the storm, winds can meet, and even exceed, the hurricane threshold: 33.5 m s^{-1} . The windstorms associated with polar lows can be quite hazardous, especially when they occur in association with low temperatures (resulting in severe wind-chill conditions). In addition, polar lows can produce blinding snowstorms with snowfall rates of perhaps 200 mm h^{-1} , leading to extremely dangerous blizzard conditions. On some occasions, they can be associated with strong and possibly severe thunderstorms, as well.

When cold air has a long fetch over the relatively warm waters of large lakes early in the cold season, yet another type of mesoscale storm can arise: the lake-effect snowstorm. The orientation of the synoptic flow determines when and where

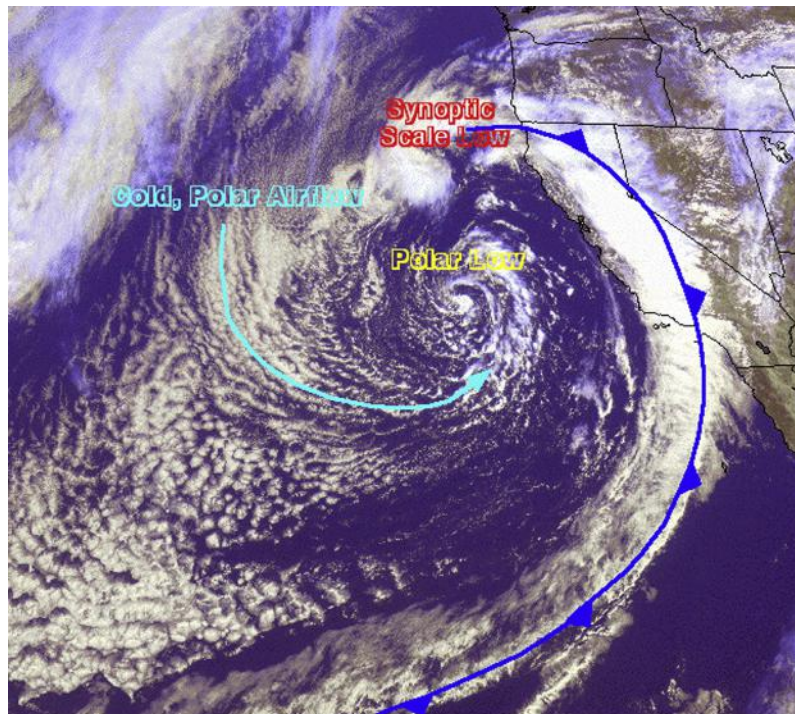


Figure 3 An example of a polar low in the cold airstream behind a wintertime synoptic-scale cold front associated with a synoptic-scale cyclone (a low-pressure center). NOAA image.

lake-effect storms develop. These snowstorms can produce sustained heavy snowfall rates, resulting in deep snow accumulations in preferred locations and sometimes are accompanied by thunder and even waterspouts.

Even when the underlying surface is more or less uniform, mesoscale storms can develop *within* synoptic-scale cyclones (Figure 3). These usually are tied to a mesoscale disturbance in the middle or upper troposphere that encounters conditions favorable for its intensification. Such systems can produce unforeseen snow and ice storms in the winter, and severe thunderstorms during the warm season. There may not be a strong cyclone near the surface in such events.

Severe Thunderstorms

Severe thunderstorms typically produce weather events that cover a wide range of size scales, from a few hundred kilometers down to just a few kilometers, or even smaller. This is because thunderstorms can occur as isolated events or in groups. In the United States, a thunderstorm-related event is considered severe when the wind gusts equal or exceed 25 m s^{-1} , or the hailstone diameters exceed 2.5 cm, or if a tornado is produced. As noted earlier, these criteria are mostly arbitrary.

A thunderstorm is composed of one or more cells, where a *cell* is the basic building block of a thunderstorm. Cells, in turn, are made up of one group of air parcels being driven upward by positive buoyancy and another being driven downward by negative buoyancy and the presence of precipitation in the air. Positive buoyancy arises in updrafts by the release of latent heat during the condensation of water vapor.

This heat release acts like the burner of a hot air balloon, reducing the density of the air in which condensation is occurring and thereby causing the air to rise. As the process continues, under the right conditions, precipitation forms in the updraft. This precipitation can produce downdrafts simply by its accumulating weight dragging downward on the surrounding air. Moreover, when precipitation falls into relatively dry air surrounding a developing storm, the evaporation of that precipitation chills the air; evaporation absorbs latent heat from the air in the same way that condensation releases that heat.

When downdrafts caused by thunderstorms reach the surface, they are forced to spread out, like pancake batter poured onto a griddle. This creates an outflow at the surface (sometimes called a *downburst*), with the outflow winds sometimes reaching the criterion for calling the thunderstorm severe. On some occasions, these outflow winds can exceed 40 m s^{-1} .

For particularly strong updrafts, the possibility of large hail formation arises. Hailstones develop in the part of the storm where supercooled water and ice crystals are both present; liquid water is said to be supercooled when its temperature is below the melting point (0°C) and the water is not yet frozen. Hailstones can become quite large, exceeding 5 cm diameters at times, and be capable of penetrating roofs, shattering windows, and occasionally creating human casualties. Even small hail can cause crop damage, of course.

In general, tornadoes form in association with severe thunderstorms. Tornadoes are intense low-pressure vortices that can produce the strongest winds of any storm; at their highest intensity, tornadic wind speeds can approach 140 m s^{-1} . Most tornadoes, however, are not that intense.

Tornadoes over bodies of water are called *waterspouts*, but there is no scientific distinction between them. Tornadoes are created in thunderstorms when pretornadic, relatively weak circulations are intensified through conservation of angular momentum.

Isolated Events

The most intense form of thunderstorm is the so-called *supercell* thunderstorm, which typically is isolated from surrounding storms. Supercells are rotating thunderstorms that develop their rotation by drawing upon the vertical wind shear in the prestorm environment. The vast majority of supercells produce some sort of severe weather: hail, damaging straight-line winds, and/or tornadoes; only about 20% of them are tornadic, however. The most violent severe weather of all types is almost always associated with supercells (Figure 4), including the majority of strong and violent (F2–F5 on the Fujita scale) tornadoes and giant hailstones (exceeding 5 cm in diameter).

Whereas the typical thunderstorm cell has a lifetime of about 20–30 min, supercells can persist for many hours. This means that all forms of severe weather from supercells can be prolonged, sometimes leaving long, wide swaths of damage (see Figure 5). The organized nature of a supercell, associated with its overall rotation, means that supercells produce a disproportionate share of the damage associated with thunderstorms. Perhaps only about 10% of all thunderstorms are supercells, but they are responsible for the majority of thunderstorm damage in areas where they occur.

Because supercell updrafts are often intense, supercells can become prolific hail producers; a noteworthy example was a supercell that hit the Dallas–Fort Worth Metroplex on the evening of 5 May 1995 with softball-sized hail and torrential rains. The damage from that one storm was estimated at \$1 billion and there were numerous hail-related injuries.

Apart from supercells, isolated thunderstorms usually are nonsevere and typically do not last very long. On rare occasions,

isolated thunderstorms can produce a brief ‘pulse’ of severe weather, usually hail or winds that are only marginally severe.

Aggregations of Thunderstorms

Thunderstorms do not always occur as isolated events. Rather, they tend to form in groups, in either lines or clusters of individual cells. The most common grouping is in lines, sometimes referred to as *squall lines*. When thunderstorm cells form in aggregations, then the collection of storms can live for a much longer time than the individual cells (which retain their 20- to 30-min lifetimes). This means that the hail and wind events produced by such groupings of thunderstorms are intermittent, rather than prolonged (as with supercells), as cells form and decay within the group. Severe weather still can go on in such cases for many hours in this intermittent fashion. The interactions between individual cells in lines and clusters of thunderstorm cells are often complicated and hard to predict, but those interactions can also be responsible for severe weather.

A particularly dangerous form of thunderstorm aggregation arises when new cells are constantly forming in one place, and tracking over the same region repeatedly, a situation called ‘training’ because the cells are like cars in a train. This means that a particular area experiences rainfall from a succession of thunderstorm cells, which can result in extremely heavy rainfall. This is the process associated with the majority of flash flood events, worldwide. In the United States, heavy rainfall is not considered to be a criterion for what is officially considered to be ‘severe’ in spite of the importance of such rainfall in flooding events. On the other hand, many other nations around the world consider heavy rainfall to be an important form of severe storm.

Severe Tropical Storms

The most obvious severe weather associated with the tropics are *tropical cyclones*. These storms are known by different



Figure 4 Supercell-associated tornado on 9 June 2005, near Zurich, Kansas. Image © 2005 C. Doswell (used by permission).

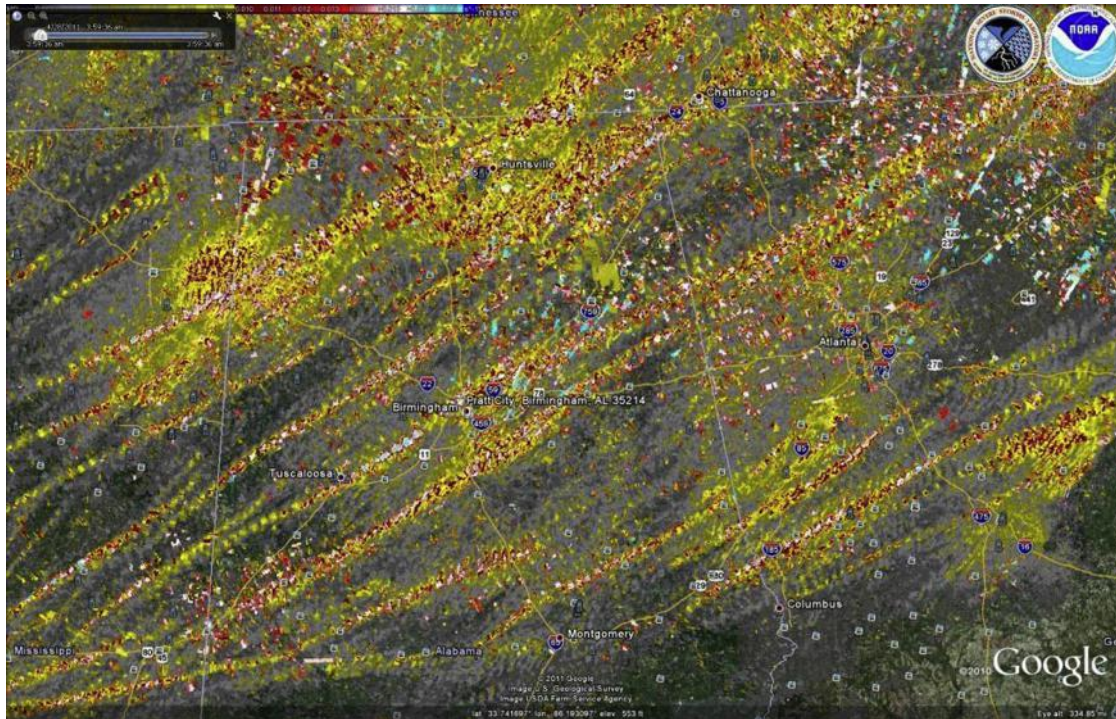


Figure 5 Map of supercell tracks associated with the outbreak of severe storms and tornadoes on 27 April 2011, over the southeastern United States. NOAA image.

names in different parts of the world: hurricanes (in North America), typhoons (in the tropical Pacific), and cyclones (in the Indian Ocean and Australia), among others. However, they all are the same phenomenon. Such storms arise when sea surface temperatures become warm, the vertical wind shear is weak, and tropical weather disturbances move through the easterly trade winds of the Tropics. They produce winds in excess of 33.5 m s^{-1} and the peak sustained winds (i.e., not gusts) can approach 90 m s^{-1} in extreme cases. The size of the region of damaging winds can vary considerably from one event to another, but winds exceeding 'hurricane force' (33.5 m s^{-1}) can be found within a circle on the order of 100 km or so in diameter. Such a large region of strong winds means that damaging wind speeds can go on for many hours.

Although they are well known for strong winds, tropical cyclones can pack a lethal combination of hazards: storm surge, heavy rainfalls, and even embedded tornadoes, as well as the more well-known strong winds. Storm surge is created by a combination of strong winds and low pressure, resulting in an elevated sea level near the center of the storm. When this surge, which can be several meters high, makes landfall, low-lying coastal regions can be inundated. The rainfall component is nothing to take lightly, either. Hurricane Mitch (Figure 6) devastated parts of Nicaragua and Honduras in 1998, mostly from flash floods and landslides. There were more than 9000 fatalities, making it the worst weather disaster in the twentieth century in the Western Hemisphere.

Tropical cyclones are usually several hundred kilometers in diameter and can last for tens of days. Their paths often

take them out of the Tropics into midlatitudes, where they can maintain their structure for a time before eventually dissipating or transforming into midlatitude cyclones. Tropical storms usually dissipate shortly after making landfall, because their energy source (warm seawater) is cut off. Nevertheless, dissipated tropical cyclonic storms can remain dangerous well after they lose their strong winds by creating an environment favorable for heavy rain-producing thunderstorms.

Relatively little is known about other types of severe storms in the Tropics. Severe thunderstorms, especially supercells, are uncommon in the Tropics because of a general lack of vertical wind shear. Of course, heavy rain-producing tropical thunderstorms are relatively common in some parts of the Tropics.

Societal Impacts and Their Mitigation

Severe storms in all their variety cause the loss of hundreds of lives and several billion dollars in property during the course of a year in the United States. It is worth noting that the United States can recover from such property damage because of its large, generally healthy economy. Severe storm economic losses in the United States typically are much less than 1% of the gross domestic product (currently several trillion dollars), so by spreading out the impact of severe storms, the areas affected can recover and rebuild. On the other hand, when severe storms (e.g., Hurricane Mitch) devastate less-developed nations with small economies, the damage to their infrastructure can be so large that it might take decades to recover.

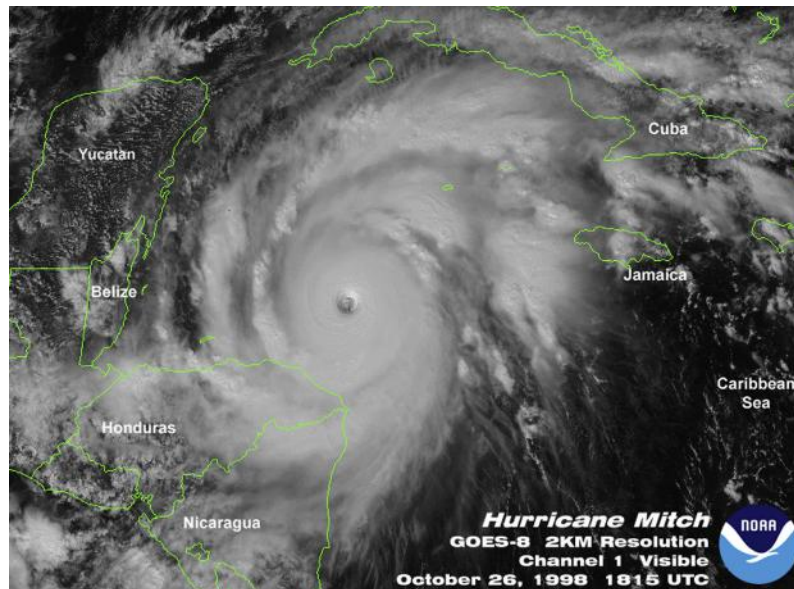


Figure 6 View from the GOES-8 geostationary satellite of Hurricane Mitch near Honduras and Nicaragua. NOAA image.

Forecasting severe storms has shown a slow increase in accuracy during the past several decades, as new technologies are leading to improved understanding and predictability. The accuracy of forecasts generally increases as the scale of the storm increases; it is possible to be more accurate with a synoptic-scale forecast than with a forecast on the scale of a single thunderstorm in most cases. There is more complete understanding of the synoptic-scale meteorology than on scales smaller than synoptic. Furthermore, forecast accuracy generally decreases with the age of the forecast, at a rate that also depends on the scale. The accuracy of a synoptic-scale forecast can stay high for a few days, whereas a forecast of a thunderstorm-scale event can remain accurate for a few tens of minutes, at most.

Property damage mitigation depends mostly on making the right preparations for the storms that are possible in a given location, well in advance of the storms. Once the storms are underway, there tends to be relatively little that can be done to prevent property damage. For example, a home built on a barrier island that can be swept by landfalling tropical cyclones is unlikely to remain undamaged for more than a few decades, at most. As the storm approaches, there is little time to do anything substantial to the home to limit structural damage. Any damage mitigation would be of a superficial nature, such as boarding up the windows. Thus, structural damage can be prevented primarily by not building in vulnerable areas. On the other hand, there are several ways in which homes can be constructed to resist tornado damage (Figure 7), unless the homeowner is unlucky enough to be hit by the most intense winds in a violent tornado. Within the damage swath produced by a violent tornado, only a few places will actually experience the most violent winds; most of the rest of the structures will encounter winds that can be resisted through appropriate construction practices.

Casualty mitigation can be a complex topic, as well. In some instances, as with tropical cyclones, evacuation is possible and



Figure 7 Damage caused by a violent tornado that hit the city of Moore, Oklahoma, on 3 May 1999. FEMA image.

may be the best way to protect lives when it is feasible. For tornadoes, having access to a suitable shelter is preferred; in situations where proper shelter is not available, the alternatives during tornadoes are not very good. For flooding situations, evacuation to higher ground is the appropriate way to prevent casualties, when time permits.

The ability to detect and predict severe storms is also important for casualty mitigation. In the United States, there has been a gradual reduction in weather-related fatalities with time, in part because there are fewer 'surprise' storms today, and in part because education about severe storm hazards has led to improved public preparations. Nevertheless, we continue to be vulnerable as a nation to disasters caused by severe storms, and complacency can be a fatal error.

See also: Aviation Meteorology: Aviation Weather Hazards.
Boundary Layer (Atmospheric) and Air Pollution: Diurnal Cycle.
Data Assimilation and Predictability: Predictability and Chaos.

Dynamical Meteorology: Baroclinic Instability; Coriolis Force; Rossby Waves; Waves. **Hydrology, Floods and Droughts:** Flooding. **Mesoscale Meteorology:** Bow Echoes and Derecho; Convective Storms: Overview; Mesoscale Convective Systems. **Mountain Meteorology:** Downslope Winds; Land and Sea Breezes; Lee Waves and Mountain Waves; Orographic Effects: Lee Cyclogenesis; Valley Winds. **Synoptic Meteorology:** Anticyclones; Cyclogenesis; Extratropical Cyclones; Fronts; Lake-Effect Storms; Polar Lows. **Weather Forecasting:** Severe Weather Forecasting.

Further Reading

- Agnone, J.C. (Ed.), 1995. *Raging Forces: Earth in Upheaval*. National Geographic Society, Washington, DC.
- Anthes, R., 1982. *Tropical Cyclones. Their Evolution, Structure and Effects*. American Meteorological Society, Boston, MA.
- Church, C., Burgess, D., Doswell, C., Davies-Jones, R. (Eds.), 1993. *The Tornado: Its Structure, Dynamics, Prediction, and Hazards*. American Geophysical Union, Washington, DC.
- Doswell III, C.A. (Ed.), 2001. *Severe Convective Storms*. American Geophysical Union, Boston, MA.
- Foote, G.B., Knight, C.A. (Eds.), 1977. *Hail: A Review of Hail Science and Hail Suppression*. American Geophysical Union, Boston, MA.
- Hill, C.E. (Ed.), 1986. *Nature on the Rampage: Our Violent Earth*. National Geographic Society, Washington, DC.
- Junger, S., 1997. *The Perfect Storm*. W.W. Norton, New York.
- Lamb, H., 1991. *Historic Storms of the North Sea, British Isles and Northwest Europe*. Cambridge University Press, Cambridge, New York.
- Lorenz, E.N., 1993. *The Essence of Chaos*. University of Washington Press, Seattle, WA.
- Ludlam, F.H., 1980. *Clouds and Storms*. Pennsylvania State University Press, University Park, PA.
- Markowski, P., Richardson, Y.P., 2010. *Mesoscale Meteorology in Midlatitudes*. Wiley-Blackwell, Hoboken, NJ.
- Shapiro, M., Grønås, S. (Eds.), 1999. *The Life Cycles of Extratropical Cyclones*. American Meteorological Society, Boston, MA.

Waterspouts

JH Golden, Forecast Systems Laboratory, NOAA, Boulder, CO, USA

© 2003 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 6, pp 2510–2525, © 2003, Elsevier Ltd.

Introduction

What is a waterspout? It is defined as an intense columnar vortex (usually containing a funnel cloud) of small horizontal extent that occurs over a body of water and is suspended from a cumuliform cloud.

Waterspouts occur most frequently in the subtropics during the warm season; more are reported in the Florida Keys than in any other place in the world. Funnel diameters range from a few up to 100 m or more, lifetimes average 5–10 min, but large waterspouts may persist for up to 1 h and are the most hazardous. Examples of typical and large Florida Keys waterspouts are shown in Figures 1 and 2. Waterspouts can also occur in groups or sequential families of up to nine from a single cumulus cloudline over a 90 min period, as shown in Figure 3.



Figure 1 'Typical' waterspout, photographed by the author from light aircraft, west of Key West, FL, on 9 July 1969.

Worldwide Climatology

Literature on waterspouts is virtually nonexistent. By far, the majority of the work on this subject comprises individual observations of waterspout occurrence from the surface, a few documented by still photographs or drawings constructed from memory. Perhaps one of the best early comprehensive surveys of waterspout structure and behavior, deduced from observations available up to that time, was made by Ferrel in his *A Popular Treatise on the Winds*. Up to the late 1960s, there remained much disagreement over certain structural features of the waterspout, especially the sense of vertical motion in and surrounding the funnel. Vortex structure within the parent cloud remains virtually unknown. Some of the more plausible waterspout models deduced from surface observations have been proposed by Bundgaard, Dinwiddie, and Rossman. These



Figure 2 Large waterspout, photographed by the author during a chance aerial encounter near Lower Matecumbe Key, FL, on 2 September 1967. Note the double-walled structure of the funnel cloud, with spray vortex and trailing wake on sea surface.

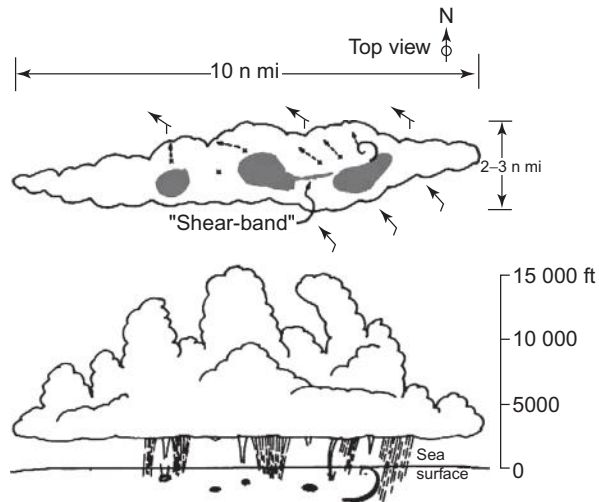


Figure 3 Schematic, composite scaled model of a waterspout-active cloudline, illustrating typical locations of waterspouts in various stages of life cycle (to be outlined later), motion relative to shower locations, and mesoscale flow (plan view above, side view below).

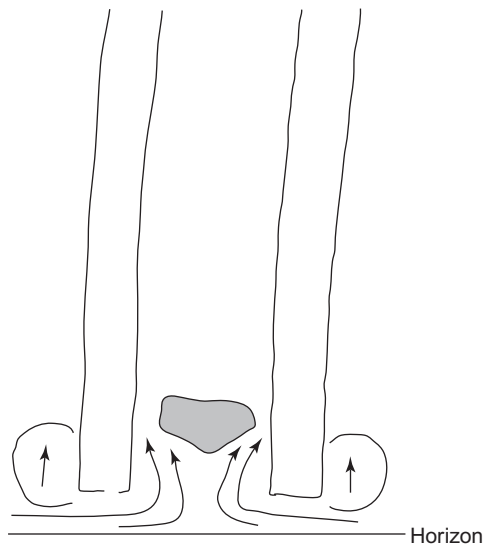


Figure 4 Waterspout model derived by Dinwiddie, 1959.

are depicted in **Figures 4–6** and illustrate some of the contrasting models of waterspout structure and radial-vertical circulation up to the 1960s. The only estimate of the pressure minimum in a waterspout was given by Chollet. He described a ship being overtaken by a waterspout during which the ship's barometer fell 21 hPa.

Gordon assembled a global waterspout frequency map, using British ship observations from 1900 to 1947. Apart from the observational bias of using only British ships and their selective routes, we note that more waterspouts occur on the western sides of the subtropical anticyclones. This distribution implies that waterspouts are most frequent in the equatorial Atlantic and Indian Oceans, the Gulf of Mexico, and the Mediterranean Sea. Gordon concluded from his data that

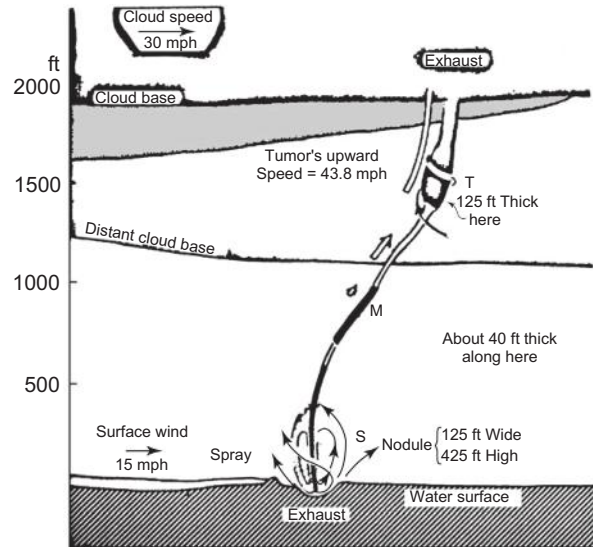


Figure 5 Schematic waterspout flow model from the frame of a 16 mm movie taken of a 1953 Tampa Bay waterspout by Bundgaard, 1953.

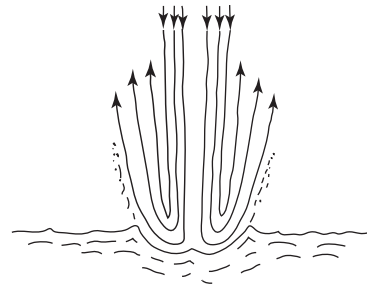


Figure 6 Rossman's, 1960 waterspout conceptual model: 'Streamline divergence at the base of a waterspout and shape of the sea surface... gives schematically the flow pattern of a waterspout foot: a depression in the water surface or a ringshaped wave, the crest of which breaks up in water droplets and forms the spray around the funnel'.

waterspouts occur generally over regions having relatively high sea and air temperatures for the latitude, where consequent unstable conditions are likely to exist in low levels.

Gordon's frequency chart illustrates that waterspouts rarely occur along the western coasts of North and South America, Europe, and Africa where cold upwelling in the ocean is common.

United States Waterspout Climatology

More recently, the author performed an assessment of waterspout frequencies along the US East and Gulf Coasts. Two primary data sources were utilized in the compilation of this waterspout climatology. All NOAA *Storm Data* reports from coastal counties during the period 1959–73 were examined for waterspouts, tornadoes, and funnel clouds. Tornadoes that passed over coastal lakes or waterways, or crossed a coastline and continued out to sea were included in the waterspout

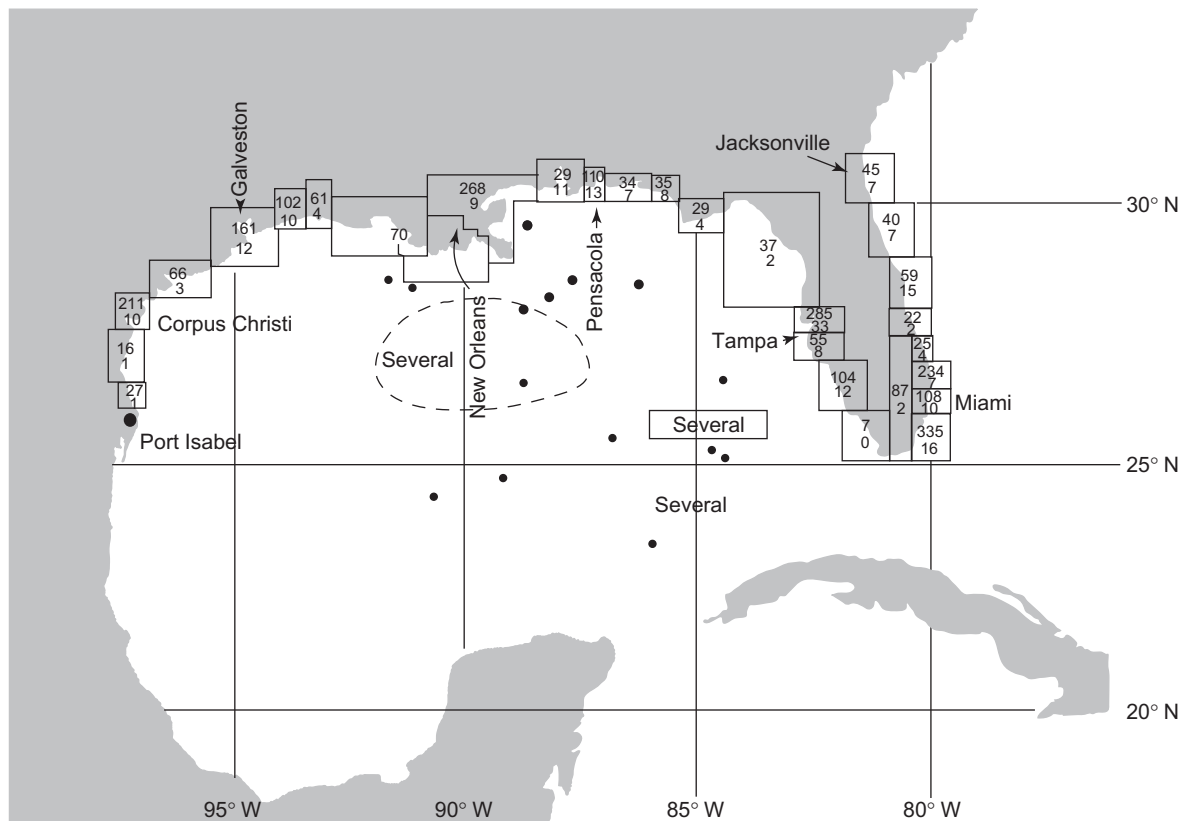


Figure 7 Waterspout climatology for US Gulf and Florida Coasts (*NOAA Storm Data*: 1953–73; ship reports: 1850–1940). Different boxed coastal regions were chosen to correspond to major population centers and geographical boundaries. The upper number refers to total waterspouts (see text) and the lower number to damaging cases. ‘Several’ implies more than 10 waterspouts.

statistics and plots. Funnel clouds that were likewise reported over water bodies were also included as probable waterspouts, using the research results by the author on the waterspout life cycle as a guide. The second data source was a large file of mariners’ reports covering almost a century, 1850–1940, for the entire western North Atlantic and Gulf of Mexico. This file was kept by past editors of the *Mariner’s Weather Log*, and is therefore probably biased toward United States vessel reports. The results for waterspout totals and damaging waterspouts for demographically partitioned areas are plotted in **Figure 7** for the US Gulf and Florida coasts, and similar results are given by the author for the US East Coast north of Jacksonville. (Waterspout statistics for the Florida Keys have been omitted from **Figure 7**, but are given below for the years 1958–68.) As noted in the Introduction, over 30 years of research has shown that waterspouts occur more frequently in the Florida Keys than anywhere else in the world. Moreover, Golden (1973) demonstrated that conventional data sources for the Keys underestimate the actual yearly waterspout population by up to an order of magnitude. This tendency is likely present in the *Storm Data* used for the Gulf and East Coasts; public apathy and lack of understanding about potential waterspout hazards are contributing factors. **Figure 7** and other similar waterspout charts of the East Coast show that the primary warm axis of the Gulf Stream and large coastal bays and inland waterways are favored regions of waterspout occurrence. More recent studies show increasing reports of waterspouts along the Southern

California coast and Great Lakes (Buffalo and Cleveland), both partly due to increased public awareness. Rare outbreaks of waterspouts have been documented in such odd northern locations as Lake Winnipeg (Canada), the Great Salt Lake (UT), Lake Tahoe (NV), and Baseline Reservoir (Boulder, CO).

Table 1 gives a list of the 10 most active areas along the entire US Gulf and East Coasts in decreasing order of reported waterspout frequency for the 15 year period. The Florida Keys experience from 50 to 500 waterspouts *each year* (at least 400 waterspouts were documented during the 1969 Lower Keys Waterspout Project). Total waterspouts and the number producing damage are given for each coastal area. Tampa Bay has the greatest number of damaging waterspouts and it is known that half or more of these originate over the Gulf of Mexico during midlatitude disturbances.

Annual/Seasonal/Diurnal Waterspout Frequencies in the Florida Keys

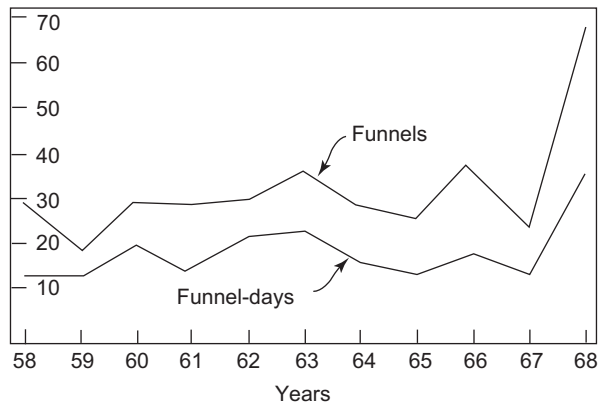
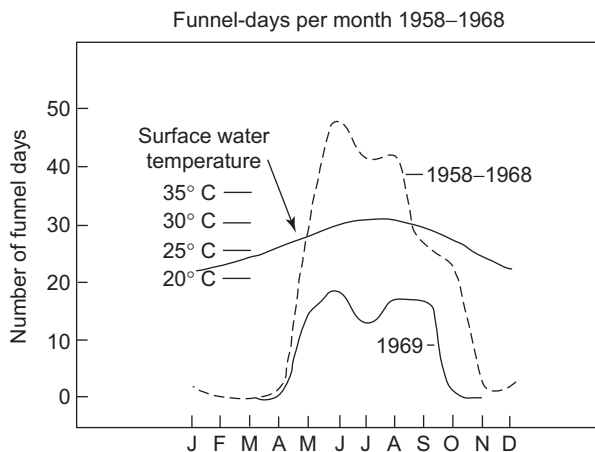
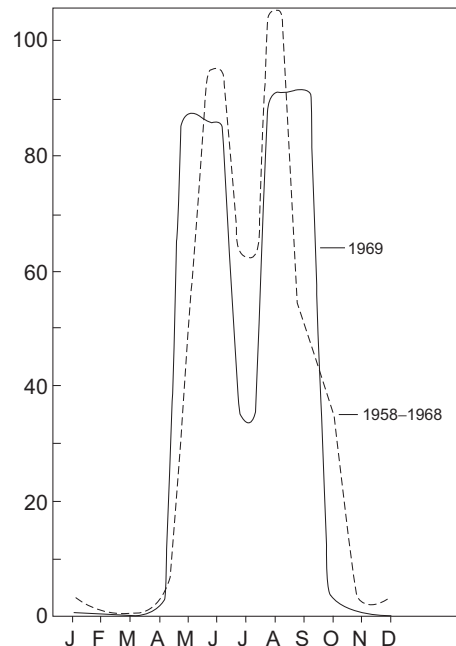
According to Gerrish and to 1969 data studied by Clemons, nearly all vortices reported as funnel clouds over water actually are weak waterspouts, and this has been confirmed by the present author’s research. **Figure 8** shows the annual totals of funnels and funnel-days from 1958 to 1968 for Key West Weather Service Office observations. A ‘funnel-day’ is defined as a day when one or more funnels were sighted. Note that

Table 1 Top 10 coastal areas for waterspout occurrence: US East and Gulf Coasts, 1959–73 (all values are 15 year totals)

Location ^a (areal coverage)	Total spouts	Spouts per unit area ($\times 10^{-4} \text{ km}^{-2}$)	Damaging cases
1. Florida Keys ^b (22 809 km ²)	> 1000	> 6572	15
2. Greater Miami, FL (10 138 km ²)	335 (++)	330	16
3. Tampa Bay, FL (6970 km ²)	235 (++++)	363	33
4. Palm Beach, FL (5069 km ²)	234	462	7
5. Corpus Christi, TX (6246 km ²)	211 (+++++)	338	≥ 10
6. Ft Lauderdale-Del Rey Beach, FL (5069 km ²)	180 (++)	355	10
7. Galveston Bay, TX (11 560 km ²)	161 (+++++)	139	12
8. Mississippi River Delta ^c , N Orleans (South of Lake Maurepas to Mississippi Delta) (14 790 km ²)	142 (++)	96	12
9. Pensacola Bay, FL (4164 km ²)	110 (+)	264	13
10a. Ft Myers, FL (12 672 km ²)	104	82	12
10b. Mississippi Sound, MI (5651 km ²)	103 (++)	182	16
10c. Port Arthur, TX, to Sabine Lake (5711 km ²)	102 (+)	179	10

^aCompare these data with [Figures 1 and 4](#).^bEstimated from field data in Golden (1973) and Rossow (1970).^cLake Pontchartrain, north of New Orleans, had 96 (+++++) total waterspout reports, one damaging (*not* included here).

Note: Each + indicates one observation of several or numerous waterspouts and was counted as only one event in the tabulations.

**Figure 8** Number frequency of funnel clouds and funnel-days per year (1958–68) from the Key West, FL, National Weather Service, NOAA data. Adapted with permission from Clemons, 1969.**Figure 9** Funnel days per month for the period 1958–68 (dashed) and for 1969 (bottom, solid) from Key West NWS data. The top solid line is the mean monthly sea surface temperatures at Key West pier during 1958–68.**Figure 10** Number of funnel clouds per month for the period 1958–68 (dashed) and for 1969 (solid) from Key West data. Note the double peak in each curve and the distinct minimum in July.

there were 35 funnel-days reported by Weather Service observers during 1968. However, including data from other sources in the Lower Keys, the number swells to 56 funnel-days observed within 30 km of Key West for the period. Lower Keys funnel sightings during 1968 far exceeded the annual frequency for the previous nine years.

Waterspouts in the Lower Keys are primarily a rainy-season phenomenon (May to October). The dashed curve ([Figure 9](#)) shows the number of funneledays per month at Key West (1958–68). June is the most active waterspout month, in terms

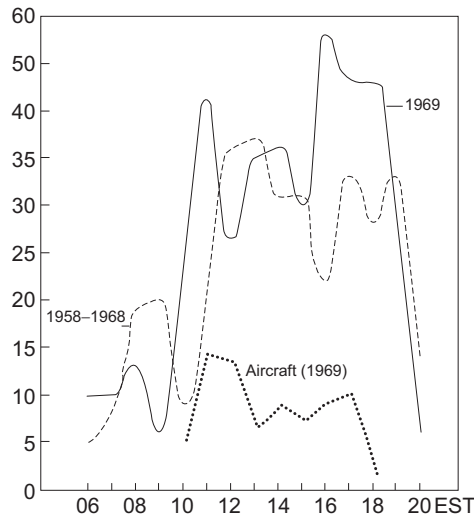


Figure 11 Diurnal trends in total number of funnels observed each hour (EST) during 1969 (solid) and 1958–68 (dashed); funnel observations from research aircraft are for 1969 (dotted).



Figure 12 Large tornadic waterspout accidentally photographed by time exposure near midnight, looking east from a balcony at N Miami Beach, 15 June 1985. © Jim Leonard, used with permission.

of funneldays, closely followed by August and July, with May, September, and October substantially less active. Note the in-phase relationship between trends in funneldays per month and monthly mean ocean water temperature at Key West. **Figure 10** (dashed curve) shows the 11 year funnel cloud totals plotted by month. From April to May, the total funnel number increases from a few to 50 and then nearly doubles from May to June (the secondary maximum).

Diurnal Occurrence and Duration

The diurnal distribution of waterspout funnels (1958–68) is given by the dashed curve in **Figure 11**. There are two primary maxima: the first near the noon hour (1130–1300 EST) with a slight decrease near 1330 EST, and the second maximum from 1630 to 1830 EST. A lower tertiary peak occurs in the early

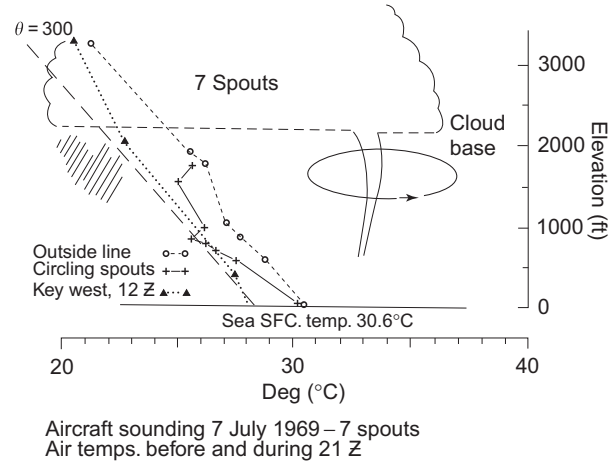


Figure 13 Temperature sounding taken from an outside-dial temperature gauge on light aircraft (read to the nearest 0.5°C) beneath a waterspout-active cloudline on 7 July 1969. Key West, FL, sounding at 1200 GMT is also plotted for comparison (9 h earlier and approximately 20 km from cloudline).



Figure 14 Mature waterspout with intense spray vortex, being chased by a large speedboat towing a water-skier, over shallow waters west of Key West, FL, in June 1977.

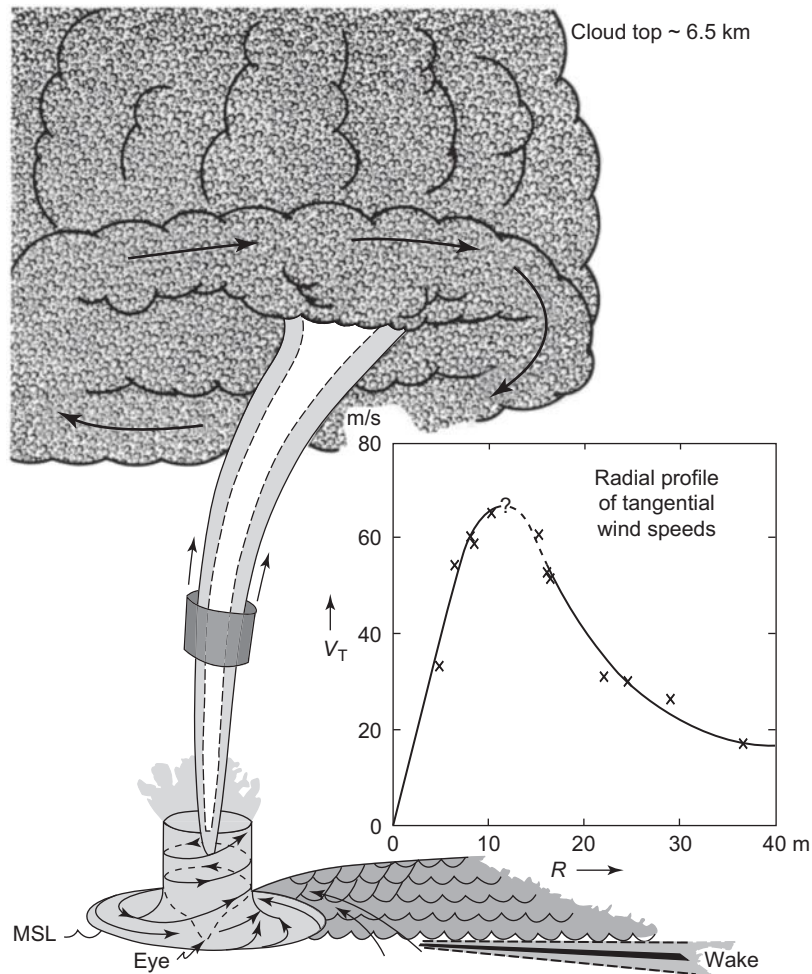


Figure 15 Proposed three-dimensional structural model of a waterspout based on detailed aircraft observations of the Lower Matecumbe Key waterspouts at close range. The vertical scale is greatly contracted. The inset shows the derived radial profile of tangential speeds through the waterspout's spray vortex, obtained by photogrammetric analysis of aircraft movies. Note the trailing 'wake' of disturbed sea water, with the wave train oriented perpendicular to and on the right side of the wake (cyclonic waterspout). See text for additional details.

morning between 0800 and 0900 EST. A few nocturnal waterspouts have been reported by Coast Guard and Navy pilots flying at low altitude and by Keys residents. A recent, extremely rare photograph of a large tornadic waterspout at around midnight is shown in Figure 12. This photograph was taken in time exposure from the balcony of a high-rise apartment in North Miami, looking east. The photographer was shooting lightning photographs and captured the large waterspout funnel and spray vortex by accident.

The average duration of funnel events (including both funnel clouds and waterspouts, 1958–67) was 14.6 min, which includes 61% of the total number; 1968 sightings averaged 13.0 min, which accounts for 66% of the total. For the 10 years 1958–67 and for the year 1968, 75% of the funnel events persisted for 20 min or less. The longest-lived funnel during the 1968 season lasted 52 min. Sixty-two minutes is the longest single event recorded (1958–67), but multiple funnel activity has been observed to last for almost 3 h from the ground.

Multiple Waterspout Events

Multiple-funnel sightings, with two or more occurring simultaneously, or more than one event on the same day, occurred on 54 of 167 funnel-days (1958–67) and on 14 of 35 funnel-days during 1968 alone. The maximum number of funnels sighted from the surface in Key West in a single day is eight (in June 1966 and August 1968). During one 45 min period of a 30 June 1969 flight, the author observed nine complete waterspouts emanating from the same cumulus cloudline. The maximum number sighted simultaneously from Key West (1958–68) is six; five have been reported several times.

Favored Weather Conditions for Waterspouts

A study was conducted using the above waterspout/funnel statistics with the NOAA Local Climatological Data for Key West, May to September, years 1964–68. High surface humidity and

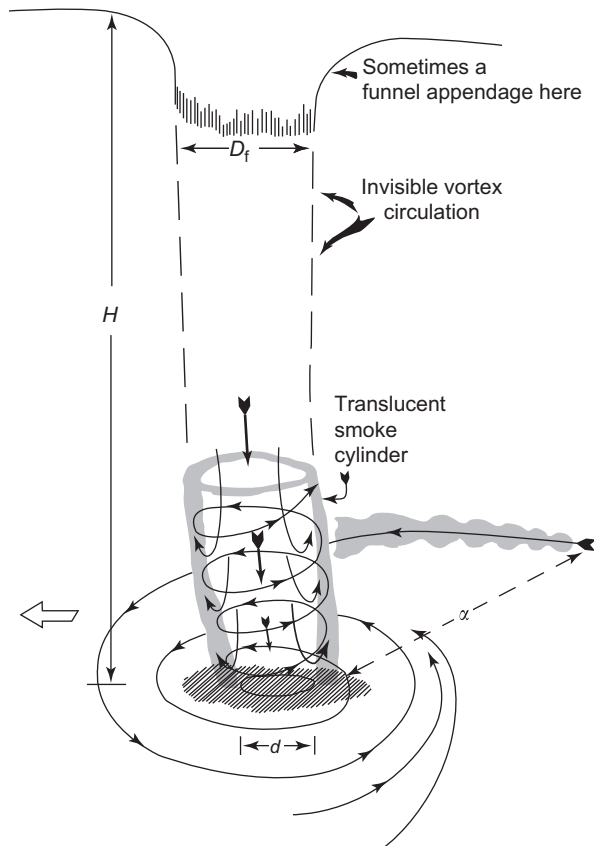


Figure 16 Composite structural-flow model for the 'dark spot stage' of the waterspout life cycle. Compare with photographs in [Figure 21\(a\)–\(c\)](#). Characteristic range of scales is H (cloud base) = 550–670 m msl, D_f (maximum funnel diameter) = 3–45 m, and a = 15–760 m.

temperature (and therefore high equivalent potential temperature) favor funnel formation. For the period 1958–68, no funnels were reported when air temperatures were below 71°F or dew point below 66°F. Waterspouts may form in midlatitude locations such as the Southern California coast and the Great Lakes during the cool season with surface temperatures and dewpoints much lower than the Keys; however, in those situations other forcing mechanisms such as upper lows or approaching surface fronts play a major role. We examined a rare outbreak of several waterspouts, one of them large and tornadic, over Lake Tahoe in the early morning of 28 September 1998. Even though surface water temperatures on the lake were no higher than 70°F, the air mass was extremely unstable due to the presence of an upper cold low over Central California.

Prevailing wind direction during June at the surface in Key West is south-easterly, with a gradual backing to east-south-easterly in October, according to the *Climatic Atlas of the United States*. In May, the waterspout season onset month, prevailing surface winds are east-south-easterly. Wind speeds of $3\text{--}5\text{ m s}^{-1}$ were recorded during 57% of all funnel events, a calm was reported once, and 16 kn was the highest speed observed during funnel events. These wind speed figures exclude gustiness caused by showers. The two most frequent wind directions actually observed during funnel events are from 120° and 080° (1964–68). Conditions favorable for

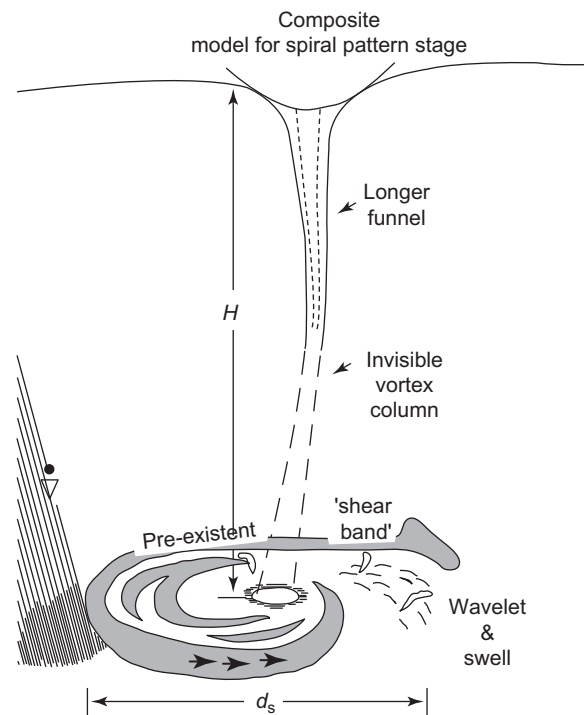


Figure 17 Composite schematic model for the spiral pattern (stage 2) in the waterspout life cycle. The vertical scale is contracted, H (cloud base height) varies from 550 to 670 m msl, and d_s varies from 150 to 920 m. Bold arrows in spiral pattern indicate that a major band evolves around the dark spot during this stage. Compare with actual case photographs in [Figure 21\(c\) and \(d\)](#).

waterspout formation occur regularly on successive days in the summer months over the Florida Keys. Two or more successive funnel days were recorded for 20 of 167 occurrences during the 10 year period, but 10 were reported during the 35 funnel-days of 1968. Consecutive-day occurrences suggest some special slow-moving type of synoptic flow regime, and this hypothesis was examined and confirmed later by the author.

What, then, are the synoptic and mesoscale conditions associated with waterspouts? The synoptic scale was found to be the controlling influence on waterspout-active convective cloudline developments in the Lower Keys. If strong anticyclonic conditions persist in the 1000 to 600 hPa layer ($\sim$ lowest 4 km), subsidence will inhibit cumulus-cloudline development over the heated islands and shallow water. Additionally, if a strong cyclonic synoptic-scale disturbance (e.g. tropical storm) affects the Keys, the strong low-level winds and vertical wind shear will disrupt the surface heating mechanism. Whenever either several waterspouts or giant, long-lived waterspouts were observed, a weak but well-defined synoptic-scale troughline in the lower tropospheric mean flow approached and passed through the Keys, causing enhanced cumulus-cloudline development.

The cloudline scale (mesoscale) is perhaps most crucial to waterspout formation. It is estimated from the 1969–72 Keys waterspout data that at least 90% of the waterspouts are spawned by rapidly building cumulus cloudlines (i.e., not isolated cumuli). It is significant that spiral patterns usually form on the flanking edge of a cloudline, with the developing

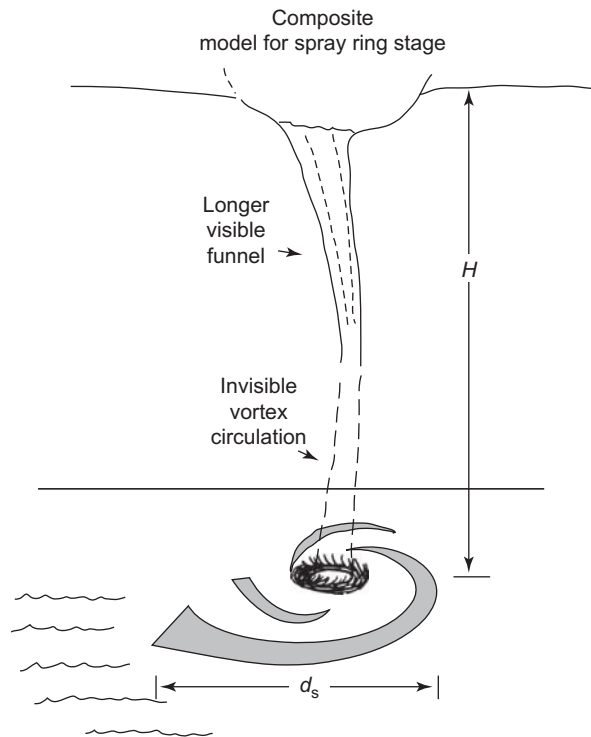


Figure 18 Composite schematic model for the spray ring (stage 3) in the waterspout life cycle. The vertical scale is contracted, H varies from 550 to 670 m msl, and d_s from 445 to 260 m. See text for more details. Compare with the actual case in [Figure 21\(e\)](#).

rain shower or two generally within a few kilometers of the epicenter. Frequently, the tail or major 'feeder band' of the spiral appears to emanate from a narrow protrusion of the downwind shower boundary. These features are summarized in [Figure 3](#). Generally, there are both cyclonic shear and curvature in the low-level mesoscale flow through the cloudline. The updrafts and subcloud-layer convergence are enhanced by converging shower outflows, and the waterspouts move in the directions of these outflows. While most of the Keys' cloudlines are generated by the low-level differential heating mechanism from the chain of islands and large surrounding envelope of shallow, warm water, others are associated with preexisting mesoscale boundaries.

Typical 'waterspout proximity soundings' at NWS Key West show that the low-level winds are light ($5\text{--}8\text{ m s}^{-1}$) from a south-west-south-east-north-east direction (most often an easterly component), and veer with height up to 600–700 hPa. The vertical wind shear implied by these soundings is generally weak, but it may be locally enhanced by outflows generated by the cloudlines themselves. There is usually abundant tropical moisture up to 700 hPa, with drier air above. Dry-adiabatic or superadiabatic lapse rates are often present below about 900 hPa. The remaining levels are convectively unstable, and lifted indices are near or below zero. No attempt has been made to construct 'mean waterspout soundings'. The reasons become clear upon examination of [Figure 13](#), showing spiral-ascent aircraft temperatures beneath and just outside a cloudline that produced seven sequential waterspouts. The Key West sounding, taken earlier that day some 20 km from the cloudline, is also

plotted. Note especially the observed superadiabatic lapse rate up to about 300 m msl. The outflow from a nearby shower apparently contributed to the greater instability beneath this cloudline (air and sea temperatures obtained during another research flight also indicated a superadiabatic lapse rate in the subcloud layer on another day). If these results hold in general, they strongly suggest that the low-level thermodynamic properties of the waterspout environment are poorly represented by conventional land-based soundings.

Waterspout Risks and Hazards

Is a waterspout merely any tornado over a water surface (as defined in the American Meteorological Society's revised *Glossary of Meteorology*) or is there some fundamental difference in structure and energetics? More to the point, most waterspouts are, indeed, nonsupercell tornadoes over a water surface. The primary risk from waterspouts is to residents, mariners, and structures in coastal areas and large inland lakes. The first documented death from a waterspout in the US was a wind-surfer along the Lake Michigan waterfront in Chicago in July 1993. There are some well-documented cases of large vessels being capsized or de-masted by large waterspouts, and some other reports of boats being swamped with deluges of heavy rain and/or sea spray.

Even though tornadic waterspouts appear to be quite rare, impressive damage may occur. The author noted that two eyewitnesses to the landfall of the second, intense waterspout at Lower Matecumbe Key, FL, claimed that a 1965 Cadillac weighing over 2 t was 'lifted a few feet off the ground and then set down again'; and what surely was a tornadic waterspout slammed into Venice, Italy, in the late evening of 4 September 1970. The whirlwind killed at least 18 persons when it lifted a crowded passenger motorboat from the water and sent it to the bottom of a lagoon in less than a minute. The wind picked up the 25 t boat, lifted it into the air, turned it around several times, and then plunged it back into the water. Most recently, the author and collaborators studied a tornadic waterspout that originated over south St Petersburg, FL, on 12 July 1995, producing F1 damage (\$200 000) and injuring one person before moving offshore and growing over Tampa Bay.

That waterspouts pose a serious threat to structures lying along their paths at landfall is beyond doubt, based on studies by the author, Fujita, and Macky. For example, Fujita and colleagues found that about one-fourth of the typhoon-associated tornadoes affecting the Japanese islands over a 22 year period *originated as intense waterspouts*. Damage statistics indicate that tornadoes of waterspout origin in Japan are stronger on the average than those originating over land. Many of the damaging tornadoes affecting the central and eastern Gulf Coast during the late fall and early spring originate over the northern Gulf of Mexico as intense waterspouts. For example, an intense waterspout made landfall on 7 February 1971 at 1618 local time and caused \$3 million in damage in Pensacola, FL. More recently, a very destructive outbreak of tornadoes in the Tampa Bay area on 3 October 1992 began when an intense tornadic waterspout moved onshore in Pinellas County. The author has shown that about 10% of the waterspouts each summer in the Keys reach

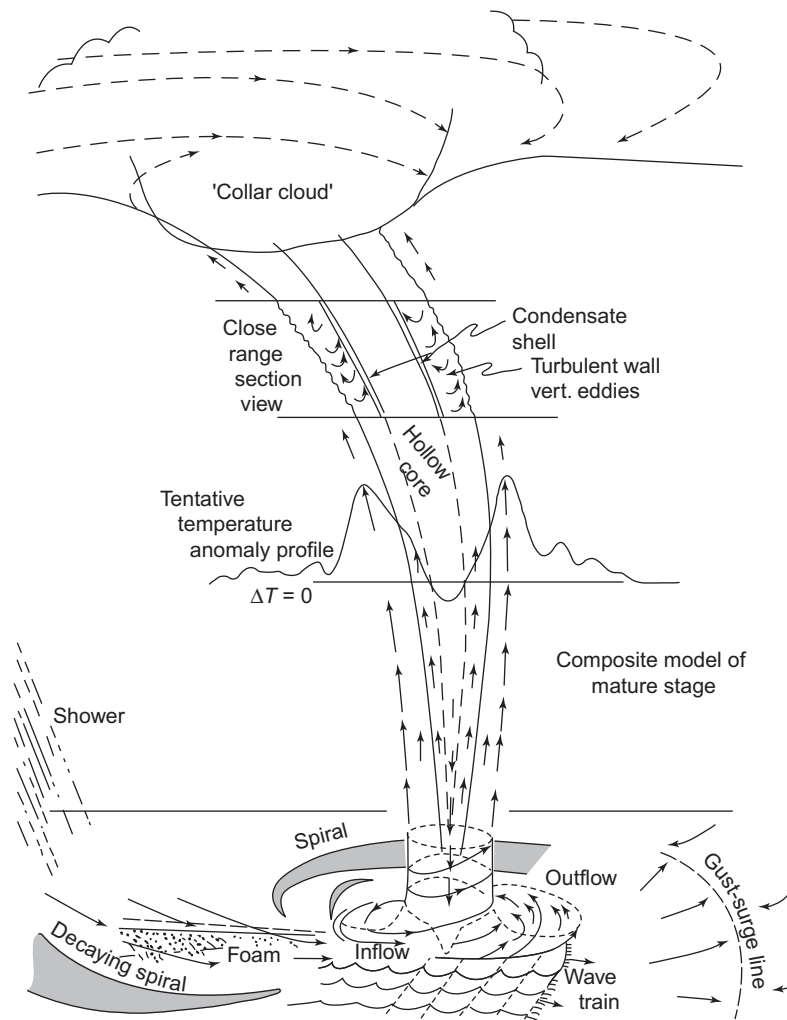


Figure 19 Composite schematic model of a mature waterspout/spray vortex (stage 4) in the waterspout life cycle. For scaling reference, the maximum funnel diameters in this stage, just below the ‘collar cloud’, range from 3 to 140 m. See text for more details on structural, flow and thermal features. Compare with actual case photographs in [Figure 21\(f\)–\(i\)](#).

tornadic size and intensity, with peak velocities in the spray vortex of 50 m s^{-1} or more.

Another hazard from waterspouts is public misconception about their own personal risk. Unfortunately, we have seen some mariners in recent years actually chasing waterspouts! A spectacular and foolish case of such a waterspout chase is shown in the aerial photograph in [Figure 14](#), which shows a large outboard motorboat chasing a fully developed intense waterspout and towing a water-skier at the same time. This illustrates a sad example of putting oneself (and others) directly in harm’s way.

Waterspout Kinematics and Dynamics

The chance aerial encounter with the three waterspouts near Lower Matecumbe Key, FL, on 2 September 1967 provided a unique opportunity to learn more about their structure and wind fields. Photogrammetry with the use of transparencies taken by the author revealed that the largest waterspout funnel ([Figure 2](#)) varied in diameter from 38 m near the parent cloud

base to 21 m at its lower end. Using the dimensions of the surface spray vortex obtained with the slides and then tracking and timing the rotation of spray plumes and particles at various radii on the aircraft movies, we obtained a tangential velocity profile through the spray vortex.

The resulting three-dimensional model synthesized from data on the two major waterspouts is shown in [Figure 15](#), which includes an inset showing the measured tangential speed profile across the spray vortex of the most intense waterspout. The reader should compare [Figure 15](#) with the earlier models depicted in [Figures 4–6](#).

The derived tangential wind speed profile through the waterspout’s spray vortex closely approximates that of a *Rankine-combined vortex*. With the ordinate axis as tangential wind speed (V_t) in meters per second and the abscissa as radius (R) in meters, we note that the maximum tangential wind speed was 65.0 m s^{-1} or 130 kn at a radius of just 12 m from the vortex center. In atmospheric dynamics, it is often assumed that the cyclostrophic wind is a valid approximation to the real wind in tropical cyclones in equatorial latitudes and in small-scale

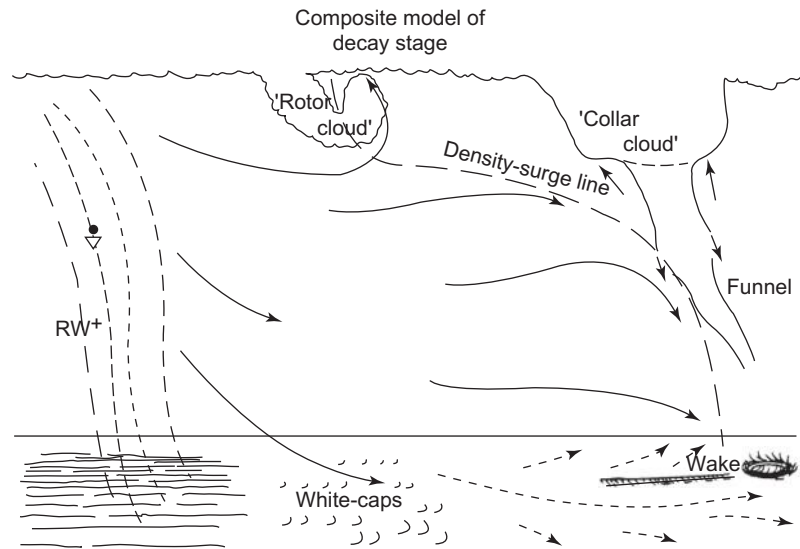


Figure 20 Composite model of decay (stage 5) in waterspout life cycle. Note that in some cases, the shower outflow line (labeled ‘density-surge line’) may be closely followed by the heavy rain shower. Also, the funnel cloud often undergoes rapid changes in shape and may become greatly contorted late in stage 5. For scaling purposes, maximum funnel diameters range from 3 to 105 m. Compare with actual case photographs in [Figure 21 \(j\), \(k\)](#).

vortices with very great wind speeds and path curvature. However, the data show that at times in a given waterspout’s life cycle, there may be horizontal accelerations of inflow into the spray vortex, especially from the right-rear quadrant relative to the direction of vortex motion. This feature would tend to obviate the possibility of a cyclostrophic balance in such cases.

The most intense waterspout documented by aircraft in 1969 was anticyclonic, and photogrammetric analysis of movies yielded peak tangential wind speeds in the spray vortex of 88 m s^{-1} at $R = 9 \text{ m}$. We also found that about 10% of the waterspouts each year in the Keys are *anticyclonic*, and a few of these may also be tornadic (peak rotational velocities in the spray vortex in excess of 50 m s^{-1}).

The Waterspout Life Cycle and Interacting Scales of Motion

The 1969 Lower Keys Waterspout Project and subsequent field programs over the intervening years reveal that all waterspouts undergo a regular life cycle composed of five discrete but overlapping stages: (1) the *dark spot stage*, characterized by a prominent light-colored disc on the sea surface, surrounded by a dark patch diffuse on its outer edges – the dark spot may or may not have a small funnel cloud above it initially, but signifies a complete vortex column extending from cloud base to sea surface; (2) the *spiral pattern stage*, the primary growth phase of the waterspout, characterized by the development of

Table 2 1969–70 statistical summary of documented stages in waterspout life cycle

Stage Major feature	1 Dark spot	2 Spiral	3 Spray ring	4 Spray vortex	5 Decay
1969: of 95	66 initially observed in stage 1	16 had spirals, and smoke flares showed circulation on this scale in several other cases	56 attained the sustained stage 3; short duration, transitional stage	51 reached this stage	51 decayed as heavy showers or cool outflow overtook spout; also 33 dark spots simply decayed by fading away
1970: of 33	All initially observed in stage 1, and 17 had an associated funnel at some time in the life cycle. $V_{\text{max}} = 10\text{--}15 \text{ m s}^{-1}$ in broad band just outside circular light-colored disc	Only 7 evolved to this stage and beyond; $d_s \sim 150\text{--}1000 \text{ m}$	$V_{\text{max}} \geq 22 \text{ m s}^{-1}$ in narrow band around dark spot periphery	86 had an associated funnel at some time in the life cycle; $V_{\text{max}} \geq 85 \text{ m s}^{-1}$ in sharply defined peak just outside ‘eye’ of spray vortex in bright band of concentrated spray	Some spiral rain curtains observed around decaying waterspouts
Duration range of stage (1969–70)	$t_1 = 1\text{--}22 \text{ min}$	$t_2 = 2\text{--}7 \text{ min}$	$t_3 = 1\text{--}2 \text{ min}$	$t_4 = 2\text{--}17 \text{ min}$	$t_5 = 1\text{--}3$; up to 7 min

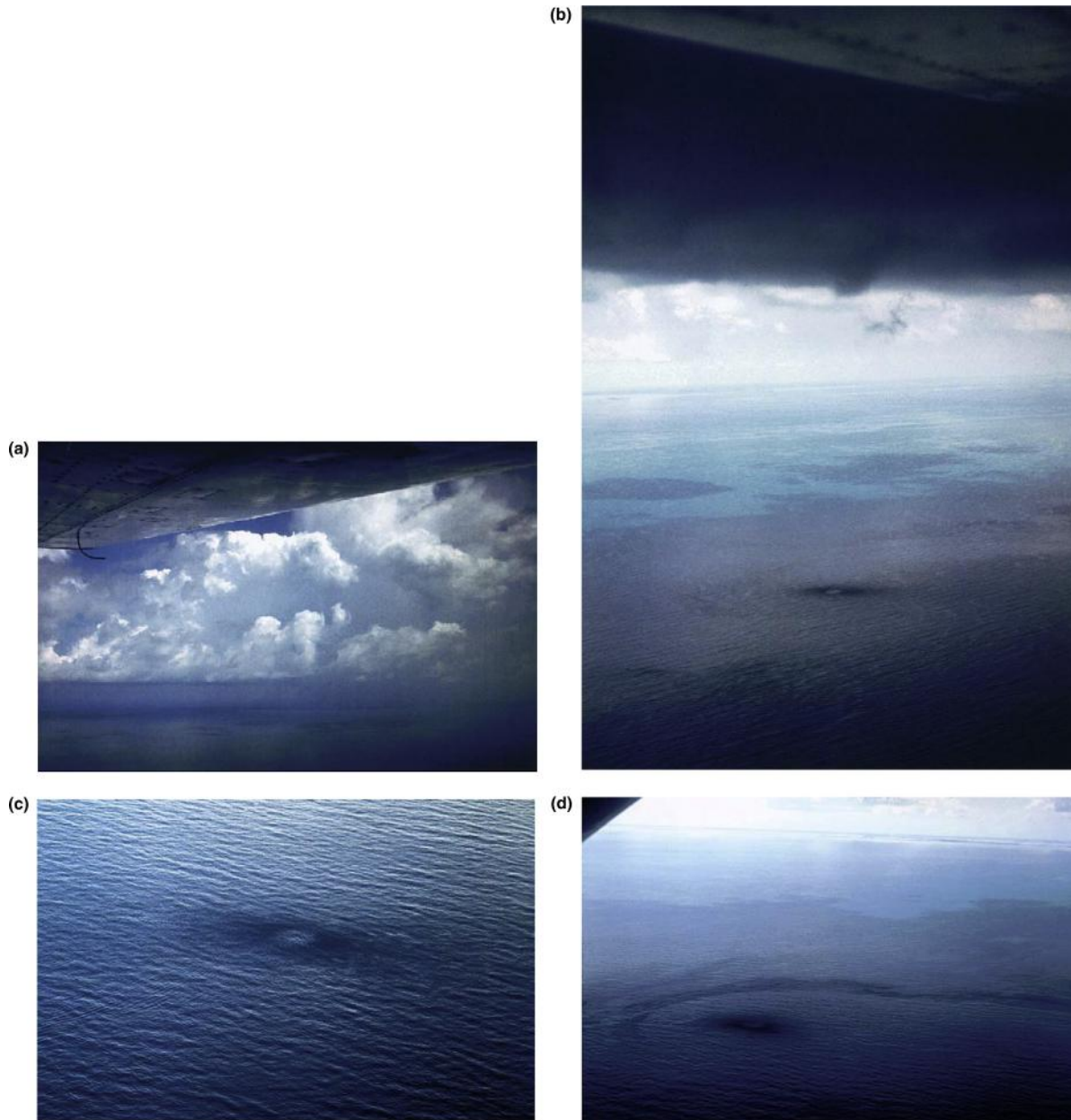


Figure 21 Sequence of photographs, by the author, of the complete life-cycle of a large, cyclonic waterspout in cumulus cloudline 20 km north of Key West, FL, in the late morning of 10 September 1969. This was the largest of nearly 100 waterspouts documented by research aircraft during the summer 1969 Lower Keys Waterspout Project. (a). Approach to building east–west oriented cumulus congestus cloudline (cloud base measured at 600 m). (b) Dark spot stage. Note the large dark spot, with a bright inner disc surrounded by a darker patch on the sea surface, and a large truncated funnel above. View from aircraft orbiting near cloud base, looking southwest with Keys in the background. Compare with Figure 16. (c) Dark spot stage – telephoto of dark spot. Note the small wind-driven waves approaching and apparently encroaching the dark spot from the left center. Water depths were 12–15 m, and only a horizontal cyclonic shear of about $\pm 15 \text{ m s}^{-1}$ could be detected across the 90 m diameter of this dark spot. Compare with Figure 16. (d) Late spiral-pattern stage, with incipient spray ring and elongation of dark ‘shear band’ around the eastern semicircle of the vortex, looking south at 1054 EST, 10 September 1969. Compare with Figure 17. (e) Preexistent shear band (far left) and newly developed spiral pattern at 1057 EST, looking west. Spray ring now fully developed and intensifying. Compare with Figure 18. (f) View of mature stage showing spray vortex, rope-like funnel and complete spiral pattern on sea surface. Looking NE at 1059 EST. (g) Looking SE at 1100 EST, with outer funnel wall developed downward from cloud base, and intense spray vortex on sea surface. (h) Large, double-walled waterspout funnel and spray vortex with trailing wake at 1103 EST. Compare with Figure 19. (i) Lower portion of the large waterspout, now with single-walled funnel, and spray sheath rising helically upward around spray vortex, with eye; looking ESE at 1105 EST. (j) Onset of decay stage, with protuberances moving around funnel walls and weakening spray vortex below; view towards south-west at 1107 EST. (k) Decay stage, with funnel narrowing and retracting, spiral pattern gone and spray vortex expanding with large eye. Note the smoke flare on the sea surface on the right and heavy showers in the background (waterspout moving away from showers, with cooler outflow intercepting waterspout). Compare with Figure 20.



Figure 21 (continued).

alternating dark- and light-colored bands spiraling around the dark spot on the sea surface; (3) the *spray ring (incipient spray vortex) stage*, characterized by a concentrated spray ring around the dark spot, with a lengthening funnel cloud above; (4) the *mature waterspout stage*, characterized by a spray vortex of maximum intensity and organization, the gradual weakening of the spiral pattern, and maximum funnel cloud length

and diameter; and (5) the *decay stage*, when the waterspout dissipates (often abruptly) as it is intercepted by the cool downdrafts from a nearby rain shower. The five stages of the waterspout life cycle are depicted conceptually in [Figures 16–20](#), respectively.

Not every waterspout observed from its inception evolved through all other stages; however, the combinations of stages



Figure 21 (continued).

1–2–3–4–5, 1–2–5, 1–3–4–5, and 1–5 have been documented. All waterspout cases documented by aircraft-based field programs during 1969–70 are summarized in [Table 2](#). Stages 1 and 4 have the greatest range of duration and often comprise the bulk of a waterspout's total lifetime. The largest and best-documented waterspout observed from aircraft during the 1969 Lower Keys Waterspout Project is shown in [Figure 21\(a\)–\(k\)](#). This sequence illustrates the evolution and key structural features of a complete waterspout life cycle.

Waterspouts' Relation to Other Convective Vortices

Other research and field projects by Peter Sinclair have shown many flow similarities between waterspouts and dust devils. The major difference seems to be one of scale and the fact that dust devils are associated with dry convection (no parent cloud system) in very hot climates. The high formation frequency and the life cycle of waterspouts in the Florida Keys result from energy and angular momentum exchanges among five scales of atmospheric circulation. These are: (a) the *funnel scale*, corresponding to the waterspout itself, with funnel diameters ranging from 3 to 150 m; (b) the new *spiral scale*, ranging from 150 m to 1 km on the sea surface; (3) the *individual cumulus-cloud scale*, ranging from less than 1 to up to 5 km in diameter; (d) the *cumulus-cloudline scale*, ranging from 5 to 100 km in

length; and (e) the *synoptic scale*, extending several hundreds of kilometers horizontally. The order of presentation of interacting scales does not imply the sense of energy or momentum transfer. In fact, a more detailed discussion of this scale-interaction process by the author suggests that the funnel scale is the end product of vorticity concentrated by larger-scale convergence fields.

The tornado life cycle resembles, in many respects, that typical of Florida Keys waterspouts. Both commence with surface evidence of vortex existence before a funnel cloud has descended a significant distance toward the surface. Approaching the mature stage, the tornado and waterspout exhibit spiral inflow characteristics with a distinct boundary between warm, moist air and cool, dry air. A schematic plan view of composited mean subcloud-layer mesoscale flow around the Union City tornado during its mature stage is given in [Figure 22](#). The cooler air mass from a nearby precipitation area apparently cuts off flow of warm, moist air into the tornado's circulation, leading to vortex decay. The visible funnel becomes thin, increasingly tilted and contorted as it dissipates. Major differences between the tornado and waterspout appear to be vortex and parent cloud scales and, to a lesser extent, vortex lifetimes and intensities. Both vortices may evolve rapidly through their respective life cycles without evolving through every stage. We note the strong resemblance, apart from scale, between the preexistent dark 'shear band' from which the

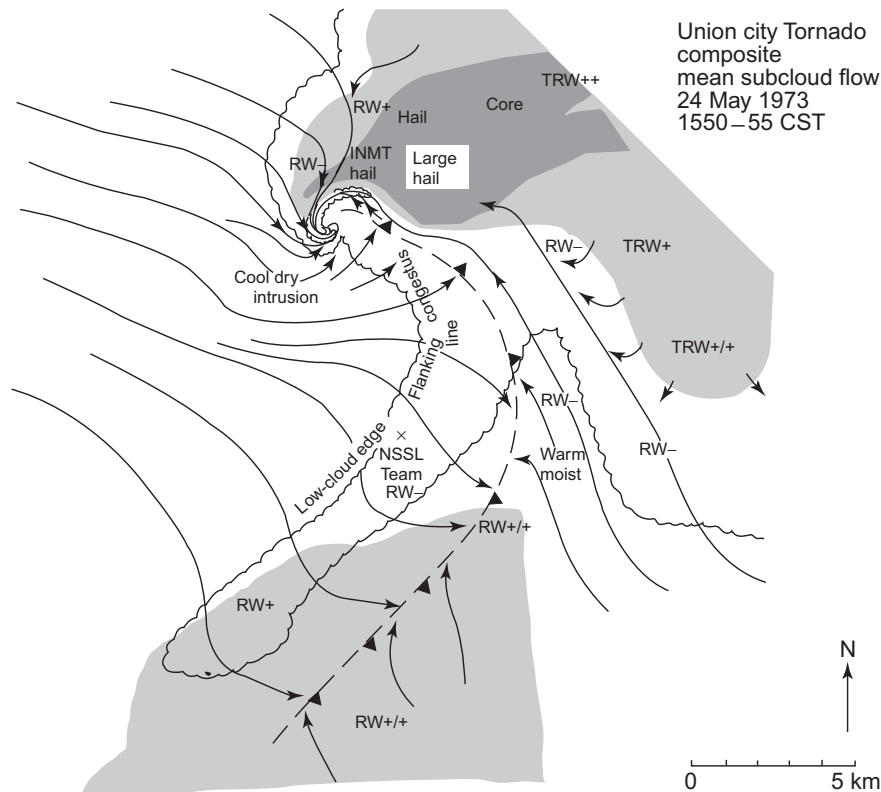


Figure 22 Composite schematic of mean subcloud flow around Union City, OK, tornado during its mature stage (1550–1555 CST), 24 May 1973. The position of the NSSL intercept team is plotted relative to the tornado ('T'), rain, and hail at Union City. RW: showers; TRW: thundershowers. Main parent storm cloud boundaries are scalloped.

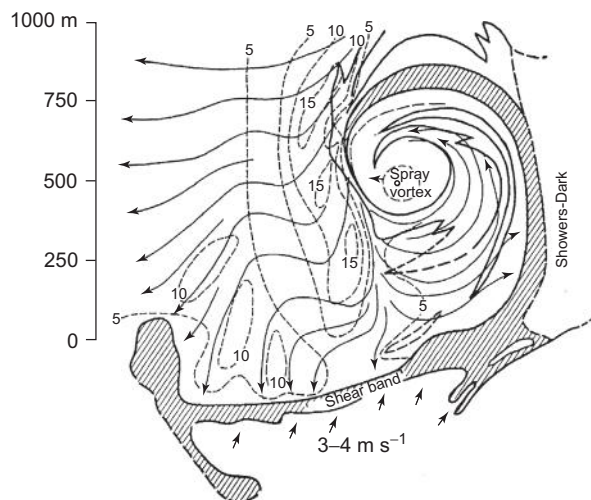


Figure 23 For the waterspout shown in **Figure 21(a)–(k)**, composite streamline (solid) and isotach (dashed in m s^{-1}) analysis of boundary layer flow on spiral scale around a mature waterspout on 10 September 1969, derived from the analysis of smoke plumes and photogrammetry of spiral features. During the period of compositing of marine flare data, we have assumed that the boundary layer circulation of the waterspout is a permanent-type system, i.e., quasi-steady state, as the spray vortex advances toward the smoke flares. At each instant of time, therefore, a streakline approximates a relative streamline.

waterspout's spiral pattern evolved (Figure 23, see also Figure 21 (d),(e),(f)) and the flanking cumulus cloudline in Figure 22 which spirals into the rotating 'wall cloud'. Both the flanking line and waterspout shear band appear to be the demarcation of gust fronts from nearby precipitation, and both signify pronounced low-level wind discontinuities with large cyclonic shear (vorticity). A recent study was made by the author and National Weather Service collaborators of a large waterspout that originated over land as a destructive tornado in St Petersburg, FL, and then tracked south-eastward over Tampa Bay with a 30 min lifetime. It was found that the tornadic waterspout's parent storm developed rapidly in response to surface mesoscale boundary interactions.

See also: **Air Sea Interactions:** Momentum, Heat, and Vapor Fluxes; Sea Surface Temperature; Surface Waves. **Aviation Meteorology:** Aviation Weather Hazards. **Boundary Layer (Atmospheric) and Air Pollution:** Convective Boundary Layer. **Dynamical Meteorology:** Vorticity. **Mesoscale Meteorology:** Cloud and Precipitation Bands; Severe Storms. **Tropical Cyclones and Hurricanes:** Overview and Theory. **Tropical Meteorology and Climate:** Equatorial Waves; Intertropical Convergence Zone; Intraseasonal Oscillation (Madden–Julian Oscillation); Tropical Climates.

Further Reading

- Bluestein, H.B., Golden, J.H., 1993. Review of tornado observations. AGU Geophysical Monograph. In: *The Tornado: Its Structure, Dynamics, Prediction and Hazards*, vol. 79. American Geophysical Union, Washington, DC.
- Brady, R.H., Szoke, E.J., 1989. A case study of nonmesocyclone tornado development in Northeast Colorado: similarities to waterspout formation. *Monthly Weather Review* 117, 843–856.
- Collins, W.G., Paxton, C.H., Golden, J.H., 2000. The 12 July 1995 Pinellas County, Florida, tornado/waterspout. *Weather and Forecasting* 15, 122–134.
- Golden, J.H., 1971. Tornadoes and waterspouts over South Florida. *Monthly Weather Review* 99, 146–154.
- Golden, J.H., 1973. Some statistical aspects of waterspout formation. *Weatherwise* 26, 108–117.
- Golden, J.H., 1974a. The life-cycle of Florida Keys' waterspouts, I. *Journal of Applied Meteorology* 13, 676–692.
- Golden, J.H., 1974b. The life-cycle of Florida Keys' waterspouts, II. *Journal of Applied Meteorology* 13, 693–709.
- Golden, J.H., 1974c. Life-cycle of Florida Keys' Waterspouts. NOAA Technical Memo, ERL-NSSL-70 (available from NTIS or NSSL, Norman, OK).
- Golden, J.H., 1977. An assessment of waterspout frequencies along the US East and Gulf Coasts. *Journal of Applied Meteorology* 16, 231–236.
- Golden, J.H., 1999. Wild waterspouts over Lake Tahoe. *Weatherwise* 52 (5), 14–19.
- Golden, J.H., Bluestein, H.B., 1994. The NOAA-National Geographic Society Waterspout Expedition (1993). *Bulletin of the American Meteorological Society* 75 (12), 2281–2288. Interested parties can also buy/rent National Geographic Society video production 'CYCLONE!' which has excellent waterspout sequence.
- Golden, J.H., Purcell, D., 1978. Life-cycle of the Union City, OK tornado and comparison with waterspouts. *Monthly Weather Review* 106, 3–11.
- Leverson, V.H., Sinclair, P.C., Golden, J.H., 1977. Waterspout wind, temperature, and pressure structure deduced from aircraft measurements. *Monthly Weather Review* 105 (6), 725–733.
- Simpson, J., Morton, B.R., McCumber, M.C., Penc, R.S., 1986. Observations and mechanisms of GATE waterspouts. *Journal of the Atmospheric Sciences* 43, 753–782.
- Wakimoto, R.M., Lew, J.K., 1993. Observations of Florida waterspouts during Ca PE. *Weather and Forecasting* 8, 412–423.

Bow Echoes and Derecho

ML Weisman, National Center for Atmospheric Research, Boulder, CO, USA

© 2003 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 1, pp 311–312, © 2003, Elsevier Ltd.

Introduction

Windstorms produced by complexes of convective storms (thunderstorms) pose a significant hazard to life and property in many places of the world, especially during the spring and summer months. The largest and most long-lived of these events have been given a generic name of 'derecho,' a term that originated in the late 1800s to refer to convective systems producing wide and long swaths of straight-line wind damage. Detailed studies of convective wind events, however, have shown that a vast majority are associated with a particular type of organized convective system, more popularly referred to as a 'bow echo.' This article describes the basic structures and environments associated with bow echoes and derechos, and further highlights some of the recent research that clarifies the mechanisms critical to their development and maintenance.

Bow echoes and derechos form a subset under the more general heading of mesoscale convective systems, which include squall lines, mesoscale convective complexes, and the like. In all of these cases, the system is envisioned to be composed of a sequence of relatively independent convective cells that contribute collectively to a larger system-scale structure. The individual convective cells can be ordinary cells, multicells, or supercells, as described elsewhere in the encyclopedia (see *Mesoscale Meteorology: Convective Storms: Overview*). In the following, the system-scale attributes that have led to these particular systems being identified for their unique form of mesoconvective organization are emphasized.

Bow Echoes

Bow echoes, originally referred to as a line echo wave pattern (LEWP), are most readily identified by a persistent bow shape

on a radar screen, and have become especially associated with the production of long, narrow swaths of damaging surface winds. Much of what is known observationally concerning bow echoes originated with Dr TT Fujita, who spent much of his career trying to characterize and understand the types of convective systems most apt to produce severe weather such as downbursts and microbursts.

A typical evolution and morphology of radar echoes associated with a severe bow echo is presented in [Figure 1](#). The system usually begins as a strong convective cell or a short line of convective cells that may be either isolated or embedded within a more extensive squall line. As the strong surface winds develop, the initial cell evolves into a bow-shaped line segment of cells, with the strongest winds occurring near the apex of the bow. Cyclonic and anticyclonic motions of the radar cells are often noted on the northern and southern ends of the bow segment, respectively. During the declining stage, the system often evolves into a comma-shaped echo with predominantly cyclonic motion of the radar echoes evident on the northern end of the system ([Figure 1\(e\)](#)).

At the core of most bow echoes is a strong surface cold pool and associated surface mesohigh, which is produced via the evaporation of rain and transport of potentially colder air to the surface from the mid-troposphere (e.g., 3–5 km above ground level (AGL)). Cold pools within severe bow echoes can be as much as 10–15 °C colder than the surrounding air mass, with surface pressure excesses reaching 5–8 mbar. In association with this cold pool is also often an intense rear-inflow jet that extends rearwards from the leading edge of the active convection and may extend in depth from 3 to 5 km AGL down to the surface. This rear-inflow jet helps transport drier, mid-level air into the precipitating region behind the leading edge of the convection, increasing the potential for strong,

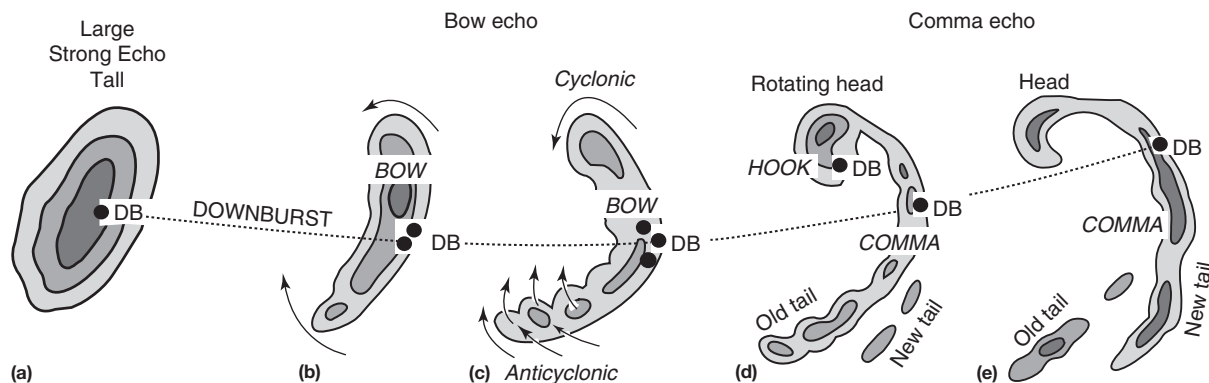


Figure 1 A typical morphology of radar echoes associated with bow echoes that produce strong and extensive downbursts, labeled DB on the figure. Reproduced from Fujita, T.T., 1978. Manual of downburst identification for project Nimrod. Satellite and Mesometeorology Research Paper No. 156. Department of Geophysical Sciences, University of Chicago.

evaporationally produced downdrafts and resulting cold pools. Such rear-inflow jets can reach magnitudes of 25 m s^{-1} or greater above the ambient mid-level flow. The cold pool and rear-inflow jet represent the primary source for the strong surface winds in such systems.

Bow echoes are observed over a wide range of scales, from tens of kilometers to over 200 km (along line length) in extreme cases. Lifetimes can range from a couple of hours to over 10 h. Although a range of scales is observed, the most intense bow echoes tend to be 40–120 km in length and have lifetimes of 4–6 h. Widespread surface winds of 25 m s^{-1} or greater are commonly observed with severe bow echoes, with extreme cases producing swaths of damaging winds of greater than 50 m s^{-1} , producing widespread falls of trees, toppling power poles, damaging buildings, and blowing vehicles off highways.

An example of bow echo that passed through Springfield, Illinois, on 6 August 1977 is presented in Figure 2. This system evolved from a relatively isolated cell to a comma-shaped echo over a 5 h period while producing a continuous swath of damaging surface winds over a 200 km path. A detailed damage survey taken during a portion of this event is presented in Figure 3, and demonstrates that the broad swath of outflow winds is often made up of a series of individual downdrafts and microbursts (see *Mesoscale Meteorology*: Microbursts). It also demonstrates the tendency of bow echoes to generate tornadoes, especially along and to the north of the apex of the bow (Figure 2). In the present case, 18 tornadoes were generated just to the north of the bow echo apex as the northernmost cell evolved into a cyclonically rotating head. The relationship between bow echoes and tornadoes has still not been adequately explained.

A wide range of radar echo configurations can be associated with developing severe bow echoes (Figure 4). One of the

common characteristics is the development of a strong low-level reflectivity gradient near the leading edge of the concave-shaped echo. Another significant feature is the presence of a weak echo channel or 'rear inflow notch' (RIN) on the backside of the bow, which often signifies the location of the intense rear-inflow jet and the likelihood of downdraft winds and possible downdraft-induced tornadoes. Also, while a bow echo is generally organized on a scale larger than a single convective cell, individual severe cells, sometimes supercellular, may be contained within the larger scale structure (e.g., note the 'intense convective cell' with the tight reflectivity gradient and hooklike appendage at the southern end of the bow echo in Figure 4).

Another example of a mature bow echo is shown in Figure 5 from 5 May 1996 near Paducah, Kentucky. In this example, a large bow-shaped convective system has two smaller scale bows embedded with the larger circulation. The Doppler winds clearly depict a large rear-inflow jet behind the core of the system (dark blue), with weak anticyclonic shear to the south of the bow and stronger cyclonic shear evident on the northern end of the bow. Additionally, the smaller embedded bows each have their own localized rear-inflow jets with associated rotational features on the ends. This event emphasizes that a range of bow-echo scales can exist, sometimes side by side, in the same basic environment. A vertical cross section of reflectivity taken through the core of the bow (Figure 6(a)) depicts strong, upright convective cells at the leading edge, with a weaker stratiform region extending rearward. A storm relative velocity cross section (Figure 6(b)) depicts front-to-rear ascending flow through the convective cells and extending aloft within the anvil, with a strong rear-inflow jet beneath the front-to-rear flow at mid-levels.

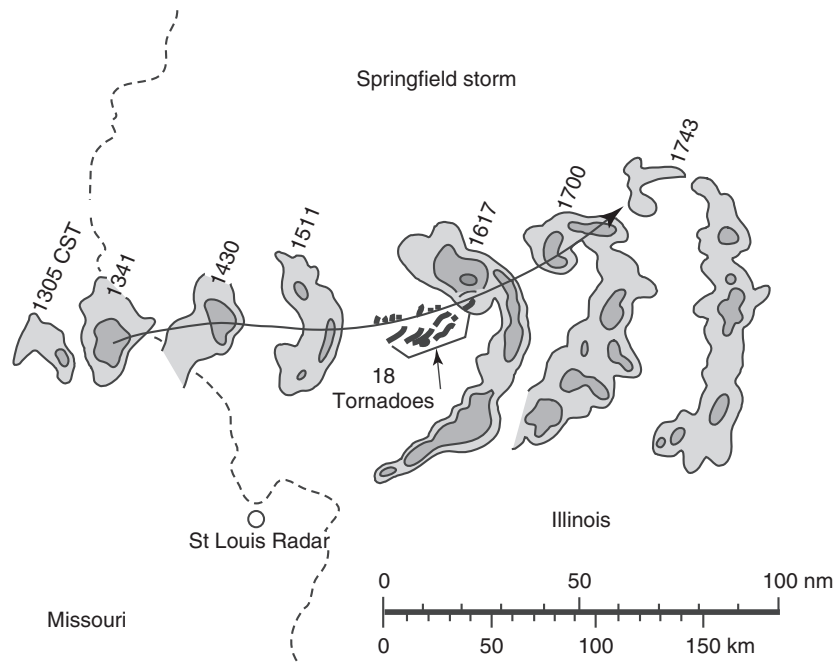


Figure 2 Evolution of radar echoes associated with Springfield downbursts and tornadoes of 6 August 1977. Reproduced from Fujita, T.T., 1978. Manual of downdraft identification for project Nimrod. Satellite and Mesometeorology Research Paper No. 156. Department of Geophysical Sciences, University of Chicago.

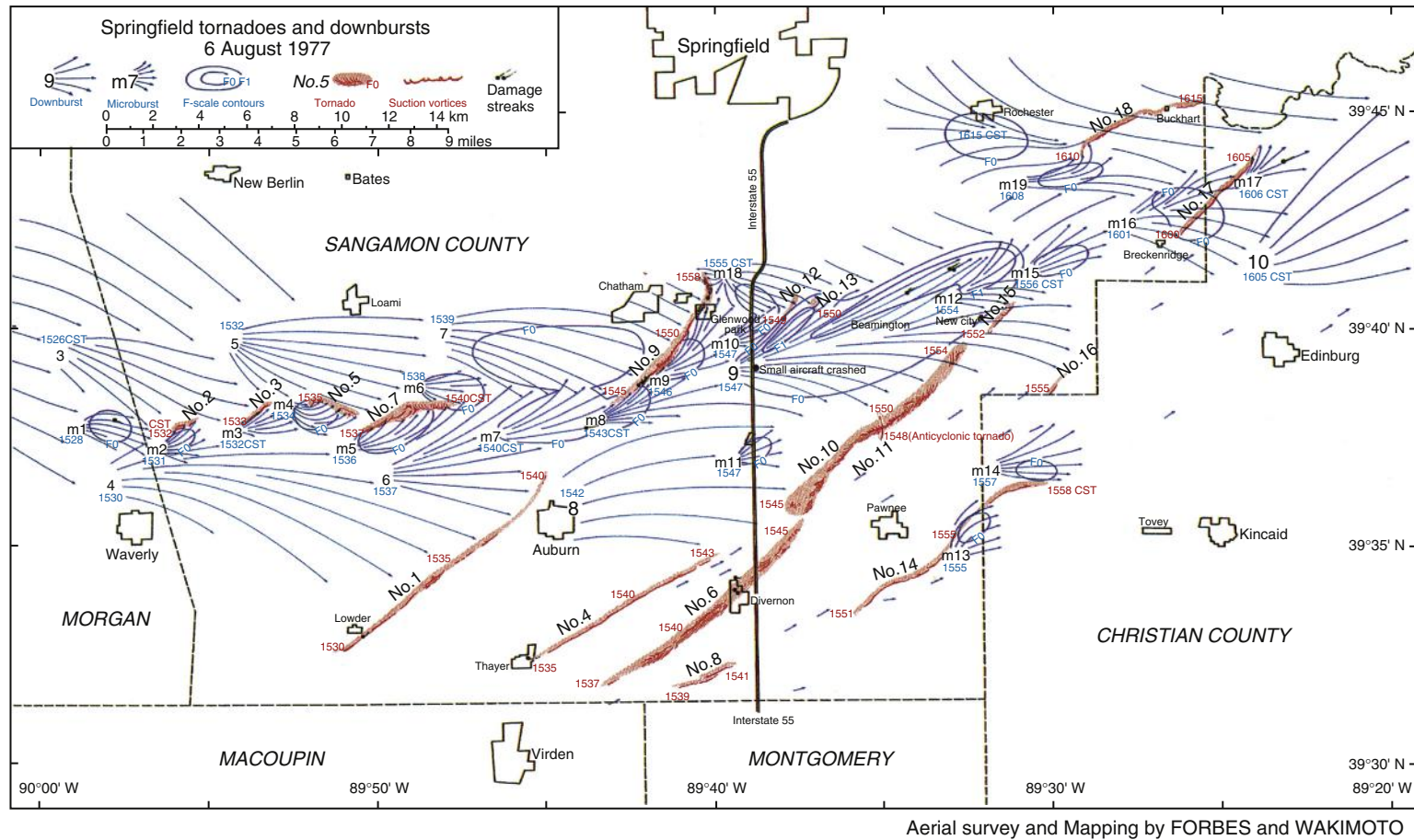


Figure 3 Eighteen tornadoes, 10 downbursts, and 17 microbursts are depicted in this map. One tornado (No. 11) was anticyclonic. Apparently, eight tornadoes formed on the left side of microbursts. No trace of downbursts was found in the vicinity of other tornadoes. Reproduced from Fujita, T.T., 1978. Manual of downburst identification for project Nimrod. Satellite and Mesometeorology Research Paper No. 156. Department of Geophysical Sciences, University of Chicago.

Derechoes

The term 'derecho' is used to describe convective systems that produce straight-line convective wind gusts greater than 26 m s^{-1} within a concentrated area with a major axis length of at least 400 km. The gusts must also show a systematic pattern of progression, with no more than 3 h elapsing between successive wind damage events. Such systems have been observed to have lifetimes of as long as 18 h, producing a swath of damaging winds hundreds of kilometers wide and 1000 km long. An example of the extent and longevity of a derecho event is presented in Figure 7. In this case a squall line produced damaging wind over a swath hundreds of kilometers wide and 1000 km long over 18 h.

Two basic patterns of radar cells are often associated with a derecho (Figure 8). The first pattern (referred to as a progressive derecho) consists of a single bowed segment of convective cells that often develop just on the cool side of a weak stationary front. The bowed feature moves parallel to the front. The second pattern (referred to as a serial derecho) consists of a longer squall line that has evolved into a series of bow echoes or LEWPs that propagate along the squall line.

While bow echoes represent one of the primary convective structures associated with derecho events, severe bow echoes occur much more frequently than derechos, producing damaging surface winds over more limited regions and for shorter time periods than included in the strict definition of a derecho.

Bow Echo and Derecho Environments and Climatology

Bow echo and derecho environments are generally characterized by large amounts of convective instability and low-level moisture. Surface dew points are commonly greater than 20°C with lifted indices averaging about -9°C . Convective available potential energy (CAPE) is generally greater than 2000 J kg^{-1} , with many cases exhibiting CAPEs greater than 4000 J kg^{-1} . Such large CAPEs support both the development of strong convective updrafts as well as strong convective downdrafts and cold pools, the latter being especially critical for the development of strong surface outflow. The development of strong cold pools can also be supported by dry mid-tropospheric conditions, although bow echoes and derechos are observed for both moist and dry mid-level conditions as long as sufficient

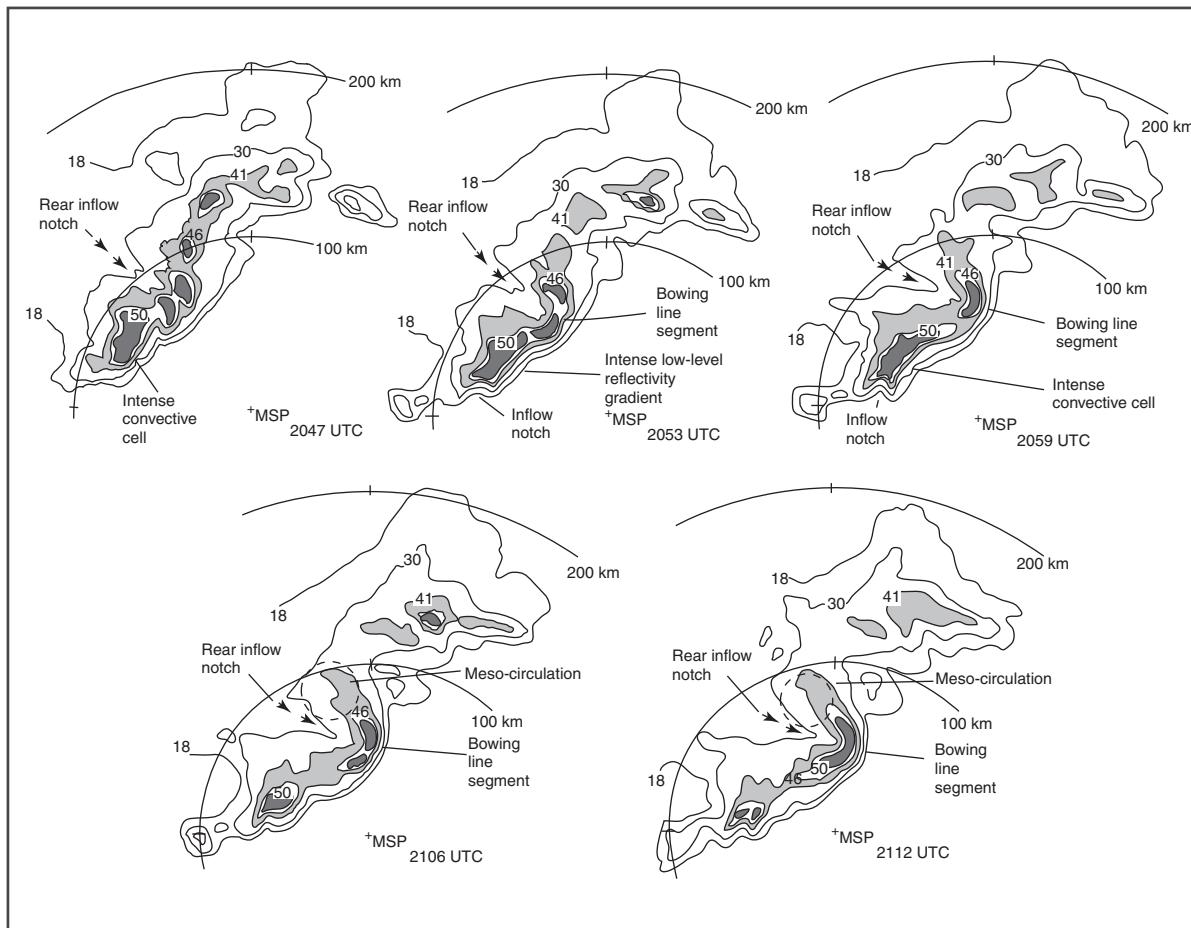


Figure 4 Radar analysis of the central Minnesota derecho between 2047 and 2112 UTC from Minneapolis-St Paul, Minnesota (MSP). Reflectivity contours are 18, 30, 41, and 46 dBZ. Shaded region represents reflectivity values greater than 50 dBZ. Reproduced from Przybylinski, R.W., 1995. The bow echo: Observations, numerical simulations, and severe weather detection methods. *Weather and Forecasting* 10, 203–218.

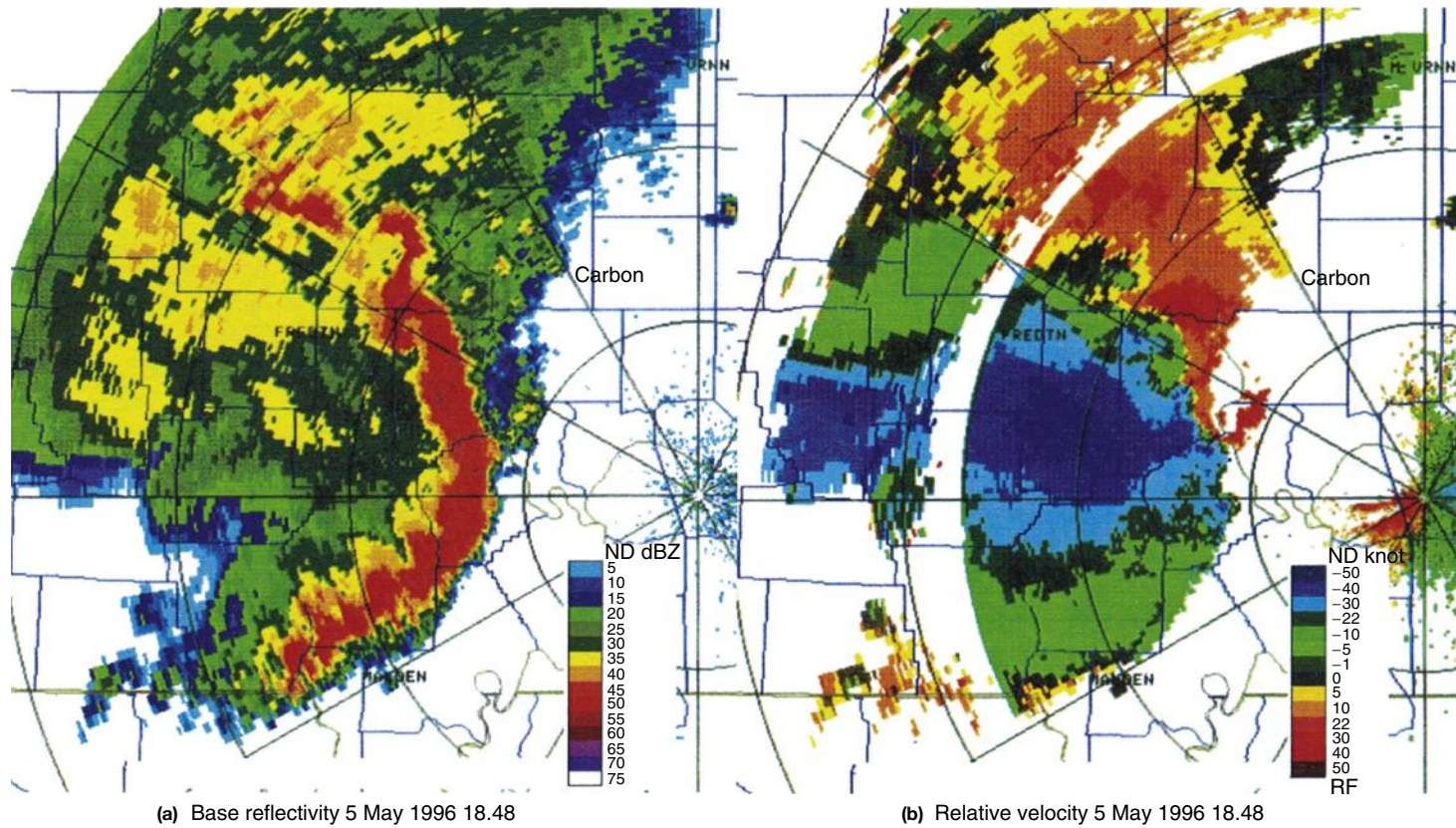


Figure 5 (a) Base reflectivity and (b) relative velocity from the Paducah WSR-88D radar at 18.48 GMT for 5 May 1996. Velocities are presented relative to a storm motion of 33 knots (17 m s^{-1}) from 280° . On the scales in the lower left of each figure, ND indicates no data and RF indicates range folding. For (b), blue colors represent flow toward the radar, while red colors represent flow away from the radar. Przybylinski, R.W., personal communication.

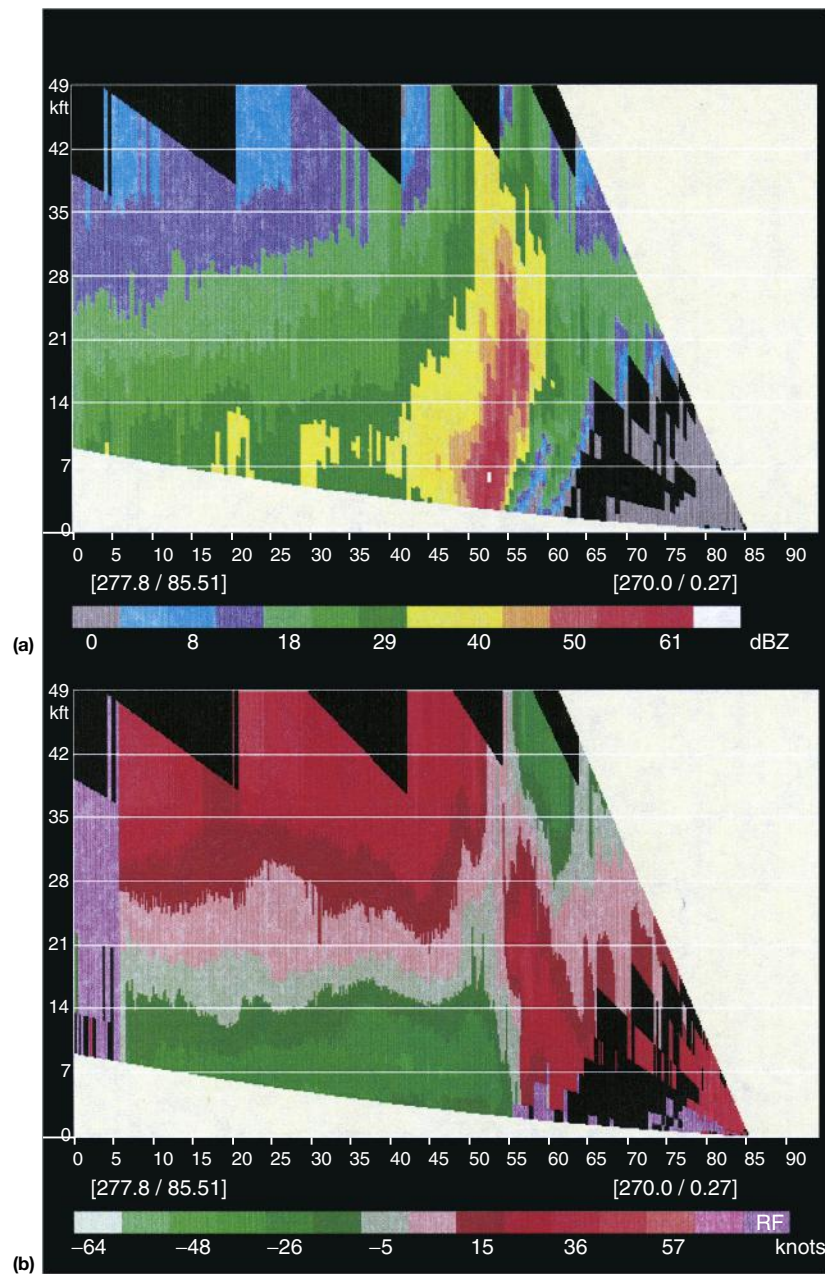


Figure 6 Vertical cross sections of (a) reflectivity and (b) storm-relative at 1855 UTC for the Paducah, Kentucky bow echo. The vertical cross sections are taken at a 277° heading from KPAH. Velocities are presented relative to a storm motion of 20 knots (10 m s^{-1}) from 277° . For (b), green colors represent flow toward the radar, while red colors represent flow away from the radar. Przybylinski, R.W., personal communication.

CAPE is available. Environmental vertical wind shear magnitudes tend to be in the moderate range for severe convective events, with about 15 m s^{-1} of shear evident between the surface and 700 mbar (roughly 0–3 km AGL), and about 20 m s^{-1} between the surface and 500 mbar (roughly 0–6 km AGL). Such shear magnitudes are generally weaker than are associated with supercell storms, although bow echoes and derechos are observed in these more strongly sheared environments as well.

Bow echoes and derechos can occur in environments with strong synoptic-scale forcing, as with severe, prefrontal squall

lines, but also occur quite often in association with more benign synoptic patterns. As presented in Figure 9, many events begin along or to the north of a weak east-to-west oriented quasistationary frontal boundary, in the vicinity of a mid-tropospheric ridge, and then move along the boundary (as in the progressive bow echo in Figure 8). The existence of a low-level jet (LJ) impinging from the south and flowing along the frontal region, along with the polar jet (PJ) oriented parallel to the front farther to the north, leads to moderate magnitudes of vertical wind shear in the lower and mid-troposphere. The addition of enhanced low-level convergence along the zone, to

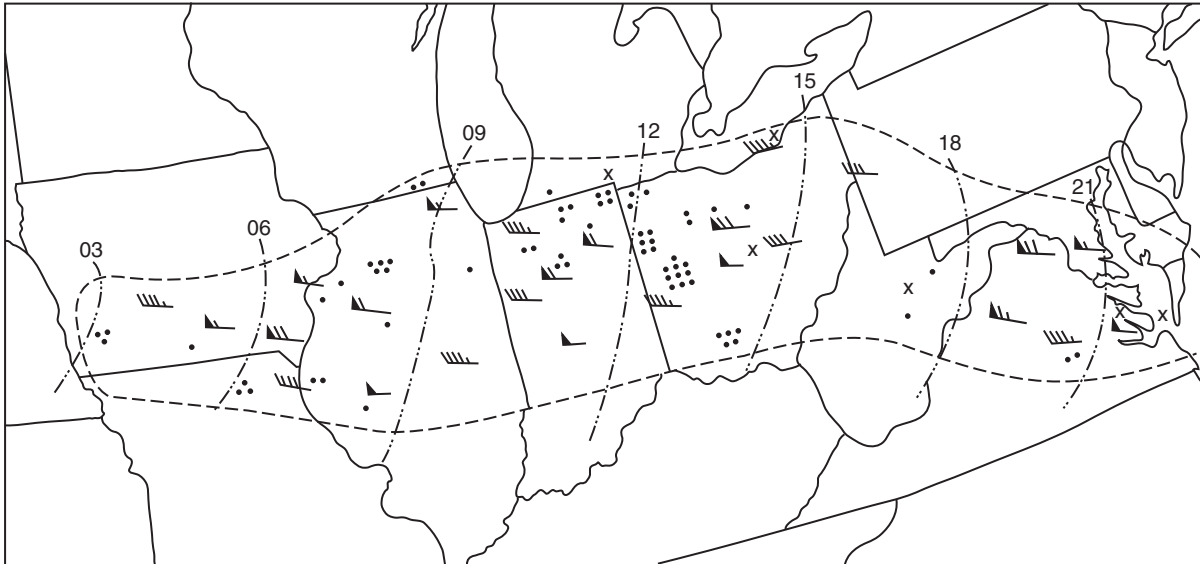


Figure 7 Area affected by the convective windstorm of 5 July 1980 (dashed line). Three-hourly squall line positions are indicated in UTC (from 03.00 to 21.00 on 5 July). Officially measured convective gusts are indicated by wind barbs (full barb signifies 5 m s^{-1} , flag signifies 25 m s^{-1}). Personal injuries (67) are indicated by dots, and each death (6) is shown by an 'x'. Reproduced from Johns, R.H., Hirt, W.D., 1987. Derechos: Widespread convectively induced windstorms. *Weather and Forecasting* 2, 32–49.

help initiate the convection, along with the associated deepening of the moist layer, to help increase the thermodynamic instability, fills out the ingredient list that promotes the development of bow echoes and derechos for this type of weather pattern.

Figure 10 shows the paths of 67 well-defined derecho events over the USA during the period of 1983–93. A primary corridor of derecho activity is evident over the upper mid-western states, with secondary corridors along an axis from Kansas through Oklahoma and Texas, and also in the southeast. As presented in Figure 11, derechos are most frequent during the spring and summer months over the USA, but can occur at almost any time of the year. A similar climatology is believed to exist for the even more frequent smaller scale, shorter lived bow echoes, although a specific study documenting bow echo occurrence has yet to be undertaken.

Numerical and Dynamical Studies

The tendency for a convective cell to evolve into a bow-shaped system of cells for certain environments is readily reproduced in numerical cloud modeling studies. Fundamentally, an updraft produces rain that falls and evaporates, thereby producing a pool of cold air that spreads along the ground. This spreading cold pool produces convergence and lifting along its leading edge that can then trigger new cells. However, rather than a cold pool producing a complete circle of new cells around the initial storm, cells are favored along a bow-shaped arc oriented perpendicular to the vertical wind shear vector. The ability to trigger new cells along this arc increases dramatically as the amount of vertical wind shear increases, and also if the wind shear is confined to the lowest 2–5 km AGL.

Figure 12 demonstrates these results for numerical model simulations of bow echoes, with and without the effects of the Earth's rotation (Coriolis forcing) included. These simulations are initiated with five convective cells along a line 150 km in length in a horizontally homogeneous environment with moderate CAPE (2200 J kg^{-1}) and strong low-level vertical wind shear (20 m s^{-1} over the lowest 2.5 km AGL; winds are kept constant above 2.5 km). In both cases, a line of strong convective cells has become established by 3 h. For the non-Coriolis case (Figure 12(a)), this line becomes significantly bow-shaped between 3 and 6 h, with strong mirror image cyclonic and anticyclonic vortices developing at mid-levels behind the northern and southern ends of the system, respectively. With Coriolis forcing added (Figure 12(b)), the northern cyclonic line-end vortex strengthens over time, while the southern anticyclonic vortex weakens, leading to a highly asymmetric system configuration by 6 h. The strengthening of the northern cyclonic vortex is directly related to the mid-level convergence of planetary rotation. Figure 12(b) also demonstrates that a range of bow-echo scales can be produced within the same convective system, very similar to the observations of the Paducah bow echo case from 5 May 1996 (Figure 5). The tendency to develop such subsystem-scale vortices within such simulations increases with increasing magnitudes and depth of the ambient vertical wind shear.

Much of the strength and structural characteristics of bow echo-type systems can be understood by considering the development of a two-dimensional circulation along a vertical cross section through the core of the convective line. This is most easily accomplished via the two-dimensional horizontal vorticity equation for inviscid, Boussinesq flow (eqn [1]), where $\eta = \partial u / \partial z - \partial w / \partial x$ and B represents the buoyancy, defined by eqn [2].

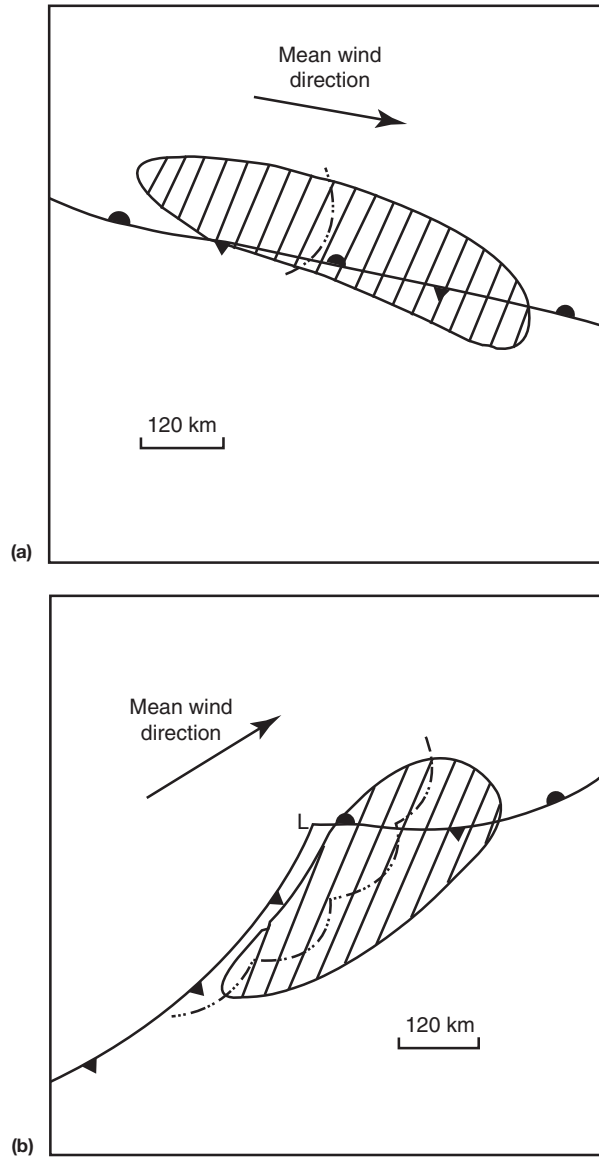


Figure 8 Schematic representation of features associated with (a) progressive and (b) serial derechos near the midpoint of their lifetimes. The total area affected by these derechos is indicated by the hatching. The frontal and squall line symbols are conventional. Reproduced from Johns, R.H., Hirt, W.D., 1987. Derechos: Widespread convectively induced windstorms. *Weather and Forecasting* 2, 32–49.

$$\frac{d\eta}{dt} = -\frac{\partial B}{\partial x} \quad [1]$$

$$B \equiv g \left[\frac{\theta'}{\bar{\theta}} + 0.61(q_v - \bar{q}_v) - q_c - q_r \right] \quad [2]$$

Here θ represents the potential temperature, q_v , q_c , and q_r represent the mixing ratio of water vapor, cloud water, and rainwater, respectively, and g is the acceleration due to gravity. Within this framework, the only source of horizontal vorticity other than the ambient vertical wind shear (e.g.,

$\eta = \partial U(z)/\partial z$, where $U(z)$ represents the ambient wind profile) is horizontal gradient of buoyancy. Thus, the analysis of the development of circulation is simplified for understanding the evolution of the buoyancy field and its interaction with the ambient shear. The two primary sources of buoyancy within a convective system are the warm convective updraft and the cold convective downdraft and cold pool. The net circulation of the system thus depends on the relative strengths of the ambient shear, the convective updrafts, and the cold pool.

From the perspective of eqn [1], a convective system initially leans in the direction of the ambient vertical wind shear vector (e.g., downshear) as the warm convective updraft feels the influence of the sheared flow (Figure 13(a)). However, as the surface cold pool develops and strengthens over time, the opposing circulation associated with the cold pool forces the system to achieve a more upright (Figure 13(b)) and then upshear-tilted (Figure 13(c)) configuration. Once the system begins to tilt upshear, a rear-inflow jet is generated in response to the buoyant front-to-rear ascending current aloft and rearward spreading cold pool at the surface. For most convective systems, this rear-inflow jet descends and spreads along the surface well behind the leading edge of the convection, enhancing the surface outflow but generally weakening the convective system. For the stronger shear, large CAPE bow echoes, however, this rear-inflow jet may remain elevated, enhancing the lifting at the leading edge of the system, and promoting an even stronger, and more long-lived convective system (Figure 13(d)). This configuration of the vertical circulation and elevated rear-inflow jet is quite similar to the 5 May 1996 case, as presented in Figure 6.

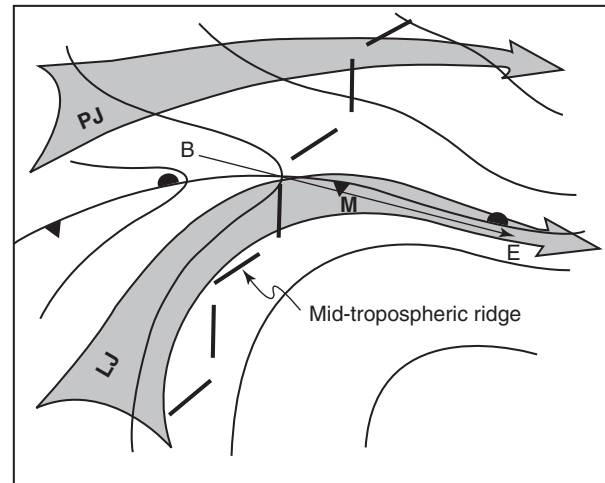


Figure 9 Idealized sketch of a mid-latitude warm season synoptic-scale situation especially favorable for development of long-lived progressive bow echo complexes producing extensive swaths of damaging winds. The line B–M–E represents the track of the bow echo complex. Thin lines denote sea level isobars in the vicinity of a quasistationary frontal boundary. Broad arrows represent LJ stream and PJ in the upper troposphere. Reproduced from Johns, R.H., Doswell III, C.A., 1992. Severe local storm forecasting. *Weather and Forecasting* 7, 588–612.

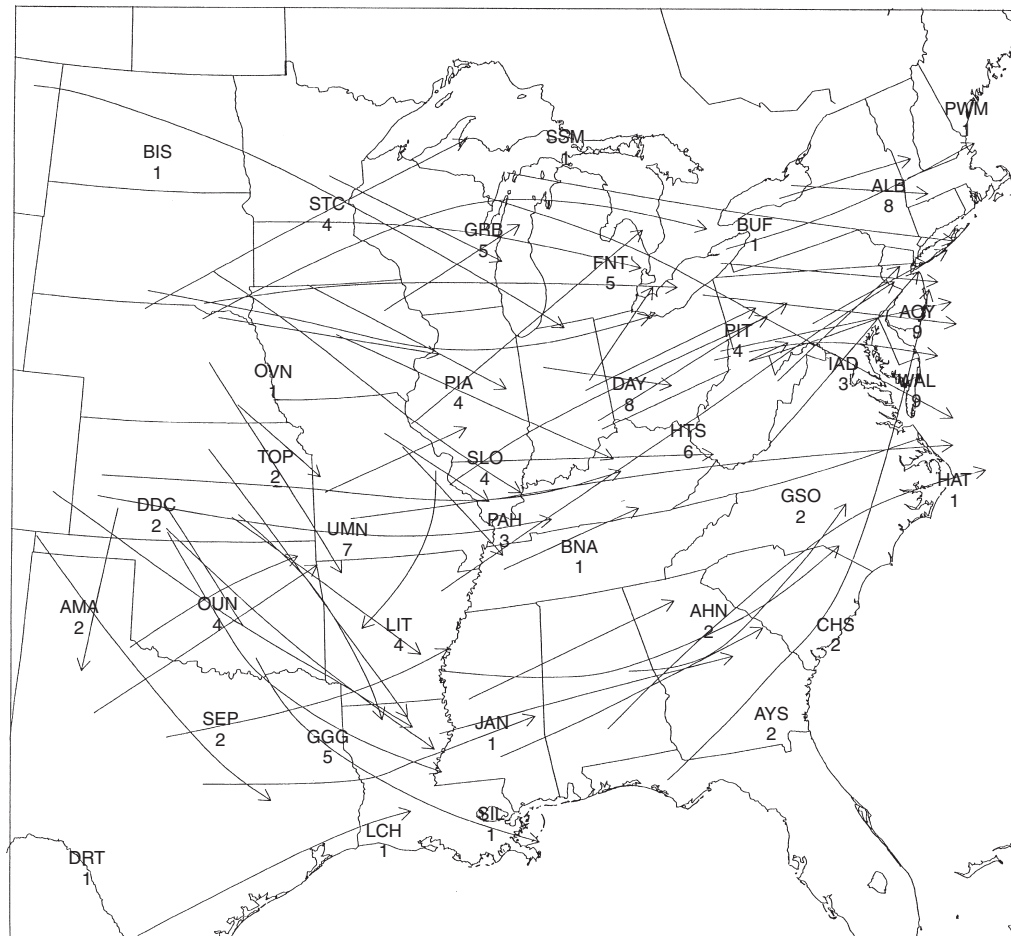


Figure 10 Graphical plot of derecho path and centroid, along with the number and location of proximity soundings, as generated for a climatological study of well-organized derecho events during the period of 1983–93 over the central and eastern USA. Adapted from Evans, J.S., Doswell III, C.A., 2000. Examination of derecho environments using proximity soundings. *Weather and Forecasting* 16, 329–342.

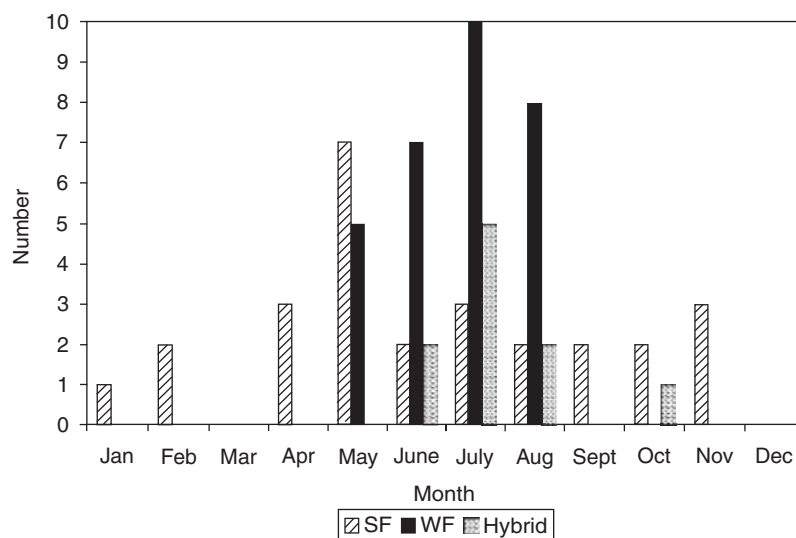


Figure 11 Monthly distribution of derecho events over the central and eastern USA for the years 1983–93, for the events shown in Figure 10. Events are also subclassified based on the relative strength of the synoptic-scale forcing, as either strong forcing (SF), weak forcing (WF), or hybrid events. Adapted from Evans, J.S., Doswell III, C.A., 2000. Examination of derecho environments using proximity soundings. *Weather and Forecasting* 16, 329–342.

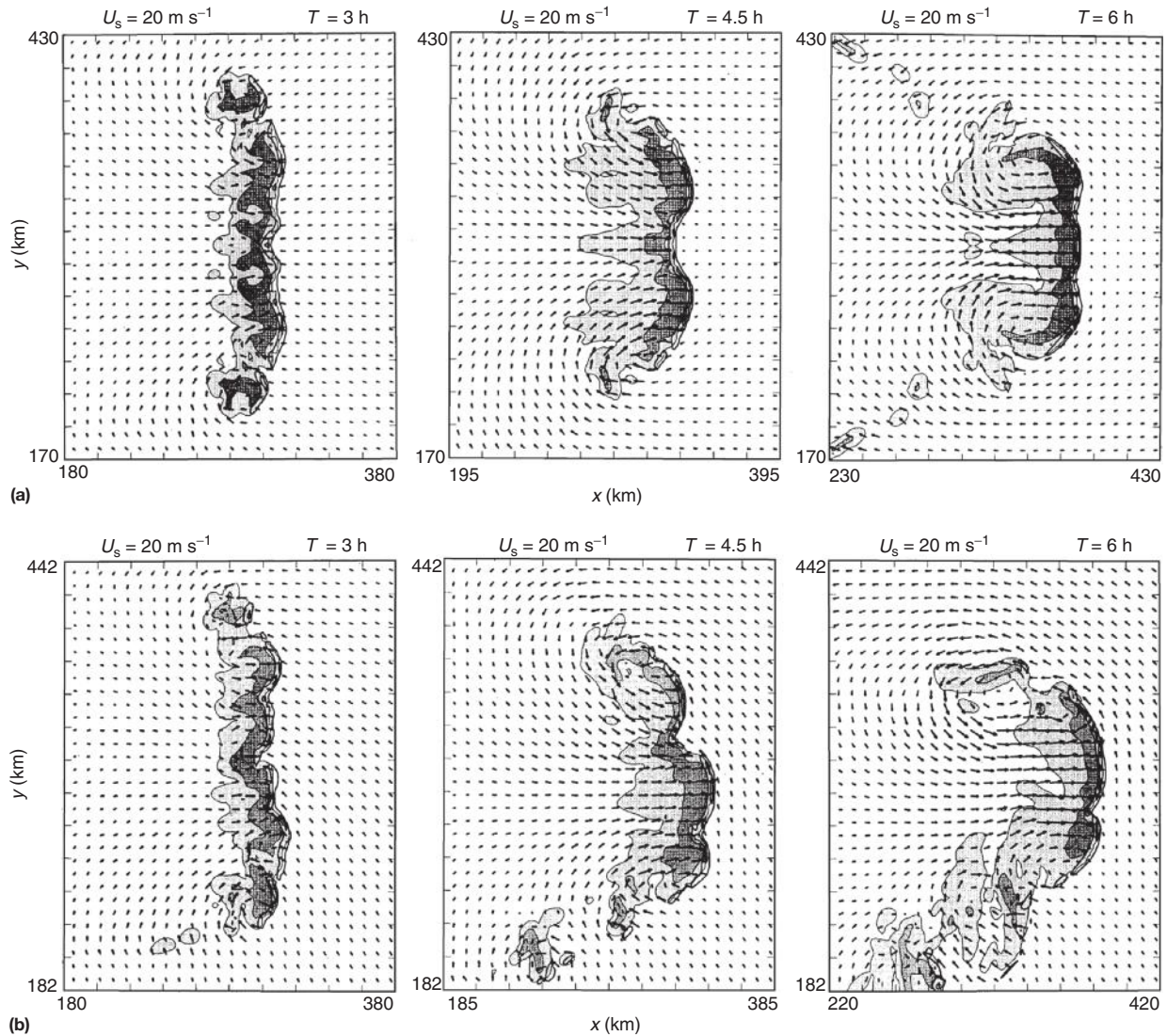


Figure 12 Horizontal cross sections of system-relative flow, rainwater mixing ratio, and vertical velocity at 2 km AGL for the $U_s = 20 \text{ m s}^{-1}$ 2.5 km shear (a) non-Coriolis and (b) Coriolis simulations at 3, 4.5, and 6 h, respectively. Vectors are presented every four grid points (8 km), with a vector length of 8 km equal to a wind magnitude of 20 m s^{-1} . The rainwater is contoured for magnitudes greater than 1 g kg^{-1} (lightly shaded) and magnitudes greater than 3 g kg^{-1} (darkly shaded). The vertical velocity is contoured at 5 m s^{-1} intervals, with the zero contours omitted. A domain speed of $u_m = 18.5 \text{ m s}^{-1}$ has been subtracted from the flowfield. Tick marks are spaced 20 km apart. Adapted from Weisman, M.L., Davis, C., 1998. Mechanisms for the generation of mesoscale vortices within quasi-linear convective systems. *Journal of the Atmospheric Sciences* 55, 2603–2622.

System strength and severity are enhanced even further through the development of the line-end vortex pair (e.g., Figure 12(a)), which focuses and strengthens the mid-level rear-inflow jet, thereby enhancing the resultant convective downdrafts and surface outflow. The source of this line-end vorticity is both the downward tilting of the ambient vertical wind shear layer as well as upward tilting of the system-generated vertical wind shear associated with the ascending updraft current and descending rear-inflow jet.

The processes described above contribute to the evolution of all convective systems, but produce severe weather

for a relatively restricted range of environmental conditions. Generally, long-lived, severe wind-producing convective systems, such as bow echoes and derechos, are produced in idealized simulations for environments with at least 2000 J kg^{-1} of CAPE and at least 10 m s^{-1} of vertical wind shear over the lowest 2–5 km AGL. However, the more coherent systems with significant bookend vortices and strong, elevated rear-inflow jets, as presented in Figure 12, are restricted to environments with at least $15\text{--}20 \text{ m s}^{-1}$ of vertical wind shear over the lowest 2–5 km AGL.

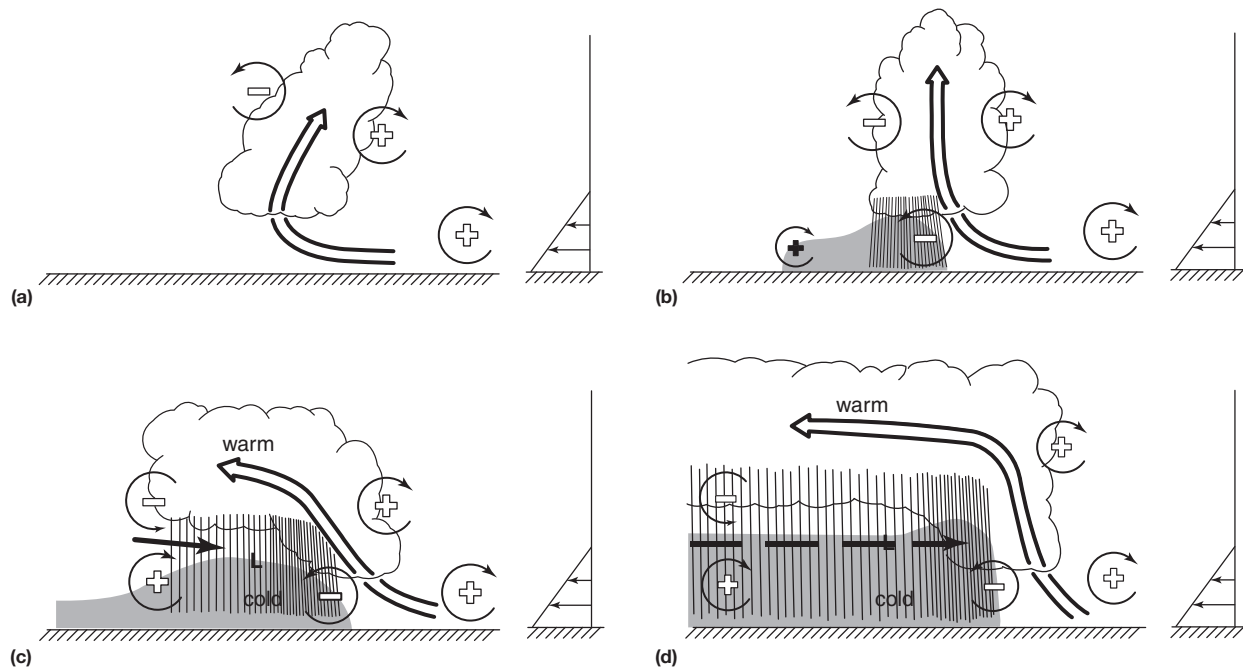


Figure 13 Four stages in the evolution of an idealized bow echo developing in a strongly sheared, large CAPE environment. The updraft current is denoted by the thick, double-lined flow vector, with the rear-inflow current is (c) denoted by the thick solid vector. The shading denotes the surface cold pool. The thin, circular arrows depict the most significant sources of horizontal vorticity, which are either associated with the ambient shear or which are generated within the convective system, as described in the text. Regions of lighter or heavier rainfall are indicated by the more sparsely or densely packed vertical lines, respectively. The scalloped line denotes the outline of the cloud. Reproduced from Weisman, M.L., 1993. The genesis of severe, long-lived bow-echoes. *Journal of the Atmospheric Sciences* 50, 645–670.

See also: Mesoscale Meteorology: Density Currents; Gust Fronts; Mesoscale Convective Systems; Microbursts.
Numerical Models: Convective Storm Modeling. **Weather Forecasting:** Severe Weather Forecasting.

Further Reading

Evans, J.S., Doswell III, C.A., 2000. Examination of derecho environments using proximity soundings. *Weather and Forecasting* 16, 329–342.
 Fujita, T.T., 1978. Manual of Downburst Identification for Project Nimrod. Satellite and Mesometeorology Research Paper No. 156. Department of Geophysical Sciences, University of Chicago.

Johns, R.H., 1993. Meteorological conditions associated with bow echo development in convective storms. *Weather and Forecasting* 8, 294–299.
 Johns, R.H., Hirt, W.D., 1987. Derechos: widespread convectively induced wind-storms. *Weather and Forecasting* 2, 32–49.
 Johns, R.H., Doswell III, C.A., 1992. Severe local storm forecasting. *Weather and Forecasting* 7, 588–612.
 Przybylinski, R.W., 1995. The bow echo: observations, numerical simulations, and severe weather detection methods. *Weather and Forecasting* 10, 203–218.
 Weisman, M.L., 1993. The genesis of severe, long-lived bow-echoes. *Journal of the Atmospheric Sciences* 50, 645–670.
 Weisman, M.L., 2001. Bow echoes: a tribute to T.T. Fujita. *Bulletin of the American Meteorological Society* 82, 97–116.
 Weisman, M.L., Davis, C., 1998. Mechanisms for the generation of mesoscale vortices within quasi-linear convective systems. *Journal of the Atmospheric Sciences* 55, 2603–2622.

Density Currents

PG Baines, University of Melbourne, Melbourne, VIC, Australia

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

Density currents (also known as ‘gravity currents’) in the atmosphere are air flows that are principally due to differences in density (i.e., temperature) between the neighboring bodies of air. Descending cold dense bodies of air are sometimes termed ‘katabatic flows’ or ‘katabatic winds’ at the ground level and ascending heated air constitutes the ‘anabatic winds.’ Such density differences cause lateral pressure gradients, which produce horizontal motion, with cold air moving over the ground into a warmer air mass. They occur on length scales ranging from several hundred meters to hundreds of kilometers. The Earth’s rotation (via the Coriolis force) is normally not important for these flows, unless they extend over distances of 100 km or more. In this article, some examples of these flows are provided, and their basic dynamical properties for flows over horizontal ground that move in one direction and that which move radially outward in two dimensions are described, and then density currents that flow downslopes, taking into account the effects of radiative forcing and environmental stratification, are discussed.

Examples

Atmospheric examples of such flows are common. They include *sea breezes*, in which heating of the air over the land by convection from the solar-heated surface below causes a lateral density difference between that and the relatively cooler air over the ocean. The onshore-flowing sea breeze is the result. Similarly, the weaker offshore nocturnal *land breeze* occurs due to radiative nocturnal cooling over the land. Radiative cooling of the air causes flow down sloping topography, forming *drainage flows*, and *valley winds*, particularly at night. Another source is thunderstorms, which contain downdrafts of cold air due to the drag of falling raindrops, and cooling because of their evaporation. On reaching the ground, these downdrafts spread and form density currents. Cold fronts usually contain one or more squall lines, which consist (by definition) of a line of thunderstorms, so that the associated downdraft-produced density currents are conspicuous features of cold fronts. Such flows are sometimes made visible by suspended dust, which mark out the features of the cold density current. Noted examples of these are seen in the Sudan (where they are known as *haboobs*), and in India, Australia, and Arizona during dry summers. In these flows the dust usually makes a negligible contribution to the density difference, but this is not the case in another example of density currents – *powder snow avalanches*. Here turbulence causes the snow to be suspended in the air, producing the density difference that causes the downflow, which in turn produces the turbulence.

Unidirectional Density Currents

The dynamical essentials of the above phenomena can be encapsulated by looking at simple idealized flows that represent them. These can best be seen from simple laboratory experiments. Here water is used to represent air, because both are effectively incompressible for motions on these length and time scales. A density current can be created in the laboratory by releasing cold water into relatively warmer water, but it is

usually simpler to work with isothermal fresh water and use dissolved salt to make the water denser. In **Figure 1** a body of dense salty water has been released into a tank of fresh water, producing a density current moving from left to right over a horizontal surface, viewed from the side. The leading part of the current consists of a ‘head’ with an overhanging leading nose. This head may be regarded as a limiting form of hydraulic jump in a cold layer of dense fluid, in which the depth of this layer upstream of the jump is zero. Behind the head the fluid has a three-layer structure. The layer of dense fluid constituting the main part of the current is at the bottom. This fluid moves faster than the head, and catches up with it. There it rises and is mixed with the surrounding lighter environmental fluid, and forms a density-stratified layer, spread out above the bottom current. Vorticity is produced in the upper part of the head by shear between the head and the ambient fluid, and this is also deposited in the stratified layer. This mixed stratified layer moves slowly in the direction of the current. It is highly turbulent immediately behind the head, where most of the mixing takes place, and this turbulence decays with distance from the head. Here the interfaces between the mixed layer and

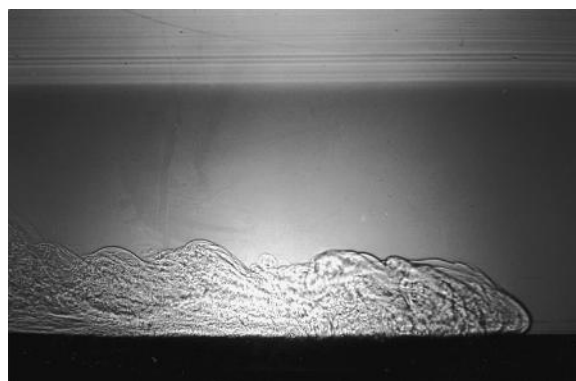


Figure 1 Shadowgraph showing the structure of the flow in and behind a density current head.

the dense fluid below and the environmental fluid above are generally stable, so that apart from the decaying turbulence little further significant mixing occurs.

Figure 1 shows a unidirectional gravity current head, averaged across its width by the shadowgraph technique. In fact, density current heads have three-dimensional structure, as shown in **Figure 2**. The leading edge contains many lumps and bumps, and these shapes change continually as the head propagates. Some small parts of the head are more advanced than others, but these then disappear and are overtaken by others. This *lobe* and *cleft* structure is due to the drag on the current by the rigid lower surface over which it propagates. This retards the lowest levels of dense fluid, causing the overhanging noses. The lighter fluid beneath these noses then rises through the dense fluid driven by its buoyancy difference, and this causes the unstable three-dimensional lobe and cleft structure that is seen in **Figure 2**. This structure can sometimes be seen in the atmospheric examples cited above, where dust or moisture make the cold air visible.

We may consider an idealized two-dimensional model of a density current, in which the volume flux Q and density ρ_1 of the dense fluid are specified at a given source location. If the density of the environmental fluid is ρ_0 , the buoyancy of the dense fluid is then $g' = (\rho_1 - \rho_0)g/\rho_0$. Behind the head, the velocity v_1 and thickness d_1 of the dense layer are approximately constant. We then have $Q = v_1 d_1$, and from dimensional analysis alone we have

$$d_1 \sim (Q^2/g')^{1/3}, \quad v_1 \sim (Qg')^{1/3} = (g'd_1)^{1/2} \quad [1]$$

The speed v_H of the head also scales with v_1 , with $v_H < v_1$. Laboratory observations of flow into a deep homogeneous environment at large Reynolds numbers give

$$v_H = 1.2(g'd_1)^{1/2}, \quad v_1 = 1.4(g'd_1)^{1/2} \quad [2]$$

which are approximately constant with time for a sustained source of fluid. The same observations also show that the total height of the head is typically $3d_1$, and the height of the nose is

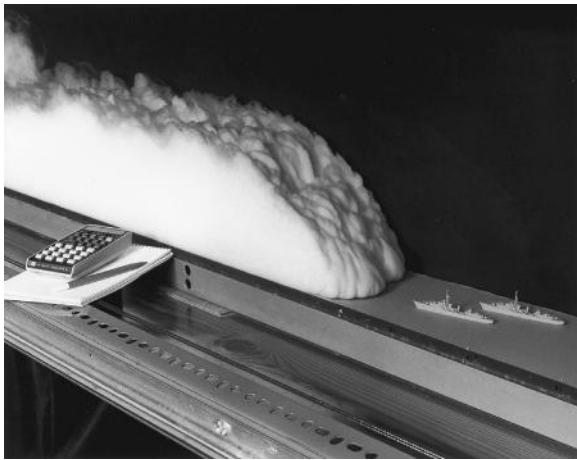


Figure 2 A laboratory simulation of a density current, in which the dense fluid has been made visible by milk. Note the lobe and cleft structure of the dense fluid interface, and the low-level transient overhanging noses.

about $0.15d_1$. These expressions apply to density current heads in general, when the buoyancy and thickness of the dense fluid approaching the head are given by g' and d_1 . This steady pattern is maintained because the driving pressure gradient is dynamically balanced by loss of momentum through mixing in the head (see below), and the following current is restrained by friction with the bottom surface and the fluid above. When the current length becomes large the thickness of the dense layer decreases gradually away from the source, providing a small pressure gradient to overcome friction.

Unidirectional Density Currents due to Collapse of a Dense Body of Fluid

Density currents may also be created by suddenly releasing a large body of dense fluid within a stationary environment of lighter fluid. This models a finite-sized source, and is readily simulated in a laboratory tank by raising a vertical barrier that separates the dense and lighter fluid. In a two-dimensional situation where the fluid spreads in the x -direction only, the flow passes successively through three stages. In the first stage of collapse of the dense fluid (of initial depth d_0 , length L_0 , and area $A = A_0 = d_0 L_0$ in a fluid of total depth D_0), termed the *slumping* stage, the front travels as a density current of constant speed and depth. However, the surface of the dense fluid behind this front is not horizontal as large amplitude internal waves propagate on it (**Figure 3**). These waves reflect from the far end (or center, for a symmetric collapsing body), and the character of the flow changes when they catch up to the front, or density current head.

Density current propagation in the atmosphere is often significantly affected by upper level stratification. For example, it is common for sea breeze fronts to propagate inland under an atmospheric inversion, which acts as an effective lid. This may be represented in experiments by a finite total depth. If this is present, the resulting motion of the ambient fluid may affect the waves on the interface with the dense fluid, and hence affect the details of the slumping behavior. In particular, the presence of a reversed flow in the upper layer causes a corresponding reduction in the speed over the ground of the collapsing front, relative to eqn [2]. The distance x_s at which the slumping stage ends and the waves reach the head, measured from the end of the tank, is then observed to be

$$x_s/L_0 = 3 + 7.4d_0/D_0 \quad [3]$$

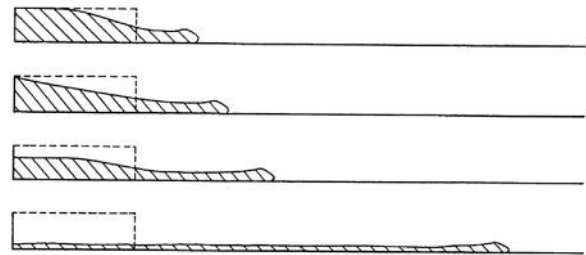


Figure 3 Schematic diagrams showing various stages at successive times in the initial slumping stage of the collapse of a rectangular volume (shown dashed) of dense fluid with $d_0/D < 1$.

After the reflected waves have caught up with the density current head the flow enters the second self-similar (inertio-gravity) stage, which is dominated by inertia, buoyancy, and mixing. Here the dense fluid collapses in the form of a rectangle of approximately uniform area, with increasing length and uniform decreasing height d_1 . This rectangular uniformity is maintained by internal waves propagating outward on the dense fluid interface. Mixing in the main body of the collapsing rectangular current is generally very small. This is because the mean gradient Richardson number at its upper boundary (on which mixing depends) is generally greater than 0.25, implying that the mean flow is stable and does not generate local mixing. Hence the density within the collapsing rectangular body of dense fluid remains largely unaltered until the fluid enters the head, where most of the mixing takes place. The total rectangular area $A(t)$ of unmixed dense fluid continually decreases because $v_1 > v_H$ in eqn [2]. During this stage, the length of the dense fluid increases as $(Ag't^2)^{1/3}$, so that the velocity of the density current head decreases with time as $(Ag'/t)^{1/3}$ because of the steady decrease of d_1 . This continues until the third stage is reached, where the motion becomes sufficiently thin and slow for viscous effects to become important, a regime that is not relevant here.

In the second stage, dense fluid enters the head from behind where it is mixed with environmental fluid. The result is spread out behind, in a layer over the dense layer. The rate of mixing within the head is roughly proportional to its mean height, and hence decreases with it. At early times in this stage ($x > x_s$), the detained mixed fluid consists mostly of the dense fluid with a small part of environmental fluid. However, as the current proceeds, this proportion reverses and near the end ($x \gg x_s$), nearly all the mixed fluid is environmental. When the end of the second stage has been reached (at $x = x_s + 29A_0^{1/2}$), and mixing has effectively ceased, the total volume of mixed fluid that has been produced is slightly more than twice the initial volume of dense fluid. The volume of the remaining unmixed dense fluid is quite small, so that overall, the dense and environmental fluids have been mixed in approximately equal proportions.

Radial Density Currents

In the atmosphere, collapsing localized bodies of dense fluid may be constrained to spread in one direction if they are spatially confined such as in a valley. However, this is usually not the case for thunderstorm downdrafts, for example, which are often free to spread horizontally without confinement. These and other localized sources of cold air in flat terrain, or in a valley that widens at a constant rate, can cause density currents that spread in the radial direction. These have a curved front or head, expanding radially away from a nominal central source. Provided the curvature of the front is not too large, the speed of the front and the fluid behind it are given approximately by eqn [2].

Such flows may be modeled by axisymmetric collapsing bodies of dense fluid, which have some properties that resemble those of one-dimensional spreading. Again, there are three stages of spreading: the slumping, inertial-buoyancy balance, and viscous stages. For a body of dense

fluid of initial height d_0 , radius R_0 , and volume V_0 , in a fluid of overall depth D , the initial collapse occurs in the 'slumping' stage, governed by wave propagation and adverse flow of the ambient fluid. Here the radial position R of the front increases at a constant speed in the range $R_0 < R < R_s$, where R_s corresponds to x_s and denotes the limit of the slumping stage. The shear between the front of dense fluid and the ambient fluid produces vorticity that is initially contained within the mixing head of the current, and as the front expands radially, this vorticity increases by vortex stretching (Figure 4). Downbursts that cause this process may occur on a range of scales – from several hundred meters or more for thunderstorm downbursts, and somewhat less than this for *microbursts*. The latter constitute particular hazards for light aircraft on take-off and landing, because they may create strong localized circular vortices with radii of 100 m or more.

The vortex intensifies as it expands, and initially may be much deeper than the following fluid. For a small initial volume of dense fluid, this expanding vortex ring of dense mixed fluid is all that is produced. But, for a larger initial volume, as the dense fluid expands and the head moves outward, the initial vortex progressively breaks up and is subsumed into the mixed layer. New vorticity is continually created at the head, as for the unidirectional currents, but this is associated with the newly mixed fluid and the stretching is much less than that in the initial vortex. In this second inertial-buoyancy phase (where $R > R_s$), gravity waves maintain approximately uniform depth of the body of dense fluid, and it collapses in the form of an axisymmetric pillbox, with the volume slowly decreasing due to mixing and detrainment behind the circular head. If the volume of dense fluid at any given time is $V = \pi R^2 d_1$, and the properties of the front are given by eqn [2], the radius R increases as

$$R \sim (g'V)^{1/4} t^{1/2} \quad [4]$$

V slowly decreases as dense fluid entering the head is mixed, and this dynamical regime continues until (in the laboratory) the viscous regime is reached.

If the source of dense fluid is maintained with constant volume flux Q_1 , the initial vortex and head form as above but the flow behind it evolves differently, as the depth d_1 and velocity v_1 of the following radial outflow are not uniform and decrease with radial distance r . The only significant mixing

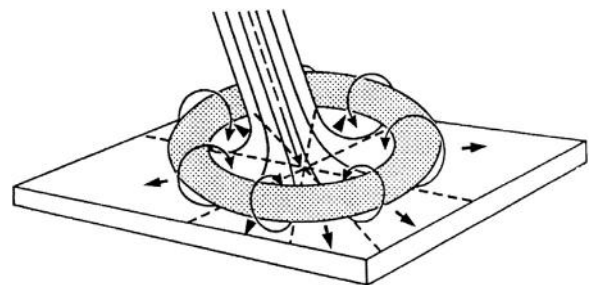


Figure 4 Schematic diagram of the ring vortex shortly after a downdraft reaches the ground.

occurs behind the head, and for steady flow, dimensional analysis gives

$$d_1 \sim \left(\frac{Q_1^2}{g' r^2} \right)^{1/3}, \quad v_1 \sim \left(\frac{Q_1 g'}{r} \right)^{1/3} \quad [5]$$

In the inertio-gravity range, the radial speed of the head is then given by

$$v_H = 0.63 \left(\frac{Q_1 g'}{t} \right)^{1/4} \quad [6]$$

Density Currents Downslopes

When the terrain is not horizontal but slopes downward, buoyancy acts directly to drive the current downward and this is stronger than the indirect effect of establishing a horizontal pressure gradient as described above. The downslope buoyancy force is now $g' \sin \theta$ per unit mass of dense fluid, where θ is the angle between the sloping bottom and the horizontal. The onset of a steady source of dense fluid at the top of the slope leads to the formation of a density current head similar to that described above (see below). In the following current the physics is different, because the flow is now unstable, unless the slope angle is very small. With a homogeneous environment, mixing now occurs along the whole length of the density current, and not just behind the head. After the passage of the head an approximately steady flow is established, with the buoyancy force being balanced primarily by *entrainment* of environmental fluid into the current from above and, to a much lesser extent, by the drag on the bottom surface. Both the size of the head and the thickness of the following current increase with downslope distance.

Behavior of the Head

For a constant supply of dense fluid of buoyancy g'_0 and flow rate Q_0 at the top of the slope, the speed of the head v_H is almost independent of slope angle and is given approximately by

$$v_H = (1.5 \pm 0.2) (g'_0 Q_0)^{1/3}, \quad 5^\circ \leq \theta \leq 90^\circ \quad [7]$$

However, in spite of this uniform speed, the size of the head increases as it moves downslope, at the rate

$$dH/dx = 0.72\theta/\pi \quad [8]$$

where H is the height (or thickness) of the head, x is downslope distance, and θ is in radians. The increased buoyancy at steeper slopes is balanced by increased entrainment that acts to keep the head speed constant, regardless of downslope distance and slope angle.

Entrainment in Downslope Flows into Uniform Environments

After the passage of the head the following flow is unstable to shear instability, and mixing with the overlying fluid results. This process becomes stronger with increasing slope angle θ . The mixed fluid above the dense layer is also denser than the environment, and hence it also moves downslope under gravity

but at a reduced speed. The thickness and volume of this mixed layer both increase with downslope distance. This combined flow may now be regarded as a single entity, and the net downslope buoyancy flux is constant with x and t , and equal to $g'_0 Q_0$. The net entrainment of environmental fluid into this overall downslope flow may be described by an entrainment coefficient E which is a function of the bulk Richardson number Ri , defined by

$$Ri = g' d \cos \theta / U^2 \quad [9]$$

where U is the mean velocity and d the mean thickness of the total flow. At each level, the velocity of inflow w_e of the environmental fluid into the downflow is given by

$$w_e = E(Ri)U \quad [10]$$

where U is the mean velocity of the downflow. E decreases monotonically with increasing Ri , from $E = 0.075$ at $Ri = 0$ to very small values for $Ri > 0.8$. In these flows, Ri is approximately constant with downslope distance, and decreases with increasing slope angle. Experiments show that U , and hence w_e , are also constant, so that the downslope flux Q and the lateral spreading increase linearly with distance. The dynamics of these flows resemble those of plumes of buoyant fluid (from, e.g., chimneys), and are generally referred to as 'plumes.'

Downslope Flows into Stratified Environments

If the environmental fluid is density stratified, the effect of stratification on density currents flowing over a horizontal surface are mostly limited to the generation of internal waves, where the current takes the form of a moving obstacle. However, for flows downslopes, if ambient stratification is present, it is a major parameter. If N is the *buoyancy frequency* of the ambient stratification, one may identify a depth D below the source, defined by $N^2 = g'_0/D$, where the ambient density equals the inflowing density (and g'_0 denotes the buoyancy of the source fluid). If there were no mixing, all of the inflowing fluid would be expected to reach and spread horizontally at this level. The speed of the head is again approximately constant over most of its distance traveled, and scales with [eqn \[7\]](#) above. The main differences from the homogeneous case concern the current following the head, which we next discuss.

There are two main flow regimes that may occur for such downslope currents. These regimes depend on the downslope buoyancy force and the frictional drag of the sloping bottom. If the bottom slope is large enough so that the buoyancy force exceeds the bottom drag, the former must also be (at least partly) balanced by mixing with the overlying fluid. This results in substantial mixing with the environment and entrainment of this mixed fluid into the downslope current, implying inflowing environmental fluid toward the current. This dynamical picture is similar to that with a homogeneous environment as described above – such flows may be regarded as 'plumes,' and as being in the *Plume regime*. This flow is maintained until the density of the downslope current approaches its mean level of neutral density, beyond which it overshoots this level due to inertia, and springs back to settle as an intrusion into the environment around this level.

If instead the bottom slope is small, so that the down-slope buoyancy force is small enough to be balanced by the bottom drag, a different flow regime results. Here the mixing with the environment is small, and the downslope current resembles a gravity current with a distinct upper surface. The mixing that does occur is associated with eddies that mostly lie outside (above) the current, and remove fluid from it in thin filaments. This process has the nature of *detrainment*, whereby fluid leaves the downflow and mixes with the environment, so that, on average, environmental fluid leaves the vicinity of the current, rather than being drawn toward it as in the entrainment case. An example of such flows is shown in Figure 5, and they are said to be in the *Gravity current regime*.

These flows are governed by two dimensionless parameters – a bulk Richardson number Ri defined as in eqn [9] but based on the dense layer only, and a parameter B , the *Buoyancy number*, defined by

$$B = QN^3/g'^2 \quad [11]$$

which is a measure of the effect of the ambient stratification. On theoretical grounds, which are consistent with experiments, the boundary between the two flow regimes is approximately

$$C_D = 0.2B^{0.4} \sin \theta \quad [12]$$

Here if the left-hand side is the larger the flow is in the gravity current regime, and if it is the smaller, the flow is in the plume regime. In principle, since B varies with the flow, the character of the flow may change with distance down the slope.

In the gravity current regime, the dense layer is observed to have approximately uniform thickness over most of its length, but its velocity mostly decreases. It loses fluid to the environment, but also entrains a little fluid from it so that its density progressively decreases, until it reaches its ambient level where it spreads out, at a vertical distance somewhat less than D below the source. Entrainment into this dense layer may be expressed in terms of an entrainment coefficient

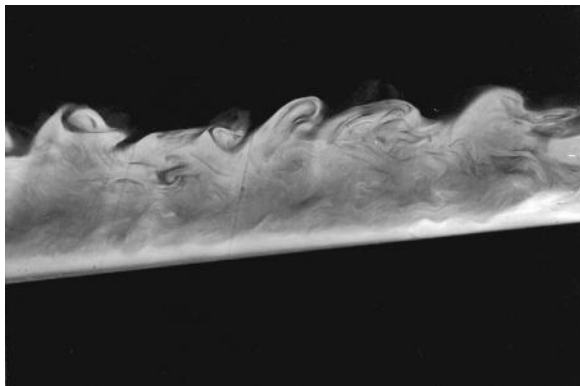


Figure 5 The turbulent, mixing interface in the body of a dense layer flowing down a slope at 6° to the horizontal, in a density-stratified environment. The dense fluid is the lightly dyed layer close to the boundary. Note the filaments extending from it, and the detrained mixed fluid above.

that depends on Ri , in a manner similar to eqn [10]. Fluid leaving the dense layer finds its own local neutral level, and this occurs continuously along the path of the current. Dense fluid from the source may therefore be distributed over a range of depths, and not all near one level as in plumes. This *detrainment* into the environment depends on both Ri and B – larger B implies larger detrainment, to the extent that all the dense fluid may be detrained before it reaches its ambient level.

Katabatic Flows Forced by Radiative Cooling

A prominent source of density currents is cooling of air near the ground through radiation. This is a common occurrence at night, particularly under clear skies, and causes drainage flows over uneven terrain. In complex terrain, this can occur at a number of source regions, at different altitudes. Depending on the circumstances, these flows may be quite strong and have depths of several hundred meters, or have speeds of only several meter by second and depths less than 50 m. Nocturnal cooling also tends to reestablish the ambient stratification at low levels that is destroyed by convection during the day. The processes described above can then cause a complex interleaving of stratified layers of air that flow progressively down several slopes, or find their own environmental level, giving flow in various different directions at different heights. Since this occurs at night, good observations of these complex flow patterns are rare.

One region where katabatic flows are common and are reasonably well observed and understood is Antarctica (and to a lesser extent, Greenland), where radiative cooling over the ice sheets sets up a perpetual drainage regime over the whole continent. Cold air produced over the central plateau drains off toward the coast in a layer that is several hundred meters thick, and several degrees colder than the air above. This effect exists throughout the year, and is particularly strong in winter. Speeds increase as the topographic gradient increases toward the coast. Here the intensity of the katabatic flows varies according to the local synoptic situation, but very strong winds of 40 m s^{-1} or more may last for days, or even weeks in certain locations. Near the coast, or up to 50 km out to sea, this shallow, intense (supercritical) air stream may undergo an internal hydraulic jump and adjust to a broader, more slowly moving (subcritical) air stream. These cold offshore katabatic flows may push the ice away from the coast, causing rapid new ice formation in the open water that takes its place, particularly in winter.

Anabatic Flows Forced by Radiative Heating

Daytime radiative solar heating of sloping terrain causes heating of the adjacent air by mixing and convection. This tends to cause flow upslope, although these flows are not as strong as the drainage flows because they depend on lateral gradients for their existence and involve mixing with the overlying fluid. Nonetheless, they are an important part of the diurnal cycle in valley flows influenced by radiative heating and cooling. They may also be important in promoting wild (bush) fire propagation up hillsides.

See also: **Arctic and Antarctic:** Antarctic Climate. **Dynamical Meteorology:** Hydraulic Flow; Kelvin–Helmholtz Instability. **Mesoscale Meteorology:** Microbursts. **Turbulence and Mixing:** Overview.

Blumen, W. (Ed.), 1990. Atmospheric Processes over Complex Terrain. Meteorological Monographs No. 45, American Meteorological Society 323 pp.
Simpson, J.E., 1997. Gravity Currents. Cambridge University Press, 244 pp.
Turner, J.S., 1986. Turbulent entrainment: the development of the entrainment assumption, and its application to geophysical flows. *Journal of Fluid Mechanics* 173, 431–471.

Further Reading

Baines, P.G., 1995. Topographic Effects in Stratified Flows. Cambridge University Press, 482 pp.
Baines, P.G., 2005. Mixing regimes for the flow of dense fluid down slopes into stratified environments. *Journal of Fluid Mechanics* 538, 245–267.

Convective Storms: Overview

ML Weisman, National Center for Atmospheric Research, Boulder, CO, USA

© 2003 Elsevier Ltd. All rights reserved.

This article is reproduced from previous edition, volume 2, pp 548–559, © 2003, Elsevier Ltd.

Introduction

Convective storms, also commonly referred to as thunderstorms, produce some of the fiercest weather on earth, including flooding rains (with rain rates up to several inches or 100 mm h^{-1}), severe surface winds (sometimes reaching magnitudes greater than 100 kn), hail (reaching the size of grapefruit), frequent lightning, and tornadoes. Individual convective cells are generally observed on scales of $5\text{--}30 \text{ km}$, and can have lifetimes ranging from $30\text{--}40 \text{ min}$ to greater than 6 h . Furthermore, groups of convective cells can become organized into larger mesoscale convective systems, such as squall lines, bow echoes, and mesoscale convective complexes, which can extend over hundreds of kilometers and, in some cases, can last for several days.

Convective storms exist under a wide variety of conditions and evolve in an equally wide variety of ways. Storm behavior is inherently dependent on the environment in which the storm grows, including thermodynamic stability, vertical wind profiles, and mesoscale forcing influences. In the following, the properties of the most basic storm types, including the ordinary cell, multicell, and supercell, are reviewed, and the fundamental physical processes that promote the various storm behaviors are explained. The knowledge of convective storms is based largely on extensive radar studies (using both conventional and Doppler radars) as well as numerical cloud modeling studies. More information on convective storms can

also be obtained from related chapters on lightning, hail, tornadoes, mesoscale convective systems, bow echoes, convective storm modeling, and severe weather forecasting.

Observed Convective Storm Types

The concept of the convective cell is fundamental to a discussion of convective storms. The convective cell will be regarded as a region of strong updraft (greater than 5 m s^{-1}) and associated precipitating downdraft having a horizontal cross section of $10\text{--}100 \text{ km}^2$, and extending in the vertical through most of the troposphere. Intense convective cells can have updrafts greater than 60 m s^{-1} , with downdrafts sometimes greater than 30 m s^{-1} . Research has shown that convective cells as observed on radar often evolve in identifiable, repeatable patterns. On the basis of these radar characteristics, conceptual models have been proposed for the most commonly observed storm types. These include the short-lived ordinary cell, multiple cell systems or 'multicell,' and supercell.

Ordinary Cell Storm

The ordinary cell represents the most basic convective storm type (Figure 1). It consists of a single updraft, which rises rapidly through the troposphere in a conditionally unstable atmosphere, producing large amounts of liquid water and ice.

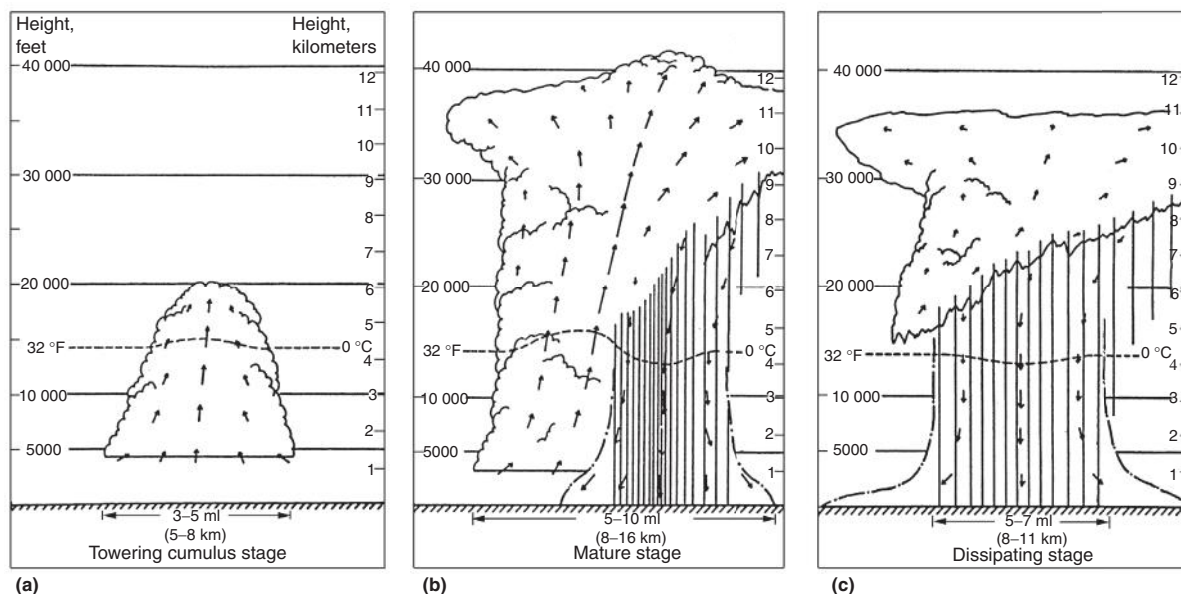


Figure 1 (a) Towering cumulus stage, (b) mature stage, and (c) dissipating stage of an 'ordinary' convective cell. Courtesy of Doswell, C.A. Adapted with permission from Byers, H.R., Braham, R.R., Jr., 1949. *The Thunderstorm*. Supt of Documents. US Government Printing Office, Washington, DC.

When the raindrops or ice particles become too heavy for the updraft to support, they begin to fall, creating a downdraft that quickly replaces the updraft. The downdraft is initially nearly saturated, but as it falls into the lower troposphere and mixes with drier air, strong evaporational cooling may occur. This cooling accelerates the downdraft (because of negative buoyancy), which spreads out horizontally as a cold pool (gust front) on reaching the surface. If the diverging outflow winds reach severe levels (greater than about 50 kn), the event is referred to as a downburst or microburst. This life cycle (Figure 1) usually takes 30–50 min to complete, and generally severe weather such as high winds or hail tends to be short-lived. Relatively weak, short-lived tornadoes do occasionally occur with ordinary cells, and are sometimes referred to as landspouts or nonsupercell tornadoes.

Multicell Storm

The multicell storm can be thought of as a cluster of short-lived ordinary cells. The cold outflows from each cell, however, combine to form a large gust front, the convergence and lifting along its leading edge being generally strongest in the downshear direction relative to the low-level (0–3 km agl) vertical wind shear vector. In most cases, this also happens to be in the direction of storm motion. This convergence and lifting can trigger new updraft development along and just behind the gust front, and new cells evolve as described in the previous subsection. Figure 2 shows this process in a vertical cross section through a multicellular hailstorm observed during the

National Hail Research Experiment. The new cell growth often appears disorganized, but occasionally occurs on a preferred storm flank. Because of their ability to renew themselves constantly through new cell growth, multicell storms often last many hours, affecting areas of thousands of square kilometers. If the storm motion is very slow, heavy local rainfall may occur, presenting the possibility of flooding. Severe surface winds in the form of downbursts or microbursts can occur with multicell storm systems, with hail and tornadoes also possible in the vicinity of strong updraft centers.

Supercell Storm

The supercell is potentially the most dangerous convective storm, often producing high winds, large hail, and long-lived tornadoes. In its purest form, it consists of a single, quasi-steady, rotating updraft and associated downdraft, which may have a lifetime of several hours. It often evolves from multicell storm systems, and even during its quasi-steady phase may comprise several small-scale rain centers embedded within a larger encompassing cellular structure. However, the general structure and evolution of the supercell suggest that it is dynamically different from ordinary convection.

A schematic of a supercell is presented in Figure 3. Unlike ordinary cells or multicell systems, supercells are often characterized by a persistent separation between the primary updraft and downdraft currents. The updraft region is generally found on the upshear side of the cloud, and is characterized by a well-defined cloud base with rapidly growing cloud turrets above.

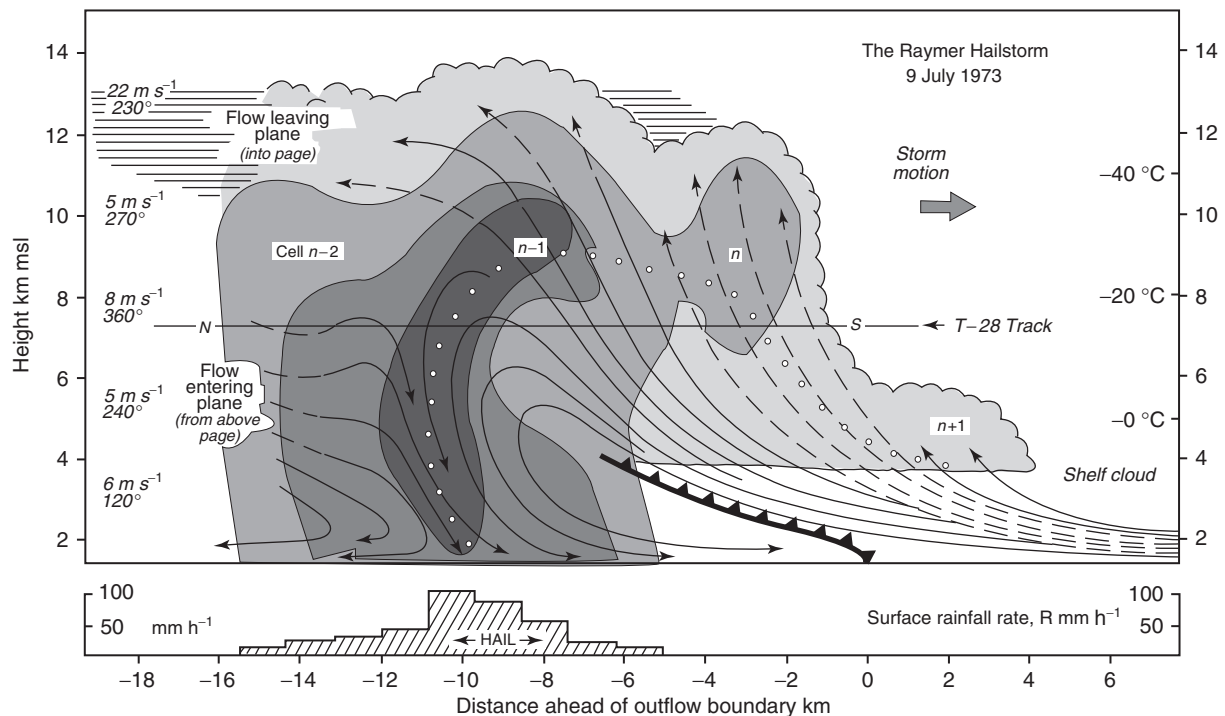


Figure 2 Vertical cross section through a multicell hailstorm, along the storm's direction of travel through a series of evolving cells ($n-2$, $n-1$, and n , $n+1$). The solid lines are streamlines of flow relative to the moving system; on the left, their broken ends represent flow into and out of the plane, and on the right, they represent flow remaining within a plane a few kilometers closer to the reader. Light shading represents the extent of the cloud, and the three darker shades represent radar reflectivities of 35, 45, and 50 dBZ. Reproduced from Browning, K.A., Fankhauser, J.C., Chalon, J.-P., et al., 1976. Structure of an evolving hailstorm. Part V: Synthesis and implications for hail growth and hail suppression. *Monthly Weather Review* 104, 603–610.

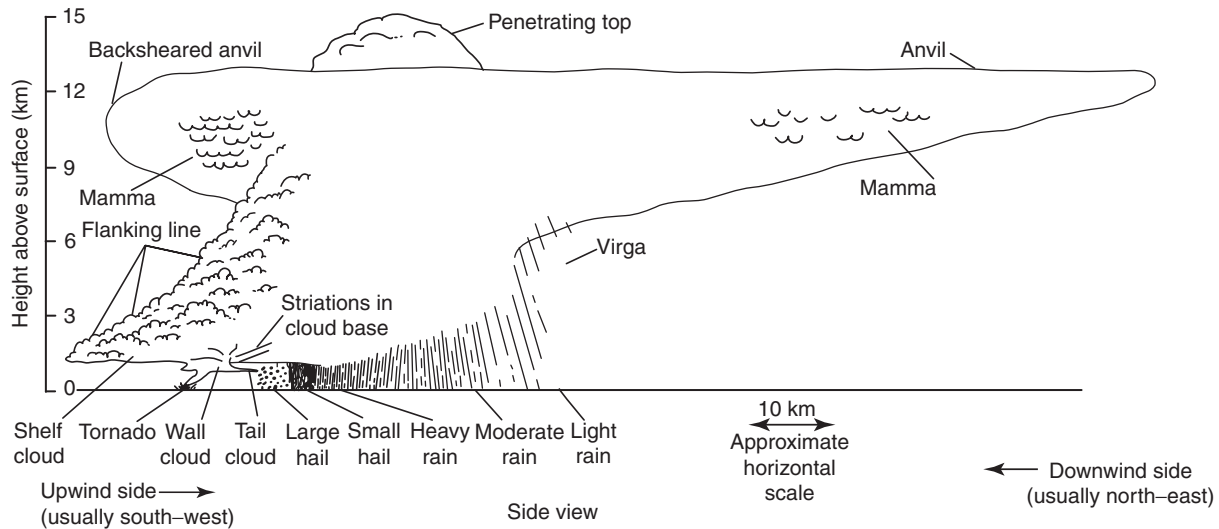


Figure 3 Schematic visual view of a mature supercell thunderstorm. Reproduced from Bluestein, H.B., Parks, C.R., 1983. A synoptic and photographic climatology of low-precipitation severe thunderstorms in the southern plains. *Monthly Weather Review* 111, 2034–2046.

This portion of the storm often exhibits pronounced cyclonic rotation. The downdraft region is found primarily downshear of the updraft region, appearing more diffuse due to the heavy precipitation. The anvil spreads predominantly downshear aloft, but in stronger storms also extends upshear somewhat as the divergence near the storm top is able to force itself upstream against the strong upper level flow. Overshooting tops are quite common for the stronger storms. Many supercells also display a stair-step-shaped flanking line extending upshear from the storm's main updraft region. A persistent lowering of the cloud base, referred to as a wall cloud, is also often observed beneath the main updraft region of the storm, and is often a precursor to the development of tornadoes.

The structure of a mature supercell as it might be observed on radar is depicted in Figure 4. The reflectivity field tends to be elongated in the direction of the mean vertical wind shear, with a hooklike appendage often appearing on the southwest flank of the storm. The midlevel reflectivity often overhangs the low-level echo, and often a bounded weak echo region (BWER) appears at middle levels above the edge of the low-level reflectivity gradient. A BWER usually indicates the presence of both strong updraft and strong rotation about a vertical axis in its vicinity.

Figure 5 presents the significant surface features commonly observed during a supercell's mature phase. The main updraft region is found straddling the hook or notch in the rain field, with two primary downdraft regions, referred to the forward flank downdraft (FFD) and rear flank downdraft (RFD), located on the downshear and upshear sides of the updraft, respectively. A surface gust front separates the cool, rainy air from the warm ambient air, with the gust front often wrapped around the southern flank of the storm due to the circulation associated with a surface mesocyclone. This rear-flank gust front can overtake the gust frontal boundary associated with the FFD, creating an occlusion of these frontal features. A tornado, if present, often forms at the tip of this occlusion (on the edge of the hook echo) on the gradient between updraft and downdraft (but within the updraft).

A time series of radar reflectivity structure for a storm that occurred on 19 April 1972 near Norman, OK (Figure 6), portrays a commonly observed trait of supercell storms. About 1 h into the storm's lifetime, the rain center appears to split into two diverging echo masses: the more intense southern storm veers to the right and slows its motion, while the northern storm moves more quickly to the northeast. Such storm splitting is common in association with supercell storms. The right mover (right relative to the direction of the ambient shear vector) is associated with a cyclonically rotating updraft, while the left mover is associated with an anticyclonically rotating updraft. Both right and left movers of a splitting storm are apt to produce severe weather such as hail and high winds, but tornadoes are rarely associated with left-moving (LM) storms.

Physical Mechanisms Controlling Convective Storm Growth and Evolution

Convective storm type and severity are strongly dependent on the environmental conditions in which the storm grows. Of particular importance is the thermodynamic instability (buoyancy) and vertical wind shear. Thermodynamic instability exerts a fundamental control on convective storm strength, as it controls the vertical acceleration of air parcels. Vertical wind shear, however, influences strongly the form that the convection might take, i.e., whether the convection evolves as short-lived ordinary cells, multicells, or supercells. In the following, the basic physical processes that contribute to the wide spectrum of observed convective storm properties are reviewed.

Buoyancy Effects

Convective storms differ dynamically from larger scale atmospheric phenomena primarily due to the much stronger vertical accelerations and resulting vertical motions (both upward and downward) that are produced. Thus, the most fundamental

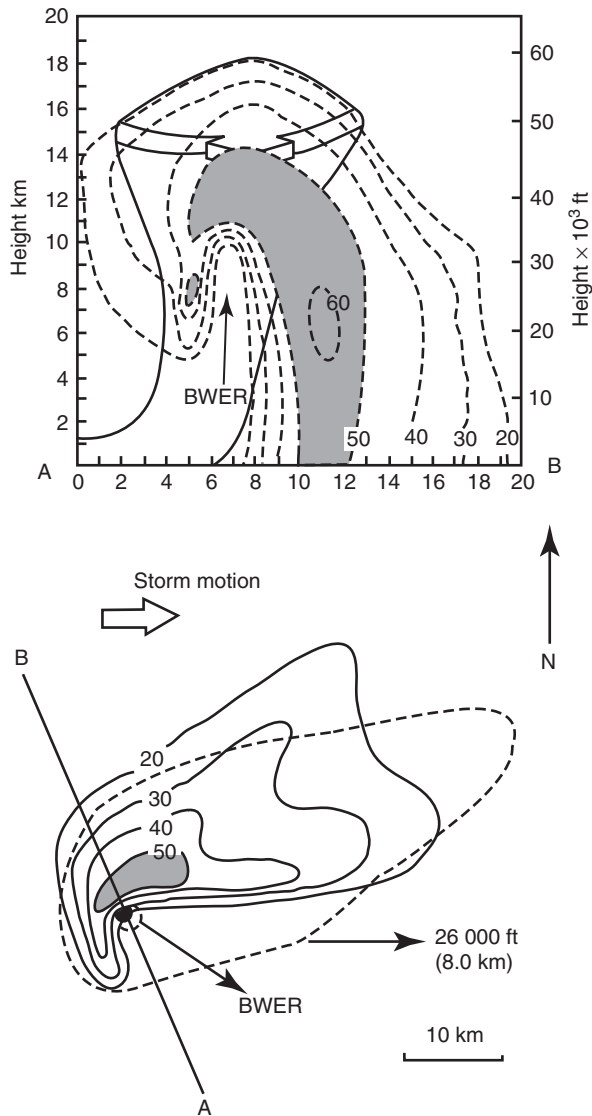


Figure 4 (Above) Vertical cross section as might be observed on a radar scope during the mature phase of an intense supercell storm. Low-level inflow, updraft, and outflow aloft (solid lines) are superimposed on the radar reflectivity (dashed lines). (Below) Composite tilt sequence. Solid lines are the low-level reflectivity contours, dashed lines outline the echo >20 dBZ derived from the middle level elevation scan, and the black dot is the location of the maximum top from the high-level scan. Adapted from Lemon, L.R., 1980. Severe Thunderstorm Radar Identification Techniques and Warning Criteria. NOAA Technical Memorandum, NWS NSSFC-3, Kansas City, MO (NTIS PB81-234809).

equation relevant to convective storm dynamics is the non-hydrostatic vertical momentum equation:

$$\frac{dw}{dt} = -C_p \bar{\theta}_v \frac{\partial \pi}{\partial z} + B \quad [1]$$

where π is a nondimensional form of the pressure, referred to as the Exner function,

$$\pi \equiv \left(\frac{p}{p_0} \right)^{R_d/C_p} \quad [2]$$

and B is the buoyancy, including the effects of water vapor and precipitation loading:

$$B \equiv g \left[\frac{\theta'}{\bar{\theta}} + 0.61(q_v - \bar{q}_v) - q_c - q_r \right] \quad [3]$$

In these equations, C_p represents the specific heat at constant pressure, p_0 is a reference surface pressure, R_d is gas constant for dry air, θ represents the potential temperature, q_v represents the water vapor mixing ratio, and q_c , q_r , and q_i represent cloud water, rainwater, and ice mixing ratios, respectively.

For an undisturbed environment (e.g., characterized by no variation of wind with height), the pressure contributions to vertical accelerations are usually very small relative to the buoyancy contributions, and are neglected. Under this assumption, an estimate of potential updraft (and downdraft) strength in a convective storm is often made by integrating the potential temperature contributions from buoyancy along a representative parcel path. For an updraft parcel, this quantity

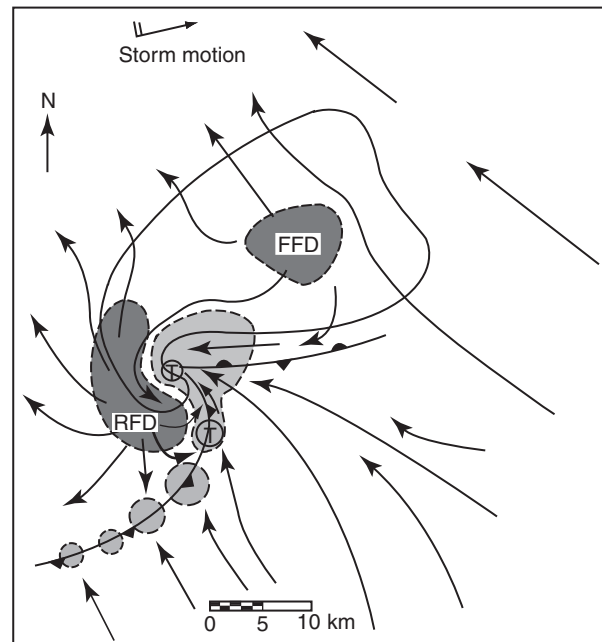


Figure 5 Schematic plan view of a tornadic supercell thunderstorm at the surface. The thick line encompasses radar echo. The thunderstorm wavelike 'gust front' structure is also depicted by the use of a solid line and frontal symbols. Surface positions of the updraft are finely stippled; FFD and RFD are coarsely stippled; and along with associated streamlines (relative to the storm). Likely, tornado locations are shown by encircled T's. The major cyclonic tornado is most probable at the wave apex, while a minor cyclonic tornado may occur at the bulge in the cold front (southern T), which also marks the favored location for new mesocyclone. Anticyclonic tornadoes, if any, are found even farther south along the cold front. Reproduced from Davies-Jones, R.P., 1985. Tornado Dynamics. In: Kessler, E. (Ed.), 1986. Thunderstorms: A Social, Scientific, and Technological Documentary. Vol. 2: Thunderstorm Morphology and Dynamics, second ed. [revised and enlarged]. University of Oklahoma Press, Norman, OK and London, UK. Adapted with permission from Lemon, L.R., Doswell III, C.A., 1979. Severe thunderstorm evolution and mesocyclone structure as related to tornado genesis. Monthly Weather Review 107, 1184–1197.

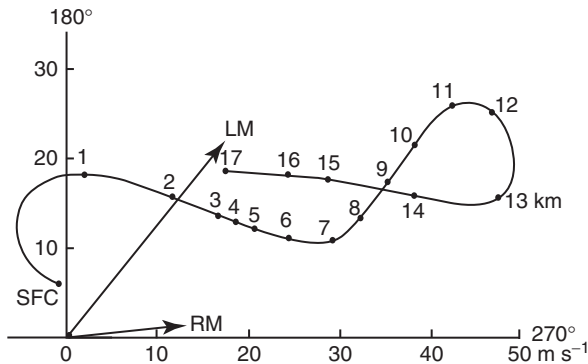
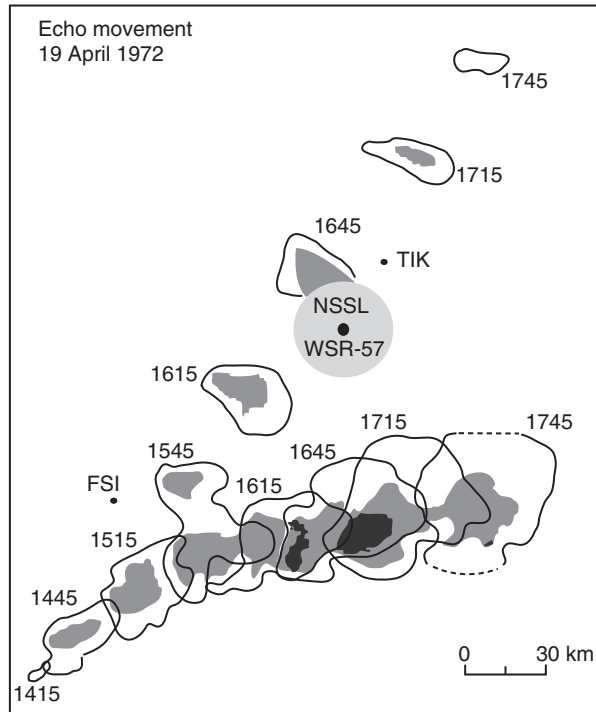


Figure 6 (Top) WSR-57 radar history of a splitting storm observed in south central Oklahoma. The solid contours indicate return greater than 10 dBZ, and the stippled regions indicate a return greater than 40 dBZ. Times adjacent to each outline are central standard time (CST). (Bottom) A hodograph representative of the storm's environment. Right moving (RM) and left moving (LM) indicate the observed motion of the RM and LM cells. Adapted with permission from Burgess, D.W., 1974. Study of a Right-Moving Thunderstorm Utilizing New Single Doppler Radar Evidence. Masters Thesis, Department of Meteorology, University of Oklahoma.

is referred to as the convective available potential energy (CAPE):

$$\text{CAPE} = \int_{\text{LFC}}^{\text{EL}} g \left(\frac{\theta'(z)}{\bar{\theta}(z)} \right) dz \quad [4]$$

where $\theta'(z)$ defines the potential temperature of a representative adiabatically ascending surface parcel, $\bar{\theta}(z)$ defines the environmental potential temperature profile, and the integral is taken over the vertical interval where the lifted parcel is warmer than its environment (usually from the level of free convection

(LFC) to the equilibrium level (EL)). This calculation is equivalent to evaluating the positive area represented on a skew-T diagram. Maximum potential temperature excesses in convective updrafts can be greater than 10 K, with magnitudes of CAPE larger than $6000 \text{ m}^2 \text{ s}^{-2}$, but generally potential temperature excesses range between 3 and 6 K, with CAPEs of 1500 and $2500 \text{ m}^2 \text{ s}^{-2}$ for moderately unstable convective days. A similar quantity can be calculated for downdraft parcels, and is referred to as DCAPE (downdraft CAPE).

By equating this CAPE (DCAPE) to vertical kinetic energy, one can then estimate the maximum updraft (downdraft) that would be expected from a given environment:

$$W_{\text{max}} = (2 \times \text{CAPE})^{1/2} \quad [5]$$

Using this relationship, a CAPE of $2500 \text{ m}^2 \text{ s}^{-2}$ would translate to a maximum possible updraft strength of 70 m s^{-1} . However, water loading, perturbed vertical pressure gradients, and mixing effects reduce these estimates by roughly 50%. Vertical motions of 60 m s^{-1} or greater have been observed in the strongest storm updrafts, but maximum downdrafts rarely exceed 30 m s^{-1} .

Vertical Wind Shear Effects

While the thermodynamic structure influences strongly the vertical accelerations in a convective storm, vertical wind shear has a strong influence on what form convection might take. In particular, short-lived ordinary cells tend to be the preferred mode of organization in weak wind shear regimes, while multicells and supercells become the respective preferred mode of organization for increasing magnitude of vertical wind shear.

The characteristics of the wind profile in this regard are best represented in the form of a hodograph, where the wind vectors at each height are plotted from the origin, and then the tips of the vectors are connected to produce a hodograph trace (Figure 7). Vertical wind shear vectors are everywhere tangent to this hodograph trace, with the length of the hodograph curve over a given depth being a direct measure of the magnitude of the wind shear over that depth.

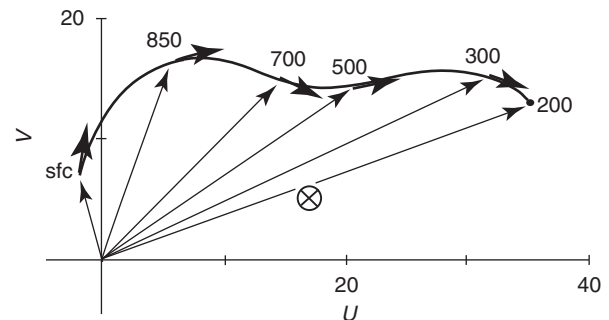


Figure 7 Composite hodograph (m s^{-1}) for tornadic supercell storms. Light arrows represent the wind vectors at each level, and heavy arrows indicate the direction of the shear vector at each level (labeled in mbar). The estimated mean storm motion is denoted by an encircled X. Reproduced from Klemp, J.B., 1987. Dynamics of tornadic thunderstorms. Annual Reviews of Fluid Mechanics 19, 369–402. Adapted with permission from Maddox, R.A., 1979. An evaluation of tornado proximity wind and stability data. Monthly Weather Review 104, 133–142.

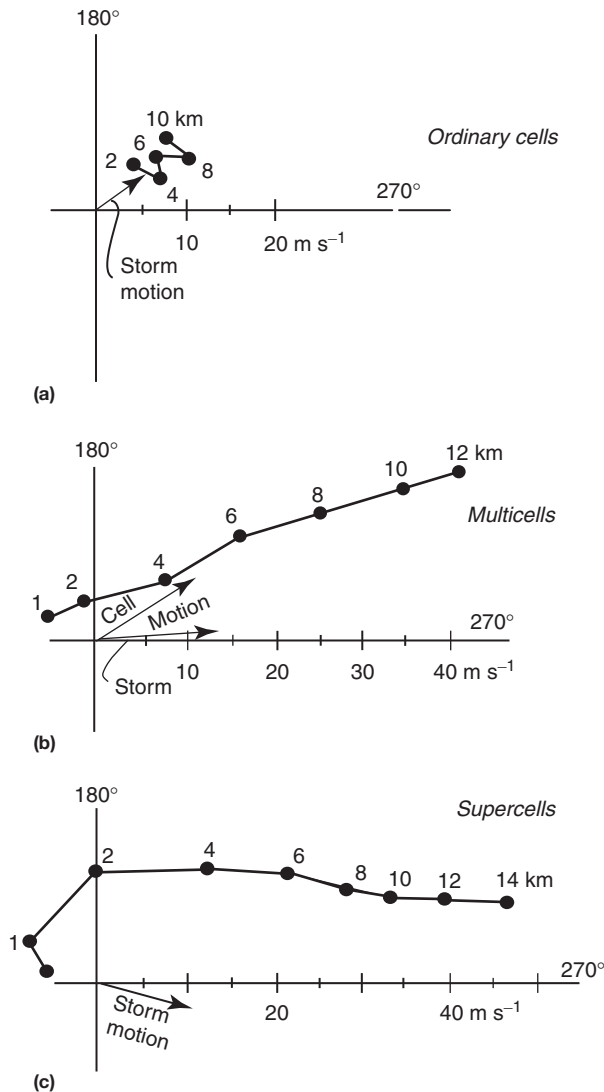


Figure 8 Typical wind hodographs for (a) ordinary cell, (b) multicell, and (c) supercell storms observed during the Alberta Hail Studies project. Adapted with permission from Chisholm, A.J., Renick, J.H., 1972. The Kinematics of multicell and supercell Alberta hailstorms. Alberta Hail Studies, Research Council of Alberta Hail Studies, Rep. 72-2, Edmonton, Canada, pp. 24–31.

The relationship between wind shear and basic storm type is demonstrated in Figure 8, which depicts composite hodographs from a study of hailstorms in Alberta, showing an increasing length of the hodograph (especially over the lowest 6 km agl) as the type of convection progresses from short-lived storms to supercells. Generally, multicell storms become more prevalent when the length of the hodograph over the lowest 4–6 km agl is greater than 10–15 m s⁻¹, with supercells becoming more prevalent when the length of the hodograph is greater than 20–25 m s⁻¹ over the lowest 4–6 km agl. Also included on the hodograph plots are observed cell motions. For ordinary cells and multiple cell systems, cell motion tends to be with the mean wind over the lower 6–8 km of the profile, appearing on or near the hodograph trace. For the supercell, however, cell motion is well off the hodograph, well to the right of a calculated mean wind from the profile. Similar off-hodograph propagation is evident for the 19 April splitting storm (Figure 6), and reflects the unique dynamic character of supercell storms, as will be described below.

Two physical mechanisms help explain the organizational capacity of vertical wind shear. The first is related to the ability of a cold pool to trigger new convective cells. The second is related to the interaction of an updraft with the environmental vertical wind shear to produce an enhanced, quasi-steady storm structure.

Cold pool–shear interactions

Cold pools are one of the most prominent features of convective storms, and have a critical role in determining whether a storm system can be maintained over a long period of time. This is due primarily to the ability of the cold pool to lift the surrounding air mass, thus serving potentially as a trigger for new convective cells. For a zero wind shear environment, the lifting along the leading edge of a cold pool is generally restricted to the depth of the cold pool's nose, as the circulation generated by the cold pool rapidly drags the lifted air rearward. If the LFC is significantly higher than the nose of the cold pool, then it is unlikely that new cells can be triggered as the cold pool propagates away from a given cell (e.g., Figure 9(a)).

This picture changes significantly with the addition of environmental low-level vertical wind shear. Associated with this vertical wind shear is an opposing circulation that can

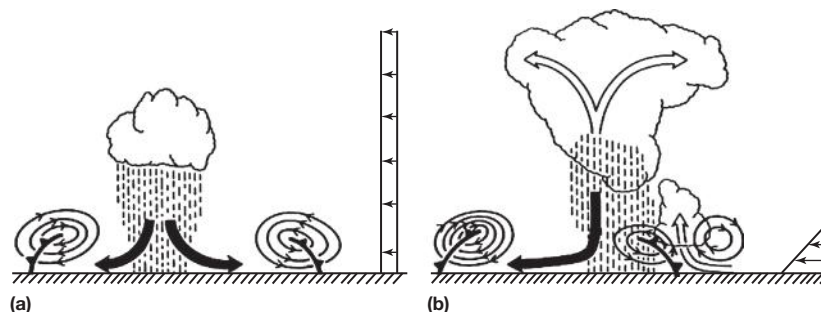


Figure 9 Schematics of cold pool–shear interactions. (a) A convective cell in a zero-shear environment produces a cold pool that propagates away from the cell. Without the presence of low-level shear, the circulation of the spreading cold pool inhibits deep lifting, and is less apt to trigger a new convective cell. (b) The presence of low-level shear counteracts the circulation of the cold pool on the downshear side, promoting deeper lifting and an enhanced potential to trigger new convective cells. Adapted from Rotunno, R., Klemp, J.B., Weisman, M.L., 1988. A theory for strong, long-lived squall lines. *Journal of the Atmospheric Sciences* 45, 463–485.

balance the circulation of the cold pool somewhat on its downshear side, producing a more vertically oriented jet of air and deeper lifting at its leading edge (e.g., Figure 9(b)). An 'optimal' state for lifting along the cold pool can be envisioned when the circulation generated by the cold pool matches the opposite circulation associated with the environmental vertical wind shear. The depth/layer of vertical wind shear most important to this process is equivalent to the depth of the cold pool itself, but deeper shear layers will also contribute somewhat to this process. Ambient shear can further enhance the strength of new cells by virtue of the fact that such cells will move downshear along with the cold pool, increasing both the relative inflow into these cells and the time over which the cells maintain their low-level convergence and feed on the warm air ahead of the gust front. All in all, the strength and longevity of multiple cell convective systems are enhanced for increasing magnitudes of ambient vertical wind shear, due primarily to the enhanced ability of the cold pool to trigger new cells.

Updraft-shear interactions

Vertical wind shear can further contribute to convective storm strength, organization, and sustenance through the interaction of the sheared flow with the convective updrafts. These effects can be both positive and negative. The negative effects are most clearly evident during the early stages of a storm's life, as clouds are observed to lean over in the direction of the mean tropospheric shear vector. This process takes vertical kinetic energy out of the accelerating buoyant plume, converting it into horizontal kinetic energy. If the shear is too strong relative to the buoyancy, a cloud can be literally torn apart.

The positive attributes of the shear are most clearly associated with the development of rotation about a vertical axis within the storm. This rotation originates through the tilting of horizontal vorticity inherent in the vertically sheared flow, as can be shown from the vertical vorticity equation:

$$\frac{d\zeta}{dt} = \omega_H \cdot \nabla_H w + \zeta \frac{\partial w}{\partial z} \quad [6]$$

where ω_H and ζ represent the horizontal and vertical components of vorticity, respectively. This process is visualized in Figure 10(a), for an isolated updraft developing in a unidirectionally sheared flow. The updraft initially deforms the ambient vortex lines upward, leading to the development of a vortex couplet at midlevels, centered on the updraft. Cyclonic vorticity is generated on the right flank of the updraft (relative to the direction of the shear vector), with anticyclonic vertical vorticity on the left flank.

The main impact of this rotation on storm structure occurs through the relationship between the velocity field and the pressure field. In particular, the localized development of rotation in a fluid is associated with lowered pressures (e.g., consider what happens when you stir a cup of coffee). For convective scales of motion, this lowering of pressure occurs whether the rotation is cyclonic or anticyclonic. If the resulting rotation at midlevels in a storm is sufficiently strong (e.g., if the storm is developing in a sufficiently sheared environment), the induced pressure deficits at midlevels will produce a significant upward-directed vertical pressure gradient force that will force the updraft to propagate to both flanks of the original cell.

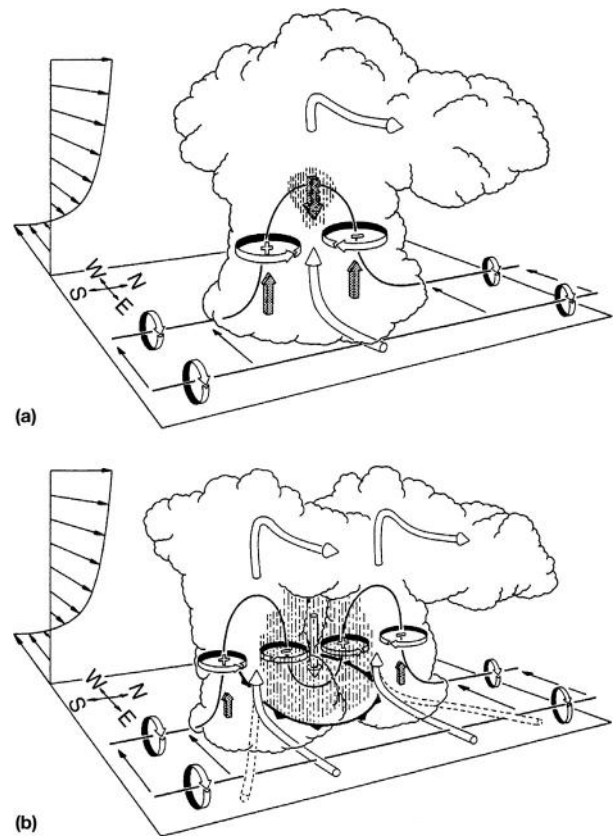


Figure 10 Schematic depicting how a typical vortex tube contained within (westerly) environmental shear is deformed as it interacts with a convective cell (viewed from the southeast). Cylindrical arrows show the direction of cloud-relative airflow, and heavy solid lines represent vortex lines with the sense of rotation indicated by circular arrows. Shaded arrows represent the forcing influences that promote new updraft and downdraft growth. Vertical dashed lines denote regions of precipitation. (a) Initial stage: vortex tube loops into the vertical as it is swept into the updraft. (b) Splitting stage: downdraft forming between the splitting updraft cells tilts vortex tubes downward, producing two vortex pairs. A new updraft is forced on the flanks of the splitting cell in response to upward-directed vertical pressure gradient forcing associated with the midlevel rotation. The barbed line at the surface marks the boundary of the cold air spreading out beneath the storm. Reproduced with permission from Klemm, J.B., 1987. Dynamics of tornadic thunderstorms. *Annual Reviews of Fluid Mechanics* 19, 369–402.

Once the updrafts propagate to the flanks (Figure 10(b)), they become more colocated with the midlevel rotation centers, which are then further enhanced by vortex stretching. The vortex tilting process continues to generate new rotation on the flanks of the storm, and the updrafts will continue to propagate toward these midlevel rotational centers. Thus, the original cell splits into mirror image cyclonic and anticyclonic storms that propagate to the right and left of the shear vector, respectively. This is the most basic process by which supercell storms may be generated and sustained.

The relationship between the velocity and pressure fields in a convective storm can be derived by taking the divergence of the momentum equations, assuming incompressibility, which

leads to the following Poisson equation for the nondimensional pressure, π :

$$\nabla \cdot (C_p \bar{\rho} \nabla \pi) = -\nabla \cdot (\bar{\rho} \mathbf{v} \cdot \nabla \mathbf{v}) + \frac{\partial B}{\partial z} \quad [7]$$

This equation can be solved for the contributions to the perturbation pressure field from the velocity and buoyancy terms on the right-hand side of eqn [7] individually, allowing the vertical momentum eqn [1] to be rewritten to reflect the contributions from velocity-related pressure perturbations and buoyancy-related processes individually as well:

$$\frac{dw}{dt} = C_p \bar{\theta}_v \frac{\partial \pi_{dn}}{\partial z} + \left(-C_p \bar{\theta}_v \frac{\partial \pi_b}{\partial z} + B \right) \quad [8]$$

The first term on the right-hand side of eqn [8] is referred to as the dynamic contribution to vertical acceleration, and includes all the effects of shear on an updraft, such as the initial tendencies for a cell to lean in the direction of the shear as well as the positive influences due to the development of rotation. The second term on the right-hand side of eqn [8] includes the usual effects of buoyancy as well as the compensating effects due to the buoyant contributions to the pressure field. For ordinary convective cells, which develop in weakly sheared environments, the buoyancy terms generally contribute 60–70% of the maximum updraft strength in a storm. However, the supercell storms, which develop in strongly sheared environments, 60–70% of the maximum updraft strength can come from the dynamic contributions, with most of this contribution coming in the lowest several kilometers of the storm. This explains why supercell storms can be unusually strong, and can persist, sometimes even in the presence of significant low-level capping inversions, as are generally observed at night.

The updraft–shear interaction processes described above are symmetric about the ambient shear vector for unidirectionally sheared environments (e.g., shear environments characterized by a straight line on a hodograph). In such cases, mirror image supercells propagating off the hodograph to the right and left of the shear vector can be produced, as demonstrated in idealized cloud model simulations presented in Figure 11(a). This symmetry is modified, however, by the addition of directional shear to the environment. If the environmental vertical wind shear vector turns clockwise with height over the lowest few kilometers agl (referred to as a clockwise-curved hodograph), as presented in Figure 11(b), the pressure forcing is enhanced on the cyclonic flank of the original cell, and a dominant cyclonically rotating supercell results from the original splitting process. However, if the environmental vertical wind shear turns counterclockwise with height (not shown), the anticyclonic member of the original split would have been favored instead. Climatologically, environmental hodographs in the vicinity of supercell storms exhibit cyclonic turning of the shear vector at low levels (e.g., consider the hodographs in Figures 7 and 8(c)), and thus cyclonically rotating supercells tend to be more common and dominant than anticyclonically rotating supercells.

Figure 12 presents the overall flow structure for a mature, cyclonically rotating supercell storm. An anticyclonically rotating supercell is the mirror image of this. The flow vectors depict the main interwoven airstreams, with the low-level flow converging from both ahead and behind of the surface gust front and rising into a deep, rotating updraft, and the midlevel

flow passing in front of and then descending behind the updraft. The updraft reaches the top of the storm, where it then diverges within the anvil, primarily in the downshear direction. While the midlevel rotation in the storm is generated via the tilting of horizontal vorticity associated with the warm, ambient environment (e.g., Figure 10(a)), the air that feeds the low-level rotation originates largely from the cold side of the surface cold pool boundary. Horizontal vorticity is generated in response to the buoyancy gradients across this boundary, as depicted by the low-level vortex lines turning toward the storm on the cold side of the forward flank gust front, and this horizontal vorticity feeds into the low-level updraft in a stream-wise sense, leading to the low-level updraft rotation. It is this low level, rotating updraft that leads to the development of significant tornadoes within supercell storms.

Summary

For convective storms, cold pool generated lifting and dynamic pressure forcing work together to produce the observed storm characteristics. The relative importance of each mechanism is dependent on the characteristics of the thermodynamic profile as well as the vertical wind shear profile of the environment in which the storm grows. A convective system may be composed of both ordinary cells and supercells simultaneously, while maintaining a general multicell character. Storm types also have a tendency to change during the lifetime of an event. For instance, an isolated supercell will often evolve into a more multicellular line of ordinary cells over time as the storm-generated cold pool and associated lifting becomes stronger and begins to dominate over the dynamic lifting effects associated with the rotating updraft. In such cases, a supercell is said to have ‘gusted out’ or ‘lined out.’ Convective storms also change character as they move into a different mesoscale environment, or when they interact with each other, as within a squall line.

While convective updraft characteristics can generally be anticipated quite well from environmental thermodynamic and shear profiles, potential downdraft and resulting cold pool characteristics are much more difficult to gauge from environmental conditions. The storm-generated downdraft and cold pool are certainly sensitive to the amount of thermodynamic instability and the distribution of moisture in the environment, but it is also sensitive to the characteristics of the precipitation that is produced within the storm. For instance, a convective cloud that predominantly produces a few large raindrops or hailstones will tend to have a weaker downdraft and cold pool than a cloud that produces a large quantity of smaller drops, due to decreased evaporation rates. Along these lines, supercell storms have been subclassified into high precipitation (HP), classic, and low precipitation (LP) varieties, based on intensity and distribution of the precipitation and the resulting strength of the system-generated cold pool. Many of these factors are discussed in companion chapters within this volume.

Climatology of Convective Storm Types

Ordinary cell and multicell storm systems are commonly observed from the tropics through midlatitudes, whenever thermodynamic instability exists and there is a sufficient

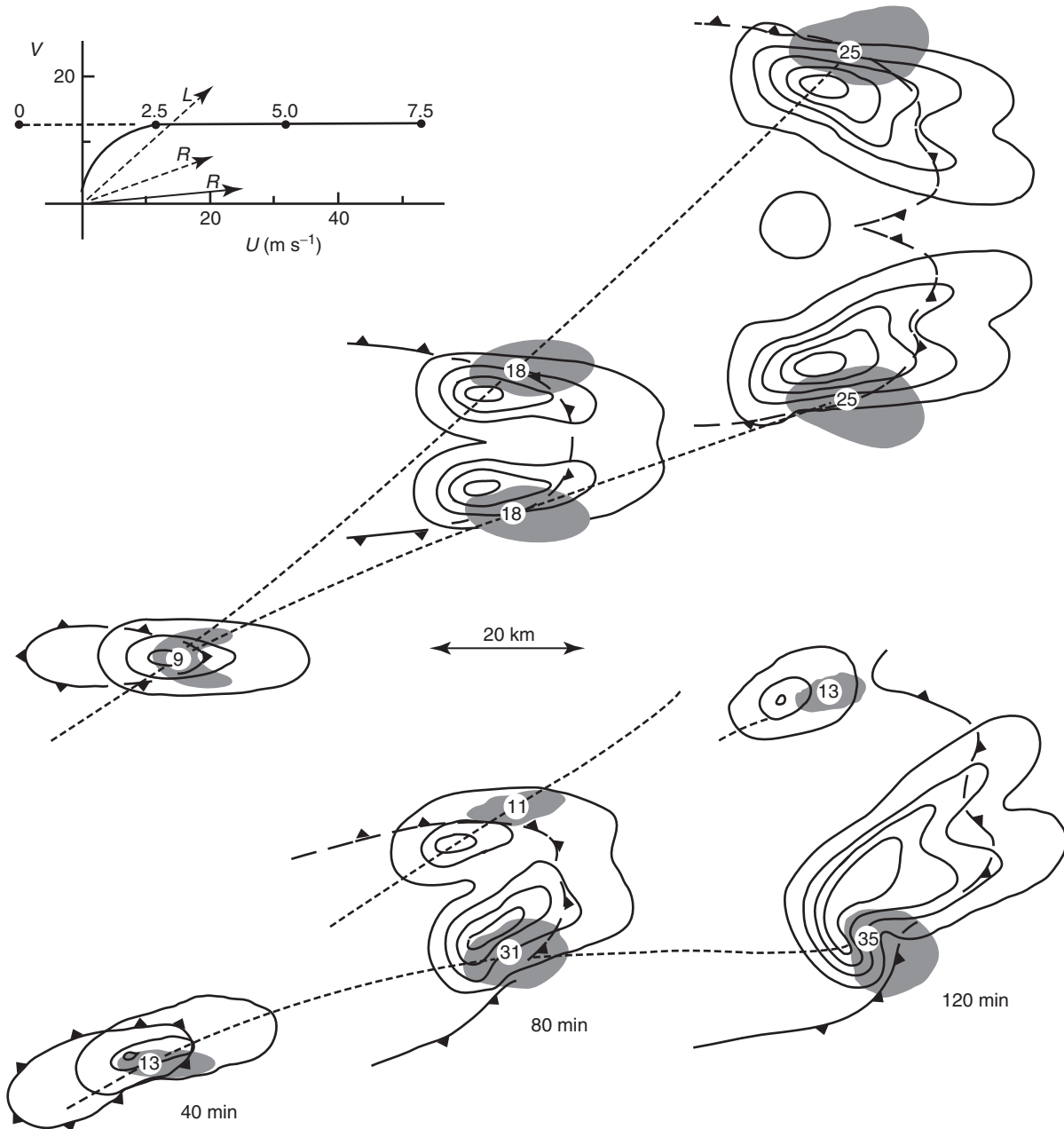


Figure 11 Plan views of numerically simulated convective storms at 40, 80, and 120 min for two environmental wind profiles (displayed at the upper left) having wind shear between the surface and 7.5 km agl. The storm system in the lower portion of the figure evolves in response to the wind profile for which the vertical wind shear vector turns clockwise with height between the ground and 2.5 km (heavy solid line in the hodograph), while the upper system develops when the shear is unidirectional (same wind profile except follow the heavy dashed line below 2.5 km). The plan view depicts the low-level (1.8 km) rainwater field (similar to radar reflectivity) contoured at 2 g kg^{-1} intervals, the midlevel (4.6 km) updraft (shaded regions), and the location of the surface cold pool boundary (barbed lines). The maximum updraft velocity is labeled (in m s^{-1}) within each updraft at each time. The dashed lines track the path of each updraft center. Arrows on the hodograph indicate the supercell propagation velocities for the unidirectional (dashed) and turning (solid) wind shear profiles. Reproduced from Klemp, J.B., 1987. Dynamics of tornadic thunderstorms. *Annual Reviews of Fluid Mechanics* 19, 369–402.

triggering mechanism for the convection. Supercell storms, however, tend to be more limited to midlatitude, continental regions, where sufficient vertical wind shear can exist in association with thermodynamic instability. Supercell storms are especially prevalent in the spring and early summer in the plains and mid-western regions of the United States, where the Gulf of Mexico supplies a source of low-level moisture to

enhance thermodynamic instability, and the frequent passage of synoptic-scale waves offers a source for the vertical wind shear. The frequency of supercell storms and the associated tornadoes in this part of the United States has led to this region being referred to as ‘Tornado Alley.’

Supercell storms can also be embedded within the rain bands of landfalling tropical storms and hurricanes. These

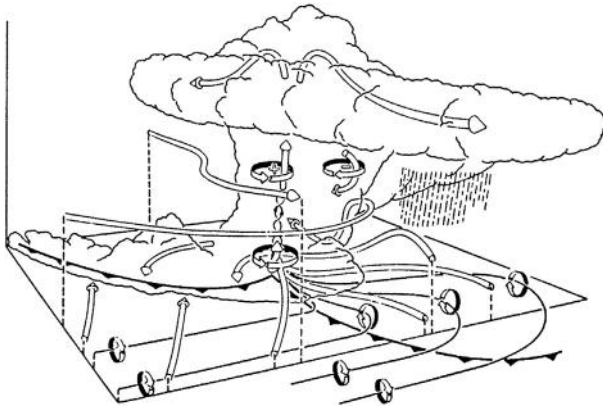


Figure 12 Three-dimensional schematic view of a mature supercell thunderstorm at a stage when low-level rotation is intensifying. The storm, viewed from the southeast, is evolving in westerly environmental wind shear. The cylindrical arrows depict the flow in and around the storm. The thick lines show the low-level vortex lines, with the sense of rotation indicated by the circular ribbon arrows. The heavy barbed line marks the boundary of the cold air beneath the storm. Reproduced with permission from Klemp, J.B., 1987. Dynamics of tornadic thunderstorms. *Annual Reviews of Fluid Mechanics* 19, 369–402.

supercells tend to be shallower than more classic supercells, as minimal instability ($\text{CAPE} < 1000 \text{ J kg}^{-1}$) is available within the associated tropical air mass. However, very strong low-level vertical wind shears are generated as the rain bands come ashore, and very intense (of order 10 m s^{-1}) updrafts can be generated in the lowest 1–2 km agl in such cells, due to the dynamic vertical pressure gradients associated with the rotating updrafts. Such shallow supercells are hypothesized to be the source of tornado outbreaks within landfalling tropical storms and hurricanes.

In the midlatitudes, convection often occurs in the warm sectors of synoptic-scale waves, in association with cold fronts and warm fronts. In the tropics, convective activity is commonly located along the intertropical convergence zone (ITCZ). In both the tropics and midlatitudes, convection also

commonly occurs along sea breeze fronts or in association with topographic features such as mountain ridges. Generally, oceanic convection tends to be weaker than continental convection, due to less thermodynamic instability over oceanic regions. Convective storms can occur at any time of the day or night, when thermodynamic instability and a trigger are available, but severe convection tends to maximize in the late afternoon and evening hours, in response to the enhanced thermodynamic instability associated with diurnal heating.

See also: Mesoscale Meteorology: Bow Echoes and Derecho; Cloud and Precipitation Bands; Density Currents; Gust Fronts; Hail and Hailstorms; Waterspouts. **Weather Forecasting:** Severe Weather Forecasting.

Further Reading

- Browning, K.A., 1977. The Structure and Mechanism of Hailstorms. Hail: A Review of Hail Science and Hail Suppression, Meteorological Monographs, vol. 16. American Meteorological Society, Boston, MA, pp. 1–43.
- Byers, H.R., Braham Jr., R.R., 1949. The Thunderstorm. Supt of Documents. US Government Printing Office, Washington, DC.
- Church, C., Burgess, D., Doswell, C., Davies-Jones, R., 1993. The tornado: its structure, dynamics, prediction, and hazards. *Geophysical Monographs* 79, 637.
- Doswell III, C.A., 1985. The Operational Meteorology of Convective Weather vol. II Storm Scale Analysis. NOAA Technical Memorandum ERL ESG-15.
- Kessler, E., 1986. Thunderstorms: A Social, Scientific, and Technological Documentary. Thunderstorm Morphology and Dynamics, second ed., Vol. 2. [revised and enlarged] University of Oklahoma Press, Norman, OK and London, UK.
- Klemp, J.B., 1987. Dynamics of tornadic thunderstorms. *Annual Reviews of Fluid Mechanics* 19, 369–402.
- Lemon, L.R., 1980. Severe Thunderstorm Radar Identification Techniques and Warning Criteria. NOAA Technical Memorandum, NWS NSSFC-3, Kansas City, MO (NTIS PB81-234809).
- Rotunno, R., Klemp, J.B., Weisman, M.L., 1988. A theory for strong, long-lived squall lines. *Journal of the Atmospheric Sciences* 45, 463–485.
- Weisman, M.L., Klemp, J.B., 1986. Characteristics of Isolated Convective Storms. *Mesoscale Meteorology and Forecasting*. American Meteorological Society, Boston, MA, pp. 331–358.

MESOSPHERE

Contents

Atomic Species in the Mesopause Region

Ionosphere

Metal Layers

Polar Summer Mesopause

Atomic Species in the Mesopause Region

MG Mlynczak, NASA Langley Research Center, Hampton, VA, USA

LA Hunt, Science Systems and Applications Incorporated, Hampton, VA, USA

© Published by Elsevier Ltd.

Synopsis

Atomic oxygen [O] and atomic hydrogen [H] play fundamental roles in the photochemistry and energetics of the Earth's mesopause region between 80 and 105 km. However, these species are exceptionally difficult to measure and must be inferred from observations of related phenomena. Long-term data sets of these two species exceeding 11 years in length are now available from satellite measurements. These observations reveal the magnitude and variability of [O] and [H], and the 11-year solar cycle is evident in their time series. Radiative and energetic constraints on the [O] concentration show the mesopause region energy balance to be consistent with global annual radiative/chemical equilibrium. This unique data set of atomic species in Earth's atmosphere enables basic tests of the radiant and chemical physics in upper atmosphere climate models.

Introduction

The determination of the abundance of atomic oxygen [O] and atomic hydrogen [H] has long been a central focus in the study of the photochemistry, energy balance, and aeronomy of Earth's mesopause region. Nominally taken to encompass the altitude regime between 80 and 105 km (pressures between 10^{-2} and 10^{-4} hPa), mesopause region density is sufficiently low to enable long-lived populations of H and O, which would otherwise completely react or recombine to form molecular species. Mesopause region atomic oxygen is generated through the photolysis of O₂ at ultraviolet wavelengths, primarily in the Schumann-Runge continuum, the Schumann-Runge bands, and at Lyman- α (121.6 nm) wavelength. Atomic hydrogen is primarily generated by photolysis of H₂O at Lyman- α .

Atomic species O and H play fundamental roles in the energy balance of the mesopause region. They influence both the heating and cooling of the region. Specifically, during the photolysis of O₂ at ultraviolet wavelengths, a substantial fraction of the energy of the absorbed photon is used to break the chemical bond of the O₂ molecule. This energy does not immediately show up as heat but rather is carried away in

chemical potential energy of the two O-atom products of photolysis. Because of the low mesopause region density, recombination of atomic oxygen (to form O₃ or O₂) will occur sometime later, likely at a different location than where the original photon deposition/photolysis occurred. When recombination does occur, the large amount of chemical energy carried by the O-atoms is released, heating the atmosphere. The existence of atomic hydrogen provides a further catalyst for heating of the mesopause region through the highly exothermic reaction of H and O₃ (Mlynczak and Solomon, 1991). This reaction is the single largest source of heat in the mesopause region. In fact, heating by exothermic chemical reactions exceeds that deposited directly by solar radiation throughout much of the mesopause region, as shown in Figure 1 (Mlynczak and Solomon, 1993).

Atomic oxygen also influences the rates of radiative emission by numerous infrared-active molecules in the mesopause region including CO₂, O₃, and OH. Due to the low density in the mesopause region, the vibrational populations of these molecules depart from local thermodynamic equilibrium (LTE) and are said to be in non-LTE. Under non-LTE, the vibrational populations are no longer given by a Boltzmann

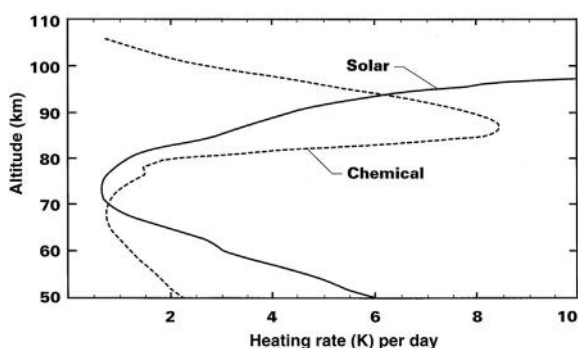


Figure 1 Rates of solar heating and heating due to exothermic chemical reactions, illustrating that exothermic reactions are the largest source of heat over much of the mesopause region. After Mlynczak, M.G., Solomon, S., 1993. A detailed evaluation of the heating efficiency in the middle atmosphere. *Journal of Geophysical Research* 98 (D6), 10517–10541. <http://dx.doi.org/10.1029/93JD00315>.

distribution at the local kinetic temperature. Collisions, radiations (both solar and terrestrial), and chemical reactions may determine the vibrational populations under non-LTE. Atomic oxygen in particular is very effective in achieving energy exchange (via collisions) between the translational energy field and the internal vibrational states of CO_2 (Rodgers et al., 1992), O_3 (West et al., 1976), and OH (Kalogerakis et al., 2011). Atomic oxygen is therefore a necessary input to the computation of infrared radiances, fluxes, and cooling rates in the mesopause region.

Knowledge of the atomic oxygen and atomic hydrogen abundances is essential to achieve an understanding of the basic structure and overall composition of the mesopause region. The challenge with measuring atomic oxygen and atomic hydrogen is that neither species has radiative emission features in the mesopause region that can be readily observed from instruments on orbiting satellites. Atomic oxygen has two fine structure lines in the far infrared, one at 63 μm and one at 145 μm . The 63- μm line has been measured from space by the CRISTA instrument carried on the Space Shuttle (Grossman et al., 2000), from suborbital rockets (Grossmann and Vollmann, 1997), and from high altitude balloons (Mlynczak et al., 2004). However, the abundance of atomic oxygen in the thermosphere above is so large that it is not possible to ‘see’ the fine structure emission at the mesopause from an orbiting satellite. Resonance lamp techniques have been used to observe atomic hydrogen from suborbital rockets (Sharp and Kita, 1987).

There is simply no traditional method to observe mesopause region atomic oxygen or atomic hydrogen from orbiting satellites. There are, however, techniques to infer the atomic oxygen and atomic hydrogen concentrations from measurements of ozone and radiative emission from the OH radical (e.g., Good, 1976). The Sounding of the Atmosphere using Broadband Emission Radiometry (SABER) instrument on the NASA Thermosphere-Ionosphere-Mesosphere Energetics and Dynamics (TIMED) satellite has been observing the terrestrial mesosphere and lower thermosphere since January 2002. SABER makes observations of the O_3 abundance and the OH emission at 2.0 μm specifically for the derivation of O and H,

day and night, in the mesopause region. These techniques are discussed below. First, the question of the allowed range of atomic oxygen will be examined. This is essential for understanding the atomic oxygen concentration in general and for providing a basis for validating atomic oxygen concentrations inferred by other means.

Energetically and Radiatively Allowed Range of Atomic Oxygen

Upper Limits on the Atomic Oxygen Concentration

The challenges associated with making direct measurements of atomic oxygen, and the uncertainties associated with inferring it from related measurements, prompt the question as to whether methods exist by which the O concentration can be constrained. Independent validation of measured atomic oxygen is essential because of the importance of O in the energy balance and photochemistry of the mesopause region.

Mlynczak et al. (2013b,c) have presented approaches by which the maximum and minimum vertical profile of global annual mean concentration of atomic oxygen can be determined. These approaches are based on the consideration of global annual mean energy balance and on observations of emission from the highly vibrationally excited hydroxyl (OH) molecule in the mesopause region, the latter analyzed in the limit of complete radiative deactivation. These approaches place fundamental constraints on the maximum and minimum global annual mean atomic oxygen concentration independent of the standard techniques used to derive and infer the atomic oxygen concentration.

The first approach proceeds from considerations of energy balance and the role that exothermic chemical reactions play in the heat budget of the mesopause region. Following Mlynczak et al. (2013b), the global annual mean, energy balance at a given altitude or pressure level in the mesopause region can be written as:

$$H(\text{O}_2) + H(\text{O}_3) + H(\text{C}) + H(\text{D}) = C(\text{IR}) \quad [1]$$

In eqn [1], $H(\text{O}_2)$ and $H(\text{O}_3)$ are the rates of heating due to the absorption of solar ultraviolet radiation by O_2 and O_3 , $H(\text{C})$ is the heating due to exothermic chemical reactions, $H(\text{D})$ is the heating rate due to dynamical processes such as gravity wave breaking, and $C(\text{IR})$ is the rate of radiative cooling due to infrared (IR) emission. Equation [1] neglects some small heating and cooling terms such as heating due to absorption of near-IR solar radiation by CO_2 and cooling by emission of infrared radiation by H_2O . These small terms are of comparable magnitude and are neglected as they do not influence the derivation of the radiatively constrained atomic oxygen.

To proceed, two points are noted. First, the heating due to exothermic chemical reactions $H(\text{C})$ and the heating due to absorption of solar ultraviolet radiation by ozone $H(\text{O}_3)$ can be expressed as functions of atomic oxygen. Equation [1] can be rewritten, neglecting $H(\text{D})$, as:

$$F(\text{O}) = C(\text{IR}) - H(\text{O}_2) \quad [2]$$

Here, $F(\text{O})$ represents the functional form of the sum of $H(\text{O}_3)$ and $H(\text{C})$ that is only a function of atomic oxygen. As shown in Mlynczak et al. (2013b), $F(\text{O})$ is a simple quadratic

equation whose coefficients depend only on standard reaction rate coefficients and enthalpies. The second point is that by neglecting $H(D)$, the O derived from eqn [2] represents an upper, radiatively allowed limit of atomic oxygen, constrained by the rates of radiative cooling by CO_2 and solar heating due to absorption by O_2 . The radiatively constrained atomic oxygen derived from $F(O)$ is independent of any of the standard techniques used to infer atomic oxygen that are discussed below. As such it not only represents an upper limit on the global annual mean atomic oxygen concentration in the mesopause region, but it also serves as an independent validation of the global mean atomic oxygen derived by other methods. The true atomic oxygen cannot be larger than that derived from eqn [2]. Otherwise, the $H(C)$ term would be too large and the mesopause region would perpetually warm due to the heat deposited from exothermic chemical reactions.

Equation [2] is evaluated using standard SABER data products of carbon dioxide radiative cooling rates and heating rates due to absorption of ultraviolet radiation by O_2 , as given in Mlynczak et al. (2013b). Figure 2 shows the radiatively constrained global annual mean atomic oxygen concentration derived from SABER for 2004 and 2008. The concentrations decrease from 2004 to 2008 due to the declining solar cycle over those years. Peak, radiatively allowed atomic oxygen concentrations near 95 km are approximately $5 \times 10^{11} \text{ cm}^{-3}$.

Lower Limits on the Atomic Oxygen Concentration

A technique for determining a radiatively allowed lower limit on the global annual mean atomic oxygen concentration will now be reviewed, following Mlynczak et al. (2013c). At night, the OH radical is produced by the reaction of H and O_3 , which is very exothermic, and produces highly vibrationally excited OH up to the $v = 9$ state. The OH molecule radiates strongly, particularly in the infrared, where the radiative lifetime of the highest lying vibrational states ($v = 8$ and $v = 9$) is about 8 ms.

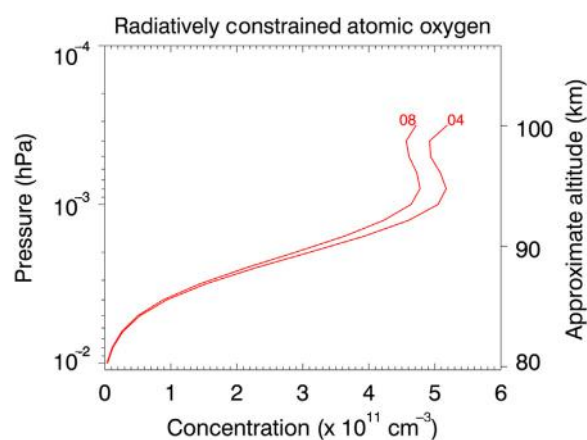


Figure 2 Global annual mean radiatively constrained atomic oxygen concentration profiles derived from SABER for 2004 and 2008. After Mlynczak, M.G., et al., 2013b. Radiative and energetic constraints on the global annual mean atomic oxygen concentration in the mesopause region. *Journal of Geophysical Research – Atmospheres* 118, 5796–5802. <http://dx.doi.org/10.1002/jgrd.50400>.

These highly vibrationally excited states are also physically quenched in collisions with O_2 , N_2 , and even O . Assuming photochemical equilibrium between the production of ozone by recombination of O and O_2 , and loss of ozone by reaction of H and O_3 , a relatively simple relation between the observed emission rate of highly vibrationally excited $OH(v)$, the rates of loss of $OH(v)$ by collisions and radiation, and the atomic oxygen concentration is established. This is the basic technique employed by SABER for deriving the atomic oxygen concentration at night from measurements of $OH(v)$ emission in a channel near $2 \mu\text{m}$ (Mlynczak et al., 2013a).

The observed rate of $OH(v)$ emission depends on three factors: the atomic oxygen concentration, the rate of spontaneous emission of radiation by the $OH(v)$, and the rate at which collisions between $OH(v)$ and other species quench the vibrational states to lower vibrational levels including the ground state, or cause reaction and loss altogether of the $OH(v)$. The rates of spontaneous emission, given by the Einstein A coefficients for each transition, are fixed – an excited state will always attempt to radiate with a lifetime specified by the A coefficient. However, for an observed $OH(v)$ emission rate, the derived atomic oxygen depends directly on the rates of quenching of the $OH(v)$ by collisions with O_2 , N_2 , and O . Specifically, the derived O is proportional to the strength of quenching for a specific observed $OH(v)$. Higher rates of collisional quenching require more atomic oxygen and vice versa. In the limit of no collisional quenching, in which the $OH(v)$ is assumed to relax solely by the spontaneous emission of radiation, the minimum atomic oxygen concentration will be derived.

Mlynczak et al. (2013c) applied this principle to the SABER observations of $OH(v)$ at night in order to derive a radiatively

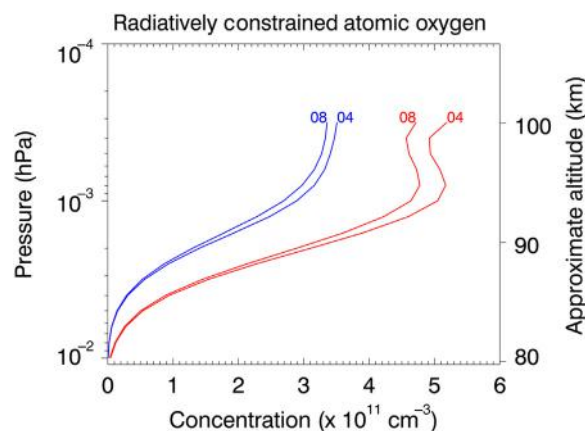


Figure 3 Minimum (blue curves) and maximum (red curves) radiatively allowed atomic oxygen concentrations for 2004 and 2008, derived from SABER. After Mlynczak, M.G., et al., 2013b. Radiative and energetic constraints on the global annual mean atomic oxygen concentration in the mesopause region. *J. Geophys. Res. Atmos.* 118, 5796–5802. <http://dx.doi.org/10.1002/jgrd.50400>; Mlynczak, M.G., Hunt, L.A., Marshall, B.T., Mertens, C.J., Russell III, J.M., Siskind, D., Thompson, R.E., Gordley, L.L., 2013c. Radiative constraints on the minimum atomic oxygen concentration in the mesopause region. *Geophysical Research Letters* 40, 3777–3780. <http://dx.doi.org/10.1002/grl.50725>.

constrained minimum atomic oxygen profile in the mesopause region. **Figure 3** shows the minimum radiatively allowed atomic oxygen (blue curves) along with the maximum radiatively allowed atomic oxygen (red curves), for the years 2004 and 2008. These are the global annual mean values in both the cases. The true global annual mean atomic oxygen concentration must fall between the blue and the red curves.

The peak minimum atomic oxygen concentrations shown in **Figure 3** near 95 km are about $3 \times 10^{11} \text{ cm}^{-3}$. This value is only about 40% smaller than the radiatively allowed maximum atomic oxygen. The conclusion is that the allowed range of global annual mean atomic oxygen is rather small, about 3×10^{11} to $5 \times 10^{11} \text{ cm}^{-3}$ at the peak. This result places fundamental limitations on the chemistry and energetics in numerical models of the middle and upper atmosphere. It also brackets the range for observations of the atomic oxygen concentration, which will now be discussed in detail.

Derivation of Atomic Oxygen from Measurements of OH(ν) and O₃

The results presented in Section [Energetically and Radiatively Allowed Range of Atomic Oxygen](#) effectively bracket the range of allowed values of the global annual mean vertical profile of atomic oxygen in the mesopause region. In this section, the approach for deriving atomic oxygen from related measurements will be given. Because the lifetime of atomic oxygen is long (approximately days to weeks) in the mesopause region, derivations of O both day and night are required. To this end, the SABER experiment was designed to derive atomic oxygen day and night, using different approaches for each, which are described in detail by [Mlynczak et al. \(2013a\)](#). These are now summarized briefly here.

For the day, photochemical equilibrium between production of ozone by recombination of O and O₂ is taken to be in balance with destruction of ozone by ultraviolet photolysis. Specifically,

$$J[\text{O}_3] = k_2[\text{O}][\text{O}_2][M] \quad [3]$$

In [eqn \[3\]](#), J is the photolysis rate (s^{-1}) of ozone, primarily entirely in the Hartley band from approximately 200 to 350 nm, M is the total number density (cm^{-3}), and k_2 is the recombination rate coefficient ($\text{cm}^6 \text{ s}^{-1}$).

For night, SABER measurements of the OH emission near 2.0 μm are used to derive O. SABER observes the $\nu = 8$ and $\nu = 9$ states of OH ([Mast et al., 2013](#)) that are formed directly upon reaction of H and O₃ to derive the atomic oxygen concentration. At night, production of ozone by recombination of O and O₂ is balanced by the loss of ozone by reaction of H and O₃. The photochemical balance for this process is written as:

$$k_4[\text{H}][\text{O}_3] = k_2[\text{O}][\text{O}_2][M] \quad [4]$$

where k_4 is the reaction rate coefficient of H and O₃. The notations k_4 and k_2 are used for consistency with [Mlynczak et al. \(2013a\)](#). The product $k_4[\text{H}][\text{O}_3]$ is directly related to the

observed rate of emission from the OH ($\nu = 8, 9$) states, which enables the O atom concentration to be derived directly from measurements of the OH(ν) emission. The O₂ and the total number densities are derived from SABER measurements of pressure and temperature. This technique for derivation of atomic oxygen was developed by [Good \(1976\)](#).

Derivation of Atomic Hydrogen from Measurements of OH(ν)

SABER measurements of the OH(ν) emission rate and of ozone can be used to derive the atomic hydrogen concentration at night. Once the atomic oxygen concentration is derived from [eqn \[4\]](#), this equation is used to directly solve for H by back substitution of the derived O into the right-hand side of the equation. SABER measurements of emission from ozone at 9.6 μm are used to provide the O₃ concentration on the left-hand side of [eqn \[4\]](#). SABER temperature and pressure data are used to compute the total number density and to compute the temperature dependence of the rate coefficients k_2 and k_4 , allowing [eqn \[4\]](#) to be solved directly for the H concentration. The SABER approach also follows directly from [Good \(1976\)](#).

Observations of Atomic Oxygen and Atomic Hydrogen

In this section, we present observations of the O and H atoms derived from SABER observations as discussed above. Zonal annual and global annual means are given because of the importance of these species in the mesopause region energy balance, and the strictest test in this regard is the global annual mean value ([Mlynczak et al., 2013b](#)). The SABER atomic oxygen and atomic hydrogen data are however available on a profile-by-profile basis, approximately 1500 profiles per day, since January 2002. The excellent precision (low radiometric noise) of the SABER radiometer ([Mlynczak et al., 2013a](#)) enables retrieval of O and H for each individual limb radiance profile measured by the instrument. The individual species vertical profiles are of high quality and may be used to study the atomic species in the mesopause region on daily and longer timescales. These data are publicly available via the link to the SABER data website given at the end of this article.

Day Atomic Oxygen

Shown in **Figure 4** are SABER day zonal annual mean atomic oxygen concentrations (left panels) and SABER day zonal annual mean volume mixing ratios (right panels) for the years 2002, 2008, and 2012. These years span the time from near the maximum to the minimum of solar cycle 23 (2002–08) and well into the rise into solar cycle 24 (2012). The volume mixing ratios are the base 10 logarithm of the actual mixing ratio due to the range of mixing ratio (three orders of magnitude) encountered in the mesopause region. Both the concentrations and mixing ratios are largely zonally symmetric in the day. Peak atomic oxygen concentrations in

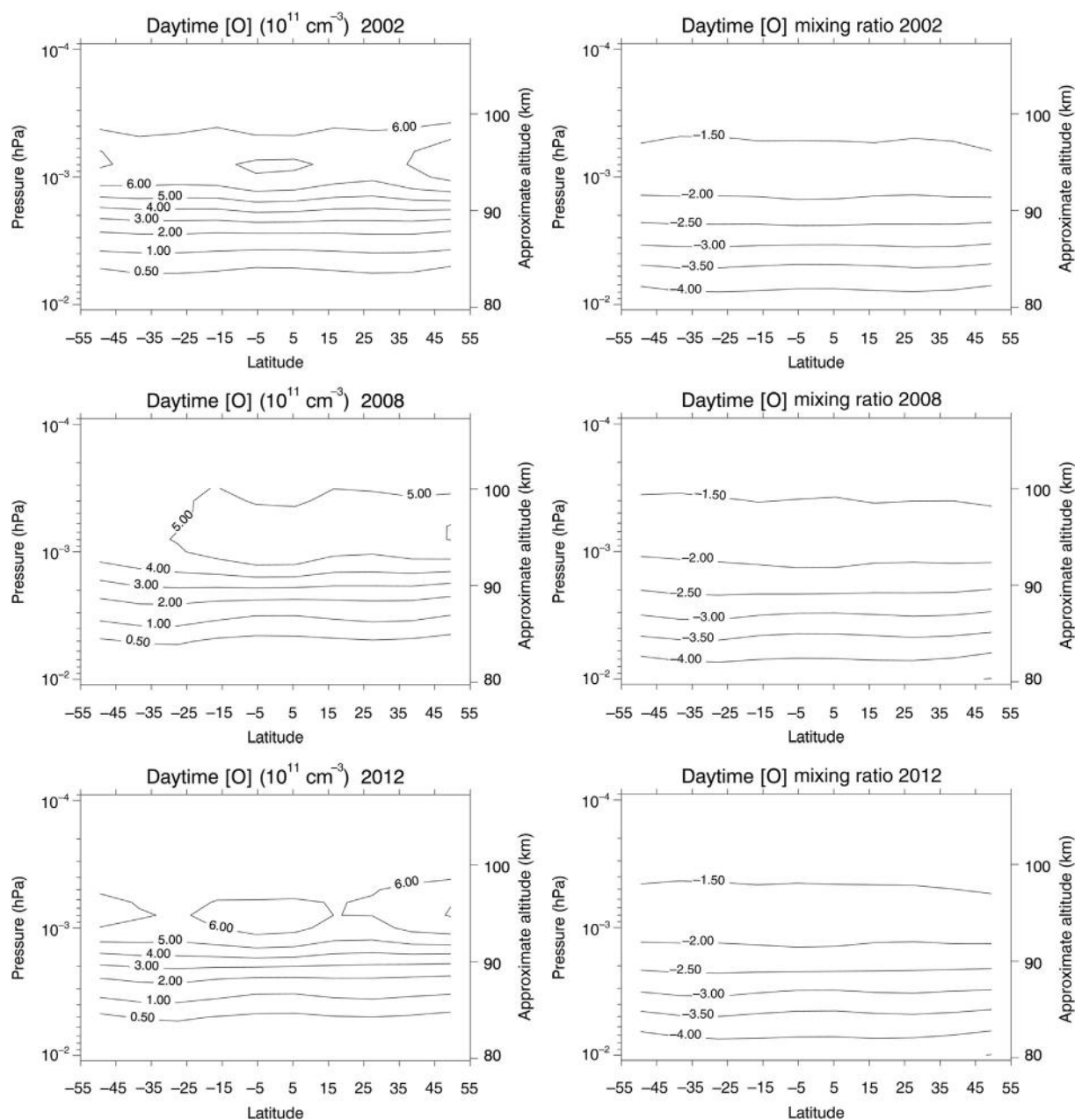


Figure 4 Day zonal annual mean atomic oxygen concentrations (cm^{-3} , left panels) and volume mixing ratios (right panels) for 2002, 2008, and 2012. Mixing ratios are base 10 logarithm of the actual volume mixing ratio. Day atomic oxygen derived from measurements of the ozone concentration. After Mlynarczyk, M.G., et al., 2013a. Atomic oxygen in the mesosphere and lower thermosphere derived from SABER: algorithm theoretical basis and measurement uncertainty. *Journal of Geophysical Research – Atmospheres* 118, 5724–5735. <http://dx.doi.org/10.1002/jgrd.50401>.

excess of $6 \times 10^{11} \text{ cm}^{-3}$ are observed near 95 km in 2002 and 2012. The observed peak concentrations in 2008 are smaller by $\sim 20\%$, and are likely a result of the occurrence of solar minimum during that time.

Figure 5 shows the day global annual mean atomic oxygen concentrations for 2002, 2008, and 2012 derived from the zonal mean values in Figure 4. The variability in day atomic oxygen is clearly evident with the 2008 global annual mean

peak concentration approximately 20% smaller than in 2002 and approximately 10% smaller than in 2012. The progression of the solar cycle is clearly illustrated in Figure 6, which shows the difference in the day global annual mean atomic oxygen concentration in each year from that in 2002. The values are negative indicating that the day global annual mean atomic oxygen is smaller each year after 2002. The largest differences occur in 2008–09 during solar minimum conditions.

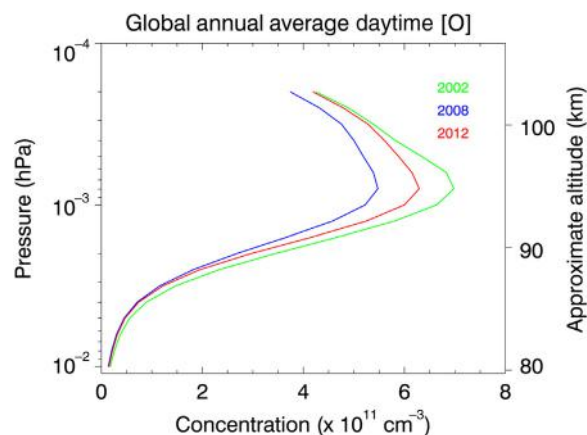


Figure 5 Day global annual mean atomic oxygen concentrations (cm^{-3}) for 2002, 2008, and 2012. Minimum concentrations occur in 2008 during the time period associated with a minimum of solar activity.

Night Atomic Oxygen

Figure 7 shows the night zonal annual mean atomic oxygen concentrations (left panels) and (base 10 logarithm) volume mixing ratios (right panels) for 2002, 2008, and 2012. The night atomic oxygen concentration exhibits a much different spatial pattern than in the day, with maxima occurring in each hemisphere near 35° latitude. This pattern may be associated with tidal activity. The night volume mixing ratios also exhibit variations with latitude, unlike the day values, which are also suggestive of a tidal origin. Nevertheless, the day and night atomic oxygen concentrations are quite similar, especially in the global mean, as noted by Mlynarczyk et al. (2013a). The concordance of the day and night concentrations, in the annual means, is remarkable, given the two disparate techniques (and

associated assumptions and parameters) used in their derivation.

Figure 8 shows the global annual mean atomic oxygen concentrations at night for 2002, 2008, and 2012. As with the day concentrations, night minimum occurs in 2008, and the relative values in 2002 and 2012 are also consistent with the day. Figure 9 shows the difference in the night global annual mean atomic oxygen concentration from 2002 for each year from 2003 to 2012. The values (as in Figure 6) are negative, indicating that the night atomic oxygen is also smaller in all years after 2002. The largest difference is again seen to occur in 2008. Of particular interest is that the differences at night in the global annual mean atomic oxygen concentration are only about one-half the magnitude of those in the day shown in Figure 6. This result suggests that the influence of the solar cycle may be different during the day than at night. In particular, since atomic oxygen is subject to dynamical transport, these results suggest that solar activity influences both the chemistry and the dynamics, resulting in a different ‘solar cycle’ response of atomic oxygen for day and for night.

Comparison between Derived and Radiatively Constrained Atomic Oxygen

The SABER atomic oxygen data set is unique in its length and in the agreement between the day and the night atomic oxygen derived from two different techniques. In this section, the SABER-derived atomic oxygen is compared with the radiatively allowed atomic oxygen concentrations given earlier. The purpose is to provide a measure of validation of the SABER-derived atomic oxygen and to place the derived abundance in context with the global annual mean energy balance of the mesopause region.

Shown in Figure 10 are the SABER global annual mean atomic oxygen concentrations (average of day and night values, labeled ‘SABER’), along with the radiatively constrained

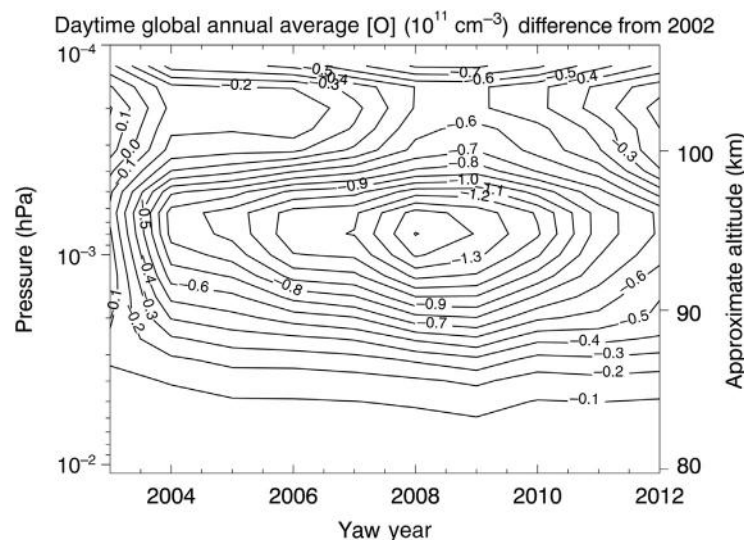


Figure 6 Time series of the difference in the global annual mean atomic oxygen concentration from 2002 through 2012. This figure shows the progression of the solar cycle in the daytime. Negative contour values indicate a decrease of atomic oxygen from 2002.

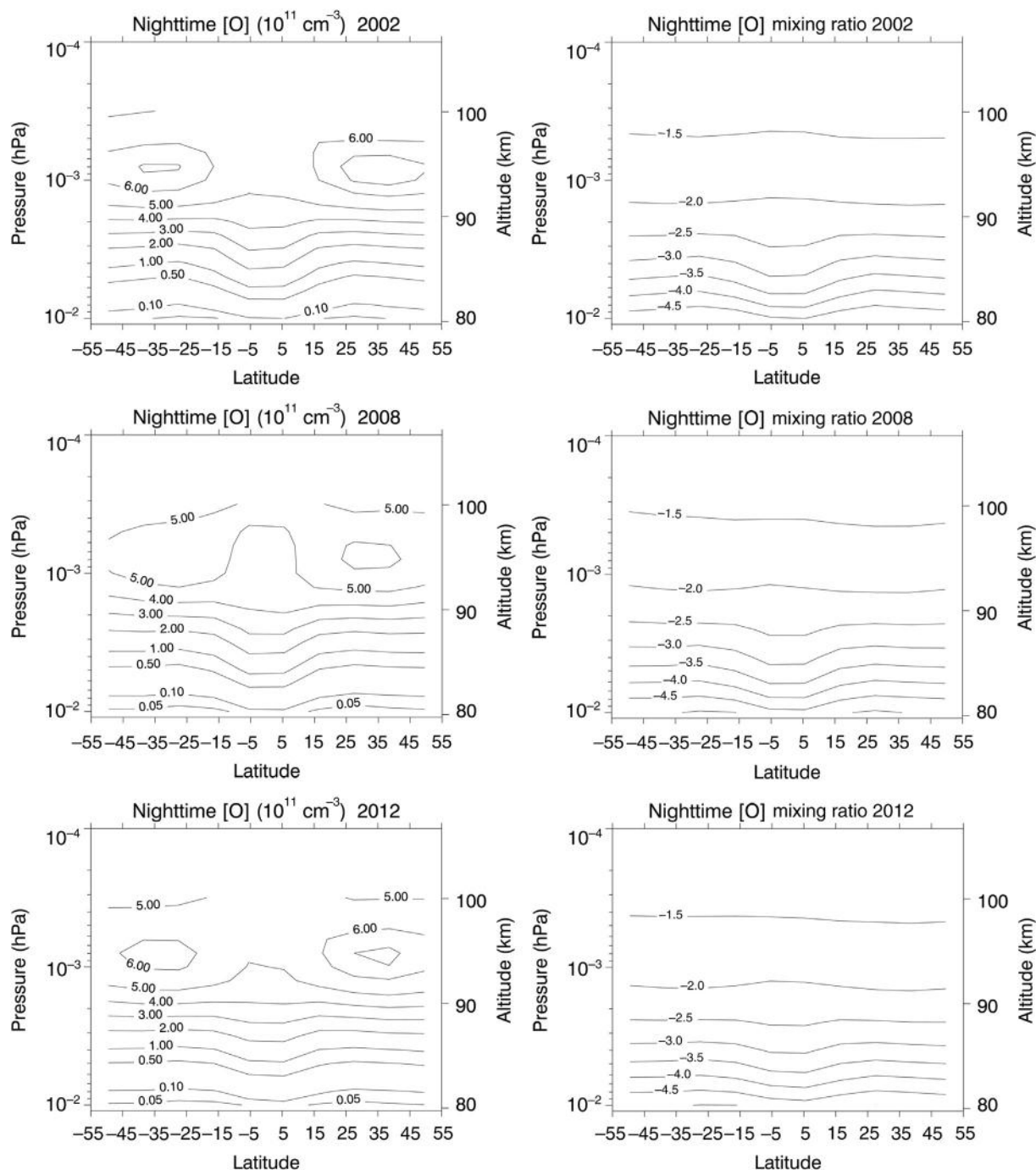


Figure 7 Night zonal annual mean atomic oxygen concentrations (cm^{-3} , left panels) and volume mixing ratios (right panels) for 2002, 2008, and 2012. Mixing ratios are the base 10 logarithm of the actual volume mixing ratio. Night atomic oxygen derived from measurements of $\text{OH}(v)$. After Mlynczak, M.G., et al., 2013a. Atomic oxygen in the mesosphere and lower thermosphere derived from SABER: algorithm theoretical basis and measurement uncertainty. *Journal of Geophysical Research – Atmospheres* 118, 5724–5735. <http://dx.doi.org/10.1002/jgrd.50401>.

minimum (labeled 'RC min') and maximum concentrations (labeled 'RC max'), for the year 2012. The SABER-derived atomic oxygen is much closer to the maximum allowed radiative constrained values than to the minimum allowed values. Above 92 km, the SABER values are approximately 15% larger than the maximum allowed radiatively constrained atomic

oxygen. The pattern is observed for all years although only 2012 is shown here. The uncertainty in both the SABER-derived and the maximum radiative constrained values is $\sim 20\%$ (Mlynczak et al., 2013a,b), so in essence, the SABER atomic oxygen agrees with the maximum radiative constrained atomic oxygen throughout the mesopause region.

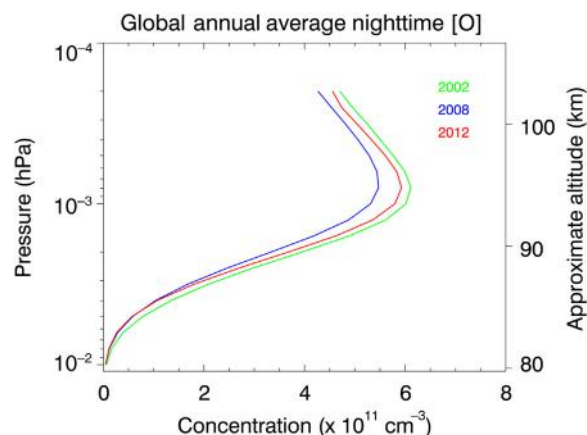


Figure 8 Night global annual mean atomic oxygen concentrations (cm^{-3}) for 2002, 2008, and 2012. Minimum night atomic oxygen occurs in 2008 during a period of minimum solar activity.

There is one major consequence of the SABER atomic oxygen concentrations at the global mean values shown in [Figure 10](#). Specifically, the data imply that the global annual mean mesosphere is near radiative/chemical equilibrium, and that heating or cooling due to dynamical processes such as gravity wave breaking are minimal in the global annual mean. This is because the *maximum* radiatively constrained atomic oxygen is consistent with the difference between radiative cooling by carbon dioxide and solar heating associated with molecular oxygen. However, as discussed in [Mlynczak et al. \(2013a\)](#), the uncertainty in the SABER-derived atomic oxygen concentration is $\sim 20\%$, and by inference, so is the global mean heating rate due to exothermic chemical reactions and absorption of solar ultraviolet radiation by ozone. This uncertainty leaves substantial room for other processes such as turbulent dissipation of breaking gravity waves to be a significant component of the global mesopause region energy budget,

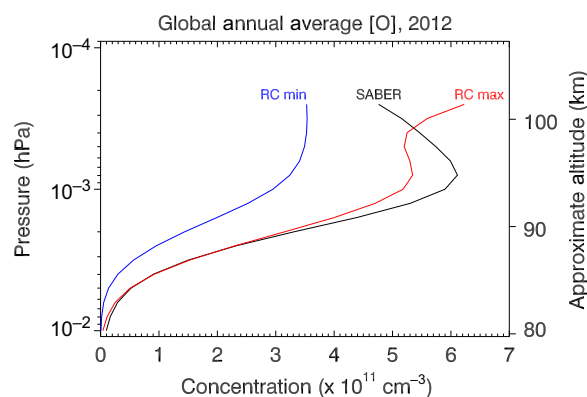


Figure 10 Comparison of SABER global annual mean atomic oxygen concentrations (day and night average) with radiatively constrained maximum (RC max) and radiatively constrained minimum (RC min) concentrations for 2012. After Mlynczak, M.G., et al., 2013b. Radiative and energetic constraints on the global annual mean atomic oxygen concentration in the mesopause region. *J. Geophys. Res. Atmos.* 118, 5796–5802. <http://dx.doi.org/10.1002/jgrd.50400>; Mlynczak, M.G., Hunt, L.A., Marshall, B.T., Mertens, C.J., Russell III, J.M., Siskind, D., Thompson, R.E., Gordley, L.L., 2013c. Radiative constraints on the minimum atomic oxygen concentration in the mesopause region. *Geophysical Research Letters* 40, 3777–3780. <http://dx.doi.org/10.1002/grl.50725>.

and remains the single largest uncertainty in quantifying the energy balance of the mesopause region. Atomic oxygen may need to be known to be better than 10% in order to further improve and quantify the role of dynamical processes on the mesopause region energy budget.

Night Atomic Hydrogen

Shown in [Figure 11](#) (left panels) are the night zonal annual mean concentrations of atomic hydrogen derived from SABER

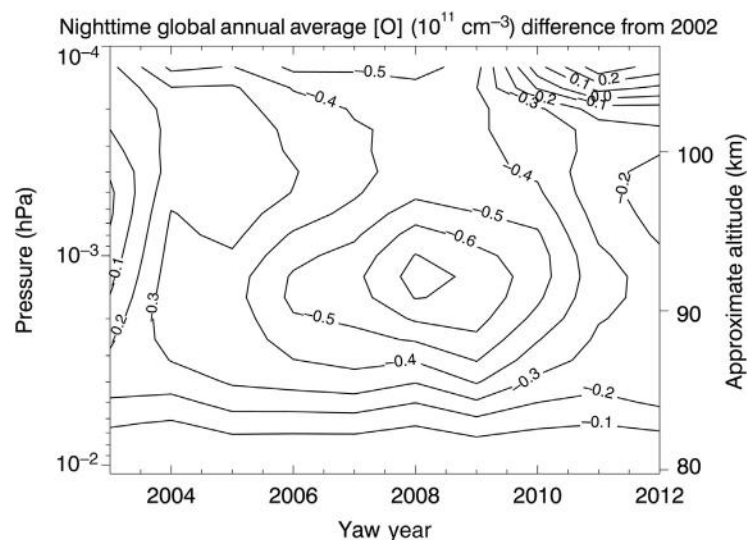


Figure 9 Time series of the difference in the global annual mean atomic oxygen concentration from 2002 through 2012. This figure shows the progression of the solar cycle at night. Peak concentration differences at night are approximately half of that during daytime (see [Figure 6](#)).

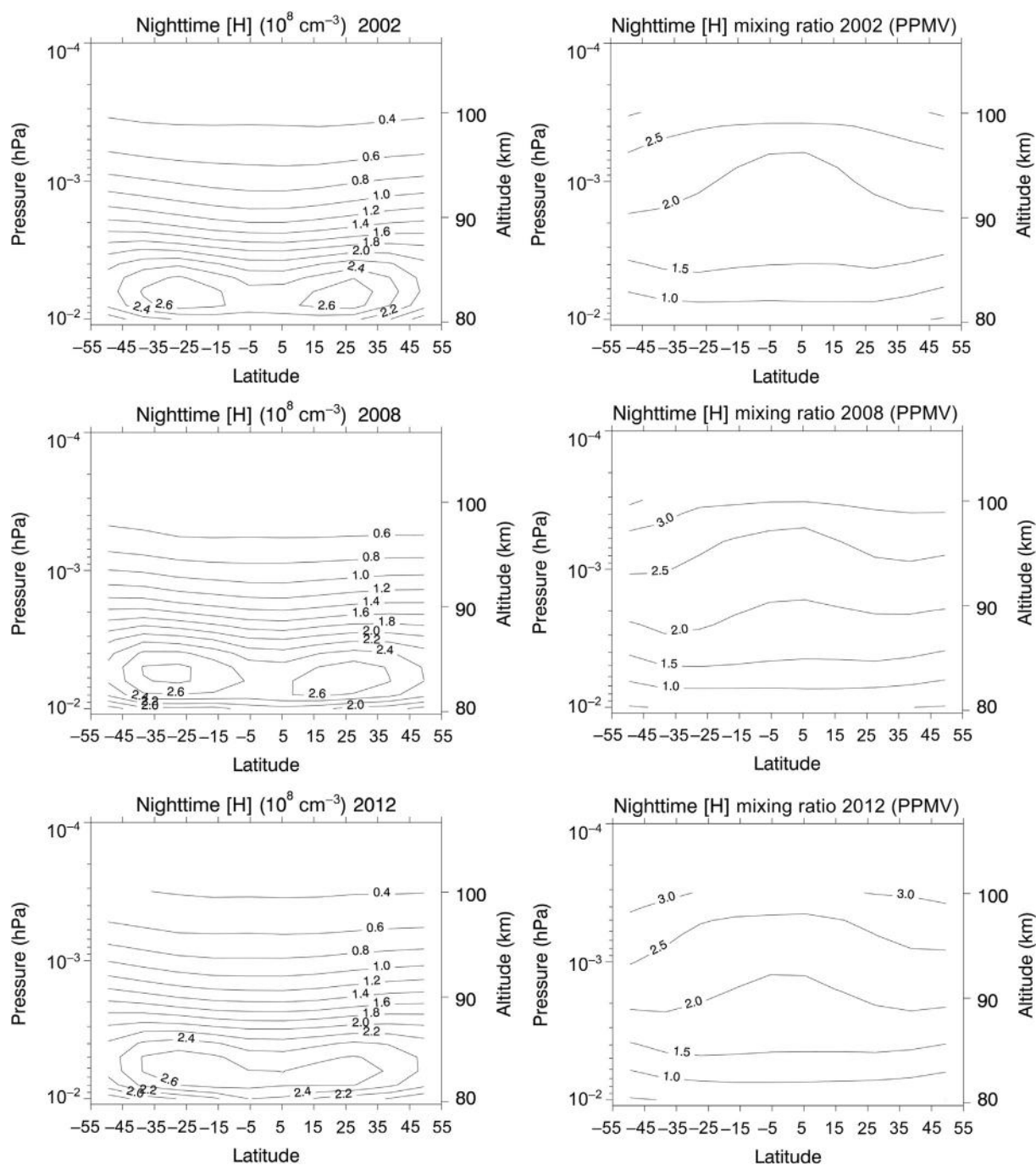


Figure 11 SABER-derived night zonal annual mean atomic hydrogen concentrations (cm^{-3} , left panels) and volume mixing ratios (parts per million, right panels) for 2002, 2008, and 2012.

for the years 2002, 2008, and 2012. The concentrations are in units of 10^8 cm^{-3} . The night H exhibits a peak concentration of $\sim 2.6 \times 10^8 \text{ cm}^{-3}$ in the lower part of the mesopause region around 84–85 km. Above 87 km, the H concentration is relatively uniform with latitude and continually decreases with altitude. The right panels of Figure 11 show the volume mixing ratio of H in parts per million. The mixing ratio is observed to increase with altitude, and is largest near the equator on a given pressure surface.

Figure 12 shows the night global annual mean H concentrations for 2002, 2008, and 2012. In contrast to atomic oxygen, the global annual mean atomic hydrogen concentrations are observed to *increase* from 2002 to 2008, and *decrease* from 2008 to 2012, i.e., directly out of phase with the solar cycle during this time. This is further evident in Figure 13 in which the time series of the difference in the night global annual mean H concentration for each year from that in 2002. In particular, there is about a 15% increase in the global annual

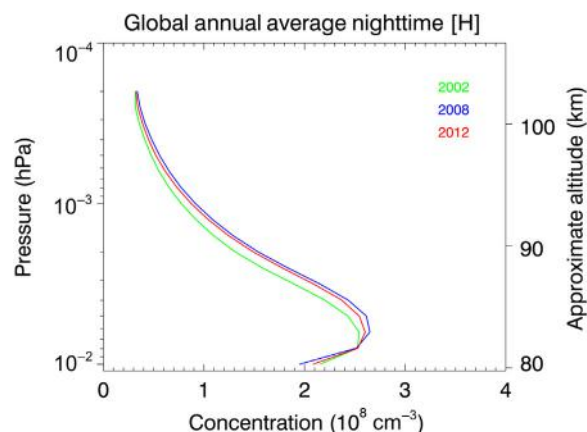


Figure 12 SABER-derived night global annual mean atomic hydrogen concentrations (cm^{-3}) for 2002, 2008, and 2012. Concentrations in 2008 are largest despite occurring during the time of minimal solar activity.

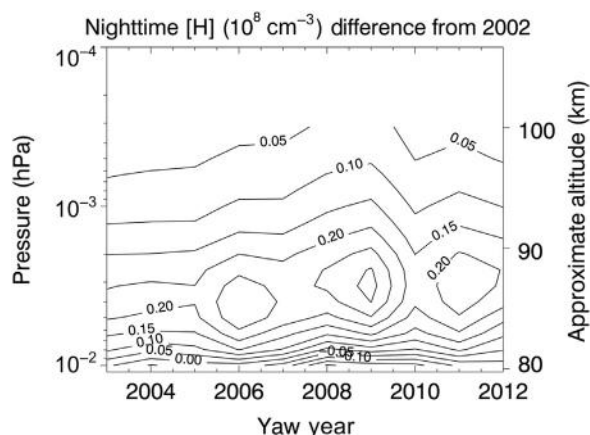


Figure 13 Time series of the difference in the SABER-derived global annual mean atomic hydrogen concentration from 2002 through 2012. This figure shows the progression of the solar cycle at night. Positive values indicate an increase in atomic hydrogen from 2002, peaking during 2008–09, which corresponds to solar minimum conditions.

H concentration from 2002 to 2009 at ~ 87 km altitude. Other SABER parameters such as O_3 and temperature, in addition to atomic oxygen, are positively correlated with the solar cycle, but H is not. This effect of increasing H with decreasing solar activity has also been observed in data from the SCIAMACHY experiment (Kaufmann et al., 2013).

Summary and Future Directions

The SABER data set of atomic oxygen and atomic hydrogen now covers nearly 12 years in length, starting from January 2002. It is presently the longest continuous set of observations of atomic species in the mesopause region. The TIMED mission has been approved for operations until at least September 2015, with

possibility for extension beyond that date. The SABER atomic species data begin near solar maximum in 2002, continue through the prolonged solar minimum in 2008, and go well into the onset and progression of solar cycle 24. Over 6 million individual profiles of atomic oxygen and atomic hydrogen are now stored on the SABER data server for scientific research.

The SABER data have also been used to provide fundamental new understanding of atomic oxygen concentrations in the mesopause region. The concept of radiatively allowed global annual mean maximum and minimum atomic oxygen concentration profiles now provides fundamental limits within which observed and modeled atomic oxygen concentrations must fall (Mlynczak et al., 2013b,c). The range is surprisingly narrow, with roughly a 40% difference between the maximum and minimum allowed peak atomic oxygen concentration. Prior to these concepts there was no direct way to validate or constrain modeled or observed values of atomic oxygen.

A major issue going forward is resolving the energy balance of the mesopause region, which is a key motivation and focus of the SABER experiment on the TIMED mission (Mlynczak, 1997). The uncertainty in atomic oxygen is a limiting factor in resolving the relative roles of chemical heating, solar heating, and dynamical heating in the mesopause region. The SABER data are consistent with a mesopause region near global mean radiative/chemical equilibrium due to the agreement between the SABER-derived atomic oxygen and the maximum radiatively constrained atomic oxygen. Uncertainties in the derived atomic oxygen presently preclude a more definitive conclusion. However, the accuracy of SABER atomic oxygen can be substantially improved if the uncertainty in the rate coefficient for the recombination of O and O_2 can be reduced. At mesospheric temperatures, this rate coefficient carries an uncertainty in excess of 20% and is the primary source of uncertainty in deriving the atomic oxygen concentrations from O_3 and $\text{OH}(v)$ and the radiative constrained values of atomic oxygen (Mlynczak et al., 2013a,b,c).

SABER atomic hydrogen data provide a previously unseen look into the long-term evolution of the mesopause region. The anticorrelation of atomic hydrogen with the solar cycle likely originates from a dynamical or thermal cause. It cannot be due to variability in solar radiation as the solar Ly- α irradiance decreases markedly over the solar cycle. Additional study with the combined SABER data set including the thermal structure, ozone, atomic species, and energetics will likely reveal the mechanism resulting in an increasing H concentration with decreasing solar activity.

The data presented above are unique for the length of the data set and for the radiometric quality that has resulted in over 6 million individual profiles of atomic oxygen and atomic hydrogen to be generated to date. Because of the central role atomic species have in the mesopause region energy budget, these data can provide fundamental assessments of the radiant and chemical physics of upper atmosphere climate and general circulation models. The concept of radiatively constrained atomic species further complements these data by providing independent means of validating the highly derived species abundances. The SABER team looks forward to extending the data set as long as possible for assessment of potential changes in atmospheric composition and energetics associated with climate change.

Acknowledgment

The authors would like to thank the National Aeronautics and Space Administration for the continued support through the TIMED mission and the SABER project, as well as through the Living with a Star and Heliophysics Guest Investigator Programs. All SABER data are freely available for download over the World Wide Web at www.saber.gats-inc.com.

References

- Good, R.E., 1976. Determination of atomic oxygen density from rocket borne measurements of hydroxyl airglow. *Planetary and Space Science* 24, 389–395.
- Grossman, K.U., Kaufmann, M., Gerstner, E., 2000. A global measurement of lower thermosphere atomic oxygen densities. *Geophysical Research Letters* 27, 1387–1390. <http://dx.doi.org/10.1029/2000GL003761>.
- Grossmann, K.U., Vollmann, K., 1997. Thermal infrared measurements in the middle and upper atmosphere. *Advances in Space Research* 19, 631–638.
- Kalogerakis, K.S., Smith, G.P., Copeland, R.A., 2011. Collisional removal of OH(X 2P, $v = 9$) by O, O₂, O₃, N₂, and CO₂. *Journal of Geophysical Research* 116, D20307. <http://dx.doi.org/10.1029/2011JD015734>.
- Kaufmann, M., Ern, M., Lehmann, C., Riese, M., 2013. The Response of Atomic Hydrogen to Solar Radiation Changes, *Climate and Weather of the Sun-Earth System*, 10. Springer, 171–188.
- Mast, J., Mlynczak, M.G., Hunt, L.A., Marshall, B.T., Mertens, C.J., Russell III, J.M., Thompson, R.E., Gordley, L.L., 2013. Absolute concentrations of highly vibrationally excited OH($v = 9 + 8$) in the mesopause region derived from the TIMED/SABER instrument. *Geophysical Research Letters* 40, 646–650. <http://dx.doi.org/10.1002/grl.50167>.
- Mlynczak, M.G., 1997. Energetics of the mesosphere and lower thermosphere and the SABER experiment. *Advances in Space Research* 20, 1177–1183.
- Mlynczak, M.G., et al., 2013a. Atomic oxygen in the mesosphere and lower thermosphere derived from SABER: algorithm theoretical basis and measurement uncertainty. *Journal of Geophysical Research* 118, 5724–5735. <http://dx.doi.org/10.1002/jgrd.50401>.
- Mlynczak, M.G., et al., 2013b. Radiative and energetic constraints on the global annual mean atomic oxygen concentration in the mesopause region. *Journal of Geophysical Research – Atmosphere* 118, 5796–5802. <http://dx.doi.org/10.1002/jgrd.50400>.
- Mlynczak, M.G., Hunt, L.A., Marshall, B.T., Mertens, C.J., Russell III, J.M., Siskind, D., Thompson, R.E., Gordley, L.L., 2013c. Radiative constraints on the minimum atomic oxygen concentration in the mesopause region. *Geophysical Research Letters* 40, 3777–3780. <http://dx.doi.org/10.1002/grl.50725>.
- Mlynczak, M.G., Martin-Torres, F.J., Johnson, D.G., Kratz, D.P., Traub, W.A., Jucks, K., 2004. Observations of the O(³P) fine structure line at 63 μ m in the upper mesosphere and lower thermosphere. *Journal of Geophysical Research* 109, A12306. <http://dx.doi.org/10.1029/2004JA010595>.
- Mlynczak, M.G., Solomon, S., 1991. Middle atmosphere heating by exothermic chemical reactions involving odd-hydrogen species. *Geophysical Research Letters* 18 (1), 37–40. <http://dx.doi.org/10.1029/90GL02672>.
- Mlynczak, M.G., Solomon, S., 1993. A detailed evaluation of the heating efficiency in the middle atmosphere. *Journal of Geophysical Research* 98 (D6), 10517–10541. <http://dx.doi.org/10.1029/93JD00315>.
- Rodgers, C.D., Taylor, F.W., Muggeridge, A.H., Lopez-Puertas, M., Lopez-Valverde, M.A., 1992. Local thermodynamic equilibrium of carbon dioxide in the upper atmosphere. *Geophysical Research Letters* 19, 589–592.
- Sharp, W.E., Kita, D., 1987. In situ measurement of atomic hydrogen in the upper mesosphere. *Journal of Geophysical Research* 92, 4319–4324.
- West, G.A., Weston Jr., R.E., Flynn, G.W., 1976. Deactivation of vibrationally excited ozone by O(³P) atoms. *Chemistry and Physics Letters* 42, 488–493.

Synopsis

This article provides an overview of the ionosphere with an introduction to early understanding of the elements and discussion of its sources and fundamental features. The article also discusses the different latitudes of the ionosphere and its relevance to space weather and as an earthquake predictor.

Introduction

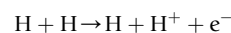
The early Greeks thought that all material was created from four elements: air, earth, fire, and water. We now know that the elements are actually combinations of protons, neutrons, and electrons and that all matter is made from various combinations of these atomic building blocks. However, when cataloging the possible *states* of matter, the ancients were not so far off. In our daily lives, the states of matter referred to as solid, liquid, and gas are clearly related to Aristotle's earth, water, and air.

But what about fire? What about a fourth state of matter? Indeed, there is a fourth state of matter, the state called 'plasma,' which could equally well have been called 'fire,' since the hotter a flame, the closer it comes to the plasma state. We dwell here for a moment on the plasma state itself, since the earth is surrounded by just such a medium – a region called the ionosphere – which is the topic of this section. In fact, we see throughout this volume that the atmosphere of the earth itself includes all four states of matter when one includes raindrops, snow, ice, and the ionospheric plasma, in addition to the gaseous component.

The list of earth, water, and air (solid, liquid, and gas) can be reordered according to the common knowledge that when a solid is heated, it becomes a liquid and then a gas in processes called 'change of phase.' At each phase change, bonds are broken to form the next phase. More energy is required at each step and, by the time one gets to the transition from gas to plasma, quite a lot of energy is required. In this final step, the phase change required actually rips an electron away from the gaseous atom or molecule, leaving a positive ion behind. Since electric charge is conserved, the new state of matter remains neutrally charged on average (equal numbers of positive ions and negative electrons), but these constituents may seldom run into each other and hence have little chance to recombine into the atomic (gaseous) state, and thus, a plasma is born.

How much energy is needed to rip apart atoms? The response is 'a few electron volts,' which is surprising at first, since we are all familiar with batteries in our radios and automobiles that operate at voltages of 1.5–12 V. But we must remember that batteries run on chemical reactions, which themselves involve exchanges of electrons between atoms and molecules, so the volt is a natural-sized unit for ionic bonds. How can we relate this unit to temperature? Suppose we have a pure gas like hydrogen; how hot must it be to become a plasma? Suppose the gas is already hot enough for H₂ molecules to separate into pure hydrogen as the bonds are broken due to collisions with each other. The proton–electron pair that makes up a hydrogen molecule has a bond requiring

13.5 electron volts to break. So for significant numbers of atom collisions to result in ionization, e.g.,



the average energy of the colliding H atoms must be the order of 13.5 electron volts. What does a volt mean? The unit of a volt is Joules per Coulomb (J C^{−1}). This means that if a 1-volt battery is capable of storing one Coulomb of charge, a total of 1 J of energy is available (enough to lift 1 kg 1 m on the earth's surface). A typical car battery can supply 100 A (100 C s^{−1}) for 1 h, so it stores about 360 000 C. Thus, the 12-V car battery stores about 4.3 million Joules (12 J C^{−1} 360 000 C), enough energy to lift a 1000 kg car to a height of 430 m. You should respect your car battery. Since an electron has only 1.6 × 10^{−19} C, 13.5 electron volts corresponds to about 2.2 × 10^{−18} J. Now we need to relate energy to temperature. Kinetic theory shows that the average energy of a particle in a gas is equal to (3/2)*k_BT*, where *k_B* is Boltzmann's constant (1.38 × 10^{−23} J deg^{−1}) and *T* is the absolute temperature in degrees Kelvin. If we set this expression equal to 2.2 × 10^{−18} J and solve for *T*, we find 106 280 K. Such a high temperature shows why it is difficult to produce and control plasmas in laboratories or fusion machines.

The sun is powered by nuclear fusion at its core and is therefore very hot; hence it follows that much of its matter is ionized. Gravity controls this fiercely hot object. The earth is much cooler, of course, and hence it is not obvious that a plasma state would exist in its environs. However, a plasma surrounds the earth, called the ionosphere. The fundamental production and loss mechanisms for the earth's ionosphere are described next, followed by more exotic sources of the plasma surrounding the earth, including the solar wind, magnetic storms, meteors, and the aurora. These sources are localized in time and space and can be linked under the umbrella of weather processes in space or, in short, space weather. Space weather is also influenced by sources of energy and momentum from the earth, the dense atmosphere below, and sources including waves from severe storms, orographic features, and earthquakes, as well as the release of stored energy via plasma instabilities.

Sources and Fundamental Features of the Ionosphere

The ionosphere is formed primarily when the most intense component of the solar spectrum – the X-rays and extreme ultraviolet (EUV) light – impacts the illuminated side of the earth. These high-energy photons strike the dayside of the earth, ionizing the upper atmosphere and losing energy in

the process. As the beam penetrates the atmosphere, the ionizing beam becomes weaker and weaker, leaving behind a layer of ionization. Part of the energy goes into heat as well as into ionizing the air, resulting in the temperature also rising to values much higher than in any part of the dense atmosphere below. Life on earth is thus protected by its upper atmosphere from these dangerous photons, just as the ozone layer absorbs the lower energy, but still harmful, ultraviolet component of the sun's spectrum.

We compare and contrast the atmosphere and ionosphere in Figure 1. The most important atmospheric parameter is temperature, which we plot vs height on the left. The key ionospheric parameter is the number of electrons (which equals the number of positive ions) per cubic centimeter. This is plotted on the right for typical nighttime and daytime conditions.

As anticipated above, the atmospheric temperature rises from its lowest value near the mesopause (near 200 K) to well over 1000 K in the thermosphere in the same height range where the daytime ionosphere is produced. A glance at the right hand side of Figure 1 shows that the ionosphere does not entirely disappear at night, even though the sunlight is no longer present to create new ionization. This is one of the key characteristics of the earth's ionosphere and explains, for example, how Marconi was first able to send wireless signals across the Atlantic Ocean at night.

To understand why some of the ionosphere remains through the night, we must consider the ion chemistry of the region.

At high altitudes (>300 km), production (P) and loss (L) of ionospheric plasma are both small. The result is a balance between diffusion and gravitation and results in the so-called hydrostatic equilibrium in which the plasma pressure (p) is of the form

$$p = p_0 e^{-h/H_p} \quad [1]$$

where 'e' is the base of the natural logarithms, h is height above some reference, p_0 is the pressure at the reference altitude, and H_p is the plasma scale height,

$$H_p = \frac{k_B T}{(M/2)g} = \frac{2k_B T}{Mg} \quad [2]$$

In eqn [2], M is the average ion mass and g is the gravitational acceleration. Equation [1] says that the pressure falls by a factor of about 2.7 with every altitude increase of H_p . M is quite close to the average mass of the neutral atmospheric particles surrounding the plasma. The factor of 2 comes from the average *plasma* mass being half the ion mass, since the electron mass is so tiny. The neutral atmosphere behaves like [1] except the neutral scale height, H_n , is half as large. We conclude that because the electrons are so light, the ionosphere extends higher into space than the neutral atmosphere surrounding it. For reference, H_n is about 50 km and H_p is about 100 km in the middle ionosphere.

At these altitudes, the composition of the atmosphere is no longer similar to the surface composition (79% N_2 , 20% O_2 + minor constituents). The atmosphere is no longer mixed, and lighter atoms can reach higher altitudes. Also, O_2 is photo-dissociated into free oxygen atoms. Figure 2 shows the composition of various atoms, molecules, and ions vs height for the midlatitude ionosphere/thermosphere. We see in this figure that oxygen becomes dominant at 200 km and hydrogen above 700 km. Similarly, the ionosphere is primarily made up of H^+ (with some He^+) at very high altitude, O^+ in the height range near the peak density, and a mixture of O_2^+ and N_2^+ and NO^+ in the lower thermosphere. Hydrogen is so light that it can escape the earth's gravity and form the earth's geocorona, a halo of hydrogen analogous to the sun's glowing corona seen during an eclipse. By chance, hydrogen and oxygen have almost identical ionization potentials, so charge exchange is a very easy process:

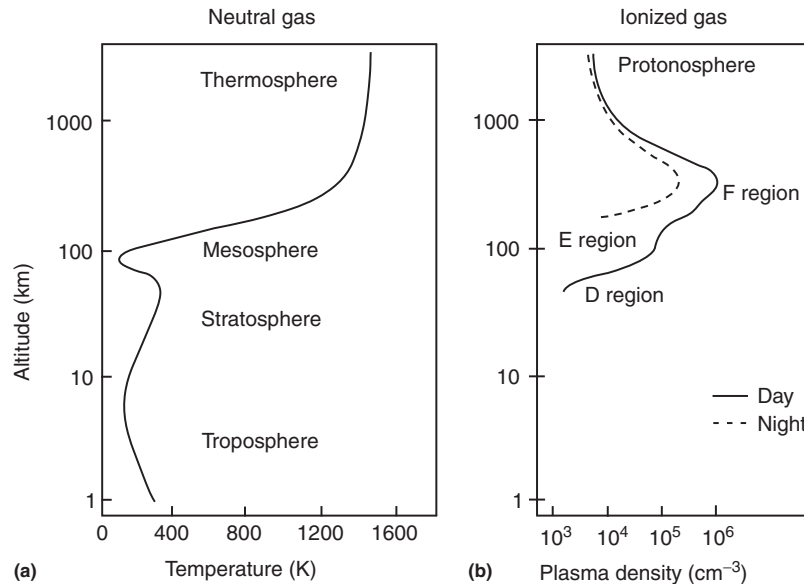


Figure 1 Typical profiles of neutral atmospheric temperature (a) and ionospheric plasma density (b) with the various layers designated. Reprinted with permission from Kelley, M.C., 2009. *Waves and instabilities at midlatitudes*. In: *The Earth's Ionosphere: Plasma Physics and Electrodynamics*, second ed., vol. 96. Academic Press, pp. 267–342. © 2009 by Academic Press.

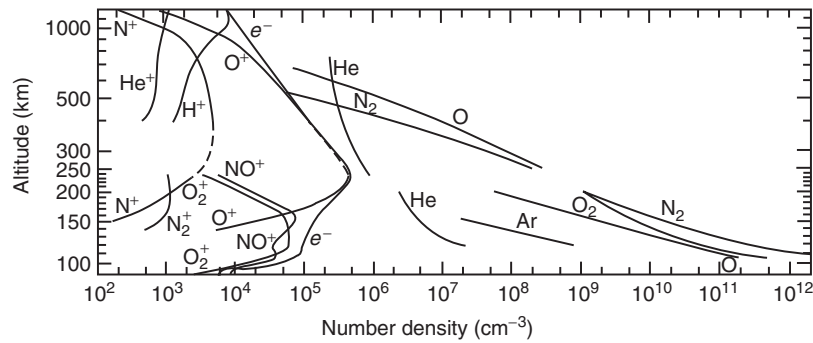


Figure 2 International Quiet Solar Year (IQSY) daytime atmospheric composition. Reprinted with permission of the MIT Press from Johnson, C.Y., 1969. Ion and neutral composition of the ionosphere. Ann IQSY 5. MIT Press, Cambridge, MA. © 1969 by MIT.

that is, if O^+ is surrounded by H gas, after a while an oxygen ion will give up its charge to form a hydrogen ion (H^+). This explains why O^+ ions formed at low altitudes during daytime become H^+ ions at very high altitudes.

Gravity and pressure are not the only forces with which the ionosphere must deal. The earth's dipole magnetic field lines force the hydrogen ions to travel on closed trajectories between the hemispheres since following the magnetic lines is easy, but moving across them is not. The particle motion is like a helix, moving in circles around the magnetic field lines while freely moving parallel or antiparallel to the direction of the field lines. This motion results in the entire region (in a toruslike shape shown in Figure 3) filling with a hydrogen plasma during the daytime (whose source is the sunlight ionization of oxygen coupled with charge exchange). During the night, this region, called the plasmasphere, starts to unload downward into the ionosphere by the reverse process, tending to maintain the oxygen plasma in the ionosphere during the night with the whole process starting over the next day. Why the plasmasphere abruptly ends at about 4 earth radii (60° magnetic latitude) is very interesting and is discussed below.

Refilling from above is not the entire reason for the ionosphere lasting all night, however. It turns out that charged atoms cannot easily recombine with an electron, since in a reaction such as



it is very difficult to simultaneously conserve energy and momentum, and the reaction rate is very small. But in the following reactions

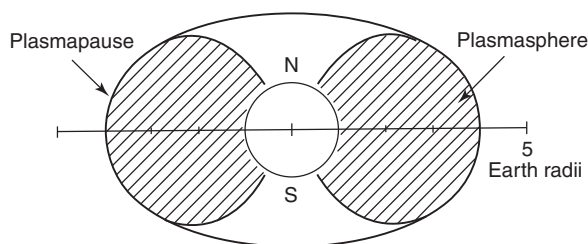
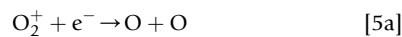
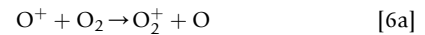


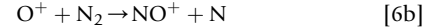
Figure 3 A toroidal region of high plasma density exists around the Earth on average within the region shown. These magnetic flux tubes are filled with plasma of ionospheric origin during the day and discharge only slowly at night.



there are two end products, and this difficulty does not arise. Reactions [5a,b] are called dissociative recombination and are very fast. This explains why the molecular ions (seen in Figure 2) at low altitudes disappear at night, leaving the O^+ plasma above as the distinct nighttime layer (shown in Figure 1). In fact, reaction [4] is so slow that O^+ is actually lost through a two-step process such as, for example, charge exchange followed by [5a]



or ion-atom interchange followed by [5b]



Both [5a] and [5b] leave oxygen in an excited state, which emits both red (630 nm) and green (557.9 nm) light that is visible from the ground to sensitive cameras. Such emissions are called airglow and give us a tool for visualizing the ionosphere, as shown in the next section.

To summarize what we have learned thus far, the ionosphere is created during the daytime by X-rays and EUV from the sun that are absorbed while heating and ionizing the outer layer of the atmosphere. This action heats the gas to temperatures over 1000 K, explaining why it is called the thermosphere. The plasma, which is primarily O^+ above 200 km, diffuses upward against gravity, reaching so high that charge exchange with the geocorona converts the ionosphere to a H^+ plasma, which can escape gravity. But it is constrained by the dipole magnetic field to a toruslike configuration, filled during the day and emptied at night. In the lower thermosphere, molecular ions dominate, but they disappear quickly after sunset, leaving a slowly decaying O^+ layer.

A Day in the Life of the Midlatitude Ionosphere

The most powerful single tool for ionospheric studies is an incoherent scatter radar that detects the microscopic fluctuation due to thermal motions in the ionosphere. The first such

instrument, and still the largest of the eleven now in use worldwide, is near Arecibo, Puerto Rico. The dish is 1 km in circumference and can ‘see’ the ionosphere out to several 1000 km altitude. In [Figure 4](#), we show the plasma content measured by the radar over a full day. The discussion in the [Section Sources and Fundamental Features of the Ionosphere](#) explains the basic character of this plot, but not at all its details. We see that the density is high during the day, even at 100 km. At night the lower ionosphere rapidly – and almost completely – disappears, so much so that to see anything at all we must change the altitude and gray scale. Then slowly during the night, the density decays to low values just before sunrise, when the cycle begins again. But what about the wiggles? Why does the layer move up and down? And what is the origin of the weak ionization layers seen at low altitudes?

These effects in large part are due to horizontal winds and waves in the thermosphere. With a very hot atmosphere on the dayside (>1000 K) in full sunlight and a very cool one at night (<800 K), it is not surprising that strong winds blow from day to night continuously all over the globe. Unlike the thick lower atmosphere, the thermosphere has no thermal inertia. The winds simply blow continuously, acting as a huge atmospheric thermal tide. Speeds of 200 m s^{-1} (720 km h^{-1}) are not uncommon. It is very hard to move the plasma across the magnetic field lines, but such winds easily move the ionosphere up and down the magnetic field lines. At Arecibo, the direction of the magnetic field is at an angle of 45° to the

vertical (pointing northward and downward). When the wind blows southward, the ionosphere moves upward, much like a ping-pong ball held up against gravity on an inclined plane by a hair dryer, as shown in [Figure 5](#).

At high altitude there is very little neutral gas, and recombination is weak. If the wind ceases or blows north, the ionosphere falls due to gravity into regions where [reaction \[5a,b\]](#) can eat away at it. So a southward wind not only elevates the ionosphere, it also keeps it high out of reach of the losses due to the thermosphere. Some of what we see at night in [Figure 4](#) is due to these winds.

The more abrupt changes in height may be due to the electrical forces that act on the ionosphere. These electric fields have associated voltages as high as 200 000 V and ionospheric currents as great as 1 000 000 A, yielding power levels of 2×10^{11} Watts, more power than any artificial generator on earth. Two major generators provide this electrification and both involve motion of a conductor across a magnetic field, exactly the manner by which generators convert mechanically rotating machines into electric energy. The solar wind is the most powerful of the two, generating hundreds of kilovolts across the earth’s polar regions and causing one of nature’s most spectacular visual displays – the aurora borealis and australis (see the next section). The second generator is the motion of the earth’s atmosphere, described in previous paragraphs. Tides, winds, and gravity waves in the atmosphere all drive currents and generate electric fields by the dynamo effect.

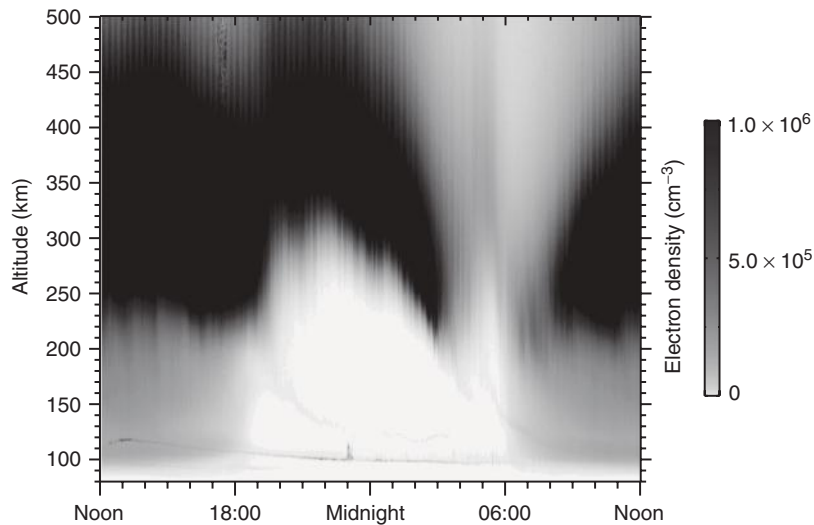


Figure 4 Ionospheric plasma density over Arecibo during a 24-h period (16–17 September 1999; time is Atlantic Standard Time). The lack of plasma below 250 km at night is due to recombination of molecular ions. The high-altitude plasma and interesting thin layers are due to ions such as O⁺, Mg⁺, and Fe⁺, which have long lifetimes.



Figure 5 An illustration showing the analogy between the midlatitude ionosphere on the right, with atmospheric winds pushing the ionosphere up the magnetic field, and a light object suspended on an inclined plane.

Because the magnetic field lines behave like conducting wires, the only voltage easily allowed is across magnetic field lines. **Figure 6** illustrates what happens if a single positive particle is subjected to orthogonal electric and magnetic fields. Initially, the particle is accelerated parallel to the electric field by the force qE .

But once it attains a velocity, the magnetic force, $q\mathbf{V} \times \mathbf{B}$, deflects it to the right. Eventually, it comes to rest and the cycle starts over. The pattern is the same as a dot on the edge of a rolling, nonslipping wheel moving to the right at a velocity E/B . A negatively charged particle (electron) starts out in the opposite direction, but is also deflected to the right, and drifts on average at the same speed as the positive ions. Since there are equal numbers of positive ions and electrons, there is no net current – just a net velocity. Above about 150 km, collisions are so rare that **Figure 6** describes the motion quite well – electric fields are one-to-one, related to the motion of the ionosphere across the magnetic field lines, whereas winds, gravity, and diffusion dominate along the direction of the magnetic field. An eastward electric field over Arecibo, for example, causes an $(\mathbf{E} \times \mathbf{B})B^{-2}$ drift northward and upward at an angle of 45° . Some of the abrupt height changes visible in **Figure 4** are due to such electric field–induced motions.

The abrupt height changes might be temporal or spatial or a combination of both; it is difficult to tell with a single measurement. But the fact that red light is emitted in [reactions \[5a\] and \[5b\]](#) allows us to visualize the plasma in two dimensions. The data in **Figure 7** were obtained on this same night using a bare, backlit charge-coupled device chip

illuminated by a fisheye (all-sky) lens. A narrow (630 ± 1 nm) filter was inserted in the path and the chip was exposed for 90 s. There were 1024×1024 pixels in the image, which, at a height of 250 km, covers a 1000-km diameter circle. The image has been corrected for the lens effects, vignetting, etc., and projected as if we were above the earth looking down rather than up (hence the map of the Caribbean in its usual geometry). We see intricate patterns of light and dark regions, with one of the dark zones positioned right over the Arecibo Observatory. The ionosphere is highly structured this night and is far different than would be predicted if only production, loss, gravity, and diffusion were operating. The sequence of images taken this night shows the dark bands surging poleward from well south of Puerto Rico and then drifting toward the west. This unexpected behavior demonstrates that we have much to learn about even the most well-behaved regions of the ionosphere.

Fire and Ice: The High Latitude Ionosphere

Other entries in the *Encyclopedia of Atmospheric Science* discuss the aurora and the earth's magnetosphere at some length. Here we discuss some of the more striking features of the ionosphere in this region. First of all, the light emissions we call the aurora also are due to the impact of energetic particles on the atmosphere, primarily electrons. Some of these particles leak in from the solar wind, but most are accelerated from the background plasma at a height near 5000 km above the earth. The earth has

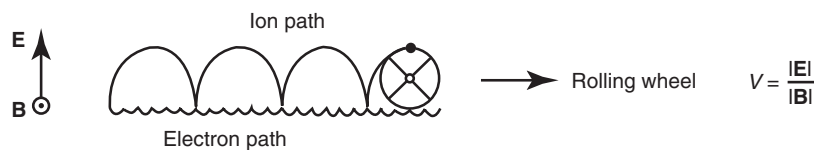


Figure 6 In crossed electric and magnetic fields in vacuum, ions and electrons exhibit the motion shown schematically. The ion path is specifically shown to be similar to that of a dot on the rim of a rolling wheel.

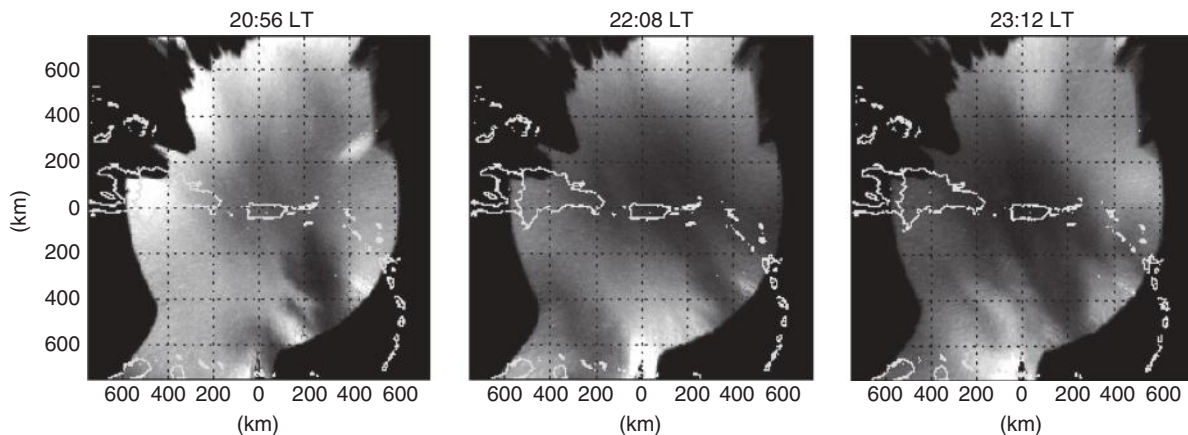


Figure 7 The 630 nm airglow for 17–18 February 1999, superimposed on a map of the Caribbean islands with Puerto Rico in the center. The regions depleted of airglow (shown black) commenced in the SE and surged to the NW on this night. Reproduced by permission of the American Geophysical Union from Kelley, M.C., Makela, J.J., Swartz, W.E., Collins, S.C., Thonnard, S., Aponte, N., Tepley, C.A., 2000. Caribbean Ionosphere Campaign, year one: airglow and plasma observations during two intense midlatitude spread-F events. *Geophysical Research Letters* 27, 2825.



Figure 8 A spectacular view of the eastern half of the United States during a major auroral display.

its own ‘cosmic ray’ generator, just as do pulsars, Jupiter, and other magnetized objects in the cosmos. **Figure 8** shows a view of the earth from 800 km during a spectacular aurora. Close inspection of the city lights shows that the aurora reached Madison, Wisconsin and covered most of the Great Lakes region.

Since intense light must be accompanied by the production of plasma, the aurora is a highly dynamic source of the ionosphere wherever and whenever it appears. **Figure 9** gives some insight into its global character. The view is from 5000 km and shows rings of light circling the top (and bottom) of the earth’s polar regions like a halo. This auroral oval waxes and wanes in size, sometimes (as seen in **Figure 8**) reaching highly populated regions, but usually limited to the mid-Arctic zone.

During certain conditions in the solar wind (when the interplanetary magnetic field is parallel to the earth’s magnetic dipole), the solar wind electric field is transmitted very efficiently along the earth’s magnetic field and throughout both polar ionospheres. At these times, a magnetic storm often occurs and great aurora results. Since the earth’s magnetic field is nearly vertical near the pole, the $(\mathbf{E} \times \mathbf{B})B^{-2}$ drifts are nearly horizontal. Thus, two huge circulation cells often occur in the ionosphere. If the magnetic storm continues for more than an hour or so, the neutral atmosphere can also be put into

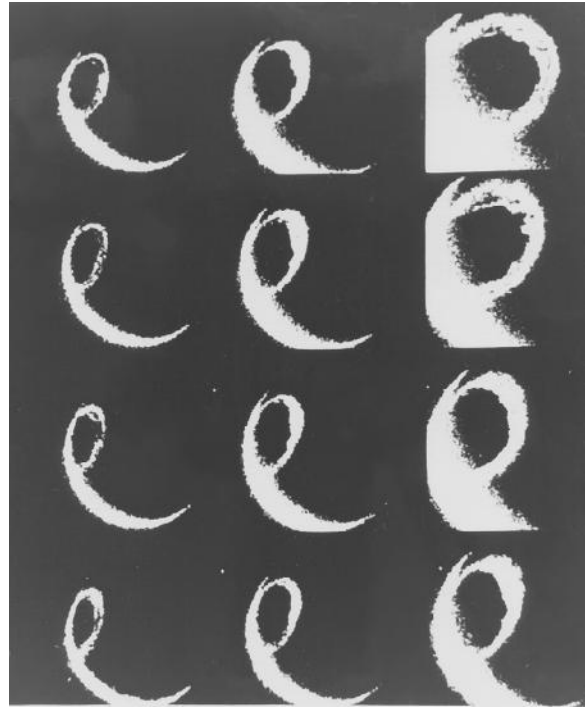


Figure 9 The sequence begins at 05:29 UT on 2 April 1982 (upper left image) as the NASA/GSFC spacecraft Dynamics Explorer 1 first views the auroral oval from the late evening side of the dark hemisphere at low northern latitudes near apogee (3.65 Earth radii altitude) and then from progressively greater latitudes as the spacecraft proceeds inbound over the auroral oval toward perigee. The poleward bulge at onset of the auroral substorm is observed beginning at 06:05 UT (fourth frame). In successive 12-min images, the substorm is observed to expand rapidly in latitude and longitude. Photograph courtesy of Frank, L.A., Craven, J.D., Rairden, R.L. University of Iowa. Reprinted with permission from Kelley, M.C., 2009. *Waves and Instabilities at Midlatitudes*. In: *The Earth’s Ionosphere: Plasma Physics and Electrodynamics*, second ed, vol. 96. Academic Press, pp. 267–342. © 2009 by Academic Press.

motion. Remarkably, the coupled solar wind and ionosphere actually put the earth’s atmosphere in motion. Heat is also generated by the electric currents, which change the global circulation.

The polar circulation rips away at the plasmasphere, reducing its size and compressing it to lower latitudes. After the storm and over a few days, the region refills with cool dense plasma out to about 4 earth radii and 60° magnetic latitude. This altitude/latitude region is where the solar wind and earthly wind dynamos have approximately equal control of the ionosphere/plasmasphere/magnetosphere system. During a great magnetic storm, the composition of the thermosphere can be seriously modified, even worldwide, creating great negative ionospheric storms wherein the ionosphere virtually disappears for a day, even in full sunlight. The atmosphere changes so much that the earth acts like some other planet, one with very little oxygen, in fact. The solar cycle maximum of 2001 promises many exciting scientific discoveries as we continue to instrument the earth and its surrounding near space regions. We also expect that space weather will become more relevant to mankind, as discussed next.

Space Weather

There are many aspects of space weather that are relevant to human habitation, particularly as we become more dependent on technology- and space-based systems. Some of these aspects involve the ionosphere and are described here. Others are related to magnetospheric phenomena, such as the killer electrons of the radiation belt, which create havoc in satellite systems, and solar proton events, which create severe radiation levels for astronauts building the space station.

Rapidly changing magnetic fields due to the vast ionospheric currents flowing in a major auroral event create electrical voltages at the surface of the earth, just as they do in an electrical transformer. Vast power grids are the perfect detector of these voltages, which are unexpected and thus can trigger the unnecessary shutdown of elements along a grid. During the great storm of the last solar cycle, the Province of Quebec went dark for 12 h because of such a surge. Today, our power grids are even more interconnected, and predictions of such conditions are becoming both of more practical importance and of some practical feasibility. Such predictions are among the first challenges of the fledgling National Space Weather Program in the United States as well as the global counterpart of this new program.

Some important space weather effects are strictly due to the earth's dynamic atmosphere without help from the solar wind. The midlatitude weather discussed earlier is of this type and is fairly rare. But near the magnetic equator (where the magnetic field lines are exactly horizontal), severe convective storms occur night after night in some seasons and longitudes. The ionosphere is so dense at the equator that such storms create havoc with communication systems

using radio waves, which must propagate through the ionosphere from satellites to the ground. In brief, in such storms the satellite signals 'twinkle,' just as starlight does passing through the turbulent lower atmosphere. This creates deep fades and distortions in satellite signals, which disrupt communications. The higher the radio wave frequency, the less ionospheric turbulence creates problems. However, even at the high frequencies of the Global Positioning System (GPS) satellites (>1 Gigahertz), ionospheric effects can occur.

Figure 10 is a space weather radar map obtained over the magnetic equator near Lima, Peru. The dark areas reveal where ionospheric turbulence exists. On this night, severe weather erupted just after sunset and lasted for several hours, its effects extending to over 1000 km altitude. Radio signals propagating through this region would be seriously degraded.

The Ionosphere as an Earthquake Predictor

Heki (2011) showed that for large earthquakes ($M > 8$), the total electron content (TEC) between GPS satellites and the ground appears to change by 10%.

However, more likely the ionosphere simply changes altitude, which in the TEC can appear to be a rise or fall. Kelley et al. (2012) explained this phenomenon as due to precursor electromagnetic waves detected prior to earthquakes (Fraser-Smith et al., 1990). The electric field in such waves will cause the ionosphere to rise or fall according to the work by Beach et al. (1997). This idea could be incredibly important, both for saving lives and economically.

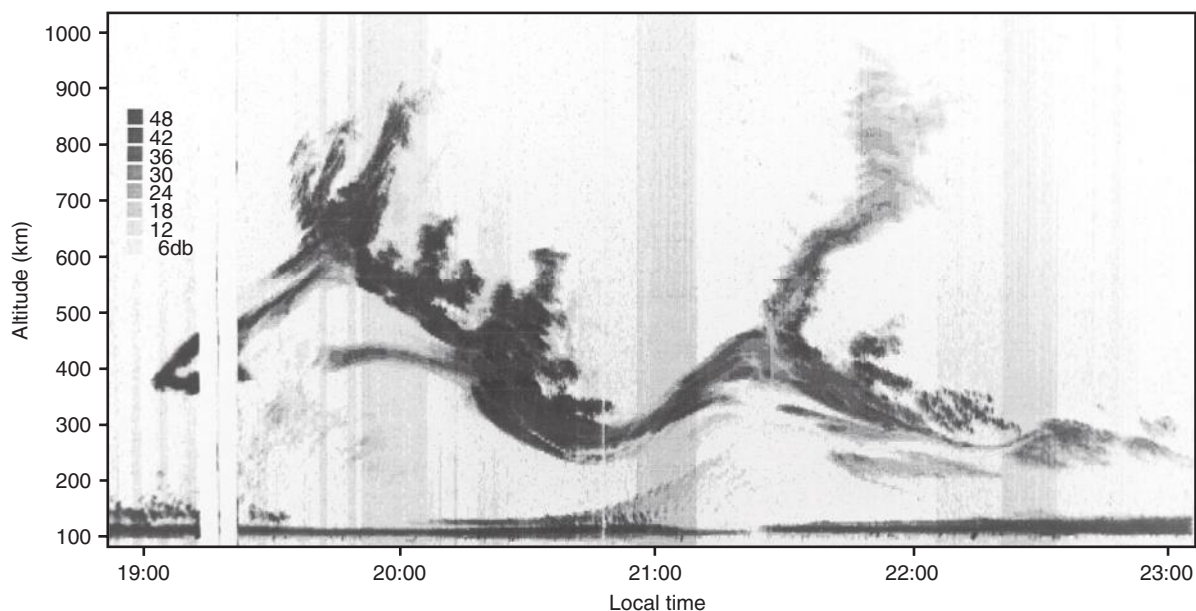


Figure 10 A space weather radar map showing regions of highly turbulent ionospheric plasma over Peru. The dark regions are analogous to clear air and thunderstorm-related turbulence in the troposphere. Reproduced by permission of the American Geophysical Union from Kelley, M.C., Larsen, M.F., LaHoz, C., McClure, J.P., 1981. Gravity wave initiation of equatorial spread F: a case study. *Journal of Geophysical Research* 86, 9087.

Summary

We have learned much about the ionosphere since its discovery near the beginning of the twentieth century, when Marconi first skipped radio waves off it all the way across the Atlantic. The Space Age rocket and satellite probes, along with powerful ground-based radars, have revealed much about its properties. New things are still being discovered, but we are also entering an age of prediction. The goal of the US National Space Weather Program is predicting the onset of severe space weather, much as the meteorology community predicts severe storms in the lower atmosphere. However, the scale of the interacting system stretches from the sun to the earth and is much more variable than the solar constant that directly heats the earth's lower atmosphere. In addition, this vast region of space has few observing stations, and the challenges of predicting its behavior are enormous. Remarkably, the ionosphere also seems to be an enormous detector of earthquakes before they happen.

See also: **Electricity in the Atmosphere:** Ions in the Atmosphere. **Global Change:** Upper Atmospheric Change. Magnetosphere. **Mesosphere:** Polar Summer Mesopause. **Radiation Transfer in the Atmosphere:** Radiation, Solar. **Satellites and Satellite Remote Sensing:** GPS Meteorology. **Solar System/Sun, Atmospheres, Evolution of Atmospheres:** Solar Winds. Thermosphere. **Turbulence and Mixing:** Turbulent Diffusion.

Further Reading

- Beach, T.L., Kelley, M.C., Kintner, P.M., Miller, C.A., 1997. Total electron content variations due to nonclassical TIDs: theory and GPS observations. *Journal of Geophysical Research* 102, 7279–7292.
- Chen, F.F., 1984. *Introduction to Plasma Physics and Controlled Fusion*, second ed. Plenum Press, New York.
- Davies, K., 1990. *Ionospheric Radio*. Peter Peregrinus, Exeter, UK.
- Fraser-Smith, A.C., Bernardi, A., McGill, P.R., 1990. Low-frequency magnetic field measurements near the epicenter of the Mg 7.1 Loma Prieta earthquake. *Geophysical Research Letters* 17 (9), 1465–1468.
- Hargreaves, J.K., 1992. *The Solar-Terrestrial Environment*. Cambridge University Press, Cambridge.
- Heki, K., 2011. Ionospheric electron enhancement preceding the 2011 Tohoku-Oki earthquake. *Geophysical Research Letters* 38, L17312. <http://dx.doi.org/10.1029/2011GL047908>.
- Kelley, M.C., 1989. *The Earth's Ionosphere*. Academic Press, San Diego.
- Kelley, M.C., Swartz, W.E., Heki, K., 2012. An explanation for earthquake precursors at ionospheric altitudes. *Nature*.
- Rishbeth, H., Garriott, O.K., 1969. *Introduction to Ionospheric Physics*. In: *International Geophysical Series*, vol. 14. Academic Press, New York.
- Schunk, R., Nagy, A., 2009. *Ionospheres: Physics, Plasma Physics, and Chemistry*, second ed. Cambridge University Press, Cambridge, UK.

Metal Layers

JMC Plane, University of Leeds, Leeds, UK

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

The ablation of interplanetary dust particles is the source of the metal atoms – Na, Fe, Mg, K, and Ca – that occur in layers between 80 and 100 km. The relative abundances of these metal atoms are quite different from their relative abundances in chondritic meteorites, mostly because elements ablate with widely different efficiencies. Sporadic E layers are concentrated layers of metallic ions and electrons that occur above 90 km and have an important influence on radio communications. Meteoric smoke particles, formed by the condensation of metallic compounds, are the major source of nuclei for noctilucent clouds.

Introduction

The ablation of interplanetary dust particles generates a daily input of about 20 tonnes of a variety of metals into the Earth's upper atmosphere. This gives rise to the layers of neutral metal atoms that occur globally at altitudes between 80 and 100 km. Several of these metals (sodium, iron, potassium, calcium, and lithium) possess suitable optical transitions and can be observed from the ground by the spectroscopic techniques of photometry and lidar. These metals occur as free atoms because above 80 km the concentration of atomic oxygen (O) exceeds that of ozone (O₃): while O₃ oxidizes metal atoms to metal oxides that then go on to form a variety of compounds such as hydroxides, carbonates, and bicarbonates, atomic oxygen and associated atomic hydrogen reduce these compounds back to metal atoms. Above 100 km the metals become ionized by charge transfer with the increasing levels of E region ions such as NO⁺ and O₂⁺.

Surprisingly, the relative abundances of the metal atoms are quite different from their relative abundances in chondritic meteorites. For example, atomic calcium is depleted by more than two orders of magnitude with respect to sodium. The metals also exhibit different seasonal behavior: the integrated column densities of all the metals peak in early winter, but sodium and iron have a marked midsummer minimum, whereas calcium and potassium have a secondary midsummer maximum and hence little seasonal variation. The explanation for these differences appears to be a combination of differential ablation (e.g., the least volatile metal, calcium, is ablated about 10 km lower in the atmosphere than sodium), and differences in the gas-phase chemistries controlling the layers.

Sporadic metal layers are observed between about 95 and 120 km. These are short-lived thin layers of neutral metal atoms that appear explosively on top of the background layers. They are most probably linked to sporadic E layers, which are concentrated layers of metallic ions and electrons that occur between 90 and 130 km and that have an important influence on radio communications.

Lidar enables the metallic layers to be observed with excellent spatial and temporal resolution. Hence, the metal atoms can be used as tracers of dynamical processes such as gravity waves and tides. Narrow-linewidth lidar observations of

the hyperfine structure of the Na- and K-D-lines have also been used to measure temperature and wind speed with excellent precision. Temperature profiles can also be obtained with a technique which uses a two-color lidar to observe the relative populations of the lowest spin-orbit states of atomic iron.

Finally, meteoric metals affect the atmosphere in a number of ways. Sporadic E layers have a significant effect on the electrical conductivity of the lower thermosphere. In the mesosphere, metallic compounds in the form of individual molecules or dust particles appear to be the major source of nuclei for noctilucent cloud formation. In the stratosphere, metal-rich dust may influence ozone chemistry, by providing condensation nuclei for sulfate particles in the Junge layer and affecting the freezing point of polar stratospheric clouds. In the troposphere, the deposition of soluble meteoric iron into the Southern Ocean may represent a significant fraction of bioavailable iron, which would have important climate feedbacks.

Meteoric Ablation as a Source of Mesospheric Metals

The magnitude of the input of interplanetary dust particles into the Earth's atmosphere is actually rather uncertain. Even recent estimates of the dust input vary from 5 to 270 tonnes per day. The input rate is so uncertain because no single technique can observe particles over the mass range from about 10⁻¹² to 1 g, which makes up the bulk of the incoming material. Furthermore, the interpretation of observations within the atmosphere, as well as deposits in ice cores and deep-sea sediments, are affected by uncertain atmospheric transport and deposition patterns. Most of the dust in the inner solar system comes from Jupiter family comets (perhaps as much as 90%). These are comets with short orbital periods of a few decades, and an aphelion close to the orbit of Jupiter. The remaining dust comes from the asteroid belt between Mars and Jupiter; and Halley family and Oort cloud comets, which have orbital periods of centuries. Dust particles from long-decayed cometary trails and the asteroid belt give rise to a continuous input of sporadic meteoroids into the earth's atmosphere. In addition, the dust-rich trails from comets, which recently crossed the earth's orbit (i.e., within the past century or so), are the source of meteor showers (e.g., the *Perseids* and *Leonids*), although

these provide a much smaller mass flux on average than the sporadic background. Dust particles drift into the inner solar system because the pressure of solar photons causes the orbital velocities of particles with a radius larger than about $1\text{ }\mu\text{m}$ to decelerate – a process known as Poynting–Robertson drag. Most of the dust mass should be contained in particles with masses in the range $1\text{--}10\text{ }\mu\text{g}$ (radius $50\text{--}150\text{ }\mu\text{m}$), which enter the terrestrial atmosphere from a near-prograde orbit with a mean speed around 14 km s^{-1} .

Interplanetary dust particles enter the atmosphere at speeds between 11 and 72 km s^{-1} if they originate within the solar system. High energy inelastic collisions with air molecules lead to sputtering of elements from the particle surface. Most particles with masses in excess of $1\text{ }\mu\text{g}$ also rapidly heat up until they reach a melting point around 1850 K . Evaporation of atoms and oxides from the molten particle – a process termed meteoric ablation – is then very rapid, usually leading to complete evaporation of the particle. The peak ablation altitude is around 90 km . However, for an individual dust particle entering the atmosphere, different elements tend to evaporate at different altitudes, a process termed differential ablation, i.e., the most volatile elements – Na and K – ablate first, followed by the main constituents Fe, Mg, and Si, and finally the most refractory elements such as Ca, Al, and Ti. The major metallic constituents of meteorites by elemental abundance are Mg 14.4% , Si 13.6% , Fe 12.1% , Al 1.2% , Ca 0.82% , Na 0.80% , Ni 0.67% , K 0.05% , and Ti 0.03% .

Techniques for Observing Metals in the Mesosphere

The first quantitative observations of metal atoms were made in the 1950s using ground-based photometers that measured the resonance fluorescence from spectroscopic transitions of the metal atoms excited by solar radiation. Emission lines from Na, K, Fe, and Ca^+ were successfully observed because these metals have extremely large resonant scattering cross sections. A large scattering cross section is essential because their concentrations relative to the general atmosphere are less than 100 parts per trillion (100×10^{-12}). Photometers are generally pointed to near zenith during twilight, when the geometrical shadow height of the Earth (the terminator) is close to mesospheric altitudes. Radiative transfer theory is used to derive the vertical concentration profile from the variation of the emission signal as the terminator passes up through the metal layer.

Photometry was superseded in the 1970s when the discovery of tunable laser sources allowed the development of the resonance lidar (laser radar) technique. In this technique, a pulsed laser beam is tuned to a strongly allowed spectroscopic transition of the metal atom of interest, and transmitted up through the atmosphere. The laser pulse is Mie- and Rayleigh-scattered, particularly in the lower atmosphere where there are aerosol layers and the pressure is greater. In the mesosphere, the pulse is resonantly scattered by the metal atoms. A small fraction of the scattered light returns to the ground, where it is collected by a telescope and measured by photon-counting. The return signal is electronically binned to provide the range and hence height resolution of the scattering layer, typically to within 40 m . The absolute metal density is calibrated from the Rayleigh-scattered cross section at a lower altitude of known

atmospheric temperature and density. Lidar has so far been used to observe Na, K, Li, Ca, Ca^+ , and Fe.

The technique has a number of important advantages over photometry. The first is that observations can be made continuously over a complete diurnal cycle, provided an astronomical-quality telescope and narrowband optical filter are employed for daytime measurements. The second is that observations can be made rapidly (typically every 60 s for the Na layer), so that the metal layers can act as tracers of atmospheric motions such as tides and gravity waves. Third, in the case of Na and K, a narrow-linewidth laser can be used in the lidar transmitter to measure temperature and wind profiles in the upper mesosphere. This is possible because the absorption spectrum of each D-line contains hyperfine structure. For example, the D_2 absorption spectrum consists of six hyperfine resonance lines, which become blended at temperatures above 6 K . The degree of blending is very sensitive to temperature, so, employing a laser with a tuning accuracy and frequency stability of at least 50 MHz to scan across the D_2 absorption spectrum, the temperature can be measured with an uncertainty $\approx 1\text{ K}$. The laser can also be scanned to the wings of the D_2 absorption peak in order to measure the Doppler width of the peak. From this the radial wind (i.e., along the line-of-sight of the lidar) can be calculated with an error of less than 3 m s^{-1} , and can then be resolved into the vertical and zonal wind components. A recent development has been to use a two-laser lidar operating at 372 and 374 nm to measure the relative populations of the spin-orbit multiplets of ground state $\text{Fe}(^5\text{D})$, which are related to temperature through the Boltzmann equilibrium.

There have also been a number of measurements by rocket-borne mass spectrometers of the concentrations of positive metallic ions in the upper atmosphere. These flights have been motivated by an attempt to establish a link between meteor showers and the abundance of metallic ions, and to study the role of metallic ions in forming sporadic E layers (see below). Metallic ions such as Mg^+ have also been observed by resonant scattering of sunlight, using spectrometers on satellites and the space shuttle. Satellite observations have also provided the first global picture of the neutral Na and Mg layers.

Observations of Metallic Species in the Atmosphere

Figure 1 shows profiles of the annual mean layers of Na, Fe, K, and Ca observed by lidar at several midlatitude locations. Note that although magnesium is the most abundant meteoric metal, atomic magnesium cannot be observed from the ground because its optical transition at 285.2 nm is obscured by the stratospheric ozone layer. The Na layer has been studied in greater detail than those of the other metals because it is the easiest metal to observe spectroscopically. The column density of Na is about $5 \times 10^9\text{ atoms cm}^{-2}$, although this can vary by a factor of 10 depending on time and location. The layer exhibits a seasonal variation with a wintertime maximum, which is also latitude dependent. For instance, at low latitudes the winter enhancement is only about 1.3:1, whereas at midlatitudes this variation increases to about 3:1 and to more than 10:1 in the polar regions. The height of the peak of the Na layer varies between 88 and 92 km , with the

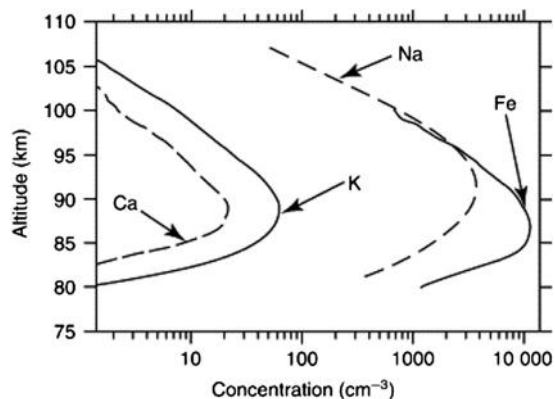


Figure 1 Vertical profiles of the annual mean concentrations of Fe, Na, K, and Ca measured by lidar at a number of midlatitude locations in the United States and Europe.

highest peak heights occurring during summer. The full width at half-maximum (FWHM) of the layer is about 10 km, and it is usually characterized by strikingly small scale-heights of 2–3 km on the top and bottom sides of the layer (the scale height is the distance over which the concentration changes by a factor of e (2.718...)).

The other neutral metal layers that have been observed in the upper atmosphere are those of Fe, K, Ca, and Li. **Figure 2** illustrates the relative column abundances of these metals to that of Na, plotted against their relative abundances in chondritic meteorites recovered from the Earth's surface. Since there is strong evidence that meteoric ablation is the major source of the metals, a good correlation might be expected. Apart from lithium for which there is a rather small observational database, the other metals are all depleted with respect to sodium. In particular, calcium is depleted by the enormous factor of 120–360 depending on season. There are other unexpected differences in the layers. Compared with the peak of the sodium layer, the lithium peak is about 4 km higher, while the potassium, calcium, and iron peaks are several kilometers lower and have considerably smaller scale-heights, as shown in **Figure 1**.

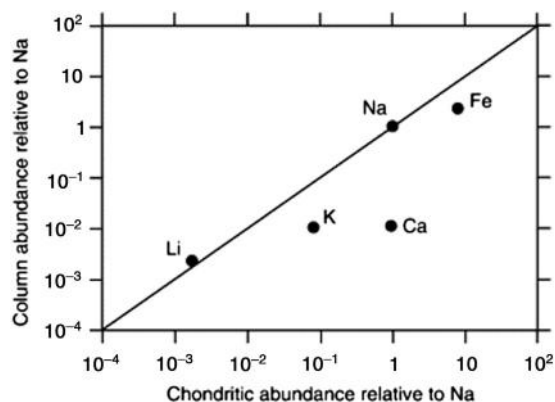
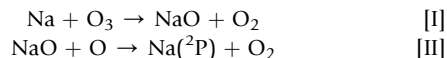


Figure 2 Relative annual average abundances of meteoric metal atoms (at midlatitudes), compared to their relative chondritic abundances.

Lidar observations have also revealed the curious phenomenon of sporadic metal layers. These are very thin, concentrated layers of neutral metal atoms that occur at altitudes between 90 and 110 km, sometimes appearing within a matter of minutes and then surviving for perhaps a few hours. The average FWHM of these sporadic layers is only about 2 km, and their peak concentrations can be as much as 40 times the peak of the background metal layer. They have also been observed with a horizontal extent of more than 1000 km. In explaining this intriguing phenomenon, it has been noted that sporadic layers commonly occur together with sporadic E layers. These are thin layers of metallic ions that can be formed by horizontal wind transport across magnetic field lines, which produces convergence of the ions into layers at null points in the wind shear. Sporadic neutral layers can then result from metal ions forming ion clusters that undergo dissociative recombination with electrons (see below). However, there may well be other mechanisms for sporadic layer formation, including auroral precipitation acting on meteoric dust particles.

Metallic Species in the Earth's Airglow

It was first reported in 1929 that radiation at 589 nm is present in the nightglow spectrum. A decade later it had been established that this radiation is due to the transition $\text{Na}(3^2\text{P}_{3/2,1/2} - 3^2\text{S}_{1/2})$ from a source located within the Earth's atmosphere. Sydney Chapman then postulated the sequence of [reactions \[I\] and \[II\]](#) to account for the emission.



Historically, there have been two significant problems associated with validating the Chapman mechanism. The first was that the rate coefficients for [reactions \[I\] and \[II\]](#) have to be fast enough to generate the measured D-line emission intensity of 50–200 R (1 Rayleigh = 10^6 photons $\text{cm}^{-2} \text{s}^{-1}$ emitted in all directions). However, laboratory measurements (see below) have now confirmed that both reactions are extremely fast. [Reaction \[I\]](#), which is rate determining, proceeds via the electron jump (or harpoon) mechanism. The second problem was the size of the branching ratio, f , for production of $\text{Na}(^2\text{P})$ in [reaction \[II\]](#). Geophysical models of the Na layer show that f has to be ~ 0.1 , in agreement with a field experiment in which a rocket carrying a sodium photometer was launched through the Na layer while a ground-based lidar observed the sodium atom concentration.

A series of elegant laboratory experiments have shown that [reaction \[I\]](#) produces NaO almost entirely in the low-lying $\text{NaO}(\text{A}^2\Sigma^+)$ excited electronic state, rather than the $\text{NaO}(\text{X}^2\Pi)$ ground state. The $\text{NaO}(\text{A})$ state has a long radiative lifetime and is not quenched efficiently, so that in the mesosphere [reaction \[II\]](#) involves both $\text{NaO}(\text{A})$ and $\text{NaO}(\text{X})$ reacting with O. Another lab experiment has shown that f for $\text{NaO}(\text{A}) + \text{O}$ is 0.14 ± 0.04 , thereby reconciling the geophysical observations with the underlying chemical physics.

A further twist to the Na D emission is that the ratio of the two lines at 589.0 and 589.6 nm, which comprise the doublet is not constant (uniquely among the various atomic and

molecular emissions in the earth's nightglow). The average ratio is about 1.67, but this can vary by $\pm 20\%$. This variation appears to be due to a competition between the reaction $\text{NaO(A)} + \text{O}$, which should generate $\text{Na}(^2\text{P}_j)$ with a $J = 3/2$ to $1/2$ propensity of 2.0, or quenching of NaO(A) to NaO(X) by O_2 . The resulting NaO(X) then reacts with O to generate $\text{Na}(^2\text{P}_j)$ with a $J = 3/2$ to $1/2$ propensity of 1.5, where these propensities are derived from theoretical statistical correlation arguments. The D-line ratio should therefore vary between 1.5 and 2.0, depending on the local ratio of O to O_2 .

Modeling of Metallic Layers

Since the only metal species that can be observed directly in the mesosphere are the atomic neutrals and ions, our understanding of the chemistry that forms the metal layers has come from a combination of laboratory studies and modeling. Up until the 1980s, laboratory measurements of the rate coefficients for reactions involving metallic species were available only for ion-molecule reactions, which were studied in ion drift tubes. However, since then the two classical techniques of flash photolysis and the fast flow tube have been applied with great success to the challenging task of studying reactions of neutral metallic species in the gas phase at the low temperatures characteristic of the upper atmosphere. In addition, photoelectron spectroscopy and molecular beams have been employed to investigate the production of excited states in exothermic reactions. While most laboratory work has concentrated on sodium chemistry, there is now a growing database on reactions of iron, magnesium, calcium, and potassium.

Figure 3 is a schematic diagram of the gas-phase chemistry of sodium that is employed in current models. Almost all of

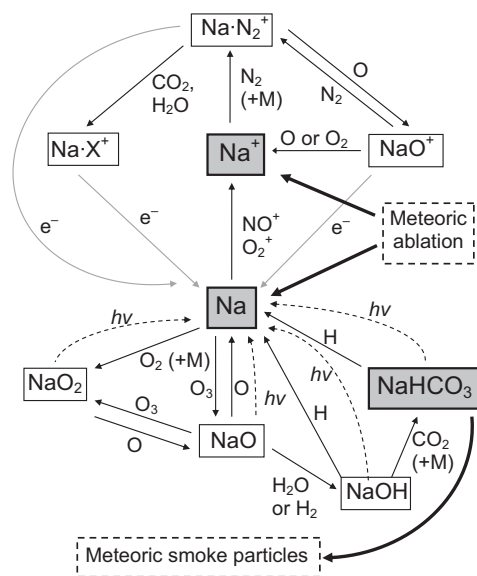


Figure 3 The significant chemical cycles of sodium in the upper mesosphere/lower thermosphere region. Major sodium species are shown in the gray boxes. Thin solid black lines indicate reactions with measured rate coefficients. The broken lines are measured photodissociation reactions. The gray solid lines are dissociative electron recombination reactions, which have not yet been studied.

these reactions have now been studied in isolation under conditions appropriate to the upper mesosphere. Figure 4 illustrates vertical profiles of the major sodium compounds, predicted from a one-dimensional model incorporating the chemistry in Figure 3. Above the atomic sodium layer at around 90 km, ion chemistry predominates. Sodium atoms are ionized mostly by charge transfer with the ambient NO^+ and O_2^+ ions, with a small contribution from solar photoionization. Dielectric recombination (i.e., $\text{Na}^+ + \text{electron} \rightarrow \text{Na}$) is a very inefficient process. Instead, Na^+ is neutralized by forming clusters, which then undergo dissociative recombination with an electron. The rate of neutralization is governed by the initial formation of $\text{Na} \cdot \text{N}_2^+$, followed by a competition between atomic O and CO_2 , the former converting the cluster back to Na^+ , and the latter forming a stable cluster that will subsequently react with an electron.

Below 90 km, Na becomes converted to the stable reservoir NaHCO_3 , via a series of steps beginning with the oxidation of Na by O_3 to NaO . As shown in Figure 3, species such as NaO , NaOH , and NaHCO_3 are converted back to Na by reaction with O and H . The rates of the reaction between atomic hydrogen and NaHCO_3 , as well as its photodissociation, have been shown in the laboratory to be slow, so that NaHCO_3 should be the major reservoir below 84 km, as shown in Figure 4. NaHCO_3 is assumed to then polymerize with other metallic compounds to form meteoric smoke (see below), hence its predicted decrease below 80 km. Note, however, that there have not yet been direct observations of NaHCO_3 in the atmosphere. It is only when O and H become abundant above 85 km that atomic sodium becomes the dominant form. Hence, the small scale-height on the underside of the Na layer mirrors the fall-off in atomic O and H .

The chemistries of iron, magnesium, and calcium are somewhat different from that of sodium. Unlike Na^+ , the ions Fe^+ , Mg^+ , and Ca^+ are chemically active, reacting with O_3 to form oxides and forming strongly bound dioxides with O_2 . The rate of neutralization of these metal ions is governed by competition between atomic oxygen and electrons for the metal oxide ions. The removal of metal atoms on the underside of the layer involves oxidation by O_3 to form neutral metal oxides, followed by recombination with O_2 , CO_2 , or H_2O to form the trioxide, carbonate, or dihydroxide, respectively.

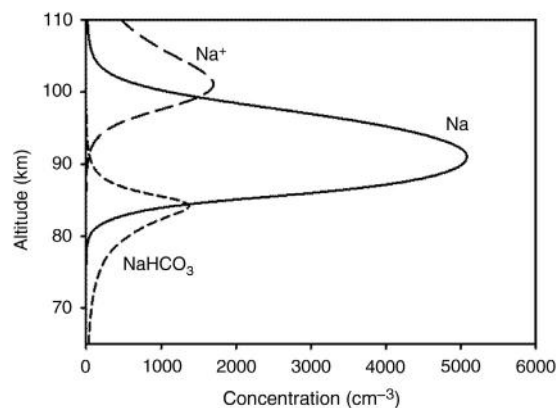


Figure 4 Diurnally averaged height profiles of Na , Na^+ , and NaHCO_3 , predicted by a one-dimensional model for January, 40°N .

Although the dihydroxide (e.g., $\text{Mg}(\text{OH})_2$) is thermodynamically most stable, it is not formed very rapidly because of the dryness of the upper mesosphere. Hence, it is not clear at present which of these compounds form the major reservoir species for these three metals.

Models using the sodium chemistry in Figure 3 are able to reproduce very satisfactorily the observed Na layer as a function of season and latitude. Models for the Mg, Fe, K, and Ca layers have also been published recently. As Figure 1 shows, although there are differences in the heights and shapes of these layers with respect to the Na layer, these differences are of second order. This implies that the enormous deviations in the metal abundances from their expected chondritic ratios (Figure 2) are not primarily due to differences in chemistry, but are due to differential meteoric ablation. Indeed, current models require iron and calcium ablation efficiencies of about 25 and 5%, respectively. Models of iron and calcium are able to explain quite satisfactorily most of the seasonal differences between these metals and sodium. However, the very surprising differences between sodium and potassium, which are from the same group in the Periodic Table, remain to be fully explained.

The formation of sporadic metal layers has also been successfully modeled using ion-molecule chemistry of the kind illustrated for Na^+ in Figure 3. Sporadic E layers tend to descend between 120 and 95 km under the influence of the diurnal tide, at a rate of a few km h^{-1} . O_3 and the general atmospheric density increases with descending height, so that molecular metallic ions (e.g., FeO^+) form more quickly. At the same time, the atomic O density falls, so the reactions involving O which convert metallic oxides back to the atomic ion (e.g., $\text{FeO}^+ + \text{O} \rightarrow \text{Fe}^+ + \text{O}_2$) get slower. Hence, these oxides tend to undergo dissociative recombination with electrons (e.g., $\text{FeO}^+ + e^- \rightarrow \text{Fe} + \text{O}$), generating a layer of neutral atoms.

Models have also been modified to explore perturbations to the metal layers induced by gravity waves. Gravity waves in this region can cause temperature variations of over 20 K and vertical displacements of several kilometers in less than an hour. The motivation for studying the coupling of metal chemistry and dynamics has been to examine whether the metals are suitable inert tracers of atmospheric motion, or whether the apparent dynamical perturbation is amplified by a fast chemical response to changes in temperature and the concentrations of minor species such as O_3 , O, and H. Several studies now show that the sodium layer is a conservative tracer of short-period gravity waves. In contrast, the underside of the iron layer appears to be controlled by fast chemistry, which may lead to significant chemical amplification.

Impact of Meteoric Metals in the Mesosphere and Stratosphere

In the mesosphere below about 85 km, metallic compounds polymerize together to form aerosols called meteoric smoke particles. Since the major meteoric constituent elements are Fe, Mg, and Si, and these appear to ablate at similar rates and altitudes, it is likely that the smoke consists of amorphous Fe–Mg silicate particles about 1 nm in radius. Smoke particles have been detected by rocket-borne charged particle detectors, since a fraction of them is charged by the uptake of ambient

electrons. Meteoric smoke has also been observed by optical extinction using a spectrometer on the Aeronomy of Ice in the Mesosphere Satellite.

It has been proposed that metallic species play a major role as the source of condensation nuclei for noctilucent clouds (also termed polar mesospheric clouds). These are ice clouds that occur between 82 and 85 km, at high latitudes during summer. They were first observed in 1886 and are increasing in frequency-of-occurrence and brightness, and spreading to lower latitudes. The clouds have therefore been studied intensively as indicators of climate change in the upper atmosphere. The changes in the clouds probably result from increasing water vapor concentrations and decreasing temperatures in the upper mesosphere. Because the upper mesosphere is a very dry region, the condensation nuclei for noctilucent clouds must be particularly effective. Two proposed sources of nuclei are hydrated metallic ions, and meteoric smoke particles, which may be as small as single metal silicate molecules with very large dipole moments. In both cases, strong electrostatic forces promote the binding of water molecules.

Smoke particles are transported within a few weeks to the winter pole, and then descend in the winter polar vortex into the stratosphere. During this period of several months, the particles should grow through coagulation to be around 40 nm in radius. Uptake of sulfuric acid on smoke particles may account for the observed depletion of this acid above 35 km. Laboratory studies show that the particles will then dissolve in the cold, concentrated sulfuric acid droplets that comprise the Junge layer in the lower stratosphere around 20 km. The high concentrations of meteoritic ions such as Fe^{2+} and Mg^{2+} which have been measured in these droplets using airborne aerosol mass spectrometry may affect the freezing point of polar stratospheric clouds, and hence impact on ozone depletion.

Meteoric smoke particles probably enter the troposphere at midlatitudes, where stratosphere–troposphere exchange is driven by atmospheric waves generated by mountain ranges and storm tracks over the North Atlantic, North Pacific, and the Southern Oceans. Within the troposphere, wet deposition (by rain and snow) of these particles appears to dominate over dry deposition. The evidence for this comes from measurements of meteoric smoke in ice cores from Greenland and central Antarctica. A recent study using a general circulation model indicates that there should be a large rate of deposition of meteoric iron into the Southern ocean (50–60° S), where the supply of bioavailable iron to phytoplankton is limited. The resulting Fe fertilization could have important climate feedbacks because phytoplankton draw down CO_2 from the atmosphere.

See also: **Chemistry of the Atmosphere:** Chemical Kinetics; Ion Chemistry. **Clouds and Fog:** Noctilucent Clouds. **Dynamical Meteorology:** Atmospheric Tides. **Global Change:** Upper Atmospheric Change. **Lidar:** Resonance. **Mesosphere:** Atomic Species in the Mesopause Region; Ionosphere; Polar Summer Mesopause. **Observations Platforms:** Rockets. **Ozone Depletion and Related Topics:** Ozone Depletion Potentials. **Radar:** Meteor Radar. **Solar System/Sun, Atmospheres, Evolution of Atmospheres:** Meteors.

Further Reading

- Ciemesha, B.R., Batista, P.P., Simonich, D.M., 1996. Formation of sporadic sodium layers. *Journal of Geophysical Research* 101, 19701–19706.
- Gardner, C.S., Taylor, M.J., 1998. Observational limits for lidar, radar, and airglow imager measurements of gravity wave parameters. *Journal of Geophysical Research* 103, 6427–6437.
- Hervig, M.E., Gordley, L.L., Deaver, L.E., Siskind, D.E., Stevens, M.H., Russell, J.M., Bailey, S.M., Megner, L., Bardeen, C.G., 2009. First satellite observations of meteoric smoke in the middle atmosphere. *Geophysical Research Letters* 36 art. no: L18805.
- Plane, J.M.C., 1991. The chemistry of meteoritic metals in the upper atmosphere. *International Reviews in Physical Chemistry* 10, 55–106.
- Plane, J.M.C., 2003. Atmospheric chemistry of meteoric metals. *Chemical Reviews* 103, 4963–4984.
- Plane, J.M.C., 2012. Cosmic dust in the earth's atmosphere. *Chemical Society Reviews* 41, 6507–6518.
- Rapp, M., Thomas, G.E., 2006. Modeling the microphysics of mesospheric ice particles: assessment of current capabilities and basic sensitivities. *Journal of Atmospheric and Solar-Terrestrial Physics* 68, 715–744.
- von Zahn, U., Gerding, M., Höffner, J., McNeil, W.J., Murad, E., 1999. Iron, calcium, and potassium atom densities in the trails of Leonids and other meteors: strong evidence for differential ablation. *Meteoritics and Planetary Science* 34, 1017–1027.
- Vondrak, T., Plane, J.M.C., Broadley, S., Janches, D., 2008. A chemical model of meteoric ablation. *Atmospheric Chemistry and Physics* 8, 7015–7031.

Polar Summer Mesopause

RH Varney and MC Kelley, Cornell University, Ithaca, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

The summer polar mesosphere is one of the most fascinating regions of the atmosphere. In full sunlight, the temperature reaches as low as 100 K, far colder than all but a few regions in the solar system. Even though the upper atmosphere is freeze dried, enough water still remains to allow ice crystals to form around meteoric dust particles, leading to the formation of the highest clouds on earth (near 85 km), which form beautiful structures seen in the polar twilight. Interestingly, these clouds were never reported until the end of the nineteenth century and have increased throughout the twentieth century. Many researchers relate this phenomenon to atmospheric change since, as the lower atmosphere heats up, the upper atmosphere cools. In addition, since methane is the main cause of hydrogen in mesospheric water vapor, its increase is a factor in the formation of these clouds. Increased interest in this region was due in part to the huge radar target this region exhibits, 10 000 000 times as large as elsewhere outside of the polar summer. Explanation of these signals has led to a new understanding of the role of dust in a plasma.

Introduction

The mesosphere is the region of the atmosphere located between the stratosphere and the thermosphere, between 50 and 90 km, in which temperature decreases with height. The transition between the mesosphere and the thermosphere is called the mesopause and is the altitude at which the temperature reaches a minimum before increasing with height in the thermosphere. The mesosphere is collocated with the D-region ionosphere, the lowest portion of the partially ionized plasma blanket that surrounds the Earth. The polar summer mesosphere is particularly fascinating. Although the mesosphere has very little water vapor, the highest clouds on Earth are found in this region. Remarkably, in the full polar summer sunlight, the temperature often reaches values as low as 110 K, with one measurement as low as 90 K. This is clearly the coldest natural temperature found on or near the Earth. Additionally, we find an intense radar scattering layer together with atmospheric motions that are dominated by poorly understood gravity waves and tides. These properties are reason enough for scientific interest in the polar mesosphere, but there are important global change aspects as well.

Noctilucent Clouds and the Temperature Anomaly

Interest in the polar summer mesosphere was started on 18 June 1885, 2 years after Krakatoa's eruption in 1883. Silvery-blue clouds were observed from the ground under twilight conditions when the Sun was below the horizon but was still illuminating the mesosphere (i.e., just before dawn or just after sunset). Using photographic triangulation, the height of these clouds was found to be about 82 km, the highest clouds ever seen on Earth. Owing to their unusual nighttime brightness, they were named noctilucent clouds (NLCs). It is believed that a major volcanic eruption introduced a considerable amount of water vapor into the stratosphere that took 2 years to be transported to the mesosphere, eventually contributing to the ice particles that formed these

clouds. The ice particles were large enough to scatter sunlight and be seen by the naked eye. These clouds are only observed during the summer months, usually presenting a wavy pattern, an effect attributed to their interaction with passing gravity waves. The ideal viewing zone is between 53 and 57° latitude because of the long twilight and the polar location of the clouds themselves. The example presented in [Figure 1](#) displays some of the NLC characteristics.

An important discovery was made during the International Geophysical Year of 1957–58. Rocket grenades launched from many locations revealed that, against all expectations, the temperature in the polar summer mesosphere is colder than in the winter polar zone. In fact, a temperature difference of about 100 K exists between summer and winter. More recent data presented in [Figure 2](#) reveal these hemispheric temperature differences quite well, along with the level of temperature fluctuation in the region, which is quite high in the winter hemisphere. This result supports speculations that the NLCs were composed by ice that forms at extremely low



Figure 1 Noctilucent clouds as observed at 2255 UTC on 19 July 1997 from Glengarnork, Ayshire, Scotland. Figure courtesy of Tom McEwan.

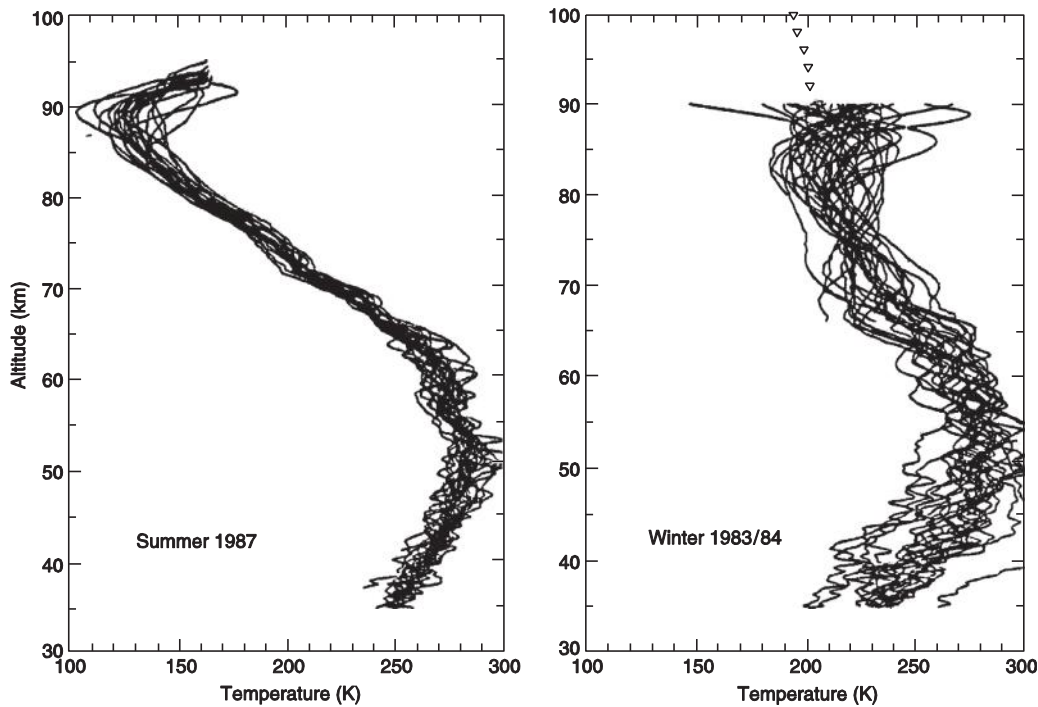


Figure 2 Temperature profile measurements taken using radar tracking of falling spheres during summer 1987 (left) and winter 1983–84 (right) over Andoya, Norway. Adapted from von Zahn, U., Meyer, W., 1989. Mesopause temperatures in polar summer. *Journal of Geophysical Research* 94, 1647.

temperatures, even at the low water vapor pressure of the mesopause zone (1–2 ppm).

New ways to study NLCs have been developed that should improve our understanding experimentally. For instance, in the early 1980s, satellite measurements detected NLCs, although, owing to the different method of detection, they were called polar mesospheric clouds (PMCs). They are believed to be the same as NLCs, with the only difference being that, from orbit, they could be observed 24 h a day. In 2007, the Aeronomy of Ice in the Mesosphere (AIM) spacecraft was launched with multiple optical instruments dedicated to studies of PMCs. **Figure 3** shows a composite image of PMCs constructed from multiple orbits of the AIM spacecraft.

Another observation method that is not hampered by lighting conditions takes advantage of the unexpectedly high radar cross section in the polar summer mesosphere. This scattering process is of considerable interest in its own right and is discussed in detail later. The latest way to monitor NLCs involves lidar, a method analogous to radar but using light waves instead of radio waves.

NLCs are usually observed only at high latitudes between 50 and 60°, but on 22 June 1999 they were observed in Boulder, Colorado (40° N), an indication that NLCs are moving south. This phenomenon could be due to cooling and/or increased water vapor caused by rising levels of methane and carbon dioxide due to human activity, a topic discussed later.

Mesospheric Dynamics

The gravity or buoyancy waves that create the interesting structure in **Figure 1** are of more than passing importance in

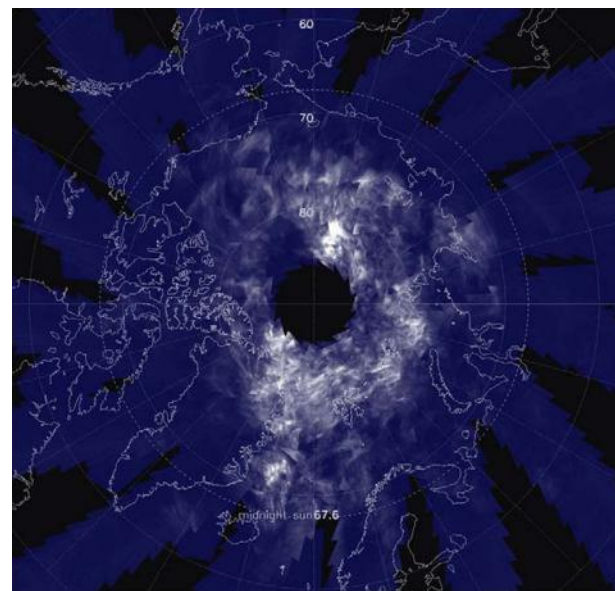


Figure 3 PMC albedo measured by the Cloud Imaging and Particle Size (CIPS) instrument on the AIM spacecraft. The above image is a composite of 15 orbits on 9 July 2007. Adapted from Rusch, D.W., Thomas, G., McClintock, W., et al., 2009. The cloud imaging and particle size experiment on the aeronomy of ice in the mesosphere mission: cloud morphology for the northern 2007 season. *Journal of Atmospheric and Solar-Terrestrial Physics* 71, 356–364.

understanding the low temperature of the polar summer mesosphere. Such waves are also called internal waves, since they can propagate easily through the atmosphere from one height range to another.

Typical sources include surface wind flow over orographic features, frontal systems, and severe storms; even earthquakes, tsunamis, and nuclear explosions have created substantial waves in the atmosphere. What makes these waves unique and important to the upper atmosphere is that they increase in amplitude as they propagate upward. This seems counterintuitive, but is actually firmly rooted in the principle of energy conservation. The kinetic energy per unit volume in a wave packet is given by eqn [1]:

$$W_{KE} = (1/2)\rho_0(\delta u)^2 \text{ J m}^{-3} \quad [1]$$

In eqn [1], δu is the perturbation wind velocity in the wave (m s^{-1}) and ρ_0 is the background atmospheric mass density (kg m^{-3}). (Since the kinetic energy of a simple molecule of mass M is $(1/2)Mv^2$, measured in joules, exchanging ρ_0 for M yields the kinetic energy density in J m^{-3} .) Similar expressions hold for other forms of energy such as heat and potential energy in a wave packet. The important issue is that, as the wave propagates upward, ρ_0 decreases drastically, exponentially, in fact, with the form for a uniform atmospheric temperature being $\rho_0(z) = \rho_0(0)e^{-z/H}$, where $H = kT/Mg$ is the atmospheric scale height (7 km at surface temperatures) and k is Boltzmann constant. So if energy is to be conserved as a wave propagates upward each time density ρ_0 decreases by a factor of 2, the wave velocity must increase by a factor of $\sqrt{2}$ or about 40%. Eventually, any given wave will reach a height where the amplitude is so great that it breaks. A good rule of thumb is that a wave will break when its internal wave perturbation velocity exceeds its propagation speed. It catches up with itself, steepens, and breaks. Figure 4 shows an analogous situation for water waves. The wave speed slows as the water becomes shallow and, when the internal perturbation velocity exceeds the wave velocity, the wave breaks.

What does this have to do with the cold summer mesopause? When waves break on a beach or in the clear air, they deposit their energy and momentum back into the local medium (swimmers on the surface know this very well). The mesosphere is so tenuous that the input of momentum from waves generated in the dense lower atmosphere is very significant. Current theories of the mesosphere argue that waves reaching these heights come from preferred directions that are different in the two hemispheres. In the summer hemisphere,

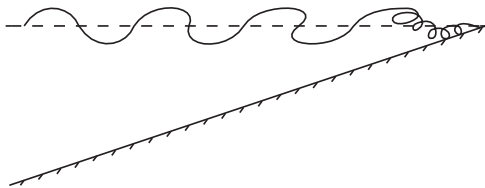


Figure 4 Analogy for gravity wave breaking similar to water waves on a beach. After Kelley, M.C., 2009. *The Earth's Ionosphere: Plasma Physics and Electrodynamics*, International Geophysics Series, vol. 96, second (Ed.), Academic Press, Burlington, MA. Reproduced with permission from Elsevier.

the waves preferentially come from the west, depositing a net eastward momentum into the medium. This spins up the atmosphere somewhat and it moves away from the pole. To conserve mass, there is a net upflow at high latitudes, resulting in adiabatic cooling. The opposite effect occurs at the winter pole and the temperature rises.

The preferential direction could arise in a variety of ways, but we discuss only one here: the so-called critical layer effect. Gravity or buoyancy wave velocities are small enough that jet stream winds can be larger than the wave propagation speed. Suppose a wave propagates upward to a height where its horizontal phase speed equals that of the background wind (height h_0 in Figure 5). At this height, the wave is not a wave at all, just some eddies in the flow, and it ceases to exist. In fact, only waves with horizontal velocities greater than u_{max} get through the jet stream at all. But waves propagating in the other direction, against the flow, are never subject to this effect and pass through easily. Since the jet stream is to the west in the summer and to the east in the winter, gravity wave filtering might explain the mechanism described above and the observed temperature asymmetry.

Modern global circulation models can include such effects by parameterizing momentum fluxes. One such model is the Thermosphere–Ionosphere–Mesosphere–Electrodynamics General Circulation Model (TIME-GCM). The calculated yearly variation of temperature at 85 km is shown in Figure 6. We can observe an asymmetry between hemispheres for solstice

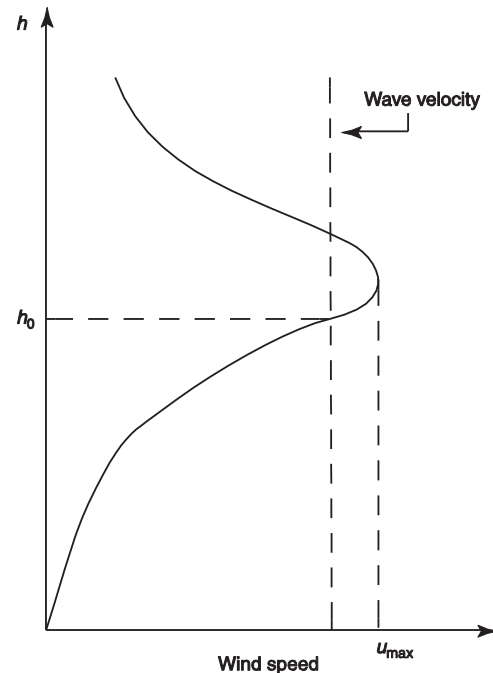


Figure 5 Internal gravity waves have phase velocities of the same order as wind speeds in the jet stream. Here we illustrate a wind profile whose peak value exceeds the phase velocity of a particular wave. In this case, an upward propagating wave will reach the critical layer where it is absorbed by the fluid. After Kelley, M.C., 2009. *The Earth's Ionosphere: Plasma Physics and Electrodynamics*, International Geophysics Series, vol. 96, second (Ed.), Academic Press, Burlington, MA. Reproduced with permission from Elsevier.

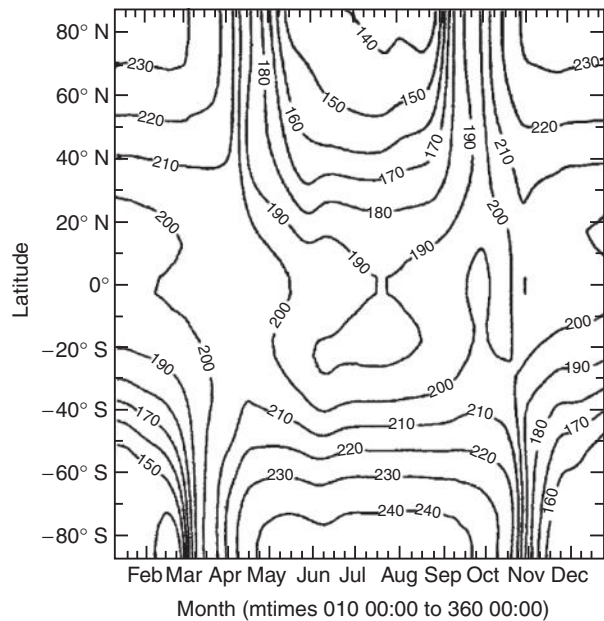


Figure 6 TIME-GCM variation of zonal average neutral gas temperature in Kelvin, over a year at 85 km. Figure courtesy of Roble, R.G., 2000. On the feasibility of developing a global atmospheric model extending from ground to the exosphere. *Atmospheric Science across the Stratopause*. Geophysical Monograph 123, American Geophysical Union, pp. 53–67.

conditions at high latitudes. The southern polar summer hemisphere seems to be warmer than the northern polar summer hemisphere by 100 K.

Polar Mesosphere Summer Echoes

As noted above, an intriguing feature of the polar summer mesosphere is the ability of the region to strongly scatter radio waves. Powerful very high frequency (VHF) (30–300 MHz) radars can detect scattering from turbulent irregularities in the mesosphere at all latitudes and in all seasons. Echoes from the polar summer mesosphere, however, exhibit numerous properties that are unlike echoes from ordinary mesospheric turbulence. Early studies of VHF radar echoes from the high-latitude summer mesosphere and lower thermosphere using the Poker Flat Mesosphere, Stratosphere, and Thermosphere (MST) radar in Alaska (65° N) showed a relatively narrow and surprisingly intense echoing layer centered at about 86 km. The echoes were characterized by their strong VHF (50 MHz) radar backscattering cross section, with backscattered powers 2–3 orders of magnitude greater than typical values observed at low or middle latitudes (in any season) or at high latitudes (in nonsummer periods). The echoes are now referred to as Polar Mesosphere Summer Echoes (PMSEs).

As observed at VHF (~50 MHz), northern hemispheric PMSEs exhibit the following characteristics:

- They display a thin but intense echoing region near the summer mesopause (~85 km) with a margin of ± 3 km.

- While some relatively strong, albeit sporadic, echoes have been reported at latitudes as low as 52°, the strongest, most continuous echoes are observed at latitudes poleward of about 65°.
- The echoes appear around mid-May, last until mid-August, and are relatively continuous.
- Both the height range and seasonal variations of PMSEs correlate reasonably well with those of the cold temperature mesopause (i.e., the coldest known atmospheric region).
- Joint observations using VHF radar and sounding rockets show that intense PMSEs often can be associated with sharp 'biteouts' in the ambient electron density.

Subsequent to their discovery in 1981, many observations related to PMSEs have been made using radar, lidar, and rockets. PMSEs have since been observed using radars at 224, 450, 500, 930, and 1295 MHz, all of which are too high in frequency to ever observe echoes from ordinary mesospheric turbulence. Figure 7 shows an example of simultaneous measurements of PMSEs at three different frequencies. These observations have helped to formulate a number of theories proposed to explain the generation of the intense radar echoes and the remarkable physical conditions associated with them. Subfields in research as disparate as dusty (icy) plasma physics, interplanetary dust cloud studies, meteor ablation, and re-coagulation science all have something in common with the polar summer mesopause region.

Radar scattering from the mesosphere is dominated by scattering from free electrons in the overlapping D-region ionosphere. The scattered radiation from many individual electrons will only be detectable if it adds coherently, which will happen if the structures in the electron density satisfy the Bragg scattering condition. For a monostatic radar, this condition requires the electron density to have structures matching half of the radar wavelength (e.g., 3 m for a 50-MHz radar). Neutral turbulence is ubiquitous in the mesosphere due to breaking gravity waves and dynamic instabilities. The turbulent eddies begin at scales of ~1 km, but cascade to progressively smaller scales until they become so small that viscosity becomes important. The spectrum of neutral turbulence is typically divided into an inertial subrange at the larger scales, where energy flows to progressively smaller scales without being destroyed, and a viscous subrange at the smallest scales where viscosity is removing kinetic energy from the turbulent eddies. The dividing line between the two subranges is near the Kolmogorov microscale, which is typically approximately 1–10 m in the mesosphere and increases with altitude. PMSEs are both intriguing and surprising because the 3-m irregularities responsible for Bragg backscatter at 50 MHz should lie within the viscous subrange of turbulence at 86 km and, as a consequence, should be strongly damped. Observations at higher frequencies indicate the presence of even smaller scales, which is even more puzzling.

Numerous different observations have demonstrated a link between PMSEs and the ice particles responsible for NLCs/PMCs. Rocket data of electron structure measurements in the medium provided the first clear indication that echoes were related to the coupling of electrons and small particles. Figure 8 shows simultaneous VHF scatter echo profiles and the electron

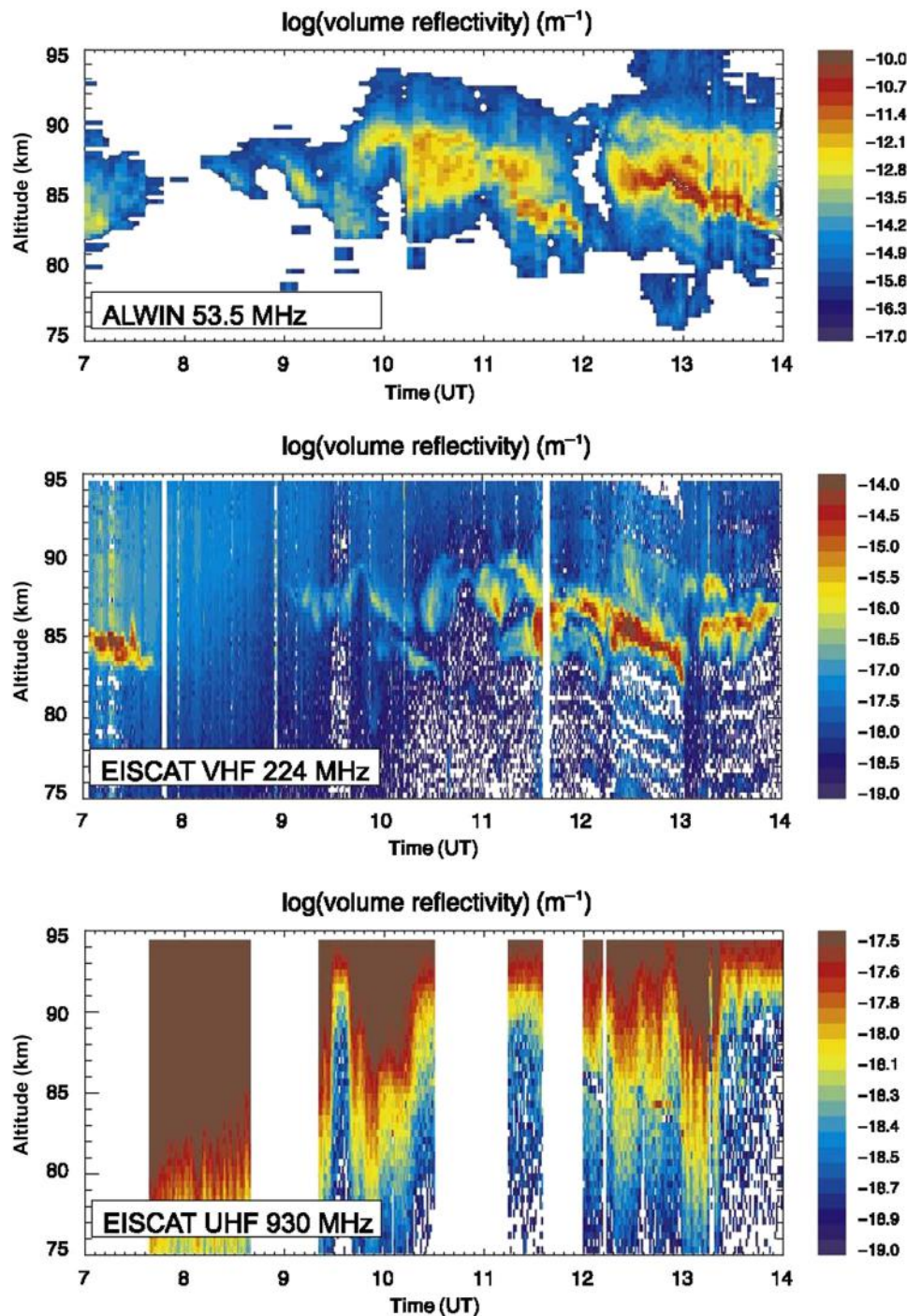


Figure 7 Observations of PMSEs from three different radars at different frequencies. For the two lower frequency radars, all of the observed strong signal is PMSE, but for the ultra high frequency (UHF) (300 to 3000 megahertz) radar (bottom panel), most of the signal is incoherent scatter from ambient D-region plasma. The UHF PMSE is visible as a small thin layer at 84 km at 1245 UT. After Rapp, M., Strelnikova, I., Latteck, R., et al., 2008. Polar mesosphere summer echoes (PMSE) studied at Bragg wavelengths of 2.8 m, 67 cm, and 16 cm. *Journal of Atmospheric and Solar-Terrestrial Physics* 70, 947–961.

density measured by the rocket during a strong event. A severe, sharp biteout in the latter is coincident with the echoing region. Biteouts like this have been observed by numerous rockets as well as by ground-based incoherent scatter radars. These biteouts form when dust or ice particles become charged by

collecting free electrons and thus are an indirect indication that these particles are present inside the PMSE layer.

Ice particles can be detected directly using optical techniques, since ice particles will scatter optical light. [Figure 9](#) is a simultaneous measurement of lidar and VHF radar

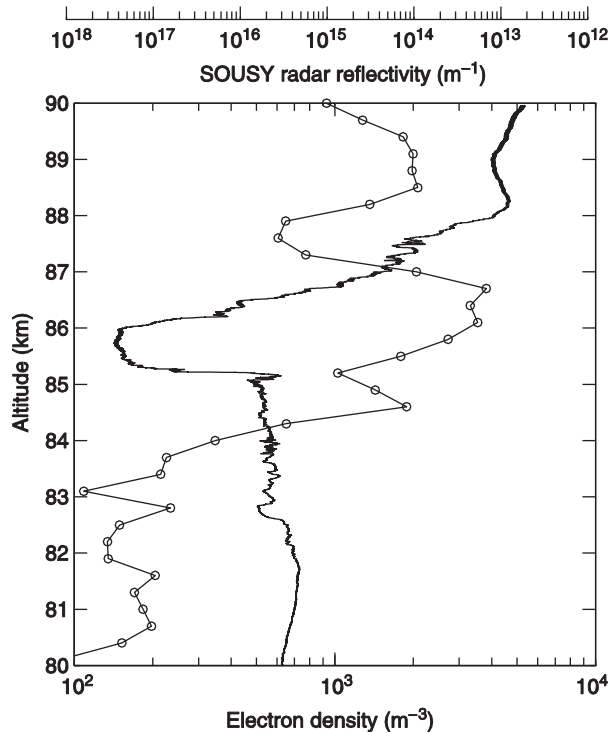


Figure 8 Comparison of rocket measurements of the electron density profile (solid line) with the simultaneous 53.5 MHz SOUSY radar observations of radar reflectivity (circles) during the MAC/SINE campaign (14 July 1987; 0929 UT). Adapted from Alcalá, C.M., Kelley, M.C., 2001. Nonturbulent layers in polar mesosphere summer mesosphere, 2. Application of wavelet analysis to VHF scattering. *Radio Science* 36, 891.

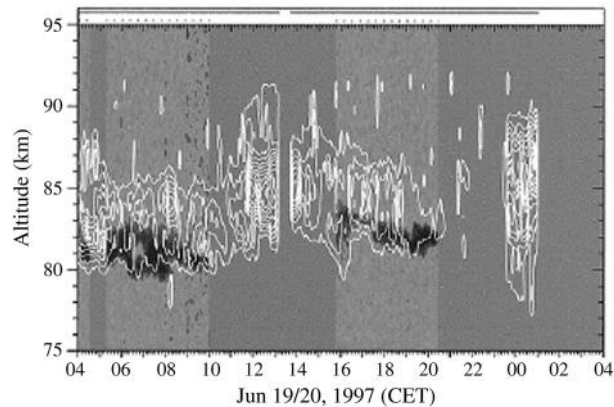


Figure 9 Relationship between PMSEs (light contours) detected by radar and NLC particles (shaded regions) detected by lidar. Adapted from von Zahn, U., Bremer, J., 1999. Simultaneous and common-volume observations of noctilucent clouds and polar mesosphere summer echoes. *Geophysical Research Letters* 26, 1521.

backscatter from the polar summer mesosphere. A lidar is a radar that uses laser light instead of radio waves. The lidar is sensitive to the largest ice particles, the same particles that can be viewed from the ground as NLCs. Notice that the lidar signal is at the lower edge of the radar signal, indicating that the radar

detects small particles as they fall, grow larger, and then are detected by the lidar just prior to sublimating as the temperature rises.

The occurrence of PMSEs in the summer polar mesosphere is undoubtedly linked to very cold temperatures. Simultaneous measurements of PMSEs with radar and of temperature with rockets showed that a sufficiently cold mesopause temperature is a necessary but an insufficient condition for the existence of PMSEs. A study of PMSE occurrence and low seasonal mesopause temperatures, the results of which are presented in [Figure 10](#), showed that the probability curve of seasonal PMSE occurrence in the Northern Hemisphere was similar to, but delayed by about 10 days from, a mean seasonal curve of low mesopause temperatures deduced from a number of measurements.

The cold mesopause/PMSE occurrence is further supported by measurements of PMSEs in the Southern Hemisphere at Machu Picchu station (62° S), which showed that PMSEs are much weaker and more sporadic in the Southern Hemisphere. One reasonable inference emerging from this observation is that the mean mesopause temperature at southern high latitudes is warmer than its northern counterpart. Indeed, preliminary examination of daily averaged summertime temperatures using satellite data (High Resolution Doppler Imager) at 84 ± 1.5 km suggests a few degrees of temperature difference between comparable summertime months found at 65° latitude, with the Southern Hemispheric temperatures being warmer. These results support previous indications of a warmer Southern Hemisphere from other satellite observations (Solar Mesosphere Explorer) when the temperatures were warmer in the Southern Hemisphere by about 4 K.

Another study supporting north-south asymmetries suggests that PMSEs can be detected at lower latitudes in the Northern Hemisphere than in the Southern Hemisphere, and it seems that PMSEs start occurring in an earlier stage in the south than in the north with respect to the solstice. These results could help in understanding why PMSEs were not detected late in the season at Machu Picchu station. Data collection at Machu Picchu station has not been possible earlier than late December for logistic reasons, thereby missing late November and early December observations when PMSE occurrence is supposed to start.

All of this discussion requires that PMSE generation somehow be associated with charged aerosols and, primarily, ice particles, suggesting that again, in addition to cold temperatures, water vapor is also necessary for PMSE generation. The differences between seasonal PMSE variation and temperature were speculated to be caused by water vapor.

Many mechanisms have been proposed as being responsible, or at least partly responsible, for PMSE generation. While it has been suggested that ice particles themselves scatter radio waves just as rain and snow enhance weather radar signals, such particles are too few and too small to create a huge signal. The observed echo strengths can only be generated by scattering from free electrons with a substantial amount of structure at the Bragg wavelength. One explanation for how such structuring could occur has received more theoretical attention and more observational evidence than any other: a modified version of turbulence theory that involves ambipolar diffusion.

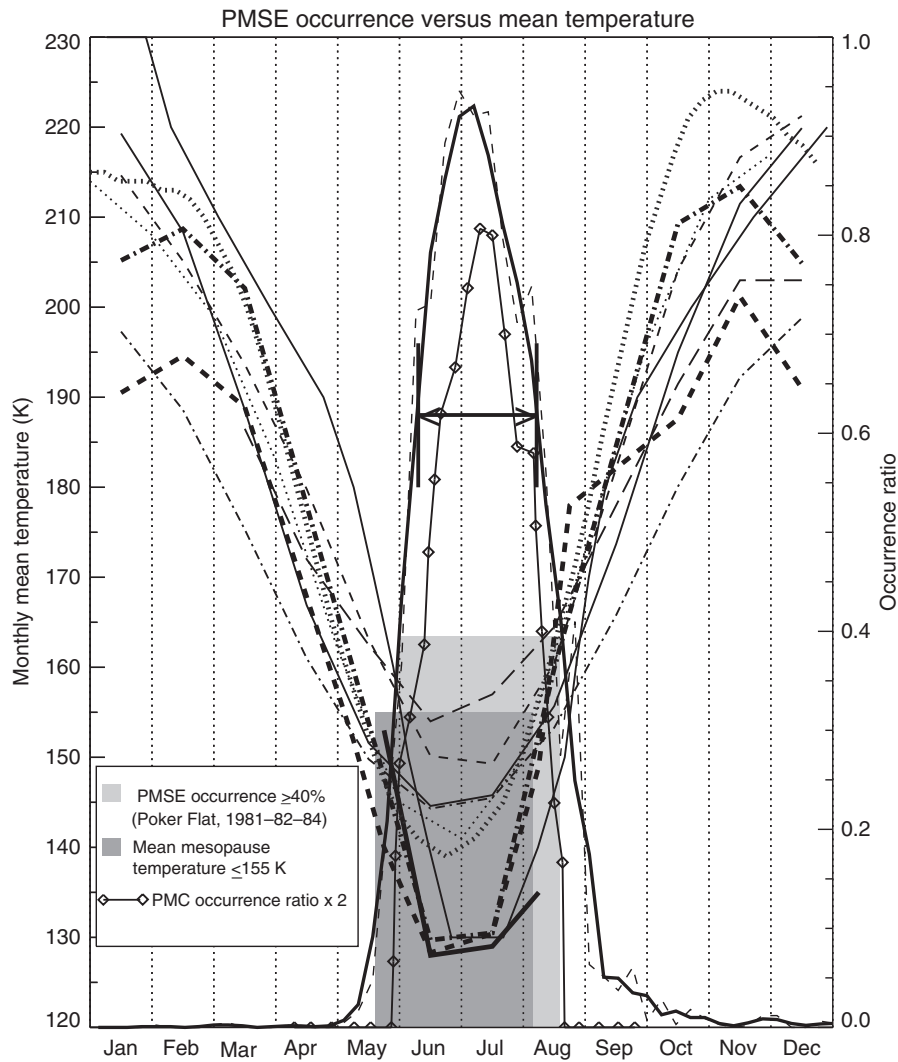


Figure 10 Comparison of polar mesospheric summer echoes (PMSEs) percentage of occurrence at Poker Flat, Alaska, and various estimates of high-latitude seasonal mesospheric temperature fluctuations. Adapted from Balsley, B.B., Huaman, M.M., 1997. On the relationship between seasonal occurrence in northern hemispheric polar mesosphere summer echoes and mean mesopause temperatures. *Journal of Geophysical Research* 102, 2021.

Free electrons in the D-region ionosphere are structured by neutral turbulence because the electron-neutral collision frequency is so high that electrons essentially follow the neutrals at these altitudes. However, this does not mean that the electron density fluctuation spectrum must be identical to the kinetic energy turbulence spectrum. Ultimately, neutral viscosity limits the smallest scales that can form in the kinetic energy spectrum, whereas electron diffusion limits the smallest scales that can form in the electron density fluctuation spectrum. A commonly used dimensionless number is the Schmidt number, Sc , the ratio of viscosity of air to the electron diffusivity. If $Sc \sim 1$, the electron density fluctuation spectrum will be similar to the neutral kinetic energy spectrum (i.e., it will have clearly defined inertial and viscous subranges). When $Sc \gg 1$, however, the electron density fluctuation spectrum will be divided into three subranges. The largest scales will be in the inertial-convective subrange. Near the Kolmogorov microscale, the spectrum will transition to the viscous-convective

subrange. In this subrange, viscosity is important but diffusion is not. Finally, the spectrum transitions to the viscous-diffusive subrange and falls sharply. The dividing line between the viscous-convective and viscous-diffusive subranges is near the Batchelor microscale, which is smaller than the Kolmogorov microscale by a factor of $1/\sqrt{Sc}$. Thus, when $Sc \gg 1$ (i.e., when the electron diffusivity is anomalously low), the electron density fluctuation spectrum can be extended to much smaller scales than usual. To explain observations of PMSEs by UHF radars, Schmidt numbers of 3000–8000 are needed.

Diffusion of electrons in a plasma works differently from diffusion in neutral gases because the electrons are charged and thus coupled to all the other charged species in the plasma via electric fields. Consider a blob of plasma that is initially electrically neutral and composed of electrons and one type of ions. The electrons want to diffuse away much more quickly than the heavier ions, but if they did, they would

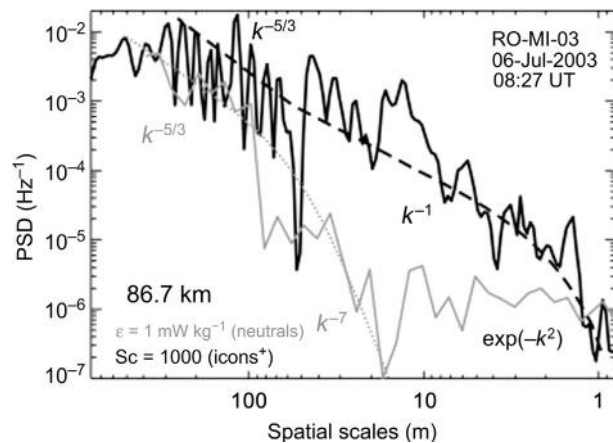


Figure 11 Simultaneous rocket measurements of a neutral kinetic energy spectrum (gray) and an ion density fluctuation spectrum (black) in the summer polar mesopause. The neutral spectrum shows a clearly defined inertial subrange at large scales ($k^{-5/3}$) and viscous subrange at small scales (k^{-7}). The ion spectrum shows a clearly defined inertial-convective subrange at large scales ($k^{-5/3}$), a viscous-convective subrange at intermediate scales (k^{-1}), and a viscous-diffusive subrange at small scales ($\exp(-k^2)$). Notice that the ion spectrum extends to much smaller scales than the neutral spectrum. After Strelnikov, B., Rapp, M., Zecha, M., et al., 2009. PMSE and E-region plasma instability: In situ observations. *Journal of Atmospheric and Solar-Terrestrial Physics* 71, 143–157.

leave behind a blob of ions with a net positive charge. An electric field forms, slowing the electron diffusion and accelerating the ion diffusion, and both species then diffuse together at the same rate. This field is known as the ambipolar electric field and the phenomenon is called ambipolar diffusion. Ambipolar diffusion is like a poorly trained dog that runs ahead and drags its owner along behind it while on a walk. In this analogy, the dog represents the electrons, the owner the ions, and the leash forcing the two to move together is the ambipolar electric field. The plasma in the polar summer mesopause is a three-component plasma consisting of electrons, ions, and charged dust or ice particles. Electrons and ions in such a plasma have two different ambipolar diffusion modes: a fast mode, close to the ion diffusion rate, and a slow mode, close to the dust/ice diffusion rate. **Figure 11** shows neutral and ion spectra measured by a rocket flying through a PMSE layer. The neutral spectrum has well-defined inertial and viscous subranges. The ion spectrum has well-defined inertial-convective, viscous-convective, and viscous-diffusive subranges. The ion spectrum is extended to much smaller scales than the neutral spectrum because of the slow ambipolar diffusion mode, which appears in the presence of charged dust/ice. The ambipolar mode can be so slow that the required Schmidt numbers of 3000–8000 for UHF PMSEs are reached if the ice particles have radii of 30 nm or more. Estimates from AIM and from ground-based multicolor lidar measurements have both observed ice particles with radii of 10–60 nm, meaning that the high Schmidt numbers required to explain UHF PMSEs using turbulence theory are readily obtainable.

On the Possible Relationship between PMSEs, NLCs, and Global Change

As time has passed, more observations of NLCs have been reported and an increasing trend has been detected. This increment in NLCs is explained by the observed increment in atmospheric methane (CH_4) and carbon dioxide (CO_2) due to human activity. A doubling increment in either of these components will produce a cooling of the thermosphere and mesosphere by about 50 and 10 K, respectively. Remember that cold temperatures are a necessary condition for NLC generation, although not the only one. About half of the mesospheric water vapor is believed to come from the photodissociation and oxidation of upwardly transported CH_4 with the chemically active radical (OH). Colder temperatures and more water vapor can produce more NLC events, so they can be used as indicators of global change.

The necessary conditions for PMSEs and NLCs or PMCs to occur appear similar: low temperatures are required and are apparently related to water vapor. The seasonal PMSE occurrence corresponds well with the high-latitude seasonal occurrence of NLCs. Recent studies have shown a correlation of PMCs with PMSEs in the Northern Hemisphere, where the mean long-term PMC occurrence ratio curve fits symmetrically inside the PMSEs occurrence ratio (**Figure 8**). A close correlation between NLCs and PMSEs has also been observed using lidar and radar data, respectively. These studies were made using a common volume and they agree most of the time. Thus, PMSEs are of particular interest in view of their frequent coincident occurrence with NLCs and the possible association of recent increased detection of NLCs with global warming trends. If we could monitor PMSEs for long periods and observe increases/decreases in these events (and their relative strength) over time, we could use such information as a possible indicator of global change.

See also: Climate and Climate Change: Carbon Dioxide. **Clouds and Fog:** Noctilucent Clouds. **Global Change:** Upper Atmospheric Change. **Gravity Waves:** Overview. **Numerical Models:** Parameterization of Physical Processes: Gravity Wave Fluxes. **Radar:** Mesosphere–Stratosphere–Troposphere and Stratosphere–Troposphere Radars and Wind Profilers.

Further Reading

- Gadsden, M., Schröder, W., 1989. *Noctilucent Clouds*. Springer-Verlag, New York.
- Rapp, M., Luebken, F.-J., 2004. Polar mesosphere summer echoes: review of observations and current understanding. *Atmospheric Chemistry and Physics* 4, 2601–2633.
- Röttger, J., 1994. Middle atmosphere and lower thermosphere processes at high latitudes studied with the EISCAT radars. *Journal of Atmospheric and Solar-Terrestrial Physics* 56, 1173–1196.
- Russell III, J.M., Bailey, S.M., Gordley, L.L., Rusch, D.W., Horányi, M., Hervig, M.E., Thomas, G.E., Randall, C.E., Siskind, D.E., Stevens, M.H., Summers, M.E., Taylor, M.J., Englert, C.R., Espy, P.J., McClintock, W.E., Merkel, A.W., 2009. The Aeronomy of Ice in the Mesosphere (AIM) mission: overview and early science results. *Journal of Atmospheric and Solar-Terrestrial Physics* 71, 289–299.
- Thomas, G.E., 1991. Mesospheric clouds and the physics of the mesopause region. *Reviews in Geophysics* 29, 553–575.

This page intentionally left blank

Second Edition

ENCYCLOPEDIA OF ATMOSPHERIC SCIENCES

Edited by
Gerald R. North **Editor-in-Chief**
John Pyle and Fuqing Zhang



6 VOLUME SET

ENCYCLOPEDIA OF **ATMOSPHERIC SCIENCES**

SECOND EDITION

VOLUME 4

This page intentionally left blank

ENCYCLOPEDIA OF **ATMOSPHERIC SCIENCES**

SECOND EDITION

EDITOR-IN-CHIEF

GERALD R NORTH

Texas A&M University, College Station, TX, USA

EDITORS

JOHN PYLE

Cambridge University, Cambridge, UK

FUQING ZHANG

Pennsylvania State University, University Park, PA, USA

VOLUME 4



Amsterdam • Boston • Heidelberg • London • New York • Oxford
Paris • San Diego • San Francisco • Singapore • Sydney • Tokyo
Academic Press is an imprint of Elsevier



Academic Press is an imprint of Elsevier
32 Jamestown Road, London NW1 7BY, UK
225 Wyman Street, Waltham, MA 02451, USA
525 B Street, Suite 1800, San Diego, CA 92101-4495, USA

Copyright © 2015 Elsevier Ltd. unless otherwise stated. All rights reserved.

No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means electronic, mechanical, photocopying, recording or otherwise without the prior written permission of the publisher

Permissions may be sought directly from Elsevier's Science & Technology Rights Department in Oxford, UK: phone (+44) (0) 1865 843830; fax (+44) (0) 1865 853333; email: permissions@elsevier.com. Alternatively you can submit your request online by visiting the Elsevier web site at <http://elsevier.com/locate/permissions>, and selecting *Obtaining permission to use Elsevier material*

Notice

No responsibility is assumed by the publisher for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions or ideas contained in the material herein, Because of rapid advances in the medical sciences, in particular, independent verification of diagnoses and drug dosages should be made

British Library Cataloguing in Publication Data

A catalog record for this book is available from the British Library

Library of Congress Catalog Number:

A catalog record for this book is available from the Library of Congress

ISBN (print): 978-0-12-382225-3

For information on all Elsevier publications
visit our website at store.elsevier.com

Printed and bound in the United Kingdom

15 16 17 18 19 10 9 8 7 6 5 4 3 2 1

		Working together to grow libraries in developing countries
www.elsevier.com • www.bookaid.org		

<i>Acquisitions Editor:</i> Simon Holt <i>Project Manager:</i> Michael Nicholls <i>Associate Project Manager:</i> Marise Willis <i>Designer:</i> Matthew Limbert

DEDICATION

This second edition of the *Encyclopedia of Atmospheric Sciences* is dedicated to the memory of James Holton who was editor-in-chief of the first edition. He was a great researcher and colleague inspiring an entire generation of atmospheric scientists.

This page intentionally left blank

CONTENTS

<i>List of Contributors</i>	xxvii
<i>Preface to the First Edition</i>	xxxix
<i>Preface to the Second Edition</i>	xli
<i>Editor Biographies</i>	xliii
<i>Guide to Using the Encyclopedia</i>	xlvi

VOLUME 1

BASIC ATMOSPHERIC STRUCTURE AND CONCEPTS **1**

Beaufort Wind Scale	1
<i>L Hasse</i>	
Wind Chill	7
<i>M Bluestein</i>	
Standard Atmosphere	12
<i>W W Vaughan</i>	

AEROSOLS **17**

Aerosol–Cloud Interactions and Their Radiative Forcing	17
<i>U Lohmann</i>	
Aerosol Physics and Chemistry	23
<i>M Kalberer</i>	
Climatology of Stratospheric Aerosols	32
<i>L W Thomason and J-P Vernier</i>	
Climatology of Tropospheric Aerosols	40
<i>N Bellouin and J Haywood</i>	
Dust	48
<i>I N Sokolik</i>	
Observations and Measurements	53
<i>P H McMurry</i>	
Role in Radiative Transfer	66
<i>G A Ban-Weiss, and W D Collins</i>	

Role in Climate Change <i>N Bellouin</i>	76
Soot <i>P Chylek, S G Jennings, and R Pinnick</i>	86
Agricultural Meteorology and Climatology <i>E S Takle</i>	92
ARCTIC AND ANTARCTIC	98
Antarctic Climate <i>J Turner</i>	98
Arctic Climate <i>M C Serreze</i>	107
Arctic Haze <i>L M Russell and G E Shaw</i>	116
AIR SEA INTERACTIONS	
Freshwater Flux <i>J Schulz</i>	122
Momentum, Heat, and Vapor Fluxes <i>P K Taylor</i>	129
Sea Surface Temperature <i>W J Emery</i>	136
Surface Waves <i>A Benilov</i>	144
AVIATION METEOROLOGY	153
Aircraft Emissions <i>R R Friedl</i>	153
Aircraft Icing <i>M K Politovich</i>	160
Aviation Weather Hazards <i>A J Bedard, Jr</i>	166
Clear Air Turbulence <i>G P Ellrod (Retired), J A Knox, P F Lester, and L J Ehernberger (Retired)</i>	177
BIOGEOCHEMICAL CYCLES	187
Sulfur Cycle <i>P Brimblecombe</i>	187
Bromine <i>R von Glasow and C Hughes</i>	194
Heavy Metals <i>T D Jickells and A R Baker</i>	201

Iodine	205
<i>L J Carpenter</i>	
BOUNDARY LAYER (ATMOSPHERIC) AND AIR POLLUTION	220
Overview	220
<i>P J Mason and D J Thomson</i>	
Air Pollution Meteorology	227
<i>X-M Hu</i>	
Coherent Structures	237
<i>F T M Nieuwstadt and J C R Hunt</i>	
Complex Terrain	242
<i>J J Finnigan</i>	
Convective Boundary Layer	250
<i>M A LeMone</i>	
Microclimate	258
<i>M W Rotach and P Calanca</i>	
Modeling and Parameterization	265
<i>A A M Holtslag</i>	
Observational Techniques <i>In Situ</i>	274
<i>E F Bradley</i>	
Observational Techniques: Remote	284
<i>W M Angevine and C J Senff</i>	
Ocean Mixed Layer	290
<i>L Kantha and C A Clayson</i>	
Stably Stratified Boundary Layer	299
<i>L Mahrt</i>	
Surface Layer	305
<i>G L Geernaert</i>	
Urban Heat Islands	310
<i>J C Luvall, D A Quattrochi, D L Rickman, and M G Estes, Jr</i>	
Diurnal Cycle	319
<i>A Betts</i>	
CHEMISTRY OF THE ATMOSPHERE	324
Chemical Kinetics	324
<i>R P Wayne</i>	
Ion Chemistry	333
<i>J L Fox</i>	
Isotopes, Stable	348
<i>C A M Brenninkmeijer</i>	
Laboratory Kinetics	356
<i>D J Donaldson and S N Wren</i>	

Methane <i>E Dlugokencky, and S Houweling</i>	363
Observations for Chemistry (<i>In Situ</i>): Ozone Sondes <i>H G J Smit</i>	372
Observations for Chemistry (<i>In Situ</i>): Particles <i>T Deshler</i>	379
Observations for Chemistry (<i>In Situ</i>): Water Vapor Sondes <i>J B Smith</i>	387
Observations for Chemistry (Remote Sensing): IR/FIR (Satellite, Balloon and Ground) <i>H Fischer and F Hase</i>	401
Observations for Chemistry (Remote Sensing): Lidar <i>G Vaughan</i>	411
Observations for Chemistry (Remote Sensing): Microwave <i>J Waters</i>	418
Principles of Chemical Change <i>R P Wayne</i>	429
Radioactivity: Cosmogenic Radionuclides <i>D Lal</i>	437
Volcanoes: Composition of Emissions <i>M T Coffey and J W Hannigan</i>	446
Tracers <i>K A Boering</i>	450

VOLUME 2

CLIMATE AND CLIMATE CHANGE	1
Overview <i>D L Hartmann</i>	1
Carbon Dioxide <i>C L Sabine and R A Feely</i>	10
Climate Feedbacks <i>A E Dessler and M D Zelinka</i>	18
Climate Prediction: Empirical and Numerical <i>S Hastenrath</i>	26
Climate Variability: Decadal to Centennial Variability <i>D G Martinson</i>	33
Climate Variability: Nonlinear and Random Effects <i>M Ghil</i>	38
Climate Variability: North Atlantic and Arctic Oscillation <i>J W Hurrell</i>	47
Climate Variability: Seasonal and Interannual Variability <i>D S Gutzler</i>	61

Energy Balance Climate Models <i>G R North and K-Y Kim</i>	69
Global Impacts of the Madden–Julian Oscillation <i>C Zhang</i>	73
Greenhouse Effect <i>G R North</i>	80
History of Scientific Work on Climate Change <i>S Weart</i>	87
Intergovernmental Panel on Climate Change <i>K E Trenberth</i>	90
Nuclear Winter <i>A Robock</i>	95
Radiative–Convective Equilibrium Climate Models <i>N O Renno and X Huang</i>	102
Volcanoes: Role in Climate <i>A Robock</i>	105
CLOUDS AND FOG	112
Cloud Modeling <i>W-K Tao and M Moncrieff</i>	112
Contrails <i>P Minnis</i>	121
Cloud Microphysics <i>D Lamb</i>	133
Classification of Clouds <i>A L Rangno (Retiree)</i>	141
Climatology <i>S Warren, R Eastman, and C J Hahn</i>	161
Measurement Techniques <i>In situ</i> <i>D Baumgardner, J-F Gayet, A Korolev, C Twohy, and J Fugal</i>	170
Fog <i>P J Croft and B Ward</i>	180
Noctilucent Clouds <i>G E Thomas</i>	189
Stratus and Stratocumulus <i>R Wood</i>	196
CRYOSPHERE	201
Glaciers, Topography, and Climate <i>A B G Bush and M P Bishop</i>	201
Permafrost <i>T E Osterkamp and C R Burn</i>	208

Sea Ice	217
<i>M C Serreze, F Fetterer, and W F Weeks (Retired)</i>	
Snow (Surface)	227
<i>M Sturm</i>	
DATA ASSIMILATION AND PREDICTABILITY	237
Data Assimilation	237
<i>A C Lorenc</i>	
Ensemble-Based Data Assimilation	241
<i>Z Meng and F Zhang</i>	
Ensemble Prediction	248
<i>R Buizza</i>	
Predictability and Chaos	258
<i>L A Smith</i>	
DYNAMICAL METEOROLOGY	265
Overview	265
<i>J R Holton</i>	
Acoustic Waves	272
<i>K E Gilbert</i>	
Atmospheric Tides	287
<i>J Oberheide, M E Hagan, A D Richmond, and J M Forbes</i>	
Balanced Flow	298
<i>M E McIntyre</i>	
Baroclinic Instability	304
<i>R Grotjahn</i>	
Coriolis Force	313
<i>D W Moore</i>	
Critical Layers	317
<i>P Haynes</i>	
Hamiltonian Dynamics	324
<i>T G Shepherd</i>	
Hydraulic Flow	332
<i>R B Smith</i>	
Inertial Instability	334
<i>J A Knox</i>	
Kelvin–Helmholtz Instability	343
<i>P G Drazin</i>	
Kelvin Waves	347
<i>B Wang</i>	
Kinematics	353
<i>D D Houghton</i>	

Laboratory Geophysical Fluid Dynamics <i>R L Pfeffer</i>	360
Lagrangian Dynamics <i>I Roulstone</i>	369
Potential Vorticity <i>M E McIntyre</i>	375
Primitive Equations <i>A A White and N Wood</i>	384
Quasigeostrophic Theory <i>H C Davies and H Wernli</i>	393
Rossby Waves <i>P B Rhines</i>	404
Solitary Waves <i>J P Boyd</i>	417
Static Stability <i>J A Young</i>	423
Stationary Waves (Orographic and Thermally Forced) <i>S Nigam and E DeWeaver</i>	431
Symmetric Stability <i>H B Bluestein</i>	446
Vorticity <i>J R Holton</i>	451
Wave-CISK <i>C S Bretherton</i>	455
Wave Mean-Flow Interaction <i>M Juckes</i>	458
Waves <i>J R Holton</i>	464

VOLUME 3

ELECTRICITY IN THE ATMOSPHERE	1
Global Electrical Circuit <i>E R Williams</i>	1
Ions in the Atmosphere <i>K L Aplin and R G Harrison</i>	9
Lightning <i>M B Baker</i>	14
Sprites <i>W A Lyons</i>	20
Forensic Meteorology <i>L E Branscome</i>	28

GENERAL CIRCULATION OF THE ATMOSPHERE	33
Overview	33
<i>J M Wallace, D W J Thompson, and P Beresford</i>	
Angular Momentum of the Atmosphere	43
<i>D A Salstein</i>	
Energy Cycle	51
<i>R Grotjahn</i>	
Weather Regimes and Multiple Equilibria	65
<i>F Molteni</i>	
Mean Characteristics	73
<i>R Grotjahn</i>	
Teleconnections	90
<i>S Nigam and S Baxter</i>	
 GLOBAL CHANGE	 110
Climate Record: Surface Temperature Trends	110
<i>P D Jones</i>	
Sea Level Change	121
<i>R S Nerem</i>	
Upper Atmospheric Change	128
<i>R G Roble</i>	
Biospheric Impacts and Feedbacks	132
<i>B A Hungate and G W Koch</i>	
 GRAVITY WAVES	 141
Overview	141
<i>D C Fritts</i>	
Buoyancy and Buoyancy Waves: Optical Observations	153
<i>M J Taylor and W R Pendleton, Jr</i>	
Buoyancy and Buoyancy Waves: Theory	160
<i>T J Dunkerton</i>	
Gravity Waves Excited by Jets and Fronts	164
<i>R Plougonven and F Zhang</i>	
Convectively Generated Gravity Waves	171
<i>T P Lane</i>	
 HYDROLOGY, FLOODS AND DROUGHTS	 180
Overview	180
<i>R C Bales</i>	
Deserts and Desertification	185
<i>V P Tchakerian</i>	
Drought	193
<i>S Quiring</i>	

Flooding	201
<i>C A Doswell III</i>	
Groundwater and Surface Water	209
<i>S Ge and S M Gorelick</i>	
Modeling and Prediction	217
<i>Z Yu</i>	
Palmer Drought Severity Index	224
<i>L Nkemdirim</i>	
Soil Moisture	232
<i>A Robock</i>	
LAND-ATMOSPHERE INTERACTIONS	240
Overview	240
<i>R E Dickinson</i>	
Canopy Processes	244
<i>P D Blanken</i>	
Trace Gas Exchange	256
<i>J N Cape and D Fowler</i>	
LIDAR	262
Atmospheric Sounding Introduction	262
<i>P S Argall and R Sica</i>	
Backscatter	270
<i>C M R Platt and R L Collins</i>	
Differential Absorption Lidar	277
<i>S Ismail and E V Browell</i>	
Doppler	289
<i>R M Hardesty</i>	
Raman	296
<i>D N Whiteman</i>	
Resonance	305
<i>C S Gardner and R L Collins</i>	
Magnetosphere	309
<i>G K Parks</i>	
MESOSCALE METEOROLOGY	316
Overview	316
<i>D J Parker</i>	
Cloud and Precipitation Bands	323
<i>R M Rauber and M Ramamurthy</i>	
Gust Fronts	331
<i>R Rotunno</i>	

Hail and Hailstorms	334
<i>C Knight, N Knight, and H E Brooks</i>	
Mesoscale Convective Systems	339
<i>A Laing</i>	
Microbursts	335
<i>R M Wakimoto</i>	
Severe Storms	361
<i>C A Doswell III</i>	
Waterspouts	369
<i>J H Golden</i>	
Bow Echoes and Derecho	384
<i>M L Weisman</i>	
Density Currents	395
<i>P G Baines</i>	
Convective Storms: Overview	401
<i>M L Weisman</i>	
MESOSPHERE	411
Atomic Species in the Mesopause Region	411
<i>M G Mlynczak and L A Hunt</i>	
Ionosphere	422
<i>M C Kelley</i>	
Metal Layers	430
<i>J M C Plane</i>	
Polar Summer Mesopause	436
<i>R H Varney and M C Kelley</i>	
VOLUME 4	
MIDDLE ATMOSPHERE	1
Planetary Waves	1
<i>A K Smith and J Perlwitz</i>	
Polar Vortex	12
<i>M R Schoeberl and P A Newman</i>	
Quasi-Biennial Oscillation	18
<i>T J Dunkerton, J A Anstey, and L J Gray</i>	
Semiannual Oscillation	26
<i>K Hamilton</i>	
Stratospheric Sudden Warmings	30
<i>A O'Neill, A J Charlton-Perez, and L M Polvani</i>	
Transport Circulation	41
<i>S E Strahan</i>	

Zonal Mean Climatology <i>P Braesicke</i>	50
MOUNTAIN METEOROLOGY	57
Overview <i>R B Smith</i>	57
Cold Air Damming <i>B A Colle</i>	62
Downslope Winds <i>D R Durran</i>	69
Katabatic Winds <i>T R Parish</i>	75
Land and Sea Breezes <i>R A Pielke, Sr</i>	80
Lee Vortices <i>C C Epifanio</i>	84
Lee Waves and Mountain Waves <i>D R Durran</i>	95
Orographic Effects: Lee Cyclogenesis <i>C Schär</i>	103
Valley Winds <i>D Zardi</i>	114
NUMERICAL MODELS	135
Chemistry Models <i>M P Chipperfield and S R Arnold</i>	135
Coupled Ocean-Atmosphere Models: Physical Processes <i>M Zhang</i>	144
General Circulation Models <i>C R Mechoso and A Arakawa</i>	153
Methods <i>J Thuburn</i>	161
Model Physics Parameterization <i>D J Stensrud, M C Coniglio, K H Knopfmeier, and A J Clark</i>	167
Parameter Estimation <i>A Aksoy</i>	181
Parameterization of Physical Processes: Clouds <i>R Forbes, C Jakob, and M Miller</i>	187
Parameterization of Physical Processes: Gravity Wave Fluxes <i>M J Alexander</i>	194
Parameterization of Physical Processes: Turbulence and Mixing <i>A Beljaars</i>	200

Spectral Models <i>F Baer</i>	212
Mesoscale Atmospheric Modeling <i>R A Pielke, Sr</i>	219
Cloud-System Resolving Modeling and Aerosols <i>W-K Tao and T Matsui</i>	222
Large-Eddy Simulation <i>C-H Moeng and P P Sullivan</i>	232
Regional Prediction Models <i>B W Golding</i>	241
Convective Storm Modeling <i>M D Parker</i>	246
OBSERVATIONS PLATFORMS	255
Balloons <i>J-P Pommereau</i>	255
Buoys <i>J M Hemsley</i>	264
Kites <i>B B Balsley</i>	268
Radiosondes <i>W F Dabberdt and H Turtiainen</i>	273
Rockets <i>M F Larsen</i>	285
OCEANOGRAPHIC TOPICS	290
General Processes <i>N C Wells</i>	290
Surface/Wind Driven Circulation <i>R X Huang</i>	301
Thermohaline Circulation <i>R X Huang</i>	315
Water Types and Water Masses <i>W J Emery</i>	329
OPTICS, ATMOSPHERIC	338
Optical Remote Sensing Instruments <i>G G Shepherd</i>	338
Airglow Instrumentation <i>M Conde</i>	346

OZONE DEPLETION AND RELATED TOPICS	353
Long-Term Ozone Changes <i>N R P Harris</i>	353
Ozone as a UV Filter <i>J E Frederick</i>	359
Ozone Depletion Potentials <i>D J Wuebbles</i>	364
Photochemistry of Ozone <i>G K Moortgat and A R Ravishankara</i>	370
Stratospheric Ozone Recovery <i>D J Hofmann and R Müller</i>	380
Surface Ozone Effects on Vegetation <i>M Ashmore</i>	389
Surface Ozone (Human Health) <i>M Lippmann</i>	397
PALEOCLIMATOLOGY	404
Ice Cores <i>E J Steig</i>	404
Varves <i>R Gilbert</i>	411
RADAR	415
Cloud Radar <i>T Uttal</i>	415
Incoherent Scatter Radar <i>M P Sulzer</i>	422
Mesosphere—Stratosphere—Troposphere and Stratosphere—Troposphere Radars and Wind Profilers <i>G Vaughan and D Hooper</i>	429
Meteor Radar <i>N J Mitchell</i>	438
Polarimetric Doppler Weather Radar <i>R J Doviak and R D Palmer</i>	444
Precipitation Radar <i>S E Yuter</i>	455
Synthetic Aperture Radar (Land Surface Applications) <i>R K Vincent</i>	470
VOLUME 5	
RADIATION TRANSFER IN THE ATMOSPHERE	1
Radiation, Solar <i>Q Fu</i>	1

Absorption and Thermal Emission <i>R M Goody and X Huang</i>	5
Cloud-Radiative Processes <i>Q Fu</i>	13
Non-local Thermodynamic Equilibrium <i>M López-Puertas and B Funke</i>	16
Scattering <i>M Mishchenko, L Travis, and A Lacis</i>	27
Ultraviolet Radiation <i>K Stamnes</i>	37
Ultraviolet, Surface <i>R McKenzie and S Madronich</i>	45
SATELLITES AND SATELLITE REMOTE SENSING	51
Aerosol Measurements <i>R A Kahn</i>	51
Earth's Radiation Budget <i>N G Loeb and B A Wielicki</i>	67
GPS Meteorology <i>S S Leroy</i>	77
Measuring Ozone from Space — TOMS and SBUV <i>R D McPeters and R S Stolarski</i>	87
Orbits <i>S Q Kidder</i>	95
Precipitation <i>G Liu</i>	107
Remote Sensing: Cloud Properties <i>P Yang and B A Baum</i>	116
Research <i>M D King</i>	128
Surface Wind and Stress <i>W T Liu</i>	138
Temperature Soundings <i>A Dudhia</i>	145
Water Vapor <i>J E Harries</i>	157
SOLAR SYSTEM/SUN, ATMOSPHERES, EVOLUTION OF ATMOSPHERES	163
Evolution of Earth's Atmosphere <i>Y L Yung, M L Wong, and E J Gaidos</i>	163
Planetary Atmospheres: Mars <i>R M Haberle</i>	168

Planetary Atmospheres: Venus <i>P J Gierasch and Y L Yung</i>	178
Solar Terrestrial Interactions: Climate Impact <i>J D Haigh</i>	183
Solar Winds <i>S T Suess and B T Tsurutani</i>	189
Meteors <i>P Jenniskens</i>	195
STATISTICAL METHODS	201
Data Analysis: Empirical Orthogonal Functions and Singular Vectors <i>C S Bretherton</i>	201
Data Analysis: Time Series Analysis <i>G R North</i>	205
STRATOSPHERIC CHEMISTRY TOPICS	211
Overview <i>J A Pyle</i>	211
Halogens <i>D Toohey</i>	215
Halogen Sources, Anthropogenic <i>A McCulloch and P M Midgley</i>	221
Halogen Sources, Natural (Methyl Bromide and Related Gases) <i>S Yvon-Lewis and J H Butler</i>	228
HO _x <i>T F Hanisco</i>	233
Hydrogen Budget <i>J E Harries</i>	238
Reactive Nitrogen (NO _x and NO _y) <i>Y Kondo</i>	242
Stratospheric Water Vapor <i>K H Rosenlof</i>	250
STRATOSPHERE/TROPOSPHERE EXCHANGE AND STRUCTURE	257
Global Aspects <i>J R Holton</i>	257
Local Processes <i>J F Lamarque and P Hess</i>	262
Tropopause <i>M Dameris</i>	269

SYNOPTIC METEOROLOGY	273
Anticyclones <i>S J Colucci</i>	273
Forecasting <i>D Mansfield</i>	280
Weather Maps <i>R Reynolds</i>	289
Cyclogenesis <i>G J Hakim</i>	299
Extratropical Cyclones <i>A Joly</i>	304
Fronts <i>D M (David) Schultz and W Blumen</i>	337
Fronts in the Lower Stratosphere <i>A L Lang</i>	344
Frontogenesis <i>D M (David) Schultz</i>	353
Jet Streaks <i>P Cunningham and D Keyser</i>	359
Lake-Effect Storms <i>P J Sousounis</i>	370
Polar Lows <i>I A Renfrew</i>	379
Thermal Low <i>R H Johnson</i>	386
 THERMODYNAMICS	 391
Humidity Variables <i>J A Curry</i>	391
Moist (Unsaturated) Air <i>J A Curry</i>	394
Saturated Adiabatic Processes <i>J A Curry</i>	398
 Thermosphere <i>S C Solomon and R G Roble</i>	 402
 VOLUME 6	
TROPICAL CYCLONES AND HURRICANES	1
Overview and Theory <i>R A Tomas and P J Webster</i>	1

Hurricane Dynamics <i>Y Wang</i>	8
Hurricane Predictability <i>J A Sippel</i>	30
Hurricanes: Observation <i>F D Marks</i>	35
Tropical Cyclogenesis <i>Z Wang</i>	57
Tropical Cyclones and Climate Change <i>T R Knutson</i>	65
Tropical Cyclones in the Western North Pacific <i>J C L Chan</i>	77
Tropical Cyclones: Secondary Eyewall Formation <i>C-C Wu and Y-H Huang</i>	85
TROPICAL METEOROLOGY AND CLIMATE	91
El Niño and the Southern Oscillation: Observation <i>N Nicholls</i>	91
El Niño and the Southern Oscillation: Theory <i>P Chang and S E Zebiak</i>	97
Equatorial Waves <i>M C Wheeler and H Nguyen</i>	102
Hadley Circulation <i>J Lu and G A Vecchi</i>	113
Intertropical Convergence Zone <i>D E Waliser and X Jiang</i>	121
Intraseasonal Oscillation (Madden–Julian Oscillation) <i>R A Madden</i>	132
Madden–Julian Oscillation: Skeleton and Conceptual Models <i>A J Majda and S N Stechmann</i>	137
Monsoon: Overview <i>J Slingo</i>	146
Monsoon: Dynamical Theory <i>P J Webster and J Fasullo</i>	151
Monsoon: ENSO–Monsoon Interactions <i>K-M Lau</i>	165
Tropical Climates <i>S Hastenrath</i>	170
Walker Circulation <i>K-M Lau and S Yang</i>	177

TROPOSPHERIC CHEMISTRY AND COMPOSITION	182
Aerosols/Particles <i>J H Seinfeld</i>	182
Aliphatic Hydrocarbons <i>J Rudolph and O Stein</i>	188
Aromatic Hydrocarbons <i>I Barnes</i>	204
Biogenic Hydrocarbons <i>A Guenther</i>	214
Cloud Chemistry <i>P Herckes and J L Collett, Jr</i>	218
H ₂ <i>U Schmidt and T Wetter</i>	226
Hydroxyl Radical <i>K C Clemitshaw</i>	232
Mercury <i>J Munthe and J Sommar</i>	239
Oxidizing Capacity <i>D H Ehhalt, F Rohrer, and A Wahner</i>	243
Peroxyacetyl Nitrate <i>H B Singh</i>	251
Sulfur Chemistry, Organic <i>I Barnes</i>	255
Volatile Organic Compounds Overview: Anthropogenic <i>R G Derwent</i>	265
TURBULENCE AND MIXING	268
Overview <i>P Haynes</i>	268
Turbulence, Two Dimensional <i>P Bartello</i>	273
Turbulent Diffusion <i>A Venkatram and S Du</i>	277
WEATHER FORECASTING	287
Marine Meteorology <i>L Xie and B Liu</i>	287
Operational Meteorology <i>D R Novak</i>	293
Seasonal and Interannual Weather Prediction <i>J P Li and R Q Ding</i>	303
Severe Weather Forecasting <i>D J Stensrud, H E Brooks, and S J Weiss</i>	313

Wildfire Weather <i>J Coen</i>	323
Inadvertant Weather Modification <i>S A Changnon</i>	332
<i>Appendix 1: Physical Constants</i>	337
<i>Appendix 2: Units and their SI Equivalents</i>	339
<i>Appendix 3: Periodic Table of the Elements</i>	340
<i>Appendix 4: The Geologic Time Scale</i>	341
<i>Index</i>	343

This page intentionally left blank

LIST OF CONTRIBUTORS

A. Aksoy
University of Miami, Miami, FL, USA; and NOAA Hurricane Research Division, Miami, FL, USA

M.J. Alexander
NorthWest Research Associates (NWRA), Boulder, CO, USA

W.M. Angevine
CIRES, University of Colorado; and NOAA Earth System Research Laboratory, Boulder, CO, USA

J.A. Anstey
University of Oxford, Oxford, UK

K.L. Aplin
University of Oxford, Oxford, UK

A. Arakawa
University of California, Los Angeles, CA, USA

P.S. Argall
The University of Western Ontario, London, ON, Canada

S.R. Arnold
University of Leeds, Leeds, UK

M. Ashmore
University of York, York, UK

F. Baer
University of Maryland, College Park, MD, USA

P.G. Baines
University of Melbourne, Melbourne, VIC, Australia

A.R. Baker
University of East Anglia, Norwich, UK

M.B. Baker
University of Washington, Seattle, WA, USA

R.C. Bales
University of Arizona, Tucson, AZ, USA

B.B. Balsley
University of Colorado, Boulder, CO, USA

G.A. Ban-Weiss
Lawrence Berkeley National Laboratory, Berkeley, CA, USA; and University of Southern California, Los Angeles, CA, USA

I. Barnes
University of Wuppertal, Wuppertal, Germany

P. Bartello
McGill University, Montréal, QC, Canada

B.A. Baum
University of Wisconsin–Madison, Madison, WI, USA

D. Baumgardner
Universidad Nacional Autónoma de México, Mexico City, D.F., Mexico

S. Baxter
University of Maryland, College Park, MD, USA

A.J. Bedard, Jr.
National Oceanic and Atmospheric Administration, Boulder, CO, USA

A. Beljaars
European Centre for Medium-Range Weather Forecasts, Reading, England

N. Bellouin
University of Reading, Reading, UK

A. Benilov
Acute Solutions, Highlands, NJ, USA

P. Beresford
European Centre for Medium-Range Weather Forecasts, Reading, UK

A. Betts
Atmospheric Research, Pittsford, VT, USA

M.P. Bishop
Texas A&M University, College Station, TX, USA

P.D. Blanken
University of Colorado at Boulder, Boulder, CO, USA

H.B. Bluestein
University of Oklahoma, Norman, OK, USA

M. Bluestein
Indiana University – Purdue University, Indianapolis, IN, USA

W. Blumen[†]
University of Colorado Boulder, Boulder, CO, USA

K.A. Boering
University of California – Berkeley, Berkeley, CA, USA

J.P. Boyd
University of Michigan, Ann Arbor, MI, USA

E.F. Bradley
CSIRO Land and Water, Canberra, ACT, Australia

P. Braesicke
Karlsruhe Institute of Technology, Karlsruhe, Germany

L.E. Branscome
Climatological Consulting Corporation, FL, USA

C.A.M. Brenninkmeijer
Max Planck Institute for Chemistry, Mainz, Germany

C.S. Bretherton
University of Washington, Seattle, WA, USA

P. Brimblecombe
University of East Anglia, Norwich, UK

H.E. Brooks
National Oceanic and Atmospheric Administration, Norman, OK, USA

E.V. Browell
STARS II Affiliate, NASA Langley Research Center, Hampton, VA, USA

R. Buizza
ECMWF, Reading, UK

C.R. Burn
Carleton University, Ottawa, ON, Canada

A.B.G. Bush
University of Alberta, Edmonton, AB, Canada

J.H. Butler
NOAA Earth System Research Laboratory, Boulder, CO, USA

P. Calanca
Agroscope Reckenholz-Taenikon, Zurich, Switzerland

J.N. Cape
Edinburgh Research Station, Midlothian, UK

L.J. Carpenter
University of York, York, UK

J.C.L. Chan
City University of Hong Kong, Hong Kong

P. Chang
Texas A&M University, College Station, TX, USA

S.A. Changnon
University of Illinois, IL, USA

A.J. Charlton-Perez
University of Reading, Earley Gate, Reading, UK

M.P. Chipperfield
University of Leeds, Leeds, UK

P. Chylek
Dalhousie University, NS, Canada

A.J. Clark
University of Oklahoma and National Oceanic and Atmospheric Administration, Norman, OK, USA

C.A. Clayson
Woods Hole Oceanographic Institution, Woods Hole, MA, USA

K.C. Clemitshaw
Imperial College of Science, Technology, and Medicine, Ascot, UK

J. Coen
National Center for Atmospheric Research, Boulder, CO, USA

M.T. Coffey
National Center for Atmospheric Research, Boulder, CO, USA

B.A. Colle
Stony Brook University – SUNY, Stony Brook, NY, USA

J.L. Collett, Jr.
Colorado State University, Fort Collins, CO, USA

R.L. Collins
University of Alaska Fairbanks, Fairbanks, AK, USA

W.D. Collins
Lawrence Berkeley National Laboratory, Berkeley, CA, USA

S.J. Colucci
Cornell University, Ithaca, NY, USA

M. Conde
University of Alaska Fairbanks, Fairbanks, AK, USA

M.C. Coniglio
National Oceanic and Atmospheric Administration, Norman, OK, USA

[†] Deceased.

- P.J. Croft
Kean University, Union, NJ, USA
- P. Cunningham
Florida State University, Tallahassee, FL, USA
- J.A. Curry
Georgia Institute of Technology, Atlanta, GA, USA
- W.F. Dabberdt
Vaisala Company, Boulder, CO, USA
- M. Dameris
Deutsches Zentrum für Luft- und Raumfahrt, Institut für Physik der Atmosphäre, Oberpfaffenhofen, Wessling, Germany
- H.C. Davies
Institute for Atmospheric and Climate Science, Swiss Federal Institute of Technology (ETH), Zürich, Switzerland
- R.G. Derwent
rdscientific, Newbury, UK
- T. Deshler
University of Wyoming, Laramie, WY, USA
- A.E. Dessler
Texas A&M University, College Station, TX, USA
- E. DeWeaver
University of Wisconsin, Madison, WI, USA
- R.E. Dickinson
University of Texas at Austin, Austin, TX, USA
- R.Q. Ding
Institute of Atmospheric Physics, Chinese Academy of Sciences, Beijing, China
- E. Dlugokencky
NOAA Earth System Research Laboratory, Boulder, CO, USA
- D.J. Donaldson
University of Toronto, Toronto, ON, Canada
- C.A. Doswell, III
University of Oklahoma, Norman, OK, USA
- R.J. Doviak
National Severe Storms Laboratory, Norman, OK, USA
- P.G. Drazin[†]
University of Bath, England, UK
- S. Du
California Air Resources Board, Sacramento, CA, USA
- A. Dudhia
University of Oxford, Oxford, UK
- T.J. Dunkerton
Northwest Research Associates, Bellevue, WA, USA
- D.R. Durran
University of Washington, Seattle, WA, USA
- R. Eastman
University of Washington, Seattle, WA, USA
- L.J. Ehernberger
National Aeronautics and Space Administration, Dryden Flight Research Center, Edwards, CA, USA
- D.H. Ehhalt
Forschungszentrum Jülich, Jülich, Germany
- G.P. Ellrod
National Oceanographic and Atmospheric Administration/National Environmental Satellite, Data, and Information Service, Granby, CT, USA
- W.J. Emery
University of Colorado, Boulder, CO, USA
- C.C. Epifanio
Texas A&M University, College Station, TX, USA
- M.G. Estes
Universities Space Research Association, Huntsville, AL, USA
- J. Fasullo
University of Colorado – Boulder, Boulder, CO, USA
- R.A. Feely
NOAA Pacific Marine Environmental Laboratory, Seattle, WA, USA
- F. Fetterer
University of Colorado, Boulder, CO, USA
- J.J. Finnigan
CSIRO Atmospheric Research, Black Mountain, ACT, Australia
- H. Fischer
Institute of Meteorology and Climate Research (IMK), Karlsruhe Institute of Technology, Karlsruhe, Germany
- J.M. Forbes
University of Colorado, Boulder, CO, USA
- R. Forbes
European Centre for Medium-Range Weather Forecasts, Reading, UK
- D. Fowler
Edinburgh Research Station, Midlothian, UK

[†] Deceased.

J.L. Fox
Wright State University, Dayton, OH, USA

J.E. Frederick
The University of Chicago, Chicago, IL, USA

R.R. Friedl
California Institute of Technology, Pasadena, CA, USA

D.C. Fritts
GATS Inc., Boulder, CO, USA

Q. Fu
University of Washington, Seattle, WA, USA

J. Fugal
Max Planck Institute of Chemistry, Mainz, Germany

B. Funke
Instituto de Astrofísica de Andalucía, CSIC, Granada, Spain

E.J. Gaidos
University of Hawaii at Manoa, Honolulu, HI, USA

C.S. Gardner
University of Illinois at Urbana-Champaign, Urbana, IL, USA

J.-F. Gayet
Université Blaise Pascal, Clermont Ferrand, France

S. Ge
University of Colorado, Boulder, CO, USA

G.L. Geernaert
US Department of Energy, Washington, DC, USA

M. Ghil
Ecole Normale Supérieure, Paris, France; and University of California, Los Angeles, CA, USA

P.J. Gierasch
Cornell University, Ithaca, NY, USA

K.E. Gilbert
University of Mississippi, University, MS, USA

R. Gilbert
Queen's University, Kingston, ON, Canada

J.H. Golden
Forecast Systems Laboratory, NOAA, Boulder, CO, USA

B.W. Golding
Met Office, Exeter, UK

R.M. Goody
Harvard University (Emeritus), Cambridge, MA, USA

S.M. Gorelick
Stanford University, Stanford, CA, USA

L.J. Gray
University of Oxford, Oxford, UK

R. Grotjahn
University of California, Davis, CA, USA

A. Guenther
Pacific Northwest National Laboratory, Richland, WA, USA

D.S. Gutzler
University of New Mexico, Albuquerque, NM, USA

R.M. Haberle
NASA/Ames Research Center, Moffett Field, Mountain View, CA, USA

M.E. Hagan
National Center for Atmospheric Research, Boulder, CO, USA

C.J. Hahn
University of Arizona, Tucson, AZ, USA

J.D. Haigh
Blackett Laboratory, Imperial College London, London, UK

G.J. Hakim
University of Washington, Seattle, WA, USA

K. Hamilton
University of Hawaii, Honolulu, HI, USA

T.F. Hanisco
Harvard University, Cambridge, MA, USA

J.W. Hannigan
National Center for Atmospheric Research, Boulder, CO, USA

R.M. Hardesty
NOAA Environmental Technology Laboratory, Boulder, CO, USA

J.E. Harries
Imperial College of Science, Technology and Medicine, Blackett Laboratory, London, UK

N.R.P. Harris
University of Cambridge, Cambridge, UK

R.G. Harrison
The University of Reading, Reading, UK

D.L. Hartmann
University of Washington, Seattle, WA, USA

F. Hase
Institute of Meteorology and Climate Research (IMK), Karlsruhe Institute of Technology, Karlsruhe, Germany

- L. Hasse
Universität Kiel, Kiel, Germany
- S. Hastenrath
University of Wisconsin, Madison, WI, USA
- P. Haynes
University of Cambridge, Cambridge, UK
- J. Haywood
Met Office, Exeter, UK
- J.M. Hemsley
National Data Buoy Center, Stennis Space Center, MS, USA
- P. Herckes
Arizona State University, Tempe, AZ, USA
- P. Hess
National Center for Atmospheric Research, Boulder, CO, USA
- D.J. Hofmann[†]
NOAA Climate Monitoring and Diagnostics Laboratory, Boulder, CO, USA
- J.R. Holton
University of Washington, Seattle, WA, USA
- A.A.M. Holtslag
Wageningen University, Wageningen, The Netherlands
- D. Hooper
Science & Technology Facilities Council (STFC), Didcot, UK
- D.D. Houghton
University of Wisconsin-Madison, Madison, WI, USA
- S. Houweling
SRON Netherlands Institute for Space Research, Utrecht, The Netherlands
- X.-M. Hu
University of Oklahoma, Norman, OK, USA
- R.X. Huang
Woods Hole Oceanographic Institution, Woods Hole, MA, USA
- X. Huang
University of Michigan, Ann Arbor, MI, USA
- Y.-H. Huang
National Taiwan University, Taipei, Taiwan
- C. Hughes
University of York, York, UK
- B.A. Hungate
Northern Arizona University, Flagstaff, AZ, USA
- J.C.R. Hunt
University College London, London, UK
- L.A. Hunt
Science Systems and Applications Incorporated, Hampton, VA, USA
- J.W. Hurrell
National Center for Atmospheric Research, Boulder, CO, USA
- S. Ismail
Science Directorate, NASA Langley Research Center, Hampton, VA, USA
- C. Jakob
Monash University, VIC, Australia
- S.G. Jennings
National University of Ireland, Galway, Ireland
- P. Jenniskens
SETI Institute, Moffett Field, CA, USA
- X. Jiang
University of California, Los Angeles, CA, USA
- T.D. Jickells
University of East Anglia, Norwich, UK
- R.H. Johnson
Colorado State University, Fort Collins, CO, USA
- A. Joly
Centre National de Recherches Météorologiques – Groupe d'étude de l'Atmosphère Météorologique, Météo-France and CNRS, Toulouse, France
- P.D. Jones
Climatic Research Unit, University of East Anglia, Norwich, UK
- M. Jukes
University of Oxford, Oxford, UK
- R.A. Kahn
NASA Goddard Space Flight Center, Greenbelt, MD, USA
- M. Kalberer
University of Cambridge, Cambridge, UK
- L. Kantha
University of Colorado, Boulder, CO, USA
- M.C. Kelley
Cornell University, Ithaca, NY, USA

[†] Deceased.

D. Keyser
*University at Albany, State University of New York,
Albany, NY, USA*

S.Q. Kidder
*Colorado State University, Fort Collins,
CO, USA*

K.-Y. Kim
Seoul National University, Seoul, Korea

M.D. King
University of Colorado, Boulder, CO, USA

C. Knight
*National Center for Atmospheric Research, Boulder,
CO, USA*

N. Knight
*National Center for Atmospheric Research, Boulder,
CO, USA*

K.H. Knopfmeier
*University of Oklahoma; and National Oceanic and
Atmospheric Administration, Norman, OK, USA*

J.A. Knox
University of Georgia, Athens, GA, USA

T.R. Knutson
*NOAA Geophysical Fluid Dynamics Laboratory,
Princeton, NJ, USA*

G.W. Koch
Northern Arizona University, Flagstaff, AZ, USA

Y. Kondo
The University of Tokyo, Tokyo, Japan

A. Korolev
*Meteorological Service of Canada, Toronto, ON,
Canada*

A. Lacis
*Goddard Institute for Space Studies, New York,
NY, USA*

A. Laing
*National Center for Atmospheric Research, Boulder,
CO, USA*

D. Lal
*Scripps Institution of Oceanography, La Jolla,
CA, USA*

J.F. Lamarque
*National Center for Atmospheric Research, Boulder,
CO, USA*

D. Lamb
*The Pennsylvania State University, University Park,
PA, USA*

T.P. Lane
*The University of Melbourne, Melbourne, VIC,
Australia*

A.L. Lang
*University of Albany – State University of New York,
Albany, NY, USA*

M.F. Larsen
Clemson University, Clemson, SC, USA

K.-M. Lau
*NASA/Goddard Space Flight Center, Greenbelt, MD,
USA*

M.A. LeMone
*National Center for Atmospheric Research, Boulder,
CO, USA*

S.S. Leroy
*Harvard School of Engineering and Applied Sciences,
Cambridge, MA, USA*

P.F. Lester
San Jose State University, San Jose, CA, USA

J.P. Li
*Institute of Atmospheric Physics, Chinese Academy of
Sciences, Beijing, China*

M. Lippmann
New York University, Tuxedo, NY, USA

B. Liu
North Carolina State University, Raleigh, NC, USA

G. Liu
Florida State University, Tallahassee, FL, USA

W.T. Liu
*California Institute of Technology, Pasadena,
CA, USA*

N.G. Loeb
NASA Langley Research Center, Hampton, VA, USA

U. Lohmann
ETH Zurich, Zürich, Switzerland

M. López-Puertas
*Instituto de Astrofísica de Andalucía, CSIC, Granada,
Spain*

A.C. Lorenc
The Met Office, Bracknell, Berkshire, UK

J. Lu
*Pacific Northwest National Laboratory, Richland, WA,
USA*

J.C. Luvall
*National Space Science and Technology Center,
Marshall Space Flight Center, Huntsville, AL, USA*

- W.A. Lyons
FMA Research Inc., Fort Collins, CO, USA
- R.A. Madden
National Center for Atmospheric Research, Boulder, CO, USA
- S. Madronich
National Center for Atmospheric Research, Boulder, CO, USA
- L. Mahrt
Oregon State University, Corvallis, OR, USA
- A.J. Majda
New York University, New York, NY, USA
- D. Mansfield
National Meteorological Center, Bracknell, UK
- F.D. Marks
Hurricane Research Division, Miami, FL, USA
- D.G. Martinson
Columbia University, Palisades, NY, USA
- P.J. Mason
Met Office, Bracknell, UK
- T. Matsui
NASA/Goddard Space Flight Center, Greenbelt, MD, USA; and University of Maryland, College Park, MD, USA
- A. McCulloch
University of Bristol, Bristol, UK
- M.E. McIntyre
University of Cambridge, Cambridge, UK
- R. McKenzie
National Institute for Water and Atmospheric Research, Lauder, Central Otago, New Zealand
- P.H. McMurtry
University of Minnesota, Minneapolis, MN, USA
- R.D. McPeters
NASA Goddard Space Flight Center, Greenbelt, MD, USA
- C.R. Mechoso
University of California, Los Angeles, CA, USA
- Z. Meng
Peking University, Beijing, China
- P.M. Midgley
M & D Consulting, Leinfelden Metternich, Germany
- M. Miller
European Centre for Medium-Range Weather Forecasts, Reading, UK
- P. Minnis
Science Directorate, NASA Langley Research Center, Hampton, VA, USA
- M. Mishchenko
Goddard Institute for Space Studies, New York, NY, USA
- N.J. Mitchell
The University of Bath, Bath, UK
- M.G. Mlynarczyk
NASA Langley Research Center, Hampton, VA, USA
- C.-H. Moeng
National Center for Atmospheric Research, Boulder, CO, USA
- F. Molteni
Abdus Salam International Centre for Theoretical Physics, Trieste, Italy
- M. Moncrieff
National Center for Atmospheric Research, Boulder, CO, USA
- D.W. Moore
Pacific Marine Environmental Laboratory, Seattle, WA, USA
- G.K. Moortgat
Max-Planck-Institute for Chemistry, Mainz, Germany
- R. Müller
Institute for Energy and Climate Research (IEK-7), Forschungszentrum Jülich, Jülich, Germany
- J. Munthe
IVL Swedish Environmental Research Institute, Göteborg, Sweden
- R.S. Nerem
University of Colorado, Boulder, CO, USA
- P.A. Newman
NASA Goddard, Space Flight Center, Greenbelt, MD, USA
- H. Nguyen
Centre for Australian Weather and Climate Research, Bureau of Meteorology, Melbourne, VIC, Australia
- N. Nicholls
Bureau of Meteorology Research Centre, Melbourne, VIC, Australia
- F.T.M. Nieuwstadt
Delft University of Technology, Delft, The Netherlands
- S. Nigam
University of Maryland, College Park, MD, USA

L. Nkemdirim
University of Calgary, Calgary, AB, Canada

G.R. North
Texas A&M University, College Station, TX, USA

D.R. Novak
Weather Prediction Center, College Park, MD, USA

A. O'Neill
University of Reading, Earley Gate, Reading, UK

J. Oberheide
Clemson University, Clemson, SC, USA

T.E. Osterkamp
University of Alaska, Fairbanks, AK, USA

R.D. Palmer
University of Oklahoma, Oklahoma, OK, USA

T.R. Parish
University of Wyoming, Laramie, WY, USA

D.J. Parker
University of Leeds, Leeds, UK

M.D. Parker
North Carolina State University, Raleigh, NC, USA

G.K. Parks
University of Washington, Seattle, WA, USA

W.R. Pendleton
Utah State University, Logan, UT, USA

J. Perlwitz
University of Colorado, Boulder, CO, USA

R.L. Pfeffer
Florida State University, Tallahassee, FL, USA

R.A. Pielke, Sr.
University of Colorado at Boulder, CO, USA

R. Pinnick
US Army Research Laboratory, Adelphi, MD, USA

J.M.C. Plane
University of Leeds, Leeds, UK

C.M.R. Platt
Colorado State University, Fort Collins, CO, USA

R. Plougonven
Ecole Polytechnique, Palaiseau, France

M.K. Politovich
National Center for Atmospheric Research, Boulder, CO, USA

L.M. Polvani
Columbia University, New York, NY, USA

J.-P. Pommereau
LATMOS, CNRS, Guyancourt, France

J.A. Pyle
University of Cambridge, Cambridge, UK

D.A. Quattrochi
National Space Science and Technology Center, Marshall Space Flight Center, Huntsville, AL, USA

S. Quiring
Texas A&M University, College Station, TX, USA

M. Ramamurthy
University Corporation for Atmospheric Research, Boulder, CO, USA

A.L. Rangno (Retiree)
University of Washington, Seattle, WA, USA

R.M. Rauber
University of Illinois at Urbana-Champaign, Urbana, IL, USA

A.R. Ravishankara
Colorado State University, Fort Collins, CO, USA

I.A. Renfrew
University of East Anglia, Norwich, UK

N.O. Renno
University of Michigan, Ann Arbor, MI, USA

R. Reynolds
University of Reading, Reading, UK

P.B. Rhines
University of Washington, Seattle, WA, USA

A.D. Richmond
National Center for Atmospheric Research, Boulder, CO, USA

D.L. Rickman
National Space Science and Technology Center, Marshall Space Flight Center, Huntsville, AL, USA

R.G. Roble
National Center for Atmospheric Research, Boulder, CO, USA

A. Robock
Rutgers University, New Brunswick, NJ, USA

F. Rohrer
Forschungszentrum Jülich, Jülich, Germany

K.H. Rosenlof
Earth System Research Laboratory, National Oceanic and Atmospheric Administration, Boulder, CO, USA

- M.W. Rotach
University of Innsbruck, Innsbruck, Austria
- R. Rotunno
National Center for Atmospheric Research, Boulder, CO, USA
- I. Roulstone
University of Surrey, Guildford, UK
- J. Rudolph
York University, Toronto, ON, Canada
- L.M. Russell
Scripps Institution of Oceanography, University of California, San Diego, La Jolla, CA, USA
- C.L. Sabine
NOAA Pacific Marine Environmental Laboratory, Seattle, WA, USA
- D.A. Salstein
Atmospheric and Environmental Research, Inc., Lexington, MA, USA
- C. Schär
Atmospheric and Climatic Science ETH, Zürich, Switzerland
- U. Schmidt
Johann Wolfgang Goethe-University, Frankfurt am Main, Institut für Meteorologie und Geophysik, Frankfurt am Main, Germany
- M.R. Schoeberl
Science and Technology Corporation, Lanham, MD, USA
- D.M. (David) Schultz
University of Manchester, Manchester, UK
- J. Schulz
Meteorological Institute, University of Bonn, Bonn, Germany
- J.H. Seinfeld
California Institute of Technology, Pasadena, CA, USA
- C.J. Senff
CIRES, University of Colorado; and NOAA Earth System Research Laboratory, Boulder, CO, USA
- M.C. Serreze
University of Colorado, Boulder, CO, USA
- G.E. Shaw
Geophysical Institute, University of Alaska, Fairbanks, AK, USA
- G.G. Shepherd
York University, Toronto, ON, Canada
- T.G. Shepherd
University of Toronto, Toronto, ON, Canada
- R. Sica
The University of Western Ontario, London, ON, Canada
- H.B. Singh
NASA Ames Research Center, Mountain View, CA, USA
- J.A. Sippel
National Aeronautics and Space Administration (NASA), Greenbelt, MD, USA
- J. Slingo
University of Reading, Reading, UK
- H.G.J. Smit
Research Centre Jülich, Jülich, Germany
- A.K. Smith
National Center for Atmospheric Research, Boulder, CO, USA
- J.B. Smith
Harvard University, Cambridge, MA, USA
- L.A. Smith
London School of Economics, Centre for the Analysis of Time Series, London, UK
- R.B. Smith
Yale University, New Haven, CT, USA
- I.N. Sokolik
Georgia Institute of Technology, Atlanta, GA, USA
- S.C. Solomon
National Center for Atmospheric Research, Boulder, CO, USA
- J. Sommar
Göteborg University, Göteborg, Sweden
- P.J. Sousounis
AIR Worldwide, Boston, MA, USA
- K. Stamnes
Stevens Institute of Technology, Hoboken, NJ, USA
- S.N. Stechmann
University of Wisconsin–Madison, Madison, WI, USA
- E.J. Steig
University of Washington, Seattle, WA, USA
- O. Stein
IEK 8: Troposphere, Research Center Juelich, Juelich, Germany

D.J. Stensrud
*National Oceanic and Atmospheric Administration,
Norman, OK, USA*

R.S. Stolarski
Johns Hopkins University, Baltimore, MD, USA

S.E. Strahan
*Universities Space Research Association, NASA Goddard
Space Flight Center, Greenbelt, MD, USA*

M. Sturm
*US Army Cold Regions Research & Engineering
Laboratory-Alaska, Fort Wainwright, AL, USA*

S.T. Suess
*NASA Marshall Space Flight Center, Huntsville, AL,
USA*

P.P. Sullivan
*National Center for Atmospheric Research, Boulder,
CO, USA*

M.P. Sulzer
Arecibo Observatory, Arecibo, PR, USA

E.S. Takle
Iowa State University, Ames, IA, USA

W.-K. Tao
*NASA/Goddard Space Flight Center, Greenbelt, MD,
USA*

M.J. Taylor
Utah State University, Logan, UT, USA

P.K. Taylor
*Southampton Oceanography Centre, Southampton,
UK*

V.P. Tchakerian
Texas A&M University, College Station, TX, USA

G.E. Thomas
University of Colorado, Boulder, CO, USA

L.W. Thomason
NASA Langley Research Center, Hampton, VA, USA

D.W.J. Thompson
Colorado State University, Fort Collins, CO, USA

D.J. Thomson
Met Office, Bracknell, UK

J. Thuburn
University of Exeter, Exeter, UK

R.A. Tomas
University of Colorado – Boulder, Boulder, CO, USA

D. Toohey
University of Colorado Boulder, Boulder, CO, USA

L. Travis
*Goddard Institute for Space Studies, New York,
NY, USA*

K.E. Trenberth
*National Center for Atmospheric Research, Boulder,
CO, USA*

B.T. Tsurutani
Jet Propulsion Laboratory, Pasadena, CA, USA

J. Turner
British Antarctic Survey, Cambridge, UK

H. Turtiainen
Vaisala Company, Helsinki, Finland

C. Twohy
Oregon State University, Corvallis, OR, USA

T. Uttal
NOAA, Boulder, CO, USA

R.H. Varney
Cornell University, Ithaca, NY, USA

G. Vaughan
University of Manchester, Manchester, UK

W.W. Vaughan
*University of Alabama in Huntsville, Huntsville,
AL, USA*

G.A. Vecchi
GFDL/NOAA, Princeton, NJ, USA

A. Venkatram
*University of California – Riverside, Riverside,
CA, USA*

J.-P. Vernier
*Science Systems and Applications, Inc., Hampton,
VA, USA*

R.K. Vincent
*Bowling Green State University, Bowling Green,
OH, USA*

R. von Glasow
University of East Anglia, Norwich, UK

A. Wahner
Forschungszentrum Jülich, Jülich, Germany

R.M. Wakimoto
*National Center for Atmospheric Research, Boulder,
CO, USA*

D.E. Waliser
California Institute of Technology, Pasadena, CA, USA

J.M. Wallace
University of Washington, Seattle, WA, USA

- B. Wang
University of Hawaii, Honolulu, HI, USA
- Y. Wang
University of Hawaii at Manoa, Honolulu, HI, USA
- Z. Wang
University of Illinois at Urbana-Champaign, Urbana, IL, USA
- B. Ward
Public Works and Natural Resources, Longmont, CO, USA
- S. Warren
University of Washington, Seattle, WA, USA
- J. Waters
California Institute of Technology, Jet Propulsion Laboratory, Pasadena, CA, USA
- R.P. Wayne
University of Oxford, Oxford, UK
- S. Weart
Center for History of Physics, American Institute of Physics, College Park, MD, USA
- P.J. Webster
Georgia Institute of Technology, Atlanta, GA, USA
- P.J. Webster
University of Colorado – Boulder, Boulder, CO, USA
- W.F. Weeks
University of Alaska Fairbanks, Fairbanks, AK, USA
- M.L. Weisman
National Center for Atmospheric Research, Boulder, CO, USA
- S.J. Weiss
National Oceanic and Atmospheric Administration, Norman, OK, USA
- N.C. Wells
University of Southampton, Southampton, UK
- H. Wernli
Institute for Atmospheric and Climate Science, Swiss Federal Institute of Technology (ETH), Zürich, Switzerland
- T. Wetter
Johann Wolfgang Goethe-University, Frankfurt am Main, Institut für Meteorologie und Geophysik, Frankfurt am Main, Germany
- M.C. Wheeler
Centre for Australian Weather and Climate Research, Bureau of Meteorology, Melbourne, VIC, Australia
- A.A. White
University of Surrey, Guildford, UK
- D.N. Whiteman
NASA Goddard Space Flight Center, Greenbelt, MD, USA
- B.A. Wielicki
NASA Langley Research Center, Hampton, VA, USA
- E.R. Williams
Massachusetts Institute of Technology, Cambridge, MA, USA
- M.L. Wong
California Institute of Technology, Pasadena, CA, USA
- N. Wood
Met Office, Exeter, UK
- R. Wood
University of Washington, Seattle, WA, USA
- S.N. Wren
University of Toronto, Toronto, ON, Canada
- C.-C. Wu
National Taiwan University, Taipei, Taiwan
- D.J. Wuebbles
University of Illinois, Urbana, IL, USA
- L. Xie
North Carolina State University, Raleigh, NC, USA
- P. Yang
Texas A&M University, College Station, TX, USA
- S. Yang
NOAA/NWS/NCEP, Climate Prediction Center, Camp Springs, MD, USA
- J.A. Young
University of Wisconsin, Madison, WI, USA
- Z. Yu
College of Hydrology and Water Resources, Hohai University, Nanjing, China; and University of Nevada Las Vegas, Las Vegas, NV, USA
- Y.L. Yung
California Institute of Technology, Pasadena, CA, USA

S.E. Yuter
*North Carolina State University, Raleigh,
NC, USA*

S. Yvon-Lewis
*Texas A&M University, College Station,
TX, USA*

D. Zardi
University of Trento, Trento, Italy

S.E. Zebiak
*International Research Institute for Climate Prediction,
Palisades, NY, USA*

M.D. Zelinka
*Lawrence Livermore National Laboratory, Livermore,
CA, USA*

C. Zhang
University of Miami, Miami, FL, USA

F. Zhang
*Pennsylvania State University, University Park,
PA, USA*

M. Zhang
*Stony Brook University, Stony Brook,
NY, USA*

PREFACE TO THE FIRST EDITION

A half century ago the American Meteorological Society published the *Compendium of Meteorology*, which in a single volume of 1334 pages summarized the state of understanding of the atmosphere at that time. A perusal of the contents of that volume indicates that although a broad range of topics was covered, the vast bulk of the volume was devoted to traditional meteorological topics such as atmospheric dynamics, cloud physics, and weather forecasting. Barely 4 percent of the volume was devoted to articles related to atmospheric chemistry or air pollution and, of course, none of the volume was devoted to techniques such as satellites and remote sensing.

As Sir John Mason aptly notes in his foreword to the present work, the atmospheric sciences have expanded in scope enormously over the past 50 years. Topics such as atmospheric chemistry and global climate change, of only marginal interest 50 years ago, are now central disciplines within the atmospheric sciences. Increasingly, developing areas within the atmospheric sciences require students, teachers, and researchers to familiarize themselves with areas far outside their own specialties. This work is intended to satisfy the need for a convenient and accessible references source covering all aspects of atmospheric sciences. It is written at a level that allows undergraduate science and engineering students to understand the material, while providing active researchers with the latest information in the field.

More than 400 scientists, from academia, government, and industry have contributed to the 330 articles in this work. We are very grateful to these authors for their success in providing concise and authoritative summaries of complex subjects. As editors, we have benefited from the chance to learn from these articles, and we believe that all students and active scientists who want to increase their knowledge of the atmosphere will benefit enormously from access to this work.

We are also grateful to the 31 members of the Editorial Advisory Board who have guided us in our coverage of the very broad range of topics represented in this encyclopedia. Their willingness to suggest topics and authors, and to carefully review draft articles has contributed significantly to our success.

The production of this multivolume encyclopedia would not have been possible without the dedicated work of the staff of the Major Reference Works group at Academic Press. We are especially grateful to the Major Reference Work Development Manager, Colin McNeil, who has worked closely with us during the entire process. Finally, we appreciate the liberal use of color figures in the printed encyclopedia.

James R Holton, Judith A Curry, and John Pyle

This page intentionally left blank

PREFACE TO THE SECOND EDITION

Since the publication of the first edition of the *Encyclopedia of Atmospheric Sciences*, significant advances in research have been achieved all across the broad and expanding spectrum of the field and related disciplines. In particular, climate science with primary input from the atmospheric research emerges as a new field and integrator of interlocking peripheral disciplines over the last decade. These events have demanded the solicitation of new and updated articles for the 2003 edition. Some articles from the earlier publication were judged to be of such a fundamental and enduring nature that they did not require modification. But huge amounts of new information from Earth-orbiting satellite observatories have brought much new insight to the field. In addition there are new findings in many areas such as the latest simulations of meteorological and climatic processes of interest as well as simulations and observations of the composition and interaction of the field's chemical constituents. While interest in the ozone hole and its ramifications may have reached a plateau, ever more understanding of the stratosphere and its role in climate change emerges. The study of past climates provides new means of testing climate models and theories. In weather prediction we see new progress on how data are to be better assimilated for much improved initialization of the forecast model leading to the promise of more accurate predictions of severe weather and tropical cyclones over longer lead times. These are just a few of the new features of the second edition.

The editors of the second edition are greatly indebted to our predecessors in the first edition. They set the outline of topics and solicited the original authors, while establishing a high standard for the content of this publication. In many cases we decided to reprint those articles or request only minor updates. Nevertheless, many articles in this edition are entirely original, based on which we also made significant reorganization of the content. We are proud of our product and hope it provides the same assistance to students, researchers, and practitioners throughout the science and engineering communities.

Editors of the second edition
Gerald R North
Fuqing Zhang
John Pyle

This page intentionally left blank

EDITOR BIOGRAPHIES



Gerald R North received his PhD in theoretical physics from the University of Wisconsin in 1966. After postdoctoral research at the University of Pennsylvania he became a faculty member in physics at the University of Missouri—St. Louis. He shifted his research focus to climate science research during his sabbatical year at the National Center for Atmospheric Research, where he won the Outstanding Paper Award in 1975. He moved to NASA Goddard Space Flight Center (GSFC) in 1978 where he was awarded the NASA Medal for Research Excellence. During his stay at GSFC, he was the proposer and first study scientist for the Tropical Rainfall Measuring Mission, which was launched in 1997 and is still orbiting in 2014.

He moved to Texas A&M University in 1986 as a university distinguished professor of atmospheric sciences where he served as department head from 1995 to 2003. He has served as editor-in-chief of the *Reviews of Geophysics* and is recognized as one of the most cited authors in geosciences (Web of Science). He has chaired and/or served on a number of national committees and is a Fellow of the American Geophysical Union, American Meteorological Society (AMS) and the American Association for the Advancement of Science, and winner of the Jule Charney Award for Research (AMS).

He has published about 150 refereed papers not including many book chapters and reviews. His books include *Paleoclimatology*, co-authored with Thomas Crowley, and *An Introduction to Atmospheric Thermodynamics* co-authored with Tatiana Erikhimova.

North's interests are focused on the use of mathematical and statistical tools to solve climate problems over a wide range of issues including: analytical solutions of simplified energy balance climate models, use of random field techniques in representing and interpreting climate data and model simulations, detection of deterministic signals in climate change, statistical analysis satellite remote sensing for mission planning and analysis of data, paleoclimate problems using simplified climate models.



John Pyle obtained a BSc in Physics at Durham University before moving to Oxford where he completed a DPhil in Atmospheric Physics, helping to develop a numerical model for stratospheric ozone studies. After a short period at the Rutherford Appleton Laboratory he moved to a lectureship at Cambridge University in 1985. In 2000 he was appointed professor of atmospheric science and since 2007 has been the 1920 professor of physical chemistry. He is a Professorial Fellow at St Catharine's College. He has been a codirector of Natural Environment Research Council's National Centre for Atmospheric Science, where he is currently Chief Scientist.

His research focuses on the numerical modelling of atmospheric chemistry. Problems involving the interaction between chemistry and climate have been addressed; these range from stratospheric ozone depletion to the changing tropospheric oxidizing capacity and have included the environmental impact of aviation, land use change, biofuel technologies, and the hydrogen economy. He has studied palaeochemistry problems as well as the projected atmospheric composition changes during the current century. He has published more than 250 peer reviewed papers.

He played a major role in building an EU stratospheric research programme in the 1990s, coordinating several major field campaigns. He has contributed to all the WMO/UNEP assessments on stratospheric ozone since the early 1980s and is now one of the four international cochairmen

of the Scientific Assessment Panel, responsible for these assessments. He was a convening lead author in the IPCC Special report "Safeguarding the ozone layer and the global climate system," published in 2006.

He was elected Fellow of the Royal Society in 2004 and an American Geophysical Union Fellow in 2011. He was awarded the Cambridge ScD degree in 2012. Other honours and awards include membership of Academia Europaea (1993), Royal Society of Chemistry (Interdisciplinary award, 1991, and John Jeyes lectureship, 2008), and the Royal Meteorological Society Adrian Gill Prize, in 2004.



Fuqing Zhang is a professor with tenure in the Department of Meteorology at the Pennsylvania State University, with a joint appointment in the Department of Statistics, along with an endowed position as the E Willard & Ruby S Miller Faculty Fellow at the College of Earth and Mineral Sciences at the Pennsylvania State University. His research interests include atmospheric dynamics and predictability, data assimilation, ensemble forecasting, tropical cyclones, gravity waves, mountain plains and sea-breeze circulations, warm-season convection, and regional-scale climate. He earned his BS and MS in meteorology from Nanjing University, China in 1991 and 1994, respectively, and his PhD in atmospheric science in 2000 from North Carolina State University.

He spent seven years as an assistant and then associate professor at Texas A&M University before coming to Penn State University as a full professor in 2008. In 2000, he spent a year and a half as a postdoctoral fellow at the National Center for Atmospheric Research. He also held various visiting scholarship appointments at various academic and research institutions including the National

Center for Atmospheric Research in Boulder, Colorado; the Navy Research Laboratory in Monterey, California; NOAA/AOML Hurricane Research Division, Miami, Florida; Peking University and Nanjing University, China; the Chinese State Key Laboratory of Severe Weather in Beijing, China; and Laboratoire de Meteorologie Dynamique, École Normale Supérieure in Paris, France.

He has authored/co-authored about 130 peer reviewed journal publications and has given more than 160 keynote speeches or invited talks at various institutions and meetings. He has served as principal investigator/co-principal investigator for 30 federal or state-sponsored research grants. He has chaired/cochaired more than 10 scientific meetings or workshops. He also served on various review or advisory panels for numerous organizations that include National Science Foundation, Office of Naval Research, NASA, NOAA, and National Academies. He has also served as editor of several professional journals including *Monthly Weather Review*, *Science China*, *Atmospheric Science Letter*, *Acta Meteorologica Sinica*, and *Computing in Science & Engineering*.

He has also received numerous awards for his research and service. Notably, in 2007 he received the **Outstanding Publication Award** from the National Center for Atmospheric Research. In 2009, was the sole recipient of the American Meteorological Society's 2009 **Clarence Leroy Meisinger Award** "for outstanding contributions to mesoscale dynamics, predictability, and ensemble data assimilation." Most recently, he received the 2014 American Meteorological Society's **Banner Miller Award** "for valuable insights into incorporating real-time airborne Doppler radar measurements via ensemble data assimilation, leading to improvements in forecasts of tropical cyclone track and intensity."

GUIDE TO USING THE ENCYCLOPEDIA

Structure of the Encyclopedia

The material in the encyclopedia is not arranged by ordinary alphabetical order, but by alphabetical order according to 49 principal topic areas taken to allow all papers belonging to each principal topic to appear together in the same volume. Within each principal subject, article headings are also arranged alphabetically, except where logic dictates otherwise. For example, overview articles appear at the beginning of a section.

There are four features that help you find the topic in which you are interested:

- i. the contents list
- ii. cross-references to other relevant articles within each article
- iii. a full subject index
- iv. contributors

i. Contents List

The contents list, which appears at the front of each volume, lists the entries in the order that they appear in the encyclopedia. It includes both the volume number and the page number of each entry.

ii. Cross-references

All of the entries in the encyclopedia have been cross-referenced. The cross-references, which appear at the end of an article as a See also list, serve four different functions:

- i. To draw the reader's attention to related material in other entries

- ii. To indicate material that broadens and extends the scope of the article
- iii. To indicate material that covers a topic in more depth
- iv. To direct readers to other articles by the same author(s)

Example

The following list of cross-references appears at the end of the article.

See also: **Biogeochemical Cycles:** Biogeochemistry of Iodine. **Stratospheric Chemistry Topics:** HO_x; Halogen Sources, Natural (Methyl Bromide and Related Gases); Halogens; Hydrogen Budget; Overview; Reactive Nitrogen (NO_x and NO_y).

iii. Index

The index includes page numbers for quick reference to the information you are looking for. The index entries differentiate between references to a whole article, a part of an article, and a table or figure.

iv. Contributors

At the start of each volume there is list of the authors who contributed to that volume.

This page intentionally left blank

MIDDLE ATMOSPHERE

Contents

Planetary Waves

Polar Vortex

Quasi-Biennial Oscillation

Semiannual Oscillation

Stratospheric Sudden Warmings

Transport Circulation

Zonal Mean Climatology

Planetary Waves

AK Smith, National Center for Atmospheric Research, Boulder, CO, USA

J Perlwitz, University of Colorado, Boulder, CO, USA

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

Planetary waves are vertically propagating disturbances that are identified by large-scale variations in the pressure, temperature, winds, and composition. Much of the large-scale longitude variability in the middle atmosphere is due to planetary waves. They can be stationary or travel in longitude with periods of a few days to a few weeks. Planetary waves interact with other waves and with the background atmosphere. Large-amplitude waves can contribute to rapid changes in wind and temperature. Planetary waves also play a role in the impact of stratospheric changes on the troposphere.

Introduction

A planetary wave is a large-scale perturbation of the atmospheric circulation that extends coherently around a full longitude circle. The perturbations have wavelike forms in the longitudinal and vertical directions and often also in the latitudinal direction. These large-scale waves are a dominant part of the spatial and temporal variability in the stratosphere, and they also make contributions at higher altitudes in the mesosphere. Most important are quasistationary midlatitude Rossby waves, which propagate upward from the troposphere and are ubiquitous but quite variable in the middle atmosphere during winter. They are important because they have significant influence on the wind speed, temperature, distribution of ozone, and other characteristics of the middle atmosphere. Other planetary waves that are also important are global traveling modes, known as normal modes, and a class of waves confined to the equatorial region.

Much of the information about planetary waves in the middle atmosphere has come from measurements made by orbiting satellites. The type of measurement that has been most used is temperature profiles. The geopotential height profile

can be determined from the temperature profile using the hydrostatic relation if the geopotential height at some point along the profile (such as at the bottom) is known. Planetary wave information has also been determined from satellite observations of horizontal winds, especially in the mesosphere. Radiosonde, radar, and other ground-based observations, while not sufficient to determine global wave structure, provide important additional information about vertical structure. The data that have been used to construct the illustrations in this article were processed by the National Center for Atmospheric Research (<http://ncar.ucar.edu/>), Modern Era Retrospective Analysis for Research and Applications (MERRA; <http://gmao.gsfc.nasa.gov/research/merra/intro.php>), and Sounding of the Atmosphere Using Broadband Emission Radiometer (SABER; <http://saber.gats-inc.com/>).

Quasistationary Midlatitude Rossby Waves

Where they occur, planetary waves are evident in all dynamical fields: temperature, wind, pressure, and density. Observations of wave structure are commonly viewed using the temperature

or the geopotential height. Geopotential height is the height of a given pressure level. **Figure 1** shows the temperature and geopotential height for a pressure level in the middle stratosphere of the Northern Hemisphere on a particular day (10 hPa on 31 January 2007). The North Pole is at the center of each panel, and the equator is around the outer edge. The Greenwich Meridian is the horizontal line extending down from the pole. Colors indicate the temperature in Kelvin (upper) and the geopotential height in kilometers (lower) of the 10 hPa pressure surface. The variations are smooth compared with the much more abrupt variability seen near the Earth's surface on a typical weather map.

The planetary waves shown in **Figure 1** are the deviations from symmetry around a longitude circle. These can be decomposed further by performing Fourier analysis in longitude at a fixed latitude and pressure level. **Figure 2** illustrates the temperature and height variations and the results of Fourier analysis at 60° N. Both wave fields are dominated by zonal wave number 1, but wave number 2 also has significant amplitude. On a polar plot such as **Figure 1**, a wave number 1 pattern tends to be evident as a displacement of the lowest heights or temperatures off of the pole and sometimes, as in the case shown, as a distinct region of high height or temperature.

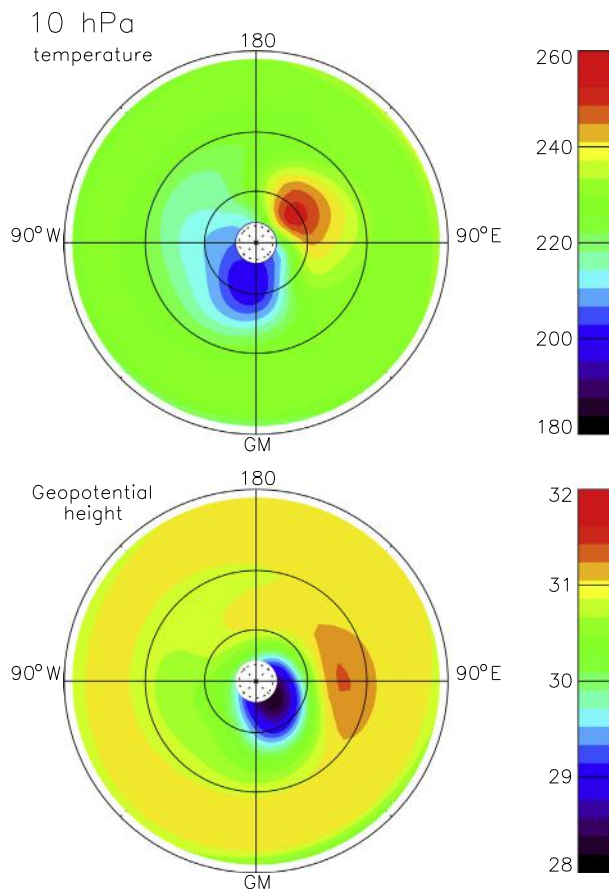


Figure 1 Polar stereographic projections (the North Pole is at the center, and the Equator is around the outer edge) of the temperature and geopotential height for the Northern Hemisphere at 10 hPa on 31 January 2007. Contours indicate the temperature in Kelvin or the height in kilometers. GM: Greenwich meridian.

Wave number 2 is evident as an elongation of the contours of the low-geopotential-height area.

For this day, pressure level, and latitude, the wave number 1 patterns in temperature and geopotential have similar phases. This is evident from the location of the maxima in **Figure 2** (both in the quadrant 90–180°) and also from the similar locations of the low temperature and low geopotential height in **Figure 1**. This is not always the case. **Figure 3** shows maps of the temperature and geopotential height for other pressure levels in the middle atmosphere. At 0.1 and 0.01 hPa, the longitude of low temperature is not near that of low geopotential height.

From geostrophic balance, air motion on a pressure surface tends to follow lines of constant geopotential height, with low values to the left in the Northern Hemisphere (counterclockwise flow around the low). One can see that following such a trajectory will move an air parcel alternately closer to the pole and farther from the pole. The time taken to complete a circuit of the pole depends on the latitude and wind speed, but it normally ranges from several days to longer than a week. The air parcel will also move vertically as it follows the circumpolar path, although the vertical excursions are much smaller than those in latitude, being typically on the order of 1 km.

Figure 4 shows the amplitude of wave numbers 1 and 2 for the case illustrated in **Figures 1** and **3**. Values at 10 hPa pressure correspond to the variations shown in **Figure 2**. Typical midlatitude Rossby waves vary in amplitude daily, but the longitude of the maximum geopotential height at a given pressure level often stays in the same quadrant. Several of the features evident in **Figure 4** are typical of wave amplitudes in the winter hemisphere. The geopotential height amplitudes are largest in middle to high latitudes near 60°; maximum amplitudes in the vertical are reached in the stratosphere, usually between 10 and 0.3 hPa, and wave number 1 has larger amplitude than wave number 2. The temperature amplitude has a minimum at the level where the geopotential height amplitude is largest. This case is atypical in that the wave number 1 amplitude is particularly large; more common values are about half of that seen on this particular day. Although planetary-scale waves have their largest amplitudes in the stratosphere, they do not disappear in the mesosphere (above ~0.3 hPa). Both temperature and geopotential height have measurable amplitudes up to the highest level shown (0.01 hPa, which is near 80 km altitude).

Figure 5 shows the variation of wave number 1 and 2 amplitudes during two Northern Hemisphere winters: 2006–07 and 2008–09. Wave amplitudes vary on a time scale of a few days. There are also large differences between different winters. Wave number 1 normally has higher amplitude than wave number 2, but exceptions have occurred, for example in January 2009. The amplitudes are not symmetric around the winter solstice. On average, the amplitude of wave number 1 is larger during January–February. The case in early 2009 is an exception; wave number 1's amplitude was small during January through March of that year.

Some of the Rossby waves generated in the troposphere by thermal contrasts between land and ocean surfaces and by flow over mountains propagate vertically and are the source of the midlatitude waves in the middle atmosphere. A substantial part of the variability seen in the wave amplitude (**Figure 5**) is

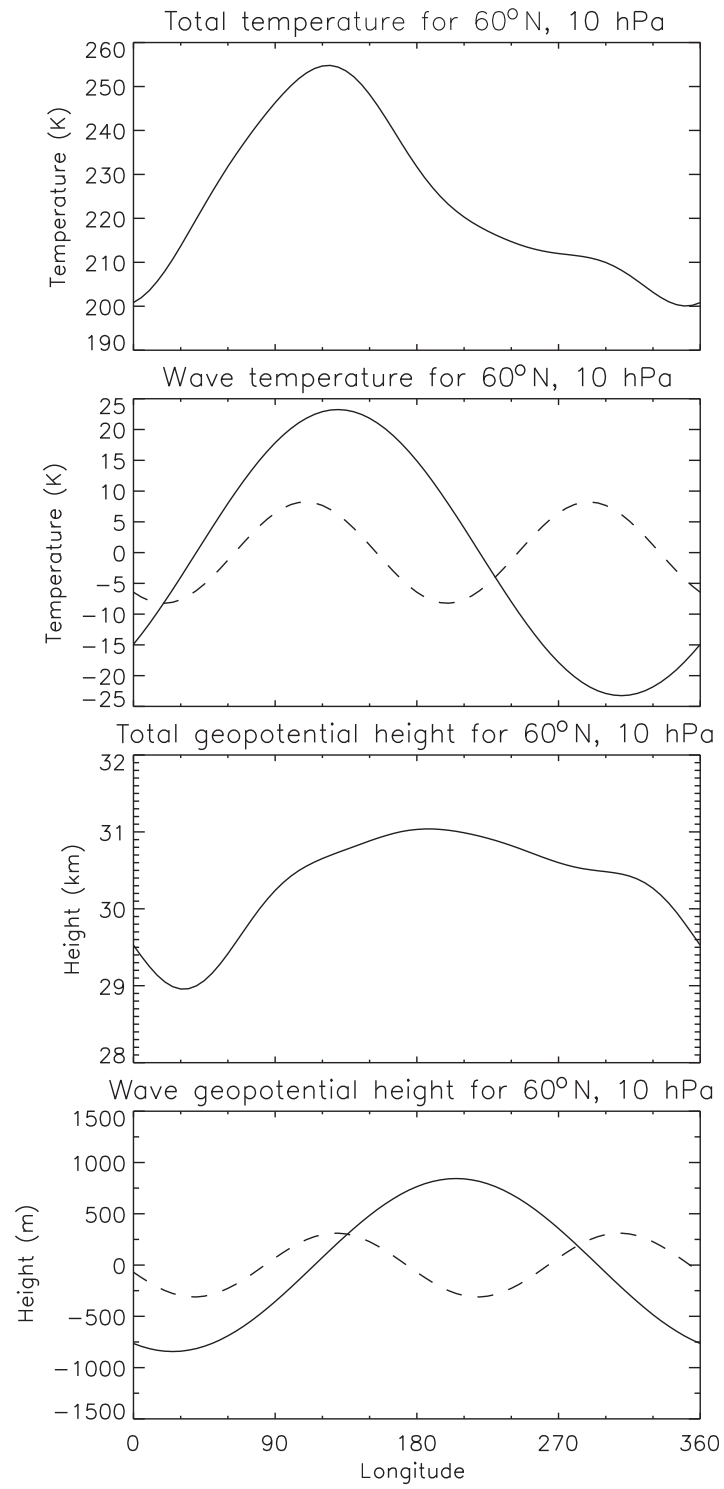


Figure 2 Variation of temperature and geopotential height with longitude and Fourier analysis showing the contributions of wave numbers 1 (solid) and 2 (dashed) at 60° N for the case shown in [Figure 1](#).

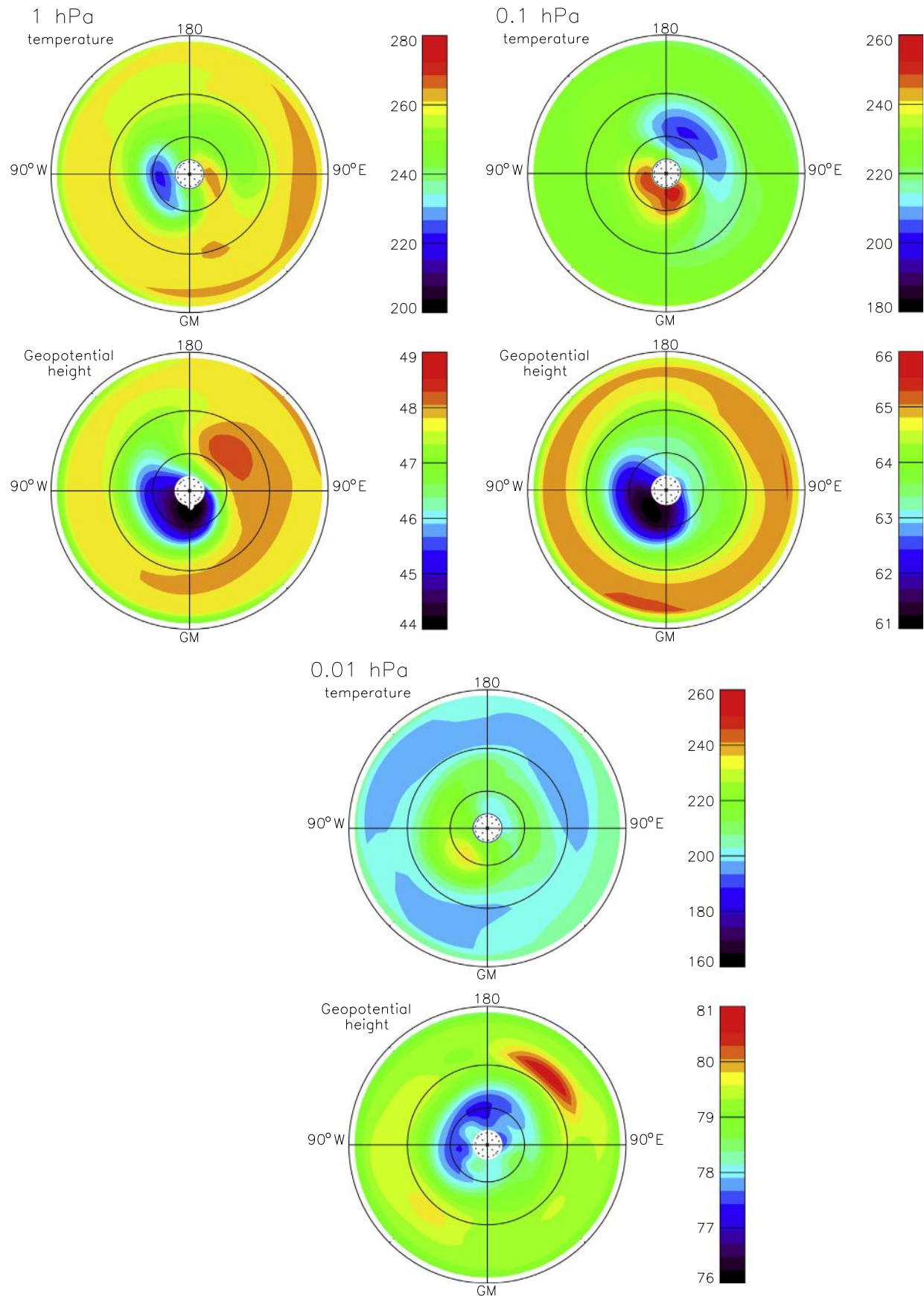


Figure 3 Polar stereographic projections of the temperature and geopotential height for the Northern Hemisphere at 1, 0.1, and 0.01 hPa on 31 January 2007.

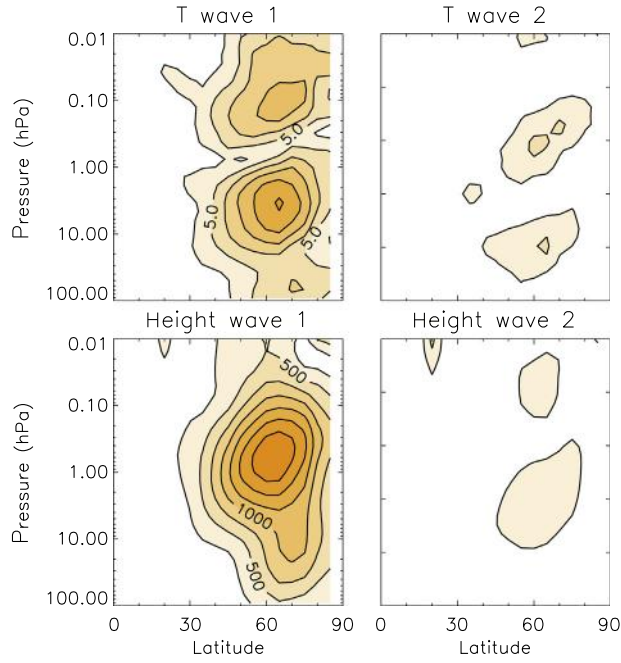


Figure 4 The amplitude of wave numbers 1 and 2 of temperature and geopotential height as a function of latitude and pressure for 31 January 2007.

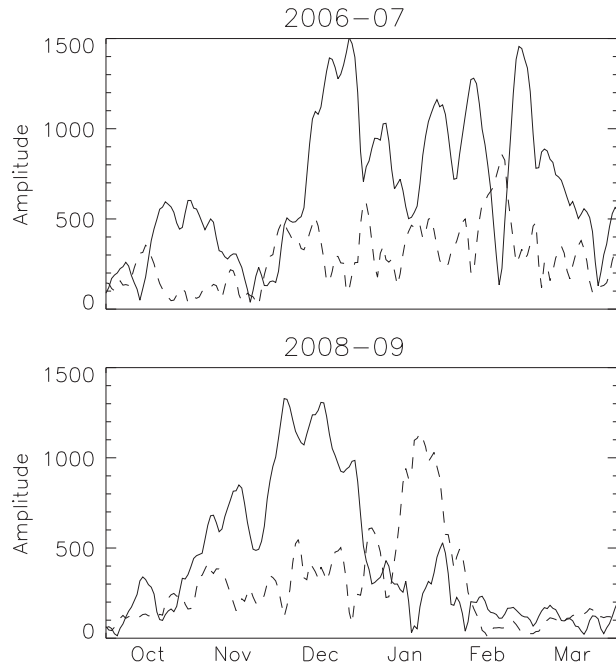


Figure 5 Variation of wave 1 and 2 geopotential height amplitude at 10 hPa with time, for the Northern Hemisphere winters of 2006–07 and 2008–09. Solid lines are wave number 1 amplitude, and dashed lines are wave number 2 amplitude.

a result of the variability of the large-scale weather patterns in the troposphere. With a few simplifications, it is straightforward to derive an equation that predicts which waves will propagate into the middle atmosphere. First, we split all fields into a zonal mean (longitudinal average) and perturbation

components and assume conditions of linearity, where the perturbation fields are significantly smaller than the zonal mean fields. The wave equation is derived from the perturbation form of the quasigeostrophic potential vorticity equation on a β -plane. In the β -plane approximation, the variation of Coriolis torque with latitude is assumed to be constant: $f(y) = f(y_0) + \beta(y - y_0)$, where f is the Coriolis torque and y is latitude. The quasigeostrophic potential vorticity equation is then given by eqn [1], where q , the quasigeostrophic potential vorticity, is defined in eqn [2].

$$\left(\frac{\partial}{\partial t} + \bar{u} \frac{\partial}{\partial x} \right) q' + v' \frac{\partial \bar{q}}{\partial y} = D \quad [1]$$

$$q = f + \frac{1}{f} \frac{\partial^2 \phi}{\partial x^2} + \frac{\partial}{\partial y} \frac{1}{f} \frac{\partial \phi}{\partial y} + \frac{f}{N^2} \frac{\partial}{\partial z} \rho \frac{\partial \phi}{\partial z} \quad [2]$$

The overbars in eqn [1] indicate zonal averages, and the primes indicate deviations from the zonal average. The planetary vorticity f has no perturbation component and is not included in q' . The zonal derivative $\left(\frac{\partial^2 \phi}{\partial x^2} \right)$ disappears for $\bar{q}$. $\bar{u}$ is zonal wind; v' is perturbation meridional wind; D represents damping due to diabatic processes, interaction with small-scale waves, and wave breaking; ϕ is the geopotential; N is the buoyancy frequency; ρ is the density; x is longitude; and z is altitude.

To solve eqn [1], we assume a waveform in the zonal dimension and in time, with wave number k and frequency ω or, equivalently, phase speed c , where $c = \omega/k$ (eqn [3]).

$$\phi' = \psi_k(y, z) \exp(ikx - i\omega t + z/2H) \quad [3]$$

Note that there is also an altitude factor $\exp(z/2H)$. This altitude factor takes into account the fact that a propagating, nondissipating wave will conserve energy, which is proportional to $|\phi'|^2$. With this conservation, the amplitude will increase with altitude because of the decrease in density (proportional to $e^{-z/H}$).

Substitution of eqns [2] and [3] into eqn [1] gives eqn [4], where Q is given by eqn [5].

$$\frac{\partial}{\partial y} \left(\frac{\partial \psi_k}{\partial y} \right) + \frac{f^2}{N^2} \frac{\partial^2 \psi_k}{\partial z^2} + Q \psi_k = D \quad [4]$$

where

$$Q = \frac{\bar{q}_y}{(\bar{u} - c)} - k^2 - \frac{f^2}{4H^2 N^2} \quad [5]$$

and

$$\bar{q}_y = \frac{\partial f}{\partial y} - \frac{\partial^2 \bar{u}}{\partial y^2} - \frac{f^2}{N^2} \frac{1}{\rho} \frac{\partial}{\partial z} \rho \frac{\partial \bar{u}}{\partial z} \quad [6]$$

H is the atmospheric scale height and N is the buoyancy frequency. Equation [4] is a two-dimensional wave equation when Q is positive (i.e., the solutions are sinusoidal in latitude and altitude). Note that the vertical and meridional scales are vastly different ($N \gg fN \gg f$; the ratio f^2/N^2 is typically 10^{-4}).

The restoring force responsible for the existence of planetary waves is provided by the zonal mean distribution of atmospheric angular momentum (or vorticity), which is determined

by the magnitude of Q . For a given zonal wave number k , the variations in Q are dominated by the meridional gradient of zonal mean potential vorticity $\bar{q}_y$ and the average wind speed $\bar{u}$. The strongest effect on the potential vorticity gradient is the gradient in planetary vorticity with latitude $\frac{\partial f}{\partial y}$, but its magnitude and sign can be significantly modulated by horizontal or vertical shear in the mean zonal wind. Outside of the tropics, $\bar{q}_y$ tends to be positive except for localized areas near the strongest wind shear.

The term Q can tell us quite a bit about the wave properties. Equation [4] has the form of a wave equation only when Q is positive. Since the background potential vorticity gradient $\bar{q}_y$ is usually positive, it is readily seen that the condition for propagation will not be met if the wind speed is less than the wave phase speed (or, for stationary waves forced by fixed features such as mountains, if the wind speed is less than zero). The locus of points where $(\bar{u} - c)$ goes to zero is known as the critical line. Normally, stratospheric winds are positive during winter and negative during summer, explaining the absence of the stationary planetary Rossby waves, for which $c = 0$, in the summer middle atmosphere when $\bar{u}$ is negative. Also note that there is a negative contribution that involves the wave number k^2 . For waves with larger wave numbers, this can become large enough that Q is negative. This explains why most Rossby waves observed in the middle atmosphere have low wave numbers (eqn [7]).

$$Q > 0 \Rightarrow k^2 < \frac{\bar{q}_y}{(\bar{u} - c)} - \frac{f^2}{4H^2N^2} \quad [7]$$

Q is also known as the refractive index squared, analogous to the geometrical optics property of the same name. Waves will be refracted in the latitude–altitude plane if there are gradients in Q . For example, for the case presented in Figure 4, Q increases toward low latitudes and has negative regions in high latitudes. Figure 6 shows contours of Q for zonal wave number 1 and, superimposed on this, arrows indicating the direction of propagation of the wave in the Northern Hemisphere's stratosphere. There is strong upward propagation in midlatitudes together with refraction toward the equator.

There is another way to look at these planetary waves that can shed light on how they interact with the background wind. Take the perturbation potential vorticity eqn [1], multiply it by the perturbation potential vorticity q' , and average zonally. This gives a conservation equation for the wave quantity q'^2 , which is the quasigeostrophic form of wave action (eqn [8]).

$$\frac{1}{2} \frac{\partial \overline{q'^2}}{\partial t} + \overline{v'q'} \frac{\partial \bar{q}}{\partial y} = \overline{q'D'} \quad [8]$$

It can be shown that the net poleward flux of potential vorticity, $\overline{v'q'}$, is proportional to the Eliassen–Palm (EP) flux divergence. The EP flux divergence is a quantity that is derived from the structure of the wave and is a precise measure of how the wave will force changes in the background atmosphere. A negative value of $\overline{v'q'}$, or equivalently of EP flux divergence, indicates that the wave is causing a negative acceleration of the background flow. In other words, the extent to which a wave can change the background average wind speed or temperature depends on $\overline{v'q'}$; if it is zero, the wave may be able to propagate

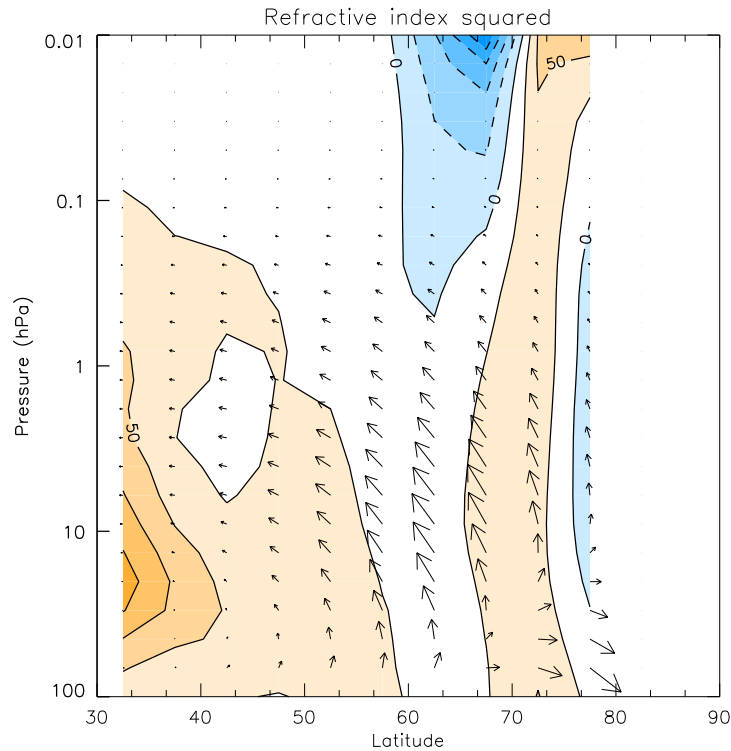


Figure 6 Contours of refractive index squared (Q) and vectors showing the direction and magnitude of energy propagation for wave number 1, for the case shown in Figure 4.

but will have no impact on the background atmosphere. From eqn [8], it is evident that $\overline{v'q'}$ will be nonzero only if the wave is not steady (first term nonzero) or is dissipating (right side of equation nonzero).

While eqn [8] is simple, its applications are important, and the phenomena it describes have a profound influence on middle-atmosphere circulation. In essence, a planetary wave will not influence the background temperature or winds and also will not cause any net transport of mass unless the wave is transient or is dissipating. In reality, planetary wave transience always occurs and dissipation can be significant, especially in the upper stratosphere and mesosphere. Planetary waves tend to reduce the wind speed and the strength of the winter-time vortex, to warm up the high-latitude stratosphere, and to induce transport of mass toward the winter pole and downward in the high-latitude stratosphere.

The largest amplitudes of planetary waves occur in the midlatitude upper stratosphere. At higher altitudes, dissipation becomes strong enough that the amplitudes begin to decrease with height. There are several processes that can dissipate the waves. The cooling by emission of heat (infrared radiation) is approximately proportional to temperature; warmer areas cool faster, thereby reducing the amplitude of the wave. Another important dissipation mechanism is interaction with small-scale waves. Propagation and dissipation of these small-scale waves will vary depending on the winds and temperatures of the planetary waves, and interactions will in turn affect the planetary wave structure. Small-scale waves and the turbulence produced by them become much more significant in the mesosphere (>60 km) and attenuate planetary waves that propagate to these levels.

As is evident from Figures 3 and 4, wave patterns are seen in both the stratosphere (from the tropopause up to about 0.03 hPa) and in the mesosphere (from 0.3 to the mesopause). Interactions of the large-scale dynamics with small-scale waves can damp wave structures but also generate waves. A planetary-scale variation in the small-scale waves that propagate into the mesosphere can exist due to planetary-scale variations in the generation of the small-scale waves. Another means of generating variations is through differences in the dissipation of the waves due to differences in winds, temperature, and so on. This means that planetary waves in the stratosphere can cause wavelike variations in the small-scale waves propagating to the mesosphere. When the small-scale waves dissipate and modify the dynamics in the mesosphere, they will generate a planetary wave pattern in the winds and temperature. The quasistationary planetary waves seen in winter above about 0.01 hPa probably contain a mix of propagating waves and wavelike variations due to interactions with small-scale waves.

Wave amplitudes also decrease toward low latitudes, even though the waves often propagate toward the equator. The waves cannot propagate where the refractive index squared is less than zero. As waves approach the low latitudes, they can undergo significant distortions as the normal restoring force associated with the polar gradient in potential vorticity $\bar{q}_y > 0$ begins to weaken. Since the wave itself can affect the potential vorticity, local regions where $q_y \leq 0$ will appear; air motion in such regions no longer experiences a restoring force but instead is unstable to perturbations in the north–south direction. Under these circumstances, there is a rapid and irreversible

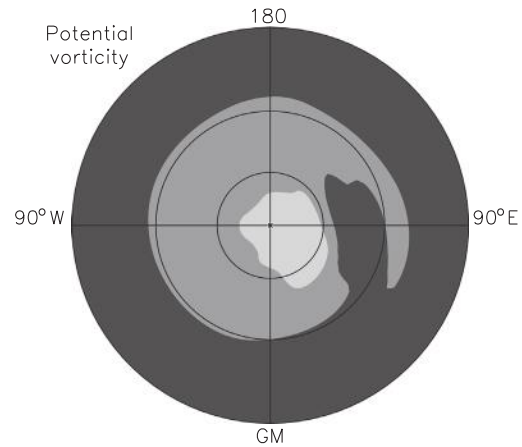


Figure 7 The potential vorticity field corresponding to Figure 1.

deformation of the potential vorticity contours on a pressure surface, and the wave is said to ‘break,’ an analogy to the breaking of waves on a beach. Figure 7 shows the potential vorticity field corresponding to Figure 1. There is evidence of a broad region of weak or reversed latitudinal gradients of potential vorticity in the vicinity of 90° E and of deformation of the contour lines as tongues of high-potential-vorticity air are drawn out into the low-potential-vorticity region. This wrapping of contours is indicative of Rossby wave breaking. It leads to dissipation of planetary waves, primarily in low latitudes, and to horizontal mixing.

Occasionally, strong amplification occurs, and the wave has a sudden strong impact on the background wind and temperature. Extreme events can alter the temperature of the polar stratosphere by as much as 40 K in a week. The strongest of these events are known as major sudden stratospheric warmings, and they are one of the most dramatic examples in the atmosphere of the interaction of waves with the mean fields. In addition to increasing the temperature, the warming events also break down the polar vortex, redistribute ozone and other chemicals, and reverse the direction of the stratospheric jet in high latitudes. Major sudden stratospheric warmings occur during about half of the winters in the Northern Hemisphere.

Although the sudden warming events are rare, the processes that lead to them occur with smaller amplitude throughout a normal winter. Midlatitude planetary waves in general act to warm the polar regions and to slow the speed of the stratospheric jet. Planetary wave amplitudes tend to be larger in the Northern Hemisphere than in the Southern Hemisphere because of the different distributions of continents and orography. As a result, Southern Hemisphere stratospheric polar temperatures are cooler, and the vortex is stronger. From an observational record extending back to 1957, only a single major sudden stratospheric warming (in 2002) has been observed in the Southern Hemisphere.

Middle-Atmosphere Planetary Waves and Stratospheric Climate

The variability of the middle-atmosphere circulation is to a large degree controlled by wave motions that originate in the

troposphere. However, there is increasing evidence from observations and atmospheric modeling that the dynamical coupling between the troposphere and the middle atmosphere is a two-way process. While our understanding about this two-way coupling and its implications has increased considerably during the last two decades, the mechanisms by which changes in mean winds or planetary waves in the middle atmosphere can influence the tropospheric flow are still an important topic of current research.

As discussed in Section Quasistationary Midlatitude Rossby Waves, the middle-atmosphere background wind determines the propagation characteristics of upward-propagating planetary Rossby waves. Two mechanisms of downward coupling between the middle atmosphere and the troposphere can be identified. The nature of downward coupling depends on the vertical and meridional structure of the background wind, which either allows planetary waves to freely propagate within the middle atmosphere or acts like a vertical reflective surface that reflects wave activity downward.

There are specific configurations of the zonal mean flow that prohibit the upward propagation above a specific altitude in the free atmosphere (Section Quasistationary Midlatitude Rossby Waves, eqn [7]). During the summer months, planetary Rossby waves are generally trapped in the troposphere because of the easterly background flow of the stratosphere. During the winter and early spring, a reflective surface in the middle atmosphere can form when either (1) the mean zonal wind is less than the phase speed of the wave (i.e., $\bar{u} < c$), or (2) the meridional gradient in the potential vorticity becomes negative ($\bar{q}_y < 0$) above the jet maximum as a result of the decrease of the zonal mean wind with height above the jet peak. From eqn [6], it is evident that a large positive value of $\frac{1}{\rho} \frac{\partial}{\partial z} \rho \frac{\partial \bar{u}}{\partial z}$ can lead to a negative value of $\bar{q}_y$, and therefore to a reflecting surface. An upward-propagating wave that reaches a vertical reflective surface will be reflected and will then propagate downward into the troposphere. The reflected wave can modify the phase and amplitude of the tropospheric wave source.

The impact of wave reflection is illustrated using observations from a period during February 2000. Figure 8 shows a sequence of daily longitude $\times$ height cross-sections (from 0 to 360° and 700 to 1 hPa) of the zonal wave structure averaged between 50 and 60° N for the period from 5 to 13 February 2000. Strong changes in the propagation characteristics of a wave with zonal wave number 1 occur during this period. During this sequence, the mean zonal flow in the middle atmosphere is characterized by negative vertical wind shear of the westerly jet (not shown) causing the formation of a vertical reflective surface in the midstratosphere. Vertical wave propagation is characterized by a phase tilt of the wave field in the vertical plane. A westward tilt with height indicates upward propagation, and an eastward tilt indicates downward propagation. During the first 2–3 days in the daily sequence (Figure 8), the wave number 1 field exhibits a clear westward phase tilt with increasing height, illustrating that the wave is propagating upward. In the following days, the amplitude of the wave number 1 field in the stratosphere decreases. The most dramatic change is in the phase of zonal wave number 1 in

the troposphere and lower stratosphere (200–20 hPa). In this region, the wave moves westward beginning on 8 February. As a result of this shift, the wave number 1 phase then tilts eastward with height, suggesting downward propagation of the wave.

Planetary wave reflection is the most direct way by which the stratospheric background wind can affect the tropospheric circulation. However, another process of downward coupling, which also involves planetary waves in the middle atmosphere, is expected to cause more large-scale effects on the tropospheric flow. Planetary waves that propagate vertically change the middle atmosphere flow in winter midlatitudes when they grow large enough to break and be absorbed. The wind changes that result from the dissipating wave in turn shift the location of the region of strongest interaction of the waves with the mean flow. As a result of this shift, there is a downward and poleward progression of the location of the background wind disturbances, which can eventually reach the tropopause level. The strength of this interaction depends both on the strength of the tropospheric wave forcing and on the background flow itself. This process, which can cause long-lasting zonal wind anomalies at the tropopause level, is often accompanied by circulation anomalies all the way to the surface. The surface anomalies resemble a pattern called annular mode, which exhibits a hemispheric-scale spatial pattern of climate variability characterized by north–south shifts in mass between polar and lower latitudes. The mechanisms by which the long-lasting zonal wind anomalies at the tropopause resulting from wave mean flow interactions in the middle atmosphere are transmitted to the surface are still not well understood.

Normal Modes

Another class of waves that can be significant in the middle atmosphere is that of Rossby normal modes, also known as free modes. These are waves that correspond to natural modes of variability of the Earth's atmosphere. Based on the size and rotation rate of the Earth and the depth of the atmosphere, there are preferred responses. From eqn [5], it is clear that the restoring force Q depends on the phase speed of the wave c . In an undamped isothermal atmosphere ($D = 0$, $T = T_0$) with no background wind ($\bar{u} = 0$ and $\bar{q}_y = f$), there can exist global solutions to eqn [4] that are finite even without forcing. These are the normal modes, which can be considered to be resonances of the global atmosphere. The vertical structure of each mode can be determined from eqn [4] when damping is omitted and no waves are introduced by the boundary conditions. While these theoretical modes are the responses that would occur in an isothermal atmosphere without damping, the actual atmospheric conditions often allow for the existence of waves that are similar to the idealized modes. A perturbation to the atmosphere that excites a spectrum of waves can include one or more of these normal modes, which then grow in amplitude with height due to the decrease in density.

A number of these modes have been identified theoretically for realistic conditions. They can reach large amplitudes in the middle atmosphere. One commonly observed mode is

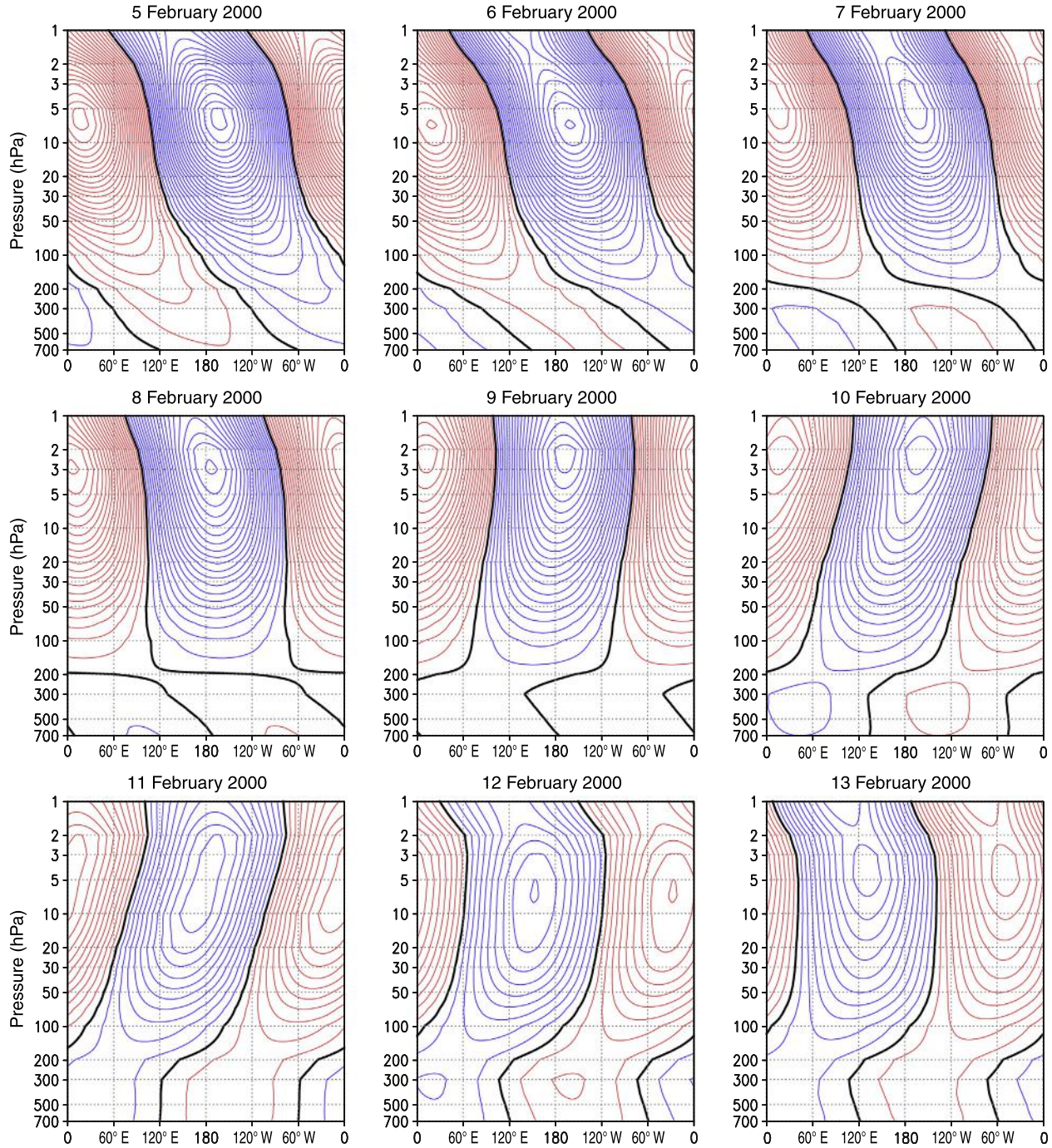


Figure 8 Longitude-height cross-sections of zonal wave one geopotential height (m) averaged between 50 and 60° N for the daily temporal sequence from 5 to 13 February 2000. Positive and negative values are in red and blue, respectively; zero values are in black. The isoline interval is 50 m.

the quasi-two-day wave, which has a period of about 2 days and a zonal wave number of 3 or 4. It is regularly observed in the mesosphere just after solstices and can attain very large amplitudes (meridional wind $\sim 30 \text{ m s}^{-1}$). Two modes with zonal wave number 1 also appear regularly, although they have smaller amplitudes: the 16-day wave, most commonly seen in the winter hemisphere, and the 5-day wave. These waves do not transport much momentum and do not

normally have a major direct impact on the global momentum balance in the stratosphere. However, they can interact with other waves such as quasistationary Rossby waves, gravity waves, and tides and thereby affect the periodicity of variability in the middle atmosphere. Because of its slow phase speed, the 16-day wave is a significant component of the atmospheric response to quasistationary disturbances.

Equatorial Waves

Another class of planetary waves occurs only near the Equator. These rely on both buoyancy forces and the conservation of potential vorticity, and they are not represented by eqn [1]. Of these, the most commonly observed in the middle atmosphere is the Kelvin wave, which is regularly seen in satellite observations with high vertical resolution. A theoretical Kelvin wave in the β -plane approximations has no meridional winds. Its structure can be determined from the momentum and thermodynamic equations under simplified conditions ($\bar{v} = v' = \bar{w} = D' = 0$). Again, we assume a waveform with zonal wave number k and frequency ω and also assume a vertical wave number m (eqn [9]).

$$\phi' = \phi_k \exp[i(kx + mz - \omega t) + z/2H] \quad [9]$$

The governing equations can then be written as eqns [10]–[12].

$$-(\omega - k\bar{u})u' + k\phi' = 0 \quad [10]$$

$$\beta y u' + \phi'_y = 0 \quad [11]$$

$$k u' - (\omega - k\bar{u}) \frac{m^2}{N^2} \phi' = 0 \quad [12]$$

Eliminating u' from eqns [10] and [11] gives the meridional structure of the wave (eqn [13], where ϕ_0 is the wave geopotential at the equator).

$$\phi' = \phi_0 \exp \left[-\frac{\beta k y^2}{2(\omega - k\bar{u})} \right] \quad [13]$$

Figure 9 shows the latitudinal structure of geopotential and zonal wind for an idealized Kelvin wave.

A range of Kelvin waves have been observed. All are traveling waves that move eastward with time. They have wave perturbations in temperature and zonal and vertical winds; the meridional winds associated with the wave are zero. Kelvin waves with low zonal wave numbers (waves 1–3) have the largest amplitudes. Some Kelvin waves have periods of 24 h and are therefore also classified as atmospheric tides.

Equations [10] and [12] can be combined to give a dispersion relation (eqn [14]), which relates the frequency of the wave (ω) to the zonal (k) and vertical (m) wave numbers. The wave

numbers have units of inverse length and are related to the wavelengths of the wave; for example, the vertical wavelength is $2\pi/m$. Lower-frequency waves will have larger vertical wave numbers or, equivalently, short vertical wavelengths.

$$(\omega - k\bar{u}) = \frac{-Nk}{m} \quad [14]$$

Low-frequency Kelvin waves that take 10–20 days to propagate around the globe occur in the lower stratosphere but are not able to propagate deep into the middle atmosphere. Eventually, they are likely to encounter a critical level when the background zonal wind speed is equal to the phase speed of the wave ($\omega - k\bar{u} = 0$), and the wave can no longer propagate. The tropical stratospheric winds are characterized by temporal variability associated with the quasibiennial (QBO) and semiannual (SAO) oscillations. The locations of the critical levels change with time owing to these oscillations in $\bar{u}$. Higher frequency Kelvin waves can propagate through the stratosphere and sometimes as high as the upper mesosphere. The deeper propagation is possible because their phase speeds exceed wind speeds that are found at lower levels (i.e., they do not encounter a critical level). Propagation is also facilitated because the faster phase speeds are associated with larger vertical group velocities and are less effectively damped.

Another equatorial mode is known as the mixed Rossby-gravity wave. This mode propagates westward and has been observed in the stratosphere. Although the mixed Rossby-gravity wave is, like the Kelvin wave, an equatorially trapped traveling wave, it has some key differences. It propagates westward and has finite meridional wind. In some ways, mixed Rossby-gravity waves and Kelvin waves in the stratosphere are complementary because they tend to be seen in different wind regimes, have zonal phase speeds in opposite directions, and carry zonal momentum of opposite signs.

When a wave encounters a critical level, where $\omega = k\bar{u}$, it can no longer propagate. The momentum carried by the wave will be absorbed at or near that point and will alter the background wind speed. Kelvin and mixed Rossby-gravity waves will propagate upward (Kelvin waves through easterly winds, and mixed Rossby-gravity waves through westerly winds) only until they reach a critical level. Mean zonal wind speeds in the equatorial stratosphere, and hence also the locations of critical

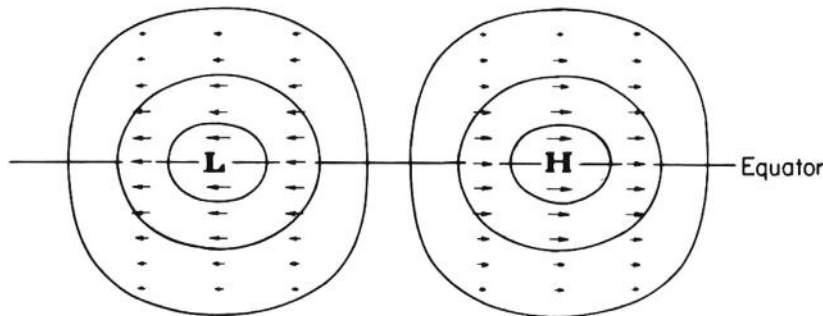


Figure 9 Longitude × latitude structure of the geopotential and horizontal winds of an idealized Kelvin wave. Reproduced with permission from Andrews, D.G., Holton, J.R., Leovy, C.B., 1987. *Middle Atmosphere Dynamics*. Academic Press, Orlando (Adapted from Matsuno, T., 1966. Quasi-geostrophic motions in the equatorial area. *Journal of Meteorological Society of Japan*, 44, 25–43.).

levels, oscillate with time on seasonal to interannual time scales, in connection with the SAO and QBO. The momentum deposited at that point actually alters the evolution of these oscillations. The QBO is believed to be driven to a large extent, and the SAO in part, by the momentum deposited by these tropical waves.

See also: **Dynamical Meteorology:** Kelvin Waves; Quasigeostrophic Theory; Rossby Waves; Stationary Waves (Orographic and Thermally Forced). **Middle Atmosphere:** Polar Vortex.

- Houghton, J.T., 1986. *The Physics of Atmospheres*, second ed. Cambridge University Press, Cambridge.
- James, I.N., 1994. *Introduction to Circulating Atmospheres*. Cambridge University Press, Cambridge.
- Labitzke, K., van Loon, H., 1999. *The Stratosphere: Phenomena, History and Relevance*. Springer-Verlag, Berlin.
- Mohanakumar, K., 2008. *Stratosphere Troposphere Interactions: An Introduction*. Springer, New York.
- Perlwitz, J., Harnik, N., 2004. Downward coupling between the stratosphere and troposphere: the relative roles of wave and zonal mean processes. *Journal of Climate* 17, 4902–4909. <http://dx.doi.org/10.1175/JCLI-3247.1>.
- Polvani, L.M., Sobel, A.H., Waugh, D.W. (Eds.), 2010. *The Stratosphere: Dynamics, Transport, and Chemistry*. American Geophysical Union, Washington.

Further Reading

- Andrews, D.G., Holton, J.R., Leovy, C.B., 1987. *Middle Atmosphere Dynamics*. Academic Press, Orlando.
- Eyring, V., Shepherd, T.G., Waugh, D.W. (Eds.), 2010. SPARC Report on the Evaluation of Chemistry-Climate Models. World Climate Research Program. <http://www.atmos.physics.utoronto.ca/SPARC>.

Polar Vortex

MR Schoeberl, Science and Technology Corporation, Lanham, MD, USA
PA Newman, NASA Goddard, Space Flight Center, Greenbelt, MD, USA

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

The polar vortex refers to a region of the winter polar stratosphere characterized by high nearly zonal westerly winds and isolation from the rest of the stratosphere and the extreme cold temperatures within the vortex that allows for cloud formation and the complex chemistry of rapid ozone loss to occur in late winter and early spring. The polar vortex extends from the tropopause, (8–11 km in altitude), to the stratopause (50–60 km in altitude). Above the stratopause the zonal winds reverse. The Southern Hemisphere (SH) polar vortex is much stronger than the Northern Hemisphere (NH) vortex due to the absence of large-scale waves – planetary waves – in the SH. Planetary waves disrupt the vortex producing a short period circulation reversal known as a Major Stratospheric Sudden Warming. Major sudden warmings occur about every other year during the NH winter, but, due to the lack of planetary wave activity, they are very rare in the SH winter. The only observed SH sudden warming occurred in 2002. The polar vortex forms in fall and persists through winter into spring at which point it breaks up giving rise to the easterly summer circulation. The breakup of the polar vortex in the spring is called the final warming. The final warming occurs near the spring equinox in the NH, but occurs 1–2 months later in the SH. There is good evidence that the Antarctic ozone depletion has produced to a more persistent SH polar vortex. The spring vortex breakup date is now late November rather than late October as was observed in the 1970–80s.

The polar vortex is the region of high-atmospheric vorticity that forms with the establishment of the winter stratospheric polar jet (Figure 1). The polar jet begins just above the tropopause in both the Northern and Southern Hemispheres reaching maximum wind speeds near the stratopause (~50 km). The polar jet arises from the strong temperature contrast between the warm tropical stratosphere and the cold polar stratosphere. The tropical stratosphere is heated by the ozone ultraviolet absorption. The polar winter stratosphere is unheated during

polar night and cools through infrared emission to space principally by the gases carbon dioxide, ozone, and water. The resultant equator to pole temperature difference creates a strong pressure gradient and, as air moves northward, the Coriolis force deflects this air eastward, creating a strong eastward-flowing jet. Above the stratopause, the temperature gradient between the tropics and the polar region reverses, and polar night jet speed decreases with altitude into the mesosphere. The winter polar vortex is seen in both hemispheres,

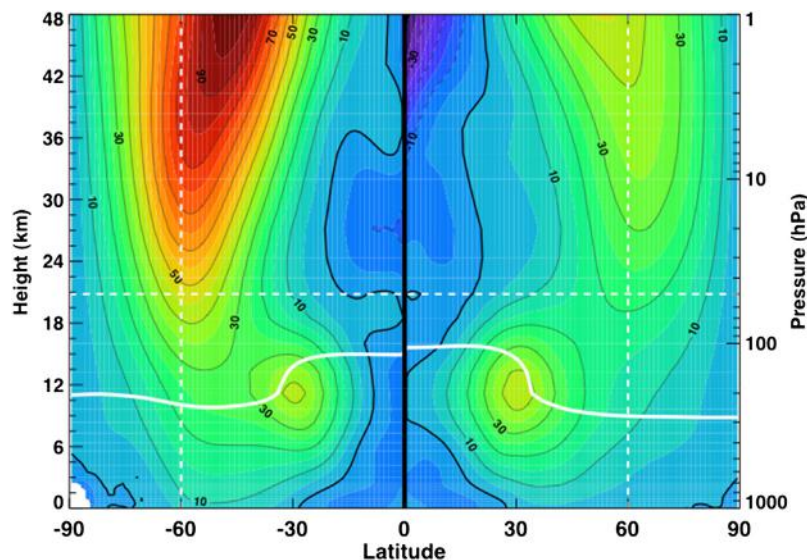


Figure 1 Contours of the wintertime zonal mean zonal (west-east) wind in m s^{-1} for July in the SH (left) and January in the NH (right) based upon 1979–2010 Modern Era Retrospective-Analysis for Research and Applications or MERRA data set (Rienecker, M., et al., 2011. MERRA: NASA's modern-era retrospective analysis for research and applications. *Journal of Climate* 24, 3624–3648. doi:10.1175/JCLI-D-11-00015.1). The log-pressure height ($7 \log(1000/p)$ km) is shown on the left with pressure in hPa on the right. The stratosphere is the region above the tropopause (indicated with the solid white lines). The polar vortex is the region poleward of the strong stratospheric wind jet. The dotted white lines indicate the locations of the cross sections shown in Figure 2.

but the Southern Hemisphere (SH) winter vortex is significantly stronger and longer lasting than its Northern Hemisphere (NH) counterpart. The polar jet almost completely isolates the polar vortex from the mid-latitude stratosphere. This vortex isolation along with the heterogeneous chemical processing on stratospheric clouds creates the conditions for severe polar ozone depletion over the winter poles. The linkage of the polar vortex to the ozone depletion has generated considerable interest and research into the formation and breakup of the polar vortex. Below we review the development, evolution, and breakup of the polar vortex. We also review sudden stratospheric warmings and the differences between the NH and SH vortices.

Formation of the Polar Vortex

The polar vortex begins to form after the fall (autumnal) equinox when solar heating of the polar ozone layer is cut off with the onset of polar night. Without solar heating, infrared (IR) cooling of the polar air mass causes temperatures to fall and air begins to descend over the polar region. By continuity, air also flows poleward as the polar jet intensifies (*see Radiation Transfer in the Atmosphere: Absorption and Thermal Emission; Scattering*). The polar vortex usually becomes well organized as a strong zonal (East-West) flow by mid-November. **Figure 2** shows the evolution of the vortex zonally or longitudinally with averaged wind speed at 50 hPa (~ 20 hPa). Satellite data obtained by the Halogen Occultation Limb Experiment aboard the Upper Atmosphere Research Satellite first showed that a significant amount of mesospheric air enters the stratosphere during the formation phase of the polar vortex. Subsequent model calculations showed that this descent was in agreement with radiative transfer theory. Polar air parcels in the midstratosphere may descend a scale height or more (~ 7 km) during the formation of the vortex. At lower altitudes, the descent is not as large, but this vortex formation period is an important mechanism for the return of air from the middle and upper stratosphere to the lower stratosphere. Because air from the upper stratosphere contains a lower concentration of organic chlorine relative to inorganic chlorine, the descending circulation enriches the inorganic chlorine content of the polar vortex and is a factor in the severity of polar ozone depletion.

The Mature Vortex

In most years, the NH polar vortex winds reach peak intensity during mid-January. It is during this period that polar stratospheric temperatures reach their lowest values. From 20 to 25 km, temperatures approach 190 K, and clouds form in the lower stratosphere. At an altitude of about 20 km the vortex has an area of 2.1×10^7 km² (2.9×10^7 in the SH). **Figure 3** shows the midwinter climatology of the NH and SH polar vortices. While the SH vortex is nearly symmetric, the NH vortex is distorted. Streamlines are displaced poleward in the North Pacific and over the Aleutian islands by an anticyclonic anomaly. This feature is sometimes called the Aleutian anticyclone.

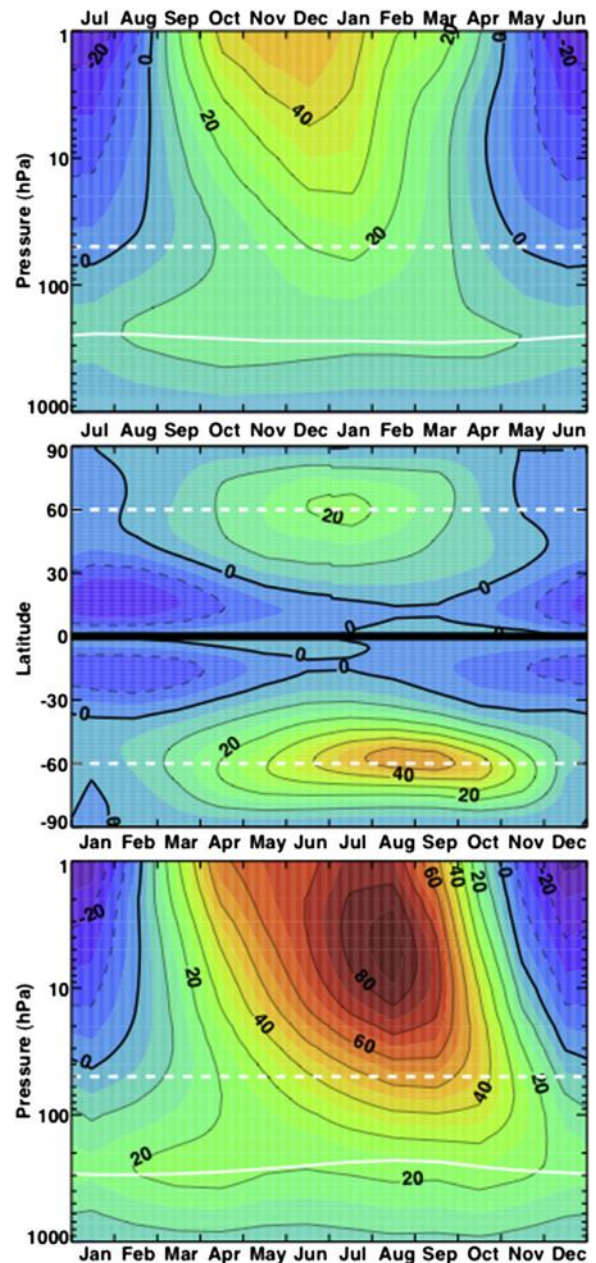


Figure 2 (Middle) Time–latitude diagram of the zonal mean wind speed at 50 hPa (~ 20 km) (see the horizontal dotted line in **Figure 1**). The top of the image has been shifted 6 months (July to June) with respect to the bottom (January to December) to emphasize the differences between the midwinter flow. Both polar jets begin to intensify just before equinox and continue to grow in strength reaching a peak in mid- to late-winter (January in NH, August in SH). (Top) 60°N Time–height cross sections and (Bottom) 60°S Time–height cross sections. Again, the 60°N cross section is shifted by 6 months for comparison. The vortex winds intensify at highest altitudes first and the isotachs appear to descend as the vortex intensifies later at lower altitudes. The NH vortex reaches peak wind speeds, on an average, between in mid-December and in mid-January. The SH vortex reaches peak wind speeds about 2 months after winter solstice.

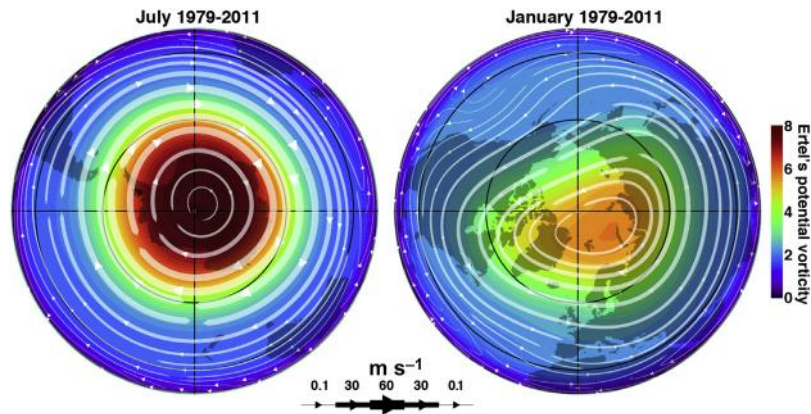


Figure 3 Streamlines of the monthly mean wind flow (white arrows) along with the structure of the vortex (Ertel's potential vorticity) on the 500 K isentropic surface (~ 20 km altitude) for the SH (left) and NH (right). The thickness of the streamlines denotes the wind speed. Note that the SH vortex is stronger (higher values of potential vorticity) and is more symmetric around the pole. The potential vorticity is computed using winds and temperatures from a 33-year average of the MERRA stratospheric analyses.

Sudden Stratospheric Warmings

The strong zonal winds of the polar vortex provide a channel for the upward propagation of Rossby waves from the troposphere. Charney and Drazin (1961) first derived a dispersion relation for vertically propagating Rossby waves and showed that only the planetary-scale waves ($\lambda > 6000$ km) could penetrate the strong westerly polar vortex winds during winter. Synoptic scale Rossby waves ($\lambda < 6000$ km), responsible for most of the tropospheric variability, would be trapped in the troposphere. Planetary-scale Rossby waves (planetary waves for short) are principally forced by airflow over topography,

although they can also arise from longitudinal variations in thermal forcing or nonlinear interaction between short-scale waves. The stationary zonal harmonic one planetary wave seen in Figure 3 displaces the polar vortex away from the North Pole. This planetary wave is linked to the flow over the Himalayan Plateau. As planetary waves propagate into the stratosphere, they tilt westward with altitude and distort the longitudinally symmetric structure of the vortex. As viewed from an observer hovering over the pole, the circumpolar jet appears to be offset from the pole, or elongated across the pole. As these waves move, it appears as if the vortex is wobbling (see Figure 4). These wobbles can grow so large that the vortex is

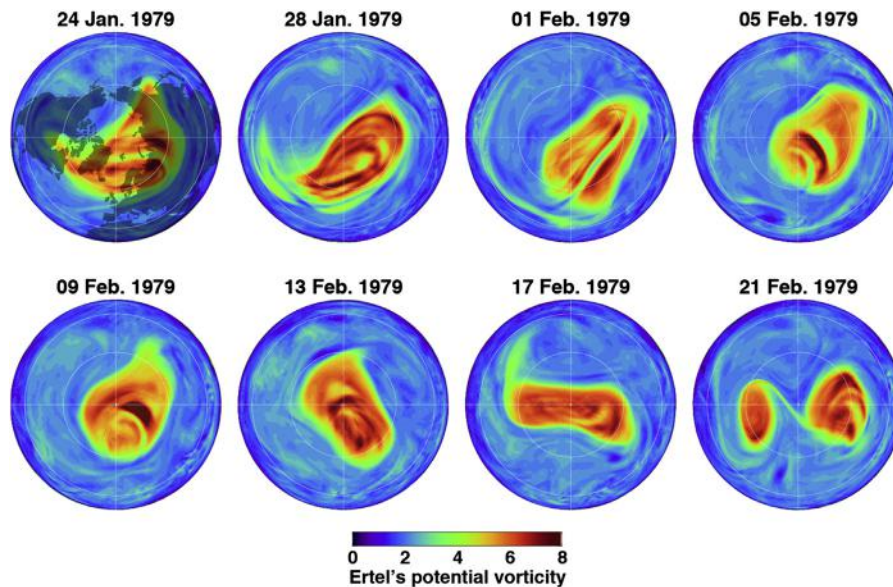


Figure 4 Sudden warming in the Northern Hemisphere is shown in this sequence of potential vorticity plots on the 500 K potential temperature surface (~ 20 km). The projection is polar orthographic (North Pole is at the center). The sequence begins with the vortex displaced off the pole by the Aleutian anticyclone. The vortex elongates, rotates, and finally splits into two fragments over the period of about 1 month.

completely pulled off of the pole so that the zonal mean midlatitude temperature gradient is reversed. This condition is called a sudden stratospheric warming and is relatively common in the NH. In 2002, the SH produced its only observed sudden warming. **Figure 5** shows the vortex structure and the streamlines for a series of dates during this September 2002 major warming. A warming disrupts the vortex and fragments of the vortex move to extrapolar latitudes. For example, the 5 October panel of **Figure 5** shows a long streamer of air that has broken off of the polar vortex following the warming and now extends westward from the vortex (about 40° E to about 90° W).

The World Meteorological Organization definition of a major sudden stratospheric warming requires a winter zonal mean wind reversal at 10 hPa. A rapid decrease in wind speed but without a reversal in the temperature gradient is called a minor warming. A stratospheric warming is not considered an instability of the vortex since none of the usual atmospheric instability conditions are met (see **Middle Atmosphere: Stratospheric Sudden Warmings**).

Sudden stratospheric warmings appear to descend from higher altitudes. Modeling studies have pointed to critical layer interaction as the mechanism for this descent. A critical layer is a region where the zonal mean wind speed is equal to the zonal phase speed of a propagating wave. An upward-propagating large-scale planetary wave increases in amplitude (as measured by the geopotential anomaly, for example) with increasing altitude as the atmospheric density decreases. Eventually, the

flow becomes so highly distorted in the upper atmosphere that the wave can no longer propagate and the wave energy and momentum are ‘dumped’ into the flow as a wave-breaking event. The strong nonlinear wave–mean flow interaction takes place at the ‘critical level’ for the stationary planetary wave or the zero zonal wind speed contour. The sudden warming disturbance appears to move downward because the secondary circulation driven by the wave interaction with the critical level extends below the level of wave breaking, decelerating the mean flow below the wave-breaking zone. The extension of the secondary circulation associated with wave breaking below the level of wave breaking is sometimes called ‘downward control.’ As the critical level moves downward, the planetary wave breaks at much lower altitudes hence the sudden warming appears to descend. A similar mechanism is responsible for the descent of the alternate easterly and westerly winds in the tropics that is called the Quasibiennial Oscillation. As the stratospheric warming penetrates into the troposphere, it can cause a shift in the climatic patterns.

A sudden warming can develop in any reasonable stratospheric numerical model that allows upward planetary wave propagation and that is forced by a planetary wave with increasing amplitude at the lower boundary. What is not understood is what causes the planetary wave amplitude to increase rapidly. Some studies have suggested that resonance-type instabilities may be responsible for the amplification. A more descriptive approach has also been useful in understanding the development of the sudden warming as shown in

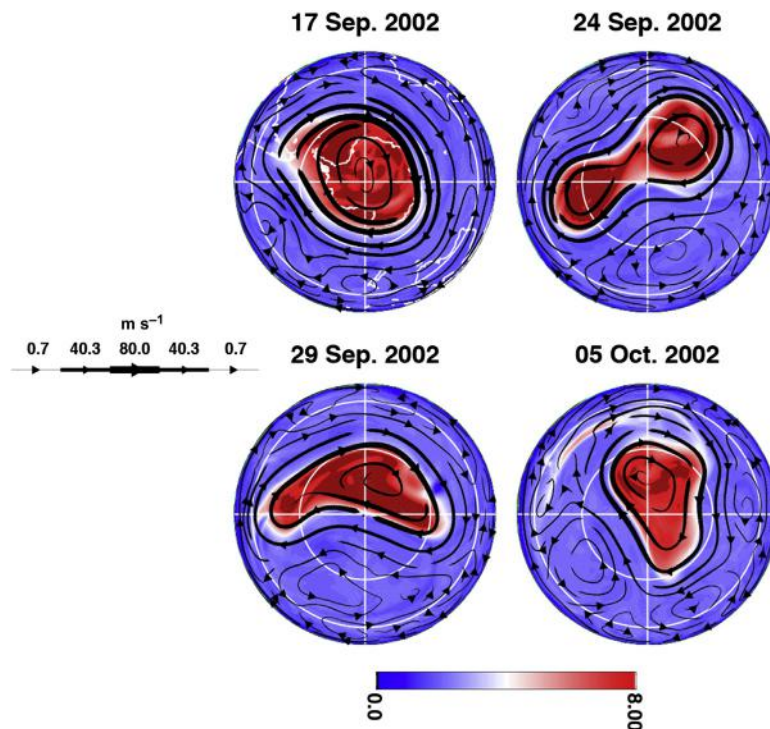


Figure 5 A sequence of streamline and potential vorticity maps display the only major stratospheric warming observed in the historic record of the Southern Hemisphere (500 K potential temperature surface, ~20 km). Note how the early map is relatively symmetric about the pole, but is highly distorted into a planetary wave-2 structure by 24 September. By 5 October, the polar vortex has moved back onto the South Pole, but winds are greatly weakened over this 19-days period. The weakening of the winds and vortex is accompanied by a large temperature increase in the stratosphere.

Figure 4. If the vortex is perturbed from a zonally symmetric position, the strong winds will begin to transport low-potential vorticity (PV) air from the tropics toward the polar region. The tropical air, which has an anticyclonic vorticity, tends to further increase the vortex displacement from the pole. Continued forcing from below causes the vortex to dip further toward the tropics and further increases the northward transport of anticyclonic air. This process continues amplifying the anticyclone generating the sudden warming event. Subsequent to the displacement of the vortex from the pole, elongation of the vortex often creates two zones where anticyclonic air can be transported northward, and a secondary ridge forms subsequently splitting the vortex in two. When the vortex is displaced

off the pole, the dominant planetary wave is zonal harmonic one. With the strong vortex winds disrupted, the poleward transport of very-low-PV air weakens, and radiative cooling of the polar air causes downward transport of high-PV air from the mid- and upper stratosphere, reestablishing the polar vortex. Depending on the time of year and altitude, the recovery usually takes a few weeks as is dictated by the IR cooling rates in the lower stratosphere. Sudden stratospheric warmings are most common in mid- to late winter, but they have been observed as early as December in the NH. From the mid-1950s to 1991, major warmings occurred about every other year, but between 1992 and 1998 there were no major warmings. A spectacular December 1998 major warming broke this lull.

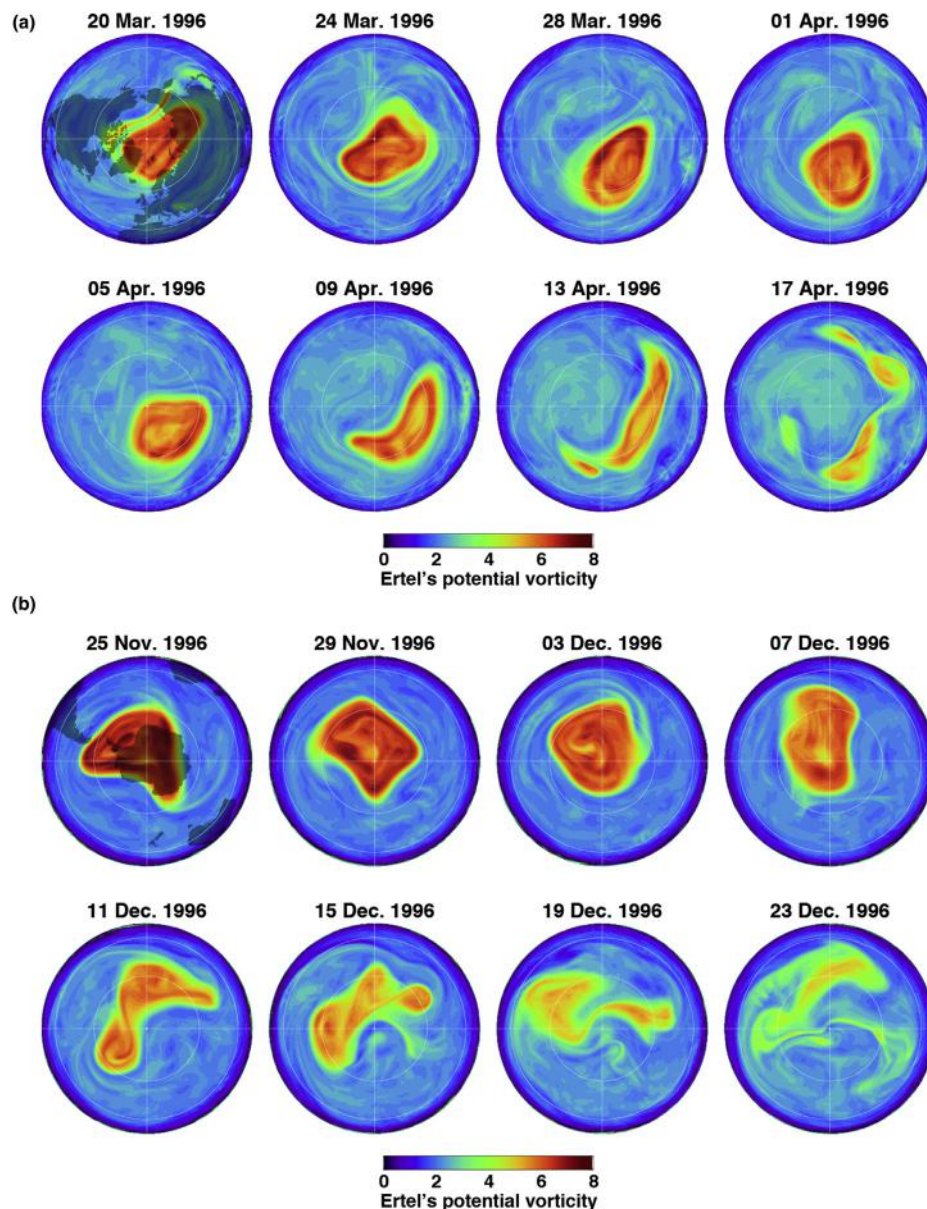


Figure 6 The final warming in the NH (a) and SH (b) at the 500 K potential temperature surface (~ 20 km). Although each year is slightly different, these 2 years typify the process. Final warmings mark the breakup of the polar vortex and the onset of the summer circulation. The appearance of a triangular pattern in the SH is not surprising since planetary wave three is forced by topography. Final warmings in the NH occur around equinox, whereas final warmings in the SH may be more than 2 months after the equinox (note the difference in the vortex duration shown in [Figure 2](#)).

Transport across the Polar Vortex Boundary

The size of the polar vortex can be measured by a number of methods. The most common is to use the contour of the peak in the PV gradient. This gradient region is seen in [Figure 4](#) as the strong transition from the higher PV (red) to lower PV values (blue). After reaching maturity in midwinter, the polar vortex area begins to decrease. This erosion process occurs as planetary waves pull PV off the vortex edge, often into spectacular filaments associated with sudden warmings. One such filament is visible in the lower left corner of the second map in [Figure 4](#). High-resolution simulations of the vortex show that the vortex is always shedding smaller filaments of vorticity. The erosion of the vortex is continuous, but it is most rapid during sudden stratospheric warmings. This erosion process mixes vortex air into the midlatitude air (see 5 February 1979, [Figure 4](#)). Occasionally, the atmospheric circulation forces midlatitude air into the interior of the vortex, but this process rarely happens above 100 hPa (16 km). Thus, the polar vortex tends to isolate vortex air from midlatitude influence, with the vortex edge acting as a mixing barrier. The isolation of the polar vortex air from midlatitude air is a key factor in the development of polar ozone depletion. Without the isolation, the high concentration of ozone-destroying chemicals could not be sustained. Below 100 hPa, the polar vortex is less organized and intrusions from the tropical troposphere are often seen. However, because air is descending within the vortex during winter, these intrusions do not affect the chemistry above 100 hPa.

Breakup of the Polar Vortex

The polar vortex typically lasts until the spring equinox in the NH and up to 2 months or more past the spring equinox in the SH. Solar heating of the polar ozone layer increases toward equinox, thus reducing the ability of the strong IR cooling to restore the temperature gradient. As during midwinter, the strong planetary activity in spring disrupts the vortex, but the warmer stratospheric temperatures that follow the warming are much closer to the radiative equilibrium temperature. Thus no strong diabatic descent of high-PV air follows the event and the vortex does not recover. This event is called the final warming. After the final warming, the circulation in the zonal mean winds remains as easterlies in the lower stratosphere, signaling the onset of the summer circulation. Fragments of the chemically perturbed vortex air mass may persist throughout the summer and have been observed in the lower stratosphere. [Figure 6](#) shows the sequence of events that lead to a final warming in the NH and SH, respectively. These dates are typical for each hemisphere, although final warmings may be later or earlier by 15 days or so.

Differences between the NH and SH Vortices

The SH polar vortex forms more rapidly, is stronger, covers a larger area, persists longer than the NH vortex ([Figure 3](#)), and, unlike its boreal counterpart, shows little year to year variability in strength ([Figures 1](#) and [2](#) middle and bottom

panels). The stronger vortex is due to the lack of planetary wave activity in the SH. This lack of wave activity is due to the relative flatness of the SH in comparison to the large-scale continental topographic features of the NH. Indeed, some of the strongest SH planetary wave features are not stationary but traveling waves. Major midwinter stratospheric warmings have been observed only once in the SH since 1979. The paucity of SH warmings is a result of weaker poleward eddy heat flux, and this weak poleward heat flux leads to a much weaker poleward and downward circulation over Antarctica. This dynamically driven downward circulation warming warms the polar stratosphere. Its absence results in much colder vortex temperatures and polar stratospheric clouds form over wide regions, giving rise to the more severe Antarctic ozone depletion. The persistent nature of the Antarctic polar vortex is also a result of a lack of planetary wave activity that can erode the Antarctic vortex. Historically, the Antarctic vortex was observed to last until about a month after spring equinox (the NH vortex lasts until around spring equinox). More recently, the lower stratospheric Antarctic vortex has been observed to persist until late November. This persistence is likely due to the ozone loss within the vortex interior, which reduces the solar heating after equinox and slows the rise of polar stratospheric temperatures. Ozone depletion also takes place within the NH vortex, but no convincing evidence yet exists that this depletion has had an effect on the timing of breakup of the NH polar vortex.

See also: Middle Atmosphere: Planetary Waves; Stratospheric Sudden Warmings; Transport Circulation. **Ozone Depletion and Related Topics:** Ozone Depletion Potentials. **Radiation Transfer in the Atmosphere:** Absorption and Thermal Emission; Scattering. **Stratosphere/Troposphere Exchange and Structure:** Global Aspects.

Further Reading

- Andrews, D.G., Holton, J., Leovy, C., 1987. *Middle Atmosphere Dynamics*. Academic Press, New York.
- Brasseur, G.P., Orlando, J.J., Tyndall, G.S., 1999. *Atmospheric Chemistry and Global Change*. Oxford University Press, New York.
- Butchart, N., Remsberg, E.E., 1986. The area of the stratospheric polar vortex as a diagnostic for tracer transport on an isentropic surface. *Journal of Atmospheric Science* 43, 1319–1339.
- Charney, J.G., Drazin, P.G., 1961. Propagation of planetary scale disturbances from the lower atmosphere into the upper atmosphere. *Journal of Geophysical Research* 66, 83–109.
- Dessler, A., 2000. *The Chemistry and Physics of Atmospheric Ozone*. Academic Press, London.
- Plumb, A., 2002. Stratospheric transport. *Journal of the Meteorological Society of Japan* 80, 793–809.
- Rienecker, M., et al., 2011. MERRA: NASA's modern-era retrospective analysis for research and applications. *Journal of Climate* 24, 3624–3648. <http://dx.doi.org/10.1175/JCLI-D-11-00015.1>.
- Schoeberl, M.R., Hartmann, D., 1991. The dynamics of the stratospheric polar vortex and its relation to springtime ozone depletions. *Science* 251, 46–52.
- Shepherd, T.G., 2007. Transport in the middle atmosphere. *Journal of the Meteorological Society of Japan* 85B, 165–191.
- World Meteorological Organization (WMO), 1986. *Atmospheric Ozone, 1985*. WMO Report 16. WMO, Geneva.

Quasi-Biennial Oscillation

TJ Dunkerton, Northwest Research Associates, Bellevue, WA, USA

JA Anstey and LJ Gray, University of Oxford, Oxford, UK

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision and update of the previous edition article by T J Dunkerton, volume 3, pp 1328–1336, © 2003, Elsevier Ltd.

Synopsis

Interannual variability of winds and temperatures in the lower and middle tropical stratosphere is dominated by the quasi-biennial oscillation (QBO). The QBO is characterized in rawinsonde observations by slowly descending easterly and westerly wind regimes, with each cycle of the oscillation lasting 2–3 years. Concentrations of ozone and other trace gases are modulated by the QBO, and its influence also extends during winter to higher latitudes in the stratosphere, and to some degree also into the troposphere.

Introduction

The quasi-biennial oscillation (QBO) of the equatorial lower and middle stratosphere is a fairly regular 2- to 3-year cycle characterized by alternating descending regimes of easterly (westward propagating) and westerly (eastward propagating) zonal wind (Figure 1). Accompanying the zonal wind QBOs are anomalies of temperature, trace constituents, and mean meridional circulation. The QBO dominates the equatorial lower and middle stratosphere and affects other regions of the atmosphere, including the tropical troposphere, tropical upper stratosphere and mesosphere, and the extratropical middle atmosphere in winter.

The QBO is more regular than El Niño/Southern Oscillation (ENSO), but is not biennial or otherwise exactly synchronized with the seasonal cycle. The oscillation is, however, modulated by the seasonal cycle in one or more ways. Onset of new QBO westerly regimes in the middle stratosphere is linked to descending westerly phases of the stratopause semiannual oscillation (SAO) (see **Middle Atmosphere: Semiannual Oscillation**). Not all SAO phases initiate a new QBO phase, obviously, since the period of the QBO is 4–6 times that of the SAO. New phases of the QBO aloft generally wait for old phases of the same sign to decay in the lower stratosphere and disappear. Below 30 hPa, the seasonal cycle modulates the final descent of old QBO phases (Figure 2). This seasonal effect is distinct from that of the SAO and is likely caused by annual cycles of upwelling and of wave fluxes entering the stratosphere from below.

The QBO period is unrelated to any known periodic forcing and is therefore difficult to understand, but we should not overlook the regular nature of the oscillation. The predictability of the QBO is remarkable, considering that wave motions responsible for the QBO (described below) occupy a broad spectrum of spatial and temporal scales, are episodic, complicated, and inherently unpredictable. The oscillation has now been observed by regular radiosonde soundings for more than half a century (26 complete cycles), with fairly constant amplitude, and by all appearances is a permanent feature of the tropical stratosphere.

QBO of Mean Zonal Wind

The prominent aspect of the QBO is an oscillation of mean zonal wind, where ‘mean’ refers either to a zonal average of

data spaced uniformly around a latitude circle or to a time average of soundings at an individual rawinsonde station (typically averaged over 1 month). Zonal or temporal averaging eliminates most high-frequency waves in the tropical stratosphere. However, not all zonal asymmetries are eliminated by time averaging. Quasi-stationary planetary waves propagate from the winter hemisphere into the middle and upper equatorial stratosphere and mesosphere. Monsoon circulations originating in the tropical troposphere penetrate the lowermost tropical stratosphere. It is unknown whether these zonally asymmetric motions are an important part of the QBO, i.e., to what extent the QBO departs from perfect zonal symmetry. The traditional view of the QBO as a zonally symmetric phenomenon is accurate enough to provide a basis for theoretical, numerical, and experimental studies of the oscillation.

The QBO is the dominant component of mean-flow variability in the equatorial lower stratosphere, and is approximately symmetric about the equator. The annual cycle of mean zonal wind is approximately antisymmetric and is easily removed from the data by subtraction of the climatology. The SAO attains measurable amplitude in the tropical middle stratosphere, and dominates the upper stratosphere and lower mesosphere. There is also a ‘mesopause SAO’ out of phase with the stratopause SAO. Both SAOs are affected by the underlying QBO.

Details of the wind oscillation are as follows. The QBO attains maximum amplitude of about 23 m s^{-1} on the Equator, between 30 and 10 hPa, diminishing in amplitude above and below this layer. The amplitude of the zonal wind QBO is approximately zero at the tropical tropopause ($\sim 17 \text{ km}$). The latitudinal profile of zonal wind amplitude is Gaussian-like with e-folding scale of $14\text{--}15^\circ$ (Figure 3). The phase of the zonal wind QBO is independent of latitude in the tropics, except that (1) the onset of westerly acceleration on the equator occurs about a month before its onset at adjacent off-equatorial latitudes, and (2) QBO extrema in the middle stratosphere are displaced a few degrees off the equator into the winter hemisphere. There is a pronounced asymmetry between easterly and westerly phases. The maximum value of westerly wind ($\sim 15 \text{ m s}^{-1}$) is about half that of the maximum easterly wind ($\sim 30 \text{ m s}^{-1}$). The descent of westerlies from 10 to 70 hPa is more rapid and uniform than the descent of easterlies. Easterly shears occasionally stall above 50 hPa; most stalls occur between July and February (e.g., during 1988 and 1989 in Figure 1). As

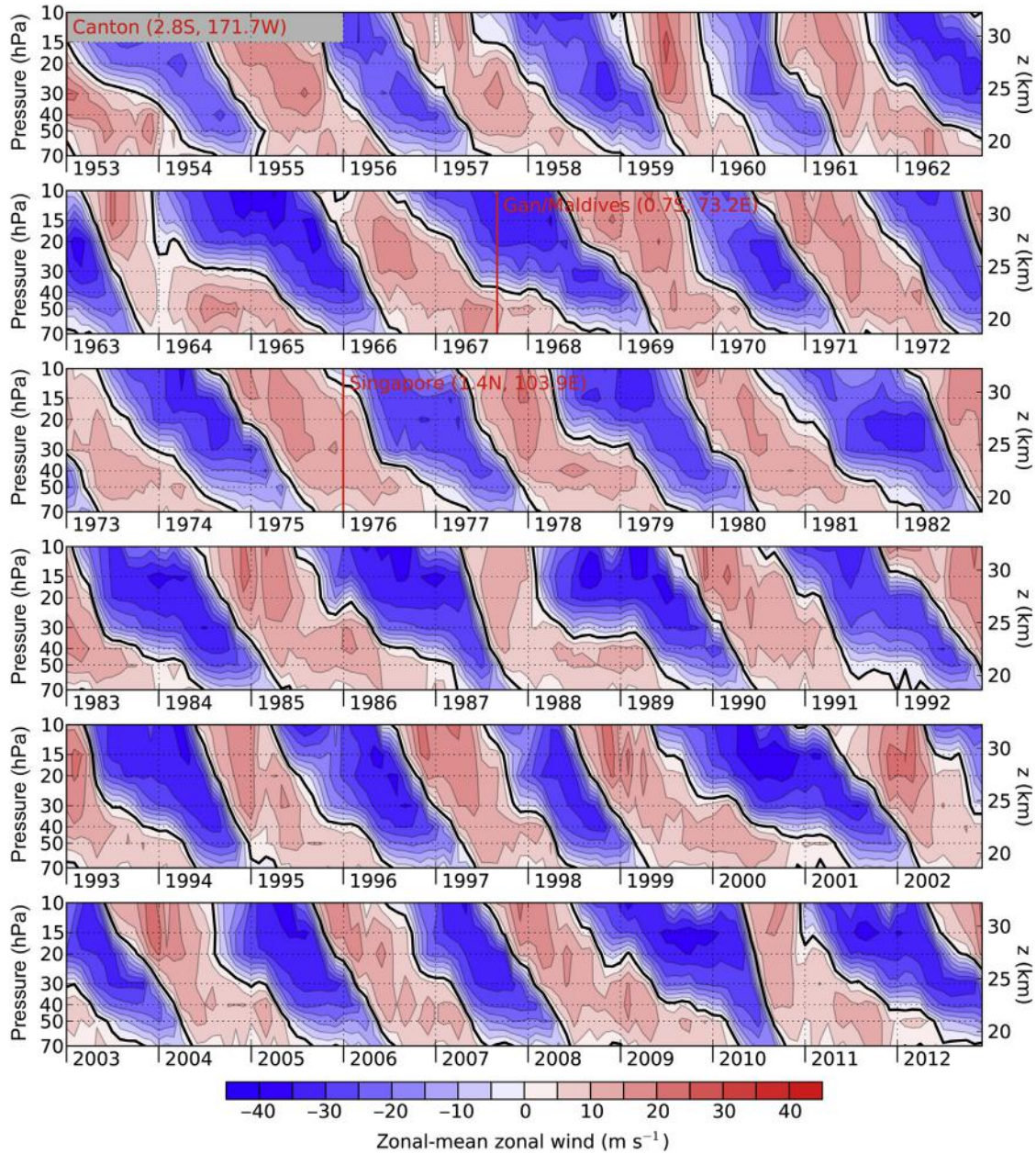


Figure 1 Time–height cross section of monthly mean zonal wind near the Equator, obtained from rawinsonde data at three locations. Data obtained from Freie Universität Berlin. Contour interval 5 m s^{-1} , with westerlies red and easterlies blue. Right vertical axis shows the approximate altitude. Red vertical lines and text indicate the rawinsonde stations from which each portion of the data is taken.

a result, QBO westerly (easterly) regimes are of longer duration than QBO easterly (westerly) regimes at 50 (10) hPa.

Easterly onsets at 50 hPa cluster in the northern late spring and early summer (Figure 2). The duration of easterly phases is typically about 12 months at this level, so westerly onsets also cluster at this time of year. Several onsets fall outside this broad time interval; hence, the QBO is modulated, but not synchronized, by the seasonal cycle. It is likely that if the seasonal cycle did not exist, the QBO would be completely asynchronous. Similarity between the average period of the QBO (~ 28 months) and integral or half-integral multiples of

the annual cycle (24 or 30 months) is coincidental. The period of the oscillation cannot be explained by the seasonal cycle.

Theory of the QBO

A theory of the QBO must account not only for its unusual period but also for equatorial westerlies in ‘superrotation,’ extending over a deep layer, without loss of amplitude from 10 to 50 hPa. A rotating atmosphere, at rest with respect to the surface, has maximum absolute angular momentum at the

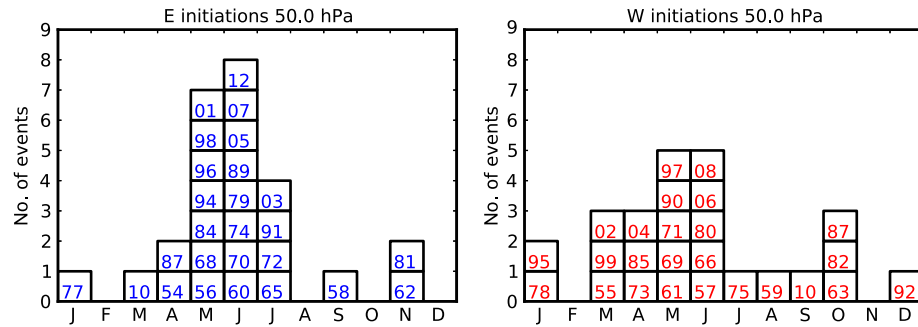


Figure 2 Seasonal distribution of QBO phase initiations at 50 hPa, defined as the first month of new easterly (E, left) or westerly (W, right) zonal wind phase. Numbers in the boxes indicate the year of the phase transition.

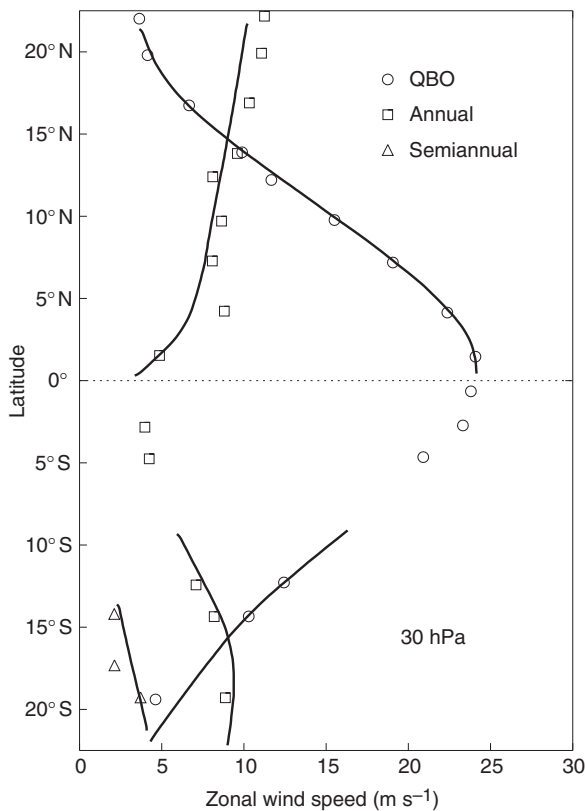


Figure 3 Latitudinal profiles of zonal wind amplitude for annual, semiannual, and QBO cycles at 30 hPa, derived from rawinsonde data. Symbols show results from individual stations, at a variety of longitudes. Solid curves were obtained from station data averaged in 2° latitude bins.

equator. Equatorial westerlies imply a larger value of absolute angular momentum, which cannot be obtained through advective rearrangement of angular momentum from a state at rest. Either superrotation must exist a priori, or momentum transport by zonally asymmetric motions is required. Equatorial westerlies exist in the SAO westerly phase, but it is unlikely that QBO westerlies are derived entirely (if at all) from SAO westerlies. It is theoretically possible for SAO westerlies to be advected downward at the equator by an anomalous diabatic circulation, but to conserve absolute angular momentum

(weighted by mass) the zone of westerlies would become much narrower than observed, and the rate of descent would be much slower. This statement, incidentally, is based on conservation of angular momentum and differs from a linear-diffusive model of zonal wind propagation described in early QBO literature. Of course, SAO westerlies also require explanation. It is thought that momentum transport by zonally asymmetric wave motions is responsible for descending easterly and westerly phases of the QBO, and for descending westerly phases of the stratopause SAO. These waves consist primarily of large-scale Kelvin, Rossby-gravity waves, inertia-gravity waves, and small-scale gravity waves.

Momentum transport by vertically propagating waves is one of two essential ingredients in a simple theory of the QBO. The other ingredient is a mechanism for switching between QBO phases at lower levels. As shown schematically in Figure 4, a spectrum of easterly and westerly waves, also propagating vertically, spontaneously create descending easterly and westerly shear zones. Momentum transported vertically by the waves is irreversibly deposited in the mean flow as the waves break or are dissipated, e.g., by radiative damping. Dissipating waves in general tend to accelerate the mean flow in the direction of their horizontal phase propagation. Given a small asymmetry in the wave spectrum propagating up from below (or a small nonzero mean flow, if the forced wave spectrum is symmetric about zero phase speed), waves of one sign will drive the mean flow in the lower part of the domain, leaving the upper part free to accelerate in the opposite direction, owing to waves with phase speed of opposite sign. There is a limiting value of mean flow that can be attained through this mechanism for any fixed spectrum of wave forcing. Thus, in order to conserve angular momentum, shear zones must descend with time. This process continues until (1) the lower shear zone cannot descend any farther, and (2) the upper shear zone descends to immediately above the lower jet. At this point, a switching mechanism is required for any further change in the mean flow. Most models assume that vertical diffusion of momentum provides such a switching mechanism, causing the lower jet to decay. Then, waves responsible for the lower jet are free to propagate vertically and create a new shear zone at upper levels. The entire process repeats, with shear zones reversed, until the lower jet (now of opposite sign) is once again destroyed by diffusion. Waves of opposite sign then propagate to upper levels, forming a new shear zone, and so on, resulting in a perpetual 'nonlinear oscillation.' The period of oscillation is determined by several

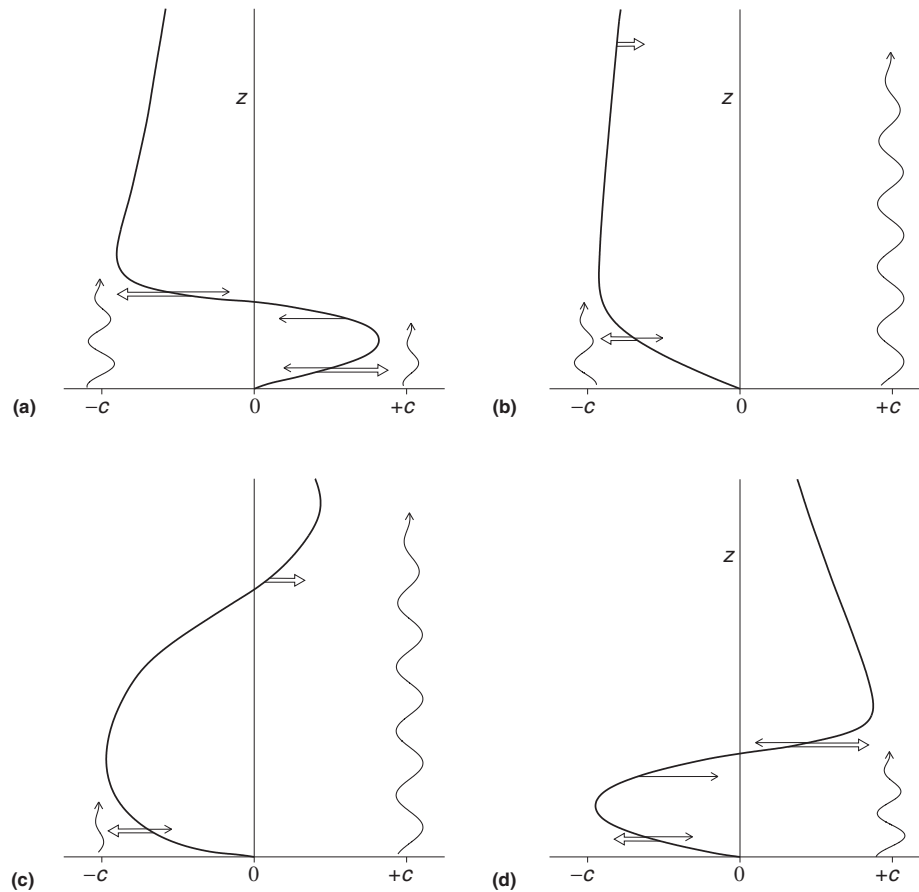


Figure 4 Conceptual model of the QBO, illustrating the vertical propagation and mean-flow acceleration due to eastward and westward waves (large arrows) and tendency due to vertical diffusion of momentum (small arrows). Panel labels (a) through to (d) indicate the temporal progression of the oscillation; see text for further explanation. After Plumb, R.A., 1984. The quasi-biennial oscillation. Holton, J.K., Matsuno, T., (Eds.), *Dynamics of the Middle Atmosphere*. Terra Scientific, Tokyo.

factors, including the total wave flux input at the lower boundary, the range of mean wind speeds attained by wave driving, and the amount of atmospheric mass affected by the oscillation. This model easily accounts for the QBO's asynchronous period, superrotation, downward propagation without loss of amplitude, and tendency for new phases to begin aloft as old phases decay at lower levels. It should be noted that the wave forcing is felt within each descending shear zone, so that the oscillation is, as it were, continually forced. The QBO, therefore, should not be regarded as a freely propagating, zonally symmetric wave. In this context, the 'lower boundary' of the QBO corresponds to the tropopause, with wave excitation occurring in the tropical troposphere.

It is likely that vertical mixing of momentum occurs intermittently or continuously within descending shear zones as a result of shear instabilities arising from the superposition of intense mean shear ($\sim 10 \text{ m s}^{-1} \text{ km}^{-1}$) and large-amplitude wave motions. Vertical shear observed in the QBO and SAO is the largest of any mean flow in the terrestrial atmosphere. The mean shear by itself, however, is insufficient for shear instabilities that require the local Richardson number to fall below $1/4$.

This conceptual model of the QBO is attributable to Richard Lindzen. In a 1968 paper coauthored with James Holton, the first numerical simulations of the QBO were described. At the

time, Lindzen's assertion regarding the role of waves did not have much observational support. The idea was intuitive, derived from (1) the known theoretical properties of gravity-wave momentum transport and critical-layer absorption (*see Dynamical Meteorology: Critical Layers*), (2) speculation that the necessary wave fluxes existed in the tropical atmosphere, and (3) the assumption that a suitable mechanism for phase switching could be found. Observational evidence of equatorial Rossby-gravity waves and Kelvin waves (*see Tropical Meteorology and Climate: Equatorial Waves*) began to accumulate in the later 1960s. It soon became evident that wave fluxes were quantitatively in the right ballpark, so there was little reason to doubt the Lindzen-Holton theory. This evidence was exploited by Holton and Lindzen to construct a modified theory in which the QBO is driven by two waves: a radiatively damped Kelvin and Rossby-gravity wave, rather than by critical-layer absorption of a broad spectrum of gravity waves. The basic QBO mechanism was unchanged.

Results of a laboratory experiment by Alan Plumb and Angus McEwan published in 1978 provided compelling evidence that an oscillation similar to that predicted by theory can occur in a weakly viscous nonrotating fluid as a result of momentum transport by vertically propagating internal gravity waves.

Today, it is thought that additional wave fluxes are required, beyond those of the Kelvin and Rossby-gravity waves, primarily because the entire oscillation occurs in a region of slow ascent due to the Brewer–Dobson circulation. The rates of mean ascent and QBO descent are comparable, but opposite in sign. A simple estimate suggests that about twice as much momentum flux is required, relative to the value without upwelling, to sustain a similar QBO. This additional flux is likely provided by a broad spectrum of gravity and inertia-gravity waves whose role in the QBO can be simulated reasonably well using the original Lindzen–Holton model. Recent observations of ascending water vapor anomalies in the equatorial lower stratosphere were used by the author to calibrate a simple QBO model, suggesting that the required total wave flux is two to four times as large as that of large-scale equatorial Kelvin and Rossby-gravity waves.

Considerably more work remains to be done to obtain a detailed picture of the wave spectrum. This problem is a challenging one because actual wave motions are spatially localized and episodic in time, occupying a wide range of spatial and temporal scales. Much of the relevant wave activity in the tropical stratosphere is attributable to deep moist convection in the tropical troposphere, which is irregular from day to day and also characterized by strong regional and seasonal variations.

Tropical Effects of the QBO

Accompanying the zonal wind QBOs are temperature anomalies (~ 4 K) in thermal wind balance with the zonal wind anomalies. Westerly (easterly) vertical shear is relatively warm (cold) and requires anomalous subsidence (upwelling) to maintain the temperature anomaly against radiative damping (Figure 5). The resulting circulation anomalies affect the dynamical QBO. Subsidence in westerly shear enhances the descent rate and vice versa. The QBO's mean meridional circulation may also limit the latitudinal extent of the oscillation. The example shown in Figure 5 was obtained from

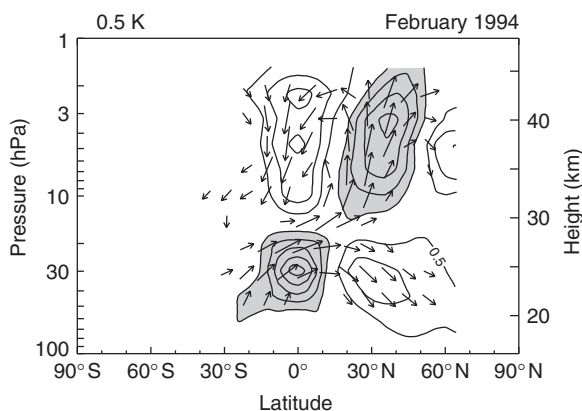


Figure 5 Temperature and mean meridional circulation anomalies associated with the QBO in February 1994, obtained from UK Met Office data. Contour interval 0.5 K, starting at this value, with values less than -0.5 K shaded. Adapted from a diagram by Bill Randel.

observations in late winter 1994. At solstice, the circulation pattern is asymmetric about the Equator, with a stronger winter cell. The QBO circulation at equinox is more symmetric.

For photochemically active trace species whose mixing ratio depends directly or indirectly on temperature, there is a QBO anomaly of mixing ratio sharing the same pattern and cross-equatorial symmetry as temperature. Ozone in the middle and upper stratosphere is photochemically controlled and provides a good example. Long-lived species such as methane, water vapor, and nitrous oxide, on the other hand, illustrate the effects of the QBO circulation anomaly. Ozone in the lower stratosphere, under dynamical control, can also be regarded as a long-lived tracer. Phase differences between QBO circulation and constituent anomalies depend on the rate of destruction by eddy mixing and photochemistry.

It has been known for several decades that column ozone is affected by the QBO. For this constituent, and for many other tracers, the relation to the dynamical QBO is somewhat difficult to understand, particularly at off-equatorial latitudes where the seasonal cycle and QBO overlap. Here, tracer anomalies are more synchronized with the seasonal cycle than is the dynamical QBO, and their temporal spectra contain additional frequencies (periods near 21 and $8\frac{1}{2}$ months) representing the nonlinear interaction of annual cycle and QBO. An example is shown in Figure 6, for methane in the middle stratosphere. A possible explanation of the ‘quasi-synchronized’ pattern of tracer anomalies in the subtropics is that the QBO’s influence is felt more strongly at a certain time of year (e.g., late winter or early spring) than at other times of year. The subtropical pattern cannot remain constant, because the relation between the QBO and annual cycle changes from one year to the next. Rather, there is a slow modulation of the subtropical pattern. The modulation period is equal to the time required for QBO onsets to march through a calendar year: e.g., 13 years, for a QBO period of 26 months. During a modulation cycle, the maximum amplitude of subtropical tracer anomaly may switch from one hemisphere to the other, and back again.

Temperature and circulation anomalies affect trace constituents directly, through photochemistry and advection. An indirect effect of the QBO is to modulate the lateral propagation and isentropic mixing due to planetary waves entering the tropics from the winter hemisphere. QBO easterlies ensure that a critical layer for quasi-stationary planetary waves is located in the winter hemisphere, well off the Equator, isolating the equatorial stratosphere from mixing due to planetary-wave breaking. QBO westerlies, on the other hand, allow quasi-stationary waves to penetrate the tropics and possibly break there, affecting the momentum balance of the QBO and the distribution of trace constituents. Modeling has shown, however, that equatorial mixing does not necessarily occur in the westerly phase. Breaking of small-amplitude waves occurs off the Equator, in both hemispheres, where the mean zonal wind is minimum. Large-amplitude waves have a wider critical layer, which may overlap the equator and cause mixing there. These two situations seem relevant to the QBO. At 50 hPa, where planetary waves are relatively small, QBO westerlies are long lived (12–18 months) without attenuation. At 10 hPa, where the waves are larger, QBO westerlies are short lived. This asymmetry could be due to the vertical propagation of an

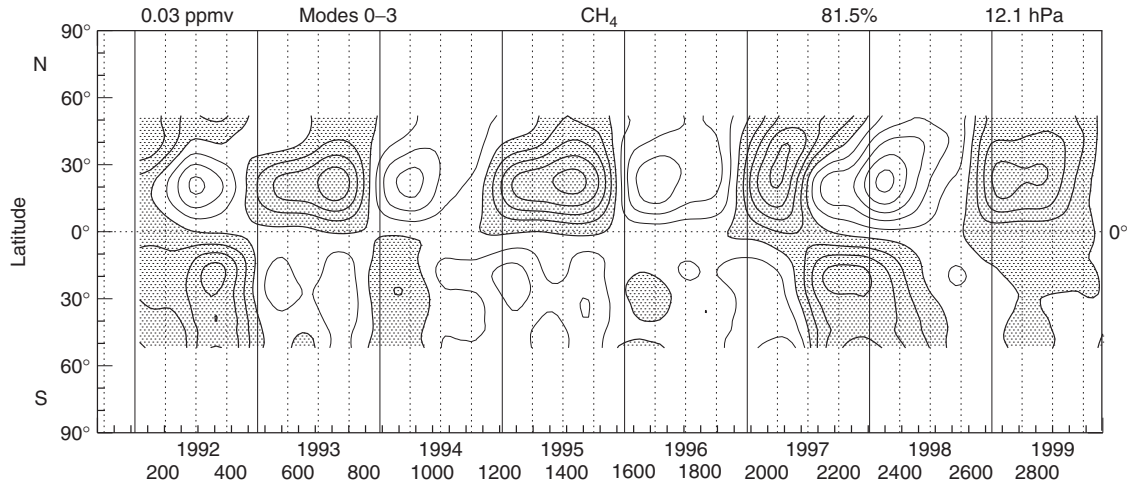


Figure 6 Time–latitude cross section of methane anomalies in the middle stratosphere, obtained from the Halogen Occultation Experiment instrument, reconstructed from the leading four Empirical Orthogonal Functions (EOFs) of deseasonalized data. Contour interval 0.03 ppmv, with negative values shaded.

asymmetric wave spectrum, or to the asymmetry of induced mean meridional circulation. There is evidence, however, that laterally propagating waves also determine the onset of QBO easterlies at 10 hPa.

The combined effects of mean circulation advection and isentropic mixing are illustrated in **Figure 7**, showing a latitude–height cross section of nitrous oxide in early 1992 obtained from the Cryogenic Limb Array Etalon Spectrometer instrument aboard the Upper Atmosphere Research Satellite. The upper part of the ‘staircase’ pattern of tracer isopleths is attributable to an asymmetric QBO circulation anomaly similar to that shown in **Figure 5**. QBO westerlies in the

middle and upper stratosphere at this time overlay easterlies in the lower stratosphere, and a subtropical westerly jet had formed near 30° N. Between this jet and the Equator was a narrow zone of mixing by planetary waves, visible in tracer maps, acting to homogenize the constituent field in this region. Similar isentropic mixing was evident in the midlatitude ‘surf zone’ at lower levels. The phase of the dynamical QBO was reversed 1 year later, along with the tracer anomalies, and there was little evidence of isentropic mixing in the tropical upper stratosphere.

Other effects of the QBO in the tropical atmosphere include a modulation of SAO westerly descent, with westerlies

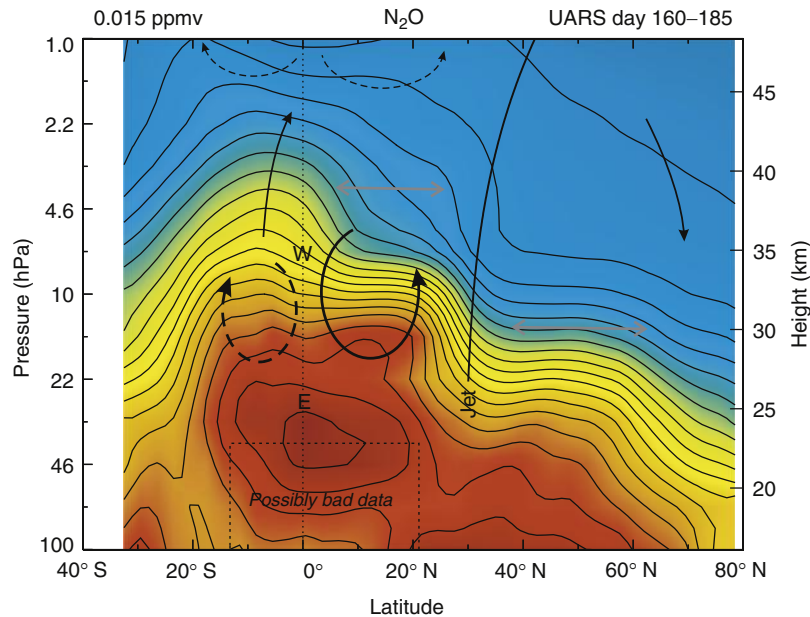


Figure 7 Latitude–height cross section of zonally averaged nitrous oxide in February and March 1992, obtained from Cryogenic Limb Array Etalon Spectrometer (CLAES) data. Contour interval 0.015 ppmv, starting with this value at upper right. The axis of the subtropical jet in late winter is indicated by the solid curve. UARS, Upper Atmosphere Research Satellite.

extending to progressively lower levels as the QBO westerly phase decays. Easterly phases of the mesopause SAO are stronger in the QBO deep westerly phase. The QBO is prevented from entering the tropical troposphere, owing to the Hadley circulation, but some effects of the QBO have been reported in this region also. The QBO may affect the latitudinal distribution of sea-level pressure, precipitation, and deep convection. Analysis of QBO-related variability in the tropical troposphere is complicated by unrelated biennial variability in this region. From the 1950s to 1980s, the QBO was believed to influence the occurrence of Atlantic hurricanes, but this relationship appears to have vanished as the observational record has lengthened.

Extratropical Effects of the QBO

The extratropical middle atmosphere undergoes a strong seasonal cycle, with a complete reversal of mean zonal wind and mean meridional circulation between summer and winter. Quasi-stationary planetary waves propagate vertically in mean westerlies that prevail during winter, and effects of the QBO on the extratropical middle atmosphere are observed primarily during this period (i.e., from late autumn to early spring). During January, mean zonal winds poleward of 45° N are stronger by about 10 m s^{-1} in the QBO 50 hPa westerly phase, relative to the easterly phase at this level (Figure 8(a)). A similar effect is seen in other winter months, although the signal tends to weaken in late winter (e.g., in March). This anomaly represents a strengthening of the stratospheric polar vortex (note the climatological wind contours shown in Figure 8(a)), and is accompanied by more intense planetary-wave activity and flux convergence at high latitudes in the 50 hPa easterly phase, which occur slightly earlier in winter (November through January). Stratospheric major warmings in the polar lower stratosphere, which occur in about half the

winters, are somewhat more likely in the easterly phase of the QBO. There is also some evidence that the QBO's influence on the polar lower stratosphere in late winter depends on the phase of the solar cycle. The 'normal' effect of the QBO described above is seen in years near solar minimum, while near solar maximum the effect is absent or perhaps reversed. There is some debate over the robustness of this solar modulation of QBO influence, but unlike in the aforementioned case of Atlantic hurricane activity, evidence of this effect is present in the observational record up to the present day. In addition to the solar cycle, there is also evidence that QBO influence on the polar vortex can be modulated by ENSO. Model studies focusing on the extratropical effect of the QBO unanimously agree with the observed tendency for more disturbed conditions and weaker mean zonal winds when the QBO in the lower tropical stratosphere (e.g., at 50 or 40 hPa) is in the easterly phase. The mechanism is not well understood, but may involve nonlinear reflection of planetary waves from a subtropical critical layer. The zero wind line is a critical surface for stationary planetary waves, and as noted in the previous section, it is displaced poleward in the easterly QBO phase, forming a narrower waveguide and resulting in more intense wave activity at high latitudes. The effect of the QBO is gradual, accumulating through a winter season, leading to a divergence of behavior between easterly and westerly phases of the QBO.

The models robustly suggest that the extratropical effect of the QBO is greatest in an intermediate range of planetary-wave forcings. At small amplitude, planetary waves have little effect on the circulation, and the QBO's influence is small. At large amplitude, planetary waves disrupt the polar vortex irrespective of the phase of the QBO. In Northern winter, planetary-wave amplitudes lie in an intermediate range, allowing the QBO's effect to be seen throughout much of the winter. In Southern winter, wave amplitudes are smaller, and the QBO's effect at high latitudes is delayed until early spring. At this time of year

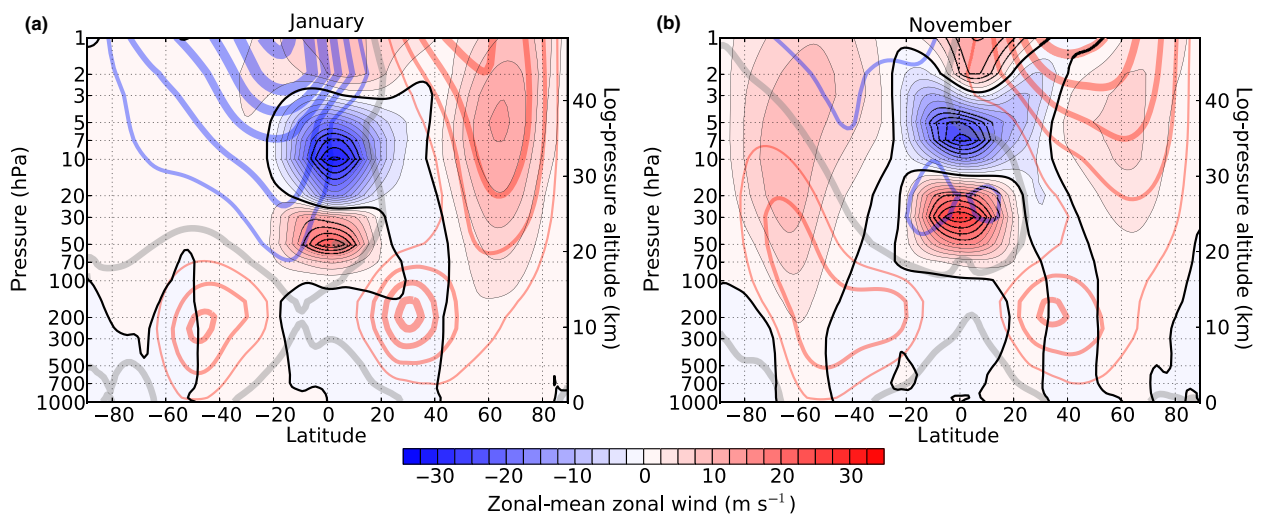


Figure 8 (a) Composite of mean zonal wind in January, showing the difference between westerly and easterly phases of the QBO at 50 hPa in January. (b) As in (a), but for November, using the QBO phase at 20 hPa in July. Contour interval 2.5 m s^{-1} , with westerlies red and easterlies blue. For reference the climatological zonal wind is shown in line contours (contour interval 10 m s^{-1}) using the same color scheme, with thicker lines for larger values, and the gray line indicating the zero wind contour separating stratospheric winter westerlies from summer easterlies.

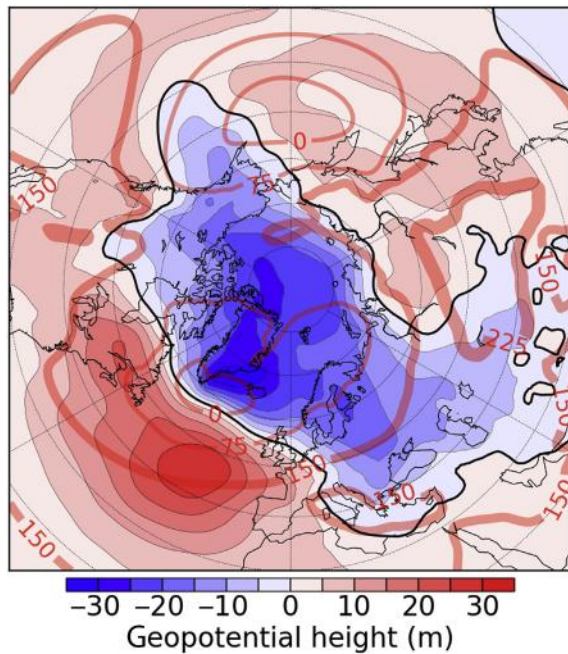


Figure 9 Extratropical surface signature of the QBO during northern hemisphere winter, shown by the January 1000 hPa geopotential height (Z1000) westerly minus easterly QBO composite difference, with QBO phase defined by 50 hPa QBO winds in January (as in [Figure 8\(a\)](#)). Contour interval 5 m, with positive anomalies red and negative anomalies blue. For reference the climatological Z1000 is shown in line contours (contour interval 75 m), with thicker lines for larger values.

the Antarctic ozone hole forms. The severity of ozone loss is affected by the QBO, with greater depletion in westerly QBO years when the air inside the polar vortex is colder and the vortex westerly winds are stronger. [Figure 8\(b\)](#) illustrates this effect for November, when the transition from winter westerlies to summer easterlies is in progress (as indicated by the climatological wind contours shown in [Figure 8\(b\)](#)). The equatorial QBO modulates a mode of variability in high northern latitudes known as the Arctic Oscillation or Northern Annular Mode (NAM). The NAM is closely related to polar vortex

warming events and propagates downward from the stratosphere to polar troposphere, implying a possible link between the QBO and high-latitude tropospheric circulation in winter. The surface manifestation of this link during January is shown in [Figure 9](#). The NAM pattern, which in its positive phase is characterized by low pressure at high latitudes surrounded by a ring of higher pressure at midlatitudes, is evident in [Figure 9](#). The signal is strongest in the Atlantic region, where it resembles the North Atlantic Oscillation (NAO) pattern. The NAO has a strong effect on weather conditions in this region, and its apparent link with the QBO may provide a useful source of NAO predictability on seasonal timescales.

See also: **Dynamical Meteorology:** Critical Layers; Kelvin Waves; Wave Mean-Flow Interaction; Waves. **Gravity Waves:** Buoyancy and Buoyancy Waves: Theory; Overview. **Middle Atmosphere:** Semiannual Oscillation. **Tropical Meteorology and Climate:** Equatorial Waves.

Further Reading

- Anstey, J.A., Shepherd, T.G., 2013. High-latitude influence of the quasi-biennial oscillation. *Quarterly Journal of the Royal Meteorological Society* 140 (678), 1–21. doi:10.1002/qj.2132.
- Andrews, D.G., Holton, J.R., Leovy, C.B., 1987. *Middle Atmosphere Dynamics*. Academic Press, Orlando, FL.
- Baldwin, M.P., Gray, L.J., Dunkerton, T.J., et al., 2001. The quasi-biennial oscillation. *Reviews in Geophysics* 39, 179–229.
- Dunkerton, T.J., 1997. The role of gravity waves in the quasi-biennial oscillation. *Journal of Geophysical Research* 102, 26053–26076.
- Gray, L.J., et al., 2010. Solar influences on climate. *Reviews in Geophysics* 48, 1–53.
- Holton, J.R., Lindzen, R.S., 1972. An updated theory for the quasi-biennial cycle of the tropical stratosphere. *Journal of Atmospheric Sciences* 29, 1076–1080.
- Lindzen, R.S., Holton, J.R., 1968. A theory of the quasi-biennial oscillation. *Journal of Atmospheric Sciences* 25, 1095–1107.
- Maruyama, T., 1997. The quasi-biennial oscillation (QBO) and equatorial waves—a historical review. *Papers in Meteorology and Geophysics* 48, 1–17.
- Plumb, R.A., 1984. The quasi-biennial oscillation. In: Holton, J.K., Matsuno, T. (Eds.), *Dynamics of the Middle Atmosphere*. Terra Scientific, Tokyo.
- Wallace, J.M., 1973. General circulation of the tropical lower stratosphere. *Reviews of Geophysics and Space Physics* 11, 191–222.

Semiannual Oscillation

K Hamilton, University of Hawaii, Honolulu, HI, USA

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

The dominant feature of the large-scale circulation near the equatorial stratopause is a strong semiannual oscillation in the prevailing winds. The stratopause zonal winds oscillate from strong easterlies around the time of the solstices to westerlies around the equinoxes. A similar oscillation, but exactly out of phase, is observed near the mesopause. The zonal wind oscillation is accompanied by similar semiannual variations in temperature and in the mean vertical motion.

Introduction

Regular observations of the winds and temperatures in the low-latitude stratosphere began only in the 1950s. Such observations were limited initially to operational balloon-borne radiosondes, which normally have a maximum ceiling of about 30 km (pressure of about 10 hPa). Once several years of such observations were available, it was clear that the variations in near-equatorial winds in the stratosphere below 30 km were dominated by a somewhat irregular oscillation with a mean period of roughly 28 months – what we now know as the quasibiennial oscillation (QBO).

Rocket observations of the winds extending to altitudes significantly higher than the balloon ceiling began in the mid-1960s at a small number of tropical stations. The black curves in [Figure 1](#) show the composite seasonal cycle of the zonal wind at several levels from the midstratosphere to the stratopause determined from many years of rocket observations at two near-equatorial stations. The result clearly displays the QBO in the

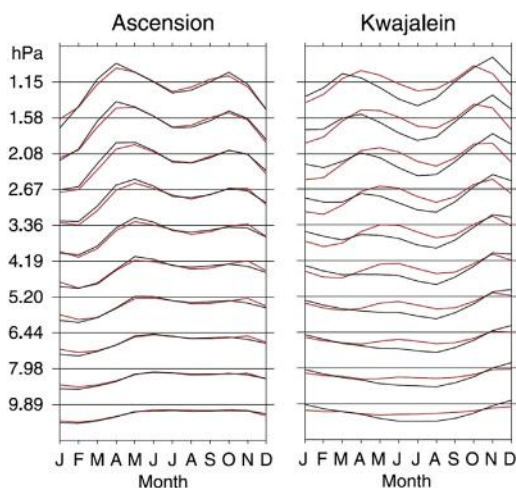


Figure 1 The long-term mean zonal wind each month of the year at different levels above Ascension Island (8S, 14W) and Kwajalein (9N, 168E). Black curves show observations from rocketsondes; red curves show observations from global reanalysis data. The vertical distance between the axes at successive levels represents 40 m s⁻¹. Reproduced from Badwin, M.P., Gray, L.J., 2005. *Geophysical Research Letters* 32, L09806. doi:10.1029/2004GL022328. Courtesy American Geophysical Union.

middle stratosphere, but also shows that the QBO becomes much less prominent above about 5 hPa (~35 km altitude), where it is replaced as the dominant signal by a strong semiannual oscillation (SAO) with peak-to-peak amplitude exceeding 50 m s⁻¹ near 1 hPa (~50 km altitude). Near the stratopause, the phase of the SAO is such that maximum easterly anomalies occur shortly after the solstices in January and July, while the maximum westerly anomalies occur around April and October.

Observations

As noted in this article, the discovery of the SAO was based on *in situ* observations from sondes launched by rockets. In recent years, these direct observations of the wind have been supplemented by gridded global reanalysis data that use all the available observations, including satellite determinations of temperature together with a sophisticated simulation model of the atmosphere. The red curves in [Figure 1](#) show the seasonal composite from many years of the reanalysis data produced by the European Centre for Medium Range Weather Forecasts (ECMWF). The ECMWF reanalysis also shows the SAO in the wind and compares reasonably well with direct observations. [Figure 2](#) is a height–time section of the seasonal composite of the equatorial zonal–mean zonal wind from many years of the ECMWF global analyses plotted from the ground up to 0.1 hPa (~65 km altitude). The phase and amplitude of the SAO are seen to vary systematically with height. Notably, there is a downward propagation of the wind reversals reminiscent of the behavior of the QBO. There is also some annual (12-month) variation evident in the wind results in [Figure 2](#). Near the stratopause, the SAO cycle in the first half of the calendar year is significantly stronger than that in the second half of the year.

Data from all available rocketsonde stations have been used to produce the plot of the amplitude of the semiannual harmonic of zonal wind, as a function of height and latitude shown here as [Figure 3](#). The dominant meridional modulation is a falloff in SAO amplitude away from the equatorial region, although there is an indication of a somewhat larger amplitude on the southern side of the equator.

The *in situ* observations of the wind in the upper stratosphere and lower mesosphere are limited to the rocket soundings, which are available from only a small number of stations, and the useful wind data are very sparse at low latitudes. The SAO is also seen in temperature observations, and,

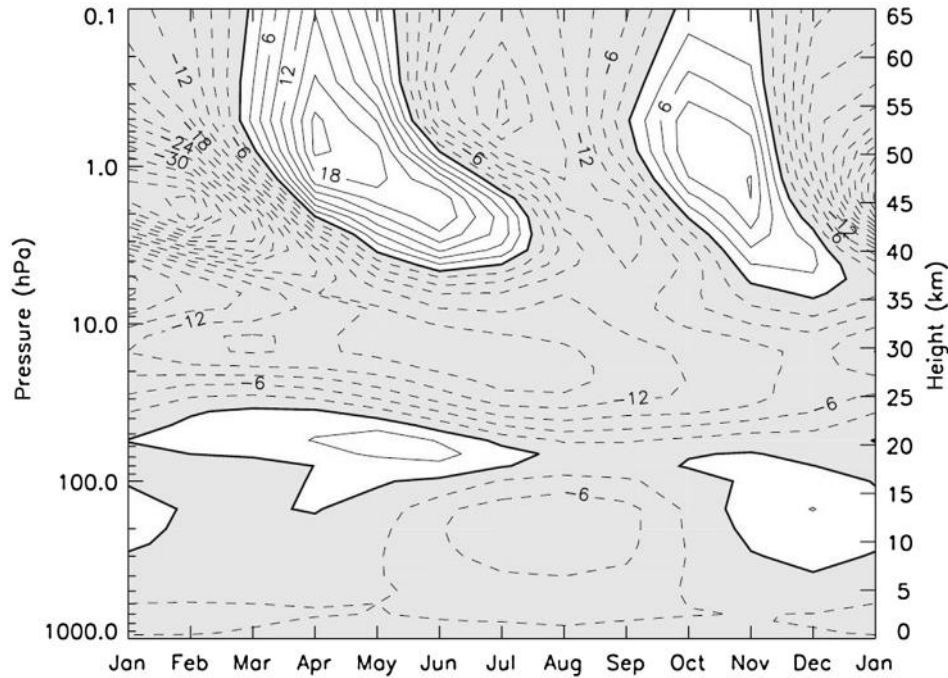


Figure 2 Height–time section of the composite seasonal cycle of the equatorial zonal-mean zonal wind from reanalysis data. The contour interval is 3 m s^{-1} , and shading denotes easterly mean winds. Reproduced from Pascoe, C.L., Gray, L.J., Crooks, S.A., Jukes, M.N., Baldwin, M.P., 2005. *Journal of Geophysical Research* 110, D08105. Courtesy American Geophysical Union.

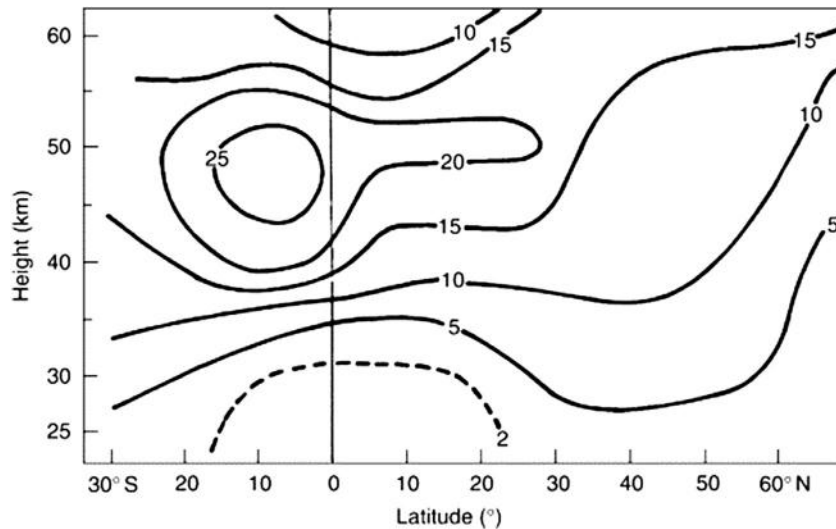


Figure 3 Height–latitude section of the amplitude of the semiannual harmonic of the zonal wind determined from long records of rocketsonde observations at 10 stations between 31° S and 64° N . The solid contours have intervals of 5 m s^{-1} . Reproduced from Hopkins, R., 1975. *Journal of Atmospheric Sciences* 32, 712–719. Courtesy American Meteorological Society.

in particular, the stratopause SAO has been observed in remote-sensing radiometer measurements of temperature from several different satellites. The equatorially trapped and roughly zonally symmetric aspects of the SAO near the stratopause appear clearly in these satellite temperature observations, as does the second equatorial maximum near the mesopause. The zonal-mean zonal wind and temperature are expected to be in approximate thermal wind balance (see **Dynamical Meteorology: Overview**), implying that, when there is westerly shear

on the equator, the mean temperature should drop off with latitude to both the north and south. Thus, westerly (easterly) shear regions in the SAO should correspond to anomalously warm (cold) temperatures centered on the equator. The available observations of the wind and temperature SAO do appear to be consistent with this expectation. As an example, **Figure 4** shows the equatorial zonal-mean temperature at several levels near the stratopause determined for each month of the year by averaging available remotely sensed satellite observations over

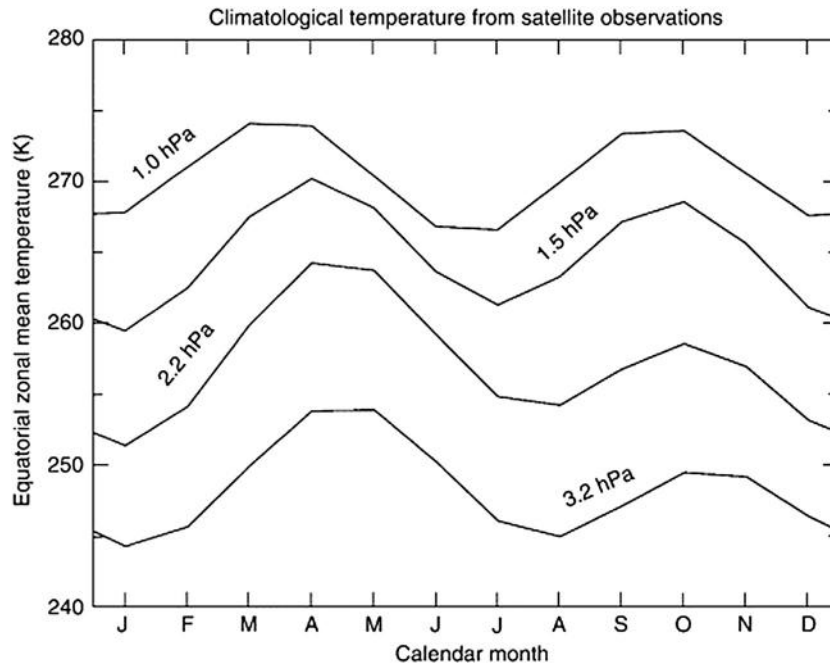


Figure 4 Climatological seasonal march of zonal-mean equatorial temperature at several levels in the upper stratosphere: 1 hPa (~ 50 km), 1.5 hPa (~ 48 km), 2.2 hPa (~ 45 km), and 3.2 hPa (~ 42 km).

several years. The seasonal march of these temperatures at each level is dominated by the SAO, and the downward phase propagation is evident.

In the upper mesosphere and lowermost thermosphere, the available *in situ* rocket data are extremely limited, although there are enough wind measurements to suggest that a second maximum in SAO amplitude occurs near the mesopause. Fortunately, at these heights, satellite-based remote-sensing observations of the horizontal wind (based on measurements of the Doppler shifts of atmospheric line spectral emissions) are available. **Figure 5** shows the zonal-mean zonal wind at the

equator as a function of time of year and height between 65 and 110 km, determined from several years of these satellite measurements. The strong SAO in the 75–90 km altitude range is easily seen. A similar pattern has been observed in low-latitude ground-based radar observations of the wind at these heights.

Dynamics of the SAO

The fact that a strong SAO in atmospheric circulation should occur near the equator is not, in itself, surprising. The sun

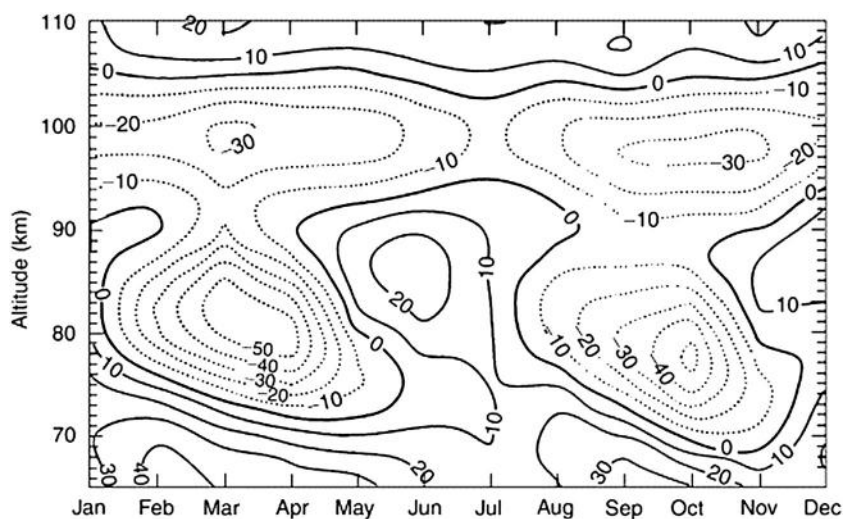


Figure 5 Composite seasonal cycle of the zonal-mean zonal wind at the equator based on several years of data from satellite remote sensing. The contour interval is 10 m s^{-1} , and dashed contours indicate easterly winds. Reproduced from Garcia, R., Dunkerton, T.J., Libermann, R.S., Vincent, R.A., 1997. *Journal of Geophysical Research* 102, 26019–26032. Courtesy American Geophysical Union.

crosses the equator twice each year, and the seasonal cycle of the daily-mean solar radiation incident on the atmosphere is dominated by a semiannual harmonic at low latitudes. However, it has been a significant challenge to account for all the observed features of the SAO. The occurrence of easterly mean wind extremes around the solstices at the stratopause level can be accounted for in terms of the annual modulation of the large-scale circulation in the meridional plane. In particular, at stratopause levels, the large-scale flow is thought to proceed from the summer hemisphere (where air is rising in the extratropical stratosphere) into the winter hemisphere (where air sinks in the extratropics) (*see Middle Atmosphere: Transport Circulation*). Consistent with this flow is a large-scale structure of the zonal wind field with strong easterlies in the summer hemisphere and westerlies in the winter hemisphere (*see Middle Atmosphere: Zonal Mean Climatology*). The effect of the cross-equatorial flow is to advect the summer easterlies onto the equator, and this effect should peak twice a year, roughly around or just after the solstices (when the cross-equatorial flow is strongest).

The westerly mean flow accelerations in the stratopause SAO are thought to arise largely from the interaction of the mean flow with vertically propagating internal gravity waves or large-scale equatorial waves generated in the lower atmosphere (*see Gravity Waves: Overview; Tropical Meteorology and Climate: Equatorial Waves*). The mutual effects of (1) the mean flow on wave propagation and (2) wave momentum transport on the mean flow lead to a pattern of downward-propagating mean jets (*see Dynamical Meteorology: Wave Mean-Flow Interaction; Middle Atmosphere: Quasi-Biennial Oscillation*). In the lower stratosphere, this is observed in both phases of the QBO, and the downward propagation is particularly characteristic of the westerly acceleration phase of the stratopause SAO (*Figure 2*).

Simple numerical models of the mean flow in the low-latitude middle atmosphere that have incorporated representations of these two mechanisms (a seasonally independent source of gravity waves or equatorial waves to account for the westerly acceleration, and the semiannually modulated effects of cross-equatorial momentum advection) have been able to simulate realistic stratopause SAO behavior. Also, these two mechanisms have been shown to operate in some comprehensive general circulation models (*see Numerical Models: General Circulation Models*) that extend into the middle atmosphere and produce somewhat realistic SAO simulations. However, there still remain some significant uncertainties concerning the stratopause SAO dynamics, including such issues as the relative role of higher frequency gravity waves versus lower frequency, large-scale equatorial waves, and the possibility of significant effects on the equatorial mean flow from stationary planetary waves (*see Dynamical Meteorology: Stationary Waves (Orographic and Thermally Forced)*) forced in the extratropics.

The uncertainties concerning the dynamics of the strong SAO near the mesopause are even greater. One plausible explanation for the mesopause SAO is that the mean wind variations of the stratopause SAO act as a filter for a broad spectrum of eastward- and westward-propagating gravity waves excited in the lower atmosphere. When there are strong westerlies at the stratopause level, the upward propagation of waves with westward phase speeds into the mesosphere is favored. When these waves are absorbed at higher levels, they will produce easterly mean flow

accelerations (*see Dynamical Meteorology: Wave Mean-Flow Interaction; Middle Atmosphere: Quasi-Biennial Oscillation*). Thus, this filtering effect leads naturally to a mean flow oscillation at higher levels that is 180° out of phase with the stratopause SAO. Once again, this mechanism has been included in simple numerical models that have produced a reasonably realistic SAO in the simulated mesopause mean flow. However, observations of the relevant wave-mean flow interactions in the real atmosphere sufficiently detailed to quantitatively confirm this mechanism are not available.

Effects on Trace Constituents

The SAO in zonal wind and temperature should be accompanied by an SAO in mean circulation in the meridional plane. The SAO variations in zonal-mean vertical velocity are much too small to be measured directly, but the accumulated effects can be seen in the distribution of the concentration of long-lived trace constituents in the stratosphere and mesosphere. As noted in this article, the occurrence of westerly shear on the equator is accompanied by anomalously warm temperatures at low latitudes. Associated with these warm temperatures will be anomalously high radiative cooling rates, and either sinking of air or at least anomalously weak rising motion near the equator.

The effects of the SAO appear clearly in observations of the zonal-mean concentration of long-lived trace constituents such as nitrous oxide (N_2O) and methane (CH_4) in the stratosphere. Both N_2O and CH_4 have chemical sources in the troposphere and are lofted into the stratosphere in the mean tropical upwelling. These constituents are subject to chemical destruction in the stratosphere, and this leads to a vertical stratification in the concentration (*see Stratospheric Chemistry Topics: Overview*). The SAO of rising and depressing the mixing ratio distribution, leading to a semiannual cycle of trace constituent concentration near the equator.

See also: *Gravity Waves: Overview. Middle Atmosphere: Quasi-Biennial Oscillation; Zonal Mean Climatology. Tropical Meteorology and Climate: Equatorial Waves.*

Further Reading

- Andrews, D.G., Holton, J.R., Leovy, C.B., 1987. *Middle Atmosphere Dynamics*. Academic Press, New York.
- Baldwin, M.P., Gray, L.J., 2005. Tropical stratospheric zonal winds in ECMWF ERA-40 reanalysis, rocketsonde data, and rawinsonde data. *Geophysical Research Letters* 32, L09806. <http://dx.doi.org/10.1029/2004GL022328>.
- Garcia, R.R., Dunkerton, T.J., Lieberman, R.S., Vincent, R.A., 1997. Climatology of the semiannual oscillation of the tropical middle atmosphere. *Journal of Geophysical Research* 102, 26019–26032.
- Hamilton, K., 1998. Dynamics of the tropical middle atmosphere: a tutorial review. *Atmosphere – Ocean* 36, 319–354.
- Hirota, I., 1980. Observational evidence of the semiannual oscillation in the tropical middle atmosphere: a review. *Pure Applied Geophysics* 118, 217–238.
- Pascoe, C.L., Gray, L.J., Crooks, S.A., Jukes, M.N., Baldwin, M.P., 2005. The quasi-biennial oscillation: analysis using ERA-40 data. *Journal of Geophysical Research* 110, D08105. <http://dx.doi.org/10.1029/2004JD004941>.

Stratospheric Sudden Warmings

A O'Neill and AJ Charlton-Perez, University of Reading, Earley Gate, Reading, UK
LM Polvani, Columbia University, New York, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by A O'Neill, volume 3, pp 1342–1353, © 2003, Elsevier Ltd.

Synopsis

This article reviews the history, dynamics, and impacts of stratospheric sudden warmings. Sudden warmings are the most dramatic meteorological phenomenon in the stratosphere and affect the distribution of trace constituents in the stratospheric and tropospheric weather.

Introduction

The stratospheric sudden warming (SSW) is the most dramatic meteorological phenomenon to take place in the stratosphere. This article will describe its main features and will give a brief history of its discovery. Examples of the various kinds of SSW will be presented and discussed. There is then a review of some of the theoretical ideas that have been advanced to explain the origin and evolution of the SSW. This includes a discussion of what triggers the SSW and what factors might affect the frequency at which it occurs. Finally, the effects of SSWs are considered, first on the distribution of trace constituents in the stratosphere, and secondly on the weather.

The Basic Facts

SSWs occur in the stratosphere of the Northern Hemisphere during winter. Temperatures rise sharply, by as much as 80 °C or more in a few days. Figure 1 shows temperature changes recorded at a high-latitude location during a strong sudden warming. Accompanying these changes, the stratopause descends over some locations by as much as 20 km, as shown in Figure 2.

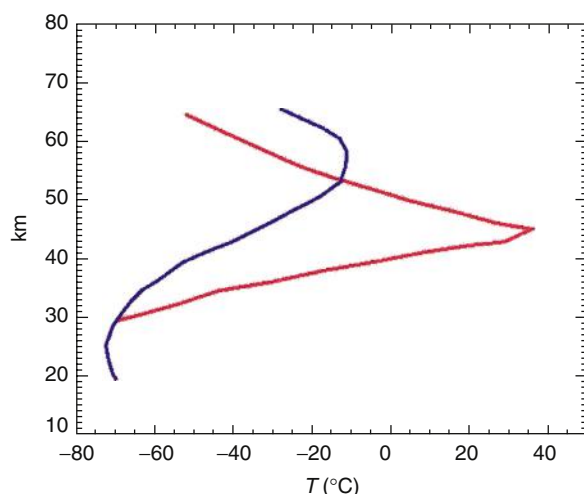


Figure 1 Typical rocketsonde observations above West Geirinish (57° N, 7° W) before (blue) and at the height of (red) a major stratospheric warming.

Before the onset of the warming, the stratospheric circulation is dominated by a cold and strong westerly polar vortex, which lies over the North Pole, covering most of the Northern Hemisphere outside the tropics (Figure 3). During a so-called major warming, this polar vortex is almost entirely broken down in a matter of days. The stratospheric circulation undergoes a dramatic change. Westerly winds are replaced by easterly winds throughout much of the stratosphere at high latitudes. The occurrence of a major warming can be detected in the stratosphere over the whole globe.

Major warmings do not occur every winter. Typically, six major warming events are observed in a decade, but there is significant variability between different decades. Over the short observational record in the stratosphere, there have been extended periods with no major warmings (1992–98) and periods with major warmings almost every year (the 2000s). So-called minor warmings occur almost every winter. Although

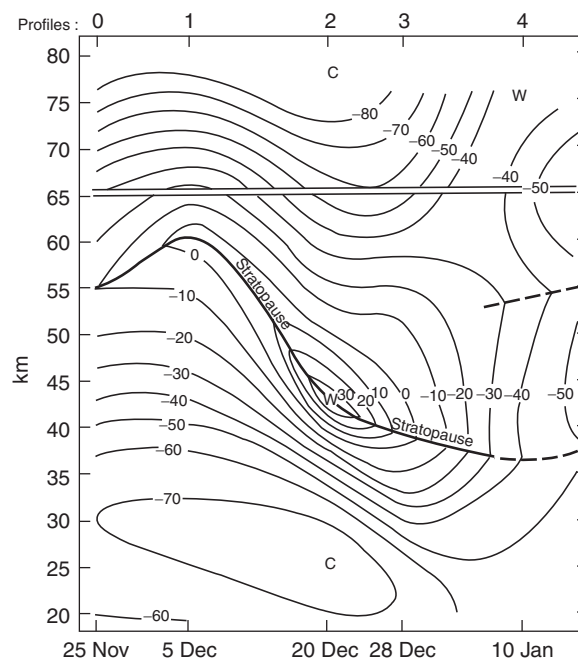


Figure 2 Schematic evolution of temperature at about 60° N during the development of a major midwinter warming. Notice that the height of maximum temperature defining the stratopause descends with time.

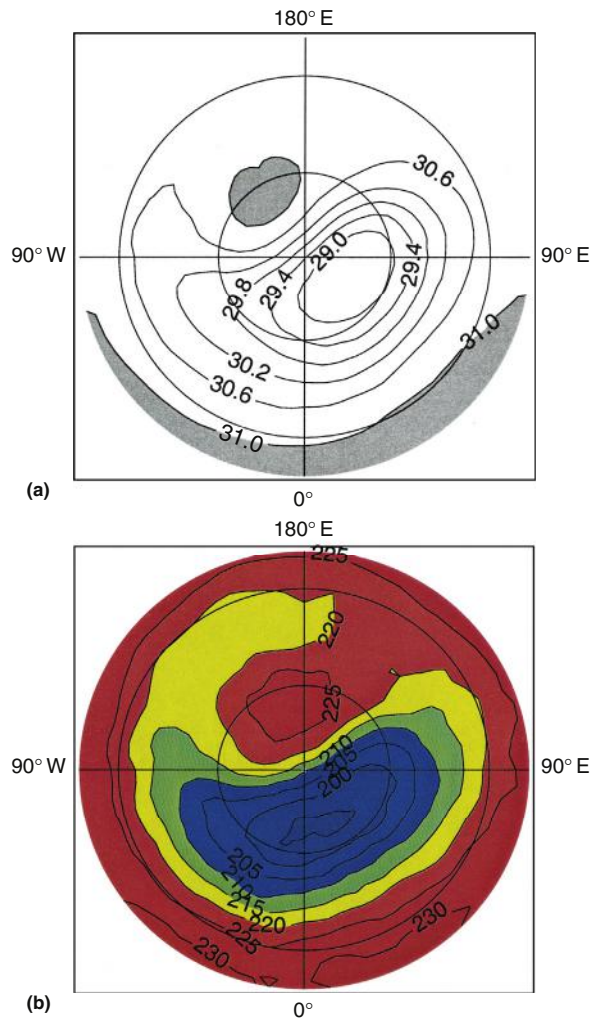


Figure 3 Typical fields of geopotential height (a) and temperature (b) for the winter stratosphere of the Northern Hemisphere at 10 hPa, near 30 km, on 1 December 1981. The polar vortex corresponds to low values of geopotential height and temperature. Winds circulate around the polar vortex in a counterclockwise direction on the figure. The Aleutian High is marked by the shaded patch near the dateline at 180° E. Winds circulate around it in a clockwise direction on the figure. Latitude circles are marked at 30° N and 60° N, and the edges of the panels are tangential to the 20° N line of latitude. Units: km for geopotential height; K for temperature.

there are large increases in temperature during such events, the polar vortex is not broken down.

Major warmings do not occur in the stratosphere of the Southern Hemisphere (though an exceptional event, resembling wave-2 major warmings in the Northern Hemisphere, occurred in September 2002). The polar vortex is much stronger than that of the Northern Hemisphere and less readily disrupted. Warmings do, however, occur in spring when the vortex is in the process of weakening during the springtime transition from westerly winds in winter to easterly winds in summer. They are called ‘final warmings,’ because the polar vortex does not recover until the following winter.

SSWs lead to a net poleward and downward transport of stratospheric air. They also bring about a vigorous mixing of air

of polar origin with air of subtropical (or tropical) origin, thereby weakening any preexisting, latitudinal gradients in the concentrations of chemical constituents.

SSWs are generated by the upward propagation of planetary-scale disturbances from the underlying troposphere. The precise nature of the links between stratosphere and troposphere during SSWs has been difficult to establish, because SSWs are highly nonlinear dynamical events.

A Brief History

In the early 1950s, knowledge of the stratosphere was still woefully inadequate. Richard Scherhag, at the newly established Free University of West Berlin, had used radiosondes to construct monthly mean maps of the stratosphere at a height of 22 km. He realized that a new American radiosonde, which remedied some shortcomings with earlier models, would allow reliable measurements of temperature to be made up to a height of 40 km or more. Starting in January 1951, he arranged for these radiosondes to be launched regularly from Tempelhof Airport in Berlin. He witnessed the first ‘explosive warming’ on 27 January 1952; it came as a complete surprise. He noted that temperatures rose to values that were not reached even at the height of summer.

Scherhag set up a group of stratospheric meteorologists at the Free University of Berlin. For decades, by using data from radiosondes and rocketsondes, this group assiduously constructed maps for the Northern Hemisphere of geopotential height and temperature at selected pressure levels in the stratosphere up to 10 hPa. (The paucity of sondes meant that such maps could not be constructed in the Southern Hemisphere.) Thanks to the efforts of this group, the nature, extent, and evolution of SSWs were revealed. Today, the radiosonde network is much improved, and satellites give continuous global coverage of the state of the atmosphere, including the stratosphere. Yet the important discovery of SSWs resulted from a single measurement made in the right circumstances.

Classification and Description of SSW

A typical state of the stratosphere of the Northern Hemisphere during winter was presented in Figure 3. Two vortices are present: the cold polar vortex, which dominates the circulation, and a persistent, but weaker, anticyclone near the dateline, referred to as the Aleutian High. The intensification of the Aleutian High and the accompanying weakening of the polar vortex, almost invariably play a major part in the evolution of SSWs.

The four following kinds of SSW have been identified: major midwinter warmings, minor warmings, Canadian warmings, and final warmings.

Major Midwinter Warmings

Major midwinter warmings occur mostly in December, January, and February. In addition to a warming of the north polar region and the reversal of the normal wintertime temperature gradient (cold pole; warm tropics), they lead to a breakdown of the polar westerly vortex, which is replaced by

an anticyclone (usually the Aleutian High). The World Meteorological Organization classifies warmings as major when the usual westerly winds of the Arctic at 10 hPa are replaced by easterlies as far south as 60° N.

The polar vortex is either displaced entirely from the pole or splits in two. These warmings are therefore often referred to respectively as ‘wave-1’ warmings and ‘wave-2’ warmings because these zonal wave numbers tend to dominate the evolution of the geopotential height field in each case. It can, however, be more useful and instructive to describe the events in terms of how the vortex evolves during the event, namely vortex displacement and vortex splitting events. Some major warmings exhibit a hybrid character, with the polar vortex being displaced asymmetrically from the pole and then split.

Wave-1/vortex displacement type

Figure 4 shows a classical example of a wave-1 type SSW, which occurred in northern winter 1986/87. The Aleutian High (near the dateline) has intensified, displacing the polar vortex and the associated cold air from the pole. Such intensification is often the result of merger between the quasi-stationary Aleutian High and an eastward-traveling anticyclone that forms to the west near 0° E and is then advected eastward around the polar vortex. Temperatures rise sharply in the strengthening jet stream between the polar vortex and the Aleutian High. This temperature rise is the result of adiabatic compression of the air as it descends (by up to about 2 km) on entering the jet stream. The polar vortex is characterized by high values of potential vorticity. It weakens rapidly as air with high values of potential vorticity is drawn away into the anticyclonic circulation (see Figure 5, described later, for an illustration of the process). As the warming subsides, the polar vortex recovers by radiative cooling of the polar air.

Figure 6 shows the accompanying evolution at 10 hPa of zonal-mean winds and temperatures (quantities averaged around latitude circles). A sharp rise in polar temperatures is evident in December, leading to a reversal of the north–south temperature gradient, and accompanied by a rapid deceleration of zonal-mean winds. The winds reverse from westerly to easterly as far south as 60° N, so the event would be classified as ‘major.’

Wave-2/vortex splitting type

The most dramatic major warmings involve a complete split in the polar vortex, followed by a rapid breakdown of one or both of the two cyclonic vortices that result from this split. The growth of the Aleutian High is accompanied by the development of a second anticyclone in the vicinity of the Greenwich Meridian at 0° E. An unusually symmetrical example of a ‘wave-2’ major warming occurred in northern winter 1984/85. Its evolution is illustrated in Figure 7. There are two developing anticyclones: the Aleutian High near 180° E and another (nonclimatological) anticyclone near 0° E. The polar vortex was split in the ‘pincer’ formed by these anticyclones, which then merged over the pole, as shown in Figure 8, bringing warm air over the polar cap. Subsequently, both of the cyclones weakened rapidly as they were stretched out around the strong anticyclone over the pole. Often, vortex splitting events are preceded by a ‘preconditioning’ of the vortex in which it is displaced from the pole and elongated. This

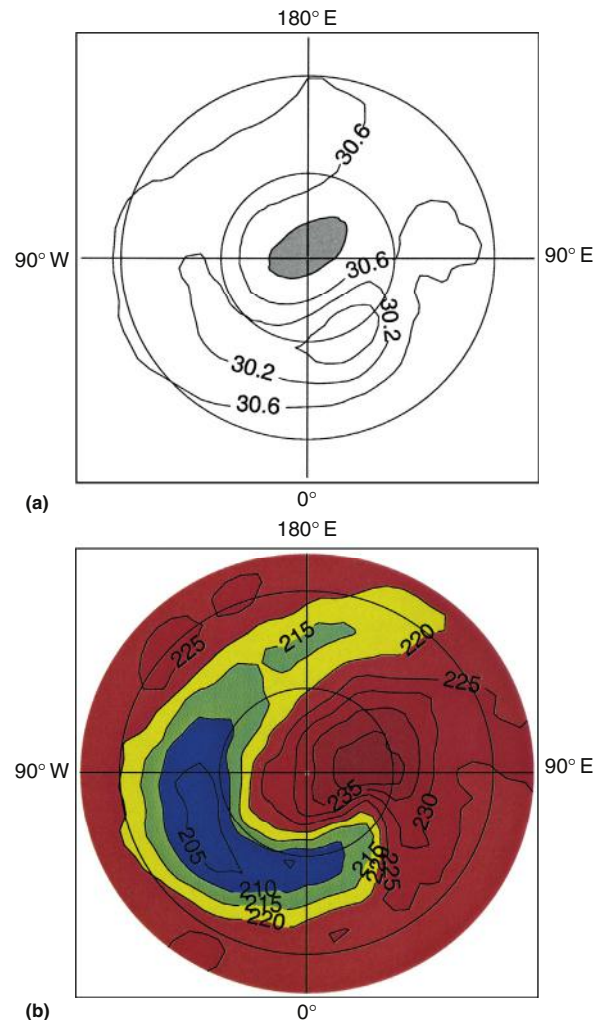


Figure 4 Fields of geopotential height (a) and temperature (b) for the Northern Hemisphere at 10 hPa on 10 December 1987 at the height of a major warming of the wave-1 type. The Aleutian High (gray shading) has moved over the pole; the polar vortex is displaced from the pole and is much weaker than normal. Temperatures over the polar cap are now higher than at lower latitudes, the reverse of normal conditions in the winter stratosphere. Units: km for geopotential height; K for temperature.

preconditioning has a strong signature in the wave number one geopotential height field meaning that separating vortex displacement and vortex splitting events purely on the basis of the amplitudes of the wave number one and two geopotential height field can be difficult. Instead, alternative methods, which focus on examining the two- and three-dimensional structure of the potential vorticity fields have proved a useful complement to traditional methods in classifying and understanding SSWs in recent years.

Minor Warmings

Minor warmings occur every winter in the stratosphere of both hemispheres. They share many of the characteristics of ‘wave-1’ major warming, except that the polar vortex is not broken

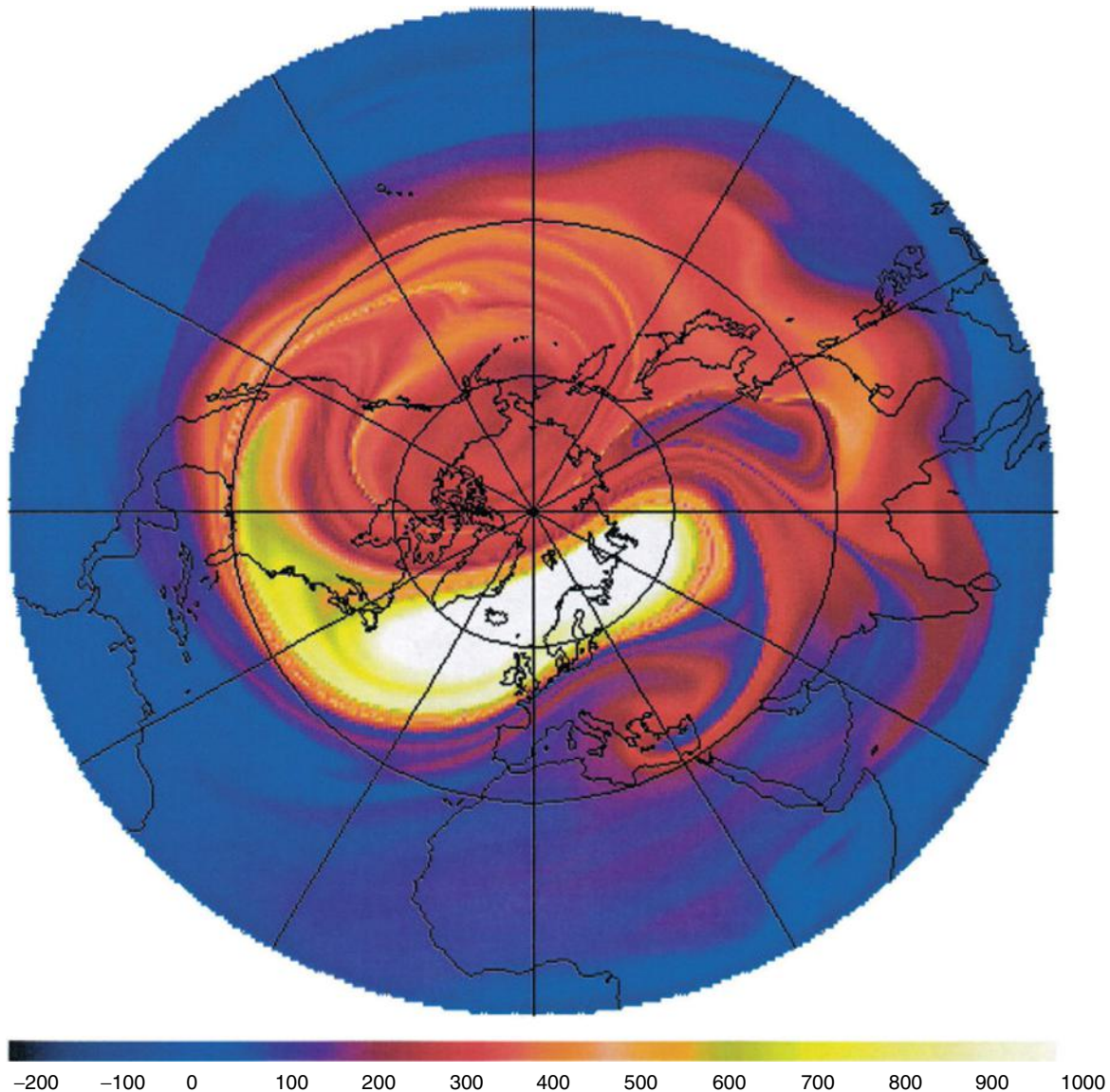


Figure 5 This figure illustrates the strong, quasi-horizontal mixing of air that occurs in the stratosphere during an SSW. It was derived by using a high-resolution Lagrangian method in which the movement of many thousands of air parcels was computed. Streamers of air (delineated by blue shading) are being drawn from low latitudes into the Aleutian High, the red area at the top of the figure containing yellow swirls. Air from the polar vortex, shaded yellow, is also being drawn into the Aleutian High where it mixes with the air from low latitudes. Through this loss to the Aleutian High of air that has high values of potential vorticity, the polar vortex weakens during an SSW.

down and the wind reversal from westerly to easterly is less extensive. Major warmings are often preceded by a series of minor warmings, which may serve to weaken the polar vortex steadily after midwinter, making it more susceptible to a complete breakdown.

One particular type of minor warming is the so-called, Canadian warming. These events occur in early winter in the stratosphere of the Northern Hemisphere, typically from mid-November to early December. (They have no counterpart in the Southern Hemisphere.) They show a special kind of structure and evolution, illustrated in Figure 9. The warm Aleutian High is advected eastward in a few days from its normal position over the dateline toward the 90° W line of longitude

over Canada. The polar vortex is displaced from the pole and strongly distorted, but does not break down. Temperature changes are modest compared with major warmings, and affect mainly the middle and lower stratosphere. As the Aleutian High collapses, the vortex regains its usual polar position.

Final Warmings

Springtime final warmings mark the transition in the stratosphere between westerly winds in winter and easterly winds in summer. This transition does not occur smoothly, as it would were it driven purely by radiative heating, but in a sporadic manner owing to dynamical variability of the vortex. A final

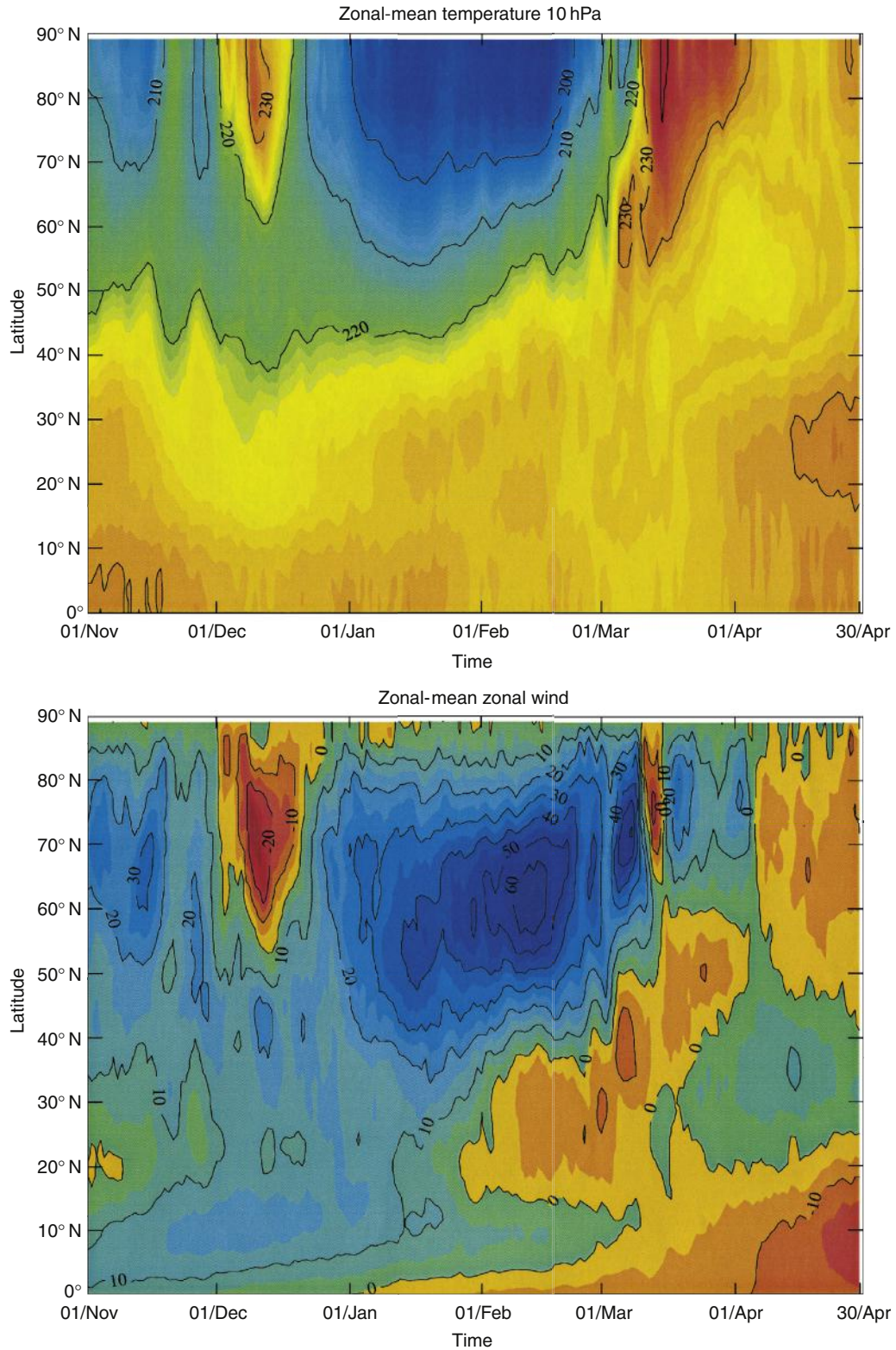


Figure 6 Evolution of zonal-mean temperature and zonal-mean wind for the Northern Hemisphere at 10 hPa from November 1987 through April 1988. The rapid rise of zonal-mean temperature in early December is accompanied by a rapid deceleration of zonal-mean wind, which changes direction from westerly to easterly as far south as 60° N. The subsequent polar cooling and midlatitude recovery of westerly winds correspond to the recovery of the polar vortex by radiative cooling after the stratospheric sudden warming.

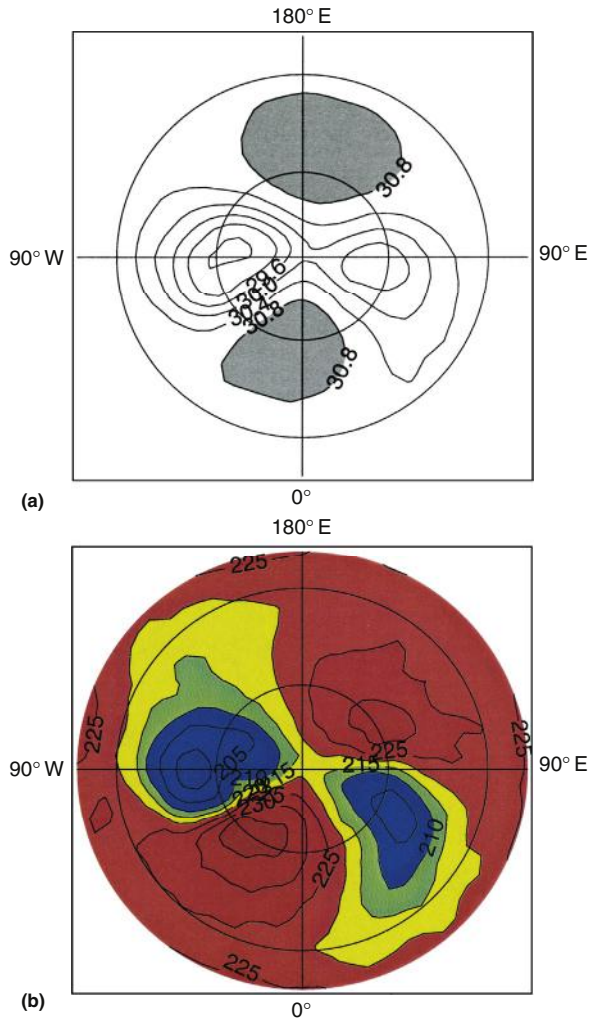


Figure 7 Fields of geopotential height (a) and temperature (b) for the Northern Hemisphere at 10 hPa on 28 December 1984, during the build up to a major warming of the wave-2 type. The Aleutian High near the dateline (shaded) is accompanied by a second (nonclimatological) anticyclone near longitude 0° E. The polar vortex is in the process of splitting into two cyclones. Units: km for geopotential height; K for temperature.

warming is said to have occurred when the latitudinal temperature gradient finally changes sign at the end of winter and easterly winds remain at middle and high latitudes. **Figure 6** illustrates these changes for a final warming in the Northern Hemisphere near the beginning of April 1988.

Final warmings are the main type of dynamical variability in the stratosphere of the Southern Hemisphere. Their evolution is very similar from year to year, and is illustrated in **Figure 10**. A quasi-stationary anticyclone appears near the dateline, typically in early October. The intense polar vortex weakens from the top of the stratosphere steadily downward, as warm air progressively replaces cold air over the polar cap. This breakdown extends down to the lower stratosphere, where a remnant of the polar vortex remains to intensify again during the following winter.

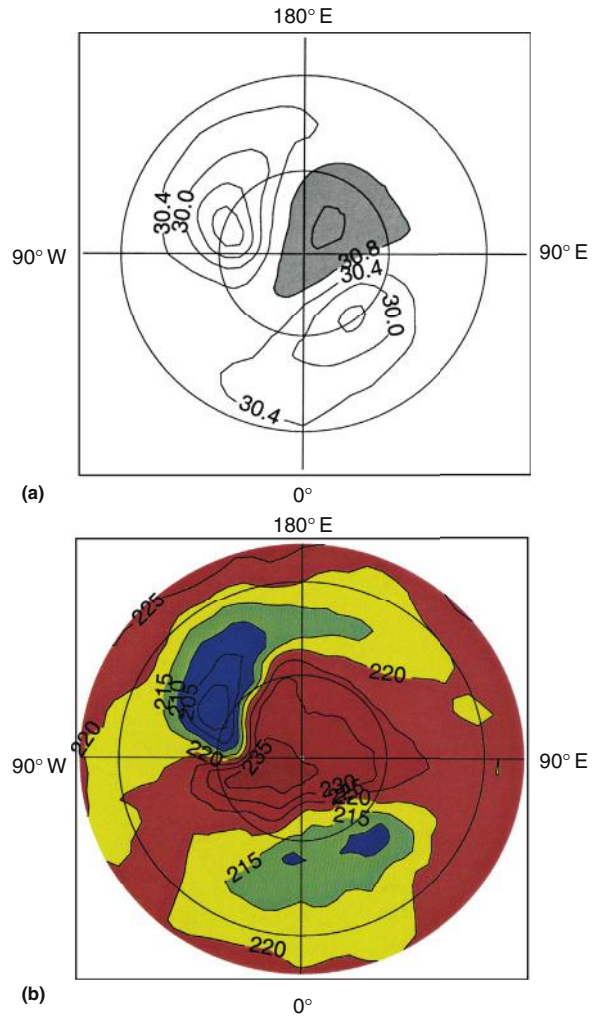


Figure 8 Fields of geopotential height (a) and temperature (b) for the Northern Hemisphere at 10 hPa on 2 January 1985 at the height of a wave-2 major warming. The two anticyclones shown in **Figure 7** have merged over the pole, and warm air has spread over the polar cap. The two cyclones are weakening rapidly as they are drawn westward around the anticyclone (clockwise on the figure). Units: km for geopotential height; K for temperature.

Recent work has suggested one further interesting contrast between the two hemispheres in the evolution of the final warming in the vertical. In the Southern Hemisphere, the final warming almost always occurs from the top-down with the vortex first breaking down at 1 hPa in November followed by subsequent breakdown at higher pressures with a final breakdown at 50 hPa at the end of December. In the more dynamically active Northern Hemisphere, the final warming can occur with two different distinct vertical structures, a top-down mode similar to that in the Southern Hemisphere and a mode in which the vortex breaks down first in the middle stratosphere (~ 10 hPa). In the top-down mode the vortex breaks down at 1 hPa in late April and at 50 hPa in mid-May. In the middle-first mode the vortex breaks down at 10 hPa in early April and subsequently at 50 hPa in late April; the vortex in the upper

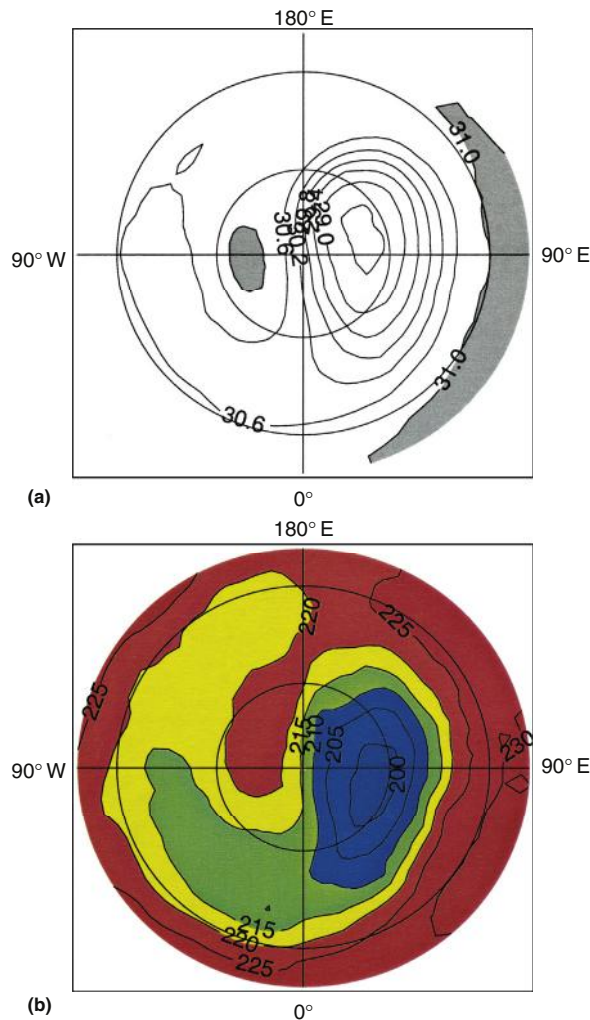


Figure 9 Fields of geopotential height (a) and temperature (b) for the Northern Hemisphere at 10 hPa on 7 December 1981 during a Canadian warming. The Aleutian High (gray shading) has been advected eastward from its usual position near the dateline to 90° W (over Canada). The polar vortex, though displaced from the pole, remains strong. Notice that the temperature changes over the polar cap are much less than those occurring during major warmings, as shown in [Figures 4](#) and [8](#). Units: km for geopotential height; K for temperature.

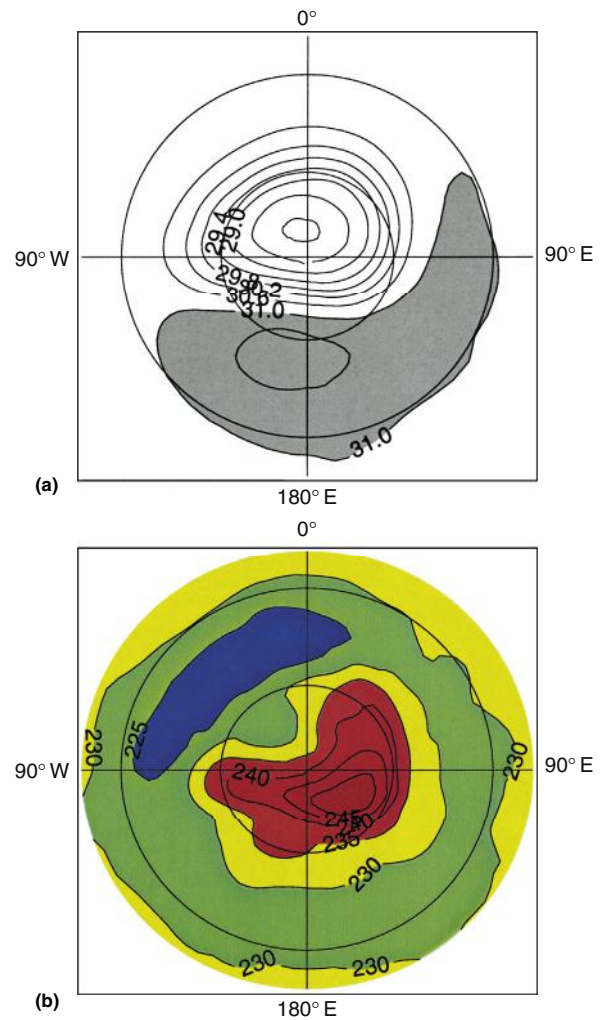


Figure 10 Fields of geopotential height (a) and temperature (b) for the Southern Hemisphere at 10 hPa on 15 October 2001 during a final warming. A quasi-stationary anticyclone (gray shading) has developed over the dateline and the polar vortex is in the process of weakening. At the level shown, warm air has already replaced cold air over the polar cap. The weakening and breakdown of the polar vortex proceeds in a 'top-down' manner from the upper stratosphere to the lower stratosphere.

stratosphere follows a similar evolution to the top-down mode with a breakdown at 1 hPa in late April.

Theories of SSW

Early Ideas

An early theory was that SSWs were due to solar 'storms.' Scherhag entitled his first paper on the phenomenon "Proof of the Influence of Solar Eruptions on the Weather in the Stratosphere." In this, he tried to link an unexpected, intense SSW to a very strong eruption on the sun. It is now known, however, that the warming he observed began much earlier than the solar event.

By the 1960s, theoretical studies were providing compelling evidence that SSWs originated from the upward propagation of planetary-scale disturbances in the troposphere, visible as a large-scale meandering of the westerly jet stream in the troposphere at midlatitudes. It was shown that the direction and strength of the winds in the stratosphere exert a strong control on the ability of these disturbances to propagate upward. The theory showed that in summer, when winds in the stratosphere are easterly, disturbances cannot propagate out of the troposphere to any great extent. This finding explains why SSWs occur only in winter and not in summer. The theory showed that during winter, when winds in the stratosphere are westerly, the longest, planetary-scale disturbances can propagate from the troposphere right through the stratosphere, but

shorter-scale disturbances, characteristic of weather systems in the troposphere, are confined to lower altitudes. This filtering effect of the westerly winds in the stratosphere explains why the geopotential height and temperature patterns associated with SSWs have characteristically large horizontal scales.

Paralleling this theoretical work, studies based on observational data showed that SSWs are accompanied by an increase in the upward energy flux from the troposphere. Building on this finding, the seminal modeling and theoretical work of Taroh Matsuno in the early 1970s provided compelling evidence for the tropospheric origin of SSWs, and set out a theory of SSWs, which is broadly accepted today.

Matsuno's Theory of SSWs

Matsuno's theory of SSWs was built around a theoretical model of stratospheric dynamics that divides the circulation into zonal-mean (averaged around a latitude circle) and wave components (departures from the zonal mean). He postulated that the mean-flow changes observed during sudden warmings, including the deceleration in zonal-mean wind and the mean temperature rise near the pole (Figure 6), are attributable to the effects of vertically propagating planetary waves forced in the troposphere by large-scale disturbances there. A theoretical finding, the Charney–Drazin nonacceleration theorem, had demonstrated that steady, conservative, linear planetary waves are incapable of inducing such mean-flow changes. Matsuno realized that the nonacceleration theorem would be violated when a transient packet of planetary waves first starts to propagate upward. Calculations showed that at the leading edge of the packet a westward body force would be exerted at midlatitudes on the zonal-mean wind. In addition, a zonal-mean temperature increase would occur at polar latitudes (as well as a zonal-mean temperature decrease at low latitudes). Thus, the mean-flow changes associated with the leading edge of a transient planetary-wave packet are qualitatively similar to those observed during SSWs.

Matsuno gave a convincing demonstration of the validity of these theoretical ideas by using a simplified numerical model to simulate the interaction between a single planetary wave and the zonal-mean flow. Wave amplitudes could be specified at the lower boundary of the model near the tropopause, so that the effects of forced, transient planetary waves could be studied. He found that the interaction between the planetary wave and the mean flow led to an evolution of mean wind and temperature fields, which in some cases was reminiscent of that occurring during SSWs. Figure 11 shows fields from an experiment in which the forcing was a planetary wave with zonal wave number 2 (high–low–high–low around a latitude circle). Westerly winds were replaced by easterly winds after about 20 days of integration, and there was an accompanying rapid temperature increase of over 80 K in the simulated stratosphere. The evolution shown in the figure is reminiscent of that seen during SSWs of the wave-2 type described earlier.

The strong association, demonstrated by Matsuno's numerical simulations, between SSWs and transient wave activity in the troposphere offers an immediate explanation of why SSWs do not occur in the Southern Hemisphere during midwinter. Planetary waves are generated by mountains and by contrasts between the temperatures of land and sea. Because much of the

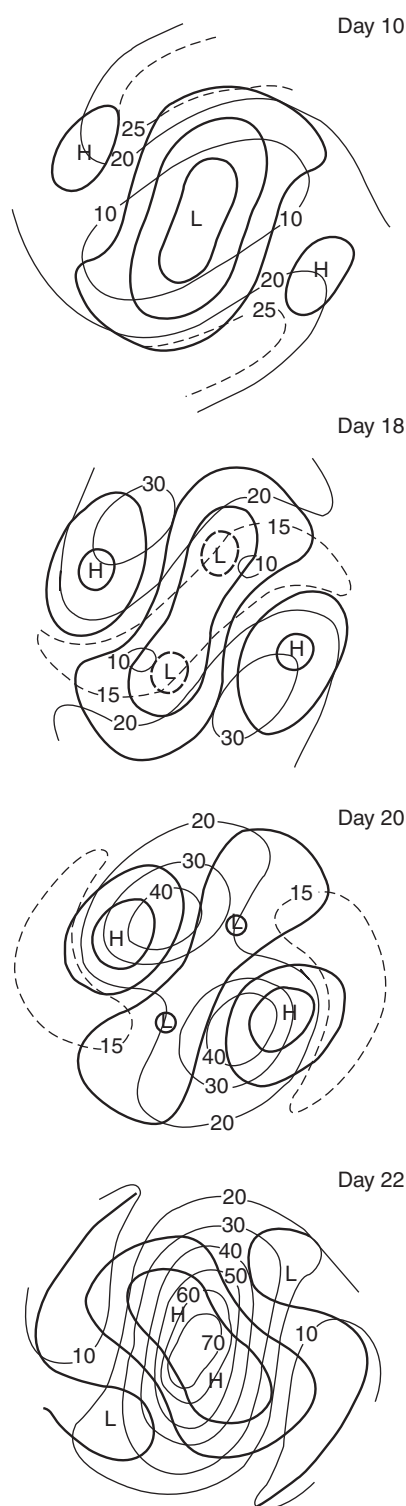


Figure 11 Results from a numerical simulation of an SSW by Taroh Matsuno using an idealized model. The figure shows polar stereographic plots of the time evolution of temperature (thin curves, marked in degrees Celsius) and height (thick curves, 0.5 km intervals) for the isobaric surface corresponding to a height of about 30 km, near 10 hPa. The contours cover the area north of 30° N. Compare this sequence of plots with the sequence shown for an actual wave-2 major warming by Figures 7 and 8.

Earth's surface in the Southern Hemisphere is oceanic, such waves are weaker than in the Northern Hemisphere. The polar vortex in the Southern Hemisphere is correspondingly stronger and more resilient to any short-term increases in planetary wave activity in the troposphere.

Additional Dynamical Aspects

Since Matsuno's seminal work, numerous SSWs have been studied utilizing global observational data from satellites. These studies have amply confirmed the premise that SSWs are associated with increases in the upward fluxes of planetary wave activity in the stratosphere.

However, two features of Matsuno's numerical experiments must be amended for detailed application to actual SSWs. First, SSWs are highly nonlinear dynamical events. The neglect of wave-wave interactions in Matsuno's simplified model means that certain aspects of stratospheric variability cannot be captured by the model. Secondly, Matsuno's experiments involve planetary wave perturbations to a stratospheric state that is initially zonally symmetric. For actual warmings, the initial asymmetry of the stratospheric state is relevant. As we have already noted, wave-1 warmings often arise when an eastward-traveling anticyclone in the stratosphere merges with the preexisting, quasi-stationary Aleutian High, leading to its intensification and an associated weakening of the polar vortex. Wave-2 warmings tend to develop from an earlier wave-1 minor warming, when the stratospheric polar vortex is already highly elongated and therefore more susceptible to splitting.

An Alternative Hypothesis – Self-Tuned Resonance

An alternative to the broadly accepted Matsuno model of warmings described above is the idea originally proposed by Alan Plumb in 1981 that warmings might occur through the resonant growth of an oscillatory mode of the polar vortex excited by the interaction of the vortex with the underlying topography. Importantly in the Plumb model, the vortex need not be forced precisely at the resonant frequency because nonlinear effects can allow the system to 'self-tune' toward the resonant state. In September 2002, a major, vortex splitting sudden warming was unexpectedly observed in the Southern Hemisphere. This has led to resurgence in interest in the dynamics of sudden warming events and in the self-tuned resonance theory in particular. An interesting hypothesis resulting from this work is the idea that vortex splitting and vortex displacement warmings might have fundamentally different underlying dynamics, in contrast to the Matsuno theory above. Vortex splitting events like the 2002 Southern Hemisphere sudden warming might be best explained as the resonant excitation of the barotropic or external vertical Rossby mode of the vortex.

Dynamical Links with the Troposphere

Although some progress has been made in establishing which features of the tropospheric circulation can produce SSWs, the precise nature of these connections remains somewhat elusive.

Broadly speaking, it appears that the strength of the stratospheric Aleutian High, the growth of which almost invariably plays a part in SSWs, is strongly influenced by the strength of a large-scale climatological feature of the tropospheric circulation, the so-called East Asian Low, located near 140° E in longitude. It is the most intense, persistent cyclonic structure in the troposphere (Figure 12). It is produced largely by the Himalayan mountain range and by land-sea contrasts in winter between the cold Asian continent and the warmer Pacific Ocean. To a first approximation, the Aleutian High can be viewed as part of a downstream Rossby wave train emanating from the East Asian Low.

Some major SSWs of the wave-1 type do appear to be associated with the anomalous intensification of the East Asian Low. It is not the case, however, that there is generally a simple, one-to-one correspondence between an intensification of the Aleutian High and a preceding intensification of the East Asian Low or other tropospheric structure.

Experiments with numerical models have revealed why the expectation of a simple correspondence between transient events in the troposphere and SSWs in the stratosphere, as envisaged in Matsuno's original theoretical scheme, can be frustrated. Experiments with fully nonlinear models show that when planetary wave amplitudes in the troposphere are set to typically observed values and are held steady (i.e., there is no imposed wave transience), the simulated stratosphere exhibits strong internal variability reminiscent of a sequence of minor warmings. In these experiments, the violation of the Charney–Drazin nonacceleration theorem is the result not of transience in wave forcing (as in Matsuno's experiments) but of nonlinearity, which results in transience of both the wave and the mean-flow response. Evidently, if steady waves in the troposphere can cause SSW-like variability in the stratosphere, then a straightforward attribution of an SSW to a transient event in the troposphere cannot always be expected. Nevertheless, it is the case that major SSWs occur when planetary wave amplitudes in the troposphere are anomalously high, even if the exact time sequence of events may be complicated.

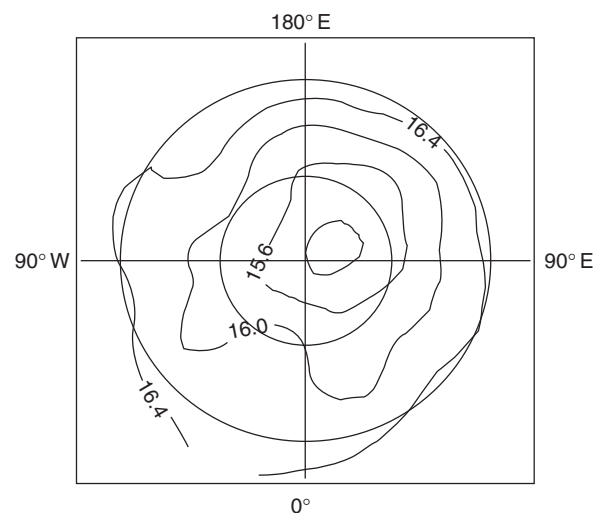


Figure 12 Field of geopotential height at 100 hPa near the tropopause on 1 December 1981. The geopotential 'low' in the top right quadrant corresponds to the East Asian Low seen at lower altitudes. Unit: km.

By contrast, attempts to link major warmings of the wave-2 type with the development of specific features in the tropospheric circulation have been more successful. The dramatic split in the polar vortex, which occurs during these events, can often be traced right down into the troposphere, where the East Asian Low is partnered by the appearance of another cyclonic circulation over North America near 90° W (Figure 13) – hence the growth of a strong wave-2 pattern in the troposphere.

Factors Affecting the Occurrence of SSWs on Interannual and Longer Timescales

Different winters show substantial differences in the frequency and nature of SSWs. While much of this variability is likely to be random in nature, the likelihood of SSW occurrence in any particular winter is thought to be influenced by several different processes in the Earth system.

1. Variability in the large-scale structure of the tropospheric circulation;
2. Variability in the tropical stratosphere; and
3. Variability of external forcings.
 - a. A major source of tropospheric variability is the El Niño Southern Oscillation phenomenon (ENSO). Studies with general circulation models have shown that the occurrence of a major SSW is enhanced during both the warm and cold phase of ENSO. This is particularly interesting since the mean state of the stratosphere is warmer/colder during El Niño/La Niña, yet the frequency of SSW is, in both cases, much higher than for neutral conditions in the east Pacific. The mechanism for this important influence of ENSO is still poorly understood, and is the subject of much ongoing research.
 - b. The other dominant source of variability in the tropical stratosphere is the quasi-biennial oscillation (QBO), which involves the quasi-periodic reversal of zonal winds from westerly to easterly and back again. Theory shows that the zonal-wind structure in the stratosphere

affects the propagation of planetary waves from the troposphere. The standard explanation for the relationship between the QBO and changes in the polar vortex is that when the QBO is in its westerly phase throughout the tropical lower stratosphere, the waves are refracted away from polar regions toward the tropics. The stratospheric polar vortex should therefore be cold and strong. When, on the other hand, the QBO is in its easterly phase, the waves are more confined to middle and polar latitudes. The polar vortex should be warmer and weaker and therefore more susceptible to breakdown by SSWs. Recent modeling work has, however, begun to question these ideas and, as in the case of ENSO, further experimentation is needed to fully understand the relationship between the QBO and the frequency of SSWs.

- c. Finally, there are many different external climate forcings, both natural and anthropogenic, that may influence the frequency of SSWs over coming decades. The most persistent of these is the forcing of the stratospheric state by the continued emission of greenhouse gases. Most modeling studies, which examine the number of SSWs in climates with significantly enhanced levels of carbon dioxide suggest that a small increase in the frequency of SSWs is likely to be observed by the end of the twenty-first century (assuming greenhouse gases continue to increase in coming decades). Two important caveats to this broad conclusion, however, are that there are significant differences among climate models in predictions of future SSW frequency, and that the magnitude of any change in SSW frequency is likely to be small compared to decadal variability in SSW occurrence. The extent to which changes in the structure and strength of the polar vortex or changes to the generation of planetary wave anomalies in the troposphere are implicated in the change in SSW remains the topic of current research.

Other external forcings, which might influence the frequency of SSWs predominantly consist of processes, which perturb the stratospheric radiative balance. These include changes in solar output, the decline and recovery of stratospheric ozone, changes to the stratospheric aerosol distribution either through volcanic eruptions or anthropogenic geo-engineering and changes to stratospheric water vapor. Since the study of SSWs as climate phenomenon is relatively recent and since the observational record is relatively short, there has been very limited work, to date, on the influence of these processes on SSW frequency.

It is certainly also important to note that because of the short observational record in the stratosphere and the fact that models have only recently begun to be able to simulate SSWs well, understanding and discovering connections between the SSW frequency and other parts of the climate system is at an early stage. This is an area of research that is likely to expand in future years.

Effects of Sudden Warmings

On the Distribution of Trace Constituents

An SSW irreversibly alters the distribution of trace constituents in the stratosphere. Air is transported over large distances in the

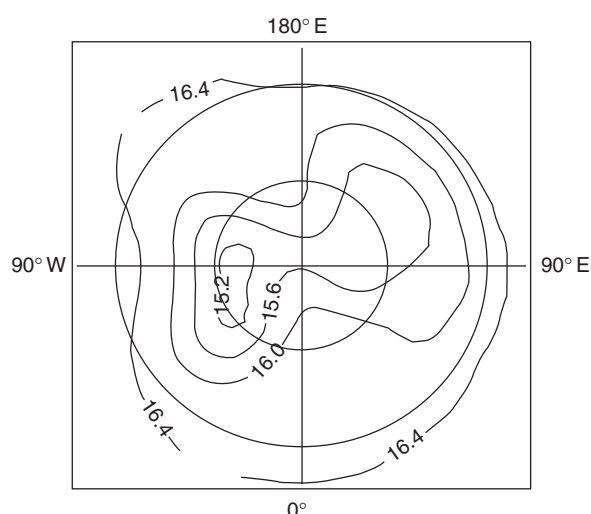


Figure 13 Field of geopotential height at 100 hPa on 28 December 1984. Note the correspondence between this field and that shown in Figure 7. Unit: km.

north–south direction and irreversibly mixed at midlatitudes. In addition, there is an enhancement of the overturning, meridional circulation, consisting of systematic ascent of air at low latitudes and systematic descent of air at high latitudes.

Figure 5 illustrates the large-scale, quasi-horizontal mixing that occurs during an SSW. As the Aleutian High (or any other anticyclone) intensifies, it entrains in air from low latitudes, sometimes from deep within the tropics. Air is also peeled off the polar vortex into the anticyclone, where it is wound into ever narrowing spirals along with the low-latitude air. Any preexisting, latitudinal gradients in chemical concentrations are reduced by this anticyclonic mixing. Chemical reactions can also occur, since previously well-separated reactive species are brought into chemical contact. The breakup of the polar vortex during a major SSW, or during a final warming, leads to the greatest mixing of polar and low-latitude air, bringing, for instance, ozone-poor air from the polar vortex to midlatitudes.

On the Weather

SSWs are the premier mode of coupling between the stratosphere and troposphere, representing both:

- An influence of the troposphere on the stratosphere through the upward propagation and breaking of tropospheric planetary waves, and
- An influence of the resulting changes to the stratospheric mean state on the troposphere on intraseasonal timescales.

In describing stratosphere–troposphere coupling, SSWs can be thought of playing a role analogous to that of ENSO in describing ocean–atmosphere coupling in the tropical Pacific. A key aspect of this coupling is that the circulation anomalies related to SSWs in the lower stratosphere persist for long time (up to 60 days) potentially providing a source of memory for extended-range and seasonal weather forecasts.

The typical response of the tropospheric circulation to an SSW event is often described as a shift of the Northern Annular Mode toward its negative phase. In simpler and more intuitive terms, this means the southward migration and weakening of

the extratropical, eddy driven jet stream. The resulting changes to surface temperature include cold anomalies over North America and Western Europe and warm anomalies over Newfoundland, Greenland, and Southern Europe.

Although this is the widely accepted view of the influence of stratospheric anomalies, several other factors must also be considered. First, some studies show that the influence of SSW events is more consistently as described above in the Atlantic basin than in the Pacific basin. Second, SSWs are also thought to bring about changes in the breaking of baroclinic waves in the troposphere, leading to change in the position and structure of the synoptic weather systems, which dominate day-to-day variations in weather at extratropical latitudes. Finally, experiments which examine if better predictions of the stratospheric state might improve tropospheric weather forecasts suggest that the quantitative size of any improvement to predictability is small but significant.

See also: Dynamical Meteorology: Overview; Waves. Middle Atmosphere: Planetary Waves; Polar Vortex; Transport Circulation; Zonal Mean Climatology.

Further Reading

- Andrews, D.G., Holton, J.R., Leovy, C.B., 1987. *Middle Atmosphere Dynamics*. Academic Press, Orlando, FL.
- Charlton, A.J., Polvani, L.M., 2007. A new look at stratospheric sudden warmings. Part I. Climatology and modelling benchmarks. *Journal of Climate* 20 (3), 449–471.
- Gill, A.E., 1982. *Atmosphere–Ocean Dynamics*. Academic Press, London.
- Labitzke, K., 1981. Stratospheric–mesospheric midwinter disturbances a summary of observed characteristics. *Journal of the Meteorological Research* 86, 9665–9678.
- Labitzke, K., van Loon, H., 1999. *The Stratosphere Phenomena, History and Relevance*. Springer, New York.
- Limpasuvan, V., Thompson, D.W.J., Hartmann, D.L., 2004. The life cycle of the northern hemisphere sudden stratospheric warmings. *Journal of Climate* 17 (7), 2584–2596.
- McIntyre, M.E., 1982. How well do we understand the dynamics of stratospheric warmings? *Journal of the Meteorological Society of Japan* 60, 37–65.
- Waugh, D.W., Polvani, L.M., 2010. Stratospheric polar vortices. In: Polvani, L.M., Sobel, A.H., Waugh, D.W. (Eds.), *The Stratosphere: Dynamics, Chemistry and Transport*. Geopress, American Geophysical Union, Washington.

Transport Circulation

SE Strahan, Universities Space Research Association, NASA Goddard Space Flight Center, Greenbelt, MD, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by W A Norton, volume 3, pp 1353–1358, © 2003, Elsevier Ltd.

Synopsis

The transport circulation is forced by upward propagating tropospheric waves that break in the stratosphere and mesosphere, depositing energy and momentum. This drives a global-scale circulation with slow ascent in the tropics balanced by poleward and downward motion in the middle and high latitudes. Observations of long-lived trace gases have elucidated the details of stratospheric transport processes. Transport plays a central role in stratospheric composition through control of the distribution of long-lived gases such as N_2O , CH_4 , and chlorofluorocarbons. Their distributions impact stratospheric ozone and climate.

Introduction

The Physical Basis for the Transport Circulation

The stratospheric meridional circulation, often referred to as the Brewer–Dobson Circulation (BDC), was deduced by Brewer and Dobson from observations of the distributions of water vapor (H_2O) and ozone (O_3). Air enters the stratosphere through the tropical tropopause where it ascends slowly, moving gradually poleward and downward. Above the stratosphere, the mesospheric circulation moves air from the summer to the winter pole. Air in the descending branch of the BDC exits the stratosphere at the extratropical tropopause

and affects tropospheric composition. Figure 1 shows a schematic of the middle atmosphere circulation.

On seasonal and longer timescales, the mean stratospheric circulation can be described in the latitude–height plane where averages are taken around latitude circles ('zonally averaged'). Transport in the middle and high latitudes has a strong seasonal cycle with large interhemispheric differences. The equations of motions governing the zonally averaged circulation can be represented using the Transformed Eulerian Mean (TEM) framework. Because the TEM description formulates complex wave-mean flow interactions in the stratosphere in terms of a single zonal wave force acting on the zonal mean

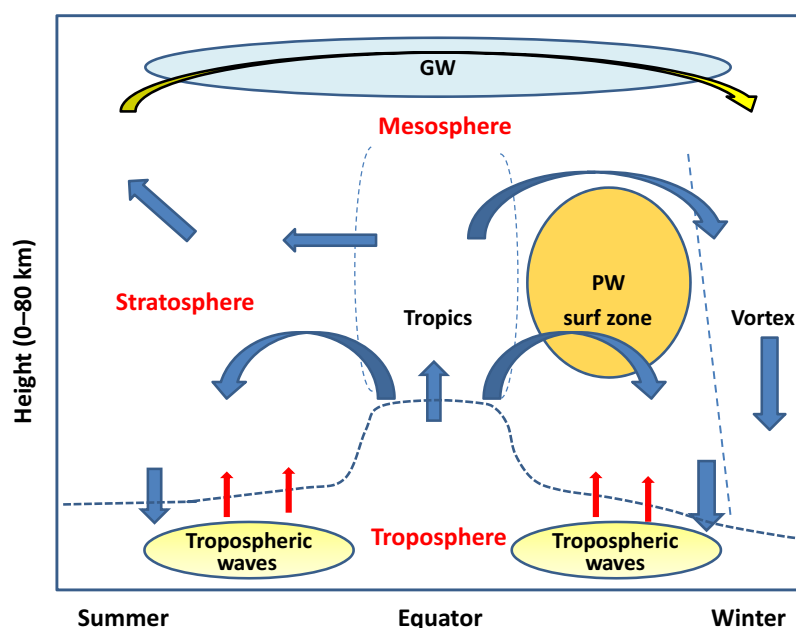


Figure 1 Schematic of transport circulation in the middle atmosphere. Transport by the Brewer–Dobson circulation (BDC) is shown by blue arrows. In the lower stratosphere, the BDC transports air from the tropics into both hemispheres, then downward through the tropopause. In the middle and upper stratosphere, the BDC has a strong poleward, downward circulation in the winter hemisphere where there is planetary wave (PW) breaking. In the summer hemisphere the upper stratospheric circulation is poleward and upward. The pole-to-pole transport in the mesosphere is shown by the yellow arrow. Gravity waves (GWs) are the major source of wave energy in the mesosphere. The year-round source of tropospheric wave energy propagating into the stratosphere is depicted by red arrows. Regions with restricted two-way transport ('mixing barriers') are indicated by vertical dotted lines. The dashed thick blue line above the troposphere is the tropopause.

flow, it provides the most useful framework for describing mean constituent transport in the extratropics (regions poleward of 30°), where planetary-scale Rossby waves provide the dominant wave forcing. The circulation can also be formulated using isentropic (constant entropy) coordinates. Isentropic surfaces, also known as potential temperature surfaces, are the quasi-horizontal layers along which long-lived stratospheric constituents are adiabatically advected.

Synoptic-scale tropospheric disturbances occur throughout the year creating waves that deposit energy and momentum in the extratropical lower stratosphere, driving the lower branch of the equator-to-pole mean circulation in each hemisphere. In the middle and upper stratosphere, the wintertime stratospheric transport is dominated by large-scale Rossby planetary waves (PWs) that are forced by topography or diabatic heating patterns and propagate upward from the troposphere. When these waves break, they mix air along quasi-horizontal isentropic (potential temperature) surfaces. The region of wave breaking between the tropics and high latitudes is known as the 'surf zone.' Waves tend to break at critical surfaces where the zonal wind is the same as the zonal phase speed of the wave. The large-scale Rossby waves are quasi-stationary (i.e., have zero phase speed) and cannot propagate through the zero wind line. In summer, stratospheric winds in the extratropics shift from Westerly to Easterly, and the PWs are no longer able to propagate into the stratosphere. Without energy and momentum from breaking waves, the summer circulation is weak. Mixing processes are slow and temperatures are near radiative equilibrium.

Inertia-gravity waves, with a broad distribution of horizontal phase speeds, propagate upward through the stratosphere year-round regardless of wind direction, depositing momentum in the upper stratosphere and mesosphere. The wave energy drives the mesosphere far from radiative equilibrium, leading to a strong circulation equatorward from the summer pole and poleward in the winter hemisphere. Diabatic cooling at the winter pole leads to strong downward transport of mesospheric air deep into the stratosphere.

Diabatic heating increases with altitude in the stratosphere primarily due to the absorption of solar UV radiation by O_3 . In spite of this thermal forcing, the stratospheric circulation is predominantly driven by the deposition of momentum from waves. Rossby wave breaking in the winter hemisphere not only mixes constituents in the midlatitude surf zone but also leads to a net poleward and downward transport. As mass sinks at high latitudes in the descending branch of the BDC, the air adiabatically warms and keeps polar temperatures above radiative equilibrium. The circulation driven by the net heating is called the diabatic circulation and it represents approximately the same global-scale meridional circulation described by the TEM framework. This wave-driven circulation is also referred to as the 'extratropical pump' because midlatitude wave driving effectively draws air poleward and downward while the tropics respond with rising motions and adiabatic cooling.

The Impact of the Transport Circulation on Trace Gases

The BDC, depicted two-dimensionally (2D) in [Figure 1](#), represents the slow net meridional (horizontal) and vertical motions of mass in the middle atmosphere. The mean

timescale for an air parcel to travel through the middle atmosphere may be up to ~ 6 years. Trace gases enter the stratosphere through the tropical tropopause. As the tropical air slowly ascends, it is exposed to increasingly higher levels of UV radiation and increasing concentrations of reactive species (i.e., radicals). The longer an air parcel has been in the stratosphere, the greater the likelihood it has traveled to high altitudes where it experienced photochemical loss. For many long-lived species, such as the tropospheric source gases, nitrous oxide (N_2O), methane (CH_4), and chlorofluorocarbons (CFCs), photochemical losses become important in the middle stratosphere (~ 30 hPa) and above, primarily in the tropics. Breaking planetary-scale waves in the winter hemisphere drive the 'extratropical pump' which transports tropical air poleward and downward into the photochemically cooler environment of the middle and high latitudes. The integrated effects of transport and exposure to reactive chemical environments control the large-scale mean distributions of long-lived trace gases as well as their lifetimes.

[Figure 2](#) shows the zonal mean structure of two long-lived trace gases, N_2O and CH_4 . The lines of constant mixing ratio (isopleths) are lifted upward in the tropics by the ascending branch and are pushed downward in the middle and high latitudes by the descending branch of the BDC. Tracer isopleths in the middle latitudes of the winter or spring hemisphere are flattened as a result of rapid mixing by breaking PWs, creating a 'surf zone.' The balance between horizontal mixing (flattening) and the diabatic circulation (steepening) produces the observed tracer slope.

Although N_2O and CH_4 have different tropospheric sources and different photochemical loss processes – including differences in the altitudes and latitudes where losses occur – their stratospheric distributions shown in [Figure 2](#) are quite similar. These observations provide compelling evidence that their features are controlled by transport and hence dynamical processes. This is the case for all long-lived trace gases, where long-lived means that the timescales for chemical production or loss are much longer than the timescales for transport. When two trace gases are controlled by transport they show a compact correlation. This has been demonstrated by numerous studies of stratospheric trace gas, many using observations from aircraft campaigns. The compact correlation between N_2O and CH_4 in the lower stratosphere is illustrated in [Figure 3](#). The compactness arises because horizontal transport occurs rapidly compared to vertical transport by the circulation. Ozone has rapid production and loss processes in the middle stratosphere but is long-lived in the lower stratosphere. It shows a compact correlation with species such as N_2O and CH_4 in the lower stratosphere.

The combined effects of the slow, mean-meridional circulation of the stratosphere and rapid isentropic mixing in the midlatitude surf zone result in three relatively separate regions in the winter hemisphere. The left panel of [Figure 4](#) shows the zonal mean distribution of Microwave Limb Sounder (MLS) N_2O data from a day in northern winter. The midlatitude surf zone is easily identified by the flattened isopleths from 450 to 1000 K (approximately 7–70 hPa) in between the tropical and polar regions. The right panel of [Figure 4](#) shows an area-weighted probability distribution function (PDF) of MLS N_2O data from tropical to polar latitudes on a single potential

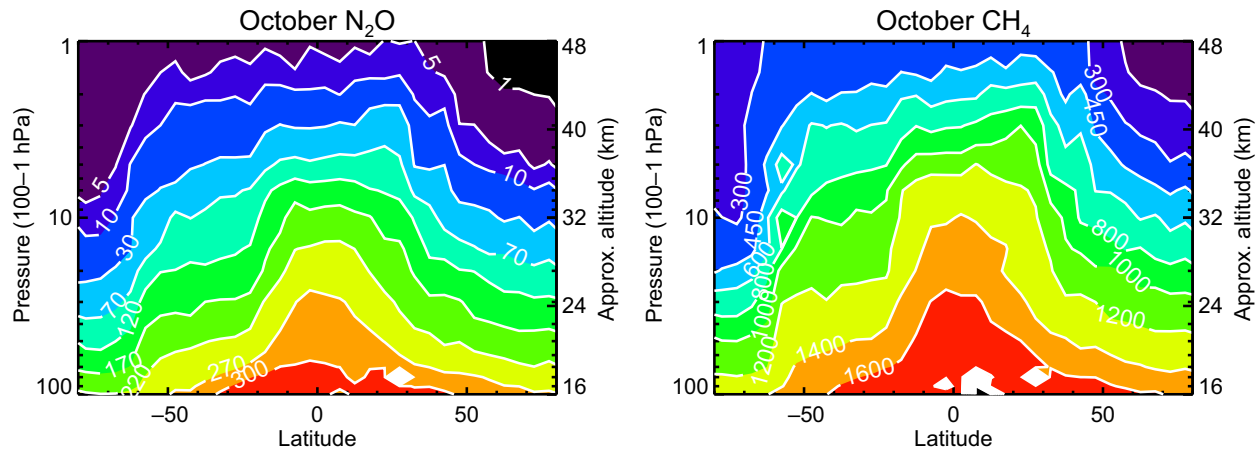


Figure 2 Zonal mean climatologies 80°S – 80°N in September and October N_2O and CH_4 (ppb) derived from SCISAT-1 Atmospheric Chemistry Experiment (ACE) observations, 2004–09. The altitude range shown spans the stratosphere. White areas indicate missing data. ACE climatology courtesy of Ashley Jones.

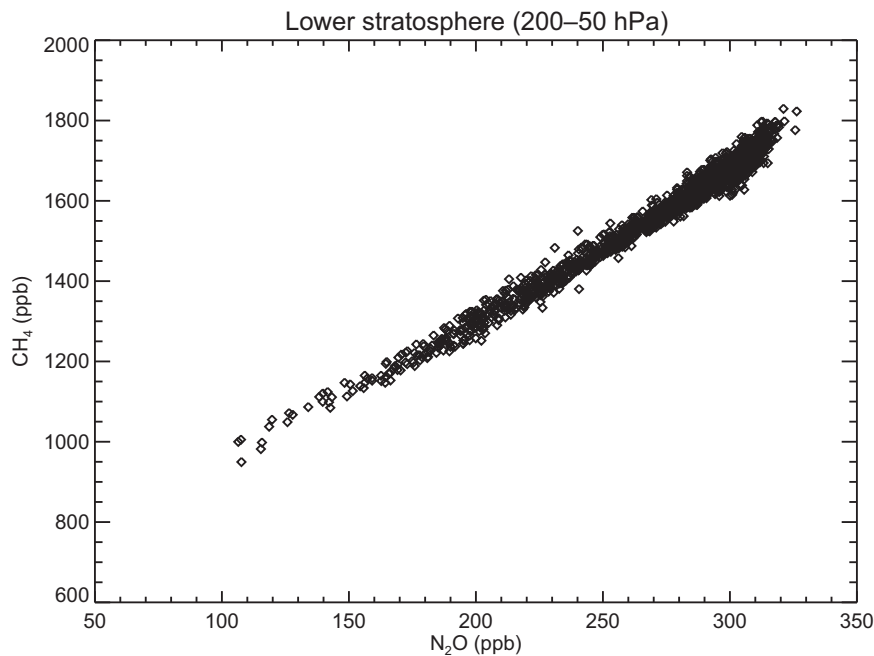


Figure 3 The compact correlation of ACE N_2O and CH_4 in September and October in the lower stratosphere and tropical upper troposphere, 200–50 hPa. The compactness is a consequence of horizontal mixing occurring on timescales shorter than vertical transport by the BDC.

temperature surface. The PDF shape reveals well-mixed regions along with a lack of mixing between them. This PDF indicates that tropical air ($\text{N}_2\text{O} > 220$ ppb) is well mixed, i.e., has a fairly narrow distribution, but is isolated from the effects of midlatitude mixing. It is not completely isolated as some mixing ratios are observed that are intermediate between tropical and midlatitude values, e.g., 210 ppb. The midlatitudes are also well mixed (approximately 120–180 ppb), but quite isolated from the lower mixing ratios found in the polar vortex (10–20 ppb). The minimum in between the midlatitude and vortex peaks is deep and wide, indicating extremely weak mixing between these regions. These minima are referred to as

the subtropical and polar vortex transport barriers. High N_2O mixing ratios in the ascending BDC are associated with recent arrival of tropospheric air in the stratosphere having little photochemical loss, while low N_2O at high latitudes found in the descending branch of the BDC is photochemically aged air. It has spent years in the stratosphere with greater exposure to photochemical loss.

The high latitude gradient develops when the vortex forms, as polar air diabatically (radiatively) cools and descends in the fall. Wave-driven radiative cooling strengthens the poleward downward circulation of the BDC. (This also increases ascent rates in the tropics.) The effect of waves on the downward circulation is

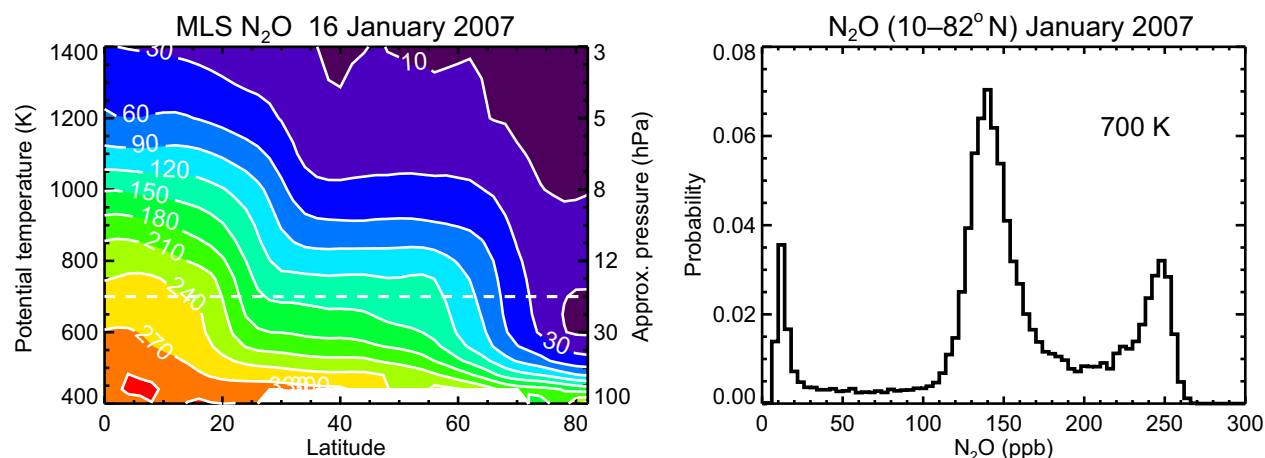


Figure 4 Left: Zonal mean N₂O observed by Aura MLS on 16 January 2007 on potential temperature (isentropic) surfaces. Approximate pressure levels are given on the right side of the plot. The broad, flat region in the midlatitudes is the well-mixed surf zone. Steep gradients near 25° and 70° reflect the isolation of the tropics and the high latitudes (vortex). The dashed line at 700 K (~18 hPa) corresponds to the data used in the probability distribution function (PDF) calculation on the right. Right: Area-weighted PDF of the N₂O observations on the 700 K surface, averaged over 10 days in January. The PDF maxima indicate three well-mixed regions: the polar vortex (10–20 ppb), the midlatitude surf zone (120–170 ppb), and the tropics (>220 ppb). The minima between the peaks indicate weak mixing between the regions.

less inside the vortex, resulting in smaller descent rates there than at the vortex edge. Because waves do not usually penetrate the vortex, vortex air becomes isolated and long-lived trace gases such as N₂O have low values in the vortex relative to the midlatitudes. The net result of these processes, mid- and high-latitude descent and isentropic mixing restricted to the midlatitudes, is to create a steep tracer gradient at the vortex edge. When the sun returns to the winter hemisphere in spring, strong wave activity disturbs the vortex leading to its eventual breakdown and mixing with midlatitude air. With the vortex gone, so are the steep trace gas gradients between mid- and high latitudes. Unlike the vortex edge, the subtropical edge of the surf zone exists year-round. In winter and spring, breaking midlatitude waves unable to penetrate the tropics lead to a region of weak horizontal mixing in the subtropics, similar to the situation at the vortex edge. The subtropical gradient persists in summer and the fall, but it is less steep because it no longer sharpened by erosion. When midlatitude wave activity is at a minimum, a PDF of the summer hemisphere would show only the midlatitude and tropical peaks with a shallow minimum between them. The strength of each hemisphere's transport circulation varies due to seasonal variations in wave driving, which leads to seasonal variations in trace gas distributions.

The strength of large-scale tropospheric waves propagating into the stratosphere can have significant interannual variations. The interaction of tropical Kelvin and gravity waves forces a downward propagating oscillation in the direction and speed of winds in the equatorial lower and middle stratosphere. The mean period of the oscillation, known as the quasi-biennial oscillation (QBO), is about 26 months. The QBO interacts with extratropical Rossby waves in a way that depends on the phase of the QBO (i.e., Easterly or Westerly). These phase-dependent interactions impact the degree of isolation of the tropics and of the polar vortex, leading to interannual variations in lower and midlatitude stratospheric composition.

Timescales for Transport in the Stratosphere

Mean age is a particularly useful quantity for understanding transport because it is a measure of the integrated effects of circulation and mixing. Stratospheric 'transit time' is the time required for an infinitesimally small fluid element to travel from its stratospheric entry point at the tropical tropopause to any given point in the stratosphere. Due to mixing and seasonal variations in the strength of the circulation, there are many possible transport pathways to a given point, thus the transit time is actually a distribution or 'spectrum' of times. An air parcel may be considered as an observable quantity composed of a large number of fluid elements, each with its own transport pathway and time. The mean age of a parcel is the average of the transit times of all the fluid elements comprising the parcel. The older the mean age of air, the more likely the air has traveled to high altitudes where it would be exposed higher UV levels and fast photochemistry. Observations of certain very long-lived trace gases, with well-measured growth rates in the troposphere, can be used to deduce transport timescales (mean age) in the stratosphere.

Stratospheric Transport Derived from Observations

Over the past 30 years, satellite datasets, along with observations from aircraft and balloon campaigns, have clarified our understanding of the transport circulation, leading to a detailed understanding of many transport processes. For example, instruments on the NASA Upper Atmosphere Research Satellite (UARS, 1991–2005), the NASA Aura satellite (2004–ongoing), and the Canadian SCISAT-1 (2003–ongoing) have provided multiyear datasets with near global coverage of long-lived trace gases including CH₄, N₂O, H₂O, HF, HCl, CFCl₃ ('CFC-11'), CF₂Cl₂ ('CFC-12'), and O₃. In some cases the instruments also

measure shorter-lived species such as NO_2 , HNO_3 , ClO , and ClNO_3 , whose abundances affect O_3 . Multiyear satellite datasets provide climatological information on stratospheric transport processes and their interannual variability. *In situ* instruments on aircraft and balloon platforms measure many of the same species observed by satellite instruments but at much higher spatial resolution, although their spatial and temporal coverage is limited. Aircraft data with spatial resolution of hundreds of meters have revealed the small scale of filamentation processes occurring at the edge of the polar vortex.

Circulation and Mixing: Age of Air

The extremely long-lived species carbon dioxide (CO_2) and sulfur hexafluoride (SF_6) have surface sources with well-measured growth rates that can be used to calculate the mean age of air in the stratosphere. Empirically, CO_2 mixing ratios and their seasonal cycle observed at the tropical tropopause, i.e., the stratospheric entry point, can be approximated by the average of surface CO_2 measurements at two tropical stations, Samoa (14°S) and Mauna Loa (19°N), delayed by 2 months. Because of the paucity of tropical tropopause measurements, the time series of CO_2 from these stations is used as a convenient proxy for the CO_2 mixing ratio at its entry point to the stratosphere. These stations have measured CO_2 for many decades, allowing growth rates to be accurately calculated. The growth rates allow CO_2 measurements in the stratosphere to be interpreted as a 'clock,' where the elapsed time between stratospheric entry and a particular point in the stratosphere can be determined by when that mixing ratio was observed at the entry point. Figure 5 illustrates how balloon-borne CO_2 measurements made over a period of more than two decades can be compared to the proxy time series at the stratospheric entry point to determine mean age at a point in the middle stratosphere. Notice that both lines have same slope, which is equal to the growth rate of surface CO_2 . The horizontal offset

between the two lines (Δx) is the mean transit time. (Due to the contribution of the oxidation of CH_4 to CO_2 , a correction of ~ 0.5 years must be applied to the mean age derived from this method.) This data analysis shows that the mean transit time to the midlatitude middle stratosphere has consistently been ~ 5.0 years for more than two decades.

Figure 6 shows mean age derived from CO_2 and SF_6 measurements from a wide range of latitudes and altitudes made over a period of more than 30 years. These data represent a wealth of knowledge about the combined effects of circulation and mixing in the midlatitudes and tropics. Mean age in the tropics is low, reflecting the recent entry of air in the stratosphere. Age increases with latitude because of increasing contributions from air descending from higher altitudes in the stratosphere (see Figure 1). Mean ages in the polar regions have large seasonal variability, reflecting the seasonal variations in the descending branch of the BDC.

Tropical Ascent and Subtropical Mixing

Aircraft campaigns in the late 1980s and the early 1990s, flying from the subtropics to the tropics in both hemispheres observed sharp meridional gradients in long-lived trace gases, e.g., N_2O , O_3 , and NO_y (a long-lived quantity equal to the sum of all reactive nitrogen species). This revealed the isolation of tropical air from the midlatitudes, providing evidence that the strong mixing in the surf zone did not penetrate the tropics. The global-scale picture of tropical isolation was uncovered using satellite measurements of H_2O , CH_4 , and temperature with the discovery of the 'water vapor tape recorder.' Water vapor becomes 'freeze dried' as it ascends through the cold tropical tropopause – a very cold region, typically 190–195 K. The extratropical pump (wave driving) that controls the strength of the BDC is strongest in northern winter and weakest in northern summer. This causes tropical tropopause temperatures to be lowest in winter and highest in summer. The seasonal cycle in tropopause temperatures modulates the

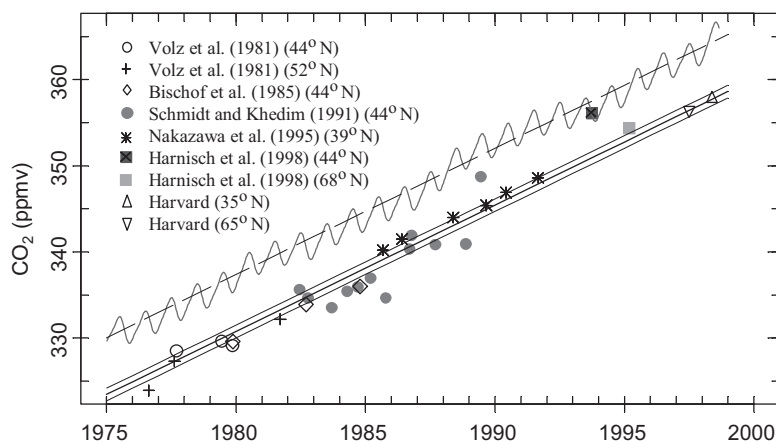


Figure 5 Proxy time series of CO_2 at the tropical tropopause (gray) and midlatitude CO_2 observations from ~ 45 to 50°N , 20–25 km (50–25 hPa) (symbols). The lines represent the best fit to the points. The temporal offset between the two lines gives the mean stratospheric transit time to 45 – 50°N , 20–25 km. From Andrews, A.E., et al., 2001. Mean ages of stratospheric air derived from *in situ* observations of CO_2 , CH_4 , and N_2O . *Journal of Geophysical Research* 106 (D23), 32295–32314. Copyright 2001, American Geophysical Union. Reproduced with permission from the American Geophysical Union.

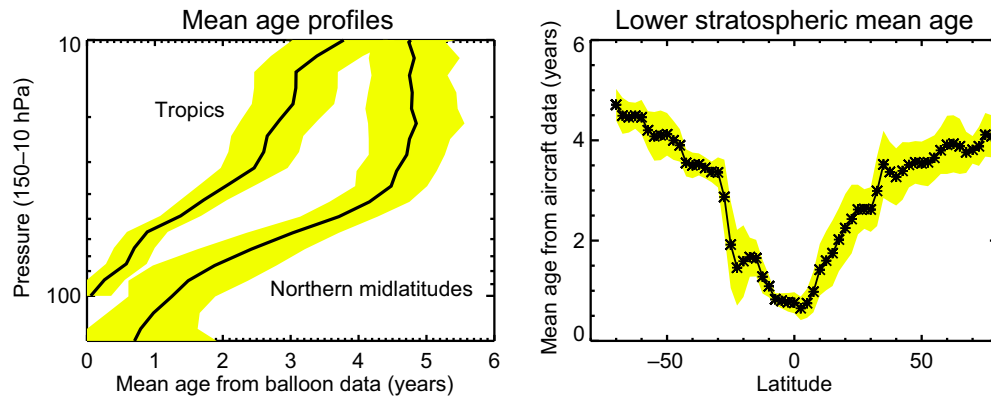


Figure 6 Left: Tropical and midlatitude (35–50° N) average mean age profiles in the lower and middle stratosphere derived from CO₂ and SF₆, and their 1σ uncertainties (yellow shading). Right: lower stratospheric mean age derived from aircraft CO₂ observations at ~20 km (~50 hPa) and 1σ uncertainties (yellow shading).

saturation vapor pressure of water entering the stratosphere. As air slowly ascends in the tropics, the temperature-modulated H₂O mixing ratios produce an upward propagating cycle, shown in Figure 7. The water maxima (orange) are mixing

ratios that passed through the tropopause when it was relatively warm (summer), while the minima (blue/green) are the low mixing ratios that ascended through a colder tropopause (winter). The maxima (or minima) appearing at ~26 km

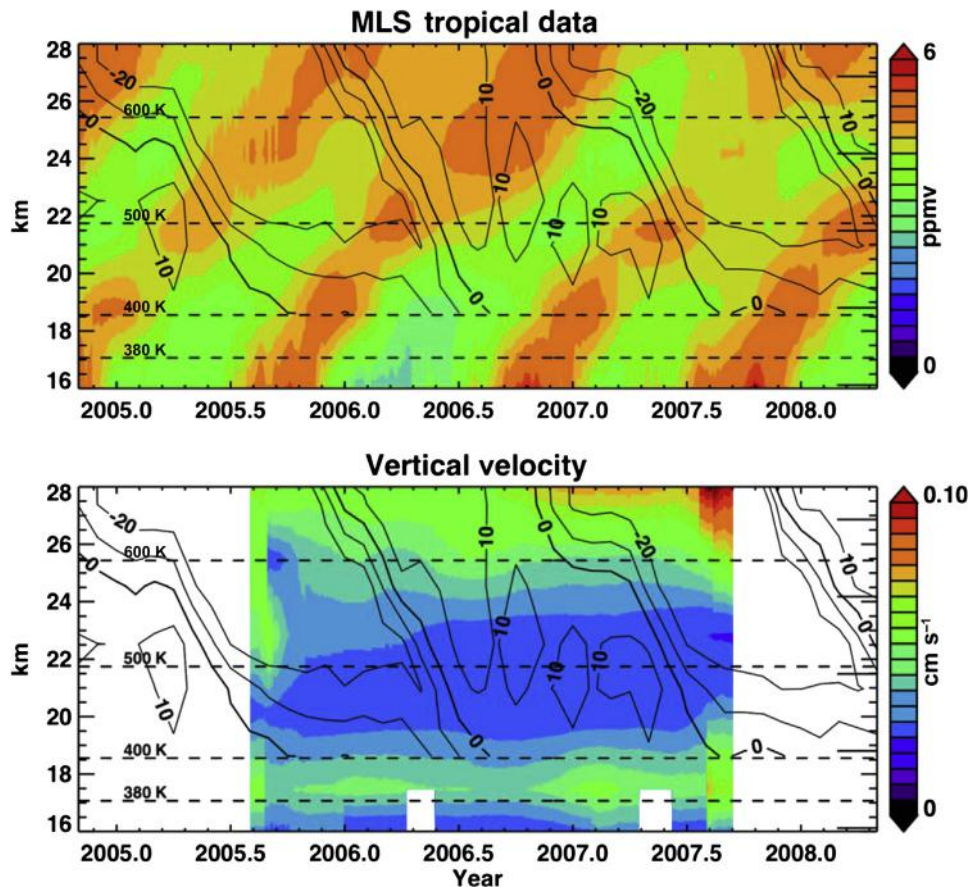


Figure 7 Top: 3.5 Years of tropical Aura MLS H₂O data showing the tape recorder effect (color contours). Bottom: Vertical velocities derived from the MLS H₂O data shown in the top panel (color contours). Approximate isentropic levels (dashed) and zonal (u) wind (solid) contours are overlaid in each panel. Easterly winds are positive. From Schoeberl, M.R., et al., 2008. Comparison of lower stratospheric tropical mean vertical velocities. *Journal of Geophysical Research* 113, D24109. <http://dx.doi.org/10.1029/2008JD010221>. Copyright 2008 by the American Geophysical Union. Reprinted with permission from the American Geophysical Union.

originated at the tropopause 1 year earlier. The vertical velocity is calculated from the upward motion of the cycles' minimum or maximum (bottom of Figure 7). Vertical diffusion is negligible and does not affect the location of the extrema. The amplitude of the cycle diminishes very slowly with height, indicating weak mixing of midlatitude air into the tropics. The tape recorder analysis provides empirical evidence for two key transport quantities: ascent rate and subtropical mixing strength as a function of height. These transport processes together effectively control mean age in the lower stratosphere.

High Latitude Descent and Isolation

The high latitude edge of the surf zone provides a similar study of the balance between rapid mixing in the midlatitude surf zone and the descending branch of the circulation. As in the subtropics, this balance results in steep gradients in long-lived trace gases at the boundary between the regions. Lower stratospheric aircraft missions in the late 1980s found very steep gradients in N_2O and other long-lived species at the edge of the Antarctic and Arctic vortices in winter, confirming the isolation of vortex air from the surf zone. Large O_3 loss occurs each austral spring inside the Antarctic vortex because of the meteorological conditions there. The first condition, low temperature, is required for the formation of polar stratospheric clouds (PSCs), which convert HCl and ClONO_2 into reactive forms of Cl while denitrifying the air. And the second condition, isolation, is necessary in order to maintain the highly perturbed chemistry required for PSC-driven O_3 loss. If there were strong mixing with midlatitude air, vortex air would be warmer, PSCs would evaporate, and HNO_3 would be resupplied to the vortex where it would consume reactive Cl .

The tape recorder shows that tropical air is isolated in all seasons, but the isolation of polar vortex air is a seasonal phenomenon. The vortex forms in the fall and breaks down in spring. The Antarctic vortex generally lasts until late spring (November) while the Arctic vortex, subject to far more wave disturbances than its Southern Hemisphere counterpart, has a final breakdown in early spring, March or April, and may experience a midwinter 'sudden stratospheric warming' (SSW). In an SSW, the existing vortex may be pushed off the pole to lower latitudes and destroyed by mixing in the surf zone. If a new vortex forms, which is likely if the SSW occurs before March, its composition will be that of the midlatitude air that replaced it. The degree of vortex isolation therefore has a major effect on Arctic stratospheric composition. The Antarctic vortex composition has less interannual dynamical variability due to weak wave driving during winter and early spring. The Arctic vortex composition shows significant interannual variability due to variations in wave driving, which affect descent as well as the permeability of the vortex edge. Both processes transport high O_3 into the polar region. Large interannual variations in the permeability of the Arctic vortex edge and circulation strength (i.e., descent) make the precise calculation of O_3 loss due to heterogeneous reactions (PSCs) difficult because the amount of O_3 transported into the vortex each winter is variable and difficult to quantify.

Figure 8 shows examples of interannual variability in the composition inside the Arctic vortex using ozone measured by the Aura MLS instrument. These data are from the 700 K

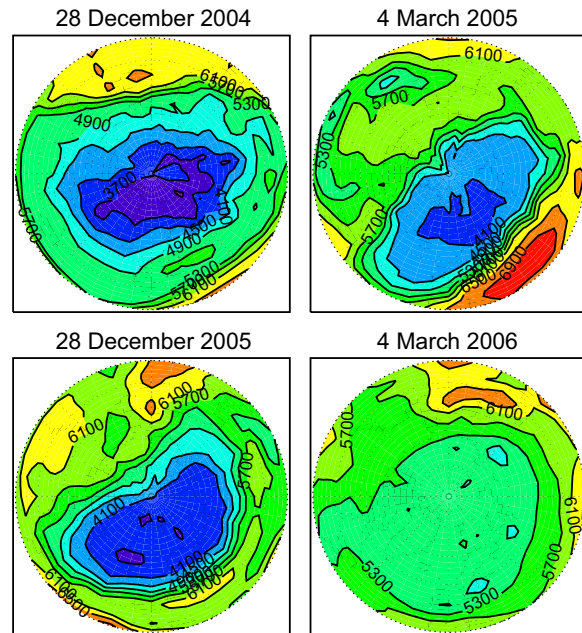


Figure 8 Left panels: Aura MLS O_3 (in ppb) on the 700 K surface (~ 15 hPa) in early winter (December 28) of 2004 and 2005. Right panels: Aura MLS O_3 on the same surface in late winter. Ozone loss from gas phase and heterogeneous chemical reactions is negligible at these levels. The large difference in March O_3 in these 2 years is due to transport.

(15–18 hPa) surface, which has little or no PSC activity. The observations shown in the left panels are from early winter of two consecutive years. The Arctic winter of 2004–05 (top panels) was very cold winter with little wave activity, while 2005–06 (bottom panels) was a relatively warm winter with frequent wave activity including a midwinter SSW. In early winter, both years show similar areas of low O_3 inside an intact, isolated vortex. Note that O_3 gradients at the vortex edge in December 2005 are steeper than 2004, a less-disturbed year. This is an example of how wave activity acts to erode and sharpen the vortex edge. The panels on the right show O_3 in early March. In March 2005, the vortex had been relatively cold and undisturbed by waves all winter while the vortex in March 2006 had recently reformed after an SSW. Ozone in March 2006 is more than 1 ppm (1000 ppb) higher than the previous year due to strong wave activity, which brought high O_3 air from the midlatitudes to the pole during winter. These differences reflect differences in advection and mixing (transport) rather than the effects of heterogeneous chemical ozone loss by PSCs.

Descent through the Extratropical Tropopause

Air leaves the stratosphere through the extratropical tropopause, approximately 30 – 90° in each hemisphere, as shown in Figure 1. The net transport by the BDC at the tropopause is downward and very little extratropical upper tropospheric air enters the stratosphere through upward motion. The tropopause acts as a transport barrier and this is confirmed through observed trace gradients at the tropopause. Carbon dioxide

(CO₂) has a ~4-ppm amplitude seasonal cycle in the Northern Hemisphere upper troposphere (UT) due to seasonal variations of sources and sinks. The amplitude of the CO₂ seasonal cycle just above the northern midlatitude tropopause is ~2 ppm and is phase lagged by several months relative to the UT cycle. If rapid mixing across the tropopause occurred, the seasonal cycles above and below the tropopause would be the same.

Improving Our Understanding of the Transport Circulation: Observations and Models

How Observations Affect Model Development and the Ability to Predict

Observational analyses such as the ones presented in the previous section provide a detailed understanding of key transport processes. The state of our understanding of the transport circulation is represented in three-dimensional (3D) chemistry climate models (CCMs). These models physically represent the chemical, dynamical, and radiative processes that control atmospheric composition. CCMs are used to predict how future changes in surface emissions of CFCs and greenhouse gasses (GHGs) may affect stratospheric O₃. In the late 1990s, several groups compared 2D (zonal mean) and 3D models against stratospheric observations in order to evaluate the realism of the models' representations of transport. Comparisons were made to mean ages derived from CO₂ and SF₆, N₂O vertical and horizontal gradients, and upward propagation of the tropical water tape recorder signal. The results revealed significant shortcomings in transport in most models, in particular, fast circulation and excessive subtropical mixing. Poor transport impacts the lifetimes of ozone-depleting substances (ODSs) in these models, which in turn affects

how quickly stratospheric O₃ is predicted to recover as the emissions of CFCs decrease.

Model evaluations using observations have led to more realistic model transport. In the past two decades, increases in computer speed and decreases in the cost of data storage have made it possible to produce simulations with higher spatial resolution that include processes that were previously parameterized or neglected (e.g., coupling between dynamics and chemistry, tropospheric chemistry). Although these computational improvements are necessary for longer and more complex simulations, new simulations would not advance our understanding or improve our ability to explain and predict if there were not observations to guide model development. Recently, a comprehensive effort was undertaken to assess the radiation, dynamics, chemistry, and transport in CCMs by using observationally derived process evaluation. The transport evaluation identified several key transport processes in the stratosphere that had to be well represented in models, including tropical ascent rates and the degree of mixing across transport barriers in the subtropics and high latitudes. Mean age is a function of both circulation and mixing strength, and is thus a useful model transport diagnostic. Figure 9 provides an example of how well recent CCMs simulate mean age in the lower stratosphere (blue lines) compared to a group of 2D and 3D models from the 1990s (yellow shading). Overall, present-day models have significantly more realistic transport compared to the older models.

The Importance of the Transport Circulation to Short-Lived Species (e.g., O₃)

Although ozone in the middle and upper stratosphere is controlled by the balance between photochemical production

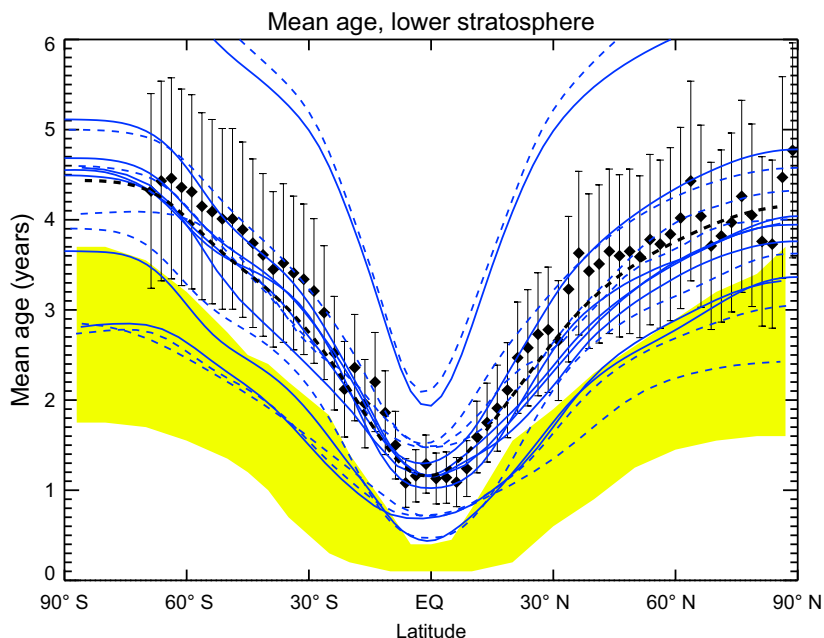


Figure 9 Comparison of observationally derived lower stratospheric mean age with models' simulations (~50 hPa). The range of model mean ages from 11 two-dimensional (zonal mean) and 7 three-dimensional models in the late 1990s is shown by the yellow shading. Blue solid and dashed lines show 15 CCMs participating in a model intercomparison in 2009. Observations and their 2 σ uncertainties are shown by the black diamonds.

and loss, the transport circulation plays an important role in the ozone distribution because it controls the distribution of many of the species involved in ozone loss processes. Radical species NO_2 and Cl play important roles in the catalytic loss cycles controlling O_3 in the middle and upper stratosphere. Their abundances depend on abundances of the reservoir ‘families’ they belong to, NO_y , and Cl_y . ‘ NO_y ’ is the sum of all reactive nitrogen species, mostly NO , NO_2 , HNO_3 , ClNO_3 , and N_2O_5 . ‘ Cl_y ’ is the sum of reactive chlorine species, primarily HCl , ClNO_3 , ClO , plus other minor species. The species within the family are in rapid photochemical equilibrium with each other, but their sum changes very slowly and is controlled by transport. If a model’s circulation is too fast, its mean age will be too young and its trace gases will have had less photochemical exposure; simulated NO_y would be lower than observed, assuming there are no errors in the chemical mechanism. Less NO_y means less NO_2 , which reduces O_3 loss. In the lower stratosphere, Cl_y affects O_3 through heterogeneous chemical loss inside the polar vortex during winter. Cl_y is created when CFCs are transported to the middle and upper stratosphere where they are photolyzed, releasing Cl . Thus Cl_y in the lower stratosphere depends on the transport circulation (i.e., the model’s ability to get CFCs up to appropriate altitudes). Models with young mean age at high latitudes are likely to have too little Cl_y , which reduces the amount of O_3 lost by PSC processes. The accurate simulation of stratospheric composition therefore depends not only on chemistry but also on the correct simulation of the transport circulation.

Understanding the Transport Circulation: Why It Matters?

Understanding the atmosphere’s present-day transport circulation is necessary in order to understand how changes in emissions of anthropogenic GHGs and ODSs will affect the ozone layer and the radiative balance of the atmosphere in the future (i.e., the climate response). Observations, such as vertical or horizontal profiles of stratospheric trace gases and their seasonal variations, reveal underlying physical processes occurring in the atmosphere. The empirical determination of transport processes based on observations provides constraints for model behavior, leading to improvements in the model’s

physical basis. If a model has a realistic representation of physical processes involved in the transport circulation, its response to changes in emissions will also be physically based, increasing the credibility of its composition and climate projections. Currently, some CCMs are able to realistically represent many essential chemical and transport processes in the stratosphere. Continued use of observations to understand stratospheric chemistry and transport processes will lead to greater realism in model simulations and greater confidence in their predictions of future circulation and composition.

See also: Gravity Waves: Overview. Middle Atmosphere: Planetary Waves; Polar Vortex; Quasi-Biennial Oscillation; Stratospheric Sudden Warmings; Zonal Mean Climatology.

Further Reading

- Andrews, A.E., Boering, K.A., Daube, B.C., et al., 2001. Mean age of stratospheric air derived from in situ observations of CO_2 , CH_4 , and N_2O . *Journal of Geophysical Research* 106, 32295–32314.
- Brasseur, G.P., Orlando, J.J., Tyndall, G.S. (Eds.), 1999. *Atmospheric Chemistry and Global Change*. Oxford University Press, New York.
- Brewer, A.W., 1949. Evidence for a world circulation provided by the measurements of helium and water vapour distribution in the stratosphere. *Quarterly Journal of the Royal Meteorological Society* 75, 351–363.
- Dobson, G.M.B., 1956. Origin and distribution of the polyatomic molecules in the atmosphere. *Proceedings of the Royal Society of London, Proceedings A* 236, 187–193.
- Hall, T.M., Waugh, D.W., Boering, K.A., Plumb, R.A., 1999. Evaluation of transport in stratospheric models. *Journal of Geophysical Research* 104, 18815–18839.
- Holton, J.R., Haynes, P.H., McIntyre, M.E., et al., 1995. Stratosphere-troposphere exchange. *Reviews of Geophysics* 33, 403–439.
- Plumb, R.A., 2002. Stratospheric transport. *J. Meteorol. Soc. Jpn.* 80, 793–809.
- Schoeberl, M.R., Douglass, A.R., Stolarski, R.S., et al., 2008. Comparison of lower stratospheric tropical mean vertical velocities. *Journal of Geophysical Research* 113. <http://dx.doi.org/10.1029/2008JD010221>.
- Stratospheric Processes and their Role in Climate (SPARC), 2010. SPARC Report on the Evaluation of Chemistry-Climate Models. In: Eyring, V., Shepherd, T.G., Waugh, D.W. (Eds.), SPARC Report No. 5, WCRP-132, WMO/TD-No. 1526. <http://www.atmos.physics.utoronto.ca/SPARC>.
- Waugh, D.W., Hall, T.M., 2002. Age of stratospheric air: theory, observations, and models. *Reviews of Geophysics* 40. <http://dx.doi.org/10.1029/2000RG000101>.

Zonal Mean Climatology

P Braesicke, Karlsruhe Institute of Technology, Karlsruhe, Germany

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by W J Randel, volume 3, pp 1358–1365, © 2003, Elsevier Ltd.

Synopsis

The term ‘middle atmosphere’ refers to the height region of approximately 10–90 km and is used to characterize the atmospheric region between the troposphere and the thermosphere. The stratosphere and mesosphere form the largest part of the middle atmosphere. This article describes slowly varying climatological features of temperature, zonal wind, and some other derived quantities. The Northern and Southern extra-tropical middle atmosphere will be contrasted in its seasonal behavior, highlighting similarities and differences during winter. The third region covered is the tropical middle atmosphere and the intra-seasonal and interannual variability determining its structure, for example, the quasi-biennial oscillation.

Introduction

The term ‘middle atmosphere’ refers to the height region of approximately 10–90 km and is used to characterize the atmospheric region between the troposphere and the thermosphere. The troposphere is dominated by weather systems and convection, and is generally well mixed. In the thermosphere, ionization and molecular diffusion become important physical processes and govern the behavior of the layer. The stratosphere and mesosphere form the largest part of the middle atmosphere and are the focus here.

This article will describe slowly varying climatological features of temperature, zonal wind (west–east wind, positive when eastward), and some other derived quantities. Much of the discussion will be using latitude–height (or pressure) cross sections of zonally averaged quantities (averaged over all longitudes at a particular latitude and height), because scaling relations in the middle atmosphere are such that the zonal wind is generally larger than the meridional wind (south–north wind, positive when northward). Note that the mean zonal flow is interconnected to the meridional (south–north) mean temperature gradient.

Notwithstanding the possibility of describing the structure of the middle atmosphere in zonal mean terms (two dimensional in space), deviations from the zonal mean are of particular importance during the winter in high latitudes. Planetary waves shape the structure of the stratospheric flow, and wave amplification and breaking waves can lead to rapid changes in the wintertime circulation (wave mean-flow interactions). Even though this behavior can be summarized in a zonal mean framework, in which large-scale temperatures, winds, and forcing mechanisms are linked, the details of this interactions (and their regional relevance) require a full three-dimensional assessment that goes beyond this contribution.

Most of this article will focus on describing zonal mean dynamics in three different latitude regions. The Northern and Southern extratropical middle atmosphere will be contrasted in its seasonal behavior, highlighting similarities and differences. The third region covered is the tropical middle atmosphere and the intraseasonal and interannual variability determining its structure.

Vertical Structure

The general vertical structure of the Earth’s atmosphere can be described using a mean temperature profile. Layers and significant levels can be identified by analyzing the height dependence of such a temperature profile (Figure 1). The lowest layer of the atmosphere is the troposphere (in Greek ‘tropos’ meaning turning) extending from the surface up to 8–17 km. Temperatures are decreasing with height and the point where the vertical gradient falls above a threshold of -2 K km^{-1} for at least 2 km is defined as the tropopause (this is the classic ‘World Meteorological Organization’ definition; other definitions also exist). The height of the tropopause varies with latitude and is lowest at the poles and highest in tropical latitudes. Above the

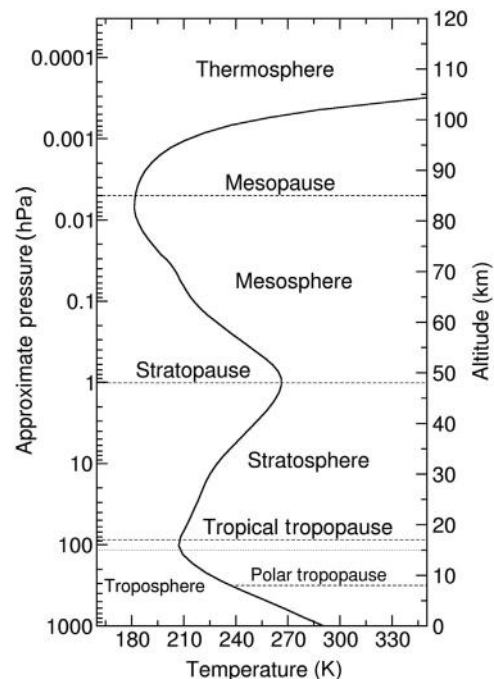


Figure 1 Mean temperature profile as a function of pressure/altitude. Data from Fleming et al., 1988.

tropopause temperatures are increasing with height forming the stratosphere (stratum: blanket). The stratosphere is the lower part of the middle atmosphere and extends to around 50 km. Its thermal structure is largely determined by the presence of ozone absorbing solar ultraviolet radiation. The temperature maximum in 50 km is called the stratopause. Above the stratopause temperatures are decreasing with height forming the mesosphere (in Greek 'mesos' meaning middle), the upper part of the middle atmosphere. The temperature minimum around 85 km is called the mesopause. Some 5–10 km above the mesopause temperatures start to increase rapidly. Here, in the thermosphere (in Greek 'thermos' meaning warm) region the governing physical processes change. Molecular diffusion, ionization, and ion drag become important players in determining the behavior of the layer.

In atmospheric science, height scales are often described using pressure. Geometric altitude (as used above) and pressure are related on a large scale by the hydrostatic approximation that states that the change of pressure with altitude is equal to the negative of density multiplied by the Earth's gravitational acceleration ($g = 9.81 \text{ m s}^{-2}$). Further assuming that air is an ideal gas, and using the ideal gas law together with the hydrostatic approximation, an exponential decay of pressure with height can be described. The pressure p at any given height z can be estimated by assuming a reference pressure p_0 and a constant scale height H : $p = p_0 e^{-z/H}$. The scale height H is defined as the specific gas constant for dry air ($R = 287 \text{ J kg}^{-1} \text{ K}^{-1}$) multiplied by temperature T and divided by g . Certainly, assuming a representative mean layer temperature for the middle atmosphere is a significant approximation, but works well for a general conversion between height and pressure. The value of H is usually assumed to be around 7000 m. Due to the exponential decrease of pressure (or density respectively) only $\sim 10\%$ of atmospheric mass are in the middle atmosphere.

Observations and Models

Most observations of the middle atmosphere are fairly recent. Today, measurements are obtained by balloon and rocket sondes, ground-based remote sensing (lidar or radar techniques), and satellite remote sensing. Routine balloon observations reaching the lower (stratospheric) parts of the middle atmosphere have been performed since the 1950s, mostly in the Northern Hemisphere and over land. Global temperature measurements from operational satellite instruments started around 1979 and have been complemented in recent decades by research satellite measurements. NASA's Upper Atmospheric Research Satellite contributed profoundly to our improved understanding of the middle atmosphere, measuring many quantities (including trace gases like ozone) from 1991 to 2002, as has European Space Agency's Envisat Mission from 2002 to 2012 and NASA's Aura Mission since 2004. In recent years, global meteorological analysis products have started to cover the lower part of the middle atmosphere, combining observational data and model forecasts in an optimal way to provide a best estimate of the state of the atmosphere. This is an ongoing development and expectations are that models will improve the coverage of the middle atmosphere in the future.

Zonally Averaged Meteorology

Starting from theoretical concepts the seasonal variation in the latitude–height structure of temperatures and zonal-mean zonal winds will be described. Derived quantities and their relation to the general circulation will be discussed, with a link to the full three-dimensional spatial structure of the atmosphere.

Theoretical Concepts

The vertical structure section introduced pressure as a vertical coordinate and explained the conversion between pressure and altitude using a constant scale height. Based on this approximation and the assumption that the horizontal motion in the middle atmosphere is quasi-geostrophic (large-scale horizontal winds result from a balance between pressure gradient and Coriolis force, and small deviations from this balance are considered), a set of two equations can be defined describing the averaged momentum and thermodynamic budget.

$$\frac{\partial \bar{u}}{\partial t} - f \bar{v}^* = \frac{e^{z/H}}{a \cos \phi} \nabla \cdot \mathbf{F}$$

$$\frac{\partial \bar{\Theta}}{\partial t} + \bar{\Theta}_z \bar{w}^* = Q$$

In this set of equations, $\bar{u}$ is the zonal-mean zonal velocity (west–east component of the three-dimensional velocity vector; positive for westerly (eastward) winds), f is the Coriolis parameter ($f = 2\Omega \sin \phi$; $\Omega = 2\pi/T$; $T = 86400 \text{ s}$), $a = 6.37 \cdot 10^6$ is the radius of Earth, t represents time, an overbar indicates a zonal mean, and the index z denotes a vertical derivative with respect to height. θ is the potential temperature of an air parcel. Assuming an adiabatic process, an air parcel that is moved from high altitudes to sea level would increase its temperature following the relationship

$$\Theta = T \left(\frac{p_0}{p} \right)^{R/c_p}$$

with p_0 being the pressure at sea level and the specific heat capacity at constant pressure for air $c_p = 1012 \text{ J kg}^{-1} \text{ K}^{-1}$.

The averaged momentum and thermodynamic budget equations (above) have two driving terms. The local tendency of the zonal wind (partial time derivative) is proportional to the divergence of a vector field $\mathbf{F}$ and the local (potential) temperature tendency is driven by a diabatic warming Q . Both equations have an advection term to balance the budget, with a meridional advection term for the wind and vertical advection term for the (potential) temperature tendency equation, where the (effective) meridional and vertical velocities are given by $\bar{v}^*$ and $\bar{w}^*$, respectively. This so-called residual mean circulation (or 'transformed Eulerian mean circulation' (TEM circulation)) is defined by

$$\bar{v}^* = \bar{v} - e^{z/H} \frac{\partial}{\partial z} \left(e^{-z/H} \frac{\bar{v}' \Theta'}{\bar{\Theta}_z} \right)$$

$$\bar{w}^* = \bar{w} + \frac{1}{a \cos \phi} \frac{\partial}{\partial \phi} \left(\frac{\bar{v}' \Theta'}{\bar{\Theta}_z} \cos \phi \right)$$

It is important to note that both residual velocity components, the meridional and the vertical, are linked through a continuity equation, even though this does not imply mass continuity in the system, and streamlines of the residual mean meridional circulation may in some places intersect the Earth's surface. For considerations regarding middle atmosphere dynamics this caveat is not of importance. The vector $\mathbf{F}$ is the so-called Eliassen–Palm (EP) flux and has two components

$$\mathbf{F} = \begin{pmatrix} F_{(\phi)} \\ F_{(z)} \end{pmatrix}$$

given by

$$F_{(\phi)} = -e^{-z/H} a \cos \phi \overline{u'v'}$$

$$F_{(z)} = e^{-z/H} a \cos \phi f \frac{\overline{v'\Theta'}}{\overline{\Theta}_z}$$

The meridional component ($F_{(\phi)}$) is proportional to a meridional momentum flux and the vertical component ($F_{(z)}$) is proportional to a meridional temperature flux. The divergence of this vector

$$\nabla \cdot \mathbf{F} = \frac{1}{a \cos \phi} \frac{\partial}{\partial \phi} (F_{(\phi)} \cos \phi) + \frac{\partial}{\partial z} (F_{(z)})$$

is the wave driving of the mean flow. Note that the horizontal wind and the temperature structure are linked and that this link can be expressed by a thermal wind relationship that relates vertical wind shear to the meridional temperature gradient. Therefore, the first-order driver for the overall circulation and its seasonal variation in the middle atmosphere is the radiative heating and cooling that is represented within Q . Ozone absorption of solar ultraviolet radiation is the major contributor to heating, while infrared emissions by carbon dioxide, ozone, and water vapor are cooling. The maximum heating is found in the summer hemisphere. The infrared cooling is mostly dependent on temperature and shows less latitudinal structure. The net radiative temperature change (cooling plus heating) has therefore a strong latitudinal structure, with net heating in the summer and net cooling in the winter hemisphere. This hemispheric asymmetry is the driver of the global circulation in the middle atmosphere.

Structures in Temperatures and Winds

The focus will be on the Northern and Southern Hemisphere winter seasons, and their similarities and differences in zonal-mean temperatures and zonal winds. This comparison works well in all regions of the middle atmosphere where the annual cycle is the dominant source of variability (extratropics), but does not help to elucidate tropical variability, discussed in the following subsection.

Figure 2 shows zonal mean temperatures for northern (left) and southern (right) hemisphere winter as a function of latitude and height (pressure). The tropopause, stratopause, and mesopause (from bottom to top) are indicated with small crosses. Areas of low temperatures (below 220 K) are shaded. In both seasons, the tropopause is highest at the equator and lowest at the poles, with a distinct temperature minimum around the tropical tropopause. Stratospheric polar temperatures are lower in southern winter compared to northern winter. The stratopause is highest in polar regions during winter. The lowest temperatures can be found around the summer mesopause in polar regions.

Figure 3 shows zonal-mean zonal winds for northern (left) and southern (right) hemisphere winter as a function of latitude and height (pressure). Distinct jet systems are visible in both hemispheres. Here, a jet is defined as a coherent structure with closed isolines for an absolute wind speed exceeding 20 m s^{-1} . The tropopause, stratopause, and mesopause are indicated by crosses as in **Figure 2**. Winds in the winter hemisphere are largely westerly (eastward) in the stratosphere and mesosphere. During summertime winds are easterly (westward) in this height regime. The tropopause intersects the tropospheric jets (westerly all year round). The wintertime (westerly) jets in the middle atmosphere tilt equatorward and are often referred to as Polar Night Jets (PNJs). Seasonal means cannot reproduce the wind structures in the tropical middle atmosphere well, because the wind variability is not dominated by an annual cycle. Instead variability is largest on a quasi-biennial or semiannual timescales (depending on height).

Note that even though the average flow in the middle atmosphere is westerly during wintertime, rapid deviations from the mean can occur and strong interannual variability of the PNJ on the Northern Hemisphere is well known. So-called

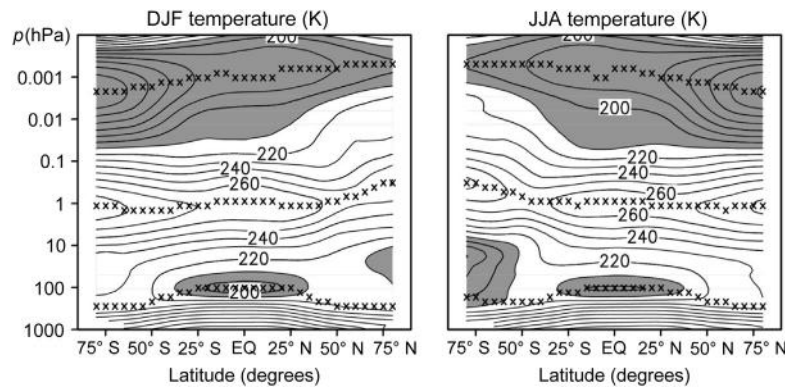


Figure 2 Zonal mean temperature as a function of latitude and pressure for December–January–February (DJF; left) and June–July–August (JJA; right). Crosses indicate the thermal tropopause, stratopause, and mesopause (from bottom to top). Data from Fleming et al., 1988.

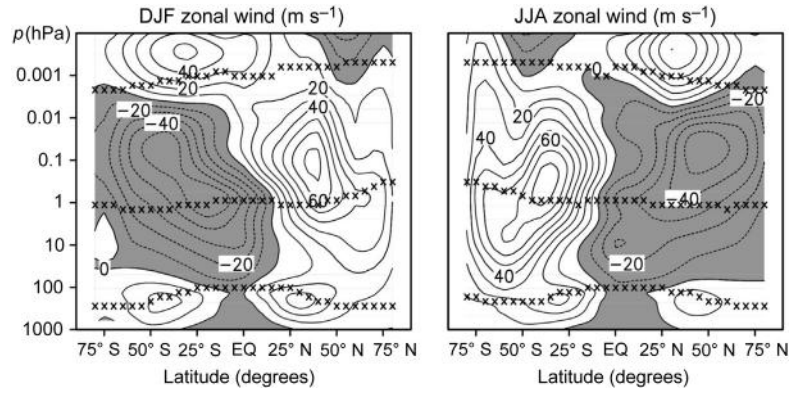


Figure 3 Zonal-mean zonal wind as a function of latitude and pressure for December–January–February (DJF; left) and June–July–August (JJA; right). Crosses indicate the thermal tropopause, stratopause, and mesopause (from bottom to top). Data from Fleming et al., 1988.

major stratospheric sudden warmings (SSWs) are linked to a reversal of the meridional temperature gradient, with stratospheric temperatures over the poles strongly increasing, and a wind reversal to easterlies in middle latitudes. The SSWs develop rapidly over a couple of days and recover over a time-scale of weeks and occur occasionally during Northern Hemisphere winter. Only one major SSW has so far been observed on the Southern Hemisphere during 2002.

Tropical Circulation (Quasi-Biennial Oscillation and Semiannual Oscillation)

In contrast to the extratropical middle atmosphere, the annual cycle is not a very dominant mode of variability in tropical latitudes. Intraseasonal and interseasonal variations are the more dominant modes of variability there. By looking at seasonal averages so far, most of the variability dominating the middle atmosphere in the tropics has been averaged out.

Figure 4 shows the zonal-mean zonal wind averaged over tropical latitudes as a function of time and height (pressure). Between 3 and 100 hPa, a quasi-periodic change of westerly and easterly (gray shaded) winds is apparent. This quasi-biennial oscillation (QBO) is largely driven by tropical waves and has an approximate periodicity of 26–28 months. The

easterly wind extremum (approximately -30 m s^{-1}) is usually stronger in magnitude than the westerly wind maximum ($\sim 15 \text{ m s}^{-1}$) and the successive phases of the QBO descend slowly ($\sim 1 \text{ km}$ per month) from the upper to the lower stratosphere. Generally, westerly winds descend more regularly and quicker than easterly winds that occasionally stop descending for a few weeks before continuing to propagate further down. Note that the westerly winds at the equator have greater angular momentum than the rotating Earth; this is an additional evidence for tropical wave forcing causing the QBO. There is a weak link to the seasonal cycle, with the onset of westerly winds in the upper stratosphere being somewhat linked to the descending westerly phase of the semiannual oscillation (SAO). The SAO near the stratopause manifest itself in a twice-yearly wind reversal (there is a mesopause SAO as well that is out of phase with the stratopause SAO) and is the dominant mode of variability in the middle atmosphere above 3 hPa for latitudes from 30° S to 30° N . It is forced by waves and momentum advection. The amplitude during the first half of the calendar year usually exceeds the amplitude in the second half. This is linked to the stronger extratropical wave forcing during the Northern Hemisphere winter, discussed below. The descending shear zones (westerly over easterly winds or vice versa) of the QBO modulate the thermal structure

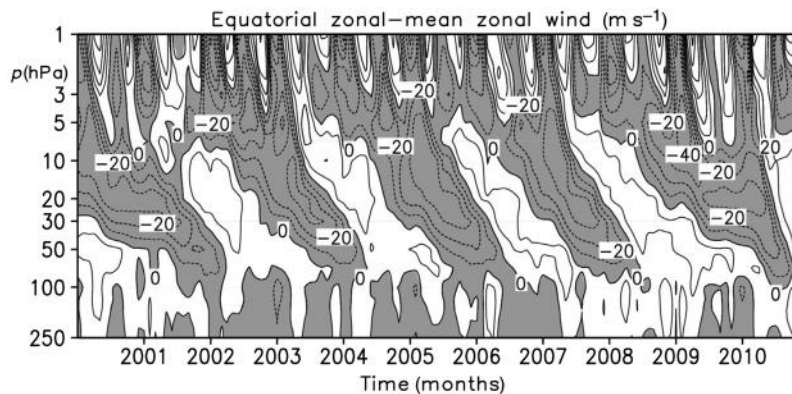


Figure 4 Equatorial zonal-mean zonal wind as a function of time and pressure. Data shown is monthly; year labels indicate beginning of the year; data from ERA-Interim. Data from Fleming et al., 1988.

of the tropical stratosphere and the ascending branch of the Brewer–Dobson circulation (BDC, discussed below). Even though the QBO itself is narrow in latitudinal extent (15°S – 15°N) it affects the middle atmosphere circulation on a hemispheric to global scale by, for example, influencing the propagation of planetary waves. Links between planetary waves and the hemispheric circulation patterns are discussed in the next section.

Large-Scale Waves and Forcing

Wave forcing and its impact on the mean circulation have already been mentioned in the theoretical section. This section will illustrate features of extratropical planetary waves in the lower middle atmosphere, highlighting important hemispheric asymmetries.

Figure 5 shows the zonal-mean zonal wind as a function of latitude and height (pressure) for both winter seasons (left) and the amplitude of planetary wave 1 (PW1; right) correspondingly. The wave amplitudes shown are the results of a Fourier decomposition of geopotential height (the height at which a certain pressure occurs). The gradient of the geopotential height determines the horizontal wind on a constant pressure surface (geostrophic approximation, see above). The data used to compile this figure is the ERA-Interim reanalysis data set. The data are produced by a state-of-the-art meteorological model, which is constrained using observational data to produce a best-guess of the state of the atmosphere at any given time. The current data set covers the period 1979 to present. So far, only the lower part of the middle atmosphere (stratosphere) is covered. As in Figure 3 the tropospheric and stratospheric jets are well captured. The hemispheric asymmetry in stratospheric jet strength is obvious, with winds exceeding 80 m s^{-1} on the Southern Hemisphere and only just reaching 60 m s^{-1} in the upper part of the stratosphere on the Northern Hemisphere. Planetary wave amplitudes are very different. The wave amplitude for a planetary wave of wave number 1 (PW1; one trough and one ridge along a longitude circle) is lower on the Southern Hemisphere ($\sim 400\text{ m}$) compared to the Northern Hemisphere. The vertical propagation of planetary waves from the upper troposphere to the stratosphere requires winds to be

westerly and not too strong (Charney–Drazin criterion). So the lower amplitude on the Southern Hemisphere is consistent with less excitation (the flow is more zonal, because of fewer orographic features in the Southern Hemisphere), and the stronger jet (the conditions for propagation are less favorable in the stratosphere). Certainly one leads to the other and larger wave amplitudes would cause a more disturbed and therefore a weaker averaged jet.

Even though a model can never fully capture reality, model output of a chemistry–climate model has been used to generate Figures 6 and 7. The model captures the principal climatological features well and provides a self-consistent description of the atmospheric flow. Here, the emphasis is on the linkages between different elements of the circulation and transport as described in the theory section.

Figure 6 shows the EP flux (see above for definition, scaled arrows) and the deceleration of the mean flow (isolines) due to resolved planetary waves as a function of latitude and height (pressure) for Northern Hemisphere winter (left) and Southern Hemisphere winter (right). The fluxes are generally upward and equatorward (where sizable with respect to the decreasing density), in line with the propagation properties of planetary waves. The deceleration is most prominent during Northern Hemisphere winter in the upper stratosphere and lower mesosphere, with much smaller values on the Southern Hemisphere during winter. This is consistent with the jet asymmetry discussed above; where deceleration is larger and the jet is weaker.

Structures in Meridional Circulation and Tracers

The residual circulation of the middle atmosphere cannot be measured directly. It can be inferred from observations with the help of models (e.g., using a data assimilation system), or it can be assessed analyzing the concentration of long-lived trace gases.

In the previous section, we focused largely on zonal-mean zonal winds and the related temperature structures. As discussed in the theory section, the slower effective vertical and meridional motion (south–north) is important as well and determines the distribution of long-lived trace gases like

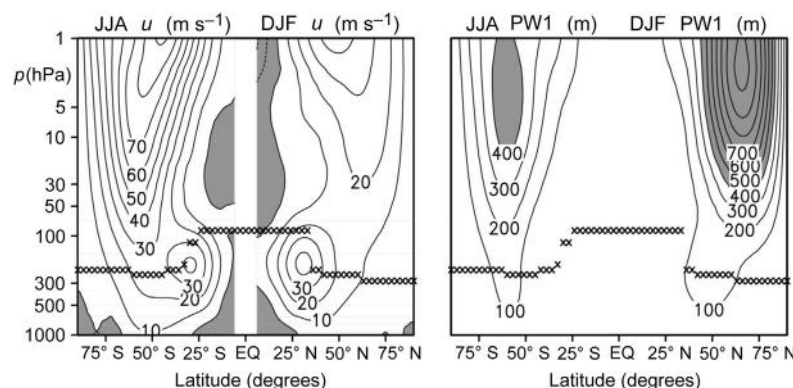


Figure 5 Zonal-mean zonal wind for the winter season as a function of latitude and pressure (left). Amplitude of planetary wave 1 (PW1) for the winter season as a function of latitude and pressure (right). December–January–February (DJF); June–July–August (JJA). Data from ERA-Interim. Data from Fleming et al., 1988.

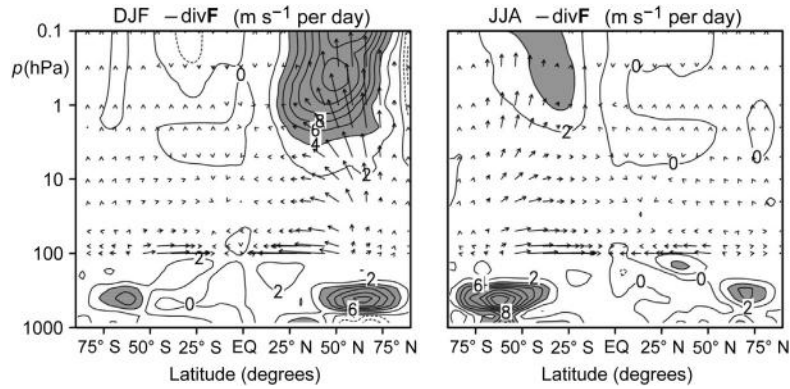


Figure 6 Eliassen–Palm flux vectors (arrows) and its convergence (isolines and shading) as a function of latitude and pressure for December–January–February (DJF; left) and June–July–August (JJA; right). Data from a chemistry–climate model for internal consistency with Figure 7. Data from Fleming et al., 1988.

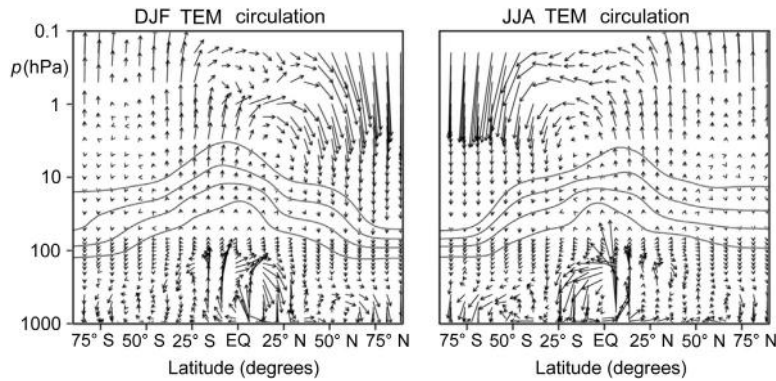


Figure 7 Residual circulation velocity vectors (scaled arrows) and isopleths of a long-lived tracer (isolines) as a function of latitude and pressure for December–January–February (DJF; left) and June–July–August (JJA; right). Data from a chemistry–climate model for internal consistency. Data from Fleming et al., 1988.

nitrous oxide (N_2O ; lifetime ~ 120 years) and methane (CH_4 ; lifetime ~ 10 years).

Figure 7 shows the TEM circulation (for definition see theory section above) with scaled arrows as a function of latitude and height (pressure) for the Northern (left) and Southern (right) Hemisphere winter. Overlaid are selected isolines of a long-lived tracer (N_2O). The above troposphere part of the TEM circulation is referred to as the BDC. It is shaping the tracer distribution of long-lived tracers in the stratosphere and mesosphere.

As can be seen from Figure 7, the ascending branch of the BDC is on the summer hemisphere, with a maximum in the subtropical middle atmosphere. Ascend is slow above the tropopause and accelerates toward the upper part of the stratosphere. While ascending, air flows predominantly to the winter pole where it descends. This overturning motion determines the meridional gradients of long-lived tracers in the winter hemisphere. Tracers with high tropospheric values (e.g., N_2O) indicate the ascending branch of the BDC by a relative concentration maximum at a given height. Isolines of tracer concentrations slope down poleward and their vertical gradient increases (Figure 7). The slope is not uniform and often shows a stepwise behavior in the subtropics and toward higher latitudes. This flattening of the slope is caused by quasi-horizontal

mixing. Even though we cannot measure the BDC directly as a velocity field, we can infer its behavior from trace gas observations. Probability density functions, as a function of potential temperature, can help to visualize different flow regimes, and trends of very inert tracers with clearly identified tropospheric sources (e.g., SF_6) can help to assess possible trends in the BDC. This is usually done by deriving a so-called age-of-air, a measure for the time an air parcel has spend in the stratosphere since entering through the tropical tropopause. Current models indicate that the BDC will strengthen under climate change (decreasing age-of-air), but strong observational evidence for this model result is still missing.

Concluding Remarks

This article summarizes the theoretical basis, observational evidence (including reanalysis data), and some model applications for characterizing and explaining the zonal mean structure and circulation of the middle atmosphere. The basic understanding of the middle atmosphere circulation is good, but some open questions remain, in particular regarding trends in the BDC. Satellite observations and advances in modeling (including data assimilation) will help to grow our

understanding further. Even though the zonal mean approach for describing atmospheric circulation is extremely useful, it is important to keep the full spatial complexity in mind when discussing rapidly changing flow regimes, for example (major) sudden stratospheric warmings.

See also: **Dynamical Meteorology:** Quasigeostrophic Theory; Wave Mean-Flow Interaction. **Middle Atmosphere:** Planetary Waves; Polar Vortex; Quasi-Biennial Oscillation; Semiannual Oscillation.

Further Reading

- Andrews, D.G., Holton, J.R., Leovy, C.B., 1987. *Middle Atmosphere Dynamics*. Academic Press, New York.
- Brasseur, G., Solomon, S., 1986. *Aeronomy of the Middle Atmosphere*. Reidel, Boston, MA.
- Labitzke, K.G., van Loon, H., 1999. *The Stratosphere: Phenomena, History and Relevance*. Springer-Verlag, Berlin.

MOUNTAIN METEOROLOGY

Contents

Overview

Cold Air Damming

Downslope Winds

Katabatic Winds

Land and Sea Breezes

Lee Vortices

Lee Waves and Mountain Waves

Orographic Effects: Lee Cyclogenesis

Valley Winds

Overview

RB Smith, Yale University, New Haven, CT, USA

© 2003 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 3, pp 1400–1405, © 2003, Elsevier Ltd.

Introduction

Mountain meteorology is the study of how mountains modify weather and climate. The subject is as old as meteorology itself. Aristotle's *Meteorologica* (c. 340 BC) included the (incorrect) speculation that mountains control the altitude range in which clouds form. Pascal's 1648 measurement of how the air pressure decreases with altitude on Puy de Dome in southern France addressed some of the most profound issues in meteorology: the weight and compressibility of air. Throughout the age of exploration, travelers and geographers described and tried to understand the various influences of mountains on climate: the temperature lapse rate (i.e., tree line, snow line, and high-altitude tundra and glaciers), reversed winds on hilltops, mountaintop clouds, wet–dry contrast across mountain ranges, the triggering of convection by hills, sheltering from winds on steep lee slopes, and gap winds. Some scholars would argue that the scientific study of mountain meteorology began with the extensive measurements of the physical conditions around Mt Blanc in the Alps by HB Saussure (1740–99) or with the reports on mountains and climate in South America and Asia by A von Humboldt (1769–1859). In the nineteenth century, most meteorology texts by authors in Europe and America contained sections on mountain climate, including those by Kaemtz from Halle (1844), Maury from Washington, DC (1855), Loomis from New Haven (1868), Tyndall from London (1872), Flammarion from Paris (1874),

and Ferrel from Kansas City (1889). In the twentieth century, as mathematical models of the atmosphere advanced, and our ability to observe the atmosphere quantitatively has improved, our understanding of the influence of mountains on the atmosphere has grown exponentially.

The subject of mountain meteorology is connected with the broader fields of ecology and geology in several ways. Many of the Earth's deserts are caused by the barrier effects of mountain ranges. Rain forests often occur on windward mountain slopes. Local ocean circulations are influenced by the cold air reaching the sea through mountain passes, and by the freshwater input to the oceans, channeled by the terrain. Major continental ice sheets have grown from small mountain glaciers. The shape and height of the mountains themselves are controlled by the intensity of orographic rain and snow, and the subsequent eroding action of stream flow and glacial scraping. Mountain meteorology is also important in human affairs, and thus there is continuing research to improve the forecasting of hazardous effects of mountains: severe winds, floods, avalanches, and air pollution.

The most obvious influence of mountains on climate and vegetation is the decrease of temperature with altitude along mountain slopes. The lapse rate along mountain slopes varies somewhat with latitude, season, and aspect, but typically takes a value of $-5\text{ }^{\circ}\text{C km}^{-1}$. Thus, a 4 km high mountain would be $20\text{ }^{\circ}\text{C}$ colder than the surrounding valleys and plains. In the midlatitude temperate zone, this difference gives the

mountaintop a polar climate, with tundra, stunted vegetation, and permanent snowfields. In the tropics, however, highland climates can be quite comfortable. In low-latitude Peru, for instance, highlands in the Andes are productively farmed for barley, corn, potatoes, and fava beans. In contrast, the eastern Peruvian lowlands have excessive temperature and rains, preventing proper soil development. The western Peruvian lowlands are an unproductive coastal desert.

The challenge and richness of the field of mountain meteorology are due in part to the fractal nature of the Earth's terrain. Major mountain ranges such as the Rockies, Andes, Himalayas, and Alps have horizontal dimensions of 1000 km, yet they contain within themselves a hierarchy of smaller scales down to at least 100 m: a factor of 10 000 in physical size. The heights of mountains vary from an arbitrary minimum of 100 m to nearly 10 km: a factor of 100. The orientations of mountain ranges also vary with respect to both the direction of the prevailing winds and the Sun's rays. The geographical position of each mountain range is highly significant, as the physical characteristics of the environment influence the way mountains modify the climate. For example, mountains in the tropics may trigger convection and thunderstorms owing to the unstable nature of a warm, moist air column, while high-latitude mountains will force smooth uplift with enhanced stratus rain and snow.

It is interesting to note that the height of the Earth's highest mountain (8.8 km) is only about one-tenth of one percent of the Earth's radius (6280 km). Thus, the Earth is nearly as smooth as the proverbial billiard ball. Given this fact, the importance of mountains on weather and climate is somewhat surprising. A partial answer to this paradox is that the Earth's atmosphere is also rather shallow with a density scale height of 8.5 km. Thus, the largest mountains reach to altitudes above most of the atmospheric mass. A more careful physical analysis allows us to identify four specific reasons for the importance of mountains in the atmosphere. These are discussed below.

Stable Stratification and Buoyancy Forces

A remarkable property of the atmosphere is its static stability, that is, an inherent resistance to vertical air motion. This stability arises from its typical temperature lapse rate ($\gamma = -6.5^\circ\text{C km}^{-1}$), which is greater than the adiabatic lapse rate ($\Gamma = -9.8^\circ\text{C km}^{-1}$). The magnitude of the static stability is characterized by the buoyancy frequency $N = [g(\gamma - \Gamma)/T]^{1/2} = 0.01\text{ s}^{-1}$, where g is the acceleration of gravity and T is the air temperature. A parcel of air displaced upward will return to its original level owing to buoyancy forces in a time $t = N^{-1}$: about 600 s or 10 min. The influence of this stability on airflow over mountains is an essential aspect of mountain meteorology.

The static stability of the atmosphere resists vertical motion, while mountain slopes try to generate vertical motion. If the mountain height is modest and the wind is strong, air will be able to climb the windward slope and reach the hilltop. In its effort to restore the air parcels to their original altitude, the buoyancy force causes the air to overshoot its equilibrium position, bringing the air rapidly down the lee slope and generating mountain waves. Through the action of mountain waves, the influence of the terrain may be felt at great vertical

and horizontal distances from the generating terrain. Mountain waves have been the subject of intensive study since the 1930s. They are of two types: vertically propagating and trapped waves. Vertically propagating mountain waves can be found at great altitudes above the mountain, even in the stratosphere (Figures 1 and 2). They usually have an irregular pattern, with a poorly defined wavelength longer than 15 km. Trapped mountain waves can occur under conditions when the Scorer parameter (the ratio of the buoyancy frequency to the wind speed) decreases with altitude, as is the case with a strong jet stream and reduced static stability in the upper troposphere. Trapped waves occur in the form of beautifully periodic waves, with wavelength from 8 to 25 km, extending 100 km or more downstream of a mountain ridge. Mathematically, trapped lee waves arise from a resonance in which internal gravity waves reflected downward by the jet stream are reflected upward again by the Earth's surface.

For higher hills or slower winds, the air may be unable to rise over the terrain or, if it does so, may generate nonlinear breaking mountain waves and turbulence. Over a long ridge, mountain wave breaking causes a transition from weak mountain waves to a strong downslope flow situation. Good examples are the Foehn in the Alps, the westerly Chinook windstorm in Colorado, and the northeasterly Bora over the Adriatic Sea. A long ridge may also create a barrier jet, an air current along the windward slope, to the left or right of the incident airstream, depending on whether the mountain lies in the Northern or Southern Hemisphere. Near isolated peaks or ridges with gaps, the airflow is forced to split and divert around the mountain causing corner or gap winds. The most famous flow splitting/gap flow phenomenon is the Mistral, a northerly wind reaching the Gulf of Lyon between the Pyrenees and the Alps. Severe downslope and gap winds can cause damage to crops and structures. Isolated high hills may generate vorticity by the action of breaking gravity waves or by sloping boundary layers. This vorticity sheds downstream to create steady or oscillating vortex wakes. The big island of Hawaii, during the steady summer Trade Winds, produces two large counter-rotating eddies in its lee (Figure 3). The small Caribbean island of St Vincent generates a long, straight wake extending for 400 km toward the coast of Central America.

Under the simplest of situations, a single parameter, the nondimensional mountain height ($H = hN/U$), is a controlling quantity in mountain airflow dynamics. In the definition of H , h is the mountain height, N is the buoyancy frequency, and U is the ambient wind speed. If $H < 1$, laminar mountain waves will exist. If $H > 1$, wave breaking or flow splitting will occur. As an example, if $N = 0.01\text{ s}^{-1}$ and $U = 10\text{ m s}^{-1}$ a mountain height of 500 m (i.e., $H = 0.5$) would generate weak mountain waves, while a mountain height of 2000 m (i.e., $H = 2$) would generate severe downslope winds or barrier jets and gap winds.

Condensation of Water Vapor

Another property of the atmosphere that sensitizes it to mountain effects is its high water vapor content. A typical relative humidity in the Earth's lower atmosphere is 70–80%. This level of humidity is maintained by a global balance between evaporation from the oceans and land, and

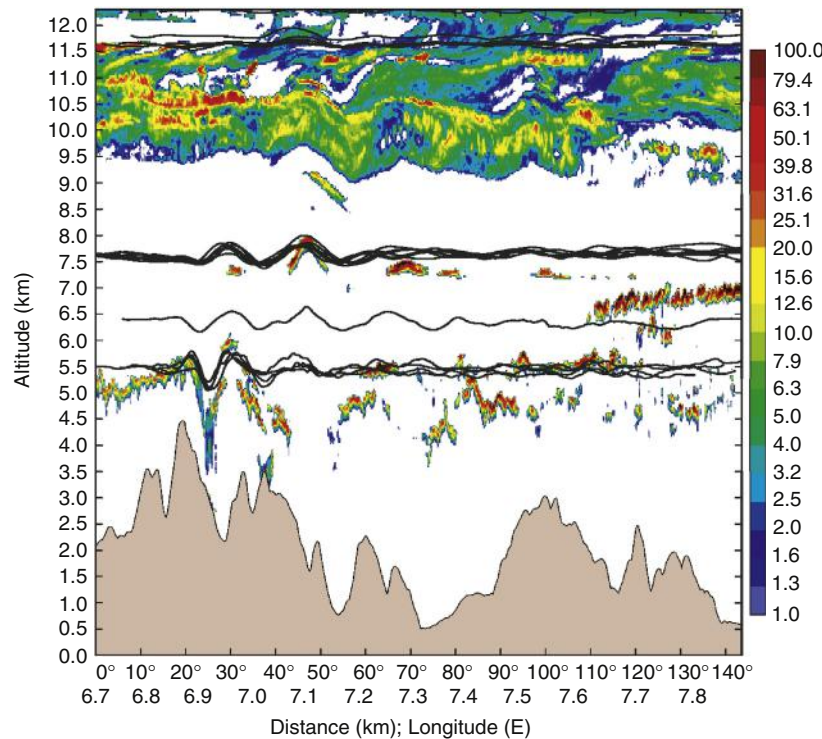


Figure 1 Mountain waves over Mt Blanc in the Alps. This diagram shows a vertical cross section of the troposphere aligned southwest to northeast over the peak of Mt Blanc. The ambient wind is from left to right in the diagram. Vertically propagating waves are seen from the disturbed patterns of clouds and from the vertical air motion sensed by a research aircraft. Color scale represents the intensity lidar backscatter, related to cloud particle density. Reproduced from Smith, R.B., Skubis, S., Doyle, J., et al. (2002) Mountain Waves over Mt. Blanc: The role of a stagnant boundary layer. *Journal of the Atmospheric Sciences* 59, 2073–2092.

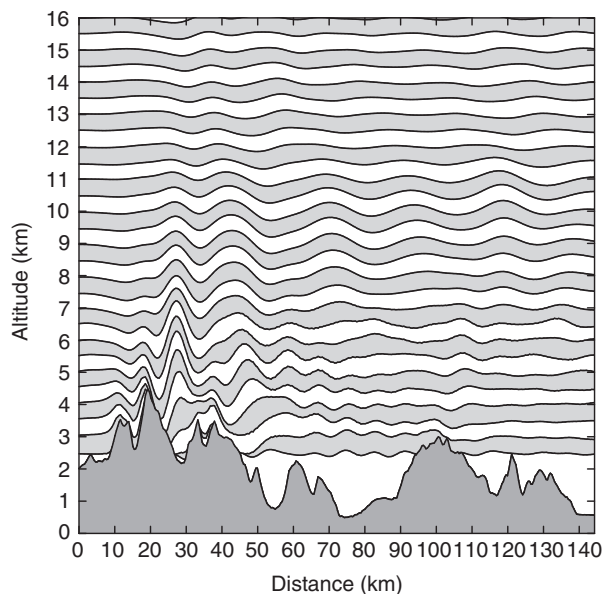


Figure 2 A theoretical calculation of mountain waves corresponding to the case illustrated in Figure 1. The lines of constant potential temperature show the patterns of vertical air motion as it passes over the Alpine terrain. Reproduced from Smith, R.B., Skubis, S., Doyle, J., et al. (2002) Mountain Waves over Mt. Blanc: The role of a stagnant boundary layer. *Journal of the Atmospheric Sciences* 59, 2073–2092.

precipitation from clouds. With such humid air, only a few hundred meters of uplift and adiabatic cooling are needed to bring the air to saturation. Thus, airflow over mountains is often associated with fog, clouds, and precipitation.

While the basic thermodynamics of the ascent of a moist parcel has been known for more than a century, the physics of orographic precipitation is poorly understood and probably varies considerably with physical scale and climatic setting. One key issue is whether clouds generated by mountain uplift can precipitate. In the 1950s, T Bergeron, a Swedish cloud physicist, pointed out that, for narrow hills and moderate wind speeds, air parcels spend so little time over the hill that there is insufficient time to convert the cloud droplets to larger precipitation-sized particles. Under this condition, orographic clouds cannot produce precipitation themselves but can only amplify existing broad-scale precipitation by a droplet scavenging process.

For midlatitude mountain ranges with widths exceeding 50 km (e.g., the Sierras, the Alps, the Southern Alps, and the Andes), the transit time across the range is sufficient for hydrometeors to form. During episodes of strong moist airflow against the range, the upslope regions experience heavy rain and/or snow. The leeward slopes in contrast experience dry, clear, descending flow – the rain shadow. In the northern midlatitude belt of westerlies, the Rocky Mountains lie across the prevailing airstream. Large annual precipitation is found on the western side (i.e., Oregon and Washington states), while

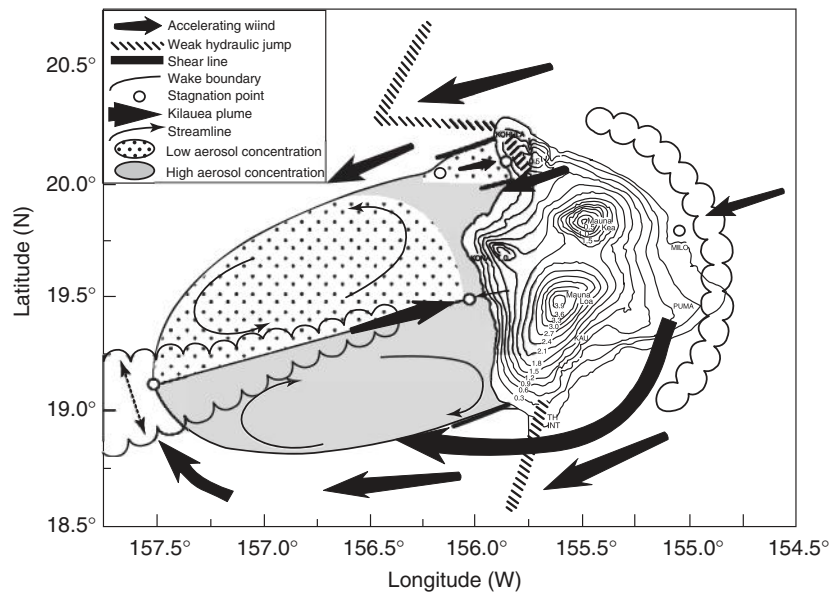


Figure 3 A schematic illustration of the eddies in the lee of Hawaii during the Trade Wind season. The patterns of airflow and aerosol concentration were mapped by research aircraft. Reproduced from Smith, R.B., Grubisic, V. (1993) Aerial observations of Hawaii's wake. *Journal of the Atmospheric Sciences* 50, 3728–3750.

a drier climate is found in the High Plains of Wyoming and Colorado. In Europe, the mountains of Scandinavia produce a wet climate in Bergen, Norway, and drier climates downstream in Oslo and Stockholm. The Alps, oriented in an east-west direction, experience wet-dry contrast during short periods of southerly or northerly flow, but no systematic climate contrast. No better example of this phenomenon exists than in the southern Andes, between latitudes 30° S and 50° S. Exposed to the prevailing westerlies in the Southern Hemisphere, the western side of the Andes receives several meters of annual rainfall, while the eastern side is dry. The south island of New Zealand provides another clear example of a wet-dry contrast. The quantitative estimation of how much water can cross over a high mountain range is still a topic of research. Spillover or drift of hydrometeors is one way in which water can bypass the effective 'cold-trap' presented by a high mountain. Another bypass mechanism is airflow through gaps. Air that finds a route through a gap is neither lifted as high nor chilled as much as air that rises over the highest peaks.

In lower latitudes, or in summer, the air column tends to be less stable to moist ascent, and the perturbing influence of mountains triggers convection. Examples of orographic triggering have been studied in Hawaii, Taiwan, and on the south side of the Alps in Italy. Convective showers over steep terrain are dangerous as the heavy rainfall can be channeled quickly into flash floods.

Solar Radiation

Another atmospheric response to mountains is caused by an uneven warming by solar radiation. An isolated mountaintop, for example, subjected to the Sun's rays, will normally develop a surface temperature that is warmer than the free atmosphere at the same elevation. The resulting horizontal temperature

gradient will generate a circulation between the hilltop and surrounding area. A good example of this thermal circulation is the daily buildup of convective clouds over the peaks in the Rocky Mountains in summer. Hikers are advised to reach the peaks by noon and start down, thus avoiding the electrical discharges from afternoon convective clouds. An opposite circulation might occur if the hilltop is snow-covered, so that it reflects most of the Sun's radiation.

Another terrain geometry of interest is the deep valley. Because of the volume displaced by the valley walls, the sunlight absorbed per unit horizontal area is used to warm a smaller volume of air than would be the case in flat terrain, thus raising the temperature more. This effect can generate a differential warming between the valley and adjacent plain, or between sections of the valley with different widths or wall steepnesses. The resulting temperature difference can generate an up-axis 'valley wind'. At night, the process is reversed and a down-axis 'mountain wind' can occur. These mountain-valley circulations are predictable and reliable, modulated mostly by the cloud cover shielding the Sun and by the seasons, varying the Sun angle.

Interesting enhancements and asymmetries in thermally driven mountain flows can occur if the Sun angle is low. Consider for example, an east-west ridge with north- and south-facing slopes, tilted at 22° from the horizontal. At a latitude of 45° north in the winter, the Sun lies about 68° south of the zenith point at local noon. Rays from the Sun would strike the south face of the ridge at 46° from normal incidence. The irradiance falling on the hillslope is given by the product of the solar constant (S) and the cosine of the angle between the Sun's rays and the direction perpendicular to the surface. For the example at hand, the irradiance on the sunlit slope is $S \cos(46^\circ) = (1380) (0.69) = 958 \text{ W m}^{-2}$. In contrast, the north-facing slope would experience tangential rays and thus collect zero irradiance. This differential heating between north- and south-facing slopes will generate its own thermal

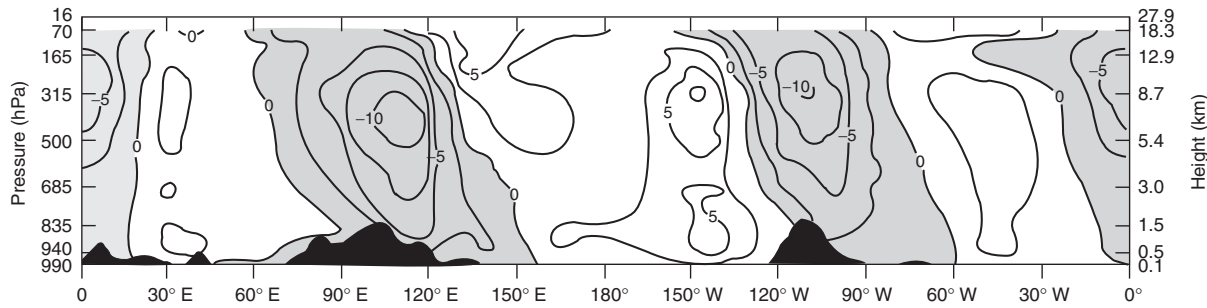


Figure 4 The stationary pattern of planetary waves induced by the surface irregularities of the Earth – oceans, continents, and mountains. These meandering bends in the jet stream and polar front control midlatitude weather. This diagram shows the southerly component of the wind vector, as a function of longitude and height. The contours are labeled with wind speed in meters per second. Reproduced from Manabe, S., Terpstra, T.B. (1974) The effects of mountains on the general circulation of the atmosphere as identified by numerical experiments. *Journal of the Atmospheric Sciences* 31, 3–42.

circulation, rising air on the heated slope and descending air along the slope in shadow. The heated air may rise above the hilltop to generate cumulus clouds.

Coriolis Force and Large-Scale Planetary Waves

One of the most interesting, but least understood, mountain effects on climate is the influence of major mountain ranges on the midlatitude planetary waves. The term ‘planetary wave’ here refers to the north–south meandering of the westerly jet stream and polar front around the globe (Figure 4). The wavelength of these waves ranges from 5000 to 20 000 km. The week-by-week location of the crests and troughs of the planetary waves modulates the temperatures and storminess of sites around the world.

It was suggested by J Charney and A Eliassen in the 1960s that the Rocky Mountains and the Tibetan Plateau could, through a Coriolis force mechanism, generate standing planetary waves in the Northern Hemisphere that would have their own influence on global climate and would interact with drifting planetary waves generated by other processes. This suggestion has received considerable attention from researchers using general circulation models (GCMs), including a clever mountain (M)/no-mountain (NM) comparison technique. The basis of this method is a supercomputer-based GCM that has been tested for its ability to predict, from first principles, the statistics of the planetary wave structure in the Earth’s atmosphere. With such a numerical model, it is a relatively simple matter to remove the Earth’s mountains from the lower boundary conditions of the model. Comparing the M and the NM runs allows one to determine the role of the mountains in the distribution of climate. The results indicate that, indeed, the amplitude and phase of planetary waves are significantly influenced by the major mountain ranges.

This result is of more than theoretical interest. Over the 4.6 billion year history of the Earth, mountain belts have been created and destroyed by repeated cycles of crustal plate collision and erosion. Thus, in addition to other climate influences such as the variations in solar intensity, orbital parameters, and atmospheric composition, the changing distribution of mountains could have modified the Earth’s climate. On the shorter timescale of a hundred thousand years, large continental glaciers

have cyclically grown and decayed. A good example is the Laurentide ice sheet in eastern Canada. These massive ice sheets behave much like mountain ranges, altering both local and global climates. The Greenland ice cap, with its peak altitude near 2800 m, plays a significant role today in modifying weather patterns and trapping water vapor in the North Atlantic region.

In the future, mountain meteorology will continue to challenge scientists with the chaotic nature of its fluid dynamics and the sensitivity of orographic effects to the precise details of the ambient flow and solar radiation fields. Powerful new observational tools such as remote sensing radar and lidar will allow three-dimensional airflow and cloud fields to be mapped. Numerical models with ever-increasing spatial resolution, run on supercomputers, will stimulate research on nonlinearity and the complex interactions between fluid dynamics, cloud physics, and radiation. If successful, these investigations will lead to improved weather prediction and increased understanding of ancient and modern climates.

See also: **Dynamical Meteorology:** Hydraulic Flow; Overview; Static Stability; Stationary Waves (Orographic and Thermally Forced); Vorticity. **Mesoscale Meteorology:** Cloud and Precipitation Bands; Mesoscale Convective Systems. **Mountain Meteorology:** Downslope Winds; Katabatic Winds; Lee Waves and Mountain Waves; Valley Winds. **Numerical Models:** Mesoscale Atmospheric Modeling. **Thermodynamics:** Saturated Adiabatic Processes.

Further Reading

- Baines, P.G., 1995. *Topographic Effects in Stratified Flows*. Cambridge University Press, Cambridge, UK.
- Blumen, W. (Ed.), 1990. *Atmospheric Processes over Complex Terrain*. Meteorological Monographs, vol. 23/45. American Meteorological Society, Boston, MA.
- Peattie, R., 1936. *Mountain Geography*. Harvard University Press, Cambridge, MA.
- Price, L., 1981. *Mountains and Man*. University of California Press, Berkeley, CA.
- Smith, R.B., 1979. The influence of mountains on the atmosphere. In: Saltzman, B. (Ed.), *Advances in Geophysics*, vol. 21. Academic Press, New York, NY, pp. 87–230.
- Smith, R.B., 1989. Hydrostatic airflow over mountains. In: Saltzman, B. (Ed.), *Advances in Geophysics*, vol. 31. Academic Press, New York, NY, pp. 1–41.
- Whiteman, C.D., 2000. *Mountain Meteorology: Fundamentals and Applications*. Oxford University Press, New York, NY.
- Wurtele, M.G., 1996. Atmospheric lee waves. *Annual Reviews of Fluid Mechanics* 28, 429–476.

Cold Air Damming

BA Colle, Stony Brook University – SUNY, Stony Brook, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

Cold air damming (CAD) occurs when cold, stable air becomes topographically blocked by relatively long, continuous mountain ranges. The cold air becomes rotationally trapped against the mountain to form a pressure wedge and low-level jet parallel to the barrier. The cold dome can be further modified by diabatic processes within the planetary boundary layer. This article summarizes the structure and evolution of CAD. The large-scale flow patterns that favor CAD are highlighted, with particular emphasis on the Appalachian Mountains and Rocky Mountains. Finally, some of the societal impacts and forecast challenges are discussed.

Introduction

Cold air damming (CAD) is the process whereby cold, stable air becomes topographically blocked and channeled adjacent to a mountain range. This results in a cold dome along the barrier, which can often be identified by a pronounced U-shaped 'wedge' in the sea-level pressure pattern (Figure 1(a)), and an enhancement in the terrain-parallel flow adjacent to the mountain slope (Figure 1(b)), commonly referred to as a barrier jet. The temperatures within the damming region are often 5–10 °C cooler than surrounding areas of equal elevation (Figure 1(a)). Damming events typically occur during the cold season when an along-barrier pressure gradient is created by a cold anticyclone passing to the north (south) of the barrier slope in the Northern (Southern) Hemisphere. When the favorable synoptic conditions persist, the cold air mass is often observed to remain in a quasi-steady state.

Long, continuous mountain ranges obstruct the low-level flow of cold air and can result in CAD. The structure and evolution of CADs are determined by the ambient flow speed and direction approaching the barrier, terrain height, low-level density stratification, diabatic effects, and latitude. The terrain-normal wind speeds, stratification, and terrain height determine whether the low-level flow is blocked by the barrier. Since CAD events typically last more than several hours, planetary rotation limits the horizontal extent of the damming and determines the evolution of the momentum balances near the terrain. Diabatic effects such as evaporative cooling and melting can also modify the CAD evolution.

CAD is common along many mountainous regions around the world such as the Appalachians and Rockies in North America. The development of the cold dome during these events results in several operational forecast challenges such as precipitation type (freezing rain, sleet, snow), precipitation enhancement along 'coastal fronts' and from ascent over the cold dome, stratus/fog, and unseasonably cold weather. Numerical models at moderate resolution (<30 km grid spacing) can resolve CAD events, but there are forecast challenges associated with the strength and dissipation of these events.

Nondimensional Flow Parameters

CAD is dependent on the amount of cold air blocked by an elongated mountain range. Flow blocking by a barrier is related to the ratio of the kinetic energy of the ambient flow to the potential energy needed to get a near surface air parcel over the mountaintop. This ratio is given by the Froude number, Fr , where

$$Fr = U/(NH). \quad [1]$$

In eqn [1], U is the speed of the average flow impinging toward the barrier below crest level, N is the Brunt–Väisälä frequency, and H is the mountain height. Fr^2 may be interpreted as the ratio of inertial forces, $U(U/H)$, to buoyancy forces, N^2H . The square of the stability, $N^2 = g/\theta_o(\Delta\theta/\Delta z)$, for these events is often given as the reduced gravity ($g' = g/\theta_o(\Delta\theta)$) divided by the depth over which $\Delta\theta$ is evaluated, since there is frequently an inversion capping the CAD. The average potential temperature beneath the inversion is given by θ_o . A large Froude number ($Fr > 1$) indicates that the flow has enough kinetic energy to surmount the barrier, while low Froude number ($Fr < 1$) favors blocked flow along the barrier. Many CAD episodes are characterized by an Fr less than 0.5, which is an indicative of significant flow blocking.

In the rotating limit (i.e., when Coriolis effects are to be considered), theoretical studies have shown that the Burger number ($B = HN/Lf$, where L is the mountain half-width and f is the Coriolis parameter) is also an important parameter to diagnose potential blocking. For situations of equal stratification and latitude, the Burger number is proportional to the steepness of the mountain slope (H/L). When $B < 1$, the flow is quasi-geostrophic/semigeostrophic and the flow can pass over the mountain; however, when $B > 1$, the barrier is considered to be 'hydrodynamically steep' and the flow is blocked by the barrier and begins to flow down the pressure gradient (Figure 2(a)).

When $Fr < 1$ or $B > 1$, blocked ageostrophic downgradient flow is forced by an along-barrier pressure gradient imposed by the synoptic-scale pressure field (Figure 2). Using scale analysis on the momentum equations, when l/L (cross-barrier length scale/along-barrier length scale) is small and the Rossby number $R_l = V/fl$ (where V is the magnitude of the along-barrier flow) is approximately equal to or greater than one, the winds

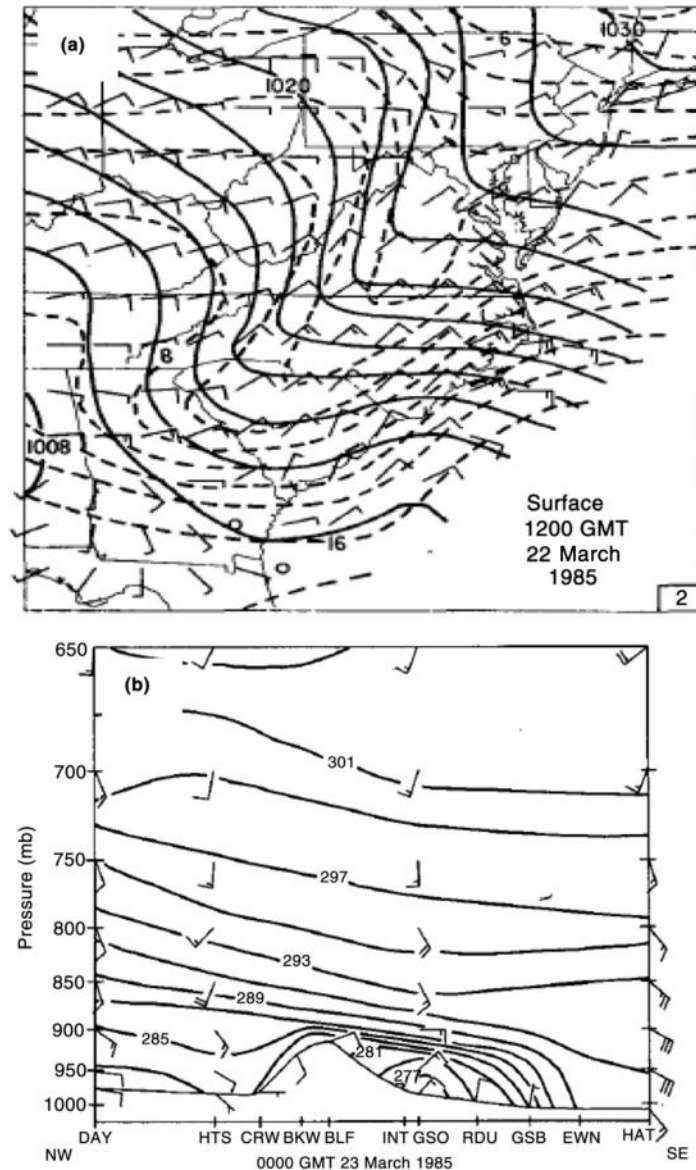


Figure 1 CAD example to the east of the Appalachians showing (a) a surface map of winds (full barb = 10 kts), temperatures (dashed lines in °C), and sea-level pressure (solid lines in hPa), as well as (b) a northwest to southeast cross section of horizontal observed winds and potential temperature (°K) for 1200 UTC 22 March 1985. Figures 6(c) and 14(b) in Bell, G.D., Bosart, L.F., 1988. Appalachian cold-air damming. *Monthly Weather Review* 116, 137–161. © American Meteorological Society. Reprinted with permission.

are primarily terrain-parallel within approximately a Rossby radius of deformation, $l_R = (g'H)^{1/2}/f$, of the barrier. For example, for the Appalachian Mountains of North America, l_R is typically around 150 km (where $H \sim 1000$ m, $g' \sim 0.25 \text{ m s}^{-2}$, and $f \sim 10^{-4} \text{ s}^{-1}$).

Within a Rossby radius of the terrain, an antitriptic balance typically develops in the along-barrier direction between friction and the pressure gradient (Figure 2(b)). Meanwhile, in the cross-barrier direction the mass field adjusts under the influence of rotation (e.g., geostrophic adjustment) such that the pressure gradient normal to the terrain balances the Coriolis force associated with the terrain-parallel winds. Since the Rossby radius depends on the latitude (as defined by f), CAD is

able to extend further away from the barrier closer to the equator given identical stratification and mountain height.

Structure and Evolution

Antecedent Conditions

The interaction of the synoptic-scale flow with an elongated mountain barrier acts to initiate CAD. Typically, significant damming events are associated with an upper level trough and associated cold surface anticyclone crossing immediately to the north (south) of the barrier in the Northern (Southern) Hemisphere (Figure 3(a)). (For purposes of illustration, only

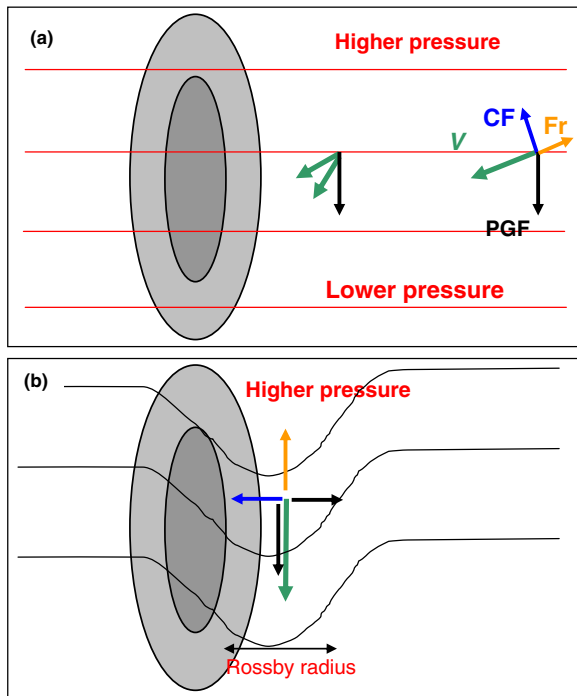


Figure 2 Illustration of the development of flow blocking in which (a) an along-barrier pressure gradient develops adjacent to a mountain (Northern Hemisphere) and a parcel well upstream of the mountain is nearly geostrophic; (b) subsequently, through flow blocking and geostrophic adjustment, a mesoscale pressure ridge develops adjacent to the mountain within a Rossby radius of deformation, and the flow has an antitropic balance in the along-barrier direction and nearly geostrophic in the cross-barrier direction.

the evolution CAD in the Northern Hemisphere is shown.) For damming along the Appalachians, there is often split westerly flow aloft, with a secondary upper level trough and associated surface cyclone situated to the southwest of the barrier. For other barriers such as the Rocky Mountains and Andes, significant cross-barrier flow at crest level to the south of the anticyclone often results in downslope warming and a lee trough (Figure 3(a)). For either situation, this high–low pressure couplet enhances the along-barrier pressure gradient.

Initiation and Mature Phases

When the low-level cold air around the anticyclone becomes blocked along the slope ($Fr < 1$), the along-barrier pressure gradient results in ageostrophic nearly terrain-parallel flow adjacent to the barrier (Figure 3(a) and 3(b)), which advects cold air southward along the barrier slope. Meanwhile, above the mountain there may be neutral or even warm advection. Because of the significant cold air source to the north, cold advection is typically much larger than upslope adiabatic cooling during damming events. The rapid decrease in cold advection with height results in a shallow cold air dome capped by an isothermal or inversion layer, which increases the stability and enhances the flow blocking. Meanwhile, a Coriolis torque acting on this downgradient flow results in an upslope flow component near the barrier slope, which helps push cold

air up the mountain slope. For some cases, the synoptic-scale low-level flow will also have a geostrophic wind component directed toward the barrier.

Because of the significant low-level stability and significant mountain slope, the Froude number (Fr) is typically $\ll 1$ and Burger number (B) is $\gg 1$, which results in blocking of the cold air before it reaches the mountaintop. This blocking prevents the development of geostrophic balance in the along-barrier direction and a wedge of cold air (cold dome) develops adjacent to the mountain. This cold dome hydrostatically results in a pressure ridge along the barrier, which extends a Rossby radius (l_R) upstream of the crest (Figure 3(c)). In the along-barrier direction, the flow becomes balanced by the pressure gradient and friction processes. Because of the strong static stability and weak vertical mixing, most of the friction is the result of surface drag. In the cross-barrier direction, a geostrophic balance develops between the cross-barrier pressure gradient and the Coriolis force.

To demonstrate this adjustment process and vertical structure of the damming process, Figure 4 shows the horizontal flow and thermal evolution in a cross section taken normal to the mountain range. Prior to the CAD event, the isentropes are nearly horizontal adjacent to the mountain with fairly weak stability near crest level. For many cases, there may be a preexisting cold stable layer from a previous damming event or a depression in the isentropes from adiabatic warming associated with downslope flow (Figure 4(a)). At this initial time, there is little flow in the along-barrier direction. Once the along-barrier pressure gradient becomes more established, ageostrophic northerlies develop and the Coriolis force turns the flow in an upslope direction (Figure 4(b)). The isentropes begin to tilt in response to the cold air being advected up the barrier. The low-level cold advection and weak upslope cooling combined with neutral or warm advection above results in a stable layer intensifying near the top of the cold dome.

During the next few hours, the mass field associated with the blocked flow adjusts to planetary rotation and extends approximately a Rossby radius upstream of the barrier (Figure 4(c)). By this time, CAD and the cold dome are firmly established adjacent to the barrier. The cold dome is deepest near the bottom of the slope, which hydrostatically corresponds to the axis of the sea-level pressure ridge (Figure 3(b)). The damming regime is often characterized by neutral or conditionally unstable lapse rates in the lower layer (cold dome) and is capped by an inversion. The terrain-parallel flow within this cold dome is maximized where the greatest tilt of the isentropes and resulting cross-barrier pressure gradient exist. The core of maximum winds (labeled J in Figure 4(c)), which can frequently exceed 20 m s^{-1} , is often referred to as a ‘barrier jet.’ The development of the barrier jet feeds more cold air into the damming region, therefore helping to sustain the cold dome.

The cold dome and barrier jet associated with damming persist as long as there is a cold air source feeding the cold dome. Meanwhile, strong warm advection may occur to the east and over the cold dome (not shown), which increases the inversion at the top of the cold dome and reduces the vertical mixing between the CAD layer and the ambient flow above. There is often ascent of relatively warmer air up and over the

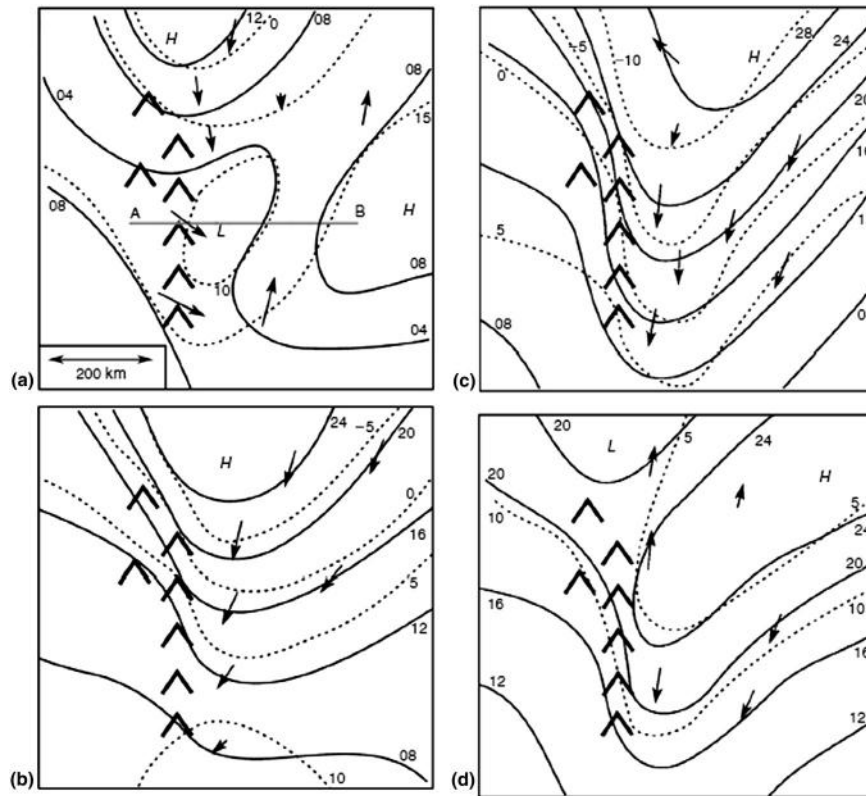


Figure 3 Horizontal evolution of CAD for the (a) initiation, (b) development, (c) mature, and (d) decay stages. Sea-level pressure (solid, every 4 mb), temperature (dashed, every 5 °C), and surface winds are shown. Line segment AB in (a) is the location of the cross section shown in Figure 4.

CAD cold dome, which can enhance the precipitation upstream of the mountain slope. For some events such as Appalachian damming, precipitation falling from the saturated flow over the cold dome results in evaporative cooling and further strengthening or maintenance of the cold dome.

Dissipation Phase

CAD dissipates when the cold surface anticyclone moves away from the barrier, and the cold air source is lost (Figure 3(d)). There are five processes that can contribute to CAD weakening: (1) Cold advection just above the mountain crest can reduce the strength of the inversion capping the cold dome and thus reduce the potential for flow blocking; (2) Solar heating, if there are limited clouds within the cold dome; (3) Near surface divergence, as pressure falls occur to the north of the barrier and the developing southerly flow is directed more opposite from the terrain-parallel flow to the south (in the Northern Hemisphere). This divergence can also reduce the cloud cover and increase the weakening from solar heating; (4) Shear-induced mixing of potentially warmer air from aloft can weaken the cold dome from the top down; and (5) An approaching warm or coastal front that mixes out the relatively shallow cold dome. In some locations, particularly for large mountain barriers such as the Rockies and the Himalayas, the cold dome may be eroded by significant

cross-barrier and downslope flow near crest level (e.g., ‘Chinook’ or ‘Foehn’ winds).

CAD Types

A few CAD damming evolutions have been identified for the Appalachian Mountains, which are related to the intensity of the CAD and the role of diabatic processes (Figure 5). The classification includes (1) Classical damming initiated by the interaction of synoptic-scale features with topography, but little diabatic processes; (2) Classical events that develop more *in situ* as a result of diabatic cooling processes (evaporation as precipitation falls into dry air); and (3) Hybrid events in which both (1) and (2) are prevalent.

The classical CAD has a large anticyclone located over the Northeast United States during the peak strength of CAD (Figure 5(b)). This surface high pressure is associated with an anomalous shortwave ridge at 500 hPa from southeast Canada to the southeastern US coast (Figure 5(d)), and a midlevel trough over the Central United States. The jet streak at 250 hPa extends eastward from the Northeast United States in a region of confluent flow to the east of the upper level ridge axis (Figure 5(a)). A surface high-pressure ridge extends southward to the east of the Appalachians (Figure 5(b)), with nearly terrain-parallel flow near the surface (not shown). Meanwhile,

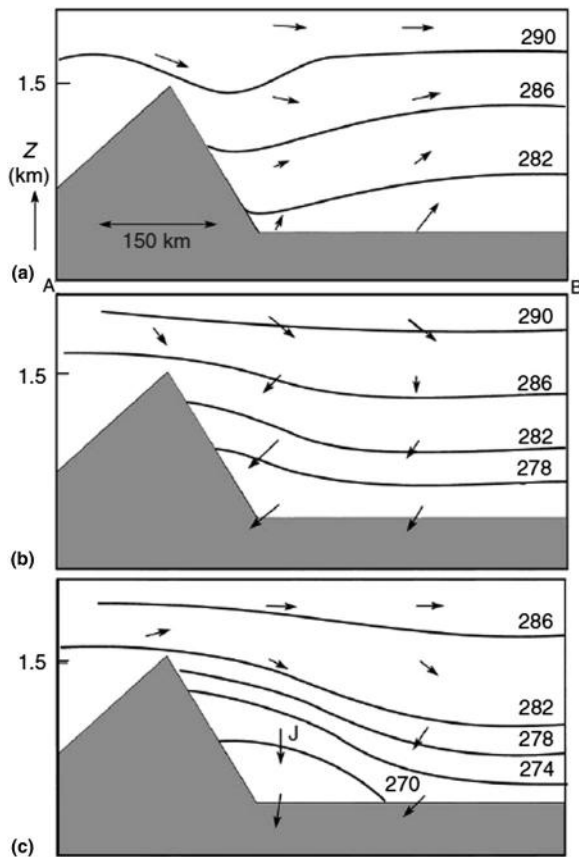


Figure 4 Vertical evolution of CAD for the (a) initiation, (b) development, and (c) mature stages for cross section AB showing the potential temperature (solid, every 4 K) and horizontal winds in the section. The location of AB is shown in [Figure 3\(a\)](#). The letter 'J' shows the position of the terrain-parallel barrier jet.

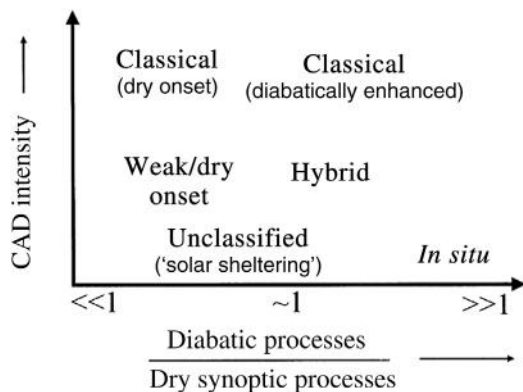


Figure 5 The CAD spectrum defined with respect to event intensity and the relative contribution of synoptic-scale dry forcing and diabatic processes. From Figure 2(b) in Bailey, C.M., Hartfield, G., Lackmann, G.M., Keeter, K., Sharp, S., 2003. An objective climatology, classification scheme, and assessment of sensible weather impacts for Appalachian cold-air damming. *Weather Forecasting* 18, 641–661. © American Meteorological Society. Reprinted with permission.

the pressure ridge at 850 hPa is centered over the Northeast United States and to the east of the Southeast United States ([Figure 5\(c\)](#)), with southwesterly flow and warm advection over the CAD.

The large-scale pattern for the hybrid CAD events is similar to that of the dry CAD events, although there are some distinct differences that result in more diabatic effects in the hybrid events. The shortwave ridge at 500 hPa is less amplified and anomalous than the dry events ([Figure 6\(b\)](#)), and there is a coupled jet structure at 250 hPa ([Figure 6\(a\)](#)), with the east of the Appalachians located in the right jet exit region along the Gulf Coast and right jet entrance region of the jet to the east of the Northeast United States. The midlevel and 850-hPa trough approaching the Appalachians is deeper in the hybrid events ([Figure 6\(c\)](#)). Meanwhile, there is a surface cyclone located over the Southeast United States ([Figure 6\(d\)](#)), with an inverted trough extending northward to the Ohio Valley. The large-scale ascent as well as moist flow around the developing cyclone ascending over the cold dome produces precipitation within the CAD region (not shown). This precipitation can enhance the cold dome through evaporative cooling and melting if the freezing level is relatively low ([Figure 7](#)).

CAD Examples around the World

There have been many documented studies of CAD around the world. The most widely studied area has been along the eastern side of the Appalachian Mountains, where the 'wedge ridge' sea-level pressure pattern is sometimes referred to as the 'Baker' ridge. Because the Appalachians are within a few hundred kilometers of the relatively warmer Atlantic Ocean, the low-level temperature gradient near the coast can be enhanced through confluence between the nearly terrain-parallel flow associated with CAD and the warmer more geostrophic flow near the coast. This frontogenetical process helps in the development of an enhanced temperature gradient near the coast called a 'coastal front.' The low-level temperature advections associated with these shallow baroclinic zones result in mixed precipitation, which can be in the form of freezing rain (ice storms) and sleet within areas of the cold dome.

CAD has recently been documented for other large mountain barriers, such as the Rockies and the Andes. For these extremely long N-S barriers, relatively cold air can propagate into tropical latitudes (10° N/S). This channeling of cold air is often referred to as a 'cold surge,' which is a cold anticyclone moving equatorward adjacent to the mountains. The primary mechanism for driving the surge southward is the development of an along-barrier pressure gradient, cold advection, and the damming evolution described above. The very high orography associated with the Rockies, Andes, and Himalayas often results in a horizontal scale of CAD that is typically much larger (>300 km) than the Appalachians and other more narrow coastal barriers. Since the Rocky Mountains are adjacent to the sloping Great Plains, flow blocking is also favored along the slope, resulting in a cold dome that extends more than 500 km east of the Rockies. As the North American surge moves southward along the mountainous Mexican coast and the large-scale slope is lost, the scale of the damming collapses to a few hundred kilometers as determined by the Rossby radius.

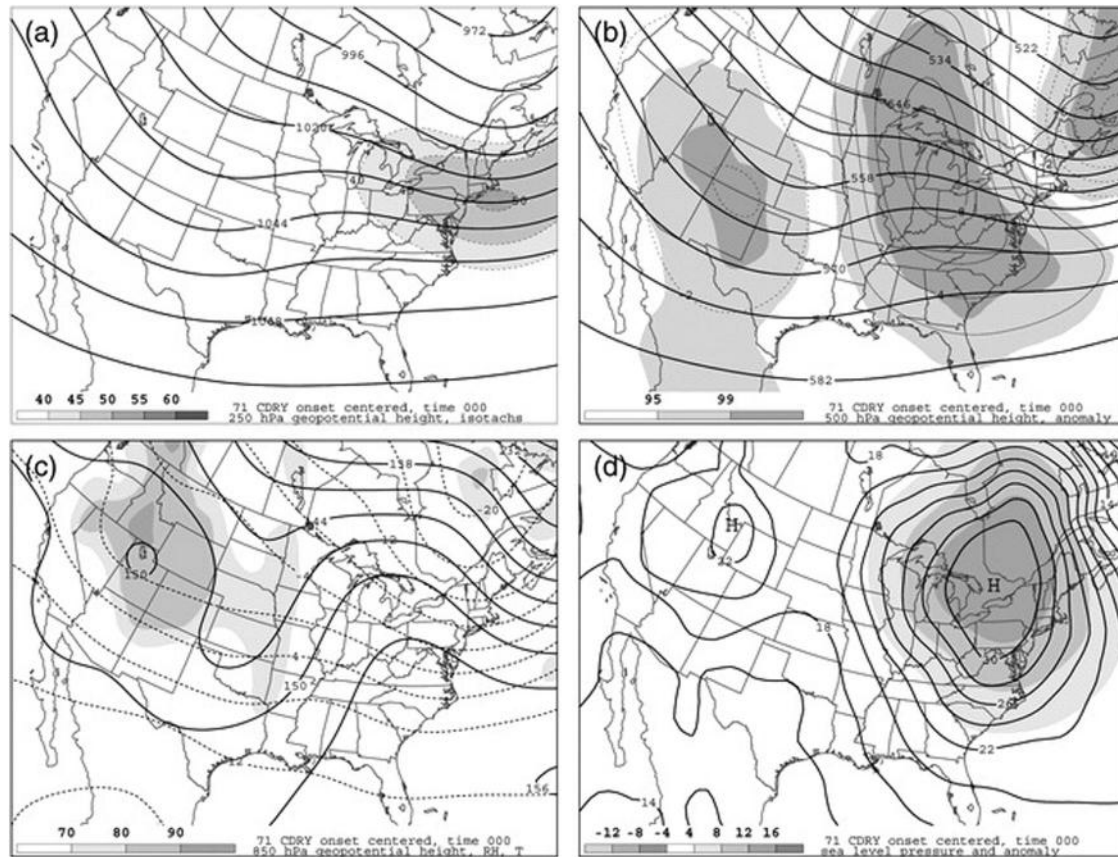


Figure 6 Composite of the classical dry CAD events for the Appalachians at the time of maximum onset showing: (a) 250-mb height (solid, every 12 dam) and isotachs (dashed in m s^{-1} and shaded as in legend at lower left of panel, with intervals below 40 m s^{-1} omitted); (b) 500-mb height (solid, every 6 dam), anomaly (thin contours, dashed (negative), and solid (positive); interval, 2 dam), and statistical significance (shaded as in legend at lower left); (c) 850-mb height (solid, every 3 dam), temperature (dashed; contour interval, 4°C), and relative humidity (shaded as in legend at lower left); (d) sea-level pressure (solid; contour interval, 2 hPa) and anomaly (shaded as in legend at lower left). From Figure 10 in Bailey, C.M., Hartfield, G., Lackmann, G.M., Keeter, K., Sharp, S., 2003. An objective climatology, classification scheme, and assessment of sensible weather impacts for Appalachian cold-air damming. *Weather Forecasting* 18, 641–661. © American Meteorological Society. Reprinted with permission.

CAD and associated barrier jets can occur for other mountainous coastal regions, such as along Vancouver Island British Columbia, southeast Alaskan coast, east slope of the Washington Cascades, and along the Eastern Sierras. During these cases, cold air is pooled up against the barrier from a previous excursion of cold air from the interior of the continent through gaps and channels in the terrain. The along-barrier pressure gradient is often enhanced with the approach of a surface front, resulting in an enhancement of the terrain-parallel ageostrophic flow (barrier jet).

For maritime climates, such as coastal California and Australia, damming occurs when cool marine air is rotationally blocked against the coastal terrain. Typically, these events are associated with a transition from warm dry offshore flow to terrain-blocked cool marine adjacent to the coast. These coastal marine surges often result in a narrow stratus tongue that extends approximately a Rossby radius off the coast.

For many years, CAD could not be adequately resolved by operational numerical weather models, such as National Centers for Environmental Prediction's (NCEP) Nested Gridded Model and the Medium Range Forecast (MRF) Model, since

horizontal grid spacings of these models were around 30–80 km. More recent higher resolution operational or research mesoscale models (i.e., NCEP's North American Mesoscale model at 12 km resolution) are able to more realistically forecast different damming events. However, there are predictability issues regarding the strength and duration of the damming event related to difficulties in simulating the diabatic effects (radiation, evaporative cooling, and melting). In particular, numerical models often scour out the cold air too quickly and thus underestimate the duration of these CAD events.

CAD Impacts

CAD can have significant societal impacts. In addition to below normal temperatures persisting for several days at times, many CAD events are associated with freezing rain and sleet, which can create hazardous travel and aviation delays. For the Southeast United States, the high impact events tend to have a more pronounced jet-entrance region over the Mid-Atlantic region and another jet streak along the Gulf Coast. There is also

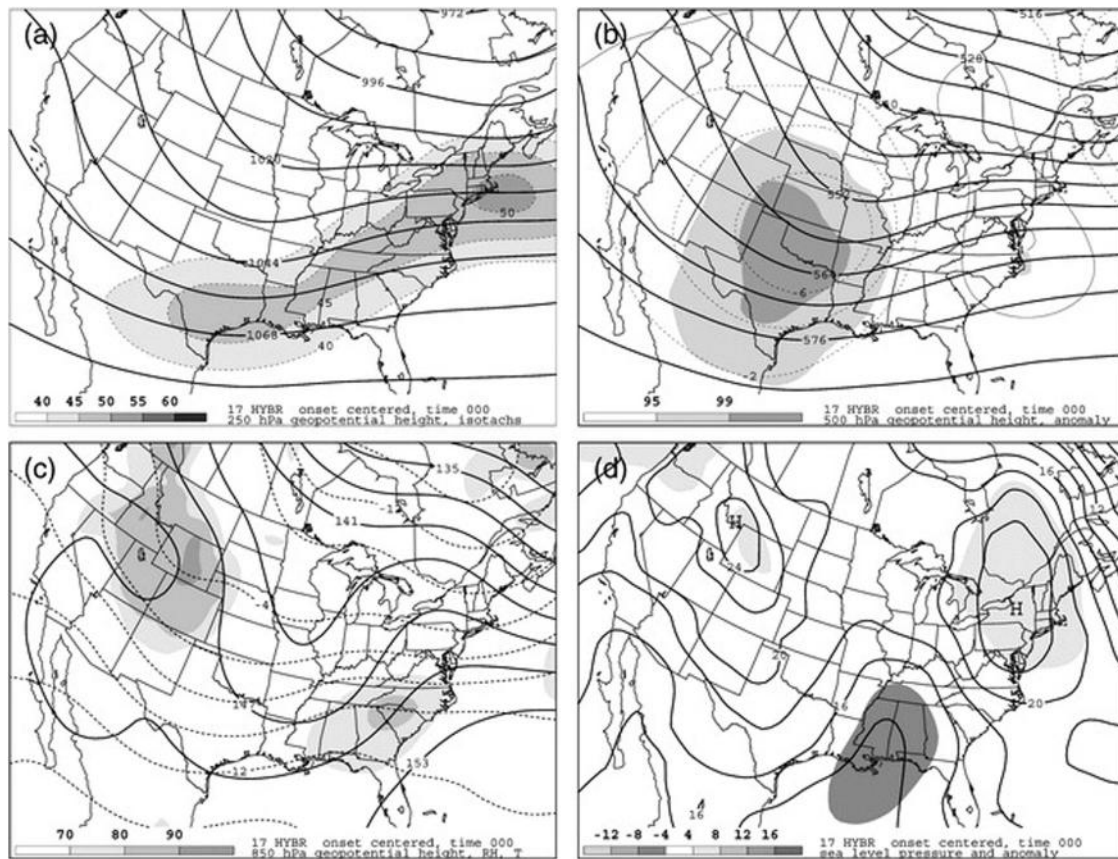


Figure 7 Same as Figure 6 except for the hybrid damming events. From Figure 13 in Bailey, C.M., Hartfield, G., Lackmann, G.M., Keeter, K., Sharp, S., 2003. An objective climatology, classification scheme, and assessment of sensible weather impacts for Appalachian cold-air damming. *Weather Forecasting* 18, 641–661. © American Meteorological Society. Reprinted with permission.

stronger warm advection and moisture advection off the Atlantic at 850 hPa than drier CAD events. This coupled jet, low-level warm advection, and weaker ridging aloft tend to create relatively strong upward motion forcing aloft and heavy precipitation. Weak CAD has been shown to occur ahead of landfalling tropical systems over the Southeast United States (Floyd, 1999), in which low-level frontogenesis between the CAD and the coastal front can enhance the precipitation well before the landfall of the tropical system.

The CAD region can also create a 'blocking front' within a Rossby radius upstream of the mountain, in which a low-level convergence boundary develops between the terrain-parallel flow and ambient flow approaching the mountain. This boundary has been shown to extend the heavy precipitation upstream of a mountain via this convergence and the flow up and over this blocked region. This has been observed to occur upstream of the Sierras, Wasatch, and European Alps.

See also: Aviation Meteorology: Aviation Weather Hazards. **Boundary Layer (Atmospheric) and Air Pollution:** Complex Terrain. **Dynamical Meteorology:** Balanced Flow; Coriolis Force; Static Stability. **Mountain Meteorology:** Overview. **Numerical Models:** Regional Prediction Models. **Synoptic Meteorology:** Anticyclones.

Further Reading

- Bailey, C.M., Hartfield, G., Lackmann, G.M., Keeter, K., Sharp, S., 2003. An objective climatology, classification scheme, and assessment of sensible weather impacts for Appalachian cold-air damming. *Weather Forecasting* 18, 641–661.
- Baker, D.G., 1970. A study of high pressure ridges to the east of the Appalachian Mountains (Ph.D. thesis). Massachusetts Institute of Technology, p. 127.
- Bell, G.D., Bosart, L.F., 1988. Appalachian cold-air damming. *Monthly Weather Review* 116, 137–161.
- Bluestein, H.B., 1993. Synoptic-Dynamic Meteorology in Midlatitudes. In: *Observations and Theory of Weather Systems*, vol. II. Oxford University Press pp. 359–362.
- Colle, B.A., Mass, C.F., 1995. The structure and evolution of cold surges east of the Rocky Mountains. *Monthly Weather Review* 123, 2577–2610.
- Forbes, G.S., Anthes, R.A., Thomson, D.W., 1987. Synoptic and mesoscale aspects of an Appalachian ice storm associated with cold air damming. *Monthly Weather Review* 115, 564–591.
- Lackmann, G., 2011. Mid-latitude Synoptic Meteorology: Dynamics, Analysis, and Forecasting. American Meteorological Society, Boston, MA, p. 345.
- Overland, J.E., 1984. Scale analysis of marine winds in straits and along mountainous coasts. *Monthly Weather Review* 112, 2532–2536.
- Pierrehumbert, R.T., Wyman, B., 1985. Upstream effects of mesoscale mountains. *Journal of Atmospheric Science* 42, 977–1003.
- Stauffer, D.R., Warner, T.T., 1987. A numerical study of Appalachian cold-air damming and coastal frontogenesis. *Monthly Weather Review* 115, 799–821.
- Xu, Q., 1990. A theoretical study of cold air damming. *Journal of Atmospheric Science* 47, 2969–2985.

Downslope Winds

DR Durran, University of Washington, Seattle, WA, USA

© 2003 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 2, pp 644–650, © 2003, Elsevier Ltd.

Introduction

Very strong surface winds sometimes develop when air flows over a high mountain ridge with a steep lee slope. Such winds are known to occur at many locations throughout the middle latitudes. Local names for these winds include the Alpine foehn, the Rocky Mountain chinook, the Croatian bora, the Santa Ana in Southern California, and the Argentine zonda. These winds are collectively referred to as *downslope* winds. Downslope winds in the lee of major mountain barriers can approach hurricane strength. (By definition, hurricanes are storms with *sustained* winds of at least 32 m s^{-1} (115 km h^{-1}).) Every few years, for example, the eastern slope of the Colorado Front Range (part of the Rocky Mountains) experiences a damaging windstorm with peak gusts as high as 60 m s^{-1} (216 km h^{-1}). An anemometer trace recorded at the National Center for Atmospheric Research in Boulder, CO, during a strong chinook is shown in [Figure 1](#).

In modern meteorological usage, downslope winds are distinguished from *katabatic* winds by the dynamical processes driving each flow. Katabatic winds usually refer to shallow gravity currents generated by the cooling of surface air over sloping terrain. Downslope winds usually refer to winds generated as a deeper layer of air is forced over topography. In contrast to katabatic winds, the diabatic cooling of air in contact with a cold surface plays no essential role in the dynamics of downslope winds.

In most downslope wind events (including the typical foehn and chinook), the onset of the downslope wind is

accompanied by an increase in the surface temperature and a drop in the dew point. Whereas the area of violent wind is limited to a relatively narrow swath along and adjacent to the lee slope, the warmer drier air mass can extend much further downstream. Nevertheless, in some cases the upstream conditions may be so cold, and the initial downstream conditions sufficiently warm, that the onset of a downslope wind brings a drop in temperature. The most well-known example of this type of cold downslope wind is the Croatian bora. Despite the difference in the evolution of the surface temperature, there does not appear to be any significant dynamical distinction between the processes responsible for the development of high downslope winds in cold and warm events.

Contours of the potential temperature observed on 11 January 1972 during an intense downslope windstorm are plotted in the vertical cross section through Boulder, CO, shown in [Figure 2](#). These contours provide a rough indication of the streamlines in the flow, which is moving from left to right. (The isentropes would be exactly identical to streamlines if the flow were steady, inviscid, adiabatic, and two-dimensional (2D).) A large-amplitude mountain wave is clearly visible in the potential temperature field just to the lee of the continental divide. The apparent horizontal displacement of the wave trough at upper levels from its position at low levels is due to a 2 h difference between the times at which observations were collected in the upper and lower flight levels. Also apparent in [Figure 2](#) is a layer of enhanced static stability around the 550 mb level in the upstream flow. When intense downslope winds develop in a deep cross-mountain flow, strong mountain waves and low-level stable layers similar to those shown in [Figure 2](#) are usually present.

The connection between mountain waves and strong downslope winds is less apparent in situations where the cross-mountain wind component reverses with height at some level in the middle or lower troposphere, as is often the case in the Croatian bora or when strong winds blow from the east down the western slopes of the Wasatch mountains in Utah. Contours of the potential temperature observed during a moderate bora along a cross section through Senj, Croatia, are shown in [Figure 3](#). The flow in this example is from right to left. A low-level inversion is once again apparent upstream of the mountains; however, no significant wave activity is present above the 3 km level. In this case, the upstream inversion is coincident with a region of strong vertical wind shear in which the cross-mountain wind component reverses direction. The level at which the cross-mountain wind component drops to zero is a critical level for steady 2D mountain waves, and any gravity waves triggered by the mountain break down and dissipate as they approach this critical level. (A critical level is a level at which the phase speed of a wave equals the speed and direction of the basic state flow.)

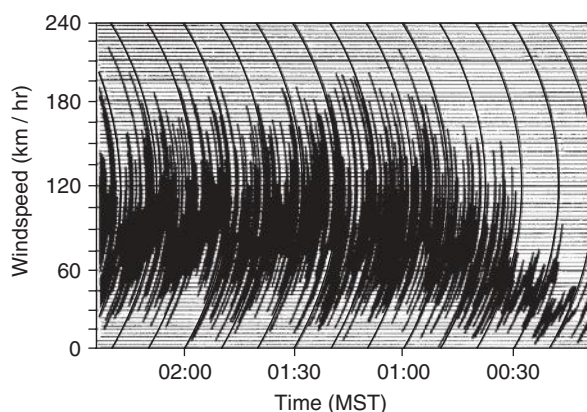


Figure 1 Anemometer trace recorded at the National Center for Atmospheric Research during the onset of the 17 January 1982 Boulder windstorm. Time reads right to left. Reproduced from Durran, D.R., 1990. Mountain waves and downslope winds. In: Blumen W (Ed.) Atmospheric Process over Complex Terrain [Figure 4.11]. American Meteorological Society. Boston, MA, pp. 59–81.

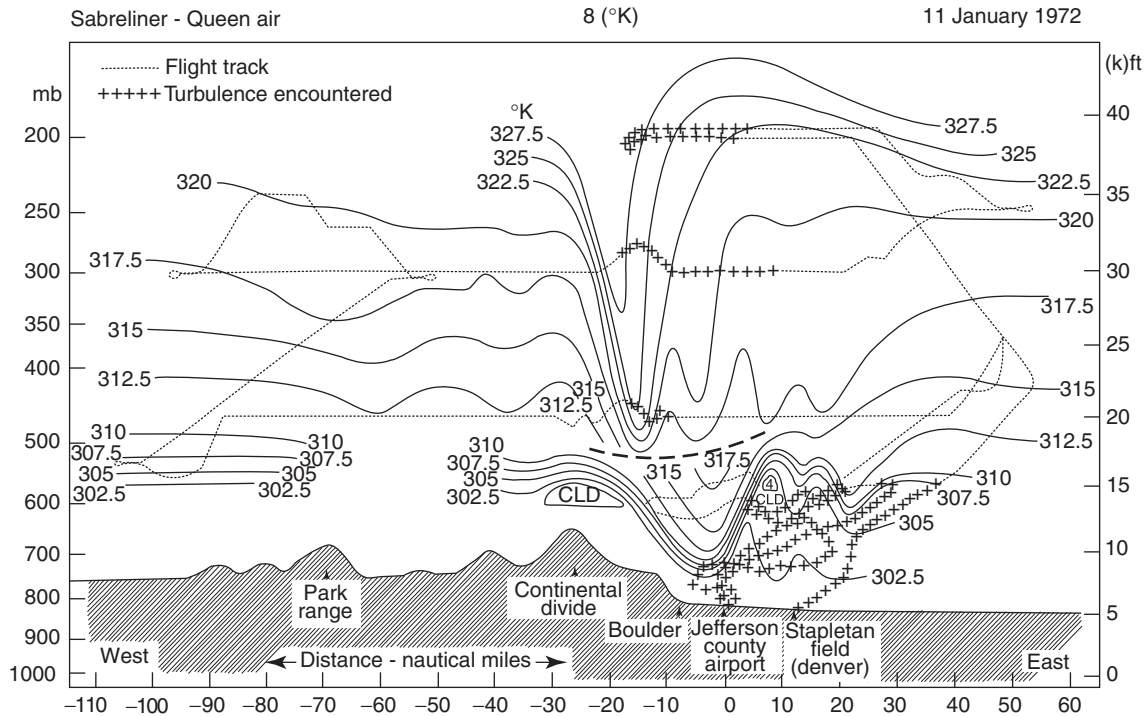


Figure 2 Cross section of potential temperature along an east-west line through Boulder, CO, from aircraft observations collected on 11 January 1972. The heavy dashed line separates data collected by two aircraft at different times. Flight tracks are indicated by the dotted lines: those segments along which significant turbulence was encountered are denoted by pluses. Reproduced from Lilly, D.K., 1978. A severe downslope windstorm and aircraft turbulence event induced by a mountain wave [Figure 7]. *Journal of the Atmospheric Sciences* 35, 59–77.

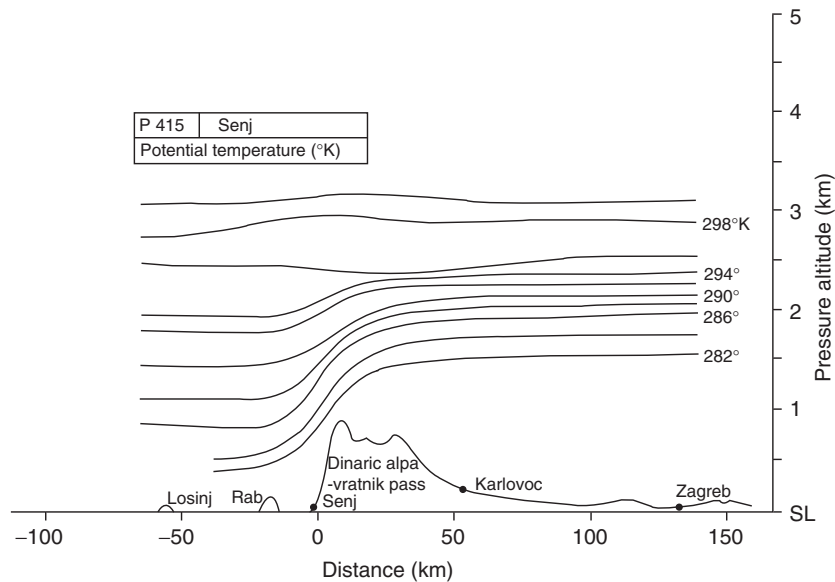


Figure 3 Cross section of potential temperature along a northeast-southwest line through Zagreb and Senj, Croatia. Reproduced from Smith, R.B., 1987. Aerial observations of the Yugoslavian bora [Figure 9(b)]. *Journal of the Atmospheric Sciences* 44, 269–297.

The Hydraulic Analog

The dynamics governing the development of strong downslope winds in the atmosphere are analogous to those governing the rapid increase in speed that occurs when water flowing over

a rock in a river undergoes a transition from a relatively slow velocity upstream to a thin layer of high-velocity fluid over the downstream face. In such circumstances, a turbulent hydraulic jump often develops downstream of the rock at the point where the high-speed flow decelerates back to the ambient

velocity of the river. Since the fundamental processes responsible for the rapid acceleration of water flowing over a rock can be explained more simply than those which govern downslope winds in the atmosphere, it shall be begun by considering the hydraulic model for a shallow layer of water flowing over an obstacle in an open channel.

Suppose a homogeneous fluid, such as freshwater, is flowing over a ridgelike obstacle. Assuming the flow is steady and that there are no variations in the coordinate direction parallel to the ridge axis, and making the hydrostatic approximation, the flow is governed by the horizontal momentum equation

$$u \frac{\partial u}{\partial x} + g \frac{\partial D}{\partial x} = -g \frac{\partial h}{\partial x} \quad [1]$$

and the mass continuity equation

$$\frac{\partial uD}{\partial x} = 0 \quad [2]$$

where x is the horizontal coordinate directed perpendicular to the ridgeline, u is the velocity in the x direction, D is the thickness of the fluid, and h is the local height of the obstacle. Using eqn [2] to substitute for $\partial u/\partial x$ into eqn [1] yields

$$(1 - F^2) \frac{\partial}{\partial x} (D + h) = \frac{\partial h}{\partial x} \quad [3]$$

Here

$$Fr = \frac{u}{\sqrt{gD}} \quad [4]$$

is the Froude number, which is the ratio of the local flow speed to the local phase speed of a linear shallow water gravity wave.

According to eqn [3], the magnitude of the Froude number determines whether the free surface rises or falls as the fluid ascends the upstream slope of the obstacle. The case $F > 1$, known as *supercritical flow*, is shown in Figure 4(a); the fluid thickens and slows as it passes over the top of the obstacle, and it reaches its minimum speed at the crest. The accelerations experienced by the fluid are qualitatively similar to those experienced by a hockey puck traversing a frictionless ridge of ice. The case $F < 1$, known as *subcritical flow*, is shown in Figure 4(b). The fluid parcel accelerations in the subcritical flow seem counterintuitive in that the fluid thins and accelerates as it crosses the top of the obstacle, reaching its maximum speed at the crest.

Why does a subcritical flow accelerate as it encounters rising bottom topography? In contrast to a frictionless hockey puck, the acceleration of a fluid parcel is determined not only by the gravity and the angle of the slope but also by the pressure gradient forces. The steady-state momentum equation (eqn [1]) requires a three-way balance between acceleration (the first term), pressure gradient forces arising from changes in the fluid depth (the second term), and the work per unit mass done against gravity while ascending the sloping topography (the third term). The value of the Froude number determines whether the work done against gravity is predominantly balanced by accelerations or by the pressure gradient force. From eqn [2]

$$\left(u \frac{\partial u}{\partial x} \right) / \left(g \frac{\partial D}{\partial x} \right) = \left(u \frac{\partial u}{\partial x} \right) / \left(-\frac{gD}{u} \frac{\partial u}{\partial x} \right) = -F^2 \quad [5]$$

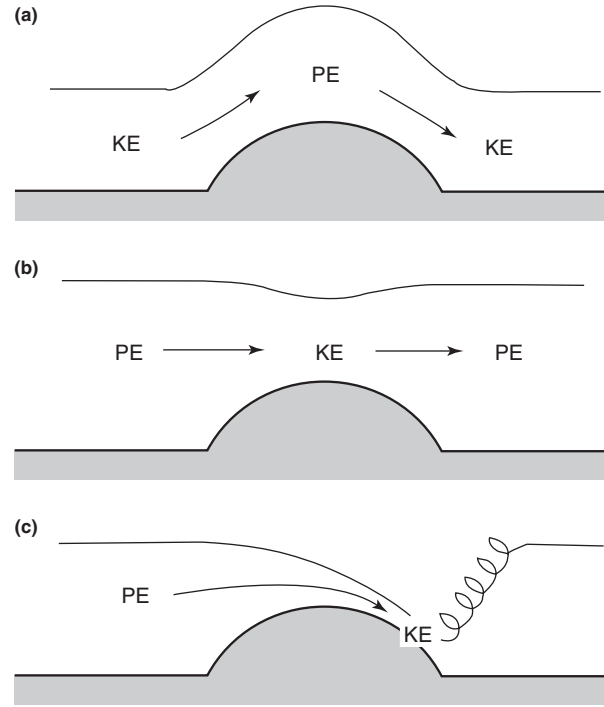


Figure 4 Behavior of shallow water flowing over an obstacle: (a) everywhere supercritical flow; (b) everywhere subcritical flow; and (c) hydraulic jump after a transition from subcritical to supercritical flow over the crest. Reproduced from Durran, D.R., 1990. Mountain waves and downslope winds [Figure 4.5]. In: Blumen, W., (Ed.) Atmospheric Process over Complex Terrain. American Meteorological Society, Boston, MA, pp. 59–81.

Thus in steady open-channel hydraulic flow, acceleration always opposes the pressure gradient force due to changes in fluid depth. Furthermore, F^2 may be interpreted as the ratio of the magnitude of the acceleration to the magnitude of the pressure gradient force generated by changes in the fluid depth. In supercritical flow ($F > 1$) acceleration dominates the pressure gradient force and the three-way balance in eqn [1] is satisfied such that fluid parcels ascending the upstream slope decelerate as they do work against gravity.

Before discussing the subcritical case, it is helpful to recast the discussion in terms of the conversions between kinetic energy (KE) and potential energy (PE). Equation [1] implies that $u^2/2 + g(D + h)$ is constant along a streamline. This is just Bernoulli's theorem for steady incompressible hydrostatic flow since the contribution of $w^2/2$ to the total KE is neglected in the hydrostatic approximation. The term $g(D + h)$ represents the combined PEs associated with the gravitational and pressure fields, as may be verified by taking the hydrostatic pressure to be zero at the top of the water and choosing the $z = 0$ level to coincide with the bottom of the channel away from the obstacle; then at an arbitrary level z ,

$$gz + \frac{p}{\rho_0} = gz + g(D + h - z) = g(D + h) \quad [6]$$

According to this generalized interpretation of PE, fluid parcels ascending the obstacle in a supercritical flow slow down as they convert KE into PE, and after passing the crest they reaccelerate as PE is converted back to KE (Figure 4(a)).

On the other hand, in subcritical flow ($F < 1$) the pressure gradient force dominates acceleration and the three-way balance in eqn [1] requires that fluid parcels accelerate in the direction opposite to the component of gravity parallel to the topography. As shown in Figure 4(b), fluid parcels ascending the obstacle accelerate as the free surface drops and PE is converted to KE. After passing the crest, the parcels decelerate as KE is converted back to PE. The disturbance centered over the obstacle in Figure 4(b) is a steady surface gravity wave.

The flow regime that serves as an analog for downslope windstorms is shown in Figure 4(c). If the flow is subcritical upstream and if a column of fluid undergoes a sufficient acceleration and experiences a sufficient decrease in thickness as it ascends toward the crest, a transition from subcritical to supercritical flow occurs at the top of the obstacle. Since the lee-slope flow is now supercritical, fluid parcels continue to accelerate as they descend, and very high velocities can be produced because PE is converted to KE during the entire time over which a fluid parcel traverses the obstacle. The deceleration that would otherwise occur in the lee-side portion of the standing gravity wave is disrupted when the flow becomes supercritical. In this case fluid parcels eventually decelerate when they pass through a turbulent hydraulic jump at some point downstream from the crest.

Application of the Hydraulic Analog to the Atmosphere

The hydraulic analog is best applied to the atmosphere in a qualitative rather than a quantitative manner. Quantitative application is hindered by the difficulty of defining a dynamically meaningful Froude number in vertically unbounded continuously stratified flow. A variety of expressions have been described as Froude numbers in the literature, but all of the simple expressions have serious deficiencies.

The parameter U/Nh_0 , where N is the Brunt–Vaisala frequency, U the wind speed, and h_0 the maximum mountain height, is sometimes referred to as the Froude number in idealized cases in which N and U are constant throughout the upstream flow. Unlike the denominator in the conventional shallow water Froude number, Nh_0 is not the horizontal phase speed of any particularly significant wave. (Nh_0 is the phase speed of a hydrostatic internal gravity wave with vertical wavelength $2\pi h_0$, but there is nothing particularly significant about this wavelength in contrast to other similar waves with wavelengths such as $5h_0$ or $6h_0$.) On the other hand, the maximum perturbation horizontal wind speed u' in linear flow over an obstacle with constant N and U scales like Nh_0 , so that $U/(Nh_0) \approx U/u'$ might be better described as a nonlinearity parameter.

When there is a strong well-defined inversion at some elevation H in the upstream flow, many authors have attempted to define a Froude number as $U/\sqrt{g'H}$, where $g' = g\Delta\theta/\theta_0$ is the 'reduced gravity,' $\Delta\theta$ is the increase in potential temperature across the inversion, and θ_0 is the mean potential temperature below the inversion. The difficulty with this approach is that it implies that the pressure gradient force is entirely determined by the vertical displacements of the inversion layer and thereby neglects the influence, on the surface pressure gradient, of vertical displacements in the stably

stratified fluid above and below the inversion. Moreover, it is also very difficult to determine a precise quantitative value for $U/\sqrt{g'H}$ in more general applications in which the wind speed is not constant below the inversion and the inversion itself may be indistinct. As a consequence, the reduced gravity shallow water model, in which F is replaced by $U/\sqrt{g'H}$ in eqn [3], will not reliably yield reasonable approximations to the speed and depth of the downslope flow in actual windstorms.

Significant downslope winds have been observed to develop in three basic situations: (1) when a standing mountain wave in a deep cross-mountain flow achieves sufficient amplitude to overturn and break down at some level in the troposphere; (2) when standing mountain waves break and dissipate at a critical level in a shallow cross-mountain flow; and (3) when there is sufficient static stability near mountaintop level in the cross-mountain flow to create high downslope winds even without wave breaking. The qualitative application of hydraulic theory to the dynamics of downslope winds centers on the idea that in all three of these cases there is a transition from wavelike behavior over the upstream slopes of the topography to a nonwavelike regime in the lee.

First, consider the case of breaking waves in a deep cross-mountain flow. The structures of the low-level horizontal velocity perturbations in a stationary 2D internal gravity wave forced by an isolated ridge are shown in Figure 5(a). In this case, the upstream wind and static stability are constant with height such that $N = 0.01047 \text{ s}^{-1}$, $U = 10 \text{ m s}^{-1}$, and $Nh_0/U = 0.6$. Streamlines for this same stationary internal gravity wave are plotted in Figure 3(a) (see Mountain Meteorology: Lee Waves and Mountain Waves). As is apparent in Figure 5(a), the detailed structures of the velocity perturbations in the internal gravity wave are somewhat different from those in the surface gravity wave schematically illustrated in Figure 4(b). In particular, the maximum perturbation surface wind speed occurs halfway down the lee slope in the internal gravity wave, whereas it occurs at the crest in the surface gravity wave. Nevertheless, both types of waves allow a fluid parcel to arrive at the ridge crest with a positive perturbation velocity (i.e., to undergo a net acceleration while ascending to the crest), and in both cases the wind speed eventually returns to its ambient value well downstream of the crest as KE is converted back to PE in the lee-side portion of the stationary gravity wave.

The enhancement of the perturbation horizontal winds along the lee slope in Figure 5(a) is too weak to create significant downslope winds. (The total wind speed increases from 10 m s^{-1} far upstream to approximately 15 m s^{-1} in the lee.) Much stronger downslope winds occur in the case shown in Figure 5(b), which is a vertical cross section of the perturbation horizontal velocity in a simulation identical to that shown in Figure 5(a), except that the height of the mountain has been doubled so that $Nh_0/U = 1.2$. The higher topography in this case forces the internal gravity wave to overturn and produces a well-mixed region of weakly reversed flow at elevations around 3 km over the lee slope. (The region of reversed flow is that, in which the horizontal perturbation velocity is less than -10 m s^{-1} .) Streamlines for this same wave-breaking case are shown in Figure 4(a) (see Mountain Meteorology: Lee Waves and Mountain Waves). Although the lee-side flow is dramatically different when the wave is breaking, the flow upstream of the crest remains consistent with that in a stationary internal gravity

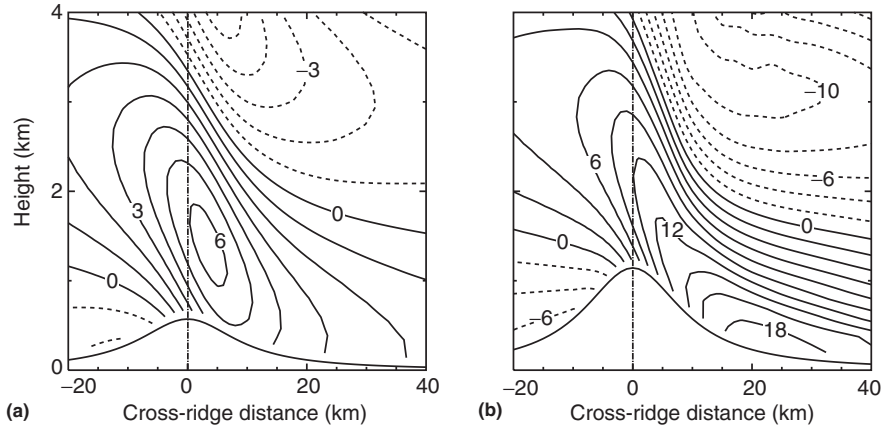


Figure 5 Perturbation horizontal velocity in flow over an isolated mountain when (a) $Nh_0/U = 0.6$, contour interval 1 m s^{-1} ; and (b) $Nh_0/U = 1.2$, contour interval 2 m s^{-1} .

wave. Linear theory for stationary internal gravity waves predicts that doubling the mountain height should double the amplitude of the perturbation horizontal velocities without changing the spatial distribution of the perturbations relative to the mountain, and this is essentially the case in the region upstream of the crest. Note, for example, the similarity of the 3 m s^{-1} contour in [Figure 5\(a\)](#) with the 6 m s^{-1} contour in [Figure 5\(b\)](#).

Since the wave in [Figure 5\(b\)](#) has become unstable and overturned above the lee slope, there is no standing gravity wave to decelerate the fluid parcels as they descend. Instead these parcels continue to accelerate as PE is converted to KE along the entire lee slope, generating strong downslope winds in which the maximum surface wind speeds ($>28 \text{ m s}^{-1}$) are approximately three times stronger than the 10 m s^{-1} flow far upstream. Wave breaking in a deep cross-mountain flow appears to have played an important role in the generation of the 11 January 1972 Boulder, CO, windstorm. The presence of breaking waves is suggested by the almost vertical orientation of the isentropes on the lee side of the trough in the upper-level wave in [Figure 2](#) and by the turbulence encountered along the flight legs through this region.

The second type of situation conducive to the development of strong downslope winds is illustrated in [Figure 3](#). In this bora event a critical level at an elevation of about 2 km disrupts the lee-side gravity wave so that, once again, fluid parcels near the surface undergo a net acceleration in the wavelike upstream flow as they ascend the mountain crest and then continue to accelerate as they convert PE to KE while descending the entire lee slope. The vertical displacement of a streamline about its initial undisturbed level $\delta(x, z)$ can be modeled with reasonable fidelity in the flow beneath the critical layer by solving the hydrostatic Long's equation

$$\frac{\partial^2 \delta}{\partial z^2} + \frac{N^2}{U^2} \delta = 0 \quad [7]$$

subject to the lower boundary condition that the streamline follow the topography

$$\delta[x, z = h(x)] = h(x) \quad [8]$$

and an upper boundary condition in which the horizontal wind speed is held constant along a 'dividing streamline'

separating the well-mixed turbulent region from the underlying high-speed flow. In the case shown in [Figure 3](#), the 294 K isentrope approximates a dividing streamline while the 296 K isentrope roughly coincides with the top of the wedge of well-mixed air downwind of the crest. Very close mathematical analogies exist between conventional shallow water hydraulic theory and the mathematical solutions to eqns [7] and [8], although there is no *simple* parameter that plays the role of the Froude number in this analogy.

The third situation that produces strong downslope winds may occur when there is high static stability at low levels in the cross-mountain flow and lower stability aloft. A prototypical example of this type is presented in [Figure 6](#), which shows contours of the perturbation horizontal velocity field and streamlines from a numerical simulation identical to that described in [Figure 5\(a\)](#), except that the Brunt–Väisälä frequency above 3 km in the upstream flow is reduced by a factor of 0.4. Comparison of the horizontal wind speed perturbations between [Figures 5\(a\) and 6\(a\)](#) shows that the perturbation horizontal winds are twice as strong and that the maximum winds have shifted to the surface along the lee slope in the two-layer flow. The amplification of the surface winds in the two-layer simulation is produced without wave breaking; in fact, the flow does not come close to stagnation.

The streamlines within the lower layer shown in [Figure 6\(b\)](#) appear similar to those in water undergoing a transition from subcritical to supercritical flow over the crest of an obstacle. Near the base of the lee slope in [Figure 6](#), the flow recovers toward ambient conditions by radiating energy downstream in a series of vertically trapped gravity waves. The removal of energy by these trapped waves is analogous to the dissipation of energy at the point where the flow recovers toward ambient downstream conditions in a hydraulic jump in the standard shallow water model ([Figure 4\(c\)](#)). Additional sensitivity studies have demonstrated that the changes in the depth of the lower layer and the maximum height of the mountain modify the two-layer flow in a manner one would expect on the basis of hydraulic theory. In particular, making the lower layer too deep or the mountain too small eliminates the transition to a high wind regime.

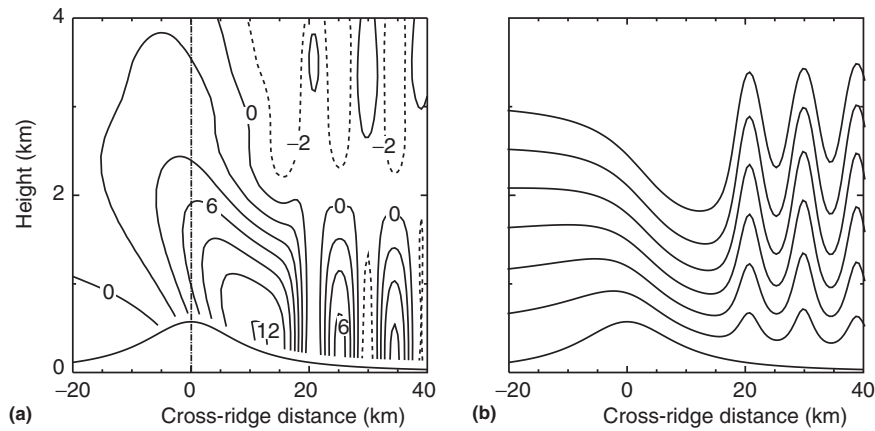


Figure 6 Two-layer flow over an isolated mountain in which the upstream value of Nh_0/U is 0.6 in the lower layer and 0.24 above: (a) perturbation horizontal velocity contour interval 2 m s^{-1} ; and (b) streamlines within the lower layer.

In actual downslope wind events the dynamical influence of a low-level stable layer may act in concert with wave breaking to generate very high winds. Indeed climatological data and numerical experiments suggest that this is often the case in Boulder windstorms. In particular, nonlinear wave amplification due to the presence of a low-level stable layer appears to have served as a necessary precursor to wave breaking during the 11 January 1972 event.

See also: **Gravity Waves:** Buoyancy and Buoyancy Waves: Optical Observations; Buoyancy and Buoyancy Waves: Theory. **Mountain Meteorology:** Katabatic Winds; Lee Vortices; Lee Waves and Mountain Waves.

Further Reading

- Durran, D.R., 1990. Mountain waves and downslope winds. In: Blumen, W. (Ed.), *Atmospheric Process over Complex Terrain*. American Meteorological Society, Boston, MA, pp. 59–81.
- Lilly, D.K., 1978. A severe downslope windstorm and aircraft turbulence event induced by a mountain wave. *Journal of the Atmospheric Sciences* 35, 59–77.
- Smith, R.B., 1987. Aerial observations of the Yugoslavian bora. *Journal of the Atmospheric Sciences* 44, 269–297.
- Smith, R.B., 1989. Hydrostatic air flow over mountains. In: Saltzman, B. (Ed.), *Advances in Geophysics*, vol. 31. Academic Press, New York, NY, pp. 1–41.

Katabatic Winds

TR Parish, University of Wyoming, Laramie, WY, USA

© 2003 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 3, pp 1057–1061, © 2003, Elsevier Ltd.

Introduction

Katabatic winds are a class of atmospheric motion in which air flow is directed down a topographic incline such as a mountainside or glacier. The term 'katabatic' is Greek in origin from katabatikos, meaning to go down, and has been used in meteorological literature since World War I. Reference to katabatic winds is reserved for air motion in the lower atmosphere, extending from the surface to a height typically below 1 km.

In the most general sense, any wind blowing down an incline can be classified as a katabatic wind. The broad definition has allowed a host of local wind features to be defined as katabatic. As air moves downslope, it is subject to adiabatic warming due to compression, since the pressure increases during the descent. Two general categories of downslope winds exist, distinguished from each other by the relative temperature of the air stream. A wind that is warmer than the air being displaced along and at the bottom of an incline is referred to as a 'foehn.' This term has its origin in the warm winds descending the slopes of the Alps in Europe. Similar downslope winds east of the Rocky Mountains of North America have been called 'chinooks', which is taken from the Native American word meaning 'snow eater' to describe the effects of such wind events. The occurrence of these winds requires particular weather conditions in the ambient atmosphere above the downslope flow that directs the air motion.

Use of the term 'katabatic' to describe warm wind features was common throughout much of the early and middle part of the twentieth century, but has decreased over the last few decades. Discussion of katabatic winds now invariably refers to the second type of downslope flow, which is a cold wind. Some air streams originate over high plateaus that are subjected to intense cooling, such as over the interior of the Antarctic or Greenland ice sheets. These air currents remain colder than the surrounding environment during the descent despite compressional heating. Gravity is the driving force for such flows since the density of the cold air is greater than that of the air it displaces. The cold downslope winds are dependent on the local terrain slope characteristics and relative density of the air stream, although weather conditions in the ambient atmosphere are often important as well.

Katabatic winds occur most frequently during nocturnal conditions, and encompass a wide range of time and distance scales. Cold air drainage flows along the sides of valleys having a distance scale on the order of 1 km and the broad continental-scale gravity flows over the great ice sheets of Antarctica and Greenland are each classified as katabatic. The following discussion will focus on the cold, dense downslope wind features for which the term katabatic is most commonly applied.

Dynamics of Katabatic Winds

Topography is the controlling factor for both the intensity and the direction of katabatic winds. This is illustrated schematically in Figure 1, which represents the atmosphere as a simple two-layer fluid. Although highly idealized, the layer model approach in Figure 1 can explain many of the characteristics of katabatic winds. Each layer has a constant density. The lower layer represents the cold, dense layer nearest to the Earth; and the upper layer represents the atmosphere undisturbed by the cooling process. A negatively buoyant lower layer most commonly develops at night when the sky is clear as strong radiative cooling of the inclined surface takes place. Horizontal pressure differences are the fundamental cause for all atmospheric motions. In the case of katabatic winds the horizontal pressure gradient force is proportional to the difference in pressure between the cooled air near the surface (p_1) and at the same horizontal level in the ambient atmosphere overlying the cooled layer (p_2). Through the hydrostatic relationship, the vertical change in pressure is dependent on the air density and the thickness of the air column. For a given height in a vertical column, pressure decreases more rapidly with height for cold air than for warm air. Applying the hydrostatic principle, p_1 will be greater than p_2 if no horizontal pressure difference exists above the cooled layer, and thus horizontal acceleration will occur in a downslope sense. By use of the hydrostatic equation,

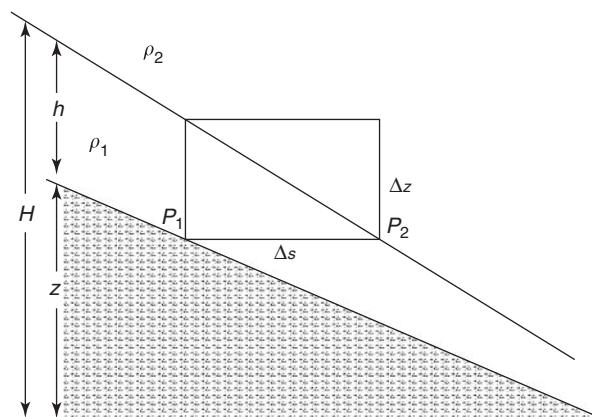


Figure 1 Cross-sectional view of two-layer fluid representative of the cold, katabatic wind layer (lower layer) beneath the undisturbed ambient atmosphere; p designates pressure, ρ designates density ($\rho_1 > \rho_2$). H is the height of the katabatic layer above a fixed reference, h is the depth of the katabatic layer above the ground, and z is the terrain height above a fixed reference.

the horizontal pressure gradient (P) in the lower layer can be defined as in eqn [1].

$$P = -g \frac{\Delta \rho}{\rho_1} \frac{\partial H}{\partial s} - \frac{1}{\rho_1} \frac{\partial p}{\partial s_{\text{top}}} \quad [1]$$

In eqn [1], g ($=9.81 \text{ m s}^{-2}$) is the acceleration of gravity, $\Delta \rho$ is the density difference ($\rho_1 - \rho_2$) between the bottom and top layers, ρ_1 is the density in the bottom layer, H is the height of the lower layer above some fixed reference level, and s is a horizontal length scale. Two terms act to force atmospheric motion in the lower layer. The first term represents the effects of cooling over the sloping surface to produce a negatively buoyant air stream. The second term depicts the pressure gradient force in the atmosphere above the lower layer in the horizontal direction along the terrain slope. It is thus evident that the horizontal pressure gradient force in the atmosphere above the cooled layer near the surface is transferred downward through the fluid.

Using the equation of state, eqn [1] may be rewritten in a more convenient form as eqn [2].

$$P = g \frac{\Delta T}{T_1} \frac{\partial H}{\partial s} - \frac{1}{\rho_1} \frac{\partial p}{\partial s_{\text{top}}} \quad [2]$$

Here ΔT is the temperature difference between the two layers ($T_1 - T_2$) and T_1 is the temperature in the bottom layer. In this form, ΔT can be viewed as the strength of the temperature inversion between the cooled layer and the ambient overlying atmosphere. In ordinary atmospheric situations, temperatures decrease with height. Under conditions of a temperature inversion, such as occurs during nocturnal conditions, the coldest air is situated closest to the ground and temperatures increase with height. The strength of the temperature inversion

can be determined from the measurement of temperature in the vertical direction, known as a sounding.

Recognizing that the height H of the density interface is just the sum of the terrain height and thickness of the lower layer, P can also be expressed as in eqn [3], where z is the height of the terrain and h is the depth of the cooled lower layer.

$$P = g \frac{\Delta T}{T_1} \frac{\partial z}{\partial s} + g \frac{\Delta T}{T_1} \frac{\partial h}{\partial s} - \frac{1}{\rho_1} \frac{\partial p}{\partial s_{\text{top}}} \quad [3]$$

The first term in eqn [3] represents the effect of a density contrast over sloping terrain. It can be seen that the size of this term is dependent on the magnitude of the density contrast and the slope of the terrain. This term is sometimes called the buoyancy force. The second term represents the influence of the gradient of thickness of the cooled layer column along the slope. For a situation in which the thickness of the lower layer decreases, as is shown schematically in Figure 1, this term will act to accelerate the downslope component of the flow. For cases in which the horizontal scales are larger than a few kilometers, the second term is generally smaller than the first term and can be neglected. This simplification can also be made if it is assumed that the inversion interface follows the terrain. Both remaining terms, the buoyancy force and the horizontal pressure gradient force in the ambient atmosphere, are important in the forcing of katabatic winds. Although developed for the two-layer model shown in Figure 1, eqn [3] can be applied directly to the atmosphere if potential temperature is substituted for temperature to incorporate the effects of compressibility.

Figure 2 is a nomogram illustrating the relationship between the inversion strength, terrain slope, and ambient pressure gradient. It can be shown that a temperature inversion

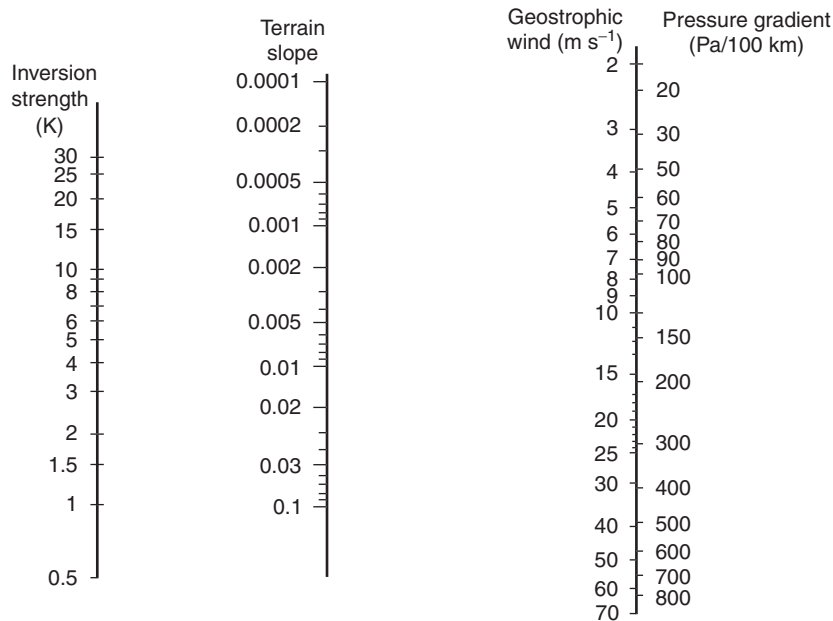


Figure 2 Nomogram showing the relationship between inversion strength, terrain slope, and the horizontal pressure gradient force. Extending a straight line between the inversion strength and terrain slope provides a measure of the pressure gradient force associated with katabatic winds. Calculations assume a katabatic layer potential temperature of 273 K; geostrophic wind magnitude is valid at 43° N.

over even gentle slopes can produce significant accelerations. Typical horizontal pressure gradients in the ambient atmosphere are on the order of 100 Pa over a horizontal distance of 100 km, which is equivalent to a geostrophic wind of approximately 8 m s^{-1} in middle latitudes. As can be seen from Figure 2, this is the same magnitude as the buoyancy force arising from a lower layer that is diabatically cooled by 10°C over an incline of approximately 1/500. For steeper inclines with modest temperature inversions, the buoyancy force can be seen to be considerably larger than typical values of the horizontal pressure gradient force in the ambient atmosphere. Because of the large magnitude of the buoyancy force and stable stratification of the atmosphere, katabatic winds respond to the local topographic configuration, often regardless of the pressure gradient in the free atmosphere.

Newton's second law governs the response of a free body to external forces acting on that body. In the case of the Earth's atmosphere, which rotates following the planetary angular velocity, it is advantageous to write Newton's second law in a coordinate system fixed to the Earth. A common method is to express the acceleration of an air parcel in terms of the specific forces (force divided by mass) to form the equation of motion. Atmospheric forces acting to accelerate air include the pressure gradient, gravity, and friction. The Coriolis force is an apparent force arising from the rotation of the Earth. The horizontal vector equation of motion for katabatic flow relative to the rotating Earth can be expressed as eqn [4].

$$\frac{D\mathbf{V}}{Dt} = \mathbf{P} - f\mathbf{k} \times \mathbf{V} + \mathbf{F} \quad [4]$$

Here D/Dt represents the time rate of change following a fluid parcel, $\mathbf{V}$ is the horizontal velocity, t is the time, f is the Coriolis parameter ($2\Omega \sin \varphi$, where Ω is the Earth's angular velocity of $7.29 \times 10^{-5} \text{ rad s}^{-1}$ and φ is the latitude), and $\mathbf{F}$ is the frictional drag. For small horizontal scales of motion of a few kilometers or timescales on the order of a few minutes, the effect of the Earth's rotation is insignificant and acceleration of katabatic flow is subject to the pressure gradient force $\mathbf{P}$ and friction. Under these conditions, the air accelerates directly downslope in the absence of strong forcing from the ambient environment. For larger scales of motion on the order of hundreds of kilometers or timescales on the order of a few hours, the Coriolis force needs to be considered. Acceleration provided by the Coriolis force is directed to the right (left) of the wind vector in the Northern (Southern) Hemisphere. For katabatic winds in which the Earth's rotation cannot be neglected, air streams are deflected across the terrain gradient, often at angles exceeding 45° from downslope. These conditions are found over the great continental ice sheets of Antarctica and Greenland. Observations show that the katabatic wind is deflected at angles across the fall line vector of the terrain that is consistent with effects of the Coriolis force.

Geographical Distribution of Katabatic Winds

Every continent on the Earth experiences katabatic winds. Occurrences are most evident during nocturnal conditions. Along the highlands and slopes of mountainous terrain, rapid

cooling of the surface takes place after sunset during periods of fair weather as the Earth's surface radiates energy that escapes to space. Air in contact with the ground also becomes cooled and hence becomes denser than air at the same horizontal level situated away from the terrain. The cold, dense air then begins to move downslope. Similarly, radiative cooling of air over an elevated plateau, especially one that is covered by snow, results in the development of a pool of cold air, and a shallow dome of high pressure becomes established. Additional accumulation of cold air or passage of extratropical cyclones may act to trigger the release of cold air through mountain gaps or valleys in the form of katabatic winds.

Descriptive names for katabatic wind events have been given to certain episodes that produce profound changes in the local weather or present peril to agricultural or other commercial activities. Nearly, all are the combined result of both cooling of the air stream and the horizontal pressure gradients in the ambient atmosphere. The mistral is a cold, northerly wind that originates in the elevated regions of the Alps, and descends into the Rhone valley and other low-lying regions along the south of France. It is best developed when low pressure forms in the Gulf of Genoa. The cold air often results in frost damage to vineyards. Bora is a name originally used to describe the cold northeast winds observed along the coast of the former Yugoslavia. This air stream originates over Russia and crosses the Carpathian and Alps mountain ranges and descends onto the usually warm shores of the Adriatic Sea. Bora cases are associated with a pressure field consisting of a cyclone near the Black Sea and an anticyclone over the European continent. Both mistral and bora are strong and gusty winds with maximum frequency of occurrence during winter. Other local katabatic winds include the oroshi, a cold wind found on the Pacific side of the mountains on the island of Japan and the coho, a westward surge of cold air off the elevated plateau regions through the Columbia Gorge along the northwest coast of the USA and southern British Columbia. Katabatic winds are not restricted to middle and high latitudes. Cold drainage flows known as *tehuantepecers* occur in the Gulf of Tehuantepec along the Pacific coast of Mexico. Katabatic winds have also been documented in South America along the eastern slopes of the Andes from northern Chile to central Colombia.

Katabatic winds are best developed over the great ice sheets of Antarctica and Greenland. The geographical position of the ice sheets and high reflectivity of the ice fields combine to limit the heating of the surface. Temperature inversions in the lowest 100 m over the elevated interior sections of Antarctica commonly exceed 25°C during nonsummer months. Observations show the surface winds to be consistently of a katabatic nature for much of the year. Surface winds over Antarctic ice slopes are nearly unidirectional, with approximately 90% of the observations from a 30° sector. The katabatic wind systems over Antarctica and Greenland are among the most persistent on the Earth.

To a first approximation, the near-steady flows over the ice sheets can be envisaged as a balance between the horizontal pressure gradient force, the Coriolis force, and the friction force. A simplified but useful approximation for the friction force is $\mathbf{F} = -k\mathbf{V}$, where k is a dimensionless constant having a magnitude between 10^{-2} and 10^{-3} . For steady conditions,

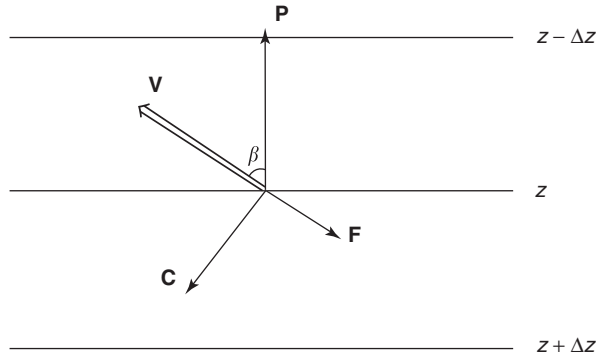


Figure 3 Horizontal view of force balance in the lower atmosphere for katabatic winds over Antarctica; z designates the terrain height, $\mathbf{V}$ is the katabatic wind vector, β is the angle between the wind and the downslope direction of the terrain, P is the horizontal pressure gradient force associated with atmospheric cooling over sloping terrain, C is the Coriolis force, and F is the friction force.

the scalar equations of motion from eqn [4] can be written as eqns [5] and [6].

$$0 = P - fv - \frac{k}{h}Vu \quad [5]$$

$$0 = fu - \frac{k}{h}Vv \quad [6]$$

Equation [5] is the component along the direction of the horizontal pressure gradient force P , which is typically in the downslope direction. Equation [6] is the equation of motion perpendicular to P . Again, V is the magnitude of the wind; $u = V \cos \beta$ and is directed along P ; and $v = V \sin \beta$ and is

directed to the left of and normal to P . The term β is the angle between P and V . The force balance arising from eqns [5] and [6] to produce the steady Antarctic katabatic wind is shown schematically in Figure 3. Equations [5] and [6] can then be solved for V and β (eqn [7]), provided estimates of the strength and depth of the temperature inversion, terrain slope, and pressure gradient force in the ambient environment are known.

$$V = \sqrt{\frac{hP \cos \beta}{k}} = \frac{P}{f} \sin \beta \quad [7]$$

For a given terrain slope, increasing values of P are accompanied by an increase in V and a decrease in β , and hence the flow is faster and directed more downslope. The observed behavior of Antarctic katabatic winds qualitatively matches that prescribed by this simplified model. Assessment of the steady-state wintertime Antarctic katabatic wind regime from eqn [7] is illustrated by streamlines in Figure 4. The force balance between pressure gradient, Coriolis force, and friction results in a katabatic wind vector that is directed across the height contours of the terrain at angles of 30–60° over the interior of the continent. Winds near the coast are directed more downslope. A counterclockwise, divergent katabatic wind pattern is present over the Antarctic ice sheet. A similar pattern can be found over Greenland, except that the motion is in a clockwise sense. By mass continuity, sinking motions must exist over the great ice sheets. A vertical circulation becomes established in the troposphere above the katabatic layer in response to the continental outflow. The strongest winds are found over the steepest ice slopes, which are generally near the coast.

Antarctic katabatic winds are not uniform, as can be seen in Figure 4. Convergence of drainage streamlines is present at various locations about the periphery of Antarctica. Cold,

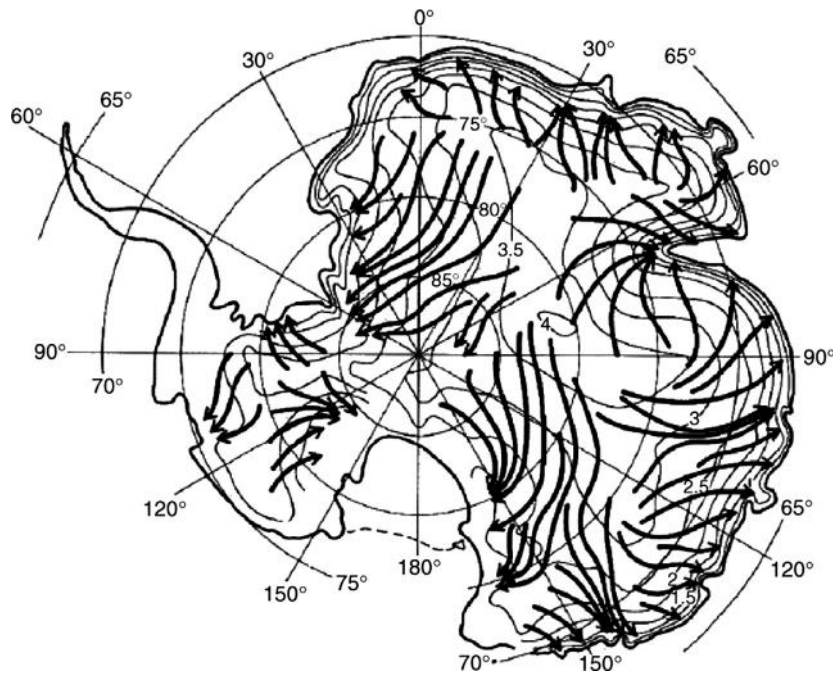


Figure 4 Streamlines of wintertime katabatic winds over Antarctica. Terrain contours (units of kilometers) of Antarctic ice sheet denoted by thin lines.

negatively buoyant air from a broad horizontal area becomes concentrated into a restricted pathway. These 'confluence zones' represent areas of enhanced cold air supplies available to feed katabatic winds downstream. The most intense and persistent katabatic winds are observed along and downstream of the axes of the confluence zones depicted in [Figure 4](#).

See also: **Arctic and Antarctic:** Antarctic Climate. **Boundary Layer (Atmospheric) and Air Pollution:** Stably Stratified Boundary Layer. **Dynamical Meteorology:** Coriolis Force; Overview. **Mountain Meteorology:** Cold Air Damming; Downslope Winds; Land and Sea Breezes; Overview; Valley Winds.

Further Reading

- Atkinson, B.W., 1981. Meso-scale Atmospheric Circulations. Academic Press, New York, NY.
- Barry, R.G., 1981. Mountain Weather and Climate. Methuen, London, UK.
- King, J.C., Turner, J., 1997. Antarctic Meteorology and Climatology. Cambridge University Press, Cambridge, UK.
- Schwerdtfeger, W., 1970. The climate of the Antarctic. In: Orvig, S. (Ed.), World Survey of Climatology, Vol. XIV. Elsevier, Amsterdam.
- Schwerdtfeger, W., 1984. Weather and Climate of the Antarctic. Elsevier, Amsterdam.
- Simpson, J.E., 1994. Sea Breeze and Local Winds. Cambridge University Press, Cambridge, UK.
- Yoshino, M.M. (Ed.), 1976. Local Wind Bora. University of Tokyo Press, Tokyo.
- Whiteman, C.D., 2000. Mountain Meteorology. Oxford University Press, Oxford, UK.

Land and Sea Breezes

RA Pielke, Sr., University of Colorado at Boulder, CO, USA

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

The physics of land and sea breezes are presented. This includes how they develop in calm conditions and the effect on these mesoscale wind circulations under larger scale wind flow. This article discusses, for example, how light offshore large-scale flow at the coast can amplify the magnitude of low wind convergence associated with a sea breeze, while a stronger wind can prevent the sea breeze from even developing. The reason that the land breeze is a shallower atmospheric feature is presented.

Of all mesoscale phenomena, sea and land breezes have been the most studied, both observationally and theoretically. This is undoubtedly a result of the geographically fixed nature of these phenomena (the location of land–water boundaries) as well as the repetitive nature of the event. The sea breeze is defined as occurring when the wind is onshore, whereas the land breeze occurs when the opposite flow exists. Detailed discussion of sea and land breezes is given in Simpson (1994, 2007), with briefer discussions in Pielke (1984, 2000), Lin (2007), and Atkinson (1981). Sea and land breezes that occur in association with larger lakes are called lake and land breezes (e.g., Neumann and Mahrer, 1975). The leading edge of the sea breeze winds is called the sea breeze front. The first numerical model of the sea breeze was completed by Estoque (1961).

During the case of calm large-scale winds and in flat terrain, it is comparatively easy to describe the diurnal variations of the coastal wind circulations. Defant (1951) presented an excellent qualitative description for this condition, which is illustrated in Figure 1. The idealized sequence of events is as follows:

1. At some time in the early morning, the pressure surfaces become flat and no winds occur (e.g., at 0800 LST – perhaps an hour after sunrise).

2. Later in the morning, mass is mixed upward over land by turbulent mixing in the unstably stratified boundary layer as well as due to the expansion of the volume of air due to its heating, creating an offshore pressure gradient at some distance above the ground (Tijm and van Delden, 1999; Nicholls and Pielke, 1994). Over water, the penetration of sunlight and resultant distribution of radiative heating with depth and the ability of water to mix minimize significant heating of the surface (e.g., at 1100 LST). The temperature of the water is not important in determining the strength of the sea breeze, as long as the air above is warmer than the water.
3. The resultant offshore movement of air above the ground near the coast creates a low-pressure region at the ground, and onshore winds (the sea breeze) develop (e.g., at 1300 LST).
4. The onshore winds transport cooler marine air over the land, thereby advecting the horizontal temperature gradient and, hence, the sea breeze inland. The distance the sea breeze travels inland depends most directly on the intensity of the total heat input to the air (Pearson, 1973; Tijm et al., 1999; Neumann, 1977) and the latitude (Rotunno, 1983) (e.g., at 1600 LST).

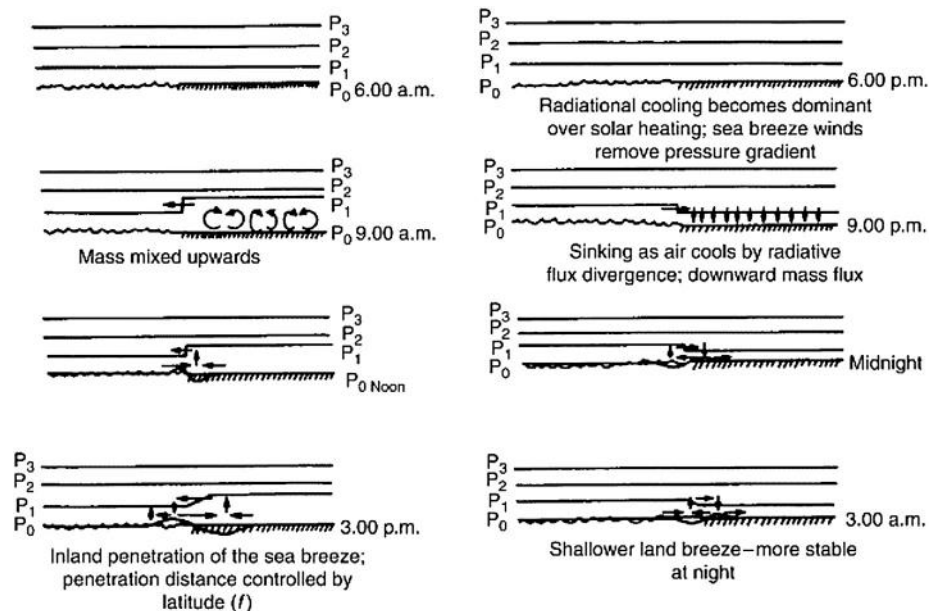


Figure 1 Schematic of the diurnal evolution of the sea and land breeze in the absence of synoptic flow. Reproduced from Pielke Sr., R.A., 1984. Mesoscale Meteorological Modeling, second ed. Academic Press, p. 676.

5. As the sun sets, longwave radiational cooling becomes dominant over solar heating, and the local wind field removes the horizontal temperature gradient. The pressure surfaces again become horizontal (e.g., at 1900 LST).
6. As longwave cooling continues and compresses, the air near the ground becomes denser and sinks. The resultant lowering of the pressure surfaces a short distance above the ground creates an onshore wind at that level (e.g., at 2200 LST).
7. In response to the loss of mass above the surface over the water, a pressure minimum develops at the ocean interface immediately off the coast. The offshore wind that then develops near the surface is called the land breeze (e.g., at 0100 LST).
8. The distance of offshore penetration of the land breeze depends on the amount of cooling over the land. Because the planetary boundary layer over land is stably stratified at night and, therefore, vertical mixing is weaker and closer to the ground, the land breeze is a shallower and weaker phenomenon than the daytime sea breeze.

There may even be a third, higher layer of flow associated with these local winds, which Tijm et al. (1999) refer to as a "return–return current."

When the coastline is irregular, local regions of enhanced or weakened low-level convergence develop, as illustrated for the daytime portion of the cycle in Figure 2. (Such zones of preferential convergence help explain the frequency of showers and thunderstorms in certain locations in south Florida during the summer, as seen, for example, in Figure 3 and discussed in Pielke et al., 1991.)

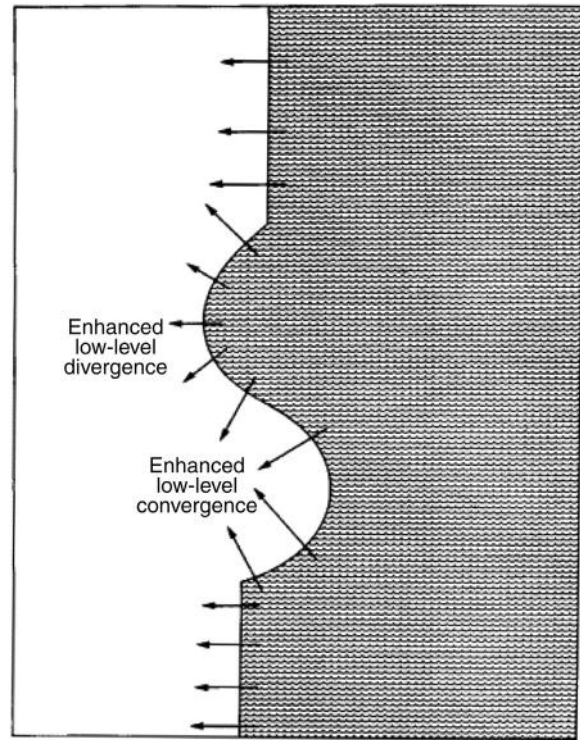


Figure 2 Schematic of the influence of the coastline configuration on a sea breeze in the absence of large-scale flow. Reproduced from Pielke Sr., R.A., 1984. *Mesoscale Meteorological Modeling*, second ed. Academic Press, p. 676.

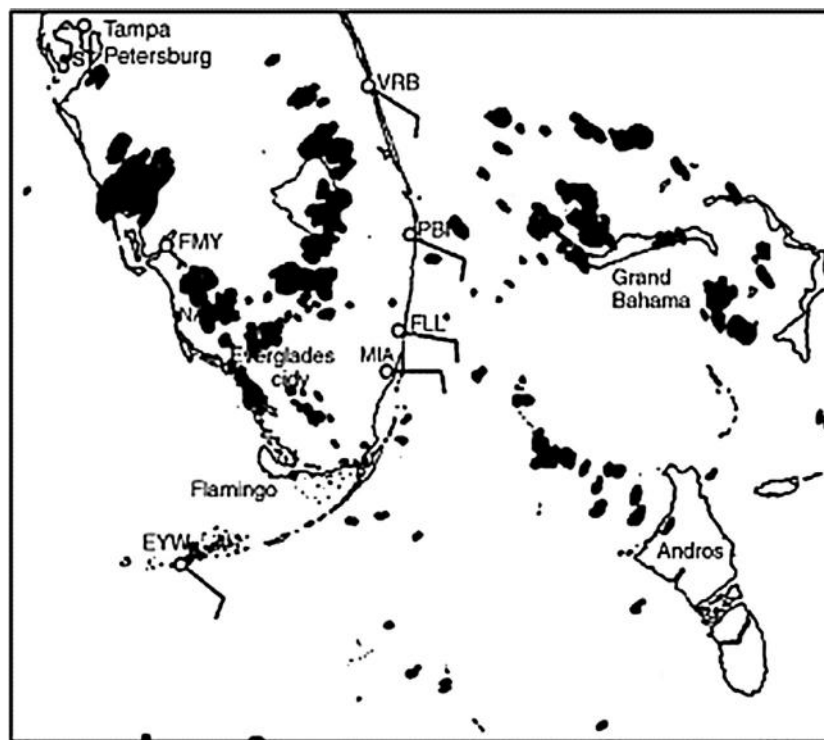


Figure 3 Radar echo coverage at 1501 EST 19 August 1971 as seen by the Miami WSR-57 10 cm radar. Reproduced from Pielke Sr., R.A., 1974. A three-dimensional numerical model of the sea breezes over south Florida. *Monthly Weather Review* 102, 115–139.

The evolution of the sea breeze is somewhat more complicated when a weak or moderate (i.e., $\lesssim 6 \text{ m s}^{-1}$) prevailing synoptic flow is included. For the two distinct situations of comparatively cold water and comparatively warm water relative to land, a synoptic wind direction from the colder to the warmer surface weakens the intensity of the local wind by diminishing the horizontal temperature gradient. By contrast, when a prevailing larger scale flow of the same strength is from the warmer to the colder surface, if the synoptic wind speed is not too strong, the temperature gradient is strengthened and the subsequent local wind flow is stronger. An example of this effect is shown in Figures 4 and 5, where the sea breeze wind convergence is more clearly evident when the large-scale wind is in the opposing direction from the sea breeze altered flow.

Examples of water that is warm relative to the land include the eastern sides of continents in the tropics and midlatitudes at night and over coastal waters during a polar outbreak. Situations with water that is cold relative to the adjacent land include the eastern sides of continents in the tropics and midlatitudes during sunny days, along the west

side of continents in which upwelling is occurring, as well as along polar coastal areas in the summer. Fog and low stratus often form over the relatively cold water in polar and upwelling ocean areas and move onshore associated with the sea breeze.

The magnitude of the effect of a particular horizontal temperature gradient can be estimated from existing observational and numerical studies. It has been found that, in the tropics and midlatitudes, a horizontal gradient of less than about 10 W m^{-2} per 30 km has only a minor influence on local wind patterns. With a gradient of 100 W m^{-2} per 30 km, however, significant effects are discernible from the statistical evaluation of observational data, whereas at 1000 W m^{-2} per 30 km, the influence on local wind patterns is very pronounced in case-by-case studies. With a nonzero large-scale wind, the heating must be greater in order for a sea breeze to develop. Using observational data, it has been shown that a sea breeze does not develop when the horizontal pressure gradient generated by the differential heating between land and adjacent water is insufficient to overcome the kinetic energy of the large-scale flow.

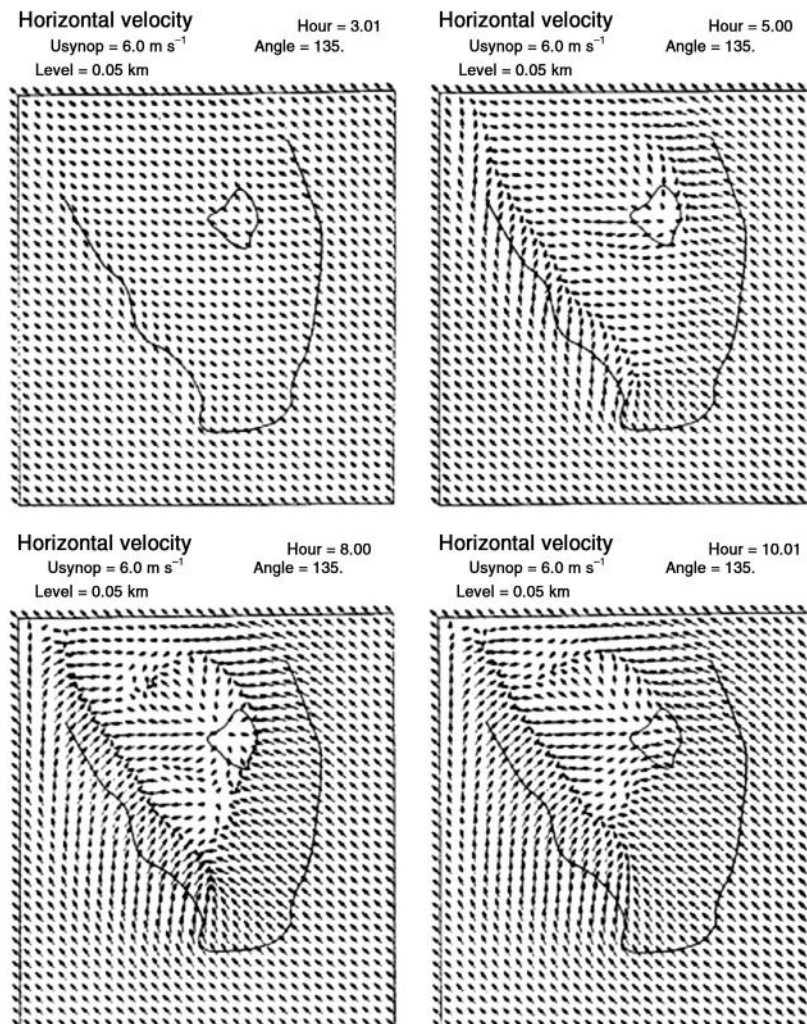


Figure 4 Horizontal wind at the 50 m level, 3, 5, 8, and 10 h after a simulated sunrise for a uniform synoptic southeast wind case over south Florida. Note how the shower pattern in Figure 3 closely corresponds to the wind convergence pattern in Figure 4. Reproduced from Pielke Sr., R.A., 1974. A three-dimensional numerical model of the sea breezes over south Florida. *Monthly Weather Review* 102, 115–139.

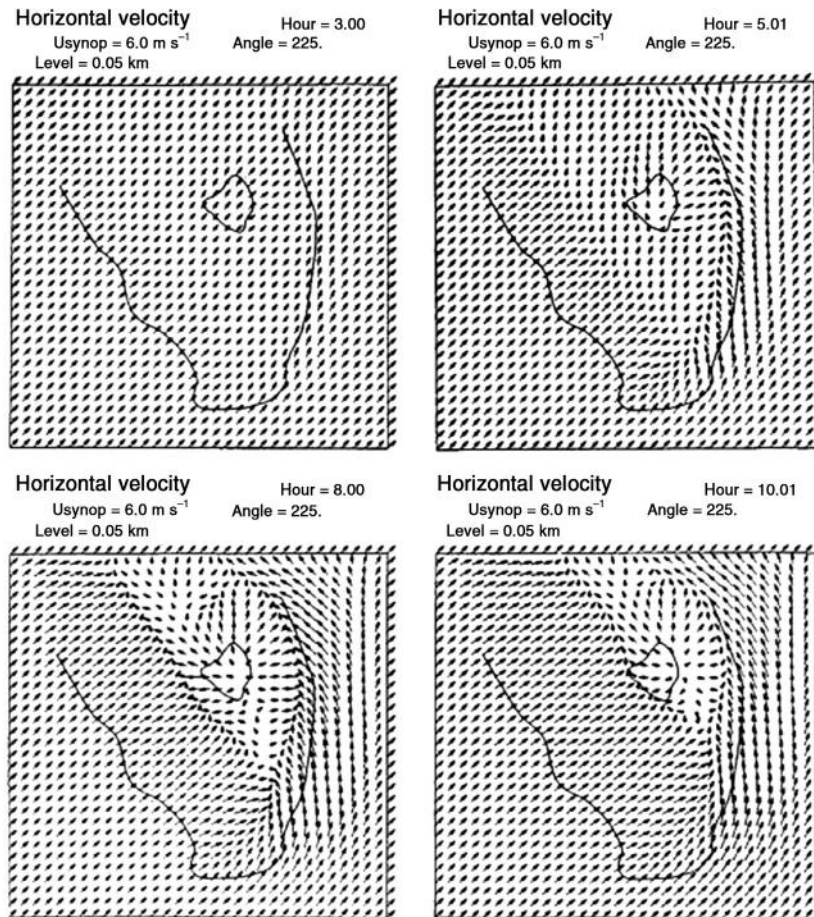


Figure 5 Same as Figure 4, except for a uniform synoptic southwest wind. Reproduced from Pielke Sr., R.A., 1974. A three-dimensional numerical model of the sea breezes over south Florida. *Monthly Weather Review* 102, 115–139.

When the coastal terrain is hilly or mountainous, sea and land breezes interact with local winds that are created as a result of the heating and cooling of this elevated terrain relative to the adjacent atmosphere at the same altitude. The sea breeze and upslope mountain flows that are created as the terrain is heated during the day, for example, can generate particularly strong onshore winds. However, the subsidence in the adjacent atmosphere caused by the upslope flow can inhibit the development of the sea breeze, resulting in an onshore wind that is less than the sum of the two winds. In addition, the intensity of combined local wind circulation tends to be less when the terrain slope is larger (Segal et al., 1983). Sea and land breezes can result in the accumulation of pollution as air recirculates over industrial and urban areas (Eastman et al., 1995).

See also: **Mesoscale Meteorology:** Mesoscale Convective Systems. **Tropical Meteorology and Climate:** Monsoon: Overview.

Bibliography

- Atkinson, B.W., 1981. *Mesoscale Atmospheric Circulations*. Academic Press. p. 495.
- Defant, F., 1951. Local winds. *Compendium of Meteorology*. American Meteorological Society, Boston, MA. pp. 655–672.
- Eastman, J.L., Pielke, R.A., Lyons, W.A., 1995. Comparison of lake-breeze model simulations with tracer data. *Journal of Applied Meteorology* 34, 1398–1418.
- Estoque, M.A., 1961. A theoretical investigation of the sea breeze. *Quarterly Journal of the Royal Meteorological Society* 87, 136–146.
- Lin, Y., 2007. *Mesoscale Dynamics*. Cambridge University Press.
- Neumann, J., 1977. On the rotation rate of the direction of sea and land breezes. *Journal of Atmospheric Sciences* 34, 1913–1917.
- Neumann, J., Mahrer, Y., 1975. A theoretical study of the lake and land breezes of circular lakes. *Monthly Weather Review* 103, 474–485.
- Nicholls, M.E., Pielke Sr., R.A., 1994. Thermal compression waves. II. Mass adjustment and vertical transfer of total energy. *Quarterly Journal of the Royal Meteorological Society* 120, 333–359.
- Pearson, R.A., 1973. Properties of the sea breeze front as shown by a numerical model. *Journal of Atmospheric Sciences* 30, 1050–1060.
- Pielke Sr., R.A., 1974. A three-dimensional numerical model of the sea breezes over south Florida. *Monthly Weather Review* 102, 115–139.
- Pielke Sr., R.A., 1984, 2000. *Mesoscale Meteorological Modeling*, second ed. Academic Press. p. 676.
- Pielke Sr., R.A., Song, A., Michaels, P.J., Lyons, W.A., Arritt, R.W., 1991. The predictability of sea breeze generated thunderstorms. *Atmosphere* 4, 65–78.
- Rotunno, R., 1983. On the linear theory of the land- and sea-breeze. *Journal of Atmospheric Sciences* 40, 1999–2005.
- Segal, M., Mahrer, Y., Pielke, R.A., 1983. A study of meteorological patterns associated with a lake confined by mountains – the Dead Sea case. *Quarterly Journal of the Royal Meteorological Society* 109, 549–564.
- Simpson, J.E., 1994, 2007. *Sea Breeze and Local Wind*. Cambridge University Press, New York. p. 234.
- Tijm, A.B.C., van Delden, A.J., 1999. The role of the sound waves in sea-breeze initiation. *Quarterly Journal of the Royal Meteorological Society* 125, 1997–2018.
- Tijm, A.B.C., van Delden, A.J., Holtslag, A.A.M., 1999. The inland penetration of sea breezes. *Contributions to Atmospheric Physics* 72, 317–328.

Lee Vortices

CC Epifanio, Texas A&M University, College Station, TX, USA

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

Lee Vortices refer to the flow of the atmosphere around mountain barriers resulting in a pair of counterrotating eddies downstream. This article discusses some of the key topographic flow types revealed by laboratory experiments, numerical modeling, and theory.

Introduction

Flow of the atmosphere past a mountain barrier often results in a pair of counterrotating eddies immediately downstream, a flow pattern referred to as *lee vortices*. In some cases, these vortices can be quite persistent, at times lasting for as long as several days. A particularly well-studied example is the island of Hawaii, which often produces stable, quasisteady wakes (Figure 1(a)). However, in other situations, the counterrotating vortex flow turns out to be unstable, and a transition occurs to a state in which vortices of alternating sign are periodically shed downstream. As the eddies drift away, they form a flow structure referred to as a *vortex street*, which is sometimes captured in satellite images. A particularly striking example is shown in Figure 1(b), for flow past an island off the coast of Chile.

Observations show that vortices and wakes are relatively common in mountainous regions, over both the land and the water, and can have important impacts on local weather conditions. Vortices that form near urban areas often influence local air quality, as pollutants are continually recirculated instead of flushing away downstream. Some prominent examples include the Melbourne Eddy (near Melbourne, Australia), the Santa Barbara Eddy (near Santa Barbara, California), and the Denver Cyclone (near Denver, Colorado). Wake eddies have also been linked to the initiation and intensification of severe storms. A well-studied example is the Denver Cyclone, which is often associated with severe

convection, resulting in hail, flooding, and tornadoes. Studies of flow past the Alps suggest that on larger scales, mountain wakes may also play a role in lee cyclogenesis, as the vortices interact with upper level troughs to promote the deepening of synoptic-scale systems.

Lee vortices develop on time scales that are short compared to a day, and have typical length scales on the order of 10–100 km. As a result, in most cases, the rotation of the Earth has only a secondary impact on the motion. Most studies of lee vortices have neglected the role of the Coriolis force (i.e., have considered nonrotating flow), and we adopt the same constraint in the following sections. There is some evidence that for longer temporal and larger spatial scales, the Earth's rotation tends to suppress vortex formation; however, in general, the effect of the Coriolis force on lee vortices has received only modest attention.

Stratified Flow Past Topography: Basic Phenomenology

Overview

Topography can lead to a wide range of flow responses, depending on the background conditions and the size of the barrier. To provide context for our lee-vortex discussion, we first present a brief overview of some of the key topographic flow types, as revealed by theory, numerical modeling, and

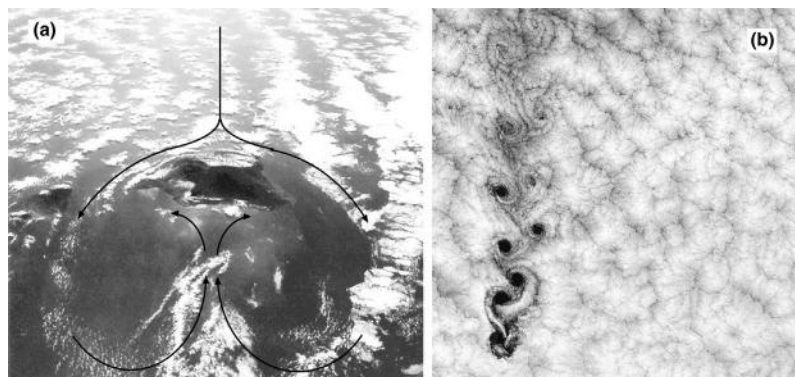


Figure 1 (a) Aerial photo of cloud layer in flow past the island of Hawaii. The cores of lee vortices feature warmer air than the surrounding flow, and thus often manifest as holes or breaks in the cloud layer. The lobes of clear air extending downstream of the island in (a) are thus the signature of a pair of counterrotating vortices. Arrows suggest the flow field as inferred from the cloud pattern. (Photo courtesy of Vanda Grubišić.) (b) Satellite image showing a vortex street downstream of Alejandro Selkirk Island, off the coast of Chile. The island is in the bottom left part of the figure.

laboratory experiments. For simplicity (and brevity), we focus on a particular prototype problem: namely, flows past an isolated obstacle, with constant upstream wind U and static stability (or Brunt–Väisälä frequency) N , and without Coriolis effects. While such a simplified model excludes certain, important types of flows (such as trapped lee waves, which depend on changes in N and U), it nonetheless provides a wide range of insights into a variety of real-world disturbances.

We suppose the obstacle shape can be described by a streamwise length scale a , a cross-stream length scale b , and a maximum height h . The key parameters governing the behavior of the flow are then (1) the nondimensional mountain height $\epsilon = Nh/U$, which measures the amplitude (or nonlinearity) of the disturbance; (2) the vertical aspect ratio $\delta = U/Na$, which measures the importance of nonhydrostatic effects; and (3) the horizontal aspect ratio $\beta = b/a$, which measures the effective width of the barrier. In most flows of interest, the vertical aspect ratio δ is relatively small (on the order of 0.1), implying the flow is essentially hydrostatic. For the most part, then, the set of control parameters can be reduced to just ϵ and β .

Laboratory, numerical, and theoretical studies suggest that our simplified prototype problem supports four main classes of flow response: (1) small-amplitude waves; (2) wave overturning and breaking; (3) upstream stagnation and flow-splitting; and (4) lee vortices. The regime diagram in Figure 2 illustrates the occurrence of these phenomena as a function of ϵ and β . Brief overviews of the classes are as follows:

Small-Amplitude Waves

When $\epsilon \ll 1$, the mountain-induced disturbance for all β takes the form of a small-amplitude mountain wave (see **Mountain Meteorology: Lee Waves and Mountain Waves** for examples of small-amplitude waves). The flow in this regime is well described by linear theoretical approaches, which are valid in the limit of small ϵ . As ϵ increases, the streamlines in the wave become more steeply inclined in the vertical, as driven by nonlinear (or finite-amplitude) effects.

Wave Breaking

As long as $\beta \gtrsim 1$ (i.e., for elongated ridges), the streamlines in the wave pattern will steepen to the point of overturning once ϵ

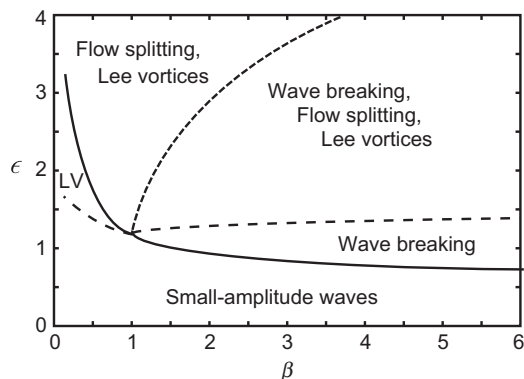


Figure 2 Schematic regime diagram for steady stratified flow past an isolated ridge, as a function of ϵ and β . Note that the actual positions and shapes of the regime boundaries will depend somewhat on obstacle shape.

exceeds some critical value $\epsilon_b(\beta)$, which is usually in the range 0.7–1.2, depending on the obstacle shape. Overturning of the streamlines places more dense fluid over less dense fluid, thus causing the wave to break and become locally turbulent. An example with overturning streamlines and wave breaking is shown Figure 3(a), as seen in a laboratory experiment with $\epsilon = 2.5$ and $\beta = 3$. Wave breaking and the associated turbulence leads to an elevated, well-mixed region above the lee slope, which behaves almost like a free surface, dynamically

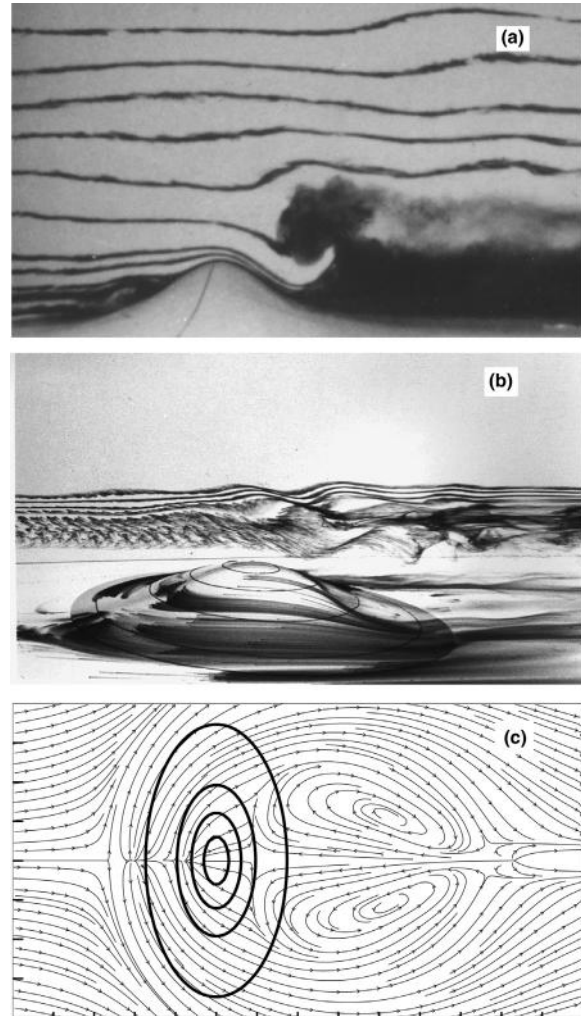


Figure 3 (a) Dye lines in the centerline plane for flow past a 3D obstacle with $\epsilon = 2.5$ and $\beta = 3$. The streamline above the lee slope has steepened to the point of overturning, causing the wave to break and the flow to become turbulent. (From Castro, I. P., Snyder, W.H., 1993. Experiments on wave breaking in stratified flow over obstacles. *Journal of Fluid Mechanics* 255, 195–211.) (b) Dye lines in the centerline plane and on the obstacle surface for flow past a ridge with $\epsilon = 2.5$ and $\beta = 2$. The basic flow is from left to right. Flow-splitting is apparent on the windward slope below the second height contour. Also apparent is a wake in the lee, with a hint of possible reversed flow along the centerline. (From Baines, P.G., 1995. *Topographic Effects in Stratified Flows*. Cambridge University Press, 482 pp.) (c) Surface streamlines from a free-slip numerical computation using an obstacle identical to that shown in (b). Note that the dye lines in (b) are suggestive of the streamlines in (c).

decoupling the low-level flow from the flow farther aloft. The flow below the well-mixed region shows a striking similarity to supercritical shallow-water flow, with a rapidly decreasing layer depth and strongly accelerated winds (see **Mountain Meteorology: Downslope Winds** for a discussion of shallow-water flow over a ridge; see also **Dynamical Meteorology: Hydraulic Flow** for a basic discussion of shallow-water dynamics). Downstream of the obstacle, the depressed streamlines abruptly return to their upstream heights, in a structure resembling a shallow-water hydraulic jump (**Figure 3(a)**).

Flow Splitting

Regardless of β , there exists a critical mountain height $\varepsilon_s(\beta)$ at which the flow stagnates on the upstream face of the ridge. The stagnation is caused by the dynamic pressure anomaly in the wave pattern, and occurs at a finite height z_s on the upstream face. Below z_s , the flow splits and passes primarily around the obstacle, instead of over, while above z_s the flow ascends to the peak. **Figure 3(b)** shows an example of upstream stagnation and flow splitting in a laboratory flow, with $\varepsilon = 2.5$ and $\beta = 2$; streamlines from a numerical computation with the same ε and β are shown in **Figure 3(c)**. Flow splitting limits the vertical displacements in the flow, and thus has an inhibitory effect on wave overturning and breaking. For narrow obstacles (i.e., $\beta \leq 1$), splitting is reached at smaller ε than breaking (i.e., $\varepsilon_s < \varepsilon_b$), and breaking is then suppressed for all ε . On the other hand, for elongated obstacles breaking occurs before splitting, and the phenomena are observed to coexist over some range in mountain heights (cf **Figure 2**).

Lee Vortices

A pair of counterrotating vortices appears in the lee of the obstacle whenever ε exceeds some critical value $\varepsilon_c(\beta)$ (**Figure 3(b)** and **3(c)**). In laboratory studies, persistent, counterrotating eddies have been observed only when upstream flow splitting is also present, as in **Figure 3(b)**. However, there is at least some evidence from numerical simulations that for certain flow conditions (particularly for $\beta < 1$), vortices may occur without upstream splitting. In any case, it seems to be true that for most obstacle shapes, splitting, and vortices occur at roughly the same obstacle height, and the two curves are taken to be coincident in **Figure 2**, at least for $\beta > 1$. Vortices with weak flow reversal are likely to be stable and quasisteady. Vortex pairs with more extensive flow reversal are expected to be unstable, and transition to a vortex-shedding state.

Boundary Layers and Free-Slip Models

Consider a laboratory experiment consisting of flow past a stationary topographic obstacle, such as the examples

illustrated in **Figure 3(a)** and **3(b)**. At the exact surface of the obstacle, the fluid is constrained to be motionless, due to frictional coupling with the solid boundary. This assumption is referred to as the *no-slip condition*, and is generally observed to be quite accurate. Immediately above the boundary is a thin shear zone, referred to as a *boundary layer*, above which the flow finally reaches its normal, free-stream speed, as illustrated in **Figure 4(a)**. The depth of the shear zone varies with the strength of the viscosity, but in general the thickness is quite small, much smaller than the size of the obstacle itself. Above the boundary layer, the gradients in the flow are much more gradual, and the associated viscous effects are correspondingly weak.

In the atmosphere, the boundary layer is more complex than that found in a laboratory experiment. Nonetheless, the basic premise of a thin shear zone still applies. For our purposes, we might define the boundary layer thickness to be the depth over which shear effects dominate those of stratification. (In the atmospheric context, this surface shear zone is usually referred to as the *surface layer*, to distinguish it from the deeper *atmospheric boundary layer*, where shear is much weaker.) For a stably stratified atmosphere, this depth is typically on the order of a few tens of meters, much smaller than the types of topographic features associated with lee vortices.

Since the boundary layer is relatively thin, in many cases we can simply ignore it, and assume the flow passes freely along the obstacle surface with no coupling to the boundary below. This approach is referred to as the *free-slip approximation*, and is generally adequate as long as the shear zone remains attached to the obstacle surface. However, in many cases the shear zone does not remain attached, but instead lifts from the obstacle through a process referred to as *boundary-layer separation*. In general, the separation occurs in a region of adverse pressure gradient, where the flow features significant convergence. This convergence causes the layer to be lifted, and ultimately detach from the surface, as illustrated schematically in **Figure 4(b)**. When the boundary layer separates, it carries significant shear vorticity to the fluid interior, and over time, this vorticity can curl into eddies, much like those seen **Figure 1**. Indeed, for many applications (such as flows past cars, airplanes, buildings, etc.), the associated wakes and vortices are entirely due to separating boundary layers.

When lee vortex streets were first captured in satellite images in the early 1960s, it was generally assumed that the vorticity of the eddies was the result of boundary-layer separation. In particular, the cases observed at the time were all found to have strong stable layers below the obstacle crest, which lead to flow splitting on the upstream face of the

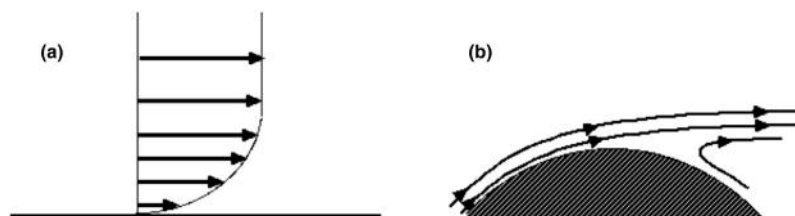


Figure 4 (a) Schematic close-up of the boundary layer at the surface of a solid obstacle. (b) Separation of the boundary layer.

obstacle, as in [Figure 3\(b\) and 3\(c\)](#). As the flow passed laterally around the barrier, it was assumed to acquire significant vertically oriented vorticity in the boundary layer, through its coupling with the obstacle surface. Once the boundary layer separated, this vorticity would be shed to the fluid interior, thus leading to eddies circulating about vertical axes.

This boundary-layer explanation remained dominant until the mid-1980s, when simulations of flow past realistic 3D topography first became computationally feasible. One of the early applications of these 3D simulations was a study of flow past the island of Hawaii, as part of the Joint Hawaii Warm Rain Project. Somewhat unexpectedly, the Hawaii researchers soon happened on an interesting discovery: namely, that even when the surface of the island is specified as free-slip, the simulations still produce realistic wakes and vortices, much like those seen in analogous laboratory experiments. As the free-slip condition completely neglects the boundary layer, these simulated vortices are clearly not the result of boundary-layer separation. About the same time, investigators studying other topographic flows found that by decreasing surface friction in their models, they could actually *increase* the strength of the simulated lee eddies. Again, this pointed to something other than the boundary-layer explanation.

Over the past two decades, researchers have mainly focused on processes other than boundary-layer separation as an explanation for lee vortex formation. As explained below, the rotation of the eddies is instead assumed to originate in the fluid interior, either through buoyancy gradients (i.e., baroclinicity), or else through turbulent stresses. In the subsequent sections, we focus mainly on this more recent body of work. That said, it should be kept in mind that for certain flow conditions (most notably, for large δ or for very large ε – say, for $\varepsilon \gtrsim 10$), boundary-layer separation may still play an important role.

Vortex Formation

As with most vortex flows, the process of lee-vortex formation is perhaps best described in terms of the evolution of the vorticity distribution.

Fundamentals

For simplicity, we take the typical lee-vortex flow to be density stratified but essentially incompressible (as is generally the case for most flows of interest). The three-dimensional (3D) vorticity vector $\boldsymbol{\zeta}$ is then governed by

$$\frac{D\boldsymbol{\zeta}}{Dt} = (\boldsymbol{\zeta} \cdot \nabla)\mathbf{u} + \frac{\nabla\rho \times \nabla p}{\rho^2} + \nabla \times \mathbf{F} \quad [1]$$

where $\mathbf{u}$ is the velocity vector, p is the pressure, ρ is the density, and $\mathbf{F}$ represents the effects of viscous or turbulent stresses. The D/Dt notation in [eqn \[1\]](#) indicates a material (or Lagrangian) derivative.

Each of the terms on the right in [eqn \[1\]](#) has a physically distinct impact on the vorticity evolution. The first term is the *stretching and tilting* term, which effectively treats the

vorticity like a small elastic band embedded in the flow. As the flow evolves, the band (and hence the vorticity) is continually deformed by the wind field, with the deformations taking the form of tilting and stretching of the band. The second term on the right is the *baroclinic* term, which describes the rotation generated when a pressure gradient acts on an air mass of variable density. For topographic (and most other) flows, this baroclinic generation is well approximated by

$$\frac{\nabla\rho \times \nabla p}{\rho^2} \approx -\mathbf{k} \times \nabla b \quad [2]$$

where $\mathbf{k}$ is the vertical unit vector and b is the buoyancy. To a good approximation, then, baroclinic generation can be thought of in terms of horizontal buoyancy gradients, and the associated tendency toward vertical motion. The final term in [eqn \[1\]](#) is the *viscous* term, which describes vorticity forcing through viscous and/or turbulent stresses.

Note that according to [eqn \[2\]](#), any vorticity generated through baroclinic effects is strictly horizontal. The development of vertical vorticity thus depends on one of two processes: (a) the direct forcing of vertical vorticity by viscous/turbulent stresses; or else (b) the tilting of horizontal vorticity into the vertical by gradients in the vertical velocity field.

Finally, a full description of the flow requires the first law of thermodynamics, which we can write in the form

$$\frac{D\theta}{Dt} = H \quad [3]$$

where θ is the potential temperature and H is proportional to the diabatic heating. Combining [eqn \[1\]](#) with [eqn \[3\]](#) (and again assuming incompressibility) then leads to the useful conservation relation

$$\frac{DQ}{Dt} = \boldsymbol{\zeta} \cdot \nabla H + (\nabla \times \mathbf{F}) \cdot \nabla \theta \quad [4]$$

where $Q = \boldsymbol{\zeta} \cdot \nabla \theta$ is the *potential vorticity* (the potential vorticity is usually defined as $\Pi = \boldsymbol{\zeta} \cdot \nabla \theta / \rho$, where ρ is the density. By this standard definition, $Q = \rho \Pi$ is then the potential vorticity per unit volume, in the same sense that ρu is the x momentum per unit volume. As this is cumbersome terminology, we refer to Q simply as the potential vorticity) (see **Dynamical Meteorology: Potential Vorticity**).

According to [eqn \[4\]](#), for flow which is both adiabatic ($H = 0$) and inviscid ($\mathbf{F} = 0$), the potential vorticity maintains a constant value along particle trajectories. In the particular case in which the upstream potential vorticity is zero (as in the example below), this further implies that the vorticity vector must remain tangent to isentropic surfaces (or surfaces of constant θ), at least until the onset of dissipative processes.

Vertical Vorticity

Much of the analysis of lee vortices has been based on idealized modeling experiments, such as that shown in [Figures 5 and 6](#). In the example shown, the flow consists of a uniform wind started impulsively from rest at time $t = 0$ (much as in a towing tank experiment). The upstream static stability is roughly analogous to a trade-wind flow, with an inversion layer extending from $h/2$ to $0.7 h$ (with $N = 0.02$), with weaker

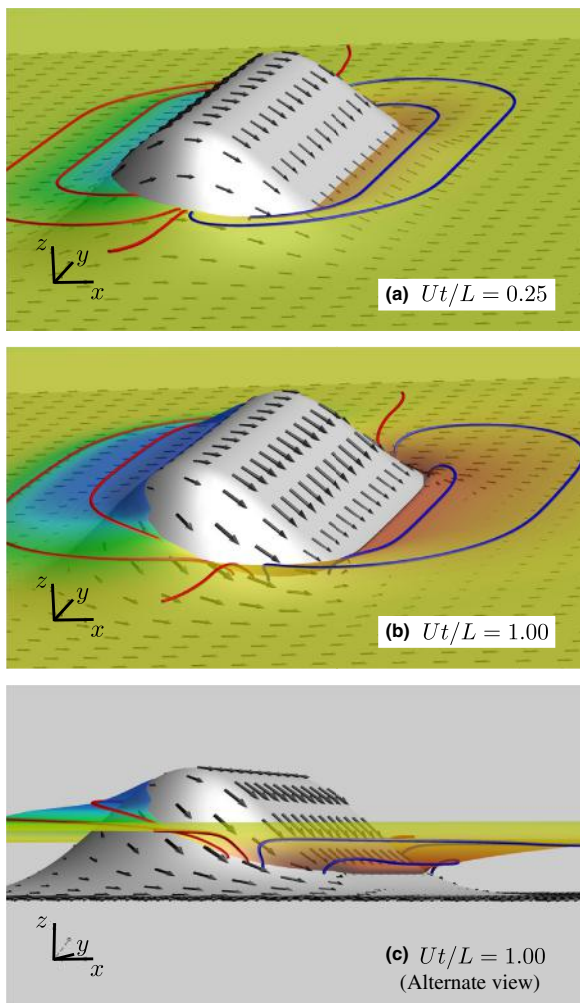


Figure 5 Early evolution for a stratified numerical simulation with flow accelerated rapidly from rest (from left to right), with an inversion layer below obstacle crest. (See text for details.) Shown is an isentropic surface near the base of the inversion, with the coloring of the surface indicating buoyancy (red colors positive, blue colors negative). The red and purple tubes show selected vortex lines running roughly parallel to the θ -surface, with the coloring of the lines indicating the direction of the vorticity. Red lines circle the obstacle in a clockwise sense, and purple lines counterclockwise. Vectors show the flow field at the ground. Times shown include (a) $Ut/L = 0.25$ and (b) $Ut/L = 1.00$, where U is the background wind and L is the obstacle half-width. (c) Same as (b), but from an alternate viewing angle that highlights the tilting of the vortex lines.

stability aloft ($N = 0.005$) and below ($N = 0.01$). Both the initial and upstream states have zero vorticity and thus satisfy $Q = 0$.

Figure 5 shows the early stages of evolution for the experiment described above, which highlights the development of circulation about a vertical axis. Shown in the figure is an isentropic surface near the base of the inversion layer, along with the buoyancy and vortex lines (a vortex line is a curve in space along which the vorticity vector is everywhere tangent. Vortex lines are thus to the vorticity field what streamlines are to the velocity) on the isentropic surface and

the velocity vectors at the ground (see the figure caption for details). Note that for ease of interpretation, the vertical scale of the figure has been exaggerated by a factor of roughly seven: in reality, the height of the obstacle is significantly less than its width.

For the early development shown in Figure 5, the basic flow structures can be described reasonably well in terms of Taylor series in time – specifically, by expanding eqns [1] and [3] as Taylor series about $t = 0$, and then grouping terms of like order in Δt . While the details of the expansion are somewhat involved, the end results suggest the following stages of development:

1. At $O(\Delta t)$, the main event is the production of buoyancy anomalies. Specifically, the ascending flow on the upstream face of the obstacle leads to negative buoyancy, while positive buoyancy is produced in the lee.
2. Once buoyancy anomalies exist, the vorticity follows through the baroclinicity term eqn [2]. The pattern of vorticity (as determined by the cross product) takes the form of two sets of loops: a set of clockwise loops on the upstream side, and a set of counterclockwise loops downstream (Figure 5(a)). The interface between the two sets is found roughly parallel to the ridgeline (at $x = 0$). At this stage, the vorticity is strictly horizontal.
3. With time, the pattern of vorticity loops is advected downstream, so that the interface between the clockwise and counterclockwise loops is found along the lee slope (Figure 5(b)).
4. Eventually (technically at $O((\Delta t)^4)$), as the vorticity pattern shifts downstream, the vorticity loops are tilted downward along the edges by the descending flow over the lee slope, thus producing vertical vorticity (Figure 5(b) and 5(c)). The sense of the tilting is to produce positive vertical vorticity on the right side of the flow (facing downstream) and negative on the left.

At the early stages shown in Figure 5, the developing flow is roughly adiabatic and inviscid, suggesting that vortex lines remain embedded in isentropic surfaces. Thus, as the θ -surface is dragged downward by the descending flow, the vortex lines on the surface are dragged downward as well. As seen below, these regions of tilting and vertical vorticity production later become the centers of the developing vortex.

Evolution and Reconnection

The subsequent evolution of the flow to the point of fully formed vortices is illustrated in Figure 6. As seen in the figure, the flow downstream of the obstacle quickly reaches a state of stagnation, which halts the downstream progression of the isentropic surface along the ground (Figures 5(b) and 6(a)). However, at the wake edges the flow remains accelerated, and the resulting shear zone then tends to deform the isentropic surface along either side, as illustrated in Figure 6(a)–6(c). This deformation of the θ -surface implies significant lateral buoyancy gradients at the wake edges, which ultimately causes the air in the wake region to spread laterally (Figure 6(c)). In time, the spreading wake air reconnects with the air outside the wake

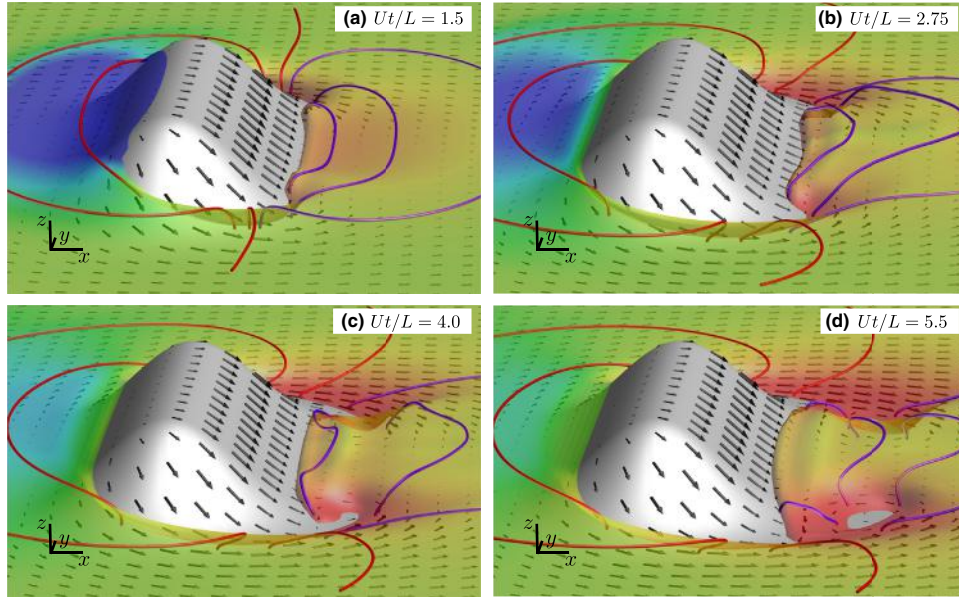


Figure 6 Same as [Figure 5](#), but for $Ut/L =$ (a) 1.5, (b) 2.75, (c) 4.0, (d) 5.5. After Epifanio, C.C., Rotunno, R., 2005. The dynamics of orographic wake formation in flows with upstream blocking. *Journal of Atmospheric Science* 62, 3127–3150.

([Figure 6\(d\)](#)), and the flow reaches its mature state, with a pair of fully recirculating lee eddies.

Interestingly, the reconnection of the flow as seen in [Figure 6\(d\)](#) has the effect of pinching off a hole in the θ -surface at the vortex center. Since the air above the inversion is warmer and (typically) drier than the air below, this pinched off area often shows up as a hole in the cloud layer in satellite images, as seen in [Figure 1\(b\)](#).

The Retracting Piston

[Figure 7](#) shows the flow evolution in the obstacle's centerline plane ($y = 0$), which highlights the development of reversed flow. For the purposes of illustration, the flow is shown in a reference frame that moves with the background wind – that is, in the reference frame of the figure, the background state is seen to be at rest, while the obstacle moves through the fluid (much as in a towing-tank experiment).

As the obstacle is set in motion, the air upstream of the obstacle below the inversion layer quickly becomes blocked ([Figure 7\(b\)–7\(d\)](#)), which allows warmer air from aloft to descend the lee slope. This descending air from above sets up a strong temperature contrast with the air downstream, which quickly collapses to form a lee-side front (or hydraulic jump-like structure), as seen in [Figure 7\(b\)–7\(d\)](#). The strong baroclinicity across the front leads to a plume of positive (i.e., into the plane of the figure) vorticity in the inversion layer downstream, which in turn implies a current in the upstream direction in the air below the inversion. As the topography is pulled away, the air below the inversion thus propagates into the space vacated by the obstacle, much as a density current would propagate into warmer air. As seen from the ground-based frame, this lee-side propagation appears as a deceleration and eventual reversal of the lee-side flow.

The flow development shown [Figure 7](#) has a close analog in the theory of shallow-water fluids, in which a piston at one end of a container of fluid is abruptly pulled away from the fluid at $t = 0$. As the piston retracts, the fluid adjusts under gravity so as to flow into the space vacated by the piston, much as the colder air downstream flows into the space vacated by the obstacle in [Figure 7](#). Interestingly, the same type of evolution (including flow reversal) occurs even in flows past 2D topography – that is, with respect to flow reversal, at least, the dynamics of wake formation is not inherently 3D.

Potential Vorticity Dynamics

A different but complementary view of lee vortices follows from a consideration of dissipative effects in the flow. As suggested by [eqn \[4\]](#), a key impact of dissipative processes is the production of potential vorticity (PV) anomalies.

The PV-Bernoulli Relation

The PV relation [eqn \[4\]](#) is often usefully considered in its equivalent conservation-flux form

$$\frac{\partial Q}{\partial t} = -\nabla \cdot \mathbf{J} \quad [5]$$

where

$$\mathbf{J} = \mathbf{u}Q - \zeta\mathbf{H} + \nabla\theta \times \mathbf{F} \quad [6]$$

is referred to as the flux of PV. According to [eqn \[5\]](#), PV is created locally wherever the PV flux [eqn \[6\]](#) is convergent or divergent. The PV flux $\mathbf{J}$ can in turn be divided into two parts: an advective part $\mathbf{u}Q$ which describes the transport of PV by the

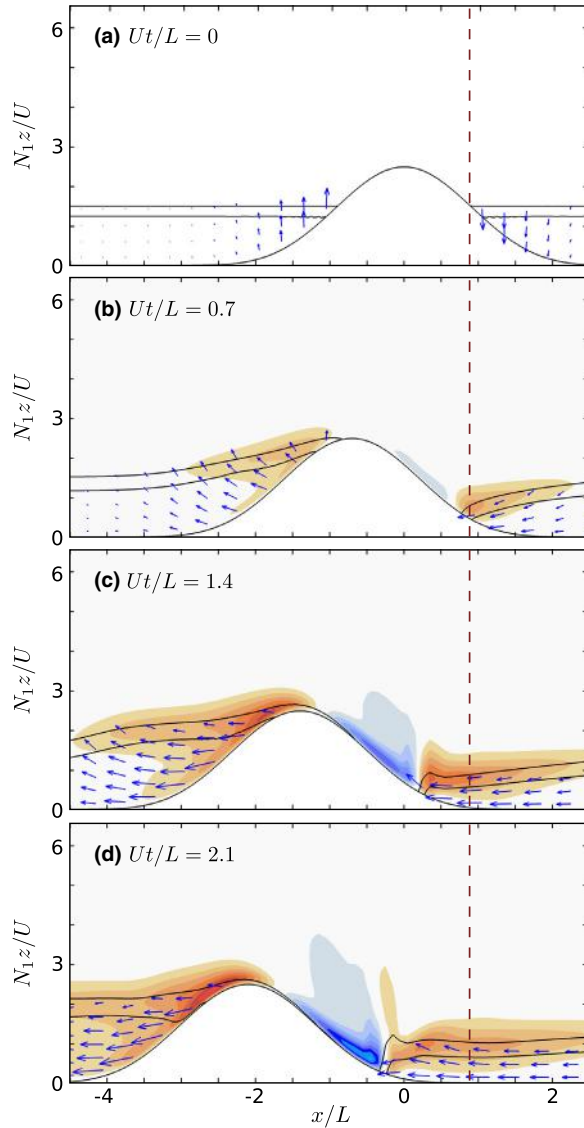


Figure 7 Flow in the centerline plane of the obstacle for the numerical simulation in [Figures 5 and 6](#). Fields are shown in a reference frame moving with the background wind, so that the obstacle appears to move leftward. Shown are the vorticity in the plane of the figure (red colors into the page, blue out of the page) and the velocity vectors in the figure's frame of reference. Black contours show isentropic surfaces at roughly the top and bottom of the inversion layer. Shown are times spanning $0 \leq Ut/L \leq 2.1$, with a time interval of $Ut/L = 0.7$. After Epifanio, C.C., Rotunno, R., 2005. The dynamics of orographic wake formation in flows with upstream blocking. *Journal of Atmospheric Science* 62, 3127–3150.

wind, and a dissipative part, whose convergence is the Lagrangian tendency in [eqn \[4\]](#).

In the case of lee vortices, the PV flux is often described in terms of its relation to the Bernoulli function

$$B = \frac{\mathbf{u} \cdot \mathbf{u}}{2} + c_p T + gz \quad [7]$$

where T is the sensible temperature, c_p is the specific heat at constant pressure, and g is the gravitational constant (see

Bernoulli Function). The Bernoulli function satisfies its own conservation relation, specifically

$$\frac{DB}{Dt} = \mathcal{J} + \mathbf{u} \cdot \mathbf{F} + \frac{1}{\rho} \frac{\partial p}{\partial t} \quad [8]$$

where $\mathcal{J} = c_p TH/\theta$ is the diabetic heat source. If the flow is steady, then the Bernoulli function remains constant along streamlines, except where acted upon by dissipation (much like the PV).

For the case of steady-state flow, it can be shown that the Bernoulli function ([eqn \[7\]](#)) and the PV flux ([eqn \[6\]](#)) are related by the simple expression

$$\mathbf{J} = \nabla \theta \times \nabla B \quad [9]$$

implying that PV generation is closely tied to the production of Bernoulli anomalies. According to [eqn \[9\]](#), at steady state the PV flux is tangent to both isentropic surfaces and surfaces of constant B . Typically, the flux is visualized on isentropic surfaces, in which case the flux vector is tangent to contours of B on the surface.

PV Generation in Wakes

The close relationship between PV fluxes and Bernoulli anomalies in orographic wakes is illustrated in [Figure 8](#). Shown in the figure are schematics of flows on an isentropic surface at steady state. In the case of [Figure 8\(a\)](#), the flow passes directly over the obstacle, with no blocking, and encounters a dissipative zone over the lee slope, presumably due to wave breaking. For [Figure 8\(b\)](#), the flow is blocked upstream and passes laterally around the obstacle, with the isentrope intersecting the terrain, and with a recirculation zone in the lee. In both cases the upstream flow has zero PV and constant B , so that any anomalies downstream are due entirely to dissipation over the obstacle.

In the case with wave breaking ([Figure 8\(a\)](#)), the dissipation in the breaking zone typically causes a reduction in the Bernoulli function for particles passing through the dissipative area. These reduced Bernoulli values are then carried away downstream, leading to a wake of reduced B , as shown in the figure. In terms of PV, the Bernoulli gradient across the breaking zone implies a PV flux from left to right (facing downstream), which according to [eqn \[6\]](#) must be mainly due to the dissipative parts of $\mathbf{J}$ (since the flux is perpendicular to $\mathbf{u}$). At the edges of the wake, the Bernoulli gradients imply PV fluxes downstream, with a positive (downstream) flux on the right and a negative (upstream) flux on the left. Since the flow downstream is largely nondissipative, these fluxes must be primarily due to the $\mathbf{uQ}$ term in [eqn \[6\]](#). Overall, then, the conceptual model is one of PV anomalies being created at either edge of the breaking zone through dissipation and then being advected away downstream, with positive PV on the right and negative on the left. In between the PV anomalies is a wake of decelerated air, with reduced values of Bernoulli function.

[Figure 8\(b\)](#) shows a slightly more subtle case, in which the flow on the upstream face of the obstacle is blocked, causing the θ -surface to intersect the terrain (as in the examples from the previous section). Downstream of the obstacle is a pair of recirculating eddies, with reduced Bernoulli function in the recirculation zones. As in the breaking case, the Bernoulli gradients at the wake edges imply fluxes of PV, with a positive (downstream) flux on the right side and negative (upstream)

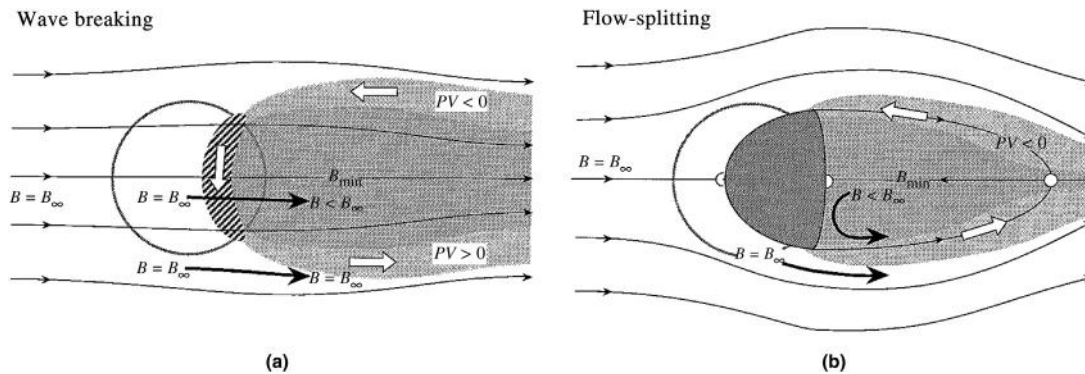


Figure 8 Schematic illustration showing PV generation in stratified flow past an obstacle. Shown are flow fields as visualized on an isentropic surface, with solid contours indicating streamlines. In (a), the isentropic surface completely covers the terrain with no blocking, while (b) shows a case with blocking on the upstream face, causing the isentropic surface to split. The hatched area in (a) is a region of turbulent wave breaking, while dark shading in (b) shows a hole in the isentropic surface, where the surface intersects the obstacle. For both panels, the shaded area downstream is a region of reduced Bernoulli function in the wake. Open arrows show the PV flux $\mathbf{J}$ associated with the Bernoulli gradient on the isentropic surface. From Schär, C., Durran, D.R., 1997. Vortex formation and vortex shedding in continuously stratified flows past isolated topography. *Journal of Atmospheric Science* 54, 534–554.

flux on the left. However, unlike the breaking case, the PV fluxes in the blocked flow are thought to be associated with reconnection, rather than breaking. Specifically, as air in the wake eddies recirculates, it eventually joins with air from outside the wake at the wake edges, thus bringing low-Bernoulli air in contact with air whose Bernoulli function remains constant at the upstream value. The resulting Bernoulli gradient implies a downstream PV transfer, with positive PV on the right and negative on the left.

Note that as presented, the cause of the Bernoulli reduction in **Figure 8(b)** is not immediately apparent. Regardless, the Bernoulli drop must be tied to dissipation in some form: the most likely causes are either dissipation in jump-like or wave-breaking features along the lee slopes (or perhaps the flanks) of the obstacle, or else weak diffusion distributed throughout the entire wake zone.

PV Banners

Orographic PV anomalies were studied as part of the Mesoscale Alpine Programme, a large international field campaign conducted during the fall of 1999. **Figure 9** shows two examples: one for flow past the Alps (**Figure 9(a) and 9(c)**) and one for the Dinaric Alps (**Figure 9(b) and 9(d)**). In both cases, the wake flow was analyzed using a variety of tools, including instrumented aircraft, dropsondes, and mesoscale model results.

As illustrated in **Figure 9(a) and 9(b)**, topographically generated PV anomalies often take the form of long, narrow plumes extending downwind of the terrain. In general, these plumes occur as positive-negative couplets, with each couplet tied to a specific feature in the terrain profile. As illustrated by the figure, a complex mountain range such as the Alps or Dinaric Alps can lead to a number of PV anomalies being generated simultaneously. In the case of the Alps (**Figure 9(a) and 9(c)**), the individual anomalies associated with specific peaks are further embedded within a larger overall PV couplet, associated with the mountain range as a whole. (Indeed,

Figure 9(c) shows just the positive branch of this larger overall couplet.)

Given their visual appearance, PV anomalies such as those shown in **Figure 9** are often referred to as *PV banners*, suggesting an analogy to pennants attached the mountain peaks. As seen along flight tracks (as in **Figure 9(d)**), these PV banners often take the form of oscillations, with crests and troughs of PV where the various PV couplets cross the track. The cross-stream dimensions of the banners appear to vary somewhat: in **Figure 9**, the smaller individual banners have widths of 10–20 km, while the larger Alps-scale banner has a scale closer to 100 km. As seen in **Figure 9(c)**, the PV tends to be limited to levels below the mountain peaks, with the strongest anomalies in the inversions at the top of the boundary layer. That said, it is reasonable to expect (but not yet shown) that similar anomalies might occur at higher altitudes, associated with breaking waves.

Results from model simulations suggest that PV banners often extend far downstream, at least as far as several 100 km. (For the particular case of Hawaii, traces of the wake can be discerned in satellite observations at incredible distances, perhaps as much as 3000 km downstream. The details of this extremely long wake remain uncertain: one possibility is that air–sea interaction plays a role.) In some cases the individual plumes associated with specific peaks appear to coalesce downstream, leading to larger vortices on the scale of the entire mountain range. For the case of the Alps, these larger vortices are then large enough to affect synoptic-scale flows, perhaps leading to enhanced lee-side cyclogenesis (see **Mountain Meteorology: Orographic Effects: Lee Cyclogenesis**).

Shallow-Water Wakes

The vortices described above for stratified flows have close analogs in shallow-water dynamics as well. **Figure 10** shows two examples: one in which the fluid layer passes completely over the obstacle (**Figure 10(a) and 10(b)**), and one in which

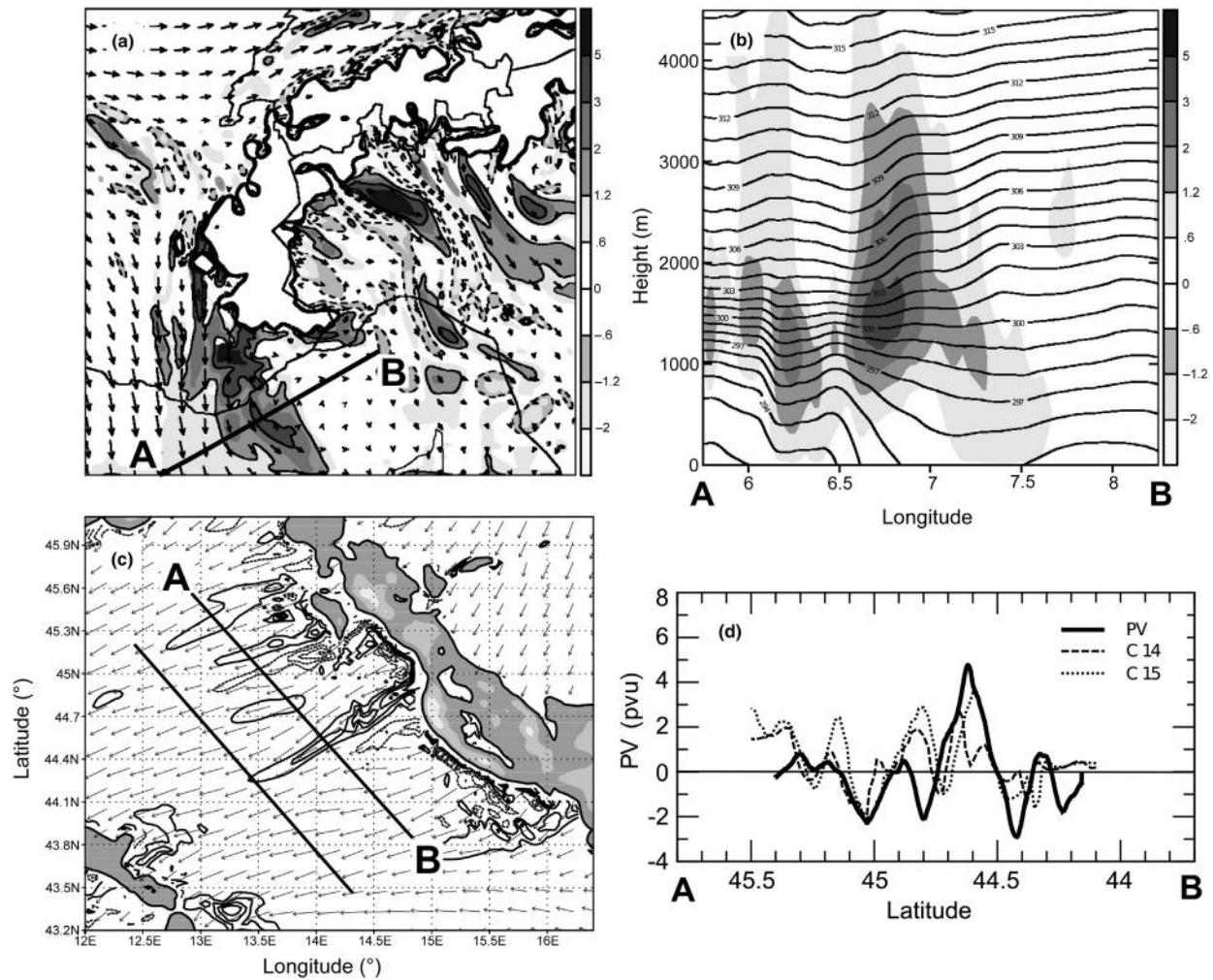


Figure 9 Orographic PV anomalies as observed during the Mesoscale Alpine Programme. (a) Flow past the Alps as computed by the MC2 model with 3-km horizontal resolution. Shown are the wind vectors and PV anomalies (solid contours/dark shading shows positive values, dashed contours/light shading shows negative) as seen on the 850 hPa pressure surface. Thick solid contour shows the intersection of the pressure surface with the terrain. (b) Cross section along the contour indicated in (a) (A to B), showing objectively analyzed fields from both aircraft and dropsonde data. Shown are PV (PVU; shaded with colorbar) and potential temperature (contours). Adapted from Schär, C., Sprenger, M., Lüthi, D., Jiang, Q., Smith, R.B., Benoit, R., 2003. Structure and dynamics of an Alpine potential-vorticity banner. *Quarterly Journal of the Royal Meteorological Society* 129, 825–855. (c) Flow past the Dinaric Alps, as computed by the COAMPS model at 3-km horizontal grid spacing. Shown are PV (solid contours positive, dotted negative) and wind vectors at a height of 600 m above sea level. Shaded areas indicate terrain features higher than 600 m. (d) PV along the contour indicated in (c) (A to B) as computed from aircraft flight level data (thick solid). Also shown are PV anomalies computed by the COAMPS model, at times bracketing the aircraft observations (dashed and dotted). Adapted from Grubišić, V., 2004: Bora driven potential vorticity banners over the Adriatic. *Quarterly Journal of the Royal Meteorological Society*, 130, 2571–2603.

the flow is blocked upstream, causing the layer to split around the obstacle (**Figure 10(c) and 10(d)**).

In the case shown in **Figure 10(a) and 10(b)**, the flow becomes supercritical as it passes over the obstacle peak, with rapidly decreasing layer depth and strongly accelerated flow along the lee slope (not shown). At the base of the obstacle, the depth abruptly recovers to nearly its upstream value in a hydraulic jump (indicated by the bold solid line), much like that seen in **Figure 3(a)**. Downstream of the jump is a pair of recirculating eddies, with plumes of vertical vorticity extending downstream along the wake edges (**Figure 10(b)**).

Hydraulic jumps are inherently dissipative features, and are thus at least broadly similar to the wave-breaking region shown

in **Figure 8(a)**. As seen in **Figure 10(b)**, the flow entering the jump acquires the shallow-water equivalent of PV, which then streams away downstream along the edges of the wake as a pair of PV banners. The jump also produces a loss in Bernoulli function, resulting in a Bernoulli deficit in the wake zone. As in the stratified case, this Bernoulli loss and the PV generation are closely related, with direct analogs to eqns [5], [6], and [9] applying at steady state.

In the case shown in **Figure 10(c) and 10(d)**, the obstacle is sufficiently large that the fluid layer is unable to ascend to the peak, thus splitting on the upstream face and passing around the sides. In the lee of the obstacle is a pair of recirculating eddies. As in **Figure 8(b)**, a pair of PV anomalies emanate from

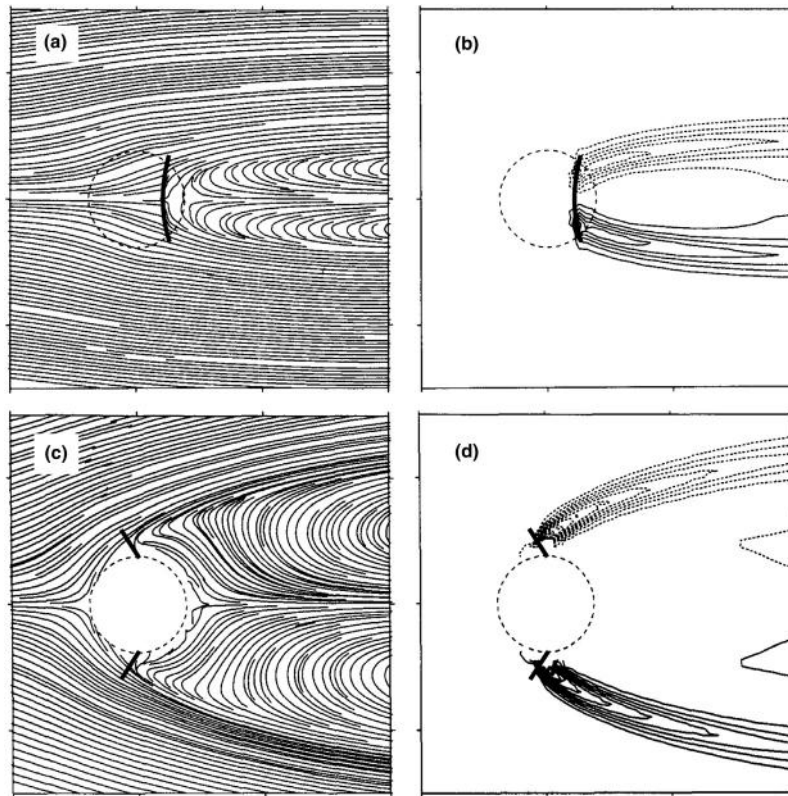


Figure 10 Lee vortices in free-slip shallow-water flow. Shown are (a) streamlines and (b) vorticity for a case in which the fluid layer completely covers the obstacle, with supercritical flow along the lee slope. (c) and (d) as in (a) and (b), but for a case in which the obstacle pierces the fluid surface, causing the flow to split on the upstream face. Heavy solid lines in all panels show the positions of hydraulic jumps. The blank region in (c) shows the part of the obstacle above water level. Modified from Schär, C., Smith, R.B., 1993. Shallow-water flow past isolated topography. Part I: Vorticity production and wake formation. *Journal of Atmospheric Science* 50, 1373–1400.

the flanks of the obstacle, where recirculating wake air joins with air passing laterally around the barrier. The flow also features a pair of weak hydraulic jumps just upstream of the reconnection points (often referred to as *flank shocks*), which are absent in Figure 8. Nonetheless, observations of flow past Hawaii suggest that similar jump-like or wave-breaking features may in fact exist in the real world.

The time shown in Figure 10(c) and 10(d) is a time when the wake has already reached a mature and quasisteady state. However, consideration of the earlier transient evolution reveals strong similarities to the example case shown above (in the Vortex Formation section), including a retracting piston response like that shown in Figure 7, and a recirculation and reconnection phase much like that in Figure 6. The flow even features a hole in the fluid layer at the vortex center, much like the hole in the isentropic surface shown in Figure 6(d).

Vortex Shedding

The previous sections describe wakes that are essentially symmetric, with a pair of attached, counterrotating vortices. However, in many cases, wakes transition to a completely different state, in which vortices are continually shed downstream, with each successive shed vortex having opposite

rotation from the last. The resulting *vortex streets* can lead to striking satellite images, such as that shown in Figure 1(b).

Vortex streets are thought to result from a pair of attached (i.e., nonshedding) vortices that becomes unstable. The instability initially manifests as an oscillation, with each eddy in the vortex pair becoming alternately stronger and weaker. However, as the oscillation grows, the eddies eventually begin to detach and drift downstream, thus transitioning to a vortex shedding state. The instability that drives this transition is effectively a property of the shear flow in the counterrotating eddies, and thus has little dependence on either the mechanism of vortex formation or the details of the flow near the obstacle. As a result, the dynamics of shedding in geophysical wakes is surprisingly similar to that found in homogenous wakes, for which there is an extensive literature.

Figure 11 shows the instability and transition to shedding for an elongated shallow-water wake, with the attached, symmetric state of the wake shown in Figure 11(a) and 11(b). The flow in Figure 11(a) and 11(b) supports two different types of instability modes. The first is a local mode that grows along the shear lines at the wake edges, and which behaves roughly like a classic inflection-point instability. However, the energy of this instability propagates rapidly away downstream, before the disturbance has a chance to grow substantially.

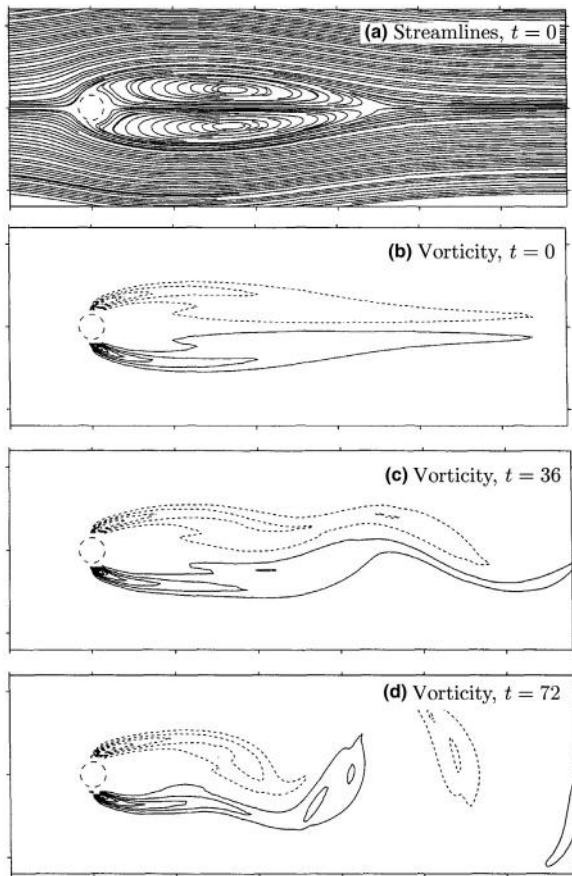


Figure 11 Wake instability and vortex shedding in shallow-water flow. The flow shown in (a) and (b) consists of a perfectly symmetric, counterrotating, stable wake, as computed numerically using an enforced symmetry condition. The symmetry condition is then removed (at time $t = 0$), which allows the growth of nonsymmetric instability modes, as illustrated in (c) and (d). Shown are the (a) streamlines and (b)–(d) vorticity at times (a), (b) $t = 0$, (c) $(gD)^{1/2}t/a = 36$, and (d) $(gD)^{1/2}t/a = 72$, where a is the obstacle half-width and D is the upstream fluid depth. Adapted from Schär, C., Smith, R.B., 1993. Shallow-water flow past isolated topography. Part II: Transition to Vortex Shedding. *Journal of Atmospheric Science* 50, 1401–1412.

The second type of instability is an antisymmetric, global oscillation, which involves coupling of the two shear lines with the reversed flow in between. Modes of this type grow in place, and are thus able to significantly disrupt the wake pattern.

As seen in **Figure 11(c)**, the instability first manifests as a wavelike oscillation near the downstream stagnation point. As the oscillation grows, the flow begins to shed isolated vorticity patches (**Figure 11(d)**), predominately from the downstream edge of the wake. Eventually the instability works its way back to the lee slope, and complete, recirculating vortices of alternating sign are shed from either side of the obstacle, with nearly perfect periodicity (not shown).

Detailed analysis shows that the growing oscillation tends to emanate from the parts of the wake with flow reversal. Wakes without reversed flow are thus expected to be stable. Similarly, if the wake features only a limited region of weak reversed flow, then the dispersion of energy away from this region may be sufficient to suppress the growth of the mode. The transition (or bifurcation) point between weak, stable wakes and stronger, unstable wakes has yet to be explored in much detail for shallow-water flow. However, it seems likely that for the free-slip and inviscid case, at least, virtually any wake with flow reversal will eventually become unstable, as the wake elongates downstream.

Simulations including viscosity and bottom friction show that both processes tend to exert a stabilizing influence, particularly for wakes with only weak flow reversal. The source of the stabilization is primarily a decrease in the length of the wake (thus reducing the extent of the reversed flow), and to a lesser degree the damping of the instability modes themselves. Of the two damping processes, it appears that bottom friction in particular is likely an important stabilizing factor in the real atmosphere. This presumably accounts for the many apparently stable vortices observed in real-world flows, such as the case of Hawaii shown in **Figure 1(a)**.

See also: **Boundary Layer (Atmospheric) and Air Pollution:** Coherent Structures; Complex Terrain; Convective Boundary Layer; Modeling and Parameterization; Observational Techniques *In Situ*; Observational Techniques: Remote; Ocean Mixed Layer; Overview; Stably Stratified Boundary Layer; Surface Layer. **Dynamical Meteorology:** Potential Vorticity; Vorticity. **Gravity Waves:** Buoyancy and Buoyancy Waves: Optical Observations; Buoyancy and Buoyancy Waves: Theory. **Mountain Meteorology:** Downslope Winds; Lee Waves and Mountain Waves; Orographic Effects: Lee Cyclogenesis; Overview.

Further Reading

- Baines, P.G., 1995. *Topographic Effects in Stratified Flows*. Cambridge University Press, 482 pp.
- Chopra, K.P., 1973. Atmospheric and oceanic flow problems introduced by islands. *Advances in Geophysics* 16, 297–421.
- Rotunno, R., Smolarkiewicz, P.K., 1991. Further results on lee vortices in low-Froude-number flow. *Journal of Atmospheric Science* 48, 2204–2211.
- Schär, C., Durran, D.R., 1997. Vortex formation and vortex shedding in continuously stratified flows past isolated topography. *Journal of Atmospheric Science* 54, 534–554.
- Schär, C., Sprenger, M., Lüthi, D., Jiang, Q., Smith, R.B., Benoit, R., 2003. Structure and dynamics of an Alpine potential vorticity banner. *Quarterly Journal of the Royal Meteorological Society* 129, 825–855.
- Smith, R.B., 1989. Hydrostatic airflow over mountains. *Advances in Geophysics* 31, 1–41.
- Smolarkiewicz, P.K., Rotunno, R., 1989. Low Froude number flow past three-dimensional obstacles. Part I: Baroclinically generated lee vortices. *Journal of Atmospheric Science* 46, 1154–1164.

Lee Waves and Mountain Waves

DR Durran, University of Washington, Seattle, WA, USA

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

The basic properties of mountain waves are discussed using linear theory for small-amplitude disturbances. Important nonlinear processes governing waves launched by larger mountains are then considered. The final topic is mountain-wave momentum flux and the interaction of these waves with the larger-scale flow.

Introduction

Buoyancy perturbations develop when stably stratified air ascends a mountain barrier. These perturbations often trigger disturbances that propagate away from the mountain as gravity (or buoyancy) waves. Gravity waves triggered by the flow over a mountain are referred to as 'mountain waves' or 'lee waves.' Mountain waves sometimes reveal their presence through dramatic cloud formations, such as smooth lenticular



Figure 1 Single lenticular cloud over Laguna Verde, Bolivia. This cloud was probably formed by a vertically propagating mountain wave. Copyright Bernhard Mühr, www.wolkenatlas.de.



Figure 2 Multiple lenticular clouds over Mývatn, Iceland, formed by trapped lee waves. Copyright Georg Müller, www.wolkenatlas.de.

clouds (see [Figures 1](#) and [2](#)) and ragged rotor clouds. Large-amplitude mountain waves can generate regions of clear air turbulence that pose a hazard to aviation. Large-amplitude mountain waves may also produce very strong winds that blow down the lee slope of ridgeline topographic barriers (see **Mountain Meteorology: Downslope Winds**).

What happens to mountain waves after they are generated? If the wave amplitude becomes large in comparison to the vertical wavelength, the streamlines in a vertically propagating mountain wave steepen and overturn in a manner roughly analogous to a breaking wave in the ocean. Such 'convective' overturning often occurs as the waves enter the lower stratosphere where they encounter increased static stability and decreasing horizontal wind speeds. The convective overturning of vertically propagating waves is also promoted by the systematic decrease in atmospheric density with height. Those waves that do not breakdown due to convective overturning before reaching the mesosphere are ultimately dissipated by the vertical transfer of infrared radiation between the warm and cool regions within the wave and the surrounding atmosphere (radiative damping).

Horizontal momentum is transported by mountain waves from the regions of wave dissipation to the surface where a net pressure force is exerted on the topography. A decelerative force is exerted on the large-scale atmospheric circulation in those regions where the wave undergoes dissipation.

The basic structure of a mountain wave is determined by the size and shape of the mountain and by the vertical profiles of temperature, wind speed, and moisture in the impinging flow. The overall character of the wave can often be predicted on the basis of linear theory, in which the mountain is assumed to be small in comparison with the vertical wavelength of the mountain wave, and such theory will be the subject of the next section. Nevertheless, nonlinear effects do exert a significant influence on the wave amplitude and are essential to the dynamics of mountain-wave dissipation in regions of wave breaking; such effects will be considered later in this article.

Linear Mountain-Wave Theory

The strongest mountain waves are forced by long quasi-two-dimensional ridges that are sufficiently narrow so that the dynamical influence of the Coriolis force can be neglected. The

basic dynamics of these waves are largely captured by the linear theory for steady two-dimensional Boussinesq flow over an obstacle, for which the linearized momentum, thermodynamic, and continuity equations may be reduced to the following single equation for the vertical velocity w ,

$$\frac{\partial^2 w}{\partial x^2} + \frac{\partial^2 w}{\partial z^2} + \ell^2 w = 0. \quad [1]$$

Here x is the horizontal coordinate perpendicular to the ridge-line; z is the vertical coordinate, and

$$\ell^2 = \frac{N^2}{U^2} - \frac{1}{U} \frac{d^2 U}{dz^2} \quad [2]$$

is the 'Scorer parameter' in which $U(z)$ is the speed of the basic-state flow and $N(z)$ is the Brunt-Väisälä frequency (or alternatively, the buoyancy frequency). In the Boussinesq limit, the Brunt-Väisälä frequency may be defined in terms of the basic-state potential temperature $\bar{\theta}(z)$, a constant reference potential temperature θ_0 , and the gravitational acceleration g , such that $N^2 = (g/\theta_0) d\bar{\theta}/dz$.

Neglecting the effects of surface friction, the velocity perpendicular to the topography must vanish at the surface of the topography $z = h(x)$. This constraint provides a lower boundary condition for eqn [1], and can be approximated as $w(x, 0) = U \partial h / \partial x$ to the same order of accuracy retained in the linearized governing equations. The atmosphere has no distinct upper boundary, so the upper boundary condition is imposed in the limit $z \rightarrow \infty$. In order to ensure the physical relevance of mathematical solutions to eqn [1] in the infinitely deep atmosphere, those solutions must satisfy one of the two possible conditions, that is, either (i) the perturbation energy density must approach zero as $z \rightarrow \infty$, or (ii) if the perturbation energy density is finite as $z \rightarrow \infty$, then the perturbation energy flux associated with each individual's vertically propagating mode must be upward. The second condition allows the representation of disturbances generated within the domain that propagate energy upward to arbitrarily great heights, but it prohibits downward propagating modes from radiating energy into the domain from infinity.

Constant Wind Speed and Stability, Sinusoidal Ridges

As a first example consider flow in a horizontally periodic domain in which $h(x) = h_0 \sin(kx)$. The lower boundary condition becomes $w(x, 0) = U h_0 k \cos kx$, and solutions to eqn [1] subject to this lower boundary condition may be written in the form

$$w(x, z) = \tilde{w}_1(z) \cos kx + \tilde{w}_2(z) \sin kx. \quad [3]$$

Substituting eqn [3] into eqn [1], one obtains

$$\frac{d^2 \tilde{w}_i}{dz^2} + (\ell^2 - k^2) \tilde{w}_i = 0 \quad i = 1, 2. \quad [4]$$

Consider the simplest possible atmospheric structure in which N and U are constant with height. Without loss of generality we will focus on the case in which $U > 0$ and $k > 0$. Since N and U are constant, $\ell^2 = N^2/U^2$ is also constant. Defining $\nu = (\ell^2 - k^2)^{1/2}$ and $\mu^2 = -\nu^2$, the solution to eqn [4] may be written as

$$\tilde{w}_i(z) = \begin{cases} A_i e^{\mu z} + B_i e^{-\mu z} & k > \ell, \\ C_i \cos \nu z + D_i \sin \nu z & k < \ell, \end{cases} \quad [5]$$

where A , B , C , and D are constants to be determined by the upper and lower boundary conditions. Note that the fundamental character of the solution depends on the relative magnitudes of the Scorer parameter and the horizontal wavenumber.

If $\ell < k$, or equivalently, if the intrinsic frequency of the wave Uk is greater than N , solutions to eqn [4] either grow or decay exponentially with height. Only the solution that decays with height is admitted by the upper boundary condition that the perturbation energy density must approach zero as $z \rightarrow \infty$. The vertical velocity satisfying eqn [1] and the upper and lower boundary conditions is

$$w(x, z) = U h_0 k e^{-\mu z} \cos kx. \quad [6]$$

On the other hand, if $\ell > k$, the solutions to eqn [4] are sinusoidal functions of z that neither amplify nor decay as $z \rightarrow \infty$. After imposing the lower boundary condition, the general solution can be expressed as

$$w(x, z) = (U h_0 k - E) \cos(kx + \nu z) + E \cos(kx - \nu z), \quad [7]$$

where the constant E is determined by the upper boundary condition. Writing the solution in the form eqn [7] makes it easy to distinguish between waves that propagate energy upward or downward by examining the relationship between the signs of the vertical and horizontal wavenumbers.

The perturbation energy in a wave propagates at the group velocity. In the constant- N -and- U case, the dispersion relation for the time-dependent generalization of eqn [1] is

$$\omega = Uk \pm \frac{Nk}{(k^2 + \nu^2)^{1/2}}, \quad [8]$$

where ω is the frequency and k and ν are the horizontal and vertical wavenumbers in an arbitrary wave of the form $\Re(e^{i(kx + \nu z - \omega t)})$. Since by assumption $U > 0$, all steady waves (for which $\omega = 0$) are associated with the negative root in eqn [8], and their vertical group velocities are

$$\frac{\partial \omega}{\partial \nu} = \frac{Nk\nu}{(k^2 + \nu^2)^{3/2}}, \quad [9]$$

implying that upward group velocity, and upward energy transport, occurs when k and ν have the same sign. The upper boundary condition therefore requires $E = 0$ in eqn [7], and when $\ell > k$, the solution to eqn [1] becomes

$$w(x, z) = U h_0 k \cos(kx + \nu z). \quad [10]$$

The difference between the case $\ell < k$ and the case $\ell > k$ is illustrated in Figure 3, which shows streamlines over a series of sinusoidal ridges in a steady flow with $N = 0.01 \text{ s}^{-1}$ and $U = 15 \text{ m s}^{-1}$. In the case shown in Figure 3(a), the topographic wavelength is 8 km and $\ell^2 < k^2$ (or equivalently $Uk > N$); the waves decay exponentially with height, and the wave crests are aligned vertically. In the case in Figure 3(b), the topographic wavelength is 40 km and $\ell^2 > k^2$ (or $Uk < N$); the waves propagate vertically without loss of amplitude, and the wave crests tilt upstream with height. The waves decay away from the forcing when the intrinsic frequency exceeds the

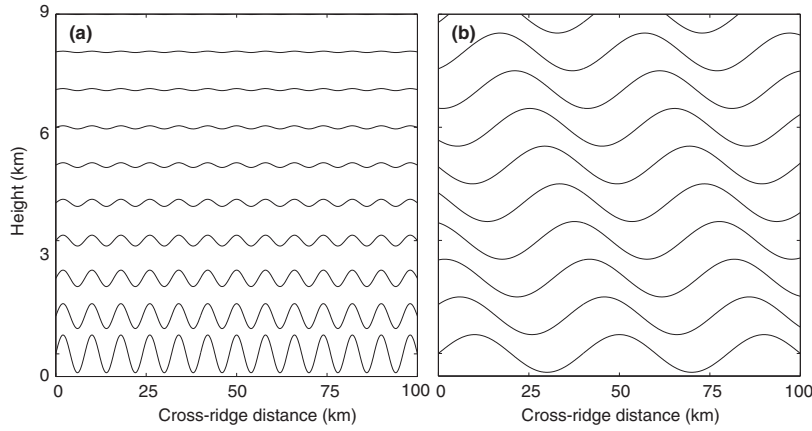


Figure 3 Streamlines in steady airflow over an infinite series of sinusoidal ridges when $N = 0.01 \text{ s}^{-1}$, $U = 15 \text{ m s}^{-1}$, and the wavelength of the topography is (a) 8 km (case $Uk > N$) or (b) 40 km (case $Uk < N$). The flow is from left to right. The lowest streamline coincides with the topography.

Brunt-Väisälä frequency ($Uk > N$) because there is no way for buoyancy restoring forces to support oscillations at such high frequencies (see **Gravity Waves: Buoyancy and Buoyancy Waves: Optical Observations; Buoyancy and Buoyancy Waves: Theory**). On the other hand, when the intrinsic frequency is less than the Brunt-Väisälä frequency, vertical propagation occurs because buoyancy restoring forces can support air-parcel oscillations along a path slanted off the vertical at an angle $\phi = \cos^{-1}(Uk/N)$. In steady mountain waves, ϕ is the angle at which lines of constant phase tilt off the vertical.

Isolated Mountain, Vertical Variations in N or U

The mountain-wave solutions [6] and [10] are only valid for air streams with constant basic-state wind speed and stability flowing across an endless series of sinusoidal ridges. If more realistic terrain profiles and atmospheric structures are considered, other linear solutions can be obtained that more strongly resemble observed mountain waves. In this section, we will describe how the wave response is influenced by isolated topography and vertical variations in atmospheric wind speed and stability.

Suppose that the mountain profile consists of a single ridge from which the terrain elevation drops to some reference level at all distances sufficiently far upstream and downstream. Just as Fourier series can be used to represent a wide variety of periodic functions with an infinite sum of sines and cosines, the isolated mountain can, under rather general conditions, be constructed from periodic functions by the use of Fourier transforms. Let $\hat{w}(k, z)$ denote the Fourier transform of $w(x, z)$ with respect to the x -coordinate, and let $\hat{h}(k)$ be the Fourier transform of the topography $h(x)$.

The k -th component of the Fourier transformed vertical velocity $\hat{w}(k, z)$ must satisfy the Fourier transform of the governing eqn [1],

$$\frac{\partial^2 \hat{w}}{\partial z^2} + (\ell^2 - k^2) \hat{w} = 0, \quad [11]$$

which has the same form as eqn [4]. The lower boundary condition transforms to $\hat{w}(k, 0) = iUk\hat{h}_0\hat{h}$. When N and U are

constant, the solution to eqn [11], subject to the appropriate upper and lower boundary conditions, is

$$\hat{w}(k, z) = ikU\hat{h}(k) \exp\left[i(\ell^2 - k^2)^{1/2}z\right], \quad k > 0. \quad [12]$$

Equation [12] is just the complex analog of eqn [5]; each Fourier component $\hat{w}(k, z)$ of the transformed vertical velocity is identical to the $\hat{w}_i$ forced by an infinite series of sinusoidal ridges having wavenumber k and amplitude $\hat{h}(k)$. The solutions obtained in the preceding section are therefore also applicable to the case of isolated topography. The only complication arises from the requirement that after the $\hat{w}(k, z)$ are determined, the total vertical velocity $w(x, z)$ must be obtained by computing an inverse Fourier transform. The relative weight attached to each individual wavenumber in the composite solution is determined by the Fourier transform of the mountain.

Streamlines for steady linear flow over an isolated ridge of the form

$$h(x) = \frac{h_0 a^2}{x^2 + a^2} \quad [13]$$

are shown in Figure 4(a) for the case $N = 0.01047 \text{ s}^{-1}$, $U = 10 \text{ m s}^{-1}$, and $Nh_0/U = 0.6$.

In this case, $Na/U \approx 10$ and the dominant horizontal wavenumbers in the Fourier transform of the topography satisfy $k^2 \ll \ell^2$, which eliminates the dependence of the vertical structure on the horizontal wavenumber in eqn [12]. As a result, all modes associated with these dominant wavenumbers have approximately the same vertical wavelength ($2\pi U/N = 6 \text{ km}$), so the streamline at 6 km approximately reproduces the mountain profile while those at 3 and 9 km are roughly the mirror-image of the topography. The solution shown in Figure 4(a) is computed numerically without making the hydrostatic assumption and is very similar to that which would be obtained in the hydrostatic limit, in which all horizontal wavenumbers have exactly the same vertical wavelength and the mountain profile is exactly reproduced by the streamline originating at the 6 km level upstream.

As suggested by Figure 4(a), when an infinitely long ridge is sufficiently wide that the flow is approximately hydrostatic

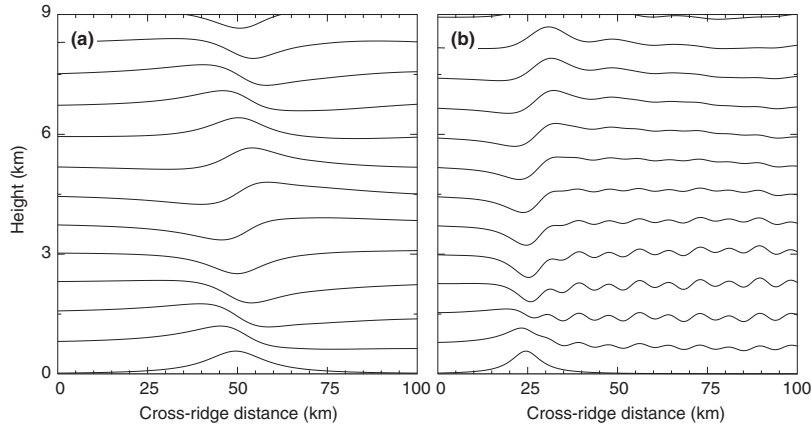


Figure 4 Streamlines in steady airflow over an isolated mountain as predicted by linear theory when (a) $a = 10$ km, N is constant, and $Nh_0/U = 0.6$; (b) $a = 5$ km, N is constant throughout each of two layers such that between the surface and 3 km $N_l h_0/U = 0.6$, and above 3 km $N_u h_0/U = 0.24$.

($Na/U \gg 1$) but still narrow enough that Coriolis forces can be neglected ($|f|a/U \ll 1$, where f is the Coriolis parameter), energetic mountain waves are found only in the region directly above the mountain. In the nonhydrostatic case, some waves do appear in the region downstream from the ridge, as can be deduced from the horizontal group velocity $\partial\omega/\partial k$, which using eqn [8] and again assuming $U > 0$, may be expressed as

$$\frac{\partial\omega}{\partial k} = \frac{\omega}{k} + \frac{Nk^2}{(k^2 + \nu^2)^{3/2}}. \quad [14]$$

The first term in eqn [14] is the phase speed, which is zero for a steady mountain wave. The second term is nonnegative, implying downstream energy propagation – except in the hydrostatic limit when the second term vanishes because $k^2/\nu^2 \rightarrow 0$. A sufficient decrease in the width of the mountain, relative to that shown in Figure 4(a), will therefore lead to the generation of nonhydrostatic waves that populate a wedge-shaped region emanating upward and downstream from the mountain. The wave energy for each component of the total solution propagates along a line whose slope is equal to the ratio of the vertical group velocity to the horizontal group velocity for that component. Quasi-uniform low-level wave trains, such as those shown in Figure 4(b), do not however, occur unless there are significant vertical variations in the wind speed and static stability.

If the vertical variations in U and N are such that the Scorer parameter decreases significantly with height, cross-ridge flow may generate a qualitatively different type of wave, the ‘trapped lee wave.’ A series of trapped lee waves (also known as resonant lee waves) are apparent extending downstream from the ridge throughout the layer $0 \leq z \leq 4$ km in Figure 4(b); a vertically propagating wave is also visible directly above the mountain. The streamlines shown in Figure 4(b) are for the linear solution to the same problem considered in Figure 4(a), except that $a = 5$ km and the static stability above 3 km is reduced by a factor of 0.4. (The Brunt-Väisälä frequencies in the upper and lower layers are thus $N_u = 0.004188$ and $N_l = 0.01047$ s⁻¹, respectively.)

A necessary condition for the existence of trapped waves in the two-layer problem is that

$$\ell_L^2 - \ell_U^2 > \frac{\pi^2}{4H^2}, \quad [15]$$

where ℓ_U and ℓ_L are the Scorer parameters in the upper and lower layers, and H is the depth of the lower layer. Equation [15] states that the difference in wave propagation characteristics in the two layers must exceed a certain threshold before the waves can be trapped. The horizontal wavenumber of any resonant lee wave in the two-layer system satisfies $\ell_L > k > \ell_U$, implying that the wave propagates vertically in the lower layer and decays exponentially with height in the upper layer. As shown in Figure 4(b), trapped waves have no tilt, even though they can propagate vertically in the lower layer. The reason for this is that wave energy is repeatedly reflected, without loss of amplitude, from the upper layer and the flat ground downstream from the mountain. As a result, the downstream disturbance is the superposition of equal-amplitude upward and downward propagating waves, a combination which has no tilt.

Nonlinear Mountain Waves

Now suppose that the mountain height is not small compared to the vertical wavelength of the mountain wave. If N and U are constant, the streamline displacement $\delta(x, z)$ in steady two-dimensional Boussinesq flow over such a ridge is still governed by a relatively simple mathematical model known as Long’s equation

$$\frac{\partial^2 \delta}{\partial x^2} + \frac{\partial^2 \delta}{\partial z^2} + \frac{N^2}{U^2} \delta = 0, \quad [16]$$

Although Long’s equation is a linear partial differential equation, it may be derived from the fully nonlinear equations without making any linearization or small-amplitude assumptions. Nevertheless, eqn [16] may also be derived by assuming the mountain is infinitesimally high and linearizing the

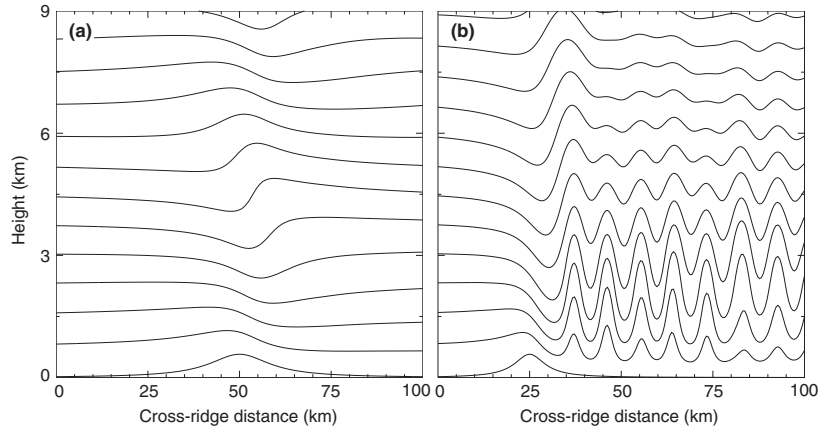


Figure 5 As in [Figure 2](#) except that the streamlines are for a fully nonlinear flow as computed using a numerical model. The trapped waves in panel b are not completely steady; the solution is shown a nondimensional time $Ut/a = 20$ after starting the flow from rest.

governing equations in the usual manner. When N and U are constant, the only difference between the linear and nonlinear solutions arises from the lower boundary condition, which requires $\delta[x, h(x)] = h(x)$ in the exact finite-amplitude case and is approximated by $\delta(x, 0) = h(x)$ in the small-amplitude limit.

As one might guess from the similarities in the governing equations, when N and U are constant the influence of nonlinear dynamics on the wave structure is often relatively minor. This similarity can be appreciated by comparing the linear solution in [Figure 4\(a\)](#) with the corresponding nonlinear solution in [Figure 5\(a\)](#), both of which show streamlines in a Boussinesq flow for which $Nh_0/U = 0.6$.

Nonlinear processes steepen the streamlines around $z = 4.5$ km, which is $3/4$ of a vertical wavelength ($3\lambda_z/4$) above the topography. Conversely, the nonlinear waves are less steep than their linear counterparts near $z = 1.5$ km, which is $\lambda_z/4$ above the mean height of the topography. Despite these modest differences in the shape of the streamlines in the linear and nonlinear waves, the wave amplitude is almost identical in both cases. Nonlinear processes do not have a dramatic impact on the waves forced by flow over an infinitely long ridge unless either (i) there are vertical variations in N and U or (ii) the mountain is high enough to force wave overturning.

The influence of nonlinear wave dynamics on the flow in the two-layer atmosphere previously considered in connection with [Figure 4\(b\)](#) is shown in [Figure 5\(b\)](#). The amplitude of the lee waves in the nonlinear solution is much larger than that in the linear solution, and in the nonlinear case significant variations are visible among the individual troughs and crests in the region $65 \leq x \leq 100$ km. As suggested by this example, and demonstrated in several observational campaigns and numerical studies, linear theory does reliably predict the amplitude of trapped lee waves generated by finite-amplitude mountains. The main shortcoming of linear theory is that it cannot capture the tendency of the nonlinear dynamics to enhance the short-wavelength Fourier components in the low-level wave field over the lee slope. The nonlinear enhancement of these short-wavelength perturbations in the first wave above the mountain often produces more forcing at the wavelength of the resonant lee waves than does the direct forcing by the topographic profile itself.

Clouds that form in regions of net upward displacement in vertically propagating hydrostatic waves may appear like the cloud in [Figure 1](#). The large single region of cloudiness parallel to the mountain crest is probably formed by air-parcel displacements qualitatively similar to those in the streamline originating near the 6-km-level in [Figure 5\(a\)](#).

Clouds that form in trapped lee waves may appear as a series of long bands parallel to the generating ridge. Such bands are often visible in satellite photos and are formed by streamline patterns qualitatively similar to those originating in the layer between 2 and 4 km in [Figure 5\(b\)](#). Nevertheless, three dimensional variations in the upstream topography often break these bands into the superposition of many lens-shaped cloud masses, such those shown in [Figure 2](#).

Returning to the discussion of how nonlinear dynamics modify the structure of mountain waves, consider the influence of wave breaking on the flow. Two examples in which the wave amplitude becomes large enough to overturn are shown in [Figure 6](#).

The case shown in [Figure 6\(a\)](#) is one with constant N and U identical to that in [Figure 5\(a\)](#), except that the mountain height is increased so that $Nh_0/U = 1.2$. (The vertical scale also extends to $z = 15$ km.) Wave overturning first begins at the $3\lambda_z/4$ level, which is the same level at which the wave faces appear to be steepened in [Figure 5\(a\)](#). As the wave begins to overturn, a $\lambda_z/2$ deep region of well-mixed stagnant fluid develops over the lee slope and begins to extend downstream. A second region of wave overturning eventually develops at a height of $7\lambda_z/4$, although the perturbations are weaker at this level due to the dissipation experienced by the wave as it propagates through the first wave-breaking level. [Figure 6](#) shows the solution at a nondimensional time (Ut/a) of 30, by which time the near-mountain solution is quasi-steady, but the layers of well-mixed fluid in the wave-breaking region continue to expand further downstream. Also shown are contours of the subgrid-scale eddy diffusivity. Regions in which the subgrid-scale diffusivity is large are regions in which the numerical model has diagnosed the presence of vigorous small-scale turbulence such as that which occurs due to wave breaking.

Although the breaking of mountain waves in an atmosphere with constant N and U has received a great deal of theoretical attention, the morphology of such flows is not representative of

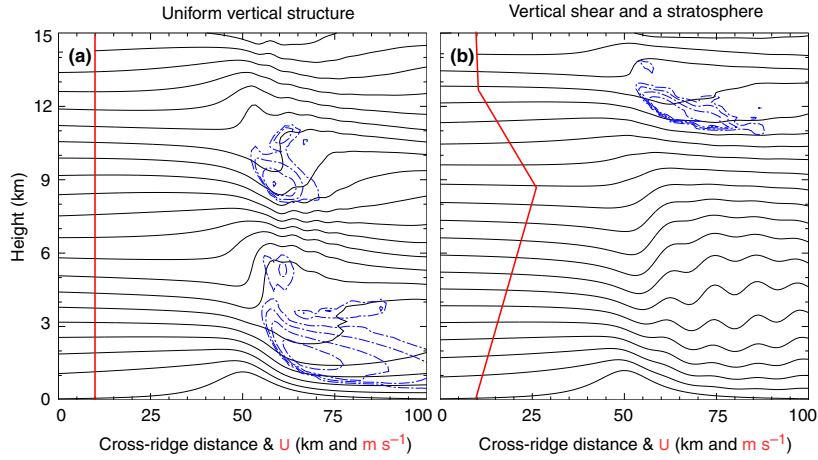


Figure 6 Streamlines (black), contours of the subgrid-scale eddy diffusivity (blue dot-dashed, at intervals of $20 \text{ m}^2 \text{ s}^{-1}$) and the vertical profile of the large-scale wind (red) for (a) the case shown in Figure 3(a) except that $Nh_0/U = 1.2$ and the vertical scale extends to 15 km; (b) as in (a) except with westerly wind shear throughout the tropopause and a realistic stratosphere (see text).

most real-world wave-breaking events, in which the wave structure is significantly modified by vertical wind shear in the upstream flow. Those ridges that run north–south in the middle latitudes are oriented perpendicular to the climatological westerly flow and are frequent generators of large-amplitude mountain waves. A prototypical example of the mountain waves generated by such ridges in a deep westerly flow is shown in Figure 6(b). The mountain profile, the surface wind speed (10 m s^{-1}), and the low-level stability (0.01047 s^{-1}) are identical to those for the case in Figure 6(a), but the upstream wind speed U increases linearly to 25 m s^{-1} at a height of 9 km. The presence of a stratosphere is modeled by increasing N to 0.02 s^{-1} above 9 km and linearly decreasing U back to 10 m s^{-1} at $z = 13 \text{ km}$. The wind speed is a constant 10 m s^{-1} above 13 km. The increase in the cross-mountain wind with height throughout the troposphere decreases the local value of the nonlinearity parameter $\varepsilon(z) = N(z)h_0/U(z)$ to a minimum just below the tropopause at $z = 9 \text{ km}$. Above the tropopause ε increases rapidly with height due to the factor-of-two increase in N and the reversal of the wind shear. As evident in Figure 6(b), these more realistic vertical variations in the upstream flow are sufficient to shift the primary region of wave breaking to the lower stratosphere (around $z = 12 \text{ km}$) and to prevent wave breaking in the troposphere.

The influence of wave breaking is highly nonlocal. In the case with constant N and U , the entire lee-side flow in the wave-breaking regime (Figure 6(a)) is dramatically different from that in the nonbreaking regime (Figure 5(a)). In particular, the surface winds above the lee slope are significantly enhanced in the wave-breaking regime (see Mountain Meteorology: Downslope Winds). The breaking waves in Figure 6(b) also exert a nontrivial influence on the low-level flow, although this influence is considerably less dramatic than that which develops as a consequence of wave breaking in Figure 6(a).

Vertical Momentum Transport

When air flowing over a mountain generates vertically propagating waves, a region of high pressure develops upstream of

the ridge crest and a region of low pressure appears in the lee. The distribution of these pressure perturbations is revealed by the along-flow variation in the spacing between the two lowest streamlines in Figures 3(b), 4(a), 5 and 6. The asymmetry in the pressure distribution across the ridge gives rise to a net pressure force on the topography that tends to accelerate the topography in the direction of the mean flow. An equal and opposite force is exerted on the mean flow by the topography.

To see how the topographically induced decelerative forcing is distributed throughout the fluid, consider the horizontal momentum eqn [17] in which $\mathbf{v}$ is the total velocity vector, p is the pressure, ρ is the density, $\mathbf{i}$ is the unit-vector along the x -coordinate, and $u = \mathbf{v} \cdot \mathbf{i}$,

$$\frac{\partial \rho u}{\partial t} + \nabla \cdot (\rho u \mathbf{v} + p \mathbf{i}) = 0. \quad [17]$$

Integrate the preceding throughout the volume between the surface $h(x)$ and an arbitrary level z_t ; use the divergence theorem; note that there is no advective momentum flux through the lower boundary, and assume that the domain is periodic in the horizontal, then

$$\frac{\partial}{\partial t} \iiint \rho u \, dV = - \iint \rho u w \, dx dy \Big|_{z=z_t} - \iint p \frac{\partial h}{\partial x} \, dx dy \Big|_{z=h}. \quad [18]$$

When vertically propagating mountain waves are present, the cross-mountain pressure drag (given by the last term in eqn [18]) must decelerate the volume-averaged flow in the layer between the surface and z_t unless that drag is balanced by a downward transfer of momentum through level z_t . This same result can be obtained for flow in nonperiodic domains under the assumption that the perturbation quantities vanish at the lateral boundaries, although caution is advised when trying to apply eqn [18] in a nonperiodic domain because non-negligible mountain-wave-induced perturbations may extend far upstream and downstream from a very long ridge.

The interaction between the mean flow and the mountain-wave induced momentum fluxes can be described more

precisely by separating the dynamical variables into an average over the domain (denoted by an overbar and taken as the representative of the synoptic-scale flow impinging on the mountain) and the perturbation about that average (denoted by a prime and assumed to represent the contributions from mountain waves generated by the flow over the ridge). If the horizontal momentum equation for two-dimensional inviscid Boussinesq flow,

$$\frac{\partial \rho_0 u}{\partial t} + \frac{\partial}{\partial x}(\rho_0 u^2 + p) + \frac{\partial}{\partial z}(\rho_0 u w) = 0, \quad [19]$$

is averaged over a periodic domain (or if it is assumed that the perturbations vanish at the lateral boundaries of a nonperiodic domain) and if $\bar{w} = 0$, one obtains

$$\frac{\partial \rho_0 \bar{u}}{\partial t} = -\frac{\partial}{\partial z}(\rho_0 \overline{u'w'}). \quad [20]$$

A decelerative forcing will therefore be exerted on the flow in those regions in which the mountain-wave-induced momentum flux is divergent, that is, where $\partial(\rho_0 \overline{u'w'})/\partial z > 0$.

The vertical profile of the momentum flux is particularly easy to describe for steady, inviscid, small-amplitude waves in a periodic domain (or in an unbounded domain in which the waves decay as $x \rightarrow \pm\infty$). The cross-mountain pressure drag in such waves is identical to the vertical momentum flux at $z = 0$, as may be seen from the steady-state version of eqn [18] in the limit $z_t \rightarrow 0$. Furthermore, a classic theorem due to Eliassen and Palm states that under the preceding assumptions $\rho_0 \overline{u'w'}$ is constant with height except at a 'critical level' at which $\bar{u} = 0$. Mountain waves are dissipated at the mean-state critical layers found in real atmospheric flows. Mountain waves are also dissipated through breaking and overturning if they attain sufficiently large amplitude due to the decrease in density with height or, as in Figure 6, if they propagate into a region in which the local value of N/U increases significantly. Small-amplitude mountain waves that propagate all the way to the mesosphere without experiencing overturning are damped by radiative heat transfer.

The Eliassen and Palm theorem implies that small-amplitude mountain waves transport a fraction of the momentum of the cross-mountain flow downward to the surface from those elevations at which the waves undergo dissipation. There will be no vertical momentum flux divergence and no forcing of the mean flow within the those layers of the atmosphere in which the waves are steady and nondissipative. The momentum fluxed downward by the waves is transferred to the topography by the cross-mountain pressure drag. Similar distributions of the vertical momentum flux are obtained even when the waves are nonlinear. For example, the vertical momentum flux profile associated with the finite-amplitude waves shown in Figure 6(b) is approximately nondivergent between the ground and the region of wave breaking in the layer $11 \leq z \leq 13$ km. In contrast, the momentum flux profile is strongly divergent in the wave breaking region, and the mean flow is subject to a significant decelerative forcing throughout this layer (see **Dynamical Meteorology: Wave Mean-Flow Interaction**). Unlike surface friction, the drag associated with mountain waves is typically exerted on the flow well above the lower boundary. Numerical experiments with general circulation models suggest that mountain-wave-induced drag plays a nontrivial role in the total momentum budget of the atmosphere.

Nonsteady Waves

The assumption that mountain waves are in steady state greatly simplifies their theoretical analysis and leads to predictions that are often in decent agreement with observations. Nevertheless, clear evidence of nonstationary behavior has also been documented, particularly in the case of trapped lee waves. Trapped waves are resonant oscillations whose amplitude and wavelength are quite sensitive to changes in the structure of the flow impinging on the mountain. Thus, rather modest changes in the large-scale wind speed and temperature profiles can produce easily observed variations in the lee wave train. Even when the large-scale synoptic forcing is essentially constant, lee-wave transience can be produced by either nonlinear wave-wave interactions or by the diurnal heating or cooling of the planetary boundary layer. Solar heating, for example, reduces the buoyancy frequency $N(z)$ in the lower atmosphere in a manner that typically tends to increase the wavelengths of trapped waves.

Although it can be more difficult to observe, changes in the large-scale flow also influence vertically propagating waves. One property of vertically propagating waves that is particularly sensitive to variations in the cross-mountain flow is the vertical profile of momentum flux. An example of this sensitivity is shown in Figure 7, which is a plot of $\rho_0 \overline{u'w'}$ as a function of time and height from a numerical simulation in which a localized jet crosses over an isolated ridge.

The large-scale winds are constant with height, and over the ridge crest they increase sinusoidally from zero to 20 m s^{-1} and then fall back to zero over a 50-h period. Throughout the domain $N = 0.01 \text{ s}^{-1}$, so at specific times during the simulation, the flow in a vertical plane perpendicular to the ridge-line

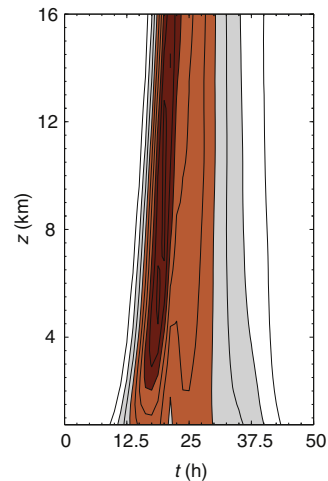


Figure 7 Horizontal-domain-averaged momentum flux $\rho_0 \overline{u'w'}$ generated by time-varying flow over a 750-m high ridge plotted as a function of time and height. Values are normalized by the momentum flux for the linear steady-state solution for waves driven by a 20 m s^{-1} flow across the same ridge. Solid lines show contours at intervals of 0.25, with values in the range (0.5,1.0), (1.0,1.75), and greater than 1.75 shaded gray, orange, and dark orange-red, respectively. Adapted with permission from Chen, C.C., Durran, D.R., Hakim, G.J., 2005. Mountain wave momentum flux in an evolving synoptic-scale flow. *Journal of Atmospheric Sciences* 62, 3213–3231.

is roughly similar to that shown in Figures 5(a) and 6(a), except that the ridge in the simulation used to produce Figure 7 is 750 m high. The actual momentum flux is normalized by the flux produced by linear waves over a mountain of the same height and shape in a steady uniform 20 m s^{-1} flow (with the same value of $N = 0.01 \text{ s}^{-1}$). As a point of reference, if nonlinear processes were negligible, and if the 50-h period over which the flow varies were sufficiently long that the waves could be considered steady, the Eliassen-Palm theorem would apply, in which case every contour drawn in Figure 7 would be a straight vertical line and the contours would be symmetric about hour 25 because the momentum flux would be linearly proportional to the wind speed. Because of the normalization, the contour at hour 25 would have a value of unity, and there would be no orange region on the plot.

Yet as apparent in Figure 7, the momentum fluxes are not symmetric about hour 25, but are much stronger when the large-scale flow is accelerating ($t < 25 \text{ h}$) than when it is decelerating ($t > 25 \text{ h}$). For example, the momentum flux is three times stronger at hour 19 than at hour 31, even though the large-scale wind above the ridge at both times is identical. Moreover, the flux aloft at hour 19 is more than twice as strong as the flux that would be predicted to occur at the time of strongest flow using a steady-state analysis. This momentum-flux enhancement may be explained by noting that each packet of wave energy launched while the mountain wave is present moves upward at the group velocity. In this example, the vertical group velocity of each packet is proportional to the large-scale cross-mountain wind speed at the time the packet was launched. As a consequence, packets launched later in the acceleration phase tend to overtake those launched earlier and accumulate above the mountain, creating higher momentum fluxes while the cross-mountain flow is accelerating.

The drag generated by the vertical divergence of mountain-wave generated momentum fluxes is currently parameterized in global atmospheric models for both weather forecasting and climate simulation. The formulation of accurate 'gravity-wave-drag' parameterizations is greatly complicated both by the nonlinearity of finite-amplitude mountain waves in (compare Figures 4(b) and 5(b)) and by the dependence of the flux on the past history of the large-scale flow.

See also: Dynamical Meteorology: Wave Mean-Flow Interaction. **Gravity Waves:** Buoyancy and Buoyancy Waves: Optical Observations; Buoyancy and Buoyancy Waves: Theory. **Mountain Meteorology:** Downslope Winds; Lee Vortices.

Further Reading

- Baines, P.G., 1995. *Topographic Effects in Stratified Flows*. Cambridge University Press, Cambridge.
- Chen, Chih-Chieh, Durran, Dale R., Hakim, Gregory J., 2005. Mountain wave momentum flux in an evolving synoptic-scale flow. *Journal of Atmospheric Sciences* 62, 3213–3231.
- Durran, D.R., 1986. Mountain waves. In: Ray, Peter S. (Ed.), *Mesoscale Meteorology and Forecasting*. American Meteorological Society, Boston, pp. 472–492.
- Eliassen, A., Palm, E., 1960. On the Transfer of Energy in Stationary Mountain Waves, vol. 22. *Geof. Publikasjoner*, pp. 1–23.
- Gill, Adrian E., 1982. *Atmosphere–Ocean Dynamics*. Academic Press, Orlando, p. 662.
- Holton, James R., 1992. *An Introduction to Dynamic Meteorology*, third ed. Academic Press, San Diego. p. 507.
- Smith, R.B., 1979. The influence of the mountains on the atmosphere. In: Saltzman, B. (Ed.), *Advances in Geophysics*, vol. 21. Academic Press, pp. 87–230.

Orographic Effects: Lee Cyclogenesis

C Schär, Atmospheric and Climatic Science ETH, Zürich, Switzerland

© 2003 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 4, pp 1602–1614, © 2003, Elsevier Ltd.

Introduction

Since the beginning of systematic studies on cyclone tracks, it has been noted that many mountain ranges act as preferred locations for cyclogenesis. An illustration of this is provided in Figure 1 in terms of cyclogenetic areas and associated storm tracks. The chart is based on hand-counted frequencies of closed isobars on sea-level pressure charts. Besides the classical Atlantic and Pacific storm tracks that emanate to the east of the continents, a range of smaller-scale cyclogenetic areas can be identified in close vicinity to topographic obstacles. Particularly worth noting is the elongated region to the lee of the Rocky Mountains, the Gulf of Genoa to the south of the Alps, and an area near the southern tip of Greenland. In the literature, many additional regions of orographic cyclogenesis are documented. In the Northern Hemisphere this includes the Atlas mountain range in Algeria, the central Japanese mountains, the Tibetan plateau, and the Altai mountains. In the Southern Hemisphere, orographic cyclogenesis is less abundant but has, for instance, been reported for the Andes and the Transantarctic mountain range.

The mountains under consideration are thus characterized by a wide range of horizontal scales, and include both 'synoptic-scale mountains' such as Greenland and the Rocky mountains, as well as 'meso-scale ranges' such as the European Alps, the Pyrenees, and the Atlas, to mention just a few. Here we will focus attention upon orographic cyclogenesis to the lee of the Rocky mountains and the Alps, and thereby cover examples from both these categories.

Synoptic and Climatological Description

Alpine Lee Cyclogenesis

The Alps are an arc-shaped mountain range ~800 km long that has a mean halfwidth of $L \approx 100$ km and an average ridge height of $H \approx 3$ km. Additional distinctive features of the range are major valleys that run predominantly north or south onto the foreland and connect at several Alpine passes, typically with altitudes between 1800 and 2500 m. The highest peaks of the Alps reach up to 4800 m.

Alpine lee cyclogenesis has been the subject of intense observational, theoretical, and numerical studies during the last decades. Numerous case studies, partly based on the Alpine Experiment (ALPEX) conducted in 1982, provide a comparatively coherent picture of the main features of an Alpine event. The pertinent synoptic setting of an Alpine lee cyclogenesis event is the approach of a cyclone toward central or northern Europe and the passage of the associated cold front with an accompanying upper-level trough toward the Alpine ridge (see Figure 2). The approach of this system initiates a highly transient phase of Alpine weather, with a duration of normally 2–3 days. Ahead of the approaching cold front, there is usually

southwesterly flow toward the Alps. The respective airstream consists of warm and moist air, is comparatively weakly stratified, and is associated with high wind velocities. It is able to rise over the Alpine ridge, leading to South Föhn to the north (Figure 2(a)). The cold air behind the cold front, however, is more stably stratified and often capped by a pronounced inversion. It is usually unable to climb the Alps but is rather deflected laterally. The orographic interception of the front is accompanied on the windward side by retardation, low-level blocking, and flow splitting; by the steepening of the baroclinic zone; by the build-up of an impressive pressure gradient across the Alpine barrier; and by the generation of the Mistral flow to the west (Figure 2(b)) and Bora to the east of the Alps (Figure 2(c)).

To the lee the cyclogenesis evolves in two phases: a first phase comprising the *in situ* rapid development of a surface mesoscale low, followed by a less rapid phase involving an increase in both the horizontal and vertical scales of the system together with its possible translation away from the immediate lee region. The second phase is characterized by the arrival in the lee of an upper-level trough. In the absence of a suitable phase relationship between the low-level incipient cyclone and the upper-level trough, the shallow firstphase vortex is unable to grow vertically and decays quickly. Deep cyclones, on the other hand, detach from the Alps and take on the characteristics of typical extratropical depressions.

On average, about 30 Alpine lee cyclones are observed per year, with maximum frequency during the spring and autumn seasons. The climatological distribution of lee cyclogenesis suggests a predominant formation near the south-western tip of the Alps over the Gulf of Genoa. This geographic concentration betokens an orographic influence. Once a deep lee cyclone has formed, it becomes an important governor of the regional weather and climate. Lee cyclones often attain maximum strength when located over northern Italy, and the associated strong southerly flow ahead of the cyclone can induce storm surges in the Adriatic Sea (which occasionally threaten Venice) and advect moist Mediterranean air toward the Alps (which substantially contributes to the annual precipitation totals in northern Italy and the eastern Alps). Following their formation, lee cyclones propagate toward the east, following one of two major tracks. The first leads across the eastern Alps into eastern Europe, while the second follows the northern border of the Mediterranean Sea. As a result, Alpine lee cyclones are crucial to the precipitation climate in the eastern Mediterranean region.

Rocky Mountain Lee Cyclogenesis

The Rocky Mountains are oriented approximately north–south from Alaska to central Mexico, and are characterized by a mean ridge height of ~3 km. Their length and width correspond to

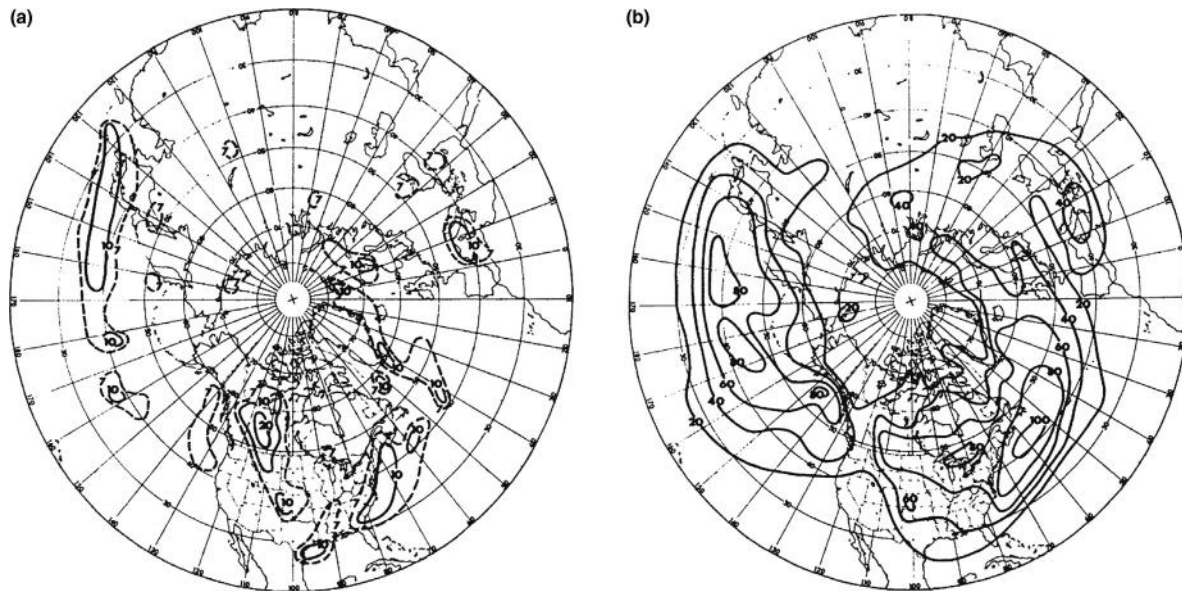


Figure 1 Cyclogenetic areas and associated storm tracks in the Northern Hemisphere for January, based on hand-counted frequency of closed isobars on conventional surface pressure charts (with 4 hPa contour interval). The two panels relate to (a) cyclogenetic regions (new closed low maintained for at least 24 h), and (b) the frequency of cyclones. The numbers relate to the counts on a $5^\circ \times 5^\circ$ grid during the 20 Januaries from 1958 to 1977. Reproduced from Whittaker, L.M., Horn, L.H., 1984. Northern hemisphere extratropical cyclone activity for four mid-season months. *International Journal of Climatology* 4, 297–310.

approximately 4000 km and 800 km in the north–south and west–east directions, respectively, and these scales exceed those of the Alps by almost an order of magnitude.

Rocky Mountain lee cyclogenesis usually starts with an anticyclone residing over the mountain, and an incident low-pressure system over the Pacific (see [Figure 3\(a\)](#)). As the Pacific low approaches the continent, it moves poleward ([Figure 3\(b\)](#)) and finally disappears ([Figure 3\(c\)](#)), while the initial high-pressure system moves south-eastward. Slightly prior to the arrival of the upper-level trough, a new surface low forms, typically 1000 km to the south of the last windward position of the Pacific low. After the upper-level trough has propagated over the Rocky Mountains, a deep development may follow. Usually the lee cyclone moves somewhat farther south before it propagates eastward and finally joins the Atlantic storm track. Strictly speaking, these events do not necessarily invoke a ‘new’ cyclone, but may be interpreted as an anomalous southward propagation of the low when crossing the Rockies (whereby the cyclone is masked by the mountain anticyclone), followed by regeneration in phase with the arrival of the upper-level trough.

Case studies have identified a series of associated mesoscale processes. When the parent cold front impinges upon the Rockies, observations demonstrate that there is frontal retardation, associated frontogenesis, and often the formation of an impressive tropopause fold that implies a descent of stratospheric air toward the cyclonic region. The relevance of Rocky Mountain cyclogenesis for the mid-west is evident from the frequent explosive developments, which often have important repercussions in terms of extreme weather events. Distinct frequency maxima are found, both in winter and summer, to the east of the highest parts of the Rockies in Alberta and Colorado.

Dynamical Mechanisms

There is a great variety of orographic depressions, ranging from small mesoscale vortices that may involve latent heating and convection, to deep lee cyclones that during their life cycle approach the typical structure of extratropical cyclones. Here attention is restricted to the latter category. The development of these deep lee cyclones involves – at least in the second phase – the typical ingredients of regular extratropical cyclogenesis, such as a vertical coupling between low-level thermal and upper-level potential vorticity (PV) anomalies (see **Synoptic Meteorology: Cyclogenesis**), sometimes affected by concomitant intensification by diabatic effects. Often lee cyclones form in an environment that would seem to support cyclogenesis even in the absence of topography. Numerical experiments with and without orography suggest that the role of the topography is often to trigger and modulate cyclogenesis, which would otherwise occur in the near vicinity and/or somewhat later in time. Despite the close relationship to regular mid-latitude cyclogenesis, a great variety of orographic mechanisms have been proposed, which may act in concert with classical cyclogenetic processes. Some of the key dynamical mechanisms are reviewed below.

Regime Diagram for Steady Flow Past Topography

While lee cyclogenesis is always a highly transient feature, essential aspects of its dynamics can be understood by resorting to the theory of steady flows past isolated topography. Here consideration is given to the dry flow of an airstream of uniform upstream velocity U and Brunt–Väisälä-frequency N toward a circular obstacle of height H , horizontal scale L , and shape $H(1 + r^2/L^2)^{-3/2}$. Assuming free-slip lower boundary

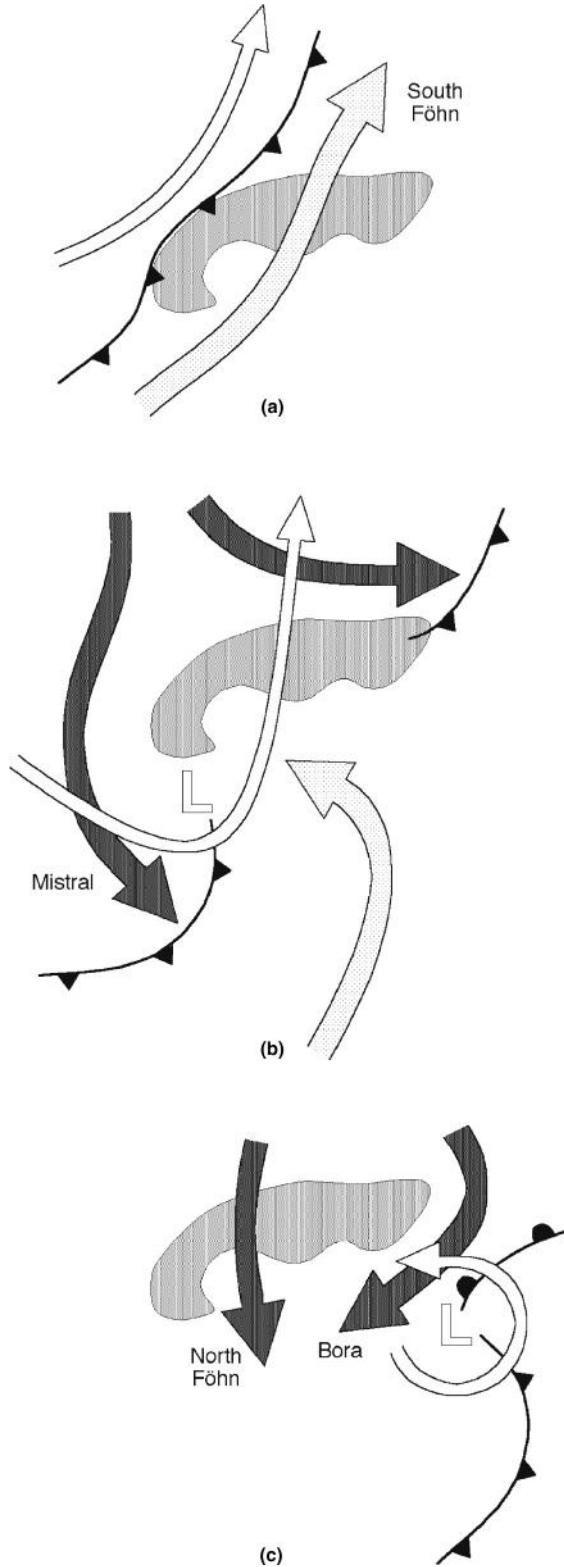


Figure 2 Schematics of an Alpine lee cyclogenesis event showing (a) incipient cold front, (b) cyclone formation, and (c) detachment of cyclone from the Alps. The bold arrows depict low-level air streams with light and dark shading indicating air streams ahead and behind the cold-front. The thin arrows show the upper-level flow.

conditions and hydrostatic dynamics, the flow-response is governed by two dimensionless parameters. These are a Rossby number based on the mountain width, (eqn [1], with f denoting the Coriolis parameter) and a dimensionless mountain height (eqn [2]), which is sometimes referred to as inverse Froude number.

$$R_o = \frac{u}{Lf} \quad [1]$$

$$\varepsilon = \frac{NH}{u} \quad [2]$$

The regime diagram spanned by these parameters is considered in Figure 4.

For the flow system described above, quasi-geostrophic solutions for an unbounded atmosphere exist provided that the dimensionless slope of the obstacle satisfies eqn [3].

$$R_o \varepsilon = \frac{NH}{Lf} < \frac{1}{2} \quad [3]$$

The respective regime boundary is included in Figure 4 as a bold line. The consideration of higher-order balance beyond the quasi-geostrophic system does not notably extend the validity of the balanced dynamics. The main feature of the quasi-geostrophic flow response is the presence of an anticyclone sitting over the mountain top. This is generated by vortex tube compression as the flow is directed over the obstacle. In the Northern Hemisphere, the presence of the mountain anticyclone implies an accelerated (decelerated) flow on the left-hand (right-hand) flank of the mountain (looking downstream). Flows with $\varepsilon > \varepsilon_{\text{cap}} = 3\sqrt{3}/2$ are in addition characterized by a stagnation point and a concomitant Taylor cap (a vertically confined region of closed streamlines), but these flows appear not very relevant to atmospheric conditions.

In the absence of background rotation ($R_o = \infty$), the dimensionless mountain height is the single control parameter in the limit of the inviscid and adiabatic dynamics with a free-slip lower boundary condition. Thus the flow response does not depend upon the slope of the obstacle, but on its height alone. Idealized numerical simulations have served to identify the major flow regimes. For configurations with $\varepsilon < \varepsilon_{\text{crit}} \approx 1.2$, where $\varepsilon_{\text{crit}}$ denotes the critical mountain height, the flow is essentially inviscid and adiabatic, and predominantly directed over (rather than around) the obstacle. In contrast, for $\varepsilon > \varepsilon_{\text{crit}}$, there is low-level flow splitting and the flow is around the obstacle. The later configuration implies the breakdown of the inviscid dynamics due to gravity-wave breaking and/or flow separation, and leads to the formation of a wake. In the presence of background rotation ($R_o < \infty$), the critical mountain height increases with the Rossby number. The respective regime boundary is included in Figure 4 as a bold line. Unlike the regime boundary for balanced solutions, it represents a sharp transition that involves a bifurcation.

To conclude the discussion, the typical range of the flow parameters (R_o, ε) has been included in Figure 4 for several major mountains of the Earth. The parameter range gives a crude estimate that takes into account atmospheric variability. The horizontal scale of the mountains considered has been estimated from topographic maps. In the case of

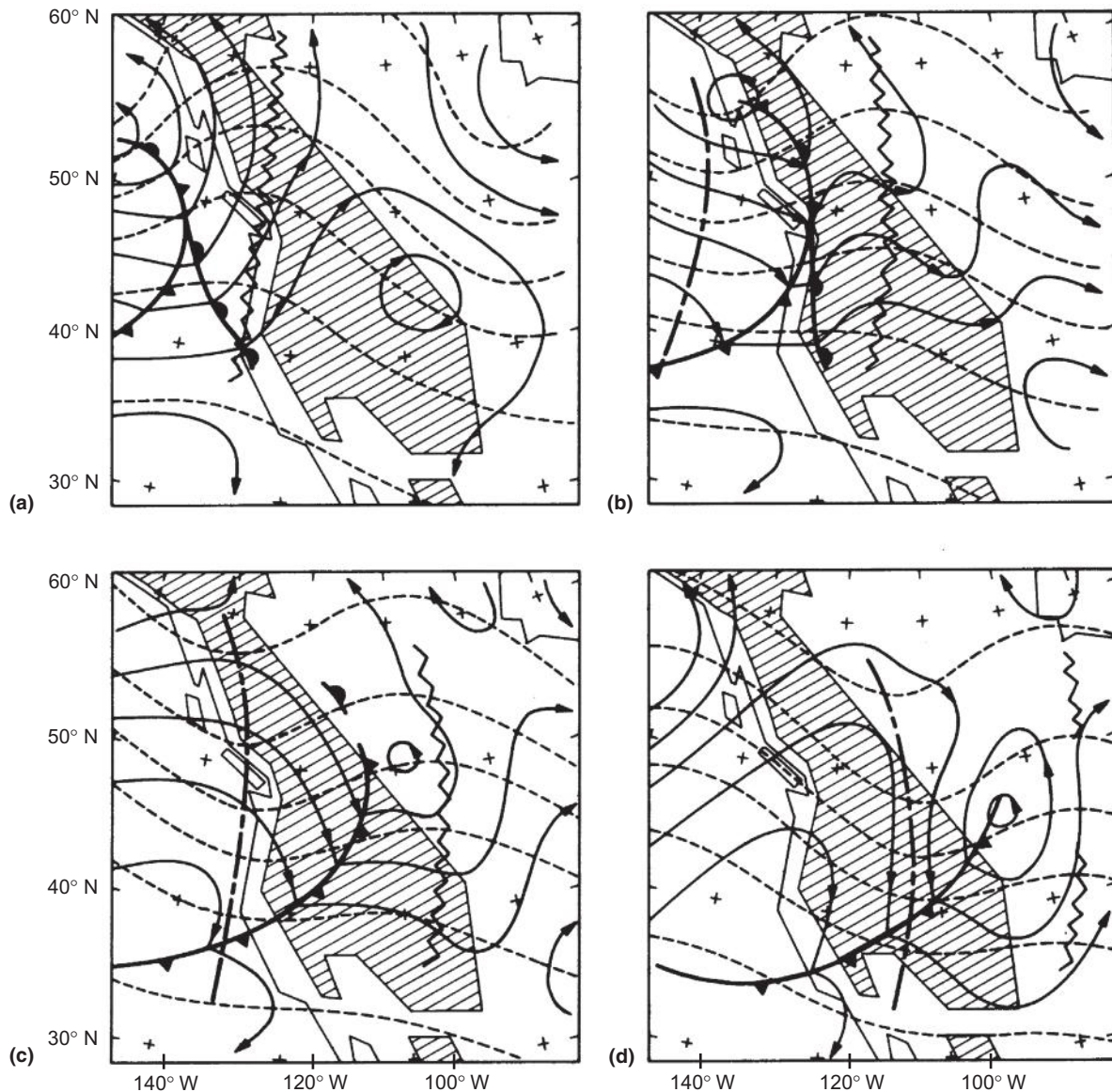


Figure 3 Schematics of Rocky Mountain Lee Cyclogenesis at approximately 12 h intervals. The shading represents a simplified orography with terrain higher than 1500 m. The heavy solid lines denote surface fronts, the solid lines the 1000 hPa geopotential, the dashed lines the 500 hPa geopotential, and the dash-dotted line the position of the upper-level trough. From Palmén and Newton, 1969.

elongated obstacles, the assigned horizontal scale L is based upon the width (rather than the length) of the mountain. For the mountain height, the maximum height has been selected in the case of isolated mountain peaks (e.g., for the Matterhorn), and the mean ridge height in the case of complex mountain ranges (e.g., for the Rocky Mountains and the Alps).

Two important inferences follow from Figure 4. First, most mesoscale mountain ranges (such as the European or Southern Alps) do not allow for a confident application of the quasi-geostrophic or balanced dynamics. For typical mid-latitude values of $f/N \approx 0.01$, the breakdown of the quasi-geostrophic dynamics occurs for obstacle slopes of $H/L \approx 0.005$. Mesoscale mountain ranges thus have horizontal scales that are about half an order of magnitude too small for confident application of the quasi-geostrophic theory. In the case of

strong flow or weak stability, quasi-geostrophic theory may, however, be qualitatively applicable to the flow past the Rocky Mountains or Greenland. Second, most major mountain ranges of the Earth are characterized by a height that often or always exceeds the critical dimensionless mountain height for flow splitting. For many mountain ranges, both flow regimes are possible, depending upon the ambient atmospheric conditions. This implies that regime transition from flowover to flow-around or vice versa may occur.

The Quasi-Geostrophic Framework in the Presence of Topography

The simplest theory for cyclogenesis in the absence of topography is due to Eady's work and applies the framework of the

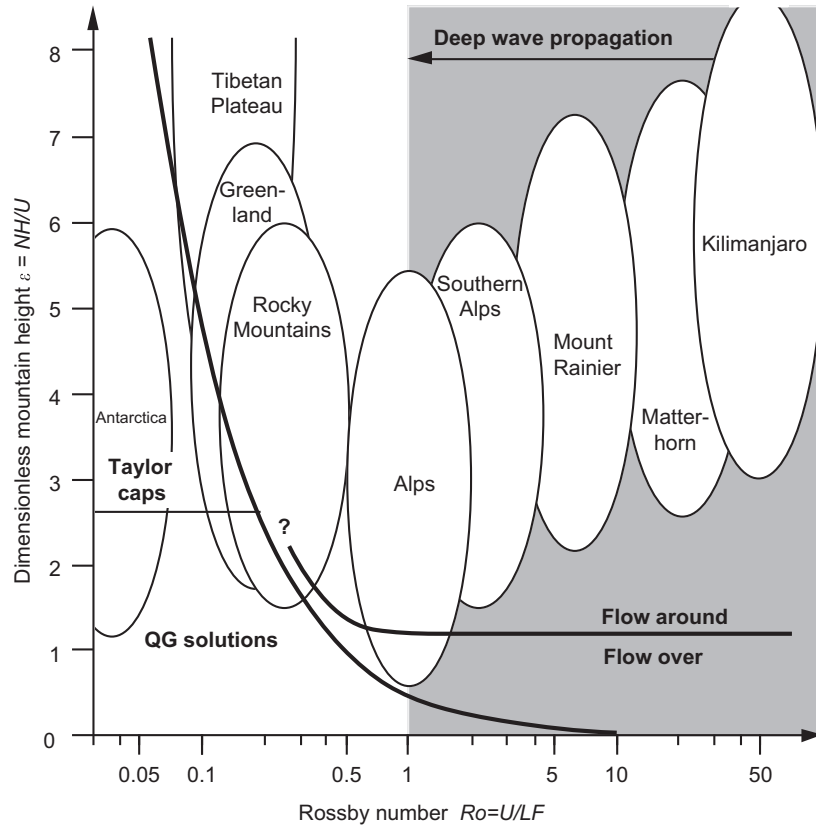


Figure 4 Regime diagram for idealized stratified flows past an isolated obstacle of height H , halfwidth L , and shape $H(1 + L^2/L^2)^{-3/2}$, for an upstream profile characterized by uniform values of U and N . Regime boundaries are indicated by bold lines (see text for details) and the typical parameter ranges of several major mountain ranges are also included.

quasi-geostrophic dynamics (see **Dynamical Meteorology: Quasigeostrophic Theory**) to a continuously stratified atmosphere of uniform PV, confined above and below by the surface and a rigid lid. Several attempts have been made to apply this framework to orographic cyclogenesis. Here we begin by considering the dynamical foundations of these studies and consider the quasi-geostrophic dynamics of an adiabatic and inviscid Boussinesq atmosphere on an f -plane. The dynamics is governed by the conservation of relative potential vorticity q^* (eqn [4]) in the interior of the fluid, the thermodynamic equation [5] on the horizontal bounding surfaces, and the hydrostatic relation [6]).

$$\frac{D_g q^*}{Dt} = 0 \quad \text{with} \quad q^* = \zeta_g + \frac{f^2}{N^2} \frac{\partial^2}{\partial z^2} \left(\frac{p^*}{\rho_0} \right) \quad [4]$$

$$\frac{D_g}{Dt} \left(\frac{g\theta^*}{\Theta} \right) + wN^2 = 0 \quad [5]$$

$$\frac{\partial}{\partial z} \left(\frac{p^*}{\rho_0} \right) = \frac{g\theta^*}{\Theta} \quad [6]$$

Here the starred quantities denote the deviation from the Boussinesq background state (ρ_0, θ_0, p_0) of constant Brunt–

Väisälä frequency N , Θ is a constant reference potential temperature, the D_g/Dt operator refers to the total derivative following the geostrophic motion (eqn [7]), and ζ_g is the relative geostrophic vorticity $\zeta_g = -\partial u_g/\partial y + \partial v_g/\partial x$.

$$(u_g, v_g) = \frac{1}{f} \left(-\frac{\partial}{\partial y}, \frac{\partial}{\partial x} \right) \frac{p^*}{\rho_0} \quad [7]$$

For the classical Eady problem, the atmosphere is confined above and below by rigid horizontal surfaces at $z = 0$ and z_T , representing the Earth's surface and a rigid tropopause. In the presence of topography, several studies have employed the 'shallow-mountain' approximation, whereby the lower boundary condition is applied at $z = 0$ rather than at the real topographic height. This yields eqn [8], with $h(x, y)$ denoting the height of the specified orography.

$$w = \begin{cases} \mathbf{v}_g \cdot \nabla h & \text{for } z = 0 \\ 0 & \text{for } z = z_T \end{cases} \quad [8]$$

In effect, the shallow-mountain approximation replaces the mountain by 'equivalent' inflow and outflow through the plane $z = 0$, while it retains other nonlinearities of the flow. In comparison with the full lower boundary condition, the shallow-mountain approximation can be shown to properly represent the mountain volume (and thus the far-field circulation), while it underestimates the mountain height and slope.

If the flow is of constant PV at an initial time $t = 0$, then eqn [4] implies that it remains so forever. In this case, the governing equations may further be simplified. Upon defining the quasi-geostrophic streamfunction [9], eqn [4] simplifies to a homogeneous elliptic equation [10].

$$\psi = \frac{1}{f} \frac{p^*}{\rho_0} \quad [9]$$

$$q^* = \nabla_h^2 \psi + \frac{f}{N^2} \frac{\partial^2 \psi}{\partial z^2} = 0 \quad [10]$$

The lower and upper boundary conditions are of Neumann type and determined by the thermodynamic equation [5] in combination with the simplified boundary conditions [8].

$$\frac{D_g}{Dt} \left(\frac{\partial \psi}{\partial z} \right) + \frac{N^2}{f} w = 0 \quad \text{for } z = 0, z_T \quad [11]$$

Some of the mechanisms to be discussed below are directly based on the quasi-geostrophic system as laid out above; for others it will be necessary to relax the quasi-geostrophic assumption for a part of the relevant processes.

Modification of Baroclinic Instability

One of the early theories of lee cyclogenesis views orographic cyclogenesis either as a modification of baroclinic instability by topography or as a scattering problem studying the effects of topography on incident growing baroclinic modes. These theories in essence represent a modification of Eady's problem. The basic state flow is thus assigned a vertically sheared baroclinic structure (eqn [12]).

$$U(z) = \Lambda z \quad [12]$$

In the absence of topography, the linearization in a periodic channel yields the classical Eady instability problem, which supports growing baroclinic modes.

In the presence of topography, the linearized version of the simplified lower boundary condition [8] is combined with [12] to give eqn [13], where dashed quantities denote the perturbation.

$$w' = \mathbf{v}'_g \cdot \nabla h \quad (z = 0) \quad [13]$$

It is important to note that the shallow-mountain approximation in combination with $U(z=0) = 0$ implies that only the perturbation can interact with the topography, but not the basic state flow itself. Anticyclonic and cyclonic vorticity may

then be generated through vortex tube compression and expansion, and this occurs when the perturbation flow is respectively up and down sloping topography. The first-order response is sketched in Figure 5 for two different geometries of an idealized elongated obstacle with a length scale of several thousand kilometers. When the topography is aligned from north to south (as for the Rocky Mountains), an incident low experiences strengthening to the north and weakening to the south over the upstream slope of the topography, and the opposite pattern downstream. When the topography is aligned from west to east (as qualitatively the case for the Alps, but note the difference in scale), the effect of the topography is to induce a low-high pressure dipole across the obstacle, the sign of which is determined by the phase of the wave.

The full mathematical problem of orographically modified normal modes may be dealt with either analytically using expansion techniques, or numerically using a continuously stratified or two-layer model. An example of a numerical solution is reproduced in Figure 6, which shows the surface streamfunction of a growing mode for the geometry of the Rocky Mountain. Upstream of the obstacle, the low drifts toward north and weakens, while a new surface low reappears to the lee in the south. This sequence of events is qualitatively comparable with that of observed Rocky Mountain lee cyclogenesis events (see Figure 3).

An important feature of the baroclinic instability mechanism of lee cyclogenesis is that the growth rates of the classical baroclinic problem are reduced by the presence of topography (since it inhibits meridional advection of low-level air). Thus, the mechanism is unable to explain the rapid growth rates that are commonly observed in the first phase of Alpine lee cyclogenesis events, for example.

Generation of Baroclinic Lee Waves

In comparison to the baroclinic instability mechanism discussed above, the baroclinic lee wave theory does not include any baroclinic growth but provides a possible explanation for the rapid initial growth and pressure drop. The theory is based on essentially the same dynamical framework as detailed above, with the following alterations. First, the rigid-lid atmosphere is replaced by an unbounded atmosphere $0 \leq z \leq \infty$. By avoiding the rigid lid (the Eady problem) and the β -plane (the Charney problem), baroclinic waves are stabilized in spite of the available potential energy. Second, the basic-state flow is replaced by eqn [14].

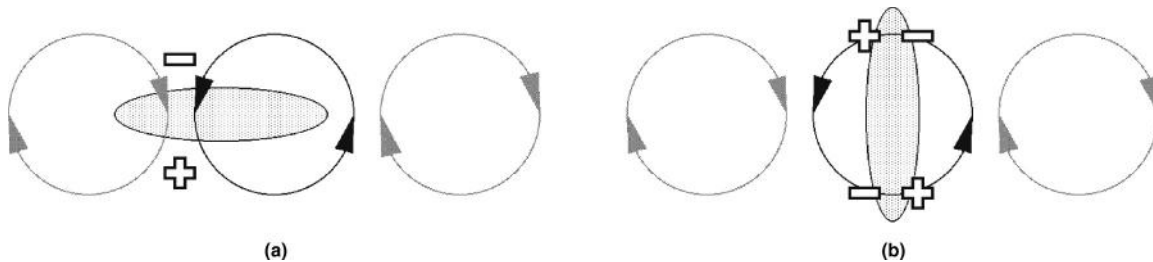


Figure 5 First-order vorticity generation (+) and destruction (−) by vortex stretching for the baroclinic instability mechanism of lee cyclogenesis. The two panels show the interaction of an Eady mode with west–east and north–south oriented topographic obstacles, respectively.

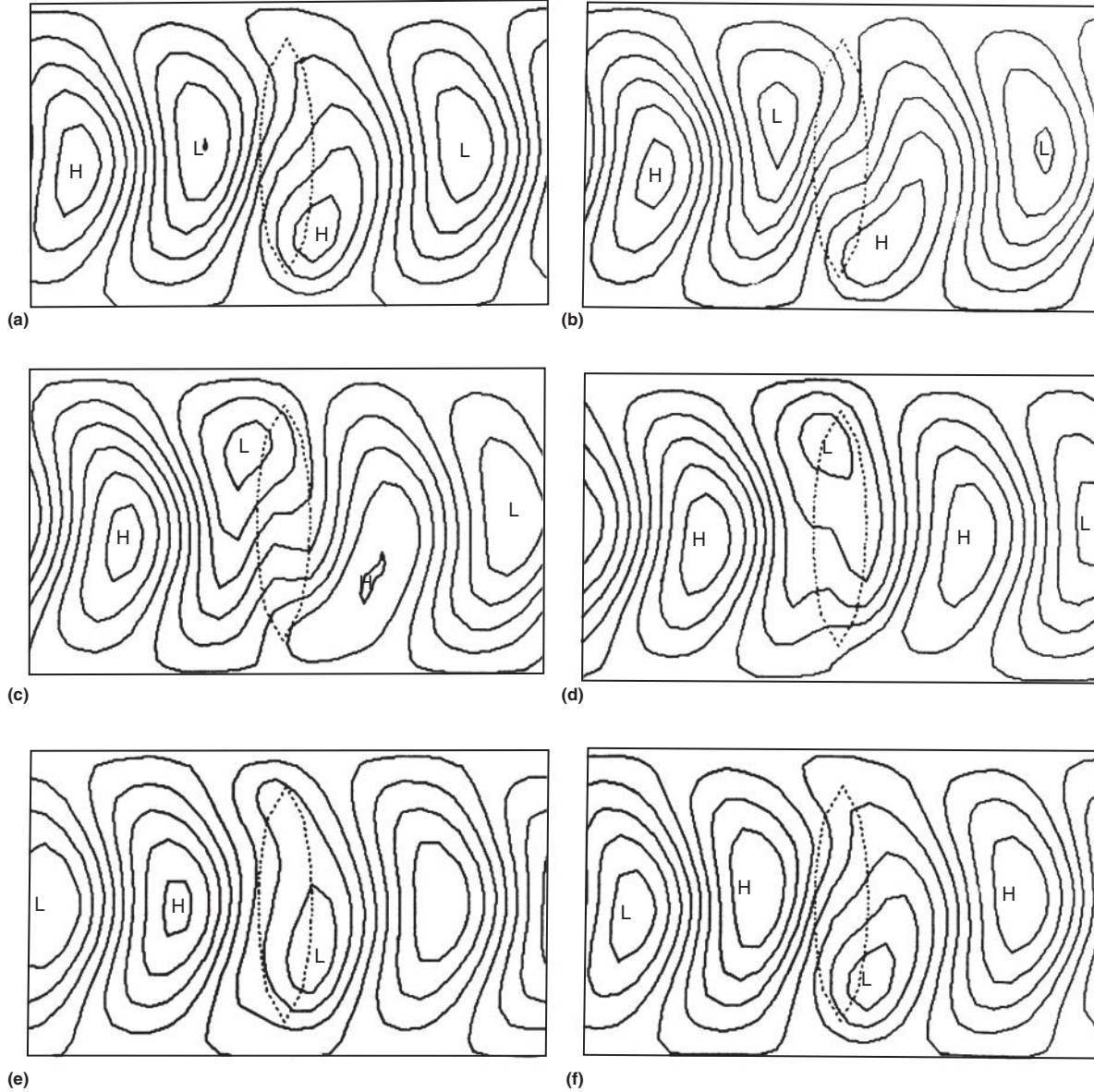


Figure 6 Example of an orographically modified normal mode for the geometry of the Rocky Mountain lee cyclogenesis. The panels (a) to (f) show the phases of the development in terms of the low-level streamfunction. Reproduced from Buzzi, A., Speranza, A., Tibaldi, S., Tosi, E., 1987. A unified theory of orographic influences upon cyclogenesis. *Meteorology and Atmospheric Physics* 36, 91–107.

$$U(z) = U_0 + \Lambda z \quad [14]$$

Unlike eqn [12], this flow may have $U(z) \neq 0$, and thus include the interaction of the basic-state flow with the underlying topography. The choice of this particular configuration allows for a clear distinction between unstable baroclinic growth (previous subsection) and orographic forcing (this subsection).

The baroclinic waves supported by this unbounded configuration are surface-trapped and decay exponentially with height. For a two-dimensional wave with wavenumber k and in absence of topography, the quasi-geostrophic streamfunction is given by eqn [15].

$$\psi'(x, z) \propto e^{-z/S} e^{i(kx + \omega t)} \quad [15]$$

In eqn [15], ω denotes the frequency, and eqn [16] gives the steering level.

$$S = \frac{f}{N} |k|^{-1} \quad [16]$$

The phase velocity of these waves is determined by the wind at the steering level (eqn [17]).

$$c_p = S(z = H) \quad [17]$$

The system supports stationary baroclinic waves provided there is a wind-reversal at some height S^* , i.e., $U(z = S^*) = 0$. The wavelength of these waves may then be obtained from eqn [16].

The baroclinic lee-wave theory views the initial rapid phase of lee cyclogenesis as the formation of a standing baroclinic

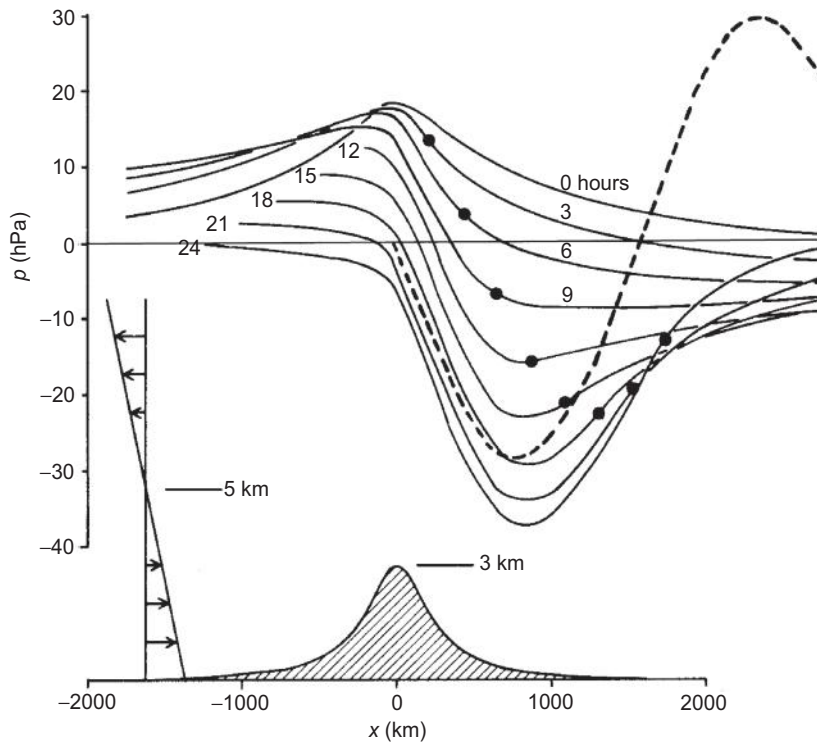


Figure 7 Transient development of a standing baroclinic wave to the lee of a two-dimensional ridge in a sheared background flow. The initial mountain anticyclone exhibits a rapid decay and there is a strong pressure drop to the lee (surface pressure (in hPa)). The dashed curve shows the asymptotic steady-state lee wave. The distance covered by the group velocity is indicated by the dot on each curve. Reproduced from Smith, R.B., 1984. A theory of lee cyclogenesis. *Journal of the Atmospheric Sciences* 41, 1159–1168.

lee wave in an initial-value problem. An example for a two-dimensional configuration (with the x -axes pointing perpendicular to an infinite ridge) is shown in [Figure 7](#). Starting from some initial conditions (here from the mountain anticyclone), a standing baroclinic wave quickly develops, with an associated rapid pressure drop of 35 hPa in 24 h. The key requirement of this mechanism is the presence of a wind-reversal with height, a configuration that is often (but not always) met in the Alpine cases but is rarely met with Rocky Mountain lee cyclogenesis.

Low-Level Blocking of Approaching Cold Air

The mechanisms discussed in the previous two subsections are entirely based on the quasi-geostrophic dynamics (assuming small Rossby numbers) and on a simplified lower boundary condition (implying absence of flow splitting). As discussed above, these conditions are rarely met with major topographic obstacles. Most orographic cyclones thus entail the violation of the balanced and the inviscid dynamics during the first phase of their development. Nevertheless, the formation of the mature cyclone during a second phase is likely to follow a largely balanced evolution. A dynamical interpretation may be sought by considering the first-phase unbalanced generation of orographic flow perturbations but assuming that their second-phase interaction with the synoptic-scale dynamics is approximately balanced. It then follows from the invertibility principle that relevant orographic perturbations must be associated with (i) a surface potential temperature anomaly (to be treated in this subsection), and/or (ii) an internal PV anomaly (to be

treated in the next subsection). These flow anomalies in essence represent a conceptual intermediary between the (unbalanced) mesoscale dynamics, and the subsequent (essentially balanced) interaction of orographically generated flow anomalies with the synoptic-scale environment.

A well-documented mechanism for the generation of surface potential temperature anomalies is the retardation and deformation of an approaching cold front. When cold air advection interacts with orographic flow splitting, a warm anomaly results within the sheltered lee of the obstacle. An example of this process is presented in [Figure 8](#). It shows an Alpine trajectory analysis derived from a real-case numerical simulation (with 14 km horizontal resolution). The starting points for the relevant trajectories were selected immediately behind an impinging cold front. Pronounced flow splitting occurs and the track of the trajectories encompasses an extended wake region that is ~ 5 K warmer than the surrounding air. It is important to realize that the scale of the resulting surface- θ anomaly may differ from that of the topography. In the example shown, the wake anomaly has an extension of almost 1000 km in the downstream direction, exceeding the Alpine width by a large factor. A surface anomaly of this size can well interact with the larger-scale balanced part of the flow evolution.

The orographic retardation of cold air advection implies geostrophic (or higher-order balanced) adjustment of the heavily distorted baroclinic configuration. This may induce a pressure drop and associated vertical motion. However, from the viewpoint of the balanced dynamics, it is the warm

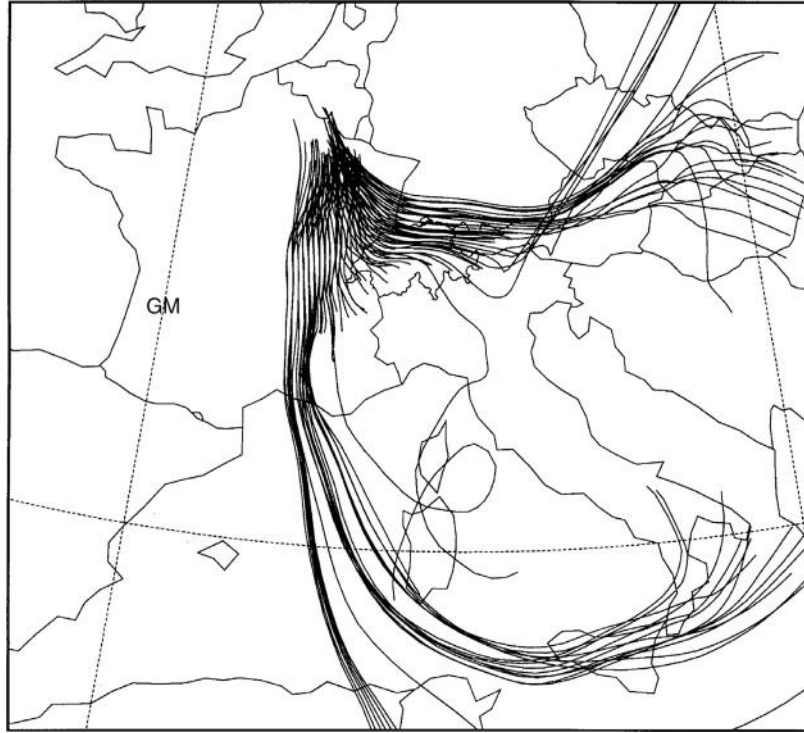


Figure 8 Splitting of an incident low-level airstream (initially located at the 850 hPa level) by the Alps. The figure derives from three-dimensional 54 h trajectory computations driven by a numerical simulation of a cold-frontal passage on 30 April 1982. Reproduced from Kljun, N., Sprenger, M., Schär, C., 2001. The modification of a front by the Alps: a case study using the ALPEx reanalysis data set. *Meteorology and Atmospheric Physics* 78, 89–105.

low-level air itself that constitutes the driving agent of this development. The implications in terms of surface pressure perturbation may be estimated from quasi-geostrophic theory. To this end, consider a circular surface- θ anomaly of the form of eqn [18], with r denoting the radius and L its horizontal scale.

$$\theta^*(r, z = 0) = \Delta\theta \left[\left(\frac{r}{L} \right)^2 + 1 \right]^{-1/2} \quad [18]$$

In an unbounded atmosphere, the associated three-dimensional quasi-geostrophic streamfunction can be found by solving eqn [10] subject to the lower boundary condition obtained from the hydrostatic relation [6], i.e.,

$$\frac{\partial \psi}{\partial z} = \frac{1}{f} \frac{g\theta^*}{\Theta} \quad [19]$$

It can easily be verified that expression [20] is the desired solution.

$$\psi(r, z) = -\frac{g\Delta\theta}{\Theta} \frac{L}{N} \left[\left(\frac{r}{L} \right)^2 + \left(\frac{N}{f} \frac{z}{L} + 1 \right)^2 \right]^{-3/2} \quad [20]$$

Thus, the surface pressure perturbation in the center of the anomaly is given by eqn [21].

$$\Delta p = -f\rho_0 \frac{L}{N} \frac{g\Delta\theta}{\Theta} \quad [21]$$

Using typical mid-latitude values for f and N , and a potential temperature anomaly with an amplitude of $\Delta\theta = 10$ K and a horizontal scale of $L = 500$ km, one finds a pressure perturbation with an appreciable amplitude of ~ 17 hPa.

Generation of Low-Level Potential Vorticity Anomalies

In addition to surface potential temperature anomalies, orographic wakes comprise low-level internal PV anomalies that may interact with the balanced synoptic-scale environment. The generation of such PV anomalies is intimately related to the generation of lee vortices (see **Mountain Meteorology: Lee Vortices**). In essence, PV in lee vortices results from the violation of PV conservation by diabatic or viscous processes. Relevant processes for orographic flows include surface friction, dissipation by either gravity-wave breaking or flow separation, and diabatic heating associated with orographic precipitation. The formation of lee vortices is familiar from mountainous islands, where they may lead to spectacular Karman vortex streets on satellite pictures.

Wakes with low-level PV features can often be identified in numerical weather prediction models. In the case of the Alps, the formation of elongated PV streamers in the wake appears to occur almost whenever there is appreciable flow past the Alpine ridge. An example is shown in the left-hand panels of **Figure 9**. These show the formation of a large number of orographic PV streamers at the 850 hPa level, which will be referred to as ‘PV banners’. The situation considered is that of a cold frontal passage followed by deep lee cyclogenesis. The figure is based on a mesoscale simulation with a horizontal resolution of 14 km.

Individual pairs of banners with anomalously positive and negative values of PV can be attributed to individual flow-splitting events, either on the scale of the whole of the Alps (primary banners) or on that of individual massifs and peaks of

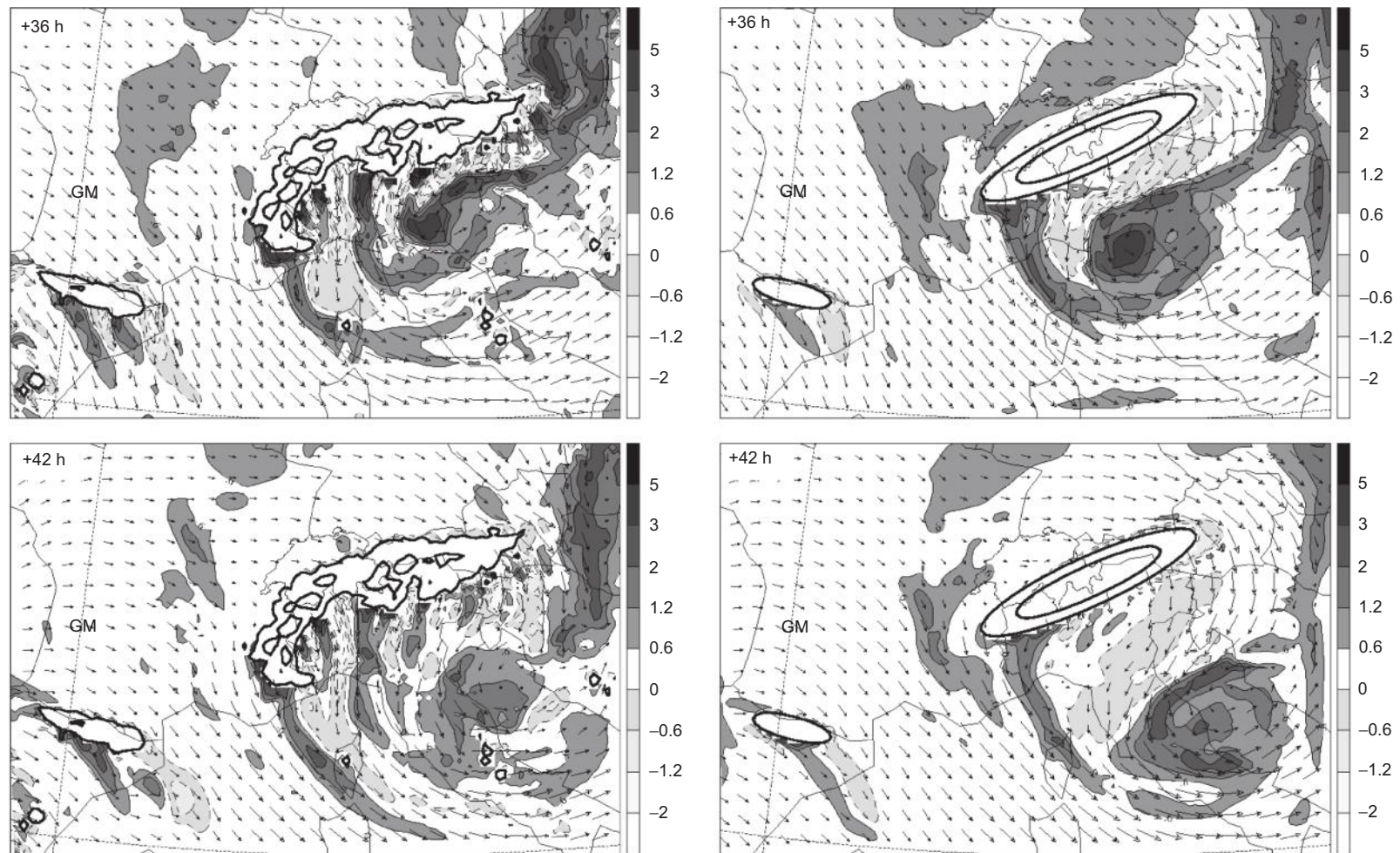


Figure 9 Potential vorticity (PV units) and horizontal wind vectors on the 850 hPa level for an Alpine lee cyclogenesis event. The data are from real-case numerical simulations using the full topography (left-hand panels) and a simplified topography (right-hand panels), respectively. The heavy lines denote topographic height at 1000 and 2200 m. Reproduced from Aebischer, U., Schär, C., 1998. Low-level potential vorticity and cyclogenesis to the lee of the Alps. *Journal of the Atmospheric Sciences* 55, 186–207.

the model topography (secondary banners). The simulated PV banners have amplitudes of several PV units, and grow in length when the orographically generated PV anomalies are advected downstream. In this way, Alpine banners can attain a length of up to 1500 km on occasions. The existence and apparent stability (with respect to shedding instabilities) of such banners has been verified observationally within the Mesoscale Alpine Programme (MAP) conducted in 1999.

The formation of a multitude of banners has implications for their far-field effects, and thus their ability to affect the balanced dynamics.

In essence, the far-field effect will be dominated by the primary (outermost) banners, while the effect of the secondary banners will decay quickly with height, as governed by the associated Rossby depth of deformation af/N , where a denotes the distance between neighboring banners. Only the primary banners may invoke deep interaction with other flow features, such as an approaching upper-level PV streamer. The right-hand panels of [Figure 9](#) suggest this kind of effect. The respective simulation is identical to that in the left-hand panels, except for the use of idealized ellipsoidal topography. It shows how the primary PV banner to the west of the Alpine ridge rolls up and contributes to the low-level PV within the initial core of the developing lee cyclone.

Synthesis

In the 1980s, the multitude of processes that can contribute to lee cyclogenesis triggered a scientific debate about lee cyclogenesis theories. A sophisticated testing methodology was developed that involves projecting the key assumptions of a theory (such as a linearization strategy and lower boundary conditions) onto a numerical simulation of a lee cyclogenesis event. The methodology is applicable to theories with well-defined mathematical stipulations (such as the modified baroclinic instability and the baroclinic lee-wave mechanisms discussed above). The application of the testing methodology has demonstrated that none of the tested mechanisms alone is sufficient to explain the simulated lee cyclones, despite some support for these mechanisms from case studies and statistical investigations.

It thus appears that a unique mechanism for lee cyclogenesis does not exist. Rather lee cyclogenesis derives from a combination of different factors, with relative contributions that vary from mountain to mountain and case to case.

Nevertheless, the blocking of low-level cold air in combination with upper-level PV advection appears to play at least some role in most of the cases. This mechanism is well compatible with the classical cyclogenesis theory in absence of topography, and applicable to many mountain ranges, among them the Alps and the Rocky Mountains. It is interesting to note that as early as 1920, orographic blocking and cold frontal retardation was mooted by the Austrian meteorologist Heinrich von Ficker as being the likely explanation for Alpine lee cyclogenesis. In the 1950s, the mechanism was revived by case studies, and in the early 1970s by the first numerical modeling work on the subject.

See also: **Dynamical Meteorology:** Balanced Flow; Baroclinic Instability; Potential Vorticity; Quasigeostrophic Theory; Vorticity. **Mountain Meteorology:** Lee Vortices. **Synoptic Meteorology:** Cyclogenesis; Extratropical Cyclones.

Further Reading

- Aebischer, U., Schär, C., 1998. Low-level potential vorticity and cyclogenesis to the lee of the Alps. *Journal of the Atmospheric Sciences* 55, 186–207.
- Bannon, P.R., 1992. A model of rocky-mountain lee cyclogenesis. *Journal of the Atmospheric Sciences* 49, 1510–1522.
- Bleck, R., Mattocks, C., 1984. A preliminary analysis of the role of potential vorticity in Alpine lee cyclogenesis. *Contributions to Atmospheric Physics* 57, 357–368.
- Buzzi, A., Speranza, A., Tibaldi, S., Tosi, E., 1987. A unified theory of orographic influences upon cyclogenesis. *Meteorology and Atmospheric Physics* 36, 91–107.
- Kijun, N., Sprenger, M., Schär, C., 2001. The modification of a front by the Alps: a case study using the ALPEx reanalysis data set. *Meteorology and Atmospheric Physics* 78, 89–105.
- Palmén, E., Newton, C.W., 1969. *Atmospheric Circulation Systems*. Academic Press, London.
- Pichler, H., Steinacker, R., 1987. On the synoptics and dynamics of orographically induced cyclones in the Mediterranean. *Meteorology and Atmospheric Physics* 36, 108–117.
- Pierrehumbert, R.T., 1986. Lee cyclogenesis. In: Ray, P.S. (Ed.), *Mesoscale Meteorology and Forecasting*, Boston: American Meteorological Society, pp. 493–515.
- Smith, R.B., 1984. A theory of lee cyclogenesis. *Journal of the Atmospheric Sciences* 41, 1159–1168.
- Taffermer, A., Egger, J., 1990. Test of theories of lee cyclogenesis—ALPEx cases. *Journal of the Atmospheric Sciences* 47, 2417–2428.
- Tibaldi, S., Buzzi, A., Speranza, A., 1989. Orographic cyclogenesis. In: *Proceedings of the Palmén Memorial Symposium on Extratropical Cyclones*, 1988. American Meteorological Society, Helsinki, pp. 107–128.
- Whittaker, L.M., Horn, L.H., 1984. Northern hemisphere extratropical cyclone activity for four mid-season months. *International Journal of Climatology* 4, 297–310.

Valley Winds

D Zardi, University of Trento, Trento, Italy

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by J. Egger, volume 6, pp 2481–2490, © 2003, Elsevier Ltd.

Synopsis

The article reviews the distinctive properties of daily periodic, thermally driven winds typically occurring in mountain valleys. The mechanisms leading to the development of the horizontal pressure gradients driving such winds are examined, emphasizing the basic processes involved, as well as their connections with key physical factors, such as solar radiation and topographic features. Interactions with other thermally driven winds, such as slope flows on valley sidewalls and mountain-plain circulations, are also discussed. Implications for phenomena associated with the water cycle and moist processes, as well as with air pollution transport, are also briefly outlined.

Introduction

Valley winds are daily periodic, thermally driven circulations which develop in mountain valleys under favorable synoptic conditions, i.e., clear sky and weak or absent upper ambient winds. A valley may be loosely defined as a rather elongated orographic channel, either confined between two elevated mountain ranges, or carved into a plain (such as a canyon). A valley geometry is typically marked by the processes that concurred to shape it. These may consist in earth crust modifications by tectonic movements, erosion by glacier expansion and retreat, terrain erosion and sediment deposition performed by rivers, long-term action of atmospheric factors, or various combinations of these processes. Valleys are usually open to adjacent plains, unlike closed basins – such as the Meteor Crater in Arizona, USA – which are completely surrounded by elevated walls. Typical geometrical elements of a valley include the floor and the sidewalls. Accordingly, among the various geometrical factors affecting valley winds, we will mainly refer to the valley floor width, the sidewall slope angle, and the sidewall crest height above the valley floor. Clear skies allow strong incoming shortwave radiation during daytime, as well as outgoing longwave radiation during nighttime, and thus favor both diurnal surface heating and nocturnal cooling, i.e., the required thermal forcing. In addition, weak or absent upper winds allow an unperturbed development of these flows, which will be then only controlled by the combination of surface thermal forcing and topographic features.

Valley winds typically blow up-valley during daytime, and down-valley during nighttime. These flows are actually the lower, and more evident, branches of closed circulations, with the respective upper branches flowing in the opposite directions above the sidewall ridge-top level (Figure 1). However, these upper equalizing flows, sometimes called *antiwinds*, are more difficult to detect: being unconfined by sidewalls, unlike the valley winds below, they can occur over larger widths, and comparable depths, so they usually display smaller mean speeds. Moreover, they are more directly exposed to the perturbing influence of upper ambient winds.

Basically valley winds develop as a consequence of horizontal pressure gradients between the interior of a valley and an adjacent plain. Such gradients arise because the air

temperature cycle, associated with the daytime warming and nighttime cooling of the atmosphere, in the interior of a valley, i.e., up to a height comparable with the mean sidewall crest elevation, typically displays at any level a larger amplitude than at the same level above an adjacent plain. Accordingly, as the vertical pressure distribution is to a large extent controlled by the hydrostatic balance, these different vertical thermal profiles produce horizontal valley–plain pressure gradients, which, reversing twice per day, drive up-valley winds during daytime and down-valley winds during nighttime (Figure 1).

Notice that valley winds are not the only thermally driven circulations, which may develop over complex terrain. Actually they may be viewed as one component of a whole system of interconnected diurnal winds, typically occurring, at different scales, under the same favorable synoptic situations, over major mountain ranges surrounded by plain areas. The simplest flows are *slope winds*, blowing up the sunlit inclines during daytime, and downward during nighttime. Up-slope flows are sometimes called *anabatic* winds, and down-slope flows *katabatic winds* (from the ancient Greek verbs '*anabainein*', '*going up*', and '*katabainein*', '*going down*'). In particular the word *katabatic wind* is extensively used to denote drainage winds blowing on glaciers and large ice surfaces in the polar regions (see **Mountain Meteorology: Katabatic Winds**). Such flows also occur along the sloping sidewalls of large valleys, and contribute, as shown in the following section, to the development of *valley winds*. At even larger scales, an overall organized air motion can be identified, conveying air toward the mountain range in the lower layers during daytime, and promoting a reversed flow during nighttime. These low-level flows are also accompanied by corresponding return upper winds, flowing in the opposite directions. Altogether this comprehensive system forms the so-called *mountain-plain circulation*.

Further cases of daily periodic thermally driven circulations are found in closed basins, as well as on isolated plateaus.

The present article mainly focuses on valley winds. Slope flows and mountain-plain circulations will be only briefly outlined, insofar as it is convenient for understanding their connections with valley winds. The interested reader will find an extensive treatment of all the above circulations in the review by Zardi and Whiteman (2013).

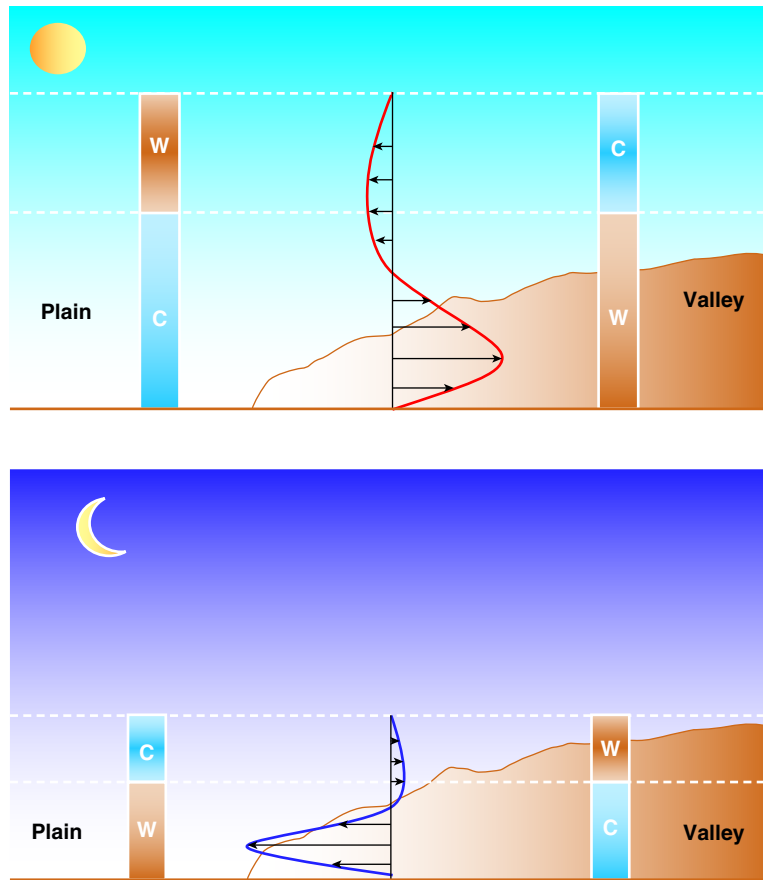


Figure 1 Idealized picture of the vertical structure associated with daytime up-valley wind (upper panel) and nighttime down-valley wind (lower panel) between a valley and an adjacent plain. The red and blue curves reproduce vertical profiles of the horizontal wind component at the valley inlet. The two columns of air – one representing of the atmosphere over the valley floor, and one above the plain – include red and blue sections, indicating layers where potential temperature is relatively warm (W) or cold (C). The free atmosphere is unperturbed by the daily cycle above the tops of the columns. Reproduced with permission from Zardi, D., Whiteman, C.D. 2013. Diurnal mountain wind systems. In: Chow, F.K., De Wekker, S.F.J., Snyder, B. (Eds.), *Mountain Weather Research and Forecasting – Recent Progress and Current Challenges*, Springer Atmospheric Sciences. Springer, Berlin, pp. 37–122. Adapted from Whiteman, C.D., 2000. *Mountain Meteorology: Fundamentals and Applications*. Oxford University Press, New York, 355pp.

Being forced by the diurnal cycle of surface radiation, all of these circulations display a similar daily periodic up-and-down reversal. Nevertheless, the cycles of the various circulations occurring in the same area may be differently phased among them. Many factors may affect these phase delays, such as the varied exposure of the underlying terrain to incoming solar radiation, the different timing and amounts of sensible heat flux allowed by unequal surface properties, and the intrinsically different reaction timescales resulting from the circulation extent (timescales are typically larger for circulations encompassing larger areas).

Also, diurnal valley winds are not the only circulations that may be found in mountain valleys, where various kinds of airflows may be produced by the action of upper winds, or pressure gradients associated with synoptic scale structures, rather than from the diurnal cycle of incoming solar radiation. The former are called *dynamically driven winds*, as opposed to the *thermally driven* valley winds, which are the main subject of the present article. Indeed, as the latter are originated from the heating and cooling of the valley atmosphere via surface energy

budgets, they are not caused by upper winds blowing above the valley, which may possibly condition, but not primarily determine, their evolution and spatial structure. In fact, thermally driven diurnal wind systems are better developed when the above external forcings are weaker. This may be argued from the panels on the first row of [Figure 2](#): here the direction of valley winds is not affected by the direction of the upper geostrophic wind, and this is the only case displaying a natural diurnal cycle. In contrast, dynamically driven valley winds are clearly marked by the action of the forcing synoptic systems. Indeed three main channelling scenarios are associated with the ambient wind above the valley, and arise either from downward transport of horizontal momentum; from channelling of ambient winds by the valley sidewalls, and the consequent alignment with the valley axis; or from pressure-driven channelling. The connection between geostrophic wind direction and the resulting channelled wind in the valley is schematically shown in [Figure 2](#) for all the above cases.

The first mechanism consists in a strong downward transport of horizontal momentum from above the valley, caused

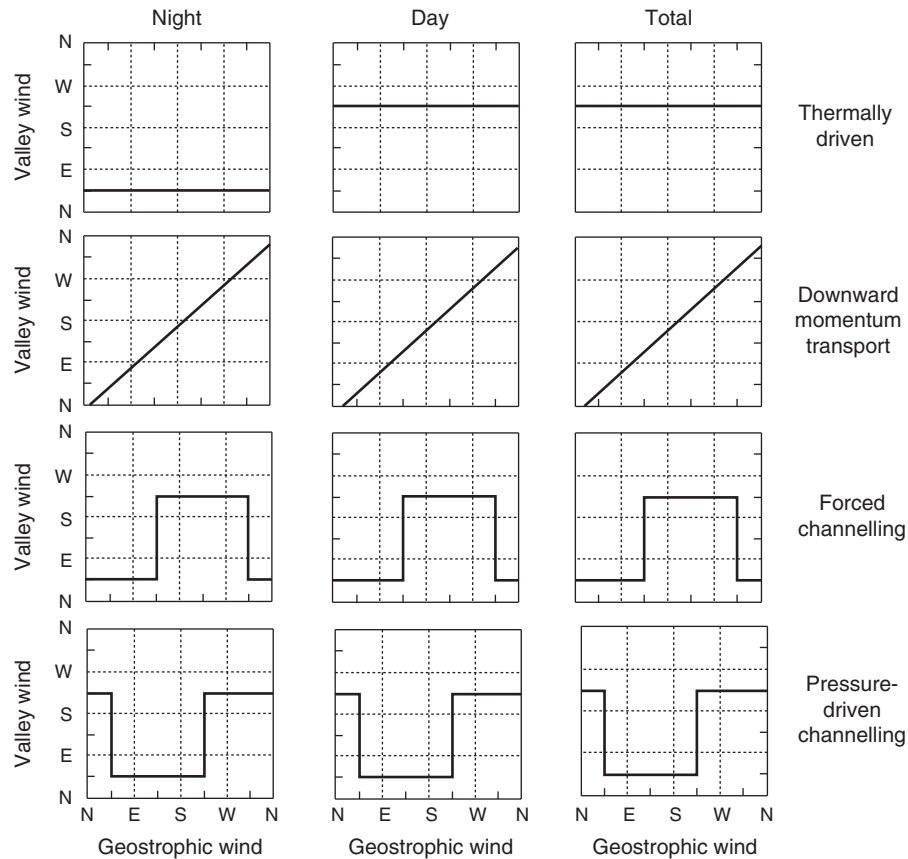


Figure 2 The relationship between above-valley (geostrophic) and valley wind directions for four possible forcing mechanisms: thermal forcing, downward momentum transport, forced channelling, and pressure-driven channelling. The valley is assumed to run from northeast to southwest. Reproduced with permission from Whiteman, C.D., Doran, J.C., 1993. The relationship between overlying synoptic-scale flows and winds within a valley. *Journal of Applied Meteorology* 32, 1669–1682.

by either turbulent mixing, gravity waves, or other mechanisms. Friction determines a slight turning of the wind from the geostrophic wind direction toward the lower pressure as the ground is approached. This coupling is most likely to occur during unstable or neutrally stratified conditions in wide valleys, with low sidewalls and flat bottom, where thermally driven winds are less likely to develop, and channelling along the valley axis by the sidewalls would be rather ineffective.

A second mechanism is known as *forced channelling*: it occurs when ambient winds, which are in geostrophic balance above the valley, are channelled by the valley sidewalls so that the wind aligns with the valley axis. In this case, the direction and speed of the valley wind depend on the sign and magnitude of the component of the ambient wind projected along the valley axis. Wind blows up or down the valley axis depending on the direction of the geostrophic wind relative to it. As a consequence, winds are predominantly aligned with the valley axis, with sudden shifts when geostrophic wind shifts across a line normal to the valley axis.

A third mechanism is *pressure-driven channelling*, which occurs when the wind in the valley is driven by the component of the geostrophic pressure gradient in the along-valley direction. This gradient will be zero only when geostrophic wind is directed along the valley axis. Winds in the valley will shift from up to down valley (or vice versa) when the geostrophic wind

direction shifts across the valley axis. Pressure-driven channelling produces winds blowing predominantly along the valley axis, as for the previous case, but the valley wind reversal occurs for geostrophic wind directions 90° different from those of the forced channelling mechanism. Notice that winds in the valley can blow in opposition to along-valley wind direction components above the valley.

The above *dynamically driven* winds may typically display strong intensities, depending on how strong the forcing synoptic situations are. In contrast, *thermally driven* valley winds generally display weak to moderate speeds. Nevertheless, some remarkable examples of strong valley winds have been observed. For instance, in the Kali Gandaki Valley in Nepal, connecting the Indian Plain to the Tibetan Plateau (Figure 3), wind speeds in the order of $5\text{--}15\text{ m s}^{-1}$ are commonly met (Figure 4). Strong and gusty wind speeds may also occur when valley winds blow through elevated rims or gaps overlooking lower lands. As an example, such a situation occurs rather regularly in spring and summertime in the Alps, at the Lakes Valley exit. The latter is shaped as an elevated saddle, opening on the steep western side of the Adige Valley, north of the city of Trento. Here the local wind known as ‘*Ora del Garda*’ – a coupled lake and valley breeze, originating from Lake Garda shores, and then channelling along the nearby Lakes Valley – outflows through the saddle, and suddenly jumps onto

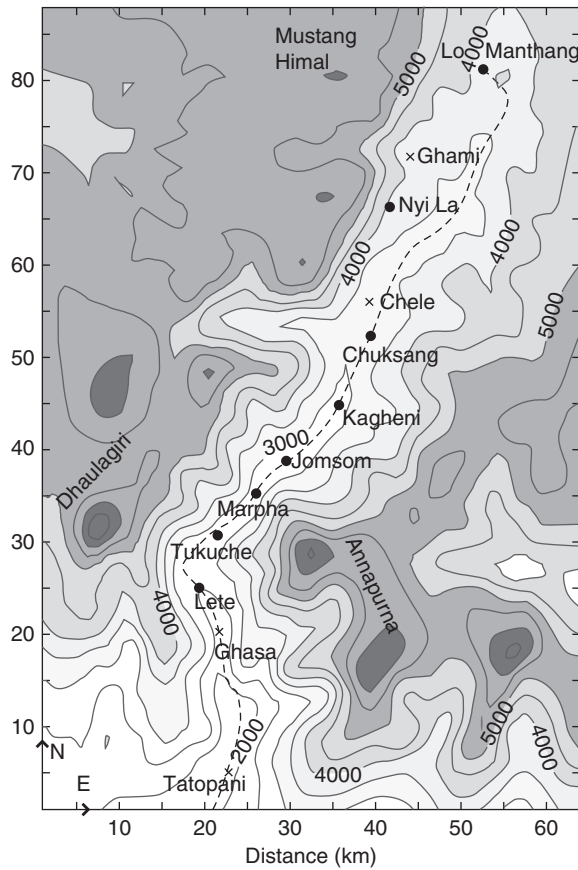


Figure 3 Map of the Kali Gandaki Valley in Nepal. Dots indicate the main villages and towns. Solid lines are height contours (labeled in m AMSL). Horizontal distances are in km. Reproduced with permission from Egger, J. et al., 2000. Diurnal winds in the Himalayan Kali Gandaki Valley. Part I: Observations. Monthly Weather Review 128, 1106–1122.

the Adige Valley floor: interacting with the local up-valley wind (as suggested in Figure 5), this outflow results in a strong and gusty wind, blowing there throughout the afternoon.

The typical diurnal cycle of valley winds follows the natural sequence of daytime heating and nighttime cooling at the

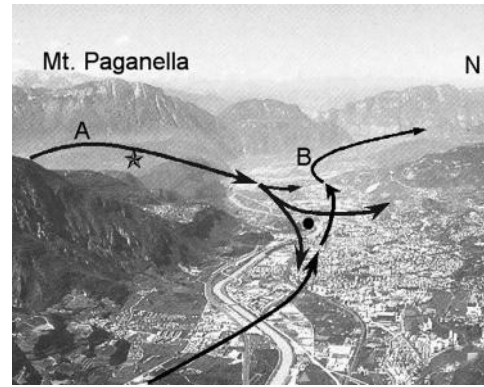


Figure 5 The area north of the city of Trento where the 'Ora del Garda' wind, blowing from the Lakes Valley (A) through an elevated saddle (★), interacts with the up-valley wind blowing in the Adige Valley from south (B). Reproduced with permission of the American Meteorological Society from de Franceschi, et al., 2002. Proceedings of the 12th Conference on Mountain Meteorology, Park City, Utah.

Earth's surface. The basic components of the valley winds are schematized in Figure 6, along with the timing of their diurnal development. Black arrows indicate along-valley flows, whereas white arrows denote all the airflow components blowing on valley cross sections. Notice that here, for the sake of simplicity, the valley is represented by a very idealized topography, and heating is assumed to occur symmetrically on both sides of the valley. This may only occur under very particular circumstances, e.g., when the valley axis lies on the same vertical plane as the sun trajectory in the sky (such as along the equator at equinoxes), and terrain properties are symmetrically distributed on the two sides of the valley, so as to make surface energy budgets symmetrically identical as well. Also forcing from upper winds should be symmetrically acting on the two sides. These simplifying assumptions give an idea of how many factors may variously affect the spatial distribution of key variables associated with valley wind development, such as temperature, wind strength, and direction.

After sunrise (Panel A) incoming solar radiation increasingly hits the ground. In a valley, direct radiation typically starts

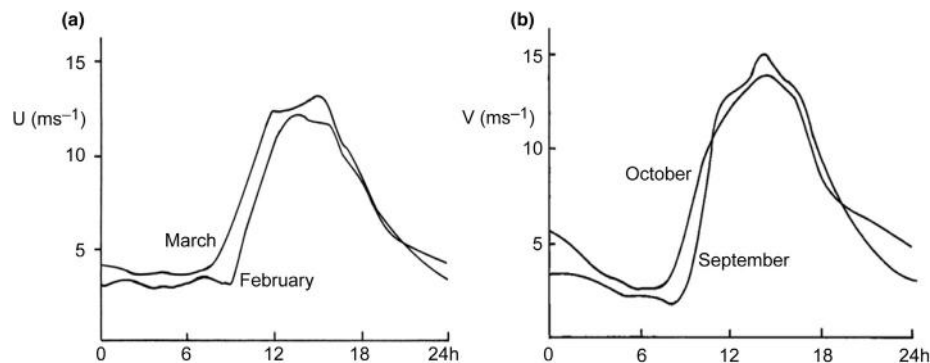


Figure 4 Monthly mean values of the hourly mean wind speed observed in the Kali Gandaki Valley in Kagbeni (2900 m AMSL) in (a) February–March and (b) September–October 1990. Measurements were taken at a height of 9 m AGL. Notice the typical diurnal cycle, with peak wind speeds in the early afternoon. Reproduced with permission from (a) Egger, J. et al., 2002. Diurnal winds in the Himalayan Kali Gandaki Valley. Part I: Observations. Monthly Weather Review 128(4), 1106–1122 and (b) Egger, J. et al., 2000. Diurnal winds in the Himalayan Kali Gandaki Valley. Part III: piloted aircraft soundings. Monthly Weather Review 130, 2042–2058.

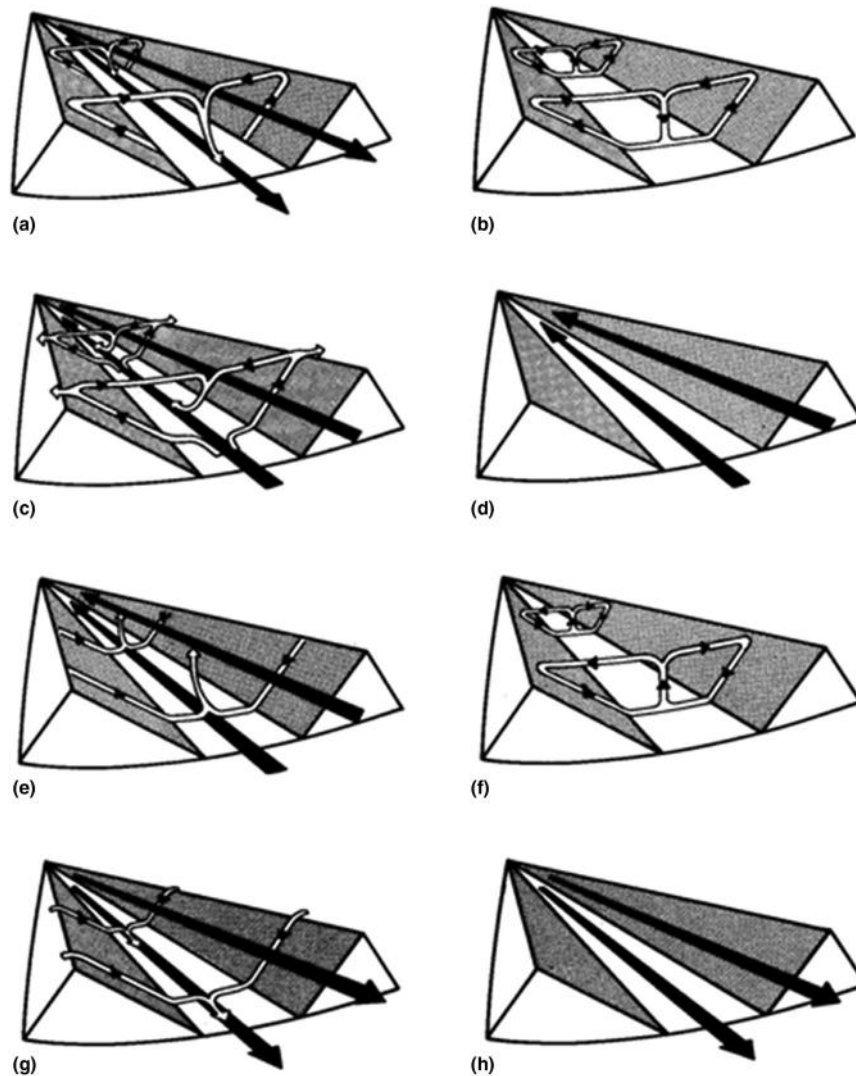


Figure 6 Diurnal cycle of valley winds: see text for explanation. Reproduced with permission from Defant, F., 1949. *Zur Theorie der Hangwinde, nebst Bemerkungen zur Theorie der Berg- und Talwinde* [A theory of slope winds, along with remarks on the theory of mountain winds and valley winds]. *Archiv für Meteorologie, Geophys. Bioclimatologie*, Ser. A 1, 421–450. [Theoretical and Applied Climatology] [English translation: Whiteman, C.D., Dreiseitl, E., 1984. *Alpine meteorology*, Translations of classic contributions by Wagner A., Ekhardt, E., Defant, F., PNL-5141/ASCOT-84-3. Pacific Northwest Laboratory, Richland, Washington, 121pp. <http://www.osti.gov/bridge/servlets/purl/6665518/6665518.pdf>.]

lighting the crests first, and then gradually reaches the valley floor. As a consequence of the energy partitioning resulting from the surface budget (which will be addressed in the following section), part of the energy input associated with solar radiation is converted into sensible heat flux, which is increasingly supplied to the air layers adjacent to the ground. On the sidewalls this heating progressively reduces the existing temperature deficit. This deficit was produced by nighttime energy loss from radiative ground cooling, and favored the development of nocturnal drainage flows down the slopes. On the contrary, as soon as the morning heating proceeds, air parcels adjacent to the sunlit slopes become warmer, and positively buoyant. As a result they promote organized upward motions conveying warmer air all along the sloping sidewalls, from the valley floor all the way up to the crest level. Here air parcels inertially rise

even above the crest level, where they cannot get heat from the ground any more. Then they either are buoyant enough to rise further and form thermals, possibly producing cumulus clouds, or, after eventually overshooting their neutral-buoyancy level, remain trapped into a horizontal return flow converging toward the valley center. Indeed this return flow is part of a compensating circulation bringing air parcels downward to the floor level, where air is diverted toward the sidewall feet to feed slope flows. This circulation contributes, along with heat locally supplied by energy partitioning at the valley floor, to remove the nocturnal stable layer on the floor, and thus to break up the ground-based nocturnal inversion, which was built up by the nighttime drainage of cold air down the sidewalls. Moreover, subsidence at the valley core produces an adiabatic compression of air parcels, and hence further contributes to raising air

temperature within the valley atmosphere. As a consequence, the overall temperature drop, and the associated density excess, produced by nocturnal cooling, are gradually overcome during the morning, as heat is increasingly supplied to the valley atmosphere. Similar processes are produced by incoming radiation over the adjacent plain, where daytime heating removes the ground-based nocturnal inversion and raises surface temperatures high enough to eventually trigger the development of atmospheric convection. However, the heating processes in the valley are quite different from those occurring on the plain. In particular, the organized circulation activated by sidewall slopes on each valley cross section is more effective in promoting a vertical heat transfer than random convective motions developing over the adjacent plain. As a consequence, at any height air temperatures throughout the valley volume are generally higher than over the plain. This contrast, increasing during the course of the day, produces a progressive reversal of the along-valley pressure gradient, and hence a reversal of the winds from down- to up-valley (Panel B). Indeed, the higher the sun over the horizon, the larger the sunlit portions of the sidewalls, and the stronger the incoming radiation per unit surface area. Accordingly, the intensity of along-valley winds typically reaches a maximum by mid-afternoon (Panel C). The strength of up-valley winds may even overwhelm slope winds, which may get increasingly embedded in them (Panel D). As soon as solar radiation starts declining, slope winds progressively weaken as well: indeed the energy input from declining solar radiation can no longer compensate the longwave net radiative loss. As a consequence, the sensible heat flux to the air layers adjacent to the slopes becomes negative, and a cooling phase begins, resulting in a progressive reversal of slope winds (Panel E). The slope winds then act as drainage flows, bringing colder air along the sidewalls to the valley floor, and thus contributing to an overall cooling of the valley atmosphere. The up-valley wind inertially continues blowing for some time after sunset, and thus maintains a residual mechanical turbulence production in the surface layer, which favors turbulent mixing and thus sustains sensible heat fluxes. Such a circulation promotes a rather efficient overall cooling of the valley atmosphere, whose lower layers rapidly become colder than the layers at equal heights above the adjacent plain. This contrast leads to a gradual reversal of the daytime plain–valley horizontal pressure gradient, and accordingly to a weakening of the up-valley wind (Panel F), and then, finally, to the onset of a down-valley wind (Panel G). This wind typically blows stronger and stronger, and in increasingly deeper layers, during the course of the night (Panel H). The resulting nocturnal circulation is maintained by the combination of cooling processes, favored by outgoing radiation and advection of colder air, which keep operating till dawn, when the cycle starts again from the beginning (Panel A).

As a result of these alternating slope and valley winds, at any point on a valley sidewall the surface wind vector displays a typical rotation during the day, as shown in Figure 7.

As implied in the above description, as far as a valley shape and layout are similar to the idealized topography reproduced in Figure 6, the main stream of the valley wind will ideally follow the along-valley axis. However, a variety of modifications in space and time – such as accelerations and decelerations, deviations, secondary circulations, bifurcations, and

confluences – may occur in connection with changes in topographic factors – such as valley cross-section area, or shape, or valley curvatures. Further modifications may derive from varied surface properties, or from interactions with concurring winds flowing along tributary valleys, as well as from changes in the external forcings, such as the varying intensity of solar radiation at the ground in space and time, or the changing action of upper winds. A full account of these effects cannot be provided here: the interested reader will find further details and literature references to various case studies in the mentioned chapter by Zardi and Whiteman (2013).

It is interesting to observe that the development of valley winds displays many analogies with the development of coastal breezes (see **Mountain Meteorology: Land and Sea Breezes**). The latter are daily periodic wind systems blowing normal to the shorelines of seas or large lakes, under the same weather situations that favor the development of valley winds. Coastal flows are produced by horizontal pressure gradients associated with the different vertical thermal structures developing over the land and over the water surface. As these contrasts occur, with opposite results, both under the daytime heating, and under the nocturnal cooling phase, they produce coastal breezes in much the same way as valley winds are produced by valley–plain contrasts, reversing twice per day.

In particular, the overall structure of both wind systems may be viewed as an organized, large-scale solenoidal vortex, rotating in such a way that its lower branch blows from the plain to the valley interior – or, respectively, from above the sea toward the onshore region – after sunrise, and gradually reverses after sunset. Indeed such circulation, as well as the associated flow vorticity, is produced by baroclinic effects, i.e., from the misalignment of pressure and density gradients associated with the above mentioned thermal contrasts (see **Air Sea Interactions: Freshwater Flux; Momentum, Heat, and Vapor Fluxes; Sea Surface Temperature; Surface Waves. Dynamical Meteorology: Vorticity**).

Physical Processes Controlling Valley Winds

In the previous section, we outlined the basic features of a typical diurnal valley wind system, providing an intuitive insight in the combination of the physical factors concurring in its development. In the present section, we will explore in more detail some of the above processes, and the mechanisms involving them, along with the factors controlling their occurrence, intensity, and duration.

Radiation and Energy Budgets

In general, thermally forced flows are driven by buoyancy effects and pressure gradients arising from air density variations associated with heating or cooling of air parcels. On the other hand, local changes in air temperature result from a combination of factors, as clearly summarized in the tendency equation for potential temperature at any point in the atmosphere:

$$\frac{\partial}{\partial t}(c_p \bar{\theta}) = -\nabla \cdot \mathbf{A} - \nabla \cdot \mathbf{H} - \frac{\bar{\theta}}{\bar{T}}(\nabla \cdot \mathbf{R} + \dot{Q}) \quad [1]$$

Here c_p is the air specific heat at constant pressure, while $\bar{\rho}$, $\bar{T}$, and $\bar{\theta}$ are the average values (with respect to turbulent

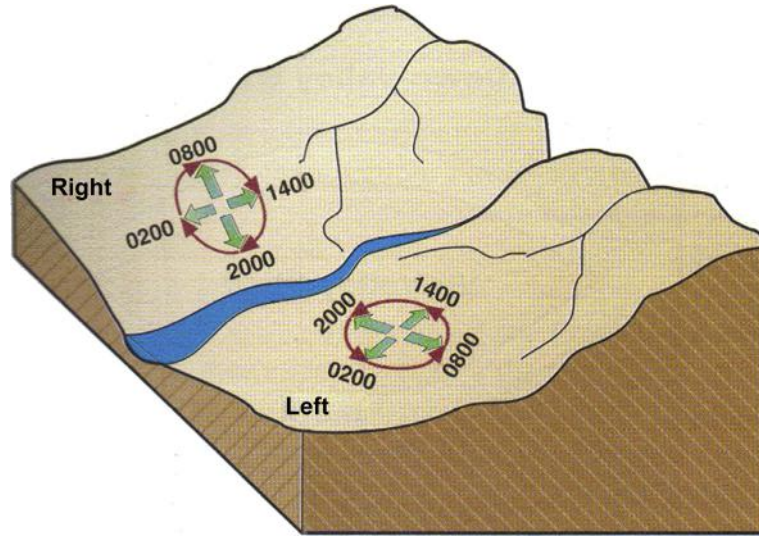


Figure 7 Diurnal rotation of the wind velocity vector resulting from the combination of slope and valley winds. Green arrows denote the velocity vector at the indicated hours, whereas red circle arrows represent the wind rotation. Reproduced with permission from Whiteman, C.D., 2000. *Mountain Meteorology: Fundamentals and Applications*. Oxford University Press, New York, 355pp.

fluctuations) of air density, temperature, and potential temperature, respectively. The term $A = c_p \bar{\rho} \bar{u} \bar{\theta}$ is the sensible heat flux associated with advection by the mean wind velocity $\bar{u}$, $H = c_p \bar{\rho} \overline{u' \theta'}$ is the turbulent sensible heat flux (produced by the coupling between turbulent fluctuations of wind velocity u' and potential temperature θ'). (Following Reynolds decomposition for turbulent flows, primed variables indicate turbulent fluctuations with respect to the mean values, indicated by an overbar.) R is the radiative heat flux, including both the shortwave and the longwave component. Finally, $\bar{Q}$ includes the rates of heating (or cooling) associated with conversion of water vapor into or from liquid water or ice.

Notice that both advective and turbulent heat fluxes imply some air motions, i.e., respectively, a mean wind velocity $\bar{u}$ and turbulent velocity fluctuations u' . However, in the ideal situations for thermally driven flows, i.e., dry atmosphere under clear sky and calm synoptic winds, little or no advection is produced by any large scale flows, and little turbulence can be produced mechanically, i.e., through the typical energy cascade from large to small eddies, via the coupling between mean wind shear and turbulent momentum fluxes. Rather, advection and turbulence are mainly associated with air motions originated by the combination of gravity with unbalanced density gradients, produced by heating or cooling. However, these heating and cooling are not primarily induced by the absorption or emission of radiation (the term $\nabla \cdot R$ in eqn [1]), which is usually negligible within the layers involved in valley winds systems. Rather they come from processes controlling energy partitioning at the surface. Typically, the surface energy budget may be expressed in terms of the components normal to the surface (labeled with 's') of the above mentioned fluxes, i.e., radiation flux R_s , and sensible heat flux H_s , along with latent heat flux L_s , and heat flux exchanged with subsurface layers, i.e., the ground heat flux G_s , as

$$R_s + H_s + L_s + G_s = 0 \quad [2]$$

In eqn [2] all fluxes toward the surface (both from the atmosphere and from ground) are positive, and all fluxes away from the surface are negative. The turbulent latent heat flux L_s is associated with evaporation (sublimation) from, or condensation (deposition) onto, the surface. It is produced by the coupling between turbulent fluctuations of the wind velocity component normal to the surface w' and specific humidity q' , i.e., $L_s = (\ell_i \bar{\rho} \overline{w' q'})_s$, ℓ_i being the appropriate latent heat (i.e., the latent heat of vaporization ℓ_v or the latent heat of sublimation ℓ_s depending on the situation) (see **Dynamical Meteorology: Acoustic Waves**). Ground heat flux depends on heat conduction in the subsurface ground (soil, rock, etc.).

The (net all-wave) surface radiation flux R_s may be decomposed into four components, namely

$$R_s = K\uparrow + K\downarrow + L\uparrow + L\downarrow \quad [3]$$

with $K\downarrow$ being the incoming shortwave radiation, including both direct and diffuse solar radiation, $K\uparrow$ the outgoing shortwave radiation (i.e., the fraction of incoming shortwave radiation reflected by the surface), $L\uparrow$ the outgoing longwave radiation emitted by the surface, and $L\downarrow$ the incoming longwave radiation emitted downward by the atmosphere (air gases, water vapor, and clouds). Such a budget occurs at any point of the ground surface. However, over complex terrain it typically displays a higher variability in space and time, following topographic effects on incoming solar radiation. A remarkable example of a typical diurnal cycle of incoming shortwave radiation in a valley is shown in Figure 8, where mean diurnal cycles are shown from measurements performed at different points in the Riviera Valley in the Swiss Alps (Figure 9).

Therefore the surface energy budget is the main factor controlling thermally driven circulations, promoting both the turbulent heat fluxes and the advective motions required to develop the whole system of valley winds. Accordingly, factors controlling this budget also affect the development of valley winds. For instance, clear skies and clean air favor

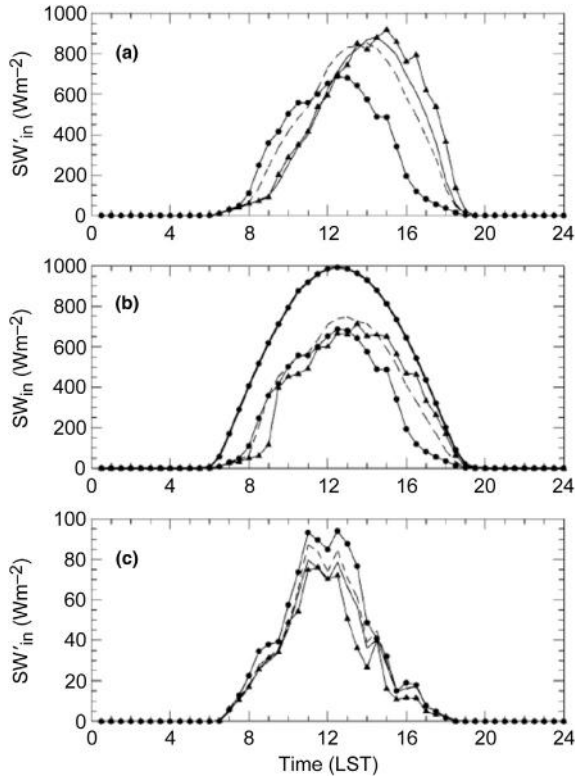


Figure 8 Diurnal cycles of incoming short-wave radiation at different measurements sites in the Riviera Valley, Switzerland (see Figure 9 for location), namely CVF (filled dots), ESM (dashed line), ESF2 (full line), and ESR (triangles). Panels (a) and (b) represent average values during valley wind days on slope-parallel and horizontal surfaces, respectively, while panel (c) refers to overcast days. The bold curve in (b) denotes the mean diurnal cycle of extraterrestrial irradiance on a horizontal surface at site CVF. Note the different vertical axes. CVF, Centered in valley floor; ESM, Eastern slope meadow; ESF2, Eastern slope forest; ESR, Eastern slope ridge. Reproduced with permission from Matzinger, N., et al., 2003. Surface radiation budget in an Alpine valley. *Quarterly Journal of the Royal Meteorological Society* 129, 877–895.

stronger energy exchanges, allowing elevated values of incoming solar radiation at surface level during daytime, as well as outgoing radiation emitted from the ground during nighttime. Hence such conditions favor both daytime heating and nighttime cooling of air layers at the earth surface. However, this heating or cooling depends on the energy budgets (eqns [2] and [3]), whose terms are controlled by soil properties – such as heat capacity, thermal conductivity, and moisture content – as well as by surface properties, such as reflectivity, emissivity, and topographic features. As a consequence, the strongest diurnal wind systems are typically found in elevated and dry environments: elevation provides a better exposure to solar radiation during the day, and strong longwave radiative loss at night, whereas dryness typically reduces shortwave attenuation from air turbidity and cloud cover, as well as the latent heat flux involved in water evaporation or condensation, thus allowing more energy excess (or deficit) to be available for more intense heating (or cooling) of the atmosphere.

Slope Winds

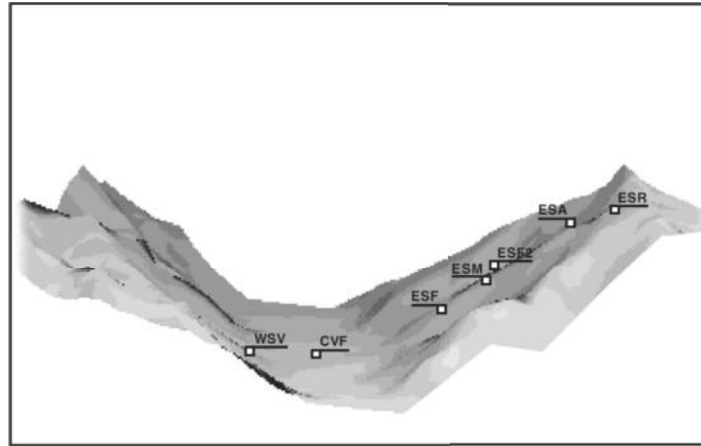
The picture of the diurnal cycle provided in Section Introduction suggests that slope flows along the valley sidewalls play a key role in the development of valley winds, as they are the earliest drivers of the whole valley wind system. Indeed, by reacting quickly to changes in the surface energy budget, following the radiation cycle at the ground, slope flows effectively promote heat and mass advection along the slopes, thus producing exchanges both in the vertical and in the cross-valley directions. As these exchanges are determined by the topographic features of the valley cross section, they occur in a much more organized way than the analogue vertical exchange processes over flat terrain: here during daytime thermal convection feeds stochastically generated thermals, resulting in an increasingly deep convective boundary layer (CBL), whereas nocturnal cooling leads to a largely uniform, increasingly stable boundary layer.

To understand the basic dynamics implied in slope flows, it is worth recalling here a prototypal model proposed by Prandtl (1942). (Strictly speaking, the original Prandtl theory accounts for a momentum flux dominated by molecular viscosity and a heat flux determined by heat conduction, i.e., for a laminar atmospheric flow. Here we propose a straightforward extension of the theory to the turbulent case, which is more likely to occur in a real atmosphere. The mathematical expression of the solution remain in fact the same, if molecular viscosity and heat diffusivity are replaced by their respective turbulent analogs, provided that they may be assumed to be constant.) Indeed let us consider an infinite plane surface tilted by an angle α over the horizontal (Figure 10), and an overlying stably stratified atmosphere, initially at rest and displaying a constant vertical potential temperature lapse rate γ . Then imagine perturbing this static equilibrium situation by imposing a prescribed constant value ΔT_s for the temperature anomaly at any point of the surface (or, equivalently, a prescribed value H_s of the sensible heat flux). Let us adopt for convenience a frame of reference, where s is the coordinate along the slope (positive up-slope) and n is the coordinate normal to the slope (positive from the surface toward the atmosphere). As the surface temperature anomaly ΔT_s is invariant with s , let us seek for a solution that does not depend on the along-slope coordinate, but only depends on n . Mass continuity then implies that the motion be perfectly parallel to the slope, i.e., there will be only one velocity component along s (positive up-slope). Moreover, let us seek for a steady-state solution. Accordingly the equations for conservation of momentum and energy at leading order reduce to:

$$0 = \beta g \sin \alpha \bar{\theta} - \frac{\partial}{\partial n} \overline{u' w'} \quad [4]$$

$$\gamma \sin \alpha \bar{u} = - \frac{\partial}{\partial n} \overline{w' \theta'} \quad [5]$$

Equation [4] represents the local balance between two forces (per unit mass). The first term represents the buoyancy force, resulting from the combination of gravity (g), topography (α), and the relative density anomaly, represented by $\beta \bar{\theta}$, where β is the coefficient of thermal expansion and $\bar{\theta}$ is the potential temperature anomaly with respect to the



Site label	Site characteristics	Height m AMSL	Inclination/ exposure	Horizon angle	
				West	East
CVF	Centered in the valley floor	250	0°	25°	3°
WSV	Western slope, vineyard	340	27°	31°	15°
ESF	Eastern slope, forest	760	31°	9°	21°
ESM	Eastern slope, meadow	1060	20°	5°	21°
ESF2	Eastern slope, forest	1030	35°	8°	23°
ESA	Eastern slope, alpine meadow (pasture)	1860	0°	3°	22°
ESR	Eastern slope, ridge	2110	40°	0°	35°

Figure 9 Cross section through the Riviera Valley, Switzerland. Labels indicate sites where radiation measurements were made under the MAP Riviera project in 1999. Characteristics of each site are listed in the table. Reproduced with permission from Matzinger, N., et al. 2003. Surface radiation budget in an Alpine valley. *Quarterly Journal of the Royal Meteorological Society* 129, 877–895.

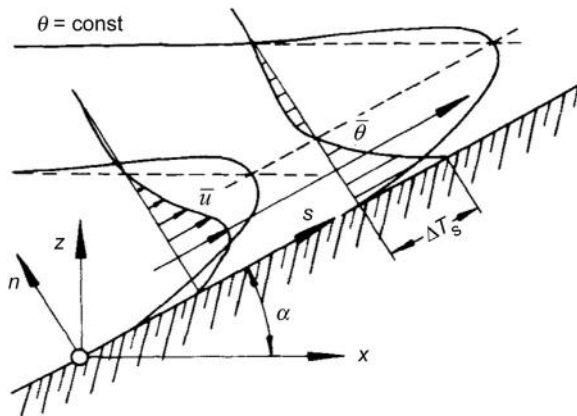


Figure 10 Profiles of velocity ($\bar{u}$) and potential temperature anomaly for up-slope flow ($\bar{\theta}$) according to Prandtl's (1942) model (see text for explanation). Adapted with permission from Schumann, 1990. Large-eddy simulation of the up-slope boundary layer. *Quarterly Journal of the Royal Meteorological Society* 116, 637–670.

unperturbed situation. This term produces a positive (up-slope) buoyancy force on lighter parcels (i.e., parcels displaying a larger potential temperature than at the same level in the unperturbed atmosphere) and vice versa. The second term represents friction effects, produced by turbulent momentum flux in the direction normal to the slope. Friction always counteracts air motions, either up-slope or down-slope, and a steady motion requires that buoyancy and friction permanently balance each other.

Equation [5] represents a local energy balance between along-slope heat advection and sensible heat flux normal to the slope. For positive $\bar{u}$ the first term reproduces an up-slope advection of potentially colder air from lower layers, thus contributing to local cooling. The second term is positive when heat flows from the surface to the atmosphere, which occurs when the surface displays a higher temperature than the overlying atmosphere layers. A similar reasoning applies, simply reversing the signs, for down-slope flows associated with surface cooling. The balance between the two requires that the wind blows up-slope on a surface which is heating the atmosphere, and down-slope when it is cooling.

To make progress toward a solution of eqns [4] and [5] we need to specify the structure of turbulent fluxes, or at least to stipulate a relationship between them and the unknown variables $\bar{u}$ and $\bar{\theta}$. One way is to postulate that turbulent transport processes behave in a similar way to their molecular analogs, such as viscous friction or heat conduction, where fluxes of momentum or heat are proportional to velocity and temperature gradients, respectively. So let us set:

$$\overline{u'w'} = -K_m \frac{\partial \bar{u}}{\partial n} \quad [6]$$

$$\overline{w'\theta'} = -K_h \frac{\partial \bar{\theta}}{\partial n} \quad [7]$$

where K_m and K_h are the eddy viscosity and eddy heat diffusivity, respectively. Assuming further that these are constant values throughout the slope flow layer, substituting in eqns [4] and [5] one gets two coupled linear equations that admit the solution found by Prandtl (1942):

$$\bar{u} = U \exp(-n/\ell) \sin(n/\ell) \quad [8]$$

$$\bar{\theta} = \Delta T_s \exp(-n/\ell) \cos(n/\ell) \quad [9]$$

where $U = N\gamma^{-1}Pr_T^{-1/2}\Delta T_s$, with $N = (\beta g\gamma)^{1/2}$ being the Brunt-Vaisala frequency of the unperturbed atmosphere and $Pr_T = K_m/K_h$ the turbulent Prandtl number.

The length scale ℓ is provided by:

$$\ell = \left(\frac{4K_h K_m}{N^2 \sin^2 \alpha} \right)^{1/4} \quad [10]$$

An example of the slope-normal profiles provided by this solution is shown in Figure 10. The potential temperature anomaly is maximum at the surface, and then its magnitude decreases exponentially with distance from the surface, reversing sign at the nodes of the cosine function. This means that some upper counterflow accompanies the development of the main flow adjacent to the surface. Indeed, as a result of friction, due to the thermal upward thrust, layers which have not themselves been heated are set in motion, so after they have risen to a new position they are colder than the particles which were there, when the stratification of the air was undisturbed. The wind strength scale U is proportional to the surface temperature anomaly ΔT_s : therefore a large, positive ΔT_s produces a strong up-slope flow, and vice versa. Figure 10 also shows how potential-isothermal lines are modified by the

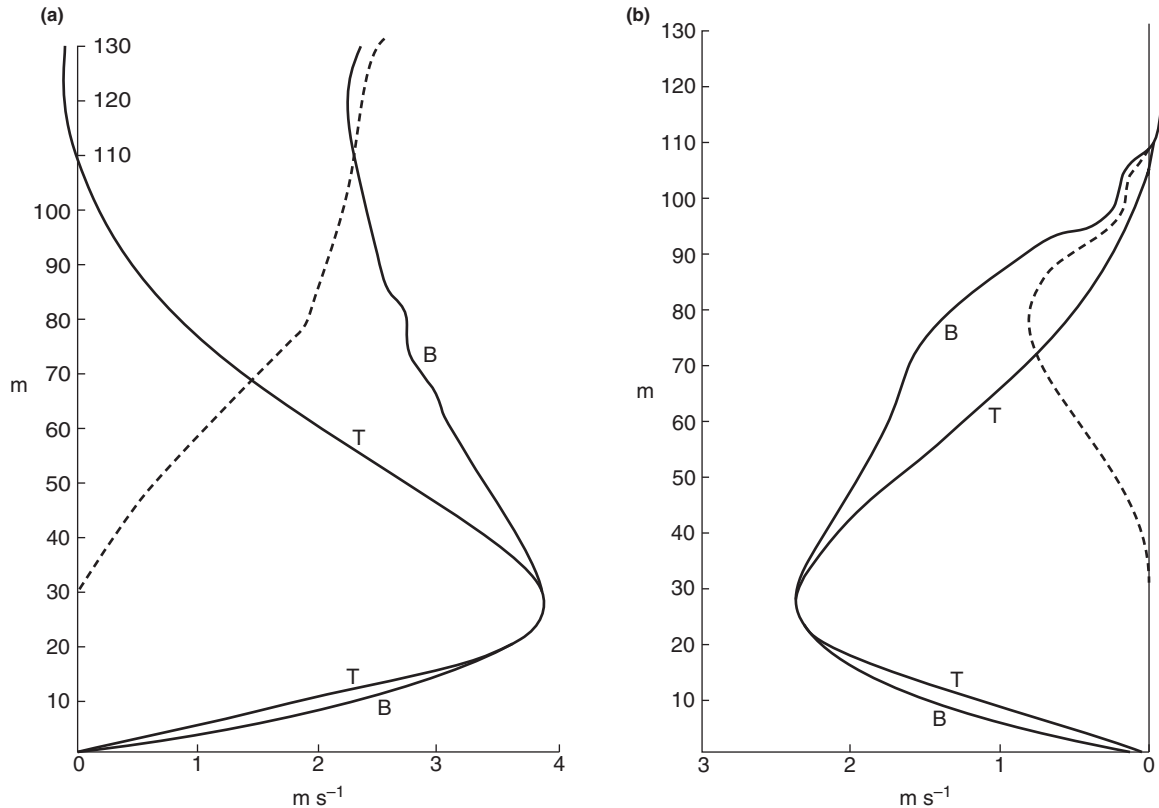


Figure 11 Velocity profiles from observations (curves B) performed with pilot balloons on the Nordkette (near Innsbruck, Austria) and theoretical predictions (curves T) based on Prandtl's (1942) model for up-slope (a) and down-slope (b) winds. Dashed lines indicate the difference between observed and theoretically calculated wind speed above the maximum. Reproduced with permission from Barry, R.G., 2008. *Mountain Weather and Climate*, third ed. Cambridge University Press, 506pp; Adapted from Defant, F., 1949. Zur Theorie der Hangwinde, nebst Bemerkungen zur Theorie der Berg- und Talwinde [A theory of slope winds, along with remarks on the theory of mountain winds and valley winds]. *Archiv für Meteorologie, Geophys. Bioclimatologie*, Ser. A 1, 421–450. [Theoretical and Applied Climatology] [English translation: Whiteman, C.D., Dreiseitl, E., 1984. *Alpine meteorology*, Translations of classic contributions by Wagner A., Ekhart, E., Defant, F., PNL-5141/ASCOT-84-3. Pacific Northwest Laboratory, Richland, Washington, 121pp.] <http://www.osti.gov/bridge/servlets/purl/6665518/6665518.pdf>.

slope flow: far from the surface the isolines are perfectly horizontal and equally spaced, reflecting the constant vertical lapse rate of potential temperature in the unperturbed atmosphere. Closer to the surface, isolines bend downward for a heated surface, associated with up-slope flow, and upward for a cooling surface, associated with down-slope flows.

The above theoretical model explains many aspects involved in the development of slope flows, and despite many idealizing assumptions (infinite plane slope, s -invariance) and approximations (linearization, simplified representation of turbulence), it captures the essentials of real slope winds, as shown by measurements taken on real slopes reported in Figure 11. The main discrepancies appear in up-slope flows, where measurements suggest a deeper layer characterized by upward velocity than estimated by the model profile. This has probably to do with the stronger upward convective motions associated with thermally generated turbulence. This idea is also supported by high-resolution numerical simulations. Figure 12 shows slope-normal profiles of along-slope wind speed and potential temperature anomaly from large eddy simulations of steady turbulent up-slope flow over an infinite slope for various slope angles ($\alpha = 2, 4, 7, 10$, and 30°). Both velocity and temperature profiles display some similarities with Prandtl's solutions. However, for angles $\alpha \leq 10^\circ$ a well-mixed layer develops, which makes potential temperature uniform for a significant depth above the surface layer, and also makes the velocity profiles rather flat.

Further results for realistic cases of down-slope flows are not reported here, as they are the subject of the article **Mountain Meteorology: Katabatic Winds** in this Encyclopedia.

Turbulence

As envisaged in the previous section, turbulent motions play a key role in enhancing momentum and energy exchanges in thermally driven flows. As discussed before, the main forcing for valley winds comes from energy and momentum exchanges at the surface, where flows are inherently turbulent, and this implies that these exchanges are stronger where turbulence is more intense. A measure of the turbulence intensity is provided by the turbulent kinetic energy (TKE). TKE is defined as the kinetic energy per unit mass associated with the turbulent velocity fluctuations, namely $TKE = \overline{u'^2} + \overline{v'^2} + \overline{w'^2}$, where u , v , and w are the velocity components with respect to some Cartesian coordinate system and an overbar denotes the average with respect to turbulence. The two main mechanisms controlling the production of TKE are the so-called *mechanical production*, arising from the coupling of momentum flux with the gradient of the mean flow, and *buoyancy production*, originated by the coupling of buoyancy flux with background stratification. While the former is always positive, i.e., always contributes to increasing TKE, the latter is positive under unstable conditions, whereas it acts to suppress turbulence under stable situations. Figure 13 shows the spatial distribution of (subgrid) TKE obtained with large eddy simulations on a valley cross section of two idealized valleys – a wider and a narrower one – at two different times, i.e., at 13 p.m. (upper panels) and at 17 p.m. (lower panels). During daytime surface heating continuously feeds buoyancy production of TKE, which is then redistributed either by larger thermals,

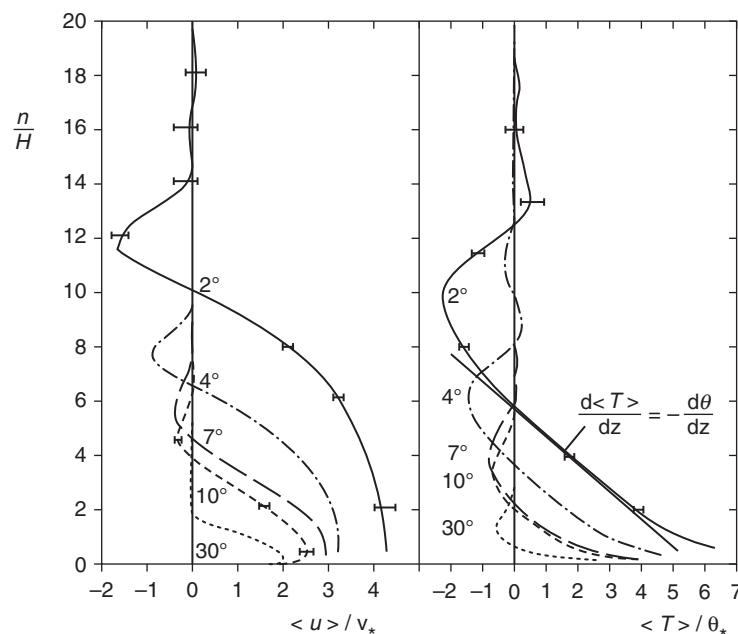


Figure 12 Profiles of mean velocity (left) and temperature (right) from Large-Eddy Simulations of turbulent up-slope flows for inclination angles $\alpha = 2, 4, 7, 10$, and 30° . Mean wind speed $\langle u \rangle$, temperature $\langle T \rangle$, and height above the slope n are normalized by the following scales: $v_* = (\beta g H_s / N)^{1/2}$, $H = v_* / N$, and $\theta_* = H_s / v_*$. With the setting adopted for the simulations ($\gamma = 0.003 \text{ K m}^{-1}$, $\beta = 1/300 \text{ K}^{-1}$, $g = 10 \text{ m s}^{-2}$, and $H_s = 0.1 \text{ K m s}^{-1}$) the scale values result in $v_* = 0.58 \text{ m s}^{-1}$, $H = 58 \text{ m}$, and $\theta_* = 0.17 \text{ K}$. The error bars signify standard deviations. The thin line in the panel on the right corresponds to the well-mixed situation. Reproduced with permission from Schumann, 1990. Large-eddy simulation of the up-slope boundary layer. Quarterly Journal of the Royal Meteorological Society 116, 637–670.

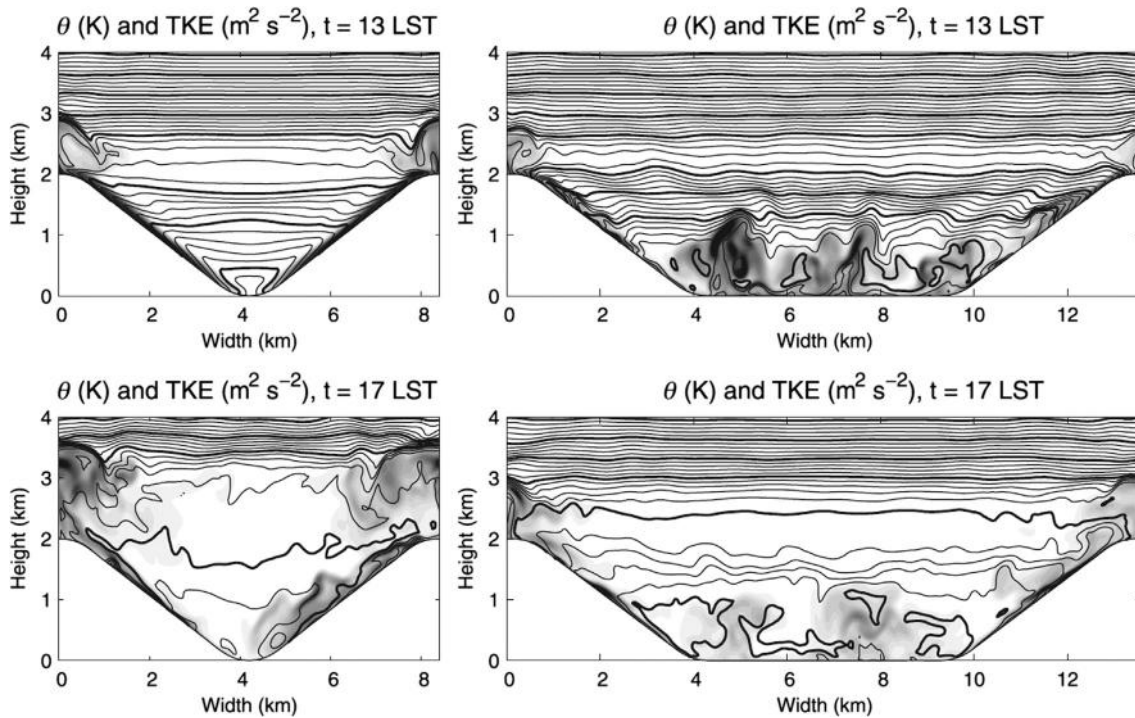


Figure 13 Consecutive mid-domain cross sections of θ (isolines every 0.2 K, thick every 1 K; $\theta = 312$ K at top) and subgrid-scale TKE (shading) in the (left) narrow and (right) wide valley runs. Subgrid-scale TKE, approximately one order of magnitude smaller than the explicitly resolved fraction, is used merely to suggest the position of the dominant turbulent eddies. TKE, Turbulent kinetic energy. Reproduced with permission from Serafin, S., Zardi, D., 2010. Daytime Heat Transfer Processes Related to Slope Flows and Turbulent Convection in an Idealized Mountain Valley. *Journal of the Atmospheric Sciences* 67, 3739–3756.

especially above the valley floor of a wide valley, or by the combined effect of slope flows and the resulting cross-valley circulation. These two processes realize a sort of competition between a more organized flow, marked by orographic features of the terrain, and randomly generated thermals on the valley floor, more similar to those typically forming on openly flat terrain. As a consequence, in a narrow valley slope

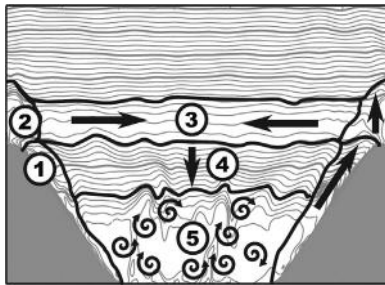


Figure 14 An idealized sketch of the cross-valley circulation induced by up-slope flows (1) ending in thermal plumes (2) at the crest tops, and in horizontal motions (3) at mountaintop level. Adiabatic subsidence (4) stabilizes the valley core, and suppresses the development of a convective boundary layer (5) on the valley floor. Arrows and whirls suggest the main features of the flow field. Reproduced with permission from Serafin, S., Zardi, D., 2010. Daytime Heat Transfer Processes Related to Slope Flows and Turbulent Convection in an Idealized Mountain Valley. *Journal of the Atmospheric Sciences* 67, 3739–3756.

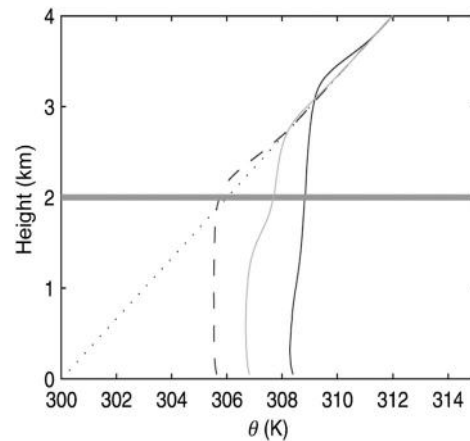


Figure 15 Comparison of potential temperature profiles over a plain (dashed black line, about 8 h after sunrise), wide valley (gray line, 7 h after sunrise), and narrow valley (continuous black line, 7 h after sunrise). The cross sections of the two valleys are the same as in Figure 13. The different times correspond to an equal energy input for all cases. The horizontal gray line indicates the level of the top of the valley sidewall. Reproduced with permission from Serafin, S., Zardi, D., 2011. Daytime development of the boundary layer over a plain and in a valley under fair weather conditions: a comparison by means of idealized numerical simulations. *Journal of Atmospheric Science* 68, 2128–2141.

flows are expected to be stronger, and thermal plumes on the sidewall crests more developed, than in a wider one.

Vertical Thermal Structure and Horizontal Pressure Gradients

The main mechanism driving valley winds originates from pressure unbalances arising at various levels between neighboring regions – namely within a valley and above an adjacent plain – as a consequence of different processes producing air heating or cooling at different levels in the two regions. The processes leading to these contrasts can be better understood by considering again an ideal valley–plain system, consisting of a simple valley with an along-valley invariant, symmetric cross section, and a horizontal floor, facing an open plain. Vertical accelerations associated with either upward or downward motions characterizing valley winds are usually small. Therefore, the vertical pressure distribution is essentially governed by hydrostatic balance. Hence horizontal pressure gradients, driving along valley flows, arise just from changes in the vertical temperature distribution occurring during the diurnal cycle in the valley atmosphere and in the adjacent plains.

To point out the relationship between the vertical thermal structure and the along-valley pressure gradient, consider a valley with a perfectly horizontal floor at the same level as the adjacent plain. Let us concentrate on the along-valley pressure distribution considering the values that pressure assumes over a vertical plane based on the along-valley axis at the valley floor. Also, let us focus on the daytime phase, as similar reasoning may be easily extended by analogy to the nighttime phase.

The mechanisms leading to the heating of the valley atmosphere over a vertical cross section are schematized in Figure 14. Up-slope flows along the sidewalls convey heated air up to the crest level, where it either produces vertically ascending thermal plumes, or gets involved in the upper return flow converging toward the valley center and feeding the subsidence motion that compensates the removal of heated air operated by up-slope flows at lower levels. These mechanisms result in a vertical thermal structure, which is remarkably different from the classical CBL profile occurring under the same weather situation and solar radiation over a plain. Figures 15 and 16 offer a comparison between the two situations, and some hints about the influence of the valley width. The vertical structure of the CBL in

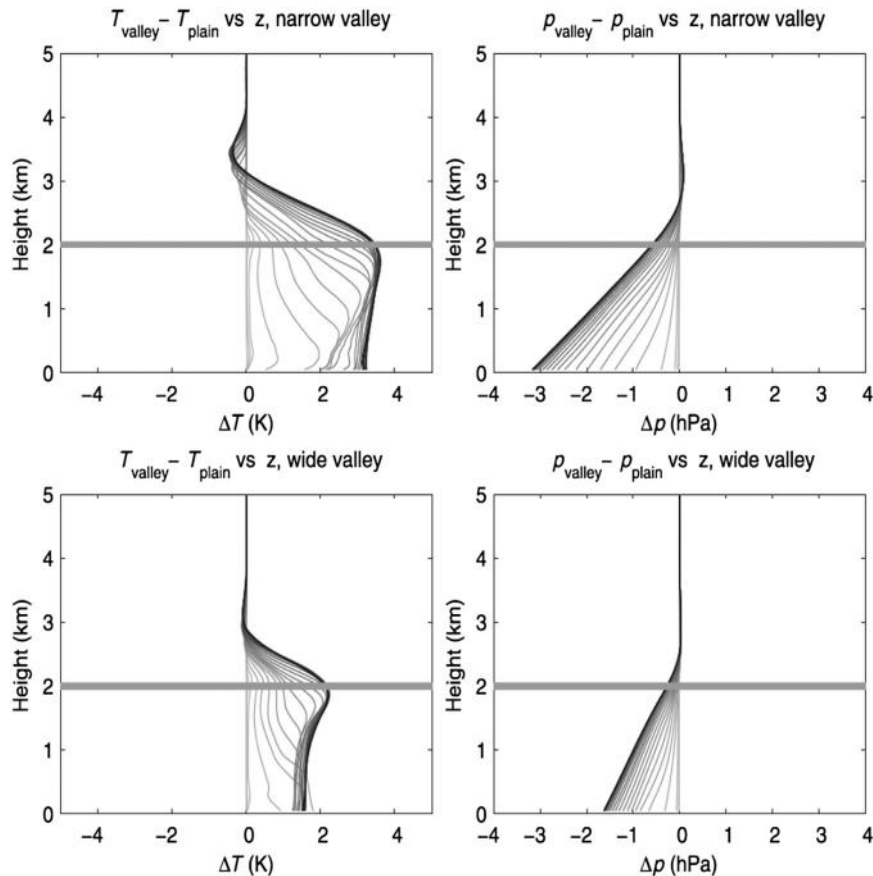


Figure 16 Daytime development of the vertical profiles of temperature and pressure differences, taken at the same heights, between the interior of an idealized valley and an adjacent plain region. Two cases are considered, representative of a narrow and a large valley (the two valley cross sections are the same as in Figure 13). Upper panels refer to the narrow valley and lower panels to the larger one. In all frames, profiles evolve in time from an unperturbed state (representative of sunrise time), where the atmosphere has the same vertical structure both in the valley and in the plain (light gray), to a final condition (12 h later) where the contrasts between the valleys and the plains are largest (black). Reproduced with permission from Serafin, S., Zardi, D., 2011. Daytime development of the boundary layer over a plain and in a valley under fair weather conditions: a comparison by means of idealized numerical simulations. *Journal of Atmospheric Science* 68, 2128–2141.

a valley displays a shallower mixed layer, topped by a deeper stable layer. The valley atmosphere is potentially warmer than over the plain throughout the depth of the boundary layer. Also, under the same forcing, the narrow valley produces a warmer atmosphere than the wider one. There are two main reasons: slope flows at the valley sidewalls promote the intake of air at the sidewall feet, and thus remove newly heated air from above the valley floor, subtracting it from getting involved in the development of the CBL. However, the wider the valley, the smaller the mean air intake per unit valley width, so this process is more effective on a narrow valley floor. As a consequence the mixed layer that develops over the valley floor is generally shallower than over a plain, and this aspect is more evident in the narrow valley. The reason consists in the fact that slope flows along the sidewalls develop almost identically in both valleys, but their effect on the core of the valley atmosphere depends on the valley cross-section aspect ratio: the larger the valley, the weaker the impact. Indeed the diversion of air from the valley toward the slope feet has less effect on a wider valley, so surface sensible heat flux can develop a CBL, which is more similar to that over a flat plain. On the other hand, the same air masses set in motion by slope flows feed the compensating subsidence at the valley top. However, mass conservation requires that the downward velocity of this sinking motion be inversely proportional to the valley width, so a weaker subsidence occurs in a wider valley.

All of these processes result in valley–plain contrasts in the vertical profiles of temperature and pressure, which are well summarized for an idealized valley case in [Figure 16](#) (see caption for explanation). Similar contrasts are also observed in real cases: [Figure 17](#) shows such a situation observed in temperature profiles from radiosoundings taken respectively at Innsbruck, in the upper Inn Valley in the Alps, and at Munich, in the adjacent Bavarian Plain.

The effect of such contrasting vertical structures on the winds is also clearly exemplified by the diurnal cycle observed in the Adige Valley, on the southern side of the Alps, facing the adjacent Po Plain ([Figure 18](#)). This valley displays a gradually sloping floor along the 150 km path connecting the surroundings of the city of Verona (91 m above mean sea level, AMSL) in

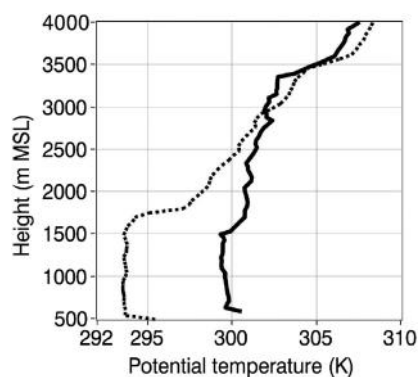


Figure 17 Vertical profiles of potential temperature from radiosoundings at Munich in the Bavarian plain (dotted line) at 519 m AMSL, and Innsbruck in the Inn Valley (solid line) at 574 m MSL, at 1200 UTC 19 July 2002. From Weissmann, et al., 2005. The Alpine mountain–plain circulation: airborne Doppler lidar measurements and numerical simulations. *Monthly Weather Review* 133, 3095–3109.



Figure 18 Map of the Adige valley in the southern Alps. Courtesy L. Giovannini.

the plain to the upper valley at Merano (330 m AMSL). Results reproduced in [Figures 19–21](#) report averaged values for all the days in which favorable weather conditions allowed a full development of valley winds in the years 2004–11. In particular, [Figure 19](#) shows diurnal mean cycles of along-valley wind speed at various stations along the valley. Amplitude and phase of local wind strength are strongly affected by local topography and land cover. Nevertheless the typical cycle of daytime up-valley wind, peaking in the afternoon, and down-valley nocturnal winds, weaker but persisting throughout the night, is clearly reproduced at all the stations. As already pointed out, this diurnal wind cycle occurs in connection with the corresponding cycle of horizontal pressure distribution. This connection is shown by the mean diurnal oscillation of surface pressure at various stations along the valley, from the plain (Verona) to the upper valley (Merano), reproduced in [Figure 20](#). Notice that the amplitude of the diurnal surface pressure cycle displays an increasing trend up-valley, with the smallest value in Verona and the largest in Merano. As a consequence the along-valley pressure distribution displays the layout shown in [Figure 21](#). The pressure change with along-valley distance from the plain is almost linear at any time. However, the higher amplitudes occurring further up-valley cause a daily periodic reversal of the pressure gradient.

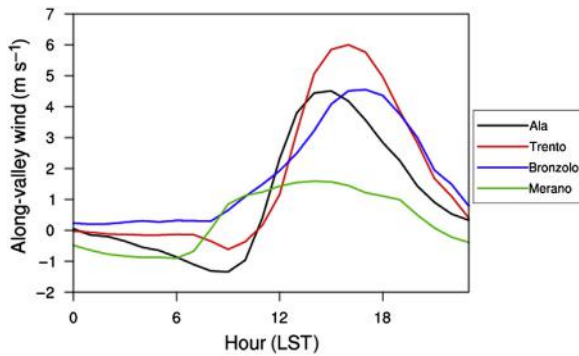


Figure 19 Average diurnal cycle of the along-valley wind strength along the Adige Valley. Courtesy L. Giovannini.

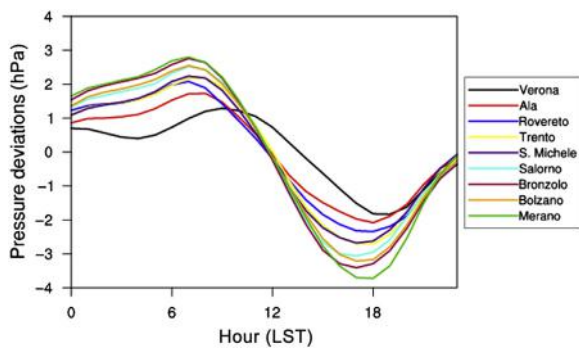


Figure 20 Average diurnal cycle of surface pressure at the stations indicated in Figure 18. Courtesy L. Giovannini.

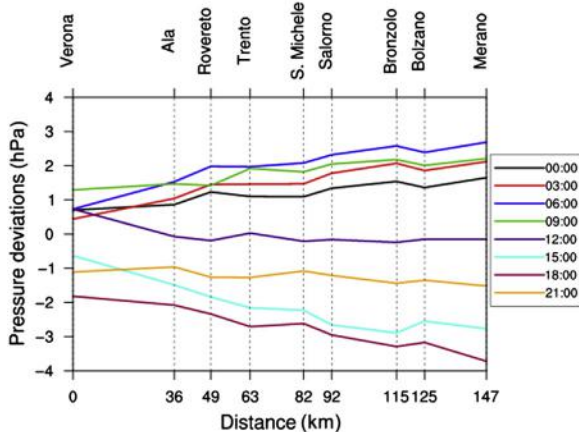


Figure 21 Average diurnal development of the surface pressure deviation (time LST in the chart). The slope of the curves gives an idea of the local horizontal pressure gradient along the valley. Courtesy L. Giovannini.

Similar pressure cycles occur at all levels within the layer affected by valley winds, and are the main drivers of the daily alternating wind strength and direction at any level. The vertical structure of these winds is strongly dependent on local

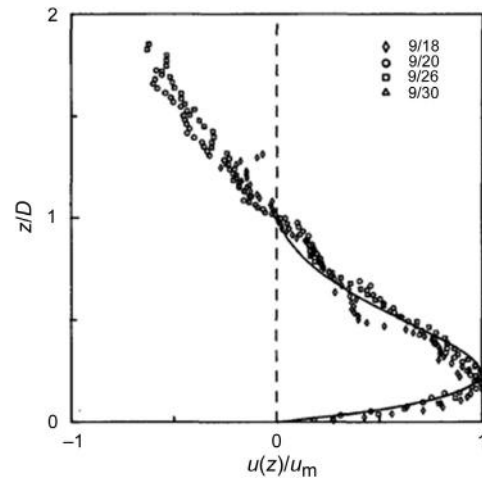


Figure 22 Vertical velocity profiles of down-valley winds normalized with the wind speed at the jet maximum u_m for the days 18, 20, 26, and 30 September 1984 in the Brush Creek Valley (Colorado). The solid line is a least-square fit of the expression $u(z)/u_m = A \exp(-Bz) \sin(\pi z/D)$, resulting in $A = 3.2$ and $B = 3.3$ m. Reproduced with permission from Clements, et al., 1989. Mean structure of the nocturnal drainage flow in a deep valley. *Journal of Applied Meteorology* 28, 457–462.

topographic features (curvatures, valley width changes, etc.), as well as on the strength and timing of the various forcing terms. As a consequence a generally valid shape for valley wind profiles, as schematized in Figure 1, cannot be easily provided. However, in some cases vertical profiles inspired by Prandtl's (1942) solution for slope flows successfully reproduced observations, as shown in Figure 22 for along-valley wind measurements in the Brush Creek Valley.

Interactions with Mountain-Plain Circulations

Daily periodic winds occurring along valleys lying in the major mountain ranges (such as the Alps, the Rocky Mountains, the Himalayan Chain, and Tibetan Plateau) are often embedded within the associated larger scale mountain-plain circulations, embracing the whole extent of the mountain range. Mountain-plain circulations are also driven by a diurnal pressure oscillation between the mountainous region and the adjacent plains. For the largest mountains, these oscillations have been observed to affect one or more modes of the atmospheric tides at global scale. The daily periodic reversal of this wind system is somewhat delayed relative to slope and valley winds because of the larger air mass involved.

Based on extensive numerical model simulations on a targeted area East of the Rocky Mountains, Wolyn and McKee (1994) identified a conceptual model of the daytime evolution of the mountain-plain circulation, which includes a sunrise state and three phases (Figure 23). The main features of the sunrise state are the bulging isentropes, the jet down the east side of the barrier and the stable core (Panel A). Divergence created by the nocturnal flows helps create the downward bulging isentropes. Phase 1 (Panel B) results from the weakening nocturnal flows interacting with surface heating, and lasts until 3–4 h after sunrise. Warming is associated with the weakening nocturnal jets, and occurs up to 20 km east of the

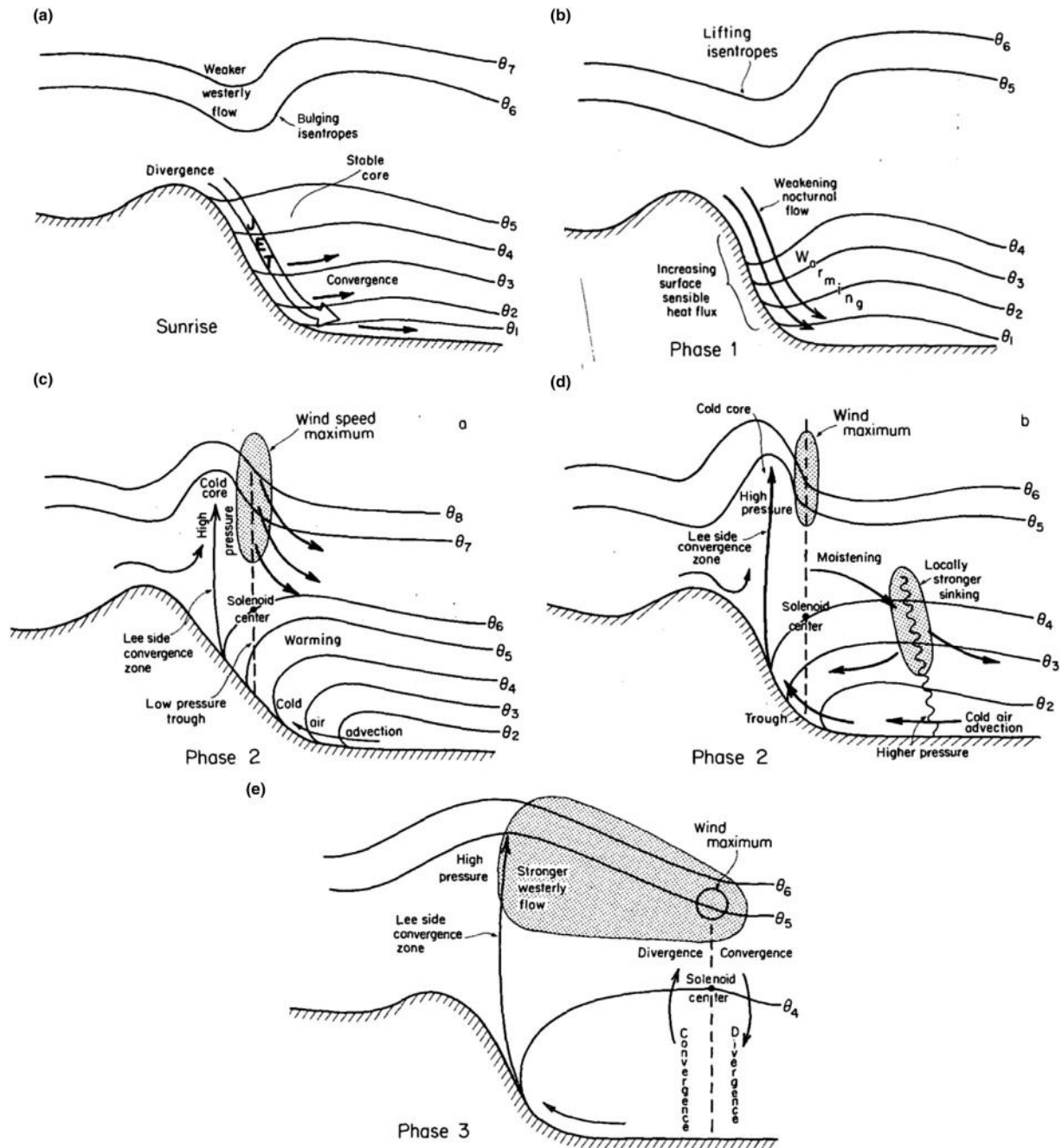


Figure 23 Conceptual daytime mountain-plain circulation model: (a) sunrise state, in which there is an interaction between nocturnal thermal and ambient flows; (b) phase 1, during which the weakening nocturnal flow interacts with surface heating; (c–d) phase 2, consisting of the developing solenoid; and (e) phase 3, the migrating solenoid. Further details are discussed in the text. Reproduced with permission from Zhang, Koch, 2000. Numerical simulations of a gravity wave event over CCOPE. Part II: Waves generated by an orographic density current. Monthly Weather Review 128, 2777–2796.

barrier base. Panel C shows the first stage of phase 2, characterized by a developing solenoidal circulation. The CBL on the eastern plains is suppressed due to horizontal cold-air advection in the CBL and the warming above the CBL. In stage 2 (Panel D) sinking and horizontal warm-air advection immediately east of the solenoid center warms the air sufficiently to create a negative pressure gradient (lower to the west) in the stable core above the CBL. This region of negative pressure

gradient expands eastward, and the reversal from negative to positive of the horizontal pressure gradient is marked by a pressure ridge: when this pressure ridge passes over a site, easterly flows appear above the CBL, strong vertical wind shear develops in the region between the strengthening up-slope and the westerly return flows, and the westward mass flux in the up-slope flow increases at a faster rate. Phase 3 (Panel E) is characterized by a migrating solenoid, whose center is located in

a pressure trough beneath the eastward-moving leading edge of the cold core. The migrating solenoid is only a transient disturbance: the main daytime circulation remains nearly stationary during this phase. Generally when the solenoid passes a location, the CBL grows explosively and the depth of the up-slope flow increases.

Atmospheric Processes Affected by Daily Periodic Mountain and Valley Winds

A variety of atmospheric processes may be affected by the diurnal cycle of valley winds. Here we concentrate on two aspects, namely processes involving water and those affecting air quality.

Water Cycle and Moist Processes

The dynamics of valley winds involve essentially dry processes. Nonetheless many valleys lie near, or contain, water bodies or humid and vegetated areas. Whenever an appreciable water content is available, moist processes may significantly affect the energy and mass budgets associated with valley flows. For instance, higher contents of soil moisture affect the partitioning of surface fluxes, and tend to increase latent heat fluxes at the expense of reduced sensible heat fluxes, thus hampering the heating of surface layers. This is the reason why thermally driven circulations are best developed in arid environments.

Rising motions of moist air may eventually lead to condensation and to the formation of clouds in the upper part

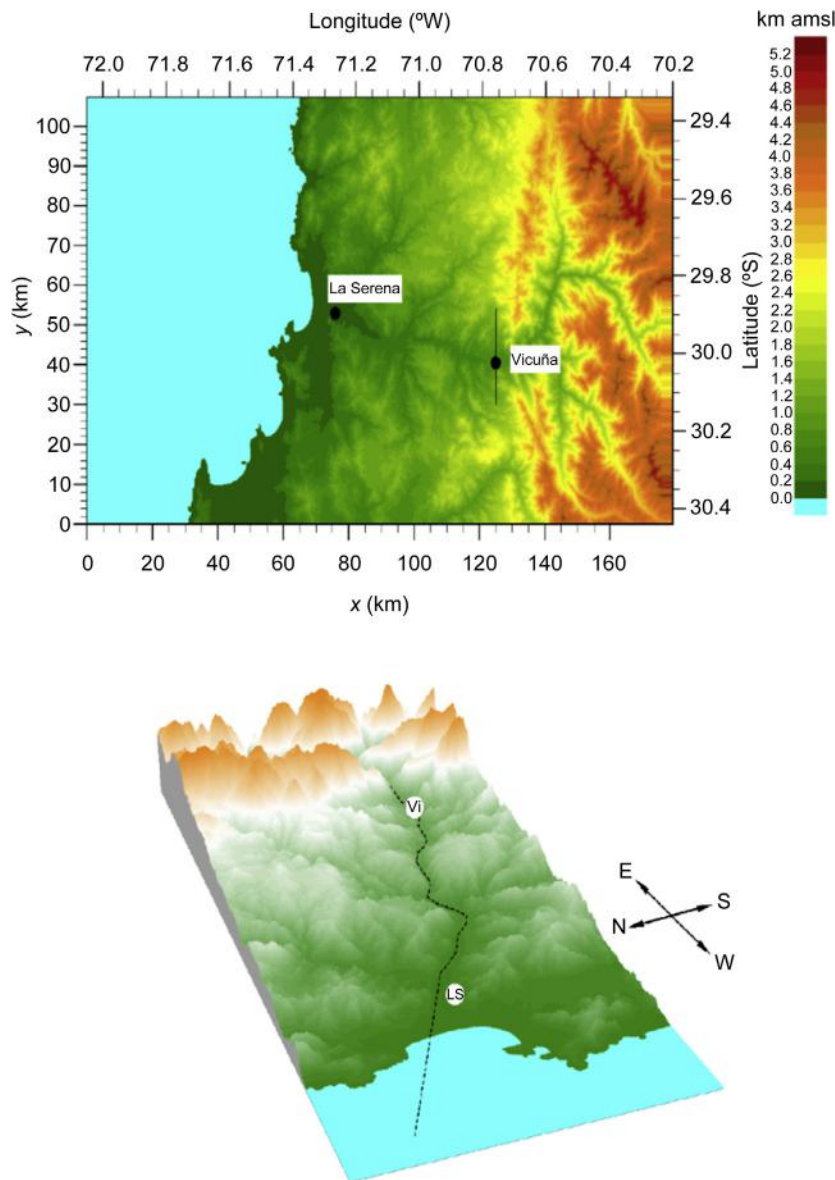


Figure 24 Map (top) and 3-D perspective (bottom) of the Elqui Valley in the Andes. The dashed line in the valley is oriented along the valley axis. 'LS' and 'Vi' indicate the positions of La Serena (145 m AMSL) and Vicuña (650 m AMSL), respectively. AMSL, above mean sea level. Reproduced with permission from Bischoff-Gauß, et al., 2008. Model Simulations of the Boundary-Layer Evolution over an Arid Andes Valley. Boundary-Layer Meteorology 128, 357–379.

of valley sidewalls or above the crests. The resulting clouds may cover the sky and thus reduce the radiative energy input, and produce showers and thunderstorms, especially in the afternoon.

Nocturnal cooling by drainage winds, especially at the valley floor, may lead to condensation. The formation of dew may be beneficial as water supply to vegetation and cultures, especially in arid areas. On the other hand, frost may produce serious damage to crops, and surface icing, as well as fog, may lead to serious risks for transportation safety.

Examples of the first case are found in some valleys on the western side of the Andes range, facing the Pacific Ocean, such

as the Elqui Valley in Chile (Figure 24). Daily sums of nocturnal dew deposition by condensation of moist air, advected from the coast by daytime up-valley winds are shown in Figure 25. The annual nocturnal sum of dew deposition amounts to 5–10 mm year⁻¹, which is about 5–10% of the mean annual precipitation, but almost of the same order of magnitude as precipitation amounts in dry years. So, as in many other arid areas, dew deposition is an important additional source of water for the natural vegetation, especially in dry years.

Also, organized thermally driven vertical motions of moist air associated with valley winds, may contribute, like other

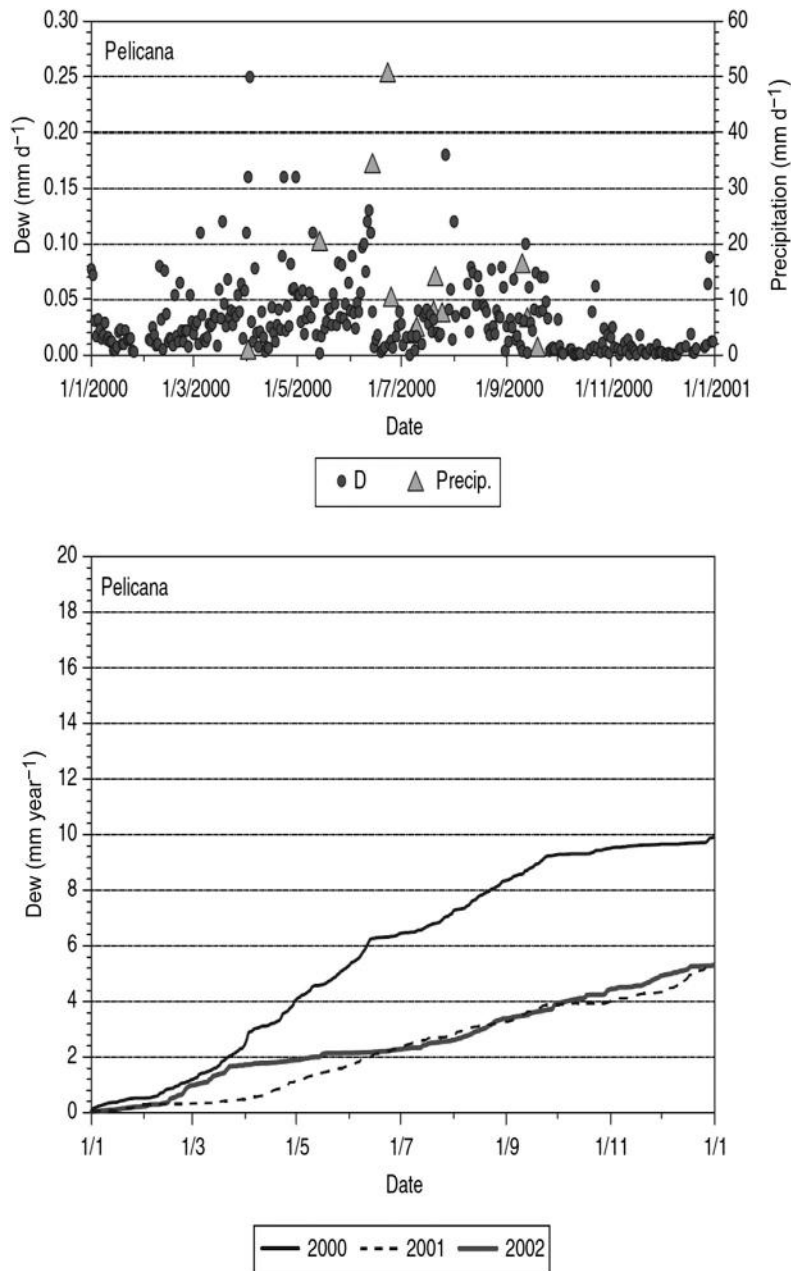


Figure 25 Daily sum of nocturnal dew deposition (D) at Pelicana and daily sum of precipitation at La Vicuña in 2000 (upper panel) and accumulated annual dew formation for 2000–02 (lower panel). Reproduced with permission from Kalthoff, et al., 2006. The energy balance, evapotranspiration and nocturnal dew deposition of an arid valley in the Andes. *Journal of Arid Environments* 65, 420–443.

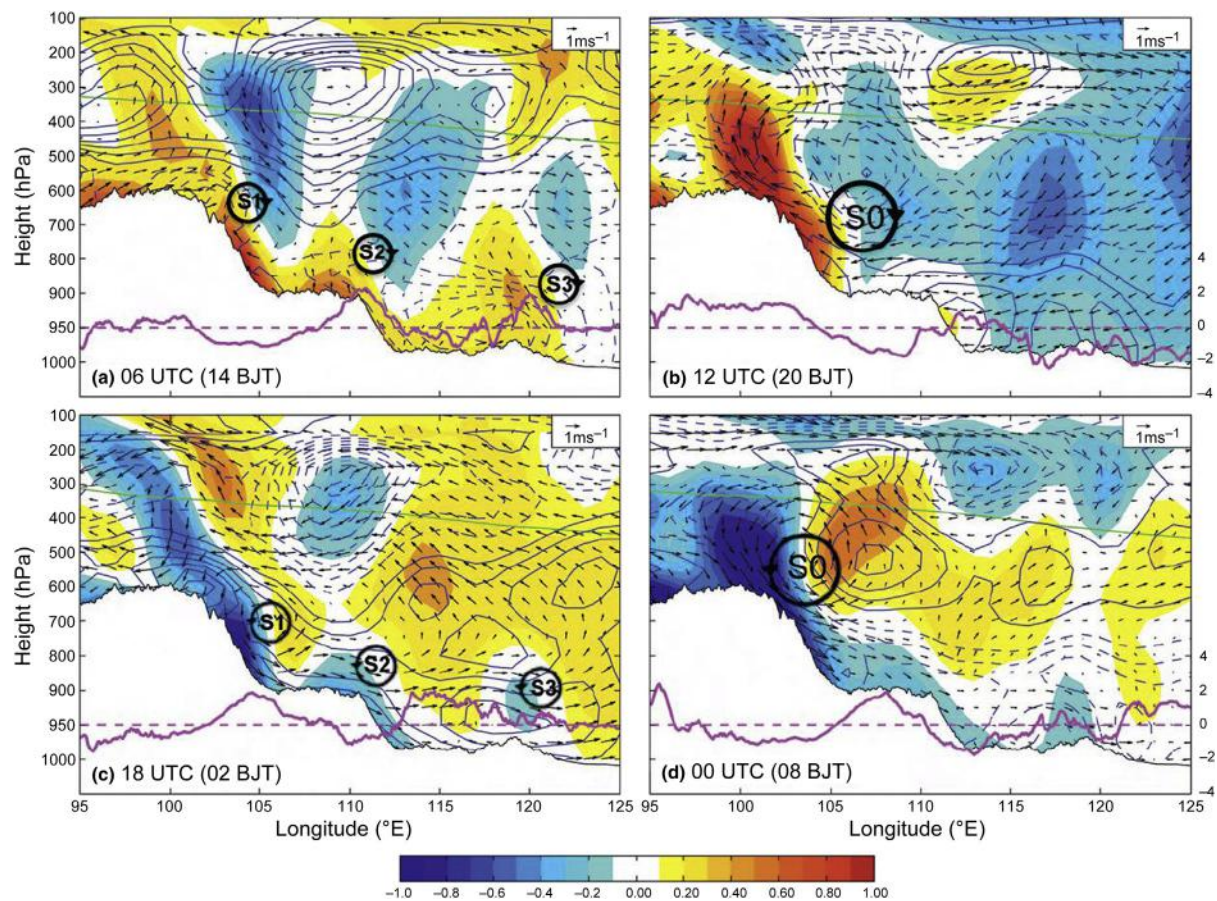


Figure 26 Vertical profiles East of the Tibetan plateau: the vertical motion deviations (unit: cm s^{-1} , colored), the perturbation vertical circulation vectors (zonal wind and 100 times of vertical velocity), and the perturbation meridional winds (0.2 m s^{-1} ; solid blue, positive; dashed, negative) latitudinally averaged between 27°N and 35°N diagnosed with GFS analyses at (a) 0600, (b) 1200, (c) 1800, and (d) 0000 UTC during the pre-mei-yu period (15 May–15 June). The pink solid curves show the averaged normalized diurnal precipitation deviations with the pink dashed straight line as the zero value. The black solid curves show the averaged terrain elevations. The green solid lines show that the zonal wind is equal to the mean diurnal propagation speed of 15 m s^{-1} . The S0, S1, S2, and S3 show the approximate solenoid centers. Reproduced with permission from Bao, et al., 2011. Diurnal variations of warm-season precipitation east of the Tibetan Plateau over China. *Monthly Weather Review* 139, 2790–2810.

combinations of thermally driven flows over complex terrain, to promote convective and orographically induced clouds, as well as the associated precipitation phenomena. Mountain-plain circulations also play a key role in moisture transport, and in the initiation of moist convection. Typically convergence of the daytime flows over the mountains produces afternoon clouds and air mass thunderstorms, and the divergent nighttime sinking motions produce late afternoon and evening clearing. A remarkable example, as to intensity and extension, is offered by the mountain-plain circulation effects on warm-season precipitation east of the Tibetan Plateau, which are strongly affected by the differential heating between the plateau, the highlands, the plains, and the ocean. **Figure 26** shows latitude-averaged vertical cross sections of typical circulation patterns associated with the diurnal cycle. In the early afternoon (**Figure 26(a)**), there are three distinct west–east solenoidal circulations in the lower- to mid-troposphere. These circulations are driven by the differential diabatic heating, with the upward branches on the highland-plateau slopes and the downward branches over the low

basins, plains, and oceans. The westernmost and strongest solenoid (S1) has the westward-tilted rising branch over the eastern slope of the Tibetan Plateau and the sinking branch over the Sichuan basin. The second solenoid in the middle (S2) has a rather shallower rising branch over the highlands along the Qinling and Wushan mountain ranges, and a more extended and broader sinking branch over the east China plains. The third solenoid (S3) has the rising branch along the coastal lands and the weak sinking branch over the nearby oceans. Each of the upward branches of the solenoids corresponds to a diurnal precipitation peak. On a larger scale there also exists a broader domain-wide vertical solenoid circulation (S0) across all three solenoids with the upward branch on the eastern Tibetan Plateau and the downward branch over the plains. In the early evening (**Figure 26(b)**), the S0 becomes the dominant mode in the cross section, with the upward motion strengthened at the eastern slope of the Tibetan Plateau and the downward motion over most of the areas eastward except for the weak upward branch of the S2 in the lower troposphere on the eastern slope of the Qinling and Wushan mountain ranges.

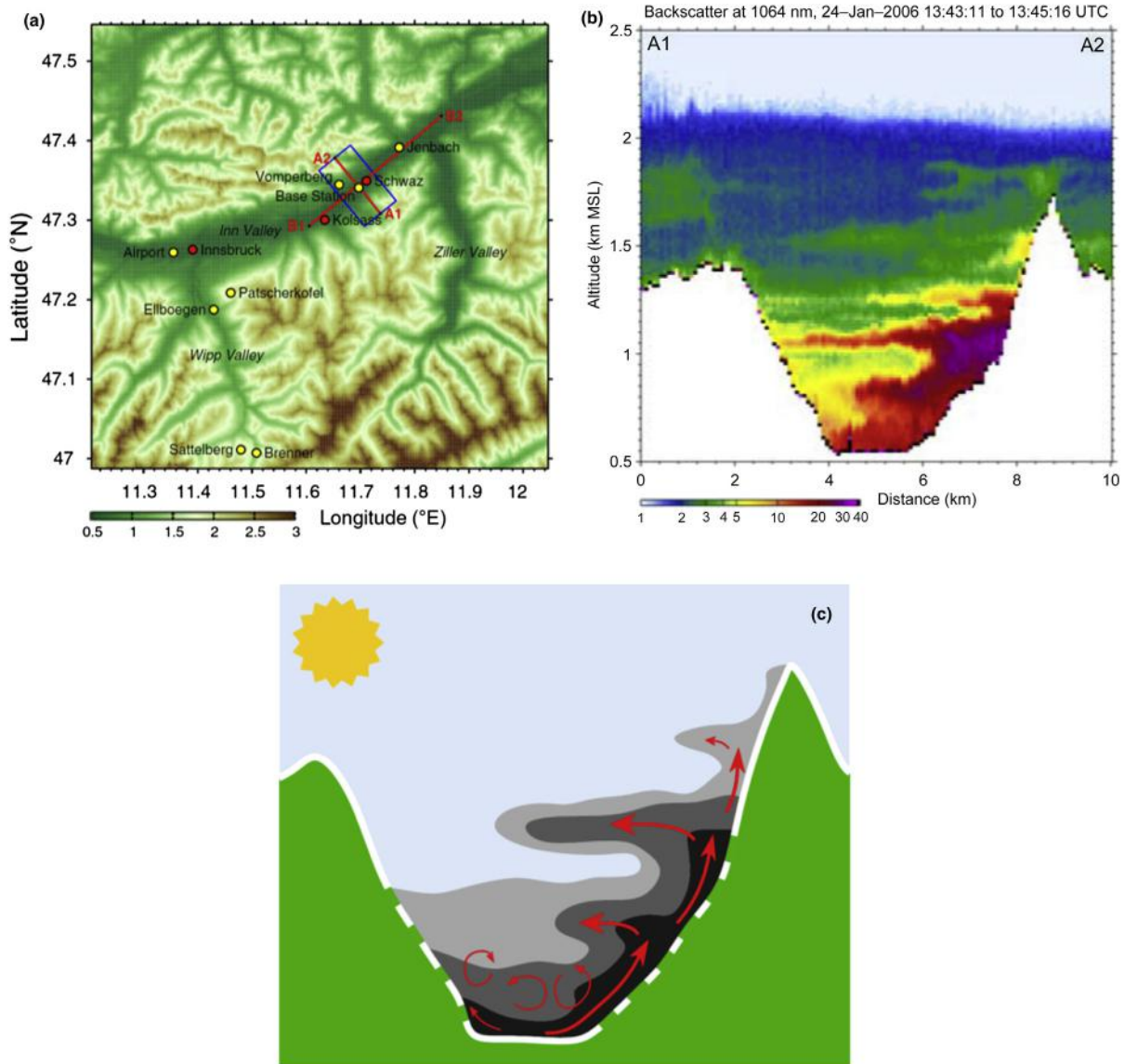


Figure 27 Selected pictures from the wintertime field measurements of pollutant dispersion in the Inn Valley (Austria). (a) Map of the target area: color scale represents elevations in km AMSL, yellow and red markers indicate the locations of measurement sites and towns. (b) Vertical transects of aerosol backscatter intensity at 1400 UTC 24 January 2006 (A1–A2 as in panel a). (c) Schematic of pollution transport processes by up-slope winds. Arrows indicate mean flow and turbulent eddies. A white solid (dashed) line indicates a closed (broken) snow cover. AMSL, above mean sea level. Reproduced with permission from Gohm, et al., 2009. Air Pollution Transport in an Alpine Valley: results from airborne and ground-based observations. *Boundary-Layer Meteorology* 131, 441–463.

Both the upward branches of S1 and S2 continue to be associated with local diurnal precipitation maxima at this hour, while the broader and stronger sinking branch over the east China plains corresponds to a broad local precipitation minimum phase in these regions at this hour. The coastal solenoid S3 is mostly absent in this early evening hour, though the sinking branch over the ocean is considerably stronger than over the coastal land. In the early morning (Figure 26(c)), the nighttime vertical circulation is nearly a complete reversal of the daytime circulation (Figure 26(a)) with the downward branches over the highland–plateau slopes and the upward branches over the low-lying plains–basins. Consequently, strong diurnal precipitation peaks (nocturnal rainfall maxima)

are observed over the Sichuan basin (part of S1) and over the east China plains (part of S2). However, the nocturnal circulation pattern may be further strengthened to peaked maximum at 2100 UTC: the nocturnal precipitation peak phase over the plains is also coincidental with a developing low-level southerly jet that transports more warm moist air to this area and contributes to the enhancement of nighttime precipitation. Boundary layer processes associated with the reduced turbulence diffusion due to ceased daytime heating are believed to be responsible for the development of the low-level nocturnal jet. A few hours after sunrise (Figure 26(d)), the vertical circulation evolves from the nocturnal pattern in Figure 26(c) to the daytime pattern in Figure 26(a), and is again dominated

by the domain-scale broad solenoid (S0) as a reversal of that in [Figure 26\(b\)](#) with the downward motion maximized at the eastern slope of the Tibetan Plateau and the upward motion in a broad area to its east maximized at the mid-troposphere.

Air Pollution

Airflows and thermal structures associated with valley winds may affect the fate of atmospheric pollutants in many ways. Typical sources of air pollution are settlements, infrastructures, and industrial activities mainly based at the valley floor or, even more frequently, on the adjacent plains. So up-valley winds may convey highly polluted air into the valleys, whereas down-valley winds may have a cleansing effect of polluted air at the valley exit over the plain during nighttime. On the other hand, up-slope flows may transport to upper levels either already formed primary pollutants, or precursors of secondary pollutants released at the valley floor. There the exposure to radiation may enhance the formation of photochemical species, which eventually get drained to lower levels by nocturnal down-slope winds.

Also the thermal structures associated with valley winds, through their effects on stability, may affect the fate of atmospheric pollutants. For example, the depressed CBL over the valley floor may reduce the mixing height, leading to higher concentrations. This may be particularly critical during wintertime, when on one hand the reduced radiative input leads to shallower mixed layers, and on the other hand, more pollutants are emitted from house heating and traffic. In a similar way the frequent occurrence of nighttime inversions, especially during wintertime, also reduces the mixing height.

[Figure 27](#) provides an example of nontrivial features characterizing wintertime pollutant dispersion patterns in the Inn Valley in the Alps.

Acknowledgments

Data from surface weather stations used for the graphics presented in [Figures 19–21](#) were kindly provided by the Hydrographic Office of the Autonomous Province of Bolzano (for Merano, Bolzano, Bronzolo, and Salomo stations), the Meteorological Office of the Autonomous Province of Trento – Meteotrentino (Rovereto station), the Edmund Mach Foundation (San Michele, Trento, and Ala stations), and the Environmental Agency of the Veneto Region (Verona station).

Lorenzo Giovannini kindly performed the climatological analysis of time series from the above data, and prepared [Figures 18–21](#).

The Author is greatly indebted to Massimiliano de Franceschi, Lorenzo Giovannini, Lavinia Laiti, Stefano Serafin, Elena Tomasi, and Felicity Hope for carefully reviewing the manuscript, and suggesting many valuable improvements.

See also: Agricultural Meteorology and Climatology. **Air Sea Interactions:** Freshwater Flux; Momentum, Heat, and Vapor Fluxes; Sea Surface Temperature; Surface Waves. **Aviation Meteorology:** Aircraft Emissions. **Boundary Layer (Atmospheric) and Air Pollution:** Complex Terrain; Convective Boundary Layer; Diurnal Cycle; Microclimate; Stably Stratified Boundary Layer; Surface Layer. **Clouds and Fog:** Fog. **Hydrology, Floods and Droughts:** Soil Moisture. **Land-Atmosphere Interactions:** Canopy Processes; Overview; Trace Gas Exchange. **Mesoscale Meteorology:** Mesoscale Convective Systems. **Mountain Meteorology:** Katabatic Winds; Land and Sea Breezes; Overview. **Numerical Models:** Parameterization of Physical Processes: Clouds; Parameterization of Physical Processes: Turbulence and Mixing. **Oceanographic Topics:** Surface/Wind Driven Circulation. **Ozone Depletion and Related Topics:** Surface Ozone Effects on Vegetation. **Satellites and Satellite Remote Sensing:** Earth's Radiation Budget; Surface Wind and Stress; Water Vapor. **Stratosphere/Troposphere Exchange and Structure:** Local Processes. **Synoptic Meteorology:** Anticyclones. **Thermodynamics:** Moist (Unsaturated) Air.

Further Reading

- Barry, R.G., 2008. *Mountain Weather and Climate*, third ed. Cambridge University Press. 506pp.
- Defant, F., 1949. Zur Theorie der Hangwinde, nebst Bemerkungen zur Theorie der Berg- und Talwinde [A theory of slope winds, along with remarks on the theory of mountain winds and valley winds]. *Archiv für Meteorologie, Geophys. Bioclimatologie*, Ser. A 1, 421–450 [Theoretical and Applied Climatology]. [English translation: Whiteman, C.D., Dreiseitl, E., 1984. *Alpine meteorology*, Translations of classic contributions by Wagner A., Ekhardt, E., Defant, F., PNL-5141/ASCOT-84-3. Pacific Northwest Laboratory, Richland, Washington, 121pp.] <http://www.osti.gov/bridge/servlets/purl/6665518/6665518.pdf>.
- Egger, J., 1990. Thermally forced flows: theory. In: Blumen, W. (Ed.), *Atmospheric Processes over Complex Terrain*. American Meteorological Society, Boston, pp. 43–57.
- Geiger, R., Aron, R.H., Todhunter, P., 2009. *The Climate Near the Ground*, seventh ed. Roman and Littlefield Publishers, Maryland, 523pp.
- Whiteman, C.D., 1990. Observations of thermally developed wind systems in mountainous terrain. In: Blumen, W. (Ed.), *Atmospheric Processes over Complex Terrain*, American Meteorological Society Meteorological Monographs, vol. 45 (23), pp. 5–42.
- Whiteman, C.D., 2000. *Mountain Meteorology: Fundamentals and Applications*. Oxford University Press, New York, 355pp.
- Wolyn, P.G., McKee, T.B., 1994. The mountain–plains circulation east of a 2-km-high north–south barrier. *Monthly Weather Review* 122, 1490–1508.
- Zardi, D., Whiteman, C.D., 2013. Diurnal mountain wind systems. In: Chow, F.K., De Wekker, S.F.J., Snyder, B. (Eds.), *Mountain Weather Research and Forecasting – Recent Progress and Current Challenges*, Springer Atmospheric Sciences. Springer, Berlin, pp. 37–122.

NUMERICAL MODELS

Contents

Chemistry Models

Coupled Ocean-Atmosphere Models: Physical Processes

General Circulation Models

Methods

Model Physics Parameterization

Parameter Estimation

Parameterization of Physical Processes: Clouds

Parameterization of Physical Processes: Gravity Wave Fluxes

Parameterization of Physical Processes: Turbulence and Mixing

Spectral Models

Mesoscale Atmospheric Modeling

Cloud-System Resolving Modeling and Aerosols

Large-Eddy Simulation

Regional Prediction Models

Convective Storm Modeling

Chemistry Models

MP Chipperfield and SR Arnold, University of Leeds, Leeds, UK

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by M P Chipperfield, volume 4, pp 1414–1423, © 2003, Elsevier Ltd.

Synopsis

The formulation of chemical models is discussed. This article summarizes the component modules of chemical models (gas phase chemistry, heterogeneous chemistry, photolysis, deposition). The chemical continuity equation is described. Trajectory, one-dimensional, two-dimensional, and three-dimensional models are discussed. Example applications in the stratosphere and troposphere are given.

Introduction

Chemical models are used to test our understanding of atmospheric chemistry, and for predictions of the future state of the atmosphere. A model will contain different modules to treat processes such as gas phase chemistry, aqueous phase chemistry, heterogeneous chemistry, photolysis reactions, and emission and deposition of species. The formulation of the model will depend on the problem being studied. The core of a model is the chemical continuity equation, which is an expression of the rate of change of a chemical species. Integrating this continuity equation permits the model to step forward in time. In general, the large computational cost of calculating atmospheric chemistry leads to a number of

approximations in models, such as using grouping species into families and assuming that some species are in steady state. Models that are used to study atmospheric chemistry range from zero-dimensional 'box' models, which may contain very detailed chemistry schemes (e.g., 3000 species), to global three-dimensional models, which may contain around 50.

Use of Models

Numerical models are a mathematical summary of our current understanding of atmospheric chemistry. A good model should contain a representation of all of the important species,

reactions, and processes relevant for the particular system that is being studied. A numerical model can then be used for:

- testing our understanding of atmospheric chemistry by comparison between model calculations and observations;
- investigating the effect of a newly discovered reaction or process on other species; and
- predicting the future state of the atmosphere based on a series of assumptions.

The components of the model (physical and chemical processes considered, number of chemical species and reactions) will depend on the problem being addressed (e.g., spatial and temporal scale). The model needs to have an appropriate domain (e.g., global, regional), resolution (e.g., size of grid boxes), and time step. The model must contain all of the necessary processes, but other factors (usually computer resources) often constrain those that can be included.

Components of a Chemical Model

In order to calculate the time-dependent concentrations of chemical species in the atmosphere, a model must contain a representation of the important chemical and physical processes. **Figure 1** illustrates these components and how they are used in each model time step to calculate the chemical concentrations.

Gas Phase Reactions

All models of atmospheric chemistry will deal with gas phase reactions between species (see **Chemistry of the Atmosphere: Chemical Kinetics**). Laboratory measurements provide data,

which can be used to calculate the rate constants for gas phase reactions in models. For bimolecular reactions, the rate constant usually depends only on temperature and can be calculated from the Arrhenius equation (eqn [1]), where k is the rate constant at a temperature T , A_e is the Arrhenius factor, E_a is the activation energy, and R is the gas constant. The values for A_e and E_a are provided by laboratory data.

$$k(T) = A_e \exp(-E_a/RT) \quad [1]$$

Other expressions are used to fit certain bimolecular reactions, which show a pressure dependency (e.g., $\text{HNO}_3 + \text{OH}$) and termolecular (three-body) reactions. Expert panels regularly review the body of chemical kinetics literature and produce reports of recommended rate constants for use in atmospheric models. As well as providing modelers with an expert analysis of photochemical data, their common use permits model results to be referenced and for predictions from different models to be more easily compared.

Liquid Phase Reactions

Models dealing with tropospheric chemistry will need to account for the uptake of gases by cloud droplets, and chemical reactions within the clouds (aqueous chemistry). This gas uptake and reaction involves several steps (diffusion of gas to droplet, dissolution in droplet, diffusion through droplet, reaction in droplet, diffusion of products, and evaporation of dissolved products at the droplet surface). Treating aqueous chemistry in a model is more complicated than gas phase chemistry. First, the lifetimes of aqueous species are usually short and the system of differential equations to solve is stiff (see below). Second, when dissolved species are removed by

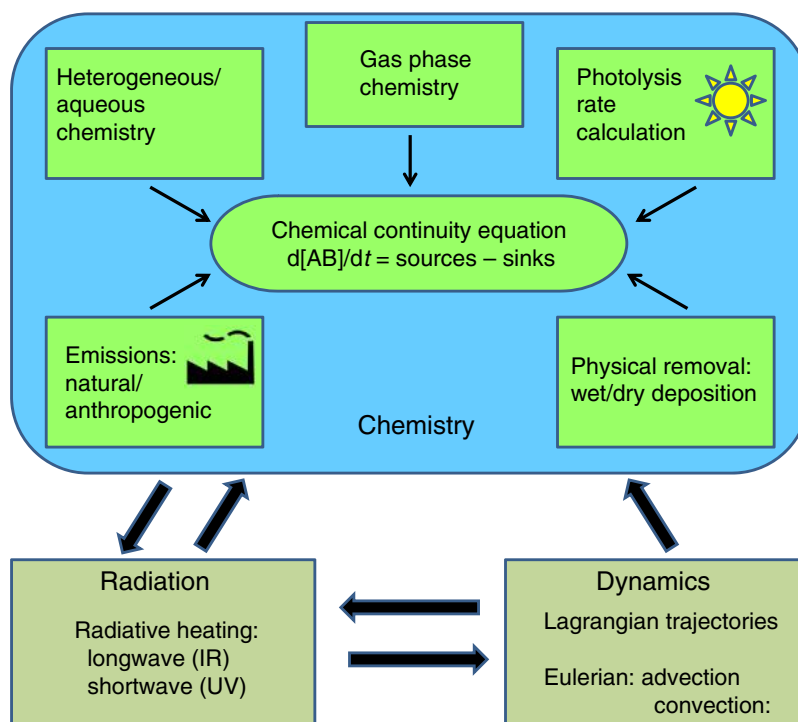
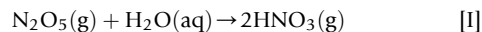


Figure 1 Components of a chemical model. A multidimensional model will also include dynamical and radiation modules. These may be combined so that the chemistry is or is not coupled (where changes in the concentrations of chemical species feedback on the radiation and dynamics).

reactions they are rapidly replaced by dissolution of more species from the gas phase. Therefore, the processes of dissolution and chemical reaction need to be solved in a coupled way.

Heterogeneous Reactions

Heterogeneous reactions involve the collision of a gas phase molecule with a solid or liquid particle, followed by a chemical reaction. An example is the hydrolysis of N_2O_5 shown by reaction [I].



This reaction is normally parametrized in models using a measured ‘reaction probability’ (γ) that an N_2O_5 molecule colliding with a surface will react, and a ‘collision frequency’ calculated using kinetic theory and the known or assumed concentration of aerosols. This treatment can also be used for heterogeneous reactions involving solid particles. An alternative treatment for liquid aerosols, when both reactants are soluble, is to treat the reaction as a liquid phase reaction (as above).

Photolysis Rate Coefficients

The photolysis rate coefficient (or photodissociation frequency), J is the first-order rate constant for the process shown by reaction [III], where h is the Planck’s constant, and ν is the frequency of the radiation.



As solar radiation is the driving force for atmospheric chemistry, the accurate calculation of J rates is an important component of models.

The photolysis rate coefficient for species AB (J_{AB}) is calculated from eqn [2], where $I(\lambda, z)$ is the photon flux at wavelength λ and altitude z in the atmosphere, and σ is the absorption cross-section.

$$J_{\text{AB}}(z) = \int I(\lambda, z) \sigma_{\text{AB}}(\lambda) d\lambda \quad [2]$$

In an atmospheric model eqn [2] is solved by replacing the integration over wavelength by a summation over discrete wavelength intervals. The World Meteorological Organization gives a list of 158 wavelength intervals, covering the range 175–850 nm, which are typically used in stratosphere–troposphere models. Fewer wavelength intervals can be used for troposphere-only models to save computer time.

The photon flux at a point in the atmosphere, $I(\lambda, z)$, is calculated using the flux at the top of the atmosphere and the Beer–Lambert law,

$$I_{\text{tr}} = I_0 \exp(-\epsilon cl) \quad [3]$$

Where I_0 is the incident radiation, I_{tr} is the transmitted radiation, ϵ is the absorption (or extinction) coefficient, c is the concentration of the absorber, and l is the path length. The flux at altitude z depends on attenuation by absorbing gases (mainly O_2 and O_3), scattering by molecules and aerosols, and reflection by the surface and clouds.

Models may calculate an instantaneous value of J , or an average value over daylight or 24 h period. An instantaneous and regularly updated value of J is necessary for a model to

reproduce the diurnal cycle of short-lived species (e.g., OH). However, this approach can be expensive computationally. For simulations of long-lived species, where it is not necessary to explicitly resolve the diurnal cycle, day- or 24-h-averaged J rates can be used. As the interactive, online calculation of J rates is nevertheless expensive, some models use precalculated ‘lookup’ tables. The J for the required conditions is interpolated from the tabulated values.

Emissions and Physical Removal

All chemical models, except those simulating isolated air masses over short periods, require a representation of the processes that input and remove chemicals from the atmosphere. Emissions of trace gases may occur through natural (e.g., vegetation) or anthropogenic processes (e.g., industrial emissions, aircraft). Expert panels regularly review the strength of these emission sources and provide data sets for input into models (e.g., the Global Emissions Inventory Activity (GEIA) of the International Global Atmospheric Chemistry (IGAC) program). Some of these emission sources (e.g., emissions of hydrocarbons from vegetation and trace gas and aerosol emissions from wildfires) have strong spatial and temporal variability driven by climatic conditions. Biogenic emissions of trace gases such as isoprene have a strong dependence on parameters such as temperature and sunlight, and so schemes are often included in models, which represent the response of such emissions to the meteorological data being fed to the model.

Chemical species may be physically removed from the atmosphere by wet or dry deposition processes. Dry deposition removes gases and particles at air–surface interfaces. Wet deposition involves the dissolution of a gas into a cloud droplet, which is then rained out.

Chemical Continuity Equation

Chemical models aim to use the known concentrations of chemical species at time t , and calculated rates of change, to calculate the concentrations at the end of the chemical time step $t + \Delta t$. The heart of the model is a ‘chemical continuity equation.’ This is an expression for the rate of change of a chemical species due to the chemical reactions that produce or destroy it. It is an expression of the conservation of mass and can be expressed as eqn [4], where P is the rate of chemical production and L is the rate of chemical loss (which usually depends on $[\text{AB}]$), and $[\]$ indicates a concentration. Other terms can be included in eqn [4] to account for other processes that affect $[\text{AB}]$ (e.g., physical removal).

$$\frac{d[\text{AB}]}{dt} = P - L[\text{AB}] \quad [4]$$

Table 1 lists a small subset of the 100 reactions that are important in the chemistry of the stratosphere (see **Stratospheric Chemistry Topics**: Overview; Halogens; Reactive Nitrogen (NO_x and NO_y); HO_x). This subset is used here to illustrate the form of a continuity equation, but note that detailed atmospheric models will include many more species and reactions. There are three reactions in Table 1, which either produce (reaction [X]) or

Table 1 Subset of gas phase reactions important in stratospheric chemistry

$\text{Cl} + \text{O}_3 \rightarrow \text{ClO} + \text{O}_2$	[III]
$\text{ClO} + \text{O} \rightarrow \text{Cl} + \text{O}_2$	[IV]
$\text{ClO} + \text{NO} \rightarrow \text{Cl} + \text{NO}_2$	[V]
$\text{ClO} + \text{NO}_2 + \text{M} \rightarrow \text{ClONO}_2 + \text{M}$	[VI]
$\text{ClONO}_2 + h\nu \rightarrow \text{Cl} + \text{NO}_3$	[VII]
$\text{OH} + \text{HO}_2 \rightarrow \text{H}_2\text{O} + \text{O}_2$	[VIII]
$\text{O}(^1\text{D}) + \text{H}_2\text{O} \rightarrow 2\text{OH}$	[IX]
$\text{OH} + \text{NO}_2 + \text{M} \rightarrow \text{HNO}_3 + \text{M}$	[X]
$\text{OH} + \text{HNO}_3 \rightarrow \text{H}_2\text{O} + \text{NO}_3$	[XI]
$\text{HNO}_3 + h\nu \rightarrow \text{OH} + \text{NO}_2$	[XII]

destroy (reactions [XI] and [XII]) HNO_3 . Based on this reaction set, the chemical continuity equation for the rate of change of the concentration of HNO_3 contains three terms and can be represented as eqn [5],

$$\frac{d[\text{HNO}_3]}{dt} = k_X[\text{OH}][\text{NO}_2][\text{M}] - k_{\text{XI}}[\text{OH}][\text{HNO}_3] - J_{\text{XII}}[\text{HNO}_3] \quad [5]$$

where k_n is the rate constant for reaction n , J_n is the photolysis frequency, and the square brackets indicate a concentration. A continuity equation can be written for each chemical species contained in the model. This gives a set of coupled first-order ordinary differential equations. In all but the very simplest cases (e.g., the decay of a radioactive tracer), an analytical solution is not possible and the coupled differential equations must be solved numerically. This system of differential equations is usually 'stiff,' i.e., the lifetimes (or time scales) of the chemical species being solved vary by several orders of magnitude (e.g., seconds to years). Therefore, sophisticated (and computationally expensive) solvers need to be used.

A chemical continuity equation similar to [4] can be written for each species contained in the reaction scheme. However, in practice a number of conceptual simplifications and numerical approximations can be made.

Photochemical Steady State

For a chemical species with a very short chemical lifetime it is not necessary, or desirable, to integrate the chemical continuity equation. The short lifetime increases the stiffness of the system and would require a short chemical time step. Computer time can be saved by placing short-lived species in steady state.

In the chemistry of the stratosphere and troposphere, the chemical lifetime of OH is of the order minutes or less. Therefore, the concentration of OH can be derived by placing it in the photochemical (or photostationary) steady state.

Based on the reaction scheme given above, this would give eqn [6].

$$\begin{aligned} \frac{d[\text{OH}]}{dt} &= 0 \\ &= 2k_{\text{IX}}[\text{O}(^1\text{D})][\text{H}_2\text{O}] + J_{\text{XII}}[\text{HNO}_3] \\ &\quad - k_X[\text{OH}][\text{NO}_2][\text{M}] - k_{\text{XI}}[\text{OH}][\text{HNO}_3] \\ &\quad - k_{\text{VIII}}[\text{OH}][\text{HO}_2] \end{aligned} \quad [6]$$

Therefore, at each time step in the model the [OH] can be derived from the calculated concentration of other species and the appropriate rate constants and photolysis rates. The calculated concentration of OH will vary throughout the diurnal cycle (e.g., as J_{XII} changes), although at each time step the instantaneous concentration is assumed to be constant. Note that as several interdependent species may be treated to be in steady state (e.g., both [OH] and $[\text{HO}_2]$ in the above example), the steady state concentration of these species should be derived iteratively.

Chemical Families

The number of continuity equations to be solved (and computational time) can be reduced by grouping closely coupled chemical species together in a family. As well as needing to solve only one continuity equation, the photochemical lifetime of the family is generally longer than the lifetimes of the individual members, producing a less stiff system (Figure 2). Finally, using chemical families has advantages in multidimensional models. Generally, it is not desirable to transport short-lived species separately as they have strong gradients (e.g., near the terminator), which can cause numerical problems (undershoots and overshoots) in advection schemes. A chemical family will generally have a smoother distribution and pose fewer problems for the advection scheme.

In stratospheric models a ClOx family is often defined as $[\text{ClOx}] = [\text{ClO}] + [\text{Cl}]$. This is justified because Cl is in rapid

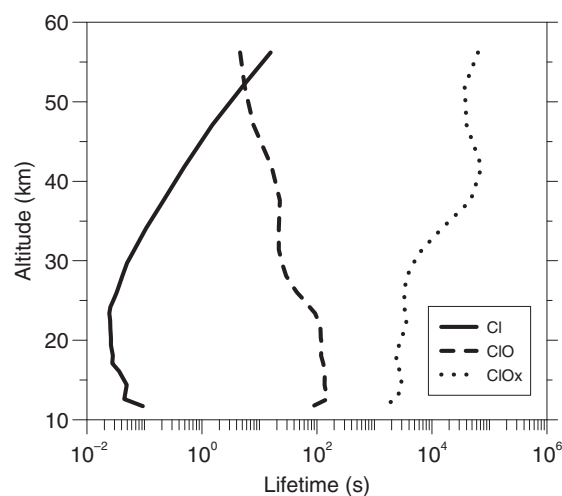


Figure 2 Photochemical lifetimes (defined as $1/(\text{first-order loss rate})$) of Cl, ClO, and ClOx ($= \text{Cl} + \text{ClO}$). The ClOx family has a much longer lifetime than the shortest lived family member, resulting in a less stiff system of equations to solve.

photochemical equilibrium with ClO, and change in the concentration of ClO will also affect Cl through the reactions, which interconvert the two. When a chemical family is used in a model, a single chemical continuity equation is written for the overall rate of change of the family. Based on the reactions given in Table 1, the continuity equation for ClOx can be expressed by eqn [7], where M represents any air molecule.

$$\begin{aligned} \frac{d[\text{ClOx}]}{dt} &= 0 \\ &= -k_{\text{VI}}[\text{ClO}][\text{NO}_2][\text{M}] + J_{\text{VII}}[\text{ClONO}_2] \end{aligned} \quad [7]$$

Note that reaction [III] for example, which simply interconverts Cl and ClO has no net effect of ClOx and does not appear in eqn [7]. The concentration of the total family must be divided among the $n - 1$ individual members. This is achieved by writing $n - 1$ steady state expressions for $n - 1$ members. In the case of the ClOx family, by placing Cl in steady state ($d[\text{Cl}]/dt = 0$) we can derive eqn [8] for the ratio of $[\text{Cl}]/[\text{ClO}]$.

$$\frac{[\text{Cl}]}{[\text{ClO}]} = \frac{k_{\text{IV}}[\text{O}] + k_{\text{V}}[\text{NO}] + J_{\text{VII}} \frac{[\text{ClONO}_2]}{[\text{ClO}]}}{k_{\text{III}}[\text{O}_3]} \quad [8]$$

Although this equation is derived by assuming Cl is in steady state, the concentration of Cl (and ClO) will vary over the model time step as ClOx changes. However, eqn [7] effectively fixes the ratio of Cl:ClO over this time step.

Care is needed when deriving these expressions for the partitioning of family members. Most of the terms in eqn [7] can be identified with reactions [III], [IV], and [V], which directly interconvert Cl and ClO. However, there is also a term involving $[\text{ClONO}_2]/[\text{ClO}]$, which is related to the two-step interconversion of ClO and Cl via the formation and photolysis of ClONO₂. It is very important to include these indirect terms as they are often associated with catalytic cycles that destroy stratospheric O₃ via the reaction [III]. In order for the model to correctly determine the O₃ loss, the calculated $[\text{Cl}]$ must be accurate.

Another chemical family commonly used in atmospheric models is 'odd oxygen,' which is defined as $\text{Ox} = \text{O}(^3\text{P}) + \text{O}(^1\text{D}) + \text{O}_3$. This family provides a very convenient way of calculating the atmospheric abundances of O₃, O(³P), and O(¹D) below about 70 km. Above this altitude the photochemical lifetime of O becomes long (due to the low air density) and so, O and O₃ can no longer be assumed to be in photochemical equilibrium.

Mechanism Reduction

The number of species and reactions involved in chemical reactions of organic species in the polluted (urban) troposphere is huge. For example, the University of Leeds Master Chemical Mechanism model contains around 3800 species and 10 000 reactions. For many practical purposes the number of reactions needs to be reduced. The methods used for reducing the number of species in urban photochemical models are: (1) the carbon-bond lumping method (when organic species are separated into a few common bond groups), (2) the surrogate species method (where species with similar reactivity are grouped together and solved as one species), and (3) the lumped species

method (where species are grouped together but the reaction rate constants for the lumped group is a mole fraction weighted average of the rate constants for the individual species).

Types of Models

A range of chemical models exists appropriate for studying different problems. In all cases the model consists essentially of a chemical 'box' model (with the description of the required chemistry) either used alone, or within an array of grid boxes in a multidimensional model.

Box Models and Trajectory Models

A chemical 'box' model solves the chemical continuity equations in a single air mass. These models can be computationally cheap, allowing detailed reaction schemes to be included (e.g., up to 500 species) and avoiding the need for numerical approximations such as chemical families. Box models can either represent a stationary, idealized air mass, or can be combined with a calculated air mass trajectory to produce a 'Lagrangian' model. Several chemical trajectory models can be integrated simultaneously to create a 'domain-filling trajectory model,' in which the number of model boxes is sufficient to fill a region of the atmosphere so that 3D distributions can be obtained.

Results from trajectory box models are generally valid over the time scale on which the approximation of no mixing into, or out of, the box is valid. This depends on the location in the atmosphere and may vary from a few days in the troposphere to a few weeks to months in the polar lower stratosphere.

Three-Dimensional Models

Three-dimensional (3D) models solve the chemistry on a longitude $\times$ latitude $\times$ altitude array of grid boxes. Dynamical processes are included that transfer chemical species between these fixed grid boxes in a so-called 'Eulerian' model. The chemical component of a 3D model is essentially a chemical box model. However, the high computational cost of a 3D model means that the reaction schemes have to be limited (e.g., to around 40–50 species) and some careful approximations used (e.g., families). The nature of 3D models, with their arrays of chemical box models, mean that they can be written to take good advantage of high performance vector and parallel computers. Nevertheless, when included in a 3D model, the cost of chemistry normally dominates the cost of other processes (e.g., radiative and dynamical calculations). Even in Earth System Models (ESMs) atmospheric chemistry is likely to be one of the most expensive components.

The chemistry and transport in a 3D model can be combined in an 'operator split' approach. In this method the chemical integration is separated from the dynamical integration and the advection of tracers. This decoupling of chemical and dynamical time steps is often more efficient as optimum time steps can be chosen for each process (e.g., the need for a short chemical time step does not imply an equally short dynamical time step). A multidimensional chemical model requires a module for

transporting chemical species – i.e., an ‘advection scheme.’ The advection scheme should ideally be: conservative, monotonic (no undershoots and overshoots), nondiffusive (maintain tracer gradients), and nondispersive (tracer features should advect independent of their scale). The advection scheme should also advect species independent of their concentration. In practice, essentially all of the tracer advection schemes in use compromise on one or more of these criteria.

Two types of 3D chemical models are commonly used. General circulation models (GCMs) are global radiative dynamical models used in numerical weather prediction and climate studies (see **Numerical Models: General Circulation Models**). Chemistry can be included in these models to produce a chemical GCM, allowing the calculation of the coupled effects of chemical and dynamical changes. In these coupled models, a chemically induced change in O_3 , for example, will affect the atmospheric heating rates, temperature, and therefore the dynamics. In turn, these dynamical changes redistribute O_3 . As GCMs calculate their own circulation, the results do not correspond to a specific day, but represent the typical behavior of the atmosphere. Therefore, the results of GCMs need to be compared with observations in a climatological sense. More recently, ESMs have been developed, which couple together processes that govern the evolution of the atmosphere, biosphere, oceans, and cryosphere. These models also contain atmospheric chemistry, and can be used to investigate potentially important Earth system feedback processes, for example between the biosphere and climate, driven by changes in atmospheric trace gas and aerosol abundances in response to changes in emissions from the biosphere.

Offline chemical transport models (CTMs) do not calculate the atmospheric circulation. Instead the wind (and temperature, humidity, cloud) fields are read in from another source (e.g., meteorological analyses). This has a number of advantages: the model is cheaper to run compared to the full GCM and, importantly, the model dynamics are constrained to the

‘real’ meteorological situation. This permits direct comparison between model calculations and observations. As the circulation in CTMs is fixed, they cannot be used for future predictions involving coupling of chemistry and dynamics. The meteorological analyses used to force CTMs come from weather services such as the European Centre for Medium-Range Weather Forecasts (ECMWF), U.K. Met Office (UKMO), or the National Centers for Environmental Prediction. They are produced as part of the routine weather prediction and now usually extend from the surface to the upper stratosphere. The accuracy of CTM results depends critically on the quality of these meteorological analyses and how they are used in the model. In the stratosphere the advection by the analyzed winds is usually the only transport process considered, while in the troposphere the model will usually need to parametrize ‘subgrid-scale’ transport processes such as convection and boundary layer mixing. **Table 2** illustrates the chemistry, which is included in a typical stratospheric CTM while **Table 3** illustrates a similar tropospheric model. The stratospheric model contains detailed halogen chemistry, while the tropospheric model contains more hydrocarbon species. Simulations of Arctic O_3 depletion from the stratospheric model are shown in **Figure 3**. **Figure 4** shows an example of CO distribution from the tropospheric model.

One-Dimensional and Two-Dimensional Models

Before computer power permitted the use of 3D chemical models one-dimensional (1D) and two-dimensional (2D) models were widely used for atmospheric studies. One-dimensional models represent variations of tracers with altitude and were the main tool in the 1970s and early 1980s. The models generally represent a global mean atmosphere at each layer and vertical motion is parametrized as a diffusion process. Clearly this is a gross approximation of the real atmosphere and these models are no longer used.

Table 2 Details of the SLIMCAT 3D stratospheric chemical transport model

Chemistry	Coupled short-lived species	Ox (= $O_3 + O(^3P) + O(^1D)$) NOx (= $N + NO + NO_2$), NO_3 , N_2O_5 , HNO_3 , HO_2NO_2 ClOx (= $Cl + ClO + Cl_2O_2$), $ClONO_2$, HCl , $HOCl$, $OCIO$ BrOx (= $Br + BrO$), $BrONO_2$, $BrCl$, HBr , $HOBr$ CH_2O , CH_3OOH
	Steady state	H, OH, HO_2 CH_3 , CH_3O_2 , CH_3O , HCO
	Source gases and long-lived species	CH_4 , N_2O , CO , H_2O , $CFCl_3$ (CFC-11), CF_2Cl_2 (CFC-12), $C_2F_3Cl_3$ (CFC-113), CHF_2Cl (HCFC-22), CH_3Cl , CH_3CCl_3 , CCl_4 , CH_3Br , $CBrClF_2$, $CBrF_3$ COF_2 , $COFCl$, HF
	Fixed Reactions	O_2 , N_2 , H_2 , CO_2 120 Gas phase 8 Heterogeneous 36 Photodissociation
Dynamics	Meteorology Tracer advection	From analyses (e.g., ECMWF, UKMO) Finite volume scheme
Resolution	Horizontal	Variable $10^\circ \times 10^\circ - 0.5^\circ \times 0.5^\circ$
	Vertical	Variable: 0.5–3 km
	Domain	Global: surface – 60 km
Computational	Language	Fortran (parallel) Parallel/vector machines (Inc. Workstations)

Table 3 Details of the TOMCAT 3D tropospheric chemical transport model

Chemistry	Coupled short-lived species	Ox (= O ₃ + O(³ P) + O(¹ D)) NOx (= NO + NO ₂ + NO ₃), N ₂ O ₅ , HNO ₃ , HO ₂ NO ₂ , HONO CH ₂ O, CH ₃ OOH, CH ₃ CHO, C ₂ H ₆ , C ₂ H ₅ OOH, C ₂ H ₅ CHO, C ₃ H ₈ , <i>i</i> -C ₃ H ₇ OOH, <i>n</i> -C ₃ H ₇ OOH, (CH ₃) ₂ CO, CH ₃ COCH ₂ OOH, CH ₃ COO ₂ NO ₂ , C ₂ H ₅ COO ₂ NO ₂ , CH ₃ ONO ₂ , C ₅ H ₈
	Steady state	H, OH, HO ₂ CH ₃ O ₂ , CH ₃ CO ₃ , MeCOCH ₂ OO, C ₂ H ₅ OO, <i>i</i> -C ₃ H ₇ OO, <i>n</i> -C ₃ H ₇ OO, C ₂ H ₅ COO ₂
	Source gases and long-lived species	CH ₄ , CO, H ₂ O
	Fixed	O ₂ , N ₂ , H ₂ , CO ₂
	Reactions	110 Gas phase 2 Heterogeneous 40 Photodissociation
Physics	Deposition	Dry and wet deposition
Dynamics	Emissions	Surface (natural and anthropogenic), lightning, aircraft
	Meteorology	From analyses (e.g., ECMWF, UKMO)
Resolution	Tracer advection	Finite volume scheme
	Horizontal	Variable 5° × 5°–0.5° × 0.5°
	Vertical	Variable: 0.5–3 km
Computational	Domain	Global: surface – 30 km (or higher)
	Language	Fortran (parallel) Parallel/vector machines (Inc. Workstations)

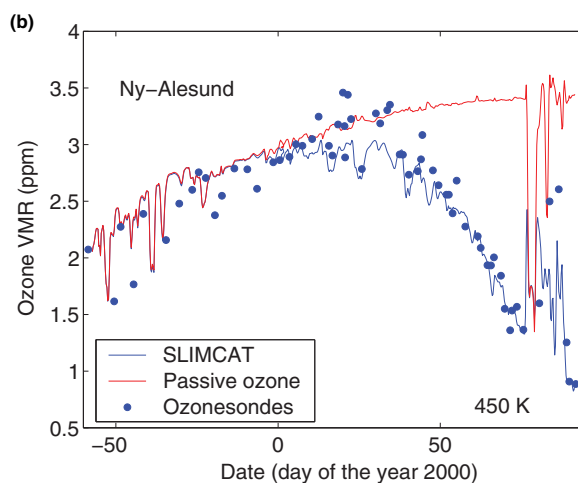
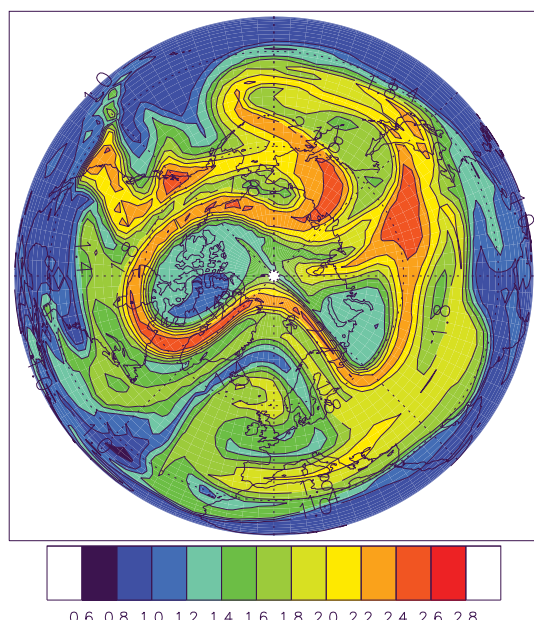
(a) SLIMCAT 3D model ozone 17 March 2000 17 km

Figure 3 Example results from the SLIMCAT three-dimensional chemical transport model (Table 2) for the Arctic winter of 1999–2000. The model was run with a horizontal resolution of 2.5° latitude × 3.75° longitude. (a) The distribution of O₃ near 17 km on 17 March 2000. The model is in good agreement with O₃ sonde observations at around 17 km (450 K potential temperature surface) at the Arctic station of Ny Alesund (78° N). (b) The difference between the model O₃ and the ‘passive’ O₃ indicates the chemical destruction since 1 January 1999. VMR, volume mixing ratio.

Two-dimensional (latitude–height) radiative dynamical chemical models calculate the zonal mean state of the atmosphere. They were the principal computational tool in the 1980–90s and are still in use. Although 2D models cannot capture the motion of stratospheric polar vortex or the longitudinal asymmetry of tropospheric surface emissions, for example, they are computationally cheap and are useful for calculating a wide range of multiyear scenarios and for sensitivity studies.

Testing Models

Given the complexities of atmospheric chemistry, and the many interactions with other processes, atmospheric chemical models can be very large programs. Much care is needed to write the code in a rigorous way to minimize the risk of errors. Generally the program should make use of all of the options available in a particular language, and on a particular machine, to test for coding errors. When a model is running it is

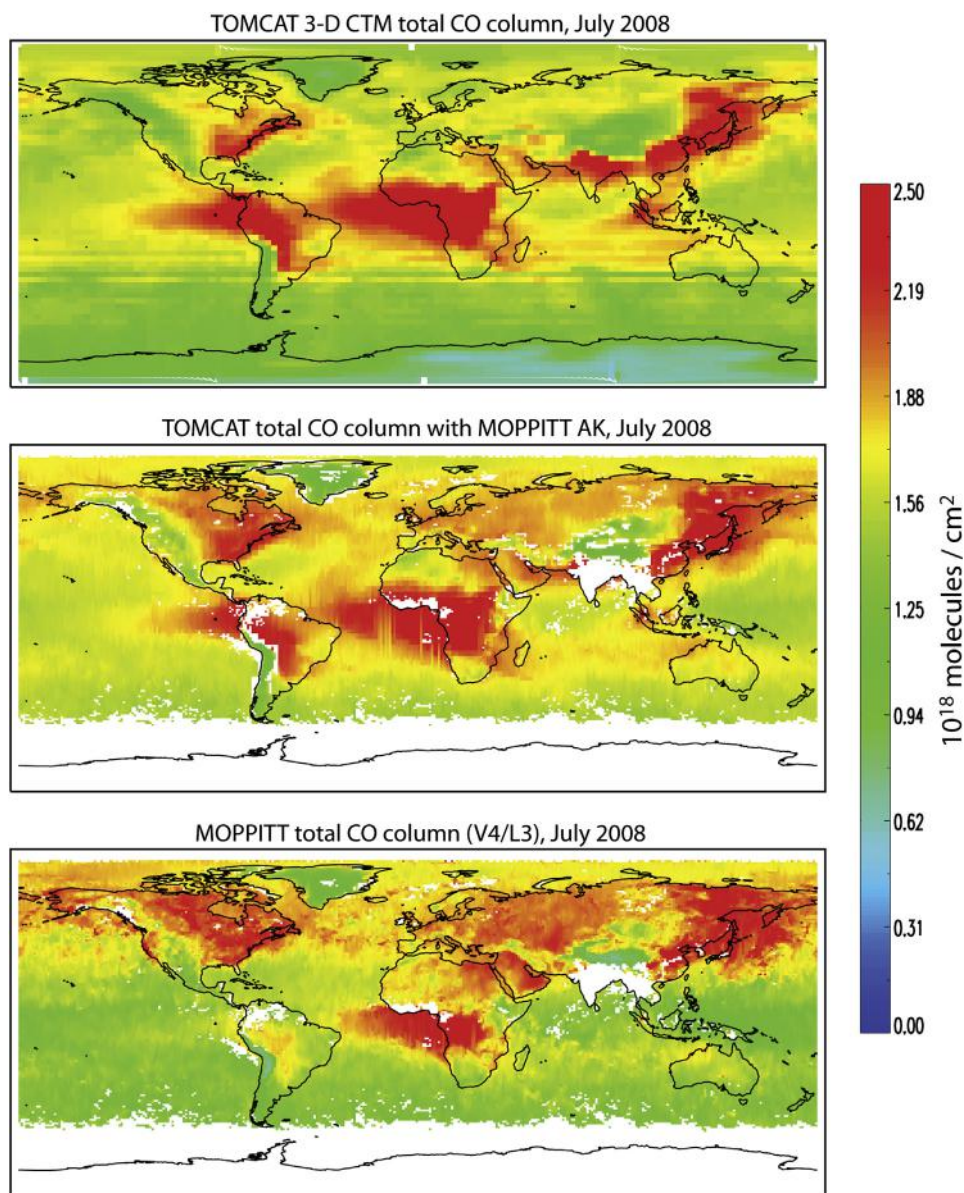


Figure 4 Example results from the TOMCAT three-dimensional tropospheric chemical transport model (Table 3) showing total column CO for July 2008 compared with satellite observations from the MOPPITT instrument. The model was run with a horizontal resolution of 2.8° latitude $\times$ 2.8° longitude. The upper panel shows direct model output while the lower panel shows the MOPPITT observations. The middle panel shows the model results sampled with the MOPPITT averaging kernels in order to give a true comparison of the model with the observations. CO is enhanced in regions of strong emission, such as the industrial regions of the northern hemisphere and in regions of tropical biomass burning. Figure courtesy of Sarah Monks, University of Leeds.

evaluated by comparison with observations to test its ability to capture processes in the real atmosphere. Periodically different atmospheric models are intercompared in international workshops to assess the uncertainties in model calculations due to differences in approach. The Stratospheric Processes and their Role in Climate program organized the chemistry-climate model (CCM) evaluation project CCMVal. This was an extensive process-based evaluation of the schemes in CCMs. All CCMs performed the same controlled experiments. Model output was compared against observations and against detailed 'benchmark' models. The performance of models was

graded. The overall result was an improved understanding of the models and a reduction of uncertainties in future predictions.

See also: **Chemistry of the Atmosphere:** Chemical Kinetics. **Numerical Models:** General Circulation Models; Methods. **Radiation Transfer in the Atmosphere:** Absorption and Thermal Emission. **Tropospheric Chemistry and Composition:** Cloud Chemistry; Volatile Organic Compounds Overview: Anthropogenic.

Further Reading

- Atkinson, R., et al., 2004. IUPAC (International Union of Pure and Applied Chemistry) evaluated kinetic and photochemical data for atmospheric chemistry: volume I – gas phase reactions of Ox, HOx, NOx and SOx species. *Atmospheric Chemistry and Physics* 4, 1461–1738.
- Brasseur, G., Solomon, S., 2005. *Aeronomy of the Middle Atmosphere*, third ed. D. Reidel Publishing, Dordrecht, Netherlands.
- Brasseur, G.P., Orlando, J.J., Tyndall, G.S. (Eds.), 1999. *Atmospheric Chemistry and Global Change*. Oxford University Press, Oxford.
- Hemispheric Transport of Air Pollution (HTAP), 2010. United Nations, New York & Geneva, Dentener, F., Keating, T., Akimoto, H. (Eds.), ECE/EB.AIR/100, ISBN: 978-92-1-117043-6.
- Jacobsen, M.Z., 2005. *Fundamentals of Atmospheric Modeling*, second ed. Cambridge University Press, Cambridge.
- Park, J., et al., 1999. *Models and Measurements Intercomparison II*. NASA Publication NASA/TM-1999-209554, Langley Research Center, Hampton, VA.
- Sander, S.P., et al., 2010. NASA/JPL, Chemical Kinetics and Photochemical Data for Use in Stratospheric Modeling. Evaluation no. 17. JPL Publication, 10–6.

SPARC, 2010. In: Eyring, V., Shepherd, T., Waugh, D. (Eds.), *Chemistry-Climate Model Validation*. SPARC Report Number 5.

WMO, 1985. World Meteorological Organization, Atmospheric Ozone. Global Ozone Research and Monitoring Project, Report No. 16, WMO, Geneva, CH 1211, Geneva 20, Switzerland.

Relevant Websites

www.cesm.ucar.edu – Community Earth System Model.
www.iupac-kinetic.ch.cam.ac.uk – IUPAC Kinetic Data.
jpldataeval.jpl.nasa.gov – JPL Kinetic Data.
mcm.leeds.ac.uk/MCM/ – Leeds Master Chemical Mechanism.
www.sparc-climate.org – SPARC.
www.see.leeds.ac.uk/tomcat – TOMCAT/SIMCAT 3-D CTM.
www.ukca.ac.uk – UKCA 3D CCM.

Coupled Ocean-Atmosphere Models: Physical Processes

M Zhang, Stony Brook University, Stony Brook, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

This article describes the concepts and parameterization methodologies of the major sources and sinks of heat, momentum, and constituents in the coupled ocean-atmosphere models, along with their limitations and uncertainties. The article also gives examples of the current performances of these models and discusses their major common biases.

The basis of coupled ocean-atmosphere models is a set of physical laws that govern the motion, temperature, pressure, density of the atmosphere and ocean and the concentrations of constituents in them: The Newton's second law governs the atmospheric winds and ocean currents; the first thermodynamic law describes the temperature change as a function of the heat sources and sinks; the law of mass conservation expresses the changes of concentration of various constituents as functions of their respective sources and sinks. These laws are written on a set of discretized grids and are averaged in the individual domains corresponding to this discretized set of grids. All sources and sinks of heat and constituents, and the effects of unresolved processes on the resolved scale fields, are referred to as physical processes in coupled ocean-atmosphere models. This article describes these processes. The differences in the parameterizations of these processes among the models are believed to be the main cause of differences in their simulations and predictions of weather, climate variability, and future climate change.

Radiative Transfer

Absorption and emission of radiation are internal sources and sinks of heat within the atmosphere and oceans. The wavelengths of radiation from the Sun span the range from alpha, gamma, and X-rays, to UV radiation, visible light, infrared radiation, extending to microwaves and radio waves. At the Sun's temperature of about 6000 K, the majority of the radiative energy from the Sun lies in the range from the UV to the near-infrared, with the peak in the visible lights of 0.4–0.7 μm

wavelengths. Radiation emitted from the atmosphere, the oceans, and the land surface is primarily in the range of infrared radiation with peaks in the band from 10 to 20 μm due to their lower temperatures. Because of the large difference of the temperatures of the Sun and the Earth and the separation of their respective radiation wavelength spectra, ocean-atmosphere models often calculate radiations from the Sun and from the Earth separately with different approximations as solar radiation and infrared radiation. They are also referred to as shortwave (solar) and longwave (infrared) radiations. Figure 1 shows the separation of radiation spectra from two emitting bodies of 6000 and 285 K representing the Sun and the terrestrial systems.

Radiative transfer in gases is highly sensitive to wavelengths due to the interaction of electromagnetic waves with the quantized molecular energy structure of gases. Absorption and emission of radiation differ for different types of gas molecules. Figure 2 shows portions of the absorption spectra in the shortwave for water vapor molecules and in the longwave for carbon dioxide molecules. These spectral properties are obtained from standard databases constructed from spectroscopy theory and laboratory measurements. Almost all atmospheric radiative transfer models use data from the HITRAN database (High-resolution Transmission molecular absorption database). HITRAN is a long-running project started by the Air Force Cambridge Research Laboratories in the late 1960s in response to the need for detailed knowledge of the infrared properties of the atmosphere. It has been continuously developed at the Harvard-Smithsonian Center for Astrophysics with participation of a large community. The HITRAN database includes spectroscopic parameters for 39 atmospheric molecules. These

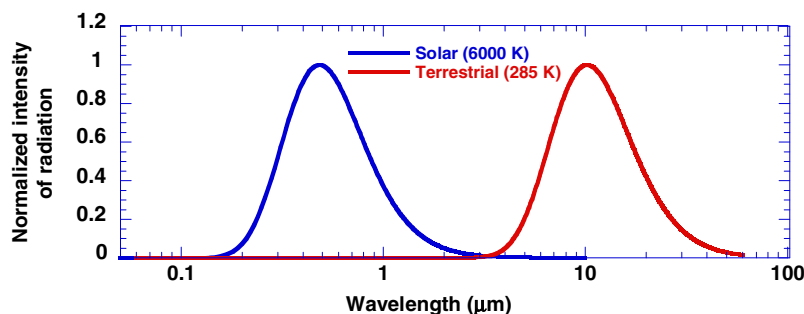


Figure 1 Wavelength dependence of normalized blackbody radiation at 6000 K (Sun) and 285 K (Earth).

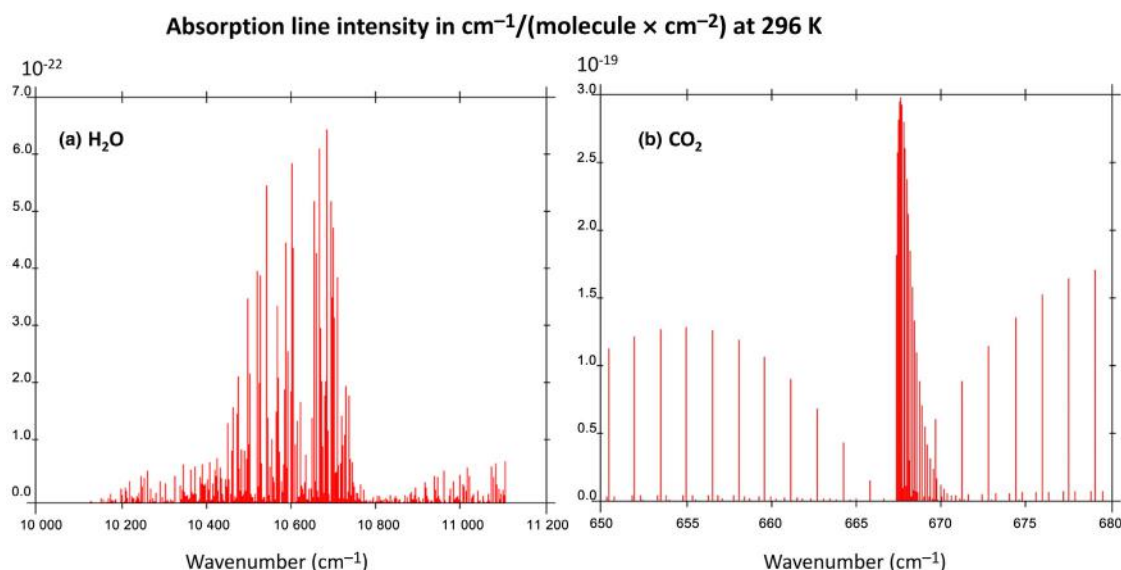


Figure 2 (a) Intensity of absorption of shortwave radiation by a water vapor molecule in the spectral range of $10\,000\text{ cm}^{-1}$ ($1\text{ }\mu\text{m}$) to $11\,200\text{ cm}^{-1}$ ($0.89\text{ }\mu\text{m}$). (b) Intensity of absorption of longwave radiation by a carbon dioxide molecule in the spectral range of 650 cm^{-1} ($15.38\text{ }\mu\text{m}$) to 680 cm^{-1} ($14.71\text{ }\mu\text{m}$).

include all molecule types that need to be considered for atmospheric radiative transfer, such as water vapor, carbon dioxide, ozone, methane, and other atmospheric trace gases.

The most comprehensive atmospheric radiative transfer models use the line-by-line calculations. They compute radiative transfer at individual wavelengths, and sum all wavelengths together to obtain the radiative energy. These models are computationally prohibitive to use in ocean-atmosphere models. Therefore, drastic simplifications are used to divide the entire shortwave and longwave spectra into no more than two dozen broad bands. These bands are selected based on the absorption behavior of all gas molecules in the atmosphere. Within each band, the atmospheric absorption, emission, and scattering are calculated by using various approximations. The most widely used approach is the so-called correlated- k method, in which the wavelength dependence of the absorptions is sorted according to the magnitudes (or g -points) into different bins; radiative transfer is then calculated in these bins, so that the wavelength dependence is minimized. The bins are summed to obtain the radiative fluxes in the bands; the bands are summed to obtain values for the whole spectra.

Radiative transfer calculation in the ocean is much simpler than that in the atmosphere. Radiation is absorbed in very short distances from the ocean surface. Only reflection and scattering at the ocean surface and absorption within about 30 m from the surface are calculated. The specification of the reflection and scattering contains uncertainties because the ocean surface is not perfectly flat. The atmospheric winds continuously induce waves of different scales at the surface. The absorption of radiation within the top layer of the ocean is also affected by the abundance of chlorophylls. The chlorophylls concentration is often specified from observational climatology with large uncertainties. Some ocean-atmosphere models do not include the treatment of chlorophylls in radiation calculation.

Radiative transfer in sea ice is treated the same as in the ocean but with different reflection, scattering, and absorption properties. Sophisticated models consider air bubbles and melted water that are trapped in the ice; they also consider the roughness of the ice surface. However, uncertainties of these variables are large.

Radiative transfer over land is calculated based on surface types. These include vegetation, snow, lake, bare soil, and urban structure. Vegetation is often categorized into various plant functional types. For each vegetation type, the areas and shapes of the leaves and the height of plants are assumed. For snow, the granule size, snow age, and area concentration of dark particle matters are needed as input. For bare soil, the color properties are specified as input. The radiative transfer calculation gives the amount of radiation that is reflected, scattered, and absorbed by the surface. Radiation in the vegetation influences the photosynthesis and canopy temperature, thus affecting water vapor evaporation and transpiration from the leaves. Uncertainties in the input data of radiative transfer over land can be large. The spatial heterogeneity of the input data is often difficult to parameterize.

Solar Radiation

Solar radiation reaching to the top of the atmosphere is calculated based on the Sun-Earth distance, the eccentricity (orbital shape of the Earth), the procession (the rotation of the tilted axis of the Earth), and the obliquity (the angle between the Earth's self-rotation axis and the orbital axis). Solar activities such as the 11-year solar cycle are considered in most models. The absolute amount of the output of energy from the Sun is not precise. Current satellite measurements give a range of mean incoming radiation to the top of the atmosphere, or the solar constant, from about 1370 to

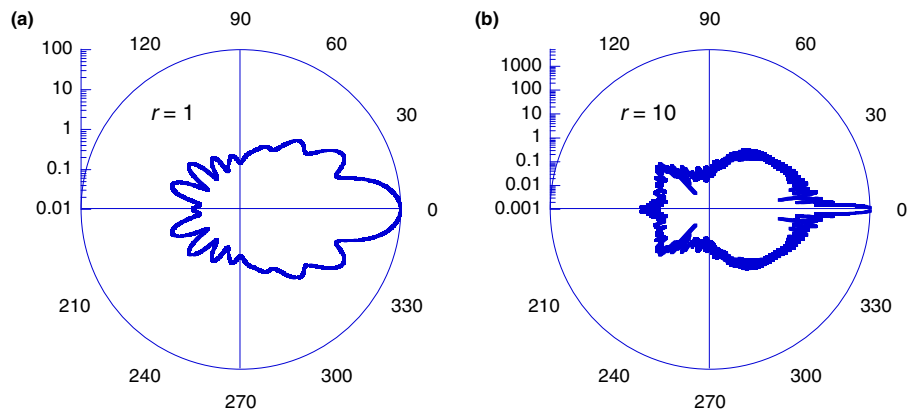


Figure 3 Angular distribution of scattered radiation of an incident visible light at $0.7 \mu\text{m}$ wavelength from the left by a liquid water particle of radius (a) $1 \mu\text{m}$, (b) $10 \mu\text{m}$.

1365 W m^{-2} . Different models may use slightly different values of the solar constant.

In the transfer of solar radiation, internal emitting sources of radiation within the atmosphere can be neglected. This is a convenient simplification. However, because the wavelengths of solar radiation are prone to strong interactions with atmospheric aerosol particles and cloud particles with sizes of $1\text{--}10 \mu\text{m}$, solar radiation is strongly scattered by these particles. The scattered radiation is further scattered by the particles, leading to multiple scattering processes until they are absorbed by the particles, atmosphere, the surface, or are scattered to outer space.

Scattering of radiation by particles is highly dependent on the size of the particles and the wavelengths of the radiation as well as the electromagnetic properties of the particles. This affects the total amount of scattered and absorbed radiation, and the directional distribution of the scattered radiation. **Figure 3** shows an example of the normalized directional distribution of scattered radiation by a cloud particle in liquid phase at two wavelengths. Because of these wavelength dependences and because of the multiple scattering processes, the calculation of solar radiation involves integration across space angles. Fortunately, the multiple scattering processes smear out some of the angular inhomogeneity, so many of the bulk simplifications, such as the use of two streams to describe the directional distribution, and the averaged optical properties of the particles, can provide accurate calculation of scattered radiations.

The largest uncertainties in the calculation of the solar radiation are due to the input data, particularly cloud particles and atmospheric aerosols. Cloud processes are crudely parameterized in current ocean-atmosphere models (see later Section on **Clouds**). Atmospheric aerosols include many types; their concentration, sizes, shapes, and time evolution represent some of the largest uncertainties in radiative calculations.

Current ocean-atmosphere models all use the plane-parallel assumption in which the radiation and the atmosphere within a grid are horizontally homogeneous. Three-dimensional radiative transfer calculations have indicated that lateral radiative fluxes may reach tens of watts per square meter at the cloud scales. It is not clear whether these

lateral fluxes are important in high-resolution climate models. Three-dimensional radiative transfer calculations will incur significantly more computational costs. It is unlikely that they will be implemented in climate models in the foreseeable future.

Infrared Radiation

The surface and the atmosphere emit and absorb infrared radiation. Even though these are internal sources, calculation of infrared radiative transfer is much simpler than that of solar radiation. This is because at the wavelengths of infrared radiation, scattering by cloud particles and the majority of aerosols can be neglected.

Scattering of infrared radiation can occur on large aerosol particles. These are typically not considered in current ocean-atmosphere models. The large particles have short residence time in the atmosphere due to fast gravitational sedimentation, so the errors of neglecting them are small. Scattering of infrared radiation can also occur on cloud particles, but because absorption of infrared radiation by liquid or ice particles is large in this wavelength range, radiation is absorbed after a few scattering events, so scattering can be safely ignored.

Clouds

Clouds strongly regulate the energy balance of the atmosphere-earth system. Cloud particles reflect solar radiation to space, acting to cool the planet; they also strongly absorb infrared radiation emitted by the surface and by the below-cloud atmosphere, acting to trap infrared radiation and warm the planet. How clouds vary in response to climate changes, including variations in their area coverage, thickness, altitude, and liquid and ice water content, is the subject of intensive research in the last three decades as the cloud-climate feedback problem. Clouds are part of the hydrological cycle of the atmosphere-earth system. They are a manifestation of the phase changes of water between vapor, liquid, and ice in the atmosphere.

Cloud processes in ocean-atmosphere models are typically represented by two components. One is the cloud macrophysics; the other is cloud microphysics. The cloud macrophysical component calculates the fractional area coverage of clouds, spatial inhomogeneity of cloud properties, grid-scale condensation and evaporation, sublimation and vaporization when only a fraction of the grid domain is occupied by clouds. The macrophysical parameterization is often based on empirical relationships and assumptions on the subgrid scale distribution of total water. These empirical relationships or assumptions are approximate. They are often made differently for different types of clouds. They should be resolution dependent, but most current models are not designed to account for their dependences on resolutions.

The cloud microphysical component calculates the time evolution of the condensed cloud mass, particle number concentration, and their sources and sinks. Some cloud microphysical models only calculate the mass amount of condensed water. These models are called one-moment models. Models that include both the mass amount and number concentration are called two-moment models. To calculate the impact of aerosol on cloud microphysics, two-moment schemes are needed. Therefore, most current generation ocean-atmosphere models use two-moment schemes. Sources and sinks of cloud mass are calculated based on grid-scale condensation or sublimation, evaporation or vaporization, conversion between cloud drops and raindrops and snow, scavenging by rain and snow, breakout of raindrops or snowflakes. Sources and sinks of number concentration are calculated based on nucleation, evaporation, and vaporization of raindrops and snowflakes, and their breakup and removal by rain and snow. Figure 4 shows an example of the processes in a two-moment scheme.

Current models treat ice clouds and mixed phase clouds very crudely. Ice clouds rely heavily on the presence of ice nuclei,

whose concentration and nucleation properties are very uncertain. Ice crystals form in different shapes as functions of the ambient air conditions and ice nuclei. Different shapes have different density and falling speed. The most common practice in current models is to use a ramp function of temperature to calculate whether the clouds are liquid, ice, or mixed phase.

Precipitation

Precipitation processes include the formation and changes of raindrops, snowflakes, and their evolution in the falling process to the ground. Some models also include categories of graupels and hails. Similar to cloud microphysical models, if a model only calculates the precipitation mass, it is called a one-moment scheme; if a model calculates both the mass and number concentration of precipitation particles, it is called a two-moment scheme. The separation of cloud particles and precipitation particles is based on observational support that the size distributions of these two types of particles are typically well separated, and so they have very different falling speeds.

Formation of precipitation is calculated from conversion of cloud drops and ice crystals, and collection of cloud particles by the larger falling precipitation particles. The conversion occurs because cloud particles have different sizes and different sedimentation velocities. Large particles fall faster than smaller particles, and so the small particles are captured by large particles. Since the size distribution of cloud particles in current models is approximate, the formulation of the cloud-to-precipitation conversion contains many empirical parameters.

One efficient mechanism for precipitation particles to grow occurs in mixed phase clouds. The saturation vapor pressure over water is greater than over ice. In mixed phase clouds, water vapor is saturated with respect to liquid droplets, but super-saturated over ice particles. As a result, liquid particles can

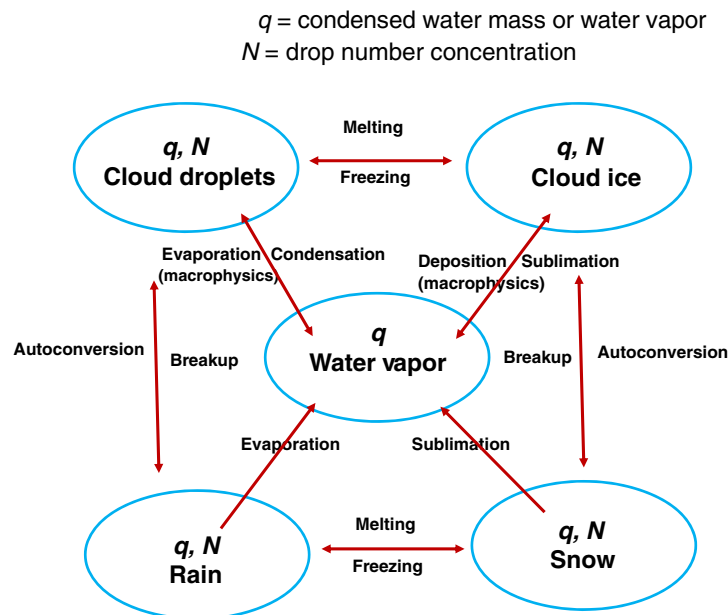


Figure 4 Schematics of some of the cloud and precipitation microphysical processes.

rapidly evaporate to sublimate to the ice particles to make them grow. This process, also called Bergeron process, is calculated in the precipitation parameterization. It also applies to ice crystals falling into supercooled liquid clouds.

Besides their formation, precipitation particles evolve along their falling path before reaching the ground. They accumulate cloud particles or smaller precipitation particles; they also experience evaporation or vaporization and breakup. The accumulation depends on the falling speed of the precipitation particles, which require information of their sizes and shapes. For raindrops, spherical shape can be safely assumed, but for snowflakes, since their shapes are highly uncertain and are crudely parameterized, the falling speed is approximate. The evaporation and vaporization of falling precipitation particles depend on the relative humidity of the ambient air, often taken as the mean condition of the grid domain, which may not be accurate.

Turbulent Mixing

Turbulent transports or mixing of heat, momentum, and constituents refer to the effects of subgrid scale processes on the resolved-scale fields. Because these subgrid scale processes operate at all scales smaller than resolved-scale fields, even in high-resolution models with grid spacing of several kilometers, turbulent transports still need to be parameterized. Turbulent fluxes are an essential part of any ocean-atmosphere model. For example, the turbulent transport of water vapor from the surface to the atmosphere is equal to the global precipitation.

Equations of turbulent fluxes can be formulated if higher order subgrid transport terms are known. The calculation of these high-order terms requires even higher order terms. The equations are therefore not closed and approximations are made to parameterize some terms. The simplest type of turbulence parameterization is a diffusion scheme, in which the turbulent flux is represented by the vertical gradient of a field multiplied by a diagnostically calculated diffusivity. These schemes are called first order model. The diffusivity is typically a specified function of vertical static stability and wind shear, or the dimensionless Richardson number. In some models it is calculated as a prescribed function of height, with a peak in the middle of the boundary layer to mimic results from large-eddy simulation models. Some models also add a counter-gradient term to the down-gradient diffusion term to account for large-eddy transport. This term is parameterized based on the source of large eddies such as surface buoyancy fluxes.

Some models prognostically calculate the turbulent kinetic energy (TKE) based on buoyancy flux, shear production, transport, and dissipation of eddy energy. This energy is then used in the calculation of the diffusivity. Other models may diagnostically calculate TKE. If the TKE is diagnostically calculated, the scheme is still first order; if it is prognostically calculated, the scheme is often called one-and-half order, since not all of second order moments are prognostically calculated. Very few models use second order or higher order closure schemes.

In the atmosphere, two separate types of turbulent transports are typically parameterized. One is for the free

atmosphere that is away from the atmospheric boundary layer (ABL); the other is for the ABL. There are no fundamental differences in the philosophy of their parameterizations, except for the degree of simplifications. In the ABL, most models first judge whether the layer is a stable boundary layer or unstable boundary layer. Different closure assumptions are used based on the stability. Convective boundary layer, in which the surface upward buoyancy flux is positive, is the most common type ABL.

In the ocean, the boundary layer turbulence is driven by wind mixing, shear of ocean currents, and static instability. The static instability can be caused by cooling of ocean surface and salinity change. The parameterization of oceanic turbulent fluxes in the vertical direction uses similar approaches as that for the atmosphere. In the horizontal direction, however, turbulent mixing in the ocean has two more complications than that in the atmosphere. First, because of the long timescale of ocean circulations and the large role of mesoscale eddies, horizontal turbulent mixing cannot be neglected. Earlier models parameterize the horizontal diffusivities of small and mesoscale eddies on the native model horizontal surfaces. Current generation models formulate them along the isopycnal or density surfaces. A second complication is the turbulent mixing along the lateral boundaries. The same equations as for the vertical boundary layer are often used, in which the turbulence is primarily driven by horizontal shear of ocean currents.

Surface Fluxes

Surface fluxes include shortwave and longwave radiative fluxes, and the heat, momentum, constituent fluxes across the interfaces between the atmosphere and the ocean or land. The radiative fluxes are calculated from the radiative transfer parameterizations described in previous sections. The turbulent heat, momentum, and constituent fluxes across the surface are calculated using similarity theory instead of the eddy diffusivities because the turbulence at the surface is strongly confined by the surface geometry.

In the similarity theory, a length scale – the Monin-Obukhov length scale – is used to normalize the height in the surface boundary layer. Empirical relationships of the vertical profiles of the winds, temperature, and constituents are obtained with respect to this normalized height based on observations. The surface fluxes appear as constant parameters in the profile relationships and in the Monin-Obukhov length scale. Vertical integration of these profiles from the surface to a reference height gives the formula to calculate the surface fluxes. The most widely known forms are

$$\tau = -\rho_{\text{air}} C_D |V_{10}| v_{10}$$

$$H = \rho_{\text{air}} C_H |V_{10}| (T_s - T_2)$$

$$E = \rho_{\text{air}} C_E |V_{10}| (q_s - q_2)$$

where τ , H , E are surface momentum stress acting on the atmosphere, surface heat flux, and constituent (such as water vapor) fluxes respectively; ρ_{air} is the density of air; V_{10} is the

wind at 10 m height; T_s , q_s are temperature and the constituent concentrations at the surface; T_2 , q_2 are temperature and constituent concentrations at 2 m; C_D , C_H , C_E are the bulk transfer coefficients of momentum, heat, and constituents. These coefficients are dependent on the fluxes themselves, so iterations are needed. In simple calculations or applications, however, they are often specified as constants.

At the ocean–atmosphere interface, net flux of freshwater is needed to calculate the salinity. The freshwater flux to the ocean is calculated by simply taking the difference of precipitation and surface evaporation.

Convection in the Atmosphere and Oceans

Convection is one type of turbulence, but because it is highly anisotropic in the vertical and horizontal directions, it is always separately parameterized. Convection schemes are used to calculate the convective transport of heat, momentum, constituents, and their sources and sinks within the convective portion of a grid cell.

Most atmospheric convection schemes use mass-flux plume models. These schemes parameterize the triggering condition of convection that determines when and where convection is activated. They also parameterize the entrainment and detrainment rates of the plume mass, the vertical distribution of the convective mass fluxes, and the cloud microphysical processes within the convective plumes.

The triggering condition of convection always includes atmospheric conditional instability as one criterion. However, models differ in using additional factors, such as the origination level of convection, whether the instability is calculated for undiluted or diluted air parcels, whether convection can be inhibited by a layer of negative buoyancy, and whether resolved-scale dynamics are considered. Lateral entrainment and detrainments can also differ greatly among different models: some use specified rates; others calculate them as functions of humidity of the ambient air; some even ignore them altogether. Convection mass fluxes are often calculated using closure assumptions, which are subject to considerable uncertainties. Cloud microphysical processes within convection are typically crudely parameterized such that when the condensed water exceeds a threshold value, the excess amount is treated as precipitation. These assumptions represent some of the largest uncertainties in current ocean–atmosphere models.

Most current convective parameterizations are diagnostic. They are not passed from one time step to another time step. Almost all schemes are designed for models with grid sizes that encompass an ensemble of cumulus clouds. As high-resolution models become more popular, these convection schemes may need structural improvements to correctly reflect their temporal and spatial effects on resolved-scale motions.

Atmospheric convective parameterizations often include separate schemes for shallow convection and for deep convection. The essential elements of these two types of schemes are the same. The differences are due to the different depth of the instability layer so that different assumptions are used for the entrainment and detrainment rates. In addition,

because deep convection is typically thicker and is with larger instability, it is associated with strong precipitation, whose evaporation causes downdrafts.

Convection parameterization in the ocean typically uses simple adjustment schemes. Vertically unstable water column is adjusted to neutral stratifications. Convection is often initiated because of density anomaly at the ocean surface, which can be caused by radiative or evaporative cooling.

Gravity Waves

Most models parameterize the effects of internal gravity waves on the resolved-scale flow. These waves cannot be resolved by the large-scale models because of their small vertical and horizontal wavelengths. These gravity waves are calculated by using information of terrain, jets, or frontogenesis in highly idealized forms. The calculation also considers saturation of gravity waves for which the vertical wavelength is small enough to cause static instability. Momentum transport by the gravity waves is then obtained and vertically differentiated to obtain the momentum wave drag.

It has been shown that the momentum wave drag has a large impact on the mean zonal wind of the stratosphere and the temperature in polar regions. Large gravity wave drag slows down the westerly jet in the stratosphere and warms the polar stratosphere. Gravity wave parameterizations often contain many empirical parameters that are subject to large uncertainties.

Aerosols

Modern ocean–atmosphere models all include aerosols as prognostic variables. Aerosols not only affect radiation directly but also impact the cloud condensation nuclei and therefore clouds. These two effects are called aerosol direct and indirect effects on climate. The role of anthropogenic aerosols – aerosols caused by human activities – in past and future climate change has been a subject of intensive research in the last two decades.

Aerosols in the models are typically categorized into groups. These often include sulfate, ammonia, sea salt, dust, black carbon, and organic carbon. The mass amount and number concentration of each group are calculated. Assumptions are made on the size distributions of aerosols and how they are mixed, since these determine the radiative properties, water activity, and chemical reactions of the aerosols.

The parameterizations of aerosols include emissions of both natural and anthropogenic sources, such as dust and sea salt from natural emissions, sulfate and ammonia from anthropogenic sources. Aerosols are also calculated to evolve with time according to gas-particle conversion, coagulation, and chemical aging in addition to transport. The gas-particle conversion involves many types of precursor gases that originated from organic sources, many of which are not well known. Sinks of aerosols are removals by dry and wet depositions.

Superparameterizations

Superparameterization refers to the use of cloud-resolving models as a substitute of some physical parameterization components (cloud macrophysical and convection schemes) in large-scale atmospheric models. Since cloud-scale dynamics are resolved and the clouds are calculated based on these resolved-scale fields, superparameterization is superior to traditional parameterizations of cumulus convection and cloud processes. The disadvantage is the more expensive computational cost.

Current superparameterizations use spatial grid sizes of about 1 km in the cloud-resolving models. Boundary-layer turbulences and cloud microphysics still rely on parameterizations. Superparameterizations have been only recently used in coupled ocean-atmosphere models for limited amount of simulations.

Performances of Ocean-Atmosphere Models

When forced with incoming solar radiation, current ocean-atmosphere models are able to simulate earthlike global distribution of surface temperature, vertical distribution of atmospheric winds and ocean currents. They can also simulate cyclones, storm tracks, stationary waves, and interannual variability of tropical ocean temperature that resembles the observed El Niño events. As two examples, Figure 5 shows the simulated annually averaged global distribution of sea-surface temperature by the Community Earth System Model Version 1

and its comparison with observation. Figure 6 shows the simulated atmospheric zonally averaged eastward winds in the December–February months and its comparison with observation. The model is able to reproduce most of the important observational features in the ocean surface temperature and in the atmospheric winds.

Ocean-atmosphere models have been also used to simulate past climate change and to make projections of future climate. To calculate past changes, the models are given time-dependent forcing fields of greenhouse gases such as CO₂, solar variability, volcanic forcing, anthropogenic aerosol changes, and land use and land cover changes. Most models are able to simulate the global temperature increase in the twentieth century. These results have been summarized in the past reports of the Intergovernmental Panel on Climate Change.

Common Biases

There are several common biases in virtually all coupled ocean-atmosphere models that researchers have been trying to solve in the last several decades. One is the so-called double ITCZ syndrome. It refers to a double intertropical convergence zone (ITCZ) simulated by the models in the annual mean precipitation in the central Pacific that is not present in observations (Figure 7). Higher spatial resolutions do not seem to improve this aspect of the model simulations. The second long-standing problem is the overall failure of most models in simulating the Madden-Julian oscillation, which is an

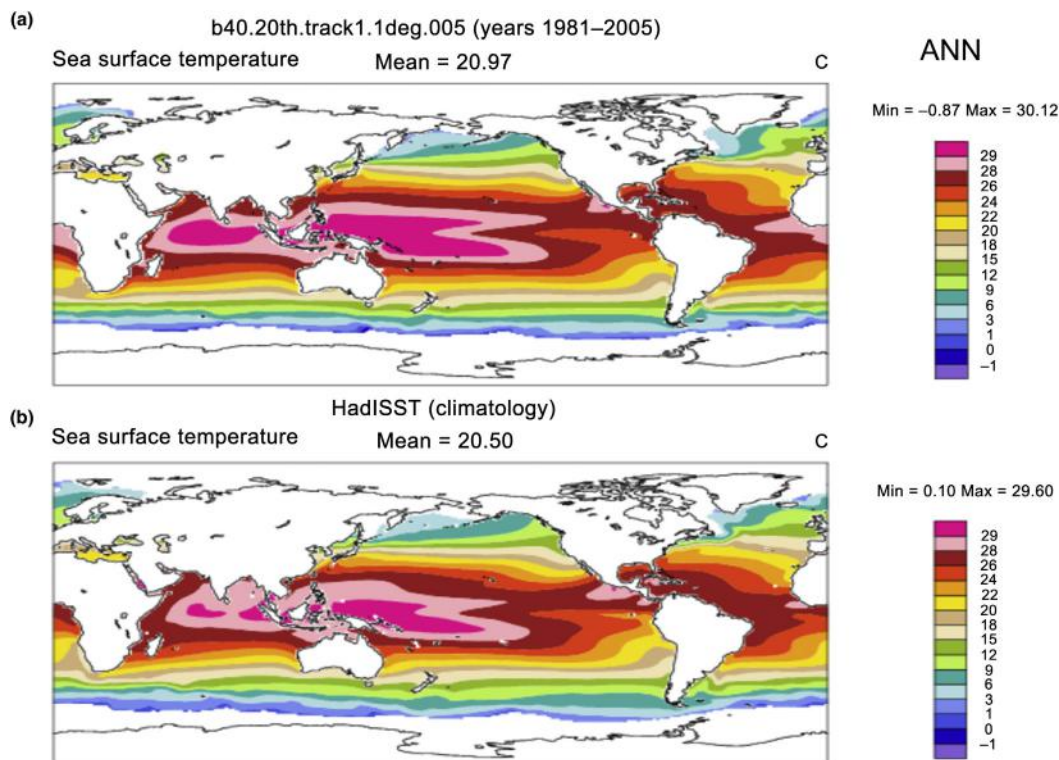


Figure 5 Annual mean sea surface temperature: (a) simulation by the Community Earth System Model Version 1; (b) observation. ANN refers to annual averages.

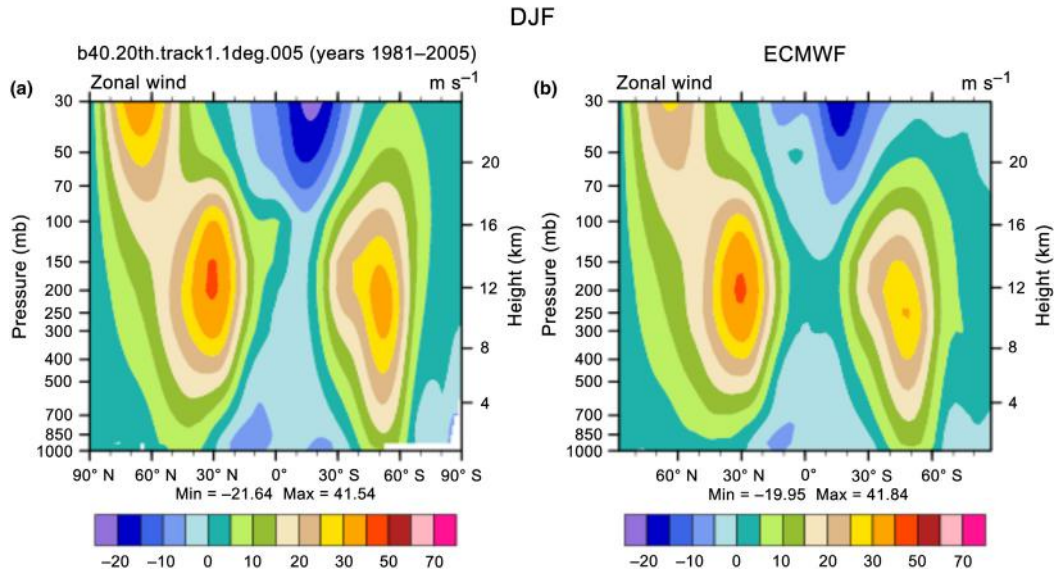


Figure 6 Height–latitude cross-section of eastward atmospheric wind averaged over all longitudes: (a) simulation by the Community Earth System Model Version 1; (b) observational estimates. DJF refers to December–January–February season. ECMWF refers to observational reanalysis at the European Center for Medium Range Weather Forecasting.

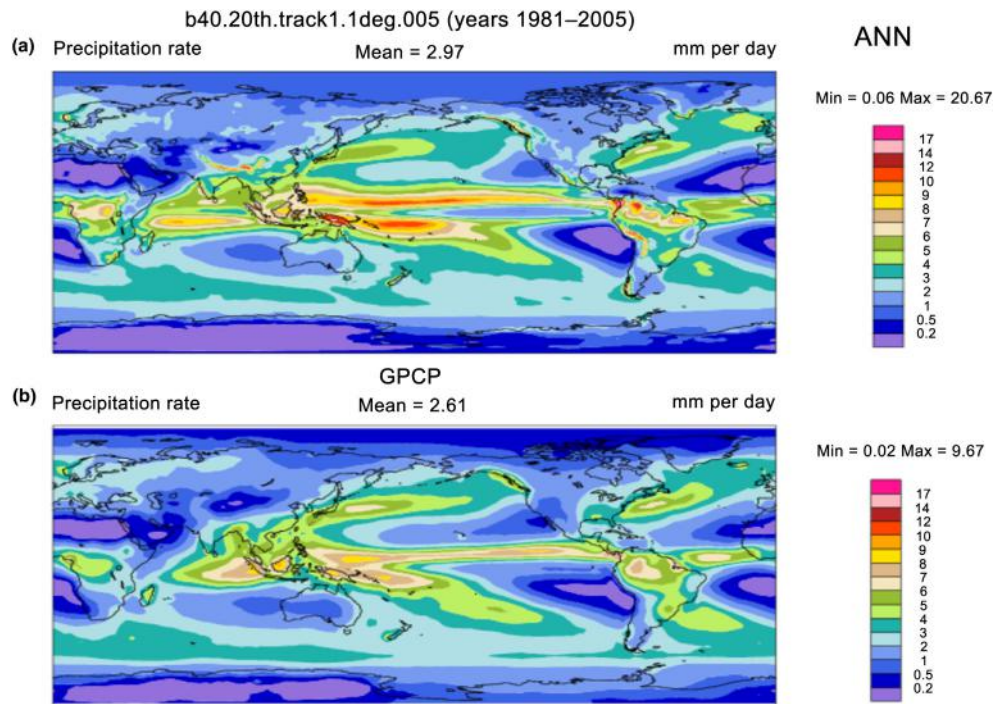


Figure 7 Annual mean precipitation: (a) simulation by the Community Earth System Model Version 1; (b) observation from the Global Precipitation Climatology Project (GPCP).

intraseasonal propagation of atmospheric circulation anomalies and convection from the equatorial eastern Indian Ocean to the western Pacific with a period of about 30–60 days. This intraseasonal oscillation has been shown to affect weather in many regions of the globe. The third common bias is the misrepresentation of the diurnal variation of precipitation over

the oceans. Models simulate peak precipitation at noontime, but observations show peak precipitation during the night over most regions of the oceans. All these biases are believed to be due to inaccurate physical parameterizations in the models.

Individual models may have other significant biases, including the simulations of El Niño and its impact, monsoon

rainfall, the deep ocean circulation, and land surface temperature. All these biases are being actively studied in the modeling centers.

See also: **Aerosols:** Role in Radiative Transfer. **Air Sea Interactions:** Momentum, Heat, and Vapor Fluxes. **Boundary Layer (Atmospheric) and Air Pollution:** Modeling and Parameterization; Ocean Mixed Layer. **Clouds and Fog:** Cloud Microphysics. **Land-Atmosphere Interactions:** Overview. **Numerical Models:** Cloud System Resolving Modeling and Aerosols; General Circulation Models; Model Physics Parameterization; Parameterization of Physical Processes: Clouds; Parameterization of Physical Processes: Gravity Wave

Fluxes; Parameterization of Physical Processes: Turbulence and Mixing. **Radiation Transfer in the Atmosphere:** Absorption and Thermal Emission; Cloud-Radiative Processes; Scattering. **Turbulence and Mixing:** Overview.

Further Reading

Randall, D.A., 2000. General Circulation Model Development, Past, Present and Future. Academic Press, p. 807.

General Circulation Models

CR Mechoso and A Arakawa, University of California, Los Angeles, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

General circulation models (GCMs) are essential tools for climate studies. The history of the general circulation modeling is broadly divided into three phases characterized, respectively, by the rapid development of theories on large-scale atmospheric motion, development of early GCMs, and expansion of internal processes and use of higher resolutions. The scope of general circulation modeling is outlined. The basic approximations, boundary conditions, discretization of the governing equations, and the problem of parameterization of physical processes are addressed. Current trends toward unification and the reduction of the dependency of model physics on grid size are mentioned.

Introduction

General circulation of the atmosphere refers to the time-averaged, planetary scale motion representing the atmosphere's long-term statistical behavior. General circulation models (GCMs) numerically integrate the equations of fluid motion to simulate the evolution, maintenance, and variations of the general circulation. GCM applications, for example, climate simulations or predictions of the second kind, do not generally emphasize the dependence of the solution on initial conditions. A comprehensive GCM can also be used as an extended numerical weather prediction (NWP) model. In NWP applications, or climate predictions of the first kind, initial conditions are crucially important.

The History of General Circulation Modeling

The start of general circulation modeling is hardly distinguishable from that of NWP. V. Bjerknes first advocated the idea of NWP in the early twentieth century. L.F. Richardson carried out a pioneering albeit unsuccessful NWP attempt in 1922. The history of general circulation modeling that would follow can be divided into a prelude and three phases (I, II, and

III, see Figure 1). The prelude was characterized by the rapid development of quasigeostrophic theories on large-scale atmospheric motions in the late 1940s.

Phase I began in 1950 when Charney, Fjörtoft, and von Neumann obtained successful 24-h forecasts of 500 hPa geopotential height using the two-dimensional quasigeostrophic vorticity equation. This success demonstrated the relevance of a simple dynamical model for daily weather changes, and dynamic meteorology and synoptic meteorology began to merge. Another epoch-making event during Phase I was the recognition that the dynamics of 'cyclones' and that of the 'general circulation' were closely linked. The numerical experiment performed by Philips in 1956 using a quasigeostrophic two-level model with prescribed heating highlighted those links. In the experiment, the large-scale components of initially random disturbances grew due to baroclinic instability, modifying the general circulation from the zonally symmetric regime to the wave regime. Correspondingly, the zonally averaged meridional circulation in the middle latitudes changed from the Hadley type to the Ferrel type, producing midlatitude surface westerlies. Philips's 1956 experiment, therefore, captured fundamental features of the observed general circulation of the atmosphere, whose causes had been more or less a matter of speculation.

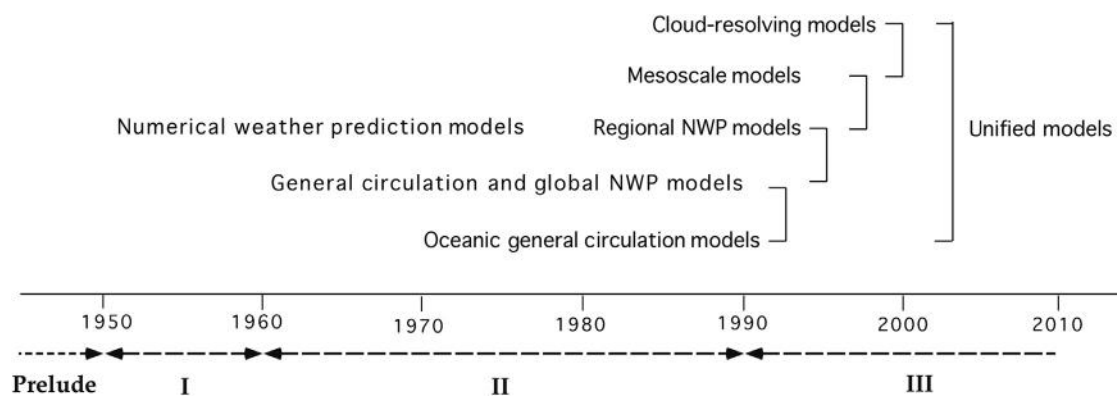


Figure 1 A chart showing the history (and near future) of numerical modeling of the atmosphere.

The beginning of Phase II of general circulation modeling roughly corresponds to the development of early GCMs, including those at the Geophysical Fluid Dynamics Laboratory (GFDL), University of California, Los Angeles (UCLA), Lawrence Livermore National Laboratory (LLNL), and National Center for Atmospheric Research (NCAR). Major changes from Phase I to Phase II are listed below:

- The use of the primitive equations (i.e., the governing equations without quasigeostrophic or similar approximations on the balance between the mass and velocity fields) became standard;
- Computational difficulties associated with the use of the primitive equations and their long-term integration were essentially overcome;
- Heating was made the result of motion, as well as the cause of motion, including water-vapor mixing ratio as a standard prognostic variable; and
- The importance of parameterized representation of subgrid processes, cumulus parameterization in particular, was clearly recognized.

Development of these early GCMs stimulated the meteorological community to look into the feasibility of a global observation and analysis experiment. The Global Atmospheric Research Program (GARP) then followed at an unprecedented scale in the atmospheric sciences. GARP tremendously stimulated and supported worldwide efforts in general circulation modeling almost throughout Phase II. During this phase, the scope and diversity of numerical modeling of the atmosphere models, including the development of oceanic GCMs and atmospheric models, focused on smaller domains/scales as shown in Figure 1.

The history of general circulation modeling is now in Phase III (see Figure 1). The beginning of this phase is marked by the development of coupled atmospheric and oceanic GCMs. More details on developments during Phase III are presented after a general discussion on the scope of general circulation modeling.

The Scope of General Circulation Modeling

The principal energy source for the general circulation is solar radiation, which influences the atmosphere rather remotely. The various heat components in the atmosphere (and in the climate system) are produced by complex interactions among the processes shown in Figure 2. (Here Large-Scale means scales larger than convection and turbulence scales, and Dynamical Processes includes associated adiabatic thermodynamical processes.) The horizontal scales of different atmospheric phenomena are shown in Figure 3. (Here the arrow indicates the smallest scale explicitly resolved by typical contemporary GCMs.) Although there is a trend to use higher resolutions as computer technology advances (dashed arrow in Figure 3), the collective effects of unresolved processes cannot be ignored and, therefore, they must be formulated in terms of the resolvable-scale prognostic variables unless the resolution is sufficiently high to explicitly simulate individual clouds. This is the 'parameterization problem,' which represents a crucial part of general circulation modeling.

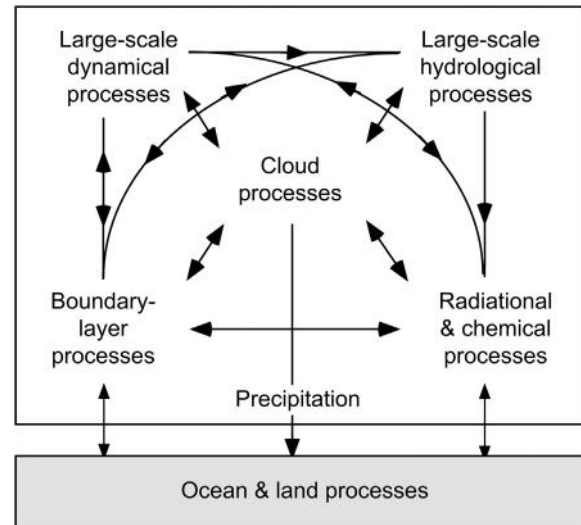


Figure 2 Basic processes in the climate system and their interactions.

Basic Approximations and Upper and Lower Boundary Conditions

Most GCMs use the 'quasistatic approximation,' in which the hydrostatic equation replaces the vertical component of the momentum equation and Coriolis and metric forces that depend on the vertical velocity are neglected in the horizontal component of that equation. The approximation eliminates vertically propagating sound waves, and is justifiable when motions have a sufficiently large horizontal scale (say, $\gg 15$ km with typical static stability). In the 'primitive equations,' the quasistatic approximation is the only major simplification on dynamics.

When the model top corresponds to the top of the atmosphere, it is reasonable to assume that there is no vertical mass flux across the upper boundary. This condition can be artificial, however, even when the upper boundary is formally placed at the top of the atmosphere, because of practical limitations on vertical resolution. At present, no simple but fully justifiable way of handling the upper boundary is available for general use in nonlinear discrete models.

For a horizontally discrete model, the effective height of the surface at the model's lower boundary may be higher than the actual height averaged over the grid box. The additional height depends on the subgrid-scale variance of orographic height. Considerations of this sort lead to the idea of 'envelope orography.' Most GCM applications during Phase II prescribed the geographical distribution of sea surface temperature (SST).

Discretization of the Governing Equations

In numerical modeling, the governing partial differential equations are replaced by a finite number of algebraic equations. Such a 'discretization' may seriously distort certain dynamical aspects of the continuous system even when a reasonably high resolution is used. The appropriate handling of this aspect is particularly important in GCM applications that require long-term integrations.

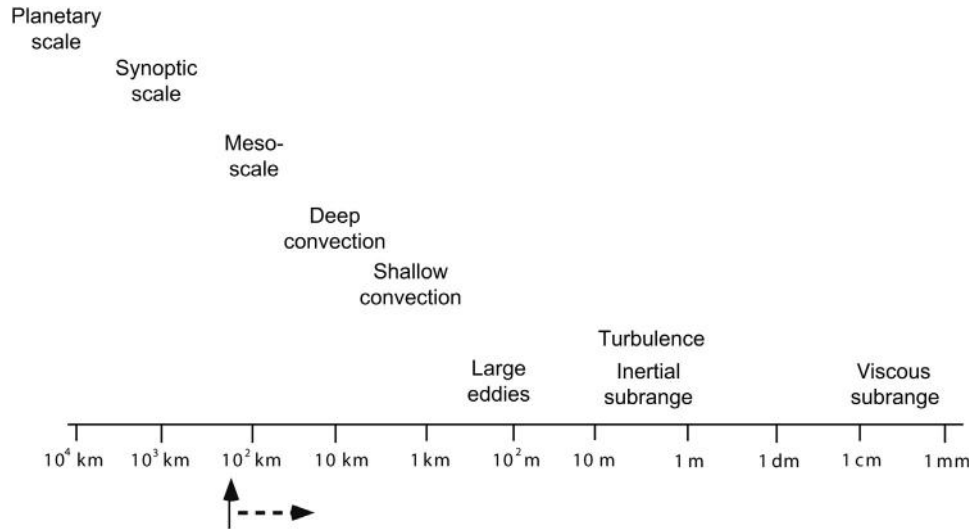


Figure 3 Horizontal scales of typical atmospheric motions.

Basic Terminology

When the Eulerian expression of the material time derivative ($D/Dt = \partial/\partial t + \mathbf{V} \cdot \nabla$) is used, discretization is made in space and time. Space discretization can be done using either *finite-difference/finite-volume methods* or *spectral methods*. In the former (or grid point) methods, differential operators in the equations are replaced by difference operators both in the horizontal and in the vertical. In spectral methods, the dependent variable is replaced by a function expressed in the horizontal as a *finite* series with global basis functions. Spectral methods are usually combined with the *transform method*, in which products in nonlinear terms are calculated at grid points in the physical space rather than as interactions of waves.

If variables at a new time level appear not only in the finite-differenced time derivatives, the scheme is *implicit*; otherwise it is *explicit*. *Semi-implicit* schemes, in which implicit expressions are used in selected terms in the prognostic equations, such as the pressure gradient force in the horizontal momentum equation, are often used to allow for longer time intervals while maintaining computational stability. A semi-implicit scheme is almost always used in the spectral method.

The *semi-Lagrangian* method is becoming increasingly popular for the advection (or the advection part of) equation. Removing the restriction on the time interval due to the Courant–Friedrichs–Lewy (CFL) condition is the major motivation for the Lagrangian approach. Semi-Lagrangian schemes consider trajectories whose arrival points coincide with grid points, which are fixed in space, while values at departure points are determined by interpolations from grid points. These schemes are generally combined with the semi-implicit method.

Vertical Coordinates and Grids

The choice of vertical coordinate can make substantial differences in a discrete system. Most existing GCMs use the σ coordinate (σ is the pressure divided by surface pressure and hence

at the lower boundary $\sigma = 1$) or a hybrid σ – p coordinate, where p is the pressure. The lower boundary condition is simple with a coordinate of this type. The pressure gradient force, however, consists of the sum of two terms with comparable magnitude and opposite sign over steep topography. Discretization errors, therefore, can be serious in the sum even when they are small in individual terms.

The most commonly used vertical grid for primitive equation models is the Lorenz grid, shown in [Figure 4\(a\)](#) for the σ coordinate. With this grid, the model atmosphere is divided into sublayers, and the ‘vertical velocity’ of the coordinate ($\dot{\sigma} \equiv D\sigma/Dt$) is defined at the interfaces between the sublayers. The budget equations for mass and three-dimensional prognostic variables, such as horizontal velocity v and potential temperature θ , are then applied to the sublayers. The Lorenz grid, therefore, is convenient for keeping track of budgets of three-dimensional variables. There is, however, a computational mode in the vertical structure of θ . This problem does not exist in the Charney–Phillips grid shown in [Figure 4\(b\)](#), which predicts θ at the interfaces where $\dot{\sigma}$ is defined.

Most GCMs use a Eulerian finite-difference system in the vertical. The number of vertical levels in current GCMs varies roughly between 30 and 60, and even higher.

Horizontal Coordinates and Grids

Most GCMs use longitude–latitude grids based on the spherical coordinates. Finite-difference methods based on such grids, however, must deal with the ‘pole problem’ as meridians converge near the poles. There are solutions to this problem but none is fully satisfactory. Spectral methods based on spherical harmonics avoid the problem, but have difficulties in advecting highly variable positive definite scalars such as moisture. When the quasiuniform spherical ‘geodesic’ grids generated from icosahedra or other Platonic solids are used, the pole problem does not exist.

Conceptually, the simulation of large-scale atmospheric flow requires the proper simulation of geostrophic adjustment

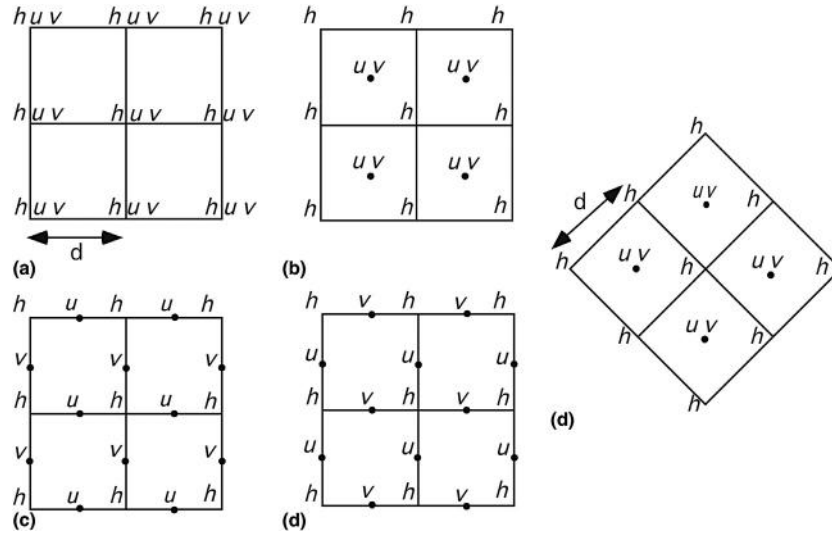


Figure 4 The Lorenz and Charney–Phillips vertical grids with the σ coordinate.

by dispersion of inertia–gravity waves and that of the slowly changing flow after geostrophic adjustment has taken place. The simulation of geostrophic adjustment can be greatly affected by the choice of horizontal grid structure through different dispersion properties of the inertia–gravity waves in the discrete system. Figure 5 shows some grid structures for the shallow water equations, where h is the free surface height, and u and v are the horizontal velocity components. The nomenclature Grid A through Grid E with h replaced by a scalar prognostic variable is often used for describing the horizontal grid structure of GCMs. The computational design for slowly changing flow is also important in making long-term integrations of the governing equations. Here, avoiding excessive nonphysical production or dissipation of quadratic functions of prognostic variables has been found to be effective.

The horizontal resolutions of current GCMs vary widely according to the intended applications. Grid sizes used for

low-resolution climate simulations are around 200 km, while for NWP the grid size can be about one order of magnitude smaller.

Parameterization of Physical Processes

GCMs include formulations of most or all of the physical processes shown in Figure 2. Those formulations critically affect the quality of the GCM simulations.

Radiation

The net radiative heating is obtained from parameterizations of radiative transfer by visible and near-infrared solar radiation and cooling by longwave terrestrial radiation. The calculations are usually made for clear and cloudy conditions.

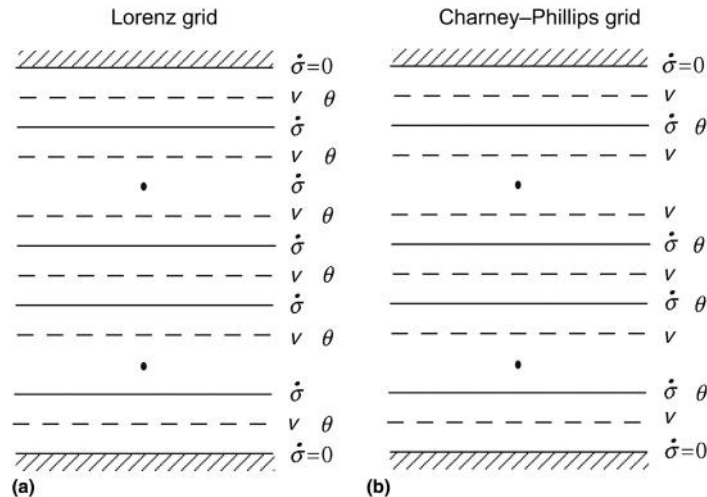


Figure 5 Distributions of the dependent variables on horizontal grids A through E.

For the solar radiation, computing absorption requires consideration of all absorption lines of atmospheric constituents at each grid point. Since this is too demanding, methods used are based on functional relations for absorption in fairly broad spectral intervals. Scattering of solar radiation must be considered. One approach to include scattering is to assume that the radiation can be divided into upward and downward streams of radiant energy ('two-stream' methods). For the longwave radiation, the principal gases considered are H_2O and CO_2 . Other important gases are O_3 , CH_4 , N_2O , and CFCs. The calculation is also performed by band models.

Radiative calculations require specification of part of the grid box covered by clouds (fractional cloud cover). There are different approaches to the formulation of fractional cloud cover, which remains one of the major uncertainties in current GCMs.

Formulation of Planetary Boundary Layer (PBL) Processes

The standard practice in numerical models of the atmosphere is to calculate surface fluxes using the bulk aerodynamical method. The Monin–Obukov surface layer similarity theory shows that the transfer coefficients depend on the surface roughness length and the bulk surface Richardson number.

In view of the vertical resolutions in GCMs, however, values in the surface layer are unknown and must be diagnosed from the values predicted above. One of the main differences between PBL parameterizations for GCMs is in the representation of PBL processes above the surface layer. Recently, the importance of representing (vertically) nonlocal effects of PBL turbulence has been widely recognized. Processes crucial for a cloud-topped PBL, however, are included only in a few GCMs.

Dry Convective Adjustment

If the temperature lapse rate exceeds the dry adiabatic lapse ($\partial\theta/\partial z$ is negative), the atmosphere is dry convectively unstable and small-scale (dry) convection will develop. In most GCMs, this process is parameterized as an instantaneous adjustment of the unstable vertical profile of θ to a neutral profile.

Condensation Processes

Most GCMs include grid-scale condensation as adjustment to a saturated state when the water–vapor mixing ratio at a grid point tends to become supersaturated.

If the temperature lapse rate exceeds the moist adiabatic lapse rate and if the relative humidity is sufficiently high, small-scale cumulus convection is expected to develop. Formulating the collective effects of such cumulus convection, which is not resolved by usual GCMs grids, in terms of resolvable-scale prognostic variables is called *cumulus parameterization*.

There are schemes in which cumulus effects are directly formulated using large-scale low-level convergence as the forcing mechanism. An increasing number of cumulus parameterizations currently used, however, formulates the effect of moist convective processes on large-scale fields as cumulus adjustment and correspondingly identify, if necessary, how large-scale processes control cumulus convection against the adjustment (large-scale forcing). This is typically done as a combination of a one-dimensional cloud ensemble model and an assumption on the quasiequilibrium states. These states can be considered as neutral or marginally unstable states from the viewpoint of moist convective instability.

There are still a number of uncertainties in formulating cloud processes for a cumulus parameterization, as illustrated by the question marks in Figure 6. During the last decade, there has been a new development, called *superparameterization* or *multiscale modeling framework (MMF)*. In this approach, cumulus parameterization is replaced with simulation of cumulus convective processes by two-dimensional cloud-resolving models (CRMs) embedded within each grid cell of the GCM. In this way, most of the cloud-scale interactions shown in Figure 2 are considered.

Parameterization of Cloud Fields

The coupling between radiative and dynamical–hydrological processes through time-dependent cloudiness has been either completely neglected or modeled very crudely even in comprehensive GCMs. This represents the major uncertainty on the role of cloud feedbacks in climate change.

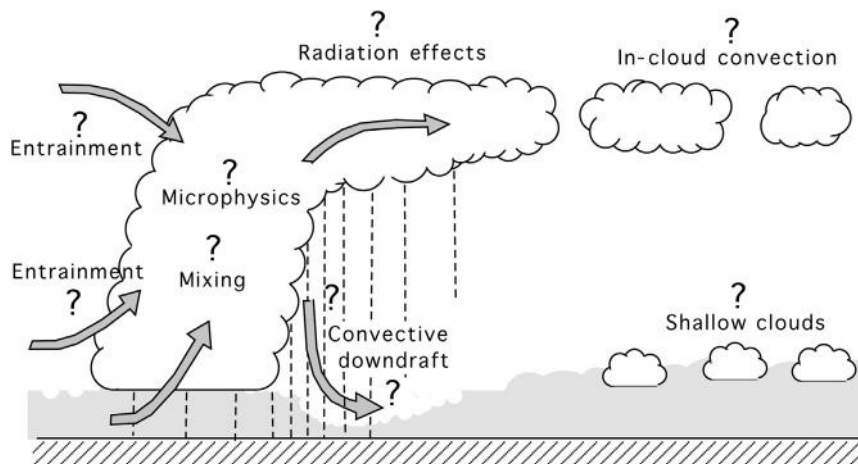


Figure 6 Outstanding questions in modeling cloud processes for cumulus parameterization.

The standard method for determining time-dependent fractional cloudiness is based on an empirical diagnostic relationship between its values and relative humidity. Even when this approach is applicable, however, prediction of mean relative humidity cannot be done without considering subgrid-scale cloud processes. A more physical determination of cloud effects on radiation involves the following problems:

1. Determination of the total amount of cloud water–ice in a grid box;
2. Determination of cloud geometry including the variation of cloud water–ice within a grid box; and
3. Determination of the radiation properties of clouds.

An increasing number of GCMs uses a ‘partially prognostic’ approach for 1, in which mean cloud water–ice mixing ratio is predicted with an advection inclusion with parameterized microphysics. A ‘fully prognostic’ scheme, which predicts cloudiness through explicitly including cloud generation processes involving 1–3 has begun to be used in some GCMs.

Subgrid-Scale Orographic Gravity Waves

Internal gravity waves forced by subgrid-scale orography can produce a drag at the surface. Near the surface, however, the deceleration effect due to this drag is compensated by convergence of the downward momentum transport by the waves. Thus the actual deceleration can happen remotely at higher levels where wave breaking takes place. Most current GCMs include a parameterization of this effect due to subgrid-scale orographic gravity waves.

Land Surface Processes

In the simplest (and earliest) approaches to represent land surface processes in GCMs moisture conditions are prescribed at the lower boundary. This may produce reasonable mean evaporation fields, but it precludes land–atmosphere interactions. The simplest interactive approach is the so-called bucket model, in which water level in a reservoir of soil moisture increases due to precipitation and decreases due to evaporation (and runoff).

More recent approaches give vegetation a more direct role in determining the surface energy and water balance, particularly by allowing stomatal conductance (and thus evaporation efficiency) to decrease in response to increased environmental stress. Recent trends include the explicit representation of spatial heterogeneity in surface characteristics within a surface element.

Microphysical Processes

Many GCMs have recently incorporated parameterized microphysical processes. These include condensational growth of cloud droplets, depositional growth of ice crystal, homogeneous, heterogeneous, and contact freezing of cloud droplets, and autoconversion of cloud droplets. They also include aggregation of ice crystals, depletion of cloud ice and cloud droplets by snow and of cloud droplets by rain, evaporation of cloud water and rain, sublimation of cloud ice and snow, and melting of cloud ice and snow.

Chemical and Aerosol Processes

Early attempts to combine GCMs and chemistry models followed an ‘off-line’ approach, in which fields generated by the former (e.g., winds) drive the latter (e.g., chemical species). More recently, GCMs and chemistry models have often been coupled ‘online,’ thus allowing for the feedbacks among dynamical, radiative, and chemical processes. In some cases, the chemistry models also calculate the advective and convective processes that transport chemical species; in others these processes are calculated by the GCM itself.

Water Isotopes

Water isotope diagnosis has been incorporated into some GCMs. Isotopes are influenced by many factors such as precipitation amount, regional geography, and distance from the sources of precipitating water. Additional difficulties for interpretation of results in terms of paleoclimate are that isotope–temperature relationships seem to vary with climate state.

GCM Applications

GCMs have two primary applications: (1) weather and climate prediction, and (2) investigations aimed to increase the understanding of the climate system and its variability. For deterministic NWP the model is initialized by combining model predicted and observed data. The current standard technique for prediction is based on performing multiple runs from slightly different initial conditions (‘ensemble integrations,’) in which each ensemble may have a few dozen members. At present, the limit of practical predictability achieved by deterministic forecasts is 7–8 days in winter, as determined by midlatitude anomaly correlation 60% score. GCMs coupled to an oceanic GCM are used for climate predictions of the first kind. It is too early to determine the limit of practical predictability in this case, but successes from several months in advance have been documented.

For the second application, the typical methodology for research starts by performing a GCM simulation long enough to achieve quasiequilibrium, which is then defined as the model’s climate. The next step repeats the simulation by altering the model component representative of the process under investigation. One of the earliest applications of this technique addressed the role played by the Himalayas on the Indian summer monsoon by comparing model climates with and without those elevations. The technique has been applied extensively to examine the global impacts on the atmosphere of SST anomalies, such as those associated with El Niño and La Niña events in the Pacific Ocean as well as similar features in the other oceans. Another application is the impact on climate of variations in surface conditions as consequence of climate change (see [Figure 7](#)). The underlying assumption in all these experiments is that one can separate the anomalies in climate from those that are ultimately associated with variations in boundary conditions.

Another important GCM application has been climate change expected as a result of the changing atmospheric composition. GCMs are particularly suitable for this problem, which involves many interactions and feedbacks. The impact of

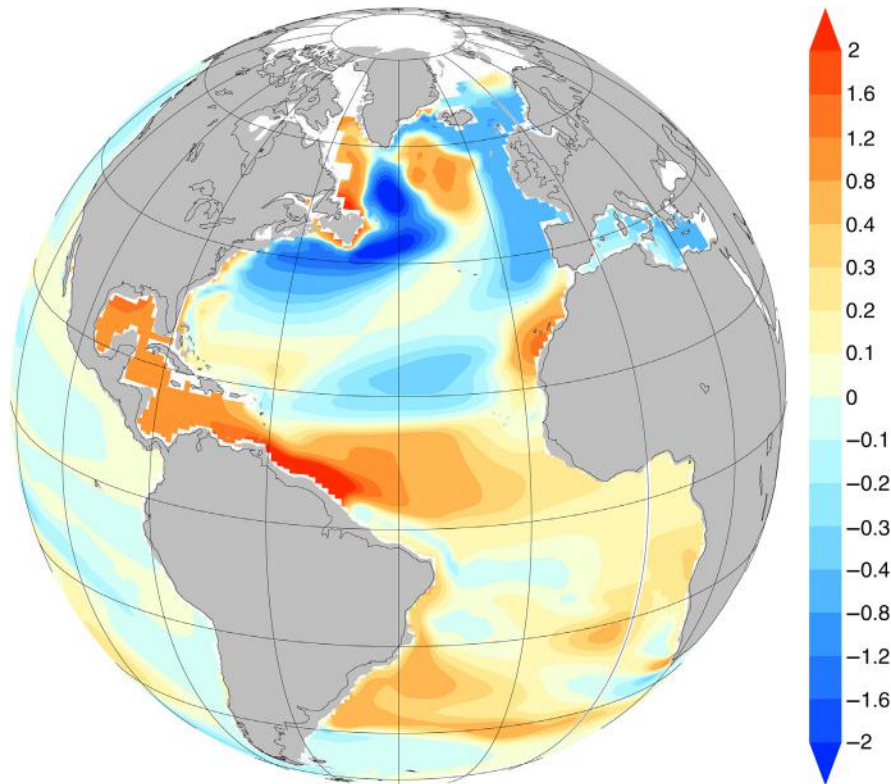


Figure 7 Ocean temperature anomalies (K) at 160-m deep simulated by the UCLA coupled atmosphere–ocean GCM as the response to melting of the Greenland ice sheet. Values correspond to the average between years 26 and 30 after the melting started. Courtesy of N. Argawal, MPI.

increased greenhouse gases has been assessed by comparing model climates with different concentrations of those gases. For these studies, atmospheric GCMs (AGCMs) have been coupled to ocean models of diverse complexity. This has allowed for exploring how the oceans can delay the effects of a greenhouse warming. More recently, the role of increasing sulfate aerosols has been questioned and GCMs have been coupled to sulfur chemistry models. This has allowed for an assessment of the indirect aerosol forcing problem, namely the impact on climate of changes in reflectivity, formation, and residence time of clouds in the presence of aerosols.

GCM simulations with prescribed concentrations of ozone in the lower stratosphere from observational data suggest that depletion of lower stratospheric ozone is the major radiative factor in accounting for the 1979–90 cooling trend in the global, annual-mean stratosphere, with a substantially lesser contribution by the greenhouse gases. Other studies have addressed the interactive radiative–chemistry–dynamical effects of ozone losses in the Antarctic polar region during the southern spring.

Computational Aspects

The computer code of all contemporary GCMs is highly optimized for vector computer architectures. Most models have also been ported to massively parallel computer architectures. AGCM codes are very heterogeneous and their optimization is not simple. The major current parameterizations operate in vertical columns of the atmosphere. If a two-dimensional grid partition

in the horizontal (longitude–latitude) is used, very little communication is required between processors for that part of the code. However, the amount of calculations (e.g., those for moist convection) can strongly vary in space and time resulting in a dynamic load imbalance. The ‘pole problem’ requires application of Fourier filters, resulting in a static load imbalance. Performances in the order of 400 Gflops have been reported.

GCM Verification

Verification of GCM performance is a challenging problem. In NWP, a standard value of performance is given by the correlation between predicted and observed fields to which the long-term model and observed climatology have been subtracted, respectively. It is currently accepted that values larger than 0.6 of such an ‘anomaly correlation’ computed by using the geopotential height field at 500 hPa correspond to ‘useful’ results. For climate simulations, early validations compared monthly means of simulated and observational fields, with emphasis on quasistationary planetary waves, eddy statistics, and monsoon circulations. Early in the 1990s, a number of AGCMs participated in a coordinated effort on verification of their performance. This was based on the comparison between outputs from simulations with the same prescribed boundary conditions corresponding to the period 1979–88. This coordinated approach has been extended to coupled atmosphere–ocean–land models in simulations that can extend up to centuries. Several efforts are also underway on the development of quantitative metrics for the success of climate simulations.

Single-column models (SCMs) have been used to test selected physical parameterizations. An SCM consists of an isolated, time-dependent GCM column forced and constrained with observational estimates of advective fluxes, and whose outputs are compared with observational data generally provided by field programs.

Current Directions

From the point of view of climate simulations, prescribing SST as in most applications during Phase II assumed the most important part of the answer and, therefore, it might have hidden crucial deficiencies of GCMs. The coupling of an AGCM with an oceanic GCM to predict SST at the beginning of Phase III started a trend toward expansion of internal processes. While there is no question about the merits of such expansion, it also demands a higher level of modeling effort. The elimination of prescribed external conditions, such as SST, uncovered model deficiencies that are serious even for simulating the present climatology. For example, most contemporary coupled atmospheric–oceanic GCMs tend to produce a climate that is significantly more symmetric about the equator than in both the atmospheric component with prescribed SSTs and the observation. Inclusion of land surface models in AGCMs revealed similar difficulties. Inclusion of other modules such as sea ice has motivated the nomenclature ‘Earth System Models.’

Another outstanding trend in Phase III is the use of higher resolutions as computer technology advances, resulting in model resolution increases. This makes the separation more ambiguous between resolvable processes, which can be highly transient, and parameterized unresolvable processes, which can only be near a statistical equilibrium. Various processes shown in [Figure 2](#) interact too strongly to be separately formulated, especially when clouds are involved. A key goal of Phase III will be a unification of those formulations and the reduction of

the dependency of model physics on grid size, which is an artificial length introduced for computational purposes. To achieve the goal, a trend to use the nonhydrostatic dynamics core and explicit simulation of cloud-scale motions even for simulating the global climate has already begun. This includes the global cloud-resolving models (GCRMs) and global models with ‘superparameterization’ or ‘MMF,’ in which the conventional cumulus parameterization is replaced by explicit simulations of convective processes using CRMs with a limited spatial dimension. The MMF has already been used for a variety of climate studies with encouraging results. Attempts to broaden the applicability of MMF with the goal to unify MMF and GCRM are also being made.

See also: **Clouds and Fog:** Cloud Modeling. **Dynamical Meteorology:** Inertial Instability; Rossby Waves; Waves. **General Circulation of the Atmosphere:** Overview. **Gravity Waves:** Overview. **Numerical Models:** Methods; Parameterization of Physical Processes: Clouds; Regional Prediction Models. **Tropical Meteorology and Climate:** El Niño and the Southern Oscillation: Observation.

Further Reading

- Donner, L., Schubert, W., Somerville, R. (Eds.), 2010. *The Development of Atmospheric General Circulation Models, Complexity, Synthesis and Computation*. Cambridge University Press, Cambridge, p. 272.
- Jacobson, M.Z., 1999. *Fundamentals of Atmospheric Modeling*. Cambridge University Press, Cambridge, p. 656.
- Randall, D.A. (Ed.), 2000. *General circulation model development: past, present and future. Proceedings of a Symposium in Honor of Professor Akio Arakawa*. Academic Press, New York.
- Schlesinger, M.E., 1988. *Physically-based Modelling and Simulation of Climate and Climatic Change*. Kluwer Academic Publishers, Dordrecht, p. 1084.
- Trenberth, K.E. (Ed.), 1992. *Climate System Modeling*. Cambridge University Press, Cambridge, p. 788.

Methods

J Thuburn, University of Exeter, Exeter, UK

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

Numerical modeling of the atmosphere is challenging because of the enormous range of space and timescales involved, and because many qualitative features, such as balance and conservation, must be accurately captured. We discuss here a number of topics that have proved important historically in the development of numerical models of the atmosphere, including spatial discretization and its relation to representation of model fields, choice of vertical and horizontal grids, time integration schemes, advection schemes, and solution methods for elliptic problems.

Special Challenges in Numerical Modeling of the Atmosphere

The atmosphere has a number of characteristics that make numerical solution of its governing equations especially challenging. Most obvious is the geometry. The atmosphere is highly anisotropic, with vertical length scales being typically much smaller than horizontal. Gravity acts in the vertical direction, and the atmosphere is strongly stratified and close to hydrostatic balance. The ratio of horizontal to vertical model resolution is commonly chosen to capture this anisotropy, and model grids and methods must be able to represent hydrostatic balance (see Section on [Grids](#) below). For global modeling, the grid and methods chosen must be able to represent the (approximately) spherical geometry of the Earth (again see Section on [Grids](#)).

The atmosphere is strongly multiscale in both space and time. In terms of spatial energy spectra, the largest scales are energetically dominant, but spectra are rather shallow, implying that, whatever the resolution of an atmospheric model, there is significant dynamical variability near the resolution limit, which requires careful handling by numerical methods, and on unresolved scales, which must be represented by 'subgrid models.' The atmosphere also supports dynamics with a huge range of timescales. Care is needed in modeling fast processes to ensure that the numerical solution is stable (see Section on [Time Integration Schemes, Stability, Courant–Friedrichs–Lewy Criterion](#) below). On the other hand, fast waves (acoustic waves, and to a lesser extent, inertio-gravity waves) tend to be energetically weak, implying that the dynamics is often slowly evolving and close to hydrostatic and geostrophic balance. The chosen numerical method must be able to represent this balance accurately. Also, certain properties of the atmosphere evolve slowly, either in a Lagrangian sense (e.g., the moisture content of an air parcel in the absence of condensation and evaporation) or in a global integral sense (e.g., the total mass or energy of the atmosphere). Numerical methods should be able to capture these conservation properties accurately (see Section on [Conservation and Scale-Selective Dissipation](#) below).

Representation of Data

In a numerical model, the fields describing the three-dimensional state of the atmosphere (density, velocity,

temperature, etc.) must be stored in a computer as a finite set of values. There are several possible ways to do this.

- *Gridpoint representation.* In gridpoint methods, a field is represented by its values sampled at a finite set of points called the grid or mesh points ([Figure 1](#), top left).
- *Finite volume representation.* Each value stored in the computer represents the value of a field averaged over a finite volume or grid cell ([Figure 1](#), top right).
- *Finite element representation.* Each field $\phi(\mathbf{x})$ is represented as a series in terms of certain basis functions $\psi_i(\mathbf{x})$, and the coefficients $\hat{\phi}_i$ are stored in the computer ([Figure 1](#), bottom left):

$$\phi(\mathbf{x}) = \sum_i \hat{\phi}_i \psi_i(\mathbf{x}). \quad [1]$$

Usually the ψ_i are simple local functions such as piecewise polynomials that are nonzero only over a small number of grid cells.

- *Spectral representation.* As for finite element methods, each field is represented as a series in terms of certain basis functions, but here the basis functions are mutually orthogonal global functions. The simplest example in one dimension is a Fourier series representation, in which the basis functions are sines and cosines ([Figure 1](#), bottom right). For modeling the atmosphere on a spherical planet, a spectral method based on spherical harmonics is often used.

Some methods may be viewed in more than one way, for example as a finite volume representation for the purpose of looking at conservation properties but as a gridpoint representation for the purpose of calculating derivatives. In more than one dimension, combinations of the above methods can be used, e.g., spectral in the horizontal and gridpoint in the vertical.

The number of data values used to represent a field defines the model resolution: a greater number of data values corresponds to higher resolution and means that smaller-scale structures in the field can be represented. For a current-day global weather forecast model, a typical horizontal resolution might be a few tens of kilometers, while a typical vertical resolution might stretch from a few meters or a few tens of meters near the surface to 2 or 3 km in the stratosphere. Climate models generally have coarser horizontal resolution to make long integrations affordable.

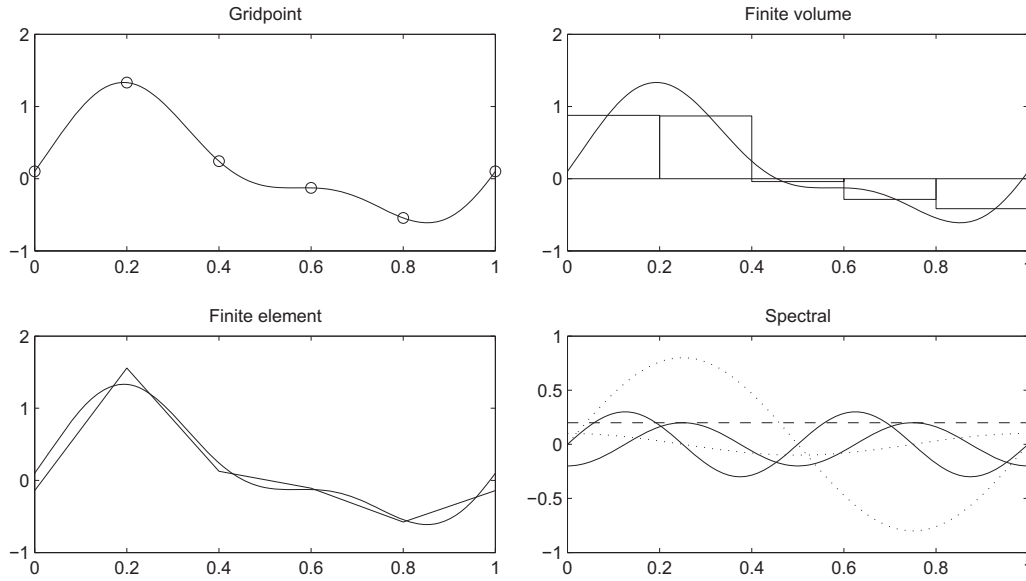


Figure 1 Illustration of different discrete methods for representing the same one-dimensional function. Top left: representation by five gridpoint values (the sixth value is the same as the first because the function is periodic). Top right: finite volume representation as five grid cell averages. Bottom left: finite element representation in terms of five piecewise linear ‘wings hat’ basis functions. Bottom right: the first five Fourier components of the spectral representation of the function.

Round-Off Error, Truncation Error, Local Truncation Error, Consistency

Numerical solution of differential equations involves approximations and hence, inevitably, errors. Round-off error arises because a finite number of digits, and hence a limited precision, are used in storing numbers in a computer. On a typical workstation using standard precision, the round-off error is about ± 3 in the eighth decimal digit. For most purposes this error is so small as to be insignificant. However, the effects of round-off error can be greatly amplified in certain calculations; for example, when adding small numbers to large ones or when taking differences of nearly equal numbers.

The approximation of a continuous field by a finite set of data values and the approximation of derivatives or integrals by algebraic operations such as differences or sums is called discretization. The errors made in these approximations are called truncation errors. They are a function of the resolution of the model, and are usually much more significant than round-off errors.

The local truncation error is the residual obtained when the true solution of a differential equation is substituted into the discrete approximation of that equation. It is a measure of how well the discretization approximates the differential equation. For example, consider the one-dimensional linear advection equation

$$\frac{\partial \phi}{\partial t} + v \frac{\partial \phi}{\partial x} = 0, \quad [2]$$

where the advecting velocity v is a constant. Let us discretize on a uniform grid with spatial interval Δx and time interval Δt so that ϕ_j^n is the numerical approximation to ϕ at $x = j\Delta x$, $t = n\Delta t$.

Approximating the space and time derivatives by simple difference formulas gives, for example,

$$\frac{\phi_j^{n+1} - \phi_j^n}{\Delta t} + v \frac{\phi_j^n - \phi_{j-1}^n}{\Delta x} = 0. \quad [3]$$

If we substitute the true solution $\phi(j\Delta x, n\Delta t)$ in place of ϕ_j^n , etc., in eqn [3] and make use of Taylor series, we find the residual, i.e., the local truncation error, as $O(\Delta x) + O(\Delta t)$. In this example the discretization is said to be first-order accurate in space and time, because both Δx and Δt appear raised to the power one in the local truncation error. A discretization is said to be consistent if the local truncation error tends to zero as $\Delta x \rightarrow 0$ and $\Delta t \rightarrow 0$.

Spatial Discretization Methods: Finite Difference, Finite Volume, Finite Element, and Spectral

A huge variety of methods are available for approximating the spatial derivatives that appear in the governing equations. The different methods are related to how the data are represented. Finite difference methods are perhaps conceptually the simplest. Rearranging Taylor series allows the construction of approximate expressions for derivatives in terms of point values of the data. Equation [3] shows a simple example in which $\partial \phi / \partial x$ at $(x = j\Delta x, t = n\Delta t)$ is approximated by $(\phi_j^n - \phi_{j-1}^n) / \Delta x$. Higher-order accurate formulas and formulas for higher derivatives can also be derived.

Finite volume methods allow one to define a local or global integral of a predicted field, and so are natural when the emphasis is on integral conservation properties. The equations

are usually written in terms of the budget of each prognostic field for each grid cell, and the method is defined by constructing a suitable approximation for the flux of each field across each cell face.

For finite element and spectral methods, the objective is to determine how the coefficients $\hat{\phi}_i$ evolve in time. One way to do this is to require that the residual, that is, the error when the approximate solution [1] is substituted into the original differential equation, should be orthogonal to each of the basis functions. This approach typically leads to a set of simultaneous equations involving $d\hat{\phi}_i/dt$ multiplied by a sparse matrix called the mass matrix. Once these equations are solved for $d\hat{\phi}_i/dt$, $\hat{\phi}_i$ may be stepped forward using a suitable time integration scheme. Discontinuous Galerkin and spectral element methods are closely related to the finite element method and have also been used in atmospheric models.

For spectral methods, the basis functions are mutually orthogonal; therefore the mass matrix is diagonal and trivial to invert. On the other hand, nonlinear terms in the governing equations would be expensive to evaluate directly in terms of the spectral coefficients $\hat{\phi}_i$. Instead, it is simpler and more efficient to transform the model fields from a spectral representation in terms of $\hat{\phi}_i$ to a gridpoint representation in terms of ϕ_i to evaluate the nonlinear terms, and then to transform the result back into the spectral representation. The transforms can be carried out efficiently with the aid of a fast Fourier transform algorithm. This method is called the spectral transform method.

Grids

Because of the strong anisotropy of the atmosphere, atmospheric models are almost invariably constructed with a certain number of levels or layers in the vertical, with essentially the same horizontal grid structure at each level. Thus it makes sense to discuss the vertical grid structure and horizontal grid structure separately.

An obvious choice for the locations of the vertical levels is to place them at specified heights. However, in the presence of mountains, the bottom boundary condition is greatly simplified if the model levels are distorted to follow the terrain (Figure 2). It is common to define the levels to be more closely spaced near the bottom boundary, to better capture details of the atmospheric boundary layer.

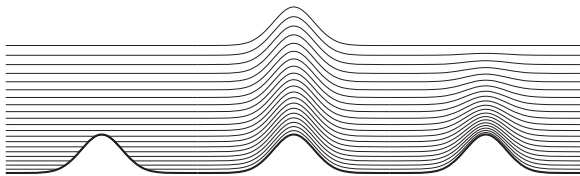


Figure 2 Schematic showing three variants of a height-based coordinate and levels and how they vary in the vicinity of a mountain. Left: flat, terrain-intersecting levels; middle: basic terrain-following levels; right: hybrid levels that are terrain-following near the surface but become flat at high altitude.

The governing equations may be expressed in a variety of different vertical coordinates, and model levels chosen at specified values of the vertical coordinate. As well as height, example vertical coordinates include pressure, mass, potential temperature, and a Lagrangian vertical coordinate (i.e., one that moves with the flow), and their corresponding terrain-following variants. Each has its own advantages and disadvantages compared with a height-based coordinate and levels.

More accurate wave propagation and a more accurate representation of hydrostatic balance can be achieved by using a staggered vertical grid, in which different variables are interleaved rather than colocated. Two types of staggered vertical grid are widely used: those with the temperature variable at the same levels as the horizontal velocity components, commonly referred to as the Lorenz grid; and those with the temperature variable staggered with respect to the horizontal wind components, commonly referred to as the Charney–Phillips grid (Figure 3). The Charney–Phillips grid provides slightly more accurate wave propagation, while energy conservation is easier to achieve on the Lorenz grid.

For a limited area model, the simplest choice for the horizontal grid is a rectangular Cartesian grid. For a global model, a longitude–latitude grid is the simplest choice. However, the clustering of resolution near the poles creates additional difficulties related to stability, convergence of elliptic solvers (see below), and scalability on massively parallel computers. This has motivated the development of models with a variety of more uniform horizontal grids. Some examples are shown in Figure 4.

Staggered horizontal grids are possible, and can give increased accuracy in representing wave propagation and balance. Some examples of staggered rectangular Cartesian grids are shown in Figure 5. The commonest staggered grids are called the Arakawa A-, B-, C-, D-, and E-grids, after the work of Arakawa and colleagues who first studied their wave dispersion properties. Analogous staggering of variables is possible on more general polygonal grids such as those shown in Figure 4.

Time Integration Schemes, Stability, Courant–Friedrichs–Lewy Criterion

Given a solution at time $t = n\Delta t$ (and perhaps at earlier time steps), we need a scheme to advance the solution to time

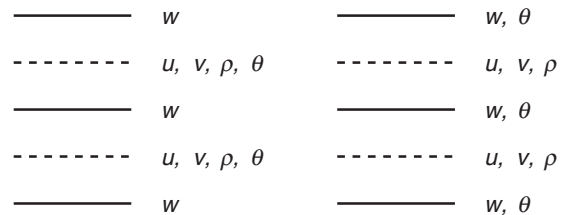


Figure 3 Schematic showing the vertical placement of the prognostic fields for examples of the Lorenz vertical grid (left) and the Charney–Phillips vertical grid (right). u , v , and w are the velocity components, ρ is density, and θ is potential temperature.

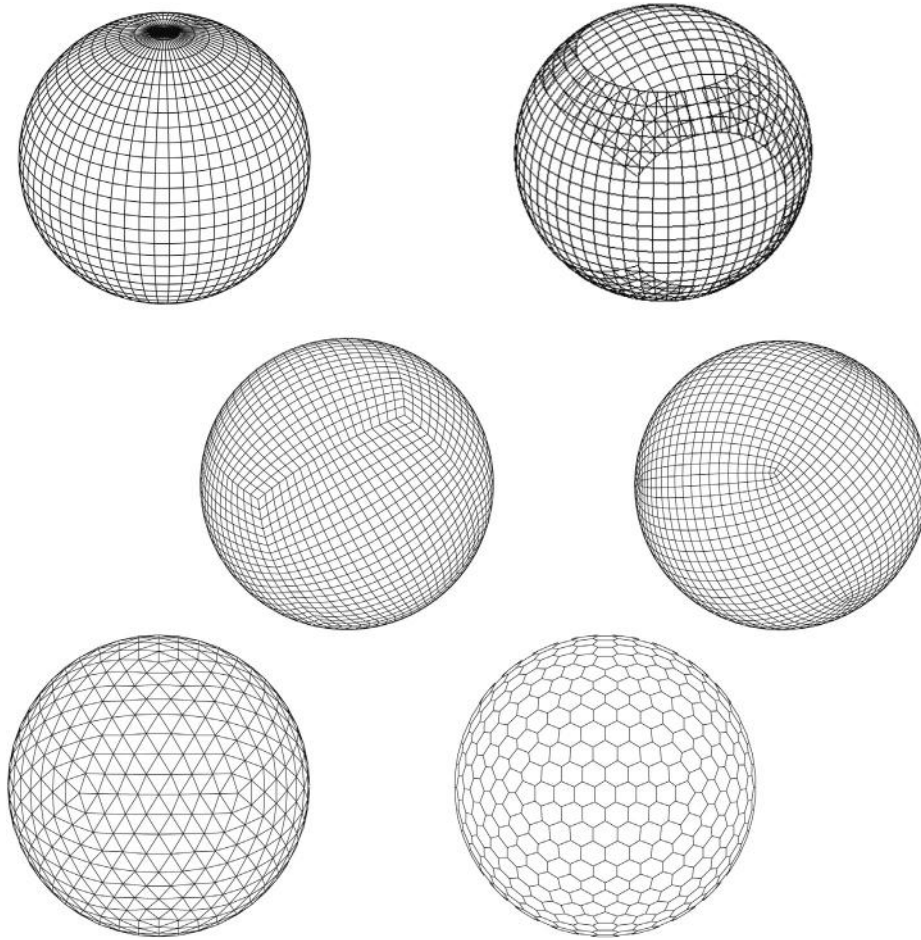


Figure 4 Some examples of spherical grids used in atmospheric models. Top left to bottom right: Longitude–latitude grid, Yin–Yang grid, equal angle cubed sphere, conformal cubed sphere, triangular icosahedral grid, hexagonal–pentagonal icosahedral grid.

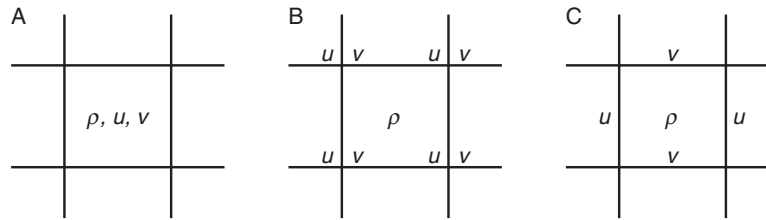


Figure 5 Schematic showing the horizontal placement of prognostic fields for three of the rectangular grids studied by Arakawa and colleagues. Left: A-grid; middle: B-grid; right: C-grid. u , v are the horizontal velocity components and ρ is the density.

$t = (n + 1)\Delta t$. A great many such schemes are possible. Consider the following differential equation to illustrate ideas:

$$\frac{d\phi}{dt} = R(\phi). \quad [4]$$

One of the simplest possible schemes approximates the time derivative by a forward difference:

$$\frac{\phi^{n+1} - \phi^n}{\Delta t} = R(\phi^n). \quad [5]$$

Equation [3] uses this time integration scheme. This is an example of an explicit scheme, so-called because the formula

may be rearranged to give an explicit expression for ϕ^{n+1} in terms of known quantities.

The scheme (5) is not very accurate. Greater accuracy can sometimes be obtained by using information from other time levels, giving a so-called multistep method. The leapfrog scheme

$$\frac{\phi^{n+1} - \phi^{n-1}}{2\Delta t} = R(\phi^n). \quad [6]$$

is one example. Adams–Bashforth schemes also fall into this category.

Another way to increase accuracy is via a multistage scheme. Here, intermediate solutions are computed at one or more

stages between step n and step $n + 1$. The Matsuno or forward-backward scheme is one example:

$$\frac{\phi^* - \phi^n}{\Delta t} = R(\phi^n); \quad \frac{\phi^{n+1} - \phi^*}{\Delta t} = R(\phi^*). \quad [7]$$

Runge-Kutta schemes are also popular multistage schemes.

An important consideration for time integration schemes is stability. There are several slightly different technical definitions of stability. For practical purposes, a scheme is unstable if it produces numerical solutions that grow in amplitude when the true solution does not grow. Explicit time integration schemes are typically stable only for sufficiently small Δt , that is, they are conditionally stable. Stable solutions for larger time steps can be obtained using implicit time integration schemes. A simple example is the trapezoidal or Crank-Nicolson scheme

$$\frac{\phi^{n+1} - \phi^n}{\Delta t} = \frac{1}{2} (R(\phi^n) + R(\phi^{n+1})). \quad [8]$$

Such schemes are called implicit because the unknowns ϕ^{n+1} appear in several places, and a system of (possibly nonlinear) simultaneous equations must be solved to step forward. See Section on [Elliptic Problems](#) below.

Stability often depends on the coupling of the spatial discretization and the time integration scheme, in particular on the satisfaction of the Courant-Friedrichs-Lewy (CFL) criterion. The CFL criterion states that a necessary condition for stability is that the numerical domain of dependence of the solution at any point in space and time should contain the physical domain of dependence. For simple schemes, such as (3), the CFL criterion often reduces to the requirement that the Courant number $c\Delta t/\Delta x$, where c is an appropriate wave or advective signal velocity, should be less than one.

Various issues must be taken into consideration when choosing a time integration scheme (in combination with a spatial discretization). The scheme should, of course, be stable, but at the same time it should not introduce excessive dissipation or damping of the solution. Phase and group propagation of all waves of meteorological interest should be accurately captured. Even waves that are not of direct interest (often the case for acoustic waves and sometimes inertio-gravity waves) must be captured well enough to represent adjustment toward balance. Multistep schemes may support computational modes, that is, numerical solutions that do not correspond to any solution of the original differential equation; some form of time filtering may be necessary to control them. Finally, the overall cost should be acceptable.

In practice, time integration schemes are often complex combinations of the kinds of scheme mentioned above. For example, early atmospheric models often used a leapfrog scheme (6) for the conservative wavelike terms combined with a forward scheme (like (5), but from step $n - 1$ to step $n + 1$) for dissipative terms, using each scheme on the terms for which it is conditionally stable. To avoid solving a global nonlinear system of equations, semi-implicit schemes treat a linearized subset of terms in the governing equations implicitly (like [eqn \[8\]](#)) and the remaining terms explicitly. Horizontally explicit, vertically implicit schemes treat only vertical derivative terms implicitly, so that only a relatively small system of equations needs to be solved for each vertical column. Finally, split explicit schemes are also popular. These avoid the need to solve

a system of simultaneous equations that arises with an implicit method while reducing computational cost by taking small time steps only for a subset of (relatively cheap) terms associated with fast waves and using longer time steps for the other (more expensive) terms.

Advection

The transport of some fluid property by fluid parcels is called advection. The importance of advection for fluid problems, and, in particular, the fact that many important quantities, such as potential temperature, potential vorticity, moisture, and long-lived chemical tracers often evolve slowly following fluid parcels – that is, they are approximately conserved in a Lagrangian sense – have led to a great deal of effort to develop accurate numerical methods for advection.

One approach to modeling advection is to combine an accurate approximation for the spatial derivatives (e.g., the $\partial\phi/\partial x$ term in [eqn \[2\]](#)) with an accurate approximation for the time derivative term (e.g., the $\partial\phi/\partial t$ term in [eqn \[2\]](#)). In this approach, better stability and accuracy is often obtained by using an upwind biased approximation for the spatial derivatives.

An alternative approach is to re-express the advection equation as

$$\frac{D\phi}{Dt} = 0, \quad [9]$$

where D/Dt indicates a derivative following a fluid parcel rather than at a fixed point in space. Semi-Lagrangian schemes follow this approach. For each gridpoint, x_j say, they first compute a departure point x_d , which is the location at step n of the fluid parcel arriving at x_j at step $n + 1$. The value of ϕ at the departure point ϕ_d^n is computed by interpolating the field ϕ^n . Then the arrival point value is set equal to the departure point value: $\phi_j^{n+1} = \phi_d^n$. Semi-Lagrangian schemes are widely used in atmospheric models because of their high accuracy, and because their stability is not limited by the advective Courant number, allowing longer time steps and making them very efficient.

It is often desirable to ensure that the numerical advection scheme does not lead to unphysical negative values of the advected field, or to spurious amplification of extrema (which could lead to spurious supersaturation in the moisture field, for example). Such a desirable property goes by a variety of names, including monotone, monotonicity preserving, total variation diminishing, shape preserving, or locally bounded. It can be achieved by suitable modifications to interpolation schemes or to fluxes, often referred to as limiters.

Elliptic Problems

Elliptic problems arise in atmospheric models when some form of balance is assumed, e.g., when determining the stream function from the quasigeostrophic potential vorticity, or when determining the pressure field for the Boussinesq or anelastic equations. They also arise when vertical vorticity and horizontal divergence are the predicted variables, and it is necessary to diagnose the velocity field from them via a stream function and velocity potential. Finally, the system of equations that

arises for the unknowns when using an implicit or semi-implicit time integration scheme typically takes the form of a spatial discretization of a Helmholtz problem

$$\phi^{n+1} - \nabla \cdot (\nu \nabla \phi^{n+1}) = \text{RHS}, \quad [10]$$

where RHS indicates known terms.

Spatial discretization of an elliptic problem leads to an equation of the form

$$\mathbf{M}\phi^{n+1} = \mathbf{B}, \quad [11]$$

where $\mathbf{M}$ is a sparse matrix and $\mathbf{B}$ is a known vector. Direct solution of eqn [11] is usually prohibitively expensive for three-dimensional problems. However, other methods are available.

If a spectral representation of ϕ was used and the basis functions could be chosen to be eigenfunctions of the Helmholtz operator, then $\mathbf{M}$ would be diagonal and the solution would be trivial. In practice the basis functions are usually chosen to be eigenfunctions of the horizontal part of the Helmholtz operator (for ν constant), e.g., spherical harmonics for a global model. Then eqn [11] separates into a number of much smaller problems whose matrices correspond to the vertical part of the Helmholtz operator; these can be solved by direct matrix inversion.

If a spectral method is not used, then eqn [11] may be solved iteratively. One type of iterative method improves ('relaxes') the solution at each gridpoint in turn, making a number of sweeps over the whole grid. Depending on the details of the relaxation, this gives rise to Jacobi, Gauss-Seidel, and successive over-relaxation methods. Convergence of these basic relaxation methods may be slow if the length scale $\nu^{1/2}$ in eqn [10] is large compared with the grid length. In this case, convergence can be greatly accelerated by using a multigrid method, in which some iterations are carried out on coarser grids than that on which the final solution is required.

Another type of iterative method finds a sequence of successively better solutions over a growing hierarchy of vector spaces, called Krylov subspaces, whose bases are generated by iteratively applying $\mathbf{M}$ to some starting vector. Depending on the details, this gives rise to methods such as conjugate gradient, biconjugate gradient, minimum residual, generalized minimum residual, and generalized conjugate residual with k steps. The convergence of Krylov subspace methods can often be greatly improved by use of an appropriate preconditioner. That is, the problem [11] is rewritten as

$$\mathbf{P}^{-1}\mathbf{M}\phi^{n+1} = \mathbf{P}^{-1}\mathbf{B} \quad [12]$$

or as

$$\mathbf{M}\mathbf{Q}^{-1}(\mathbf{Q}\phi^{n+1}) = \mathbf{B} \quad [13]$$

for some matrix $\mathbf{P}$ or $\mathbf{Q}$ whose inverse is easily calculated.

Conservation and Scale-Selective Dissipation

Good Lagrangian conservation properties can often be achieved in a numerical model by the choice of an accurate

advection scheme, as discussed above. Depending on the application, integral conservation properties may also be important. Conservation of mass of air, mass of moisture and chemical tracers, energy, angular momentum, entropy, and potential enstrophy (among others) have all been addressed by model developers. Conservation of mass is straightforwardly achieved by the use of a finite volume method for the density or tracer density equation. Other conservation properties typically require some special measures in the numerical methods, and may only be achieved approximately.

A variety of processes can lead to the growth of variability or 'noise' in a numerical model on marginally resolved scales. These include physically realistic processes such as the transfer of energy and potential enstrophy to different length scales through the nonlinearity of the governing equations, as well as numerical errors such as dispersion errors and aliasing (the misrepresentation of variability at one, unresolved, scale as variability at another, resolved, scale). Other contributors include the action of physical parameterization schemes, and small-scale forcing, for example, due to orography. To control such a build up of noise, all atmospheric models include some form of scale-selective dissipation that preferentially damps small scales. This may be in the form of extra dissipation terms explicitly added to the governing equations, or in the form of inherently dissipative numerical approximations. Care is needed to ensure that such dissipation is sufficient to control the buildup of grid-scale noise without leading to excessive damping or adversely affecting conservation properties.

See also: **Dynamical Meteorology:** Acoustic Waves; Balanced Flow; Lagrangian Dynamics; Overview; Potential Vorticity; Primitive Equations. **Gravity Waves:** Buoyancy and Buoyancy Waves: Theory. **Numerical Models:** General Circulation Models; Mesoscale Atmospheric Modeling; Regional Prediction Models. **Turbulence and Mixing:** Turbulence, Two Dimensional.

Further Reading

- Durrant, D.R., 1999. Numerical Methods for Wave Equations in Geophysical Fluid Dynamics. Springer-Verlag, New York.
- Haltiner, J.G., Williams, R.T., 1980. Numerical Prediction and Dynamic Meteorology. Wiley, Chichester.
- Lauritzen, P.H., Jablonowski, C., Taylor, M.A., Nair, R.D., 2011. Numerical Techniques for Global Atmospheric Models. Springer-Verlag, Berlin, Heidelberg.
- Pielke, R.A., 1984. Mesoscale Meteorological Modeling. Academic Press, London.
- Press, W.H., Teukolsky, S.A., Vetterling, W.T., Flannery, B.P., 1992. Numerical Recipes in FORTRAN 77: The Art of Scientific Computing, second ed. Cambridge University Press, Cambridge.
- Staniforth, A., Thuburn, J., 2012. Horizontal grids for global weather and climate prediction models: a review. Quarterly Journal of the Royal Meteorological Society 138, 1–26.
- Thuburn, J., 2008. Some conservation issues for the dynamical cores of NWP and climate models. Journal of Computational Physics 227, 3715–3730.

Model Physics Parameterization

DJ Stensrud and MC Coniglio, National Oceanic and Atmospheric Administration, Norman, OK, USA

KH Knopfmeier and AJ Clark, University of Oklahoma and National Oceanic and Atmospheric Administration, Norman, OK, USA

© Published by Elsevier Ltd.

Synopsis

The parameterizations of physical processes that cannot be explicitly represented by numerical weather prediction models are critical model components and their improvement is one reason why numerical models are becoming more skillful. Parameterization schemes relate the effects of the chosen physical process to the model variables using an algorithmic or statistical approach. Typical parameterization schemes used in operational forecast models represent the effects of shortwave and longwave radiation, cloud cover, soil–vegetation–water–atmosphere transfer, urban areas, planetary boundary layer, convection, microphysics, and orographic drag. Brief overviews of some of these parameterizations are provided, along with a comparison of scheme results to highlight recent advances and remaining challenges.

Introduction

Numerical weather prediction models are ubiquitous in meteorology and the atmospheric sciences. These models form the cornerstone of operational weather forecasting and are used to study a wide variety of phenomena from thunderstorms to the El Niño Southern Oscillation. Numerical models also are beginning to play a role in seasonal forecasting and are a key component in studies of global climate change. Arguably, the most important components of these models are the subgrid-scale parameterization schemes that represent physical processes the model cannot resolve explicitly. These parameterization schemes determine the amounts of shortwave and longwave radiation that reach the Earth's surface; the partitioning of energy into sensible and latent heat fluxes from the Earth's surface to the atmosphere; the evolution and depth of the planetary boundary layer (PBL); the development and evolution of clouds; and the amount of rainfall that reaches the ground. Parameterization schemes are important because they have a very large influence on the resulting model behavior. If one desires to improve model capability and/or model forecast skill, then one must have a deep understanding of model parameterization schemes.

The development of parameterization schemes follows the reductionist approach, whereby schemes are developed separately for individual physical processes and it is assumed that the sum total of all the schemes in a model is capable of representing the behavior of the atmosphere. The schemes almost always treat the effects of the subgrid-scale physical processes within the vertical column of each model grid cell (Figure 1) and only rarely deal with changes in the horizontal direction. The vertical projection of parameterizations is a natural outcome of the physical processes represented, which tend to rearrange mass and momentum in the vertical direction. Parameterization schemes relate the effects of the chosen physical process to the model variables using an algorithmic or statistical approach. The development of a scheme tends to occur independently of other schemes and often involves testing of scheme performance against observational data sets, which may be limited in geographic extent. When inserted into a numerical model, the scheme is then

assumed to produce reasonable results anywhere on the globe and likely for conditions in which it was never fully tested. However, the overall success of operational numerical weather prediction indicates that the reductionist approach has worked very well. The continued presence of forecast failures suggests that understanding the interactions between parameterization schemes may be needed to provide further forecast improvements.

The development of new parameterization schemes appears to be accelerating. Articles in American Meteorological Society journals with the word 'parameterization' in the title have more than doubled from 104 in the decade from 1980 to 1989 to 228 in the decade from 2000 to 2009. Community models, such as the Weather Research and Forecasting (WRF) model, allow users to select between multiple parameterization schemes for each different physical process contained in the model. This situation has both positive and negative aspects. The positive side is that a wise user can select parameterization schemes that are best suited for a particular problem being investigated. The negative side is that the selection of a poor parameterization scheme could lead not only to an inaccurate weather forecast but also to the wrong conclusion when conducting a phenomenological study. Thus, the need to understand parameterization schemes has increased and model users must accept greater responsibility to make good scheme choices.

The study of parameterization also opens a window for us to examine our most fundamental understandings of atmospheric processes. A physical process is distilled to its most essential components when put into a parameterization scheme, as only a limited amount of complexity is possible due to computational constraints. Each scheme starts from a set of assumptions that depend to some extent upon the configuration of the numerical model for which it is designed. A common assumption used by nearly all parameterization schemes is the horizontal grid spacing of the numerical model. As model grid spacing continues to decrease as computer power increases, some schemes are no longer needed and the assumptions behind other schemes may be violated unless the scheme is updated to function properly at the smaller grid spacing. This situation presents substantial

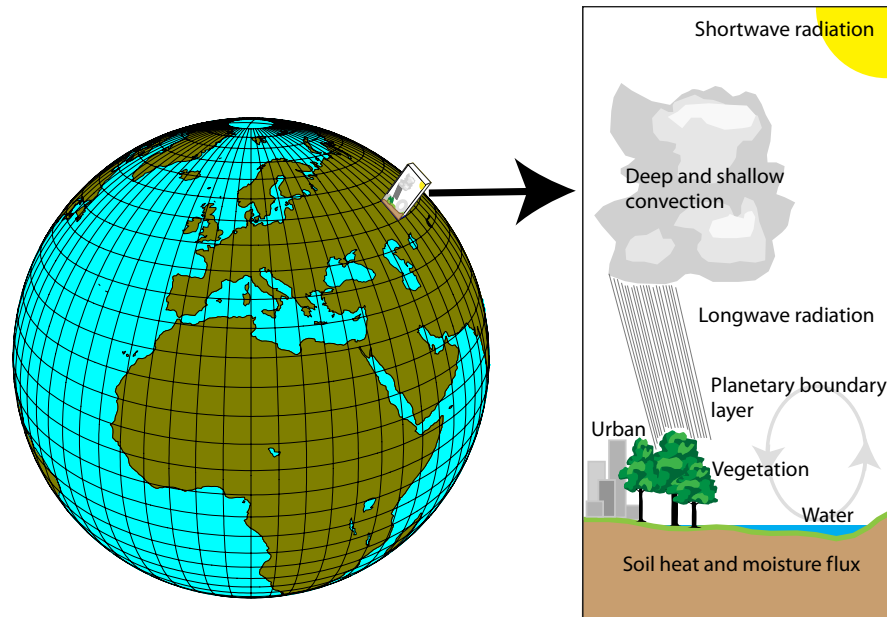


Figure 1 Conceptual illustration of a vertical column of a grid cell within a numerical weather prediction model. Some of the important parameterization schemes, including shortwave and longwave radiation, deep and shallow convection, planetary boundary layer, soil–vegetation–water–atmosphere transfer, and urban effects are shown.

challenges to the community as we seek to balance the benefits gained from smaller grid spacing with the costs associated with updating parameterization schemes.

One of the more significant changes in the past decade is the increased attention to parameterization and the more thoughtful verification and comparison of scheme behavior. In this article, a few of the more common parameterization scheme types found in numerical weather prediction models are briefly discussed to provide some perspective on the challenges to parameterization. Since it is impossible to discuss all the parameterization schemes available for a given physical process, the basic concepts behind the schemes are outlined and output from several schemes are compared to highlight the importance of these schemes to the resulting numerical forecasts. For all the parameterization scheme comparisons, the Advanced Research WRF model is used with all other parameterization schemes held constant and identical initial and boundary conditions. Thus, the differences described later are due to the cumulative effects of the different parameterization schemes on the model forecasts.

Shortwave and Longwave Radiation

Two of the most important and computationally expensive parameterization schemes for climate and weather models are the schemes that predict shortwave and longwave radiation. Shortwave radiation is produced by the sun, passes through the atmosphere where it is both absorbed and scattered, and eventually reaches the Earth's surface. Approximately 40% of the electromagnetic energy emitted by the sun is in the visible region of 0.4–0.7 μm , with 10% in shorter wavelengths

(ultraviolet) and 50% in longer wavelengths (near infrared). Some of the incoming shortwave radiation is reflected back to space, due to the albedo effect, with the remainder absorbed by vegetation and the ground surface. Rayleigh scatter dominates for the shorter wavelengths, producing the blue color of the sky, whereas absorption occurs in clouds for the longer wavelengths.

The surface albedo over land depends upon soil type, soil moisture, vegetation type and vegetation coverage, and human changes to the landscape. The albedo of vegetation and the amount of vegetation coverage change throughout the plant life cycle. Wet soils have a lower albedo than dry soils. In colder climates, newly fallen snow can produce an albedo near 0.9, which then decreases as the snow melts and ages. The albedo over water depends upon the solar zenith angle – the angle between a line pointing toward the sun and the vertical – with the albedo approaching 1 for large zenith angles and approaching 0 for small zenith angles. The presence of clouds also dramatically alters the albedo. Clouds reflect and absorb shortwave radiation and the reduction of shortwave radiation under overcast skies can be considerable. Aerosols created from sea spray, fires, dust, chemical reactions, boreal forests, pollution, and volcanic eruptions also influence shortwave radiation through absorption and scattering. Aerosols can act as cloud condensation nuclei for cloud droplets, thereby helping to create clouds and modifying the albedo.

The surface albedo in many models is determined by a database of land use type and is held constant throughout the forecasts. Satellite analyses and surface reports are used to create maps of snow cover. The presence and amount of cloud is either predicted explicitly by the model or parameterized using other model variables such as relative humidity and vertical motion. Aerosol optical depth is measured at the

ground and can be estimated by satellites, although not all models include aerosol effects.

Parameterization schemes for shortwave radiation have increased in sophistication in the past few decades. The shortwave irradiance is divided into direct and diffuse contributions, where the direct component is from photons that have not been scattered. Calculation of the direct component follows Beer's Law, with the irradiance at a specific frequency reaching a given vertical level related to the distance traveled through the atmosphere (path length) and the optical depth that depends upon the absorber characteristics and concentration in the layer traversed. Integration over all frequencies yields the total shortwave irradiance. Narrow- or broad-band approaches can be used for the integration over frequency and the correlated- k method, in which the absorption coefficients are rearranged into ascending order over a given frequency interval, can be used to increase integration speed and accuracy. Calculation of the diffuse component is much more difficult, owing to the effects of multiple scattering and leads to an integration over both frequency and zenith angle for upward and downward directed fluxes.

A comparison of two commonly used shortwave radiation schemes from a 6-h forecast valid at 1800 UTC 2 April 2006, produced by the WRF model using 20-km grid spacing, indicates that the amount of radiation that reaches the Earth's surface is similar over a majority of the model domain (Figure 2). Even under clear skies, however, differences often exceed 30 W m^{-2} . The largest differences stretch from the Gulf of Mexico northward to the Midwestern states in association with larger values of precipitable water, 1000–500 hPa mean relative humidity, and model-produced rainfall (Figure 3). These are areas where cloud cover is diagnosed in the model, thereby reducing the shortwave radiation reaching the surface. The maximum difference between the two schemes at any grid point is 770 W m^{-2} while the mean absolute difference over all grid points is 79 W m^{-2} . The larger differences in the two schemes are related to how they treat cloud effects.

Longwave (terrestrial) radiation is produced at the Earth's surface and extends from the near infrared through the infrared portions of the spectrum with a peak near $11 \mu\text{m}$. Numerous atmospheric gases absorb and emit longwave radiation, including carbon dioxide, water vapor, ozone, methane, and nitrous oxide. Some of these gases are well mixed in the atmosphere, while others can have large horizontal and vertical gradients. Other atmospheric gases such as nitrogen and oxygen are almost totally unaffected by infrared light. Clouds also are very efficient absorbers and emitters of longwave radiation, which is particularly important since clouds cover over half the globe at any given moment.

Parameterization schemes for longwave radiation are similar to those for shortwave radiation, although the range of frequencies that need to be integrated is larger and scattering effects can be neglected. A two-stream approach is used whereby the upward and downward radiative fluxes are calculated separately and either narrow- or broad-band methods are used to integrate the radiative transfer equation over all frequencies. The equations appropriate for calculating the longwave radiative flux are

$$F_U(z) = \int_0^\infty \pi B_\nu(0) \tau_\nu^f(z, 0) d\nu + \int_0^\infty \int_0^z \pi B_\nu(z') \frac{d\tau_\nu^f}{dz'}(z, z') dz' d\nu, \quad [1]$$

$$F_D(z) = \int_0^\infty \int_z^\infty \pi B_\nu(z') \frac{d\tau_\nu^f}{dz'}(z, z') dz' d\nu, \quad [2]$$

where F_U and F_D are the upward and downward fluxes through height z , B_ν is the Planck function, ν is frequency, and τ_ν^f is the diffuse transmission function defined over a hemisphere. The diffuse transmission function depends upon the concentration of the attenuating gas over the integration depth and its absorption coefficient, which is a function of temperature and pressure. The first term in F_U represents the attenuation of the longwave radiation emitted from the ground surface, while the second term represents the longwave radiation emitted by the atmosphere. The single term in F_D represents the longwave radiation emitted by the atmosphere. Longwave radiation parameterization schemes differ in how these integrals are calculated and how the absorption coefficients are handled.

Comparisons between line-by-line radiative transfer models and longwave parameterization schemes have led to numerous improvements, including the use of the correlated- k approach. However, as with the shortwave radiation schemes, differences between longwave schemes are most discernable when clouds are present (Figure 4). The largest differences between two longwave radiation schemes largely overlap the regions of largest differences in the shortwave schemes (Figure 2). Differences of up to 98 W m^{-2} in longwave fluxes are indicated, with a mean absolute difference of 11 W m^{-2} over all grid points.

Soil–Vegetation–Water–Atmosphere Transfer

Gradients in soil and vegetation conditions, as well as gradients in the radiation fluxes that reach the surface, lead to horizontal gradients in surface sensible and latent heat fluxes from the ground surface to the atmosphere. The partitioning of sensible and latent heat fluxes directly influences near surface temperature, moisture, and winds as well as the depth of the PBL. Strong horizontal gradients in sensible heat flux can lead to nonclassical mesoscale circulations, such as vegetation breezes, and may influence low-level cloud development. The harvesting of crops dramatically changes vegetation conditions over a matter of days, providing a clear human influence on near surface conditions. The presence of buildings and urban areas also influence the partitioning of the sensible and latent heat fluxes and can lead to the development of urban heat islands. The effects of urban areas are sometimes calculated in separate parameterizations owing to the importance of forecasts in urban areas to energy demand and the sheer size of the population that resides in urban environments.

The strong influence of vegetation on the partitioning of sensible and latent heat fluxes creates a number of challenges to land surface parameterization. The movement of water from the soil to the roots to the stomata in the leaves and hence to

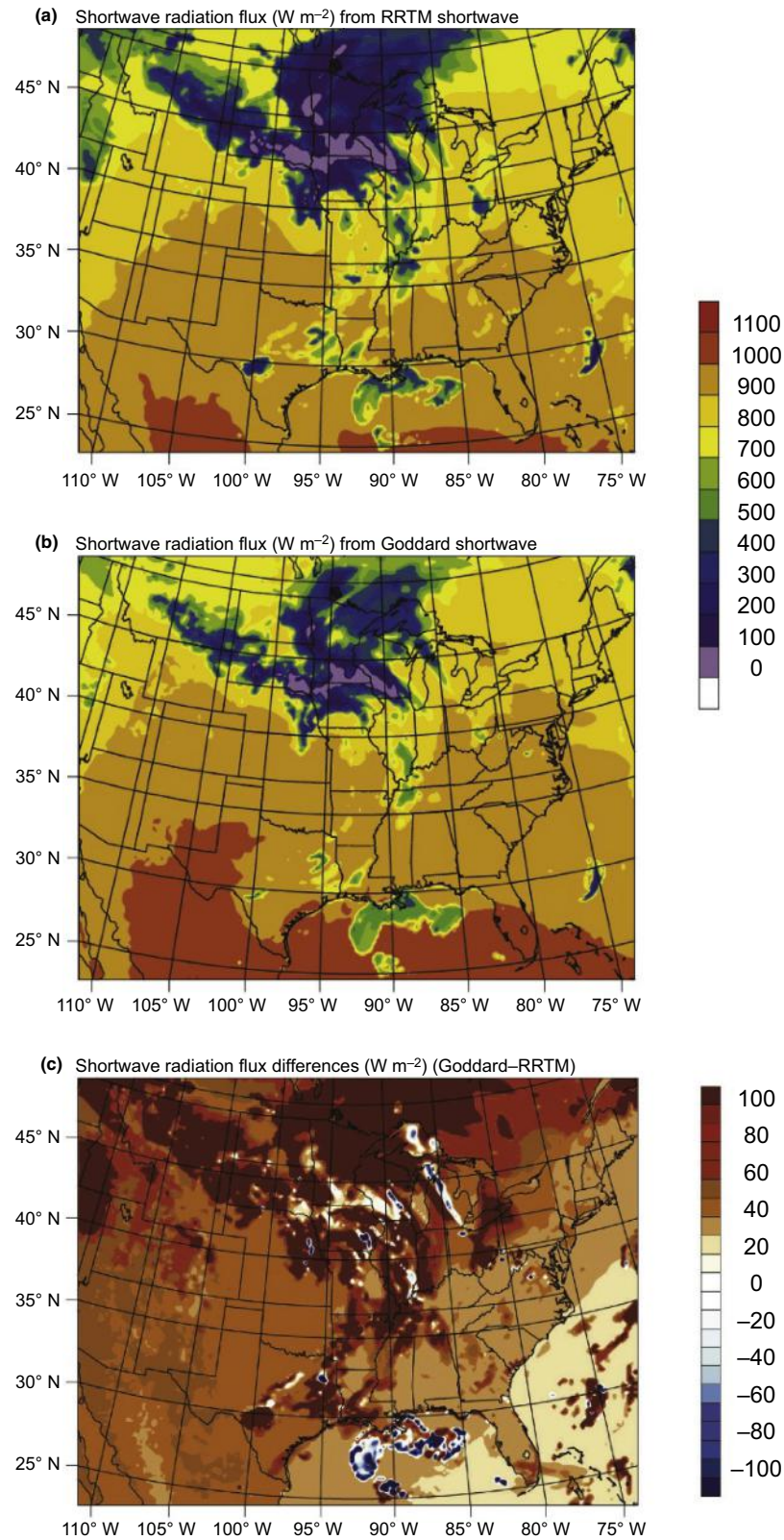


Figure 2 Shortwave radiation flux (W m^{-2}) reaching the Earth's surface at 1800 UTC 2 April 2006 (6-h forecast) calculated using the WRF model with 20-km grid spacing with the (a) rapid radiative transfer model (RRTM) and (b) Goddard shortwave parameterizations. The difference (Goddard-RRTM) is shown in (c). Values indicated by color bars.

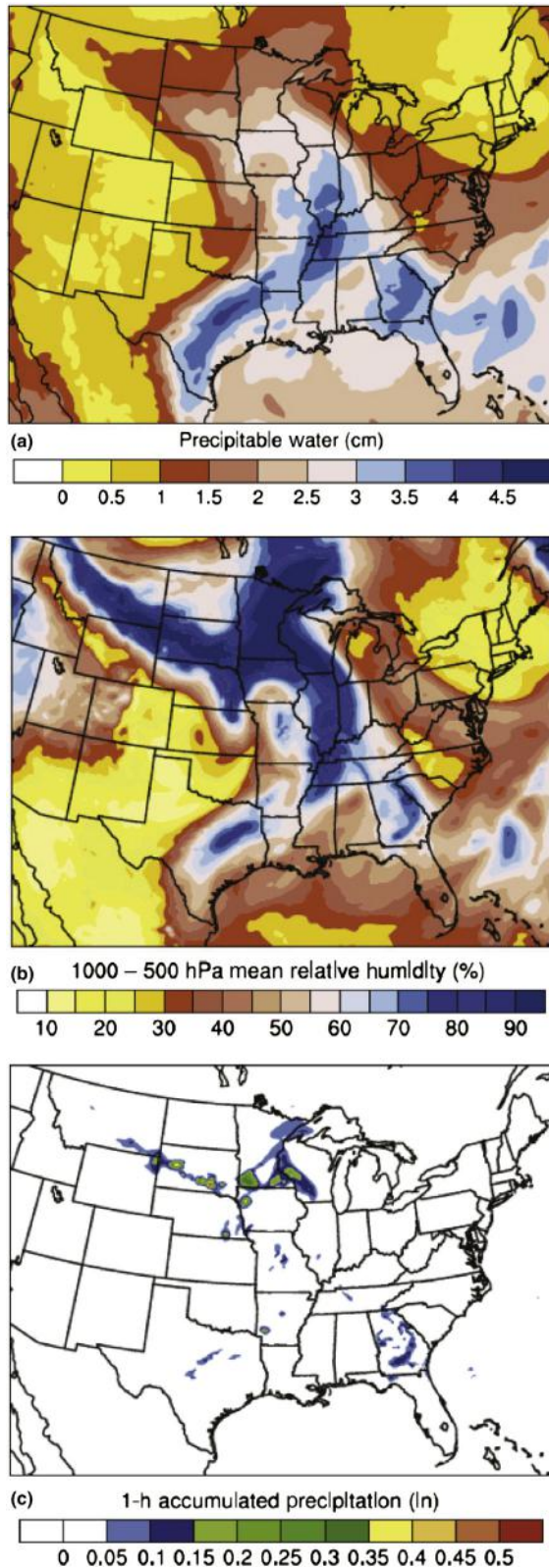


Figure 3 Fields of (a) precipitable water (cm), (b) 1000–500 hPa mean relative humidity (%), and (c) 1-h accumulated rainfall (inches) valid at 1800 UTC 2 April 2006 (6-h forecast) using the WRF model with the RRTM shortwave parameterization. Values indicated by color bars.

the atmosphere must be represented in both weather and climate models. Yet plants have developed strategies to minimize water loss. Whereas the need for plants to obtain carbon dioxide for photosynthesis drives the exposure of saturated plant tissues to the atmosphere, the majority of land surface models simply assume that the stomata will remain open until the water supply from the soil is cut off. Newer schemes are starting to include calculations of photosynthesis that directly influence vegetation transpiration.

Another challenge to land surface parameterization is the accurate depiction of the initial condition of the land and vegetation states. Soil moisture can have large variations over small distances due to changes in rainfall, soil type, terrain slope, vegetation, and soil conditions. Unfortunately, there are few *in situ* measurements of soil moisture and temperature, and satellite observations can only provide information on soil moisture in the first few centimeters below the surface. Thus, the initial soil moisture and temperature values often are provided by land data assimilation schemes that force a land surface parameterization with observations of rainfall, surface meteorological observations, and satellite estimates of incoming radiation. The use of observations to force the land surface parameterization is found to yield reasonable soil moisture and temperatures when compared against available observations. Vegetation conditions can change dramatically from season to season and year to year, with changes in the timing of spring green up to several weeks typical. Farmers also rotate their crops or leave the fields fallow, altering the vegetation type. While vegetation coverage and type can be estimated from satellite, many models use a monthly vegetation climatology that may differ significantly from conditions on the ground. One of the more important parameters in land surface models is the vegetation fraction – the fraction of the model grid cell where midday solar radiation is intercepted by actively transpiring vegetation. The vegetation fraction directly influences the latent heat flux by weighting the bare soil and canopy transpiration. Differences of 30% or larger between an interpolated monthly climatology of vegetation fraction and real-time estimates from satellite are quite common.

A comparison of two land surface schemes indicates that while the general patterns of sensible and latent heat fluxes are similar (Figure 5), differences of 100 W m^{-2} or more can be present at individual grid points. The pattern of latent heat flux matches well with the climatological vegetation fraction used by the schemes (Figure 6) with larger values of vegetation fraction leading to larger latent heat flux. In contrast, the regions with the largest values of sensible heat flux are located across the southwestern portion of the model domain and correspond with low values of vegetation fraction and larger values of shortwave radiation. Thus, the largest sensible heat fluxes occur where there is little vegetation, few clouds, and strong incoming solar radiation. The important role of vegetation is seen through binning the latent heat flux differences by the vegetation fraction value. Results indicate that the two land surface schemes yield the largest mean differences in latent heat flux for values of vegetation fraction less than 40% and greater than 60% (Figure 7). The differences are due to the different formulations for bare soil and canopy transpiration.

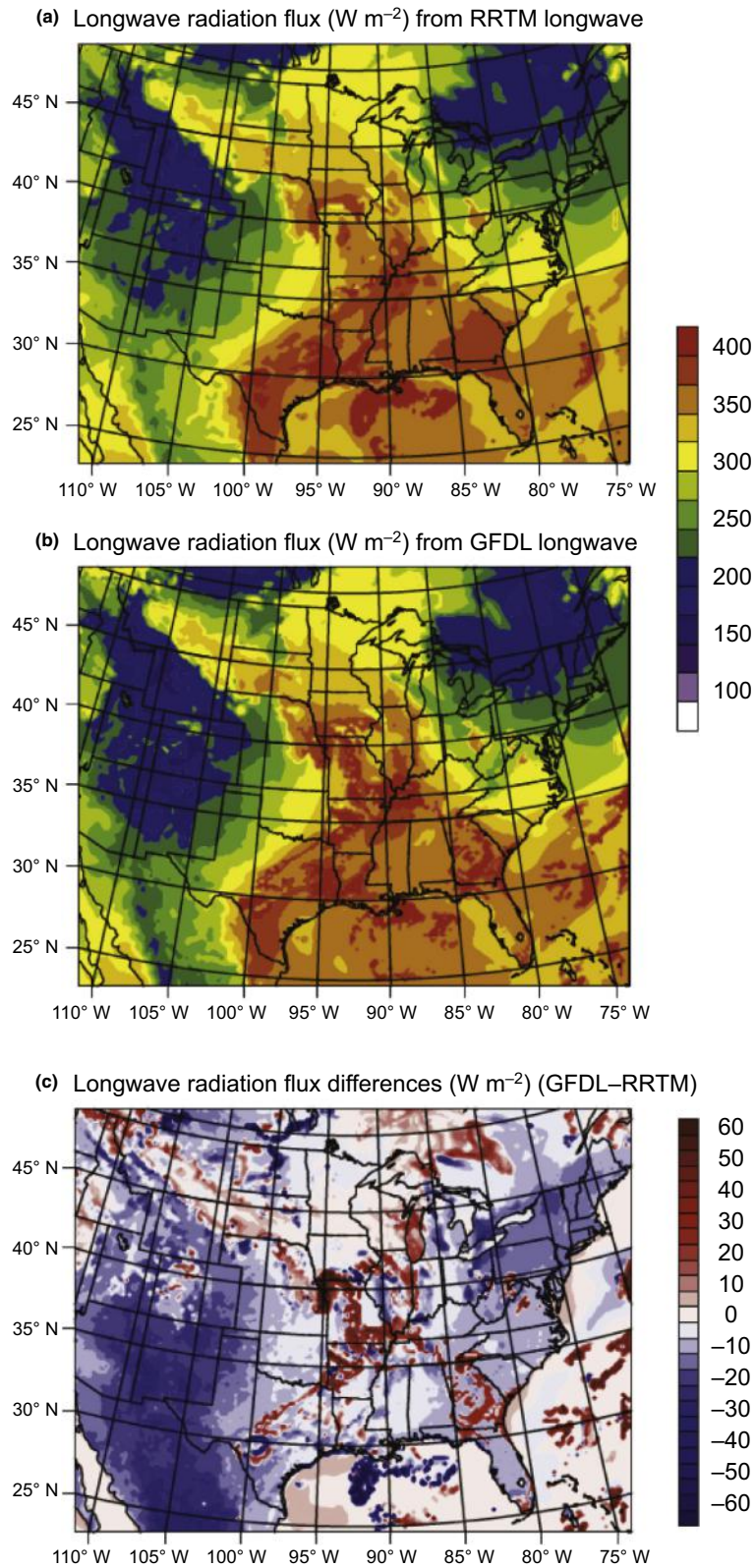


Figure 4 Longwave radiation flux (W m^{-2}) reaching the Earth's surface at 1800 UTC 2 April 2006 (6-h forecast) calculated using the WRF model with 20-km grid spacing with the (a) rapid radiative transfer model (RRTM) and (b) Geophysical Fluid Dynamics Laboratory (GFDL) longwave parameterizations. The difference (GFDL–RRTM) is shown in (c). Values indicated by color bars.

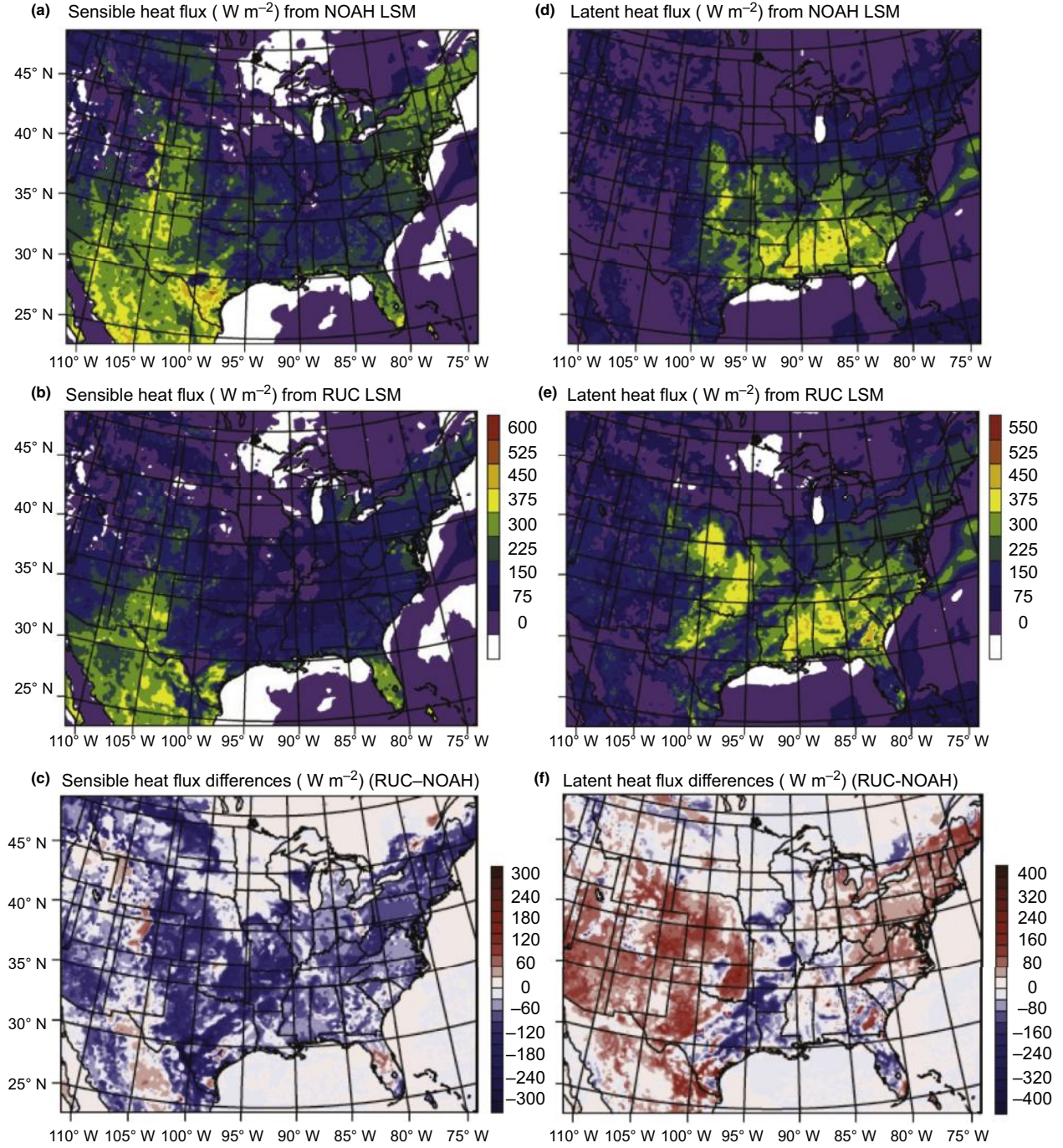


Figure 5 Sensible and latent heat fluxes (W m^{-2}) at 1800 UTC 2 April 2006 (6-h forecast) calculated using the WRF model with 20-km grid spacing. Sensible heat fluxes from the (a) NOAA, (b) Rapid Update Cycle (RUC), and (c) their difference (RUC–NOAH). Latent heat fluxes from the (d) NOAA, (e) RUC, and (f) their difference (RUC–NOAH). Values indicated by color bars.

Over bare soil, a common approach is to define the latent heat flux Q_{EB} as

$$Q_{EB} = (1 - \sigma_f) \left(\frac{\Theta_1 - \Theta_w}{\Theta_{fc} - \Theta_w} \right)^a E_p, \quad [3]$$

where E_p is the potential evaporation, σ_f is the vegetation fraction (0–1), Θ_1 is the volumetric water content within the top-most soil layer, Θ_{fc} is the volumetric water content field capacity, Θ_w is the wilting point – the volumetric water content at which plants can no longer bring water out of the

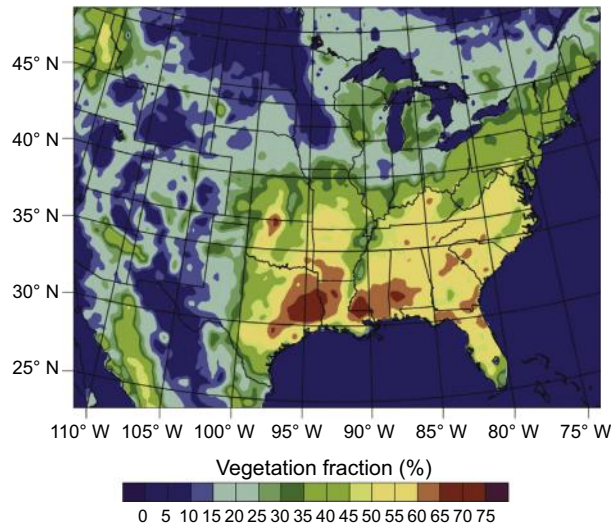


Figure 6 Vegetation fraction (%) valid at 2 April 2006 as interpolated from a monthly climatology.

soil – and α is a scaling parameter. Thus, the latent heat flux is determined largely by the soil moisture value in the top-most soil layer. For the two schemes compared, α is set to 1 for the RUC scheme and is set to 2 for the NOAH scheme. While the calculations of E_p are not identical between the two schemes, the comparisons are consistent with the different values of α leading to larger latent heat fluxes from bare soil in the RUC scheme than in the NOAH scheme as seen in **Figure 7** for low values of vegetation fraction. The differences in latent heat flux for the higher values of vegetation fraction are due in part to the NOAH scheme including the effects of solar radiation, vapor

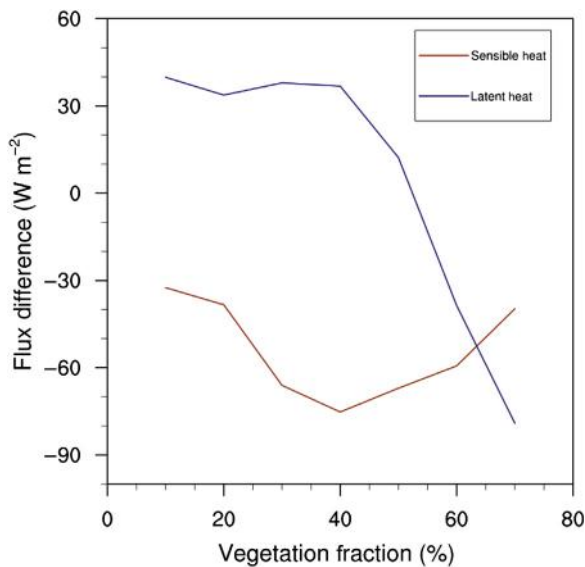


Figure 7 Mean flux difference (W m^{-2}) (RUC–NOAH) for sensible and latent heat flux values as a function of vegetation fraction (%) calculated using WRF model forecasts valid at 1800 UTC 2 April 2006 (6-h forecast).

pressure deficit, and air temperature on the canopy resistance, whereas the RUC scheme primarily uses only the root-zone soil moisture.

At first blush, one would think that calculating heat fluxes over water would be easier than over land, since ocean surface temperature varies slowly and there is no vegetation to complicate the parameterization. Indeed, for short range forecasts ocean water temperatures are often assumed constant. Challenges arise instead from the presence of waves, which have a large influence on the fluxes through alterations of the surface roughness. Sea spray is produced by wind gusts, bursting bubbles, and breaking waves, producing an atmospheric layer above the ocean in which water droplets evaporate and their temperatures come into equilibrium with the environment. This situation leads to alterations in the sensible and latent heat fluxes and produces a net drag on the airflow. Observations upon which to develop a parameterization are very challenging over the ocean, especially under high wind conditions.

Planetary Boundary Layer

The PBL is influenced directly by the Earth's surface and responds to surface fluxes with a timescale of an hour or less. Thus, the behaviors of the land surface and the PBL are tightly coupled. During the daytime, the boundary layer can extend to several kilometers above the ground and be fully turbulent throughout its depth, whereas at nighttime the boundary layer may be as shallow as a few tens of meters and turbulence is intermittent. These large changes in boundary layer depth during the diurnal cycle present unique challenges to parameterization.

The daytime PBL tends to be dominated by strong surface heat flux and vertical mixing due to large thermals – bubbles of buoyant air that start near the ground and rise into the boundary layer – that are roughly as wide as the boundary layer is deep. As these thermals impinge upon the top of the boundary layer they overshoot their level of neutral buoyancy and then sink back into the boundary layer bringing with them curtains of air from above. This entrainment of air from above the boundary layer into the boundary layer is a key factor in determining boundary layer depth and structure. The potential temperature is often well mixed (nearly constant with height) in the absence of clouds; moisture and momentum are not often well mixed. Surface sensible heat flux and entrainment both act to warm the boundary layer, while surface latent heat flux and entrainment often act in opposition since entrainment tends to bring drier air into the boundary layer. When clouds form within the boundary layer, the turbulence is modified and sometimes the turbulence within the cloud decouples from the turbulence within the rest of the boundary layer. Turbulence is intrinsic to the boundary layer and the largest boundary layer eddies provide turbulence its distinguishing characteristics.

The nocturnal PBL is dominated by longwave radiational cooling at the surface and episodic mixing events due to wind shear, drainage flows, and gravity waves. The nocturnal boundary layer is stable with potential temperature increasing with height. The lack of a strong controlling influence on the nocturnal boundary layer, and the ever-shifting interplay

between turbulent mixing and viscous and buoyant dissipation, leads to a challenging parameterization problem.

It is helpful to classify PBL parameterizations into two broad types: local and nonlocal closure schemes. Local closure schemes relate the unknown boundary layer variables to model variables at nearby vertical grid points. Many local schemes require the explicit prediction of turbulent kinetic energy, thereby adding a new variable to the model that must be integrated in time. The most significant concern with local schemes is that they assume mixing is related to the local turbulent kinetic energy, whereas observations indicate that large eddies produce most of the mixing. Nonlocal closure schemes relate the unknown boundary layer variables to model variables anywhere within and near the boundary layer. A nonlocal perspective often is shown to better explain the turbulence characteristics of a boundary layer. However, as with all parameterizations, a number of assumptions are needed to develop a nonlocal closure scheme that may not be well suited for all environments. Nonlocal schemes typically are used for daytime boundary layers and revert to a local scheme at night.

The difference between local and nonlocal closure can be illustrated using the turbulence diffusion equation for potential temperature (θ), in which the tendency due to boundary layer effects is expressed by

$$\frac{\partial \theta(z)}{\partial t} = \frac{\partial}{\partial z} \left[K_\theta(z) \left(\frac{\partial \theta(z)}{\partial z} - \gamma_\theta(z) \right) \right], \quad [4]$$

where K_θ is the eddy diffusivity coefficient and γ_θ is a correction to the local gradient that incorporates the contributions of the large-scale eddies. For a local closure scheme, the values of K_θ and γ_θ are complicated functions of the turbulent kinetic energy at the same vertical level. No information from lower or higher levels directly influences these terms. In contrast, for a nonlocal closure scheme the values of K_θ and γ_θ are vertical curve fits to values derived from large eddy simulations or observations, resulting in a specified vertical profile shape that depends upon the depth h of the boundary layer, the convective velocity scale w^* , and the entrainment coefficient k_e , often resembling

$$K_\theta(z) = w^* h \left(\frac{z}{h} \right)^{4/3} \left(1 - \frac{z}{h} \right)^2 \left(1 + \frac{k_e z}{h} \right). \quad [5]$$

Thus, the structure of the boundary layer, and in particular its depth and assumptions regarding entrainment, has a direct influence on the eddy diffusivity used in the parameterization. An equation for h is typically included in a nonlocal scheme in comparison to diagnosing boundary layer height from the turbulent kinetic energy as done within a local closure scheme.

Comparisons of five PBL schemes highlight the differences in scheme behavior and support the notion that local and nonlocal closure schemes produce consistent differences (Figure 8). The local closure schemes (MYJ, QNSE, MYNN) have lower mean boundary layer potential temperature, higher mean boundary layer mixing ratio, lower boundary layer depths, and higher values of surface-based convective available potential energy (CAPE) than the two nonlocal closure schemes (ACM2, YSU). One might hope that these consistent differences yield a clear conclusion that one type of scheme performs better than the other. Unfortunately, the comparison

results are not so clear. The local closure schemes better match the observations for boundary layer depth and surface-based CAPE, while the nonlocal closure schemes better match the observations of mean boundary layer potential temperature and mixing ratio (Figure 8). Based upon the temporal paucity of current standard observations with which to verify boundary layer structures, it is easy to argue that improvements to boundary layer schemes are greatly hampered by the lack of more frequent observations.

A more persuasive illustration of the differences in boundary layer schemes are predicted soundings from the 23-h forecasts of the same five schemes valid at Topeka, Kansas, on 2300 UTC 10 May 2011 (Figure 9). The two nonlocal closure schemes yield the deepest boundary layers as easily seen in the mixing ratio profiles. The local closure schemes predict a shallower boundary layer and much higher low-level mixing ratios. The effects of the PBL schemes on the mixing ratios between 800 and 700 hPa is dramatic and is routinely seen when multiple boundary layer schemes are compared. Also notice the differences in the low-level warm layer at the top of the boundary layer, which is not well predicted by any of the schemes.

Convection and Microphysics

Moist convection occurs in many sizes and shapes and plays a very large role in the success or failure of numerical model forecasts. Convection can be deep, such as thunderstorms or squall lines, or shallow as often found over the oceans off the west coasts of continents. Deep convection vertically spans much of the troposphere and produces precipitation, thereby acting to warm and dry the environment. Shallow convection vertically spans only a small portion of the troposphere and acts to cool and moisten the environment in the upper-half of the cloud and warm and dry the lower half of the cloud layer. Shallow convection produces no net warming or drying, as it is nonprecipitating.

The effects of convection can be treated implicitly using a convective parameterization scheme in which the effects of convection are parameterized prior to the occurrence of saturation at the model grid point. These schemes are used when model grid spacing exceeds 20 km. As model grid spacing drops below 4 km or so, then convection is treated explicitly by including parameterizations for microphysical processes that produce cloud water, rainwater, cloud ice, snow, graupel, and hail. A gray zone exists for model grid spacing between 4 and 20 km, since convective parameterization is needed to produce reasonable results, yet some of the assumptions used to develop convective schemes may no longer hold. In addition, some convective schemes are designed to work in tandem with a microphysics scheme as they act to encourage grid-scale saturation and the development of mesoscale circulations associated with stratiform convective regions.

There are a wide variety of convective parameterization schemes and numerous ways to classify their behavior. One can separate the schemes into deep-layer and low-level control schemes based upon the vertical extent of the forcing that controls the convective activity. Deep-layer control schemes often tie the development of convection to the creation of

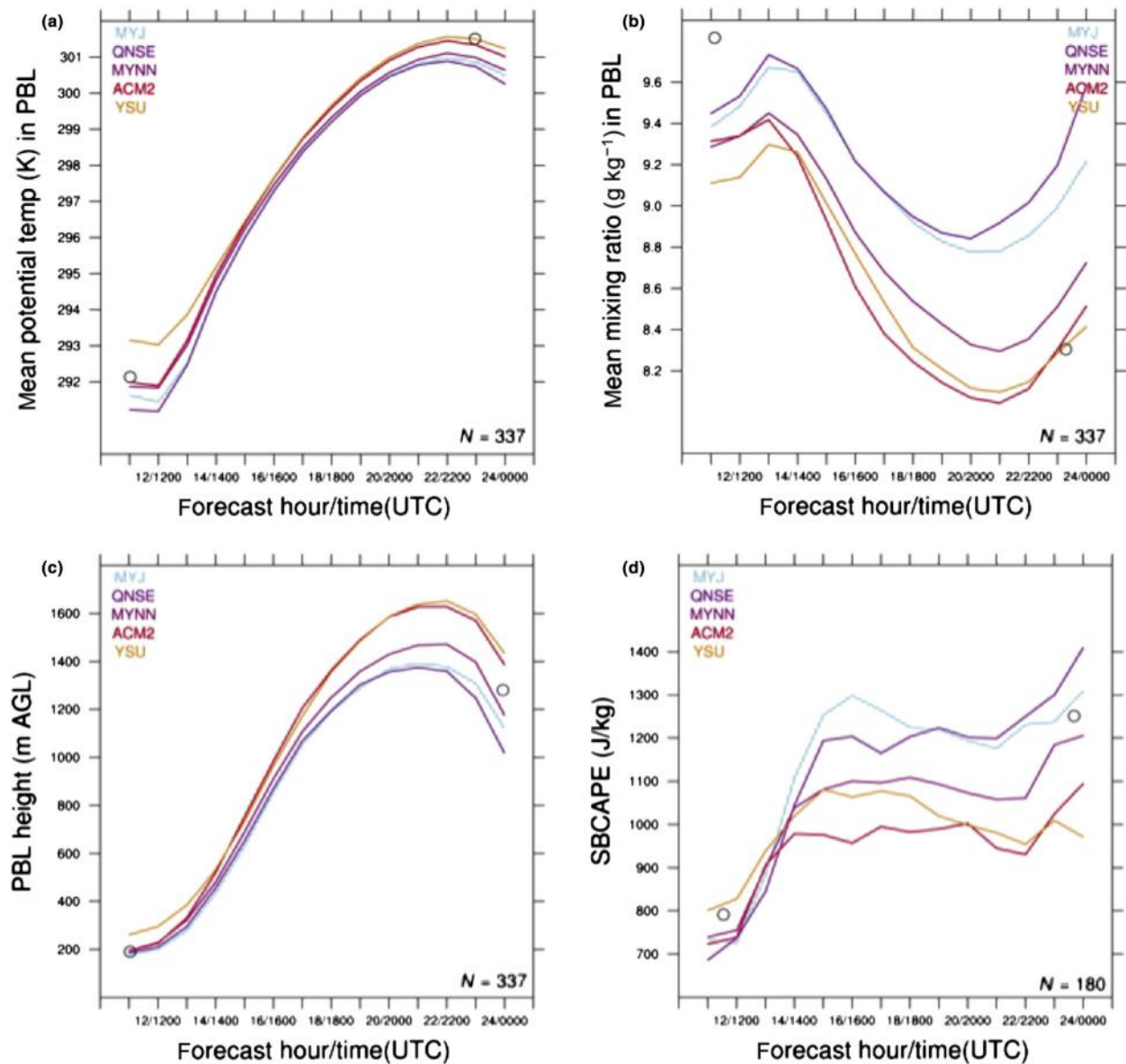


Figure 8 Mean boundary layer parameters computed over 25 days at selected radiosonde sites across the eastern two-thirds of the United States from five WRF model forecasts that vary only by PBL scheme for 12–24-h forecasts valid 1200 to 0000 UTC using data from the storm-scale ensemble forecast (SSEF) during 2011. Parameters are (a) mean boundary layer potential temperature (K), (b) mean boundary layer mixing ratio (g kg^{-1}), (c) mean boundary layer depth (m), and (d) mean surface-based CAPE (J kg^{-1}). The number of soundings comprising the mean at each hour is shown in the bottom right of each panel. The open circle denotes the mean of the radiosonde observations at the approximate time of launch. Local closure schemes are MYJ, QNSE, and MYNN, while nonlocal closure schemes are ACM2 and YSU. All WRF forecasts started at 0000 UTC in the period 9 May through 10 June 2011. The SSEF uses 4-km grid spacing and is run by the Center for Analysis and Prediction of Storms.

CAPE such that convection immediately consumes the CAPE created by the model. Low-level control schemes allow for CAPE to be stored and transported before it is activated by the scheme as these schemes focus on how convection is initiated through the removal of convective inhibition. However, one could also distinguish between convective schemes based upon whether or not scheme behavior is sensitive to moisture or instability, or whether or not the scheme tendencies are static or dynamic. The complexity of convective forms and sizes makes

classifying convective schemes difficult and illustrates the breadth of parameterization approaches developed.

One of the more significant advances in the past decade has been the fairly rapid move toward convection-allowing models that only use a microphysics parameterization to represent convective effects. These models no longer need to use a convective parameterization, and so avoid the many well-known limitations associated with the parameterization of deep convection. Results indicate that convection-allowing

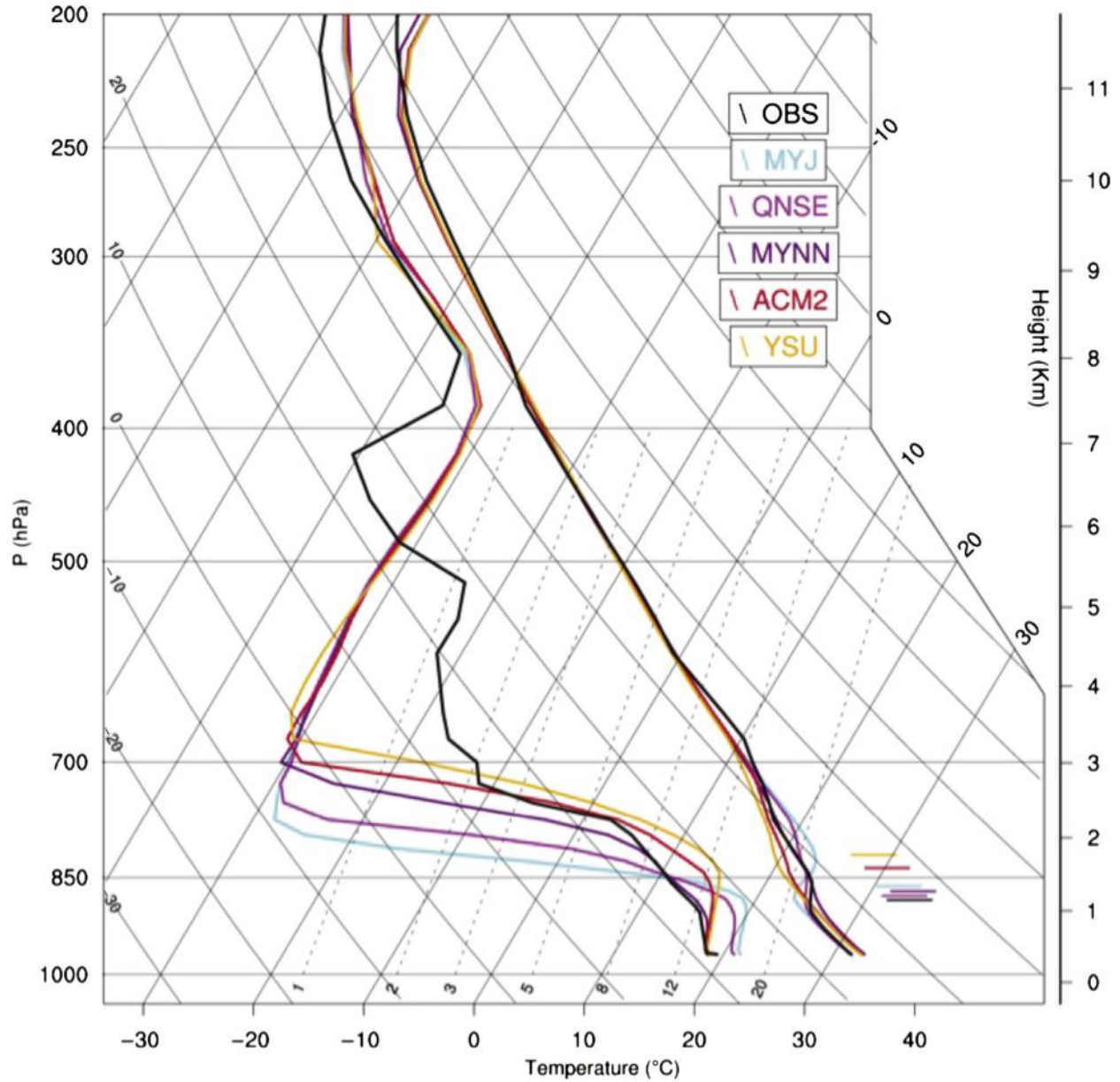


Figure 9 Skew- T log p diagram of an observed sounding taken at Topeka, Kansas at 2306 UTC 10 May 2011 (black line) and the corresponding profiles from five 23-h WRF model forecasts that vary only by PBL scheme. The horizontal lines to the right of the temperature trace show the diagnosed PBL height for each sounding. Local closure schemes are MYJ, QNSE, and MYNN, while nonlocal closure schemes are ACM2 and YSU. Forecasts produced by the 4-km SSEF system run by the Center for Analysis and Prediction of Storms.

models are much more capable of reproducing the observed behaviors of deep convection, such as the nocturnal propagation of convection across the central United States during the warm season, the development of surface cold pools and stratiform rain regions, and yield more accurate predictions of rainfall amounts. These models have also been very successful in producing realistic convective modes, or the forms of convection, such as an isolated thunderstorm, a supercell thunderstorm, a line of convective cells, or a bowing line of convective cells. However, we are also learning that the model forecasts are very sensitive to the microphysics parameterization.

A comparison of composite reflectivity forecasts from six microphysics schemes valid at 0200 UTC 10 June 2011 illustrates the differences often seen (Figure 10). All the schemes include predictive equations for cloud water, rainwater, ice, snow, and graupel, except for the Ferrier scheme that only has equations for cloud water, rainwater, and snow. There is general agreement that convection stretches from southwest to northeast but the intensity of the convective line and extent of the stratiform regions varies greatly among the schemes. Some only produce intense convection to the north, while others have intense convection along the entire convective zone.

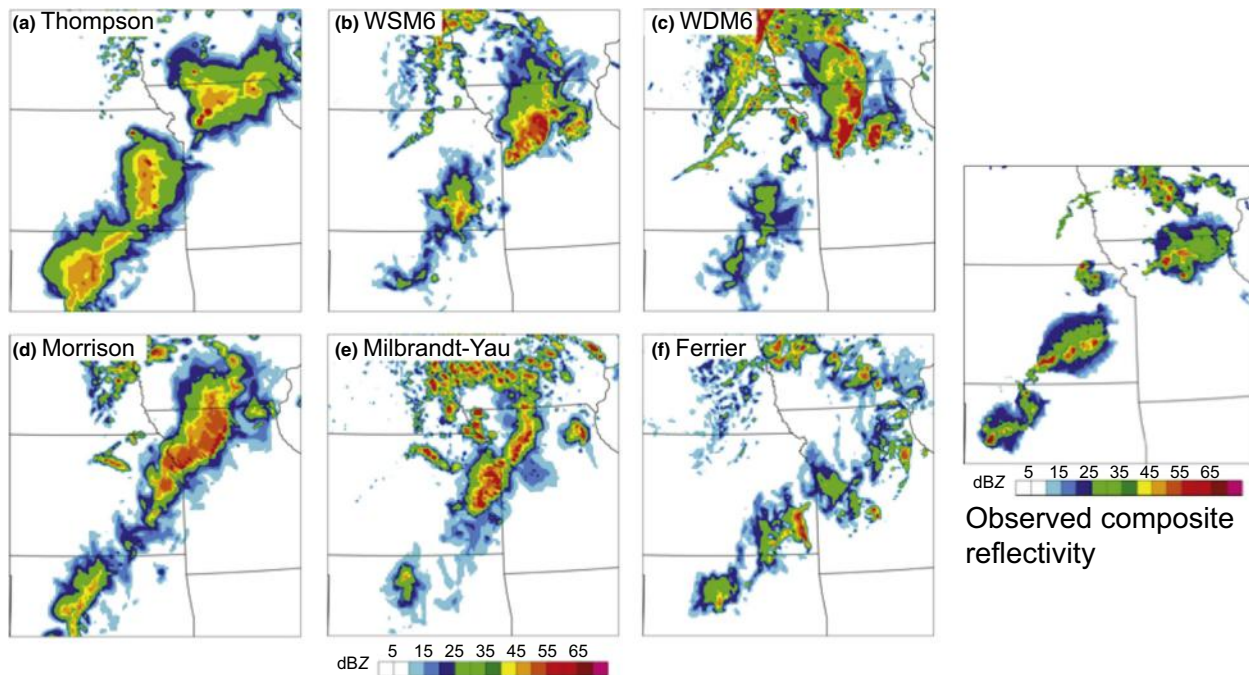


Figure 10 Simulated composite reflectivity from 26-h forecasts valid at 0200 UTC 10 June 2011 calculated from six WRF model forecasts that vary only by microphysics scheme. The schemes are (a) Thompson, (b) WSM6, (c) WDM6, (d) Morrison, (e) Milbrandt-Yau, and (f) Ferrier. The observed composite reflectivity from the National Severe Storms Laboratory National Mosaic and Multi-Sensor QPE product is also shown. The Ferrier scheme is the only scheme that does not include graupel. The WSM6 and Ferrier schemes are single-moment bulk schemes, while the other schemes are double-moment bulk schemes for at least some of the hydrometeor types.

When compared to the observed composite reflectivity, none of the forecasts is a perfect match.

The associated 2-m temperature and 10-m wind forecasts valid at the same time again show dramatic differences between the schemes. The pools of evaporatively cooled downdraft air that spread out horizontally underneath precipitating clouds, commonly called cold pools, vary from being fairly weak in the Thompson scheme to very strong and extensive in the Ferrier scheme (Figure 11). The 10-m wind fields also show vastly different features, with the WSM6 and WDM6 schemes producing a strongly divergent wind field near the center of the convective zone and the rest of the schemes having a weakly divergent wind field in the same region. The column-integrated graupel mixing ratio also varies widely between the schemes, with the Milbrandt-Yau scheme having the largest coherent region of graupel.

The ability of four of these microphysics schemes to accurately predict rainfall location over many cases provides another interesting comparison. Results indicate that the Thompson scheme at forecast hour 30 has no coherent spatial bias in the location of rainfall forecasts in excess of 0.5 inches, whereas the other schemes predict rainfall too far to the south or southeast (Figure 12). These spatial biases likely are due to cold pools that are too strong due to excessive evaporation.

Future Directions

Forecasts from numerical weather prediction models have shown steady improvement over the past decades, leading to

people relying on forecasts more and more to help make personal and economic decisions. When significant snowstorms are predicted, airlines now cancel flights and move aircraft out of the affected region. When supercell thunderstorms are predicted, emergency managers take action to keep citizens safe by deploying storm spotters and responders ahead of the storms. The success of numerical weather prediction is due in large part to the improved parameterization schemes that are essential components of numerical weather prediction models. Without parameterization schemes, forecasts from numerical models would be nearly useless for meeting the weather forecast needs of society.

As model grid spacing decreases, the need for some parameterization schemes will disappear. At grid spacing below 4 km, there is no clear need for convective parameterization. At grid spacing below 50 m, there is no clear need for boundary layer parameterization. However, parameterizations for radiation, soil-vegetation-water-atmosphere transfer, and microphysics will always be needed as these physical processes operate on the molecular scale.

One of the challenges of the coming decade is how to handle the gray zones in which the physical process begins to be represented explicitly by the model but is not entirely represented. This is already occurring for convective parameterization for grid spacing below 10 km or so. But it also will occur for boundary layer parameterization when grid spacing approaches 1 km and the model begins to explicitly produce boundary layer circulations. These circulations represent the model trying to produce the large eddies observed in the

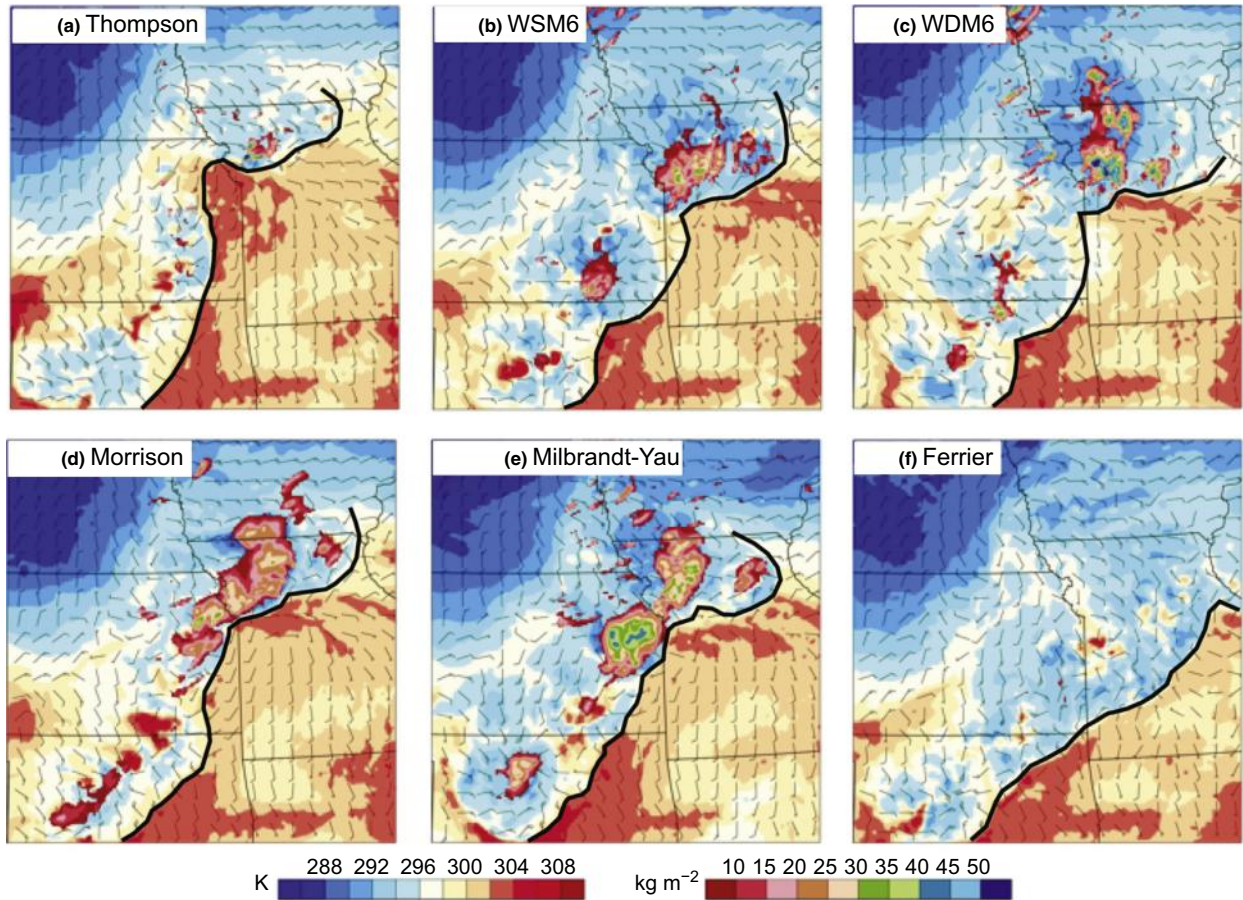


Figure 11 Fields of 2-m temperature (color filled by the scale on the bottom left), 10-m winds (full barb = 5 m s^{-1}), and column-integrated graupel mixing ratio (scale on the bottom right) from 26-h forecasts valid at 0200 UTC 10 June 2011 calculated from six WRF model forecasts that vary only by microphysics scheme. The schemes are (a) Thompson, (b) WSM6, (c) WDM6, (d) Morrison, (e) Milbrandt-Yau, and (f) Ferrier. The black line depicts the eastern edge of the surface cold pool.

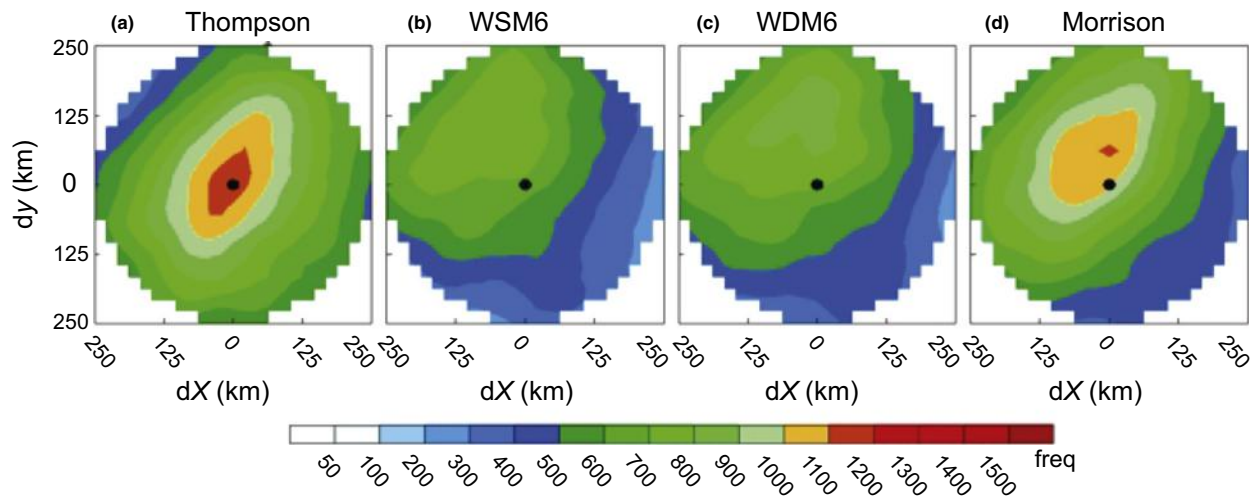


Figure 12 Composite frequencies of observed rainfall greater than 0.50 inch relative to grid points forecasting rainfall greater than 0.50 inch at forecast hour 30 from SSEF members using (a) Thompson, (b) WSM6, (c) WDM6, and (d) Morrison microphysics parameterizations. Boldface dot in each panel marks the center of the composite domain. From Clark, A.J., et al., 2012. An overview of the 2010 Hazardous Weather Testbed experimental forecast program spring experiment. Bulletin of the American Meteorological Society 139, 55–74.

boundary layer and so will yield greater vertical mixing. Yet these large pseudo-eddies will not be able to account for all the mixing that occurs in the boundary layer, and so some additional boundary layer parameterization will be needed.

The challenges for radiation parameterization are different. Most parameterizations today assume that the radiation streams are directed vertically through a single model grid cell as depicted in [Figure 1](#). For a grid spacing of 40 km, this is a very reasonable assumption. As grid spacing drops below 10 km, however, the sunlight that reaches the ground in one grid cell actually passes through several neighboring grid cells. Similar arguments apply for longwave radiation. Thus, modifications to the radiation parameterizations clearly are needed as grid spacing decreases. The effects of terrain slope on radiation also are neglected in many models and deserve further attention.

The parameterization of microphysical processes faces yet another type of challenge. The past few years have seen the development of a number of double-moment bulk microphysical schemes in which both particle mixing ratio and particle number concentration are predicted. These double-moment schemes should allow the schemes to represent correctly the microphysical responses within a wider variety of environments, as particle mixing ratio and concentration can vary independently of each other. The challenge is that the number of tunable parameters in these parameterizations is large and verification data are limited. It may be that information on hydrometeor type, as can be estimated from dual-polarization radars, will be helpful in tuning double-moment microphysics schemes.

There are parameterizations for a number of other physical processes that have not been discussed. Parameterizations exist for fog, orographic drag, shallow convection, urban areas, and

the effects of wind turbines. These are all valid physical processes or effects that may be needed in a numerical model to increase its capability or forecast skill. The need for accurate parameterizations to propel advances in forecast skill suggests that continued and increased attention to model parameterization will be needed.

See also: **Clouds and Fog:** Cloud Modeling. **Cryosphere:** Permafrost. **Data Assimilation and Predictability:** Ensemble Prediction. **Hydrology, Floods and Droughts:** Modeling and Prediction. **Land-Atmosphere Interactions:** Overview. **Numerical Models:** Convective Storm Modeling; General Circulation Models; Large-Eddy Simulation; Mesoscale Atmospheric Modeling; Parameterization of Physical Processes: Clouds; Parameterization of Physical Processes: Gravity Wave Fluxes; Parameterization of Physical Processes: Turbulence and Mixing; Regional Prediction Models. **Weather Forecasting:** Seasonal and Interannual Weather Prediction.

Further Reading

- Kalnay, E., 2003. *Atmospheric Modeling, Data Assimilation and Predictability*. Cambridge University Press, Cambridge.
- Pielke, R.A., 1984. *Mesoscale Meteorological Modeling*. Academic Press, New York.
- Stensrud, D.J., 2007. *Parameterization Schemes: Keys to Understanding Numerical Weather Prediction Models*. Cambridge University Press, Cambridge.
- Trenberth, K.E. (Ed.), 1992. *Climate System Modeling*. Cambridge University Press, Cambridge.
- Warner, T.T., 2011. *Numerical Weather and Climate Prediction*. Cambridge University Press, Cambridge.

Parameter Estimation

A Aksoy, University of Miami, Miami, FL, USA and NOAA Hurricane Research Division, Miami, FL, USA

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

The state of the art of parameter estimation in atmospheric sciences is discussed. It is common in numerical weather prediction models to find parameterizations to represent subgrid-scale atmospheric processes such as radiation, cloud microphysics, convection, and turbulence. According to a widely accepted opinion, errors in such parameterizations are major contributors to model error. Parameter estimation aims to reduce such errors by utilizing atmospheric observations within the data assimilation framework. Although early work on parameter estimation mostly utilized variational techniques, with the advent and progress of ensemble-based data assimilation systems, an increasing number of parameter estimation studies based on ensemble-based techniques have recently emerged. The goal here is to highlight the state of the art in parameter estimation through the lens of these most recent atmospheric science publications.

Introduction

Parameter estimation in the field of atmospheric sciences refers to the determination of the best values of certain parameters in a numerical model through data assimilation or other similar techniques. The practice therefore is intimately tied to addressing model deficiencies due to inaccurate parameters. The approach is sometimes also referred to as the inverse modeling problem, although, from the viewpoint of parameter estimation, the distinction is mostly semantic. In some publications, one also encounters the alternative term parameter identification.

Parameters in numerical models can be part of processes that are either explicitly resolved or parameterized at the sub-grid scale. In the former, parameters are part of the model dynamical core of a model and are directly related to the physical changes in momentum and heat. Some examples of such parameters are the angular speed of the rotation of the Earth, the gravitational acceleration, the gas constant for dry air, and the specific heat of dry air at constant pressure. Such parameters are generally associated with universal physical processes and their values are well known with high accuracy.

In the latter case, subgrid-scale parameterizations can lead to large numerical model errors. This is due to two main reasons: (1) limited understanding and observations of the processes lead to large uncertainties in parameter values that quantify these processes; and (2) crude representation of the parameterized processes within one grid volume (or spectral truncation wavelength) leads to parameters that represent processes of multiple spatial and temporal scales. In numerical models, the most common subgrid-scale processes that are parameterized include turbulence in the planetary boundary layer (PBL), moist convection, phase changes of water (microphysical processes), and radiative transfer between the Earth's surface, atmosphere, and space. Each of these has a class of parameterizations with multiple proposed schemes (algorithms). These parameters are not known with high accuracy and must be estimated.

One can therefore envision two main purposes for doing parameter estimation. On the one hand, detailed and targeted observations and advanced data assimilation techniques can be

used to gain a better understanding of the parameterized physical processes themselves. This would aid in the tuning/calibration of numerical models during their development. If a numerical model becomes operational, one can also perform a similar procedure by augmenting the model state with the parameters to be estimated. Although both approaches are procedurally similar, in the former, one is more concerned with better understanding of the physical processes themselves, whereas the other aims to improve numerical forecasts through data assimilation by acknowledging that forecast errors are due to uncertainties in both the initial conditions and the model.

The early literature of parameter estimation in atmospheric sciences has generally focused on the use of variational data assimilation schemes. More recently, and with the proliferation and success of ensemble-based data assimilation systems in providing high-quality analyses at a wide range of atmospheric scales, an increasing number of studies based on these techniques have emerged. The goal here is to summarize the state of art in parameter estimation through the lens of these most recent studies.

Parameter Estimation Methodology

The general methodology of parameter estimation follows closely that of state-only data assimilation. Therefore, further reading in variational and ensemble-based data assimilation techniques are strongly recommended.

The first step in parameter estimation generally involves obtaining an augmented state vector that consists of both the state (control) variables and the parameters to be estimated. Since, by definition, there is no dynamical feedback from the model state to the parameters, the traditional algorithms of data assimilation are needed to be supplemented by specific measures so that optimal solutions for the parameters can be obtained. In variational data assimilation, explicit penalty terms are included in the cost function for parameters. Furthermore, in four-dimensional data assimilation, the adjoint model equation is augmented by an explicit term that involves the parameters. In ensemble-based data assimilation, appropriate parameter perturbations are introduced to obtain

one-way covariance information between parameters and the observed model state. Additionally, measures are taken to maintain sufficient parameter spread throughout data assimilation cycles.

In **Figure 1**, the general methodology of parameter estimation in a simplified schematic is illustrated. **Figure 1(a)** shows how a first guess that is sensitive to parameter perturbations is generated. First, the numerical model is advanced using perturbed parameters. In ensemble-based techniques this is achieved by advancing each ensemble member using a perturbed parameter value. Here, it is also assumed that there exists initial condition uncertainty, as it is standard in all data assimilation applications (hence the distribution in the ‘initial state’ balloon). In variational techniques, a ‘first guess’ parameter value is used to obtain the first guess for the model state. At this stage, observation operators are also applied to the first guess model state to obtain the first guess in observation space. In the second stage (**Figure 1(b)**), regression is carried out between the first guess and the parameter to be updated, to obtain a linear statistical relationship between the (observed) model state and parameters. In ensemble-based techniques, this involves the computation of sample covariances between observations and parameters. In variational techniques, adjoint models are used. Finally, the update is performed (**Figure 1(c)**) to obtain a new parameter value informed by the latest available observations. In ensemble-based techniques, this is done by projecting observation-first-guess differences onto the parameter space through the regression relationship. In variational techniques, a cost function minimization procedure that accounts for parameter variability is applied. It is also common to repeat this workflow in subsequent analysis cycles where model advances take into account the updated parameter values.

Parameter estimation is generally an extension of data assimilation in terms of its methodology. At the same time, it is unique in that there is only a one-way dynamical interaction between parameters and the model state. In other words, there is an increased possibility of low signal-to-noise ratio between parameters and observations in the presence of potentially strong dynamical interactions within model state variables themselves. In the next section, issues related to the difficulty of extracting useful information from observations to successfully update parameters are discussed in more detail.

Parameter Identifiability

Parameter estimation is, in essence, the problem of inverse mapping from model space to the space of parameters. In the atmospheric science literature, the term parameter identifiability has been used to denote how easy it is to find unique solutions of the inverse problem for unknown parameters from available observations of the model state. An argument is made that nonuniqueness and instability of the identified processes may contribute to the ill posedness of the problem. A more categorical explanation for parameter identifiability argues that three factors can be thought to contribute to it (see **Figure 2**, for a schematic illustration):

- Observability:** Changes produced by parameter variations do not necessarily project sufficiently strongly onto model/observation space. Strictly speaking, this may be due to both insufficient model sensitivity to and/or incompatible model-observation mapping with the parameters. The end result of realistic measurement errors leading to unrealistically large changes in estimated parameter values is the same. Lack of observability is illustrated in **Figure 2(a)** by the relatively flat slope of the regression line between the observed model state and parameters. In such a scenario, small observation differences are projected onto large parameter differences and may result in unphysical or unrealistic parameter values in the update.
- Simplicity:** Sometimes, even with sufficient model sensitivity, the mapping between the parameter space and model space can be such that a simple one-to-one relationship does not exist between the model output and the parameter to be estimated, which is generally caused by the nonlinearity in how the model responds to the expected parameter variations. In such situations, a unique solution may not exist to the inverse problem and the optimal parameter value may never be found. Lack of simplicity is illustrated in **Figure 2(b)** by the quadratic relationship between the observed model state and parameters. If the parameter space is not carefully constrained (i.e., the regression is limited to only one side of the quadratic relationship), the probability that the same observation differences are projected onto multiple distinct parameter

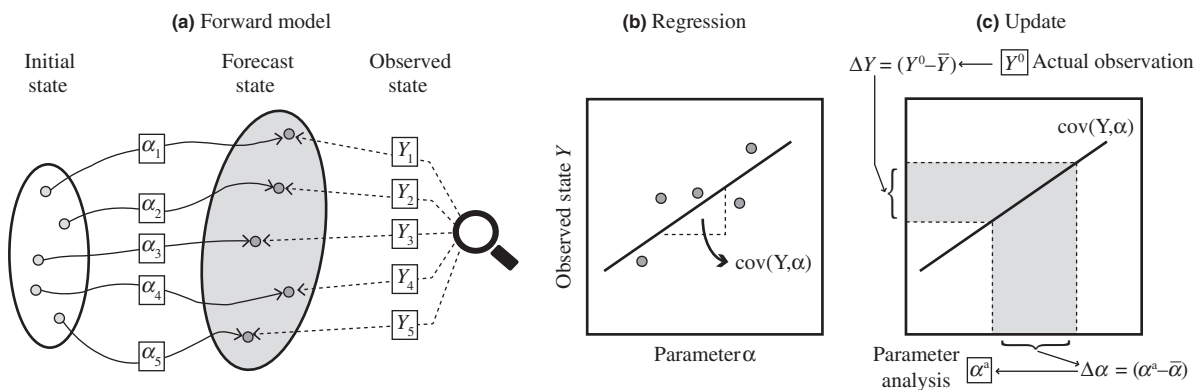


Figure 1 Illustration of the typical workflow of parameter estimation (using the ensemble method as an example). (a) Forward model. (b) Regression. (c) Update. Symbols α and Y represent the parameter to be estimated and the observation variable, respectively.

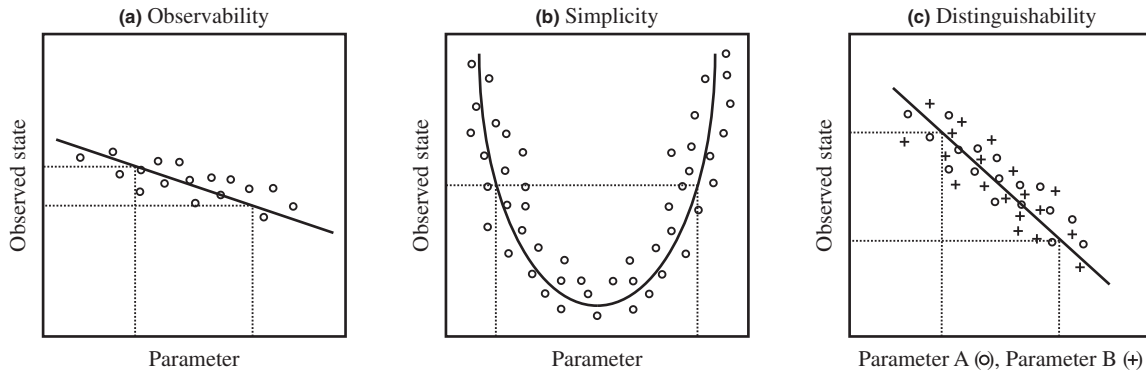


Figure 2 Illustration of types of parameter identifiability and how they can negatively impact successful parameter estimation. (a) Observability. (b) Simplicity. (c) Distinguishability. The thick curves in the panels represent the statistical regression relationship between parameters and observed state. Thin dashed lines represent selected parameter–observation pairs on the regression curves.

differences is increased, which can result in unrealistic parameter updates.

- c. *Distinguishability*: In other situations, various parameters may have similar aggregate effects on the model state. Therefore, observations related to the involved model output may lead to adjustments in the wrong set of parameters. Lack of distinguishability is illustrated in [Figure 2\(c\)](#), where very similar statistical relationships exist between the observed model state and two different types of parameters. In such a situation, given model response may be interpreted as being the result of perturbations in a wrong parameter, causing certain parameters to be updated toward unrealistic values, whereas those that should be updated remain mostly unchanged.

Successful parameter estimation is intimately linked to whether and how each of these factors is addressed. For example, observability implies that careful analysis must be carried out to ensure that the model exhibits sufficient sensitivity to the parameters to be estimated. At this stage, it is also imperative to choose variables that best reflect the sensitivity between model space and parameter space. Next, the proposed parameter space must be carefully examined to avoid non-simple observation–parameter relationships. Finally, when multiple parameters are estimated, cross correlations among the parameters must be carefully identified so that parameters that induce similar effects on the observed fields are eliminated. Only when all of these conditions are met, sufficient parameter identifiability can be achieved and successful parameter estimation can be performed. Some suggestions to improve identifiability are (1) change the observing system (i.e., address observability); (2) modify the model to eliminate the source of nonuniqueness (i.e., address simplicity); or (3) altogether modify the construction of the inverse method, which can be achieved by introducing new information to distinguish and avoid the behavior in the modeled processes that result in the nonuniqueness in the first place.

When parameter variations are considered in the context of uncertain initial and boundary conditions, sensitivity of the model state to parameters can decrease substantially, leading to heightened identifiability issues. This can be investigated

through the magnitude and linearity of the signal in the ensemble spread due to parameter perturbations. If lack of identifiability is detected, scaling parameter perturbations to increase the resulting ensemble spread could be helpful, but the positive effects would be unavoidably constrained if the linearity in ensemble spread is limited. Multiple, simultaneously uncertain parameters are also argued to be a source of negative impact on identifiability.

Earlier studies have investigated parameter identifiability mostly in the context of model sensitivity. Some studies approached identifiability both with the use of a response function and correlations between model output and uncertain parameters. A response function can be constructed to represent the mean-square distance between the model solution and observations averaged over all observation points. Normalizing then by observation variance provides a metric that allows for the comparison of model response as a result of various parameter perturbations. Computing the response function in observation space also addresses the observability aspect of identifiability. Another metric to investigate sensitivity is root-mean-squared correlation, which represents the magnitude of spatially averaged absolute sample correlation between the model state and parameters. The root-mean-squared correlation metric can also be extended to observation space to account for the observability aspect.

Other Challenges for Parameter Estimation

Besides the difficulties arising from identifiability issues, other challenges also exist for parameter estimation. One such important problem arises from the fact that parameters in a numerical model are not ‘dynamical’ by definition, by which it is meant that there is no feedback from the evolving model state back to parameters. In ensemble-based data assimilation, this creates the immediate problem of filter divergence: When the ensemble spread becomes exceedingly small, the first guess is given increasing weight by the data assimilation scheme, which results in new observations to have less weight on the analysis. Without dynamical feedback, parameter spread is destined to become small enough to lead to filter divergence unless explicit measures are taken to maintain the desired

level of ensemble spread in the parameters to be estimated. A common such measure is to ‘inflate’ analysis variance when it drops below a threshold level.

The other aspect of the nondynamical nature of model parameters is that they are generally specified ‘globally,’ i.e., they are assigned the same values over the entire computational domain of the model. In real-data applications with large, three-dimensional observational datasets, this results in the overdetermination of parameters and manifests itself in the form of excessive noise in updated parameters (updated parameters that appear to follow random walk). To counter this issue, one method, ‘spatial updating’ updates the parameters first horizontally as two-dimensional arrays using covariance localization, and then uses spatial averaging to obtain the updated global parameter values. Another approach is a data selection technique, which uses only a certain subset of observations that exhibit the largest ensemble correlations with the parameters.

When simultaneous state and parameter estimation is carried out, it has been shown by some that the state estimation errors dominate the uncertainties of the model state during the early stages of data assimilation so that the covariance information between the model state and parameters becomes unreliable. To remedy the situation, some studies used a delayed approach in which parameter estimation after a number of data assimilation cycles when the model state is likely to be constrained by observations, and covariances between the model state and parameters are likely to be reliable. Other studies chose not to update the state at all, instead focusing on updating parameters, and approached the filter divergence and overdetermination issues by assimilating observations individually at their exact time. The final global parameter value is then achieved by averaging over time and ensemble members.

Other challenges to successful parameter estimation pertain to the lack of knowledge on how best to sample parameter space. In most situations, suitable values of parameters are barely known and there is very little knowledge, if any, on the bounds and the nature of uncertainty that surrounds them. Although, as demonstrated by some studies, parameter uncertainty can take many shapes and forms in the form of non-Gaussian, multimodal distributions, in more realistic situations, such assumptions are difficult to make. This has generally led to the assumption of prior parameter distributions with limited information content, which generally takes the form of bounded uniform distributions or Gaussian distributions. However, since parameter values are generally expected to remain bounded and positive definite, Gaussian prior probability density functions (PDFs) are not always suitable choices. Approaches in the literature have varied to address these situations; methods such as transforming parameters logarithmically to mimic log-normal PDFs, assumption of beta distributions that are naturally bounded, and trigonometric transformations for bounded parameters are used.

As the number of parameters simultaneously estimated increases, a new issue also arises to effectively sample the multidimensional parameter space in an ensemble of limited size. Though most studies sample parameters independently from respective assumed prior PDFs, this does not necessarily guarantee that the joint parameter space is effectively sampled. One method to increase the effectiveness of sampling, the

Markov chain Monte Carlo algorithm, revisits high-probability regions of the parameter space in an iterative manner. Although this technique reduces the computational burden of effective sampling by a few orders of magnitude, for sufficiently many parameters, the technique may still be computationally unfeasible with complex numerical models. As an alternative, a Latin hypercube sampling strategy is suggested, which works in a normalized bounded space but provides independent parameter distributions with even sampling. These samples can then be transformed into any other PDF if nonuniform distributions are to be assumed.

Another challenge for parameter estimation arises from the nature of the complex and highly nonlinear relationship between individual parameters and the aggregate model response to them. In many situations, the same parameters simultaneously impact various processes, and the interactions among the processes make isolating individual impacts difficult. One proposed approach assigns multiplicative weights to the processes that contribute to the total outcome of a parameterization scheme, and introduces uncertainties to these weights. The advantage of this approach is that it is a natural way of looking at model error, in the sense that individual processes themselves are better understood physically and therefore lend themselves suitably to a subjective assessment of the misrepresentation of the parameterized aspect of the model. Furthermore, the additive nature of the processes themselves may prove advantageous in obtaining a linear model response to the perturbations of the multiplicative weights.

Future Advances

Research in parameter estimation is likely to advance on several avenues. First of all, studies that systematically investigate all dimensions of parameter identifiability are needed to obtain a complete picture of the challenges and limitations facing parameter estimation. Comparisons between parameter-related model sensitivity and initial condition- or boundary condition-related model sensitivity should be carefully made to take into account the timescales at which these various sources of sensitivity act upon the atmospheric state. For fair comparison, it is also advisable to note that model sensitivity to initial or boundary conditions is the result of perturbations that are usually obtained from continuously cycling data assimilation systems that do not account for parameter uncertainty, whereas, for practical reasons, parameter perturbations can only be introduced to the most recent initial and boundary condition ensembles.

There is also a need for a unified vision for ‘correlation localization’ that places the updating of state variables and parameters on a common ground. In state-only data assimilation, it is natural to visualize the influence of observations onto the model state in geographical space. For global model parameters, such a natural localization space does not exist. Although some studies have proposed ad hoc techniques to manage this dichotomy, a systematic approach that puts on equal ground the localizations for model state and model parameters has yet to emerge.

Finally, a fundamental shift from focusing on individual parameters, whose effects on the model state are usually

obscure because of the many dependent nonlinear processes, toward focusing on those actual processes that contribute to the particular subgrid-scale parameterizations may be prudent. It is possible to control the contributions from these individual processes by assigning multiplicative weights to their final output that are to be estimated. This philosophy also enables parameter estimation to become a more holistic approach to counter model error. After all, it may be more difficult to justify parameter estimation as a legitimate means of treating model error when estimated parameters are usually empirical and not well observed. When the focus shifts to the processes themselves, the uncertainty can be expressed in the natural space of known physical processes and the estimation may then effectively inform on the relative importance of individual processes under observed atmospheric conditions.

Summary

The state of the art of parameter estimation in atmospheric sciences is reviewed. Parameter identifiability, defined as the ease of finding unique solutions of the inverse problem for unknown parameters from available observations of the model state, is composed of the three dimensions of observability, simplicity, and distinguishability. The major challenges in parameter estimation are discussed as the nondynamical nature of model parameters, difficulty of updating globally specified parameters using spatially and temporally varying observations, distinguishability in the presence of initial and boundary condition uncertainty, lack of knowledge of the probabilistic nature of parameter uncertainty, and effective sampling of multidimensional parameter space. Some research directions for the near future are suggested.

Acknowledgments

The author is grateful for the valuable inputs obtained from Drs Robert Rogers and Sim Aberson of NOAA Hurricane Research Division.

See also: **Boundary Layer (Atmospheric) and Air Pollution:** Modeling and Parameterization. **Data Assimilation and Predictability:** Data Assimilation; Ensemble Prediction; Ensemble-Based Data Assimilation. **Numerical Models:** Methods; Model Physics Parameterization; Parameterization of Physical Processes: Clouds; Parameterization of Physical Processes: Turbulence and Mixing; Regional Prediction Models.

Further Reading

See Navon (1998) for a comprehensive review of parameter identifiability and early literature on variational parameter estimation. Zhu and Navon (1999) provide some further technical details on the variational approach to parameter estimation. Good examples for recent studies on ensemble-based parameter estimation are Aksoy et al. (2006a,b), Tong and Xue (2008a,b), Hu et al. (2010), Jung et al. (2010), Godinez et al. (2012), Posselt and Bishop (2012), and van Lier-Walqui et al. (2013).

Nielsen-Gammon et al. (2010) discuss in detail the three dimensions of parameter identifiability. All three factors are evaluated there for various parameters of a PBL parameterization scheme: observability is deduced from the magnitude of ensemble standard deviations of model fields when individual parameters are perturbed, simplicity is deduced from the nature of scatter plots between model fields and individual parameters, and distinguishability is deduced from the overall level of correlations between model fields and individual parameters.

Posselt and Vukicevic (2010), using a one-dimensional model of convection and cloud microphysics, employed the Markov chain Monte Carlo approach to effectively sample the high-probability regions of parameter space spanned by multiple parameters. That way, they were able to obtain joint parameter-state PDFs, which, in some cases, exposed relationships that were highly nonlinear and multimodal, resulting in parameter identifiability issues.

Hacker et al. (2011) investigated parameter identifiability in the presence of uncertain initial and boundary conditions. The limitations to identifiability from simultaneous uncertain parameters are studied by Nielsen-Gammon et al. (2010), Posselt and Vukicevic (2010), Tong and Xue (2008b), and Aksoy et al. (2006a). To investigate sensitivity, Tong and Xue (2008a) suggested a response function and correlations between model output and uncertain parameters, whereas Aksoy et al. (2006a) used the metric root-mean-squared correlation.

Aksoy et al. (2006a) introduced the method of inflating parameter variance when it drops below a threshold level (also see Tong and Xue, 2008b; Jung et al., 2010; Zhang et al., 2012). The two counter measures for the overdetermination problem when estimating global parameters, the spatial updating method and the correlation-dependent data selection technique, were introduced by Aksoy et al. (2006a) and Tong and Xue (2008b), respectively. Jung et al. (2010) and Zhang et al. (2012) used the delayed parameter estimation approach to improve the sensitivity of the model state to multiple, simultaneously uncertain parameters. Godinez et al. (2012) chose to only update the parameters and used the technique of updating parameters at their exact time. To obtain prior distributions for bounded parameters, Tong and Xue (2008b) introduced logarithmic transformations, Hacker et al. (2011) applied beta distributions, and Nielsen-Gammon et al. (2010) used trigonometric transformations. The two methods to effectively sample the multiple parameter space, the Markov chain Monte Carlo algorithm, and the Latin hypercube sampling strategy, are suggested by Posselt and Vukicevic (2010) and Hacker et al. (2011), respectively.

van Lier-Walqui et al. (2013) proposed the process approach, where the weights of process outcomes are estimated rather than empirical parameters themselves.

References

- Aksoy, A., Zhang, F., Nielsen-Gammon, J.W., 2006a. Ensemble-based simultaneous state and parameter estimation with MM5. *Geophysical Research Letters* 33, L12801. <http://dx.doi.org/10.1029/2006GL026186>.
- Aksoy, A., Zhang, F., Nielsen-Gammon, J.W., 2006b. Ensemble-based simultaneous state and parameter estimation in a two-dimensional sea-breeze model. *Monthly Weather Review* 134, 2951–2970.
- Godinez, H.C., Reisner, J.M., Fierro, A.O., Guimond, S.R., Kao, J., 2012. Determining key model parameters of rapidly intensifying Hurricane Guillermo (1997) using the ensemble Kalman filter. *Journal of Atmospheric Sciences* 69, 3147–3171.
- Hacker, J.P., Snyder, C., Ha, S.-Y., Pocerich, M., 2011. Linear and non-linear response to parameter variations in a mesoscale model. *Tellus* 63A, 429–444.
- Hu, X.-M., Zhang, F., Nielsen-Gammon, J.W., 2010. Ensemble-based simultaneous state and parameter estimation for treatment of mesoscale model error: a real-data study. *Geophysical Research Letters* 37, L08802. <http://dx.doi.org/10.1029/2010GL043017>.
- Jung, Y., Xue, M., Zhang, G., 2010. Simultaneous estimation of microphysical parameters and the atmospheric state using simulated polarimetric radar data and an ensemble Kalman filter in the presence of an observation operator error. *Monthly Weather Review* 136, 1649–1668.
- Nielsen-Gammon, J.W., Hu, X., Zhang, F., Pleim, J.E., 2010. Evaluation of planetary boundary layer scheme sensitivities for the purpose of parameter estimation. *Monthly Weather Review* 138, 3400–3417.
- Posselt, D.J., Bishop, C.H., 2012. Nonlinear parameter estimation: comparison of an ensemble Kalman smoother with a Markov chain Monte Carlo algorithm. *Monthly Weather Review* 140, 1957–1974.
- Posselt, D.J., Vukicevic, T., 2010. Robust characterization of model physics uncertainty for simulations of deep moist convection. *Monthly Weather Review* 138, 1513–1535.

- Tong, M., Xue, M., 2008a. Simultaneous estimation of microphysical parameters and atmospheric state with simulated radar data and ensemble square root Kalman filter. Part II: sensitivity analysis and parameter identifiability. *Monthly Weather Review* 136, 1630–1648.
- Tong, M., Xue, M., 2008b. Simultaneous estimation of microphysical parameters and atmospheric state with simulated radar data and ensemble square root Kalman filter. Part II: parameter estimation experiments. *Monthly Weather Review* 136, 1649–1668.
- van Lier-Walqui, M., Vukicevic, T., Posselt, D.J., 2014. Linearization of microphysical parameterization uncertainty using multiplicative process perturbation parameters. *Monthly Weather Review* 142 (1), 401–413.
- Zhang, S., Liu, Z., Rosati, A., Delworth, T., 2012. A study of enhanceive parameter correction with coupled data assimilation for climate estimation and prediction using a simple coupled model. *Tellus* 64A, 10963. <http://dx.doi.org/10.3402/tellusa.v64i0.10963>.

Parameterization of Physical Processes: Clouds

R Forbes, European Centre for Medium-Range Weather Forecasts, Reading, UK

C Jakob, Monash University, VIC, Australia

M Miller, European Centre for Medium-Range Weather Forecasts, Reading, UK

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by C Jakob, M Miller, volume 4, pp 1692–1698, © 2003, Elsevier Ltd.

Synopsis

Clouds are important for weather forecasting and climate prediction not only for their direct role in the hydrological cycle, but also for their interactions with radiation and the dynamics of the atmosphere. The formation and dissipation of clouds depends on the interaction of many small-scale processes, which must be parameterized in atmospheric models. General concepts of cloud parameterization and the hierarchy of approaches commonly used over the years are discussed, as well as contemporary issues for cloud scheme development.

Introduction

At any given time, clouds cover between 60 and 70% of the globe and for most of mankind they are an everyday experience. Clouds exert various influences on the Earth–atmosphere system, of which the most important are:

- modification of the radiative fluxes in the atmosphere and at the Earth's surface;
- release and consumption of latent heat related to phase changes of water either directly inside the clouds or in precipitation generated in them;
- transport of heat, moisture, momentum, and atmospheric trace constituents over large distances in the vertical in convectively generated clouds; and
- modification of the surface hydrology through precipitation generated in clouds.

For a more detailed discussion of these cloud effects the reader is referred to other articles in the encyclopedia (*see Clouds and Fog: Classification of Clouds; Climatology; Measurement Techniques In Situ. Mesoscale Meteorology: Convective Storms: Overview. Radiation Transfer in the Atmosphere: Cloud-Radiative Processes*) and in the Further Reading section at the end of this article.

Given the importance of the various influences clouds have in the evolution of both the atmosphere and the surface, it is immediately obvious that these effects need to be included in atmospheric models that are used for the simulation of climate and the prediction of weather. As described in the articles dealing with general circulation models (GCMs) and numerical weather prediction (*see Numerical Models: General Circulation Models; Regional Prediction Models*), these models seek numerical solutions to the hydrodynamic equations that govern atmospheric motions.

Various numerical techniques can be applied to achieve this goal, but all of them ultimately involve splitting the area over which the model is applied into 'boxes' of finite size in both the horizontal and the vertical. While the continuous differential equations describe atmospheric motions on all spatial and temporal scales, their discrete form can only describe processes on spatial scales of the order of twice the grid length or larger. Processes that occur in clouds cover a wide range of spatial

scales, from micrometers for the condensation and evaporation of individual cloud droplets, through a few hundred meters in the case of fair weather cumulus clouds, up to several hundred kilometers for the cloud systems associated with extratropical baroclinic systems. Hence, describing the detailed dynamics of individual clouds would require model grid box sizes on the order of meters or less. Current computing power as well as difficulties in finding the necessary initial conditions at these spatial scales prohibit the use of such grid box sizes in global atmospheric models. In reality, typical horizontal grid lengths in contemporary global models range from the order of 10 km in numerical weather prediction applications to more than 250 km in climate modeling. Processes that act on scales smaller than these grid sizes are normally referred to as subgrid-scale processes and are, per se, not represented in the solution of the finite difference equations. Many of these processes do, however, affect the dynamic and thermodynamic states of the atmosphere on larger spatial scales. Obvious examples are the large amounts of water vapor, heat, and momentum that are transported by turbulent and convective motions. Since an explicit description of the subgrid-scale processes is prohibited, only their statistical effects on the grid box-mean state can be taken into account. Since the numerical solution of the model equations allows the atmospheric state to be known only on scales on the order of the grid box size, the description of these statistical effects has to be expressed in terms of those scales. The technique involved is generally referred to as parameterization.

To describe the main effects clouds have on the atmosphere as outlined above, the following cloud-related quantities need to be known:

- horizontal coverage of cloud, normally referred to as cloud fraction;
- vertical extent of the clouds;
- sources and sinks of cloud condensate including condensation, evaporation/sublimation, and conversion into precipitation and fallout;
- phase of the condensate;
- particle size and shape;
- in-cloud distribution of condensate; and
- amounts of heat, water vapor, and momentum that are transported in convective clouds.

This list implies scales much smaller than the typical resolution of most atmospheric models. The problem of representing clouds in large-scale atmospheric models is therefore one in parameterizing their overall effects on the resolved scales.

There are a number of problems to be overcome in the parameterization of clouds. First, there exists a variety of cloud types, such as stratocumulus clouds at the top of convective boundary layers, vast cloud systems associated with extratropical disturbances, deep convective systems that may or may not be organized, and upper-tropospheric cirrus clouds. These different cloud types are formed, maintained, and dissipated by different physical processes, such as convection, radiative cooling, small-scale turbulence, large-scale ascent or descent, and cloud microphysical processes that lead to the generation of precipitation. Many of these processes are poorly understood and act on scales smaller than those resolved in a large-scale model, which makes them the subject of physical parameterization themselves. Furthermore, the radiative effects of clouds depend on a large number of different cloud parameters that all need to be described accurately to ensure their correct treatment in the radiation parameterization.

It is worthwhile pointing out that, because of their distinctive properties in terms of significant small-scale heat, water vapor, and momentum transport, cumulus and cumulonimbus clouds have been recognized as being of particular importance. This has led to (an artificial) separation of the description of the vertical transport and condensation effects from the radiative effects of convective clouds in what is known as cumulus convection parameterizations. As will be briefly discussed below, recent efforts in improving cloud parameterizations involve attempts to overcome this artificial split between convective and more 'stratiform' cloud processes. Furthermore, the details of the radiative transfer in clouds are normally dealt with in radiation parameterizations. Thus, typical cloud parameterizations need to:

1. describe the generation and dissipation of clouds and the precipitation formed in them; and
2. provide the radiation parameterization with the necessary information to evaluate the cloud effects on the radiative fluxes, most prominently the area coverage and cloud condensate content.

Before a brief overview of how the problem of cloud parameterization can be addressed, some general concepts for any type of cloud parameterization will be outlined.

General Concepts in Cloud Parameterization

The sizes of many of the observed clouds are often significantly smaller than the typical sizes of model grid boxes in GCMs. Even on integration over all individual clouds in an area comparable to those grid sizes, one finds from observations that often the area is only partially covered with cloud. Since this has important consequences, especially for the radiative cloud effects, almost all cloud parameterizations describe the fractional coverage of a model grid box with cloud as one of their key parameters. Since cloud fraction is such a fundamental concept that is used in many different ways across

a whole variety of cloud parameterizations, it seems worthwhile to highlight the general implications of the concept of fractional cloud cover.

Assuming that clouds form whenever the specific humidity locally exceeds its saturation value, which occurs if sufficient cloud condensation nuclei (CCN) are available (see below), fractional cloud cover implies that certain parts of a model grid box become supersaturated before others. This has several implications. One of them is that clouds exist in the model grid box before the grid-mean relative humidity reaches the saturation value of 100%. This has been used in many cloud parameterizations to determine the cloud fraction by defining a critical relative humidity, RH_{crit} , above which clouds exist in a grid box and a functional relationship that increases cloud cover from zero below RH_{crit} to one when the entire grid box is saturated. It should be obvious that the definitions of both RH_{crit} and the functional relationship are far from unique and for many years cloud parameterization was nothing more than attempting to find and refine such definitions.

Another consequence of considering cloud fraction is that there must exist a distribution of humidity and temperature around their mean value in a grid box. The knowledge of these variations would in fact be sufficient to describe the cloud field within a grid box. Figure 1 provides an illustration of this idea using the concept of total water q_t , which is the sum of the humidity q and condensate amount q_l . In a one-dimensional model grid box, both the total water q_t and the saturation humidity q_s (dependent only on temperature and pressure) are assumed to be nonuniform. In those areas where the humidity exceeds the saturation value ($q_t > q_s$), clouds are assumed to exist and the sum of the cloud areas (c) divided by the size of the grid box (x) is the total cloud fraction, a , where $a = c/x$. The mathematical technique used to describe these variations describes joint probability distribution functions (PDFs) for a temperature variable and a humidity variable. Unfortunately, the distribution functions are neither known well enough nor expected to be unique and will depend on many different physical processes. Nevertheless, the introduction of the idea of

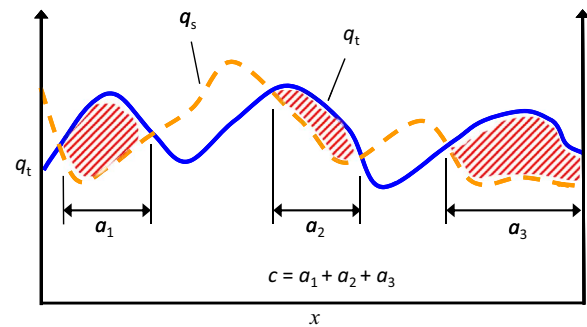


Figure 1 Schematic of the existence of clouds in the supersaturated areas of a one-dimensional model grid box. The x -axis represents space. The solid line (q_t) shows the value of total water as a function of location within the grid box. The dashed line (q_s) shows the saturation value of specific humidity. Areas in which $q_t > q_s$ represent clouds, indicated by the hatched areas. The total cloud fraction, c , is equal to sum of the cloudy areas a_1 , a_2 , and a_3 .

distributions provides a conceptual framework for the development of cloud parameterizations.

As well as the representation of subgrid heterogeneity, a cloud parameterization can also represent the microphysical processes that describe phase changes and precipitation formation. The degree of complexity depends very much on the computational power available, the application, and our knowledge of the microphysical processes themselves. One of the microphysical processes to be described in any cloud parameterization is the condensation process. This theoretically involves the description of two distinct processes: the nucleation of cloud particles and their initial growth by diffusion of water vapor toward the nucleated particles. It is well known that the main warm-phase nucleation process in the atmosphere is that of heterogeneous nucleation of cloud water droplets on small aerosol particles, usually referred to as CCN (see **Aerosols: Aerosol–Cloud Interactions and Their Radiative Forcing; Clouds and Fog: Cloud Microphysics**). In the presence of abundant CCN in the atmosphere, condensation occurs whenever the relative humidity with respect to water exceeds its saturation value by often less than 1%, while in the absence of CCN large values of supersaturation (several hundred percent) are needed to allow homogeneous nucleation of sufficiently large droplets. In order to avoid the complex treatment of nucleation processes, most cloud parameterizations to date assume that CCN are always available in sufficient numbers and the condensation problem reduces to converting any supersaturation directly to cloud water. For ice particles, nucleation in the atmosphere can occur via heterogeneous or homogenous nucleation. Supersaturations with respect to ice are frequently observed in the upper troposphere, complicating the parameterization of ice clouds. Many other microphysical processes may also be represented in cloud parameterizations with varying degrees of complexity, such as the conversion of cloud to precipitation (autoconversion, accretion, aggregation), evaporation and melting, but these are discussed elsewhere (see **Clouds and Fog: Cloud Microphysics**).

Common Approaches to Cloud Parameterization

The previous section has established the reasons why cloud processes need to be parameterized in atmospheric models.

The main effects of clouds were found to be their influence on the radiative fluxes, their latent heat effects, and the ability to transport heat, moisture, and momentum in case of convective clouds. It was also established that all but the radiative effects of convective clouds are treated in a separate convection parameterization, which is not the subject of this article. The role of clouds for atmospheric models was recognized early on, although in the first models it was mainly the latent heat effects that were considered to be important. This section gives a brief overview of the major steps in the history of cloud parameterization. Various approaches are considered in the context of the major effects that need to be described in models. Each of the periods of development in cloud parameterization can be assessed using the following four questions.

1. How are nonconvective condensation processes on subgrid scales described?
2. How are the radiation effects of the clouds derived after answering (1)?
3. How are the convection and cloud parameterizations linked?
4. How are the microphysical processes that lead to precipitation generation described?

Table 1 provides an overview over the timeline of key aspects of the treatment of cloud-related processes in atmospheric models.

Early Condensation Schemes

In the development of early GCMs in the 1960s, the latent heat effects of both convective and nonconvective condensation processes needed to be considered. Furthermore, since models included an evolution equation for a humidity variable, unphysical states of supersaturation needed to be avoided in the evolution of the model variables. Therefore, a simple but effective condensation scheme was introduced into the models. Its basic idea was to readjust back to saturation any possible supersaturated states occurring on the grid scale at the end of a model time step. The condensate thus formed was removed instantaneously as precipitation. Hence, although condensation processes and therefore their latent heat effects were described, it was not clouds but precipitation that was formed during the condensation. A similarly simple description of

Table 1 An overview of the historic evolution of key aspects of cloud parameterization. The symbols are defined as follows: $\bar{q}$ is the grid-mean specific humidity; q_s is the grid mean of its saturation value; a represents cloud fraction with a_{cu} describing the contribution from convectively generated clouds to that value; l represents the condensate content, with l_{cu} again describing that in convective clouds; RH is the grid-mean relative humidity and CP is the rate of convective precipitation

	Modeling period			
	1960/1970s	1970/1980s	1980/1990s	Now and beyond
Condensation (nonconvective)	$\bar{q} > q_s$	$\bar{q} > q_s$	l prognostic function of outcome of processes	l prognostic function of the processes themselves
Radiation effects	Prescribed zonal mean albedo and emissivity of clouds	$a = f(\text{RH})$ l prescribed	$a = f(\text{RH})$ l prescribed	a prognostic l prognostic
Convection	No cloud interaction	$a_{cu} = f(\text{CP})$ l_{cu} prescribed	$a_{cu} = f(\text{CP})$ l_{cu} prescribed	Condensate and mass as sources for a and l
Microphysics	None	None	Simple bulk microphysics	Complex bulk microphysics

convection was used in which the temperature lapse rate for saturated grid columns was not allowed to exceed that of a moist adiabat. Any condensate formed in this 'moist convective adjustment' process was also removed as precipitation. The role that radiation effects of clouds play in the general circulation was considered small, so that most early GCMs used prescribed zonally averaged cloud albedos and emissivities as input for their radiation calculations. Since all condensate was removed as precipitation, no description of microphysical processes was necessary; hence, early GCMs described only condensation processes with no cloud interaction whatsoever. In fact one could argue from today's point of view that early GCMs did not parameterize clouds but precipitation. The first column in [Table 1](#) represents this period in the evolution of cloud parameterization.

Diagnostic Cloud Schemes

It was soon recognized that the radiative effects of clouds might play a crucial role in the general circulation of the atmosphere. The next generation of cloud parameterizations was therefore aimed at providing some interaction of cloudiness and the other model variables. This was usually achieved by parameterizing the cloud fraction as a function of relative humidity. This type of parameterization had already been proposed for early models but it was not used in GCMs until the 1980s. The reasons for this are not entirely obvious, but the difficulties of validating the model predictions of cloud fraction and the rather limited computing power available at the time were factors.

Relative humidity schemes rely on the concept that if the grid-mean relative humidity exceeds a threshold value RH_{crit} , usually on the order of 80%, it is likely that some part of the grid volume has already reached saturation and therefore clouds start to form. If the grid-mean relative humidity reaches 100%, the entire grid box is assumed to be covered with clouds. Since all models using this approach still used the description of condensation as before, the radiative and latent heat effects of clouds were entirely decoupled. Furthermore, since condensation occurred only for grid-mean values of relative humidity above 100% but clouds existed before that, the amount of condensate needed for the description of the radiative effects of the model clouds was simply prescribed.

The development of more complex convection parameterizations allowed convectively generated clouds to be described as a function of the results of the convection parameterization. This was often achieved by linking the cloud fraction to the precipitation produced in the convection scheme and again prescribing the condensate content. The simple removal as precipitation of any moisture in excess of the saturation humidity makes the description of microphysical processes unnecessary.

This type of cloud parameterization is usually referred to as the 'diagnostic' approach, since the main cloud parameters (cloud fraction and condensate amount) are diagnosed using the grid-averaged quantities, and is represented by the second column in [Table 1](#). Over the years, the basic relative humidity approach was developed, by introducing additional predictors such as vertical motion and inversion strength at the top of convective boundary layers, into the cloud fraction description.

It is noteworthy that this approach provides reasonable estimates of many of the main observed cloud patterns and can be made to work well by adjusting the many free parameters in the parameterization. This, together with a low computational cost, made it a widely used parameterization approach right up to the mid-1990s.

Prognostic Condensate

One of the major drawbacks of the diagnostic approach described above is the obvious disconnection of the cloud latent heat effects from the radiative effects. Sundqvist introduced an additional prognostic model equation for cloud condensate, previously only applied in cloud-scale modeling, establishing this link for a GCM parameterization. By explicitly predicting the amount of condensate formed, a link to the radiative impact of the clouds could be created through the direct use of the predicted condensate in the radiation calculations. A consistent diagnostic treatment of cloud fraction was also introduced in which the cloud fraction remains a function of the grid-mean relative humidity, which is now directly influenced by the condensation processes that are allowed to occur before grid-mean saturation is reached.

The description of convective clouds remained unaltered by Sundqvist's approach. One immediate consequence that should play a major role in the further development of cloud parameterizations is that the conversion of some of the cloud condensate to precipitation needs to be described. Very simple descriptions of the autoconversion process together with some intuitive parameterization of the precipitation-enhancing collection and Bergeron-Findeisen mechanism were used. Although simple, the use of a parameterization scheme of this kind for the first time acknowledged the need to describe microphysical processes as part of the cloud parameterization problem.

Statistical Schemes

In parallel to the introduction of what is now usually known as 'the Sundqvist parameterization,' another approach emerged, based on ideas originally applied in much higher-resolution cloud models. Here, the parameterization of clouds is based on the idea outlined above that the existence of clouds on a subgrid scale requires that the humidity and its saturation value be somehow distributed around their grid-mean values. The knowledge of their PDFs is therefore sufficient to describe both cloud fraction and condensate content within a grid box. The most common use of this idea is by means of a joint PDF for a temperature variable and a humidity variable. Since it was originally developed for the description of nonprecipitating boundary layer clouds, conservative thermodynamic variables such as liquid water potential temperature and total mixing ratio are often preferred.

[Figure 2](#) illustrates the general idea of this approach. Liquid water potential temperature and total mixing ratio are assumed to be distributed with a joint PDF. A saturation curve for a given grid-mean temperature and pressure is then drawn. All the values of the PDF that lie above this saturation curve represent clouds and the cloud fraction and condensate content can be calculated by integrating over this part of the distribution.

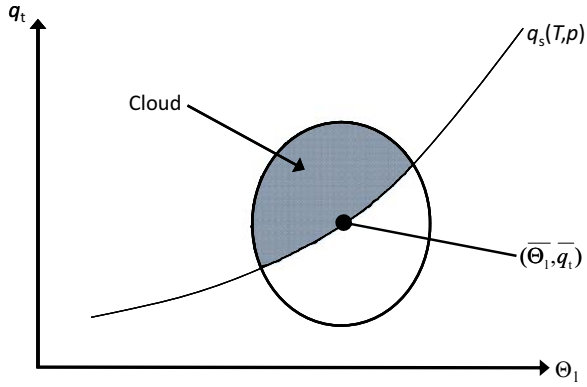


Figure 2 Schematic diagram of a possible distribution of liquid water potential temperature, Θ_l ; and total water, q_l ; in a model grid box and its implication for clouds.

The crucial question for a successful application in GCMs is the definition of the distribution function itself. Different approaches were taken here using either fully prescribed and fixed PDFs or simple links of some of the distribution parameters (e.g., variance) to the turbulence parameterization. A critical issue for the use of the PDF of variables is that their conservation breaks down in the presence of precipitation, which needs to be addressed for implementation in a GCM.

Since it is obvious that cloud fraction and cloud condensate content within a grid box do depend on PDFs as used in the parameterization, this approach for parameterization appears promising if the evolution of the PDF can be predicted from the evolution of the resolved scales. Note that since the result of a PDF-based parameterization is a condensate content and a cloud fraction, there is a similar requirement as for the Sundqvist scheme to describe the microphysical processes for the conversion of cloud to precipitation.

Fully Prognostic Schemes

In the early 1990s a new approach to cloud parameterization emerged, in which both the time evolution of the cloud condensate and that of cloud fraction are described using prognostic equations,

$$\frac{\partial l}{\partial t} = A(l) + S(l) - D(l) \quad [1a]$$

$$\frac{\partial a}{\partial t} = A(a) + S(a) - D(a) \quad [1b]$$

In eqns [1a] and [1b], l is the grid-mean condensate content and a is the cloud fraction. $A(l, a)$ represents the advection of the two variables, $S(l, a)$ represents any sources of condensate or cloud fraction, and $D(l, a)$ represents their dissipation. This approach was pioneered by Tiedtke and has been introduced into a number of GCMs.

More recently, research has been focusing on combining the fully prognostic approach with that used in statistical schemes. Here, instead of predicting grid-mean condensate and cloud fraction, the moments of a probability density function (mean, variance, skewness) are used as prognostic model variables and

the relevant cloud parameters are deduced from the PDF as in the traditional statistical cloud parameterizations.

Contemporary Issues

Convectively Generated Clouds

Both the introduction of a prognostic variable for the description of cloud condensate and the use of a PDF condensation scheme solve the problem of linking the latent heat effects of clouds with the macroscopic parameters entering the radiation calculations. A major remaining problem in both approaches is that they do not include clouds produced by convective processes as an integral part of their formulation. In models using either of these two cloud parameterization approaches, convective clouds are usually still treated as they were in diagnostic cloud parameterizations.

A variety of approaches for tackling this problem have been devised since then. The most common approach used in the schemes solving a prognostic equation for the condensate is to treat water substance detrained from convective updrafts as a source of liquid water for the 'stratiform' clouds. The exact nature of the link depends on the definition of 'detrainment' and can vary for different schemes. Although using 'detrained' condensate from convection as a source for cloud condensate has become a standard way of linking convection and radiation through cloud formation, the variety of different ad hoc techniques used points to a lack of understanding of how exactly this link should be represented. A further major problem is how to represent the cloud fraction resulting from the detrainment process. Recent parameterizations have attempted to derive consistent treatments of both condensate and cloud fraction from convection. Despite the progress made in this area, the inclusion of clouds generated by convective processes remains an uncertain area of active research.

Process-Oriented Approaches

More and more contemporary cloud parameterizations have moved from what can be described as an integrating approach to a process-oriented treatment of clouds. The difference between the two approaches is illustrated in Figure 3. Figure 3(a) summarizes the concept of integrating cloud parameterizations. Various physical processes, such as resolved scale ascent, convection, turbulence, etc. modify one or several resolved variables and/or their tendency. Those resolved quantities (e.g., relative humidity or its tendency) are then used to evaluate the evolution of the model clouds. A major drawback of this approach is that the effects of parameterized processes, such as convection, that contribute directly to cloud formation and dissipation are first 'integrated' onto the grid scale only to be reinterpreted for subgrid-scale cloud processes.

In contrast, in a process-oriented approach (Figure 3(b)) each potentially cloud-modifying process, resolved (e.g., large-scale ascent) or parameterized (e.g., convection) directly alters the model's cloud variables as well as other resolved-scale model variables. In this way, information available at the level of other physical parameterizations can be directly used in the cloud scheme and the clouds become a more integral part of the parameterization package. The physically more

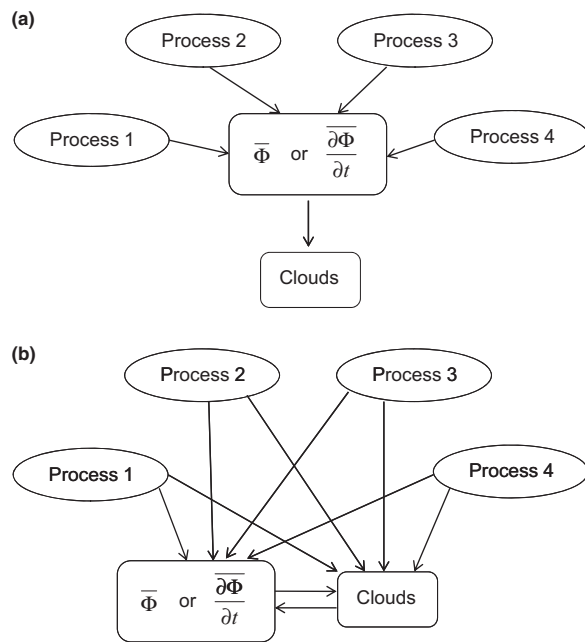


Figure 3 Schematic of the different approaches to cloud parameterization: (a) the principles of ‘integrating’ cloud schemes; (b) the process-oriented approach. Note that arrows indicating the obvious direct interactions between individual processes other than cloud processes have been omitted for clarity.

appealing process-oriented approach to cloud parameterization significantly raises the level of complexity of the parameterization, since the influence that each physical process exerts on the model clouds needs to be explicitly described.

Cloud Microphysics, Subgrid Heterogeneity, and Model Resolution

Most recently, the attention in cloud parameterization has shifted significantly toward the treatment of cloud microphysics. This has been facilitated by increased computing power and the availability of sophisticated microphysics parameterizations from cloud-resolving and mesoscale numerical models. Although increased sophistication in describing cloud and precipitation processes in GCMs is certainly justified, the transplantation of a microphysics scheme from a cloud-resolving model to a GCM is not without problems. This is mainly due to the scales at which the forcing terms (input variables) for the microphysical scheme are available and to the difference in time steps used by the different models. Microphysical processes are highly nonlinear and their parameterization has to rely on the knowledge of the local amount of condensate. In GCMs often only the grid-mean value (or cloud-mean value if cloud fraction is a model variable) for condensate is known, which results in modifications of microphysical constants in order to achieve reasonable cloud condensate and precipitation amounts. The desire to improve on this is leading to increasing emphasis on the representation of subgrid heterogeneity for a more realistic description of the microphysical interactions at scales smaller than the grid scale.

Another trend is the continuing increase in resolution of GCMs. With higher horizontal and vertical resolution, the

concept of cloud fraction actually becomes less important, the forcing becomes better resolved (e.g., updraft strength), and grid-point values of cloud condensate become more representative of local conditions. Hence, as resolution increases, it is common to put less emphasis on the representation of subgrid cloud heterogeneity and apply more complex and physically more realistic parameterizations of cloud microphysics. However, a single GCM can be used with widely varying model resolutions for different applications and there is therefore a continuing requirement for cloud parameterizations with a realistic treatment of both subgrid heterogeneity and microphysics that work effectively and efficiently across a wide range of scales.

Representing the Ice-Phase, Mixed-Phase, and Cloud–Aerosol Interactions

There are still many uncertainties in our knowledge of cloud and precipitation processes and how to represent these in atmospheric models. Ice-phase microphysics is particularly complex due to the number of ways in which ice particles can nucleate and the widely varying particle shapes (habits) observed in the atmosphere. Improving the representation of microphysical processes in the ice-phase, as well as their interactions with supercooled liquid water in the mixed-phase, is an ongoing challenge. There is an increased interest in aerosol–cloud–radiation interactions, particularly in the context of climate change, and although there remains many uncertainties, there is active research to improve our understanding of cloud–aerosol interactions and their representation in atmospheric models.

Evaluating with Observations

In recent years there has been an increase in the number of high quality observations of cloud and precipitation, particularly from remote sensing radar and lidar both at fixed locations on the ground and globally from satellite. These observations are providing a wealth of information on the distribution of cloud and precipitation and their variability in space and time, and an unprecedented opportunity to evaluate models, increase our understanding and improve the parameterization of cloud and precipitation.

See also: **Aerosols:** Aerosol–Cloud Interactions and Their Radiative Forcing. **Clouds and Fog:** Classification of Clouds; Climatology; Cloud Microphysics; Cloud Modeling; Measurement Techniques *In Situ*; Stratus and Stratocumulus. **Numerical Models:** Cloud-System Resolving Modeling and Aerosols; Convective Storm Modeling; General Circulation Models; Regional Prediction Models.

Further Reading

- Cotton, W.R., Anthes, R.A., 1989. *Storm and Cloud Dynamics*. Academic Press, San Diego.
- Emanuel, K.A., 1994. *Atmospheric Convection*. Oxford University Press, New York.
- Emanuel, K.A., Raymond, D.J., 1993. The representation of cumulus convection in numerical models. *American Meteorological Society: Meteorological Monographs* 24 (46), 1–246.

- Houze Jr., R.A., 1993. *Cloud Dynamics*. Academic Press, San Diego.
- Lamb, D., Verlinde, J., 2011. *Physics and Chemistry of Clouds*. Cambridge University Press, Cambridge.
- Liou, K.N., 1992. *Radiation and Cloud Processes in the Atmosphere*. Oxford University Press, New York.
- Pruppacher, H.R., Klett, J.D., 1997. *Microphysics of Clouds and Precipitation*. Kluwer Academic Publishers, Dordrecht.
- Rogers, R.R., Yau, M.K., 1996. *A Short Course in Cloud Physics*, third ed. Butterworth-Heinemann, Oxford.
- Straka, J.M., 2009. *Cloud and Precipitation Microphysics*. Cambridge University Press, Cambridge.
- Sundqvist, H., 1978. A parameterization scheme for nonconvective condensation including prediction of cloud water content. *Quarterly Journal of the Royal Meteorological Society* 104, 677–690.
- Tiedtke, M., 1993. Representation of clouds in large-scale models. *Monthly Weather Review* 121, 3040–3061.

Parameterization of Physical Processes: Gravity Wave Fluxes

MJ Alexander, NorthWest Research Associates (NWRA), Boulder, CO, USA

© 2015 Elsevier Ltd. All rights reserved.

Nomenclature

$\hat{c}$ Intrinsic phase speed.

ϵ Intermittency.

f Coriolis parameter.

F Vertical flux of horizontal momentum.

F_s Saturated momentum flux magnitude.

Fr Froude number.

H Density scale height.

h_0 Orographic height parameter.

J/c_p Thermal forcing.

$\hat{k}$ Unit vector in the vertical.

K_{zz} Vertical diffusion coefficient.

(k, m) (Horizontal, vertical) wave numbers.

λ_z Vertical wavelength.

N Buoyancy frequency.

$\hat{\omega}$ Intrinsic frequency.

Φ Geopotential.

Pr Prandtl number.

ρ Density.

σ_h Subgrid scale topography standard deviation.

R Gas constant.

t Time.

T Temperature.

$\bar{u}$ Wind in the direction of wave propagation.

V Horizontal wind vector.

ω Vertical wind.

X Momentum forcing.

z Vertical distance.

Synopsis

Gravity waves can carry momentum and energy fluxes vertically across deep regions of the atmosphere ranging from the surface to the thermosphere. Dissipation of these fluxes can lead to net changes in the momentum and energy budgets of the surrounding fluid. These processes are important to global circulation, and parameterization approaches for global circulation models are summarized. Constraining information from observations is also briefly summarized along with outstanding issues and recent developments.

Introduction

Gravity waves are oscillations that cause perturbations in the winds, temperature, density, and pressure fields in the atmosphere. They are also called buoyancy waves (*see Gravity Waves: Buoyancy and Buoyancy Waves: Theory*) because the restoring force for their oscillation is the buoyancy force associated with vertical parcel displacements in a stably stratified fluid. Because the density of the atmosphere decreases exponentially with height, vertically propagating gravity waves have amplitudes that tend to grow exponentially with height in the absence of dissipation. Therefore, even small-amplitude waves in the lower atmosphere may have very large effects at high altitudes. Gravity waves carry momentum and energy fluxes, so their dissipation can lead to net changes in the momentum and energy budgets of the surrounding fluid.

The effects of dissipation of gravity wave momentum fluxes on the larger-scale circulation must be parameterized in global models because the resolution required to model them directly is prohibitively fine for most weather forecasting and climate applications. The waves important to the circulation in the atmosphere have horizontal wavelengths ranging from about ten to thousands of kilometers. The largest of these can be resolved in today's global models. However, the vertical wavelength of a gravity wave will vary substantially with height

owing to the effects of wind shear, and it is important to represent vertical wavelengths as small as 1 km. Gravity waves with periods as short as 5–10 min can carry significant momentum flux vertically. Further, the sources of these waves include processes that are also parameterized and/or poorly resolved, namely, convective heating, fine-scale topography, imbalance in the jet stream, and frontal structures. Direct modeling of the spectrum of gravity waves in global models is therefore still not feasible at the present time.

Parameterization of Gravity Wave Effects

Parameterization of gravity waves generated by flow over fine-scale topography is now widely used in atmospheric climate and weather forecasting models. When the effects of these waves were introduced in climate models, they significantly reduced an eastward wind bias in the Northern Hemisphere winter troposphere and lower stratosphere that had previously tended to grow worse with improved resolution. The parameterized orographic gravity wave drag provided a realistic dissipation process that transported momentum from the surface to the free atmosphere.

Global models of the atmosphere that include many levels above the tropopause also require parameterization of gravity

waves from other sources. Parameterized orographic waves are characterized by phase speed $c = 0$ relative to the ground (stationary waves), while waves from other sources may have wide-ranging phase speeds (nonstationary waves). The detailed characteristics of the nonstationary waves are not well understood. Parameterizations of these tend to assume a simple latitudinally and/or seasonally varying source spectrum of wave phase speeds. The effects of these waves are most pronounced in the mesosphere, where the drag they exert on the middle atmosphere jets drives a strong summer-to-winter meridional circulation that reverses the temperature gradient that would be predicted from consideration of solar heating alone. In the stratosphere, gravity wave forcing also drives a mean meridional circulation, though the contribution of gravity waves is secondary to the planetary wave forcing. Dissipation of gravity wave energy fluxes causes important vertical mixing of heat and constituents in the upper mesosphere and lower thermosphere, and this is a very important process in the oceans.

The parameterization of gravity wave effects has therefore focused on estimation of two terms for inclusion in the fundamental fluid equations solved in global models.

- A body forcing term $\mathbf{X}$ for the momentum equation, e.g., eqn [1].

$$\frac{D\mathbf{V}}{Dt} = -f\mathbf{k} \times \mathbf{V} - \nabla\Phi + \mathbf{X} \quad [1]$$

Here $\mathbf{V}$ is the vector wind, f is the Coriolis parameter, $\mathbf{k}$ is the unit vector in the vertical, and Φ is the geopotential (see **Dynamical Meteorology: Overview** for additional discussion). The meaning of eqn [1] is that the acceleration of a fluid parcel $D\mathbf{V}/Dt$ is equal to the sum of forces per unit mass. The three forces shown here are the Coriolis force ($-f\mathbf{k} \times \mathbf{V}$), the pressure gradient force ($-\nabla\Phi$), and the wave-driven force $\mathbf{X}$.

- An ‘eddy diffusion coefficient’ K_{zz} for describing vertical mixing effects on temperature and trace constituents in the thermodynamic equation and in equations describing tracer transport, e.g., eqn [2].

$$\frac{DT}{Dt} + \frac{wN^2H}{R} = \frac{J}{c_p} + \frac{1}{\rho} \frac{\partial}{\partial z} \left[\rho K_{zz} \left(\frac{\partial T}{\partial z} + \frac{g}{c_p} \right) \right] \quad [2]$$

This is the thermodynamic energy equation including a diffusive mixing term due to unresolved gravity waves. T is the temperature, w is the vertical velocity, N is the buoyancy frequency, H is the scale height, R is the gas constant, and J/c_p defines the thermal forcing (again see **Dynamical Meteorology: Overview**). Gravity wave mixing effects are considered only in the vertical because the atmosphere is generally treated as horizontally homogeneous over the scale of the unresolved waves. A similar mixing term can be applied to conservation equations for trace constituents.

The eddy diffusion term is only known to be important at mesospheric and lower thermospheric altitudes, while the momentum forcing term is important at levels ranging from the upper troposphere to the lower thermosphere. Momentum diffusion and direct heating terms can also arise from gravity wave dissipation, but these have been neglected in most parameterization applications. However, the heating term is

now believed to be important locally in the mesosphere energy budget (see **Gravity Waves: Overview** for additional discussion of the physics of these effects and their importance in the atmosphere).

Simplifying Assumptions

Parameterizations in global models assume that wave propagation is one dimensional and purely vertical. Effects of the Coriolis force on waves with the lowest intrinsic frequencies are also usually neglected ($\hat{\omega}^2 \gg f^2$, where $\hat{\omega}$ is the wave intrinsic frequency). The hydrostatic approximation is also generally made, which assumes that the horizontal wavelength is much longer than the vertical wavelength, or $\hat{\omega}^2 \ll N^2$. With these assumptions, momentum forcing is proportional to the vertical gradient of the vertical flux of horizontal momentum (or Reynolds stress) $\mathbf{F} = \rho(\overline{u'w'}, \overline{v'w'})$ carried by the wave and eqn [3] applies,

$$\mathbf{X} = -\frac{1}{\rho} \frac{\partial \mathbf{F}}{\partial z} \quad [3]$$

The parameterization problem then reduces to (1) specifying the gravity wave flux $\mathbf{F}_0$ at some initial height z_0 usually taken to be the source altitude somewhere in the troposphere or near the tropopause, then (2) computing the dissipation as a function of height or $\mathbf{F}(z)$. Dissipation could be due to any number of effects including radiative damping, nonlinear wave-wave interactions, wave breaking due to convective or dynamical instability, and molecular diffusion. Different parameterizations make different assumptions about which dissipation processes are important, but for calculating of the momentum forcing term $\mathbf{X}$, only the net effect of these processes on $\mathbf{F}(z)$ is important. Figure 1 shows a schematic profile of momentum flux for a wave carrying positive momentum flux dissipating as a function of height. This dissipation via eqn [3] results in a positive force. The direction of the force will always be such as to accelerate or ‘drag’ the background wind speed toward the wave intrinsic phase speed.

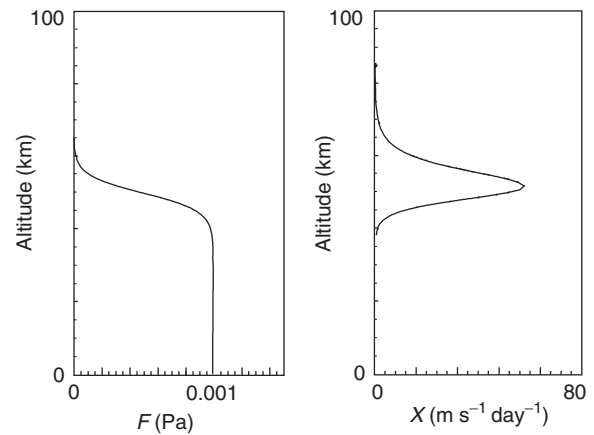


Figure 1 An example vertical profile of momentum flux $\mathbf{F}$ and the force $\mathbf{X}$ resulting from the dissipation of a wave with initial positive momentum flux (and positive intrinsic phase speed) as a function of height.

When waves break down, mixing is likely, and the eddy diffusion term can be important. High-resolution numerical model studies have also shown that the specific characteristics of the wave at the point of breaking or the nature of the instability process may strongly affect the degree of mixing. This makes the relationship between the momentum forcing and the vertical diffusion uncertain. Given the changes observed in gravity waves with height described in article **Gravity Waves: Overview**, the relationship between X and K_{zz} may even vary dramatically with height between the tropopause and the upper atmosphere.

Lindzen-Type Parameterizations

The most commonly applied parameterizations are based on a formulation by Richard Lindzen, first outlined in 1981, and these are generally referred to as ‘Lindzen’ or ‘Lindzen-type’ parameterizations. These treat the spectrum of gravity waves as a discrete set of simple plane waves varying only in their phase speeds c and propagation directions. Each wave mode is assigned an amplitude at a specified source level z_0 . As the waves propagate vertically, they are Doppler shifted and refracted by vertical variations in the background wind. Doppler shifting is described by the change in the intrinsic phase speed $\hat{c}(z)$ or frequency $\hat{\omega}(z)$ associated with vertical shear in the component of horizontal wind in the direction of the wave propagation $\bar{u}$ (eqn [4]). k is the horizontal wave number.

$$\hat{c}(z) = \frac{\hat{\omega}(z)}{k} = c - \bar{u}(z) \quad [4]$$

Refraction describes the corresponding change in vertical wave number m with height (eqn [5]).

$$m(z) = \frac{N(z)k}{\hat{\omega}(z)} \quad [5]$$

Refraction also occurs with changes in N . When the wave amplitude exceeds the threshold for convective instability (see **Gravity Waves: Overview**), the wave amplitude is assumed to saturate at that threshold value and grow no further. This assumption gives an estimate of wave dissipation as a function of height. In terms of the momentum flux, the saturation limit can be written as eqn [6].

$$|F| \leq F_s = \text{Fr}^2 \frac{\rho k}{2N} |c - \bar{u}|^3 \quad [6]$$

Here, F_s is the saturated value of the momentum flux and Fr is the critical Froude number (generally assumed ~ 1). Applying eqn [3] then gives the momentum force in the saturated region (eqn [7]).

$$X = \varepsilon \frac{k(c - \bar{u})^3}{2N} \left(\frac{1}{H} + \frac{3d\bar{u}/dz}{c - \bar{u}} \right) \quad [7]$$

Vertical gradients in N have been neglected, and the shear term is also sometimes neglected. The factor ε has been described as an efficiency factor or intermittency. For parameterization of nonzero phase speed waves, the sources are generally specified to be globally uniform, or varying in simple ways, because we lack better knowledge of the detail of how to treat these waves. The factor ε then is meant to describe the

fractional coverage of waves and wave sources in space and time. It is applied only in the force eqn [7] and not to the flux and saturation condition eqn [6] because the local amplitudes of the waves should determine the breaking levels and saturation regions. If ε is small, the spatially and temporally averaged amplitudes of gravity waves will be much smaller than the local amplitudes.

A vertical diffusion coefficient, derived by assuming the wave dissipation described by eqn [6] occurs because local mixing effects maintain the local temperature gradient at the adiabatic temperature lapse rate (the limit for convective instability). This assumption gives eqn [8].

$$K_{zz} = \frac{1}{\text{Pr}} \frac{(c - \bar{u})X}{N^2} \quad [8]$$

Note that in eqn [8] the uncertain Prandtl number (Pr) has been included explicitly. This factor describes the ratio of thermal to momentum diffusion and has been assigned values ranging from 1 to 10 in model applications. While the force X always has the same sign as the intrinsic phase speed $\hat{c} = c - \bar{u}$, the vertical diffusion coefficient is always positive.

Orographic Wave Parameterization

For orographic wave drag parameterizations, the wave phase speed is set to $c = 0$, which simplifies eqns [6]–[8] to eqns [9]–[11].

$$|F| \leq F_s = \varepsilon \text{Fr}^2 \frac{k}{2N} |\bar{u}|^3 \quad [9]$$

$$X = \varepsilon \frac{k\bar{u}^3}{2N} \left(\frac{1}{H} - \frac{3}{\bar{u}} \frac{d\bar{u}}{dz} \right) \quad [10]$$

$$K_{zz} = -\frac{1}{\text{Pr}} \frac{\bar{u}X}{N^2} \quad [11]$$

The efficiency factor ε is applied not only to the force eqn [10] but also to the local fluxes eqn [9] because here ε describes processes that limit wave amplitudes at the source, such as flow blocking. The spatial and temporal intermittency is also explicit in the source description since waves are only generated over topography and their amplitudes will depend on the background wind speed.

Many variants exist for how to specify the topographic source flux F_0 , but one of the simplest that is still commonly applied in global models assumes the source is a function of the topographic height standard deviation σ_h of subgrid scale topography within a model grid box, e.g., eqn [12], where the subscripts refer to values at the source level close to the surface.

$$F_0 = -\frac{\varepsilon k}{2} \rho_0 N_0 V_0 h_0^2 \quad [12]$$

The amplitude $h_0 = \min(2\sigma_h, \text{Fr}V_0/N_0)$, where the Froude number limit crudely accounts for blocking effects.

Spectral Parameterizations

Parameterization methods for describing a full spectrum of gravity wave effects are now in use, and are an essential

component of chemistry–climate models. Spectral parameterizations are intended to describe waves generated by a collection of sources, and are generally applied together with an orographic wave drag scheme. With discrete parameterizations like the Lindzen-type scheme with only a few wave phase speeds specified, the drag effects on the mean flow can be very sudden or discrete in time and in the horizontal plane because of the dependence of the breaking criteria (eqn [6] or [9]) on $(c - \bar{u})$. The spectral parameterizations can provide smoother gravity wave effects for incorporation in global models seeking to describe the climatological effects of gravity waves on the mean flow.

The mechanism proposed to describe wave dissipation as a function of height varies among the different spectral parameterizations, and each also has restrictions on the properties of the wave spectrum that can be specified at the source level. The current uncertainty that exists in how to specify the gravity waves at their sources is large enough that it remains

difficult to separate source differences in these models from differences in the description of wave dissipation. Figures 2 and 3 serve to illustrate this problem.

Figure 2(a) shows normalized energy spectra that differ only in the shape of the spectrum at long vertical wavelengths $25 \text{ km} \geq \lambda_z \geq 2.5 \text{ km}$. Long vertical wavelength waves $>10 \text{ km}$ are more difficult to observe in the lower stratosphere near gravity wave sources. The three spectra in Figure 2(a) are meant to describe the range of uncertainty in the long-wavelength portion of the spectrum. Figure 2(b) shows the three corresponding momentum flux spectra as a function of intrinsic phase speed $\hat{c}$, where it has been assumed that the waves propagate in the zonal direction and the spectrum is symmetric. The uncertainty in long vertical wavelength waves translates to an uncertainty in waves with high intrinsic phase speeds. The conversion from energy to momentum flux requires knowledge of the intrinsic frequency, which is also difficult to determine from observations, adding additional

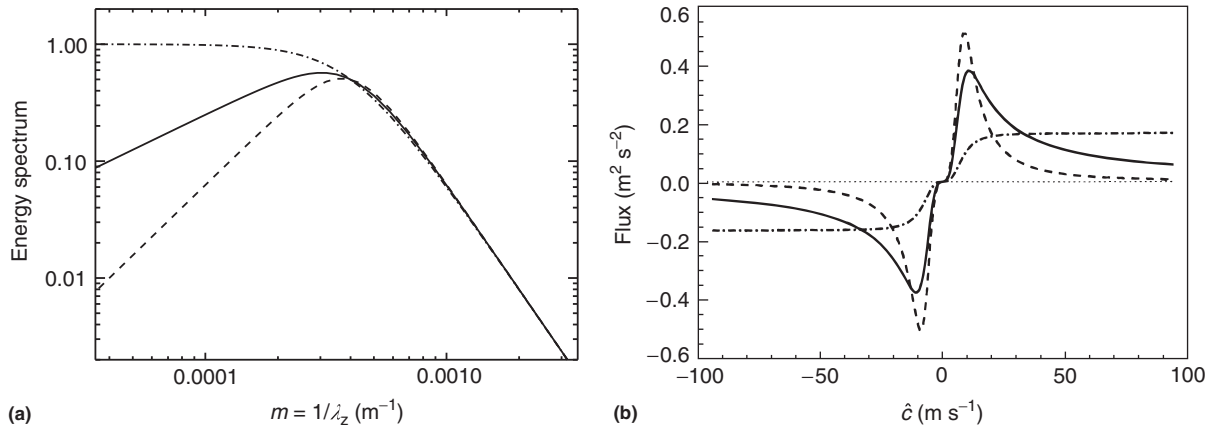


Figure 2 Gravity wave spectra for (a) energy as a function of vertical wave number m and (b) momentum flux per unit density as a function of intrinsic phase speed $\hat{c}$. The dashed, solid, and dash-dotted lines in each panel represent spectra that vary at low wave numbers as m^2 , m^1 , and m^0 , respectively.

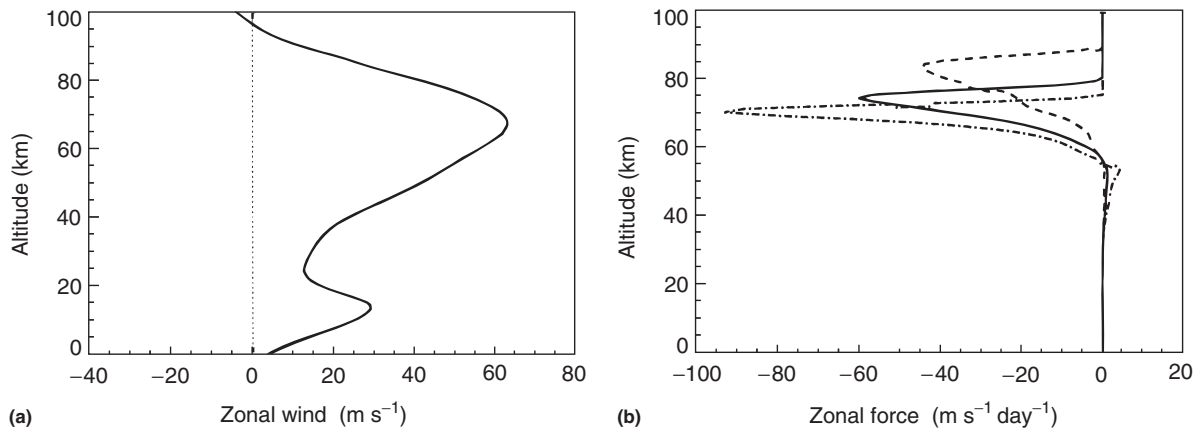


Figure 3 Vertical profiles of (a) zonal wind input into the parameterized momentum force calculation and (b) the three force profiles that result from the three different momentum flux spectra in Figure 2(b). To derive the force, a horizontal wavelength of 200 km was assumed, and the efficiency factors for each spectrum were chosen so that the integrated momentum flux in each spectrum was the same (equal eastward and westward fluxes of 0.001 Pa).

uncertainty. The three curves have been chosen to illustrate cases with different dependences on high and low intrinsic phase speed waves.

Results of these spectra input to a gravity wave parameterization calculation of the momentum forcing are shown in Figure 3. The integrated momentum flux input is specified to be a constant among the three cases and the source altitude is set at 5 km. The parameterization employed is similar to the Lindzen-type parameterizations, but the spectrum is broken into small increments $\Delta\hat{c} = 1 \text{ m s}^{-1}$, and instead of employing the saturation condition eqn [6] to describe the dissipation as a function of height, each spectral element loses all of its momentum flux at the breaking level.

Figure 3(a) shows the zonal wind profile assumed, and the three zonal force profiles corresponding to the three phase speed spectra are shown in Figure 3(b). The peaks between 60 and 90 km are most sensitive to the large negative intrinsic phase speed portion of the spectrum that has been so difficult to quantify with observations. These differences arise primarily because of the uncertainty in our understanding of this high phase speed portion of the gravity wave spectrum in nature.

Constraints on Gravity Wave Momentum Fluxes

Existing constraints on gravity wave momentum fluxes and forces on the mean flow used to constrain gravity wave drag parameterizations are reviewed here. Every measurement is limited to observing a specific range of gravity wave properties, while other possible waves in the spectrum will not be visible. This leads to the expectation that each observational constraint may be low biased to some extent.

Measurements Local to Sources

Observations of gravity waves just above a localized known source can provide some constraints for parameterizations. Relatively few observations, however, are useful for deriving momentum fluxes. Fewer still can also specify the wave phase speeds, horizontal wave numbers, and propagation directions required to accurately predict their effects on the atmosphere at higher levels. A still smaller number has the capacity to provide the sort of general information needed to parameterize gravity wave sources and how they vary in changing conditions geographically and seasonally.

Observation of gravity waves by radar and aircraft in the lower stratosphere has provided gravity wave momentum fluxes above topographic and convective sources. (Observations have generally been reported as horizontal and vertical wind covariances in units of $\text{m}^2 \text{s}^{-2}$. Those reported here are in $\text{Pa} = \text{kg m}^{-1} \text{s}^{-2}$, which have been multiplied by density ρ . This allows easier comparison of observations at different altitudes because in the absence of dissipation, wind covariances will grow exponentially with height in proportion to ρ^{-1} .) Observations from satellite platforms have also provided some information on gravity wave momentum fluxes to date. These have shown local momentum fluxes. When wave activity was observed over topography of approximately $0.02\text{--}1 \text{ Pa}$. Orographic waves are not always observed to be stationary, but

generally have slow phase speeds and very low frequencies relative to the ground. Model studies have shown that time-varying winds and nonlinear effects can lead to the generation of orographic waves with nonzero phase speeds. Horizontal wavelengths of orographic waves should be related to the scales of surface roughness. Scales observed with relevance to parameterization of circulation effects range from approximately ten to hundreds of kilometers.

Waves observed over deep convective clouds have been reported with momentum fluxes approximately $0.01\text{--}0.2 \text{ Pa}$. Short horizontal wavelengths approximately $10\text{--}100 \text{ km}$ have been observed by aircraft. From satellite measurements, waves from convection have been inferred by proximity to deep clouds and by characteristic concentric ring-shaped patterns. These have shown wavelengths ranging from a few tens to hundreds of kilometers. Longer horizontal wavelengths approximately 500 to several thousand kilometers have been inferred from balloon soundings at tropical locations where convection is the likely source. These are low frequency inertia-gravity waves, and although their horizontal scales are resolvable in most global models, their vertical wavelengths are often too short, and their sources are still poorly represented. Phase speeds for convectively generated waves have also been reported to vary greatly from near zero to approximately 50 m s^{-1} , although the uncertainty in phase speed estimates can be large.

Estimates of mean flow forcing derived from observations of momentum flux remain too uncertain to provide meaningful constraints. However, given the strength of localized fluxes observed and some reasonable assumptions, it is likely that localized forces $\mathbf{X}$ occur at times in the mesosphere with magnitudes of at least hundreds of meters per second per day.

Fluxes and properties of waves appearing above regions of imbalance in the jet stream have been inferred from a few case studies, but no generalization of their properties can yet be described. Model studies suggest that a wide range of horizontal scales are emitted from this source, but phase speeds relative to the ground tend to be small, a few meters per second, being tied to the motion of the underlying synoptic systems. Although aircraft measurements have been used to identify enhanced variance above frontal sources due to waves with scales less than $\sim 200 \text{ km}$, further information on wave properties needed to constrain parameterizations is lacking.

Constraints on Long-Term Average Momentum Fluxes

Longer term averaged gravity wave momentum fluxes provide additional valuable constraints for parameterizations, and together with the local flux constraints above, they illuminate intermittency. Global maps of momentum flux in the lower stratosphere have been computed from satellite measurements and long-duration balloon data that reveal seasonal and geographic patterns. In general, the largest fluxes occur in winter at mid-to-high latitudes. Fluxes near the equator are generally smaller, and a secondary maximum occurs at summer subtropical latitudes. Within these seasonal patterns, localized peaks appear over topography in winter, with monthly- and seasonal-mean peak values of $5\text{--}20 \times 10^{-3} \text{ Pa}$. Away from topography, long-term mean winter fluxes have typical values of approximately $1\text{--}3 \times 10^{-3} \text{ Pa}$. In summer, peaks over

convection are observed with values $\sim 2 \times 10^{-3}$ Pa and background values $< 1 \times 10^{-3}$ Pa. Extremely low values approaching noise limits occur at high summer latitudes poleward of $\sim 60^\circ$.

Although these mean fluxes are at least 10–100 times smaller than localized values, vertical propagation through decreasing atmospheric density and dissipation at high altitudes in the mesosphere and above can lead to large drag forces. Estimates of these forces derived from observations remain too uncertain to provide any meaningful constraints. Global models with self-consistent momentum budget calculations predict forces due to gravity waves of tens of meters per second per day near the mesopause, and in the lower stratosphere due to strong orographic wave forcing, values on the order of $1 \text{ m s}^{-1} \text{ day}^{-1}$.

Recent Developments

Gravity waves generated by flow over orography are parameterized in most global climate and weather forecasting models. Recent developments include more realistic description of the wave properties and momentum fluxes at the source level F_0 , including more realistic descriptions of flow blocking and downstream vortex formation effects, and coupling to surface friction. Recent studies have related some of these changes to the frequency of occurrence of sudden stratospheric warmings in models and to surface climate response patterns. Climate studies have also found that parameterized orographic wave drag responds to long-term changes in zonal winds near the tropopause in a way that leads to acceleration of the stratospheric mean meridional transport circulation.

These results underscore the need for gravity wave parameterizations applied in climate models to respond to long-term changes in climate in realistic ways. All parameterization methods currently respond to changes in zonal wind and stability in the middle atmosphere. In orographic schemes, the source flux F_0 (eqn [12]) also responds to climate changes in near-surface winds and stability. Conversely, most spectral schemes intended to represent other non-orographic gravity waves have sources remain fixed. Some models are now including a variety of different non-orographic gravity wave sources with fluxes that do respond to changes in climate. Such ‘non-orographic source parameterizations’ have been developed for both convective sources and frontal sources and have been applied in a few climate models. These source parameterizations have been based on two-dimensional model studies of convection and frontogenesis. While these convective and frontal source parameterizations represent an advance by allowing the sources to respond in realistic ways to climate changes, the many assumptions made in their application remain unvalidated, leaving large uncertainties in the model responses at present.

Intermittency in the gravity wave fluxes is generally treated with a simple tuning parameter, such as ε in eqn [7]. Intermittency is in reality much more complex: Observed wave amplitudes span at least a factor of 100 with the largest

amplitude waves occurring as extremely rare events. Although in an average sense the momentum fluxes from large- and small-amplitude events may be similar, their effects on the atmosphere will be very different. Large rare events will break at lower altitudes and may force localized responses and/or emit substantial secondary waves, while ubiquitous small events may break at higher altitudes and instead force circulations more similar to the climatological mean effects described in the introduction. New interest in stochastic parameterization methods may open opportunities to include effects of intermittency in more realistic ways in future model applications.

See also: **Dynamical Meteorology:** Overview. **Gravity Waves:** Buoyancy and Buoyancy Waves: Optical Observations; Buoyancy and Buoyancy Waves: Theory; Overview.

Further Reading

- Alexander, M.J., Dunkerton, T.J., 1999. A spectral parameterization of mean-flow forcing due to breaking gravity waves. *Journal of the Atmospheric Sciences* 56, 4167–4182.
- Alexander, M.J., Geller, M., McLandress, C., Polavarapu, S., Preusse, P., Sassi, F., Sato, K., Eckermann, S., Ern, M., Hertzog, A., Kawatani, Y., Pulido, M., Shaw, T., Sigmond, M., Vincent, R., Watanabe, S., 2010. Recent developments in gravity wave effects in climate models, and the global distribution of gravity wave momentum flux from observations and models. *Quarterly Journal of the Royal Meteorological Society* 136, 1103–1124.
- Butchart, N., Cionni, I., Eyring, V., Waugh, D.W., Akiyoshi, H., Austin, J., Brühl, C., Chipperfield, M.P., Cordero, E., Dameris, M., Deckert, R., Frith, S.M., Garcia, R.R., Gettelman, A., Giorgetta, M.A., Kinnison, D.E., Li, F., Mancini, E., McLandress, C., Pawson, S., Pitari, G., Plummer, D.A., Rozanov, E., Sassi, F., Scinocca, J.F., Shepherd, T.G., Shibata, K., Tian, W., 2010. Chemistry-climate model simulations of 21st century stratospheric climate and circulation changes. *Journal of Climate* 23, 5349–5374.
- Garcia, R.R., Marsh, D.R., Kinnison, D.E., Boville, B.A., Sassi, F., 2007. Simulation of secular trends in the middle atmosphere, 1950–2003. *Journal of Geophysical Research* 112. <http://dx.doi.org/10.1029/2006JD007485> (see Appendix A).
- Hamilton, K. (Ed.), 1997. *Gravity Wave Processes, Their Parameterization in Global Climate Models. Series I, Global Environment Change*, vol. 50. Springer-Verlag, Heidelberg.
- Holton, J.R., 1982. The role of gravity wave induced drag and diffusion in the momentum budget of the mesosphere. *Journal of the Atmospheric Sciences* 39, 791–799.
- Holton, J.R., 1992. *An Introduction to Dynamic Meteorology*. Academic Press, San Diego, CA.
- Lindzen, R.S., 1981. Turbulence and stress owing to gravity wave and tidal breakdown. *Journal of Geophysical Research* 86, 9707–9714.
- McFarlane, N.A., 1987. The effect of orographically excited gravity wave drag on the general circulation of the lower stratosphere and troposphere. *Journal of the Atmospheric Sciences* 44, 1775–1800.
- McLandress, C., 1998. On the importance of gravity waves in the middle atmosphere and their parameterization in general circulation models. *Journal of Atmospheric Solar-Terrestrial Physics* 60, 1357–1383.
- Webster, S., Brown, A.R., Cameron, D.R., Jones, C.P., 2003. Improvements to the representation of orography in the Met Office Unified Model. *Quarterly Journal of the Royal Meteorological Society* 129, 1989–2010.

Parameterization of Physical Processes: Turbulence and Mixing

A Beljaars, European Centre for Medium-Range Weather Forecasts, Reading, England

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

The representation of subgrid turbulence in meteorological models for numerical weather prediction and climate is discussed. The following topics are addressed: (1) The importance of turbulence parametrization in large scale models; (2) the choice of conserved variables for mixing in the vertical; (3) the representation of surface fluxes in large scale models; (4) commonly used boundary layer formulations, namely, bulk schemes, local closure, profile formulations, and higher order models; and (5) the cloudy boundary layer.

Introduction

Turbulence is a prominent feature in the atmosphere, particularly in the atmospheric boundary layer. Turbulence length scales vary typically from a few millimeters to a few hundreds of meters, which are smaller than the grid spacing in models of the atmosphere. However, these small-scale motions have a strong impact on the evolution of the mean flow and therefore a parametrization scheme is needed in models that simulate the flow in the atmosphere. A mathematical framework to describe the effect of turbulence on the mean flow can be obtained by separating variables into a mean and a fluctuating part: the so-called Reynolds decomposition. The resulting equations for the mean flow contain variances and covariances of fluctuating quantities that are unknown. This is known as the closure problem. Expressing the unknown quantities in terms of mean flow variables is called turbulence parametrization.

The turbulent transport terms are among the dominant terms in the atmospheric boundary layer equations and govern

the exchange of heat, moisture, and momentum with the surface. In a climate model, it is for instance through the turbulence scheme that the model knows about the sea surface temperature and the heat and moisture input from the surface. Also the momentum sink due to drag at the surface is predominantly controlled by the turbulence parametrization. **Figure 1** shows a latitude height cross section of the mean moisture tendency from turbulent diffusion in a global model. Although turbulence can occur anywhere in the atmosphere (e.g., as clear-air turbulence near the jet stream), the main tendencies are near the surface, illustrating the importance of turbulence for exchange with the surface.

Models of the atmosphere cover a wide range of applications, e.g., canopy flow models, flow over complex topography, air pollution dispersion models, internal boundary layer models, mesoscale models, numerical weather prediction (NWP) models, and climate models. All these models have turbulence schemes to describe the effect of turbulent transport on the mean flow. Dependent on the application, the requirements may vary and it is beyond the scope of this article to cover all options.

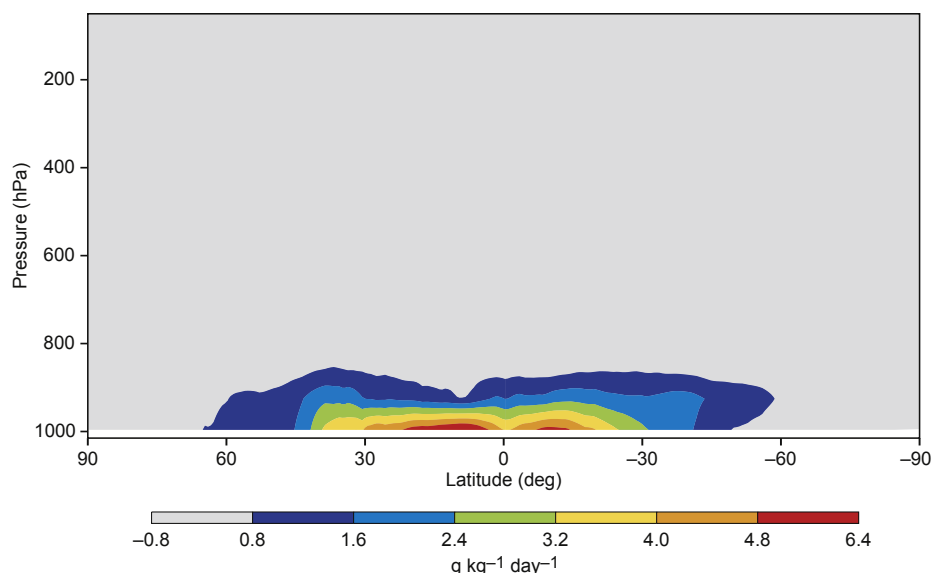


Figure 1 North–South vs height cross section of the zonal mean moisture tendencies in the European Centre for Medium-Range Weather Forecasts (ECMWF) numerical weather prediction model. The units are $\text{g kg}^{-1} \text{ day}^{-1}$ with a contour interval of 0.5. Courtesy of Ernst Klinker, ECMWF.

However, most turbulence schemes are built on the same principles. Here we will focus on NWP and climate models.

Turbulence is a very complex and nonlinear flow phenomenon for which no general theory exists. Therefore turbulence schemes rely heavily on insight in the physics of turbulent flow, empirical relations from observations, and on similarity arguments to represent observations. With the increased availability of computer power large eddy simulation models have also become important in providing 'data' to replace or complement observations.

Turbulence can be generated by wind shear or by buoyancy, the latter acting as a strong modulator of the diurnal cycle over land. An atmospheric boundary layer heated from the surface during the daytime is highly turbulent through buoyancy production, whereas the nighttime boundary layer can be very calm through the suppression of turbulence by stable stratification.

Buoyancy generation of turbulence is not always due to heating from the surface. Unstable stratification can also be generated, e.g., by differential advection or by radiative cloud top cooling. Stratocumulus for instance can maintain a well-mixed structure because the cloud top cooling leads to buoyancy-generated turbulence and strong vertical mixing.

Variables

For the description of turbulent transport in the vertical, it is necessary to use thermodynamic variables that are conserved for adiabatic ascent or descent. For the dry case, suitable variables are potential temperature θ or dry static energy s together with specific humidity q (see also Table 1 for the definition of symbols).

Table 1 Symbols and units

Symbols	Description	Units
C_m	Transfer coefficient for momentum	—
C_{mn}	Neutral transfer coefficient for momentum	—
C_h	Transfer coefficient for heat	—
C_{hn}	Neutral transfer coefficient for heat	—
C_q	Transfer coefficient for moisture	—
C_{qn}	Neutral transfer coefficient for moisture	—
C_p	Specific heat at constant pressure	$\text{J kg}^{-1} \text{K}^{-1}$
d/dt	Total derivative $\partial/\partial t + U\partial/\partial x + V\partial/\partial y + W\partial/\partial z$	s^{-1}
E	Turbulence kinetic energy	$\text{m}^2 \text{s}^{-2}$
F_m	Louis momentum stability function of Ri	—
F_h	Louis heat stability function of Ri	—
g	Gravitation constant	m s^{-2}
h	Boundary layer depth	m
k	Von Kármán constant	—
l	Mixing length	m
L	Obukhov length	m
L_v	Latent heat of vaporization of water	J kg^{-1}
p	Pressure	Pa
p_o	Surface pressure	Pa
q	Specific humidity	kg kg^{-1}
q_l	Liquid water/ice content	kg kg^{-1}
q_t	Total water content	kg kg^{-1}
q_{sat}	Saturation specific humidity	kg kg^{-1}
R	Gas constant	$\text{J kg}^{-1} \text{K}^{-1}$

(Continued)

Table 1 Symbols and units—cont'd

Symbols	Description	Units
R_i	Gradient Richardson number	—
R_{ib}	Bulk Richardson number	—
s	Dry static energy	J kg^{-1}
s_l	Liquid water static energy	J kg^{-1}
s_v	Virtual dry static energy	J kg^{-1}
t	Time	s
T	Temperature	K
U	Velocity in x -direction	m s^{-1}
V	Velocity in y -direction	m s^{-1}
W	Velocity in z -direction	m s^{-1}
u_*	Friction velocity	m s^{-1}
w_*	Free convection velocity scale	m s^{-1}
w_e	Entrainment velocity	m s^{-1}
x	Horizontal coordinate	m
y	Horizontal coordinate	m
z	Vertical coordinate	m
z_{om}	Roughness length for momentum	m
z_{oh}	Roughness length for heat	m
z_{oq}	Roughness length for moisture	m
Greek symbols	Description	Units
Δ	Temperature jump at boundary layer top	K
ε	Turbulence dissipation rate	$\text{m}^2 \text{s}^{-3}$
ϕ_m	MO gradient stability function for momentum	—
ϕ_h	MO gradient stability function for heat	—
γ	Lapse rate above boundary layer	K m^{-1}
ψ_m	MO integral stability function for momentum	—
ψ_h	MO integral stability function for heat	—
θ	Potential temperature	K
θ_l	Liquid water potential temperature	K
θ_v	Virtual potential temperature	K
ρ	Density	kg m^{-3}
λ	Asymptotic mixing length	m
ν	Kinematic viscosity	$\text{m}^2 \text{s}^{-1}$
Symbols	Subscripts, superscripts etc. for variable X	
X_l	X -Value at lowest model level	
X_s	X -Value at surface	
X_m	Mixed layer value of X	
X_o	X -Value at the surface (for fluxes)	
X_h	X -Value at boundary layer top	
X'	Fluctuation of X	
$\bar{X}$	Averaged X	
$\vec{X}$	Vector X	

$$\theta = T(p_o/p)^{R/c_p}, \quad [1]$$

$$s = c_p T + gz. \quad [2]$$

In these expressions, T represents the temperature, p is the pressure, p_o is the reference pressure (usually 1000 hPa), R is the gas constant, c_p is the specific heat capacity of air at constant pressure, g is the gravitational acceleration, and z is the height above the surface. For turbulence in cloudy situations, it is necessary to consider the latent heat release in saturated parcels during adiabatic ascent and descent. Suitable conserved variables are liquid water potential temperature θ_l or liquid water static energy s_l and total water content q_t .

$$\theta_1 = \theta - \left(\frac{L_v \theta}{c_p T} \right) q_1, \quad [3]$$

$$s_1 = c_p T + gz - L_v q_1, \quad [4]$$

$$q_t = q + q_1, \quad [5]$$

where L_v is the latent heat of vaporization of water, and q_1 is the specific liquid water content. When clouds are present (i.e., $q_1 > 0$), q is equal to the saturation value q_{sat} at temperature T .

To quantify buoyancy and stability, a variable is needed that reflects the density of the fluid. For instance, static stability is determined by the density of a fluid parcel moved adiabatically to a reference height in comparison with the density of the surrounding fluid. The virtual potential temperature θ_v and the virtual dry static energy s_v are often used for this purpose:

$$\theta_v = \theta(1 + 0.61q - q_1), \quad [6]$$

$$s_v = c_p T(1 + 0.61q - q_1) + gz. \quad [7]$$

Surface Layer

The surface layer, also called the constant flux layer, is the lower part of the boundary layer where the fluxes are close to the surface values. The concept is an asymptotic one, with the constant flux layer assumption being valid only for heights that are small compared to the boundary layer height. In practice the criterion of a maximum of 10% deviation of the surface flux is often used. With a linear decrease of fluxes from the surface to the top of the boundary layer, it implies that 10% of the boundary layer depth can be considered to be the surface layer. So in a 1000-m deep daytime mixed layer, the surface layer has a depth of the order of 100 m, whereas in the stable nighttime boundary layer, which is only a few hundred meters deep, the surface layer can be as shallow as 10 m.

It is a common practice in turbulence models to use Monin–Obukhov (MO) similarity between the surface and the lowest model level. The theory of MO similarity is well established and the corresponding empirical functions have been the subject of extensive observational studies (Högström, 1988). Although MO similarity works well, in principle it is limited to stationary flow over homogeneous terrain. This is a serious limitation as flow in, e.g., NWP and climate models over real terrain, which is seldom homogeneous (Table 1).

The standard way of expressing (kinematic) surface fluxes of momentum $\overline{u'w'}_0$ and $\overline{v'w'}_0$, sensible heat $\overline{w'\theta'}_0$, and moisture $\overline{w'q'}_0$ into wind, temperature, and moisture differences over the surface layer is with the help of transfer coefficients (e.g., Brutsaert, 1982; Stull, 1988; Garratt, 1992):

$$\overline{u'w'}_0 = -C_m |\vec{U}_1| U_1, \quad [8]$$

$$\overline{v'w'}_0 = -C_m |\vec{U}_1| V_1, \quad [9]$$

$$\overline{w'\theta'}_0 = C_h |\vec{U}_1| (\theta_s - \theta_1), \quad [10]$$

$$\overline{w'q'}_0 = C_q |\vec{U}_1| (q_s - q_1), \quad [11]$$

where C_m is the transfer coefficient for momentum (drag coefficient), C_h is the transfer coefficient for heat, C_q is the transfer coefficient for moisture, $|\vec{U}_1|$ is the absolute horizontal wind speed at the lowest model level, θ_1 and q_1 are potential temperature and specific humidity at the lowest model level, and θ_s and q_s are potential temperature and specific humidity at the surface. (Over water θ_s and q_s are determined by the surface temperature and its saturation specific humidity (over the ocean humidity is only 98% of the saturation value due to salinity effects); over land, a surface scheme is needed to obtain θ_s and q_s .) In accordance with MO theory, the transfer coefficients can be written in terms of profile functions containing a logarithmic part, with roughness lengths as surface characteristics, and a stability function describing the effect of stability as a function of the Obukhov length:

$$C_m = \frac{k^2}{[\ln(z_1/z_{om}) - \Psi_m(z_1/L)]^2}, \quad [12]$$

$$C_h = \frac{k^2}{[\ln(z_1/z_{om}) - \Psi_m(z_1/L)][\ln(z_1/z_{oh}) - \Psi_h(z_1/L)]}, \quad [13]$$

$$C_q = \frac{k^2}{[\ln(z_1/z_{om}) - \Psi_m(z_1/L)][\ln(z_1/z_{oq}) - \Psi_h(z_1/L)]}, \quad [14]$$

where L is the Obukhov length ($= -u_*^3/(\overline{w'\theta'}_{vo} kg/\theta_v)$); k is the von Karman constant (0.4); $u_* = \{(\overline{u'w'}_0)^2 + (\overline{v'w'}_0)^2\}^{1/4}$ is the friction velocity; z_{om} , z_{oh} , and z_{oq} are the roughness lengths for momentum, heat, and moisture; z_1 is the height of the lowest model level (in many models between 10 and 30 m); and $\Psi_{m,h}$ are the stability functions for momentum and heat/moisture. Over land where z_1 is not necessarily large compared to z_{om} , a displacement height equal to z_{om} needs to be added to z_1 , but over the ocean this effect is negligible because $z_1 \gg z_{om}$. The transfer coefficients can also be written as

$$C_m = C_{mn} F_m(\text{Ri}_b, z_1/z_{om}, z_1/z_{oh}), \quad [15]$$

$$C_h = C_{hn} F_h(\text{Ri}_b, z_1/z_{om}, z_1/z_{oh}), \quad [16]$$

$$C_q = C_{qn} F_h(\text{Ri}_b, z_1/z_{om}, z_1/z_{oh}), \quad [17]$$

where the neutral transfer coefficients are defined by the logarithmic part of the profile functions

$$C_{mn} = \frac{k^2}{[\ln(z_1/z_{om})]^2}, \quad [18]$$

$$C_{hn} = \frac{k^2}{[\ln(z_1/z_{om})][\ln(z_1/z_{oh})]}, \quad [19]$$

$$C_{qn} = \frac{k^2}{[\ln(z_1/z_{om})][\ln(z_1/z_{oq})]}, \quad [20]$$

and F_m and F_h are the stability functions dependent on bulk Richardson number $\text{Ri}_b = (g/\theta_v) z_1 (\theta_{vs} - \theta_{v1}) / |\vec{U}_1|^2$, and the

ratios z_1/z_{om} and z_1/z_{oh} . It can easily be shown that Ri_b is related to z_1/L , so the functions F_m and F_h can be derived from the MO stability functions for which the main body of empirical information exists. In these expressions, the stability functions F_m and F_h also have a formal dependence on z_{oq} , which has been neglected here, because z_{oq} is often equal to z_{oh} and the stability dependence on moisture is weak.

An extension of the air-surface transfer framework has been to add a gustiness velocity in the definition of the absolute wind speed (Beljaars, 1995):

$$|\vec{U}_1| = (U_1^2 + V_1^2 + \beta w_*^2)^{1/2}, \quad w_* = \left(\overline{hw'\theta'}_{vo} g / \theta_v \right)^{1/3} \quad [21]$$

with h for boundary layer height, w_* is the free convection velocity scale, $\overline{hw'\theta'}_{vo}$ is the surface buoyancy flux, and β an empirical parameter. To include the effect of gustiness on the friction velocity, the friction velocity needs to be redefined as well

$$u_* = C_m^{1/2} (U_1^2 + V_1^2 + \beta w_*^2)^{1/2}. \quad [22]$$

Observational and model studies suggest β -values ranging from 0.8 to 1.25. The idea is that the horizontal wind (U_1, V_1) and the large-scale components of the horizontal wind fluctuations (scaling with w_*) control the air-surface coupling. Horizontal wind fluctuations do not obey MO similarity (they scale with h) and therefore this part of the air-surface transfer is an extension to MO theory. The gustiness effect is only important at low wind speeds since values of w_* vary typically between 0.5 and 1.5 $m\ s^{-1}$. However, it is necessary to include gustiness in the air-surface transfer to avoid the singularity at zero wind speed and to obtain a proper free convection limit.

The formalism described above leaves two aspects undefined namely; (1) the neutral transfer coefficients defined by roughness lengths, and (2) the stability functions. The stability functions have been subjects of a number of experimental programs and are reasonably well known for homogeneous terrain (Högström, 1988). Popular functions for unstable situations are

$$\Psi_m = 2 \ln \left\{ \frac{x+1}{2} \right\} + \ln \left\{ \frac{x^2+1}{2} \right\} - 2 \arctan(x) + \frac{\pi}{2}, \quad [23]$$

$$\Psi_h = 2 \ln \left\{ \frac{x^2+1}{2} \right\}, \quad \text{with } x = (1 - 16z/L)^{1/4}. \quad [24]$$

In the neutral limit with $|L| \rightarrow 0$, $\Psi_{m,h} \rightarrow 0$ (leaving just the logarithmic part of the transfer coefficients in eqns [12]–[14]) and $F_{m,h} \rightarrow 1$. For stable situations, the following functions are adequate for most applications (Beljaars and Holtslag, 1991)

$$-\Psi_m = a \frac{z}{L} + b \left\{ \frac{z}{L} - \frac{c}{d} \right\} \exp(-dz/L) + \frac{bc}{d}, \quad [25]$$

$$-\Psi_h = \left\{ 1 + \frac{2az}{3L} \right\}^{3/2} + b \left\{ \frac{z}{L} - \frac{c}{d} \right\} \exp(-dz/L) + \frac{bc}{d} - 1, \quad [26]$$

with $a = 1$, $b = 0.667$, $c = 5$, and $d = 0.35$. The evidence for having different Ψ_m and Ψ_h functions is very weak in the fully turbulent regime (i.e., $z/L < 1$), but necessary in very stable situations. With $\Psi_m = \Psi_h$, it can easily be shown that a critical Richardson number exists, beyond which no turbulence can exist. Practically, it is important that the conversion from Richardson number to z/L , which is necessary in large-scale models, does not become singular. The functions in eqns [25] and [26] guarantee such a regular behavior.

The last parameters that need specification in order to describe air-surface transfer are the roughness lengths for momentum, heat, and moisture. Roughness lengths are considered to be surface characteristics, but specifying them can be very difficult. The roughness lengths are as much part of the air-surface transfer problem as the transfer functions themselves.

For land, empirical tables are used that relate roughness lengths to vegetation cover and type (e.g., 0.03 m for smooth grass and 1–2 m for forest). However, heterogeneity in land cover has a big influence. Intermittent obstacles (e.g., lines of trees) can change a smooth surface into a surface that exerts a strong drag on the atmospheric flow. The standard way of handling surface heterogeneity in large-scale models is to define so-called effective roughness lengths. The effective roughness lengths are the ones that give the correct area-averaged flux in a large-scale model. The concept is justified by observations over heterogeneous terrain where far away from the surface a logarithmic profile is observed with a slope that corresponds to u_*/k , where the friction velocity is the square root of the *area-averaged* kinematic momentum flux at the surface (Mason, 1988). Far away means at such a distance from the surface that the internal boundary layers from the surface heterogeneities have blended. Another example of surface heterogeneity is subgrid orography, which exerts form drag on the flow. The effect of subgrid orography is often included in large-scale models through an enhanced roughness length. The orographic contribution to the ‘effective roughness length’ can be very large. In areas with orography, the roughness length for momentum can go up to a few hundred meters.

Much less is known about the roughness lengths for heat and moisture. For homogeneous low vegetation they tend to be an order of magnitude smaller than the roughness length for momentum, but the effect of terrain heterogeneity is less obvious. Indications are that the transfer coefficients for heat and moisture are not enhanced because the equivalent of pressure drag on obstacles (form drag) does not exist for heat and moisture, and therefore the roughness lengths for heat and moisture decrease when the roughness length for momentum has a heterogeneity or orographic enhancement. There are also indications that the roughness lengths for heat and moisture are not really surface properties but also depend on environmental parameters such as solar elevation, wind, and stability.

The notion that surface roughness lengths are an integral part of the air-surface transfer problem and the fact that roughness lengths over land can be rather uncertain have inspired other approaches. An example is the free convection transport theory for heat and moisture proposed by Stull (1994). It has transfer laws that do not depend on surface

characteristics, but the empirical coefficients are not the same for land and ocean.

Also, the use of an effective roughness for subgrid orography is rather unsatisfactory because the z_{om} values can become much larger than the height of the lowest model level. This makes the specification of roughness lengths for heat and moisture rather intractable. As an alternative for the effective roughness length concept, Beljaars et al. (2004) therefore proposed to implement turbulent orographic form drag as an additional tendency profile on model levels.

Roughness lengths over the ocean are much smaller than over land (between 10^{-5} and 10^{-3} m). Since ocean waves make the surface aerodynamically rough and since waves are wind driven, it is not surprising that the ocean roughness length for momentum scales predominantly with the friction velocity. A common expression is

$$z_{om} = C_{ch} u_*^2 / g + 0.11 \nu / u_*, \quad [27]$$

$$z_{oh} = 0.40 \nu / u_*, \quad [28]$$

$$z_{oq} = 0.62 \nu / u_*, \quad [29]$$

where the first term in eqn [27] represents the so-called Charnock relation for rough ocean waves with C_{ch} , the empirical Charnock parameter (common values range from 0.01 to 0.03). The second term with ν/u_* represents the scaling for molecular exchange, with ν for the kinematic viscosity of air ($1.5 \cdot 10^{-5} \text{ m}^2 \text{ s}^{-1}$). Neutral transfer coefficients in accordance with eqns [18], [20], [27], and [29] are given for the 10-m reference level in Figure 2 as a function of wind speed. There is a clear increase of C_{mn} with wind speed, whereas the transfer coefficient for moisture is fairly constant. Both transfer coefficients increase at low winds due to the ν/u_* scaling. The shown characteristics as a function of wind speed correspond very well with observations, although recent studies indicate that the transfer coefficients also depend on the state of the ocean waves often characterized by the wave age (Janssen, 2004).

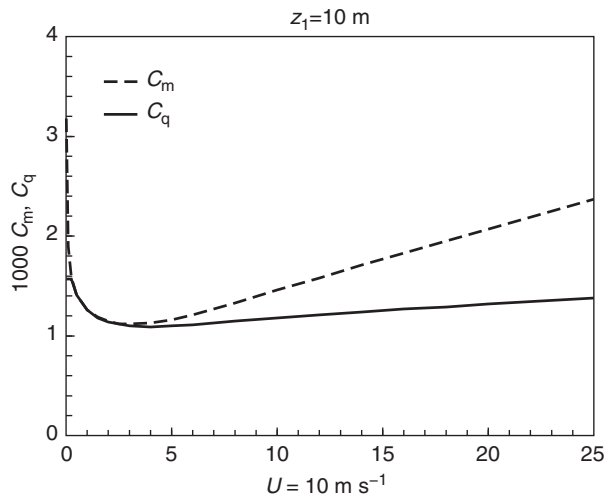


Figure 2 Neutral transfer coefficients for momentum (dashed) and moisture (solid) as a function of wind speed for the 10-m reference level.

Bulk or Slab Model for the Dry Mixed Layer

Turbulent mixing is strong in the unstable boundary layer due to heating from the surface and therefore the outer part of the boundary layer (BL) (above the surface layer) tends to be well mixed in conserved variables (particularly θ but also U , V , and q). An obvious way of describing the well-mixed boundary layer is by describing the evolution of the BL by its depth, its mixed layer values U_m , V_m , θ_m , and q_m , and the jump in these conserved variables at the top of the BL (Driedonks, 1982, Figure 3 illustrates this bulk model). Simple rate equations can be derived for potential temperature with a mixed layer value θ_m and a depth h :

$$h \frac{d\theta_m}{dt} = \overline{w'\theta'}_0 - \overline{w'\theta'}_h, \quad [30]$$

$$\frac{dh}{dt} = w_e, \quad [31]$$

$$\frac{d\Delta\theta}{dt} = \gamma w_e - \frac{d\theta_m}{dt}, \quad [32]$$

$$\overline{w'\theta'}_h = -w_e \Delta\theta, \quad [33]$$

where $\overline{w'\theta'}_0$ is the kinematic surface heat flux, w_e the entrainment velocity, $\overline{w'\theta'}_h$ the flux at the top of the mixed layer, $\Delta\theta$ the temperature jump at the top of the mixed layer, and γ the lapse rate above the boundary layer. The total derivatives in eqns [30]–[33] include advection. This is not a closed set of equations; a closure equation is needed for entrainment flux $\overline{w'\theta'}_h$ or $\overline{w'\theta'}_{vh}$. This term has been studied extensively with tank experiments and in the atmosphere. Its magnitude is related to the amount of turbulent kinetic energy that is produced in the mixed layer and the fraction that is diffused

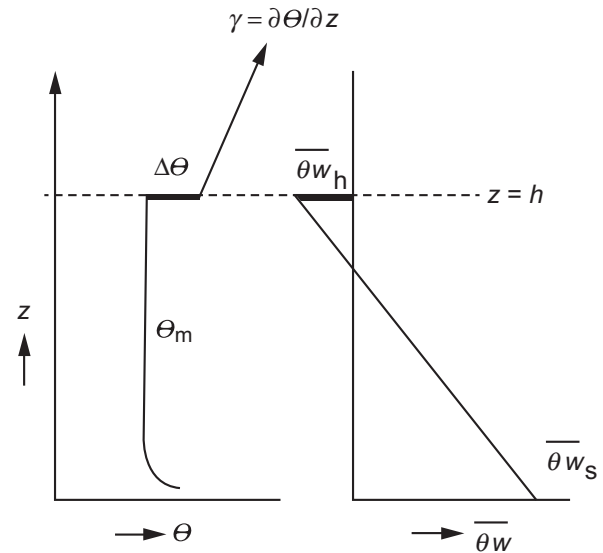


Figure 3 Schematic profile of potential temperature (θ) and kinematic heat flux profiles ($\overline{w'\theta'}$) in a simple slab model of the mixed-boundary layer.

into the inversion layer where it erodes the inversion. It is therefore plausible that this term scales with the total production of kinetic energy in the mixed layer. The following closure equation combines a number of production mechanisms in an empirical way

$$h(g/\theta_v)\overline{w'\theta'}_{vh} = B_1 u_*^3 + B_2 |\Delta \vec{U}|^2 w_e + B_3 w_*^3 + B_4 \sigma_w^3. \quad [34]$$

The empirical coefficients are by no means known accurately, but typical values are: $B_1 = 2.5$, $B_2 = 0.7$, $B_3 = 0.2$, and $B_4 = 1.5$. The first term represents the production of turbulence by shear in the mixed layer and scales with the surface friction velocity, the second one represents the production by shear in the inversion and scales with the velocity jump $\Delta \vec{U}$ over the inversion, the third term is often the dominant one as it represents the buoyancy production in the mixed layer, and the last term, scaling with the variance of the vertical velocity, represents the energy needed to spin up the calm air that is entrained into the mixed layer. Often in practical applications only the dominant third term is used, which corresponds to the well-known closure assumption $\overline{w'\theta'}_{vh} = 0.2\overline{w'\theta'}_{vo}$.

The advantage of bulk models is that they are simple and realistic (at least for well-mixed quantities), and do not rely on high vertical resolution. However, this type of model can be difficult to implement in a large-scale model as it involves a moving lower boundary (the top of the mixed layer). An alternative is to use the boundary layer height as a floating extra level to enhance the vertical resolution of a model at the inversion height where it is most needed. Furthermore, slab models have difficulties in handling less well-mixed boundary layers such as stable boundary layers, baroclinic boundary layers, and quantities that are not so well mixed such as wind and specific humidity.

Irrespective of whether mixed layer models are of practical importance, they are certainly of conceptual importance. The bulk model clearly illustrates the budget aspect of the well-mixed atmospheric boundary layer: the time evolution of the mixed-layer value is determined by the flux at the surface and the flux at the top. This is an aspect that should not be forgotten in more complex models: it is important to focus on the parametrization of the flux at the surface and on the flux in the inversion layer. The flux in between is automatically linear if sufficient mixing is present in the scheme.

K-Closure Based on Local Stability

A fairly simple closure for models that resolve the boundary layer explicitly is K -closure. This scheme (Louis, 1979) has become very popular in NWP and climate models. It follows MO similarity and in analogy with molecular diffusion, fluxes are assumed to be proportional to gradients

$$\overline{u'w'} = -K_m \frac{\partial U}{\partial z}, \quad \overline{v'w'} = -K_m \frac{\partial V}{\partial z}, \quad [35]$$

$$\overline{w'\theta'} = -K_h \frac{\partial \theta}{\partial z}, \quad \overline{w'q'} = -K_h \frac{\partial q}{\partial z}. \quad [36]$$

The exchange coefficients for momentum K_m and heat and moisture K_h are not a property of the fluid but depend on the

flow. Surface layer similarity provides expressions for K_m and K_h near the surface, which are

$$K_m = \frac{l^2}{\phi_m^2} \left| \frac{d\vec{U}}{dz} \right|, \quad K_h = \frac{l^2}{\phi_m \phi_h} \left| \frac{d\vec{U}}{dz} \right|. \quad [37]$$

The turbulence length scale $l = kz$ in the surface layer, with k for the von Kármán constant (0.4), and ϕ_m and ϕ_h are the MO stability functions for momentum and heat/moisture. This scheme is the differential form of what has been described above for the surface layer with $\phi_{m,h} = (1 - \eta \partial \Psi / \partial \eta)$, and $\eta = z/L$. Although these stability functions strictly apply only to the surface layer, they are also used above the surface layer, but the length scale has to be limited for large z , which is done by applying an asymptotic value λ in the following way (Blackadar, 1962)

$$\frac{1}{l} = \frac{1}{kz} + \frac{1}{\lambda}. \quad [38]$$

Typical values for λ are in the range of 30–300 m. This formulation works quite well for boundary layer turbulence but may be too active (dependent on the choice of stability functions) in the upper troposphere where clear-air turbulence can be significant. Therefore, some versions of the scheme gradually reduce the length scale in the upper troposphere by multiplying the length scale resulting from eqn [38] by, e.g., $\beta + (1 - \beta)/(1 + z/\lambda_a)$ with $\beta = 0.2$ and $\lambda_a = 4000$ m, which reduces the length scale to 20% of eqn [38] for heights well above 4000 m.

The expressions in eqn [37] contain the MO stability functions ϕ_m and ϕ_h , which are functions of l/kL (i.e., z/L in the surface layer), depend on fluxes. However, in a numerical model it is necessary to express the eddy diffusion coefficients in terms of model variables. The following relation between the gradient Richardson number and l/kL can be used

$$Ri = \frac{g}{\theta_v} \frac{d\theta_v/dz}{|d\vec{U}/dz|^2} = \frac{l}{kL} \frac{\phi_h}{\phi_m^2}. \quad [39]$$

The computational procedure is as follows: (1) compute the Ri number from the model variables, (2) solve eqn [39] for l/kL , and (3) substitute the result in eqn [37] to obtain the eddy diffusion coefficients. Depending on the precise form of the stability functions it may be necessary to solve eqn [39] iteratively or to store the relation between Ri and l/kL in a lookup table.

A numerically more efficient way of formulating this scheme is to put the empirical stability information in functions of the Richardson number (Louis, 1979). The eddy diffusion coefficients are expressed now in the following way

$$K_m = l^2 F_m \left| \frac{d\vec{U}}{dz} \right|, \quad K_h = l^2 F_h \left| \frac{d\vec{U}}{dz} \right|, \quad [40]$$

where F_m and F_h are empirical functions of the Richardson number and by definition $F_m = \phi_m^{-2}$ and $F_h = (\phi_m \phi_h)^{-1}$. Equation [40] is actually the closure equation that is used in many NWP and climate models. This formulation is known as the ‘Louis scheme.’ It is fully consistent with MO similarity in the surface layer and also with the concept of ‘local scaling’

for the stable boundary layer (Nieuwstadt, 1984). Typical stability functions are

$$\begin{aligned} \text{Ri} < 0 : F_m &= 1 - \frac{2b\text{Ri}}{1 + c_h(-\text{Ri})^{1/2}}, \\ F_h &= 1 - \frac{3b\text{Ri}}{1 + c_h(-\text{Ri})^{1/2}}, \end{aligned} \quad [41]$$

$$\begin{aligned} \text{Ri} > 0 : F_m &= \frac{1}{1 + 2b(1 + d\text{Ri})^{-1/2}}, \\ F_h &= \frac{1}{1 + 3b(1 + d\text{Ri})^{1/2}}, \end{aligned} \quad [42]$$

$$c_h = l^2/(3z^2), \quad b = 5, \quad c = 5, \quad d = 6. \quad [43]$$

An unsatisfactory aspect of the Louis scheme is that the stability functions $F_m = F_m(\text{Ri})$ and $F_h = F_h(\text{Ri})$ in NWP and climate models are inspired by large-scale model performance

rather than by observational studies. For the unstable boundary layer, the results are not very sensitive to the precise form of the stability functions, because the diffusion coefficients are large and the result will always be a well-mixed layer. However, in the stable boundary layer, results are very sensitive to the stability functions. Figure 4 shows two versions of functions that have been used in the European Centre for Medium-range Weather Forecasts (ECMWF) operational weather prediction model (Louis, Tiedtke, and Geleyn (LTG) after Louis et al., 1982; and a revised version of LTG) compared with functions that are observationally based for stable situations (Beljaars and Viterbo, 1998). The stability functions that are typically used in numerical models have much more mixing at high Richardson numbers than observed, but application of the observed MO functions leads to too stable near-surface layers as the result of a positive feedback with the land surface temperatures. The reason for the poor performance of observationally based functions in numerical models is a matter of speculation and subject of ongoing research. A number of reasons have been suggested (1) gravity waves at the top of the boundary layer

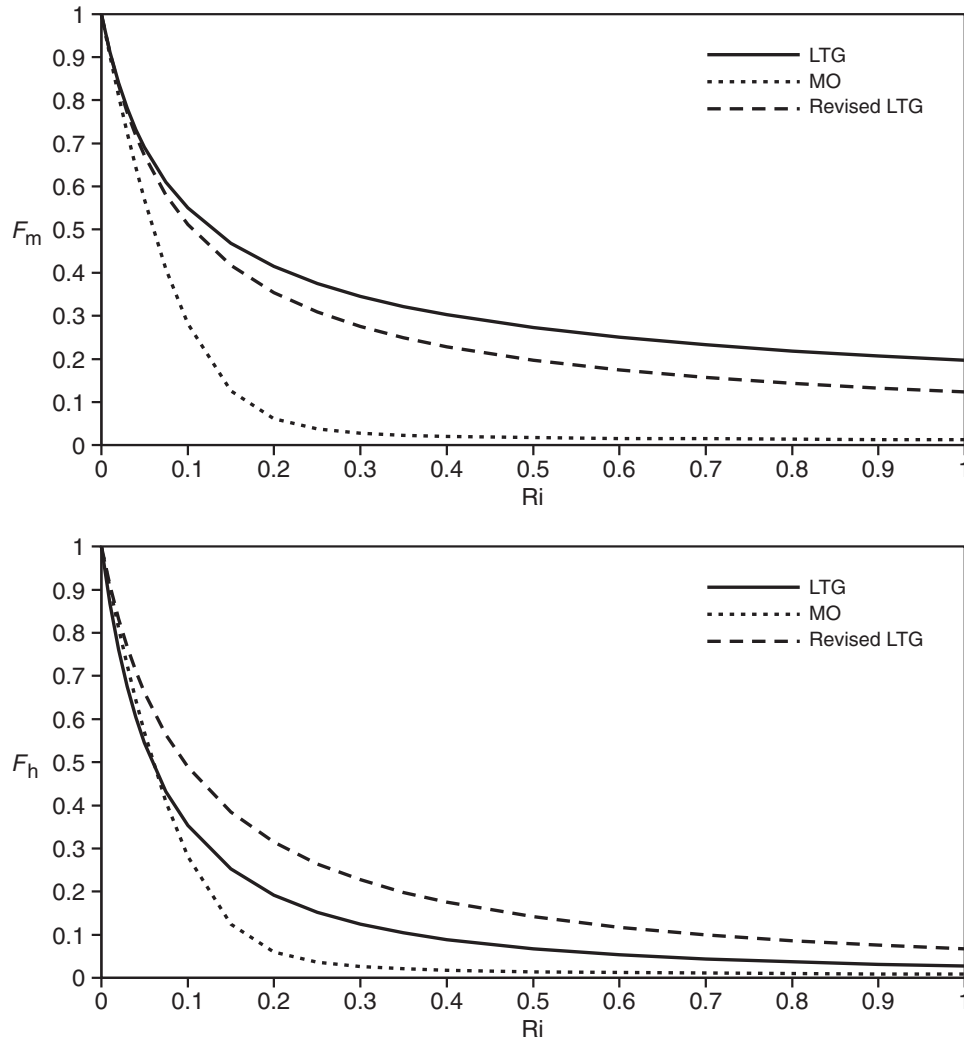


Figure 4 Stability functions for momentum and heat for the stable boundary layer as a function of the Richardson number. Three versions are shown: LTG and revised LTG (both of which have been used in the ECMWF model), and one that is derived from observationally based MO functions.

propagating into the stable boundary layer might break and cause extradiusion, (2) if the nighttime cooling is not distributed uniformly in the horizontal, mesoscale motion can maintain a higher level of mixing, and (3) effects of heterogeneity average nonlinearly and enhance the mixing. More research is obviously needed to clarify the missing physics and to find new formulations that are adequate for large-scale models. Currently, there are some interesting developments suggesting that nonlocal aspects play a role. The idea is that the stability above the stable boundary layers affects the mixing in the stable boundary layer by gravity wave interaction (Zilitinkevich, 2002).

Other unsatisfactory aspects of the Louis scheme are that for the mixed layer the fluxes can be upgradient (countergradient effects), and that entrainment at the top of the mixed layer is virtually zero with this formulation.

K-Profile Closure

An interesting closure which has been proposed by Troen and Mahrt (1986) for low-resolution models, as an alternative to the Louis scheme is the K-profile closure. In this scheme, the eddy diffusivity concept is used (eqns [35] and [36]), but the diffusion coefficients are specified as an integral profile for the entire boundary layer. The expressions for the stable boundary layer and for the unstable surface layer ($z/h < 0.1$) are

$$K_m = kw_s z \left(1 - \frac{z}{h}\right)^2, \quad [44]$$

$$K_h = kw_s z \left(1 - \frac{z}{h}\right)^2 Pr^{-1}, \quad [45]$$

$$w_s = u_* / \phi_h, \quad Pr = \phi_h / \phi_m \quad [46]$$

In the unstable outer layer, the same expressions are used except that w_s and Pr are evaluated at $z = 0.1h$ and that a countergradient term is introduced in the heat and moisture equations (the countergradient term is an extension of the eddy diffusivity concept and reflects the effect of large eddies in the convective boundary layer that cause transport irrespective of the local gradient):

$$\overline{w'\theta'} = -K_h \left(\frac{\partial \theta}{\partial z} - \gamma_\theta \right), \quad \overline{w'q'} = -K_h \left(\frac{\partial q}{\partial z} - \gamma_q \right), \quad [47]$$

$$\gamma_\theta = C \frac{(\overline{w'\theta'})_0}{w_s h}, \quad \gamma_q = C \frac{(w'q')_0}{w_s h}, \quad C = 8. \quad [48]$$

If the stability functions obey free convection scaling, this closure will automatically provide a velocity scale $w_s = (u_*^3 + C_1 w_*^3)^{1/3}$, with $C_1 = 0.6$ and $w_* = (h \overline{w'\theta'_{vo}} g / \theta_v)^{1/3}$. An essential part of this scheme is the diagnosis of the boundary layer top h , for which the following expression is used

$$h = \frac{Ri_c (U_h^2 + V_h^2)}{(g/\theta_v)(\theta_{vh} - \theta_{vs})} \quad [49]$$

Parameter Ri_c is a critical Richardson number with a numerical value of about 0.5, θ_{vs} is a virtual potential temperature somewhere in the surface layer (say 10 m above

the surface), and θ_{vh} is the virtual potential temperature at height h . For the convective boundary layer, the temperature θ_{vs} is augmented with the temperature excess that thermals have with respect to their surroundings:

$$\theta_{vs} = \theta_v + D(\overline{w'\theta'_v})_0 / w_s, \quad D = 6.5. \quad [50]$$

Different versions of this scheme are in use (Holtslag and Boville, 1993; Beljaars and Viterbo, 1998). The original scheme as described here leads to a too rapid boundary layer growth and therefore different modifications have been proposed. One of them is not to use the absolute wind speed at boundary layer height but to take the difference between the wind at level h and the wind at the 10-m level (Vogelezang and Holtslag, 1996).

The boundary layer top entrainment is not very well controlled in the original K-profile scheme and depends on the vertical distribution of levels. Therefore, it has been proposed not to use the K-profile in the inversion, but to parametrize the diffusion in the entrainment layer by imposing a virtual temperature flux of 20% of the surface value (as is done in bulk models).

The K-profile has a few major advantages. The first is its robustness from the numerical point of view. The reason is that no oscillations in K-profile can occur because it is specified (the Louis scheme has the well-known difficulty that it can easily lead to nonlinear instabilities for long time steps; see Kalnay and Kanamitsu, 1988; Beljaars, 1992). The second advantage is that the nonlocal aspects of the scheme allow for entrainment at the boundary layer top. In the original formulation, the entrainment is highly dependent on the vertical discretization, but imposing an entrainment flux cures the problem reasonably well. Finally, the scheme has a simple way of representing the countergradient effects, which are physically realistic although it is not entirely clear yet how important they are in real applications.

Figure 5 illustrates the effect of boundary layer top entrainment on the mixed layer moisture budget. Two versions of the ECMWF model are compared with observations in short range forecasts. The first version uses the Louis scheme (with LTG functions), but as discussed in the previous section, this scheme has no entrainment at all. The second scheme uses K-profile closure with an imposed entrainment rate of 20%. It is clear that the entrainment of dry air from above the boundary layer leads to a different moisture balance in the mixed layer and to a better correspondence with observations.

Higher Order Closure

The closure problem can be taken to another level by writing equations for the second moments. The resulting equations have triple correlations and pressure velocity correlations that are unknown and need closure assumptions. Many different formulations have been proposed; most of them are truncated versions of the full second-order equations. A good overview is given by Mellor and Yamada (1982) and a number of implementations in large-scale models exist. The simplest 'higher' order closure is the one in which the turbulence

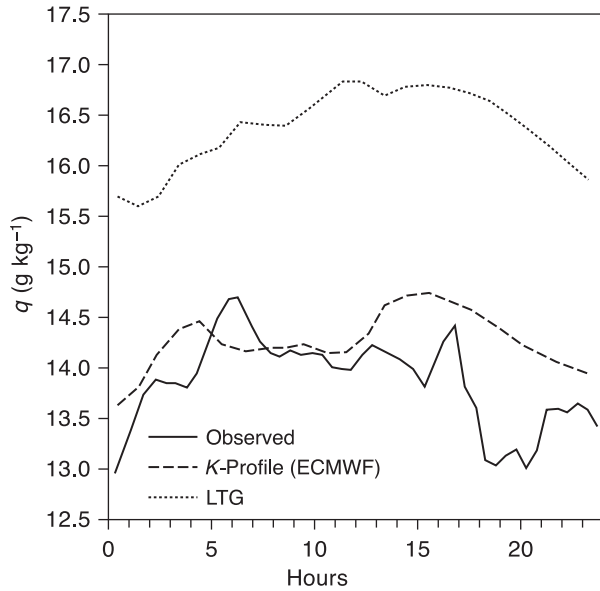


Figure 5 Composite of nine diurnal cycles of First ISLSCP (International Satellite Land Surface Climatology Project) Field Experiment (FIFE) observations of specific humidity in August 1987 in comparison with 0–24 h forecasts from two versions of the ECMWF model (the LTG scheme and the K -profile scheme with a specified entrainment of 20% at the boundary layer top).

kinetic energy equation is used to estimate the velocity scale in the turbulent diffusion coefficients for momentum, heat, and moisture.

$$\frac{dE}{dt} = -\overline{u'w'} \frac{\partial U}{\partial z} - \overline{v'w'} \frac{\partial V}{\partial z} - \frac{g}{\theta_v} \overline{w'\theta'_v} - \frac{\partial}{\partial z} \left(\overline{E'w'} + \overline{p'w'}/\rho \right) - \varepsilon, \quad [51]$$

$$K_m = c_m l_m E^{1/2}, \quad K_h = c_h l_h E^{1/2}. \quad [52]$$

At the right-hand side of eqn [51], the first two terms represent the mechanical production of turbulence kinetic energy by shear, the third term is the buoyancy term which is a source term in unstable situations and a sink term in stably stratified flow, the fourth is the diffusion of turbulent energy by pressure fluctuations and by turbulence (it cancels out when vertically integrated), and the last term (ε) is the dissipation of turbulence kinetic energy by molecular friction (conversion into heat). Parameters c_m and c_h are empirical constants.

Before the energy equation can be used it needs closure equations for the diffusion term and for the dissipation term. Also the length scales need to be specified. The diffusion term can be closed by making the energy diffusion proportional to the gradient of the energy profile (K -diffusion) and by selecting a diffusion coefficient proportional to the ones for momentum or heat. The dissipation term is always parametrized in terms of the velocity and length scales of the energy containing eddies (i.e., $\sim E^{3/2}/l$) following the idea that the large eddies control the amount of energy that is transferred on the energy cascade toward the dissipating scales. The main advantage of the turbulent kinetic energy approach is that it has a natural way of generating boundary layer top entrainment. The energy produced by shear and buoyancy (mainly in the bulk of the boundary layer) is transported upward by the diffusion term into the inversion layer where it is absorbed by the negative buoyancy term (see Figure 6 for an illustration of the energy budget in the convective boundary layer).

Many different formulations exist for the length scale, varying from simple empirical functions of, e.g., height above the surface, stability, and boundary layer depth, to full prognostic equations. A popular alternative to the prognostic equation for the length scale is a rate equation for the dissipation rate with the length scale derived from $l \sim E^{3/2}/\varepsilon$. The advantage is that the length scale is a parameter that varies with environmental conditions (e.g., stability). However, parametrization of the unknown terms in the dissipation as

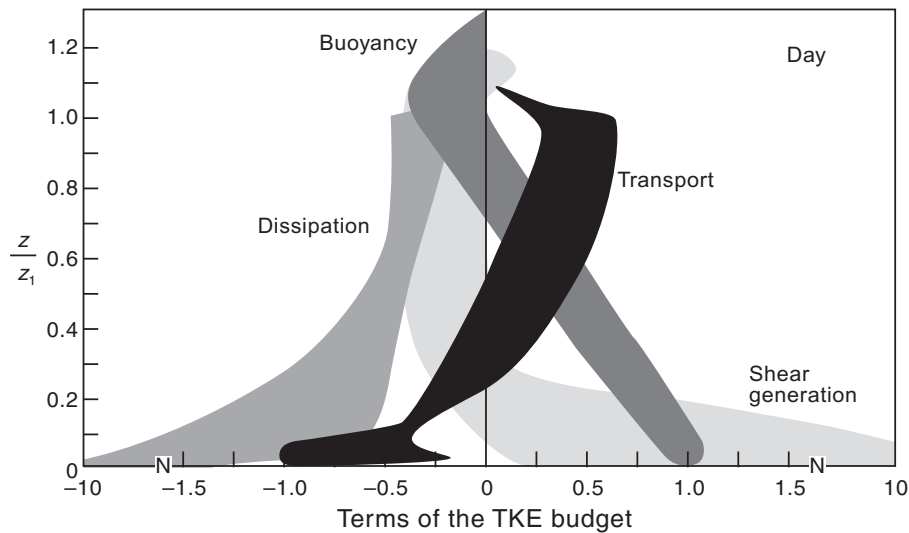


Figure 6 Illustration of the different terms in the kinetic energy budget in the unstable boundary layer. After Stull, R.B., 1988. *An Introduction to Boundary Layer Meteorology*. Kluwer, Dordrecht, Holland.

well as length scale equations is very difficult and data to support such parametrizations are virtually nonexistent.

Simplification of the full second-order equations to algebraic equations is an attractive alternative compared to the full equations. The eddy diffusivity terms and the countergradient terms come out rather naturally. Many flavors of such approximations exist and it is beyond the scope of this article to discuss them. A good overview of different levels of approximation is given in Mellor and Yamada (1982) and Holtslag and Duynkerke (1998).

The level of complexity that is needed depends very much on the application. Important aspects in NWP and climate models are entrainment, stable boundary layer diffusion, interactions with the cloud scheme, and perhaps countergradient effects. In a simulation of the internal boundary layer after a step change in surface roughness, simple models work well for the wind profiles, but energy and length scale (or ϵ) equations are needed to simulate the perturbed relations between fluxes and profiles. If one is interested in the evolution of the flow over a low hill, simple closure may do, but for the realistic simulation of the stress, it is necessary to use an advection equation for the stress (Beljaars et al., 1987).

Cloudy Boundary Layer

Boundary layer clouds have a strong impact on the boundary layer structure. The latent heat release due to condensation affects the buoyancy generation of turbulence, and the radiative effects of the clouds can change the energy budget of the boundary layer completely. Mixed layers that are driven by surface heating in the dry case can equally be maintained by radiative cooling at the cloud top as in a stratocumulus-topped boundary layer. Fog also becomes well mixed due to radiative cooling at the top as soon it is sufficiently thick. Shallow cumuli (e.g., in the trades) are very important for the drying of the boundary layer in the subtropics (boundary layer ventilation; Tiedtke, 1988) and therefore have a strong influence on the hydrological cycle. Finally, clouds are an important forecast

product in NWP and are an important modulator of the atmospheric radiative budget.

Standard turbulence schemes can be extended rather easily to situations with full cloud cover. An example is stratocumulus or fog. In this case, it is necessary to formulate the diffusion in terms of moist conserved variables (e.g., θ_1 and q_1) and to express the buoyancy in these variables. The latter is the main complication as due to the latent heat release in clouds and the water loading the formulation changes at the cloud base (see eqns [6] and [7] and Stull, 1988 for a discussion). A stratocumulus layer can be well mixed in the sense that the moist conserved variables are fairly uniform from the subcloud layer to the cloud top as the result of mixing through the entire boundary layer mainly driven by cloud top cooling. The key aspect in the modeling of a stratocumulus layer is the cloud top entrainment, which is not well understood. However, simple empirical relations exist (Lock, 1998). They couple the buoyancy flux in the inversion layer to the cloud top radiative cooling and the inversion strength (see Duynkerke, 1998; Lock, 1998) rather than to the surface buoyancy flux as in the dry mixed layer. That the boundary layer top entrainment is highly relevant in large-scale models is clear from the experience that models fill up large areas over the ocean with clouds when entrainment is not present or too small, and that stratocumulus disappears altogether when the top entrainment is too strong.

Shortwave radiation is another important factor in the evolution of a stratocumulus layer. In contrast to longwave radiation, which acts over a very shallow layer near the cloud top, solar radiation penetrates much deeper and therefore has the effect of enhancing buoyancy and decoupling the cloud layer from the subcloud layer. This can lead to breakup of the cloud due to two mechanisms. First, a stable layer near cloud base can be created which shuts off the moisture supply from the surface. Second, the enhancement of entrainment leads to additional drying and warming of the cloud layer. Proper modeling of the effects of solar radiation, the diurnal cycle, and decoupling is still very difficult.

Figure 7 illustrates the energy budget in a stratocumulus layer from a simulation with a turbulence energy equation

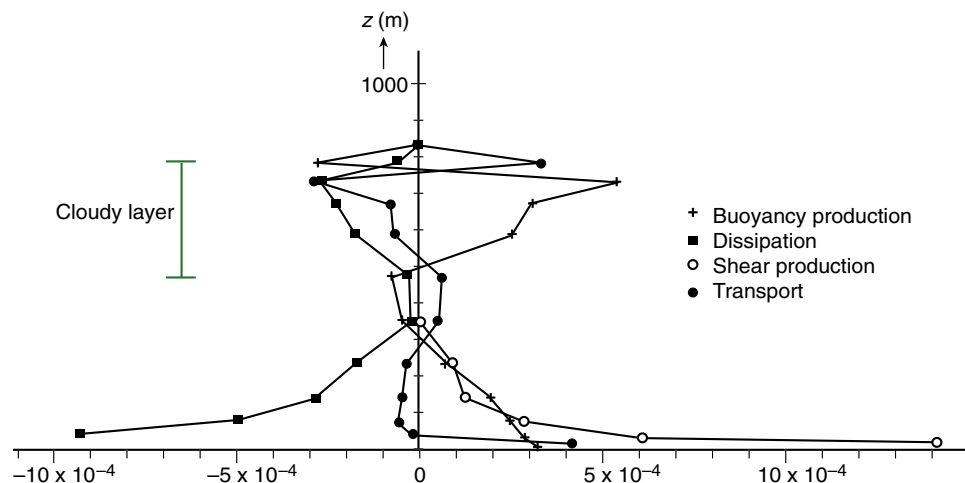


Figure 7 Profile of the different terms in the turbulence kinetic energy budget from a simulation of a stratocumulus layer by Duynkerke, P.G., Driedonks, A.G.M., 1987. A model for the turbulent structure of the stratocumulus-topped atmospheric boundary layer. *Journal of Atmospheric Sciences* 44, 43–64.

and a dissipation equation. The advantage is that no velocity and length scales need to be selected as they are both dynamically determined by the two equations. It is clear from the figure that the structure of the energy production is very complex with a sharp jump in the buoyancy term at cloud base. The diffusion term (labeled 'transport' in Figure 7) transports energy from the areas with production to stably stratified areas, e.g., the inversion. Dependent on the fluxes at the surface and at the top, the buoyancy minimum near cloud base can be negative, i.e., turbulence is damped here and so-called decoupling can occur. It is obvious that bulk mixed layer models, which can be used for a stratocumulus layer, break down when decoupling occurs.

Much more complex is the situation where partial cloudiness occurs as in fair weather cumulus or trade cumulus. Cumulus clouds are the result of condensation of the most powerful updrafts in the subcloud layer, and play an active role through the latent heat release in the cloud. There are two basic ways of handling the effects of shallow cumulus on the mixing: (1) to have a separate scheme to represent the mixing in the cloud layer in addition to a dry boundary layer scheme (e.g., Tiedtke, 1988), and (2) to have a moist turbulence model with a partial condensation scheme for the clouds (Cuijpers and Bechtold, 1995; Tompkins, 2002).

The idea of having a shallow convection scheme separate from turbulent diffusion is inspired by the observation that the processes are rather different. Although moist updrafts originate in the dry boundary layer, only some of the dry thermals make it into a cloud. Furthermore, analysis of turbulence generated by large eddy models suggests that transport in a shallow convection layer is most suitable for a mass flux scheme, whereas turbulent diffusion is more difficult to represent by mass flux transport only. The disadvantages of such a process splitting are that it creates difficulties in the transition from one process to the other at cloud base and that the transition to stratocumulus is not very well handled. In the latter case, a moist boundary layer scheme would be needed but such a scheme overlaps too much with a shallow convection mass flux scheme. Improving cloud and convection schemes and their coupling is an active area of research. Implementations have been made with intelligent switching between schemes (Brown et al., 2008) and by making shallow convection and boundary layer diffusion part of the same scheme through the so-called eddy-diffusion/mass flux approach (Siebesma et al., 2007; Köhler et al., 2011; Neggers et al., 2009).

The second way of handling clouds in the boundary layer is to use information on the fluctuations in T and q to decide about the cloud cover (as illustrated in Figure 8). The turbulent diffusion can be formulated in terms of moist variables and the partial cloudiness can be diagnosed making use of distribution functions for temperature and total water. Research has focused on the shape of the distribution functions and on how to couple them to turbulence schemes (Smith, 1990; Tompkins, 2002; Bechtold et al., 1995). Experience with large-scale models is that turbulence schemes tend to underestimate the variances (e.g., in T and q), which are needed to support a partial condensation scheme (see Smith, 1990). A possible explanation is that mesoscale variability has a nonnegligible effect on cloud fraction.

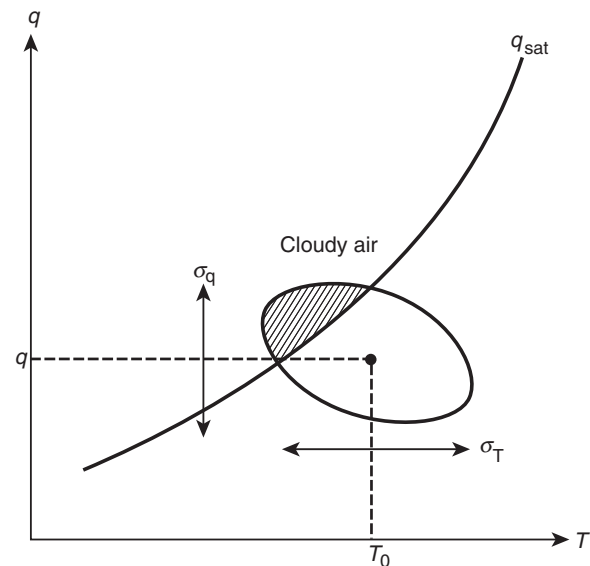


Figure 8 Illustration of how information on the variability of q and T can be used in a partial condensation scheme to estimate cloud fraction.

See also: **Air Sea Interactions:** Momentum, Heat, and Vapor Fluxes; Sea Surface Temperature. **Boundary Layer (Atmospheric) and Air Pollution:** Complex Terrain; Convective Boundary Layer; Stably Stratified Boundary Layer; Surface Layer. **Climate and Climate Change:** Energy Balance Climate Models. **Land-Atmosphere Interactions:** Overview. **Radiation Transfer in the Atmosphere:** Cloud-Radiative Processes.

References

- Bechtold, P., Cuijpers, J.W.M., Mascart, P., Trouilhet, P., 1995. Modelling of trade wind cumuli with a low-order turbulence model: toward a unified description of Cu and Sc clouds in meteorological models. *Journal of Atmospheric Sciences* 52, 455–463.
- Beljaars, A.C.M., 1992. Numerical schemes for parametrizations. In: *ECMWF Seminar Proceedings 9–13 September 1991. Numerical Methods in Atmospheric Models*, vol. II, pp. 1–42.
- Beljaars, A.C.M., 1995. The parametrization of surface fluxes in large scale models under free convection. *Quarterly Journal of the Royal Meteorological Society* 121, 255–270.
- Beljaars, A.C.M., Holtslag, A.A.M., 1991. Flux parametrization over land surfaces for atmospheric models. *Journal of Applied Meteorology* 30, 327–341.
- Beljaars, A.C.M., Viterbo, P., 1998. The role of the boundary layer in a numerical weather prediction model. *Royal Netherlands Academy of Arts and Sciences*. In: Holtslag, A.A.M., Duynkerke, P. (Eds.), *Clear and Cloudy Boundary Layers*. North Holland Publishers, Amsterdam, pp. 287–304.
- Beljaars, A.C.M., Walmsley, J.L., Taylor, P.A., 1987. A mixed spectral finite difference model for neutrally stratified boundary-layer flow over roughness changes and topography. *Boundary-Layer Meteorology* 38, 273–303.
- Beljaars, A.C.M., Brown, A.R., Wood, N., 2004. A new parametrization of turbulent orographic form drag. *Quarterly Journal of the Royal Meteorological Society* 130, 1327–1347.
- Blackadar, A.K., 1962. The vertical distribution of wind and turbulent exchange in a neutral atmosphere. *Journal of Geophysical Research* 67, 3095–3102.
- Brown, A.R., Beare, R.J., Edwards, J.M., Lock, A.P., Keogh, S.J., Milton, S.F., Walters, D.N., 2008. Upgrades to the boundary-layer scheme in the Met Office numerical weather prediction model. *Boundary-Layer Meteorology* 128, 117–132.

- Cuijpers, J.W.M., Bechtold, P., 1995. A simple parameterization of cloud water related variables for use in boundary layer models. *Journal of Atmospheric Sciences* 52, 2486–2490.
- Driedonks, A.G.M., 1982. Models and observations of the growth of the atmospheric boundary layer. *Boundary-Layer Meteorology* 23, 283–306.
- Duynkerke, P.G., 1998. About entrainment? Royal Netherlands Academy of Arts and Sciences. In: Holtslag, A.A.M., Duynkerke, P.G. (Eds.), *Clear and Cloudy Boundary Layers*. North Holland Publishers, Amsterdam, pp. 287–304.
- Duynkerke, P.G., Driedonks, A.G.M., 1987. A model for the turbulent structure of the stratocumulus-topped atmospheric boundary layer. *Journal of Atmospheric Sciences* 44, 43–64.
- Högström, U., 1988. Non-dimensional wind and temperature profiles in the atmospheric surface layer: a re-evaluation. *Boundary-Layer Meteorology* 42, 55–78.
- Holtslag, A.A.M., Boville, B.A., 1993. Local versus nonlocal boundary-layer diffusion in a global climate model. *Journal of Climate* 6, 1825–1842.
- Janssen, P., 2004. *The Interaction of Ocean Waves and Wind*. Cambridge University Press, Cambridge, UK; New York.
- Kalnay, E., Kanamitsu, M., 1988. Time schemes for strongly nonlinear damping equations. *Monthly Weather Review* 116, 1945–1958.
- Köhler, M., Ahlgrim, M., Beljaars, A., 2011. Unified treatment of dry convective and stratocumulus-topped boundary layers in the ECMWF model. *Quarterly Journal of the Royal Meteorological Society* 137, 43–57.
- Lock, A.P., 1998. The parametrization of entrainment in cloudy boundary layers. *Quarterly Journal of the Royal Meteorological Society* 124, 2729–2753.
- Louis, J.F., 1979. A parametric model of vertical eddy fluxes in the atmosphere. *Boundary-Layer Meteorology* 17, 187–202.
- Louis, J.F., Tiedtke, M., Geleyn, J.-F., 1982. A short history of the operational PBL-parameterization at ECMWF. Workshop on Boundary Layer Parameterization, November 1981. ECMWF, Reading, England.
- Mason, P.J., 1988. The formation of areally-averaged roughness lengths. *Quarterly Journal of the Royal Meteorological Society* 114, 399–420.
- Mellor, G.L., Yamada, T., 1982. Development of a turbulence closure model for geophysical fluid problems. *Reviews of Geophysics and Space Physics* 20, 851–875.
- Neggers, R.A., Köhler, M., Beljaars, A.C.M., 2009. A dual mass flux framework for boundary layer convection, Part I: Transport. *Journal of Atmospheric Sciences* 66, 1465–1487.
- Nieuwstadt, F.T.M., 1984. The turbulent structure of the stable, nocturnal boundary layer. *Journal of Atmospheric Sciences* 41, 2202–2216.
- Siebesma, A.P., Soares, P.M.M., Teixeira, J., 2007. A combined eddy-diffusivity mass-flux approach for the convective boundary layer. *Journal of Atmospheric Sciences* 64, 1230–1248.
- Smith, R.N.B., 1990. A scheme for predicting layer clouds and their water content in a general circulation model. *Quarterly Journal of the Royal Meteorological Society* 116, 435–460.
- Stull, R.B., 1988. *An Introduction to Boundary Layer Meteorology*. Kluwer, Dordrecht, Holland.
- Stull, R.B., 1994. A convective transport theory for surface fluxes. *Journal of Atmospheric Sciences* 51, 3–22.
- Tiedtke, M., 1988. Parameterization of cumulus convection in large-scale models. In: Schlesinger, M.E. (Ed.), *Physically-Based Modelling and Simulation of Climate and Climatic Change – Part I*. Kluwer Academic Publishers, Dordrecht, pp. 375–431.
- Tompkins, A.M., 2002. A prognostic parameterization for the sub-grid scale variability of water vapor and clouds in large-scale models and its use to diagnose cloud cover. *Journal of Atmospheric Sciences* 59, 1917–1942.
- Troen, I., Mahrt, L., 1986. A simple model of the atmospheric boundary layer; sensitivity to surface evaporation. *Boundary-Layer Meteorology* 37, 129–148.
- Vogelezang, D.H.P., Holtslag, A.A.M., 1996. Evaluation and model impacts of alternative boundary-layer height formulations. *Boundary-Layer Meteorology* 81, 245–269.
- Zilitinkevich, J.S., 2002. Third-order transport due to internal waves and non-local turbulence in the stably stratified surface layer. *Quarterly Journal of the Royal Meteorological Society* 128, 913–925.

Further Reading

- Beljaars, A.C.M., 1995. The impact of some aspects of the boundary layer scheme in the ECMWF model, ECMWF Seminar Proceedings on: *Parameterization of Sub-Grid Scale Physical Processes*, September 1994, Reading.
- Brutsaert, W.A., 1982. *Evaporation into the Atmosphere*. Reidel, Dordrecht, Holland.
- Garratt, J.R., 1992. *The Atmospheric Boundary Layer*. Cambridge University Press, New York.
- Holtslag, A.A.M., Duynkerke, P.G. (Eds.), 1998. *Clear and Cloudy Boundary Layers*. North Holland Publishers, Amsterdam.
- Workshop Proceedings, 1993. *Parameterization of the Cloud Topped Boundary Layer*, available from ECMWF, Shinfield Park, Reading, England.

Spectral Models

F Baer, University of Maryland, College Park, MD, USA

© 2003 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 5, pp 2099–2107, © 2003, Elsevier Ltd.

Introduction

With the advent of digital computers, weather forecasting was cast as a computational problem based on the fundamental prediction equations of fluids. Since analytic solutions are unavailable, approximations evolved to convert the differential equations to numerical equations suitable for computation on large computing machines. From this perspective, the concept of modeling was conceived. Thus weather forecasting – and more recently climate prediction – is approached by defining a numerical ‘model,’ and solutions to this model are sought. A variety of models have been developed over time to meet this goal, and the ‘spectral model’ is one of these.

The atmosphere is represented by variables describing molecular composites of its gases; the primary variables are velocity, temperature, density, water content in all phases, and aerosols. These variables are considered to be distributed continuously in three-dimensional space and to vary with time. The evolution of these variables in time may be determined at each point in space (the Eulerian method) or by following the particles through time (the Lagrangian method) and both methods are in use. The differential equations defining the future state of the variables are based on physical and dynamical principles, some well-known and others under study. These principles include the equations of motion (the Navier–Stokes equations), conservation of mass, an equation for change in entropy, equations for changes in water substance in its various phases, and chemical equations for changes of aerosols.

To define the notation of this article, these equations are presented below (*see Dynamical Meteorology: Primitive Equations*). Using the Eulerian reference, the time derivative is taken locally at each point in the fluid. The motion of the fluid is determined by an equation for the vector velocity $\mathbf{V}$ relative to the rotating Earth in all three space dimensions (eqn [1]).

$$\frac{\partial \mathbf{V}}{\partial t} = -(\mathbf{V} \cdot \nabla) \mathbf{V} - 2\Omega \times \mathbf{V} - \frac{1}{\rho} \nabla p - g\mathbf{k} + \mathbf{F} \quad [1]$$

Here, Ω is the angular velocity of the Earth; ρ and p are density and pressure, respectively, at each atmospheric point; g is the gravitational acceleration in the $\mathbf{k}$ (unit vertical vector) direction; and $\mathbf{F}$ comprises all frictional forces per unit mass. Conservation of mass is represented by the equation of continuity and the system thermodynamics are described by changes in entropy as in eqns [2]–[4].

$$\frac{\partial \rho}{\partial t} = -\nabla \cdot \rho \mathbf{V} \quad [2]$$

$$\frac{\partial s}{\partial t} = -\mathbf{V} \cdot \nabla s + \frac{1}{T} q(q_v, q_i, q_l, a_j \dots); \quad 1 \leq j \leq J \quad [3]$$

$$\frac{\partial q_k}{\partial t} = Q_k \quad [4]$$

In eqn [3], s represents specific entropy, q is the rate of heating per unit mass, and T is the temperature. Additionally, q depends on the heating rates associated with water vapor (q_v), ice (q_i), liquid water (q_l), aerosols (a_j for $j = j_1$), and other factors such as radiation. Each of the variables q_j and a_j has its own prediction (eqn [4]), where the Q_k represent complex parametric formulae relating some or all of the dependent variables. This entire system of equations constitutes the basis for selecting the ‘model’ that is integrated in time to predict the future state of the fluid. Additional features needed to complete the model are boundary conditions, initial conditions, scale truncation, external forces, and computational resources.

The final step in constructing a model is to select a technique to convert the basic nonlinear differential equations that describe the forecast system (eqns [1]–[4]) into a numerical form suitable for computation and integration on a digital computer. Finite differencing in both the time and space dimensions was the first method attempted. Since the vertical and horizontal dimensions in the atmosphere have unique properties, they may be and often are considered separately. Given that at any given height in the atmosphere a closed spherical surface exists on which the dependent variables describing the fluid are prescribed and predicted, the spectral method, which assigns a set of known continuous orthogonal functions over the domain to represent these variables, may be applied. When all the variables are described in this way, the resulting equations are integrated over the global domain, leading to a set of ordinary nonlinear differential equations in time and on each vertical level. Concurrently, differentiation in the vertical space coordinate and time is normally, but not universally, transformed to finite differences. The spectral method is most appropriate for the larger space scales since the functions usually used are global. However, regional models can be cast in the spectral framework if the boundary conditions are suitable. Alternative methods that have been applied include finite elements and spherical geodesic grids.

In comparison with other modeling techniques, the spectral method has no pole problem; its resolution is essentially homogeneous and isotropic; it allows for a simple solution of the Helmholtz equation in various settings; and, with an appropriate choice of the transform grid, it produces nearly alias-free solutions. In addition to these advantages, it is also very computationally efficient. On the basis of these virtues, it has had a long run of success and has been the method of choice at many modeling centers.

Computational Methods

Models represented by finite differences are often denoted as grid point models and the grids for these models have been selected in a variety of ways. Despite their popularity, these models have many problems leading to significant computational errors, and the spectral method with its simple lateral boundary conditions over the globe is a natural alternative. Both methods are applied in the horizontal space domain, and are combined with an alternate discretization in both the vertical and time domains. The techniques were developed with the prediction equations represented in the Eulerian framework, that is, all calculations are made locally in the space domain, including time extrapolation of the dependent variables.

Although the structural characters of the two methods are substantially different, they can be cast in a similar representational form allowing for more systematic comparison. To elucidate this similarity, consider the dependent variables presented in the prediction equations (eqns [1]–[4]) represented by the vector $\mathbf{B} = \{B_b\} = (\mathbf{V}\rho s q_v q_1, q_i, a_j \dots)^T$, where T represents transpose. The dimensions of $\mathbf{B}$ are determined by the number of variables in the system; let that be N . As the equations stand, the left-hand side of the set is simply $\partial \mathbf{B} / \partial t$ and the right-hand side can be summarized by a vector $\mathbf{F}$ with the same dimension as $\mathbf{B}$ to yield the following system (eqn [5]).

$$\frac{\partial \mathbf{B}}{\partial t} = \mathbf{F}(\mathbf{B}, \mathbf{r}, t) \quad [5]$$

$\mathbf{F}$ depends both differentially and nonlinearly on $\mathbf{B}$, the space coordinates $\mathbf{r}$, and time. Altering these equations by a transformation with the linear matrix operator $\mathbf{L}$ leads to the more general form (eqn [6]) for the prediction system.

$$\mathbf{L} \frac{\partial \mathbf{B}}{\partial t} = \tilde{\mathbf{F}}(\mathbf{B}, \mathbf{r}, t) \quad [6]$$

Consider first the finite difference process applied to this system. Selecting a three-dimensional grid with M points to approximate the continuum in space with suitably prescribed boundary conditions, and a difference operator to describe derivatives, $\mathbf{B}$ is represented at each of the points and has dimensions $(N \times M)$; if the values of $\mathbf{B}$ are available at some initial time, a numerical integration can proceed. The matrix $\mathbf{L}$ becomes by virtue of the difference operator an $(N \times M) \times (N \times M)$ matrix, which can in principle be inverted, and $\mathbf{F}$ also becomes a numerical vector with $N \times M$ elements after utilization of the difference operator at each grid point. Using a circumflex to represent numerical vectors and matrices at grid points, the finite difference system is written as eqn [7].

$$\frac{\partial \hat{\mathbf{B}}}{\partial t} = \hat{\mathbf{L}}^{-1} \hat{\mathbf{F}}(\hat{\mathbf{B}}, \hat{\mathbf{r}}, t) \quad [7]$$

The solution is thus reduced to a matrix computation, provided a numerical scheme is introduced to step the variables forward in time, and the resulting computational errors and stability issues are dependent on the numerical and physical approximations made.

The spectral method uses a different approach. Given a continuous domain over which the model variables are to be evaluated, a set of linearly independent global functions that are continuous over the domain with at least continuous first and second derivatives are selected. The model variables B_b are

expanded in these functions with time-dependent coefficients. Thus, instead of a set of values for the B_b at each grid point $(i\Delta x_1, j\Delta x_2, k\Delta x_3)$ one has eqn [8], where Z_a are the global expansion functions (with their requisite properties).

$$B_b(\mathbf{r}, t) = \sum_{a=1}^{M_e} B_{b,a}(t) Z_a(\mathbf{r}) \quad [8]$$

The choice of these functions is arbitrary, but some guidelines may optimize their selection. It would be ideal to select functions that fit the observation points of the expanded variables exactly, but the distribution of observations is not sufficiently uniform to make this feasible. The expansion functions might be chosen to fit statistics of observations interpolated to a more uniform grid such that the least number of functions (N) was required to describe most of the variance of the variables at those points. Additionally, functions could be chosen that fit boundary conditions most efficiently and/or with convenient orthogonalization properties.

For application to the prediction system, eqn [8] is introduced into eqn [6]. To maintain the exact form of eqn [6], the series given by eqn [8] must be infinite. Using a truncated form creates the spectral model, and also generates errors analogous to those from reduction to a grid (eqn [7]). Selection of an optimum truncation is therefore a significant issue.

The operator L_b , often used with the spectral method, is a diagonal matrix with L_b elements because the system is always linearly decoupled. The scalar spectral representation of eqn [6] is thus eqn [9] and the variables remain nonlinearly coupled in the functions $\tilde{F}_b$.

$$L_b \frac{\partial \mathbf{B}_b}{\partial t} = \tilde{F}_b(\mathbf{B}, \mathbf{r}, t) \quad [9]$$

Substitution of eqn [8] into eqn [9] leads to the error eqn [10].

$$\sum_{a=1}^{M_e} \left(\frac{\partial B_{b,a}}{\partial t} L_b Z_a \right) - \tilde{F}_b = \varepsilon_b \quad [10]$$

To solve this system for the unknown expansion coefficients $B_{b,a}$, multiply eqn [10] by suitable test functions $\hat{Z}_k(\mathbf{r})$ and require the integral over the space domain to vanish, a least-squares error minimization procedure. The test functions must be continuous over the domain and can be arbitrary. In practice, they are frequently chosen to be the expansion functions, but this is not required. With this approximation, the prediction equations for the expansion coefficients become eqn [11], yielding $N \times M_e$ equations for the unknown quantities, $\partial B_{b,a} / \partial t$, which can be solved for $B_{b,a}$ at future times using a suitable time extrapolation procedure.

$$\sum_{a=1}^{M_e} \left(\frac{\partial B_{b,a}}{\partial t} \int L_b Z_a \hat{Z}_k dS \right) - \int \tilde{F}_b \hat{Z}_k dS = 0 \quad [11]$$

To cast eqn [11] in a form more comparable to the finite difference eqn [7], let $\mathbf{B}_b = (B_{b,a})$ and $\mathbf{Z} = (Z_a)$, both vectors with M_e elements. Additionally, assume that the test functions can be similarly represented, i.e., $\hat{\mathbf{Z}} = (\hat{Z}_k)$. Since the functions F_b are implicitly functions of $(\mathbf{r}, t)$ (see eqn [5]), their projection onto the expansion functions yields eqn [12].

$$F_b = \sum_a F_{b,a}(t) Z_a = \mathbf{Z}^T \mathbf{F}_b \quad [12]$$

Generating the coefficients $F_{b,a}$ is nontrivial, resulting from nonlinear combinations of the expansion coefficients $B_{b,a}$ and efficient procedures will be discussed subsequently. Using the defined vectors, eqn [11] becomes eqn [13], representing M_e equations for the expansion coefficients of each dependent variable.

$$\int \hat{\mathbf{Z}}_b \mathbf{Z}^T dS \cdot \frac{\partial \mathbf{B}_b}{\partial t} = \int \hat{\mathbf{Z}} \mathbf{Z}^T dS \cdot \mathbf{F}_b \quad [13]$$

To combine the N equations of eqn [13] into one expression, define the $M_e \times M_e$ matrices $A_b \equiv \int \hat{\mathbf{Z}}_b \mathbf{Z}^T dS$ and $A \equiv \int \hat{\mathbf{Z}} \mathbf{Z}^T dS$, and then create $(N \times M_e) \times (N \times M_e)$ matrices $A_L = \text{diag}(A_b)$ and $A_R = \text{diag}(A)$. Extended vectors for the expansion coefficients to include all the variables can be constructed such that $B_s = (B_b)$ and $F_s = (F_b)$, leading to an equation (eqn [14]) formally identical to the finite difference eqn [7].

$$\frac{\partial B_s}{\partial t} = A_L^{-1} A_R F_s \quad [14]$$

The corresponding grid point values from this spectral representation may be calculated at each point $(i\Delta x_1, j\Delta x_2, k\Delta x_3)$ for each dependent variable B_b by use of eqn [8].

Spectral Modeling

Since most significant prediction models represent their dependent variables on a grid of points in the vertical and use nonspectral methods on that grid, the subsequent discussion of the spectral method will focus on the horizontal domain of the model representation. This requires that the variables B_b be represented on K surfaces in the vertical, with the surfaces separated by the grid intervals, and the variables described in those surfaces by eqn [8].

When selecting appropriate spectral functions for the expansion eqn [8], in addition to fitting observations well, the functions should also be chosen with the properties of the system in mind. Several conditions have been accepted as suitable requirements. First, require the functions Z_a to satisfy the eigenvalue problem (eqn [15]).

$$L_b Z_a = -c_{b,a} Z_a \quad [15]$$

In practice, the selection of L_b almost always represents a conversion of wind components to vorticity and divergence, which is given by a linear differential operator. Application of this operator in eqn [15] leads to a variety of useful and simple functions. The second condition is to require the expansion functions to be orthogonal and normal over the domain in a Hermitian sense (eqn [16]).

$$\int Z_i Z_j^* dS = \delta_{i,j} \quad [16]$$

This condition is reasonably simple to satisfy, since most function sets can be orthogonalized. Finally, the test functions when selected as the expansion functions do not lead to a significant loss of generality, thus this condition is uniformly imposed as $\hat{\mathbf{Z}} = \mathbf{Z}$. Utilizing these three conditions greatly simplifies the calculations required to perform each prediction time step since both integrals in eqn [13] become diagonal matrices.

A variety of functions have been used for the expansion (eqn [8]), most satisfying the conditions just enumerated, with the selection depending on the degree of generality desired to approximate the general system (eqn [14]). When the atmosphere is represented on a channel with rigid boundaries at fixed northern and southern latitudes short of the poles, double Fourier series in latitude and longitude are found to be convenient expansion functions. They satisfy the boundary conditions easily and, because of the very simple addition rules for these functions, nonlinear products are rapidly calculated. For the full global domain approximated by spherical surfaces over the Earth, the obvious expansion functions that satisfy the boundary conditions are surface spherical harmonics (often denoted as solid harmonics), and they have become the functions of choice for spectral modeling. Surface spherical harmonics are constructed as the product of associated Legendre polynomials and complex exponential functions. Selecting coordinates in spherical surfaces such that $\mu = \sin \varphi$, where φ is the latitude, and λ is the longitude, normalized Legendre polynomials represent the latitudinal structures with the form of eqn [17].

$$P_n^m(\mu) \equiv \left[(2n+1) \frac{(n-m)!}{(n+m)!} \right]^{1/2} \frac{(1-\mu^2)^{m/2}}{2^n n!} \times \left(\frac{d}{d\mu} \right)^{n+m} (\mu^2 - 1)^n \quad [17]$$

These are polynomials of degree n with $n - m$ roots in the domain $-\pi/2 < \varphi < \pi/2$ and m roots at the poles. Together with Fourier series in longitude the solid harmonics are given by eqn [18].

$$Y_{n,m}(\lambda, \mu) = P_{n,m}(\mu) e^{im\lambda} \quad [18]$$

These are the complex expansion functions Z_a used in eqn [8] for the horizontal structures. All functions vanish at the poles except the zonal ones ($m = 0$), and these remain finite there. The indices (n, m) define the roots of the functions and thus may be considered scaling elements; that is, the larger the indices, the smaller the scales represented by the functions. An example is given in Figure 1, which shows the cellular structure of the function for fixed n and various values of m . The total number of cells over the domain remains the same because some of the roots appear at the poles, but the cell structures differ. It is convenient to represent the indices as a single complex index, say $\alpha = (n + im)$. The functions are orthogonal over their respective domains and normalized; this is expressed (in a Hermitian sense) as eqn [19] with integration taken over the surface of the unit sphere.

$$\frac{1}{4\pi} \int Y_\alpha Y_{\alpha'}^* dS = \delta_{\alpha,\alpha'} \quad [19]$$

The asterisk signifies complex conjugation, and δ is the Kronecker delta. If $L_b \equiv \nabla^2$ (the Laplacian operator), substitution of Y_α for Z_a in eqn [15] yields the eigenvalues (eqn [20]).

$$c_\alpha = n(n+1) \quad [20]$$

Thus, solid harmonics satisfy the conditions desired for suitable expansion functions.

Most atmospheric variables (B_b) are sufficiently smooth that, when expanded in these functions, the series converges

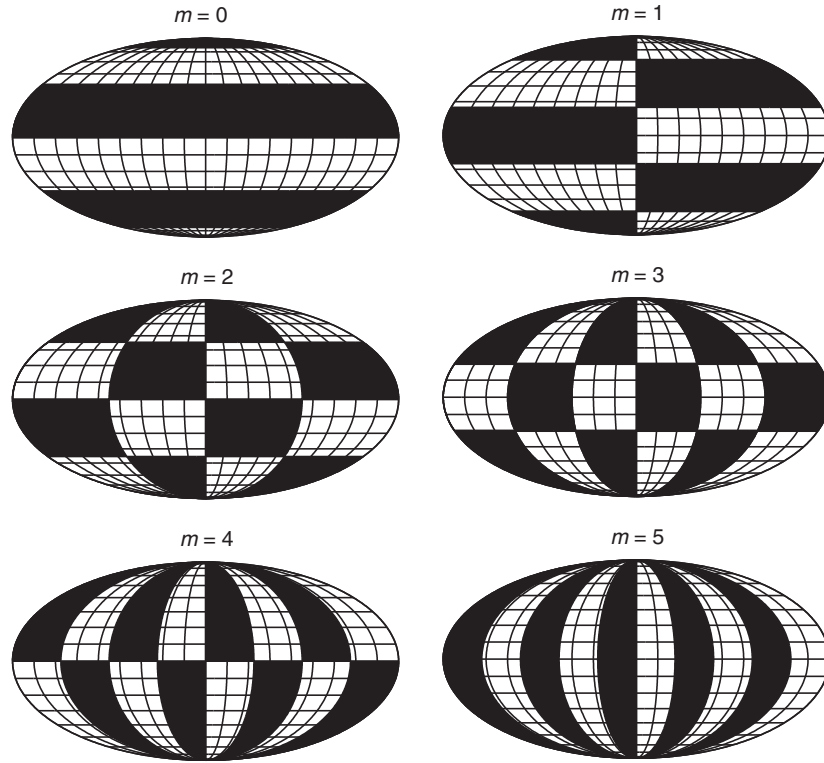


Figure 1 Cellular structure of solid harmonic functions for $n = 5$ and all allowed values of m . Reproduced from Baer, F., 2000. Numerical weather prediction. In: Zelkowitz, M.V. (Ed.), *Advances in Computers*, vol. 52. Academic Press, London, UK, pp. 91–157.

rapidly. That expansion takes the form eqn [21], where z_k is any selected vertical level and the series truncates at M_e .

$$B_b(\lambda, \mu, z_k, t) = \sum_{\alpha} B_{b,\alpha,k}(t) Y_{\alpha}(\lambda, \mu) \quad [21]$$

The range of (α) is $n \geq 0$ and, because of the complex nature of Fourier series, m takes both positive and negative values. When eqn [21] is introduced into eqn [14] and suitably integrated over the space domain, the resulting equations become a set of ordinary nonlinear differential equations in time for the expansion coefficients.

Spectral Vorticity Model

To better understand the details of this methodology, it is advantageous to simplify eqn [14] by approximations but still maintain a system that can describe the elements of the technique. The simplest appropriate system is represented by the barotropic vorticity equation (BVE). Consider a barotropic fluid, which exists if the thermodynamic variables are uniquely related to one another and are independent of position in the fluid. In this setting, fluid motions need consideration in only one horizontal surface and are independent of height. Assuming further that the fluid is incompressible, it is then also three dimensionally nondivergent. If no divergence is introduced at the upper and lower boundaries, no divergence exists in any horizontal surface. Finally, under the condition of hydrostatic equilibrium, the vertical component of velocity can be ignored. The horizontal velocity is then represented by two scalar variables, which themselves may be transformed to any other two scalar functions; because rotation plays such a major role in

atmospheric motions, vorticity and divergence are universally chosen. For the approximations stated, the divergence vanishes and hence the velocity is represented uniquely by the vorticity and the prediction equation for vorticity derived from eqn [1] is denoted the BVE.

Applying these approximations to eqn [1] and ignoring friction, the simplified equation of motion is eqn [22], where the subscript 2 denotes two dimensionality.

$$\frac{\partial \mathbf{V}_2}{\partial t} = -(\mathbf{V}_2 \cdot \nabla) \mathbf{V}_2 - 2\mathbf{\Omega} \times \mathbf{V}_2 - \frac{1}{\rho(p)} \nabla_2 p \quad [22]$$

The Earth's vorticity is expressed here as $2\mathbf{\Omega} = f\mathbf{k}$ with $f = 2\Omega \sin \phi$, the Coriolis parameter, and j is the latitude. Transform the velocity to rotation and divergence by the definitions (eqns [23a]–[23c]).

$$\mathbf{V}_2 = \mathbf{k} \times \nabla \psi + \nabla \chi \quad [23a]$$

$$\nabla \cdot \mathbf{V}_2 = \nabla^2 \chi \equiv \delta \equiv \text{divergence} \quad [23b]$$

$$\mathbf{k} \cdot \nabla \times \mathbf{V}_2 = \nabla^2 \psi \equiv \zeta \equiv \text{relative vorticity} \quad [23c]$$

The equation for predicting the vorticity (BVE) is established by applying the operator $\mathbf{k} \cdot \nabla \times$ to eqn [22] and substituting eqns [23a]–[23c], as in eqn [24].

$$\begin{aligned} \frac{\partial \zeta}{\partial t} &= -\mathbf{V}_2 \cdot \nabla \eta = -\mathbf{k} \times \nabla \psi \cdot \nabla \eta = -J(\psi, \eta) \\ \eta &\equiv \zeta + f \equiv \text{absolute vorticity} \end{aligned} \quad [24]$$

This equation represents a very simplified atmosphere but contains prominent features of the full atmospheric system and is a useful tool for evaluating prediction techniques.

Nondimensionalizing eqn [24] using the Earth's radius (a) for space and its rotation rate (Ω) for time, and noting that the Coriolis parameter becomes $f = 2\mu$, eqn [24] in terms of the stream function (ψ) is then eqn [25].

$$\frac{\partial \nabla^2 \psi}{\partial t} = -2 \frac{\partial \psi}{\partial \lambda} - F(\psi) \quad [25a]$$

$$F(\psi) \equiv \frac{\partial \psi}{\partial \lambda} \frac{\partial \nabla^2 \psi}{\partial \mu} - \frac{\partial \psi}{\partial \mu} \frac{\partial \nabla^2 \psi}{\partial \lambda} \quad [25b]$$

Indeed, $\psi = B$, the only variable remaining of the set B_n in eqn [21] and for only one k level. Equation eqn [25] contains a linear term and two quadratic nonlinear terms; these latter terms constitute F , the remains of F_b in [12]. A representation in terms of expansion coefficients $\psi_\alpha(t)$ is attained using eqn [15] for the Laplacian operator, eqn [21] for the expansion of ψ , and eqn [12] for expansion of F , yielding eqn [26].

$$\sum_{\alpha} c_{\alpha} Y_{\alpha}(\lambda, \mu) \frac{\partial \psi_{\alpha}(t)}{\partial t} = 2i \sum_{\alpha} m_{\alpha} \psi_{\alpha} Y_{\alpha}(\lambda, \mu) + \sum_{\alpha} F_{\alpha} Y_{\alpha}(\lambda, \mu) \quad [26]$$

As a final step, eqn [26] is multiplied by the test functions (in this case the conjugates of solid harmonics) and integrated over the unit sphere, noting orthogonality. This results in the prediction equation for each of the expansion coefficients (eqns [27a] and [27b]).

$$\frac{\partial \psi_{\alpha}(t)}{\partial t} = 2im_{\alpha} c_{\alpha}^{-1} \psi_{\alpha}(t) + F_{\alpha}(t) \quad [27a]$$

$$F_{\alpha}(t) = \int F(\psi) Y_{\alpha}^*(\lambda, \mu) dS \quad [27b]$$

It is evident how eqns [27a] and [27b] can be extended to involve more dependent variables and any number of levels in the vertical. However, if more variables exist in the system, these variables will be coupled nonlinearly through the coefficient F_{α} .

Suppose that the series for α is truncated at M_e as suggested. This implies that all values of ψ_{α} for $\alpha > M_e$ vanish. However, on calculating the nonlinear product $F(\psi)$, $\alpha \leq 2M_e$ coefficients F_{α} are generated; thus at each time step the number of nonvanishing coefficients could double. This complication is resolved in the spectral method by always ignoring all computations for $\alpha > M_e$.

The truncation of α at M_e is somewhat intricate since, from eqn [17], $n \geq 0$ and $n \geq |m|$, whereas $-m_{\max} \leq m \leq m_{\max}$. The set of all allowed indices is best described by the intersections of integers in a grid on an n, m plane as depicted in Figure 2. The allowed points fall on an infinite triangle bounded by the lines $n = \pm m$, but it is sufficient to present only the triangle for $m \geq 0$. All sequential values of n and m beginning at the origin are generally selected to satisfy convergence requirements for the dependent variables that they represent, but a relationship between maximum values must be chosen. Two options are preferred. The first, denoted as rhomboidal truncation, has a maximum value of $m_{\max} \equiv M$ (specified) and allows for all

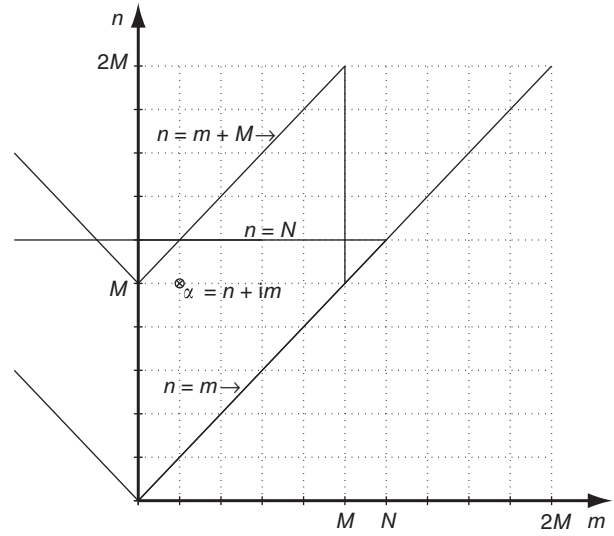


Figure 2 The domain and allowable range of indices m and n for triangular and rhomboidal truncations. Reproduced from Baer, F., 2000. Numerical weather prediction. In: Zelkowitz, M.V. (Ed.), *Advances in Computers*, vol. 52. Academic Press, London, UK, pp. 91–157.

values of $n \leq |m| + M$ for each $|m| \leq M$. The corresponding figure (this configuration describes a parallelogram) is represented in Figure 2 and the notation is written as, for example, R30 if $M = 30$. The advantage of this truncation is that each planetary wave m is represented by the same number of expansion coefficients, thereby allowing equal resolution for all waves. However, since the energy of atmospheric flow decreases rapidly with increasing wave number (m), resolution of the shorter waves may be less important than for the longer waves. This observation leads to triangular truncation, in which $n \leq N$ for each $|m| \leq M$, with $N \geq M$, a predetermined integer. Usually N is selected equal to M and this option is described as a triangle in Figure 2 with the notation T30 if $N = 30$, for example. In terms of scaling, this truncation has some advantages. The ultimate choice for truncation is to optimize the resolution of the model in terms of the number of scales included and to minimize the computing requirements by selecting the fewest degrees of freedom compatible with resolution.

Interaction Coefficient Method

Since all prediction models are computationally intensive, the spectral method must compete in the efficient utilization of available computing resources. It is apparent from eqns [27a] and [27b] that most of the computing time required involves the calculation of the coefficients F_{α} and much effort has gone into optimizing this calculation. Early attempts followed the procedure of substituting the expansion series eqn [21] for ψ into eqns [25a] and [25b] to represent $F(\psi)$ and calculating F_{α} from eqns [27a] and [27b]. This results in eqns [28a] and [28b].

$$F_{\alpha}(t) = \frac{i}{2} \sum_{\beta} \sum_{\gamma} \psi_{\beta}(t) \psi_{\gamma}(t) I_{\alpha, \beta, \gamma} \quad [28a]$$

$$I_{\alpha, \beta, \gamma} \equiv (c_{\beta} - c_{\gamma}) \times \int \left(m_{\beta} Y_{\beta} \frac{\partial Y_{\gamma}}{\partial \mu} - m_{\gamma} Y_{\gamma} \frac{\partial Y_{\beta}}{\partial \mu} \right) Y_{\alpha}^* dS \quad [28b]$$

The indices β and γ go over the same range as α , which is determined by the selected truncation, and the integration is over the unit sphere. The integrals $I_{\alpha,\beta,\gamma}$ are denoted as interaction coefficients and have exact solutions. Applying eqns [28a] and [28b] in eqns [27a] and [27b] shows that the time change of any expansion coefficient of the set α depends on the coupling of all the coefficients allowed in the spectral domain (refer to Figure 2) and each couple is weighted by its own interaction coefficient. Since each index consists of two real numbers, the set of interaction coefficients can be as large as the largest allowed index to the sixth power. In actuality, because of the simple addition rules for trigonometric functions, the integration over longitude reduces this by one order. The vector of these coefficients can be stored and needs to be computed only once. However, the number of multiplications that must be performed at each time step is daunting as the truncation limit becomes large.

The more complex system (eqn [14]) can be represented identically to eqns [27a] and [27b] by simply increasing the number of expansion coefficients to include additional variables. But a shortcoming of using interaction coefficients concerns the convergence rate for the series of several dependent variables included in the general set (B_b) when expanded in global functions, in particular liquid water and precipitation. Significant truncation errors may ensue with time integration utilizing such functions.

Transform Method

A technique denoted as the transform method is an alternate procedure for calculating the coefficients F_α , yielding the same (or better) results than the interaction coefficient method. This technique involves the transformation of the integrand in eqn [27] onto a special numerical grid and solving the integral by quadrature. If the grid is selected appropriately, the integral is evaluated exactly and at a great reduction in computing cost. In the longitudinal direction, the quadrature is most conveniently done by a trapezoidal formula since it is known that eqn [29] holds.

$$\frac{1}{2\pi} \int_0^{2\pi} e^{im\lambda} d\lambda = \frac{1}{J} \sum_{j=1}^J e^{im\lambda_j} \quad [29]$$

The summation is taken over an equally spaced grid of points λ_j , and uses twice the number of points as the maximum wave number. Since the functions in latitude are Legendre polynomials, a Gaussian quadrature is preferred. In this case, the quadrature is such that eqn [30] holds and is exact if the polynomial H is of degree $2K - 1$ or less.

$$\int_{-1}^1 H(\mu) d\mu = \sum_{k=1}^K G_k(\mu, K) H(\mu_k) \quad [30]$$

The G_k are Gaussian weights and the grid points μ_k are the roots of the Legendre polynomial $P_K(\mu)$. The appropriate grid for this calculation contains all allowed values (λ_j, μ_k) as specified. The range of the grid points is determined by the functions of the integrand in eqns [27a] and [27b].

The derivatives in $F(\psi)$ (see eqn [25]) must be taken before evaluating the function on the grid. Based on eqns [18] and [17], the differentiation with λ is straightforward, but the

μ -derivative requires more information. The Legendre polynomials satisfy the differential eqn [31], where the coefficients b_α are constants, and this defines the latitudinal derivatives.

$$(1 - \mu^2)^{1/2} \frac{dP_\alpha}{d\mu} = b_\alpha P_{\alpha-1} - b_{\alpha+1} P_{\alpha+1} \quad [31]$$

Following this procedure, $F(\psi)$ is reduced to a quadratic series over the indices (β, γ) in terms of the complex exponential functions in longitude and the associated Legendre polynomials in latitude, both of which can be evaluated on the specified grid. The actual calculation proceeds as follows. First, the quadrature over longitude is taken (eqn [32]), where the sum goes over the value $J = 3M - 1$ if triangular truncation is chosen.

$$\begin{aligned} Fm_\alpha(\mu, t) &= \frac{1}{2\pi} \int F(\psi(\lambda, \mu, t)) e^{-im_\alpha \lambda} d\lambda \\ &= \frac{1}{J} \sum_{j=1}^J F(\psi(\lambda_j, \mu, t)) e^{-im_\alpha \lambda_j} \end{aligned} \quad [32]$$

The calculation is made over those latitudes μ specified from the quadrature (eqn [33]).

$$\begin{aligned} F_\alpha(t) &= \frac{1}{2} \int Fm_\alpha(\mu, t) P_\alpha(\mu) d\mu \\ &= \sum_{k=1}^K G_k(\mu_k, K) Fm_\alpha(\mu_k, t) P_\alpha(\mu_k) \end{aligned} \quad [33]$$

Since the polynomial under summation in eqn [33] is $H(\mu)$ and is the product of three Legendre polynomials less one order, and each has a maximum order of N , it can be shown that $K = (3N - 1)/2$.

Analysis of the computing requirements for eqns [32] and [33] indicates that the maximum number of calculations is proportional to N^3 , significantly less than the N^5 needed by the interaction coefficient method. When using the transform method with those variables that have unacceptable convergence properties yet contribute to eqn [7], their series representation is not essential. Their input is included directly into the quadrature formula by their distribution on the transform grid. Since all the forcing functions are summed over the grid before quadrature is completed, any singularities from individual terms are smoothed out and their effects are minimized.

History

Since the 1960s, spectral models have become by far the most popular representation for describing the global atmospheric prediction equations in computational form. They overcome many of the limitations inherent in finite difference models. Most international prediction centers have adopted this modeling procedure. Canada and Australia implemented the model in 1976, the National Meteorological Center of National Oceanic and Atmospheric Administration (NOAA) did so in 1980, the French in 1982 and the European Center for Medium-Range Weather Forecasts (ECMWF) in 1983. As an example of how the models have evolved, production spectral models at ECMWF have grown in resolution from T63 in 1983 to T213 in 1998 with experiments currently running at T319.

See also: **Boundary Layer (Atmospheric) and Air Pollution:** Modeling and Parameterization. **Data Assimilation and Predictability:** Data Assimilation; Ensemble Prediction; Predictability and Chaos. **Numerical Models:** Mesoscale Atmospheric Modeling; Methods; Regional Prediction Models. **Weather Forecasting:** Seasonal and Interannual Weather Prediction.

- Boyd, J.P., 2000. Chebyshev and Fourier Spectral Methods, seconded. Dover, New York.
- Krishnamurti, T.N., Bedi, H.S., Hardiker, V.M., 1998. An Introduction to Global Spectral Modeling. Oxford University Press, Oxford.
- Machenhauer, B., 1991. Spectral methods. In: Numerical Methods in Atmospheric Models, vol. 1. European Center for Medium-range Weather Forecasts, Reading, UK, pp. 3–86.
- Washington, W.M., Parkinson, C.L., 1986. An Introduction to Three-Dimensional Climate Modeling. University Science Books, Mill Valley, CA.

Further Reading

- Baer, F., 2000. Numerical weather prediction. In: Zelkowitz, M.V. (Ed.), Advances in Computers, vol. 52. Academic Press, London, pp. 91–157.

Mesoscale Atmospheric Modeling

RA Pielke, Sr., University of Colorado at Boulder, CO, USA

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

Mesoscale systems in the atmosphere are those in which the instantaneous pressure field can be determined accurately by the temperature field, but the winds, even in the absence of surface frictional effects, are out of balance with the horizontal pressure gradient force. The framework of mesoscale models is overviewed and shows that the models include a physics base core comprising the pressure gradient force, advection, and gravity, and all other physical processes that are parameterized using tuned modular representations of turbulence, longwave and shortwave radiation, cumulus, and stable cloud processes. Computational methods, lateral and initial boundary conditions, and model validation are some of the topics discussed.

Introduction

Atmospheric mesoscale systems are identified as those in which the instantaneous pressure field can be determined accurately by the temperature field, but the winds, even in the absence of surface frictional effects, are out of balance with the horizontal pressure gradient force. The pressure field, under this situation, is said to be 'hydrostatic.' Larger scale atmospheric features (which are called 'synoptic' weather features), in contrast, have a wind field that is close to a balance with the horizontal pressure gradient force. These large-scale winds are said to be near gradient wind balance. Atmospheric features that are smaller than the mesoscale have pressure fields in which wind acceleration is a significant component (which is referred to as the dynamic wind). The pressure gradient that causes this dynamic wind is called the nonhydrostatic pressure.

Atmospheric mesoscale models are based on a set of conservation equation for velocity, heat, density, water, and other trace atmospheric gases and aerosols. The equation of state used in these equations is the ideal gas law. The conservation of velocity equation is derived from Newton's second law of motion as applied to the rotating Earth. The conservation of heat equation is derived from the first law of thermodynamics. The remaining conservation equations are written as a change in an atmospheric variable (e.g., water) in a Lagrangian framework where sources and sinks are identified.

Each of these conservation equations can be written to represent the changes following a parcel of velocity, potential temperature (entropy), water in its three phases, other atmospheric gases and aerosols, and mass, including source-sink terms. Models, however, seldom express the conservation relations in a Lagrangian framework. The chain rule of calculus is used to convert to an Eulerian framework.

Several assumptions are typically made in the conservation equations. These include the neglect of small-scale fluctuation of density except when multiplied by gravity (this is called the Boussinesq approximation), the neglect of vertical acceleration relative to the differences between the vertical pressure gradient force and gravity (referred to as the hydrostatic assumption), and the neglect of all molecular transfers.

The first two assumptions have not been made in recent years in the models, however, since the numerical equations are actually easier to solve without these assumptions. Nonetheless, the spatial and temporal scales of mesoscale systems result

in the two assumptions being excellent approximations with respect to mesoscale-sized systems. The third assumption is justified since advection is much more significant at transfers of heat, momentum, water, and other chemical species, than molecular motion on the mesoscale.

These conservation relations that are written as a set of simultaneous, nonlinear differential equations, unfortunately, cannot be used without integrating them over defined volumes of the atmosphere. These volumes are referred to as the model 'grid volume'. The region of the atmosphere for which these grid volumes are defined is called the 'model domain.' The integration of the conservation relations produces 'grid volume averages,' with point-specific values of the variables called 'subgrid-scale values.' The 'resolution' of data is limited to two grid intervals in each spatial direction.

The result of the grid volume averaging produces equations for the local time derivative of the grid volume-resolved variable which includes 'subgrid-scale fluxes.' An assumption that is routinely made in all mesoscale models (usually without additional comment) is that the grid volume average of subgrid-scale fluctuations is zero. This assumption, often referred to as 'Reynold's averaging,' is actually only accurate when there is a clear spatial scale separation between subgrid scale- and grid volume-resolved quantities.

Mesoscale model equations have been solved in a Cartesian coordinate framework. Each coordinate in this system is perpendicular to the other two coordinates at every location. Most mesoscale models, however, transform to a generalized vertical coordinate. The most common coordinates involve some form of terrain-following transformation, where the bottom coordinate surface is terrain height or terrain surface pressure. The result of these transformations is that the new coordinate system is not orthogonal, in general. Unless this nonorthogonality is small, the correct treatment of non-hydrostatic pressure effects in mesoscale models requires the use of tensor transformation techniques, as opposed to the separate use of the chain rule on each component of velocity, separately. The use of generalized coordinate systems introduces additional sources for errors in the models, since the interpolation of variables to grid levels becomes more complicated.

The model variables also need to be defined on a specified grid mesh. When all dependent variables are defined at the same grid points, the grid is said to be 'nonstaggered.' When dependent variables are defined at different grid points, the grid

is called a 'staggered grid.' The grid meshes can also be defined with smaller grid increments in one region surrounded by coarser grid increments. Such a grid is referred to as a 'nested grid.' If the grid increments vary at all locations, with the finest grid in a specified volume, the grid is called a 'stretched grid.'

The subgrid-scale fluxes in mesoscale models are parameterized in terms of resolvable variables. Turbulence theory, as observed from atmospheric field campaigns over horizontally homogeneous landscapes, and for undisturbed atmospheric conditions, is the basis for all mesoscale model representations of the vertical subgrid-scale flux terms. The vertical fluxes are parameterized differently when the lowest 50 m or so of the atmosphere are unstably stratified and when it is stably stratified. The planetary boundary is typically represented by three layers: a thin layer of a few centimeters near the surface where laminar fluxes are important (called the 'laminar layer'), a layer above which extends upward tens of meters where wind direction with altitude is ignored (referred to as the 'surface layer'), and the remainder of the boundary layer where the winds approach the free atmospheric value (referred to as the 'transition layer'). Disturbed (unsteady) boundary layers are not parameterized accurately, however, by existing parameterizations. The effect of land surface heterogeneity has been included on the subgrid scale only as a weighting of the surface layer fluxes by the fractional coverage of each land surface type. This technique is called the 'mosaic' or 'tile' subgrid-scale surface flux parameterization.

In contrast to the vertical fluxes, horizontal subgrid-scale fluxes in mesoscale models have no physical basis. They are included only to horizontally smooth the model calculations.

The representation of the source-sink terms in the conservation equations can be separated into two basic types: those that are derived from basic concepts and those that are parameterized. The only basic source-sink terms in mesoscale models that are derived from fundamental physical concepts are the pressure gradient force, advection, and gravity. Neither of these two effects involves adjustable (i.e., tunable) coefficients, which is one method to separate fundamental terms in the conservation equation from a parameterization. The remainder of the source-sink terms needs to be parameterized. Almost all parameterizations currently used in these models are either box or vertical column representations.

The radiative flux terms are typically separated into shortwave and longwave fluxes. The shortwave fluxes, also called 'solar fluxes,' are separated into direct and diffuse irradiance. The direct irradiance is the nonscattered flux, whereas the diffuse irradiance is the scattered radiative flux from the Sun. The direct irradiance is sometimes further separated into visible and near-infrared components. In cloudy model atmospheres, parameterizations based on cloud liquid water content, or more crudely on arbitrary attenuation based on relative humidity in the model, are used. Typically, only diffuse irradiance is permitted for overcast model conditions. Some models weight the fluxes for partly cloudy skies, using weighted parameterizations for both clear and overcast sky conditions. Polluted atmospheres also require parameterization on their effect on solar irradiance, although only a few mesoscale models have explored this issue.

Longwave irradiance is from the Earth's surface and from within the atmosphere. Scattering of longwave radiative fluxes

is typically ignored, such that only upwelling and downwelling irradiances are parameterized. This type of parameterization is called a 'two-stream approximation.' The major absorbers and emitters represented in mesoscale model parameterizations are liquid and ice clouds, water vapor, and carbon dioxide. Clouds are usually parameterized as black bodies to longwave irradiance. The water vapor and carbon dioxide are represented by the path length through the atmosphere, and their concentrations along their path. As with solar radiative fluxes, mesoscale models seldom include parameterizations of longwave irradiance associated with pollution. This neglect is partially a result of the dependency of the absorption, transmissivity, and scattering of both solar and longwave irradiance on the specific chemical composition and size spectra of the pollution.

The phase changes of water and this effect on the conservation of heat source-sink term are separated into stable cloud, cumulus convective cloud, and precipitation parameterizations. Stratiform cloud parameterizations range in complexity from algorithms which instantaneously precipitate rain (or snow) when the model relative humidity exceeds a user-specified relative humidity (referred to as a 'dump bucket' scheme), to individual conservation equations for several categories of hydrometers (e.g., cloud water, rain water, ice crystals, snow, graupel, and hail). For the larger hydrometeors, a nonzero, finite terminal fall velocity is usually specified. More detailed microphysical representations, where cloud hydrometer spectra are classified into more size class intervals (called 'microphysical bin parameterizations') are also used.

The parameterization of cumulus cloud rainfall utilizes some form of one-dimensional cloud model. These are called 'cumulus cloud parameterization schemes.' Their complexity ranges from instantaneous readjustments of the temperature and moisture profile to the moist adiabatic lapse rates when the relative humidity exceeds saturation, to representations of a set of one-dimensional cumulus clouds with a spectra of radii. These parameterizations typically focus on deep cumulus clouds, which produce the majority of rainfall and diabatic heating associated with the phase changes of water. Cumulus cloud parameterizations remain one of the major uncertainties in mesoscale models, since they usually have a number of tunable coefficients, which are used to obtain the best agreement with observations. Also, since mesoscale model resolution is close to the scale of thunderstorms, care must be taken so that the cumulus parameterization and the resolved moist thermodynamics in the model do not 'double count' this component of the source-sink terms.

The grid volume-averaged conservation equations are nonlinear and, therefore, must be solved using numerical approximation schemes. The solution techniques include finite difference, finite element, interpolation (also called semi-Lagrangian), and spectral methods. Both temporal-spatial terms and the source-sink terms must be represented by these approximation schemes. An important aspect of mesoscale models is that only advection and the pressure gradient force involve horizontal gradients explicitly. All other model terms, including each of the source-sink terms, are one-dimensional column models or point values.

Finite difference schemes involve some form of truncated Taylor series expansion. The finite element technique uses

a local basis function to minimize the numerical error, whereas the spectral method utilizes global basis functions. A spectral method has the advantage that differential relations are converted to algebraic expressions. The semi-Lagrangian scheme is based on fitting interpolation equations to data at a specific time and advecting the data with model winds.

Mesoscale models have predominately utilized the finite difference and (for advection) the semi-Lagrangian approaches. A few groups have applied the finite element method, but its additional computational cost has limited its use. The spectral method, which is most valuable for models without lateral boundaries, has not generally been used since mesoscale models have lateral boundaries.

The use of numerical approximations introduces errors. Linear stability analyses show that it is impossible to create a numerical solution scheme which accurately represents both amplitude change over time and speed of motion (advection and gravity wave propagation) for features that are shorter than about four grid increments in each spatial direction. In addition, products of the variables (which are a nonlinear terms) produce transfers of spatial scales to larger and smaller scales. The inability of the numerical model to sample the smallest scales (less than two grid intervals) results in the information (e.g., winds and temperature) erroneously appearing at a larger spatial scale. This error is called 'aliasing' and unless corrected can result in an incorrect accumulation of atmospheric structure at the wrong spatial scale. For these reasons, the term 'model resolution' should be reserved for features that are at least four grid intervals in each direction.

To integrate the models forward in time, the variables must be initialized. These values are called 'initial conditions.' Observed data, or a combination of observed data and previous model calculations, are typically used to initialize the mesoscale models. The insertion of data during a model calculation is called 'four-dimensional data assimilation (4DDA).' Lateral, top, and bottom boundary conditions are also needed for the duration of the model calculations. Lateral boundary conditions in mesoscale models can be idealized for theoretical studies (e.g., cyclic boundary conditions), or derived from large-scale observations, such as the NCEP Reanalysis or from larger scale model simulations (which is referred to as dynamic downscaling). Mesoscale models are often strongly influenced by the lateral boundary conditions, such that their accurate representation is a necessary condition for an accurate mesoscale simulation.

The top boundary conditions are similar to the lateral boundary condition and must be accurately represented. Most mesoscale models extend into the stratosphere, in order to minimize the effect of the model top on the mesoscale simulation. Damping zones at the model top (referred to as an 'absorbing layer') are usually inserted so that upward propagating model simulated gravity waves do not erroneously reflect from the artificial model top.

The surface boundary is the only surface of a mesoscale model which is physically based. This surface is typically separated into ocean (and fresh water lakes) and land surfaces. Ocean and lake surfaces can be represented simply as prescribed sea surface temperatures or using mesoscale atmospheric models coupled to mesoscale ocean, lake, and/or sea ice models. Over land, the ground is separated into bare soil and vegetated land. Soil-vegetation-atmosphere transfer

schemes (SVATS) have been introduced to represent the fluxes of velocity, heat, moisture, and other trace gases between the atmospheres and the surface. Most SVATS include the effect on water flux of transpiration. Recently, vegetation dynamical processes, such as plant growth have been included in longer term (months to seasons) mesoscale model calculations.

Model performance is assessed in several ways. The comparison of observations with model results using statistical skill tests is a major assessment tool. A complication of these evaluations is that observations have a different sampling volume (e.g., a point) than the model grid volume. Comparisons of simplified (usually linearized) version of numerical models with analytic theory have been completed to test the accuracy of linear components of the model. Several models can be intercompared to assess what features they have in common, and which they do not. The mass and energy budgets of the mesoscale models, if they are each calculated in two separate manners, provide an opportunity to check the internal consistency of the model. Peer-reviewed scientific publications and the availability for scrutiny of the model source code provide two additional valuable procedures to assess the value of the mesoscale model and the degree to which the programmed model logic agrees with the mathematical formulations presented in the literature. Proposals have been made to standardize the model computer codes, in order to assist in their more general use.

Mesoscale models have been applied to two basic types of mesoscale systems: those found primarily by initial and lateral boundary conditions (referred to as synoptically forced mesoscale systems) and those forced by surface boundary conditions (referred to as surface-forced mesoscale systems). Of the latter type, there are mesoscale systems that are caused when terrain is an obstacle to the flow (referred to as 'terrain-forced' or orographic mesoscale systems) and those generated by horizontal gradients in sensible heating of the surface (called 'thermally forced' mesoscale systems).

With the improvement in computational power, global models will soon approach mesoscale spatial and temporal resolution (which requires horizontal grid increments of ~ 1 km). This high resolution will eliminate lateral boundary conditions as a component in the accurate simulation of mesoscale atmospheric features models.

Further Reading

- Cotton, W.R., Bryan, G., Van den Heever, S., 2009. Storm and Cloud Dynamics. In: International Geophysics, second ed. vol. 99. Academic Press.
- Fedorovich, E., Rotunno, R., Stevens, B. (Eds.), 2004. Atmospheric Turbulence and Mesoscale Meteorology. Cambridge University Press.
- Kalnay, E., 2003. Atmospheric Modeling, Data Assimilation and Predictability. Cambridge University Press.
- Pielke Sr., R.A., 2002. Mesoscale Meteorological Modeling, second ed. Academic Press, San Diego, CA.
- Pielke, R.A., 2013. Mesoscale Meteorological Modeling, third ed. Academic Press, San Diego, CA.
- Pielke Sr., R.A., Stokowski, D., Wang, J.-W., Vukicevic, T., Leoncini, G., Matsui, T., Castro, C., Niyogi, D., Kishtawal, C.M., Biazar, A., Doty, K., McNider, R.T., Nair, U., Tao, W.K., 20 February 2007. Satellite-based model parameterization of diabatic heating. EOS, vol. 88 (8), 96–97.
- Warner, T., 2011. Numerical Weather and Climate Prediction. Cambridge University Press.
- Whiteman, C.D., 2000. Mountain Meteorology Fundamentals and Applications. Oxford University Press.

Cloud-System Resolving Modeling and Aerosols

W-K Tao, NASA/Goddard Space Flight Center, Greenbelt, MD, USA

T Matsui, NASA/Goddard Space Flight Center, Greenbelt, MD, USA; and University of Maryland, College Park, MD, USA

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

Aerosols and their effects on clouds and precipitation are one of the key components of the climate system and the hydrological cycle. Cloud-system resolving models (CRMs) with explicit microphysics can allow interactive processes between aerosols, cloud and precipitation for convective precipitation systems, hurricanes and other severe weather systems. CRMs can be used to improve our understanding of: (1) the impact of cloud condensation nuclei (CCN) and Ice nuclei (IN) on cloud and precipitation microphysics and (2) the associated important cloud dynamic processes. High-resolution observations from remote sensing would be a necessary part of the CRM validation through state-of-the-art multi-instrument satellite simulators.

Convective Precipitation Systems and Aerosols

Convective precipitation systems (CPSs) contribute to a major fraction of global rainfall distributions and latent heat release in the atmosphere, and their associated stratiform and anvil cover play one of the foremost roles in the Earth's radiation budget. Deep convection transports aerosols, mass, and momentum from planetary boundary layer (PBL) to middle- and upper-troposphere, and play important roles in atmospheric chemistry processes. CPSs could also produce strong wind gusts, downburst, hailstones, heavy rainfall, and lightning that are all harmful consequences on daily human life. Thus, understanding of the formation of the CPS characteristics is imperative for weather prediction and climate projection for our society.

Primal fuels (energy) of CPSs are surface turbulent heat fluxes and large-scale circulation (i.e., instability in atmosphere). Depending on the heat source and surrounding environmental conditions, surface air mass, hereafter denoted as *parcel*, can be lifted up to tropopause (~16 km above ground level) with the bubbly turbulent forms, and eventually shows the characteristic of anvil shape in the maturity stage (see more details in Houze, 1993). During the convection process, a parcel experiences complex cloud microphysics processes interacting with atmospheric aerosols.

When a parcel ascends from near surface, saturation is attained to form cloud droplets upon cloud condensation nuclei (CCN), which are derived from various water-soluble atmospheric aerosols:

$$N_{CCN} = \int_{D_{CRIT}}^{\infty} n(D) dD \quad [1]$$

where D is aerosol diameter; $n(D)$ is aerosol number concentrations; D_{CRIT} is a critical diameter for a given supersaturation rate, solute mass and types, and temperature based on the Köhler equation (which describes the equilibrium saturation ratio over the solute drop surface as a function of the drop radius). In general, a large number of water-soluble aerosol particles tend to increase the CCN number and, thus, cloud droplets. The large number of small cloud droplets competes for available moisture for condensation growth: thereby, cloud

droplets stay in relatively small size, which inhibits collision and coalescence processes to generate submillimeter-sized drizzle drops. Consequently, the large number concentrations of cloud droplets can inhibit warm-rain initiation.

If the parcel is lifted further, these small cloud droplets could be cooled to less than 0 °C without immediate freezing. In this temperature zone, aerosol particles start acting as ice nuclei (IN). The surface of supercooled cloud droplets start contacting with freezing aerosol particles through diffusiophoresis, thermophoresis, and Brownian motion or drop evaporation, and then cloud droplets are catalyzed to form embryo of ice crystals (i.e., contact/evaporation nucleation). Ice crystals can also be formed through immersing freezing aerosol particles with supercooled cloud droplets (i.e., immersion nucleation). Besides these nucleation pathways, ice crystals can also be formed directly upon nonsoluble aerosols by deposition (i.e., primal nucleation) or freezing right after condensation (i.e., condensation nucleation). Without interacting with nonsoluble aerosols, pure cloud droplets must be cooled down to at least -38 °C to form ice crystals (i.e., homogeneous nucleation). These various ice nucleation processes release latent heat by fusion and deposition that may further assist lifting the parcel.

Ice crystals keep growing by deposition of water vapor and/or aggregation processes with other ice crystals. In the convective core, the supercooled cloud droplets collide and freeze upon snowflakes surface (i.e., riming process) to form higher-density snowflakes. Once riming dominates, these particles are denoted as graupel or hail, which induce intensive cold-phase rainfall (see more details in microphysics from Pruppacher and Klett, 1997). While much more effort is required to understand the source of the IN precursor, DeMott et al. (2010) collected various field measurements, and found that measured IN concentrations, N_{IN} , can be well predicted from the concentrations of large-sized aerosol number concentrations ($N_{aero0.5}$) through the following equation (DeMott et al., 2010).

$$N_{IN} = a T^b N_{aero0.5}^{(cT+d)} \quad [2]$$

where

$$N_{aero0.5} = \int_{0.5\mu m}^{\infty} n(D) dD$$

D is the diameter of aerosols; $n(D)$ is aerosols number concentrations for a given diameters; T is air temperature; a , b , c , and d are fitted coefficients. Equation [2] is based on the summary of *in situ* measurements implying that the large concentrations of large-size aerosols in colder air temperature tend to increase the production of ice crystals and release more latent heat in the middle- to upper-troposphere.

Increased ice particles aloft could enhance the amount of ice transported to upper-troposphere. At this level, strong environmental wind together with convection-induced wind can horizontally spread out cloud particles to form the prominent anvil shape. Thus, enhanced ice production due to high aerosol concentrations could enhance the high cloud coverage, which tends to block incoming shortwave radiation from the Sun and trap outgoing long-wave radiation from the Earth's surface.

Overall, high concentrations of aerosols serving as CCN or IN could (1) delay onset of warm-rain process, (2) increase amount of supercooled water and riming process, (3) enhance ice concentrations and aloft, and (4) therefore, invigorate the cold-rain process. However, with enough energy and updraft speed, liquid cloud droplets could be lifted to the mixed-phase zone with little chance of warm-rain process regardless of variability of aerosol loading. More importantly, the presence of vertical wind shear organizes deep moist convection into the long-lasting larger forms, so-called mesoscale convective systems (MCSs). MCSs propagate for a given background wind and internal gravity wave, and a combination of updraft and downdraft could enhance ice production significantly (see more details in Houze, 1993). In this system, mesoscale cloud dynamics could dominate microphysics processes, secondary convections, and subsequent rainfall amount. Since aforementioned hypothesis of aerosol-CPS interaction would strongly depend on a case-by-case basis, first of all, it would require validating the observations extensively in time and space.

Satellite-Based View of Microphysics in CPSs

The aerosol-CPS interactions can be viewed from the satellite remote sensing on a global scale. Satellite-retrieved aerosols and

reanalysis-derived thermodynamic fields are averaged over the June–July–August period in 2006. Figure 1(a) depicts undiluted plume height (UPH) estimated from atmospheric reanalysis (MERRA: Modern Era Retrospective-Analysis for Research and Applications). Deeper moist convection more likely occurs over a higher-UPH area, such as Pacific Warm Pool, whereas it is inhibited over a lower-UPH area, such as a subtropical continental area in the southern hemisphere. Figure 1(b) depicts distribution of aerosol optical thickness (AOT) retrieved from Multiangle Imaging Spectro Radiometer (MISR) on the Terra satellite. In general, higher AOTs can be linked to larger concentrations of aerosols and CCN/IN concentrations.

In the corresponding period, precipitation radar (PR) 14 GHz echo from the Tropical Rainfall Measuring Mission (TRMM) satellite have been sampled on every 1×1 -gridded box over the same domain, and estimated maximum echo-top height (H_{ET}) during the 3-month period (Figure 1(c)). Elevated PR echo represents deeper convection due to raining of particles (snow and graupel) aloft into middle- to upper-troposphere, since the TRMM PR sensor is sensitive to rain-drops or relatively large ice particles. In a similar sampling manner, TRMM Microwave Imager (TMI) polarization-corrected brightness temperature at 85 GHz ($PCTb_{85}$) has been sampled. Figure 1(d) shows the depression in brightness temperature ($dPCTb_{85} = PCTb_{85,max} - PCTb_{85,min}$). With the presence of high-density solid precipitating drops (graupel or hail), microwave emission at 85 GHz from the Earth's surface can be scattered back to the surface; thus, larger $dPCTb_{85}$ generally represents the presence of a large amount of rimed ice particles (see more details in Matsui et al., 2009). In short, the higher PR H_{ET} and the larger $dPCTb_{85}$ represent the CPSs associated with the enhanced cold-rain process, which appear to be distributed in the high UPH regions. However, the very large PR H_{ET} (>15 km) and $dPCTb_{85}$ (>80 K) are mainly distributed over the continent with high aerosol loading estimated from MISR, including in middle to West Africa, northwest India, and the central to western portion of the United States.

From the brief satellite-based view, Figure 1(a)–1(d) supports the hypothesis of aerosol-CPSs interaction that

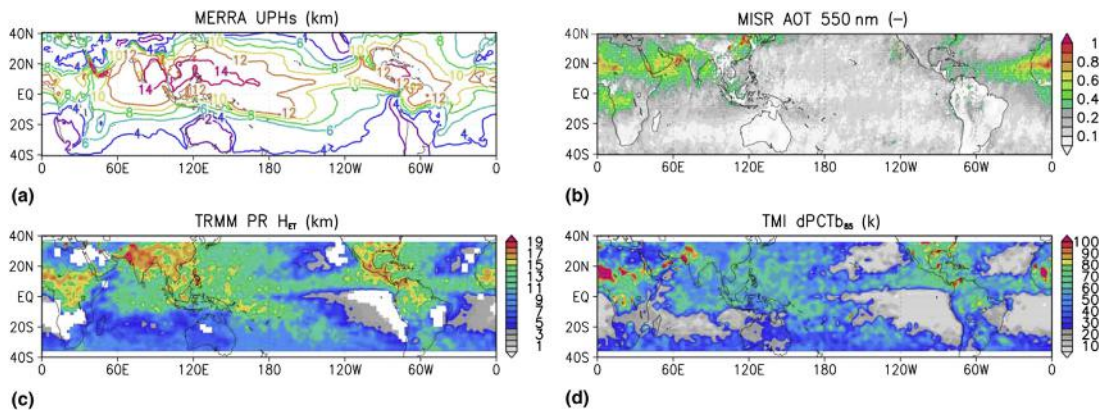


Figure 1 (a) MERRA-estimated undiluted plume height (UPH, defined as the maximum height that a near-surface parcel can be lifted by buoyant force without entraining surrounding air), (b) MISR-derived AOT, (c) TRMM PR-derived maximum H_{ET} , representing storm height, and (d) TMI-derived polarization-corrected brightness temperature depression at 85 GHz ($dPCTb_{85}$), most likely represents column amount of rimed particles.

higher aerosol concentrations tend to invigorate the cold-phase rainfall process in deep moist convection. However, a problem of observational analysis is the interpretation of statistical correlation and time evolution (life cycle) of CPSs. Satellite-based examination relies on a statistical summary of instantaneous observations; technically, it does not follow the life cycle of CPSs. Indeed, statistical analysis cannot conclude cause-and-effect relationship between aerosols and CPSs. Therefore, together with observational support, cloud-scale resolving atmospheric models (horizontal grid spacing of 0.25–3.0 km) with full microphysical interactions are also required to investigate details in the time evolution of and interactions between cloud–precipitation microphysics, meso-scale dynamics, and aerosols.

Cloud-System Resolving Modeling and Microphysics

Cloud-System Resolving Model

One of the most promising methods to test the representation of the detailed aerosol–cloud interaction process is to use explicit microphysics in cloud-system resolving models (CRMs) (see a recent review by Tao and Moncrieff, 2009). There are several major advantages in using CRMs to study the interactive processes between cloud, precipitation, and aerosol. For example, (1) unlike convective parameterization in climate models or general circulation models, CRMs employ a high-resolution, nonhydrostatic dynamic core to simulate the life cycle of CPSs; (2) the use of a fully explicit microphysics

scheme (liquid and ice) and a fine horizontal resolution can provide realistic evolution of cloud structures, optical properties, and latent heat release, which are crucial for determining the atmospheric energy budgets. One major limitation of CRMs is its inability to represent interactions between cloud/cloud system-scale dynamics and synoptic dynamics due to relative limited domains coverage.

Microphysics

Two-moment bulk and spectral-bin microphysics schemes are generally required to study the impact of the CCN, giant CCN (GCCN), and IN on cloud and precipitation formations. Typically, two-moment bulk microphysics schemes predict both mass and number concentrations of each hydrometeor with more than 25 microphysical transfers occurring among water vapor, liquid, and ice particles. These processes include growth of ice crystals by vapor deposition and riming, the aggregation of ice crystals, the formation of graupel and hail, the growth of graupel and hail by the collection of supercooled raindrops, the shedding of water drops from hail, the rapid growth of ice crystals in the presence of supercooled water, and the melting and sublimation of all forms of ice (Figure 2). With increasing computer power, explicit bin microphysical schemes have been developed for CRMs to study cloud–precipitation–aerosol interaction. One of the major differences between the two-moment bulk and spectral-bin microphysics schemes concerns the representation of the cloud particles size (Figure 3). Unlike assumed shape (such as

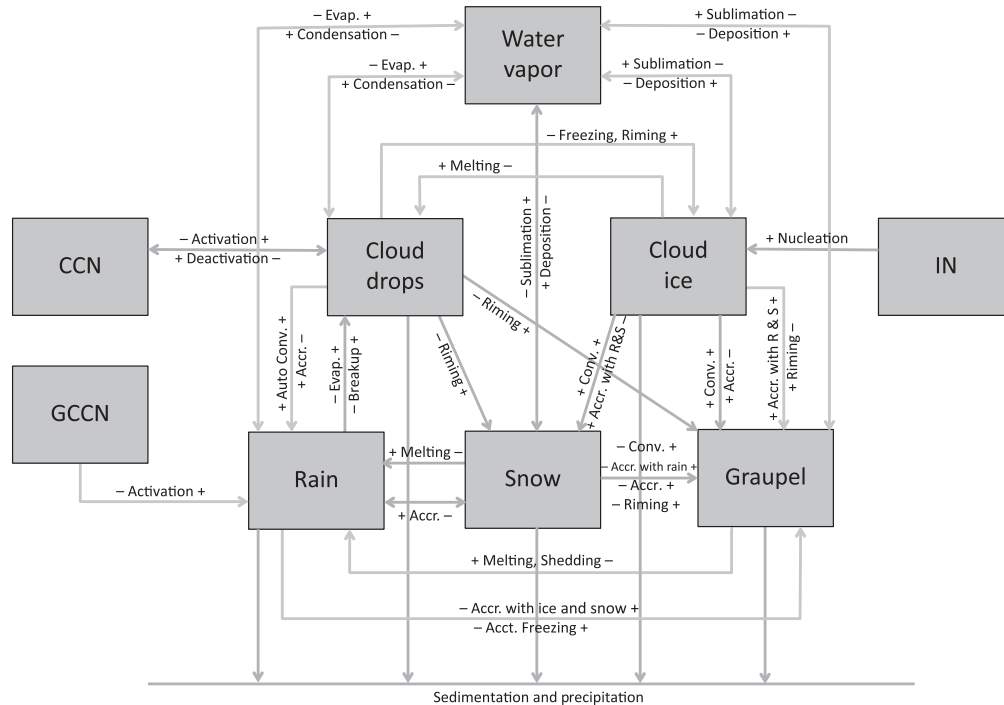


Figure 2 Schematics of the bulk microphysical processes in the typical two water and three-class ice scheme. Boxes represent the bulk classes of water and aerosol particles, and the arrows represent conversion pathways with plus and minus signs indicating direction of the named conversion process. In addition to prediction the mass of cloud water species (cloud drops, rain, cloud ice, snow, and graupel), the number of concentration of cloud water species is also predicted. Adapted from Cheng, C.-T., Wang, W.-C., Chen, J.-P., 2010. Simulation of the effects of increasing cloud condensation nuclei on mixed-phase clouds and precipitation of a front system. *Atmospheric Research* 96, 461–476. <http://dx.doi.org/10.1016/j.atmosres.2010.02.005>.

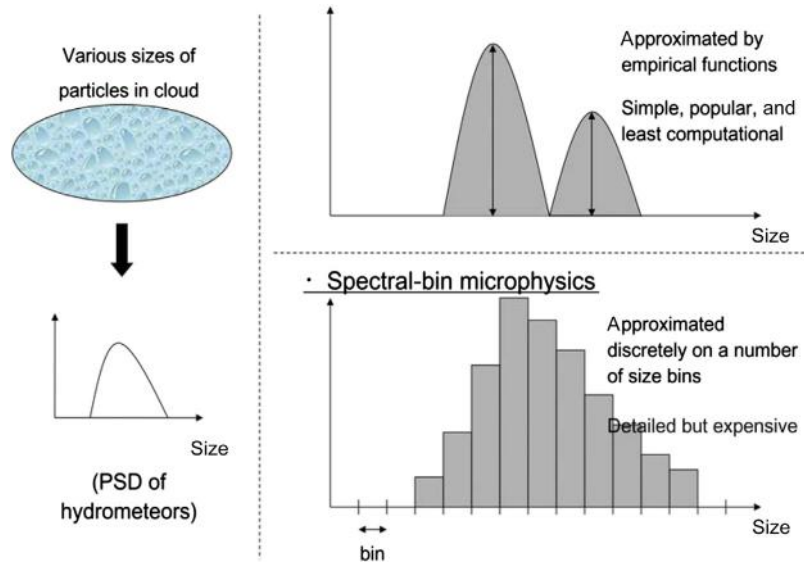


Figure 3 Numerical representation of PSD between bulk- and spectral-bin microphysics. Provided by Dr. Takamichi Iguchi. Adapted from Tao, W.-K., Chen, J.-P., Li, Z.-Q., Wang, C., Zhang, C.-D., 2012. The Impact of Aerosol on convective cloud and precipitation. *Reviews in Geophysics* 50, RG2001. doi:10.1029/2011RG000369.

gamma distributions) of particle size distributions (PSDs) in two-moment bulk microphysics, each type is described by a discretized size distribution containing as much as 30 or more size sectors (bins) (Figure 3). The formulation of the explicit microphysical processes is based on solving equations of mass advection by condensation growth and stochastic collision kinetics for the size distribution functions of water droplets (cloud droplets and raindrops together as one category). Ice particles are much more complicated due to their different shapes, so they are often classified into various growth habits such as columnar, platelike, dendrites, snowflakes, graupel, and frozen drops. Spectral-bin microphysics has also been used to improve the performance of bulk microphysics.

Results

CRMs have been used to examine the role of aerosols on CPSs. These modeling studies had many differences in terms of model grid configuration (two- or three-dimensional), domain sizes, grid spacings (150–3000 m), microphysics (i.e., two-moment bulk, spectral-bin), turbulence (simple or high order), radiation (with or without radiation effect), lateral boundary conditions (i.e., closed, radiative open, or cyclic), types of CPSs (isolated convection, tropical, or midlatitude squall lines), and model integration time (e.g., 2.5–48 h). Among the cloud-resolving modeling studies, the most striking difference is that cumulative precipitation can either increase or decrease in response to higher concentrations of CCN.

CCN Effect on Cloud Droplets Spectra

It is commonly believed that clouds in a clean environment (low CCN concentrations) produce fewer cloud droplets with

larger sizes due to greater condensational and collectional growth, leading to a broader size spectrum in comparison to a polluted environment (high CCN concentrations). Figure 4 shows the CRM with spectral-bin microphysics simulated PSDs of cloud droplets under low and high CCN concentrations for an oceanic and a continental convective case. Smaller cloud droplets with narrow PSDs were well predicted under a polluted environment (2520 of N_{CCN}), while larger cloud droplets with wider PSDs were predicted under a clean environment (600 of N_{CCN}). These basic features are well predicted by CRM simulations with either two-moment or spectral-bin microphysics in good agreement with the hypothesis and observations (Pruppacher and Klett, 1997), indicating that the cloud microphysics depends strongly on cloud–aerosol interactions. The broader or narrower PSDs could have an impact on the drizzle processes; i.e., the smaller cloud droplets would reduce the chance to form raindrops from cloud drop coalescence. Figure 4 also shows that cloud droplets with the diameter greater than 40 μm are not present in a polluted environment.

Impact of CCN on Precipitation

CRMs have been used to quantify the aerosol impact on rainfall associated with deep CPSs under different environments as well as different types of convective systems. Warm-phase rain suppression in the high CCN concentration (i.e., polluted environment) is evident in almost all case studies but only within the first hour during CRM integrations (Figure 5). In clean environments, warm-rain processes are initiated earlier than polluted environments, and rain reaches the ground early in all the clean-environment cases. This result clearly suggests that the microphysical processes dominate the initial stage of cloud evolution and rainfall. This is because that compared to the case in polluted conditions,

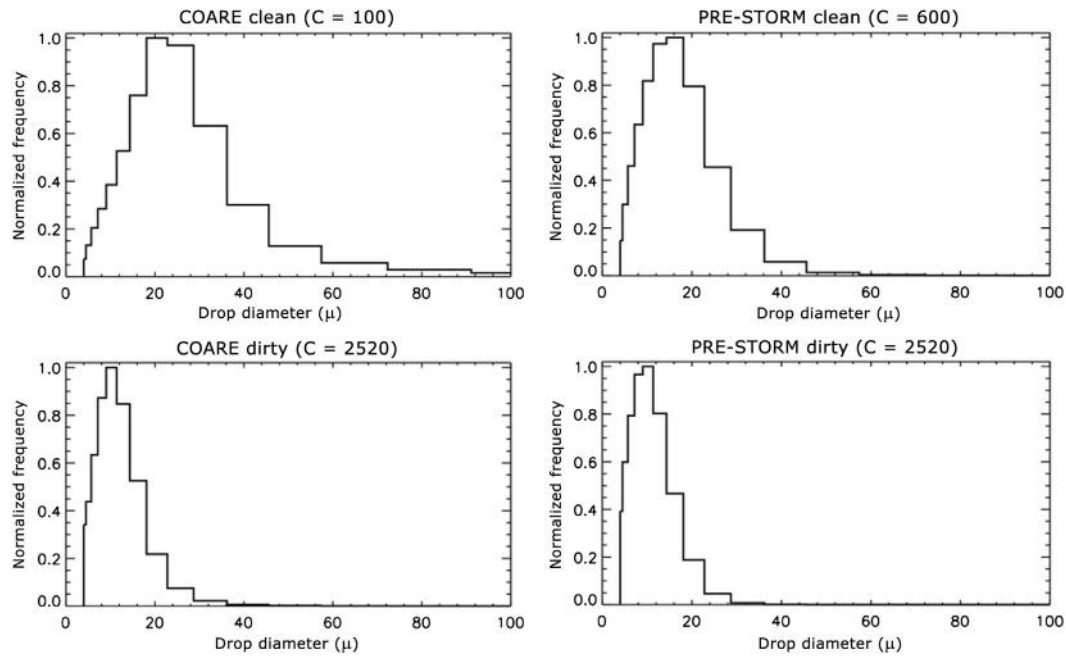


Figure 4 CRM-simulated cloud drop size distributions. The simulations are for two MCSs, a tropical oceanic squall system observed during TOGA COARE (*Tropical Ocean and Global Atmosphere Coupled Ocean-Atmosphere Response Experiment*, which occurred over the Pacific Ocean warm pool from November 1992 to February 1993), and a midlatitude continental squall system observed during PRE-STORM (*Preliminary Regional Experiment for STORM-Central*, which occurred in Kansas and Oklahoma during May–June 1985). It should be noted again that spectral-bin microphysics does not discretize cloud droplets and drizzle droplet classes. Liquid drop class is discretized in mass (size) bins, but continuous from tiny cloud droplets to larger raindrops. Adapted from Tao, W.-K., Li, X.X., Khain, A., Matsui, T., Lang, S., Simpson, J., 2007. The role of atmospheric aerosol concentration on deep convective precipitation: cloud-resolving model simulations. *Journal of Geophysical Research* 112, D24S18. <http://dx.doi.org/10.1029/2007JD008728>.

clouds in a clean environment (low CCN concentration) produce fewer droplets, larger sizes developed due to greater condensational and collectional growth, leading to a better chance to form raindrops from cloud-drop coagulation as seen in Figure 4.

When CRM simulations are integrated over the mature stage of the deep convection, the effect of the elevated CCN concentration on rainfall ranges from rain suppression in the midlatitude continental squall-line case, to little effect in the Florida sea-breeze case, and to rain enhancement in the tropical squall-line case. These results suggest that model simulations of the whole life cycle of CPSs should be examined in order to assess the impact of aerosols on precipitation processes in various forms of CPSs under their embedded environments. These results will also show the complexity of aerosol–cloud–precipitation interactions within CPSs.

Another advantage of using CRM to study aerosol–precipitation interaction simulations is to identify the factors that determine how aerosol can exert an influence on rainfall. A scheme was proposed to classify the aerosol effects on clouds and cloud systems under different moist environments (Figure 6). It shows that if the air relative humidity is high (low), then the condensation gain (loss) is large and the condensation loss (gain) is low, that could lead to an increase (decrease) of precipitation. In addition, the CRM results show that the aerosol impact on precipitation depends on the types of CPSs. A decrease in the precipitation with an increase in aerosol concentration usually occurs for isolated cumulus

clouds, and cloud systems develop within a relatively dry environment and/or within large wind shear or stratocumulus clouds. The CRM simulations also found that the increase in CCN concentrations tends to suppress precipitation under strong wind shear but enhance precipitation under weak wind shear. On the other hand, an increase in precipitation with an increase in CCN concentrations often occurs for clouds forming in a moist environment, such as coastal zones or within cloud ensembles, tropical squall line.

Physical Processes of CCN Effect on Precipitation

The CRM results have been used to examine the physical processes that could lead to aerosol-induced changes in precipitation. In general, three mechanisms were proposed to explain the *enhancement of precipitation by changing (either increasing or decreasing) the aerosol concentration*. The first mechanism is the stronger updrafts/downdrafts resulting from enhanced latent heat release, when CCN suppress warm-rain formation, and thus retains more supercooled liquid water to be frozen above 0 °C isotherm levels. This effect could be termed as the *latent heat – dynamic effect* (see Figure 7). The CRM results found that for cases having enhanced precipitation with a high CCN concentration, the clouds are usually associated with stronger updrafts/downdrafts as well as stronger convergence in the boundary layer, which provides for a better chance to trigger the secondary convections and prolong the lifetime of the CPSs.

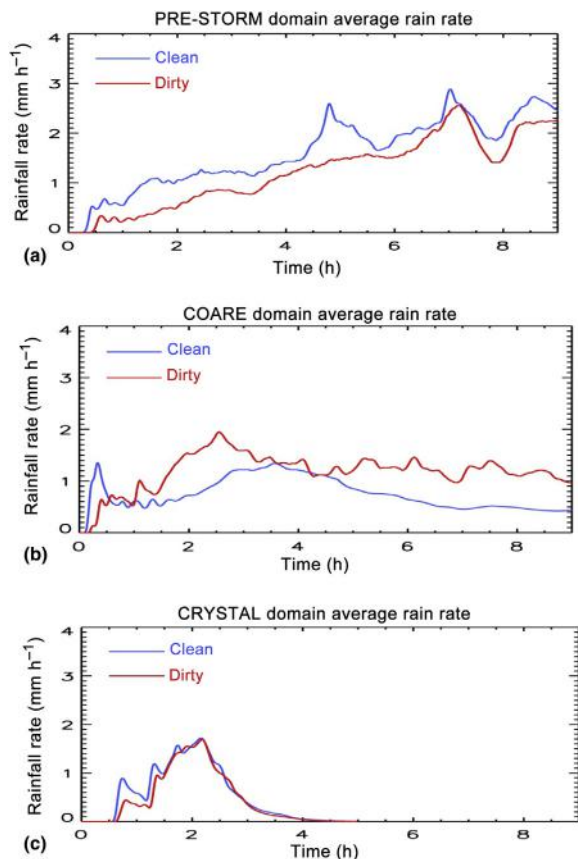


Figure 5 Time sequences of CRM estimated domain mean surface rainfall rate (mm h^{-1}) for (a) PRE-STORM, (b) TOGA COARE, and (c) CRYSTAL cases. The solid and dashed line is for clean and dirty condition, respectively. Adapted from Tao, W.-K., Li, X.X., Khain, A., Matsui, T., Lang, S., Simpson, J., 2007. The role of atmospheric aerosol concentration on deep convective precipitation: cloud-resolving model simulations. *Journal of Geophysical Research* 112, D24S18. <http://dx.doi.org/10.1029/2007JD008728>.

The second mechanism is the stronger evaporative cooling from a large number of concentrations of smaller raindrops under a high CCN-concentration condition. Rainfall evaporative cooling creates cold-air domes (domes?) under the stratiform rain area, denoted as cool pool. Enhanced evaporative cooling strengthens the near-surface cool pool. When interacting with the lower-level wind shear, colder cool pool induces the stronger low-level convergence, producing more vigorous convection that ultimately leads to enhanced surface precipitation. This positive feedback mechanism could be termed as *the cool pool effect*. It appears to be frequently occurring in the oceanic convective cases, in which the evaporative cooling in the lower troposphere is stronger for the polluted scenario compared to the clean scenario (Figure 8). Note that the stronger evaporative cooling occurs in the developing stage of the organized convective system.

The third mechanism is the CCN effect on ice microphysics. When CCN concentrations are elevated, cloud droplets tend to have lower collision efficiency due to smaller sizes. This reduces accretion and autoconversion, leading to less warm-rain process (Effect-A). Simultaneously, smaller cloud droplets, aloft in a mixed-phase zone, reduce collection efficiency with snow aggregates or ice crystals that reduce the riming process. If this process dominates, it may eventually reduce the cold-rain process (Effect-B). On the other hand, more supercooled cloud droplets aloft can enhance the riming process, if it does not lower collection efficiency (Effect-C). Thus, depending on changes in collection efficiency for the riming process, elevated CCN concentrations may lead to either less cold-rain (Effect-B) or a more cold-rain scenario (Effect-C). Regardless of the riming effect, more ice aloft due to enhanced supercooled cloud droplets can generally increase the amount of total ice-phased hydrometers, leading to a more cold-rain process (Effect-D). These processes can be termed as ice-microphysics effect (Figure 9). Total precipitation amount could be enhanced or reduced, depending on the

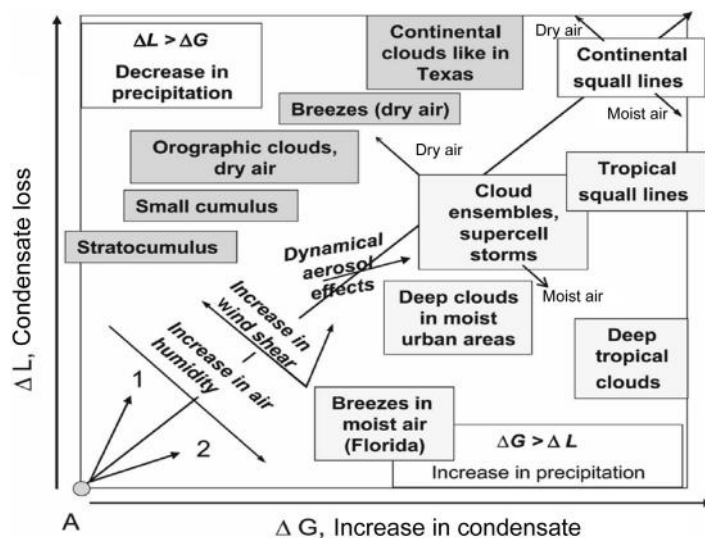


Figure 6 A schematic diagram of the aerosol effects on clouds and cloud systems of different types. The zone above the diagonal corresponds to a decrease in precipitation with the aerosol concentration. The zone below the diagonal corresponds to an increase in precipitation with the increase in the aerosol particle concentration. Adapted from Khain, A., BenMoshe, N., Pokrovsky, A., 2008. Factors determining the impact of aerosols on surface precipitation from clouds: an attempt at classification. *Journal of Atmospheric Sciences* 65 (6), 1721–1748.

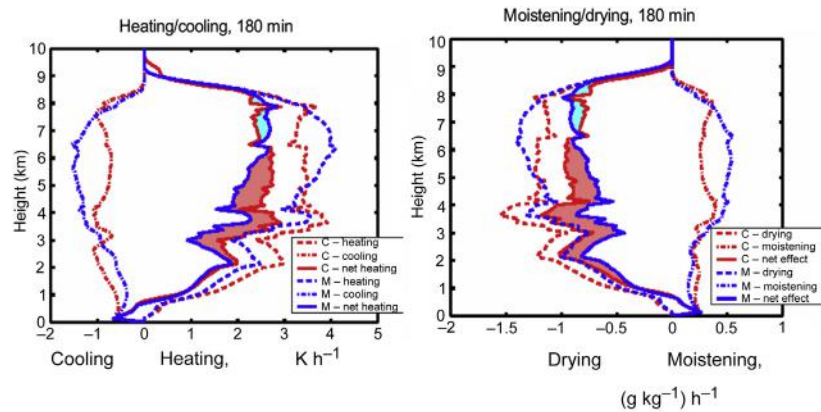


Figure 7 Vertical profiles of horizontal averaged heating and cooling (left) and moistening and drying (right) in the hail storm in the Southern Germany calculated for low (blue curves) and high (red curves) aerosol concentrations during 180 min. Areas marked pink/light blue denote zones where net heating and drying are larger/smaller in the high aerosol concentration case, respectively. Adapted from Khain, A., BenMoshe, N., Pokrovsky, A., 2008. Factors determining the impact of aerosols on surface precipitation from clouds: an attempt at classification. *Journal of Atmospheric Sciences* 65 (5), 1721–1748.

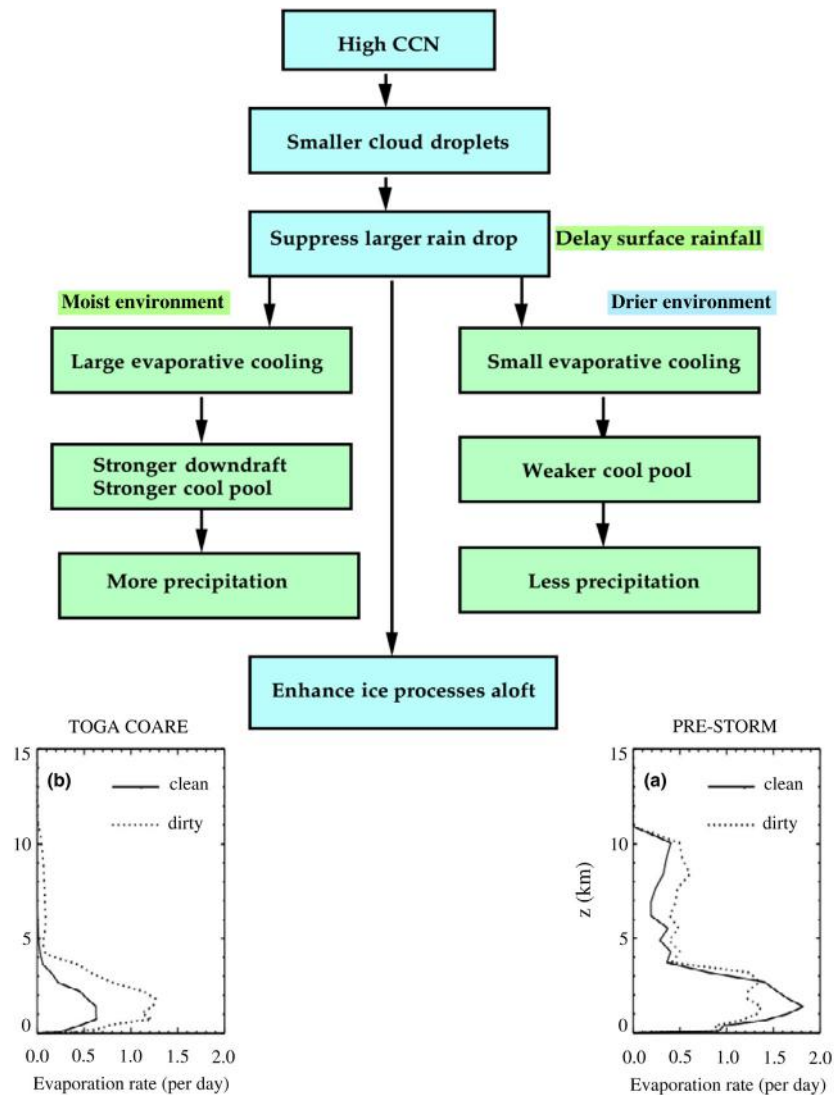


Figure 8 Schematic diagram showing the physical processes that lead to either enhancement (oceanic convective case) or suppression (midlatitude continental convective case) of precipitation in a dirty environment. Their respective evaporative cooling simulated by the clean and dirty cases are also shown. Note that the differences in cooling occurred in the early stage of model simulations. Adapted from Tao, W.-K., Li, X.X., Khain, A., Matsui, T., Lang, S., Simpson, J., 2007. The role of atmospheric aerosol concentration on deep convective precipitation: cloud-resolving model simulations. *Journal of Geophysical Research* 112, D24S18. <http://dx.doi.org/10.1029/2007JD008728>.

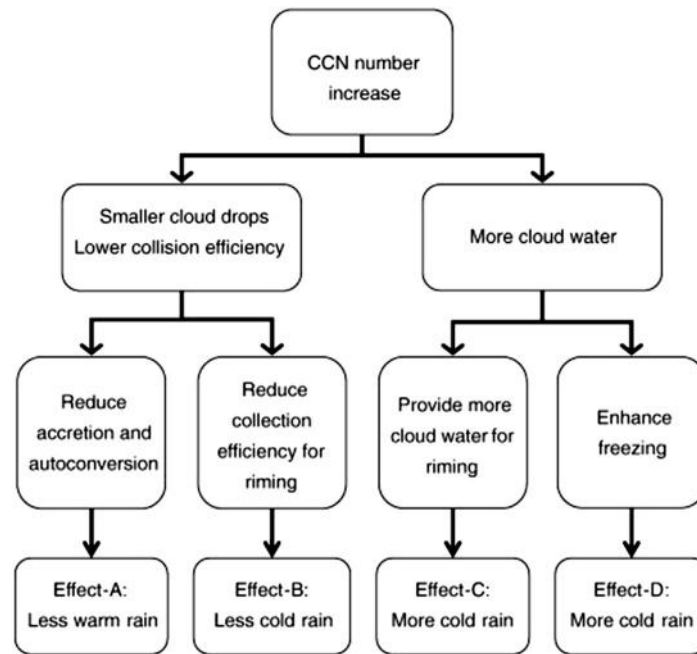


Figure 9 Schematic diagram of the major CCN effects on warm rain and cold rain. Adapted from Cheng, C.-T., Wang, W.-C., Chen, J.-P., 2010. Simulation of the effects of increasing cloud condensation nuclei on mixed-phase clouds and precipitation of a front system. *Atmospheric Research* 96, 461–476. <http://dx.doi.org/10.1016/j.atmosres.2010.02.005>.

domination of these four microphysics paths in different types of CPSs.

CCN Effect on Convective Precipitation Using Nested Cloud/Regional-Scale Model

Regional-scale limited-area models (i.e., Regional Atmospheric Modeling System, Weather Research and Forecasting Model, and Pennsylvania State University–National Center for Atmospheric Research Mesoscale Model) with fine resolution (utilizing interactive nesting technique) have also been used to study the impact of aerosols on convective precipitation events associated with the Florida sea breeze, urban heat island effect, and tropical cyclone (TC)/hurricane. An advantage of this approach is that the model initial and lateral boundary conditions are provided by large-scale analyses with realistic meteorological fields, and model simulations can be conducted with realistic terrain and heterogeneous land-surface characteristics.

The regional-scale model has been used to examine the sensitivity tests of aerosol concentrations on a convective storm over the peninsula of Florida. The results found that the increase in CCN concentration led to convective invigoration and the formation of stronger secondary clouds. Simulations of rain events over the whole peninsula for this day showed significant invigoration of squall lines. There was an increase in precipitation rate and precipitation amount for a squall line that formed in the vicinity of the east coast of Florida. At the same time, continental CCN concentrations resulted in a 5% reduction in precipitation over the whole computational domain (containing a significant fraction of Florida) vs maritime values. On the other hand, the regional-scale model

results also showed that different combinations of CCN, GCCN, and IN result in different amounts and temporal patterns of cloud water/ice contents and rainfall. The high CCN reduces cumulative precipitation by 22% compared to low CCN. Also noted were, high-GCCN and IN enhanced surface precipitation for the first 6 h of integration due to the initial broadening of the cloud droplet spectra. However, the total (12-h integration) accumulated precipitation was greatest for the clean (low CCN, GCCN, and IN) case. This could be explained by a rapid wet deposition of GCCN for the first 6 h of integration.

High aerosol concentrations in urban environments could affect precipitation variability by providing an enhanced source of CCN. The regional model has been used to examine the sensitivity of urban-induced convective clouds over and downwind of St. Louis, MO. The results indicate that downwind convergence (dynamic processes) induced by urban land cover appears to be the dominant factor in determining whether or not moist convection actually develops downwind of St. Louis. Once moist convection is initiated, urban-enhanced aerosols play a major role in determining the microphysical and dynamical characteristics of convective storms, particularly when background aerosol concentrations are low (Figure 10). Complicated relationships and feedbacks between microphysical and dynamical processes obscure generalized understandings (i.e., a linear relationship) of urban-enhanced aerosol effects on precipitation.

Note that the regional-scale models could explicitly represent locally induced mesoscale dynamics (i.e., sea-breeze convergence and urban heat island convergence). This is important because model-simulated cool pool can interact with those circulations, introducing another level of dynamic

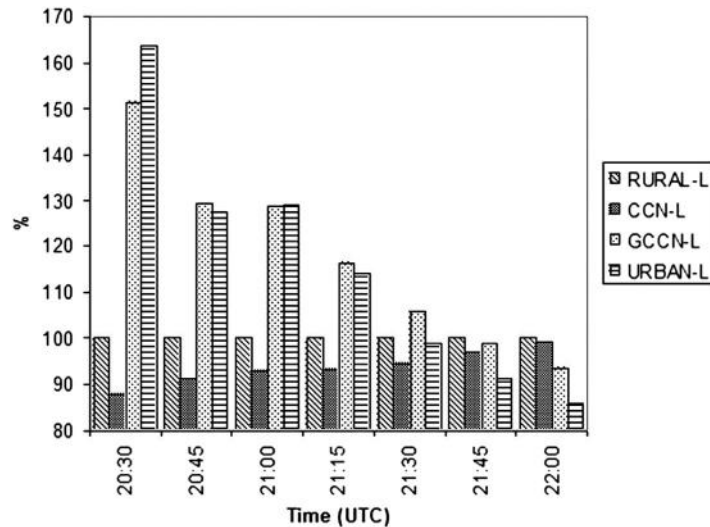


Figure 10 Time series of the accumulated volumetric precipitation in the downwind region expressed as a percentage of the RURAL-clean background simulation. RURAL-L stands for the simulation with clean CCN and GCCN background aerosol; CCN-L for the effects of the increase of CCN concentrations only over clean background concentration, GCCN-L for the effects of polluted GCCN concentrations only relative clean background concentration, URBAN-L stands for the effects of CCN and GCCN concentration relative to a clean background aerosol. Adapted from Van den Heever, S., Carrio, G.C., Cotton, W.R., DeMott, P.J., Prenni, A.J., 2006. Impacts of nucleating aerosol on Florida storms. Part I: mesoscale simulations. *Journal of Atmospheric Sciences* 63, 1752–1775.

complexity. The regional-scale model also showed that the response of convective rainfall to urban-enhanced aerosols becomes stronger when the background aerosol concentrations are low.

The regional-scale model has also been used to examine the impact of dust in the Saharan Air Layer (SAL) acting as CCN on the evolution of a TC. The results showed that the dust in the SAL as CCN could affect the simulated TC intensity by 22 hPa depending on CCN concentrations (Figure 11). The high CCN concentration could weaken the TC intensity. It can also affect eyewall development directly through the release of latent heat when activated and subsequent growth of cloud droplets and

indirectly through modulating rain band development. The development of rain bands tended to promote latent heat release away from the eyewall, block the surface inflow and enhance cold pools. However, results are no longer monotonic with increasing CCN. The impact of CCN on storm intensity was very sensitive to the background GCCN vertical profile and presumably other environmental factors. The regional-scale model was also used to investigate the potential impact of aerosols ingested into Katrina's circulation during its passage through the Gulf of Mexico on Hurricane Katrina's (2005) structure and intensity. The result was shown that continental aerosols invigorated convection largely at TC periphery, which led to its weakening prior to landfall. The minimum pressure increased by 15 hPa, and the maximum velocity decreased up to 15 m s^{-1} .

IN Effect on Precipitation Processes

Some types of aerosol can be transported by convective updrafts from PBL to middle- and upper-troposphere and can be served as IN. Only a few CRM-based studies have been identified for the effects of the IN on precipitation processes because of the limited understanding of ice formation and IN precursors (Pruppacher and Klett, 1997). *The CRM results indicated that the increase of IN concentration and, thus, heterogeneous ice nucleation would result in an enhancement in updraft due to latent heat release from added diffusive growth of increased ice crystals. Such an effect was also identified to enhance the homogeneous nucleation, i.e., to make CCN more effective in influencing the ice nucleation. Since the dominant mechanism in providing ice particles for cirrus anvil in the modeled case is still the homogeneous nucleation, the increase of IN concentration would lead to enlarged anvil coverage and increased stratiform precipitation. This finding*

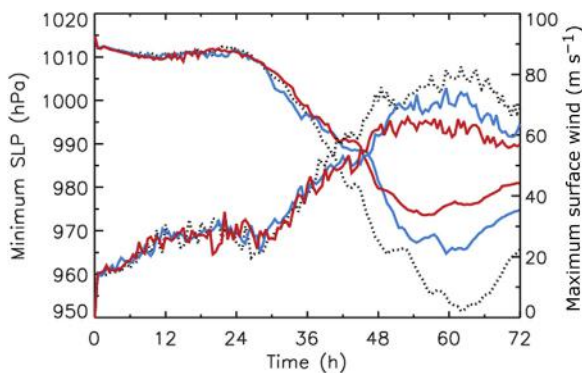


Figure 11 Temporal evolutions of the minimum sea-level pressure MSLP and maximum surface wind for 'clean' (dotted line), 'polluted' (1000 cm^{-3} , thin solid line), and 'double' (2000 cm^{-3} , thick solid line). Adapted from Zhang, H., McFarquhar, G.M., Saleeby, S.M., Cotton, W.R., 2007. Impacts of Saharan dust as CCN on the evolution of an idealized tropical cyclone. *Geophysical Research Letters* 34, L14812. doi:10.1029/2007GL029876.

could imply that a linkage from IN concentration to anvil particle concentration, coverage, and therefore radiation effects. Their results could also have potential climate consequences through cloud-radiative effect.

Conclusion

Most of the CRM-based studies highlighted that high CCN concentrations could suppress initial warm-phase precipitation processes. However, high CCN concentrations could also enhance cold-phase precipitation processes depending on the surrounding environments potentially through interacting with latent heat, cool pool, and ice-microphysics effects. These simulation results show that the nonlinear interactions between aerosols and CPSs can occur. Indeed, fully understanding the interactive processes between aerosol, IN, and CPSs is still lacking at present due to limited observations. More case studies are required to further investigate the aerosol impact on different CPS events under different thermodynamic environments. In almost all CRM or regional-model studies, idealized, uniform, or single composite CCN concentrations were employed in the CRM simulations, and IN concentrations were not linked to the aerosol concentrations. A horizontally uniform distribution of CCN was also used in the regional-scale-modeling studies. A nonhomogeneous CCN distribution, consistent with the nonhomogeneous initial meteorological conditions, will be required to assess aerosol–precipitation interactions using regional-scale models in the future. In addition to IN and GCCN, the chemistry of CCN needs to be considered in future modeling of aerosol–precipitation interactions.

In addition, many CRM studies did not often compare model results with observed cloud structures, organization, radar reflectivity, and rainfall. It may require major field campaigns to gather the data necessary to both initialize and validate the models in terms of meteorological, cloud–precipitation, and aerosol parameters using *in situ* cloud property observations, radar, lidar, and microwave remote-sensing data. Although CRM-simulated results can provide valuable quantitative estimates of the microphysics effects of aerosols, CRMs can only simulate clouds and cloud systems typically over a relatively small domain, which does not feedback to general circulations. Close collaboration between the global and CRM communities is needed in order to expand the CRM results to a regional and global perspective.

CRMs together with observations should be a promising approach for studying the physical processes associated with aerosol–cloud–precipitation interactions. Modern CRMs operate with realistic, albeit not complete, microphysical parameterizations and simulate the evolution, structure, and life cycles of CPSs. Because of the range of scales resolved by modern CRMs, high-resolution observations from space-based remote sensing is becoming an ever more necessary part of the model validation through the state-of-art multiinstrument satellite simulators (Matsui et al., 2009).

See also: **Air Sea Interactions:** Momentum, Heat, and Vapor Fluxes. **Boundary Layer (Atmospheric) and Air Pollution:** Modeling and Parameterization. **Clouds and Fog:** Cloud Microphysics. **Mesoscale Meteorology:** Mesoscale Convective Systems; Overview. **Numerical Models:** Mesoscale Atmospheric Modeling; Methods; Parameterization of Physical Processes: Clouds. **Observations Platforms:** Balloons; Radiosondes. **Radar:** Polarimetric Doppler Weather Radar; Precipitation Radar. **Radiation Transfer in the Atmosphere:** Cloud-Radiative Processes. **Satellites and Satellite Remote Sensing:** Precipitation. **Thermodynamics:** Saturated Adiabatic Processes. **Tropical Cyclones and Hurricanes:** Overview and Theory. **Turbulence and Mixing:** Overview.

Further Reading

- Convective Precipitation Systems Including Thunderstorms, Mesoscale Convective Systems
CRM Aerosol-Deep Convection Including Review
Cheng, C.-T., Wang, W.-C., Chen, J.-P., 2010. Simulation of the effects of increasing cloud condensation nuclei on mixed-phase clouds and precipitation of a front system. *Atmospheric Research* 96, 461–476. <http://dx.doi.org/10.1016/j.atmosres.2010.02.005>.
DeMott, P.J., Prenni, A.J., Liu, X., Kreidenweis, S.M., Petters, M.D., Twohy, C.H., Richardson, M.S., Eidhammer, T., Rogers, D.C., 2010. Predicting global atmospheric ice nuclei distributions and their impacts on climate. *PNAS* 107, 11217–11222 published ahead of print June 7, 2010. <http://dx.doi.org/10.1073/pnas.0910818107>.
Ekman, A.M.L., Engström, A., Wang, C., 2007. The effect of aerosol composition and concentration on the development and anvil properties of a continental deep convective cloud. *Quarterly Journal of the Royal Meteorological Society* 133B (627), 1439–1452.
Houze Jr., R.A., 1993. *Cloud Dynamics*. Academic Press, San Diego, 573 pp.
Khain, A., BenMoshe, N., Pokrovsky, A., 2008. Factors determining the impact of aerosols on surface precipitation from clouds: an attempt at classification. *Journal of Atmospheric Sciences* 65 (6), 1721–1748.
Matsui, T., Zeng, X., Tao, W.-K., Masunaga, H., Olson, W., Lang, S., 2009. Evaluation of long-term cloud-resolving model simulations using satellite radiance observations and multifrequency satellite simulators. *Journal of Atmospheric and Oceanic Technology* 26, 1261–1274.
Microphysics of Clouds and Precipitation and Precursors of Cloud Droplets and Ice Crystals
Pruppacher, H.R., Klett, J.D., 1997. *Microphysics of Clouds and Precipitation*. D. Reidel, Norwell, MA, 954 pp.
Satellite View of Precipitation Microphysics
Series of CRM Studies
Tao, W.-K., Moncrieff, M., 2009. Multi-scale cloud-system modeling. *Reviews in Geophysics* 47 (4), RG4002. <http://dx.doi.org/10.1029/2008RG000276>.
Tao, W.-K., Li, X.X., Khain, A., Matsui, T., Lang, S., Simpson, J., 2007. The role of atmospheric aerosol concentration on deep convective precipitation: cloud-resolving model simulations. *Journal of Geophysical Research* 112, D24S18. <http://dx.doi.org/10.1029/2007JD008728>.
Tao, W.-K., Chen, J.-P., Li, Z.-Q., Wang, C., Zhang, C.-D., 2012. The Impact of Aerosol on convective cloud and precipitation. *Reviews in Geophysics* 50, RG2001. <http://dx.doi.org/10.1029/2011RG000369>.
Uncertainties: Bulk Approach, Chemistry, Large-Scale Feedback, Parameterization Of In, Organic Carbon
Van den Heever, S., Carrio, G.C., Cotton, W.R., DeMott, P.J., Prenni, A.J., 2006. Impacts of nucleating aerosol on Florida storms. Part I: mesoscale simulations. *Journal of Atmospheric Sciences* 63, 1752–1775.
Zhang, H., McFarquhar, G.M., Saleeby, S.M., Cotton, W.R., 2007. Impacts of Saharan dust as CCN on the evolution of an idealized tropical cyclone. *Geophysical Research Letters* 34, L14812. <http://dx.doi.org/10.1029/2007GL029876>.

Large-Eddy Simulation

C-H Moeng and PP Sullivan, National Center for Atmospheric Research, Boulder, CO, USA

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

Large-eddy simulation (LES) is a numerical technique for integrating spatially filtered equations of motion that describe high-Reynolds number time-evolving, three-dimensional turbulence. The spatial filtering cuts off the high frequency or small-scale part of the turbulence spectrum. Virtual turbulence generated by LES has been shown to be a surrogate for measurements of turbulent flow fields. LES is widely used for turbulence research and for applications. This article describes the LES technique, and reviews its contributions to and use in atmospheric sciences.

Introduction

Turbulence consists of a three-dimensional chaotic motion that spans a wide range of scales of motion, which increases with the Reynolds number (*see Turbulence and Mixing: Overview*). A numerical integration of the full (unfiltered) equations (i.e., the Navier–Stokes equations) that explicitly calculates all scales of turbulent motion is known as direct numerical simulation (DNS). However, DNS can simulate only low-to-medium Reynolds-number turbulence (where the range of scales is not very broad) with today's computer power. Low-to-medium Reynolds number flows are typical of wind tunnel laboratory experiments. The largest DNS performed today uses $\sim 10^{10}$ grid points, which is still insufficient to simulate turbulence with a very wide range of scales. As an example, consider the atmospheric planetary boundary layer (PBL). The largest turbulent eddies in the PBL are on the order of kilometers and the smallest on the order of millimeters; the entire scale range spans more than six orders of magnitude. To numerically integrate the full Navier–Stokes equations for a turbulent PBL requires at least 10^{18} numerical grid points (i.e., 10^6 in all three directions). This is far beyond today's computing capacity or that in the foreseeable future.

Given this limitation, only a portion of the scale range can be explicitly resolved. The obvious choice is to resolve just the most important scales of the flow of interest and approximate the other scales. This is the philosophy behind large-eddy simulation (LES). For PBL turbulence, as an example, the most important scales (for most meteorological applications) are large eddies which contain most of the turbulent kinetic energy (TKE) (thus called energy-containing eddies) and are responsible for the majority of the turbulent transport. A simulation that explicitly calculates (or resolves) large eddies while approximately representing the effects of smaller ones is LES (Wyngaard, 1984, 2010). As the grid resolution of LES becomes finer, a wider range of turbulent eddies is resolved, less are parameterized, and LES-generated flows become more representative of the entire flow field. LES is a compromise between DNS (in which all turbulent fluctuations are resolved) and traditional Reynolds-averaging approach (in which all turbulent fluctuations are parameterized and only ensemble-averaged statistics are calculated).

The LES technique was developed by Jim Deardorff at the National Center for Atmospheric Research in the late 1960s.

His first LES calculation was performed using a computer that allowed for only $32 \times 32 \times 32$ (32 768) grid points. On today's machines, calculations with 10^6 – 10^7 grid points are a common practice and 10^9 – 10^{10} computations are possible on massively parallel machines. As computer power increases, the LES numerical technique will have a much broader application and more accurate solutions. LES has been developed and used extensively also by the engineering fluid dynamics community, but in this article we restrict our description to LESs of atmospheric flows, with an emphasis on the PBL. Most LES research by the PBL community focuses on applications that include buoyancy, rotation, rough surface, ocean waves, canopy, entrainment, radiation, and/or condensation.

The LES Technique

Governing Equations and Filtering Procedures

The Navier–Stokes equations for an incompressible fluid are

$$\frac{\partial u_i}{\partial t} = -\frac{\partial u_i u_j}{\partial x_j} + X_i - \frac{1}{\rho} \frac{\partial p}{\partial x_i} + \nu \frac{\partial^2 u_i}{\partial x_j^2} \quad [1]$$

where the velocity field u_i satisfies the continuity equations

$$\frac{\partial u_i}{\partial x_i} = 0. \quad [2]$$

In eqns [1] and [2], u_i are flow velocities in the three spatial directions, i.e., $i = 1$ and 2 for the horizontal directions and $i = 3$ for the vertical direction, X_i are the i th-component of body forces, ρ is the air density, p is the pressure fluctuation, ν is the kinematic viscosity of the fluid, t is time, and x_i are the spatial coordinates. For PBL applications, the major body forces are gravity and Coriolis forces (*see Dynamical Meteorology: Coriolis Force*) and hence $X_i = g_i \theta / T_0 - f \epsilon_{ij3} u_j$, where the gravitational acceleration g_i is nonzero only in the x_3 (or z) direction, θ is the virtual potential temperature, T_0 is the temperature of some reference state, and f is the Coriolis parameter. The body force X_i is obtained by expanding eqn [1] over a reference state of hydrostatic equilibrium and also using the Boussinesq approximation (*see Dynamical Meteorology: Primitive Equations*). A numerical integration of eqns [1] and [2] is called DNS. For LES the governing eqns [1] and [2] need to be spatially filtered.

Spatially filtered Navier–Stokes equations are derived by decomposing all dependent variables, for example the velocity field u_i , into $u_i = \tilde{u}_i - u_i''$ where $\tilde{u}_i$ is the filter-scale component and u_i'' is the subgrid scale (SGS) (or subfilter). The filter-scale or resolved-scale variable is defined as

$$\tilde{u}_i(x_i) = \iiint_{\text{volume}} u_i(x'_i) G(x_i - x'_i) dx'_i \quad [3]$$

where G is a three-dimensional (low-pass) filter function (typically, a Gaussian, top-hat, or sharp-wave-cutoff filter). A turbulent flow field obtained from a point sensor illustrates the filtering and decomposition process. In Figure 1, the red curve denotes the time varying total flow field. Application of the filter operator shown in eqn [3] to the total flow field yields a smoother field called the filter scale (or resolved scale) as indicated by the blue curve. The difference between the total and resolved-scale components is the subfilter scale or SGS, shown by the green curve. The SGS component is much smaller in magnitude and is of much higher frequency, compared to the resolved-scale component.

Applying the filtering procedure, term-by-term, to eqn [1] leads to the equations that govern large (resolved-scale) eddies:

$$\frac{\partial \tilde{u}_i}{\partial t} = -\frac{\partial \tilde{u}_i \tilde{u}_j}{\partial x_j} - \frac{\partial \tau_{ij}}{\partial x_j} + \frac{g_i}{T_0} \tilde{\theta} - f \varepsilon_{ij3} \tilde{u}_j - \frac{1}{\rho_0} \frac{\partial \tilde{p}}{\partial x_i} + \nu \frac{\partial^2 \tilde{u}_i}{\partial x_j^2}, \quad [4]$$

where the SGS stress is defined as $\tau_{ij} = \tilde{u}_i \tilde{u}_j - \tilde{u}_i \tilde{u}_j$. For geophysical turbulence, the last term (molecular viscosity) is negligibly small compared to the other terms and hence can be ignored.

So far in deriving eqn [4] for LES calculations, no approximations have been made. However, to solve eqn [4] we need an SGS model to describe the SGS stress τ_{ij} .

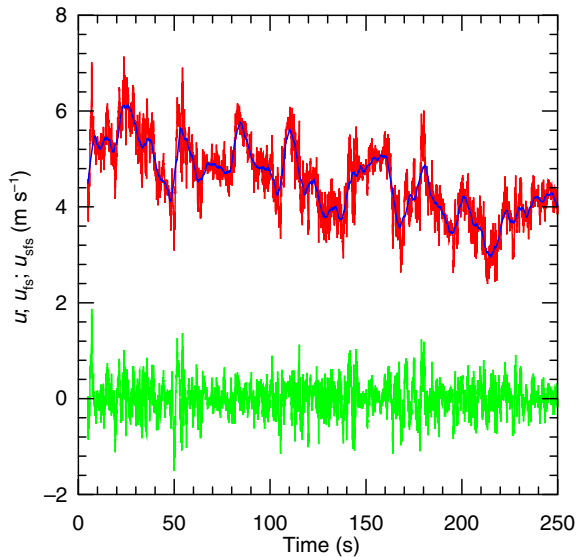


Figure 1 A sketch illustrating the filtering and decomposition procedure. The red curve shows the time evolution of a turbulent flow field obtained from a point sensor. Applying the filtering procedure yields its filter-scale field (blue curve) and subfilter-scale (or SGS) field (green curve).

Subgrid-Scale (SGS) Models

An SGS model relates the SGS stress (or SGS fluxes) to resolved-scale variables, allowing eqn [4] to be integrated. This is a closure problem. The closure assumption made in an SGS model creates the only uncertainty – other than numerical errors – to LES-generated turbulent flows. This uncertainty may become severe in regions where small eddies are dominant (or play an important role), for example near a rough-wall boundary and perhaps in the entrainment zone of the PBL. Nevertheless, in regions where large turbulent eddies play the major role, LES-generated flows (and their derived statistics such as variances and fluxes) have been shown to be insensitive to the SGS model. This is true in the bulk of the PBL where small turbulent eddies act mainly as passive motions, passing energy downscale toward dissipation. This is a turbulence process known as the energy cascade. Thus, a simple SGS model that dissipates energy properly is usually sufficient to represent the net effect of small eddies – for most meteorological applications – if the LES grid resolution is properly chosen.

The most widely used SGS models for the PBL are the Smagorinsky–Lilly (S–L) and Deardorff’s TKE models. They are similar in that both are based on SGS TKE budgets and both relate SGS stresses to resolved-scale strain tensors as

$$\tau_{ij} = -2K_M S_{ij} \quad [5]$$

where the strain tensor is $S_{ij} = (\partial \tilde{u}_i / \partial x_j + \partial \tilde{u}_j / \partial x_i) / 2$. SGS heat fluxes are similarly related to local gradients of the resolved temperature field as

$$\tau_{\theta i} = -K_H \frac{\partial \tilde{\theta}}{\partial x_i} \quad [6]$$

The S–L model

Without the buoyancy effect the SGS eddy viscosity K_M and diffusivity K_H are expressed as

$$K_M = (c_S \Delta_S)^2 S, \quad [7]$$

and

$$K_H = \frac{K_M}{P_r} \quad [8]$$

where c_S is the Smagorinsky constant, Δ_S is a filtered length scale often taken to be proportional to the grid size, S is the magnitude of the strain tensor, $S = (2 S_{ij} S_{ij})^{1/2}$, and P_r is the SGS Prandtl number. One of the most important features of the S–L model is that the SGS fluxes are nonlinear functions of the resolved strain rate, a crucial difference from the viscous (molecular) stress–strain relationship. For PBL applications, the Smagorinsky constant c_S is often set to 0.18–0.25 and the SGS Prandtl number to $\sim 1/3$ to satisfy Kolmogorov inertial-subrange theory.

To include the buoyancy effect, the K_M expression in the original Smagorinsky model is modified to depend on local Richardson number R_i (the ratio of buoyancy to shear production terms of the TKE budget),

$$K_M = (c_S \Delta_S)^2 S \left(1 - \frac{R_i}{R_{ic}} \right)^n \quad [9]$$

where R_{ic} is the critical Richardson number often set between 0.2 and 0.4, and $n = 1/2$ is often used. When the local Richardson number reaches the critical value, turbulence within that grid cell vanishes and the eddy viscosity is shut off.

The Deardorff TKE model

Deardorff extended the S–L eddy viscosity model to include the full SGS TKE e :

$$\frac{\partial e}{\partial t} = -\frac{\partial \tilde{u}_j e}{\partial x_j} - \frac{\partial (\tilde{u}_j' e + \tilde{u}_j' p'')}{\partial x_j} - \tilde{u}_i' \tilde{u}_j' \frac{\partial \tilde{u}_i}{\partial x_j} + \frac{g}{T_0} \tilde{u}'' \theta'' - \varepsilon \quad [10]$$

and then relates K_M and K_H to SGS TKE as

$$K_M = c_K l \sqrt{e} \quad [11]$$

and

$$K_H = [1 + (2l/\Delta_S)] K_M \quad [12]$$

where c_K is a diffusion coefficient to be determined and l is another SGS length scale, taken as the minimum of the stability-corrected scale and the grid scale:

$$l = \min[0.76 \sqrt{e/N^2}; \Delta_S] \quad [13]$$

where $N = \sqrt{(g/T_0)(\partial\theta/\partial z)}$ is the Brunt–Vaisala frequency.

The terms on the right-hand side of eqn [10] represent, in order, advection of SGS TKE by the resolved-scale motion, turbulence and pressure transports, local shear production (i.e., nonlinear scrambling), local buoyancy production or consumption, and molecular dissipation. In solving eqn [10], the transport term is usually approximated as

$$\tilde{u}_j' e + \tilde{u}_j' p'' = -2K_M \frac{\partial e}{\partial x_j} \quad [14]$$

and the molecular dissipation rate as

$$\varepsilon = c_e e^{3/2} / l \quad [15]$$

where c_e is a dissipation coefficient.

Under the assumption that all SGS motions lie within the inertial subrange, the SGS parameters c_K and c_e can be derived assuming that the SGS motions are isotropic and the energy spectrum has a $-5/3$ spectral slope (Moeng and Wyngaard, 1988). Commonly used values for PBL applications are $c_K \sim 0.10$, and $c_e \sim 0.19 + 0.74 l/\Delta_S$. With these model parameters, LESs are, in a way, forced – in an ensemble-mean sense – to drain energy at a rate sufficient to produce a $-5/3$ spectral slope near the filter cutoff scale.

These two SGS models are closely related. The S–L model can be derived by keeping just the last three terms on the right-hand side in eqn [10]. The SGS constants of the two models are related as $c_S^4 = c_K^3/c_e$ (if the stability correction of the SGS length scale is neglected).

Deficiency of the eddy-viscosity SGS models

The above SGS models are based on ensemble average concepts but are used inside of LES on an instantaneous basis, i.e., to represent SGS effects at every grid point and at every time step. Laboratory studies and DNS provide evidence that small-scale turbulent motions are anisotropic and intermittent and that locally the energy transfer can either be forward scatter

(from large to small scales) or backscatter (from small to large scales), which causes deviations from the equilibrium $-5/3$ law. Eddy-viscosity SGS models also assume that SGS stresses and strains are perfectly aligned (eqn [5]), and hence the local dissipation rate $\varepsilon = -\tau_{ij} S_{ij}$ is always positive thus preventing backscatter of energy. These deficiencies of eddy-viscosity models have motivated continued development of new SGS models, including (1) stochastic models where a random field is imposed at the SGS level thus permitting a backscatter of energy, (2) dynamic models where the Smagorinsky coefficient is dynamically predicted using a resolved field filtered at two different scales, and (3) velocity estimation models that attempt to model the SGS velocity fluctuations u_i' instead of SGS stresses τ_{ij} .

The deficiency of the S–L and Deardorff SGS models is most evident in the surface layer of the PBL. Based on a large body of measurements and scaling arguments, the vertical gradients of the mean fields usually obey Monin–Obukhov (M–O) similarity theory (see **Boundary Layer (Atmospheric) and Air Pollution: Surface Layer**). However, LESs using the S–L or Deardorff SGS model fail to reproduce the vertical profiles of the mean fields predicted by M–O theory, particularly for shear-driven and stable PBLs. One reason for this shortcoming is that near the surface, small eddies dominate so that almost all of the turbulent eddies are SGS in LES; few motions are actually resolved. This deficiency has been improved somewhat using SGS models that include either a backscatter effect or a contribution from the mean shear near the surface (Sullivan et al., 1994).

Numerical Setup, Methods, and Boundary Conditions

The choice of LES grid and domain size depends on the physical flow of interest and the computer capability. LES differs from other meteorological models in that its resolved-scale (or grid-scale) turbulent motion has about the same characteristic scale in all three directions, and hence requires a grid mesh that is close to isotropic. Most computers today can easily perform an LES of about $100 \times 100 \times 100$ grid points. From these grid points, an LES domain is then chosen to resolve several largest (dominant) turbulent eddies and at the same time resolve eddies as small as possible into the inertial-subrange scales. For example, for a convective PBL with 1 km depth, a $5 \text{ km} \times 5 \text{ km} \times 2 \text{ km}$ domain of LES with $100 \times 100 \times 100$ grid points would cover 3–5 large dominant eddies in each horizontal direction and at the same time resolve small eddies down to about $100 \text{ m} \times 100 \text{ m} \times 40 \text{ m}$ in size assuming model resolution is twice the grid size. For the stable PBL where dominant eddies are smaller, a smaller domain (and consequently a finer grid) is preferred.

Numerical truncation errors and specification of boundary conditions add uncertainties to all numerical models including LES. Most PBL-LES codes use finite difference methods in all three directions to compute derivatives, although some LESs employ a spectral (Fourier) representation in x – y planes taking advantage of the horizontally homogeneous nature of the PBL and computational efficiency of using fast Fourier transformation. Sharp gradients in flow variables can exist at the top of the PBL because of the presence of a strong, stably stratified overlying layer, which leads to oscillations (dispersion errors)

when finite differencing methods are used. To overcome this flaw, sign preserving (monotone) schemes are often used for scalar transport to maintain physical realizability – at the expense of introducing more numerical diffusion (see **Numerical Models: Methods**).

The surface boundary condition in LES uses M–O similarity theory to relate surface fluxes to resolved-scale variables at the first grid level – at every grid point. Note that for PBL applications, LES cannot possibly resolve the viscous layer, which is less than a centimeter above the surface. The lowest grid level of a PBL-LES lies in the inertial sublayer, which is referred to as the surface layer. The primary empirical input parameter in M–O theory is the surface roughness height, which varies from less than 0.0001 m for a smooth sea to more than 0.1 m for heavily wooded terrain. This rough-wall boundary condition is different from the smooth-wall condition in engineering flows. Caution should be used, however, because M–O theory describes ensemble-mean flux-gradient relationships in the surface layer (see **Boundary Layer (Atmospheric) and Air Pollution: Surface Layer**) and may not apply well at the local LES grid scale. This problem becomes more acute when the LES horizontal grid size is comparable to or smaller than the height of the first grid level.

The upper boundary of a typical LES domain is usually set to be well above the PBL top to avoid influences of boundary conditions on simulated PBL flows. At the top of the domain, turbulence is negligible and a no-stress condition is applied. Because turbulent motions in the PBL may excite gravity waves in the stably stratified inversion layer, a mechanism for allowing gravity waves to escape at the top of the domain is often applied. This includes applying a radiation condition or adding a wave-absorbing sponge layer at the top of the model domain.

For lateral boundary conditions, almost all PBL LESs use periodic boundary conditions because there is no adequate theory to define chaotic flow fields at an open boundary. However, periodic boundary conditions are clearly inappropriate for horizontally inhomogeneous cases, particularly for PBLs over complex terrain or under severe weather conditions. Recently, researchers have been using the nesting technique to avoid the use of periodic boundary conditions. We will get back to this issue later in **Weather Model Nesting**.

Visualization of LES-Generated Flows

The solution of eqn [4] consists of three-dimensional, time-evolving, chaotic flow fields. An example of such a flow field is shown in **Figure 2** where the vertical velocity of an LES solution of a free convective PBL (using $512 \times 512 \times 512$ grid points) is presented. The total domain of the LES is 5 km $\times$ 5 km in horizontal and 2 km in the vertical. The horizontal view shows a spokelike, irregular polygonal structure near the surface, similar to those observed in Rayleigh–Bernard convection experiments. This spokelike feature is most evident in the free convective PBL where there is no mean wind. Intersections between polygons are local horizontal convergence regions and hence are sites to form strong updrafts (purple colors).

Strong updrafts can be seen from the two vertical plan views. These updrafts carry warm surface air and that is how turbulence can effectively transport heat (and other species

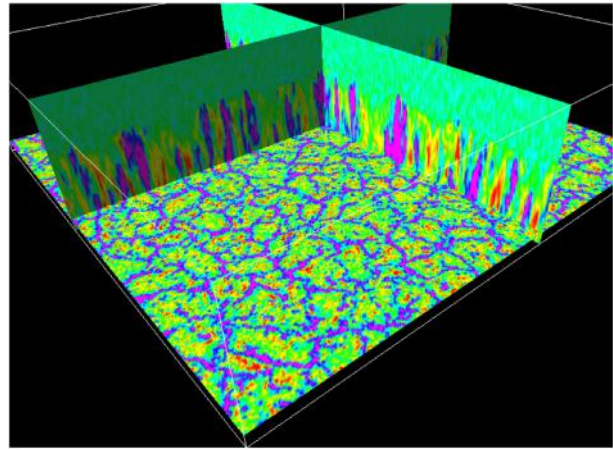


Figure 2 Contours of vertical-velocity field from an LES of the free convective PBL, shown on a horizontal plane near the surface and also on two vertical cross sections. The purple color indicates strong updrafts and red for strong downdrafts. The total domain is 5 km in both horizontal directions and 2 km in the vertical.

coming from the surface) upward. Updrafts are more intense and occupy a narrower area than downdrafts; this is known as a positively skewed vertical-velocity field, a unique feature of buoyancy-driven turbulence. Strong updrafts penetrate into the capping inversion and in the process engulf wisps of warm inversion air into the PBL. These wisps of air are subsequently entrained and mixed into the PBL. This penetration-leads-to-entrainment is a phenomenon that has been documented with radar and lidar observations and convection tank experiments (see **Boundary Layer (Atmospheric) and Air Pollution: Convective Boundary Layer**).

Turbulent structures are quite different in a stably stratified (or night-time) PBL. **Figure 3** shows the vertical-velocity field from an LES of a stable PBL. In the stable PBL, buoyancy consumes TKE that leads to much weaker turbulence compared

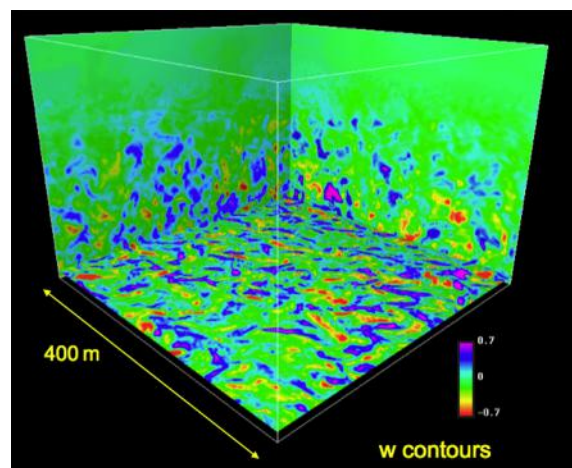


Figure 3 Contours of the vertical-velocity field from an LES of a stably stratified PBL, shown on a horizontal plane near the surface and also on two vertical cross sections. The purple color indicates strong updrafts and red for strong downdrafts. The total domain is 400 m in all three directions.

to the convective PBL. There is no spokelike structure and no strong updrafts. There are numerous small-scale interacting turbulent patches. (Field measurements show that the dominant turbulent eddy size in a typical stable PBL is on the order of several tens of meters. So for this simulation, the total domain is set to 400 m in all three directions using $200 \times 200 \times 200$ grid points.)

Statistics Derived from LES Flows

Instantaneous flow fields from LESs are chaotic, so it is their collective effects (or statistics) like variances or fluxes that are useful for applications. Moment statistics can be readily calculated from 3D LES data by correlating the local fluctuations among variables and averaging them over space and/or time. Then the vertical profiles or distributions of these statistics can be systematically documented for various PBL regimes generated under various large-scale conditions. For example, the TKE budgets calculated from LESs (Moeng and Sullivan, 1994) show differences between the shear and buoyancy-driven PBL regimes (Figure 4). In a shear-driven PBL, shear production nearly balances molecular dissipation with all other terms remaining small, while in the convective PBL, the TKE budget is dominated not only by the buoyancy production and molecular dissipation but also by the turbulent and pressure transports; these features are consistent with field observations and laboratory experiments.

A unique feature of LES is the ability to obtain pressure statistics. Pressure fluctuations are difficult, if not impossible, to measure in the field, yet they play an important role in determining moment statistics, such as pressure transport in the TKE budget and the return-to-isotropy behavior for velocity variances. The LES-generated pressure field, which remains to be verified from observation when available, provides a unique tool to estimate important pressure-related statistics.

One should be cautious in using statistics constructed from LES flows, however. Some statistics, especially higher moments, may be sensitive to the LES grid resolution, domain size, and SGS models. A necessary but not sufficient rule of thumb is to

accept or consider only the statistics that are insensitive to the LES grid resolution or SGS modeling.

Applications to Atmospheric Turbulence

Significant Accomplishments

LES has become a prominent research tool in advancing our understanding of the structure and physics of PBL turbulence. Before Deardorff's first LES calculations in the early 1970s, researchers believed that the proper velocity and length scales for PBL statistics were the friction velocity u_* and the length scale u_*/f , where f is the Coriolis parameter. Based on his LES calculations, Deardorff discovered that turbulence statistics in the convective PBL are better described by the convective velocity scale $w^* = [(g/T_0)z_i\overline{w\theta_0}]^{1/3}$ and the PBL depth z_i , where $\overline{w\theta_0}$ is the surface buoyancy flux (Deardorff, 1972). (Note that the overbar here denotes ensemble averages, which can be calculated as spatial and time averages from LES-generated flows.) This new finding, now known as mixed-layer scaling, makes it possible to collapse observed data collected from convective PBLs under various environments to form universal vertical profiles. For example, measurements of vertical flux of TKE $\overline{wE}$ and vertical-velocity variance $\overline{w^2}$ from research aircrafts at various heights, for various surface heat fluxes, form universal profiles only when these statistics are normalized by w^3 and w^2 , respectively, and also when they are shown as functions of the normalized height z/z_i (Figure 5, from Lenschow et al., 1980) (see **Boundary Layer (Atmospheric) and Air Pollution: Convective Boundary Layer**).

LES also has provided a revolutionary discovery about plume dispersion in the convective PBL. The release of a tracer from an elevated source within an LES-generated convective PBL shows that the maximum mean concentration in the plume at first descends until the plume intercepts the ground before it rises (Figure 6). The descent of the elevated plume maximum is due to the greater areal coverage of downdrafts, i.e., the positively skewed vertical-velocity field (Lamb, 1978). This finding, also observed at about the same time in the Willis and Deardorff tank experiments, has an important application to air pollution; the result can be used to predict the location and magnitude of the maximum surface concentration of emissions downstream. It provided the basis for the revision of short-range dispersion models in the 1980s (see **Turbulence and Mixing: Turbulent Diffusion**).

Another breakthrough from LES is the discovery of the asymmetry of turbulent diffusion from area sources at the surface and top of the convective PBL. Any passive, conservative scalar can be linearly decomposed into two conceptual scalar fields: top-down (which is emitted at the PBL top and has 0 flux at the surface) and bottom-up (which is emitted at the surface and has 0 flux at the PBL top). Under quasi-steady state, the fluxes of the top-down and bottom-up scalars are both linear in height and hence, after normalization by their respective boundary flux, are symmetric about the mid-PBL. LES shows that the mean gradients of the top-down and bottom-up concentrations, after normalization by w^* , z_i and the appropriate boundary flux, are *not* symmetric about the mid-PBL (Wyngaard and Brost, 1984). While the top-down gradient function remains positive throughout the whole PBL, the

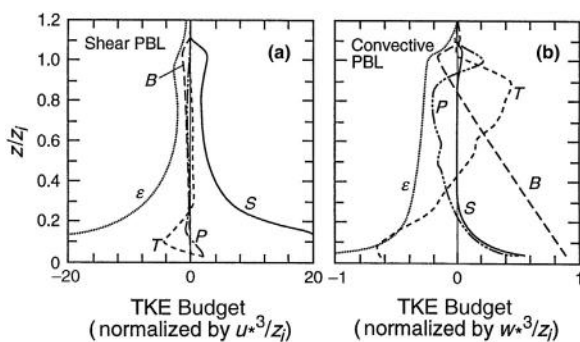


Figure 4 Vertical distributions of the terms in the TKE budget from LESs of (a) a shear generated PBL and (b) a convective PBL. Notations: B buoyancy production, S shear production, T turbulent transport, P pressure transport, and ϵ molecular dissipation rate. Reproduced with permission from Moeng, C.-H., Sullivan, P.P., 1994. A comparison of shear- and buoyancy-driven planetary boundary layer flows. *Journal of Atmospheric Sciences* 51, 999–1022.

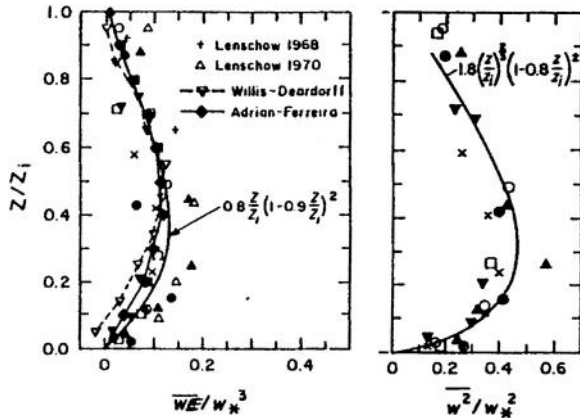


Figure 5 Observed profiles of TKE flux and vertical-velocity variance, both scaled with z_i and w_* . Adapted with permission from Lenschow, D.H., Wyngaard, J.C., Pennel, W.T., 1980. Mean-field and second-moment budgets in a baroclinic, convective boundary layer. *Journal of the Atmospheric Sciences* 37, 1313–1326.

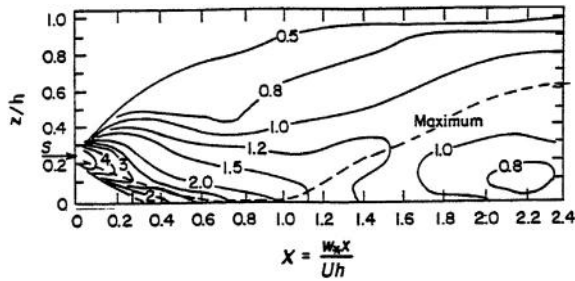


Figure 6 Contours of normalized crosswind-integrated concentration as a function of normalized height and downwind distance from an elevated source in a convective PBL as predicted from LES. Here h is the PBL depth, U the mean wind, and x downwind distance from the source. Adapted with permission from Lamb, R.G., 1978. A numerical simulation of dispersion from an elevated point source in the convective boundary layer. *Atmospheric Environment* 12, 1297–1304.

bottom-up mean gradient is positive in the lower half of the PBL but becomes negative in the top half (Moeng and Wyngaard, 1984). The negative gradient in bottom-up scalar indicates the counter-gradient transport feature, where flux and mean concentration gradient have the same sign. This asymmetric feature of the gradient functions results in different top-down and bottom-up eddy diffusivities, where the latter becomes ill-defined in the mid-PBL. Thus, in the convective PBL, a scalar emitted from the surface diffuses differently from one emitted from the top.

One of the major contributions of LES to the meteorological community is the development or calibration of PBL parameterizations, referred to as PBL models (see **Boundary Layer (Atmospheric) and Air Pollution: Modeling and Parameterization**) for climate and weather forecasting models. Various PBL parameterizations have been proposed but few of them have been evaluated or verified because field observations are often incomplete for this application. LES solutions have been used as valuable datasets to examine many closure assumptions in existing PBL schemes. This includes assumptions made

in second-order closure modeling; mass flux and entrainment/detrainment closures for mass flux modeling; entrainment-rate closure assumptions in mixed-layer modeling; and counter-gradient effects in eddy-diffusivity models. Recently, LES has been extended to study not just parameterizations of PBL turbulence but all SGS motions in cloud-resolving (cloud-system-resolving) models.

Recent LES Research for Atmospheric Science

Earlier LES work focused mainly on idealized cloud-free, flat-terrain convective PBL. This flow regime is most suited to LES because of the presence of large thermal plumes with no other complicated physical processes (like radiation and latent heating) involved. In recent years, however, LES has been expanded to study more complicated and difficult PBL regimes that are relevant to climate and severe weather predictions, or more recently wind-energy applications. We list several topics below.

Stratocumulus-Topped PBLs

One climatologically important PBL regime is the stratocumulus-topped PBL where its extensive cloud cover can significantly alter the solar radiation input to the Earth's surface. For this PBL regime, LES needs to include effects of latent heating (i.e., phase change of water substance) and radiation processes, which unfortunately introduce more uncertainties into LES. In particular, the entrainment process of this PBL regime is very difficult to simulate because it depends sensitively on the SGS model and numerical schemes (Stevens et al., 2005). Many LES practitioners working on this topic have been involved in GCSS (GEWEX Cloud-System Study) Boundary Layer Cloud Working Group, with the goal of using LES as database to improve representations of stratocumulus clouds in climate and weather prediction models.

Weather model nesting

Limited by computer power, traditional atmospheric models have been used to perform single-scale simulations. For example, weather models simulate only weather-scale motions of several tens of kilometers in size, while LES resolves just large turbulent eddies of several hundreds of meters in size. The nonresolved part of motions in traditional weather models are either prescribed or roughly parameterized. This is clearly not suitable for applications where weather and turbulent motions strongly overlap or interact. One way to deal with this problem is the development of multiple-scale modeling that nests a turbulence-resolving model (i.e., LES) inside a weather prediction model. The outer domains with coarser grid resolutions resolve weather-scale systems, while the inner domains consist of a fine-grid LES to resolve turbulent motion. With two-way nesting, scale interactions of weather and turbulence motions are allowed.

The nesting technique may also solve the periodic-boundary-condition problem. With nesting, the lateral boundary conditions of LES are provided by their adjacent outer-domain flow fields. Thus no periodic boundary condition is imposed; such LES can be applied to horizontally

inhomogeneous, nonperiodic conditions. Built on community weather models, notably the Weather Research and Forecasting (WRF) model, this type of multiple-scale modeling has shown promising results when applied to idealized or uniform-surface cases (Zhu et al., 2010). It remains a challenge when applied to complex terrain or severe weather situations (see **Boundary Layer (Atmospheric) and Air Pollution: Complex Terrain**). (For further reading see the WRF website at <http://www.wrf-model.org/index.php>.)

Deep cloud systems

With massive parallel computing, LES is no longer limited to idealized PBL applications. For example, as part of the research funded by the NSF Center for Multiscale Modeling for Atmospheric Processes, a very large-domain LES, covering an area of about $205 \text{ km} \times 205 \text{ km}$ in the horizontal and about 27 km in the vertical, was performed to simulate a tropical deep convective system (Khairoutdinov et al., 2009). It used $2048 \times 2048 \times 256$ (about 10^9) grid points and ran on thousands of processors. This LES resolves a broad range of scales that includes mesoscale organization, gravity waves, deep and shallow clouds, all the way down to energy-containing turbulence eddies. This is another kind of multiple-scale modeling but it does not use the nesting approach, and therefore is more accurate. However, it is done at the expense of huge computer resources; a 24-h simulation of this LES took about 400 000 h on IBM's BlueGene/L supercomputer.

A satellite view of the simulated cloud field shown in **Figure 7** (adopted from Khairoutdinov et al., 2009) reveals a detailed structure of the deep convective system. It shows mesoscale organization, deep and shallow clouds (bright white color), thin cirrus clouds (light blue), and random turbulent motions. This computationally expensive LES is now used as a benchmark to study how large clouds interact with small clouds and turbulence. It is also used to develop SGS parameterizations for cloud-resolving models (Moeng et al., 2010). (For this LES see www.cmmmap.org.)

Large-domain LES is also used to simulate hurricane and midlatitude squall line systems.

Marine boundary layers with resolved waves

There is a clear impetus for using LES to attack increasingly complex flows for a variety of applications. Air-sea interaction and the coupling between winds and currents with surface gravity (water) waves is an example where LES has provided new information about turbulent flow dynamics and has also guided the interpretation of data collected in field campaigns.

To simulate atmospheric turbulence above a three-dimensional time-dependent surface wave field, the LES governing equations are written in a transformed wave-following coordinate system (Sullivan and McWilliams, 2010). The new complexity introduced here is that the gridlines of the mesh translate vertically adapting to the surface movement. The LES system of equations is augmented by an additional equation governing the grid movement. LES of the atmospheric marine boundary layer with varying stratification illustrate the importance of wind-wave directionality and so-called 'wave age,' i.e., the ratio of a reference wind speed and a characteristic

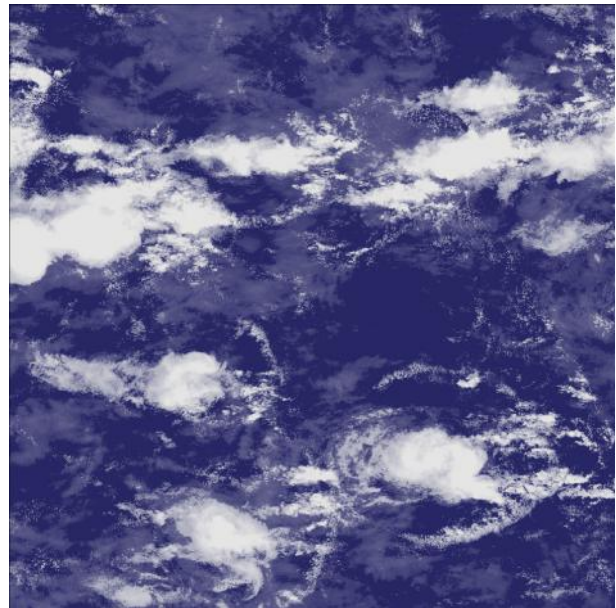


Figure 7 Visualization of simulated cloud scene over an area of about $205 \text{ km} \times 205 \text{ km}$ generated from the large-domain LES. The image represents visible albedo estimated from the liquid and ice water paths. The cloud arcs around some of the deep clouds are shallow clouds lifted up by the gust fronts at the edges of spreading cold pools. White color indicates thick clouds, semitransparent gray-blue indicates thin cirrus clouds, and dark-blue color shows the cloud-free regions. Adapted with permission from Khairoutdinov, M.F., Krueger, S.K., Moeng, C.-H., Bogenschutz, P.A., Randall, D.A., 2009. Large-eddy simulation of maritime deep tropical convection. *Journal of Advances in Modeling Earth Systems* 1, Art # 15, 1–13.

wave speed in a surface wave height spectrum. Depending on wave age, the winds are slowed or accelerated by the action of the waves. The mean wind profile, turbulence variances, and vertical momentum flux are thus dependent on the nature of the wave field; the LES predicted dependence of vertical momentum flux on wave age is also found in observations. LES results with moving waves show important differences compared with rough-wall boundary layers and flow over stationary bumps. LES of upper ocean boundary (or mixed) layers also incorporates surface wave effects by including the Stokes drift associated with the wave field. The latter appears as a new forcing term in LES, a so-called vortex force, which is responsible for the generation of Langmuir circulations. Langmuir circulations are potent coherent structures that interact with background ocean turbulence to produce enhanced mixing and entrainment in the ocean boundary layer.

Observations of SGS variables

The basic assumptions used to derive SGS closure models and constants for LES, discussed previously, are often violated in actual LES implementations. Flow near rough boundaries, regions with stable stratification, and the addition of new dynamical processes, e.g., vegetation, hills, and clouds, modulate the background turbulence and introduce new time and length scales that are not accounted for in conventional SGS closures.

Recently, novel observations have been carried out in the atmospheric surface layer with the objective of measuring and quantifying the structures and statistics of SGS variables. The basic principle of the Horizontal Array Turbulence Studies (HATS) is to sample turbulent fields by orienting an array of point sensors, usually sonic anemometers, perpendicular to the main flow direction. Then by adopting Taylor's hypothesis in the streamwise direction, a 'horizontal plane of turbulence' can be constructed. A two-dimensional filter (in horizontal directions) is applied to the wind and temperature fields that mimic the filtering in LES codes. If a sufficient number of sensors are employed, then the fields can be double filtered to yield additional information about the coupling between resolved and SGS correlations. Further, if a vertically stacked array of sensors is used in the field (Figure 8) vertical derivatives of resolved and SGS variables can be acquired. HATS field campaigns have been carried out over flat rough surfaces, moving ocean waves, in a canopy of trees, and over snow. In these studies, the atmospheric stratification varies from unstable to neutral to stable. (For these field data, see <http://data.eol.ucar.edu/codiac/projs?SGS00>.)

The statistics of SGS variables, from the HATS experiments, are studied similar to their conventional ensemble average counterparts for means and Reynolds-averaged fluxes and

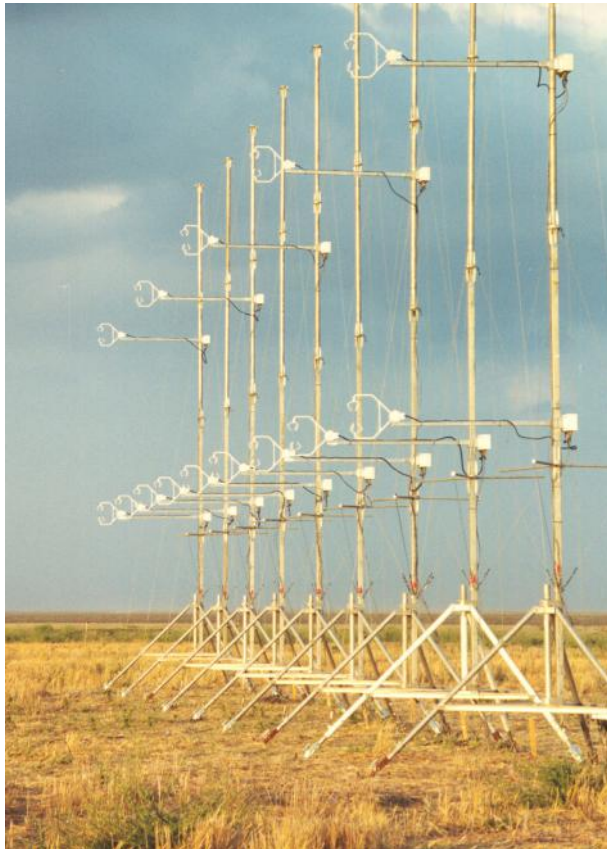


Figure 8 Twin horizontal arrays of sonic anemometers with 9 and 5 sensors displaced vertically.

variances. Analysis of SGS momentum and scalar fluxes from HATS (Sullivan et al., 2003) shows that the scale separation between the largest energy-containing eddies Λ and the filter scale Δ is a fundamental critical parameter. The variation of the SGS fluxes and variances for different filter widths, height above the surface, and atmospheric stratification collapse when plotted as a function of Λ/Δ . When the ratio Λ/Δ is large (>10) then the SGS variables are more isotropic and can be modeled using expressions described previously. In an intermediate range of scale separation ($1 < \Lambda/\Delta < 10$), the SGS fluxes become increasingly anisotropic, and more complex closures are required (Wyngaard, 2004). When $\Lambda/\Delta < 1$ then the resolved motions are small compared to the SGS and second-order closure modeling is appropriate. HATS datasets can be used to estimate the transfer of energy between the resolved and SGS field and viscous dissipation in eqn [10] as well as closure constants. The observations show that the Smagorinsky constant c_s in eqn [7], (or c_K in eqn [11]) is dramatically reduced near rough boundaries.

Future Challenges

In reality, the PBL is much more complicated than what has been simulated by LES so far. Much of the complication arises from the heterogeneous nature of the underlying surface. The earth's land surface is characterized by spatially varying patches, undulating terrain, and urban development, which can induce circulations that interact with, and hence change turbulence dynamics. Complex surface conditions in particular affect the very stable PBL where turbulence is no longer continuous but becomes intermittent in space and time (see **Boundary Layer (Atmospheric) and Air Pollution: Stably Stratified Boundary Layer**). The ability to simulate turbulence transition becomes crucial. Wave-current couplings and wave breaking also lead to complex air-sea interactions. All these complex surface conditions can significantly influence turbulent transport in many meteorological applications, such as air pollution, vegetation growth, cloud formation, and hurricane development.

LES is now being adapted to realistic PBL regimes embedded in multiscale meteorological flows. Including complexities, however, introduces additional uncertainties in LES solutions. It is important to examine and validate the fidelity of LES of complicated flows against observations. LES has also been used to study interactions of turbulence with cloud microphysics, biochemistry, and aerosols. Caution should be exercised in these applications since the interactions may depend critically on small-scale turbulent motions, which are SGS in LES.

See also: Agricultural Meteorology and Climatology. **Boundary Layer (Atmospheric) and Air Pollution:** Complex Terrain; Convective Boundary Layer; Modeling and Parameterization; Stably Stratified Boundary Layer; Surface Layer. **Dynamical Meteorology:** Coriolis Force; Primitive Equations. **Numerical Models:** Methods. **Turbulence and Mixing:** Overview; Turbulent Diffusion.

References

- Deardorff, J.W., 1972. Numerical investigation of neutral and unstable planetary boundary layers. *Journal of Atmospheric Sciences* 29, 91–115.
- Khairoutdinov, M.F., Krueger, S.K., Moeng, C.-H., Bogenschutz, P.A., Randall, D.A., 2009. Large-eddy simulation of maritime deep tropical convection. Large eddy simulation of maritime deep tropical convection. *Journal of Advances in Modeling Earth Systems* 1, Art #15, 1–13.
- Lamb, R.G., 1978. A numerical simulation of dispersion from an elevated point source in the convective boundary layer. *Atmospheric Environment* 12, 1297–1304.
- Lenschow, D.H., Wyngaard, J.C., Pennell, W.T., 1980. Mean-field and second-moment budgets in a baroclinic, convective boundary layer. *Journal of the Atmospheric Sciences* 37, 1313–1326.
- Moeng, C.-H., Sullivan, P.P., 1994. A comparison of shear- and buoyancy-driven planetary boundary layer flows. *Journal of the Atmospheric Sciences* 51, 999–1022.
- Moeng, C.-H., Wyngaard, J.C., 1984. Statistics of conservative scalars in the convective boundary layer. *Journal of the Atmospheric Sciences* 41, 3161–3169.
- Moeng, C.-H., Wyngaard, J.C., 1988. Spectral analysis of large-eddy simulations of the convective boundary layer. *Journal of the Atmospheric Sciences* 45, 3573–3587.
- Moeng, C.-H., Sullivan, P.P., Khairoutdinov, M.F., Randall, D.A., 2010. A mixed scheme for subgrid-scale fluxes in cloud-resolving models. *Journal of the Atmospheric Sciences* 67, 3692–3705.
- Stevens, B., et al., 2005. Evaluation of large-eddy simulations via observations of nocturnal marine stratocumulus. *Monthly Weather Review* 133, 1443–1462.
- Sullivan, P.P., McWilliams, J.C., 2010. Dynamics of winds and currents coupled to surface waves. *Annual Review of Fluid Mechanics* 42, 19–42.
- Sullivan, P.P., McWilliams, J.C., Moeng, C.-H., 1994. A subgrid-scale model for large-eddy simulation of planetary boundary-layer flows. *Boundary-Layer Meteorology* 71, 247–276.
- Sullivan, P.P., Horst, T.W., Lenschow, D.H., Moeng, C.-H., Weil, J.C., 2003. Structure of subfilter-scale fluxes in the atmospheric surface layer with application to large-eddy simulation modeling. *Journal of Fluid Mechanics* 482, 101–139.
- Wyngaard, J.C., 1984. Large-Eddy Simulation: Guidelines for Its Application to Planetary Boundary Layer Research. U.S. Army Research Office Contract 0804, pp. 122.
- Wyngaard, J.C., 2004. Toward numerical modeling in the Terra Incognita. *Journal of the Atmospheric Sciences* 61, 1816–1826.
- Wyngaard, J.C., 2010. *Turbulence in the Atmosphere*. Cambridge University Press, New York, p. 393.
- Wyngaard, J.C., Brost, R.A., 1984. Top-down and bottom-up diffusion of a scalar in the convective boundary layer. *Journal of the Atmospheric Sciences* 41, 102–112.
- Zhu, P., Albrecht, B.A., Gbate, V.P., Zhu, Z., 2010. Multiple-scale simulations of stratocumulus clouds. *Journal of Geophysical Research* 115, D23201.

Relevant Websites

<http://www.wrf-model.org/index.php>.
<http://cmmmap.colostate.edu/>.
<http://data.eol.ucar.edu/codiac/projs?SGS00>.

Regional Prediction Models

BW Golding, Met Office, Exeter, UK

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

Regional numerical weather prediction (NWP) models are used by many National Meteorological and Hydrological Services to forecast the detailed weather over their areas of responsibility for the next few days. Their main advantages are in achieving much higher resolution, with limited computer power, than is possible with a global model and in delivering forecasts more quickly and frequently. They require information from a global model to provide the forecast for the boundary of the regional domain. Advanced NWP centers use regional models with grid lengths ranging from 1 to 3 km.

Introduction

Numerical weather prediction (NWP) models are at the heart of the modern production of weather forecasts. They use the mathematical equations governing motion of a thin shell of gas on a rotating sphere, together with the equations describing changes of the state of water between vapor, liquid, and ice, and supplemented by additional computations to represent the effects of the sun's radiation, of clouds and of the turbulence generated by surface topography. In addition, the conditions at the bottom of the atmosphere must be specified. A constant temperature, obtained from recent observations, is usually specified over the oceans, but over land the evolution with time must be predicted using equations that describe the interaction between radiative heating or cooling and evaporation at the surface, and the heat and water content of the soil beneath, taking account of the effects of vegetation on these processes. The result is a comprehensive description of the state of the atmosphere and the earth's surface at regular intervals in space and time.

Regional prediction models are NWP models that predict the evolution of the atmosphere overlying a limited part of the earth's surface. Such models are used so that, with limited computer power, the atmosphere overlying this part of the earth's surface can be modeled with the highest possible resolution. For simplicity, we exclude variable resolution global NWP models, which are used by some weather services as an alternative method of achieving this aim. We also exclude those models designed for simulation rather than forecasting, since they are typically designed to focus on the study of particular processes or meteorological events, and cannot properly be termed prediction models. A key distinguishing feature of a prediction model is that it has a data assimilation component to specify the initial state.

Whereas global NWP models have boundaries only at the bottom and top of the atmosphere, a regional model has lateral boundaries at the edges of the region covered. Conditions at these boundaries may be held fixed, may be prescribed from another model covering a larger area, or may be interactive, passing information both from and to a model covering a larger area. The study of how to optimally specify the boundaries of regional models has been very important in their evolution, and has also led some scientists to abandon them in favor of variable resolution global models, as mentioned above.

Requirement

The need for regional models comes out of the requirements placed on National Weather Services (NWSs) for forecasts of high impact weather and for detailed forecasts that distinguish the weather forecast at different locations characteristic of the topographical variability of a country (**Figure 1**). Weather systems that can have high impact include thunderstorms and

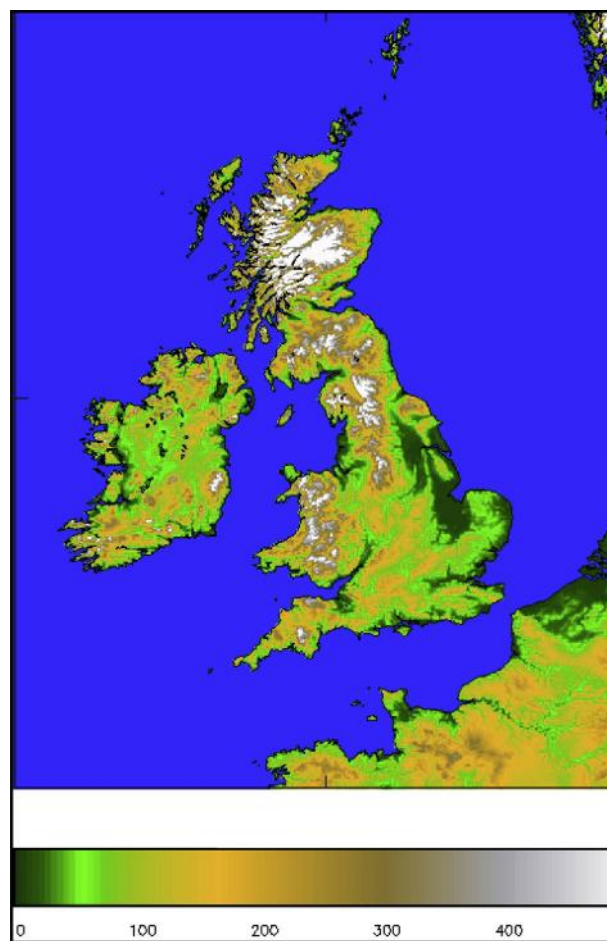


Figure 1 Representation of land surface height in the UK convective scale model (heights in meters).

squall lines, both of which are typically less than 10 km across, and fog, which is often constrained by valleys, also of small size. For countries with coastal or mountainous regions, the local variations of importance to citizens also occur at very small scales: typically less than 10 km separating a headland from a bay or a mountain from a valley. These requirements place a strong pressure on the weather forecasting agencies to forecast at the finest possible resolution and, with a given amount of computer power, this can best be achieved using a regional model embedded in a coarser resolution global model, the latter being run either locally or often at a remote center. NWSs are also required to provide weather information for all parts of their country, so regional models are usually configured so as to encompass the whole of the country plus a border region where the forecast may be distorted by the boundary conditions. As a result of these two conflicting requirements, it is often the case that the highest resolution regional models are run by the smaller countries.

The requirements placed on NWP models are intimately bound up with progress toward using automated methods in the preparation of weather forecasts, and therefore match to a large degree the meteorological requirements of the weather forecasts themselves. The dominant driver for development of regional models has been the goal of Quantitative Precipitation Forecasting. While the capability of models to predict precipitation has steadily improved, it is only in recent years that even regional NWP models have been able to resolve the processes that are relevant to flood producing rainfall. Apart from rainfall, regional NWP models are also able to forecast all of the elements of a typical public weather forecast: temperature, wind, cloud, snow, hail, and fog. The skill of models varies with the location, the lead time and the variable of interest. For some variables, the raw model forecasts can be improved using automated algorithms based on either the recent historical performance (e.g., using model output statistics) or the conditions at the start of the forecast (e.g., using nowcasting techniques).

For longer lead times, the area of the globe that influences the local weather becomes larger, and so the distortions to the forecast due to the artificial boundary become more important. These distortions counteract the benefit of higher resolution. Since the forecasts become less accurate anyway at longer lead times, most regional models are only used for short range weather forecasting, up to about 2 days ahead. However, some recent advances in high-resolution modeling mean that these models can provide useful additional information at much longer ranges. This use of regional models is termed 'dynamic downscaling' to indicate that the information on the evolution of the atmosphere is all coming from the global model, but the regional model is computing the result of the large-scale weather systems interacting with the smaller scale local topography. It is also possible to estimate these interactions using 'diagnostic downscaling' methods, but they are only able to represent a very simplified set of interactions compared to those that can be represented by the NWP model.

Development

The first NWP models were all regional models due to lack of computer power and of observations to drive global models.

Indeed it was not until the late 1970s that the first global NWP models were introduced, although global atmospheric circulation models had been run for some time previously to support atmospheric research.

Since there were no global models to provide boundary conditions, early models had to use fixed boundary conditions. This resulted in a severe limitation on the lead time for which forecasts were useful, since the effects of these erroneous boundary conditions propagated rapidly into the forecast domain. The earliest models were balanced models, obtained using the quasigeostrophic approximation. This resulted in a pronounced simplification and allowed the powerful theory of quasigeostrophic flow dynamics to be used in designing them. The later move to the so-called primitive equations allowed more accurate representation of important processes, but considerably complicated the model structure, and especially the processes required to produce an initial state from which a realistic prediction could be made. After several years the main types of model split the equations so that the smooth evolution obtained with the balanced equations could be integrated cheaply using long time steps, while the fast motions could be integrated less precisely, but in such a way as to ensure the stability of the solution. Two techniques have been widely used for this, the so-called split-explicit and split semi-implicit methods. It was at this time, also, that modelers divided into those using atmospheric variables defined at grid points, and those using a Fourier decomposition of the distribution of the variable. The latter, termed the spectral method, has proved very valuable for global modeling, but few centers have adopted it for regional models.

An important issue for all limited area models was the method of specification of the boundary conditions. There are important mathematical constraints on this, arising from the atmospheric equations being solved. Important work in the 1970s demonstrated that the specification should be different at inflow boundaries from that at outflow boundaries, but that the assessment of the type of boundary should take account not just of the mean wind, but also of the speed of traveling pressure waves. The complexity of such solutions deterred operational modelers, who generally specified all available variables, and then controlled any resulting disturbance using diffusion. In the end, it was shown that specification of a mathematically correct set of conditions is not possible for the primitive equations. Following from this, research into more pragmatic methods resulted in the technique most widely used today, in which the boundary values are blended into the regional model across several rows of the grid, with decreasing weight toward the interior. This technique suppresses the generation of unrealistic small-scale gravity waves while ensuring that information on the scale of the forcing model is transferred into the interior of the regional model.

As computers became more powerful, there was pressure to extend the boundaries so as to enable realistic longer forecasts to be run. On the other hand, there was also a desire to use finer resolution grids so as to increase the detail in the area for which the forecast was being made. The almost universally adopted solution has been to use a coarse grid covering the whole globe to forecast for the longest lead times possible, and to nest inside it a fine grid regional model covering a domain surrounding the required forecast area.

In some cases, two nested models have been used to enable the finest possible resolution to be used over the immediate forecast area, without having too large a change in resolution at the boundaries. However, it is recognized that each boundary introduces distortions, so this is not generally favored.

Given the cost of high performance computers, an option increasingly being adopted is for countries to combine resources in running a shared global model from which each obtains boundary conditions for its own regional model run at whatever resolution the computer power they can afford will allow.

Although much used for experimental simulations, two-way nested models typically place too great restrictions on the forecasting process and so are not used in regional NWP. Apart from centers that use boundary conditions from another center, which obviously makes two-way nesting impossible, other objections include the need to run the models simultaneously and the need for the models to have the same specification. Typically, global models run after waiting several hours for data from around the world, whereas regional models can run more quickly using local data. The specification of a regional model may well be different for various reasons, including the fact that different numerical techniques work best over different domains, as well as to enable the model to focus on the atmospheric processes most important in the region of interest. Notwithstanding these drivers, there are considerable advantages in the 'seamless' approach to atmospheric modeling, in which multiple configurations of a single model are used for different domains and lead times. With this approach, the scientific expertise in atmospheric radiation processes, for instance, only needs to be applied once to develop a representation for the model, and it can then be used in all configurations, whether global or regional, short range or climate.

Recent development has focused on the so-called convective scale models, with grid lengths of about 1 or 2 km, in which convective storms are represented explicitly rather than being represented by a statistical average of their effects on the larger scale flow. While considerable success has been gained with regional models running at these scales, the theoretical foundation for their performance is much less well understood than for the coarser resolution models and this has undoubtedly slowed down the development of some aspects, particularly the fitting of the initial state to available observations.

While most global models and many regional models represent the vertical structure of the atmosphere using the hydrostatic approximation, which assumes vertical accelerations are small, an increasing number of centers are now using nonhydrostatic models. This is necessary for models that represent convective processes explicitly. The theoretical foundation for these models is significantly more complex than it is for hydrostatic models, and there is, as yet, no agreement on the best form of the equations to be used, some allowing the presence of sound waves, and others excluding them. As with the earlier move from quasigeostrophic to the primitive equations, the methods of treatment of the faster motions and slower motions also differ between models.

Looking to the future, there will undoubtedly continue to be a desire to move to even smaller grid models so as to capture more of the structure of convective storms and local weather systems. However, there are other aspects of the coarser

resolution global models that also have a claim on some of the computer power that will become available in the next few years.

Several weather services are already experimenting with ensembles of convective scale models to try to gain information on the level of certainty in the storm predictions. For the current generation of global models, ensembles are run by perturbing both the initial state of the forecast and the parameters of the forecast model to estimate the sensitivity of the forecast. Techniques have been developed to gain the maximum information about the sensitivity of forecasts of the main weather systems that use as few perturbed forecasts as possible. For a regional NWP model, the primary source of perturbation is likely to be in the weather systems crossing the boundaries from the global model. However, added to this will be the sensitivity of the local weather systems to local perturbation. These are initiated by different processes from the large-scale weather systems, so the perturbations will need to be optimized quite differently.

Another area of development in global models is to extend the processes represented in the models to include those that alter the composition of the atmosphere, including chemical interactions, and to incorporate a dynamic representation of the oceans in place of the fixed sea surface conditions. These developments are well advanced in the climate modeling community, who have coined the term 'earth system model' to describe such a combination of atmosphere, ocean, and land surface in a full interactive model. The benefits are particularly important in long climate forecasts because the ocean has modes of variability on timescales of years to decades. However, such coupling is also of importance for short range forecasting with regional models. Shallow seas respond much more quickly to solar heating than the deep ocean, and the resulting gradients of sea temperature can be stretched out by tidal currents around coasts. Emissions of gases and aerosols from cities and industries can have significant effects on cloud and fog formation. Heavy rain, channeled through rivers, provides an enhanced freshwater source to coastal waters, creating salinity gradients that drive ocean processes. All of these linkages require a coupled regional model for their proper representation.

The other area of very active development is the coupling of impact models to regional NWP models. Impact models may represent physical, biological, social, or economic processes, but the thing they have in common is that they convert the weather variables output by the NWP model into impacts that are of importance to society. With the recent advances in the resolution of regional NWP models, there are now many more impacts of the weather for which the weather forecasts have the precision required to attempt their direct prediction. A good example is flooding, for which forecasts used to rely almost exclusively on upstream river measurements. With the advent of convective scale precipitation forecasts, there is considerable interest in assessing their ability to drive flood forecasting models, especially for rapidly responding catchments that were excluded from previous forecasting approaches.

Examples

The regional NWP model used by the Met Office in the United Kingdom is a convective scale configuration of its general

purpose Unified Model, which is also used for global prediction and climate modeling. It is a nonhydrostatic, grid point model, with one-way boundary conditions provided from a global model with a 25 km grid. The domain is defined as a rectangle on a reprojection of the earth's surface obtained by rotating the poles so that the British Isles appear to be near the equator. The horizontal grid has a spacing of about 1.5 km covering the British Isles and a small surrounding area. This area is made as large as possible by using a stretched grid, which expands up to 4 km. In the vertical, there are currently 70 levels, which follow the shape of the land/sea surface at low levels, but which become flatter at higher levels, until they are surfaces of constant pressure at the top.

The model uses the nonhydrostatic, compressible equations of atmospheric flow in semi-implicit form, with the unbalanced motions integrated implicitly. Sophisticated submodels calculate the influence of clouds, solar and terrestrial radiation, surface vegetation, and turbulence. A full forecast up to 36 h ahead takes about 15 min on an IBM P7 supercomputer.

The initial state for each forecast is obtained by incorporating observations of surface and upper air conditions into the model. These observations are made at fixed times by NWSs, by ships at sea and by aircraft in flight. Together with the observations made by satellites they are exchanged between all of the world's weather services using a sophisticated global data network. The Met Office model uses a three-dimensional variational method for assimilating the observational information

into the model state. In addition, the model state is 'nudged' toward precomputed analyses of cloud and precipitation obtained from radar and satellite imagery, together with lightning fixes and surface cloud and precipitation observations. The aim of the assimilation process is to create an initial model state that evolves smoothly and accurately in the early part of the forecast, avoiding the so-called spin-up problem when the precipitation, for instance, may be deficient in the first few hours of forecast.

A 36 h forecast is produced every 3 h.

The main products from the model are charts of near surface temperature, wind and visibility, cloud amount and height, and precipitation amount, which are used as guidance by forecasters (Figure 2). However, the results are also used as the basis for generation of automated forecasts, which are now competitive with manually generated forecasts for many purposes.

COSMO (CONsortium for Small-scale MOdeling) is a regional model that has been developed by a consortium of European weather services and is implemented in several different configurations in member countries (Figure 3). The German Weather Service uses a convective scale configuration, called COSMO-DE (CONsortium for Small-scale MOdeling-DEutsche), with a horizontal grid spacing of about 2.8 km. Trials of this configuration showed particular skill in the prediction of severe rainfall events leading to floods. They are currently trialing an ensemble prediction system based on this model configuration.

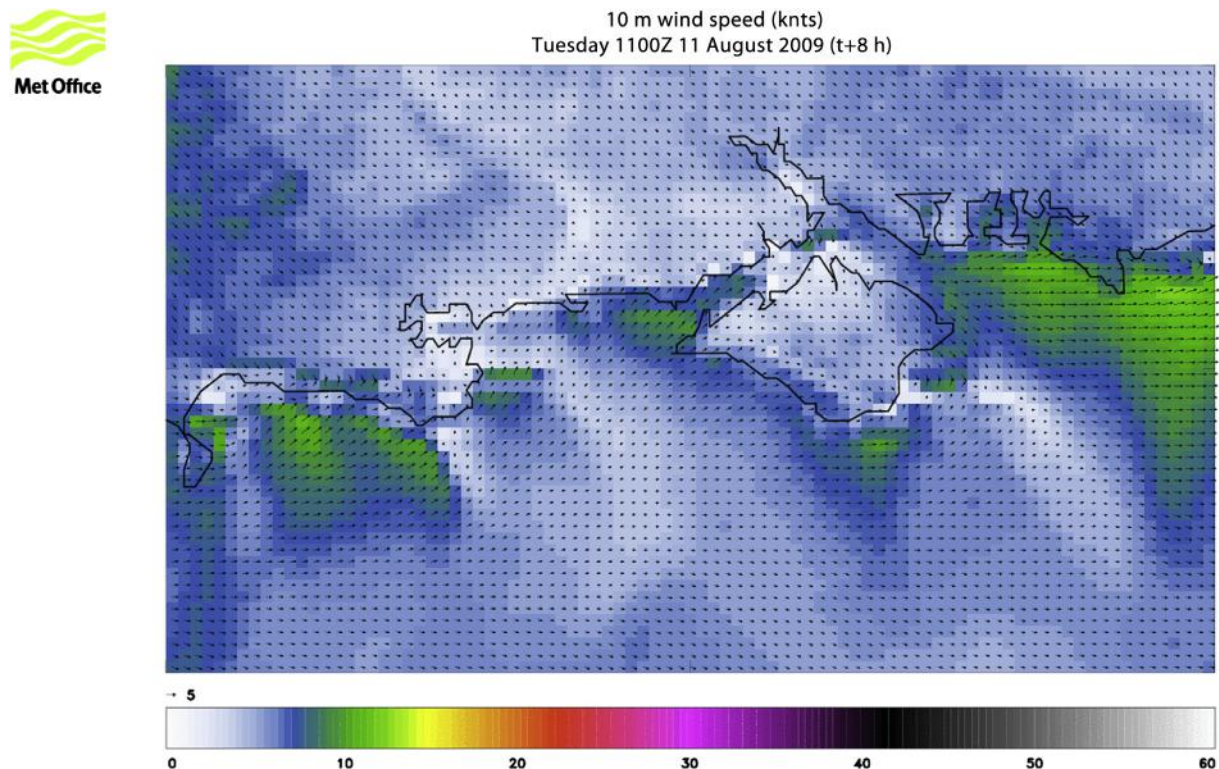


Figure 2 Example of a wind speed and direction prediction at full resolution for part of the south coast of England from the operational 1.5 km grid model of the UK Met Office showing detailed wind responses to the headlands of Portland and Swanage and through the channel between the Isle of Wight and Southampton. (Wind speeds are in knots.)

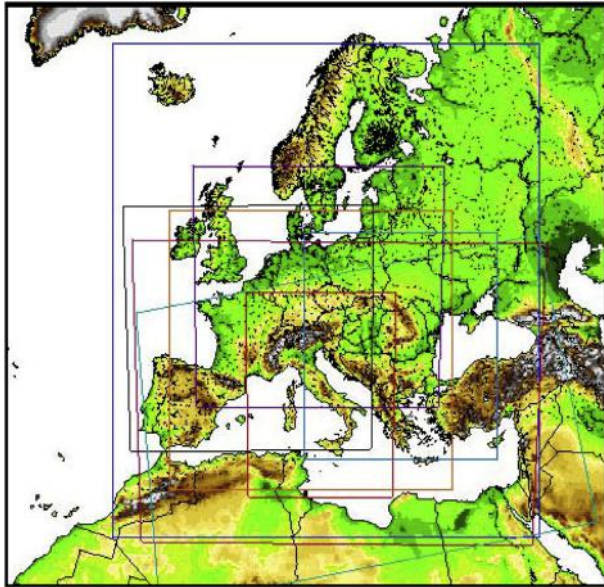


Figure 3 Some European domains implemented by operational users of the COSMO model showing a variety of compromises between covering a large area and focusing high resolution on a single country.

The United States National Centre for Environmental Prediction runs a regional model for the whole of the North American continent and a higher resolution model for the contiguous states. Current operational models have a grid length of 12 km, but trials are at an advanced stage to move to a 3 km grid configuration.

The HIRLAM (high resolution limited area model) has been developed by a consortium of Nordic and other European countries. It is a grid point, hydrostatic model. The participating countries use it in a variety of configurations according to their operational requirements.

See also: **Air Sea Interactions:** Momentum, Heat, and Vapor Fluxes. **Boundary Layer (Atmospheric) and Air Pollution:** Modeling and Parameterization. **Data Assimilation and Predictability:** Data Assimilation. **Dynamical Meteorology:** Balanced Flow; Primitive Equations; Quasigeostrophic Theory. **Numerical Models:** Mesoscale Atmospheric Modeling; Methods. **Weather Forecasting:** Operational Meteorology. **Numerical Models:** Parameterization of Physical Processes: Clouds; Parameterization of Physical Processes: Gravity Wave Fluxes; Parameterization of Physical Processes: Turbulence and Mixing.

Further Reading

Pielke, R.A., 1984. *Mesoscale Meteorological Modeling*. Academic Press.
 Ray, P.S. (Ed.), 1986. *Mesoscale Meteorology and Forecasting*. American Meteorological Society.

Convective Storm Modeling

MD Parker, North Carolina State University, Raleigh, NC, USA

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

Numerical models have been central to a wide range of advances in the knowledge of convective storm dynamics. Many long-standing research findings emerged from carefully controlled sensitivity experiments in early, highly idealized model configurations. Over time, convective storm simulations have become increasingly realistic, thanks to decreases in grid spacing (permitting resolution of subcloud turbulent eddies), more sophisticated treatments of precipitation microphysics, and increased understanding of how often-neglected processes (e.g., radiative fluxes) may influence the conclusions from traditional idealized models.

Introduction

The use of numerical models is extremely popular in the convective storms research community for a number of reasons. Firstly, the governing equations for convection do not have analytic solutions except in very specialized cases. And, secondly, the small-scale details of convective storms are rather hard to observe. High-resolution remote sensing and surface measurements are indeed available from a handful of field experiments. However, the primary wind and thermodynamic variables are usually measured too sparsely to support computation of temporal and spatial derivatives needed for studies of storm dynamics, and *in situ* thermodynamic observations are almost nonexistent aloft. In contrast, numerical model output is dense and conveniently gridded, making it possible to directly compute physically meaningful tendencies and budgets at very high resolution.

Models are also internally consistent dynamical laboratories, providing an extremely effective means for controlled hypothesis testing; inputs can be varied across model experiments (something not possible with observed real-world cases). Idealized models also provide the possibility for reducing the complexity of meteorological problems, in the interest of isolating and understanding important processes. Historically, theory and modeling of convective storms have advanced hand in hand. New theories can readily be tested in models (whereas they are often challenging to validate with conventional observations); in turn, theory is often advanced by attempting to explain sensitivities and behaviors produced by models.

Operational numerical weather prediction (NWP) with models has also recently realized the capability to forecast convective storms explicitly, albeit coarsely (such models are commonly referred to as ‘convection permitting’). This approach has produced some notable successes in predicting timing of convective initiation and the subsequent predominant convective mode. The numerical underpinnings of NWP forecasts are fundamentally the same as those for process study simulations, although the resolution of NWP model runs is usually coarser (due to operational time demands). Whereas real-time convection-permitting forecasts are of great operational importance, the majority of the model-driven advances in our understanding of storm behavior have emerged from more idealized simulations configured for research.

Typical Experimental Design

All of the models currently used for convective storms research could be considered cloud models. In principle, one can use any model that is nonhydrostatic (since vertical accelerations are generally large in convective storms) and represents moist processes, provided it has subgrid-scale and physical parameterizations appropriate for the grid spacings at which convective storms are typically simulated. Most modern cloud models are fully compressible and use techniques such as timestep splitting in order to retain stability while simulating the effects of acoustic waves. Almost all such models use an Eulerian staggered grid (the Arakawa C grid), and perform finite differences using the grid volume-averaged equations. The coordinates may be either Cartesian or terrain-following, and the vertical coordinate may be either height based or mass (pressure) based. Because very high resolution is generally demanded for convective storm simulations, techniques such as grid stretching (primarily in the vertical) and grid nesting (primarily in the horizontal) are often used in order to construct large domains in a way that is computationally affordable.

Starting with the original convective storm simulations that emerged in the 1970s (many using the Klemp–Wilhelmson model, perhaps the first well-established code for the specific purpose of convective storm modeling), typical experiments have employed highly idealized configurations, including horizontally homogeneous initial conditions. In such cases, the simulated convective evolution is understood to be a fundamental response to a representative environmental sounding. Two rather popular initial conditions for convective storm simulations are the sounding from the 20 May 1977 Del City, Oklahoma tornado, and the idealized analytic sounding created by M. Weisman and J. Klemp (known colloquially as the ‘Weisman–Klemp’ sounding) (Figure 1). These soundings are useful benchmarks for sensitivity tests because they are easy to incorporate and are known to produce robust storms in a variety of numerical models. As a part of such sensitivity tests, simplified wind profiles have also been frequently used, including straight-line, quarter-turn, and half-turn hodographs of varying lengths (Figure 1).

Homogeneous model environments are convenient for this kind of experimentation because they allow the user to independently vary the initial thermodynamic and wind profiles without needing to adjust background horizontal gradients

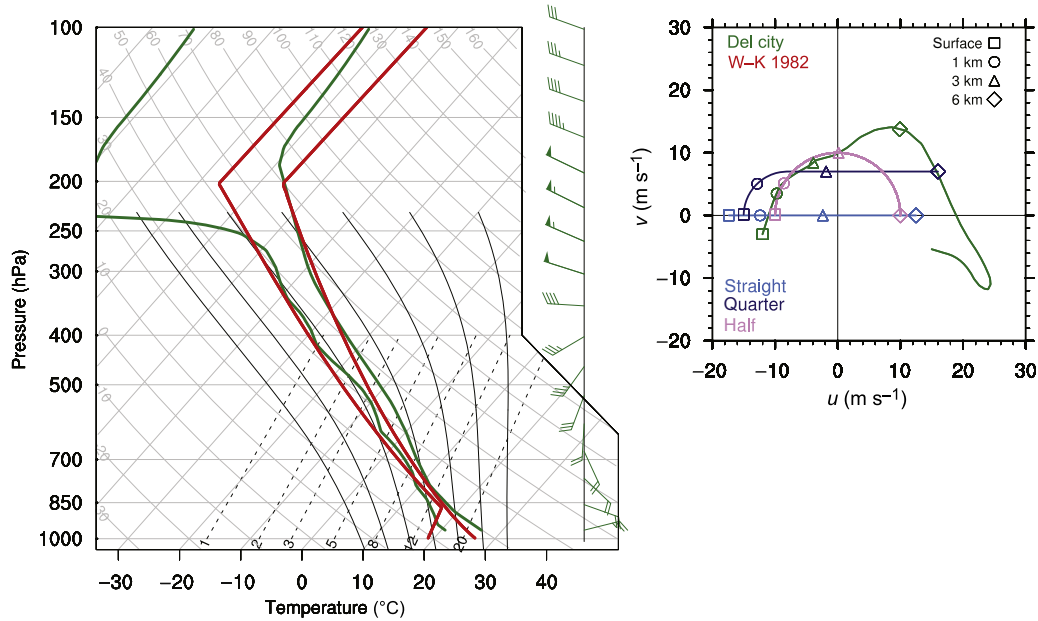


Figure 1 Skew- T log- p diagram and hodograph plot for several traditional soundings and wind profiles commonly used in idealized convective storms models. The temperature, humidity, and vertical wind profiles for the 20 May 1977 Del City, Oklahoma, tornado are plotted in green. The temperature and humidity profiles of the idealized Weisman–Klemp sounding are plotted in red. Several additional idealized wind profiles (often used in tandem with the Weisman–Klemp sounding) are plotted on the hodograph, including a simple straight-line profile (light blue), a quarter-turn supercell-like profile (dark blue), and a half-circle supercell-like profile (purple). The altitudes of the surface, 1, 3, and 6 km data points, are annotated with symbols on the hodograph curves as shown.

(i.e., to maintain large-scale thermal wind balance). Open (or ‘radiative’) lateral boundary conditions can be easily used, and even more simplified mirror or periodic conditions are possible, where the geometry of the problem supports it (an example using a periodic boundary condition is shown in [Figure 2\(b\)](#)). When combined with a free-slip model bottom, such boundary conditions allow the homogeneous environment to remain steady in time, and simple simulation results to become very nearly Galilean-invariant.

These traditional idealized model experiments have typically included few (or no) parameterizations of processes other than subgrid-scale turbulence and precipitation microphysics. Although it is now known that surface fluxes, planetary rotation, and radiation are potentially important to convective evolution, such processes were long neglected in convective storm models for reasons of computational affordability. Even today, highly idealized approaches remain popular because simpler simulations allow for controlled hypothesis testing using a small number of well-known inputs. Indeed, because of the highly nonlinear interactions in full-physics storm simulations, a wide range of even more idealized simulations (which might generally be called ‘toy model experiments’) has also been devised. Such simulations might not even include an actual convective storm (perhaps only a stormlike heat source or sink).

Using whatever degree of simplification, a single idealized model run is not especially useful by itself because its volatility is unknown. Since many details are omitted from the model configuration, exact correspondence to an observed case is unlikely. Additionally, artificial forcing is often needed in order

to produce convective development, particularly in horizontally homogeneous environments (which lack preexisting fronts and mesoscale circulations). Common artificial triggers include idealized warm thermals, cold pools, convergence zones, and surface fluxes. Typically, after storms have developed in response to the artificial trigger, research then focuses on how the storms freely evolve within the environment. However, owing to the somewhat improvisational initial forcing, only a basic credibility check of the results is often possible. Therefore, trends among multiple runs tend to be more meaningful than the specific output values from any one run. The true strengths of the idealized approach lie in robust sensitivity testing and determination of basic cause–effect relationships.

A great deal has been learned from such simulations, and they continue to have pedagogical value. However, advances in understanding and in computer power have increasingly driven the field toward more sophisticated approaches. Hybrid idealized approaches have emerged in which some horizontal or temporal variability is permitted within the context of an environment that is still closely controlled. Also, a complementary body of work has emerged using case study simulations. In such simulations, direct observations (or analyses from larger scale models) are used for the model’s initial and boundary conditions. As a result, the forcing for storms is a natural property of the large-scale environment (no artificial triggering is needed), and the expectation is that model output that will correspond to specific observations from the chosen case. In order to accomplish such realism, a model must include many more components (e.g., representations

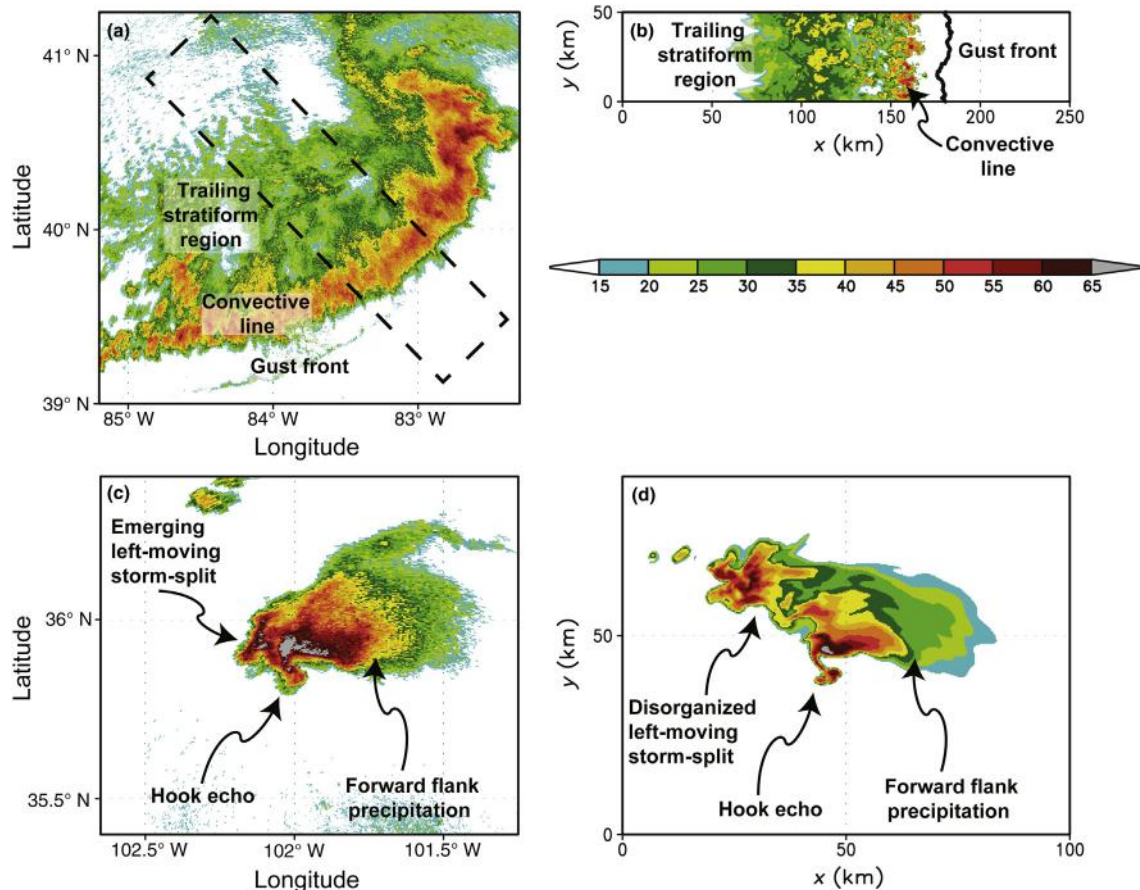


Figure 2 Plan views of radar reflectivity (shaded, dBZ) for (a) an observed squall line (Wilmington, Ohio, radar from 2113 UTC on 29 June 2012), (b) an idealized simulated squall line (using the traditional Weisman–Klemp sounding and a straight-line wind profile as shown in Figure 1), (c) an observed supercell (Amarillo, Texas, radar from 2327 UTC on 18 May 2010), and (d) an idealized simulated supercell (using the traditional Del City, Oklahoma, tornado sounding and wind profile shown in Figure 1). The size of the domain for panel (b) is shown by the dashed rectangle in panel (a). The horizontal scales are identical in panels (c) and (d). Characteristic features of squall lines and supercells are annotated. Because there are no clear-air radar targets to produce a ‘fine line’ in the squall line simulation, in panel (b) the gust front is instead indicated by a black contour of potential temperature perturbation (θ') = -2 K. All of the simulations were performed using a 250-m grid spacing. The simulated squall line in panel (b) is an example of using a periodic lateral boundary condition in the y -direction to minimize the computational expense of the domain.

of terrain, land cover, radiation, and boundary layer flux profiles).

Because numerical models (and the atmospheric equations themselves) have a sensitive dependence on initial conditions, and because there is often very little information for initial conditions on the scales of convective storms themselves, in practice it is still very difficult to replicate an observed case. Even so, such models can still be useful as internally consistent proxies for the kinds of storms that develop in a particular regime. Additionally, with advances in data assimilation techniques that have the capability to ingest Doppler radar observations, it is becoming possible to constrain case study simulations and produce results that match the placement and structure of observed storms more closely. Ensembles of simulations have also increasingly been used to quantify probabilities and uncertainties present in case study modeling results, providing a more statistical view of convective behavior. The configurations of case study simulations are also generally more appropriate for direct assessment of the

strengths and weaknesses of operational convection-resolving NWP models.

Explicit Representation vs Parameterization of Important Processes

If an important process is too small in scale to be adequately represented by a model’s grid spacing, then a ‘parameterization’ is developed (based upon theory and/or observations) that allows the model fields to respond as though the missing process was present. Models used for convective storms research are always convection permitting (i.e., the storms are not parameterized). However, the fact that the convection is permitted (occurring on the model grid) is not necessarily the same as saying that the convection is well resolved (properly represented). As in all model simulations, great caution is needed to avoid attaching significance to simulated features that are poorly resolved.

Because there is generally not sufficient computer power to run convective storm simulations at grid spacings where the solutions are ‘converged’ (that is, no longer sensitive to further decreases in grid spacing), grid spacings are generally chosen based upon the scales of the processes thought to predominate the problem of interest. Extensive testing has shown that gross mesoscale structures of convective systems begin to be resolved with horizontal grid spacings of 1–4 km (Figure 3). Because the vertical gradients of most variables tend to be comparatively large, such simulations still typically use vertical grid spacings that are <1 km. However, current conventional wisdom is that a proper representation of convective storms requires explicit treatment of the largest eddies that account for entrainment and mixing within clouds and at cloud edges. Such eddies first emerge in simulations as grid spacings decrease from roughly 500 toward 100 m (Figure 3). Sensitivity tests have shown that simulations with grid spacings >500 m have too little turbulent kinetic energy on the scales of importance to convective clouds (Figures 3 and 4).

Even within highly idealized convective storm simulations using small grid spacings, the treatments of turbulence and precipitation microphysics are of great importance. A number

of frequently neglected processes such as surface fluxes and radiation also require parameterization if they are to be included in convective storm models.

Treatment of Turbulence

Turbulent mixing is a fundamental part of convective motions in the atmosphere, and yet this important process is only marginally resolved in most modern convective storm simulations. It is still rather rare for full convective storm simulations to have grid spacings that would be consistent with large eddy simulations (LES, commonly having grid spacings below 100 m), mainly due to computational expense. When grid spacings on the order of tens of meters have been used, it has typically been in the context of shallower clouds (boundary layer cumuli and congestus, stratocumuli), or for detailed simulations of tornadoes (without their parent thunderstorms).

Even in true LES models, a subgrid-scale parameterization is needed to represent the effects of turbulent motions smaller than what the model can explicitly predict. The typical treatment assumes that the smallest resolved scales fall within the

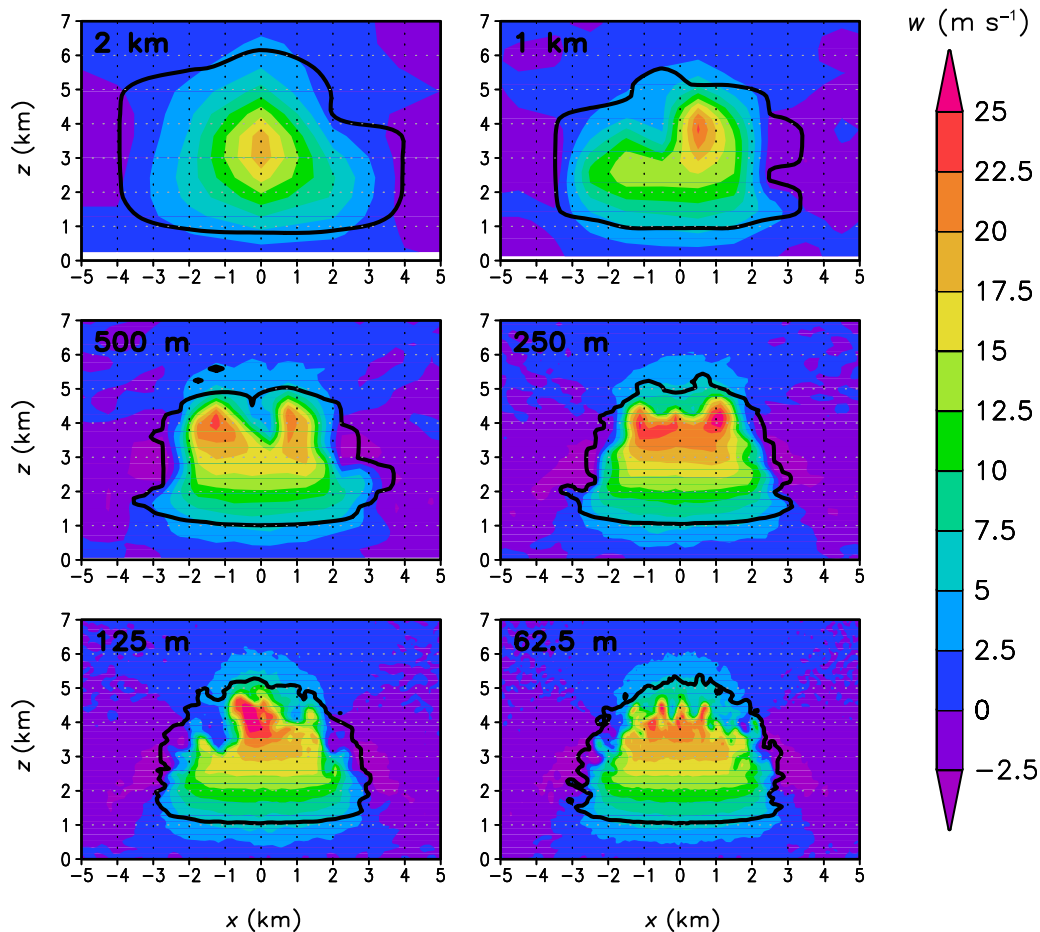


Figure 3 Contemporaneous vertical cross sections of vertical velocity (shaded) and cloud outline (black contours) for a developing convective storm in simulations with varying grid spacings (as labeled in each panel). Each simulation uses the traditional Weisman–Klemp sounding shown in Figure 1, a resting base state, and an identical warm thermal trigger with additional random noise inserted into the initial potential temperature field.

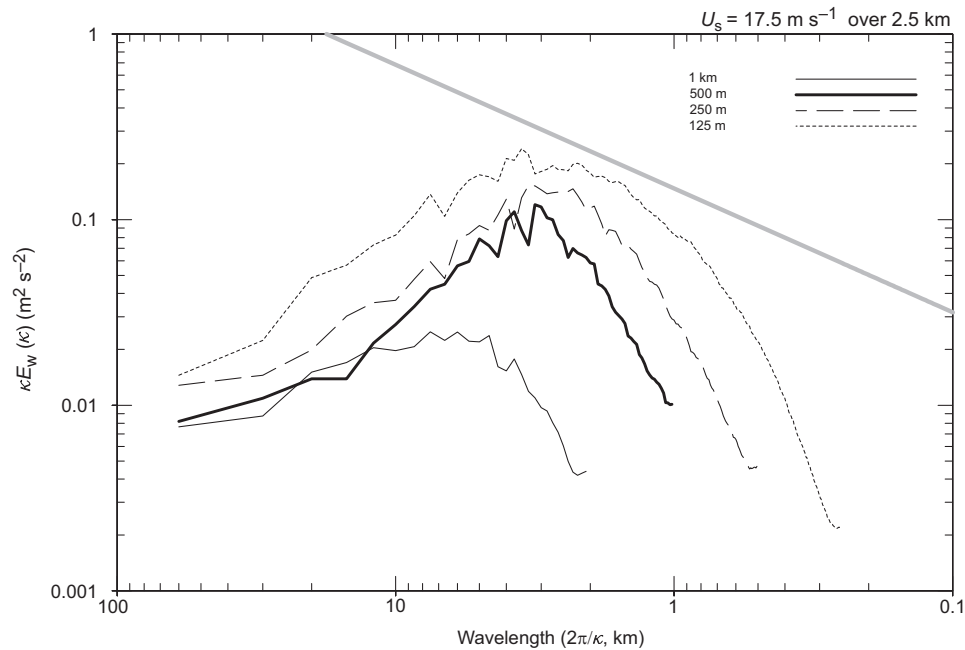


Figure 4 Contemporaneous along-line vertical velocity spectra (kinetic energy per unit mass) at 5 km above ground for a suite of idealized squall line simulations using different grid spacings (keyed to the line styles as shown). A thick gray line corresponding to a $\kappa^{-5/3}$ spectrum is included to illustrate behavior that is expected in an inertial subrange. Reproduced with permission from Bryan, G.H., Wyngaard, J.C., Fritsch, J.M., 2003. Resolution requirements for the simulation of deep moist convection. *Monthly Weather Review* 131, 2394–2416.

inertial subrange, such that basic ‘K-theory’ for eddy viscosity (or diffusivity) can be applied to represent the mean downscale energy transfer within the unresolved scales. This kind of parameterization is used in almost all convective storm models. However, it appears that the characteristic $\kappa^{-5/3}$ slope of an inertial subrange (at scales beneath the energy peak on a log–log spectrum plot, where κ is the wave number) may not be well depicted in storm simulations with grid spacings ≥ 500 m (Figure 4, where the gray line has a slope of $-5/3$ for reference).

The downward slopes steeper than $\kappa^{-5/3}$ to the far right in Figure 4 are due to excessive energy losses from numerical filtering, which limits the effective resolution of the model; on coarser grids, this limitation prevents faithful depiction of even the largest eddies within the inertial subrange, and the $\kappa^{-5/3}$ regime vanishes. Based on such analyses, grid spacings of roughly 250 m (or less) seem to be the minimal requirement for appropriate convective storm modeling with LES-like sub-grid-scale parameterizations. Computer power has now increased to the point that such runs are practical for most experiments. However, smaller scale convective features (e.g., tornadoes) require still finer grid spacings than this.

Treatment of Cloud and Precipitation Microphysics

The treatment of cloud and precipitation microphysics is entirely parameterized in all modern models, and is one of the most important components of convective storm simulations. Microphysics is extremely influential on the dynamics, contributing to heating in regions of condensation, riming, and freezing, to cooling in regions of evaporation, melting, and

sublimation, and to downward accelerations from hydrometeor loading. The proper representation of hydrometeors is also important for the correct treatment of radiation (i.e., shading and absorption of upwelling longwave radiation) in models that include such processes.

Microphysical parameterization also may be the largest source of uncertainty in modern convective storms models. Recent studies have shown large sensitivities to seemingly minor features such as details of the hail category (whose subsequent fall speeds partly determine the horizontal footprint of convective precipitation) and treatment of raindrop breakup (since drop sizes have a major influence on fall speeds and evaporation rates). Unfortunately, many such components are rather poorly constrained by measurements. Therefore, substantial uncertainties exist even in the most modern microphysical parameterizations.

First-generation microphysical treatments (a popular example in convective storms modeling is the well-known ‘Kessler scheme’) included only bulk treatments of water vapor, cloud droplets (which are assumed to move with the flow and have a negligible fall speed), and raindrops (which are assumed to have a nonzero fall speed). Most such schemes assume that cloud droplets are monodisperse and that raindrops have the commonly observed inverse-exponential (Marshall–Palmer) size distribution. The slope of the drop size distribution is generally related to the predicted rainwater mixing ratio with the idea that, as mixing ratio increases, the number of small drops is relatively unchanged while the number of large drops increases the most (proportionally). Such schemes usually assume that cloud water is produced instantaneously when supersaturation exists, whereas rain is

produced over time from cloud water (e.g., via collision and coalescence). Physical relationships and empirical formulations are then used to represent the rates of important processes such as conversion of cloud to rain, evaporation, and fallout.

Such liquid-only parameterizations produced storms that were recognizable, and enabled early studies to diagnose fundamentals such as the mode of convective organization. However, some of the simulated features were unrealistic, especially because of the exclusion of ice. The subsequent addition of bulk treatments for cloud ice and snow (in a similar fashion to cloud water and rain) was shown to improve the realism of convective storm anvils as well as zones of moderate (stratiform) precipitation within mesoscale convective systems (MCSs). These improvements are largely attributable to the much smaller fall speeds of snow (compared to rain), but also because the inclusion of freezing and melting adds realism to the vertical structures of updrafts and downdrafts. Recently, some advanced schemes have further differentiated between unique subcategories of snow (columns, plates, dendrites, aggregates, etc.), presumably because the growth and fallout properties of ice are sensitive to the predominant ice crystal habit.

At the very least, most modern bulk parameterizations also include a rimed ice category for added realism. This category is ordinarily tuned to represent either hail or graupel, with the primary difference being the assumed density of the particles (traditionally, 900 kg m^{-3} for hail vs 400 kg m^{-3} for graupel). This choice in turn influences the computed fall speeds of the rimed ice. For many historical convective storm simulations, hail has been preferred. Hail particle trajectories are more clearly distinct from snow, providing more realistic horizontal distributions of light vs heavy precipitation in supercells and squall lines. The choice of hail also generally leads to more realistic (weaker) surface cold pools due to the larger fall speeds and smaller surface area per unit volume (in comparison to graupel). Of course, in nature there is a spectrum of densities for rimed ice particles and, in stratiform regions of MCSs for example, the less dense graupel category may be more realistic. Some recently developed parameterizations now either predict both categories of rimed ice, or else allow for variable graupel/hail density as a predicted function in the model (based upon the assumed hail growth regime).

Contemporary studies have also suggested that different regions within storms are better described by different ‘intercepts’ of the inverse-exponential size distribution. Such results imply that it may not be possible to configure a single-moment scheme (which predicts only the hydrometeor mixing ratios) to properly handle the entire system. Therefore, a present trend in convective storms modeling is toward multimoment microphysical parameterizations. Double-moment schemes predict total number concentrations (in addition to mixing ratios), which allows for diagnosis of the ‘intercept’ of the inverse-exponential size distribution. Therefore, the concentrations of large and small drops can evolve somewhat independently from one another. This provides a better representation of processes such as size sorting and the preferential evaporation of smaller drops. Recently introduced triple-moment parameterizations allow even more extravagant descriptions of the size distribution (i.e., deviations from the simple inverse-exponential relation). Tests have shown that some

multimoment schemes produce more realistic convective cold pools and reflectivity structures than single-moment schemes (Figure 5). However, these higher order schemes also have many more free parameters. It is not always clear whether such settings are well constrained by observations or theory, and yet they have been shown to produce very large model sensitivities.

Some modern models are also now beginning to predict the concentrations of condensation aerosols interactively, using them as input to the precipitation scheme. Aerosols may influence the hydrometeor size distributions, both in terms of initial droplet concentrations and in terms of the rates and locations of subsequent growth by coalescence and riming. In turn, as the sizes and spatial arrangements of hydrometeors change, the footprint of potential cooling in the storm also may change. These effects have been shown to be nonzero, although there is still debate about their relative importance in comparison to the basic environmental ingredients that strongly control storm dynamics.

Apart from the above bulk microphysical schemes, there are also a number of models using ‘bin-resolving’ treatments that represent the hydrometeor size distributions through the use of discrete ‘bins.’ These bins are defined based on ranges of mass or size (with mass doubling being a typical approach for determining the thresholds). Because no a priori shape is assumed for the particle size distributions, they are free to evolve independently, with the mass and/or number concentrations of hydrometeors in each bin being continually predicted. Up until recently, the number of calculations required for each individual bin was considered too computationally expensive, and the approach was thus rare. However, with advances in computer power, it has become more practical (although still requiring limits in domain size). Some researchers have also used ‘bin-emulating’ schemes, which achieve much of the accuracy of bin schemes (while maintaining the computational efficiency of a bulk scheme) by only temporarily partitioning a bulk size distribution into bins and then deriving microphysical tendencies from predetermined lookup tables.

Treatment of Other Subgrid-Scale Processes

The inclusion of surface fluxes, realistic boundary layers, and radiative forcing is increasingly viewed as both practical and desirable for convective storm simulations. Historically, such features have often been ignored both for computational efficiency and to simplify interpretation of the primary dynamics involved in convection–environment interactions. For example, in runs with surface fluxes of momentum (i.e., surface drag), the traditional use of a homogeneous base state becomes complicated because there is no large-scale pressure gradient to oppose deceleration of the flow by surface drag. Naturally, real storms do occur in environments where surface fluxes of buoyancy and momentum drive an actively mixed boundary layer; also, real storms experience feedbacks from shortwave shading produced by their cloudiness, as well as net longwave cooling at their cloud top levels. The addition of parameterizations for these processes has begun to provide a clearer understanding of their roles in storm dynamics.

It appears likely that surface fluxes (particularly of momentum) are necessary to properly represent the

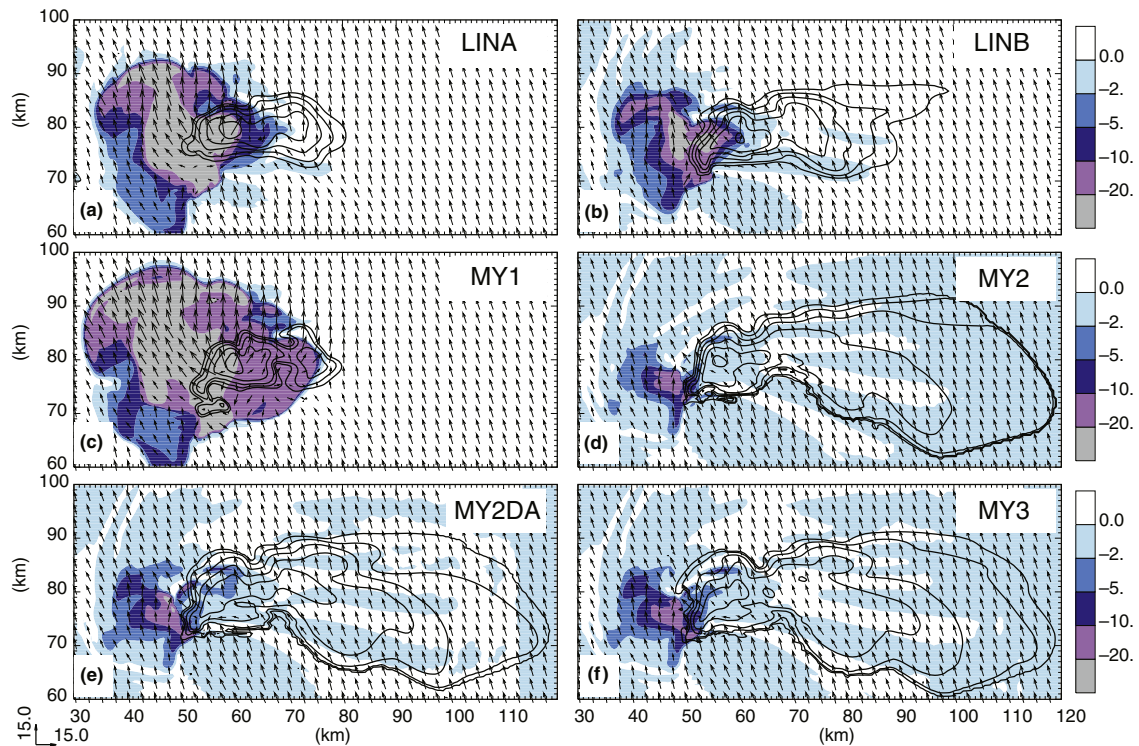


Figure 5 Contemporaneous plan views of surface equivalent potential temperature perturbation (shaded), reflectivity (contoured with a 10-dBZ increment), and horizontal ground-relative wind vectors (plotted every 2.5 km, 1 step = 15 m s^{-1}) for simulations using a 500-m grid spacing and the following microphysical parameterizations (all of which include ice): (a) the Lin–Farley–Orville single-moment scheme, (b) the Lin–Farley–Orville single-moment scheme with a reduced rain intercept (producing larger drops on average), (c) the Milbrandt–Yau single-moment scheme, (d) the Milbrandt–Yau double-moment scheme, (e) the Milbrandt–Yau double-moment scheme with a diagnostic relationship for the shape of the size distributions (simplistically emulating a triple-moment scheme), and (f) the full triple-moment Milbrandt–Yau scheme. Adapted with permission from Dawson, D.T., Xue, M., Milbrandt, J.A., Yau, M.K., 2010. Comparison of evaporation and cold pool development between single-moment and multimoment bulk microphysics schemes in idealized simulations of tornadic thunderstorms. *Monthly Weather Review* 138, 1152–1171.

tornadogenesis process as it occurs in nature. Indeed, some recent studies have found surface friction to be a leading term in the calculated vorticity budget. Realistic variations in soil moisture within such runs may also dramatically modulate the environmental boundary layer relative humidity (and thus convective available potential energy (CAPE) and evaporative potential). Other recent results also suggest that, although the net effect of boundary layer circulations may be modest, storms become increasingly unsteady as they interact with boundary layer rolls and eddies in simulations with heterogeneous surface fluxes. Finally, it appears that stabilization from convective anvil shading can have a noticeable influence on both low-level stability and the low-level vertical wind profile (due to changes in vertical mixing). Such lower tropospheric changes can directly impact the storm dynamics.

When the aforementioned processes are included, some of the traditional tenets of convective storm modeling have to be reexamined. For example, in classical simulations with free-slip bottom boundaries (i.e., no surface fluxes), the dynamics are presumed to be Galilean invariant: this enables scientists to run simulations in which the storm remains conveniently stationary within the model grid. However, surface fluxes are sensitive to the ground-relative wind, and summed radiative tendencies are influenced by the grid-relative motion of cloud.

So, larger domains are generally needed in order to use native wind profiles in simulations with these processes included.

Further parameterizations of processes like electrification and lightning are also sometimes used, although these tend to be passive end products of the dynamical and microphysical fields, and do not feed back upon them (with the exception that lightning parameterizations are sometimes coupled to chemistry models, a topic not addressed in detail here).

Historical Applications and Results

Convective storm models have a history of success in replicating most observed features of storms. In turn, analysis of output from such models has helped to spell out the ways in which convective evolution is linked to the environmental profiles of winds, temperature, and humidity. Because of their great societal impact, squall lines (including MCSs) and supercells have received the most attention (Figure 2), with much investigation of their dynamics and environmental sensitivities.

The earliest models of convective storms were for ‘proof of concept,’ and mostly used either axisymmetric or 2D (slab) symmetric geometry with only crude microphysics. From such

models, basic explanations for processes such as downdraft forcing and density current evolution began to emerge. Subsequent 2D and 3D runs (beginning in the late 1970s) established the important role of vertical wind shear in determining storm type. It became clear that, in undisturbed environments, disorganized ordinary cells would occur in weak vertical wind shear, with multicells in moderate shear (e.g., [Figure 2\(a\)–2\(b\)](#)) and supercells in strong shear (e.g., [Figure 2\(c\)–2\(d\)](#)). Sensitivity tests showed that the transitions between these storm types can be surprisingly abrupt. Likewise, early simulations began to establish the important role of CAPE (a metric that was not nearly so widely used prior to the convective storm modeling era) upon storm structure, intensity, and outflow production. Subsequent climatologies have largely validated the parameter sensitivities suggested by models, and in turn the models have helped to explain the physical reasons why the sensitivities exist.

Squall Lines and Mesoscale Convective Systems

Although almost every modern simulation is now carried out in 3D, squall lines were one of the first storm types to be studied extensively because their natural structures (with much larger across-line gradients than along-line gradients, e.g., [Figure 2\(a\)](#)) could be reasonably represented in affordable 2D slab-symmetric models (or smaller periodic 3D ‘channel’ models, as in [Figure 2\(b\)](#)). Early modeling results helped to articulate the role of interactions between the system outflow and the lower tropospheric line-perpendicular environmental vertical wind shear. A combination of simulations and theory explained the tendency for squall lines to exhibit continued redevelopment of convection on the down-shear edges of their outflows, the propensity for squall lines to produce severe winds and heavy precipitation in strongly sheared environments, and the common observation that squall lines tend to evolve toward a structure with trailing stratiform precipitation as their cold pools intensify.

Once such fundamentals were established, subsequent modeling work helped to articulate how important features such as rear inflow jets, midlevel mesoscale convective vortices, and near-surface line-end vortices form. Many of these phenomena had already been capably studied using observations, but model simulations provided the first truly controlled experiments that isolated the most important processes. In tandem, some extremely simple (toy model) simulations revealed that the vast majority of mesoscale circulations within and around squall lines and MCSs can be regarded as simple responses to the footprint of latent heating in the system. In understanding the dynamics of these mesoscale features, a comprehensive conceptual model emerged for the genesis of bow echoes and severe surface winds as well. The important roles of both line-parallel vertical wind shear and upper tropospheric vertical wind shear also became increasingly apparent from simulations, and such findings enabled a more realistic representation of a wider range of observed MCS structures and behaviors.

The era of widespread 2D squall line simulations came to a close shortly after the turn of the century, as the dynamics of the constituent convective cells were repeatedly shown to be decidedly 3D. Models revealed that many squall lines possess

low-level slablike lifting, but discrete cellular updrafts farther aloft, with the height and abruptness of this transition being very strongly linked to the cold pool strength. The regular pulsing of the individual updrafts in squall lines was further linked to a self-driven mechanism that cuts the cells off the gust front, and modeling studies also revealed that gust front lifting can produce moist absolutely unstable layers, which in turn break down into embedded 3D roll-like circulations. Although such results have not generally invalidated the original conclusions from most of the pioneering 2D modeling studies, they provide a much clearer explanation for many of the small-scale features that are regularly observed (e.g., with radar) in real-world squall lines and MCSs.

Recently modeling work has also advanced the knowledge of several operationally important subsets of squall lines. Nocturnal and elevated (that is, having inflow well removed from the surface) MCSs are common in nature, and simulations have provided a clearer understanding of the mechanisms responsible for sustaining squall lines during the overnight hours and in environments where a surface cold pool does not develop (or is quite weak). Such studies have highlighted the important roles of gravity waves and bores in MCS maintenance, and the intricate interplay between stability and vertical wind shear in such cases. Finally, model simulations have largely driven the discussion of possible mechanisms for the formation of intense surface mesovortices within the gust fronts of mature squall lines, which remain poorly understood and present significant hazards to society.

Supercells and Tornadoes

Once 3D simulations became practical, a great many idealized simulations focused on understanding the basic dynamics and environmental sensitivities of supercells. Much of the earliest evidence for supercellular storm splitting came from simulations (and the behavior was subsequently observed to be quite common, e.g., [Figure 2\(c\) and 2\(d\)](#)). Confounding some intuitive speculations, models further showed that storm splitting was primarily a dynamically driven process, and did not require the presence of a precipitation-driven downdraft. Many far-reaching impacts of dynamic pressure effects upon supercell motion, intensity, and longevity were largely demonstrated with models, including the understanding that right-moving supercells are favored over left-moving supercells in most midlatitude Northern Hemisphere environments because the vertical wind shear vector turns clockwise with height.

Appreciation of these modeled dynamical influences also led to the maturation of diagnostic pressure analysis in convective storms research, and the related partitioning of accelerations into buoyant and dynamic components. Such approaches demonstrated, for example, that the sounding-based approach of inferring vertical motion from CAPE was likely to substantially underestimate the strength of supercell updrafts, which profit from additional upward accelerations due to dynamical contributions.

Prior to the era of 3D storm modeling, the origins of vertical vorticity and its subsequent evolution in supercells had long been the subject of interest. Gridded model output fields permitted computation of reasonably accurate trajectories and

budgets, demonstrating the source regions for air parcels in different parts of the storm, and enabling the assessment of circulation around material circuits that are advected by the model winds. Such techniques confirmed that the primary source of midlevel updraft rotation was the horizontal vorticity of the sheared background environment, whereas the source for supercells' low-level mesocyclones was revealed to be baroclinic generation of horizontal vorticity in the forward flank. This knowledge led to a new generation of tornadogenesis hypotheses invoking baroclinity and emphasizing the importance of the storm downdrafts and outflow.

It also had been long surmised that supercells interacting with preexisting outflow boundaries or fronts might intensify (and become more likely to produce a tornado). The beneficial contribution of enhanced low-level shear on the cool sides of such boundaries (as compared to the hindering effect of increased stability) was cleanly isolated in controlled model experiments that studied the same storm both with and without a boundary crossing.

Many such modeling studies have assessed tendencies in the low-level rotation of supercells, and have speculated about the possible ramifications to tornado formation. However, proper representation of a tornado likely requires model grid spacings on the order of 10 m. Therefore, most historical tornado simulations focused only on the vortex itself within a small domain, omitting the parent supercell for computational economy. Such studies have revealed a great deal about tornado structure and dynamics, but the real-world context for tornado formation was missing. Amazingly, some of the most central tenets of modern tornado theory emerged from very simple (toy model) simulations that demonstrated the need for a downdraft in order to produce a surface vortex in environments without preexisting vertical vorticity.

Early attempts to model full supercells with tornado-like vortices normally involved grid nesting near the tornado; unfortunately, it was never clear how the abrupt change in model resolution influenced the production and strength of the simulated vortex. The computers of today (and the future) make it possible to use high resolution over a much greater fraction of the model domain and for a much longer period of simulation. As an increasing number of studies attempt this approach, proper treatments of surface drag and small-scale turbulence are emerging as critical components of the experiments.

Future Directions

A great deal has been learned from convective storms modeling since the 1970s, and yet some very basic processes remain mysterious. For example, the genesis mechanisms of tornadoes and gust front mesovortices are still not well understood, and it is not yet clear how storms develop and maintain themselves in a variety of inhospitable environments. Given ongoing difficulties in making the small-scale measurements needed to address such questions, models will likely remain a cornerstone of convective storms research.

As computer power advances, the trends toward decreasing grid spacing and more sophisticated scale-appropriate parameterizations will no doubt continue. In addition, greater numbers of simulations will be possible for an individual experiment or

forecast. This will enable scientists to explore a much broader parameter space (in terms of convective ingredients), and also to exploit ensemble techniques (which specifically quantify the uncertainties in experiments and forecasts).

Larger ensembles also add value to modern data assimilation techniques (e.g., the ensemble Kalman filter). For example, assimilation of radar data into both case study and idealized model simulations has already become practical. Radar data may be assimilated at intervals of a minute or less, and preliminary results suggest very good agreement between observed and simulated storms. Such techniques will provide additional observational constraints upon research simulations, and will open up the possibility of deterministic or probabilistic short-term severe weather forecasting using convection-permitting models.

Regional forecast models already marginally resolve convection, and regional climate models will soon. This will likely open up new linkages to the forecasting of hydrology and severe weather, and to prediction of changes in convective weather under future climate scenarios. Meanwhile, idealized research simulations will continue to incorporate more realistic features, including long-neglected processes such as radiation, boundary layer circulations, and storm interactions with environmental heterogeneity and terrain. It will be of interest to learn how such complexity influences our view of convection, which currently is largely based on sensitivities to environmental ingredients within highly simplified model configurations.

See also: **Clouds and Fog:** Cloud Microphysics; Cloud Modeling. **Mesoscale Meteorology:** Convective Storms: Overview; Hail and Hailstorms; Mesoscale Convective Systems; Severe Storms. **Numerical Models:** Cloud-System Resolving Modeling and Aerosols; Large-Eddy Simulation; Mesoscale Atmospheric Modeling; Methods; Model Physics Parameterization.

Further Reading

- Bryan, G.H., Morrison, H., 2012. Sensitivity of a simulated squall line to horizontal resolution and parameterization of microphysics. *Monthly Weather Review* 140, 202–225.
- Bryan, G.H., Wyngaard, J.C., Fritsch, J.M., 2003. Resolution requirements for the simulation of deep moist convection. *Monthly Weather Review* 131, 2394–2416.
- Dawson, D.T., Xue, M., Milbrandt, J.A., Yau, M.K., 2010. Comparison of evaporation and cold pool development between single-moment and multimoment bulk microphysics schemes in idealized simulations of tornadic thunderstorms. *Monthly Weather Review* 138, 1152–1171.
- Klemp, J.B., Wilhelmson, R.B., 1978. The simulation of three-dimensional convective storm dynamics. *Journal of the Atmospheric Sciences* 35, 1070–1096.
- Meng, Z., Zhang, F., 2011. Limited-area ensemble-based data assimilation. *Monthly Weather Review* 139, 2025–2045.
- Morrison, H., Milbrandt, J., 2011. Comparison of two-moment bulk microphysics schemes in idealized supercell thunderstorm simulations. *Monthly Weather Review* 139, 1103–1130.
- Pielke, R.A., 2002. *Mesoscale Meteorological Modeling*, second ed. Academic Press, New York.
- Stensrud, D.J., 2007. *Parameterization Schemes: Keys to Understanding Numerical Weather Prediction Models*. Cambridge University Press, New York.
- Weisman, M.L., Klemp, J.B., 1982. The dependence of numerically simulated convective storms on vertical wind shear and buoyancy. *Monthly Weather Review* 110, 504–520.
- Wilhelmson, R.B., Wicker, L.J., 2001. Numerical modeling of severe local storms. In: Doswell, C. (Ed.), *Severe Convective Storms*, Meteorological Monograph, vol. 28 (50). American Meteorological Society, Boston, pp. 123–166.

OBSERVATIONS PLATFORMS

Contents

Balloons

Buoys

Kites

Radiosondes

Rockets

Balloons

J-P Pommereau, LATMOS, CNRS, Guyancourt, France

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

Balloons are unique tools for studying the atmosphere. In contrast to satellites providing a global view of the Earth's atmosphere at broad resolution, they allow investigating atmospheric processes at high resolution. Following the meteorological balloon-sondes at the end of the nineteenth century, powerful vehicles have been designed using plastic films available since 1947. Since then a number of balloon platforms have been designed spanning from zero pressure balloons flying up to 40 km altitude, spherical or pumpkin super-pressure balloons and Infrared Montgolfier allowing long duration flights for several months, and tethered balloons for studying the lower atmosphere.

Introduction

Though a large number of high-performance satellite instruments have been placed into orbit during the last decades, balloons are used more frequently than ever in studying the atmosphere, for several reasons. Indeed, if satellites are unique in providing a global view of the Earth's atmosphere, they also suffer several limitations. First, because of atmospheric attenuation and clouds, chemical species as well as meteorological parameters can only be observed from satellites with difficulty at altitudes below 20 km, i.e., in the lowermost stratosphere where ozone depletion takes place, and in the troposphere where the impact of human activities could be the largest. Moreover, the vertical resolution of the measurements from satellite is limited to 1–2 km at best, and several important chemical species could not be derived by remote sensing techniques. Finally, satellites have a limited lifetime. Except for ozone, at present they do not allow the long-term monitoring of the composition of the atmosphere, a vital necessity in understanding the impact of anthropogenic sources and the relation between chemistry and climate.

In contrast, currently available balloons allow a variety of *in situ* and remote sensors, making measurements from the

ground to about 40 km, to be carried at relatively low cost, thus allowing the repetition of the measurements over a long period. Importantly also, balloon flights can be performed within a shorter time frame than that required for the development of space projects, allowing the rapid checking of new ideas, concepts, or instruments for further use in space and later the validation of the measurements of spaceborne instruments.

Furthermore, long-duration balloon systems of various types are currently available or under development, which should allow greater use of balloons in atmospheric studies on a global scale at a relatively low cost.

Just as meteorological satellites did not replace the need for several hundred daily radiosonde ascents, so space instruments did not replace the use of balloons, but in contrast resulted in their considerable development. This, together with the use of modern space technology in the instruments and payloads design, i.e., Iridium satellite transmission, Global Positioning System (GPS), explains why scientific ballooning has been taken up by space agencies in most countries.

The objective of this article is to give an overview of the unmanned balloon systems currently available or under development for atmospheric research. After a brief historical recall of scientific ballooning and of those conceptual aspects needed to understand how balloons work and their

limitations, the main systems currently available will be described from the point of view of the scientific user.

History

Balloons are not new vehicles. They all still rely on one of the two concepts both flown for the first time in 1783 in France: hot air by the Montgolfier brothers and gas by Charles and Robert. As a first step for about a century, these balloons were used primarily for spectacular events during celebrations, were all manned, and were used little for science. The first observations of scientific interest were those of Robertson in 1803 and of Gay-Lussac and Biot in 1804. They demonstrated the decrease of temperature, pressure, and moisture with height and the constant composition of air up to an altitude of 7000 m. But progress was slow, limited by the maximum altitude acceptable to the pilots. Only a handful of ascents of scientific interest were performed during the nineteenth century. Barral and Bixio in 1850 in France discovered the existence of ice particles – cirrus clouds – forming at low temperatures at 7000 m. In 1852, Welsh and in 1862, Glaisher in the United Kingdom extended observations to a record altitude of 8800 m. Then, Sivel, Croce-Spinelli, and Tissandier in France in 1875 made the first spectroscopic observations of astronomical bodies above the dense atmosphere at 8600 m, during an ascent in which the first two men perished. Finally, Berson in Germany, reaching 9150 m in 1894 and 10 300 m in 1901, measured a record low temperature of -47.91°C .

Though manned flights were continued into the next century (Piccard in 1931 reached 15 781 m, followed shortly after by Prokofiev, Goudonov, and Brinbam in the Soviet Union at 18 500 m, Settle in the United States at 18 665 m, and several others until Kittinger reached 30 000 m in the United States in the 1960s), the greatest progress in atmospheric science and meteorology was to come from the use of unmanned balloons.

The first series of unmanned ascents for studying the upper atmosphere was performed by Hermite and Besançon in 1892 in France; they used an onboard recording thermometer, barometer, and hygrometer designed by the meteorological instrument manufacturer Richard. These were recovered after the flight. This was followed by the installation of the first upper air sounding station at Trappes near Versailles by the meteorologist Teisserenc de Bort. Thanks to this effort, the stratosphere was discovered in 1898. The stratosphere was shown to be a region, where the temperature did not continuously decrease to the absolute zero around 50 km as thought before, but instead leveled off or even increased above 12–13 km. The next important technical step forward was that of Assmann in Germany, who in 1901 suggested the use of small rubber dilatable balloons and a parachute for safe recovery of the instruments. With such a system, a record altitude of 37 700 m was reached at Pavia in Italy in 1912. The further significant step is due to Hergesell, who in Germany in 1910 performed the first wind sounding using a theodolite on the ground to follow the horizontal motion of the balloon during ascent, while the altitude of the balloon was reckoned using a simple stopwatch. But the major breakthrough in atmospheric science was the invention of the radiosonde by Idراع and Bureau in

1929, who added a radio transmitter to send the temperature, pressure, humidity, and wind information in real time, thus eliminating the need to wait for an unpredictable and sometimes very long recovery of the sonde. Though many improvements have been added since (e.g., neoprene balloon material, much more sensitive sensors, miniaturized electronics, Omega, and later the GPS for the location of the sonde), the radiosondes in use today in the upper-air network of the World Meteorological Organization are basically the same as in 1929.

However, because of the restricted load permitted below rubber balloons, or the weight and therefore the altitude limitation of manned systems, the use of balloons for science other than meteorology remained limited until the arrival of plastic film in 1947 developed by Winzen at General Mills in the United States. Astronomers or cosmic ray scientists as well as atmospheric scientists were immediately interested in the new technique to carry heavy payloads above the absorbing atmosphere. Modern scientific ballooning had started by the late 1950s in the United States under the direction of Ney at the University of Minneapolis, later transferring to the National Center for Atmospheric Research (NCAR), and then to the National Aeronautic and Space Administration (NASA). Soon after, scientific ballooning activities were also started in France by Blamont at the Centre National de la Recherche Scientifique, later transferring to the Centre National d'Etudes Spatiales (CNES). Thenceforth, the technology propagated rapidly in the 1960s in the Soviet Union, Japan, India, Indonesia, Brazil, and Argentina, though complete balloon manufacturing capabilities were not available in most of the latter countries.

A variety of balloons were progressively made available to atmospheric scientists, and ranged from open or zero-pressure balloons carrying heavy payloads for a few hours at high altitude to long duration for a few weeks or months in the lower atmosphere. Though their performances and uses varied, they all followed the same physical principles, which are recalled below before describing current available platforms.

Balloon Concepts

The lift of a balloon, derived from Archimedes' principle, can be expressed as

$$V(\rho_{\text{air}} - \rho_{\text{gas}}) = \Sigma M_s(1 + f)$$

where V is the volume, ρ_{air} and ρ_{gas} are the densities of air and the lifting gas (hydrogen, helium, or hot air), respectively, ΣM_s is the sum of solid masses (balloon envelope and payload), and f is the free lift. Getting off the ground requires a positive free lift or excess gas of generally 15–20% of the total $M_s(1 + f)$. Because of the exponential decrease of pressure with altitude, the balloon expands continuously in volume during the ascent, until it reaches its float altitude. While reaching this level the excess gas corresponding to the free lift has to be evacuated or else retained by a high-strength material – otherwise the balloon bursts.

In one type of balloon, the zero-pressure, large open ducts at the base of the envelope vent the excess gas, so that the

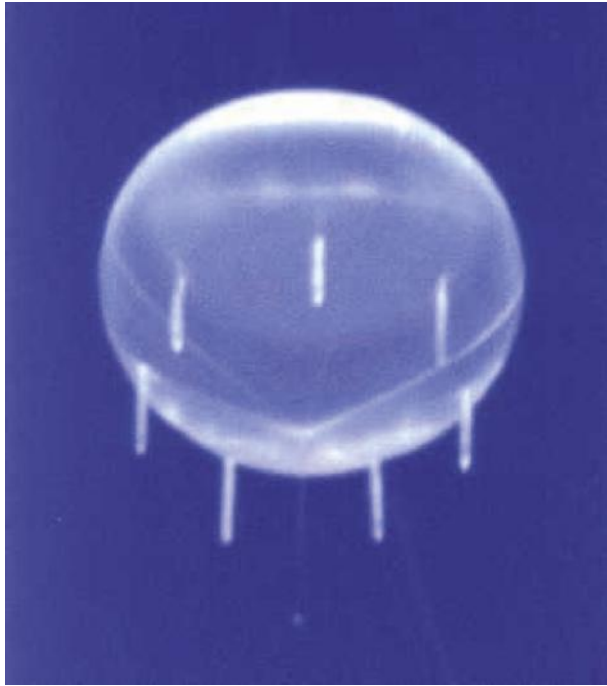


Figure 1 Zero-pressure balloon of 2 000 000 m³ at float altitude. The excess gas is vented by seven open ducts at the bottom of the envelope. © NASA.

pressure differential from inside and to outside remains small (Figure 1). In the other, the superpressure sealed balloon, the excess gas converts into a differential pressure that is retained by the high strength, and thus necessarily heavy material.

Since the strength on the envelope is very small, a zero-pressure balloon can be manufactured in light 20–40 μm thin polyethylene material. Its volume can be as large as required – up to 2 million m³ for carrying a load of several hundred kilograms up to 40 km. However, there are two limitations: launch operations and flight duration. Though efficient methods (crane or auxiliary balloon) have been developed to keep the payload off the ground at liftoff, the size of the balloon, which is only partially inflated at ground (300 m high for the largest balloons), makes a launch very sensitive to surface wind. The maximum acceptable wind speed is generally 2–3 m s⁻¹ at 50 m height, or even at 200 m for the largest balloons, which sometimes means waiting for weeks for the right conditions. The second limitation is the duration of the flight. The cooling at sunset makes the balloon to contract and descend, and this can be counteracted only by an irreversible drop of ballast. The duration of the flight is thus limited by the amount of ballast, which must be carried only at the expense of the scientific payload. Zero-pressure balloon flights are generally limited to few hours. However, long-duration cruises of up to 4–5 days across the Atlantic or the Soviet Union have been effected through carrying a large amount of ballast or alternatively by replacing the solid ballast by a reservoir of liquid helium to compensate the loss of gas. Longer flights have also been achieved in the specific case of permanent day or night in polar areas, thus requiring little dropping of ballast. Taking advantage of such conditions, zero-pressure circumnavigations of 2–3 weeks (1 month in 2002) have been achieved over Antarctica by NASA in the summer and also by the Japanese.

Long-duration flights generally require totally different approaches. Two successful ideas, both currently in use, have been suggested: superpressure balloons and Infrared Montgolfier. The first promoted by Lally at NCAR in the 1960s is derived from a concept of small spherical paper balloons used by the Japanese during World War II. The superpressure balloon is a sealed sphere in which the excess gas corresponding to the free lift converts into overpressure. The pressure increases during daytime solar heating and decreases with the cooling at night. Since the volume and the mass of the system are constant, the balloon flies at a constant density (isopycnic) level. It remains aloft until leakage reduces the overpressure to zero in the night cold, when the balloon irreversibly drops. The system is limited by the thickness and thus the weight of the material – which is generally polyester fabric or polyethylene composite – that is required to maintain the overpressure. Since the stress of the envelope increases with the balloon's radius, the size of a superpressure balloon for a given material is limited to the volume at which the balloon can carry its own unladed weight. For a spherical balloon of polyester, the limit is around 12 m diameter, allowing a payload of 30 kg at around 19 km altitude.

To overcome this limitation, a new design has been suggested in France (CNES Stratospheric Super-Pressure Balloons), in the United States (NASA Ultra Long Duration Balloons (ULDB)), and also in Japan, which consists of a 'pumpkin-shaped' balloon. In this design, Kevlar tendons take the meridional stress while the longitudinal stress is reduced by the small local radius of the pumpkin lobes or gores, which allows a reduction in the density of the balloon material. Though this solution seems promising, there are technical difficulties still to be overcome, particularly in the arrangement between the tendons and the envelope when reaching float altitude. Though a successful 3-month flight was achieved in 1992 by CNES in South Africa, other attempts have failed. The same happened at NASA with its ULDB Program started in 1998. Although some test flights have been successful in Antarctica during the polar day, all other attempts in Australia or Kiruna have failed. However, there is no doubt that once these problems are solved, the pumpkin shape could offer a unique opportunity for suborbital long circumnavigations of large payloads at high altitude for astronomy and the remote observation of the atmosphere.

A totally different concept, suggested by the present author and Hauchecorne in the late 1970s, is the use of a hot-air balloon heated at night by the Earth's thermal emission, and therefore named Infrared Montgolfier (MIR). This is achieved by adding a very low thermal-emission aluminum layer to the upper half of the infrared absorbing polyester envelope, thus preventing the balloon from radiating toward space, the lower half being in transparent polyethylene. In this arrangement, the temperature differential of 20–30 °C between the air inside the balloon and that ambient air in the cold stratosphere keeps the MIR stable at night. The extra lift provided by solar heating during daytime makes the balloon ascend and thus purge a large part of the lifting gas. But since this gas is air, it can be replaced during the evening descent by simply keeping open a large mouth at the bottom of the balloon. The system is thus reversible. The only limitation is during flying over very cold and low-emissive clouds, such as high-altitude anvils in tropical regions, which reduces the temperature differential between the inside and outside by a few degrees, which is not

enough to keep the balloon aloft for the whole night. In its present design, the MIR of 45 000 m³ volume carries 60 kg at 28 km during daytime and between 18 and 24 km at night, depending on the cloud cover, for an average duration of 3 weeks, with a record flight of 2 months in the tropics.

The various techniques above could also be combined. An example occurred in the recent record manned flight around the world, which used a helium balloon on top of a hot-air system, both covered by an aluminized layer to reduce the radiative heat loss and carrying a ballast of propane to feed a burner in case of major cooling.

Current Scientific Balloon Systems and Their Use

Based on the above concepts, a variety of balloon systems are available to the scientist wanting to study the Earth's, or another planet's atmosphere, the characteristics and performances of which are summarized in Table 1. They are all described below from the point of view of the user, together with examples of their application.

Open Zero-Pressure Balloons

The most commonly used scientific balloon is the zero-pressure balloon (Figure 2). Its volume varies from 50 000 m³ to 2 Mm³ or more, which can carry between 100 and 2000 kg (the record is 3600 kg) between 25 and 40 km (the record is 42 km); see Figure 3. It is equipped with powerful telemetry and remote control transmissions of up to 500 bps. For many purposes, the gondola is stabilized or oriented toward the Sun or an astronomical object, or even stabilizes with respect to a magnetic or geographic heading. In most applications, flight is limited to within the telemetry range from the control station of 300–400 km that is for duration of a few hours, though it can be extended by using a downrange receiving station. The size of the balloon (up to 300 m high at liftoff), as well as the weight of the payload, requires the use of mobile crane or else auxiliary balloons to keep the gondola off the ground during the launch operation, which has to be performed by a well-trained team. The descent of the payload for recovery below one or several parachutes terminates the flight. However, and this is of great interest for atmospheric research, the ascent or descent speed and altitude of the balloon can also be adjusted by alternately

Table 1 Performances of scientific balloons

Balloon	Volume (m ³)	Payload (kg)	Altitude (km)	Duration
Zero pressure	5000–2 000 000	50–2000	25–40	Hours–days
Pressurized sphere	50–600	2–20	12–19	Weeks–months
Pressurized tracers	3–8	2–3	1–5	Weeks
Pressurized pumpkin	1 000 000	1000	35	(Goal 100 days)
Infrared Montgolfier	45 000	60	27 (day) 18 (night)	Weeks–months

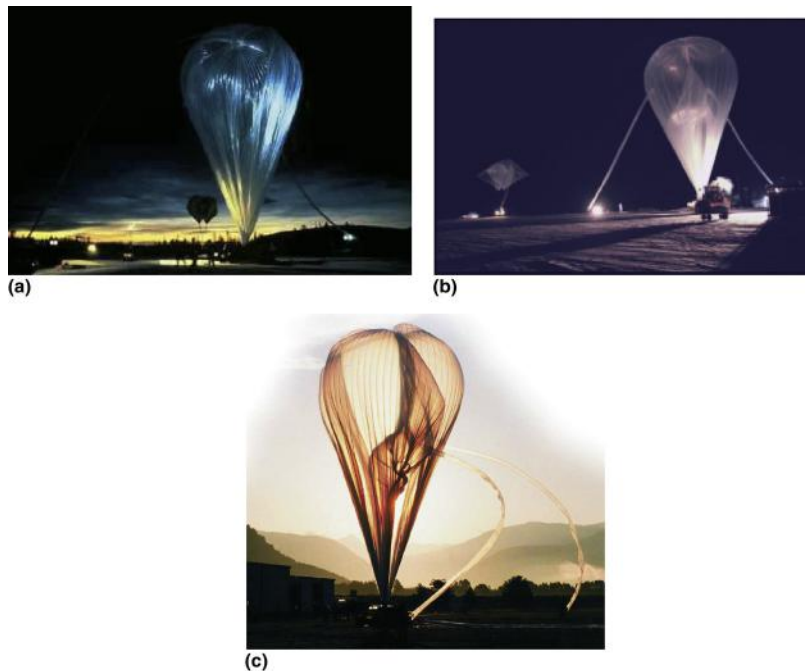


Figure 2 (a) and (b) Inflation of a 100 000 m³ zero-pressure balloon in the polar night in northern Sweden in January 2000 for studying stratospheric clouds and ozone depletion. The two small balloons in the background are auxiliary balloons used to keep the payload off the ground during liftoff. © CNRS. (c) Early morning inflation of a 100 000 m³ zero-pressure balloon at Gap in the Alps. © CNES.



Figure 3 1 000 000 m³ zero-pressure balloon ready for release at McMurdo, Antarctica, in December 2001. The 1100 kg payload is suspended from a mobile crane. Taking advantage of the polar day, a record flight of 1 month was achieved. © NASA.

valving the gas and dropping the ballast. This facility is now in common use for studying a specific layer, e.g., stratospheric clouds in polar areas, or setting a slow descent speed across the whole stratosphere for air sampling or chemical analysis.

In the United States, balloons manufactured by Raven/Aerostar are operated by the NASA Columbia Scientific Balloon Facility (25 flights per year, with payloads of up to 2000 kg) at Palestine, Texas; Fort Sumner, New Mexico; McMurdo, Antarctica; Alice Spring, Australia; and Kiruna, Sweden. In France, they are manufactured by Zodiac Espace and operated by the CNES space agency at the Centre de Lancement de Ballons at Aire sur l'Adour or Gap in the summer in southern France (20 flights per year, with maximum payloads of 500 kg for safety reasons) as well as in Kiruna, Sweden. Several other countries are also conducting balloon programs on their own or in cooperation with the United States and France: Japan, India, China, Canada, Australia, Norway, Sweden, Italy, Brazil, Argentina, and Indonesia. The Soviet Union has had an extensive program in the past, including 4- to 5-day flights from European Russia to Siberia in the winter and from Kamchatka to the Urals in the summer, but this has been suspended for the moment.

Depending on scientific objectives, both NASA and CNES frequently carry out series of balloon launches at remote sites at

all latitudes in Canada, Brazil, Australia, Alaska, Sweden, Antarctica, etc. Among others, recent examples are the series of winter campaigns conducted since 1990 by the CNES for the European Commission at the Swedish ESRANGE facility at Kiruna for studying stratospheric ozone depletion; those included THESEO-SOLVE in the winter of 2000, which involved both CNES and NASA.

The use of zero-pressure balloons in atmospheric science is twofold: (1) for the remote observation from the float level of the vertical profile of chemical species by techniques similar to that in use from orbit; (2) for *in situ* measurement, during ascent or slow descent, of aerosols, stratospheric clouds particles, tracer gases, and chemical species by a variety of techniques close to those in use in the laboratory. Remote sensing techniques include absorption measurements in the UV, visible, and infrared regions by solar, star, or lunar occultation, and emission techniques in the IR and microwave wavelengths. Most of the atmospheric instruments subsequently placed in orbit, such as the ATMOS Fourier Transform Spectrometer on the Space Shuttle, the Microwave Limb Sounder onboard the NASA Upper Air Research Satellite, or the MIPAS Fourier Transform Spectrometer on the ESA ENVISAT satellite, have been flown first on balloons. Besides the new science which could be derived a long time ahead from the long lead time



Figure 4 Small, light, zero-pressure balloon of 10 000 m³ ready for launch at the ESRANGE facility at Kiruna in northern Sweden during the Ozone THESEO-SOLVE campaign in February 2000. Five complementary scientific instruments for measuring chemicals and aerosols by a variety of techniques are shown ready to be flown together in separate packages, which at launch slip along the launch pad. © CNRS.

development of a space project, the advantage is the advanced maturity of the retrieval algorithms as well as later, the validation of the measurements from space by direct comparison with collocated balloon observations.

But the unique contribution of balloons to atmospheric science is the *in situ* measurement of a number of parameters and species, and their altitude and time variations, which cannot be done by other methods, particularly in the stratosphere. Examples include the profiles of the organic chlorine and bromine species (chlorofluorocarbons and halons) responsible for the destruction of ozone; the profile of OH, the hydroxyl radical involved in many chemical cycles; the composition of the crystals of polar stratospheric clouds, which triggers the ozone loss process in polar areas; and high-resolution profiles of long-lived species or tracers needed to understand transport processes. The list of instruments flown includes mass spectrometers, gas chromatographs, samplers, fluorescence-resonance chemical reactors, aerosols, and condensation nuclei counters, tunable diode lasers associated to multiple path-length cells, etc.

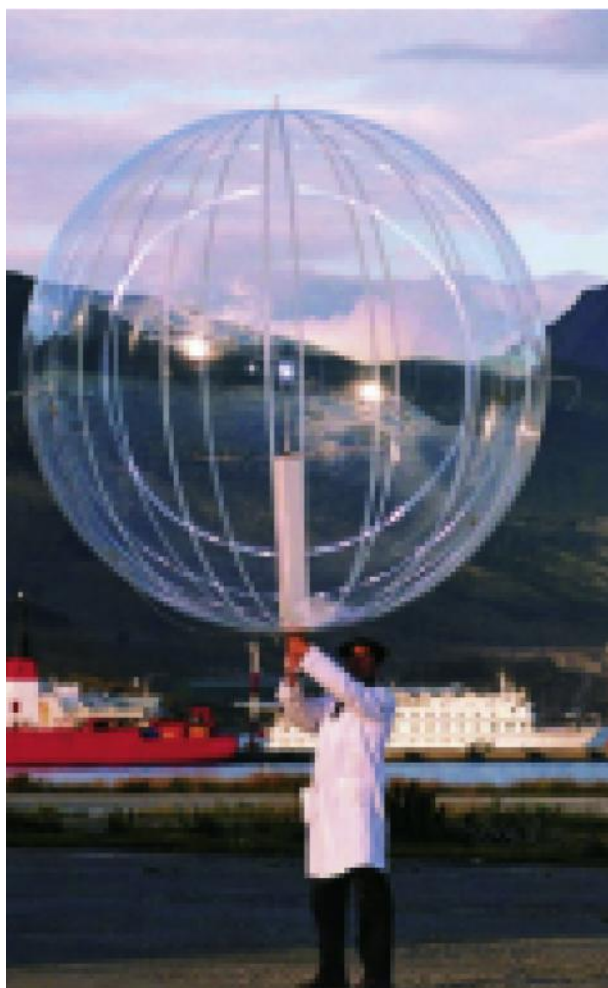


Figure 5 Launch of a tracer superpressure balloon at Ushuaia, Argentina, in February 2000 for studying the meteorology of the lower atmosphere over the Southern Atlantic. The electronics are located inside the balloon for protection against rainfall.

Light, Small, Zero-Pressure Balloons

Large zero-pressure balloons as described above are powerful tools, but since they are extremely sensitive to surface wind, and hence cannot often be launched when required, they are difficult to use in atmospheric science. In addition, their operation needs heavy equipment and large facilities, while finally their relatively high cost does not allow them to be flown as frequently as is desirable for recording atmospheric variability. An alternative approach, explored during recent years by CNES, is the reduction of the size of the balloons (Figure 4). This is achieved by using lighter material, 16 μm and more recently 12 μm thick, but reinforced at the gores assembly, and also by reducing dramatically the weight of the payload by applying miniaturization techniques to scientific instruments as well as to operational subsystems. Finally, launch techniques have also been revised. The payloads are



(a)



(b)

Figure 6 (a) Launch of a 10 m diameter superpressure constant-level balloon at Kiruna, northern Sweden, in January 1999 for studying the dynamics of the stratosphere. © CNES. (b) 10 m diameter super-pressure balloon under testing at the CNES facility at Toulouse. © CNES. CNES, Centre National d' Etudes Spatiales.

designed to be dragged, perhaps on a sledge, on the launch pad, thus not requiring auxiliary balloons or a crane. In such design, a 10 000 m³ balloon carrying 100 kg of apparatus at 32 km could be launched in a surface wind at up to 7 ms⁻¹ (30 km h⁻¹) that is in most meteorological conditions. This flexibility meets a variety of scientific requirements such as those imposed by sunset or sunrise, the presence of polar stratospheric clouds, satellite overpass, tropical storms, etc.

A number of lightweight atmospheric instruments in the 5–20 kg range have been developed in Europe and in the United States for use with these smaller, lighter balloons: UV-visible solar or moon occultation spectrometers, tunable diode lasers, air samplers, aerosol counters, gas chromatographs, and resonance fluorescence reactors, measuring a broad range of chemical species. Such instruments are frequently flown together on the same balloon from unprepared or wind-exposed ranges, in order to study polar ozone chemistry, stratospheric clouds, tropical convection, seasonal and diurnal cycles, long-term trends, satellite validations, etc. However, for safety and programmatic reasons, CNES small balloon and long duration MIR and super-pressure balloon flights are no more feasible preventing the continuation of the above successful projects. It is hoped that they could restart in the future.

Tethered Balloons

Basically, tethered balloons are open balloons that are only a little pressurized, the excess gas being vented through a valve. Though the concept is simple, tethered balloons have been of little use in atmospheric research except in the lowermost boundary layer. Attempts to build permanent atmospheric observatories in the lower stratosphere around 20 km in France in the 1980s, using newly available Kevlar cables, failed totally,

mainly because of gusty winds but also because of air-safety constraints. The use of tethered balloons seems thus to be limited to the lowermost 1–2 km levels in light wind and in areas with little air traffic. However, a promising system for studying the sea–air interface in remote oceanic regions is the drifting balloon attached to a floating guide rope as tested in 2000 by the CNES in the tropical Atlantic.

Superpressure Balloons

As noted earlier, superpressure balloons were developed originally in the United States at NCAR for studying the meteorology of the upper troposphere. They were spherical balloons of 43 µm thick polyester and 2.70 m diameter, able to carry a 2.5 kg payload at 200 hPa (12 km). In the late 1960s and early 1970s, more than 400 balloons were flown by NCAR within a GHOST project from New Zealand in the frame of the international Global Atmospheric Research Program. In addition, another 400 of similar size were flown by the French from South America using a dedicated satellite, EOLE, to track them with a new receiving system at the origin of the ARGOS satellite data collection in broad use in ballooning nowadays. On both projects, the average flight duration was around 3 months, with a record of 1.5 years.

In order to meet the scientific demand, superpressure balloon programs have evolved in two directions: small tracer balloons in the planetary boundary layer and relatively large ones for studying transport in the lower stratosphere. The first type, with the payload mounted inside the envelope to protect the electronics from rainfall (Figure 5) were, for example, flown for several weeks in the 1980s over the Indian Ocean for investigating the atmospheric circulation associated to the monsoon. The experiment was repeated in 2000 in India,



Figure 7 Artist's view of the ULDB pumpkin superpressure balloon at float. Designed for carrying 1 t at 35 km altitude during 100 days, such balloons could allow the remote sensing of the atmosphere on a global scale at far less cost than that of using satellites. ULDB, Ultra Long Duration Balloons. © NASA.

during an INDOEX international project, as well as in Argentina, and more recently in the Indian Ocean, West Africa, and the Mediterranean.

The second type of balloon was the large spherical one flown in the lower stratosphere (Figure 6). But as explained earlier, and shown for example by the unsuccessful attempt by NCAR in the 1980s to develop a 24 m diameter balloon to carry a dropsondes package at 24 km, these balloons are at the limit of technology. The largest system currently existing at CNES is a 12 m diameter balloon carrying a 30 kg payload at

19 km for studying the dynamics of the winter Antarctic stratosphere.

Ultra Long Duration Balloons

Unless a new, high-strength material appears in the future, the size and therefore the altitude of spherical superpressure balloons will be limited. The only way to overcome the limit is to change completely the architecture of the balloon and adopt a pumpkin shape (Figure 7). A large pumpkin balloon

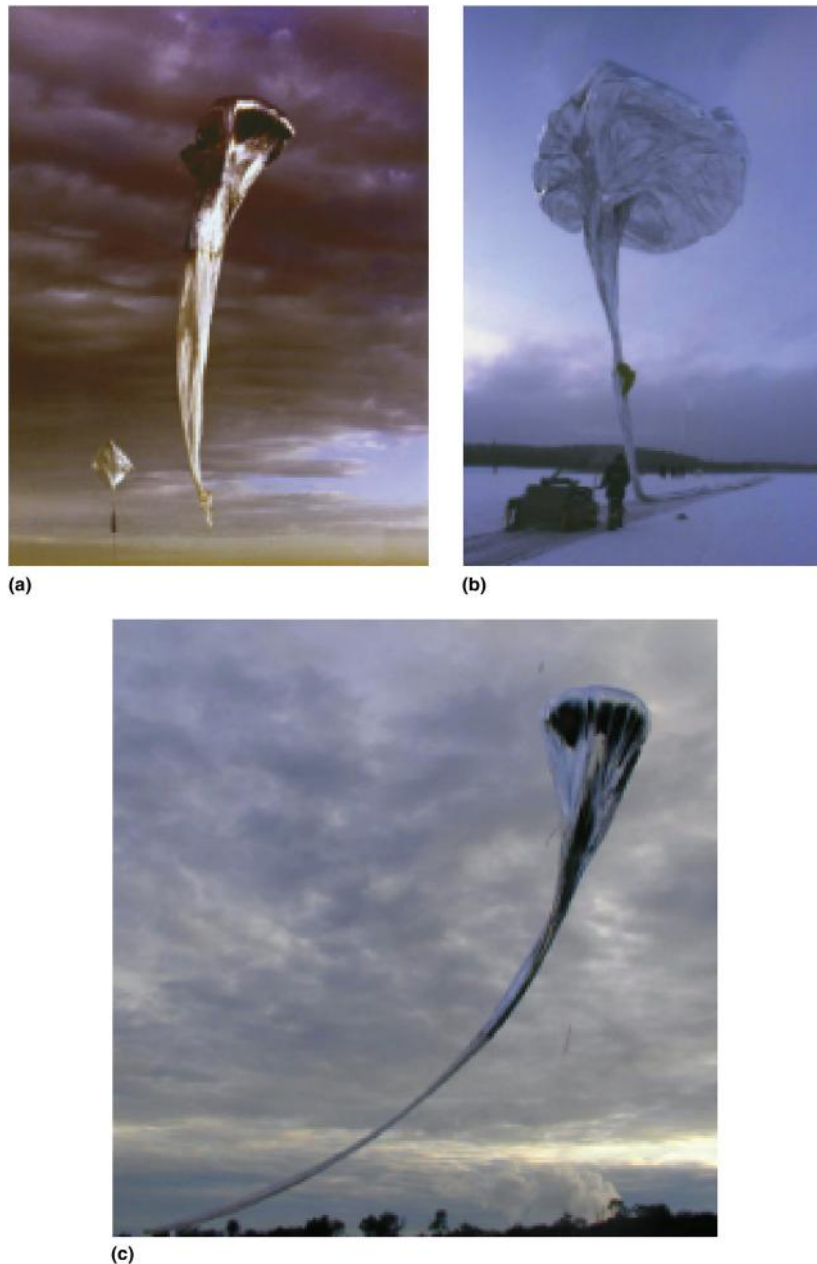


Figure 8 (a) Test flight of MIR 5600 m³ prototype in the early 1980s at Pretoria, South Africa. © CNES. (b) Release of an aluminized 45 000 m³ MIR at Kiruna in February 2000 for studying ozone depletion in the winter Arctic stratosphere. After 18 days the flight was terminated and the payload recovered in Russia. © CNES. (c) Release of an aluminized 45 000 m³ MIR in Brazil in February 2001. The balloon flew for 34 days and two circumnavigations over the tropics before recovery in northern Argentina. © CNES.

of 58 m diameter for carrying 1.6 tons at 33.5 km for 100 days is under development at NASA since 1998. Though a few test flights of smaller balloons have performed successfully in Antarctica and in the Arctic, the vehicle is still not fully operational illustrating the difficulty of the project. However, the program is continuing. There is no doubt that when successful ULDBs, renamed recently as superpressure balloons, will be powerful tools in enabling atmospheric observation to be made by new passive or active remote sensing instruments long before such instruments are able to be placed in orbit.

Infrared Montgolfier

Because of the smaller lift of hot air compared with hydrogen or helium, the MIR ([Figure 8](#)) requires a larger volume than gas balloons for the same payload. But, in turn, it is a zero-pressure balloon and so the stress applied on the material is weak. The MIR is thus a robust balloon whose duration is limited only by the capacity to fly over very low-emitting high-altitude clouds. In its present design of 45 000 m³ volume, the MIR available at CNES in France can carry a payload of 60 kg for several weeks around the world. It has been flown successfully for more than 2 months at the tropics and for up to 3 weeks in the Arctic in the winter and spring, where the limitation comes more from the restriction of flights to north of 55° N for air safety reasons, than from adverse meteorology.

Taking advantage of the diurnal change of altitude from 18 to 20 km during nighttime to 27–28 km during the day, a number of atmospheric observations could be performed, ranging from passive or active remote sensing by solar occultation or lidar, to profiling *in situ* by a variety of techniques. In its current design, the MIR is fully operational. However, further tests of thinner and lighter balloon material are in progress for extending the duration of flight. When successful, and together with new, powerful low-orbit satellite transmissions, it will be a unique tool for combining *in situ* and remote measurements for studying several processes of importance at work in the global stratosphere and climate.

Planetary Ballooning

Though the composition of their atmospheres is different than that of the Earth, several planets in the solar system could also

be explored by balloon systems. Of particular interest are those totally covered by clouds and thus where the surface is invisible from space, such as Venus and Titan. But this could apply also to the study of the meteorology and the surface of Mars, where a balloon could help observe a variety of landscapes with a single station. Several concepts have been studied in the past and sometimes fully developed and tested in France, the United States, and the Soviet Union. Among them, the most advanced space mission projects were a heavy, 10 m diameter Kevlar cloth superpressure balloon for Venus and a thin 6 μ m superpressure balloon for Mars landing at night and dragging a long instrumented guide rope on the surface. Unfortunately and for various technical and programmatic reasons, they have been all canceled but one: a highly successful Teflon balloon designed and flown by the Russian Institute for Cosmic Research (IKI) in the atmosphere of Venus during the Russian–US–French VEGA mission in 1985. Teflon was chosen instead of polyester because of the presence of sulfuric acid clouds in the Venusian atmosphere, resulting in a new concept of sealed, superpressurized, and slowly expandable balloon. Injected on the night side of the planet at 53 km, in the high-altitude 4-day circulation, the balloon performed beautifully for 48 h as expected, allowing the observation of pressure, wind, and temperature change along a night–day transition and for the full following day. Though this success still remains unique, several concepts of zero pressure, super pressure, and Montgolfier autonomous systems, named Aerobots for Robotic Balloons, continue under study in the United States and France, for further planetary exploration missions.

See also: Chemistry of the Atmosphere: Observations for Chemistry (*In Situ*): Water Vapor Sondes. **Observations Platforms:** Buoys; Kites; Radiosondes; Rockets.

Further Reading

Details on balloon national programs can be found at balloon.cnes.fr:8180 for France, www.wff.nasa.gov/~code820 for the US, www.isas.ac.jp/info/balloone3.html for Japan, www.dan.sp-agency.ca/ for Canada, www.rocketrange.no for Norway, www.ssc.se/ for Sweden, and www.das.inpe.br/slb/ for Brazil.

Descriptions of most recent balloon systems and test flight results could be found in the series of proceedings of the biennial symposia on European Rocket and Balloon Programs and Related Research (European Space Agency Publication Division, ESTEC, Noordwijk, The Netherlands).

Buoys

JM Hemsley, National Data Buoy Center, Stennis Space Center, MS, USA

© 2003 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 4, pp 1438–1443, © 2003, Elsevier Ltd.

Introduction

The need to collect real-time meteorological data to be used in forecasting is commonly accepted as important to the protection of life and property. Indeed, one can find observers or packages of automated instruments tasked with collection of meteorological data spread, often densely, across many countries of the world. Unfortunately, the world's oceans are much less densely populated with observing systems. It is difficult to acquire meteorological data at sea. Observers aboard ships are sometimes too busy to take observations when they are most needed. Satellite sensors can be foiled by clouds. The best solution for collecting data at sea is from buoys, either moored or drifting. This article will discuss the types of buoys often used, what data are collected or might be collected, and the advantages and disadvantages of the general buoy types. Because of the author's experience, the article will emphasize the systems used by the National Oceanic and Atmospheric Administration, National Weather Service, National Data Buoy Center (NDBC) of the United States of America.

Purpose and Types

A number of nations collect meteorological data from moored buoys in support of their warning and forecast missions. These include the United States and Canada, which have large networks of moored buoys; the United Kingdom and France, whose networks are somewhat smaller; and nations like Japan, Korea, Australia, New Zealand, and India, which have even smaller networks. In addition, most of these nations, as well as South Africa, Brazil, Germany, Ireland, and The Netherlands, have used drifting buoys in their data collection efforts. Buoys are used in the open ocean, in inland seas, and in major seaways. They are particularly important to forecasters on the eastern boundaries of the oceans and where tropical and extratropical storms are common. Often, data from a buoy are important in the proper forecasting of storms forming over the oceans. Buoy data sometimes provide surface pressure and wind data indicating strengthening 12–18 h before satellite imagery. It is unfortunate that these data are so scarce compared to data from stations on land.

Generally, there are two types of buoys, moored and drifting. A discussion of the types of buoys in each category is appropriate.

Moored Buoys

Hulls

There has been considerable evolution of moored buoy types, which started quite large both to ensure stability at sea and to house the large numbers of batteries, a diesel generator or two, and diesel tanks needed to power the sensors and transmitters.

Break throughs in solar power technology and microelectronics have resulted in the size of buoys being greatly reduced and their capabilities being considerably enhanced.

Stability is an important characteristic of moored buoys. They must provide a stable platform for the acquisition of data, be able to withstand severe sea conditions without capsizing, and protect the electronics needed to control the collection, processing, and transmission of the data. Early buoys were often quite large, first 12 m in diameter, then 10 m. These steel buoys were survivable and had considerable room for the systems required for data collection in the early 1970s. Because they were large, though, they were expensive and difficult to deploy and recover. Eventually, an alternative was found that provided good survivability and adequate space, especially after improvements in solar power, instrumentation, and communications allowed for significant reductions in power requirements. That alternative was a boat-shaped aluminum hull, called a NOMAD when it was developed by the US Navy. It is the buoy hull now often used by NDBC and the Canadian Atmospheric Environment Service at their deep ocean stations. With its mooring attached at the bow, the buoy tends to align itself with the direction of the waves, making it much more survivable in high, steep waves. The highest significant wave height ever measured by an NDBC NOMAD buoy was 16.9 m, measured in the north-east Pacific Ocean. The significant wave height is the average of the highest 1/3 of waves. According to a generally accepted rule of thumb, the highest waves in a storm can be twice the significant wave height, meaning that the buoy probably survived waves as high as 30 m or more.

The same improvements that allowed for the development of NOMAD as a replacement for the huge discus buoys in the deep ocean encouraged the development of smaller buoys for both meteorological and oceanographic use. One of the principal innovators in buoy technology was the Woods Hole Oceanographic Institution in Woods Hole, Massachusetts, USA. Their development of an aluminum discus buoy hull of 3 m diameter made possible the collection of data from networks of moored buoys. The 3 m buoy and its derivatives, while not inexpensive, are considerably more economical than either the large discus buoys or NOMADS. A direct descendent of the Woods Hole 3 m buoy is used extensively in the United States and Canadian networks. A few 12 m and 10 m buoys can still be found operationally deployed, but the majority of buoys now used for the collection of meteorological data are either NOMADS or 3 m discus buoys. Examples of those buoys may be seen on the NDBC web site, (<http://www.ndbc.noaa.gov>).

Not all buoys used for meteorological data collection are exactly like those used in the United States and Canada. One considerable network, deployed in the tropical Pacific and Atlantic oceans by the US National Oceanic and Atmospheric Administration's Pacific Marine Environmental Laboratory, in

conjunction with French and Japanese agencies, uses a toroid buoy. These fiberglass-wrapped foam hulls are less expensive than aluminum hulls, but they are also less capable buoys. In fact, their primary purpose is to collect oceanographic data, with only limited meteorological data collection capabilities. With their moorings containing oceanographic instrumentation, the buoys are replaced each year. This is necessary for good oceanographic data collection, but it would be considered unduly costly for long-term collection of meteorological data.

Other networks use a variety of buoy hull types that they have determined to be best suited to their needs. The buoys used in the United Kingdom, Japan, and Korea are quite different from those of the United States and Canada, but they perform very well for their weather services.

Moorings

A mooring is a system comprising a chain, cable, and/or rope and an anchor of some type that keeps the moored buoy on station. One of the most significant differences between moored buoys intended to collect meteorological data and those intended for oceanographic investigations is the mooring they use. As mentioned previously, oceanographic data collection buoys are replaced frequently, at least in comparison with meteorological buoys. Their oceanographic sensors are susceptible to fouling and, therefore, must be changed or cleaned periodically. Meteorological buoys are often a part of an operational observing network, and frequent replacement of buoys is much too expensive. Both the electronic and the mechanical systems used on the buoys must therefore be long-lived. Moorings are an important component of buoy longevity.

Considerable effort has been put into mooring design, modeling, and field testing in order to develop moorings that will provide a simple, long-term, deep water capability at reasonable cost. Designers of moorings for environmental observations had the advantage that they did not need the tight watch circle (the circle around the buoy's anchor within which the buoy drifts) required of navigation buoys marking a channel. The watch circle for a meteorological buoy was able to be wider because of the consistency of meteorological data over a relatively few miles of ocean. Because of that flexibility, the design of the mooring system was more open to innovation. Two general types of moorings are typically used for meteorological buoys: semi-taut and inverse catenary. Semi-taut moorings are common in shallow water and are most often made either of all chain, in waters up to about 90 m deep, or a combination of chain and, typically, nylon line of 5 cm or more in diameter. Semi-taut moorings using chain and nylon are used in water depths generally from 60 to 600 m. They consist of a length of chain immediately under the buoy, then nylon line, followed by another length of chain attached to the anchor. Semi-taut moorings involving nylon line are probably the most difficult to service, because they are often under tension at sea, essentially becoming taut moorings.

In an effort to design a mooring that allowed easier servicing or replacement of a buoy, NDBC developed its two types of inverse catenary moorings. These moorings have been used successfully in water depths as great as 6000 m and are routinely used in depths greater than 5000 m. The

mooring consists of chain immediately under the buoy; then nylon as in the taut mooring; then at a design depth, often in the range of 500–700 m, a length of polypropylene line or glass balls is inserted; then more nylon line to the bottom where it connects with chain and the anchor. The purpose of the polypropylene rope or glass balls is to provide a buoyant section in the mooring. This serves two purposes. First, it allows for some error in the expected depth of water. Should the anchor settle in water deeper than intended, the mooring will have enough excess length to allow for a successful deployment. More important is the isolation of the lower portions of the mooring from the buoy's motions. Mooring failures associated with wear most often happen where two hard components, such as chain links or chain and shackles come into contact with each other. In the upper mooring, there is considerable movement associated with the buoy's motions. That is not considered a problem, since the upper chain can be inspected when the buoy is serviced and replaced as necessary. Since inverse catenary moorings are almost never retrieved, that is not the case with the chain at the lower end of the mooring, in as much as 6000 m of water. The buoyant section of the mooring successfully isolates the lower mooring so it is not affected by buoy motion, reducing wear on the chain and other hard components at or near the sea bottom. An inverse catenary mooring was once recovered by the NDBC after having been deployed for nearly ten years. It was found that the lower portions of the mooring suffered almost no wear and that the nylon rope had retained nearly its design breaking strength.

Drifting Buoys

Drifting buoys, as the name implies, are those that move with the winds and currents in the ocean. Some, when deployed with drogues, are Lagrangian drifters, allowing ocean current speed and direction to be derived from their positions over time. Many are deployed without drogues, making them less useful in the determination of circulation patterns. Because drifting buoys can easily be deployed by ship or aircraft, they are often used in remote ocean areas for some basic meteorological data collection. Because they are considerably less expensive than moored buoys, they are also used in lieu of or to supplement moored buoy networks.

There are two types of drifting buoys most widely used today. The first, the FGGE (First Global GARP Experiment) buoy was extensively used in the Tropical Ocean and Global Atmosphere (TOGA) Research Program. These buoys were generally cylindrical, 3.7 m in length, and with a floatation collar at the waterline. The FGGE buoys have approximately 1.5 m of their length above the waterline. The second type of drifting buoy is the Surface Velocity Profiler – Barometer (SVP-B). These buoys are spheres on the order of 41 cm in diameter with a barometer port that stands approximately half the buoy's diameter above its top surface. They were designed to be Lagrangian drifters (SVP) and have been modified to provide some meteorological data (SVP-B).

Drifting buoys are packaged for deployment either from ships or from aircraft. In both cases, they are transported in containers that will come apart when soaked in sea water. This allows a sailor to simply drop the drifting buoy over the side of

the ship or an aircraft crewman to deploy it from the plane. For delivery by air, the drifting buoys are, of course, fitted with a parachute to ensure they survive the fall into the ocean.

Measurements

Moored Buoys

Moored buoys are important systems for maintaining weather and sea watch for warning and forecast updates. They provide sea truth for forecast model initialization and verification, warning and forecast verification, and radar and satellite data validation and calibration.

There is both consistency and variability in the measurements made on moored buoys. With few exceptions, wind speed, gust, and direction, atmospheric pressure, and air temperature are collected; this is a consistent feature among nearly all national programs. Wind and pressure data are considered important enough by NDBC that they have redundant sensors to increase the data return. It is also important that the buoys report their location in order to ensure that the buoy is not adrift. Other data often collected include relative humidity; solar radiation; precipitation (amount and/or rate); oceanographic measurements, including wave height, period, energy spectral information, and, sometimes, direction; sea surface temperature; and ocean current (point or profile). In these lie the variations among the data collection networks. Different nations have identified different data priorities.

Drifting Buoys

Drifting buoys provide a more global capability for acquiring and reporting environmental observations. They are useful for scientific and climate programs, weather forecasting, and hurricane tracking.

Drifting buoys have considerably less capability than moored buoys, principally because of their size and limited power. All drifting buoys report their location. This was what they were originally designed to report, so that their drift direction and rate might be calculated. The first drifting buoys to report meteorological data reported only atmospheric pressure and air and sea surface temperature. The size of these FGGE buoys, though, allowed the addition of a Savonius Rotor to measure wind speed. After some design work and prototype testing, a wind fin and compass were added to provide wind direction. In some drifting buoys, accelerometers or other devices have been added to provide wave data.

The SVP-B was a logical result of the SVP. Oceanographers were able to adapt the SVP to add a barometer port above the buoy and a moisture filter to protect the barometer installed inside the buoy. As with the FGGE buoy, the SVP-B also collects sea surface temperature. This relatively inexpensive buoy has allowed the deployment of many more drifting buoys in the remote portions of the oceans. They have been the mainstay of international buoy programs in the South Atlantic and Indian Oceans, as well as in the Arctic and Antarctic. Even with only pressure data reported, there have been several studies performed by the South African Weather Bureau that have shown the value of even just a few pressure data points

from drifting buoys in improving forecasts during the winter storm season. Development of the SVP-B continues. There are efforts to add wind speed measurement capability to the buoy using underwater acoustic sensors, hydrophones that measure the sound of bubbles made by the wind. Wind direction might be measured using a wind fin on the barometer port. The potential for wind measurement from the SVP-B is an exciting prospect.

Data Transmission and Distribution

Data from buoys are useful for warnings and forecasts only if those data are available in real time to the forecasters. It is important that the data be transmitted frequently and distributed immediately.

Moored Buoys

While moored buoys are designed to report via either geostationary or polar orbiting satellites, most moored buoy networks use the geostationary satellites to transmit their data to shore. In North America, the satellites used are the Geostationary Operational Environmental Satellites, one in the 'east' and one in the 'west'. European data are transmitted via Meteosat, and Asian data are transmitted via the Japanese and Indian satellites. Data are typically transmitted hourly through the satellites to a ground station. The data are then quality controlled using automated algorithms and the good data are distributed through the Global Telecommunications System to forecasters around the world. Some networks distribute raw data without attempts at quality control. In those cases, the forecasters must understand that the data have to be compared with other available information before they can be accepted as correct.

Drifting Buoys

Drifting buoys may be deployed or may drift outside the footprint of the geostationary satellites. Together with their limited power, this means that they must often depend on the polar orbiting satellites for data transmission. Data from drifting buoys are typically transmitted to the Polar Orbiting Environmental Satellites as they pass overhead. The data are stored until the satellite passes in 'sight' of one of the receiving stations around the world. After being downloaded, the data are quality checked using automated algorithms and released on the Global Telecommunications System. Raw data can also be acquired on Local User Terminals and used immediately by a forecaster with access to that terminal.

Advantages and Disadvantages of Moored and Drifting Buoys

As might be expected, the advantages of moored buoys are often the disadvantages of drifting buoys, and vice versa. In designing a network for marine meteorological measurement, these advantages and disadvantages must be weighed. In some cases, a moored buoy network is clearly preferable. Sometimes deployment of drifting buoys is more desirable. A mixture of moored and drifting buoys can provide a solution to some needs.

Moored Buoys

Advantages

Moored buoys offer some distinct advantages over drifting buoys. Most importantly, they provide high-quality, real-time data in scheduled data messages when reporting via geostationary satellites. They are a mature technology, so there are few failures because of unproven technology; they have demonstrated their performance over years of use and millions of data messages per year. Moored buoys have, on average, a 3-year lifetime before they have to be replaced. The buoy hulls used by most weather services have virtually unlimited lifetimes, since they can be recovered and refurbished many times. Some of the NOMAD buoys being used by NDBC, for example, were built in the 1950s and are still serviceable. Moored buoys are at fixed locations, allowing the accumulation of a long time-series of data from a constant location, important for understanding climate. Finally, the buoys typically have enough reserve buoyancy and power to allow measurement capability to be added.

Disadvantages

The primary disadvantage of moored buoys is their cost to build and operate. A network of moored buoys is an expensive proposition, although when a per-measurement view is taken, they are often competitive with any other buoy system. The advantage that moored buoys offer by being in a fixed location can also be seen as a disadvantage, since they are designed to be unable to drift with currents. Because of their size, moored buoys require significant deployment resources and their deployments must be well planned. This means that a moored buoy cannot usually be deployed in front of a major storm. Rapid response is not an option.

Drifting Buoys

Advantages

A principal advantage of drifting buoys is their relative low cost, which allows them to be used as an expendable system, although some can be recovered and refurbished. They are also very valuable in remote ocean areas, where servicing a buoy is difficult and expensive. The drifting buoys now in use have a 1- to 2- year lifetime, allowing them to provide data of acceptable quality from a long track across the ocean. Their size and packaging allow the use of many different types of deployment resources. This also means that they can be used for rapid response to developing storms, such as hurricanes, or deployed in arrays of flexible design. Some drifting buoys can also be moored, making them a very flexible buoy system. Finally, they can be drogued to act as Lagrangian drifters.

Disadvantages

Because drifting buoys report through the polar orbiting satellites, their data are not provided on a schedule. This results in some of the data not being available in real time, reducing the usefulness of the data to operational forecasters. The buoys are not in fixed locations, so they do not provide useful climate data for certain applications. There are also trade-offs in data

quality because of the size of the buoys and their intimate contact with the ocean surface. Compared to moored buoys, some drifting buoys have a relatively short life. Finally, because of their size, it is difficult to add sensors and increase their capabilities.

Conclusions

Buoys provide data from a significantly under-measured portion of the Earth's surface. Their data are critical to marine warnings and forecasts, as well as to forecasts over considerable portions of the continents, often many hundreds of miles from the coast. The importance of meteorological data from buoys is recognized worldwide, with a number of nations maintaining networks of moored buoys or deploying drifting buoys in considerable numbers. Buoys are dependable, proven technology.

See also: **Air Sea Interactions:** Freshwater Flux; Momentum, Heat, and Vapor Fluxes; Sea Surface Temperature; Surface Waves. **Data Assimilation and Predictability:** Data Assimilation. **Mesoscale Meteorology:** Severe Storms. **Oceanographic Topics:** Surface/Wind-Driven Circulation; **Tropical Cyclones and Hurricanes:** Hurricanes: Observation. **Tropical Meteorology and Climate:** El Niño and the Southern Oscillation: Observation. **Weather Forecasting:** Severe Weather Forecasting.

Further Reading

- Axys Environmental Consulting Ltd, 1996. *Meteorological and Oceanographic Measurements from Canadian Weather Buoys*. British Columbia, Victoria.
- Bentley, A.N., Imes, D.W., 1993. Development of the UK Meteorological Office's Open Ocean Data Buoy. In: *Proceedings, 2nd Data Collection System User's Conference*, Athens.
- Berteaux, H.O., 1991. *Coastal and Oceanic Buoy Engineering*. Woods Hole Oceanographic Institution, Woods Hole, MA.
- Breaker, L.C., Gilhousen, D.B., Burroughs, L.D., 1998. Preliminary results from long-term measurements of atmospheric moisture in the marine boundary layer in the Gulf of Mexico. *Journal of Atmospheric and Oceanic Technology* 15, 661–676.
- Gilhousen, D.B., 1987. A field evaluation of NDBC moored buoy winds. *Journal of Atmospheric and Oceanic Technology* 4 (1), 94–104.
- Gilhousen, D.B., 1988. Quality control of meteorological data from automated marine stations. In: *Preprint Volume of the Fourth International Conference on Interactive Information and Processing Systems for Meteorological, Oceanographic, and Hydrology*, Anaheim, CA.
- Gilhousen, D.B., 1993. Recent field evaluations of a wind-measuring drifting buoy. In: *Preprint Volume of the Eighth Symposium on Meteorological Observations and Instrumentation*, Anaheim, CA.
- Gilhousen, D.B., 1994. The value of NDBC observations during March 1993's 'Storm of the Century'. *Weather Forecasting* 9 (2), 255–264.
- Gilhousen, D.B., 1998. Improved real-time quality control of NDBC measurements. In: *Preprint Volume of the Tenth Symposium on Meteorological Observations and Instrumentation*, Phoenix, AZ.
- Meindl, E., 1996. *Guide to Moored Buoys and Other Ocean Data Acquisition Systems*. World Meteorological Organization Data Buoy Cooperation Panel Technical Document No. 8. WMO, Geneva.
- Painting, D.J., 1986. Development of an operational moored buoy network. In: *Proceedings of Marine Data Systems '86*. Marine Technology Society.

Introduction

Although kites have been used both for practical purposes and for entertainment for thousands of years, their use for atmospheric research dates back only two hundred years or so. Information derived from kite-borne measurements comprised our first real systematic study of the free atmosphere. The use of kites for atmospheric research peaked sharply around the beginning of the twentieth century, with near-daily observations being made at many observatories throughout the world. Techniques developed during those early times provided a blueprint for subsequent research using data gathered from balloons, airplanes, and, ultimately, satellites. The use of kite-borne meteorological data collection has been renewed during the past decade and currently occupies a welldefined niche in meteorological research. There are a number of advantages provided by having an instrumented kite flying for extended periods a few kilometers above the Earth's surface that cannot be provided by any other means. The full height range of this region is difficult to access more or less continuously by any other technique.

A Brief History of Atmospheric Research Using Kites

Professor Alexander Wilson and his student Thomas Melville, both of the University of Edinburgh, Scotland appear to have initiated the first kite-borne atmospheric measurements in 1749. Wilson and Melville launched a string of paper kites on a single tether, each kite carrying a thermometer attached to its tail. All of the thermometers were enclosed in 'bushy tassels of paper' to insure a soft landing, and were released from their respective tails by smoldering 'fuses'. If gathered quickly after their recovery, the thermometers yielded a rough estimate of the temperature profile up to heights where the highest kite was observed 'disappearing at times among the white summer clouds'.

These early temperature 'profiling' measurements in Scotland were followed in 1752 in Philadelphia by the well-known atmospheric electricity experiment of Ben Franklin. No further kite-borne atmospheric experiments were reported for almost seventy years, until Captain Parry on his second Arctic expedition to Igloolik in the Canadian Arctic in 1822 attached a thermometer to a paper kite, which achieved an altitude of about 130 m. Although he was attempting to measure the lapse rate, he found that the temperature was relatively constant. In a similar measurement made at the same location in 1836–37, however, Admiral Back on HMS *Terror* raised a kite to three times that altitude from the deck of his frozen-in ship, and detected a 5.5 degree (Celsius) temperature decrease.

The first successful measurement of cloud-base heights was made some three years later by meteorologist J. P. Espy, a member of the Franklin Kite Club in Philadelphia. While Espy

only intended to verify his theoretical calculations of cloud-base heights, he also detected strong updrafts just below the clouds as well as evidence of an electrically charged atmosphere. These successes were followed in 1876 in New Jersey by Cleveland Abbe, who flew a kite to determine the depth of the sea breeze.

Early measurements of horizontal wind profiles were begun in the mid-1880s in England by E. Douglas Archibald. Archibald is considered by many to be the first person to make systematic use of kites for atmospheric measurements. During these early measurements, he used a series of tandem-connected kites to profile the wind to over 395 m. Archibald used a self-recording kite-borne anemometer, and flew his kites on a piano-wire tether, following the suggestion of Sir William Thompson (Lord Kelvin).

Following Archibald, most early meteorological kites were flown on piano-wire tethers. Piano-wire provided greater strength per unit weight and diameter than any other tether. While piano-wire tethers enabled systematic measurements to altitudes of many kilometers, the use of a conducting tether proved to be quite dangerous. The danger arose from the fact that the atmosphere is strongly electrically charged, even under clear weather conditions, and the tether provided a convenient discharge path between the kite and the ground. At least one instance has been recorded in the scientific literature of the electrocution of an operator/technician flying a kite during a pending thunderstorm.

Actual measurements of the atmospheric electric field during this same era were made by Alexander McAdie at Blue Hill (Massachusetts, USA) and L. Weber at Breslau (Germany). These studies, which incorporated kite-borne electrometers, built on the pioneering work of Franklin and Espy.

On the basis of these early developments in kiteborne atmospheric research, a series of kite and balloon observatories were constructed during the waning years of the nineteenth century and the early part of the twentieth century. These observatories were dedicated primarily to systematically profiling winds, temperatures, humidity, and pressure to better understand the weather. The famous Blue Hill Observatory near Boston (a privately funded observatory built by A. Lawrence Rotch) played a major part in the development of meteorological kite technology. For example, it was from Blue Hill that W. A. Eddy in 1894 succeeded in launching the first recording thermograph. Eddy's success was followed a year later at the observatory by the launching of a recording meteorograph (temperature, pressure, and humidity). On a somewhat grander scale, the US Agricultural Department's Weather Bureau (predecessor to the US Weather Service) established 17 sites east of the Rocky Mountains beginning around 1898, where kites were flown on a daily basis (weather permitting) until 1933.

Additional outstanding contributions to scientific kite research were made around the turn of the century at a number of observatories in Europe, as well as from other groups around

the world. L. P. Teisserenc de Bort not only founded the Trappes Meteorological Observatory (kite and balloon studies) near Paris, but was exceedingly active in making balloon and kite-borne measurements in Hald (Denmark) and aboard ships in the North Sea. The Lindenberg Observatory near Berlin, which still exists today, was originally established by Professor R. Assman as an observatory for kite-borne measurements. In addition, a number of smaller facilities were established throughout England after the turn of the century. Data from these facilities added appreciably to a knowledge of atmospheric conditions over the British Isles. Other observatories were established in subsequent years in Russia, Egypt, and India, while extended campaigns using kites and balloons were made to such far-flung places as East Africa, the Bahamas, Java, and Antarctica. In support of the global scale of these activities, the International Conference on Aeronautics held in Paris in 1896, established 'international days' – usually the first Thursday of every month – to coordinate kite and balloon flights in all countries.

With regard to kite-borne observations from boats, Teisserenc de Bort (Trappes) and Lawrence Rotch (Blue Hill) cooperated in an observational program aboard Rotch's converted steam launch *Otaria*. These observations extended from the Azores to the Equator in the central Atlantic. Their ocean-going observations followed Rotch's initial trans-Atlantic voyage between Boston and England, during which he flew instrumented kites on five out of a possible eight days. During this period Professor Hergesell (Germany) also flew a kite from a launch owned by Count von Zeppelin on the Bodensee, and made atmospheric measurements from the *Princess Alice*, a yacht owned by the ruler of Monaco, in the Atlantic. The outstanding work of W. H. Dines (an early president of the Royal Meteorological Society) off the west coast of Scotland during 1898–1903 also provided strong impetus for the kites-from-boats technology and for studying the atmosphere over the ocean.

Although it is difficult at this point to determine the absolute level of activity in kite-borne atmospheric research during this early period, it is clear that essentially all of the detailed information on the first few kilometers of the atmosphere around the globe before 1930 was derived from kite-borne instruments. What is certain, however, is that well over one hundred publishing scientists were actively involved in this technology during the three to four decades centered on the turn of the century. During this period, tens of thousands of atmospheric kites were launched around the globe, from the Arctic to the Antarctic and from hundreds of locations on all of the remaining continents as well.

The development of both inexpensive balloonsondes and aircraft in the early 1930s was instrumental in the decline of the use of kites for atmospheric research. After about 1930, systematic atmospheric measurements by kites were viewed as too labor-intensive and expensive. M. W. Harrington, Chief of the US Weather Bureau in 1893, estimated the cost of a single kite-borne profile to ~ 4.5 km to be \$16, roughly \$320 in today's dollars. In addition, beginning in the early 1930s, kites were seen as impediments to the burgeoning aircraft activity.

With regard to the altitudes obtainable by meteorological kites, the Blue Hill Observatory prior to the turn of the century had an impressive multiyear record of lofting meteorological

kites with recording instruments to altitudes well in excess of 3 km. These altitudes were typically reached by connecting a series of separate kites along the same tether. Other observatories listed comparable heights during the same era. These accomplishments should be contrasted with the highest record height of 9740 m reported by the Lindenberg Observatory in 1919 using tandem kites.

The Current Status of Meteorological Research Using Kites

Renewed interest in the use of kites for atmospheric research has become apparent over the past decade or so. This interest has arisen primarily because of the inherent difficulty in making fine-scale, continuous, *in situ* measurements in the first few kilometers of the Earth's atmosphere. Available techniques, which include meteorological towers, tethered balloons, low-flying aircraft, and remote sensing systems (satellites and ground-based radar and lidars), all have significant restrictions. Towers do not extend far enough into the region to provide a number of important answers. Tethered balloons are typically limited to 1–2 km, and then only under light-to-moderate wind conditions. Low-flying aircraft are unable to examine small-scale features because of their speed, and there are obvious safety limitations involved in flying too low. Finally, remote sensing instruments are limited in height resolution and normally measure only specific quantities.

In contrast, present-day kite platforms offer a relatively inexpensive means of obtaining more-or-less continuous measurements of a large variety of critical quantities over a single site. Measured quantities include chemical constituents as well as atmospheric dynamics. Perhaps the best summary of the potential of kite-borne technology for atmospheric science is expressed by stating that it is virtually impossible to obtain continuous, high-resolution measurements of any atmospheric quantity for extended periods (i.e., for many hours) over a single location at altitudes greater than about 2 km by any other means.

Present-day Kite Platforms, Tethers, Winches, and Capstans

The current technology for kite-borne measurements is very similar to that used in the early days, except that present-day systems use state-of-the-art materials and the data-gathering instruments are more complex. Although a variety of kites have been used in current research, the kites used most often are of moderate aspect ratio, ram-air-filled parafoils designed for stability, portability, and reasonable payload capability at high elevation angles. (TALA kites have also been used successfully to measure wind speed at kite level by measuring the tension along the tether.) Parafoils have no rigid supporting structures, are easily transportable, and are typically constructed of nylon or Kevlar-strengthened Mylar with Kevlar or Spectra bridles. Kite sizes range from small kites of a few square meter areas for lofting light payloads, to larger payload systems having areas of 15–20 m². **Figure 1** shows a typical parafoil used in atmospheric studies.

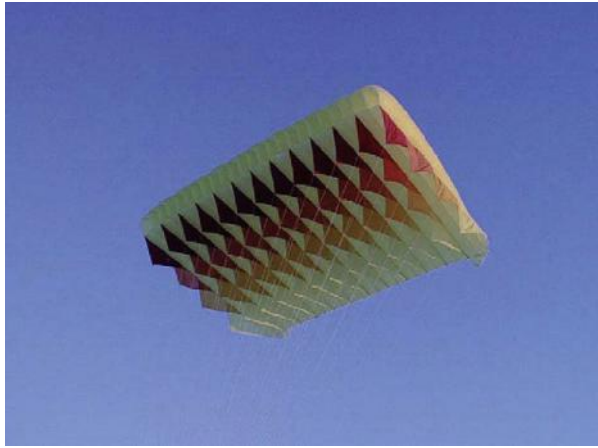


Figure 1 Large parafoil kite used for atmospheric studies. Typically, the payload is attached some distance below the kite on a separate line attached to the tether.

In contrast to the piano-wire kite tethers used in the early measurements, present-day kite tether is normally Kevlar, Spectra, or some equivalent high-strength modern fiber. Typical tether diameters range from 1 to 3 mm. Proper tether selection is very important, since the weight of the tether can be a significant fraction of the payload weight. Also, tether drag increases significantly with the line diameter. Since the tether length is typically twice the height of the kite, both the tether weight and the drag are critical factors in achieving maximum kite heights. Typical tether breaking strengths lie in the range 100–400 kg.

For the kite sizes and line tensions outlined above, it is imperative to make use of some sort of winching system to raise and lower the kites under a wide variety of wind conditions. Typical requirements include the ability to allow the kite to be let out rapidly at low line tension and to bring it in reasonably rapidly under heavy wind conditions, i.e., at high line tension. An additional requirement is to be able to hold the kite on a fixed tether length for extended periods. It is also important to remember the inadvisability of winching the tether directly onto a take-up reel. One problem in using a direct take-up reel as a winch is that each additional wrap on the take-up reel under tension dramatically increases the inward force on the reel. Thus, the thousands of turns of Kevlar put directly onto a take-up reel at high tension will collapse the reel unless it is carefully designed. This problem was effectively demonstrated in 1899 at Blue Hill, when a \$10 000 heavy steel winch purchased for winding in a piano-wire kite tether was crushed beyond recognition when it was first put into operation.

In view of this problem, many meteorological kite systems incorporate a capstan to reduce the line tension prior to spooling the tether onto a take-up reel. A few turns around a capstan greatly reduces the line tension and thereby allows the line to be spooled easily onto a take-up reel. Capstans can be electrically driven or driven by gas or diesel engines. At remote field sites it is possible to use the rear wheel of a vehicle that has been replaced by a capstan, after blocking up the rear axle. In this operation, the vehicle's engine supplies the

necessary power, while the transmission enables the line to be reeled out or in rapidly. In addition, the line length can be fixed indefinitely by either the vehicle's brake or park system. As the tether comes off the capstan it can be wound at low tension around the take-up reel (a reel driven by a small gasoline engine or by a comparable electric motor using the car's battery as a power source). An artist's sketch of a vehicle-driven winch system lofting a parafoil kite (or a balloon) is shown in Figure 2.

Meteorological Payloads and Data Collection

Kite-borne payloads for meteorological measurements are limited by both weight and power requirements. A general rule of thumb is that the payload weight should not exceed 10 kg including its power source. (This is a practical and not a fundamental limitation. Man-lifting kites lifted payloads well in excess of 100 kg to high altitudes at the turn of the century.) In addition, the physical size of a payload must be reasonably small and/or aerodynamically shaped for minimal drag. Furthermore, the payload package must operate reliably within the temperature, pressure, and humidity constraints imposed by the environment. Other than that, a payload can consist of any type of instrumentation that meets the criteria outlined below.

Payload data collection can be accomplished by payload-to-ground telemetry, provided that such a capability is included in

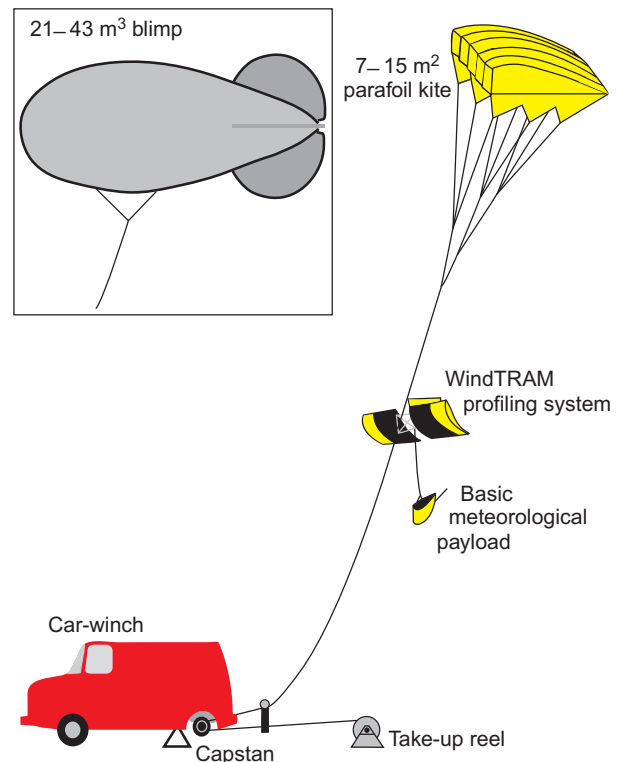


Figure 2 Sketch of a winch system for kite/balloon-borne atmospheric studies that uses the rear wheel of a vehicle for motive power. Reproduced from Balsley, B., Jensen, M.L., Frehlich, R.G., 1998. The use of state-of-the-art kites for profiling the lower atmosphere. *Boundary Layer Meteorology* 87, 1–25.

the instrument payload. Alternatively, data can be stored onboard digitally for subsequent downloading after the payload has landed. The choice between these two possibilities involves consideration of the quantity of data being collected along with the possible need to control the payload position based on the on-line observations.

The standard meteorological measurements that have been made using kite-borne platforms since the turn of the twentieth century include vertical profiles of temperature, pressure, relative humidity, and wind. One example of the inherent vertical resolution of kite-borne wind speed and temperature profiles is shown in [Figure 3](#). In addition to these quantities, measurements that have been made over the past decades include both chemical properties (e.g., water vapor and ozone) and dynamic properties (e.g., atmospheric waves, instabilities, and turbulence properties). Future measurement possibilities are primarily limited by the payload limitations.

Payload Power Sources

The most obvious payload power source is batteries. Since most instruments can be designed to use low voltages (e.g., 5–18 VDC), it is possible to use batteries that are capable of maintaining the required voltage and current capabilities under relatively cold conditions (i.e., -20°C).

Alternative energy sources include both lightweight wind-powered generators carried aloft and solar cells. Wind-powered generators weighing roughly 1 kg can supply a few watts of power on a continuous basis. This technique can be useful, since a reasonably strong wind is all but certain if the kite is in the air. Moreover, since the wind blows night and day, only minimal power storage is necessary. Solar cells, on the other hand, are lighter in weight but require considerable storage for nighttime operations, since they supply power only under daylight conditions. Also, solar panels need to be oriented relative to the Sun's position.

A third possible energy source is the Earth's global electric field. Under clear-air conditions, vertical electric field magnitudes at typical kite heights are of the order of tens of volts per meter. Unfortunately, the source impedance of this 'supply' is quite high and extraction of reasonable amounts of power is problematic. Also, although the global electric field is predictable under clear-weather conditions, the presence of local convective activity produces greatly enhanced, and typically reversed, electric fields. Little work has been done to date to develop the Earth's electric field as a power source for kite-borne studies, although early attempts to extract power from the electric charges carried down a kite wire succeeded in driving a very small electric motor.

The WindTRAM

A recent innovation by a team at the University of Colorado, the WindTRAM (Tethered Rover for Atmospheric Measurements) has added an interesting new dimension to kite-borne atmospheric measurements. This development uses the kite as a 'sky hook', i.e., as a relatively stable, semi-fixed point high above the Earth's surface. With the kite at altitude, measurements can be made using instruments suspended below the WindTRAM. As shown in [Figure 2](#), the TRAM itself is basically a wing-like device that uses wind power to travel up and down the tether. The TRAM wing has an inverted airfoil to provide negative lift when the TRAM is traveling downward (away from the kite). Travel direction is controlled using conventional radio-control (RC) systems. Typical speeds along the tether are $3\text{--}9\text{ m s}^{-1}$, corresponding to vertical velocities of $2\text{--}6\text{ m s}^{-1}$. One advantage of the WindTRAM is that continuous raising and lowering of the kite to obtain profiles is obviated, so that profiles can be made more rapidly. One disadvantage of the technique is that wind speeds need to be greater than 5 m s^{-1} at all heights between the ground and the kite. This requirement is typically not met during nighttime conditions, when winds aloft are reasonable but the ground winds are either weak or nonexistent.

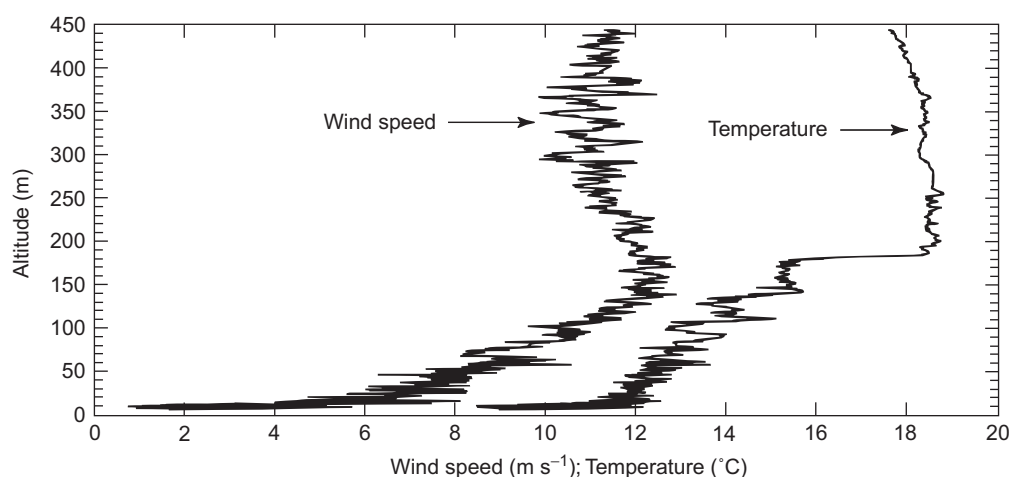


Figure 3 Measurements obtained using kite-borne sensors. This figure shows vertical profiles of wind speed and temperature measured from a sensor suspended below a kite over Kansas, USA at 02:13 LT on 18 October 1999 (ascent).

Advantages and Disadvantages of Data Gathering Using the Kite Technique

One major advantage of using kites for meteorological measurements is that they are relatively simple, lightweight, portable, devices that can operate in winds ranging from roughly 5 up to about 20 m s^{-1} . They are also capable of carrying reasonably sized payloads to altitudes of at least a few kilometers, and can remain aloft for days under ideal conditions. Moreover, the payloads are reusable, in contrast to those carried aloft by untethered balloonsondes that are normally not recovered. Kites can be flown either from land-based locations or from boats, with boat flights being somewhat preferable (except for the more confined working conditions), since the boat's velocity controls the effective wind seen by the kite. This last capability opens possibilities of inexpensive lower-atmospheric data collection at mid-ocean locations.

The primary restriction on the type (or types) of data collection lies with payload weight limitations. Multiple sensors can easily be carried aloft by the same platform. Also, many similar sensors can be attached at preset distances along a secondary line suspended from the main tether. This technique provides a more nearly vertical profile of the measured quantity. During ascent/descent periods, multiple sensors also provide a short-time history of such quantities as the sensors pass sequentially through the same regions. Alternatively, a pulley system on a secondary line can carry sensors from the ground to kite heights for more extended vertical profiles. Data gathered by meteorological kite systems can provide excellent height and time resolution profiles from the ground up to at least a few kilometers on a relatively continuous basis. Alternatively, measurements can, if required, be made for extended periods at a given altitude over a single location. Finally, atmospheric sampling is 'clean', in that the payload can be suspended well below the kite to minimize contamination and to minimize undesired 'blocking' effects that can occur in other types of atmospheric sampling.

There are a number of disadvantages to using kites for atmospheric sampling. One fundamental disadvantage is that kites require winds of at least $5\text{--}7 \text{ m s}^{-1}$ at kite level to remain aloft. Furthermore, kites are not all-weather systems and cannot fly during periods of high convective activity or in strong storm conditions. More importantly, perhaps, the principal difficulty encountered in many campaigns involves obtaining permission from the FAA (or its foreign equivalent for non-US flights) to fly the meteorological kite with payload higher than the normally allowed few hundred meters above the ground. Kites flying somewhat above these heights are of reasonable concern for small aircraft flying in the area. For greater heights, the problem extends to regularly scheduled commercial aircraft. This problem can normally be circumvented by obtaining a NOTAM (NOTice to AirMen) that prevents aircraft from temporarily flying in the area, or by flying in restricted areas in which aircraft are permanently prohibited from flying.

See also: **Aviation Meteorology:** Clear Air Turbulence. **Boundary Layer (Atmospheric) and Air Pollution:** Observational Techniques *In Situ*. **Chemistry of the Atmosphere:** Observations for Chemistry (*In Situ*): Particles. **Clouds and Fog:** Measurement Techniques *In Situ*. **Tropospheric Chemistry and Composition:** Cloud Chemistry.

Further Reading

- Balsley, B., Jensen, M.L., Frehlich, R.G., 1998. The use of state-of-the-art kites for profiling the lower atmosphere. *Boundary Layer Meteorology* 87, 1–25.
 Conover John, H., 1990. *The Blue Hill Meteorological Observatory: The First Hundred Years, 1885–1985*. American Meteorological Society, Boston, MA.
 Hart Clive, 1982. *Kites: An Historical Survey*. Paul P. Appel, Mt. Vernon, NY.
 Mickle, R.E., Cook, N.J., Hoff, A.M., et al., 1988. The Askervein Hill Project: vertical profiles of wind and turbulence. *Boundary Layer Meteorology* 43, 143–169.

Radiosondes

WF Dabberdt, Vaisala Company, Boulder, CO, USA

H Turtiainen, Vaisala Company, Helsinki, Finland

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by W F Dabberdt, R Shellhorn, H Cole, A Paukkunen, J Hörhammer, V Antikainen, volume 5, pp 1900–1913, © 2003, Elsevier Ltd.

Synopsis

The radiosonde is a balloon-borne device that measures the vertical profile of meteorological variables (called a sounding) and transmits the data to a ground-based receiving and processing station. In 2012, there were 92 operational or synoptic radiosonde stations in the United States, which made a daily average of over 200 soundings; and there were over 800 radiosonde stations worldwide, which made a daily average of more than 1300 soundings. There are two primary purposes of upper-air soundings: to analyze and describe current weather patterns, and to provide inputs to short- and medium-range computer-based numerical weather prediction models. One very important, specialized use of atmospheric soundings is in support of forecasting hurricane movement and intensity. Special radiosondes called dropwindsondes are launched from weather reconnaissance aircraft to observe atmospheric structure within the hurricane as well as in the downwind area. Other uses of radiosonde data include scientific research, climate studies, air pollution investigations, aviation operations, and defense applications.

Introduction

The radiosonde is a balloon-borne device that measures the vertical profile of meteorological variables and transmits the data to a ground-based receiving and processing station. These profiles are typically obtained twice everyday and are the core of the ground-based weather observing system that provides inputs to numerical forecast models. The expendable sensor package routinely measures the variation with altitude of temperature, humidity, and pressure as the balloon ascends from the land or ocean surface to heights up to about 30 km (a pressure altitude of about 11 hPa). When the device also measures winds, it is more properly called a rawinsonde, although the term radiosonde is commonly applied to both. The height profile of these meteorological variables constitutes an upper-air sounding that is known as a radiosonde observation (RAOB). In some cases, a balloon without a radiosonde is tracked by either optical or radar techniques in order to measure only winds. This type of balloon is known as a pilot balloon or simply a pibal, but it is not a radiosonde.

In 2012, there were 92 operational or synoptic radiosonde stations in the United States, which made a daily average of over 200 soundings; whereas there were over 800 radiosonde stations worldwide ([Figure 1](#)), which made an average of over 1300 soundings everyday in support of weather forecast activities. Additional soundings are made for specialized purposes of which defense applications are the most significant. The global numbers of RAOB and pibal soundings are down considerably from their peak daily values in 1988 of 1660 and 964, respectively. The approximately half a million radiosondes used annually are manufactured by about 15 companies worldwide. Of these, the Vaisala company, headquartered in Helsinki, Finland, is the largest manufacturer. Vaisala was founded in 1936 by Professor Vilho Vaisala, who in 1931 invented one of the world's first radiosondes (see [Appendix](#)).

Since 1957 all stations make their soundings at the same times, 00.00 and 12.00 UTC, although many stations outside the United States and Europe have reduced soundings to one per day because of budgetary constraints. Countries launching operational radiosondes are members of the World

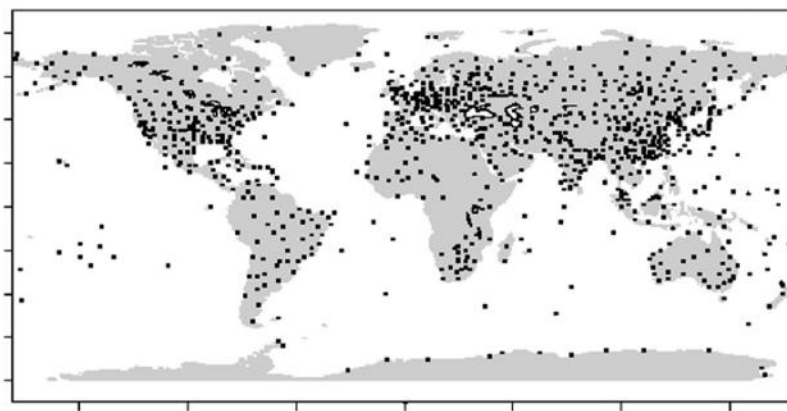


Figure 1 Global radiosonde station network.

Meteorological Organization's (WMO) World Weather Watch program; as such, they freely share their sounding data with each other. Shortly after an operational upper-air sounding is completed, a standard data message is prepared and made available to all nations using the Global Telecommunications System (GTS). Earlier the standard method was the so-called TEMP message (WMO standard format for upper-level pressure, temperature, humidity, and wind report from a fixed land station) reporting meteorological conditions at various standard or so-called mandatory (pressure) levels as well as at significant levels, which represent levels where prescribed changes in meteorological conditions occur. Currently, TEMP messages are being replaced by binary universal form for the representation of meteorological data (BUFR) code, which contains more detailed information of the sounding, as well as associated metadata.

There are two primary purposes of upper-air soundings: to analyze and describe current weather patterns, and to provide inputs to short- and medium-range computer-based weather forecast models. One very important, specialized use of atmospheric soundings is in support of forecasting hurricane movement. Special radiosondes called dropwindsondes are launched from weather reconnaissance aircraft to observe atmospheric structure in the core of the hurricane as well as in the area downwind of the storm itself (i.e., in the direction of the steering winds). These dropwindsonde measurements were the single most important factor in a 16–30% decrease in hurricane track forecast errors over the period 1982–93 (Burpee et al., 1996). And during 1997, the then-new GPS dropsondes reduced hurricane track forecast errors up to 32% and intensity forecast errors up to 20% in the five tropical cyclones that were studied (Aberson and Franklin, 1999). More recently, Weissmann et al. (2011) report typhoon track forecast improvements using the NCEP and WRF models of 20–40% during the 2008 THORPEX Pacific Asian Regional Campaign, but smaller improvements associated with the ECMWF and JMA model track forecasts. Other uses of radiosonde data include climate studies, air pollution investigations, research, aviation operations, and defense applications. The radiosonde continues to be the backbone of an eclectic suite of measurement technologies (measurements both remote and *in situ* that are made from ground-based, airborne, and satellite platforms) used to provide data for input to numerical weather prediction models.

Radiosonde Operations

The radiosonde is carried aloft by a balloon as part of a flight train (Figure 2). The balloon itself may be made of either natural rubber (latex) or synthetic rubber (neoprene). The mass of the flight train, the desired ascent rate, the type of gas used, and the maximum height of the sounding determine the size of the balloon. Operational radiosonde systems typically use balloons that weigh anywhere from 300 to 1200 g; they are filled to ensure an ascent rate of 300 m min^{-1} (5 m s^{-1}). Hydrogen is the most commonly used gas to inflate the balloon, due to its lifting capacity, although helium and natural gas are sometimes used for special applications. The flight train consists of five components: (1) the balloon; (2)

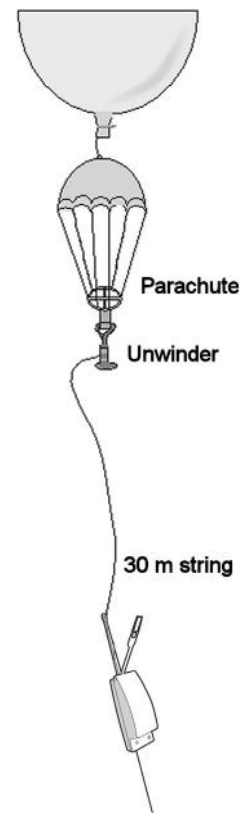


Figure 2 Typical radiosonde flight train, including balloon, parachute, unwinder mechanism, separation line, and radiosonde. © Vaisala company. Reproduced with permission.

a parachute to bring the radiosonde safely back to Earth after the balloon bursts; (3) 20–60 m of nylon separation line that isolates the radiosonde's sensors from potential water vapor and thermal contamination by the balloon; (4) a de-reeler to let out the nylon line after launch; and (5) the radiosonde itself. A few countries such as the United States and Switzerland actively seek to recover and then reuse their radiosondes. In the United States, it is estimated that 10–15% are reused after extensive refurbishment, while in Switzerland, more than 60% are recovered and reused.

In addition to the radiosonde, the sounding system includes ground equipment to receive, process, and forward the meteorological information. The ground equipment receives the data signals from the radiosonde, performs data filtering and quality control, and then produces the output data sets in the form of messages, graphics, or data listings. Typically, the system consists of a sounding processor unit, receiving antenna, radiosonde ground check unit, and a computer (see Figure 3).

Components of the Modern Radiosonde

The radiosonde is an electronics unit that comprises three major sections: a suite of sophisticated meteorological sensors, signal-processing electronics, and a radio transmitter to relay the measurements back to a receiver at the radiosonde launch station. The meteorological measurements are made at intervals

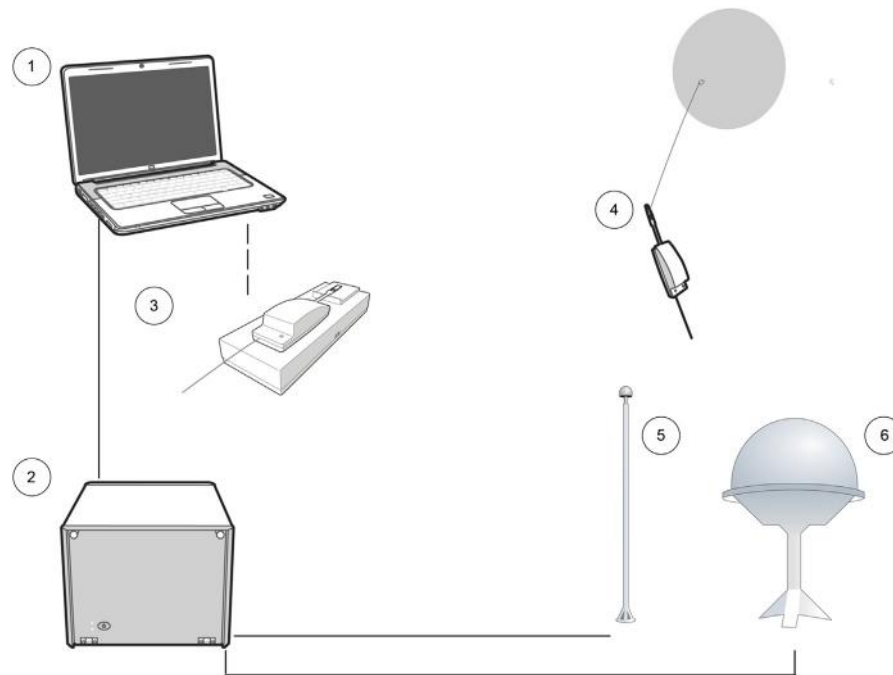


Figure 3 Components of a radiosonde sounding system. 1. Sounding workstation; 2. sounding processing subsystem; 3. ground check set; 4. radiosonde; 5. GPS antenna; 6. UHF antenna. © Vaisala company, Helsinki, Finland. Reproduced with permission.

that vary from 1 to 6 s, depending on the type and manufacturer of the radiosonde. The meteorological community has been assigned two radio frequency bands for use in transmitting meteorological data: 400.15–406 MHz and 1668.4–1700 MHz. These bands are under continuing pressure from the telecommunications industry, which seeks to use them for commercial, nonmeteorological purposes. All of the world's radiosondes are required to meet certain performance standards that have been established by the WMO (see Table 1). Figure 4 illustrates four different radiosondes currently in use around the world.

Overview of Thermodynamic Sensors

Thermodynamic sensor types vary widely among radiosondes currently in use throughout the world. Temperature sensors are of four designs: capacitance sensors, thermistors, metallic resistive sensors, and thermocouples. The two common humidity

elements are carbon hygriators and planar thin-film capacitance sensors. Pressure measurements are typically made with either an aneroid cell or a micromechanical silicon pressure sensor. Also, in some recent designs the pressure sensor is replaced with pressure reading calculated from GPS-derived height, temperature, and humidity, using the hydrostatic equation. There are about a dozen different radiosonde designs presently in use. As radiosondes have become more advanced, their changes have also created special challenges to climatologists seeking to piece together a consistent and homogeneous, multidecadal global database to analyze and understand climate change. As a result, climate researchers must account for biases in the historical records due to changes in instrumentation and observing methods, many of which have poor or no documentation. In the United States alone, these changes have been varied and significant. Four distinctly different humidity sensors have been in use since 1943. Temperature measurements have undergone major changes, including sensor type, size, coating, exposure to the air

Table 1 Accuracy requirements (expressed as standard error) for upper-air measurements for synoptic meteorology

Variable	Range	Accuracy requirement
Pressure	Surface to 5 hPa	± 1 hPa
Temperature	Surface to 100 hPa	± 0.5 K
	100–5 hPa	± 1 K
Relative humidity	Troposphere	$\pm 5\%$ (RH)
Wind direction	Surface to 100 hPa	$\pm 5^\circ$ for wind speed $< 15 \text{ m s}^{-1}$
		$\pm 2.5^\circ$ for wind speed $> 15 \text{ m s}^{-1}$
Wind speed	100–5 hPa	$\pm 5^\circ$
	Surface to 100 hPa	$\pm 1 \text{ m s}^{-1}$
	100–5 hPa	$\pm 2 \text{ m s}^{-1}$
Geopotential height of significant levels	Surface to 100 hPa	$\pm 1\%$ near the surface decreasing to $\pm 0.5\%$ at 100 hPa

World Meteorological Organization, 2008. Guide to Meteorological Instruments and Methods of Observation, seventh ed. Publication No. 8. WMO, Geneva.



Figure 4 Examples of radiosondes in current use around the world. Figures (a–c) from Wikipedia Commons; (d) from © Vaisala company, Helsinki, Finland. Reproduced with permission.

stream, and corrections to account for radiation biases. At present, the US National Weather Service uses radiosondes from two different manufacturers, each having its own distinct set of pressure, temperature, and humidity sensors.

In that the Vaisala company is the largest manufacturer of the world's radiosondes, added emphasis is given below to aspects of the design of its radiosondes and sensors.

Thermodynamic Sensors: Pressure, Temperature, and Humidity

Traditionally, all sensors used with Vaisala radiosondes have been of the capacitance type. This, however, changed in 2013 with the introduction of Vaisala RS41, which has a resistive temperature sensor. In capacitive sensors, changes in pressure, temperature, and humidity result in changes in the capacitance information from each sensor, which in turn is changed to a frequency signal by using sensor transducer electronics. Sensor frequency measurements are compared with the frequencies of reference capacitance transducers, and these in turn are converted to physical measurements based on factory calibration measurements.

In the case of pressure, the distance between capacitance plates changes as atmospheric pressure changes, causing a change in the measured capacitance. Older pressure sensors used an aneroid or bellows-type sensor that responds

mechanically to pressure changes. Modern pressure transducers are very small silicon, micromechanical pressure sensors. Pressure sensors also have a temperature dependence that is compensated by factory calibration of the sensor. Modern GPS technology has also made possible a design, where the pressure sensor is replaced with GPS-derived pressure reading, calculated from GPS height, temperature, and humidity using the hydrostatic equation. Vaisala RS41-SG is the first radiosonde model utilizing this technology.

In capacitive temperature sensors, temperature is measured by the change in the dielectric constant of the sensor. In the latest Vaisala model, RS41, the capacitive temperature sensor is replaced with a resistive platinum sensor, specially designed for stability and robustness. Essential aspects of modern temperature and humidity sensors (and their supporting members such as the sensor boom; see Figure 5) are their different coatings and treatments to minimize solar heating and improve water repellency.

The approach used to measure humidity in all Vaisala radiosondes is also based on changes in the dielectric constant. The humidity sensor is manufactured using thin-film technology. The dielectric material is a very thin layer of a special proprietary polymer that has an optimum combination of measurement properties, including stability, repeatability, hysteresis, response time, and temperature dependence. Thin-film humidity sensors are calibrated to provide output in terms of percent relative humidity with respect to water; the



Figure 5 Vaisala RS41 radiosonde with sensor boom. © Vaisala company, Helsinki, Finland. Reproduced with permission.

temperature dependence is compensated by use of temperature-dependent calibration coefficients determined from factory calibration tests.

The accuracy of radiosonde data is a combination of multiple factors: sensor performance, related transducer electronics, mechanical construction of the sonde and sensor housing, sensor and sensor-boom coatings and treatments, calibration technology, and calibration and correction algorithms. In addition to issues of radiosonde performance, the uncertainty of upper-air measurements includes sampling considerations, such as the density of the observation network, time interval between observations, and the homogeneity of the atmosphere. Together, these instrumental and environmental factors govern the accuracy and representivity of the observations.

Specialized Radiosonde Sensors

Some radiosonde manufacturers offer optional sensors to make supplemental environmental measurements. Additional electronics are used to interface the supplemental sensors to the radiosonde. Ozone concentration is the most common supplemental measurement. Radiosonde measurements of ozone are made worldwide, although at fewer stations and typically only once in a day or less often. The most common radiosonde ozone sensor is the electrochemical type. Other supplemental measurements in use today include dew point, optical backscattering by fine particles, electric field, and video imaging of particles and hydrometeors. Most advanced radiosonde ground systems effectively support both synoptic and research users, and offer options for post-ascent data calculation and analysis of supplemental measurements.

The Vaisala ozonesonde consists of an electrochemical ozone sensor connected to an interface unit and a radiosonde. Consequently, humidity, pressure, temperature, and geopotential height can be measured simultaneously with ozone sampling. Upper-air winds are also measured. This lightweight, balloon-borne instrument is capable of measuring the vertical distribution of atmospheric ozone up to a pressure altitude of 3 hPa. The uncertainty of the ozone measurement is of order 5–10% of the local value. The electrochemical concentration cell (ECC) detects ozone on the basis of an iodine–iodide oxidation–reduction or redox electrode reaction in a neutral-buffered solution. The sensor consists of an ECC that contains two platinum electrodes immersed in separate potassium iodide solutions of different concentrations that are separate anode and cathode chambers. The chambers are linked with an ion bridge. As the air containing ozone flows into the cathode solution, a chemical reaction occurs and the platinum electrodes carry electrons between the cells of the sensor. An electrical current is generated in proportion to the rate at which ozone enters the cell. The ozone concentration is determined from the electric current measurement using an equation that considers the airflow rate, air pressure, and pump temperature. The interface can also be used with other sensor types, such as the Brewer–Mast sensor. However, the Brewer–Mast sensor is nowadays almost totally replaced with the widely used ECC-type sensor, which is also more accurate.

Overview of Windfinding

There are several techniques for measuring winds only with a balloon or with a combination of balloon and radiosonde. When a radiosonde measures winds it is called a radio wind sonde or rawinsonde, and the windfinding methods vary widely. In all cases, the winds are determined by measuring the drift of the balloon. One class of wind measurement techniques tracks the balloon externally using one of the three methods: (1) optical systems use a theodolite to visually track the balloon's azimuth and elevation; (2) radio theodolites track a radio signal sent from a transmitter on the radiosonde, again to obtain azimuth and elevation information; and (3) radar systems track a radar retroreflector suspended from the balloon to obtain slant range, azimuth, and elevation. The second class of wind measurement techniques uses various navigation systems. Earlier the LORAN-C navigation system and various very low frequency (VLF) systems, such as the Russian ALPHA system and the US Navy's VLF system were commonly used. Currently, Global Positioning System (GPS) technology has replaced these older navigation systems; a GPS receiver directly measures the drift velocity of the balloon and hence the wind. Two major advantages of the GPS-based techniques are the high accuracy and precision of the wind measurements, and the worldwide coverage of GPS.

Tracking Techniques

Optical Tracking Methods

One of the earliest methods for determining the winds aloft was to visually or optically track small balloons, called pibals.

This method was developed in the mid-1870s using a small expendable balloon tracked with a small telescope. The small optical device, similar to a surveyor's transit, is called a theodolite and can accurately measure elevation and azimuth angles. If the balloon's height can be determined then its position can be found by trigonometry. There are basically two pilot balloon techniques still in use: (1) single-theodolite and (2) double-theodolite techniques. In the former, the elevation and azimuth angles of the balloon are measured at regular intervals (typically once in a minute). Balloon altitude is determined by assuming a constant ascent rate that is determined from the size and free lift of the balloon. Balloon position is then calculated from the height and the azimuth and elevation measurements. Tracking the balloon during a nighttime observation is accomplished by attaching a light stick or a small battery-powered light. In the double-theodolite technique, two theodolites are located at a known distance apart (the baseline) and simultaneous observations taken of the balloon at given time intervals. By measuring the azimuth and elevation angles of the balloon from the two known positions, the three-dimensional balloon position can be determined by the law of sines. The double-theodolite method enables accurate measurements of the balloon position without assuming a constant rate of ascent for the balloon, which can be a source of error. In this method, the baseline distance needs to be accurately measured and should be at least one-fifth of the maximum range to the balloon. The baseline should also be perpendicular to the prevailing winds. The method is not routinely used because of baseline restrictions and the cost and difficulty of coordinating two sets of observers.

Radio Theodolite, Radar, and GPS Methods

Another tracking technique used for determining winds is called radio direction finding (RDF). During World War II, the US Army Signal Corps developed the first RDF system, called the SCR-658. This system operated at 400 MHz and used two separate operators to steer a large antenna array to determine the direction of the radiosonde transmitter. A more modern RDF antenna automatically tracks the 1680 MHz telemetry signal transmitted from the radiosonde. The antenna azimuth and elevation data are sent to a computer at the ground station along with the pressure height data from the radiosonde to determine the change in radiosonde position (winds) during flight. The RDF technique (Figure 6) is the radio frequency equivalent of the optical theodolite method, and the tracking system is called a radio theodolite. There are different types of RDF antennas, including 2–3 m diameter dish antennas and phased array flat-plate antennas. RDF systems can resolve the azimuth and elevation angles to within 0.05° . If the upper level winds are high then the radiosonde will be a long distance away, resulting in the antenna elevation angle being near the horizon. At stations that experience high winds, the radiosondes can be equipped with a transponder to measure slant range or distance to the radiosonde. Winds can then be determined using azimuth, elevation, slant range, and height of the radiosonde. A similar method for tracking the radiosonde uses a radar reflector on the balloon flight train so that it can be tracked by

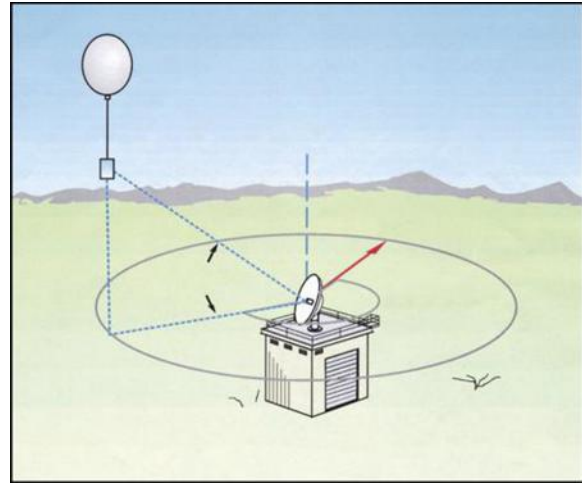


Figure 6 Angle-dependent tracking system.

windfinding radar. Slant range to the radiosonde is measured by the radar as well as azimuth and elevation angles. Radar windfinding is a common method used in many countries around the world; in 1998 about 45% of the stations used radar, as tallied in Oakley (1998).

In some countries, a combination RDF-transponder method is used, called secondary radar. The parabolic or array-type RDF antenna transmits a short pulse that is received by the radiosonde. The radiosonde then 'wakes up' and retransmits the pulse, encoding the temperature and humidity data, which are received by the ground-based RDF antenna. The RDF antenna azimuth and elevation angles are measured and the slant range is determined from the travel time of the pulse. Secondary radar systems use radiosondes that do not have a pressure sensor as pressure is calculated from the hydrostatic equation.

GPS Windfinding

LORAN or Omega NAVAID usage for radiosonde windfinding is no longer being used or supported and has been replaced by GPS. GPS was conceived in the early 1970s for the US Department of Defense, and is operated by the US Air Force. The GPS system became fully operational in late 1995. There are 24 satellites in six-orbital planes spaced 60° apart. The satellites are in a 200 km circular orbit, with an inclination angle of 55° and a periodicity of 12 h. At any time or place in the world, there are 6–11 GPS satellites 5° or more above the horizon and hence usable for GPS windfinding. There are two primary GPS techniques for determining winds from radiosondes. The GPS signals cannot be retransmitted from the radiosonde back to the ground because the bandwidth of the 1575 MHz (called the L1 band) GPS carrier signal is too wide (B2.0 MHz). The worldwide civilian use of GPS has become so great that many manufacturers produce inexpensive small GPS receivers, each the size of a credit card that can decode the navigation message every second and produce accurate



Figure 7 USAF C-130 Hurricane Hunter launching a GPS dropsonde.

three-dimensional position coordinates as well as speed and heading. A second, less expensive method uses a codeless receiver in the radiosonde that measures only the Doppler shift of the carrier frequency. The Doppler shift has two components: (1) the Doppler shift due to the satellite motion (i.e., the largest component) and (2) the Doppler shift due to radiosonde movement. The radiosonde receiver sends the Doppler information back to the ground data system. The ground data system must have a local GPS receiver that can decode the GPS message and independently measure the Doppler shift from each satellite. The satellite Doppler shift is subtracted from the radiosonde Doppler shift and the difference yields the radiosonde motion.

Specialized Types of Radiosonde Systems

Dropsonde

The dropsonde is the airborne counterpart to the conventional radiosonde (sometimes also called an upsonde). Dropsondes are ejected from research and reconnaissance aircraft and float to earth on a special balloonlike parachute. Current state-of-the-art dropsonde sensors include capacitance fine-wire sensors to measure temperature, capacitance silicon pressure sensors, and GPS receivers to measure winds. Humidity is measured with a pair of thin-film capacitance sensors that are heated alternately to avoid condensation on descent from colder to warmer air. All measurements are made twice every second, while the 400 g dropsonde falls at an initial rate of about 25 m s^{-1} at 15 km altitude, decreasing to about 10 m s^{-1} at sea level. Dropsonde data are transmitted by radio from the sonde to a data system in the aircraft. Atmospheric soundings from dropsondes provide the ability to measure conditions over remote areas such as the oceans, polar regions, and sparsely inhabited landmasses; they also provide a means to obtain soundings in and around severe weather systems, such as

hurricanes. Atmospheric soundings obtained from dropsondes during hurricane reconnaissance flights have improved the accuracy of forecasts of hurricane landfall by about 20% over the decade of the 1990s. Dropsondes were first developed in the 1960s for the US Navy and Air



Figure 8 GPS dropsonde descending on its parachute.

Force for hurricane reconnaissance and were an adaptation of radiosonde technology. These early dropsondes were heavy – about 2.5 kg – and did not have inherent wind-finding capability; windfinding at that time still used only radar or RDF. With the development of Omega NAVAJD windfinding technique for the radiosonde, it became possible to incorporate that technology into the dropsonde. This occurred in 1974 when the National Center for Atmospheric Research (NCAR) developed an Omega Dropwindsonde (ODW) for use in the Global Atmospheric Research Program's Atlantic Tropical Experiment. In 1982, the Air Force adopted the ODW system for hurricane reconnaissance and this system was used until the early 1990s. In 1985, NCAR began development of a smart (i.e., microprocessor-based), lightweight digital dropsonde that

incorporated LORAN and Omega windfinding. The Omega version of this dropwindsonde was adopted by the US Air Force in the early 1990s for its hurricane reconnaissance mission (Figure 7). The next major improvement in dropsonde technology occurred in 1995 when NCAR completed development and testing of a new GPS dropsonde with codeless GPS windfinding capability and an advanced aircraft data system (AVAPS). In 1996, NCAR licensed Vaisala Company of Woburn, Massachusetts, to commercialize production and sales of the GPS dropsonde (Figure 8) and AVAPS. In the relatively short time, the GPS dropsonde has been in use it has found research applications in the determination of hurricane structure and motion, the study of clear-air turbulence associated with upper level jet stream structure, and observing strategies for midlatitude weather



Figure 9 Remote-controlled GPS dropsonde launcher system installed on the NASA ER-2 high-altitude weather research aircraft.

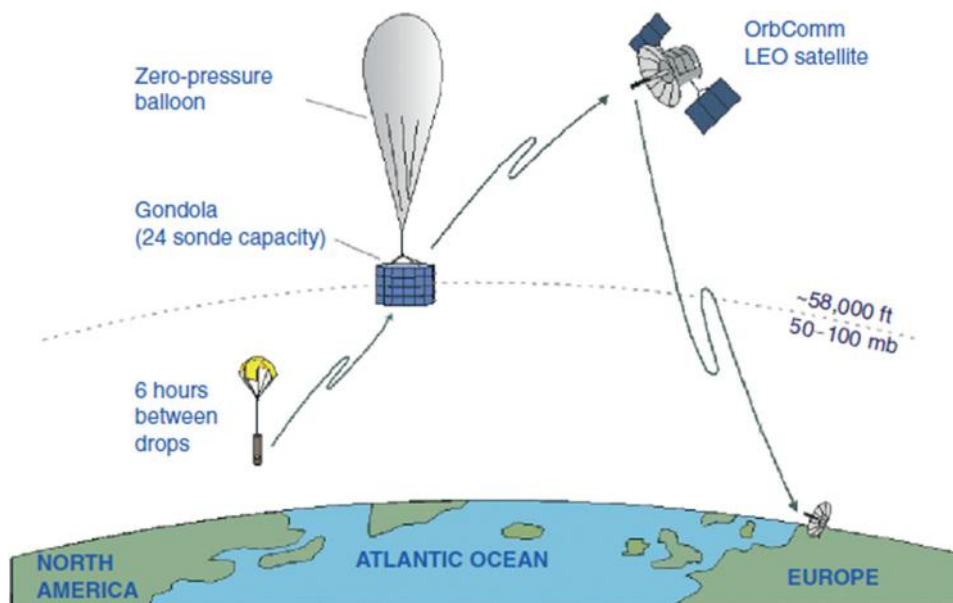


Figure 10 The driftsonde system concept.

forecasting. Current adaptations of the GPS dropsonde technology are focusing on launches at higher altitudes – including the lower stratosphere – as well as autonomous launches that eliminate the need for operators to launch the sonde and record the data, and offer the promise that it will be possible one day to obtain operational dropwindsonde profiles from commercial aircraft. [Figure 9](#) shows a test dropsonde launcher mounted on the underside of an ER-2 high-altitude weather research aircraft.

Driftsonde System

Improvements in short- and medium-range synoptic-scale weather forecasts will depend on improved upper-air soundings over the data-sparse regions of the Northern and the Southern Hemispheres. Progress toward this objective will require the optimal use of existing data sources, creative new observing methods, and improved numerical methods for data assimilation. The driftsonde system is being developed as a cost-effective sounding system that could fill these critical gaps in data coverage over oceanic and remote arctic and continental regions. The driftsonde concept seeks to obtain a large number of high-vertical-resolution GPS dropsonde profiles through the lower stratosphere and the entire troposphere by autonomous launching of dropsondes from specially designed balloon platforms. The driftsonde system includes a polyethylene carrier balloon with an attached gondola ([Figure 10](#)) that carries a payload of up to 24 GPS dropsondes. The carrier balloon ascends to between 50 and 100 hPa (20 and 16 km), and then drifts in the prevailing stratospheric winds for up to 5 days, deploying dropsondes at prescribed and special times over data-sparse regions of interest. The first application of the driftsonde system was in support of The Hemispheric Observing System Research and Predictability Experiment (THORPEX). The driftsonde system has been deployed in three field experiments associated with THORPEX (Cohn et al., 2013): AMMA (African Monsoon Multidisciplinary Analyses, 2006), T-PARC (THORPEX Pacific Asian Regional Campaign, 2008), and Concordiasi field experiment (lower stratosphere and troposphere study above Antarctica, 2010).

Unmanned Sounding Systems

Autosonde

A special type of ground equipment is the unmanned, automated sounding system (Autosonde); see [Figure 11](#). It was first introduced by the Vaisala company in 1992. These automatic sounding systems have become an important tool in the drive for meteorological data availability. Using them, weather services can extend the coverage of their upper-air networks to geographically remote, hard-to-reach locations and decrease sounding costs. A typical Autosonde system has the capacity to perform automatically for 24 consecutive soundings, meaning that in synoptic use (2 soundings per day) it needs to be visited and restocked only with interval of 12 days. There are currently about 60 unmanned Vaisala sounding systems in operation globally.

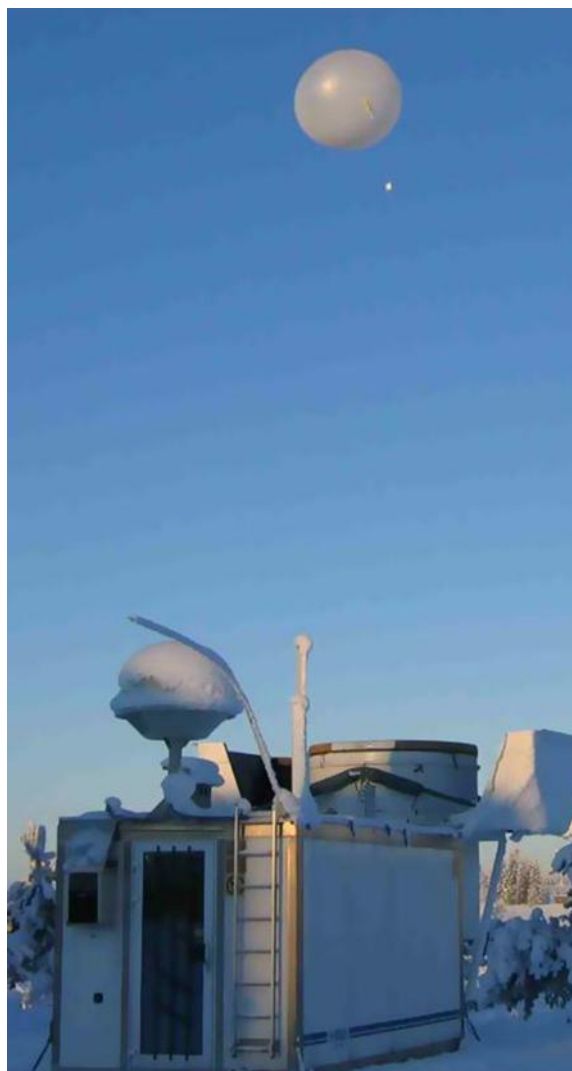


Figure 11 Equipped with special cold climate kit, the Autosonde systems can be used in temperatures below -40°C . © Vaisala company, Helsinki, Finland. Reproduced with permission.

Automated Shipboard Aerological Program

Gathering synoptic weather observations over the oceans is an important complement to land-based meteorological upper-air observations. The Automated Shipboard Aerological Program (ASAP) is a multinational effort initiated by Canada in 1982 to obtain upper-air soundings over the oceans. Radiosondes are launched from commercial ships of opportunity using a specially designed launch system ([Figure 12](#)) that permits flight trains to be launched in high-wind conditions. The upper-air sounding data from the radiosonde are sent back to the shipboard ASAP system where the data are processed in near real time to create a TEMP SHIP message (the WMO standard format for upper-level pressure, temperature, humidity, and wind report from a sea station). This message is the ocean equivalent of the TEMP message generated for land-based RAOB systems. The ASAP system sends the message to a geostationary satellite that relays the information to the GTS, which then transmits it to the numerical weather prediction



Figure 12 The ASAP system is housed in either a standard 3 or 6 m (shown above) sea container with a specially designed hatch to enable routine radiosonde launches in sustained wind up to 25 m s^{-1} (gusts up to 35 m s^{-1}).

centers around the world. The ASAP program had its beginning in June 1981, when Canada decided to discontinue its weather ship program owing to the high costs of operating and maintaining the ocean weather ship PAPA located in the Gulf of Alaska at 50° N , 145° W . The original intent was to replace the weather ship data with satellite observations; however, persistent cloudiness in areas such as the Gulf of Alaska and the North Atlantic coupled with the lack of surface weather data, made this goal impossible to attain. To remedy this problem, the Atmospheric Environment Service (AES) of Environment Canada, the National Weather Service of National Oceanic and Atmospheric Administration (NOAA), and National Center for Atmospheric Research (NCAR) established a joint ASAP project to develop a modular, mobile, moderately priced, upper-air sounding system. This system, when placed on commercial vessels (ships-of-opportunity) routinely crossing the Pacific Ocean and the Atlantic Ocean, provides real-time upper-air soundings that complement those of the global land-based upper-air network. The ASAP program operated by AES Canada started in the spring of 1982 with one commercial ship (a Japanese automobile carrier) that operated from Vancouver, British Columbia, to Japan. By 2012, it had evolved into an international program coordinated by EUMETNET, operating 19 ASAP units and delivering 4700 radio soundings per year.

Reference Radiosondes

In 2007, the WMO's Global Climate Observing System (GCOS) laid out the need for a GCOS Global Reference Upper-Air Network (GRUAN). They concluded (Seidel et al.,

2009) that shortcomings in the current upper-air measurement network do not satisfy the accuracy and detail of observations needed to specify climate variability and changes above the Earth's surface. This deficit greatly impacts the ability to accurately assess and predict climate change. The overall goal of GRUAN is to establish about 30 stations that will use reference grade radiosondes in addition to other instrumentation to represent climate around the world. Currently (2013), the network consists of 15 initial GRUAN stations around the globe. The network construction and operations are coordinated by the GRUAN Lead Centre at the Meteorological Observatory Lindenberg (Germany), run by the German National Meteorological Service, the Deutscher Wetterdienst.

Climate variables to be observed with the highest priority are temperature, water vapor, and pressure. The vertical range for these observations is between the surface and the middle to upper stratosphere. Currently, the largest challenge is the observation of water vapor in the upper troposphere and lower stratosphere, and considerable development in sensing technologies will be needed to expand the current capabilities to observe this climate variable using *in situ* sounding instrumentation. One of the ongoing efforts in this direction is the Reference Radiosonde Program launched by Vaisala company in 2009. To fulfill the accuracy requirements, the Vaisala Reference Radiosonde RR01 will use an extremely sensitive, new type of capacitive dew point sensor to complement the standard radiosonde sensors. Currently (2013), Vaisala RR01 has entered the beta testing phase.

Appendix

Historical milestones leading to the development of the modern meteorological radiosonde

<i>Year</i>	<i>Milestone</i>
1643	Evangelista Torricelli invents the barometer in Florence, Italy.
1648	French mathematician Blaise Pascal observes the decrease of atmospheric pressure with altitude.
1749	Alexander Wilson, Glasgow, Scotland, uses kites to study the variation of temperature with altitude.
1783	The French Montgolfier brothers Joseph-Michel and Jacques-Étienne invent the hot-air balloon.
1783	Jacques Alexandre Césaire Charles, Paris, France, uses a manned balloon to make the first measurements of variations of pressure and temperature with altitude.
1784	Englishman John Jeffries, London, and Frenchman Jean-Pierre Blanchard begin the systematic study of the atmosphere using manned balloons.
1804	French physicists Louis Gay-Lussac and Jean Baptiste Biot ascend to 7 km in a balloon and discover that water vapor decreases with altitude.
1822	Englishmen Sir Edward Parry and the Rev. George Fisher use kites with recording thermometers to study the Arctic atmosphere.
1847	William Radcliff Birt is the first to measure winds aloft (and temperature) with a kite flown from Kew Observatory, London.
1892	Frenchmen H. Hermite and G. Besançon launch the first free-flying weather balloon with mechanical recording system (the 'meteorograph').
1893	Lawrence Hargrave, Sydney, Australia, invents the box kite; by end of the decade, many major observatories are using box kites routinely to measure the atmosphere; they include Blue Hill (near Boston, Massachusetts), the Central Physical Observatory (Moscow), Trappes (near Paris), Kew (London), Lindenberg (Germany), and Ilmala (Helsinki).
1900	British scientist W.H. Dines invents the mechanical meteorograph design that is widely used until 1939.
1901	Richard Assmann, Germany, is first to use 'extensible' rubber balloons for free-flying soundings with meteorographs.
1917	Germans F. Herath and M. Robizsch use the 'telemeteorograph' to transmit meteorological data from a kite using the steel kite cable as the signal cable.
1920	US Weather Bureau and Army Air Corps establish a program of daily upper-air soundings using airplanes at 20 locations nationwide.
1921	US Weather Bureau establishes a kite network for routine upper-air observations; this remains in operation until 1933.

<i>Year</i>	<i>Milestone</i>
1927	M.R. Bureau and M. Idrac (France) invent the 'shortwave' (RF) tube-type transmitter and publish a paper describing the flight of their first balloon-borne sonde (although it is unclear whether any meteorological variables were actually measured). Their paper is the first documented use of the term 'radiosonde,' which they attribute to H. Hergesell (president of the international aerological commission).
1929	January 17: M. Idrac and M.R. Bureau test the first free-flying radiosonde, called the 'Thermoradio,' with a bimetallic temperature sensing element to transmit temperature data to a ground station.
1930	January 30: P.A. Moltchanov (Russia) uses a radiosonde to measure temperature and pressure to a height of 10 000 m from Slutsk. From 1930 to 1936, several thousand soundings were made in the USSR with the Moltchanov radiosonde. May 8: M.R. Bureau launches a radiosonde measuring temperature and pressure from Trappes, France, reaching an altitude of 14 400 m. May 22: P. Duckert (Germany) flies the first radiosonde measuring pressure, temperature, and humidity to a height of 15 000 m from the Aerological Observatory at Lindenberg.
1931	December 30: Prof. Vilho Väisälä (Finland) flies a radiosonde from Helsinki telemetering temperature to the ground up to a height of 7 km; like Duckert, Väisälä used the measuring elements to control the capacitance of the radio oscillator circuit.
1936	July 30: Prof. Väisälä establishes the Vaisala company and delivers the first commercial order for 20 radiosondes, delivered to Prof. Carl Gustav Rossby at the Massachusetts Institute of Technology.
1974	The National Center for Atmospheric Research (Boulder, Colorado) develops the dropsonde, a special radiosonde that is launched from research aircraft and measures winds, pressure, temperature, and humidity while descending on a parachute.
1976	The Vaisala company (Helsinki) introduces the first computer-controlled upper-air sounding systems.
1982	The US National Oceanographic and Atmospheric Administration begins routine use of dropsondes for hurricane research; 1 year later, the US Air Force initiates its hurricane reconnaissance program.
1995	The first commercial radiosonde systems using the satellite Global Positioning System to measure winds are introduced by the Atmospheric Instrumentation Research company (Boulder, Colorado) and the Vaisala company (Helsinki).
2005	The GCOS Reference Upper-Air Network (GRUAN) development is launched.

(Continued)

See also: Observations Platforms: Balloons; Kites; Rockets.
Weather Forecasting: Operational Meteorology.

Further Reading

- Aberson, S.D., Franklin, J.L., 1999. Impact on hurricane track and intensity forecasts of GPS dropwindsonde observations from the first-season flights of the NOAA Gulfstream-IV jet aircraft. *Bulletin of the American Meteorological Society* 80 (3), 421–427.
- Beelitz, P., 1954. Radiosonden. VEB Verlag Technik, Berlin, Germany.
- Burpee, R.W., Aberson, S.D., Franklin, J.L., Lord, S.J., Tuleya, R.E., 1996. The impact of omega dropwindsondes on operational hurricane track forecast models. *Bulletin of the American Meteorological Society* 77 (5), 925–933.
- Cohn, S.A., et al., 2013. Driftsondes. Providing in situ long-duration dropsonde observations over remote regions. *Bulletin of the American Meteorological Society* 94 (11), 1661–1674.
- Federal Meteorological Handbook No. 3 Rawinsonde and Pibal Observations, FCM-H3-1997, 1997. Office of the Federal Coordinator of Meteorology, Washington, DC.
- Hock, T.R., Franklin, J.L., 1999. The NCAR GPS dropwindsonde. *Bulletin of the American Meteorological Society* 80 (3), 407–420.
- Oakley, T., 1998. Instruments and Observing Methods. World Meteorological Organization Report No. 72. WMO, Geneva.
- Seidel, D.J., Berger, F.H., Diamond, H.J., Dykema, J., Goodrich, D., Immler, F., Murray, W., Peterson, T., Sisterson, D., Sommer, M., Thorne, P., Vömel, H., Wang, J., 2009. Reference upper-air observations for climate: rationale, progress, and plans. *Bulletin of the American Meteorological Society* 90, 361–369.
- Shea, D.J., Worley, S.J., Stern, I.R., Hoar, T.J., 1994. An Introduction to Atmospheric and Oceanographic Data. Report TN-40411A. National Center for Atmospheric Research, Boulder, CO.
- Weissmann, M., et al., 2011. The influence of assimilating dropsonde data on typhoon track and midlatitude forecasts. *Monthly Weather Review* 139, 908–920.
- World Meteorological Organization, 2008. Guide to Meteorological Instruments and Methods of Observation. Publication No. 8, seventh ed. WMO, Geneva.

Rockets

MF Larsen, Clemson University, Clemson, SC, USA

© 2003 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 4, pp 1449–1454, © 2003, Elsevier Ltd.

Introduction

The influence of the lower atmosphere on our daily lives is immediately clear because of the weather processes that occur there. The processes that occur in the upper atmosphere are also important, although that did not become evident until the space age began at the end of World War II. As we become more dependent on space and satellite communications, as well as traditional radio communications, understanding those processes becomes critical. The ionosphere, which covers the region from approximately 80 km to altitudes above 1000 km, is a mixture of neutral gas constituents and charged particles. The plasma, as it is known, and the electrodynamic forces that drive it can produce small-scale fluctuations in the charged particle densities that can affect radio and satellite communications significantly. The temperatures, composition, and density of the neutral gas component can also have large variations in response to both local processes and changing activity on the Sun. Such effects can alter satellite and other spacecraft trajectories enough to be a concern. Measuring the properties of the atmosphere at high altitudes where the dynamical properties are both interesting and important but where the densities are low enough to make measurements difficult has been a continuing challenge, especially in the period since World War II. The measurement techniques used extensively include ground-based remote sensing with radars and optical techniques, measurements from satellites, and measurements from suborbital rockets, also known as sounding rockets. This article will focus on the techniques used to obtain measurements from the rocket platform.

The layers of the atmosphere closest to the surface of the Earth are the troposphere and stratosphere. Above 45 to 50 km altitude, where the stratosphere ends, the temperature again decreases with altitude in the layer known as the mesosphere, followed by a region of steadily increasing temperature above 90 km, known as the thermosphere, where the absorption of extreme ultraviolet radiation from the sun is primarily responsible for the temperature increase. Radiation from the Sun also leads to the ionization of neutral particles in that region which, because of the importance of the charged constituents, is also known as the ionosphere.

Between the surface of the Earth and the region near 100 km altitude, the density of the atmospheric gas decreases by a factor of approximately 1 000 000, and traditional techniques for making measurements *in situ* from platforms such as balloons or aircraft do not work, since the platforms cannot be supported by the atmosphere. New possibilities for probing the upper-altitude regions became available at the end of World War II when rockets developed by the Germans during the war and captured by the Allies became available for scientific investigation.

History

Sounding rocket studies of the upper atmosphere began in 1946 shortly after the conclusion of World War II. At the close of the war in Europe, the Allies captured the remaining German supply of V-2 rockets that were developed at Peenemunde and used in the bombing of Britain. The V-2 was 14 m in length, weighed almost 13 000 kg, and had a range of over 300 km. In all, 67 complete rockets were assembled at White Sands, New Mexico, from captured parts. 60 were ultimately launched between 1946 and 1952, and a number of these flights were used to make measurements in the upper atmosphere. In a vertical flight, the V-2 could reach an altitude exceeding 160 km, i.e., well into the ionosphere.

It was clear from the outset that the supply of V-2 rockets was limited, and development of rockets designed specifically to be sounding rockets started almost immediately. These included the WAC Corporal which was first developed before the war and the Aerobee. The new rockets had limited capabilities, however. The WAC Corporal, for example, was only able to reach an altitude of 70 km with a payload of 11 kg, and the Aerobee could reach 100 km with a 70 kg payload. By comparison, the V-2 could carry a 1000 kg payload to an altitude of 160 km. The Viking was the first sounding rocket that provided capabilities comparable to the V-2, with a maximum performance that allowed it to reach an apogee of 220 km with a payload weight of 340 kg. In all, 47 V-2 rockets, 91 Aerobees, 8 Vikings, and 1 V-2-boosted WAC Corporal were fired at White Sands in the upper-air program between the spring of 1946 and the fall of 1952. The newer designs during that period focused increasingly on the solid propellants that are still used today rather than the liquid fuels, such as the ethyl-alcohol-based fuel used in the V-2. Solid fuel made storage and handling much safer and simpler.

Launches from White Sands of various sounding rockets continued throughout the period up to 1956 at which time planning for what became known as the International Geophysical Year (IGY) began in earnest. Part of the emphasis of the IGY was to provide detailed global observations of the upper atmosphere using both ground-based and rocket measurements. A series of launches were carried out in the year leading up to 1958, with a large number of launches in the International Geophysical Year itself. By this time the importance of the near space region for communications, satellites, and weapons systems was clear, and the National Aeronautics and Space Administration (NASA) was established on 29 July 1958 as a central organization that would act as the focus for space research. At that time, much of the sounding rocket activity designed to provide critical measurements of upper-atmosphere parameters was moved to Wallops Station on the east coast of the Delmarva Peninsula in Virginia. Other NASA facilities also continued to support sounding rocket activity, as did the research labs within the military branches. Eventually all sounding rocket activity in the United States became

consolidated at the Wallops Flight Facility in 1980. Since the NASA sounding rocket program began in 1958, there have been over 2800 science flight missions.

The first sounding rocket program for studies of the upper atmosphere was developed and continues today in the United States, but a number of other countries have also had and continue to have extremely active rocket programs. Ongoing programs include those in Japan, India, Norway, Sweden, Canada, and Brazil, all of which have active permanent launch facilities. Other countries that have carried out launches at various times in the past include Germany, France, Italy, Spain, Denmark, Australia, and Pakistan. In addition to the fixed launch facilities, a number of remote sites have also been used on a campaign basis for special-purpose launches. These have included Peru, Greenland, Puerto Rico, and North Africa. The maximum launch rate in the United States occurred in the mid-1960s with over 190 launches in 1965. The program currently provides 20–30 flight opportunities per year with launches from a variety of launch sites throughout the world. The level of activity worldwide varies, although the trends are the same as in the United States, with a decrease in the number of launches today as compared with the 1970s, for example.

In part the launch rate has decreased because of the improvements and greater availability of ground-based remote sensing and satellite techniques and in part because the size and complexity of the rocket payloads has increased, so that the rockets being flown today are, in fact, sophisticated spacecraft that cost more and require a longer time to design, manufacture, and test. In the early days of the sounding rocket program immediately after World War II, the payload instrumentation was often quite simple, with a single sensor and a payload weight of a few kilograms. As the technology has advanced and the science questions have become more complicated, the payloads have also become larger and more complex. Sounding rockets today often carry suites of 5–10 different instruments, have attitude control systems to orient the payload in different directions during the flight, and in some cases even have independent sections with instruments and telemetry that separate from the main payload so that large-scale gradients or fluctuations in the medium can be measured.

The increase in mass and volume has led to the requirement for more powerful vehicles. Some vehicles are made specifically for scientific applications. Notable examples are the Black Brant rockets made by Bristol Aerospace in Canada and the Japanese S-series rockets. The United States uses a combination of science vehicles and military surplus rockets such as the Taurus, the Nike, and the Orion for the scientific sounding rocket experiments. Many of the rockets used today are two- or even three-stage vehicles, and payload diameters are generally 14–17 inches (36–44 cm). The Black Brant XI, for example, is a three-stage vehicle that can reach altitudes close to 600 km with a 300 kg payload. The vehicles used today for sounding rocket studies use solid fuel exclusively.

Safety requirements dictate that launches can take place only in isolated locations, since a sufficiently large area has to be available for the unguided rockets to impact safely without danger to people or property. This condition is satisfied in a few places where established ranges exist and in some remote and isolated areas. Active sounding rocket ranges within the United States, for example, include White Sands, New Mexico, at the

site of the first atomic weapons tests, Wallops Island, Virginia, on the east coast of the Delmarva Peninsula, and Poker Flat, Alaska, located just north of Fairbanks.

An unusual chapter in the history of the sounding rocket program occurred in the late 1950s and early to mid-1960s when gun-launched payloads were developed and used. Both 5-inch and 16-inch guns were built. The latter consisted of two 16-inch Navy ship gun barrels welded end-to-end and reinforced with steel support web to minimize the barrel droop when the gun was elevated. The propulsion was provided by bags of gunpowder of the type used in standard artillery firings. Since there was no need for a rocket motor, the vehicle consisted of the payload only. The primary launch sites were the island of Barbados in the Caribbean and the test facility at Yuma, Arizona, although some tests of the smaller gun were also carried out at the NASA facility at Wallops Island, Virginia. The 16-inch gun in particular provided a simple and relatively inexpensive system for getting payloads into the ionosphere. In addition to the relatively low cost, the other primary advantage of the system was that a series of launches could be carried out in rapid succession. It was estimated that the gun could be cleaned and reloaded in a little over half an hour. A number of experiments were carried out in which payloads were launched at 1–2-hour intervals throughout the night. The main disadvantage of the gun-launch approach was that the payload accelerations, i.e., *g*-forces, during launch were as much as 1000 times greater than those produced by a conventional rocket motor. As a result, there were significant ongoing problems with telemetry systems and even the simplest type of instrumentation, and there was only limited success. The most fruitful use of this technique was for chemical tracer release experiments which provide a visual means for tracking the movement of the atmosphere, i.e., the winds, and do not require electronics. More than 65 wind profile measurements using trimethyl aluminum (TMA) as a tracer of the neutral flow were carried out from Barbados alone. By the mid-1960s there was heavy emphasis on the use of more conventional rocket motor systems which allowed more complicated and fragile electronics to be used, and the gun launch program was brought to a close. Recently there has been some renewed interest in the technique because of the relatively low cost and because smaller and much more rugged electronic devices are possible now with the integrated systems that can be made using modern solid-state electronics.

Techniques

Although sounding rockets have been a common element in a number of middle and upper-atmosphere studies since the program started after World War II, the instrumentation used on the rockets has varied greatly. In fact, the number of techniques is too large to list them all here. Some of the more common and representative measurements will be described, however.

The majority of the payload instruments used make measurements with an electronic sensor and relay the measured value to a receiving station on the ground via a telemetry radio link. In the early days of the program, analog modulation of the radio signal was used to relay the information. Telemetry systems have evolved together with digital electronics, and

modern systems now transmit information as digitally encoded signals at kilohertz and even megahertz rates.

Examples of electronic instruments used on rocket payloads can be as simple as a thermistor for measuring temperatures or more complicated, such as a Langmuir probe for measuring electron densities, a double probe for measuring the strength of the electric field, or a mass spectrometer for measuring the composition of the neutral or ionized component of the atmosphere. These techniques will be described in more detail below.

Temperature Profile Measurements

One objective of the sounding rocket program from the earliest days in the late 1950s was to develop an inexpensive vehicle that could carry meteorological instruments above 30 km, the maximum altitude of balloon flights. Prior to 1970, the Arcasonde was used extensively. After 1970 the Super Loki, which is less wind-sensitive at takeoff than the Arcas, was used almost exclusively. These small rockets can easily be handled by one person and can carry a payload of a few kilograms to altitudes of 90 to 95 km. The relatively simple and inexpensive vehicles have been used extensively for routine sampling of the stratosphere and mesosphere to obtain density, temperature, and wind profiles. The actual measurement is usually made on the downward portion of the flight and is telemetered, i.e., sent by radio waves to a receiver on the ground. The temperature system is based on a decelerator called a 'Starute' that is released at apogee. The Starute is a combination of an inflatable sphere with an inflatable torus surrounding the center to enhance the stability of the platform as it descends. The payload consists of a bead thermistor to measure the temperature profile. The wind profile is estimated by tracking the horizontal movement of the Starute with radar during the descent.

Electron Density Profile Measurements

One of the earliest instruments flown on sounding rockets was the Langmuir probe, which is used to measure the electron density profile in the atmosphere. The first flight was in the late 1940s, and the same instrument is still used today in many payloads, although with improvements in the electronic design. The probe consists of an electrode with an applied potential that is inserted into the plasma. The current flowing from the plasma to the electrode is then measured. Since the current density is related to the plasma density, the current measurement can be converted to an ambient plasma density value. In the Langmuir probe measurement, the voltage is often swept over a range of values and the voltage-current characteristic of the probe is measured. By sweeping the voltage, both the electron density and the electron temperature can be measured.

Although the basic principle of the Langmuir probe is simple, there are numerous complications in the design of an actual instrument. The shape of the electrode can affect the measurement. Variations in the conductivity of the electrode due to imperfections in the metal or contaminants on the surface can bias the measurements. Effects of the flow past the electrode due to the orientation of the sensor relative to the rocket orientation can also affect the measurement.

Electric Field Measurements

A type of instrument used extensively to measure electric fields in the atmosphere is the double probe, which has what are essentially two Langmuir probes at floating potential. Generally these are pairs of conducting spheres on nonconducting booms deployed symmetrically about the spacecraft. For a body immersed in a plasma, the floating potential is such that the net current to the body is zero. If a background electric field exists in the medium then a potential difference will be induced between the sensors. In reality, the potential difference has contributions from both the ambient field and the field generated by the motion of the sensor system across magnetic field lines. If the motion of the rocket is known accurately, that contribution to the measured potential difference can be subtracted, thus giving a measurement of the ionospheric electric field. Complications include the effects of potentials induced by photoemissions. The latter will cancel if the sensors are completely symmetric and if the conducting surfaces are as uniform and similar as possible. Other potential problems include the effects of the rocket wake and the currents drawn by the voltmeter used to measure the potential difference.

Composition Measurements

Mass spectrometers have been used extensively throughout the history of the sounding rocket program to measure the composition or concentration of various constituents in the atmosphere. The basic instrument is an adaptation from particle accelerator techniques used in the laboratory. Charged particles are allowed to enter the instrument through an aperture on the surface of the rocket payload. As the particles traverse the instrument chamber, their path can be changed by applying electric or magnetic fields within the chamber, and the actual trajectories will depend on the charge to mass ratio. By varying the fields appropriately, the trajectories can be changed so that only the particles with a certain charge-to-mass ratio strike the detector. The measured current is then proportional to the number of particles. The technique obviously requires that the particles be charged, which would be the case for ions. A neutral mass spectrometer operates on the same principle, but the gas particles entering the chamber are charged first by exposing them to an ion source near the aperture.

Potential problems arise if the density of the gas entering the instrument is too high, since collisions can occur between particles so that the trajectories are altered. A typical operating pressure is 1.333×10^{-4} hPa, which is achieved only at altitudes far above 100 km. Measurements at lower altitudes require that the pressure inside the mass spectrometer be pumped down to the appropriate level. The first instrument with a high-speed cryopump to measure the ion concentration and composition in the middle and upper atmosphere was already flown in 1963.

Photometer Measurements

Photometer measurements rely on the natural light emissions from atmospheric gases to infer the density profile of the

emitting gases. When an atom or molecule is excited by absorption of solar radiation, collisions, or chemical processes, it produces light in a very narrow spectral range as it relaxes to a less excited state. Since each kind of molecule or atom has very specific transitions, the intensity of the light that is measured also gives a measure of the total number of atoms that are emitting. A photometer mounted so that it looks toward zenith as the rocket flies through the atmosphere will measure the contribution from all the emitters located above the rocket at any given instant. This information can be used to construct a profile of the number of emitters as a function of altitude.

By using liquid nitrogen to cool the photometers, the background noise in the instruments can be reduced considerably, and photomultiplier tubes can be used to boost the signal detectability. Such modern improvements in the instrumentation have resulted in extremely sensitive instruments with large dynamic ranges.

Other Techniques

There are a few techniques that do not rely on telemetry. One used extensively is the chemical release technique in which a chemical tracer is released from the rocket payload as it traverses a range of altitude in the atmosphere. The tracer is then tracked photographically from two more sites on the ground so that triangulation techniques can be applied to the combined photographic or video image data in order to determine the motion of the tracers. Some of the chemicals that have been used include barium, strontium, lithium, sodium, and trimethyl aluminum (TMA). All of these produce a visible cloud or trail than can be seen with the naked eye in nighttime conditions and therefore can be photographed with reasonably short exposures of a few seconds' duration. The star background in the photographs is used to determine the look direction to the tracer from a given site. The look direction information from several sites can then be combined to give the absolute location of the tracer at a given instant, and the change in position with time gives the velocity.

The various tracers used extensively include barium, which ionizes when exposed to sunlight and thus provides a way to track the motion of ionization in the ionosphere. Since the barium released has to be in sunlight but the observer has to be in darkness in order to be able to see the tracer, such experiments are usually carried out at twilight when those conditions are fulfilled. The barium is mixed with explosive that is detonated to produce a cloud release that appears red to the observer. Lithium and sodium produce resonant scatter that are red and green, respectively, when exposed to sunlight, but both remain neutral and thus provide a means of tracking the nonionized component of the atmosphere. The payload consists of thermite mixed with the chemicals that is ignited to start the release. The burn raises the temperature of the metals sufficiently to vaporize them so that the material is released as a vapor trail along the rocket trajectory. As with barium, the requirement for sunlight limits the time of the measurements to twilight.

The tracer used in more payloads than perhaps any other is trimethyl aluminum (TMA). The chemical has the property that it reacts spontaneously with oxygen or water. At surface

pressures and densities, TMA produces a rapid burn when exposed to oxygen in the atmosphere, but at altitudes in the ionosphere where the oxygen density is much lower the reaction is much slower. The resulting chemiluminescence is bright enough to be seen with the naked eye at night for 5 minutes or longer. Since TMA is neutral, measurements of the wind profiles can therefore be obtained at any time during the night, not just at twilight.

Another type of measurement uses radio beacons to measure the total electron content between the rocket and a receiver on the ground. Beacon systems operate on the principle that the effective path length between a radio transmitter and a receiver are affected by ionization in the intervening region. Specifically the refractive index in a plasma is less than the refractive index of free space, so that fluctuations of integrated electron density induce different variations in the phase path at each frequency. By using dual-frequency radio transmitters on a rocket and measuring the difference in the phase of the two received signals on the ground, the total electron content along the path between the rocket and the receiver can be determined. By making measurements of the phase difference throughout the flight of the rocket, information about both the horizontal and the vertical gradients in the electron densities can be obtained. No telemetry is required for this instrument.

Conclusion

Much of what was learned about the upper atmosphere during the first few decades after World War II was discovered as a result of sounding rocket measurements. More recently satellite technology has become increasingly important for making extensive measurements from the upper troposphere to the lower thermosphere and extending to the upper part of the thermosphere in some cases. The suborbital sounding rockets continue to fill a special niche by covering the gap between the measurements provided by various ground-based remote sensing techniques and sensing satellite measurements both *in situ* and remote. The sounding rocket experiments flown today can be justified on the basis of the science objectives alone, but the sounding rocket payloads are also a relatively inexpensive testbed for instrumentation that can later be flown on satellites. Sounding rockets thus continue to provide a method for obtaining high-resolution *in situ* measurements in near space at relatively low cost. The obvious limitations are the requirement for a rocket range where the vehicles can be launched safely and can impact without danger to either the public or property, and the relatively short duration of the measurement.

See also: **Basic Atmospheric Structure and Concepts:** Standard Atmosphere. **Chemistry of the Atmosphere:** Observations for Chemistry (*In Situ*): Particles. **Electricity in the Atmosphere:** Ions in the Atmosphere. **Mesosphere:** Ionosphere. **Optics, Atmospheric:** Airglow Instrumentation. **Radar:** Mesosphere–Stratosphere–Troposphere and Stratosphere–Troposphere Radars and Wind Profilers. **Satellites and Satellite Remote Sensing:** Research. Thermosphere.

Further Reading

- Goldberg, R.A., 1986. Handbook for MAP, vol. 19. SCOSTEP Secretariat, University of Illinois, Urbana, IL.
- Newell, H.E., 1953. High Altitude Rocket Research. Academic Press, New York.
- Pfaff, R.F., Borovsky, J.E., Young, D.T. (Eds.), 1998. Measurement Techniques in Space Plasmas: Fields. American Geophysical Union, Washington, DC.
- Pfaff, R.F., Borovsky, J.E., Young, D.T. (Eds.), 1998. Measurement Techniques in Space Plasmas: Particles. American Geophysical Union, Washington, DC.
- Pfaff, R.F., 1996. In-situ measurement techniques for ionospheric research. In: Kohl, H., Ruester, R., Schlegel, K. (Eds.), Modern Ionospheric Science. Max-Planck-Institut für Aeronomie, Katlenburg-Lindau, pp. 459–551.
- Wallace, H.D., 1997. Wallops Station and the Creation of an American Space Program, NASA SP-4311. NASA History Office, Office of Policy and Plans, Washington, DC.

OCEANOGRAPHIC TOPICS

Contents

General Processes

Surface/Wind Driven Circulation

Thermohaline Circulation

Water Types and Water Masses

General Processes

NC Wells, University of Southampton, Southampton, UK

© 2003 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 4, pp 1528–1540, © 2003, Elsevier Ltd.

Introduction

This article will discuss first what is meant by the term ocean circulation; second, how the ocean circulation is determined by measurements and dynamical processes; and finally, the consequences of this circulation on the Earth's climate.

What is Ocean Circulation?

The ocean circulation in its simplest form is the movement of sea water through the ocean, which principally transfers temperature and salinity from one region to another. Temperature differences between regions give rise to heat transfers. Similarly, differences in salinity produce transfers of salt. On the time scale of the ocean circulation, the inputs and exports of salt into and out of the ocean make a negligible contribution to overall salinity and so variations in salinity occur by the addition and removal of fresh water into and out of the ocean.

Two major processes control the ocean circulation: the action of the wind and the action of small density differences within the ocean, produced by differences in temperature and salinity. The former process is the wind-driven circulation, while the latter is the thermohaline circulation. Although it is useful to separate these two processes to better understand the ocean circulation, they are not independent from each other.

The ocean circulation is in reality a very complex system, because the flows are not steady in time or space. They are turbulent flows that show variability on scales from the largest scale of the ocean basins to the smallest scales where the energy is finally dissipated as heat. This turbulent structure of the ocean means there are fundamental limitations on the predictability of its behavior.

Because of this inherent complexity, oceanographers have approached ocean circulation by using a combination of

observational methods, including ships, buoys, and satellites, combined with the mathematical methods of dynamical oceanography. This integrated approach allows hypotheses to be made that can be tested by comparison with observations. Furthermore, mathematical models of the ocean circulation, based on the dynamical principles, can be constructed and tested against observations.

This article will consider how we measure ocean circulation, how we determine the major processes at work, and the consequences of the ocean circulation on the climate system.

How Do We Determine the Ocean Circulation?

The determination of ocean currents involves measurement of the displacement of an element of fluid over a measured time interval. The position of the measurement is defined mathematically in a Cartesian coordinate system (**Figure 1**) in which x is positive eastward direction (lines of constant latitude), y is positive northward direction (toward the geographic North Pole), and z is positive upward; $z = 0$ corresponds to mean sea level. Without ocean currents and tides, the sea level would be an equipotential surface – that is, one of constant potential energy. The z coordinate is perpendicular to the equipotential surface. The origin is the intersection of the Greenwich meridian (Universal meridian) and the Equator with mean sea level.

The coordinates of a parcel of water may be determined by the Global Positioning System (GPS). This satellite-based system provides a horizontal position with an accuracy of better than 100 m, which is sufficient for large-scale flows in the ocean. Large-scale flows are at least 10 km in spatial scale and have time scales of at least a day. A pressure device attached to a current meter normally determines the vertical position.

There are two mathematical methods for defining the displacement of the fluid. One is to measure the velocity of the

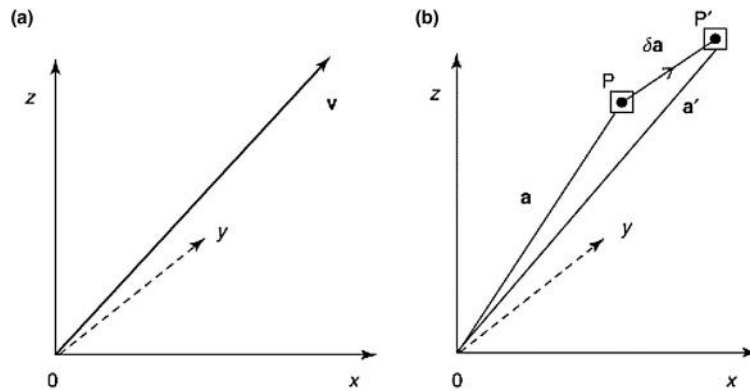


Figure 1 (a) Eulerian and (b) Lagrangian specifications of flow. For both systems, $+x$ is the eastward direction, $+y$ is the northward direction, and z is vertically upward. (a) The Eulerian specification involves measurement of the velocity ($\mathbf{v}$) of the fluid at a fixed point (x, y, z) . (b) The Lagrangian specification involves measurement of the velocity of a fluid element: $\mathbf{a}$ is the position vector from the origin to the fluid element. An element of fluid moves a small distance $\delta \mathbf{a}$ in a time δt , from P to P' . Velocity is $\delta \mathbf{a} / \delta t$. In the infinitesimal limit, $\delta \mathbf{a} / \delta t \rightarrow d\mathbf{a} / dt = \mathbf{v}(\mathbf{a}, t)$.

fluid at a fixed point in the ocean, and the other is to follow the element of the fluid and to measure its velocity as it moves through the ocean. The first method is known as an Eulerian description and the second is a Lagrangian description of flow. In principle, the two methods are independent of each other. This means that an Eulerian measurement cannot provide Lagrangian currents, nor can Lagrangian measurements provide Eulerian currents.

Having defined the two mathematical methods, we will now consider how we can measure currents in practice. Initially, these methods will address only the measurement of the horizontal flow. The vertical flow is difficult to measure directly, and will be discussed later in this article.

First we will consider the Eulerian method. The measurement of the flow at a fixed point in the ocean is only straightforward when a fixed position can be maintained, for instance when a current meter attached to the bottom of the ocean or to a pier on the coast. Most measurements have to be made well away from land. This is achieved by attaching the current meter to a mooring, which is attached to weights and then deployed (Figure 2). The position of the mooring can be determined from GPS. The current meter may be a rotary device or an acoustic device. The rotary current meter measures the number of revolutions over a fixed period, while the acoustic type measures the change in frequency of an emitted sound pulse caused by the ocean current (i.e., it uses the Doppler effect). Moorings may be deployed for periods of up to 2 years. In the analysis of the record, it is normal to remove the high-frequency variability of less than 1 day, caused by tides, by filtering of the data.

A Lagrangian measurement of current can be determined by following an element of water with a float. The horizontal displacement of the water over a small interval of time defines the Lagrangian current. Figure 3 shows typical float designs. The position of the float can be determined by two methods. A float that has a surface satellite transmitter/receiver can have its position determined by GPS, while a subsurface float would use an acoustic navigation system. Some floats can descend to a predetermined depth, maintain that depth for a few weeks and then return to the surface for a position fix. This technique

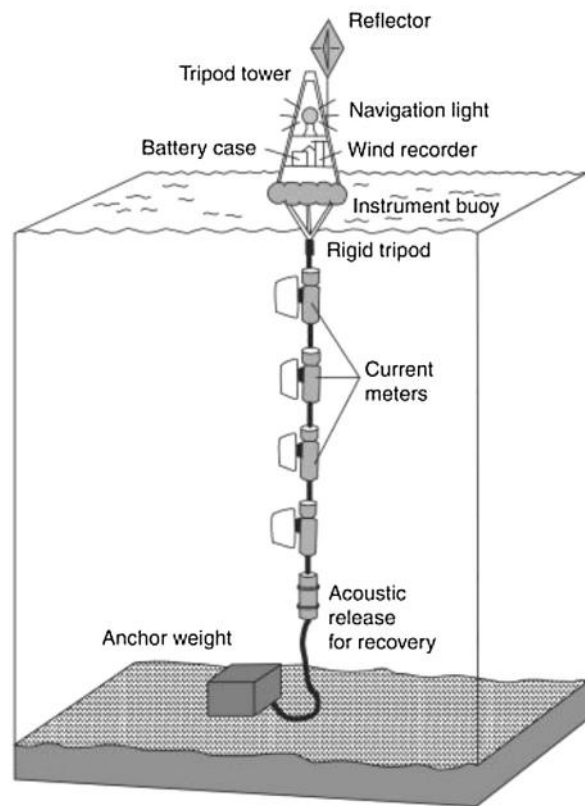


Figure 2 A typical current meter mooring.

allows the current to be measured down to depths of 1 km below the surface.

Each method gives different information on the flow field. A mooring will give a time series of the horizontal current, whereas a float will give the trajectory of the horizontal displacement of the parcel. It is worth remarking that most of the information on the surface ocean circulation has come from mariners' observations of the ships set, a method that has been used since the nineteenth century. However, these

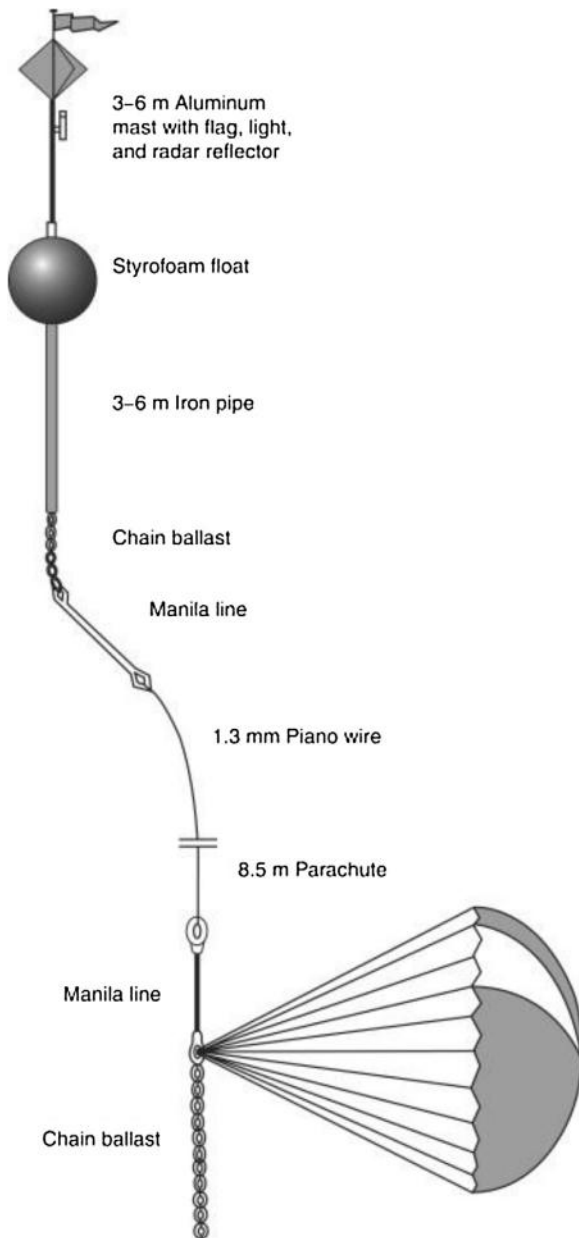


Figure 3 A typical drifter with a parachute drogue at a few meters below the surface. It will follow the current at the depth of the parachute.

measurements have their limitations since they are neither Eulerian nor Lagrangian measurements and additional influences (e.g., wind effects on the ship) may cause errors.

This information can be analyzed in many different ways to discern the major current systems. From a set of moorings deployed for a few years across, say, the Gulf Stream, we can determine the mean flow (i.e., the average of all the current measurements) and we can determine its variability. The mean flow could be calculated over a particular period of time. This period is limited by the period of deployment, which is of the order of 2 years. This is rather short for a climatological mean, and a much longer period of 10 years is desirable. A few longer time-series of currents have been determined for the Gulf

Stream in the Florida Straits and for the Antarctic Circumpolar Current in the Drake Passage.

Recall that ocean currents are turbulent and therefore have variability on a whole range of time scales. Hence the mean flow gives no information on the variability of the flow. However, we can calculate the statistics of the flow, based on the kinetic energy.

The kinetic energy (KE) per unit volume is defined as in eqn [1].

$$KE = \frac{1}{2} \rho (u^2 + v^2) \quad [1]$$

In eqn [1], ρ is the density of the sea water and u and v are the eastward and northward components, respectively, of the horizontal flow.

If the time mean current is defined as $\bar{u}$, and the deviation from $\bar{u}$ at any time as u' , we can define the mean kinetic energy (KEM) and eddy kinetic energy (EKE) by eqns [2] and [3]

$$KEM = \frac{1}{2} \rho (\bar{u}^2 + \bar{v}^2) \quad [2]$$

$$EKE = \frac{1}{2} \rho (u'^2 + v'^2) \quad [3]$$

These two numbers give quantitative measures on the mean and variability of the flow respectively. The ratio EKE/KEM gives a measure the relative variability of the flow. If the ratio is very much less than 1 then the flow is steady; if the ratio is approximately equal to 1 then the flow is very variable. Figure 4 shows the variability of the flow in the Agulhas Current, which is an intense and highly variable current off the coast of South Africa.

Although the EKE/KEM ratio gives a measure of the variability of the current, it does not give any idea of the exact time or space scales over which the current is varying. For example, the current may show a slow change from one season to another or it may show faster variation due to eddies. To address this variation we can use time-series analysis, such as Fourier analysis, to determine the KE of the flow for different time periods. Fourier analysis produces a spectrum of the KE, either in frequency for a time-series or in wavenumber for a spatial variation in flow. Figure 5 shows the analysis of a time-series into its component frequencies. If the current were varying on all time scales, the spectrum would be flat; if there were only one dominant period, the spectrum would peak at that one frequency. This particular analysis shows that the current is varying at the tidal frequency and the inertial frequency, both at the high-frequency end of the spectrum. The inertial frequency is given by $2\Omega \sin \phi$, where Ω is the rotation rate of the Earth and ϕ is the latitude. At the lower-frequency end, which corresponds to the ocean circulation frequency, there is a broad band of high kinetic energy. This band is due to eddies that cause fluctuations of currents on time scales of weeks to months.

For these mean climatological currents, our knowledge has been augmented by the application of the dynamic method. This method is based on the observation that large-scale ocean currents are in geostrophic balance over large areas of the ocean. Geostrophic balance means that the Coriolis force balances the horizontal pressure gradient force. The geostrophic flow is a good approximation to the flow in the

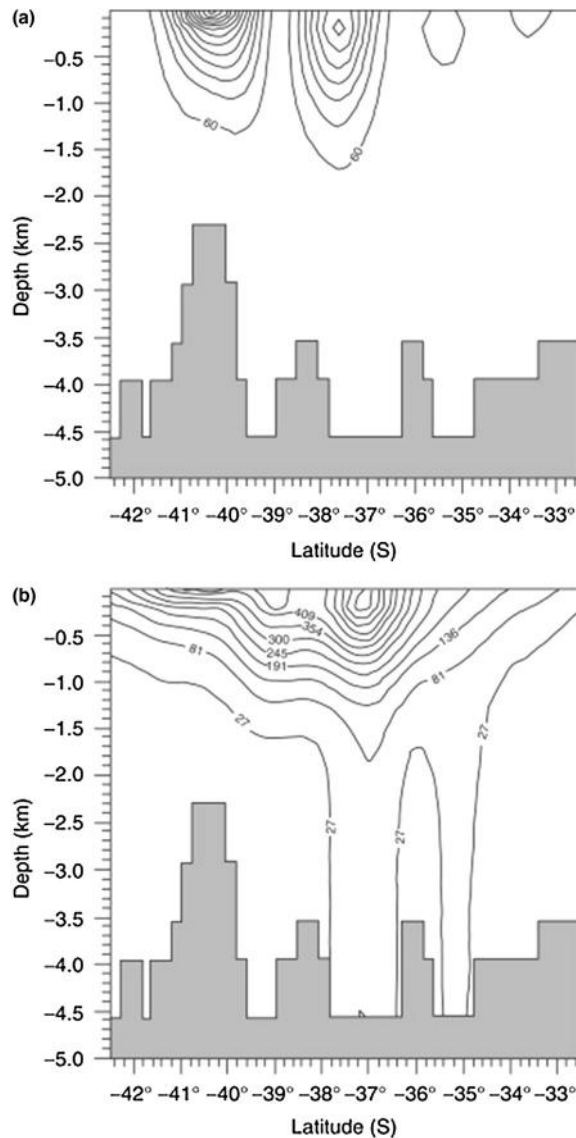


Figure 4 The mean kinetic energy (KEM) and the eddy kinetic energy (EKE) in a north-south slice through the Agulhas Current System at 14.4° E. Contour units are $\text{cm}^2 \text{s}^{-2}$. The KEM maximum corresponds to the mean position of the Agulhas Return Current (eastward flow) between 40° and 41° S, and the Agulhas Current (westward flow) between 37° and 38° S. The EKE distribution is much broader than that of KEM, which shows the large horizontal extent of the flow variability. The ratio EKE/KEM is typically about 1/3, which indicates a very variable current system. Reproduced from Wells NC, Ivchenko V, and Best SE (2000) *Journal of Geophysical Research* 105, 3233–3246.

interior of an ocean outside the equatorial region. The horizontal pressure gradient is dependent on the slope of the ocean surface and the horizontal variation of the density distribution within the ocean. The geoid is an equipotential surface that would be represented by the sea level of a stationary ocean; ocean currents cause deviations in sea level from the geoid. In the future, the slope of the ocean surface may be determined by satellite measurements of the sea surface height and the geoid, but at present we do not have an accurate geoid at sufficiently

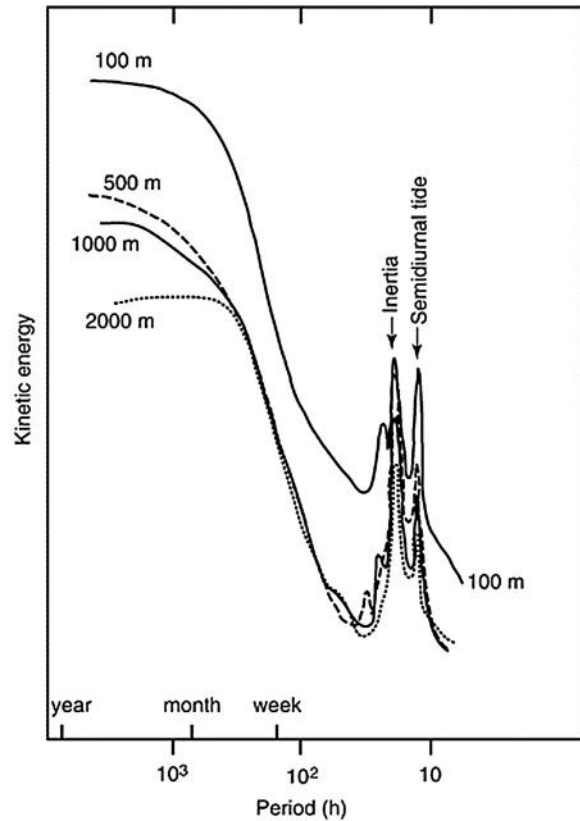


Figure 5 Frequency spectrum of kinetic energy from four depths at Site D (39° N, 70° W), north of the Gulf Stream. Note the two high-frequency peaks, coinciding with the inertial period (19 h) and the semidiurnal tide (12.4 h). Reproduced from Rhines PB (1971) *Deep Sea Research* 18: 21–26.

high resolution to measure the sea surface slope. The horizontal variation of density can be determined from temperature, salinity, and pressure measurements that have been made over large ocean areas during the last century. The dynamic method allows determination of the vertical shear of the horizontal geostrophic current; therefore, additional measurements are required to determine the absolute geostrophic current. For example, if the current has been measured at a particular depth, the dynamic method can be referenced to that depth and the vertical profile of current can be obtained.

The recent World Ocean Circulation Experiment (WOCE) hydrographic program has provided more measurements of the ocean than all the previous hydrographic programs and will give the most comprehensive assessment of the climatological horizontal ocean flow to date.

Recall that we cannot measure the vertical circulation of the ocean directly because it is technically too difficult to do so at present. Current indirect methods used to determine the vertical circulation rely on the use of mathematical approaches, such as dynamical models, or the use of chemical tracers.

Observations of temperature and salinity can be inserted into a mathematical ocean general circulation model (see Box 1) which allows the 3D circulation to be determined, subject to limitations in the accuracy of the model.

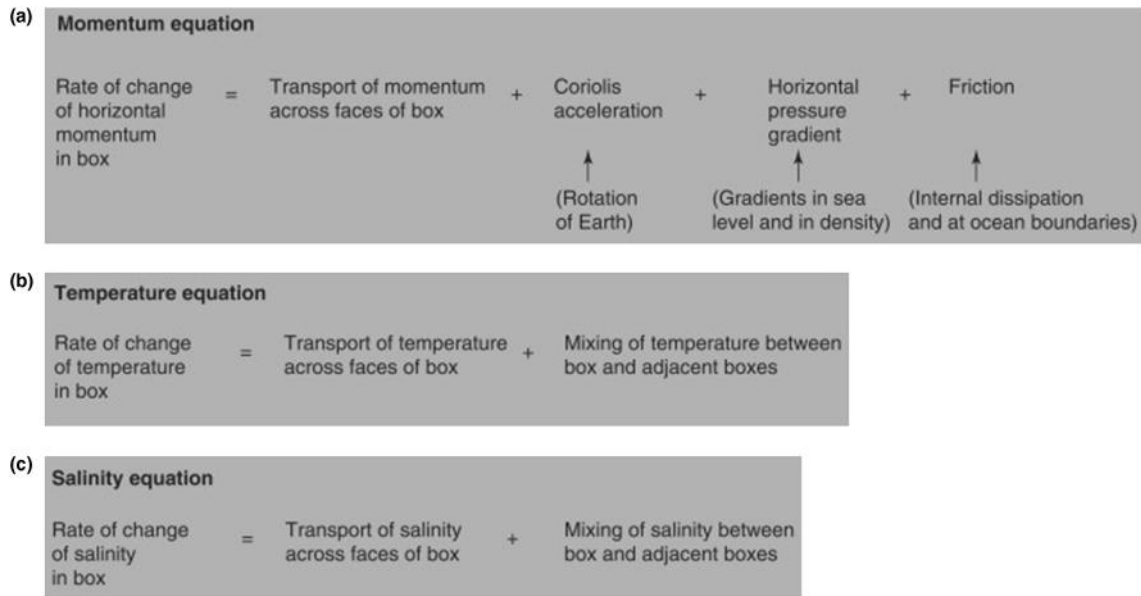


Figure 6 The basic equations for an ocean general circulation model. Reproduced from Summerhayes CP, and Thorpe SA (1996) *Oceanography: An Illustrated Guide*. London: Manson Publishing.

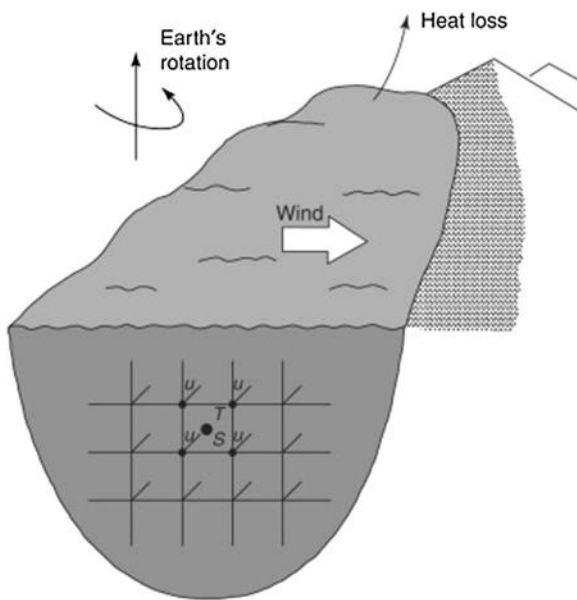


Figure 7 A schematic of the model boxes in an ocean general circulation model. The equations for momentum are solved at corners of the boxes (u), while the temperature (T), and salinity (S) equations are solved at the centers of boxes. The model is forced by climatological wind stress, surface heat fluxes, and fresh water fluxes. Reproduced from Summerhayes CP, and Thorpe SA (1996) *Oceanography: An Illustrated Guide*. London: Manson Publishing.

Chemical tracers have been injected inadvertently into the ocean from nuclear tests in the 1960s and from industrial processes (e.g., chlorofluorocarbons). Naturally occurring tracers such as ^{14}C also exist. These tracers can be measured with high accuracy in a few laboratories around the world and, from their distributions at different times, the 3D circulation

can be estimated. This method reveals the time history of the ocean circulation wherever the tracer is measured. This is very different information from that provided by the methods previously discussed, but nonetheless it can reveal unique aspects of the flow. For example, nuclear fallout deposited in the surface layer of the Nordic seas in the 1960s was located in the deep western boundary current 10 years later.

Wind-Driven and Thermohaline Circulation

The wind-driven circulation is considered first. The surface layer of the ocean is directly driven by the surface wind stress and is also subject to the exchange of heat and fresh water between ocean and atmosphere. This layer, which is typically less than 100 m in depth, is referred to as an Ekman layer. A steady wind stress causes a transport of the surface water 90° to the right of the wind direction in the Northern Hemisphere and 90° to the left in the Southern Hemisphere because of the combined action of the wind stress on the ocean surface and the Coriolis force. These Ekman flows can converge and produce a downwelling flow into the interior of the ocean. Conversely, a divergent Ekman transport will produce an upwelling flow from the interior into the surface layer (see [Figure 8](#)).

This type of flow is known as Ekman pumping, and is directly related to the Curl of the wind stress. It is of fundamental importance for the driving of the large-scale horizontal circulation in the upper layer of the ocean. For example, between 30° and 50° latitude the climatological westerly wind drives an Ekman flow equatorward, while between 15° and 30° latitude the Trade Winds drive an Ekman flow polewards. At about 30° latitude the flows converge and sink into the deeper ocean. Before we can discuss the influence of Ekman pumping on the interior ocean circulation, we need to consider role of density.

Box 1 An Ocean General Circulation Model

An ocean general circulation model comprises a set of mathematical equations that describe the time-dependent dynamical flows in an ocean basin. The basin is discretized into a set of boxes of regular horizontal dimensions but variable thickness in the vertical dimension. The horizontal flow (northward and eastward components) is predicted by the momentum equation (Figure 6(a)) at the corners of each box (Figure 7).

The forcing for the flow may come from the wind stress (the frictional term in the momentum equation) and from the surface buoyancy fluxes, arising from heat and fresh water (precipitation+runoff-evaporation) exchange with the atmosphere and adjacent land masses. These buoyancy fluxes change the temperature and salinity in the surface layer of the ocean. The surface water masses are then subducted into the interior of the ocean by the vertical and horizontal components of the flow, where they are mixed with other water masses.

The processes of transport and mixing are described by the temperature and salinity equations (Figure 6(b) and 6(c)) at the center of each ocean box (Figure 7). From these two equations, the sea water density, and thence the pressure, can be obtained for each box. The horizontal pressure gradient is then determined for the momentum equation, while the vertical velocity is calculated from the horizontal divergence of the flow. This set of time-dependent equations can then be used to describe all the dynamical components of the flow field, providing suitable initial and boundary conditions are specified.

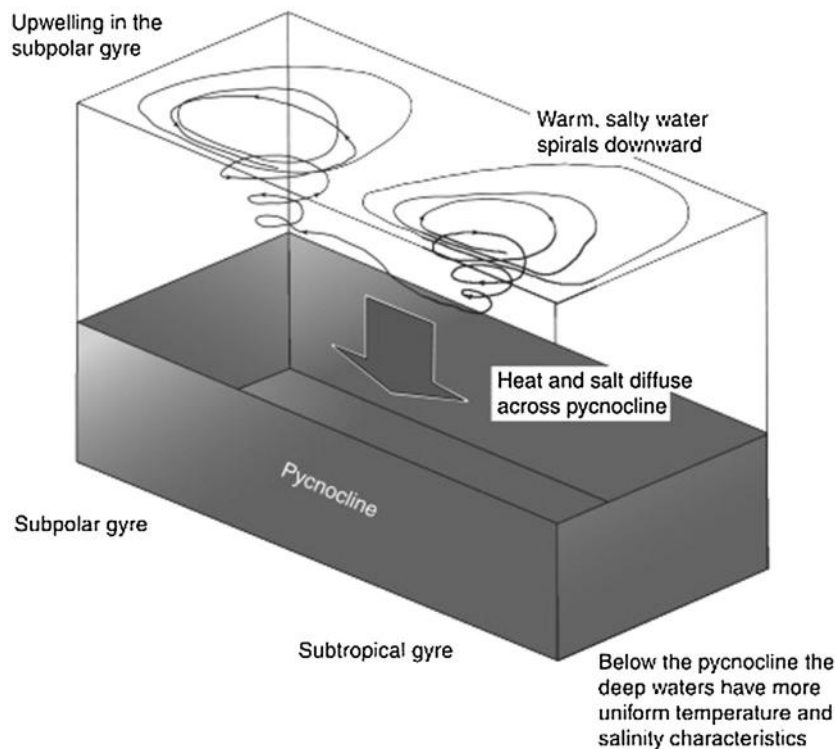


Figure 8 A schematic representation of the wind-driven circulation in the subpolar and subtropical gyre of an ocean basin. The wind circulation causes a convergence of Ekman transport to the center of the subtropical gyre and downwelling into the interior. Conversely, in the subpolar gyre there is a divergence of the Ekman transport and upwelling from the interior into the surface layer. This Ekman pumping is responsible for the gyre circulations (see text for details). The western boundary currents are depicted by the closeness of the streamlines. They are caused by the poleward change in the Coriolis force known as the beta effect. Reproduced from Bean MS (1997) PhD thesis, University of Southampton.

As we move from the surface to the deepest layers of the ocean, the density of sea water increases. From hydrographic measurements of density, we can map the horizontal variation of the depth of a chosen density surface. These constant-density surfaces are known as isopycnals. They have the important property that the flow tends to move along these surfaces and therefore the variations in the depth of these surfaces gives us a picture of the horizontal flow in the deep ocean, away from the surface layer and benthic layer. The isopycnal surfaces dip down in the center of the subtropical gyre at about 30° . The formation of this lens of light warm

water is related to the climatological distribution of surface winds, which produce a convergence of Ekman transport toward the center of the gyre and a downwelling of surface waters into the interior of the ocean. At the center of the lens, the sea surface domes upwards, reaching a height of 1 m above the sea surface at the rim. Owing to hydrostatic forces, the main thermocline is depressed downward to depths of the order of 500–1000 m (Figure 9).

The surface horizontal circulation flows anticyclonically around the lens with the strongest currents on the western edge, where the slope of the density surface reaches a maximum.

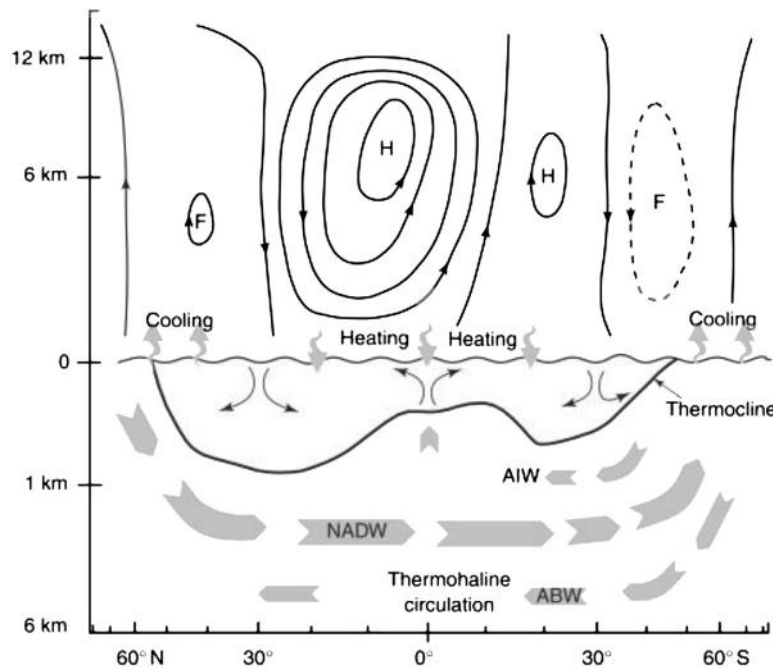


Figure 9 A representation of the meridional average section through the atmosphere for December–February. The cells are the Hadley cell (H) and the Ferrel cell (F). The strength of the cells is represented by the solid contours which are in units of 40 Mt s^{-1} , whilst the dashed contours are in units of 20 Mt s^{-1} . Note the predominantly downward motion at $\sim 30^\circ$ latitude, associated with the subtropical anticyclones, and the strong upward motion at equatorial latitudes which is associated with the Inter-Tropical Convergence Zone. A meridional transect through the Atlantic Ocean, showing the position of the main thermocline. The small arrows represent the wind driven downwelling (Ekman pumping) at $\sim 30^\circ$ latitude, and the equatorial upwelling, which occurs within and above the main thermocline. The North Atlantic Deep Water (NADW) is produced in the Nordic Seas and is the predominant water mass by volume. The Antarctic Intermediate Water (AIW) is produced at $\sim 50^\circ \text{ S}$ and by virtue of its salinity is lighter than the NADW. In contrast Antarctic Bottom Water is the most dense water mass in the world's ocean and is formed in the Weddell and Ross Seas. These deep flows form part of the thermohaline circulation. The vertical scales are exaggerated in the lower troposphere and in the upper ocean. The horizontal scale is proportional to the area of the Earth's surface between latitude circles.

These are geostrophic currents, where there is a balance between the Coriolis force and the horizontal pressure gradient force. Generally, the circulation in the subtropical gyres is clockwise in the Northern Hemisphere and anticlockwise in the Southern Hemisphere. These large-scale horizontal gyres are ultimately caused by the climatological surface wind circulation and are found in all the ocean basins.

The surface layer is also subject to heating and cooling, and the exchange of fresh water between ocean and atmosphere, both of which will change the density of the layer. For example, heat is lost over the Gulf Stream on the rim of the light water lens of the subtropical gyre. Recall that flow tends to follow isopycnal layers and these layers will slope downward toward the center of the gyre. Cooling of the waters in the Gulf Stream leads to the sinking of surface waters to produce a water mass known as 18°C water. This water, which is removed from the surface layer, will slowly move along the isopycnal layers into the thermocline. As it moves clockwise around the gyre it will be subducted in to the deeper layers of the thermocline, in a spiral-like motion (Figure 8). The deepest extent of the main thermocline is located in the subtropical gyre to the west of Bermuda on the eastern edge of the Gulf Stream rather than in the center of the ocean basin. This asymmetry of the gyre is related to the beta effect – the change of the Coriolis parameter with latitude.

The subtropical gyres are one of the best-studied regions of the ocean, and our understanding is therefore most developed in these regions. These gyres occur in the surface and thermocline regions of the ocean and are primarily controlled by the wind circulation, with modifications due to heating and cooling of the surface. The question now arises why we observe thermoclines in the ocean. For example, why is the warm water not mixed over the whole depth of the ocean and why is the average ocean temperature about 3°C .

To explain the observed behavior, we need to consider the thermohaline circulation, which is generated by small horizontal differences in density, due to temperature and salinity, between low and high latitudes. How does it work? Consider an ocean of uniform depth and bounded at the Equator and at a polar latitude. We will assume it has initially a uniform temperature and is motionless (we will ignore for the moment the effect of salinity on density). This hypothetical ocean is then subject to surface heating at low latitudes and surface cooling at high latitudes. In the lower latitudes the warming will spread downward by diffusion, while in high latitudes the cooling will spread downward by convection, which is a much faster process than diffusion. The heavier, colder water will induce a higher hydrostatic pressure at the ocean bottom than will occur at low latitudes. The horizontal pressure gradient at the ocean bottom is directed from the high latitudes to the lower

latitudes, which will induce an equatorward abyssal flow of polar water. The flow can not move through the equatorial boundary of our hypothetical ocean and therefore will upwell into the upper layer of the tropical ocean, where it will warm by diffusion. The flow will then return poleward to the high latitudes, where it downwells into deepest layers of the ocean to complete the circuit. It is found that the downwelling occurs in narrow regions of the high latitudes, while upwelling occurs over a very large area of the tropical ocean. This hypothetical ocean demonstrates the key role of the deep horizontal pressure gradient, caused by horizontal variations in density, for driving the flow.

To explain the observed thermohaline circulation, this hypothetical ocean has to be modified to take into account the Coriolis force, which causes the deep abyssal currents to flow in narrow western boundary currents, the effect of salinity on the density (the haline component of the flow), asymmetries in the buoyancy fluxes between the Northern and Southern Hemispheres, and the complex bathymetry of the ocean basins.

We will now give a descriptive account of the thermohaline circulation. The deepest water masses in the ocean have their origin in the polar seas. These seas experience strong cooling of the surface, particularly in the winter seasons. In the North Atlantic, there are connections through the Nordic seas to the Arctic Ocean, through which sea ice flows. Heat energy melts the ice in the North Atlantic and the meltwater gives rise to further cooling. There are two effects on the density of the water: Cooling increases the density whereas surface freshening, due to ice melt, decreases the density of the water. The former process usually dominates the density and hence denser waters are produced. These dense, cold waters flow into the Atlantic through the East Greenland and West Greenland Currents and then into the Labrador Current. These cold waters mix and sink beneath the warm North Atlantic Current.

In addition to surface polar currents, we also have deep ocean currents. The cold saline water entering from the Nordic seas mixes as it sinks to the abyssal layers of the ocean and moves southward as a deep current along the western boundary of the Atlantic. This water mass is known as NADW (North Atlantic Deep Water) and is the most prominent and voluminous of all the deep water masses in the global ocean. It flows into the Antarctic Circumpolar Current, from where it flows into the Indian and Pacific Oceans. In addition to NADW, colder denser water – Antarctic Bottom Water (AABW) – enters the Southern Ocean from the Antarctic shelf seas. It is not as voluminous as NADW but it flows northward in the deepest layers into the Atlantic, where it can be distinguished as far north as 30°. These deep flows upwell into the thermocline and surface waters where they return to the North Atlantic. This global thermohaline circulation has been termed the global conveyor circulation to signify its role in transporting heat and fresh water (Figure 10) around the planet.

How does this circulation explain the thermocline? We can estimate the rate at which these cold, deep abyssal waters are produced and we know for a steady state in the ocean that production has to be balanced by removal. A large-scale upwelling of the abyssal waters into the thermocline produces this removal. Our simple conceptual picture is of warm thermocline water mixing downward, balanced by a steady upwelling of the cold abyssal layers. Without the

upwelling, the warm waters would mix into the deepest layers of the ocean.

The Role of Fresh Water in Ocean Circulation

In the discussion so far we have seen that the wind-driven circulation and the thermohaline circulation are major components of ocean circulation, which are ultimately driven by the surface wind stress and buoyancy fluxes. Buoyancy fluxes are the net effect of heat exchange and the fresh water exchange with the overlying atmosphere. We have seen that heat exchange is a major process explaining the existence of both the thermocline and the deep abyssal water, but what is the role of the fresh water in ocean circulation?

In the subtropics there is net removal of fresh water by evaporation. This increases the salinity of the water, which, in turn, increases the density of the thermocline waters. Normally this effect is opposed by heating, which lightens the water. However, in the Mediterranean and the Red Sea, evaporation produces saline waters, which by virtue of their salinity and cooling in winter, sink to the deepest layers of the basins. At the Straits of Gibraltar, this dense saline layer flows out beneath the incoming fresher and cooler Atlantic water. This Mediterranean water forms a distinct layer of high-salinity water in the eastern Atlantic Ocean. Similar behavior occurs at Bab el Mandeb adjacent to the Gulf of Aden.

The influence of fresh water is more substantial in the polar oceans. A given amount of fresh water will have a greater effect on density at low temperatures than at high temperatures, because the thermal expansion of sea water decreases with decreasing temperature. At higher latitudes, we have a net addition of fresh water into the oceans, which arises from the excess of precipitation over evaporation and the melting of sea ice moving toward the Equator from the polar regions.

The addition of fresh water adds buoyancy to the surface layer, while cooling removes buoyancy; therefore, the fresh water will tend to reduce the effect of the cooling. In the Arctic Ocean, the surface layer is colder but less dense than the warmer layer at ~100 m and therefore is in equilibrium. This stable halocline in the Arctic Ocean reduces the vertical heat flux into the deep ocean.

In the subpolar oceans, the addition of fresh water reduces the density of the surface layer and can reduce the prevalence of deep convection. This happened in the late 1960s when fresh water, probably from excessive ice in the Arctic Basin, melted in the subpolar gyre. The effect on the thermohaline circulation is unknown, but it is believed from modeling studies that the drop in the production of deep waters reduced the thermohaline circulation of the ocean.

What Are the Consequences of this Circulation on the Climate System?

The effects of the ocean circulation on the climate can be understood in terms of the heat capacity of the ocean. The heat capacity of a column of sea water only 2.6 m deep is equivalent to that of a column of the whole atmosphere and therefore the ocean heats and cools on a long time scale compared with the atmosphere.

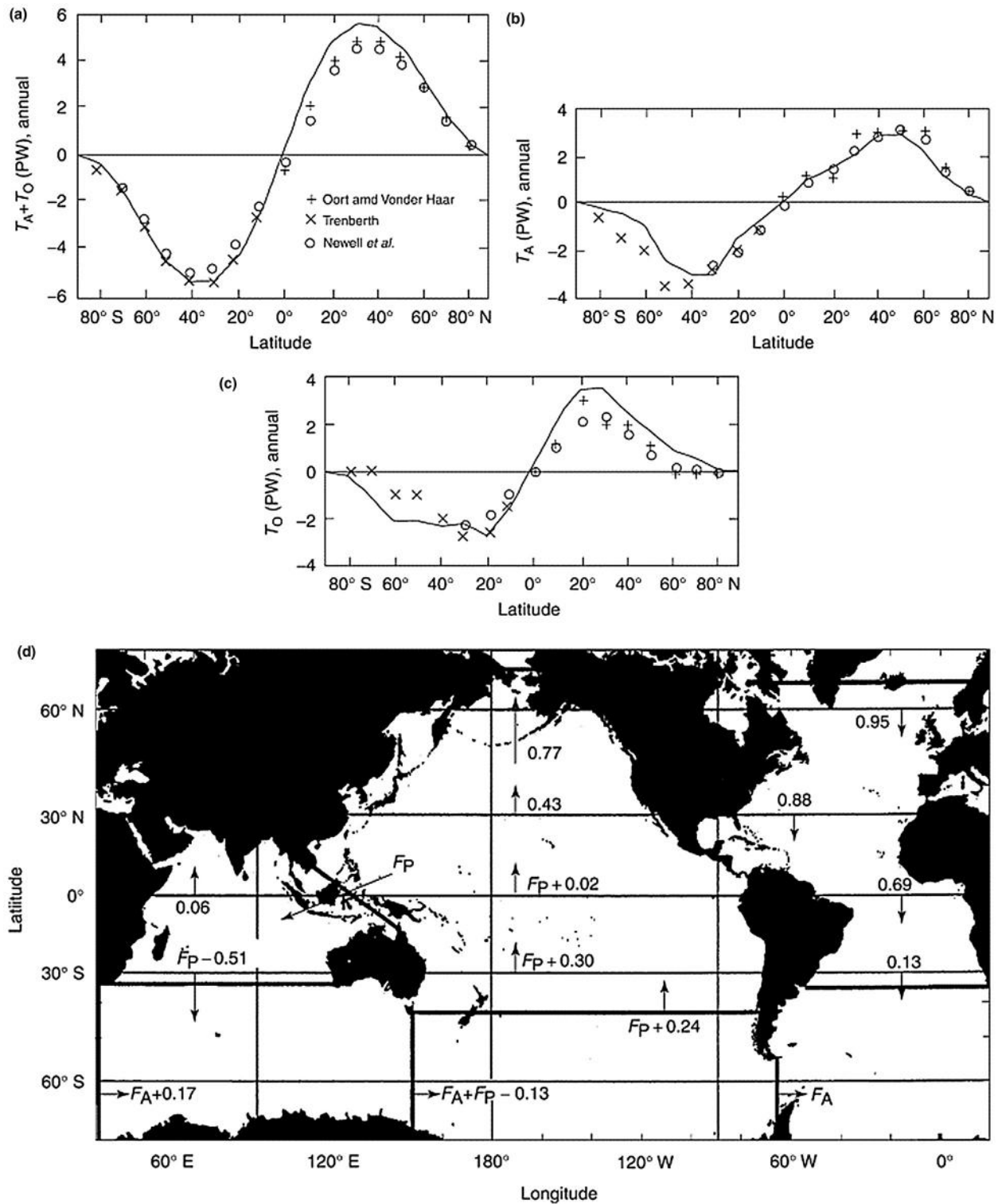


Figure 10 (a) Poleward transfer of heat (a) by ocean and atmosphere together ($T_A + T_O$), (b) by atmosphere alone (T_A), and (c) by ocean alone T_O . The total heat transfer (a) is derived from satellite measurements at the top of the atmosphere; that of the atmosphere alone (b) is obtained from measurements of the atmosphere; and (c) is calculated as the difference between (a) and (b). 1 petawatt (PW) = 10^{15} W. (d) An estimate of the transfer of fresh water (10^9 kg s^{-1}) in the world ocean. In polar and equatorial regions precipitation and river runoff exceed evaporation, and hence there is an excess of fresh water, while in subtropical regions there is a water deficit. A horizontal transfer of fresh water is therefore required between regions of surplus to regions of deficit. F_P and F_A refer to the fresh water fluxes of Pacific-Indian throughflow and of the Antarctic Circumpolar Current in the Drake Passage, respectively. (a)–(c) Reproduced from Carrissimo BC, Oort AH, and von der Haar TH (1985) *Journal of Physical Oceanography* 15: 85. (d) Reproduced from Wijffels SE, Schmitt R, Bryden H, and Stigebrandt A (1992) *Journal of Physical Oceanography* 22: 158.

We know that there is poleward gradient of temperature, which is driven by the thermal radiation imbalance between the low and high latitudes. In response to this temperature gradient, we require a flow of heat from the warmer to cooler latitudes. Both the atmosphere and ocean circulations transfer this heat from low to high latitudes by a variety of mechanisms. In the low latitudes of the atmosphere there is the Hadley cell, which transfers low-temperature air in the lowest levels via the Trade Winds toward the Equator and transfers warmer air poleward in the upper troposphere (Figure 9). At higher latitudes, anticyclones and cyclones and their accompanying upper-air jet streams transfer heat poleward. In the ocean, the wind-driven Ekman currents transfer heat as surface waters move across latitude circles. This water is returned deeper in the ocean at a different temperature from that of the surface water. The ocean gyres carry heat toward higher latitudes since the poleward flows of the western part of the gyres are warmer than the equatorward flows in the eastern parts of the gyres. Finally, and not least, there is the contribution of the thermohaline circulation, which transports warm surface and thermocline waters to the highest latitudes and returns cold water to lower latitudes. Figure 10 shows the heat transport and fresh water in the ocean.

A major difference between the atmosphere and the ocean is the relative speed of their circulation. The atmosphere circulation is a fast system, responding on time scales of days to weeks. For example, weather systems in temperate latitudes grow and decay on time scales of a few days. By contrast, the ocean tends to be slower in its response. The fastest parts of the system are the surface Ekman layers, which respond to changes in the surface wind circulation on a time scale of 1 or 2 days. Changes in wind circulation can cause planetary waves that will change sea level and surface temperature on monthly to seasonal time scales. In particular, the equatorial oceans respond to the surface wind stress on seasonal time scales, which allows a strong coupling between the ocean and atmosphere to take place. This gives rise to phenomena such as the El Niño Southern Oscillation. The subtropical gyres respond to changes in the wind circulation on decadal time scales, while the deep thermohaline circulation responds on millennial time scales. There is some evidence for rapid changes of local parts of the thermohaline circulation on time scales of 50 years.

Observations of the deep ocean are far fewer in number than at the ocean and land surface. The longest continuous data set is from a deep station at Bermuda that commenced operations in 1954. Observations from cruises in the earlier part of the century are of unknown quality and it is therefore difficult to know whether differences are due to the use of different instruments or to real changes in the ocean. It is only since the 1950s that such changes have been measured accurately. Figure 11 shows changes in the temperature for that period across the Atlantic. These changes are of the order of a few tenths of a degree over periods of 15 years. Recall that the heat capacity of oceans is very much larger than atmosphere: hence these changes in temperature involve very significant changes in the heat content of the ocean. The World Ocean Circulation Experiment from 1990 to 1997 has provided measurements of ocean properties such as temperature, salinity, and chemical tracers as well as current

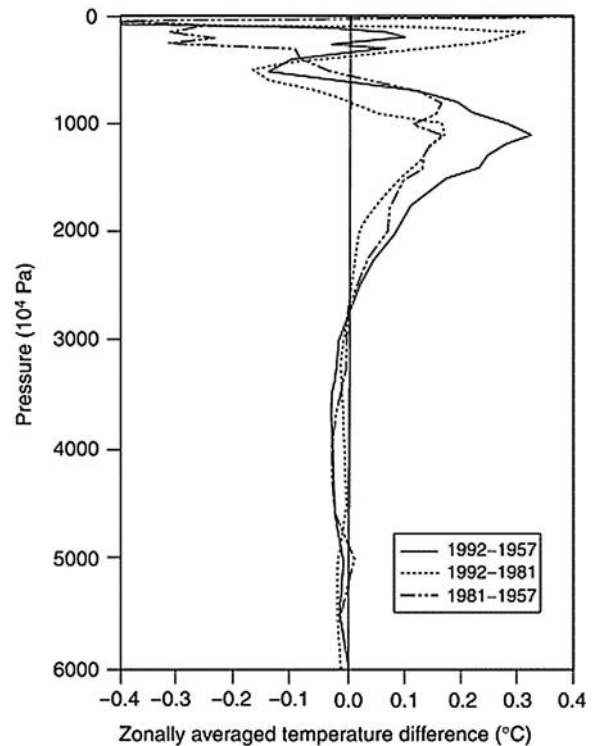


Figure 11 Temperature changes (°C) in the subtropical North Atlantic (24° N), 1957–1992. The measurements have been averaged across 24° N between North Africa and Florida. Reproduced from Parrilla G, Lavin A, Bryden H, Garcia M, Millard R (1994) Rising temperatures in the Subtropical North Atlantic Ocean. *Nature* 369: 48–51.

measurements on a global scale. This set of high-quality measurements will provide the baseline from which future changes in ocean circulation can be determined.

In contrast to the brief record of deep ocean observations, sea surface temperature measurements of reasonable quality go back to the late nineteenth century. These measurements can be used to assess changes in the surface layers. Salinity measurements are fewer and not as reliable, but changes can be still detected.

Salinity measurements in the Mediterranean over the last century have shown a warming of the Western Mediterranean Deep Water of 0.1°C and increase of 0.05 in salinity. These reasons for this change are not known, but it has been speculated that the change in salinity may be attributed to a reduction in the fresh water flow due to the damming of the Nile and of rivers flowing into the Black Sea.

An important question recently identified is the stability of the thermohaline circulation. The thermohaline circulation is driven by small density differences and therefore changes in the temperature and salinity arising from global warming may alter the thermohaline circulation. In particular, theoretical modeling of the ocean circulation has shown that the thermohaline circulation may be reduced or turned off completely when significant excess fresh water is added to the subpolar ocean. In the event of the thermohaline circulation being significantly reduced or stopped, it may take many centuries before it returns to its present value.

In view of the current levels of uncertainty, it is necessary to continue to monitor the ocean circulation, as this will provide the key to the understanding of the present circulation and enhance our ability to predict future changes in circulation.

See also: **Dynamical Meteorology:** Quasigeostrophic Theory.
Oceanographic Topics: Surface/Wind-Driven Circulation; Thermohaline Circulation; Water Types and Water Masses.
Tropical Meteorology and Climate: El Niño and the Southern Oscillation: Observation; El Niño and the Southern Oscillation: Theory; Walker Circulation.

Further Reading

- Gill, A.E., 1982. *Atmosphere–Ocean Dynamics*. Academic Press, London. p. 666.
Siedler, G., Church, J., Gould, J., 2001. *Ocean Circulation and Climate*. Academic Press, London.
Summerhayes, C.P., Thorpe, S.A., 1996. *Oceanography: An Illustrated Guide*. Manson Publishing, London.
Wells, N.C., 1997. *The Atmosphere and Ocean: A Physical Introduction*, second ed. Wiley, Chichester.

Surface/Wind Driven Circulation

RX Huang, Woods Hole Oceanographic Institution, Woods Hole, MA, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by P Bogden and C A Edwards, volume 4, pp 1540–1549, © 2003, Elsevier Ltd.

Synopsis

Surface motions in the upper ocean include surface waves and wind-driven currents. These motions are driven by surface winds. There is a wave boundary layer in the upper ocean where surface waves dominate. Below this surface boundary layer, the ocean can be conceptually separated into several layers, including the Ekman layer, the mixed layer and the main thermocline, and the thick layer below. These layers are characterized by different dynamics, but they may overlap. The most outstanding feature is the main thermocline, and it can be treated as the base of the wind-driven circulation in the upper ocean.

Introduction

Winds on the sea surface provide the major source of energy responsible for motions in the upper ocean over broad spatial and temporal scales, from surface waves and small-scale turbulent motions to large-scale oceanic currents. Although people walking on the beach or on board of ships can easily observe surface waves and small-scale turbulence and currents, the large-scale currents can be studied through well-planned scientific observations only. Wind-driven circulation is a key player in regulating the sea surface temperature and the air–sea heat flux; thus, wind-driven circulation is an important component of the climate system. This article is focused on the dynamic structure of the wind-driven circulation; hence, the feedback to the atmosphere is not discussed here.

Wind Stress Pattern

Wind stress on the sea surface is one of the most important driving forces for the oceanic circulation. Wind stress generates small-scale surface waves first; through wave–wave interaction, energy is transferred in phase space, leading to surface waves of long wavelength and large amplitude. However, in comparison with the larger-scale currents, surface waves are considered as small-scale problems and they are not considered as parts of wind-driven circulation discussed in this article.

Winds on the sea surface represent the velocity structure at the base of the atmospheric boundary layer. As such, they are directly linked to the circulation above the atmospheric boundary layer, and their pattern reflects the overall structure of the atmospheric general circulation; thus, they have remarkable global-scale structure, [Figure 1](#), where the annual mean winds stress is displayed. The most outstanding feature of the wind stress over the global oceans is the strong westerlies at midlatitudes of both hemispheres, which is the surface expression of the jet stream of the atmospheric circulation. At low latitudes, the trade wind dominates; at the equator, easterlies dominate in the Pacific and Atlantic Oceans, but in the Indian Ocean relatively weak westerlies dominate.

For a long time, winds were measured through ship-board instruments over the global oceans. Since the advance of satellite technology, wind speed and direction over the global ocean can be directly inferred from microwave measurements

over the sea surface (such as the QuikSCAT data). The common practice in oceanography is to calculate wind stress using the bulk formulae $\tau_w = \rho_{\text{air}} C_d |\vec{u}_{\text{atm}} - \vec{u}_{\text{oce}}| (\vec{u}_{\text{atm}} - \vec{u}_{\text{oce}})$, where ρ_{air} is the air density, $C_d \approx 10^{-3}$ is the empirical drag coefficient, $\vec{u}_{\text{atm}} - \vec{u}_{\text{oce}}$ is the velocity difference between the air and water within the boundary layers. Note that C_d is not a constant, and it varies with the wind speed. The bulk formulae are based on many *in-situ* measurements; however, the accuracy of the bulk formulae remains to be improved, especially for the case with strong wind. In addition, although in many previous applications the effect of oceanic current were not included in the calculation of wind stress, recent studies indicated that the correction due to the ocean current should be taken into consideration.

Wind stress changes with time, and it has a noticeable seasonal cycle over most parts of the world ocean. In particular, seasonal cycle of wind stress in the Indian Ocean and adjacent oceanic regions is quite strong. For example, wind over the Somali and Vietnamese coasts blows in opposite directions during different seasons, [Figure 2](#). The seasonal cycle of wind stress in these areas strongly affects the local upwelling and the strength and direction of coastal currents.

Furthermore, wind energy input into surface waves is a strong nonlinear function of wind stress, so that the high frequency components of wind stress make a critical part of contribution to the wind energy input into the ocean. Although monthly mean wind stress can be used to calculate the wind-driven circulation with reasonable results, mixed layer property calculation requires wind stress products with high temporal resolution up to 6-hourly wind.

Surface Circulation in the World Ocean

Apart from the surface wave boundary, the upmost part of the ocean is occupied by the Ekman layer where the frictional force is balanced by the Coriolis force. Below this relatively thin upper layer, currents in most parts of the world oceans are organized in the form of gigantic gyres, as shown in [Figure 3](#). There are strong subtropical gyres in the North Pacific, North Atlantic, South Pacific, South Atlantic, and the South Indian Oceans. At high latitudes, there are subpolar gyres in the North Pacific and North Atlantic Oceans, and Weddell Sea Gyre and Ross Sea Gyre at the southern edge of the South Ocean.

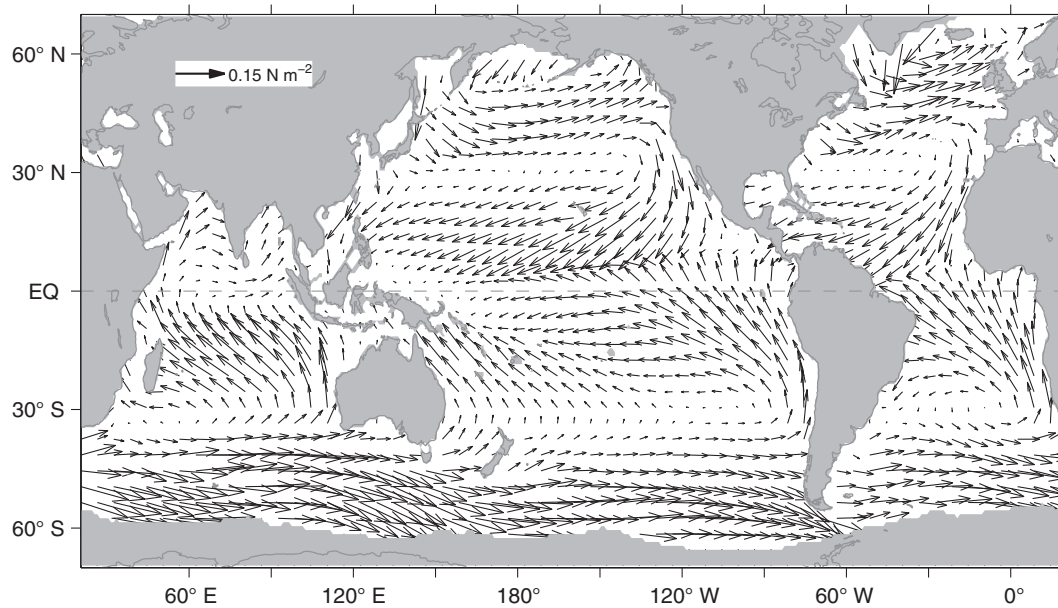


Figure 1 Annual mean wind stress, based on QuikSCAT data.

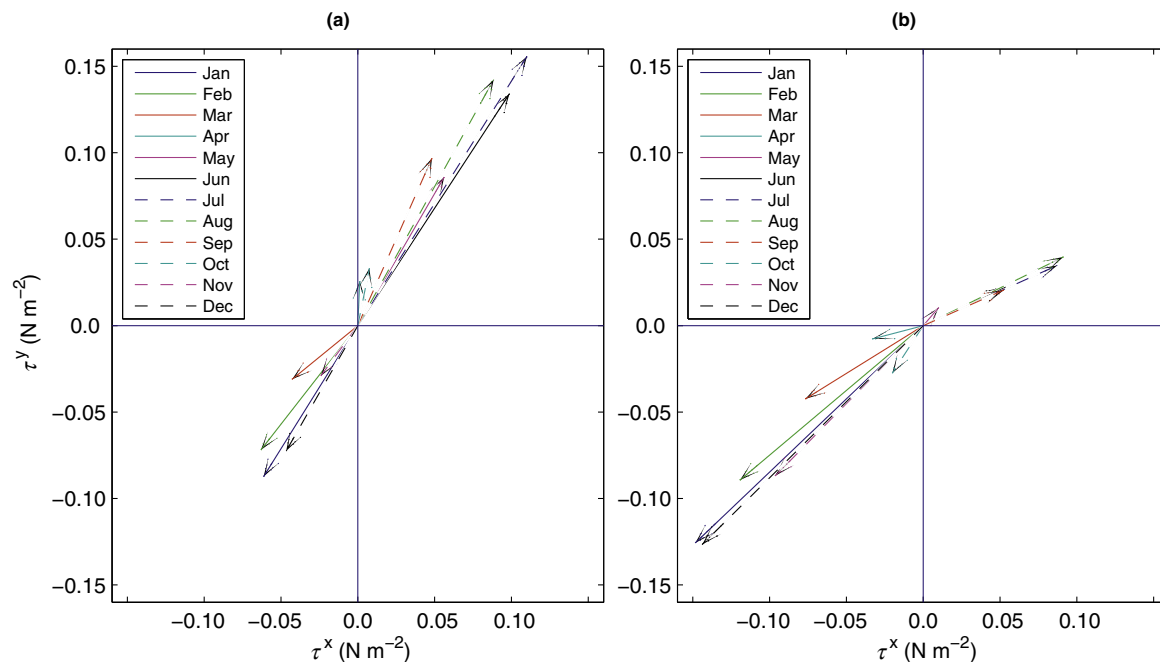


Figure 2 Seasonal cycle of wind stress in the (a) Somali coast (at 50.125° E, 5.125° N) and (b) South of Vietnam (at 108.125° E, 10.125° N), based on WOA09 data.

These gyres are directly driven by surface winds. Many oceanic currents observed in the oceans are part of the organized wind-driven circulation discussed in this article, including the Gulf Stream in the North Atlantic Ocean, the Kuroshio in the North Pacific Ocean, the Brazil Current in the South Atlantic Ocean, the East Australian Current in the South Pacific Ocean, and the Agulhas in the Southwest Indian Ocean.

One of the most important features of these wind-driven gyres is the western intensification of the currents,

as schematically shown in [Figure 3](#). Although most currents in the ocean interior move relatively slowly, these western boundary currents can reach the speed in excess of 1 m s^{-1} . Most importantly, they can carry a huge amount of water. For example, the Gulf Stream transports can reach to 150 Sv ($1 \text{ Sv} = 10^6 \text{ m}^3 \text{ s}^{-1}$). In addition, they carry a large amount of heat and thus play a critically important role in setting up the poleward heat transport in the Earth's climate machinery.

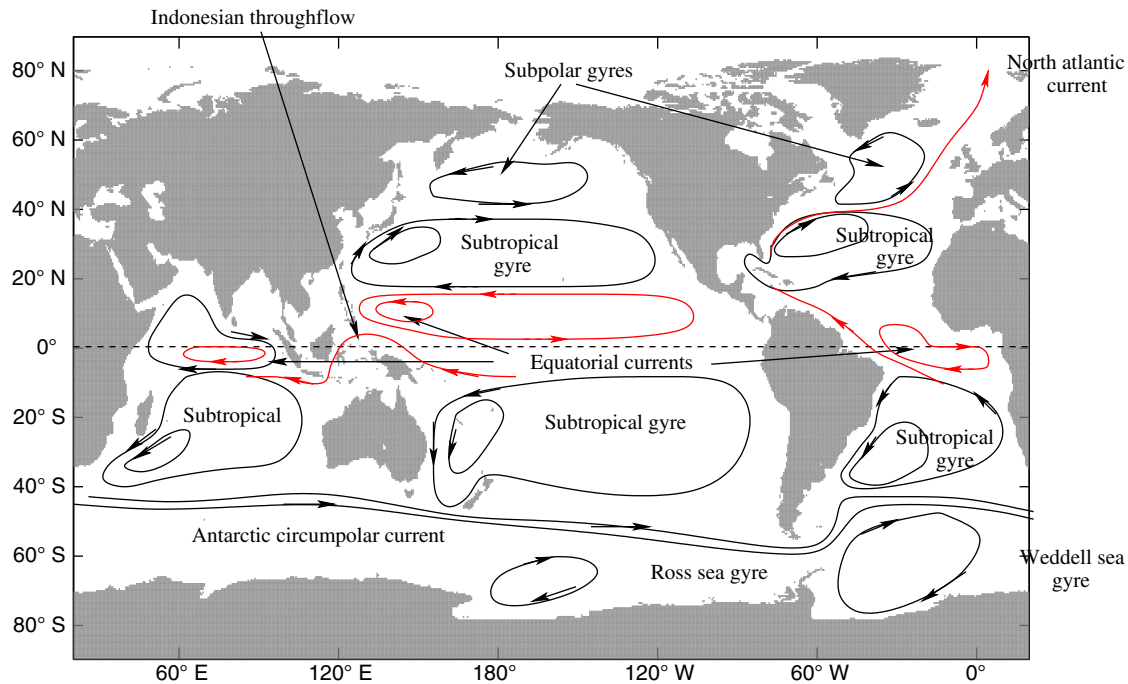


Figure 3 Major surface currents in the world oceans.

The fast-moving western boundary currents were discovered long time ago by marine navigators. Benjamin Franklin and Timothy Folger in 1769–70 created the first chart of the Gulf Stream, which was widely used by navigators going on cruises between European countries and North America. Even at the present day, their chart of the Gulf Stream seems quite accurate in prescribing the large-scale feature of this fast-moving current system. The dynamical explanation of such fast-moving currents or the so-called western intensification had to wait for more than 170 years following this chart.

At low latitudes, there are strong equatorial current systems in the Pacific, Atlantic, and Indian Oceans, and they play critical roles in the oceanic circulation and climate system.

There is also a strong Antarctic Circumpolar Current (ACC), which is the only circum-earth current system under the present continental setting. In addition, there are many other surface currents, which connect surface circulation in different parts of the world oceans, such as the Indonesian Throughflow and the North Atlantic Current.

The shape of subtropical gyres is somewhat similar to the pattern of wind stress in the subtropical basins. The similarity between the circulation patterns in the atmosphere and oceans is deeply rooted in the dynamics. In fact, as will be explained shortly, the formulation of the gyration is the consequence of potential vorticity balance in a closed basin; thus, the wind stress curl is the essential ingredient of wind-driven gyres.

Thermal Structure in the Upper Ocean

The circulation in the upper ocean is closely related to the density structure. Seawater density is controlled by temperature and salinity; however, over much of the warm near-surface waters, temperature dominates. Therefore, density structure

can also be inferred from temperature structure in the oceans. There is a mixed layer on the top of the ocean, where the temperature, salinity, and density is vertically nearly homogenized. Water properties and depth of the mixed layer have a profound annual cycle.

The mechanical energy required to sustain the turbulent motions in the mixed layer is provided by the wind stress applied to the sea surface. In particular, wind stress inputs about 60 TW ($1 \text{ TW} = 10^{12} \text{ W}$) of mechanical energy into the surface waves. The exact amount of this energy remains a topic of intensive research, and its pathway in the ocean remains unclear. However, it is believed that most part of this energy is dissipated within the mixed layer, leaving only a quite small portion of it to be transformed into the deeper part of the ocean, probably in forms of near-inertial oscillations. In addition, convection due to surface cooling or salt rejection during sea ice formation can also provide kinetic energy sustaining turbulent motions in the mixed layer. It is to emphasize that, however, the kinetic energy associated with turbulent motions due to convection is converted from the gravitational potential energy originally stored in the system. In fact, convection in the ocean cannot create mechanical energy.

Another outstanding feature in the thermal structure in the upper kilometer of the world ocean is the main thermocline. The thermocline is defined as a layer within the water column where the vertical gradient of temperature is a local maximum. There are four major types of thermoclines in the ocean: diurnal, seasonal, main, and abyssal thermoclines. The diurnal thermocline is associated with the diurnal cycle of the mixed layer and it exists in the upper few tens of meters in the ocean; the seasonal thermocline is associated with the seasonal cycle of the mixed layer, and it extends from the base of the diurnal thermocline to the depth of a couple of hundred meters except

in some special locations where wintertime convection can reach to great depths. The main thermocline is below the seasonal thermocline at depths of 200–800 m. As a result, it is not greatly affected by the seasonal cycle of the mixed layer; hence, it is also called the permanent thermocline. In addition, in part of the world oceans there is abyssal thermocline associated with abyssal circulation.

Wind-driven circulation can exist in either a homogeneous ocean or a stratified ocean. Wind-driven circulation in a homogeneous ocean is very weak because the volumetric flux of the circulation is uniformly distributed over the entire depth of the ocean. On the other hand, due to the existence of strong stratification associated with the main thermocline, wind-driven circulation is mostly confined to an upper moving layer above the main thermocline. As a result, currents associated with wind-driven circulation in a stratified ocean can be greatly enhanced.

In most parts of the world oceans, seawater density is primarily controlled by temperature, with salinity playing a secondary role. Therefore, the main thermocline is also closely linked to the main pycnocline. Note that dynamically the main pycnocline is more directly relevant to the dynamics of the oceanic circulation; however, the main thermocline itself is closely linked to temperature changes in the atmosphere, hence thermocline is a term often used in scientific study of the oceanic circulation and climate changes.

The typical structure of the main thermocline can be seen through an east–west temperature section. As shown in Figure 4, the main thermocline is located around the depth of 200 m at the eastern boundary, and gradually slopes down to the depth of 600 m in the North Pacific Ocean and 800 m in the North Atlantic Ocean. As shown in Figure 4, warm water in the upper ocean is separated from the cold water in the deep ocean by a layer of relative shape temperature gradient associated with the main thermocline.

In the interior part of the ocean, large-scale currents obey geostrophy. Using geostrophy and thermal wind relation, the

direction of large-scale current can be inferred from temperature or density sections as follows. Since strong wind-driven current is confined in the upper kilometer of the ocean, one can assume that water at great depth is motionless. For example, one can assume that horizontal velocity at the depth of 2.5 km is negligible, and the corresponding horizontal pressure gradient is nearly zero at this depth. As shown in Figure 4, water on the right-hand side of the ocean basins is colder and thus denser than that on the left-hand side. Using the thermal wind relation and the hydrostatic approximation, one comes to the conclusion that at depth shallower than 2.5 km pressure gradient force in the oceanic interior is pointed eastward. According to geostrophy, to balance the pressure gradient force the Coriolis force must point westward; thus, current in the ocean interior should move equatorward.

On the other hand, near the western boundary the slope of isopycnal flips sign and becomes quite steep, indicating strong and poleward narrow western boundary currents. Thus, temperature and the corresponding density structure shown in Figure 4 can be interpreted as the sign of the western intensification.

Density structure along the meridional section reveals another important dynamic feature, Figure 5. The center of the wind-driven circulation on each isopycnal surface is roughly the deepest part of the isopycnal surface. As shown in Figure 5, the center of the circulation moves northward with increasing density. This is called the poleward intensification of the wind-driven subtropical circulation. This is also linked to the recirculation of the subtropical gyre, as will be discussed shortly.

In general, near the equator thermocline is much shallower. Since the Coriolis force vanishes near the equator, the equatorial thermocline is directly linked to the equatorial zonal wind stress. In a steady state, at the sea level the zonal wind stress on surface is balanced by the pressure gradient force associated with the zonal sloping sea surface. In both the Pacific and Atlantic Oceans, easterlies prevail near the equator.

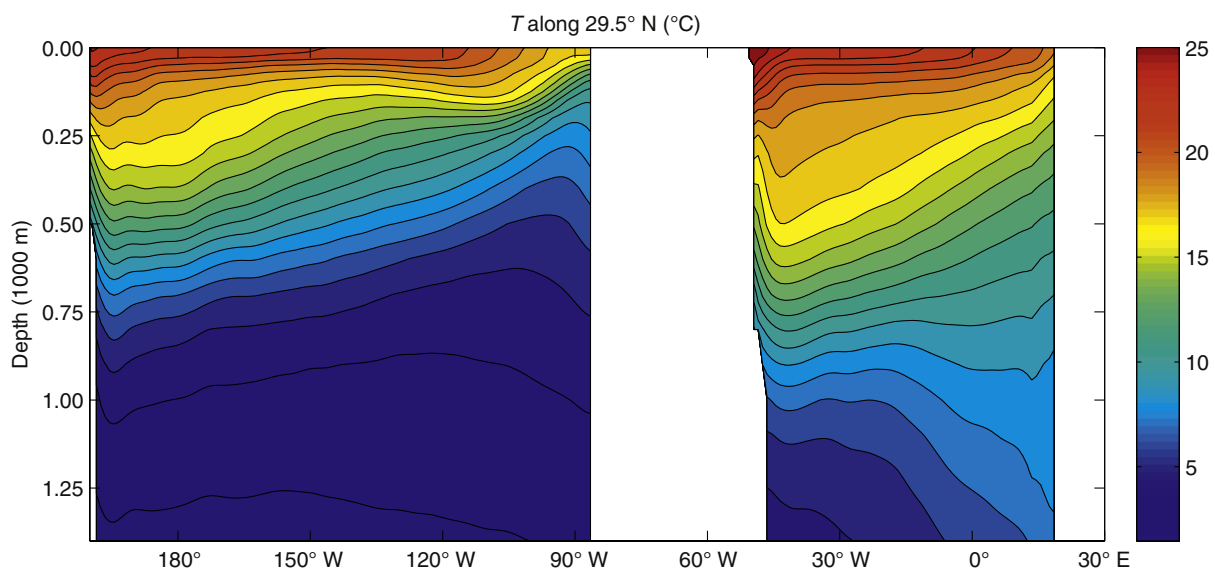


Figure 4 Temperature section along 29.5° N, based on WOA09 data.

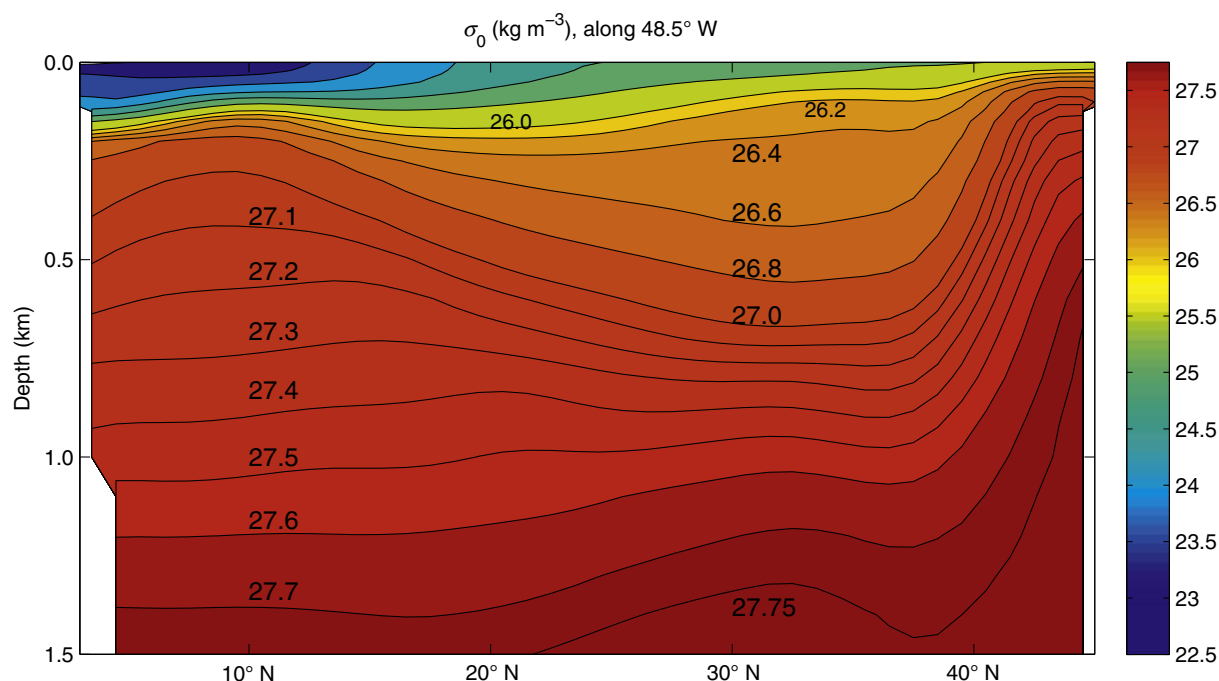


Figure 5 A potential density section along 48.5° W, based on WOA09 data.

As a result, equatorial thermocline in these two basins slopes downward from east to west. On the other hand, the main thermocline in the Indian Ocean slopes down eastward because wind in the Indian Ocean is primarily westerly. This combination of equatorial winds leads to a condition favorable for the formation of a large warm pool with the main thermocline sits at the depth of 200 m in the western equatorial Pacific. In the east side, there is a cold tongue in both the equatorial Pacific and Atlantic Oceans, upper panel of [Figure 6](#).

In the meridional section, the main thermocline appears in the form of a dumbbell, lower panels of [Figure 6](#). The bowl-shaped main thermocline at midlatitudes in these three basins is outstanding. It is clear that the main thermocline is deepest in the South Indian Ocean. The main thermocline in the North Atlantic Ocean is deeper than that in the North Pacific Ocean.

The depth of the main thermocline in the world oceans is closely linked to the wind-driven circulation in the upper ocean. In the world oceans, there are five subtropical gyres and they can be clearly identified from the basin-scale bowl-shaped main thermocline, as shown in [Figure 7](#). Note that the main thermocline is a conceptual layer only; thus, such a subsurface vertical temperature gradient maximum may not exist at any specific location. As will be shown shortly, the simple reduced gravity model predicts that the depth of the main thermocline is proportional to the zonal integration of the Ekman pumping rate and inversely proportional to the stratification in the upper ocean. For example, in the Pacific Ocean relatively low salinity in the upper ocean leads to a relatively strong stratification in the upper ocean and thus a shallow main thermocline. On the other hand, high salinity water in the Atlantic and Indian Ocean leads to relatively weak stratification and relatively deep main thermocline.

Theory of the Wind-Driven Circulation

Ekman Layer

There is an Ekman layer in the upper ocean. The Ekman layer is defined as the surface boundary layer in which the frictional force is balanced by Coriolis force. Ekman in 1905 first formed the idea of such a boundary layer in the ocean. Winds input a large amount of mechanical energy, on the order of 3 TW, which is used to maintain motions in the Ekman layer against friction. Within the Ekman layer, the wind stress is transformed downward through eddy-induced horizontal momentum flux. A major uncertain part of the Ekman layer theory is the vertical eddy viscosity A_v . The classical theory of Ekman layer assumes that A_v is isotropic and has a constant value over the whole depth of the layer. Under such an assumption, the horizontal velocity for the Ekman layer in a steady state appears in the form of a spiral, quite similar to that in the atmospheric boundary layer.

Observing the Ekman layer in the ocean was a great challenge due to the stance of strong surface waves and turbulence in the upper ocean. Ekman layer predicted by the classical theory was confirmed through observation only in the 1980s. Observations indicated that vertical eddy viscosity A_v is not constant. In fact, *in-situ* observations indicated that vertical eddy viscosity decays with increasing depth following some negative power laws or exponentially decaying laws. If A_v is not constant or isotropic, the shape of the Ekman spiral can be different from that predicted by the classical theory.

The horizontal volume transport integrated over the depth of the Ekman layer is independent of the vertical eddy viscosity; this volume transport is perpendicular to the wind stress and pointing to the right-hand side (in the Northern Hemisphere); it can be written in the form of $\vec{V}_{\text{Ekman}} = -\vec{z} \times \vec{\tau} / f \rho_0$, where τ is the wind stress, f is the Coriolis parameter, ρ_0 is mean water

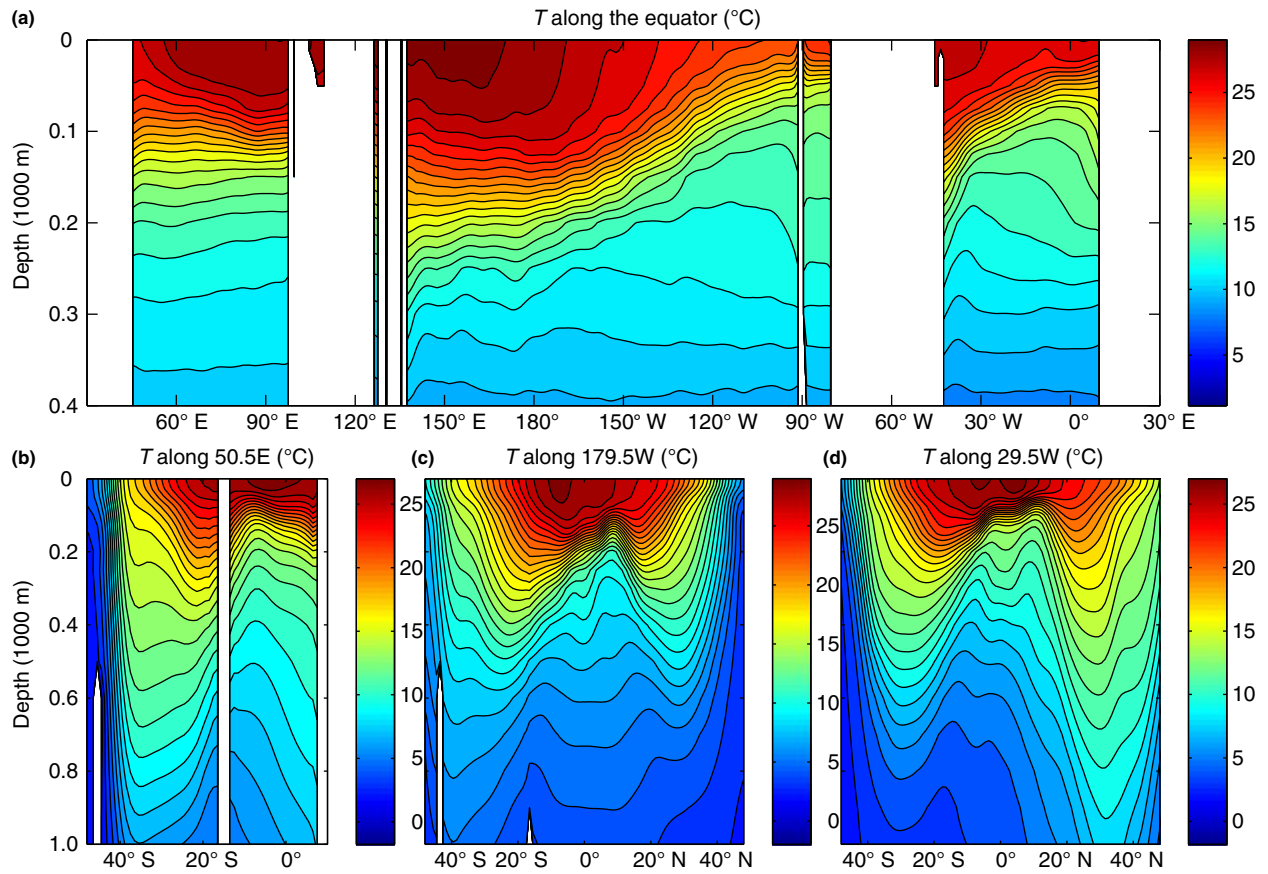


Figure 6 Temperature section along the equator (a) and meridional section through three basins (b,c,d), based on WOA09 data.

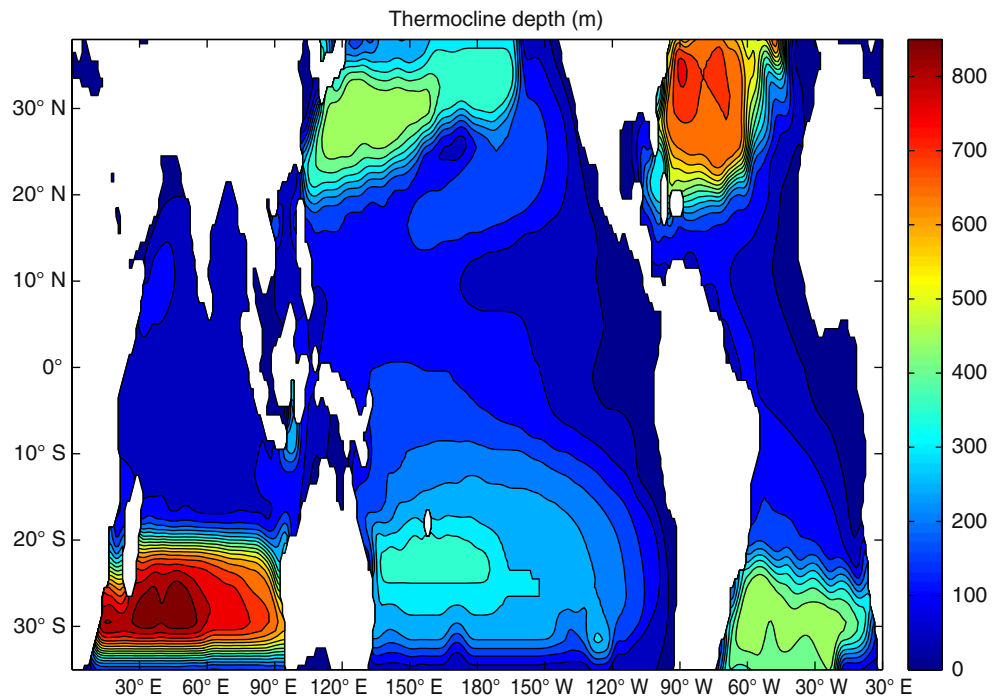


Figure 7 Depth of the main thermocline in the world oceans, based on WOA09 data.

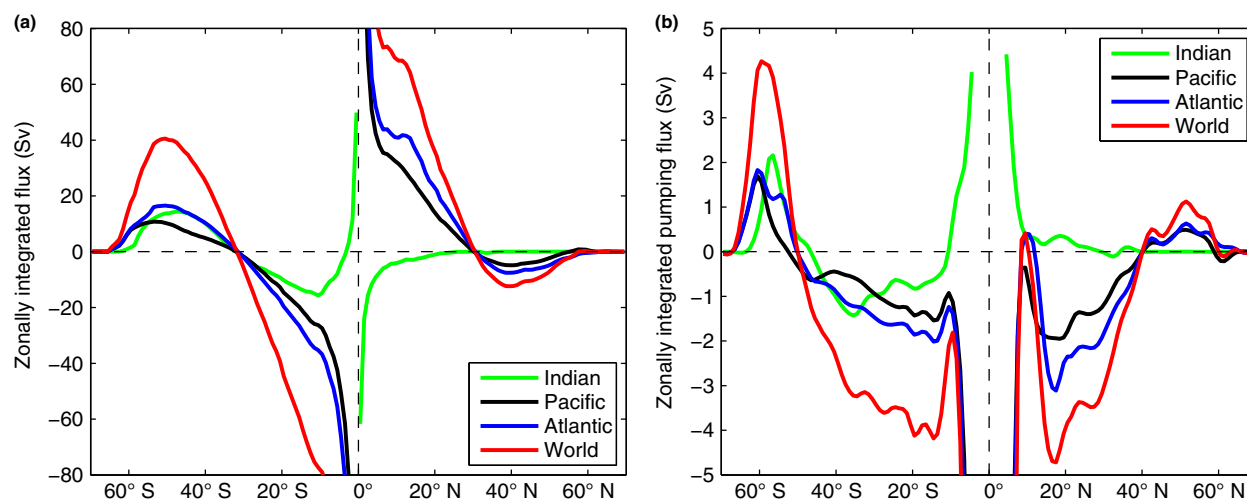


Figure 8 Meridional Ekman transport (a) and Ekman pumping (b) in the world oceans, based on QuikSCAT data.

density. The zonally integrated Ekman transport in the three basins is on the order of 10–20 Sv (Figure 8(a)). As it is inversely proportional to the Coriolis parameter, it becomes unbounded near the equator, indicating that the theory of Ekman layer does not apply to the equatorial band.

Thus, there are large horizontal mass transports in the upper ocean, which can play key roles in transporting heat, fresh-water, and other tracers. Since it is directly proportional to the wind stress, such transports can change in response to variability in the atmospheric circulation and climate.

Due to variation of both wind stress and Coriolis parameter f , Ekman transport varies with geographic location. The convergence (divergence) of Ekman transport leads to the Ekman pumping, $w_{\text{Ekman}} = \text{curl}(\tau/f\rho_0)$. In general, Ekman pumping velocity is quite small, on the order of 10^{-6} m s^{-1} , which is equivalent to 0.08 m per day or 30 m per year. However, in subtropical basin interiors this seemingly small vertical velocity leads to an equatorward geostrophic flow in the subsurface layer, and thus dynamically sets up the gigantic wind-driven circulation in the subtropical basins; while in the subpolar basin, Ekman pumping is upward, and it leads to a poleward geostrophic flow in the subsurface layer and thus the cyclonic subpolar gyres.

Coastal upwelling/downwelling is induced by long-shore wind. If wind blows along the coast, off-shore (or on-shore) Ekman transport must be compensated by upwelling (downwelling) along the coast. Coastal upwelling can bring nutrient-rich water from depth to the surface; thus, high productivity and good fishing grounds along some of the coastlines are closely linked to strong along-shore wind. Since along-shore wind often changes with the season, the strength of coast upwelling also has strong seasonal cycle, and hence the biological productivity. Some of the sites of strongest seasonal cycle of the coastal upwelling are shown in Figure 2. The seasonal cycle of wind stress is strong in the Indian Ocean. In particular, the seasonal cycle of wind stress east of Somalia is very strong, so that coastal upwelling there has a very strong seasonal cycle. In fact, the seasonal cycle of wind is so strong that the direction of the Somali current reverses during the seasonal cycle.

Sverdrup Transport of the Wind-Driven Circulation

The simplest way to describe the wind-driven circulation is to treat the circulation in the upper ocean in terms of a single layer; thus, the wind stress is to be treated as a body force uniformly distributed within this layer. The layer integrated volume flux satisfies the Sverdrup relation:

$$\beta h v = (\partial \tau^y / \partial x - \partial \tau^x / \partial y) / \rho_0. \quad [1]$$

where $\beta = df/dy$, h is the layer thickness, v is the meridional velocity, (τ^x, τ^y) are the wind stress components, and ρ_0 is the constant reference density. The Sverdrup relation is essentially a potential vorticity equation. According to this equation, negative wind stress curl in the subtropical basin drives an equatorward flow in the basin interior. A zonal integration of this relation leads to the Sverdrup streamfunction (or Sverdrup transport). The Sverdrup transport includes contributions due to surface Ekman transport and the geostrophic transport above the thermocline layer. As an example, Figure 9 shows the Sverdrup transport in the Pacific Ocean. There are clearly the subtropical gyres in both the North and South Pacific Oceans, with the maximum streamfunction on the order of 40 Sv. In addition, there is a subpolar gyre at the northern high latitudes and the equatorial circulation system at low latitudes. Note that the Sverdrup relation is valid for the steady state circulation only.

Theories of the Wind-Driven Gyres

The energetics of wind-driven circulation

Wind-driven gyres are the direct result of surface wind forcing. These gigantic circulation systems are sustained by mechanical energy input from the surface winds. Wind energy input to the ocean can be separated into several categories. First and most importantly, the total amount of mechanical energy input to the large-scale surface currents is estimated as 1 TW, and most of such energy is put into the South Ocean and other regions of fast currents, such as the Gulf Stream and Kuroshio. Second, the wind energy input to the Ekman layer is about 3 TW; however,

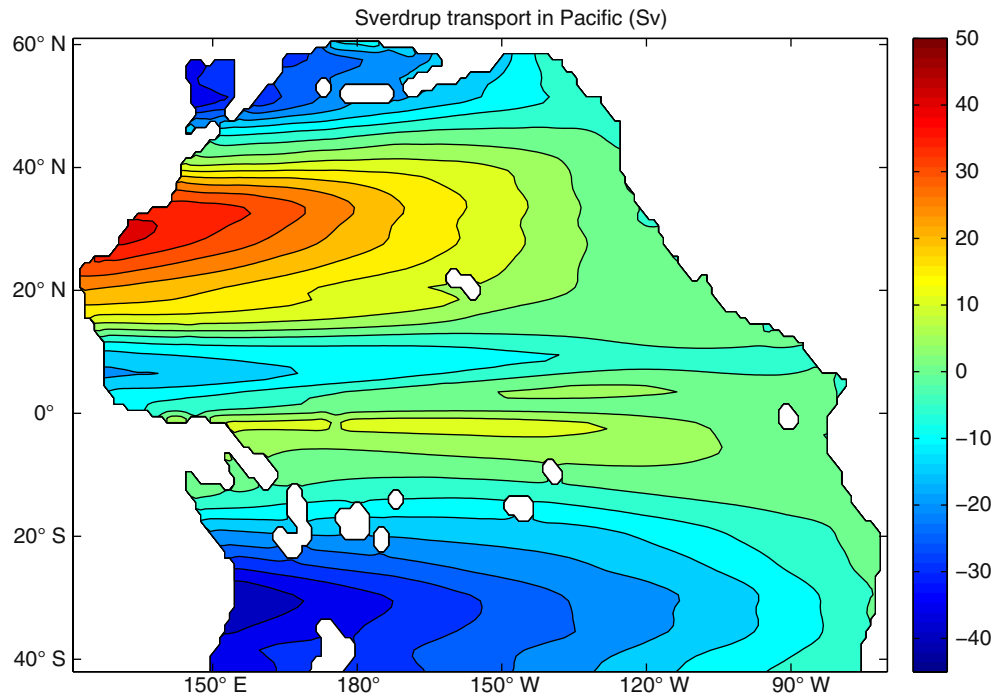


Figure 9 Sverdrup transport in the Pacific Ocean, based on QuikSCAT data.

most of this energy is used to overcome the friction in the Ekman layer. Third, surface waves receive approximately 60 TW from the surface wind; however, this huge amount of mechanical energy is mostly dissipated within the surface wave boundary layer. There is the Langmuir circulation in the upper ocean, which is closely linked to surfaced waves. Surface waves and Langmuir circulation play key roles in regulating the surface layer dynamics and the air-sea interaction. In addition, a small amount of mechanical energy received by surface waves can penetrate through the base of the mixed layer and thus contribute to mixing in the subsurface layer; but, the pathway and this energy remains unclear at this time. Fourth, the sea level atmospheric pressure varies with time; combining with the vertical motions of the sea surface, this leads to a mechanical energy input to the ocean. The exact amount of this energy input remains unclear, and current estimate puts it on the order of 0.01–0.04 TW. However, the effect of sea level pressure change is mostly projected into the barotropic mode in the oceans; thus, its effect on the surface motions is small and may be negligible.

The reduced gravity model

The structure of the wind-driven circulation can be explored in terms of the simple reduced gravity model. The basic idea is to treat the main pycnocline as a step function in density coordinate. Assume that the upper and lower layers have constant density ρ_1 and ρ_2 , and the lower layer is infinitely deep and motionless. A commonly used parameter in such a model is the reduced gravity, defined as $g' = g(\rho_2 - \rho_1)/\rho_0$. The simple formulation of the reduced gravity model allows either analytical or straightforward numerical solutions. For simplicity, solutions shown here are obtained through

numerical integration. The model is started from initial states with a fixed amount of warm water in the upper layer and it is forced by a cosine wind stress.

The first case is for a model on the f -plane, i.e., $f = \text{const.}$ is assumed. As shown in upper panels of [Figure 10](#), both the layer thickness and transport of this solution are symmetric with respect to the E–W and N–S directions. The second case is for a model on a beta-plane (i.e., $f = f_0 + \beta(y - y_0)$), and the corresponding solution is asymmetric with respect to the E–W direction, lower panels in [Figure 10](#). In fact, the current near the western boundary is much stronger than in the interior, and this phenomenon is called the western intensification. The dramatic contrast between a model on the f -plane and a model on the beta-plane was first discovered by Stommel in 1948, who recognized the important meaning of such a difference and made a link between this phenomenon and the Gulf Stream and other strong currents observed in the world oceans.

The reason of the western intensification can be explained in terms of the simple reduced gravity model. For the steady flow in the oceanic interior, the lowest order balance of the circulation can be examined in a beta-plane model as follows. Assuming a steady state and omitting the inertial terms and friction terms, the vorticity equation is reduced to Sverdrup relation [eqn \[1\]](#). Thus, the streamfunction satisfies the Sverdrup relation

$$\psi = (\partial \tau^x / \partial y - \partial \tau^y / \partial x)(x_e - x) / \beta \rho_0, \quad [2]$$

and the layer thickness satisfies the following equation

$$h^2 = h_e^2 + \frac{2f^2}{g'\rho_0\beta} \left[\left(\frac{\tau^x}{f} \right)_y - \left(\frac{\tau^y}{f} \right)_x \right] (x_e - x), \quad [3]$$

where h_e is the layer thickness along the eastern boundary, $\beta = df/dy$ is the meridional gradient of the Coriolis parameter.

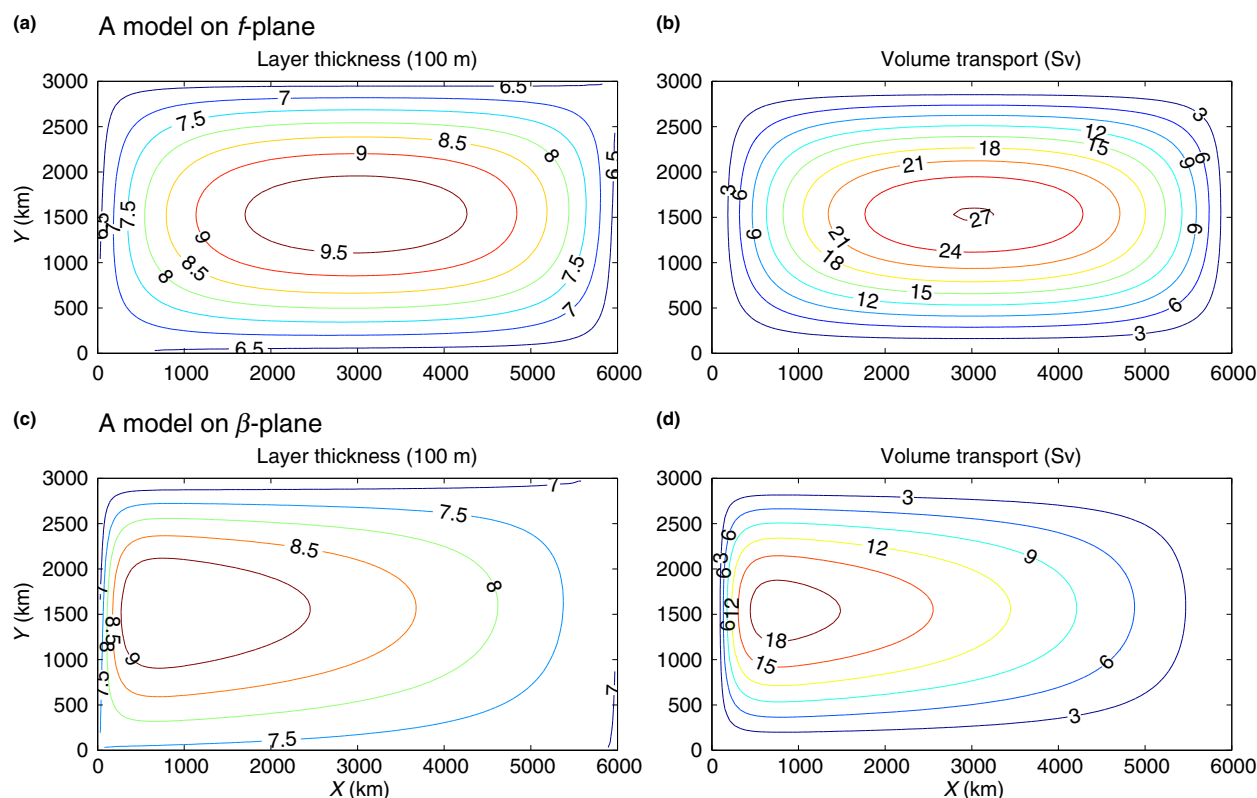


Figure 10 Layer thickness (a, c) and volume transport (b, d) for a model on an f -plane and a model on β -plane.

The advantage of the reduced gravity model is that it provides a clear description for both the volumetric transport and the depth of the main thermocline. Note that although the quasi-geostrophic theory has been widely used for the description of wind-driven gyres in many previous published textbooks and papers, such a description is inaccurate because the wind-driven circulation involves large deviation of stratification, so that the wind-driven gyration is beyond the validity of the basic assumptions made in the quasi-geostrophic theory.

In a single-moving-layer model for a subtropical basin, the equatorward transport in the basin interior must be closed through the addition of either a western/eastern boundary layer, which can transport the mass poleward. Furthermore, in the Northern hemisphere there is a large amount of negative vorticity input from wind stress over the subtropical basin. In a steady state, the basin-integrated vorticity budget must be balanced; thus, there should be a large source of positive vorticity along the lateral boundaries of the basin.

A simple dynamical analysis indicates that only the western boundary can play the role of generating the positive vorticity and thus balancing the vorticity budget for the wind-driven circulation. The existence of western boundary layer manifests in the form of the so-called western intensification.

Stommel first postulated a boundary layer for a reduced gravity model in terms of bottom friction. He assumed that friction is linearly proportional to the horizontal velocity in the moving layer. In more accurate terminology, his bottom friction can be generalized as the interfacial friction, which is assumed to be linearly proportional to the velocity difference

between the upper and lower layers. Such an interfacial friction can be interpreted as a crude parameterization of baroclinic instability. Another possible type of boundary current is the lateral friction model postulated by Munk. In addition, Charney and Morgan postulated the inertial boundary layer theory.

The structure of western boundary currents postulated in these theories can be examined analytically. Since the western boundary current is rather narrow in the cross-stream direction, the corresponding control equations can be simplified by the standard boundary layer technique. In fact, the cross-stream momentum equation can be simplified in terms of the semi-geostrophic approximation, and simple analytical solutions can be obtained, which can be used to illustrate the essential dynamics of the western boundary current.

All these boundary layer theories can be used to close the wind-driven gyres. A close reexamination reveals that western boundary currents observed in the ocean are primarily controlled by the inertial terms, i.e., these boundary currents are essentially inertial western boundary layer in nature. The frictional force is of importance only within a relatively thin sublayer.

The advantage of the reduced gravity model is that it can predict the basin-wide distribution of the main thermocline depth. The structure of the wind-driven circulation in a closed basin on a beta-plane is shown in the lower panels of Figure 10, where the wind-driven circulation in the subtropical gyre consists of the interior part and the Stommel frictional western boundary current. In this example, the horizontal distribution of the thermocline depth can be seen clearly in

a reduced gravity model, as shown in **Figure 10(c)**. For most part of the basin away from the western boundary the solution is described by the interior dynamics discussed above. On the other hand, the strong northward current appears along the western boundary, i.e., the western intensification.

Layer model with outcropping

An advantage of the reduced gravity model is its capability of capturing the strong nonlinearity associated with horizontal variability of stratification, in particular for the case with isopycnal outcropping. Due to the strong wind forcing, surface heat and freshwater fluxes, isopycnals outcrop at high latitudes. A fundamental assumption made in the quasi-geostrophic theory is that variation of the stratification in the horizontal direction is very small; thus, such theory is not suitable for describing the large-scale dynamics associated with layer outcropping. For a model with finite amount of warm water in the upper layer, in part of the basin the layer thickness becomes thinner and thinner as the wind forcing is enhanced. However, when the continuity equation for the layer thickness is transformed into finite difference forms, using the commonly used central difference scheme, the layer thickness may become negative. In order to avoid such situation, the so-called positive-definite scheme should be used.

A typical solution with outcropping is shown in the upper panels of **Figure 11**, where the upper layer vanishes along the outcrop line near the northwest corner of the basin. North of the outcrop line the lower layer outcrops; within the framework of the single-moving-layer model, there is no motion within the outcrop window. A strong internal boundary current is formed along the edge of the outcrop line, which mimics the

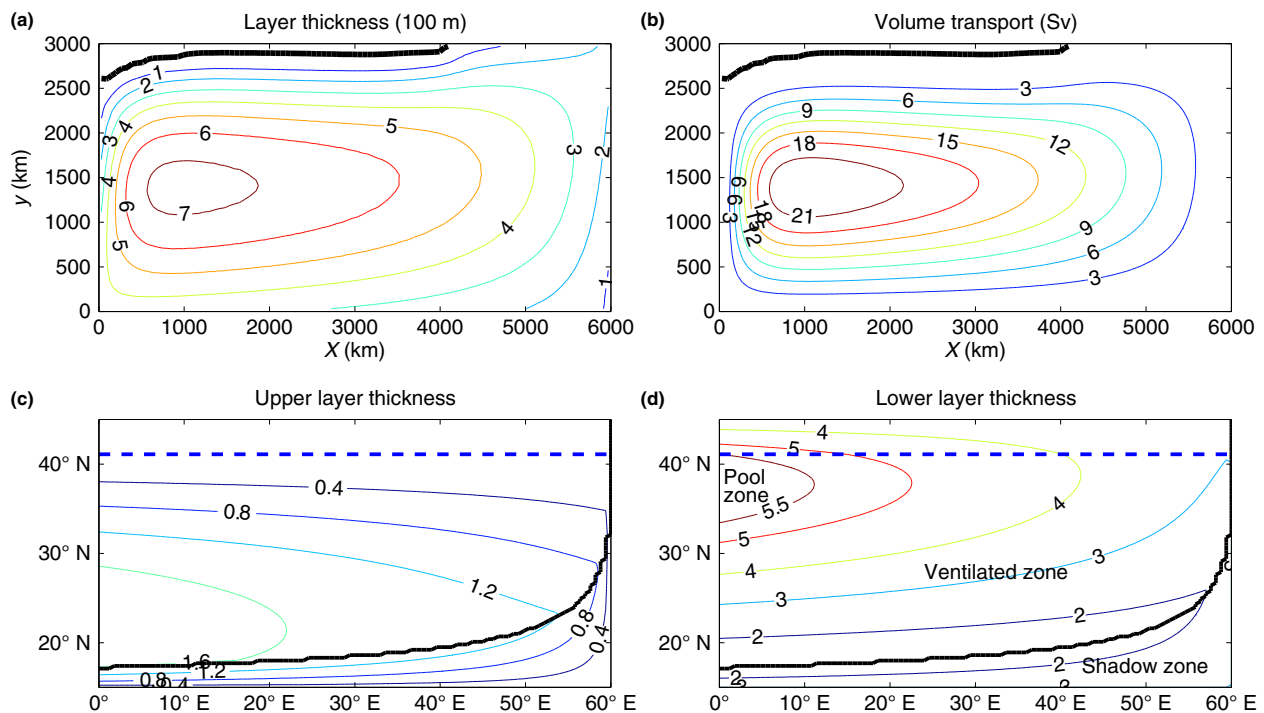
strong western boundary currents after separation, such as the Gulf Stream and Kuroshio. Simulating the case with isopycnal outcropping is a critically important step forward in simulating the wind-driven circulation and the associated density structure in the ocean.

The ventilated thermocline

Although simple reduced gravity models can provide essential information about the wind-driven circulation, such single-moving layer models give no information about the vertical structure of the circulation in the upper ocean.

In pursuing the structure of the wind-driven circulation, the theory of thermocline gradually formed, which is aimed at explaining the three-dimensional structure of the wind-driven circulation. From the beginning of thermocline theory, two paradigms developed independently: the diffusive thermocline by Stommel and Robinson and ideal-fluid thermocline by Welander. The diffusive thermocline theory interprets the main thermocline as an internal thermal boundary layer, and thus emphasizes the critical role of diffusion in forming the main thermocline. On the other hand, the ideal-fluid thermocline theory interprets the main thermocline in terms of ideal-fluid theory, and emphasizes the critical role of adiabatic deformation of the stratification set up by a background thermohaline circulation. Welander actually produced some simple and elegant solutions based on the ideal-fluid thermocline theory.

In the beginning, the theory of diffusive thermocline was pursued by many researchers, apparently due to the seemingly more complete dynamical framework by incorporating the diffusion term. Due to the complicated dynamics involved, exact analytical solutions for the diffusive thermocline were



hard to find. Instead, most studies were aimed at finding different kinds of similarity solutions. The searching for similarity solutions was summed up by Veronis in 1970s. On the other hand, the ideal-fluid thermocline theory was unpopular because it seemed rather incomplete for not including the diffusion terms.

The major deficit of similarity solutions is that they cannot satisfy the essential dynamical constraints, such as the Sverdrup relation; thus, searching for nonsimilarity solutions was a main research frontier. In addition, recent *in-situ* measurements indicate that diffusion in the main thermocline is rather weak; thus, the theory of ideal-fluid thermocline regains its momentum.

One of the conceptual difficulties in pursuing an ideal-fluid theory for the thermocline is as follows. In an ideal-fluid ocean, the interfacial friction must be small and negligible; in such a framework, how are the subsurface layers set in motion? The major breakthroughs took place in early 1980s. Rhines and Young postulated the potential vorticity homogenization theory. They argued that under strong wind forcing, potential vorticity in the subsurface layers is homogenized. As a result, closed potential vorticity contours appear in the subsurface layer, along which water parcels can move freely according to the ideal-fluid thermocline theory. Furthermore, under the assumption of infinitesimal dissipation of potential vorticity, potential vorticity within the close contours are homogenized toward the value along poleward boundary of the circulation. Therefore, a unique solution, which is stable to small perturbations, exists.

Another major breakthrough is the theory of ventilated thermocline. In the stratified ocean, most isopycnals outcrop at high latitudes in winter. The outcropping phenomenon can be clearly seen even in annual mean meridional section of temperature and density, as shown in Figures 5 and 6. Iselin in 1939 first postulated the idea that water masses are formed at

the sea surface in late winter, and subsequently pushed downward through Ekman pumping. His physical insightful idea and Welander's framework of ideal-fluid thermocline were not pursued for a long time. Apparently inspired by the success of the potential vorticity homogenization theory, these two ideas were combined and extended into a beautiful theory of the ventilated thermocline by Luyten, Pedlosky, and Stommel. They formulated the model in terms of three-moving layers plus a deep stagnant layer in the abyss. At lower latitudes, the uppermost layer is exposed to Ekman pumping; however, with the increase of latitudes, upper layers outcrop and lower layers are directly exposed to Ekman pumping. At higher latitudes, Ekman pumping drives the outcropping lower layers in motions, and ventilation and subduction take place.

Ventilation and subduction are the basic elements of the wind-driven circulation in a stratified ocean (Figure 12). The ocean is conceptually separated into four layers; an Ekman layer of the top (not drawn in this sketch), the upper and lower layer below the Ekman layer, and a thick and stagnant layer at the bottom. Wind stress induces the Ekman pumping at the base of the Ekman layer. The upper/lower layer is directly forced when it is exposed to Ekman pumping, as indicated by vertical arrows on the top of the layer. The upper layer outcrops along the zonal outcrop line; thus, north of the outcrop line the lower layer is directly driven by the Ekman pumping. Since the lower layer is exposed to the surface force, we also call this as the ventilation of the lower layer. As a result, there is an anti-cyclonic circulation in this layer, indicated by the blue curved arrows. Poleward of the outcrop line the meridional volume transport in this layer satisfies the Sverdrup constraint.

Equatorward of the outcrop line, the lower layer moves underneath the upper layer, and this is called subduction. Since there is motion in the lower layer north of the outcrop line, there is no reason why it should stop movement after it is subducted; thus, it should continue its anticyclonic movement

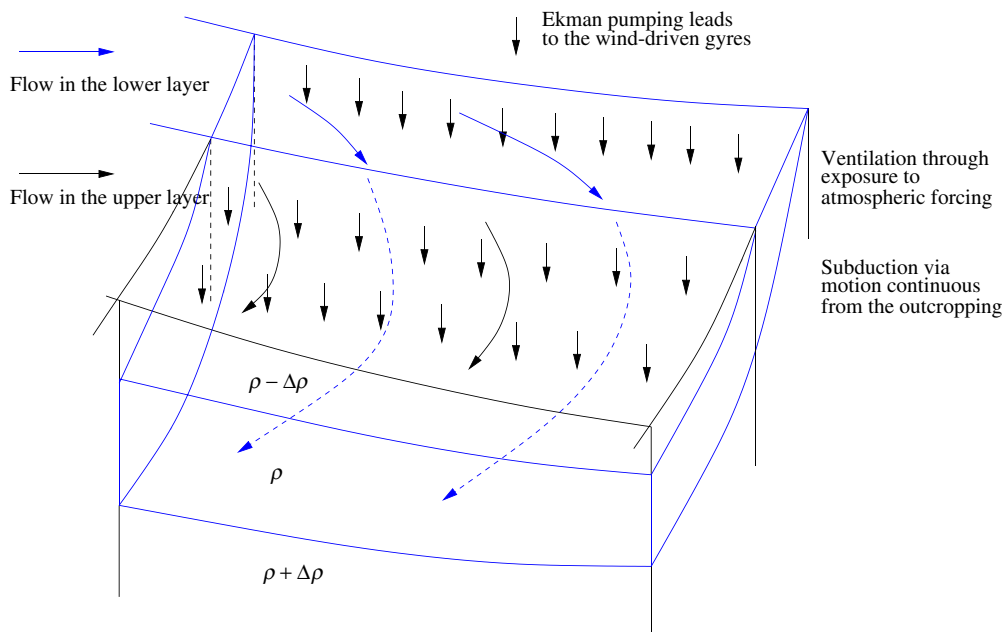


Figure 12 Sketch of ventilation and subduction.

after subduction, as indicated by the dashed lines with arrow. South of the outcrop line, the upper layer is directly forced by Ekman pumping. Because both layers are in motion, in this regime the Sverdrup constraint should apply to the meridional volume transport integrated over the total depth of these two layers. In order to determine the solution in this regime, one more dynamical constraint is required. For ideal-fluid motion, the potential vorticity in the second layer should be conserved when it is not directly forced by Ekman pumping. A simple and elegant solution of this problem is postulated in the ventilated thermocline theory. In the ocean interior, relative vorticity is negligible for large-scale motions; thus, potential vorticity for a two-moving layer model is $q_1 = f/h_1$ and $q_2 = f/h_2$. In fact, keeping the potential vorticity of the second layer at its value when it is subducted along the outcrop line gives the additional dynamical constraint for solving the problem.

The theories of potential vorticity homogenization and ventilated thermocline represent major breakthroughs in understanding the three-dimensional structure of wind-driven circulation. In particular, the discovery of shadow zone, the ventilated zone, and the pool zone brought about completely new physical insight for the wind-driven circulation in the upper ocean.

As an example, a two-layer ventilated thermocline model is shown in the lower panels of [Figure 11](#) for a northern hemisphere model ocean, in which the zonal dashed lines indicate the outcropping line of the upper layer. The model is forced by a simple sinusoidal Ekman pumping field, which works on the layer exposed to the upper surface forcing, i.e., the lower layer north of the outcrop line and the upper layer south of the outcrop line. North of the outcrop line, the lower layer is exposed to the Ekman pumping, and an anticyclonic circulation can be clearly seen north of the outcrop line. South of the outcrop line, the lower layer is subducted; however, it continues its south-westward movement as shown in [Figure 11\(d\)](#). The upper layer has zero thickness north of the outcrop line; but, south of the outcrop line it is directly exposed to Ekman pumping.

As shown in [Figure 11\(d\)](#), geostrophic flow in the lower layer after subduction can be separated into three dynamical zones: the pool zone near the western boundary, the shadow zone near the eastern boundary, and the ventilated zone in the middle. An important conceptual breakthrough in the ventilated thermocline theory is the existence of a shadow zone within the subsurface layer near the eastern boundary. A strong kinematic condition along the eastern boundary is the no-flow-penetration condition along the eastern boundary. Thus, the subsurface layer should have a constant thickness h_2 along the eastern boundary. If the eastern boundary were a streamline, then the corresponding potential vorticity f/h_2 should be constant along the eastern boundary. Because the Coriolis parameter declines equatorward, f/h_2 cannot be constant. Therefore, the eastern boundary cannot be a streamline for the subsurface layer; instead, there should be a shadow zone near the eastern boundary where the subsurface layers are stagnant. From hydrographic data, the existence of shadow zone can be inferred from the appearance of extremely low oxygen levels at the depth of 1000 m and near the eastern boundary at low latitudes.

There are two other dynamical zones in [Figure 11\(d\)](#). Potential vorticity in the ventilated zone is sent up along the

outcrop line when the lower layer is subducted. Pool zone is defined by streamlines emerging from the western boundary. As such, potential vorticity in the pool zone is not set up by wind-driven forcing within the basin. Instead, it is set up by either potential vorticity homogenization or other dynamics.

The original Sverdrup relation applies for a single-moving-layer model only. For a model with multiple moving layers, it is extended to an integral constraint for the meridional volume flux for all the moving layers. The ventilated thermocline theory was generalized to a theory for the continuously stratified ocean by Huang. Such a model provides density and current structure in the ocean, which are comparable with observations.

Note that the Sverdrup relation applies for the steady circulation only. There are two important issues related to the application of this relation. First, the steady solution of wind-driven circulation in a basin is established after the wave adjustment. The volume flux over the whole depth of the ocean is established after the barotropic Rossby waves passing through. Since barotropic Rossby waves move quite fast, the timescale for the barotropic circulation to be established is in the order of 7–10 days for a tropical basin.

However, the baroclinic Rossby waves move relatively slowly, with the speed in the order of 0.1 m s^{-1} . Thus, for the midlatitudes of the North Atlantic or the North Pacific Ocean, the steady circulation of the first baroclinic mode takes about 10–20 years to be established. In general, the time-dependent solution of the wind-driven circulation can be calculated by the integrating the time delayed Ekman pumping rate.

Furthermore, the contribution due to eddies is not included in the Sverdrup relation. In the ocean, eddies give rise to nonlinear contribution to the meridional volume flux. As a result, meridional volume flux across the zonal section can deviate from the Sverdrup relation.

Combination of Surface and Deep Currents

Currents in the North Atlantic Ocean

Although surface currents are mostly controlled by wind stress, the thermohaline circulation, which occupies the whole depth of the ocean also manifests in the upper ocean. The circulation in the North Atlantic Ocean is a good example. As shown in [Figure 13](#), there is a warm current crossing the equator. This current moves northward and joins the warm water in the Gulf Stream, which exists as an internal boundary current separating the subtropical gyre to the south and the subpolar gyre to the north. It is important to note that the wind-driven gyre in the North Atlantic Ocean consists of the linear Sverdrup interior and the recirculations. The recirculation regimes are located roughly where the surface expression of the western boundary current emerges from the western boundary region by separation from the coast, where the nonlinearity of the circulation is not negligible. As a result, the Sverdrup dynamics does not apply and eddy activity associated with the nonlinear inertial terms become large and dominating. Instead of the laminar fluid like current, at any given time the Gulf Stream encompasses many large-amplitude eddies.

Furthermore, the total volumetric flux in the Gulf Stream recirculation regime is in the order of 150 Sv, much larger than

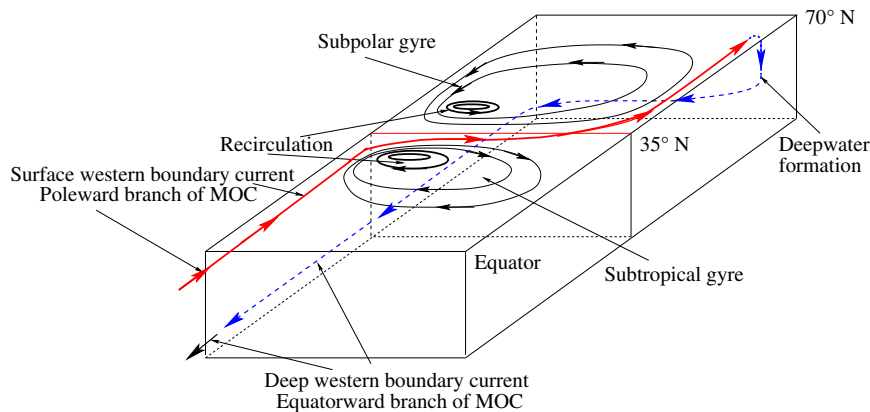


Figure 13 Combination of surface wind-driven circulation and deep circulation in the North Atlantic Ocean.

the value predicted from the linear Sverdrup theory. This large volume transport includes contribution of 30–35 Sv due to the linear wind-driven circulation, about 15–20 Sv from the returning flow of the deep thermohaline circulation, and approximately 90–100 Sv due to the nonlinear dynamics (in forms of the southern recirculation and the north recirculation), **Figure 13**. The return flow of the thermohaline circulation continues its poleward movement and appears in the form of the North Atlantic Current, which moves to the high latitude part of the North Atlantic basin and eventually feeds the deepwater formation.

The Antarctic Circumpolar Current

There is a strong circum-earth current system in the South Ocean, and it is called the ACC. There is no meridional boundary in the Southern Ocean. As a result, there is no zonal pressure gradient force to maintain any meridional geostrophic flow; thus, the classical Sverdrup dynamics does not apply to ACC. The formation of ACC depends on many dynamical factors, including wind stress, wind stress curl, the shape of coastline and bottom topography, the surface thermohaline forcing, and, most importantly, the contribution of mesoscale eddies. In particular, mesoscale eddies play capital roles in setting the structure of ACC. ACC plays an important role in regulating the global oceanic circulation and climate. However, the complete description of the ACC dynamics requires an in-depth discussion and it is beyond the scope of this article.

Mesoscale Eddies

Most classical theories of wind-driven circulation treat the circulation in terms of laminar fluids, with the roles of eddies neglected. The framework of three-dimensional structure of gyre-scale wind-driven circulation was completed in 1980s, represented by the multilayer ventilated thermocline theory and its extension to the case of continuously stratified ocean. These theories provided the lowest order structure of the wind-driven circulation and laid down the foundation for the further development of oceanic circulation.

With the advance in technology in observation, theory, and numerical models, the situation has changed rapidly over the

past decade. In fact, physical oceanography is now entering the eddy resolving era. By definition, two kinds of eddies are now the focus of research. The mesoscale eddies have horizontal scales from 10 to 500 km and vertical scales from tens to hundreds of meters, and the submesoscale eddies have horizontal dimensions on the order of 1–10 km and vertical scales on the order of tens of meters or smaller.

The ocean is a turbulent environment, and eddy motions are one of the fundamental aspects of oceanic circulation. In fact, it is estimated that the total amount of eddy kinetic energy is about 100 times larger than that of the mean flow. The roles of mesoscale and submesoscale eddies in the oceanic circulation and climate remain to be explored.

It is expected that with the great technical advances in satellite observation and global observation program like ARGO, eddy study is pushed forward with a great speed. Studies of these eddies, including observations, theory, laboratory experiments, and parameterization in numerical models, will be the most productive research frontiers for the next 10–20 years.

See also: Air Sea Interactions: Surface Waves. Boundary Layer (Atmospheric) and Air Pollution: Ocean Mixed Layer. Oceanographic Topics: Thermohaline Circulation. Satellites and Satellite Remote Sensing: Surface Wind and Stress.

Further Reading

- Chelton, D.B., Schlax, M.G., Samelson, R.M., 2011. Global observations of nonlinear mesoscale eddies. *Progress in Oceanography* 91, 167–216.
- Ekman, V.W., 1905. On the influence of the earth's rotation on ocean currents. *Arkiv för Matematik, Astronomi och fysik* 2, 1–52.
- Gill, A.E., Green, J.S.A., Simmons, A.J., 1974. Energy partition in the large-scale ocean circulation and the production of mid-ocean eddies. *Deep Sea Research* 21, 499–528.
- Huang, R.X., 2010. *Ocean Circulation, Wind-Driven and Thermohaline Processes*. Cambridge Press.
- Luyten, J., Pedlosky, J., Stommel, H.M., 1983. The ventilated thermocline. *Journal of Physical Oceanography* 13, 292–309.
- Olbers, D., Borowski, D., Völker, C., et al., 2004. The dynamical balance, transport and circulation of the Antarctic Circumpolar Current. *Antarctic Science* 16 (4), 439–470.
- Pedlosky, J., 1996. *Ocean Circulation Theory*. Springer-Verlag, Heidelberg.

- Pedlosky, J., 2006. A history of thermocline theory. In: Jochum, M., Murtugudde, R. (Eds.), *Physical Oceanography Developments since 1950*. Springer, New York, pp. 139–152.
- Price, J., Weller, R.A., Schudlich, R.R., 1987. Wind-driven ocean currents and Ekman transport. *Science* 238, 1534–1538.
- Qiu, B., 2001. Kuroshio and Oyashio currents. In: *Encyclopedia of Ocean Sciences*. Academic Press, pp. 1413–1425.
- Rhines, P.B., Young, W.R., 1982. A theory of the wind-driven circulation. I. Mid-ocean gyres. *Journal of Marine Research* 40 (Suppl.), 559–596.
- Richardson, P., 1980. Benjamin Franklin and Timothy Folger's first printed chart of the Gulf Stream. *Science* 207 (4431), 643–645. <http://dx.doi.org/10.1126/science.207.4431.643>.
- Stommel, H., 1948. The western intensification of wind-driven ocean currents. *Transactions of the American Geophysical Union* 29, 202–206.
- Veronis, G., 1969. On theoretical models of the thermocline circulation. *Deep Sea Research* 16 (Suppl), 301–323.
- Weller, R.A., Bigorre, S.P., Lord, J., Ware, J.D., Edson, J.B., 2012. A surface mooring for air–sea interaction research in the Gulf Stream. Part I: Mooring design and instrumentation. *Journal of Atmospheric and Oceanic Technology* 29, 1363–1376. <http://dx.doi.org/10.1175/JTECH-D-12-00060.1>.

Thermohaline Circulation

RX Huang, Woods Hole Oceanographic Institution, Woods Hole, MA, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by J R Toggweiler and R M Key, volume 4, pp 1549–1555, © 2003, Elsevier Ltd.

Synopsis

Thermohaline circulation is a major component of the oceanic general circulation. Thermohaline circulation includes water mass formation at the surface, its transportation through the ocean interior, and the return flow to the source region. Thermohaline circulation can have multiple states and catastrophic transitions from one state to the other, and such changes are closely linked to climate changes on the Earth. External mechanical energy is crucial in maintaining the circulation against internal dissipation; thus, mixing parameterization based on external mechanical energy constraint is the key for numerical simulation.

Introduction

What is thermohaline circulation? Literally, thermohaline circulation is the component of oceanic circulation that is directly linked to the density difference induced by temperature and salinity difference. In many previously published textbooks, thermohaline circulation is defined as the circulation driven by density difference induced by surface thermal and haline forcing. According to the new paradigm, thermohaline circulation is a circulation driven by mechanical stirring, which transports mass, heat, freshwater, and other properties in the meridional and zonal directions. Mechanical stirring is supported by external sources of mechanical energy from wind stress and tidal dissipation. In addition, surface heat and freshwater fluxes are necessary for setting up the circulation.

The Cold Deepwater

At low latitudes, surface water is rather warm. The discovery of cold water in the deep ocean at low latitudes by Henry Ellis, captain of Halifax, in 1751 came as a great surprise. Deepwater at low latitudes is much colder than the corresponding lowest temperature at the sea surface in winter; thus, the source of such water mass must be formed at high latitudes where cold water mass can be formed in winter. Tracing cold water formed at high latitudes and its movement to low latitudes led to the theories of thermohaline circulation in the world oceans.

The bottom of the world oceans is covered by cold water originating from a very few narrow sites at high latitudes, where severe winter conditions produce the coldest water in the oceans. Since oxygen solubility is high at low temperature, high oxygen content also indicates the recently formed water mass. Hence, both low temperature and high oxygen concentration can be used as the indication of deepwater formation.

Over the past century, extensive observation data over the world's oceans have been accumulated. Since seawater is slightly compressible, temperature of a water parcel should be slightly increased during a downward adiabatic movement. Hence, the concept of potential temperature is widely used in the description of deep circulation. Potential temperature is defined as the temperature of a water parcel, if it is moved adiabatically (and without changing its salinity) to the sea surface. Potential temperature distribution on the bottom of

the world's oceans is shown in [Figure 1](#). From this figure, the following characteristics of bottom water temperature are readily seen.

Cold water on the bottom is formed around Antarctica, primarily in the Weddell Sea and Ross Sea. The cold water mass formed around the edge of the Antarctic continent that sinks to the bottom of the world oceans is called Antarctic Bottom Water (AABW). From these source regions, bottom water is carried northward and eastward by currents and eddies.

Cold bottom water spreads northward in each basin, mostly in the form of deep western boundary current. Bottom water temperature in the South Atlantic Ocean is the coldest among all basins. In the South Atlantic Ocean, only the Brazil Basin receives AABW directly. Due to topographic setting, the Angola Basin is close to the cold bottom water source from the south. In fact, AABW's effluence to this basin is through a narrow gap near the equator, where relatively cold water moves eastward and finally reaches the Angola Basin from the northern opening passage.

At the northern end of the North Atlantic basin there are different sources of relatively cold water, which originate from the Nordic Seas ([Figure 1](#)). These dense water masses overflow the narrow passages in the Greenland–Scotland Ridge and become the source waters of the deepest component of the North Atlantic Deepwater (NADW).

Over the past century, theories of thermohaline circulation have been developed in order to explain the general circulation related to the formation and spreading of bottom/deepwater and the connection with water in the upper ocean where surface thermohaline forcing prevails. The goal in this article is to explain the physical phenomena and the dynamical theories of the thermohaline circulation in the world oceans.

Water Masses Signature Seen from a Meridional Section

Although thermohaline circulation is often connected to the bottom/deepwater formation and transport, it encompasses many other types of water masses. As an example, the characters of such water masses can be seen clearly from a meridional section ([Figure 2](#)).

From the potential temperature section, the AABW is marked by the low temperature water originated from the edge of Antarctica. The existence of other water masses can be also seen from the salinity section. The AABW is marked by

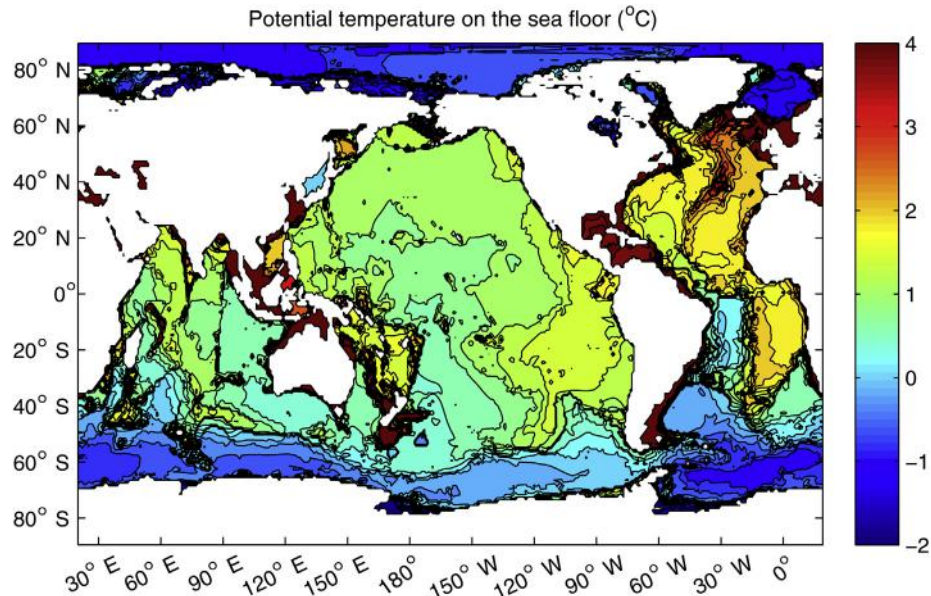


Figure 1 Potential temperature on the sea floor, based on WOA09 data.

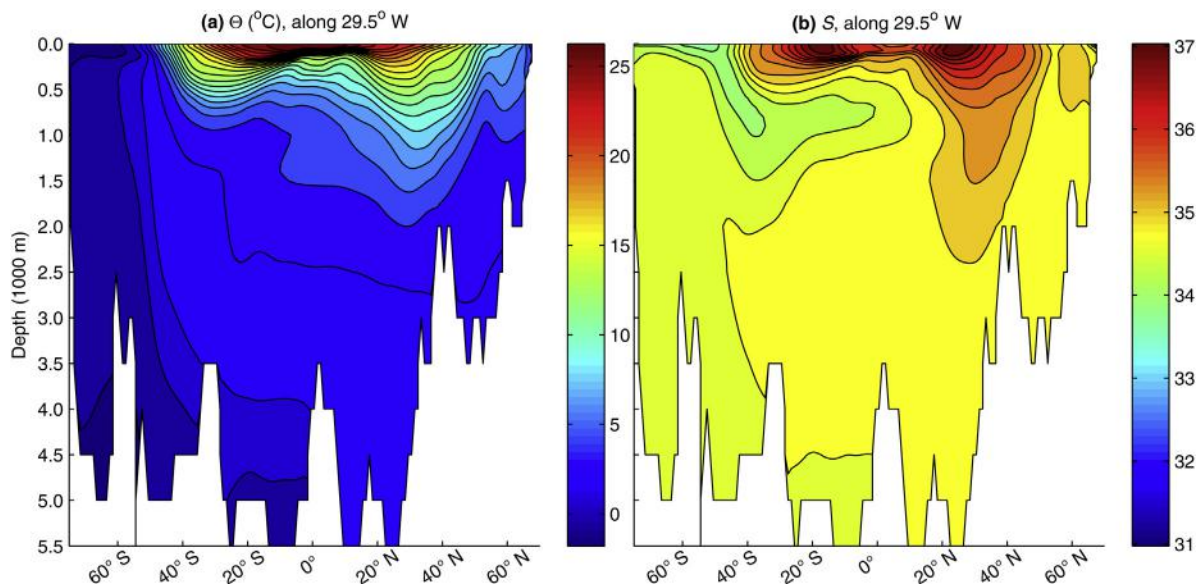


Figure 2 (a) Potential temperature and (b) salinity along 29.5° W, based on WOA09 data.

relatively low salinity. At a shallower level in the southern hemisphere, the Antarctic Intermediate Water (AAIW) marked by a tongue of salinity minimum extends from the surface to the depth of 1 km. Below this tongue of salinity minimum and above the relatively fresh layer of AABW on the bottom is the relatively salty water associated with the NADW, which extends from the North Atlantic Ocean all the way across 40° S and at a depth of 2–3 km. This high salinity tongue is mostly due to the strong evaporation at the surface and sinking of the resulting salty water at middle latitudes in the North Atlantic Ocean.

Surface Buoyancy Forcing

Surface Heat Flux

Surface heat flux is a major forcing for the thermohaline circulation; it is the sum of the four terms: the incoming short-waves solar radiation, the outgoing sensible heat flux, latent heat flux, and long-wave radiation. The net air–sea heat flux map is shown in [Figure 3](#). There is a strong heat flux into the ocean along the equatorial band, in particular the cold tongues in both the Pacific and Atlantic Oceans. Both the Kuroshio and Gulf Stream are the major sites of heat loss to the atmosphere.

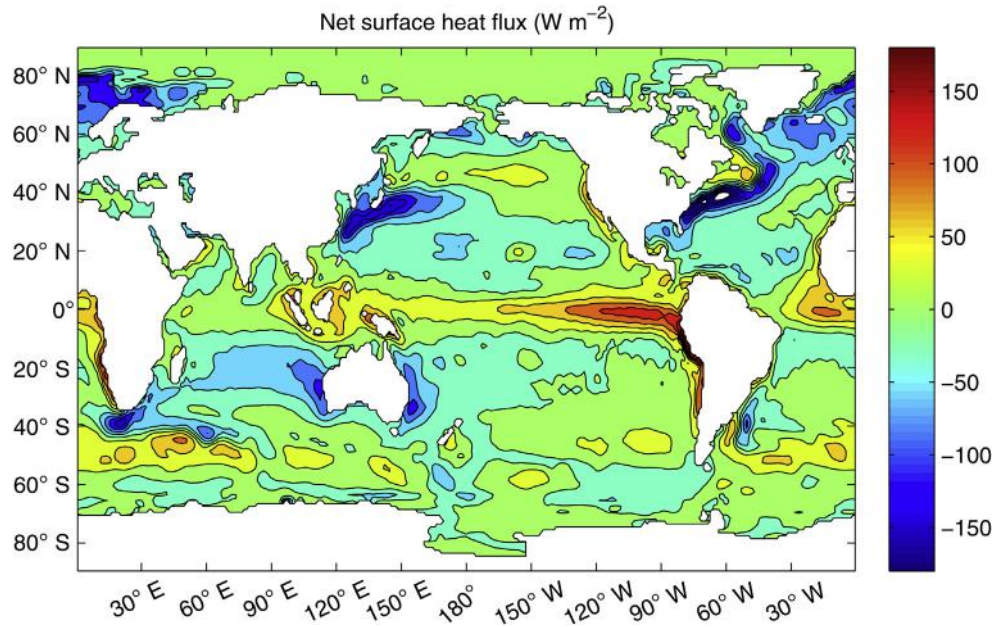


Figure 3 Net annual-mean surface heat flux, based on NCEP–NCAR (1979–2010) data.

In addition, the western coast of South America appears as a heat absorption band, which is due to the low sea surface temperature associated with strong coastal upwelling. The high latitude Atlantic Ocean appears as a heat sink, which is related to the strong meridional overturning (Figure 3).

The net heat flux is asymmetric with respect to the equator. In fact, in the Indian Ocean and South Atlantic Ocean, the net heat flux is downward, i.e., the ocean gains heat, instead of losing heat. The net air–sea heat flux distribution implies that there is a meridional heat transport in the ocean; otherwise, the underlying ocean would continuously cool or warm, depending on the sign of the heat flux. In addition, there is a strong zonal transport of heat in the ocean. The zonal heat flux is intimately linked to the oceanic currents, including both the wind-driven circulation and thermohaline circulation.

Surface Freshwater Flux

The surface freshwater flux includes evaporation and precipitation, plus river runoff. Evaporation provides the moisture in atmosphere, and it brings heat from low-latitude ocean to the atmosphere, where water vapor is carried poleward. Water vapor carries a large amount of latent heat, and this is one of the vital mechanisms of poleward heat transport in the climate system. Water vapor in the atmosphere eventually condenses and releases the latent heat content, returning to the oceans or land as precipitation. Freshwater flux through the air–sea interface plays vital roles in regulating the hydrological cycle in the ocean. In particular, freshwater flux is the key ingredient in controlling salinity distribution in the oceans. Hence, freshwater flux is one of the key players in regulating the thermohaline circulation.

Strong evaporation appears in the subtropics and the western boundary current system in both hemispheres. In particular, both the Gulf Stream and Kuroshio System are the

regions with the maximal evaporation rate in the global oceans. The oceans receive the returning water as precipitation from the atmosphere, plus river runoff. The major regions of strong precipitation include the equatorial Pacific Ocean and the South Pacific Convergence Zone, which extends southeastward in the South Pacific Ocean, with the annual-mean precipitation rate of more than 4 m per year. In addition, there are large regions of strong precipitation in the equatorial Indian Ocean and Atlantic Ocean (Figure 4).

The difference in evaporation and precipitation is what really affects the haline circulation in the oceans. The pattern of the net freshwater flux across the air–sea interface is dominated by two features: the strong precipitation bands and the strong and relatively narrow regions of strong evaporation. Overall, the equatorial band is dominated by a net gain of freshwater. In particular, there is strong net freshwater flux, on the order of 3 m per year, into the western equatorial Pacific Ocean and the eastern equatorial Indian Ocean. On the other hand, the eastern parts of the five subtropical basins in both hemispheres appear as deserts in the oceans, where the annual-mean net freshwater loss to the atmosphere is on the order of 1–1.5 m per year.

Surface Buoyancy Flux

Surface thermohaline forcing is often converted into the density flux, which is defined as

$$F_\rho = -\rho_0(\alpha F_T - \beta F_S)$$

where $\alpha = -\frac{1}{\rho_0} \frac{\partial \rho}{\partial T} \Big|_{P,S}$, $\beta = \frac{1}{\rho_0} \frac{\partial \rho}{\partial S} \Big|_{P,T}$ are the thermal expansion coefficient and saline contraction coefficient, ρ is seawater density, ρ_0 is the constant reference density; $F_T = Q/\rho_0 c_p$, Q is the net heat flux into the ocean, c_p is the heat capacity of water; $F_S = (E - P)S/(1 - 0.001S)$, $E - P$ is evaporation minus precipitation, S is sea surface salinity in unit of part per thousand.

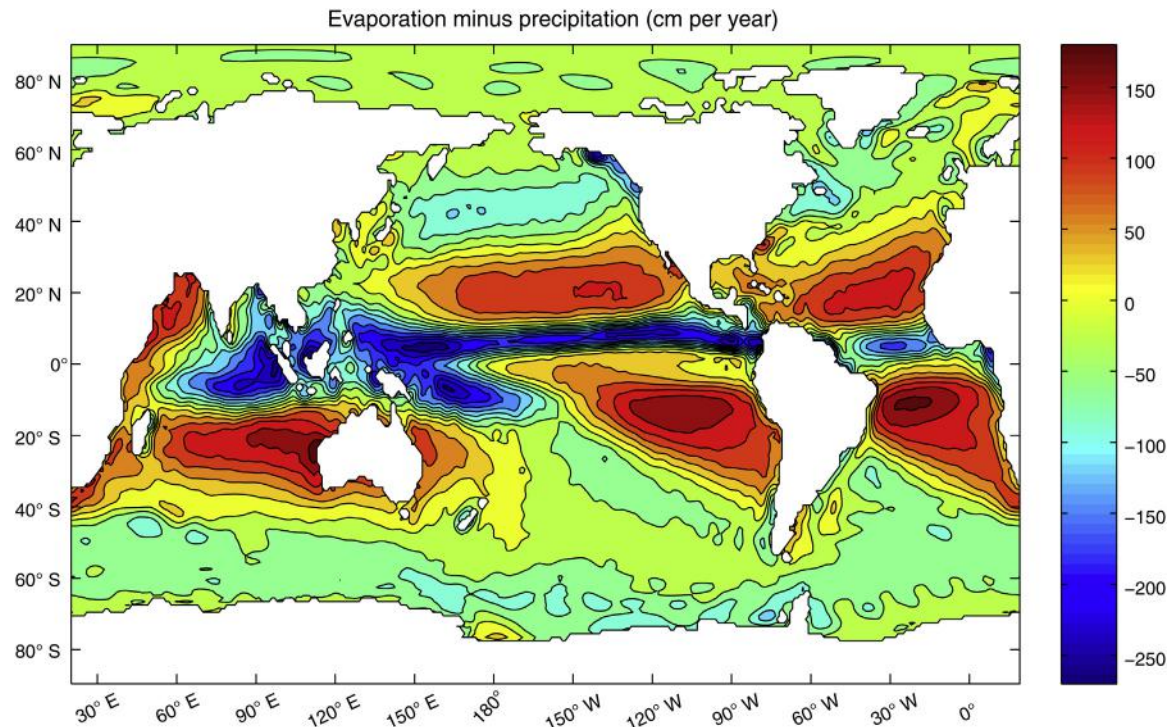


Figure 4 Annual-mean evaporation minus precipitation, based on NCEP–NCAR data.

Density flux is primarily controlled by the net air–sea heat flux; thus, its distribution shares features similar to the net air–sea heat flux, with opposite signs. Near the equator, the density flux is negative, indicating that surface heating reduces water density; this is true especially in the cold tongue of the eastern equatorial Pacific Ocean. In both the Gulf Stream and Kuroshio, density flux is positive, reflecting the strong air–sea heat loss to the atmosphere.

Overall, density flux over the most part of the North Atlantic Ocean is positive, but it is negative in the northern North Pacific Ocean (Figure 5). This difference reflects the fact that North Atlantic Ocean is much warmer than the North Pacific Ocean. As a result, there is strong evaporation in the North Atlantic Ocean, even at high latitudes. Such a difference is primarily due to the existence of a strong thermohaline overturning circulation in the Atlantic Ocean. The regions of positive density flux are closely linked to the site of water mass formation, including mode water formation and deepwater formation. This map does not include the contribution associated with sea-ice formation; thus, the site of bottom water formation in the Weddell Sea cannot be clearly identified from this map.

Water Mass Formation

Bottom Water Formation

Under the current climate condition, bottom water is formed in the Weddell Sea, labeled as number 6 station in Figure 6. Due to the cold and dry air from the Antarctic continent, sea ice is formed along the edge of the continent, in particular in the Weddell gyre. Salt rejection during the sea-ice formation

increases the salinity of the underlying water parcels. This heavier water mass flows down the Antarctic continental slope, mixing with the ambient water on its way.

There are many other sites in the world’s oceans where dense water masses are formed through local air–sea interaction. Although some of these water masses may have density higher than the source water from the Weddell Sea, they cannot sink to the bottom of the world oceans. This is due to a very special thermodynamic property of seawater, the so-called thermobaricity. The most important point in connection with the thermobaricity is that the cold and relatively freshwater is more compressible than the warm and relatively saline water. Assume that there are two water masses with the same density at the sea surface, except that one is warm and salty and the other is cold and fresher. Because of thermobaricity, the first water parcel can be heavier than the second one at greater depth.

The source water from the Weddell Sea is cold and relatively fresh. Due to the difference in compressibility, it can be further compressed and thus form the densest water in the bottom of the world oceans. On the other hand, although the NADW and other types of water masses formed at the sea surface may be heavier at the sea surface, they are less compressible, and, as such, cannot sink to the bottom of the world oceans.

Deepwater Formation

Under the current climate condition, the most prominent type of deepwater in the oceans of the world is the NADW. There are two different ways of this deepwater formation. In Type 1, deepwater can be formed in the open ocean, typically in the

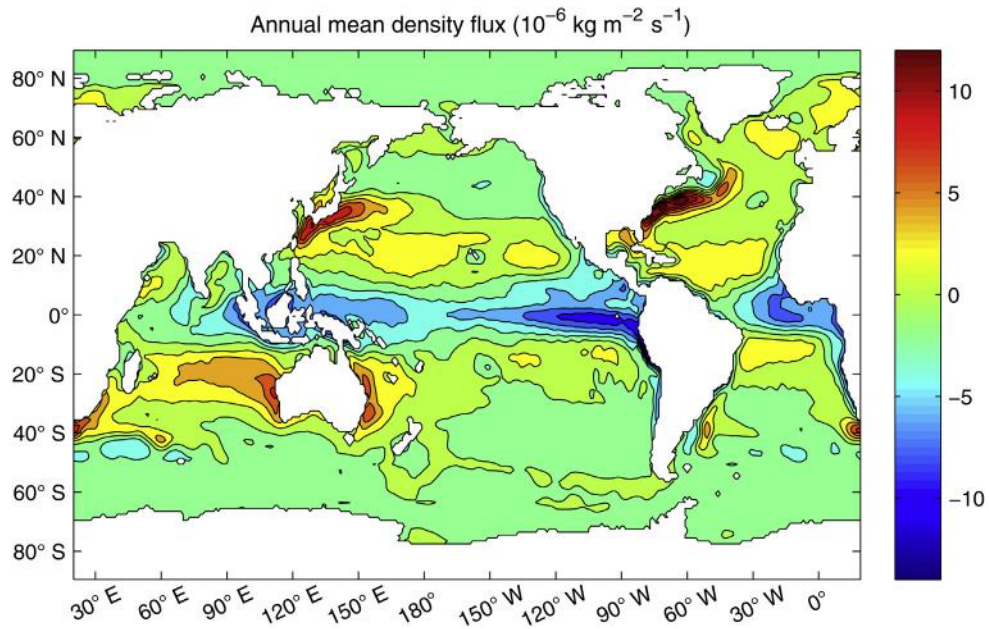


Figure 5 Annual-mean density flux at the sea surface, based on NCEP–NCAR data.

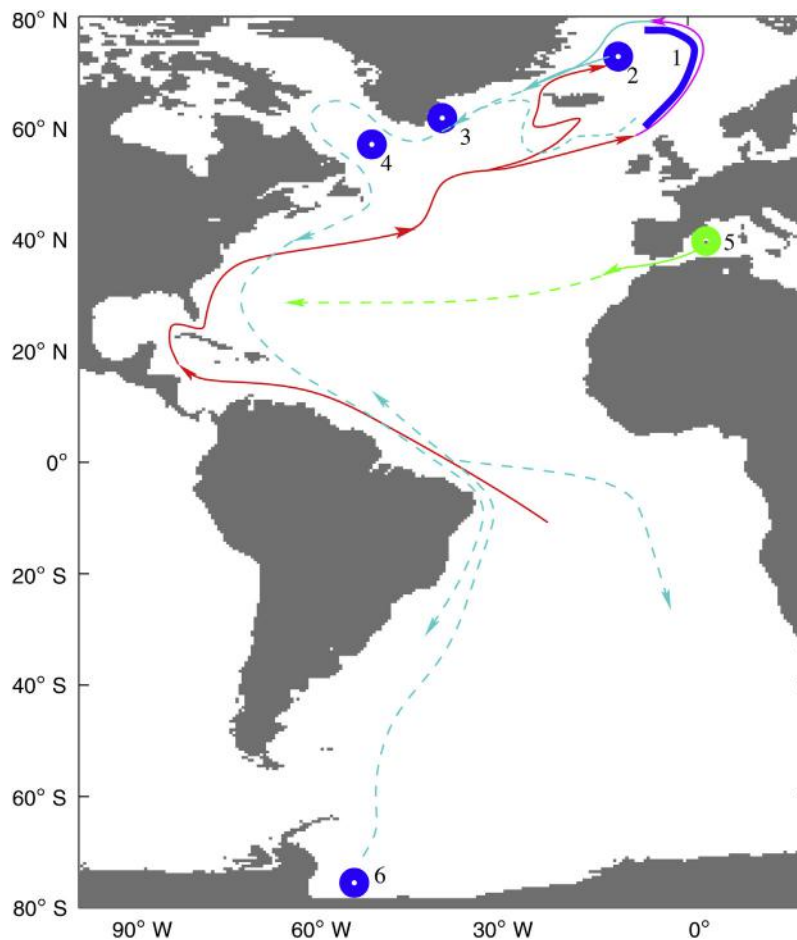


Figure 6 Deep/bottom water formation in the Atlantic sector and the related currents. Heavy blue curve and circles indicate sites of bottom/deep-water formation; heavy green circle and curve indicate Mediterranean water formation and its pathway; red curves with arrows depict warm surface currents; light blue curves with arrows indicate deep boundary currents.

center of a cyclonic gyre, where the stratification is weak. In late winter, strong surface cooling can induce chimneylike density structures in the middle of the gyre, where dense water is formed and sinks to a deeper part of the water column. In Type 2, water advected by oceanic currents can gradually increase in density due to cooling and salt rejection along the path of the current. The dense water so formed along the path of the currents sinks to depth and thus constitutes the source of deepwater.

NADW formed under the current climate conditions can be further separated into an intermediate and deep component. The intense air–sea forcing in the Nordic Seas, together with the oceanic preconditioning, provides a suitable environment for the origin of the deep component. Both Type 1 and Type 2 formations occur here. In particular, the inflowing subtropical-origin Atlantic water is densified as it circumnavigates the Nordic Seas within the boundary current system (Type 2 formation, denoted by the heavy blue curve in [Figure 6](#)), ultimately exiting Denmark Strait in the East Greenland Current (light blue solid curve in [Figure 6](#)). In addition, open-ocean convection occurs in the Greenland and Iceland Seas (Type 1 formation), which is believed to supply dense water to both the Faroe Bank and Denmark Strait overflows. A newly discovered current, the North Icelandic Jet, carries the transformed water from the Iceland Sea into Denmark Strait (depicted by number 2 station in [Figure 6](#) and the light blue line that emanates from it).

The intermediate component of NADW has two primary sources. On the east side of the southern tip of Greenland, the steep topography of the coast leads to strong local winds. This, in conjunction with cold wintertime air temperatures, drives middepth convection and water mass formation (labeled by the number 3 station in [Figure 6](#)). In addition, wintertime cooling in the western Labrador Sea leads to the formation of Labrador Sea Water, which is believed to be the major source of the intermediate component of NADW (number 4 station in [Figure 6](#)).

NADW is advected southward by the deep western boundary current in the North Atlantic and eventually reaches the southern hemisphere, as depicted by the slight blue curves with arrows in [Figure 6](#).

An additional source of deepwater formation is in the Mediterranean Sea, where wintertime convection creates a water mass known as Mediterranean water (MW, labeled station 5 in [Figure 6](#)). MW overflows the Strait of Gibraltar into the eastern North Atlantic Ocean. Due to the high salinity at the sea surface, MW is heavier than the source waters of the AABW and NADW. However, because it is warm and salty, MW is less compressible than these deep/bottom water sources. As a result, MW cannot sink to the bottom of the world oceans; instead, it equilibrates near a depth of 1000–1500 m in the North Atlantic Ocean (depicted by the green curve with arrows in [Figure 6](#)).

Mode Water Formation

Thermohaline circulation is a technical term, which should include many types of circulation related to density difference induced by thermohaline forcing; thus, it includes many different types of water mass formation. In addition to

the bottom and deepwater masses formation discussed above, there are other types of water masses formed at much shallow depth. Water mass formation is distributed non-uniformly in the temperature–salinity space. In fact, much of water mass formation is concentrated around some rather narrow regions in the space of temperature and salinity. In general, these water masses are formed and circulated at relatively shallow depth, and they are often called mode water, such as Subtropical Mode Water, Subpolar Mode Water, and AAIW.

Mode water formation takes place primarily at the air–sea interface. This process is often called subduction – which takes place in later winter at the base of the mixed layer. This is a combination of downward pumping due to the wind-driven circulation in the upper ocean and the thermohaline process in the upper ocean. After its formation, mode water is carried out by the three-dimensional thermohaline circulation, modifying its properties on the path gradually through turbulent and eddy mixing. Mode water eventually loses its identity through obduction, which is a process opposite to subduction. Mode water formation and its variability are closely related to the climate condition, in particular the air–sea interface exchanges of heat and freshwater flux.

The typical depth range of mode water circulation is within the top kilometers of the ocean. As a result, mode water formation, including their site and annual rate, is closely linked to climate change on interannual and decadal timescales.

The major features of the meridional circulation in the Atlantic Ocean are shown in [Figure 7](#). Note that circulation in the ocean is truly three-dimensional in nature. Therefore, when reading such a sketch, one should not forget that there are many other important components of circulation excluded from the sketch. Most importantly, wind-driven gyres primarily appear in the form of horizontal gyration, which is excluded from such a meridional sketch.

The most important components of the thermohaline circulation in the Atlantic section include two cells. First, there is a deep cell related to the AABW. Water associated with this cell originates from the edge of Antarctic and moves northward along the bottom of the ocean, crossing the equator and moving all the way to the lower latitudes in the North Atlantic Ocean. Along this path, it is gradually mixed with the relatively light water of NADW through diapycnal mixing driven by tides and internal waves.

Second, a slightly shallower overturning cell is linked to NADW. This cell starts from high latitudes in the North Atlantic Ocean. Because the NADW is lighter than the AABW, it cannot sink to the sea floor; instead, it sinks to the depth range of 4 km and spreads southward. NADW and AABW encounter each other mostly in the southern hemisphere. The strong westerlies in the southern hemisphere drive the strongest upwelling system in the world oceans. Through this upwelling system the NADW is moved upward all the way to the sea surface, including its water mass properties modified by diapycnal mixing along its pathway.

On much shallow levels, the thermohaline circulation appears in the form of mode water ventilation and subduction and the subsequent meridional overturning circulation, as depicted by red and pink color arrows in [Figure 7](#).

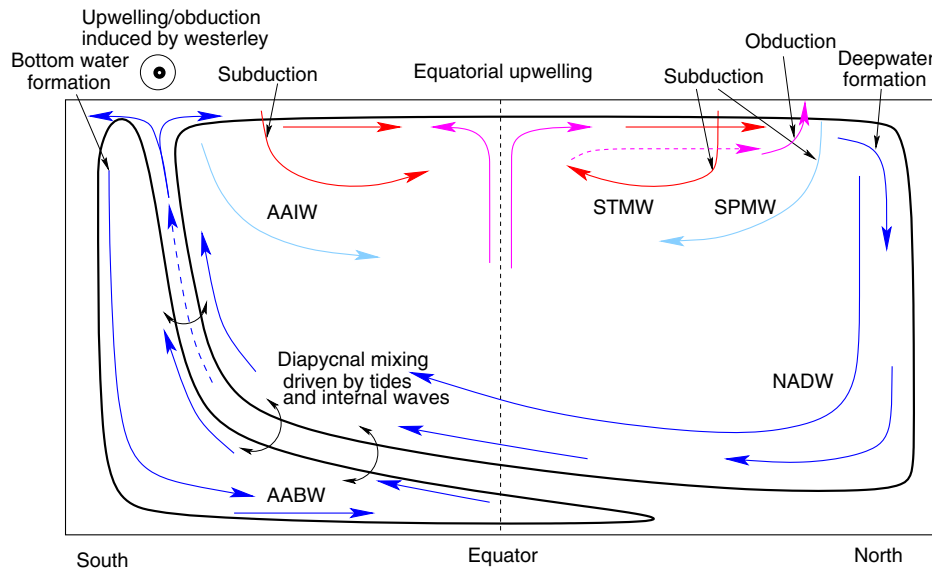


Figure 7 Sketch of water mass formation and meridional overturning cells in the Atlantic Ocean.

Theory of Deep Circulation

After bottom and deepwaters are formed in the source regions, they move through the world's oceans in the form of deep circulation. In general, there are three key components of the deep circulation.

First, most sources of bottom/deepwater mass are formed in marginal seas, which are separated from the open ocean by narrow channels. As these source waters flow through the narrow channels, they overflow some relatively shallow ridges. When deepwater overflows shallow ridge, the current speed can exceed the corresponding signal speed of internal gravity waves. Thus, a phenomenon called the internal hydraulic jump takes place, which is also called deepwater fall. The volumetric flux of such deepwater fall can be quite large, on the order of a few Sv ($1 \text{ Sv} = 10^6 \text{ m}^3 \text{ s}^{-1}$), and the corresponding elevation drop of a few hundred meters. In comparison, the water falls on land are much smaller. For example, the well-known Niagara Falls has a volumetric flux of $3000 \text{ m}^3 \text{ s}^{-1}$, and the corresponding elevation drop is about 56 m.

After leaving the source region, bottom/deepwater circulates the world oceans. The basic theory of deep circulation was postulated by Stommel and Arons in a series of semi-narian papers published in the 1960s. They assumed that the ocean had a flat bottom with no topography and there was a uniform upwelling in the basin interior. For a model on a beta-plane and in a steady state, the linear vorticity equation predicts that water from the source region must flow through the world oceans in the form of deep western boundary currents. On the other hand, in the basin interior the linear potential vorticity balance requires that there is a universal poleward flow.

Their theoretical prediction of the existence of deep western boundary currents was quickly confirmed by field observations, and this was called one of the most remarkable advances in dynamical oceanography. Over the past half century, many deep boundary currents were discovered through well-planned

hydrographic surveys in the world oceans. The existence of deep boundary currents was considered as a solid proof of the theory. On the other hand, the poleward flow in the oceanic interior predicted by the theory has not been tested through observations, and this is due to the fact that the poleward velocity predicted by the theory is so small that it was rather difficult to measure.

Stommel and Arons theory of deep circulation dominated the study of deep circulation for more than 30 years. With the great advance in the ocean measurement technique, theory, and numerical models, their theory was reexamined. The conclusions of their theory are a simple logic consequence of the basic assumptions. However, there is no reason a priori, why the upwelling in the ocean interior should be uniform. There are steep bottom topographic features and the circulation is not necessary in a steady state.

Recent studies based on *in-situ* observations and numerical simulations with high resolution reveal the following issues. There are indeed deep western boundary currents as predicted by the Stommel and Arons theory; however, due to eddies and topographic features, such boundary currents can be different from the simple laminar boundary current predicted by their theory. In the ocean interior, instead of the universal poleward flow, the deep basin can be occupied by zonal alternative deep jets with strong eddies. In simple words, the deep circulation is much more complicated and rich in dynamical features than the simple deep circulation pattern predicted by the Stommel and Arons theory.

Freshwater and Haline Circulation

Although the role of thermal forcing is quite familiar for most people, the role of freshwater forcing and the haline circulation are rather unfamiliar. Freshwater flux through the air-sea interface appears in the form of evaporation and precipitation, plus the river runoff. In addition, salt rejection during sea-ice

formation and freshwater resulting from ice melting are also important sources of freshwater flux in the upper ocean.

Evaporation and precipitation often appear as intermittent processes, but the thermohaline circulation in the oceans is much more persistent; thus, the idea that such infrequent freshwater flux can be linked to the thermohaline circulation seems nonintuitive. However, thermohaline process has a timescale much longer than the typical timescale of the intermittence related to surface evaporation and precipitation. Consequently, over the much longer timescale, the effect of surface evaporation and precipitation appears as well organized surface forcing.

In fact, Hough in 1897 made the first attempt of theorizing the oceanic current. In a study of tidal circulation, he postulated a simple model for the evaporation- and precipitation-driven flow on a sphere. His model is rather simple and does not take friction into consideration. In 1933, Goldsbrough made a great improvement in the freshwater-driven circulation. To find a closed solution driven by freshwater flux he assumed a rather special evaporation and precipitation pattern, in which the net evaporation and precipitation is zero along any fixed latitude of a basin. His solution was further improved by Stommel, who postulated to close the freshwater-driven circulation by adding the western boundary current. Thereby, the circulation driven by freshwater flux is often called the Goldsbrough–Stommel solution.

Although Goldsbrough and Stommel laid the foundation of the freshwater-driven circulation, their theory was mostly neglected over a long time. The primary reason of such situation is due to the fact that the strength of the circulation driven by freshwater flux alone is on the order 1 Sv, which is at least one order of magnitude smaller than the circulation observed in the oceans.

A close reexamination reveals the reason of such a problem. The seawater is loaded with salt, and there is strong mixing of salt (or freshwater) in the oceans. Goldsbrough–Stommel theory was formed for the freshwater only without considering the role of salt and mixing. If the ocean has no salt, or if there is no mixing of salt in the ocean, the only circulation driven by surface freshwater flux should be that predicted by the Goldsbrough–Stommel theory. In reality, due to strong mixing in the ocean the freshwater induced circulation has two components: a barotropic component as predicted by the Goldsbrough and Stommel and a baroclinic component which is at least one order of magnitude stronger than the barotropic component.

Therefore, for the thermohaline circulation the air–sea interface freshwater flux plays roles as important as the surface heat flux. From a more general point of view, freshwater flux through the air–sea interface is part of the hydrological cycle in the climate system. Hydrological cycle can appear in quite different forms, such as moisture and clouds in the atmosphere, precipitation and evaporation through the air–sea interface, and water transported through the oceans.

Among many roles played by the hydrological cycle, poleward heat transport is a key component of the climate machinery. Poleward heat flux in the climate system is carried out by the general circulation in the atmosphere and oceans. Traditionally, poleward heat flux is separated into two parts. In the atmosphere, the poleward heat flux consists of the heat flux

carried by the dry atmosphere and the heat flux carried out by the moisture. The rest of the poleward heat flux is carried out by the oceanic general circulation.

However, the moisture circulation in the climate system is essentially an atmosphere–ocean coupled mode. At low and mid-latitudes, moisture is carried by the atmospheric circulation to high latitudes, where it is transformed by precipitation (rain and snow) and enters into the oceans. The equatorward return flow of the global hydrological cycle is mostly through the oceanic currents, or the meridional overturning circulation in the ocean. It is clear that without the oceanic thermohaline circulation the moisture transport in the atmosphere would be incomplete. Thus, one can also separate the poleward heat flux into three components: the sensible heat flux carried by the dry atmosphere, the sensible heat flux carried by the oceanic current, plus a third component carried by the moisture transport in the atmosphere and the return haline circulation in the oceans. Comparing with the traditional way of classifying the poleward heat flux, this three-component view may provide more accurate description of the relevant physics.

Sandstrom Theorem and Framework for Thermohaline Circulation

Sandstrom Theorem

Atmospheric general circulation can be interpreted as a heat engine, and the great success in atmospheric science led people to believe that the equator–pole temperature difference in the ocean can drive a strong thermohaline circulation. In fact, the theory of thermohaline circulation discussed in many previous publications were based on the idea that ocean is a heat engine.

Starting from late the 1990s, people began to realize that thermohaline circulation may work in a way fundamentally different from the general circulation in the atmosphere. In fact, more than 100 years ago, Sandstrom postulated a framework for thermally forced circulation in the environment of gravitational field. He argued that the thermally forced circulation of a system can be idealized as a Carnot cycle (Figure 8). When the system is heated from above and cooled from below, no mechanical energy can be generated from the cycle. As such, the system cannot be self-sustained against the friction, so there is no circulation. On the other hand, if the system is heated from below and cooled from above, net mechanical energy can be generated through each cycle. Thus, system can be self-sustained against the internal friction. Sandstrom also carried out laboratory experiments and found that when the heating source was below the cooling source, there was strong circulation; however, when the heating source was above the cooling source, there was no circulation.

The atmosphere receives the solar radiation, which is primarily absorbed by the ocean and land at the lower boundary first and then reflected back, i.e., it is a system heated from below. The outer space works as the cooling source at the upper boundary; hence, the atmosphere is cooled from above. According to Sandstrom theorem, the atmosphere is a heat engine.

The ocean is mostly heated and cooled from the upper surface. Due to thermal expansion, sea level at low latitudes, where heating takes place, is about 1 m higher than the sea level

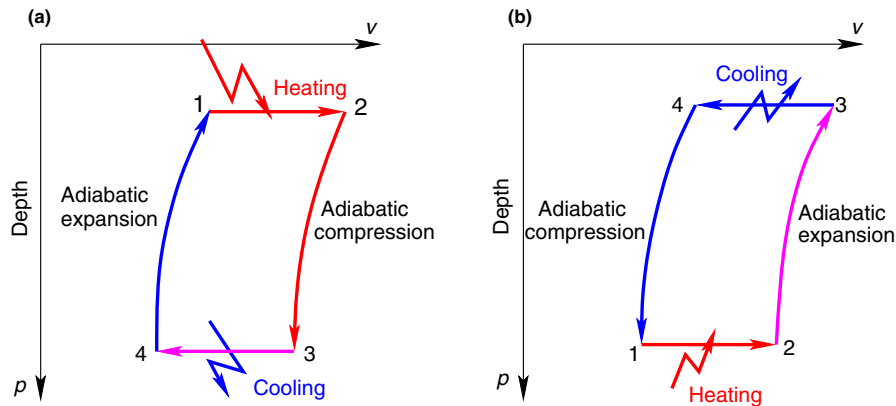


Figure 8 Two settings of thermal engine illustrating the Sandstrom theorem. (a) When heating is above cooling the circulation cannot be self-sustained. (b) When heating is below cooling the circulation can be self-sustained.

at high latitudes where cooling takes place. Cooling and heating from the same geopotential level is called horizontal differential heating. The behavior of a system under horizontal differential heating is currently a research frontier.

According to the Sandstrom theorem, there should be no convectively driven circulation in the ocean. There is, however, strong overturning circulation in the world's oceans, which seems in contradiction to the Sandstrom theorem. According to the new theoretical framework, the existence of strong overturning circulation in the ocean can be explained as a circulation driven by external source of mechanical energy.

School of Pushing and School of Pulling

The framework of thermohaline circulation can be classified into different schools.

School of pushing

Deepwater formation pushes deep currents and thus drives the thermohaline circulation.

This is the classical school for the thermohaline circulation, in which thermohaline circulation is thought to be driven by surface thermohaline forcing, in particular the surface cooling/heating. Surface cooling produces dense water that sinks to a great depth. The high latitude ocean is filled up with cold and dense water from surface to bottom. Combining with the warm and light water in the upper ocean at low latitudes, this creates a pressure force in the abyssal ocean, driving cold bottom water toward low latitudes and thus pushing the meridional circulation (Figure 9(a)). This school is based on the connection between meridional overturning circulation and pressure difference induced by surface thermohaline forcing and deep-water formation.

For example, in a two-box model Stommel assumes that the circulation rate is proportional to the north-south pressure difference. Furthermore, the proportional constant relating to the pressure difference and overturning rate is assumed to be invariant under different climate conditions.

The school of pushing views the thermohaline circulation as buoyancy controlled. Many people believe that the strong circulation induced by sudden cooling may well be a strong

support for this theory. However, a close examination reveals that the strong circulation after the onset of sudden cooling is due to the release of a large amount of gravitational potential energy during the convective adjustment. In fact, the total amount of mechanical energy for the mean state is greatly reduced during such sudden cooling. Furthermore, the circulation would decline gradually, if there were no continuous supply of external source of mechanical energy.

School of pulling

Pulling by deep mixing

Deep mixing removes cold water from the abyss and maintains the stratification and circulation.

The major problem of the school of pushing is as follows. If there were no external source of mechanical energy to sustain mixing, cold water would pile up in the ocean; as a result, the solution is eventually reduced to a very weak circulation. To maintain a sizable circulation, the cold water in the abyss should be removed. Deep mixing sustained by tidal mixing can transform cold water into warm water in the deep ocean, creating room for newly formed deepwater and thus pulling the thermohaline circulation.

Pulling by wind stress

The Southern westerlies pull cold water from the deep ocean and thus maintains the global thermohaline circulation.

Under the current topographic and climate setting, a strong Ekman upwelling around the latitude band of 50–60° S exists, which is closely related to the Antarctic Circumpolar Current. Due to this strong upwelling, NADW is pulled up to the upper ocean. In the upper ocean, wind stress continues to input mechanical energy through Ekman layer and surface waves, and water properties are gradually modified in the surface mixed layer on the way northward.

These two ways of pulling can work together as follows. Wind-driven upwelling and mixing in the upper ocean can build up the major parts of the thermohaline circulation and water mass transformation in the world's oceans, while the remaining parts of the water mass transformation in them are relatively small and can be accomplished by external mechanical energy from tidal mixing (Figure 9(b)).

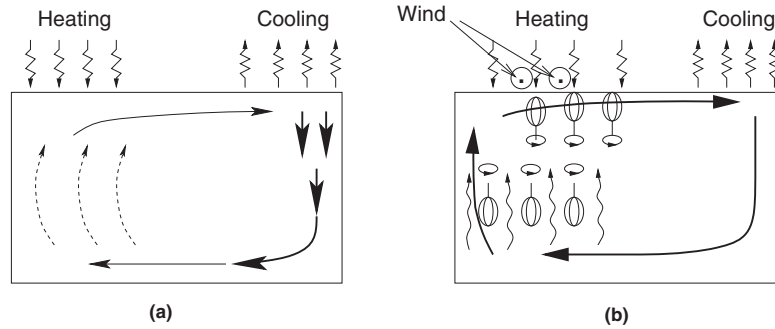


Figure 9 Two schools of the thermohaline circulation in the world oceans. (a) Pushing by deepwater formation. (b) Pulling by deep mixing and wind stress.

Three Paradigms

Thermohaline circulation in the ocean is a very complex system; thus, it can be studied from quite different angles.

Buoyancy paradigm

Thermohaline circulation is driven by horizontal buoyancy difference or deepwater formation; thus, energetically the ocean is driven by surface thermohaline forcing. The oceans receive huge amount of thermal energy in the form of short-wave radiation (60.8 PW, $1 \text{ PW} = 10^{15} \text{ W}$), and sending back heat flux in forms of latent heat (35 PW), long wave radiation (21.2), and sensible heat (4.7 PW), **Figure 10(a)**.

In simple box models the circulation rate is assumed to be linearly proportional to the north-south density difference. In many oceanic general circulation models and other numerical models used in climate study, the diapycnal diffusivity is treated as a fixed parameter of the model. The common practice in choosing such parameter is by tuning the model to reproduce an overturning rate that matches the present-day observations.

The salient points of this paradigm are: the diffusivity is considered as fixed parameters intrinsic to the model. The same diffusivity is used in the model to simulate circulation under different climate conditions, such as the circulation during the last glacial maximum or circulation under the global warming scenario for the next 100 years.

Mechanical energy paradigm

Buoyancy forcing in the upper surface cannot generate mechanical energy required to overcome friction/dissipation associated with circulation. In order to maintain the circulation against friction/dissipation, external sources of mechanical energy are needed. The major external sources of mechanical energy sustaining the oceanic circulation include wind stress energy (64 TW) and tidal dissipation (3.5 TW) (**Figure 10(a)**). Accordingly, the thermohaline circulation is a mechanical conveyor transporting mass, heat, and freshwater fluxes, and it is directly driven by the external source of mechanical energy. Diapycnal mixing is subject to the mechanical energy constraint.

Since the sources of mechanical energy change with the atmospheric wind conditions and tidal dissipation, diapycnal diffusivity should change with climate and tides. However, the circulation rate is not necessarily linearly proportional to the amount of mechanical energy sustaining the circulation. It is important to emphasize that the mechanical energy paradigm is relatively young, and the relevant theories and parameterization will take a long time and much effort to be developed. With recent interest in this new paradigm, it will become more mature and competent in the near future.

Entropy paradigm

Entropy balance is one of the fundamental thermodynamic laws governing the universe, including the oceanic circulation. In a quasi-steady state, the total amount of energy fluxes

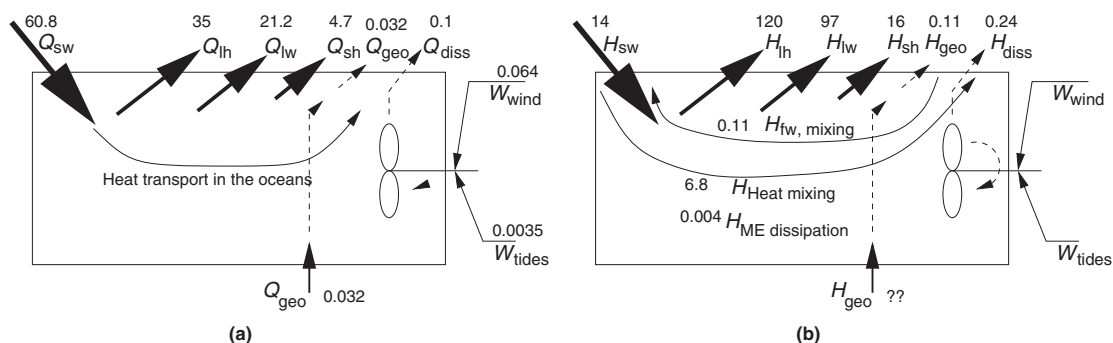


Figure 10 Energy and entropy fluxes in the world oceans. (a) Energy fluxes in the world oceans (in 10^{15} W). (b) Entropy fluxes in the world oceans (in 10^{12} W K^{-1}). (Question marks indicate highly uncertain value.) Updated from Huang, R.X., 2010. *Ocean Circulation, Wind-Driven and Thermohaline Processes*. Cambridge Press.

entering and leaving the system should be equal (Figure 10(a)). The balance of entropy flux is totally different. The oceanic circulation is a dissipated system, so that it generates entropy internally. As a result, the circulation is associated with the negative entropy fluxes. Thus, entropy is not conserved, i.e., the entropy flux carried by short-wave radiation into the system ($H_{sw} = 14 \times 10^{12} \text{ W K}^{-1}$) is smaller than the entropy flux leaving the system ($H_{lh} = 120 \times 10^{12} \text{ W K}^{-1}$ for latent heat; $H_{lw} = 97 \times 10^{12} \text{ W K}^{-1}$ for long wave radiation; $H_{sh} = 16 \times 10^{12} \text{ W K}^{-1}$ for sensible heat) (Figure 10(b)). The removal of entropy is implicitly part of the air-sea heat exchange, including the incoming short-wave solar radiation of low entropy and the high entropy fluxes associated with outgoing heat fluxes. This net entropy flux plays a vitally important role in maintaining the ocean circulation against the accumulation of entropy associated with heat/salt diffusion in the ocean.

There are additional sources of entropy. Heat transported from regions of high temperature to regions of low temperature is one of the important sources of internal entropy production. The poleward heat transport is in the order of 1.5–2 PW. The entropy production due to oceanic heat transport is estimated as $6.8 \times 10^{12} \text{ W K}^{-1}$. The rate of entropy production due to freshwater mixing in the oceans is $0.11 \times 10^{12} \text{ W K}^{-1}$. These estimates may serve as the theoretical lower bound of entropy production due to heat transport and hydrological processes in the world's oceans; a detailed balance and its geographic distribution of these terms remain unclear.

According to the new paradigm of thermohaline circulation, mechanical energy involved in thermohaline circulation comes from the external sources of wind/tides. Thus, the total mechanical energy dissipation rate is equal to the rate of input from the external sources of mechanical energy. As a result, for the world's oceans the total entropy production due to momentum dissipation in the world oceans is estimated at $H_{diss} = 0.24 \text{ TW K}^{-1}$. However, most of this energy is dissipated in the surface mixed layer and shallow seas. The total amount of external mechanical energy sustaining the turbulent mixing in the subsurface ocean interior is in the order of 1 TW. Hence, the entropy production associated with mechanical energy dissipation, which in turn is associated with mixing in the ocean interior, is on the order of 0.004 TW K^{-1} .

Mechanical Energy and the World Ocean Circulation

Mechanical Energy Balance

As discussed above, although there is a large amount of thermal energy going through the ocean, it cannot be efficiently converted into mechanical energy sustaining the oceanic circulation against the dissipation. On the other hand, although mechanical energy from external sources is 1000 times smaller than the thermal energy flux, it can be used to compensate the dissipation associated with oceanic circulation. Thus, in terms of maintaining the oceanic circulation against dissipation, external sources of mechanical energy is the real force driving the oceanic circulation. In this sense, surface thermohaline forcing works as the necessary preconditions for the thermohaline circulation.

Two of the most important items are the tidal contribution and the wind stress energy input to the surface geostrophic

currents. Munk and Wunsch gave the first rough estimate of tidal dissipation based on tidal assimilation with altimetry data. The net dissipation rate of the surface tides is 3.5 TW ($1 \text{ TW} = 10^{12} \text{ W}$). Most of tidal dissipation takes place in the marginal seas, and there is roughly 0.9 TW left for the tidal dissipation in the open oceans. Tidal dissipation is primarily through internal wave breaking and converting into small-scale turbulence. In a stratified flow, vertical mixing pushes light fluid downward and dense fluid upward; thus, it can convert turbulent kinetic energy into the potential energy of the mean state. The efficiency of such conversion is low, in the order of 0.15–0.2. Therefore, the gravitational potential energy generated by tidal mixing in the open ocean is approximately 0.18 TW.

The other important item is wind stress input to the surface geostrophic current, which is estimated at 0.9 TW. The pathway of this energy remains unclear at this time. Surface wind also contributes mechanical energy into the Ekman layer (3.1 TW) and a huge amount of energy into the surface waves (in the order of 60 TW). However, most of such energy is dissipated within the surface wave boundary layer and the frictional Ekman layer; how much of the mechanical energy from surface wind can penetrate through the base of the mixed layer remains unclear till now.

In summary, our knowledge of mechanical energy balance in the world's ocean remains preliminary. Due to the complicated nature of the relevant physics, we do not yet know how many items are on the list. In a sense, the energy diagram shown in Figure 11 should be taken as a rough estimate only. In fact, many values cited in this diagram are no better than within a factor of two.

What Controls the Strength of the Meridional Overturning Cells?

Simple scaling analysis suggests that under the assumption of constant meridional density difference, the strength of the meridional overturning circulation is proportional to one-third power of the vertical eddy diffusivity. In a stably stratified ocean, vertical (or more accurately, diapycnal) mixing pushes light water downward and heavy water upward; thus, it requires external source of mechanical energy.

Most oceanic general circulation models were run under the assumption that vertical eddy diffusivity was a model parameter, which the modeler could tune more or less arbitrarily, 20 years ago. As the theory of thermohaline circulation advanced over the last two decades, it is now commonly accepted that external sources of mechanical energy from tides and winds must be used as the integral constraint for parameterizing eddy diffusion.

Climate changes can affect the thermohaline circulation in different ways. First, climate changes can alternate the surface thermohaline forcing, such as the air-sea surface heat flux and freshwater flux. Because of such changes in the surface boundary conditions, the thermohaline circulation can be alternated in response. For example, surface freshening may induce thermohaline catastrophe and surface cooling may induce a rapid enhancement of thermohaline overturning circulation. However, in order to maintain a steady state of circulation, the external mechanical energy sustaining the circulation is required.

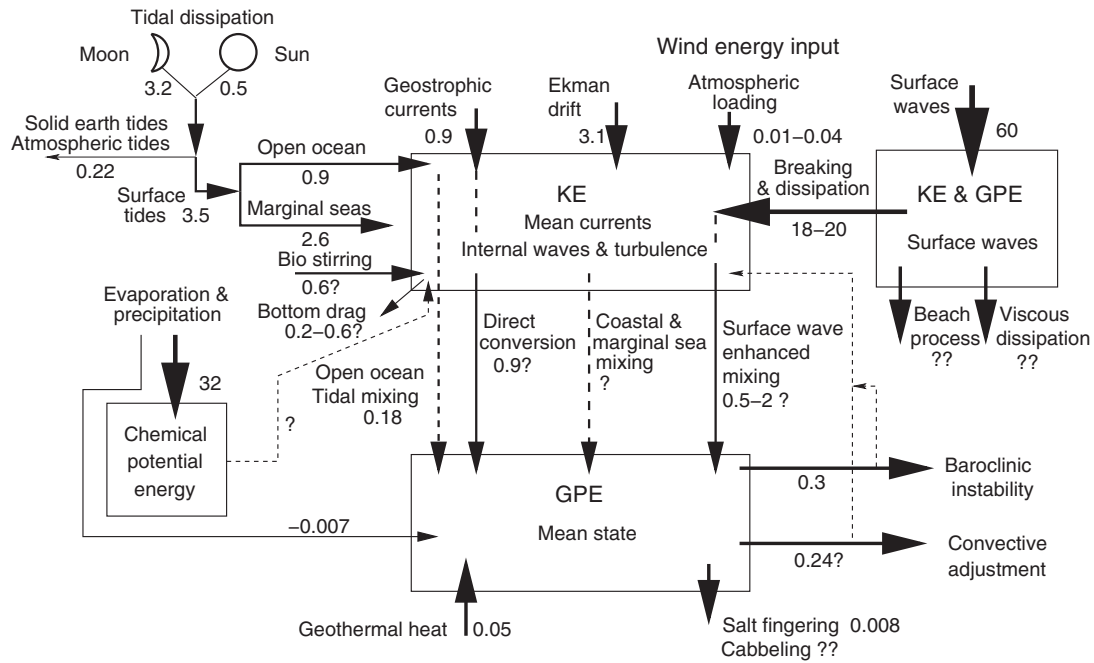


Figure 11 Mechanical energy diagram for the world oceans. (Question marks indicate uncertain values.)

Second, the external mechanical energy from tides and winds can also change with climate. With changes in the external sources of mechanical energy, the thermohaline circulation must be changed in response, because the energy sustaining the circulation against friction is alternated.

Multiple States and Thermohaline Catastrophe

Multiple States

Thermohaline circulation can be studied using simple box models. In 1961, Stommel idealized thermohaline circulation in terms of a two-box model, in which the low-latitude ocean is represented by a single box and the high latitude ocean is represented by another box. These two boxes are linked by imaginary pipes and subject to relaxation conditions for both temperature and salinity. Stommel assumed that relaxation time for salinity was much longer than that of temperature.

This seemingly oversimplified model provides much of the dynamical structure essential for thermohaline circulation in the world's oceans. Introducing two new variables $T = T_1 - T_2$ and $S = S_1 - S_2$, the problem is reduced to solving a simple set of ordinary differential equations of two unknown variables; the solution of this problem can be found in the traditional T - S diagram used in oceanography.

Most importantly, the model has three steady states, as shown in Figure 12. There is a stable thermal mode, which is characterized by sinking of cold and relatively freshwater at the high-latitude box; this is compensated by the equatorward return flow and upwelling at the low-latitude box. The thermal mode is characterized by a relatively fast overturning rate. There is a stable haline mode, which is characterized by sinking of the relatively warm and salty water at the low-latitude box. There is

a compensated flow through upwelling at the high-latitude box. The haline mode is characterized by a relatively slow overturning rate. In addition, there is an unstable thermal mode. Although Stommel's model seems oversimplified, it captures the essential dynamics of the thermohaline circulation, such as the multiple solutions, and these dynamical features have been reproduced in many much more complicated models.

Stommel's box model was designed for a single hemisphere ocean, and it was extended into a three-box model for a two-hemisphere ocean by Rooth. In the two-hemisphere ocean, there are pole-pole modes in which circulation can be primarily controlled by temperature-induced density difference or salinity-controlled density difference. In fact, the modes symmetric with respect to the equator are unstable, and tend to drift toward the pole-pole modes.

Thermohaline Catastrophe

The multiple steady states of thermohaline circulation can be unstable, and catastrophic transition from one steady state to another steady state can take place. Using a numerical oceanic general circulation, F. Bryan discovered the catastrophic transitions from a steady symmetric solution in a two-hemisphere ocean model to a steady asymmetrical solution. Since thermohaline circulation can carry a substantial amount of thermal energy and transport it to high latitudes, the catastrophic transitions of the thermohaline circulation can bring about abrupt changes in the climate condition in the Earth. In particular, such abrupt transitions can be tricked by adding a freshwater cap at the deepwater formation site at high latitudes. Due to its close connection with the freshwater cap, such catastrophic transitions are often called the halocline catastrophe.

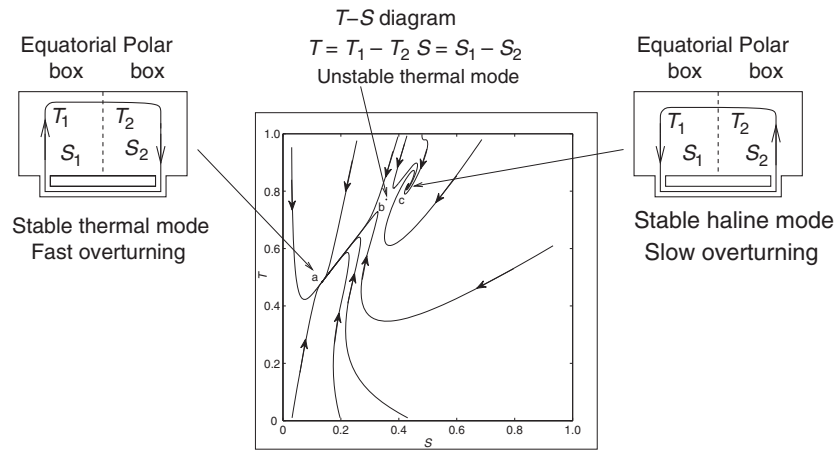


Figure 12 Stommel's two-box model.

Thermohaline circulation plays a critical role of transporting heat poleward, which is a substantial component of the climate system. Past climate proxies indicated that climate system went through many abrupt changes over the past. Therefore, the multiple states and catastrophic changes of thermohaline circulation have capital implication for climate change. For example, Nanabe and Stouffer found that the atmosphere-ocean coupled system can have two steady states, one corresponding to the current climate setting and the other to a state, in which the North Atlantic Ocean is much colder and fresher. The maximum temperature and salinity differences are 5°C and 3‰ .

Thermohaline circulation can have multiple states and oscillations over a broad range of spatial and temporal scales. For example, thermohaline circulation component in connection with the freshwater flux can oscillate in decadal timescale. Furthermore, thermohaline circulation can oscillate on the timescale of thousands of year, a phenomenon called flushing. These types of oscillation are closely linked to the three models discussed above, namely Stommel's two-box model, F. Bryan's two-hemispheric numerical model, and the Manabe-Stouffer coupled model. These models materialized different degrees of idealization for the climate system. However, the multiple stable equilibrium states obtained from these models share fundamentally similar structure. Paleoclimatic records indicate that the meridional circulation in the Atlantic basin has gone through On/Off cycles; thus, one of the most important frontiers in oceanic circulation and climate study is whether such an abrupt change in the North Atlantic circulation may happen in the near future. In particular, the main focus is the potential impact of changes of hydrological cycle and freshwater flux in the climate system.

A likely scenario of global warming is the intensification of hydrological cycle. With more evaporation at low latitudes and more precipitation at high latitudes, the meridional salinity gradient can be enhanced, producing a meridional pressure difference against that due to thermal forcing. As a result, the meridional overturning cell associated with the NADW may be slowed down and even be interrupted. Accordingly, many of

the state-of-art climate models predict that meridional overturning circulation in the Atlantic basin will be substantially reduced over the next hundred years.

In view of global climate changes, within the next 30–50 years the Arctic Ocean may be ice free in summer time. Without the ice, the large amount of relatively freshwater in the Arctic Ocean could be carried out to the North Atlantic Ocean, where such large amount of freshwater may intricate a halocline catastrophe, similar to what has been simulated by F. Bryan and others.

Although climate models have been improved greatly over the past, these models need to be further improved. In particular, the oceanic component of most current climate models is based on the old theoretical assumption that vertical diffusivity is invariant under different climate conditions. In the point of view of the new energy theory, wind stress, and to a lesser degree, the tidal dissipation can be quite different under different climate conditions. Thus, results from such model experiments remain questionable. It is expected that with the rapid progress toward unraveling the mystery of mixing and oceanic circulation, we will be able to simulate and predict the oceanic circulation and climate more accurately in the near future.

See also: Air Sea Interactions: Freshwater Flux; Surface Waves. **General Circulation of the Atmosphere:** Energy Cycle. **Oceanographic Topics:** Surface/Wind Driven Circulation; Water Types and Water Masses. **Satellites and Satellite Remote Sensing:** Surface Wind and Stress.

Further Reading

- Huang, R.X., 2010. *Ocean Circulation, Wind-Driven and Thermohaline Processes*. Cambridge Press.
- Kuhlbrodt, T., Griesel, A., Montoya, M., Levermann, A., Hofmann, M., Rahmstorf, S., 2007. On the driving processes of the Atlantic meridional overturning circulation. *Reviews of Geophysics* 45, RG2001. <http://dx.doi.org/10.1029/2004RG000166>.
- Liu, L.L., Huang, R.X., 2012. The global subduction/obduction rates, their interannual and decadal variability. *Journal of Climate* 25, 1096–1115.

- Manabe, S., Stouffer, R.J., 1988. Two stable equilibria of a coupled ocean–atmosphere model. *Journal of Climate* 1, 841–866.
- Marshall, J., Speer, K., 2012. Closure of the meridional overturning circulation through Southern Ocean upwelling. *Nature Geoscience* 5, 171–180. <http://dx.doi.org/10.1038/ngeo1391>.
- Munk, W.H., Wunsch, C., 1998. Abyssal recipes II: energetics of the tidal and wind mixing. *Deep Sea Research* 45, 1977–2010.
- Pickart, R.S., Spall, M.A., Ribergaard, M.H., Moore, G.W.K., Milliff, R.F., 2003. Deep convection in the Irminger Sea forced by the Greenland tip jet. *Nature* 424, 152–156.
- Schmitz Jr., W.J., 1996. World Ocean Circulation. In: *Some Global Features/North Atlantic Circulation*, vol. I. Woods Hole Oceanographic Institution. Technical Report WHOI-96-03.
- Stommel, H.M., 1961. Thermohaline convection with two stable regimes of flow. *Tellus* 13, 224–230.
- Stommel, H., Arons, A.B., 1960. On the abyssal circulation of the world ocean – I. Stationary planetary flow patterns on a sphere. *Deep Sea Research* 6, 140–154.
- Talley, L.D., 2013. Closure of the global overturning circulation through the Indian, Pacific, and Southern Oceans: schematics and transports. *Oceanography* 26 (1), 80–97.
- Talley, L.D., Pickard, G.L., Emery, W.J., Swift, J.H., 2011. *Descriptive Physical Oceanography, an Introduction*. Elsevier.
- Våge, K., Pickart, R.S., Spall, M.A., Moore, K., Valdimarsson, H., Torres, D.J., Erofeeva, S.Y., Nilsen, J.E.O., 2013. Revised circulation scheme north of the Denmark Strait. *Deep Sea Research* 79, 20–39.
- Warren, B.A., 1981. Deep circulation of the world ocean. In: Warren, B.A., Wunsch, C. (Eds.), *Evolution of Physical Oceanography*. Massachusetts Institute of Technology Press, Cambridge.
- Wunsch, C., Ferrari, R., 2004. Vertical mixing, energy, and the general circulation of the oceans. *Annual Review of Fluid Mechanics* 36, 281–314.

Water Types and Water Masses

WJ Emery, University of Colorado, Boulder, CO, USA

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

Historically, water masses were defined in terms of their temperature–salinity (TS), oxygen, and nutrient properties. The TS characteristic diagram became one of the primary methods of specifying water masses in terms of their TS extremes or singular TS values. Globally, TS curves are found to vary greatly closer to the surface. In the deep ocean, the global water mass distribution is related to the ‘global conveyor belt’ that circulates throughout the world ocean.

Introduction

Much of what is known today about the currents of the deep ocean has been inferred from studies of the water properties such as temperature, salinity, dissolved oxygen, and nutrients. These are quantities that can be observed with standard hydrographic measurement techniques that collect temperatures and samples of water with a number of sampling bottles strung along a wire to provide the depth resolution needed. Salinity or ‘salt content’ is then measured by an analysis of the water sample, which, when combined with the corresponding temperature value at that ‘bottle’ sample, yields temperature and salinity as functions of the depth of the sample. Modern observational methods have in part replaced this sample bottle method with electronic profiling systems, at least for temperature and salinity, but many of the important descriptive quantities such as oxygen and nutrients still require bottle samples accomplished today with a ‘rosette’ sampler integrated with the electronic profiling systems. These new electronic profiling systems have been in use for over 30 years, but still the majority of data useful for studying the properties of the deep and open ocean come from the time before the advent of modern electronic profiling systems. This knowledge is important in the interpretation of the data since the measurements from sampling bottles have very different error characteristics than those from modern electronic profiling systems.

This article reviews the mean properties of the open ocean, concentrating on the distributions of the major water masses and their relationships to the currents of the ocean. Most of this information is taken from published material, including the few papers that directly address water mass structure, along with the many atlases that seek to describe the distribution of water masses in the ocean. Coincident with the shift from bottle sampling to electronic profiling is the shift from publishing information about water masses and ocean currents in large atlases to the more routine research paper. In these papers, the water mass characteristics are generally only a small portion, requiring the interested descriptive oceanographer to go to considerable trouble to extract the information he or she may be interested in. While water mass distributions play a role in many of today’s oceanographic problems, there is very little research directed at improving our knowledge of water mass distributions and their changes over time.

What is a Water Mass?

The concept of a ‘water mass’ is borrowed from meteorology, which classifies different atmospheric characteristics as ‘air masses.’ In the early part of the twentieth century, physical oceanographers also sought to borrow another meteorological concept by separating the ocean waters into ‘warm’ and ‘cold’ water spheres. This designation has not survived in modern physical oceanography, but the more general concept of water masses persists. Some oceanographers regard these as real, objective physical entities, building blocks from which the oceanic stratification (vertical structure) is constructed. At the opposite extreme, other oceanographers consider water masses to be mainly descriptive words, summary shorthand for pointing to prominent features in property distributions.

The concept adopted for this discussion is squarely in the middle, identifying some ‘core’ water mass properties that are the building blocks. In most parts of the ocean, the stratification is defined by mixing both vertical and horizontal orientations of the various water masses that advect into the location. Thus, in the maps of the various water mass distributions, a ‘formation region’ is identified where it is believed that the core water mass has acquired its basic characteristics at the surface of the ocean. This introduces a fundamental concept first discussed by Iselin (1939), who suggested that the properties of the various subsurface water masses were originally formed at the surface in the source region of that particular water mass. Since temperature and salinity are considered to be ‘conservative properties’ (property is changed only at the sea surface), these characteristics would slowly erode as the water properties were advected at depth to various parts of the ocean.

Descriptive Tools: The Temperature–Salinity Curve

Before focusing on the global distribution of water masses, it is appropriate to introduce some of the basic tools used to describe these masses. One of the most basic tools is the use of property versus property plots to summarize an analysis by making extrema easy to locate. The most popular of these is the temperature–salinity (TS) diagram, which relates density to the observed values of temperature and salinity. Originally, the TS curve was constructed for a single hydrographic cast and thus related the TS values collected for a single bottle sample with the salinity computed from that sample. In this way, there was

a direct relationship between the TS pair and the depth of the sample. As the historical hydrographic record expanded, it became possible to compute TS curves from a combination of various TS profiles. This approach amounted to plotting the TS curve as a scatter diagram (Figure 1), where the salinity values were then averaged over a selected temperature interval to generate a discrete TS curve. The TS curve shown in Figure 1, which is an average of all of the data in a 10° square just northeast of Hawaii, shows features typical of those that can be found in all TS curves. As it turned out, the TS pair remained the same while the depth of this pair oscillated vertically by tens of meters, resulting in the absence of a precise relationship between TS pairs and depth. As sensed either by 'bottle casts' or by electronic profilers, these vertical variations express themselves as increased variability in the temperature or salinity profiles, while the TS curve continues to retain its shape, now

independent of depth. Hence, a composite TS curve computed from a number of closely spaced hydrographic stations no longer has a specific relationship between temperature, salinity, and depth.

As with the more traditional 'single-station' TS curve, these area average TS curves can be used to define and locate water masses. This is done by locating extrema in salinity associated with particular water masses. The salinity minimum in the TS curve of Figure 1 is at about 10°C , where there is a clear divergence of TS values as they move up the temperature scale from the coldest temperatures near the bottom of the diagram. There are two separate clusters of points at this salinity minimum temperature, with one terminating at about 13°C and the other transitioning to the warmest temperatures. It is this termination of points that results in a sharp turn in the mean TS curve and causes a very wide standard deviation. These

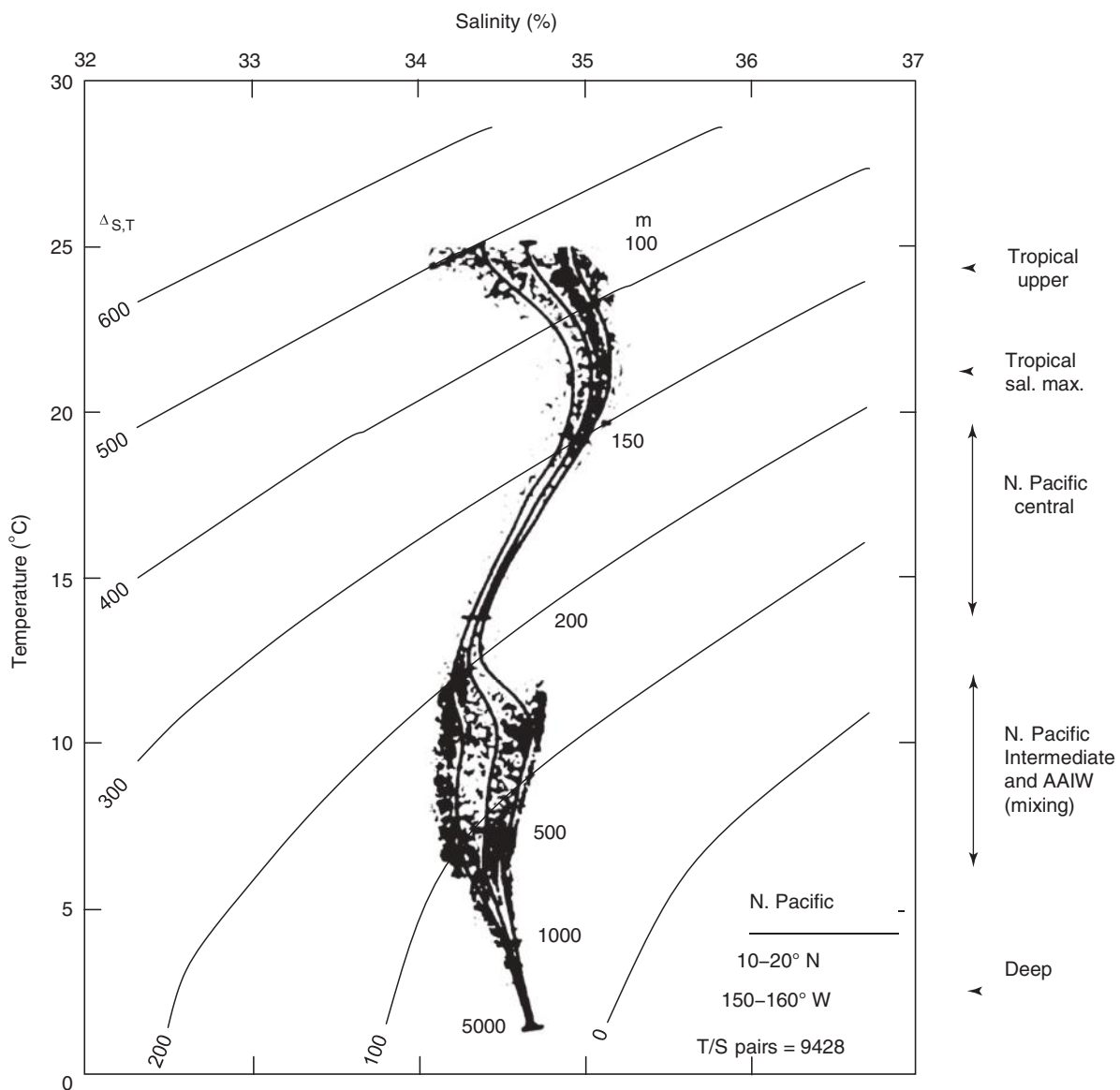


Figure 1 Example of TS 'scatterplot' for all data within a 10° square, with the mean TS curve (center line) and curves for one standard deviation in salinity on either side.

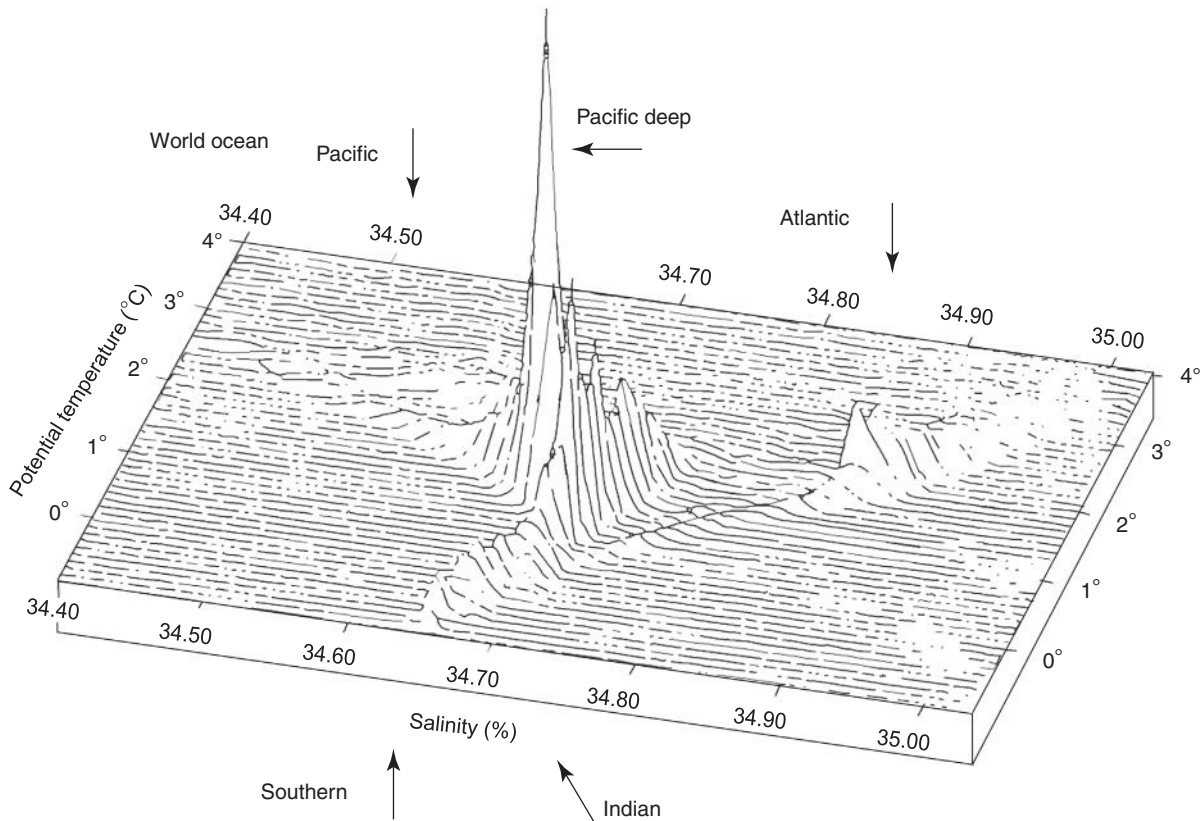


Figure 2 Simulated three-dimensional T–S–V (V – volume) diagram for the cold-water masses of the world ocean.

two clusters of points represent two different intermediate-level water masses. The relatively high salinity values that appear to terminate at 13 °C represent the Antarctic Intermediate Water (AAIW) formed near Antarctica, reaching its northern terminus after flowing up from the south. The coincident less salty points indicate the presence of North Pacific intermediate water moving south from its formation region in the northern Gulf of Alaska.

While there is no accepted practice in water mass terminology, it is generally accepted that a ‘water type’ refers to a single point on a characteristic diagram such as a TS curve. As introduced in this article, ‘water mass’ refers to some portion or segment of the characteristic curve, which describes the ‘core properties’ of that water mass. In the example given here, the salinity characteristics of the two intermediate waters were salinity minima, which were the overall characteristic of the two intermediate waters. We note that the extrema associated with a particular water mass may not remain at the same salinity value. Instead, as one moves away from the formation zone for the AAIW, which is at the oceanographic ‘polar front,’ the sharp minimum that marks the AAIW water that has sunk from the surface to about 1000 m in depth starts to erode, broadening the salinity minimum and slowly increasing its magnitude. By comparing conditions of the salinity extreme at a location with salinity characteristics typical of the formation region, one can estimate the amount of the source water mass that is still present at the distant location. Called the ‘core layer’ method, this procedure was a crucial development in the early study of the ocean water masses and long-term mean currents.

Many variants of the TS curve have been introduced over the years. One particularly instructive form is a ‘volumetric TS curve.’ Here, the oceanographer subjectively decides just how much volume is associated with a particular water mass. This becomes a three-dimensional relationship, which can then be plotted in a perspective format (Figure 2). In this plot, the two horizontal axes are the usual temperature and salinity, while the elevation represents the volumes with those particular TS characteristics. For this presentation, only the deeper water mass characteristics have been plotted, which can be seen by the restriction of the temperature scale from –1.0 to 4.0 °C. Arrows have been added to show just which parts of the ocean various features have come from. That the Atlantic is the saltiest of the oceans is very clear, with a branch to high salinity values at higher temperatures. The most voluminous water mass is the Pacific deep water, which fills most of the Pacific Ocean below the intermediate waters at about 1000 m.

Global Water Mass Distribution

Before turning to the TS curve description of the water masses, it is necessary to indicate the geographic distribution of the basic water masses. The reader is cautioned that this article treats only the major water masses, which most oceanographers accept and agree upon. If a particular region is of interest, close inspection will reveal a great variety of smaller water mass classifications; these can be almost infinite, as higher resolution is obtained in both horizontal and vertical coverages.

Table 1 presents the TS characteristics of the world's water masses. In the table are listed the area name, the corresponding acronym, and the appropriate temperature and salinity range. Recall that the property extreme erodes as it moves away from the source region, so it is necessary to define a range of properties. This is also consistent with the view that a water mass refers to a segment of the TS curve rather than a single point.

As is traditionally the case, the water masses have been divided into deep and abyssal waters, intermediate waters, and upper waters. While the upper waters have the largest property ranges, physically they occupy the least amount of ocean volume. The reverse is true of the deep and bottom waters, which have a fairly restricted range but occupy a substantial portion of the ocean. Since most ocean water mass properties are established at the ocean's surface, those water masses that spend most of their time isolated far from the surface will erode the least and have the longest lifetime. Surface waters, on the other hand, are strongly influenced by fluctuations at the ocean surface, which rapidly erode the water mass properties. In mean TS curves, as in [Figure 1](#), the spread of the standard deviation at the highest temperatures reflects this influence from the heat and freshwater flux exchange that occurs near and at the ocean's surface.

Accompanying the table are global maps of water masses at all three of these levels. The upper waters in [Figure 3](#) have

the most complex distribution, with significant meridional and zonal changes. A 'best guess' at the formation regions for the corresponding water mass is indicated by the hatched regions. For its relatively small size, the Indian Ocean has a very complex upper water mass structure. This is caused by some unique geographic conditions. First is the monsoon, which completely changes the wind patterns twice a year. This causes reversals in ocean currents, which also influence the water masses by altering the contributions of the very saline Arabian Gulf and the fresh Bay of Bengal to the main body of the Indian Ocean. All of the major rivers in India flow to the east and discharge into the Bay of Bengal, making it a very fresh body of ocean water. To the west of the Indian subcontinent is the Arabian Sea with its connection to the Persian Gulf and the Red Sea, both locations of extremely salty water, making the west side of India very salty and the east side very fresh. The other upper ocean water masses in the Indian Ocean are those associated with the Antarctic circumpolar current, which are found at all of the longitudes in the Southern Ocean.

As the largest ocean basin, the Pacific has the strongest east-west variations in upper water masses, with east and west central waters in both the North and South Hemispheres. Unique to the Pacific is the fairly wide band of the Pacific Equatorial Water, which is strongly linked to the equatorial

Table 1 Temperature–salinity (TS) characteristics of the world's water masses

Layer	Atlantic Ocean	Indian Ocean	Pacific Ocean
Upper waters (0–500 m)	Atlantic Subarctic Upper Water (ASUW) (0.0–4.0 °C, 34.0–35.0%) Western North Atlantic Central Water (WNACW) (7.0–20.0 °C, 35.0–36.7%) Eastern North Atlantic Central Water (ENACW) (8.0–18.0 °C, 35.2–36.7%) South Atlantic Central Water (SACW) (5.0–18.0 °C, 34.3–35.8%)	Bengal Bay Water (BBW) (25.0–29 °C, 28.0–35.0%) Arabian Sea Water (ASW) (24.0–30.0 °C, 35.5–36.8%) Indian Equatorial Water (IEW) (8.0–23.0 °C, 34.6–35.0%) Indonesian Upper Water (IUW) (8.0–23.0 °C, 34.4–35.0%) South Indian Central Water (SICW) (8.0–25.0 °C, 34.6–35.8%)	Pacific Subarctic Upper Water (PSUW) (3.0–15.0 °C, 32.6–33.6%) Western North Pacific Central Water (WNPWCW) (10.0–22.0 °C, 34.2–35.2%) Eastern North Pacific Central Water (ENPCW) (12.0–20.0 °C, 34.2–35.0%) Eastern North Pacific Transition Water (ENPTW) (11.0–20.0 °C, 33.8–34.3%) Pacific equatorial water (PEW) (7.0–23.0 °C, 34.5–36.0%) Western South Pacific Central Water (WSPCW) (6.0–22.0 °C, 34.5–35.8%) Eastern South Pacific Central Water (ESPCW) (8.0–24.0 °C, 34.4–36.4%) Eastern South Pacific Transition Water (ESPTW) (14.0–20.0 °C, 34.6–35.2%)
Intermediate waters (500–1500 m)	Western Atlantic Subarctic Intermediate Water (WASIW) (3.0–9.0 °C, 34.0–35.1%) Eastern Atlantic Subarctic Intermediate Water (EASIW) (3.0–9.0 °C, 34.4–35.3%) Antarctic Intermediate Water (AAIW) (2–6 °C, 33.8–34.8%) Mediterranean Water (MW) (2.6–11.0 °C, 35.0–36.2%) Arctic Intermediate Water (AIW) (–1.5 to 3.0 °C, 34.7–34.9%)	Antarctic Intermediate Water (AAIW) (2–10 °C, 33.8–34.8%) Indonesian Intermediate Water (IIW) (3.5–5.5 °C, 34.6–34.7%) Red Sea–Persian Gulf Intermediate Water (RSPGIW) (5–14 °C, 34.8–35.4%)	Pacific Subarctic Intermediate Water (PSIW) (5.0–12.0 °C, 33.8–34.3%) California Intermediate Water (CIW) (10.0–12.0 °C, 33.9–34.4%) Eastern South Pacific Intermediate Water (ESPIW) (10.0–12.0 °C, 34.0–34.4%) Antarctic Intermediate Water (AAIW) (2–10 °C, 33.8–34.5%)
Deep and abyssal waters(1500 m bottom)	North Atlantic Deep Water (NADW) (1.5–4.0 °C, 34.8–35.0%) Antarctic Bottom Water (AABW) (–0.9 to 1.7 °C, 34.64–34.72%) Arctic Bottom Water (ABW) (–1.8 to –10.5 °C, 34.88–34.94%)	Circumpolar Deep Water (CDW) (1.0–2.0 °C, 34.62–34.73%) Circumpolar surface waters	Circumpolar Deep Water (CDW) (0.1–2.0 °C, 34.62–34.73%) Subantarctic Surface Water (SASW) (3.2–15.0 °C, 34.0–35.5%) Antarctic Surface Water (AASW) (–1.0 to 1.0 °C, 34.0–34.6%)

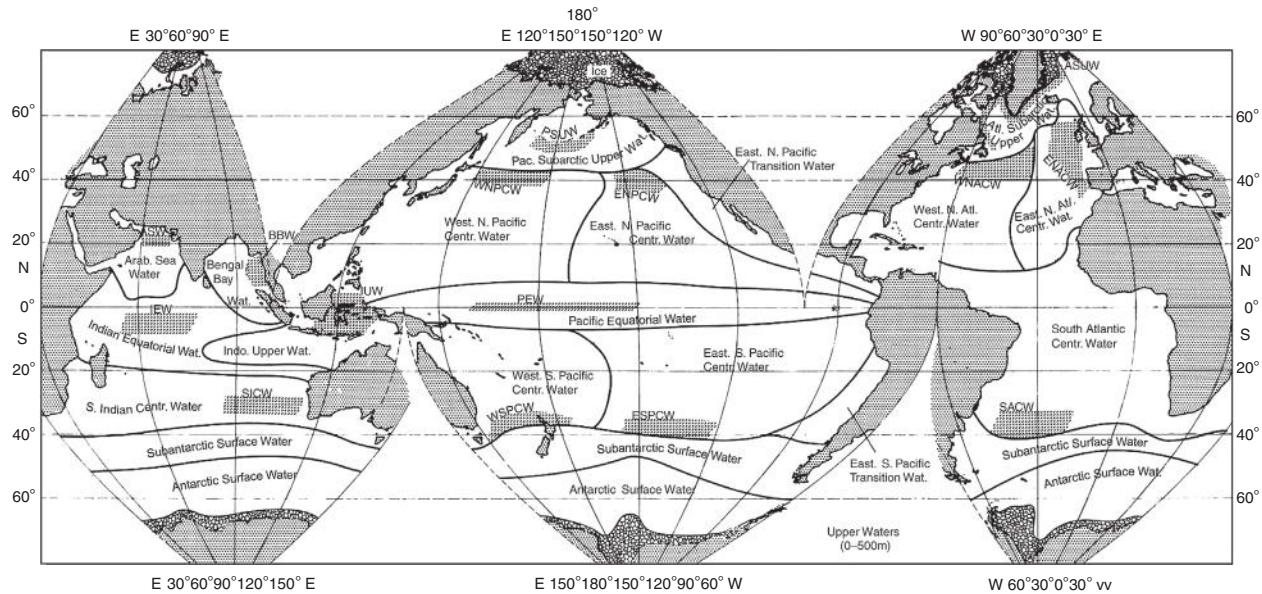


Figure 3 Global distribution of upper waters (0–500 m). Water masses are in abbreviated form with their boundaries indicated by solid lines. Formation regions for these water masses are marked by cross-hatching and labeled with the corresponding acronym title.

upwelling, which may not exist in El Niño years. None of the other two ocean basins have this equatorial water mass in the upper ocean. The Atlantic Ocean has northern hemisphere upper water masses that can be separated east–west, while the South Atlantic upper water mass cannot be separated east–west into two parts. Note the interaction between the North Atlantic and the Arctic Ocean through the Norwegian Sea and Fram Strait. Also in these locations are found the source regions for a number of Atlantic water masses. Compared with the other two oceans, the Atlantic has the most water mass source regions, which produce a large part of the deep and bottom waters of the world ocean.

The chart of intermediate water masses in [Figure 4](#) is much simpler than that of the upper ocean water masses in [Figure 3](#). This reflects the fact that there are far fewer intermediate waters, and those that are present fill large volumes of the intermediate-depth ocean. The North Atlantic has the most complex horizontal structure of the three oceans. Here, intermediate waters form at the source regions in the northern North Atlantic. One exception is the Mediterranean Intermediate Water, which is a consequence of climatic conditions in the Mediterranean Sea. This salty water flows out through the Straits of Gibraltar at about 320 m depth, where it then descends to at least 1000 m, and maybe a bit more. It now sinks

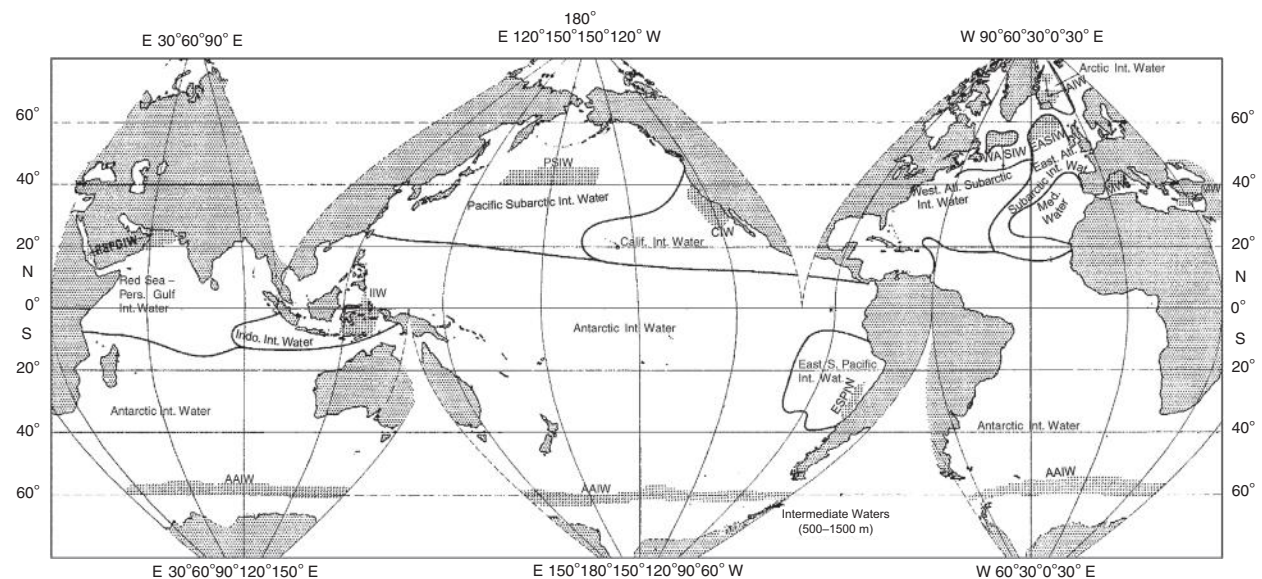


Figure 4 Global distribution of intermediate water (550–1500 m). Lines, labels, and hatching follow the same format as described for [Figure 3](#).

below the vertical range of the less saline AAIW, instead joining with the higher salinity of the deeper North Atlantic Deep Water (NADW), which maintains the salinity maximum indicative of the NADW.

In the Southern Ocean, the formation region for the AAIW is marked as the location of the oceanic Polar Front, which is known to vary considerably in strength and location, moving the formation region north and south. That this AAIW fills a large part of the ocean can be clearly seen in all of the ocean basins. In the Pacific, the AAIW extends north to about 20° N, where it meets the North Pacific Intermediate Water (NPIW), as noted from Figure 1. The AAIW reaches about the same latitude in the North Atlantic, but it reaches only about 5° S in the Indian Ocean. In the Pacific, the northern intermediate waters are mostly from the North Pacific, where the NPIW is formed. There is, however, another intermediate water of smaller volume that is formed in the transition region west of California, mostly as a consequence of coastal upwelling. A similar intermediate water formation zone can be found in the South Pacific mainly off the coast of South America, which generates a minor intermediate water mass.

The deep and bottom waters mapped in Figure 5 are restricted in their movements to the deeper reaches of the ocean. For this reason, the 4000 m depth contour has been plotted in Figure 5, and a good correspondence can be seen between the distribution of bottom water and the deepest bottom topography. Some interesting aspects of this bottom water can be seen in the eastern South Atlantic. As the dense bottom water makes its way north from the Southern Ocean, in the east it runs into the Walvis Ridge, which blocks it from further northward extension. Instead, the bottom water flows north along the west of the mid-Atlantic ridge and, finding a deep passage in the Romanche Gap, flows eastward and then south to fill the basin north of the Walvis Ridge. A similar complex pattern of distribution can be seen in the Indian

Ocean, where the east and west portions of the basin fill from the south separately because of the central ridge in the bottom topography. In spite of the requisite depth of the North Pacific, the Antarctic Bottom Water (AABW) does not extend as far northward in the North Pacific. This means that some variant of the AABW, created by mixing with other deep and intermediate waters, occupies the most northern reaches of the deep North Pacific. Because the North Pacific is essentially 'cut off' from the Arctic, there is no formation region of deep and bottom water in the North Pacific.

The three-dimensional TS curve of Figure 2 indicated that the most abundant water, mass marked by the highest peak in this TS curve, corresponded to Pacific Deep Water. In Table 1, there is listed something called 'Circumpolar Deep Water' in the deeper reaches of both the Pacific and Indian Oceans. This water mass is not formed at the surface but is instead a mixture of NADW, AABW, and the two intermediate waters present in the Pacific. The AABW forms in the Weddell Sea as the product of very cold, dense freshwater flowing off the continental shelf. It then sinks and encounters the upwelling NADW, which adds a bit of salinity to the cold freshwater, making it even denser. This very dense product of Weddell Sea shelf water and NADW becomes the AABW, which then sinks to the very bottom and flows out of the Weddell Sea to fill most of the bottom layers of the world ocean. It is probable that a similar process works in the Ross Sea and some other areas of the continental shelf to form additional AABW, but the Weddell Sea is thought to be the primary formation region of AABW.

Summary TS Relationships

As pointed out in this article, one of the best ways to detect specific water masses is with the TS relationship, whether it is computed for single hydrographic casts or from a historical

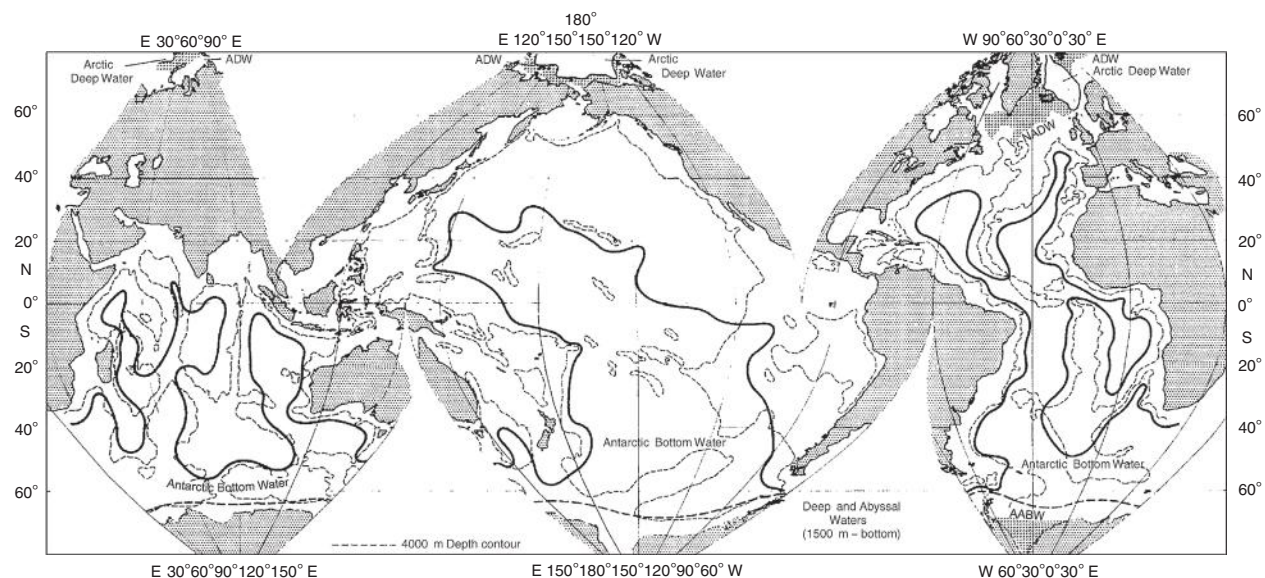


Figure 5 Global distribution of deep and abyssal waters (1500 m–bottom). Contour lines describe the spreading of abyssal water (primarily AABW). The formation of NADW is indicated again by hatching, and its spreading terminus, near the Antarctic, by a dashed line, which also suggests the global communication of this deep water around the Antarctic.

accumulation of such hydro casts. Here, traditional practice is followed, and the summary TS curves are divided into the major ocean basins starting with the Atlantic (Figure 6). Once again, the higher salinities typical of the Atlantic can be clearly seen. The highest salinities are introduced by the Mediterranean outflow marked as MW in Figure 6. This joins with water from the North Atlantic to become part of the NADW, which is marked by a salinity maximum in these TS curves. The AAIW is indicated by the sharp salinity minimum at lower temperatures. The source water for the AAIW is marked by a dark square in the figure. The AABW is a single point, which now does not represent a 'water type' but, rather, a water mass. The difference is that this water mass has very constant TS properties that are represented by a single point in the TS curves. Note that this is the densest water on this TS diagram (the density lines are shown as the dashed curves in the TS diagram marked with the value of σ).

The rather long segments stretching to the upper temperature and salinity values represent the upper waters in the Atlantic. While this occupies a large portion of the TS space, it covers only a relatively small part of the upper ocean when compared to the large volumes occupied by the deep and bottom water masses. From this TS diagram, it can be seen that the upper waters are slightly different in the South Atlantic, the East North Atlantic, and the West North Atlantic. Of these differences, the South Atlantic differs more strongly from the other two than they do from each other.

By comparison with Figure 6, the Pacific TS curves (Figure 7) are very fresh, with all but the highest upper water mass having salinities below 35‰. The bottom property anchoring this curve is the circumpolar deep water (CDW), which is used to identify a wide range of TS properties that are known to be deep and bottom water but that have not been

identified in terms of a specific formation region and TS properties. As with the AABW, a single point at the bottom of the curves represents the CDW. The relationship between the AAIW and the PSIW can be clearly seen in this diagram. The AAIW is colder and saltier than is the PSIW, which is generally a bit higher in the water column, indicated by the lower density of this feature. There are no external sources of deep salinity like with the Mediterranean Water in the Atlantic. Instead, there is a confusing plethora of upper water masses that clearly separate the east–west and north–south portions of the basin. So, we have Eastern North Pacific Central Water (ENPCW) and Western North Pacific Central Water (WNPCW), as well as Eastern South Pacific Central Water (ESPCW) and Western South Pacific Central Water (WSPCW).

The 'central waters' all refer to open-ocean upper water masses. The more coastal water masses such as the Eastern North Pacific Transition Water (ENPTW) are typical of the change in upper water mass properties that occurs near the coastal regions. The same is also true of the South Pacific. In general, the fresher upper-layer water masses of the Pacific are located in the east, where river runoff introduces a lot of freshwater into the upper ocean. To the west, the upper water masses are saltier, as shown by the quasilinear portions of the TS curves corresponding to the western upper water masses. The Pacific Equatorial Water (PEW) is unique in the Pacific probably due to a well-developed equatorial circulation system. As seen in Figure 7, the PEW TS properties lie between those of the east and west central waters.

The Indian Ocean TS curves in Figure 8 are quite different from those of either the Atlantic or the Pacific. Overall, the Indian Ocean is quite a bit saltier than the Pacific but not quite as salty as the Atlantic. Also like the Atlantic, the Indian Ocean receives salinity input from a marginal sea as the Red Sea

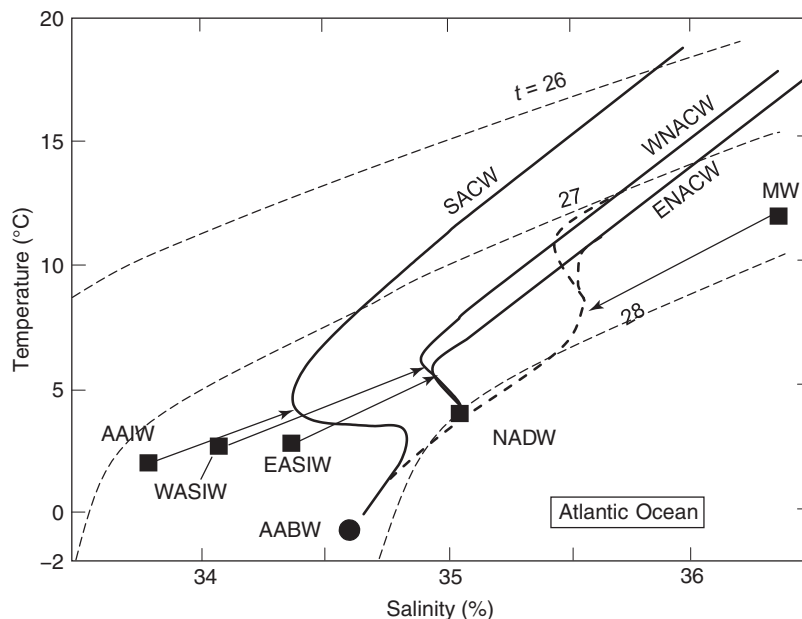


Figure 6 Characteristic temperature–salinity (TS) curves for the main water masses of the Atlantic Ocean. Water masses are labeled by the appropriate acronym, and core water properties are indicated by a dark square with an arrow to suggest their spread. The cross-isopycnal nature of some of these arrows is not intended to suggest a mixing process but merely to connect source waters with their corresponding characteristic extrema.

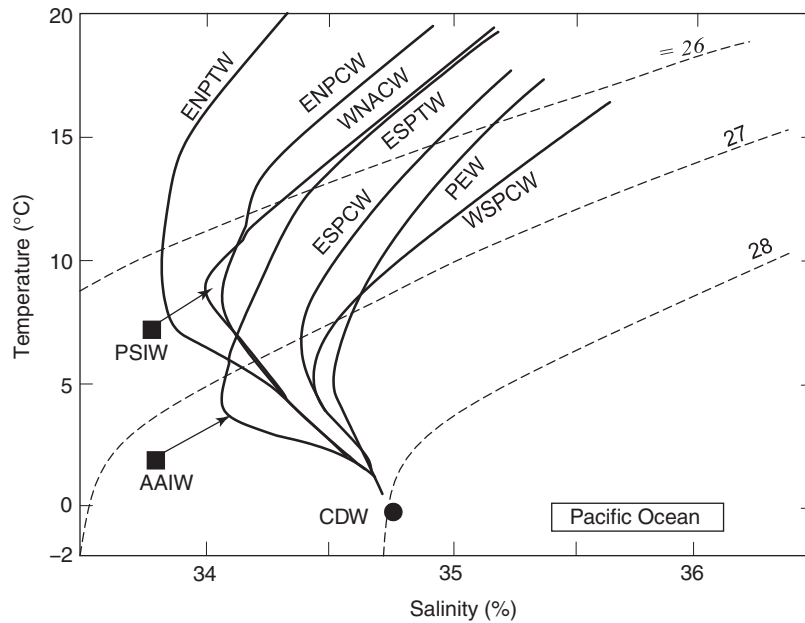


Figure 7 Characteristic TS curves for the main water masses of the Pacific Ocean.

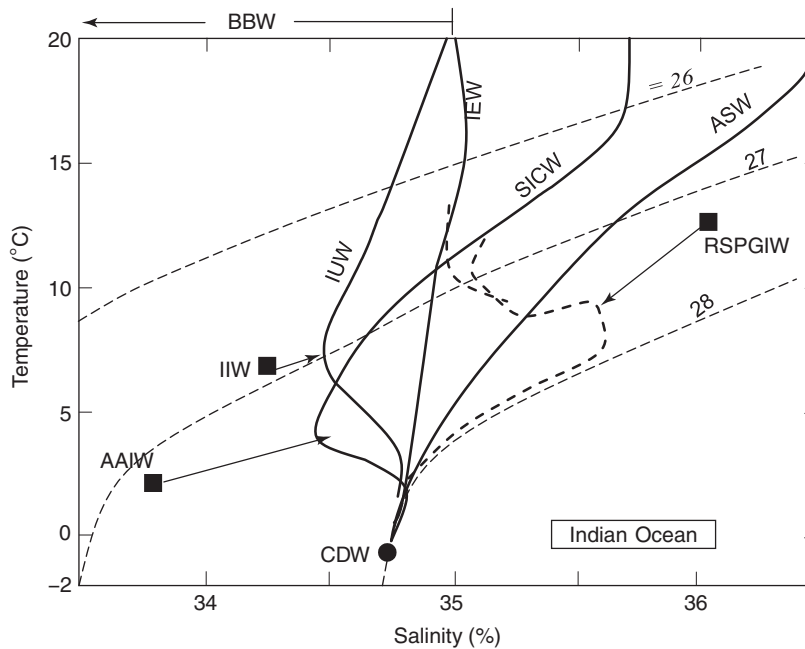


Figure 8 Characteristic TS curves for the main water masses of the Indian Ocean. All labels as in [Figure 6](#).

deposits its salt-laden water into the Arabian Sea. Its presence is noted in [Figure 8](#) as the black box marked RSPGIW (for Red Sea–Persian Gulf intermediate water). Added at the sill depth of the Red Sea, this intermediate water contributes to a salinity maximum that is seasonally dependent.

The bottom water is the same CDW that we saw in the Pacific. Unlike the Pacific, the Indian Ocean equatorial water masses are nearly isohaline above the point representing the CDW. In fact, the line that represents the Indian Equatorial Water (IEW) runs almost straight up from the CDW at about

0.0 °C to the maximum temperature at 20 °C. There is expression of the AAIW in the curve that corresponds to the South Indian Central Water (SICW). A competing Indonesian Intermediate Water (IIW) has higher temperature and higher salinity characteristics, which result in it having an only slightly lower density, creating the weak salinity minimum in the curve transitioning to the Indian Upper Water (IUW). The warmest and saltiest part of these TS curves represents the Arabian Sea Water (ASW) on the western side of the Indian subcontinent.

Discussion and Conclusion

The descriptions provided in this article cover only the most general of water masses, their core properties, and their geographic distribution. In most regions of the ocean, it is possible to resolve the water mass structure into even finer elements describing more precisely the differences in temperature and salinity. In addition, other important properties can be used to specify water masses not obvious in TS space. While dissolved oxygen is often used to define water mass boundaries, care must be taken as this nonconservative property is influenced by biological activity and the chemical dissolution of dead organic material falling through the water column. Nutrients also suffer from modification within the water column, making their interpretation as water mass boundaries more difficult. Characteristic diagrams that plot oxygen against salinity or nutrients can be used to seek extrema that mark the boundaries of various water masses.

The higher vertical resolution property profiles possible with electronic profiling instruments also make it possible to resolve water mass structure that was not even visible with the lower vertical resolution of earlier bottle sampling. Again, this complexity is merited only in local water mass descriptions and cannot be used for global-scale descriptions. At this global scale, the descriptive data available from the accumulation of historical hydrographic data are adequate to map the large-scale water mass distribution, as has been done in this article.

See also: **Boundary Layer (Atmospheric) and Air Pollution:** Ocean Mixed Layer. **General Circulation of the Atmosphere:** Mean Characteristics; Overview. **Oceanographic Topics:** General Processes; Surface/Wind Driven Circulation; Thermohaline Circulation.

Further Reading

- Emery, W.J., Meincke, J., 1986. Global water masses: summary and review. *Oceanologica Acta* 9, 383–391.
- Iselin, C.O'D., 1939. The influence of vertical and lateral turbulence on the characteristics of the waters at mid-depths. *Transactions of the American Geophysical Union* 20, 414–417.
- Mamayev, O.I., 1975. Temperature–salinity analysis of world ocean waters. In: *Elsevier Oceanography Series*, vol. 11. Elsevier Scientific Pub. Co., Amsterdam. p. 374.
- Pickard, G.L., Emery, W.J., 1992. *Descriptive Physical Oceanography*, fifth ed. Pergamon Press, Oxford, England.
- Reid, J.L., 1973. Northwest Pacific Ocean water in winter. In: *The Johns Hopkins Oceanographic Studies*, vol. 5. Johns Hopkins Press, Baltimore, MD. p. 96.
- Sverdrup, H.U., Johnson, M.W., Fleming, R.H., 1941. *The Oceans*. Prentice-Hall Inc, Englewood Cliffs, NJ. p. 1087.
- Worthington, L.V., 1976. On the North Atlantic circulation. In: *The Johns Hopkins Oceanographic Studies*, vol. 6. Johns Hopkins Press, Baltimore, MD. p. 110.
- Worthington, L.V., 1981. The water masses of the world ocean: some results of a fine-scale census (Chapter 2). In: Warren, B.A., Wunsch, C. (Eds.), *Evolution of Physical Oceanography*. MIT Press, Cambridge, MA, pp. 42–69.

Contents

Optical Remote Sensing Instruments

Airglow Instrumentation

Optical Remote Sensing Instruments

GG Shepherd, York University, Toronto, ON, Canada

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

The optical spectroscopic instruments employed in remote sensing of the atmosphere were invented more than 100 years ago, and it is remarkable how little the fundamental concepts embodying these instruments have changed. However, the technologies involved in using these concepts are totally different. About 50 years ago, photoelectric detection of photons became possible and with this an understanding of what the responsivity of an optical system really means. This caused a revolution in the field that can be called interferometric spectroscopy, at just about the same time as the space age began. About 25 years ago, array detectors, particularly CCDs, began to be used in space instruments. The application of existing instruments as spectral imagers changed the technology again. In this article, the fundamental concepts are reviewed, their characteristics and then their application, including imagers for remote sensing, are described.

Introduction

The sun that sustains the environment also makes it possible to study and monitor the atmosphere through optical remote sensing, by taking advantage of its incredibly broad electromagnetic spectrum. The shorter wavelengths, in the far and extreme ultraviolet, are absorbed in the high atmosphere, driving its photochemistry, which in turn produces its own spectrum of emitted radiation, called airglow. Light in the visible region reaches the Earth's surface because the primary atmospheric constituents of molecular oxygen and nitrogen are transparent to solar radiation but the minor constituents such as carbon dioxide, ozone, and water vapor do absorb at specific wavelengths, providing a means of detecting them from the light absorbed or reflected back into the space. Superimposed on this is molecular and particulate scattering from the atmosphere. The sun provides an indirect but equally important method of optical remote sensing from the atmosphere's thermal infrared radiation, from those minor constituents, or from the Earth's surface itself. As a result, instruments covering a wide range of wavelengths can be used to determine species concentrations in the upper atmosphere, from the troposphere through the stratosphere and mesosphere, up into the uppermost region, the thermosphere.

What the researcher needs to know in order to deduce these concentrations is the number of photons emitted (the volume emission rate), or absorbed, per unit volume of the

atmosphere. What is observed is the line integral (the integrated emission rate) of these quantities along the line of sight, but from a set of line integrals, the true height profiles can be retrieved through numerical inversion.

In addition to retrieving constituent concentrations, derived quantities can be determined as well. Thermal emission from the atmosphere is a function of its temperature and the species' concentrations, following the physics of blackbody radiation. If the species has a known mixing ratio (fractional concentration in the atmosphere) as is the case for CO₂, the temperature may be determined. Once this is known, the concentrations of unknown minor constituents may be obtained from their thermal emission. Temperature profiles may also be determined from the molecular scattering of the atmosphere, called Rayleigh scattering, as the observed scale height is a function of the temperature. In the upper atmosphere, rotational temperatures, which accurately represent kinetic temperatures, can be obtained from the relative strengths of rotational lines for molecular emission and from atoms from their line widths, where those are dominated by the Doppler shifts of their thermal motion. In a similar manner, winds may be obtained from the Doppler shifts corresponding to the bulk motion of the observed atmosphere, superimposed on their random motions.

All of this capability depends on the ability of optical remote sensing instruments to observe the atmospheric spectra with the required responsivity and spectral resolution. Going

back over 100 years, the diffraction grating spectrometer along with the Fabry–Perot and Michelson interferometers were well-known instruments. With the advent of photoelectric detection, a large number of alternative approaches were explored, but the recent needs in remote sensing have brought these three back to the fore again, including an interesting hybrid, in part because of the current emphasis on imaging. In this article, the conceptual bases of the different instruments are described and the factors involved in choosing the best instrument for a particular application are discussed.

Responsivity and Superiority of Optical Spectroscopic Systems

The responsivity R of an optical system is the ratio of its output signal, say the number of electrons collected during the observation, to the input radiance. It is given by $R = A \Omega \tau q T$, where A is the area of the instrument opening aperture, Ω is the solid angle of observation, τ is the instrument transmittance, q is the detector quantum efficiency, and T is the time duration of the observation. Since the radiance is what is presented by the atmosphere, the researcher has control only over the responsivity, which determines the error of the observation. The $A\Omega$ was called the *étendue* by Pierre Jacquinot, a pioneer in understanding the determining factors in responsivity, and it is absolutely fundamental as τ and q are determined purely by technical considerations. Ω is the spatial resolution element required for the measurement and A is as large as necessary, compatible with the available budget but also with the allowed instrument volume and mass. But Ω cannot be chosen at will; it is restricted by the dispersive element within the basic spectroscopic system. Normally one or more telescopes transform the solid angle of the field of view on the atmosphere to its value within the dispersing element. But the *étendue* here is the same as in the field of view; *étendue* is conserved throughout an optical system. The reason for the linkage between solid angle and spectral resolution is that for any dispersing element the transmitted wavelength, λ , changes with off-axis angle; the spectral resolution $d\lambda$ is often described in terms of the ‘resolving power,’ simply $\mathcal{R} = \lambda/d\lambda$. Jacquinot showed that for each spectroscopic instrument, the value of $\Omega \mathcal{R}$ has a fixed value; for the interferometers he considered, both Fabry–Perot and Michelson, the value is 2π . Jacquinot described this restriction in terms of what he called the ‘luminosity resolving power product,’ luminosity being $A\Omega$. In ‘Spectral Imaging of the Atmosphere,’ the author proposed that $\Omega \mathcal{R}$ be called the ‘superiority’ (S) of the instrument, because it does not include the arbitrary factor A .

Jacquinot’s statement that the superiority of each type of optical instrument is fixed was formulated when spectroscopic instruments had a single field of view, and all of the corresponding output photons were collected by a single detector; for those circumstances, the statement is absolute. However, in subsequent developments, additional factors were recognized. The first of these is ‘multiplexing’ and the second is ‘field widening.’ As well, the considerations change for imaging instruments, which are now very much at the forefront of optical remote sensing. All of these factors are described in what follows for specific instruments. These are the (1)

diffraction grating spectrograph (DGS), (2) Fabry–Perot spectrometer (FPS), (3) scanning Michelson interferometer generally known as the Fourier transform spectrometer (FTS), (4) Doppler Michelson interferometer (DMI), and (5) spatial heterodyne interferometer (SHI). It is difficult to be consistent with the nomenclature. The author prefers to call an instrument a spectrometer if its immediate output is a spectrum and a spectrograph where the entire spectrum is recorded at once, originally on film, now on array detectors. A device which produces an interferogram, which requires a Fourier transformation to obtain a spectrum, is herein called an interferometer. However, in the above list, historical precedent is recognized.

The Diffraction Grating Spectrograph

The diffraction grating spectrometer, as described by Jacquinot, has a superiority $S = \beta \tan i$, where β is the angular length of the slit and i is the angle of incidence onto the grating (for simplicity a Littrow arrangement is used where the angles of incidence and diffraction are equal). From basic considerations S must be around 0.1, which is much less than the 2π for the FPS and the FTS. This profound result is what propelled interferometric spectroscopy into prominence in the 1960s, the effects of which are still being incorporated into flight experiments.

However, in remote sensing instruments, the DGS, in which the exit slit is replaced by an array detector, has produced some impressive results. One example is the OSIRIS instrument on the Odin satellite, another is the SCIAMACHY instrument on the ENVISAT satellite, another is the MAESTRO instrument on Canada’s SCISAT satellite, and for the future the OMPS (Ozone Mapping and Profiler Suite) instrument is now in flight on the NASA NPP (National Polar-orbiting Partnership satellite) launched in 2011.

Suppose a CCD detector is employed, with 256×256 pixels. Also suppose that the slit width corresponds to the size of 1 pixel so that the slit corresponds to one column of pixels. Each of these 256 columns can simultaneously measure 256 wavelengths, which increases the responsivity by a factor of 256, this enhancement is called multiplexing. But as an imager, the spatial resolution element can now be just 1 pixel, reducing the signal by $1/256$, compensating exactly for the multiplex increase. Thus, a one-dimensional image is obtained with the same responsivity as for the single-field whole-slit measurement. Multiplexing is therefore an important factor to consider.

Fabry–Perot Spectrometer

The FPS comprises a cavity, called an etalon, formed by two parallel surfaces with coatings of reflectivity slightly less than unity, within which multiple reflections take place. At angles and wavelengths where the multiply-reflected waves are in phase the transmitted signal is strong, and where they are out of phase the light is reflected back to the source. Thus, the etalon acts as a filter, transmitting wavelength λ , as given by eqn [1], where m is the order of interference, n is the refractive index of the cavity medium, normally air, t is

the thickness of the cavity, and θ is the angle of the ray with the etalon normal, measured inside the cavity.

$$m\lambda = 2nt \cos \theta \quad (1)$$

The quantity $2nt$ is called the optical path difference (OPD); it is a measure of the delay time between reflected rays expressed as the distance in vacuum the light would travel in the same time. The quantity m is the number of wavelengths that fit within the cavity. For most applications, the order of interference varies from about 1000 (low resolution) to 100 000 (high resolution). The interference filter is a particular type of low-resolution etalon, so the Fabry–Perot etalon has a wide range of applications. For each value of m , λ is different; thus, the etalon has a spectral transmittance resembling a Dirac comb, with passbands equally separated in the spectrum by an amount known as the free spectral range, $\Delta\lambda$. These multiple passbands must be dealt with in some way if the spectrum is to be measured unambiguously. Since λ is a constant for a constant θ , the transmitted radiation for a given wavelength/order forms a ring of constant θ about the etalon normal in the focal plane. Each order produces a ring of a different radius, forming the familiar Fabry–Perot ring pattern.

The shape of the spectral passband function (its overall transmittance F) is best expressed in terms of wave number, $\sigma = \lambda^{-1}$, where r is the individual surface reflectance, τ the transmittance, and Δ is the OPD of the etalon; $F(\sigma)$ is given by

$$F(\sigma) = \frac{[\tau/(1-r)]^2}{1 + [4r/(1-r)^2]/\sin^2 \pi\sigma\Delta} \quad (2)$$

The transmittance peaks where $\sin^2 \pi\sigma\Delta = 1$, which occurs whenever σ increases by $(1/\Delta)$. The transmittance function is shown in **Figure 1**, for an air-spaced etalon of 2.5 cm spacing having $r = 0.9$; the free spectral range is $1/(2 \times 2.5) = 0.2 \text{ cm}^{-1}$. The full width at half maximum (FWHM) is equal to the free spectral range divided by the finesse f , given by **eqn [3]**,

$$f = \frac{\Delta\sigma}{\text{FWHM}} = \frac{\pi\sqrt{r}}{1-r} \quad (3)$$

which for the reflectance of 0.9 used in **Figure 1** is equal to 29.8.

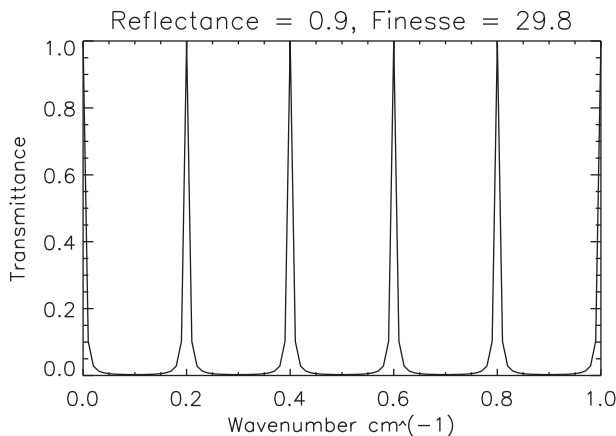


Figure 1 FPS passband response function for an etalon with reflectance of 0.9, corresponding to a finesse of 29.8.

Although the finesse is a constant for a given etalon, the spectral passband width may be chosen by adjusting the etalon separation. This flexibility is one of the advantages of the FPS, which in its simplest form is just an etalon followed by a lens, with a circular aperture at its focal point and a detector behind it. The sharpness of the passband peaks allows the FPS to be used as a spectrometer, but to produce a spectrum the etalon must be scanned. According to **eqn [1]**, to change λ one must change n or t or θ . The refractive index n can be changed by changing the gas pressure in an air-spaced etalon. For a high-resolution etalon, about 1 atmosphere of pressure ($\sim 100 \text{ kPa}$) change is required to scan the spectrum by one order; this was a common technique for ground-based instruments. Varying t involves the mechanical movement of one of the etalon plates; in current instruments, this is done with piezoelectric positioners cemented between the plates. For high accuracy, capacitors are used to sense the spacing and control the piezoelectric voltage in a feedback system. This approach works well at high or low resolution, since a movement of $\lambda/2$ causes a scan through one free spectral range.

To create a single passband with an FPS, multiple etalons can be used. A triple etalon system was used for the High-Resolution Doppler Imager (HRDI) on the Upper Atmosphere Research Satellite (UARS), for middle atmosphere wind measurement.

Finally, λ can be changed by changing θ , and this is readily achieved through the use of an array detector, in which radial distance is equivalent to wavelength, though with a quadratic scale. However, this has the very nice consequence that equal spectral widths correspond to equal areas on the detector array. The disadvantage of this approach is that the rings become very narrow at the edges of the array, requiring small pixels, and in the readout, the readout noise is accumulated for each pixel. This limitation has been relaxed through the use of the CLIO (circle to line imaging optical system), in which a hollow cone is used to image the circular fringes into straight lines that correspond to the pixel layout on the CCD. Because the CCD charges can be binned, that is, added together in a noise-free way on the chip before reaching the output register, this reduces the noise significantly in the final spectrum.

The superiority $S = 2\pi$ for an air-spaced etalon is increased to $S = 2\pi n^2$ where solid etalons are employed, such as with interference filters.

Scanning Michelson Interferometer (Fourier Transform Spectrometer)

The Michelson interferometer shown in **Figure 2** also forms a cavity of thickness t , but this is done in a different way from the FPS. The incoming light is split into two beams created in a beamsplitter as shown in **Figure 2**, each of which suffers a transmission and reflection through the beamsplitter before recombination and registration at the detector. The OPD is the same as given in **eqn [1]** for the FPS but in operation the mirror M_1 is moved over relatively large distances and the output signal as a function of Δ is measured from 0 to some maximum value. The output record, with the constant term removed, is called an interferogram. For a monochromatic signal of infinitesimal line width, the output is simply a sinusoid of constant amplitude,

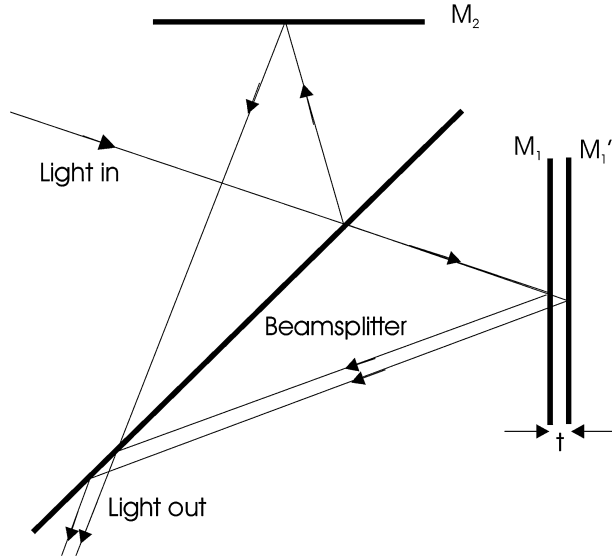


Figure 2 An ordinary Michelson interferometer with a beamsplitter, a fixed mirror M_2 , and a movable mirror M_1 , which is shown in two positions, indicated by M_1 and M_1' . For position M_1 , the combined rays are collinear, corresponding to 0 OPD, while for position M_1' , the rays are displaced, producing an OPD of $2nt \cos \theta$, where θ is the off-axis angle of the rays.

one cycle for every $\lambda/2$ of mirror motion. If the spectral line has a finite width, then the amplitude of the cosinusoid decays with increasing OPD because the different components of the line progressively get out of phase. If the spectrum contains two different wavelengths, there will be two superposed cosinusoids of different periods; for a complex spectrum, many cosinusoids will be superimposed. The spectrum therefore may be retrieved by a Fourier transform as in eqn [4].

$$F(\sigma) = \int_{-\infty}^{\infty} I(\Delta) \cos(2\pi\sigma \Delta) d\Delta \quad (4)$$

In actual practice, it is more complicated than this. The interferogram cannot be measured to infinity but only to some $\Delta_{\max}$. A real instrument has dispersion in its optical elements so that the cosinusoids are phase shifted by amounts that are dependent on wavelength, and the phase must be retrieved. This can be done by measuring from $-\Delta_{\max}$ to $+\Delta_{\max}$ and doing both sine and cosine transforms, but this is an inefficient process. It is preferable and feasible to use a small part of the interferogram about 0 OPD to determine the phase shifts, and to use this to create a corrected interferogram that is symmetric about 0, and for which a cosine transform may be used.

This instrument, which was used by Michelson to measure the separation of very closely spaced lines and the widths of spectral lines, has enjoyed a spectacular increase in attention and use, beginning in the 1960s. As noted, its $S = 2\pi$ gave it a substantial responsivity but it also has a multiplex advantage as every spectral element is simultaneously observed for the full duration of the observation. Examples of highly successful observations are from the ATMOS instrument, flown on the space shuttle, and the ACE-FTS, flown on Canada's SCISAT spacecraft. The challenge in any such instrument is the precision movement of the mirror. For ACE-FTS, this was done using a dual pendulum approach; although not shown in the figure,

the two prisms were mounted on two arms of a pendulum, rotating about a fixed pivot as is shown in the instrument configuration of Figure 3. The method of observation of the absorption spectrum of the atmosphere was solar occultation, in which the sun is observed as it sets or rises at the Earth's limb. With this large radiance, high resolution and ample responsivity were accomplished with a very small instrument (35 kg). Other impressive examples are the MIPAS instrument on ENVISAT, which observed thermal emission from the atmosphere with a resolution of 0.035 cm^{-1} , requiring a large instrument (170 kg), and the Japanese FTS on the GOSAT spacecraft.

Doppler Michelson Interferometer

The Michelson interferometer has the remarkable property that it can be field widened. Figure 2 shows an ordinary Michelson interferometer set to 0 OPD for mirror M_1 ; the rays emerge collinearly, making the OPD independent of angle. When the mirror is moved to M_1' , the emerging rays are no longer collinear and the OPD becomes a strong function of off-axis angle. In Figure 4, an OPD is introduced through the use of a refractive plate in one arm, but the rays still emerge collinearly and the OPD is nearly independent of angle; 'nearly' because the position of the virtual mirror does depend weakly on angle. Analytically the field widening is expressed in eqn [5] for the more general case of refractive materials of thickness t_1 and t_2 in the two arms, having refractive indices of n_1 and n_2 , respectively.

$$\frac{\Delta}{2} = n_1 t_1 - n_2 t_2 - \frac{\sin^2 i}{2} \left(\frac{t_1}{n_1} - \frac{t_2}{n_2} \right) - \frac{\sin^4 i}{8} \left(\frac{t_1}{n_1^3} - \frac{t_2}{n_2^3} \right) + \dots \quad (5)$$

If $t_1 n_2 = t_2 n_1$, the instrument is said to be field widened as the term in $\sin^2 i$ vanishes and the path difference is then independent of angle to fourth order as shown in the equation.

The superiority of the field-widened Michelson interferometer is given by $S = 4\pi n^2 \sqrt{\mathfrak{R}}$. This over rules the Jacquinot statement as the superiority is not constant; it increases with increasing resolving power, which makes it ideal for high-resolution measurements, such as Doppler measurements of line width (temperature) and shift (winds). The off-axis angle inside the instrument can be as much as 6° or even larger.

This approach was used in the WIND Imaging Interferometer (WINDII) flown on NASA's UARS mission from 1991 to 2003 for the measurement of winds in the upper atmosphere, using airglow as a Doppler target. By choosing appropriate pairs of glasses for the two arms, the interferometer was made achromatic and thermally stabilized. It was fabricated as a 'solid' interferometer, which is all glass components cemented together except for an air space in front of the phase stepping mirror. This mirror was stepped in OPD phase increments of 90° to allow the measurement of spectral line phase and thus the wind, from the cosinusoids of single airglow lines.

Spatial Heterodyne Spectroscopy

This approach has become recognized within the last decade as able to fill significant niches in the capability of atmospheric remote sensing. Remarkably it comprises elements from almost

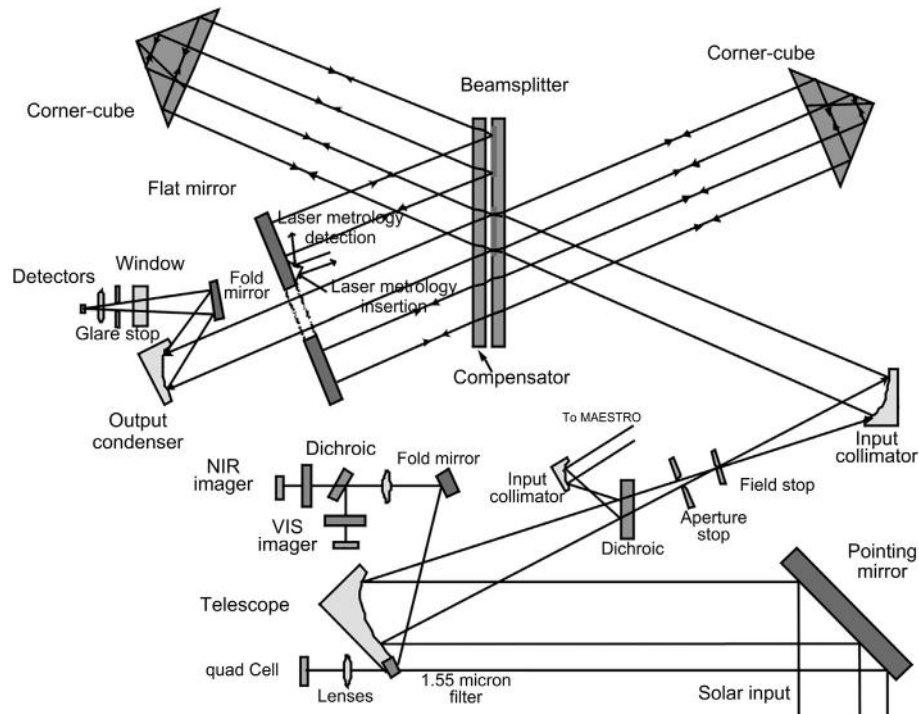


Figure 3 The ACE-FTS instrument flown on the Canadian SCISAT mission. The cube corner prisms were mounted on the two arms of a dual pendulum so that as one prism approached the beamsplitter the other receded. This doubled the OPD for a given reflector motion. The double-pass layout gave a further doubling. With permission from Spectral Imaging of the Atmosphere, Elsevier. Courtesy of ABB Bomem.

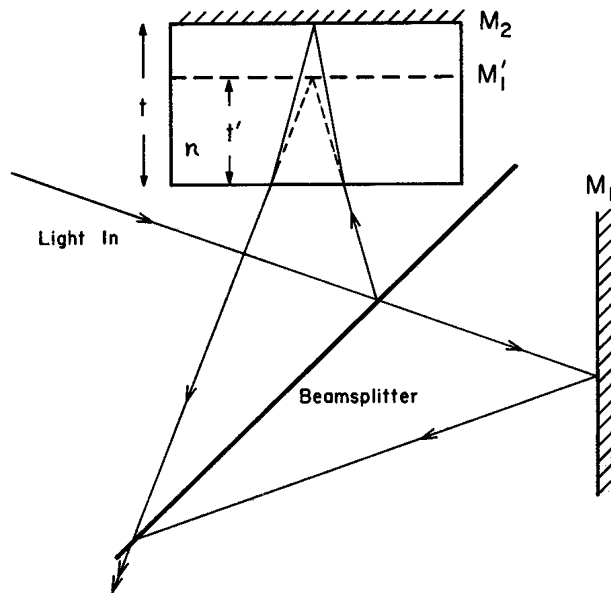


Figure 4 A field-widened Michelson interferometer. The ray paths are similar to those in Figure 2(a) except the mirror M_2 has been replaced by a glass plate, with the apparent reflection taking place from the virtual mirror within it. The rays emerge colinearly so the OPD is nearly independent of incident angle, even though there is a finite path difference.

every instrument discussed up to this point. It dates back to an instrument called SISAM (Spectromètre Interférentiel à Sélection par l'Amplitude de Modulation) by its inventor, Pierre Connes, of Jacquinot's laboratory. Connes recognized that by replacing the mirrors of an ordinary Michelson interferometer with diffraction gratings, the Littrow wavelength would be unmodulated, while other wavelengths would, say by oscillating a retardation plate inside the instrument. Then by rotating the gratings in synchronism, one could bring different wavelengths sequentially into zero modulation and so have a spectrometer – with no need for Fourier transforms. Unfortunately, rotating the gratings in synchronism proved to be impractical. The motivation at that time was to avoid the use of the large computers that were then required for the FTS instruments. Some 40 years later, Fred Roesler and John Harlander implemented a practical version in which the output was imaged onto a two-dimensional array. It can be shown, as in Figure 5, that the wave fronts coming from the two gratings cross at an angle, creating an OPD along the wave fronts. For the Littrow wavelength, this angle is 0, yielding a flat field across the detector (no modulation). For a wavelength different by $\delta\sigma$, the output beams from the two gratings create a co-sinusoidal pattern of one spatial wavelength across the detector. Increasing the wavelength difference to $2\delta\sigma$ increases the angle between the wave fronts and creates two spatial waves across the detector. Thus, a spatial interferogram is recorded, which can be Fourier transformed to yield a spectrum with a number of spectral elements corresponding to one-half the number of pixels across the detector (by Shannon's theorem). The SHS device is an FTS with no moving parts. A true interferogram is obtained, the only difference being that zero

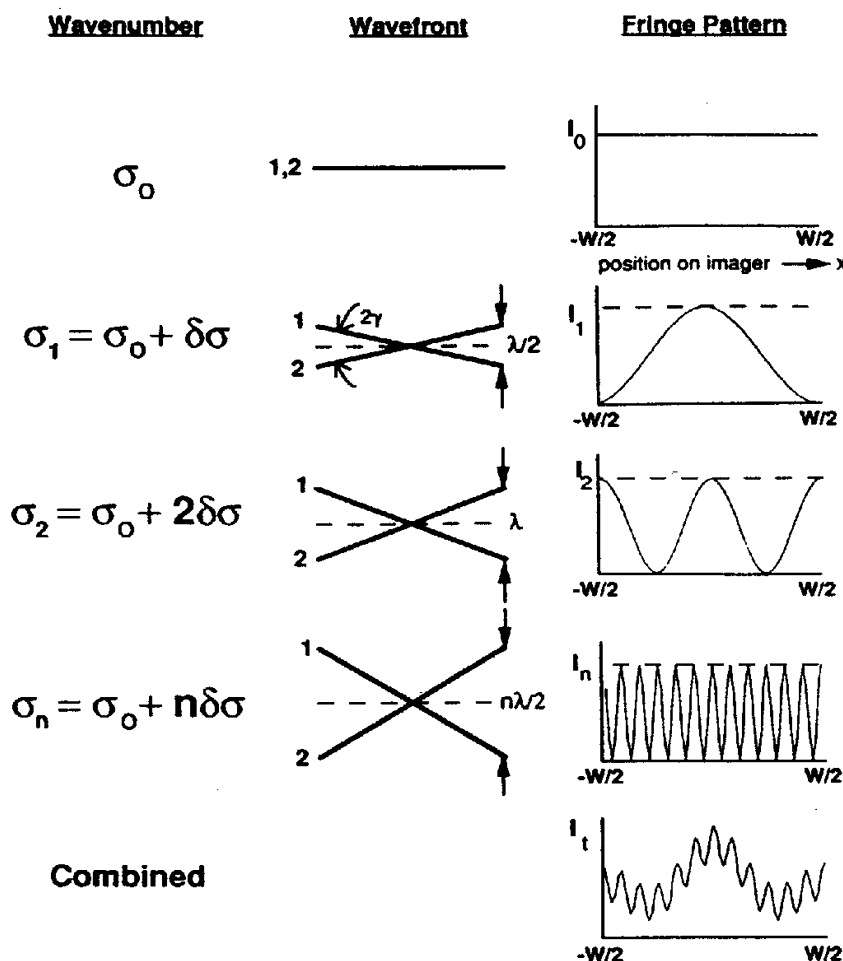


Figure 5 Explanation of the SHS approach after Harlander et al. (1992).

frequency is now the heterodyne frequency. The spectral range is limited, but this is offset by the fact that the desired wavelength range can be selected at will because of the heterodyne capability. In remote sensing terms, a separate instrument is needed for every species to be observed, but this is acceptable given the very high responsivity as this device can also be field widened.

The first flight SHS instruments were fabricated using the same 'solid' instrument approach pioneered by WINDII, where all the optical elements are cemented together, possible because there are no moving parts. These instruments are more formally said to be 'monolithic.' Englert and colleagues at the US Naval Research Laboratory have demonstrated that the SHS configuration may be operated at a near fixed but large OPD by inserting a plate of glass into one arm, as was done with WINDII, allowing the measurement of winds. In this case, a region of the interferometer is recorded, from which a true spectrum cannot be obtained, as 0 OPD is not included. However, the winds can be obtained from the phase shifts, as for WINDII. The advance over WINDII is, however, that the phase shifts can be obtained simultaneously for more than one line. This approach is currently being investigated by the author and colleagues as the optimum way to measure winds in the stratosphere,

using ozone as a Doppler target. For ozone, even with a narrow filter, it is difficult to isolate a single spectral line. But also, for wind measurements, the atmospheric lines and the phase calibration spectral lamp line can be recorded in the same interferogram, reducing the susceptibility to thermal phase drift.

Imaging from Space

As implied earlier, almost all remote sensing optical instruments now incorporate imaging, whether it be conventional nadir viewing imaging (mapping) or imaging vertical altitude profiles at the Earth's limb, both from a low-Earth orbit. A more challenging example is whole-Earth imaging from geostationary orbit (GEO) as has been done with GOES-X and will be done with GOES-R. An even more challenging approach is imaging the polar cap from highly elliptical orbits (HEO) of which the Molniya orbit is the best known. In imaging it is necessary to think of responsivity per pixel.

For a simple camera imager, it can be shown that the responsivity per pixel depends only on the f/number of the

camera lens and the area of the pixel. However, in order to achieve this, the solid angle defined by the camera must be accepted by the dispersing element; if this is not possible then an alternative approach must be sought. Some of these are indicated below, where the five types of instruments discussed in the previous sections are evaluated for their imaging capabilities.

Diffraction Grating Spectrograph

The DGS has a low superiority as discussed earlier, but its geometry has some unique advantages that can be applied to imaging. As shown in Figure 6(a), the optical input is a slit that is imaged on the target. Because the 'diffraction' effect sends different wavelengths off the grating at different angles, the output is a rectangle that can be imaged onto an array detector, also as shown in Figure 6(a). One dimension of the detector is wavelength, the other space – the output is a one-dimensional image in which for each spatial resolution element (pixel) the entire spectrum covered is available. To create a two-dimensional image in nadir viewing the slit can be moved along the spacecraft track perpendicular to its long axis simply by the motion of the spacecraft, called push broom scanning. For

GEO or HEO orbits, a scanning mirror can be used to scan the slit across the target. If the étendue is insufficient to cover the entire image then a cross-scanning mirror can be added to achieve multiple strip scans. For GEO, for example, the Earth is scanned in strips with two mirror motions, one horizontal and one vertical, each strip having a length determined by the angular dimensions of the slit. This is feasible because long scan times are permissible, e.g., 15 min. To cover the spectrum, a typical scenario is to use multiple spectrographs, each defining one spectral region, with the specific spectral bands determined by the placement of the array detectors at the exits of each spectrograph.

Fabry–Perot Spectrometer

As already discussed, the FPS category includes interference filters and the off-axis wavelength change with angle follows the same basic eqn [1]. If the FPS is placed in front of the camera lens then each pixel will have its own passband function which must be taken account in the image analysis. Another approach is to use telecentric optics; image-space telecentric lenses have an exit pupil infinitely far in front of the lens, the result of which is that all of the chief rays (the rays passing through the center of the aperture stop) have a 0 angle of incidence. This minimizes the off-axis angle effect through the FPS and makes the passband the same for all pixels. This is a simple and feasible method for low resolution as shown in Figure 6(b).

One can alternatively utilize the off-axis effect for wavelength scanning. HRDI used the radial dimension for Doppler shifts (and winds) while using a gimbaled telescope for scanning. Working at significant off-axis angles with an FPS means covering multiple orders; this is limited by the fact that the outermost rings in the FPS ring pattern become very narrow, making demands on the detector pixel size. The CLIO configuration as discussed earlier offers a solution to this problem; as employed in NASA's TIMED mission in the TIDI (Timed Doppler Interferometer) which utilized different orders for the four different telescope directions. For two-dimensional imaging, a raster-type scan is required, as indicated in Figure 6(c).

Fourier Transform Spectrometer

The scanning Michelson interferometer, or FTS, covers a wide spectral range at high resolution and enjoys the multiplex advantage as well as a superiority of 2π , but the imaging capability is limited by how far one can go off-axis and still obtain a valid interferogram, as illustrated in Figure 6(c). Subimages could be created by acquiring suites of interferograms near the optical axis and then be used to create a full two-dimensional image by a raster scan using two mirrors as shown in the figure.

Doppler Michelson Interferometer

Because of its field widening, the DMI is a true imager in that it can tolerate sufficiently large off-axis angles to meet the solid angle requirements of the camera for most applications. The spectral information, which for WINDII was the interferogram

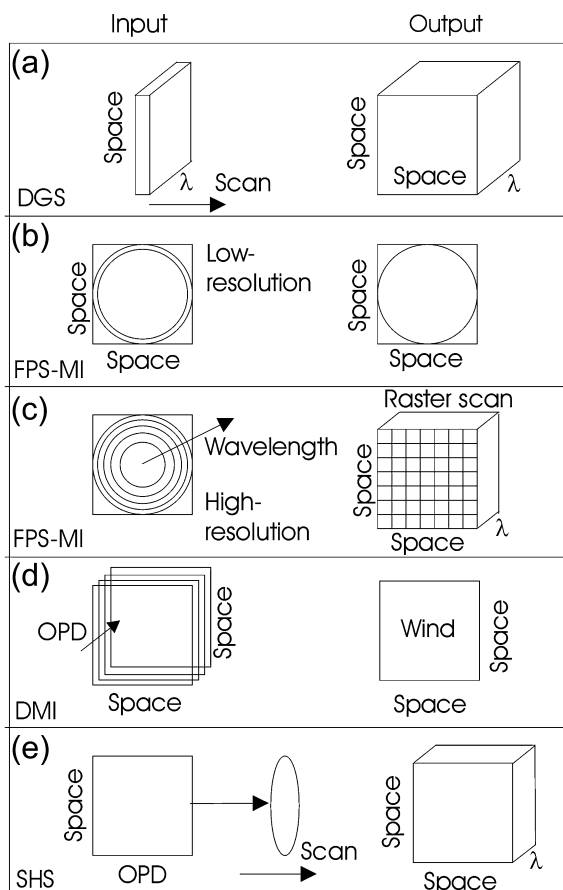


Figure 6 Depiction of ways in which the five different spectroscopic devices discussed in this manuscript may be implemented as spectral imagers. (a) The diffraction grating spectrograph, (b) the low-resolution FPS, (c) the high-resolution FPS, (d) the DMI, and (e) the SHS.

phase, was obtained by acquiring four images with different OPD values using a stepping mirror. This has the disadvantage that the images are taken sequentially, rather than simultaneously, but has the advantage that the derived quantity, the wind, was obtained independently from each pixel, yielding a two-dimensional image of wind as is shown in [Figure 6\(d\)](#). In practice, because of the low available data rate and the desire to measure winds at low emission rates, the WINDII horizontal information was collapsed onto a single profile.

Spatial Heterodyne Spectroscopy

The SHS instrument is a powerful version of the FTS, operating over limited specific wavelength bands and as such, the motivation to employ it for imaging, as for the FTS, is very strong. It is similar to the DGS in that with a two-dimensional array detector one acquires a complete interferogram (and thus spectrum) for each row of pixels. This works well for limb viewing where a single vertical profile is all that is required although there is a potential problem in that the light contributing to the interferogram comes from a spatial region spread out horizontally across the limb. While the atmosphere is approximately horizontally homogeneous it is not entirely so. The solution to this problem is to use an anamorphic imaging system in which one of the two imaging directions is compressed with respect to the other as shown in [Figure 6\(e\)](#) so that a smaller region of atmosphere is used to generate each interferogram. If this is done then the compressed direction is analogous to the DGS slit and the scanning methods described for it above may be used.

Summary

Over the 50 years or so of the space age, a wide variety of optical devices has been used for the remote sensing of the atmosphere. However, over time there has been a trend toward a small number of basic instruments each configured for a specific type of measurement. Those discussed here are (1) the diffraction grating spectrometer, (2) the FPS, (3) the scanning Michelson interferometer/Fourier transform spectrometer, (4) the DMI, and (5) the spatial heterodyne interferometer. A comparison of the inherent responsivities of these instruments has been presented, based on the Jacquinot principle of superiority. When applied to imaging instruments, the

considerations are somewhat different as the imposition of the required solid angle of observation by the imager itself dominates. Nevertheless the dispersing device must be capable of meeting the same solid angle requirement. This can always be achieved if the instrument is sufficiently large or if appropriate mirror scanning methods are used. It is expected that the approaches to optical remote sensing will stabilize in the coming decade.

See also: Chemistry of the Atmosphere: Observations for Chemistry (Remote Sensing): IR/FIR (Satellite, Balloon and Ground); Observations for Chemistry (Remote Sensing): Lidar; Observations for Chemistry (Remote Sensing): Microwave. **Satellites and Satellite Remote Sensing:** Aerosol Measurements; GPS Meteorology; Measuring Ozone from Space: TOMS and SBUV; Precipitation; Surface Wind and Stress; Temperature Soundings; Water Vapor.

Further Reading

- Bernath, P.F., et al., 2005. Atmospheric Chemistry Experiment (ACE): Mission overview. *Journal of Geophysical Research* 32, L15S01. doi:10.1029/2005GL022386.
- Connes, P., 1958. Spectromètre Interférentiel à Sélection par l'Amplitude de Modulation. *Journal de physique et Radium* 19, 215–222.
- Englert, C.R., Babcock, D.D., Harlander, J.M., 2007. Doppler asymmetric spatial heterodyne spectroscopy (DASH): concept and experimental demonstration. *Applied Optics* 46, 7297–7307.
- Harlander, J., Reynolds, R.J., Roesler, F.L., 1992. Spatial heterodyne spectroscopy for the exploration of diffuse interstellar emission lines at far-ultraviolet wavelengths. *The Astrophysical Journal* 396, 730–740.
- Hays, P.B., 1990. Circle to line interferometer optical system. *Applied Optics* 29, 1482–1489.
- Hays, P.B., Abreu, V.J., Dobbs, M.E., Gell, D.A., Grassl, H.F., Skinner, W.R., 1993. The high resolution Doppler imager on the Upper Atmosphere Research Satellite. *Journal of Geophysical Research* 98, 10713–10723.
- Jacquinot, P., 1950. New developments in interference spectroscopy. *Reports on Progress in Physics* 23, 267–312.
- Llewellyn, E.J., et al., 2004. The OSIRIS instrument on the Odin spacecraft. *Canadian Journal of Physics* 82, 411–422.
- McElroy, C.T., et al., 2007. The ACE-MAESTRO instrument on SCISAT: description, performance, and preliminary results. *Applied Optics* 46, 4341–4356.
- Shepherd, G.G., 2002. *Spectral Imaging of the Atmosphere*. Academic Press, London.
- Shepherd, G.G., Thuillier, G., et al., 1993. WINDII: the wind imaging interferometer on the upper research satellite. *Journal of Geophysical Research* 98, 10725–10750.

Airglow Instrumentation

M Conde, University of Alaska Fairbanks, Fairbanks, AK, USA

© 2003 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 5, pp 2078–2095, © 2003, Elsevier Ltd.

Introduction

The airglow and, in polar regions, the aurora are faint and generally diffuse optical glows originating from photochemical reactions occurring in the Earth's upper atmosphere at altitudes above ~ 70 km. During daytime and twilight there is an additional source of airglow due to resonant scattering of sunlight. The airglow provides a means of remote sensing the conditions within its atmospheric source region. Optical instruments are used to measure properties of the emitted radiation, from which properties of the atmosphere may be inferred.

All optical instruments measure light intensity and, by repeated measurement, can also measure the time variation of intensity. Further, most instruments are designed to resolve the intensity distribution of the incident radiation with respect to one or more other parameters such as viewing direction, wavelength, or polarization state.

Airglow processes emit light over a wavelength range from ultraviolet through infrared. The propagation and detection requirements for light vary greatly over this range. Also, the dispersing components that are required differ, depending upon the spectral character of the incident radiation and upon how it is to be resolved. Together, these considerations have resulted in an enormous variety of optical instrumentation for studying the airglow. Some of the major instrument types will be described here, and a brief discussion of some important design considerations will be presented.

Common Instruments

Photometers

Photometers, which measure optical brightness within a single field of view, are the simplest optical instruments for measuring the airglow. Most photometer applications include a narrow-band filter, to isolate a single spectral emission feature. Ideally, a narrow-band photometer would also sample some wavelength adjacent to the emission line of interest, to estimate the brightness of background light that arises for example from the Moon, streetlights, twilight, etc. Many remote sensing techniques use ratios of brightness of two or more emission lines. The required measurements can be made either by swapping several filters in a single photometer or by operating several complete photometers in parallel.

Although applications exist for wide-field photometers, fields of view subtending at most a few degrees are more common. Such narrow-field photometers may be placed on steerable mounts, or be set to view in some fixed direction (typically in either the geographic or the geomagnetic zenith.) One very common configuration is to arrange for one or more narrow-field, narrow-band photometers to view the sky via a spinning mirror that sweeps their field(s) of view along a north–south meridian. This yields a sequence of latitudinal

'cuts' through the two-dimensional airglow or auroral luminosity pattern, as depicted in [Figure 1](#).

2-D Imagers

Airglow and auroral imagers are similar to photometers, except that they use a two-dimensional imaging detector to resolve the airglow brightness with respect to two angular directions in the sky. Any field of view is possible, up to a 180° all-sky view. Once again, an airglow imager operating at visible wavelengths would normally use a narrow-band filter to select a particular emission line. In the near infrared it is possible to use a broader bandpass filter to select groups of lines from OH rotational bands.

Spectrometers

Simple spectrometers measure the spectral distribution of incident radiation from a single field of view. Various optical components can be used to disperse the light according to wavelength, including prisms, gratings, Fabry–Perot etalons, or Michelson interferometers. Two levels of resolution are useful for airglow and auroral applications. For applications that merely require line intensities, spectral resolutions of 0.1 to several nanometers are adequate. Much higher resolutions (3 pm or less) are required to determine the Doppler shifts and Doppler widths of spectral lines. These results can then be used to estimate upper-atmospheric winds and temperatures respectively.

The world's most sensitive optical spectrometers are those attached to large astronomical telescopes. During observations, celestial targets are usually positioned upon the spectrometer slit so that a portion of the adjacent 'dark sky' is also included; this is used for background subtraction from the celestial spectrum. As the telescope must look through the upper atmosphere, the background measurement also yields an airglow spectrum with unparalleled sensitivity (it is not uncommon for a large telescope to integrate a single spectrum for several hours). This technique has allowed emission lines whose very existence was once a mere theoretical prediction to be observed for the first time in airglow spectra.

Hyperspectral Imagers

Hyperspectral imagers combine the spectral resolution of spectrometers with the angular resolution of conventional imagers. That is, one- or two-dimensional images are recorded, with the spectral distribution of the source radiation independently resolved at each image pixel. Subsequently, the spectrum recorded in each pixel is analyzed to derive an estimate of some geophysical parameter (such as atmospheric wind or temperature). These parameter estimates can then be projected back onto the longitude and latitude locations of the original image pixels. The result is a geographic map of

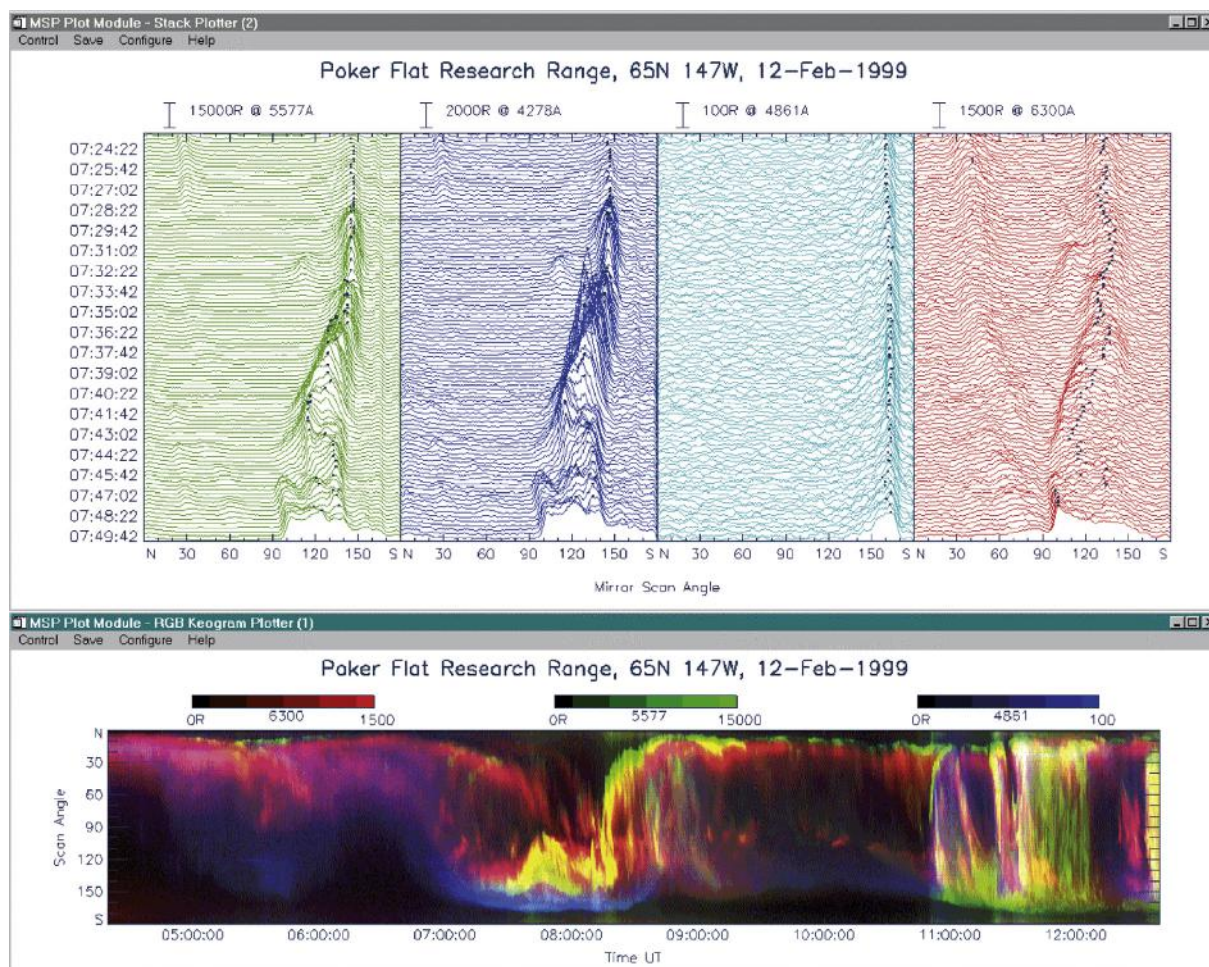


Figure 1 Data recorded by a four-channel meridian-scanning photometer. The four upper panels show, for four different optical wavelengths, the brightness of airglow and auroral emissions as functions of viewing angle along the local geomagnetic meridian. A viewing angle of 0° corresponds to the geomagnetic north horizon, whereas 180° corresponds to the south horizon. Brightness data from 100 successive mirror scans are shown in a 'stacked' format, with corresponding sample times indicated along the left-side axis. The lower panel shows a longer time-history for three of the four channels. The horizontal axis in this presentation represents time, whereas the vertical axis represents viewing angle along the meridian. Each pixel in this image is brightness-modulated in three colors. The red brightness is proportional to the measured sky brightness at $\lambda = 630.0$ nm. Conversely, green and blue colors indicate the measured brightness at $\lambda = 557.7$ nm and $\lambda = 486.1$ nm, respectively.

the spectrally derived parameter values. **Figure 2** is an example of spectra of the $\lambda = 630.0$ nm thermospheric emission line, recorded by an all-sky-viewing Fabry-Perot hyperspectral imager.

Daytime Airglow Spectrometers

During the day, airglow emissions are seen from the ground superimposed upon a blue-sky background of scattered sunlight, which, for small wavelength intervals, exhibits approximately the same spectrum as direct sunlight. This atmospherically scattered sunlight is not only spectrally complex; it is also very much brighter than the airglow. For example, the left panel of **Figure 3** shows that the $\lambda = 630$ nm daytime airglow feature appears superimposed on a corresponding Fraunhofer absorption line, but with only about 1% to 2% of the brightness of the solar continuum.

It is possible to isolate the daytime airglow spectrum, by subtracting a suitably normalized version of the solar background spectrum (**Figure 3(b)**). However, this requires a spectrometer with a single passband and extremely high spectral resolution ($R \geq 200000R$ for the $\lambda = 630$ nm daytime airglow). These requirements can be met by a multiple-etalon Fabry-Perot spectrometer such as the one used to obtain **Figure 3** or, possibly, by a grating spectrometer operating in Echelle mode to achieve the maximum possible wavelength dispersion.

Commonly Observed Wavelengths

Airglow and auroral spectra are complex, exhibiting many discrete emission lines. **Table 1** shows some of those most commonly observed by airglow and auroral instruments.

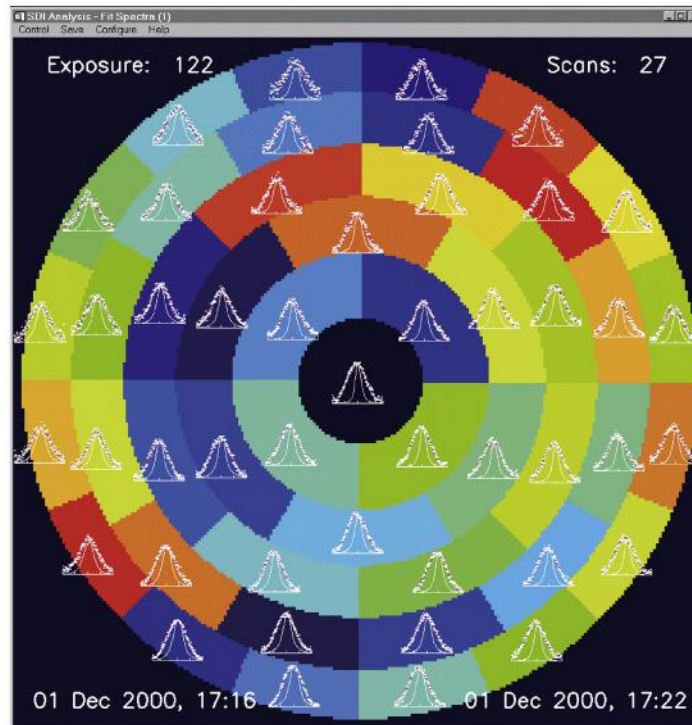


Figure 2 Spectra of the atmospheric $\lambda = 630.0$ nm optical emission recorded by a ground-based all-sky-viewing Fabry–Perot hyperspectral imager. The circular region represents the instrument’s field of view in the sky. Signal processing in the instrument divides this field of view into 47 independent regions. Here, the regions are rendered using different colors, allowing the extent of each one to be discerned. The central region maps to viewing the sky in the zenith, whereas the outer regions map to viewing directions close to the horizon. An independent spectrum of the $\lambda = 630.0$ nm emission is measured from each region, spanning a wavelength interval of only ~ 10 pm. These measured spectra are depicted using small white crosses. The magenta curves depict numerically fitted model spectra, whereas the continuous white curves depict the spectra obtained by viewing a helium–neon laser. The increased width of the sky spectra compared with the laser spectra can be used estimate the temperature of the emitting atoms in the upper atmosphere.

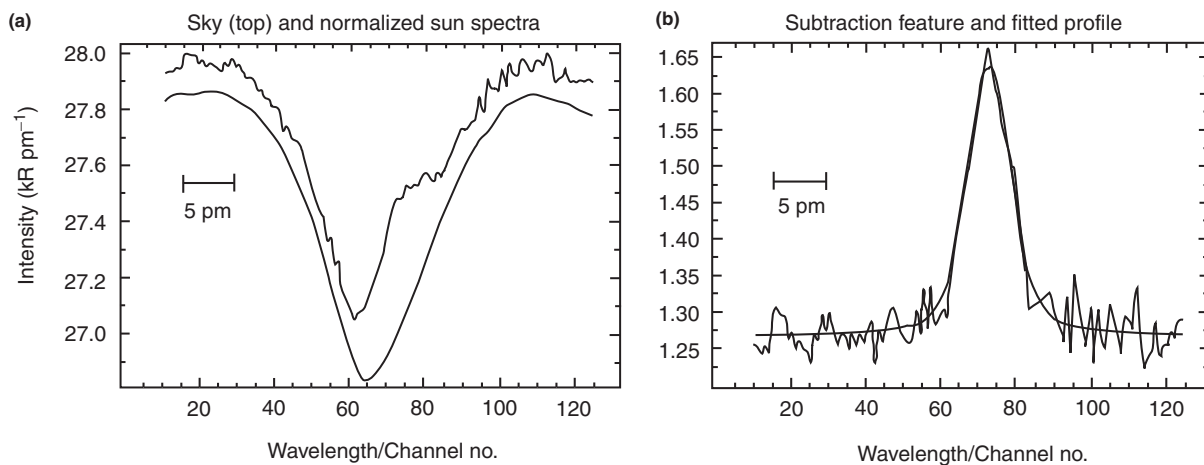


Figure 3 (a) Spectra of sunlight and of daytime blue skylight spanning a wavelength interval of ~ 40 pm and centered near $\lambda = 630.0$ nm. The broad dip seen in both spectra is due to a Fraunhofer absorption line. The smaller hump in the sky spectrum roughly centered on channel 70 is due to daytime airglow. Note that the zero intensity level lies far below the bottom axis in this figure. (b) The spectrum of daytime airglow, obtained by subtracting the sun spectrum from the sky spectrum. The smooth curve superimposed upon the measured data is the result of fitting a numerical model spectrum.

Table 1 Spectral emission lines commonly observed by airglow or auroral optical instruments

Wavelength (nm)	Species	Height	Comments
557.7	O	97–140 km	97 km for airglow, higher for aurora
630.0	O	180–250 km	Airglow and aurora
427.8	N ₂	95–110 km	Aurora only
589.0/589.6	Na	90 km	Emitted by sodium atoms ablated from meteoric dust particles
840.0	OH	87 km	One of many OH lines
732.0	O ⁺	200–250 km	Useful for ion drift
486.1	H	100–120 km	Proton aurora

Viewing Locations

The vast majority of airglow instruments are ground-based, and offer long-duration measurements at minimum cost. However, ground-based observations are limited in several important ways, including the following:

- A ground-based site can view only a limited geographic region.
- The lower atmosphere is opaque to ultraviolet radiation at wavelengths shorter than ~ 300 nm, making airglow or aurora at these wavelengths unobservable from the ground;
- The upper atmosphere must be viewed through the troposphere, which is often cloudy and, when sunlit, appears much brighter than the airglow (due to scattered sunlight);
- It is difficult to recover an altitude profile of the emission intensity from ground-based viewing, particular for diffuse and featureless airglow layers.

For applications that must overcome one or more of these limitations, instruments can be flown (albeit at considerable expense) on balloons, aircraft, sounding rockets, low-orbiting satellites, or distantly orbiting satellites.

A trend since the early 1990s has been the use of computerized tomography to combine observations from several different viewing locations to recover the three-dimensional structure of airglow or auroral features. This method is an extension of ground-based triangulation that was originally used to determine the heights of atmospheric optical emission features.

Design Considerations for Airglow Instruments

General Characteristics

To facilitate discussion of some general principles, consider the ‘generic’ airglow instrument depicted in [Figure 4](#). All optical instruments must include at least one detector that records the light intensity incident upon it. The detector area may be divided into one or more independent elements, usually termed ‘pixels’. Prior to the 1960s, most optical detection was done with photographic film. Since that time various optoelectronic technologies have largely superseded photographic recording, except in specialist applications. Film continues to excel in applications requiring the combination of a large

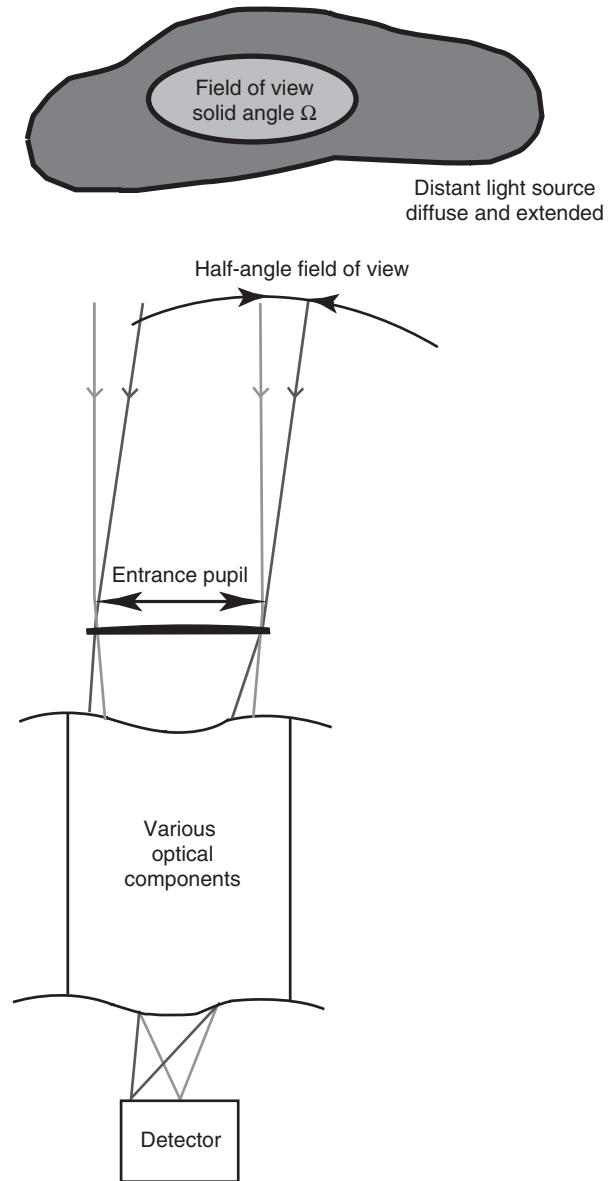


Figure 4 A schematic depiction of a ‘generic’ airglow instrument observing the distant sky. This figure illustrates the detector, the entrance pupil, and the field of view in the sky.

sensitive area (more than ~ 1000 mm²) and a very large number of independent pixels (more than a few million).

Some combination of optical components is placed in front of the detector. Light enters this system through an aperture known as the entrance pupil. Note that the entrance pupil itself need not be a physical object. Frequently, it is merely a virtual image of some physical aperture that is located further back in the optical path.

The optical components relay light from the entrance pupil to the detector(s). Although [Figure 1](#) indicates that these components have formed a sharp image of the sky onto the detector, this need not be the case in general.

An instrument with a nonzero optical throughput must accept radiation over a nonzero range of incidence angles, for

both of the angular dimensions subtended by a distant scene (such as the atmospheric airglow viewed from the ground). The angular region accepted by the instrument is known as its field of view. The solid angle subtended by the field of view at the entrance pupil is usually denoted as Ω .

Light-Gathering Power

Nighttime airglow is extremely dim at visible wavelengths. The ground-level photon flux is comparable to that from a candle at ~ 100 m altitude. Thus, by far the major design issue for any airglow instrument is to achieve adequate sensitivity for the intended measurement.

The light-gathering power of any optical system is limited by a quantity known as its 'etendue', denoted by E . Physically, this is given by the product of the area of the entrance pupil, A , multiplied by the solid angle observed, Ω , i.e.

$$E = A\Omega$$

This fundamental limit determines the maximum number of photons that an instrument can collect per unit time. E is conserved within the optical path, which means that the etendue of a practical instrument is usually determined by limitations of optical components located somewhere in the optical path behind the entrance pupil. The dimensions of E are $(\text{length})^2 \times (\text{angle})^2$, and the conventional unit used is the square centimeter steradian.

Conventionally, the surface brightness of an extended airglow source is measured in units called Rayleighs. The number of photons an instrument with an etendue of $E \text{ cm}^2 \text{ sr}$, viewing a source with a uniform brightness of L Rayleighs throughout its field of view, will gather in one second is given by

$$\Phi = EL \frac{10^6}{4\pi}$$

(Inevitable optical losses within the instrument mean that substantially fewer of these photons will actually be detected.) Note that units of Rayleighs merely indicate a number flux of photons; no restriction is made regarding their spectral distribution. Thus, the brightness of an isolated narrowband spectral line emission can be measured directly in Rayleighs, provided that it is assumed that the photon flux is integrated over all wavelengths within the emission's spectral envelope. By contrast, the brightness of a broadband or continuum source, such as the daytime blue sky, must be specified in units of Rayleighs per unit wavelength interval.

As an example, assume that an instrument with a circular entrance pupil of 50 mm diameter and a circular 1° full-angle field of view is observing the entire spectral envelope of the $\lambda = 630$ nm airglow, which is emitting a typical midlatitude nighttime intensity of 50 Rayleighs. The solid angle viewed is

$$\begin{aligned}\Omega &\approx \frac{\pi\theta^2}{4} \\ &= \pi \times (1^\circ \times \pi/180)^2 / 4 \text{ sr}\end{aligned}$$

and the etendue is

$$E = \pi \times 2.5^2 \times \frac{\pi(1^\circ \times \pi/180)^2}{4} = 4.70 \times 10^{-3} \text{ cm}^2 \text{ sr}$$

The number of photons admitted to the instrument per second is then

$$\begin{aligned}4.70 \times 10^{-3} \text{ cm}^2 \text{ sr} \times 50 \times (10^6/4\pi) \text{ cm}^{-2} \text{ sr}^{-1} \text{ s}^{-1} \\ = 1.8 \times 10^5 \text{ s}^{-1}\end{aligned}$$

No matter what optical components are placed behind the entrance pupil, this is the maximum possible rate of photon detection.

Many airglow instruments apply optical components that disperse the incident radiation into multiple independent 'channels' – separated, for example, according to wavelength or viewing direction. While there are good reasons to do this, it must be understood that each individual channel only receives a corresponding fraction of the total incident photon flux. For example, if a square detector consisting of (say) 256×256 pixels is replaced by one of the same size, but divided into 512×512 pixels, then the modified instrument must observe for four times as long to achieve the same signal level in each pixel.

Optical Losses

No optical components are 100% efficient. Losses of at least a few percent occur at every optical surface owing to absorption, scattering, unwanted reflections, or imperfect geometric alignment. If there are components (such as filters) with limited spectral bandwidth, further losses may occur if these are poorly matched spectrally, either with each other or with the spectrum of the incident radiation. Finally, no detector can record 100% of the photons incident upon it. The fraction that are recorded (known as the 'quantum efficiency') varies between a few percent and $\sim 80\%$ for the types of detector typically used in airglow instruments. Taken together, these losses result usually in an actual rate of photon detection that is only a small fraction of the theoretical maximum based on the system etendue and the source brightness.

Calibration

Depending on the application, absolute calibration of the instrument's wavelength and/or intensity response may be required. Of these two, wavelength calibration is by far the easier, as emission lines from inexpensive spectral lamps provide highly precise and stable wavelength references.

Certain experiments (such as Doppler wind measurement) require wavelength measurements with extremely high precision. For example, the Doppler shift from a wind of 3 m s^{-1} corresponds to a wavelength change of only one part in 10^8 . Unfortunately, several of the airglow lines used for wind measurements are emitted by atomic transitions from long-lived metastable states. (Two examples of this are the $\lambda = 557.7$ nm and $\lambda = 630.0$ nm atomic oxygen lines, whose parent states have radiative lifetimes of ~ 0.7 and ~ 110 s respectively.) It is difficult to obtain a non-shifted wavelength reference for these emissions in the laboratory, because their parent states are quenched by collisions with container walls for any lamp of reasonable size. In the upper atmosphere, by contrast, there are no walls, and radiation from long-lived states is possible. Fortunately, it is usually possible to derive a zero-velocity wavelength reference from the data themselves.

One common method uses the (very reasonable) assumption that the long-term average vertical wind in the upper atmosphere should be close to zero.

Two related calibrations frequently required for spectrometry are the instrumental wavelength dispersion and the instrument's spectral response function. The dispersion can often be calculated *a priori* from the instrumental design parameters or, if a spectrum can be recorded over a sufficient wavelength range, from the positions of two or more calibration lines. The spectral response function can be obtained by observing a laboratory source known to emit a very narrow spectral feature. Suitable sources include lasers and certain unstructured spectral lines of high-mass atoms, such as the $\lambda = 546.1$ nm emission from mercury-198. If the spectral response is desired at some wavelength far removed from that available for calibration then instrumental design parameters must be used to calculate the required wavelength transformation.

Perhaps surprisingly, intensity calibrations are much more difficult. Instruments are calibrated with respect to secondary references, which are usually incandescent lamps that have themselves been calibrated in a national standards facility. An incandescent lamp approximates a point source, whereas for intensity calibration it is desirable to fill the instrument's field of view with a uniform brightness. Thus, in practice, the instrument is arranged to view lamp radiation scattered off an extended Lambertian screen. Applying calibration data for both the lamp and screen allows calculation of the screen brightness in Rayleighs per unit wavelength. When an instrument views the screen then the resulting signal represents the integral under the product of the screen's emission spectrum multiplied by the instrument's spectral response function. This integral has units of Rayleighs; it is related to the total screen brightness within the instrumental passband by a scalar calibration constant. Thus, provided the instrument's spectral response function is known, this procedure assigns the observed response to a calculated number of Rayleighs, which is the desired intensity calibration. Of course, the procedure must be repeated for all wavelengths of interest. It should also be repeated over a range of screen brightness, as many instruments have a nonlinear intensity response.

Signal Requirements

In an analogy to the need to observe a nonzero solid angle, every measurement must be made by integrating the incident photon flux for a nonzero interval of time. A primary figure of merit for any optical instrument used in low-light applications is the exposure time required per measurement – the shorter the time, the better the performance. This time is limited ultimately by the Poisson statistics associated with counting discrete photons. That is, the number of photon detections, N , needed to measure an intensity to a precision of p percent must satisfy

$$p \leq \frac{100\sqrt{N}}{N} \quad \text{or} \quad N \geq \left(\frac{100}{p}\right)^2$$

For example, if we wanted to measure intensity to a precision of 2%, we would need to count 50^2 photons. If the instrument that we considered earlier had an overall detection efficiency of

10%, then it would require an integration time of $50^2/(0.1 \times 18 \text{ kHz})$, or 1.4 s, to achieve this precision for the assumed 50 Rayleigh source.

Frequently, there are other sources of intensity uncertainty beyond those arising solely from the Poisson statistics of counting the signal photons. A widely applicable expression for the actual measurement uncertainty after t seconds of photon counting is

$$\sigma_N^2 = st + Bt + R^2$$

where s is the detection rate of signal photons, B is the occurrence rate of 'background' counts, and R is a one-time noise penalty incurred in certain detectors during read-out. Background counts can arise from stray light entering the instrument, from electronic noise within the detector, or from both. In this case, the integration time required to measure the signal intensity to a precision of p percent must satisfy the relation that

$$\frac{100\sqrt{st + Bt + R^2}}{st} \leq p$$

so

$$t \geq 100 \times \frac{50s + 50B + \sqrt{(2500s^2 + 5000sB + 2500B^2 + p^2s^2R^2)}}{p^2s^2}$$

Applying this to the earlier example, but now with $B = 500$ Hz and $R = 20$ counts, would increase the required integration time from 1.4 to 1.9 s. Clearly, longer integration times are needed as additional sources of noise become significant. Setting both $B = 0$ and $R = 0$, and substituting $st = N$, shows that this expression then reduces to that for Poisson statistics of the signal photons alone.

The Dimension Problem

Typical optical detectors resolve at most two dimensions at once (by recording an image). However, many airglow applications require resolution over more dimensions than this. For example, the Doppler width of an airglow emission line can be used to infer atmospheric temperature. A map of Doppler linewidths across a wide field of view in the sky would yield a corresponding map of atmospheric temperature. However, to achieve this in practice requires recording a complete line spectrum for each pixel in the sky image. Observing the time evolution of a temperature map would require resolution in four dimensions – time, wavelength, and two angular dimensions across the sky.

It is often possible to map the detector's imaging dimensions onto any two of the desired measurement dimensions. However, to resolve more than two dimensions we must use some form of multiplexing, either in time or else across the detector spatially.

Some imaging devices can actually resolve a third dimension, at least crudely. Color film is an example of this, as are new superconducting tunnel junction devices that can resolve wavelength by counting the number of Cooper pairs that are dissociated per incident photon.

See also: **Lidar:** Atmospheric Sounding Introduction. Magnetosphere. **Mesosphere:** Ionosphere. **Observations Platforms:** Rockets. **Optics, Atmospheric:** Optical Remote Sensing Instruments. **Radiation Transfer in the Atmosphere:** Ultraviolet Radiation. **Satellites and Satellite Remote Sensing:** Research. Thermosphere.

Malacara, D. (Ed.), 1988. Physical Optics and Light Measurements. Academic Press, San Diego, CA.
Vaughan, J.M., 1989. The Fabry–Perot Interferometer: History, Theory, Practice and Applications. Hilger, Bristol.
Workman, J., Springsteen, A.W., 1998. Applied Spectroscopy: A Compact Reference for Practitioners. Academic Press, San Diego, CA.

Further Reading

Chamberlain, J.W., 1961. Physics of the Aurora and Airglow. Academic Press, New York.
Parker, S.P. (Ed.), 1988. Optics Source Book. McGraw-Hill, New York.

OZONE DEPLETION AND RELATED TOPICS

Contents

Long-Term Ozone Changes

Ozone as a UV Filter

Ozone Depletion Potentials

Photochemistry of Ozone

Stratospheric Ozone Recovery

Surface Ozone Effects on Vegetation

Surface Ozone (Human Health)

Long-Term Ozone Changes

NRP Harris, University of Cambridge, Cambridge, UK

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

Continuous measurements of ozone have been made for several decades, with the longest record starting in 1926. While the longest records are from ground-based stations, global coverage was only achieved in 1979 with instruments carried on satellites. Together these show the following general features. In polar regions, severe ozone depletion occurs over Antarctica each Austral spring (the Antarctic ozone hole), while only in some years do large ozone losses occur over the Arctic in Boreal spring. Smaller long-term decreases over the midlatitudes in both hemispheres, with the larger losses occurring in the Southern Hemisphere. There are no significant ozone trends over the tropics. Changes in chemical composition, particularly chlorine and bromine, and changes in atmospheric dynamics have influenced ozone. It is important to distinguish between these and other factors in order to assess the effect of the international phasing out of the emissions of ozone-depleting substances and the possible influence of climate change on stratospheric ozone and UV radiation.

Background

Accurate determination of the magnitude and nature of trends in ozone is needed if the relative contributions of the various possible causes of decadal changes in ozone are to be distinguished. The Antarctic ozone hole is a good example where ozone measurements provided tight constraints on the possible causes of the rapid springtime depletion. High-quality measurements of the total ozone column (integrated from the ground to the top of the atmosphere) since the 1950s provided an excellent climatology, which showed that the decline first observable in the late 1970s was a new phenomenon and not part of the natural variability. These same measurements showed that total ozone dropped rapidly from the end of August and that extensive loss did not take place in the polar night, but only upon the return of sunlight. Measurements of the vertical distribution of ozone soon confirmed the latter point and, more importantly, showed that the loss occurred in the lower stratosphere, principally at altitudes between 15 and

20 km. By early October, the atmosphere between 14 and 21 km altitude in the core of the vortex is practically devoid of ozone. The average total ozone columns in Antarctic spring (September to November) in the 1990s were about 60% lower than those in the 1960s. During the 1990s and 2000s, the Antarctic ozone hole of late winter and spring continued unabated, with a single exception in 2002. That year apart the final warming of the vortex, which effectively marks the end of the ozone hole, has taken place a month or so later in the last two decades than in the 1960s to the 1980s. These observations tightly constrain possible explanations, especially when allied with the chemical and aerosol measurements made since the 1970s.

There is no other region where available ozone measurements provide such tight constraints. While this is partly due to the large size of the ozone loss over Antarctica that dwarfs the uncertainties in the ozone measurements and partly due to the relative homogeneity inside the Antarctic vortex, it is also testament to the perseverance of the people who made the

measurements in such an adverse environment. A record of similarly high quality and duration is not available for most of the globe, and our knowledge of global ozone trends is limited by the length and quality of the available ozone record. This limitation is compounded by the smaller magnitude of the trends outside Antarctica. Reviews of the understanding of global ozone trends are presented in the series of the World Meteorological Organization (WMO) assessments.

Long-Term Ozone Measurements

The importance of ozone lies principally in its ability to absorb UV-B radiation (light with wavelengths <310 nm). Because of this, little UV-B radiation penetrates to the Earth's surface. A decrease in the total amount of ozone in the atmosphere (most of which is in the stratosphere) would lead to higher UV-B levels at the Earth's surface than occurs naturally and so lead to damage in life-forms because they are not adapted to it. This same property (absorption of UV radiation) can be used to measure atmospheric ozone, with more ozone leading to greater absorption of UV and less ozone leading to less absorption. Most of the measurements referred to here are based on this basic principle. Applying it in practice is not straightforward, as proper account must be taken of factors such as atmospheric particles and sulfur dioxide, which also interact with UV radiation.

Our picture of the multidecadal variations in total ozone comes from the ground-based network, principally Dobson spectrophotometers, which was mainly in the northern midlatitudes in the early years. Reliable records in this region have been available from several locations since the early 1960s. The quality of the global record is patchy, but it has been greatly improved over the last 30 years with the introduction of internationally agreed operating procedures and quality assessment. The longest continuous record of measurements is from Arosa in Switzerland, where measurements of ozone have been made since the 1920s. [Figure 1](#) shows the measurements from Arosa since the 1920s.

It is now over 30 years since the Nimbus 7 satellite was launched which included the TOMS and SBUV sensors in its

payload. The two instruments worked for unexpectedly long times until 1993 and 1990, respectively. Subsequent versions of both sensor types have been flown, and it is possible to combine records from the different instruments, provided great care is taken with their absolute calibrations. In practice, the only independent means of assessing this is by comparison with the ground-based stations, a process that inevitably reduces the true independence of the ground and satellite systems. Other satellite instrument series have been operational since the 1990s. It is important that the various satellite systems are used in complementary ways that serve to increase the quality of the world's ozone observing system. Experience has shown the great value of the well-run, closely calibrated network of ground-based instruments that operate within WMO's Global Ozone Observing System ([Figure 2](#)) and the need to continue this in future. The average global ozone distribution is also shown, with high values at higher latitudes.

The limits on statistical significance imposed by natural variability and particularly by uncertainty in instrumental calibration do not necessarily equate with scientific importance. Two particular instances stand out – ozone over the tropics where a small change in ozone could have large radiative and climatic impact, and ozone in the lower stratosphere which is hard to measure with good precision and accuracy on the global scale.

Ozone Changes

Recent analyses of total ozone data confirm the now familiar features of the stratospheric ozone depletion that has occurred since the 1970s. At midlatitudes in the Northern Hemisphere, there has been a statistically significant decline in the total ozone in all seasons ([Figures 1](#) and [3](#)) with larger reductions in winter and spring than in summer and autumn. The loss of ozone at southern midlatitudes is also statistically significant throughout the year, albeit with a smaller seasonal variation. No statistically significant trends in total ozone have been observed in the tropics. These features can all be clearly seen in [Figure 3](#) that shows five different sets of ozone measurements.

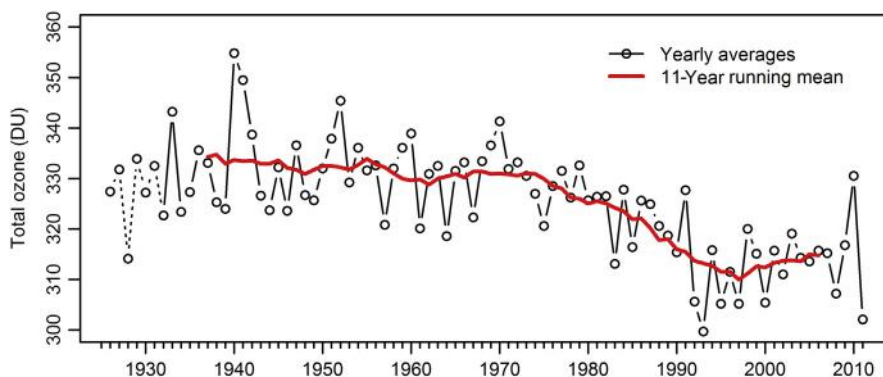


Figure 1 Total ozone measurements at Arosa, Switzerland, from 1926 to 2011. The thin solid line and dots show the annual averages, with a dotted line prior to 1932 to indicate the large fractions of missing data at that time. The thicker line is the 11-year running mean. The measurements are performed by MeteoSwiss. Figure courtesy of H. Rieder.

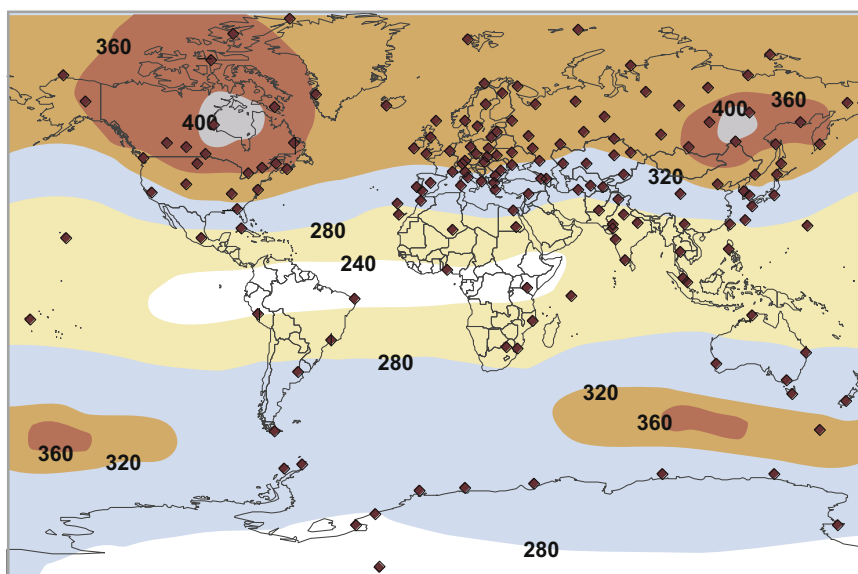


Figure 2 The average global ozone distribution with low values in the tropics and higher values at higher latitudes. The diamonds illustrate the ground-based stations where ozone measurements are made within the World Meteorological Global Atmospheric Watch (WMO-GAW) network. Courtesy of WMO.

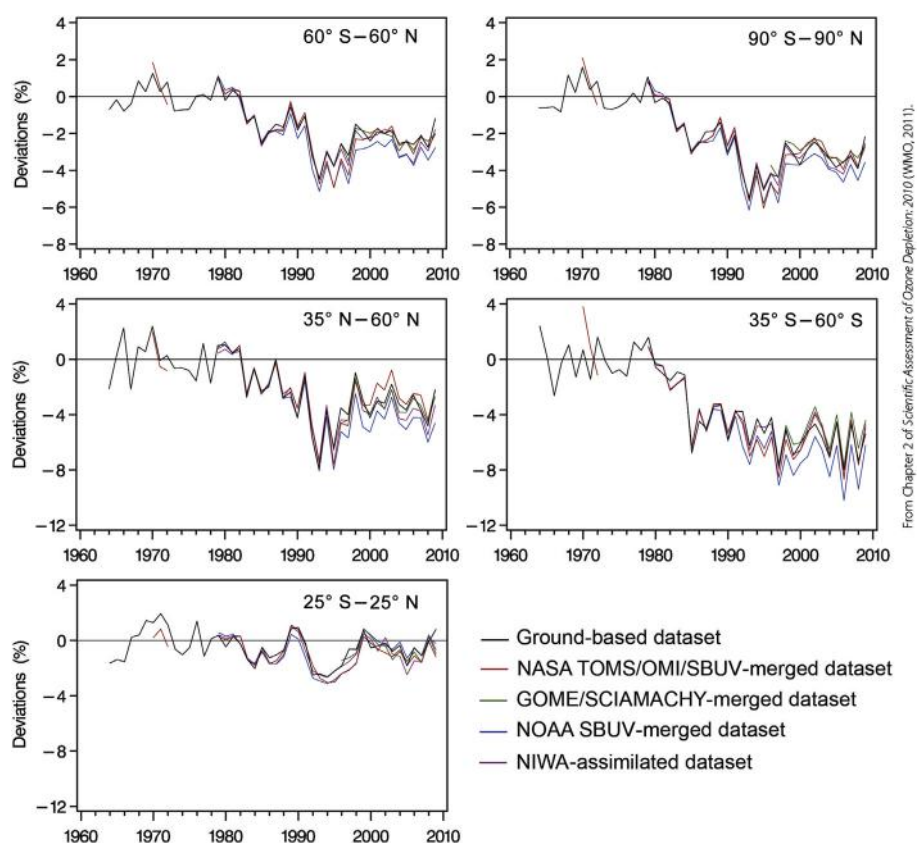


Figure 3 Annual mean area-weighted total ozone deviations from the 1964–80 means for five latitude bands, estimated from different global data sets. Deviations are expressed as percentages of the ground-based time average for the period 1964–80. Figure from Chapter 2 of WMO Scientific Assessment of Ozone Depletion: 2010. World Meteorological Organisation Global Ozone Research and Monitoring Project, Report No. 52, Geneva, Switzerland, 2011.

These deseasonalized time series show remarkable consistency with excellent stability and agreement to about $\pm 1\%$.

The top two plots (60°S – 60°N and 90°S – 90°N) are different measures of the global total ozone amount: both are used because there are only a few measurements in polar regions and these are overwhelmingly made in summer when sunlight is present. Both show a decrease in total ozone, starting around 1980. The largest decrease (4–6%) occurred in the early to mid-1990s, which is ascribed to the presence of large amounts of volcanic aerosol following the eruption of Mt Pinatubo in 1991. However, since the late 1990s (when the effects of Mt Pinatubo had largely dissipated), the global total ozone amount has not changed much, being 2–4% below the pre-1980 levels.

A similar and somewhat magnified behavior is observed in the record from northern midlatitudes (35°N – 60°N) where the influence of Mt Pinatubo is seen most strongly. However, there is a marked contrast with the behavior observed over southern midlatitudes (35°S – 60°S) where there is no increase in the years following the Mt Pinatubo eruption. In recent years, total ozone values in northern and southern midlatitudes have stabilized at respectively about 3.5 and 6% lower than the 1964–80 average. In both hemispheres, there is little sign of increase in recent years.

No statistically significant trends in total ozone have been found in the tropics (25°S – 25°N), though there are indications of a small decrease in the 1990s around the time of Mt Pinatubo.

When looking at changes of a few percent that result from several causes, there is a danger of over-interpreting changes of 1% – after all this is the estimated stability of the observational network for total ozone. It is clear, given the sizes of the changes and the measurement uncertainty, that care must be taken in discussions of the possible ‘recovery’ of the ozone layer. For example, when interpreting the ozone record, the assumptions underlying the analysis must be valid. The statistical models used to calculate ozone trends attempt to account for a number of sources of variations in ozone. The standard approach is to describe the effects of natural phenomena such as the quasi-biennial oscillation (QBO) and the 11-year solar cycle by assuming a linear relation between the strength of these phenomena and total ozone. The ozone trends used to be assumed to be linear, but this is no longer the case now that the concentration of ozone-depleting substances (ODSs) is decreasing. The details of the statistical analysis would be of solely academic interest if the relationship of total ozone with the various influences was well established, but in practice this is not the case. The impact of volcanic eruptions is particularly hard to assess in this regard as their impact depends on location, strength, and timing as well as the level of ODS in the atmosphere.

Having said all that, there is clearly considerable scientific and public interest in being able to identify a recovery in ozone as a result of the ODS reductions occurring after the implementation of the Montreal Protocol. One of the places where this signal might be most easily discernible is the upper stratosphere, where (1) the effects of the ODS-related chemical ozone loss are relatively large and well understood; and (2) the measurements are of high quality. Figure 4 shows a composite of ground-based and satellite measurements taken at five

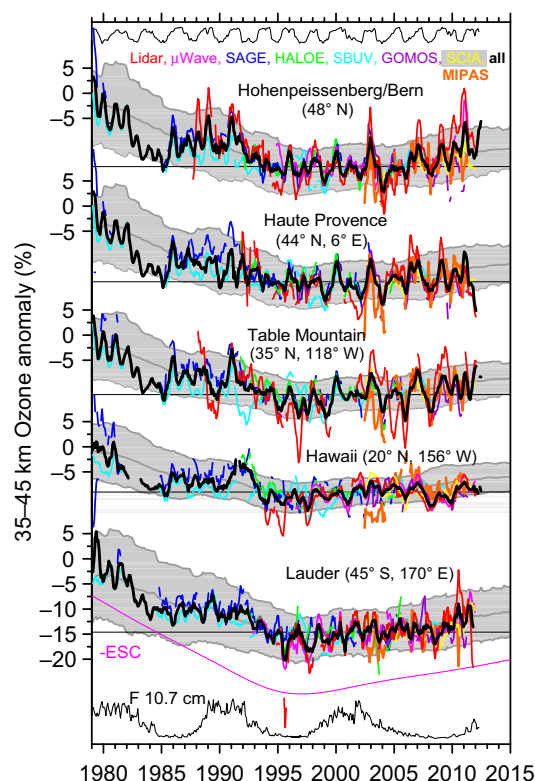


Figure 4 Ozone anomalies at altitudes of 35–45 km from 1979 to mid-2012 for eight data sets at five ground-based stations in the Network for the Detection of Atmospheric Composition Change. The gray underlay shows the range for an ensemble of climate-chemistry models. The thin black lines at the top and bottom show negative 10 hPa zonal wind at the equator as a proxy for the quasi-biennial oscillation, and 10.7 cm solar flux as a proxy for the 11-year solar cycle, respectively. The thin magenta line near the bottom shows inverted effective stratospheric chlorine (ESC) as a proxy for ozone destruction by chlorine. Updated from Chapter 2 of WMO Scientific Assessment of Ozone Depletion: 2010. World Meteorological Organisation Global Ozone Research and Monitoring Project, Report No. 52, Geneva, Switzerland, 2011. Figure courtesy of W. Steinbrecht.

ground stations in northern midlatitudes, tropical latitudes, and southern midlatitudes. The most obvious feature is the upper stratospheric ozone decline from 1980 to the mid-1990s. It is also clear that decline has stopped and ozone has stabilized since 1995–96. However, no statistically significant trend is found in this period and it is not yet possible to attribute any change in ozone trends to changes in ODS.

Large changes were also seen at lower altitudes (between 10 and 25 km) in the extratropics with no significant change at 30 km. It is hard to assess the recent changes at these altitudes, partly because there are large natural variations that have strong regional dependencies and partly because the available measurements do not give a coherent, long-term picture. This is a concern because changes of ozone in the tropical lower stratosphere and upper troposphere are of the great potential importance to climate change.

Finally, it is interesting to compare what happens over the Arctic with the dramatic changes that have occurred over Antarctica – and the picture is certainly more complex.

The observed ozone losses in the Arctic vary a great deal from winter to winter, with large losses in cold winters when the processes leading to chemical ozone loss are favored. Large ozone losses (20–30% in the column and 40–70% at altitudes around 18 km) caused by chemistry have occurred in the colder winters. Most notable was the ozone loss in spring 2011, which was the closest yet to an Antarctic ozone hole. Here, over 80% of the ozone between 18 and 20 km was destroyed, and the column ozone loss was similar to that which occurs in the Antarctic. Important differences remain, including the facts that ozone amounts in the Arctic are naturally higher than in the Antarctic (so that a greater loss is required to reach the same level) and the Arctic is not cold every year so that such large losses will remain exceptional.

Discussion

It is clear that chlorine and bromine compounds resulting from ODS breakdown lead to ozone depletion, but it is also clear that they do not account quantitatively for all of the observed trends. There is no doubt that rapid chemical ozone loss occurs over the Antarctic and Arctic in winter/spring. The signal is large and a whole host of chemical measurements allow unambiguous identification of perturbed chlorine and bromine chemistry as the cause. There is equally no doubt that the ozone-depleted air is subsequently mixed over larger areas and leads to reduced ozone amounts over midlatitudes.

Other factors are also involved. Statistical analyses of total ozone trends routinely include terms to describe the influence of the solar cycle and the QBO on total ozone. With over 30 years of global satellite data and over 50 years of ground-based observations – i.e., longer than the periods of these or other natural phenomena that could affect ozone – any total ozone changes resulting from the increase in ODS are reasonably robust to possible errors in the description of these phenomena. However, identification of any recovery (based on 10–15 years of measurements since the ODS maximum) still requires correct attribution of these factors.

In order to separate the various influences, close attention is paid to examining the time series of ozone measurements in order to extract information on possible mechanisms from the observed interannual variations. For example, 2D and particularly 3D models of stratospheric chemistry and dynamics have been used to investigate this, with their simulations including the effect of observed aerosol amounts on the trends calculated for the chlorine and bromine changes and observed changes in stratospheric transport. The calculated ozone changes following the major volcanic eruptions in the 1980s and the 1990s are consistent with variations in the observed total ozone record. In particular, a difference in behavior is found in the northern and southern midlatitudes. The reasons for this difference include a greater loss in the Antarctic that dilutes the ozone amount more in southern midlatitudes that does the smaller loss in the Arctic; a greater effect of Mt Pinatubo on the Northern Hemisphere; and a greater decadal variability in the dynamic influence on ozone in the Northern Hemisphere.

Long-term changes in the circulation of the atmosphere connected with climate change are and will be causing long-

term changes in total ozone. A number of changes are already occurring that will have important consequences for the stratosphere and stratospheric ozone. For example, changes in temperature have been observed, with episodic jumps associated with volcanic eruptions superimposed on a long-term cooling trend; an increase in tropopause height (and decrease in tropopause pressure) has been observed globally which at a simple level means that the stratosphere is getting smaller; and there is a widening of the tropical belt. These changes are all associated with climate change and are all having important, though incompletely understood consequences for stratospheric ozone amounts and distribution.

So, looking to the future, total ozone will be subjected to chemical and dynamical influences. The major chemical influence is from the declining levels of ODS as a result of the success of the implementation of the Montreal Protocol. The secondary chemical influence from is the likely changes in the chemically important, greenhouse gases, methane (CH_4) and nitrous oxide (N_2O): the main uncertainty in predicting these changes arises from uncertainties in their future emissions. However, the dynamical influences associated with climate change could be greater and are currently known with far less certainty.

Unambiguous identification of the effect of the Montreal Protocol on ozone observations has been a ‘holy grail’ for a few years as there is real value to providing definitive proof of a success of environmental legislation. However, with so many competing influences on ozone to account for, it is extremely hard to achieve. Scientists have looked at regions where the influence of ODS-related chemistry is particularly strong such as the upper stratosphere (Figure 4) and the Antarctic ozone hole. To date, they have not succeeded. In the upper stratosphere, for example, the observed cooling affects the speed of the ozone chemistry in such a way that more ozone is calculated to be present when the ODS are gone than before they were introduced. The sustained increase in N_2O concentration is perturbing the chemistry further. The Antarctic may be a better place to look, but even there it is likely to be decades before an ozone recovery can be rigorously identified.

See also: Chemistry of the Atmosphere: Observations for Chemistry (*In Situ*): Ozone Sonde. **Ozone Depletion and Related Topics:** Ozone Depletion Potentials; Ozone as a UV Filter; Photochemistry of Ozone; Surface Ozone (Human Health); Surface Ozone Effects on Vegetation.

Further Reading

- Farman, J.C., Gardiner, B.G., Shanklin, J.D., 1985. Large losses of total ozone in Antarctica reveal seasonal ClOx/NOx interaction. *Nature* 315, 207–210.
- Harris, N.R.P., Kyrö, E., Staehelin, J., Brunner, D., Andersen, S.B., Godin-Beekmann, S., Dhomse, S., Hadjinicolaou, P., Hansen, G., Isaksen, I., Jrrar, A., Karpetchko, A., Kivi, R., Knudsen, B., Krizan, P., Lastovicka, J., Maeder, J., Orsolini, Y., Pyle, J.A., Rex, M., Vanicek, K., Weber, M., Wohltmann, I., Zanis, P., Zerefos, C., 2008. Ozone trends at northern mid- and high latitudes – a European perspective. *Annals of Geophysics* 26, 1207–1220. www.ann-geophys.net/26/1207/2008/.

- Manney, G.L., et al., 2011. Unprecedented Arctic ozone loss in 2011. *Nature* 478, 469–475. <http://dx.doi.org/10.1038/nature10556>.
- Solomon, S., 1999. Stratospheric ozone depletion: a review of concepts and history. *Reviews of Geophysics* 37, 275–316.
- Stahelin, J., Harris, N.R.P., Appenzeller, C., Eberhard, J., 2001. Ozone trends: a review. *Reviews of Geophysics* 39, 231–290.
- WMO Scientific Assessment of Ozone Depletion: 2010. World Meteorological Organisation Global Ozone Research and Monitoring Project, Report No. 52, Geneva, Switzerland, 2011.

Relevant Websites

- http://www.antarctica.ac.uk/about_antarctica/geography/ozone.php — BAS ozone page.
- <http://ozonewatch.gsfc.nasa.gov/> — NASA OZONEwatch.
- <http://www.ndsc.ncep.noaa.gov/> — NDACC.
- <https://www.wmo.int/pages/prog/arep/gaw/ozone/index.html> — WMO ozone page.

Ozone as a UV Filter

JE Frederick, The University of Chicago, Chicago, IL, USA

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

Atmospheric ozone absorbs incoming solar radiation, where the attenuation is greatest in the ultraviolet part of the spectrum. The absorbing properties of ozone are expressed in values of the absorption cross section, and these provide the basis for computing the transmission of the Earth's atmosphere as a function of wavelength. As a consequence of attenuation by ozone, essentially no solar energy reaches the Earth's surface at wavelengths shorter than 295–300 nm, while ground-level radiation at wavelengths between 300 and 315 nm is sensitive to the ozone content of the atmosphere.

Introduction

Historically, interest in atmospheric ozone (O_3) emphasized the molecule's role in absorbing incoming ultraviolet sunlight. As a consequence of the attenuation provided by ozone, little or no solar radiation reaches the Earth's surface at ultraviolet wavelengths between approximately 200 and 300 nm. At shorter wavelengths, from 200 to 240 nm, absorption by the diatomic oxygen molecule (O_2) also takes place, while at wavelengths longer than 300 nm the attenuation provided by ozone weakens, although absorption occurs at wavelengths that extend through the visible region.

The Sun emits radiant energy over a broad range of the electromagnetic spectrum, while the human eye responds only to the portion that lies at visible wavelengths from 400 to 780 nm. Approximately 46% of the Sun's total energy lies in the visible. Another 46% exists in the infrared at wavelengths longer than 780 nm, while only 8% of the Sun's radiant energy exists at wavelengths shorter than 400 nm in the ultraviolet. Despite this relatively modest absolute energy, the ultraviolet is responsible for driving much of the chemistry of the Earth's atmosphere as well as for degradation of artificial materials and effects on living things at the Earth's surface.

In the late nineteenth century, scientists believed that the Sun emitted radiation in the ultraviolet, although the human eye does not respond to energy at these wavelengths. Measurements designed to detect ultraviolet radiation revealed the expected levels of sunlight reaching the ground at wavelengths from approximately 315 to 400 nm, a spectral region called the UV-A. However, at shorter wavelengths the instruments observed an extremely steep decline in energy. This spectral region, which is marked by a precipitous drop in ground-level radiation as wavelength decreases, is known as the UV-B. In standard usage the UV-B encompasses wavelengths from 280 to 315 nm, although for practical purposes, no solar energy reaches the Earth's surface at wavelengths shorter than 295–300 nm.

The scarcity of UV-B radiation at the ground presented a dilemma for scientists who first observed this behavior. The steep drop in radiant energy with decreasing wavelength was not consistent with the spectrum of radiation emitted by hot objects observed in the laboratory. The visible radiation received from the Sun suggested an object whose absolute temperature was in the vicinity of 5700–5800 K, but a radiator at this temperature should have a larger emission in the UV-B than was

present in sunlight at the Earth's surface. The resolution to the problem was that the Sun did indeed emit energy in the UV-B portion of the spectrum, but somewhere between the Sun and the surface of the Earth much of this energy was absorbed.

In 1881 the English chemist William Hartley created ozone in the laboratory by shooting electrical sparks through a container of dry air. He then directed ultraviolet light from a lamp through this ozone. He observed that the ozone absorbed the very same ultraviolet wavelengths that were absent from sunlight at the ground. Both the spontaneous formation of ozone in the laboratory and the selective wavelength-dependent absorption of ultraviolet radiation by this ozone led Hartley to suggest that ozone formed in the Earth's atmosphere, and absorption by ozone was responsible for the near absence of the UV-B component of sunlight at the ground. Subsequent studies in the early twentieth century demonstrated that atmospheric ozone resided somewhere above the tops of tall mountains, and optical methods eventually confirmed that most of the ozone exists in the stratosphere, between 10 and 50 km in altitude.

Figure 1 illustrates the ground-level solar ultraviolet spectral irradiance over the wavelength range 290–340 nm. Irradiance

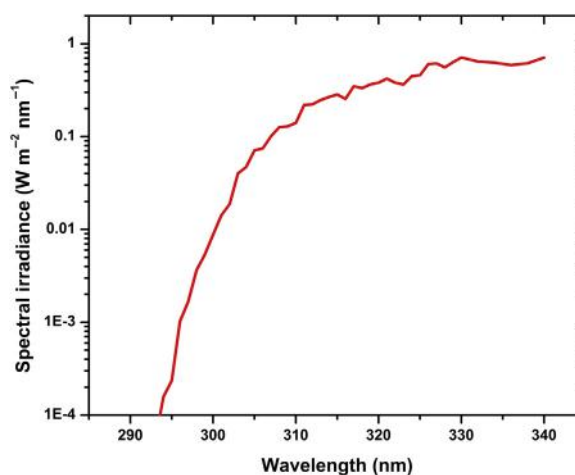


Figure 1 Ultraviolet solar irradiance striking a horizontal surface at the ground over the wavelength range 290–340 nm. The values, based on model calculations, refer to latitude 40° N at solar noon in mid-July under clear skies.

is defined as the energy per unit time that strikes a unit horizontal area per unit wavelength interval, where the typical unit is $\text{W m}^{-2} \text{ nm}^{-1}$. These results come from a numerical model that treats absorption and scattering of sunlight as a function of wavelength as the solar beam passes through the atmosphere. The values in Figure 1 refer to latitude 40° N in mid-July at noon under clear skies. In the UV-B the solar irradiance at the ground decreases by two orders of magnitude as wavelength declines from 315 to 297–298 nm. This behavior results from absorption by atmospheric ozone.

Molecular-Scale Interactions

The absorption of solar UV-B light is a manifestation of processes that occur on the molecular scale. An ozone molecule consists of three oxygen atoms held together by two bonds, which arise from the sharing of electrons. The length of each bond is 0.126 nm, indicating the characteristic dimension of the molecule. Ultraviolet sunlight interacts with the molecule by breaking one of the bonds leading to photodissociation, written as $\text{O}_3 + h\nu \rightarrow \text{O}_2 + \text{O}$. In this process a quantum of light, denoted by $h\nu$, is destroyed and the energy goes into breaking one of the two bonds in the original molecule.

To describe a photodissociation, it is convenient to view a beam of sunlight as if it consists of a stream of massless particle-like entities called photons. The energy of one photon is $h\nu$, where ν is the frequency of the light and h is the Planck's constant. Alternatively, the energy of a photon is hc/λ , where c is the speed of light and λ is the wavelength. To dissociate an ozone molecule, the energy of the photon must exceed a specific minimum amount, called the binding energy ozone, which is determined by the strength of the molecular bond to be broken. There is a one-to-one correspondence between photons absorbed and ozone molecules dissociated; that is, absorption of one photon leads to the dissociation of one ozone molecule. In the process, the photon is annihilated, and its energy is expended in breaking one of the bonds in the ozone molecule. The binding energy of the molecule defines the maximum wavelength of light that is capable of causing photodissociation, called the dissociation wavelength. In the case of ozone, the dissociation wavelength lies in the infrared part of the spectrum, and any photon whose energy exceeds the binding energy can break ozone into O_2 and O . Any excess photon energy appears in kinetic energy or in excited states of the products. In practice, the photodissociation of ozone becomes efficient only for photons that are much more energetic than the binding energy and that correspond to light in the ultraviolet part of the electromagnetic spectrum.

The absorption cross section of ozone describes the interaction of radiation with the molecule and the wavelength-dependent values reflect the strength of absorption across the spectrum. Laboratory measurements are able to determine the magnitude of the cross section, and these empirical values provide a means to compute the transmission of the atmosphere throughout the ultraviolet spectral region. Figure 2 depicts a beam of light incident on a layer of ozone at an angle θ to the vertical. Let $L_0(\lambda)$ be the incident radiance at wavelength λ , and $L_T(\lambda)$ be the transmitted radiance that exits from the bottom of the slab. Radiance has a unidirectional

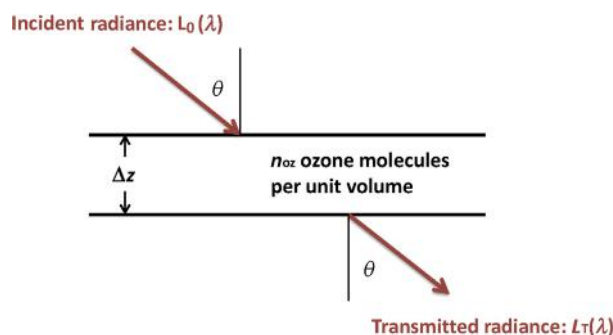


Figure 2 Attenuation experienced by a beam of ultraviolet light in passing through a layer of ozone.

character, whereas irradiance, as in Figure 1, sums over energy propagating downward from all locations in the sky. The layer in Figure 2 has a thickness Δz in centimeters and contains ozone in the amount n_{oz} molecules per cubic centimeter. When absorption is the only process operating, the relationship between the transmitted and incident radiance is given by Beer's law (eqn [1]).

$$L_T(\lambda) = L_0(\lambda) \exp[-\sigma(\lambda) n_{\text{oz}} \Delta z / \cos \theta] \quad [1]$$

The quantity $n_{\text{oz}} \Delta z / \cos \theta$ specifies the number of ozone molecules per unit area measured along the slant path taken by the beam of light through the layer. Equation [1] is essentially a definition of the absorption cross section of ozone, $\sigma(\lambda)$, whose dimension is area, typically expressed in square centimeters. The cross section describes the strength of absorption per ozone molecule as a function of wavelength. If λ_D is the dissociation wavelength, then $\sigma(\lambda) = 0$ for all $\lambda > \lambda_D$, and at these wavelengths a beam of light is unaltered in passing through a layer of ozone. For $\lambda < \lambda_D$ the degree of attenuation depends on the magnitude of the absorption cross section, which varies dramatically with wavelength.

In principle, it is possible to compute the absorption cross section, given a complete knowledge of the molecular structure of ozone, although in practice this is neither practical nor feasible. Instead, applications in atmospheric science make use of absorption cross sections determined empirically. Here a monochromatic light source illuminates a laboratory chamber that contains a known concentration of ozone. Then from measurements of the radiance incident on and transmitted through the chamber, eqn [1] allows deducing the cross section at the wavelength of incident light. By varying the wavelength, one can infer values of the absorption cross section over the entire ultraviolet spectral region of interest in atmospheric applications.

Figure 3 depicts the absorption cross section of ozone measured over the wavelength range 200–340 nm, encompassing the UV-C, defined as wavelengths less than 280 nm, the UV-B, and the shortest wavelengths of the UV-A. The cross section in the 315–340 nm portion of the UV-A is very small compared to that in the UV-B, and a dependence on temperature is apparent. At wavelengths longer than 340 nm the attenuation provided by ozone is negligible, although dissociation is energetically possible throughout the remainder of the ultraviolet and visible portions of the spectrum. The

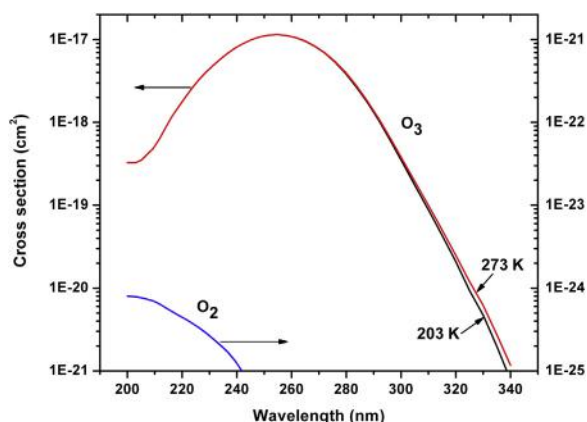


Figure 3 The absorption cross section of ozone and of molecular oxygen over the wavelength range 200–340 nm. The vertical scale on the left applies to ozone and that on the right to molecular oxygen.

wavelengths shown in [Figure 1](#), 290–340 nm, overlap the longer wavelengths in [Figure 3](#). A comparison of the two figures shows that the steep increase in ultraviolet irradiance with wavelength over this spectral range coincides with, and is a consequence of, the dramatic decrease in the absorption cross section of ozone as wavelength increases.

[Figure 3](#) shows a peak in the absorption cross section of ozone near 250 nm, with a sharp decline toward both longer and shorter wavelengths. Based solely on this behavior, one would expect to observe sunlight reaching the Earth's surface near 200–210 nm, since the cross section at these wavelengths is close to that at 300 nm. In fact, no solar radiation survives passage through the atmosphere at 200–210 nm because at wavelengths shorter than 240 nm molecular oxygen begins to absorb, and the photodissociation $\text{O}_2 + h\nu \rightarrow \text{O} + \text{O}$ occurs. Although the absorption cross section of O_2 , also shown in [Figure 3](#), is extremely small, a factor of one hundred thousand to one million less than that for ozone (see the right-hand scale for O_2), the atmosphere contains a very large amount of O_2 . The consequence of the combined absorption by O_2 and O_3 is that incoming sunlight over the entire UV-C is removed at altitudes far above the ground. The absorption by O_2 in photodissociation becomes stronger as wavelength decreases below 200 nm, and at much shorter wavelengths, less than 100 nm, absorptions by molecular nitrogen (N_2) and O_2 lead to ionization in the Earth's atmosphere at altitudes greater than 100 km. All of the above processes combined ensure that the surface of the Earth is totally shielded from the highest-energy photons in the solar spectrum.

Transfer of Solar Radiation through the Atmosphere

The transfer of UV-B and UV-A sunlight through the atmosphere is more complicated than depicted in [Figure 2](#) and [eqn \[1\]](#) because absorption by ozone is not the only process that alters the incoming solar beam. In passing through an atmospheric layer, incident light is both partially absorbed and also scattered. Scattering refers to a change in direction of propagation, with no change in the amount of energy

contained in the radiation field. Most atmospheric molecular scattering is by the major gases, N_2 and O_2 , since they are vastly more abundant than ozone. In addition, the atmosphere contains suspended small particles, especially in the lowest several kilometers of altitude, and these are important in scattering radiation.

At the top of the atmosphere, solar radiation at the wavelengths absorbed by ozone is incident as a direct beam coming from one specific direction in space. As this beam propagates downward, it is attenuated to an extent determined by the cross section of ozone at the wavelength considered, by the abundance of ozone molecules, and by scattering. But rather than destroying radiant energy, scattering redirects the sunlight into all possible angles over the upper and lower hemispheres. The area of physics known as radiative transfer develops the concepts and mathematical methods needed to compute the attenuated direct solar radiation and the scattered radiation moving in any direction at each altitude throughout the atmosphere as well as at the ground.

Scattering of radiation by objects much smaller than the wavelength of light, where molecules are prime examples, is called Rayleigh scattering, named after Lord Rayleigh who studied the process in detail. The attenuation of a direct beam of sunlight by Rayleigh scattering is described by an equation analogous to Beer's law ([eqn \[1\]](#)). In this case, the cross section for scattering is inversely proportional to wavelength raised to the fourth power ($1/\lambda^4$), and the appropriate number density is the total number of atmospheric molecules per unit volume in the layer. However, as the direct beam is attenuated, a scattered radiation field, with components propagating in all directions, is created. Rayleigh scattering explains the blue color of the sky on a clear day, since the shortest visible wavelengths, in the blue and violet spectral regions, scatter more efficiently than do longer wavelengths, in the yellow or red. The strong inverse wavelength dependence of the Rayleigh scattering cross section implies that ultraviolet radiation is scattered even more strongly than visible light. As a consequence, on a clear day much of the ultraviolet sunlight received at the ground is in the form of scattered radiation from the entire upper hemisphere as opposed to being concentrated in the direct beam from the Sun. Scattering by particles and liquid droplets acts in a way similar to that by molecules except that the wavelength dependence is much weaker than in the Rayleigh case. When particles and droplets are abundant, the sky takes on a whitish color, an indication that all wavelengths in sunlight are scattering with roughly equal efficiency.

Atmospheric scattering alters the amount of solar ultraviolet energy absorbed by ozone. This can be deduced from [Figure 2](#) and [eqn \[1\]](#). The attenuation of a beam of light depends on the ozone content of the layer in [Figure 2](#) and on the angle θ at which the radiation is incident. In the absence of scattering, all of the radiant energy would be concentrated at the single angle θ defined by the location of the Sun in the sky. With scattering, however, radiation impinges on a layer from all directions. From [eqn \[1\]](#), the attenuation by ozone is different for each value of θ . As a result, calculations of solar ultraviolet irradiance at the Earth's surface must treat the coupled actions of absorption and scattering at all altitudes throughout the atmosphere.

Although models that treat the transfer of ultraviolet sunlight through the atmosphere must include the complications identified above, eqn [1] still identifies the major variables that determine the solar irradiance at the Earth's surface. When applied to the atmosphere, θ in Figure 2 is the solar zenith angle, defined as the angle between the instantaneous position of the Sun and the local vertical. As the Earth rotates, the solar zenith angle at any location varies from 90° at sunrise to a minimum value at midday, called solar noon, and back to 90° at sunset. The apparent motion of the Sun relative to a point fixed on the Earth's surface is accompanied by a changing path length taken by solar radiation through the atmospheric ozone layer. The minimum path length and maximum UV-B irradiance at the ground occur at solar noon. An increase in the solar zenith angle, as occurs in the hours after solar noon, is accompanied by an increase in absorption by ozone and a decrease in ground-level UV-B irradiance. In addition to the changing magnitude of absorption, simple geometrical factors are significant as well. A fixed amount of solar energy is spread over a larger horizontal area at the Earth's surface as the angle of incidence, θ , increases. This leads to a decline in the energy received per unit horizontal area irrespective of absorption by ozone.

The dependence of solar zenith angle on latitude, season, and local time leads to large systematic variations in solar ultraviolet irradiance at the ground. Figure 4 illustrates the behavior of irradiance at several wavelengths in the UV-B and UV-A as functions of solar time during a clear afternoon at latitude 40° N in July, where 12:00 denotes solar noon, the time of minimum solar zenith angle. The irradiance at all wavelengths declines rapidly as the Sun gets closer to the horizon. At 340 nm, the irradiance is almost independent of the ozone amount, and the value at noon exceeds that at 17:00 h by a factor of three. This variation arises from increased atmospheric backscattering as the solar zenith angle increases acting in combination with the geometrical effect noted above. As wavelength decreases, absorption by atmospheric ozone becomes increasingly important and leads to a more pronounced dependence of irradiance

on solar zenith angle. At a wavelength of 305 nm, the noontime irradiance exceeds that at 17:00 h by a factor of 19, and at 300 nm this factor increases to 119.

Changes in Atmospheric Ozone Amounts and Ultraviolet Radiation

Concerns over depletion of the ozone layer arose in the 1970s, when scientists realized that chemicals released by various anthropogenic activities could lead to a slow decline in ozone amounts in the stratosphere. The importance of this depletion arose from ozone's role as an absorber of biologically damaging ultraviolet sunlight. Exposure to radiation in the UV-B, and to a lesser extent in the UV-A, is responsible for a variety of negative effects, including certain types of skin cancers, cataracts, and inhibition of plant development. Equation [1] implies that a decline in the number of ozone molecules in a column of the atmosphere would be accompanied by an increase in solar ultraviolet irradiance reaching the Earth's surface. If decreases in the ozone abundance continued over a period of decades, one would expect an increase in the incidence of the biological effects identified above. The relationship between ultraviolet radiation exposure and a specific biological response can be exceedingly complex and is often poorly understood on a quantitative basis. However, the increase in ground-level ultraviolet irradiance associated with a known change in atmospheric ozone content is straightforward to predict.

Figure 5 presents the changes in ultraviolet spectral irradiance reaching the ground for -10 and $+10\%$ changes in the column ozone amount. The vertical scale is the ratio of irradiance computed for an altered ozone amount to that for the unperturbed amount, where the reference state is 40° N latitude at noon in mid-July as shown in Figure 1. The wavelength scale extends from 290 nm, where negligible radiation reaches the Earth's surface, to 340 nm, where absorption by ozone is extremely weak. As wavelength decreases, the sensitivity of

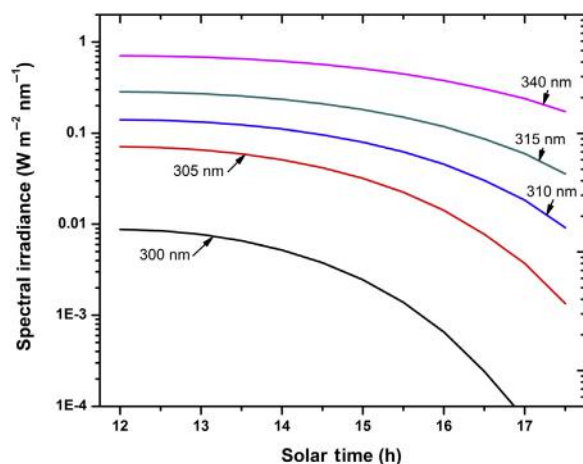


Figure 4 Variation of ground-level solar ultraviolet irradiance at various wavelengths with solar time during an afternoon for latitude 40° N in mid-July under clear skies.

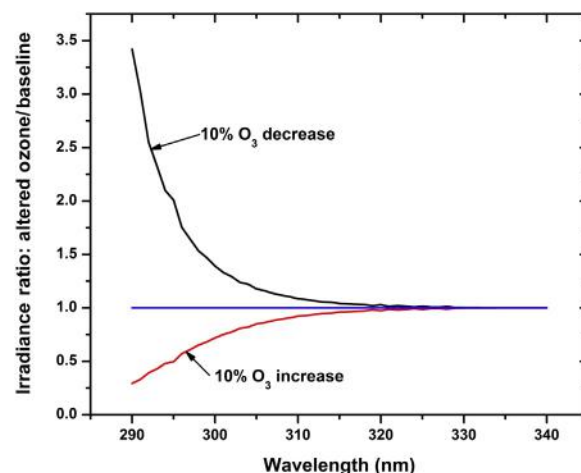


Figure 5 Sensitivity of ground-level ultraviolet irradiance to changes in the column ozone abundance of $+10$ and -10% about the unperturbed amount for latitude 40° N at noon in mid-July.

ground-level radiation to changes in the ozone amount increases rapidly. This type of behavior is qualitatively consistent with eqn [1] combined with the wavelength dependence of the absorption cross section in Figure 3, although realistic quantitative results must include the effects of scattering.

At the short-wavelength end of the UV-A, near 315 nm, a $\pm 10\%$ change in column ozone leads to change in irradiance of less than 5%, where the sign of the change in irradiance is opposite to that in ozone. The sensitivity to changes in ozone weakens with increasing wavelength, and at wavelengths longer than 330 nm, one can neglect absorption. As wavelength decreases from 315 nm through the UV-B, the sensitivity to changes in ozone rises dramatically. For noontime summer conditions, a 10% change in column ozone is accompanied by a 10% change in irradiance of the opposite sign at a wavelength of 308–309 nm. This increases to a doubling or halving near 295 nm, the shortest wavelengths that are readily measurable at the ground.

The largest percentage changes in Figure 5 occur at wavelengths where the absolute irradiances that reach the ground are quite small, and this is a direct consequence of the wavelength-dependent behavior of the ozone absorption cross section. One might question whether the extremely small irradiances in the UV-B are significant in view of the much larger values in the UV-A. Numerous studies have confirmed the biological importance of UV-B radiation at levels that exist in the natural environment. These investigations consistently show that the shorter-wavelength UV-B sunlight is more effective in causing damage to living tissues than are longer wavelengths in the UV-A, although radiation throughout both regions must be considered. When one combines wavelength-dependent biological sensitivities with the spectral variation of solar irradiance at the ground, the UV-B is generally the most effective region in causing biological damage, despite the small absolute energy levels. Hence, the possibility of a major

change in the ozone content of the Earth's atmosphere and the accompanying increase in ground-level UV-B radiation were a legitimate cause for concern.

Although depletion of the ozone layer is no longer a major issue, ultraviolet solar radiation remains an important environmental parameter. The chemistry of the lower atmosphere is driven by the UV-B radiation that survives passage through the stratospheric ozone layer. In this way, absorption of solar ultraviolet radiation by ozone influences the chemical makeup of the air that surrounds us. In addition, long-term exposure to natural levels of ultraviolet sunlight will continue to have biological significance.

See also: **Climate and Climate Change:** Volcanoes: Role in Climate. **Ozone Depletion and Related Topics:** Ozone Depletion Potentials; Photochemistry of Ozone.

Further Reading

- Brasseur, G.P., Orlando, J.J., Tyndall, G.S. (Eds.), 1999. *Atmospheric Chemistry and Global Change*. Oxford University Press, New York.
- Frederick, J.E., 2008. *Principles of Atmospheric Science*. Jones and Bartlett, Sudbury, MA.
- Giese, A.C., 1976. *Living with Our Sun's Ultraviolet Rays*. Plenum Press, New York.
- Seinfeld, J.H., Pandis, S.N., 1998. *Atmospheric Chemistry and Physics*. Wiley, New York.
- van der Leun, J., Tang, X., Tevini, M. (Eds.), 2003. *Environmental Effects of Ozone Depletion and Its Interaction with Climate Change: 2002 Assessment*. United Nations Environment Programme, Nairobi.
- Warneck, P., 1988. *Chemistry of the Natural Atmosphere*. Academic Press, San Diego.
- Weiler, C.S., Penhale, P.A. (Eds.), 1994. *Ultraviolet Radiation in Antarctica: Measurements and Biological Effects*. American Geophysical Union, Washington, DC.
- Worrest, R.C., Caldwell, M.M. (Eds.), 1986. *Stratospheric Ozone Reduction, Solar Ultraviolet Radiation and Plant Life*. Springer-Verlag, Berlin.

Ozone Depletion Potentials

DJ Wuebbles, University of Illinois, Urbana, IL, USA

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

Ozone depletion potentials (ODPs) is an extremely useful metric for policy considerations and scientific analyses for evaluating the relative effects of human emitted gases affecting the global atmospheric environment. ODPs arose as a simplified means for determining the relative ability of a chemical to destroy ozone. This concept for a relative measure of the cumulative impact on stratospheric ozone from trace gas emissions was found to be particularly useful by policymakers in their considerations to develop policy to protect stratospheric ozone from further destruction from emissions of halocarbons.

Introduction

Weighing functions, such as ozone depletion potentials (ODPs), are extremely useful tools in policy considerations and scientific analyses for evaluating the relative effects of human emitted gases affecting the global atmospheric environment. The scientific concept of ODPs originally arose as a simplified means for determining the relative ability of a chemical to destroy ozone (Wuebbles, 1981, 1983). This concept for a relative measure of the cumulative impact on stratospheric ozone from trace gas emissions was found to be particularly useful by policymakers in their considerations to develop policy to protect stratospheric ozone from further destruction from emissions of chlorofluorocarbons (CFCs), halons, and other relatively long-lived halocarbons. As a result, the ODP concept is an integral part of national and international considerations on ozone protection policy, including the international Montreal Protocol and its Amendments and the US Clean Air Act. The US Environmental Protection Agency and other organizations such as the United Nations Environment Programme now require evaluations of ODPs for all newly considered halocarbons, particularly for those under consideration as replacement compounds for CFCs and halons and uses as refrigerants, foam blowing agents, solvents, fire retardants, and other relevant applications where significant emissions could occur into the atmosphere.

Emissions of gases containing chlorine, bromine, or iodine can particularly affect stratospheric ozone (but not the emissions of water soluble gases like HCl that would likely be washed out of the atmosphere and not reach the stratosphere). Chlorine, bromine, and iodine react extremely efficient with ozone in catalytic processes, and these atoms can destroy thousands of ozone molecules if they reach the stratosphere. If a compound does not contain chlorine, bromine, or iodine, then it less likely to affect the stratosphere. Exceptions could occur if the compound can achieve such significant use that it would become a source of stratospheric nitrogen oxides, hydrogen oxides, or sulfuric particles.

The Concept of ODPs

The ODP of a gas is defined as the integrated change in total ozone per unit mass emission of the gas, relative to the change in total ozone per unit mass emission of CFC-11, one of the

gases of most concern to ozone change. Alternatively, the ODP can be derived by using a constant emission calculated to steady state relative to the same for CFC-11. Numerically, the two approaches are equivalent. Thus, the ODP of gas X is defined by

$$\text{ODP}(X) = \frac{\Delta\text{O}_3(X, E)}{\Delta\text{O}_3(\text{CFC} - 11, E)} \quad [1]$$

where $\Delta\text{O}_3(X, E)$ is the steady state percent change in stratospheric ozone burden due to emission at the Earth's surface of species X at E (kg year^{-1}). As a relative measure, ODPs are thought to be subject to fewer uncertainties than estimates of the absolute percentage of ozone depletion caused by different gases.

ODPs as used in the international scientific assessments on stratospheric ozone sponsored by the World Meteorological Organization (e.g., WMO, 1992, 1995, 1999, 2003, 2007, 2011) are generally determined by two different means: calculations from models, primarily from models of global atmospheric chemistry and physics, and a semiempirical approach developed by Susan Solomon and colleagues (Solomon et al., 1992, 1994). The two approaches give similar results. Until roughly the year 2000, zonally averaged, two-dimensional models of the global atmosphere were primarily used in the model determination of ODPs. More computationally intensive three-dimensional models, that represent variations with longitude as well as latitude and altitude, and that more fully represent the physics and chemistry of the atmosphere, are now used for such studies, and are the modeling tools of choice for determining ODPs. The numerical models used in such studies attempt to account for known chemical and physical processes affecting chemical species in the troposphere and stratosphere.

The compounds are assumed to enter the atmosphere at ground level, be transported in the atmosphere by dynamical processes, and react via photolysis or through reaction with other atmospheric constituents. The reactivity of the particular compound depends on its molecular structure. Generally, the compounds of interest are likely to either undergo photolytic breakdown by ultraviolet (UV) or near-UV light, react with OH in the troposphere and stratosphere, or react with atomic oxygen. There are several pathways for delivering the final products of interest, e.g., the chlorine or the bromine, to the stratosphere, as shown in Figure 1. The original compound, the source gas, can be transported directly to the stratosphere,

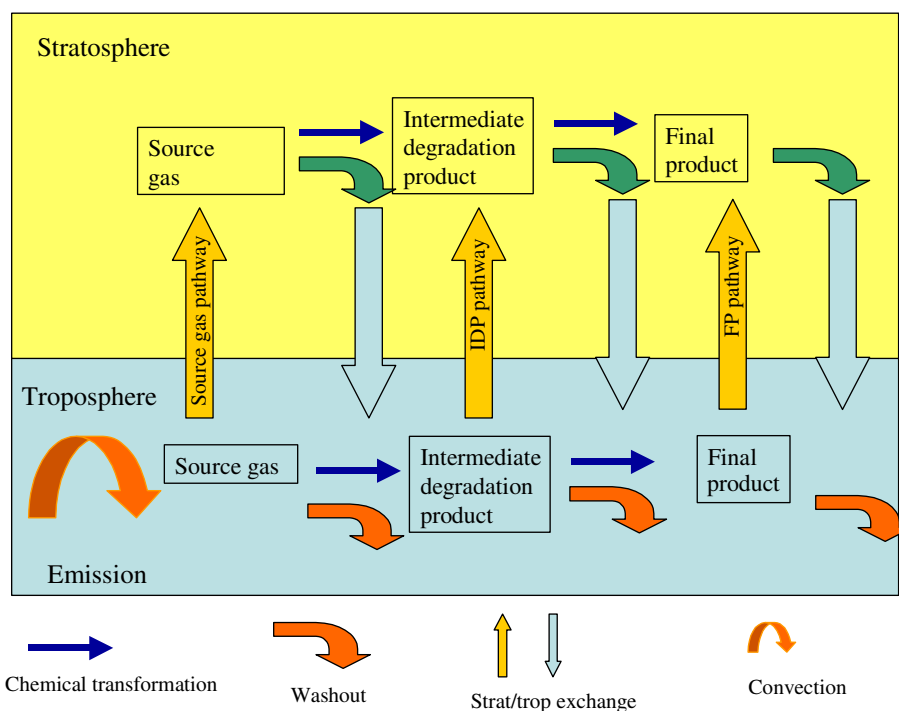


Figure 1 Schematic diagram of the multiple pathways that can occur, particularly for short-lived gases, to deliver halogen atoms, the final product, to the stratosphere. The dark yellow arrows indicate transport into the stratosphere, while the light blue arrows represent transport out of the stratosphere. Adapted from WMO (World Meteorological Organization), 2003. Scientific Assessment of Ozone Depletion: 2002. Global Ozone Research and Monitoring Project Report No. 47. Geneva, Switzerland, 498 pp.

where it reacts to form intermediate degradation products that then react to release the halogen atoms, or, for some reactions, release the halogen atoms directly, that can react with ozone. The long-lived gases that are well mixed in the troposphere, those with atmospheric lifetimes greater than about 1 year, have almost all of their release of halogen atoms occur in the stratosphere. For the short-lived gases, those with atmospheric lifetimes less than 1 year, there is generally enough reactivity in the troposphere that the intermediate reaction products and the final products are often produced in the troposphere. If the products are water soluble, then a high fraction can be removed by washout processes, particularly if the products are produced in the lower troposphere. However, some of these products can be transported to the stratosphere, thus allowing the halogen atoms to be available for reactions with ozone. Various research studies indicate that, for some compounds, a high fraction of the halogen reaching the stratosphere does so in the form of reaction products or as halogen atoms (Ko et al., 1997; Youn et al., 2010; Patten and Wuebbles, 2010, 2011; Wuebbles et al., 2001, 2009, 2011).

Figure 2 shows some of the important considerations in modeling the relevant chemical and physical processes to determine the stratospheric ozone depletion from emission of a halogenated compound. The processes in the large dotted box are not important for long-lived gases, but can be extremely important for short-lived gases.

A major uncertainty in the models is the amount of tropospheric hydroxyl, OH. Since few reliable measurements of OH are available, the global distribution has not been directly measured. As a result, given the importance of the atmospheric

lifetime in determining the ODP for a substance, those gases where reaction with tropospheric OH is the primary loss mechanism, as it is for many of the replacement compounds, a scaling to the partial lifetime of CH_3CCl_3 due to its reaction with tropospheric OH is often used. This scaling is generally not needed in current three-dimensional models. As discussed below, short-lived compounds, those with an atmospheric lifetime shorter than about half a year, require special consideration.

The semiempirical approach for determining ODPs is based on direct measurements of selected halocarbons and other trace species in the stratosphere. Thus, it relates ODP values to measured quantities that can also be used to evaluate the validity of the model-calculated results. The semiempirical derivation requires knowing the physical properties of the halogenated compound, its total lifetime (including loss in the ocean or on soil in addition to reactivity in the atmosphere), a factor representing the distribution of inorganic halogen released in the stratosphere obtained from observations of the compound or other compounds with similar reactivity, and an analysis of the catalytic efficiency for ozone destruction, determined relative to chlorine. The definition of the ODP of compound X using the semiempirical approach is:

$$\text{ODP}_{\text{semi-emp}}(\text{X}) = (\text{fractional release factor}) \times \frac{\mu_{\text{CFC-11}}}{\mu_{\text{X}}} \times \frac{\tau_{\text{X}}}{\tau_{\text{CFC-11}}} \times \frac{n_{\text{X}}}{n_{\text{CFC-11}}} \times \alpha \quad [2]$$

where the fraction τ is the lifetime, n is the number of halogen atoms ($n = 3$ for CFC-11), and α is the relative effectiveness of

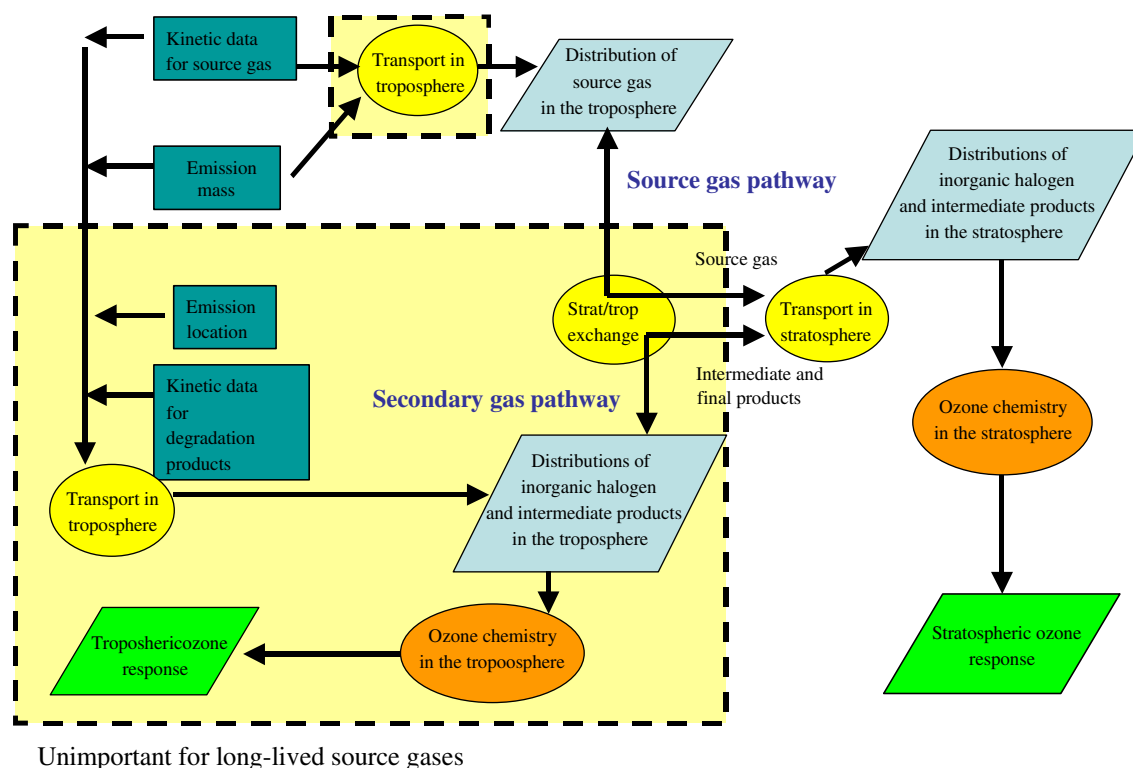


Figure 2 Information needed for modeling the effects of emission of halogenated compounds on stratospheric ozone. Items in rectangles are compound specific. Items in ovals are common to all compounds. Items in parallelograms are model predictions. Adapted from WMO (World Meteorological Organization), 2003. Scientific Assessment of Ozone Depletion: 2002. Global Ozone Research and Monitoring Project Report No. 47. Geneva, Switzerland, 498 pp.

either bromine or iodine to destroy ozone relative to chlorine (current models suggest globally averaged α values of 60 for bromine (Daniel et al., 2011) and much larger values for iodine, as much as 500–1000). The α values are multiplied by the number of bromine or iodine atoms. While bromine and iodine atoms are much more effective than chlorine at destroying ozone, particularly in the lower stratosphere, various forms of the halogen fluorine, such as F and CF_3 , have very little effect on ozone (Solomon et al., 1995).

Note that the semiempirical approach only works for long-lived gases because it does not account for the halogen that can reach the stratosphere as intermediate reaction or final products. Therefore, this approach will be in error for short-lived halogenated compounds.

In the semiempirical approach, the observed fractional dissociation is used to determine the amount of chlorine and bromine (or iodine) released and is then compared with the observationally derived ozone loss distribution, with the assumption that the ozone loss results only from halogen chemistry. The correlation between different compounds is determined on the basis of their relative reactivity in the troposphere and stratosphere. This semiempirical method avoids some of the demanding requirements of accurate numerical simulation of source gas distributions (i.e., of the CFCs, hydrochlorofluorocarbons (HCFCs), and other compounds) and of the resulting ozone destruction. However, the semiempirical approach also depends on the accuracy of the measurements used with this approach. As mentioned earlier, the results for

long-lived gases from current models compare well with the semiempirical derivation of the ODPs.

Since ODPs are defined in terms of the steady-state ozone change (or alternatively as the integrated cumulative effect on ozone of constant emissions), they are not representative of the relative, transient effects expected for compounds during the early years of emission. For model calculations assuming constant emissions, time-dependent ODPs can also be defined where the changes in ozone for the compound and for CFC-11 are calculated as a function of time on a per unit mass emitted basis. These time-dependent ODPs provide information on the shorter timescale effects of a compound on ozone (e.g., see WMO, 1992, 1995). In the first few years after emission, the time-dependent ODP for a compound can be often be larger than the eventual steady-state value. However, the steady-state values generally are preferred and are used in regulatory considerations.

Derived ODPs

By definition, the ODP for CFC-11 is 1.0. The calculated ODPs for other CFCs are all greater than 0.4. The ODPs for halons are all extremely large, much greater than 1.0, reflecting the high reactivity of bromine with ozone. The ODPs for the HCFCs being used or considered as CFC or halon replacements are all small, with values of 0.01–0.05 and sometimes less. The effect on ozone from a unit mass

Table 1 Steady-state ODPs for some of the important chlorocarbons and bromocarbons of concern to stratospheric ozone

Gas	Chemical formula	ODP
CFCs		
CFC-11	CCl ₃ F	1.0
CFC-12	CCl ₂ F ₂	0.82
CFC-113	CCl ₂ FCFClF ₂	0.85
CFC-114	CClF ₂ CClF ₂	0.58
CFC-115	CF ₃ CClF ₂	0.57
Bromocarbons		
Methyl bromide	CH ₃ Br	0.66
Halon-1301	CF ₃ Br	15.9
Halon-1211	CF ₂ ClBr	7.9
HCFCs		
HCFC-22	CHClF ₂	0.04
HCFC-123	CF ₃ CHCl ₂	0.01
HCFC-124	CF ₃ CHClF	0.02
HCFC-141b	CH ₃ CCl ₂ F	0.12
HCFC-142b	CH ₃ CClF ₂	0.06
HCFC-225ca	CF ₃ CF ₂ CHCl ₂	0.02
HCFC-225cb	CClF ₂ CF ₂ CHClF	0.03
Others		
Carbon tetrachloride	CCl ₄	0.82
Methyl chloroform	CH ₃ CCl ₃	0.16
Methyl chloride	CH ₃ Cl	0.02

Daniel et al. in WMO (World Meteorological Organization), 2011. Scientific Assessment of Ozone Depletion: 2010. Global Ozone Research and Monitoring Project Report No. 52. P.O. Box 2300. Geneva, Switzerland, 516 pp.

emission of one of these HCFCs would correspondingly be less than a 100th of the effect on ozone than the CFC or halon they would replace. ODPs for all the hydrofluorocarbons (HFCs), perfluorocarbons (PFCs), and for sulfur hexafluoride are near zero, owing to the low reactivity of their dissociation products with ozone. **Table 1** shows ODPs for several of the gases of most concern to ozone. Replacement compounds like HCFCs and HFCs (not shown because their ODPs are close to zero) have much smaller ODPs than the CFCs and halons.

Short-Lived Compounds: Special Consideration

As discussed above, the ODP concept was originally developed for CFCs and other halocarbons whose atmospheric lifetime are sufficiently long (>1 year) so that their mixing ratios are uniform in the troposphere. However, a number of compounds now being used or being proposed as replacements for substances controlled under the Montreal Protocol have extremely short atmospheric lifetimes, on the order of days to a few months. Emissions of these compounds, generally referred to as very short-lived substances (VSLS), present special problems in determining ODPs. Most uses of these compounds are at northern midlatitudes, in the United States and Canada, Europe, or in Asia. After emission, these compounds must be transported to the tropics where they can then be transported into the stratosphere. On the other hand, emissions from locations in the tropics could have a much higher probability of reaching the stratosphere. As a result, for very short-lived compounds, ODPs are defined as a function of location (latitude, longitude) and sometimes the time (usually season) of emission. It is not adequate to treat these chemicals as if they were uniformly emitted at all latitudes and longitudes as normally done for long-lived gases. Thus, for VSLS, there is no single ODP value, but instead there should be a matrix of values, dependent on location and perhaps time, if, for example, there is a seasonal variation in the use and emission of the compound.

ODPs for a number of short-lived compounds have now been derived (see **Table 2**). An important example of a VSLS is n-propyl bromide (also referred to as 1-bromopropane, CH₂BrCH₂CH₃ or simplified as 1-C₃H₇Br or nPB). This compound, useful as a solvent, has been estimated to have an atmospheric lifetime of about 20 days annually averaged for emissions globally and 25 days for emissions from 30 to 60° N due to its reaction with hydroxyl, OH. Because nPB contains bromine, any amount reaching the stratosphere has the potential to affect concentrations of stratospheric ozone. Several studies (Olsen et al., 2000; Bridgeman et al., 2000; Wuebbles et al., 2001, 2009, 2011; Patten and Wuebbles, 2010, 2011) have noted that the idea of a single lifetime is

Table 2 Estimated ODPs for various short-lived halocarbons based on three-dimensional chemistry-climate modeling studies

Gas	Emissions latitudes	Lifetime (days)	ODP
C ₃ H ₇ Br (nPB)	60° S–70° N	19.6	0.011
	30–60° N	24.7	0.0049
C ₂ Cl ₄ (PCE)	30–60° N	111	0.0060
C ₂ HCl ₃ (TCE)	30–60° N	13	0.00035
C ₃ H ₂ F ₃ Br (BTP)	60° S–60° N	4.3	0.0052
	30–60° N	7.0	0.0028
CH ₃ I	30–60° N	13.6	0.017
CF ₃ I	30–60° N	5.0	0.008

Youn, D., Patten, K.O., Wuebbles, D.J., Lee, H., So, C.-W., 2010. Potential impact of iodinated replacement compounds CF₃I and CH₃I on atmospheric ozone: a three-dimensional modeling study. *Atmospheric Chemistry and Physics* 10, 10129–10144; Patten, K.O., Wuebbles D.J., 2010. Atmospheric lifetimes and ozone depletion potentials of trans-1-chloro-3,3,3-trifluoropropylene and trans-1,2-dichloroethylene in a three-dimensional model. *Atmospheric Chemistry and Physics* 10, 10867–10874; Patten, K., Khamaganov, V., Orkin, V., Baughcum, S., Wuebbles, D., 2011. OH reaction rate constant, IR absorption spectrum, ozone depletion potentials and global warming potentials of 2-Bromo-3,3,3-Trifluoropropene. *Journal of Geophysical Research* 116, D24307. <http://dx.doi.org/10.1029/2011JD016518>; Wuebbles, D.J., Youn, D., Patten, K., Wang, D., Martínez-Avilés, M., 2009. Metrics for ozone and climate: three-dimensional modeling studies of ozone depletion potentials and indirect global warming potentials. In: Zerefos, C., Contopoulos, G., Skalkas, G. (Eds.), *Twenty Years of Ozone Decline*, Springer Publishing, Dordrecht, The Netherlands. <http://dx.doi.org/10.1007/978-90-481-2469-5>; Wuebbles, D.J., Patten, K., Wang, D., Youn, D., Martínez-Avilés, M., Francisco, J., 2011. Three-dimensional model evaluation of the ozone depletion potentials for n-propyl bromide, trichloroethylene and perchloroethylene. *Atmospheric Chemistry and Physics* 11, 2371–2380.

flawed for very short-lived compounds – mixing ratios of nPB reaching the tropical upper troposphere depend strongly on the latitude and season of the emissions. These studies found that tropical emissions were twice as likely to enter the stratosphere as high-latitude emissions. Wuebbles *et al.* (2001) showed that up to 0.5% of the Br emitted as nPB at the surface entered the stratosphere, with only 20–30% of Br in the form of nPB. The remainder was in the form of reaction products or inorganic bromine. These studies also emphasize the importance of knowing the degradation products in quantifying their contribution to halogen loading in the stratosphere.

After emission, generally over continents at midlatitudes in the Northern Hemisphere, short-lived chemicals like nPB are not expected to have well-mixed concentrations in the troposphere. Winds and other mixing processes in the atmosphere would gradually transport these gases to the tropics where they can most effectively be transported to the lower stratosphere. At the same time, because of the short lifetime, reaction with OH and other chemical loss processes would convert a significant fraction of the chemical emitted to its degradation products. The processes affecting resulting degradation products could be quite important as they could act as vehicles for transporting chlorine or bromine to the lower stratosphere. The inorganic bromine compounds produced from tropospheric nPB degradation, if not sufficiently removed by wet deposition, could contribute further to stratospheric bromine. The question of how much of the halogen that would be available to affect stratospheric ozone therefore depends greatly on the complex effects of transport and chemical processes in the troposphere.

Evaluations of ODPs for short-lived compounds with three-dimensional chemical transport models of the troposphere and stratosphere are needed for such short-lived gases. In theory, these models should be able to effectively determine the amount of a short-lived chemical compounds and its degradation products reaching the stratosphere and the resulting potential effects on ozone. The models that have complete representations of stratospheric chemical and physical processes are computationally intensive.

Table 2 compares the ODP values derived for several chemicals with short atmospheric lifetimes using three-dimensional models with complete representation of tropospheric and stratospheric processes. In general, the ODPs for such short-lived compounds tend to be much smaller than those for the ODPs and their first generation replacements, the HCFCs (as seen by comparing with **Table 1**). With the use of CFCs and HCFCs now largely banned because of the concerns about stratospheric ozone, replacements for those compounds need to be found. Given their very short lifetimes, the direct effects on climate from such short-lived compounds also are very small. As a result, it is likely that industry will be turning to consideration of such very short-lived compounds for the various uses as refrigerants, foam blowing agents, solvents, fire retardants, and other relevant applications.

See also: Ozone Depletion and Related Topics: Long-Term Ozone Changes; Photochemistry of Ozone; Stratospheric

Ozone Recovery. Stratospheric Chemistry Topics: Halogen Sources, Anthropogenic.

Bibliography

- Bridgeman, C.H., Pyle, J.A., Shallcross, D.E., 2000. A three-dimensional model calculation of the ozone depletion potential of 1-bromopropane (1-C3H7Br). *Journal of Geophysical Research* 105, 26493–26502.
- Daniel, J.S., Velders, G.J.M., Morgenstern, O., Toohey, D.W., Wallington, T.J., Wuebbles, D.J., *et al.*, 2011. A focus on information and options for policymakers. In: WMO/UNEP, Scientific Assessment of Ozone Depletion: 2010. Research and Monitoring Project Report No. 52. World Meteorological Organization Global Ozone.
- Ko, M.K.W., Sze, N.-D., Scott, C.J., Weisenstein, D.K., 1997. On the relation between stratospheric chlorine/bromine loading and short-lived tropospheric source gases. *Journal of Geophysical Research* 102, 25507–25517.
- Olsen, S.C., Hannegan, B.J., Zhu, X., Prather, M.J., 2000. Evaluating ozone depletion from very short-lived halocarbons. *Geophysical Research Letters* 27, 1475–1478.
- Patten, K., Khamaganov, V., Orkin, V., Baughcum, S., Wuebbles, D., 2011. OH reaction rate constant, IR absorption spectrum, ozone depletion potentials and global warming potentials of 2-Bromo-3,3,3-Trifluoropropene. *Journal of Geophysical Research* 116, D24307. <http://dx.doi.org/10.1029/2011JD016518>.
- Patten, K.O., Wuebbles, D.J., 2010. Atmospheric lifetimes and ozone depletion potentials of trans-1-chloro-3,3,3-trifluoropropylene and trans-1,2-dichloroethylene in a three-dimensional model. *Atmospheric Chemistry and Physics* 10, 10867–10874.
- Solomon, S., Burkholder, J.B., Ravishankara, A.R., Garcia, R.R., 1994. Ozone depletion and global warming potentials of CF₃I. *Journal of Geophysical Research* 99, 20929–20935.
- Solomon, S., Mills, M.J., Meiht, L.E., Pollack, W.H., Tuck, A.F., 1992. On the evaluation of ozone depletion potentials. *Journal of Geophysical Research* 97, 825–842.
- Solomon, S., Wuebbles, D., *et al.*, 1995. Ozone depletion potentials, global warming potentials, and future chlorine/bromine loading, Chapter 13. In: Scientific Assessment of Ozone Depletion: 1994, Global Ozone Research and Monitoring Project Report No. 37. World Meteorological Organization, P.O. Box 2300, Geneva, Switzerland.
- WMO (World Meteorological Organization), 1992. Scientific Assessment of Ozone Depletion: 1991. Global Ozone Research and Monitoring Project Report No. 25. WMO, P.O. Box 2300, Geneva, Switzerland.
- WMO (World Meteorological Organization), 1995. Scientific Assessment of Ozone Depletion: 1994. Global Ozone Research and Monitoring Project Report 37. WMO, P.O. Box 2300, Geneva, Switzerland.
- WMO (World Meteorological Organization), 1999. Scientific Assessment of Ozone Depletion: 1998. WMO, P.O. Box 2300, Geneva, Switzerland. ISBN: 92-807-1722-7.
- WMO (World Meteorological Organization), 2003. Scientific Assessment of Ozone Depletion: 2002. Global Ozone Research and Monitoring Project Report No. 47. P.O. Box 2300, Geneva, Switzerland, 498 pp.
- WMO (World Meteorological Organization), 2007. Scientific Assessment of Ozone Depletion: 2006. Global Ozone Research and Monitoring Project Report No. 50. P.O. Box 2300, Geneva, Switzerland, 572 pp.
- WMO (World Meteorological Organization), 2011. Scientific Assessment of Ozone Depletion: 2010. Global Ozone Research and Monitoring Project Report No. 52. P.O. Box 2300, Geneva, Switzerland, 516 pp.
- Wuebbles, D.J., 1981. The relative efficiency of a number of halocarbons for destroying stratospheric ozone. Lawrence Livermore National Laboratory, report UCID-18924.
- Wuebbles, D.J., 1983. Chlorocarbon production scenarios: potential impact on stratospheric ozone. *Journal of Geophysical Research* 88, 1433–1443.
- Wuebbles, D.J., Patten, K.O., Johnson, M.T., Kotamarthi, R., 2001. New methodology for ozone depletion potentials of short-lived compounds: n-propyl bromide as an example. *Journal of Geophysical Research* 106, 14551–14571.
- Wuebbles, D.J., Youn, D., Patten, K., Wang, D., Martínez-Avilés, M., 2009. Metrics for ozone and climate: three-dimensional modeling studies of ozone depletion potentials and indirect global warming potentials. In: Zerefos, C., Contopoulos, G., Skalkas, G. (Eds.), *Twenty Years of Ozone Decline*. Springer Publishing, Dordrecht, The Netherlands. <http://dx.doi.org/10.1007/978-90-481-2469-5>.
- Wuebbles, D.J., Patten, K., Wang, D., Youn, D., Martínez-Avilés, M., Francisco, J., 2011. Three-dimensional model evaluation of the ozone depletion potentials for n-propyl bromide, trichloroethylene and perchloroethylene. *Atmospheric Chemistry and Physics* 11, 2371–2380.
- Youn, D., Patten, K.O., Wuebbles, D.J., Lee, H., So, C.-W., 2010. Potential impact of iodinated replacement compounds CF₃I and CH₃I on atmospheric ozone: a three-dimensional modeling study. *Atmospheric Chemistry and Physics* 10, 10129–10144.

Further Reading

- Brasseur, G., Orlando, J.J., Tyndall, G.S. (Eds.), 1999. *Atmospheric Chemistry and Global Change*. Oxford University Press, New York.
- Makhijani, A., Gurney, K.R., 1995. *Mending the Ozone Hole: Science, Technology, and Policy*. MIT Press, Cambridge, MA, 360 pp.
- National Academy of Sciences, 1997. *Fire Suppression Substitutes and Alternatives to Halon for U.S. Navy Applications*. National Academy Press, Washington, D.C.
- Wuebbles, D.J., 1995. Weighing functions for ozone depletion and greenhouse gas effects on climate. *Annual Review of Energy and the Environment* 20, 45–70.
- Wuebbles, D.J., Calm, J.M., 1997. An environmental rationale for retention of endangered chemical species. *Science* 278, 1090–1091.
- Wuebbles, D.J., Ko, M.K.W., 1999. Summary of EPA/NASA Workshop on the Stratospheric Impacts of Short-lived Gases, March 30–31, Washington, D.C. Available through the U.S. Environmental Protection Agency, Stratospheric Protection Division, Washington, D.C.
- Wuebbles, D.J., Jain, A., Kotamarthi, R., Naik, V., Patten, K.O., 1999. Replacements for CFCs and halons and their effects on stratospheric ozone. In: Nathan, T.R., Cordero, E. (Eds.), *Recent Advances in Stratospheric Processes*. Research Signpost, India.

Photochemistry of Ozone

GK Moortgat, Max-Planck-Institute for Chemistry, Mainz, Germany

AR Ravishankara, Colorado State University, Fort Collins, CO, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by A R Ravishankara, volume 4, pp 1642–1649, © 2003, Elsevier Ltd.

Synopsis

The photodissociation of ozone in the UV–visible is reviewed and the photochemical processes are identified. Especially the formation of atomic oxygen atom in the first electronically excited state, $O(^1D)$, plays an important role in the lower atmosphere, since it is the dominant precursor of the formation of OH radicals in the lower atmosphere. Experimental studies of the determination of the $O(^1D)$ quantum yield, predominantly performed in the wavelength range 305–355 nm, are presented. The atmospheric implications of the wavelength and temperature dependence of new $O(^1D)$ quantum yields are discussed.

Introduction

Ozone is one of the most important constituents of the Earth's atmosphere (from here on just referred to as the atmosphere). It plays several major roles in the atmosphere that include shielding the Earth's surface from harmful UV radiation, acting as a greenhouse gas, and generating reactive free radicals such as OH, which oxidize and remove pollutants from the atmosphere. Each of these important roles involves the absorption of photons by ozone and hence the importance of ozone photochemistry.

Role of Ozone in the Atmosphere

Absorption Spectrum

Ozone absorbs radiation from the vacuum ultraviolet (VUV) throughout the UV and visible regions and into the infrared. However, photodissociation of ozone in the atmosphere is mostly due to absorption of visible and shorter wavelength radiation. The absorption spectrum of ozone in the visible and UV regions is shown in [Figure 1](#) and is based on the data from a variety of sources.

The ozone absorption spectrum in the 200–900 nm region can be attributed to four systems: the Hartley band (UV), the Huggins bands (near UV), the Chappuis band (visible), and the Wulf band (near infrared). [Orphal \(2003\)](#) and [Sander et al. \(2011\)](#) have published reviews of available spectra measured over the temperature range 200–300 K. Absorption shown in [Figure 1](#) involves the excitation of ozone from the ground electronic state to an excited electronic state from which it dissociates to an oxygen atom and an O_2 molecule. It is, however, possible for ozone to split apart into three O atoms below ~ 197 nm; however, such a three-way dissociation is not important in the atmosphere except at very high altitudes. It is also possible for ozone to reemit the absorbed light (fluoresce); such emission seldom occurs.

Special Role of $O(^1D)$ Atoms in the Atmosphere

Generally speaking, species in excited states do not play a major role in the chemistry of the Earth's lower atmosphere (i.e., below about 40 km). The most notable exception is the oxygen

atom in the first electronically excited state, $O(^1D)$. The primary reason for the extreme importance of $O(^1D)$, which is present in very small concentrations, is that it reacts with highly unreactive atmospheric species, such as H_2O and N_2O , to create free radicals. Free radicals are very reactive and are the primary initiators of chemical changes in the lower atmosphere. Specifically, the OH radical in the stratosphere and troposphere and NO (and eventually all nitrogen oxides) in the stratosphere are produced mostly from the reactions of $O(^1D)$.



The OH radical is immensely important in the atmosphere for the following reasons: (1) OH is the most important initiator of the atmospheric degradation of the majority of natural and anthropogenic trace gases in the troposphere, (2) OH reactions provide the pathways that interconvert active and inactive forms of catalysts that are responsible for stratospheric ozone removal, (3) OH is directly involved in catalytic destruction of stratospheric ozone, and (4) OH is critical for initiating the sequence of reactions that photochemically generate tropospheric ozone. Similarly, nitric oxide (NO) is a crucial ingredient of the stratosphere. Once formed via the reaction of $O(^1D)$ with N_2O , it takes part in direct ozone destruction, suppresses the ozone destruction by halogens, and is a key chemical reactant. Therefore, $O(^1D)$ is a critical reactant in the Earth's atmosphere. The major source of $O(^1D)$ in the atmosphere below ~ 40 km is the photolysis of ozone in the Hartley and Huggins bands that are indicated in [Figure 1](#). (Note: The ground state of oxygen atom, $O(^3P)$, does not react rapidly with most atmospheric species and plays a much less important role in transforming chemicals in the atmosphere. But, its recombination with O_2 to make ozone is the primary pathway for ozone formation throughout the atmosphere.)

Photodissociation of Ozone

Definition of Absorption Cross Sections, Quantum Yield, and Photolysis Frequency

Before examining the photochemistry of ozone, it is necessary to define a few terms that are commonly used. The 'absorption

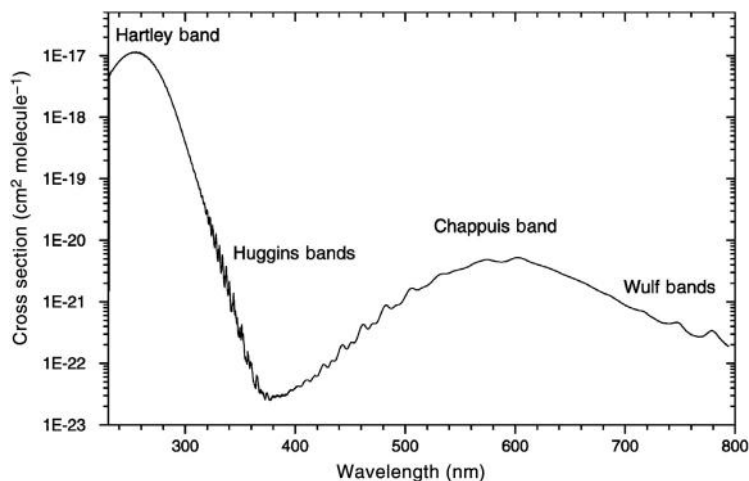


Figure 1 Absorption cross section of ozone between 230 and 800 nm at room temperature, taken from Burrows, J.P., Dehn, A., Deters, B., Himmelmann, S., Richter, A., Voigt, S., Orphal, J., 1999. Atmospheric remote-sensing reference data from GOME: 2. Temperature-dependent absorption cross sections of O₃ in the 231–794 nm range. *Journal of Quantitative Spectroscopy & Radiative Transfer* 61, 509–517; Burkholder, J.B., Talukdar, R.K., 1994. Temperature dependence of the ozone absorption spectrum over the wavelength range 410–760 nm. *Geophysical Research Letters* 21, 581–584; Malicet, J., Daumont, D., Charbonnier, D.J., Parisse, C., Chakir, A., Brion, J., 1995. Ozone UV spectroscopy. II. Absorption cross-sections and temperature dependence. *Journal of Atmospheric Chemistry* 21, 263–273; Molina, L.T., Molina, M.J., 1986. Absolute absorption cross sections of ozone in the 185– to 350 nm wavelength range. *Journal of Geophysical Research* 91, 14501–14508; Voigt, S., Orphal, J., Bogumil, K., Burrows, J.P., 2001. The temperature dependence (203–293 K) of the absorption cross sections of O₃ in the 230–850 nm region measured by Fourier-transform spectroscopy. *Journal of Photochemistry and Photobiology Chemistry* 143, 1–9. The band systems, commonly referred to as Hartley, Huggins, Chappuis, and Wulf bands, are shown in the figure taken from the review of Orphal, J., 2003. A critical review of the absorption cross-sections of O₃ and NO₂ in the ultraviolet and visible. *Journal of Photochemistry and Photobiology A Chemistry* 157, 185–209. See also the Spectral Atlas (Keller-Rudek, H., Moortgat, G.K., Sander, R., Sörensen, R., 2013. The MPI-Mainz UV/VIS spectral atlas of gaseous molecules of atmospheric interest. *Earth System Science Data* 5, 365–373. www.uv-vis-spectral-atlas-mainz.org) for a comprehensive set of data on the absorption cross section of ozone.

cross section, denoted by σ and usually given in units of cm² molecule^{−1}, is a measure of the ability of a molecule to absorb photons. The absorption cross sections can vary from zero to roughly 1×10^{-16} cm² molecule^{−1}. In some exceptional cases, for example in atoms, absorption cross sections can be much higher. An absorption cross section of $\sim 10^{-18}$ to 10^{-17} cm² molecule^{−1} is considered to be large. So, as shown in Figure 1, ozone is a strong absorber in the UV region and a weak absorber in the visible region. The extent to which the light is absorbed is given in terms of the absorption cross section, the concentration of the molecules (n , in molecule cm^{−3}), and the path length (l , in cm) by Beer–Lambert’s law, which states that

$$\frac{I}{I_0} = e^{-n\sigma l} \quad [3]$$

where I and I_0 are the intensities of transmitted and incident light, respectively.

The second quantity of interest is the ‘quantum yield’ for a given process, usually denoted by Φ , which is the fraction of molecules that have absorbed light that result in a specific process or lead to a specific product. The quantum yield for the dissociation of a molecule varies between 0 and 1. The quantum yield for the production of a photoproduct is also usually somewhere between 0 and 1, but can be greater than unity in some cases. For example, if O₃ photodissociates to give three atoms, the quantum yield

for O atoms is three, while the quantum yields for ozone loss is one.

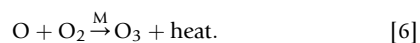
The formation rate in the atmosphere (also called ‘photolysis frequency’) of O(¹D) from the photolysis of ozone is defined as follows

$$J(\text{O}^1\text{D}) = \int_{\lambda} \sigma(\lambda, T) \Phi(\lambda, T) F(\lambda) d\lambda \text{ in units of s}^{-1} \quad [4]$$

where $\sigma(\lambda, T)$ and $\Phi(\lambda, T)$ are the absorption cross section of ozone and the O(¹D) quantum yield at the photolysis wavelength λ and temperature T , and $F(\lambda)$ is the solar actinic flux in the units of photons cm^{−2} nm^{−1} s^{−1}.

Products of Ozone Photolysis

Photodissociation of ozone in the visible region, the Chappuis band, is a major contributor to the stratospheric production of O atoms in their electronic ground state (O(³P)). However, these atoms almost always combine with O₂ to form ozone with release of thermal energy (heat). Thus, the absorption of visible radiation by ozone converts light into heat and this conversion is particularly important in maintaining the temperature structure of the stratosphere.



In the first equation, $h\nu$ represents a photon of light of frequency ν and in the second equation M represents an inert atmospheric gas such as N_2 and O_2 . In most regions of the atmosphere, the interconversion between O and O_3 is so rapid that they are often treated together as one entity called 'odd oxygen,' denoted O_x .

The photodissociation of O_3 has been studied for many decades (Wayne, 1987). It has been shown that O_3 photodissociates to give an O atom and an O_2 molecule. The key questions are: (1) Does ozone dissociate every time it absorbs a visible or UV photon? (2) What energy states of O and O_2 are produced from ozone photodissociation? (3) How do the yields of the photofragments change with temperature, pressure, and wavelength? Clearly, the production of $O(^1D)$ or $O(^3P)$ leads to vastly different consequences in the atmosphere, as noted earlier. Understanding why these species are formed at the specific wavelengths is also of interest. From various studies that have been carried out to date, it appears that ozone always dissociates when it absorbs a visible or UV photon, at least at pressures encountered in the Earth's atmosphere. Therefore, we can assume that the quantum yield for ozone dissociation is unity. This is true at all atmospheric pressures and, hence, the quantum yields are independent of pressure. Because the excited ozone produced upon absorption of UV and visible radiation exists for only a very short time, $<10^{-13}$ s, the identity of the molecule that surrounds this ozone does not matter, as the excited state has no time to collide before it dissociates. In other words, the composition of the atmosphere does not change the quantum yield for the dissociation of ozone.

The energetic thresholds for the production of various products from the photodissociation of ozone are shown in Table 1. This table merely shows what is energetically possible at 298 K. However, the nature of the excited state of ozone produced upon light absorption determines which set of products is possible at a given wavelength and temperature. Since several different excited states can be created by absorption of light at one wavelength, it is possible to get multiple sets of products. These photochemical reaction pathways have been elucidated via experiments over the past few decades (Wayne, 1987; Matsumi and Kawasaki, 2003).

Table 1 Energetic and wavelength threshold for the production of various electronic states of O and O_2 in the photodissociation of ozone

Products	Threshold energy (kJ mol ⁻¹)	Threshold wavelength (nm)
$O(^3P) + O_2(^3\Sigma)$	106.6	1118.4
$O(^3P) + O_2(^1\Delta)$	199.1	599.2
$O(^3P) + O_2(^1\Sigma)$	263.5	452.6
$O(^1D) + O_2(^3\Sigma)$	296.1	402.8
$O(^1D) + O_2(^1\Delta)$	388.5	307.0
$O(^1D) + O_2(^1\Sigma)$	452.9	263.3
$O(^3P) + O(^3P) + O(^3P)$	605.1	197.1

Threshold energies and wavelengths are for 298 K and assume that ozone is in the ground vibrational state. $O_2(^3\Sigma)$ is the ground state of molecular oxygen, with the first and second excited states of O_2 being $O_2(^1\Delta)$ and $O_2(^1\Sigma)$, which lie 94.1 and 156.8 kJ mol⁻¹ above the ground state.

Formation of $O(^1D)$ via Ozone Photolysis in the Atmosphere

Over the past 35 years, the photodissociation of ozone has been carefully examined to quantify how much $O(^1D)$ is produced in the atmosphere subsequent to UV absorption by ozone. This examination was stimulated by the suggestion of Levy (1971) that OH radicals can indeed be formed in the lower atmosphere, where H_2O photolysis to produce OH is not possible. He suggested that the reactions involving the photolysis of ozone to produce $O(^1D)$ followed by the reaction of a fraction of the $O(^1D)$ with H_2O that produces OH to be a source of OH (eqn [1]).

The main atmospheric constituents, N_2 and O_2 , quench almost all the $O(^1D)$ to $O(^3P)$ in the atmosphere. We will first discuss the photochemistry in the UV region, i.e., below 400 nm in the Hartley (220–310 nm) and Huggins bands (310–370 nm), and then in the visible region, the Chappuis band (400–700 nm).

Because of absorption by overhead ozone, the actinic flux available for ozone photolysis in the lower stratosphere and the troposphere decreases very rapidly below ~ 330 nm. Stratospheric ozone acts as a UV shield by absorbing most of the biologically active UV radiation. Since light below ~ 290 nm essentially does not reach the lower stratosphere and the troposphere, the critical wavelengths for the photodissociation of ozone in this altitude region is restricted to ~ 290 – 350 nm. The upper limit for this wavelength range comes about because the absorption cross section of ozone in the Huggins bands becomes so small at, or beyond, 350 nm that, even with the increased solar actinic flux (by more than four orders of magnitude), the photodissociation of ozone becomes negligible. This is precisely the wavelength region where $O(^1D)$ production decreases from near-unity values around 305 nm to near-zero values around 350 nm, as shown in Figure 2. Therefore, the calculated atmospheric $O(^1D)$ production rate is very sensitive to the wavelength dependence of the quantum yield for its production in this wavelength range. It is also this region where the UV absorption cross sections of ozone $\sigma(O_3)$ and the quantum yields Φ for $O(^1D)$ production are highly sensitive to temperature. Thus, all the factors that together determine the $O(^1D)$ production rate are simultaneously changing rapidly with wavelength in this critical region. Therefore, accurate determinations of these parameters are thus crucial. Since the absorption cross section are reasonably well known and available light levels can be calculated, the uncertainty in photochemistry of ozone is the main concern for calculating the $O(^1D)$ production rate.

Quantum Yield for Formation of $O(^1D)$ in the Photolysis of Ozone

Early Quantum Yield Measurements

Until the mid-1990s, based on many previous data, it was assumed that the $O(^1D)$ production decreases monotonically from a near-unity value at ~ 290 nm to zero by ~ 310 nm. The threshold for the energetically allowed channel to produce $O(^1D)$ and $O_2(^1\Delta)$, the first electronically excited state of O_2 molecule, is ~ 307 nm for all ozone in the ground vibrational state (Table 1). However, one would expect the threshold to be shifted to a longer wavelength (~ 310 nm) because a fraction of the ozone

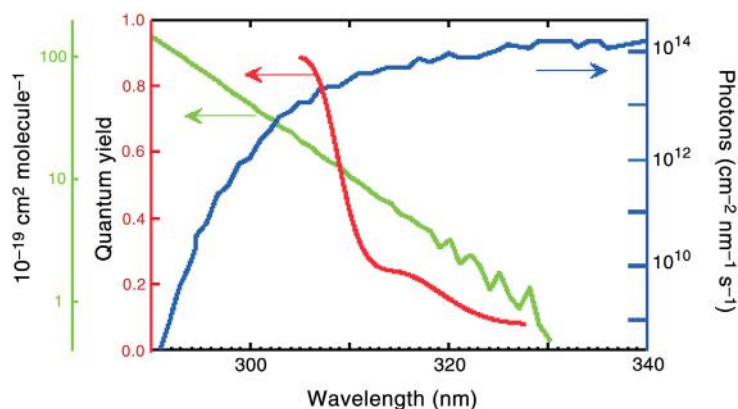
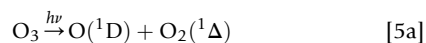


Figure 2 A representation of the various factors that determine the rate of $O(^1D)$ production in the atmosphere. The green line is the absorption cross section of ozone (note the logarithmic scale). The red curve is the current quantum yield (see [Figure 4](#) for recently determined values). The blue line is the approximate calculated solar flux at the surface for equinox, a solar zenith angle of 40° , US standard atmosphere conditions, an overhead ozone column of 300 DU, and a surface albedo of 5%.

molecules will have internal (extra) energy in the form of vibrational energy and hence increase the threshold wavelength.



It was assumed that the quantum mechanical electronic spin conservation rules would require the coproduct of $O(^1D)$ to be $O_2(^1\Delta)$ and that for $O(^3P)$ to be $O_2(^3\Sigma)$, since the upper state of ozone accessed by absorption is a singlet state formed via the strongly allowed transition from a singlet ground state.

The earliest measurements of $O(^1D)$ quantum yield in ozone photodissociation were those of [Jones and Wayne \(1970\)](#) as shown in [Figure 3](#). Their results agreed with the above expectation of zero quantum yields around 320 nm that increased to unity around 300 nm. Later studies of Moortgat and coworkers ([Arnold et al. \(1997\)](#); [Moortgat and Kudszus \(1978\)](#); [Moortgat and Warneck \(1975\)](#)) showed that the quantum yield changed with wavelengths above 305 nm to produce the shape shown in

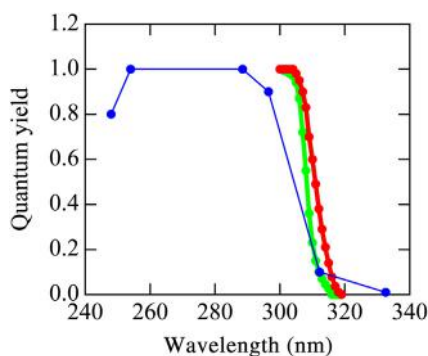


Figure 3 Earliest measurement of the quantum yield for production of $O(^1D)$ in the photodissociation of O_3 by [Jones and Wayne \(1970\)](#) (blue curve). The other two curves are the determinations by Moortgat and coworkers at 298 K (red) and 230 K (green). Adopted from the data in the paper of Jones, I.T.N., Wayne, R.P., 1970. The photolysis of ozone by ultraviolet radiation, IV. The effect of photolysis wavelength on primary steps. *Proceedings of the Royal Society of London A* 319, 273–287 and Moortgat, G.K., Kudszus, E., 1978. Mathematical expression for the $O(^1D)$ quantum yield from the O_3 photolysis as temperature (230–320 K) and wavelength (295–320 nm). *Geophysical Research Letters* 5, 191–194.

[Figure 3](#). In these studies, which were state-of-the-art at that time, the measured quantum yields for $O(^1D)$ reached zero by 330 nm and this variation was not very sensitive to temperature in the wavelength range of 305–330 nm. Many other studies followed and were generally in agreement with this conclusion. The studies by Moortgat and coworkers were a major advancement over those of earlier studies. They had to use indirect methods since powerful lasers for photolysis were not yet available and direct detection of $O(^1D)$ was difficult, especially when the absorption cross sections and quantum yields were very low, for example, beyond 320 nm. Yet, these determinations of the $O(^1D)$ quantum yields enabled the calculation of $O(^1D)$, and hence of the OH radical, production rates in the atmosphere and firmly established Levi's mechanism for OH production in the atmosphere.

Presence of a 'Tail' in the $O(^1D)$ Quantum Yield

There were, however, some laboratory data that suggested the presence of a 'tail' in $O(^1D)$ quantum yield (i.e., a nonzero quantum yield) beyond the energetic threshold. [Adler-Golden et al. \(1982\)](#) had clearly pointed out that vibrationally excited ozone, which is naturally present in the atmosphere, could give $O(^1D)$ via the spin-allowed channel well beyond the 307 nm energetic threshold calculated for the ground state of ozone. Thus, the vibrational excitation of ozone decreased the energy barrier for $O(^1D)$ production and extended this photochemical process to longer wavelengths. [Michelsen et al. \(1994\)](#) showed that the effect of the vibrationally excited ozone in the atmosphere was to give nonzero quantum yields, which varied with temperature, beyond the thermodynamic limit of 307.0 nm (see [Table 1](#)). The quantum yield in this wavelength region from atmospheric ozone was expected to decrease with decreasing temperature because the fraction of vibrationally excited ozone molecule would decrease with temperature.

The introduction of improved direct $O(^1D)$ detection methods and availability of tunable laser light sources since the mid-1990s started a new period of intensive study. [Matsumi et al. \(2002\)](#) present an excellent survey of the direct and indirect techniques used for $O(^1D)$ atom detection. [Armerding et al.](#)

(1995), Talukdar et al. (1998), Bauer et al. (2000), and Silvente et al. (1997) introduced $O(^1D)$ detection by laser-induced fluorescence (LIF) to monitor OH products of $O(^1D)$ reactions. Takahashi et al. (1996b, 1998) and Taniguchi et al. (2000) applied direct VUV photolysis LIF for $O(^1D)$ detection at 115 nm and $O(^3P)$ at 130 nm. Chemical ionization combined with mass spectrometry has been used by Smith et al. (2000) to measure $O(^1D)$, while Hancock and coworkers (Ball et al. 1997; Denzer et al. 1997, 1998; Ravishankara et al. 1998) introduced direct resonance-enhanced multiphoton ionization. Very recently, Vranckx (2009) has used another chemiluminescence method to detect $O(^1D)$. All these techniques and measurements of the $O(^1D)$ quantum yield – with direct and indirect detection methods – showed clearly the existence of the ‘tail.’ Moreover, they showed that $O(^1D)$ is produced well beyond the 310 nm threshold proving that the quantum yield does not go to zero even at wavelengths as long as ~ 330 nm even when the temperature is low enough to eliminate the existence of a significant fraction of vibrationally excited ozone.

$O(^1D)$ Quantum Yield Determinations

The absolute quantum yield for $O(^1D)$ has been measured at various wavelengths between 300 and 350 nm in some studies while other studies have measured the quantum yields at various wavelengths relative to a benchmark wavelength. Here, we use 308.0 nm as the benchmark to place all measurements on an absolute scale.

Absolute $O(^1D)$ Quantum Yield at 308.0 nm

Overall, the dataset for the shape of the $\Phi(O(^1D), \lambda)$ fall-off curve at wavelengths between ~ 308 and ~ 340 nm is in reasonable

agreement but exact values as a function of wavelength and temperature differed considerably. Furthermore, the wavelength dependent quantum yield values of $\Phi(O(^1D), \lambda)$ were measured relative to different reference wavelengths. Therefore, a standard reference wavelength at 308.0 nm was established for the $O(^1D)$ quantum yield. Since the photolysis yield for O_3 on absorption of one photon is unity, and the threshold for $3O(^3P)$ is 197.1 nm at 298 K (Table 1), the amount of $O(^1D)$ produced can be compared with the total amount of oxygen atoms formed for an absolute determination. An absolute yield of $\Phi(O(^1D)) = 0.79 \pm 0.12$ at 308.0 nm, independent of temperature, has been measured by several groups using two different techniques. A more accurate value of $\Phi(O(^1D)) = 0.804 \pm 0.048$ was obtained at 308.0 nm by Takahashi et al. (2004), confirming the previous measurements. Here we normalize all the results to a $\Phi(O(^1D), 308.0 \text{ nm}) = 0.79$.

All previous relative $\Phi(O(^1D))$ measurements were put on an absolute scale by an independent group of principal investigators of these recent studies working in the area of ozone photodissociation. This panel conducted a rigorous evaluation of the experimental data for $O(^1D)$ production in the wavelength range 306–328 nm and temperature range 200–320 K (Matsumi et al., 2002). The quantum yield of eight studies were renormalized using $\Phi(O(^1D)) = 0.79$ at 308.0 nm and averaged to produce the $\Phi(O(^1D), \lambda)$ as a function of wavelength, as shown in Figure 4.

Temperature Dependence

All these recent data agree as to how the $O(^1D)$ quantum yield changes with wavelength and substantiate the existence of the temperature-dependent ‘tail’ and of the spin-forbidden photodissociation pathway. The renormalized data were used to obtain the best-fit parameters for the wavelength range

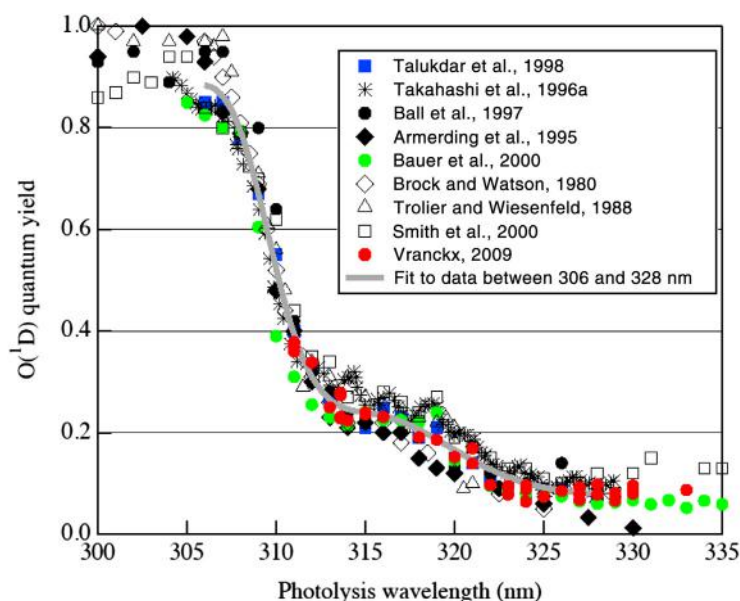
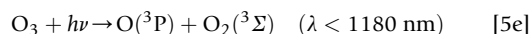
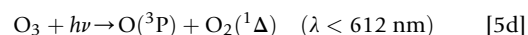
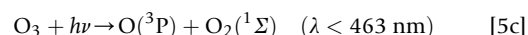
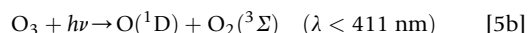
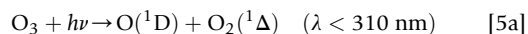


Figure 4 Wavelength dependence of the $O(^1D)$ quantum yields in the photolysis of O_3 at 298 K. The values reported by nine studies have been renormalized to a quantum yield of 0.79 at 308.0 nm. The best fit of these data is shown as the gray line. After Matsumi, Y., Comes, F.J., Hancock, G., Hofzumahaus, A., Hynes, A.J., Kawasaki, M., Ravishankara, A.R., 2002. Quantum yields for production of $O(^1D)$ in the ultraviolet photolysis of ozone: recommendation based on evaluation of laboratory data. *Journal of Geophysical Research* 107. <http://dx.doi.org/10.1029/2001JD000510>.

306–328 nm and the temperature range 200–320 K. Because of the large number of studies upon which the evaluation is based, the averaged 298 K data were given a larger weight in the fitting procedure than the data at other temperatures. The NASA/JPL panel (Sander et al., 2011) recommends the fitting expression derived by Matsumi et al. (2002), using three Gaussian terms and a constant term representing the spin-forbidden channel, for use in atmospheric modeling. Figure 5 shows the wavelength dependent $O(^1D)$ quantum yield at various temperatures in the range 205–321 K.

Physical Processes of $O(^1D)$ Formation

In the wavelength region of 290 to ~ 340 nm, the following processes are possible based on energetics (the threshold wavelengths are shown for 0 K):



The nonzero $O(^1D)$ yield beyond the threshold for spin-allowed channel (process [5a]) has been attributed by Takahashi et al. (1996a) to the existence of a previously unobserved spin-forbidden channel [5b] for $O(^1D)$ production, i.e., formation of $O(^1D) + O_2(^3\Sigma)$. The occurrence of the spin-forbidden photodissociation to produce $O(^1D)$ with a temperature-independent near constant quantum yield up to 335 nm is suggested by various studies. The process [5b] was reported

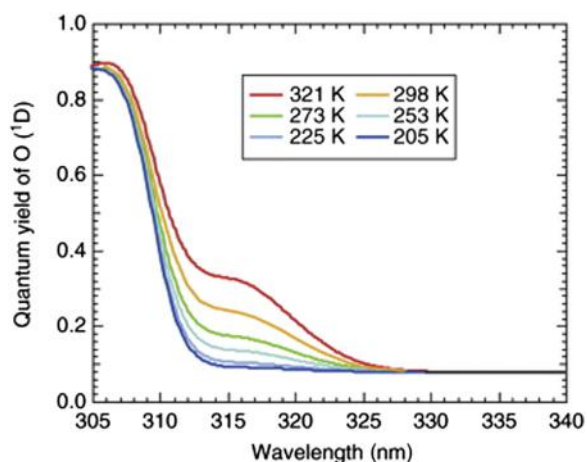


Figure 5 The quantum yield for $O(^1D)$ as a function of wavelength at a few temperatures. These are the values obtained from a detailed evaluation of the existing data by Matsumi, Y., Comes, F.J., Hancock, G., Hofzumahaus, A., Hynes, A.J., Kawasaki, M., Ravishankara, A.R., 2002. Quantum yields for production of $O(^1D)$ in the ultraviolet photolysis of ozone: recommendation based on evaluation of laboratory data. *Journal of Geophysical Research* 107. <http://dx.doi.org/10.1029/2001JD000510>.

to extend even to 375 nm by Bauer et al. (2000). Other spin-forbidden products in the photodissociation of ozone have also been observed, such as $O_2(^1\Delta)$ photofragments in the range 310–328 nm [5d] by Denzer et al. (1997, 1998) and Ravishankara et al. (1998), and $O_2(^1\Sigma)$ photofragments in the range 335–352 nm [5c] by O’Keeffe et al. (1998). Thus, it appears that the spin-forbidden channels are more prevalent than previously believed.

The observed $O(^1D)$ quantum yield wavelength dependence clearly shows that three processes occur. Matsumi et al. (2002) summarized the contributions of the three processes in Figure 6: (1) Ozone in the ground vibrational level absorbs light below ~ 310 nm to give $O(^1D)$ and $O_2(^1\Delta)$. This process [5a], which decreases very rapidly as the wavelength increases above the threshold, is the largest contributor to the rapid change in the quantum yield with wavelength and is not very sensitive to temperature (Region I). (2) Ozone present in vibrationally excited states photodissociates at wavelengths longer than 310 nm. It also contributes below 310 nm, but its contribution is not significant compared to the dissociation from the ground state. The vibrational energy of ozone allows $O(^1D)$ production beyond 310 nm, to approximately 320 nm, and is an allowed process. As the temperature is lowered, the contribution of this channel decreases as the population of vibrationally excited ozone decreases (Region II). (3) Ozone excited from its ground vibrational state photodissociates to $O(^1D) + O_2(^3\Sigma)$, a spin-forbidden process [5b]. The contribution of this channel, which is small but significant, appears to be essentially independent of temperature and wavelength (Region III).

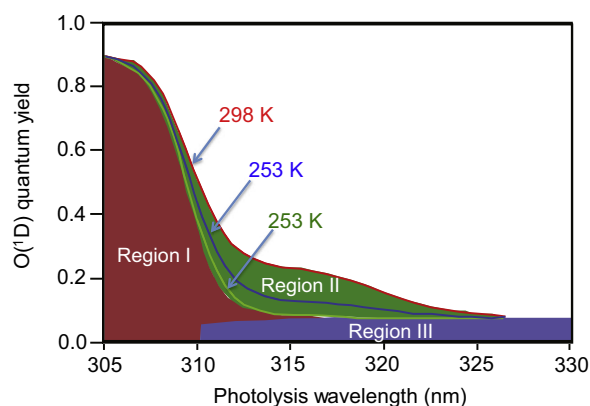


Figure 6 Contributions made by the different O_3 photolysis channels to the $O(^1D)$ quantum yield. Region I corresponds to the $O(^1D)$ formed via process $O(^1D) + O_2(^1\Delta)$; region II (vertical hatching) represents the contribution of the hot band excitation of this process at 298 K, while region III (diagonal hatching) corresponds to the contribution of $O(^1D)$ via the spin-forbidden channel $O(^1D) + O_2(^3\Sigma)$. The solid lines are the recommended $O(^1D)$ quantum yield for 203, 253, and 298 K in the wavelength range 305–330 nm. Adapted from Matsumi, Y., Comes, F.J., Hancock, G., Hofzumahaus, A., Hynes, A.J., Kawasaki, M., Ravishankara, A.R., 2002. Quantum yields for production of $O(^1D)$ in the ultraviolet photolysis of ozone: recommendation based on evaluation of laboratory data. *Journal of Geophysical Research* 107. <http://dx.doi.org/10.1029/2001JD000510>.

Measurements of $O(^1D)$ Quantum Yield in the Region 305–375 nm

In some of the mentioned studies, it was observed that $\Phi(O(^1D), \lambda)$ does not drop to zero at wavelengths longer than 328 nm. The spin-forbidden channel leads to a temperature- and wavelength-independent yield of $O(^1D)$. Smith et al. (2000) reported a nearly constant $\Phi(O(^1D)) \sim 0.12$ between 328 and 338 nm, while Bauer et al. (2000) found a constant yield of 0.064 ± 0.006 up to 375 nm. The evaluation panel recommended a value $\Phi(O(^1D)) = 0.08 \pm 0.04$ independent of temperature up to 338 nm.

Since the $O(^1D)$ formation rate in the atmosphere at large solar angles is sensitive to the $O(^1D)$ quantum yield up to about 340 nm, and since the wavelength range 310–335 nm has large contribution to $O(^1D)$ production, further studies of the quantum yields in this region would be beneficial. Because of the very low absorption cross sections of ozone in this region, experimental measurements are very difficult. Beyond ~ 330 nm, the majority of the ozone dissociated to yield $O(^3P) + O_2(^3\Sigma)$, the spin-allowed channel [5e]. Currently, it is not clear if the small spin-forbidden $O(^1D)$ production [5b] channel extends to ~ 400 nm, the energetic threshold for this process (Table 1).

This motivated a recent study to examine the $O(^1D)$ yield over the wavelength range 311–333 nm using a novel technique for detection for both $O(^1D)$ and $O(^3P)$ based on CF_2 chemiluminescence (Vranckx, 2009). The results of this study match the evaluation given by Matsumi et al. (2002). They also confirmed the participation of the vibrational enhanced photolysis channel [5a] $O(^1D) + O_2(^1\Delta)$ up to 325 nm, and established the formation of a constant $O(^1D)$ quantum of $= 0.081$ above 325 nm from the spin-forbidden process [5b]

$O(^1D) + O_2(^3\Sigma)$. The results from the study by Vranckx are also shown in Figure 4. It can be seen that the new results match previous results and the current NASA/JPL recommendation (Sander et al., 2011).

Atmospheric Implications of the New $O(^1D)$ Quantum Yields

The significance of these changes in $O(^1D)$ quantum yields is very important to atmospheric calculations. One measure of this importance is the change in the rate of $O(^1D)$ production in the atmosphere as measured by the first-order rate coefficient for this process. Figure 7 illustrates the change in the calculated $O(^1D)$ production rate when the processes discussed above are included. There are many situations in the atmosphere where the available wavelengths are restricted to greater than 310 nm, where the importance of the ozone photodissociation via the excited ozone and the spin-forbidden channels become dominant. Such situations include high solar zenith angles and large overhead ozone columns, both common at high latitudes during late fall to early springtime. Of course, high solar zenith angles occur everyday at all sunlit locations at least for a short period.

Ozone Photochemistry in the Region 150–305 nm

Throughout the Hartley band continuum at wavelengths below 305 nm, ozone photolysis occurs via the two spin-allowed processes [5a] and [5e], and the quantum yield for $O(^1D)$ is large and close to unity. However, it appears that the quantum yield of $O(^1D)$ never reaches unity, that is, $O(^3P)$ is always

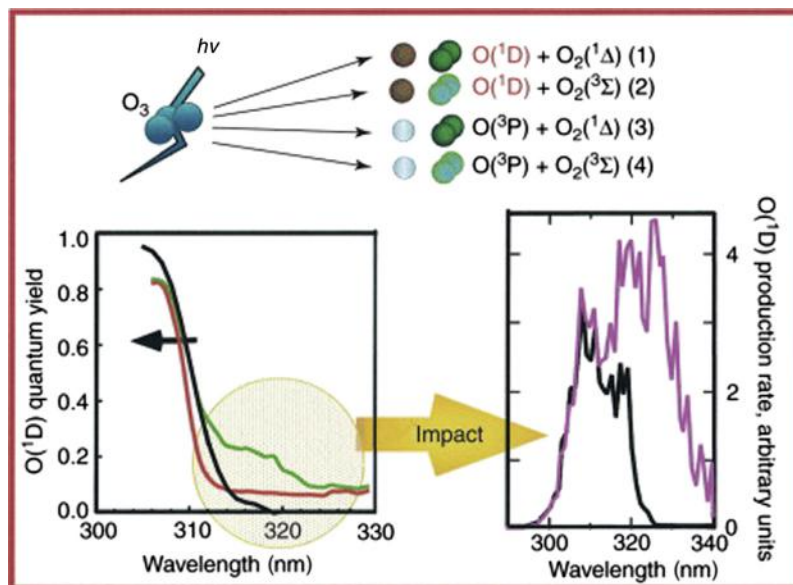


Figure 7 The impact of including the formation of $O(^1D)$ production at wavelengths longer than ~ 320 nm. The black curve on the left panel is the quantum yield for the formation of $O(^1D)$ as a function of wavelength that was used before the discovery of the tail and the spin forbidden channels. The red and green curves are the quantum yields for $O(^1D)$ at 298 and 200 K, based on the best evaluation. The inclusion of these quantum yields changes the $O(^1D)$ production rate from the black curve to the pink curve in the right panel. Adapted from Ravishankara, A.R., Hancock, G., Kawasaki, M., Matsumi, Y., 1998. Atmospheric photochemistry of ozone: surprises and recent lessons. *Science* 280, 60–61.

produced, at least down to 193 nm. The $O(^1D)$ quantum yield in the photolysis of O_3 was determined at 297 K in the wavelength range 230–308 nm by Takahashi et al. (2002) relative to $\Phi(O(^1D)) = 0.79$ at 308.0 nm as a reference. The $O(^1D)$ quantum was found to be almost independent of the photolysis wavelength over the Hartley band (~ 0.91). The results of this study are shown in Figure 8 together with the data of other previous reported studies at a few discrete wavelengths. In several follow-up studies in the range 193–225 nm, the $O(^1D)$ quantum yield was found to decrease from 0.90 ± 0.12 to 0.48 ± 0.03 at 193 nm (Nishida et al., 2004; Turnipseed et al., 1991). The earlier study of Taherian and Slanger (1985) shows that the $O(^1D)$ quantum yield continues to decrease with decreasing wavelength with a value of ~ 0.3 at 156 nm (This study also showed the formation of oxygen atoms in its second excited state, $O(^1S)$). The recent NASA/JPL panel recommended a value of $\Phi(O(^1D)) = 0.90$ in the range 220–305 nm independent of temperature.

The coproduct of $O(^3P)$ is assumed to be $O_2(^1\Sigma)$, with a large amount of internal and translational energy. Stranges et al. (1995) characterized the highly excited triplet states of O_2 , and additional channels corresponding to $O_2(^1\Delta)$ and $O_2(^1\Sigma)$ by the photolysis of O_3 at 193 nm. However, in the face of the prevalence of spin-forbidden processes at longer wavelengths, it may be useful to characterize the electronic state of O_2 that is produced in the range 200–300 nm. Sufficient work has not been done in this wavelength region to elucidate fully the quantum yield of $O(^1D)$. The wavelength region of 220–305 nm is not of major significance to the lower atmosphere because most light in this region is absorbed by ozone before it reaches the mid and lower stratosphere. Above ~ 40 km, O_2 photolysis starts to contribute significantly to $O(^1D)$ production and generation via ozone photolysis decreases in relative importance since the O_2 is so much more abundant than ozone.

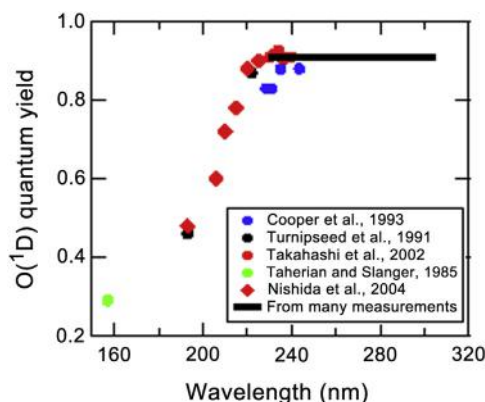


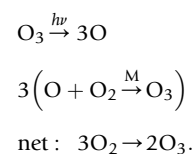
Figure 8 Quantum yield values for the $O(^1D)$ formation in photolysis of ozone in the Hartley band as a function of photolysis wavelength. Original data obtained by various groups are plotted (see legend). The solid line represents the current NASA/JPL recommended value of 0.90. Sander, S.P., Abbatt, J., Barker, J.R., Burkholder, J.B., Friedl, R.R., Golden, D.M., Huie, R.E., Kolb, C.E., Kurylo, M.J., Moortgat, G.K., Orkin, V.L., Wine, P.H., 2011. Chemical Kinetics and Photochemical Data for Use in Atmospheric Studies, Evaluation No 17. JPL Publication 10-6, Jet Propulsion Laboratory, Pasadena. <http://jpl.dataeval.jpl.nasa.gov>.

Formation of $O(^3P)$

Below 200 nm, where a new band of ozone appears, the photochemistry of ozone is not as well understood as in the 200–350 nm region. There are some indications that ozone photolysis in this region, specifically at 193 nm, can yield three oxygen atoms.



Production of three O atoms from ozone could be due to one of two processes: (1) the excited ozone molecule splits apart into three O atoms or (2) the excited atom splits apart into an O atom and an O_2 molecule that is sufficiently excited to decompose into two O atoms. If the production of three ground state O atoms in ozone photolysis is prevalent in this region, as pointed out by Turnipseed et al. (1991), photodissociation of O_3 can actually lead to ozone production.



Such ozone production will be, however, limited to high in the atmosphere because O_2 shields the lower levels from wavelengths less than 200 nm.

The identities of the electronic states of the products of the photodissociation of ozone in the Chappuis band have not been extensively studied. However, all the available data to date suggest that $O(^3P)$ is the primary product with a quantum yield of unity. It also appears that the coproduct of this $O(^3P)$ is $O_2(^3\Sigma)$, the ground electronic state of O_2 . This conclusion is based mostly on the observation that the quantum yield for the removal of ozone in the photolysis of pure ozone in the Chappuis band is close to two. If other electronic states of O_2 were produced, one would expect a larger quantum yield for the removal of ozone. Even though the flux of $O(^3P)$ produced by ozone photolysis in the visible is large and the abundance of $O(^3P)$ is orders of magnitude larger than that of $O(^1D)$, $O(^3P)$ does not play a major role in chemical transformations within the atmosphere. This is primarily because of the low reactivity of $O(^3P)$ compared to that of $O(^1D)$. However, visible photolysis of ozone to produce $O(^3P)$ followed by its reaction with O_2 to regenerate O_3 (the primary fate of $O(^3P)$ in the atmosphere) is a major pathway for the conversion of solar light to heat. It plays a significant role in determining the vertical temperature structure in the stratosphere.

Conclusions

The absorption cross sections and pathways for the dissociation of O_3 have been well established via experimental studies. The atmospheric importance of ozone and its dissociation pathways has been the primary driver for the large amount of experimental work that quantified absorption cross sections and quantum yields for various products. We have summarized the experimentally derived information in this article. Clearly,

the simple ozone molecule has a very complex photochemistry and this complexity profoundly affects the chemistry of the atmosphere. We just note that theoretical attempts to understand the photochemistry of ozone started decades back and there have been many theoretical studies carried out to date to understand ozone photodissociation. A recent article reviews the new theoretical investigations in the photodissociation of ozone in the four major absorption bands (Grebenshchikov et al., 2007). The complexity of ozone photodissociation has made it difficult to theoretically elucidate the dissociation processes and quantify their occurrence. Advancements in theoretical understanding would be beneficial for enabling predictions where experimental data are limited and extending our understanding to the photochemistry of other molecules.

See also: Basic Atmospheric Structure and Concepts: Standard Atmosphere. **Chemistry of the Atmosphere:** Chemical Kinetics; Laboratory Kinetics; Observations for Chemistry (*In Situ*): Ozone Sondes; Principles of Chemical Change. **Climate and Climate Change:** Greenhouse Effect. **Lidar:** Differential Absorption Lidar. **Mesosphere:** Atomic Species in the Mesopause Region; Metal Layers. **Numerical Models:** Chemistry Models. **Optics, Atmospheric:** Optical Remote Sensing Instruments. **Ozone Depletion and Related Topics:** Long-Term Ozone Changes; Ozone Depletion Potentials; Ozone as a UV Filter; Stratospheric Ozone Recovery; Surface Ozone (Human Health); Surface Ozone Effects on Vegetation. **Radiation Transfer in the Atmosphere:** Radiation, Solar; Ultraviolet Radiation. **Satellites and Satellite Remote Sensing:** Measuring Ozone from Space: TOMS and SBUV. **Stratosphere/Troposphere Exchange and Structure:** Global Aspects. **Stratospheric Chemistry Topics:** Halogens; HO_x ; Overview; Reactive Nitrogen (NO_x and NO_y). **Tropospheric Chemistry and Composition:** Hydroxyl Radical; Oxidizing Capacity.

References

- Adler-Golden, S.M., Schweitzer, E.L., Steinfeld, J.I., 1982. Ultraviolet continuum spectroscopy of vibrationally excited ozone. *Journal of Chemical Physics* 76, 2201–2209.
- Armerding, W., Comes, F.J., Schülke, B., 1995. $\text{O}(^1\text{D})$ quantum yield of ozone photolysis in the UV from 300 nm to its threshold and at 355 nm. *Journal of Physical Chemistry* 99, 3137–3143.
- Arnold, I., Comes, F.J., Moortgat, G.K., 1977. Laser flash photolysis: quantum yield of $\text{O}(^1\text{D})$ formation from ozone. *Chemical Physics* 24, 211–217.
- Ball, S.M., Hancock, G., Martin, S.E., Pinot de Moira, J.C., 1997. A direct measurement of the $\text{O}(^1\text{D})$ quantum yields from the photodissociation of ozone between 300 and 328 nm. *Chemical Physics Letters* 264, 531–538.
- Bauer, D., D'Ottone, L., Hynes, A.J., 2000. $\text{O}(^1\text{D})$ quantum yields from O_3 photolysis in the near UV region between 305 and 375 nm. *Physical Chemistry Chemical Physics* 2, 1421–1424.
- Brock, J.C., Watson, R.T., 1980. Laser flash photolysis of ozone: $\text{O}(^1\text{D})$ quantum yields in the fall-off region 297–325 nm. *Chemical Physics* 46, 477–484.
- Burkholder, J.B., Talukdar, R.K., 1994. Temperature dependence of the ozone absorption spectrum over the wavelength range 410–760 nm. *Geophysical Research Letters* 21, 581–584.
- Burrows, J.P., Dehn, A., Deters, B., Himmelmann, S., Richter, A., Voigt, S., Orphal, J., 1999. Atmospheric remote-sensing reference data from GOME: 2. Temperature-dependent absorption cross sections of O_3 in the 231–794 nm range. *Journal of Quantitative Spectroscopy & Radiative Transfer* 61, 509–517.
- Cooper, I.A., Neill, P.J., Wiesenfeld, J.R., 1993. Relative quantum yield of $\text{O}(^1\text{D}_2)$ following ozone photolysis between 221 and 243.5 nm. *Journal of Geophysical Research* 98, 12795–12800.
- Denzer, W., Hancock, G., Pinot de Moira, J.C., Tyley, P.L., 1997. Direct observation of spin-forbidden formation of $\text{O}(^1\text{D})$ in the near-UV photolysis of ozone. *Chemical Physics Letters* 280, 496–500.
- Denzer, W., Hancock, G., Pinot de Moira, J.C., Tyley, P.L., 1998. Spin-forbidden dissociation of ozone in the Huggins bands. *Chemical Physics* 231, 109–120.
- Grebenshchikov, S.Y., Qu, Z.W., Zhu, H., Schinke, R., 2007. New theoretical investigations of the photodissociation of ozone in the Hartley, Huggins, Chappuis, and Wulf bands. *Physical Chemistry Chemical Physics* 9, 2044–2064.
- Jones, I.T.N., Wayne, R.P., 1970. The photolysis of ozone by ultraviolet radiation, IV. The effect of photolysis wavelength on primary steps. *Proceedings of the Royal Society of London A* 319, 273–287.
- Keller-Rudek, H., Moortgat, G.K., Sander, R., Sörensen, R., 2013. The MPI-Mainz UV/VIS spectral atlas of gaseous molecules of atmospheric interest. *Earth System Science Data* 5, 365–373. www.uv-vis-spectral-atlas-mainz.org.
- Levy, H., 1971. Normal atmosphere: large radical and formaldehyde concentrations predicted. *Science* 173, 141–143.
- Malicet, J., Daumont, D., Charbonnier, D.J., Parisse, C., Chakir, A., Brion, J., 1995. Ozone UV spectroscopy. II. Absorption cross-sections and temperature dependence. *Journal of Atmospheric Chemistry* 21, 263–273.
- Matsumi, Y., Kawasaki, M., 2003. Photolysis of atmospheric ozone in the ultraviolet region. *Chemistry Reviews* 103, 4767–4781.
- Matsumi, Y., Comes, F.J., Hancock, G., Hofzumahaus, A., Hynes, A.J., Kawasaki, M., Ravishankara, A.R., 2002. Quantum yields for production of $\text{O}(^1\text{D})$ in the ultraviolet photolysis of ozone: recommendation based on evaluation of laboratory data. *Journal of Geophysical Research* 107. <http://dx.doi.org/10.1029/2001JD000510>.
- Michelsen, H.A., Salawitch, R.J., Wennberg, P.O., Anderson, J.G., 1994. Production of $\text{O}(^1\text{D})$ from photolysis of O_3 . *Geophysical Research Letters* 21, 2227–2230.
- Molina, L.T., Molina, M.J., 1986. Absolute absorption cross sections of ozone in the 185- to 350 nm wavelength range. *Journal of Geophysical Research* 91, 14501–14508.
- Moortgat, G.K., Kudzus, E., 1978. Mathematical expression for the $\text{O}(^1\text{D})$ quantum yield from the O_3 photolysis as temperature (230–320 K) and wavelength (295–320 nm). *Geophysical Research Letters* 5, 191–194.
- Moortgat, G.K., Warneck, P., 1975. Relative $\text{O}(^1\text{D})$ quantum yields in the near UV photolysis of ozone at 298 K. *Zeitschrift für Naturforschung* 30a, 835–844.
- Nishida, S., Taketani, F., Takahashi, K., Matsumi, Y., 2004. Quantum yield for $\text{O}(^1\text{D})$ production from ozone photolysis in the wavelength range of 193–225 nm. *Journal of Physical Chemistry A* 108, 2710–2714.
- O'Keeffe, P., Ridley, T., Wang, S., Lawley, K.P., Donovan, R.J., 1998. Photodissociation of ozone between 335 and 352 nm to give $\text{O}_2(^1\Sigma_g^-) + \text{O}(^3\text{P})_J$. *Chemical Physics Letters* 298, 368–374.
- Orphal, J., 2003. A critical review of the absorption cross-sections of O_3 and NO_2 in the ultraviolet and visible. *Journal of Photochemistry and Photobiology A Chemistry* 157, 185–209.
- Ravishankara, A.R., Hancock, G., Kawasaki, M., Matsumi, Y., 1998. Atmospheric photochemistry of ozone: surprises and recent lessons. *Science* 280, 60–61.
- Sander, S.P., Abbatt, J., Barker, J.R., Burkholder, J.B., Friedl, R.R., Golden, D.M., Huie, R.E., Kolb, C.E., Kurylo, M.J., Moortgat, G.K., Orkin, V.L., Wine, P.H., 2011. Chemical Kinetics and Photochemical Data for Use in Atmospheric Studies. Evaluation No 17. JPL Publication 10-6, Jet Propulsion Laboratory, Pasadena. <http://jpl.dataeval.jpl.nasa.gov>.
- Silvente, E., Richter, R.C., Zheng, M., Saltzman, E.S., Hynes, A.J., 1997. Relative quantum yields for $\text{O}(^1\text{D})$ production in the photolysis of ozone between 301 and 336 nm: evidence for the participation of a spin-forbidden channel. *Chemical Physics Letters* 264, 309–315.
- Smith, G.D., Molina, L.T., Molina, M.J., 2000. Temperature dependence of $\text{O}(^1\text{D})$ quantum yields from the photolysis of ozone between 295 and 338 nm. *Journal of Physical Chemistry* 104, 8916–8921.
- Stranges, D., Yang, X., Chesko, J.D., Suits, A.G., 1995. Photodissociation of ozone at 193 nm by high-resolution photofragment translational spectroscopy. *Journal of Chemical Physics* 102, 6067–6077.
- Taherian, M.R., Slinger, T.G., 1985. Product and yields from O_3 photodissociation at 1576 Å. *Journal of Chemical Physics* 83, 6246–6250.
- Takahashi, K., Hayashi, S., Matsumi, Y., Taniguchi, N., Hayashida, S., 2002. Quantum yields of $\text{O}(^1\text{D})$ formation in the photolysis of ozone between 230 and 308 nm. *Journal of Geophysical Research* 107 (D20), ACH-11. <http://dx.doi.org/10.1029/2001JD002048>.
- Takahashi, K., Hayashi, S., Suzuki, T., Matsumi, Y., 2004. Accurate determination of the absolute quantum yield for $\text{O}(^1\text{D})$ formation in the photolysis of ozone at 308 nm. *Journal of Physical Chemistry A* 108, 10497–10501.
- Takahashi, K., Kishigami, M., Matsumi, Y., Kawasaki, M., Orr-Ewing, A.J., 1996a. Observation of the spin-forbidden $\text{O}(^1\text{D}) + \text{O}_2(^3\Sigma_g^-)$ channel in the 317–327 nm photolysis of ozone. *Journal of Chemical Physics* 105, 5290–5293.

- Takahashi, K., Matsumi, K., Kawasaki, M., 1996b. Photodissociation processes of ozone in the Huggins band at 308–326 nm: direct observation of $O(^1D_2)$ and $O(^3P_1)$ products. *Journal of Physical and Chemical Reference Data* 100, 4084–4089.
- Takahashi, K., Taniguchi, N., Matsumi, Y., Kawasaki, M., Ashfold, M.N.R., 1998. Wavelength and temperature dependence of the absolute $O(^1D)$ yield from the 305–329 nm photodissociation of ozone. *Journal of Chemical Physics* 108, 7161–7172.
- Talukdar, R.K., Longfellow, C.A., Gilles, M.K., Ravishankara, A.R., 1998. Quantum yields of $O(^1D)$ in the photolysis of ozone between 289 and 329 nm as a function of temperature. *Geophysical Research Letters* 25, 143–146.
- Taniguchi, N., Takahashi, K., Matsumi, Y., 2000. Photodissociation of O_3 around 309 nm. *Journal of Physical Chemistry A* 104, 8936–8944.
- Trolier, M., Wiesenfeld, J.R., 1988. Relative quantum yield of $O(^1D_2)$ following ozone photolysis between 275 and 325 nm. *Journal of Chemical Physics* 93, 7119–7124.
- Turnipseed, A.A., Vaghjiani, G.L., Gierczak, T., Thompson, J.E., Ravishankara, A.R., 1991. The photochemistry of ozone at 193 and 222 nm. *Journal of Chemical Physics* 95, 3244–3251.
- Voigt, S., Orphal, J., Bogumil, K., Burrows, J.P., 2001. The temperature dependence (203–293 K) of the absorption cross sections of O_3 in the 230–850 nm region measured by Fourier-transform spectroscopy. *Journal of Photochemistry and Photobiology Chemistry* 143, 1–9.
- Vranckx, S., 2009. Kinetic Studies of O_3 -Photolysis and Subsequent Atmospheric $O(^1D)$ Reactions (Ph.D. thesis), University of Leuven, Belgium.
- Wayne, R.P., 1987. The photochemistry of ozone. *Atmospheric Environment* 21, 1683–1694.

Further Reading

- Brasseur, G.P., Orlando, J.J., Tyndall, G.S., 1999. *Atmospheric Chemistry and Global Change*. Oxford University Press, Oxford.
- Calvert, J.G., Pitts Jr, J.N., 1966. *Photochemistry*. Wiley, New York.
- Finlayson-Pitts, B.J., Pitts Jr, J.N., 1986. *Atmospheric Chemistry: Fundamentals and Experimental Techniques*. Wiley, New York.
- Meier, R.R., Anderson, G.P., Cantrell, C.A., Hall, L.A., Lean, J., Minschwaner, K., Shetter, R.E., Shettle, E.P., Stamnes, K., 1997. Actinic radiation in the terrestrial atmosphere. Review paper. *Journal of Atmospheric Solar-Terrestrial Physics* 59, 2111–2157.
- Okabe, H., 1978. *Photochemistry of Small Molecules*. Wiley, New York.
- Wayne, R.P., 2000. *Chemistry of the Atmospheres*, Third Edition. Oxford University Press, Oxford.

Stratospheric Ozone Recovery

DJ Hofmann[†], NOAA Climate Monitoring and Diagnostics Laboratory, Boulder, CO, USA

R Müller, Institute for Energy and Climate Research (IEK-7), Forschungszentrum Jülich, Jülich, Germany

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by D J Hofmann, volume 5, pp 2202–2208, © 2003, Elsevier Ltd.

Synopsis

Stratospheric ozone is depleted as a result of chemical loss caused by artificial chemicals. Maximum losses occur over the poles with the strongest loss in Antarctica (ozone hole). As a result of the Montreal Protocol, human-produced chemicals causing ozone depletion decline now in the atmosphere, so that a recovery of the ozone layer is expected. In the upper stratosphere (about 30–50 km), the previously observed decline has stopped since a few years, but an unequivocal detection of ozone recovery in other parts of the stratosphere is not possible yet.

Introduction

The phenomenon of depletion of the stratospheric ozone layer by human-produced chemicals has been dealt with in other parts of this Encyclopedia (*see Ozone Depletion and Related Topics: Ozone Depletion Potentials; Ozone as a UV Filter; Photochemistry of Ozone; Climate and Climate Change: Volcanoes: Role in Climate; Ozone Depletion and Related Topics: Surface Ozone (Human Health); Surface Ozone Effects on Vegetation*). Here, we deal with the recovery of the ozone layer – the reasoning behind the predictions that the ozone layer will recover to a state not exactly as it was prior to about 1980 (when the effects of ozone depletion clearly emerged) but to a state in which the threat of harmful ultraviolet radiation increases is no longer an environmental concern.

The subject of the recovery of the stratospheric ozone layer was dealt with in the most recent World Meteorological Organization (WMO)/United Nations Environment Programme (UNEP) Scientific Assessment of Ozone Depletion: 2010. Moreover, the reader is referred to Chapter 12.4 of the WMO/UNEP Scientific Assessment of Ozone Depletion: 1998 (*see Further Reading*) for an extensive discussion of stratospheric ozone recovery. We will summarize here why recovery is expected and discuss the models used to predict the course of recovery. Further, the measurements related to recovery of the ozone layer will be updated and the first signs of a recovery of the ozone layer will be discussed.

Ozone loss in the polar regions during spring is much more severe than the reduction that has occurred at midlatitudes since about 1980. This is related mainly to the fact that in order for chemical ozone destruction to proceed rapidly, very low temperatures and the presence of surfaces for heterogeneous chemistry are required. In most of the global stratosphere, temperatures are too high for rapid heterogeneous reactions of chlorine-bearing species. Conditions for rapid ozone depletion occur in polar winter and spring in association with very low temperatures and stratospheric aerosol and cloud particles together with the onset of springtime sunlight following the dark winter.

Shortly after the discovery of the Antarctic ozone hole in 1985, expeditions to Antarctica in 1986 and 1987 to determine the cause of the springtime ozone depletion resulted in considerable public awareness of the phenomenon. Even now, each austral spring (September–October) finds the Antarctic ozone hole in the news, with reports that either it was ‘not as bad’ as last year or that ‘it was worse than last year.’ In addition, expeditions to study Arctic ozone loss have demonstrated that major ozone loss may occur when Arctic springs are unusually cold. Indeed, the Arctic spring of 2011 showed unusually low temperatures until late in spring were observed, which resulted in severe ozone losses, reaching levels hitherto only observed in Antarctica (Manney et al., 2011).

These events have resulted in considerable confusion concerning the eventually expected outcome of this phenomenon. In actuality, the year-to-year fluctuations in the severity of the ozone hole have been relatively small in recent years, as can be seen in [Figure 1](#), where the total column ozone as measured at the South Pole during the later half of October and at the Halley station (75° S) in October are shown. Adequate sunlight for measurements of total ozone with the Dobson ozone spectrophotometer is available only after mid-October at the South Pole. Variations in the magnitude of Antarctic ozone loss since the last ~15 years have been dominated by variations in temperature and the dynamics of the polar vortex, in which the winter–spring ozone depletion process is confined; in particular, in 2002.

Although the Arctic stratosphere does not get as cold in winter as does the Antarctic stratosphere, substantial ozone loss occurs in cold Arctic springs. Arctic ozone loss is much more variable from year to year than in Antarctica, with little or no loss occurring in warm winters and in winters with a very early breakup of the vortex ([Figure 2](#)). The strongest Arctic ozone loss so far was observed in the extremely cold Arctic spring of 2011. The future of Arctic ozone depletion will depend on a number of factors, including climate change and, possibly, an artificial injection of sulfate particles into the stratosphere to counteract climate change (geoengineering), which is discussed.

At midlatitudes, ozone losses are much smaller. The smaller loss, together with large ozone fluctuations related to transport, makes the detection of midlatitude ozone loss more difficult than in the polar regions. Since midlatitude temperatures are not low enough for rapid heterogeneous reactions of

[†] Deceased.

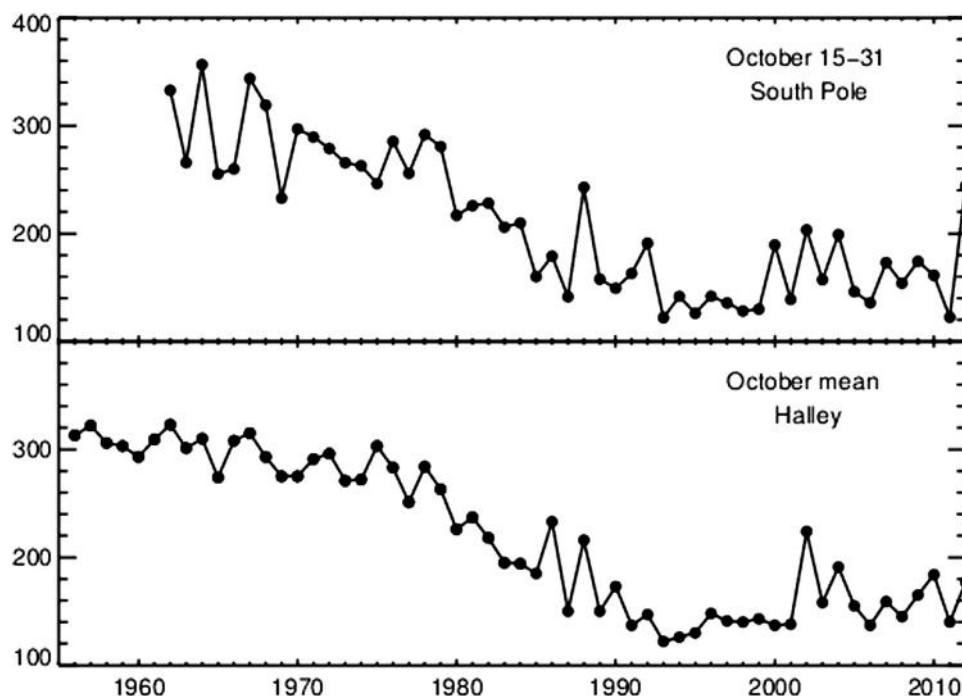


Figure 1 Dobson spectrophotometer total column ozone measurements at South Pole station for the 15–31 October period and at Halley station (75° S) for October. Reliable data at South Pole are not available prior to 15 October owing to lack of sunlight for the measurement.

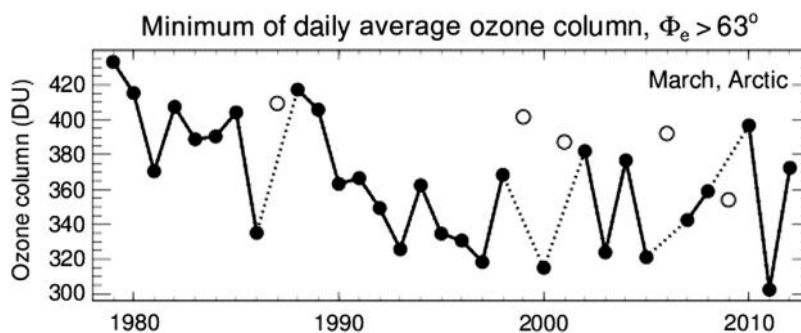


Figure 2 Minimum total ozone in the Arctic in March (calculated the minimum of daily average column ozone poleward of 63° S equivalent latitude). Years in which the polar vortex broke up before March are shown as open symbols. Updated from WMO, 2011. Scientific Assessment of Ozone Depletion: 2010. Global Ozone Research and Monitoring Project, Report No. 52. World Meteorological Organization. Figure courtesy of Jens-Uwe Grooß, Forschungszentrum Jülich.

chlorine-bearing species, and thus for widespread chlorine activation, chemical ozone loss rates are much lower than in the polar regions in spring. Following major volcanic eruptions, sulfuric acid aerosol droplets become important in the heterogeneous chemical process that leads to the enhanced destruction of ozone. **Figure 3** demonstrates the degree of ozone loss experienced across midlatitudes of the United States and shows how the fluctuations related to transport rival the losses experienced since 1980. Enhancement in ozone loss in 1992–93 is believed to be related to the stratospheric aerosol injected into the stratosphere by the eruption of Pinatubo in June 1991.

Global measurements and projected future abundances of the chlorine- and bromine-bearing gases believed responsible for most of the ozone depletion (referred to as ozone-depleting substances (ODSs)) are shown in **Figure 4**. These data indicate

that the combined equivalent effective stratospheric chlorine (EESC) abundance (EESC is a measure of the stratospheric halogen loading and its potential for ozone depletion and is calculated from the surface abundances of chlorine- and bromine-bearing ODS) peaked in the stratosphere in the late 1990s (several years after the peak loading of ODSs in the troposphere).

Thus, there is no reason to expect the Antarctic ozone hole or global ozone depletion to become significantly worse than at present. However, recovery of EESC and thus of the ozone layer to pre-1980 levels is not expected until the middle of the twenty-first century. Model predictions of climate change suggest that climate change should accelerate ozone recovery (in particular in the upper stratosphere) because climate change leads to lower stratospheric temperatures, which cause

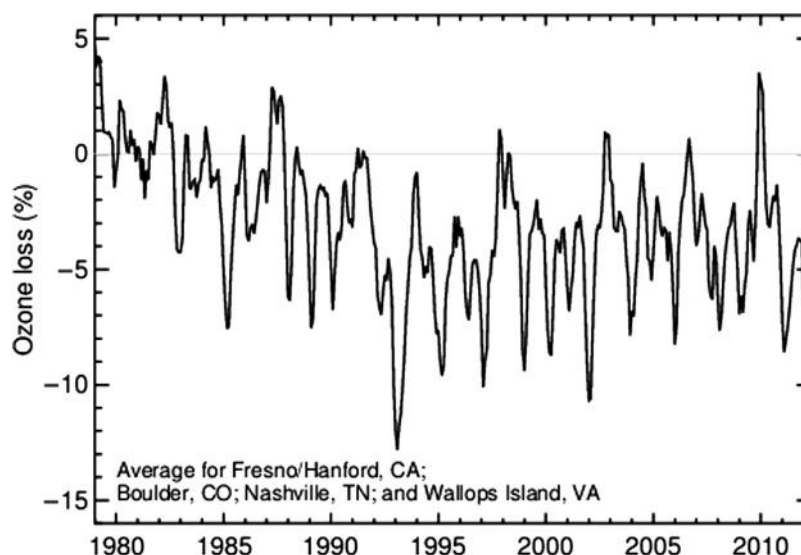


Figure 3 Monthly average total column ozone deviations from the pre-1979 mean at four Dobson spectrophotometer stations across midlatitudes of the United States. The large reduction in 1992–93 is partially related to the Pinatubo volcanic eruption.

gas phase catalytic ozone loss cycles to proceed slower. In the Arctic, where a cooler stratosphere would enhance heterogeneous processing and exacerbate ozone depletion, climate change could delay ozone recovery. Major volcanic eruptions, which supply aerosol particles to the stratosphere, aid the heterogeneous chemistry of halogen ozone loss and will cause a delay in ozone recovery. Clearly, the road to recovery will not be smooth, but it appears that the remedy has been found and it is likely that the phenomenon of stratospheric ozone depletion will not get substantially worse than at present. But how long will the recovery of the ozone layer take?

Ozone Recovery Defined

When speaking about ‘ozone recovery’ here, we are concerned with the way the atmosphere responds to a reduction of the stratospheric halogen loading as a result of the Montreal Protocol and its amendments and adjustments. The first stage of ozone layer recovery is defined (WMO, 2011) as a slowing of ozone decline (i.e., a cessation in the worsening of ozone depletion), identified as the occurrence of a statistically significant reduction in the rate of ozone decline due to changing EESC, i.e., due to changing halogen loading. EESC values are represented as a weighted sum of reactive chlorine and bromine in the stratosphere, accounting for the greater effectiveness of bromine in destroying ozone and thus allowing an estimate how the potential of ODSs to destroy ozone varies over time.

This first stage is observed in the upper stratosphere since about the past decade. Ozone abundances in midlatitudes and in Antarctic spring have similarly ceased to decline since about a decade (Figures 1 and 3), but an unequivocal attribution to changes in the stratospheric halogen loading is difficult. The second stage of ozone recovery is defined as the onset of ozone increases and is identified as the occurrence of statistically significant increases in ozone above the previous minimum

values caused by declining EESC (WMO, 2011). The unambiguous detection of such a recovery is difficult because of the requirement to detect a statistically significant ozone increase, above natural variability, that is occurring slowly over a long period of time. Measurement stability and comparability between multiple instruments over a 20-year or longer period is thus required to detect recovery. These observations could be further confused by occasional volcanic eruptions that will cause ozone depletion to increase for 2–3 years during which the aerosol particles from the eruption slowly fall out of the stratosphere. At the 40-km region, where the chemistry-affecting ozone is relatively simple and volcanic effects are absent, recovery might be observable most easily. However, as ozone in this region contributes only a few percent to the total column, recovery at 40 km should not be interpreted as evidence for the recovery of the global ozone layer. Observation of ozone recovery is important because it can show that the implementation of regulations on ODSs, established by the Montreal Protocol and its amendments and adjustments, was an effective course to follow.

Modeling Ozone Recovery

Projections of future levels of stratospheric ozone are based on the results of three-dimensional chemistry climate models (CCMs) that allow a realistic description of the polar wind system (polar vortex) and that contain a comprehensive representation of stratospheric heterogeneous and gas-phase chemistry. Further, equivalent stratospheric chlorine (ESC) values are predicted in CCMs describing the potential of ODSs to destroy ozone. While both EESC and ESC are designed as a measure of essentially the same quantity, ESC may be calculated directly from models as a function of altitude, latitude, longitude, and time. For the purposes of this general discussion of ozone recovery, EESC and ESC are essentially equivalent. The range of total ozone changes and the range of changes in ESC since 1960

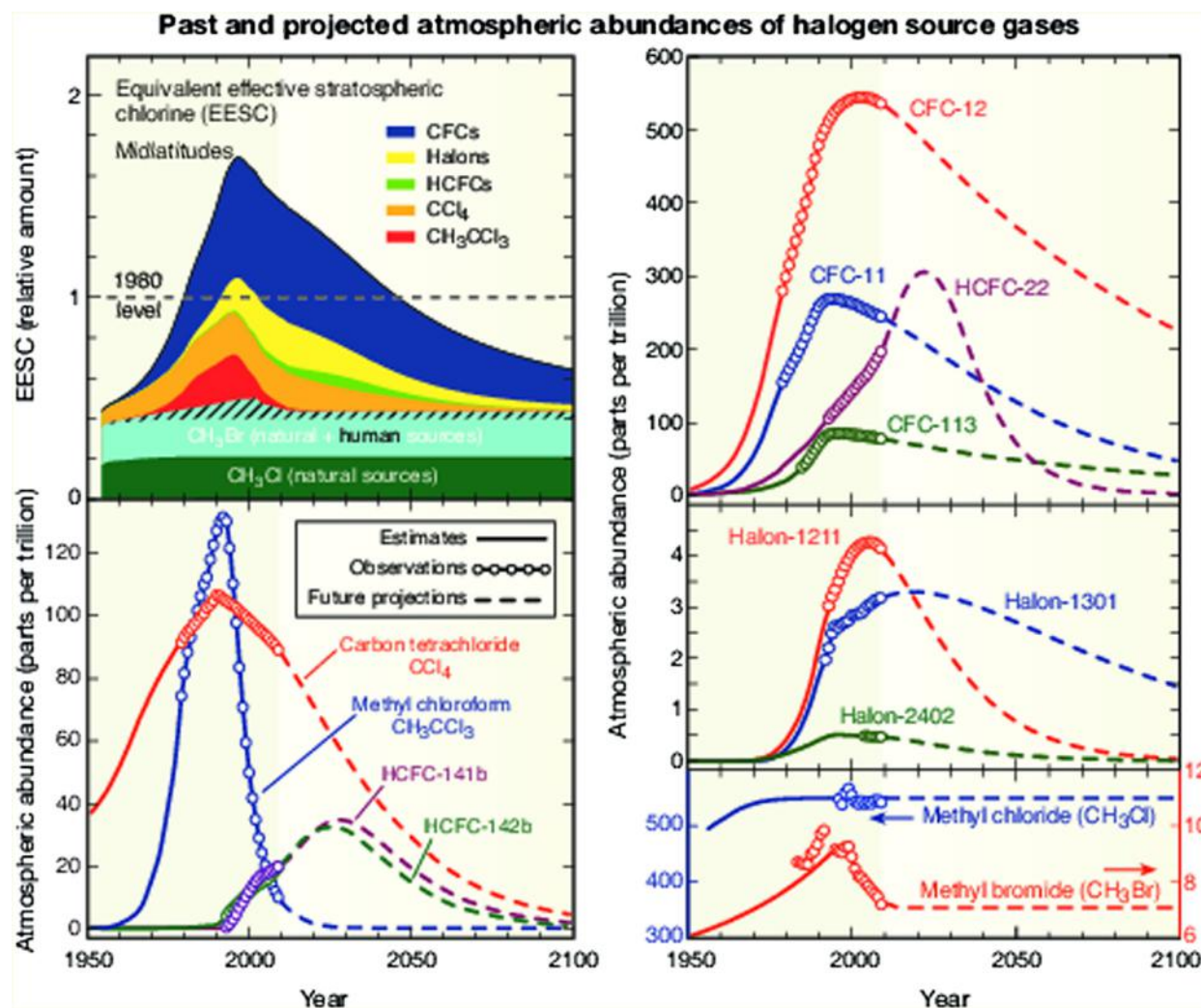


Figure 4 Measurements and projected future abundances of global chlorine- and bromine-bearing compounds that are included in the EESC calculation shown at the top left panel of the figure. Taken from WMO, 2011. Scientific Assessment of Ozone Depletion: 2010. Global Ozone Research and Monitoring Project, Report No. 52. World Meteorological Organization.

as derived from a variety of CCMs is shown in Figure 5. Comparison of model results simulating recent ozone changes with observations increases the confidence in projections of future stratospheric ozone, although, of course, there remain sources of uncertainties in CCM simulations.

Current CCM simulations predict that global annually averaged total column ozone returns to 1980 levels before the middle of the century, which is earlier than when the stratospheric halogen loading returns to 1980 levels (Figure 5). This accelerated recovery is driven by changes in climate, in particular decreasing stratospheric temperatures. Simulated changes in tropical total column ozone between 1960 and 2100 are generally small, but 2100 tropical ozone is predicted to be slightly lower than that in 1960. For the midlatitudes, CCMs predict a different development in the two hemispheres. In the northern midlatitudes, the models predict a return of annually averaged total ozone values to 1980 values between 2015 and 2030, while for Southern Hemisphere midlatitudes the return to 1980 values is predicted to occur between 2030 and 2040.

The difference between southern and northern midlatitudes is linked to a simulated strengthening of the poleward transport of ozone in the Northern Hemisphere and the impact of Antarctic ozone loss in the Southern Hemisphere. Spring total ozone values in the Arctic are projected to return to 1980 values two to three decades before the polar halogen loading returns to 1980 levels and to exceed 1980 levels thereafter. As CCMs generally underestimate the Arctic ozone loss, however, there is the possibility that the predicted return is too early. The latest time for recovery to 1980 conditions is projected for the Antarctic, namely recovery to 1980 conditions after mid-century. However, also in the Antarctic, ozone is predicted to return to 1980 conditions before the stratospheric halogen loading in this region returns to 1980 levels (Figure 5).

All the models depend on halogen levels declining as prescribed by the amended and adjusted Montreal Protocol. This includes future replacement of presently unregulated hydrofluorocarbons, halons, and other bromine-bearing compounds such as methyl bromide. If the emissions of these compounds

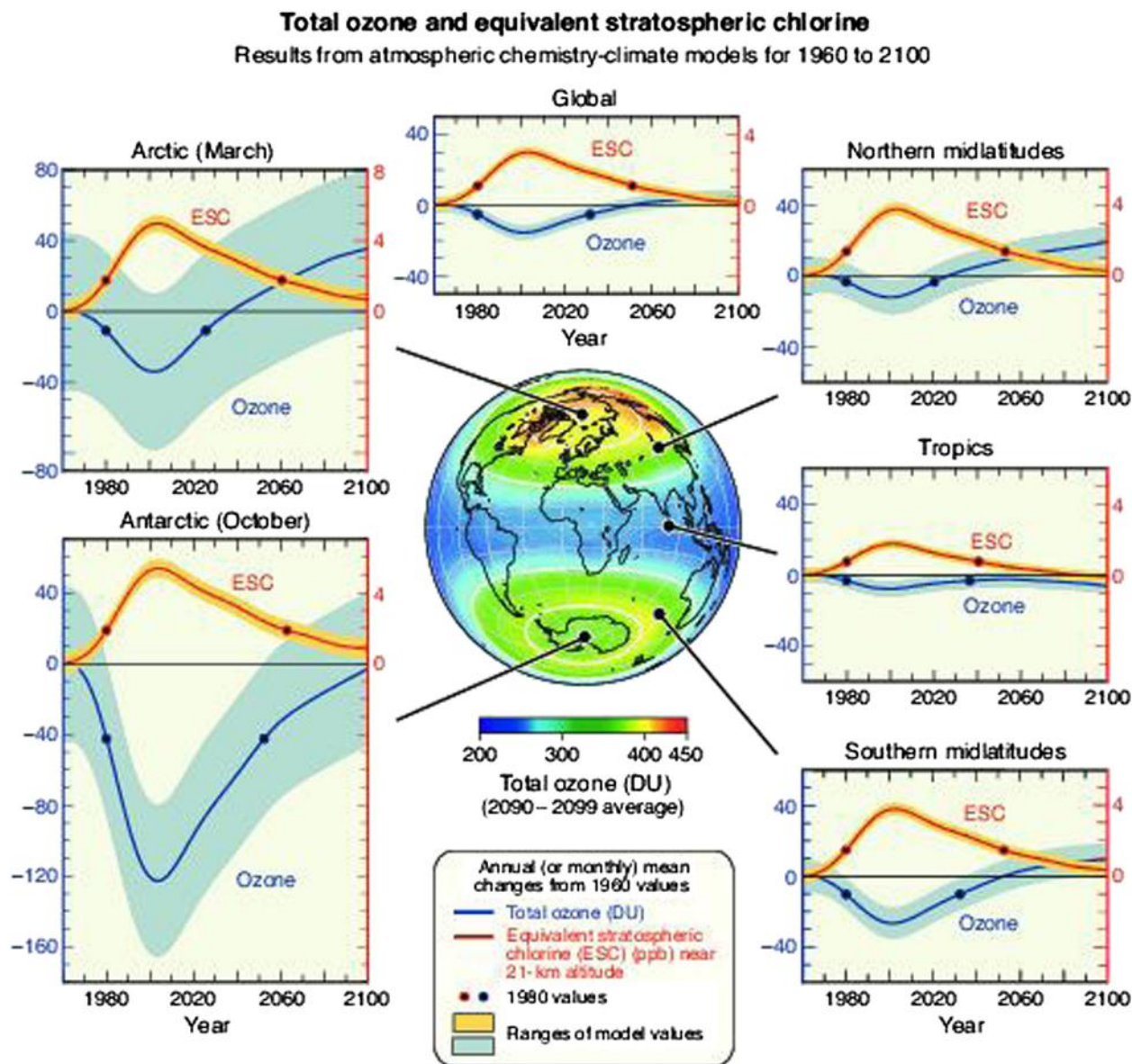


Figure 5 Long-term changes in total ozone and ESC. Values shown here are deduced from the simulations of CCMs that take into account the effects of ODSs and climate change on stratospheric ozone. Taken from WMO, 2011. Scientific Assessment of Ozone Depletion: 2010. Global Ozone Research and Monitoring Project, Report No. 52. World Meteorological Organization.

do not decline as prescribed by the Protocol, because of continued production and/or emission in developing countries who were provided special dispensations in the Protocol, then the predictions of recovery will, of course, not be accurate.

Observing the Recovery

The long-term variability of stratospheric ozone is generally estimated using multiregression statistical models that quantify the relationship between ozone and a variety of different explanatory variables describing both natural and anthropogenic forcings. Important examples for the natural forcings are the annual cycle of ozone, the 11-year solar cycle, the

quasi-biennial oscillation (QBO), and the volcanic effects. These forcings are extracted from the long-term time series simultaneously with trend components representing the effect of ODSs on ozone. The EESC is commonly used as the proxy representing the ODS effect in statistical models and different statistical techniques are used to test the statistical significance of the EESC-related terms (WMO, 2011).

Upper Stratosphere

In the upper stratosphere, a region where ozone changes are closely controlled by changes in EESC, the first stage of recovery has been reached, namely a statistically significant deviation of ozone from the previously observed linear decline. Until the

late 1990s, upper stratospheric ozone in midlatitudes declined by several percent per decade, but since then, the steep decline has ceased and ozone levels have increased by about 2.5% until 2011 (Figure 6; WMO, 2011). Upper stratospheric ozone is affected by parameters other than the change of ODSs, in particular by changing upper stratospheric temperatures and methane abundances. However, the good agreement of observed ozone decline between 1980 and the late 1990s with CCM simulations provides evidence that increases in ODSs over this period are responsible for the observed behavior (WMO, 2011). Likewise, the good fit of the observed ozone anomalies by EESC after the most important natural sources of ozone variability have been removed, provides strong evidence that the slowing of the upper stratospheric ozone decline can be attributed to the changes in ODSs.

Midlatitude Lower Stratosphere

At midlatitudes, unequivocal detection of ozone recovery is difficult. Here, the signal of ODS-driven ozone change is small

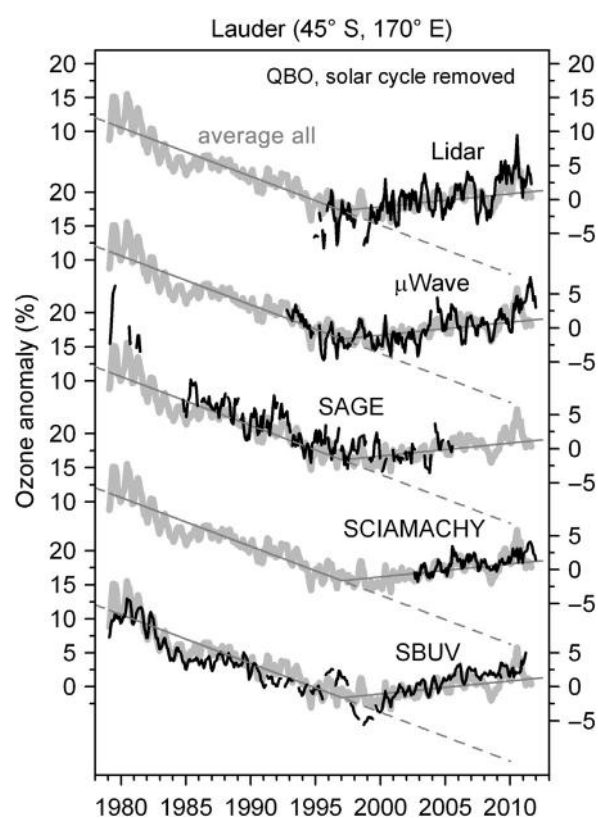


Figure 6 Ozone anomalies above the southern midlatitude station Lauder (New Zealand) averaged over the altitude range from 35 to 45 km after subtraction of the estimated QBO and solar cycle effect. For the satellite experiments Stratospheric Aerosol and Gas Experiment (SAGE), Solar Backscatter Ultraviolet Instrument (SBUV), and Halogen Occultation Experiment (HALOE), zonal mean data are shown. The underlaid thick gray line is the average anomaly record. The gray trend line gives the 1979–96 linear trend of the average record at each station, extrapolated after 1996. Natural variability due to the QBO and solar cycle is removed. Based on Steinbrecht et al. *Journal of Geophysical Research* 11, D10308, doi:10.1029/2005JD006454; figure courtesy of Wolfgang Steinbrecht, Deutscher Wetterdienst.

compared to natural variability, which makes difficult the detection of an ozone increase. However, observations show that total ozone in the midlatitudes (30–60°) of both hemispheres have remained at the same level for the period 1996–2009; namely 6 and 3.5% below the 1964–80 average for the Southern and Northern Hemispheres, respectively. The observed ozone decline in the lower stratosphere (20–25 km) between 1980 and 1996 is about 7% in both hemispheres. For the period 1996–2009, no statistically significant change is observed in the lower stratosphere in the Southern Hemisphere and a small increase by about 2.5% in the Northern Hemisphere (WMO, 2011).

Antarctica

In recent years, ozone has been totally destroyed in the heart of the ozone hole region at 15–20 km, which means that, in these regions, concentrations of active chlorine are greater than is required to destroy all the ozone available. Thus, this region is not expected to be an early indicator of the beginning of ozone recovery. However, at both the horizontal and vertical boundaries of the ozone hole region ozone loss is not saturated and thus in these regions there is an opportunity for detection of the beginning of recovery.

The horizontal extent of the ozone hole can best be observed by nadir-viewing satellite instruments, which detect total column ozone. It has been customary to use the 220 Dobson unit (DU) contour to define the outer boundary of the springtime Antarctic ozone hole as this is the value at which a steep gradient in ozone exists with substantial depletion internal and minimal depletion external to the 220 DU contour. Figure 7 shows the magnitude of the area interior to this contour, averaged for the springtime period from 9 September to 13 October, as a function of time since regular satellite measurements began in 1979. Since the early 1990s, this parameter has been in the range of $22 (\pm 2) \times 10^6 \text{ km}^2$. This is equivalent to the area poleward of about 66° S latitude. All but the tip of the Antarctic Peninsula lies internal to this area, so that the ozone hole defined in this manner covers essentially all of Antarctica.

At the boundary of the depletion region, stratospheric temperatures are not as low as internal to the boundary and thus the heterogeneous chemistry, which is responsible for chlorine activation, does not proceed as readily. Thus, assuming that temperatures will not change substantially with time, the area enclosed by the springtime 220 DU contour at maximum depletion could be a sensitive indicator of the beginning of ozone recovery in Antarctica. Here, dynamically induced variability must be discriminated from the change of ozone loss driven by reductions in ODSs. The most notable example of dynamically induced changes in the area enclosed by the 220 DU contour is the year 2002 (Figure 7), when a sudden stratospheric warming occurred in the Antarctic stratosphere.

The vertical extent of the ozone hole can be observed with balloon-borne instruments and has been monitored annually since 1986 at the South Pole. Figure 8 shows vertical ozone profiles measured at the South Pole during the ozone hole maximum depletion at the beginning of the current continuous measurement period in 1986 and during the ozone hole period

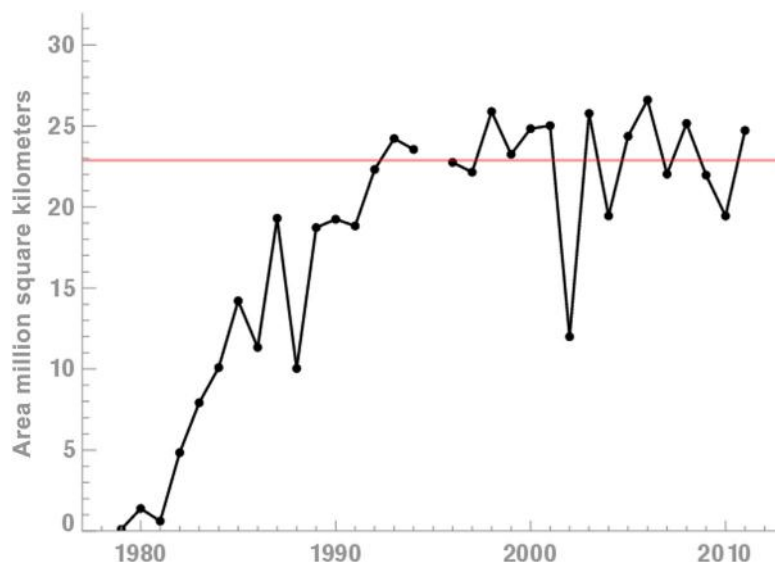


Figure 7 Geographical area of the 220 DU contour over Antarctica between 9 September and 13 October from Total Ozone Mapping Spectrometer (TOMS) satellite data. Red line indicates the area of the Antarctic continent. Figure courtesy of Paul Newman, NASA Goddard Space Flight Center.

in 2000. These profiles are compared to profiles measured at the South Pole prior to the advent of the ozone hole phenomenon during the 1967–71 period. The progression toward total ozone destruction in the 15- to 20-km region is clear. It also is clear that ozone depletion has progressed to higher altitudes during the 1986–2000 period, with a sharp top to the ozone hole at about 21 km in 1986 and about 24 km in 2000. Barring major temperature trends in this region, a decline in the top of the ozone hole would be an indicator of ozone recovery in Antarctica. However, near complete removal of ozone up to altitudes of 21 km has been observed in recent years and no observations of a decline in altitude of the top of the ozone hole have been reported (WMO, 2011; Hofmann et al., 2009).

Another parameter that will be sensitive to a decay of the stratospheric halogen loading is the rate of ozone loss in the main ozone loss region (14–21 km) during September, the period when ozone is declining rapidly (Figure 9). Pre-1990 values were in the range of 2.4 DU per day. In recent years, the value has been about 3 DU per day with large year-to-year variations, but no definitive signs of ozone recovery (Hofmann et al., 2009).

Arctic

The detection of the recovery of Arctic springtime ozone loss is expected to be more difficult than elsewhere in the stratosphere because models suggest that strong ozone losses in the Arctic are possible for many years owing to the more dynamic situation in the Arctic compared to the Antarctic (WMO, 2011). The interannual variability in the degree of springtime Arctic ozone depletion is much too large to allow any simple observations of the beginning of recovery of the Arctic ozone layer. Indeed, in March 2011, ozone concentrations in the Arctic stratosphere were the lowest ever recorded and ozone loss was

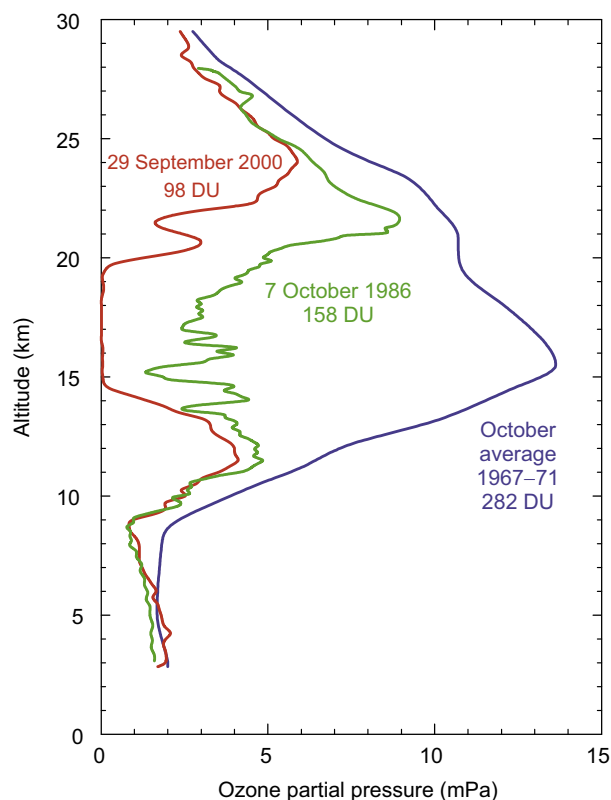


Figure 8 Ozone vertical profiles obtained with balloon-borne ozone-sondes at South Pole Station at the time of maximum ozone depletion in spring. Shown are profiles for the years 1986 (when continuous measurements began) and 2000. The recent measurements are compared to those made during the 1967–71 period.

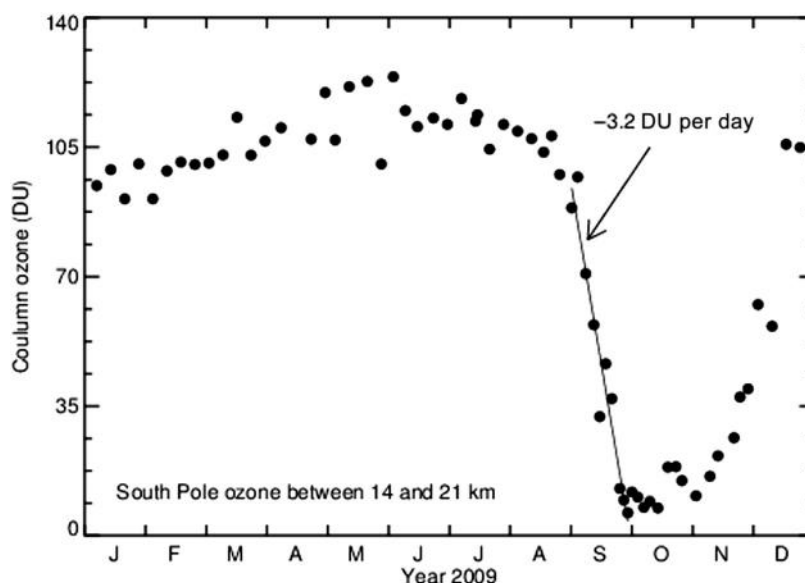


Figure 9 South Pole Station ozone between 14 and 21 km (the region of maximum ozone depletion) as a function of time during the year 2009, with a determination of the ozone loss rate during September.

comparable to that regularly observed in the Antarctic (Figure 2; Manney et al., 2011).

Summary and Conclusions

In the upper stratosphere, the first stage of recovery has been reached, namely a statistically significant deviation of ozone from the previously observed linear decline due to changing halogen loading. Ozone abundances in midlatitudes and in Antarctic spring have similarly ceased to decline since about a decade, but an unequivocal attribution to changes in the stratospheric halogen loading is difficult. The observed ozone decline in the lower stratosphere (20–25 km) between 1980 and 1996 is about 7% in both hemispheres and a small increase by about 2.5% is observed in the Northern Hemisphere thereafter.

There are different diagnostics that allow the detection of recovery of stratospheric ozone in Antarctic spring: the area of the 220 DU contour in late September/early October, the vertical extent of the area of extreme ozone depletion, and the rate of ozone loss (in DU per day) in September. All these measures show a certain natural variability and a cessation of ozone decline since more than a decade, but no clear signs of recovery caused by a reduction of the stratospheric halogen loading. Arctic ozone loss is very variable from year to year and the lowest ever recorded ozone concentrations occurred in March 2011.

A substantial recovery of the stratospheric ozone layer due to a reduction of the stratospheric halogen loading is projected by model simulations, but the resulting changes in ozone will depend strongly on geographical region. Simulated changes in tropical total column ozone between 1960 and 2100 are small. In the midlatitudes, the models predict a return of annually averaged total ozone values to 1980 values between 2015

and 2030 and in 2030–40, in the Northern and Southern Hemispheres, respectively. Spring total ozone values in the Arctic are projected to return to 1980 values in about 2040, two to three decades before the polar halogen loading returns to 1980 levels and to exceed 1980 levels thereafter. For the Antarctic, recovery to 1980 conditions is predicted to occur after midcentury.

However, the atmosphere – because of climate change – will develop into a different state than in 1980 over the course of the century, i.e., over the time period when ozone recovery will occur. Thus, the ozone layer in 2100 will have recovered from the effects of ODSs but, due to climate change, will be in a different state than it was in 1960 when the effects of ODSs on the ozone layer were not yet noticeable.

See also: Ozone Depletion and Related Topics: Ozone Depletion Potentials; Ozone as a UV Filter; Photochemistry of Ozone; Surface Ozone (Human Health); Surface Ozone Effects on Vegetation.

Further Reading

- Farman, J.C., Gardiner, G.G., Shanklin, J.D., 1985. Large losses of total ozone in Antarctica reveal seasonal ClO_x/NO_x interaction. *Nature* 315 (1985), 207–210.
- Hassler, B., Bodeker, G.E., Solomon, S., Young, P.J., 2011. Effects on Antarctic total ozone observations at various stations. *Geophysical Research Letters* 38, L01805. <http://dx.doi.org/10.1029/2010GL045542>.
- Hofmann, D.J., Oltmans, S.J., Harris, J.M., Johnson, B.J., Lathrop, J.A., 1997. Ten years of ozonesonde measurements at the South Pole: implications for recovery of springtime Antarctic ozone. *Journal of Geophysical Research* 102, 8931–8943. <http://dx.doi.org/10.1029/96JD03749>.
- Hofmann, D.J., Jonson, B.J., Oltmans, S.J., 2009. Twenty-two years of ozonesonde measurements at the South Pole. *International Journal of Remote Sensing* 30, 3995–4008.

- Hood, L.L., 2000. Trends in lower stratospheric circulation and their effects on column ozone trends at northern midlatitudes during the 1979–1998 period. In: Proceedings of the Quadrennial Ozone Symposium. Hokkaido University, Sapporo, pp. 49–50.
- Hood, L.L., McCormick, J.P., Labitzke, K., 1997. An investigation of dynamical contributions to midlatitude ozone trends in winter. *Journal of Geophysical Research* 102, 13079–13093.
- IPCC, 2007. In: Houghton, J.T., et al. (Eds.), *Climate Change 2007 the Science of Climate Change*. Cambridge University Press, Cambridge.
- Manney, G.L., et al., 2011. Unprecedented Arctic ozone loss in 2011. *Nature* 478, 469–475. <http://dx.doi.org/10.1038/nature10556>.
- Montzka, S.A., Butler, J.H., Elkins, J.W., 1999. Present and future trends in the atmospheric burden of ozone-depleting halogens. *Nature* 398, 690–694.
- Müller, R. (Ed.), 2012. *Stratospheric Ozone Depletion and Climate Change*. Royal Society of Chemistry. <http://dx.doi.org/10.1039/9781849733182>. ISBN: 978-1-84973-002-0.
- Newman, P.A., Oman, L.D., Douglass, A.R., Fleming, E.L., Frith, S.M., Hurwitz, M.M., Kawa, S.R., Jackman, C.H., Krotkov, N.A., Nash, E.R., Nielsen, J.E., Pawson, S., Stolarski, R.S., Velders, G.J.M., 2009. What would have happened to the ozone layer if chlorofluorocarbons (CFCs) had not been regulated? *Atmospheric Chemistry and Physics* 9, 2113–2128. <http://dx.doi.org/10.5194/acp-9-2113-2009>.
- Shindell, D., Rind, D., Lonergan, P., 1998. Increased polar stratospheric ozone losses delayed eventual recovery due to increasing greenhouse gas concentrations. *Nature* 392, 589–592.
- WMO, 1999. *Scientific Assessment of Ozone Depletion: 1998*. Global Ozone Research and Monitoring Project, Report No. 44. World Meteorological Organization.
- WMO, 2011. *Scientific Assessment of Ozone Depletion: 2010*. Global Ozone Research and Monitoring Project, Report No. 52. World Meteorological Organization.
- Yang, E.S., Cunnold, D.M., Newchurch, M.J., Salawitch, R.J., McCormick, M.P., Russell, J.M., Zawodny, J.M., Oltmans, S.J., 2008. First stage of Antarctic ozone recovery. *Journal of Geophysical Research* 113 (D10), D20308. <http://dx.doi.org/10.1029/2007JD009675>.

Surface Ozone Effects on Vegetation

M Ashmore, University of York, York, UK

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

Current concentrations of surface ozone in many parts of the world can reduce crop yields, impair forest growth, and change plant species composition. Predicted increases in northern hemispheric background ozone concentrations are likely to increase these impacts over the first half of this century. Ozone effects on vegetation are modified by climate and CO₂ concentrations, while ozone itself can change ecosystem carbon budgets. For these reasons, surface ozone needs to be considered as an important element of global environmental change over this century.

Introduction

The effects of surface ozone on vegetation were first demonstrated in the Los Angeles Basin in the 1950s. Since then, extensive field and experimental research has demonstrated the wider impacts of ozone on crops and forests, especially in North America and western Europe. Surface ozone is now a global problem, with evidence of a trend of rising background northern hemispheric concentrations and increasing evidence of the impacts of ozone on vegetation in Asia, Africa, and Latin America. During a photochemical smog episode, a range of chemical species is formed, but there is little doubt that ozone is by far the most important of these in terms of effects on vegetation. Although evidence exists of the effects of other oxidants, such as peroxyacetyl nitrate and hydrogen peroxide, on vegetation, these are of minor significance compared with effects of ozone.

It is important to emphasize that ozone is just one environmental factor influencing vegetation, and that there are important interactions between ozone and climatic factors, soil conditions, and pests and diseases. Thus, any assessment of its impact on vegetation needs to consider the pollutant in the wider environmental context. Hence, assessment of ozone impacts on vegetation in different regions of the planet, and of the potential impact of the projected increases in background surface ozone concentrations during this century, must consider interactions with varying climatic conditions.

This article aims to provide an overview of surface ozone effects on vegetation. The first three sections provide an overview of methods of investigation, of the key mechanisms of damage, and of the relationships between ozone exposure and plant response. This is followed by two sections that provide more specific information about the two key impacts of surface ozone – on food production and on ecosystem services. This article concludes with consideration of possible future impacts of ozone, providing a global perspective and considering interactions with other components of global environmental change.

Methods of Assessing Ozone Impacts on Vegetation

A variety of methods have been used to study the impacts of ozone on vegetation, ranging from controlled laboratory chambers, which clearly link cause and effect but do so under

highly unrealistic conditions, through to field observations, which provide no certainty about the link between cause and effect. The most important source of information on the impacts of ozone on crop yield has been the use of exposure chambers in the field, in which ozone exposures can be modified and controlled, either by filtration of ambient pollutants or by the addition of ozone in temporal patterns that simulate ambient concentration profiles. The most commonly used type of chamber, the open-top chamber, provides microclimatic conditions that are comparable, but not identical to, those outside. [Figure 1](#) shows one such facility, in which different ozone treatments can be assigned to replicate chambers.

However, open-top chambers do modify the microclimate, for example, by increasing air and soil temperatures. They also significantly change the interactions between plants, insects, and fungal pathogens, while their size prevents their use for large tree species. One alternative method is the use of open-air fumigation systems, in which ozone is released over outdoor plots; this allows multiyear exposure of ecosystems under unmodified climatic conditions, the opportunity to study interactions between ozone exposure and extreme stress events such as storms, and realistic interactions with pest and disease organisms. [Figure 2](#) shows an example of such experiment in a field of soybean; ozone is released from different points within the ring depending on the wind direction, and replicate rings (not shown in [Figure 2](#)) can be assigned to different ozone treatments. However, such systems can only be used to examine the effects of increasing ozone concentrations above background, and they cannot be used to assess the effects of reducing ozone exposures. For crop species, ozone-protectant chemicals such as ethylene diurea (EDU) have also proved an effective means of identifying the effects of ambient ozone exposure.

There are enormous technical and logistical challenges and costs to exposing more mature trees to ozone over long periods of time under field conditions. However, two studies have done so in recent years, and have provided important new insights into the long-term effects of ozone exposure. The first of these (the AspenFACE field fumigation study in the northern US) applied elevated ozone and elevated CO₂ concentrations for a period of over a decade to communities of three important North American trees – aspen, birch, and maple – that were grown from seedlings. [Figure 3](#) provides an aerial photograph of this experiment, showing rings that have been assigned to different ozone and CO₂ concentrations. The second of these studies, based in the Kranzberg forest, in Germany, applied



Figure 1 Overview of the University of Newcastle open-top chamber facility at Close House, west of Newcastle-upon-Tyne, UK. The 16 open-top chambers are assigned to different controlled ozone treatments, depending on experimental objectives.



Figure 2 Overview of the University of Illinois SoyFACE experimental site, situated near Urbana, Illinois, US. Each ring within the soybean crop is assigned to a different controlled ozone and CO₂ treatment, depending on experimental objectives. Image taken from <http://www.igb.illinois.edu/soyface>.

ozone for 8 years to a mature mixed spruce and beech canopy that was planted about 60 years ago.

The limited duration of any chamber or field release experiment means that field observations may be the only feasible way of assessing the long-term impacts of ozone over decades. Indeed, field observations have played an important role in assessment of ozone effects on vegetation, especially where concentrations are high enough to produce characteristic visible symptoms of injury. However, spatial and temporal associations between plant response and ozone

exposure are much more difficult to demonstrate in regions with lower concentrations, because gradients in ozone exposure are often confounded with gradients in climatic and other factors.

Mechanisms of Impact

Ozone has impacts at different levels of organization, from the cellular to individual organs and plants, through to plant



Figure 3 Aerial view of the AspenFACE experimental site, situated at the Harshaw Experimental Forest near Rhinelander, Wisconsin, US. Each ring is assigned to a different controlled ozone and CO₂ treatment. Image taken from <http://aspenface.mtu.edu>.

communities and ecosystems. The effects of surface ozone can be conceptualized as having four stages:

1. Ozone is transferred from the boundary layer of the atmosphere to sites of damage within the leaf. There is little evidence that ozone deposition to external plant surfaces or to soil surfaces has any adverse effect; rather it is the flux of ozone through the stomatal pores, which regulate the exchange of gases between the atmosphere and the intercellular spaces within the leaf that is important. However, direct effects of ozone, not mediated through damage within the leaf, may occur on reproductive structures (e.g., pollen, flowers, and fruits).
2. Ozone causes damage and physiological changes within the leaf. The primary site of damage is thought to be the cell membrane, either directly due to ozone or indirectly due to secondary radical species formed by reactions with organic molecules within the leaf. At high concentrations, damage to the cell membrane induced by short-term exposure to ozone can cause cell collapse, leading to visible leaf injury. The most important long-term effect of ozone exposure within the leaf is a reduction in net CO₂ fixation as a result of both reduced photosynthetic activity and the increased rates of respiration, which are associated with maintenance and repair of damaged tissue.
3. Effects of ozone within the leaf have a range of other effects within the plant. Most importantly, ozone often reduces the translocation of assimilates to other parts of the plant, such as the roots and grains. The rate of production of new leaves is accelerated in some species in response to ozone stress, thus allowing the maintenance of plant growth rates, but

this is often at the expense of assimilate allocation to, and growth of, the root system and of storage organs. Ozone also accelerates the rate of leaf senescence, which in turn reduces the period of leaf growth with a positive carbon balance. The reduced carbon assimilation typically leads to reduced growth rates. Whether these in turn lead to reduced yields of agricultural or forest crops will depend on the allocation of assimilates to the harvested part of the plant, compared with other organs.

4. Effects of ozone within the plant have a wider range of ecological implications. These include effects on species composition and diversity, biogeochemical cycles, soil chemistry, and soil biology. Some of these effects are a direct consequence of the reduced photosynthesis and carbon assimilation in the leaf, but other mechanisms, such as interference with hormonal signaling by ozone, are also important. To provide one example of the range of effects that might occur within a simple experimental ecosystem, a study of a constructed meadow in Finland reported that ozone caused visible leaf injury, delayed the onset of spring growth, reduced leaf and root growth, delayed flowering, reduced the number of berries, reduced rust infection of the leaves, reduced nitrogen levels in the soil, reduced emissions of carbon dioxide (CO₂) and nitrous oxide from the soils, and reduced the microbial biomass in the soil.

The extent to which ozone entry into the leaf can lead to adverse effects on plant growth and yield depends on a range of defense and compensatory mechanisms. Within the leaf, antioxidant systems can reduce the impact of ozone. In particular, the levels of ascorbate, and other antioxidants, in the cell wall

may be critical in reducing the flux of molecules of ozone or secondary free radicals to sites of damage in the cell membrane. There is also clear evidence that exposure to ozone can result in upregulation of antioxidant systems in the plant through effects on gene expression.

The effect of a given ozone concentration is not fixed, but depends on a wider range of factors, which can be grouped into two broad categories. The first category is external factors, such as climate, soil type, and CO₂ concentration, which modify the responses of vegetation to ozone. As examples of the interactions with climate:

- Severe visible injury to sensitive species is most often found in periods when high ozone levels are associated with high atmospheric humidity, enhancing the uptake of ozone into the leaf.
- The effect of ozone on wood production and stem growth is lower on soils with a lower water-holding capacity, where drought conditions are more frequent, and greater in moist years than in dry years.
- Most plant species show little injury if exposed to ozone in the dark, when stomata are closed and there is little flux of ozone into the leaves, but in those species, which do retain open stomata at night, ozone damage may be greater, because of the lower antioxidant capacity within the leaf in the dark.

The second category is internal factors. Physiological, anatomical, and biochemical factors that vary between, and within, plant species, strongly influence the size of any impacts of ozone. It is long established that there is a large variation between species in their sensitivity to ozone. There is also considerable genetic variation in sensitivity within species, and gene loci associated with an increased tolerance of ozone stress have been identified. For crop species, the difference in sensitivity between cultivars may reflect inadvertent selection in the breeding process, although there is little evidence that modern crop cultivars, bred under greater ozone concentrations, are more tolerant of ozone. Indeed, some data suggest the opposite; for example, that modern wheat cultivars are more sensitive than those bred earlier in the twentieth century, possibly because characters that are associated with higher yield potentials are also associated with higher ozone sensitivity. Outside the agronomic environment, there is evidence that natural selection is operating to favor more ozone-tolerant genotypes in areas with higher ozone exposures.

Exposure-Response Relationships

While the mechanisms leading to adverse impacts of surface ozone on vegetation can be described qualitatively, policy assessments aim to quantify the size of the regional impacts of ozone, and the benefits of measures to reduce precursor emissions. This has typically involved an empirical approach, defining the relationships between ozone exposure and a limited range of measures of plant response related to plant growth or yield. A major issue in defining such exposure-response relationships for the cumulative effects of ozone over one or more growing seasons is the method used to summarize the seasonal ozone exposure. Different exposure

indices have been devised, describing different components of the seasonal ozone exposure, often giving higher weights to periods with higher concentrations, which are assumed to have a greater biological impact.

For annual crops, the main approach has been to fit statistical relationships between seasonal ozone exposure and final crop yield, based on experimental data, typically derived using open-top chambers. The first national program to determine exposure-response relationships for crops was the US National Crop Loss Assessment Network, which was established in the late 1970s and examined 10 crops, representing 85% of the cropped area of the United States. The experiments involved a range of ozone concentrations to generate relationships between ozone exposure and crop yield. Ozone exposure was expressed as the seasonal mean ozone concentration during either 7 or 12 h during the day. For a seasonal 7-h mean ozone concentration of 50 ppb, compared with a baseline of 20–25 ppb, estimated yield reductions for the most sensitive crops, such as soybean and cotton, exceeded 10%; for winter wheat they were slightly under 10%; and for the most tolerant crops, such as rice, barley, and sorghum, they were below 5%.

In western Europe, a similar experimental approach was applied to a smaller number of crops. A new cumulative seasonal ozone exposure index was developed, termed the AOT40, which sums the positive differences between hourly mean ozone concentrations and 40 ppb during all daylight hours over the growing season. The use of this index allowed a common linear exposure-response relationship to be fitted to spring wheat data from experimental studies in six different countries, with a 10% yield loss corresponding to an AOT40 value of about 5000 ppb h.

However, caution is needed in interpreting the exposure-response relationships derived from such experiments. In open-top chamber exposure-response experiments, ozone is added simultaneously in all treatments under the same climatic conditions. In contrast, when the impacts of ozone in different locations are compared, for example, across the continent of Europe, the very different climates may modify the impacts of ozone. For example, in the hotter and drier climates of southern Europe, the same ozone exposure may have less effect than under the cool moist climate of Scandinavia. Hence, although the highest ozone concentrations have the greatest impact under experimental conditions, under field conditions it is not necessarily the highest ozone concentrations that are most significant.

One reason for this is that highest ozone concentrations tend to occur under meteorological conditions that limit ozone flux through the stomata into the leaf. Since ozone flux from the atmospheric boundary layer into the leaf is a critical factor for adverse effects on vegetation, environmental factors that influence this flux may modify ozone impacts. Under most conditions, the most important factor controlling ozone flux is the stomatal conductance of the plant canopy; thus, high vapor pressure deficits or high soil moisture deficits, which are often associated with high ozone concentrations, tend to reduce ozone impacts.

Accordingly, rather than relate plant response to external ozone concentrations or exposure indices, it is mechanistically preferable to use the estimated flux of ozone into the leaf. There

is now a significant body of evidence, both statistical and biological, that an index based on the modeled accumulated seasonal O_3 flux through the plant stomata is better related to effects of O_3 on crop yield and forest growth than one based on external concentrations. Use of a flux-based index gives a very different spatial distribution of risk than does use of exposure-based indices, and evidence suggests that a flux-based evaluation is a better representation of the spatial distribution of impacts on vegetation across Europe. Flux-based dose-response relationships have now been developed for some of the most important environmental effects of O_3 – on crop yields, on carbon storage in forest trees, and on the food quality of pasture for grazing animals – and have been applied to assess the benefits of precursor emission reductions across Europe.

Effects on Food Production

Visible leaf injury to crop species has been reported in the field on many occasions since the first confirmed reports in the late 1950s in the United States. A list of crop species for which documented evidence of visible injury in the field following episodes of high ozone concentrations has been obtained in Europe includes beans, beetroot, chicory, clover, clementine, courgette, grapevine, maize, oats, onion, parsley, peach, peanut, potato, radish, soybean, tobacco, tomato, watermelon, and wheat. This indicates the wide range of crop species that can be affected. [Figure 4](#) provides an illustration of some of these typical symptoms of visible injury. In some of these cases, the appearance of visible injury on leaves would directly reduce the economic value of the crop, but in most cases it is the long-term effect on the yield of the harvested part of the crop that is important. Visible leaf injury to shrub and forest species has also been reported in the field in Europe and North America, especially on sensitive deciduous species, such as poplar, ash, and black cherry.

There have been many experimental studies to demonstrate the significant impact of ozone exposures on crop yield in western Europe and North America. The exposure-response studies described in the previous section have been integrated with databases on crop distributions and ozone concentrations to assess the benefits to agriculture of measures to reduce ozone

concentrations. In the United States, it was estimated that a national decrease in ozone concentrations of 40% would provide a net annual economic benefit of \$3000 million, or 2.8% of the national production. A similar recent exercise in Europe estimated that reducing current ozone exposures would have direct economic benefits to agriculture of about €2000 million annually. Such national or international estimates disguise the much greater losses in particular crops and regions. For example, yield losses due to ozone levels in the 1980s in California were estimated to be in the range 20–25% for sensitive crops such as cotton, bean, and onions, an order of magnitude higher than the overall national estimate.

Such estimates have many uncertainties, including how well the exposure-response functions represent the real responses in different regions. They ignore the possible indirect effects of ozone through altered pest and pathogen damage; ozone has been shown to have both positive and negative effects on the prevalence of pests and diseases. Several major systemic pesticides reduce the impacts of ozone, and this may be a significant factor in the field. The estimates also ignore effects on crop quality; for example, ozone has been shown to reduce the oil content of oilseed rapeseeds, in addition to seed yield, producing a significant additional economic loss for the producer. In summary, these estimates can provide only an approximate guide to the possible regional economic significance of ozone.

In recent years, it has been possible to link global models of ozone production with exposure-response relationships to estimate effects on global food production. The results of one such exercise are shown in [Figure 5](#). This estimated that global yields of four staple crops (wheat, rice, maize, and soybean) were reduced by between 2 and 16% by modeled ozone concentrations in the year 2000. This is broadly consistent with a recent meta-analysis of experimental data for potato, barley, rice, wheat, bean, and soybean, which estimated that, at the current global ambient ozone concentrations of 30–50 ppb, mean yield losses for the different crops range from 5 to 19%.

Importantly, there are significant regional variations in the estimated yield losses, with those in densely populated parts of South and East Asia, for example, being substantially greater than the global average. There is an increasing body of experimental evidence to quantify effects on crop yields in these regions, and recent analysis of these experimental data has

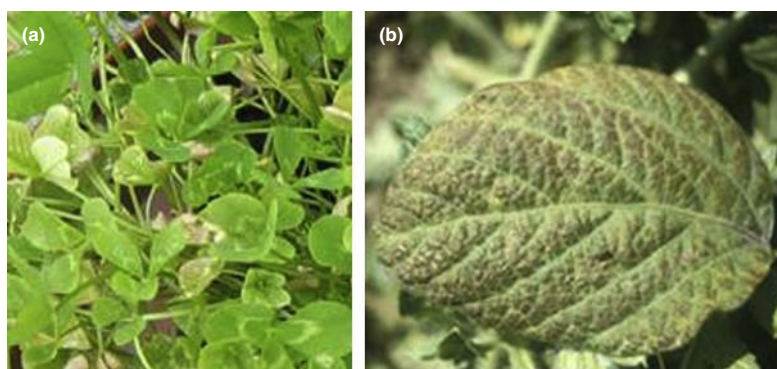


Figure 4 Typical symptoms of visible leaf injury caused by elevated ozone concentrations in the field in Europe, showing leaves of (a) clover and (b) soybean. Image taken from <http://icpvegetation.ceh.ac.uk>.

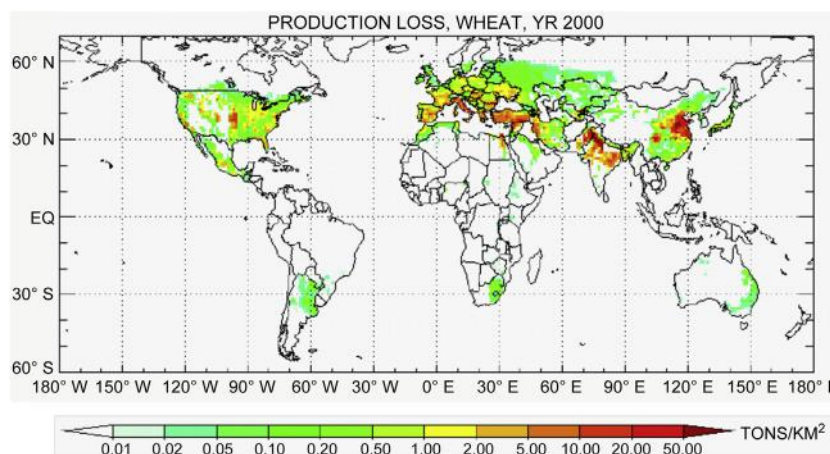


Figure 5 Modeled global distribution of wheat yield loss for the year 2000. Reproduced with permission from van Dingenen, R., Dentener, F.J., Raes, F., Krol, M.C., Emberson, L., Cofala, J., 2009. The global impact on crop yields under current and future air quality legislation. *Atmospheric Environment* 43, 604–618.

suggested that the effects of ozone on yields of wheat and rice are greater than would be predicted based on exposure–response relationships derived from European and North American data. This evidence highlights ozone as a significant threat to crop yields in these regions.

Effects on Ecosystem Services

Managed and unmanaged ecosystems provide a wide range of services for human society. Over the past decade, ecologists have developed methods to categorize these services, and have begun to quantify and value the impacts of anthropogenic stresses, such as air pollution, on them. These ecosystem services can be divided into:

1. Provisioning services, e.g., the food, fuel, and fiber provided by ecosystems.
2. Regulating services, e.g., the ways in which ecosystems contribute to regulation of climate, water flow and quality, and pollination.
3. Cultural services, e.g., aesthetic and educational qualities of ecosystems, and their contribution to tourism and recreation.

The major effects of surface ozone on ecosystem services are described below. This summary does not include every type of ecosystem service, but focuses on those for which there is clear evidence of the adverse effects of ozone.

In terms of provisioning services, the effects of ozone on arable food production were described in the previous section, but effects on forests are also important. Numerous experimental studies in chambers have demonstrated the adverse effects of ozone on the growth of seedlings or saplings over periods of 1–5 years. Meta-analyses allow the identification of consistent effects of ozone in these studies, including reduced wood yield, shorter height, and smaller trunk diameter. Although some studies have shown effects of both drought and elevated CO_2 in reducing the impacts of ozone, these effects are not consistently found in the meta-analysis. Overall, the

evidence from this analysis suggests that current global ozone concentrations reduce tree biomass by about 7%.

However, there are major questions over how the results from these short-term experiments of 1–5 years' duration can be extrapolated over a tree's lifetime. Contrasting results from two long-term open-air fumigation experiments illustrate the uncertainties in these extrapolations. In the AspenFACE study, results after the first few years showed a clear effect of ozone in reducing the growth of aspen trunks, but after 12 years there was no effect of ozone on aspen trunk growth. The reason for this is that the original planting included aspen clones differing in sensitivity to ozone, and the tolerant clones had a selective advantage over the sensitive clones in elevated ozone. Hence after 12 years, the effect of ozone was dominated by that of the tolerant clones, and was lower than that in the early years of the experiment. In contrast, doubling the ambient ozone concentrations in the Kranzberg free-air release system led to almost a 50% reduction in trunk growth in beech, but not in spruce, a greater effect than initial observations of changes in stem diameter suggested. The difference between the outcomes in these two unique experiments has enormous implications for long-term modeling of ozone effects on forest production.

Furthermore, even these long-term experiments cannot capture the complexity of ozone effects on forest ecosystems over periods of decades. The area where the long-term impacts of ozone stress on forest community composition have been most intensively studied is in the mountains around Los Angeles, where effects began to be observed in the 1960s. The dominant species of these mixed-conifer forests prior to European settlement (ponderosa pine and Jeffrey pine) are also the most sensitive species to ozone. Both species have shown severe ozone injury and reduced needle lifetimes, which are associated with reduced radial growth, or years with missing growth rings. Trees affected by ozone are more susceptible to attack by bark beetles, which are often the direct cause of mortality; outbreaks of bark beetles are associated with drought years. Regeneration in these forests is now greater for species such as white fir, which are more resistant to ozone, a similar mechanism to that which led to the domination of tolerant

aspen clones in the AspenFACE experiment. Current fire-exclusion policies favor replacement of ponderosa pine and Jeffrey pine by more fire-sensitive species that also happen to be more ozone tolerant. This example illustrates the complex interactions with management practices, pest outbreaks, and climate that may influence the long-term impacts of ozone on forests.

Effects of ozone on the provisioning services provided by pastures are also economically important. Most studies suggest that total harvested production is only affected at very high ozone exposures, but that shifts in species composition, and in particular a reduction in the proportion of legume species in the harvested material, can occur at lower exposures. This has been shown to reduce the nutritional quality of the crop for grazing animals, although no study has yet shown directly a link between ozone exposure and meat or milk production.

The changes in forest production as a provisioning service also have implications for the regulating ecosystem services, and in particular climate regulation. Terrestrial ecosystems, and forests in particular, are an important component of the global carbon cycle, and reductions in the amount of carbon sequestered in ecosystems will lead to increases in atmospheric CO₂ concentrations, and hence more rapid warming of the planet. The reductions in wood production described above will reduce carbon sequestration and climate regulation, but three other effects of ozone also need to be considered:

- Effects of ozone on the sequestration of carbon below-ground are important, as well as effects on sequestration above-ground. For example, in the AspenFACE experiment, elevated ozone halved the formation of the stable fraction of soil carbon, which is not readily exchanged with the atmosphere, while the Kranzberg experiment showed that ozone reduced the CO₂ sink strength both above-ground and below-ground.
- Ozone may modify the flux of CO₂ from soils to the atmosphere through respiration. In both the AspenFACE and Kranzberg experiments, an increased rate of soil respiration was reported, but other experiments have reported no effects, or negative effects, of ozone on soil respiration.
- Ozone exposure can modify fluxes to the atmosphere of other radiatively active gases. For example, ozone alters fluxes of nitrous oxide and methane from rice paddies and peatlands, although the size and direction of these effects remain uncertain.

Overall, there is a large potential effect of ozone on climate regulation, but inconsistencies in the results from different studies, and the lack of integrated estimates of effects on the whole ecosystem carbon budget, mean that the size of these effects cannot be confidently predicted.

Ozone can reduce transpiration rates from a plant community, both because of reduced leaf area per unit ground area, and because it often causes a closure of the stomata. However, recent evidence has demonstrated that ozone exposure also interferes with the hormonal signals that control the normal closure of stomata in response to soil drying. Hence ozone exposure causes a loss of stomatal control of water loss by the plant, and can exacerbate the adverse effects of drought conditions. Although few studies have upscaled these findings to a whole stream catchment,

there is evidence that ozone can modify the ecosystem regulation of water flows provided by forests.

Pest and disease regulation is an important ecosystem service that may be modified by ozone, but the direction of these effects is very uncertain, with evidence that direct effects on the pest or disease organisms, modification of leaf surface properties, changes in food quality, increased predation, and changes in production of defense compounds are all important. Within a single experiment, a range of effects on pests and diseases may be reported. For example, effects of ozone within the AspenFACE study include increased aphid numbers, increased growth of tent caterpillars, and increased rust infection of leaves.

Ozone can affect both the regulating and cultural services provided by many natural, or seminatural, plant communities. For example, in plant communities such as traditional hay meadows or woodlands, seasonal flowering is an important cultural service attracting tourists and providing aesthetic value, as well as being linked to the regulating service of pollination. Ozone can both reduce and delay flowering, leading to detrimental effects on these ecosystem services.

Many native plant species, when grown alone, are as sensitive to short-term experimental ozone exposures as the most sensitive crop species. The range in sensitivity between and within these plant species means that competition between and within species may be modified by ozone, leading to changes in species composition and, potentially, the loss of sensitive species or genotypes, with a loss of genetic resources, another important ecosystem service. The available evidence from short-term chamber experiments with simple sown species mixtures, or transplanted cores taken from the field, clearly demonstrates that more sensitive species are at a competitive disadvantage under elevated ozone. However, as for forests, such experiments are of limited value in assessing the long-term impact of ozone on biodiversity in the complex plant communities found in the field.

Two recent long-term open-air fumigation experiments on upland grassland communities in Switzerland and the UK show contrasting effects of ozone. In the former, there were few significant effects of ozone on species composition of a species-rich grassland, suggesting that such systems are relatively tolerant to ozone. In contrast, the latter showed significant shifts in species composition in an upland hay meadow undergoing a change in management. One potential reason for the greater response in this system is the fact that a hay rattle, a keystone species in these communities (i.e., one that modifies the response of other species, in this case, through selective parasitism), appears to be very sensitive to ozone. Hence the precise effects of ozone on the ecosystem services provided by such seminatural plant communities may be hard to predict and depend on unique characteristics of individual communities.

Ozone as a Component of Future Global Environmental Change

There is evidence from Europe and North America that peak concentrations of ozone during photochemical episodes have declined over the last two decades, reflecting the effect of regional strategies to control precursor emissions. However, there is also evidence that, over the same period, annual mean

concentrations have increased. This is believed to reflect an increase in northern hemispheric background tropospheric concentrations as a result of increased precursor emissions at a hemispheric scale. Decreased peak ozone concentrations would be expected to reduce the extent of visible injury and the size of any reductions in plant growth. However, increased background concentrations over the range 30–50 ppb will increase effects on vegetation. Model projections over this century suggest that, while ozone concentrations over much of the developed world will have decreased by 2050, those in developing countries will continue to increase, as will northern hemispheric background concentrations. However, very few experiments have been designed to assess the effect of these changing exposure patterns.

Furthermore, any analysis of the implications of these future changes in surface ozone for global food production, and for global ecosystem services, needs to be conducted, not in isolation, but in the context of a wider assessment of global environmental change, for four major reasons:

1. Elevated CO₂ concentrations reduce the effect of ozone on plant growth and crop yield. This is because they decrease stomatal conductance, reducing the flux of ozone to sites of damage in the leaf, and they also increase the leaf antioxidant defense capacity. However, there are exceptions. For example in the AspenFACE study referred to earlier, elevated CO₂ concentrations prevented any adverse effects of increased ozone exposure on tree growth and many other variables. However, the effects of ozone of leaf senescence, bud burst, insect and fungal damage, and regulation of water loss were not prevented by exposure to elevated CO₂.
2. Ozone effects may be modified by climate, while ozone exposure may exacerbate the impact of climatic stress on vegetation. While drier and warmer conditions might reduce the flux and short-term impacts of ozone episodes, decreased root growth under ozone stress might increase the impact of droughts. These climatic interactions may also be indirect. For example, warmer winters may be associated with earlier bud burst or crop emergence, leading to greater exposure to elevated hemispheric background ozone concentrations at a sensitive early stage of plant development.
3. Ozone exposure can modify the global carbon cycle, with important indirect effects on global warming, since it reduces photosynthesis, growth, and carbon sequestration. Models that have scaled up from experimental studies on young trees suggest that the adverse effects of global ozone concentrations on forest production could significantly reduce the size of the terrestrial carbon sink. One model estimate suggests that the indirect effect of ozone on radiative forcing, caused by reducing the land carbon sink by 143 PgC over this century, may be as large as the direct radiative forcing due to ozone. However, these models have tended to only consider effects on above-ground growth, and need to be extended to include effects of ozone on the full carbon cycle, including carbon sequestered in the soil and the fluxes of CO₂ and other radiatively active gases from the soil.
4. Ozone may act in combination with climate and environmental change, and changes in population and food

consumption, to threaten global food security. Global models suggest that the proportion of the world's crops exposed to ozone levels above threshold exposures for significant effects on yield will increase significantly over the next two decades. Much of this increased area of agricultural production at risk is likely to be in the rapidly developing regions of Asia, Latin America, and Africa. The economic and social implications of widespread loss of yield of staple crops could be very serious in countries with rapidly increasing populations and loss of productive land. Unfortunately, there is little direct evidence of the significance of ozone impacts on agriculture in these areas. There is an urgent need to extend current risk assessment methods for ozone to these regions of the planet, in order to link this issue more effectively with wider policy evaluation.

In conclusion, surface ozone effects on vegetation are likely to increase in severity at a global scale over the first half of this century, although there will be major regional variations. Ozone needs to be considered as an important component of global environmental change that will have major impacts on the planet's vegetation cover, alongside the changing effects of CO₂ concentrations, climate, and other environmental stresses.

See also: Ozone Depletion and Related Topics: Long-Term Ozone Changes; Photochemistry of Ozone; Surface Ozone (Human Health).

Further Reading

- Ashmore, M.R., 2005. Assessing the future global impacts of ozone on vegetation. *Plant Cell Environment* 28, 949–964.
- Emberson, L.D., Ashmore, M.R., Murray, F. (Eds.), 2003. *Air Pollution Impacts on Crops and Forests*. Imperial College Press, London.
- Feng, Z., Kobayashi, K., 2009. Assessing the impact of current and future concentrations of surface ozone on crop yield with meta-analysis. *Atmospheric Environment* 43, 1510–1519.
- Fuhrer, J., 2009. Ozone risk for crops and pastures in present and future climates. *Naturwissenschaften* 96, 173–194.
- Karnosky, D., Pregitzer, K.S., Zak, D.R., Kubiske, M.E., Hendrey, G.R., Weinstein, D., Nosal, M., Percy, K., 2005. Scaling ozone responses of forest trees to the ecosystem level in a changing climate. *Plant Cell Environment* 28, 965–981.
- Matyssek, R., Karnosky, D.F., Wieser, G., Percy, K., Oksanen, E., Grams, T.E.E., Kubiske, M., Danke, D., Pretzsch, H., 2010. Advances in understanding ozone impact on forest trees: messages from novel phytotron and free-air fumigation experiments. *Environmental Pollution* 158, 1990–2006.
- Mills, G., Hayes, F., Simpson, D., Emberson, L., Norris, D., Harmens, H., Buker, P., 2011. Evidence of widespread effects of ozone on crops and semi-natural vegetation in Europe (1990–2006) in relation to AOT40 and flux-based risk maps. *Global Change Biology* 17, 592–613.
- Royal Society, 2008. *Ground-level Ozone in the 21st Century: Future Trends, Impacts and Policy Considerations*. Science Policy Report 15/08. The Royal Society, London.
- Sitch, S., Cox, P.M., Collins, W.J., Huntingford, C., 2007. Indirect radiative forcing of climate change through ozone effects on the land-carbon sink. *Nature* 448, 791–794.
- van Dingenen, R., Dentener, F.J., Raes, F., Krol, M.C., Emberson, L., Cofala, J., 2009. The global impact on crop yields under current and future air quality legislation. *Atmospheric Environment* 43, 604–618.
- Wittig, V.E., Ainsworth, E.A., Naidu, S.L., Karnosky, D.F., Long, S.P., 2009. Quantifying the impact of current and future tropospheric ozone on tree biomass, growth and physiology and biochemistry: a quantitative meta-analysis. *Global Change Biology* 15, 396–424.

Surface Ozone (Human Health)

M Lippmann, New York University, Tuxedo, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

Ozone (O_3) is formed in ambient air at the earth's surface by photochemical reactions during daylight hours, and is inhaled by people engaged in normal outdoor activities, with much greater dosages during strenuous physical activity and their associated increases in lung ventilation. The dosage indoors is much lower due to O_3 's poor penetration into indoor spaces, its rapid reaction with indoor surfaces, and minimal lung ventilation during sedentary activities. The inhaled O_3 reacts with lung epithelia from the trachea to the respiratory acinus, and (1) causes transient reductions in pulmonary capacity and (2) increases in pulmonary inflammation and permeability. These effects are highly variable from person-to-person, and are considered to be adverse effects in sensitive subjects. There are also cumulative effects from chronic exposures in humans and laboratory animals, but these effects have not been as well characterized as the acute effects.

Introduction

Ozone (O_3) is almost entirely a secondary air pollutant, formed in the atmosphere through a complex photochemical reaction sequence requiring reactive hydrocarbons (HCs), nitrogen dioxide (NO_2), and sunlight. O_3 can only be controlled by reducing ambient air concentrations of HC, NO_2 , or both. Motor vehicles are one of the major categories of sources of HC, NO_2 , and nitric oxide (NO). NO and NO_2 combined, referred to as NO_x , along with HC, have been the target of control efforts and major reductions (>90%) have been achieved in HC emissions per vehicle. Substantial reductions in NO_x emissions from motor vehicles and power plants are now being implemented.

We know a great deal about O_3 chemistry and have developed highly sophisticated O_3 air quality models. Unfortunately, the models, and their applications in control strategies, have clearly been inadequate in terms of meeting national and local air quality goals. We also know a great deal about some of the health effects of O_3 . However, much of what we know relates to transient, apparently reversible effects that follow acute exposures lasting from 5 min to 8 h. These effects include changes in lung capacity, flow resistance, epithelial permeability, and reactivity to bronchoactive challenges; such effects can be observed within the first few hours after the start of the exposure and may persist for many hours or days after the exposure ceases. Repetitive daily exposures over several days or weeks can exacerbate and prolong some of these transient effects, while ameliorating others.

Decrements in respiratory functions such as forced vital capacity (FVC), as measured by the air volume that can be expelled from fully inflated lungs with a maximal effort, and forced expiratory volume in the first second of a vital capacity maneuver (FEV_1) fall into the category where adversity begins at some specific level of pollutant-associated reduction. However, there are clear differences of opinion on what the threshold of adversity ought to be, especially for the more sensitive members of the population.

For O_3 , the only significant exposure route is inhalation, and exposure can be defined as the concentration at the nose and mouth. Local outdoor concentrations are reduced in the vicinity of heavy vehicular traffic, owing to scavenging of O_3 by

NO. On the other hand, less trafficked areas downwind may have higher O_3 concentrations because of the enrichment of the air mass with motor vehicle exhaust, precursor chemicals, and because of active photochemistry. Thus, outdoor O_3 concentrations can be either higher or lower than those measured at central monitoring sites. Indoor concentrations of O_3 are almost always substantially lower than those outdoors because of efficient scavenging by indoor surfaces and the lack of indoor sources.

Exposures are the only one determinant of O_3 dose, which are also determined by the volume of air inhaled and the pattern of uptake of O_3 molecules along the respiratory tract. When people work or exercise outdoors, and thereby increase their rate of ventilation, the contribution of outdoor exposure to total O_3 dose generally becomes the dominant factor. The dose to target tissues in the respiratory acini (the region from the terminal bronchioles through the alveolar ducts) increases even more with exercise than does total respiratory tract dose, since O_3 penetration to the acini increases with tidal volume and flow rate.

Types of Studies

Controlled Human Exposures

There is a very large database from studies in which selected human volunteers were exposed to O_3 in purified laboratory air for specific times. In most studies, a series of exposures took place in random order, and included exposures at one or more concentrations, as well as a sham exposure to purified air. Many involved prescribed periods and intensities of exercise during the exposure interval. The most commonly measured effects were changes in forced expiratory flow rates and volumes and/or changes in airway resistance and compliance.

The advantages of controlled human exposure studies are (1) the opportunity to carefully select and carefully characterize the subjects, whether they be healthy normals, asthmatics, smokers, etc.; (2) the willingness and ability of most volunteer subjects to perform various levels and durations of exercise during the exposures; (3) the ability to deliver and monitor the preselected challenge atmospheres during the exposure; (4) the ability of the subjects to reproducibly perform respiratory

maneuvers required for some functional assays and to provide information on symptomatic responses; and (5) avoidance of the need to make interspecies extrapolations.

The major limitations of controlled human exposures are that (1) ethical constraints limit the challenges and effects of assays that can be performed (i.e., challenges that produce only transient functional changes); (2) the numbers of repetitive challenges and assays are limited by subject tolerance and cooperation; and (3) the number of subjects that can be studied is limited by the generally large costs of performing the studies and/or by the availability of sufficient numbers of subjects with the desired characteristics.

Natural Human Exposures

There is also a human database from studies of the responses of natural populations to exposures to ambient air pollutants. Among the more difficult challenges of such studies some are (1) identifying an accessible population at risk whose relevant exposures can be defined and adequately characterized; (2) collecting an adequate amount of suitable quality-assured data on exposure and responses; and (3) collecting sufficient data on identifiable characteristics of the individuals and of their environmental exposures to other agents that may influence the response variables and thereby confound any of the hypothesized pollutant exposure–response relationships that may be present.

The basic design premise in field studies is to maximize the signal-to-noise ratio for the pollutant exposure–response relationships. The noise on the response side of the relationships has been the focus of much work, and guidance on these aspects is available from the American Thoracic Society. There has been less focus on the reduction of the noise in the exposure variables. An opportunity was created by the recognition that the summer haze is regional in scale and is enriched in secondary air pollutants such as O_3 and sulfuric acid (H_2SO_4), both of which form gradually during daylight hours in air masses containing diluted primary pollutants transported over long distances from industrial, power plant, and motor vehicle sources, especially SO_2 , NO_2 , and HC. Therefore, studies of populations were undertaken in communities remote from local sources of air pollution, primarily in wooded regions where ammonia (NH_3) sources, which would rapidly neutralize the H_2SO_4 , would also be minimal.

In studies of acute responses to the inhalation of secondary air pollutants, it is not possible to identify a nonexposed (control) population. On the other hand, the concentrations of the secondary pollutants have large temporal variations. There are both diurnal variations, with peak concentrations generally occurring in the afternoon, and day-to-day variations in concentration associated with the trajectory of the air mass over pollutant-precursor source areas, atmospheric stability, intensity of incident solar radiation, temperature, and humidity. Thus, the volunteer subjects are exposed to different concentration profiles each day.

Population-Based Studies of Chronic Health Effects

Neither controlled human exposure studies in the laboratory nor natural human exposure studies in the field can provide any direct information on chronic effects of prolonged human

exposures to O_3 . The only way to get such information is to compare group-average measures of function, symptom frequency, lost activity days, hospital admissions, clinic visits, medical diagnoses, etc. with estimates of area-wide chronic exposure intensity.

Because of the large number of possible confounding factors and the difficulty of properly classifying exposures, very large populations must be studied in order to find significant associations between group-average exposures and effects. Any statistically significant effects that are attributed to O_3 would tend to be underestimated because of the influence of the confounding factors.

Alternatively, they could be spurious if the effects are really caused by variables that are colinear with O_3 .

Controlled Exposures of Laboratory Animals

The most convenient and efficient way to study mechanisms and patterns of response to inhaled O_3 , and of the influence of other pollutants and stresses on these responses is by controlled exposures of laboratory animals. The transient functional responses to acute exposures can be measured and the differences in response among different animal species and between them and humans similarly exposed can be determined. Responses that require highly invasive procedures or serial sacrifice can also be obtained to gain information that cannot be obtained from studies on human volunteers. Finally, long-term exposures in animals can produce cumulative responses that determine the pathogenesis of chronic disease. Other advantages of studies on animals are the ability to examine the presence of, and basis for, variations in response that are related to age, sex, species, strain, genetic markers, nutrition, the presence of other pollutants, and so on. As in controlled human exposure studies, the concentrations and durations of the exposure can be tightly controlled, as can the presence or absence of other pollutants and environmental variables. Another important advantage of controlled animal studies is that relatively large numbers of individuals can be exposed simultaneously, creating the possibility of detecting responses that affect only a limited fraction of the population.

Among the significant limitations to the use of exposure-response data from animal studies, human risk assessments are quite limited ability to interpret the animal responses in relation to likely responses in humans who might be exposed to the same or lower levels. Controlled chronic exposure protocols can be very labor intensive and expensive, which tend to limit the number of variables that can effectively be examined in any given study.

Effects of Single Exposures to Ozone

Respiratory Mechanical Function Responses

There are more data on respiratory function responses than on any of the other coincident responses to short-term O_3 inhalation. It is well established that the inhalation of O_3 causes concentration-dependent mean decrements in exhaled volumes and flow rates during forced expiratory maneuvers, and that the mean decrements increase with increasing depth of

breathing and the duration of the exposure. There is a wide range of reproducible responsiveness among healthy subjects and functional responsiveness to O_3 is no greater, and usually lower, among cigarette smokers, older adults, and patients with chronic obstructive pulmonary disease.

Healthy children at summer camps with active outdoor recreation programs had greater decrements in lung function than children exposed to O_3 at comparable concentrations in chambers for 1 or 2 h. Furthermore, their activity levels were considerably lower than those of the children exposed in the chamber studies while performing very vigorous exercise. Since it is well established that functional responses to O_3 increase with levels of physical activity and ventilation, the greater responses in the camp children had to be caused by other factors, such as greater cumulative exposure, or by the potentiation of the response to O_3 by other pollutants in the ambient air.

A study of children with moderate to severe asthma was performed at a summer camp in New England, and the decrements in peak expiratory flow rates associated with ambient O_3 concentrations were similar in magnitude to those reported for healthy children at other summer camps in the Northeastern United States. However, the level of physical activity of the asthmatic children, and hence their O_3 intake, was much lower. Also, the asthmatic children had less reserve functional capacity. Thus, the level of health concern for such comparable functional decrements is much greater.

Field studies have also been done of functional responses of adults engaged in relatively brief (0.5 h) recreational activities outdoors in the presence of varying levels of O_3 . The magnitudes of the functional decrements per unit of ambient O_3 concentration were similar to those observed in volunteers exposed while exercising vigorously for 1 or 2 h in controlled chamber exposure studies.

The observations from the field studies in the children's camps stimulated the Environmental Protection Agency (EPA) Clinical Studies Laboratory in Chapel Hill, NC, USA, to undertake a chamber exposure study of adult male volunteers involving 6.6 h of O_3 exposure. Moderate exercise was performed for 50 min per hour for 3 h in the morning and again in the afternoon. The functional decrements become progressively greater after each hour of exposure. The effects were transient in that there were no residual functional decrements on the following day. The decrements in FEV_1 after 6.6 h of exposure to O_3 concentrations of 120 parts per billion (ppb) averaged 13.6% and were comparable to those seen previously in the same laboratory on similar subjects following 2 h of intermittent heavier exercise at 220 ppb. Follow-up studies were done on adult males with 6.6 h exposures at 80 and 100 ppb, as well as at 120 ppb. Those at 80 and 100 ppb showed lesser changes that also became progressively greater after each hour of exposure.

Evidence of the time scale for the biological integration of O_3 exposure can also be deduced from the rate at which the effects dissipate. In one study, young adult females were exposed in a protocol that produced a mean decrement in FEV_1 of 21%. After 18 h, their mean decrement was down to 4%, and at 42 h it was 2%.

The time scale for the biological integration of the effects of a single O_3 exposure has also been examined in studies on laboratory animals. Rats can provide a good test model for the observed human responses to O_3 , even though they are a less

sensitive species than humans. The lesser responses are consistent with the lesser retention of inhaled O_3 by rats.

The large interindividual variability of O_3 -induced functional responses is not yet understood, and functional responses in individuals do not correlate well with other responses. Based on the large database accumulated over a 10-year period at the EPA Clinical Studies Laboratory, O_3 concentration explained 31% of the variance in FEV_1 responses, and subject age explained only another 4%. For individuals exposed two or more times for 6.6 h, 47% of those exposed to 120 ppb had an FEV_1 decrement of 10% or greater. Respiratory function effects can accumulate over many hours, and an appropriate averaging time for transient functional decrements caused by O_3 is 6 h. This was a major factor for the change in the averaging time for the primary O_3 standard in the United States from 1 to 8 h. Another factor was the recognition that O_3 exposures in ambient air can have broad peaks with 8-h averages equal to 90% of the peak 1-h averages. While there is a great deal of knowledge about O_3 exposure-respiratory function response in humans, as summarized above, we still know very little about the mechanisms responsible for the responses.

Effects on Aerosol Dispersion

As discussed above, acute exposures of humans to low levels of O_3 cause decreases in lung function. However, there is also a potential for small airways responses that do not produce measurable changes in conventional respiratory function indices. In tests of aerosol dispersion, an aerosol bolus is injected into the tidal breath at three different depths, and the primary measure of bolus dispersion is the expired bolus half-width (HW). The HW increased significantly following O_3 exposure relative to air exposure accounted for less than 25% of the variance. Changes in aerosol dispersion seen with O_3 exposure were related to changes in turbulent mixing and/or regional time constants in the small airways, suggesting a possible O_3 effect in that region of the lung as well as effects in the larger airways that produce the respiratory function decrements.

Symptomatic Responses

Respiratory symptoms have been closely associated with group mean pulmonary function changes in adults acutely exposed in controlled exposures to O_3 and in ambient air containing O_3 as the predominant pollutant. In controlled 2-h O_3 exposures, some heavily exercising adult subjects experienced cough, shortness of breath, and pain on deep inspiration at 120 ppb O_3 , although the group mean response was statistically significant only for cough. Above 120 ppb O_3 , respiratory and nonrespiratory symptoms have included throat dryness, chest tightness, substernal pain, cough, wheeze, pain on deep inspiration, shortness of breath, dyspnea, lassitude, malaise, headache, and nausea. The prolonged exposure studies involving 6.6 h of exposure at concentrations between 80 and 120 ppb also produced significant increases in respiratory symptoms including cough and pain on deep inspiration. Although O_3 causes symptomatic responses in adults at current peak levels, there is a large set of data indicating that such responses do not occur in healthy children.

These laboratory results are consistent with the results obtained in field studies of healthy children at summer camps, which failed to find any symptomatic responses despite the occurrence of relatively large decrements in function that were proportional to the ambient O₃ concentrations.

Epidemiological studies have provided evidence of qualitative associations between ambient oxidant levels <100 ppb and symptoms in children and young adults, such as throat irritation, chest discomfort, cough, and headache. Symptoms reported in individuals exposed to O₃ in purified air are similar to those found for ambient air exposures except for eye irritation, a common symptom associated with exposure to the ambient mixtures of photochemical oxidants, which has not been reported for controlled exposures to O₃ alone.

Morbidity

Associations between ambient air pollutants and respiratory morbidity were examined using the Health Interview Survey, a large cross-sectional database collected in the United States by the National Center for Health Statistics. The results indicate an association between fine particulate pollutants and both minor restrictions in activity and respiratory conditions severe enough to result in work loss and bed disability in adults. Ozone, on the other hand, was associated only with the more minor restrictions.

Studies examining associations between the daily average concentrations of ambient air pollutants and daily hospital admissions for respiratory disease have found consistent associations in summer between hospital admissions for respiratory disease and daily levels of sulfate, O₃, and temperature, but no comparable associations for a group of non-respiratory conditions. The effect of ambient O₃ on hospital admissions for respiratory causes ranges from one to three total respiratory admissions per day per 100 ppb O₃ per 10⁶ persons, or from a relative risk per 100 ppb O₃ of 1.1–1.36.

Another index of respiratory morbidity in asthmatics is physician-authorized use of medication. In the previously cited study of children with moderate to severe asthma, the camp physician authorized supplemental asthma medication to children in the group at a rate proportional to the ambient O₃ concentration.

Effects on Athletic Performance

It has been more than four decades since epidemiological evidence from Southern California suggested that the percentage of high school track team members failing to improve performance increased with increasing oxidant concentrations in the hour before a race. The effects may have been related to increased airway resistance or to associated discomfort, which may have limited motivation to run at maximal levels. Subsequent controlled exposure studies of heavily exercising competitive runners have demonstrated decreased function at 200–300 ppb.

There are animal models for decreased performance during O₃ exposure. Rats and mice were exposed for 6 h to O₃ at 80, 120, 250, or 500 ppb while housed in running wheels. Running in both species decreased in a concentration-related manner during exposure to O₃, with the decrease being greater with

increasing time of exposure. The decrease in running activity produced by O₃ persisted for several hours after exposure.

Effects on Airway Reactivity

Exposure to O₃ can also alter the responsiveness of the airways to other bronchoconstrictive challenges as measured by changes in respiratory mechanics. Tests involving exposures of healthy subjects to 80, 100, and 120 ppb for 6.6 h produced, respectively, 56, 89, and 121% increases in methacholine responsiveness. Increased responsiveness to methacholine was also seen with 1 h at 350 ppb in young adult males with allergic rhinitis after exposure to O₃. In subjects with asthma, changes in FEV₁, methacholine responsiveness, and allergen responsiveness did not correlate with each other. The O₃-induced responsiveness does not appear to be dependent on the presence of inflammatory cells in the small airway wall.

Effects on Airway Permeability

The effects of inhaled O₃ on respiratory epithelial permeability were studied in healthy, nonsmoking young men exposed for 2 h to purified air and 400 ppb O₃ while performing intermittent treadmill exercise at 67 l min⁻¹. Seventy-five minutes after the exposures, the pulmonary clearance of a radioisotope-labeled organic molecule [^{99m}Tc]DTPA, was measured as an index of epithelial permeability. O₃ exposure sufficient to produce decrements in the respiratory function of human subjects also caused an increase in permeability. An increased permeability could facilitate the uptake of other inhaled toxicants and/or the release of inflammatory cells such as neutrophils onto the airway surfaces.

Effects on Airway Inflammation

O₃-induced airway reactivity to methacholine is also associated with neutrophil influx into the airways and with changes in cyclooxygenase metabolites of arachidonic acid. Inflammatory and biochemical changes in the airways also occur following O₃ exposure. The inflammatory process caused by O₃ exposure is promptly initiated and persists for at least 18 h. Cells and enzymes capable of causing damage to pulmonary tissues were increased, and the proteins that play a role in the fibrotic and fibrinolytic processes were elevated as a result of O₃ exposure.

The weight of the evidence from these results, showing both functional and biochemical responses in humans and laboratory animals that accumulate over multiple hours and persist for many hours or days after exposure ceases, is clear and compelling. Both functional changes and inflammatory processes were shown to occur in humans following exposures to 100 ppb O₃ for 6.6 h, whereas higher concentrations were required to elicit comparable responses in rats. Thus, the rat data, which provide evidence of other effects as well, appear to provide conservative indications of effects in humans.

Effect of Single and Multiday Exposures on Particle Clearance from the Lungs

The effect of 2-h exposures to 200 or 400 ppb O₃ with intermittent light exercise on the rates of tracheobronchial

mucociliary particle clearance was studied in healthy adult males. The 200 ppb O₃ exposure produced a significant acceleration of particle clearance in peripheral airways but failed to produce a significant reduction in FVC, suggesting that significant changes in the ability of the deep lung to clear deposited particles take place before significant changes in respiratory function take place.

Effects of Single and Multiday Exposures on Lung Infectivity

Both *in vivo* and *in vitro* studies in animals have demonstrated that O₃ can affect the ability of the immune system to defend against infection. Increased susceptibility to bacterial infection has been reported in mice at 80–100 ppb O₃ for a single 3-h exposure. Related alterations of the pulmonary defenses caused by short-term exposures to O₃ include impaired ability to inactivate bacteria in rabbits and mice and impaired macrophage phagocytic activity, mobility, fragility and membrane alterations, and reduced lysosomal enzymatic activity. Some of these effects have been shown to occur in a variety of species including mice, rats, rabbits, guinea pigs, dogs, sheep, and monkeys.

Effects of Multiday and Ambient Episode Exposures

Since single exposures lasting for an hour or more at current peak ambient O₃ levels produce measurable biological responses in healthy humans, and since there is a high probability that one high-O₃ day will be followed by several more high-O₃ days, it is important to know the extent to which the effects accumulate or progress over multiple days. The data on functional adaptation are largely, but not exclusively, based on studies in human volunteers, whereas the database on biochemical and structural changes caused by O₃ is based entirely on studies in laboratory animals.

It is well established that repetitive daily exposures, at a level that produces a functional response upon single exposure, result in an enhanced response on the second day, with diminishing responses on third and fourth day and virtually no response by fifth day.

For repeated 6.6 h per day exposures to 120 ppb O₃, the peak functional response occurs on the first day, with progressively lesser responses after the second, third, and fourth days of exposure. However, for these same subjects, their responsiveness to methacholine challenge peaked on the second day, and remained elevated throughout all 5 days of exposure. The persistent elevation of airway responsiveness is one reason to discount the view held by some that the functional adaptation phenomenon indicates that transient functional decrements are not an important health effect. Additional evidence comes from research in animals showing that persistent damage to lung cells accumulates even as functional adaptation takes place.

Chronic Effects of Ambient Ozone Exposures

The chronic effects database includes a quite limited amount of information on human effects and a more substantial volume

of data on effects seen in laboratory animals undergoing chronic exposures.

Controlled Laboratory Exposure Studies: Human Responses

A study in Southern California provided evidence for a seasonal adaptation of lung function. In this study, a group of subjects selected for their relatively high functional responsiveness to O₃ had much greater functional decrements following 2 h of exposure to O₃ at 180 ppb with intermittent exercise in a chamber in the spring than they did in the following fall or winter, although their responses in the following spring were equivalent to those in the preceding spring. These findings suggest that some of the variability in acute response coefficients reported for earlier controlled human exposures to O₃ in chambers could have been related to seasonal variations in responsiveness, which, in turn, may be related to a long-term adaptation to chronic O₃ exposure.

Epidemiological Studies

Reports of populations of children living in Southern California indicate that chronic air pollutant exposures can affect lung development as measured by growth in respiratory function. However, O₃ was not among the pollutants that correlated with lung development. Rather, the effect was associated with particulate matter, NO₂, and acidic vapors.

There was some limited evidence for chronic effects of O₃ based upon an analysis of pulmonary function data in a national population study in 1976–80 – the second National Health and Nutrition Examination Survey. Using ambient O₃ data from nearby monitoring sites, a highly significant O₃-associated reduction in lung function was seen for people living in areas where the annual average O₃ concentrations exceeded 40 ppb.

A prospective study of a cohort of Seventh Day Adventists living in California reported an O₃-associated increase in the number of men developing adult-onset asthma, but not of women. It was suggested that the effect in men might have been due to their greater time spent out of doors and engaged in strenuous activities that could have resulted in more O₃ inhalation.

The limited evidence available to date for the effect of O₃ on longevity is also largely negative. A study utilizing the Public Use Sample containing data on 2 million individuals in the United States obtained both death certificate data and air pollution network data in eight regions of the United States. Highly significant and consistent associations with mortality were found for SO₄²⁻. Significant, but weaker and less consistent associations were seen for SO₂ and CO. No significant associations were seen for O₃ or NO₂.

Controlled Laboratory Exposure Studies: Animal Responses

The highest O₃ dose is received at the acinus, where the terminal bronchioles lead into alveolar ducts, and a series of studies have shown that the effects of inhaled O₃ on lung structure are also greatest in this region. Morphometric techniques selectively focused on this limited region of the lung showed that significant changes occurred in the alveoli

just distal to the terminal bronchioles in rats exposed for 12 h per day for 6 or 12 weeks to 120 or 240 ppb O₃. In both juvenile and adult rats, there were significant increases in the numbers of alveolar type I and type II epithelial cells and alterations in the interstitium and endothelium. From physiological studies of rats that were simultaneously exposed, there were significant increases in the vital capacity and end-expiratory volume that suggested alterations in distensibility of the lung tissue.

Studies at relatively low O₃ concentrations have also been done in monkeys exposed to O₃ for 8 h per day for 6 or 90 days to 150 or 300 ppb O₃. Responses included ciliated cell necrosis, shortened cilia, and secretory cell hyperplasia with less stored glycoconjugates in the nasal region. Respiratory bronchiolitis observed at 6 days persisted to 90 days of exposure. Even at the lower concentration of 150 ppb O₃, nonciliated bronchiolar cells appeared hypertrophied and increased in abundance in respiratory bronchioles.

For some chronic effects, intermittent exposures can produce greater effects than those produced by a continuous exposure regime that results in higher cumulative exposures. Two groups of 7-month-old male monkeys were exposed to 250 ppb for 8 h per day either daily or, in the seasonal model, daily in alternate months during a total exposure period of 18 months. A control group breathed only filtered air. Monkeys from the seasonal exposure model, but not those exposed daily, had significantly increased total lung collagen content, chest wall compliance, and inspiratory capacity. All monkeys exposed to O₃ had respiratory bronchiolitis with significant increases in related morphometric parameters. Lung growth was not completely normal in either exposed group. Thus, long-term effects of oxidant air pollutants that have a seasonal occurrence may be more dependent on the sequence of polluted and clean air than on the total number of days of pollution, and estimations of the risks of human exposure to seasonal air pollutants from effects observed in animals exposed daily may underestimate long-term pulmonary damage.

Summary and Conclusions

The apparently reversible effects that have followed acute exposures lasting from 5 min to 6.6 h include changes in lung capacity, flow resistance, epithelial permeability, and reactivity to bronchoactive challenges. These effects may persist for many hours or days after the exposure ceases. Repetitive daily exposures over several days or weeks can exacerbate and prolong these effects.

Most of the data available on transient functional effects of O₃ were obtained from controlled human exposure studies and field studies of limited duration. Such studies can provide information on chronic pollutant effects only to the extent that prior exposures affect the transient response to single-exposure challenges. Furthermore, interpretation of the results of such tests is limited by our generally inadequate ability to characterize the nature and/or magnitude of the prior chronic exposures. Most of the limited data we have on the effects of chronic O₃ exposures on humans come from epidemiological studies.

Epidemiological studies offer the prospect of establishing chronic health effects of long-term O₃ exposure in relevant populations and offer the possibility that the analyses can show the influence of other environmental factors on responses to O₃ exposure. On the other hand, the strengths of any of the associations may be difficult to establish firmly because of the complications introduced by uncontrolled cofactors that may confound or obscure the underlying causal factors.

In terms of functional effects, we know that single O₃ exposures to healthy nonsmoking young adults at concentrations in the range 80–200 ppb produce a complex array of pulmonary responses including decreases in respiratory function and athletic performance, and increases in symptoms, airway reactivity, neutrophil content in lung lavage, and rate of mucociliary particle clearance. Responses to O₃ in purified air in chambers occur at concentrations of 80 or 100 ppb when the exposures involve moderate exercise over 6 h or more and require concentrations of 180 or 200 ppb when the duration of exposure is 2 h or less. On the other hand, 45% mean FEV₁ decrements have been seen at 100 ppb of O₃ in ambient air for children at summer camps and for adults engaged in outdoor exercise for only 0.5 h. The apparently greater responses to O₃ in ambient air may be related to the presence of, or prior exposures to, acidic aerosols, but further investigation of this hypothesis is needed.

Exposure over successive days of adult humans in chambers to O₃ at current high ambient levels leads to a functional adaptation in that the responses are attenuated by the third day, and are negligible by the fifth day. On the other hand, a comparable functional adaptation in rats does not prevent the progressive damage to the lung epithelium. Daily exposures of animals also increase other responses in comparison to single exposures, such as a loss of cilia, a hypertrophic response of Clara cells, alterations in macrophage function, and alterations in the rates of particle clearance from the lungs.

The plausibility of accelerated aging of the human lung from chronic O₃ exposure is greatly enhanced by the results of a series of chronic animal exposure studies in rats and monkeys. There is little reason to expect humans to be less sensitive than rats or monkeys. On the contrary, humans have a greater dosage delivered to the respiratory acinus than do rats for the same exposures. Another factor is that the rat and monkey exposures were administered to confined animals with little opportunity for heavy exercise. Thus, humans who are active outdoors during the warmer months may have greater effective O₃ exposures than the test animals. Finally, humans are exposed to O₃ in ambient mixtures. The potentiation of the characteristic O₃ responses by other ambient air constituents seen in the short-term exposure studies in humans and animals may also contribute toward the accumulation of chronic lung damage from long-term exposures to ambient air containing O₃.

The lack of a more definitive database on the chronic effects of ambient O₃ exposures in humans is a serious failing that must be addressed with a long-term research program. The potential impacts of such exposures on public health deserve serious scrutiny and, if they turn out to be substantial, strong corrective action. Further controls on ambient O₃ exposure will be extraordinarily expensive and will need to be very well justified.

Acknowledgments

This article is a greatly condensed version of a comprehensive review appearing as the second entry in the Further Reading list that follows. Readers can consult that review for more detail and the citation of the supporting references.

See also: **Aviation Meteorology:** Aircraft Emissions. **Ozone Depletion and Related Topics:** Photochemistry of Ozone; Surface Ozone Effects on Vegetation.

Further Reading

- Ashmore, M.R. University of Bradford, Bradford, West Yorkshire, UK.
Lippmann, M., 2009. Ozone. In: Lippmann, M. (Ed.), *Environmental Toxicants: Human Exposures and Their Health Effects*, third ed. Wiley, New York.
Lippmann, M., Schlesinger, R.B., 2000. Toxicological bases for the setting of the health-related air pollution standards. *Annual Reviews in Public Health* 21, 309–333.

Contents

Ice Cores

Varves

Ice Cores

EJ Steig, University of Washington, Seattle, WA, USA

© 2003 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 4, pp 1673–1680, © 2003, Elsevier Ltd.

Introduction

Ice cores drilled in glaciers and ice sheets are used to investigate atmospheric chemistry, climate and glacier dynamics in the past. Precipitation that falls as snow in the polar regions, and at high altitudes in the tropical and temperate regions, if it survives summer warmth, is eventually transformed to glacier ice. Aerosols and soluble gases that are scavenged from the atmosphere by precipitation (wet deposition), or accumulate on the snow surface as dry deposition, are preserved within the ice. Atmospheric gases are also preserved in the form of air bubbles trapped within the ice. Ice cores are retrieved by means of a mechanical or thermal drill that cuts an annulus around a central, vertical core that is typically 10–20 cm in diameter and a few tens of meters to several thousand meters in length. If the relationship between length along the core and time in the past is known, time-series of atmospheric composition can be derived. The best-known example is the record of carbon dioxide from Antarctic ice cores, which documents significant changes in concentration of this important greenhouse gas over the last several hundred thousand years of Earth history. Time-series have also been developed from ice cores for many other trace gases and aerosol species, and, indirectly, of atmospheric temperature, atmospheric dynamics, solar variability, and marine and terrestrial biogeochemical cycles. Because most of this information is unavailable from any other source, ice cores play a central role in our understanding of paleoclimate. Data from ice cores also serve as a baseline of natural variability against which contemporary changes in the atmosphere may be evaluated.

Glacier Ice and Ice Core Dating

Ice core studies are conducted at perennially cold sites where little or no melting occurs. Examples include Antarctica, the interior portions of Greenland, the small ice caps of Arctic

Canada, and the high mountains of the Andes and the Himalaya. Under such conditions, there is net snow accumulation in most years, and a combination of mass loading and metamorphism leads to compaction of the snow and eventual transition to glacier ice. Buildup of ice masses reaches from hundreds of meters thick in alpine settings to several kilometers in the great polar ice sheets of Antarctica and Greenland. This buildup is balanced by the slow deformation of glacier ice under pressure. Flow by internal deformation and sliding along the bed brings ice into warmer, lower-elevation environments where it is lost to melting, sublimation, or iceberg calving. Because of compaction and flow, snow layers of original thickness λ at the surface become thinner with depth. The amount of time t for a layer to reach depth z cannot generally be determined directly, but can be estimated with a geophysical ice flow model, used to calculate a thinning function $\psi = \lambda(z, 0)/\lambda(0, -t)$, if the net snow accumulation rate $\dot{b}(-t)$ is known. Ice core drilling sites are usually chosen near an ice divide (analogous to a drainage divide in hydrology), where the dominant strain direction is vertical. This greatly simplifies the calculation of ψ , which in this case will be approximately exponential. Under favorable circumstances, age and layer thickness can be determined by direct visual observation, using the alternating layering of coarse-grained, dusty, and wind-packed summer snow and relatively uniform winter snow. Counting of visible annual layers and seasonal cycles in chemistry, coupled with the identification of volcanic eruptions of known age (using ash chemistry and sulfate concentration anomalies preserved in the ice), allowed researchers to determine the age–depth relationship in cores from Summit, Greenland, to a precision better than 2% for the last 10 000 years (10 ky) and 5% for the last 50 ky. Precise determination of layer thickness and age also permitted the calculation of an accurate time series of accumulation (Figure 1).

Typical snow densities at the surface of an ice sheet are $\sim 300 \text{ kg m}^{-3}$, increasing to $\sim 900 \text{ kg m}^{-3}$ a few tens of meters below the surface. The zone of intermediate density, referred to

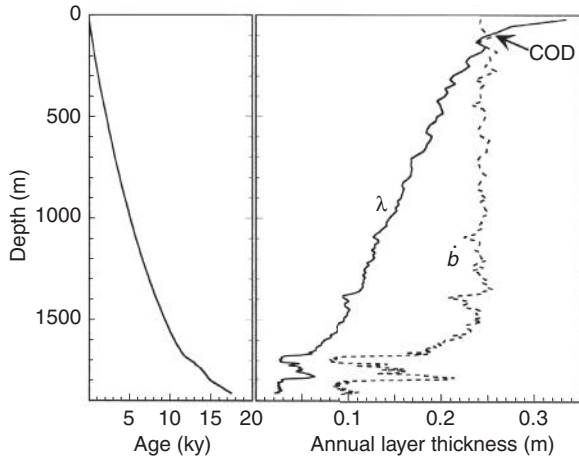


Figure 1 Depth versus age, measured annual layer thickness (λ) and calculated ice-equivalent annual accumulation ($\dot{b}$) in the GISP2 (Greenland Ice Sheet Project 2) ice core from Summit, central Greenland. The air bubble close-off depth (COD) is approximately at the intersection of $\dot{b}$ and λ .

as the firn, is a permeable layer through which atmospheric gases can move. As air channels in the firn become closed off, gases are isolated from the overlying atmosphere and are eventually trapped as air bubbles. Bubble close-off occurs at a density near 820 kg m^{-3} , which generally corresponds to a depth of 60–100 m. At a given depth, the age of the gas trapped within the ice is always younger than the ice itself because of advective and diffusive transport within the firn. The amount of time (eqn [1]) corresponding to the bubble-close-off depth (COD) is a nonlinear function of snow accumulation rate and temperature.

$$\Delta \text{age}(z) = t_{\text{ice}}(z) - t_{\text{gas}}(z) = t_{\text{ice}}(\text{COD}) \quad [1]$$

At high-accumulation sites like central Greenland ($\dot{b} > 200 \text{ kg m}^{-2} \text{ y}^{-1}$), Δage can be a few hundred years or less. At Vostok Station in central East Antarctica ($\dot{b} < 30 \text{ kg m}^{-2} \text{ y}^{-1}$), Δage is several thousand years. Several Antarctic ice core records have been dated indirectly by correlation with the layer-counted Greenland records, using variations in globally well-mixed trace gases as time markers (Figure 2). The major source of error in this cross-dating technique is uncertainty in the value of Δage .

Stable Isotopes and Temperature

Time-series of past temperature change are obtained from ice cores using the ratios of stable isotopes of water. The major features of the global distribution of water isotopes can be represented by a simple Rayleigh distillation approximation of the hydrological cycle that works particularly well at high latitudes. Seasonal variations in temperature (T) are reflected in the $^{18}\text{O}/^{16}\text{O}$ and D/H ratios of atmospheric water vapor and precipitation according to nearly linear relationships (eqns [2] and [3]).

$$\frac{d(\delta^{18}\text{O})}{dT} = \alpha \quad [2]$$

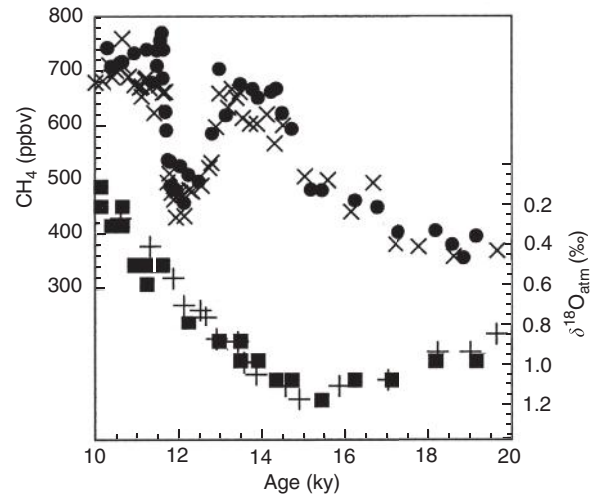


Figure 2 Concentrations of methane (CH_4) and isotopic ratio of molecular oxygen ($\delta^{18}\text{O}$ of O_2) from ice cores at Summit (GISP2), Greenland ($\bullet$, $\blacksquare$) and Taylor Dome, Antarctica ($\times$, $+$) between 10 and 20 ky ago. The oscillation in CH_4 between 13 ky and 11.5 ky is coincident with the Younger Dryas cold oscillation observed in paleoclimate records from around the North Atlantic region. The overall decrease in $\delta^{18}\text{O}$ of atmospheric O_2 (plotted on a reversed scale, by convention) reflects the increase in sea level from the influx of isotopically light (low $\delta^{18}\text{O}$) water from melting Northern Hemisphere ice sheets. Adapted with permission from E J Steig et al. (1998) Science 282: 92–95.

$$\frac{d(\delta D)}{d(\delta^{18}\text{O})} = 8 \quad [3]$$

The average value of α is about 0.7 ‰ K^{-1} . (‰ = parts per thousand; δ refers to the deviation in ‰ from the $^{18}\text{O}/^{16}\text{O}$ or D/H ratio of a standard, usually Standard Mean Ocean Water). Where temperatures are warm or snow accumulation rates are low, diffusion of water molecules within the firn layer will erase seasonal cycles after a few years to decades. Annual and inter-annual variations, however, are generally well preserved and provide a means of extending instrumental temperature records back in time. Regional or global climate events such as the Little Ice Age between c. AD 1400–1900, and the transition from the cold temperatures of the last glacial period, about 20 ky ago, to the current, Holocene interglacial period are clearly seen in the isotopic time-series (Figure 3). Also evident in Arctic ice cores are very rapid isotopic changes, known as Dansgaard–Oeschger (D–O) events after their discoverers. The calculation of temperature from isotope ratios is based on the assumption that the slope, α , of the δ/T relationship is constant. Calibration of the isotope paleothermometer has been accomplished using spatial gradients on instrumental time-series, and with paleotemperatures calculated from measured temperatures in liquid-filled boreholes, using geophysical inverse methods. The calibration results show that α varies by about a factor of 2 over glacial–interglacial time scales because of changes in the seasonality of precipitation, moist air mass transport trajectories, and ocean surface conditions. Where such changes can be quantified, $\delta^{18}\text{O}$ and δD may be considered reliable quantitative proxies for past temperature

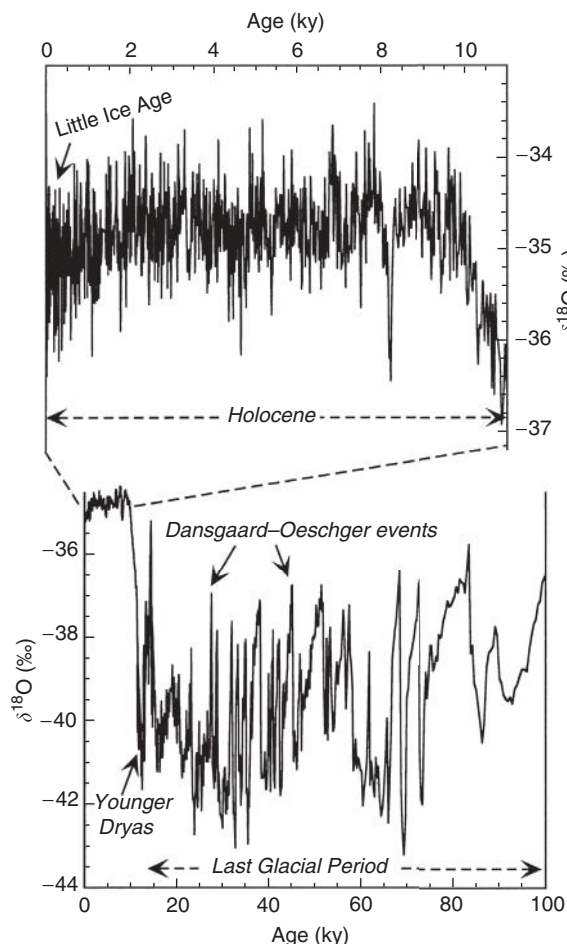


Figure 3 Oxygen isotope ratios of water ($\delta^{18}\text{O}$) from the GISP2 ice core, Greenland, from 100 ky ago to the present. The transition from the highly variable glacial period to the relative stable Holocene interglacial occurs at ~ 11.5 ky. The most recent 11 ky are shown on an expanded scale.

change. Ocean surface conditions, specifically, can be investigated using variations in 'deuterium excess', defined in

$$d = \delta\text{D} - 8 \times \delta^{18}\text{O} \quad [4]$$

Deviations of d from the average meteoric water value of 10‰ reflect the difference in sensitivity to phase changes of δD versus $\delta^{18}\text{O}$. Ice core time series of d are used to identify the ocean source areas for polar and high-altitude water vapor, and to quantify changes in evaporation temperature and humidity at those source areas.

Trace Gases

Ice cores provide an essentially direct record of past concentrations of atmospheric gases. Depending on the magnitude of Δage , each gas sample analyzed typically represents an average of a few hundred years. Measurement of CO_2 from Greenland ice cores has been problematic because of the relatively warm temperature and high carbonate content of the ice. That CO_2

records from Antarctic ice cores are reliable is demonstrated by the excellent agreement with direct atmospheric measurements during the period of overlap (Figure 4). Over long time scales CO_2 is linearly related (correlation coefficient, $r^2 > 0.7$) to stable-isotope records of air temperature (Figure 5). Natural variations over the last 420 ky remained within 180–300 ppmv, well below the early twenty-first century value of > 370 ppmv. Analysis of data from multiple ice and ocean sediment cores indicates a lead of temperature over CO_2 of perhaps a few hundred years for glacial–interglacial time scales. However, the phase relationship appears to be nonstationary and its average value remains poorly known because of large uncertainties in Δage .

Concentrations of other atmospheric trace gases such as methane (CH_4) and nitrous oxide (N_2O) show long-term variations similar to those of CO_2 . Comparisons of records from ice cores in Antarctica and Greenland allows determination of changes in the inter-polar gradients of these gases, which reflects different biogenic production rates in the Northern versus Southern Hemispheres. Methane additionally shows rapid variations that correspond to the fast warming and cooling of the D–O events observed in $\delta^{18}\text{O}$ records. Measurements of $^{15}\text{N}/^{14}\text{N}$, $^{40}\text{Ar}/^{36}\text{Ar}$, and $^{84}\text{Kr}/^{36}\text{Ar}$ ratios reveal the preservation of thermally driven isotope ratio anomalies that permit independent calculation of the magnitude of temperature change and the value of Δage during these events. The results indicate that some of the D–O warmings were extremely rapid ($> 1^\circ\text{C}/\text{y}$) and of even larger magnitude (> 15 K) than would be predicted from $\alpha \sim 0.7\text{‰K}^{-1}$, and show that temperature leads CH_4 by no more than a few decades.

Another important set of data obtained from ice core air bubbles is the $^{18}\text{O}/^{16}\text{O}$ ratio of atmospheric molecular oxygen (O_2), reported as $\delta^{18}\text{O}_{\text{atm}}$ to distinguish it from $\delta^{18}\text{O}$ of ice. The primary control on $\delta^{18}\text{O}_{\text{atm}}$ is the isotopic composition of average surface ocean water ($\delta^{18}\text{O}_{\text{sw}}$), which in turn is determined to first order by global ice volume – the amount of water stored on land in the form of ice sheets and glaciers. Time-series

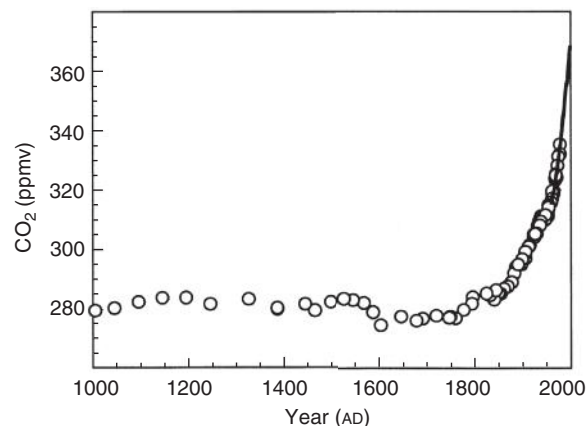


Figure 4 Carbon dioxide (CO_2) concentrations ($\circ$) from air bubbles in an ice core from Law Dome, Antarctica, for the last 1000 years compared with the direct measurements of atmospheric CO_2 (—) at Mauna Loa, Hawaii since AD 1959.

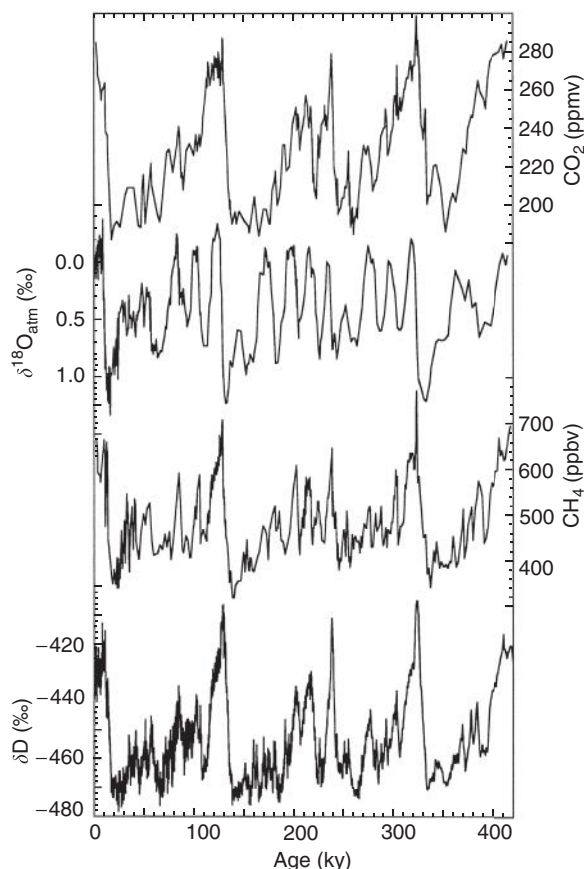


Figure 5 Concentrations of CO₂, δ¹⁸O of O₂, and methane (CH₄) in air bubbles and hydrogen isotopes ratios (δD) in ice from the Vostok ice core, central East Antarctica, for the last 420 ky.

of δ¹⁸O_{atm} from ice cores are used to tie ice core chronologies to each other and to marine sediment records of δ¹⁸O_{sw}. The connection between δ¹⁸O_{atm} and ocean water δ¹⁸O_{sw} is indirect, via photosynthesis and respiration by phytoplankton; there is a consequent variability in the lag between δ¹⁸O_{sw} and δ¹⁸O_{atm}, known as the 'Dole effect', of up to ~2000 years.

Aerosols and Soluble Gas-Phase Species

The major ions present in most ice cores are Na⁺, H⁺, Ca²⁺, Mg²⁺, Cl⁻, NO₃⁻ and SO₄²⁻. Typical concentrations are at the ppb (10⁻⁹) level. In coastal areas of the polar ice sheets, Na⁺, Cl⁻, Mg²⁺, Ca²⁺, and SO₄²⁻ represent > 80% of the ionic budget. Nitrate and non-sea-salt sulfate dominate the budget at more inland sites. All of these species show major variations that parallel those in stable isotopes and trace gas concentrations (Figure 6). Determining whether observed changes in concentration reflect changes in wet and dry deposition processes or real changes in atmospheric mixing ratios can be problematic. For example, a large fraction of the glacial-interglacial change in SO₄²⁻ concentrations probably reflects dilution by the twofold increase in snow accumulation rate at the beginning of the Holocene. The set of transfer functions describing air-to-snow relationships for all chemical

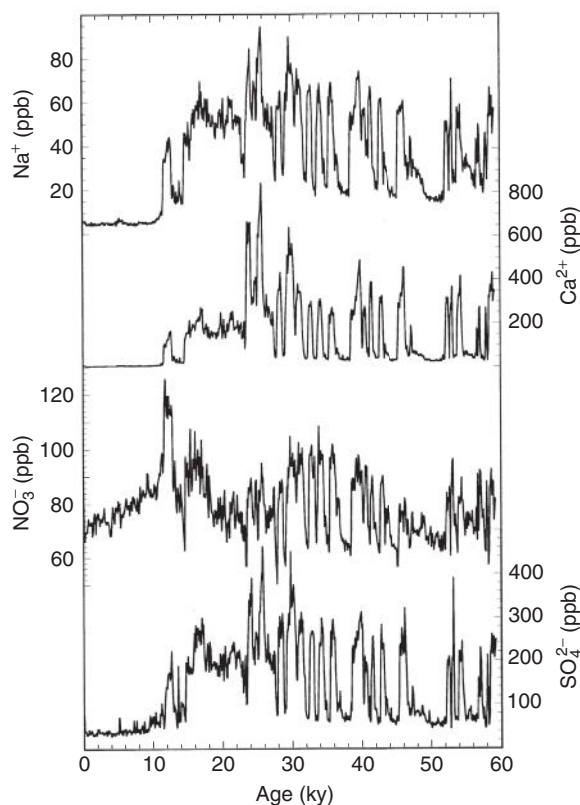


Figure 6 Concentrations of sodium, calcium, nitrate, and sulfate ions in the GISP2 ice core, Greenland, from 60 ky ago to the present.

species can be expressed as a formal mathematical inverse problem. Because of the high level of covariance in most ice core time-series, this problem must currently be considered under-terminated for most species. Nevertheless, significant changes in atmospheric mixing ratios have clearly occurred and are particularly large for terrestrially derived ions such as Ca²⁺.

Increases in terrestrial species reflect increases in the overall dustiness of the atmosphere due to enhanced continental aridity and aeolian transport during cold climate periods. This effect is amplified by increased wind speeds and the consequent greater efficiency of transport through the troposphere to the surfaces of glaciers and ice sheets. Specific source regions for continental dust have been identified through the use of isotopic and rare-earth element analysis of mineral grains preserved in the ice. Increased concentrations of marine-derived species in ice cores reflect increases both in transport efficiency and in biogenic production rates. More than 50% of SO₄²⁻ in inland Antarctic snow is derived from the oxidation of dimethyl sulfide, an algal waste product. The ratio of sulfate to methanesulfonic acid, an intermediate oxidation product that is preserved in the ice, has been used to identify biogenic versus nonbiogenic sources of SO₄²⁻. Changes in the biogenic contribution of SO₄²⁻ are of particular interest because of the role that sulfate aerosols play in climate forcing via modification of cloud radiative properties. Emissions from volcanoes are an important source of SO₄²⁻ concentration anomalies in ice cores; comparison with the

stable-isotope record of temperature provides strong independent evidence for the cooling effect of sulfate aerosols. In central Greenland cores, time-series analysis shows that geochemical variations can be resolved into two nearly orthogonal components that are dominated by marine and terrestrial species, respectively. The ratio of marine to terrestrial components is believed to reflect variations in the relative strength of meridional versus zonal tropospheric flow, possibly related to changes in the strength of the polar vortex.

Species such as hydrogen peroxide (H_2O_2), hydrochloric acid (HCl), and nitric acid (HNO_3), which have an appreciable vapor pressure at ambient surface temperatures, additionally complicate the interpretation of ice core geochemical records. While deposition rates are controlled by kinetic processes in clouds and at the snow surface, loss rates following deposition are dominated by the tendency toward equilibrium with the overlying atmosphere. Final concentrations archived in the ice core are determined by the extent to which equilibrium is reached. The approach to equilibrium depends on a combination of snow accumulation rate, temperature and, for photoreactive species, irradiance levels. Ambient oxidative conditions determined by hydroxyl radical ($\dot{\text{O}}\text{H}$) and ozone (O_3) mixing ratios also play a role. The measurement of multiple reactive and nonreactive species and their isotopic composition in ice cores offers a means for obtaining quantitative information on the oxidizing capacity of the atmosphere in the past.

Cosmogenic Radionuclides and Geomagnetic and Solar Variability

Rare isotopes produced in the atmosphere by cosmic rays, including ^{10}Be , ^{26}Al , ^{36}Cl , and ^{14}C , are all present in ice cores in measurable concentrations. These cosmogenic radionuclides (CRNs) are of interest as a means to extend the record of geomagnetic and solar variability beyond the instrumental and historical time periods and as a source of information on snow accumulation and atmospheric dynamics. Solar modulation of CRN production is reflected in variations in concentration that have dominant periodicities at ~ 11 years and ~ 90 years, and that closely track the record of atmospheric ^{14}C variations, independently derived from tree-ring studies (Figure 7). Over longer time scales, geomagnetic modulation may also be important in affecting CRN deposition rates. Differences among records from Arctic, tropical, and Antarctic ice cores can in principle be used to separate solar and geomagnetic modulation of CRN production, since the latter is unimportant at high latitudes. In practice, the difficulty of separating production variations from meteorologically influenced changes in deposition rate has prevented an unequivocal geomagnetic record from being obtained from CRN measurements. An exception is the large excursion observed in ^{10}Be and ^{36}Cl in several ice cores, which coincides with the Laschamp geomagnetic excursion 40 ky ago, when the field strength may have approached zero. Most other large changes in CRN concentrations reflect changes in snow accumulation rate. Assuming that production rates are approximately constant over time scales greater than those expected from solar variability (i.e. >1000 years), ^{10}Be can be used as an independent

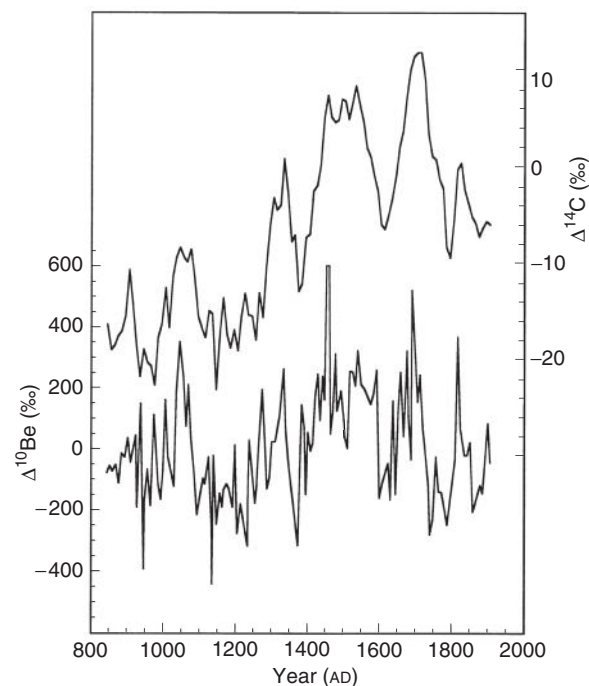


Figure 7 Anomalies of ^{10}Be in an ice core from the South Pole compared with atmospheric ^{14}C anomalies from analyses of tree rings between AD 800 and 1950. The Δ notation refers to relative deviation in parts per thousand. Adapted with permission from E. Bard et al., *Earth and Planetary Science Letters* 150: 453–462.

variable in separating changes in atmospheric loading of other aerosol species from changes in precipitation scavenging and dry deposition efficiency. The assumption of constant production is reasonable in Antarctica, where limited meridional air-mass exchange prevents a significant amount of geomagnetically modulated, low-latitude-produced ^{10}Be from reaching the ice sheet surface. Because significant CRN production occurs in the stratosphere (about two-thirds of the total production), CRN can also provide information on past stratosphere/troposphere exchange processes. This has been accomplished qualitatively, using the magnitude of solar-modulated ^{10}Be variations, and the ratio of ^{36}Cl to ^{10}Be , which have different tropospheric lifetimes, as tracers of air mass origin. Radiocarbon (^{14}C) is a special case of CRN in ice cores because the predominant fraction is produced *in situ* by neutron activation of ^{14}N . Concentrations of ^{14}CO and $^{14}\text{CO}_2$ in air bubbles provide a determination of snow accumulation rate that agrees well with independent evidence.

The Causes of Ice Ages and Rapid Climate Change Events

Geochemical time-series from ice cores have had a profound influence on our understanding of climate change. Particularly important is the opportunity provided by ice cores to evaluate the characteristics of paleoclimate time-series in the frequency domain. The high-precision, independent dating achieved with ice cores has allowed a more rigorous test of Milankovich

theory than was previously possible with ocean sediment cores, and confirms that climate change exhibits significant power at orbital frequencies. The observation from the Vostok ice core that CO_2 varies approximately linearly with temperature addresses a critical problem for Milankovich theory, which is that the amplitude of solar insolation variations resulting from Earth's orbital changes is too small to alone account for the large magnitude of climate changes. At least 50% of the required amplification of small insolation changes over glacial–interglacial cycles can be attributed to observed CO_2 variations alone. The $\delta^{18}\text{O}_{\text{atm}}$ record shows a strong coherence with precession (~ 23 ky period) supporting the proposed linear relationship between orbital forcing and global ice-sheet volume.

Ice core data also demonstrate that large-magnitude climate changes have occurred in the past that are not related to Milankovich forcing. Although there is no evidence for rapid CO_2 changes in the past (variations originally identified in the Greenland Dye 3 ice core are now known to be artifacts), the D–O events clearly reflect real, rapid changes in atmospheric composition and dynamics. The D–O events are evident not only in $\delta^{18}\text{O}$ and CH_4 in Greenland ice cores, but also in dust and aerosol concentrations, and have been identified in ocean sediment cores from the North Atlantic region, from chemical tracers of deep water circulation, and from sea surface temperature. These events recur throughout the last glacial period with a nominal frequency of $1/1500 \text{ y}^{-1}$. There is considerable variation in the length of this cycle: in most data sets the spectral power of D–O events rises significantly above red noise only for bandwidths of several hundred years. On average, every fourth D–O event is associated with a coarse-grained ‘Heinrich’ layer in North Atlantic sediment cores. Heinrich layers reflect the transport of debris by icebergs from the margins of the ice sheets that covered much of Europe and North America during the last glaciation. Many paleoclimatologists believe that comparison of paleoclimate records from both hemispheres is a key to understanding the D–O events and their association with ocean thermohaline circulation and the dynamics of large ice sheets. Quasi-periodic climate changes with comparable timing have been identified in South American and Antarctic ice cores and Southern Ocean sediment cores, but these are generally smaller in amplitude and, with a few exceptions, do not show the rapid warmings characteristic of D–O events.

Natural Climate Variability and the Anthropogenic Impact on the Atmosphere

Ice core records provide an important baseline against which anthropogenic changes to the atmosphere can be measured. The high temporal resolution obtained with ice core geochemical and isotopic measurements has been used to extend records of the North Atlantic Oscillation and other important climate indices beyond the short period available from instrumental records. Arrays of ice cores covering the last 200–2000 years at annual resolution have been used to document the spatial patterns of interannual variability of both Arctic and Antarctic climate, for which instrumental data are particularly limited. On longer time scales, stable-isotope

records show that temperatures during the twentieth century were slightly cooler, on average, than during the peak warm intervals of both the last interglacial period (~ 125 ky ago) and the Holocene. A mid-Holocene warm period is especially pronounced in ice cores from northern Greenland and the Canadian Arctic. The ice cores provide little evidence, however, for temperatures as high as the last two decades of the twentieth century having occurred at any time in the past 100 000 years. Similarly, ice core records show that the background of natural variability in atmospheric chemistry is large. However, increases in most chemical species in the last 100–200 years appear to be unprecedented in magnitude and rate of change. Particularly clear signals of the anthropogenic impact on the chemistry of the troposphere include the following examples. Ice core measurements show that, relative to the AD 1900 value, concentrations of CH_4 have doubled, CO_2 has increased by 25%, and N_2O by 10%. In Northern Hemisphere ice cores, SO_4^{2-} concentrations increased by more than a factor of 4 between 1900 and 1970. This is in good agreement with the documented increase in SO_2 emissions from industrialized nations. Northern Hemisphere ice cores also document an overall decrease in SO_4^{2-} since 1980, probably reflecting abatement measures in the United States and Europe. Nitric acid deposition has similarly increased through most of the twentieth century and continues to increase. About 50% of the current NO_3^- deposited to Arctic snow can be attributed directly to anthropogenic NO_x emissions. Concentrations of heavy metals in Northern Hemisphere ice cores have increased since the late eighteenth century in parallel with growth in use since the industrial revolution. Isotopic studies have been used to identify source regions for these contaminants and show that United States emissions contributed about two-thirds of the deposition of lead from the atmosphere over Greenland until the late 1970s when the use of leaded gasoline declined. Finally, analysis of ice cores from several Arctic and temperate alpine locations shows that deposition rates of organochlorine pesticides remain at levels comparable to those of the 1970s when production and use was much higher than at present. The continued high concentration of these compounds in Arctic snowfall likely reflects their long residence times in soils.

See also: **Arctic and Antarctic:** Arctic Climate. **Climate and Climate Change:** Carbon Dioxide; Climate Variability: North Atlantic and Arctic Oscillation. **Cryosphere:** Sea Ice; Snow (Surface). **Land-Atmosphere Interactions:** Trace Gas Exchange. **Paleoclimatology:** Varves. **Synoptic Meteorology:** Anticyclones.

Further Reading

- Alley, R.B., Bender, M.L., 1998. Greenland ice cores: frozen in time. *Scientific American* 278, 80–85.
- Bales, R.C., Wolff, E.W., 1996. Chemical Exchange between the Atmosphere and Polar Snow. In: NATOASI Series 1, vol. 43. Springer Verlag, Berlin.
- Cuffey, K.M., Brook, E.J., 2000. Ice sheets and the ice-core record of climate change. In: Jacobson, M.C., Charlson, R.J., Rodhe, H., Orians, G.H. (Eds.), *Earth System Science – from Biogeochemical Cycles to Global Change*. Academic Press, London, pp. 459–494.

Ice Core Studies of Global Biogeochemical Cycles. In: Delmas, R.J. (Ed.), NATO ASI Series 1, vol. 30. Springer-Verlag, Berlin.

Greenland Summit Ice Cores: Greenland Ice Sheet Project 2/Greenland Ice Core Project. In: Hammer, C., Mayewski, P.A., Peel, D., Stuiver, M. (Eds.), Special issue of *Journal of Geophysical Research*, vol. 102. American Geophysical Union, Washington DC no. C12.

Hondoh, T. (Ed.), 2000. *The Physics of Ice Core Records*. Hokkaido University Press, Sapporo.

Legrand, M., Mayewski, P., 1997. Glaciochemistry of polar ice cores: a review. *Reviews of Geophysics* 35, 219–243.

Oeschger, H., Langway, C.C. (Eds.), 1989. *The Environmental Record in Glaciers and Ice Sheets*. Wiley, New York.

Paterson, W.S.B., 1994. *The Physics of Glaciers*. Pergamon, New York.

Petit, J.R., Jouzel, J., Raynaud, D., et al., 1999. Climate and atmospheric history of the past 420,000 years from the Vostok ice core, Antarctica. *Nature* 399, 429–436.

Varves

R Gilbert, Queen's University, Kingston, ON, Canada

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by A Weinheimer and F Biondi, volume 4, pp 1680–1685, © 2003, Elsevier Ltd.

Synopsis

A varve is a demonstrably annual marine or lacustrine sedimentary deposit. Varves provide a good method of dating past events. Because the deposition of sediment is related to environmental processes that are ultimately controlled by climate, the physical, chemical, or biological characteristics of a varved record provide a good proxy for climate and climate change.

Introduction

A varve is a deposit of sediment that can be shown to have accumulated during about 1 year. It is distinguished from deposits above and below by rhythmicity in one or more observable characteristics occurring in response to seasonal fluctuations of physical, chemical, or biological processes. The term derives from the Swedish word *varv*, meaning a layer, for example of stitches in a knitted garment or a lap around a racecourse. It was coined by de Geer (e.g., 1912) from his work on the Quaternary glacial-lacustrine deposits of Sweden (Figure 1(a)). The value of varves as climatic and paleoclimatic proxies lies in (1) demonstration that they are annual deposits and therefore may be used to determine absolute chronology, and (2) assessment that their sedimentary properties relate demonstrably to hydroclimatic processes. Although a varve is commonly defined as a couplet of light and dark sediment (Figure 1(b)) and so appears much like a tree ring, the spatial and temporal variability of environmental processes means that many varves are much more complex (Figure 1(c)). This often makes counting them more difficult, and so may make determination of chronology uncertain, but they offer a potentially rich source of information on regional climate as well as on specific events of weather, such as severe storms.

Formation of Varves

The depositional environment in which varves form is normally aquatic, although varves may occur subaerially, for example as a result of seasonally varying aeolian processes or snowfall (deposition of the crystalline mineral H_2O) in the accumulation areas, especially mid- and low-latitude alpine glaciers. In the latter case, melt or windblown dust in summer distinguishes depositions in each year as the snow transforms into glacial ice.

Aquatic varves have been described from many settings in lakes and the sea, but they all fall into one of three classes. *Clastic varves* normally consist of laminae of fine silt and clay-sized sediment deposited from suspension during periods of limited inflow of water and sediment, and periods of diminished processes of distribution in the water body. Coarse silt and sand layers are deposited in response to abundant inflow and vigorous circulation in the water body. Commonly, these couplets cannot be distinguished in fresh samples of wet

sediment, but on drying, the sand and coarse silt take on a light tone, while the hydrophilic clay remains moist and dark. Thus, with careful attention during drying, clastic varves are easily distinguished by eye (Figure 1(a)–1(c)), and photographic records are normally the best for analysis of thick varves, while thin sections are commonly used for submillimeter-sized varves.

The deposition of fine- and coarse-grained layers in a clastic varve is enhanced by processes within the water body. In lakes experiencing sediment-laden inflow, turbidity currents efficiently deliver bursts of sediment that are deposited nearly instantaneously as graded beds, and thus distinct layers are associated with each, often daily, event. Complex varves commonly form in this manner. The silty sand layers within the dark winter clay layers shown in Figure 1(c) result from the incursion of intense autumn and winter storms from the Pacific Ocean into the mountains. Snowmelt augmented by warm rain generates large floods that deliver bursts of coarser sediment to the lake, where turbidity currents efficiently transport them to the lake floor. The fine-grained laminae are deposited from settling through the water column. According to Stokes' law, a 1 μm diameter particle requires about 3 years to settle through a 100 m deep lake. Recent studies have demonstrated that the formation of small flocs in fresh water greatly increases settling velocity and allows the fine-grained sediments to accumulate during several months of quiet conditions, which are appropriate for the formation of the winter cap on a varve. In salt water, much larger flocs are formed and are removed from the water column in a few days to weeks. This may be a reason why clastic varves are uncommon in marine settings.

Biogenic varves form by the seasonal deposition of organic material derived from land (e.g., pollen production in spring) or originating in the water body itself (most commonly as blooms of diatoms or other aquatic organisms). Deposition of darker terrigenous organic residue and inorganic sediment separates the lighter biogenic laminae and so forms varves. These are the most common type of marine varves (Figure 1(d)); they usually form in basins with anoxic bottom water that prevents the establishment of a benthos that would bioturbate and destroy varves as they formed. Biogenic varves are also found in lakes, but here anoxic conditions are less important because the lacustrine benthos is commonly sparse or absent. Nevertheless, some of the best varve sequences occur in strongly stratified meromictic lakes that have almost no

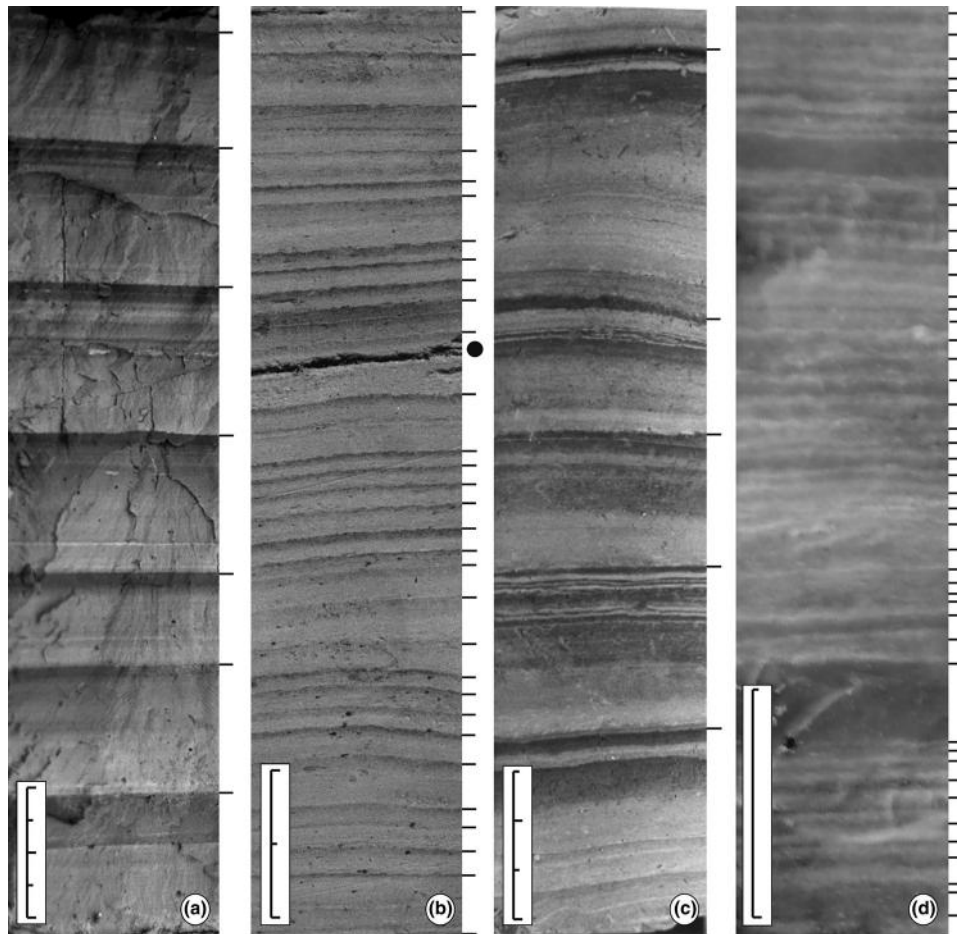


Figure 1 (a) Section *in situ* of early Holocene glacialacustrine varves from Bergsbrunna, Sweden. These are part of the 12 000-year record first described by de Geer in the late nineteenth century. (b) Simple varves in a core of modern sediment from glacially dammed Ape Lake, British Columbia. Although these varves consist predominately of light and dark couplets because sediment input is in a melt-water pulse directly from the damming glacier and other sources of sediment are unimportant, variability within some varves can be distinguished. The dot indicates a crack formed during drying of the core. (c) Complex varves in a core from montane Lillooet Lake, British Columbia, where input is from large alpine glaciers up to 80 km distant, as well as from a variety of mountainside and valley-bottom sources. Interplay between maritime and continental air masses in the Cordillera contributes to the complexity. (d) Lithified Miocene marine varves in a hand specimen from the Monterey Formation of Southern California. The diffuse light-toned laminae are the result of annual diatom blooms. In each case, the varves are distinguished by tick marks to the right, and the scales are in 5 mm divisions.

currents at depth and where dissolved oxygen is low or absent. Anoxia also slows the decay of organic carbon that may form the distinctive lamina in some biogenic varves.

Chemogenic varves form in lakes as a result of the seasonal precipitation of salts, especially calcium carbonate, from a supersaturated solution created during summer when water temperature is high or when the uptake of carbon dioxide by aquatic vegetation is greatest. Coarse particulate carbonate normally settles through the water and becomes buried sufficiently quickly that most is not redissolved in the undersaturated cold-water deep in these lakes. In winter, the whole lake is sufficiently cold that carbonate does not precipitate, and even a small amount of terrigenous sediment is sufficient to produce a distinctive lamina to form a varve. Biogenic and chemogenic processes are more likely to form simple varves because they commonly occur only once per year, whereas processes associated with the deposition of clastic sediment are often of shorter duration and higher frequency.

Varve Chronology

Because the boundaries of a varve are normally placed at the most sharply distinguished transition between different forms of sediment (Figure 1), and because this usually occurs in spring or summer (associated with the onset of distinct processes of higher energy or activity), a varve does not correspond to a calendar year. As well, because the onset of seasonal events such as nival or glacial melt, monsoon rain, a diatom bloom, or precipitation of salts varies from year to year, sometimes by weeks or even by a month or more, neither does a varve represent exactly 1 year of deposition. These distinctions normally do not affect the use of varves as a measure of absolute chronology, but they may become significant when varves are used as proxy for particular seasonal events.

The value of varves as a dating tool depends on demonstration that their cyclicity is annual. Dates obtained from

varve counting may be related to independent evidence such as (1) radiometric dates, (2) the correlation of thickness time series with other proxies such as tree rings, coral bands, or glacial ice layers, (3) time-stratigraphic markers such as tephra, (4) measured or calculated rates of sediment accumulation, for example in sediment traps or from sub-bottom acoustic or seismic surveys, or (5) measured hydroclimatic variables. However, many accounts of varves depend only on the inference that strong, regular cycles must be annual, notwithstanding other regular cycles such as tides, daily temperature, rainfall and stream discharge, and even frequent storms that may also produce strong, regular cyclic effects. When the annual character cannot be demonstrated, cyclic deposits should be referred to as *rhythmites*.

Counting varves, like tree rings, is a simple procedure, although complex varves especially offer the likelihood of significant overestimation of age (Figure 1(c)). As well, if seasonal variation is sufficiently muted in some years, for example during a cool, dry summer when inflow of water and sediment is reduced, or precipitation of salts fails to occur, a varve may not be recognized, and thus, the age may be underestimated. Such errors can be reduced by comparison of time series records from a number of sites in the same or nearby water bodies, of other time series such as tree rings, or of independent counts of the same varve record by different observers. Techniques such as dot counting, which was developed by dendrochronologists, are now commonly used to determine varve chronologies.

The longest varve records, especially from marine sediments, extend over many thousands of years. Lacustrine records tend to be shorter, and fewer ancient records exist because lakes exist for shorter periods, varves tend to be thicker in lakes (Figure 1) and so fewer are recovered in a single core, and the heavy coring equipment available on large ships normally cannot be brought onto lakes. Nevertheless, a few long lacustrine records have been compiled by deep drilling from fixed platforms, or by linking the variations in thickness in modern lakes to those in progressively older lakes, in the manner that dendrochronologists link tree ring records from modern trees to ancient logs.

Varves as Hydroclimatic Proxy

Figure 2 suggests that there are a number of at least partially independent steps between weather and climate, and the record derived from varved sediment, each of these steps may limit the relation between these end members and so the value of varves as proxy. The production of terrestrial sediment by weathering and erosion involves processes that are at least as much a function of the geology of the parent material, geomorphic forces (e.g., colluvial activity and glaciation), and human factors (e.g., land use) as are the processes of climate. Exhaustion of large sediment supplies such as those produced by extensive glaciation may occur over thousands of years following retreat of the glaciers, and so the character of varves may change slowly and independently of climate. Fluvial transport sorts sediments spatially and temporally by size and composition, commonly storing materials for lengthy periods before delivery to the lake or sea. Hysteresis in the relation

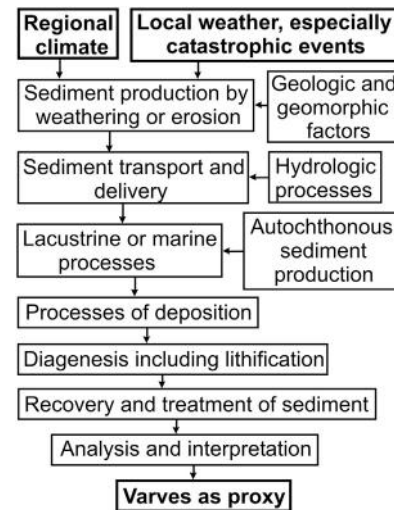


Figure 2 Conceptual model showing the relation of a varve record to the ultimately controlling factors of weather and climate. The model is not comprehensive, and only the principal links are shown. Concepts are described in this article.

between water and sediment discharges further distorts the relation between hydrological factors and sediment transport in the short term.

Processes in a water body such as waves, the formation of chemical precipitates, circulation, and currents determine how sediment is distributed through the water body. Deposition of sediment is determined by processes such as settling from suspension, flocculation, gravity flow, and focusing of sediment from other parts of the water body, all of which have only weak links to climatic control. Diagenetic processes begin to alter the sediment immediately following deposition. These include dewatering and compaction so that earlier varves become progressively thinner, and chemical changes and physical disturbance occur. Lithification of sediment into sedimentary rock under high heat and pressure further changes the properties of the deposits.

Significant disturbance also occurs during recovery of the sediment from the lake or sea. Gravity corers compact and distort the sediment as they penetrate, for example by causing bowing down of the laminae due to friction with the side walls in varying amounts depending on the shear strength of individual laminae (Figures 1(b), 1(c), and 3). Piston corers may seriously distort sediment by suction or by contact with an unsecured piston during recovery. Improper transportation and storage may further disturb the sediment. During drying to best reveal the varves, shrinkage of the sediment leads to the formation of cracks and fissures (Figure 1(b)), and complete drying to prepare thin sections may cause larger cracks such as the large dark layers in the section shown in Figure 3. The use of X-radiography lessens this problem and, especially with marine varves, may produce superior results.

Despite all these limitations, many varve sequences do make excellent proxies for past general conditions of climate as well as for individual events of weather such as major storms. Most commonly, the time series of varve thickness is linked by

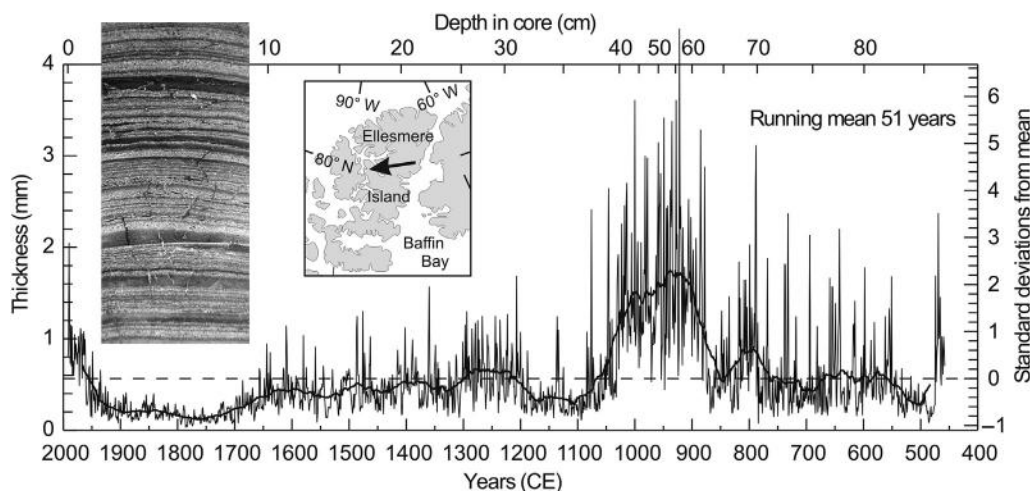


Figure 3 Varve record from Romulus Lake, Ellesmere Island, Nunavut. A photograph of a thin section of 40 mm of the core from 70.2 to 74.2 cm depth containing 68 varves is inset. It shows the variability in varve thickness and the disturbance (bowing) due to coring. The location of the lake is marked by an arrow on the inset map.

inference to climatic controls, as illustrated by **Figure 3**. Romulus Lake in the Canadian High Arctic has a concentration of salt three times that of seawater due to its isolation from the Arctic Ocean during isostatic uplift and due to the rejection of brine by growing permafrost in the drainage basin newly exposed by the uplift. The very strong stratification by density in this meromictic lake provides ideal conditions for the formation of varves. A relation to varying climate during the more than 1500-year record is inferred from understanding of periglacial conditions in the region. During the Medieval Warm Period centered on the eleventh century, higher temperatures and possibly greater rainfall caused the active layer to thicken and so melt the ice-rich permafrost beneath to produce major slope failures. This resulted in large pulses of water heavily charged with fine-grained marine silts and clays being delivered to the lake. Because this process is irregular in space and time, the variability in varve thickness is much greater during this period. Subsequent cooling of the climate, along with exhaustion of sediment as a result of these events, caused the permafrost to reestablish, the active layer to thin, and sediment delivery to decrease. Thus, the varves became thinner and the time series more regular. This culminated in the Little Ice Age of the eighteenth century, when conditions were coldest and sediment delivery least. A significant rise in thickness during the twentieth century may relate to conditions of global warming. If so, conditions equaling or exceeding those of the Medieval Warm Period may be expected in future. Similar, although less dramatic, effects have been reported from other arctic lakes as well as from ice-core records. It is in this convergence of evidence from a number of proxies that the strength of this method lies.

A second method for the use of varves as climatic proxy involves the development of a calibration based on the characteristics of varves, including thickness, stratigraphy, or composition, compared with instrumental records of weather and climate, or of observed processes of sediment production, delivery, and deposition. Statistical relations between, for example, varve thickness and mean annual or seasonal temperature or precipitation commonly do not produce

results in which more than 50–70% of the variability of varve thickness (the coefficient of variation, r^2) is accounted for by the measured variables of weather. Somewhat better results have been obtained in the relation between varve thickness and measured delivery of water or sediment to the water body because this involves fewer steps in the cascade (**Figure 2**). Studies of intraannual variation in the stratigraphy of varves have been linked to individual events of weather or climate. For example, each of the laminae in **Figure 1(c)** can be associated with discharge events in the inflowing river related to snow melt, glacial melt (and, by implication, summer temperature), or rainfall, including winter storms. Even though these and similar observed or implied relations have been determined in many studies, application to a varve time series from a different setting would be inappropriate because the relations are not universal. Even application of the calibration from modern varves to those that are from the same location but older than the instrumental record must be performed cautiously because the links shown in **Figure 2** are not temporally stable.

See also: Biogeochemical Cycles: Heavy Metals; Sulfur Cycle. Climate and Climate Change: Climate Variability: Decadal to Centennial Variability. Paleoclimatology: Ice Cores.

Further Reading

- Baumgartner, T.R., Michaelsen, J., Thompson, L.G., Shen, G.T., Soutar, A., Casey, R.E., 1989. The recording of interannual climatic change by high-resolution natural systems: tree-rings, coral bands, glacial ice layers, and marine varves. *American Geophysical Union Geophysical Monographs* 55, 1–14.
- de Geer, G., 1912. A geochronology of the last 12,000 years. In: *Proceedings of the International Geological Congress, Stockholm, 1910*, vol. 1, pp. 241–253.
- Fisher, C.G., 1990. Bibliography and Inventory of Holocene Varved and Laminated Marine Sediments. NOAA Paleoclimate Publications Series Report No. 1.
- Hodder, K.R., Gilbert, R., Desloges, J.R., 2007. Glacilacustrine varved sediment as alpine hydroclimatic proxy. *Journal of Paleolimnology* 38, 365–394.
- Kemp, A.E.S. (Ed.), 1996. *Palaeoclimatology and Palaeoceanography from Laminated Sediments*. Geological Society, London, Special Publication No 116.

RADAR

Contents

Cloud Radar

Incoherent Scatter Radar

Mesosphere–Stratosphere–Troposphere and Stratosphere–Troposphere Radars and Wind Profilers

Meteor Radar

Polarimetric Doppler Weather Radar

Precipitation Radar

Synthetic Aperture Radar (Land Surface Applications)

Cloud Radar

T Uttal, NOAA, Boulder, CO, USA

© 2003 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 4, pp 1795–1802, © 2003, Elsevier Ltd.

Introduction

Cloud radars (**Figure 1**) are active remote sensors that transmit at millimeter wavelengths and are particularly suited for determining the properties of nonprecipitating clouds such as stratus, altostratus, and cirrus. Millimeter-wave radars typically operate near 35 GHz ($\lambda = 8.6$ mm, Ka-band) or near 94 GHz ($\lambda = 3.3$ mm, W-band), which are frequency ‘window’ regions where absorption by atmospheric gases are at a minimum. Despite this, attenuation from atmospheric liquid and vapor is more of an issue for these short-wavelength radars than for longer-wavelength (centimeter scale) weather radars. Therefore, cloud radars often have range limitations of 20 km or less and are not suitable for observing conditions of moderate to heavy rainfall. The primary advantage of shorter-wavelength radars lies in their sensitivity to small ice particles and cloud droplets that can be two orders of magnitude smaller than typical raindrops. Some millimeter cloud radar systems are multiparameter; in other words, they transmit and receive electromagnetic waves at more than one wavelength or polarization.

Millimeter-wave radars are still primarily research instruments, typically unique prototypes designed and operated by university, government, or private research organizations. The rapid development of millimeter-wave cloud radars in the 1990s has been largely driven by the need to observe characteristics of nonprecipitating clouds in order to determine cloud impacts on global atmospheric radiation and consequences for climate predictions by large global climate models. In addition, scanning and airborne millimeter radars are used to make measurements necessary for understanding the physics of precipitation and the mechanisms by which precipitation forms. Cloud radars also are providing promising technologies for some operational applications, such as remotely sensing aircraft icing conditions (**Figure 2**).



Figure 1 Portable cloud radar installation. Photo credit Tom Ayers.

Physics

Radar backscattering cross-sections (σ) for small particles vary inversely with the radar wavelength (λ) to the fourth power ($\sigma \propto \lambda^{-4}$). This physical property is the basis of the utility of short-wavelength radars for observing small particles and it enables cloud radars to utilize small antennas and to transmit peak power levels that are many orders of magnitude lower than those of weather radars. The beam width (θ) of a radar is related to the wavelength (λ) and the radar antenna diameter (A) by the relationship $\theta \propto \lambda A^{-1}$. This relationship means that, relative to antenna size, shorter-wavelength radars can obtain narrower beam widths; this, combined with short power pulses, allows for high-resolution observations (on the order of tens of meters) capable of resolving the fine-scale structure of clouds. With their lower power and small antenna requirements, cloud radar systems are ideal for portable systems that can be deployed in field experiments, in remote locations, as well as on ship, aircraft, and satellite platforms.

Received radar power (P) is measured in decibels (dB) referenced to 1 milliwatt (dBm) or $10 \log_{10} (P/\text{mW})$. The log scale is necessitated by the fact that returned power can vary over 8 orders of magnitude. The total received radar power is a function of three factors: distributed cloud target characteristics (effective reflectivity factor measured in dBZ_e referenced to $1 \text{ mm}^6 \text{ m}^{-3}$), the distance of the target(s) from the radar (r), and the radar constant (RC), which is based on the radar hardware characteristics (eqn [1]).

$$P(\text{dBm}) = Z_r(\text{dBZ}_r) - 20 \log r(\text{km}) - RC(\text{dB}) \quad [1]$$

This equation indicates that a smaller radar constant (RC) will result in more sensitive radar, since for a given atmospheric return (Z_r) at a given range (r), a smaller value of RC results in a stronger received signal (P). Thus, a critical goal of radar hardware design is minimization of the radar constant by optimally designing such factors as transmitted power level and antenna size. The radar reflectivity factor in dBZ_r can be converted to standard radar units of $\text{mm}^6 \text{ m}^{-3}$ using eqn [2].

$$Z_e = 10^{(\text{dBZ}_r/10)} \quad [2]$$

For Rayleigh scattering conditions ($D \ll \lambda$), the equivalent radar reflectivity factor of a population of hydrometeors is related to the diameters (D) of equal volume spherical particles (mm) and particle size distribution (N). For water droplets Z_r is given by eqn [3] and for ice particles by eqn [4], where ρ_0 is the ice particle bulk density in g cm^{-3} . These two equations are the basis for obtaining cloud microphysical properties from radar reflectivities.

$$Z_e = \int_0^\infty N(D) D^6 dD \quad [3]$$

$$Z_e = 0.2 \int_0^\infty N(D) [D \rho_0^{0.3333}]^6 dD \quad [4]$$

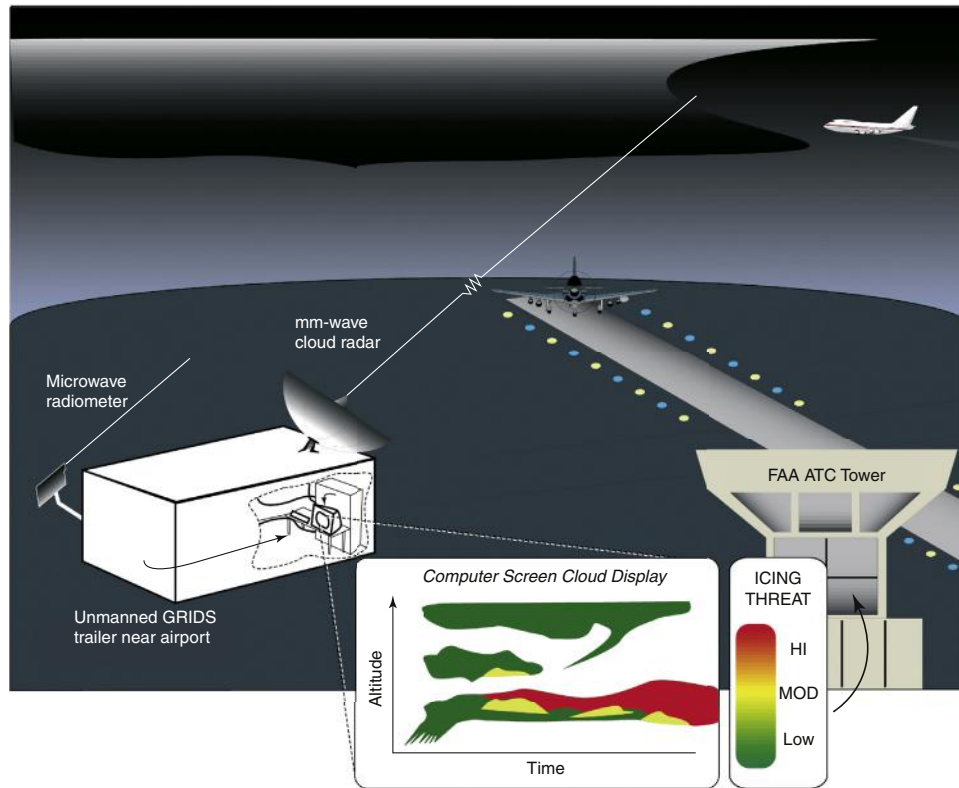


Figure 2 Use of cloud radar to sense aircraft icing conditions. Figure courtesy of Brooks Martner.

Typically, cloud radars also measure small Doppler shifts between the transmitted and received signal, which enables the calculation of the radial speed of cloud hydrometeors away from or towards the radar. Velocity measurements can be achieved with a precision on the order of 5 cm s^{-1} ; this is precise enough to be useful in measuring the fall velocities of ice crystals or cloud droplets as well as the turbulence characteristics of clouds. Some systems also record the complete Doppler spectra, in other words, the frequency distribution of Doppler velocities of the billions of hydrometeors within each radar sampling volume. Others record only the width of the Doppler spectrum in order to reduce the size of the collected data set.

History

One of the earliest meteorological uses of millimeter-wave radars was by the US Air Force, which operated 35 GHz radars at airbases as aids to aviation operations and forecasting in the 1960s. These non-Doppler systems successfully determined bulk cloud geometry, for instance cloud layer altitudes; however, these systems were eventually decommissioned owing to repeated failures with hardware. Some of the systems and parts from the systems were transferred to university research groups, but in general, the late 1970s and early 1980s were a period during which work with millimeter-wave radars declined. In the early 1980s, advances in solid-state, low-noise electronics inspired a number of research groups to develop both surface-based and airborne cloud radars and to demonstrate their usefulness for elucidating the properties of non-precipitating and lightly precipitating clouds. In particular, scanning and depolarization capabilities were added to greatly increase the utility of these systems. In addition to the engineering advances, innovative techniques for combining radar and lidar measurements, as well as radar and radiometer measurements (microwave and infrared), were developed that have led to the ability to estimate values of important microphysical and optical cloud properties. These include, but are not limited to, crystal/droplet size, concentration, water content, and cloud optical depth. Radars with polarization capabilities provide additional information for determining cloud crystal or droplet shape and orientation, and by inference, phase. The 1990s were also the first decade when a few millimeter-wave radars began operation on a continuous and long-term basis for climate research.

Meteorological Applications

Cloud Structure

Nonprecipitating and lightly precipitating clouds are predominant features of the Earth's physical environment; however, quantitative measurements of seemingly straightforward features such as cloud height and thickness and the number of cloud layers have historically been extremely difficult to obtain. Without radar measurements, the primary alternative for observing these important cloud properties are detection of cloud bases from laser ceilometers operating at airports and weather stations, cloud tops estimates from satellite radiance measurements, and qualitative visual

estimates by human observers of the heights of clouds and number of layers. In contrast, cloud radars provide comprehensive and detailed information on the geometrical morphology of clouds within the constraints of attenuation at one end of the detection scale and nondetection of extremely small cloud hydrometeors at the other. In interpreting cloud boundaries, it must be considered that there is often no significant difference in radar reflectivities between cloudy regions and adjacent regions of precipitation; these regions must be distinguished by the change in structure of the Doppler velocities, the Doppler spectra, or with additional measurements as might be provided by another instrument such cloud base ceilometer or lidar. Figure 3 shows an example of simultaneous, complex, multilayered cloud scenes from operational cloud radars in Barrow, Alaska (top panel) and in Lamont, Oklahoma (bottom panel) for a 24-hour period on 29 July 1998. The labels on the figure indicate a number of interesting features detected by the radars including cirrus, melting-layer bright bands, light drizzle, altostratus, boundary-layer stratus, and insects.

Cloud Water Contents and Crystal and Droplet Size Determination

A number of techniques that utilize cloud radar measurements, either alone or in combination with additional measurements from lidars or radiometers, can be used to retrieve information on cloud properties. The term 'retrieve' in this context indicates the process of theoretically inferring physical cloud microphysical and optical aspects (such as cloud water contents, crystal or droplet sizes, and cloud optical depths) from some combination of directly measured quantities (such as radar reflectivity, Doppler velocities, radiometric brightness temperatures, and lidar backscatter). These methods can be divided into those that use radar data only, those that use radar and radiometer data, and those that use radar data in combination with lidar measurements.

The earliest radar-only retrieval methods were based on empirical relationships between radar reflectivity factor and ice water content (IWC) or liquid water content (LWC) of the form given in eqn [5], where the a and b coefficients were typically based upon physical regressions from either *in situ* cloud ice crystal or droplet measurements from aircraft or from ground observations.

$$\text{IWC or LWC} = a \times (Z_e)^b \quad [5]$$

Although useful, these methods tend to have a wide range of variability in values of the a and b coefficients, depending on the specific individual cloud data sets that are used to determine the coefficients.

The next generation of retrieval techniques for ice cloud parameters used radar reflectivity measurements to obtain a cloud layer average Z_e , and infrared radiometer measurements of the brightness temperature of downwelling radiation in the $10 \mu\text{m}$ to $11.5 \mu\text{m}$ spectral band to calculate cloud optical depth. Using the fact that layer-mean reflectivity and the brightness temperature are proportional to the sixth and the second moments of the cloud particle size distribution, respectively, cloud layer averages of crystal sizes can be estimated. The optical depth calculation can also be used to

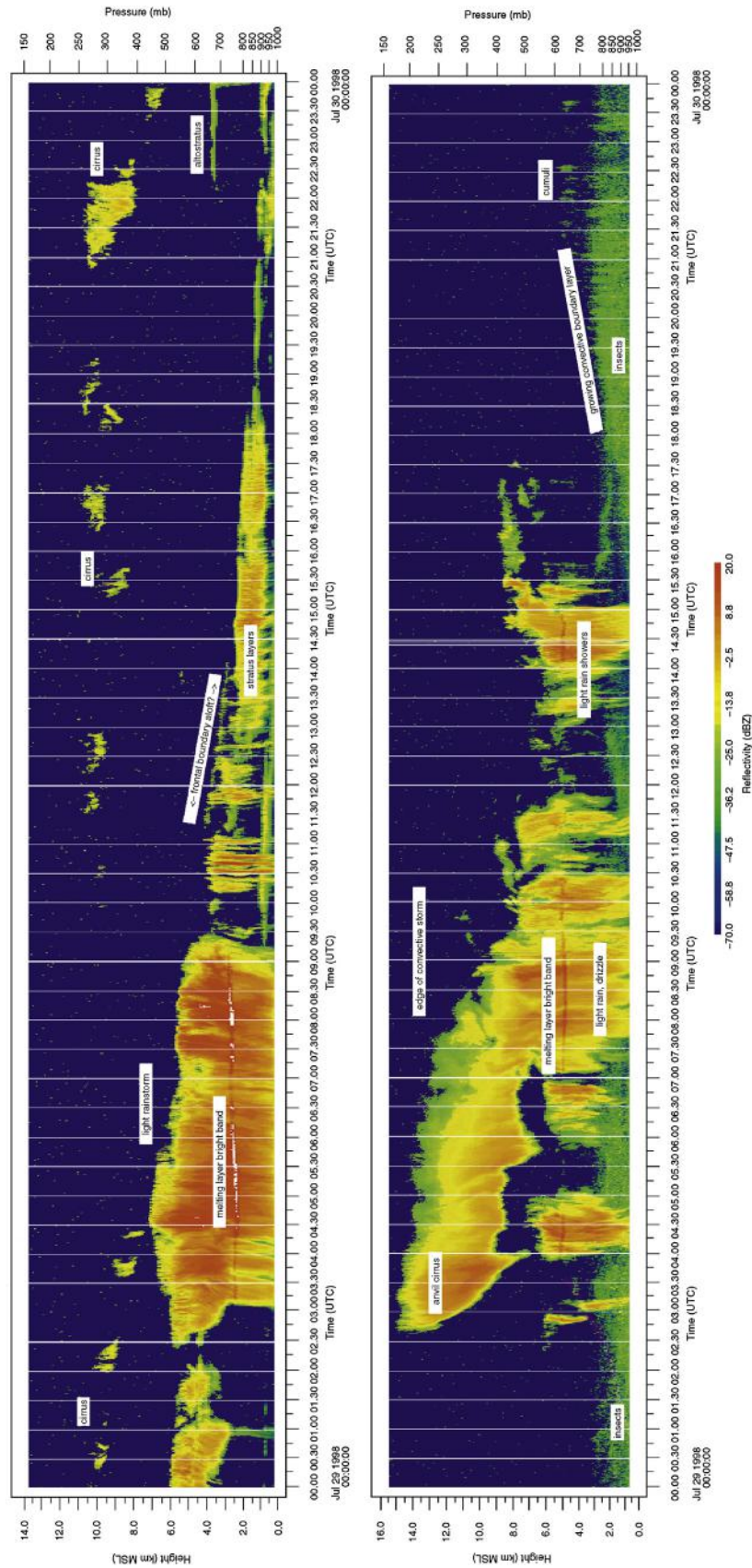


Figure 3 Simultaneous multilayered cloud scenes produced by operational cloud radars in Barrow, Alaska (top panel) and Lamont, Oklahoma (bottom panel). Figure courtesy of Brooks Martner.

determine a specific value for the coefficient a , and a fixed value can be reasonably assumed for the coefficient b , which in general has a smaller degree of variability (between 0.55 and 0.75). Eqn [5] can then be applied through the depth of the cloud to calculate profiles of ice water content. From these profiles of ice water content, mean particle diameters are calculated by using an assumed relationship between ice crystal size and bulk density. Additional adjustments to this technique involve the use of the microphysical profiles to calculate downwelling solar irradiances with a radiative transfer model, comparison to measured values, and iteration to achieve consistency with the additional solar measurements.

A drawback to the techniques that incorporate infrared radiometer data is that retrievals are limited to situations in which ice clouds are optically thin and are not observationally obscured by intervening liquid clouds which obfuscate the radiometer measurements. The need for retrievals that can be applied to a broader range of cloud types and multilayer situations has led to the most recent development in radar-based retrievals that utilize radar Doppler velocities. In this method, Doppler velocities are averaged over time periods sufficient to average out the effect of vertical air motions to yield estimates of crystal fall speeds. The crystal fall speeds can then be used to estimate crystal sizes and ice water contents.

The development of retrievals for liquid clouds has paralleled that of ice clouds, with estimates of optical depth from infrared radiometers being replaced by measurements of column-integrated liquid-water path from microwave radiometers. The original empirical relationships of the form described by eqn [5] were followed by techniques that used the profile of radar reflectivity, in combination with microwave radiometric liquid-water path, to determine a profile of liquid-water contents based on the assumption that the drop size distribution was lognormal and the droplet concentration was constant through the depth of the cloud. A second technique uses *a priori* data sets from aircraft collected in an ensemble of clouds similar to those for which the retrievals are to be

performed. From these, empirical regressions are determined between both radar reflectivity and liquid-water content, as well as between radar reflectivity and droplet sizes. This second technique can be used when microwave radiometer data is unavailable or it is unclear which cloud layers or levels contain the liquid layers. Figure 4 shows an example of cloud-water contents that have been retrieved for a low-level liquid water cloud using radar and microwave radiometer, and an overlying upper-level ice cloud using radar-only measurements of reflectivity and Doppler velocity.

Another completely different approach to retrievals of cloud properties is to use cloud radar in combination with cloud lidar. Radar and lidar retrieval techniques use Mie scattering theory and accounts for partial attenuation of lidar signals. This enables theoretical calculation of radar and lidar backscatter cross-sections, and the ratios of the backscatter cross-sections can be then computed as a function of particle size. These computed curves can be used to interpret measured radar and lidar cross-sections in terms of ice crystal sizes.

Note that the methods described here have been developed for either all-ice clouds or all-liquid clouds. In the real environment many clouds have mixed phase elements. Since these cloud types have different microphysical and radiative characteristics than the simpler single-phase clouds, their characterization remains a priority research issue for the cloud radar research community.

Crystal and Droplet Shape and Orientation

Electromagnetic radiation consists of perpendicular electric and magnetic fields that can be directionally oriented or 'polarized'. By convention, the polarization of an electromagnetic wave refers to the direction of the orientation of the electric field. Cloud hydrometeors will depolarize (change the orientation) of an electromagnetic wave when it is scattered back to the radar receiver as a function of hydrometeor shape (with implications for cloud phase), aspect ratio, orientation, and

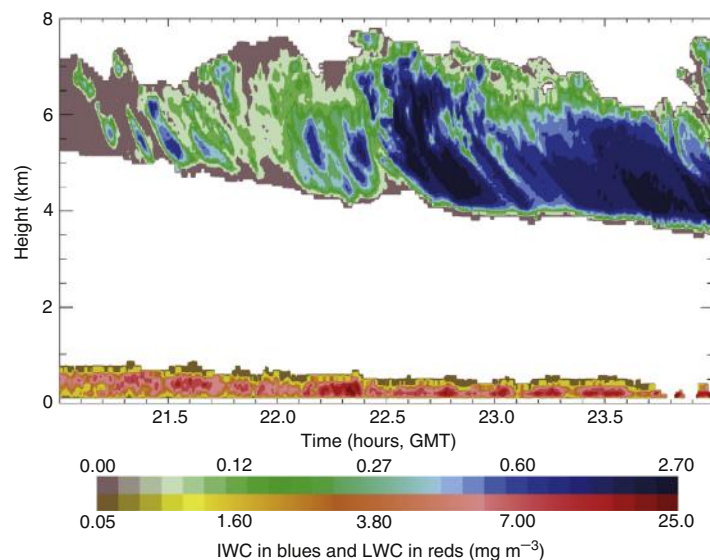


Figure 4 Ice-water and liquid-water content of clouds determined using a combination of radar-only and radar-radiometer retrieval techniques.

bulk density. Thus information on these cloud parameters can be obtained by examining how polarization of the received signal has changed with respect to the transmitted signal. To measure this change, the radar must transmit and/or receive in two separate polarizations, allowing the calculation of depolarization ratios. Such radars are referred to as polarimetric or dual polarization radars. Various polarization states ranging from linear to circular are possible, depending on how vertically and horizontally polarized pulses are mixed in the transmission scheme. Common schemes include utilization of linear depolarization ratios (LDR), circular depolarization ratios (CDR), differential reflectivities (ZDR), and the more recently developed 45° slantwise depolarization ratios (SLDR). These are summarized in Table 1, along with power ratio ranges that can result from different types of cloud targets.

Theoretically, perfect spheres (such as very small cloud droplets) will have no depolarizing effect on electromagnetic waves, resulting in a depolarization ratio of $-\infty$, while scatterers with large length-to-width aspect ratios (such as chaff) will be highly polarizing, with depolarization ratios approaching 0 dB. Meteorological targets such as drizzle, dendrites, plates, columns, and graupel will have depolarization ratios in between these values. It should be noted that, in practice, it is very difficult to generate or measure electromagnetic waves with perfect circular or linear depolarization, and the lower limit for depolarization ratios is determined by the inability to perfectly isolate the power leakage between different polarization states generated by the radar. This lower limit is referred to as the antenna cross-talk limit, or antenna cancellation ratio

(denoted as $-\infty$ in Table 1) and will vary as a function of the type of polarization chosen. Therefore, in practice, it is optimum to choose polarization states for which a maximum dynamic range can be achieved since rather than $-\infty$, the true lower threshold will be a system limited finite value.

Because certain ice crystal types have preferred fall orientations, the polarization ratios observed by radar for the same target type can also vary as a function of the elevation angle at which the radar views the cloud. For this reason, scanning dual-polarization radar through a cloud can be a very powerful tool. Figure 5 shows 45° slantwise depolarization ratios as a function of radar elevation. This figure clearly demonstrates the utility of this radar polarization scheme for determining hydrometeor type.

Platforms

The comparatively low power requirements, small antenna size requirements, and robust electronics of millimeter-wave cloud radars enable the design of systems that are suitable for deployment in portable containers and on vehicles, aircraft, and ships. New engineering designs have also resulted in unattended systems that can operate continuously, and an increasing number of such systems are collecting data around the globe both in long-term atmospheric monitoring programs and in short-term field experiments. Present plans indicate that the first space-based cloud radar will be launched sometime between 2004 and 2010. Such space-based cloud radars, together

Table 1 Common radar polarization schemes

Target	Linear depolarization ratio (LDR) [10 log (Z_{HV}/Z_{HH})]	Differential reflectivity (ZDR) [10 log (Z_{HH}/Z_{VV})]	Circular depolarization ratio (CDR) [10 log (Z_{LL}/Z_{LR})]	45° Slantwise depolarization ratio (SLDR) [10 log ($Z_{(135)(45)}/Z_{(45)(45)}$)]
Cloud droplets	$-\infty$ dB	0 dB	$-\infty$ dB	$-\infty$ dB
Ice crystals	$-\infty$ to -10 dB	0 to 4 dB	$-\infty$ to -7 dB	$-\infty$ to -7 dB

Note: for each Z_{XY} , Z is the power, X is the received polarization, and Y is the transmitted polarization, where H = horizontal linear, V = vertical linear, L = left-hand circular, R = right-hand circular, 135 = linear rotated 135° from horizontal, 45 = and linear rotated 45° from horizontal.

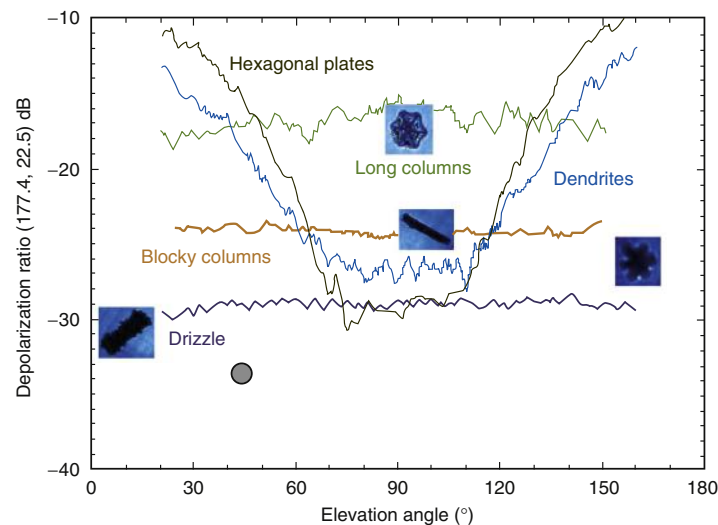


Figure 5 45° slantwise depolarization ratios for different hydrometeor types as a function of elevation angle. The different hydrometeor types are clearly distinguished by this method. Courtesy of Roger Reinking.

with expanding surface networks of cloud radars, indicate that these instruments will become an increasingly routine and integral part of cloud research and cloud monitoring.

See also: **Clouds and Fog:** Measurement Techniques *In Situ*.

Lidar: Backscatter. **Radar:** Meteor Radar; Polarimetric Doppler Weather Radar; Precipitation Radar.

Further Reading

Batten, L.J., 1959. Radar Meteorology. University of Chicago Press, Chicago.

Batten, L.J., 1973. Radar Observation of the Atmosphere. University of Chicago Press, Chicago.

Clothiaux, E.E., Miller, M.A., Albrecht, B.A., et al., 1995. An evaluation of a 94 GHz radar for remote sensing of cloud properties. *Journal of Atmospheric and Oceanic Technology* 12, 201–229.

Doviak, R.J., Zmic, D.S., 1984. *Doppler Radar and Weather Observations*. Academic Press Inc, Orlando, FL.

Goddard, E.E., Strauch, R.G., 1983. *Radar Observation of Clear Air and Clouds*. Elsevier Science Publishers B.V, Amsterdam.

Kingley, S.P., Quegan, S., 1992. *Understanding Radar Systems*. McGraw-Hill, London.

Kropfli, R.A., Kelly, R.D., 1996. Meteorological research applications of mm-wave radar. *Meteorology and Atmospheric Physics* 59, 105–121.

Moran, K.P., Martner, B.E., Post, M.J., et al., 1998. An unattended cloud-profiling radar for use in climate research. *Bulletin of the American Meteorological Society* 79, 443–455.

Rinehart, R.E., 1997. *Radar for Meteorologists*, Third ed. Rinehart Publications, Grand Forks, ND.

Incoherent Scatter Radar

MP Sulzer, Arecibo Observatory, Arecibo, PR, USA

© 2003 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 4, pp 1812–1819, © 2003, Elsevier Ltd.

Introduction

The incoherent scatter radar (ISR) technique is a powerful ground-based tool used to measure various properties of the ionized part of the upper atmosphere called the ionosphere. All radars transmit radio waves at a target and receive much weaker waves generated when electrons, the lightest charged component of the target matter, accelerate in response to the incident waves and reradiate the signal. Reflection and scatter are terms used to describe the reradiation, depending upon the degree and nature of the organization of the electrons in the target. Incoherent scatter returns come from free electrons in the ionospheric gas, or plasma, usually with a strong influence from the ions. ISRs can be used to measure electron and ion temperatures and velocities, and the number densities of the electrons and the various ions. ISR has remained a useful technique for ionospheric studies during the last 40 years, because a complete, accurate, and elegant theory describes the spectrum of the scattered signal, and because inexpensive and easily implemented digital signal processing makes the use of new radar techniques practical and allows new and better data analysis methods. Data analysis consists of comparing measured spectra with model spectra, adjusting the model parameters for a good match using nonlinear least-squares fitting. The primary task of ISRs is the global study of the effects of energy inputs into the ionosphere and upper atmosphere, that is, solar radiation and particles entering along the Earth's magnetic field lines from above, and energy, often in the form of waves, from the denser atmosphere below. ISRs verify many aspects of the behavior of the ionospheric plasma, including plasma instabilities generated naturally and artificially. ISRs also make measurements in the middle atmosphere, and can function as MST (mesosphere–stratosphere–thermosphere) radars in the lower atmosphere.

The required hardware is large: one or more antennas, 30–300 m in diameter, and a powerful transmitter, 1 MW or more, capable of transmitting for several percent of the time. Most radars are monostatic, using the same antennas for transmission and reception. Each ISR is the result of different compromises in design, trading away the less needed characteristics in order to lower the cost. The Arecibo radar (18.3° N, 293.2° E) obtains a very high sensitivity at 430 MHz by using a fixed spherical dish 305 m in diameter constructed in a sinkhole in Puerto Rico. Energy focuses to a line rather than a point as with a parabolic antenna, and the original design used a long line feed antenna. Today, a Gregorian feed with secondary and tertiary reflectors is used to correct the spherical aberration. With such a large antenna the feed is moved rather than the dish, with zenith angles of 20° attainable at Arecibo. The Jicamarca observatory in Peru (−12.0° N, 283.1° E) uses a large array of dipoles operating at 50 MHz to obtain a similar

sensitivity. It has a limited range of pointing angles nearly perpendicular to the magnetic field and is used to study the electrodynamics and other features of the ionosphere at the magnetic equator.

Radars requiring full sky coverage use movable parabolic dishes, and accept lower sensitivity as a result of the smaller antenna. The Millstone Hill facility in Massachusetts, USA (42.6° E, 288.5° N) operates at 440 MHz and has two parabolic dishes, one fixed in the vertical direction (68 m) and the other (46 m) rapidly steerable over most of the sky. The Sondrestrom facility in Greenland (67.0° N, 309.0° E) operates at about 1100 MHz and also uses a steerable parabolic dish (32 m). This radar has moved from the Stanford Campus to Alaska and then to Greenland in response to the needs of the scientific community. These four radars form a longitudinal chain operated under the CEDAR (Coupling, Energetics, and Dynamics of Atmospheric Regions) program of the US National Science Foundation (NSF).

The EISCAT (European Incoherent Scatter) Association, a consortium of six European countries and Japan, operates the newest facilities. These are 931 MHz radar with transmitting and receiving facilities in Tromsø, Norway (69.6° N, 19.2° E) and receiving sites in Kiruna, Sweden, and Sodankylä, Finland, a so-called tristatic system. The Tromsø site also has a monostatic 224 MHz system using a cylindrical paraboloidal antenna with mechanical steering in one plane and electrical pointing by means of phasing in the other plane. The newest EISCAT radar is located in the Svalbard archipelago on the island of Spitzbergen (78° N, 20° E) and has two parabolic dishes. The Institute of Ionosphere in Kharkov, Ukraine, operates a facility at 150 MHz with a 100 m fixed vertical dish and a 25 m steerable dish. The Institute of Solar-Terrestrial Physics operates a radar near Irkutsk (53° N, 103° E) in Russia (Siberia) which steers by changing the frequency between 154 and 162 MHz. The EISCAT, NSF, and Kharkov radars frequently operate together on 'World Days' that allow global studies of the ionosphere and space weather events. Originally designed for middle atmospheric studies, the MU radar near Shigaraki, Japan, has regularly been used in the ISR mode in recent years.

Elementary Description: Operation and Scattering Mechanism

Figure 1 shows the range–time diagram for an ISR spectral measurement in which the radar emits a pulse of length T that illuminates the electrons at range h_0 during a time interval T_0 , and at h_1 during a later time interval T_1 . The overlap of these time intervals is not a problem when the requirements for range resolution are modest.

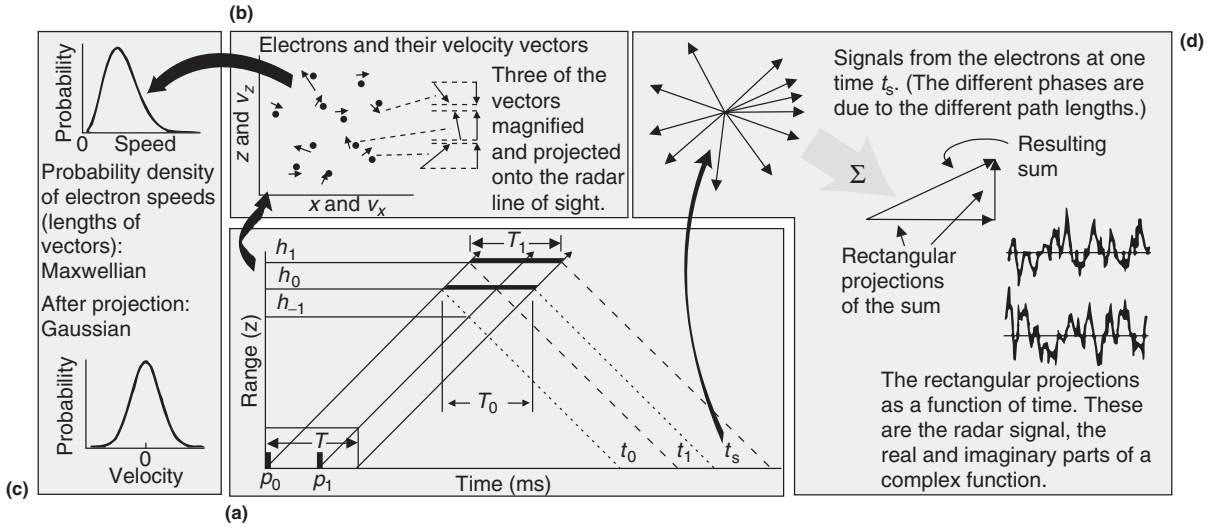


Figure 1 (a) A range–time diagram in which a pulse of length T is transmitted, illuminating electrons at range h_0 for time interval T_0 and range h_1 for time interval t_1 . The substantial overlap in time might or might not be a problem. (b) The electrons at h_1 have apparently random locations and velocities. The radar is sensitive to the component along the line of sight (z axis). (c) The Maxwellian probability density function associated with the speed of the electrons and the Gaussian function associated with the line of sight component. (d) The signal at time t_s is the sum of that from many electrons with random phase. The radar analyzes the rectangular projections of the sum. Part (a) also shows two short pulses p_0 and p_1 . The signal at each time is the sum of narrow ranges resulting from both pulses.

Nearly all of the power passes through the ionosphere, but a small amount scatters from the electrons. Consider N electrons in a small range increment around h_1 . They move with speeds distributed according to a Maxwellian distribution and with apparently random directions. The received radar signal is the sum of the very small signals scattered from all the individual electrons, and consideration of the phase path length for each electron from the transmitting antenna and back to the receiving antenna gives eqn [1] for the return signal.

$$S(t) = \sum_{i=0}^{N-1} a e^{i(\omega_0 t - k r_i(t))} \quad [1]$$

where S is the summed signal; a comes from the radar equation and depends on factors such as the power and antenna gain of the radar, the distance to the electrons, and possibly the polarization of the waves; ω_0 is the angular frequency of the radar, $k = 4\pi/\lambda$ is the scattering radar wave number for backscatter for a radar wavelength λ ; and r_i is the line of sight distance from the radar to the i -th electron. The receiver circuitry removes the rotation of the signal at the rate ω_0 so that the time variation in the signal is due to the motion of the electrons. It is apparent that eqn [1] is the spatial Fourier transform of the scattering medium along the line of sight, providing a convenient abstraction from the actual charge distribution. If there is only one electron with constant velocity along the radar line of sight, then the change in the phase of the received signal is uniform in time, and S is a sinusoid like the transmitted signal, but Doppler shifted to a new frequency.

If the electron velocities are uncorrelated, the phase angles in the sum are random, and the time-averaged power from the volume is just N times the power from a single electron, a quantity easily computed from elementary electromagnetism. The cross section of a single electron is given by $4\pi r_e^2$, where

$r_e = 2.8 \times 10^{-15}$ m is the classical electron radius, and so an electron looks like a target of 9.85×10^{-29} m. If one illuminates a cube 20 km on a side that contains an electron number density 1×10^{12} m⁻³, the total cross section is 7.9×10^{-4} m², equivalent to a sphere with a radius of 1.6 cm. Also the received radar signal is a noise-like process, and so the information about the medium is contained in the average properties of the process, the average total power, and the distribution of this power in frequency. If each electron moves with constant velocity during the time it is illuminated by the radar signal, then the spectrum is the bell-shaped Gaussian function. This follows from the Maxwellian distribution of electron speeds determined by T_e , the electron temperature, and m_e , the mass of the electron, because the velocity distribution of electrons determines the distribution of power at the corresponding Doppler shift. Actually, the electrical interaction between the ions and the electrons narrows the spectrum and determines the pulse length T required for adequate frequency resolution. Nevertheless, the Gaussian electron spectrum, as well as similar ones associated with the ions, is important in the complete analysis.

Measurement Techniques

Techniques for improving the resolution use short pulses or a modulated pulse, often with a constant transmitted amplitude, but a binary phase pattern (0° and 180°). The goal is to obtain statistical estimates of the spectrum as a function of range that are as accurate as possible and have sufficient range resolution. The signal power must be large and the receiver and sky noise small, and the number of independent samples from the noise-like signal must be large so that measurements of its

average properties are accurate. In addition to the use of a modulated pulse, sometimes transmissions occur on different frequencies at different times during the pulse, or even on several frequencies at once.

If the range resolution obtained using a pulse length of T , set by frequency resolution requirements, is not good enough, one solution is to use sequences of two or more short pulses, but then the average transmitted power decreases. It is also possible to use modulated long pulses, explained here using a simple analogy. If only the total power and not the spectrum is of interest, then obtaining the performance of a short pulse from a properly modulated long pulse by decoding the signal is easy, and a very similar method allows decoding of the set of lag profiles used in computing the spectral estimates. A lag profile is the product of the signal and a delayed version of the complex conjugate of the signal, while the set comprises lag profiles of all useful delays. This section analyzes pulse compression of the signal, the double (short) pulse technique, and the use of pulse compression on lag profiles to measure the spectrum with good range resolution.

A rectangular radar pulse of length T and unity peak power has the same total energy as one of length T/M and power M . The power spectrum of the shorter pulse contains a lot more energy in the higher frequencies. For the long pulse to give the same resolution as the short one it must have the same energy at the higher frequencies and a practical, but approximate, method to produce this is to use phase modulation, such as the Barker code (Figure 2). Then the major difference between the two pulses is the relative phases of the sinusoidal components of their spectra. Changing the phase of the spectrum of the signal corresponds to a linear filtering operation in the time domain, and the long pulse can be converted to a short pulse, or 'decoded' in this way. The received signal before decoding consists of copies of the code, one for each target, shifted by the range to the target and added. The addition of the target signals is linear, and so it can be decoded in the receiver after the target signals add, obtaining the same results as operating on each signal individually. The result is almost like using a short pulse of high peak power, as the simulation in the right-hand part of

Figure 2 shows. It used matched filter decoding, which has the best signal-to-noise ratio, but it also has small false echoes, or sidelobes, at a level of about 1% of the main echo or lobe.

The target must remain close to stationary while illuminated by the coded pulse if the decoding is to occur accurately. For lower atmosphere radars where the spectrum of the scatter is very narrow, decoding can extend over multiple pulses, allowing the use of complementary codes that have canceling sidelobes. In the ionosphere, the spectrum widens as the temperature rises and lighter ions dominate. In the topside of the ionosphere, pulse compression is less practical.

The power spectrum of a random process such as ionospheric radar signals contains the same information as its Fourier transform, the autocorrelation function (ACF). The ACF is the sum of lag products obtained by multiplying samples of the process from one time with the complex conjugates of those from later times. Just as the radar signal at a particular time is the result of scatter from a range of heights depending upon the length of the radar pulse and how it is modulated, a lag profile is made up of lag products from a similar range of heights. The set of lag profiles for all delays contains all possible information about the ACF.

Figure 3 shows the results of a simulation where the model ionospheric density increases linearly with altitude and has two very narrow, closely spaced layers of electron density enhancement, a situation that might occur in the E region. If the transmission is two short pulses spaced by three times their width, where the pulse and layer spacing correspond, as in Figure 3(a), two of the four signals from the layers are coincident, and the overlapping signal is twice as high as the others, demonstrating the simple addition of powers resulting from the lack of correlation of the scatter from different ranges. The noise rejection bandpass filter spreads the signal in height gives the waveforms pointed shapes.

Figure 3(b) shows the lag profile with a delay equal to the layer spacing, showing only one return from each layer just as required. Although Figure 1 shows that the returns at times t_0 and t_1 from short pulses p_0 and p_1 are each the sum of scatter from two heights, they have only one height in

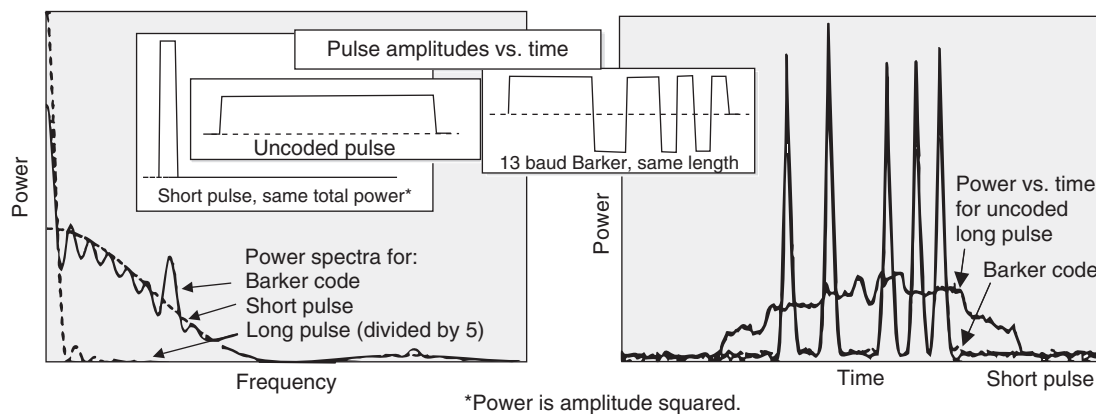


Figure 2 A radar can transmit the uncoded long pulse, but needs to transmit a pulse 13 times shorter with 13 times the peak power and the same total power. This is very expensive; instead it modulates the long pulse with the Barker code and uses signal processing techniques to get nearly the same results. The modulated pulse must have a spectrum as close as possible to that of the short pulse. The simulation at the right shows that the signal from five narrow targets looks almost the same with the short pulse and the Barker-coded pulse.

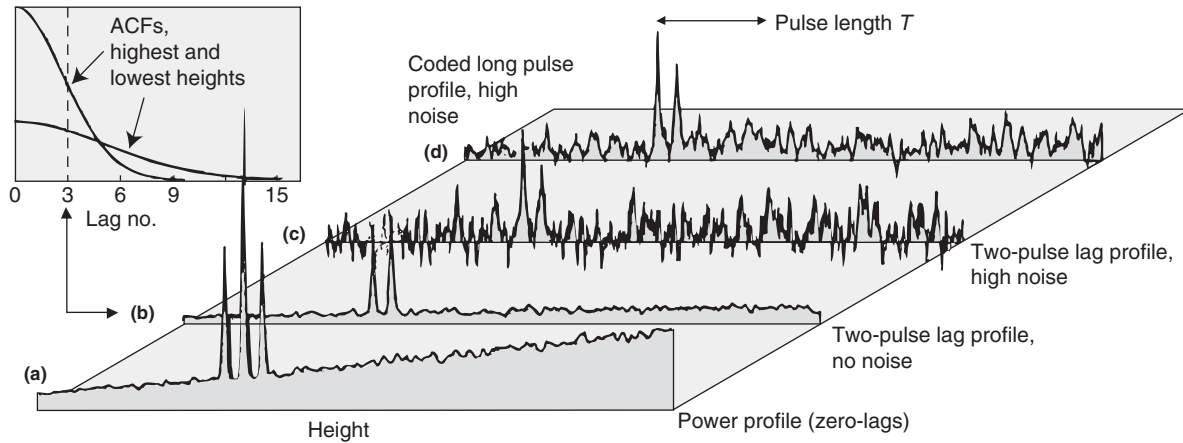


Figure 3 (a) The power profile resulting from a pair of short pulses with spacing three times their width and a pair of thin layers with corresponding spacing. (b) The lag profile for a spacing of 3; there is only a single return from each layer! This desirable result occurs because scattered signals from different heights are not correlated. (c) and (d) Comparison of the performance of the double pulse technique and the random codes technique (coded long pulse) for a case with high sky and system noise. The latter has less noise on the profile and obtains all of the different lags at once, resulting in more than 50 times better time resolution.

common. (A diagonal line downward-to-the right defines the heights and times of scattered signal entering the receiver simultaneously.) Thus the product of the two returns has only one term involving the same height, and only it has correlation. However, the terms that do not correlate still contribute random clutter, and thus increase the noise on the profile. It is possible to transmit the two pulses on orthogonal polarizations at the Jicamarca radar, eliminating the extra noise, but it is inevitable with more than two pulses. The filter used to reduce the effect of receiver and sky noise also spreads the signals in time, introducing a spread of delays into the lag profile. Shortening the pulse or correction during analysis of the data eliminates this problem.

Varying the pulse spacing allows different lag delays to be measured, giving all the necessary information about the random process. This simple technique provides excellent measurements, but it is slow because sampling of each lag occurs sequentially, and also the pair of short pulses generally uses only a small fraction of the available transmitter power. Special sequences using three or more pulses partially address both of these problems. Several ambiguity-free lag profiles occur simultaneously from these multiple pulse sequences, which have the property that each of the several delays measured by pairs of pulses in the sequence occurs only once. However, multiple pulse sequences are still sparse, even when interleaving several sequences on different frequencies, a technique used at the EISCAT facilities.

Suppose a long phase-coded pulse is transmitted where the length of a baud, the shortest section of constant phase, is equal to the short pulse and the total length is 16 Bd. The lag profile for a delay of 3 Bd contains lag products at one time covering a height range of 13 Bd, and the lag products from one height are present for a time of 13 Bd. All of the many products involving more than one height have zero average values owing to lack of correlation, ignoring the effect of the noise rejection filter. (The simulations do include this small effect.) The lag profile is a sum of lag products from all heights but all with the same delay, and this suggests that one should decode the lag

profile just like the Barker-coded signal. However, the sequence of signs given by the product of the transmitted code with itself shifted by 3 Bd determines the code, and thus the code is different for each delay. That is, if a code is present, new codes can be generated by multiplying it by itself shifted by some integral number of bauds. The ACF of each of these new codes has a central peak and sidelobes. Any code that has sidelobes nearly as low as a Barker code is a good code. Unfortunately, there is no code which generates good codes for each shift. This deficiency can be made up by using a sequence of codes such that the generated codes for each shift have sidelobes which cancel or tend to do so, like the complementary sequences described above. The sequence can be as long as needed, since only the stationarity of the ACF rather than the signal is important. One useful solution is to use a different random code on each transmitted pulse. Products of shifted random sequences are also random sequences, and so the sidelobes are random and approach zero as more pulses are used, as shown in Figure 4. Exact cancellation is not necessary since it is not possible to remove the clutter power, but only to remove its systematic effect. However, it is remarkable that short sequences of codes with no more members than twice the number of bauds can achieve perfect cancellation of the average value of lag product sidelobes. These sequences, called alternating codes, are found with very clever computer search techniques. Both these techniques – alternating codes and the random codes or coded long pulse technique – are in regular use. They give nearly identical results when used correctly, and the choice of technique depends upon practical considerations.

Figure 3(c) and 3(d) shows the improvement of these techniques over the double pulse when receiver and sky noise dominate. Part (c) shows the double pulse measurement for an accumulation of 1000 transmitted pulses. Each of the other delays requires the same number of transmissions. Part (d) shows the results for the random codes, and there are two differences affecting the relative speeds. First, all the different delays are available simultaneously with the random codes; and second, the random noise is lower than the double pulse.

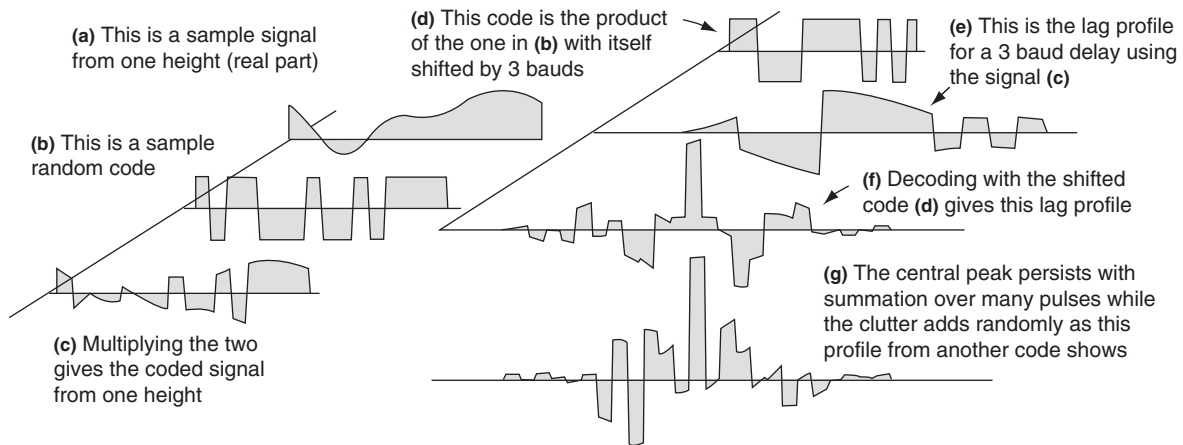


Figure 4 The various waveforms associated with the coded long pulse technique. In the receiver, the signal from one height (a) appears to be multiplied by the code (b), resulting in a modulated signal (c). If the code (b) is shifted by 3 Bd and multiplied by itself, then the resulting code (d) determines the lag profile (e). Decoding with the code (d) gives lag profile (f). The central peak is the signal that is wanted; the rest of the profile is undesirable clutter that becomes relatively small with summation over many radar pulses.

As a result of both of these factors, the random codes are more than 50 times faster than the double pulse.

It is very complicated to determine the exact relative merits of the techniques when neither clutter nor noise dominates. However, it is clear that in the clutter-dominated case they are all nearly the same, with the noted exception for the double pulse at Jicamarca.

The Incoherent Scatter Spectrum

The elementary analysis above ignored the electrostatic interaction among the various components of the plasma; the interaction between the electrons and ions is very important. **Figure 5** shows several plots computed by means of a program that includes this interaction. Part (a) shows a sequence using the Arecibo $\mathbf{k}$ vector, with the electron density n_e increasing by a factor of 2. Each plot is the ratio of the spectrum and its associated n_e , so that an invariant shape would give identical plots. The lowest, 25 cm^{-3} , gives a spectrum that is almost exactly Gaussian, but there is a very small unresolved narrow feature at zero frequency. However, as n_e increases to easily measurable values, power shifts into the narrow feature and the Gaussian shape is lost. Part (b) shows the frequency scale expanded by 100 times to resolve the narrow feature. As n_e increases, the spectral ratios become identical. It would appear that all of the power has shifted to the narrow feature, called the ion line, with constant shape; actually there is a small amount of power in the plasma line, a very narrow feature at the plasma resonance frequency (about 9 MHz for $n_e = 10^6 \text{ cm}^{-3}$). Fast electrons produced by photoionization or other causes enhance this line, making it a useful complement to the ion line.

Part (c) shows the ion line also; here, $n_e = 10^6 \text{ cm}^{-3}$ for all cases but T_e varies as T_i is held constant. As T_e/T_i changes from 1 to 8, the spectrum becomes more peaked; the location of the peak depends upon the temperature of the electrons and the mass of the ion. Part (d) shows the effect of changing the ion mass. Lighter ions result in wider spectra with similar shapes;

O^+ , He^+ , and H^+ have masses of 16, 4, and 1, respectively. With two or more ions present, the shape differs from the sum of the spectra for the single ions, particularly if the ion masses are similar.

One of the several approaches to constructing a theory containing the electron-ion interaction uses the Nyquist theorem, familiar to electrical engineers for determining the random noise level in lossy circuits but also useful for plasma fluctuations in a generalized form. The different species (electrons and each kind of ion) all have admittance functions that combine to form the spectrum of a sort of 'circuit.' The dressed particle approach calculates the effect of each particle on each of the other species, and so the total effect of a particle is the sum of its intrinsic charge and the induced charges. Finding the induced charge requires a perturbation solution of the Vlasov equation. The motion of the individual charges determines the effect of the intrinsic charge, like the simple Gaussian electron spectrum found above and by a similar method for the ions. It turns out that consideration of the induced charges gives integrals over the velocity density function that are very similar to the ones for the intrinsic charge, and so in effect these single species spectra combine in a fairly simple way to give the complete incoherent scatter spectrum. The ion line is scattered from ion acoustic waves at the thermal level, while the plasma line is scattered from Langmuir waves, possibly enhanced above the thermal level by photoelectrons or other causes. It is also possible to include the effects of the magnetic field and most types of collisions are also possible.

The Debye length is an important parameter in the equation for the spectrum. It defines a scale length that determines the behavior of the plasma. For smaller lengths, the thermal energy of the plasma is sufficient to keep the species independent, but for longer lengths the electrical interactions are strong enough to tightly couple the species. Thus the Debye length increases with temperature, but decreases with the concentration of charges, qualitatively explaining the spectra of **Figure 5(a)** and **5(b)**.

The Maxwellian velocity distribution is established by collisions between pairs of particles, either of the same or of

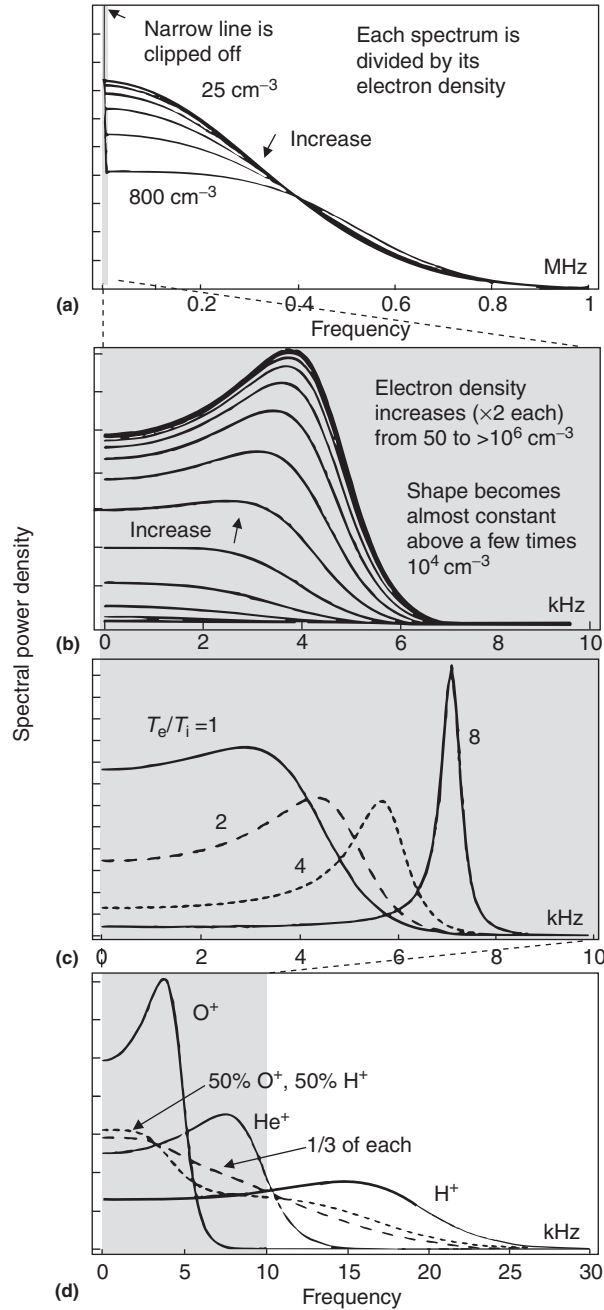


Figure 5 Various incoherent scatter spectra. Parts (a) and (b) show the ratios of the spectrum with the corresponding electron density. Part (a) shows that the wide Gaussian spectrum occurs only for very small electron densities. At higher densities that can be easily seen with a radar, the power shifts into a narrow line, resolved in (b). Part (c) shows what happens as T_e changes while $T_i = 1000$ K and $n_e = 10^6$ cm⁻³. Part (d) shows the effects of various ratios of different ions with the other parameters as in part (c) with $T_e/T_i = 1$.

different species. Collisions are so infrequent that they have no significant effect on the spectrum of the scatter in the F region for most ISRs, but this is not so at lower altitudes in the E and D regions. The effect on the single species spectrum is simple to understand in principle, but it can be difficult to compute with

good accuracy. When a particle changes direction and speed frequently in a time interval in which its progress along the radar line of sight is small, as measured by the phase change of the signal, its apparent progress is slowed. In the limit of many collisions, it is like the so-called random walk (or drunk sailor's walk) in which the progress from a starting point is proportional to the square root of time rather than linear. Thus the spectrum is narrowed. There is a second effect resulting from the sudden changes in the speed along the radar line of sight that is very much like modulating the carrier of an FM radio station. This causes the energy to spread to higher frequencies, but at a reduced power level. Figure 6(a) shows the net effect for four different values of ion-neutral collision frequency for the Arecibo case using the Bhatnagar-Gross-Krook (BGK) collision model.

Figure 6(b) shows the effect of the Earth's magnetic field using $\mathbf{k}$ vectors for the Jicamarca radar. The gyroradius for electrons is significantly smaller than the radar wavelength, and so the only effective motion is along the field lines. As the radar looks more closely to perpendicular, an electron has to move very far before it significantly affects the phase of the radar signal. This takes a long time, and so the electron single species spectrum becomes narrow. The ion spectrum normally determines the width of the incoherent scatter spectrum since it is narrower than the electron spectrum, and so as the angle moves from parallel toward perpendicular there is very little effect at first. However, eventually the electron spectrum becomes narrower than the ion spectrum

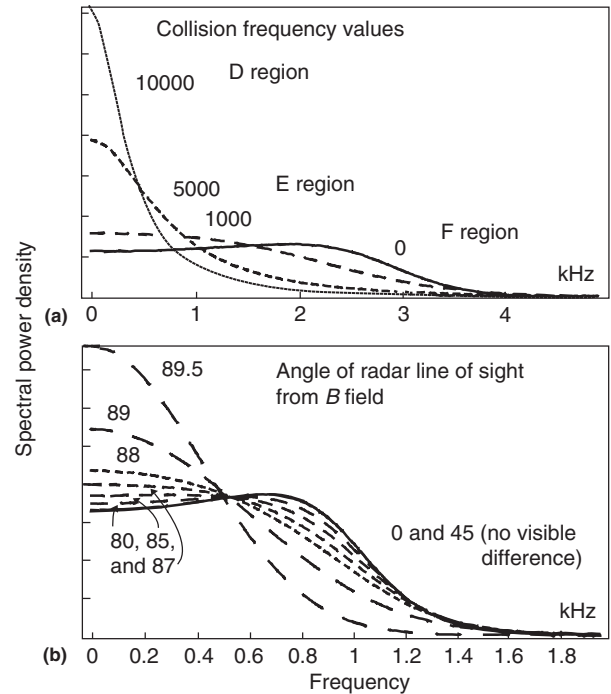


Figure 6 (a) The effects of collisions on the spectrum for four different ion-neutral collision frequencies. The plot indicates ionospheric regions where the collision frequencies are typical. Part (b) shows how the backscatter spectrum varies as the angle between the radar $\mathbf{k}$ vector and the magnetic field approaches perpendicular. These curves are only approximately correct since the effects of Coulomb collisions are not included.

and takes control as the ions lose their influence on the electrons. The fact that the electrons must move so far before the phase of the radar signal changes has another surprising effect discovered recently. The effects of Coulomb collisions on an electron due to other electrons and ions are significant, a rare case where collisions matter in the F region. These collisions narrow the spectrum by about a factor of 2 for an angle of 0.25° , and strongly affect T_e/T_i to more than 2° from perpendicular.

See also: **Dynamical Meteorology:** Atmospheric Tides; Waves. **Mesosphere:** Ionosphere; Metal Layers; Polar Summer Mesopause. **Middle Atmosphere:** Zonal Mean Climatology. **Radar:** Mesosphere–Stratosphere–Troposphere and Stratosphere–Troposphere Radars and Wind Profilers; Meteor Radar. Thermosphere.

Further Reading

- Cook, C.E., Bernfeld, M., 1993. Radar Signals, an Introduction to Theory and Application. Artech House, Norwood, MA.
- Evans, J.V., 1969. Theory and practice of incoherent scatter radar by Thomson scatter radar. *Proceedings of the IEEE* 57 (4), 496–530.
- Goldston, R.J., Rutherford, P.H., 1995. Introduction to Plasma Physics. Institute of Physics Publishing, Bristol, UK.
- Hunsucker, R.D., 1991. Radio techniques for probing the terrestrial ionosphere. In: *Physics and Chemistry in Space*, vol. 22. Springer-Verlag, Berlin.
- Kelley, M.C., 1989. The Earth's Ionosphere, Plasma Physics and Electrodynamics. Academic Press, San Diego, CA.
- Nygren, T., 1996. Introduction to Incoherent Scatter Measurements. Invers Oy, Sodankylä, Finland.
- Skolnik, M.I., 1970. Radar Handbook. McGraw-Hill, New York, NY.
- Skolnik, M.I., 1988. Radar Applications. IEEE Press, New York, NY.
- Stix, T.H., 1962. The Theory of Plasma Waves. McGraw-Hill, New York, NY.

Mesosphere–Stratosphere–Troposphere and Stratosphere–Troposphere Radars and Wind Profilers

G Vaughan, University of Manchester, Manchester, UK

D Hooper, Science & Technology Facilities Council (STFC), Didcot, UK

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by R F Woodman, volume 5, pp 1825–1833, © 2003, Elsevier Ltd.

Synopsis

Wind profiling radars measure very high frequency or ultra high frequency radiation scattered by the clear atmosphere from around 70 m to 20 km altitude in the stratosphere and troposphere, and from 70 to 90 km in the mesosphere. The scattering mechanisms include turbulence, Fresnel scattering, and scattering from free electrons. Wind profiles are typically measured with a vertical resolution of 75–300 m every 15–30 min, with accuracy of $1\text{--}2\text{ m s}^{-1}$. The signal power and spectral width of the radar echoes contain information on atmospheric layers and turbulence, respectively. Radars are used in networks and in conjunction with other instruments for many applications.

Introduction

Radar systems were originally developed to detect solid targets, so the potential of radar for detecting precipitation in the atmosphere was realized very early on. Less obvious was the potential application of radio waves scattered by the air itself – such mysterious echoes were dubbed ‘angels’ in the early days. Unlike the scattering of light waves, which results from direct interaction between the photons and individual air molecules, the scattering of radio waves depends on irregularities in the refractive index of the air on scales comparable to the radar wavelength, typically centimeters to meters. Turbulence plays a prominent role in generating such structures, allowing radar to be used for remote sensing of turbulence. Furthermore, the irregularities move with the mean flow of the air, so by measuring the Doppler shift of the scattered radiation the radar can measure the wind speed component along the beam. Combining wind speed measurements in three (or more) directions provides a measure of wind velocity, and in the continuous measurement of vertical profiles of winds this technique has found its most common application. Uniquely, these radars can continuously measure the vertical component of wind, a parameter of particular interest to meteorologists.

The use of very high frequency (VHF) (usually around 50 MHz) and ultra high frequency (UHF) (around 400–500, 915, or 1215 MHz) radars to measure vertical profiles of winds and turbulence in the atmosphere originated in the early 1970s. Mesosphere–Stratosphere–Troposphere (MST) radars were developed from instruments originally designed to study the lower ionosphere. To reach these altitudes a large radar system is required, with an antenna array typically around 10^4 m^2 and powerful radio transmitters; the Aberystwyth MST radar in the United Kingdom, shown in [Figure 1](#), is a typical example. As the technique has developed, however, its main applications have been found in operational meteorology, where the required altitude coverage is restricted to the troposphere and lowermost stratosphere (Stratosphere–Troposphere radars) or to even lower altitudes. Boundary layer profilers, measuring in the bottom 2–3 km of the atmosphere, are now commercially available from a number of manufacturers

worldwide (e.g., [Figure 2](#)). These rely on UHF frequencies, bringing down the cost of the radar and permitting readily transportable (trailer-mounted) systems. Despite the differences in size and operating frequency, however, the same underlying principles apply to all these radar systems.

Scattering Mechanisms

The refractive index of air may be approximated as follows:

$$n = 1 + 10^{-6} \left(77.6 \frac{p}{T} + 3.75 \cdot 10^5 \frac{e}{T^2} \right) \quad [1]$$

where p is the dry pressure in hPa, e is the partial pressure of water vapor in hPa, and T is the temperature in K. The term in brackets is usually termed the refractivity N .

In the lower atmosphere, up to ~ 20 km altitude, fluctuations in atmospheric density (manifest in the temperature and humidity terms) determine the scattering cross section. Above this height scattering from the neutral atmosphere is too faint to be detected, but in the lower ionosphere the much larger cross section of free electrons results in a second region of radar echoes, extending in practice from about 70 to 90 km. Mesospheric echoes have a lower altitude limit determined by electron density and an upper limit determined by viscosity, which increases with height, eventually limiting the inner scale of turbulence (the boundary between the inertial and viscous scales) to values greater than the radar wavelength (see below).

The classic interpretation of UHF and VHF radar echoes is that they arise from turbulence in the atmosphere. In the presence of a large-scale vertical gradient in refractive index, turbulence acts to bring air parcels with different temperature and humidity into close proximity, setting up a random field of fluctuations in refractive index covering the entire spectrum between the inner and outer scales of turbulence. If the outer scale (thickness of a turbulent layer) is much larger than the radar wavelength, Bragg scattering applies – the scattering cross section results entirely from the Fourier component with wave vector $2\mathbf{K}$, where $\mathbf{K}$ is the wave vector of the incident electromagnetic wave. This means the radar is sensitive to fluctuations



Figure 1 The Mesosphere–Stratosphere–Troposphere radar at Aberystwyth, Wales, UK (52.4° N, 4.01° W). This radar has operated continuously since 1997, measuring wind profiles between 2 and ~16 km. It operates at 46.5 MHz, with 400 Yagi antennas covering an area 100 × 100 m.

on the scale of half the wavelength – around 3 m in the case of VHF radars and 10 cm in the UHF case.

When turbulence fills the radar range gate, the radar reflectivity η is related to the refractive index structure constant c_n^2 by:

$$\eta = 0.38 c_n^2 \lambda^{-1/3} \quad [2]$$

where λ is the radar wavelength. c_n^2 may be further expressed as:

$$c_n^2 = a \varepsilon^{-1/3} K M^2 \quad [3]$$

where a is a constant (approximately equal to 3.3), ε is the turbulent energy dissipation rate (in W m^{-3}), K is the eddy

diffusion coefficient, and M is the potential refractive index gradient, derived using quantities (potential temperature θ and specific humidity q in kg kg^{-1}) that are conserved in adiabatic vertical displacement (see Vaughan and Worthington, 2000 for further details).

$$M = -77.6 \cdot 10^{-6} \frac{p}{T} \left(\frac{\partial \ln \theta}{\partial z} \left[1 + \frac{15\,500\,q}{T} \right] - \frac{7750}{T} \frac{\partial q}{\partial z} \right) \quad [4]$$

In this model the received radar power P becomes $\text{Pr}^2 \propto c_n^2 \lambda^{-1/3}$, with r^2 accounting for the inverse square law in radiation intensity from scattering (r is the range), and the



Figure 2 UHF radar (1.29 GHz) built by Degreane and operated by the UK's Atmospheric Measurement Facility. It has three panels, pointing to the vertical and 15° off vertical, and is mounted on a trailer, which can be towed to different locations for field campaigns. The UK Mesosphere–Stratosphere–Troposphere radar is shown in the background.

constant of proportionality depending on the radar transmitted power, antenna area and gain, receiver efficiency, and assumptions made about the turbulent scattering processes.

Such is the conceptual simplicity of the Bragg scattering theory that many radar studies, particularly at UHF wavelengths, assume that the echo power is a direct measure of turbulent intensity. Early in the development of VHF radars, however, it became obvious that this could not explain all the observations. Almost all VHF echo profiles were found to be at least mildly aspect sensitive, in other words the echo power decreased rapidly as the radar beam was directed away from the zenith. Furthermore, the spectral width of the echo power was very narrow. To explain this, the theory of Fresnel scattering was proposed. Here, the vertical profile of refractive index is considered to comprise ‘sheets,’ i.e., step changes over vertical distances commensurate with the radar wavelength but horizontally coherent over the radar beam width. If these sheets are very numerous, and distributed randomly in the vertical, Bragg scattering will again apply in that the Fourier component matching $2K$ will be preferentially scattered. In this case, the radar reflectivity η is given by:

$$\eta = (\delta r)^2 \frac{F(\lambda)^2}{16} M^2 \quad [5]$$

where δr is the range resolution of the radar and $F(\lambda)$ is a function relating refractive index variations on the small-scale to the

large-scale gradient measurable, for instance, by a radiosonde. Unlike echoes from turbulence, these echoes are specular because of the sheetlike nature of the scattering irregularities. Evidence for the existence of sharp gradients in refractive index on small vertical scales has come from balloon flights with specially designed temperature sensors, and a correspondence has been shown between the distribution of such steps and echo power from a nearby VHF radar. However, the horizontal extent of the sharp gradients is not known, and the nature of specular reflection of electromagnetic radiation remains controversial.

These two models of reflectivity are best considered as limiting cases, with the actual mechanism at any time depending on the state of the atmosphere and the radar wavelength. At VHF, specular echoes predominate so that the Fresnel model is most appropriate in most conditions. However, the echoes are rarely purely specular, and the degree of anisotropy (rate at which echo power decreases with zenith angle) is highly variable. Models of anisotropic turbulence have been proposed to explain this phenomenon, but so far none of them challenges the conceptual simplicity of the two described above. At UHF wavelengths, particularly in the convective daytime boundary layer, turbulent (isotropic) scatter dominates and echo power is related to turbulent intensity – but UHF profilers measure almost as effectively at night as in the daytime, depicting in particular residual layers of turbulence as shown in Figure 3.

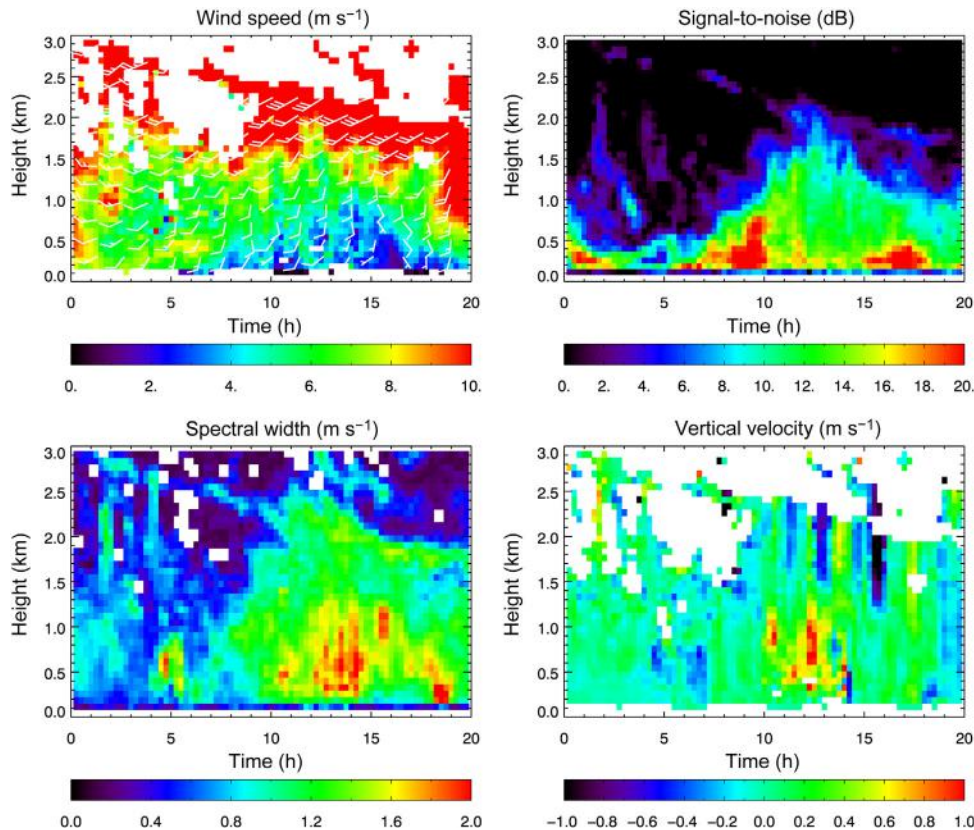


Figure 3 Example of measurements from the boundary-layer profiler shown in Figure 2, located at Achern in Germany on 24 June 2007. Note the growth of the convective boundary layer during the morning, shown both in signal-to-noise ratio and spectral width, and the residual layer between 3 and 2.5 km at night.

Both VHF and UHF wind profilers detect scattering from precipitation, but at VHF this echo seldom exceeds the clear-air backscatter. The reverse is the case at UHF, where because of the shorter wavelength Rayleigh scattering from droplets and ice particles is much stronger. As with conventional rain radars, a bright band may be observed at the melting level at UHF. Vertical velocities measured by these radars during precipitation events are those of the falling particles – but because the particles are moved horizontally by the wind, the derived horizontal velocities match the wind components. Indeed, this can present an advantage by extending the vertical coverage of the radar. In clear air, a UHF profiler seldom measures above ~ 3 km; during the passage of a front, for example, winds may be measured up to 10 km or even higher. At VHF, heavy convective rain events can (because of the size of the droplets) provide both a clear air and a particle peak in the Doppler spectrum, both of which can be measured separately if required.

In contrast to UHF radars, where precipitation always enhances the echo power, stratiform or frontal rain has the effect of reducing VHF echo power (e.g., Figure 4). In such events, the droplets are too small to be observed, and the reduction in power most probably occurs because the falling raindrops ‘wash out’ structure in refractive index on the scale of the radar wavelength. The function F in eqn [5] linking M

measured on the scale of a radar range gate and the smaller-scale fluctuations that scatter the radio waves is clearly different in clear and precipitating air.

Whereas suitable fluctuations in refractive index in the troposphere and lower stratosphere are present throughout the day (permitting continuous wind measurements), those in the mesosphere are more sporadic, since free electrons in this part of the atmosphere generally occur in daytime and the echoes are patchy in space and time.

Wind Measurements

The most common method for measuring winds with radar is the Doppler beam swinging technique. With this technique, the radar beam is emitted and received by the same antenna, and is sequentially directed vertically and at off-vertical angles close to the zenith (e.g., at four orthogonal azimuths directed at 5 – 15° to the zenith). By measuring the Doppler shift of the received echo relative to that transmitted, and combining the velocities derived along the different radial directions, a profile of the three components of wind may be constructed. Typical dwell times along each radial direction are 10 – 20 s, so a profile may take several minutes to measure. The technique therefore implicitly assumes both horizontal uniformity of the wind

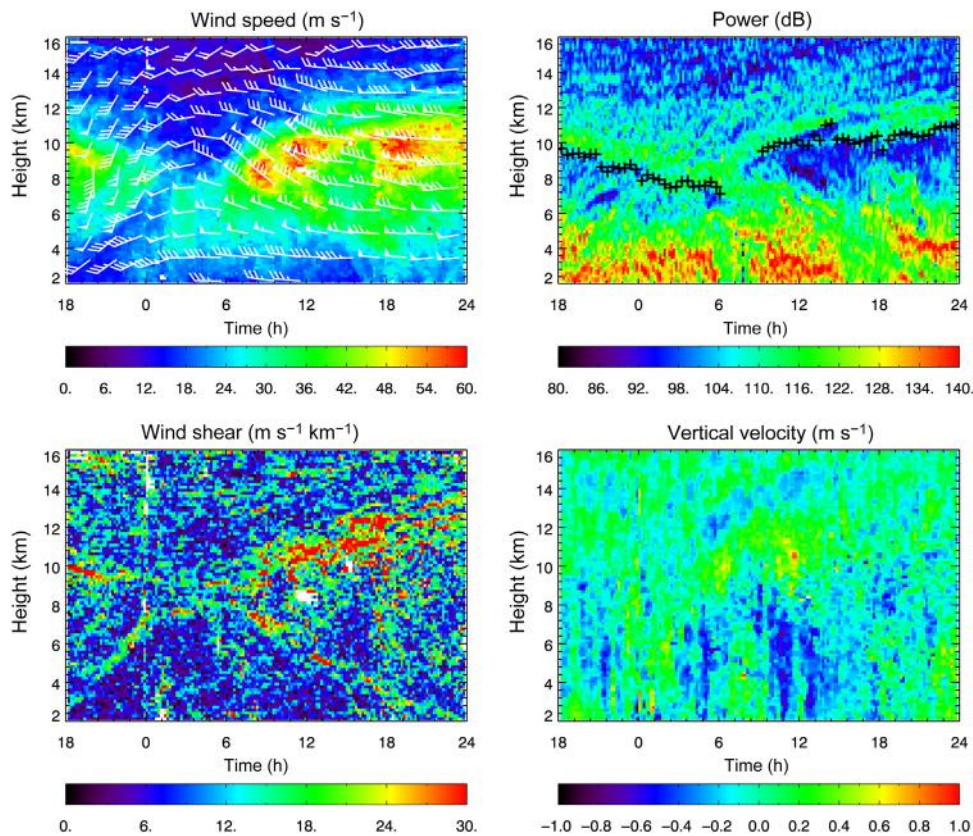


Figure 4 Example of measurements from the VHF radar at Aberystwyth for the period of 18 h on 4 September 2011 to 2359 h on 5 September 2011, showing: (i) passage of a cold front between 18 and 24 on 4th (ascending layer of high wind shear), (ii) passage of warm front on the 5th (descending layer of high wind shear), (iii) mountain waves (large values of vertical velocity lasting several hours), (iv) convection (short-lived maxima in vertical velocity near 0 h), (v) the tropopause (black crosses), (vi) suppression of radar power by rain (6–8 and 16–19 h, below 6 km).

within the volume containing the various beams, and stationarity in time over the measurement cycle. In the troposphere and lower stratosphere, the horizontal scale corresponds to a few kilometers, and in the mesosphere to a few tens of kilometers. This is generally sufficient for meteorological measurements but for mesospheric applications an alternative technique, called spaced antenna drift, is sometimes used. Here, one transmitted beam is coupled to a number (three or more) of receiving antennas, and wind velocities calculated through correlation analysis of the measured signals. Further discussion of this technique (which has also been used occasionally for tropospheric and lower stratospheric radars) may be found in [Hocking \(2011\)](#).

The distinguishing physical feature of wind profiling radars compared to the popular perception of radar is the absence of a steerable dish. The long wavelengths needed to receive clear-air echoes preclude mechanical steering other than in exceptional cases, so the basic construction becomes an array of antennas to which the electromagnetic radiation can be fed from one or more transmitters. The array shown in [Figure 1](#), for example, comprises 400 Yagi antennas over an area 100×100 m, which are fed with five transmitters located in the building next to it. The typical peak power is 160 kW with a two-way half-power beam width of 2.2° at 46.5 MHz. For a VHF radar of this type, beam steering is achieved by changing the phase of the transmitted wave across the array. In this case, the phase is changed by switching lengths of cable into the transmitter line, but a more flexible arrangement is to use distributed (lower power) transmitters, each serving a few antennas, where the phase can be controlled in software. The large array, and the need to protect the sensitive receiver from the transmitted pulse, means there is a dead time in measurement with these systems corresponding to around 1–2 km altitude; wind measurements are not possible at altitudes much lower than this.

For UHF radars of gigahertz frequencies and above, a much smaller array is needed to achieve the same beam width and the dead time between transmission and receiving is less; wind measurements are possible above around 70 m with these radars. Phase switching may be used to direct the beam as at VHF, or a different approach may be used – fixed panels containing arrays of antenna elements directed vertically and at orthogonal azimuths off vertical. An example of such a system is shown in [Figure 2](#). Otherwise the basic principles of UHF and VHF radars are very similar.

An ideal wind profiler will have high vertical and temporal resolution and high accuracy. In practice, a number of factors limit these qualities, and there are trade-offs between them. Vertical resolution is limited by the duration of the transmitted pulse, which in turn is determined by the radio bandwidth available to the radar. Regulatory authorities are reluctant to issue licenses for wide-bandwidth transmitters, so the pulse length normally used is around 1 μ s, corresponding to a vertical range of 150 m (at UHF shorter pulses can be used, giving around half this vertical range). In radar terms, resolution refers to the ability to distinguish two hard targets a certain distance apart, which is not applicable for a continuous medium, so this vertical range is often used as a measure of the vertical resolution. It should be kept in mind that the actual resolution will depend both on atmospheric conditions and on the application defining that resolution.

A device commonly used to improve signal-to-noise ratio while preserving range resolution is pulse coding. Here, a longer pulse is emitted, with the phase of the electromagnetic wave being flipped by 180° at (say) 1 μ s intervals. Complementary codes have been devised, which allow the full range resolution of the radar to be recovered by decoding this information in the receiver while preserving the power advantage of the longer pulse.

To derive a radial wind value in a given range gate from a wind profiler, the basic method is to measure the echo signal in phase and in quadrature with the local oscillator at time intervals $\Delta\tau$. A time series of these measurements is then converted into a spectrum using Fourier analysis, and the first moment of the spectrum used to measure the Doppler shift (the complex signal values are needed to allow the Fourier analysis to distinguish motion toward and away from the radar). Clearly the Nyquist frequency, determining the maximum unambiguously measurable radial wind velocity, $V_{\max}$, is $f_N = 1/2\Delta\tau$ cycles per second. Since the Doppler shift is $2V/\lambda$ for radial wind speed V , this corresponds to $V_{\max} = \lambda/4\Delta\tau$. A VHF radio transmitter is typically operated at a pulse repetition frequency of around 1 kHz; any higher and there is the possibility of echoes from one pulse being contaminated with high-altitude echoes from the previous one. With $\lambda = 6$ m, $\Delta\tau = 1$ ms corresponds to $V_{\max} = 1500$ m s $^{-1}$ – far in excess of any conceivable atmospheric velocity. This provides an opportunity to use coherent averaging – averaging over tens to hundreds of pulses – to improve signal-to-noise ratio and provide values of $\Delta\tau$ closer to 0.3 s likely to be needed in practice.

The length of the Fourier transform determines the resolution of the derived spectrum (again by the Nyquist theorem), which is one constraint on its accuracy. A longer time series provides a better resolved spectrum – but this introduces another error, arising from the assumption of stationarity. Measuring a sample of signal returned over ~ 10 s (a typical value used in practice) and deriving a single spectrum from it assumes that the atmosphere above the radar does not change its properties during this time. Of course this is not the case in practice, and the choice of operating parameters is a compromise between the desire for resolution and good signal-to-noise ratio, and the desire to measure a representative wind.

At UHF, the value of $\Delta\tau$ needed to keep $V_{\max}$ to a reasonable value is much smaller, and much shorter time series are needed for the same velocity resolution. Pulse repetition frequencies can also be higher since echoes from the high atmosphere may be neglected. Here, incoherent averaging is often employed to improve signal-to-noise ratio – averaging of a number of successive spectra before calculating the spectral parameters. Also common in UHF radars is the use of different modes – short pulses for detailed measurements at low levels and longer pulses to extend the vertical coverage.

Once the spectrum is calculated, the derivation of atmospheric parameters is in principle simple. Given a bell-shaped spectral feature occupying, say, a fifth of the measured spectrum, its integral (zeroth moment) gives the echo power, its first moment the radial velocity and its second moment the spectral width – after first subtracting the noise power derived from the remaining four-fifths of the spectrum. In practice matters are rarely this simple. Ground clutter, for example,

affects all radars to some extent so that points around zero Doppler shift must be removed from the analysis – thus rendering the radar unable to measure very light winds. For UHF radars the problem is much worse – these receive echoes from birds, swaying power lines, and other moving targets that contaminate spectral values well away from zero. At least the radar echoes from well-developed boundary layer turbulence are close to Gaussian in shape, and amenable to interpretation as a mean velocity with some variance around it. VHF echoes from the free atmosphere often look nothing like a Gaussian spectrum, frequently being multimodal.

For this reason useful winds from both VHF and UHF wind profilers involve some degree of averaging. Although a raw wind measurement may be made every 1–2 min, reliable winds are usually provided every 15–30 min. At UHF, consensus averaging algorithms are used, to eliminate the frequent wild outliers derived from the raw spectra. At VHF, simple averaging of the winds is usually enough. With this proviso, the accuracy of radar wind profiles when compared to near-coincident radiosondes is found to be in the range 1–2 m s⁻¹. Systematic errors also affect the measurements, particularly at VHF where the anisotropy of the echoes from the atmosphere effectively draws an off-vertical beam closer to the zenith than its nominal position. Corrections are normally made for such errors, but uncertainties of a few percent remain. For example, regular comparisons are made at the UK Met Office between wind profiles assimilated into the operational forecast model and the background field, both as a bias and root mean square (rms) difference. They find that radiosondes and the Aberystwyth MST radar wind profiles have almost identical biases (<1 m s⁻¹) and rms differences (2–3 m s⁻¹) with respect to the model.

Signal Power and Turbulence

Equations [2] and [5] both contain the term M^2 – whatever the model, the radar echo power is proportional to the square of the potential refractivity gradient. Therefore, there is a close relationship between the measured echo power and sharp gradients of humidity and potential temperature (eqn [4]). Indeed, eqn [4] shows that these two terms are of opposite sign – so that maximum power is expected for layers where $\partial\theta/\partial z$ is positive and $\partial q/\partial z$ is negative. These conditions are routinely met at the tropopause and the boundary layer top, both of which can be accurately identified in VHF and UHF profiles, respectively, provided the atmospheric feature is itself well defined. Examples of each are shown in Figures 3 and 4, respectively. Detailed comparisons of radar power with M^2 derived from radiosonde profiles show that the relationship is qualitative rather than quantitative with a great deal of scatter around the underlying correlation. Nevertheless, it remains the case that the information content of VHF echo power profiles is yet to be fully exploited.

At UHF, the echo power during daytime clearly identifies the mixed layer, with its characteristic growth and decay (Figure 3). Given the unambiguous turbulent scattering filling the radar echo volume, it is possible therefore to derive c_n^2 from the power, if the radar system constants are known (i.e., the radar is calibrated). However, in most cases wind profiling

radars are not calibrated since this is not necessary for the wind measurements that are their primary product. Quantitative information on turbulence is therefore derived from the spectral width: the second moment of the spectrum.

Turbulence within a volume of air creates a spread of instantaneous velocities, the magnitude of which depends on the strength of the turbulence. This is detected by the radar as a range of Doppler velocities – a broad echo spectrum. Equations have been derived linking the turbulent energy dissipation rate ε and the eddy diffusion coefficient K to the radial velocity variance measured by the radar, $\overline{V'^2}$:

$$\varepsilon = f_1 \overline{V'^2} \omega_b \quad [6]$$

$$K = \frac{f_2 \varepsilon}{\omega_b^2} \quad [7]$$

where ω_b is the Brunt–Vaisala frequency and f_1 and f_2 are constants. There is considerable debate about the values of these constants; Hocking (2011) recommends $f_1 = 0.27$ and $f_2 = 0.25$.

Unfortunately, turbulence is not the only effect that can broaden a radar echo spectrum, the most prominent competing effects being beam broadening and shear broadening. Beam broadening is a result of the finite angular width of a radar beam; for a nominally vertical beam, for example, the edges will extend a degree or more away from the vertical. A horizontal wind in these circumstances will have a component in the radial direction at the edge of the beam, causing a broadening of the measured (vertical) Doppler spectrum. Shear broadening is more direct: a gradient in the horizontal velocity along a beam produces a range of radial wind components within the scattering volume unrelated to any turbulence. Methods have been devised to evaluate and remove these effects, but since they are usually the dominant source of broadening such methods are only partially successful. The exception is in the daytime boundary layer, where the strong and generally isotropic turbulence means that it dominates the spectra. At VHF, turbulence can only normally be detected in deep convective clouds or in thin layers of shear-driven turbulence, most commonly found around the jet stream.

Mesospheric Measurements

Above about 25 km, the echo power from the molecular atmosphere becomes too weak to detect even with the most powerful radars. This remains the case throughout the stratosphere, until the lower altitudes of the ionosphere are reached. Here, scattering from electrons provides an echo allowing VHF radars in particular to be used to study the altitude region between 70 and 90 km. The basic technique is no different to that already described, but particular features of the mesospheric echoes need to be taken into account.

Of particular importance for mesospheric studies is the availability of turbulent fluctuations in electron density. The inner scale of turbulence is around 5 m at 80 km, which means that UHF wavelengths are well into the viscous subrange in the mesosphere and should not be strongly scattered. At VHF, however, the radar measurement scale of ~3 m can lie in the

inertial subrange and so provide measurable scatter. Between 70 and 80 km, sporadic VHF echoes can be measured from the sunlit atmosphere throughout the year – free electrons in this part of the atmosphere are produced by short-wavelength solar illumination and do not persist into the hours of darkness. Wind measurements derived from these echoes enable studies of gravity waves, planetary waves, and tides that dominate the dynamics of the mesosphere.

Considerations of the inner scale of turbulence suggest that VHF echoes should not be measured much above 80 km. By contrast this is where the strongest mesospheric echoes by far are found! These are the polar mesospheric summer echoes (PMSEs), which occur at high latitudes around midsummer. At this time of year, the mesopause is at its coldest and indeed the polar summer mesopause can reach 140 K, easily the coldest region of the Earth's atmosphere. These cold temperatures result in noctilucent clouds of tiny ice particles that, like PMSEs, can extend to temperate latitudes around the solstice. Observations of PMSEs show that the echoes are highly variable, both in altitude and in time, and although generally coexisting with noctilucent clouds they are not exactly correlated with them. Typically, they are found as thin layers of high echo power between 80 and 90 km (e.g., Figure 5), most prominent at midday. Several theories for producing PMSEs have been advanced but this is still an area of active research.

Applications of Wind Profilers

One of the particular strengths of radar wind profilers is their ability to measure the vertical component of the wind. This finds its most fruitful application in studies of convection and atmospheric gravity waves: small-scale processes with vertical velocities of order 1 m s^{-1} or greater. By contrast, synoptic-scale vertical velocities, typically of order 1 cm s^{-1} , are very difficult to measure with wind profilers – for example, given a maximum radial velocity of 5 m s^{-1} ($\Delta\tau = 0.3 \text{ s}$), a 512-point Fourier transform would be needed just to provide a raw resolution of 1 cm s^{-1} , requiring a dwell time of around 2.5 min – far longer than the 10–20 s normally used. Even with adequate resolution, very long averaging times ($\sim 10 \text{ h}$) are required to remove the smaller-scale features, and experience

shows that systematic biases arise, which render the final result questionable.

Gravity Waves

Gravity waves are ubiquitous in the free atmosphere where the static stability is generally positive, and the full spectrum of gravity wave motions, from the Brunt–Väisälä period to the inertial period f_i , may be measured with a VHF radar. Of particular importance are mountain waves – standing gravity wave patterns forced by topography, and occurring in the majority of synoptic environments even over small hills. Air flowing through these patterns oscillates vertically at a frequency at or close to ω_b , but the pattern itself changes much more slowly – with a typical timescale of several hours as the synoptic environment itself changes. Consequently, a VHF radar fixed in one location will measure a pattern of vertical motion that changes over this timescale – which must not be confused with genuine synoptic-scale vertical motion. VHF radar measurements of mountain waves are valuable for testing numerical models of wave propagation and for identifying layers where the waves break – typically at altitudes where the wind component along the wave propagation vector is zero (a so-called critical layer). Enhanced spectral width in such layers is a measure of turbulence as mentioned above.

At the other end of the gravity wave spectrum are inertia–gravity waves – waves where the Coriolis acceleration causes individual air parcels to trace out ellipses in the plane of propagation of the wave. Because of the low frequency, these planes are almost horizontal, and inertia–gravity waves are detected using the horizontal wind components rather than the vertical. These waves are very common in midlatitudes, especially in the lower stratosphere where maxima and minima in wind components descend over time, and may be tracked over many hours (or even days in extreme cases). They are associated with jet streams, suggesting that the process of geostrophic adjustment is important in their formation. The temporal continuity of wind profiler measurements provides the most unambiguous way of measuring this poorly understood class of gravity wave.

The ability of wind profilers to measure coincident horizontal and vertical wind components provides a way to measure momentum fluxes from gravity waves, which are

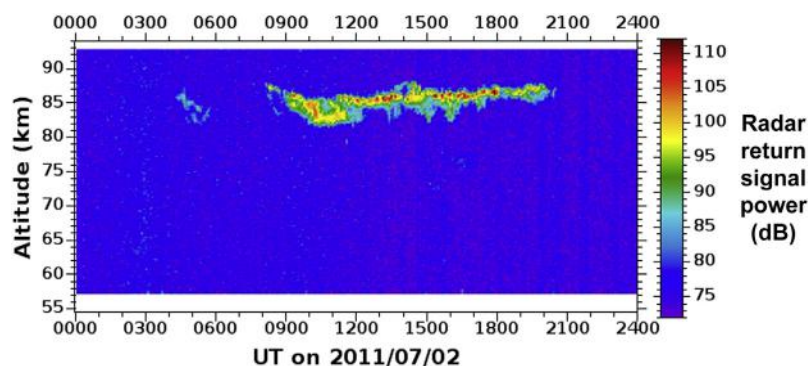


Figure 5 Example of mesospheric summer echo observed in midlatitudes by the Aberystwyth MST radar. Courtesy of the NERC MST Radar Facility. MST, Mesosphere–Stratosphere–Troposphere.

required for parameterization of the contribution of gravity waves to the momentum budget, particularly in the stratosphere and mesosphere.

Convection

Convection is particularly evident in data from boundary layer profilers, where the growth and decay of the mixed layer during daytime may be clearly observed in echo power, vertical velocity, and spectral width (e.g., Figure 3). With a typical time resolution of 15 min or more, the radars do not resolve individual thermals, but in unstable environments where deeper convection forms, the resulting plumes are clearly visible. If convection is sufficiently deep to result in precipitation, the strong particle echoes dominate the weaker clear-air scattering, as described above.

Deep convection is also clearly visible in VHF radar echoes, with similar characteristics to the UHF echoes except that particle scattering is only appreciable for very large storms. Typically, deep convection reduces the ability of a VHF profiler to measure winds because of the loss of horizontal homogeneity in the wind field.

Atmospheric Layers

The use of radar echo power to identify the boundary layer top and the tropopause has already been introduced in Section [Signal Power and Turbulence](#), but radars can be used to detect other layered structures as well. Fronts, for example, show up clearly as sloping layers (with time) of enhanced wind shear beneath a jet stream, often with maxima in echo power as well. Indeed, tropopause folds, which are upper-level fronts containing a layer of stratospheric air sandwiched between layers of tropospheric air, are particularly clearly observed – the echo power maximum in this case corresponding to the enhanced stability and negative humidity gradient at the base of the fold (eqn [4]). In the stratosphere, layers of turbulence may be detected above a jet stream or in regions of gravity wave breaking – demonstrating the sporadic and local nature of turbulent mixing in this part of the atmosphere.

Radar Networks

Individually, wind profilers can depict in great detail the dynamics of the atmosphere at one location, but do not reveal horizontal variations unless further assumptions are made, for example, about the speed of passage of a front. To address this deficiency, a network of profilers is required. Such networks have been established by meteorological agencies in a number of countries, for example, that of the National Oceanic and Atmospheric Administration (NOAA) in the United States, Japan Meteorological Agency in Japan, and the E-WINPROF network in Europe. Wind profiles are readily assimilated into numerical weather forecasts, and the continuous coverage available from wind profilers makes them particularly suitable for assimilation into mesoscale models: with grid spacing of a few kilometers the detail in these models far outstrips that of the conventional observing network. Most network radars are of the UHF type – both because of cost and because the observational requirement is most acute in the boundary

layer – but VHF radars of varying size and complexity are also included (e.g., the Aberystwyth MST radar).

In the mesosphere, networks of radars can investigate the propagation of planetary waves and tides – disturbances with horizontal scales of thousands of kilometers. MST radars are used here in conjunction with other radar types, for example, meteor radars or medium frequency radars.

Conclusions

Wind profiling radar is by now an established technology, which has moved from specialist research applications to operational observations and from single instruments to coordinated networks of radars. Profilers are manufactured commercially by a number of companies worldwide, for example, Scintec and Degreane at UHF and ATRAD and Genesis at VHF, although the most powerful systems capable of mesospheric observations remain in the research domain. The great strength of wind profilers is the reliability of the technology – operational coverage of 98% over the course of a year is possible, and measurements are not affected by the presence of cloud or degraded by precipitation. The radars can operate autonomously and the data processed and distributed in real time – an essential attribute for operational meteorology. At the same time the technique is flexible and can be tailored to specific research applications – for example, the use of frequency-domain interferometry for enhanced vertical resolution, or data recording at high time resolution for PMSE observations.

Wind measurements from profilers are well characterized at stratospheric and tropospheric altitudes, but the scattering mechanisms, especially at VHF, are still only partially understood – in particular the nature of the predominant aspect-sensitive echoes. For this reason, echo power profiles at VHF are not as well exploited as at UHF where active turbulence dominates. Spectral width, corrected for various factors like beam broadening and shear broadening, is generally accepted to provide a measure of turbulent intensity for strong turbulence, but further work is needed to establish the validity of such expressions as eqns [6] and [7]. In the mesosphere, the scattering mechanism for PMSEs is still the subject of debate, and there have been fewer opportunities to validate MST radar winds by comparison with other techniques.

Novel applications of wind profilers come when they are combined with other techniques. The most obvious is radio acoustic sounding, a combination of wind profiler and sodar, which is able to measure virtual temperature profiles in the lowest few kilometers of the atmosphere (the radar measures the speed of the sound wave, providing a direct measure of temperature). Many of the NOAA network radars, for example, have this capability. Lidar is another active remote sensing technique, which complements a wind profiler – for example, in measurements of cloud base in operational meteorology, or stratospheric temperature profiles for gravity wave studies. Airglow cameras provide an all-sky perspective on mesospheric gravity waves, and complementary use of other mesospheric radars has already been noted. Further developments of wind profiler applications are under way in profiling humidity (using microwave radiometers) and measuring cloud dynamics (in conjunction with vertically pointing cloud radars).

This article has provided an overview of the technique of wind profiling and a brief summary of the major applications. For further information, the reader is advised to consult the references provided below or up-to-date reference books on radar meteorology.

See also: Radar: Cloud Radar; Precipitation Radar.

Further Reading

- Doviak, R.J., Zrnic, D.S., 1993. *Doppler Radar and Weather Observations*, second ed. Academic Press, San Diego, California.
- Gage, K.S., Balsley, B.B., Green, J.S., 1981. Fresnel scattering model for the specular echoes observed by VHF radar. *Radio Sciences* 16, 1447–1453.
- Hocking, W.K., 2011. A review of mesosphere–stratosphere–troposphere (MST) radar developments and studies, circa 1997–2008. *Journal of Atmospheric and Solar-Terrestrial Physics* 73, 848–882.
- Hooper, D.A., Arvilius, J., Stebel, K., 2004. Retrieval of atmospheric static stability from MST radar return signal power. *Annals of Geophysics* 22, 3781–3788.

- Lawson, J., Vaughan, G., Schultz, D.M., 2011. Classifying fronts in data from a VHF wind-profiling radar. *Atmospheric Science Letters* 12, 375–380.
- Luce, H., Crochet, M., Dalaudier, F., 2001. Temperature sheets and aspect-sensitive radar echoes. *Annals of Geophysics* 19, 899–920.
- Ottersten, H., 1969. Radio backscattering from the turbulent clear atmosphere. *Radio Sciences* 4, 1251–1255.
- Vaughan, G., Worthington, R.M., 2000. Effects of humidity and precipitation on VHF radar vertical beam echoes. *Radio Sciences* 35, 1389–1398.
- Vaughan, G., Worthington, R.M., 2007. Inertia-gravity waves observed by the UK MST radar. *Quarterly Journal of the Royal Meteorological Society of London* 133, 179–188.

Relevant Websites

- <http://www.metoffice.gov.uk/science/specialist/cwinde/profiler/> – E-WINPROF network of wind profilers.
- <http://www.jma.go.jp/jma/en/Activities/observations.html> – JMA network of wind profilers.
- <http://www.profiler.noaa.gov/npn/> – NOAA network of wind profilers.
- <http://mst.nerc.ac.uk/> – UK MST radar.

Meteor Radar

NJ Mitchell, The University of Bath, Bath, UK

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

Meteor radars detect the ionized trails left by meteors in the mesosphere and lower thermosphere. The trails are advected with the winds at the heights where they form and so can be used as tracers of atmospheric motion. Meteor radars typically measure zonal and meridional winds with time and height resolutions of about 1 h and 3 km, respectively. The radars are often simple and robust, making them ideal for extended deployment at remote sites. They are well suited to studies of background winds, tides, and planetary waves. Recent advances offer the possibility of making statistical measurements of gravity wave fluxes.

Introduction

Meteors, colloquially ‘shooting stars,’ can be among the most beautiful and striking phenomena of the naked-eye night sky. Measuring radio waves reflected from the ionized trails that meteors create has proved a powerful tool in investigations of meteors and the region of the atmosphere in which they occur. Meteor trails are carried by the winds at the height where they form, and so High-Frequency (HF) and Very-High-Frequency (VHF) radar Doppler measurements of the drift velocities of radio meteors allow determination of these winds – essentially using the meteors as tracers of atmospheric motion. Most radio meteor echoes are detected at heights between ~ 70 and 110 km. This is the upper part of the middle atmosphere and spans the upper mesosphere, the mesopause, and the lower thermosphere. These regions are otherwise notoriously difficult to investigate experimentally because they are too high for *in situ* measurements other than those made by rockets.

Meteor radars have thus played an important role in elucidating many aspects of the dynamics of this part of the atmosphere, particularly with regard to the mean winds, tides, and planetary waves that dominate the flow at these heights. Many types of radar can detect meteors, but here we will concentrate on purpose-built meteor radars used to investigate the dynamics and structure of the mesosphere and lower thermosphere. Particular attention will be given to the nature of the radio reflections from meteors and to illustrating the capabilities, strengths, and limitations of these radars for studying the dynamics of the meteor region of the atmosphere.

Meteor Radar in Context

Radio studies of meteors prior to the 1970s were mostly concerned with various aspects of meteor astronomy and with understanding the physical mechanisms responsible for the observed properties of meteor echoes. Once the basic physical mechanisms had been uncovered, it was realized that it is possible to use the drifting of radio meteors as tracers of atmospheric motion at meteor heights. Early successes included some of the first measurements of the general circulation of the mesosphere and lower thermosphere and the identification of various large amplitude tides and planetary wave modes that are major features of the dynamics of the atmosphere at these heights. Coordinated studies with

instruments at different locations allowed determination of the zonal and meridional structure of some of these winds, tides, and planetary waves.

Significant early work was often achieved even by simple radars operating without any ability to determine echo heights and so restricted to measuring winds representing a mean across the depth of the meteor region, weighted by the vertical distribution of meteor echoes (see below).

In the 1980s and early 1990s, the development of other techniques, particularly medium frequency (MF, or partial reflection) radar, mesosphere–stratosphere–troposphere (MST) radar, incoherent scatter (IS) radar, and resonance and Rayleigh backscatter lidar, to some extent overshadowed meteor radar as a tool for atmospheric studies.

In the twenty-first century, however, there has been something of a renaissance in meteor radar work. This is partly because of the development of new radio meteor techniques, and partly because sophisticated, commercially produced, purpose-built meteor radars are now available at relatively modest cost. Such systems offer accurate echo height finding and are capable of continuous operation over periods of many years with minimal maintenance, making them well suited to operation at remote sites. These factors have combined to greatly increase the number of meteor radars deployed. Presently, about 35 dedicated meteor radars operate around the world on a near-continuous basis, the great majority of which are modern commercially produced systems.

Radio Echoes from Meteors

A meteor occurs when a particle of interplanetary material, a meteoroid, enters the Earth’s atmosphere and ‘burns up.’ The majority of meteors detected by a typical meteor radar are caused by particles of less than 1 mm radius. Meteor entry speeds are in the range ~ 11 – 72 km s^{-1} and so rapid frictional heating occurs as the particle encounters the atmosphere. Once the meteoroid surface temperature reaches values of $\sim 1850 \text{ K}$, the surface ablates and significant mass loss results. The visible light emitted from the meteor arises mostly from the de-excitation of these excited meteoroid atoms lost from the particle’s surface. Inelastic collisions between these ablated atoms and the surrounding air molecules then produce an approximately cylindrical ionized trail behind the meteoroid particle.

The free electrons in this trail are able to scatter incident radio waves and so the meteor can be detected by radar systems. Although it is possible to detect a radio echo associated with the advancing ‘head’ of the meteor, atmospheric science measurements generally use the specular reflection observable in a direction perpendicular to the ionized trail axis. A full treatment of meteor physics and meteor scattering theory is quite involved and so here the theory will only be outlined briefly.

The structure of the ionized trail initially left by the meteor is a function of the entry speed, entry angle, and initial mass of the meteoroid particle. The trail’s initial radial structure consists of a core of high ionization density surrounded by less dense ionization in an outer region. The initial ionized trail radius is defined by the root mean square value of the radial position of the ablated ions after they have undergone the 10 or so collisions necessary to reach thermal equilibrium.

Atmospheric mean free path increases with height, and so the formative trail radius will be greater at greater heights. The initial trail radius also varies with meteor speed, but for a typical 40 km s^{-1} meteoroid the trail will form with an initial radius of $\sim 0.3, 0.6,$ and 1.2 m at heights of $75, 90,$ and 105 km , respectively. The trail itself will typically span some $10\text{--}15 \text{ km}$ in height. The electron line density of the trail is also a function of meteoroid mass and entry angle, but the majority of observed values are less than $\sim 10^{14} \text{ m}^{-1}$.

The amplitude of the radio echo received will rise rapidly to a maximum as the meteor moves past the specular reflection point, with most of the recorded signal being from the first Fresnel zone. The echo behavior after this maximum depends strongly on the electron line density. If the line density is below $\sim 10^{13} \text{ m}^{-1}$, then the radio wave fully enters the trail and the scattered signal that is received results from the individual contributions of all the electrons in the trail.

However, the trail radius will increase with time because of ambipolar diffusion. Within a few seconds the radius will have become sufficiently large that for typical radio wavelengths of some $\sim 5\text{--}10 \text{ m}$, destructive interference from scatter at different depths within the trail will extinguish the echo amplitude. The echo amplitude will thus decrease rapidly from the peak value. Such meteor echoes are known as underdense, and the time of decay to half peak amplitude, $t_{1/2}$, is related to the ambipolar diffusion coefficient, D , and radio wavelength, λ , by eqn [1].

$$t_{1/2} = \frac{\lambda^2 \ln 2}{16\pi^2 D^2} \quad [1]$$

Figure 1 presents the amplitude–time profile of a typical underdense echo. In this case, the echo lasts for $\sim 1 \text{ s}$ after the peak amplitude is reached before returning to the noise level again. This characteristic short-lived echo profile is very useful in distinguishing genuine meteor echoes from potentially spurious signals such as those from lightning or those produced by reflections from aircraft.

This destructive interference phenomenon also limits the greatest height at which underdense meteors can be detected using a particular radio wavelength, since at some upper height the trails will form with radii sufficiently great for destructive interference to occur instantly. This is sometimes known as the underdense echo ceiling. Here we should also note that

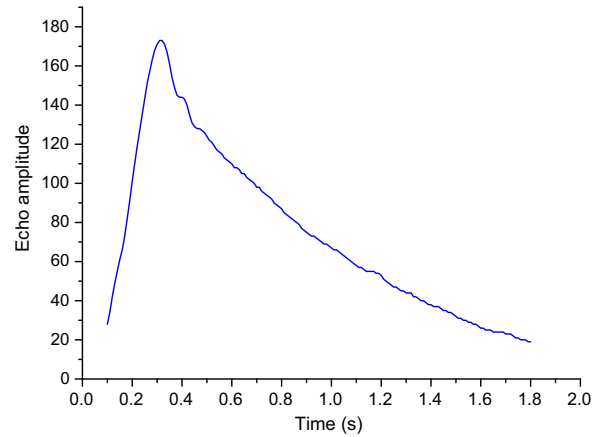


Figure 1 The amplitude profile of a characteristic underdense meteor echo, in this case recorded by the Advanced Meteor Orbit Radar (AMOR) in New Zealand (Jack Baggaley, private communication).

magnetic and electric fields may modify the behavior of meteor trails somewhat. For instance, magnetic fields result in anisotropic diffusion at heights above $\sim 95 \text{ km}$.

A minority of echoes will come from meteors with electron line densities greater than $\sim 10^{15} \text{ m}^{-1}$. In this case, secondary scattering between electrons becomes significant and the radio wave is reflected from the surface of the meteor trail as from a metal surface (i.e., dielectric constant <0). Such echoes are known as overdense, and do not show the rapid fall off in amplitude of the underdense echo. Overdense echoes may persist for many seconds, and the lifetime is determined by recombination processes that ultimately extinguish the meteor ionization.

Overdense echoes are less useful because the echo amplitude may fluctuate in a rather random manner as wind shears distort the trail and give rise to constructive and destructive interference from multiple specular reflection points. This makes unambiguous discrimination of the echo from noise more difficult. Such distortions may also cause a specular reflection point to move along the distorting trail, giving a spurious drift velocity. Electron line densities of $\sim 10^{14} \text{ m}^{-1}$ yield transitional behavior between the two echo types.

Typical Meteor Radars

A wide variety of radar powers, modes of operation, and antenna patterns can be used to detect meteors. Typically, a purpose-built system might be a pulse radar having the following parameters:

Operating frequency	25–60 MHz
Peak power	6–200 kW
Duty cycle	1–15%
Pulse repetition frequency	200–2000 Hz

The operating frequency must be selected to be high enough so that interference from ionospheric scatter is not significant,

but low enough so that the underdense echo ceiling does not become so low in height that it significantly reduces meteor count rates. Pulse repetition frequencies must be high enough that sufficient samples are recorded from a given underdense echo to determine its parameters before the echo amplitude decays back to the noise level. The set of parameters determined for each individual meteor echo would usually include the following: range, elevation and azimuth angles, power, echo decay time, and radial drift speed (which is measured by Doppler techniques).

Typically, the distribution of radial drift speeds is approximately Gaussian with a full width at half maximum of $\sim 55 \text{ m s}^{-1}$. An uncertainty in individual radial drift speeds of $< 2 \text{ m s}^{-1}$ might be attained for at least 50% of meteors recorded and nearly all would have uncertainties of $< 5 \text{ m s}^{-1}$.

Various antenna configurations can be used. So-called all-sky systems are now almost universally used in purpose-built meteor radars designed for atmospheric science. These systems use low-gain antennas to illuminate a large volume of the meteor region around the radar. Echo azimuth and elevation angles are usually determined by an array of receiver antennas acting as an interferometer. Horizontal wind speed and direction for a particular height range are then calculated by a least-squares fitting procedure applied to the individual radial drift measurements made over all azimuth angles. Note that these derived winds will thus be means over the horizontal extent of the radar's meteor collecting volume.

Beam systems will typically use relatively narrow beams and deduce horizontal flow speed and direction by ascribing the radial velocities measured in a particular beam to the azimuth of that beam's axis. If height information is not required, such a system needs no interferometer and can be particularly simple and inexpensive. The resulting winds are then taken to represent a mean over the meteor region weighted by the vertical distribution of meteor echoes.

Usually, this would mean considering the winds to be representative of heights of $\sim 90\text{--}95 \text{ km}$. Some radar systems designed primarily for other purposes, such as the HF Super-DARN radars, can make meteor measurements of this type.

Patrol-type radars may operate at low power and only switch to high power once a meteor is detected, or may rotate transmitted power between different beams, dwelling in one direction only once a meteor has been detected.

Finally, we should note that the specular nature of the radio reflection means that the majority of meteors falling into a radar's collecting volume will still not be recorded simply because the trail/radar geometry does not permit a perpendicular reflection, even for so-called all-sky radars. Regardless of the particular beam geometry used, the individual radial meteor drifts are used to deduce a horizontal flow in the atmosphere under the reasonable assumption that the vertical flow velocities are very small compared to the horizontal velocities on timescales comparable to the achievable time resolution (generally $\sim 1 \text{ h}$). Further, it is usually assumed that at any particular height the atmospheric flow is uniform over the entire horizontal extent of the meteor collecting volume. Vertical flows cannot be measured easily by meteor radars.

The strengths and limitations of meteor radars as tools for investigating the atmosphere are thus largely a consequence of the distribution of the recorded meteor echoes in space and time. To illustrate this, we will consider data recorded by a very typical modern all-sky radar, in this case a commercially produced SkiYmet radar located at Rothera in the Antarctic (68° N , 68° W). This particular radar operates at a frequency of 32.5 MHz , has a peak power of 6 kW , a duty cycle of 15%, and a pulse repetition frequency of 2144 Hz . It uses an array of five receiver antennas acting as an interferometer to allow determination of echo azimuth and elevation angles, which in combination with range information allow echo height determination. [Figure 2](#) shows one of the crossed-element Yagi



Figure 2 One of the receiver antennas of a typical all-sky meteor radar. In this case, the radar is at Rothera (68° N , 68° W) in the Antarctic.

receiver antennas used, a rather typical antenna for this type of radar. Note that, although radars of differing transmitted power, antenna polar diagram, radio frequency, etc. will give slight variations in these distributions, their general character is representative of nearly all purpose-built meteor radars.

We now consider the spatial distribution of meteors recorded by the radar. Figure 3 shows the distribution of the 4617 individual meteors recorded on a particular day, projected onto a horizontal plane. The meteors are distributed in a circular pattern around the radar at horizontal distances up to 300 km. Allowing for the circular gaps in the distribution (caused by the receivers being short-circuited during the transmission of successive radio pulses) the peak in the horizontal range distribution of meteors is at about 120 km. The vertical distribution of these meteor echoes is illustrated by the histogram in Figure 4. Note how the distribution is strongly peaked at a height near 90 km, and how very few meteors are recorded below 80 km or above about 100 km. These figures highlight two key differences between the data from a typical meteor radar and those from, say, MF, MST, and IS radars.

First, these other techniques generally produce measurements that are essentially volume averages, usually resulting from scatterers distributed throughout a volume defined by the intersection of the transmitted beam with the height range of interest. Because the transmitted beams are quite narrow, the measured wind fields are deduced from a volume of atmosphere perhaps only a few kilometers in horizontal extent. In contrast, meteor radar measurements arise from quite localized scatter from individual short-lived meteors, but the meteors themselves are rather randomly distributed within a much larger collecting volume in the meteor region, which may extend horizontally for several hundred kilometers around the radar. The winds derived for a particular height range will

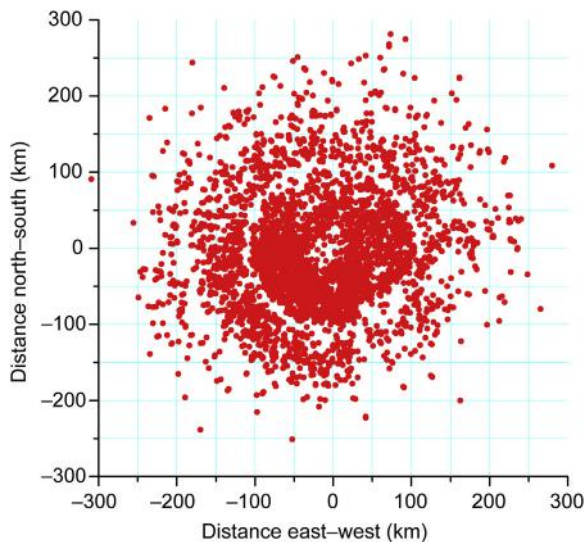


Figure 3 The distribution of 4617 individual meteors detected by an all-sky meteor radar at Rothera (68° N, 68° W) in the Antarctic, recorded on 28 August 2008. The meteor locations have been projected onto a horizontal plane with the radar at the central position. The clear rings in the distribution mark where the receiver antennas are short-circuited during the transmission of successive transmitter pulses.

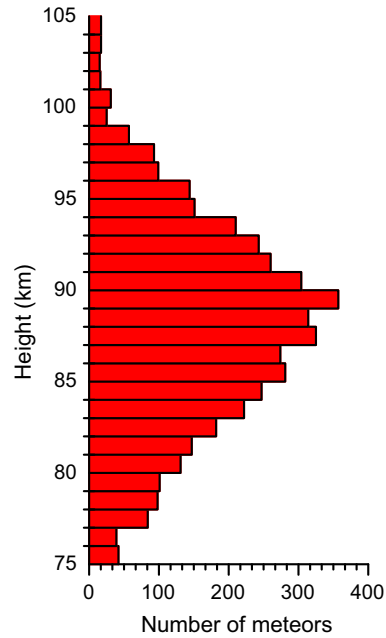


Figure 4 The vertical distribution of the meteors presented in Figure 3.

therefore represent the bulk motion of a slab of atmosphere that is only a few kilometers deep but that has a horizontal extent of several hundred kilometers. This means that a meteor radar cannot easily resolve structure on horizontal scales less than a few hundred kilometers.

Second, the vertical depth of atmosphere from which sufficient meteors are recorded for reliable calculation of winds is comparatively small. In the example presented here, winds can be effectively calculated only for heights between about 80 and 98 km, because above and below these heights there are too few meteors for reliable wind estimates to be made. Note that meteor radars using a beam, rather than all-sky, configuration will still share these properties, since to ensure adequate meteor count rates the beams must be quite broad.

Finally, there is also a distinct diurnal cycle in the count rate of meteor echoes. This is largely an astronomical phenomenon resulting from the Earth's orbital motion and rotation, which causes it to sweep up meteors on its leading hemisphere, whereas a meteor must overtake the Earth in order to hit the planet's trailing hemisphere. The effect is strongest at lower latitudes and for beam systems. The net result is that meteor count rates will generally be lowest in the second part of the day, although complex geometrical and instrumental considerations will greatly influence the magnitude of the diurnal variation recorded by any particular radar.

As an example, Figure 5 presents normalized hourly mean count rates for three meteor radars. The first is an all-sky radar operated at Esrange (68° N), the second is a twin beam radar operated in the United Kingdom (52° N); and the third is an all-sky radar operated on Ascension Island (8° S) in the equatorial Atlantic. It can be seen that even in the favorable all-sky, high-latitude case the ratio of maximum to minimum hourly meteor count rates is about 2:1, and for the midlatitude beam

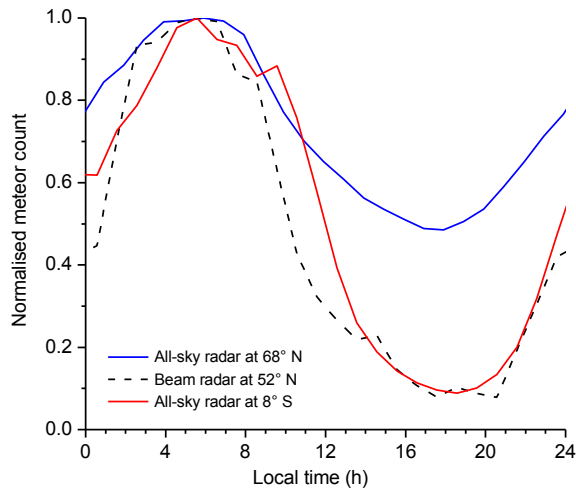


Figure 5 The diurnal cycle in hourly normalized meteor count rate for three meteor radars: the ESRANGE all-sky radar at 68° N (solid); a two-beam radar at 52° N in the United Kingdom (dashed); and a second all-sky radar at 8° S on Ascension Island (dotted).

system and the equatorial all-sky system it is about 10:1. This effect means that care must be taken in deriving winds for the low-count-rate part of the day (since there are simply fewer meteors to work with).

These distributions of meteors in space and time determine the resolution of the resulting wind measurements. They will generally have reasonable height and time resolutions of ~ 3 km and ~ 1 h, respectively, but poor horizontal resolution (although, as noted, not all radars actually determine echo heights). This rather unexciting sounding aspect of the data is counterbalanced by the almost unequalled ability of meteor radars to record continuous high-quality data sets over many years, without reliance on particular geomagnetic conditions, weather, or time of day.

All these factors influence the particular aspects of mesosphere/lower thermosphere dynamics that can be studied easily using meteor radar winds. Generally, the mean flow, tides, and planetary waves all have horizontal scales so large that the poor horizontal resolution of the radars is irrelevant. The vertical scale sizes and wavelengths of these phenomena are also significantly larger than the normal height resolution of a few

kilometers, and so meteor radars can make detailed studies of the vertical structure of these features across the meteor region. Meteor radars are thus well suited to the study of mean winds, tides, and planetary waves and have made very significant contributions in this area. This is particularly so because the radar hardware is often quite simple and well suited to prolonged operation over periods of years at remote sites with only limited maintenance.

An example of the mesospheric winds measured by a meteor radar is presented in Figure 6. The figure shows hourly mean zonal winds recorded by the all-sky meteor radar at Bear Lake Observatory (42° N, 112° W) in Utah, USA. The figure reveals a motion field dominated, in this case, by the oscillations of the 12-h tide, which increase in amplitude with height and display a high degree of short-term variability.

Unlike the large horizontal scales associated with winds, tides, and planetary waves, the horizontal wavelengths of gravity waves can often be comparable to or smaller than the horizontal extent of the meteor collecting volume. Self-cancellation thus occurs between different parts of the wave phase measured across the meteor collecting volume. In effect, the individual gravity waves cannot be resolved. Because of this, meteor radars have not usually been used in studies of gravity waves.

Recently, however, analysis methods have been developed to address this problem. These methods use the radial velocities measured for individual meteor echoes within the meteor collecting volume to determine statistical properties of the total gravity wave field in the meteor collecting volume, without having to resolve individual gravity waves. In particular, estimates can be made of the degree of gravity wave activity (measured by the variance of the perturbation winds), the anisotropy of the wave field propagation azimuths and the vertical fluxes of horizontal momentum. The latter estimates require measurement of the vertical perturbation velocities induced by gravity waves. Because of this, radars optimized for gravity wave measurements will sometimes use antenna configurations designed to enable the detection of meteors at higher elevation angles, such that the radial velocity will then provide information on vertical winds. Measuring momentum fluxes at different heights then allows an estimation of the divergence of the momentum flux and thus the associated acceleration of the mean flow, which is a parameter of great interest in understanding the large-scale dynamics of the mesosphere.

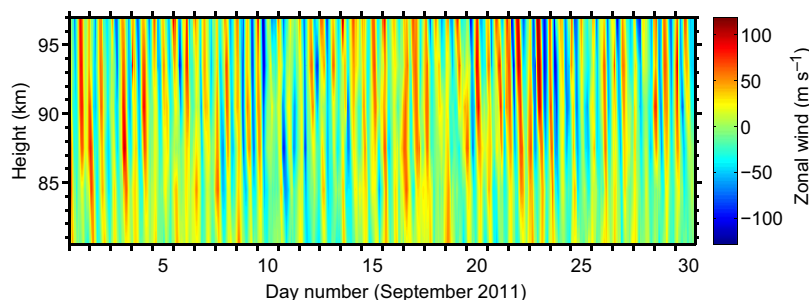


Figure 6 The zonal winds recorded in September 2011 by a meteor radar at Bear Lake Observatory (42° N, 112° W) in Utah, USA. The data have a time resolution of 1 h and a height resolution of about 3 km. The winds are dominated by the 12-h tide, which reaches large amplitudes in autumn at middle latitudes.

Other Measurements Using Meteor Radars

Meteor radars have also been used extensively for studies of the astronomical aspects of meteors, including measuring meteor fluxes, size distributions, entry speeds, and orbits. Other work has expanded the capabilities of the technique for atmospheric science measurements. In particular, various methods have been used to relate the ambipolar diffusion coefficient (measured from echo decay times) either to temperature–pressure-related quantities such as T^2/P or to wave-induced fractional perturbations of temperature or pressure. The poor quality of available absolute pressure estimates for the upper mesosphere prevents these methods from being used to infer absolute temperature values.

However, a recent extension of these methods has seen the application of a scale height analysis of ambipolar diffusion coefficients to determine absolute atmospheric temperature at the height of peak meteor counts without any reliance on knowledge of pressure. This capability is potentially very useful because ground-based measurements of atmospheric

temperature at meteor heights are otherwise extremely difficult to make on an extended basis.

See also: **Mesosphere:** Metal Layers; Polar Summer Mesopause. **Radar:** Incoherent Scatter Radar; Mesosphere–Stratosphere–Troposphere and Stratosphere–Troposphere Radars and Wind Profilers; Synthetic Aperture Radar (Land Surface Applications). **Solar System/Sun, Atmospheres, Evolution of Atmospheres:** Meteors.

Further Reading

- Celpecha, Z., Borovička, J., Elford, W.G., et al., 1998. Meteor phenomena and bodies. *Space Science Reviews* 84, 327–471.
- Hocking, W.K., Fuller, B., Vanderpeer, B., 2001. Real time determination of meteor-related parameters utilizing modern digital technology. *Journal of Atmospheric Solar-Terrestrial Physics* 63, 155–169.
- McKinley, D.W.R., 1961. *Meteor Science and Engineering*. McGraw-Hill, New York.
- Roper, R.G., 1984. MWR – meteor wind radars. *MAP Handbook* 13 (10), 124–134.

Polarimetric Doppler Weather Radar

RJ Doviak, National Severe Storms Laboratory, Norman, OK, USA

RD Palmer, University of Oklahoma, Oklahoma, OK, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by R J Doviak, M E Frazier Doviak, volume 4, pp 1802–1812, © 2003, Elsevier Ltd.

Glossary

A Echo amplitude.

ADC Analog-to-digital converter.

ATSR Alternate transmit simultaneous receive.

B_6 Receiver bandwidth.

D_e Equivalent diameter of a raindrop.

E_h, E_v Electric field components polarized horizontally, vertically.

c Speed of light.

$E[x]$ Mean or expected value of x .

f_t Transmitted frequency.

f_d Doppler frequency.

$I(r, t)$ Real part of $V(r, t)$.

j $\sqrt{-1}$, imaginary unit.

K_w The complex dielectric factor for water.

m Integer index.

mT_s Sample time.

$n(\vec{r})$ Density of scatterers.

P_{av} Average power transmitted.

P_t Peak power transmitted.

PAR Phased array radar.

PPAR Polarimetric phased array radar.

NWRT National Weather Radar Testbed.

$Q(r, t)$ Imaginary part of $V(r, t)$.

r Range.

s_{hv} Complex backscattering matrix element.

STALO Stabilized local oscillator.

STSR Simultaneous transmit simultaneous receive.

t Time.

T_s Pulse repetition time (sampling time).

$U(x)$ $U = 1$ if $0 < x < \tau_v$; otherwise $U = 0$.

$V(r, t)$ Echo voltage.

V_6 Resolution volume.

v_r Radial velocity.

WSR Weather surveillance radar.

Z_h, Z_v The reflectivity factor measured with H, V polarized waves.

Z_{DR} Differential reflectivity factor.

ϕ_0 Beam azimuth.

ϕ Beam azimuth relative to north.

θ_e Beam elevation angle.

θ_0 Zenith angle.

θ_1 One-way beamwidth at the half-power level.

λ Wavelength.

σ_v Spectrum width.

σ_τ Correlation-time along mT_s .

τ_c Correlation-time along τ_s .

τ_s Range-time.

τ_t Transmitted pulse width.

ψ Sum of phase shifts within the radar and the scatterer.

ψ_e Echo phase.

Synopsis

The theory of, and sample results from, Polarimetric Doppler Weather Radar are presented so readers can have an explanation of the foundations upon which weather radars are used by nations to provide to their citizens timely warnings of weather hazards, to make quantitative measurements of depths of precipitation fall, and to classify precipitation types.

Introduction

The term ‘Doppler’ is in honor of the Austrian physicist, Christian Johann Doppler, who first explained the principle of the ‘Doppler effect’ in 1842. The Doppler effect is the increase or decrease in the pitch or frequency of sound waves when a source of waves is moving toward or away from the listener. This is quite evident to an observer who is listening to the blare of an automobile horn as it passes by. The observer hears first a relatively high frequency wave and then a marked drop in frequency.

Electromagnetic waves (e.g., light and radio waves) emitted from galaxies also have Doppler shifts. Spectral lines of emitted radiation are Doppler shifted to lower frequencies because most galaxies are moving away in our expanding universe; blue spectral lines are shifted to lower frequency red lines, the so-called red shift.

Unlike sound waves, which have vibrations in the direction of propagation, electromagnetic waves have vibrations in a plane transverse to the direction of propagation. Through careful design of the radar, the direction of the electric field in

the plane of oscillation can be controlled thereby creating a particular *polarization*. The polarization of the electromagnetic waves has a profound effect on their interaction with scatterers (e.g., hydrometeors, insects, refractive index perturbations, etc.) in the atmosphere.

Doppler weather radars owe their acceptance by the weather services around the world to the fact that radio waves having wavelengths on the order of centimeters can penetrate extensive fields of precipitation (e.g., hurricanes) to map and reveal, like an X-ray photograph, the internal structure and motion of weather systems. Optical and infrared waves do not penetrate very far into clouds and precipitation.

The next section shows how radar uses the Doppler effect to measure the radial velocity of scatterers in the earth’s atmosphere, and how it measures the polarimetric properties of radiation backscattered from these scatterers to improve estimates of rainfall and to distinguish various types of hydrometeors (e.g., rain, snow, hail, etc.). In Section [Fundamentals of Polarimetric Doppler Radar](#), the basic theory of weather radar is provided for backscatter from a single hydrometeor. The theory is then extended to a volume of hydrometeors in Section [Volume Scattering](#), and samples of polarimetric and Doppler data are shown to emphasize the usefulness of these measurements to observe weather phenomena. Finally, the authors’ perspective on the future of weather radar is provided.

The Doppler Shift and Polarimetric Properties of Scatterers

Scatterers in the atmosphere produce measurable Doppler shifts if the scatterers are illuminated with narrow band radiation (e.g., a radio wave transmitted at a single frequency, f_i). Radio waves incident on a scatterer force electromagnetic vibrations in the scatterer. If the scatterer is at constant range r from the radar, vibrations are at f_i . If the scatterer moves toward the radar, the internal vibrations will be faster because the wave’s apparent propagation speed relative to the scatter is faster and thus the approaching scatterer experiences more rapid fluctuations of the incident waves. Thus the backscattered radiation (i.e., echoes) received by the radar will have a frequency higher than f_i .

Doppler weather radars typically transmit microwave radiation in bursts of short duration, τ_t (e.g., 10^{-6} s or 1 μ s). These radars are called pulsed Doppler radars to distinguish them from those that emit continuous waves such as radars used by police to detect the speed of automobiles. Pulsed Doppler radars can measure both range r and radial velocity v_r of scatterers.

The US National Weather Service (NWS) network of Doppler weather radars emits 1.57 μ s duration pulses of 10-cm wavelength $\lambda = c/f_t$ radiation (i.e., similar to the λ used in microwave ovens; c is the speed of light) to map r and v_r of hydrometeors. These radars are called WSR-88Ds, where WSR is an acronym of Weather Surveillance Radar. The number 88 denotes the year (1988) a contract was issued for a limited production and testing; the first unit for routine operation was installed in the late 1992 and the last radar of a network of about 160 radars was installed in 1997; the letter *D* represents its Doppler capability; polarimetric upgrades were completed in 2013.

The measurement of the Doppler shift in a single back-scattered pulse (i.e., an echo from an object, small compared to $c\tau_t$) of microwaves is not practical. Instead, the change of the echo’s phase angle from one transmitted pulse to the next is measured. Phase angle is proportional to the scatterer’s range and equals twice the number of wavelengths, and portions thereof, between the radar and the scatterer. Phase angle measures quite precisely (i.e., to fractions of a wavelength) the distance to the scatterer, but there is ambiguity because the phase can be only determined within a wavelength. Nevertheless, the change of the echo phase, from pulse to pulse, is used to measure v_r , the time rate of change of range to the scatterer.

If all hydrometeors were spherical, polarization would not play a major role in the design of weather radar. In stark contrast, hydrometeors can be found in a variety of shapes, sizes, orientations, and size distributions. Simple visual observations of snow, for example, exemplify the vast array of possibilities. Even raindrops, which are often thought to be spherical, exhibit deformation due to the drag force. This effect becomes gradually more important with increasing drop size, as illustrated in [Figure 1](#). Note that liquid cloud droplets with diameters on the order of microns are nearly spherical and thus equally backscatter linearly polarized waves of any orientation – e.g., horizontal *H* or vertical *V* polarization. The equivalent diameter D_e is the diameter of a spherical drop having the same volume as the deformed drops shown in the figure. Also shown in the figure is the *differential reflectivity*, Z_{DR} , a measure of the ratio of backscatter from *H* and *V* polarized waves.

Nonhydrometeorological scatterers found in the atmosphere (e.g., insects, birds, aircraft, etc.) typically are nonspherical. For these reasons, polarimetric signatures of scatterers are important for weather radar design and can be used for more precise characterization of their properties, leading to accurate rainfall estimates and the capability of distinguishing rain from snow, hail, insects, etc.

Polarimetric weather radar produces electromagnetic oscillations in a plane – the plane of polarization – orthogonal to the direction of wave propagation. In this plane, the electric

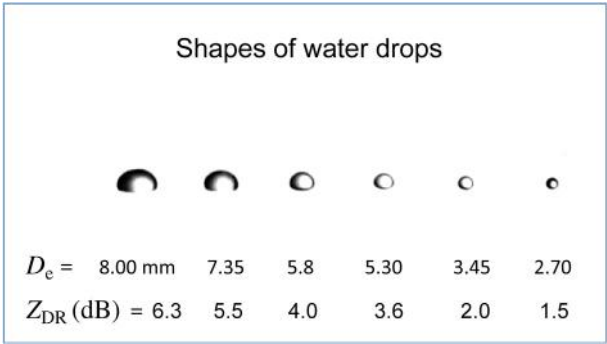


Figure 1 Shapes of raindrops falling in still air and experiencing drag force deformation. D_e is the equivalent volume diameter of a spherical drop. Z_{DR} (dB) is the differential reflectivity in decibels for each drop back-scattering 10-cm-wavelength radiation. This figure is adapted from one in Pruppacher, H.R., Beard, K.V., 1970. A wind tunnel investigation of the internal circulations and shape of water drops falling at terminal velocity in air. Quarterly Journal of the Royal Meteorological Society 96, 247–256.

field vector can be decomposed into two components having orthogonal directions, H and V , the horizontally and vertically polarized components. For H polarization, the vibrations are parallel with the ground, but for the V polarizations the vibrations lie in a vertical plane as illustrated in Figure 2. Due to the nonspherical shape of most hydrometeors, the backscattered signals from each of these linear polarizations will be different; quantification of these differences is the foundation of polarimetric radar.

H and V polarization is often implemented using the simultaneous transmit simultaneous receive (STSR) mode, as illustrated in Figure 2. As the name implies, electromagnetic waves with both H and V polarizations are transmitted simultaneously. In the STSR mode, the backscattered radiation from the scatterers is also received simultaneously in H and V receivers. Another mode alternates, from pulse to pulse, the H and V transmissions, but has simultaneous H and V reception (alternate transmit simultaneous receive (ATSR)). The major advantage of the ATSR configuration is its capability to measure *cross-polar* signals, where backscatter from a horizontally polarized transmitted pulse is received with vertical polarization or vice versa. Such cross-polar measurements do contain meteorological information but are extremely weak compared to the *copolar* signals (e.g., H polarized echoes are received when H polarized pulses are transmitted). Implementation challenges of the ATSR mode with high-powered transmitters contributed to the decision by the NWS to use the STSR mode for the WSR-88D network (Doviak et al., 2000).

Consider a single stationary hydrometeor at range r . The effects on the polarimetric signature can be divided into two components: *propagation effects* and *backscattering effects*. If no precipitation exists between the radar and the hydrometeor, propagation effects can be ignored and any polarimetric signature would be due solely to backscattering effects. For clarity, we will ignore propagation effects. In this case, the backscattered electric field is described by the following matrix equation (Doviak and Zrnić, 2006, eqn [8.39])

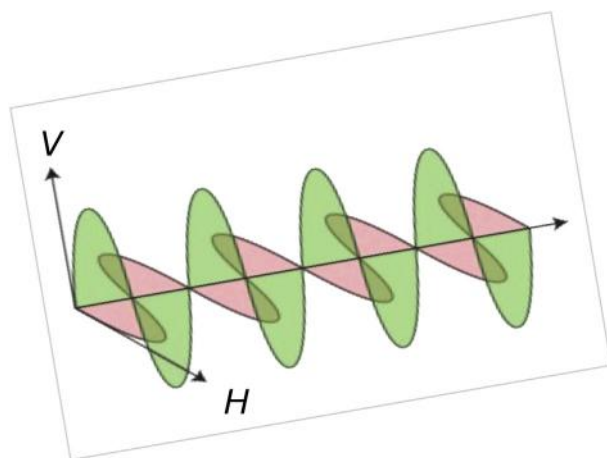


Figure 2 Electromagnetic waves propagating with H and V polarizations aligned in time (i.e., *in phase*). When two waves with H and V polarizations are simultaneously transmitted in phase, the resulting composite electromagnetic field oscillates along a 45° angle – called *slant linear 45°* – in the plane of polarization.

$$\begin{bmatrix} E_h \\ E_v \end{bmatrix}^b = \begin{bmatrix} s_{hh} & s_{hv} \\ s_{vh} & s_{vv} \end{bmatrix} \begin{bmatrix} E_h \\ E_v \end{bmatrix}^i \frac{\exp(-jkr)}{r} \quad [1]$$

where E_h and E_v are the electric field components in the h and v directions, the superscripts i and b correspond to the incident and backscattered fields, respectively, $j = \sqrt{-1}$, and s_{hv} is the complex backscattering cross section when the incident wave is vertically polarized and the horizontal component of the backscattered electric field is measured. The other elements of the backscattering matrix follow accordingly. s_{hv} is called complex because it has magnitude – proportional to the size, shape, and orientation of the hydrometeor – and phase. The exponent of the exponential factor in the equation is also a phase term. The magnitudes and phases fluctuate in time because the hydrometeor's size, shape, and orientation change due to turbulence and collisions. By independently transmitting and receiving horizontal and vertical polarizations (e.g., ATSR mode), it is possible to measure all the elements of the backscattering matrix, although the cross-polar terms (s_{hv} and s_{vh}) are one to two orders of magnitude weaker than the copolar terms (s_{hh} and s_{vv}). In contrast, the STSR mode allows relatively easy measurement of only the *copolar* terms (s_{hh} and s_{vv}) but sophisticated coding schemes (Chandrasekar and Bharadwaj, 2009) are required to estimate s_{hv} .

Fundamentals of Polarimetric Doppler Radar

The fundamentals of polarimetric Doppler radar can be more easily explained by discussing a simplified block diagram of polarimetric homodyne radar (Figure 3). Weather radar has an oscillator called the STALO (stabilized local oscillator), which generates a very pure continuous microwave (i.e., a spectral line) at a frequency, f_c . This continuous wave is converted to a sequence of microwave pulses by the modulator. Frequency f_c is called the carrier frequency because it carries the modulation pulse. The transmitted microwave pulses, of time duration τ_t and spaced T_s (≈ 1 ms) are intensified by a high power amplifier (e.g., a klystron) to produce a large peak pulse power, P_t , often close to 1 MW. The average power $P_{av} = (\tau_t/T_s)P_t$, however, is typically a kilowatt or less, about that used in microwave ovens. The transmitted pulse of vertically polarized microwaves of spatial length $c\tau_t$ is in a beam, formed by the parabolic reflector antenna, directed at azimuth and elevation angles, $\phi = 90^\circ - \phi_0$ and $\theta_e = \frac{\pi}{2} - \theta_0$, respectively (radar meteorologists measure azimuth, ϕ clockwise from north). Weather radar beams typically have a circular half-power angular width $\theta_1 \approx 1^\circ$. It should be recognized that the reflector antenna, which determines the overall quality of the polarimetric signals, is the key component of the radar system. A feed placed at the focal point of the parabolic reflector emits the H and V electromagnetic energy onto the reflector.

It is difficult to make direct measurements of the H , V microwave echoes, so most receivers shift f_c either to zero – as done by a homodyne receiver – so that weather echoes can be sampled and digitized by analog-to-digital converter (ADC) electronics of the digital receiver (Figure 3) or, more commonly, to an intermediate frequency where the ADC can also be applied. The digitized data are processed by computers and results are displayed and recorded.

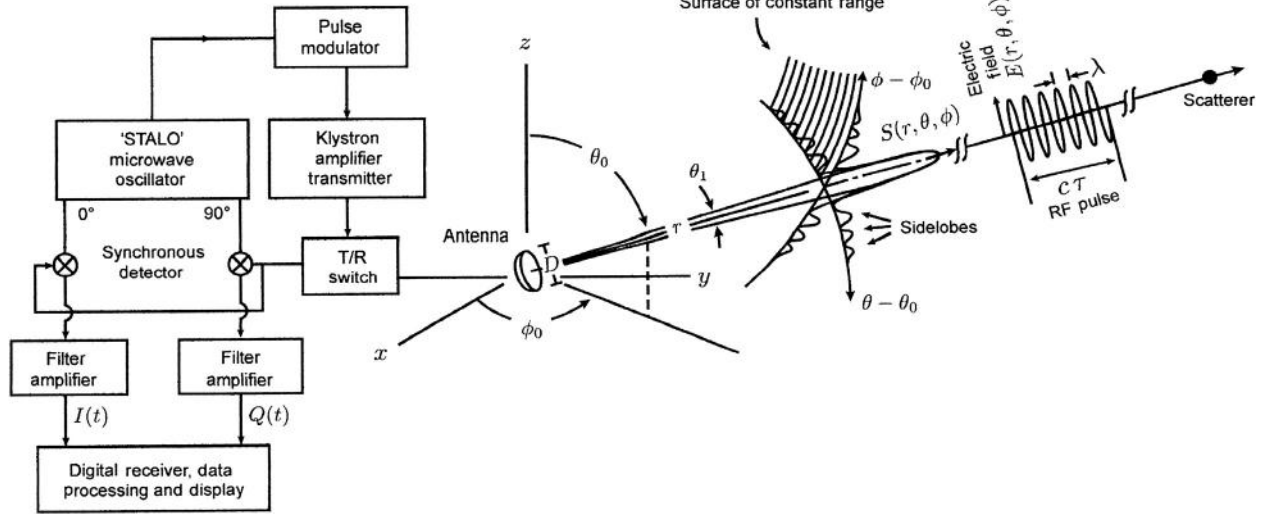


Figure 3 Block diagram for a single V channel – waves polarized in a vertical plane are transmitted – of homodyne polarimetric Doppler radar. For STSR mode, the klystron-amplified pulse would be sent to H (not shown here) and V channels of transmit/receive (T/R) switches and then to the dual-polarimetric antenna. The H and V receive channels (i.e., synchronous detectors, etc.) would have separate electronics to receive H , V weather echoes independently.

With advances in digital technology, modern receivers that digitize the signals at the shifted intermediate frequency and translate these signals to baseband (i.e., $f_i = 0$) are becoming ubiquitous. These so-called *digital receivers*, or *software-defined receivers*, have several advantages including a perfect quadrature relation between the in-phase and quadrature signals, and a flexible design. Since the receiver characteristics are controlled in software, updates can easily be incorporated as advances in signal processing become operational.

In a homodyne receiver, the echo of pulsed microwaves is directly passed to a synchronous detector (Figure 3) and mixed (multiplied) with signals from the STALO. A synchronous detector is required to resolve both the echo's magnitude and sign of its phase angle or Doppler shift. The two STALO signals are sinusoid and cosinusoid producing the *in-phase* $I(t)$ and *quadrature* $Q(t)$ signals, respectively.

To explain the pulse-to-pulse shift in echo phase and its relation to v_r , two cases are examined for the ATSR mode: a simple case of a stationary scatterer at r , and a more complex case when the scatterer is allowed to move. In the simple case, the copolar echo voltage (e.g., V_{hh}) at the receiver's input is essentially a scaled replica of the transmitted pulse and is given by

$$V_{hh}(r, t) = A s_{hh} \exp \left\{ j \left[2\pi f_t \left(t - \frac{2r}{c} \right) + \psi \right] \right\} U \left(t - \frac{2r}{c} \right) \quad [2a]$$

$$= A s_{hh} \left[\cos \left(2\pi f_t t - \frac{4\pi r}{\lambda} + \psi \right) + j \sin \left(2\pi f_t t - \frac{4\pi r}{\lambda} + \psi \right) \right] U \left(t - \frac{2r}{c} \right) \quad [2b]$$

The cosine and sine terms are the real and imaginary components of $V_{hh}(r, t)$ oscillating at the rate f_t . A is the echo amplitude, s_{hh} is the copolar backscattering matrix element for

HH polarization, $2\pi f_t \left(t - \frac{2r}{c} \right) + \psi$ is the microwave echo phase, t is the time the echo is received after the emission of the transmitted pulse, and ψ is the sum of phase shifts within the radar and the scatterer. The pulse function $U \left(t - \frac{2r}{c} \right)$ is unity when its argument $t - \frac{2r}{c}$ is between 0 and τ_t , and is zero otherwise. Equation [2a] is an exponential form and a compact way of expressing the microwave echo.

As just discussed, it is difficult to process the echo voltage at the transmitted microwave frequency, f_t so the microwave echo is downconverted to baseband (i.e., zero carrier frequency) using the synchronous detector (or digital receiver). An expression for this baseband signal can be obtained from eqn [2b] by setting $f_i = 0$. Thus, the echo voltage at the receiver's output (HH polarization, in this case) is

$$V_{hh}(r, t) = I_{hh}(r, t) + jQ_{hh}(r, t), \quad [3a]$$

where

$$I_{hh}(r, t) = A s_{hh} \cos \psi_e U \left(t - \frac{2r}{c} \right), \quad [3b]$$

$$Q_{hh}(r, t) = A s_{hh} \sin \psi_e U \left(t - \frac{2r}{c} \right), \quad [3c]$$

are the in-phase and quadrature components, and the echo phase is defined as

$$\psi_e = -\frac{4\pi r}{\lambda} + \psi \quad [4]$$

For simplicity, henceforth the polarimetric notation subscript on the in-phase and quadrature signals is dropped and will be used only where necessary for clarity.

Sampling in Range and Time

As shown in eqn [3], the echo voltage is a function of both range r and time t . Because pulses are transmitted every T_s ,

echoes from a stationary scatterer will periodically appear at $t = \frac{2r}{c} + mT_s$, where $m = 1, 2, 3, \dots$ defines each echo pulse. Because the scatterer's range is not known, a search for echoes is made by circuits that sample, at rates typically $> \tau_s^{-1}$, for echoes and the sampling process is reset every T_s . The T_s interval is defined as *range-time* τ_s ($0 \leq \tau_s \leq T_s$) because the location of the echo within T_s defines the range r to the scatterer. If the scatterer moves, not only will the echo pulse change its location along τ_s , but ψ_e will also change according to eqn [4]. Both echo position along τ_s and ψ_e can in principle be used to measure the change in scatterer location, and thus indirectly its radial velocity. However, ψ_e change is a more accurate measure of changes in scatterer location. For example, a change δr of r by $\lambda/4$ (e.g., 2.5 cm for the WSR-88D) causes ψ_e to change by 180° , a large angular change, whereas the change $\delta\tau_s$ along $\tau_s = 2r/c$ is $\delta\tau_s = 2\delta r/c$ (i.e., 1.67×10^{-10} s), a tiny fraction of τ_s . Thus, scatterer motion is measured by changes in ψ_e . As a consequence, the pulsed Doppler radar is an amplitude and phase sampling system (i.e., samples of A_{shh} and ψ_e at τ_s are obtained at T_s intervals). Range-time τ_s determines the range to the scatterer, and changes in ψ_e for echoes sampled at τ_s are measured from pulse to pulse along *sample-time*, mT_s . Because of this sampling process, the sampled echo voltage is written in the following form

$$V(\tau_s, mT_s) = I(\tau_s, mT_s) + jQ(\tau_s, mT_s). \quad [5]$$

This sampled echo voltage can be represented as a vector on the *Argand diagram* (e.g., Figure 4b). The carrier-shifted vector has the amplitude $|V(\tau_s, mT_s)|$ and echo phase ψ_e (positive when measured counter clockwise (ccw) from the $I(\tau_s, mT_s)$ axis).

Echo Phase Shift

Equation [4] shows ψ_e changes with time as the scatterer's r changes at a rate corresponding to its v_r . This time rate of change of ψ_e is

$$\frac{d\psi_e}{dt} = -\frac{4\pi}{\lambda} \frac{dr}{dt} = -\frac{4\pi}{\lambda} v_r = 2\pi f_d. \quad [6a]$$

where,

$$f_d = -\frac{2v_r}{\lambda} \quad [6b]$$

is the Doppler frequency f_d . Although it is possible to calculate a radial velocity for each copolar signal, v_r is the same for each polarization.

As an example, the $I(\tau_s, mT_s)$ and $Q(\tau_s, mT_s)$ components from a 10-cm-wavelength weather radar illuminating both stationary and moving scatterers are shown in Figure 4(a) as a function of τ_s for five successive transmitted pulses spaced $T_s = 768 \times 10^{-6}$ s. The echoes from the moving scatterer clearly exhibit a systematic change caused by v_r . Amplitudes of $I(\tau_s, mT_s)$ and $Q(\tau_s, mT_s)$ can change from a positive maximum to a negative maximum if the scatterer moves $\lambda/4$ in T_s . As shown earlier, under this same condition, the echo's range time τ_s changes less than 2×10^{-10} s, a shift too small to be detected in Figure 4(a).

Let us now calculate f_d and v_r from the $I(\tau_s, mT_s)$ and $Q(\tau_s, mT_s)$ changes seen in Figure 4 for the moving scatterer.

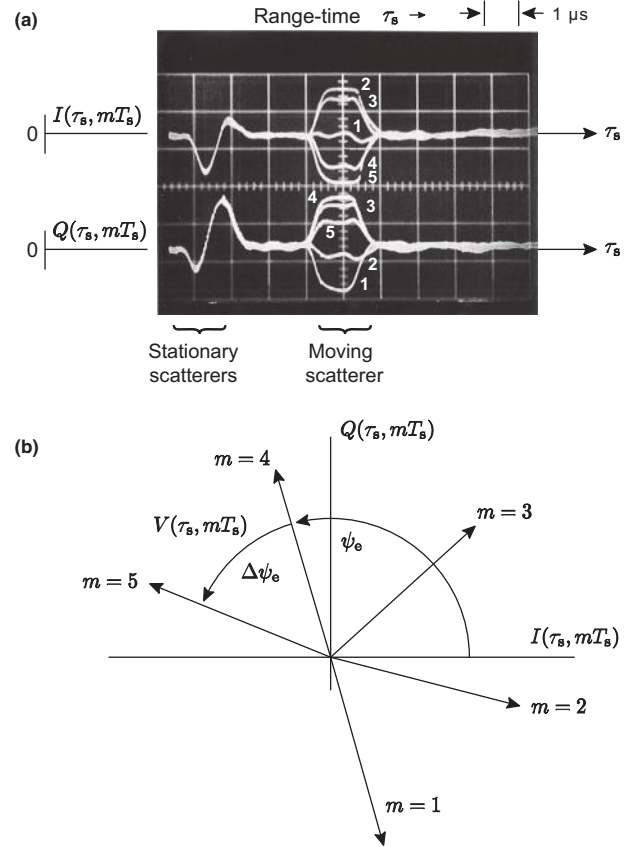


Figure 4 (a) In-phase and quadrature signals as a function of range time τ_s for five successive T_s intervals superimposed to show the echoes' relative change for stationary and moving scatterers. (b) An Argand diagram of the five echo samples in (a) for the moving scatterer.

This scatterer's five samples are sketched on the Argand diagram in Figure 4b. We see that the echo vector samples rotate ccw about the origin, and by convention this corresponds to a positive Doppler shift. Because the echo phase change $\Delta\psi_e$ between pulses 1 and 2 is approximately 60° during T_s ,

$$f_d(\text{Hz}) = 60^\circ \times \frac{1 \text{ cycle}}{360^\circ} \times \frac{1}{T_s}$$

computes to be about +217 Hz. Using eqn [6b], v_r computes to be about -10.9 m s^{-1} .

It should be mentioned that pulsed Doppler radar inherently limits both the range and velocities that can be measured unambiguously. If the radar transmits a uniform periodic sequence of pulses, which is normally the case, it is not possible to determine which transmitted pulse produced which echo. Furthermore, the periodic transmitted pulse sequence also introduces velocity ambiguities. These problems are well illustrated in a like article in the first edition of this Encyclopedia. There are several techniques to mitigate these ambiguities and the reader is referred to them for further reading (e.g., Bharadwaj and Chandrasekar, 2007; Doviak and Zrnić, 2006; Cho, 2005; Sachidananda and Zrnić, 2003; Torres et al., 2004).

Volume Scattering

Up to this point, we have discussed scattering from a single scatterer – both stationary and moving. This theory is fundamentally the same as what underpins radar used to remotely detect aircraft, ships, missiles, planets, etc. Whereas most radars are designed to detect and characterize a countable number of discrete objects (i.e., aircraft, automobile, etc.), weather radars need to characterize a countless number of hydrometeors distributed over vast volumes of the atmosphere. The echo power from each hydrometeor is very weak – a 1 mm raindrop at $r = 100$ km returns a power less than 10^{-22} W. Although each scatterer's echo is weak and usually not detectable, the sum of echo powers from an extremely large number of hydrometeors can return a strong continuum of echoes as the transmitted pulse propagates through the field of hydrometeors.

Polarimetric Doppler Weather Radars are used to estimate and map the hydrometeors' fields of reflectivity factor (proportional to the number density and size of the hydrometeors), reflectivity-weighted mean radial velocity, and polarimetric quantities. From these fields, radar meteorologists derive estimates of the fall rate and accumulation of precipitation, classify types of scatterers, and resolve storm hazards.

Other types of atmospheric radars such as wind profilers are designed to measure scatter from refractive index perturbations – caused by minute temperature and humidity variations – to measure winds in the lower atmosphere. Echoes from refractive index perturbations are called Bragg scatter (Doviak and Zrnić, 2006, Chapter 11). Mesosphere, stratosphere, and troposphere radars use large antennas (e.g., 100×100 m²), long wavelengths (i.e., $\gg 1$ m), and high average power (e.g., few tens of kilowatts) to detect Bragg scatter from the weak refractive index perturbations in the upper atmosphere (Fukao and Hamazu, 2013). Polarimetric Doppler Weather Radar can also detect Bragg scatter, but observations are limited to the troposphere, and conditions of strong refractive index perturbations.

After a delay – the time for the transmitted pulse to propagate to the near boundary of the volume of hydrometeors plus time for the first echoes to arrive at the radar – echoes are continuously received (Figure 5) during a time interval equal to twice the time it takes the microwave pulse to propagate across the volume of hydrometeors. Because one cannot resolve the echo from each scatterer, the radar's digital processors use ADCs to sample – at uniformly spaced T_s

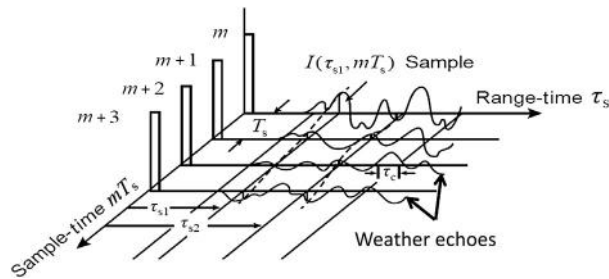


Figure 5 Idealized traces for $I(\tau_s, mT_s)$ of weather echoes from a distribution of scatterers. The traces represent $I(\tau_s, mT_s)$ for the m th T_s interval; dashed traces connect samples (i.e., the vertical line segments) at each τ_s .

intervals – I and Q weather echo amplitudes and to convert these to numbers.

Definition of the Resolution Volume

For each sample in range time ($\tau_s = \tau_{s1}, \tau_{s2}, \tau_{s3}, \dots$), there is a resolution volume V_6 , at range $c\tau_s/2$, which is enclosed by the surface on which the angular and range weighting functions (determined by θ_1 , τ_v , and the receiver's bandwidth B_6) are $1/4$ their peak value (i.e., 6 dB below the peak); the surface has an angular width θ_1 , and, for receiver bandwidth $B_6 > \tau_v^{-1}$, a range width about equal to $c\tau_v/2$. The scatterers within V_6 are those that significantly contribute to the sampled voltage at delay τ_s (Figure 5); those outside V_6 usually make significantly smaller contributions.

The Echo Voltage

The echo voltage sample $V_{hh}(\tau_s, mT_s)$ from a volume of scatterers is the summation of the echo voltage from each scatterer. Thus it can be shown the baseband echo is

$$V_{hh}(\tau_s, mT_s) = \sum_n A^{(n)} s_{hh}^{(n)} \exp \left\{ j \left[-\frac{4\pi r^{(n)}}{\lambda} - \frac{4\pi v_r^{(n)}}{\lambda} mT_s + \psi \right] \right\} \quad [7]$$

where the superscript (n) denotes the n th scatterer at an initial range $r^{(n)}$. ADCs obtain digital samples of echoes from several hundred volumes along the beam. Examples of $I(\tau_s, mT_s)$ and $Q(\tau_s, mT_s)$ time series for echoes from scatterers in two independent V_6 s are shown as a function of index m in Figure 6. One V_6 returns echoes from quasi-stationary scatterers (e.g., a cell tower, trees, etc.); these echoes (called ground clutter because, for weather radar applications, it is unwanted and clutters signals from weather) produce slowly varying voltages. The second V_6 contains hydrometeors having radial velocities causing rapid variations of the echo voltage.

Spectral Analysis of the Echo Voltage

Temporal variations of the echo voltage in sample time mT_s hold important Doppler information about the scatterers within the resolution volume. For each range time, τ_s , echo voltage samples from M transmitted pulses and for each polarization are compiled over the *dwell time* (MT_s , the time to transmit all M pulses used in the analysis) chosen to investigate these variations. This so-called *spectral analysis* is usually accomplished using the discrete Fourier transform of the echo voltage. The main product of the Fourier transform is the power spectral density – also called the *Doppler spectrum* in the weather radar community. The Doppler spectrum provides the reflectivity-weighted distribution of radial velocities of scatterers mostly within the resolution volume. Examples of Doppler spectra are provided in Figure 6 corresponding to the series of M echo voltage samples from two V_6 s. Doppler spectra can be obtained for both H and V polarization, providing the possibility of velocity-dependent polarimetric characterization.

For the ground clutter case shown in Figure 6, the Doppler spectrum shows a prominent peak at zero Doppler velocity

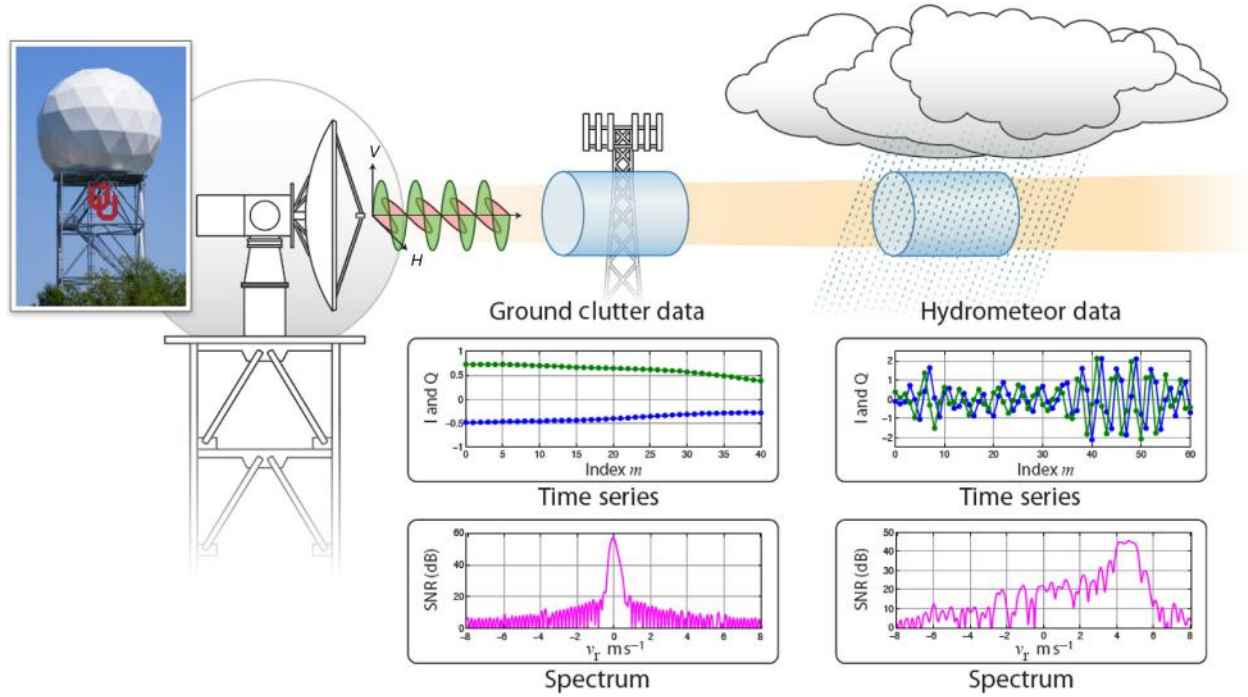


Figure 6 OU-PRIME weather radar showing idealized H and V waves propagating along the beam and two V_6 s (i.e., quasi-cylindrical resolution volumes), and time series traces $I(\tau_s, mT_s)$ and $Q(\tau_s, mT_s)$ from the two V_6 s, one in a region of a ground scatterer – a tower or trees that can sway causing I and Q to change slowly with mT_s – and the other in precipitation having moving scatterers. Each trace is a sequence of echo samples (dots) spaced mT_s ; the green trace is I and the blue trace is Q . Also shown are the Doppler spectra of echo samples from the two V_6 s.

because ground objects are mostly stationary. For weather, the peak of the Doppler spectrum estimates the heaviest weighted mean radial velocity of hydrometeors. More interestingly, the shape of the spectrum holds a wealth of information about inhomogeneities within the resolution volume, such as wind shear, multiple hydrometeor types, etc. This information can be quite useful when studying phenomena such as tornadoes, squall lines, microbursts, or any weather events with a high degree of structure.

Spectral Moments and Polarimetric Products

Neither echo voltage nor Doppler spectra are typically recorded or displayed in real time; rather, digitized samples from both H and V polarizations (i.e., $V_{hh}(\tau_s, mT_s)$ and $V_{vv}(\tau_s, mT_s)$) are collected over the desired dwell time and sent to the radar's signal processor which extracts estimates of spectral moments and polarimetric parameters. Due to the typically uniform distribution of reflectivity, turbulence, and mean wind shear within V_6 , and the Gaussian shape of the beam's illumination, the Doppler spectrum often has a Gaussian shape. In this case, only three parameters completely describe the spectrum – the zeroth, first, and second moments. The zeroth moment is the area under the Doppler spectrum and is proportional to the backscattered power for each polarized wave from a particular V_6 . The first moment (i.e., the mean radial velocity $E[v_r]$) is the integral of the product of the Doppler velocities multiplied by the power spectrum normalized by the total spectral power, essentially providing the 'center of mass' of the Doppler

spectrum. The square root of the second moment of the Doppler spectrum is the spectrum width, σ_v , which is proportional to the spread of v_r within V_6 ; thus, it is a measure of turbulence and mean wind shear. It should be noted that separate moment estimates are obtained for both polarizations. Although the zeroth moments can differ significantly, as will be described below, the radial velocity and spectrum width do not have significant difference for H and V polarized waves.

Through careful calibration, the copolar backscattering matrix elements (s_{hh} and s_{vv}) can be derived from the complex echo signals with H and V polarizations. In turn, s_{hh} and s_{vv} are used to calculate important polarimetric products. The zeroth moment is used to calculate the H and V reflectivity factors (Doviak and Zmíć, 2006, section 8.5.2.2),

$$Z_h = \left(\frac{4\lambda^4}{\pi^4 |K_w|^2} \right) n(\vec{r}) \langle |s_{hh}|^2 \rangle \quad [8a]$$

$$Z_v = \left(\frac{4\lambda^4}{\pi^4 |K_w|^2} \right) n(\vec{r}) \langle |s_{vv}|^2 \rangle \quad [8b]$$

where $n(\vec{r})$ is the density of scatterers, and K_w is the complex dielectric factor for water. The angle brackets in eqn [8] indicate averaging over the ensemble of drop shapes, sizes, and canting angles (alternatively averaging over mT_s); for spherical hydrometeors $\langle |s_{hh}|^2 \rangle = \langle |s_{vv}|^2 \rangle$. For larger drops, ice crystals, hail stones, etc., these polarimetric parameters can be quite different providing information about the shape (e.g., Figure 1) of the hydrometeors. To quantify this

difference, the logarithm of the ratio of Z_h and Z_v is taken to produce

$$Z_{DR} \equiv 10 \log_{10} \left(\frac{Z_h}{Z_v} \right) = 10 \log_{10} \left(\frac{\langle |s_{hh}|^2 \rangle}{\langle |s_{vv}|^2 \rangle} \right) \quad [9]$$

Because s_{hh} is expected to be larger than s_{vv} for drops deformed by drag forces as shown in Figure 1, typical values of Z_{DR} for rain varies from 0 to +4 dB. Snow and ice crystals typically orient themselves horizontally, have large variation in refractive index, and can have large aspect ratios, so Z_{DR} varies from 0 to +5 dB, depending on crystal habit.

Another important polarimetric product is the complex correlation coefficient

$$\rho_{hv}(nT_s) = \frac{\langle s_{vv}([m+n]T_s) s_{hh}^*(mT_s) \rangle}{\langle |s_{vv}(mT)|^2 \rangle^{1/2} \langle |s_{hh}(mT)|^2 \rangle^{1/2}} \quad [10]$$

a measure of the similarities of the time changes in s_{hh} and s_{vv} . The parameter that is most commonly displayed is $\rho_{hv}(0)$, the magnitude at zero lag (typically the magnitude symbols $|x|$ are not displayed). $\rho_{hv}(0)$ is directly computed for STSR mode, but requires further processing and assumptions for the ATSR mode. Because $\rho_{hv}(0)$ is a normalized variable, it is independent of radar calibration. For spherical hydrometeors, $\rho_{hv}(0)$ should be close to unity but can be affected by the quality of the radar hardware (e.g., reflector, feed, etc.) and by the chosen polarimetric data acquisition mode. For more complex scatterer shapes (e.g., tornado debris), $\rho_{hv}(0)$ is reduced from

unity – a fact that can be used for hydrometeor classification and tornado detection, for example.

As stated earlier, we have ignored propagation effects when describing polarimetric products, and have focused on only backscattering effects. Under certain conditions and depending on the radar wavelength, propagation effects can be important. The interested reader is referred to Doviak and Znić (2006) or Bringi and Chandrasekar (2001) for a more complete coverage of the topic.

Examples of Polarimetric Doppler Radar Data

Near the National Weather Center on its research campus, the University of Oklahoma (OU) has a high-resolution, 5-cm-wavelength polarimetric Doppler radar called OU-PRIME for *Polarimetric Radar for Innovations in Meteorology and Engineering* (Palmer et al., 2011, also see Figure 6 insert). The name was coined to emphasize the fact that the radar is used for collaborative research and education in both science and engineering.

On 10 May 2010, a tornado outbreak occurred in Oklahoma producing numerous strong tornadoes. Unfortunately, three deaths were reported along with significant property damage. Several of the more powerful tornadoes formed very near OU-PRIME (less than 1 km) providing polarimetric radar data of unprecedented resolution and quality. An example of the data is provided in Figure 7 in what is called a *plan position indicator* display, where the radar beam is kept at a constant

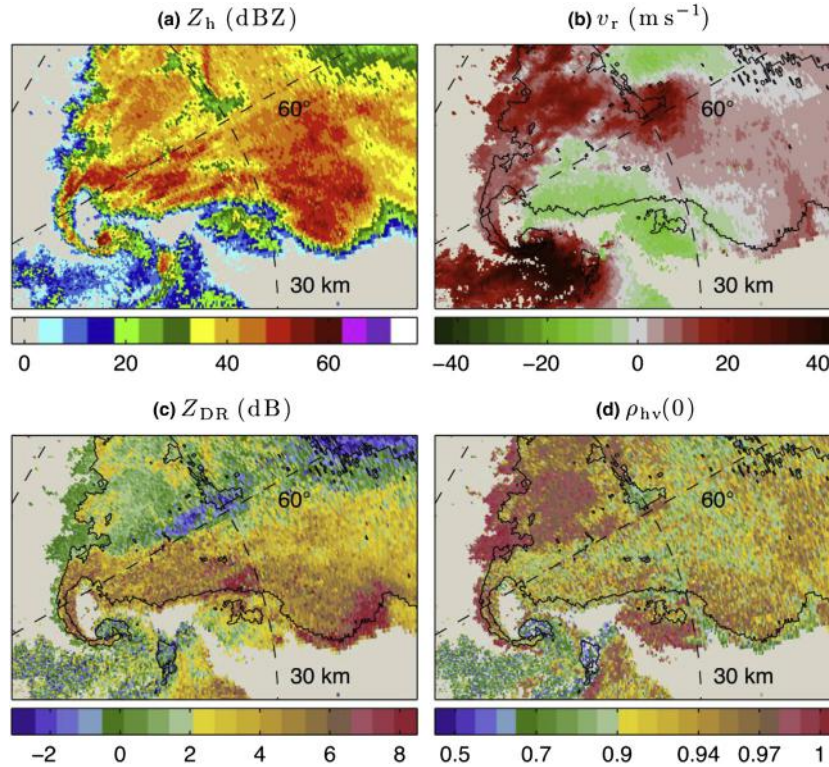


Figure 7 OU-PRIME measurements of (a) reflectivity factor, Z_h , (b) radial velocity, v_r , (c) differential reflectivity, Z_{DR} , and (d) correlation coefficient, $\rho_{hv}(0)$ at 1.0° elevation angle from the sector scan started at 2245 UTC on 10 May 2010. The radar is located off the lower left of the figure and a 30-km range ring provided for orientation. On panels b, c, and d, a 35-dBz contour line is provided to show features of the various parameters relative to the Z_h field.

elevation angle while rotating in azimuth. For the 10 May case, the four major products of a polarimetric radar are Z_{hr} , v_r , Z_{DR} , and $\rho_{hv}(0)$. The most prominent feature of the reflectivity data (panel a) is the spiral of high reflectivity, the so-called hook echo often associated with a tornado. The tornado itself is likely located at the tip of the 'hook' where there is a small blob of high reflectivity (likely due to tornadic debris) with an 'eye' of lower reflectivity. The radial velocity field (panel b) shows the well-known velocity couplet of out- ($+v_r$) and inbound ($-v_r$) radial velocities (red and green, respectively) caused by the rotation of the tornado. Z_{DR} data (panel c) exhibit very high values (red area, the so-called Z_{DR} Arc; Kumjain and Ryzhkov, 2008), indicative of large drops near the storm's eastern edge. The correlation coefficient field $\rho_{hv}(0)$, shown in panel (d), illustrates the classic debris signature of low $\rho_{hv}(0)$ values near 0.6 at the tip of the hook echo (Ryzhkov et al., 2002).

Through the combination of the various signatures of the polarimetric data, it is possible to greatly improve the performance of tornado detection algorithms and hydrometeor classification techniques. After extensive research with polarimetric radar, operational WSR-88Ds have been recently upgraded to have polarimetric capability. With this advent, new scientific applications are sure to be developed over the coming years.

Future Weather Radar – Phased Array Antennas

The history of weather radar is replete with examples of significant steps forward in technology because of developments for military applications. Prime examples include advances in transmitter technology (e.g., magnetron) and Doppler capabilities. The future of weather radar will very likely show a continuation of this trend.

One technology, which has been used in military radar for decades and is just now being tested in the weather radar community, is electronic beam scanning using *phased array antennas* (Weber et al., 2007; Zrnić et al., 2007). As previously described, conventional weather radars provide complete coverage of storms by mechanically rotating a parabolic reflector antenna. Since the antenna is a very large and heavy structure, it is typically rotated continuously in azimuth and stepped in elevation without rapid starts or stops of the antenna. By doing so, the mechanical stresses on the gears and motors are minimized, prolonging the life of the entire system and reducing maintenance costs. Phased array antennas operate in a completely different way without any mechanical motion.

Each face of a phased array antenna is made up of thousands of individual radiating elements spread over the quasi-circular aperture of the array, which is usually a planar surface (Figure 8). The beam formed by each face azimuthally scans $\pm 45^\circ$ about a vertical plane containing the broadside direction (i.e., the line perpendicular to the array surface). To observe a full 360° coverage, four phased array panels are needed. Each element of the array is a transmitter and receiver and each has a commanded amplitude and phase; the amplitude distribution across the aperture controls the beamwidth and sidelobe levels, and the phase

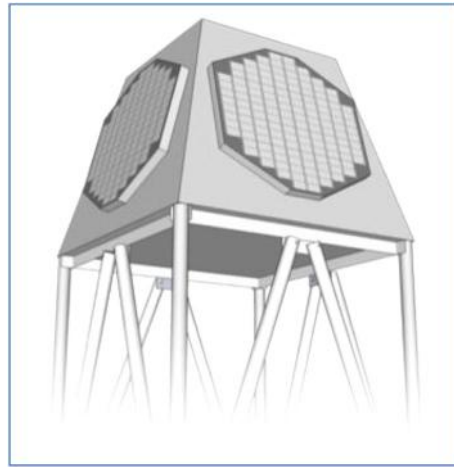


Figure 8 Artist's depiction of a four-panel planar phased array weather radar (PAR).

distribution controls the direction of the beam. If the phase is uniform across the array, the surface of constant phase of the radiated field is perpendicular to the broadside direction and the beam is along the broadside direction. If the phases of the array elements are adjusted so that the surface of constant phase is tilted relative to the array face, the beam will be pointed in a direction perpendicular to the constant phase surface.

These commanded phase shifts can be generated and changed electronically from pulse to pulse allowing extremely rapid beam steering with no mechanical motion of the antenna. Furthermore, the beam agility provided by electronic steering allows adaptive scanning so the beam needs only scan stormy regions of interest rather than having equal dwell times on regions of clear skies (Heinselman and Torres, 2011). Such a capability for weather PAR exists with the National Weather Radar Testbed – located on the campus of the OU – having a phased array antenna on loan from the US Navy, and a WSR-88D transmitter.

Because of this capability, phased array weather radars (PARs) hold promise for improved observations of rapidly evolving severe weather phenomena. Furthermore, a beam from the same panel but at a different frequency can be pointed in different directions allowing multimission applications (e.g., a single antenna could be used to both observe weather and track aircraft). Also PAR has the capability to operate with more than one beam thus increasing the update rate of data to improve forecasts and warnings (Stensrud et al., 2009; Dawson et al., 2012; Yussouf et al., 2013; Doviak and Zhang, 2013).

As with any new technology, significant challenges exist, such as combining radar polarimetry with phased array antennas. Most military radars use single polarized waves whereas the US NWS has accepted radar with dual polarization capability to better measure rainfall, classify precipitation, and issue warnings. Thus, one such challenge is to develop a polarimetric phased array radar (PPAR) that has equal or better capability than polarimetric weather radars using parabolic reflector antennas. The PPAR (Figure 8) has beamwidths

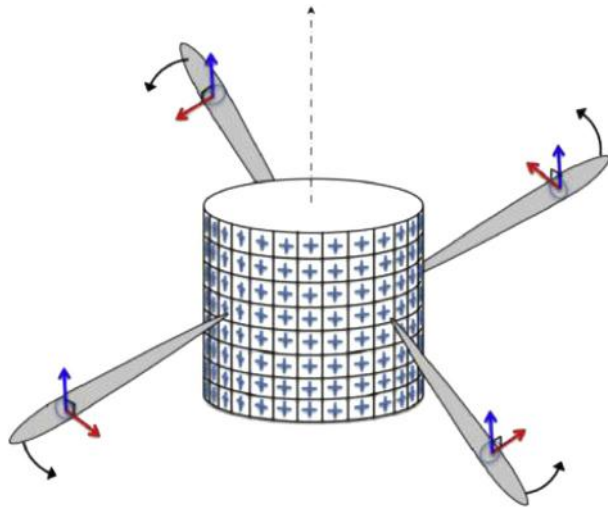


Figure 9 Artist's depiction of CPPAR with pairs of crossed dipoles for each array element (from Zhang G., Doviak, R.J., Zrnić, D.S., Palmer, R.D., Lei, L., Al-Rashid, Y., 2011. Polarimetric phased-array radar for weather measurement: a planar or cylindrical configuration? *J. Atmos. Oceanic Technol.*, 28, 63–73.). Shown in red (blue) is the $H(V)$ polarized wave and four simultaneous beams are assumed. Beams at other frequencies can be added and steered independently for simultaneous measurements in other missions; thus forming a multi-mission PAR.

and polarimetric wave properties that depend on the direction of the beam and calibration is much more complicated than needed for the existing weather radars (Zhang et al., 2009) The development of PPAR for quantitative weather measurements opens many research opportunities to overcome some of these complexities.

A solution under study is the use of a cylindrical polarimetric phased array radar (CPPAR, Figure 9; Zhang et al., 2011). With this phase array antenna the phases of the elements in a quadrant are adjusted to produce a plane surface of constant phase so that the beamwidth and polarimetric properties are the same independent of azimuth angle. By commutating the amplitude and phase commands around the periphery of the cylinder, the beam scans in azimuth without changes in beamwidth or polarimetric properties. Nevertheless, there are changes in these properties when the beam is elevated, but for weather applications the most important beam directions are those close to the ground and perpendicular to the axis of the cylinder. It remains to be shown that PPAR or CPPAR can exploit their rapid scanning capability while providing weather data with the same quality presently achieved with reflector antennas. This notwithstanding, it is possible that over the next decades PAR will be fully developed and used for operational networks around the world.

Acknowledgment

The authors express their appreciation to Dr Boon Leng Cheong, of the Advanced Radar Research Center, University of Oklahoma, who provided the figures.

See also: Radar: Mesosphere–Stratosphere–Troposphere and Stratosphere–Troposphere Radars and Wind Profilers; Meteor Radar; Precipitation Radar. **Weather Forecasting:** Severe Weather Forecasting.

Further Reading

- Bharadwaj, N., Chandrasekar, V., 2007. Phase coding for range ambiguity mitigation in dual-polarized Doppler weather radars. *Journal of Atmospheric and Oceanic Technology* 24 (8), 1351–1363.
- Bringi, V.N., Chandrasekar, V., 2001. *Polarimetric Doppler Weather Radar: Principles and Applications*. Cambridge Press, Cambridge, New York.
- Chandrasekar, V., Bharadwaj, N., 2009. Orthogonal channel coding for simultaneous co- and cross-polarization measurements. *Journal of Atmospheric and Oceanic Technology* 26, 45–56.
- Cho, J.Y.N., 2005. Multi-pri signal processing for the terminal Doppler weather radar. Part II: range–velocity ambiguity mitigation. *Journal of Atmospheric and Oceanic Technology* 22 (10), 1507–1519.
- Dawson, D.T., Wicker, L.J., Mansell, E.R., Tanamachi, R.L., 2012. Impact of the environmental low-level wind profile on ensemble forecasts of the 4 May 2007 Greensburg, Kansas, tornado storm and associated mesocyclone. *Monthly Weather Review* 140, 696–716.
- Doviak, R.J., Zrnić, D.S., 2006. *Doppler Radar and Weather Observations*, second ed., 3rd Prt. Dover Publications, Mineola, New York.
- Doviak, R.J., Zhang, G., 2013. Short course on phased array radar polarimetry. In: 36th Conference on Radar Meteorology of American Meteorological Society. Boston, MA.
- Doviak, R.J., Bringi, V., Ryzhkov, A., Zahrai, A., Zrnić, D.S., 2000. Considerations for polarimetric upgrades to operational WSR-88D radars. *Journal of Atmospheric and Oceanic Technology* 17, 257–278.
- Fukao, S., Hamazu, K., 2013. *Radar for Meteorological and Atmospheric Observations*. Springer, Japan. <http://dx.doi.org/10.1007/978-4-431-52334-3>. ISBN: 978-4-431-54334-3 (eBook).
- Heinselman, P., Torres, S., 2011. High-temporal-resolution capabilities of the National Weather Radar Testbed Phased-Array Radar. *Journal of Applied Meteorology and Climatology* 50 (3), 579–593. <http://dx.doi.org/10.1175/2010JAMC2588.1>.
- Kumjian, M.R., Ryzhkov, A., 2008. Polarimetric signatures in supercell thunderstorms. *Journal of Applied Meteorology and Climatology* 48, 1940–1961.
- Palmer, R.D., Bodine, D., Kumjian, M., Cheong, B., Zhang, G., Cao, Q., Bluestein, H.B., Ryzhkov, A., Yu, T., Wang, Y., 2011. Observations of the 10 May 2010 tornado outbreak using OU-PRIME: potential for new science with high-resolution polarimetric radar. *Bulletin of the American Meteorological Society* 92 (7), 871–891.
- Pruppacher, H.R., Beard, K.V., 1970. A wind tunnel investigation of the internal circulations and shape of water drops falling at terminal velocity in air. *Quarterly Journal of the Royal Meteorological Society* 96, 247–256.
- Ryzhkov, A., Burgess, D., Zrnić, D., Smith, T., Giangrande, S., 2002. Polarimetric analysis of a 3 May 1999 tornado. Preprints, 22nd Conf. on Severe Local Storms, Hyannis, MA American Meteorological Society 14.2. Available online at: <http://ams.confex.com/ams/pdfpapers/47348.pdf>.
- Sachidananda, M., Zrnić, D.S., 2003. Unambiguous range extension by overlay resolution in staggered PRT technique. *Journal of Atmospheric and Oceanic Technology* 20 (5), 673–684.
- Stensrud, D.J., Xue, M., Wicker, J.L., Kelleher, K.E., Foster, M.P., Schaefer, J.T., Schneider, R.S., Benjamin, S.G., Weygandt, S.S., Ferree, J.T., Tuell, J.P., 2009. Convective-scale warn-on-forecast system. *Bulletin of the American Meteorological Society* 90 (10), 1487–1499.
- Torres, S.M., Dubel, Y.F., Zrnić, D., 2004. Design, implementation, and demonstration of a staggered PRT algorithm for the WSR-88D. *Journal of Atmospheric and Oceanic Technology* 21 (9), 1389–1399.
- Weber, M.E., Cho, J.Y.N., Herd, J.S., Flavin, J.M., Benner, W.E., Torok, G.S., 2007. The next-generation multi-mission U.S. surveillance radar network. *Bulletin of the American Meteorological Society* 88 (11), 1739–1751.
- Yussouf, N., Mansell, E.R., Wicker, L.J., Wheatley, D.M., Stensrud, D.J., 2013. The ensemble Kalman filter analyses and forecasts of the 8 May 2003 Oklahoma city tornadic supercell storm using single—and double—moment microphysics schemes. *Monthly Weather Review* 141, 3388–3412.

Zhang, G., Doviak, R.J., Zrníc, D.S., Crain, J., Stainman, D., Al-Rashid, Y., 2009. Phased array radar polarimetry for weather sensing: a theoretical formulation for bias correction. *IEEE Transactions on Geoscience & Remote Sensing* 47, 3679–3689.

Zhang, G., Doviak, R.J., Zrníc, D.S., Palmer, R.D., Lei, L., Al-Rashid, Y., 2011. Polarimetric phased-array radar for weather measurement: a planar or cylindrical configuration? *Journal of Atmospheric and Oceanic Technology* 28, 63–73.

Zrníc, D., Kimpel, J.F., Forsyth, D.E., Shapiro, A., Crain, G., Ferek, R., Heimmer, J., Benner, W., McNellis, T.J., Vogt, R.J., 2007. Agile-beam phased array radar for weather observations. *Bulletin of the American Meteorological Society* 88 (11), 1753–1766.

This page intentionally left blank

Precipitation Radar

SE Yuter, North Carolina State University, Raleigh, NC, USA

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

Precipitation radars scan their vicinity to determine the location, intensity, and structure of showers and storms. Several different radar variables are used to determine different aspects of the precipitation. Radar reflectivity measures the number and size of particles. In radars that transmit and receive horizontally as well as vertically polarized energy, differential reflectivity measures the shape of particles and specific differential phase measures the associated liquid water content. Information from the different variables is used to determine hydrometeor type and to estimate rain rate.

Introduction

Precipitation radars are widely used to determine the location, size, and intensity of precipitating storms. Ground-based scanning precipitation radars are used in short-term weather and flood forecasting, and to estimate the distribution and amount of cumulative rainfall over a region. The weather services of many countries have networks of operational radars that monitor precipitation near population centers. The output from these operational radar networks can be combined to provide a picture of the distribution of precipitation over synoptic-scale regions. Mobile precipitation radar on aircraft provides pilots with information to navigate safely around dangerous regions of heavy rain, hail, and turbulence. Precipitation radars are also used to map the three-dimensional structure of storms. Spaceborne precipitation radar on low-orbit satellite maps the structure and distribution of precipitation around the globe over periods of months and years.

The British and Americans first developed weather radar during World War II. The precipitation radar transmits a pulse of electromagnetic energy via an antenna. When the transmitted energy encounters a target, such as a raindrop, part of the transmitted energy is scattered back toward the antenna, where it is received and amplified. The time delay between the original pulse transmission and the receipt of the backscattered energy is used to deduce the distance to the reflector. The relationship between the range-corrected, backscattered, returned power and the size and number of the reflecting targets is the physical foundation for interpreting precipitation radar data.

The frequency of electromagnetic waves f in Hz (s^{-1}) is defined as $f = c/\lambda$, where c is the speed of light and λ is the wavelength in meters. Radar frequencies are divided into several

bands as shown in Table 1. The choice of frequency for precipitation radar is a tradeoff between the practical constraints of size, weight, cost, and the relation between the wavelength and the size of the target hydrometeors. Theoretical considerations favor the choice of the longer S-band and C-band wavelengths for many precipitation applications. However, use of these longer wavelengths is not always practical. The beam width for circular antennas is proportional to λ/d , where d is the antenna diameter. Longer wavelengths necessitate a larger antenna to obtain a focused beam of the same angular beam width. Larger antennas are heavier, require more powerful motors to rotate them, and are more expensive than smaller antennas. Shorter X, K_u , and K_a band wavelengths are often utilized in mobile precipitation radars deployed on spacecraft, aircraft, ships, and trucks where size and weight are more constrained as compared to stationary ground-based radars.

Precipitation Radar Components

The precipitation radar consists of a transmitter, receiver, transmit–receive switch, antenna, and display (Figure 1). In this simplified diagram, the processing of the electromagnetic signals into output suitable for display is included in the display block. A phase detector is often included to measure Doppler velocity. The radar transmitter contains a modulator that switches the transmitter on and off to form discrete pulses. The radar sends out a pulse and then switches to the receiver to “listen” for echoes from precipitation and other atmospheric targets. The range to the targets is obtained by comparing the time of pulse transmission to the time the backscattered pulse is received. In precipitation radars, the pulses are transmitted at a pulse repetition frequency (PRF) of ~ 300 – 1300 Hz and each

Table 1 Precipitation radar frequencies and wavelengths

Band designation	Nominal frequency (GHz)	Nominal wavelength (cm)	Applications
S	2–4	15–8	Surface-based radars
C	4–8	8–4	Mobile and surface-based radars
X	8–12	4–2.5	Mobile and surface-based radars
K_u	12–18	2.5–1.7	Mobile and spaceborne radars
K_a	27–40	1.1–0.75	Mobile and spaceborne radars

Adapted from Skolnik, M.I., (Ed.), 1990. Radar Handbook. McGraw-Hill, New York and Rinehart, R.E., 2010. Radar for Meteorologists, fifth ed. Ronald E. Rinehart, Fargo.

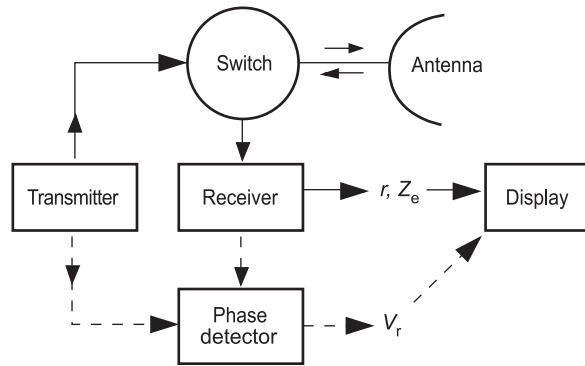


Figure 1 Simplified hardware block diagram of a precipitation radar. The non-Doppler portion of the system yields the range r and equivalent reflectivity factor (Z_e) of the target. Dashed lines connect parts of the system included in a Doppler radar that additionally measures the radial velocity of the target (V_r). Adapted from Houze, R.A., 1993. *Cloud Dynamics*. Academic Press, New York, p. 108, © Elsevier.

pulse is on the order of $1 \mu\text{s}$ (10^{-6} s) in duration. The time between transmitted pulses limits the maximum range (r_{max}) the electromagnetic pulse can travel before the next pulse is transmitted.

$$r_{\text{max}} = \frac{c}{2\text{PRF}} \quad [1]$$

The receiver detects and amplifies the received signals and averages the characteristics of the returned pulses over defined time periods. Typical peak transmitted power for operational precipitation radars is 10^5 – 10^6 W. Typical received power is 10^{-10} W. The transmit–receive switch protects the sensitive receiver from the powerful transmitter. Without the switch, the radar transmitter would burn out the receiver. In practice, the transmit–receive switch is not perfect and a small amount of transmitted energy leaks into the receiver.

Radar antennas focus transmitted energy and direct it along a narrow angular beam. For scanning radars, this direction is often described in terms of an elevation angle relative to the ground and an azimuth angle relative to north. Moving the antenna points the axis of the beam in different directions, and permits scanning of two- and three-dimensional regions of the atmosphere. The antenna shape and radar wavelength determines the radar beam size and shape. The magnitude of the radar energy is largest along the center of the beam and decreases outward with increasing angular width. The beam width is defined as the angular width where the power is exactly half the maximum power. Most precipitation radars utilize a circular parabolic antenna for both transmission and reception.

The main purpose of the display is to distinguish scatterers at different ranges. A basic scanning radar display will usually indicate the compass angle and range to the radar echo in polar coordinates. **Figure 2** shows the range-corrected received power as a function of range along a single pointing direction of the antenna. This example illustrates that not all the energy received at the radar is backscattered from meteorological targets. Transmitter leakage and ground clutter from nearby nonmeteorological targets such as trees and buildings are present at close ranges in the display. The signal from a point

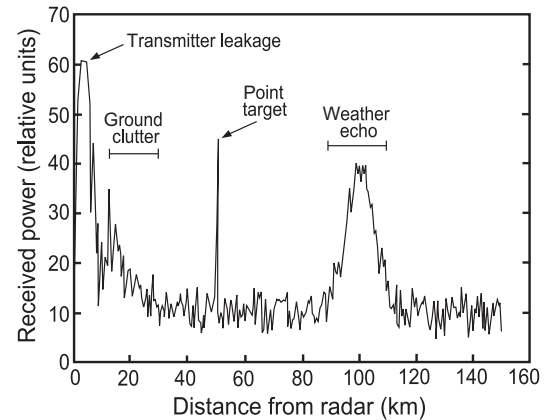


Figure 2 Two-dimensional radar display of range-corrected received power as a function of range showing signal from transmitter leakage, ground clutter, point target, and weather echo. See text for further details. Adapted from Rinehart, R.E., 2010. *Radar for Meteorologists*, fifth ed. Ronald E. Rinehart, Fargo, with permission.

target such as a radio tower is present at 50 km range. The wider signal associated with meteorological echo is present between 90 and 115 km range.

The Radar Equation

The radar equation expresses the relationship between the transmitted power and backscattered received power from precipitation targets in terms of the radar's hardware characteristics and the distance between the transmitter and the target. In this section, the radar equation and radar reflectivity will be derived by first making some simplifying assumptions and then gradually refining the terms to more accurately represent the electromagnetic theory underpinning precipitation radars.

Isolated Scatterers

The amount of power incident (P_i) at an isotropic target of cross-sectional area A_t at range r_1 from an isotropic transmitter is

$$P_i = \frac{P_t A_t}{4\pi r_1^2} \quad [2]$$

in which P_t is the transmitted power (**Figure 3(a)**). Transmitted power and received power are commonly expressed in units of watts or dBm. The latter represents the ratio of power (P) in watts relative to 1 mW ($10 \log_{10}[P/10^{-3} \text{ W}]$). In typical operational precipitation radars, the minimum detectable signal is approximately -100 dBm and the peak-transmitted power is ~ 90 dBm.

Assuming the target does not absorb any power and radiates the energy it receives isotropically, the power received at range r_2 from the target by a receiver of effective cross-sectional area A_e (**Figure 3(a)**) is

$$P_r = \frac{P_t A_t}{4\pi r_1^2} \left(\frac{A_e}{4\pi r_2^2} \right) \quad [3]$$

When the transmitted energy is focused with an antenna, the ratio of the power per unit area along the axis of the focused

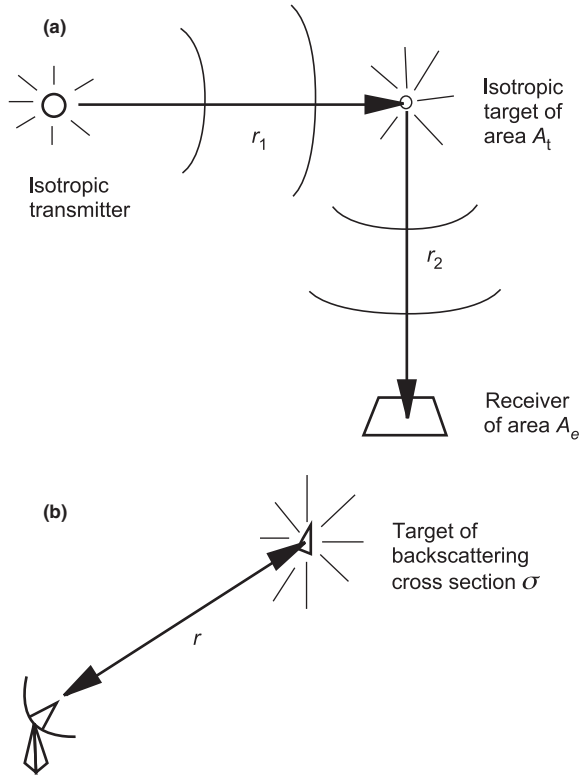


Figure 3 (a) Schematic of energy from an isotropic transmitter (P_t) traveling distance r_1 to an isotropic target of area A_t . The energy received at the target (P_r) is then isotropically radiated by the target and travels distance r_2 to a receiver of area A_e . (b) In precipitation radars, an antenna is used to focus the energy into a particular direction. The received energy is scattered by the target at range r . The backscattered portion of the energy corresponding to cross section σ is received by the antenna.

beam to the power per unit area of an isotropic transmitter is a measure of antenna gain (g). For a circular, parabolic antenna typically used on precipitation radars, the antenna gain can be approximated as a function of the horizontal (θ_H) and vertical (θ_V) beam widths in radians, or for the receiving antenna as a function of the radar wavelength and A_e .

$$g \approx \frac{\pi^2}{\theta_H \theta_V} \approx \frac{4\pi A_e}{\lambda^2} \quad [4]$$

The antenna gain is usually of large magnitude and is often expressed as the ratio relative to an isotropic antenna in decibel units, $G = 10 \log_{10}(g/1)$. For example, the antenna gain specification is 45 dB for the US National Weather Service WSR-88D operational radars, which means the antenna focuses energy about 30 000 times better than an isotropic transmitter. The use of a directional antenna for transmission leads to modification of the term representing transmitted power in eqn [3] from P_t to $g_t P_t$, where g_t is the gain of the transmitting antenna. Usually, precipitation radars utilize a single directional antenna for both transmitting and receiving which permits the simplifications of $r_1 = r_2 = r$ in eqn [3] and $g_r = g_t = g$.

Real world scatterers are usually not isotropic, and the size of the backscattered cross-sectional area A_t is usually not equal to the physical size of the scatterer. The amount of scattered

energy is dependent not only on the radar wavelength and incident power at the range of the target but also on the target characteristics of size, shape, composition, three-dimensional angle between the target and transmitter, and target velocity. The backscattering cross section (σ) is defined as the apparent area that, if scattered isotropically, would return to the receiver an amount of power equal to the power actually received. Table 2 shows values for σ for diverse types of single scatterers.

Making substitutions for P_t , A_e , and A_t into eqn [3] yields the equation for the power received by a precipitation radar with a directional antenna from a single target of backscattering cross section σ at range r (Figure 3b).

$$P_r = \frac{P_t g^2 \lambda^2 \sigma}{(4\pi)^3 r^4} \quad [5]$$

Distributed Scatterers

Precipitation particles such as raindrops, snowflakes, hail, and graupel act as distributed scatterers in the volume of atmosphere illuminated by the precipitation radar. The resolution volume (V_{res}) illuminated by a transmitted pulse along the beam is approximated by a cylinder defined by the beam widths and pulse length of the radar hardware and range to the volume,

$$V_{res} \approx \pi \left(r \frac{\theta_H}{2} \right) \left(r \frac{\theta_V}{2} \right) \frac{c\tau}{2} \quad [6]$$

in which τ is the pulse length in seconds. Antenna beam widths are assumed to be small, such that the small angle approximation $\theta \approx \sin \theta$ is valid. The resolution volume is defined as the incremental volume along the beam from which scattered energy is received simultaneously at the radar. A one-half pulse length factor (i.e., $\tau/2$), rather than τ , is needed since the backscattered energy from the front edge of the pulse at time $t_o + \tau/2$ and range $r_o + c\tau/2$ arrives back at the radar at the same instant as backscattered energy from the trailing edge of the pulse at time $t_o + \tau$ and range r_o . To account for the actual distribution of power within a beam generated by a circular parabolic antenna, a correction factor of $1/(2 \ln 2)$ is applied to eqn [6] yielding

$$V_{res} = \frac{\pi r^2 \theta_H \theta_V c\tau}{16 \ln 2} \quad [7]$$

For operational precipitation radars, typical beam widths of $1-2^\circ$ and pulse lengths of $0.5-1 \mu s$ result in resolution volumes of order $10^7-10^8 m^3$ at 60 km range.

Table 2 Backscattering cross sections at $\lambda = 10$ cm

Object	$\sigma (m^2)$
C-54 aircraft	10–1000
Human	0.14–1.05
Weather balloon, seagull	10^{-2}
Small birds	10^{-3}
Bee, dragonfly	3×10^{-6} to 10^{-7}
Water sphere ($D = 2$ mm)	1.8×10^{-10}
Free electron	8×10^{-30}

Adapted from Doviak, R.J., Zrnic, D.S., 2006. Doppler Radar and Weather Observations, second ed. Dover Publications, New York.

The backscattered signal from a volume of randomly distributed scatterers is the sum of the signals scattered by each of the targets $\sum_{i=0}^n \sigma_i$. The summation of the backscattered cross sections from precipitation scatterers in a unit volume is called the *reflectivity* and is defined as $\sum_{\text{vol}} \sigma_i$. As

individual particles move with respect to one another within the resolution volume, the summation of the signals varies slightly from pulse to pulse. The received power is usually averaged over 50 or more consecutive pulses to yield an estimate that is independent of the fluctuations. The final form of the radar equation implemented in the signal processors of precipitation radars combines eqns [5] and [7] and substitutes the radar reflectivity in the resolution volume, $V_{\text{res}} \sum_{\text{vol}} \sigma_i$, for σ . The radar processor calculates and displays

the averaged received power from the set of scatterers within the resolution volume at a particular range. The equation below is a general form of the radar equation valid for scatterers of all sizes.

$$\bar{P}_r = \frac{P_t g^2 \lambda^2 \theta_H \theta_V c \tau}{1024 (\ln 2) \pi^2 r^2} \sum_{\text{vol}} \sigma_i \quad [8]$$

The 'gate spacing' parameter of the radar processor controls the number and spacing of individual reflectivity estimates along the radar beam. It can be set to a value that combines the signal from several consecutive resolution volumes to increase the effective volume over which the average returned power is computed (eqn [8]). In many radars, the pulse length (τ), PRF, gate spacing, and the number of gates sampled (i.e., maximum range) are variables that can be modified by the radar operator. In scanning radars, the radar operator also specifies the set of elevation angles and range of azimuth angles to be illuminated by the antenna. The specification of the variable radar parameters is called the 'radar scan strategy.'

Association of the Radar Signal to Precipitation Characteristics

Radar Reflectivity of a Volume of Precipitation

To be of value in precipitation studies, the average returned power measured by the radar must be related to the physical characteristics of the precipitation particles within the resolution volume. The backscattering cross section of a single water drop (σ_d) increases monotonically when the diameter D is less than $\sim \lambda/16$ such that

$$\sigma_d = \frac{\pi^5}{\lambda^4} |K|^2 D^6 \quad [9]$$

in which $|K|^2$ is the complex dielectric factor. Equation [9] is referred to as the Rayleigh approximation of the backscattering cross section. When the diameter of the drop, D , is greater than $\lambda/16$ for liquid drops, Mie or optical scattering occurs. In contrast to Rayleigh scattering, under conditions of Mie scattering, the backscattered returned power fluctuates as the size of the scatterer increases (Figure 4).

For precipitation applications, it is preferable to use longer wavelengths to encompass as large a range of raindrop diameters as possible within the scattering regime where the

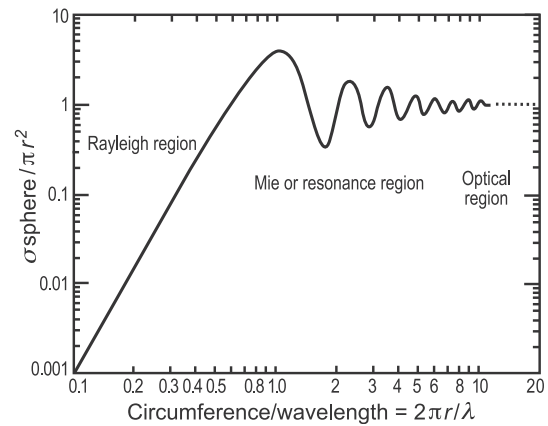


Figure 4 Normalized backscattering cross-sectional area of a perfectly conducting sphere as a function of circumference divided by radar wavelength λ . Since water drops are not perfectly conducting, the transition from Rayleigh to Mie scattering for the radar reflectivity of spherical water drops occurs at $2\pi r/\lambda \sim 0.2$. From Skolnik, M.I., (Ed.), 1990. Radar Handbook. McGraw-Hill Professional, New York, material is reproduced with permission of the McGraw-Hill Companies.

Rayleigh approximation is valid. By definition, raindrops have diameters >0.2 mm. Most rainfall in midlatitudes consists of raindrops <5 mm in diameter. Drops as large as 8 mm in diameter have been observed but are rare since large drops are unstable and tend to break up into smaller drops. Figure 5 shows the subset of raindrop diameters where Rayleigh scattering occurs as a function of wavelength. Drop size distributions (DSDs) associated with several rain rates are shown for comparison. At S-band, rain rates <10 mm h^{-1} are either entirely or nearly entirely within the Rayleigh regime. The proportion of drops within the raindrop size distribution that are in the Mie region increases with increasing rain rate. Side-by-side radars that matched in all characteristics except wavelength would observe different Z values for a given set of raindrops if the larger drops are within the Rayleigh regime for one radar and within the Mie regime for the other radar.

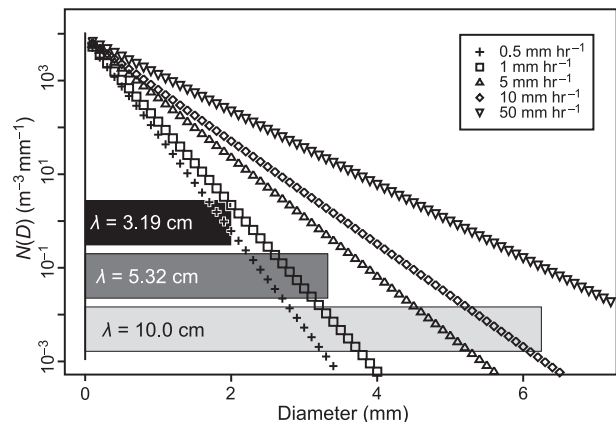


Figure 5 Comparison of reference raindrop size distributions for 0.5, 1, 5, 10 and 50 mm h^{-1} with range of Rayleigh regime diameters (horizontal bar) for S-band (3 GHz, maximum $D = 6.25$ mm), C-band (5.64 GHz, maximum $D = 3.3$ mm), and X-band (9.4 GHz, maximum $D = 2$ mm) radar frequencies.

Replacing σ_i with σ_d in eqn [8] yields the radar equation for spherical drops under conditions when the Rayleigh approximation is valid.

$$\bar{P}_r = \frac{P_t g^2 \lambda^2 \theta_H \theta_V c \tau}{512 (2 \ln 2) \pi^2 r^2} \sum_{\text{vol}} \frac{\pi^5 |K|^2 D_i^6}{\lambda^4} \quad [10]$$

Equation [10] can be rearranged to group the numerical constants and parameters of the radar hardware together to form the 'radar constant' (C).

$$\bar{P}_r = \left(\frac{P_t g^2 \theta_H \theta_V c \tau \pi^3}{1024 \ln 2 \lambda^2} \right) \frac{|K|^2}{r^2} \sum_{\text{vol}} D_i^6 = C \frac{|K|^2}{r^2} \sum_{\text{vol}} D_i^6 \quad [11]$$

The radar reflectivity factor (Z) is defined as

$$Z = \sum_{\text{vol}} D_i^6 = \sum n_i D_i^6 = \int N(D) D^6 dD \quad [12]$$

The discrete form of the definition is used in calculating Z from *in situ* measurements such as aircraft particle probes or disdrometers that resolve n_i , the number of drops per unit volume of atmosphere, in several discrete diameter intervals. In the continuous definition to the right, $N(D)$ is the number of particles per unit volume in the diameter range D to $D + dD$.

Rearranging eqn [11] yields the definition of Z in units of $\text{mm}^6 \text{m}^{-3}$ in terms of variables measured by the radar, and constants associated with the radar hardware and particular gate spacing.

$$Z = \frac{\bar{P}_r r^2}{C |K|^2} \quad [13]$$

In evaluating eqn [13], a value is needed for $|K|^2$. Precipitation particles consist of water, ice, or a combination of the two. The complex dielectric factor $|K|^2$ is 0.93 for water and 0.197 for ice. Within a particular resolution volume of the radar, the individual meteorological scatterers contributing to Z could be composed of water, ice, or melting ice. The scatterers could also include nonmeteorological targets such as insects, birds, and chaff (i.e., highly reflective man-made scatterers such as thin metal strips). There is no current method to be certain of the value of $|K|^2$. The convention is to assume that all the scatterers are composed of liquid water and to define the equivalent radar reflectivity factor Z_e as

$$Z_e = \frac{\bar{P}_r r^2}{C(0.93)} \quad [14]$$

Table 3 Typical values of observed radar reflectivity factor in various types of precipitation

Radar reflectivity factor (dBZ)	Associated precipitation types
−30 to 0	Marginally detectable precipitation
0–10	Drizzle, very light rain, light snow
10–30	Light to moderate rain and heavier snow
30–60	Moderate to heavy rain
40–55	Graupel
55–70	Hail
30–50	Melting snow particles

By this definition, a single spherical drop with a diameter of 1 mm within a 1 m^3 volume of air has $Z_e = 1$. Equivalent radar reflectivity factor is usually expressed in dBZ such that $\text{dBZ}_e = 10 \log_{10}(Z_e/1)$. It is a common usage to refer to the display of dBZ_e values as 'radar reflectivity factor.' It is usually specified in context whether 'radar reflectivity factor' refers to dBZ or $\text{mm}^6 \text{m}^{-3}$ values. The difference between 30 and 20 dBZ is 10 dB following the decibel convention. The typical ranges of radar reflectivity factor values in different types of precipitation are shown in Table 3.

The difference in dielectric constant between liquid and ice, equivalent to 6.7 dB for a particle of the same size and shape, can be used to detect phase transitions in some circumstances. When rain encounters a sufficiently deep low level layer of air below 0°C it freezes into sleet. Figure 6 shows time–height data from a vertically pointing radar for a storm that produced sleet. The phase transition is shown by the ~ 7 dB reduction of radar reflectivity factor along a time varying boundary between 800 and 2000 m altitude.

The physical interpretation of radar reflectivity factor is more complicated in regions of echo containing mixed phase particles than in regions containing exclusively snow or rain. Partially melted ice particles can have different electromagnetic properties compared to equivalent-sized particles of only ice or water. When snow melts within a narrow layer, a reflectivity maximum in the vertical associated with a concentration of partially melted particles can be discerned if the observing radar has sufficient spatial resolution and sensitivity. This local maximum in reflectivity, just below the 0°C level, is often referred to as a 'bright band.' Figure 7 shows how the apparent width of the bright band increases with increasing range and increasing resolution volume size. Nonuniform beamfilling

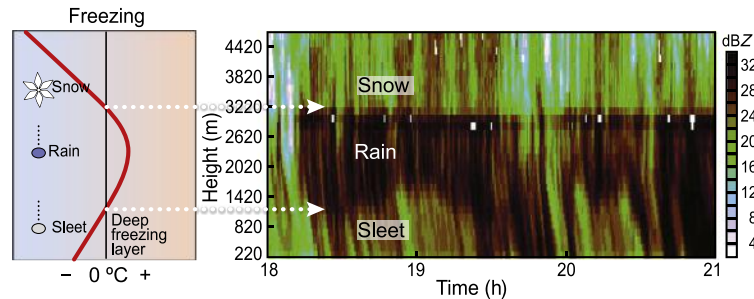


Figure 6 Time–height plot of radar reflectivity factor from a vertically pointing K_u -band radar obtained on 13 December 2007 at Stony Brook, NY. The corresponding temperature profile and hydrometeor types are shown on the left. In this example, almost all the rain freezes into sleet yielding a sharp transition in observed reflectivity associated with the changing dielectric constant. Data courtesy of Brian Colle, Stony Brook University.

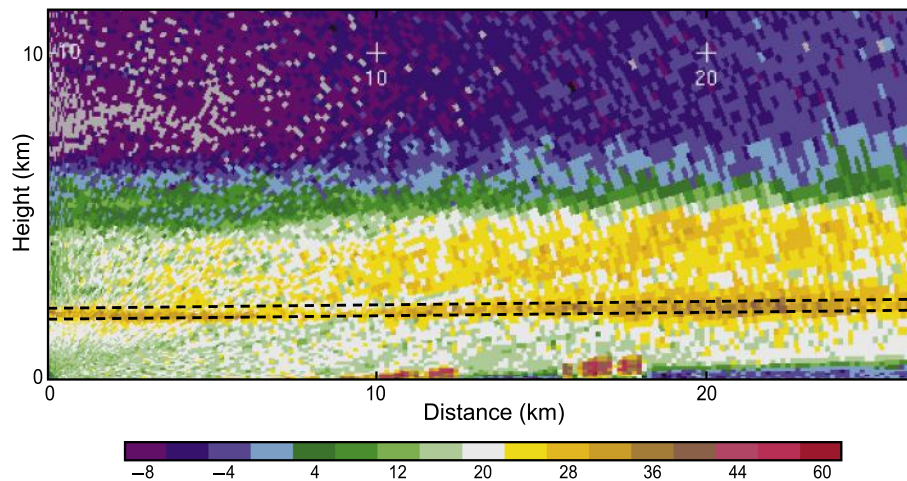


Figure 7 C-band radar cross section through a wide stratiform precipitation area in the central valley of California on 0148 UTC 3 December 2010. Tick marks are at 10 km intervals. The radar's observation of the melting layer bright band degrades with increasing range from the radar (to the right). Image courtesy of David Kingsmill, University of Colorado, CIRES and NOAA ESRL.

yields high reflectivities in volumes partially overlapping the physical melting layer even though that layer occupies only a portion of the resolution volume.

Beamfilling

The spatial scale of the resolution volume is important in interpreting radar data, since electromagnetic sensors cannot distinguish the spatial distribution of scatterers within a volume. **Figure 8** shows six unit resolution volumes containing different spatial orientations of nine scatterers of the same material, shape, and size. Since radar reflectivity factor (eqn [12]) is only a function of the size and number of the scatterers per unit volume, the radar reflectivity factor values for all six volumes are equal. By convention, it is assumed that the beam is filled with uniformly distributed scatterers (**Figure 8(a)**). Nonuniform beamfilling occurs when the assumption of uniformly distributed scatterers is false. High

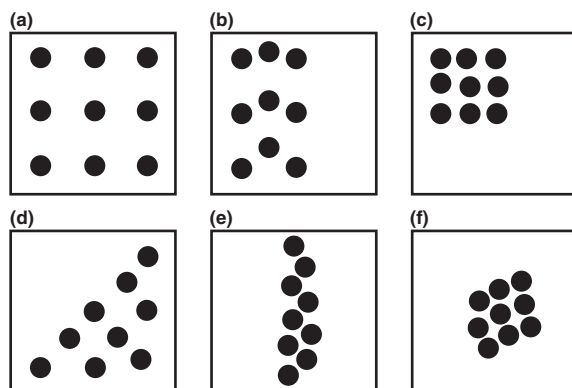


Figure 8 Set of six unit resolution volumes (a–f) with the same radar reflectivity factor. All volumes contain the same number of identical targets. The spatial distribution of targets within a radar-observed resolution volume is unknown and is usually assumed to be similar to the uniform distribution shown in (a). (b–f) are examples of nonuniform beamfilling.

time and spatial resolution radar data show that rainfall contains many rapidly varying small-scale structures (**Figure 9**). Corrections for nonuniform beamfilling are an area of active research. These corrections are especially important for large resolution volumes such as those associated with spaceborne radars. An important component of the radar design process is the balancing of the spatial scale of the features of interest and the physical characteristics of the sensor that determine the scale of the resolution volumes.

Attenuation: Energy Losses along the Radar Beam

The path along the radar beam to the target resolution volume and back to the radar contains air molecules and possibly cloud and precipitation particles. These particles absorb and scatter the radar energy, reducing both the incident energy at the target and the backscattered energy arriving at the radar compared to what their values would be if the intervening medium were free space. Attenuation is the total power extracted from the wave and is expressed mathematically as the sum of the power

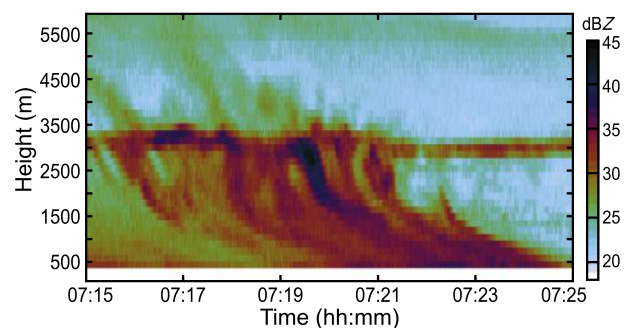


Figure 9 Fine details of the reflectivity variability within 10 min of S-band vertically pointing radar data obtained on 20 September 1999 at Locarno, Switzerland. Profiles were obtained every second. Taking into account the advection speed of the storm and the radar beam width, a volume corresponding to between 1 and 2 km altitude and 5 min is approximately equivalent to 0.1 km^3 .

absorbed and the power scattered by the intervening particles. Attenuation is a function of radar wavelength, and the size, shape, and composition of the intervening particles. Attenuation accumulates as the wave moves from the radar to the target and back from the target to the radar. Following Beer's law, the incremental reduction in received power with incremental distance ds caused by attenuation is

$$\frac{d\bar{P}_r}{ds} = -2k_L(s)\bar{P}_r \quad [15]$$

in which $k_L(s)$ is the attenuation coefficient (in units of inverse length) over an incremental volume along a path centered at range s . Although $k_L(s)$ is called a coefficient, it is not a constant. Rather it represents the combined influences of attenuation by gases, cloud, and precipitation at a particular location described by range s . The factor 2 is needed since the radar energy travels the same path twice. The received power accounting for the attenuation is the integral solution of eqn [15] from ranges 0 to r .

$$\bar{P}_r = \bar{P}_{r0} \exp\left(-2 \int_0^r k_L(s) ds\right) \quad [16]$$

in which $\bar{P}_{r0}$ is the power that would have been received without attenuation. Attenuation is often expressed as the ratio of the attenuated received power to the nonattenuated received power in units of dB km^{-1} . By utilizing the identity $\log_{10} X = 0.434 (\ln X)$ and defining the specific attenuation as $k(s) = (10 \log_{10} e) k_L(s) = 4.34 k_L(s)$ in units of dB km^{-1} , eqn [16] becomes

$$10 \log_{10} \frac{\bar{P}_r}{\bar{P}_{r0}} = 4.34 \left(-2 \int_0^r k_L(s) ds \right) = -2 \int_0^r k(s) ds \quad [17]$$

The specific attenuation can also be defined in terms of the DSD, and the extinction cross section, $\sigma_e(D)$, which represents the attenuation contributed by a particular particle size D , using

$$k(r) = \int_0^\infty \sigma_e(D) N(D, r) dD \quad [18]$$

in which $N(D, r)dD$ is the number density of hydrometeors per particle size interval dD per unit volume within the incremental volume centered on range r . In practice, specific attenuation is often estimated as a function of the equivalent reflectivity without attenuation $Z_e(r)$.

$$k(r) = \alpha Z_e(r)^\beta \quad [19]$$

in which α and β are coefficients, which are a function of the drop size distribution $N(D)$.

Equation [16] and the definition for $k(s)$, can be incorporated into eqn [13] to yield the equation for the radar reflectivity factor in situations where the attenuation is not negligible.

$$\begin{aligned} Z_e(r) &= \frac{\bar{P}_r r^2}{C|K|^2} \exp\left(2 \int_0^r k_L(s) ds\right) \\ &= \frac{\bar{P}_r r^2}{C|K|^2} \exp\left(\frac{2\alpha}{4.34} \int_0^r Z_e(s)^\beta ds\right) \end{aligned} \quad [20]$$

At radar wavelengths, attenuation by atmospheric gases is dominated by oxygen and water vapor absorption. Gaseous attenuation varies with the radar wavelength, path length, and depth of the troposphere penetrated by the radar beam. Two-way gaseous attenuation curves have been calculated using a standard atmosphere for a range of radar beam elevation angles and radar wavelengths. The overall effect of gaseous attenuation is small at S-band and C-band wavelengths, of order 0.01 dB km^{-1} one way for a 10-cm wavelength radar, but increases as wavelength decreases. A correction for gaseous attenuation is usually made automatically in the radar's signal processor.

The attenuation caused by cloud droplets $<0.2 \text{ mm}$ in diameter is dominated by absorption. For radar wavelengths $\geq 5 \text{ cm}$, attenuation by cloud particles is sufficiently minimal to be ignored. For X-band and K-band wavelengths, the one-way attenuation due to cloud droplets is of the order of 0.1 dB km^{-1} per g m^{-3} liquid water content and is of sufficient magnitude to be of concern. However, unlike the situation with gaseous attenuation, there is no standard model for the distribution of cloud droplets as their distribution is discontinuous and chaotic. Because of the complexities in estimating the actual distribution of cloud particles along a particular radar beam, an automatic correction for attenuation by cloud particles is difficult to implement.

Attenuation in rainfall and snow is a function of the radar wavelength, temperature, particle type, and particle-size distribution. The attenuation in rain can be calculated as a function of reflectivity for specified DSDs, temperatures, and radar wavelengths as shown in Figure 10 and Table 4. At

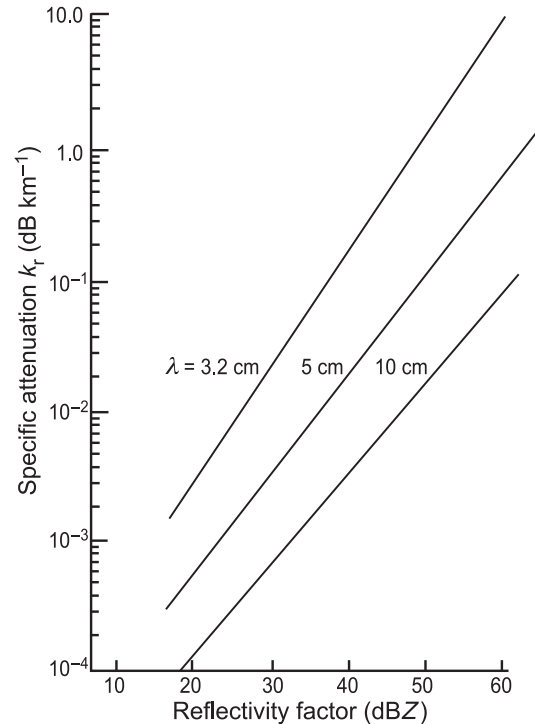


Figure 10 Specific attenuation versus radar reflectivity at $\lambda = 3.2, 5$, and 10 cm . Computations assume $T = 18^\circ\text{C}$ and Laws and Parson drop size distribution. Adapted from Doviak, R.J., Zrnic, D.S., 2006. Doppler Radar and Weather Observations, second ed. Dover Publications, New York, with permission from Dover Publications.

Table 4 Attenuation in rain for several wavelengths as a function of radar reflectivity factor (Z) at $T = 0^\circ\text{C}$

Radar wavelength (cm)	Specific attenuation in units of dB km^{-1}
2.0	$7.15 \times 10^{-4} Z^{0.725}$
3.21	$2.9 \times 10^{-4} Z^{0.72}$
5.5	$1.12 \times 10^{-4} Z^{0.62}$
10.0	$3.0 \times 10^{-5} Z^{0.62}$

Adapted from Battan, L.J., 1973. Radar Observation of the Atmosphere. University of Chicago Press, Chicago.

reflectivities >40 dBZ, which can occur within heavy rainfall, attenuation can become significant at C-band and X-band. S-band energy can become significantly attenuated when passing through regions containing hail or exceptionally heavy rainfall ($Z > 55$ dBZ). **Figure 11** shows a comparison of X-band and S-band reflectivities obtained in a squall line in Oklahoma. While both radars are in reasonable agreement regarding the reflectivity pattern between the squall line and the radar, there is a significant difference in the reflectivity pattern behind the squall line. The attenuated X-band radar reflectivities indicate a much smaller area of precipitation and weaker intensities behind the squall line compared to the non-attenuated S-band data.

Reflectivity by itself provides insufficient information to correct for attenuation since the attenuation is a function of the real rain field that is imperfectly measured because of attenuation. The rainfall attenuation correction is sensitive to the actual distribution of reflectivity and assumptions about the DSDs along the radar beam. Under special circumstances, such as downward-pointing precipitation radars on aircraft and satellites, a reference reflectivity such as the reflectivity of the ocean surface can be used to estimate path-integrated attenuation. Attenuation as a function of range can be

estimated with a dual frequency radar using the differential attenuation between the two wavelengths (**Figure 10**). Corrections based on horizontally and vertically polarized differential phase measurements are also under development. When attenuation corrections are not available, the reduction in reflectivity due to attenuation has to be accounted for in the data interpretation.

Location of the Radar Beam in the Atmosphere

Standard Refraction

The classic formula for refraction is Snell's law, which describes the bending of a light ray at the interface of two media. The ratio of the incident velocity of the wave (V_i) to the refracted velocity of the wave (V_r) is $V_i/V_r = \sin(i)/\sin(r)$, where i is the angle of incidence and r is the angle of refraction. In the earth's atmosphere there is no interface, but the nonuniform vertical distribution of water vapor pressure and temperature refracts the radar beam, changing its direction of propagation compared to a radar beam in free space. The change in direction of the beam is defined in terms of the vertical gradient of the index of refraction dn/dh where $n = c/v$, h is height, c is the speed of light, and v is the velocity of the wave in the medium. At precipitation radar frequencies for a standard atmosphere, $dn/dh \approx -4 \times 10^{-8} \text{ m}^{-1}$, sufficiently large to noticeably bend the radar beam downward compared to its free-space path. The negative value of dn/dh implies that v increases with increasing height.

If the index of refraction were constant with height or zero, the radar beam would curve upward with increasing range relative to the surface of the earth because of the earth's curvature. Precipitation radar data are usually interpreted in a ground-relative frame of reference. The following equation for the height of the radar beam above the earth's surface (h)

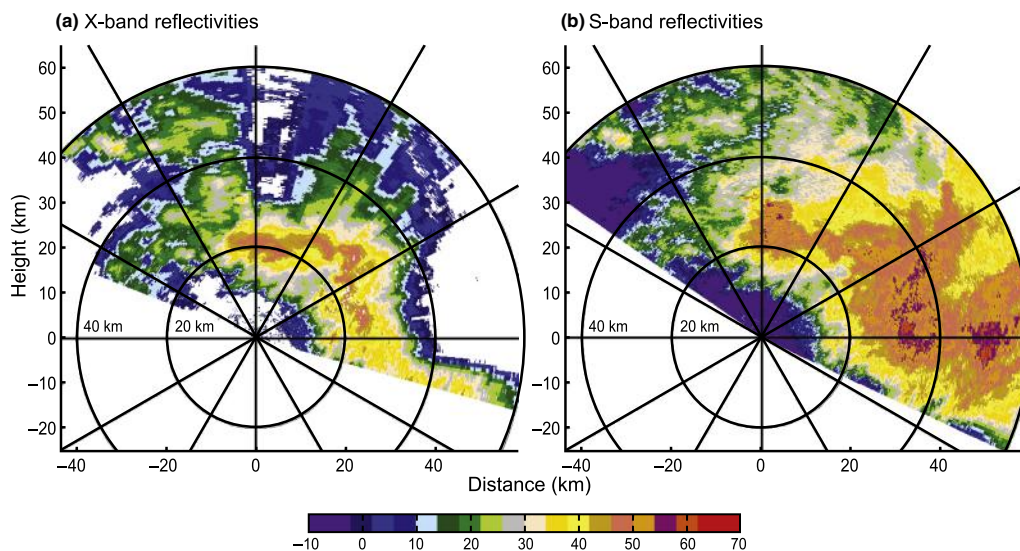


Figure 11 Comparison of X- and S-band radar reflectivities in a 1.5° elevation angle scan obtained from side-by-side X-band and S-band radars on 2342 UTC 15 June 2002 in the Oklahoma panhandle. The strong reflectivities within the squall line to the northeast of the radar attenuate the X-band signal (a). The S-band reflectivity data (b) reveal a larger area of precipitation behind the squall line than the X-band radar reflectivity data. Range rings are at 20 km intervals. Image courtesy of Scott Ellis, NCAR.

takes into account both the refraction by a standard atmosphere and the curvature of the earth.

$$h \cong h_0 - \frac{4}{3}R + \sqrt{r^2 + \left(\frac{4}{3}R\right)^2 + 2r\frac{4}{3}R \sin \theta_e} \quad [21]$$

in which θ_e is the elevation angle of the radar beam, h_0 is the height of the radar antenna, r is the slant range from the radar, and R is the radius of the earth at the latitude of interest. The arc distance, s , along the surface of the earth is defined as

$$s \cong \frac{4}{3}R \sin^{-1} \left(\frac{r \cos \theta_e}{\frac{4}{3}R + h} \right) \quad [22]$$

Figure 12 shows the radar beam center height (shaded lines) and beam width relative to the arc distance along the earth's surface for a US National Weather Service precipitation scan strategy (0.95° radar beam width and elevation angles: 0.5, 1.45, 2.4, 3.35, 4.3, 6.0, 9.9, 14.6, 19.5°). The radar location is indicated by the circle at 0 km height. Standard atmospheric refraction and the earth's curvature have been taken into account to calculate the beam path following eqns [21] and [22].

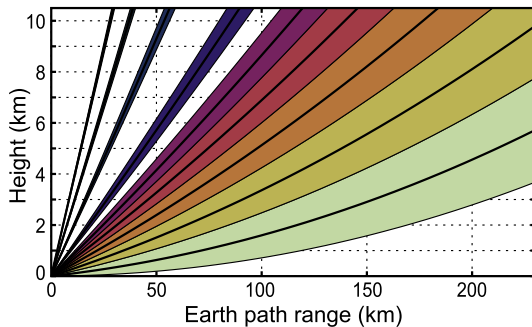


Figure 12 Radar beam center height (shaded lines) and beam width relative to the arc distance along the earth's surface for a US National Weather Service precipitation scan strategy (0.95° radar beam width and elevation angles: 0.5, 1.45, 2.4, 3.35, 4.3, 6.0, 9.9, 14.6, 19.5°). The radar location is indicated by the circle at 0 km height. Standard atmospheric refraction and the earth's curvature have been taken into account to calculate the beam path following eqns [21] and [22].

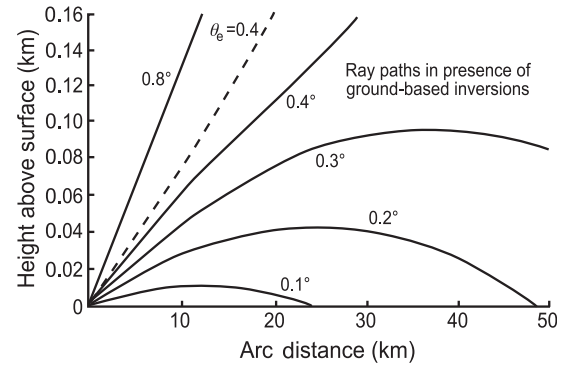


Figure 13 Ray paths in arc distance coordinates for radar beams in conditions with a standard atmosphere (dashed line) and for an atmosphere with a temperature inversion from the surface to 100-m altitude (solid lines). The arc distance coordinate compensates for the curvature of the earth and atmospheric refraction such that the ray path for the elevation angle $\theta_e = 0.4^\circ$ for a standard atmosphere is shown as a nearly straight line (dashed line). In comparison, ray paths for 0.4° and lower elevation angles in conditions with a low-level temperature inversion are refracted toward the surface. The lower the elevation angle, the greater the distance the ray path traverses within the 100-m thick inversion, and the greater the refraction of the beam downward.

moisture and temperature inversions and hence nonstandard refraction include nocturnal radiation cooling, advection of warm dry air over cooler bodies of water, and downdrafts of cool moist air that reach the surface within precipitating storms.

When the radar beam undergoes superrefraction, the height of the radar resolution volume at a particular range is lower than it would be under conditions of standard refraction. A radar beam elevation angle that usually clears nearby mountain peaks may intersect them under conditions of nonstandard refraction. A radar echo at a mountain peak can be difficult to discern as a nonmeteorological echo associated with anomalous propagation, as it is common for orographic precipitation to form over mountain peaks. In the extreme case, when the ray path is pointed into the ground, a strong echo, associated with the large cross section of the earth's surface, appears on the radar display.

A characteristic of nonmeteorological echo associated with anomalous propagation is that these echoes are relatively stationary compared to echoes associated with precipitating storms. A comparison of low elevation radar scans from several consecutive times can often reveal anomalous propagation. Automatic detection and removal of nonmeteorological echo associated with anomalous propagation includes methods involving polarization radar variables.

Polarization Variables Associated with Precipitation Estimation

Small raindrops, <1 mm in diameter, are spherical. Larger raindrops are deformed by aerodynamic drag into horizontally oriented oblate spheroids. An oblate spheroid is the body of revolution formed when an ellipse with minor axis dimension (a) and major axis dimension (b) is rotated about its minor axis. Raindrops usually fall with their maximum dimension oriented horizontally. This orientation may be temporarily disturbed by

turbulence, drop collision, or aerodynamic instability. The ratio between the horizontal and vertical dimensions of larger drops results in different electromagnetic properties of the scattered energy when the incident energy is horizontally versus vertically polarized. A special type of precipitation radar, called a polarimetric radar (also known as polarization diversity radar), is designed to measure these properties by transmitting and receiving radiation in more than one polarization. Ongoing research has shown that polarization radar variables involving the differential amplitude and phase of the received power at orthogonal polarizations can be related to the physical characteristics of the precipitation. Of these variables, two commonly used in precipitation applications are differential reflectivity (Z_{DR}), related to the axis ratio of the precipitation particles, and specific differential propagation phase shift (K_{DP}), which is a function of the concentration, composition, and size of only the anisotropic particles in a sampling volume.

Differential Reflectivity

To obtain radar reflectivity, energy is transmitted and received at the same polarization, usually horizontal. Differential reflectivity is the difference between the horizontally transmitted, horizontally received reflectivity factor (Z_{HH}) and the vertically transmitted, vertically received reflectivity factor (Z_{VV}).

$$Z_{DR} = 10 \log_{10} \frac{Z_{HH}}{Z_{VV}} \quad [23]$$

Differential reflectivity is a measure of the reflectivity-weighted mean axis ratio (a/b) of precipitation particles in a resolution volume. As raindrops increase in volume, D increases, the shape of the drop becomes more oblate, the axis ratio decreases, and the associated Z_{DR} value increases. For spherical raindrops or spherical ice particles, the axis ratio $a/b = 1$ and $Z_{DR} \sim 0$. Of particular importance to precipitation studies is the capacity of Z_{DR} obtained at nearly horizontal elevation angles to distinguish among regions containing rain, hail, and graupel, and the layer of melting snow particles above the rain layer. Table 5 describes typical Z_{DR} values associated with several basic precipitation types. Z_{DR} is also a function of the dielectric constant and hence composition of the particle. For a given axis ratio, the Z_{DR} value of a raindrop will be larger than for an ice particle of the same size. Graupel and hail particles usually either tumble or are axially symmetric and have an associated Z_{DR} value near 0. Negative values of Z_{DR} have been reported for regions containing hail and graupel but the detailed

Table 5 Typical ranges of observed differential reflectivity values in several types of precipitation

Differential reflectivity (dB)	Associated precipitation types
−0.5 to 0.5	Marginally detectable precipitation
−0.5 to 0.5	Drizzle, very light rain, light snow
>1	Moderate rain and heavier snow
0.5–4	Moderate to heavy rain
−2 to 0.5	Hail and graupel
0.5–4	Melting snow particles

Adapted from Doviak, R.J., Zmric, D.S., 2006. Doppler Radar and Weather Observations, second ed. Dover Publications, New York; Straka, J.M., Zmric, D.S., Ryzhkov, A., 2000. Bulk hydrometeor classification and quantification using polarimetric radar data: synthesis of relations. Journal of Applied Meteorology 39, 1341–1372.

physics underlying these measurements is not well understood. Melting snow appears to the radar as large particles with the dielectric properties of water. These particles often have very large axis ratios and large Z_{DR} values.

Differential reflectivity is independent of the absolute radar calibration since it is the difference between Z_{HH} and Z_{VV} but it is sensitive to relative calibration. In practice, Z_{HH} and Z_{VV} are often measured by parallel hardware with different calibration constants necessitating two calibration measurements. When viewed vertically, raindrops of all sizes appear circular and have an associated Z_{DR} value equal to 0. A Z_{DR} bias, accounting for the relative difference in calibration between the H and V polarizations can be estimated by pointing the radar beam directly upward in rain.

Differential Phase Variables

As an electromagnetic wave traverses a precipitation-filled volume, incident energy is scattered back toward the radar and forward along the beam. The forward scattered (propagated) component of the wave acquires a phase shift compared to the free-space component of the wave transmitted from the radar. Within horizontally oblate raindrops, the propagating horizontally polarized wave undergoes a larger phase shift per unit length and travels more slowly than the vertically polarized wave. After passing through a volume filled with horizontally oblate raindrops, the horizontally polarized wave will have a larger propagation phase shift than the vertically polarized wave (Figure 14). The one-way differential propagation phase (ϕ_{DP}) is defined as the difference between the propagation phase shift of the horizontally transmitted, horizontally received energy (ϕ_{HH}), and the propagation phase of the vertically transmitted, vertically received energy (ϕ_{VV}).

$$\phi_{DP} = \phi_{HH} - \phi_{VV} \quad [24]$$

As the radar wave traverses a region of precipitation filled with oblate drops, ϕ_{DP} accumulates with increasing range. To remove range effects, ϕ_{DP} is differentiated with respect to r to yield specific differential propagation phase (K_{DP}) in units of degree per kilometer.

$$K_{DP} = \frac{d\phi_{DP}}{dr} \quad [25]$$

Often measurements of ϕ_{DP} are noisy and K_{DP} is integrated over several kilometers to obtain a usable signal.

K_{DP} values can be related to the liquid content and axis ratio of the drop by the following equation.

$$K_{DP} = \left(\frac{180}{\lambda} \right) 10^{-3} cW \left(1 - \left[\frac{a}{b} \right]_m \right) \quad [26]$$

in which λ is the radar wavelength, $c \sim 3.75$, W is the liquid water content in g m^{-3} , and $(a/b)_m$ is the mass-weighted mean axis ratio. For a given liquid water content, K_{DP} increases with decreasing radar wavelength within the precipitation radar frequency band. K_{DP} is not affected by attenuation since it is based on the measurement of the phase shift of the wave rather than the amplitude of the returned power. This basis on phase shift allows a K_{DP} signal to be obtained when the radar beam is partially blocked such as in mountainous terrain. Applications of K_{DP} include distinguishing the rain portion of rain–hail

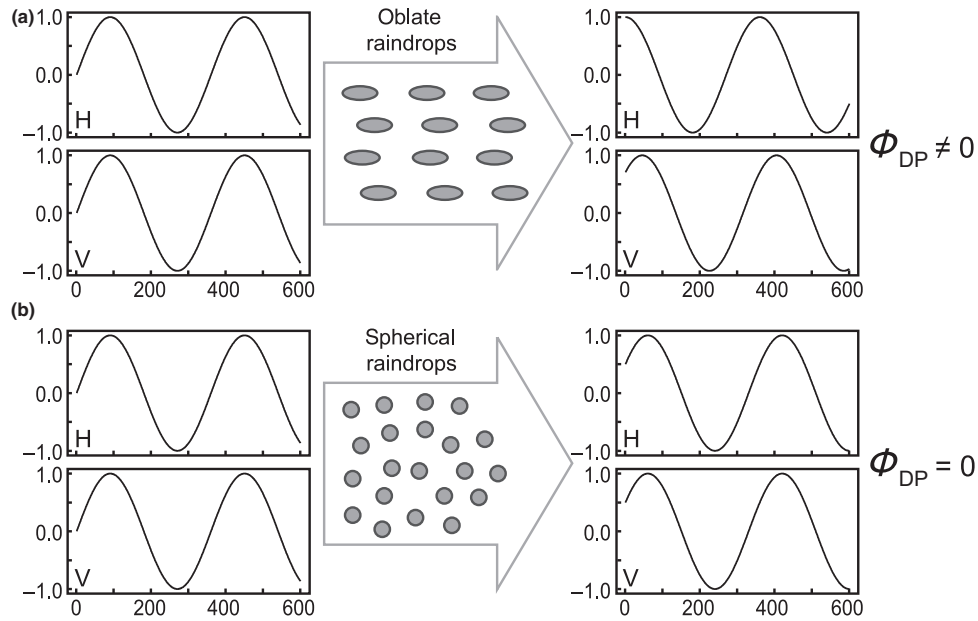


Figure 14 Sketch of the propagation phase shift of horizontally and vertically polarized electromagnetic waves passing through a precipitation-filled radar resolution volume. Left column shows the horizontally (H) and vertically (V) polarized waves prior to entering the resolution volume (middle column). Right column shows the H and V polarized waves after leaving the resolution volume. The H and V polarized waves are assumed, for this example, to be in phase prior to entering the resolution volume. (a) When the waves encounter a resolution volume containing horizontally oblate raindrops, the phase of the horizontally polarized wave is shifted more than the vertically polarized wave and $\phi_{DP} \neq 0$. (b) When the waves encounter spherical particles such as small raindrops or hail, the H polarized wave and V polarized wave are shifted the same amount and $\phi_{DP} = 0$.

Table 6 Typical ranges of observed S-band K_{DP} values in several precipitation types

K_{DP} ($^{\circ}\text{km}^{-1}$)	Precipitation types
–0.5 to 0.5	Marginally detectable precipitation
–0.5 to 0.5	Drizzle, very light rain, snow
–0.5 to 1	Moderate rain
0.5 to 5	Moderate to heavy rain
–0.5 to 1	Hail
–0.5 to 1	Melting snow particles

Adapted from Doviak, R.J., D. S. Znic, 1993. Doppler radar and weather observations, Academic Press, New York, p. 562; Straka, J.M., Znic, D.S., Ryzhkov, A., 2000. Bulk hydrometeor classification and quantification using polarimetric radar data: synthesis of relations. *Journal of Applied Meteorology* 39, 1341–1372; and Bringi, V.N., Chandrasekar, V., 2001. *Polarimetric Doppler Weather Radar*. Cambridge University Press, Cambridge.

mixtures and estimating the liquid water content of oblate raindrops. A disadvantage of K_{DP} is its insensitivity to precipitation composed of small spherical raindrops, where $D < 1$ mm and $a/b = 1$, associated with low liquid water contents and low rain rates. Table 6 describes typical S-band K_{DP} values of several basic precipitation types.

Precipitation Characterization and Estimation from Radar Data

Hydrometeor Classification

Combining complementary information from radar-observed variables can yield more information on hydrometeor types than a single variable alone. As described above, Z provides information on the size and number of particles, Z_{DR} values are

a function of particle shape (axis ratio) and size, and K_{DP} is proportional to liquid water content in rain and melting hail.

The combination of Z and Z_{DR} is often used to distinguish hail from heavy rain. This distinction is important for operational warnings as well as for rain mapping. Figure 15 shows a commonly used hail parameter called HDR. Radar data with values of Z and Z_{DR} above the line are classified as hail while those below the line are classified as rain. This figure also illustrates the impact of Mie scattering (resonance) on Z and Z_{DR} values at C-band (5.64 GHz) and X-band (9.4 GHz). The dots represent Z and Z_{DR} values computed from observed DSDs. Using the S-band (3 GHz) dots as representing values in the Rayleigh regime, one can see that resonance effects yield larger Z_{DR} values for a given Z value particularly at C-band.

An application of hydrometeor classification is shown in Figure 16. The different columns show Z , Z_{DR} , K_{DP} , and hydrometeor classification for an intense summer storm in Colorado. At the initial time (Figure 16, top row) near $x = -15$ km and $y = -40$ km, there is an area of heavy rain ($Z \sim 57$ dBZ and $Z_{DR} \sim 2.3$ dB) with a high liquid water content ($K_{DP} \sim 3.4^{\circ}\text{km}^{-1}$). Eleven minutes later (Figure 16, bottom row), the high Z region has intensified to greater than 60 dBZ and shifted slightly to the south. The corresponding near zero Z_{DR} value indicates large hail. The locally heavier rainfall (K_{DP} local maximum) is slightly to the north of the hail core.

Rainfall Mapping with Precipitation Radar: Role of the Raindrop-Size Distribution

As described in the sections above, measured precipitation radar variables can be related to several parameters of the

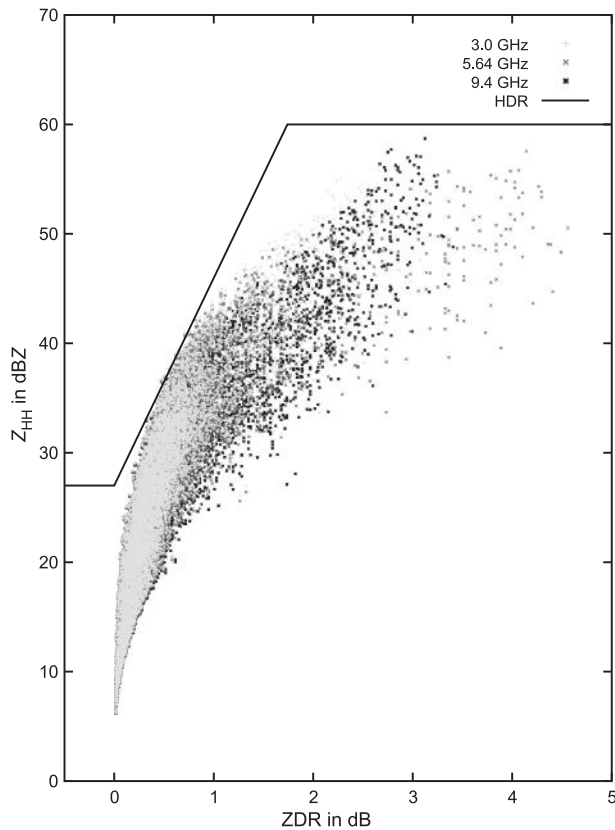


Figure 15 Computed Z_{DR} versus Z_{HH} for three frequencies based on observed drop size distributions obtained in Locarno, Switzerland during September and October 1999. The wider spread of points at C-band (5.64 GHz) compared to S-band (3.0 GHz) and X-band (9.4 GHz) indicates stronger resonance at C-band for $Z > 35$ dBZ. The line marked HDR indicates the boundary between hail (above the line) and rain (below the line) based on the combined Z and Z_{DR} hail identification parameter. From Aydin, K., Seliga, T.A., Balaji, V., 1986. Remote sensing of hail with a dual linear polarization radar. *Journal of Applied Meteorology and Climatology* 25, 1475–1484. Plot courtesy of Martin Hagen, DLR.

raindrop size distribution $N(D)$. The raindrop size distribution is generally accepted to be a truncated exponential function of the form,

$$N(D) = N_0 D^\mu \exp(-3.67D/D_0) \quad [27]$$

in which D is the diameter of a spherical drop with volume equal to that of the actual raindrop, μ has a value between -3 and 8 , D_0 is the median volume diameter of the distribution, and N_0 is a function of D_0 , μ , and the total drop concentration. DSDs and drop shapes have been measured and modeled. These studies have yielded a range of values for and a range of relationships among N_0 , D_0 , μ , and the drop axis ratio as a function of D . These uncertainties translate into uncertainties in relations among parameters of $N(D)$.

Moments of the DSD can be related to one another by an equation of the form, $\bar{y} = \alpha \bar{x}^\beta$, where x and y are different moments of the DSD, and α and β are real-valued coefficients. Table 7 lists several moments of the DSD. The most commonly applied association between moments of the DSD is the Z – R relation, which translates the variability of the DSD as measured by the radar in terms of Z in $\text{mm}^6 \text{m}^{-3}$, into variability in terms of the desired quantity R in mm h^{-1} . The Z – R relation is traditionally expressed in the form $Z = aR^b$, although Z is the independent variable. There are numerous Z – R mappings in the literature. Each relation is based on different empirical data sets, methodologies, and assumptions regarding the coefficients in eqn [27]. Some selected relations are shown in Table 8. The Z -conditional relative uncertainty in rain rate varies from ~ 10 to 50% as a function of Z for the four Z – R relationships in Table 8. These Z -conditional uncertainties represent a combination of several error sources and differences in precipitation among regions. Z – R methods usually truncate input Z at ~ 55 dBZ to mitigate large errors in estimated rain rate associated with hail.

Alternate methods to map rainfall using S-band, C-band, and X-band radar data utilize measurements of Z_{DR} and K_{DP} in addition to, or in place of, Z . K_{DP} is related to the fourth moments 4.24–5, depending on a number of factors of the

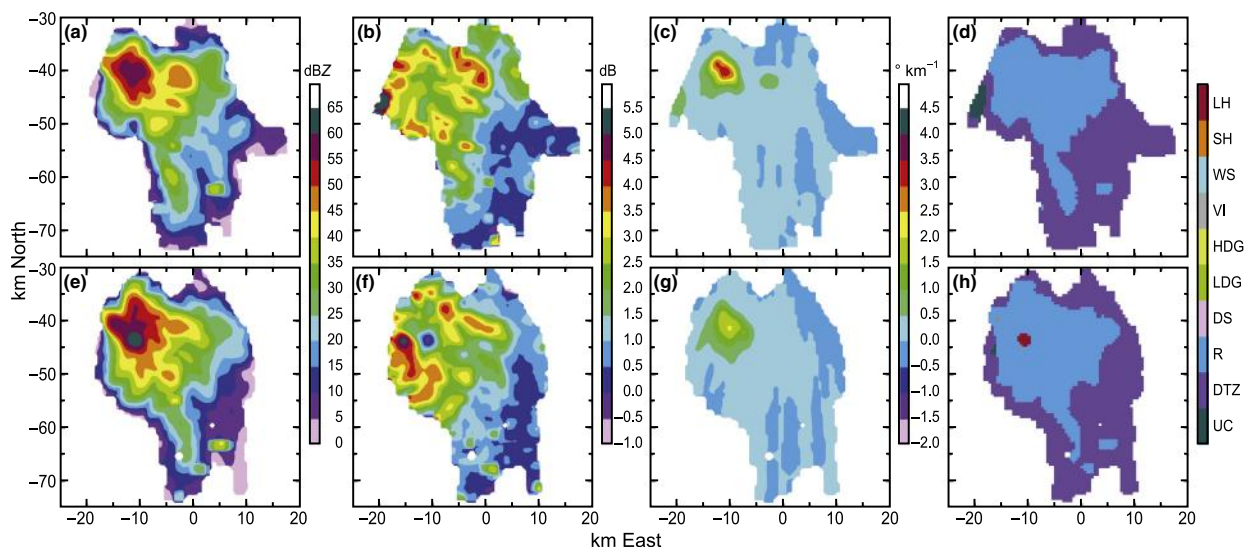


Figure 16 Radar data and hydrometeor classification obtained by the CSU-CHILL S-band radar in Colorado on 29 June 2007. Top row is for 0033 UTC. Bottom row is for 0044 UTC. (a) and (e) Radar reflectivity, (b) and (f) ZDR, (c) and (g) KDP, and (d) and (h) hydrometeor classification (LH – large hail, SH – small hail, WS – wet snow, VI – vertically aligned ice, HDG – high-density graupel, LDG – low-density graupel, DS – dry snow, R – rain, DTZ – drizzle, UC – unclassified). Figure courtesy of Brenda Dolan and Pat Kennedy, Colorado State University.

Table 7 Moments of the drop size distribution related to rainfall estimation

Measured or estimated property	Functional form in terms of $N(D)$
Total drop concentration (m^{-3})	$N_t = \int_0^\infty N(D) dD$
Radar reflectivity factor ($\text{mm}^6 \text{m}^{-3}$)	$Z = \int_0^\infty N(D) D^6 dD$
Liquid water content ($\text{mm}^3 \text{m}^{-3}$)	$W = \frac{\pi}{6} \int_0^\infty N(D) D^3 dD$
Rainfall rate (mm h^{-1})	$R = \frac{3.6\pi}{6000} \int_0^\infty N(D) D^3 V(D, T, P) dD$ in which $V(D, T, P)$ is the particle fall speed which is a function of the diameter of the particles, air temperature, and pressure

Table 8 Selected Z - R relationships

Source	Functional form in terms of $N(D)$
Midlatitude	$Z = 220R^{1.6}$
Tropical GATE	$Z = 230R^{1.25}$
Swiss Meteorological Agency	$Z = 315R^{1.5}$
US National Weather Service Default	$Z = 300R^{1.4}$

Marshall, J.S., Palmer, W.M.K., 1948. The distribution of raindrops with size. *Journal of Meteorology* 5, 165–166.

DSD and can be used to estimate R using a power law of the form $\bar{y} = \alpha \bar{x}^\beta$. Other methods employ several simultaneously measured parameters of the DSD to constrain the coefficients in eqn [27]. The derived DSD is used in turn to estimate the desired parameter R . Several assumptions, with attendant uncertainties, are needed to make the associations among the

polarimetric radar-measured bulk properties and the functional form of the DSD. Use of either Z_{DR} or K_{DP} to constrain the DSD coefficients requires assumptions about the drop axis ratio as a function of D . Neither Z_{DR} nor K_{DP} are sensitive to the small spherical drops within the DSD, so the distribution of the smaller drops must be assumed as a function of the derived

Table 9 Factors contributing to uncertainties in rainfall mapping from precipitation radar data. The magnitudes are for operational scanning radars with typical resolution volumes of 1 km^3 . The trend in uncertainty with increasing range is defined in terms of the same meteorological or nonmeteorological circumstance being present at closer versus farther ranges from the radar

Type of error	Factor	Nature of effect and situations where it is significant	Trend of uncertainty with increasing range from radar
Noise	Precision of measured $\bar{P}_r$	Varies with PRF and scan rate. Typically $\sim 1 \text{ dB}$.	Not a function of range.
Persistent bias	Calibration of $\bar{P}_r$	Typically $\sim 2 \text{ dB}$ for calibrated radars. Depends on priorities and resources.	Not a function of range.
Persistent bias	Changes in DSD between lowest radar measurement and the surface	Varies between 0 and 3 dB km^{-1} depending on vertical structure of precipitation.	Generally increases with height difference between lowest radar beam and the ground. Increases with range.
Persistent bias	Ground clutter and beam blockage	Varies with radar site and its relation to regional geography.	Shielded area increases with increasing range.
Intermittent bias	Signal enhancement by melting particles	Partially melted particles have Z values 1–5 dB larger than completely melted particles.	Effect limited to ranges where radar beam intersects melting layer.
Intermittent bias	Anomalous propagation	Associated with temperature and moisture inversions. Changes beam height in relation to its standard atmosphere value. In extreme, beam intersects ground yielding high Z value.	Magnitude increases with smaller beam elevation angles, increasing range, and increasing strength of inversion.
Intermittent bias	Nonmeteorological echo such as sea clutter, birds, and insects	Increases Z above its meteorological value. In extreme, indicates presence of heavy rain when there is none present.	Generally decreases with range, as beam is located higher from near surface sources.
Intermittent bias	Presence of hail	Increases measured Z to value 4–10 dB higher than that of associated rainfall.	Decreases once resolution volume becomes larger than pocket of hail.
Intermittent bias	Strong downdrafts	Increase R for same Z . Equivalent to reducing Z by 3–5 dB.	Decreases once resolution volume becomes larger than downdraft.
Intermittent bias	Attenuation	Reduces measured Z from its actual value. Increases with decreasing wavelength. Varies with distribution of precipitation along radar beam.	Accumulates with range.
Intermittent bias	Z -conditional variations in the DSD	Often difficult to isolate from other systematic errors. Uncertainty equivalent to variations in $Z \leq 2.5 \text{ dB}$.	Decreases with increasing size of resolution volume.

Adapted from Austin, P., 1987. Relation between measured reflectivity and surface rainfall. *Monthly Weather Review* 115, 103–107.

distribution of the larger oblate drops. Additional assumptions regarding the variability of the coefficients in the DSD, particularly μ , are often utilized.

Practical Issues Contributing to Rain Map Uncertainty

The quantitative estimation of errors associated with rain mapping from radar data is severely hampered by the lack of an independent, precise, and accurate value with which to compare. Additionally, rapidly changing conditions within precipitating storms make it difficult to repeat a measurement under the same conditions. Independent sources of random and systematic error in rain mapping are described in Table 9. Error sources for polarimetric methods are similar with a few exceptions. Z_{DR} and K_{DP} measurements are not biased by the presence of hail. K_{DP} is not influenced by attenuation. Both Z_{DR} and K_{DP} are sensitive to mild nonstandard refraction such that the height of the radar beam is lower in the atmosphere compared to its height in a standard atmosphere. Methods using polarimetric variables can detect when the nonstandard refraction is sufficiently strong such that the radar beam intersects the ground.

An additional source of uncertainty is associated with nonuniform beamfilling (Figure 8). For example, when an isolated 1 km^3 raining cell is contained within an otherwise

empty 8 km^3 resolution volume, the measured reflectivity for the entire resolution volume will be much smaller than the reflectivity within the isolated raining cell. When the resolution volume is not uniformly filled (i.e., rain is not homogenous in the volume), an uncertainty arises when computing the average rain rate from the average reflectivity. The reason is that while $\bar{y} = \alpha \bar{x}^\beta$, $\bar{y} \neq \alpha \bar{x}^\beta$ for $\beta \neq 1$, as is usually the case for Z - R relations. This type of uncertainty is of concern when computing R from Z for volumes that are larger than the scale of the rainfall variability (Figure 9) and when comparing different moments of the DSD obtained at different spatial resolutions, for example, ground-based radar and satellite passive microwave measurements, or radar and rain gauge data.

A comparison of the relative uncertainties associated with several independent sources of error in rain mapping is shown in Figure 17. The data shown are related to the monthly rainfall accumulation for October 2000 for the region centered on Kwajalein Atoll, Marshall Islands. In this analysis, uncertainties associated with the radar calibration, the Z - R relation, and a vertical profile correction are treated separately. Figure 17(a) shows a best estimate of monthly rainfall. The other panels represent the difference between high and low values for the particular source of uncertainty. There is a large variation in

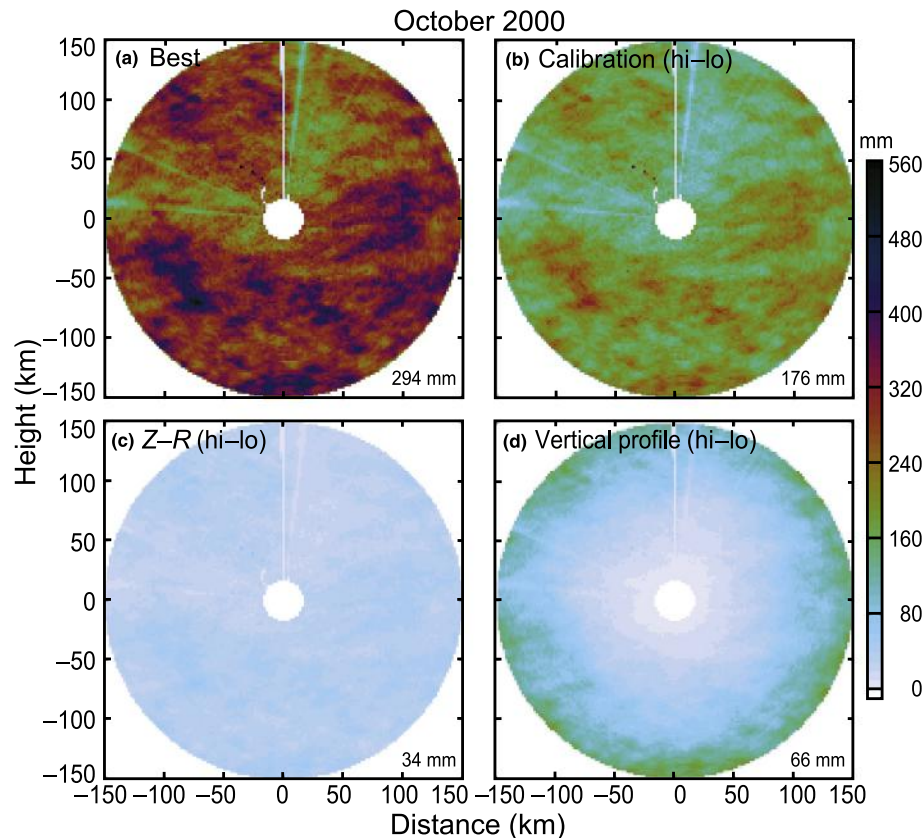


Figure 17 Monthly rain accumulation and uncertainties derived from reflectivity data collected by the KPOL S-band radar on Kwajalein Island, Marshall Islands for October 2000. Average accumulation over the region is shown in the lower right of each panel. (a) The best estimate for the month. The other panels show the difference between maps computed with (b) high (+2 dB) minus low (−2 dB) estimates of the calibration correction, (c) high ($Z = 160R^{1.5}$) minus low ($Z = 190R^{1.5}$) estimates of the Z - R relationship, and (d) high minus low estimates of a vertical profile correction. The difference maps are computed by using the best estimates of all factors except the one considered in that panel.

rainfall accumulation (regional average 176 mm) associated with small uncertainties in the radar calibration, in this case ± 2 dB (Figure 17(b)). The difference between different vertical profile corrections increases with increasing range, with the largest variations at the outer ranges of the domain (Figure 17(d)). In comparison, the uncertainty in Z – R relations derived from *in situ* drop size measurements obtained with a disdrometer are a smaller source of error (Figure 17(c), regional average = 34 mm).

See also: **Mesoscale Meteorology:** Cloud and Precipitation Bands; Mesoscale Convective Systems. **Radar:** Cloud Radar; Polarimetric Doppler Weather Radar.

References

- Austin, P., 1987. Relation between measured reflectivity and surface rainfall. *Monthly Weather Review* 115, 103–107.
- Aydin, K., Seliga, T.A., Balaji, V., 1986. Remote sensing of hail with a dual linear polarization radar. *Journal of Applied Meteorology and Climatology* 25, 1475–1484.
- Marshall, J.S., Palmer, W.M.K., 1948. The distribution of raindrops with size. *Journal of Meteorology* 5, 165–166.
- Straka, J.M., Zmic, D.S., Ryzhkov, A., 2000. Bulk hydrometeor classification and quantification using polarimetric radar data: synthesis of relations. *Journal of Applied Meteorology* 39, 1341–1372.
- Atlas, D. (Ed.), 1990. *Radar in Meteorology*. American Meteorological Society, Boston.
- Battan, L.J., 1973. *Radar Observation of the Atmosphere*. University of Chicago Press, Chicago.
- Bringi, V.N., Chandrasekar, V., 2001. *Polarimetric Doppler Weather Radar*. Cambridge University Press, Cambridge.
- Doviak, R.J., Zmic, D.S., 2006. *Doppler Radar and Weather Observations*, second ed. Dover Publications, New York.
- Houze, R.A., 1993. *Cloud Dynamics*. Academic Press, New York.
- Meischner, P. (Ed.), 2004. *Weather Radar*. Springer, Berlin.
- Meneghini, R., Kozu, T., 1990. *Spaceborne Weather Radar*. Artech House, Norwood.
- Rinehart, R.E., 2010. *Radar for Meteorologists*, fifth ed. Ronald E. Rinehart, Fargo.
- Skolnik, M.I. (Ed.), 1990. *Radar Handbook*. McGraw-Hill, New York.

Further Reading

Synthetic Aperture Radar (Land Surface Applications)

RK Vincent, Bowling Green State University, Bowling Green, OH, USA

© 2003 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 5, pp 1851–1858, © 2003, Elsevier Ltd.

Introduction

Synthetic aperture radar (SAR) images are produced by an active system that sends a microwave signal from a sensor platform to the ground and detects backscattered waves that the ground reflects directly back to a receiver on the same platform, which can be borne aloft by either airplanes or satellites. When the source and receiver are on the same platform, the radar is said to be monostatic. If the source and receiver are on different platforms, the radar is said to be bistatic. Commercial SAR systems are monostatic and always collect images to the side of the flight path of the sensor platform, unlike most multispectral imaging systems, which commonly look straight down and are passive (consisting only of receivers of reflected sunlight and emitted thermal infrared radiation). Figure 1 demonstrates these three different viewing geometries.

Although the currently available commercial SAR satellites, ERS-1 and Radarsat, send and receive only one frequency of microwaves, it is possible to build multifrequency SAR systems that send and receive several frequency ranges of radar signals simultaneously, similar to the way multispectral scanners collect several spectral bands of visible, reflective infrared, and thermal infrared wavelengths simultaneously. SAR systems that send out both horizontally and vertically polarized microwaves and receive both the horizontally and vertically polarized reflected components of each type of incident signal on the ground are called multipolarization radars. Most multispectral systems ignore polarization measurements of reflected sunlight and emitted thermal infrared radiation, although both are present in the passive signals coming from the ground. This article is devoted to applications of multifrequency and multipolarization radars, which have been used much less frequently than monofrequency SAR data in the past. However, future uses of multifrequency and multipolarization SAR systems are likely to increase, owing to their greater information content compared to monofrequency SAR systems, as this article will demonstrate.

Comparison of Multifrequency SAR with Multispectral, Short-Wavelength Images

Visible wavelengths (0.40–0.67 μm) of electromagnetic radiation from sunlight interact with the outer electron shells of transition metal ions in pigments, and thermal infrared wavelengths (4–14 μm) of heat emitted from objects on the Earth's surface interact with ions bound in crystalline lattices (Vincent, 1997). Thus, variations in the way electromagnetic radiation is reflected from or emitted by an object on the Earth's surface as a function of wavelength yield information about the chemical composition of the object, though visible light tells mostly about trace element content and thermal infrared radiation tells more about the bulk chemistry of the object. SAR systems employ microwave wavelengths (in the centimeter-to-meter range) that are at least 1000 times longer than even thermal infrared radiation, so radar waves cannot interact well either with pigments or with the crystalline structure of natural materials. Consequently, radar images contain much less information about chemical composition of an observed object than do either thermal or visible images.

However, radar images are much more sensitive than infrared or visible images to two parameters of an observed object: water content and physical extent (size and shape) of the object. These are the parameters for which microwaves (and radar images) have an advantage over all other wavelengths of electromagnetic radiation. Before they are discussed in the next two sections, however, it is instructive to review the relationship between wavelength and frequency for electromagnetic radiation, of which microwaves are examples. From the elementary physics equation of distance equals the velocity $\times$ time (or period, which is the reciprocal of the frequency), the relationship between wavelength and frequency of electromagnetic waves is given by eqn [1].

$$\lambda = \frac{c}{\nu} = \frac{30 \times 10^9 (\text{cm s}^{-1})}{\nu (\text{s}^{-1})} \quad [1]$$

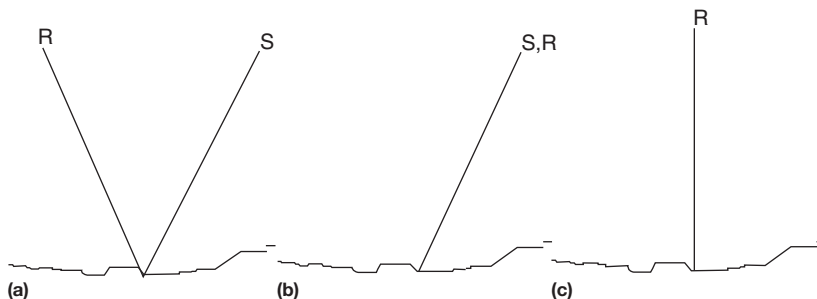


Figure 1 Viewing geometry for signal source (S) and receiver (R) for (a) bistatic radar, where source and receiver are on different platforms (any angle between S and R is acceptable, as long as R and S are both above the ground); (b) monostatic radar, where source and receiver are on the same platform; and (c) passive receiver for visible/infrared sensors of a LANDSAT satellite.

where ν is the frequency in cycles per second, or hertz, and c is the velocity of light (cm s^{-1}), which is the velocity at which all electromagnetic waves travel, including microwaves. For example, a microwave of frequency 1 GHz (10^9 Hz or 1 GHz) has a wavelength of 30 cm (within the range of L-band radar), and a microwave of frequency of 10 GHz has a wavelength of 3 cm (within the range of X-band radar). Thus, multifrequency (many frequencies) and multispectral (many wavelengths) are synonymous adjectives of radar.

The Effect of Water Content of an Observed Object on Radar Backscatter

Because water has a very high dielectric constant for most microwave wavelengths, it reflects microwaves very efficiently as a specular reflector, i.e., such that most reflection is off the first surface and the angle of reflectance equals the angle of incidence. Calm, standing water reflects almost all incident microwaves forward at the specular angle (equal to the angle of incidence), and reflects almost none of them in the direct

backscatter direction (where the receiver is located for SAR images), which makes calm water appear very dark in SAR images, regardless of the wavelength or frequency of the radar.

The effect of variations in water content of soils, or soil moisture (located above the water table), on radar backscatter is not as simple as in the case of standing water. Fortunately, this subject has been well addressed in a paper by Mancini and colleagues (see [Further Reading](#)) that reports on a laboratory experiment on the use of active microwave observations to estimate volumetric soil moisture content within an accuracy of 5%. A sandy loam with two surfaces (smooth and rough) was wetted and dried to generate different soil moisture profiles (with depth), and radar measurements were made at three different incidence angles (11° , 23° , and 35°) over a frequency ranging from 1 to 10 GHz, with 11.25 MHz step in frequency. [Figure 2](#) shows the resulting plots of radar backscatter coefficient versus frequency of the microwaves for both rough and smooth surfaces of the sandy loam at 11° and 35° angles of incidence, where w14 refers to the end of a wetting profile (water being added) that resulted in a 22.9% moisture content at a depth of 2.5 cm; d13 is a drying profile (water being

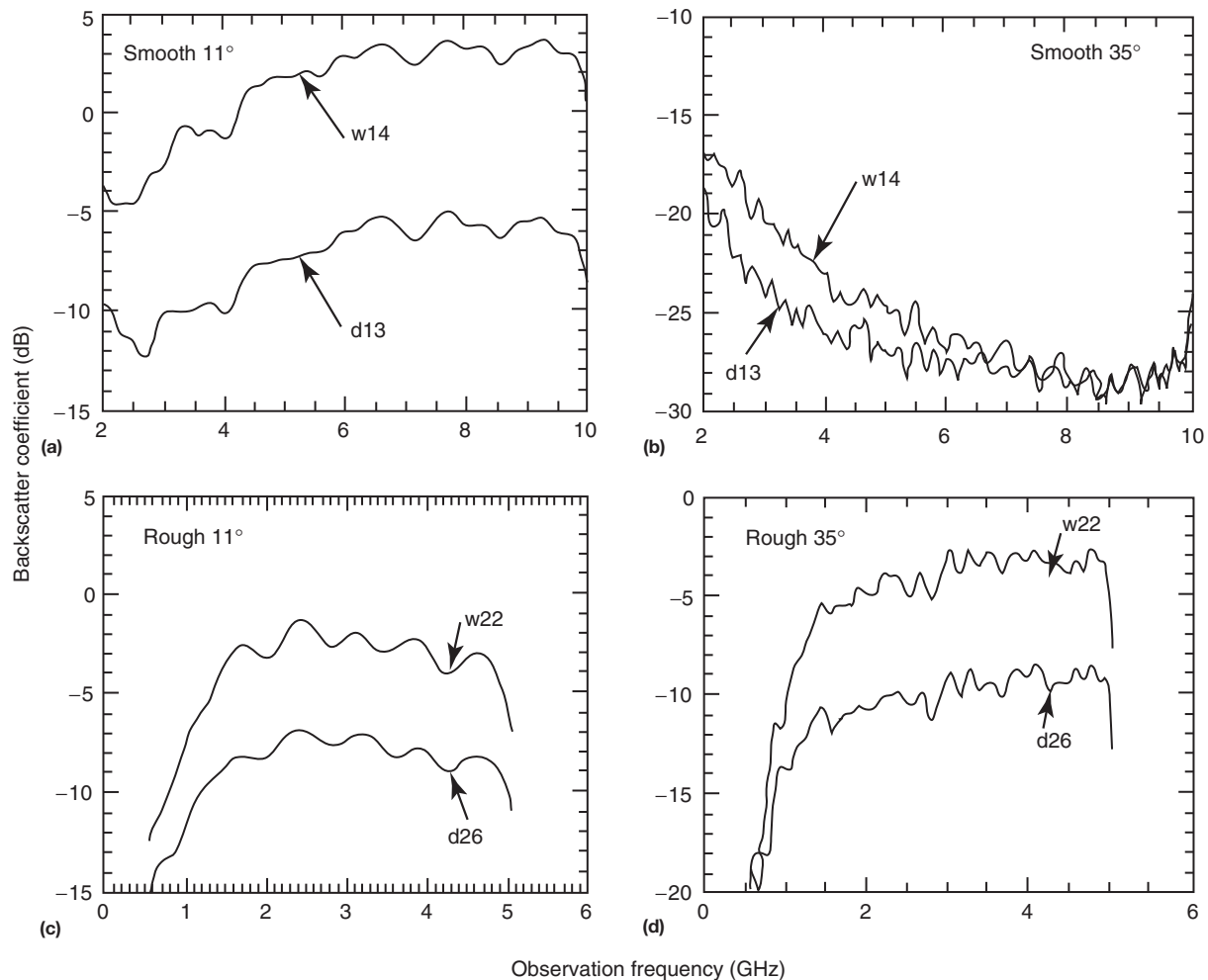


Figure 2 Radar data for smooth and rough surfaces with the extreme moisture conditions for 11° incidence angle and 35° incidence angle. Reproduced with permission from Mancini, M., Hoeben, R., Troch, P.A., 1999. Multifrequency radar observations of bare surface soil moisture content: a laboratory experiment. *Water Resources Research* 35 (6), 1827–1838.

removed by evaporation) ending with a 1.9% moisture content at a depth of 2.5 cm; w22 is a wetting profile ending with a 17.8% moisture content of 10.6% at a depth of 4 cm; and d26 is a drying profile ending with a moisture content of 10.6% at a depth of 4 cm.

Although the two left-most plots in [Figure 2\(a\) and 2\(c\)](#) demonstrate that at an angle of incidence of 11° there is a large difference between wet and dry soils, they also demonstrate that there is little frequency dependence in that difference. Thus, for an incidence angle of 11° (relatively small incidence angles), it is possible to determine soil moisture equally well at all frequencies. Multifrequency radar offers no advantage for soil moisture estimations from SAR measurements. However, the absolute value of the ratio of the difference between the backscatter coefficients at 5 and 2.5 GHz and the sum of the same two backscatter coefficients is much greater for the smooth surface than for the rough surface for all reported values of soil moisture. This indicates that multifrequency radar at low incidence angles could be used to determine relative surface roughness, regardless of soil moisture.

The top-right plot in [Figure 2\(b\)](#) shows that when smooth surfaces are viewed at higher incidence angles (35° in this case), specular reflection dominates volume reflection for frequencies greater than 6 GHz, and soil moisture no longer has much effect on the backscatter coefficient. Volume reflection dominates the rough surface reflection at an incidence of 35° , just as it did with both the smooth and rough samples of the 11° incidence measurements. In all cases above, wetter soils return more backscattered microwaves than drier soils at all frequencies, for a given surface roughness and incidence angle. When volume reflection dominates, as it does in all of [Figure 2](#) except for the smooth, 35° incidence case ([Figure 2\(b\)](#)), backscatter rises with increasing frequency, up to approximately 5 GHz. Lower incidence angles correspond to higher direct backscatter. The greatest signal return from soils, wet or dry, toward a monostatic radar receiver can be obtained by a low incidence angle and a frequency of approximately 5 GHz, as indicated by the results of the study of Mancini and colleagues.

Another feature in [Figure 2](#) is the oscillatory pattern in the backscatter coefficient as a function of frequency, which is especially obvious in the lower left plot ([Figure 2\(c\)](#)) for a rough surface viewed at an incidence angle of 11° . Mancini and colleagues claim that this is not caused by noise and state that it might be caused by the volume scattering component of the reflectance interfering with the specular reflectance component from the surface. If so, it would be possible to relate this oscillation to an observation depth, as discussed by Troch and colleagues. Therefore, multifrequency radar may be able to determine the observation depth of a soil through the oscillatory nature of the backscatter coefficient as a function of frequency. More research is needed to confirm this potentially important application of multifrequency radar.

The Effect of Size and Shape of an Observed Object on Radar Backscatter

The second parameter for which radar is more sensitive than visible or infrared radiation is the physical extent (size and

shape) of an observed object. In geological terms, size is the difference between gravel, cobblestones, and boulders on the surface. When the object is a layer of soil or rock, the thickness of the layer is considered to be its 'size.' Visible and reflective infrared electromagnetic radiations have wavelengths of the order of 0.5–4.0 mm, within the size range of clay-sized particles but not of sand-sized grains (about 60–1000 mm in size). Penetration of an electromagnetic wave into typical Earth materials will be on the order of a few wavelengths, which means that visible and infrared radiations typically penetrate rocks and soils less than the diameter of a human hair, which is approximately 100 mm. Microwaves, however, have much longer wavelengths than visible and infrared radiations, on the order of 1 cm to 1 m. They can interact with the shapes and sizes of gravel, cobbles, and boulders in different ways, depending on whether the wavelength of the microwaves selected for the radar is X-band (about 3 cm), L-band (about 30 cm), or P-band (about 70 cm). Microwaves can also penetrate dry soils to depths of meters, with greater penetration by longer wavelength radiation.

Louis Dellwig reported on the earliest investigations into the application of multifrequency radar to geological mapping more than three decades ago. He compared airborne radar images of K-band (sub-cm), Ka-band (0.84 cm), C-band (6.73 cm), and P-band (70 cm) wavelengths, collected at different times over an arid area near Písgah Crater, California. He made two significant observations from the multifrequency radar data set. First, he noted that windblown silts and sands were penetrated by the long-wavelength radar (P-band), which was reflected by the partly buried basaltic lava underneath. The shorter wavelength radar did not penetrate these aeolian deposits. Second, he noted that an alluvial fan with a surface composed of rock fragments that were mostly (85%) less than 5 cm in diameter was dark-toned (smooth) in the P-band radar image, but was bright and well-defined (rough) on the K-band radar image. This investigation ably demonstrated that longer wavelength radars penetrate deeper and are largely unaffected by pebbles and cobbles at the surface.

The author's PhD thesis was in this same Písgah Crater area; and in 1972, Ben Drake (a fellow geologist at Willow Run Laboratories of the University of Michigan, which became ERIM in January of 1973) simultaneously collected X-band and L-band images of the Lavic Lake region, which were field tested. At dark spots on the L-band images (where nothing appeared in the X-band images) were found ancient, hand-dug wells around the periphery of Lavic Lake (a dry lake), with abundant Native American pottery shards near them. It was assumed that the L-band radar had detected the top of the shallow water table. It was also noted that a basaltic lava outcrop in the dry lake with no visible connection to the main Písgah Crater lava flow had an apparent connection in the L-band image, but not in the X-band image. Only half a meter could be dug with a manual post-hole digger and dynamite was used to blast down to approximately 1.5 m, but the buried lava tube, the presence of which was believed to be there (but not proved), was never found. Dellwig's conclusion that "the real value of long-wavelength radar... is not its penetrative capability or any other single factor per se, but its unique total response to terrain characteristics, as compared to a shorter wavelength radar system" was agreed upon.

Much more recently, Abdelsalam and Stern reported on a NASA Space Shuttle Space Radar Laboratory experiment that simultaneously collected L-band (24 cm), C-band (6 cm), and X-band (3 cm) SAR data over the Sahara and Arabian deserts. They were able to locate the Keraf Suture, where an ancient collision between the Nubian Shield and the Nile Craton once occurred. The suture had not previously been mapped because windblown sand obscured it at visible and infrared wavelengths. They also found that the multifrequency radar data were able to separate igneous rocks, Cretaceous-aged sedimentary rocks, and Precambrian basement rocks for geological mapping, owing to their variations in surface roughness at these three wavelengths of active microwave radiation.

Vegetation can also provide differences in surface roughness that multifrequency radar can map. Podwysocki and colleagues found that Ka-band (0.86 cm) and X-band (3.5 cm) images displayed bright returns in the Escalante Desert of South Central Utah that had scattered cover of greasewood plants, which average 60–90 cm in height. Aeolian sand sheets and present-day playa surfaces appeared nearly black (smooth). In this case the vegetation, not visible on L-band (23.5 cm) radar images, provided a rough surface to the shorter wavelengths and a smooth surface to the longest wavelength radar.

This prompts the question whether differential penetration of vegetative canopies could provide useful information about both the vegetation and the terrain underneath the canopy. Another Willow Run Laboratories/ERIM experiment by Ben Drake in the early 1970s involved his placing corner reflectors on the ground in a deciduous forest in Michigan during the height of the growing season. When this was overflown by an X-band and L-band radar systems (simultaneous data collection), he examined separate images of both radar systems and could not detect the corner reflectors at either wavelength.

However, 2 years after Drake's experiment, Phil Jackson and coworkers created a spectral ratio image (really a ratio gray map produced with symbols on computer paper instead of on film) of X-band to L-band images from the same airborne radar system for a forested region (deciduous) in southeastern Kentucky (near Middlesboro) that showed evidence of canopy penetration by radar. [Figure 3\(a\)](#) and [3\(b\)](#) shows X-band and L-band images, respectively, of a larger region around the immediate study area. [Figure 3\(c\)](#) shows gray maps of the L-band, X-band, and a spectral ratio image of the two radar bands (L/X) for the immediate study area inside the dashed box in the two images in [Figure 3\(a\)](#) and [3\(b\)](#). The arcuate feature designated by 1's in the L-band and X-band images of [Figure 3\(c\)](#) is a 'bench,' or reclaimed area where coal mining has occurred on the steep hillside. It has a rough texture, with shale fragments ranging in size from 3 to 60 cm (though the majority of fragments were in the 3–15 cm range) and covered by grass. The dark 'bench' designated by 2 in the L-band and X-band images of [Figure 3\(c\)](#) (but bright in the L/X ratio image, like the surrounding terrain) was not level, but consisted of several slopes (less than 51°) parallel to the radar look direction. A flat field with rock fragments in approximately the 3–15 cm size range and covered by grass is designated with a 3 in the L-band and X-band images of [Figure 3\(c\)](#). Note that this flat area is brighter in the X-band image than in the L-band image, which is consistent with the theory that surfaces with roughness equal to or greater than half a wavelength appear rough (bright

return in the X-band image) and surfaces with roughness less than half a wavelength appear smooth (dark return in the L-band image). Above the '3' is a small, cross-like bright return on the L-band image that is not discernible on the X-band image. This was a bog with mud and some standing water, with a sparse stand of high weeds of 90–120 cm height. In the spectral ratio, gray map is an arcuate feature designated with a 4 that was found to be an elongated, elliptical depression. This is a real feature on the ground that was hidden in the radar shadow of the L-band and X-band images of [Figure 3\(c\)](#). The ratio image more clearly portrays the topographic information in the presence of apparent shadowing and slope factors, which cause the individual radar signals to vary greatly. The light-toned linear feature above the '5' in the spectral ratio image is an old prospecting cut near the top of the tree-covered mountain. It is a cut about 2 m wide, with a wall approximately 1.0–1.3 m high. The cut at the time of radar data collection was covered completely by trees that were not as mature as those on either side of it but were sufficiently mature not to have been visibly different from their surroundings in an aerial photograph of the area. It appeared to be seen that a ground feature in the spectral ratio image is hidden by trees in the two single-band SAR images of [Figure 3\(c\)](#).

These two experiments bring an important remote sensing dictum to the fore: linear and nonlinear image processing of coregistered spectral bands yields much more information than examination of images of the separate bands can yield. The reasons for this lie in the complexity of reflections of electromagnetic radiation from the surface and from the volume of the medium forming the interface with the air through which the radiation has propagated. There is usually both a surface specular component and a volume component to the reflected radiation. Because first surface reflectance is usually much less dependent on wavelength than the volume reflectance, the spectral ratioing of two bands of radar data tends to suppress the surface reflectance, thereby enhancing the wavelength dependence of the volume reflectance. The surface reflectance is often higher than the volume reflectance, which makes it dominate a single-band image. The ratio image, however, enhances the volume reflectance. What information is embedded in this volume reflectance for forested terrain? There are many layers of information: the total thickness and water content of leaves on the trees, the sizes of twigs and branches on the trees, the thickness of fallen leaves on the ground, and the soil moisture and roughness of the soil underneath the leaves. In the spectral ratio example above, it was either the different limb sizes of the different aged trees or reflections from the rocky ground underneath the trees that was permitted to observe the prospector's cut on the hill.

In seasonal wetlands covered by deciduous forests, it may be possible to use multifrequency radar to determine how far the water table lies beneath the fallen leaves on the forest floor. Generally speaking, however, the complexity of forest canopies creates a need for multifrequency radar. With the great variety of scatterers in forested terrain, more wavelengths (spectral bands) of radar data are required from which to separate the effects of those scatterers on backscattered microwaves.

The effect of an object's shape on radar backscatter is most pronounced when radar waves of like and crossed polarizations are employed for imaging (which is called

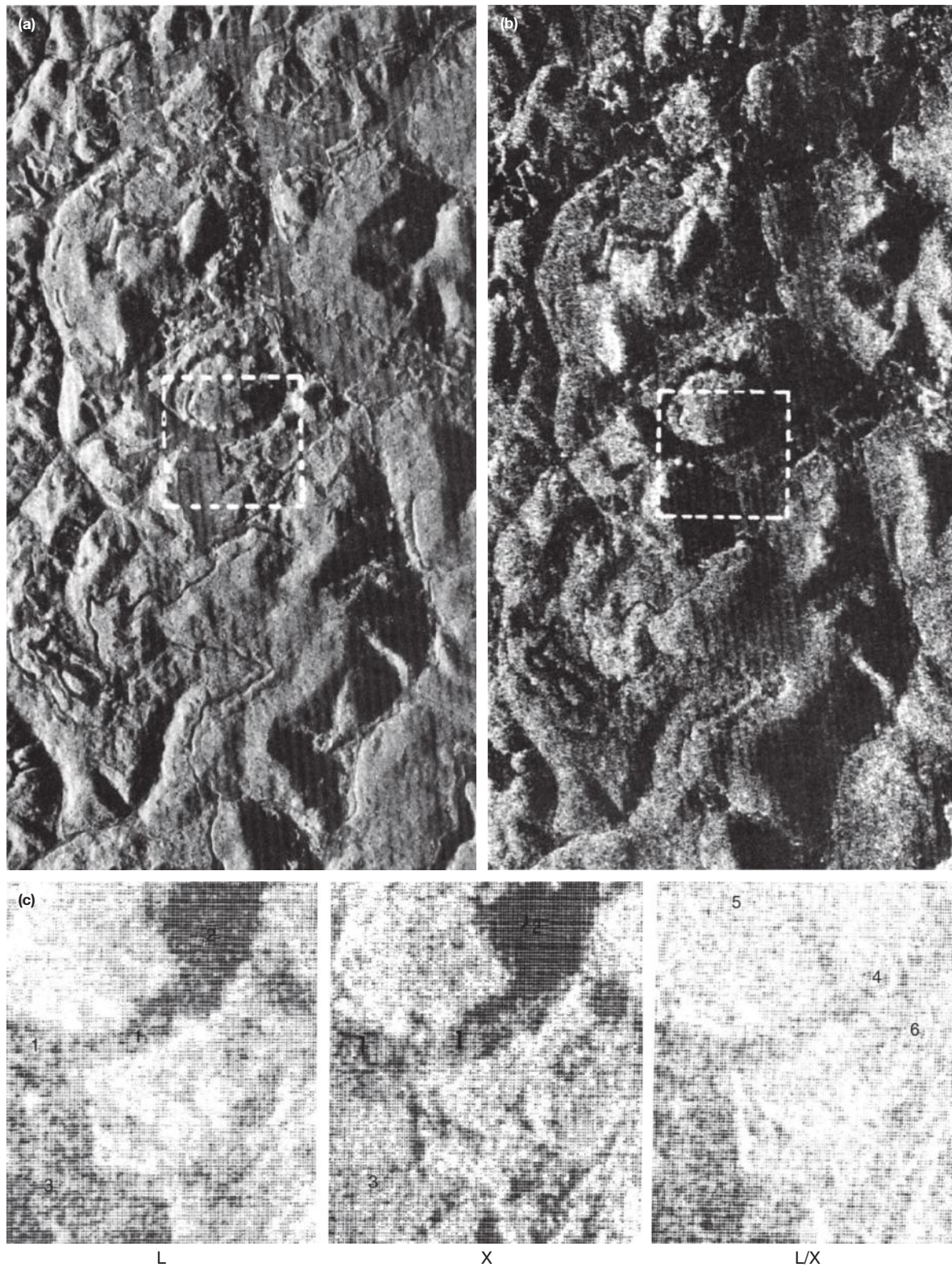


Figure 3 Synthetic aperture radar (SAR) data for southeastern Kentucky, near Middleboro. (a) X-band parallel polarization (HH); (b) L-band parallel polarization (HH); and (c) L-band (HH) image, X-band (HH) image, and L(HH)/X(HH) spectral ratio image from left to right, respectively, for the study area shown inside the dashed boxes in images (a) and (b). Reproduced from Jackson, P., Vincent, R.K., Willock, L., et al., 1975. Remote Sensing of Strip Mines – Final Report, ERIM Report No. 108500-14-F to the US Dept of Interior-Bureau of Mines (Grant No. S0241056), Environmental Research Institute of Michigan, PO Box 618, Ann Arbor, Michigan, pp. 34–40.

multipolarization radar) for whatever wavelength of microwave radiation to which it is applied. For instance, Ulaby found that sending and receiving vertically polarized radar waves (called VV for 'vertically filtered source and vertically filtered receiver') provided better classification of agricultural crops (corn in particular) than HH (horizontally filtered source and receiver), or cross-polarized (VH or HV) configurations. The vertical stalks of corn backscatter more vertically polarized radar waves than wave of other polarization orientations, and the radar waves stay vertically polarized on the way back to the detector. Because there are other types of crops that also have vertical stalks (such as sugarcane), multipolarization is an important capability for SAR systems that are intended to map agricultural crops. However, there are many other uses of multipolarized SAR data, which apply whenever the observed object has a preferred shape orientation. One example that has been reported in the literature is the detection of power line cables with VV polarization at angles away from normal incidence, when the backscatter was found to be proportional to the number and diameter of the strands on the surface of the cables. This is an important safety issue for low-altitude airplanes. Another example of multipolarization radar application (also multifrequency radar) is for the mapping of snow depth, liquid water content, and ice crystal size. Radar backscatter at 35 and 95 GHz frequencies, with every possible polarization combination of source and receiver, was experimentally measured and a numerical model was developed and adapted to fit the observations of snow-covered ground.

Forestry applications have already been shown to benefit greatly from multifrequency, multipolarization radar. Dobson and colleagues employed C-band, L-band, and P-band SAR data to study maritime pine plantations near Landes, France and near Duke, NC, USA. They found that they could determine the biomass of the coniferous forests, until the biomass saturated at a biomass level that depends on the radar frequency, ranging from 1001 ha⁻¹ at L-band wavelength to 2001 ha⁻¹ at P-band wavelength. The C-band radar was relatively insensitive to aboveground biomass, probably because of its lesser ability to penetrate the surface of the canopy, compared to the longer wavelength radars. Dobson then showed that X-band and C-band radars employed together were most effective for determining the crown biomass of a forest in northern Michigan. These two wavelengths, combined with L-band SAR, made it possible to determine tree height (0–23 m with 2.4 m RMS error), basal area (0–72 m² ha⁻¹ with 3.5 m² ha⁻¹ RMS error), dry trunk biomass (0–19 kg m⁻² with 1.1 kg m⁻² RMS error), dry crown biomass (0–6 kg m⁻² with 0.5 kg m⁻² RMS error), and total aboveground biomass (0–25 kg m⁻² with 1.4 kg m⁻² RMS error).

Summary and Conclusions

Past applications of multifrequency and/or multipolarization radar have included mapping of geological features, forests, agricultural crops, electric power lines, and snow cover. From experience of these applications, there appears to be additional future applications and general conclusions about multifrequency, multipolarization radar systems:

1. Although multifrequency radar offers no advantage for soil moisture estimations from SAR measurements, it could be used at low incidence angles to determine relative surface roughness, regardless of soil moisture.
2. Multifrequency radar may be able to determine the observation depth of a soil, through the oscillatory nature of the backscatter coefficient as a function of frequency.
3. Generally speaking, the complexity of forest canopies and agricultural crops creates a need for multifrequency, multipolarization radar, because the great variety of scatterers in forested terrain and agricultural fields require more wavelengths (spectral bands) and polarizations of radar data from which to separate the effects of those scatterers on backscattered microwaves.
4. Linear and nonlinear image processing of coregistered spectral bands yields much more information than examination of images of the separate bands can yield.

Radar systems, even with multifrequency, multipolarization capabilities, do not have to be used to the exclusion of data from other wavelength regions. As Ulaby has pointed out, the combination of multifrequency/multipolarization satellite radar data with LANDSAT visible/reflective infrared and thermal infrared data will lead to more accurate crop classifications from space. Not only do they provide better classifications when employed together but radar data can also penetrate clouds, thereby providing critical information during the growing season, even when clouds obscure the ground for shorter wavelengths of electromagnetic radiation. A multifrequency, multipolarization radar satellite in an orbit offset from and synchronous with a LANDSAT TM orbit would provide data from the visible, reflective infrared, thermal infrared, and microwave wavelength regions, collected simultaneously. The ability to map chemical composition variations on the ground while simultaneously mapping soil moisture would permit users to map trafficability for providing advice to farmers about the best times for farm machinery to be taken into the field without getting stuck, as well as for many other trafficability applications. Agriculture, forestry, geology, snow cover, power line, and trafficability mapping would all benefit from such a system.

See also: Chemistry of the Atmosphere: Observations for Chemistry (Remote Sensing): IR/FIR (Satellite, Balloon and Ground); Observations for Chemistry (Remote Sensing): Microwave. **Cryosphere:** Snow (Surface). **Hydrology, Floods and Droughts:** Soil Moisture.

Further Reading

- Abdelsalam, M.G., Stern, R.J., 1997. Utility of multifrequency imaging radar for hydrocarbon exploration in arid regions. AAPG Annual Meeting Abstracts, vol. 6, p. 1.
- Dellwig, L.F., 1969. An evaluation of multifrequency radar imagery of the Pishgah Crater Area, California. *Modern Geology* 1, 65–73.
- Dobson, M.C., Ulaby, F.T., Letoan, T., et al., 1992. Dependence of radar backscatter on coniferous forest biomass. *IEEE Transactions on Geoscience and Remote Sensing* 30 (2), 412–415.
- Dobson, M.C., Ulaby, F.T., Pierce, L.E., et al., 1995. Estimation of forest biophysical characteristics in Northern Michigan with SIR-C/X-SAR. *IEEE Transactions on Geoscience and Remote Sensing* 33 (4), 877–895.

- Jackson, P., Vincent, R.K., Willock, L., et al., 1975. Remote Sensing of Strip Mines – Final Report, ERIM Report No. 108500-14-F to the US Dept of Interior-Bureau of Mines (Grant No. S0241056), Environmental Research Institute of Michigan, PO Box 618, Ann Arbor, Michigan, pp. 34–40
- Mancini, M., Hoeben, R., Troch, P.A., 1999. Multifrequency radar observations of bare surface soil moisture content: a laboratory experiment. *Water Resources Research* 35 (6), 1827–1838.
- Podwysocki, M.H., Power, M.S., Koslow, M.H., 1985. Use of multifrequency radar images for botanical and lithologic mapping in an arid terrain. In: *Proceedings of the International Symposium on Remote Sensing of Environment, Fourth Thematic Conference in Remote Sensing for Exploration Geology*. ERIM, Ann Arbor, Michigan, p. 555.
- Sarabandi, K., Pierce, L., Oh, Y., Ulaby, F.T., 1994. Powerlines-radar measurements and detection algorithm for polarimetric SAR images. *IEEE Transactions on Aerospace and Electronic Systems* 30 (2), 632–643.
- Troch, P.A., Vandersteene, F., Su, Z., et al., 1997. Estimating microwave observation depth in bare soil through multifrequency scatterometry investigated. In: *1st EMSL User Workshop Proceedings*. Joint Research Centre, Ispra, Italy, pp. 45–53.
- Ulaby, F.T., 1981. Microwave response of vegetation. In: Kahle, A.B., Weill, G., Carter, W.D. (Eds.), 1980. *Advances in Space Research, COSPAR Interdisciplinary Scientific Commission A, 23rd Plenary Meeting in Budapest, Hungary*, vol. 1, no. 10, pp. 55–70.
- Ulaby, F.T., Siqueira, P., Nashashibi, A., Sarabandi, K., 1996. Semi-empirical model for radar backscatter from snow at 35 and 95 GHz. *IEEE Transactions on Geoscience and Remote Sensing* 34 (5), 1059–1065.
- Vincent, R.K., 1973. A Thermal Infrared Ratio Imaging Method for Mapping Compositional Variations among Silicate Rock Types. Ph.D. Thesis, Department of Geology and Mineralogy, University of Michigan, Ann Arbor, Michigan.
- Vincent, R.K., 1975. The potential role of thermal infrared multispectral scanners in geological remote sensing [Invited Paper]. *Proceedings of the IEEE* 63, 137–147.
- Vincent, R.K., 1997. *Fundamentals of Geological and Environmental Remote Sensing*. Prentice-Hall, Upper Saddle River, NJ.

Second Edition

ENCYCLOPEDIA OF ATMOSPHERIC SCIENCES

Edited by
Gerald R. North **Editor-in-Chief**
John Pyle and Fuqing Zhang



6 VOLUME SET

ENCYCLOPEDIA OF **ATMOSPHERIC SCIENCES**

SECOND EDITION

VOLUME 5

This page intentionally left blank

ENCYCLOPEDIA OF **ATMOSPHERIC SCIENCES**

SECOND EDITION

EDITOR-IN-CHIEF

GERALD R NORTH

Texas A&M University, College Station, TX, USA

EDITORS

JOHN PYLE

Cambridge University, Cambridge, UK

FUQING ZHANG

Pennsylvania State University, University Park, PA, USA

VOLUME 5



Amsterdam • Boston • Heidelberg • London • New York • Oxford
Paris • San Diego • San Francisco • Singapore • Sydney • Tokyo
Academic Press is an imprint of Elsevier



Academic Press is an imprint of Elsevier
32 Jamestown Road, London NW1 7BY, UK
225 Wyman Street, Waltham, MA 02451, USA
525 B Street, Suite 1800, San Diego, CA 92101-4495, USA

Copyright © 2015 Elsevier Ltd. unless otherwise stated. All rights reserved.

No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means electronic, mechanical, photocopying, recording or otherwise without the prior written permission of the publisher

Permissions may be sought directly from Elsevier's Science & Technology Rights Department in Oxford, UK: phone (+44) (0) 1865 843830; fax (+44) (0) 1865 853333; email: permissions@elsevier.com. Alternatively you can submit your request online by visiting the Elsevier web site at <http://elsevier.com/locate/permissions>, and selecting *Obtaining permission to use Elsevier material*

Notice

No responsibility is assumed by the publisher for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions or ideas contained in the material herein, Because of rapid advances in the medical sciences, in particular, independent verification of diagnoses and drug dosages should be made

British Library Cataloguing in Publication Data

A catalog record for this book is available from the British Library

Library of Congress Catalog Number:

A catalog record for this book is available from the Library of Congress

ISBN (print): 978-0-12-382225-3

For information on all Elsevier publications
visit our website at store.elsevier.com

Printed and bound in the United Kingdom

15 16 17 18 19 10 9 8 7 6 5 4 3 2 1

		Working together to grow libraries in developing countries
www.elsevier.com • www.bookaid.org		

<i>Acquisitions Editor:</i> Simon Holt <i>Project Manager:</i> Michael Nicholls <i>Associate Project Manager:</i> Marise Willis <i>Designer:</i> Matthew Limbert

DEDICATION

This second edition of the *Encyclopedia of Atmospheric Sciences* is dedicated to the memory of James Holton who was editor-in-chief of the first edition. He was a great researcher and colleague inspiring an entire generation of atmospheric scientists.

This page intentionally left blank

CONTENTS

<i>List of Contributors</i>	xxvii
<i>Preface to the First Edition</i>	xxxix
<i>Preface to the Second Edition</i>	xli
<i>Editor Biographies</i>	xliii
<i>Guide to Using the Encyclopedia</i>	xlvi

VOLUME 1

BASIC ATMOSPHERIC STRUCTURE AND CONCEPTS **1**

Beaufort Wind Scale	1
<i>L Hasse</i>	
Wind Chill	7
<i>M Bluestein</i>	
Standard Atmosphere	12
<i>W W Vaughan</i>	

AEROSOLS **17**

Aerosol–Cloud Interactions and Their Radiative Forcing	17
<i>U Lohmann</i>	
Aerosol Physics and Chemistry	23
<i>M Kalberer</i>	
Climatology of Stratospheric Aerosols	32
<i>L W Thomason and J-P Vernier</i>	
Climatology of Tropospheric Aerosols	40
<i>N Bellouin and J Haywood</i>	
Dust	48
<i>I N Sokolik</i>	
Observations and Measurements	53
<i>P H McMurry</i>	
Role in Radiative Transfer	66
<i>G A Ban-Weiss, and W D Collins</i>	

Role in Climate Change <i>N Bellouin</i>	76
Soot <i>P Chylek, S G Jennings, and R Pinnick</i>	86
Agricultural Meteorology and Climatology <i>E S Takle</i>	92
ARCTIC AND ANTARCTIC	98
Antarctic Climate <i>J Turner</i>	98
Arctic Climate <i>M C Serreze</i>	107
Arctic Haze <i>L M Russell and G E Shaw</i>	116
AIR SEA INTERACTIONS	
Freshwater Flux <i>J Schulz</i>	122
Momentum, Heat, and Vapor Fluxes <i>P K Taylor</i>	129
Sea Surface Temperature <i>W J Emery</i>	136
Surface Waves <i>A Benilov</i>	144
AVIATION METEOROLOGY	153
Aircraft Emissions <i>R R Friedl</i>	153
Aircraft Icing <i>M K Politovich</i>	160
Aviation Weather Hazards <i>A J Bedard, Jr</i>	166
Clear Air Turbulence <i>G P Ellrod (Retired), J A Knox, P F Lester, and L J Ehernberger (Retired)</i>	177
BIOGEOCHEMICAL CYCLES	187
Sulfur Cycle <i>P Brimblecombe</i>	187
Bromine <i>R von Glasow and C Hughes</i>	194
Heavy Metals <i>T D Jickells and A R Baker</i>	201

Iodine	205
<i>L J Carpenter</i>	
BOUNDARY LAYER (ATMOSPHERIC) AND AIR POLLUTION	220
Overview	220
<i>P J Mason and D J Thomson</i>	
Air Pollution Meteorology	227
<i>X-M Hu</i>	
Coherent Structures	237
<i>F T M Nieuwstadt and J C R Hunt</i>	
Complex Terrain	242
<i>J J Finnigan</i>	
Convective Boundary Layer	250
<i>M A LeMone</i>	
Microclimate	258
<i>M W Rotach and P Calanca</i>	
Modeling and Parameterization	265
<i>A A M Holtslag</i>	
Observational Techniques <i>In Situ</i>	274
<i>E F Bradley</i>	
Observational Techniques: Remote	284
<i>W M Angevine and C J Senff</i>	
Ocean Mixed Layer	290
<i>L Kantha and C A Clayson</i>	
Stably Stratified Boundary Layer	299
<i>L Mahrt</i>	
Surface Layer	305
<i>G L Geernaert</i>	
Urban Heat Islands	310
<i>J C Luvall, D A Quattrochi, D L Rickman, and M G Estes, Jr</i>	
Diurnal Cycle	319
<i>A Betts</i>	
CHEMISTRY OF THE ATMOSPHERE	324
Chemical Kinetics	324
<i>R P Wayne</i>	
Ion Chemistry	333
<i>J L Fox</i>	
Isotopes, Stable	348
<i>C A M Brenninkmeijer</i>	
Laboratory Kinetics	356
<i>D J Donaldson and S N Wren</i>	

Methane <i>E Dlugokencky, and S Houweling</i>	363
Observations for Chemistry (<i>In Situ</i>): Ozone Sondes <i>H G J Smit</i>	372
Observations for Chemistry (<i>In Situ</i>): Particles <i>T Deshler</i>	379
Observations for Chemistry (<i>In Situ</i>): Water Vapor Sondes <i>J B Smith</i>	387
Observations for Chemistry (Remote Sensing): IR/FIR (Satellite, Balloon and Ground) <i>H Fischer and F Hase</i>	401
Observations for Chemistry (Remote Sensing): Lidar <i>G Vaughan</i>	411
Observations for Chemistry (Remote Sensing): Microwave <i>J Waters</i>	418
Principles of Chemical Change <i>R P Wayne</i>	429
Radioactivity: Cosmogenic Radionuclides <i>D Lal</i>	437
Volcanoes: Composition of Emissions <i>M T Coffey and J W Hannigan</i>	446
Tracers <i>K A Boering</i>	450

VOLUME 2

CLIMATE AND CLIMATE CHANGE	1
Overview <i>D L Hartmann</i>	1
Carbon Dioxide <i>C L Sabine and R A Feely</i>	10
Climate Feedbacks <i>A E Dessler and M D Zelinka</i>	18
Climate Prediction: Empirical and Numerical <i>S Hastenrath</i>	26
Climate Variability: Decadal to Centennial Variability <i>D G Martinson</i>	33
Climate Variability: Nonlinear and Random Effects <i>M Ghil</i>	38
Climate Variability: North Atlantic and Arctic Oscillation <i>J W Hurrell</i>	47
Climate Variability: Seasonal and Interannual Variability <i>D S Gutzler</i>	61

Energy Balance Climate Models <i>G R North and K-Y Kim</i>	69
Global Impacts of the Madden–Julian Oscillation <i>C Zhang</i>	73
Greenhouse Effect <i>G R North</i>	80
History of Scientific Work on Climate Change <i>S Weart</i>	87
Intergovernmental Panel on Climate Change <i>K E Trenberth</i>	90
Nuclear Winter <i>A Robock</i>	95
Radiative–Convective Equilibrium Climate Models <i>N O Renno and X Huang</i>	102
Volcanoes: Role in Climate <i>A Robock</i>	105
CLOUDS AND FOG	112
Cloud Modeling <i>W-K Tao and M Moncrieff</i>	112
Contrails <i>P Minnis</i>	121
Cloud Microphysics <i>D Lamb</i>	133
Classification of Clouds <i>A L Rangno (Retiree)</i>	141
Climatology <i>S Warren, R Eastman, and C J Hahn</i>	161
Measurement Techniques <i>In situ</i> <i>D Baumgardner, J-F Gayet, A Korolev, C Twohy, and J Fugal</i>	170
Fog <i>P J Croft and B Ward</i>	180
Noctilucent Clouds <i>G E Thomas</i>	189
Stratus and Stratocumulus <i>R Wood</i>	196
CRYOSPHERE	201
Glaciers, Topography, and Climate <i>A B G Bush and M P Bishop</i>	201
Permafrost <i>T E Osterkamp and C R Burn</i>	208

Sea Ice	217
<i>M C Serreze, F Fetterer, and W F Weeks (Retired)</i>	
Snow (Surface)	227
<i>M Sturm</i>	
DATA ASSIMILATION AND PREDICTABILITY	237
Data Assimilation	237
<i>A C Lorenc</i>	
Ensemble-Based Data Assimilation	241
<i>Z Meng and F Zhang</i>	
Ensemble Prediction	248
<i>R Buizza</i>	
Predictability and Chaos	258
<i>L A Smith</i>	
DYNAMICAL METEOROLOGY	265
Overview	265
<i>J R Holton</i>	
Acoustic Waves	272
<i>K E Gilbert</i>	
Atmospheric Tides	287
<i>J Oberheide, M E Hagan, A D Richmond, and J M Forbes</i>	
Balanced Flow	298
<i>M E McIntyre</i>	
Baroclinic Instability	304
<i>R Grotjahn</i>	
Coriolis Force	313
<i>D W Moore</i>	
Critical Layers	317
<i>P Haynes</i>	
Hamiltonian Dynamics	324
<i>T G Shepherd</i>	
Hydraulic Flow	332
<i>R B Smith</i>	
Inertial Instability	334
<i>J A Knox</i>	
Kelvin–Helmholtz Instability	343
<i>P G Drazin</i>	
Kelvin Waves	347
<i>B Wang</i>	
Kinematics	353
<i>D D Houghton</i>	

Laboratory Geophysical Fluid Dynamics <i>R L Pfeffer</i>	360
Lagrangian Dynamics <i>I Roulstone</i>	369
Potential Vorticity <i>M E McIntyre</i>	375
Primitive Equations <i>A A White and N Wood</i>	384
Quasigeostrophic Theory <i>H C Davies and H Wernli</i>	393
Rossby Waves <i>P B Rhines</i>	404
Solitary Waves <i>J P Boyd</i>	417
Static Stability <i>J A Young</i>	423
Stationary Waves (Orographic and Thermally Forced) <i>S Nigam and E DeWeaver</i>	431
Symmetric Stability <i>H B Bluestein</i>	446
Vorticity <i>J R Holton</i>	451
Wave-CISK <i>C S Bretherton</i>	455
Wave Mean-Flow Interaction <i>M Juckes</i>	458
Waves <i>J R Holton</i>	464

VOLUME 3

ELECTRICITY IN THE ATMOSPHERE	1
Global Electrical Circuit <i>E R Williams</i>	1
Ions in the Atmosphere <i>K L Aplin and R G Harrison</i>	9
Lightning <i>M B Baker</i>	14
Sprites <i>W A Lyons</i>	20
Forensic Meteorology <i>L E Branscome</i>	28

GENERAL CIRCULATION OF THE ATMOSPHERE	33
Overview	33
<i>J M Wallace, D W J Thompson, and P Beresford</i>	
Angular Momentum of the Atmosphere	43
<i>D A Salstein</i>	
Energy Cycle	51
<i>R Grotjahn</i>	
Weather Regimes and Multiple Equilibria	65
<i>F Molteni</i>	
Mean Characteristics	73
<i>R Grotjahn</i>	
Teleconnections	90
<i>S Nigam and S Baxter</i>	
 GLOBAL CHANGE	 110
Climate Record: Surface Temperature Trends	110
<i>P D Jones</i>	
Sea Level Change	121
<i>R S Nerem</i>	
Upper Atmospheric Change	128
<i>R G Roble</i>	
Biospheric Impacts and Feedbacks	132
<i>B A Hungate and G W Koch</i>	
 GRAVITY WAVES	 141
Overview	141
<i>D C Fritts</i>	
Buoyancy and Buoyancy Waves: Optical Observations	153
<i>M J Taylor and W R Pendleton, Jr</i>	
Buoyancy and Buoyancy Waves: Theory	160
<i>T J Dunkerton</i>	
Gravity Waves Excited by Jets and Fronts	164
<i>R Plougonven and F Zhang</i>	
Convectively Generated Gravity Waves	171
<i>T P Lane</i>	
 HYDROLOGY, FLOODS AND DROUGHTS	 180
Overview	180
<i>R C Bales</i>	
Deserts and Desertification	185
<i>V P Tchakerian</i>	
Drought	193
<i>S Quiring</i>	

Flooding	201
<i>C A Doswell III</i>	
Groundwater and Surface Water	209
<i>S Ge and S M Gorelick</i>	
Modeling and Prediction	217
<i>Z Yu</i>	
Palmer Drought Severity Index	224
<i>L Nkemdirim</i>	
Soil Moisture	232
<i>A Robock</i>	
LAND-ATMOSPHERE INTERACTIONS	240
Overview	240
<i>R E Dickinson</i>	
Canopy Processes	244
<i>P D Blanken</i>	
Trace Gas Exchange	256
<i>J N Cape and D Fowler</i>	
LIDAR	262
Atmospheric Sounding Introduction	262
<i>P S Argall and R Sica</i>	
Backscatter	270
<i>C M R Platt and R L Collins</i>	
Differential Absorption Lidar	277
<i>S Ismail and E V Browell</i>	
Doppler	289
<i>R M Hardesty</i>	
Raman	296
<i>D N Whiteman</i>	
Resonance	305
<i>C S Gardner and R L Collins</i>	
Magnetosphere	309
<i>G K Parks</i>	
MESOSCALE METEOROLOGY	316
Overview	316
<i>D J Parker</i>	
Cloud and Precipitation Bands	323
<i>R M Rauber and M Ramamurthy</i>	
Gust Fronts	331
<i>R Rotunno</i>	

Hail and Hailstorms	334
<i>C Knight, N Knight, and H E Brooks</i>	
Mesoscale Convective Systems	339
<i>A Laing</i>	
Microbursts	335
<i>R M Wakimoto</i>	
Severe Storms	361
<i>C A Doswell III</i>	
Waterspouts	369
<i>J H Golden</i>	
Bow Echoes and Derecho	384
<i>M L Weisman</i>	
Density Currents	395
<i>P G Baines</i>	
Convective Storms: Overview	401
<i>M L Weisman</i>	
MESOSPHERE	411
Atomic Species in the Mesopause Region	411
<i>M G Mlynczak and L A Hunt</i>	
Ionosphere	422
<i>M C Kelley</i>	
Metal Layers	430
<i>J M C Plane</i>	
Polar Summer Mesopause	436
<i>R H Varney and M C Kelley</i>	
VOLUME 4	
MIDDLE ATMOSPHERE	1
Planetary Waves	1
<i>A K Smith and J Perlwitz</i>	
Polar Vortex	12
<i>M R Schoeberl and P A Newman</i>	
Quasi-Biennial Oscillation	18
<i>T J Dunkerton, J A Anstey, and L J Gray</i>	
Semiannual Oscillation	26
<i>K Hamilton</i>	
Stratospheric Sudden Warmings	30
<i>A O'Neill, A J Charlton-Perez, and L M Polvani</i>	
Transport Circulation	41
<i>S E Strahan</i>	

Zonal Mean Climatology <i>P Braesicke</i>	50
MOUNTAIN METEOROLOGY	57
Overview <i>R B Smith</i>	57
Cold Air Damming <i>B A Colle</i>	62
Downslope Winds <i>D R Durran</i>	69
Katabatic Winds <i>T R Parish</i>	75
Land and Sea Breezes <i>R A Pielke, Sr</i>	80
Lee Vortices <i>C C Epifanio</i>	84
Lee Waves and Mountain Waves <i>D R Durran</i>	95
Orographic Effects: Lee Cyclogenesis <i>C Schär</i>	103
Valley Winds <i>D Zardi</i>	114
NUMERICAL MODELS	135
Chemistry Models <i>M P Chipperfield and S R Arnold</i>	135
Coupled Ocean-Atmosphere Models: Physical Processes <i>M Zhang</i>	144
General Circulation Models <i>C R Mechoso and A Arakawa</i>	153
Methods <i>J Thuburn</i>	161
Model Physics Parameterization <i>D J Stensrud, M C Coniglio, K H Knopfmeier, and A J Clark</i>	167
Parameter Estimation <i>A Aksoy</i>	181
Parameterization of Physical Processes: Clouds <i>R Forbes, C Jakob, and M Miller</i>	187
Parameterization of Physical Processes: Gravity Wave Fluxes <i>M J Alexander</i>	194
Parameterization of Physical Processes: Turbulence and Mixing <i>A Beljaars</i>	200

Spectral Models <i>F Baer</i>	212
Mesoscale Atmospheric Modeling <i>R A Pielke, Sr</i>	219
Cloud-System Resolving Modeling and Aerosols <i>W-K Tao and T Matsui</i>	222
Large-Eddy Simulation <i>C-H Moeng and P P Sullivan</i>	232
Regional Prediction Models <i>B W Golding</i>	241
Convective Storm Modeling <i>M D Parker</i>	246
OBSERVATIONS PLATFORMS	255
Balloons <i>J-P Pommereau</i>	255
Buoys <i>J M Hemsley</i>	264
Kites <i>B B Balsley</i>	268
Radiosondes <i>W F Dabberdt and H Turtiainen</i>	273
Rockets <i>M F Larsen</i>	285
OCEANOGRAPHIC TOPICS	290
General Processes <i>N C Wells</i>	290
Surface/Wind Driven Circulation <i>R X Huang</i>	301
Thermohaline Circulation <i>R X Huang</i>	315
Water Types and Water Masses <i>W J Emery</i>	329
OPTICS, ATMOSPHERIC	338
Optical Remote Sensing Instruments <i>G G Shepherd</i>	338
Airglow Instrumentation <i>M Conde</i>	346

OZONE DEPLETION AND RELATED TOPICS	353
Long-Term Ozone Changes <i>N R P Harris</i>	353
Ozone as a UV Filter <i>J E Frederick</i>	359
Ozone Depletion Potentials <i>D J Wuebbles</i>	364
Photochemistry of Ozone <i>G K Moortgat and A R Ravishankara</i>	370
Stratospheric Ozone Recovery <i>D J Hofmann and R Müller</i>	380
Surface Ozone Effects on Vegetation <i>M Ashmore</i>	389
Surface Ozone (Human Health) <i>M Lippmann</i>	397
PALEOCLIMATOLOGY	404
Ice Cores <i>E J Steig</i>	404
Varves <i>R Gilbert</i>	411
RADAR	415
Cloud Radar <i>T Uttal</i>	415
Incoherent Scatter Radar <i>M P Sulzer</i>	422
Mesosphere—Stratosphere—Troposphere and Stratosphere—Troposphere Radars and Wind Profilers <i>G Vaughan and D Hooper</i>	429
Meteor Radar <i>N J Mitchell</i>	438
Polarimetric Doppler Weather Radar <i>R J Doviak and R D Palmer</i>	444
Precipitation Radar <i>S E Yuter</i>	455
Synthetic Aperture Radar (Land Surface Applications) <i>R K Vincent</i>	470
VOLUME 5	
RADIATION TRANSFER IN THE ATMOSPHERE	1
Radiation, Solar <i>Q Fu</i>	1

Absorption and Thermal Emission <i>R M Goody and X Huang</i>	5
Cloud-Radiative Processes <i>Q Fu</i>	13
Non-local Thermodynamic Equilibrium <i>M López-Puertas and B Funke</i>	16
Scattering <i>M Mishchenko, L Travis, and A Lacis</i>	27
Ultraviolet Radiation <i>K Stamnes</i>	37
Ultraviolet, Surface <i>R McKenzie and S Madronich</i>	45
SATELLITES AND SATELLITE REMOTE SENSING	51
Aerosol Measurements <i>R A Kahn</i>	51
Earth's Radiation Budget <i>N G Loeb and B A Wielicki</i>	67
GPS Meteorology <i>S S Leroy</i>	77
Measuring Ozone from Space — TOMS and SBUV <i>R D McPeters and R S Stolarski</i>	87
Orbits <i>S Q Kidder</i>	95
Precipitation <i>G Liu</i>	107
Remote Sensing: Cloud Properties <i>P Yang and B A Baum</i>	116
Research <i>M D King</i>	128
Surface Wind and Stress <i>W T Liu</i>	138
Temperature Soundings <i>A Dudhia</i>	145
Water Vapor <i>J E Harries</i>	157
SOLAR SYSTEM/SUN, ATMOSPHERES, EVOLUTION OF ATMOSPHERES	163
Evolution of Earth's Atmosphere <i>Y L Yung, M L Wong, and E J Gaidos</i>	163
Planetary Atmospheres: Mars <i>R M Haberle</i>	168

Planetary Atmospheres: Venus <i>P J Gierasch and Y L Yung</i>	178
Solar Terrestrial Interactions: Climate Impact <i>J D Haigh</i>	183
Solar Winds <i>S T Suess and B T Tsurutani</i>	189
Meteors <i>P Jenniskens</i>	195
STATISTICAL METHODS	201
Data Analysis: Empirical Orthogonal Functions and Singular Vectors <i>C S Bretherton</i>	201
Data Analysis: Time Series Analysis <i>G R North</i>	205
STRATOSPHERIC CHEMISTRY TOPICS	211
Overview <i>J A Pyle</i>	211
Halogens <i>D Toohey</i>	215
Halogen Sources, Anthropogenic <i>A McCulloch and P M Midgley</i>	221
Halogen Sources, Natural (Methyl Bromide and Related Gases) <i>S Yvon-Lewis and J H Butler</i>	228
HO _x <i>T F Hanisco</i>	233
Hydrogen Budget <i>J E Harries</i>	238
Reactive Nitrogen (NO _x and NO _y) <i>Y Kondo</i>	242
Stratospheric Water Vapor <i>K H Rosenlof</i>	250
STRATOSPHERE/TROPOSPHERE EXCHANGE AND STRUCTURE	257
Global Aspects <i>J R Holton</i>	257
Local Processes <i>J F Lamarque and P Hess</i>	262
Tropopause <i>M Dameris</i>	269

SYNOPTIC METEOROLOGY	273
Anticyclones <i>S J Colucci</i>	273
Forecasting <i>D Mansfield</i>	280
Weather Maps <i>R Reynolds</i>	289
Cyclogenesis <i>G J Hakim</i>	299
Extratropical Cyclones <i>A Joly</i>	304
Fronts <i>D M (David) Schultz and W Blumen</i>	337
Fronts in the Lower Stratosphere <i>A L Lang</i>	344
Frontogenesis <i>D M (David) Schultz</i>	353
Jet Streaks <i>P Cunningham and D Keyser</i>	359
Lake-Effect Storms <i>P J Sousounis</i>	370
Polar Lows <i>I A Renfrew</i>	379
Thermal Low <i>R H Johnson</i>	386
 THERMODYNAMICS	 391
Humidity Variables <i>J A Curry</i>	391
Moist (Unsaturated) Air <i>J A Curry</i>	394
Saturated Adiabatic Processes <i>J A Curry</i>	398
 Thermosphere <i>S C Solomon and R G Roble</i>	 402
 VOLUME 6	
TROPICAL CYCLONES AND HURRICANES	1
Overview and Theory <i>R A Tomas and P J Webster</i>	1

Hurricane Dynamics <i>Y Wang</i>	8
Hurricane Predictability <i>J A Sippel</i>	30
Hurricanes: Observation <i>F D Marks</i>	35
Tropical Cyclogenesis <i>Z Wang</i>	57
Tropical Cyclones and Climate Change <i>T R Knutson</i>	65
Tropical Cyclones in the Western North Pacific <i>J C L Chan</i>	77
Tropical Cyclones: Secondary Eyewall Formation <i>C-C Wu and Y-H Huang</i>	85
TROPICAL METEOROLOGY AND CLIMATE	91
El Niño and the Southern Oscillation: Observation <i>N Nicholls</i>	91
El Niño and the Southern Oscillation: Theory <i>P Chang and S E Zebiak</i>	97
Equatorial Waves <i>M C Wheeler and H Nguyen</i>	102
Hadley Circulation <i>J Lu and G A Vecchi</i>	113
Intertropical Convergence Zone <i>D E Waliser and X Jiang</i>	121
Intraseasonal Oscillation (Madden–Julian Oscillation) <i>R A Madden</i>	132
Madden–Julian Oscillation: Skeleton and Conceptual Models <i>A J Majda and S N Stechmann</i>	137
Monsoon: Overview <i>J Slingo</i>	146
Monsoon: Dynamical Theory <i>P J Webster and J Fasullo</i>	151
Monsoon: ENSO–Monsoon Interactions <i>K-M Lau</i>	165
Tropical Climates <i>S Hastenrath</i>	170
Walker Circulation <i>K-M Lau and S Yang</i>	177

TROPOSPHERIC CHEMISTRY AND COMPOSITION	182
Aerosols/Particles <i>J H Seinfeld</i>	182
Aliphatic Hydrocarbons <i>J Rudolph and O Stein</i>	188
Aromatic Hydrocarbons <i>I Barnes</i>	204
Biogenic Hydrocarbons <i>A Guenther</i>	214
Cloud Chemistry <i>P Herckes and J L Collett, Jr</i>	218
H ₂ <i>U Schmidt and T Wetter</i>	226
Hydroxyl Radical <i>K C Clemitshaw</i>	232
Mercury <i>J Munthe and J Sommar</i>	239
Oxidizing Capacity <i>D H Ehhalt, F Rohrer, and A Wahner</i>	243
Peroxyacetyl Nitrate <i>H B Singh</i>	251
Sulfur Chemistry, Organic <i>I Barnes</i>	255
Volatile Organic Compounds Overview: Anthropogenic <i>R G Derwent</i>	265
TURBULENCE AND MIXING	268
Overview <i>P Haynes</i>	268
Turbulence, Two Dimensional <i>P Bartello</i>	273
Turbulent Diffusion <i>A Venkatram and S Du</i>	277
WEATHER FORECASTING	287
Marine Meteorology <i>L Xie and B Liu</i>	287
Operational Meteorology <i>D R Novak</i>	293
Seasonal and Interannual Weather Prediction <i>J P Li and R Q Ding</i>	303
Severe Weather Forecasting <i>D J Stensrud, H E Brooks, and S J Weiss</i>	313

Wildfire Weather <i>J Coen</i>	323
Inadvertant Weather Modification <i>S A Changnon</i>	332
<i>Appendix 1: Physical Constants</i>	337
<i>Appendix 2: Units and their SI Equivalents</i>	339
<i>Appendix 3: Periodic Table of the Elements</i>	340
<i>Appendix 4: The Geologic Time Scale</i>	341
<i>Index</i>	343

This page intentionally left blank

LIST OF CONTRIBUTORS

A. Aksoy
University of Miami, Miami, FL, USA; and NOAA Hurricane Research Division, Miami, FL, USA

M.J. Alexander
NorthWest Research Associates (NWRA), Boulder, CO, USA

W.M. Angevine
CIRES, University of Colorado; and NOAA Earth System Research Laboratory, Boulder, CO, USA

J.A. Anstey
University of Oxford, Oxford, UK

K.L. Aplin
University of Oxford, Oxford, UK

A. Arakawa
University of California, Los Angeles, CA, USA

P.S. Argall
The University of Western Ontario, London, ON, Canada

S.R. Arnold
University of Leeds, Leeds, UK

M. Ashmore
University of York, York, UK

F. Baer
University of Maryland, College Park, MD, USA

P.G. Baines
University of Melbourne, Melbourne, VIC, Australia

A.R. Baker
University of East Anglia, Norwich, UK

M.B. Baker
University of Washington, Seattle, WA, USA

R.C. Bales
University of Arizona, Tucson, AZ, USA

B.B. Balsley
University of Colorado, Boulder, CO, USA

G.A. Ban-Weiss
Lawrence Berkeley National Laboratory, Berkeley, CA, USA; and University of Southern California, Los Angeles, CA, USA

I. Barnes
University of Wuppertal, Wuppertal, Germany

P. Bartello
McGill University, Montréal, QC, Canada

B.A. Baum
University of Wisconsin–Madison, Madison, WI, USA

D. Baumgardner
Universidad Nacional Autónoma de México, Mexico City, D.F., Mexico

S. Baxter
University of Maryland, College Park, MD, USA

A.J. Bedard, Jr.
National Oceanic and Atmospheric Administration, Boulder, CO, USA

A. Beljaars
European Centre for Medium-Range Weather Forecasts, Reading, England

N. Bellouin
University of Reading, Reading, UK

A. Benilov
Acute Solutions, Highlands, NJ, USA

P. Beresford
European Centre for Medium-Range Weather Forecasts, Reading, UK

A. Betts
Atmospheric Research, Pittsford, VT, USA

M.P. Bishop
Texas A&M University, College Station, TX, USA

P.D. Blanken
University of Colorado at Boulder, Boulder, CO, USA

H.B. Bluestein
University of Oklahoma, Norman, OK, USA

M. Bluestein
Indiana University – Purdue University, Indianapolis, IN, USA

W. Blumen[†]
University of Colorado Boulder, Boulder, CO, USA

K.A. Boering
University of California – Berkeley, Berkeley, CA, USA

J.P. Boyd
University of Michigan, Ann Arbor, MI, USA

E.F. Bradley
CSIRO Land and Water, Canberra, ACT, Australia

P. Braesicke
Karlsruhe Institute of Technology, Karlsruhe, Germany

L.E. Branscome
Climatological Consulting Corporation, FL, USA

C.A.M. Brenninkmeijer
Max Planck Institute for Chemistry, Mainz, Germany

C.S. Bretherton
University of Washington, Seattle, WA, USA

P. Brimblecombe
University of East Anglia, Norwich, UK

H.E. Brooks
National Oceanic and Atmospheric Administration, Norman, OK, USA

E.V. Browell
STARS II Affiliate, NASA Langley Research Center, Hampton, VA, USA

R. Buizza
ECMWF, Reading, UK

C.R. Burn
Carleton University, Ottawa, ON, Canada

A.B.G. Bush
University of Alberta, Edmonton, AB, Canada

J.H. Butler
NOAA Earth System Research Laboratory, Boulder, CO, USA

P. Calanca
Agroscope Reckenholz-Taenikon, Zurich, Switzerland

J.N. Cape
Edinburgh Research Station, Midlothian, UK

L.J. Carpenter
University of York, York, UK

J.C.L. Chan
City University of Hong Kong, Hong Kong

P. Chang
Texas A&M University, College Station, TX, USA

S.A. Changnon
University of Illinois, IL, USA

A.J. Charlton-Perez
University of Reading, Earley Gate, Reading, UK

M.P. Chipperfield
University of Leeds, Leeds, UK

P. Chylek
Dalhousie University, NS, Canada

A.J. Clark
University of Oklahoma and National Oceanic and Atmospheric Administration, Norman, OK, USA

C.A. Clayson
Woods Hole Oceanographic Institution, Woods Hole, MA, USA

K.C. Clemitshaw
Imperial College of Science, Technology, and Medicine, Ascot, UK

J. Coen
National Center for Atmospheric Research, Boulder, CO, USA

M.T. Coffey
National Center for Atmospheric Research, Boulder, CO, USA

B.A. Colle
Stony Brook University – SUNY, Stony Brook, NY, USA

J.L. Collett, Jr.
Colorado State University, Fort Collins, CO, USA

R.L. Collins
University of Alaska Fairbanks, Fairbanks, AK, USA

W.D. Collins
Lawrence Berkeley National Laboratory, Berkeley, CA, USA

S.J. Colucci
Cornell University, Ithaca, NY, USA

M. Conde
University of Alaska Fairbanks, Fairbanks, AK, USA

M.C. Coniglio
National Oceanic and Atmospheric Administration, Norman, OK, USA

[†] Deceased.

- P.J. Croft
Kean University, Union, NJ, USA
- P. Cunningham
Florida State University, Tallahassee, FL, USA
- J.A. Curry
Georgia Institute of Technology, Atlanta, GA, USA
- W.F. Dabberdt
Vaisala Company, Boulder, CO, USA
- M. Dameris
Deutsches Zentrum für Luft- und Raumfahrt, Institut für Physik der Atmosphäre, Oberpfaffenhofen, Wessling, Germany
- H.C. Davies
Institute for Atmospheric and Climate Science, Swiss Federal Institute of Technology (ETH), Zürich, Switzerland
- R.G. Derwent
rdscientific, Newbury, UK
- T. Deshler
University of Wyoming, Laramie, WY, USA
- A.E. Dessler
Texas A&M University, College Station, TX, USA
- E. DeWeaver
University of Wisconsin, Madison, WI, USA
- R.E. Dickinson
University of Texas at Austin, Austin, TX, USA
- R.Q. Ding
Institute of Atmospheric Physics, Chinese Academy of Sciences, Beijing, China
- E. Dlugokencky
NOAA Earth System Research Laboratory, Boulder, CO, USA
- D.J. Donaldson
University of Toronto, Toronto, ON, Canada
- C.A. Doswell, III
University of Oklahoma, Norman, OK, USA
- R.J. Doviak
National Severe Storms Laboratory, Norman, OK, USA
- P.G. Drazin[†]
University of Bath, England, UK
- S. Du
California Air Resources Board, Sacramento, CA, USA
- A. Dudhia
University of Oxford, Oxford, UK
- T.J. Dunkerton
Northwest Research Associates, Bellevue, WA, USA
- D.R. Durran
University of Washington, Seattle, WA, USA
- R. Eastman
University of Washington, Seattle, WA, USA
- L.J. Ehernberger
National Aeronautics and Space Administration, Dryden Flight Research Center, Edwards, CA, USA
- D.H. Ehhalt
Forschungszentrum Jülich, Jülich, Germany
- G.P. Ellrod
National Oceanographic and Atmospheric Administration/National Environmental Satellite, Data, and Information Service, Granby, CT, USA
- W.J. Emery
University of Colorado, Boulder, CO, USA
- C.C. Epifanio
Texas A&M University, College Station, TX, USA
- M.G. Estes
Universities Space Research Association, Huntsville, AL, USA
- J. Fasullo
University of Colorado – Boulder, Boulder, CO, USA
- R.A. Feely
NOAA Pacific Marine Environmental Laboratory, Seattle, WA, USA
- F. Fetterer
University of Colorado, Boulder, CO, USA
- J.J. Finnigan
CSIRO Atmospheric Research, Black Mountain, ACT, Australia
- H. Fischer
Institute of Meteorology and Climate Research (IMK), Karlsruhe Institute of Technology, Karlsruhe, Germany
- J.M. Forbes
University of Colorado, Boulder, CO, USA
- R. Forbes
European Centre for Medium-Range Weather Forecasts, Reading, UK
- D. Fowler
Edinburgh Research Station, Midlothian, UK

[†] Deceased.

J.L. Fox
Wright State University, Dayton, OH, USA

J.E. Frederick
The University of Chicago, Chicago, IL, USA

R.R. Friedl
California Institute of Technology, Pasadena, CA, USA

D.C. Fritts
GATS Inc., Boulder, CO, USA

Q. Fu
University of Washington, Seattle, WA, USA

J. Fugal
Max Planck Institute of Chemistry, Mainz, Germany

B. Funke
Instituto de Astrofísica de Andalucía, CSIC, Granada, Spain

E.J. Gaidos
University of Hawaii at Manoa, Honolulu, HI, USA

C.S. Gardner
University of Illinois at Urbana-Champaign, Urbana, IL, USA

J.-F. Gayet
Université Blaise Pascal, Clermont Ferrand, France

S. Ge
University of Colorado, Boulder, CO, USA

G.L. Geernaert
US Department of Energy, Washington, DC, USA

M. Ghil
Ecole Normale Supérieure, Paris, France; and University of California, Los Angeles, CA, USA

P.J. Gierasch
Cornell University, Ithaca, NY, USA

K.E. Gilbert
University of Mississippi, University, MS, USA

R. Gilbert
Queen's University, Kingston, ON, Canada

J.H. Golden
Forecast Systems Laboratory, NOAA, Boulder, CO, USA

B.W. Golding
Met Office, Exeter, UK

R.M. Goody
Harvard University (Emeritus), Cambridge, MA, USA

S.M. Gorelick
Stanford University, Stanford, CA, USA

L.J. Gray
University of Oxford, Oxford, UK

R. Grotjahn
University of California, Davis, CA, USA

A. Guenther
Pacific Northwest National Laboratory, Richland, WA, USA

D.S. Gutzler
University of New Mexico, Albuquerque, NM, USA

R.M. Haberle
NASA/Ames Research Center, Moffett Field, Mountain View, CA, USA

M.E. Hagan
National Center for Atmospheric Research, Boulder, CO, USA

C.J. Hahn
University of Arizona, Tucson, AZ, USA

J.D. Haigh
Blackett Laboratory, Imperial College London, London, UK

G.J. Hakim
University of Washington, Seattle, WA, USA

K. Hamilton
University of Hawaii, Honolulu, HI, USA

T.F. Hanisco
Harvard University, Cambridge, MA, USA

J.W. Hannigan
National Center for Atmospheric Research, Boulder, CO, USA

R.M. Hardesty
NOAA Environmental Technology Laboratory, Boulder, CO, USA

J.E. Harries
Imperial College of Science, Technology and Medicine, Blackett Laboratory, London, UK

N.R.P. Harris
University of Cambridge, Cambridge, UK

R.G. Harrison
The University of Reading, Reading, UK

D.L. Hartmann
University of Washington, Seattle, WA, USA

F. Hase
Institute of Meteorology and Climate Research (IMK), Karlsruhe Institute of Technology, Karlsruhe, Germany

- L. Hasse
Universität Kiel, Kiel, Germany
- S. Hastenrath
University of Wisconsin, Madison, WI, USA
- P. Haynes
University of Cambridge, Cambridge, UK
- J. Haywood
Met Office, Exeter, UK
- J.M. Hemsley
National Data Buoy Center, Stennis Space Center, MS, USA
- P. Herckes
Arizona State University, Tempe, AZ, USA
- P. Hess
National Center for Atmospheric Research, Boulder, CO, USA
- D.J. Hofmann[†]
NOAA Climate Monitoring and Diagnostics Laboratory, Boulder, CO, USA
- J.R. Holton
University of Washington, Seattle, WA, USA
- A.A.M. Holtslag
Wageningen University, Wageningen, The Netherlands
- D. Hooper
Science & Technology Facilities Council (STFC), Didcot, UK
- D.D. Houghton
University of Wisconsin-Madison, Madison, WI, USA
- S. Houweling
SRON Netherlands Institute for Space Research, Utrecht, The Netherlands
- X.-M. Hu
University of Oklahoma, Norman, OK, USA
- R.X. Huang
Woods Hole Oceanographic Institution, Woods Hole, MA, USA
- X. Huang
University of Michigan, Ann Arbor, MI, USA
- Y.-H. Huang
National Taiwan University, Taipei, Taiwan
- C. Hughes
University of York, York, UK
- B.A. Hungate
Northern Arizona University, Flagstaff, AZ, USA
- J.C.R. Hunt
University College London, London, UK
- L.A. Hunt
Science Systems and Applications Incorporated, Hampton, VA, USA
- J.W. Hurrell
National Center for Atmospheric Research, Boulder, CO, USA
- S. Ismail
Science Directorate, NASA Langley Research Center, Hampton, VA, USA
- C. Jakob
Monash University, VIC, Australia
- S.G. Jennings
National University of Ireland, Galway, Ireland
- P. Jenniskens
SETI Institute, Moffett Field, CA, USA
- X. Jiang
University of California, Los Angeles, CA, USA
- T.D. Jickells
University of East Anglia, Norwich, UK
- R.H. Johnson
Colorado State University, Fort Collins, CO, USA
- A. Joly
Centre National de Recherches Météorologiques – Groupe d'étude de l'Atmosphère Météorologique, Météo-France and CNRS, Toulouse, France
- P.D. Jones
Climatic Research Unit, University of East Anglia, Norwich, UK
- M. Jukes
University of Oxford, Oxford, UK
- R.A. Kahn
NASA Goddard Space Flight Center, Greenbelt, MD, USA
- M. Kalberer
University of Cambridge, Cambridge, UK
- L. Kantha
University of Colorado, Boulder, CO, USA
- M.C. Kelley
Cornell University, Ithaca, NY, USA

[†] Deceased.

D. Keyser
*University at Albany, State University of New York,
Albany, NY, USA*

S.Q. Kidder
*Colorado State University, Fort Collins,
CO, USA*

K.-Y. Kim
Seoul National University, Seoul, Korea

M.D. King
University of Colorado, Boulder, CO, USA

C. Knight
*National Center for Atmospheric Research, Boulder,
CO, USA*

N. Knight
*National Center for Atmospheric Research, Boulder,
CO, USA*

K.H. Knopfmeier
*University of Oklahoma; and National Oceanic and
Atmospheric Administration, Norman, OK, USA*

J.A. Knox
University of Georgia, Athens, GA, USA

T.R. Knutson
*NOAA Geophysical Fluid Dynamics Laboratory,
Princeton, NJ, USA*

G.W. Koch
Northern Arizona University, Flagstaff, AZ, USA

Y. Kondo
The University of Tokyo, Tokyo, Japan

A. Korolev
*Meteorological Service of Canada, Toronto, ON,
Canada*

A. Lacis
*Goddard Institute for Space Studies, New York,
NY, USA*

A. Laing
*National Center for Atmospheric Research, Boulder,
CO, USA*

D. Lal
*Scripps Institution of Oceanography, La Jolla,
CA, USA*

J.F. Lamarque
*National Center for Atmospheric Research, Boulder,
CO, USA*

D. Lamb
*The Pennsylvania State University, University Park,
PA, USA*

T.P. Lane
*The University of Melbourne, Melbourne, VIC,
Australia*

A.L. Lang
*University of Albany – State University of New York,
Albany, NY, USA*

M.F. Larsen
Clemson University, Clemson, SC, USA

K.-M. Lau
*NASA/Goddard Space Flight Center, Greenbelt, MD,
USA*

M.A. LeMone
*National Center for Atmospheric Research, Boulder,
CO, USA*

S.S. Leroy
*Harvard School of Engineering and Applied Sciences,
Cambridge, MA, USA*

P.F. Lester
San Jose State University, San Jose, CA, USA

J.P. Li
*Institute of Atmospheric Physics, Chinese Academy of
Sciences, Beijing, China*

M. Lippmann
New York University, Tuxedo, NY, USA

B. Liu
North Carolina State University, Raleigh, NC, USA

G. Liu
Florida State University, Tallahassee, FL, USA

W.T. Liu
*California Institute of Technology, Pasadena,
CA, USA*

N.G. Loeb
NASA Langley Research Center, Hampton, VA, USA

U. Lohmann
ETH Zurich, Zürich, Switzerland

M. López-Puertas
*Instituto de Astrofísica de Andalucía, CSIC, Granada,
Spain*

A.C. Lorenc
The Met Office, Bracknell, Berkshire, UK

J. Lu
*Pacific Northwest National Laboratory, Richland, WA,
USA*

J.C. Luvall
*National Space Science and Technology Center,
Marshall Space Flight Center, Huntsville, AL, USA*

- W.A. Lyons
FMA Research Inc., Fort Collins, CO, USA
- R.A. Madden
National Center for Atmospheric Research, Boulder, CO, USA
- S. Madronich
National Center for Atmospheric Research, Boulder, CO, USA
- L. Mahrt
Oregon State University, Corvallis, OR, USA
- A.J. Majda
New York University, New York, NY, USA
- D. Mansfield
National Meteorological Center, Bracknell, UK
- F.D. Marks
Hurricane Research Division, Miami, FL, USA
- D.G. Martinson
Columbia University, Palisades, NY, USA
- P.J. Mason
Met Office, Bracknell, UK
- T. Matsui
NASA/Goddard Space Flight Center, Greenbelt, MD, USA; and University of Maryland, College Park, MD, USA
- A. McCulloch
University of Bristol, Bristol, UK
- M.E. McIntyre
University of Cambridge, Cambridge, UK
- R. McKenzie
National Institute for Water and Atmospheric Research, Lauder, Central Otago, New Zealand
- P.H. McMurtry
University of Minnesota, Minneapolis, MN, USA
- R.D. McPeters
NASA Goddard Space Flight Center, Greenbelt, MD, USA
- C.R. Mechoso
University of California, Los Angeles, CA, USA
- Z. Meng
Peking University, Beijing, China
- P.M. Midgley
M & D Consulting, Leinfelden Metternich, Germany
- M. Miller
European Centre for Medium-Range Weather Forecasts, Reading, UK
- P. Minnis
Science Directorate, NASA Langley Research Center, Hampton, VA, USA
- M. Mishchenko
Goddard Institute for Space Studies, New York, NY, USA
- N.J. Mitchell
The University of Bath, Bath, UK
- M.G. Mlynarczyk
NASA Langley Research Center, Hampton, VA, USA
- C.-H. Moeng
National Center for Atmospheric Research, Boulder, CO, USA
- F. Molteni
Abdus Salam International Centre for Theoretical Physics, Trieste, Italy
- M. Moncrieff
National Center for Atmospheric Research, Boulder, CO, USA
- D.W. Moore
Pacific Marine Environmental Laboratory, Seattle, WA, USA
- G.K. Moortgat
Max-Planck-Institute for Chemistry, Mainz, Germany
- R. Müller
Institute for Energy and Climate Research (IEK-7), Forschungszentrum Jülich, Jülich, Germany
- J. Munthe
IVL Swedish Environmental Research Institute, Göteborg, Sweden
- R.S. Nerem
University of Colorado, Boulder, CO, USA
- P.A. Newman
NASA Goddard, Space Flight Center, Greenbelt, MD, USA
- H. Nguyen
Centre for Australian Weather and Climate Research, Bureau of Meteorology, Melbourne, VIC, Australia
- N. Nicholls
Bureau of Meteorology Research Centre, Melbourne, VIC, Australia
- F.T.M. Nieuwstadt
Delft University of Technology, Delft, The Netherlands
- S. Nigam
University of Maryland, College Park, MD, USA

L. Nkemdirim
University of Calgary, Calgary, AB, Canada

G.R. North
Texas A&M University, College Station, TX, USA

D.R. Novak
Weather Prediction Center, College Park, MD, USA

A. O'Neill
University of Reading, Earley Gate, Reading, UK

J. Oberheide
Clemson University, Clemson, SC, USA

T.E. Osterkamp
University of Alaska, Fairbanks, AK, USA

R.D. Palmer
University of Oklahoma, Oklahoma, OK, USA

T.R. Parish
University of Wyoming, Laramie, WY, USA

D.J. Parker
University of Leeds, Leeds, UK

M.D. Parker
North Carolina State University, Raleigh, NC, USA

G.K. Parks
University of Washington, Seattle, WA, USA

W.R. Pendleton
Utah State University, Logan, UT, USA

J. Perlwitz
University of Colorado, Boulder, CO, USA

R.L. Pfeffer
Florida State University, Tallahassee, FL, USA

R.A. Pielke, Sr.
University of Colorado at Boulder, CO, USA

R. Pinnick
US Army Research Laboratory, Adelphi, MD, USA

J.M.C. Plane
University of Leeds, Leeds, UK

C.M.R. Platt
Colorado State University, Fort Collins, CO, USA

R. Plougonven
Ecole Polytechnique, Palaiseau, France

M.K. Politovich
National Center for Atmospheric Research, Boulder, CO, USA

L.M. Polvani
Columbia University, New York, NY, USA

J.-P. Pommereau
LATMOS, CNRS, Guyancourt, France

J.A. Pyle
University of Cambridge, Cambridge, UK

D.A. Quattrochi
National Space Science and Technology Center, Marshall Space Flight Center, Huntsville, AL, USA

S. Quiring
Texas A&M University, College Station, TX, USA

M. Ramamurthy
University Corporation for Atmospheric Research, Boulder, CO, USA

A.L. Rangno (Retiree)
University of Washington, Seattle, WA, USA

R.M. Rauber
University of Illinois at Urbana-Champaign, Urbana, IL, USA

A.R. Ravishankara
Colorado State University, Fort Collins, CO, USA

I.A. Renfrew
University of East Anglia, Norwich, UK

N.O. Renno
University of Michigan, Ann Arbor, MI, USA

R. Reynolds
University of Reading, Reading, UK

P.B. Rhines
University of Washington, Seattle, WA, USA

A.D. Richmond
National Center for Atmospheric Research, Boulder, CO, USA

D.L. Rickman
National Space Science and Technology Center, Marshall Space Flight Center, Huntsville, AL, USA

R.G. Roble
National Center for Atmospheric Research, Boulder, CO, USA

A. Robock
Rutgers University, New Brunswick, NJ, USA

F. Rohrer
Forschungszentrum Jülich, Jülich, Germany

K.H. Rosenlof
Earth System Research Laboratory, National Oceanic and Atmospheric Administration, Boulder, CO, USA

- M.W. Rotach
University of Innsbruck, Innsbruck, Austria
- R. Rotunno
National Center for Atmospheric Research, Boulder, CO, USA
- I. Roulstone
University of Surrey, Guildford, UK
- J. Rudolph
York University, Toronto, ON, Canada
- L.M. Russell
Scripps Institution of Oceanography, University of California, San Diego, La Jolla, CA, USA
- C.L. Sabine
NOAA Pacific Marine Environmental Laboratory, Seattle, WA, USA
- D.A. Salstein
Atmospheric and Environmental Research, Inc., Lexington, MA, USA
- C. Schär
Atmospheric and Climatic Science ETH, Zürich, Switzerland
- U. Schmidt
Johann Wolfgang Goethe-University, Frankfurt am Main, Institut für Meteorologie und Geophysik, Frankfurt am Main, Germany
- M.R. Schoeberl
Science and Technology Corporation, Lanham, MD, USA
- D.M. (David) Schultz
University of Manchester, Manchester, UK
- J. Schulz
Meteorological Institute, University of Bonn, Bonn, Germany
- J.H. Seinfeld
California Institute of Technology, Pasadena, CA, USA
- C.J. Senff
CIRES, University of Colorado; and NOAA Earth System Research Laboratory, Boulder, CO, USA
- M.C. Serreze
University of Colorado, Boulder, CO, USA
- G.E. Shaw
Geophysical Institute, University of Alaska, Fairbanks, AK, USA
- G.G. Shepherd
York University, Toronto, ON, Canada
- T.G. Shepherd
University of Toronto, Toronto, ON, Canada
- R. Sica
The University of Western Ontario, London, ON, Canada
- H.B. Singh
NASA Ames Research Center, Mountain View, CA, USA
- J.A. Sippel
National Aeronautics and Space Administration (NASA), Greenbelt, MD, USA
- J. Slingo
University of Reading, Reading, UK
- H.G.J. Smit
Research Centre Jülich, Jülich, Germany
- A.K. Smith
National Center for Atmospheric Research, Boulder, CO, USA
- J.B. Smith
Harvard University, Cambridge, MA, USA
- L.A. Smith
London School of Economics, Centre for the Analysis of Time Series, London, UK
- R.B. Smith
Yale University, New Haven, CT, USA
- I.N. Sokolik
Georgia Institute of Technology, Atlanta, GA, USA
- S.C. Solomon
National Center for Atmospheric Research, Boulder, CO, USA
- J. Sommar
Göteborg University, Göteborg, Sweden
- P.J. Sousounis
AIR Worldwide, Boston, MA, USA
- K. Stamnes
Stevens Institute of Technology, Hoboken, NJ, USA
- S.N. Stechmann
University of Wisconsin–Madison, Madison, WI, USA
- E.J. Steig
University of Washington, Seattle, WA, USA
- O. Stein
IEK 8: Troposphere, Research Center Juelich, Juelich, Germany

D.J. Stensrud
*National Oceanic and Atmospheric Administration,
Norman, OK, USA*

R.S. Stolarski
Johns Hopkins University, Baltimore, MD, USA

S.E. Strahan
*Universities Space Research Association, NASA Goddard
Space Flight Center, Greenbelt, MD, USA*

M. Sturm
*US Army Cold Regions Research & Engineering
Laboratory-Alaska, Fort Wainwright, AL, USA*

S.T. Suess
*NASA Marshall Space Flight Center, Huntsville, AL,
USA*

P.P. Sullivan
*National Center for Atmospheric Research, Boulder,
CO, USA*

M.P. Sulzer
Arecibo Observatory, Arecibo, PR, USA

E.S. Takle
Iowa State University, Ames, IA, USA

W.-K. Tao
*NASA/Goddard Space Flight Center, Greenbelt, MD,
USA*

M.J. Taylor
Utah State University, Logan, UT, USA

P.K. Taylor
*Southampton Oceanography Centre, Southampton,
UK*

V.P. Tchakerian
Texas A&M University, College Station, TX, USA

G.E. Thomas
University of Colorado, Boulder, CO, USA

L.W. Thomason
NASA Langley Research Center, Hampton, VA, USA

D.W.J. Thompson
Colorado State University, Fort Collins, CO, USA

D.J. Thomson
Met Office, Bracknell, UK

J. Thuburn
University of Exeter, Exeter, UK

R.A. Tomas
University of Colorado – Boulder, Boulder, CO, USA

D. Toohey
University of Colorado Boulder, Boulder, CO, USA

L. Travis
*Goddard Institute for Space Studies, New York,
NY, USA*

K.E. Trenberth
*National Center for Atmospheric Research, Boulder,
CO, USA*

B.T. Tsurutani
Jet Propulsion Laboratory, Pasadena, CA, USA

J. Turner
British Antarctic Survey, Cambridge, UK

H. Turtiainen
Vaisala Company, Helsinki, Finland

C. Twohy
Oregon State University, Corvallis, OR, USA

T. Uttal
NOAA, Boulder, CO, USA

R.H. Varney
Cornell University, Ithaca, NY, USA

G. Vaughan
University of Manchester, Manchester, UK

W.W. Vaughan
*University of Alabama in Huntsville, Huntsville,
AL, USA*

G.A. Vecchi
GFDL/NOAA, Princeton, NJ, USA

A. Venkatram
*University of California – Riverside, Riverside,
CA, USA*

J.-P. Vernier
*Science Systems and Applications, Inc., Hampton,
VA, USA*

R.K. Vincent
*Bowling Green State University, Bowling Green,
OH, USA*

R. von Glasow
University of East Anglia, Norwich, UK

A. Wahner
Forschungszentrum Jülich, Jülich, Germany

R.M. Wakimoto
*National Center for Atmospheric Research, Boulder,
CO, USA*

D.E. Waliser
California Institute of Technology, Pasadena, CA, USA

J.M. Wallace
University of Washington, Seattle, WA, USA

- B. Wang
University of Hawaii, Honolulu, HI, USA
- Y. Wang
University of Hawaii at Manoa, Honolulu, HI, USA
- Z. Wang
University of Illinois at Urbana-Champaign, Urbana, IL, USA
- B. Ward
Public Works and Natural Resources, Longmont, CO, USA
- S. Warren
University of Washington, Seattle, WA, USA
- J. Waters
California Institute of Technology, Jet Propulsion Laboratory, Pasadena, CA, USA
- R.P. Wayne
University of Oxford, Oxford, UK
- S. Weart
Center for History of Physics, American Institute of Physics, College Park, MD, USA
- P.J. Webster
Georgia Institute of Technology, Atlanta, GA, USA
- P.J. Webster
University of Colorado – Boulder, Boulder, CO, USA
- W.F. Weeks
University of Alaska Fairbanks, Fairbanks, AK, USA
- M.L. Weisman
National Center for Atmospheric Research, Boulder, CO, USA
- S.J. Weiss
National Oceanic and Atmospheric Administration, Norman, OK, USA
- N.C. Wells
University of Southampton, Southampton, UK
- H. Wernli
Institute for Atmospheric and Climate Science, Swiss Federal Institute of Technology (ETH), Zürich, Switzerland
- T. Wetter
Johann Wolfgang Goethe-University, Frankfurt am Main, Institut für Meteorologie und Geophysik, Frankfurt am Main, Germany
- M.C. Wheeler
Centre for Australian Weather and Climate Research, Bureau of Meteorology, Melbourne, VIC, Australia
- A.A. White
University of Surrey, Guildford, UK
- D.N. Whiteman
NASA Goddard Space Flight Center, Greenbelt, MD, USA
- B.A. Wielicki
NASA Langley Research Center, Hampton, VA, USA
- E.R. Williams
Massachusetts Institute of Technology, Cambridge, MA, USA
- M.L. Wong
California Institute of Technology, Pasadena, CA, USA
- N. Wood
Met Office, Exeter, UK
- R. Wood
University of Washington, Seattle, WA, USA
- S.N. Wren
University of Toronto, Toronto, ON, Canada
- C.-C. Wu
National Taiwan University, Taipei, Taiwan
- D.J. Wuebbles
University of Illinois, Urbana, IL, USA
- L. Xie
North Carolina State University, Raleigh, NC, USA
- P. Yang
Texas A&M University, College Station, TX, USA
- S. Yang
NOAA/NWS/NCEP, Climate Prediction Center, Camp Springs, MD, USA
- J.A. Young
University of Wisconsin, Madison, WI, USA
- Z. Yu
College of Hydrology and Water Resources, Hohai University, Nanjing, China; and University of Nevada Las Vegas, Las Vegas, NV, USA
- Y.L. Yung
California Institute of Technology, Pasadena, CA, USA

S.E. Yuter
*North Carolina State University, Raleigh,
NC, USA*

S. Yvon-Lewis
*Texas A&M University, College Station,
TX, USA*

D. Zardi
University of Trento, Trento, Italy

S.E. Zebiak
*International Research Institute for Climate Prediction,
Palisades, NY, USA*

M.D. Zelinka
*Lawrence Livermore National Laboratory, Livermore,
CA, USA*

C. Zhang
University of Miami, Miami, FL, USA

F. Zhang
*Pennsylvania State University, University Park,
PA, USA*

M. Zhang
*Stony Brook University, Stony Brook,
NY, USA*

PREFACE TO THE FIRST EDITION

A half century ago the American Meteorological Society published the *Compendium of Meteorology*, which in a single volume of 1334 pages summarized the state of understanding of the atmosphere at that time. A perusal of the contents of that volume indicates that although a broad range of topics was covered, the vast bulk of the volume was devoted to traditional meteorological topics such as atmospheric dynamics, cloud physics, and weather forecasting. Barely 4 percent of the volume was devoted to articles related to atmospheric chemistry or air pollution and, of course, none of the volume was devoted to techniques such as satellites and remote sensing.

As Sir John Mason aptly notes in his foreword to the present work, the atmospheric sciences have expanded in scope enormously over the past 50 years. Topics such as atmospheric chemistry and global climate change, of only marginal interest 50 years ago, are now central disciplines within the atmospheric sciences. Increasingly, developing areas within the atmospheric sciences require students, teachers, and researchers to familiarize themselves with areas far outside their own specialties. This work is intended to satisfy the need for a convenient and accessible references source covering all aspects of atmospheric sciences. It is written at a level that allows undergraduate science and engineering students to understand the material, while providing active researchers with the latest information in the field.

More than 400 scientists, from academia, government, and industry have contributed to the 330 articles in this work. We are very grateful to these authors for their success in providing concise and authoritative summaries of complex subjects. As editors, we have benefited from the chance to learn from these articles, and we believe that all students and active scientists who want to increase their knowledge of the atmosphere will benefit enormously from access to this work.

We are also grateful to the 31 members of the Editorial Advisory Board who have guided us in our coverage of the very broad range of topics represented in this encyclopedia. Their willingness to suggest topics and authors, and to carefully review draft articles has contributed significantly to our success.

The production of this multivolume encyclopedia would not have been possible without the dedicated work of the staff of the Major Reference Works group at Academic Press. We are especially grateful to the Major Reference Work Development Manager, Colin McNeil, who has worked closely with us during the entire process. Finally, we appreciate the liberal use of color figures in the printed encyclopedia.

James R Holton, Judith A Curry, and John Pyle

This page intentionally left blank

PREFACE TO THE SECOND EDITION

Since the publication of the first edition of the *Encyclopedia of Atmospheric Sciences*, significant advances in research have been achieved all across the broad and expanding spectrum of the field and related disciplines. In particular, climate science with primary input from the atmospheric research emerges as a new field and integrator of interlocking peripheral disciplines over the last decade. These events have demanded the solicitation of new and updated articles for the 2003 edition. Some articles from the earlier publication were judged to be of such a fundamental and enduring nature that they did not require modification. But huge amounts of new information from Earth-orbiting satellite observatories have brought much new insight to the field. In addition there are new findings in many areas such as the latest simulations of meteorological and climatic processes of interest as well as simulations and observations of the composition and interaction of the field's chemical constituents. While interest in the ozone hole and its ramifications may have reached a plateau, ever more understanding of the stratosphere and its role in climate change emerges. The study of past climates provides new means of testing climate models and theories. In weather prediction we see new progress on how data are to be better assimilated for much improved initialization of the forecast model leading to the promise of more accurate predictions of severe weather and tropical cyclones over longer lead times. These are just a few of the new features of the second edition.

The editors of the second edition are greatly indebted to our predecessors in the first edition. They set the outline of topics and solicited the original authors, while establishing a high standard for the content of this publication. In many cases we decided to reprint those articles or request only minor updates. Nevertheless, many articles in this edition are entirely original, based on which we also made significant reorganization of the content. We are proud of our product and hope it provides the same assistance to students, researchers, and practitioners throughout the science and engineering communities.

Editors of the second edition
Gerald R North
Fuqing Zhang
John Pyle

This page intentionally left blank

EDITOR BIOGRAPHIES



Gerald R North received his PhD in theoretical physics from the University of Wisconsin in 1966. After postdoctoral research at the University of Pennsylvania he became a faculty member in physics at the University of Missouri—St. Louis. He shifted his research focus to climate science research during his sabbatical year at the National Center for Atmospheric Research, where he won the Outstanding Paper Award in 1975. He moved to NASA Goddard Space Flight Center (GSFC) in 1978 where he was awarded the NASA Medal for Research Excellence. During his stay at GSFC, he was the proposer and first study scientist for the Tropical Rainfall Measuring Mission, which was launched in 1997 and is still orbiting in 2014.

He moved to Texas A&M University in 1986 as a university distinguished professor of atmospheric sciences where he served as department head from 1995 to 2003. He has served as editor-in-chief of the *Reviews of Geophysics* and is recognized as one of the most cited authors in geosciences (Web of Science). He has chaired and/or served on a number of national committees and is a Fellow of the American Geophysical Union, American Meteorological Society (AMS) and the American Association for the Advancement of Science, and winner of the Jule Charney Award for Research (AMS).

He has published about 150 refereed papers not including many book chapters and reviews. His books include *Paleoclimatology*, co-authored with Thomas Crowley, and *An Introduction to Atmospheric Thermodynamics* co-authored with Tatiana Erikhimova.

North's interests are focused on the use of mathematical and statistical tools to solve climate problems over a wide range of issues including: analytical solutions of simplified energy balance climate models, use of random field techniques in representing and interpreting climate data and model simulations, detection of deterministic signals in climate change, statistical analysis satellite remote sensing for mission planning and analysis of data, paleoclimate problems using simplified climate models.



John Pyle obtained a BSc in Physics at Durham University before moving to Oxford where he completed a DPhil in Atmospheric Physics, helping to develop a numerical model for stratospheric ozone studies. After a short period at the Rutherford Appleton Laboratory he moved to a lectureship at Cambridge University in 1985. In 2000 he was appointed professor of atmospheric science and since 2007 has been the 1920 professor of physical chemistry. He is a Professorial Fellow at St Catharine's College. He has been a codirector of Natural Environment Research Council's National Centre for Atmospheric Science, where he is currently Chief Scientist.

His research focuses on the numerical modelling of atmospheric chemistry. Problems involving the interaction between chemistry and climate have been addressed; these range from stratospheric ozone depletion to the changing tropospheric oxidizing capacity and have included the environmental impact of aviation, land use change, biofuel technologies, and the hydrogen economy. He has studied palaeochemistry problems as well as the projected atmospheric composition changes during the current century. He has published more than 250 peer reviewed papers.

He played a major role in building an EU stratospheric research programme in the 1990s, coordinating several major field campaigns. He has contributed to all the WMO/UNEP assessments on stratospheric ozone since the early 1980s and is now one of the four international cochairns on

the Scientific Assessment Panel, responsible for these assessments. He was a convening lead author in the IPCC Special report "Safeguarding the ozone layer and the global climate system," published in 2006.

He was elected Fellow of the Royal Society in 2004 and an American Geophysical Union Fellow in 2011. He was awarded the Cambridge ScD degree in 2012. Other honours and awards include membership of Academia Europaea (1993), Royal Society of Chemistry (Interdisciplinary award, 1991, and John Jeyes lectureship, 2008), and the Royal Meteorological Society Adrian Gill Prize, in 2004.



Fuqing Zhang is a professor with tenure in the Department of Meteorology at the Pennsylvania State University, with a joint appointment in the Department of Statistics, along with an endowed position as the E Willard & Ruby S Miller Faculty Fellow at the College of Earth and Mineral Sciences at the Pennsylvania State University. His research interests include atmospheric dynamics and predictability, data assimilation, ensemble forecasting, tropical cyclones, gravity waves, mountain plains and sea-breeze circulations, warm-season convection, and regional-scale climate. He earned his BS and MS in meteorology from Nanjing University, China in 1991 and 1994, respectively, and his PhD in atmospheric science in 2000 from North Carolina State University.

He spent seven years as an assistant and then associate professor at Texas A&M University before coming to Penn State University as a full professor in 2008. In 2000, he spent a year and a half as a postdoctoral fellow at the National Center for Atmospheric Research. He also held various visiting scholarship appointments at various academic and research institutions including the National

Center for Atmospheric Research in Boulder, Colorado; the Navy Research Laboratory in Monterey, California; NOAA/AOML Hurricane Research Division, Miami, Florida; Peking University and Nanjing University, China; the Chinese State Key Laboratory of Severe Weather in Beijing, China; and Laboratoire de Meteorologie Dynamique, École Normale Supérieure in Paris, France.

He has authored/co-authored about 130 peer reviewed journal publications and has given more than 160 keynote speeches or invited talks at various institutions and meetings. He has served as principal investigator/co-principal investigator for 30 federal or state-sponsored research grants. He has chaired/cochaired more than 10 scientific meetings or workshops. He also served on various review or advisory panels for numerous organizations that include National Science Foundation, Office of Naval Research, NASA, NOAA, and National Academies. He has also served as editor of several professional journals including *Monthly Weather Review*, *Science China*, *Atmospheric Science Letter*, *Acta Meteorologica Sinica*, and *Computing in Science & Engineering*.

He has also received numerous awards for his research and service. Notably, in 2007 he received the **Outstanding Publication Award** from the National Center for Atmospheric Research. In 2009, was the sole recipient of the American Meteorological Society's 2009 **Clarence Leroy Meisinger Award** "for outstanding contributions to mesoscale dynamics, predictability, and ensemble data assimilation." Most recently, he received the 2014 American Meteorological Society's **Banner Miller Award** "for valuable insights into incorporating real-time airborne Doppler radar measurements via ensemble data assimilation, leading to improvements in forecasts of tropical cyclone track and intensity."

GUIDE TO USING THE ENCYCLOPEDIA

Structure of the Encyclopedia

The material in the encyclopedia is not arranged by ordinary alphabetical order, but by alphabetical order according to 49 principal topic areas taken to allow all papers belonging to each principal topic to appear together in the same volume. Within each principal subject, article headings are also arranged alphabetically, except where logic dictates otherwise. For example, overview articles appear at the beginning of a section.

There are four features that help you find the topic in which you are interested:

- i. the contents list
- ii. cross-references to other relevant articles within each article
- iii. a full subject index
- iv. contributors

i. Contents List

The contents list, which appears at the front of each volume, lists the entries in the order that they appear in the encyclopedia. It includes both the volume number and the page number of each entry.

ii. Cross-references

All of the entries in the encyclopedia have been cross-referenced. The cross-references, which appear at the end of an article as a See also list, serve four different functions:

- i. To draw the reader's attention to related material in other entries

- ii. To indicate material that broadens and extends the scope of the article
- iii. To indicate material that covers a topic in more depth
- iv. To direct readers to other articles by the same author(s)

Example

The following list of cross-references appears at the end of the article.

See also: **Biogeochemical Cycles:** Biogeochemistry of Iodine. **Stratospheric Chemistry Topics:** HO_x; Halogen Sources, Natural (Methyl Bromide and Related Gases); Halogens; Hydrogen Budget; Overview; Reactive Nitrogen (NO_x and NO_y).

iii. Index

The index includes page numbers for quick reference to the information you are looking for. The index entries differentiate between references to a whole article, a part of an article, and a table or figure.

iv. Contributors

At the start of each volume there is list of the authors who contributed to that volume.

This page intentionally left blank

RADIATION TRANSFER IN THE ATMOSPHERE

Contents

Radiation, Solar

Absorption and Thermal Emission

Cloud-Radiative Processes

Non-local Thermodynamic Equilibrium

Scattering

Ultraviolet Radiation

Ultraviolet, Surface

Radiation, Solar

Q Fu, University of Washington, Seattle, WA, USA

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

Almost all known physical and biological cycles in the Earth's system are driven by the electromagnetic radiation emitted by the Sun, known as solar radiation. Solar radiation is also the major cause of climate change that is truly from outside the Earth's system.

Introduction

The Sun as an average star is a typical main-sequence dwarf of spectral class G-2. Its radius is 6.960×10^8 m. The mean distance between the Sun and the Earth is 1.496×10^{11} m and is known as the astronomical unit. Solar radiation is the electromagnetic radiation emitted by the Sun. Almost all known physical and biological cycles in the Earth system are driven by the solar radiation reaching the Earth. Solar radiation is also the cause of climate change that is truly exterior to the Earth system.

Solar Spectrum and Total Solar Irradiance

The distribution of solar radiation as a function of the wavelength is called the solar spectrum, which consists of a continuous emission with some superimposed line structures. The Sun's total radiation output is approximately equivalent to that of a blackbody at 5770 K. The solar radiation in the visible and infrared spectrum fits closely with the blackbody emission at this temperature. However, the ultraviolet (UV) region ($<0.4 \mu\text{m}$) of solar radiation deviates greatly from the visible and infrared regions in terms of the equivalent blackbody temperature of the Sun. In the interval $0.1\text{--}0.4 \mu\text{m}$, the equivalent blackbody temperature of the Sun is generally less than 5770 K with a minimum of about 4500 K at about $0.16 \mu\text{m}$. The deviations seen in the solar spectrum are a result

of emission from the nonisothermal solar atmosphere. Of the radiant energy emitted from the Sun, approximately 50% lies in the infrared region ($>0.7 \mu\text{m}$), about 40% in the visible region ($0.4\text{--}0.7 \mu\text{m}$), and about 10% in the UV region ($<0.4 \mu\text{m}$).

The total solar irradiance (TSI) is the amount of solar radiation received outside the Earth's atmosphere on a surface normal to the incident radiation per unit time and per unit area at the Earth's mean distance from the Sun. The TSI is an important value for the studies of global energy balance and climate. Direct measurements of TSI can be made only from space, and more than 30-year record has been obtained based on overlapping satellite observations. The most accurate value of TSI during the solar minimum period is $1360.8 \pm 0.5 \text{ W m}^{-2}$ according to measurements from the Total Irradiance Monitor on NASA's Solar Radiation and Climate Experiment and a series of new radiometric laboratory tests.

The TSI is not in fact perfectly constant, but varies in relation to the solar activities. Beyond the very slow evolution of the Sun, a well-known solar activity is the sunspots, which are relatively dark regions on the surface of the Sun. When solar activity is high as indicated by the number of sunspots, the TSI increases. The periodic change in the number of sunspots is referred to as the sunspot cycle, and takes about 11 years, the so-called 11-year cycle. The cycle of sunspot maxima having the same magnetic polarity is referred to as the 22-year cycle. The Sun also rotates on its axis once in about 27 days. Satellite observations suggest that TSI values increase by approximately 0.12%

(1.6 W m^{-2}) between solar minimum and maximum. Such change might be too small to directly cause more than barely detectable changes in the tropospheric climate: empirical analysis indicates a global surface temperature increase of $\sim 0.1 \text{ K}$ as the TSI increases from the minimum to maximum of recent solar activity cycles. However, some indirect evidence indicates that the changes of TSI related to sunspot activity may have been significantly larger over the last several centuries. Furthermore, solar variability is much larger (in relative terms) in the UV region, and induces considerable changes in the chemical composition, temperature, and circulation of the stratosphere, as well as in the higher reaches of the upper atmosphere.

Distribution of Solar Insolation at the Top of the Atmosphere

The solar insolation is the actual amount of solar radiation incident upon a unit horizontal surface over a specified time for a given locality. It depends strongly on the solar zenith angle θ_0 and also on the ratio (d/d_m) of the actual distance to the mean distance of the Earth from the Sun. The solar irradiance at the top of the atmosphere may be expressed by

$$F = S(d_m/d)^2 \cos \theta_0 \quad [1]$$

where S is the TSI. The solar zenith angle depends on the latitude, day of year, and time of day, and is given by

$$\cos \theta_0 = \sin \lambda \sin \delta + \cos \lambda \cos \delta \cos h \quad [2]$$

where λ is the latitude, δ the solar declination, and h the hour angle. The hour angle is 0 at solar noon and increases by 15° for every hour away from solar noon. The solar zenith angle is 90° at sunset and sunrise. Then the solar insolation for a specified period of time between t_1 and t_2 is given by

$$Q = \int_{t_1}^{t_2} F(t) dt. \quad [3]$$

The daily insolation at the top of the atmosphere can be determined by integration of eqn [3] over a day. For a given day of the year, the solar declination and the ratio d/d_m can be determined from standard astronomical formulas. Under present astronomical conditions, the solar declination varies from $23^\circ 27'$ on 21 June to $-23^\circ 27'$ on 22 December, while $(d_m/d)^2$ ranges from 1.034 3 on 3 January to 0.967 4 on 5 July. The values of the daily insolation at the top of the atmosphere are presented in Figure 1 as a function of latitude and time of year.

It is clear that the distribution of solar insolation at the top of the atmosphere depends on the Earth's elliptical orbit around

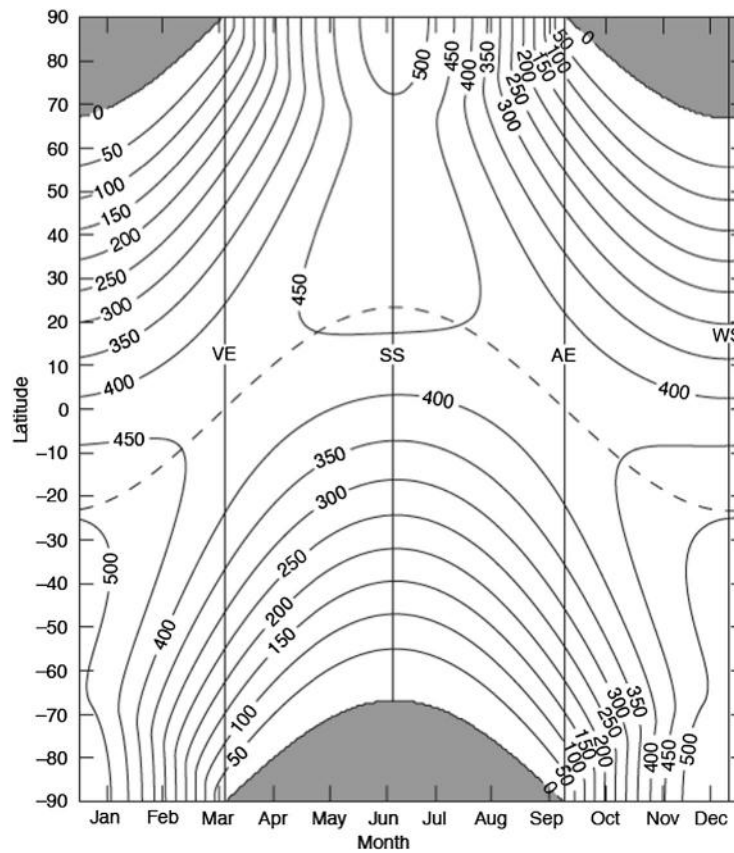


Figure 1 Daily solar insolation at the top of the atmosphere as a function of latitude and day of year. The shaded areas denote zero insolation. The positions of vernal equinox (VE), summer solstice (SS), autumnal equinox (AE), and winter solstice (WS) are indicated with solid vertical lines. Solar declination is shown with a dashed line. Adapted with permission from Liou, K.N., 2002. *An Introduction to Atmospheric Radiation*. Academic Press, San Diego, CA.

the Sun through δ and d/d_m . Because of the gravitational attraction between the Earth and other planets, the orbital parameters including the eccentricity of the orbit, the tilt of the angle, and the longitude of the perihelion vary with characteristic periods of about 100 000, 41 000, and 21 000 years, respectively. The variations of these orbital parameters of the Earth may constitute a cause for climate changes as those experienced during the Pleistocene ice ages.

Scattering and Absorption of Solar Radiation in the Earth–Atmosphere System

Solar radiation entering the Earth's atmosphere is absorbed and scattered by atmospheric gases, aerosols, clouds, and the Earth's surface. The absorbed radiation is added directly to the heat budget; whereas the scattered radiation is partly returned to space and partly continues its path through the Earth–atmosphere system where it is subject to further scattering and absorption. The fraction of the incident solar radiation that is reflected and backscattered to space is called the albedo. We might speak of the albedo of the entire Earth or of individual surfaces with reference either to monochromatic radiation or to the total incident solar radiation. In this last sense the albedo of the Earth as a whole is about 0.30.

Effects of Atmospheric Gases on Solar Radiation

The scattering of solar radiation by air molecules can be described by a theory developed by Rayleigh who showed that the amount of scattering is inversely proportional to the fourth power of the wavelength, when the sizes of particles are much smaller than the wavelength of the incident radiation. We see blue sky because atmospheric molecules scatter solar radiation much more in the blue than in the red part of the spectrum. In fact, the sky is made visible through the scattering process. On the other hand, sunsets and sunrises appear reddish because the blue light in the direct light is removed by scattering during the long path through the atmosphere, leaving the remaining reddish colors of the spectrum.

Atmospheric gases also absorb solar radiation in selected wavelength bands. The UV radiation with wavelengths shorter than $0.3 \mu\text{m}$ is lethal to the biosphere. The UV radiation in the interval $0.2\text{--}0.3 \mu\text{m}$ is mainly absorbed by O_3 in the stratosphere. The small amount of radiation with wavelengths shorter than $0.2 \mu\text{m}$ is absorbed at higher levels by O_2 , N_2 , O , and N . The photochemical processes due to absorption of solar UV radiation involving various forms of oxygen are critical in determining the amount of ozone in the stratosphere. The absorption spectrum of O_2 between 0.2 and $0.26 \mu\text{m}$ is weak but of significance in the formation of ozone.

In the troposphere, the absorption of solar radiation occurs in the visible and near infrared regions, owing primarily to H_2O , CO_2 , O_2 , and O_3 . The absorption in the visible, however, is very weak. Figure 2 shows the depletion of solar radiation in a clear atmosphere. The top curve is the solar spectrum at the top of the Earth's atmosphere and the lower curve represents the spectrum at the sea level; the shaded area gives the combined effects of scattering and absorption of solar radiation

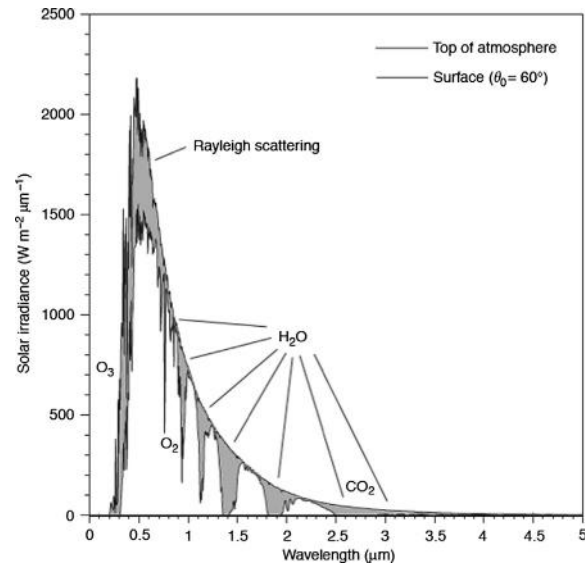


Figure 2 Solar spectrum at the top of the atmosphere and at the surface for a solar zenith angle of 60° in a clear atmosphere. Absorption and scattering regions are indicated. Adapted with permission from Liou, K.N., 2002. *An Introduction to Atmospheric Radiation*. Academic Press, San Diego, CA.

by atmospheric gases. It is evident that the depletion of solar radiation is dominated by ozone absorption in the UV, Rayleigh scattering in both UV and visible, and water vapor absorption in the near infrared regions.

Effects of Aerosols on Solar Radiation

Aerosols are suspensions of liquid and solid particles in the atmosphere, excluding clouds and precipitation. The aerosol particle sizes range from 10^{-4} to $10 \mu\text{m}$, falling under the following broad categories: sulfates, black carbon, organic carbon, dust, and sea salt. Aerosol concentrations and compositions vary significantly with time and location. Visibility measurements reflect the aerosol concentration at ground level. The visual range can vary from a few meters to 200 km, depending on the proximity to sources, the strength of the sources, and atmospheric conditions.

Aerosols scatter and absorb solar radiation. Sulfate aerosols scatter primarily solar radiation and cause cooling of the Earth–atmosphere system. The increase in the reflected solar radiation at the top of the atmosphere due to such nonabsorbing aerosols is nearly identical to the reduction in the solar radiation at the surface. Carbonaceous aerosols (black carbon and organics) absorb and scatter solar radiation. The presence of black carbon aerosols results in the absorption of solar radiation, which reduces the solar radiation reaching the surface. At the same time, these aerosols absorb the upward solar radiation reflected from below and reduce the solar radiation reflected to space. Therefore, the effect of black carbon aerosols opposes the cooling effect of other aerosols at the top of the atmosphere, whereas at the surface all aerosols reduce the solar radiation. The changes arising from the aerosol scattering

and absorption of solar radiation are referred to as their direct radiative forcing. Aerosols can also modify the solar radiation through their role in cloud condensation and as ice nuclei, an effect known as aerosol indirect radiative forcing.

Aerosol particles in the atmosphere are produced both in nature and by people. A global aerosol optical depth of about 0.12 is suggested. These aerosols increase the reflected solar radiation at the top of the atmosphere by about 3 W m^{-2} globally. Anthropogenic sources contribute significantly to the global aerosol optical depth. Global anthropogenic emissions of sulfates, organics, and black carbon even exceed natural sources. Such a large perturbation of the global aerosol loading due to human activities may significantly modify regional and global climates.

Effects of Clouds on Solar Radiation

Clouds regularly cover about 65% of the Earth, and occur in various types. Some, such as cirrus in the tropics and stratus near the coastal areas and in the Arctic are climatologically persistent. Like aerosols, clouds show substantial spatial and temporal variations.

Clouds are the most important regulator of solar radiation. By reflecting incoming solar radiation back to space, they cool the Earth–atmosphere system – the so-called cloud albedo effect. Clouds also absorb solar radiation in the near infrared region. The cooling of the Earth–atmosphere system due to the cloud albedo effect occurs primarily at the surface. The solar albedo of clouds depends substantially on cloud type and cloud form, as well as the solar zenith angle.

The most straightforward and simple diagnostic measure of the impact of clouds on solar radiation is the short-wave cloud forcing which is defined as the difference of the net solar irradiances at the top of the atmosphere between all-sky and cloudless conditions. Here the net irradiance is the incoming solar radiation minus the reflected radiation. Satellite measurements suggest that the global short-wave cloud forcing is about -45 W m^{-2} . Short-wave cloud forcings are maximized (about -120 W m^{-2}) in the summer hemisphere at about latitude 60° where solar input is large and low clouds are abundant, with a secondary maximum in tropics. Note that the magnitude of short-wave cloud forcing is about 10 times as large as those for a CO_2 doubling. Hence small changes in the cloud-radiative forcing fields can play a significant role as a climate feedback mechanism.

Solar Radiation at the Earth's Surface

The radiation coming directly from the Sun received at the Earth's surface is called direct solar radiation. The amount of scattered radiation coming from all other directions is called diffuse solar radiation. The sum of both components as received on a horizontal surface is called global solar radiation.

A significant fraction of the incoming solar radiation is reflected back by the surface. The surface albedo, defined as the ratio of the reflected over the incoming radiation, depends on the nature of the surface, solar zenith angle, and wavelength. For a water surface the albedo is about 0.06, whereas for snow the albedo is about 0.6–0.8. The albedo of bare sea ice is about 0.4. Since large areas of Earth are covered by water, snow, and sea ice, changes in the snow and sea ice cover can have a significant impact on the global albedo.

Bare land surfaces have typical surface albedo of 0.1–0.35, with the highest value for the desert sand. Albedos of most vegetation surfaces fall in the range of 0.1–0.25. The albedo for green vegetation depends greatly on wavelengths, reflecting strongly in the near infrared but absorb in the ultraviolet and visible regions.

Annual Global Mean Energy Budget of Solar Radiation

The energy budget of solar radiation can be derived by combining observations and modeling studies, which shows combined effects of atmospheric gases, aerosols, clouds, and surfaces. Under the annual global mean condition, the incident solar radiation at the top of the atmosphere is 340 W m^{-2} . Of this incident solar radiation, 75 W m^{-2} is absorbed during passage through the atmosphere. A total of 100 W m^{-2} is reflected back to space: 23 W m^{-2} from the surface and 77 W m^{-2} from clouds and aerosols and atmosphere. The remaining 165 W m^{-2} are absorbed at the Earth's surface. It is noted that while the incoming and reflected solar irradiances at the top of the atmosphere are constrained by satellite observations, uncertainties may exist for the partitioning of the absorbed solar radiation between the atmosphere and the surface on the global scale.

See also: Radiation Transfer in the Atmosphere: Absorption and Thermal Emission; Non-Local Thermodynamic Equilibrium; Scattering. Satellites and Satellite Remote Sensing: Earth's Radiation Budget.

Further Reading

- Goody, R.M., Yung, Y.L., 1989. *Atmospheric Radiation*. Oxford Press, New York.
- Kopp, G., Lean, J.L., 2011. A new, lower value of total solar irradiance: evidence and climate significant. *Geophysical Research Letters* 38, L01706. <http://dx.doi.org/10.1029/2010GL045777>.
- Liou, K.N., 2002. *An Introduction to Atmospheric Radiation*. Academic Press, San Diego.
- Peixoto, J.P., Oort, A.H., 1992. *Physics of Climate*. Springer-Verlag, New York.
- Stephens, G.L., Li, J.L., Wild, M., Clayson, C.A., Loeb, N., Kato, S., L'Ecuyer, T., Stackhouse, P.W., Lebsock, M., Andrews, T., 2012. An update on Earth's energy balance in light of the latest global observations. *Nature Geoscience* 5, 691–696.
- Thomas, G.E., Stamnes, K., 1999. *Radiative Transfer in the Atmosphere and Ocean*. Cambridge Press, Cambridge.
- Wallace, J.M., Hobbs, P.V., 2006. *Atmospheric Science: An Introductory Survey*. Academic Press, New York.

Absorption and Thermal Emission

RM Goody, Harvard University (Emeritus), Cambridge, MA, USA
X Huang, University of Michigan, Ann Arbor, MI, USA

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

This article covers the physical processes responsible for solar and thermal radiation as well as the physical laws governing such processes, properties of gaseous absorptions, common algorithms used to solve the radiative transfer in the absence of scattering, radiative properties of the Earth's atmosphere, and general thermodynamics in the context of the radiation field.

Solar and Thermal Radiation

In the final analysis, all physical or biological processes that take place on Earth owe their existence to the absorption of radiation from the Sun. In the absence of this radiation, the Earth would be cold and lifeless, at a temperature close to that of the cosmos, 2.73 K.

Solar radiation corresponds approximately to a blackbody emission temperature of 5780 K. Solar radiation reaching the Earth is, in part, scattered back to space. The remainder is absorbed and heats the Earth, which, in turn, emits thermal radiation to space. A long-term balance between absorption and emission leads to an average terrestrial emission temperature of approximately 255 K (substantially less than the observed surface temperature).

Internally to the atmosphere, thermal radiation is emitted and absorbed, transporting energy from warmer levels to colder levels. Consequently, the radiation field is characterized by the radiance, $I_{\nu,\lambda}(l)$, defined as the rate of flow of energy in the l -direction, for a unit area perpendicular to l , for a unit solid angle, and for a unit range of frequency or wavelength (ν or λ). The units are $\text{W m}^{-2} \text{sr}^{-1} (\nu \text{ or } \lambda)^{-1}$, where sr stands for steradian.

The Interaction Process

The interaction between a photon and an air molecule can be described in varying degrees of complexity, but a simplified picture of a one-stage process conveys the general nature of absorption and scattering (see Figure 1).

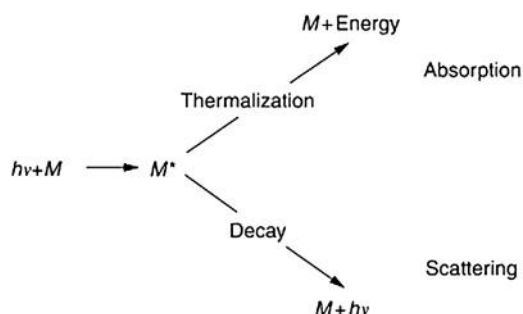


Figure 1 A molecule–photon interaction. The branching depends on the ratio of the collisional thermalization rate to the spontaneous decay rate.

The molecule is assumed to exist in one of two quantized states, M and M^* . M represents the state before interaction with the photon ($h\nu$). We will assume that this initial state is in thermodynamic equilibrium (i.e., the populations of energy levels are determined by Boltzmann's law). The first stage in the interaction is for the photon to disappear and the molecule to be raised to the excited state, M^* (which will not be in Boltzmann equilibrium).

In the second stage of the interaction, several things may happen to the excited molecule, M^* , including decomposition and chemical reactions. Leaving chemistry aside, there are two generic possibilities. First, collisions with other molecules may be so rapid that the excess energy in M^* is thermalized and the molecule returns to an equilibrium state, but with slightly more internal energy in the thermal reservoir. This process is absorption.

The alternative process is for the excited molecule to decay spontaneously, as all excited molecules will if left undisturbed for long enough. This again returns the molecule to an equilibrium state, but now a photon is emitted that is the same as the incident photon, except with different direction and different polarization. The internal energy of the thermal reservoir is unchanged. This is a coherent scattering process. The branching between absorption and scattering depends upon the ratio of the times for collisional thermalization and spontaneous decay. For all bands that will be discussed in the section Gaseous Absorption, this ratio is very small in the troposphere and stratosphere, and absorption is the dominant mode of interaction between molecules and radiation in the lower atmosphere.

The absorption coefficient per molecule, $k_{\nu,\lambda}$, is defined with respect to those photons that disappear in the interaction process (i.e., are absorbed). It is the cross-section for an absorbing collision between a photon and a molecule, and its units are square meters (m^2). Experimentally, it is the fractional change in radiance for radiation traversing an absorption tube in which it interacts with one molecule per unit area of tube cross-section (eqn [1]).

$$dI_{\nu,\lambda} = -nk_{\nu,\lambda}I_{\nu,\lambda}ds = +I_{\nu,\lambda}d\tau_{\nu,\lambda} \quad [1]$$

Here, n is the molecular number density, ds is the infinitesimal absorbing path, and $\tau_{\nu,\lambda}(= \int_{\text{path}} nk_{\nu,\lambda}ds)$ is the optical path. Note that, by convention, the direction of the path is always defined in such a way that $dI_{\nu,\lambda} = -nk_{\nu,\lambda}I_{\nu,\lambda}ds$.

Although the internal energy must be described differently, an analogous situation and analogous definitions exist when a single molecule is replaced by a single aerosol particle.

Again, both scattering and absorption can take place, although the branching is not controlled by the molecular collision rate. In this article, discussion will be limited to gaseous absorption (particle and droplet scattering are treated elsewhere; *see Aerosols: Role in Radiative Transfer*; and *Radiation Transfer in the Atmosphere: Scattering*).

Thermal Emission

Thermal emission is the complementary process to absorption: internal energy is transferred spontaneously from the thermal reservoir to photons in the radiation field. In Boltzmann equilibrium, states above the ground state have finite populations and there is a probability that decay by spontaneous emission will take place, emitting a photon. As a consequence, there is a steady state of photons associated with the interior of a constant-temperature enclosure or cavity. Cavity or black-body radiation is, according to Kirchhoff's laws, isotropic and the same as that emitted by a perfectly black body at the same temperature; the radiance is designated $B_\lambda(T)$, where T is the absolute temperature.

All meteorological and climatological studies assume that air molecules are in a state of local thermodynamic equilibrium, whereby Boltzmann equilibrium is obeyed on a local basis. Consequently, equilibrium radiation laws can be employed. But this may not be correct at low pressures, when collisions are infrequent. In the upper atmosphere, a nonequilibrium approach to radiation and fluid mechanics is necessary.

$B_\lambda(T)$ is governed by Kirchhoff's and Planck's laws. It can be expressed as a function of its maximum value (eqn [2]) and the dimensionless number $x = hc/k_B\lambda T$, where h , k_B , and c are, respectively, Planck's constant, Boltzmann's constant, and the speed of light.

$$\frac{B_\lambda}{B_\lambda(\max)} = 4.717 \times 10^{-2} \frac{x^5}{e^x - 1} \quad [2]$$

$B_\lambda(\max)$ is given by eqn [3]:

$$B_\lambda(\max) = 4.1069 \times 10^{-6} T^5 \text{ W m}^{-3} \text{ sr}^{-1} \quad [3]$$

The wavelength of the maximum emission is given by Wien's displacement law (eqn [4]).

$$\lambda(\max) = 2.8978 \times 10^{-3} T^{-1} \text{ m} \quad [4]$$

For $T = 5780 \text{ K}$, the maximum emission is at $0.501 \mu\text{m}$, in the center of the visible spectrum; for $T = 255 \text{ K}$, the maximum is at $11.36 \mu\text{m}$, in the middle infrared spectrum.

The integral of the radiance over all wavelengths is given by the Stefan-Boltzmann relation (eqn [5]), where σ is the Stefan-Boltzmann constant, equal to $5.670 \times 10^{-8} \text{ W m}^{-2} \text{ K}^{-4}$.

$$\int_0^\infty B_\lambda(T) d\lambda = B(T) = \frac{\sigma}{\pi} T^4 \quad [5]$$

From Kirchhoff's laws, the radiance emitted from an infinitesimal path ds in thermodynamic equilibrium is given by eqn [6].

$$dI_\lambda = +nk_\lambda B_\lambda(T) ds = -B_\lambda(T) d\tau_\lambda \quad [6]$$

Gaseous Absorption

Bands and Line Spectra

The Earth's atmosphere has a rich absorption spectrum of hundreds of thousands of lines and continuum features extending from the ultraviolet to the microwave spectra, belonging to many different atmospheric gases. Most of the absorptions are vibration-rotation bands, for which the band center is determined by a vibrational transition, with simultaneous rotational transitions forming branches of discrete lines on either side of the center. In the absence of a vibrational transition, lines forming a pure rotation band can exist in the far infrared and microwave spectra. In one case (the oxygen red and infrared bands), an electronic transition is also involved in the infrared spectrum, but usually electronic transitions are to be found in the visible and ultraviolet spectra, and are continua rather than line spectra.

Line and band strengths (S_{line} and S_{band} , which are integrals of the absorption coefficient over a line or band) vary over an enormous range, depending upon the nature of the molecule and the transition. They are proportional to the state population of the lower state involved, which is governed by Boltzmann's law and can be highly temperature dependent.

Figure 2 shows two typical spectral regions at high resolution. Figure 2(a) shows the regular structure of a simple linear molecule. Figure 2(b) shows the more complex structure exhibited by a molecule with three different moments of inertia. Both have been calculated from the HITRAN database (<http://www.cfa.harvard.edu/hitrان/>). HITRAN is a compilation of spectroscopic parameters with relevance to the emission and transmission of light within the atmosphere. The latest version of HITRAN, HITRAN2008, contains information on 2 713 968 spectral lines for 39 different molecules and their associated isotopologues, ranging from the ultraviolet to the microwave spectrum. The data are freely available upon request via the HITRAN website. GEISA (Gestion et Etude des Informations Spectroscopiques Atmospheriques) is a similar database that is more frequently used in Europe; it is also freely available online, via the Center for Atmospheric Chemistry Products and Services hosted at the Institut Pierre Simon Laplace, France (<http://ether.iplsl.jussieu.fr>). The latest version of GEISA, GEISA 2009, lists 3 794 488 lines for 50 molecules (111 isotopic species) in total.

HITRAN data for each line include molecule, line strength, line frequency, molecular parameters for the transition, lower state energy (for Boltzmann calculations), linewidths for air- and self-broadening collisions, temperature dependence of the linewidth, references, and error codes. In general, most lineshape information and data on continua must be supplied from other sources. Figure 3 shows a low spectral resolution composite (i.e., the lines are unresolved) of the atmospheric transmission for bands of the most important atmospheric gases.

Line Profiles

The line profile is the absorption coefficient divided by the line strength (eqn [7]), where ν_0 is the frequency of the line center.

$$f(\nu - \nu_0) = \frac{k_\nu}{S_{\text{line}}} \quad [7]$$

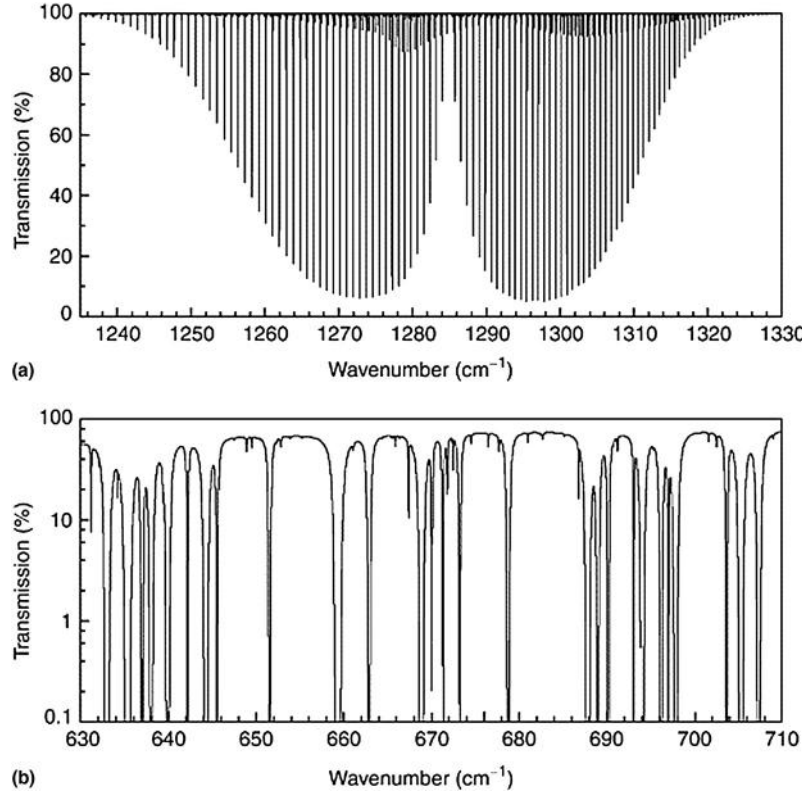


Figure 2 High-resolution spectra. The spectra are calculated from a line-by-line model based on GENLN (general purpose line-by-line atmospheric transmittance and radiance model), using the HITRAN (high-resolution transmission) database. (a) The fundamental vibration–rotation band of N_2O at 1285 cm^{-1} ($1 \text{ cm}^{-1} = 10^4/\lambda(\mu\text{m})$). The absorption path is the atmosphere above 21 km for a zenith angle of 30° . (b) The H_2O rotation band near 670 cm^{-1} . The absorption path is for the entire atmosphere at a zenith angle of 30° . The concentration of water vapor is 3% of a typical terrestrial concentration.

Since lines are formed during close collisions, the theory is extremely complicated, and use is commonly made of empirical corrections to two simple approximations: the Michelson–Lorentz profile and the Doppler profile.

The Michelson–Lorentz profile is an approximation to the line profile of a stationary molecule bombarded by others, with a mean time, t , between collisions. The profile is given by eqn [8], where A is a small line shift parameter and α_L is the Michelson–Lorentz linewidth, given by eqn [9].

$$f_L(\nu - \nu_0) = \frac{\alpha_L}{\pi[(\nu - \nu_0 - A\alpha_L)^2 + \alpha_L^2]} \quad [8]$$

$$\alpha_L = \frac{1}{2\pi t} \quad [9]$$

According to eqn [9], the linewidth is proportional principally to the air density, and therefore to the pressure, hence the term ‘pressure broadening.’

The Doppler profile is appropriate to a collision-free, thermally agitated molecule. The Doppler profile is given by eqn [10], with α_D given by eqn [11] where α_D is the Doppler linewidth and m is the mass of the absorbing molecule.

$$f_D(\nu - \nu_0) = \frac{1}{\pi^{1/2}\alpha_D} \exp\left[-\left(\frac{\nu - \nu_0}{\alpha_D}\right)^2\right] \quad [10]$$

$$\alpha_D = \frac{v_0}{c} \left(\frac{2k_B T}{m} \right) \quad [11]$$

For a water vapor line at 200 cm^{-1} and a temperature of 300 K , the Doppler width is $3.4 \times 10^{-4} \text{ cm}^{-1}$. A typical Michelson–Lorentz width with air as a perturber is 0.05 cm^{-1} at 10^5 Pa . The two widths are approximately equal for a pressure of $\sim 10^3 \text{ Pa}$, which occurs near 30 km altitude in the atmosphere. Below 30 km , Doppler broadening may be neglected; above 30 km , it may be important. A convolution between the two shapes, known as the Voigt profile, is frequently used for taking both broadening mechanisms into account.

Note that the Michelson–Lorentz profile is derived assuming the simplest treatment of collisional broadening. More detailed treatments give complicated profiles, such as the Ben–Reuven (which considers the quantum mechanical treatment of collisional broadening), Van Vleck–Weisskopf, and kinetic lineshapes, the latter two of which are special cases of the Ben–Reuven profiles. The kinetic lineshape is sometimes also known as the Zhevakin–Naumov or Gross lineshape. The discrepancies between such complicated lineshapes and the Michelson–Lorentz profile can be significant in the far wings of absorption lines. Theory and laboratory measurements indicate that the Michelson–Lorentz profile is quite precise for displacements from line centers less than $\sim 2 \text{ cm}^{-1}$. At greater

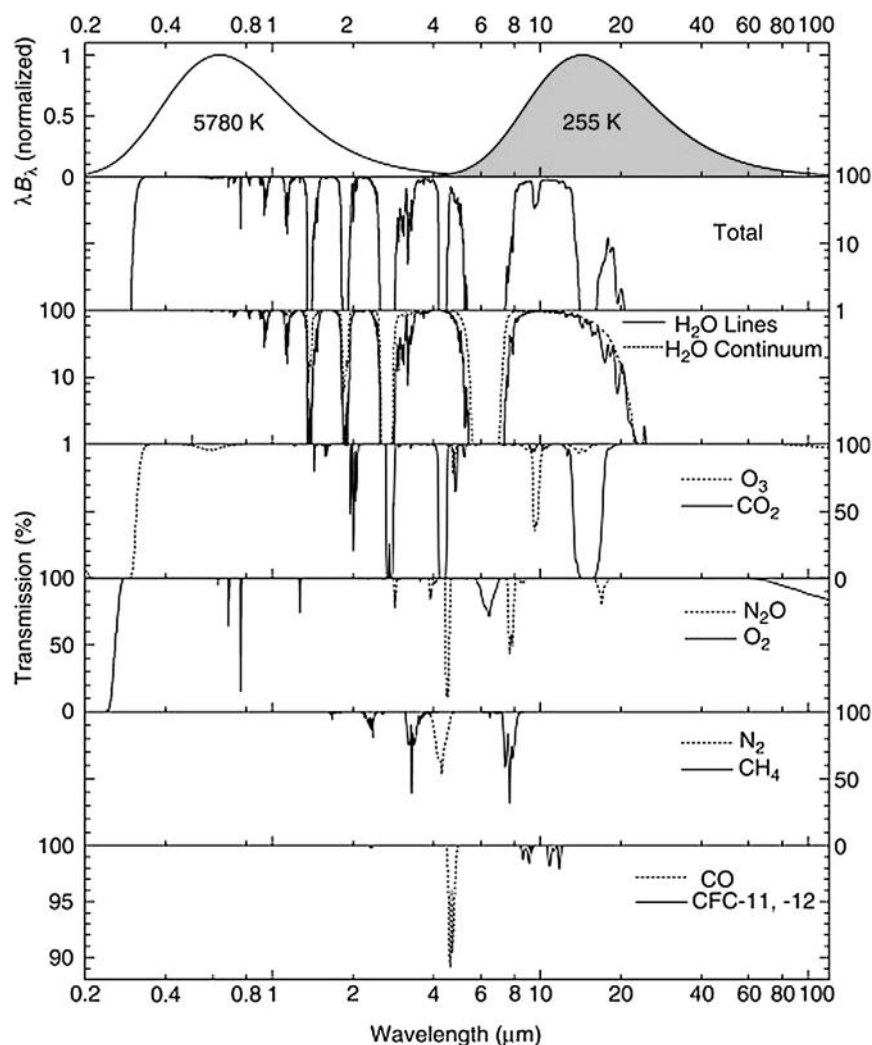


Figure 3 Vertical transmission of the entire atmosphere for the most important atmospheric absorbers; spectral resolution 10 cm^{-1} . The first panel shows blackbody radiances corresponding to solar and terrestrial temperatures. The scales are such that areas correspond to energies. The two blackbody curves have been normalized to have equal areas. The transmission scales differ from panel to panel and alternate on left and right sides. The third panel shows the water vapor line and continuum transmissions as distinct species.

displacements, important deviations have been measured for water vapor and carbon dioxide lines. These empirical departures from the Michelson–Lorentz profile must be incorporated into radiation algorithms by the user. Given complete line profiles and the data in HITRAN or GEISA, it is possible to calculate the monochromatic absorption coefficient due to lines at any frequency in the terrestrial spectrum.

Continua

The visible and ultraviolet absorptions of O_3 and O_2 are the results of electronic transitions to unquantized upper states, and they show no line structure.

Between 8 and $13\text{ }\mu\text{m}$ is a translucent region in the water vapor spectrum in which the most important absorption is a continuum formed by far wings of very strong lines in the adjacent bands. For one numerical program, the water continuum is formally defined to be the residual absorption of

wings more than 25 cm^{-1} from line centers. Water vapor lines and continua, defined in this way, are shown as two distinct spectra in Figure 3. The $10\text{ }\mu\text{m}$ continuum is of great importance for climate studies because it controls a spectral region where energy can readily escape from the lower atmosphere to space.

Nitrogen, oxygen, and carbon dioxide molecules are all symmetric, and some or all vibration–rotation transitions are strictly forbidden. In close collisions, however, these symmetries are broken and transitions can take place for a very short time. Transitions for which this is of interest are the fundamental vibrational transitions of nitrogen and oxygen and the pure rotational transitions of oxygen, nitrogen, and carbon dioxide. Because a very short time is spent during collisions, the lines are very wide, and they overlap to form a featureless continuum.

Polymers may be formed during close collisions. Polymer bands have been detected in the visible and near-infrared spectra of oxygen. Again, absorption is continuous because of the short lifetime of the polymer.

Continua behave differently from line spectra. Line absorptions saturate as the path length increases, and for long paths the dependence upon path length becomes very slow. Continua do not saturate as readily as line spectra, and their relative importance increases at long absorption paths.

Radiance Algorithms

Equations [1] and [6] may be combined to give the net change of radiance for simultaneous absorption and thermal emission. The result is Schwarzschild's equation of transfer, eqn [12]:

$$\frac{dI_\nu}{d\tau_\nu} = I_\nu - B_\nu(T) \quad [12]$$

Equation [12] is incomplete. In general, the blackbody radiance should be replaced by a multiple-scattering and absorption source function.

If we integrate eqn [12] along a path in the l-direction, with the optical path (τ) varying from $\tau = 0$ at the place where the radiance is calculated, to a boundary at $\tau = \tau'$, where the radiance is given as a boundary condition, we obtain eqn [13]:

$$I_\nu(l, \tau = 0) = I_\nu(l, \tau = \tau') \exp(-\tau') + \int_0^{\tau'} B_\nu(\tau) \exp(-\tau) d\tau \quad [13]$$

The boundary is either space (for a downward-directed radiance) or the Earth's surface (for an upward-directed radiance). Equation [13] is valid for both thermal and solar radiation, with the difference that, in the solar spectrum (0.15–5 μm), the thermal source function is usually neglected, and the second term on the right of eqn [13] omitted. The source of solar radiation in the atmosphere (in the absence of scattering) is the first (boundary) term on the right of eqn [13].

From the solution of eqn [13], it is possible to construct the radiation flux by integrating over all directions, and its divergence (the negative heating rate, also frequently denoted as the cooling rate). The radiative heating rate is the fundamental drive for the atmosphere, and its calculation with adequate precision is an important task for radiation studies.

Equation [13] is a simple quadrature. Nevertheless, it is a major numerical task to perform it efficiently, flexibly with respect to changing input data, and accurately. In the United States, two tested programs are available: LBLRTM, originally developed at the US Air Force Research Laboratory and now at the Atmospheric and Environmental Research Radiative Transfer Working Group (<http://rtweb.aer.com/lblrtm.html>); and GENLN, developed at Oxford University and the National Center for Atmospheric Research (e-mail contact: edwards@ucar.edu). Both programs are monochromatic (so-called line-by-line codes). They are based upon the HITRAN database, adding those data on lineshapes and continua that are needed to complete the calculation. No physical approximations are employed. Both programs (and also Moderate-resolution Atmospheric Transmission (MODTRAN), discussed in this article) can be combined with scattering options.

Although modern computers can step rapidly through the thermal or solar spectrum with frequency steps smaller than the

narrowest lines, line-by-line calculations are too slow for weather or climate models and applications in remote sensing, especially hyperspectral remote sensing. These usually employ band model calculations, in which eqn [13] is integrated with respect to frequency over many lines, sometimes over entire vibration-rotation bands. The exponential functions are then replaced by more complex functions based partly on laboratory measurements and partly on theoretical analysis. Such methods are of limited accuracy, although, in recent years, the main sources of error have been overcome by means of different fast parameterization techniques. One group of techniques predicts the transmittances or radiances by using a limited number of monochromatic transmittances or radiances. The most well-known of these is the correlated- k distribution technique. The exponential-sum fitting transmittance, spectral-mapping, and optimal spectral-sampling techniques also belong to the same group. One advantage of using such techniques is the full compatibility with scattering options. Another group of techniques employs efficient schemes to predict the layers' optical depth (with or without convolution by the instrument response function) and then computes transmittances or radiances accordingly. Also, a faster model has been developed based on the principal component analysis, which takes advantage of the orthogonal properties of the principle components to compress the spectral information, and then the model predicts only the scores of principle components instead of radiances or transmittances at each frequency channel. These techniques are critical for better use of current hyperspectral observations taken from space, which normally contain radiances measured at thousands of channels.

MODTRAN is a band model program with a maximum resolution of 0.1 cm^{-1} . The latest version, MODTRAN5, was developed by the US Air Force Research Laboratory and Spectral Sciences Inc. (<http://modtran.org>), and also has options for both the traditional random-band model and correlated- k distribution technique. It is commonly used when speed is a consideration.

Radiation in the Earth's Atmosphere

Interaction with the Sun and with Space

Absorption of solar radiation in the atmosphere and emission of the atmosphere to space are concentrated in layers (Chapman layers). The altitude of the maximum of a Chapman layer for a given wavelength depends on the absorption coefficient (see Figure 4).

Boundary Fluxes

Figure 5 shows net solar and thermal fluxes for 100 km and for the Earth's surface, in midlatitudes. The differences between these two fluxes represent the total radiant energy gained or lost by the atmosphere. The calculation omits all scattering; consequently (and unrealistically), the solar flux is for the direct solar beam only. The global lower atmosphere, as a whole, loses the radiant energy, which is, in general, balanced by the latent heat release from the condensations of water vapor.

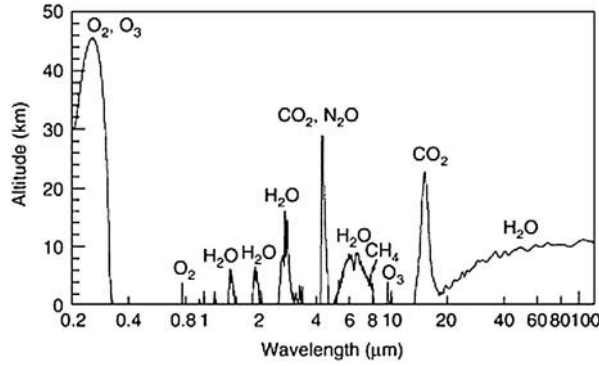


Figure 4 Altitudes of solar absorption and thermal emission. Heights of Chapman layers and primary absorbers for vertical radiances.

Heating Rates

Figure 6 shows the breakdown of the energy gained and lost to the atmosphere from **Figure 5**, with respect to altitude and principal species. Imbalances are the drivers for dynamical processes.

Radiative Equilibrium

The thermal terms in **Figure 6** are more sensitive to air temperature than are the solar terms. Consequently, the atmospheric temperature may be adjusted until the solar and thermal terms in **Figure 6** balance. This state (radiative equilibrium; see **Figure 7**) can exist only in an idealized, motionless atmosphere. Nevertheless, it offers an important heuristic basis for studying the state of the climate system (CS).

Thermodynamics of the Radiation Field

The thermodynamics of the radiation field is characterized by two intensive variables, temperature and pressure, and two extensive variables, energy and entropy; they are not independent, and all can be expressed in terms of the energy and the frequency.

Temperature

The brightness temperature is obtained by inverting eqn [2] with the ambient radiance in place of B_ν , and it is the temperature with which the radiation stream is in equilibrium.

Pressure

The radiation pressure is equal to one-third of the energy density. It can cause work to be done during an interaction with matter in motion, but, for atmospheric velocities, this work is negligible compared to changes of energy. Consequently, the heat interaction experienced by the atmosphere is the loss of the internal energy of the radiation (i.e., the negative divergence of the radiative energy flux).

Entropy

The entropy radiance (L_ν , defined in the same manner as the energy radiance) is a single-valued function of the energy radiance. For cavity radiation, the entropy radiance can be integrated over all frequencies to give eqn [14].

$$L = \int_0^\infty L_\nu d\nu = \frac{4}{3} \frac{\sigma}{\pi} T^3 \quad [14]$$

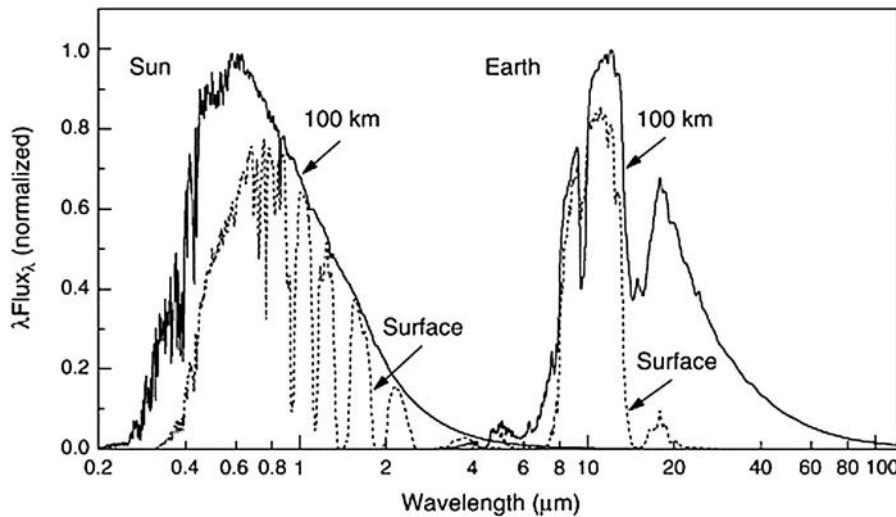


Figure 5 Upward thermal and downward solar net fluxes at 100 km and at the Earth's surface. Vertical fluxes have been calculated from MODTRAN. The conditions approximate global averages. The solar fluxes are for the direct solar beam only (zenith angle, 75.8°). All scattering and reflection terms are omitted. If scattering and reflection were included, the solar fluxes would be scaled back. The solar and thermal fluxes are plotted in such a manner that areas are proportional to energies, and scaled so that their maxima are equal. In absolute units, the areas of these curves are as follows: Sun, 100 km, 339.0 W m^{-2} ; Sun surface, 187.5 W m^{-2} ; Earth, 100 km, 253.9 W m^{-2} ; and Earth surface, 103.9 W m^{-2} . The differences between the 100 km and the surface fluxes represent the total energy gained (solar) or lost (thermal) by the atmosphere.

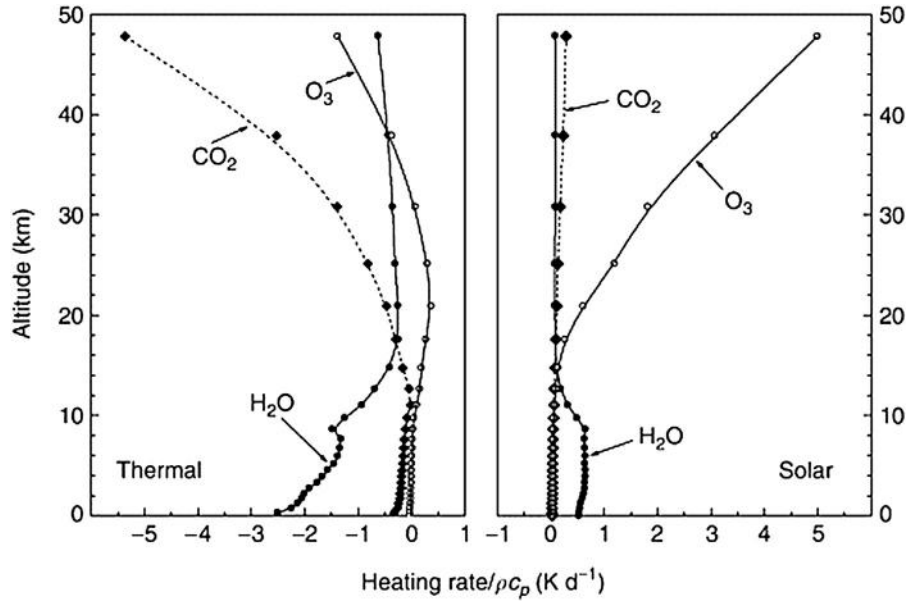


Figure 6 Atmospheric heating rates. These plots show the breakdown on the flux differences from [Figure 5](#) by altitude and by major species. The heating rate is divided by the volumetric heat capacity (ρc_p) and is expressed in K d^{-1} . The data correspond to those used in [Figure 5](#). The calculation was performed using a band model developed for the general circulation model CCM2 (the National Center for Atmospheric Research's Community Climate Model).

The absorption and emission of radiation (except in a constant-temperature enclosure) have an irreversible component. This arises because the brightness temperature of the radiation field will generally differ from the local kinetic temperature, and this difference must be thermalized through collisions. This is a disequilibrium process and, since it is assumed that the atmosphere is in local thermodynamic equilibrium, the

thermalization must, for consistency, be assumed to take place outside the CS. Consequently, the entropy increase that results from this irreversible process is accounted for by an increase of entropy in the radiation field, and has no direct significance for climate studies.

There is also a reversible component to the change of entropy during an interaction that, for the radiation field, is equal to but also the opposite sign of the corresponding change for the atmosphere. The rate of change of specific entropy for the atmosphere is given by [eqn \[15\]](#), where $\dot{q}_{\text{rad}}$ is the specific heating rate of the atmosphere.

$$\dot{s}_{\text{rad}} = \frac{\dot{q}_{\text{rad}}}{T} \quad [15]$$

The balance equation for entropy production in the entire CS may be written in the following form:

$$\int_{\text{CS}} (\dot{s}_{\text{mol}} + \dot{s}_{\text{rad}}) \rho dV = 0 \quad [16]$$

where dV is an element of volume, ρ is the density, and $\dot{s}_{\text{mol}}$ is the rate of change of specific entropy from irreversible molecular dissipative processes in the fluid. From the second law of thermodynamics, $\int_{\text{CS}} \dot{s}_{\text{mol}} \rho dV > 0$ so that $\int_{\text{CS}} \dot{s}_{\text{rad}} \rho dV < 0$ (i.e., radiation must extract entropy from the CS). According to [eqn \[15\]](#), this may be achieved if solar heating takes place at a higher temperature (lower altitude) than cooling by thermal radiation; this condition follows because most of the solar radiation is absorbed by the Earth's surface (see [Figure 5](#)).

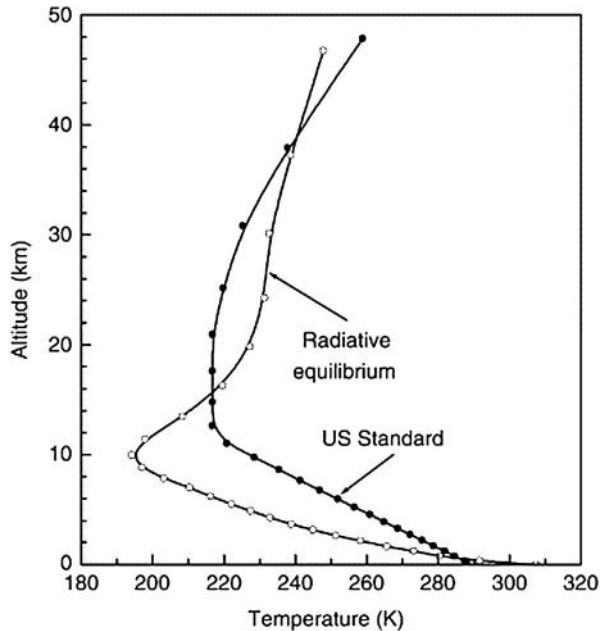


Figure 7 Radiative equilibrium temperatures. Compared to the US standard atmosphere. Solar conditions correspond to those employed in [Figure 6](#). The CCM2 radiation model was used.

Units and Nomenclature

SI units are used throughout. All nonstandard nomenclatures are defined in the text.

See also: **Aerosols:** Role in Radiative Transfer. **Chemistry of the Atmosphere:** Observations for Chemistry (*In Situ*): Particles. **General Circulation of the Atmosphere:** Energy Cycle. **Radiation Transfer in the Atmosphere:** Non-Local Thermodynamic Equilibrium; Scattering; Ultraviolet Radiation. **Tropospheric Chemistry and Composition:** Aerosols/Particles.

Further Reading

- Chandrasekhar, S., 1960. Radiative Transfer. Dover Publications, New York.
- Goody, R., 1995. Principles of Atmospheric Physics and Chemistry. Oxford University Press, New York.
- Goody, R.M., Yung, Y.-L., 1989. Atmospheric Radiation: Theoretical Basis, second ed. Oxford University Press, New York.
- Hansen, J.E., Travis, L.D., 1974. Light scattering in planetary atmospheres. Space Science Reviews 16, 527–610.
- Herzberg, G., 1945. Infrared and Raman Spectra of Polyatomic Molecules. Van Nostrand, New York.
- Herzberg, G., 1950. Spectra of Diatomic Molecules. Van Nostrand, New York.
- Liou, K.-N., 2002. An Introduction to Atmospheric Radiation, second ed. Academic Press, New York.
- Pivovonsky, M., Nagel, M.R., 1961. Tables of Blackbody Radiation Functions. Macmillan, New York.
- Planck, M., 1959. The Theory of Heat Radiation. Dover Publications, New York.
- Sobolev, V.V., 1975. Light Scattering in Planetary Atmospheres. Pergamon Press, Oxford.
- Stephens, G.L., 1994. Remote Sensing of the Lower Atmosphere: A Introduction. Oxford University Press, New York.
- Townes, C.H., Schawlow, A.L., 1975. Microwave Spectroscopy. Dover Publications, New York.

Cloud-Radiative Processes

Q Fu, University of Washington, Seattle, WA, USA

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

Clouds consist of water droplets with typical diameters of $\sim 20\ \mu\text{m}$ or ice crystals with a typical size of $\sim 100\ \mu\text{m}$ suspended in the atmosphere. Cloud-radiative processes are the interactions between these cloud particles and electromagnetic radiation, including the scattering, absorption, and emission processes. Radiative transfer, which serves as a mechanism for exchanging energy within the Earth-atmosphere system and between this system and the rest of the universe, is strongly modified by clouds. An understanding of cloud-radiative processes and their effects on radiative transfer is fundamentally important in climate studies and remote sensing applications.

Refractive Indices of Water and Ice

The refractive indices of pure water and ice, as functions of wavelength, are needed to determine the scattering and absorption properties of cloud particles. The complex refractive index $m(\lambda) = m_r(\lambda) + im_i(\lambda)$, where λ is the free-space wavelength, m_r is the real part of the refractive index, and m_i is the imaginary part. The real part of a refractive index is the ratio of the free-space speed of light to the phase speed of an electromagnetic wave in the medium. The imaginary part of a refractive index is related to the absorption coefficient by $4\pi m_i/\lambda$. For a real refractive index, only scattering can take place. For a complex index, both scattering and absorption are possible.

The refractive indices for water and ice have been measured extensively, and are comprehensively reviewed and tabulated. Figure 1 shows the real and imaginary refractive indices for water and ice as functions of wavelength. The real refractive indices of water exhibit large deviations from those of ice for wavelengths greater than about $10\ \mu\text{m}$. The imaginary refractive index of ice

shows a maximum in absorption at about $1.6\ \mu\text{m}$, where water exhibits a minimum. On the other hand, water has much greater absorption than ice at wavelengths larger than $10\ \mu\text{m}$.

The imaginary refractive indices for both ice and water are negligibly small in the UV and visible regions, but increase greatly in the near-infrared.

Light Scattering and Absorption by Cloud Particles

Scattering is a physical process by which a particle in the path of an electromagnetic wave continuously abstracts energy from the incident wave and reradiates it in all directions. The angular distribution of the scattered energy from a particle is described using a scattering phase function. Scattering is often accompanied by absorption, which results in the transfer of energy from the radiation field to the heat budget. The beam of radiation may be attenuated by both absorption and scattering, which together are called the extinction. In the field of light scattering and radiative transfer, it is customary to use the term

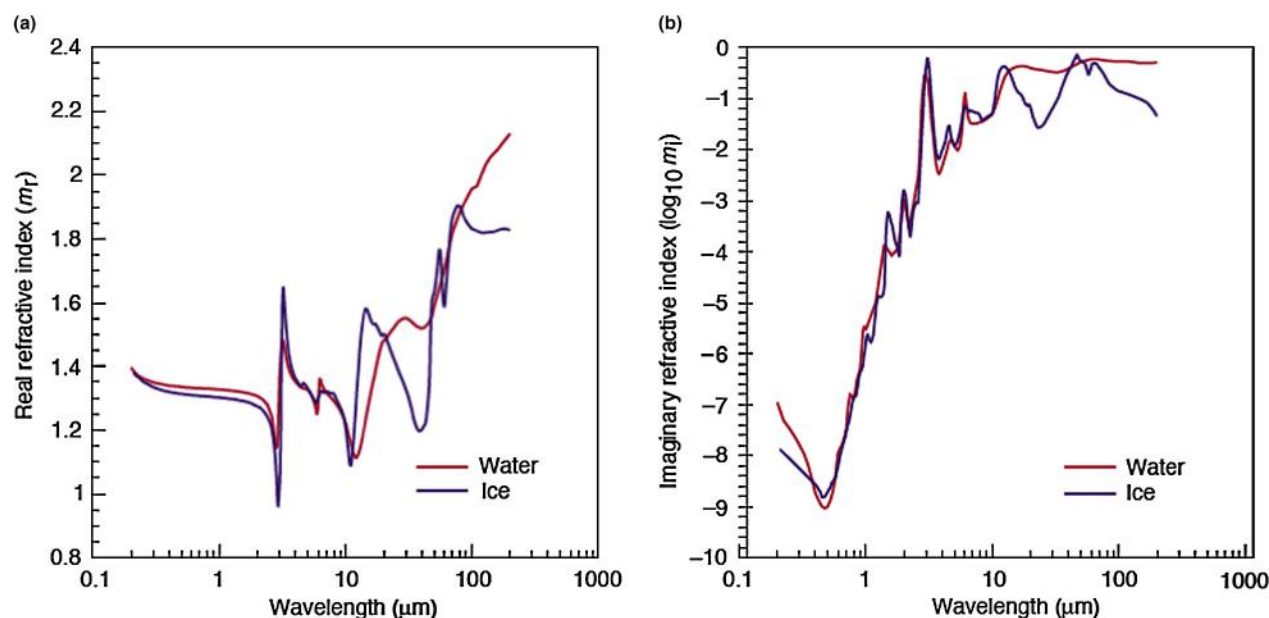


Figure 1 Refractive indices for water (red lines) and ice (blue lines) as functions of wavelength: (a) Real part and (b) imaginary part.

'cross-section,' which is analogous to the geometric cross-sectional area of a particle, to denote the amount of energy removed from the original beam by the particle. An extinction cross-section is the sum of the scattering and absorption cross-sections. The extinction (scattering or absorption) efficiency is the ratio of an extinction (scattering or absorption) cross-section to a particle's geometric cross-sectional area.

Light scattering and absorption by cloud particles depend on the size and shape of the particles, the wavelength of radiation, the particles' refractive indices, and the viewing geometry. The interactions of radiation and liquid droplets in water clouds are well described by Mie theory, a complete solution of Maxwell's equations for a plane electromagnetic wave incident on a dielectric sphere. By considering the scattered field at a large distance from a sphere, we can derive the scattering and extinction cross-sections and the scattering phase function.

Figure 2 shows the extinction (Q_e), scattering (Q_s), and absorption (Q_a) efficiencies and the asymmetry factor as functions of the size parameter, defined as $2\pi r/\lambda$, where r is the radius of the sphere. The asymmetry factor is the first moment of the scattering phase function, which is positive or negative respectively as the particle scatters more energy into the forward or backward direction. The refractive indices used are those of water at wavelengths of 0.55 and 11 μm . In the limit of large size parameters, the extinction efficiency approaches a value of 2.

For $m_i = 0$, there is no absorption, so that $Q_s = Q_e \cdot Q_s$ (or Q_e) in this case shows a series of major maxima and minima and ripples. The major maxima and minima are due to interference of light diffracted and transmitted by the sphere, whereas the ripple arises from edge rays that are grazing and traveling the sphere, spewing off energy in all directions. Q_s approaches 2 for large size parameters, which applies to the interaction of cloud droplets with visible radiation. Since all

visible wavelengths are scattered nearly equally well, clouds therefore appear white.

At 11 μm , the absorption efficiency is on the order of 1 for the size parameters larger than ~ 6 . Therefore, cloud droplets absorb nearly all of the infrared radiation incident on them. They act essentially as black bodies. Scattering efficiencies in the infrared with large size parameters result mainly from the diffraction around the sphere.

In clouds, there is usually a distribution of droplet sizes. Suppose that $N(r)dr$ is the number of droplets per unit volume in the radius range r to $r + dr$. If the scatterers are sufficiently far apart (many wavelengths) that they act independently, the extinction coefficient of cloud droplets, which is the total extinction cross-sections of cloud droplets per unit volume, is given by $\beta_e = \int \pi r^2 Q_e N(r) dr$, where Q_e is the extinction efficiency. The scattering (absorption) coefficient is obtained similarly but by replacing the extinction efficiency with the scattering (absorption) efficiency. The ratio of scattering to extinction coefficients is called the single-scattering albedo of cloud droplets. In the geometric limit (i.e., the limit of large size parameters), we have $\beta_e = \frac{1.5 \text{ LWC}}{\rho_w r_e}$, where

LWC is cloud water content (i.e., $\frac{4}{3} \pi \rho_w \int r^3 N(r) dr$), r_e the effective radius (i.e., $\int r^3 N(r) dr / \int r^2 N(r) dr$), and ρ_w the liquid water density. The vertical integration of the extinction coefficient gives the cloud optical depth that is the total extinction cross-sections of cloud droplets in a vertical column per unit surface area.

Cirrus clouds are composed almost exclusively of nonspherical ice crystals, with a wide range of shapes, such as columns, plates, hollow columns, bullet rosettes, aggregates, and so on. The familiar halos associated with cirrus clouds

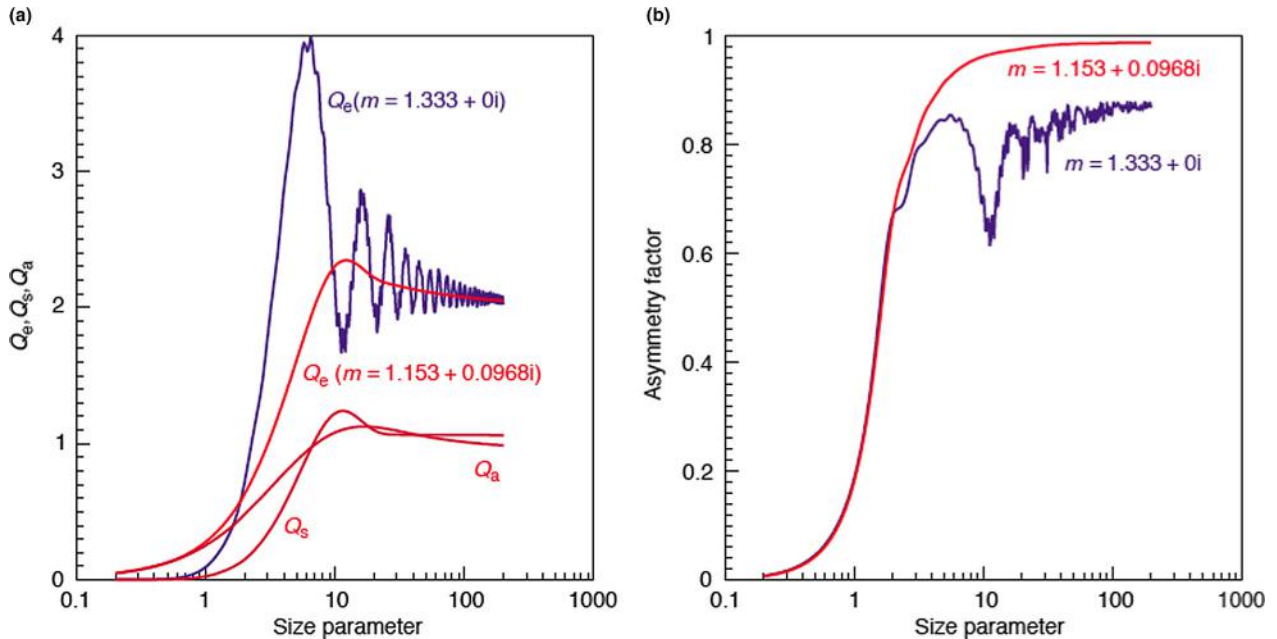


Figure 2 (a) Extinction (Q_e), scattering (Q_s), and absorption (Q_a) efficiencies and (b) the asymmetry factor of spheres as functions of size parameters. The refractive indices (m) of 1.333 and $1.153 + 0.0968i$ are those of water at wavelengths of 0.55 and 11 μm .

result from prismatic crystals with the hexagonal structure. Unlike the scattering of light by spherical water droplets, there is no general solution for light scattering and absorption by nonspherical ice particles. Geometric optics, essentially ray tracing, is often used with nonspherical ice particles having large size parameters, while numerical solutions of Maxwell's equations are used with small size parameters. Key gaps, however, still exist in our knowledge of light scattering and absorption by complex ice particles.

Radiative Transfer in a Cloudy Atmosphere

The influence of clouds on atmospheric radiation fields is governed by a radiative transfer equation. If the intensity of radiation I_λ becomes $I_\lambda + dI_\lambda$ after traversing a thickness ds in the direction of its propagation, then we can write eqn [1], where $\beta_{e,\lambda}$ is the extinction coefficient for radiation of wavelength λ , and j_λ is the source-function coefficient. The first term on the right-hand side of eqn [1] represents the reduction in the intensity caused by absorption and scattering of radiation by the atmosphere and clouds; the second term represents the contribution from emission of the layer plus a result of the radiation redirected from all other directions by the scattering process into the direction under consideration. By defining the source function $J_\lambda = j_\lambda/\beta_{e,\lambda}$, we obtain eqn [2].

$$dI_\lambda = -\beta_{e,\lambda} I_\lambda ds + j_\lambda ds \quad [1]$$

$$\frac{dI_\lambda}{\beta_{e,\lambda} ds} = -I_\lambda + J_\lambda \quad [2]$$

Knowledge of scattering and absorption coefficients and the scattering phase function is required for determination of the source function. In local thermodynamic equilibrium, the source function associated with an emission process is proportional to the Planck function, a known function of frequency and temperature. The scattering source function involves an integral over all directions of incidence, which transforms the radiative transfer equation into an integrodifferential equation. This feature of the scattering process greatly increases the complexity of radiation calculations. Many efficient and accurate methods of solving the radiative transfer equation have been developed for numerical models and remote sensing applications by assuming that clouds are uniform and infinite in the horizontal, which is called the plane-parallel cloud assumption.

For the calculation of radiative transfer in cloudy atmospheres, we must also consider gaseous absorption. The number of monochromatic calculations that are needed for molecular line spectra is so great that they can be used only for rare occasions. The correlated-k method solves this problem by approximating frequency integrals over line spectra by sums over a few finite intervals of new independent variables.

Clouds reflect and absorb solar radiation. Solutions of the radiative transfer equation can be used to determine the radiative properties of clouds under the plane-parallel assumption. If the cloud particle size distribution and vertical distribution of humidity are given, then the cloud albedo and absorption of solar radiation depend on the cloud water path and the solar zenith angle. The cloud water path is the total mass of cloud water in

a vertical column of atmosphere per unit of surface area. The cloud optical depth is proportional to the cloud water path and inversely proportional to the effective radius of cloud droplets.

The cloud albedo increases with the cloud water path and also with the solar zenith angle. It increases with the cloud water path most rapidly for smaller values of cloud water paths, and approaches a limit for very large cloud water paths. The absorption of solar radiation by plane-parallel clouds decreases with increasing solar zenith angle, but increases with the cloud water path. For a typical water droplet size distribution and a solar zenith angle of 60° , the cloud albedo can range from $\sim 25\%$ to over 80% , and its absorption from ~ 3 to $\sim 11\%$, as the cloud water path increases from 10 to 1000 gm^{-2} .

The cloud albedo and absorption of solar radiation are also sensitive to the cloud particle size. Keeping the cloud water path fixed, the albedo is larger for smaller droplets, mainly because these present a larger surface area for the same mass.

Clouds effectively absorb and emit terrestrial radiation. With cloud water paths greater than about 20 gm^{-2} , water clouds become opaque to terrestrial radiation. For most water clouds, it is a good approximation to assume that cloud surfaces absorb and emit terrestrial radiation like black bodies. Unlike water clouds, cirrus clouds are usually partially transparent to terrestrial radiation owing to their larger cloud particle sizes and smaller cloud water paths. Recent studies also suggest that the neglect of scattering processes in the infrared for cloudy atmospheres leads to a few Wm^{-2} errors in estimating the outgoing longwave radiation.

Contrary to the plane-parallel cloud assumption, real clouds are horizontally finite and inhomogeneous. With an overhead sun, solar radiation may leak from the sides of clouds, which leads to increased transmission and absorption of solar radiation throughout the atmosphere. At large solar zenith angles, the reflected and absorbed solar radiation increase, owing to the increased interception and reflection of solar radiation by cloud sides. Finite clouds also have significant effects on the transfer of long-wave radiation.

See also: **Aerosols:** Aerosol–Cloud Interactions and Their Radiative Forcing. **Clouds and Fog:** Classification of Clouds; Climatology; Measurement Techniques *In Situ*. **Numerical Models:** Parameterization of Physical Processes: Clouds.

Further Reading

- Bohren, C.F., Huffman, D.R., 1983. *Absorption and Scattering of Light by Small Particles*. Wiley Press, New York.
- Goody, R., 1995. *Principles of Atmospheric Physics and Chemistry*. Oxford University Press, New York.
- Hartmann, D.L., 1994. *Global Physical Climatology*. Academic Press, San Diego.
- Liou, K.N., 1992. *Radiation and Cloud Processes in the Atmosphere: Theory, Observation, and Modeling*. Oxford University Press, New York.
- Liou, K.N., 2002. *An Introduction to Atmospheric Radiation*. Academic Press, San Diego.
- Mishchenko, M.I., Hovenier, J.W., Travis, L.D., 2000. *Light Scattering by Nonspherical Particles*. Academic Press, San Diego.
- Salby, M.L., 1996. *Fundamentals of Atmospheric Physics*. Academic Press, San Diego.
- Thomas, G.E., Stamnes, K., 1999. *Radiative Transfer in the Atmosphere and Ocean*. Cambridge Press, Cambridge.

Non-Local Thermodynamic Equilibrium

M López-Puertas and B Funke, Instituto de Astrofísica de Andalucía, CSIC, Granada, Spain

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

This article introduces the concept of non-local thermodynamic equilibrium (non-LTE) in planetary atmospheres and provides a basic theoretical framework of non-LTE radiative transfer and statistical equilibrium of molecular states. Accurate and approximate solutions of the non-LTE problem are discussed. A climatology of non-LTE populations of the most important vibrational levels of the atmospheric species is provided. Examples of non-LTE applications in general circulation models and the remote sensing of the Earth's and other planetary atmospheres are given.

Introduction

The atmosphere, viewed as a thermodynamic system, is never in true thermodynamic equilibrium at any level. A given parcel of air is always in the process of changing its properties through the exchange of particles and photons with its surroundings, through collisions between molecules, and through chemical changes. It can, however, be in a state known as 'local' thermodynamic equilibrium (LTE), a concept that permits calculations of radiative transfer in the atmosphere, and its energy exchanges (both internally and with its upper and lower boundaries), to be made relatively simply. In general, LTE applies if the populations of the energy levels within a molecule are the same, or nearly the same, as they would be under true thermodynamic equilibrium conditions. LTE occurs when collisions are so frequent that the energy level populations depend predominantly on the local kinetic temperature, as defined by the Maxwellian statistics of molecular motion. It is worth noting that this definition of LTE does not impose restrictions upon the radiation field of the emission involved. That is, unlike in strict thermodynamic equilibrium, in LTE the local radiation field is not necessarily described by Planck's function at the local temperature. Hence, LTE is compatible with a net gain or loss of radiative energy (heating or cooling) by the gas in the transition in question, provided that collisions are quick enough to supplement the energy sink and to keep the population of the excited state coupled with the translational reservoir.

Conversely, non-LTE arises when the 'internal' temperature of a representative molecule, as determined by the statistics of the relative populations of the vibrational and rotational energy levels, becomes different from the 'external' temperature, as determined by the statistics of the velocity distribution of the molecules making up a parcel of gas. Any difference depends primarily on the mean rate at which collisions take place between the individual molecules, which in turn depends mainly on the number density, and thus on the pressure and temperature of the sample. Since the pressure dependence usually dominates, non-LTE effects in atmospheres are mainly of importance at the higher altitudes, that is, in the upper stratosphere and above.

Basic non-LTE theory and most theoretical models of planetary atmospheres deal with the transfer of energy as (primarily infrared) radiation within an atmosphere with a fixed composition and no fluid dynamics. A refinement

would be to include basic photochemistry (i.e., the situation where photon-molecule encounters lead to the breakup of the molecule and the subsequent formation of a different stable or quasistable species). This is, of course, another example of a nonequilibrium process, which depends on the nature of the photon field and often the internal state of the molecule also.

Important applications of non-LTE theory in the atmosphere include the calculation of thermal cooling and solar heating rates. Dynamical models of the upper atmosphere are very limited if they do not include realistic calculations of the heating and cooling that drive the motions, and these are strongly affected by non-LTE. Furthermore, the growth of infrared detector technology in recent decades has allowed sounding of the middle and upper atmosphere using the limb emission measured by infrared sensors on satellites. The interpretation of these data in terms of temperature and composition profiles depends on non-LTE radiative transfer theory, especially in the middle and upper atmosphere.

Breakdown of Local Thermodynamic Equilibrium

The concept of LTE was introduced by Schwarzschild in 1906 in his study of stellar atmospheres. He realized that, in a star just as in a planetary atmosphere, any individual parcel of gas is not isolated, and, in principle, no equilibrium can be defined. However, the concept of equilibrium is so valuable in practice that it is worth looking for situations where it can be assumed as a close approximation to reality. Thus, a rotational or vibrational energy level is in LTE if its population is given by Boltzmann's law at the local kinetic temperature, T , that is,

$$\frac{n_{v,r}}{n_0} = \frac{g_{v,r}}{g_0} \exp\left(-\frac{E_{v,r}}{kT}\right) \quad [1]$$

where $n_{v,r}$ is the number density of the upper (vibrational or rotational) state, $E_{v,r}$ is its energy, n_0 is the number density of the ground state, $g_{v,r}$ and g_0 are their respective degeneracies, and k is the Boltzmann constant. The emission from the level is then characterized by the Planck function, $B(T)$, at temperature T .

When the population of the level deviates from Boltzmann's law, we say that the energy level is in non-LTE or has a non-LTE population. This can be characterized by introducing an 'excitation' (vibrational or rotational) temperature $T_{v,r}$, defined by

eqn [1] with T replaced by $T_{v,r}$. It then follows that if the vibrational or rotational temperature of an energy level differs from the local kinetic temperature, the level is in vibrational or rotational non-LTE. Similarly, the radiance emitted by the level considered is said to be non-LTE emission and so is the cooling or heating it produces.

For a real atmospheric parcel containing different kinds of molecules, some can be in LTE and others not. In fact, it is possible (and, at intermediate pressures like those found in the upper stratosphere, quite usual) for one internal energy state to be in LTE, and another not, in the same individual molecule (see Table 1). The reason for the dependence on height is the dependence on pressure of the collision frequency, and the different efficiencies for transferring energy to vibrational modes during collisions. In the atmosphere, pressure and therefore frequency of collisions fall off with height. Thus, we should expect a given transition to be more likely to be in non-LTE in the upper atmosphere.

Another important parameter influencing the height at which non-LTE occurs is the energy of the transition. For those corresponding to small energy jumps, the average number of collisions required to keep the levels in equilibrium is smaller, so they can be in LTE up to higher altitudes in the atmosphere. For example, the CO₂ 4.3 μm band starts to depart from LTE at approximately 50 km, while the O(³P) states that emit at 63 μm and the purely rotational levels of virtually all atmospheric molecules, emitting in the microwave part of the spectrum, are in LTE up to the high thermosphere.

There are two common ways in which the internal states of any given molecule can ‘know’ what the ensemble of molecules in the gas is doing. One is through collisions, and the other through exchanging photons (Figure 1). If the former dominates, then LTE is normally guaranteed, but not necessarily

where the populations are controlled by the transfer of radiation. Under optically thin conditions, the absolute intensity, directional distribution, and frequency spectrum of the radiation field may have little or no relation to the Planck function at the local kinetic temperature, and hence drives the level populations to non-LTE. The reverse may be true under optically thick conditions, but then collisions are usually plentiful as well.

Depending on the processes at work, we can generally distinguish between the following non-LTE situations:

1. Classical non-LTE, where, in the absence of a strong radiative source, thermal collisions are not fast enough to supply the energy lost by spontaneous emission and so the population of the excited state is smaller than that corresponding to LTE (e.g., CO₂(0,1¹,0) in Figure 2). The level at which a transition departs from LTE can be estimated in such cases as that in which the collisional relaxation time is of the same order as the spontaneous radiative lifetime, particularly if the transition is optically thin at this height. For optically thick conditions, a better approximation is the altitude at which the thermal losses equal the spontaneous emission multiplied by the probability of photon escape to space.
2. The internal atmospheric radiation field can cause a breakdown of LTE when it is responsible for overpopulating the excited level with respect to the Boltzmann distribution. This situation is common for molecular states that give rise to weak infrared bands in the cold upper mesosphere, since their populations are sensitive to the flux of infrared photons from the warm lower atmosphere (e.g., the (0,1¹,0) level of the ¹⁶O¹²C¹⁷O isotopologue (627) and H₂O(010) in Figure 2).

Table 1 Approximate altitude of non-LTE departure of the most important energy levels of atmospheric species

Species	Band origin wavelength (μm)	Energy level	Approximate altitude ^a (km)
CO ₂	15	01 ¹ 0	90
CO ₂	15	02 ² 0	80
CO ₂	15	03 ³ 0	70
CO ₂	10	00 ⁰ 1	50
CO ₂	4.3	00 ⁰ 1, 01 ¹ 1	50
CO ₂	4.3	02 ⁰ 1, 02 ² 1, 10 ⁰ 1	40
CO ₂	2.7	02 ⁰ 1, 10 ⁰ 1	40
CO	4.6	1	40
O ₃	10	001, 100	70
O ₃	10	$\nu_3 = 2,3; 4; \nu_3 > 4$	60; 50; 30
O ₃	14.8	0,1,0	70
H ₂ O	6.3	010; 020	60; 40
H ₂ O	2.7	100, 001	10
CH ₄	7.6	ν_4	70
CH ₄	6.5	ν_2	70
NO	5.3	1	20
N ₂ O	17	010, 020	80
N ₂ O	7.8	100	75
N ₂ O	4.5	001	50
NO ₂	6.2	001; $\nu_3 > 1$	50; 10
O(³ P)	63	O(³ P ₁)	>200

^aThe altitude of LTE breakdown depends on the atmospheric conditions (latitude and season) and the solar illumination (see Figures S2–S27). For some levels, it changes considerably from day to night. The values are given here for daytime, when this altitude is lower. Note that non-LTE effects are observed at lower tangent heights in limb radiances.

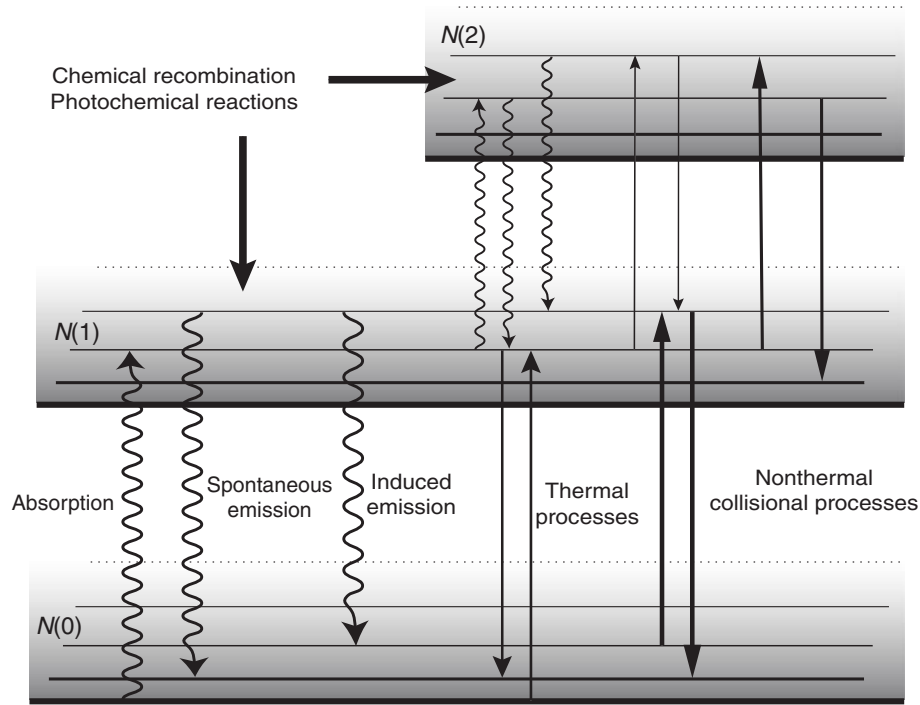


Figure 1 Processes affecting the populations of vibrational levels in the atmosphere.

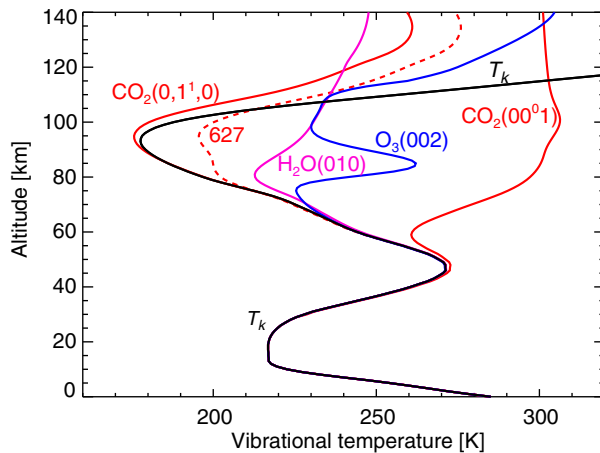


Figure 2 Vibrational temperatures for some species and levels for a typical kinetic temperature profile T_k illustrating the different processes causing non-LTE populations.

3. Non-LTE breaks down when the absorption of the strong solar radiation field is the prime process for the population of the excited levels. This occurs mainly for states that emit in the near-infrared part of the spectrum (e.g., the (0,0,1) level of CO_2 in [Figure 2](#)).
4. There are also situations where molecules are excited and de-excited by processes such as chemical recombination, photochemical reactions, electronic–vibrational and vibrational–vibrational energy transfers, dissociative recombination, and charged particles collisions (e.g., $\text{O}_3(\nu_3 = 2)$ in [Figure 2](#) is an example of the chemical recombination that forms O_3). These are also largely responsible for the

atmospheric emissions that constitute the terrestrial day-glow, night-glow, and auroral spectra, which historically have not been referred to as non-LTE emissions; however, since the populations of the excited states that produce these phenomena are far from a Boltzmann distribution at the local kinetic temperature, they should come under this heading.

Basic Non-LTE Theory

Historically, the solution of non-LTE situations has been linked to the development of radiative transfer theory and to the description of the radiating properties of matter. After the first treatment by Milne in 1930 of radiative transfer under non-LTE conditions for stellar atmospheres, this subject has been extensively studied in astrophysics, mainly to interpret stellar spectra and to study line formation in hot stars.

In 1949, Spitzer first pointed out the possibility that at the low pressures found in the upper atmosphere, the radiative field could upset the state of LTE. No quantitative treatment was given, however. The first application of a non-LTE formulation in the terrestrial atmosphere was by Curtis and Goody in their 1956 study of the CO_2 15 μm cooling rate in the mesosphere. They formulated the problem for the simplest case of a two-level transition, including only thermal collisions and radiative processes. At the same time, coinciding with the development of the first electronic computers, Curtis devised a linear parameterization of the radiative transfer equation to calculate the heating rates induced by the CO_2 15 μm bands.

To obtain the population of n_2 and n_1 , and then the vibrational temperatures from [eqn \[1\]](#), as shown in [Figure 2](#),

we consider the statistical equilibrium equation. Let us consider that the two levels are connected (Figure 1) by

1. radiative processes (spontaneous emission, induced emission, and absorption) through the vibration-rotation band (2-1);
2. thermal collisions or vibrational-translational (V-T) processes of this form:

$$k_t : n(2) + M \rightleftharpoons n(1) + M + \Delta E \quad [2]$$

where M is any air molecule, and $\Delta E = E_2 - E_1 = h\nu_0$ is the energy difference of the upper and lower levels;

3. nonthermal collisional processes of this kind:

$$k_{vv'} : C(v) + n(1) \rightleftharpoons n(2) + C(v') + \Delta E_v \quad [3]$$

$$k_{ev} : N^* + n(1) \rightarrow n(2) + N + \Delta E_{ev} \quad [4]$$

$$k_c : A + B + M \rightarrow n(2) + M \quad [5]$$

at rates $k_{vv'}$, k_{ev} , and k_c . $C(v)$ is an atmospheric molecule excited in vibrational level v before, and in v' after, the collision; and N^* is an excited atom or a molecule electronically or vibrationally excited. $\Delta E_v = E_v - E_{v'} - h\nu_0$ and $\Delta E_{ev} = E^* - h\nu_0$ are the energy excesses, with E^* being the energy of N^* .

The population of n_2 is then given by

$$\frac{n_2}{n_1} = \frac{B_{12}\bar{L}_{\Delta\nu} + p_t + p_{nt}}{A_{21} + B_{21}\bar{L}_{\Delta\nu} + l_t + l_{nt}} \quad [6]$$

where A_{21} , B_{21} , and B_{12} are the Einstein coefficients for spontaneous emission, induced emission, and absorption, respectively. $p_t = k'_t[M]$ is the thermal collisional production, where k'_t is the rate coefficient of process (eqn [2]) in the reverse direction, related to k_t by a detailed balance:

$$\frac{k'_t}{k_t} = \frac{g_2}{g_1} \exp\left(-\frac{E_2 - E_1}{kT}\right) \quad [7]$$

$p_{nt} = k_{vv}[C(v)] + k_{ev}[N^*] + k_c[A][B]/n_1$ is the specific production rate of n_2 molecules due to the nonthermal processes (eqns [3-5]), and $l_t = k_t[M]$ and $l_{nt} = k'_{vv}[C(v')]$ are the specific loss rates of n_2 in thermal (eqn [2]) and nonthermal (eqn [3]) processes. $\bar{L}_{\Delta\nu}$ is the mean radiance averaged over the spectral interval of the band, $\Delta\nu$, defined by

$$\bar{L}_{\Delta\nu} = \frac{1}{4\pi S} \int_{\Omega} \int_{\Delta\nu} L_{\nu} k_{\nu} d\nu d\Omega \quad [8]$$

where k_{ν} is the absorption coefficient, $S = \int k_{\nu} d\nu$ is the band strength, and Ω is the solid angle.

In order to get the population n_2 , we need to know the mean radiance $\bar{L}_{\Delta\nu}$, which is governed by the radiative transfer equation that, for a stratified atmosphere, takes the following form:

$$\mu \frac{dL_{\nu}(z, \mu)}{dz} = -k_{\nu} n_a [L_{\nu}(z, \mu) - J_{\nu}(z)] \quad [9]$$

where z is altitude; $\mu = \cos(\theta)$, with θ being the angle measured from the vertical; n_a is the number density of absorbing molecules; and J is the source function of the (2-1) band.

To complete the system, we need the expression that relates the source function with the population of the upper and lower states given, for an infrared transition, by

$$J_{\nu_0} = \frac{2h\nu_0^3}{c^2} \left[\frac{n_1 g_2}{n_2 g_1} - 1 \right]^{-1} \quad [10]$$

Thus, from the solution of the statistical equilibrium (eqn [6]) and the radiative transfer (eqn [9]), we can obtain the population of the upper state n_2 and the source function J_{ν_0} .

In addition to the non-LTE populations, we are also generally interested in the heating and cooling rates produced by the transitions between the energy states. The heating rate h (or, more precisely, the divergence of the radiative flux) of the band arising from the levels n_2 and n_1 can then be computed, once n_2 and the source function J_{ν_0} are known, by using

$$h = 4\pi S n_a [\bar{L}_{\Delta\nu} - J_{\nu_0}] \quad [11]$$

When interested in only the cooling and heating rates but not the populations, one may solve the problem by solving the statistical equilibrium equation (eqn [6]) and the radiative transfer equation in the form of eqn [11].

Simple Solutions of Non-LTE

One of the simplest cases of non-LTE is to assume that the excitation from absorption of radiation is negligible while radiative losses are important. The problem is then reduced to a 'local' solution. In this case, there are several approaches depending on which radiative losses we consider. The most extreme case is to assume that all photons emitted are lost and none are reabsorbed. The approach is usually called 'total escape to space.' In this case, the population can be approached by

$$\frac{n_2}{n_1} = \left(\frac{l_t}{A_{21} + l_t} \right) \frac{\bar{n}_2}{\bar{n}_1}, \quad J_{\nu_0} \approx \frac{B_{\nu_0}}{1 + A_{21}/l_t} \quad [12]$$

where the overbar denotes LTE populations. Under these conditions, we expect non-LTE to set in at those altitudes where l_t is significant or smaller than A_{21} . The density of n_2 will then be smaller than that corresponding to LTE.

For these conditions of no significant radiative excitation, the heating rate can be approximated by

$$h_{12} \approx -n_2 A_{21} h\nu_0 \quad [13]$$

and, including n_2 from eqn [12], we get

$$h_{12} \approx -\left(\frac{A_{21} l_t n_1}{A_{21} + l_t} \right) \frac{\bar{n}_2}{\bar{n}_1} h\nu_0 \quad [14]$$

A similar approach is to assume that only half of the emitted photons are lost; this is usually called 'escape to space.' This is normally interpreted as follows: photons emitted to the hemisphere above the considered layer are lost, while those emitted to the downward denser atmosphere are fully recovered by absorption from the layers below. The expressions for the non-LTE populations and heating rates are those of eqns [12] and [14] but include dividing A_{21} by 2.

A further refinement along this line of 'local' approaches is the 'cool-to-space' approach. This introduces the concept of the 'escape probability' function that gives the probability that

a photon, emitted at an altitude z , escapes to space in any direction in the upper hemisphere. This is defined by

$$\Gamma(z) = \frac{1}{S} \int_{\Delta\nu} \int_0^1 k_\nu(z) \exp[-\tau_\nu(z, \infty, \mu)] d\mu d\nu \quad [15]$$

where $\tau_\nu(z, \infty, \mu)$ is the optical depth at direction $\mu = \cos(\theta)$. The non-LTE populations, source function, and heating rate for this approach are also given by eqns [12] and [14] by multiplying A_{21} by $\Gamma/2$. These expressions tend to those for the ‘escape-to-space’ approach, where the band becomes optically thin and Γ tends to unity. For optically thick conditions, $\Gamma \rightarrow 0$ and hence $J \rightarrow B$, the upper level remains in LTE even though the radiative relaxation time may be significantly smaller than the time between thermal collisions. This approach describes very well the departure from LTE of levels where the band originating from them is optically thick and a few collisions per second are then enough to keep the levels in LTE. A typical example is the $\text{CO}_2(0,1,0)$ level in the atmospheres of Mars and Venus. Figure 3 illustrates the cooling rates for the CO_2 15 μm fundamental band in the atmosphere under those approximations and compared to the full non-LTE solution.

A further approximation, plausible at high altitudes, is to assume that the thermal collisional losses are much smaller than the radiative emissions (i.e., $A_{21} \gg l_t$). Then, eqns [12] and [14] reduce to

$$\frac{n_2}{n_1} \approx \frac{l_t}{A_{21}} \frac{g_2}{g_1} \exp\left(-\frac{h\nu_0}{kT}\right) \quad [16]$$

$$h_{12} \approx -l_t n_1 \frac{g_2}{g_1} \exp\left(-\frac{h\nu_0}{kT}\right) h\nu_0 \quad [17]$$

Note that the heating rate (which is actually cooling) is solely controlled by the specific rate of thermal collisions, l_t , independent of the value of A_{21} , and therefore independent of whether the radiative losses are A_{21} or $A_{21}/2$.

There is another common situation where we have a strong radiation source, usually the Sun, which is so strong that the radiative exchange between layers can be neglected and we still have a ‘local’ solution. Productions by thermal collisions can then also be neglected, but not the losses. In this case, the population can be approached by

$$\frac{n_2}{n_1} = \frac{I}{A_{21} + l_t} \quad [18]$$

where I is the photoabsorption coefficient given by

$$I(z) = \int_{\Delta\nu} \Phi_{\nu,\infty} k_\nu(z) \exp\left[-\int_z^\infty \frac{k_\nu(z') n_a(z')}{\cos \chi} dz'\right] d\nu \quad [19]$$

where $\Phi_{\nu,\infty}$ is the solar flux at the top of the atmosphere, and χ is the solar zenith angle. Note that the curvature of the Earth has been neglected here. The heating rate can be obtained, from eqns [10], [11], and [18], by

$$h_{12} \approx \frac{l_t}{A_{21} + l_t} n_1 I h\nu_0 \quad [20]$$

In the highest altitudes, where thermal losses are negligible, $l_t \rightarrow 0$, the heating rate tends to be zero. That is, all of the radiation absorbed from the Sun is reemitted and no net kinetic energy is transferred to the atmospheric molecules. Opposite, at the lowest regions, where $l_t \gg A_{21}$, almost all of the solar energy absorbed, $n_1 I h\nu_0$, is converted to kinetic energy.

Non-LTE Models

The theory outlined in this article has been incorporated into models that calculate the non-LTE populations of the vibrational levels of the atmospheric species, and that derive from these the cooling and heating rates produced by the transitions that occur between them. The populations are useful mainly for

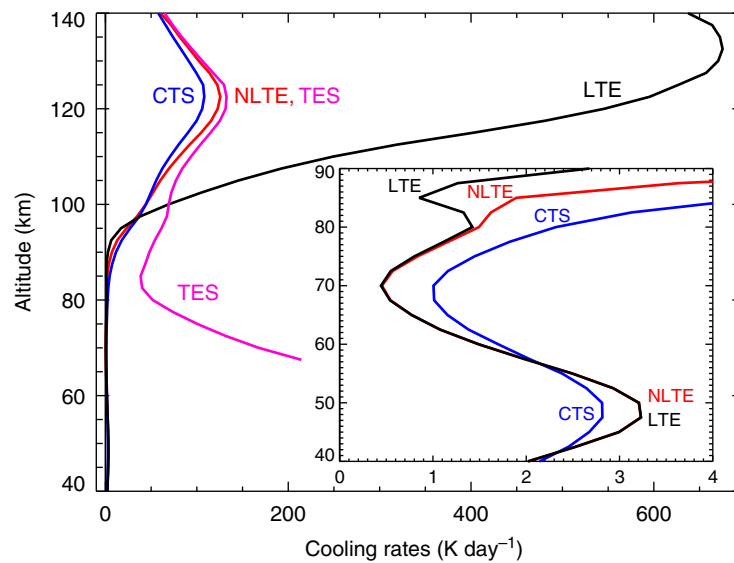


Figure 3 Cooling rates of the CO_2 15 μm fundamental band in the atmosphere for LTE, non-LTE, and the ‘total scape to space’ (TES) and ‘cool-to-space’ (CTS) approximations.

the interpretation of remote sensing data, while heating rate calculations are an essential component of the general circulation models of the upper atmosphere.

The non-LTE models solve the system formed by the statistical equilibrium equation for each of the molecular vibrational energy levels, and the radiative transfer equation for the significant transitions that take place between them. All thermal and nonthermal collisional processes involved in excitation and deexcitation of the levels must be included in the statistical equilibrium equation (eqn [6]). The radiative transfer equation (eqn [9]) must, in the non-LTE case, take account of (1) spontaneous emission; (2) the absorption of radiation from external sources, especially tropospheric, stratospheric, and solar radiation; and (3) the transfer of photons between each atmospheric layer and every other layer. The last term should incorporate, as for the LTE case, the frequency variation of the atmospheric absorption coefficient, which includes a large number of vibration-rotation lines with shapes that vary with temperature and pressure. The solution requires the integration of the radiative transfer equation over frequency and over solid angle to obtain the flux transmittances, followed by the solution over altitude of the coupled system of radiative transfer and statistical equilibrium equations. When the V-V energy exchange between vibrational levels with non-LTE populations is very fast, the coupled equations of statistical equilibrium for all the affected levels and of radiative transfer for all bands originating from those levels have to be solved simultaneously in order to obtain accurate populations.

This is achieved by inversion, iteration, or a combination of the two. A popular and flexible method for performing the numerical integration is that originally suggested by Curtis, which is based on the replacement of the integration over height by a summation over the atmospheric layers. It provides both the populations of the energy levels and the cooling and heating rates, and gives information about cooling to space and the radiative exchange between layers. The Curtis matrix method can be used to calculate cooling rates under LTE conditions as well, and in fact it was originally derived for that purpose. Another commonly used approach is the Lambda iteration, which alternates statistical equilibrium equation calculations, involving all energy levels, with radiative transfer calculations, involving all atmospheric layers. This technique is a rigorous approach, although it usually has the disadvantage of slow convergence in non-LTE situations driven by strong absorption and emission (i.e., optically thick conditions). Nevertheless, different acceleration strategies have been devised and employed in practical applications. In practice, when many energy levels and transitions are involved and are strongly coupled by collisional and/or radiative processes, models tend to use a combination of both techniques.

Non-LTE radiative transfer models represent a formidable computational task, and approximations need to be made whenever possible. Those that are often assumed include neglecting stimulated emission, assuming rotational LTE, grouping levels with very similar energies into resonant sets that can be treated as a single level, and neglecting overlapping between spectral lines of different bands. The most difficult situation is that where there is overlapping between the lines of different bands. Because of the different source

functions in each band, it then becomes necessary to evaluate the radiative transfer equation for each fine-mesh frequency point.

The usual inputs for the non-LTE models are the pressure, temperature, and composition profiles; the details of the relevant collisional thermal and nonthermal processes; and the spectroscopic line data to perform the radiative transfer calculation. Knowledge of the collisional activation and deactivation rates, energy transfer (V-T, V-V, or electronic to vibration) collisional rates and efficiencies, and nascent distributions in chemical reactions is the key input to the models, and the accuracy of the computed populations heavily relies on their uncertainties. In that sense, the models require laboratory measurements of those parameters in many cases, not to mention their dependencies on temperature that can range from 100 to 1000 K. For certain species and/or energy levels, the collisional rates have not been measured (e.g., for the deexcitation of high-energy states). Then, approximations are frequently used, as the harmonic oscillator approach. Spectroscopic line data are also very important. This is normally taken from compilations as HITRAN and GEISA. However, it is rather common that, for transitions from high energy levels and hot and overtone bands, information is missing, and approaches and theoretical approximations must then be used.

The typical outputs of the models are the non-LTE populations, frequently expressed in terms of the vibrational temperatures as described in this article and shown in Figure 2, and cooling or heating rates, like those shown in Figure 3.

Since the first model in 1956 for the CO₂ 15 μ m band, models have been developed for many other vibrational energy levels and infrared bands of this species; those of O₃, H₂O, N₂, NO, CO, and OH; and, more recently, those of CH₄, NO₂, N₂O, HNO₃, O₂, and HCN.

Non-LTE Populations

The non-LTE populations are normally used to calculate the emitted radiances or absorption spectra measured by remote sensing instruments, using a further radiative transfer code often referred to as the 'forward model' to distinguish it from the inverse procedure that is used to work back from radiances or spectra to retrieve the atmospheric variables of interest.

CO₂ is an example of an important atmospheric molecule that has many vibrational levels and whose populations are interdependent over extended atmospheric regions and originate many important transitions at 15, 10, 4.3, and 2.7 μ m. The first vibrational levels of N₂ and O₂ also need to be considered with those of CO₂ because, due to their resonances with the CO₂($v_1, v_2, 1$), H₂O(0, $v_2, 0$), and CH₄(v_2, v_4) levels, they play an important role in redistributing the excitation energy between these states via V-V collisional transfer (see Figure S1). The solution of this system requires, then, a complex and interactive model. Non-LTE populations are normally given in the form of vibrational temperatures, which are easier to understand physically, have a quick grasp of their departure from LTE, and, as population ratios relative to LTE, are very useful for radiance computations (see eqn [25]).

The variability of non-LTE populations in the atmosphere depends principally on the kinetic temperature structure and illumination conditions. Thus, the departure from LTE as well as the vibrational temperatures of the CO₂ energy levels vary significantly with the latitude, season, and illumination conditions of the atmosphere (see Figures S2 and S3). Non-LTE population climatologies for the energy levels of other important atmospheric species (e.g., H₂O, O₃, N₂O, CO, CH₄, O₂, NO, NO₂, HNO₃, OH, N₂, and HCN) are nowadays also available (see Figures S4–S27).

Cooling and Heating Rates

The reduction in collisional excitation that occurs in non-LTE situations leads to less efficient transfer between thermal kinetic energy and radiation, and then generally reduces the rate at which the atmosphere cools to space. Less obviously, it can affect the way in which the energy from the Sun is absorbed as well: the lack of frequent collisions prevents the energy absorbed from the solar radiation from being efficiently thermalized, and so influences the net heating rates. A quantitative understanding of non-LTE is therefore of value not only to improve our knowledge of processes in the upper atmosphere and to derive accurate temperature and constituent abundances from remote sensing measurements, but also to understand the energy budget of the region.

The non-LTE effect on cooling can be easily illustrated for the typical situation in the atmosphere where a band is optically thin and its upper level is excited only by thermal collisions and deexcited by spontaneous radiation (see eqn [16]). Then, the non-LTE and LTE cooling rates, $q = -h$, are related, from eqns [13] and [16] and eqn [1] for n_2 in LTE, by

$$\frac{q}{q^{\text{LTE}}} = \frac{k_t[M]}{A_{21}} \quad [21]$$

Since k_t and A_{21} are nearly constant with height while the number density of the air molecule $[M]$ acting as a collision partner falls off, the cooling rate ratio rapidly decreases with altitude. Figure 3 illustrates this situation for the CO₂ 15 μm fundamental band. Although of much smaller significance, the LTE cooling is, however, smaller in the transition region. This is due to the larger absorption of the emission coming from the adjacent warmer layers above which partially mitigates the cooling.

The bands contributing most to cooling in the infrared are, in order of importance, the CO₂ emissions at 15 μm , the O₃ bands near 9.6 μm , and the far-infrared pure rotational and 6.3 μm H₂O bands. NO 5.3 μm emission is dominant in the thermosphere, and the O(³P) 63 μm emission is also important in this region, above around 200 km. The CO₂ 15 μm and NO 5.3 μm in particular, are in non-LTE in the middle and upper atmosphere.

The effect of non-LTE on the heating is mainly through the resonant scattering of incoming solar radiation, principally in the CO₂ bands near 2.7 and 4.3 μm , and through fluorescence effects in which energy is absorbed and later reradiated, usually at a different wavelength. The CO₂ near-infrared (near-IR) bands' heating is important because it peaks at altitudes where absorption of solar radiation by O₂ and O₃ is very small.

Neglecting the CO₂ heating can lead to mean upper mesospheric temperatures about 2–8 K cooler.

The general circulation models require fast and accurate parameterizations of the cooling and heating rates, since the large computer time consumed by the full non-LTE models cannot be afforded. Several parameterizations have been developed, mainly for the Earth's and Mars' atmospheres. In these, it is common to divide the atmosphere into regions where approximations are accurate, such as the LTE region; the uppermost non-LTE region, where the cool-to-space approximation is accurate; and one or two intermediate regions, where other non-LTE approximations are made. Another factor is the number of bands contributing to the cooling. In LTE, all of them are grouped, since they have the same source function, $B(T)$. However, in the non-LTE region, each band has a different source function, usually with different transition altitudes from non-LTE to LTE. This additionally complicates the calculations in those intermediate regions, where the parameterizations are less accurate. More recently, reduced non-LTE models have been devised, limiting the inversion of the radiative transfer and statistical equilibrium equations at those intermediate regions, and they apply the LTE and cool-to-space approximations to the lowest and highest regions, respectively.

Remote Sensing of the Non-LTE Atmosphere

The first experimental studies of non-LTE in the atmosphere were made in the 1960s with infrared measurements from rocket payloads. They assumed an increased importance in the 1970s, when non-LTE radiance measurements began to be made by instruments on orbiting satellites. When the nature and the degree of the non-LTE processes affecting the observed quantity (usually thermally emitted radiance or atmospheric transmittance) are unknown, the measurements can be used to study the process in detail, and to validate theoretical models such as those described in this article. However, it is increasingly the case that the effect of non-LTE processes on the observed quantity is reasonably well understood, and the measurements are being used to determine temperature or composition. Then, the presence of non-LTE effects (even if the underlying theory is well understood) will increase the complexity of the interpretation of the data, and add an extra source of error and uncertainty. Sometimes, however, non-LTE can help by increasing the signal at high altitudes, especially when there is significant solar pumping, as is the case for CO at 4.7 μm and CO₂ at 4.3 μm , for example. The vertical coverage possible with an instrument of a given sensitivity can be greatly extended by this effect, although obviously only on the sunlit side of the planet.

Consider the radiance measured by a satellite instrument with a response function $F(\nu)$ over a finite frequency range $\Delta\nu$ at a given observation point x_{obs} . This is given for the case of limb viewing by

$$L(x_{\text{obs}}) = \int_{\Delta\nu} F(\nu) \int_{x_{\infty}}^{x_{\text{obs}}} J_{\nu}(x) \frac{d\mathcal{T}_{\nu}(x)}{dx} dx d\nu \quad [22]$$

where x is a position along the limb path line-of-sight and the integration is performed from the furthest extent of the atmosphere at the limb point, x_∞ , to the observation point at x_{obs} , passing through the tangent height of the observation point, z , and the monochromatic transmittance between x and x_{obs} is given by

$$\mathcal{T}_\nu(x, x_{\text{obs}}) = \exp\left(-\int_x^{x_{\text{obs}}} k_\nu(x') n_a(x') dx'\right) \quad [23]$$

It is useful to relate those equations to the case of LTE, which is simply done by replacing the expressions for the source function and the absorption coefficient, given by

$$\frac{J_\nu}{B_\nu} = \frac{n_2}{\bar{n}_2} \frac{k_\nu^{\text{LTE}}}{k_\nu} \quad [24]$$

and

$$\frac{k_\nu(x)}{k_\nu^{\text{LTE}}(x)} = \frac{n_1}{\bar{n}_1} \left[1 - \frac{g_1 n_2}{g_2 n_1}\right] \left[1 - \exp\left(-\frac{h\nu_0}{kT}\right)\right]^{-1} \quad [25]$$

Therefore, the emission measurements are affected by non-LTE mainly through the source function and, to a lesser extent, through the absorption coefficient (see eqns [24] and [25]). Thus, non-LTE is very important for emission measurements, particularly those measured with a limb geometry, because the relative contribution from the upper atmospheric layers, where non-LTE is more common, is larger. The non-LTE effects in nadir radiances (measured viewing near the vertical) are of less significance, since most of the radiation comes from the dense troposphere and lower stratosphere, where most emissions are in LTE. An exception is the CO_2 4.3 μm bands at daytime and NO at 5.3 μm in the thermosphere.

The effect of non-LTE on absorption spectra, where

$$L_\nu(x_{\text{obs}}) = F(\nu) L_\nu(x_\infty) \mathcal{T}_\nu(x_\infty, x_{\text{obs}}) \quad [26]$$

is much less than for emission, since it enters into the transmission function only through the absorption coefficient, and not through the source function.

In particular, non-LTE is negligible for the fundamental bands, most commonly used in absorption experiments, since the population of the ground state is not significantly affected by non-LTE (only through the vibrational partition function). Also, since most of the molecules are in the ground state at atmospheric temperatures, absorption in the hot bands (those whose lower states are not the ground state) is very weak, particularly in the tenuous upper regions where non-LTE dominates. There is, however, an important exception: the lowest vibrational level of CO_2 ($0,1^1,0$), which is responsible for the 15 μm emission. The population of this level in the atmosphere can be derived from high spectral resolution absorption measurements in the ($0,1^1,1 \leftarrow 0,1^1,0$) 4.3 μm band. From this, and using also the fundamental ($0,0,1 \leftarrow 0,0,0$) 4.3 μm band, the population of $\text{CO}_2(0,1^1,0)$, the CO_2 abundance, and the kinetic temperature (from the rotational distribution) can be retrieved simultaneously. This offers an important advantage to the analysis of emission spectra, where a simultaneous measurement of the kinetic temperature and the abundance

of the species considered in the ground and in the excited states is rarely available.

Analysis of Non-LTE Emissions

Here, we describe how atmospheric measurements have been used to improve our knowledge of non-LTE in the atmosphere. Rocket experiments in the 1970s such as the Spectral Infrared Rocket experiment (SPIRE) or the Middle Atmosphere Program/Winter in Northern Europe (MAP/WINE) campaign allowed for the first quantitative analysis of infrared non-LTE emissions (mainly in the 15 and 4.3 μm regions). Later on, instrumentation flown on the space shuttle (e.g., the Cryogenic Infrared Spectrometers and Telescopes for the Atmosphere (CRISTA) and the Cryogenic Infrared Radiance Instrumentation for Shuttle (CIRRIS)) led to important advances thanks to the increased amount of observed data. In particular, CIRRIS-1A provided valuable information on thermospheric NO emissions at 5.3 μm that allowed the study of the breakdown of rotational LTE and chemical excitations of NO due to the reaction of atomic nitrogen with molecular oxygen.

The first satellite observations of non-LTE emissions were taken by the Limb Infrared Monitor of the Stratosphere (LIMS) and the Stratospheric and Mesospheric Sounder (SAMS) on NIMBUS 7, launched in 1978. The measurements by Cryogenic Limb Array Etalon Spectrometer (CLAES) and the Improved Stratospheric and Mesospheric Sounder (ISAMS) radiometer on the Upper Atmosphere Research Satellite (UARS), taken in the 1990s, showed systematic day–night differences in several infrared channels that are attributable to non-LTE effects. For example, the daytime signal of the ISAMS 6.3 μm band of water vapor at 70 km observed at high effective spectral resolution is 2–3 times larger than that at nighttime; this factor is too large to be accounted for by any reasonable change in the concentration of water vapor. CLAES provided experimental evidence for the excitation of the CO_2 ν_3 mode by energy transfer from electronically excited $\text{O}(^1D)$ through $\text{N}_2(1)$.

The Michelson Interferometer for Passive Atmospheric Sounding (MIPAS) on Envisat, operating during 2002–12, was a very suitable instrument for studying atmospheric non-LTE emissions because of its wide spectral coverage, allowing for measuring emissions from a given compound in different spectral regions; its high spectral resolution, enabling the discrimination of emissions from different species and between different bands of the same species; and its high sensitivity, allowing measurement of the emission in the upper atmosphere where non-LTE emissions are most important. All this has led MIPAS to provide us with significant improvements in our knowledge of the non-LTE atmospheric infrared emissions of many species such as H_2O , CH_4 , CO_2 , O_3 , NO_2 , NO , and N_2O . For example, MIPAS has provided the first evidence of $\text{CH}_4(\nu_4)$ non-LTE emissions in the mesosphere. Its high spectral resolution has also allowed researchers to derive new collisional rates between the CO_2 02^01 , 02^21 , and 10^01 states that emit the 4.3 μm second hot bands (see Figure 4).

MIPAS has taken the most unequivocal spectra of H_2O in the mesosphere, clearly demonstrating the distinct non-LTE daytime emissions from the fundamental and hot bands at 6.3 μm . Furthermore, it has provided the most accurate

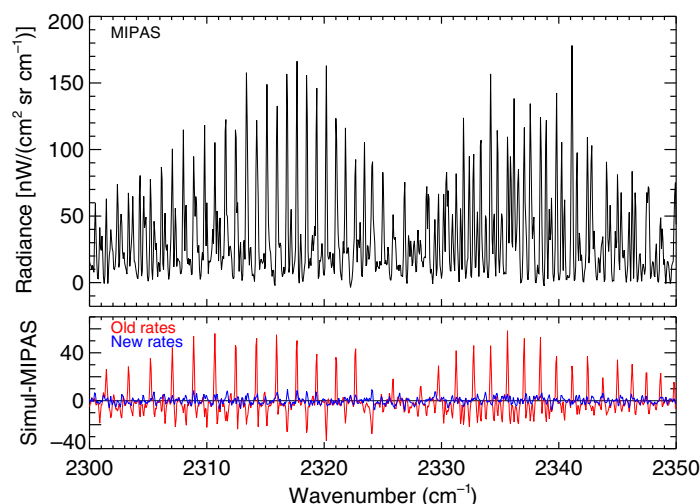


Figure 4 High-resolution limb spectra of the Earth's atmosphere near $4.3\ \mu\text{m}$ at a tangent height of 70 km. The upper panel shows the spectrum taken by MIPAS on 1 January 2009 at (24° S , 60° E) and a solar zenith angle of 62.7° . The lower panel shows the residuals (the residuals are the differences of the MIPAS measurements – the calculations) for two non-LTE simulations performed with the 'standard' collisional rates for CO_2 (red) and those derived to fit the measured spectra (blue).

collisional rate for the relaxation of $\text{H}_2\text{O}(020)$ by O_2 . In addition, MIPAS has given the first experimental confirmation of the non-LTE processes exciting the $\text{NO}(1)$ level, emitting near $5.3\ \mu\text{m}$, in the daytime stratosphere. Measurements of the NO rotational and spin transitions in the $5.3\ \mu\text{m}$ emission in the thermosphere, both under quiescent and geomagnetically perturbed conditions, have confirmed theoretical calculations of NO rotational and spin non-LTE populations. MIPAS has also made it possible to throw light on the puzzle about whether or not the $\text{NO}_2(\nu_3)$ levels are non-LTE excited in the daytime stratosphere. A recent analysis has shown that the photochemical excitation of $\text{NO}_2(\nu_3 = 1-4)$ levels is a factor of 50 smaller than previously thought, leading to small non-LTE deviations for these levels. Finally, measurements of the non-LTE emissions of the carbon monoxide first hot band, $\text{CO}(2-1)$, near $4.7\ \mu\text{m}$ and of $\text{N}_2\text{O}(001)$ near $4.5\ \mu\text{m}$ have also been observed by MIPAS for the first time.

The still-operational Sounding of the Atmosphere Using Broadband Emission Radiometry (SABER) experiment on the NASA TIMED mission has allowed researchers to constrain the $\text{OH}-\text{N}_2-\text{CO}_2$ vibration-vibration energy transfer, an important driver of the nighttime non-LTE enhancement of CO_2 $4.3\ \mu\text{m}$ emissions, thanks to its simultaneous observation of CO_2 and OH emissions at $4.3\ \mu\text{m}$ and $1.6-2.0\ \mu\text{m}$, respectively.

As mentioned in this article, most of the measurements used to study non-LTE have been taken in emission. An exception is the use of solar occultation spectroscopy by the Atmospheric Trace Molecule Spectroscopy (ATMOS) spectrometer to study the non-LTE population of $\text{CO}_2(0,1^1,0)$.

Non-LTE Retrievals

The improved knowledge of non-LTE processes gained from experimental studies in conjunction with the rapidly growing computing power of modern data-processing facilities are the key for the operational retrieval of temperature and atmospheric composition from space-borne non-LTE emission measurements, and hence for the extension of the remotely

sounded altitude region toward the mesosphere and lower thermosphere. Non-LTE retrieval algorithms require the expensive computation of the non-LTE populations of the sounded infrared emitters under the actual atmospheric conditions. Since the populations typically depend on the retrieval parameters (temperature or composition), these computations have to be repeated in each step of the iterative inversion procedure.

Non-LTE retrieval schemes have been successfully applied to the data processing of space-borne atmospheric remote sensing experiments such as ISAMS on UARS, the CRISTA experiment flown on the space shuttle, and the recent MIPAS-Envisat and SABER-TIMED instruments. These non-LTE measurements have greatly improved our knowledge of the temperature structure and distributions of CO , CO_2 , O_3 , H_2O , OH , NO , and NO_2 in the middle and upper atmosphere. Figure 5 shows some examples of quantities retrieved under non-LTE from MIPAS spectra.

Non-LTE in Planetary Atmospheres

Non-LTE situations are not limited to the Earth's atmosphere. The atmospheres of the other terrestrial planets, Mars and Venus, consist of nearly pure carbon dioxide, making the study of non-LTE processes in that molecule of crucial importance for understanding the thermal structure of their upper atmospheres. Also, the different physical and chemical conditions in these atmospheres relative to the Earth offer scenarios where different combinations of processes may be at work, and comparison with the terrestrial case can provide deeper insights into the non-LTE problem. For example, the weak hot and isotopic bands of CO_2 are much more important for radiative cooling in the Venusian atmosphere than on the Earth, and LTE applies up to much higher altitudes in the $15\ \mu\text{m}$ bands. The larger concentration of CO_2 means that these bands are optically thicker at a given pressure level, less radiative energy is lost, and a slower thermal collisional rate is enough to keep their populations in LTE. Another interesting example is the

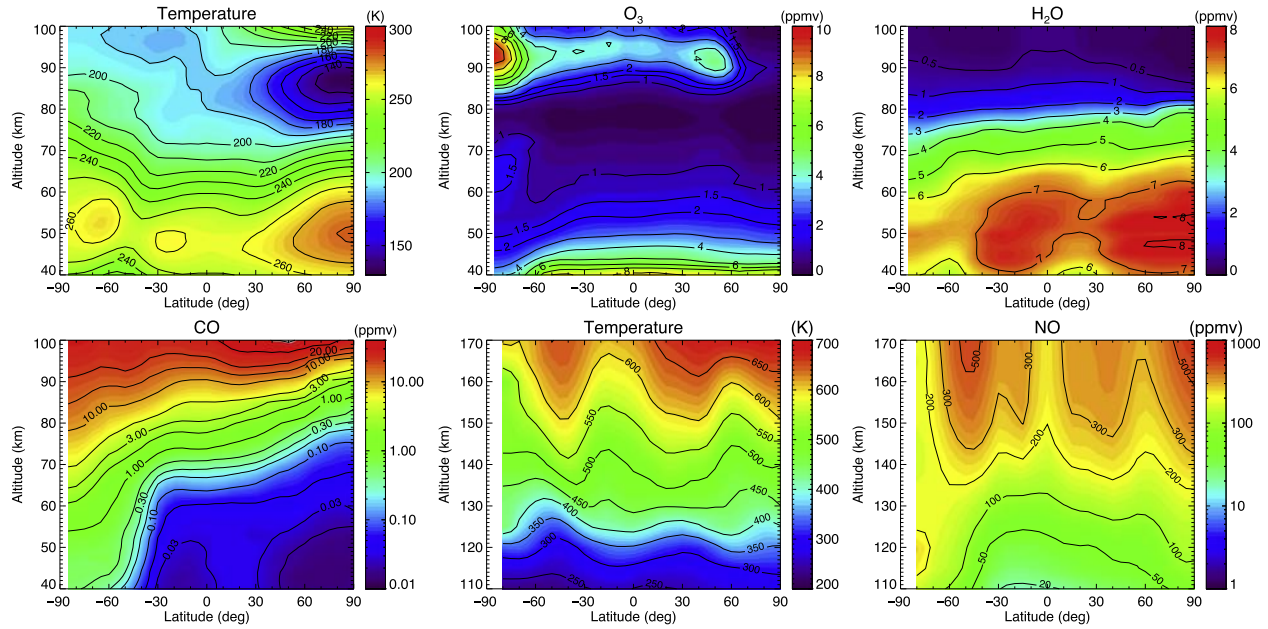


Figure 5 Monthly zonal mean of temperature, O₃, H₂O, and CO volume-mixing ratios (VMRs) in the middle atmosphere, as well as daytime temperature and NO VMR in the thermosphere retrieved considering non-LTE from MIPAS spectra taken on July 2009.

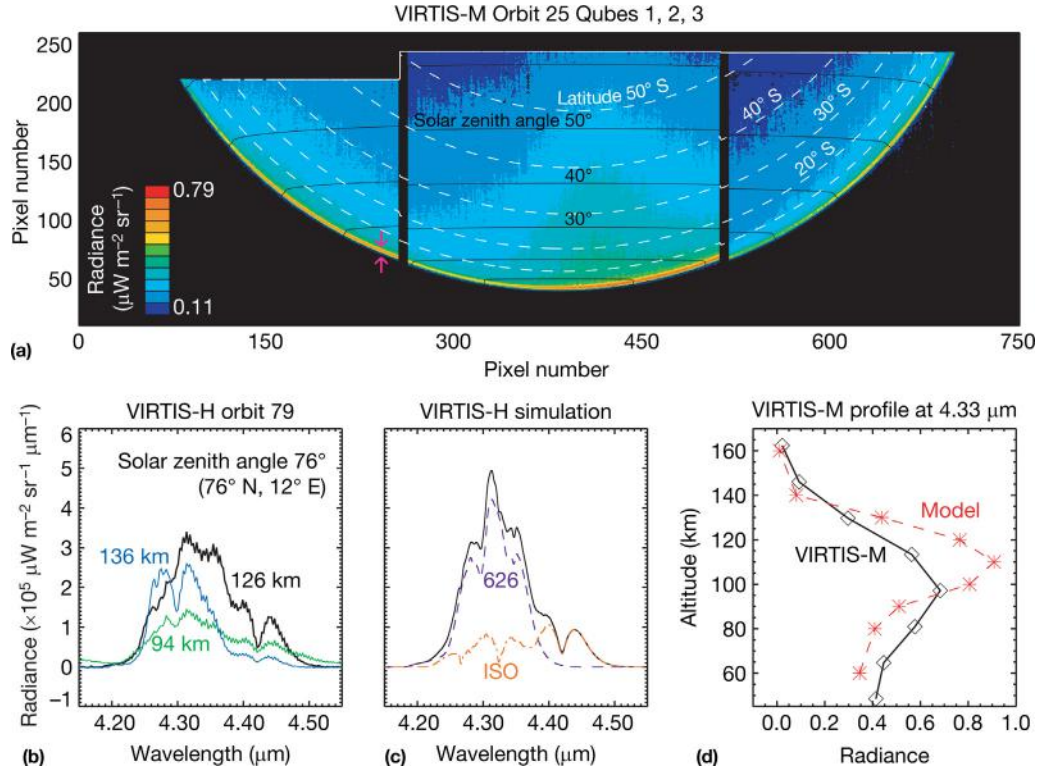


Figure 6 CO₂ non-LTE emission near 4.3 μm on the dayside of Venus. (a) Limb and nadir measurements taken by VIRTIS-M at 4.33 μm. (b) VIRTIS-H spectra taken in the limb at several tangent heights. The peaks correspond to different vibrational bands of CO₂. (c) Simulated non-LTE limb spectra for the most abundant (626) and minor isotopes (ISO) of CO₂ at a tangent height of 126 km. (d) Limb emission profiles at the location shown by the red arrows in (a). After Drossart, P., Piccioni, G., Gérard, J.C., López-Valverde, M.A., Sanchez-Lavega, A., Zasova, L., et al., 2007. A dynamic upper atmosphere of Venus as revealed by VIRTIS on Venus Express. *Nature* 450, 641.

appearance of the natural laser emission in the 10 μm bands of CO_2 in the Martian atmosphere, which is absent on Earth because of the lower relative abundance of CO_2 , and is weak in Venus because of its warmer atmosphere.

Differences in the compositions of the terrestrial planet atmospheres are important in other ways. The major atmospheric constituent on Earth, N_2 , acts as an intermediary through its first vibrationally excited state to redistribute the energy of CO_2 ν_3 levels excited by absorption of solar radiation. The second most abundant species, O_2 , plays a similar role for distributing the vibrational energy of water vapor and methane. On Mars and Venus, it is the CO_2 itself that, through much faster near-resonant vibrational–vibrational energy exchange, distributes the absorbed solar energy internally and makes efficient thermal cooling possible at 15 μm . All studies of the cooling and heating rates in the Venusian and Martian atmospheres have stressed the important role of atomic oxygen: the thermal collisions between CO_2 and $\text{O}(^3\text{P})$ are crucial for the cooling of the upper regions of a CO_2 atmosphere.

For Venus and Mars, only a few measurements were available until the beginning of the twenty-first century. A limited amount of spectra were taken on the dayside limb of Venus by NIMS–Galileo during its fast flyby of Venus in 1991. Recently, however, the PFS and OMEGA instruments on Mars Express, launched in 2003, and especially VIRTIS on Venus Express, launched in 2005 (see Figure 6), are supplying the first systematic sounding (nadir and limb) of these atmospheres in the mid-infrared (mid-IR).

The atmospheres of the giant planets and some of their satellites, especially Titan, contain radiatively active gases such as methane, hydrogen cyanide, carbon monoxide, and various other hydrocarbons. These gases play an important role in the thermal balance of their upper atmospheres, analogous to that of CO_2 on Earth. Thus, non-LTE models for planetary atmospheres other than Earth were originally developed to study energy balance problems. During the last decade, however, the availability of the European Space Agency's Infrared Space Observatory–Short Wavelength Spectrometer (ISO–SWS) observations, the high spectral resolution ground-based measurements in the near-IR, and the measurements of the Visible and Infrared Mapping Spectrometer (VIMS) instrument onboard Cassini have triggered modeling efforts for deriving temperature and composition from these atmospheres. That includes non-LTE modeling for the vibrational levels of CH_4 , CO , HCN , C_2H_2 , C_2H_6 , CH_3D , PH_3 , and NH_3 in the mid- and near-IR.

In the very tenuous atmospheres, consisting of water vapor, carbon monoxide, methane, and other gases, which make up the comae of comets, the pressures are so low that LTE never applies anywhere. In this case, not only vibrational but also rotational non-LTE have to be considered in the analysis of composition measurements, even those made in the microwave part of the spectrum.

See also: **Chemistry of the Atmosphere:** Laboratory Kinetics; Observations for Chemistry (Remote Sensing): IR/FIR (Satellite, Balloon and Ground). **Climate and Climate Change:** Carbon Dioxide. **Global Change:** Upper Atmospheric Change. **Mesosphere:** Atomic Species in the Mesopause Region.

Numerical Models: General Circulation Models. **Observations Platforms:** Rockets. **Ozone Depletion and Related Topics:** Photochemistry of Ozone. **Radiation Transfer in the Atmosphere:** Absorption and Thermal Emission. **Satellites and Satellite Remote Sensing:** Research; Temperature Soundings. **Solar System/Sun, Atmospheres, Evolution of Atmospheres:** Planetary Atmospheres: Mars; Planetary Atmospheres: Venus. Thermosphere.

Appendix A. Supplementary data

Supplementary data related to this article can be found at <http://dx.doi.org/10.1016/B978-0-12-382225-3.00339-X>.

Further Reading

- Andrews, D.G., Holton, J.R., Leovy, C.B., 1987. *Middle Atmosphere Dynamics* (Chapter 2). Academic Press, London.
- Bougher, S.W., Hunten, D.M., Roble, R.G., 1994. CO_2 cooling in the terrestrial planet thermospheres. *Journal of Geophysical Research* 99, 14609.
- Curtis, A.R., Goody, R.M., 1956. Thermal variation in the upper atmosphere. *Proceedings of the Royal Society of London A236*, 193.
- Drossart, P., 2005. Infrared spectroscopy of planetary atmospheres. *C. R. Physique* 6, 817–824.
- Drossart, P., Piccioni, G., Gérard, J.C., López-Valverde, M.A., Sanchez-Lavega, A., Zasova, L., et al., 2007. A dynamic upper atmosphere of Venus as revealed by VIRTIS on Venus Express. *Nature* 450, 641.
- Feofilov, A., Kutepov, A., 2012. Infrared radiation in the mesosphere and lower thermosphere: energetic effects and remote sensing. *Surveys in Geophysics* 33, 1231–1280.
- Fomichev, V.I., 2009. The radiative energy budget of the middle atmosphere and its parameterization in general circulation models. *Journal of Atmospheric and Solar–Terrestrial Physics* 71, 1577–1585.
- Funke, B., López-Puertas, M., Stiller, G.P., Clarmann, von, T., Höpfner, M., 2001. A new non-LTE retrieval method for atmospheric parameters from MIPAS–Envisat emission spectra. *Advances in Space Research* 27 (6–7), 1099–1104.
- Funke, B., López-Puertas, M., García-Comas, M., et al., 2012. GRANADA: a generic radiative transfer and non-LTE population algorithm. *Journal of Quantitative Spectroscopy and Radiative Transfer* 113, 1771–1817.
- García-Comas, M., Funke, B., López-Puertas, M., Bermejo-Pantaleón, D., Glatthof, N., et al., 2012. On the quality of MIPAS kinetic temperature in the middle atmosphere. *Atmospheric Chemistry and Physics* 12 (13), 6009–6039.
- Goody, R.M., Yung, Y.L., 1989. *Atmospheric Radiation: Theoretical Basis*. Oxford University Press, Oxford.
- López-Puertas, M., Taylor, F.W., 2001. *Non-LTE Radiative Transfer in the Atmosphere*. World Scientific Publishing Company Pte, Washington, DC.
- López-Puertas, M., Funke, B., Gil-Lopez, S., López-Valverde, M.A., Clarmann, von, T., et al., 2005. Atmospheric non-local thermodynamic equilibrium emissions as observed by the Michelson Interferometer for Passive Atmospheric Sounding (MIPAS). *C. R. Physique* 6, 848–863.
- López-Valverde, M.A., López-Puertas, M., Funke, B., Gilli, G., García-Comas, M., et al., 2011. Modeling the atmospheric limb emission of CO_2 at 4.3 μm in the terrestrial planets. *Planetary and Space Science* 59 (10), 988–998.
- Milne, E.A., 1930. Thermodynamics of stars. In: *Handbuch der Astrophysik*, vol. 3, Part I. (Chapter 2) 65, Springer, Heidelberg.

Relevant Websites

- <http://www.cfa.harvard.edu/hitran/>.
- <http://www.pole-ether.fr/etherTypo/?id=950&L=0>.
- <http://www.iaa.es/wpuertas/granada.html>.

Scattering

M Mishchenko, L Travis, and A Lacis, Goddard Institute for Space Studies, New York, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

Sunlight illuminating the Earth's atmosphere is scattered by gas molecules and suspended particles, giving rise to blue skies, white clouds, and optical displays such as rainbows and halos. By scattering and absorbing the shortwave solar radiation and the longwave radiation emitted by the underlying surface, cloud and aerosol particles strongly affect the radiation budget of the terrestrial climate system. As a consequence of the dependence of scattering characteristics on particle size, morphology, and composition, scattered light can be remarkably rich in information on particle properties and thus provides a sensitive tool for remote retrievals of macro- and microphysical parameters of clouds and aerosols.

Introduction

A parallel beam of light, or of any electromagnetic radiation, propagates in a vacuum without a change in its characteristics. However, interposing a particle into the beam causes two fundamental effects. First, the particle may convert some of the energy contained in the electromagnetic field into other forms of energy such as heat. This phenomenon is called absorption. Second, the directional propagation of electromagnetic energy and its polarization state get modified. This phenomenon is called scattering.

The scattering and absorption characteristics of an isolated particle are often complex functions of the particle's size, morphology, and composition. They can be determined by obtaining a numerically exact solution of the Maxwell equations or by using a suitable experimental technique. Direct computer solutions of the Maxwell equations become much more involved and are often impracticable for a compound object in the form of a cloud of particles. In such cases, one has to use a physically based asymptotic solution of the Maxwell equations called the radiative transfer equation (RTE).

Sunlight incident on the Earth's atmosphere is scattered by gas molecules and suspended particles, giving rise to blue skies, white clouds, and various optical displays such as rainbows, halos, and the glory. By scattering and absorbing the shortwave solar radiation and the longwave radiation emitted by the underlying surface, cloud and aerosol particles strongly affect the radiation budget of the terrestrial climate system. As a consequence of the dependence of scattering characteristics on particle size, morphology, and composition, scattered light can be remarkably rich in implicit information on particle properties and thus provides a sensitive tool for remote retrievals of macro- and microphysical parameters of clouds and aerosols.

Electromagnetic Scattering by a Fixed Particle

To explain the fundamental concept of electromagnetic scattering by a fixed particle, let us assume that the electromagnetic field is time harmonic, which allows one to fully describe it at any moment in time everywhere in space as the solution of the frequency-domain Maxwell equations. Specifically, it is convenient to factor out the time-harmonic dependence of the

complex electric and magnetic fields: $\mathbf{E}(\mathbf{r}, t) = \exp(-i\omega t) \mathbf{E}(\mathbf{r})$ and $\mathbf{H}(\mathbf{r}, t) = \exp(-i\omega t) \mathbf{H}(\mathbf{r})$, where $\mathbf{r}$ is the position vector of the observation point, t is time, ω is the angular frequency, and $i = (-1)^{1/2}$. The actual electric and magnetic fields are obtained by taking the real part of the respective complex fields. The field amplitudes $\mathbf{E}(\mathbf{r})$ and $\mathbf{H}(\mathbf{r})$ can be found from the following curl equations:

$$\left. \begin{aligned} \nabla \times \mathbf{E}(\mathbf{r}) &= i\omega\mu_0 \mathbf{H}(\mathbf{r}) \\ \nabla \times \mathbf{H}(\mathbf{r}) &= -i\omega\epsilon_1 \mathbf{E}(\mathbf{r}) \end{aligned} \right\} \text{inside } V_{\text{EXT}}, \quad [1]$$

$$\left. \begin{aligned} \nabla \times \mathbf{E}(\mathbf{r}) &= i\omega\mu_0 \mathbf{H}(\mathbf{r}) \\ \nabla \times \mathbf{H}(\mathbf{r}) &= -i\omega\epsilon_2(\mathbf{r}, \omega) \mathbf{E}(\mathbf{r}) \end{aligned} \right\} \text{inside } V_{\text{INT}}. \quad [2]$$

Here, V_{INT} is the cumulative 'interior' volume occupied by the scattering particle; V_{EXT} is the infinite exterior region, which is assumed to be homogeneous, linear, isotropic, and non-absorbing; the host medium and the particle are assumed to be nonmagnetic; μ_0 is the permeability of a vacuum; ϵ_1 is the real-valued electric permittivity of the host medium; and $\epsilon_2(\mathbf{r}, \omega)$ is the complex permittivity of the particle. Since the first relations in eqns [1] and [2] yield the magnetic field provided that the electric field is known everywhere, the solution of the Maxwell equations is usually sought in terms of only the electric field. To have a unique solution, eqns [1] and [2] must be supplemented by appropriate boundary conditions at the particle surface as well as by the so-called radiation conditions at infinity.

Note that although the amplitudes $\mathbf{E}(\mathbf{r})$ and $\mathbf{H}(\mathbf{r})$ do not depend on time explicitly, they can fluctuate randomly if the electromagnetic field is quasimonochromatic. However, such fluctuations are assumed to occur much more slowly than the time-harmonic oscillations described by the factor $\exp(-i\omega t)$, which justifies the use of the frequency-domain Maxwell equations at any given moment.

Let us now assume that in the absence of the particle, the electromagnetic field is given by the simplest solution of the Maxwell equations in the form of a plane electromagnetic wave propagating in the direction of the wave vector $\mathbf{k}^{\text{inc}}$:

$$\mathbf{E}^{\text{inc}}(\mathbf{r}) = \mathbf{E}_0^{\text{inc}} \exp(i\mathbf{k}^{\text{inc}} \cdot \mathbf{r}) \text{ everywhere in space.} \quad [3]$$

As shown schematically in Figure 1(a), eqn [3] represents the transport of electromagnetic energy from one point to another in the absence of the particle and embodies the concept of

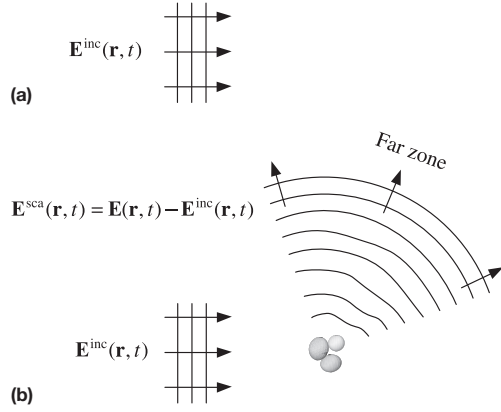


Figure 1 Electromagnetic scattering by a fixed particle. In this case, the particle is an aggregate consisting of three monomers in contact.

a perfectly monochromatic parallel beam of light. The presence of the particle modifies the electromagnetic field that would exist otherwise. It is this modification that is called *electromagnetic scattering*.

It is customary to call the difference between the total field in the presence of the particle (i.e., $E(\mathbf{r})$) and the original field that would exist in the absence of the object (i.e., $E^{\text{inc}}(\mathbf{r})$) the ‘scattered field’ and denote it as $E^{\text{sca}}(\mathbf{r})$ (see **Figure 1(b)**). Thus, the total field in the presence of the particle is intentionally represented as the sum of the respective incident (original) and scattered fields:

$$E(\mathbf{r}) = E^{\text{inc}}(\mathbf{r}) + E^{\text{sca}}(\mathbf{r}). \quad [4]$$

Of course, one can think of incident fields other than a plane wave and thereby generalize the concept of scattering. In this regard, an especially convenient framework is provided by the so-called volume integral equation which follows from the Maxwell equations and incorporates the boundary and radiation conditions:

$$E(\mathbf{r}) = E^{\text{inc}}(\mathbf{r}) + k^2 \int_{V_{\text{INT}}} d\mathbf{r}' \vec{G}(\mathbf{r}, \mathbf{r}') \cdot E(\mathbf{r}') [m^2(\mathbf{r}') - 1], \quad [5]$$

where $m(\mathbf{r}') = [\epsilon_2(\mathbf{r}', \omega)/\epsilon_1]^{1/2}$ is the refractive index of the interior relative to that of the host exterior medium; $k = |\mathbf{k}^{\text{inc}}| = \omega(\epsilon_1 \mu_0)^{1/2} = 2\pi/\lambda$ is the wave number; λ is the wavelength; $\vec{G}(\mathbf{r}, \mathbf{r}')$ is the free-space dyadic Green’s function; and $E^{\text{inc}}(\mathbf{r})$ is any physically realizable solution of the Maxwell equations for an infinite homogeneous medium. One can see that eqn [5] expresses the total field everywhere in space in terms of the total internal field. The latter is not known in general and must be found by solving eqn [5] either analytically or numerically.

Far-Field Scattering

Equation [5] can be used to show that at a distance from the particle greatly exceeding its size, in the so-called far zone, the scattered field becomes an outgoing spherical wave (**Figure 1(b)**). By placing the origin O of the spherical

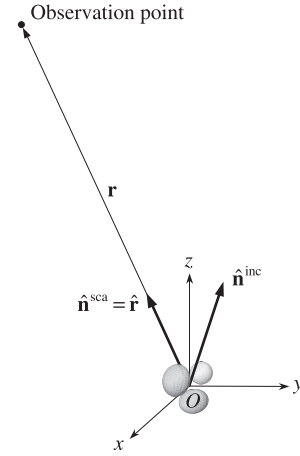


Figure 2 Scattering in the far zone of the particle.

coordinate system at the geometrical center of the particle (**Figure 2**), we have:

$$E^{\text{sca}}(\mathbf{r}) \xrightarrow{r \rightarrow \infty} \frac{\exp(ikr)}{r} E_1^{\text{sca}}(\hat{\mathbf{n}}^{\text{sca}}), \quad [6]$$

where $r = |\mathbf{r}|$ is the distance from the origin, $\hat{\mathbf{n}}^{\text{inc}} = \mathbf{k}^{\text{inc}}/k$ is a unit vector in the incidence direction, and $\hat{\mathbf{n}}^{\text{sca}} = \mathbf{r}/r$ is a unit vector in the scattering direction.

The electric and magnetic field vectors of the scattered spherical wave are perpendicular to the scattering direction, while those of the incident plane wave are normal to the incidence direction. It is therefore convenient to denote by $\mathbf{E}$ a two-element column formed by the θ - and φ -components of either electric field vector:

$$\mathbf{E} = \begin{bmatrix} E_\theta \\ E_\varphi \end{bmatrix}. \quad [7]$$

As usual, the polar (zenith) angle $0 \leq \theta \leq \pi$ is measured from the positive z -axis, while the azimuth angle $0 \leq \varphi < \pi$ is measured from the positive x -axis in the clockwise direction when looking in the direction of the positive z -axis. The use of this notation allows us to write the scattered field as:

$$E^{\text{sca}}(r\hat{\mathbf{n}}^{\text{sca}}) = \frac{\exp(ikr)}{r} \mathbf{S}(\hat{\mathbf{n}}^{\text{sca}}, \hat{\mathbf{n}}^{\text{inc}}) \mathbf{E}_0^{\text{inc}}, \quad [8]$$

where $\mathbf{S}$ is the 2×2 so-called amplitude scattering matrix expressing the θ - and φ -components of the scattered spherical wave in those of the incident plane wave. This relation plays a key role in the theory of electromagnetic scattering.

Optical Observables

The typically high frequency of time-harmonic electromagnetic oscillations makes it virtually impossible to measure the electric and magnetic fields associated with the incident and scattered waves using traditional optical instruments. Therefore, in order to make the theory applicable to analyses of actual observations, the scattering phenomenon must be characterized in terms of derivative quantities that can be measured directly (i.e., optical *observables*). The conventional approach to

address this problem is to use four real-valued quantities I , Q , U , and V , which have the dimension of monochromatic energy flux (Wm^{-2}) and fully characterize a transverse electromagnetic wave inasmuch as it is subject to practical optical analysis. These quantities, called the Stokes parameters, are always defined with respect to a plane containing the direction of wave propagation, form the four-element Stokes column vector $\mathbf{I}$, and carry information about both the total intensity I and the polarization state of the wave. The Stokes parameters are intentionally defined such that the rapidly oscillating time-harmonic factor $\exp(-i\omega t)$ vanishes upon multiplication by its complex-conjugate counterpart: $\exp(-i\omega t)[\exp(-i\omega t)]^* \equiv 1$, where the asterisk denotes complex conjugation.

In the case of scattering in the far zone, both the incident plane wave and the outgoing scattered spherical wave are transverse. This allows one to define the corresponding sets of Stokes parameters:

$$\mathbf{I}^{\text{inc}} = \begin{bmatrix} I^{\text{inc}} \\ Q^{\text{inc}} \\ U^{\text{inc}} \\ V^{\text{inc}} \end{bmatrix} = \frac{1}{2} \sqrt{\frac{\epsilon_1}{\mu_0}} \begin{bmatrix} E_{0\theta}^{\text{inc}} (E_{0\theta}^{\text{inc}})^* + E_{0\phi}^{\text{inc}} (E_{0\phi}^{\text{inc}})^* \\ E_{0\theta}^{\text{inc}} (E_{0\theta}^{\text{inc}})^* - E_{0\phi}^{\text{inc}} (E_{0\phi}^{\text{inc}})^* \\ -E_{0\theta}^{\text{inc}} (E_{0\phi}^{\text{inc}})^* - E_{0\phi}^{\text{inc}} (E_{0\theta}^{\text{inc}})^* \\ i [E_{0\phi}^{\text{inc}} (E_{0\theta}^{\text{inc}})^* - E_{0\theta}^{\text{inc}} (E_{0\phi}^{\text{inc}})^*] \end{bmatrix}, \quad [9]$$

$$\mathbf{I}^{\text{sca}}(r\hat{\mathbf{n}}^{\text{sca}}) = \begin{bmatrix} I^{\text{sca}} \\ Q^{\text{sca}} \\ U^{\text{sca}} \\ V^{\text{sca}} \end{bmatrix} = \frac{1}{r^2} \frac{1}{2} \sqrt{\frac{\epsilon_1}{\mu_0}} \begin{bmatrix} E_{1\theta}^{\text{sca}} (E_{1\theta}^{\text{sca}})^* + E_{1\phi}^{\text{sca}} (E_{1\phi}^{\text{sca}})^* \\ E_{1\theta}^{\text{sca}} (E_{1\theta}^{\text{sca}})^* - E_{1\phi}^{\text{sca}} (E_{1\phi}^{\text{sca}})^* \\ -E_{1\theta}^{\text{sca}} (E_{1\phi}^{\text{sca}})^* - E_{1\phi}^{\text{sca}} (E_{1\theta}^{\text{sca}})^* \\ i [E_{1\phi}^{\text{sca}} (E_{1\theta}^{\text{sca}})^* - E_{1\theta}^{\text{sca}} (E_{1\phi}^{\text{sca}})^*] \end{bmatrix}. \quad [10]$$

Then the responses of well-collimated polarization-sensitive radiometers located in the far zone of the particle can be described in terms of the 4×4 phase and extinction matrices as follows.

In the absence of the particle (Figure 3(a)), radiometer 2 registers no signal, whereas radiometer 1 reacts to the incident plane wave:

$$\text{Signal 1} = \Delta S \mathbf{I}^{\text{inc}}, \quad [11]$$

$$\text{Signal 2} = 0, \quad [12]$$

where ΔS is the area of the objective lens. In the presence of the particle (Figure 3(b)), radiometer 2 reacts only to the scattered spherical wave, and its polarized reading is fully characterized by the product of the phase matrix $\mathbf{Z}$ and the Stokes column vector of the incident wave:

$$\begin{aligned} \text{Signal 2} &= \Delta S \mathbf{I}^{\text{sca}}(r\hat{\mathbf{n}}^{\text{sca}}) \\ &= \frac{\Delta S}{r^2} \mathbf{Z}(\hat{\mathbf{n}}^{\text{sca}}, \hat{\mathbf{n}}^{\text{inc}}) \mathbf{I}^{\text{inc}}, \quad \hat{\mathbf{n}}^{\text{sca}} \neq \hat{\mathbf{n}}^{\text{inc}}. \end{aligned} \quad [13]$$

The elements of the phase matrix have the dimension of area and are quadratic combinations of the elements of the

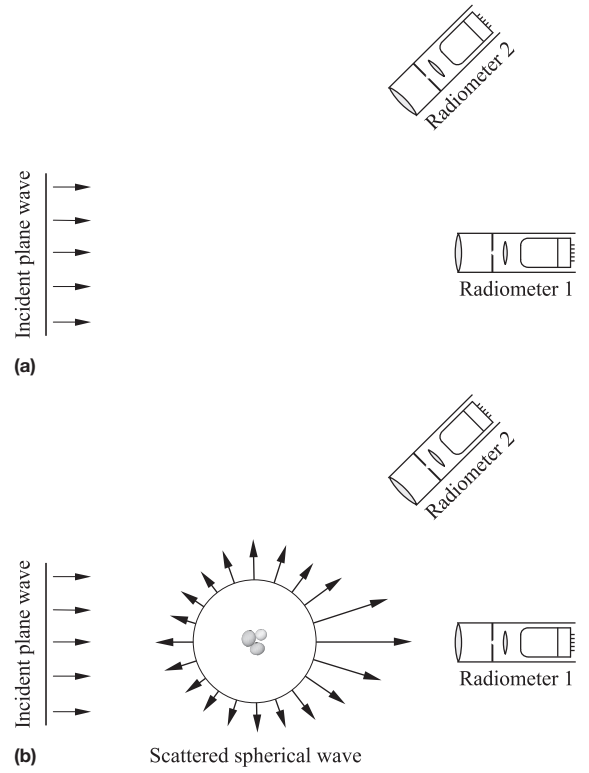


Figure 3 The readings of well-collimated polarization-sensitive radiometers in the presence of the particle differ from those in the absence of the particle.

amplitude scattering matrix $\mathbf{S}(\hat{\mathbf{n}}^{\text{sca}}, \hat{\mathbf{n}}^{\text{inc}})$. One can see that, in general, the phase matrix relates the Stokes parameters of the incident and scattered waves defined with respect to different reference planes: the meridional plane of the incidence direction $\hat{\mathbf{n}}^{\text{inc}}$ and that of the scattering direction $\hat{\mathbf{n}}^{\text{sca}}$, respectively.

Unlike radiometer 2, radiometer 1 in Figure 3(b) is facing the incident light, and, accordingly, its polarized reading consists of three parts:

1. the one due to the incident wave;
2. the one due to the forward-scattered wave; and
3. the one due to the interference of the incident wave and the wave scattered by the object in the exact forward direction:

$$\text{Signal 1} = \Delta S \mathbf{I}^{\text{inc}} + \frac{\Delta S}{r^2} \mathbf{Z}(\hat{\mathbf{n}}^{\text{inc}}, \hat{\mathbf{n}}^{\text{inc}}) \mathbf{I}^{\text{inc}} - \mathbf{K}(\hat{\mathbf{n}}^{\text{inc}}) \mathbf{I}^{\text{inc}}. \quad [14]$$

The third part is described by minus the product of the extinction matrix $\mathbf{K}$ and the Stokes column vector of the incident wave. The elements of the extinction matrix also have the dimension of area and are linear combinations of the elements of the forward-scattering amplitude matrix $\mathbf{S}(\hat{\mathbf{n}}^{\text{inc}}, \hat{\mathbf{n}}^{\text{inc}})$. Placing radiometer 1 sufficiently far from the particle makes the second term on the right-hand side of eqn [14] negligibly small.

In many respects, the measurement situation depicted in Figures 3(a) and 3(b) embodies the concept of electromagnetic scattering. Indeed, it demonstrates that in the absence of the particle, radiometer 2 would measure no signal, while the

signal measured by radiometer 1 allows one to measure the Stokes vector of the incident wave $\mathbf{I}^{\text{inc}}$. In the presence of the particle, the readings of both radiometers change. The reading of radiometer 2 is now proportional to the Stokes column vector of the scattered spherical wave, while the polarization signal measured by radiometer 1 is modified in two ways. First, the total measured electromagnetic power is attenuated as a combined result of the scattering of electromagnetic energy by the object in all directions and, possibly, the transformation of electromagnetic energy into other forms of energy (such as heat) inside the object. Second, the attenuation rates for the four Stokes components of the measured signal can, in general, be different. The latter effect is typical of objects lacking perfect spherical symmetry and is called dichroism.

Thus, to describe far-field scattering means, in effect, to quantify the differences between the readings of radiometers 1 and 2 in the presence of the object and in the absence of the object. This quantification can be fully achieved in terms of the phase and extinction matrices, which depend on object characteristics such as size, shape, refractive index, and orientation. Both matrices can be readily computed provided that the amplitude scattering matrix is already known.

In the case of quasimonochromatic fields, eqns [11]–[14] remain valid provided that now the Stokes column vectors of the incident and scattered fields are defined as averages of the right-hand sides of eqns [9] and [10] over a time interval much longer than the typical period of random fluctuations.

Derivative Quantities

There are several derivative quantities that are often used to describe various observable manifestations of electromagnetic scattering. The product of the extinction cross-section multiplied by the intensity of the incident plane wave yields the total attenuation of the electromagnetic power measured by radiometer 1 in Figure 3(b) owing to the presence of the particle. This means that the extinction cross-section depends, in general, on the polarization state and propagation direction of the incident wave and is given by:

$$C_{\text{ext}}(\hat{\mathbf{n}}^{\text{inc}}) = \frac{1}{I^{\text{inc}}} [K_{11}(\hat{\mathbf{n}}^{\text{inc}})I^{\text{inc}} + K_{12}(\hat{\mathbf{n}}^{\text{inc}})Q^{\text{inc}} + K_{13}(\hat{\mathbf{n}}^{\text{inc}})U^{\text{inc}} + K_{14}(\hat{\mathbf{n}}^{\text{inc}})V^{\text{inc}}]. \quad [15]$$

The product of the scattering cross-section multiplied by the intensity of the incident plane wave yields the total far-field power of the scattered wave:

$$C_{\text{sca}}(\hat{\mathbf{n}}^{\text{inc}}) = \frac{1}{I^{\text{inc}}} \int_{4\pi} d\hat{\mathbf{n}}^{\text{sca}} [Z_{11}(\hat{\mathbf{n}}^{\text{sca}}, \hat{\mathbf{n}}^{\text{inc}})I^{\text{inc}} + Z_{12}(\hat{\mathbf{n}}^{\text{sca}}, \hat{\mathbf{n}}^{\text{inc}})Q^{\text{inc}} + Z_{13}(\hat{\mathbf{n}}^{\text{sca}}, \hat{\mathbf{n}}^{\text{inc}})U^{\text{inc}} + Z_{14}(\hat{\mathbf{n}}^{\text{sca}}, \hat{\mathbf{n}}^{\text{inc}})V^{\text{inc}}]. \quad [16]$$

This implies that C_{sca} also depends on the polarization state as well as on the propagation direction of the incident wave. The absorption cross-section is defined as the difference between the extinction and scattering cross-sections:

$$C_{\text{abs}}(\hat{\mathbf{n}}^{\text{inc}}) = C_{\text{ext}}(\hat{\mathbf{n}}^{\text{inc}}) - C_{\text{sca}}(\hat{\mathbf{n}}^{\text{inc}}) \geq 0. \quad [17]$$

All optical cross-sections have the dimension of area. Finally, the dimensionless single-scattering albedo is defined as the ratio of the scattering and extinction cross-sections:

$$\tilde{\omega}(\hat{\mathbf{n}}^{\text{inc}}) = \frac{C_{\text{sca}}(\hat{\mathbf{n}}^{\text{inc}})}{C_{\text{ext}}(\hat{\mathbf{n}}^{\text{inc}})} \leq 1. \quad [18]$$

An important particular case of the phase matrix is the scattering matrix, defined by:

$$\mathbf{F}(\Theta) = \mathbf{Z}(\theta^{\text{sca}} = \Theta, \varphi^{\text{sca}} = 0; \theta^{\text{inc}} = 0, \varphi^{\text{inc}} = 0), \quad 0 \leq \Theta < \pi, \quad [19]$$

where Θ , traditionally called the scattering angle, is the angle between the incidence and scattering directions. It is easy to see that the scattering matrix relates the Stokes parameters of the incident and scattered waves defined with respect to the same so-called scattering plane (i.e., the plane through the incidence and scattering directions).

Scattering by a Random Many-Particle Group

Although we have been so far discussing electromagnetic scattering by a ‘single particle,’ the concept of electromagnetic scattering and eqns [1]–[5] remain valid irrespective of the specific morphology of the scattering object. In particular, they are valid for what a human eye could classify as a ‘collection of discrete particles.’ Examples of such ‘many-particle’ objects are clouds, particulate surfaces, and particle suspensions. In all such cases, the electromagnetic field perceives a morphologically complex ‘many-particle’ object at any moment in time as one scatterer in the form of a specific spatial distribution of the relative refractive index throughout the cumulative interior volume V_{INT} in eqns [2] and [5].

However, the numerically exact computer solution of the Maxwell equations becomes prohibitively expensive when the size parameter of the object (i.e., the product of the wave number k and the radius of the smallest circumscribing sphere of the object) exceeds ~ 100 . Furthermore, the concept of far-field scattering is inapplicable in the majority of practical situations involving very large many-particle groups such as clouds. Indeed, radiometers are often positioned inside or relatively close to the cloud (Figure 4) (i.e., in its near zone). As a consequence, one has to resort to an approximate computational technique and often abandon the attractively simple formulas of far-field scattering.

Two conventional approaches widely used to treat electromagnetic scattering by random particle groups are the single-scattering approximation (SSA) and the radiative transfer theory (RTT). The SSA is applicable to a relatively small, ‘optically tenuous’ group of N particles viewed from a distance much greater than the entire size of the group. In this case, eqns [11]–[14] remain valid provided that (1) the scattering signal is accumulated over a time interval long enough to establish full statistical randomness of the group; and (2) the phase, extinction, and scattering matrices of the entire group are also averaged over time. Then:

$$\langle \mathbf{Z}(\hat{\mathbf{n}}^{\text{sca}}, \hat{\mathbf{n}}^{\text{inc}}) \rangle_t = N \langle \mathbf{Z}(\hat{\mathbf{n}}^{\text{sca}}, \hat{\mathbf{n}}^{\text{inc}}) \rangle_{\xi}, \quad [20]$$

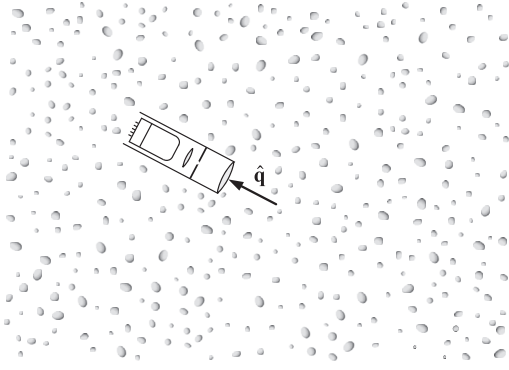


Figure 4 A well-collimated polarization-sensitive radiometer is placed inside a cloud of particles.

$$\langle \mathbf{K}(\hat{\mathbf{n}}^{\text{inc}}) \rangle_t = N \langle \mathbf{K}(\hat{\mathbf{n}}^{\text{inc}}) \rangle_\xi, \quad [21]$$

$$\langle \mathbf{F}(\Theta) \rangle_t = N \langle \mathbf{F}(\Theta) \rangle_\xi, \quad [22]$$

where $\langle \mathbf{Z}(\hat{\mathbf{n}}^{\text{sca}}, \hat{\mathbf{n}}^{\text{inc}}) \rangle_\xi$, $\langle \mathbf{K}(\hat{\mathbf{n}}^{\text{inc}}) \rangle_\xi$, and $\langle \mathbf{F}(\Theta) \rangle_\xi$ are the *single-particle* phase, extinction, and scattering matrix, respectively, averaged over all physically realizable particle states ξ in the group. The state of a particle indicates collectively its size, refractive index, shape, and orientation (i.e., all physical characteristics except the position). Note that it is the assumption of full ergodicity of the random scattering process that allows one to replace time averaging by ensemble averaging in eqns [20]–[22]. Obviously, the time averages of the extinction, scattering, and absorption cross-sections of the entire random particle group can be expressed similarly in terms of the respective ensemble-averaged single-particle cross-sections. The single-scattering albedo of the group is then given by:

$$\tilde{\omega}(\hat{\mathbf{n}}^{\text{inc}}) = \frac{\langle C_{\text{sca}}(\hat{\mathbf{n}}^{\text{inc}}) \rangle_\xi}{\langle C_{\text{ext}}(\hat{\mathbf{n}}^{\text{inc}}) \rangle_\xi}. \quad [23]$$

The RTT is an expressly near-field theory that has recently been derived directly from the Maxwell equations and allows one to model the response of a well-collimated radiometer located inside or relatively close to a random multiparticle scattering object (Figure 4). Among the conditions of applicability of the RTE are the asymptotic requirement $N \rightarrow \infty$; the ‘low-density’ requirement, according to which every particle must be located sufficiently far from all the other particles; and the assumption that the scattering signal is accumulated over a time interval long enough to establish full ergodicity of the random scattering process. The integro-differential form of the RTE reads:

$$\hat{\mathbf{q}} \cdot \nabla \tilde{\mathbf{I}}(\mathbf{r}, \hat{\mathbf{q}}) = -n_0 \langle \mathbf{K}(\hat{\mathbf{q}}) \rangle_\xi \tilde{\mathbf{I}}(\mathbf{r}, \hat{\mathbf{q}}) + n_0 \int_{4\pi} d\hat{\mathbf{q}}' \langle \mathbf{Z}(\hat{\mathbf{q}}, \hat{\mathbf{q}}') \rangle_\xi \tilde{\mathbf{I}}(\mathbf{r}, \hat{\mathbf{q}}'), \quad [24]$$

where n_0 is the average particle number density, $d\hat{\mathbf{q}}'$ is an elementary solid angle centered on the unit vector $\hat{\mathbf{q}}'$, and $\tilde{\mathbf{I}}(\mathbf{r}, \hat{\mathbf{q}})$ is the four-component so-called specific Stokes column vector. Unlike the Stokes column vectors [9] and [10], $\tilde{\mathbf{I}}(\mathbf{r}, \hat{\mathbf{q}})$ has the dimension of radiance ($\text{Wm}^{-2} \text{sr}^{-1}$). Thus, knowledge of the ensemble-averaged single-particle phase and extinction matrices is also required in order to solve the RTE.

The fundamental importance of the RTT is that its solution, $\tilde{\mathbf{I}}(\mathbf{r}, \hat{\mathbf{q}})$, directly quantifies the response of a well-collimated polarization-sensitive radiometer oriented along the unit vector $\hat{\mathbf{q}}$ at the observation point $\mathbf{r}$ (Figure 4). Furthermore, the integral $\int_{4\pi} d\hat{\mathbf{q}} \hat{\mathbf{q}} \tilde{\mathbf{I}}(\mathbf{r}, \hat{\mathbf{q}})$ gives the time-averaged Poynting vector describing the direction and rate of the local electromagnetic energy transport. Thus, the RTT is directly applicable to solving both remote-sensing and radiation budget problems.

Symmetries

In general, all elements of the extinction and phase matrices entering eqns [13] and [14] can be nonzero, which implies that the intensity I registered by radiometers 1 and 2 in Figure 3(b) can depend on all four Stokes parameters of the incident field rather than only on its intensity. This fact emphasizes the vectorial (rather than scalar) character of electromagnetic scattering. In particular, dichroism results in different attenuation rates for different polarization components of the incident field. This causes, for example, depolarization of radar signals propagating through precipitation. The scattered wave recorded by radiometer 2 in Figure 3(b) also has polarization characteristics different from those of the incident field, thereby making polarimetry a sensitive particle characterization technique.

In many cases of practical interest, the mathematical structure of the ensemble-averaged extinction, phase, and scattering matrices becomes much simpler. This happens, for example, when (1) the distribution of particle orientations in a random group during the measurement is uniform, and (2) each particle in the group has a plane of symmetry and/or is accompanied by its mirror counterpart. Then the average extinction, scattering, and absorption cross-sections and the single-scattering albedo become independent of the direction of propagation and polarization state of the incident field. The average extinction matrix is diagonal and given by:

$$\langle \mathbf{K}(\hat{\mathbf{n}}^{\text{inc}}) \rangle_\xi \equiv \langle \mathbf{K} \rangle_\xi = \langle C_{\text{ext}} \rangle_\xi \text{diag}[1, 1, 1, 1]. \quad [25]$$

The average phase matrix satisfies certain useful symmetry relations and depends on only the difference between the azimuthal angles of the scattering and incidence directions rather than on their specific values. The average scattering matrix has a simple block-diagonal structure with only six independent elements:

$$\langle \mathbf{F}(\Theta) \rangle_\xi = \begin{bmatrix} \langle F_{11}(\Theta) \rangle_\xi & \langle F_{12}(\Theta) \rangle_\xi & 0 & 0 \\ \langle F_{12}(\Theta) \rangle_\xi & \langle F_{22}(\Theta) \rangle_\xi & 0 & 0 \\ 0 & 0 & \langle F_{33}(\Theta) \rangle_\xi & \langle F_{34}(\Theta) \rangle_\xi \\ 0 & 0 & -\langle F_{34}(\Theta) \rangle_\xi & \langle F_{44}(\Theta) \rangle_\xi \end{bmatrix}. \quad [26]$$

Furthermore,

$$\langle F_{12}(0) \rangle_\xi = \langle F_{12}(\pi) \rangle_\xi = 0, \quad \langle F_{34}(0) \rangle_\xi = \langle F_{34}(\pi) \rangle_\xi = 0. \quad [27]$$

For spherically symmetric particles, the number of independent scattering matrix elements reduces to four owing to

the identities $\langle F_{22}(\Theta) \rangle_\xi \equiv \langle F_{11}(\Theta) \rangle_\xi$ and $\langle F_{44}(\Theta) \rangle_\xi \equiv \langle F_{33}(\Theta) \rangle_\xi$. As a consequence, measurements of the linear backscattering depolarization ratio

$$\delta_L = \frac{\langle F_{11}(\pi) \rangle_\xi - \langle F_{22}(\pi) \rangle_\xi}{\langle F_{11}(\pi) \rangle_\xi + \langle F_{22}(\pi) \rangle_\xi}, \quad 0 \leq \delta_L \leq 1 \quad [28]$$

and the closely related circular backscattering depolarization ratio

$$\delta_C = \frac{\langle F_{11}(\pi) \rangle_\xi + \langle F_{44}(\pi) \rangle_\xi}{\langle F_{11}(\pi) \rangle_\xi - \langle F_{44}(\pi) \rangle_\xi} = \frac{2\delta_L}{1 - \delta_L} \geq \delta_L \quad [29]$$

are among the most reliable means of detecting particle nonsphericity.

Measurement and Computation of Single-Particle Characteristics

Evaluation of the Earth's radiation balance and analyses of remote-sensing observations require accurate quantitative knowledge of average single-particle optical characteristics as functions of particle size, morphology, and composition. This knowledge can be based on theoretical computations or experimental measurements, both approaches having their strengths and limitations. Theoretical modeling does not involve expensive instrumentation and allows switching to another particle shape, size, or refractive index by changing a few lines in a computer code. However, numerically exact computations for realistic polydispersions of morphologically complex particles are costly, if even possible, and are often replaced by computations for idealized shapes, whereas approximate calculations often have uncertain accuracy and range of applicability. Experimental measurements deal with real particles, but require complex and expensive hardware and may be difficult to interpret.

Measurements

Detectors of visible and infrared light are usually polarization insensitive, so their response is determined by only the first Stokes parameter of the incoming beam. In order to measure all elements of the scattering matrix, one must insert into the beam various optical elements that can vary the polarization state of light before and after scattering in a controllable way (Figure 5). The use of high-frequency sinusoidal modulation in the time of the polarization of light before scattering combined with intensity normalization and followed by lock-in detection increases the measurement accuracy and yields several elements from only one detected signal. The measurement procedure is repeated at different scattering angles in order to determine the angular profile of the scattering matrix.

Scattering measurements using visible and infrared light benefit from the availability of sensitive detectors (photo-multipliers and avalanche semiconductor photodiodes), intense sources of radiation (lasers), and high-quality optical elements. They involve relatively inexpensive and portable instrumentation and can be performed in the field nearly as well as in the laboratory. However, they often suffer from poor advance knowledge of microphysical characteristics of scattering particles, thereby making difficult comparisons of experimental and theoretical results. The arrangement of the source of light and the detector precludes measurements at scattering angles close to 0° and 180° , which makes problematic the absolute measurement of the (1,1) element of the scattering matrix and the scattering cross-section.

The main idea of the microwave analog technique is to manufacture a centimeter-sized scattering object with desired shape and refractive index, measure the scattering of a microwave beam by this object, and finally extrapolate the results to other wavelengths (e.g., visible or infrared) by keeping the ratio of size and wavelength fixed. This allows one to determine so-called scale-invariant characteristics such as the phase

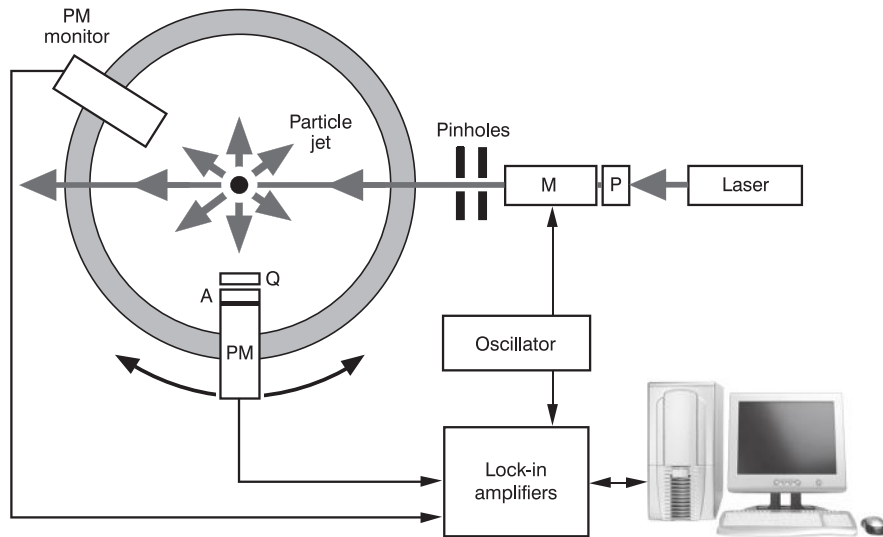


Figure 5 Schematic view of an experimental scattering setup using visible or infrared light. The laser beam passes several optical elements before and after scattering and is detected by the photomultiplier. The latter is mounted on a circular rail and can be moved around the particle jet in order to cover a wide range of scattering angles. A, polarization analyzer; M, electro-optic modulator; P, polarizer; PM, photomultiplier; Q, quarter-wave plate.

function $p(\Theta) = 4\pi \langle F_{11}(\Theta) \rangle_{\xi} / \langle C_{\text{sca}} \rangle_{\xi}$ or ratios of the elements of the scattering matrix. Microwave measurements allow a wide coverage of scattering angles, including the exact forward and backward directions, and a much greater degree of control over the target size, shape, and orientation than do optical measurements. However, the microwave measurements require more cumbersome and expensive instrumentation and large measurement facilities. Furthermore, they are performed for only one particle size, shape, and orientation at a time, thereby making ensemble averaging a time-consuming process.

Theoretical Techniques

All of the needs of a practitioner dealing with electromagnetic scattering by spherical particles may be well served by the exact and highly efficient Lorenz–Mie theory, which is the result of applying the separation-of-variables method to the Maxwell equations in spherical coordinates. There are extensions of the Lorenz–Mie theory applicable to concentric layered spheres.

For nonspherical particles, numerically exact computations must resort to more general and complex solutions traditionally divided into two broad categories. Differential equation methods compute the scattered field by solving the Maxwell equations, subject to appropriate boundary conditions, in the frequency domain or in the time domain. Integral equation methods are based on the volume or surface integral counterparts of the Maxwell equations; the boundary conditions are included in the solution automatically. Perhaps the most popular and widely used techniques are the *T*-matrix method, the discrete dipole approximation, and the finite-difference time-domain method. These techniques have somewhat different ranges of applicability in terms of particle morphology, refractive index, and size relative to the wavelength.

The practical importance of approximate treatments of light scattering diminishes as various exact techniques mature and become applicable to a wider range of problems, and as computers become ever more powerful. However, at least one phenomenological approach is unlikely to become obsolete in the near future because its accuracy can only be expected to improve as the ratio of the particle size to the wavelength grows, while numerically exact theoretical techniques for nonspherical particles cease to be practical whenever this ratio exceeds a certain threshold. This so-called geometrical optics approximation assumes that the particle size is much larger than the wavelength and that the incident plane wave can be represented as a collection of ‘independent parallel rays.’ The history of each ray impinging on the particle surface is traced using Snell’s law and Fresnel formulas (Figure 6). Sampling all escaping rays into predefined narrow angular bins yields a quantitative representation of the particle scattering and absorption properties. The ray-tracing pattern is supplemented by computation of the Fraunhofer diffraction of the incident wave on the particle projection.

Solution of the RTE

Based on the specific problem at hand, the RTE [24] must be supplemented by appropriate boundary conditions. For example, a standard model of the atmosphere is a multilayer

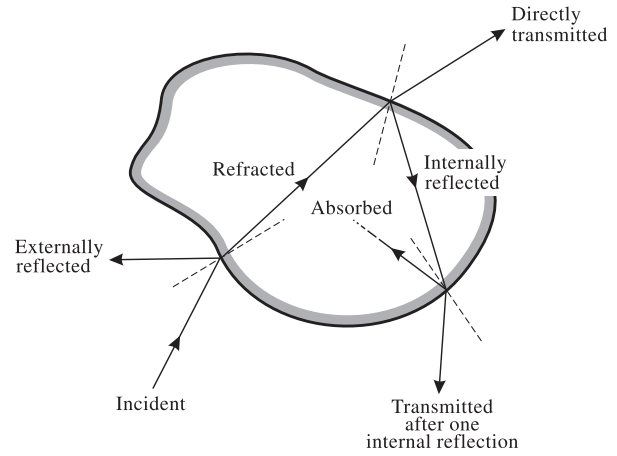


Figure 6 Ray-tracing diagram.

plane-parallel system illuminated from above by solar radiation and bounded from below by a horizontally homogeneous reflecting surface. Then the RTE can be solved numerically using efficient techniques such as the adding and doubling, discrete ordinate, and invariant imbedding methods. Complex horizontally and vertically inhomogeneous models can be handled using the less efficient but more flexible Monte Carlo technique.

Despite the expressly vectorial nature of electromagnetic scattering, the RTE [24] is often replaced by a simplified scalar version in which one keeps only the first Stokes parameter (i.e., the intensity) and only the (1,1) elements of the extinction and phase matrices. Although it is much easier to solve the scalar RTE, it is important to remember that it has no physical justification and can result in significant errors in the computed intensity.

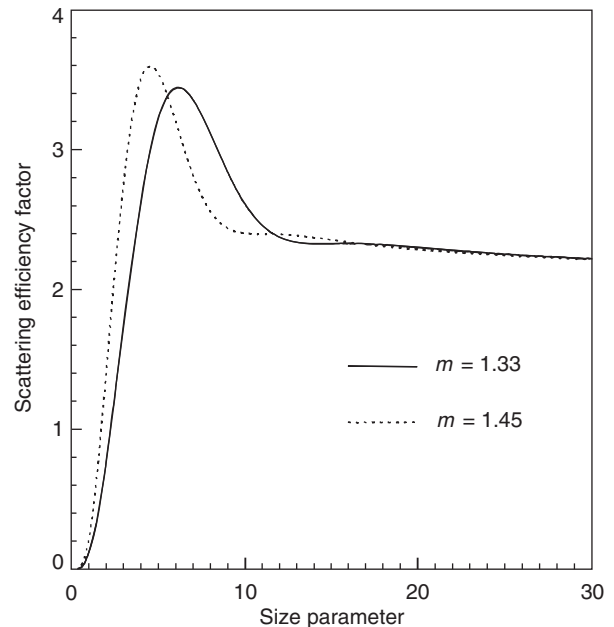


Figure 7 Scattering efficiency factor versus size parameter for poly-disperse spherical particles with refractive indices 1.33 and 1.45.

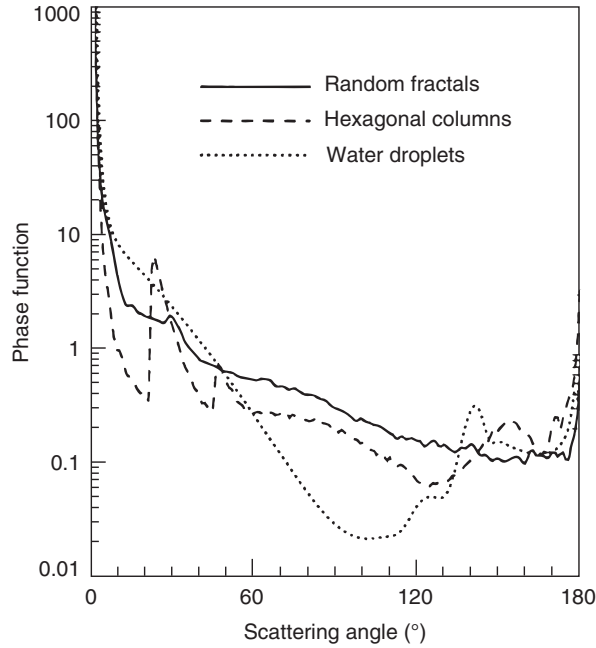


Figure 8 Phase functions for water cloud droplets, hexagonal ice columns, and randomly shaped ice crystals.

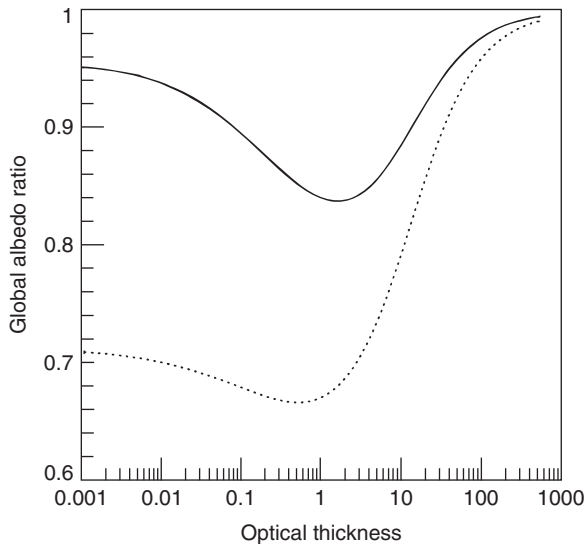


Figure 9 Global albedo of a liquid water cloud relative to that of an optical-thickness-equivalent ice cloud composed of irregular particles (dotted curve) and hexagonal columns (solid curve).

Examples of Scattering by Clouds and Aerosols

The few following examples serve to illustrate some key features of electromagnetic scattering phenomena and their use in remote sensing of aerosols and clouds and in radiation budget computations. **Figure 7** shows the scattering efficiency factor Q_{sca} versus size parameter x for a narrow size distribution of spherical particles with refractive indices $m = 1.33$ and 1.45 (typical of water and sulfate aerosols at visible wavelengths,

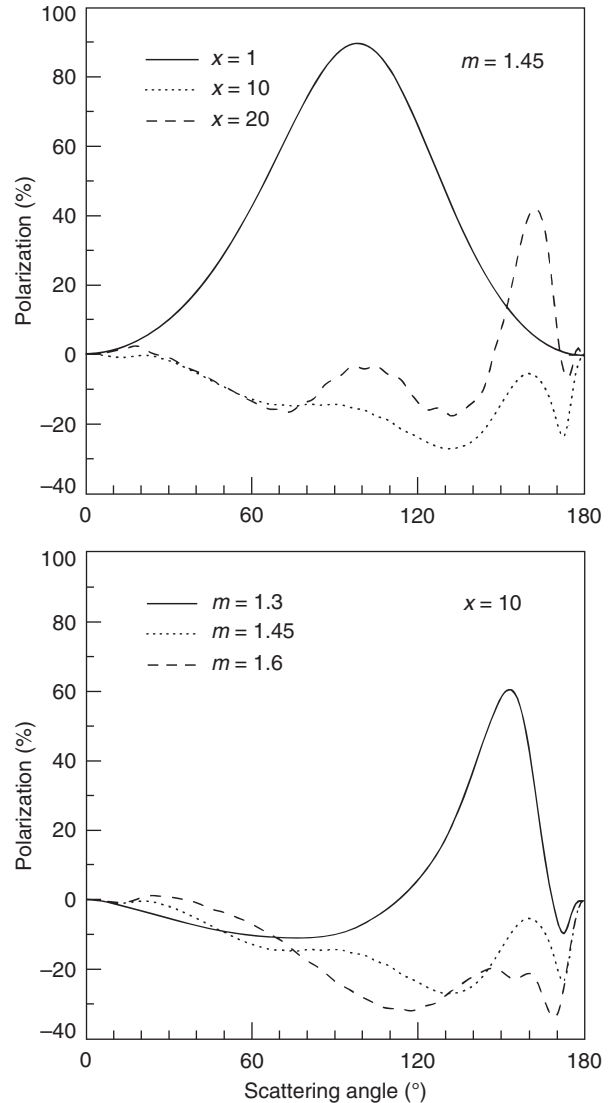


Figure 10 Linear polarization versus scattering angle for polydisperse spherical particles with varying size parameters x and refractive indices m .

respectively). The size parameter is defined as $x = 2\pi r_{\text{eff}}/\lambda$, where r_{eff} is the effective radius of the size distribution. The scattering efficiency is defined as the ratio of the average scattering cross-section to the average area of the particle geometrical projection. One can see that for wavelength-sized particles ($x \sim 5$), $\langle C_{\text{sca}} \rangle_{\xi}$ can exceed the particle geometrical cross-section by a factor greater than 3.5. As the particle size becomes much larger, Q_{sca} tends to the asymptotic geometrical-optics value of 2, with equal contributions from the rays striking the particle and the light diffracted by the particle projection. For particles much smaller than the wavelength, $Q_{\text{sca}} \propto \lambda^{-4}$, as first demonstrated by Lord Rayleigh and hence called Rayleigh scattering. The presence of a well-defined maximum in the scattering efficiency curve for a relatively narrow polydispersion (**Figure 7**) creates the possibility of an infrequent phenomenon for which aerosol particles of just the right size have a lower extinction efficiency factor in the blue than that at the larger wavelengths in the red. Thus, in contrast to the familiar

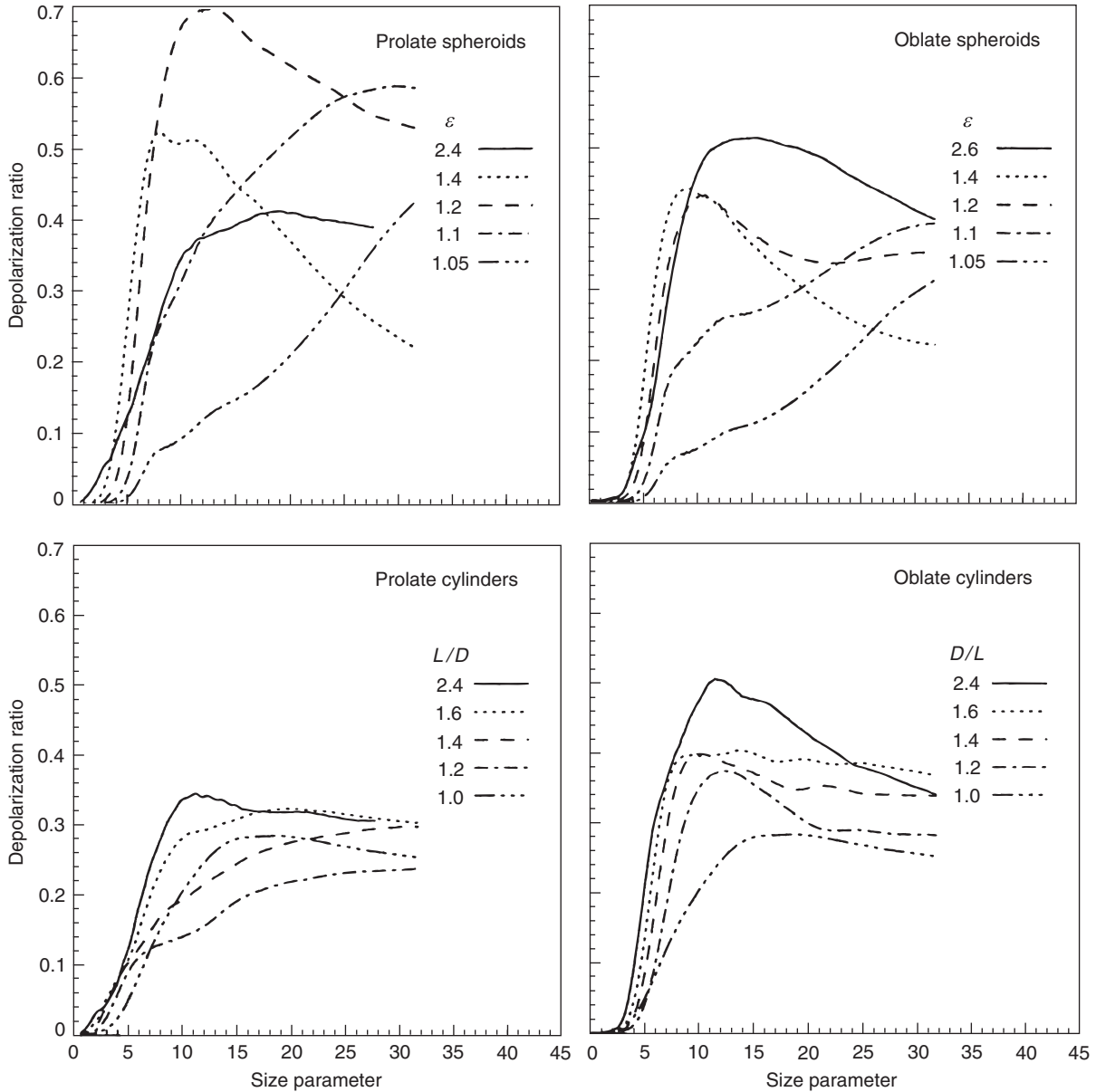


Figure 11 Linear depolarization ratio versus surface-equivalent sphere size parameter for polydisperse, randomly oriented ice spheroids and cylinders. The refractive index is 1.311. ϵ is the ratio of the largest to the smallest axes of a spheroid. The shapes of prolate and oblate cylinders are specified by length-to-diameter L/D and diameter-to-length D/L ratios, respectively.

reddening of the setting sun owing to enhanced Rayleigh scattering, a sufficiently narrow size distribution of aerosol particles in the atmosphere can produce a blue cast to the sun or moon and is perhaps responsible for the implied rarity associated with the phrase ‘once in a blue moon.’

The dotted curve in Figure 8 shows the phase function typical of spherical cloud droplets at visible wavelengths. The strong concentration of light at $\Theta = 0^\circ$ is produced by Fraunhofer diffraction of light on the particle projection, whereas the feature at $\Theta \sim 140^\circ$ is the primary rainbow generated by rays that have undergone only one internal reflection (Figure 6). The slight change of the rainbow angle with wavelength caused by dispersion gives rise to spectacularly

colorful rainbows often observed during showers illuminated by the sun at an altitude lower than about 40° . The enhanced intensity at $\Theta \sim 180^\circ$ is called the glory and can be seen from an airplane as a series of colored rings around the shadow cast by the airplane on the cloud top.

The dashed curve in Figure 8 depicts the phase function typical of randomly oriented pristine hexagonal ice crystals. The concentrations of light at $\Theta \sim 22^\circ$ and 46° are the primary and secondary halos attributed to minimum angles of deviation by 60° and 90° ice prisms. These features represent only two of many optical phenomena associated with regularly shaped ice crystals. Since cirrus clouds rather often fail to exhibit halos, the majority of real ice crystals appear to have

highly irregular shapes and rough rather than flat surfaces. Such particles are better characterized by featureless phase functions like the one shown in Figure 8 by the solid curve and computed for a random-fractal model of ice crystals.

Large numerical differences between the three phase functions depicted in Figure 8 can cause significant differences in bidirectional reflectance of optical-thickness-equivalent water and ice clouds (optical thickness is defined as the average extinction cross-section per particle times the column particle number concentration). This, in turn, may lead to significant errors in the retrieved cloud optical thickness if remote-sensing reflectance measurements are inverted using an incorrect particle model.

Figure 9 illustrates the effect of particle shape on the global cloud albedo (defined as cloud reflectance averaged over all incidence and reflection directions) at visible wavelengths. The quantity $(1 - \text{cloud albedo})$ determines how much solar energy is absorbed or transmitted by the atmosphere and is an important climatological parameter. It is seen that for the same optical thickness, clouds composed of irregular ice crystals have the largest albedo, whereas those composed of water droplets are the least reflective. This result can be explained by very large differences between the respective phase functions at side-scattering angles, which are well seen in Figure 8.

Figure 10 illustrates the ratio $P(\Theta) = -\langle F_{12}(\Theta) \rangle_{\xi} / \langle F_{11}(\Theta) \rangle_{\xi}$ of the scattering matrix elements (called the degree of linear polarization of the scattered light for unpolarized incident light) for polydisperse spheres with different refractive indices and size parameters. The obvious significant variability of polarization with m and r_{eff} (or λ) makes it a very sensitive indicator of the particle microphysical properties. Furthermore, since P is a ratio of two intensities, it can be measured to a much greater precision than intensity. These two factors explain the remarkable potential of photopolarimetry as a remote-sensing tool for aerosol and cloud particle characterization.

Finally, Figure 11 demonstrates the linear depolarization ratio [28] for polydisperse, randomly oriented, nonspherical ice particles. It is well seen that the growth of the size parameter from 0 to about 10 rapidly changes δ_L from zero to large values sometimes approaching the theoretical limit $\delta_{L,\text{max}} = 1$. This

behavior of backscattering depolarization makes it useful for sizing aerosol, cloud, and precipitation particles by performing multiwavelength lidar and radar measurements. It is also obvious that although backscattering depolarization is a reliable indicator of the presence of nonspherical particles, its magnitude is not always a good measure of the degree of particle nonsphericity.

See also: **Aerosols:** Role in Radiative Transfer. **Clouds and Fog:** Classification of Clouds; Cloud Microphysics; Contrails. **Lidar:** Atmospheric Sounding Introduction. **Radar:** Precipitation Radar. **Satellites and Satellite Remote Sensing:** Aerosol Measurements.

Further Reading

- Chandrasekhar, S., 1960. Radiative Transfer. Dover, New York.
- Davis, A.B., Marshak, A., 2010. Solar radiation transport in the cloudy atmosphere: a 3D perspective on observations and climate impacts. Reports on Progress in Physics 73, 026801.
- Goody, R.M., Yung, Y.L., 1989. Atmospheric Radiation. Oxford University Press, New York.
- Hansen, J.E., Travis, L.D., 1974. Light scattering in planetary atmospheres. Space Science Reviews 16, 527–610.
- Lenoble, J. (Ed.), 1985. Radiative Transfer in Scattering and Absorbing Atmospheres: Standard Computational Procedures. Deepak, Hampton, VA.
- Liou, K.N., 2002. An Introduction to Atmospheric Radiation. Academic Press, San Diego, CA.
- Mishchenko, M.I., 2011. Directional radiometry and radiative transfer: a new paradigm. Journal of Quantitative Spectroscopy and Radiative Transfer 112, 2079–2094.
- Mishchenko, M.I., Hovenier, J.W., Travis, L.D. (Eds.), 2000. Light Scattering by Nonspherical Particles. Academic Press, San Diego, CA.
- Mishchenko, M.I., Travis, L.D., Lacis, A.A., 2002. Scattering, Absorption, and Emission of Light by Small Particles. Cambridge University Press, Cambridge.
- Mishchenko, M.I., Travis, L.D., Lacis, A.A., 2006. Multiple Scattering of Light by Particles. Cambridge University Press, Cambridge.
- Stephens, G.L., 1994. Remote Sensing of the Lower Atmosphere. Oxford University Press, New York.
- van de Hulst, H.C., 1957. Light Scattering by Small Particles. Wiley, New York.
- Wendisch, M., Yang, P., 2012. Theory of Atmospheric Radiative Transfer. Wiley-VCH, Weinheim, Germany.

Ultraviolet Radiation

K Stammes, Stevens Institute of Technology, Hoboken, NJ, USA

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

This article provides a review of ultraviolet radiation in the atmosphere–ocean system with emphasis on the basic interaction with molecules and suspended particles in both media as well as the reflection and refraction occurring at the air–water interface. A comparison between measured and computed irradiances in the ocean is provided to illustrate the utility of the modeling. The model is also used to discuss the impact of an ozone depletion at high latitudes the primary production in the water column.

Introduction

Life on Earth began with light. Before oxygen developed in the Earth's atmosphere the ocean as well as fresh water bodies served as a suitable environment for the evolution of Earth's early life forms, because water provided protection from the damaging ultraviolet radiation (UVR) from the Sun. Photoinduced chemical reactions led to photosynthesis. Oxygen produced by photosynthetic bacteria led to the formation of ozone, and eventually to life forms that could develop also on land, because the ozone layer provided vital protection against harmful UVR.

In the primordial atmosphere, and today as well, the formation of *photochemically active* species is initiated by ultraviolet and visible solar radiation through the process of *photolysis*, in which molecules are split up into atoms and smaller molecules. Photolysis of the oxygen molecule (O_2) leads to the formation of oxygen atoms (O). Ozone is then formed when an oxygen atom and an oxygen molecule combine to yield O_3 . Chemical reactions catalyzed by photolysis are responsible for the destruction of atmospheric ozone. The bulk of the ozone resides in the stratosphere (15–50 km), where its abundance is determined by a balance between production and loss processes. The ozone production depends on the photolysis of O_2 , which is proportional to the product of the solar flux of dissociating UVR and the density of the O_2 gas (see eqn [7]). The solar UV light intensity falls off rapidly with depth into the atmosphere, due to absorption and scattering by atmospheric gases, whereas the number density of O_2 molecules decreases exponentially with height. Thus, the production of ozone will have a maximum at that height where the two curves cross each other.

This article provides a description of how UVR interacts with molecules and suspended matter in the atmosphere and in the ocean. It gives a basic description of this interaction, as well as examples of how UV penetration into the ocean is affected by the stratospheric ozone layer, and by suspended particles in the ocean.

Spectrum of Electromagnetic Radiation from the Sun

An overview of the various parts of the solar spectrum is provided in Table 1. The spectral variable is the wavelength $\lambda = c/\nu$, where c is the speed of light and ν is the frequency (s^{-1}) or (Hz). In the UV

and visible spectral range, λ is expressed in nanometers ($1 \text{ nm} = 10^{-9} \text{ m}$). The irradiance in each spectral range is listed as well as the known percentage solar variability, defined as the maximum minus minimum divided by the minimum.

Figure 1 shows the UV, visible, and near-infrared part of the spectral solar irradiance (wavelengths shorter than 1000 nm) measured on board an earth-orbiting satellite, above the atmosphere. Spectra of ideal blackbodies at several temperatures are also shown in Figure 1. Requiring that the total energy emitted be the same as a blackbody, one finds that the Sun's effective temperature is 5778 K. If the radiating layers of the Sun had a uniform temperature at all depths, its spectrum would match one of the theoretical blackbody curves exactly. Therefore, the deviations are the result of emission from a nonisothermal solar atmosphere. Some of the more important aspects of the UV/visible spectrum are: (1) most of the emission arises within the *photosphere* where the Sun's visible optical depth reaches unity, while the finer structure is due to *Fraunhofer absorption* by gases in the cooler (higher) portions of the photosphere; (2) for $125 < \lambda < 380 \text{ nm}$, the effective radiating temperature falls to values as low as 4500 K, due to increased numbers of overlapping absorbing lines. At still shorter wavelengths, some of the emission originates in the hotter *chromosphere*, which overlies the photosphere, and the effective temperature increases; and (3) the UV irradiance depends noticeably upon the solar cycle, being more intense at high than at low solar activity.

Sun–Earth Geometry

Since the Earth moves around the Sun in an elliptical orbit, the Earth–Sun distance, R_n , varies throughout the year, such that on day number d_n (0 for 1 January, 364 for 31 December)

$$R_n = R_0(a_0 + a_1 \cos \theta_n + a_2 \sin \theta_n + a_3 \cos 2\theta_n + a_4 \sin 2\theta_n)^{-1/2} \quad [1]$$

where R_0 is the average Earth–Sun distance. The coefficients are $a_0 = 1.000110$, $a_1 = 0.034221$, $a_2 = 0.001280$, $a_3 = 0.000719$, $a_4 = 0.000077$, and $\theta_n \equiv 2\pi d_n/365$ (radians). R_n varies by about 3.4% from its minimum value on about 3 January to its maximum value on about 3 July. Thus, the variation in R_n^2 and therefore in the extraterrestrial solar irradiance, is about 6.9%. This variation implies that the UV exposure is almost 7% larger

Table 1 Subregions of the spectrum

Subregion	Irradiance (W m^{-2})	Solar variability	Comments
Far UV ($100 < \lambda < 200 \text{ nm}$)	<1	7–80%	Dissociates O_2 . Discrete electronic excitation of atomic resonance lines.
Middle UV, or UV-C ($200 < \lambda < 280 \text{ nm}$)	6.4	1–2%	Dissociates O_3 in intense Hartley bands. Potentially lethal to biosphere.
UV-B ($280 < \lambda < 320 \text{ nm}$)	21.1	$<1\%$	Some radiation reaches surface, depending on O_3 optical depth. Damaging to biosphere. Responsible for skin erythema.
UV-A ($320 < \lambda < 400 \text{ nm}$)	85.7	$<1\%$	Reaches surface. Benign to humans. Scattered by clouds, aerosols, and molecules.
Visible, or PAR ($400 < \lambda < 700 \text{ nm}$)	532	$\leq 0.1\%$	Absorbed by ocean, land. Scattered by clouds, aerosols, and molecules. Primary energy source for biosphere and climate system.
Near IR ($0.7 < \lambda < 3.5 \mu\text{m}$)	722	–	Absorbed by O_2 , H_2O , CO_2 in discrete vibrational bands.

PAR, photosynthetically active radiation. Adapted with permission from Thomas, G.E., Stamnes, K., 1999. Radiative Transfer in the Atmosphere and Ocean. Cambridge University Press, Cambridge.

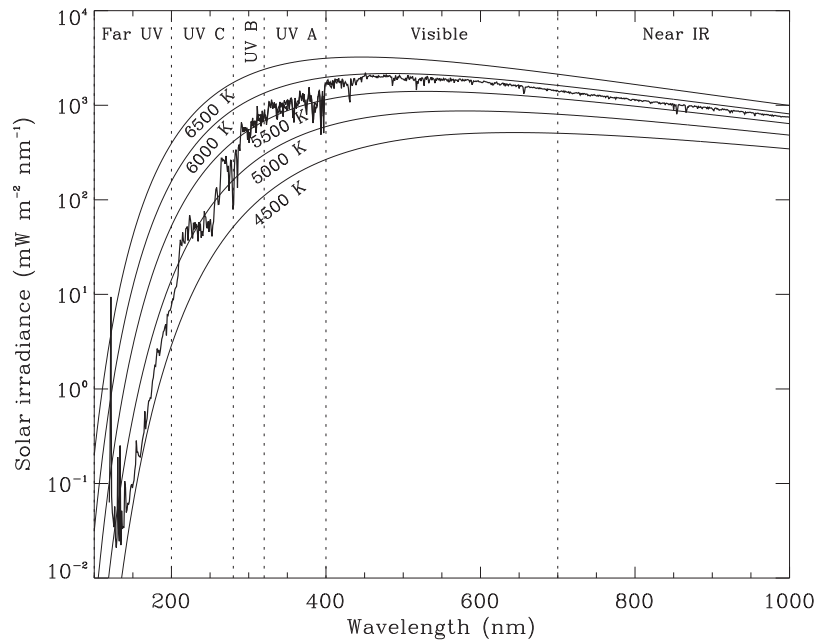


Figure 1 Extraterrestrial solar irradiance, measured by a spectrometer on board an Earth-orbiting satellite. The UV spectrum ($119 < \lambda < 420 \text{ nm}$) was measured by the SOLSTICE instrument on the UARS satellite (modified from a diagram provided by G.J. Rottmann, private communication, 1995). The vertical lines divide the various spectral subranges defined in Table 1. The smooth curves are calculated blackbody spectra for a number of emission temperatures.

in the Southern Hemisphere in January, than at the same latitude in the Northern Hemisphere in July. The illumination of the Earth's surface depends on the *solar zenith angle*, θ_0 , which is the angle between the local vertical and the direction of the center of the solar disk. The complementary angle between the local horizon and the Sun is called the *solar elevation angle*. The solar zenith angle depends on the time of the day, the time of the year, and geographic location. Thus, the solar illumination varies on diurnal, seasonal, as well as the 11-year solar cycle timescales.

Interaction of UV Light with Absorbing and Scattering Media

To describe the interaction of light with the atmosphere and the ocean (or freshwater bodies), it is convenient to think of such

media as composed of radiatively active 'particles' that absorb and scatter light. In the atmosphere, these particles consist of molecules, aerosols, and cloud droplets or ice crystals. In the ocean, they consist of molecules (although strictly speaking the scattering by pure water is caused by molecular density fluctuations) and hydrosols (suspended particles of organic and inorganic origin).

Definitions of Irradiance and Radiance

The spectral net irradiance F_ν within a small frequency range ν to $\nu + d\nu$ is the net energy d^3E crossing a surface element dA per unit time and per unit frequency: $F_\nu = \frac{d^3E}{dA dt d\nu} = F_\nu^+ - F_\nu^- [\text{W} \cdot \text{m}^{-2} \cdot \text{Hz}^{-1}]$. Here, we have introduced the spectral

hemispherical irradiances $F_\nu^+ = d^3E^+/dA dt d\nu$ and $F_\nu^- = d^3E^-/dA dt d\nu$.

The *directional dependence* of the energy flow is obtained by considering the energy d^4E that flows within a solid angle $d\omega$ around direction $\hat{\Omega}$ in the time interval dt and within the frequency range $d\nu$. The *spectral radiance* I_ν is defined as the energy per unit area per unit solid angle, per unit frequency, and per unit time that passes through a surface element dA : $I_\nu = d^4E/\cos\theta dA dt d\omega d\nu$ ($\text{W}\cdot\text{m}^{-2}\cdot\text{sr}^{-1}\cdot\text{Hz}^{-1}$), where θ is the angle between the surface normal $\hat{n}$ and the direction of propagation $\hat{\Omega}$. From these definitions, the spectral net irradiance can be expressed as: $F_\nu = F_\nu^+ - F_\nu^- = \int_+ d\omega \cos\theta I_\nu - \int_- d\omega \cos\theta I_\nu = \int_{4\pi} d\omega \cos\theta I_\nu$, where the subscripts $+$ and $-$ on the integral signs denote integration over each hemisphere. Finally, we define the mean intensity as: $\bar{I}_\nu = (1/4\pi) \int_{4\pi} d\omega I_\nu$, which is simply the radiance averaged over the sphere. Integration over all frequencies yields the *net irradiance* $F = \int_0^\infty d\nu F_\nu$ and the *actinic radiation* $4\pi\bar{I} = 4\pi \int_0^\infty d\nu \bar{I}_\nu$ both in units of ($\text{W}\cdot\text{m}^{-2}$).

Absorption, Scattering, and Extinction by Molecules and Particles

It is found experimentally that if a beam of light is incident on a thin layer of thickness ds consisting of radiatively active constituents, then the light is attenuated so that the differential loss in radiance is $dI_\nu = -k(\nu)I_\nu ds$, where k is called the extinction coefficient. Integration yields

$$I_\nu(s, \hat{\Omega}) = I_\nu(0, \hat{\Omega}) \exp[-\tau_s(\nu)]. \quad [2]$$

Here, $\hat{\Omega}$ denotes the propagation direction of the beam, and the dimensionless extinction optical path or opacity along the path s is given by $\tau_s(\nu) \equiv \int_0^s ds' k(\nu)$. The extinction optical path of a mixture of scattering/absorbing molecules and particles is defined as the sum of the individual scattering optical path, $\tau_{sc}(\nu)$, and the absorption optical path, $\tau_a(\nu) = \tau_{sc}(\nu) + \tau_a(\nu)$, where $\tau_{sc}(\nu) = \int_0^s ds' \sum_i \sigma^i(\nu, s') = \int_0^s ds' \sigma(\nu, s')$ and $\tau_a(\nu) = \int_0^s ds' \sum_i \alpha^i(\nu, s') = \int_0^s ds' \alpha(\nu, s')$. The sum is over all optically active species, and $\sigma^i(\nu, s')$ and $\alpha^i(\nu, s')$ (m^{-1}) are the individual scattering and absorption coefficients. The total extinction coefficient is $k(\nu, s') = \alpha(\nu, s') + \sigma(\nu, s')$.

Absorption by molecules in the Earth's atmosphere is due to radiatively active trace gases. For UVR and visible light the most important gas is ozone, but oxides of sulfur and nitrogen may (depending on location) have a significant effect on the UVR penetration. The molecular absorption coefficient is: $\alpha_m(\lambda) = \alpha_{n, \text{O}_3} n_{\text{O}_3}$, where the subscript m stands for molecules, and α_{n, O_3} and n_{O_3} are the ozone absorption cross section and number density, respectively (Figure 2(a) and 2(b)). At wavelengths longer than 310 nm the ozone absorption is due to the *Huggins bands*, whereas the spectrally broad but weak absorption between 450 and 700 nm is due to the *Chappuis bands* (Figure 2(a)). Scattering by molecules in the atmosphere is proportional to the gas density. Thus, the scattering coefficient due to scattering by air molecules is: $\sigma_m(\lambda) = \sigma_{n, \text{Ray}}(\lambda) n_m$, where $\sigma_{n, \text{Ray}}(\lambda) \sim \lambda^{-4}$ is the Rayleigh scattering cross section, and n_m is the total air number density (Figure 2(b)).

Absorption by pure water results from mutual interactions between the intermolecular forces. Calculation of absorption cross sections from first principles is very difficult. Thus, laboratory and *in situ* measurements become essential for establishing the absorption coefficient $\alpha_{n, w}(\lambda)$ for pure water (Figure 2(c)). Scattering in pure water is due to clusters of molecules and ions, much smaller than the wavelength of light, and from an optical point of view the clusters are uncorrelated. Thus, the wavelength dependence of the scattering cross section for pure water is similar to the Rayleigh scattering cross section for molecules in the atmosphere, and the scattering coefficient for pure water can be approximated by $\sigma_w(\lambda) = \left(\frac{129}{\lambda}\right)^{4.32}$, λ in (nm).

Gaseous Absorption and Penetration Depth

Figure 3 shows that significant O_2 absorption occurs in the *Schumann-Runge bands* between 195 and 175 nm. At shorter wavelengths this feature merges into the very intense *Schumann-Runge continuum* (130–175 nm).

An instructive way of looking at the effects of gaseous absorption on incoming radiation is through the concept of *penetration depth*, defined as the height at which the solar radiation reaches optical depth unity, for a clear atmosphere exposed to an overhead Sun. The UV penetration depth for wavelengths shorter than 320 nm is shown in Figure 4. The smallest penetration depths, occurring in the thermosphere, arise from the very high absorption coefficients of air in the X-ray ($\lambda < 10$ nm) and extreme-UV ($10 < \lambda < 100$ nm) spectral range. The far-UV solar radiation between 100 and 200 nm (see Table 1) is absorbed in the lower thermosphere and mesosphere (50–120 km). Middle-UV, or UV-C ($200 < \lambda < 280$ nm) atmospheric absorption is dominated by the intense *Hartley bands* due to O_3 . In the absence of O_3 , the Earth's land and ocean surface would be irradiated directly by UV-C to the detriment, if not the total eradication of surface life. The absorption of this important energy band causes the inversion in the Earth's stratospheric (15–50 km) temperature profile.

Light at UV wavelengths penetrating to the troposphere and surface, is customarily divided into two groups: UV-B (280–320 nm) and UV-A (320–400 nm). Living organisms are much more sensitive to damage by UV-B radiation than by radiation in the more benign UV-A region. Penetration of UV-B radiation is very sensitive to the total column abundance of O_3 . At longer wavelengths, the photosynthetically active radiation (PAR) spans the spectral range between 400 and 700 nm. This spectral range has the greatest clear-air transparency (except for radio waves), because it contains only the weak absorption features of O_2 and O_3 . The eyes of humans (and animals) are most sensitive to light in this spectral range, a fact of extreme importance for evolutionary adaptation.

UV Light Transport in the Atmosphere–Ocean System

If we are interested primarily in energy transport, it is sufficient to consider the azimuthally averaged radiance $I_\nu(z, u)$, where z

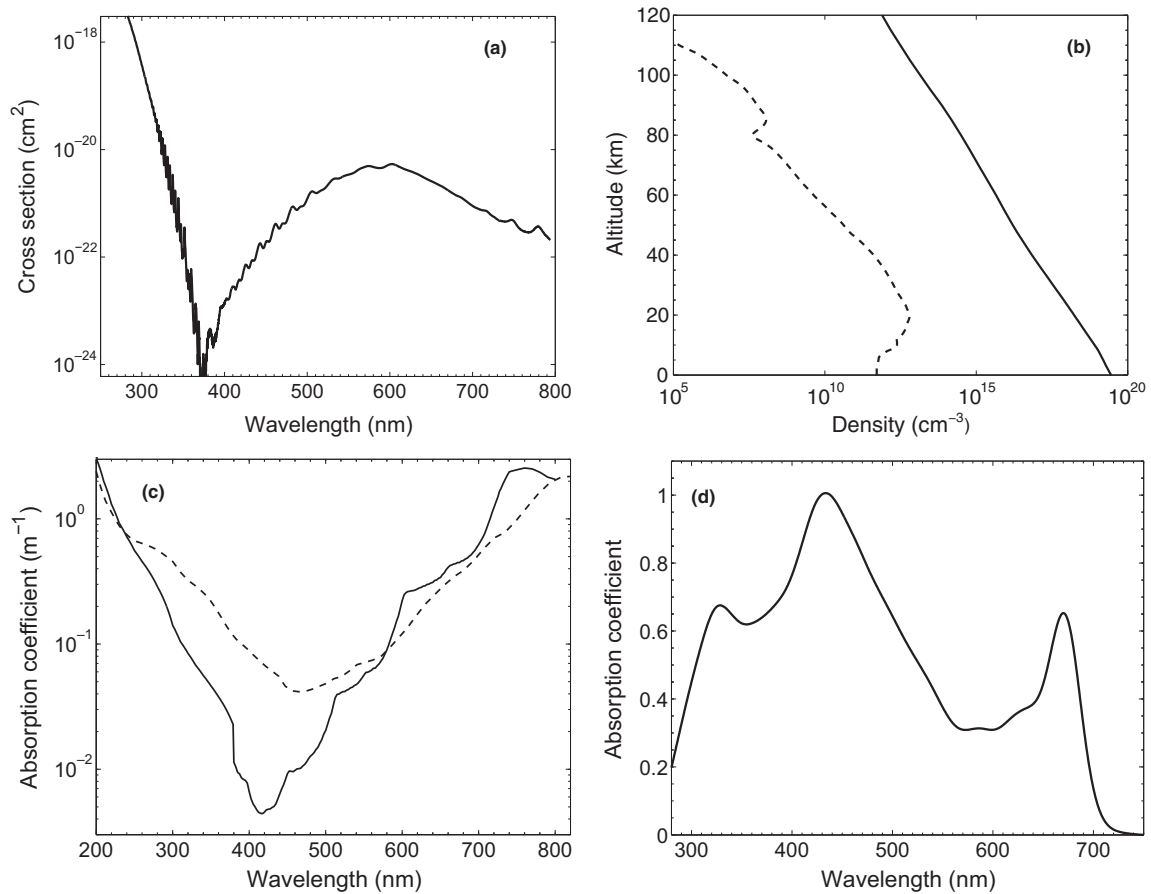


Figure 2 (a) Absorption cross section of ozone. (b) Number density of atmospheric ozone (dashed line) and total air number density (solid line). (c) Absorption spectrum for pure ice (dashed line), and for pure water (solid line). (d) Chlorophyll-specific absorption spectrum normalized at 440 nm.

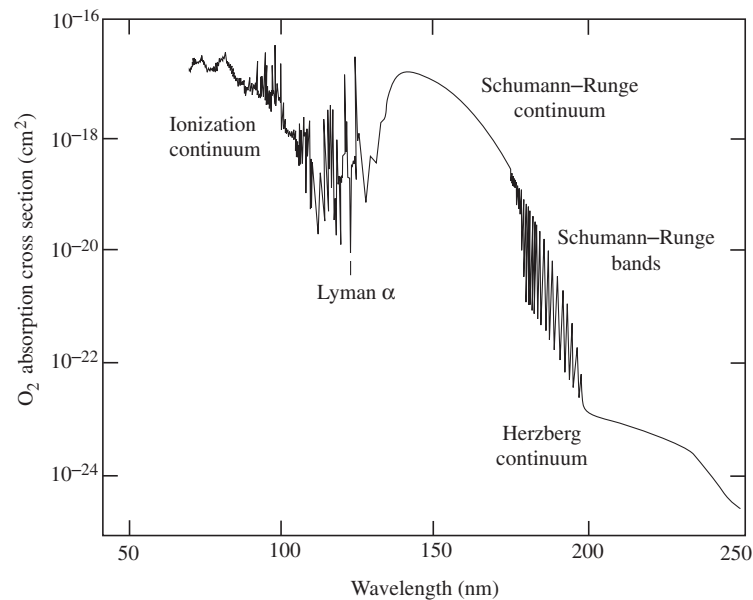


Figure 3 O_2 absorption cross section illustrating the various band and continua (see text). The deep absorption line at 121.6 nm corresponds almost exactly with the strong solar hydrogen Lyman- α line. Adapted with permission from Thomas, G.E., Stamnes, K., 1999. Radiative Transfer in the Atmosphere and Ocean. Cambridge University Press, Cambridge.

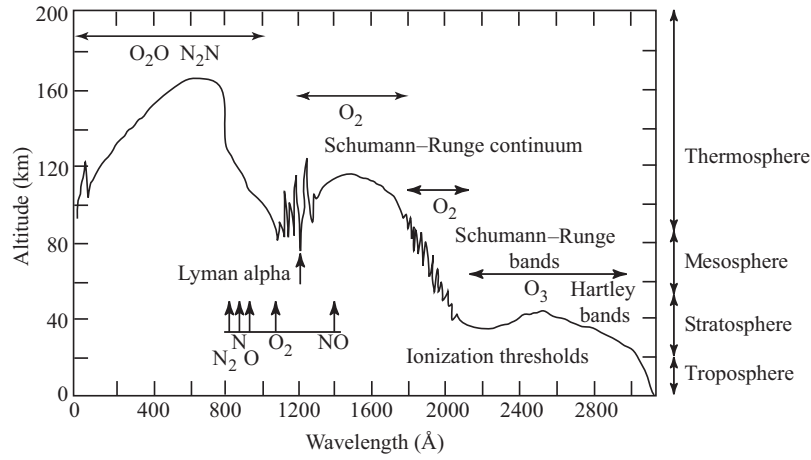


Figure 4 Atmospheric penetration depth vs wavelength. Horizontal arrows indicate the molecule (and band) responsible for absorption in that spectral region. Vertical arrows indicate the ionization thresholds of the various species. Adapted with permission from Thomas, G.E., Stamnes, K., 1999. Radiative Transfer in the Atmosphere and Ocean. Cambridge University Press, Cambridge.

denotes the level in the medium (height in the atmosphere or depth on the ocean), and $u = \cos \theta$, θ being the polar angle. It is convenient to split the radiation field into two parts: (1) the *direct* solar beam, which is exponentially attenuated upon passage through the atmosphere and ocean, and (2) the *diffuse* or scattered radiation.

According to eqn [2] the penetration of the direct solar beam through the atmosphere may be written as (dropping the subscript ν) $I_{\text{sol}}(z) = I_{\text{sol}}(\tau(z)) = F^s e^{-\tau}$. Here, F^s is the solar irradiance (normal to the solar beam direction) incident at the top of the atmosphere. Unless the Sun is close to the horizon, it is sufficient to use *plane geometry*, assuming that the atmosphere

and the ocean are stratified media, for which the optical properties vary only in the vertical. Then the *vertical* optical depth τ_ν is defined as $\tau_\nu(z) \equiv \tau_s \mu_0 \equiv \sum_i \int_z^\infty dz' k(\nu, z')$, or $d\tau_\nu(z) = -k(\nu, z) dz$, where $\mu_0 = \cos \theta_0$, and θ_0 is the solar zenith angle as illustrated in Figure 5. In plane geometry, the diffuse radiation in the medium is described by:

$$u \frac{dI_\nu(z, u)}{dz} = -k(\nu, z) I_\nu(z, u) + \frac{\sigma(\nu, z)}{2} \int_{-1}^1 du' p(z, u', u) I_\nu(z, u') + S_\nu^*(z, u). \quad [3]$$

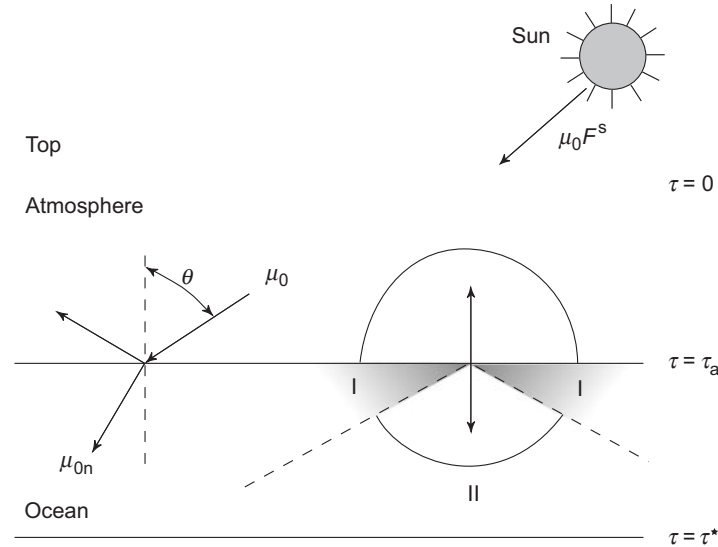


Figure 5 Schematic illustration of two adjacent media with a flat interface such as the atmosphere overlying a calm ocean. The atmosphere has a different index of refraction ($m_r \approx 1$) than the ocean ($m_r \approx 1.33$). Adapted with permission from Thomas, G.E., Stamnes, K., 1999. Radiative Transfer in the Atmosphere and Ocean. Cambridge University Press, Cambridge.

Equation [3] has a simple physical interpretation. The term on the left side is the change in the radiance along the slant path dz/u . The first term on the right side is the loss of radiation out of the beam due to extinction. The second term is the gain due to multiple scattering, and the normalized angular scattering cross section, $p(z, u', u)$, gives the probability that light is scattered from direction u' into direction u . The third term is the solar pseudo-source, proportional to the attenuated solar beam, which 'drives' the diffuse radiation. Instead of the geometrical distance dz it is customary to use the nondimensional optical depth, $d\tau(z) = -k(z)dz$ as the independent variable. Then, we may rewrite eqn [3] as follows (dropping the subscript ν):

$$u \frac{dI(\tau, u)}{d\tau} = I(\tau, u) - \frac{a(\tau)}{2} \int_{-1}^1 du' p(\tau, u', u) I(\tau, u') - S^*(\tau, u) \quad [4]$$

where the single-scattering albedo is defined as $a(\tau(z)) \equiv \sigma(z)/k(z)$. The source term $S^*(\tau, u)$ is defined below.

Surface Reflection and Transmission

Knowledge of the reflectance of underlying land and ocean surfaces is necessary for calculating the diffuse radiation field. Also, in shallow waters, the reflectance of the ocean bottom influences the diffuse radiation in the water and the radiation leaving the water surface. The reflectance and transmittance of a surface depend upon both the angles of incidence and reflection or transmission. In atmospheric radiative transfer, it is frequently assumed that the underlying land or ocean surface reflects incoming radiation isotropically. Such a surface is called a *Lambert* reflector, and the reflectance is called the surface albedo, ρ_L . However, most natural surfaces are non-Lambertian. Thus, we could relax the assumption of a Lambert reflector and use instead the bidirectional reflection distribution function, if this function is known. For the coupled atmosphere-ocean system it is preferable to consider two strata, one for the atmosphere, and one for the ocean, but with different indices of refraction. The basic radiance, defined as I/m^2 where m is the real part of the index of refraction, is the relevant quantity when we consider two strata with different values of m . If we assume for simplicity that the interface is flat and smooth (a calm ocean), then the basic radiance must satisfy Snell's law and Fresnel's equations. As illustrated in Figure 5, the downward radiation distributed over 2π sr in the atmosphere will be restricted to a cone less than 2π sr (referred to as region II in Figure 5) after being refracted across the interface into the ocean. Beams outside the refractive region in the ocean are in the total reflection region (referred to as region I in Figure 5). The demarcation between the refractive and the total reflection region in the ocean is given by the critical angle $\mu_c = \sqrt{1 - 1/m_r^2}$, where $m_r = m_{\text{ocean}}/m_{\text{air}}$ is the index of refraction of the ocean relative to the air.

Since the radiation field in the ocean is driven by solar radiation passing through the air-water interface, eqn [4]

applies also in the ocean, but the solar pseudo-sources are different. In the atmosphere, we have

$$S_{\text{air}}^*(\tau, u) = \frac{a(\tau)}{4\pi} p(\tau, -\mu_0, u) e^{-\tau/\mu_0} + \frac{a(\tau)}{4\pi} \rho_s(-\mu_0; m_r) p(\tau, \mu_0, u) e^{-(2\tau_s - \tau)/\mu_0} \quad [5]$$

where the first term is the usual solar beam pseudo-source, and the second term is due to the reflection occurring at the interface, which is proportional to $\rho_s(-\mu_0; m_r)$, the specular reflection coefficient. The pseudo-source in the ocean is just the attenuated solar beam refracted through the interface:

$$S_{\text{ocean}}^*(\tau, u) = \frac{a(\tau)}{4\pi} \frac{\mu_0}{\mu_t} \tau_s(-\mu_0; m_r) p(\tau, -\mu_t, u) e^{-\tau/\mu_0} e^{-(\tau - \tau_s)/\mu_t} \quad [6]$$

where $\tau_s(-\mu_0; m_r)$ is the beam transmittance through the interface, and μ_t is the cosine of the solar zenith angle in the ocean, which is related to μ_0 by Snell's law $\mu_t \equiv \mu_t(\mu_0, m_r) = \sqrt{1 - (1 - \mu_0^2)/m_r^2}$. With the use of the appropriate pseudo-sources for the atmosphere [$S_{\text{air}}^*(\tau, u)$] and ocean [$S_{\text{ocean}}^*(\tau, u)$] eqn [4] can now be solved subject to boundary conditions at the top of the atmosphere and the bottom of the ocean. However, in addition we must properly account for the reflection and transmission through the interface by requiring that I/m^2 satisfies Snell's law and Fresnel's equations.

Photolysis Rate and Biological Dose Rate

In atmospheric photochemistry, the *photolysis rate* is defined as the local rate of a photoabsorption process leading to a specific photodissociation event. We express the photolysis rate coefficient for the photodissociation of a particular species i as follows:

$$J_i \equiv \int_{\nu_c}^{\infty} d\nu 4\pi (\bar{I}_\nu / h\nu) \alpha^i(\nu) \eta^i(\nu) \quad [s^{-1}]. \quad [7]$$

Here, $\alpha^i(\nu)$ is the photoabsorption cross section, $\eta^i(\nu)$ ($0 \leq \eta^i \leq 1$) is the *quantum yield* or efficiency by which the absorbed radiative energy produces the photodissociation, ν_c is the minimum frequency corresponding to the threshold energy for the photoabsorption, $\bar{I}$ is the mean intensity, and $4\pi \bar{I}_\nu / h\nu$ is the density of photons at a given frequency. Photochemists frequently use the term *actinic flux* for the quantity $4\pi \bar{I}_\nu$.

The rate at which a surface receives radiative energy capable of initiating certain biological processes is obtained by weighting the received radiation by a specific spectral function $A(\nu) < 1$ called the *action spectrum*, which gives the efficiency of a particular process, for example, the UV 'kill rate.' The rate at which a flat surface is 'exposed' is called the *dose rate*

$$D \equiv \int_0^{\infty} d\nu A(\nu) F_\nu^- \quad [W \cdot m^{-2}] \quad [8]$$

where F_ν^- is the incident irradiance. The *radiation dose* is defined to be the total time-integrated amount of energy received (usually over 1 day) $\int dt D(t)$.

Modeling UV Penetration into the Ocean

UV penetration in the ocean is strongly influenced by small plankton and thus by biological productivity, which provides a close link between biological and optical oceanography. Important aspects of the ozone depletion issue include the effects of increased UV levels on algae, plankton, and fish larvae. As sources of atmospheric sulfur compounds involved in cloud formation, plankton may indirectly affect atmospheric transmission, thereby linking atmospheric radiative transfer with ocean biology.

The impact of decreased ozone levels on aquatic ecosystems may be assessed by a radiation model that provides a solution of the radiative transfer equation for the coupled atmosphere–ocean system. The atmosphere is assumed to be separated from the ocean by a plane interface as explained above. The relative refractive index for air is taken to be unity and that of the ocean to be 1.33. To account for the vertical inhomogeneity of the atmosphere and water, the two media are divided into a sufficient number of layers to resolve the changes in optical properties with height in the atmosphere and depth in the ocean. To estimate the UV penetration through this coupled atmosphere–ocean system, we need the spectral distribution of the radiation incident at the top of the atmosphere (see Figure 1) as well as the optical properties of the atmosphere and water media. For a clear atmosphere, the optical properties are determined mainly by ozone absorption (see Figure 3), and molecular (Rayleigh) scattering.

Phytoplankton dwell in the top layers of the water column, the euphotic zone, because of their requirement for PAR radiation to drive photosynthesis. In the euphotic zone, they would be exposed to any increase in UVR. If all other factors remain constant, ozone depletion would lead to increased transmission of UV-B radiation through the atmosphere and into the water column. Model results indicate that UV-B radiation is significantly absorbed in the first couple of meters into the water. At high latitudes (70°), an ozone depletion of 30% (compared to normal) will increase UV-B exposure 10 m below the surface by as much as 33% on 1 October in the Southern Ocean, and by 23% at summer solstice. Thus, the relative amount of UV-B increase in the water column due to ozone depletion is most pronounced in spring, which happens to be the time when polar ozone depletion is most severe. Chlorophyll pigment reduces the penetration of UV-B radiation into the water. The larger the chlorophyll concentration, the less the UV-B transmission; however, the UV-B to PAR ratio is relatively unaffected by the chlorophyll concentration. Thus, species that depend on a certain level of PAR for photosynthesis and therefore adjust their depth in the water to optimize PAR will be exposed to a similar level of UV-B regardless of chlorophyll content.

Measured and Computed UV Irradiance in the Ocean

Underwater spectral irradiance was measured *in situ* with a UV/visible spectrometer submersed into the ocean off Palmer Peninsula, Antarctica. Figure 6 compares the measured and

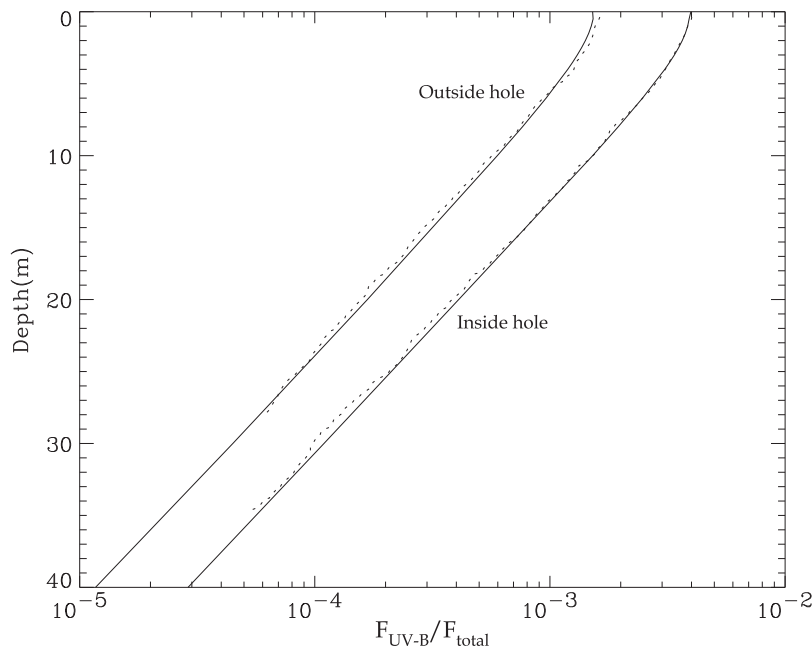


Figure 6 Comparison between model computations (solid lines) and measurements (dotted lines) of depth vs $F_{\text{UV-B}}/F_{\text{total}}$. Inside the ozone hole, the ozone abundance was 150 DU, the solar zenith angle was 56° , and the vertical distribution of chlorophyll concentration was 0.57 mg m^{-3} from the surface to 20 m depth, 0.47 mg m^{-3} below 20 m. Outside the ozone hole, the ozone abundance was 350 DU, the solar zenith angle was 57° , and the vertical distribution of chlorophyll concentration was 1.9 mg m^{-3} from the surface to 10 m depth, 1.6 mg m^{-3} from 10 to 20 m, and 1.5 mg m^{-3} below 20 m.

computed ratio of the irradiance integrated across the UV-B range ($F_{\text{UV-B}}$, 280–320 nm) to that integrated across the complete measured range (F_{total} , 280–700 nm). Ratios for both undepleted ozone levels (350 DU, labeled ‘outside hole’) and depleted levels (150 DU, labeled ‘inside hole’) are shown. How well can this model reproduce the underwater downwelling irradiance, when it is constrained to yield the correct surface irradiance by adjusting the cloud/aerosol optical depth (which was not measured)? We note that there is good agreement between computed and measured ratios below the surface, but some curvature in the computed ratio just below the surface is absent in the ratio inferred from the measurements. A possible reason for this discrepancy is the neglect of surface waves in the model, which assumes a plane atmosphere–ocean interface.

Solar zenith angle has an important influence on UVR reaching the earth’s surface. It was overcast during the measurements, but the solar zenith angle was almost the same for the measurements taken inside and outside the ozone hole. Since the optical properties of clouds and aerosols depend weakly on wavelength, the impact of clouds and aerosols on the ratio ($F_{\text{UV-B}}/F_{\text{total}}$) is expected to be small. This circumstance allows us to investigate the impact of changes in ozone abundance on the surface and submarine irradiance ratio. Ozone depletion will increase the surface and underwater UV irradiance. Although the vertical distributions of chlorophyll in the water were different under and outside the ozone hole, the impact of this difference on the vertical variation in the ratio $F_{\text{UV-B}}/F_{\text{total}}$ is small. Thus, the vertical attenuation coefficients are nearly the same inside and outside the ozone hole. Therefore, if UV-B exposure is doubled at the surface, it will be doubled at all depths, and the critical depth above which UV-B damage may occur will be correspondingly deeper in the water column.

Impact of Ozone Depletion on Primary Productivity

Many marine organisms are sensitive to UVR. The extent to which these marine organisms will be able to adapt to expected increases in UV exposure is uncertain due to a sparsity of measurements. Increased levels of UV-B radiation may impact phytoplankton communities by (1) initiating changes in cell size and taxonomic structure, (2) reducing the productivity, (3) influencing the protein content, dry weight, and pigment concentration, (4) inducing chloroplast damage, and (5) directly affecting the proteins of the photosynthetic apparatus.

To explore the potential impact of ozone depletions, one may use a simple model, in which marine photosynthesis (or primary production, ignoring respiratory losses) is parameterized as follows

$$P = P_s \left(1 - e^{I_p/I_s}\right) \frac{1}{1 + I^*} \quad [9]$$

where P_s is the maximum photosynthesis in the absence of UVR inhibition, and I_p is the photosynthetically utilizable radiation (PUR):

$$I_p = \int_{400}^{700} \bar{I}(\lambda) a^*(\lambda) d\lambda. \quad [10]$$

Here, $\bar{I}(\lambda)$ is the mean intensity (scalar irradiance) and $a^*(\lambda)$ is the normalized algal absorption spectrum. In eqn [9], I_s is the PUR saturation level, and $1/(1 + I^*)$ describes the inhibition caused by UVR, where I^* is

$$I^* = \int_{280}^{400} \varepsilon(\lambda) \bar{I}(\lambda) d\lambda, \quad [11]$$

and $\varepsilon(\lambda)$ is the action spectrum for UVR-induced inhibition of photosynthesis.

The solar irradiance $\bar{I}(\lambda)$ that drives and inhibits primary photosynthesis depends on several environmental factors including the total ozone column amount, solar elevation, and depth in the ocean, as well as the presence of clouds, ice, and snow. Employing the radiative transfer model described above for the coupled atmosphere–sea ice–ocean system, and assuming total ozone column amounts of 400 and 200 DU, one may compute $\bar{I}(\lambda)$ at the bottom of the snow-covered sea ice and at various depths of open water, and the resulting $\bar{I}(\lambda)$ values may be used in eqn [9] to calculate the primary productivity. The results of such a study shows that (1) an ozone depletion increases not only harmful UVR, but also beneficial PUR; (2) at high latitudes the benefits of increased PUR for phytoplankton under sea ice and below a certain depth in the ocean dominates over the damage caused by increased UVR; (3) a large fraction of the primary production in the polar regions is caused by ice algae growing in environments well protected from UVR; and (4) the primary production in the polar regions could increase by as much as 1% for a 50% ozone depletion.

See also: Ozone Depletion and Related Topics: Ozone as a UV Filter. **Radiation Transfer in the Atmosphere:** Radiation, Solar.

Further Reading

- Gao, W., Schmoldt, D., Slusser, J.R. (Eds.), 2009. *UV Radiation in Global Climate Change: Measurements, Modeling and Effects on Ecosystems*. Springer-Verlag GmbH Berlin, Heidelberg and Tsinghua University Press, Beijing.
- Hamre, B., Stamnes, J.J., Frette, Ø., Erga, S.R., Stamnes, K., 2008. Could stratospheric ozone depletion lead to enhanced aquatic primary production in the polar regions? *Limnology and Oceanography* 53, 332–338.
- Kirk, J.T.O., 1994. *Light and Photosynthesis in Aquatic Ecosystems*, second ed. Cambridge University Press, Cambridge, NY.
- Meier, R.R., Anderson, G.P., Cantrell, C.A., Hall, L.A., Lean, J., Minschwaner, K., Shetter, R.E., Shettle, E.P., Stamnes, K., 1997. Actinic radiation in the terrestrial atmosphere. *Journal of Atmospheric and Solar-Terrestrial Physics* 59, 2111–2157.
- Rottman, G.J., Woods, T.N., Sparr, T.P., 1993. Solar stellar irradiance comparison experiment I. Instrument design and operation. *Journal of Geophysical Research* 98, 10667–10678.
- Smith, R.C., Prezelin, B.B., Baker, K.S., Bidigare, R.R., Boucher, N.P., Coley, T., Karentz, D., McIntyre, S., Matlick, H.A., Menzies, D., Ondrusek, M., Wan, Z., Waters, K.J., 1992. Ozone depletion: ultraviolet radiation and phytoplankton biology in Antarctic waters. *Science* 255, 952–959.
- Thomas, G.E., Stamnes, K., 1999. *Radiative Transfer in the Atmosphere and Ocean*. Cambridge University Press, Cambridge, UK.
- Zerefos, C.S., Bais, A.F. (Eds.), 1997. *Solar Ultraviolet Radiation: Modeling, Measurements and Effects*. Springer-Verlag, Berlin.

Ultraviolet, Surface

R McKenzie, National Institute for Water and Atmospheric Research, Lauder, Central Otago, New Zealand
S Madronich, National Center for Atmospheric Research, Boulder, CO, USA

© 2003 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 6, pp 2474–2480, © 2003, Elsevier Ltd.

Introduction

The extraterrestrial solar radiation is dominated by visible and near-infrared wavelengths, and only a small portion of the energy is within the UV region. Furthermore, only a fraction of this UV radiation incident at the top of the atmosphere is transmitted to the Earth's surface. This transmitted UV radiation is nevertheless of great importance, because the photon energies in this spectral region are sufficient to damage bonds in molecules such as DNA, the 'building block of life'. This has led to concerns that increases in UV radiation resulting from ozone depletion will have further detrimental environmental effects. These include health effects in humans (e.g., damage to skin and eyes), damage to aquatic and terrestrial plants and ecosystems, and damage to materials. Small changes in ozone can lead to relatively large changes in UV radiation reaching the Earth's surface: it has been estimated that for each 1% reduction in ozone, the incidence rate for skin cancer will increase by 2–3% if all other factors remain unchanged.

Here the factors controlling UV irradiance at the Earth's surface are discussed. The concept of weighted UV irradiances is introduced along with methods to assess the sensitivity to ozone change. These are applied particularly to the recently defined UV Index – a standardized measure used to disseminate UV information to the public. Geographical and temporal changes in weighted UV are presented.

Solar UV Radiation and Its Transmission Through the Earth's Atmosphere

UV radiation is subdivided into three wavelength bands: UV-A (315–400 nm) radiation, UV-B (280–315 nm) radiation (some authors use 320 nm as the boundary between UV-A and UV-C radiation), and UV-C (100–290 nm) radiation. UV-A radiation is largely unaffected by ozone. UV-C comprises less than 0.6% of the incident solar spectrum at the top of the atmosphere, and it is effectively absorbed high in the Earth's atmosphere by oxygen (below ~240 nm) and by ozone (below 280 nm). UV-B radiation penetrates to the surface, but it is strongly absorbed by atmospheric ozone, whose

absorption cross-section increases rapidly towards shorter wavelengths in this spectral region. Consequently UV-B radiation received at the Earth's surface is highly variable and shows a strong dependence on solar zenith angle (SZA) and ozone amount (Table 1), leading to large geographical and temporal variabilities. As with other spectral regions of solar radiation, it is also modulated by other factors, as follows:

1. Variability in solar output. The greatest changes in solar output occur in UV-C radiation, which is not transmitted to the Earth's surface. Paradoxically, increases in solar output (e.g., associated with 27-day solar rotation, or the 11-year solar cycle) usually produce slight reductions in UV at the surface because they result in increased ozone production.
2. Seasonal differences in Sun–Earth separation. For a given solar zenith angle, irradiances in December and January will be approximately 7% larger than in June and July owing to the elliptical orbit of the Earth about the Sun.
3. Scattering and absorption in clouds. This generally reduces UV radiation at the surface, but under partly cloudy conditions with the Sun unobscured, significant increases are also possible.
4. Scattering and absorption by aerosols. This generally reduces UV radiation.
5. Rayleigh scattering by air molecules. This is much more important in the UV region than the visible region since it has a wavelength dependence.
6. Reflections from the surface. In the UV most natural surfaces have reflectivities (or albedo) less than 5%. However, fresh snow can have a reflectivity of >90%, leading to a greatly increased diffuse UV contribution.
7. Altitude. In unpolluted conditions, UV-B irradiance increases by approximately 5% per kilometer because of reduced Rayleigh scattering. This rate of increase can be increased appreciably when the effects of snow cover and boundary layer extinctions are included.

Because of its strong dependence on sun angle there are large seasonal variations in UV radiation, which become larger with increasing latitude. Seasonal and latitudinal variations in ozone are also important. Globally, the average total

Table 1 Approximate contributions (W m^{-2}) to solar energy from the UV-A and UV-B regions at 1 AU (cloudless skies, no aerosols; solar constant = 1370 W m^{-2})

Conditions	UV-B (280–315 nm)	UV-B (280–320 nm)	UV-A (315–400 nm)
Extraterrestrial	17.7	21.3	86.4
Earth surface, SZA = 0°, ozone = 300 DU	2.23	4.07	66.5
Earth surface, SZA = 60°, ozone = 300 DU	0.41	0.95	27.0
Earth surface, SZA = 60°, ozone = 450 DU	0.22	0.63	26.7
Earth surface, SZA = 60°, ozone = 100 DU	1.17	1.95	27.6

1 AU = 1 Astronomical Unit—the mean Earth–Sun separation.

ozone column amount is approximately 300 Dobson units ($1 \text{ DU} = 2.69 \times 10^{16} \text{ molecule cm}^{-2}$), with a normal range of around 250 to 450 DU. In recent years, there has been a significant reduction in ozone. Ozone column amounts below 100 DU have been observed in the Antarctic spring. However, they occur only at relatively large solar zenith angles, so although UV intensities have increased greatly, they usually remain moderate by global standards. The UV intensities are greater in the tropics, where noon time solar zenith angles are smaller and where total column ozone amounts tend to be smaller. The highest intensities occur at high altitudes (particularly over snow-covered surfaces) in the tropics, where losses by Rayleigh and aerosol scattering are minimized.

Exposure to UV Radiation

The human eye responds to light with wavelengths from about 790 nm (red) to 430 nm (violet). Ultraviolet (beyond-violet) wavelengths are shorter than the human eye 'sees'. Sunlight is received as direct rays and as diffuse light, i.e. skylight which has been scattered by the atmosphere. In the visible region the direct beam component dominates over the diffuse component whereas in the UV-B region, where scattering processes are more efficient, the diffuse component is usually dominant. The sky is blue because air molecules scatter the shorter-wavelength blue light more than the red light. UV radiation is scattered even more than blue light. If we could 'see' in the UV, the Sun would appear as a dull disk in a uniformly bright sky. Being shaded from the Sun's direct rays provides only partial protection from UV exposure because of the high level of diffuse UV light. In the visible region most natural surfaces have relatively high reflectivity, so that reflected light makes an important contribution, whereas in the UV region this is not the case.

Small doses of UV radiation are beneficial to humans, leading for example to the production of vitamin D. However, larger doses can be harmful, leading to ocular damage, immune suppression, and damage to the skin. Skin-damaging effects include ageing, erythema (skin reddening), sunburn, and some forms of skin cancer. For overhead sun conditions and for relatively low ozone amounts (e.g. in the tropics), damage to the most sensitive skin types can occur with exposures times of ~ 10 min.

Typical window glasses transmit less than 10% of sun-burning ultraviolet light, and sun block creams work by absorbing or reflecting UV rays. The SPF rating of sunscreens gives an indication of their effectiveness as UV blockers. For example, an SPF of 15 means that it should take 15 times as long before skin damage occurs as it would for damage to occur to unprotected skin (i.e., the cream, when correctly applied, should block $\sim 93\%$ of the radiation that causes skin damage).

Spectral Irradiance and UV Index

When the spectrum of UV radiation is examined in detail it becomes apparent that there are large variabilities within the broad categories discussed above. Typical clear sky spectra are shown in Figure 1(a) (note the logarithmic y-axis scale) for overhead sun assuming no ozone absorption, and for two solar

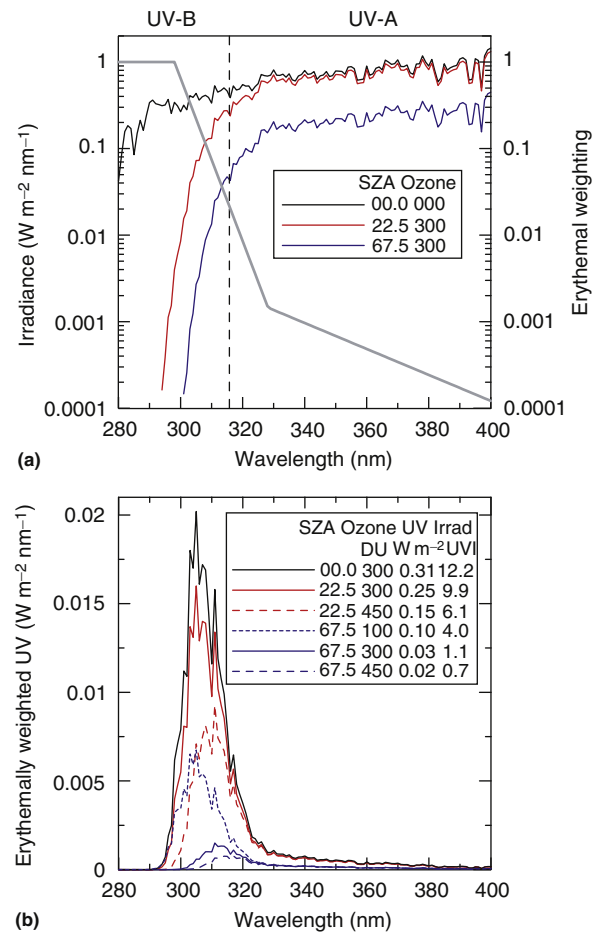


Figure 1 (a) Clear-sky spectra (at 1 AU) for no ozone and overhead sun (black), and for 300 DU ozone to represent the spectral irradiance that would be seen at solar zenith angles of 22.5° and 67.5° (red and blue respectively, corresponding to noon at midwinter and midsummer at a midlatitude site). The gray line is the erythemal action spectrum defined in the text, which is normalized to unity $\lambda \leq 298 \text{ nm}$ (unit less). (b) The spectra weighted by the erythemal action spectrum. Also shown are weighted irradiances for 100 and 450 DU. Corresponding values of the erythemally weighted UV Irradiance and UV Index are given.

zenith angles with 300 DU ozone. These two solar zenith angles correspond to noon time at midlatitudes in the summer and winter, and demonstrate that during the summer months midlatitude sites experience UV comparable to that in the tropics, but in winter the irradiances at midlatitudes are much less. At wavelengths less than 290 nm, absorption by ozone has reduced the irradiances by ~ 5 orders of magnitude. The detailed structure in the spectra is due principally to absorptions by gases in the solar atmosphere (Fraunhofer lines).

For many biological processes, the damaging effects of UV radiation become progressively more severe at shorter wavelengths. One commonly used weighting function to describe this effect is the so-called erythemal action spectrum for damage to human skin (Figure 1(a)). Spectra of erythemally weighted UV irradiances are shown in Figure 1(b). The peak contributions occur in the UV-B wavelength range, with the relative contribution from the UV-A becoming more important at larger SZA and larger ozone amounts. The weighted

irradiance for $\text{SZA} = 22.5^\circ$ are nearly a factor of 10 more than for $\text{SZA} = 67.5^\circ$, a difference much larger than if the ozone is changed from 300 DU to 450 DU. The curve for $\text{SZA} = 67.5^\circ$, ozone = 100 DU gives an indication of the weighted irradiance under the Antarctic ozone hole conditions. With this extremely low ozone amount there is a shift in the peak to shorter wavelengths.

Information on UV intensities is disseminated to the public in terms of the erythemally weighted UV irradiance (Figure 1(b)). The erythral weighting function applied to the spectrum involves an arbitrary normalization to unity at wavelengths shorter than 298 nm, so erythemally weighted UV is not strictly defined in terms of an SI unit. Furthermore, when UV information is provided to the public, another normalization is applied to provide a number, called the UV Index. Historically, the reason for this normalization choice was to give a maximum UV Index of about 10 in Canada, where the unit was first used.

$$\text{UV Index} = 40 \int I(\lambda)w(\lambda)d\lambda$$

where λ is the wavelength in nm, $I(\lambda)$ the irradiance in W m^{-2} , and $w(\lambda)$ the erythral weighting function defined (Figure 1(a)) as:

$$\begin{aligned} w(\lambda) &= 1.0 && \text{for } 250 < \lambda \leq 298 \text{ nm;} \\ w(\lambda) &= 10^{0.094(298-\lambda)} && \text{for } 298 < \lambda \leq 328 \text{ nm;} \\ w(\lambda) &= 10^{0.015(139-\lambda)} && \text{for } 328 < \lambda \leq 400 \text{ nm;} \\ w(\lambda) &= 0 && \text{for } \lambda > 400 \text{ nm.} \end{aligned}$$

While there may be uncertainties about the biological applicability of this erythral action spectrum, its great advantage is that it is mathematically defined and therefore its detailed shape is unambiguous. This is important in the UV region, where the steeply sloping spectrum spans several orders of magnitude (Figure 1(a)). Although the UV Index was developed to represent damage to human skin, it may be applied to other processes, since many biological UV effects have similar action spectra.

In reality, the UV Index is an open-ended scale, and outside the protection of the Earth's atmosphere the UV Index is ~ 300 (depending on the lower wavelength limit of the integration). Table 2 shows how the UV Index varies as a function of ozone and solar zenith angle. When applying this table to geographic regions, it is also necessary to apply corrections for variations in Sun–Earth separation.

Sensitivity of UV to Ozone Change

The sensitivity of UV radiation to atmospheric ozone is sometimes expressed as a so-called radiation amplification factor (RAF), defined as

$$\text{RAF} \equiv (\delta R/R)/(-\delta \Omega/\Omega)$$

where R and δR are, respectively, the dose rate (i.e., irradiance) and its change when the ozone column Ω changes by a small amount $\delta \Omega$.

For larger changes it is more accurate to use a power-law formulation:

$$R_2/R_1 = (\Omega_2/\Omega_1)^{-\text{RAF}}$$

where R_1 and R_2 are the dose rates for ozone columns Ω_1 and Ω_2 respectively.

The RAFs for several action spectra and weighting of UV radiation, computed for clear skies, are shown in Table 3 (processes with small RAFs are omitted).

Global and Regional Patterns

Estimates of UV irradiance are available from direct measurements at the surface, or from model calculations. In recent years satellite measurements of ozone have been used in conjunction with satellite-derived cloud parameters (e.g., reflectances) to estimate UV irradiances and UV dosage at the Earth's surface. For example, Figure 2 shows a global mean annual climatology of erythral UV dose, including cloud effects based on 14 years of measurements between 1979 and 1992 from the Total Ozone Mapping Spectrometer (TOMS) instrument on the NIMBUS-7 satellite. The largest doses occur at the Equator, particularly in cloud-free and high altitude regions. In polar regions, the mean daily doses are small because of the low Sun elevations experienced at these locations. However, during the period of the springtime ozone hole, UV doses in Antarctica can be comparable to those at midlatitudes.

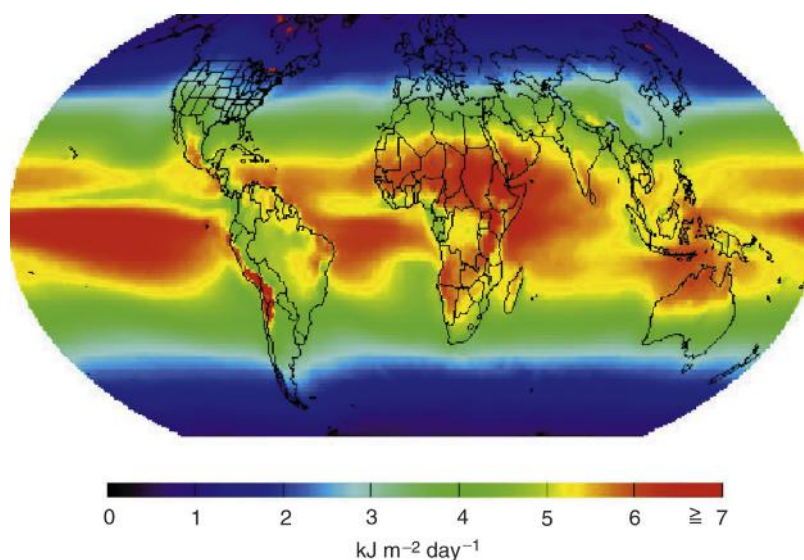
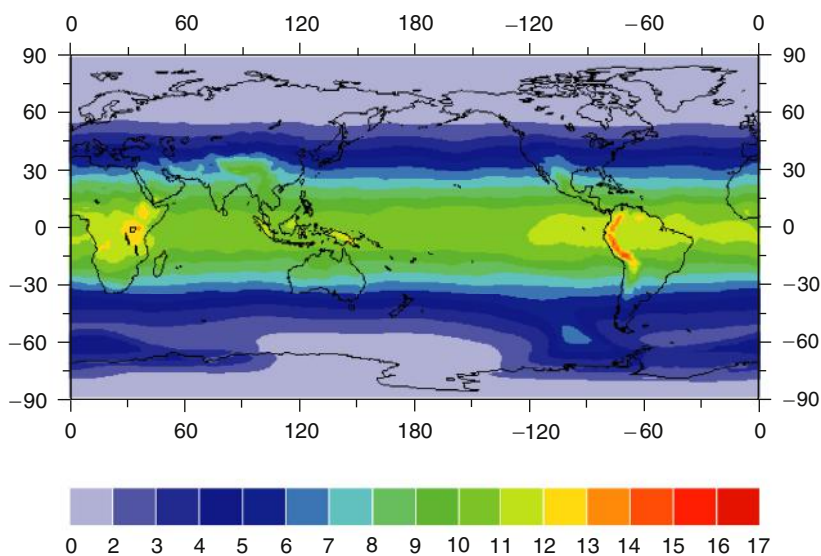
While these global products have been validated at a few locations, the sensors used do not sample radiation from the boundary layer. Consequently, a constant extinction by tropospheric aerosols is often assumed, so the products are unable to resolve differences due to variation in tropospheric pollution. For example, during summer months the UV measured at the surface in unpolluted Southern Hemisphere

Table 2 UV Index as a function of ozone and SZA at 1 AU (altitude 0.00 km, aerosol optical depth 0.00, albedo 0.05)

Ozone (DU)	Solar zenith angle (°)									
	0.00	10.00	20.00	30.00	40.00	50.00	60.00	70.00	80.00	90.00
100	43.47	41.90	37.43	30.69	22.65	14.65	7.82	3.11	0.74	0.05
150	28.16	27.09	24.08	19.58	14.29	9.13	4.82	1.92	0.47	0.03
200	20.16	19.39	17.19	13.93	10.13	6.46	3.42	1.37	0.35	0.02
250	15.42	14.82	13.13	10.63	7.74	4.94	2.63	1.08	0.28	0.02
300	12.35	11.87	10.52	8.52	6.21	3.98	2.14	0.89	0.24	0.02
350	10.23	9.84	8.72	7.08	5.17	3.34	1.81	0.77	0.21	0.02
400	8.71	8.37	7.43	6.04	4.43	2.87	1.58	0.68	0.19	0.02
450	7.56	7.28	6.46	5.27	3.87	2.53	1.40	0.61	0.18	0.01

Table 3 RAFs for several action spectra

<i>Effect</i>	<i>RAF for Jan 30° N, 290 DU</i>	<i>RAF for July 30° N, 305 DU</i>
Erythema	1.1	1.2
DNA damage	2.0	1.9
Caldwell's generalized plant damage	2.0	1.6
Tropospheric photolysis: $O_3 + h\nu \rightarrow O(^1D) + O_2$	2.1	1.8
Phytoplankton motility inhibition	1.9	1.5
Damage to cornea of the eye	1.2	1.1
Damage to lens of the eye (cataract)	0.8	0.7
Immune suppression damage	1.0	0.8
UV-A (315–400 nm)	0.03	0.02
UV-B (290–315 nm)	1.25	0.99
UV-B (290–320 nm)	0.87	0.71

**Figure 2** Global mean daily UV erythemal dose, including cloud effects from a 14-year climatology of ozone 1979–92. Data from the National Center for Atmospheric Research, USA.**Figure 3** World clear sky UV Index at midday for 5 October 2000. Data from the Australian Bureau of Meteorology.

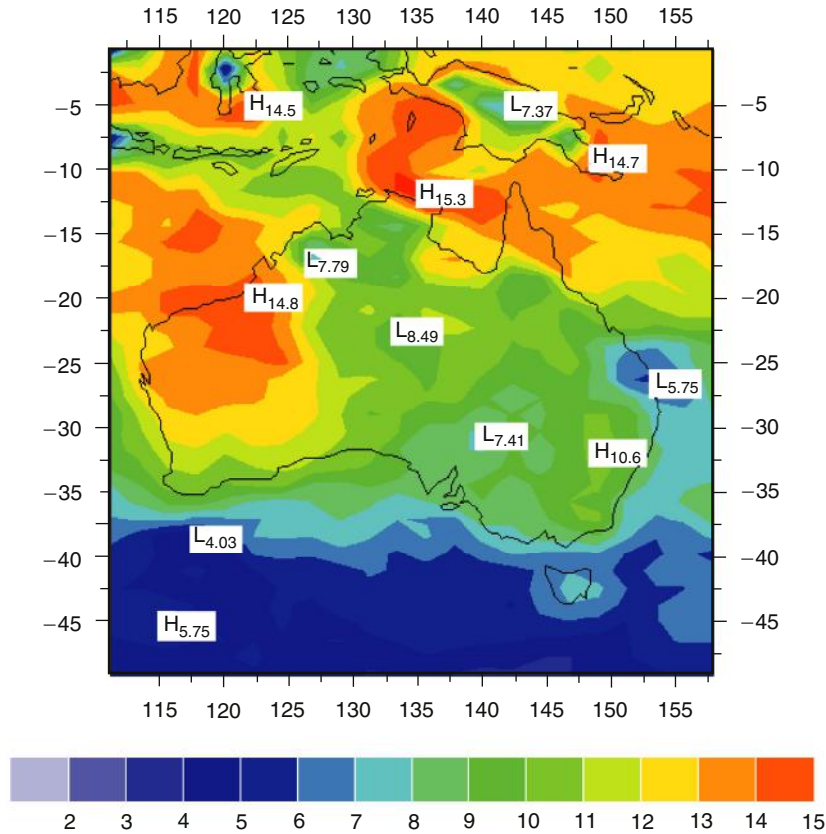


Figure 4 Sample map of noon time UV Index including cloud effects over the Australian region for December 1996. H and L refer to local maxima and minima. Data from the Australian Bureau of Meteorology.

sites is approximately 40% greater than at equivalent latitudes in Europe, whereas the differences derived from satellite observation appear closer to the 15% expected from differences in ozone and Sun–Earth separation. Daily estimates of UV doses are available on the internet.

In several countries, agencies provide UV Index information to the public either through the media, or through direct internet access. The UV Index is calculated with radiative transfer models using measured ozone fields (usually from satellite observations). An example of a global map of the clear-sky UV index is shown in Figure 3. This model includes estimates of the effects of topographical variations, leading to local UV maxima in high-altitude regions. The map shown is from 5 October 2000, a period when the Antarctic ozone hole was weakening and was displaced towards South America, bringing enhanced UV radiation to that sector.

For the purposes of providing forecasts of UV Index for the next day, some agencies include refinements to advect ozone fields and to include estimates of the modulation by clouds. At this stage, products of this type appear to be available on a regional basis only (e.g. Figure 4).

For the purposes of reporting the time evolution of UV Index from the previous day, calculated values can be compared with observations at selected sites. A sample of this is shown in Figure 5. This plot shows how the presence of clouds can increase or decrease the instantaneous UV irradiances. The morning was overcast, resulting in reduced UV irradiances.

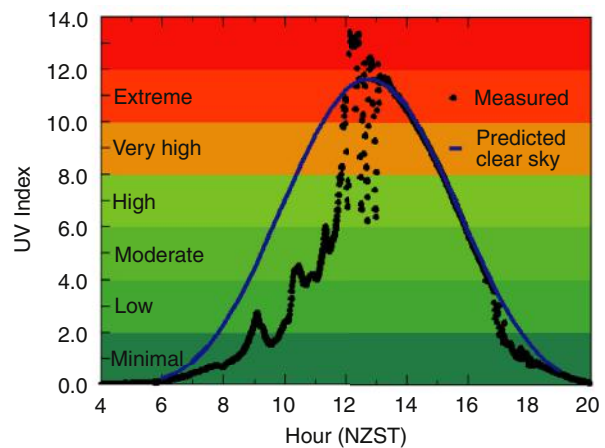


Figure 5 Measured UV Index at Lauder, Central Otago, New Zealand, 9 January 2000, compared with predictions from a clear-sky model using ozone from the NOAA Tiros Operational Vertical Sounder (TOVS) instrument. Data from the National Institute for Water and Atmospheric Research, New Zealand.

During the period when the clouds cleared around midday, there were large departures about the clear-sky predictions in both senses. This was followed by a period of several hours in the afternoon when there was good agreement between the clear-sky model and the measurements.

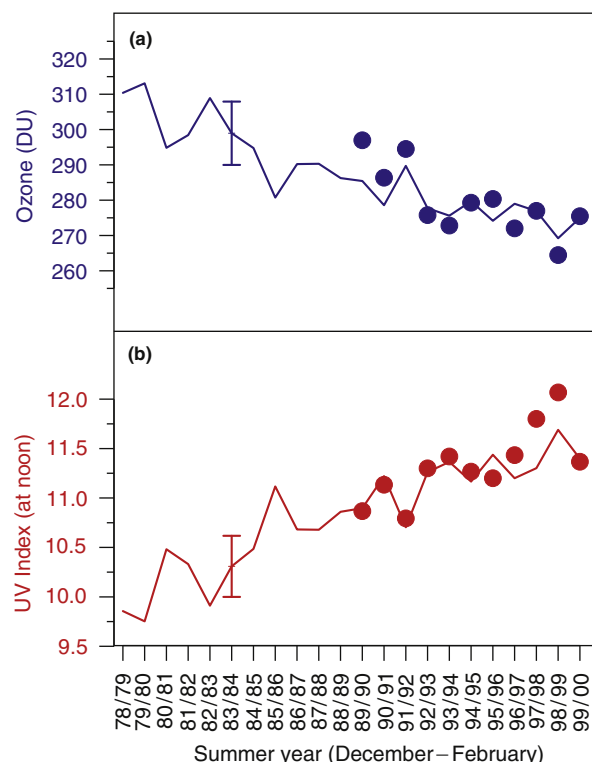


Figure 6 Mean ozone (a) and estimated UV Index (b) at Lauder, New Zealand, for summers from the summer of 1978–79 through the summer of 1999–2000. Summer is defined as the period from December through February. The solid line in (a) shows the changes in summertime ozone that have occurred since the 1970s. The solid line in (b) shows the deduced changes in clear-sky UV expected from these changes in ozone. The symbols (from 1989–90 on) show measured values of ozone and the summertime peak UV Index, both derived from the UV spectroradiometer. The error bar shows the estimated uncertainty in measurement. Adapted from McKenzie RL, Connor BJ and Bodeker GE (1999) Increased summertime UV observed in New Zealand in response to ozone loss. *Science* 285: 1709–1711.

Long-Term Trends

Since satellite observations of ozone began, significant decreases in ozone amounts have been observed outside the tropical region. These have been most pronounced at high latitudes, particularly in the Southern Hemisphere. Prior to 1990, there were few reliable long-term measurements of UV

available. Since then there has been a rapid growth in instrument development and deployment. These instruments range from those that continuously monitor weighted UV irradiances at low spectral resolution through to scanning spectrometers which acquire UV irradiances at spectral resolutions less than 1 nm. However, because of the difficulties of maintaining instrumentation to measure UV with sufficient accuracy, and because of natural variability in cloud cover, there have been few measurements which have clearly demonstrated the expected increases in UV radiation resulting from ozone depletion. Usually it is necessary to restrict the data set to demonstrate the increases in UV due to ozone while suppressing the variability due to clouds. **Figure 6** shows one example to illustrate the long-term reductions in ozone and the attendant changes in peak summertime erythral UV that have occurred at a clean site in New Zealand (45° S).

See also: **Ozone Depletion and Related Topics:** Ozone as a UV Filter. **Radiation Transfer in the Atmosphere:** Cloud-Radiative Processes; Radiation, Solar; Scattering; Ultraviolet Radiation. **Solar System/Sun, Atmospheres, Evolution of Atmospheres:** Solar Terrestrial Interactions: Climate Impact.

Further Reading

- Albritton, D.L., Aucamp, P.J., Megie, G., Watson, R.T. (Eds.), 1999. Scientific Assessment of Ozone Depletion: 1998. WMO, Geneva. World Meteorological Organization Global Ozone Research and Monitoring Project, Report No. 44.
- Herman, J.R., McKenzie, R.L., Diaz, S.B., Kerr, J.B., Madronich, S., Seckmeyer, G., 1999. Ultraviolet radiation at the Earth's surface. In: Albritton, D.L., Aucamp, P.J., Megie, G., Watson, R.T. (Eds.), UNEP/WMO Scientific Assessment of the Ozone Layer: 1998. WMO, Geneva. pp. 9.1–9.46. World Meteorological Organization Global Ozone Research and Monitoring Project, Report No. 44.
- Lemus-Deschamps, L., Rikus, L., Gies, P., 1999. The operational Australian ultraviolet index forecast 1997. *Meteorological Applications* 6, 241–251.
- McKenzie, R.L., Blumthaler, M., Booth, R., et al., 1995. Surface ultraviolet radiation. In: Albritton, D.L., Watson, R.T., Aucamp, P.J. (Eds.), UNEP/WMO Scientific Assessment of Ozone Depletion: 1994. pp. 9.1–9.22. World Meteorological Organization Global Ozone Research and Monitoring Project, Report No. 37. Geneva: WMO.
- Tevini, M., 1993. UV-B Radiation and Ozone Depletion: Effects on Humans, Animals, Plants, Microorganisms, and Materials. Lewis Publishers, Boca Raton, FL.
- United Nations Environment Programme, 1998. Environmental Effects of Ozone Depletion: 1998 Assessment. UNEP, Nairobi.
- Zerefos, C.S., Bais, A.F. (Eds.), 1997. Solar Ultraviolet Radiation. Modelling, Measurements and Effects. NATO ASI Series, vol. 152. Springer, Berlin.

SATELLITES AND SATELLITE REMOTE SENSING

Contents

Aerosol Measurements

Earth's Radiation Budget

GPS Meteorology

Measuring Ozone from Space – TOMS and SBUV

Orbits

Precipitation

Remote Sensing: Cloud Properties

Research

Surface Wind and Stress

Temperature Soundings

Water Vapor

Aerosol Measurements

RA Kahn, NASA Goddard Space Flight Center, Greenbelt, MD, USA

© Published by Elsevier Ltd.

Synopsis

Because aerosols vary on many spatial and temporal scales and exhibit a diversity of environmental impacts, satellite remote sensing makes an essential contribution to the study of airborne particles. Since the very first orbiting imagers began observing Earth, the scope of satellite data products has provided inspiration, qualitative indications, and, increasingly, quantitative constraints on the regional and global influences that aerosols exert. Major advances in this field have taken place in the last decade, providing better constraints on atmospheric processes, short-term forecasting, and climate modeling. Further advances can be expected from greater integration of satellite and suborbital data with models.

Dedication: This article is dedicated to the memory of Dr Yoram J. Kaufman, who coauthored the entry for *Satellite Remote Sensing: Aerosol Measurements* in the previous edition of this Encyclopedia. Yoram's many creative ideas greatly enriched this field, and his untimely passing in June 2006 was a great loss to our community.

Introduction

Aerosols are solid or liquid particles suspended in the air, and those observed by satellite remote sensing are typically between about 0.05 and 10 μm in size. (Note that in traditional aerosol science, the term 'aerosol' refers to both the particles and the medium in which they reside, whereas for remote sensing, the term commonly refers to the particles only. In this article, we adopt the remote-sensing definition.) They originate from a great diversity of sources, such as wildfires, volcanoes, soils and desert

sands, breaking waves, natural biological activity, agricultural burning, cement production, and fossil-fuel combustion. They typically remain in the atmosphere from several days to a week or more, and some travel great distances before returning to Earth's surface via gravitational settling or washout by precipitation. Many aerosol sources exhibit strong seasonal variability, and most experience interannual fluctuations. As such, the frequent, global coverage that space-based aerosol remote-sensing instruments can provide is making increasingly important contributions to regional and larger-scale aerosol studies.

Aerosols affect Earth's energy balance through direct radiative forcing – scattering sunlight back to space, which increases the top-of-atmosphere (TOA) albedo over most surfaces, producing net surface cooling. Darker particles absorb some incoming light, warming the ambient atmosphere, changing cloud properties locally, and possibly altering regional atmospheric circulation patterns. Aerosols also exert 'indirect' effects on clouds, as they provide cloud condensation nuclei and ice

nuclei essential for cloud particle formation, and can thus mediate cloud microphysical properties and regional water cycles. In addition, the near-surface aerosol concentration is a factor affecting local air quality. Many of these topics are explored elsewhere in this Encyclopedia (*see Aerosols: Climatology of Tropospheric Aerosols; Aerosol Physics and Chemistry; Role in Radiative Transfer; Aerosol–Cloud Interactions and Their Radiative Forcing; Satellites and Satellite Remote Sensing: Precipitation*).

The current article focuses on what has been and can be learned about aerosols from space-based remote sensing. Compared to the much older field of *in situ* aerosol sampling, satellite remote sensing is a blunt object, offering at present snapshots of the horizontal and vertical distribution of aerosol amounts over land and water, typically at horizontal spatial resolution of several kilometers, and loose classification of aerosol type under favorable retrieval conditions. Its essential contribution is the scope of coverage, which, even in the early days of satellite imagery in the 1960s, helped establish the transcontinental mineral dust pathway from the North African desert to the Caribbean (**Figure 1**), and has since provided regional and global contexts for a broad range of leading climate and health-related aerosol questions. More recently, advanced passive imagers and active lidar systems have been flown successfully in space; we tell their story here.

Aerosol Remote Sensing over Oceans

Two of the main challenges for satellite aerosol remote sensing are (1) separating the surface from the atmospheric contributions to the TOA observations, and (2) identifying the light-scattering properties of the particles, which are related to aerosol type. The first of these is dramatically reduced where satellite aerosol retrievals are performed over dark, uniform water surfaces, so early efforts to retrieve aerosol column amounts from space were performed over oceans.

Advanced Very High-Resolution Radiometer, Sea-Viewing Wide Field-of-View Sensor, and Geostationary Satellites

From a single red-band spectral channel centered at 0.63 μm , where the ocean surface tends to be dark, together with an assumed optical model for the particles involved, estimates of total-column aerosol amount were derived from Advanced Very High-Resolution Radiometer (AVHRR) reflectance measurements. These revealed the seasonal patterns of major dust, smoke, and pollution aerosol plumes on a global scale (**Figure 2**), providing both inspiration for improved aerosol measurement and modeling, and some actual constraints on chemical transport models that aimed at simulating, and to some extent predicting, the environmental impacts of airborne particles. AVHRR instruments began collecting continuous data from space in 1981, and they offer a substantial time series of global aerosol distributions. Some more recent algorithms incorporate a second AVHRR spectral channel, making it possible to derive limited particle size constraints. But poor radiometric calibration accuracy and the small number of relatively broad spectral bands on these wide-swath, single-

view instruments limited the quantitative application of the data to aerosol research. Similar products related to aerosol amount have been derived with higher temporal resolution from the US National Oceanic and Atmospheric Administration's (NOAA) Geostationary Environmental Satellites (GOES), the European Space Agency's (ESA) Spinning Enhanced Visible and Infrared Imager (SEVIRI) geostationary instrument, and also the US National Aeronautics and Space Administration's (NASA) polar-orbiting Sea-viewing Wide Field-of-view Sensor (SeaWiFS), a space-based multispectral imager that has much narrower and better-calibrated spectral bands than the AVHRR.

MODerate-Resolution Imaging Spectroradiometer

Whereas the intensity of sunlight reflected by aerosols is closely related to column amount, the spectral dependence of the aerosol-reflected component contains some information about particle size. The MODerate-resolution Imaging Spectroradiometers (MODIS), part of the NASA Earth Observing System (EOS), represents a second generation of multispectral imagers, having 36 spectral channels spanning 0.4–14.4 μm , high radiometric calibration accuracy and stability, and subkilometer pixel resolution. These features make it possible to produce a significantly higher quality constraint on the optically equivalent aerosol column amount, generally reported as the extinction aerosol optical depth (AOD), a measure of the amount of light removed from an incident beam at a given wavelength, due to scattering in all directions as well as absorption by particles. The first MODIS instrument was launched with the EOS *Terra* satellite and began acquiring data in early 2000; a second MODIS began its mission on EOS *Aqua* in June 2002. Over the ocean, MODIS TOA, scattered-light measurements from six visible and near-infrared spectral channels are interpreted in terms of AOD and fine-mode AOD fraction (FMF), along with estimates of the fine- and coarse-mode effective sizes, providing global coverage approximately every 2 days (**Figure 3**). A standard dark ocean surface model is assumed, including sun-glint exclusion and wind-dependent whitecap reflectance, and particle properties are selected from a predetermined list of likely aerosol types. The FMF is helpful in identifying aerosol type, as mechanically produced desert dust and maritime aerosols formed by breaking waves tend to be dominated by 'coarse-mode' particles larger than a micron in diameter, whereas the populations of smoke, pollution, and other combustion and biogenic aerosols fall largely into the submicron 'fine mode.'

Over-Ocean AOD Trends and the Visible Infrared Imaging Radiometer Suite

One important application of the regional- to global-scale AOD time series derived from satellite observations has been the identification of trends. Aerosol amounts vary on many spatial and temporal scales, so determining systematic tendencies requires relatively long, high-precision data records, having sufficient spatial coverage and temporal frequency to account for measurement anomalies and isolated events such as volcanic ash and wildfire smoke plumes.

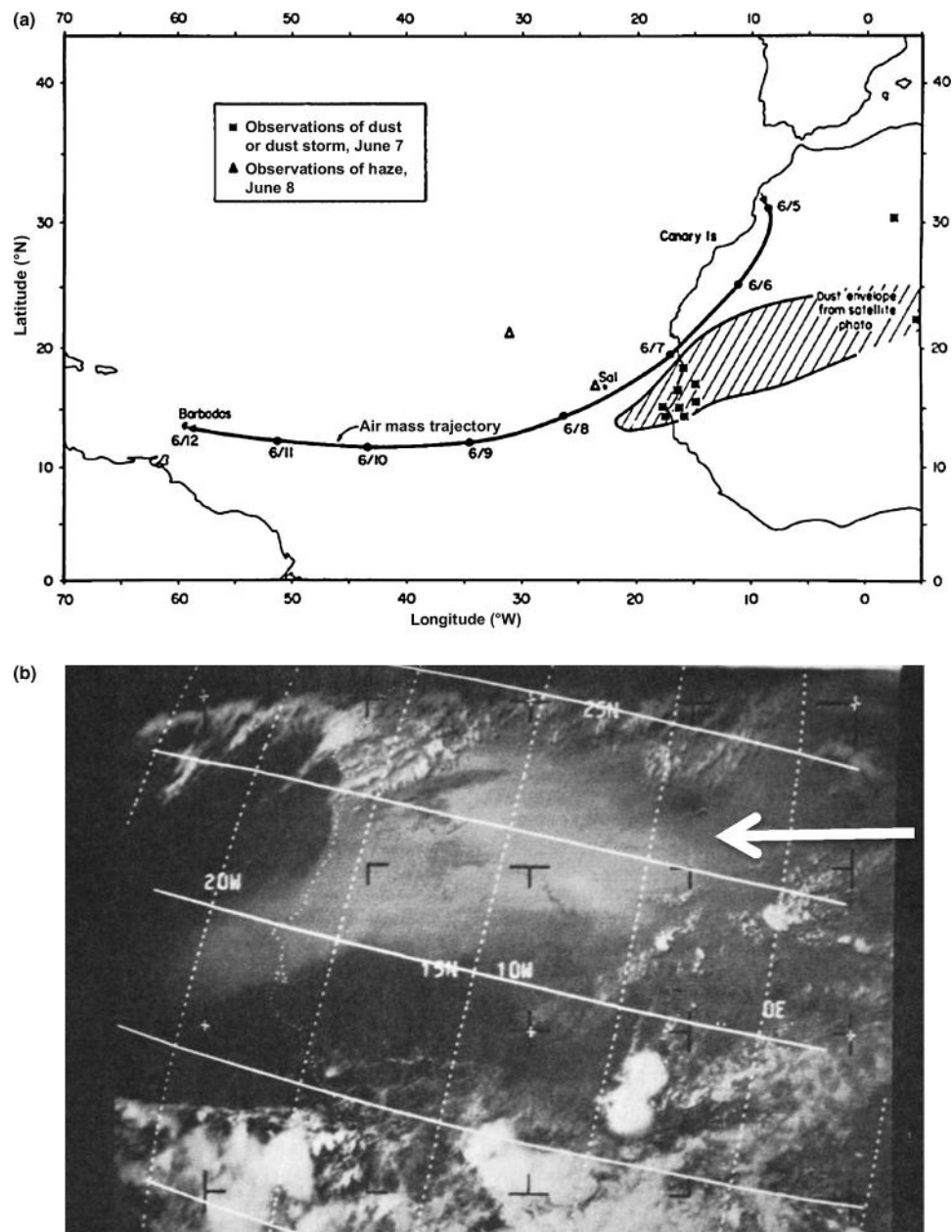


Figure 1 (a) Trajectory of a Saharan dust plume, tracked over 8 days from the source region in North Africa across the Atlantic Ocean to the Caribbean, based on images from a wide-swath vidicon camera aboard the polar-orbiting ESSA 5 satellite. (b) ESSA 5 satellite image of the plume (indicated by white arrow) over Mauritania, the western Sahara Desert, and the eastern Atlantic on 7 June 1967. Dust particles more than $20\ \mu\text{m}$ in diameter from this storm were recovered subsequently in Barbados. Reproduced from Prospero, J.M., Bonatti, E., Schubert, C., Carlson, T.N., 1970. Dust in the Caribbean atmosphere traced to an African dust storm. *Earth Planetary Science Letters* 9, 287–293.

Such analysis was first performed with 25 years of AVHRR data and subsequently, with the first decade of carefully filtered MODIS AOD retrievals (Figure 4). The globally averaged over-ocean trend derived between 2000 and 2009 was negligible, but some significant regional AOD increases and smaller decreases were found, in most cases traced to changes in human activity. To continue this time series, a broad-swath, multispectral Visible Infrared Imaging Radiometer Suite (VIIRS) imager, having capabilities in some

respects similar to those of MODIS, was launched on the National Polar-orbiting Operational Environmental Satellite System Preparatory Project (NPP; now the Suomi National Polar-orbiting Partnership, or Suomi NPP) satellite in 2011, the first in a series planned for future NOAA operational polar-orbiting satellites.

Retrieving AOD over land with comparable accuracy from space-based observations is more difficult, but significant progress has been made in this area as well.

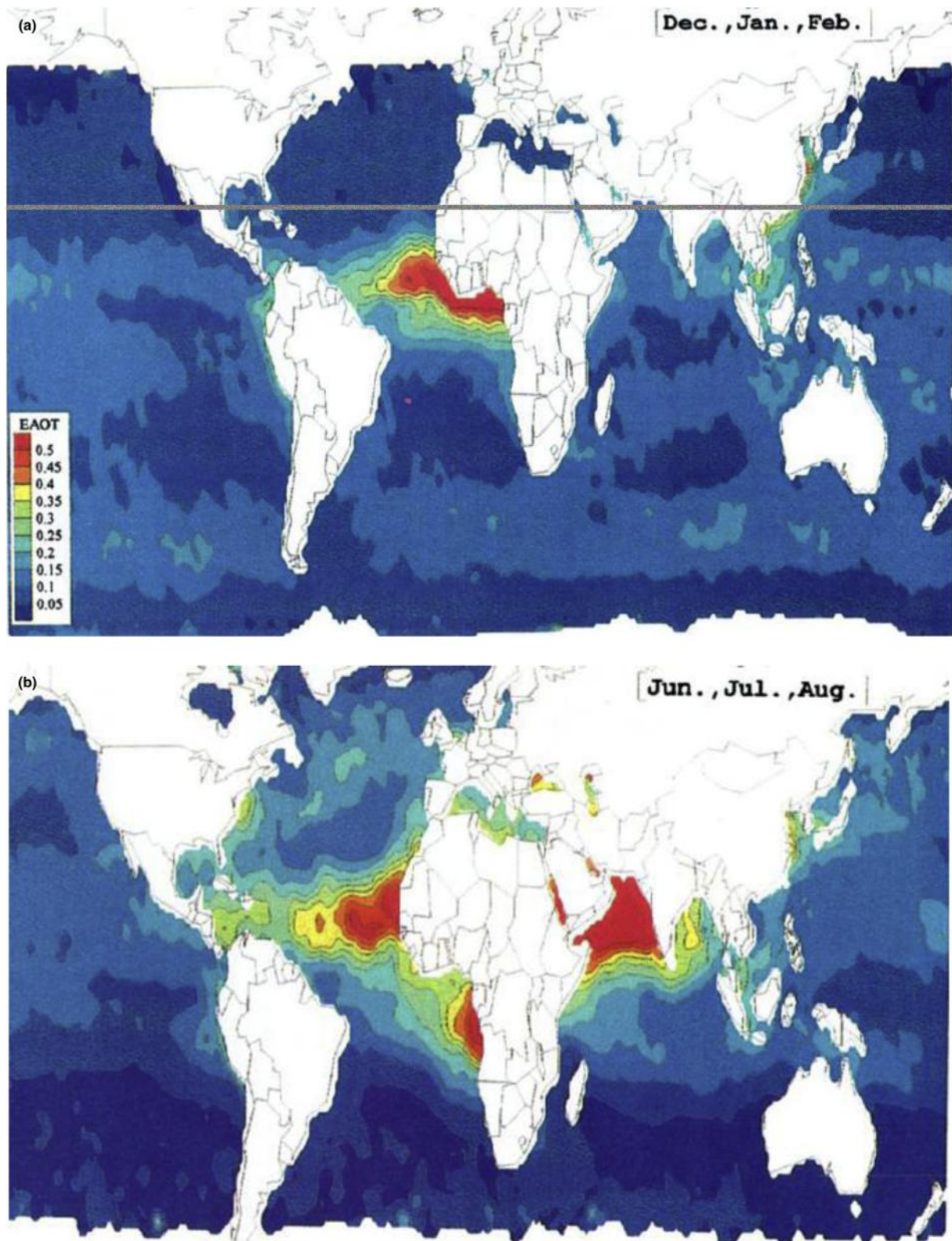


Figure 2 Global, over-ocean estimates of total-column aerosol amount for the northern (a) winter and (b) summer seasons, derived from 2 years (July 1989–June 1991) of midvisible reflectance measurements from the Advanced Very High-Resolution Radiometer (AVHRR) instruments aboard NOAA polar-orbiting satellites. The winter peak in grassland burning produces a smoke plume over the Atlantic Ocean west of the Sub-Saharan region, whereas in summer, dust from North Africa and smoke from Central Africa produce plumes over the adjacent water, and dust from the Middle East blankets the Arabian Sea. Pollution sources off the east coast of China vary little between these two seasons, whereas pollution off the east coast of the United States is more prominent in summer. Reproduced from Husar, R.B., Prospero, J.M., Stowe, L.L., 1997. Characterization of tropospheric aerosols over the oceans with the NOAA advanced very high resolution radiometer optical thickness operational product. *Journal of Geophysical Research* 102, 16889–16909.

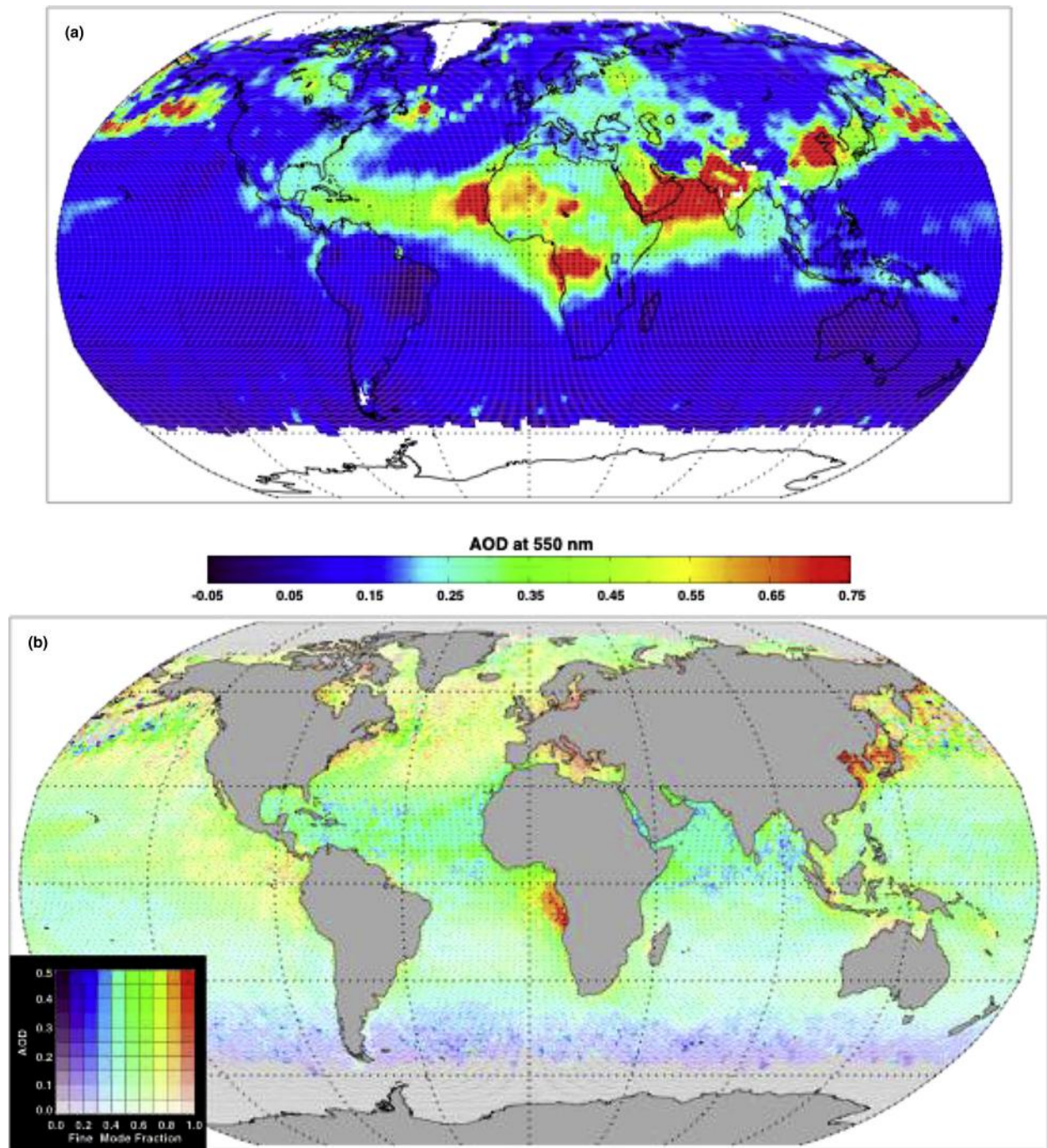


Figure 3 MODIS global monthly average aerosol products for July 2010. (a) AOD from the combination of Dark Target and Deep Blue algorithms; and (b) fine-mode fraction (FMF) over ocean, with AOD, related to the confidence with which FMF can be determined, encoded as the saturation of the color applied, and all derived with the Dark Target algorithm. Provided by the MODIS Team, NASA Goddard Space Flight Center.

Aerosol Remote Sensing over Land

The land surface of Earth is generally brighter and more variable than that of the ocean. To retrieve aerosol amount and type over land from satellite observations, the typically

smaller aerosol-reflected component of upwelling radiation must be distinguished from the surface-reflected component. Since the early days of satellite aerosol observation, a range of approaches has been conceived, developed, and applied to address this challenge.

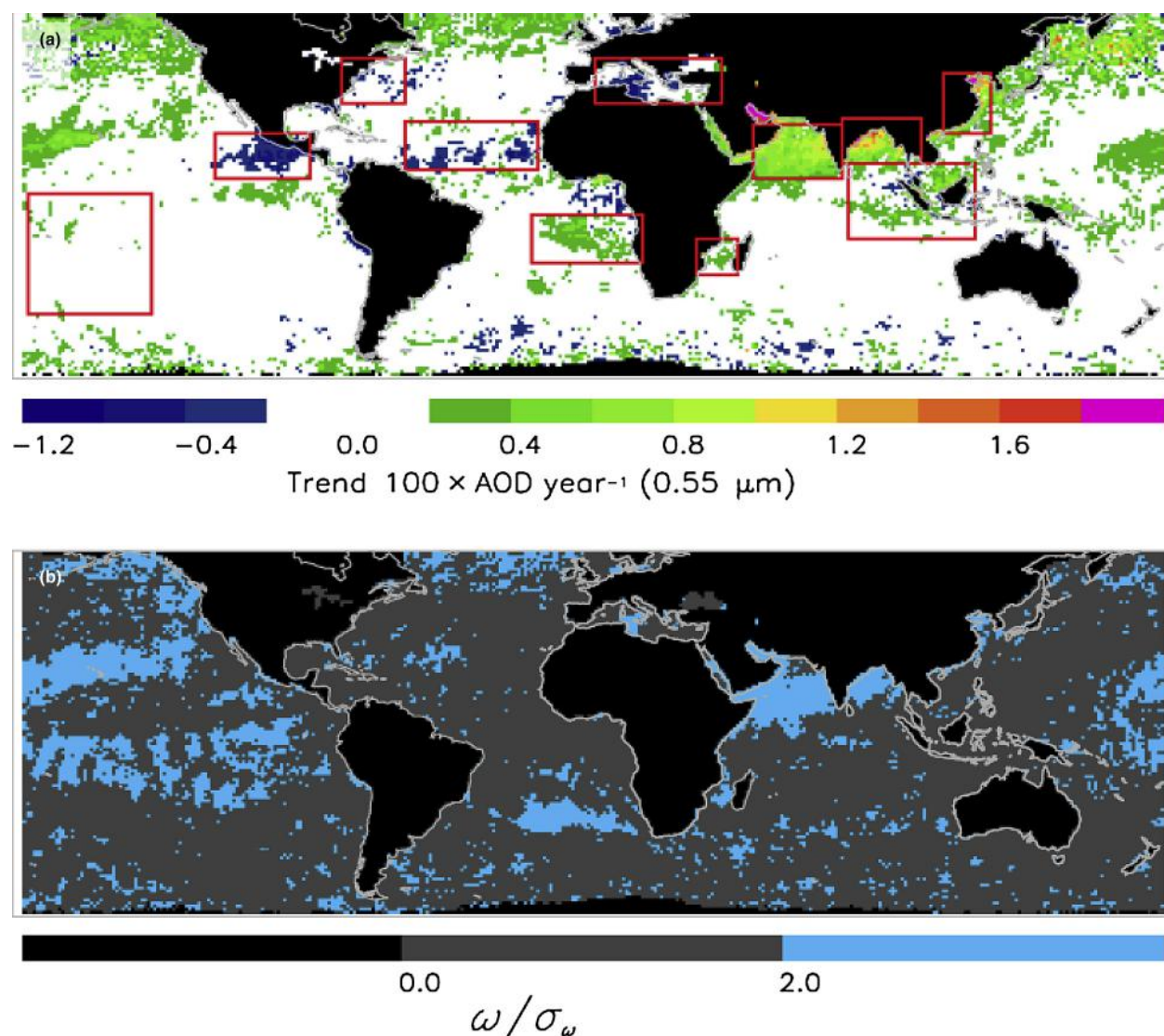


Figure 4 Global and regional, over water: (a) midvisible AOD trends, and (b) associated confidence levels, derived from 10 years (2000–09) of MODIS operational aerosol products. The MODIS time series was first filtered to remove outliers and reduce possible cloud contamination, and was calibrated using less frequent but more accurate MISR and AERONET observations. Confidence levels were assessed as the derived trend (ω), normalized by the estimated time series standard deviation (σ_ω), on $1^\circ \times 1^\circ$ grid cells. Increasing AOD trends are found over the Bay of Bengal, east coast of Asia, and Arabian Sea, whereas weaker AOD decreases were derived from Central America, the east coast of North America, and the west coast of Africa. Reproduced from Zhang, J., Reid, J.S., 2010. A decadal regional and global trend analysis of the aerosol optical depth using a data-assimilation grade over-water MODIS and Level 2 MISR aerosol products. *Atmospheric Chemistry and Physics* 10, <http://dx.doi.org/10.5194/acp-10-1-2010>.

Total Ozone Mapping Spectrometer Instruments and the Ozone Mapping Instrument

Atmospheric gas molecules scatter ultraviolet (UV) light very efficiently, partly obscuring the surface as viewed from space, and Earth's surface tends to be darker in UV light than in visible light. So the differential absorption in two UV spectral channels from the Total Ozone Mapping Spectrometer (TOMS) instruments, which began taking data in 1979, and subsequently from the Ozone Mapping Instrument (OMI), has been interpreted as an Aerosol Index, a qualitative measure of aerosol

amount over land and water. With some assumptions and constraints on aerosol vertical distribution and on aerosol UV absorption properties (usually represented by the single-scattering albedo (SSA), which is the ratio of the fraction of light scattered to that scattered and absorbed by the particles at a given wavelength), these data also yield the AOD (Figure 5). As the retrieval is based on aerosol absorption of the upwelling UV radiation, it tends to be less sensitive to near-surface aerosol. A similar approach was used for the European Global Ozone Monitoring Experiment (GOME) satellites.

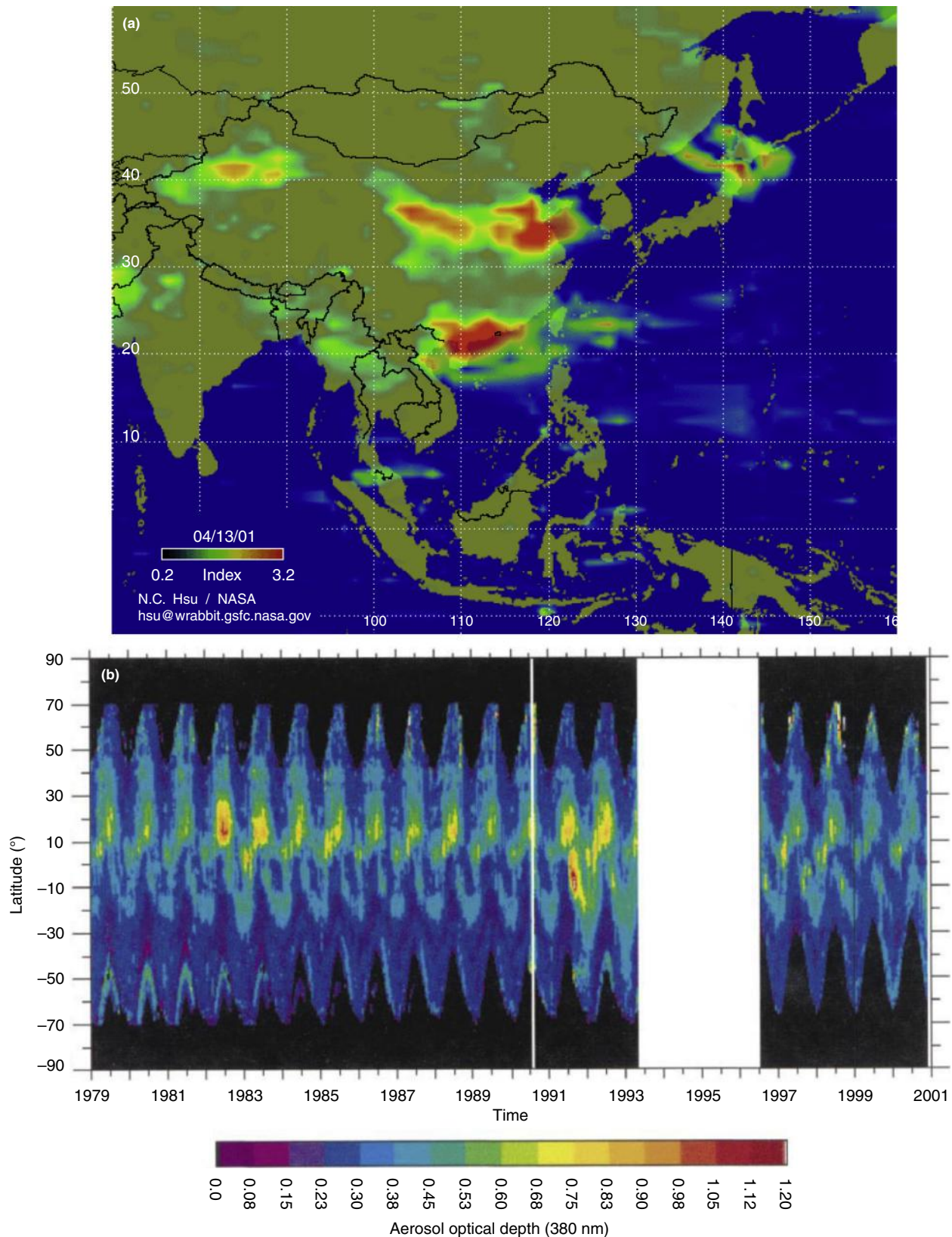


Figure 5 TOMS aerosol data products. (a) Aerosol index (AI) map over South and Southeast Asia and surrounding water, for 13 April 2001. Dust over Central Asia, smoke and pollution aerosol over Southeast Asia, and dust combined with pollution over East-Central Asia stand out. Reproduced from Hsu, N.C., Herman, J. R., Bhartia, P. K., Seftor, C. J., Torres, O., Thompson, A. M., Gleason, J. F., Eck, T. F., Holben, B. N., 1996. Detection of biomass burning smoke from TOMS measurements. *Geophysical Research Letters* 23, 745–748. (b) Zonal-average AOD time series at 380 nm, covering 1979–92 and 1996–2000. Note the prominent volcanic plumes from El Chichon in 1982 and Mt. Pinatubo in 1991, each lasting many months. Reproduced from Torres, O., Bhartia, P.K., Herman, J.R., Sinyuk, A., Ginoux, P., Holben, B.N., 2002. A long-term record of aerosol optical depth from TOMS observations and comparison to AERONET measurements. *Journal of Atmospheric Science* 59, 398–413.

MODIS Dark Target, Deep Blue, and Multiangle Implementation of Atmospheric Correction

Several approaches have been used to obtain AOD over land from the MODIS instruments. The Dark Target method over land relies on the 2.1 μm MODIS channel, in which atmospheric gas and aerosol opacity are generally low, to provide a constraint on the surface reflectance. An empirical relationship is then used to extrapolate the surface reflectance to visible wavelengths, where the AOD is determined; a similar approach is used for the ESA's Medium Resolution Imaging Spectrometer (MERIS) and is planned for the next generation of US geostationary satellite multispectral imagers. Over land, the aerosol type used in the algorithm is assumed, based on climatology derived from the global Aerosol Robotic Network (AERONET) of surface sun and sky-scanning photometers (Figure 6). AERONET also provides high-quality AOD measurements used to validate many satellite AOD products. As MODIS also has a relatively short-wavelength Deep Blue

channel at 0.41 μm , a variant of the UV absorption technique provides AOD from that instrument over brighter surfaces, such as desert. The combination of MODIS's dark water, Dark Target land, and Deep Blue AOD products is illustrated in Figure 3(a). A third approach for extracting AOD over land from MODIS is called Multiangle Implementation of Atmospheric Correction (MAIAC). It relies on detecting temporal variations in the observed radiance. The changing AOD, Ångström exponent (i.e., the spectral dependence of the AOD; specifically, minus the spectral slope of AOD in log-log coordinates), and surface angular reflection properties are extracted from 16-day time series of MODIS imagery, capturing multiple views of the region at different angles.

Along-Track Scanning Radiometer-2 and Multiangle Imaging SpectroRadiometer

As Earth is viewed at steeper angles, the atmospheric contribution to the TOA reflectance systematically increases, and the

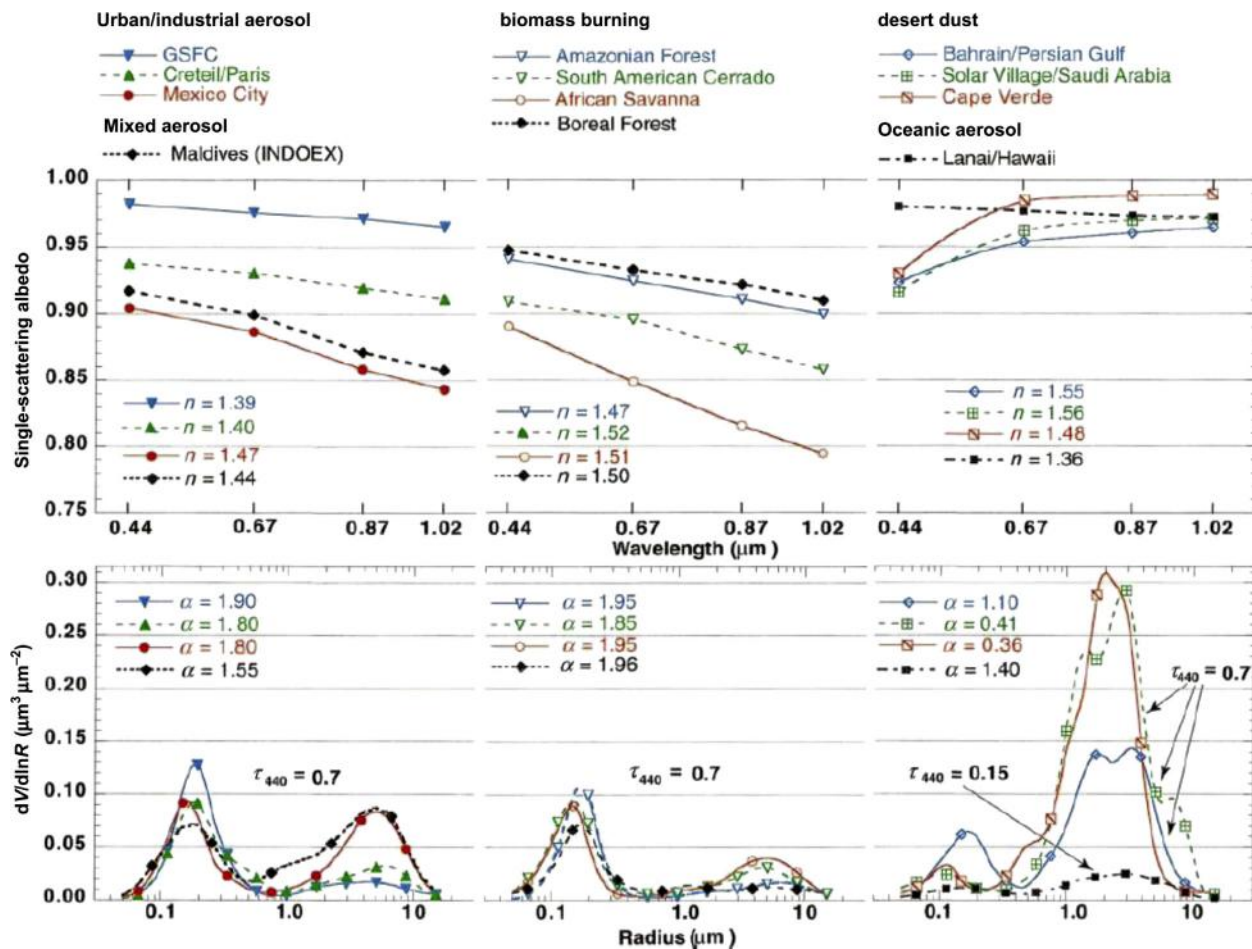


Figure 6 Averaged particle spectral single-scattering albedos (SSA) and size distributions for several sites dominated by pollution, smoke, dust, and oceanic aerosols, as derived from AERONET surface-based sun photometer measurements. AOD at 440 nm for the cases shown are indicated as τ_{440} ; the real part of the particle refractive index is given as n , and the spectral dependence of the AOD (Ångström exponent), evaluated at 440 and 870 nm, is designated α . Note how the size distributions vary with aerosol type, and how SSA varies considerably even within the urban-industrial and biomass-burning categories. Reproduced from Dubovik, O., Holben, B.N., Eck, T.F., Smirnov, A., Kaufman, Y.J., King, M.D., Tanré, D., Slutsker, I., 2002. Variability of absorption and optical properties of key aerosol types observed in worldwide locations. *Journal of Atmospheric Science* 59, 590–608.

surface is increasingly obscured. Multiangle observations make it possible to separate surface from atmosphere based on the varying air mass factors through which the observations are made. The ESA's Along-Track Scanning Radiometer-2 (ATSR-2) imagers made use of this approach with a two-angle configuration beginning in 1995; and in 2000, the NASA EOS's Multiangle Imaging SpectroRadiometer (MISR), with nine cameras pointed at angles ranging from 70° aft, through nadir, to 70° forward along the orbit track, began operations. The AOD over land and ocean is produced from both ATSR-2 and MISR instrument data sets; ATSR also reports the Ångström exponent, and MISR provides a classification of aerosol 'type' under favorable retrieval conditions, based on loose aerosol size, shape, and SSA constraints that can be derived from the multiangle, multispectral data (Figure 7).

POLarization and Directionality of the Earth's Reflectances

Polarization is an additional property of light from which information about a scene can be extracted. The inclusion of polarization sensitivity with multispectral, multiangle

capabilities allows the French Space Agency's (CNES) POLarization and Directionality of the Earth's Reflectances (POLDER) series of imagers to retrieve fine-mode and total AOD over land and water. The analysis takes advantage of the relative spectral independence of polarized reflectance for most land surfaces, and the greater polarization of light scattered by smaller and more spherical particles, such as smoke or pollution, compared to larger, nonspherical desert dust. The POLDER instruments began acquiring data in 1997, and they combine spectral, angular, and polarization information to monitor aerosol (Figure 8).

Particle Properties

In general, having constraints on particle properties improves the accuracy of AOD results for both the UV absorption and the various scattering retrieval techniques. Knowing the aerosol type is also critical for source attribution, the assessment of aerosol radiative forcing, the determination of material fluxes, and other applications that depend upon knowing the chemical or physical nature of the particles. So in addition to the multispectral

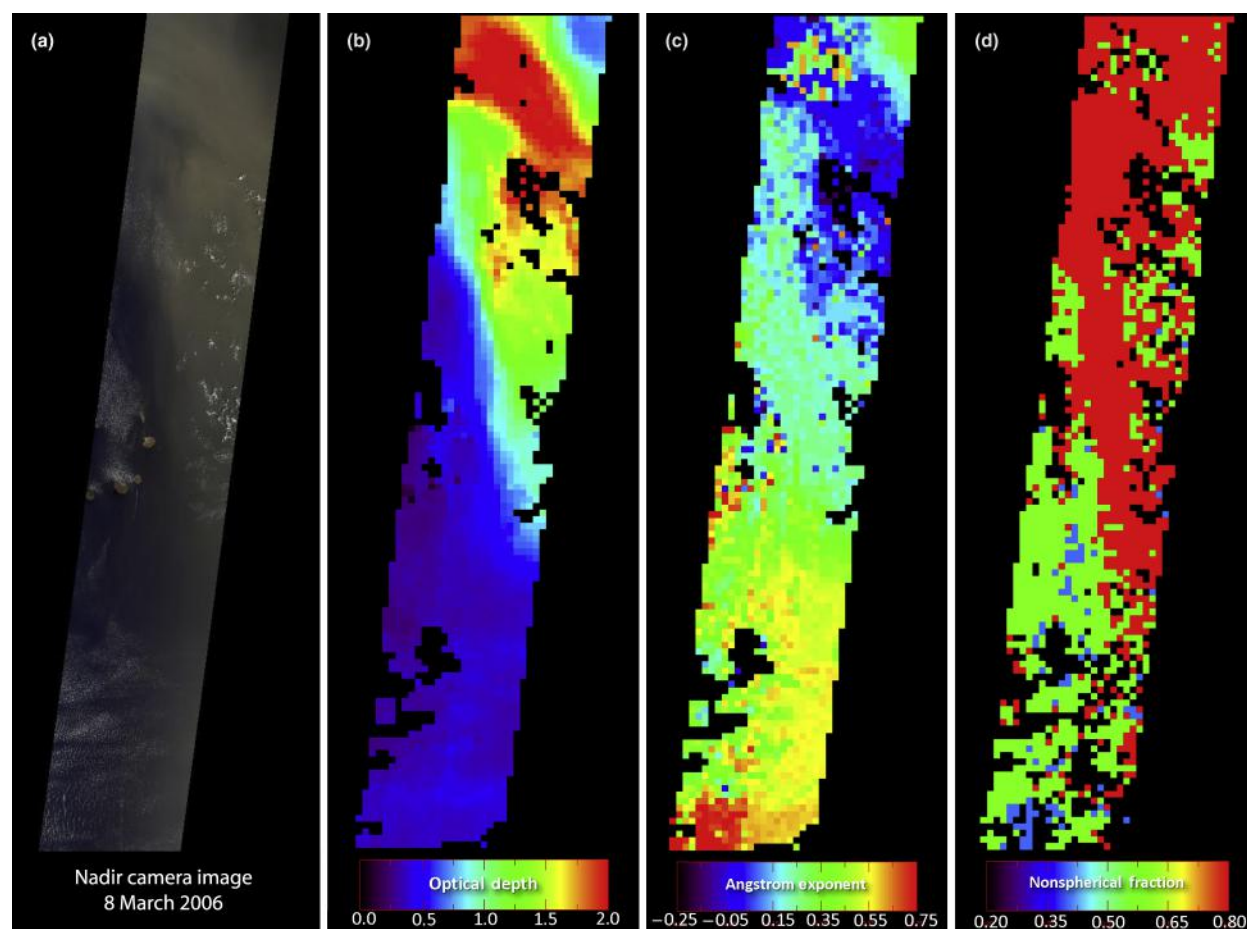


Figure 7 MISR (a) true-color, nadir-viewing image of a Sahara Desert dust plume over the Atlantic Ocean north and east of the Cape Verde Islands (which are visible in the center of the image); and retrieved (b) AOD, (c) Ångström exponent, and (d) fraction AOD nonspherical. North is roughly toward the top of the images, and the swath is about 380 km wide. The plume has higher AOD, and the dust particles are larger (i.e., a smaller Ångström exponent) and more nonspherical compared to background values. Provided by the MISR Team, NASA Goddard Space Flight Center and Jet Propulsion Lab/Caltech.

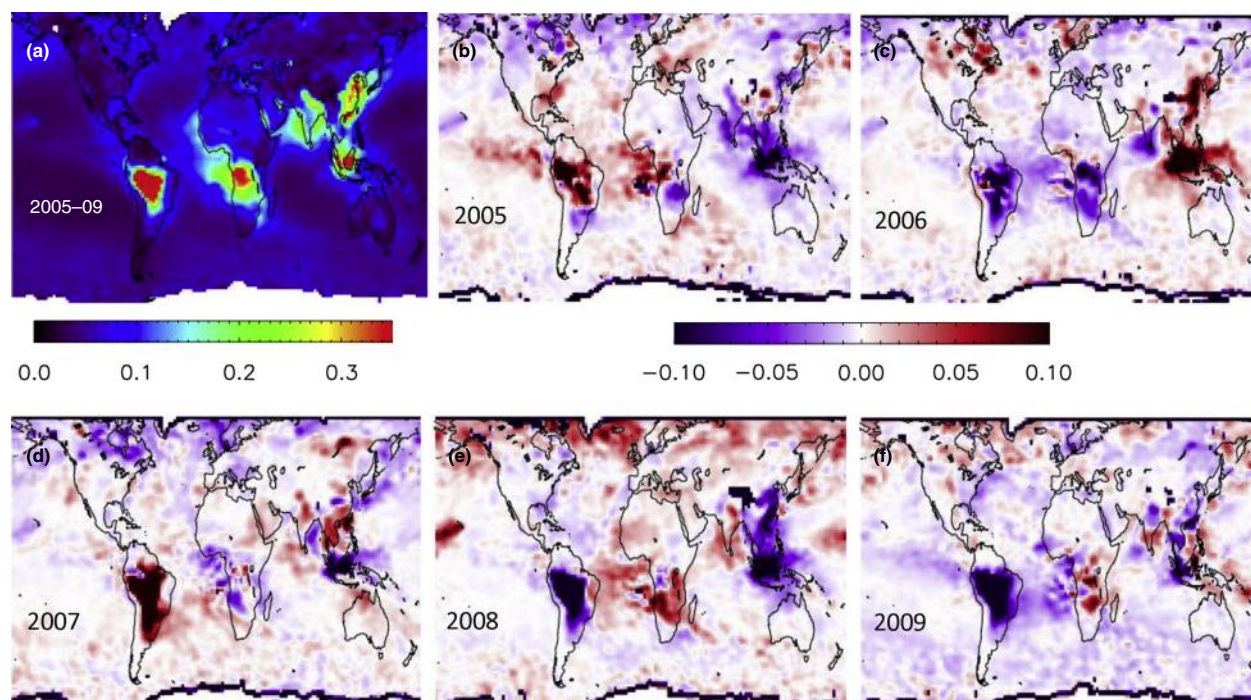


Figure 8 Three-month (September through November), global fine-mode AOD at 550 nm over land and water, derived from POLDER observations: (a) averaged over 5 years (2005–09); and (b–f) AOD anomaly for each individual year. Blue in the anomaly panels indicates that the AOD for that year was lower than the longer-term mean, whereas red indicates higher values. Note the high interannual variability in the biomass-burning regions of Brazil, Southern Africa, Indonesia, and the northern boreal latitudes. Adapted from Tanré, D., Breon, F.M., Duzé, J.L., Dubovik, O., Ducos, F., Francios, P., Goloub, P., Herman, M., Lifermann, A., Waquet, F., 2011. Remote sensing of aerosols by using polarized, directional and spectral measurements within the A-Train: the PARASOL mission. *Atmospheric Measurement Techniques* 4, 1383–1395, <http://dx.doi.org/10.5194/amt-4-1383-1395-2011>.

assessment of particle size, and the multiangle multispectral determination of particle type with and without polarization, several other techniques for constraining properties from space have been demonstrated, at least on a case-by-case basis.

The *critical reflection* technique relies on the idea that if particles reflect more light at a given view angle and wavelength than the surface below, the TOA reflectance will increase with AOD, whereas if the particles reflect less light than the surface below, the TOA reflectance will decrease as AOD increases. So if a layer of aerosol that is uniform in amount and type overlies a surface of varied albedo, some parts of the scene will be brightened more than others by the aerosols. From images of the scene on several days when the aerosol type is similar but the AOD is different, the SSA of the particles can be deduced. As aerosol properties for a given source in a given season tend to be repeatable, even if the amount of aerosol varies considerably, this is a credible approach for deriving aerosol type information over variable land surfaces in some locations.

Other methods combine the capabilities of several instruments to help constrain underdetermined remote-sensing retrievals. For example, the TOMS and OMI UV AOD retrievals are sensitive to particle SSA and aerosol vertical distribution, and pixels several tens of kilometers in size from these instruments can be affected by undetected subpixel clouds. Combining TOMS or OMI observations with near-coincident MODIS AOD and relatively high-spatial-resolution cloud clearing can produce more accurate

constraints on SSA. When aerosol vertical distribution from an active sensor such as the NASA EOS's Cloud-Aerosol Lidar and Infrared Pathfinder Satellite Observations (CALIPSO) (Figure 9) is included, even tighter constraints can be derived on aerosol properties and scene conditions. Similarly, data from the two-angle-viewing AATSR, providing AOD and cloud identification at relatively high spatial resolution, have been combined with high-spectral-resolution observations from the SCanning Imaging Absorption spectroMeter for Atmospheric CHartographY (SCIAMACHY), both aboard ESA's Envisat, to distinguish mineral dust, sea salt, soot, and water-soluble aerosol types.

With recent technological advances and the demonstrated value of multiangle, multispectral, polarizing imagery covering UV through visible into near-infrared light, future satellite aerosol instruments will likely combine all these capabilities into a single passive, broad-swath, remote-sensing imager.

Constraints on Climate Models

Synthesizing the collection of observations into a global picture while taking account of the relative strengths, limitations, and gaps in the various data records is a challenge in itself. But it is an essential step in applying the measurements to short-term forecast models and to long-term climate models, which are

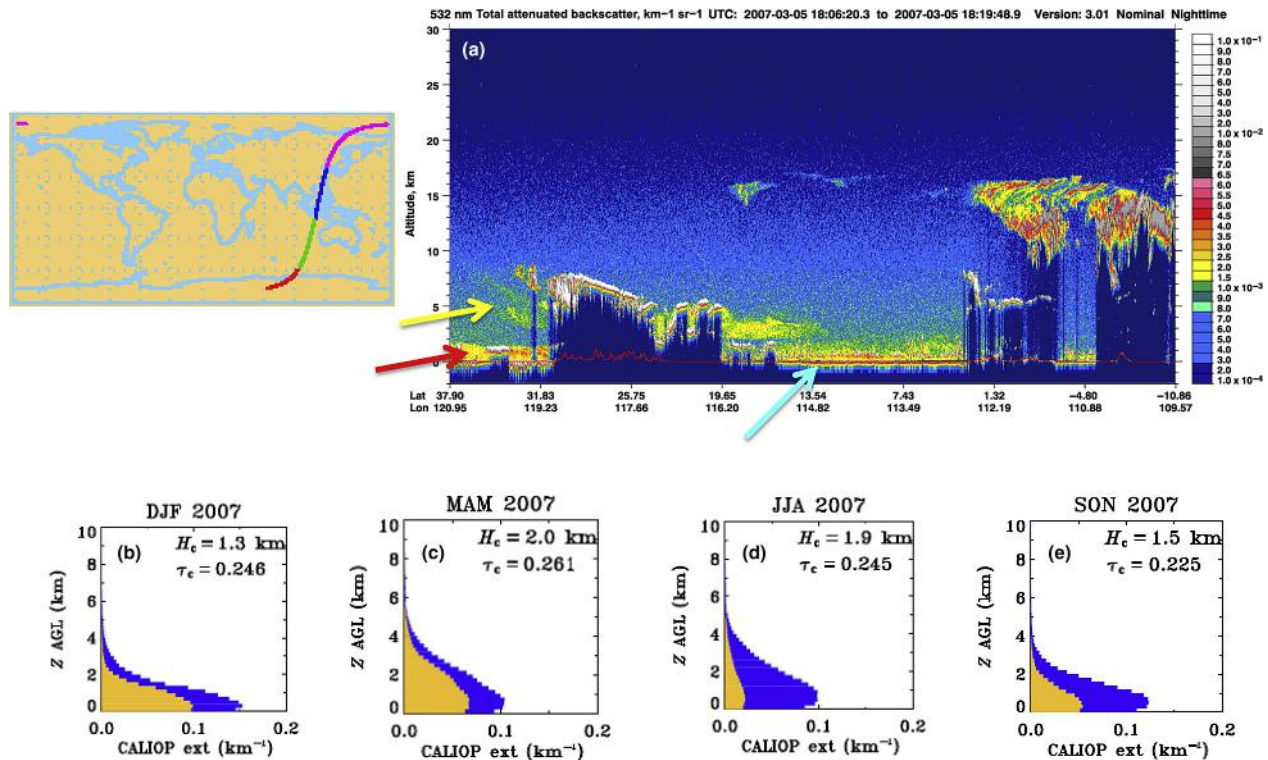


Figure 9 CALIPSO vertical profiles. (a) Backscatter curtain plot, showing cloud and aerosol layers over the region highlighted in dark blue in the adjacent context map. Spectral and depolarization lidar signals identify transported dust (yellow arrow) overlying a surface layer of polluted continental aerosol (red arrow). Clean maritime aerosol and maritime aerosol mixed with dust and pollution particles (cyan arrow) are found farther south along the orbit track. Reproduced from the CALIPSO Team (<http://www-calipso.larc.nasa.gov>; browse image from 05 March 2007), NASA Langley Research Center. (b–e) Seasonally aggregated dust (orange) and nondust, mainly smoke and pollution (blue) vertical extinction profiles over eastern China in 2007. Note how the vertical extent is a minimum in the autumn (SON) and winter months (DJF), and that the dust amount peaks in the winter and spring (MAM) seasons. The total-column AOD (τ_c) and effective aerosol scale height (H_e) are indicated in each panel. Reproduced from Yu, H., Chin, M., Winker, D.M., Omar, A.H., Liu, Z., Kittaka, C., Diehl, T., 2010. Global view of aerosol vertical distributions from CALIPSO lidar measurements and GOCART simulations: regional and seasonal variations. *Journal of Geophysical Research* 115, D00H30, <http://dx.doi.org/10.1029/2009JD013364>.

the primary tools for diagnosing many atmospheric processes and for making predictions. One approach has been to assimilate the observations into models, ‘nudging’ the model AOD based on available measurements. The frequent (1- to 2-day) global coverage of the MODIS AOD data has been successfully used to improve several-day AOD model forecasts, once the satellite data set was filtered to remove outliers and retrievals that might have been affected by unscreened cloud, leaving about 50% of the original points.

For longer-term climate modeling, more uniform global products are generally required. Monthly, global maps of AOD were created from the combination primarily of MODIS, MISR, and surface sun photometer AERONET data. These are being used to constrain leading numerical climate models. In this case, the frequent coverage of MODIS over water, the approximately once-weekly but generally more accurate overland AOD from MISR, and the AERONET data (which are the most accurate but provide only spot sampling) were the main components aggregated into a unified AOD product (Figure 10).

Knowing aerosol vertical distribution is especially important for climate modeling, as aerosol reflection and absorption of sunlight at different levels in the atmosphere not only alter the surface energy budget but can also affect the atmospheric

stability structure, regional-scale atmospheric circulation, and aerosol impacts on clouds. The vertical distribution is also critical for assessing aerosol transports, as aerosols residing above the atmospheric boundary layer tend to stay aloft longer, travel farther, and have greater environmental impact. The CALIPSO lidar generates a curtain of backscatter profiles at two wavelengths and has polarization sensitivity, with up to about 330 m horizontal resolution and 30 m vertical resolution. Some vertically resolved aerosol type classification is also possible with these data. Coverage is limited to the width of the lidar beam, which means that most aerosol sources are missed, but the great sensitivity of the observations to very thin aerosol layers far downwind of sources provides accurate snapshots that can be aggregated into general climatological constraints (Figure 9(b)–(e)).

Aerosols are usually introduced into models by providing an inventory of *injection heights*, *source strengths*, and locations for different aerosol types. Aerosol injection heights, derived from the parallax in multiangle imaging views of aerosol plumes, can provide such information (Figure 11). Unlike lidar, stereo imaging requires aerosol plumes to exhibit discrete features that can be tracked in multiple angular views, so the technique applies mainly within a few hundred kilometers of

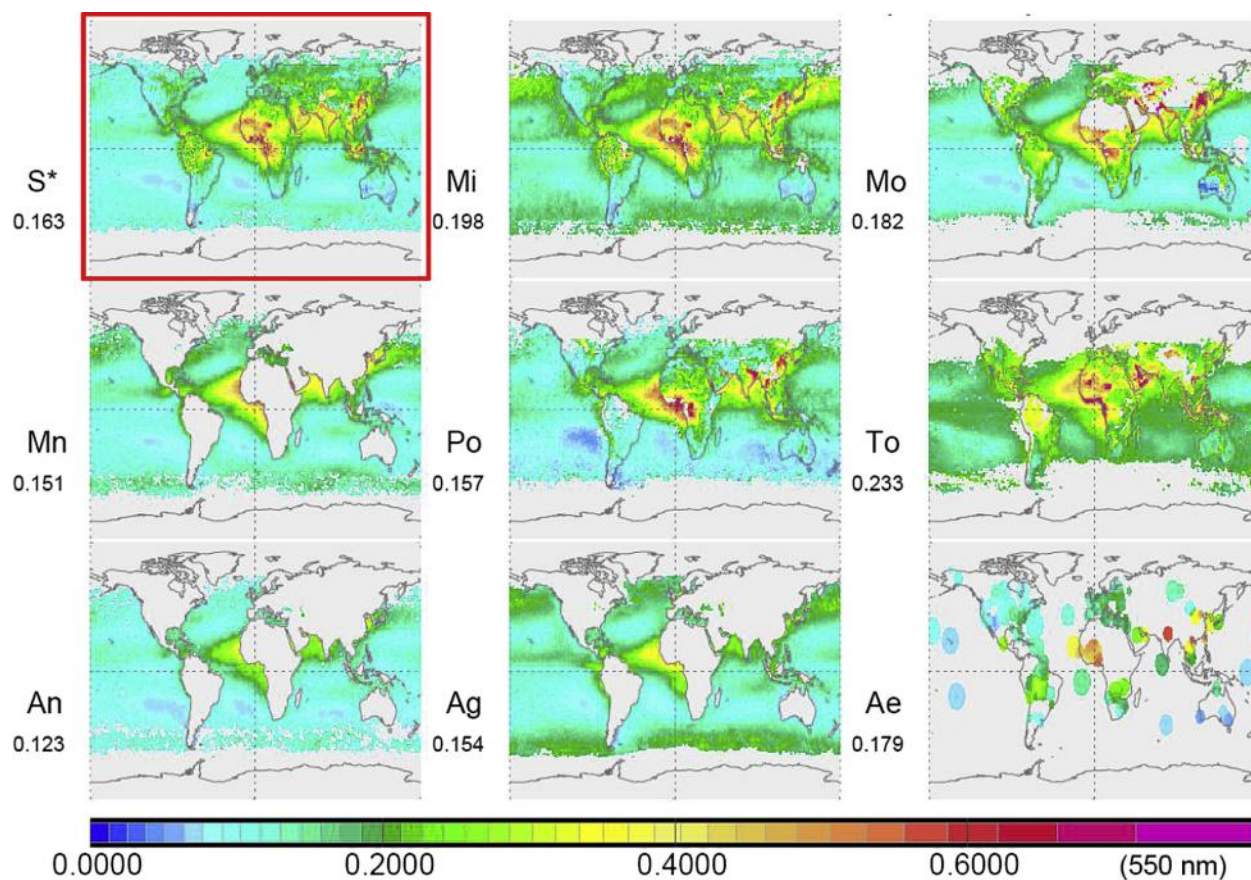


Figure 10 Maps of multiyear, annual-average, midvisible aerosol optical depth (AOD) from multiple remote-sensing data sets. A ‘synthesis’ of these AOD products (S^*), primarily from MISR, MODIS, and AERONET, used extensively for model validation, is highlighted in red. The global-average AOD for each data set is given below its label. Mi, MISR; Mo, MODIS land and water; Mn, MODIS over-ocean; Po, POLDER; To, TOMS; An, Ag, AVHRR (one- and two-channel retrieval algorithms, respectively); and Ae, AERONET ground-based sun photometer network. Reproduced from Kinne S., Schulz, M., Textor, C., Guibert, S., et al., 2006. An AeroCom initial assessment – optical properties in aerosol component modules of global models. *Atmospheric Chemistry and Physics* 6, 1815–1834.

major aerosol sources such as wildfires, volcanoes, and places where desert dust storms form. As a result, the lidar and stereo-imaging techniques are complementary; as aerosols tend to travel in discrete layers, the upwind injection heights from multiangle imaging and downwind layer heights from space-based lidar combine to provide powerful constraints on model simulations of aerosol vertical distribution.

Efforts to determine aerosol source strength from satellite observations all involve tight coupling between the available measurements and the aerosol transport models themselves. The inverse modeling approach takes the observed AOD distribution over a wide area and, in effect, runs the model backward to derive the sources. An alternative method involves running the model forward for a range of assumed source strengths and whatever constraints on injection height are available, and determines which assumed values best match the observed snapshot of AOD spatial distribution at the appropriate time step. Both methods have been demonstrated for individual cases (Figure 12), motivating continuing work to generalize these results.

One of the main applications for which frequent, global satellite aerosol observations are required is direct aerosol radiative forcing, the net change in energy flux (e.g., in W m^{-2}) at the surface produced directly by aerosol scattering and absorption. Uncertainty in the amount of surface cooling that aerosols produce is a limiting factor in determining the ability of climate models to predict changes in global mean surface temperature. Given the accuracy with which the radiative warming of long-lived greenhouse gases can be calculated, the required accuracy on aerosol properties to bring this factor into line is quite stringent. It is estimated that instantaneous, mid-visible AOD measurements need to be accurate to about 0.02 over much of the globe. Current capabilities are of the order of 0.05 or 20% of AOD, whichever is larger, over land, and somewhat better over dark water. Some improvement is expected as the algorithms used for current operational instruments are refined, but a substantial advance will likely require a next-generation space-based instrument, combined with better constraints on particle microphysical properties that will probably require a systematic program of aircraft and

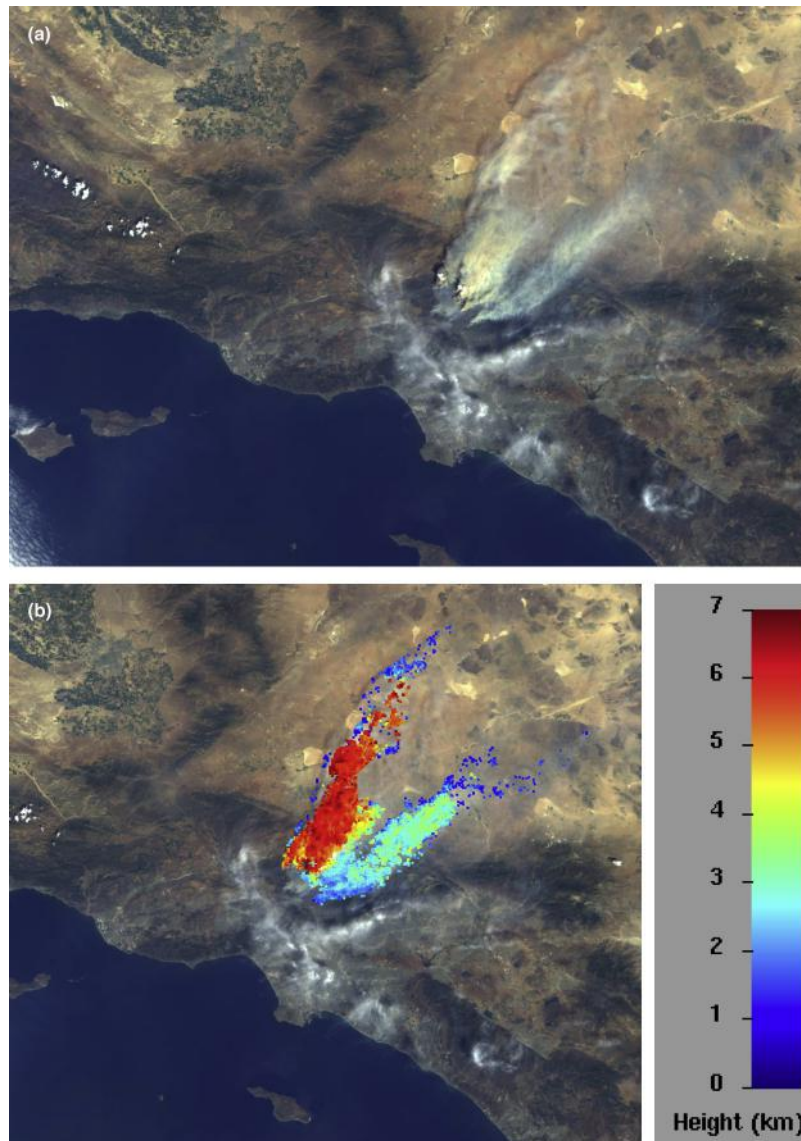


Figure 11 Satellite views of the station fire that burned in the Los Angeles, California, area in late August and early September 2009. (a) True-color image from the MISR nadir-viewing camera on 30 August 2009, showing smoke plumes from multiple active fires, and an arc-shaped cloud upwind of the fire front, possibly formed by incoming air forced to rise over the buoyant air mass heated by the fires. Pyrocumulus clouds, formed over several highly convective spots in the burning region, appear as small white dots above the smoke plume. (b) MISR stereo-derived plume heights, reported as values above sea level, for individual pixels 1.1 km in size. Provided by the MISR Team, Jet Propulsion Lab/Caltech and NASA Goddard Space Flight Center.

surface-based direct sampling along with continued surface-based remote sensing (such as AERONET).

Applications to Cloud Formation and Air Quality

One of the most challenging questions to which satellite observations have been applied is assessing aerosol impacts on clouds, often called aerosol ‘indirect’ effects, as distinct from the direct radiative forcing that aerosols produce. Aerosols are essential for cloud droplet and ice particle formation, and they

serve as collection sites for water molecules, referred to as cloud condensation nuclei (CCN) and ice nuclei (IN), respectively.

Aircraft instruments are generally best suited to study indirect effects in detail, as they can sample the tens-to-hundreds-of-meters spatial scales and minutes-to-hours temporal scales that capture the cloud development process. Nevertheless, satellite instruments have a contribution to make, taking advantage of the frequent global coverage they offer, for exploring larger-scale patterns. With imaging from passive remote sensing, cloud albedo, cloud droplet radius, and cloud optical depth can be mapped, along with cloud top

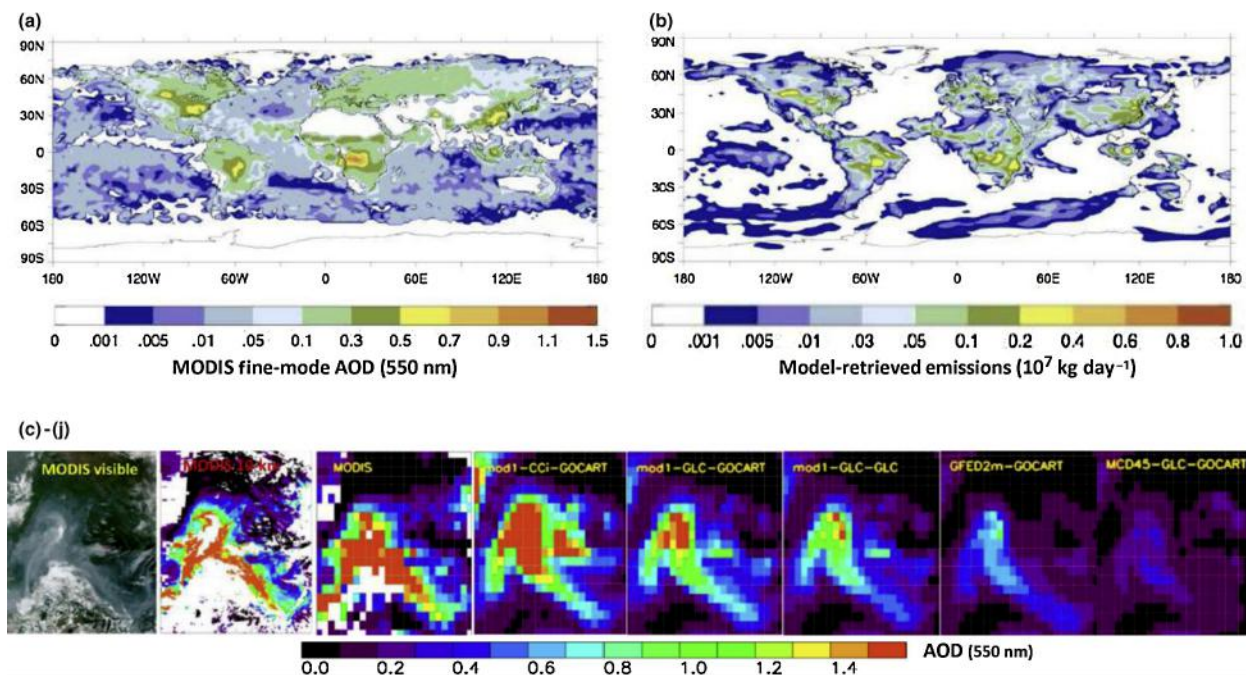


Figure 12 Characterizing aerosol source strength. (a) Average of daily global MODIS fine-mode AOD maps covering 18–30 August 2000. (b) Retrieved emission source location and strength from inverse run of the GOCART chemical transport model at $2^\circ \times 2.5^\circ$ resolution, constrained by (a). Reproduced from Dubovik, O., Lapyonok, T., Kaufman, Y.J., Chin, M., Ginoux, P., Kahn, R.A., Sinyuk, A., 2008. Retrieving global aerosol sources from satellites using inverse modeling. *Atmospheric Chemistry and Physics* 8, 209–250. (c–j, second row) Successive panels showing a MODIS-visible image of a Siberian wildfire smoke plume on 20 July 2006; a MODIS-retrieved AOD snapshot from the standard product at 10 km spatial resolution; MODIS AOD averaged to the GOCART model $1^\circ \times 1.5^\circ$ grid; and AOD snapshots taken at the MODIS overpass time from five GOCART model runs, each initialized with different, commonly used parameterizations for smoke source strength. For wildfires in different biomes around the globe, different initialization choices performed systematically better compared to the corresponding MODIS observations. Reproduced from Petrenko, M., Kahn, R.A., Chin, M., Soja, A., Kucsera, T., Harshvardhan, 2012. The use of satellite-measured aerosol optical depth to constrain biomass burning emissions source strength in the global model GOCART. *Journal of Geophysical Research* 117, D18212, <http://dx.doi.org/10.1029/2012JD017870>.

temperature and pressure, aerosol amount, and aerosol type. However, typical CCN populations are skewed toward particle sizes that remote-sensing techniques cannot distinguish from atmospheric gas molecules. So efforts have been made to identify proxies derived from those parts of the aerosol size spectrum that can be retrieved from space. In addition, the aerosol retrieval process itself is hampered by the presence of clouds – particles residing beneath clouds, in the critical cloud droplet formation regions, usually cannot be detected from orbit, and aerosol retrievals in the vicinity of clouds are often complicated by the influence of large relative-humidity gradients and cloud-scattered light. As a result, most space-based studies of indirect effects amount to mining the satellite data for correlations that test theoretical expectations about how aerosols impact clouds.

Figure 13(a)–(c) presents examples of correlations between aerosol amount and cloud droplet size, in two situations where there are both perturbed and unperturbed observations of the same environments. All other things being equal, increasing aerosol amount is expected to result in smaller cloud droplets (commonly called the ‘first indirect effect’) and, if the amount of condensed water in the cloud remains constant, higher cloud albedo. Globally, this qualitative relationship is observed in some regions more than others (Figure 13(d)); the most reliable demonstrations

involve liquid water clouds relatively near the surface. Other consequences of aerosol impacts on clouds include longer cloud lifetimes and reduced precipitation, due to smaller cloud droplets that are less likely to grow to raindrop size, although these have been more difficult to demonstrate, let alone quantify, with satellite remote sensing.

However, space-based measurements have observed cloud ‘invigoration’ (Figure 13(e)). Here, higher concentrations of CCN reduce the size of cloud droplets, inhibiting droplet coalescence until the droplets pass the freezing level. The resulting extra release of energy invigorates the cloud, increasing the cloud fraction and enhancing precipitation. Other recent work has demonstrated a relationship between aerosol concentration and open versus closed convective cell formation, and has shown a correlation between AOD and lightning occurrence as observed from space.

Air quality is monitored primarily with surface-based sampling instruments that directly observe the near-surface particle concentration and can obtain detailed information about particle size distribution and chemical composition. But as it is feasible to cover only a minuscule fraction of the planet with surface samplers, satellite observations can again play a significant role, simply by providing frequent maps of AOD during acute air pollution events. They are also beginning to contribute to longer-term exposure studies, especially when

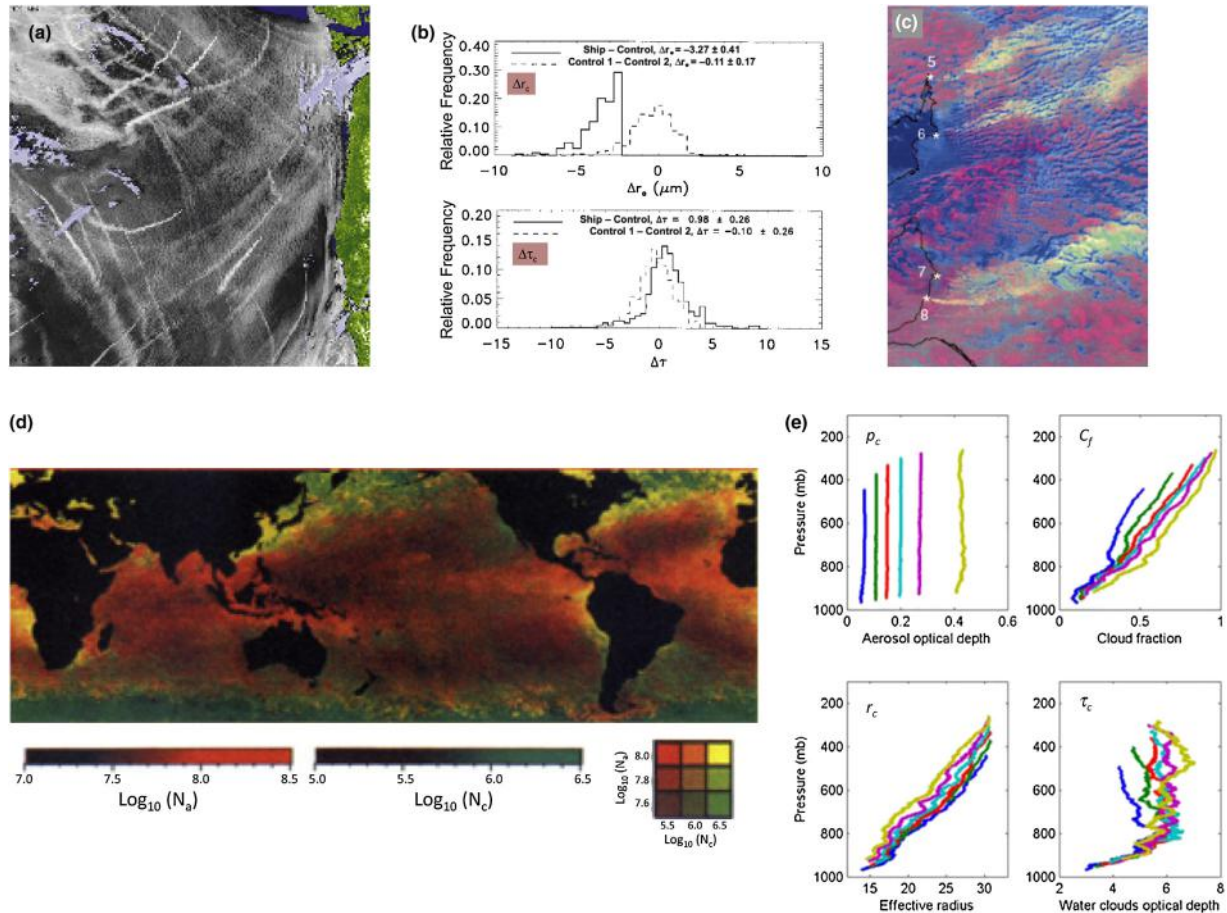


Figure 13 Indirect effects of aerosols on clouds. (a) Ship tracks off the coast of California, as viewed by AVHRR. (b) Retrieved cloud droplet radius (r_c) and cloud optical depth (τ_c) differences between observations within the polluted ship tracks and those in the surroundings, showing a decrease in droplet size in the polluted tracks. Reproduced from Coakley J.A., Walsh, C.D., 2002. Limits to the aerosol indirect radiative effect derived from observations of ship tracks. *Journal of Atmospheric Science* 59, 668–680. (c) Red color indicates regions where large droplets were retrieved in this false-color AVHRR image encoding two infrared channels and one visible spectral channel, whereas yellow trails emanate from point sources of smoke that produce smaller droplets in this fairly uniform cloud deck over south Australia. Reproduced from Rosenfeld, D., 2000. Suppression of rain and snow by urban and industrial air pollution. *Science* 287, 1793–1796. (d) Illustration of the correlation between retrieved particle number concentration (N_a) and cloud droplet concentration (N_c) globally, based on AVHRR measurements, aggregated over 4 months during 1990; yellow indicates high N_c in the presence of large N_a , whereas regions of high N_c despite small N_a appear red. Reproduced from Nakajima T., Higurashi, A., Kawamoto, K., Penner, J.E., 2001. A possible correlation between satellite-derived cloud and aerosol microphysical parameters. *Geophysical Research Letters* 28, 1171–1174. (e) Evidence from MODIS retrievals for invigoration of Atlantic convective clouds; panels show (clockwise from upper left) cloud top pressure (p_c), cloud fraction (C_f), cloud droplet effective radius (r_c), and cloud optical depth (τ_c) as a function of elevation with AOD encoded in colors, increasing from blue to red to purple and green. Reproduced from Koren, I., Kaufman, Y.J., Rosenfeld, D., Remer, L.A., Rudich, Y., 2005. Aerosol invigoration and restructuring of Atlantic convective clouds. *Geophysical Research Letters* 32, <http://dx.doi.org/10.1029/2005GL023187>.

combined with chemical transport model results that provide particle vertical distribution and chemical speciation that are otherwise lacking observationally (Figure 14).

Conclusions

Because aerosol amount and type vary on such a wide range of spatial and temporal scales, and exhibit such a diversity of environmental impacts, satellite remote sensing makes an essential contribution to the study of airborne particles. Since the very first orbiting imagers began observing Earth, the scope of satellite data products has provided inspiration, qualitative

indications, and, increasingly, quantitative constraints on the regional and global influences that aerosols exert. Major advances in this field have taken place in the last decade since the previous edition of the *Encyclopedia of Atmospheric Science* was published. New satellite-derived aerosol offerings include monthly global AOD climatologies and time series, near-source and downwind aerosol vertical distribution measurements, regional aerosol type discrimination, aerosol source strengths, and correlative analyses showing the indirect effects of aerosols on clouds. Much of this advancement has come from the current generation of satellite instruments, increasingly sophisticated aerosol field measurement campaigns, and ground-based instrument networks, and from combining

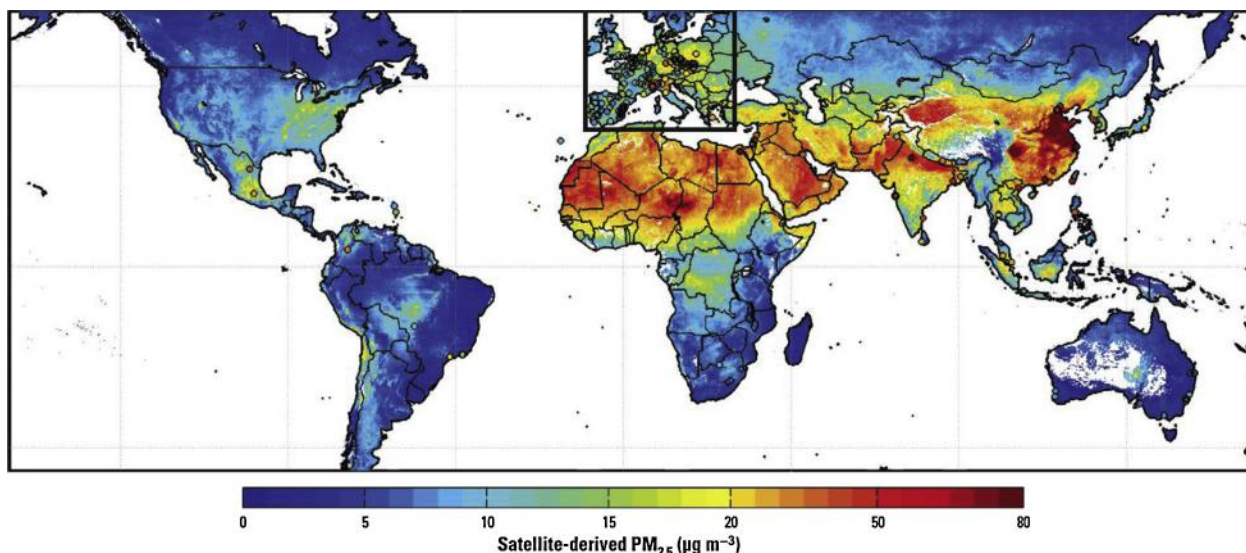


Figure 14 Six-year (2001–06), global-average, near-surface concentration of particles smaller than $2.5\ \mu\text{m}$ in diameter ($\text{PM}_{2.5}$) over land, derived from MISR plus MODIS total-column AOD, combined with vertical distribution from the GEOS-Chem chemical transport model. The vertical distributions were validated using CALIPSO profiles, and derived $\text{PM}_{2.5}$ can be compared with surface station values, which are shown using the same color scale, superposed where available within black circles in this figure. Reproduced from Van Donkelaar, A., Martin, R.V., Brauer, M., Kahn, R., Levy, R., Verduzco, C., Villeneuve, P.J., 2010. Global estimates of ambient fine particulate matter concentrations from satellite-based aerosol optical depth: development and application. *Environmental Health Perspectives* 118, 847–855.

data from multiple sources with models. Further advances can be expected as retrieval algorithms continue to be refined, as data and models are analyzed and combined in new and innovative ways, and, eventually, as next-generation space-based aerosol remote-sensing instruments are deployed.

See also: **Aerosols:** Aerosol Physics and Chemistry; Aerosol–Cloud Interactions and Their Radiative Forcing; Climatology of Tropospheric Aerosols; Observations and Measurements; Role in Radiative Transfer. **Satellites and Satellite Remote Sensing:** Precipitation; Remote Sensing: Cloud Properties.

Further Reading

- Intergovernmental Panel on Climate Change (IPCC), 2007. In: Solomon, S., Qin, D., Manning, H., Chen, Z., Marquis, M., Averyt, K., Tignor, M., Miller, H. (Eds.), *The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, Cambridge.
- Kahn, R.A., 2012. Reducing the uncertainties in direct aerosol radiative forcing. *Surveys in Geophysics* 33, 701–721. <http://dx.doi.org/10.1007/s10712-011-9153-z>.
- Kaufman, Y.J., Tanré, D., Boucher, O., 2002. A satellite view of aerosols in the climate system. *Nature* 419, 215–223.
- Ramanathan, V., Crutzen, P.J., Kiehl, J.T., Rosenfeld, D., 2001. Aerosols, climate, and the hydrological cycle. *Science* 294, 2119–2124.
- U.S. Climate Change Science Program (CCSP) Synthesis and Assessment Product 2.3, 2009. In: Chin, M., Kahn, R.A., Schwartz, S. (Eds.), *Atmospheric Aerosol Properties and Climate Impacts*, p. 116.

Earth's Radiation Budget

NG Loeb and BA Wielicki, NASA Langley Research Center, Hampton, VA, USA

© Published by Elsevier Ltd.

Synopsis

This article discusses the Earth's radiation budget and its role within the climate system. A brief summary of how Earth's radiation budget is determined from satellite observations is provided and the geographical, diurnal, and seasonal variations in its components are shown. The relationship between interannual variations in Earth's radiation budget and their relationship to natural fluctuations in the climate system such as the El Niño-Southern Oscillation are explored, as are the regional patterns of change during the past 12 years. This article also discusses the important role of clouds in modulating Earth's radiation budget at the top-of-atmosphere, surface, and within the atmosphere.

Introduction

The Earth's average temperature is relatively constant from year to year because the Earth emits as much energy to space as it absorbs from the Sun. Without this balance, the Earth would heat up or cool down. Energy from the Sun arrives at the Earth as electromagnetic waves of light that carry radiant energy, or radiation, which can be reflected, absorbed, or transmitted upon entering the Earth-atmosphere system. Most of the radiant energy from the Sun reaches Earth at visible wavelengths (primarily 0.3–2.5 μm), whereas radiation emitted by Earth occurs at wavelengths in the infrared region (primarily 3.5–100 μm). Earth's energy balance is influenced by the composition of the atmosphere, clouds, aerosols, and surface type. A steady increase in the concentration of atmospheric carbon dioxide, methane, and other trace gases since the beginning of the Industrial Revolution is causing an imbalance in the Earth's energy budget. In turn, the Earth has warmed by approximately 0.8 $^{\circ}\text{C}$ during the period 1901–2010. Far less certain are how clouds, aerosols, and the Earth's surface have changed, and how these changes have influenced the Earth's energy balance.

The regional pattern of net radiation (absorbed solar minus emitted thermal radiation) drives the atmospheric and oceanic circulations. The tropics absorb more radiation from the Sun than they emit, while the opposite is true at the poles. To make up for this latitudinal imbalance, the atmosphere and ocean transport energy from the Equator to the poles. The global average radiative imbalance between the surface and atmosphere determines how much energy is available to drive the hydrological cycle and the exchange of sensible and latent heat between the surface and atmosphere. Understanding the vertical and spatial distribution of radiation and how it changes with time is thus critical for understanding climate.

Satellite Observations

There are two major types of radiometers for measuring the reflected solar and emitted thermal radiation from Earth. Wide field-of-view nonscanning radiometers observe from horizon to horizon using either a spherical or flat-plate-type radiometer. Scanning radiometers measure radiation within a relatively

narrow field of view and provide regional coverage of Earth by scanning across the Earth from limb to limb. Scanners provide far better spatial resolution than nonscanners, but they require more complex algorithms to convert the measured radiances to radiative fluxes.

The first attempts at observing Earth's radiation budget from satellites are from the late 1950s, during the early days of the satellite era. The Explorer 7 satellite provided the first usable nonscanning wide field-of-view measurements of the total (reflected solar and emitted thermal) radiation from Earth. The 1960s was a period of rapid advances with the Television Infrared Observational Satellite series, which brought together broadband Earth Radiation Budget (ERB) instruments and five-channel scanning radiometers on the same platform for the first time. Unfortunately, the early satellite missions had lifetimes limited to roughly 1 year or less. It was not until the Nimbus-7 ERB mission in the late 1970s that an extended multiyear ERB data set was collected. ERB observed total solar irradiance and reflected solar and emitted thermal measurements with both scanner and nonscanner instruments. Another satellite mission that provided an extended record is the Earth Radiation Budget Experiment (ERBE), which consisted of scanning and nonscanning radiometers, as well as solar radiometers. The ERBE nonscanner wide field-of-view instrument on the Earth Radiation Budget Satellite (ERBS) was operational from 1985 to 2000. During the 1990s, there were two launches of the Scanner Radiation Budget (ScaRaB) instrument onboard the Meteor and Resurs satellites, which lasted less than 1 year due to the spacecraft anomalies.

Currently, there are three satellites providing global ERB observations. Clouds and the Earth's Radiant Energy System (CERES) scanner instruments are flying aboard the Terra (December 1999–present), Aqua (May 2002–present), and Suomi-NPP (October 2011–present) satellites. Each of the CERES instruments is flying with a multichannel imager, which provides higher spatial resolution data for determining coincident cloud and aerosol property information along with CERES top-of-atmosphere (TOA) radiation. This combination of instruments enables radiative fluxes at the surface and within the atmosphere to be determined at global scales. The CERES processing also utilizes narrowband visible and infrared measurements aboard geostationary satellites to provide additional information about the diurnal cycle of clouds and

radiation between 60° S and 60° N. Since the latter are intended for weather applications, they must be routinely calibrated against higher quality imager and CERES data in order to minimize the impact of instrumental artifacts on the radiation budget record.

Until the early 2000s, ERB missions were limited to either precessing or Sun-synchronous satellite orbits. This changed when Geostationary Earth Radiation Budget (GERB) instruments were launched aboard the Meteosat-8, Meteosat-9, and Meteosat-10 satellites. GERB provides reflected solar and emitted thermal radiation every 15 min over the Meteosat domain (60° S– 60° N and 60° W– 60° E). Additionally, another copy of ScaRaB was launched in October 2011 onboard Megha-Tropiques in a precessing orbit covering the tropics only. The combination of CERES, GERB, and ScaRaB flying concurrently provides unprecedented opportunities for intercalibration and improved diurnal sampling of the Earth's radiation budget.

Figure 1 shows a continuous 31-year record of tropical (20° S– 20° N) TOA broadband outgoing longwave (LW) radiation between 1979 and 2010 from nonscanner and scanner instruments. Figure 1(a) shows rather marked jumps of up to 3 W m^{-2} among the different satellites owing to

absolute calibration differences, which are within measurement uncertainty. Because there is overlap, the entire record can be placed on a common radiometric scale (Figure 1(b)). Increases in tropical mean radiation of up to 5 W m^{-2} are observed during major El Niño–Southern Oscillation (ENSO) events such as the 1997–98 El Niño. The record also provides quantitative data on the LW effect of the Mount Pinatubo eruption and subsequent recovery.

Global Energy Budget

A schematic representation of the Earth's energy budget is depicted in Figure 2. On average, 340 W m^{-2} of solar energy (or, equivalently, about three-and-one-half 100 W light bulbs every 1 m^{-2} over the surface of the Earth) reach the top of the atmosphere. Approximately one-third is reflected to space by clouds, aerosols, air molecules, and the Earth's surface. The Earth–atmosphere system loses approximately 240 W m^{-2} of infrared radiant energy to space. Owing to increased concentrations of CO_2 and other greenhouse gases, Earth is absorbing 0.5 W m^{-2} more radiation than it is emitting to space.

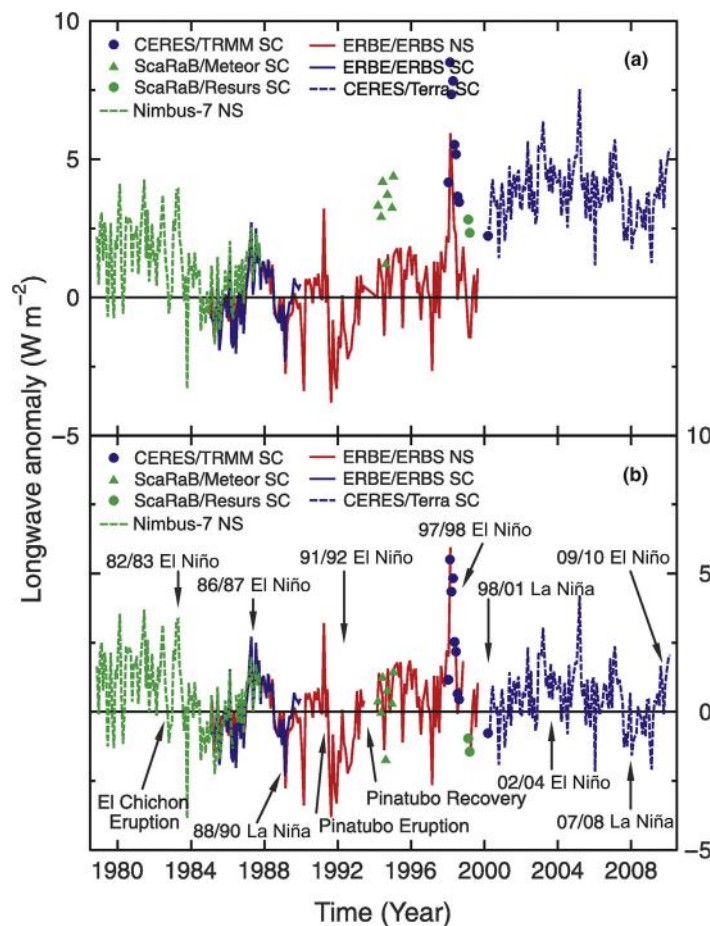


Figure 1 LW TOA flux anomalies for 20° S– 20° N from November 1978 to February 2010 (a) with no overlap correction, and (b) with overlap correction based upon an ERBS Nonscanner (WFOV Edition3_Rev1; red solid line), a Nimbus-7 Nonscanner (green dashed line), an ERBS Scanner (blue solid line), a CERES Terra cross-track (SSF1deg-lite_Ed2.5; blue dashed line), a CERES/TRMM Scanner (Edition2; blue circle), a ScaRaB/Meteor Scanner (green triangle), and a ScaRaB/Resurs Scanner (green circle). Anomalies are defined with respect to the 1985–89 period.

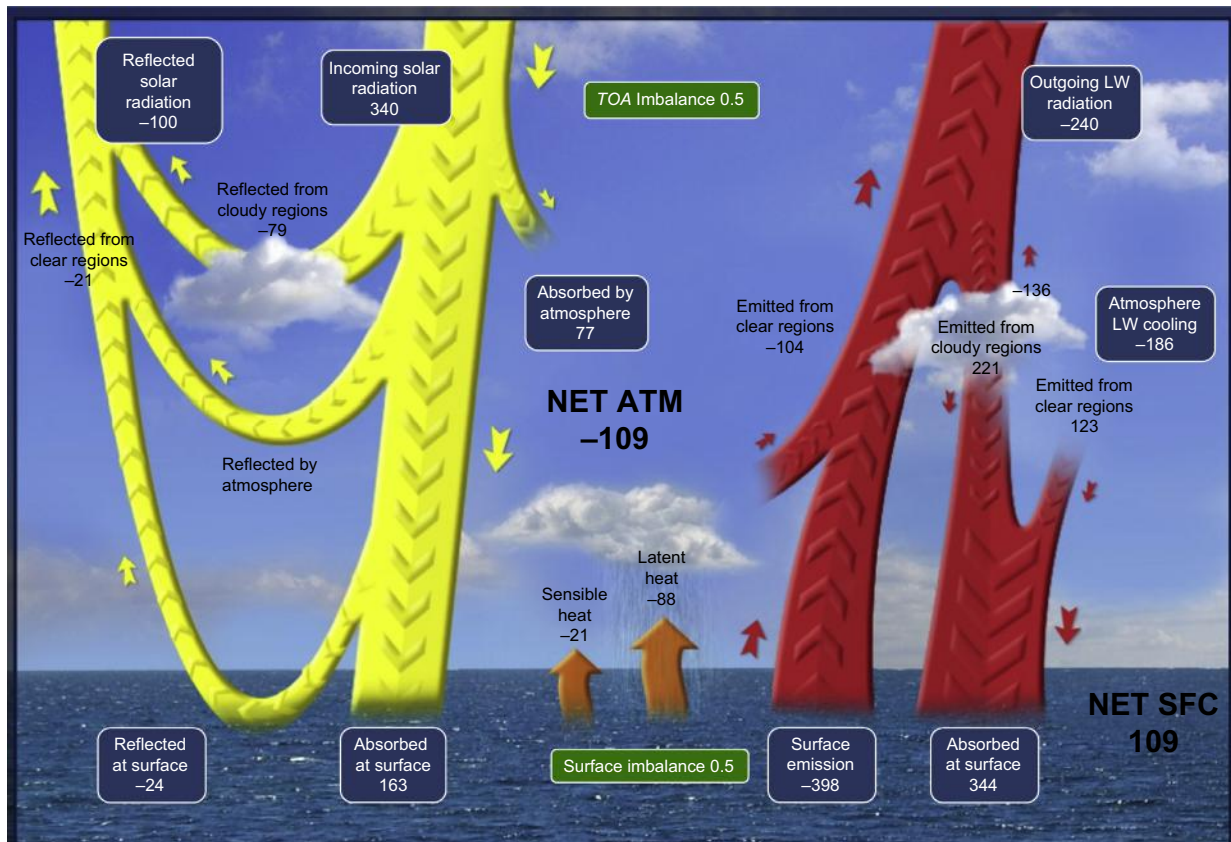


Figure 2 Earth's energy budget.

Approximately 90% of the excess energy from increased concentrations of greenhouse gases is stored in the ocean, and the remainder heats the atmosphere and land, melting snow, and ice. The planetary imbalance and its time variation are determined from a combination of *in situ* ocean temperature measurements to depths of 2000 m and satellite observations of the radiation budget.

The atmosphere absorbs 77 W m^{-2} of the Sun's energy, and 163 W m^{-2} is absorbed at the surface. The latter is used to evaporate water and heat the lower atmosphere. Most of the 240 W m^{-2} in infrared radiation emitted to space at the TOA comes from emission by air molecules and clouds, and a smaller amount ($\sim 40 \text{ W m}^{-2}$) is emitted from the surface and transmitted to space in the atmospheric 'window' wavelength regions. Note that a much larger amount of infrared radiation (398 W m^{-2}) is emitted upward from the surface. The reason why so little of the surface infrared radiation gets through the atmosphere is because atmospheric gases (mainly water vapor and carbon dioxide) and clouds absorb and then reemit 344 W m^{-2} back down to the surface. This process, known as the atmospheric greenhouse effect, is the reason why the average observed temperature at the Earth's surface is 15°C instead of -18°C , the surface temperature in the absence of a greenhouse effect. Summing the reflected solar and emitted thermal radiation terms (net radiation) reveals that there is a surplus of 109 W m^{-2} of radiant energy at the Earth's surface and a deficit of the same amount in the atmosphere. In order to compensate for this radiant energy imbalance, there is

a transfer of 88 W m^{-2} of latent and 21 W m^{-2} of sensible heat from the surface to the atmosphere.

Geographical Variations

The geographic distribution of annual average radiation at the TOA (**Figure 3**) exhibits marked variability, particularly as a function of latitude. Absorbed solar radiation (**Figure 3(a)**) reaches 360 W m^{-2} near the Equator and decreases to 50 W m^{-2} at the poles. The latitudinal difference is a consequence of Earth's geometry: the annually averaged incoming solar radiation per unit area of the Earth's surface is much greater at the Equator than at the poles. The emitted thermal radiation (**Figure 3(b)**) also declines markedly between the Equator and the poles, but the magnitude is half as large, ranging from 290 W m^{-2} at the Equator to 140 W m^{-2} at the poles. In both the absorbed solar and emitted thermal radiation patterns, the Intertropical Convergence Zone (ITCZ) is apparent as a narrow discontinuous band of convective storms resulting in reduced absorbed solar radiation and outgoing thermal radiation. Convective regions are also apparent in **Figure 3(b)** over central Africa, over the Indian Ocean and western tropical Pacific Ocean, and over the Amazon region of South America. In addition to emitting the least amount of radiation, the poles exhibit the greatest hemispheric contrast, with Antarctica emitting up to 50 W m^{-2} less infrared radiation than the Arctic at some latitudes. The net radiation at the TOA (**Figure 3(c)**) also

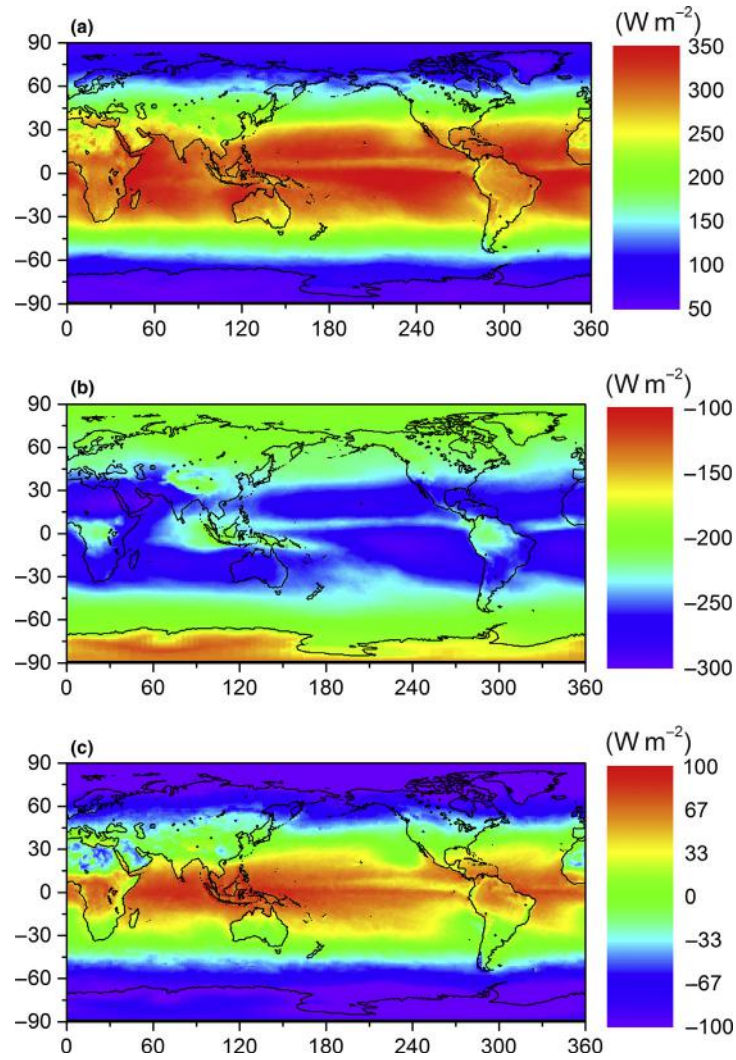


Figure 3 Geographic distribution of annual average TOA (a) absorbed solar, (b) emitted LW, and (c) net radiation for 2010 (positive downward).

exhibits marked variability with latitude, ranging from a maximum of 105 W m^{-2} in the tropics to a minimum of -130 W m^{-2} over Antarctica. The maximum net radiation into the Earth-atmosphere system occurs at the Equator over the east Indian Ocean. Net radiation is lower in the eastern Pacific Ocean and the Atlantic Ocean in the subtropics due to persistent low clouds that are effective at both reflecting solar radiation and emitting infrared radiation.

The TOA radiation averaged around latitude zones (Figure 4) shows that latitudes within 36° of the Equator gain more radiant energy than they lose through emission, while the opposite is true at higher latitudes. This imbalance drives the general circulation in the atmosphere and ocean, which transports warm air and water in the tropics poleward in order to compensate for the latitudinal net radiation imbalance. In the tropical atmosphere, a mean overturning circulation in the meridional plane, known as the Hadley circulation, is set up with ascending air near the Equator, poleward flow aloft, subsidence in the subtropical regions, and a return flow equatorward near the surface. The Hadley circulation weakens in the midlatitudes, where extratropical cyclones (i.e., weather)

provide the necessary transport of energy to the poles. In the ocean, both wind-driven surface and deep ocean circulations play an important role in meridional ocean heat transport, but the ocean contributes a much smaller portion of the required poleward transport compared to the atmosphere.

Diurnal Variations

The shape of the diurnal cycle of absorbed solar radiation at the TOA depends primarily upon solar insolation (which, in turn, depends upon latitude and season), the diurnal cycle of clouds, and surface albedo. For emitted thermal radiation, the diurnal cycle is strongly influenced by surface temperature, atmospheric profiles of temperature and humidity, and clouds. To illustrate how the diurnal cycle varies in different regions, Figure 5 shows diurnal cycles of absorbed solar, emitted thermal, and net radiation over a Saharan desert region in Algeria, over marine stratocumulus clouds off the west coast of Namibia, and over a land convection over Zambia for January 2005. The diurnal cycle in absorbed solar and net radiation

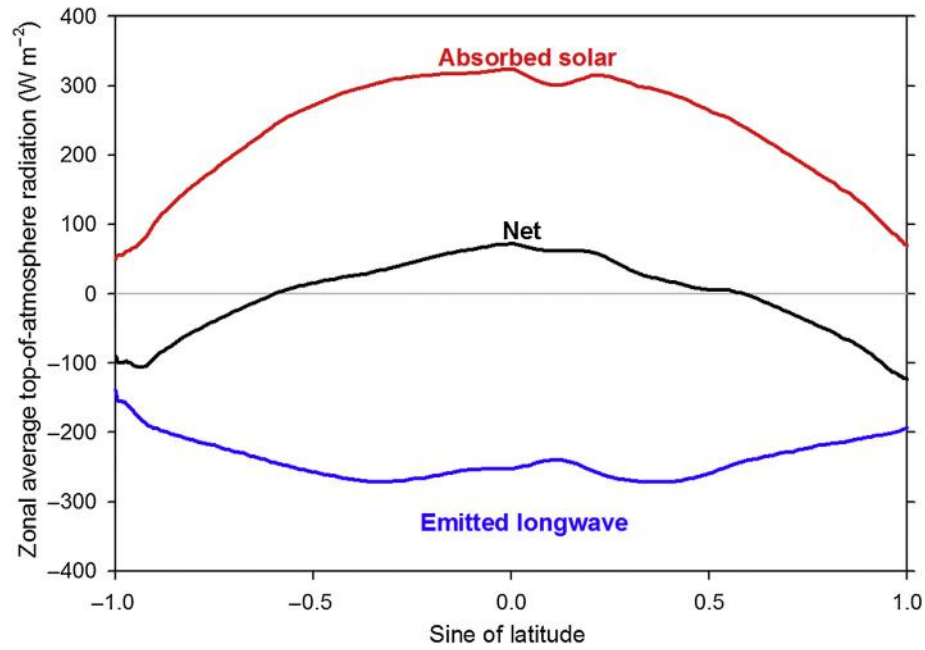


Figure 4 Zonal annual average of TOA radiation against sine of latitude for year 2010. Separate averages are provided for absorbed solar, emitted longwave, and net radiation (positive downward).

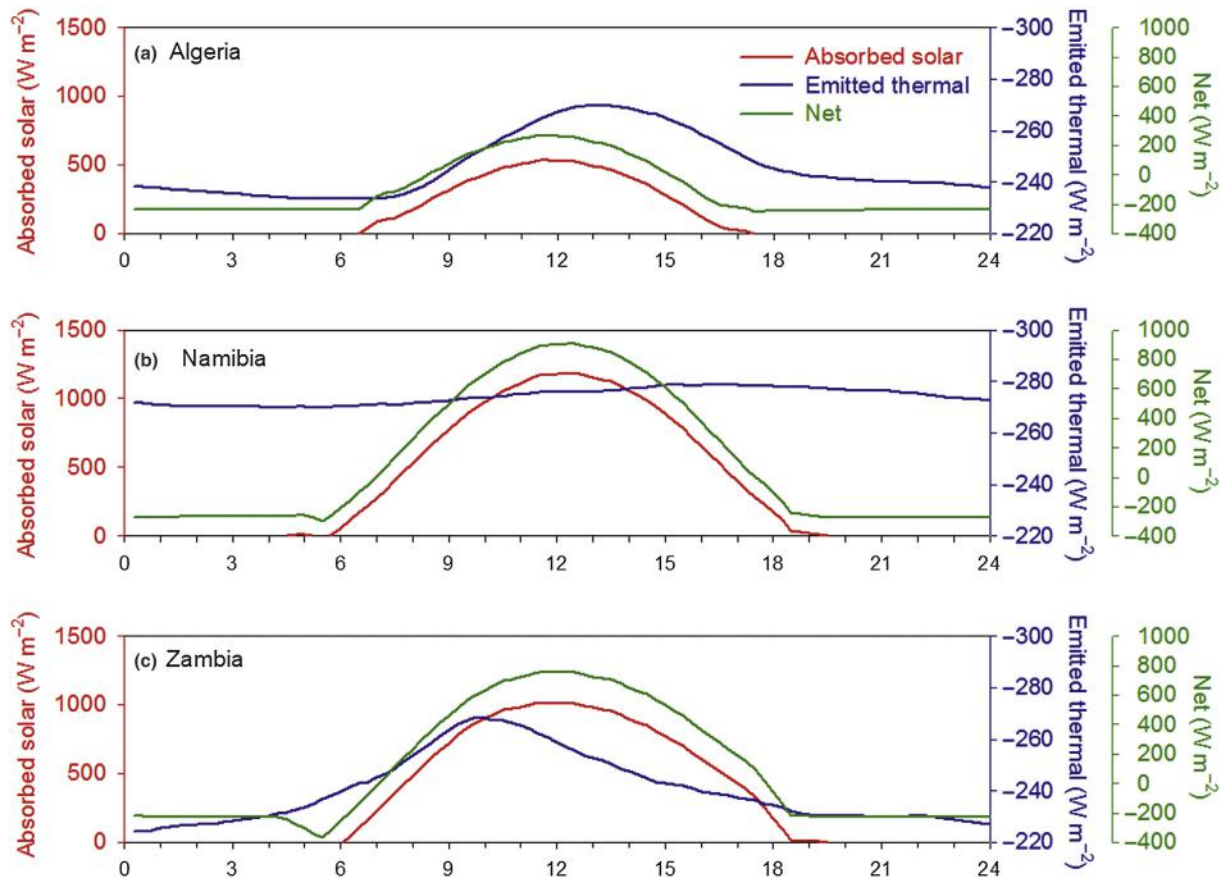


Figure 5 Monthly mean diurnal cycles of absorbed solar, emitted thermal, and net radiation for (a) the Saharan desert (25° N– 26° N, 25° E– 26° E), (b) a marine stratocumulus region (10° S– 11° S, 10° E– 11° E), and (c) land convection (0° S– 1° S, 35° E– 36° E) for January 2005.

exhibits a half-sine shape, peaking around noon, with up to 1000 W m^{-2} of solar radiation absorbed by Earth. The amplitude in absorbed solar radiation for the desert region is smaller than the other two regions because it receives less solar radiation in January due to its more northern latitude and because it has a higher surface albedo than the other regions when the incoming solar radiation reaches a maximum around noon. The peak in emitted thermal radiation over the desert region occurs at 1300 GMT and is fairly symmetric. The diurnal range during this month over the desert region is 36 W m^{-2} . In contrast, emitted thermal radiation over the marine stratocumulus region is only approximately 9 W m^{-2} because the temperature of the low clouds is similar to that of the ocean surface, resulting in little sensitivity in TOA emitted thermal radiation to changes in cloud cover. Over the land convection region, there is a marked asymmetry in the shape of the diurnal cycle, with a maximum emitted thermal radiation occurring in the morning around 1000 GMT, followed by a rapid decrease as convection builds in response to surface warming.

Seasonal Variations

The Earth moves around the Sun in an elliptical orbit with a 3.3% difference in distance between the time when the Earth is closest to the Sun (perihelion; 3 January) and when it is furthest from the Sun (aphelion; 4 July). This results in 7% more solar radiation reaching Earth in January than in July, as illustrated in Figure 6. Also shown in Figure 6 is the seasonal cycle in global mean albedo, defined as the ratio of reflected-to-incident solar radiation. Since clouds are highly reflective (with albedos typically ranging from 0.4 to 0.8) and cover approximately 65% of the Earth, they have a dominant effect on planetary albedo, reflecting approximately 20% of the solar energy reaching the planet, compared to 6% reflected from

molecules and aerosols, and 4% from the surface. Global albedo has a maximum in December, a minimum in August, and weaker secondary minima and maxima in April and May, respectively.

The seasonal cycles in absorbed solar, emitted thermal, and net radiation are shown in Figure 7. The variation in absorbed solar radiation is controlled by solar insolation and Earth's seasonal cycle in albedo, while net radiation variability also depends upon emitted thermal radiation. The Earth absorbs more solar radiation than it emits to space between October and April, when it is closer to the Sun, and emits more than it absorbs between May and September, when it is further away from the Sun. The minimum in net radiation is -10 W m^{-2} in June, 1 month prior to the minimum in solar insolation in July, and the maximum is 8 W m^{-2} in February, 1 month after the maximum in solar insolation in January. The 1-month offset from the minimum and maximum in solar insolation is due to the global albedo seasonal cycle. Remarkably, despite the large 18 W m^{-2} seasonal range in net radiation, the annual average net radiation is only approximately 0.5 W m^{-2} (Figure 2). While the mean TOA solar insolation for January–March exceeds that in October–December by 1 W m^{-2} , the absorbed solar and net radiation are 3 W m^{-2} larger during January–March. The reason is because the global mean albedo for January–March is 0.007 lower than in October–December (Figure 6).

The seasonal cycle range in global mean emitted thermal radiation is 8 W m^{-2} . The Earth emits a maximum of 244 W m^{-2} to space in August, and a minimum of 236 W m^{-2} during December. The annual cycle of emitted thermal radiation is small over the global ocean (less than 2 W m^{-2}), so variations over land are the dominant contributor to the shape of the seasonal cycle in the global mean. The maximum occurs in the Northern Hemisphere summer because the Sun is more intense during that season (due to Earth's tilt about its orbital

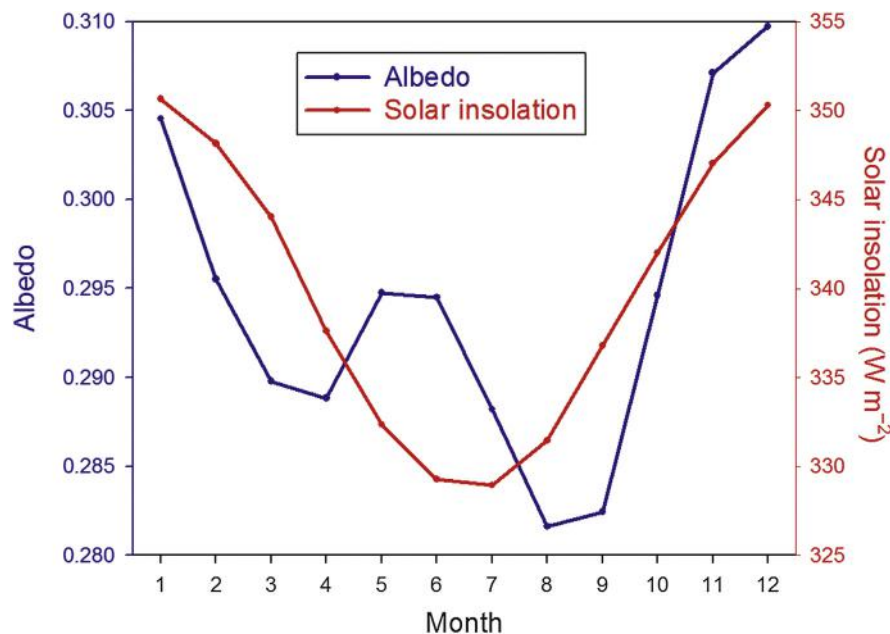


Figure 6 Seasonal cycle in TOA global albedo and solar insolation for 2010.

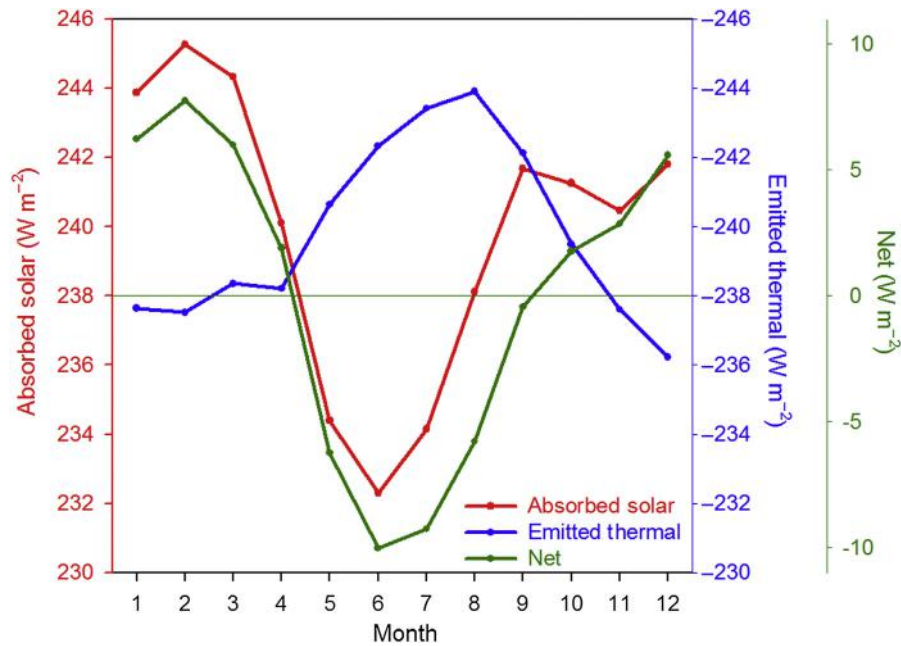


Figure 7 Annual cycle in global absorbed solar, emitted thermal, and net radiation for 2010.

plane) and most of the Earth's land mass resides in the Northern Hemisphere.

Interannual Variations and Trends

Pronounced variations in the seasonal cycle mask smaller but climatologically important year-to-year variations. Interannual anomalies are determined by subtracting the climatological monthly mean of a given month (determined from all years in the time series) from the monthly mean of each individual year for the same calendar month. **Figure 8(a)–8(c)** shows global interannual anomalies in absorbed solar, emitted thermal, and net radiation from March 2000 to June 2012 from CERES. The monthly anomalies (thin line) exhibit variability of approximately 0.5 W m^{-2} (one-standard deviation), largely due to natural variability in the climate system (ENSO, the Northern Atlantic Oscillation, weather systems, etc.). This natural variability masks any systematic trends in the data, for example those due to climate system responses to anthropogenic forcing. For these more subtle changes to be observed, a much longer observation time is needed.

During the positive phase of ENSO (El Niño), the climate system generally releases radiative energy (negative anomalies in net radiation), while the opposite is true during the negative ENSO phases (La Niña). The most prolonged ENSO event in **Figure 8** was the 2008–09 La Niña event. This period was characterized by 22 consecutive months of reduced emitted thermal radiation in the tropics (between June 2007 and March 2009), and positive anomalies in absorbed solar radiation anomalies in both the tropics and globally during much of 2008 and early 2009. The peak net incoming TOA flux anomaly occurred in June 2009, shortly after the transition from La Niña to El Niño conditions, reaching 2.5 W m^{-2} in the tropics and 1.5 W m^{-2} globally. Tropospheric air and sea surface

temperatures are generally cooler during La Niña events, even though the climate system is taking up energy. This excess energy is stored temporarily in subsurface waters at depths between 0 and 300 m, and it resurfaces during El Niño events, where it interacts with the atmosphere.

Regional patterns of change in TOA radiation between March 2000 and June 2012 (**Figure 9(a)–9(c)**) are most pronounced in the equatorial Pacific Ocean and in the Arctic Ocean. Because the record is dominated mainly by weak El Niño events prior to 2007 and much stronger La Niña events after 2007 (**Figure 9(d)**), regional trends in the tropical western Pacific Ocean show negative anomalies in absorbed solar radiation, because convection shifts more westward during La Niña events. Similarly, more solar radiation is absorbed over the central Pacific Ocean with time due to reduced cloud cover, which also results in more thermal emission of infrared radiation. The two effects largely cancel in net radiation (**Figure 9(c)**), however.

Over the Arctic Ocean, **Figure 9(a)** shows large increases in absorbed solar radiation, exceeding 10 W m^{-2} over a large area to the north of North America. While there is also an increase in emitted thermal radiation to space, the net radiation change remains positive over most of the Arctic. These changes accompany a steady decline in Arctic sea ice extent over the past several decades: satellite observations of the monthly average sea ice extent in September show a decline of 13% per decade. The process is self-perpetuating. Because of the large area of open water in late summer, the ice during the following spring tends to be thinner and therefore more vulnerable to melting, leading to further reductions in the sea ice extent. Depending upon their physical and optical properties, clouds can strongly modulate the melting by reflecting much of the solar radiation that would otherwise be absorbed at the surface, and by absorbing and reemitting surface infrared radiation (i.e., the greenhouse effect).

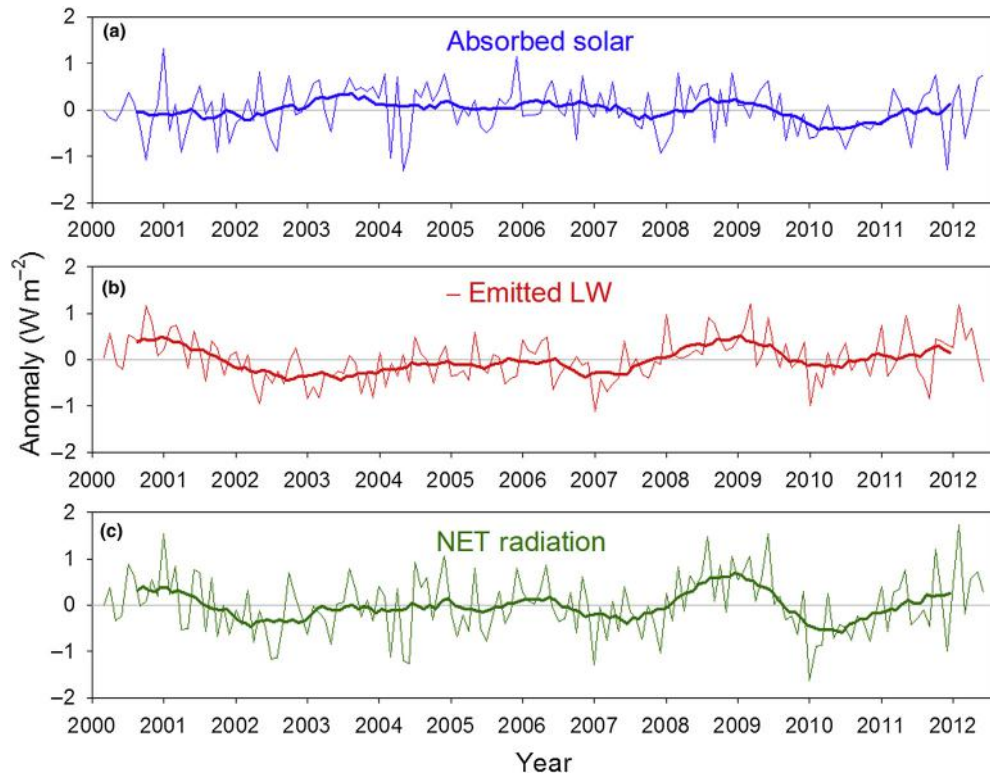


Figure 8 Global anomalies in all-sky TOA (a) absorbed solar, (b) emitted thermal, and (c) net radiation from March 2000 through June 2012. Thin lines correspond to monthly anomalies; solid line is a 12-month running mean through monthly anomalies.

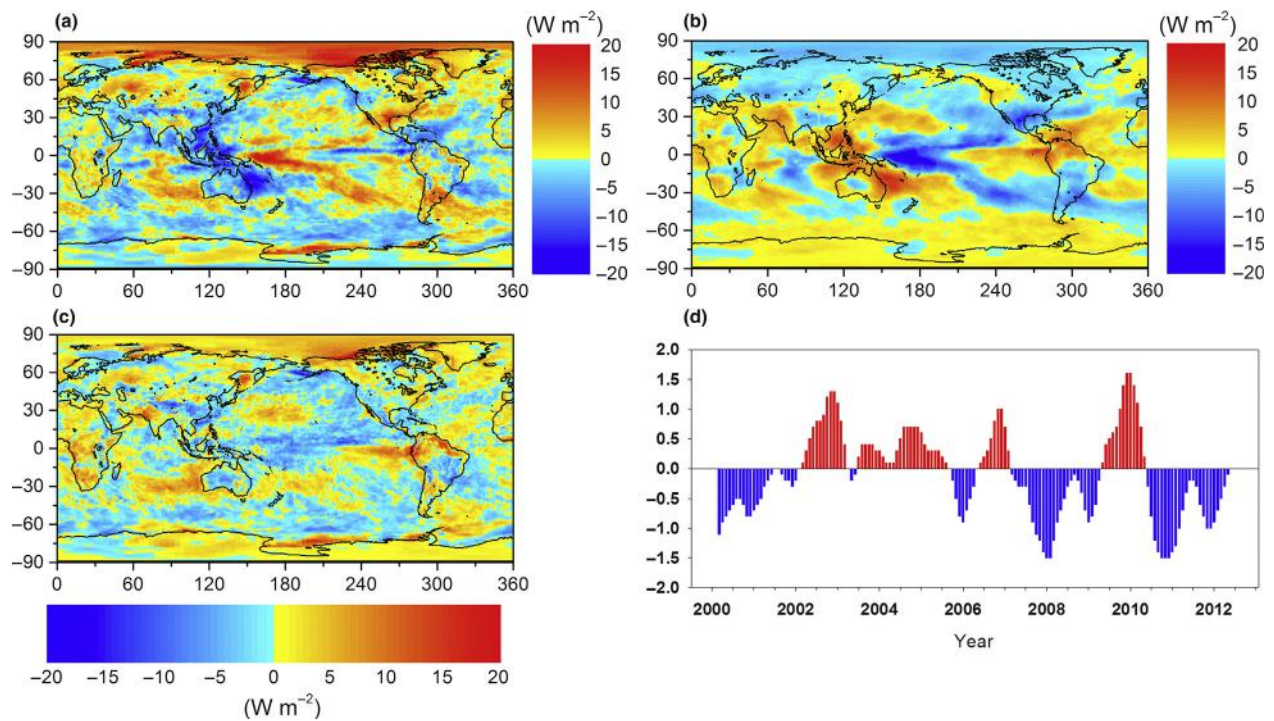


Figure 9 Regional trends in anomalies in all-sky TOA (a) absorbed solar, (b) emitted thermal, and (c) net radiation from March 2000 through June 2012. (d) Multivariate ENSO index.

Impact of Clouds on Earth's Radiation Budget at the TOA, at the Surface, and within the Atmosphere

Clouds have an important role in modulating the TOA, within-atmosphere, and surface radiation budgets at both visible and terrestrial infrared wavelengths. They reduce the amount of solar radiation absorbed by the climate system by increasing the Earth's albedo, and they decrease the loss of terrestrial infrared radiation to space through their greenhouse effect. A common approach for quantifying the influence of clouds on Earth's radiation balance is to compute the difference between the reflected solar and/or emitted thermal radiation observed under all-sky and cloud-free conditions. This difference, called cloud radiative effect (CRE), depends upon the physical and optical properties of clouds, such as their fractional coverage, optical thickness, height, and microphysical properties.

Figure 10(a)–10(c) shows the net CRE at the TOA, within the atmosphere, and at the surface. In regions where the net CRE is negative (positive), clouds have a cooling (warming) effect. At the TOA, extensive areas of net radiative cooling by clouds is observed over the subtropical eastern Pacific Ocean

off the coasts of California and Peru, over the subtropical Atlantic Ocean off the coast of Angola, over the midlatitude southern oceans between 40° S and 60° S, and over the northern Pacific and Atlantic Oceans' storm regions. Compared to cloud-free conditions, clouds in these regions are more effective at reducing how much solar radiation is absorbed by the system than they are at trapping infrared radiation through their greenhouse effect (Figure 10(a)). At the surface, their net radiative impact is much smaller due to a near cancelation of their influence on reflected solar and downward emitted thermal radiation (Figure 10(c)). This is because the base of these clouds is relatively close to the Earth's surface. The within-atmosphere CRE, given by the difference between the TOA and surface CREs, is generally negative for these clouds. This is particularly true for the stratocumulus clouds over the subtropical eastern Pacific and Atlantic Oceans, which typically reside within 2 km of the ocean surface.

A marked negative net CRE is observed at the surface in areas of convective systems, such as over the eastern Indian Ocean and the western tropical Pacific Ocean, along the Equator in the ITCZ, and in land areas over the Amazon and south central

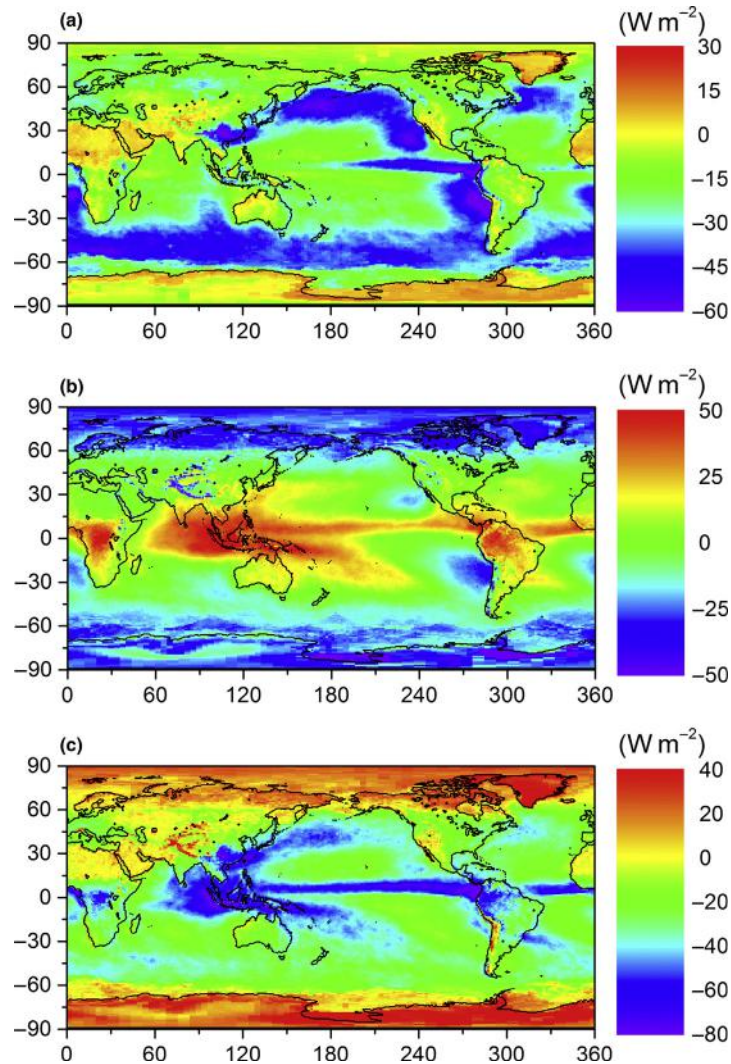


Figure 10 Net cloud radiative effect (a) at TOA, (b) within atmosphere, and (c) at surface for 2010.

Africa. These regions are largely characterized by a combination of relatively narrow vertically extensive convective cores that extend from the surface to the stratosphere, and horizontally extensive but thinner anvil and cirrus clouds, which reside in the upper troposphere at temperatures well below freezing. The latter reduce the amount of solar radiation reaching the surface but have little impact on the downwelling thermal radiation at the surface because of their cold temperatures. As a consequence, the net CRE at the surface is negative (cooling). In contrast, there is a near cancelation between reflected solar and emitted thermal effects at the TOA, as is clearly seen in **Figure 10(a)**. As a consequence, the net CRE within the atmosphere is strongly positive in the convective regions.

Polar regions show a strong positive net CRE at the surface. Since low-altitude polar clouds have similar reflectance to bright snow and ice surfaces, infrared CRE dominates the polar regions, resulting in strong warming at the surface (**Figure 10(c)**) and strong cooling in the atmosphere (**Figure 10(b)**).

Table 1 shows the global average CRE for reflected solar, emitted thermal, and net radiation. On average, the net radiative effect of clouds is that of cooling by 21 W m^{-2} compared to cloud-free conditions, at both the TOA and surface. Approximately 47 W m^{-2} more solar radiation is reflected at TOA compared to a cloud-free Earth due to the presence of clouds, and there is 51 W m^{-2} less solar radiation transmitted to the surface. Clouds reduce the outgoing thermal radiation at the TOA compared to clear conditions by 26 W m^{-2} , and they contribute approximately 30 W m^{-2} more downwelling thermal radiation at the surface. Remarkably, the solar wave and LW effects of clouds cancel within the atmosphere, despite the rather marked geographical differences (**Figure 10(b)**). It is also worth noting that there is a strong vertical gradient in the radiative cooling and heating effects of clouds: they radiatively cool at lower altitudes and warm at the upper troposphere.

The Future of ERB Observations

Because Earth's radiation budget is a fundamental property of the climate system that ultimately determines how energy heats or cools the planet, it is critically important that these measurements continue for many decades to come. Furthermore, climate models project that global temperatures will be $2\text{--}4.5^\circ\text{C}$ warmer than at the beginning of the Industrial Revolution. To place this range into perspective, the average global mean temperature during the last Ice Age was 5°C cooler than present. It is well recognized that cloud changes and their influence on the Earth's radiation budget in a warmer climate comprise the largest source of uncertainty in state-of-the-art climate models. Observations of CREs provide the only meaningful constraints on the climate

model representation of cloud processes. Unfortunately, the record of reliable global observations is still too short to use the observations to discriminate what climate models are placing Earth on the right climate trajectory. The short record is dominated by ENSO and other natural variations in the climate system that enable scientists to test only how well the models represent these short-lived fluctuations.

Extending the ERB record requires a commitment by the international community. Currently, there are plans in the United States to extend the CERES record through the end of the 2010s and into the early 2020s. A CERES instrument will fly on the first Joint Polar Satellite System satellite (JPSS-1) in 2016, and plans are underway to continue the ERB record in the next decade on JPSS-2 in 2021. There are no known plans by other countries to make ERB observations at the time of this writing.

See also: Boundary Layer (Atmospheric) and Air Pollution: Diurnal Cycle. Climate and Climate Change: Climate Variability: Seasonal and Interannual Variability; Greenhouse Effect; Overview. Clouds and Fog: Climatology. Radiation Transfer in the Atmosphere: Absorption and Thermal Emission; Cloud-Radiative Processes; Radiation, Solar; Scattering.

Further Reading

- Fasullo, J.T., Trenberth, K.E., 2008a. The annual cycle of the energy budget. Part I: global mean and land-ocean exchanges. *Journal of Climate* 21, 2297–2312.
- Fasullo, J.T., Trenberth, K.E., 2008b. The annual cycle of the energy budget. Part II: meridional structures and poleward transports. *Journal of Climate* 21, 2313–2325.
- Harries, J.E., et al., 2005. The geostationary Earth radiation budget project. *Bulletin of the American Meteorological Society* 2005, 945–960.
- Hartmann, D.L., 1994. *Global Physical Climatology*. Academic Press, San Diego, CA.
- House, F.B., Gruber, A., Hunt, G.E., Mecherikunnel, A.T., 1986. History of satellite missions and measurements of the Earth radiation budget. *Reviews in Geophysics* 24, 357–377.
- Hunt, G.E., Kandel, R., Mecherikunnel, A.T., 1986. A history of presatellite investigations of the Earth's radiation budget. *Reviews in Geophysics* 24, 351–356.
- Kato, S., Loeb, N.G., Rutan, D.A., Rose, F.G., Sun-Mack, S., Miller, W.F., Chen, Y., 2012. Uncertainty estimate of surface irradiances computed with MODIS-, CALPSO-, and CloudSat-derived cloud and aerosol properties. *Surveys in Geophysics* <http://dx.doi.org/10.1007/s10712-012-9179-x>.
- Loeb, N.G., Wielicki, B.A., Doelling, D.R., Smith, G.L., Keyes, D.F., Kato, S., Manalo-Smith, N., Wong, T., 2009. Towards optimal closure of the Earth's top-of-atmosphere radiation budget. *Journal of Climate* 22, 748–766.
- Ramanathan, V., Cess, R.D., Harrison, E.F., Minnis, P., Barkstrom, B.R., Ahmad, E., Hartmann, D., 1989. Cloud-radiative forcing and climate: results from the Earth Radiation Budget Experiment. *Science* 243, 571–623.
- Rossow, W.B., Zhang, Y.-C., 1995. Calculation of surface and top of atmosphere radiative fluxes from physical quantities based on ISCCP data set 2. Validation and first results. *Journal of Geophysical Research* 100 (D1), 1167–1197.
- Stephens, G.L., Li, J., Wild, M., Clayson, C.A., Loeb, N.G., Kato, S., L'Ecuyer, T., Stackhouse Jr., P.W., Lebsock, M., Andrews, T., 2012. An update on Earth's energy balance in light of the latest global observations. *Nat. Geosci.* 5, 691–696.
- Trenberth, K.E., Fasullo, J.T., Kiehl, J., 2009. Earth's global energy budget. *Bulletin of the American Meteorological Society* 2009, 923–933.
- Wielicki, B.A., Cess, R.D., King, M.D., Randall, D.A., Harrison, E.F., 1995. Mission to planet earth: role of clouds and radiation in climate. *Bulletin of the American Meteorological Society* 76, 2125–2153.
- Wong, T., Wielicki, B.A., Lee III, R.B., Smith, G.L., Bush, K.A., Willis, J.K., 2006. Reexamination of the observed decadal variability of the earth radiation budget using altitude-corrected ERBE/ERBS nonscanner WFOV data. *Journal of Climate* 19, 4028–4040.
- Zhang, Y.-C., Rossow, W.B., Lacis, A.A., Oinas, V., Mishchenko, M.I., 2004. Calculation of radiative fluxes from the surface to top of atmosphere based on ISCCP and other global data sets: refinements of the radiative transfer model and the input data. *Journal of Geophysical Research* 109, D19105. <http://dx.doi.org/10.1029/2003JD004457>.

Table 1 Global average radiative effect of clouds on reflected solar (SW), emitted thermal or longwave (LW), and net (NET) radiation at the top of the atmosphere (TOA), within the atmosphere (ATM), and at the surface (SFC) from March 2000 to February 2010

	SW	LW	NET
TOA	−47	26	−21
ATM	4	−4	0
SFC	−51	30	−21

GPS Meteorology

SS Leroy, Harvard School of Engineering and Applied Sciences, Cambridge, MA, USA

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

The signals of the Global Positioning System (GPS) encounter refraction in the Earth's atmosphere. In ground-based GPS meteorology, GPS signals are observed at a ground station, and the data can be inverted for precipitable water with accuracy and precision of 2%. In GPS radio occultation (RO), GPS signals are observed in low Earth orbit after traversing the atmosphere, and the data can be inverted for profiles of the microwave refractivity, pressure, and temperature. Between 8 and 30 km, the retrieved refractivity is precise to 0.2%. Both ground-based GPS and GPS RO are ideally suited to climate monitoring and numerical weather prediction.

Introduction

When a microwave signal propagates through the Earth's atmosphere, it is retarded by an index of refraction that is greater than 1; therefore, a signal emitted by a satellite of the Global Positioning System (GPS) is delayed in its arrival at a receiver with respect to its arrival if no atmosphere were present. The index of refraction varies substantially in the atmosphere, with most of the gradient in the vertical direction. Not only is the signal delayed by the atmosphere, but also it is bent in the direction of the gradient to effectively minimize the travel time from the transmitter to the receiver. In GPS meteorology, which is the sounding of the atmosphere for meteorological purposes using the signals of the GPS, the retardation and the refraction of the signal are the two phenomena that enable meteorological sounding of the atmosphere. The two techniques of GPS meteorology are *ground-based GPS meteorology* and *GPS radio occultation (RO)*. Ground-based GPS meteorology is the recording of a GPS signal by a GPS receiver on the Earth's surface. Its primary information is a quantification of the vertically integrated amount of water vapor above the GPS receiver. GPS RO is the recording of a GPS signal by a GPS receiver in a low Earth orbit (LEO). Its primary information is a profile of the microwave index of refraction as a function of height in the atmosphere. Both methods of GPS meteorological sounding have come to play an important role in numerical weather prediction, in atmospheric dynamics, and in climate monitoring. There is little bending present in ground-based GPS meteorology because the GPS signal propagates only a short distance through the atmosphere. Bending of the signal is a pronounced effect in GPS RO because the GPS signal propagates a great distance nearly orthogonally to the gradient of the index of refraction. Processing GPS RO data is consequently a more complex endeavor than processing ground-based GPS data. Both ground-based GPS meteorology and GPS RO are described in this article.

GPS is a constellation of satellites deployed by the US Air Force for the primary function of high-precision positioning. Nominally, there are at least 24 satellites in the GPS system at any given time, but because of continuing replenishment of GPS with new satellites bearing improved technology, there are usually several more than 24 GPS satellites in operation contemporaneously. Each satellite of GPS illuminates the Earth with a set of microwave L-band signals. In GPS

meteorology, the most relevant of the signals is the 'coarse acquisition' (C/A) signal because of its amplitude and because it is modulated by a publicly available pseudorandom code. The C/A signal is broadcast at the L1 frequency, which is 1.57542 GHz. Several other signals are also transmitted, one also at the L1 frequency but encrypted by a secure and publicly unavailable pseudorandom code, and another at the L2 frequency that is secured by the same pseudorandom code. The L1 C/A and the secure L2 signals are tracked in GPS meteorological applications. On average, a GPS receiver on the ground tracks approximately six GPS satellites at any given time.

Other constellations of satellites intended for precise positioning are currently being deployed and should also be useful for meteorological purposes. The European Space Agency is deploying the Galileo system, the Peoples' Republic of China the Compass system, and Russia the GLONASS system. All the positioning systems are referred to collectively as the Global Navigation Satellite Systems (GNSS), and it has become common to refer to the sounding techniques based on GNSS as ground-based GNSS and GNSS RO. At present, GPS remains the only operational GNSS, so only ground-based GPS and GPS RO are discussed here.

Ground-Based GPS Meteorology

In ground-based GPS meteorology, the delay of GPS signals by the atmosphere as received at a ground station is inverted for column-integrated water vapor (IWV), or precipitable water. Because ground-based GPS meteorology is based on timing measurements and is ultimately calibrated using atomic clocks, this technique for measuring precipitable water is widely regarded as the most accurate one and is therefore ideally suited to climate monitoring. It is unfeasible to obtain global coverage of ground-based GPS meteorology because such measurements can be obtained only at fixed stations on the ground where GPS receivers connected to fixed antennas have been placed.

GPS signals traveling through the atmosphere take longer to arrive at a GPS receiver on the ground than if no atmosphere were present. The excess time is multiplied by the speed of light to form the path delay, which has a dimension of distance. The path delay increases with an increasing mass

of air between the transmitting GPS satellite and the receiver, and also increases with increasing water vapor concentration along the path of the ray. When the transmitting GPS satellite is exactly at zenith, or directly overhead, the path delay is simply the integral of the microwave index of refraction n less 1 in altitude:

$$\text{ZPD} = \int_{\text{receiver}}^{\text{transmitter}} (n - 1) dz. \quad [1]$$

Relating the zenith path delay (ZPD) to atmospheric quantities requires an expansion of the microwave index of refraction in terms of state variables of the atmosphere, which is given by:

$$n - 1 = \left(c_1 \rho_d + c_2 \rho_v + c_3 \frac{\rho_v}{T} \right) \times 10^{-6}, \quad [2]$$

in which ρ_d is the density of dry air, ρ_v the density of water vapor, and T the atmospheric temperature. The constants are $c_1 = 222.76 \text{ m}^3 \text{ kg}^{-1}$, $c_2 = 299.0 \text{ m}^3 \text{ kg}^{-1}$, and $c_3 = 1.743 \times 10^6 \text{ K m}^3 \text{ kg}^{-1}$. This expression for the microwave index of refraction is based on a combination of first principles and empirical measurement and is accurate to 0.1%. Vertical integration of the first term in parentheses is proportional to surface pressure because it is a hydrostatic integral. As a result, the ZPD of eqn [1] is separable into a component that is due to surface pressure (the *zenith hydrostatic delay* (ZHD)) and a component that is due to water vapor density (the *zenith wet delay* (ZWD)). The ZHD is the product of $2.2715 \text{ mm hPa}^{-1}$ ($2.2715 \times 10^{-5} \text{ m Pa}^{-1}$) and the surface pressure. This makes the ZWD:

$$\text{ZWD} \approx 10^{-6} \int \left(c_2 \rho_v + \frac{c_3 \rho_v}{T} \right) dz. \quad [3]$$

The ZWD itself is related to the column IWV density. Defining a mean temperature T_m that is the inverse of the water vapor column-weighted inverse temperature,

$$T_m = \frac{\int_{\text{surface}}^{\infty} \rho_v dz}{\int_{\text{surface}}^{\infty} (\rho_v / T) dz}, \quad [4]$$

the column IWV is related to the ZWD by:

$$\text{IWV} = \text{ZWD} \times 10^6 \left(c_2 + \frac{c_3}{T_m} \right)^{-1}. \quad [5]$$

The column IWV is the mass per unit area of water vapor above the ground station. It is related to precipitable water by division of the density of water, 1000 kg m^{-3} . Once the ZPD is determined at a ground station, the ZHD as determined by a measurement of surface pressure is subtracted from the ZPD, and this last equation is applied to the remaining ZWD to retrieve column IWV. The ZHD is $\sim 2.3 \text{ m}$ and can be estimated with the same fractional uncertainty as that of the pressure measurement. The ZWD ranges from 0 to 0.40 m and is known much less precisely than the ZHD. Thus, ground-based GPS meteorology is used to retrieve column IWV. Figure 1 shows a month-long time series of ZWD and column IWV for a GPS ground station.

There are several complicating factors in the retrieval of column IWV by ground-based GPS meteorology. The most dominant one is the specification of the mean temperature T_m . The mean temperature is a function of the vertical distribution of water vapor and temperature and therefore varies in time and space. A single best estimate of the mean temperature is 260 K, but the true value of T_m varies with a standard deviation of 20 K according to location and season, which propagates into a 15% uncertainty in column IWV. However, a measurement of surface air temperature at the location of the GPS receiver reduces the uncertainty in the retrieved column IWV to 2% because the vertical lapse rate of temperature in the troposphere varies little.

Another complicating factor in ground-based GPS meteorology is the mapping of the off-zenith measurement of path delay, or the *slant path delay*, to a ZPD. In the process of ground-based GPS meteorology, the actual measurement is the slant path delay because the transmitting GPS satellite is never exactly overhead. The slant path delay measurements must be mapped to ZPD. The mapping differs from a simple multiplication by cosine zenith angle because the GPS signal does not follow a straight line through the atmosphere when

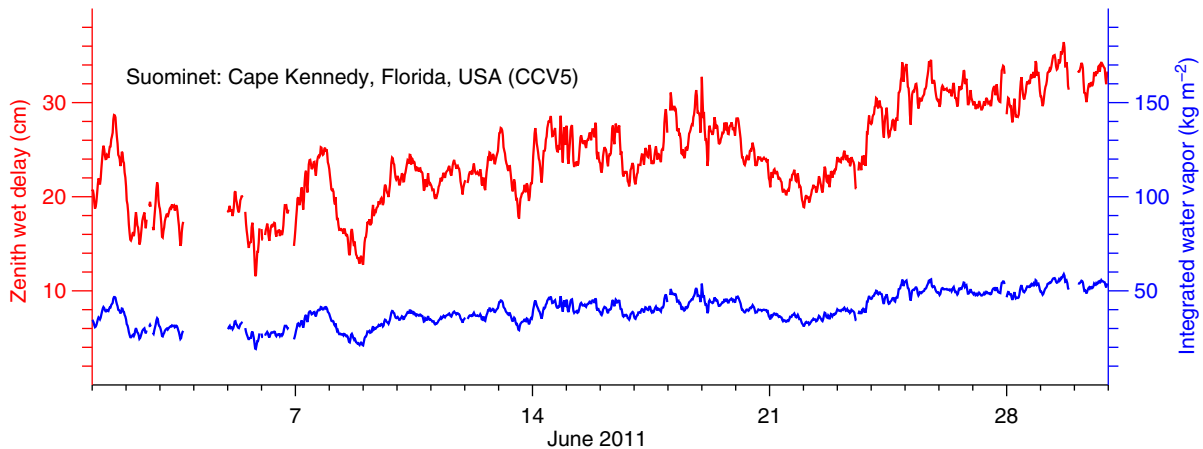


Figure 1 Zenith wet delay (ZWD) and column-integrated water vapor (IWV) every half hour from the SuomiNet station at Cape Kennedy, Florida, United States, for the month of June 2011.

the GPS satellite is not at zenith. The slant path delay P for a GPS signal is:

$$P = \int_{\text{receiver}}^{\text{transmitter}} nds - L, \quad [6]$$

in which L is the straight-line distance between the transmitting GPS satellite and the ground-based receiver and the integral is along the path of the ray. The strong variation of the index of refraction in altitude causes GPS rays to bend downward as they propagate through the atmosphere, making the path of the integral longer than the straight line connecting the transmitter and the receiver. Hence, the path delay is positive. There is no single expression that can map non-ZPD to ZPD with absolute precision based on the GPS data alone. Mapping functions have been proposed out of necessity, and they take the following form:

$$P = \text{ZHD} \times M_H(\zeta) + \text{ZWD} \times M_W(\zeta). \quad [7]$$

The function $M_H(\zeta)$ maps the ZHD to the slant path delay, and the function $M_W(\zeta)$ maps the ZWD to the slant path delay. Many expressions have been formulated for the two mapping functions, the most complex ones involving many parameters besides the zenith angle ζ and thereby obtaining greater precision. For zenith angles less than $\sim 75^\circ$, the two mapping functions take the same form, and they can map path delay to ZWD with an uncertainty of 0.7 cm (0.007 m). Mapping functions are too imprecise for zenith angles greater than $\sim 75^\circ$, and so ground-based GPS meteorology typically restricts observations to zenith angles less than 75° .

By comparison to other more direct measurements of precipitable water, ground-based GPS measurements of precipitable water are precise to 2 mm (0.002 m) plus 2% of the precipitable water. The measurements are accurate to 2 mm (0.002 m).

Several properties of ground-based GPS meteorology make it an attractive technique for measuring column IWV. First, knowledge of column IWV should constrain the prediction of precipitation in major storms, and so ground-based GPS meteorology is useful in hazard mitigation. Second, it can operate nearly autonomously, yielding estimates of column IWV every 10–30 min. Third, it is mostly insensitive to the presence of clouds, which corrupt the retrieval of column IWV from most remote-sensing data. Fourth, GPS receivers suited to ground-based GPS meteorology are typically inexpensive, their cost being approximately US\$1 million. Finally, because ground-based GPS meteorology is ultimately a timing measurement, which can be calibrated against atomic clocks, it is ideally suited to long-term climate monitoring. Its primary limitation is that it is ground based, meaning that it cannot practically obtain global coverage. Its primary competition for climate monitoring, aside from GPS RO (discussed further in this article), is passive microwave sounding and water vapor radiometry. While passive microwave sounding and radiometry from space do obtain global coverage, their sensitivity to clouds inhibits the precision of the retrieval of column IWV. Furthermore, because passive microwave sounding from space is a measurement of millimeter-wave radiation emitted from the Earth's atmosphere, it is much more difficult to calibrate than GPS

measurements. Water vapor radiometry is difficult to calibrate as well, because it is a broadband measurement.

At present, many networks of ground-based GPS receivers are used to monitor column IWV, and some have been deployed specifically for that purpose. Assimilation of slant path delay into numerical weather prediction generally shows improvement in the forecasting of precipitable water and precipitation out to 12 h in the future but weaker impact further into the future. Results have been less conclusive regarding predictions of precipitation amount, temperature, geopotential height, and winds. As of 2008, the time series of precipitable water measurements have not covered a period long enough to detect anything more than a marginally significant upward trend. Figure 2 is a map of the GPS ground stations of SuomiNet, which is dedicated to monitoring column IWV. It is just one of many such networks worldwide.

GPS RO

Whereas ground-based GPS meteorology is based on the reception of GPS signals with a receiver on the Earth's surface, GPS RO is based on the reception of GPS signals with a receiver in LEO. From the perspective of a LEO satellite with an orbital period of ~ 100 min, each GPS satellite can appear to rise or set on the Earth's horizon several times daily. During these events, the Earth's atmosphere is said to occult the GPS satellite, and data are collected from the time that the GPS ray is 100 km above the Earth's surface until the Earth's surface itself occults the signal. This is a single RO. The signal's phase and amplitude as a function of time are the recorded data. Since the atmosphere has a microwave index of refraction that mostly decreases with height, the transmitted signal is refracted downward as it transects the atmosphere (cf. Figure 3). The bending of the signal increases the optical path delay well beyond what would have been expected if no atmosphere were present. Data on the optical path delay can be transformed to an estimate of the total angle through which the signal is bent, which in turn can be transformed into a profile of the refractive index of the atmosphere. Subsequent information can be incorporated to retrieve profiles of temperature, pressure, and water vapor as functions of geopotential height.

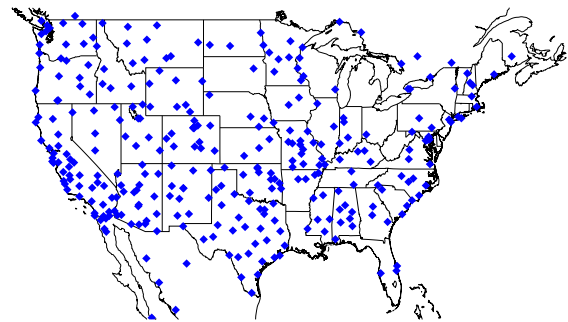


Figure 2 A map of the locations of the ground stations of SuomiNet, used in monitoring column-integrated water vapor in the United States.

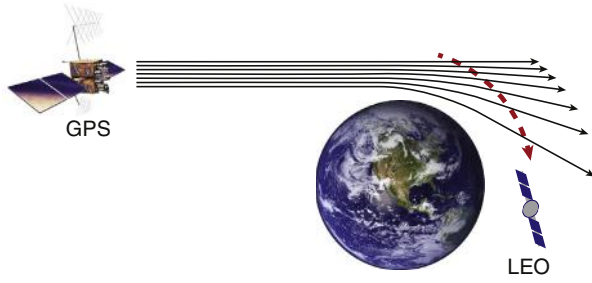


Figure 3 Spreading of GPS rays by refraction in the Earth's atmosphere. A spacecraft in LEO (dashed red line) receives the bent GPS signal.

As in ground-based GPS meteorology, GPS RO is a timing measurement that can be absolutely calibrated with an accuracy of <1 part in 10^{15} . It, too, is insensitive to clouds. Because it is space based, the coverage is global but also nonuniform. The speed of a LEO satellite in its orbit and the large number of satellites in the GPS constellation combine to yield many hundreds of RO events daily for a single LEO satellite. GPS RO is at its most precise and accurate in the region spanning approximately 8–30 km in height, commonly referred to as the upper troposphere and lower stratosphere. Below that region, the complexity of signal dynamics in a highly heterogeneous water vapor field prevents regular and precise tracking of the GPS signal. Above that region, the residual ionospheric influence and the weak bending of the GPS signal in the presence of noise degrade the accuracy and precision of the measurement. Because it records the phase of the GPS signal during an occultation event, methods paralleling those of holography can be applied to obtain vertical resolution in the atmosphere that is Fraunhofer limited rather than Fresnel limited. Accounting for smoothing of retrieved quantities, the vertical resolution is approximately 100–200 m.

The briefest description of the relationship between the refractive properties of the atmosphere and the measured phase and amplitude in a GPS RO is Hamiltonian optics, in which the eikonal Φ is the path delay (in radians) as a function of space and time and the Hamiltonian is the angular frequency $H = \omega = |\mathbf{k}|c/n(\mathbf{q})$, where $\mathbf{k}$ is the wavevector, c the speed of light, and $n(\mathbf{q})$ the index of refraction as a function of the coordinates $\mathbf{q}$. Since the Earth is close to spherically symmetrical on scales of tens to hundreds of kilometers in the horizontal, it is appropriate to consider the orbital azimuth θ as a Hamiltonian coordinate (cf. Figure 4 for geometry). The

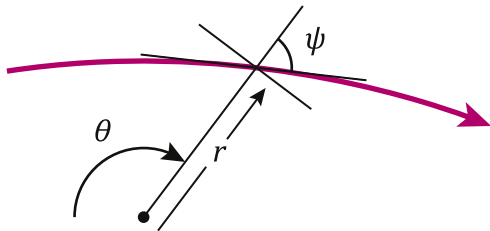


Figure 4 Geometry of the ray path. The orbital azimuth is θ , the orientation of the ray with respect to the local vertical is ψ , and the radial coordinate is r .

momentum conjugate to the orbital azimuth is $p_\theta = -(\partial\Phi/\partial\theta) = nrk_0 \sin \psi$, in which k_0 is 2π divided by the vacuum wavelength of the GPS carrier and ψ is the angle between the radial and propagation directions. In Hamiltonian systems, the time tendency of momentum is $dp_\theta/dt = -\partial H/\partial\theta$, which becomes:

$$\frac{dp_\theta}{ds} = k_0 \frac{dn}{d\theta}, \quad [8]$$

in which s measures distance along the path of the ray. Consequently, $nr \sin \psi$ is a constant along the path of the ray if the atmosphere is perfectly spherically symmetric meaning n does not depend on θ . The conservation law $nr \sin \psi = a$ is known as Bouguer's law, and, by trigonometry, a is the impact parameter, or the ray's asymptotic miss distance from the center of curvature of the Earth at the transmitter or the receiver (cf. Figure 5 for geometry). The angle through which the ray bends in the course of its propagation, in differential form along the path of the ray, is $d\alpha = d\psi + d\theta$. The application of trigonometry and Bouguer's law gives the following differential form for the bending angle along the path:

$$d\alpha = -\tan \psi \frac{dn}{n} = \pm \frac{a}{\sqrt{n^2 r^2 - a^2}} \frac{dn}{n}. \quad [9]$$

The $\pm$ is positive for downward propagation into the atmosphere and negative for upward propagation out of the atmosphere. Integrating along the path of the ray gives a function defining the bending angle as a function of the impact parameter:

$$\alpha(a) = -2a \int_{r_{\min}}^{\infty} \frac{d \ln n/dr}{\sqrt{n^2 r^2 - a^2}} dr. \quad [10]$$

This is the forward model for the observation given the profile of the index of refraction as a function of radial distance $n(r)$. The profile of the index of refraction is a function of pressure, temperature, and humidity in the atmosphere as prescribed by eqn [2].

In GPS RO, while the fundamental observable is the phase and amplitude of the GPS signal as a function of time, the derived quantity upon which all subsequent processing is based is the bending angle as a function of the impact parameter $\alpha(a)$. An Abelian integral transform of the forward model for the bending angle is applied to retrieve the profile of the index of refraction as a function of the impact parameter:

$$\ln n(a) = -\frac{1}{\pi} \int_a^{\infty} \frac{\alpha(a')}{\sqrt{a'^2 - a^2}} da'. \quad [11]$$

The dependence of the retrieved index of refraction on the impact parameter of the ray is converted to dependence on radial distance by determining the value of the radial distance at the ray's minimum radius, which is the largest r that satisfies $n(r) \cdot r = a$. Spherical symmetry is assumed, leading to Bouguer's law and the above derivation.

For the Earth's atmosphere, the refractivity, or $N = (n - 1) \times 10^6$, is roughly 330 at the Earth's surface. In the tropics, it can be as large as 400. While the 'wet' terms (c_2 and c_3 terms of eqn [2]) exceed 30% of the dry term (c_1 term of eqn [2]) only in the tropics, in the lower

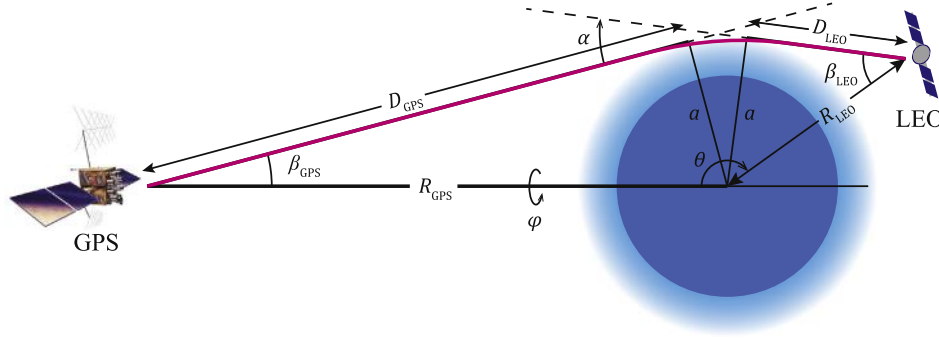


Figure 5 Geometry of a GPS radio occultation. See the text for the description of the quantities. The impact parameter a and the radial distances R_{GPS} and R_{LEO} to the GPS and LEO satellites, respectively, refer to the local center of curvature of the Earth's surface nearest the closest approach of the ray, which strays from the Earth's center by a few tens of kilometers.

atmosphere the wet terms combined are far more variable than the dry term. The typical bending angle at the Earth's surface is 1° , but it becomes much larger in the presence of critical refraction and near-critical refraction (discussed further in this article).

RO: Geometric Optics

The bending angle as a function of impact parameter can be derived from the observation of the phase of the GPS signal during a RO event and subsequently transformed to a profile of the index of refraction. The difference between the rate of change of the signal's optical path and the rate of change of the distance between the GPS and LEO satellites is known as the excess phase rate and is easily determined by the measurement of phase and the satellites' positions. The excess phase rate is nonzero only if the ray is refracted, causing retardation and bending of the signal. The excess phase rate $\dot{L}$ is related to the GPS satellite's and the LEO satellite's velocities by:

$$\dot{L} = \mathbf{v}_{\text{LEO}} \cdot (\mathbf{s}_{\text{LEO}} - \mathbf{s}) - \mathbf{v}_{\text{GPS}} \cdot (\mathbf{s}_{\text{GPS}} - \mathbf{s}), \quad [12]$$

in which $\mathbf{v}_{\text{LEO}}$ and $\mathbf{v}_{\text{GPS}}$ are the velocities of the LEO and GPS satellites, $\mathbf{s}_{\text{LEO}}$ and $\mathbf{s}_{\text{GPS}}$ are unit vectors in the direction of the ray path at reception by the LEO satellite and at transmission by the GPS satellite, and $\mathbf{s}$ is the unit vector pointing directly from the GPS to the LEO satellite. An assumption of spherical symmetry, which guarantees that the impact parameter for the ray is the same on the GPS and LEO sides of the ray's path, permits the determination of the bending angle $\alpha = \cos^{-1}(\mathbf{s}_{\text{LEO}} \cdot \mathbf{s}_{\text{GPS}})$ and the impact parameter $a = |\mathbf{s}_{\text{LEO}} \times \mathbf{R}_{\text{LEO}}|$, in which $\mathbf{R}_{\text{LEO}}$ is the coordinate vector of the LEO satellite with respect to the occultation's center of curvature.

In eqns [10]–[12], a few approximations have been made for the sake of presentation. First, the GPS and LEO satellites are assumed to travel through a vacuum in which the index of refraction is exactly 1. Second, relativistic effects are assumed to be small and hence are neglected. The former is assumed in all processing of GPS RO to date. The latter is not assumed: relativistic effects are fully accounted for in the processing of GPS RO.

RO: Physical Optics

Several complications arise in the application of the integral transform of eqn [11] for retrieving profiles of the index of refraction. Aside from spherical asymmetry of the index of refraction, the three most significant complications are atmospheric multipath, diffraction, and critical refraction. Atmospheric multipath occurs when more than one ray connects the occulted GPS satellite and the LEO satellite. It typically occurs in the lower troposphere, particularly in the tropics, when layers of very humid air overlie layers of much drier air. Each ray is associated with a different phase delay rate, and the observed signal experiences a beating phenomenon because of the rays' interference with each other. Diffraction arises when geometric optics – as presented in the equations in this article – loses its validity. Geometric optics is valid as long as the refracting medium has no significant structure on scales smaller than the Fresnel scale, which, for RO, is ~ 1 km (~ 1000 m). When a pronounced structure on smaller scales is encountered, such as the tropopause or the top of the planetary boundary layer, high-frequency ringing of the excess phase delay measurement and its amplitude result. Critical refraction, sometimes called 'superrefraction,' occurs in those layers for which the product $n(r) \cdot r$ is greater than $n(r) \cdot r$ at a level higher in the atmosphere. The index of refraction in a critical layer is indeterminate because no ray ever reaches its minimum in a critical layer, and the index of refraction beneath it is always underestimated. Critical layers are most often associated with maritime subsidence and marine stratocumulus.

The effects of diffraction and multipath can be corrected using a variety of techniques that have been implemented, which are commonly referred to as physical optics. These techniques are based on the same principle that makes holography possible. Whereas most remote-sensing techniques measure the amplitude or intensity of radiation from an incoherent source, RO tracks the phase of a refracted signal. The phase information can be used to reconstruct the electric field anywhere in postprocessing. Four methods of physical optics processing presented to date are canonical transform (types 1 and 2), full-spectrum inversion, and phase matching. All are conditioned on the WKBJ approximation, which requires that nonwave structures of the field occur on scales much larger

than the wavelength of the carrier signal. Both types of canonical transform consider field position and wavevector as coordinate-momentum pairs of a Hamiltonian system and transform to the impact parameter and bending angle, another Hamiltonian coordinate-momentum pair, by canonical transform. Full-spectrum inversion associates specific frequencies of a Fourier transform of the complete RO signal with individual rays, even if multiple rays are present at a given instant. Phase matching is a variation of full-spectrum inversion that accounts for noncircular orbits. All are rooted in the theory of Fourier integral operators.

There is no known method to recover the index of refraction in and below critical refraction layers accurately with RO data alone. In addition, there is no agreed upon method to retrieve atmospheric properties accurately in the presence of spherical asymmetry. In the case of the latter, the impact parameter is not the same on the GPS and LEO sides of the occultation, and $nr \sin \psi$ is not constant along the ray.

RO: Signal Amplitude

In the geometric optics approximation, the amplitude of the received signal is strongly related to the geometry of the occultation event and the bending-angle profile dictated by the atmosphere according to eqn [10]. It is derived by conserving radiant flux from transmission from the GPS satellite to reception by the LEO satellite (cf. Figure 5 for the geometry and angles). The radiant flux dF emitted from the GPS satellite inside a tube surrounding the ray is:

$$dF = I_{\text{GPS}} \cdot \sin \beta_{\text{GPS}} d\varphi d\beta_{\text{GPS}}, \quad [13]$$

in which I_{GPS} is the radiant intensity of the emitted GPS radiation. By conserving the radiant flux in the tube to the receiver, it is related to the irradiance E_{LEO} of the signal received at the LEO satellite. At the LEO satellite, the radiant flux dF is:

$$dF = E_{\text{LEO}} \cdot R_{\text{LEO}}^2 \cos \beta_{\text{LEO}} \sin \theta d\varphi d\theta, \quad [14]$$

in which E_{LEO} is the irradiance of the signal at the LEO satellite. The radiant flux dF of eqn [14] is equal to the radiant flux of eqn [13] if the signal is nowhere attenuated along its path. By geometry, $R_{\text{LEO}} \sin \beta_{\text{LEO}} = R_{\text{GPS}} \sin \beta_{\text{GPS}} = a$ and $\beta_{\text{GPS}} + \beta_{\text{LEO}} + \theta - \alpha = \pi$. Differentiation and substitution give the equation for irradiance at the LEO satellite:

$$E_{\text{LEO}} = I_{\text{GPS}} \frac{a}{R_{\text{GPS}} R_{\text{LEO}}} \frac{1}{D_{\text{GPS}} + D_{\text{LEO}}} \frac{1}{|Q| \sin \theta}, \quad [15]$$

in which $D_{\text{GPS}} = (R_{\text{GPS}}^2 - a^2)^{1/2}$ and $D_{\text{LEO}} = (R_{\text{LEO}}^2 - a^2)^{1/2}$ are the distances from the GPS satellite to the Earth's limb and from the LEO satellite to the Earth's limb, respectively. The defocusing factor Q is given in terms of the reduced limb distance $D \equiv (1/D_{\text{GPS}} + 1/D_{\text{LEO}})^{-1}$:

$$Q \equiv 1 - D d\alpha/da. \quad [16]$$

The differential bending angle, $d\alpha/da$, is almost everywhere negative because ray bending increases with depth in the atmosphere, and thus the defocusing factor is almost always greater than 1. For GPS RO, the term that varies the most over the course of a single RO event is the defocusing factor Q . The differential bending is illustrated as the spreading of rays in

Figure 3. The deeper the ray penetrates into the atmosphere, the weaker the signal becomes. The amplitude of the signal is proportional to the square root of the irradiance at the LEO satellite; thus, it is proportional to $Q^{-1/2}$.

The expressions given in this article for the signal amplitude do not strictly apply in the presence of diffraction and atmospheric multipath. In cases of atmospheric multipath, multiple signals connect the occulted GPS satellite and the LEO satellite, and those signals interfere with each other, resulting in a beating pattern in amplitude and phase. When the multiple rays overlap each other, a strong diffraction pattern occurs. In situations like these, the processing of the data is best handled by physical optics, the effect of which is to set the reduced limb distance $D \rightarrow 0$ and hence the defocusing factor $Q \rightarrow 1$. The postprocessed amplitude of the signal does not vary with impact parameter.

The amplitude of the received RO signal is related to the apparent vertical motion of the signal in the atmosphere. From the viewpoint of the LEO satellite, the image of the GPS satellite appears to descend into the Earth's atmosphere as a setting occultation progresses and is described as the time rate of change of the impact parameter, $\dot{a}$. As the ray connecting the GPS and LEO satellites experiences stronger bending, the ray's apparent descent slows. If no ray bending were present, the apparent vertical motion of the ray would be $\dot{a} = -D\dot{\theta}$, in which $\dot{\theta}$ is the rate of angular separation (with respect to the center of curvature) of the GPS and LEO satellites in orbit. In the presence of differential bending, however, conservation of radiant energy leads to the actual apparent vertical motion of the image of the transmitter to be as follows:

$$\dot{a} = -D\dot{\theta}/Q. \quad [17]$$

The reduced limb distance for most RO missions is $D \approx 3000$ km (3×10^6 m). In the absence of defocusing, the maximum descent rate for RO events is 3000 m s^{-1} .

RO: Calibration

While GPS RO is sometimes referred to as 'self-calibrating,' there still is a procedure for assuring the accuracy and precision of the timing measurements that constitute a RO observation. If geometric optics processing is applied, the determination of $\alpha(a)$ depends on highly precise and accurate determination of the vector velocities of the GPS and LEO satellites. If physical optics processing is applied, the integrals to be computed rely upon highly precise determination of the satellites' positions as a function of time. In either case, precise knowledge of the satellites' orbits is obtained by precise orbit determination, a computation in which data collected over ~ 24 h periods by a worldwide network of GPS receivers on the ground are used to determine the orbits of the GPS satellites. Also in precise orbit determination, data collected by the LEO satellite from nonocculted GPS satellites are used to determine the orbit of the LEO satellite.

The measurement of the excess phase rate must be calibrated either by using a suitably stable atomic clock onboard the LEO satellite or by referencing a clock on a nonocculted GPS satellite during each RO event. The latter is accomplished

by measuring the phase of the signal of the nonocculted GPS satellite at the same rate as the excess phase delay of the occulted GPS satellite. This procedure for calibration is called *single differencing*. If the clocks onboard the GPS satellites are not deemed sufficiently stable for calibration of a RO event, the signals of both the occulted GPS and the referenced GPS satellites are observed by a GPS receiver on the ground that is tied to an ultrastable atomic clock. This method of calibration is called *double differencing* (cf. Figure 6). In most applications of GPS RO retrieval, single differencing has been found to be sufficiently precise and accurate, thus eliminating the complication of finding a ground station that tracks both the occulted GPS satellite's signal and the reference GPS satellite's signal during the RO event. In *zero differencing*, the clock onboard the LEO satellite is sufficiently stable by itself that a reference nonocculted GPS satellite need not be observed by the LEO satellite.

Before the ray connecting the occulted GPS satellite to the LEO satellite enters the Earth's atmosphere, it must first propagate through the ionosphere. The Earth's ionosphere is highly dispersive, introducing an inverse square dependence on the frequency of the carrier signal to the expression for the index of refraction. The refraction of rays in the ionosphere lends RO well suited to profiling the density of electrons in the ionosphere, but in the neutral atmosphere it contributes a source of error in the retrieval of temperature, pressure, and humidity. As part of the process of calibrating a RO event, the receiver onboard the LEO satellite tracks two signals of GPS rather than just one. The primary, strongest signal is L1 C/A, with a carrier frequency of 154×10.23 MHz; the secondary signal is the encrypted L2 signal, with a carrier frequency of 120×10.23 MHz. Dispersion in the ionosphere causes each of

the two signals to take a different path through the atmosphere. The standard ionospheric calibration is a linear combination of the bending-angle profiles for L1 and L2 to yield an ionosphere-corrected bending-angle profile $\alpha_{LC}(a)$:

$$\alpha_{LC}(a) = \frac{154^2 \cdot \alpha_{L1}(a) - 120^2 \cdot \alpha_{L2}(a)}{154^2 - 120^2}. \quad [18]$$

While this ionospheric calibration is standard in retrieval algorithms, it is nevertheless difficult to apply because the L2 signal is modulated by a secure encryption. Modern-day RO receivers track L2 using software that correlates it either against an L1 signal that is subject to the same encryption as L2 or by semicodeless tracking. As a result, the L2 signal is much weaker than the L1 C/A signal and is typically tracked to only ~ 10 km altitude. A modernized GPS L2, called L2C, that is modulated by a publicly available code is becoming operational and should alleviate the difficulties of tracking L2 that have been experienced to date.

RO: Retrieval of Atmospheric State Variables

There is no single way to convert profiles of the index of refraction to profiles of temperature, pressure, and humidity. The hydrostatic equation constrains the solution, but there is still one too many unknown variables, a problem that is known as the *wet-dry ambiguity* in GPS RO. Many techniques, almost all of which are a type of variational data assimilation, have been incorporated to resolve the wet-dry ambiguity. Qualitatively, GPS RO can be considered a humidity sounder in the lower troposphere and a temperature sounder above. Prior uncertainty in humidity propagates to a greater (lesser) uncertainty in the index of refraction than does prior uncertainty in temperature in the lower troposphere (upper troposphere and stratosphere).

The most basic retrieval schemes do not attempt a resolution of the wet-dry ambiguity. Instead, they retrieve pressure and temperature as if there were no water vapor present and call those quantities *dry pressure* and *dry temperature*. Dry pressure is always greater than kinetic pressure, and the difference increases the moister the environment becomes. Dry temperature, in contrast, is always less than the kinetic temperature. In climate research, efforts are being made to treat the index of refraction, dry pressure and dry temperature, in addition to bending angle as state variables that can be directly compared to model output of the same quantities.

The next step in sophistication in the retrieval of atmospheric state variables is to assume either temperature or humidity in order to retrieve the other. In this approach, in the upper troposphere and stratosphere, a profile of the water vapor partial pressure is assumed from an analysis of numerical weather prediction (NWP) or climate reanalysis in order to retrieve temperature. Likewise, a profile of temperature is assumed in order to retrieve specific humidity in the lower troposphere. The level of transition is ad hoc. One retrieval system has set the level of transition as the highest tropospheric level at which the temperature is 250 K.

With increasing sophistication, humidity and temperature are retrieved by one-dimensional variational assimilation (1DVAR). Variational assimilation, as an application of Bayes's

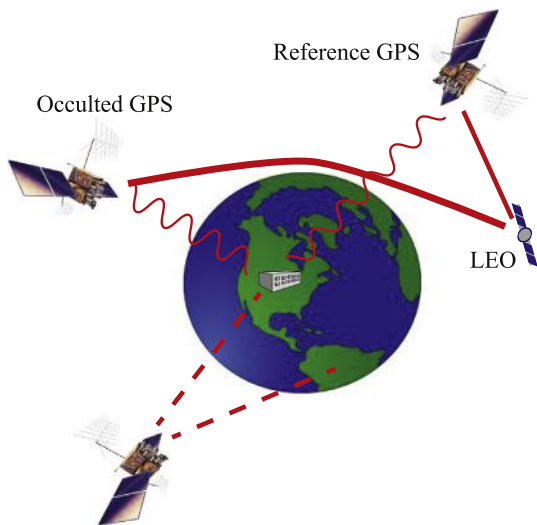


Figure 6 Calibration by single differencing requires observation of the 'reference GPS' by the LEO satellite. Calibration by double differencing requires observation of both the 'occulted GPS' and the reference GPS by a ground station such as one tied to the time standard at the US National Institutes of Standards and Technology (NIST) facility in Boulder, Colorado. All GPS satellites are used for precise orbit determination of the LEO satellite.

theorem, blends prior knowledge of temperature and humidity, usually taken from a numerical weather prediction or analysis, with data in a way that optimizes the posterior knowledge of both temperature and humidity. 1DVAR objectively quantifies the more qualitative approach (described in this section) that RO is a humidity sounder in the lower troposphere and a temperature sounder everywhere above. The forms of variational assimilation used in numerical weather prediction (3DVAR, 4DVAR, and the Ensemble Kalman Filter) follow the same principle as 1DVAR. Figure 7 shows a 1DVAR retrieval of one RO profile obtained by the Constellation Observing System for Meteorology, Ionosphere and Climate (COSMIC) project.

At present, the following institutions have published retrievals of RO data:

- The COSMIC Project Office, University Corporation for Atmospheric Research, Boulder, CO, United States (www.cosmic.ucar.edu).
- The NASA Jet Propulsion Laboratory, Pasadena, CA, United States (genesis.jpl.nasa.gov).
- The European Organization for the Exploitation of Meteorological Satellites (EUMETSAT) Radio Occultation Meteorology Spacecraft Application Facility, Copenhagen, Denmark (romsaf.org).
- The Alfred Wegener Centre of the University of Graz, Austria (globclim.org).
- The Geophysical Research Centre, Potsdam, Germany (gfs-potsdam.de).

RO: Spatial Resolution

The vertical resolution of a retrieval of GPS RO depends on the processing method. If geometric optics was used in the estimation of bending angle as a function of impact parameter, then the vertical resolution is determined by the vertical dimension of the

first Fresnel zone. In the absence of bending, the size of the Fresnel zone is $(2\lambda D)^{1/2}$. In the presence of bending, the Fresnel zone is reduced in the vertical direction, thus increasing the vertical resolution. The amount by which it is reduced in the vertical direction can be determined by the amplitude of the signal, which is directly proportional to the area of the Fresnel zone. Since the amplitude of the signal is proportional to the inverse square root of the defocusing factor, the vertical resolution of GPS RO is $(2\lambda D/Q)^{1/2}$. For L1, the wavelength of the carrier signal is 0.19 m, and hence the vertical resolution is $1070 \text{ m} \cdot Q^{-1/2}$. The transverse horizontal resolution is also the Fresnel size, and it is unaffected by differential bending of the ray.

If physical optics is applied in the reconstruction of $\alpha(a)$, the effective reduced limb distance tends to zero and hence so does the Fresnel zone. In fact, Fresnel diffraction is not the limiting factor in the vertical resolution, but Fraunhofer diffraction is. In Fraunhofer diffraction, the angular resolution of the antenna is the wavelength of the signal divided by the horizontal dimension of the antenna. In GPS RO, the antenna is effectively the path traversed by the LEO spacecraft perpendicular to the direction of the ray during which data are recorded. If data are recorded for 60 s for a LEO spacecraft moving 7 km s^{-1} (and $\beta_{\text{LEO}} \approx 65^\circ$), the Fraunhofer-limited resolution is 3.2 m. In practice, though, the vertical resolution is dictated by the limited amount of data collected and the smoothing algorithms applied to reduce noise in the retrieval. For most applications of physical optics processing, the vertical resolution is effectively 100–200 m.

The horizontal resolution of GPS RO can only be arbitrarily defined since it is a limb-sounding technique. If one considers the horizontal path through which an occultation ray experiences 68% of its bending, the horizontal resolution is $2(HR_e)^{1/2} \approx 390 \text{ km}$ ($3.9 \times 10^5 \text{ m}$), in which the scale height H is $\sim 6000 \text{ m}$ and the radius of the Earth R_e is 6378 km ($6.378 \times 10^6 \text{ m}$). Using the scale height is appropriate in the estimation of the horizontal resolution of GPS RO because most of the contribution to the integral of eqn [11] and the

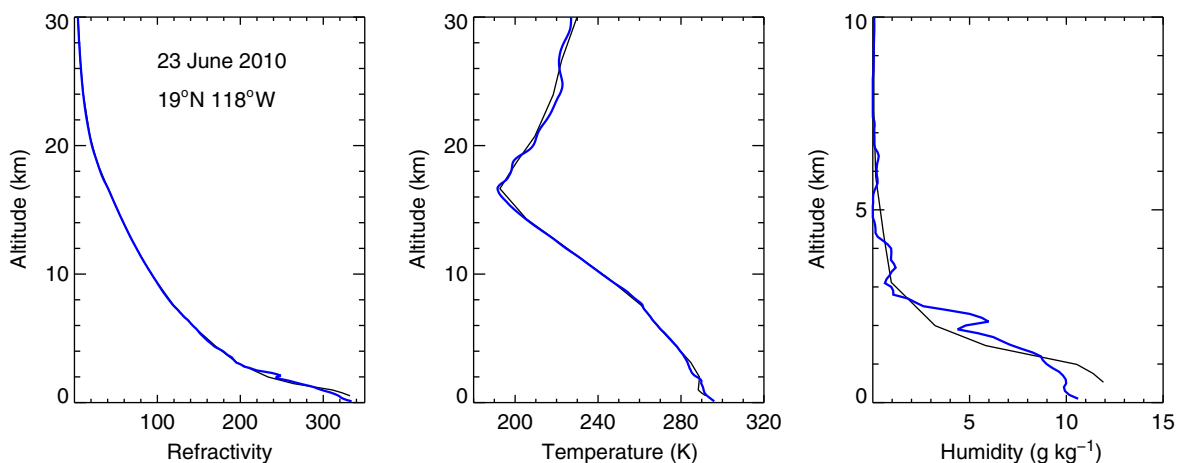


Figure 7 Retrieval of refractivity, temperature, and specific humidity from a radio occultation obtained by the COSMIC project, 23 June 2010, at 19°N , 118°W . The altitude is the geometric distance from a best-fit ellipsoid to the Earth's surface. The retrieval of temperature and specific humidity was obtained by 1DVAR at the COSMIC Project Office. The background profile is taken from operational analyses of the European Centre for Medium-Range Weather Forecasting (ECMWF) (black). The occultation data is blue.

hydrostatic integral are contributed over a scale height in the vertical. If smaller structures in the vertical are of greater interest, then their vertical dimensions are more suitable than the scale height in the estimation of the horizontal resolution of GPS RO.

RO: Sources of Uncertainty

The sources of uncertainty in GPS RO have been documented extensively. They come in two distinct families: one pertaining to the upper stratosphere and the other to the lower troposphere. In the case of the former, the vertical structure of the error has an envelope that is inversely proportional to pressure, and thus error decreases exponentially with depth with an e -folding scale given by the pressure scale height. Chief among those errors are residual uncertainty in ionospheric calibration, and uncertainty in initiating the transform of eqn [11] and in initiating the hydrostatic integral. All told, the uncertainty at 50 km altitude is 3% in refractivity or 300 m in geopotential height after smoothing over 3000 m. Receiver noise contributes a random error of 2% over the same smoothing interval. In the case of the latter, the dominant source of error is spherical asymmetry. Water vapor is highly heterogeneous in the Earth's atmosphere, leading to spherical asymmetry in the index of refraction. The uncertainty is approximately 1% in refractivity and 60 m in geopotential height in the lowest 5 km of the atmosphere.

Almost every source of error is anticipated to be random in nature; thus, averaging many RO soundings together may yield a climatology that is unbiased. In the case of ionospheric residual, a consequence of incomplete ionospheric calibration, a systematic component to the error may exist depending on the phase of the 11-year solar cycle. Comparison of climatological averages constructed by several independent retrieval centers shows that the structural uncertainties induced by initialization of the transform of eqn [11] and the hydrostatic integral dominate.

In summary, GPS RO performs best in the 8–30 km height region, where the refractivity N can be retrieved with a precision of 0.2% and an accuracy that is an order of magnitude better.

RO: Scientific Applications

GPS RO, because of its several unusual properties, has been used for many applications. Foremost among these is numerical weather prediction because of the density and global distribution of RO data, the data's sensitivity to geopotential height, their insensitivity to clouds, and their high precision and absolute calibration. Many of RO's properties make it comparable in information content to radiosondes but with the advantage of dense coverage of the Earth's Southern Hemisphere. In numerical weather prediction, GPS RO has already greatly improved analysis and prediction, especially in the Southern Hemisphere. Improvement in the prediction of precipitation has proven elusive, but that may well be a consequence of inadequate spatial resolution of the forecast model as much as it is a consequence of the long limb-path of the GPS RO signal. GPS RO has also been used in climate studies, revealing a marginally significant climate trend over the record

of GPS RO data. Finally, GPS RO has been used in studies of atmospheric dynamics, directly revealing wave generation by convection, tropopause variability, the atmospheric tides, the atmospheric signal of El Niño and the Southern Oscillation, and stratospheric internal gravity wave activity.

GPS RO Missions

GPS RO commenced in 1995 and continues to the present. Because the equipment for GPS RO sounding is inexpensive, most of the missions have been by opportunity wherein the GPS receivers are secondary instruments on satellites whose prime instruments have been technologically more complex and hence substantially more expensive. The first proof-of-concept GPS RO mission was GPS/Met, a mission of opportunity. Other missions of opportunity since then have been Ørsted, Sunsat, SAC-C, CHAMP, GRACE, TerraSAR-X, TanDEM-X, and Oceansat-2. The operational GPS RO missions have been the six-satellite COSMIC constellation and Metop-A. More missions are planned.

The various missions have achieved varying degrees of performance according to their design and implementation. Individual occultation events can be categorized as 'setting' or 'rising,' depending on whether the occulted GPS satellite appears to be setting into the Earth's limb or rising into it. The former happens in the antiveloc direction of the LEO satellite, and the latter in the forward-velocity direction. Many missions observe only the setting events because they configured only antiveloc-oriented occultation antennas. Also, missions have implemented different algorithms for tracking GPS signals during occultation events, with some performing better than others especially during instances of multipath and diffraction. The need to remove ionospheric influence dictates tracking both the L1 and L2 GPS signals, and L2 is notoriously difficult for civilian GPS receivers to track for the reasons given in this article. Poorer tracking of L2 implies degraded removal of ionospheric influence. The mission that best tracked L2 was the first, GPS/Met, because the US Air Force disabled the secure encryption on L2 for four month-long periods in 1995–97 for the sake of GPS/Met RO. Beginning with Metop-A and Oceansat-2, the antennas used to track RO signals are phased arrays in order to greatly augment the signal-to-noise ratio of RO signals. This is expected to enhance the accuracy and precision of retrieval in the lowest few kilometers of the atmosphere.

See also: **Clouds and Fog:** Stratus and Stratocumulus. **Data Assimilation and Predictability:** Data Assimilation. **Dynamical Meteorology:** Hamiltonian Dynamics; Kelvin Waves. **Global Change:** Upper Atmospheric Change. **Gravity Waves:** Overview. **Numerical Models:** General Circulation Models.

Further Reading

- Anthes, R., 2011. Exploring earth's atmosphere with radio occultation: contributions to weather, climate and space weather. *Atmospheric Measurement Techniques* 4, 135–212.
- Bevis, M., Businger, S., Herring, T., Rocken, C., Anthes, R., Ware, R., 1992. GPS meteorology: remote sensing of atmospheric water vapor using the Global Positioning System. *Journal of Geophysical Research* 97, 15787–15801.

- Fjeldbo, G., Kliore, A., Eshleman, V., 1971. The neutral atmosphere of Venus as studied with the Mariner V radio occultation experiments. *The Astronomical Journal* 76, 123–140.
- Gorbunov, M., Lauritsen, K., 2004. Analysis of wave fields by Fourier integral operators and their application for radio occultations. *Radio Science* 39. <http://dx.doi.org/10.1029/2003RS002971>.
- Hajj, G., Kursinski, E., Romans, L., Bertiger, W., Leroy, S., 2002. A technical description of atmospheric sounding by GPS occultation. *Journal of Atmospheric and Solar-Terrestrial Physics* 64, 451–469.
- Healy, S., Eyre, J., 2000. Retrieving temperature, water vapour and surface pressure information from refractive-index profiles derived by radio occultation: a simulation study. *Quarterly Journal of the Royal Meteorological Society* 126, 1661–1683.
- Kursinski, E., Hajj, G., Schofield, J., Linfield, R., Hardy, K., 1997. Observing Earth's atmosphere with radio occultation measurements using the Global Positioning System. *Journal of Geophysical Research* 102, 23429–23465.
- Nilsson, T., Elgered, G., 2008. Long-term trends in the atmospheric water vapor content estimated from ground-based GPS data. *Journal of Geophysical Research* 113. <http://dx.doi.org/10.1029/2008JD010110>.
- Xie, F., Wu, D., Ao, C., Kursinski, E., Mannucci, A., Syndergaard, S., 2010. Super-refraction effects on GPS radio occultation refractivity in marine boundary layers. *Geophysical Research Letters* 37. <http://dx.doi.org/10.1029/2010GL043299>.

Measuring Ozone from Space – TOMS and SBUV

RD McPeters, NASA Goddard Space Flight Center, Greenbelt, MD, USA

RS Stolarski, Johns Hopkins University, Baltimore, MD, USA

© Published by Elsevier Ltd.

Synopsis

Space-based measurements of ozone on a global basis began with the Backscatter Ultraviolet instrument flown on Nimbus 4 in 1970. A series of similar instruments flown by National Aeronautics and Space Administration and the National Oceanic and Atmospheric Administration have documented the development of an ozone hole over the Antarctic each October, and a drop in global ozone beginning in the 1980s. This important program for the monitoring of global ozone levels is expected to show the slow recovery of ozone over the coming decades as a result of international agreements to regulate the release of ozone-depleting chemicals.

Introduction and History

In the beginning, only a few scientists were interested in ozone. Popular interest came later, in the late 1970s, when the possible destruction of the ozone layer by chlorofluorocarbons (CFCs) was realized. The idea that ozone could be measured quantitatively from a satellite was first put forward in 1957 by Singer and Wentworth. In 1967, Dave and Mateer published a theory for the derivation of the total column amount of ozone from a satellite backscatter instrument. This theory was used for the interpretation of the data from the first instrument to measure ozone from space, the Backscatter Ultraviolet instrument, launched on the Nimbus 4 satellite in 1970. It made nearly global measurements for 2 years and then operated more sporadically for an additional 5 years. When Farman's paper on low ozone in the Antarctic was published in 1985, ozone maps from the Total Ozone Mapping Spectrometer (TOMS) revealed that this was a continent-wide phenomenon and not local. This is the power of satellite measurements of ozone.

The modern data record of global ozone measured from space starts with the launch of the Nimbus 7 satellite in 1978. The satellite carried two ozone-measuring instruments, TOMS and the Solar Backscatter Ultraviolet (SBUV) instrument. TOMS was a satellite-borne instrument that measured the total column amount of ozone in the atmosphere by measuring the ultraviolet (UV) sunlight scattered from the atmosphere, while SBUV was a companion instrument that measured how that ozone was distributed with altitude. The TOMS instrument measured the backscattered UV radiation at six wavelengths from 312 to 380 nm. By comparing light at a wavelength strongly absorbed by ozone to light at a wavelength weakly absorbed by ozone, an accurate measurement of ozone can be made. TOMS was a single monochromator with a scanning mirror that allowed the instrument to make measurements at 35 scan angles from left to right across the ground track of the satellite. TOMS was thus able to measure over the entire sunlit portion of the globe each day. The SBUV instrument was a double monochromator designed to measure backscattered radiation at 12 wavelengths from 255 to 340 nm. Because the shorter wavelengths penetrate only to the upper stratosphere, these wavelengths can be used to deduce both the upper-stratospheric concentration profile of ozone and the total column amount of ozone, but only along the nadir track of the satellite.

The TOMS instrument on Nimbus 7 made measurements for more than 14 years. The instrument finally failed in May 1993. A second TOMS instrument was launched on the Russian Meteor 3 satellite in 1991. This cooperative project was the first time an American instrument was flown on a Soviet spacecraft. This instrument made total ozone measurements until the end of 1994. A third TOMS instrument was launched on the Japanese ADEOS satellite in 1996, but the satellite power array failed after only 7 months of operation. A fourth TOMS instrument was launched on the Earth Probe satellite, also in 1996. Designated EP-TOMS, it operated through December 2005. A final TOMS instrument, QuikTOMS, was lost in 2001 when it failed to achieve orbit.

The successor instrument to the TOMS series of instruments is the Ozone Monitoring Instrument (OMI) launched on the Aura spacecraft in July 2005. OMI, which was built by the Dutch for flight on Aura, is an imaging spectrometer that measures over the range of 270–500 nm. The Ozone Mapping Profiler Suite (OMPS) is also an imaging spectrometer, designed to continue the US program of ozone monitoring begun by TOMS and SBUV. OMPS was launched on the Suomi-NPP satellite in 2011.

A series of SBUV/2 instruments has been flown by the National Oceanic and Atmospheric Administration (NOAA) to monitor long-term changes in ozone. Instruments were flown on NOAA 9 (1985–98), NOAA 11 (1989–2000), NOAA 14 (1995–2006), NOAA 16 (2001–present), NOAA 17 (2002–13), NOAA 18 (2005–12), and NOAA 19 (2009–present).

The Theory of Ozone Measurement

TOMS and SBUV can measure ozone from space because ozone absorbs very strongly in the UV. By comparing a measurement at a wavelength strongly absorbed by ozone to one weakly absorbed by ozone, an accurate estimate of the amount of ozone in the atmosphere can be made from space. The instruments measure sunlight scattered by the atmosphere and reflected from clouds and the ground.

Rayleigh Scattering

Light from the Sun penetrates into the atmosphere, with most of the visible light reaching the ground. Light is scattered by the

molecules that make up the atmosphere in a process called Rayleigh scattering, named after Lord Rayleigh, who first described it in the late nineteenth century. The probability of Rayleigh scattering depends inversely on the fourth power of the wavelength (λ^{-4}). Thus, an UV photon of 300 nm wavelength is 16 times more likely to be scattered than a visible photon of 600 nm wavelength. This is why the sky is blue. When we look at the sky away from the Sun, blue light is much more likely to be scattered toward us than is red light.

Surface Reflection and Clouds

Radiation that does reach the ground can be absorbed or reflected by the surface. The probability of reflection depends on the nature of the surface. In the UV range, the Earth is a very poor reflector. The UV reflectivity of the ocean in the 300–350 nm region of the spectrum is only about 4%. Most land surfaces have similarly low reflectivities, no more than 5% except in desert areas. Only areas covered by ice and snow have very high reflectivities, reaching 90% in the Antarctic.

When clouds are present, radiation reaching them is reflected back to space with high efficiency. Cloud reflectivities can reach 80–90% for thick clouds. Solar radiation that is reflected by clouds does not pass through the part of the atmosphere below the cloud and has no opportunity to be absorbed by the ozone below the clouds. The TOMS measurement is thus a measurement of the ozone above the cloud layer. Fortunately, this is a small effect since 90% of the ozone in the atmosphere is in the stratosphere. Only part of the 10% of the ozone column in the troposphere will be masked by clouds. This amount can be estimated from climatology, so that the measurement can be transformed into a fairly accurate estimate of the total column of ozone in the atmosphere.

Aerosols

Aerosols (dust particles in the atmosphere) also scatter radiation, further adding to the atmosphere's overall reflectivity. Their scattering does not follow the λ^{-4} dependence of Rayleigh scattering but is close to a λ^{-1} dependence. This means that when dust is in the atmosphere, the sky appears more nearly white. Aerosols do not affect our ability to measure ozone. However, the multiple reflectivity wavelengths can be used to deduce some information about the properties of aerosols. Measurement of the deviations from the expected result for a Rayleigh-scattering atmosphere can be used to determine an aerosol index and used to track dust and smoke plumes (see the section [Some Results from TOMS Measurements](#)).

Ozone Absorption

Sunlight can be absorbed in the atmosphere by a variety of molecules. The principal absorber of UV light in the Earth's atmosphere is ozone. Absorption in the UV by ozone is so strong that a few parts per million of ozone remove all of the sunlight at wavelengths shorter than about 300 nm before they can reach the ground. We are thus provided with a shield from the high-energy radiation that could break important DNA bonds in living cells. The absorption of UV by ozone is the

property that we generally use to measure the amount of ozone in the atmosphere. From the ground, we can look upward and measure how much radiation reaches us at a wavelength that is absorbed by ozone. We can compare this to the radiation received at a nearby wavelength that is not absorbed by ozone to determine the amount of ozone that must be between us and the Sun. Such measurements have been used to measure ozone since the 1920s. From satellites, we can look down and measure the radiation that is scattered back out of the atmosphere through Rayleigh scattering and again compare the amount at an absorbed wavelength with an unabsorbed (or less absorbed) wavelength.

Description of the TOMS Retrieval Algorithm

TOMS measures UV light scattered from the atmosphere and the Earth and clouds. An algorithm is needed to infer ozone from these measurements. The instrument looks downward at the Earth but also uses a diffuser plate to look at direct sunlight before it enters the atmosphere. The basic measured quantity is the ratio of the backscattered radiance to the direct solar irradiance. This is usually expressed as the N -value, or logarithm of the ratio:

$$N = -100 \log \left(\frac{I_0}{F} \right) \quad [1]$$

where F is the solar irradiance at the particular wavelength and I_0 is the backscattered radiance. TOMS measures the light scattered from the Earth–atmosphere system (I_0) at a range of wavelengths, and measures F , the direct sunlight, by periodically deploying a reflective diffuser plate in front of the instrument.

Using the ratio of backscattered radiation to direct solar flux cancels some of the main instrumental errors; that is, the instrument throughput is the same for each measurement and cancels when the ratio is taken. The reflectivity of the diffuser plate affects the solar irradiance measurement, but not the backscattered radiance measurement. If a pair of wavelengths is used in the analysis, then the diffuser reflectivity can be canceled out in the ratio if that reflectivity is the same for both wavelengths. Thus, we form the pair N -value as in eqn [2].

$$\begin{aligned} N_p &= N(\lambda_1) - N(\lambda_2) \\ &= \log \left(\frac{I_{01}}{F_1} \right) - \log \left(\frac{I_{02}}{F_2} \right) \end{aligned} \quad [2]$$

These N -values reflect the effects of scattering, reflection, and absorption. [Figure 1](#) illustrates the dependence of the N -value on wavelength, clearly showing that an ozone signal can be derived from the data. The actual algorithm used for the TOMS retrieval uses a radiative transfer code based on the early work of Dave. Forward calculations are carried out for a matrix of parameters, including total ozone. These then form a lookup table that is interpolated to derive total ozone.

Description of the SBUV Retrieval Algorithm

SBUV is similar to TOMS but is used to measure the altitude distribution of ozone. Like TOMS, its measurements are used as

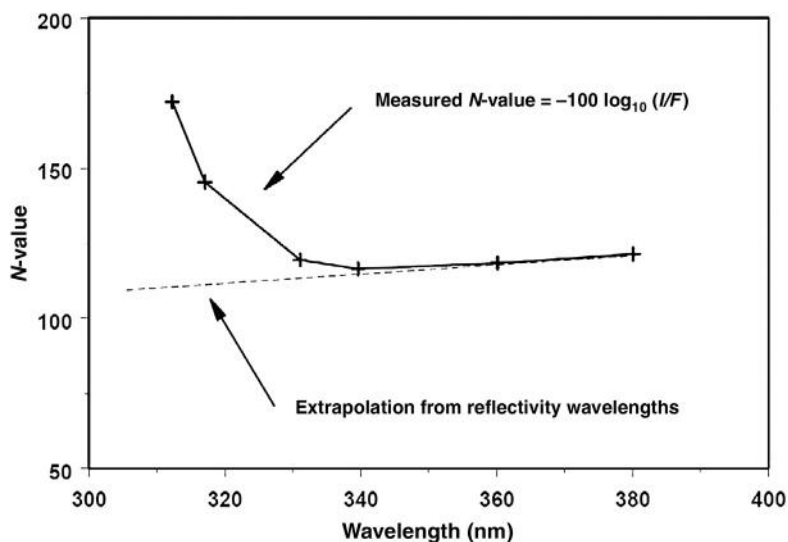


Figure 1 Illustration of the dependence of N -value on wavelength. The N -values for all of the TOMS measurements for 1 day (1 January 1985) within 1° of latitude of 35° N were averaged to make the plot. The linear straight line is fitted to the three longest wavelengths to illustrate an extrapolation to shorter wavelengths. The actual TOMS algorithm uses a full radiative transfer code to determine this extrapolation. The difference between the short-wavelength N -values and the extrapolation represents the absorption by ozone.

N -values, but covering wavelengths further into the UV range – 12 wavelengths from 250 to 400 nm. The key is that the very short wavelengths do not penetrate to the ground. At 255 nm, light penetrates to only about 50 km altitude; at 290 nm, light penetrates to 40 km; and 302 nm penetrates to 30 km. Thus, the SBUV wavelength scan is equivalent to an altitude scan of ozone. In the actual profile retrieval, the Rodgers optimal estimation technique varies a first-guess ozone profile until the set of measured N -values is matched.

TOMS and SBUV

Instruments

The TOMS instruments are single, fixed monochromators with exit slits at six near-UV wavelengths. The slit functions are triangular with a nominal 1 nm bandwidth. The order of individual measurements is determined by a chopper wheel. As it rotates, openings at different distances from the center of the wheel pass over the exit slits, allowing measurements at the different wavelengths. The order was not one of monotonically increasing or decreasing wavelength; instead, the wavelengths were interleaved to minimize the effect of spacecraft movement on the ozone retrieval. The advantage of the more modern OMI and OMPS instruments is that they measure the complete spectrum at one time rather than sequentially measuring six wavelengths.

A ground aluminum diffuser plate is deployed to reflect sunlight into the instrument for measurement of the solar irradiance. This diffuser plate was shared by TOMS and SBUV experiment on the Nimbus 7 satellite. It was normally deployed once a week for TOMS solar irradiance measurements, in addition to the SBUV deployments. On the Earth Probe TOMS, a triple diffuser was used to eliminate the problem of the diffuser darkening as it was exposed to UV light in space. One diffuser was used daily, another was used weekly,

and the third was used only a couple of times a year to serve as the clean reference.

SBUV is a double monochromator. Because light at a short wavelength like 270 nm is 1000 times weaker than light at a longer wavelength like 320 nm, scattered light within the instrument itself is a problem. By dispersing the light twice, this scattered light can be reduced. Over 32 s, SBUV scans from 250 to 400 nm in 12 steps. Because more light is needed to measure the weak signal at short wavelengths, SBUV only measures along the orbit track, rather than scanning from side to side as TOMS does to cover the entire Earth.

Orbit

The Nimbus 7, ADEOS, Earth Probe, Aura, and NPP satellites were all in Sun-synchronous polar orbits. The nearly circular orbit is oriented perpendicular to the plane of the Earth's orbit around the Sun such that the satellite comes over the South Pole of the Earth toward the Equator, crosses the Equator near local noon, and then passes over the North Pole onto the nightside of the Earth. The satellite crosses the Equator again on the nightside at near midnight local time. By the time the satellite comes back onto the dayside, the Earth has rotated for approximately 90 min, and the satellite passes over a point at the Equator that is 27° of latitude to the west of the previous orbit but, again, at local noon. In this way, the satellite orbits 15 times per day, fixed relative to the Sun, and the Earth rotates underneath so that the satellite sees the whole of the surface of the Earth within a 24-h period.

This is a qualitative description of the orbits. Actually, for the purpose of orbit stability, the satellite does not pass exactly over the pole. A Sun-synchronous orbit requires an orbital inclination of approximately 98° , which gives a maximum orbit latitude of about 80° . A slightly inclined orbit precesses just enough to stay exactly in the noon plane over the course of the year. From this orbit, TOMS can see the pole itself by

scanning to the far right or left. The Meteor 3 spacecraft was in a polar orbit but was not Sun synchronous. Its Equator-crossing time drifted from near noon to near sunset and back to near noon in a 220-day cycle. The NOAA series of satellites carrying SBUV/2 instruments were also in Sun-synchronous orbits, but ones that drifted slowly away from noon over a few years.

Geometry and Timing

The instrument field of view for TOMS is $3 \times 3^\circ$. At an altitude of 950 km for Nimbus 7, this projects to a nadir spot size on the surface of $50 \text{ km} \times 50 \text{ km}$. Earth Probe was launched initially into a 500 km orbit, which resulted in a nadir spot size of 26 km. In December 1997, it was boosted to an altitude of 740 km, increasing the nadir spot size to 40 km.

For each of the TOMS instruments, a mirror scans perpendicular to the orbital plane in 35 steps of 3° . The scan angles range from 51° on the right side of the spacecraft nadir to 51° on the left (relative to the direction of flight). At the end of the scan, the mirror returns to the first position and begins another scan. For Nimbus 7, the cross-track scans from consecutive orbits overlapped, creating a completely filled global map of the sunlit part of the Earth each day. The lower altitude of the Earth Probe TOMS results in small areas between orbits near the Equator where no measurements are

made (as in Figure 2). The location of these gaps shifts from day to day so that no place fails to be measured over the span of a few days.

During the cross-track scan, each of the 35 measurement locations is observed for 200 ms. The total duration time for a single scan is 7.8 s, during which time the satellite travels approximately 40 km. One orbit consists of nearly 400 cross-track scans or 13 000 measurements. Fifteen orbits result in about 190 000 measurements of total ozone every day. SBUV is similar to TOMS but does not scan. SBUV looks only at a 200 km square field of view directly beneath the spacecraft.

The imaging instruments, OMI and OMPS, project an entire strip of the Earth onto a charge-coupled device (CCD) detector. As the satellite moves, this strip maps out the Earth. This is known as a push-broom system. The image on the CCD consists of the strip of the Earth below the satellite (imaged in the x direction) and the wavelength (imaged in the y direction).

Some Results from TOMS Measurements

Ozone Maps

The original purpose for building TOMS was its capability to map global ozone on a daily basis to help understand its relationship to changes in the meteorology of the atmosphere.

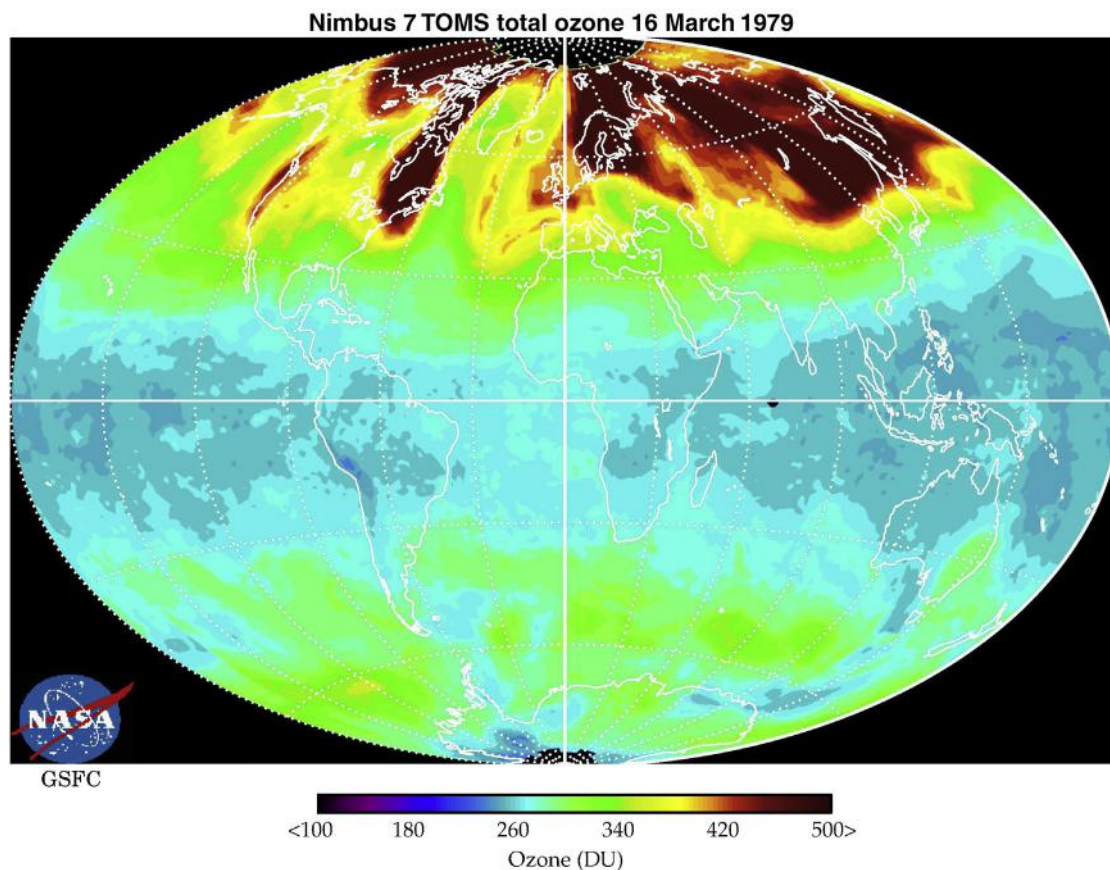


Figure 2 Global ozone mapped by TOMS on 16 March 1979, soon after launch. Note regions of very high ozone in the Arctic and low ozone near the Equator.

The problem of the relationship of total ozone to meteorology goes back to Dobson in the 1920s. Dobson had six of his spectrophotometers built and distributed throughout Europe to examine this problem. He found that when a high-pressure system was present, ozone was low; and when a low-pressure system was present, ozone was high. TOMS was designed to map the entire sunlit portion of the globe in a single day (Figure 2).

When the discovery of the ozone hole was announced in 1985, TOMS was immediately used to map the extent of the ozone-depleted region (Figure 3). Using TOMS, the daily progress of the hole could be followed. These maps demonstrated that the ozone-depleted region rotated around the pole, was distorted by the meteorology, and was finally broken up by a series of wave events that eroded the polar vortex. TOMS can produce similar maps of ozone over the Arctic polar region (not shown). The maps clearly demonstrate the day-to-day and year-to-year variability of ozone over the Arctic. As global warming changes the dynamics of the atmosphere, such maps will show the effect on the global ozone distribution.

Ozone Trends

While the Nimbus 7 TOMS instrument was originally designed to map ozone on a daily basis to study day-to-day variability in

total ozone, by the 1990s TOMS data were also being used as part of a satellite-based measurement system for detecting long-term trends in stratospheric ozone. A number of features in the TOMS measurements made it possible to detect calibration drifts of the instrument well enough that small changes in ozone could be detected – trends as small as 1% per decade (2 Sigma).

More recently, a unified long-term time series of global ozone, a Merged Ozone Data Set, has been created using data from multiple SBUV instruments. Data from the series of nine National Aeronautics and Space Administration and NOAA SBUV/2 instruments were reprocessed with a coherent calibration covering the period 1970–72 and 1979–2012. The advantage of concentrating on SBUV instruments is the long time span covered by instruments of similar design and their high accuracy. Because the profile retrieval uses wavelengths that have high sensitivity to ozone, the total column ozone derived by integrating the retrieved profile is estimated to have accuracy better than 1%.

Figure 4 compares ozone measured by satellite instruments since 1979 with that measured by ground-based instruments in Arosa, Switzerland, since 1926. This puts the current ozone decrease in the context of the historical record and emphasizes that both satellites and ground-based instruments are seeing the same ozone changes. Satellite instruments are necessary

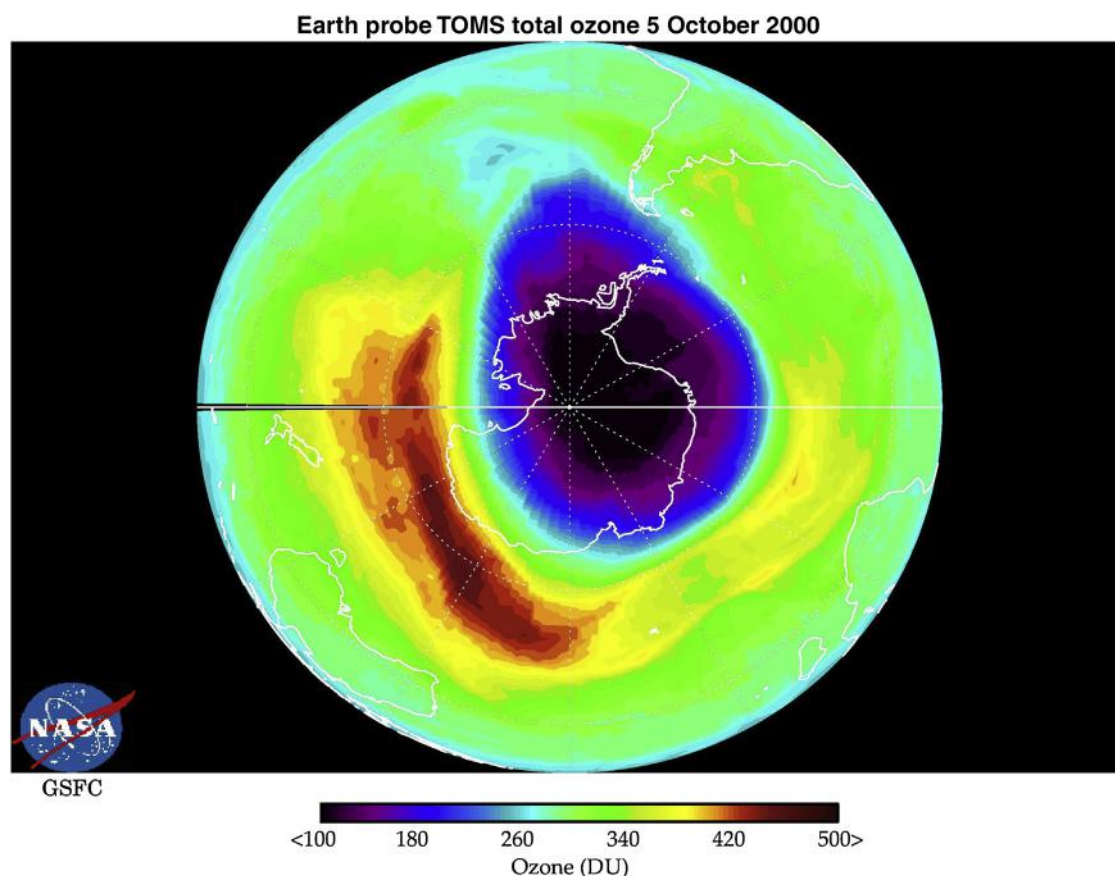


Figure 3 Single-day (5 October 2000) ozone map over the Antarctic. Dark blue to purple shades near pole indicate where total ozone amount is less than 220 DU, a common definition for the region of the Antarctic ozone hole. The map is a polar orthographic projection with the South Pole at the center and the equator at the outer boundary.

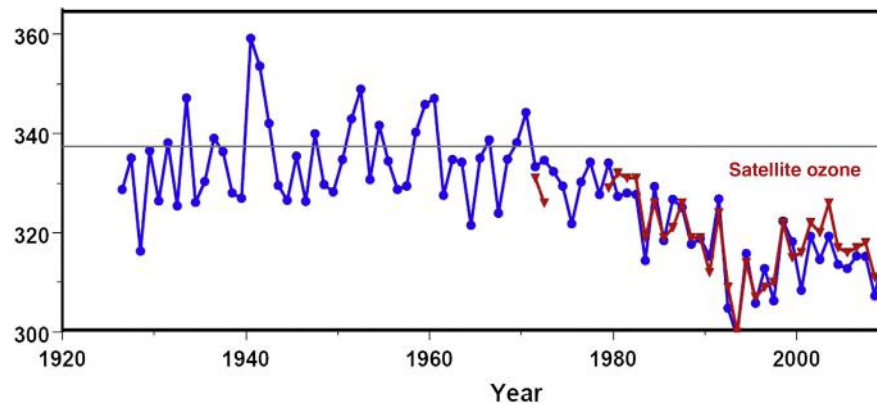


Figure 4 The long-term change in ozone as measured by a series of satellite instruments (in red) compared with ozone measured by ground-based instruments in Switzerland (in blue) since 1926.

to measure global ozone change and show that such ozone decreases are not a local phenomenon.

Dust and Smoke

TOMS measures the reflectivity of the Earth–atmosphere system at several wavelengths not absorbed by ozone. If the atmosphere were perfectly clean, the backscattered radiation received by the satellite could be determined from a Rayleigh-scattering calculation that would predict a specific ratio of radiation between two wavelengths. Aerosols disturb this ratio in a predictable manner: one direction for absorbing aerosols, and the opposite for nonabsorbing aerosols. Using these facts, the TOMS data have been used to determine an ‘aerosol index.’ Reasonable assumptions about the nature of the aerosols lead to global maps of the spread of dust from deserts and smoke from biomass burning in Africa and South America. There are now 40 years’ worth of such data from the TOMS instruments and now OMI and OMPS.

Figure 5 is a map of the aerosol index for 10 March 2006, showing dust from deserts in China being carried across the Pacific. Similar maps are used to track dust from the Sahara Desert crossing the Atlantic Ocean, smoke from forest fires, and ash from volcanoes.

Reflectivity

The basic TOMS measurement is of reflected radiation at six wavelengths. The longer of these wavelengths are not affected by ozone absorption and are thus a measure of the reflectivity of the atmosphere in the UV. The algorithm calculates the expected backscattered radiation from a pure Rayleigh-scattering atmosphere. Deviations from this expectation are driven primarily by clouds and secondarily by aerosols. The deviations caused by clouds, which are different spectrally than the deviations caused by aerosols, can be represented as a percentage reflectivity.

The scene reflectivity of the Earth at blue and UV wavelengths is low over most surfaces (except ice and snow), and are almost independent of the seasonal changes in vegetation on land and in the oceans. This makes it ideal for examining changes in radiation reflected back to space from

changes in cloud and aerosol amounts, especially as affected by the start of climate change. The aerosol index shown in **Figure 5** (in color) is overlaid onto a map of cloud reflectivity (white).

Volcanoes

When a volcano erupts, it releases large quantities of ash and sulfur dioxide (SO_2), both of which can be mapped by TOMS, OMI, and the OMPS total column mapper. SO_2 is an even stronger absorber of UV light than ozone, but with a very different spectral signature. When El Chichon erupted in 1982, SO_2 absorption produced a false apparent enhancement of TOMS ozone. When this was realized, the TOMS algorithm was modified to distinguish ozone from SO_2 . **Figure 6** shows a plume of SO_2 from the eruption of the Grimsvotn volcano in Iceland on 21 May 2011. TOMS (and OMI and OMPS) can easily track plumes of SO_2 from volcanic eruptions. The effect of SO_2 from a volcano is usually short lived because SO_2 is converted rapidly to sulfuric acid aerosols. Because they have full spectral data, OMI and OMPS can even detect SO_2 from power plants and other pollution sources in the lower troposphere.

UV at Surface

Many of the concerns about ozone depletion are related to the increase in UV radiation received at the surface of the Earth. Increases in UV would lead to increased incidence in skin cancer (among humans) and possible damage to the biosphere. TOMS measures the outgoing, absorbed UV radiation and the reflectivity due to clouds and aerosols. These data can be combined with a radiative transfer model to estimate the UV flux at the surface daily across the globe. TOMS and OMI data have been used to generate detailed maps of average UV flux over the Earth.

Tropospheric Ozone

TOMS measures the total column amount of ozone with some adjustments for the inefficiency of the penetration of UV sunlight into the boundary layer. In the tropics, most of the

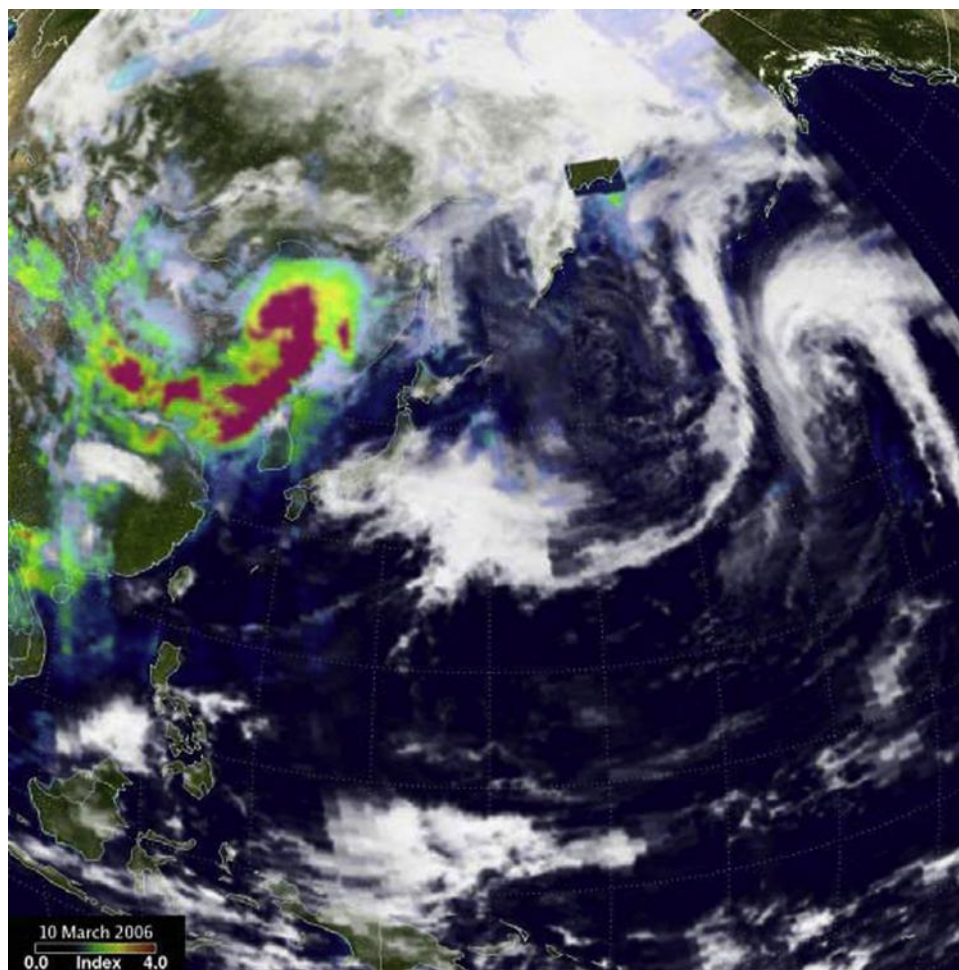


Figure 5 Aerosol index measured by an OMI on 10 March 2006 showing dust blowing out of deserts in China. Here, aerosol index is mapped (in color) with reflectivity over a world map.

variability of total ozone around a circle of constant latitude is in the troposphere rather than the stratosphere. Several schemes have been developed for taking advantage of this property of the total ozone measurements to derive tropical tropospheric ozone column amounts. The first of these combined the TOMS measurements with concurrent measurements from the Stratospheric Aerosol and Gas Experiment (SAGE) occultation measurements of the stratospheric amount. The difference between TOMS total column ozone and SAGE stratospheric ozone will be the amount of ozone in the troposphere. Similarly, OMI measurements of ozone have been combined with ozone profile measurements by the microwave limb sounder, also on the Aura spacecraft, to give very accurate measurements of tropospheric ozone. Alternatively, the stratospheric ozone amount can be estimated directly from TOMS measurements above the location of the highest clouds. That amount can be subtracted from the total ozone measured in clear areas, yielding a tropical map of column tropospheric ozone.

Application of these techniques for deriving tropical tropospheric ozone gives maps showing the ozone generated

by the products of biomass burning. The ozone development and transport can be seen far downwind from the burning source.

Summary and Future of TOMS and SBUV Measurements

We now have more than 40 years' worth of global total ozone data from TOMS instruments, and ozone profile measurements from SBUV. The Earth Probe TOMS was the last of the series, but OMI on Aura and the OMPS total column mapper have continued the measurement series. The SBUV/2 instrument on NOAA 19 is the last of the SBUV series, but the OMPS nadir profiler continues the ozone profile measurements. OMI and the OMPS instruments represent a new generation of ozone-mapping instruments, instruments that use CCD arrays to image the Earth at a large number of wavelengths. These instruments allow us to continue the long time series of ozone measurements to document the expected recovery of the Earth's ozone layer.

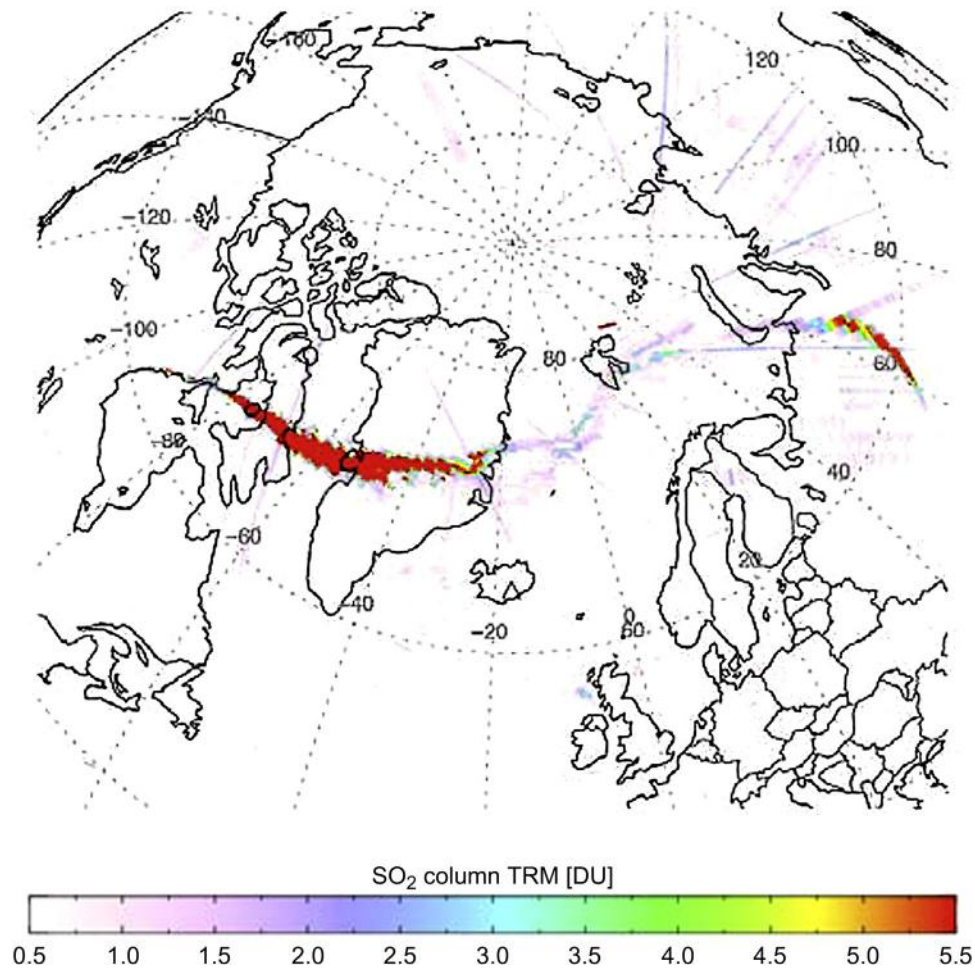


Figure 6 An OMI map of sulfur dioxide shows that SO_2 from the 21 May 2011 eruption of the Grimsvotn volcano in Iceland has been transported over Greenland 2 days later.

See also: **Aerosols:** Climatology of Tropospheric Aerosols.
Ozone Depletion and Related Topics: Long-Term Ozone Changes.
Satellites and Satellite Remote Sensing: Research.

Further Reading

- Dave, J.V., Mateer, C.L., 1967. A preliminary study on the possibility of estimating total atmospheric ozone from satellite measurements. *Journal of the Atmospheric Sciences* 24, 414–427.
- Farman, J.C., Gardiner, B.G., Shanklin, J.D., 1985. Large losses of total ozone in Antarctica reveal seasonal ClO_x/NO_x interaction. *Nature* 315, 207–210.
- McPeters, R.D., et al., 1998. Earth Probe Total Ozone Mapping Spectrometer (TOMS) Data Products User's Guide. NASA Technical Publication 1998–206895. National Aeronautics and Space Administration, Goddard Space Flight Center, Greenbelt, MD.

- McPeters, R., et al., 2008. Validation of the Aura Ozone Monitoring Instrument total column ozone product. *Journal of Geophysical Research* 113. <http://dx.doi.org/10.1029/2007JD008802>.
- Rodgers, C.D., 2000. *Inverse Methods for Atmospheric Sounding, Theory and Practice*. In: *Series on Atmospheric, Oceanic and Planetary Physics*, vol. 2. World Scientific.
- Singer, S.F., Wentworth, R.C., 1957. A method for the determination of the vertical ozone distribution from a satellite. *Journal of Geophysical Research* 62, 299–308.
- Stammes, P., Snee, M., de Haan, J.F., Veefkind, J.P., Wang, P., Levelt, P.F., 2008. Effective cloud fractions from the Ozone Monitoring Instrument: theoretical framework and validation. *Journal of Geophysical Research* 113. <http://dx.doi.org/10.1029/2007JD008820>.
- Stolarski, R.S., Krueger, A.J., Schoeberl, M.R., McPeters, R.D., Newman, P.A., Alpert, J.C., 1986. Nimbus-7 satellite measurements of the springtime Antarctic ozone decrease. *Nature* 322, 808–811.
- Yang, K., et al., 2010. Direct retrieval of sulfur dioxide amount and altitude from spaceborne hyperspectral UV measurements: theory and application. *Journal of Geophysical Research* 115. <http://dx.doi.org/10.1029/2010JD013982>.

Orbits

SQ Kidder, Colorado State University, Fort Collins, CO, USA

© 2003 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 5, pp 2024–2038, © 2003, Elsevier Ltd.

Introduction

To fully understand and use data from meteorological satellites, it is necessary to understand the orbits in which satellites are constrained to move and the geometry with which they view the Earth. This article begins with a review of basic physical principles, which reveal the shape of a satellite orbit and how to orient the orbital plane in space. This knowledge allows us to calculate the position of a satellite at any time. Orbit perturbations and their effects on satellite orbits are then discussed. Finally, the geometry of satellite tracking and Earth location of the measurements made from the satellites are explored.

Newton's Laws

Isaac Newton discovered the basic principles that govern the motions of satellites and other heavenly bodies.

Newton's Laws of Motion

1. Every body will continue in its state of rest or of uniform motion in a straight line except insofar as it is compelled to change that state by an impressed force.
2. The rate of change of momentum is proportional to the impressed force and takes place in the line in which the force acts.
3. Action and reaction are equal and opposite.

Since momentum is the product of the mass of a body and its velocity, Newton's second law is the familiar eqn [1], where F is force, m is mass, a is acceleration, v is velocity, and t is time.

$$F = ma = m \frac{dv}{dt} \quad [1]$$

In addition, Newton gave us the functional form of the force that determines satellite motion in the law of gravitation.

Newton's Law of Universal Gravitation

The force of attraction between two point masses m_1 and m_2 separated by a distance r is given by eqn [2], where G is the Newtonian (or universal) gravitation constant ($6.67259 \times 10^{-11} \text{ N m}^2 \text{ kg}^{-2}$).

$$F = \frac{Gm_1m_2}{r^2} \quad [2]$$

Consider the simple circular orbit shown in Figure 1. Assuming that the Earth is a sphere, we can treat it as a point mass. The centripetal force required to keep the satellite in a circular orbit is mv^2/r , where v is the orbital velocity of the satellite. The force of gravity that supplies this centripetal force is Gm_em/r^2 , where m_e is the mass of the Earth

($5.97370 \times 10^{24} \text{ kg}$) and m is the mass of the satellite. Equating the two forces gives eqn [3].

$$\frac{mv^2}{r} = \frac{Gm_em}{r^2} \quad [3]$$

Division by m eliminates the mass of the satellite from the equation, which means that the orbit of a satellite is independent of its mass. The period of the satellite is the circumference of the orbit divided by the velocity (eqn [4]).

$$T = \frac{2\pi r}{v} \quad [4]$$

Substituting eqn [4] in eqn [3] gives eqn [5] for the period.

$$T^2 = \frac{4\pi^2}{Gm_e} r^3 \quad [5]$$

A typical weather satellite orbits 833 km above the Earth's surface. Since the equatorial radius of the Earth is 6378.137 km, the orbit radius is about 7211 km. Substituting in eqn [5] yields a period of about 6094 s or 102 min.

As a second example, we calculate the radius required for a satellite in geosynchronous orbit, that is, an orbit in which the satellite has the same angular velocity as the Earth ($7.292115 \times 10^{-5} \text{ rad s}^{-1}$). The angular velocity of a satellite is given by eqn [6].

$$\xi = \frac{2\pi}{T} \quad [6]$$

Substituting eqn [6] in eqn [5] gives eqn [7] for the radius.

$$r^3 = \frac{Gm_e}{\xi^2} \quad [7]$$

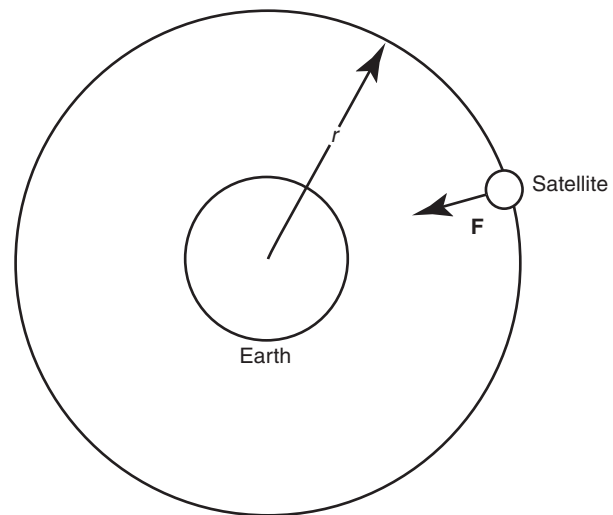


Figure 1 A circular satellite orbit.

Therefore, the required radius for a geosynchronous orbit is about 42 164 km, or about 35 786 km above the Earth's surface.

Keplerian Orbits

Although a circular orbit is the goal for most meteorological satellites, in general satellites do not travel in perfect circles. The exact form of a satellite's orbit may be derived from Newton's laws of motion and the law of universal gravitation. The results of this derivation are neatly summarized in Kepler's laws and in Kepler's equation.

Kepler's Laws

Johannes Kepler died 12 years before Newton was born and, thus, did not have the advantage of Newton's work. Kepler formulated his laws by analyzing data on the position of the planets. This task was complicated by the rotation of the Earth and the motion of the Earth about the Sun, which make planetary motions seem very complex. In modern form, Kepler's laws may be stated as follows.

1. All planets travel in elliptical paths with the Sun at one focus.
2. The radius vector from the Sun to a planet sweeps out equal areas in equal times.
3. The ratio of the square of the period of revolution of a planet to the cube of its semimajor axis is the same for all planets revolving around the Sun.

The same laws apply if we substitute satellite for planet and Earth for Sun. Equation [5] is a statement of Kepler's third law for the special case of a circular orbit.

Ellipse Geometry

The parameters that are used to specify satellite orbits are based in part on geometric terminology. Figure 2 illustrates the geometry of an elliptical orbit. The point where the satellite most closely approaches the Earth is termed the perigee, or more generally the perifocus. The point where the satellite is farthest from the Earth is called the apogee or apofocus. The

distance from the center of the ellipse to the perigee (or apogee) is the semimajor axis (denoted by the symbol a). The distance from the center of the ellipse to one focus (to the center of the Earth) divided by the semimajor axis is the eccentricity (ϵ). For an ellipse, the eccentricity is a number between zero and 1 ($0 < \epsilon < 1$). A circle is an ellipse with zero eccentricity. The equation for the ellipse, that is, the path that the satellite follows, is given in polar coordinates with the center of the Earth as origin by eqn [8].

$$r = \frac{a(1 - \epsilon^2)}{1 + \epsilon \cos \theta} \quad [8]$$

The angle θ (see Figure 3) is the 'true anomaly' and is always measured counterclockwise (the direction of satellite motion) from the perigee.

Kepler's Equation

A satellite in a circular orbit has uniform angular velocity. By Kepler's second law, however, a satellite in an elliptical orbit cannot have uniform angular velocity; it must travel faster

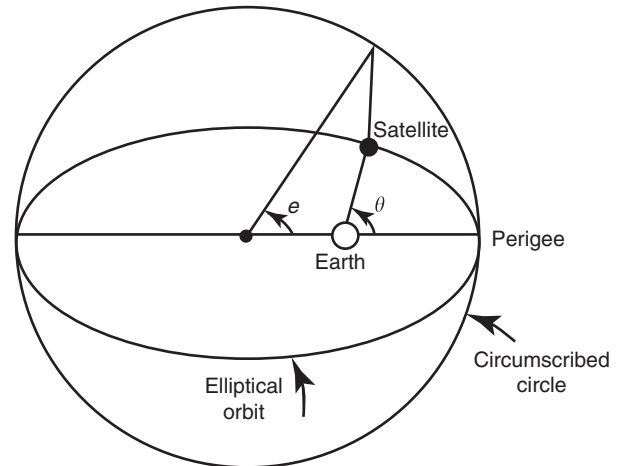


Figure 3 The geometric relationship between true anomaly (θ) and eccentric anomaly (e).

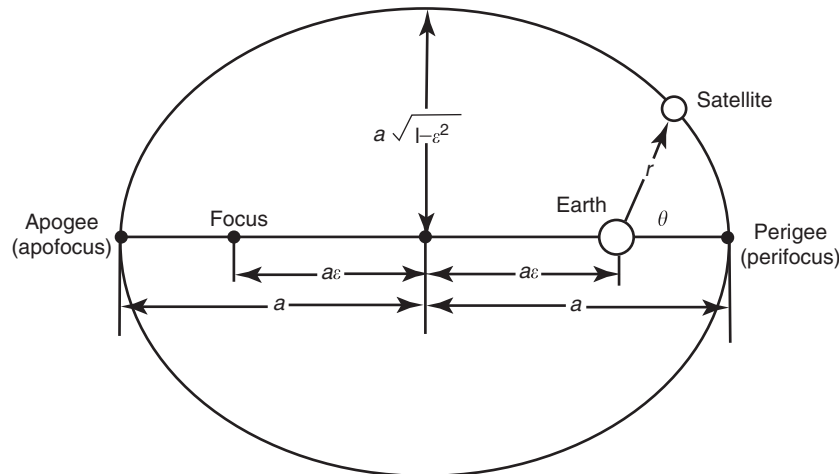


Figure 2 Elliptical orbit geometry.

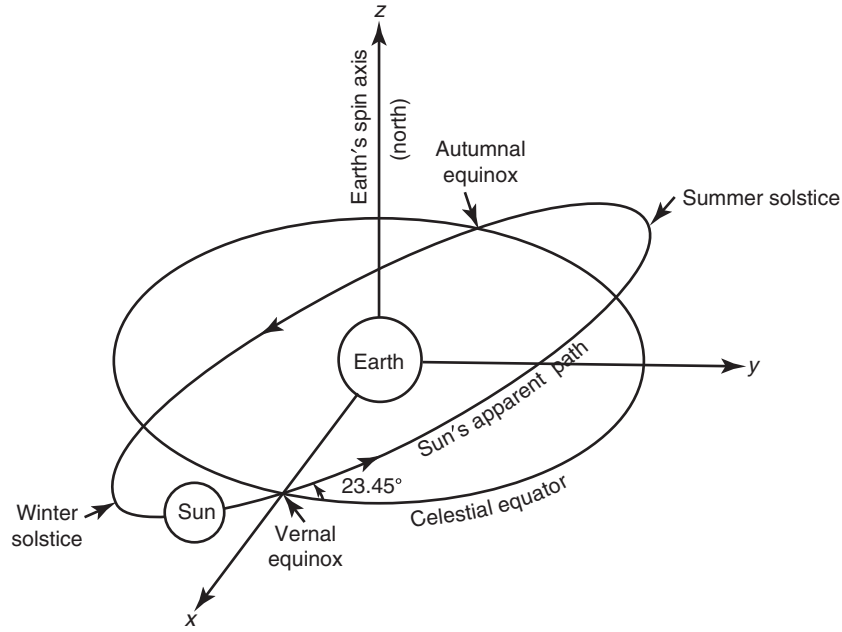


Figure 4 The right ascension–declination coordinate system.

when it is closer to the Earth. The position of the satellite as a function of time can be found by applying Kepler's equation as eqn [9].

$$M = n(t - t_p) = M_0 + n(t - t_0) = e - \varepsilon \sin e \quad [9]$$

Here M is the mean anomaly, an angle that increases linearly in time at the rate n , called the mean motion constant, given by eqn [10].

$$n = \frac{2\pi}{T} = \sqrt{\frac{Gm_e}{a^3}} \quad [10]$$

By definition, M is zero when the satellite is at perigee; therefore, t_p is the time of perigee passage. Time t_0 is called the epoch time. M_0 is called the mean anomaly true of epoch, that is, the mean anomaly at the epoch time t_0 . The angle e is the eccentric anomaly. It is geometrically related to the true anomaly (Figure 3) through eqns ([11a] and [11b]).

$$\cos \theta = \frac{\cos e - \varepsilon}{1 - \varepsilon \cos e} \quad [11a]$$

$$\cos e = \frac{\cos \theta + \varepsilon}{1 + \varepsilon \cos \theta} \quad [11b]$$

Given a , ε , and t_p (or M_0 and t_0), one can calculate r and θ at any time t using eqns ([8], [9], [10], [11a] and [11b]).

Orientation in Space

By calculating r and θ at time t , we have positioned the satellite in the plane of its orbit. Now we must position the orbital plane in space. To do this requires the definition of a coordinate system. This coordinate system must be an inertial coordinate system, that is, a nonaccelerating system in which Newton's laws of motion are valid. A coordinate system fixed to the rotating Earth is not such a system. We will adopt an

astronomical coordinate system called the right ascension–declination coordinate system.¹ In this system (Figure 4), the z -axis is aligned with the Earth's spin axis. The x -axis is chosen such that it points from the center of the Earth to the Sun at the moment of the vernal equinox, when the sun is crossing the equatorial plane from the Southern Hemisphere to the Northern Hemisphere.² The y -axis is chosen so as to make it a right-handed coordinate system. In this system, the declination of a point in space is its angular displacement measured northward from the equatorial plane, and the right ascension is the angular displacement, measured counterclockwise from the x -axis, of the projection of the point in the equatorial plane (Figure 5).

Three angles are used to position an elliptical orbit in the right ascension–declination coordinate system: the inclination angle, the right ascension of ascending node, and the argument of perigee (Figure 6).

The inclination angle (i) is the angle between the equatorial plane and the orbital plane. By convention, the inclination angle is zero if the orbital plane coincides with the equatorial plane and if the satellite rotates in the same direction as the Earth. If the two planes coincide but the satellite rotates opposite to the Earth, the inclination angle is 180° . Prograde orbits are those with inclination angles less than 90° ; retrograde orbits are those with i greater than 90° .

¹ Because the origin of this coordinate system moves about the Sun with the Earth, it is not truly inertial. However, the Sun's gravity causes the satellite to rotate around the Sun as does the Earth. Therefore, the satellite acts as if the right ascension–declination coordinate system were inertial.

² This x -axis is also referred to as the First Point of Aries because it used to point at the constellation Aries. Because of the influence of the Sun and Moon on the nonspherical Earth, the Earth's spin axis precesses like a top with a period of 25 781 years. This causes the vernal equinox to change. Today, the x -axis points to the constellation Pisces, but it is still referred to as the First Point of Aries.

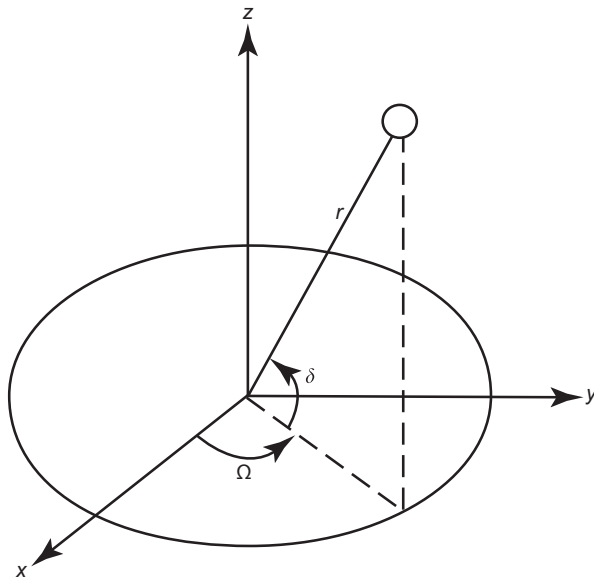


Figure 5 Coordinates used in the right ascension–declination coordinate system: right ascension (Ω), declination (δ), and radius (r).

The ascending node is the point where the satellite crosses the equatorial plane going north (ascends). The right ascension of this point is the right ascension of ascending node (Ω). It is measured in the equatorial plane from the x -axis (vernal equinox) to the ascending node. In practice, the right ascension of ascending node has a more general meaning. It is the right ascension of the intersection of the orbital plane with the equatorial plane; thus, it is always defined, not just when the satellite is actually at an ascending node.

Finally, the argument of perigee (ω) is the angle measured in the orbital plane between the ascending node (equatorial plane) and the perigee.

Orbital Elements

The parameters just discussed for locating a satellite in space are collectively known as the classical orbital elements; they

Table 1 Classical Orbital elements

Element	Symbol
Semimajor axis	a
Eccentricity	e
Inclination	i
Argument of perigee	ω_0
Right ascension of ascending node	Ω_0
Mean anomaly	M_0
Epoch time	t_0

are summarized in [Table 1](#). These parameters may be determined by optical, radar, or radio ranging observations. They are carefully determined by various agencies and are available over the Internet for most satellites. There is some variation in how the orbital elements are specified. Some agencies, for example, substitute true anomaly for mean anomaly. Also, in less formal descriptions of satellite orbits, the height of the satellite above the Earth's surface is substituted for the semimajor axis. Since the Earth is not round, the height of a satellite varies according to its position in the orbit. Such heights are converted into semimajor axis by adding the equatorial radius of the Earth.

Orbits in which the classical orbital elements (except M) are constant are called Keplerian orbits. Viewed from space, Keplerian orbits are simple. The satellite moves in an elliptical path with the center of the Earth at one focus. The ellipse maintains a constant size, shape, and orientation with respect to the stars ([Figure 7](#)). Perhaps surprisingly, the only effect of the Sun's gravity on the satellite is to move the focus of the ellipse (the Earth) in an elliptical path around the Sun (the Earth's orbit).

Viewed from the earth, Keplerian orbits appear complicated because the Earth rotates on its axis as the satellite orbits the Earth ([Figure 8](#)). The rotation of the Earth beneath a fixed orbit results in two daily passes of the satellite near a point on the Earth (assuming that the period is substantially less than a day and that the inclination angle is greater than the latitude of the point). One pass occurs during the ascending portion of the orbit; the other occurs during the descending portion of

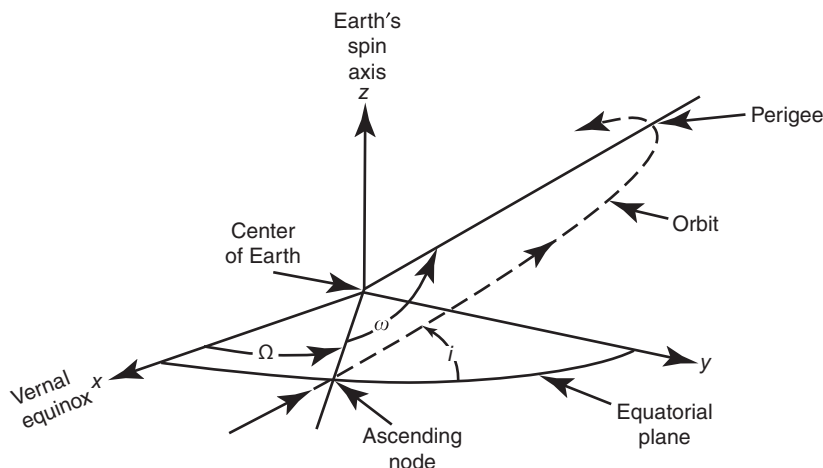


Figure 6 Angles used to orient an orbit in space.

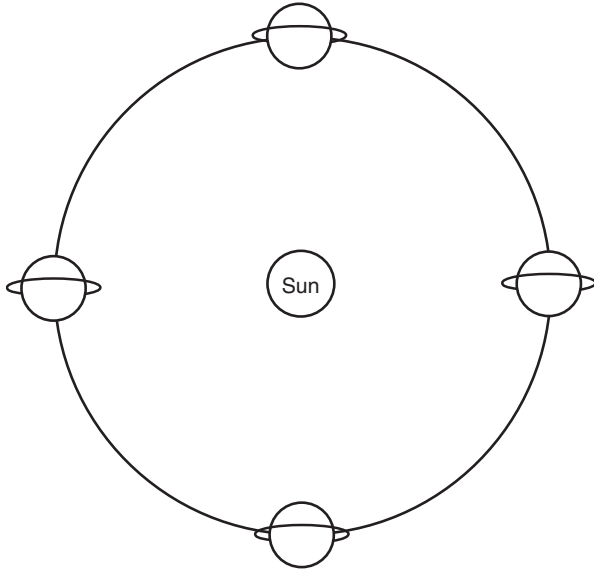


Figure 7 The change with season of a Keplerian orbit.

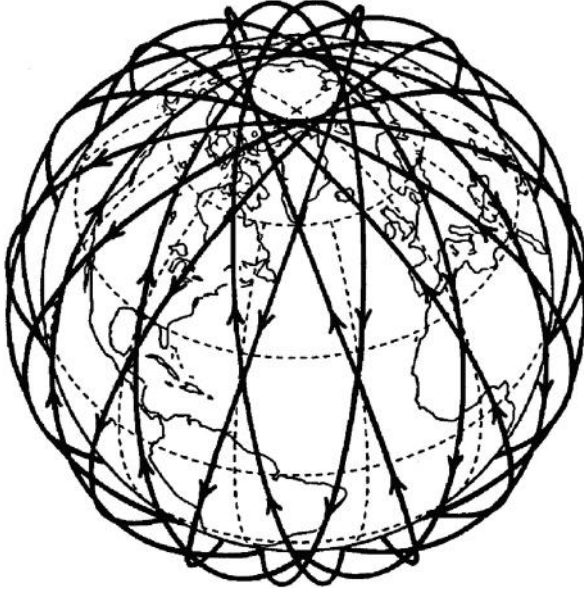


Figure 8 The orbit of a representative satellite as viewed from a point rotating with the Earth.

the orbit. This usually means that one pass occurs during daylight and one during darkness.

Orbit Perturbations

Although satellites travel in nearly Keplerian orbits, these orbits are perturbed by a variety of forces (Table 2). Forces arising from the last five processes are small and can be viewed as causing essentially random perturbations in the Orbitally they are dealt with simply by periodically (1) observing the orbital elements and (2) adjusting the orbit with on-board thrusters. Forces due to the nonsphericity of the Earth cause secular (linear with time) changes in some of the orbital

Table 2 Orbit-perturbing forces

Force	Source
Nonspherical gravitational field	Nonspherical, nonhomogeneous earth
Gravitational attraction of auxiliary bodies	Moon, planets
Radiation pressure	Sun's radiation
Particle flux	Solar wind
Lift and drag	Residual atmosphere
Electromagnetic forces	Interaction of electrical currents in the satellite with the earth's magnetic field

elements. These changes can be predicted theoretically and indeed are useful.

The gravitational potential of the Earth is a complicated function of the Earth's shape, the distribution of land and ocean, and the density of crustal material. As a first-order correction to a spherical shape, we may treat the Earth as an oblate spheroid of revolution. In cross-section, the Earth is approximately elliptical. The distance from the center of the Earth to the Equator is, on average, 6378.137 km, whereas the distance to the poles is 6356.752 km. The gravitational potential of the Earth is given approximately by eqn [12], where r_{ee} is the equatorial radius of the earth, δ is the declination angle, and J_2 (1.08263×10^{-3}) is the coefficient of the quadrupole term.

$$U = -\frac{Gm_e}{r} \left[1 + \frac{1}{2} J_2 \left(\frac{r_{ee}}{r} \right)^2 (1 - 3 \sin^2 \delta) + \dots \right] \quad [12]$$

The higher-order terms are more than two orders of magnitude smaller than the quadrupole term and will not be considered here, although they are necessary for very accurate calculations.

A satellite travels at a slightly different speed in this gravitationally perturbed orbit. The time rate of change of the mean anomaly is given by the mean motion constant (n) in the unperturbed orbit and by the anomalistic mean motion constant ($\bar{n}$) in a perturbed orbit. Considering only the quadrupole term we have eqn [13].

$$\frac{dM}{dt} = \bar{n} = n \left[1 + \frac{3}{2} J_2 \left(\frac{r_{ee}}{a} \right)^2 (1 - \epsilon^2)^{-3/2} (1 - \frac{3}{2} \sin^2 i) \right] \quad [13]$$

When the inclination angle is less than 54.7° or greater than 125.3° , $\bar{n}$ is greater than n , and the satellite orbits faster than it would in an unperturbed orbit. For inclination angles between 54.7° and 125.3° , the satellite orbits more slowly than it otherwise would.

The rate of change of the right ascension of ascending node is given by eqn [14].

$$\frac{d\Omega}{dt} = -\bar{n} \left[\frac{3}{2} J_2 \left(\frac{r_{ee}}{a} \right)^2 (1 - \epsilon^2)^{-2} \cos i \right] \quad [14]$$

The rate of change of the argument of perigee is given by eqn [15].

$$\frac{d\omega}{dt} = \bar{n} \left[\frac{3}{2} J_2 \left(\frac{r_{ee}}{a} \right)^2 (1 - \epsilon^2)^{-2} \left(2 - \frac{5}{2} \sin^2 i \right) \right] \quad [15]$$

The other three orbital elements, a , ϵ , and i , undergo small, oscillatory changes that may be neglected.

The anomalistic period of a perturbed orbit is simply that given by eqn [16].

$$\bar{T} = \frac{2\pi}{\bar{n}} \quad [16]$$

However, because M is measured from perigee, the anomalistic period is the time for the satellite to travel from perigee to moving perigee. Of more use is the synodic or nodal period, $\tilde{T}$, which is the time for the satellite to travel from one ascending node to the next ascending node. An exact value of $\tilde{T}$ must be calculated numerically; however, eqn [17] holds to very good approximation.

$$\tilde{T} = \frac{2\pi}{\bar{n} + (d\omega/dt)} \quad [17]$$

In summary, the first-order effects of the nonspherical gravitational potential of the Earth consist of a slow, linear change in two of the classical orbital elements, the right ascension of ascending node and the argument of perigee, and a small change in the mean motion constant. Table 3 shows orbital elements for some representative satellites.

Meteorological Satellite Orbits

Nearly all meteorological satellites are in one of two orbits, Sun-synchronous or geostationary, but other orbits are also useful.

Sun-Synchronous Orbits

As shown in Figure 7, in a Keplerian orbit the angle between the Sun and the plane of a satellite's orbit changes because the orbital plane is fixed while the Earth rotates around the Sun. This causes the satellite to pass over an area at different times of the day. For example, if the satellite passes over near noon (1200) and midnight (2400) in the spring, it will pass over near 0600 and 1800 in the winter. Fortunately, the perturbations

caused by the nonspherical earth can be employed to keep the Sun–Earth–satellite angle nearly constant.

The Earth makes one complete revolution about the Sun (2π radians) in one tropical year (31 556 925.9747 s). Thus, the right ascension of the Sun changes at the average rate of $1.991064 \times 10^{-7} \text{ rad s}^{-1}$ ($0.9856473^\circ \text{ day}^{-1}$). If the inclination of the satellite is correctly chosen, the right ascension of its ascending node can be made to precess at this rate. An orbit that is so synchronized with the Sun is called a Sun-synchronous orbit. For a satellite with a semimajor axis of 7221 km and zero eccentricity, eqn [14] requires an inclination of 98.75° to be Sun-synchronous. Figure 9 shows the change with season of a Sun-synchronous orbit.

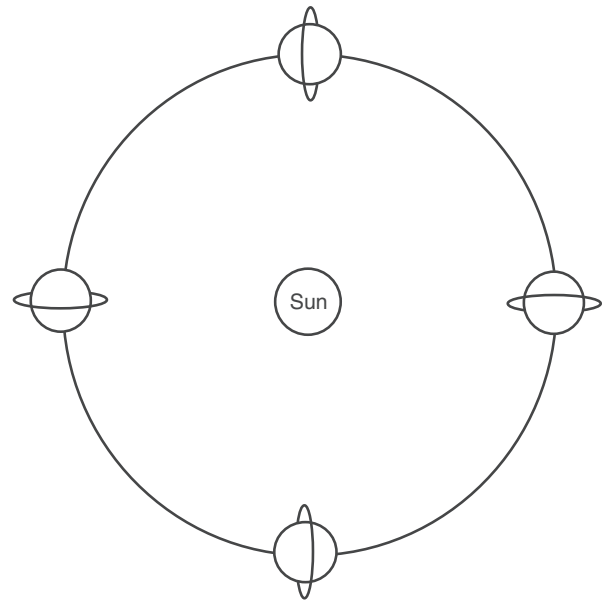


Figure 9 The change with season of a Sun-synchronous orbit.

Table 3 Orbital elements of representative satellites

Satellite					Right ascension of ascending node		Argument of perigee		Mean anomaly		Nodal period (min)
Name	ID	Semimajor axis (km)	Inclination (deg)	Eccentricity	Value (deg)	Motion (deg day ⁻¹)	Value (deg)	Motion (deg day ⁻¹)	Value (deg)	Motion (deg day ⁻¹)	
INSAT 3B	00016B	42 165.44	0.08	4.846×10^{-4}	288.81	−0.0134	233.08	0.0268	266.58	360.98	1435.97
GOES 8	94022A	42 164.66	0.16	3.691×10^{-4}	104.85	−0.0134	61.66	0.0268	104.11	360.99	1435.93
GOES 10	97019A	42 166.53	0.25	3.304×10^{-4}	276.33	−0.0134	248.95	0.0268	44.58	360.97	1436.03
METEOSAT7	97049B	42 164.70	0.54	4.900×10^{-5}	296.54	−0.0134	68.78	0.0268	339.61	360.99	1435.94
GMS 5	95011B	42 166.75	0.58	1.647×10^{-4}	65.29	−0.0134	259.00	0.0268	161.45	360.97	1436.04
FENGYUN 2B	00032A	42 167.40	0.94	9.130×10^{-5}	264.49	−0.0134	164.52	0.0268	20.73	360.96	1436.07
ELEKTRO	97069A	42 171.69	3.25	5.438×10^{-4}	81.41	−0.0134	138.84	0.0267	202.39	360.90	1436.29
TRMM	97074A	6 729.00	34.98	1.923×10^{-4}	6.60	−6.7736	293.05	9.7410	159.10	5666.31	91.33
UARS	91063B	6 948.65	56.98	5.552×10^{-4}	45.52	−4.0224	102.59	1.7883	87.93	5395.39	96.05
ERBS	84108B	6 953.02	57.00	8.553×10^{-4}	330.37	−4.0122	94.57	1.7807	59.04	5390.29	96.14
MOLNIYA 3-4	98040A	26 554.87	63.08	7.285×10^{-1}	126.09	−0.1391	279.98	0.0038	221.47	722.21	717.79
METEOR 3-6	94003A	7 572.34	82.56	1.542×10^{-3}	252.61	−0.7073	350.18	−2.5018	335.10	4740.49	109.41
TERRA	99068A	7 077.71	98.18	3.067×10^{-4}	326.40	0.9847	104.03	−3.1085	207.77	5245.62	98.88
QUICKSCAT	99034A	7 180.38	98.63	3.750×10^{-5}	73.99	0.9868	0.98	−2.9189	264.67	5133.63	101.04
NOAA 15	98030A	7 189.40	98.63	1.168×10^{-3}	277.24	0.9828	44.01	−2.9058	223.53	5123.98	101.23
FENGYUN 1C	99025A	7 233.57	98.73	1.495×10^{-3}	289.04	0.9733	30.07	−2.8357	135.06	5077.16	102.16

Epoch time = 0000 UTC 6 September 2000.

The subsatellite point is the point on the Earth's surface that is directly between the satellite and the center of the Earth. The ground track of a satellite is the path of the subsatellite point. **Figure 10** shows the ground track for three orbits of the Sun-synchronous NOAA 11 satellite.

Geostationary Orbits

Earlier we calculated the radius of a geosynchronous orbit to be 42 164 km. Perturbations due to the nonspherical Earth, however, require a slight adjustment in this figure. The adjustment is small because the radius of geosynchronous orbit is about 6.6 Earth radii and the correction terms are inversely proportional to the square of this ratio. For a geosynchronous orbit with zero eccentricity and zero inclination, eqns [6], [13], [15] and [17] require a semimajor axis of 42 166.3 km.

The terms geosynchronous and geostationary are often used interchangeably. In fact, they are not the same. Geosynchronous means that the satellite orbits with the same angular velocity as the Earth. A geostationary orbit is geosynchronous, but it is also required to have zero inclination angle and zero eccentricity. Geostationary satellites, therefore, remain essentially motionless above a point on the Equator. They are classified by the longitude of their subsatellite point.

Second-order perturbations cause a geostationary satellite to drift from the desired orbit. Periodic maneuvers, performed as frequently as once a week, are required to correct the orbit. These maneuvers keep operational geostationary satellites very close to the desired orbit.

Figure 11 shows the ground track of a typical geostationary satellite.

Other Orbits

Geostationary and Sun-synchronous are only two of an infinity of possible orbits. Others have been and will become useful for meteorological satellites.

The Earth Radiation Budget Satellite (ERBS) was launched from the Space Shuttle and orbits at an altitude of 600 km with an inclination angle of 57° . It was placed in this orbit so that it would precess with respect to the Sun and sample all local times over the course of a month.

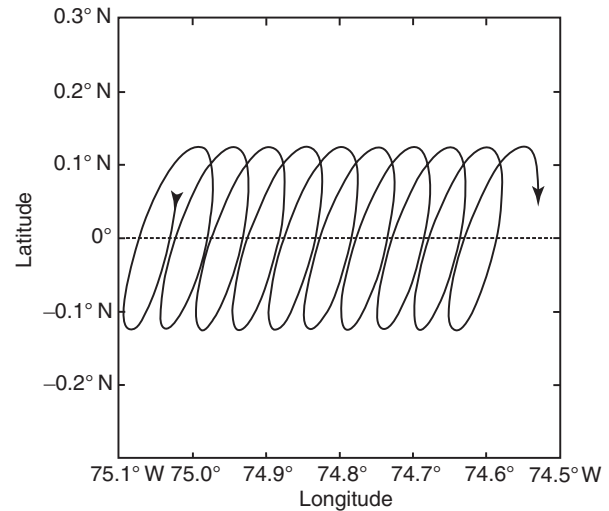


Figure 11 The ground track of a typical geostationary satellite (ten orbits of GOES 8).

Meteor satellites fly in low Earth orbit with inclination angles of about 82° . Molniya communications satellites fly in highly elliptical orbits. It has been suggested that this orbit would be useful for meteorological observations of the high latitudes. The Molniya orbit has an inclination angle of 63.4° , at which the argument of perigee is motionless; thus, the apogee, from which measurements are made, stays at a given latitude. The semimajor axis is chosen such that the satellite makes two orbits while the Earth turns once with respect to the plane of the orbit. The eccentricity is made as large as possible so that the satellite will stay near apogee longer. However, the eccentricity must not be so large that the satellite encounters significant atmospheric drag at perigee. A semimajor axis of 26 554 km and an eccentricity of 0.72 result in a perigee of 7378 km (1000 km above the Equator), an apogee of 45 730 km (39 352 km above the Equator), and a period of 717.8 min. The attractiveness of this orbit is that it functions as a high-latitude, part-time, nearly geostationary satellite. For about 8 h centered on apogee, the satellite is synchronized with the Earth so that it is nearly stationary in the sky. For a meteorological satellite in a Molniya orbit, the rapid imaging

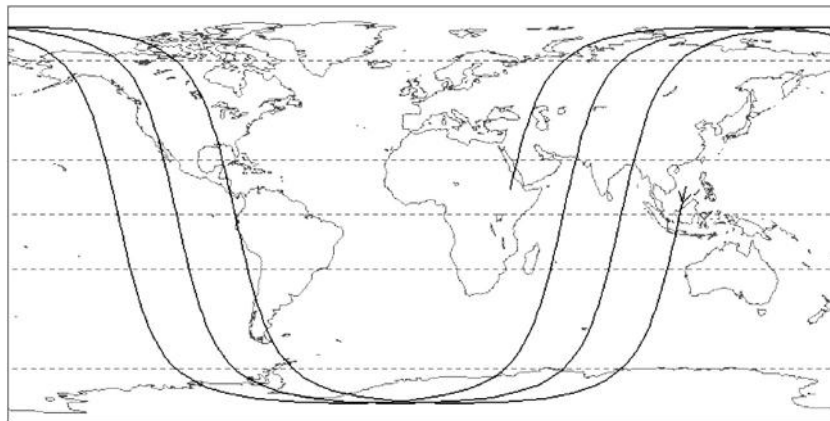


Figure 10 The ground track of a typical Sun-synchronous satellite (three orbits of NOAA 15).

capability, which is so useful from geostationary orbit, would be available in the high latitudes.

As meteorological satellite instruments become more specialized, more custom orbits are likely to be used.

Satellite Positioning, Tracking, and Navigation

It is important to be able to calculate the position of a satellite in space, to track it from Earth, and to know where its instruments are pointing. These topics are discussed in turn in this section.

Positioning in Space

To locate a satellite in a perturbed orbit at time t , one needs current values of the orbital elements. The three constant elements, a , e , and i , are taken directly from a recent bulletin. Such bulletins are available from a variety of sources, and many are available on the Internet. The other three, M , Ω , and ω , are calculated according to eqns ([18a], [18b] and [18c])

$$M = M_0 + \frac{dM}{dt}(t - t_0) \quad [18a]$$

$$\Omega = \Omega_0 + \frac{d\Omega}{dt}(t - t_0) \quad [18b]$$

$$\omega = \omega_0 + \frac{d\omega}{dt}(t - t_0) \quad [18c]$$

Next, the satellite is located in the plane of its orbit; that is, the true anomaly θ and the radius r are calculated. This is done

by (1) solving for e using Kepler's equation [9]; (2) calculating θ using eqn [11a]; and (3) calculating r using eqn [8]. (For a circular orbit, this step is simplified because the mean anomaly, the eccentric anomaly, and the true anomaly are identical, and r is constant.)

A vector can now be constructed that points from the center of the Earth to the satellite in the right ascension–declination coordinate system. The Cartesian coordinates of this vector are given by eqn [19].

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} r \cos \theta \\ r \sin \theta \\ 0 \end{pmatrix} \quad [19]$$

At this point, the orbital ellipse is assumed to lie in the x - y plane (the equatorial plane) with the perigee on the positive x -axis (Figure 12(a)).

In the next three steps, the vector is rotated so that the orbital plane is properly oriented in space. First, the vector is rotated about the z -axis through the argument of perigee (Figure 12(b)). This rotation is conveniently accomplished by multiplying the vector by a rotation matrix (eqn [20]).

$$\begin{pmatrix} x' \\ y' \\ z' \end{pmatrix} = \begin{pmatrix} \cos \omega & -\sin \omega & 0 \\ \sin \omega & \cos \omega & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} \quad [20]$$

$$= \begin{pmatrix} x \cos \omega - y \sin \omega \\ x \sin \omega + y \cos \omega \\ z \end{pmatrix}$$

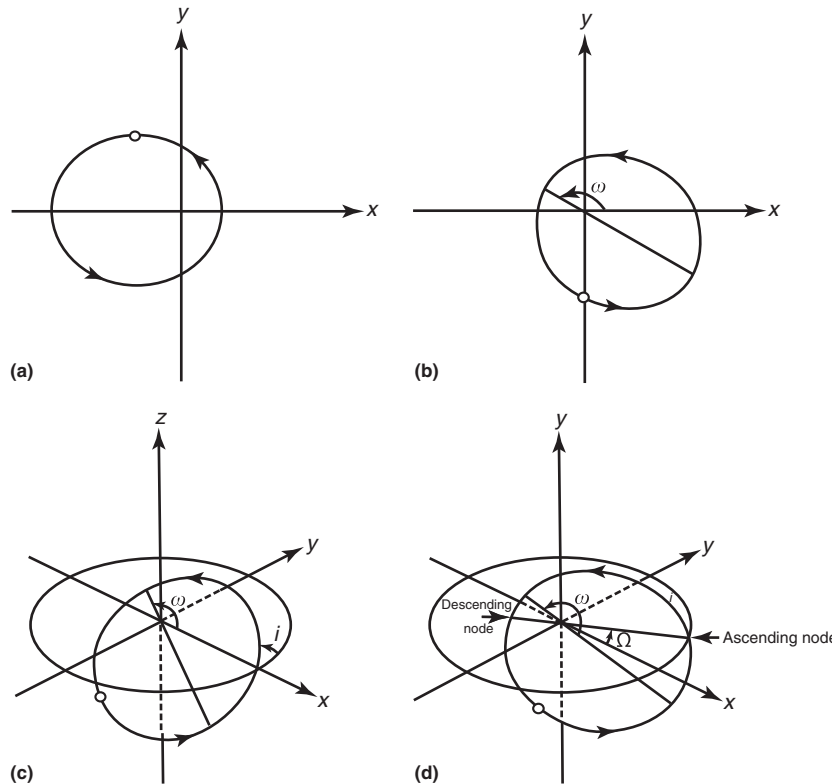


Figure 12 Rotations used to position a satellite in its orbit: (a) the satellite in the plane of its orbit; (b) rotation about the z -axis through the argument of perigee (ω); (c) rotation about the x -axis through the inclination angle (i); and (d) rotation about the z -axis through the right ascension of ascending node (Ω).

Second, the vector is rotated about the x -axis through the inclination angle (Figure 12(c)) as in eqn [21].

$$\begin{aligned} \begin{pmatrix} x'' \\ y'' \\ z'' \end{pmatrix} &= \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos i & -\sin i \\ 0 & \sin i & \cos i \end{pmatrix} \begin{pmatrix} x' \\ y' \\ z' \end{pmatrix} \\ &= \begin{pmatrix} x' \\ y' \cos i - z' \sin i \\ y' \sin i + z' \cos i \end{pmatrix} \end{aligned} \quad [21]$$

Third, the vector is rotated about the z -axis through the right ascension of the ascending node (Figure 12(d)) as in eqn [22].

$$\begin{aligned} \begin{pmatrix} x''' \\ y''' \\ z''' \end{pmatrix} &= \begin{pmatrix} \cos \Omega & -\sin \Omega & 0 \\ \sin \Omega & \cos \Omega & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x'' \\ y'' \\ z'' \end{pmatrix} \\ &= \begin{pmatrix} x'' \cos \Omega - y'' \sin \Omega \\ x'' \sin \Omega + y'' \cos \Omega \\ z'' \end{pmatrix} \end{aligned} \quad [22]$$

The vector (x''', y''', z''') is the location of the satellite in the right ascension–declination coordinate system at time t . This vector may be converted into the radius, declination, and right ascension of the satellite through eqns [23a], [23b] and [23c].

$$r_s = \sqrt{x'''^2 + y'''^2 + z'''^2} = r \quad [23a]$$

$$\delta_s = \sin^{-1} \left(\frac{z'''}{r_s} \right) \quad [23b]$$

$$\Omega_s = \tan^{-1} \left(\frac{y'''}{x'''} \right) \quad [23c]$$

Finally, it is useful to calculate the latitude and longitude of the subsatellite point. Assuming that the Earth is a sphere, the latitude (known as the geocentric latitude) is simply equal to the declination (eqn [24]).

$$\text{Geocentric latitude} = \varphi_s = \delta_s \quad [24]$$

For more precise measurements, the latitude corrected for the nonspherical shape of Earth (the geodetic latitude) is usually used (eqn [25]).

$$\begin{aligned} \text{Geodetic latitude} &= \varphi_g \\ &= \tan^{-1} \left[\left(\frac{r_{ee}}{r_{ep}} \right)^2 \tan \delta_s \right] \end{aligned} \quad [25]$$

where r_{ep} is the polar radius of the Earth. The longitude of the subsatellite point (λ_s) is the difference between the right ascension of the satellite and the right ascension of the prime meridian (0° longitude), which passes through Greenwich, England (Figure 13) (eqn [26]).

$$\lambda_s = \Omega_s - \Omega_{\text{Greenwich}} \quad [26]$$

The right ascension of Greenwich is given in some satellite bulletins, and it can be calculated from eqn [27], where Δt is the time difference in days from 0000 UTC 1 January 2000.

$$\Omega_{\text{Greenwich}} = 99.9643^\circ + 360.9856376^\circ \Delta t \quad [27]$$

Since the rotation rate changes very slightly, owing to the actions of the wind and ocean currents, eqn [27] must be updated periodically.

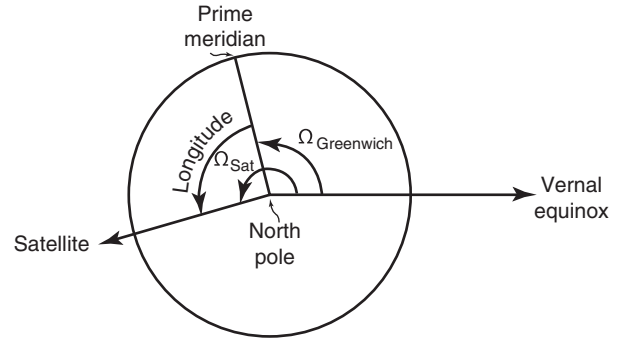


Figure 13 The relationship between Earth longitude and right ascension.

The inverse problem of finding when a satellite passes over (or close to) a particular point is solved iteratively by (1) estimating the time, (2) calculating the position of the satellite, and (3) correcting the time estimate. Steps 2 and 3 are repeated until a satisfactory solution is found.

The above method can be streamlined in two ways. First, some of the rotations can be combined. Start as above by updating the orbital elements and calculating r_s and θ at time t . Locate the satellite on the x -axis at distance r_s from the center of the Earth (eqn [28]).

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} r_s \\ 0 \\ 0 \end{pmatrix} \quad [28]$$

Define Γ , the argument of latitude, to be the angle, measured in the orbital plane, from the ascending node to the satellite as in eqn [29], where θ is the true anomaly and ω is the argument of perigee.

$$\Gamma = \theta + \omega \quad [29]$$

Rotate this vector about the z -axis through the argument of latitude. Rotate again about the x -axis through the inclination angle. Finally, rotate about the z -axis through an angle equal to the right ascension of the satellite less the right ascension of Greenwich. Equations [23a], [23b] and [23c] now yield r_s , latitude φ_s , and longitude λ_s . This method is useful for the navigation problem discussed below.

A second way to streamline these equations is to combine them, which results in eqns [30a], [30b] and [30c].

$$r_s = r \quad [30a]$$

$$\varphi_s = \delta_s = \sin^{-1}(\sin \Gamma \sin i) \quad [30b]$$

$$\begin{aligned} \lambda_s &= \tan^{-1} \left(\frac{\sin \Gamma \cos i}{\cos \Gamma} \right) + \Omega_0 - \Omega_e(t_0) \\ &\quad - \left(\frac{d\Omega_e}{dt} - \frac{d\Omega}{dt} \right) (t - t_0) \end{aligned} \quad [30c]$$

Here r_s is the distance of the satellite calculated with eqn [8]; φ_s and λ_s are its latitude and longitude, respectively. $\Omega_e(t_0)$ is the right ascension of Greenwich at the epoch time, and therefore, $\Omega_0 - \Omega_e(t_0)$ is the longitude of ascending node at the epoch time. The quantity $(d\Omega_e/dt - d\Omega/dt)$ is the relative Earth rotation rate, that is, the rotation rate of the Earth relative to the orbital plane.

Tracking

A list of time versus position of a celestial body is called an 'ephemeris' (plural: ephemerides). To track a satellite, one must be able to point one's antenna at it. The elevation angle, measured from the local horizontal, and the azimuth angle, measured clockwise from the north, can be calculated from the ephemeris data as follows.

Suppose the subsatellite point is at latitude φ_S and longitude λ_S , and that the satellite is at radius r_S from the center of the Earth. Suppose also that the antenna is located at latitude φ_e , longitude λ_e , and radius r_e (the radius of the earth). The Cartesian coordinates of the satellite are then given by eqn [31].

$$\mathbf{r}_S = \begin{pmatrix} x_S \\ y_S \\ z_S \end{pmatrix} = \begin{pmatrix} r_S \cos \varphi_S \cos \lambda_S \\ r_S \cos \varphi_S \sin \lambda_S \\ r_S \sin \varphi_S \end{pmatrix} \quad [31]$$

The Cartesian coordinates of the antenna are given by eqn [32].

$$\mathbf{r}_e = \begin{pmatrix} x_e \\ y_e \\ z_e \end{pmatrix} = \begin{pmatrix} r_e \cos \varphi_e \cos \lambda_e \\ r_e \cos \varphi_e \sin \lambda_e \\ r_e \sin \varphi_e \end{pmatrix} \quad [32]$$

The difference vector $\mathbf{r}_D \equiv \mathbf{r}_S - \mathbf{r}_e$ points from the antenna to the satellite (Figure 14). Assuming a spherical earth, the vector $\mathbf{r}_e$ points to the local vertical (Figure 15). The cosine of the satellite's zenith angle ζ (the complement of the elevation angle) is given by eqn [33].

$$\cos \zeta = \frac{\mathbf{r}_e \cdot \mathbf{r}_D}{|\mathbf{r}_e| |\mathbf{r}_D|} \quad [33]$$

Finding the azimuth angle is a little more difficult. First, we need to find two vectors in the tangent plane at the antenna. The first points north (eqn [34]).

$$\mathbf{r}_N = \begin{pmatrix} x_N \\ y_N \\ z_N \end{pmatrix} = \begin{pmatrix} -\sin \varphi_N \cos \lambda_N \\ -\sin \varphi_N \sin \lambda_N \\ \cos \varphi_N \end{pmatrix}. \quad [34]$$

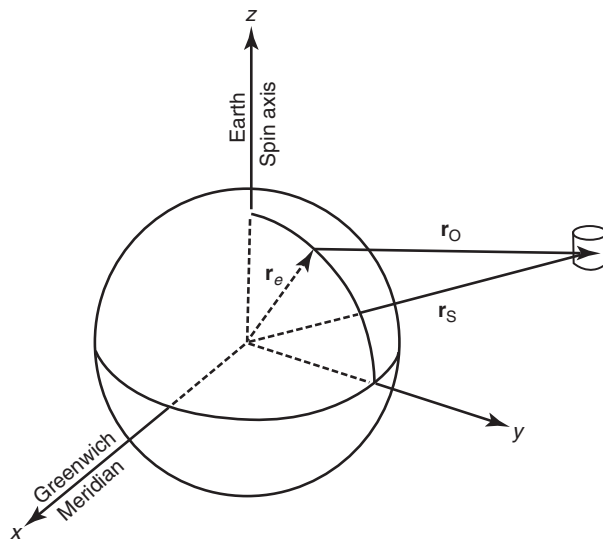


Figure 14 Satellite tracking geometry.

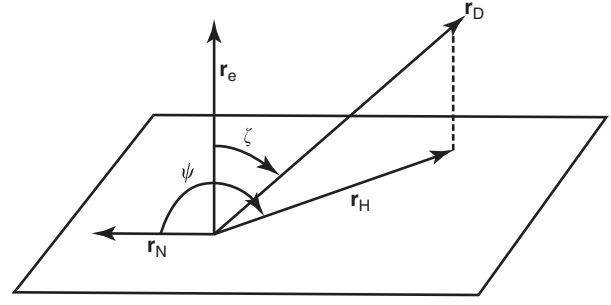


Figure 15 Definition of zenith angle (ζ) and azimuth angle (ψ).

The second is the horizontal projection of $\mathbf{r}_D$. We define unit vectors in the directions of $\mathbf{r}_e$ and $\mathbf{r}_D$ as in eqns [35a] and [35b].

$$\hat{\mathbf{r}}_e \equiv \frac{\mathbf{r}_e}{|\mathbf{r}_e|} \quad [35a]$$

$$\hat{\mathbf{r}}_D \equiv \frac{\mathbf{r}_D}{|\mathbf{r}_D|} \quad [35b]$$

The required horizontal vector is given by eqn [36].

$$\begin{aligned} \mathbf{r}_H &= \mathbf{r}_D - (\hat{\mathbf{r}}_e \cdot \mathbf{r}_D) \hat{\mathbf{r}}_e = \mathbf{r}_D - |\mathbf{r}_D| \cos \zeta \hat{\mathbf{r}}_e \\ &= |\mathbf{r}_D| (\hat{\mathbf{r}}_D - \cos \zeta \hat{\mathbf{r}}_e) \end{aligned} \quad [36]$$

The azimuth angle ψ is then given by eqn [37].

$$\cos \psi = \frac{\mathbf{r}_N \cdot \mathbf{r}_H}{|\mathbf{r}_N| |\mathbf{r}_H|} \quad [37]$$

One must be careful when taking the inverse cosine. If the satellite is west of the antenna, ψ will be greater than 180° . It must also be noted that these equations assume a spherical Earth. Fortunately, most receiving antennas are insensitive to the slight errors that this assumption causes.

Navigation

In addition to knowing where a satellite is in its orbit, it is necessary to know the Earth coordinates (latitude, longitude) of the particular scene it is viewing. The problem of calculating the Earth coordinates is known as the navigation problem; fundamentally, it is a complex geometry problem. It requires an accurate knowledge of where the satellite is in its orbit, the orientation of the satellite (its attitude), and the scanning geometry of the instrument involved.

Suppose that at a particular time a satellite is at position (x_s, y_s, z_s) with respect to the center of the Earth in the right ascension–declination coordinate system. Suppose further, that the telescope is pointing in a direction specified by declination δ_t and right ascension Ω_t . A unit vector in the direction in which the telescope is pointing is given by eqn [38].

$$\begin{pmatrix} x_T \\ y_T \\ z_T \end{pmatrix} = \begin{pmatrix} \cos \delta_T \cos \Omega_T \\ \cos \delta_T \sin \Omega_T \\ \sin \delta_T \end{pmatrix} \quad [38]$$

Figure 16 shows that the ray from which the telescope receives radiation (that is, the line in space through the

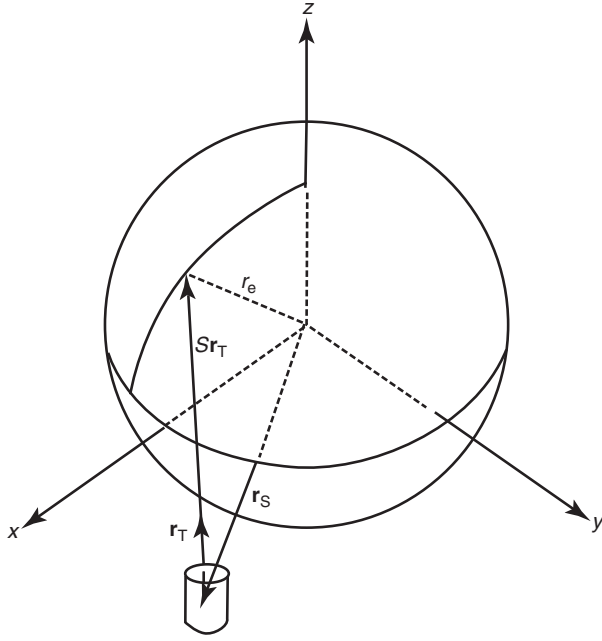


Figure 16 Navigation geometry.

satellite and in the direction of the telescope) is given by eqn [39],

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} x_s + sx_T \\ y_s + sy_T \\ z_s + sz_T \end{pmatrix} \quad [39]$$

where s is the distance from the satellite.

The location at which this ray strikes the spherical Earth is the solution of eqn [40].

$$(x_s + sx_T)^2 + (y_s + sy_T)^2 + (z_s + sz_T)^2 = r_e^2 \quad [40]$$

This is a quadratic equation in s that has no real roots if the ray does not intersect the Earth or two real roots if it does. The smaller root is to be chosen; the larger root represents the location from which the ray re-emerges from the opposite side of the Earth. When the ray is just tangent to the Earth, the two roots are equal. After a solution for s has been found, eqn [39] gives the Cartesian coordinates in the right ascension–declination coordinate system of the point on the Earth's surface being viewed.

One way to specify the telescope pointing vector is to use the pitch, roll, and yaw angles familiar from aircraft flight. Position the satellite on the positive x -axis at distance r_s from the center of the Earth. Let the satellite be traveling in the x - y plane with the positive z -axis on the left; that is, the satellite is traveling eastward in the equatorial plane. Orient the satellite so that its 'nose' is pointing in the $\hat{y}$ -direction (not parallel to the velocity vector), the left 'wing' is pointing in the $\hat{z}$ -direction, and 'up' is in the $\hat{x}$ -direction. Let the telescope begin by pointing straight down toward the center of the Earth, that is in the direction given by eqn [41].

$$\begin{pmatrix} x_T \\ y_T \\ z_T \end{pmatrix} = \begin{pmatrix} -1 \\ 0 \\ 0 \end{pmatrix} \quad [41]$$

Pitch is defined as the angle of rotation about the z -axis; positive is in the sense of the nose pointing up.³ The matrix that accomplishes this rotation is given in eqn [42].

$$\text{Pitch rotation matrix} = \begin{pmatrix} \cos P & -\sin P & 0 \\ \sin P & \cos P & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad [42]$$

A positive pitch angle causes the telescope to point in the along-track direction ahead of the subsatellite point.

Roll is defined as the angle of rotation about the y -axis; positive is in the sense of the left wing pointing up. The matrix that accomplishes this rotation is given in eqn [43].

$$\text{Roll rotation matrix} = \begin{pmatrix} \cos R & 0 & -\sin R \\ 0 & 1 & 0 \\ \sin R & 0 & \cos R \end{pmatrix} \quad [43]$$

A positive roll angle causes the telescope to point in the cross-track direction, to the left of the subsatellite point.

Yaw is defined as the angle of rotation about the x -axis; positive is in the sense of the nose pointing right. The matrix that accomplishes this rotation is given in eqn [44].

$$\text{Yaw rotation matrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos Y & -\sin Y \\ 0 & \sin Y & \cos Y \end{pmatrix} \quad [44]$$

A nonzero yaw angle does nothing to a telescope pointing straight down; however, if the pitch or roll angles are nonzero, a positive yaw angle moves the telescope in a clockwise direction around the subsatellite point.

Common scanning schemes can easily be described with these angles. A cross-track scanner can be described by a roll angle that increases linearly in time (right-to-left scanning) or decreases linearly in time (left-to-right scanning). A conical scanning instrument can be described by a constant pitch angle followed by a yaw angle that increases linearly in time (for clockwise scanning). Finally, a geostationary scanner can be described by a stepped roll angle followed by a pitch angle that increases (west-to-east scanning) or decreases (east-to-west scanning) with time. Corrections in the pitch, roll, and yaw angles need to be applied if the satellite is not aligned as indicated above.

After the appropriate pitch, roll, and yaw rotations have been applied to the initial telescope pointing vector (eqn [41]), eqn [40] yields the distance to the point being observed, and eqn [39] yields the coordinates of the point in the right ascension–declination coordinate system. Now, both the observed point and the satellite can be positioned by (1) rotating the vectors about the z -axis through the argument of latitude, (2) rotating about the x -axis through the inclination angle, and (3) rotating about the z -axis through the right ascension of ascending node less the right ascension of Greenwich. Equations [23a], [23b] and [23c] yield the latitude and longitude of the point. Finally, corrections may need to be applied for the nonspherical Earth and for the height of the terrain or the object being observed.

³ Note that the axes of rotation described here are dependent on the satellite being at the specified position and orientation in the right ascension–declination coordinate system.

The inverse problem, that of finding which satellite datum corresponds to a selected latitude and longitude, is solved iteratively. Each observation has a time associated with it, which determines all of the above angles. First, a time of observation is estimated, and the latitude and longitude of the point being observed at that time are calculated. Then the time is incremented and a new point is calculated. This process is iterated until a satisfactory solution is found.

Space–Time Sampling

To select an orbit for a satellite or a scan pattern for a particular instrument, several questions must be answered: What areas will the orbit and scan pattern allow the instrument to observe? How often will an area be observed? At what local times will the observations be made? At what viewing zenith and azimuth angles will the observations be made? What will be the solar zenith and azimuth angle when the area is being observed? These questions are all aspects of what is called space–time sampling. Using the equations in this article plus some easily acquired equations that describe the position of the Sun, these questions can be answered.

See also: Chemistry of the Atmosphere: Observations for Chemistry (Remote Sensing): IR/FIR (Satellite, Balloon and Ground); Observations for Chemistry (Remote Sensing): Microwave. **Oceanographic Topics:** Surface/Wind-Driven Circulation. **Satellites and Satellite Remote Sensing:** Aerosol Measurements; GPS Meteorology; Measuring Ozone from Space – TOMS and SBUV; Precipitation; Temperature Soundings; Water Vapor.

Further Reading

- Brouwer, D., Clemence, G.M., 1961. *Methods of Celestial Mechanics*. Academic Press, New York.
- Chen, H.S., 1985. *Space Remote Sensing Systems: An Introduction*. Academic Press, San Diego.
- Dubyago, A.D., 1961. *The Determination of Orbits*. Macmillan, New York.
- Escobal, P.R., 1965. *Methods of Orbit Determination*. Wiley, New York.
- Goldstein, H., 1950. *Classical Mechanics*. Addison-Wesley, Reading.
- Kidder, S.Q., Vonder Haar, T.H., 1995. *Satellite Meteorology: An Introduction*. Academic Press, San Diego.

Precipitation

G Liu, Florida State University, Tallahassee, FL, USA

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

This article explains the fundamental principle and retrieval methodology of measuring precipitation from satellites. Radiative signatures of precipitation are first examined at visible, infrared, and microwave spectra using radiative transfer theory with a real-world example. Techniques developed for precipitation retrievals are then introduced, including those using only a single wave spectrum and those combining multiple wave spectra from multiple satellites. Topics on measuring rainfall by space-borne radars and research on snowfall remote sensing are also introduced in light of the recent advancement in observing precipitation by active and high-frequency microwave sensors.

Introduction

Precipitation is one of the least well-measured atmospheric parameters, especially over the vast oceanic regions on the globe. There are two major obstacles that contribute to the lack of comprehensive global precipitation measurements. First, there are few surface-based observations over the oceanic areas, which cover about two-thirds of the Earth's surface. Second, precipitation is highly variable in both time and space compared to atmospheric variables such as temperature and pressure. Rainfall measured by a rain gauge at a given location can be significantly different from that measured just a couple of hundred meters away; similarly, rainfall measured at a given time can be significantly different from that measured just minutes earlier or later. As a result, a high-density rain gauge network is required in order to reasonably measure rainfall over a given area if one attempts to derive the area rain total from rain gauge observations alone. Such a high-density rain gauge network is generally not available even in well-developed countries and regions. Although ground-based radars can provide better spatial and temporal coverage than rain gauges, well-calibrated radars are available only in limited land regions in developed countries. These limitations related to surface-based measurements make satellite remote sensing indispensable.

Satellite remote sensing of precipitation is based on the radiative intensities emitted or reflected by cloud and precipitating hydrometeors. For infrared and microwave wavelengths, the radiative intensity is often expressed in terms of brightness temperature, defined as the temperature that is required to match the measured intensity to the Planck blackbody function. Brightness temperature in the infrared spectrum often represents the physical temperature of the cloud top because most clouds are optically thick for infrared radiation. Microwave radiation, however, can penetrate through cloud and rain layers, and its intensity reflects the integrated contribution by all water drops and ice particles in the atmospheric column. In the visible spectrum, the measured radiative intensity is due to the reflection of sunlight by the cloud and surface features. The dependence on sunlight limits the utility of visible sensing to daylight hours. Although visible radiation has a deeper penetration than infrared radiation, the visible reflectivity still reflects only the top portion of clouds.

Methods for passive satellite remote sensing of precipitation may be divided into the following three categories, based on

whether the information received by a satellite reflects the physical properties near the cloud top or over the whole atmospheric column: (1) sensing by visible and infrared radiation; (2) sensing by microwave radiation; and (3) sensing by a combination of visible, infrared, and microwave radiations. Visible and infrared methods are physically indirect because precipitation is derived from the radiative properties near the cloud top. In comparison, microwave techniques use more direct information on the vertical distribution of hydrometeors. In this article, the precipitation signatures at different wave spectra are first discussed using an actual satellite observation. The principles of satellite remote sensing under each of the aforementioned categories are then explained, omitting specific details of any particular retrieval algorithm.

Active sensing by space-borne radars is another important development for satellite remote sensing of precipitation in recent years. The basic principle of radar sensing from space is essentially the same as that sensing from ground (i.e., the radar reflectivity–rainfall rate relationship). The topic of measuring precipitation by ground-based radars is covered more thoroughly in other places of this book; the explanation of the measuring methodology given in this article will then be brief. So far, there have been two space-borne radars that are capable measuring precipitation: the Tropical Rainfall Measuring Mission's Precipitation Radar instrument and CloudSat's Cloud Profiling Radar. The former was designed to measure rainfall in the tropics, and the latter, while designed for measuring cloud water, is also suitable to measure light precipitation and snowfall.

The Radiative Signatures of Precipitation

The radiative signatures of precipitation may be understood by examining the observations of a hurricane shown in **Figures 1 and 2**, which display data simultaneously collected by a number of sensors on the Tropical Rainfall Measuring Mission's satellite. **Figure 1** is a thermal infrared image over the southeastern Pacific Ocean. The image covers an area approximately 720 km wide by 3050 km long. **Figure 2** shows the observed radiative properties at the satellite nadir along line A–B (see **Figure 1**), which crosses the outer cloud band of the hurricane. The parameters shown in **Figure 2** include a space radar-derived rainfall rate cross-section, near-surface

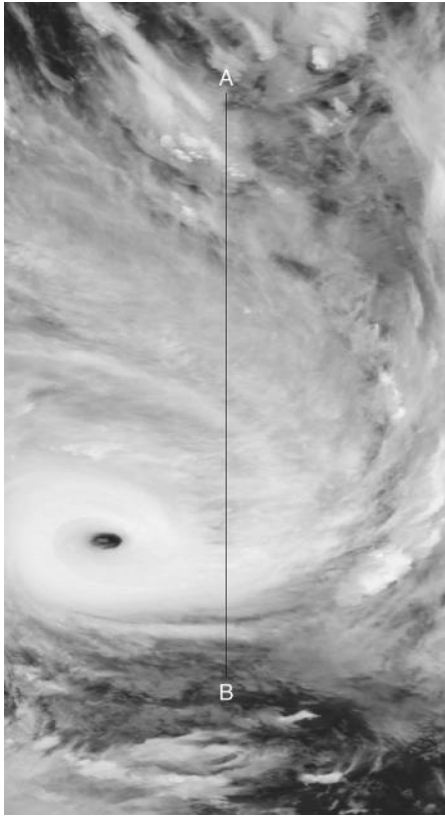


Figure 1 Thermal infrared image of a hurricane over the southeastern Pacific Ocean observed by an infrared radiometer on a Tropical Rainfall Measuring Mission satellite. The image covers an area approximately 720 km wide by 3050 km long on 5 January 1998.

rain (also derived from the space radar; because of surface contamination for space radar measurements under 1 km, the mean rainfall rate between 1 and 2 km above sea level is used here), visible reflectivity at $0.63 \mu\text{m}$, and brightness temperatures at infrared ($11 \mu\text{m}$) and microwave (19 and 85 GHz, horizontal polarization) wavelengths. The radar data indicate that the major rainfall area is on the right side in Figure 2, which corresponds to the deep clouds to the right of the hurricane's eye. Compared to those in the cloud-free area near point B, radiometric properties for the rainy areas show the following features: high visible reflectivities, low infrared brightness temperatures, high brightness temperatures at 19 GHz, and low brightness temperatures at 85 GHz. Most of the areas covered by the spiral cloud are actually not associated with rain, although clouds in those areas have low infrared brightness temperature and high visible reflectivity. It is the microwave brightness temperatures that most closely follow the radar-observed rainfall variation. The radiative properties shown here comprise the fundamental basis for satellite remote sensing of precipitation. They are explained theoretically in this section.

Radiative Transfer Equation

The radiative energy received by a satellite radiometer is either emitted (terrestrial radiation) or reflected (solar radiation) by the Earth-atmosphere system. Over 99% of the solar

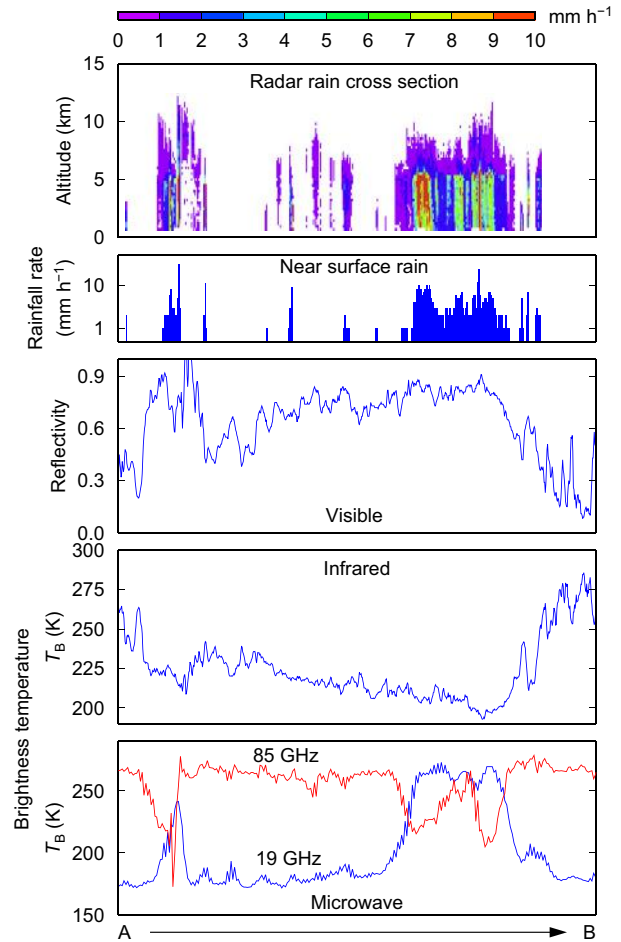


Figure 2 Radiative properties of hurricane clouds along line A–B shown in Figure 1, including a distance–height cross-section of the rainfall rate from radar, near-surface rainfall rate, visible reflectivity, and brightness temperatures at $11 \mu\text{m}$ (infrared) and 19 and 85 GHz (microwave).

(terrestrial) radiative energy is distributed in the spectra with wavelengths shorter (longer) than $4 \mu\text{m}$. Therefore, it is common that solar and terrestrial radiations are treated separately in radiative transfer models. For a plane-parallel atmosphere, the radiative transfer equation at a given wavelength can be expressed by eqn [1].

$$\mu \frac{dI(\tau, \mu)}{d\tau} = I(\tau, \mu) - \frac{\omega_0}{2} \int_{-1}^{+1} I(\tau, \mu') P(\mu, \mu') d\mu' - J(\tau) \quad [1]$$

In eqn [1], $J(\tau) = (1 - \omega_0)B(\tau)$ for terrestrial radiation, $J(\tau) = \frac{\omega_0}{4\pi} F_0 e^{-\tau/\mu_0} P(\mu, \mu_0)$ for solar radiation, τ is optical depth defined to be zero at the top of the atmosphere and increases with lowering altitude, μ and μ' are the cosine of the emergent and incident zenith angles of the wave, $I(\tau, \mu)$ denotes the radiance at optical depth level τ and emergent direction μ , ω_0 is the single-scatter albedo defined by the ratio of the scattering to extinction coefficients, μ_0 is the cosine of the solar zenith angle, P is the scattering phase function describing the angular distribution of scattered radiation, F_0 is the solar flux at the top of the atmosphere, and $B(\tau)$ is the Planck's function at

optical depth τ . Satellite-received radiances may be understood by solving radiative transfer equation [1].

Visible Reflectivity

In the visible spectrum, reflectivity increases with cloud optical depth, which is approximately proportional to liquid water path (vertically integrated liquid water) if the effective particle size remains constant. Therefore, clouds with higher values of liquid water path are more reflective. Generally, these clouds are also more likely to be associated with precipitation. This is the underlying principle for sensing precipitation by visible reflectivity. However, the sensitivity of visible reflectivity to liquid water path decreases with the increase of optical depth, and becomes virtually insensitive when optical depth is larger than 100, a value that a precipitating cloud easily exceeds. Consequently, instead of sensing the entire vertical column, reflected radiation at a visible wavelength represents the microphysical properties only near the top portion of a cloud. Therefore, the relation between the visible reflectivity and the rainfall rate at the surface is rather indirect.

Infrared Brightness Temperature

Cloud droplets are very absorptive in the thermal infrared spectrum. A consequence of this high absorption is that the cloud top may be viewed as the surface of a blackbody having a temperature the same as the air temperature at the level of the cloud top. Therefore, the infrared brightness temperature indicates the cloud top temperature except for in the case of optically thin cirrus clouds. At least in the tropics, most rainfalls are associated with well-developed convective systems that have tall cloud tops. Statistically, colder infrared brightness temperatures often indicate higher rainfall rates at the surface. However, shallow convections with warm cloud tops do sometimes produce substantial rainfall. On the other hand, nonprecipitating cirrus clouds do not produce rainfall, although they also have cold cloud top temperatures. Problems associated with the infrared sensing of rainfall are more serious in the midlatitudes, where most precipitation is produced by frontal stratiform clouds, for which cloud top temperature and precipitation are less correlated than for deep tropical convections.

Microwave Brightness Temperature

Microwave radiation received by satellite sensors measures the integrated radiative effects by the surface, atmospheric gases, and hydrometeors. Microwave brightness temperatures may either increase or decrease with increasing rainfall rate, depending on the frequency and the cloud microphysical properties. To understand the microwave signatures, consider an idealized rain cloud that contains raindrops below freezing level and ice particles above. Although accurate estimation of satellite-received radiation requires solving a radiative transfer model (eqn [1]) with absorption and multiple scattering, the primary radiative signature may be understood by examining the following approximation in eqn [2].

$$T_B \approx T_s [1 - e^{-2\tau_w} (1 - \varepsilon_s)] e^{-\tau_i} \quad [2]$$

In eqn [2], T_B is the brightness temperature received by the satellite; T_s is the surface temperature; τ_w and τ_i are the optical depths for the raindrops and ice particles, respectively; and ε_s is the surface emissivity. Note that radiance is proportional to brightness temperature at the microwave spectrum according to the Rayleigh-Jeans approximation. Therefore, brightness temperature T_B is used in eqn [2] to replace radiance I in eqn [1]. For simplicity, emission from atmospheric gases is not included in this equation, although its contribution is important, particularly near water vapor and oxygen-absorbing frequencies (e.g., 22 and 60 GHz). The emissivities for land and water surface are roughly 1 and 0.5, respectively, for microwave frequencies commonly used for precipitation retrievals. Consider the two situations given in the 'Low-frequency (<20 GHz) microwave radiation' and 'High-frequency (>80 GHz) microwave radiation' subsections.

Low-frequency (<20 GHz) microwave radiation

If the frequency is sufficient low, scattering by ice particles aloft becomes negligible (i.e., $\tau_i \approx 0$). Equation [2] then becomes eqn [3].

$$\begin{aligned} T_B &\approx T_s && \text{over land} \\ T_B &\approx T_s [1 - e^{-2\tau_w} (1 - \varepsilon_s)] && \text{over ocean} \end{aligned} \quad [3]$$

It is seen that rainfall cannot be detected over land by low-frequency microwaves because of the high surface emissivity. Since ocean surface temperature and emissivity generally do not vary dramatically, the small spatial or short temporal-scale variation of brightness temperature in eqn [3] can be attributed to the change in optical depth of raindrops, τ_w , which is approximately proportional to integrated total rainwater amount. Because total rainwater is closely related to the surface rain, low-frequency microwave brightness temperature over the ocean provides a relatively direct representation of rainfall rate. This positive correlation between rainfall rate and brightness temperature is shown in Figure 2 for 19 GHz. Because the increase of brightness temperature is due to the emission by raindrops, the rainfall signature in low microwave frequencies is called an emission signature. Figure 3 depicts brightness temperatures calculated by a radiative transfer model for nadir viewing at 19 GHz for various assumed freezing level heights. Brightness temperature increases with rainfall rate until reaching a maximum that indicates the saturation of microwave radiation. If rainfall rate further increases beyond the saturation point, the brightness temperature starts to decrease. The saturation problem prevents higher rainfall rates from being retrieved using microwave emission signatures.

High-frequency (>80 GHz) microwave radiation

For high-frequency microwaves, scattering by ice particles aloft is no longer negligible; rather, it becomes the dominant signature of the rain cloud. The optical depth due to raindrops, τ_w , usually is so large at high frequencies that $e^{-2\tau_w} \approx 0$. Equation [2] then becomes eqn [4].

$$T_B \approx T_s e^{-\tau_i} \quad [4]$$

Brightness temperature decreases with increasing optical depth of ice particles. The low value of the imaginary part of the ice dielectric constant determines that scattering is the dominant process for the interaction between ice particles and

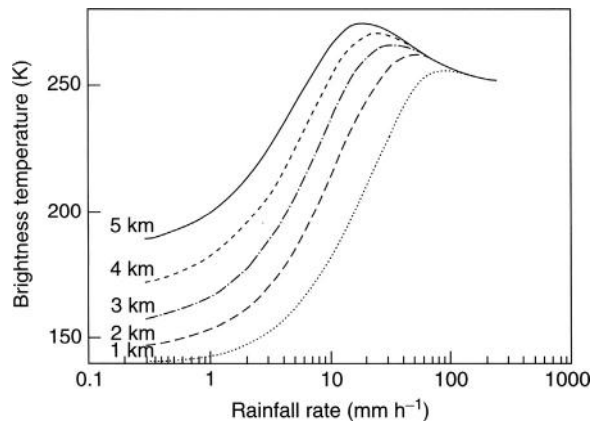


Figure 3 Brightness temperature for nadir viewing over an ocean surface at 19 GHz for various assumed freezing levels calculated by a radiative transfer model. Adapted with permission from Wilheit, T.T., 1986. Some comments on passive microwave measurement of rain. In: *Bulletin of the American Meteorological Society*, vol. 67. American Meteorological Society, Boston, pp. 1226–1232.

microwave radiation. Since a scattering cross-section is approximately proportional to the sixth power of the particle diameter, large and dense ice particles contribute the most to τ_i . Therefore, lower brightness temperatures at high microwave frequencies indicate more large ice particles aloft, which is commonly an indication of heavier rainfall rates at the surface. This relation is seen in [Figure 2](#) for 85 GHz microwave observations. Since the decrease of brightness temperature is caused by ice scattering, the rainfall signature at high microwave frequencies is called the scattering signature. Compared to the microwave emission signature, the scattering signature is a relatively indirect indication of surface rainfall. In [Figure 4](#), the model-simulated brightness temperatures at 92 GHz are shown for a 45° viewing angle. The brightness temperature falls more than 100 K for a rainfall rate increasing from 1 to 15 mm h⁻¹ given a 1 km ice layer. It must be cautioned that the magnitude of the brightness temperature depression due to ice scattering depends greatly upon ice particle size, shape, and density, for which there have not been sufficient observations available so far.

Sensing by Visible and Infrared Measurements

Since visible reflectivity and infrared brightness temperature are physically indirect indicators of surface rainfall, satellite retrieval techniques based on visible and infrared measurements are generally based on regression. That is, surface rainfall data, measured by rain gauges, radars, or both, are considered to be true values; colocated satellite-measured radiative properties are regressed against the true values to derive a statistical expression relating surface rainfall to satellite measurements. The true rainfall data are usually available for only limited regions and periods. As a result, most of these algorithms are subject to significant errors in regions where the climatological conditions are different. In addition, because of the statistical nature of these algorithms, retrieval accuracy generally increases with the increase of averaging area and time.

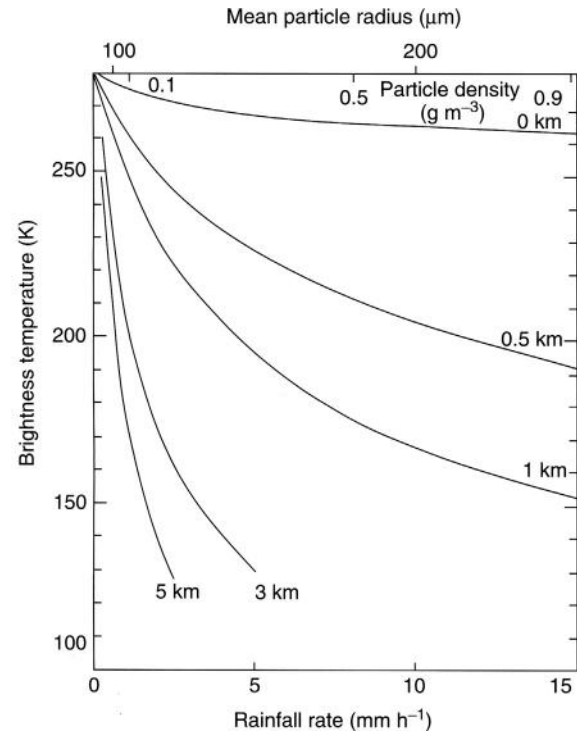


Figure 4 Brightness temperature calculations at 92 GHz (horizontal polarization, 45° view angle) for various thicknesses of the ice layer as a function of rainfall rate (lower abscissa). The two upper abscissae give the mean particle radius and the particle density corresponding to the rainfall rate through Marshall–Palmer size distribution. Adapted with permission from Wilheit, T.T., 1986. Some comments on passive microwave measurement of rain. In: *Bulletin of the American Meteorological Society*, vol. 67. American Meteorological Society, Boston, pp. 1226–1232.

Simple Regression

The notion that colder cloud tops usually correspond to heavier rainfall leads to the most straightforward approach: simply regressing rainfall rate against infrared radiation. This type of approach has mostly been done using broadband outgoing longwave radiation, instead of narrow-band brightness temperature within the atmospheric window (8–12 μm). A variety of equations relating rainfall rate to outgoing longwave radiation have been derived. Nonlinear functions are commonly used to account for the nonlinearity of the relation between the two parameters.

Area Time Integral Techniques

Studies of surface radar and rain gauge observations have shown that the total volume of rain falling over a sufficiently large area and for a long enough time period can be well predicted by the so-called area time integral (ATI) (eqn [5]).

$$ATI = \int A(X > X_{th}) dt \quad [5]$$

where X is the measured radiative property, $A(X > X_{th})$ is the area with X exceeding a threshold X_{th} , and the integration is over time t . The most utilized, infrared algorithm that uses this

principle is the GOES Precipitation Index (GPI), which gives the rain total over an area of 2.5° latitude by 2.5° longitude and for a time period of Δt (eqn [6]).

$$\text{GPI} = R_0 A (T_B < T_{B0}) \Delta t \quad [6]$$

$R_0 = 3 \text{ mm h}^{-1}$ and $T_{B0} = 235 \text{ K}$ are determined by comparing satellite measurements with ground radar observations over the tropical Atlantic Ocean. In essence, eqn [6] states that only the clouds with top temperatures colder than 235 K produce rain, and their average rainfall rate is 3 mm h^{-1} . This algorithm works reasonably well within the tropical belt of $30^\circ \text{ S} - 0^\circ \text{ N}$ for the monthly rain total. Errors increase dramatically toward high latitudes, particularly during cold seasons.

A number of other techniques have been similarly developed, but they take into account rain types and storm development stages. It is well established that deep convection more often produces heavy rainfall than stratiform clouds. The most common technique for determining cloud types relies on horizontal texture information of satellite images, such as identifying a local minimum in infrared brightness temperature imagery as the convective center and the surrounding relatively smooth portion as stratiform. If we divide observations over an area into several types, the rain total may be expressed by eqn [7], where R_i is the average rainfall rate for type i , which covers an area fraction of A_i and a time duration of Δt_i .

$$R = \sum_i R_i A_i \Delta t_i \quad [7]$$

This method is known as the 'cloud indexing' technique.

Observations also show that for thunderstorms, the rainfall rate peaks while its cloud area is growing rapidly, and rainfall is much reduced at the time of maximum cloud area. In techniques that include cloud life history, a different average rainfall rate R_i will be assigned for different development stages. This method is known as the 'life-history' technique.

Bispectral Techniques

Both infrared and visible measurements have important deficiencies in detecting rainfall. For example, stratus clouds are highly reflective but do not rain as much, or as often, as cumulonimbus clouds. Also, cirrus cloud tops are cold but do not produce rainfall. Bispectral techniques seek to combine information from visible and infrared measurements to obtain the optimal rainfall retrieval. In one such method, two lookup tables are first generated using coincident satellite and ground truth data. One of the tables is the probability of rain p_{ij} , determined by the ratio of raining cases to all cases in the infrared brightness temperature bin i and visible reflectivity bin j . Another table gives the mean rainfall rate r_{ij} , derived only from raining cases in the same two-dimensional bin. The variation of the rain possibility and mean rainfall rate in the reflectivity–brightness temperature space are schematically shown in Figure 5. For a given pixel whose brightness temperature falls in the i th and whose reflectivity falls in the j th bin, rainfall rate may be determined by eqn [8].

$$R = p_{ij} r_{ij} \quad [8]$$

Most bispectral methods attempt to retrieve instantaneous rainfall rate by constantly updating the lookup tables using

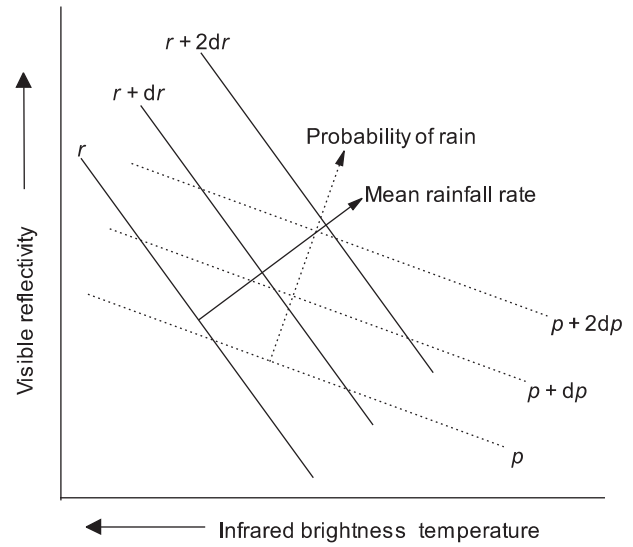


Figure 5 Schematic illustration of the probability of rain (p) and mean rainfall rate (r) in a two-dimensional diagram of visible reflectivity and infrared brightness temperature. dp and dr are positive increments.

radar, rain gauge, or even satellite microwave observations as truth. At least in theory, the bispectral methods should be superior to infrared-only or visible-only methods. However, this superiority has not been convincingly demonstrated, partially because visible data are available for only a fraction of the day, and partially because many other uncertainties still remain, such as the quality of truth data and the collocation of satellite and surface data.

Sensing by Microwave Measurements

Owing to its physical directness, microwave sensing of precipitation has received particular attention since the late 1970s. Except for only a few pure regression-type algorithms, a characteristic of the microwave methods is that they rely on radiative transfer models either at the algorithm-developing stage or during the retrieval computation. Through a radiative transfer model, microwave brightness temperatures are directly connected to the amount and distribution of precipitating hydrometeors. The microwave algorithms may be grouped into the categories of emission-based techniques, scattering-based techniques, techniques using both emission and scattering, and techniques based on the a priori radiation–rain database.

Emission-Based Techniques

The emission signature provides the most direct physical relation between rainfall and brightness temperature. Data from frequencies under 20 GHz are primarily used for this type of algorithm, although higher frequencies are sometimes included to minimize atmospheric water vapor and/or surface effects. The relation between brightness temperature and rainfall rate may be derived from radiative transfer model calculations by specifying the atmospheric temperature and humidity profiles, cloud liquid water content, rain layer thickness, and size

distribution of raindrops. The most commonly used raindrop size distribution is the so-called Marshall–Palmer distribution, in which the number concentration decreases exponentially with drop size. Rain layer depth may be assumed to be freezing level height, although the validity of this assumption deserves further investigation. There are several problems associated with emission-based algorithms: (1) They may be applied only over oceans; the high land surface emissivity prevents emission signatures from being detected by a low-frequency microwave radiometer. (2) Brightness temperature saturates for heavy rains. This problem is particularly serious for tropical regions where the rain layer is deep. (3) A nonuniform rainfall rate across the beam causes underestimation of the rainfall rate. As shown in Figure 3, the brightness temperature versus rainfall rate relation is highly nonlinear. The spatial resolution of a satellite pixel for microwave radiometers is on the order of several tens of kilometers; the rain field within one satellite pixel is generally inhomogeneous. If $R = R(T_B)$ is the theoretical relation between brightness temperature T_B and rainfall rate R for a homogeneous rain field, the retrieval resulting from the field-of-view-averaged brightness temperature $R(\bar{T}_B)$ does not equal the field-of-view-averaged rainfall rate $\bar{R}(T_B)$. Instead, in the case of microwave emission, it is always true that $R(\bar{T}_B) < \bar{R}(T_B)$ (i.e., the technique underestimates rain rate).

Scattering-Based Techniques

The scattering signature is physically less directly related to precipitation than the emission signature because it is an indication of the ice amount above freezing level. Frequencies higher than 80 GHz are primarily used for scattering-based algorithms. Algorithms have been developed either based on statistically regressing brightness temperatures to surface rainfall measurements or using results of radiative transfer models. The advantage of scattering-based algorithms is that they can be applied over both ocean and land. However, this type of algorithm has greater errors for regions where warm rains have a significant contribution.

Techniques Using Both Emission and Scattering

For rain associated with a shallow rain layer, scattering-based techniques fail to work because of lack of ice scattering. For heavy rainfall with deep rain layers, emission-based techniques cannot correctly determine rainfall rate because brightness temperature saturates. A better solution is to take advantage of both emission and scattering signatures by combining them in a single algorithm. One such algorithm is a ‘microwave index’ defined for the US Defense Meteorological Satellite Program’s Special Sensor Microwave/Imager according to eqn [9].

$$MWI = (1 - D/D_0) + 2(1 - PCT/PCT_0) \quad [9]$$

Here, $D = T_{B19V} - T_{B19H}$ is the depolarization at 19 GHz, and $PCT = 1.818T_{B85V} - 0.818T_{B85H}$ is the polarization-corrected brightness temperature at 85 GHz. In brightness temperature T_{Bvp} , the subscript v depicts frequency and the subscript p depicts polarization (V for vertical and H for horizontal). D_0 and PCT_0 are the rain threshold values for D and PCT . The first term in eqn [9] is the emission signature, and the second term is the scattering signature. Because D decreases monotonically

with the increase of rainfall rate, it represents the emission signature better than 19 GHz brightness temperatures themselves. Figure 6 depicts the 19 and 85 GHz brightness temperatures and microwave index for a viewing angle of 53° . The results are calculated from a radiative transfer model assuming that a typical tropical profile of hydrometeors for deep convection agreed. The microwave index relates to rainfall rate monotonically without saturation. An alternative way to combine the two signatures is to use the emission signature until brightness temperature at low frequency saturates, then to use the scattering signature at higher rainfall rates.

Techniques Based on the A Priori Radiation–Rain Database

Knowing the surface emissivity and vertical distributions of atmospheric temperature, humidity, and hydrometeors, brightness temperatures at satellite-observing frequencies can be calculated by a radiative transfer model. However, inversely, from brightness temperatures observed at several frequencies, multiple values of surface rainfall rate can be realized since several different combinations of the surface conditions and the atmospheric and hydrometeor distributions may result in the same upwelling radiation emerging from the top of the atmosphere. To solve this nonuniqueness problem in the retrieval, a commonly used technique is to apply a priori knowledge on rain profiles as a constraint, so that the retrieval algorithm searches solutions only within the ‘likely range.’ This technique generally consists of the following retrieval procedures. First, a large database of vertical profiles of hydrometeors and their corresponding brightness temperatures at satellite-observing frequencies must be prepared. The variety and the frequency of occurrence of rain profiles in the database should represent what occur in nature. This database may be prepared using actual observations, numerical cloud and radiative transfer model simulations, or a combination of both. However, because of the lack of observational data, this database has usually been constructed with simulated results from numerical cloud and radiative transfer models. In performing the retrieval, the set of brightness temperatures that most closely matches the satellite observed is selected, and the hydrometeor profiles corresponding to the best match are determined to be the retrieval. The retrieval produces not only rainfall rate at the surface but also its vertical distribution.

One implementation of this technique is based on the Bayes’ theorem, in which the posterior probability density function $p_{\text{post}}(\vec{R}|\vec{T}_B)$ of a rain profile $\vec{R}$ for a given set of measured brightness temperatures $\vec{T}_B$ can be expressed by eqn [10].

$$p_{\text{post}}(\vec{R}|\vec{T}_B) = \frac{p(\vec{T}_B|\vec{R})p_{\text{prior}}(\vec{R})}{\int_{\text{database}} p(\vec{T}_B|\vec{R})p_{\text{prior}}(\vec{R})d\vec{R}} \quad [10]$$

Here, $p_{\text{prior}}(\vec{R})$ is the prior probability density function of rain profile $\vec{R}$, and $p(\vec{T}_B|\vec{R})$ is the conditional probability density function of the set of brightness temperatures $\vec{T}_B$ given the rain profile $\vec{R}$. The integration in the denominator of eqn [10] is done over the database. Commonly, $p(\vec{T}_B|\vec{R})$ is assumed to be a multivariate normally distributed function with its maximum where the brightness temperatures in the database match satellite-observed values, and its standard deviation is proportional to the instrumental and radiative

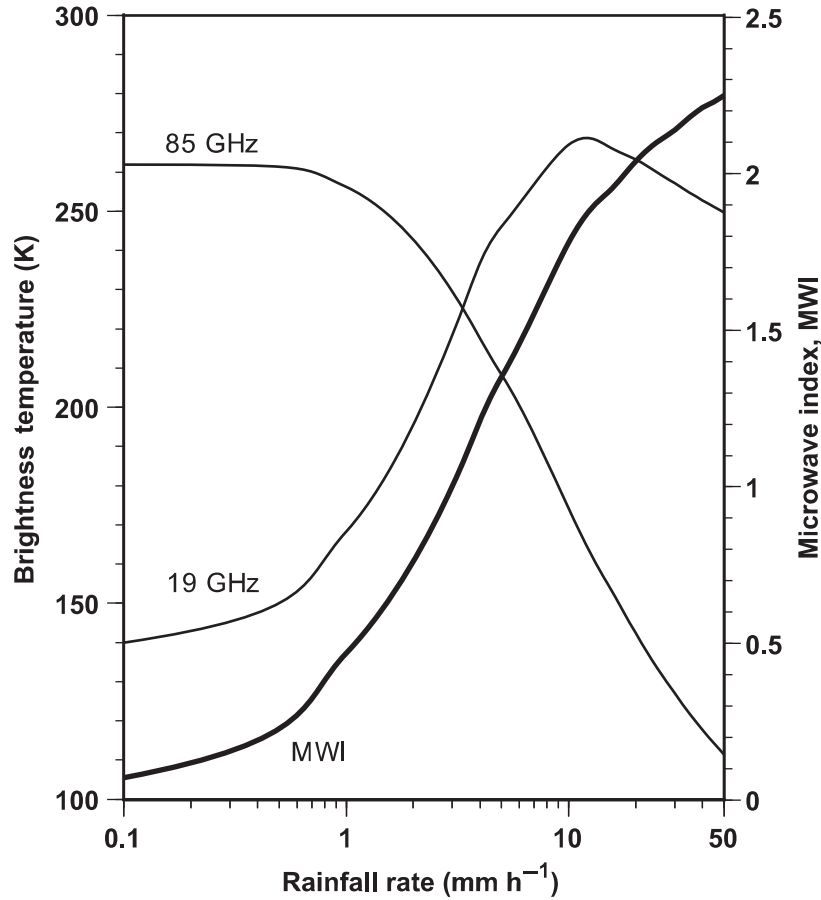


Figure 6 Radiative transfer model calculated brightness temperatures at 19 and 85 GHz, and microwave index (MWI) over ocean as a function of rainfall rate for tropical convective rains. A viewing of 53° is assumed, and the brightness temperatures shown are for horizontal polarization.

transfer model uncertainties. The posterior probability density function $p_{\text{post}}(\vec{R}|\vec{T}_B)$ gives the likelihood of a rain profile based on the observed brightness temperatures and the a priori knowledge. The retrieved rain profile $\hat{R}$ and its uncertainty σ_R^2 may be expressed by eqns [11] and [12], respectively.

$$\hat{R} = \int_{\text{database}} \vec{R} p_{\text{post}}(\vec{R}|\vec{T}_B) d\vec{R} \quad [11]$$

$$\sigma_R^2 = \int_{\text{database}} (\vec{R} - \hat{R})^2 p_{\text{post}}(\vec{R}|\vec{T}_B) d\vec{R} \quad [12]$$

The a priori database-based techniques have the advantage of fully using physical relations between cloud microphysics and microwave radiation. With more observational data becoming available in the future to build the database of hydrometeor profiles, this approach is expected to play a more significant role in satellite remote sensing of precipitation.

Combination of Multichannel and Multiplatforms

Currently, it is practical to put microwave radiometers only on low-altitude polar-orbiting satellites to ensure useful spatial resolution. The frequency of observation by a single satellite at

a certain area on Earth is unacceptably low (1–2 times a day) for determining rainfall accumulation. As a result, although microwave techniques work better for instantaneous rainfall rate retrieval, they do not outperform visible and infrared techniques on daily or monthly time scales, because visible and infrared measurements are more frequent. Therefore, combining measurements from multiple wave bands and multiple platforms has been proposed. One proposed approach is to increase the number of satellites that carry microwave sensors, so that local sampling frequency will be increased to an acceptable level. With increasing international collaboration, this proposal is expected to become reality in the near future, such as with the Global Precipitation Measurement project. The current solution has been to combine visible, infrared, and microwave measurements from available satellites. Visible and infrared measurements have the advantage of ample coverage, while microwave measurements have the advantage of physical directness. The combined techniques use microwave retrievals as truth to constantly train visible and infrared algorithms, while the trained visible and infrared algorithms are used to fill the gap left by microwave measurements. Figure 7 shows the global distribution and the zonally averaged values of rainfall rate derived by the Global Precipitation Climatology Project using 30 years (1979–2008) of multichannel multisatellite data (surface rain gauge data are also included). High rainfall rates

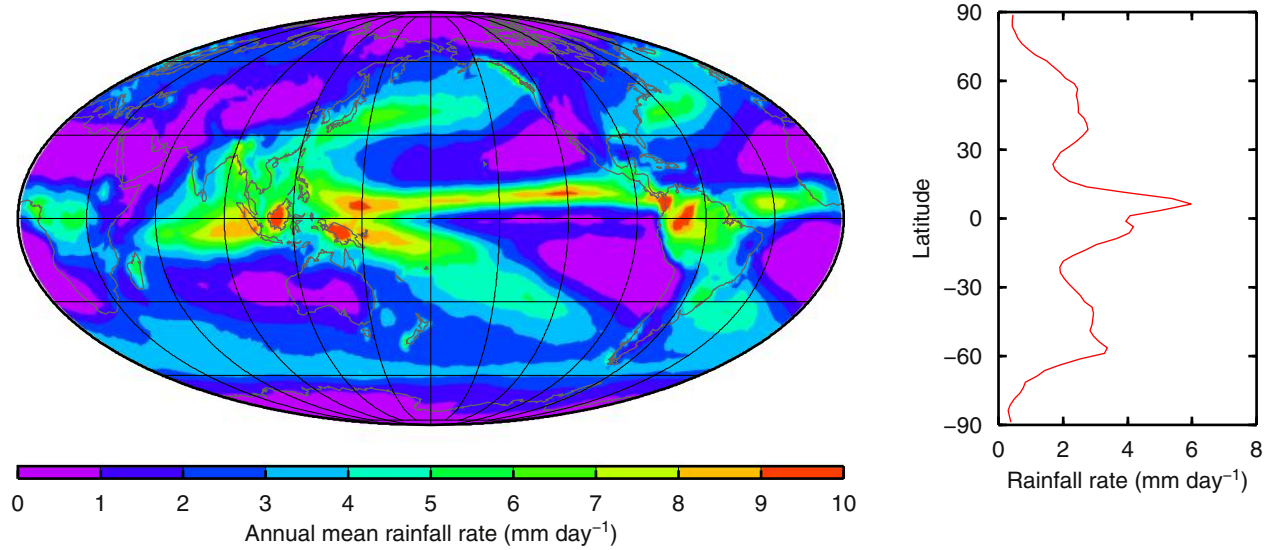


Figure 7 Global rain map (left) and its zonal distribution (right) derived from 30 years (1979–2008) of Global Precipitation Climatology Project (GPCP) data.

are observed in the intertropical convergence zone and along the midlatitude storm tracks. The zonal mean rainfall rate in the intertropical convergence zone is about 6 mm day^{-1} , or over 2 m rain total for a year. The secondary maxima of the zonal mean rainfall rate occur in the midlatitudes in both hemispheres, with a magnitude of about half of that in the intertropical convergence zone.

Sensing by Space-Borne Radars

Active remote sensing by space-borne radars provides the three-dimensional structure of precipitation. Tropical Rainfall Measuring Mission Precipitation Radar, operating at 13.8 GHz ($\sim 2.2 \text{ cm}$ wavelength), is the first precipitation radar instrument onboard a satellite. An example of this radar's observation of hurricane rainfall has been shown in Figure 2.

In contrast to radiometers that passively receive energy emitted or reflected by the Earth's surface and atmospheric constituents, radars actively transmit radiative energy, then receive the portion of energy scattered back. The backscattering intensity is represented by radar reflectivity factor Z , as defined by eqn [13].

$$Z = \frac{\lambda^4}{\pi^5 |K|^2} \int_0^\infty n(D) \sigma_b(D) dD \quad [13]$$

Here, λ is wavelength, K is a parameter determined by the refractive index of the scattering particles, $n(D)$ is the particle size distribution, and $\sigma_b(D)$ is the backscattering cross-section of particle diameter D . If the particle diameter is much smaller than the wavelength (Rayleigh-scattering regime), the backscattering cross-section will be proportional to D^6 , and radar reflectivity factor Z may be simplified to eqn [14].

$$Z = \int_0^\infty n(D) D^6 dD \quad [14]$$

Rainfall rate R , on the other hand, is the water volume flux defined by eqn [15].

$$R = \frac{\pi}{6} \int_0^\infty n(D) D^3 v(D) dD \quad [15]$$

$v(D)$ is the terminal velocity and is approximately related to drop size by $D^{0.5}$ for water drops. Comparing eqns [14] and [15], it is found that Z and R are approximately the sixth and 3.5th moment of the particle size distribution, respectively. Therefore, the Z - R relationship is not fixed but varies with the particle size distribution, which is the major factor contributing to the uncertainty in radar-derived rainfall rate. In order to reduce the uncertainty, dual-wavelength radar measurements have been designed for future Global Precipitation Measurement missions, in which at least one of the radar wavelengths must be short enough not to be in the Rayleigh-scattering regime, so that its radar reflectivity factor does not follow the relation given in eqn [14]. Two pieces of information can be obtained by the dual-wavelength radar measurements: the amount of rainwater and the mean size of raindrops. Rainfall rate retrieval algorithms can then be better constrained by using both the rainwater amount and drop size information.

Satellite Remote Sensing of Snowfall

At high latitudes during cold seasons, precipitation often takes the form of snowfall, which generally is more difficult to retrieve from satellite observations. A number of factors contribute to these difficulties. First, due to their low density, scattering by snowflakes is usually weak, which hampers snowfall detectability by both radars and radiometers. Second, for passive microwave observations, the scattering signal due to snow accumulation on ground is hardly distinguishable from that of falling snow. Third, unlike water drops, snowflakes are nonspherical; their scattering properties have not been well understood.

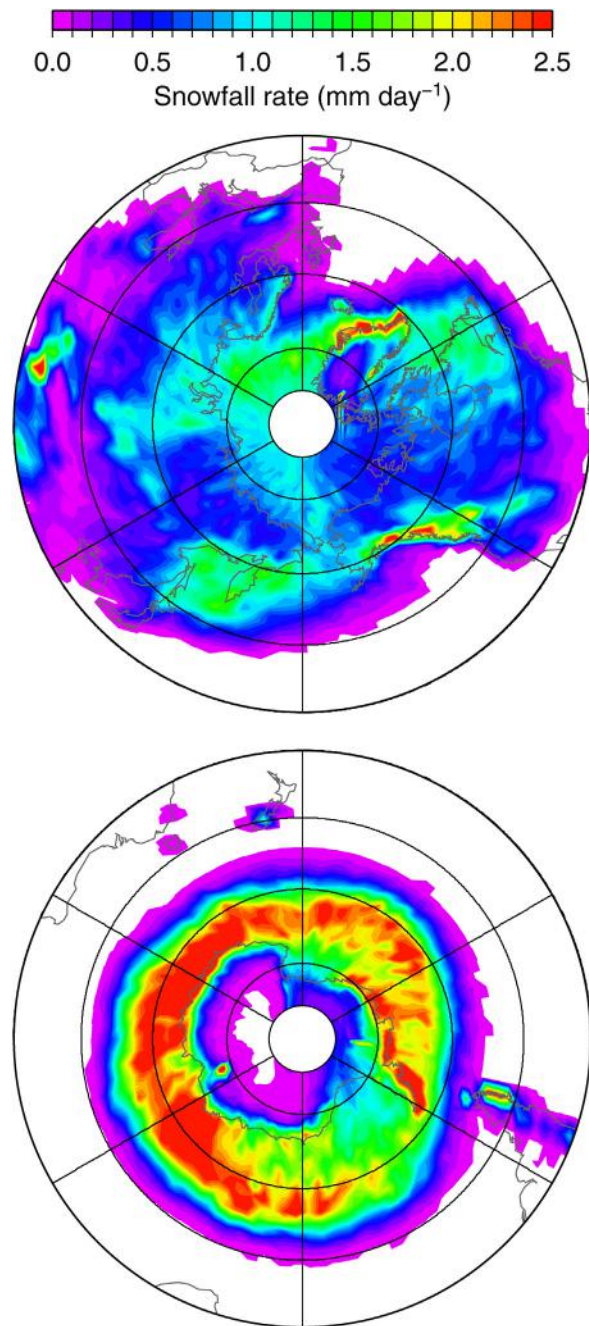


Figure 8 Snowfall maps derived from 4.5 years (June 2006–December 2010) of CloudSat radar data. Top: 30° N–90° N; bottom: 30° S–90° S.

Active and passive techniques for satellite remote sensing of snowfall have been developed using current microwave observations, although both techniques are still in their infancy. High-frequency microwave observations from 89 to 183 GHz have been used in the passive sensing technique. Similar to retrieving rainfall over land, the reduction of upwelling brightness temperature due to ice scattering is the primary signature for snowfall rate retrieval. An active technique for sensing snowfall has been developed using CloudSat's Cloud Profiling Radar data. This radar operates at 94 GHz (3.2 mm wavelength) and has a minimum detectable

radar reflectivity of -26 dBZ. It was originally designed for measuring liquid and ice water content in clouds. However, the low attenuation due to snowflakes together with the radar's high detectability make it suitable for measuring snowfall as well. A snowfall map derived from 4.5 years (June 2006–December 2010) of CloudSat data is shown in [Figure 8](#). In deriving this map, radar reflectivity data at 1 km above the surface have been used since data below 1 km altitude are often contaminated by surface returns. The snowfall map displays the following features. In the Southern Hemisphere, there is an almost zonally orientated high snowfall zone centered at approximately 65° S. In the Northern Hemisphere, however, heavy snowfall is mostly locked to geographical locations related to storm tracks. Additionally, the zonally averaged snowfall rate reaches about 2 mm day^{-1} , which is about one-third of the zonally averaged rainfall values found in the tropics, signifying the importance of snowfall in the hydrological cycle.

See also: Radar: Cloud Radar; Precipitation Radar. **Radiation Transfer in the Atmosphere:** Absorption and Thermal Emission; Scattering. **Satellites and Satellite Remote Sensing:** Research.

Further Reading

- Adler, R.F., Huffman, G.J., Chang, A., Ferraro, R., Xie, P., Janowiak, J., Rudolf, B., Schneider, U., Curtis, S., Bolvin, D., Gruber, A., Susskind, J., Arkin, P., 2003. The Version 2 Global Precipitation Climatology Project (GPCP) Monthly Precipitation Analysis (1979–Present). *Journal of Hydrometeorology* vol. 4. American Meteorological Society, Boston, pp. 1147–1167.
- Arkin, P.A., Ardanuy, P.E., 1989. Estimating climatic-scale precipitation from space: a review. *J. Clim.* vol. 2. American Meteorological Society, Boston, pp. 1229–1238.
- Barrett, E.C., Martin, D.W., 1981. *The Use of Satellite Data in Rainfall Monitoring*. Academic Press, London. p. 340.
- Grody, N.C., 1993. Remote sensing of the atmosphere from satellites using microwave radiometry. In: Janssen, M.A. (Ed.), *Atmospheric Remote Sensing by Microwave Radiometry*. John Wiley & Sons, Inc., New York, pp. 259–314.
- Kidder, S.Q., Vonder Haar, T.H., 1995. *Satellite Meteorology*. Academic Press, London. p. 466.
- Kummerow, C., Simpson, J., Thiele, O., Barnes, W., Chang, A.T.C., Stocker, E., Adler, R.F., Hou, A., Kakar, R., Wentz, F., Ashcroft, P., Kozu, T., Hong, Y., Okamoto, K., Iguchi, T., Kuroiwa, H., Im, E., Haddad, Z., Huffman, G., Krishnamurti, T., Ferrier, B., Olson, W.S., Zipser, E., Smith, E.A., Wilheit, T.T., North, G., Nakamura, K., 2000. The status of the tropical rainfall measuring mission (TRMM) after two years in orbit. *Journal of Applied Meteorology* 39. American Meteorological Society, Boston, pp. 1965–1982.
- Liu, G., 2008. Deriving snow cloud characteristics from CloudSat observations. *Journal of Geophysical Research* 113. D00A09 <http://dx.doi.org/10.1029/2007JD009766>
- Smith, E.A., Kummerow, C., Mugnai, A., 1994. The emergence of inversion-type precipitation profile algorithms for estimation of precipitation from satellite microwave measurements. In: *Remote Sensing Review*, vol. 11. Harwood Academic Publishers, Switzerland. pp. 211–242.
- Stephens, G.L., 1994. *Remote Sensing of the Lower Atmosphere: An Introduction*. Oxford University Press, New York. p. 523.
- Stephens, G.L., Vane, D.G., Boain, R.J., Mace, G.G., Sassen, K., Wang, Z., Illingworth, A.J., O'Connor, E.J., Rossow, W.B., Durden, S.L., Miller, S.D., Austin, R.T., Benedetti, A., Mitrescu, C., and the CloudSat Science Team, 2002. The CloudSat mission and the A-TRAIN: a new dimension to space-based observations of clouds and precipitation. *Bulletin of the American Meteorological Society* vol. 83. American Meteorological Society Boston, pp. 1771–1790.
- Wilheit, T.T., 1986. Some comments on passive microwave measurement of rain. In: *Bulletin of the American Meteorological Society*, vol. 67. American Meteorological Society, Boston. pp. 1226–1232.

Remote Sensing: Cloud Properties

P Yang, Texas A&M University, College Station, TX, USA

BA Baum, University of Wisconsin–Madison, Madison, WI, USA

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

Clouds constitute a unique and important component of the atmosphere. This article briefly reviews the methods of inferring cloud-top height, determining cloud thermodynamic phase, and retrieving cloud microphysical and optical properties (specifically, the effective particle size and optical thickness). Some examples based on observations made by a passive spaceborne sensor (the Moderate Resolution Imaging Spectroradiometer) and an active spaceborne sensor (the Cloud–Aerosol Lidar with Orthogonal Polarization) are illustrated.

Introduction

On any given day, clouds cover about 65% of the planet. In a fairly stable atmosphere, clouds may be cellular in appearance (i.e., cumuliform) or may appear in sheets (i.e., stratiform) that may extend over large horizontal distances. While these clouds may extend over wide areas, their typical geometric thickness is less than 1 km. In unstable atmospheres, clouds may extend from near the planet's surface to the upper troposphere. As most of the tropospheric water vapor resides near the surface, where temperatures tend to be relatively warm, low-level clouds tend to be composed of water droplets and are generally opaque to the viewer. The opacity is denoted in terms of a quantity known as optical thickness, or optical depth, and is a dimensionless measure of light attenuation caused by the scattering and absorption of energy by atmospheric particles. Clouds forming near the tropopause reside at very cold temperatures and are typically composed of ice particles. For clouds at intermediate heights between the planetary boundary layer (~ 1 km above the surface) and the middle troposphere, clouds may be composed of a mixture of supercooled water and ice particles. Water and ice clouds interact with solar radiation differently and have a large influence on the Earth's radiative energy budget. The energy budget is composed of both solar and terrestrial radiation components. Solar radiation spans from ultraviolet ($\lambda < 0.4$ μm , where λ is the wavelength) to infrared (IR) wavelengths ($\lambda > 5$ μm). A portion of the incoming solar radiation may be absorbed at the surface and within the atmosphere by clouds, aerosols, water vapor, and other trace gases such as carbon dioxide and methane. Subsequently, absorbed solar radiation is reemitted at longer wavelengths ranging from 5 to 100 μm .

Data from operational polar-orbiting and geostationary meteorological satellites are analyzed routinely for global cloud macrophysical properties such as cloud height, phase (water, ice, or some mixture of both), and microphysical and optical properties such as optical thickness and the effective particle size. Global cloud observations based on satellite measurements serve many uses. In numerical weather models, where the time scale of interest is on the order of hours to days, satellite-derived cloud and clear-sky properties from the geostationary satellites can serve as initial conditions for the models, that is, where the clouds are at some given time, their

height, and other properties. Numerical weather models may be regional in extent, covering a specific area such as North America, or global, in which case global and near real-time clouds and clear-sky properties are required for initialization of the models. Monthly, annual, or decadal averages of satellite-derived cloud properties are also useful for comparing with results from global climate models where the time scale of interest is much longer than for weather prediction models. For this type of use, cloud properties need to be collected, analyzed, and ultimately reduced to a global-gridded and time-interpolated product. An example of such a product would be one where each of the cloud properties retrieved during the course of a month is reduced to a monthly average with a time resolution of every 3–6 h. One of the primary issues in building a decadal climatology based on satellite observations is that the satellite sensor calibration needs to be very accurate. Since the advent of meteorological satellites, beginning around 1980, a long line of weather satellites have come into or out of service. Once in space, the platforms are subject to very harsh environments that can modify the sensor calibration over time, and for polar-orbiting platforms, the orbit can degrade over time. The derivation of a decadal record of cloud properties requires constant attention to sensor calibration.

To date, meteorological satellites have recorded information over the Earth at a limited number of wavelengths through the use of specially designed filter radiometers. The filters only allow radiation over a very narrow wavelength range to pass through to the detectors. Such narrowband wavelengths are typically chosen in atmospheric 'windows,' where the atmospheric constituents such as water vapor and carbon dioxide least attenuate the energy along the path to/from the surface, through the atmosphere, and finally to the satellite. At a minimum, operational satellite data are recorded at a visible (VIS) wavelength (e.g., 0.65 μm), a medium-wave-infrared (MWIR) wavelength (3.82 μm), and an IR wavelength (11 μm). Radiances at VIS and near-IR wavelengths are often converted to reflectances whose values range from 0 to 1. IR radiances are often converted to brightness temperatures (BTs) through application of the Planck function. Because of the huge volumes of data collected by satellites, the data reduction effort can become quite complex. In this article, we will discuss some of the available methods to infer cloud properties such as

cloud-top pressure, phase, optical thickness, and the effective particle size.

Cloud-Top Pressure–Height–Temperature

Over the past several decades, a number of approaches have been developed to infer cloud-top heights from satellite multispectral data. Actually, the literature provides a wealth of different research-grade algorithms, but very few have been fully developed and adopted for routine, operational processing of global data. For operational data processing, the assumption is made that only a single cloud layer is present in any individual field of view (FOV). Both surface observations and spaceborne lidar or radar measurements indicate that multilayered clouds occur frequently. If the uppermost cloud layer is optically thick, then a passive satellite sensor cannot sense the presence of lower level cloud layers. If, however, the upper cloud layer is optically thin, such as cirrus, then there is some potential for the presence of a lower level cloud layer to modify the radiances observed by the satellite sensor, causing errors in the assessment of the cloud properties for that FOV.

Another assumption generally made when inferring the cloud height is that there is a well-defined cloud-top boundary. For low-level water clouds, such as stratocumulus or cumulus, the cloud-top boundary is well defined. For high-level clouds, such as cirrus, this assumption is more problematic as the cirrus layer can be geometrically thick but with very sparse ice particles throughout the layer, which is another way of saying the cloud is optically thin.

The clouds that require the most attention in operational retrievals are those that reside either near the tropopause (high-level clouds) or near the surface. Some low-level clouds occur in atmospheres with temperature inversions. Proper placement of cloud-top heights requires that there be some knowledge of the atmospheric temperature profile, and numerical models are somewhat deficient on this in many cases. Given the many assumptions that need to be made, e.g., that an FOV contains only a single-layered cloud, is not optically thin at the top of the cloud layer, and that the temperature profile contains no surprises, there are some general approaches to inferring cloud height that are in use.

On many satellite platforms, measurements are obtained at wavelengths located in the 15- μm wavelength region, a region in which atmospheric transmission is dominated by atmospheric CO_2 . As the wavelength increases from 13.3 to 15 μm , the atmosphere becomes more opaque due to CO_2 absorption, thereby causing each channel to be sensitive to a different portion of the atmosphere. This sensitivity is demonstrated in Figure 1, which shows weighting functions at several Moderate Resolution Imaging Spectroradiometer (MODIS) channels located at wavelengths ranging from 12 to 14 μm . Each channel has a peak in its weighting function that occurs at a different pressure level than the other channels. The 12- μm channel is shown for comparison – note that its weighting function peaks at the surface. This is a ‘window’ channel that is insensitive to CO_2 . In the 1970s, Moustafa Chahine, William Smith Sr., and Martin Platt developed a technique known as CO_2 slicing to infer cloud-top pressure from radiances measured at wavelengths between 13.3 and 14.2 μm . In principle, the CO_2 slicing method is based on the

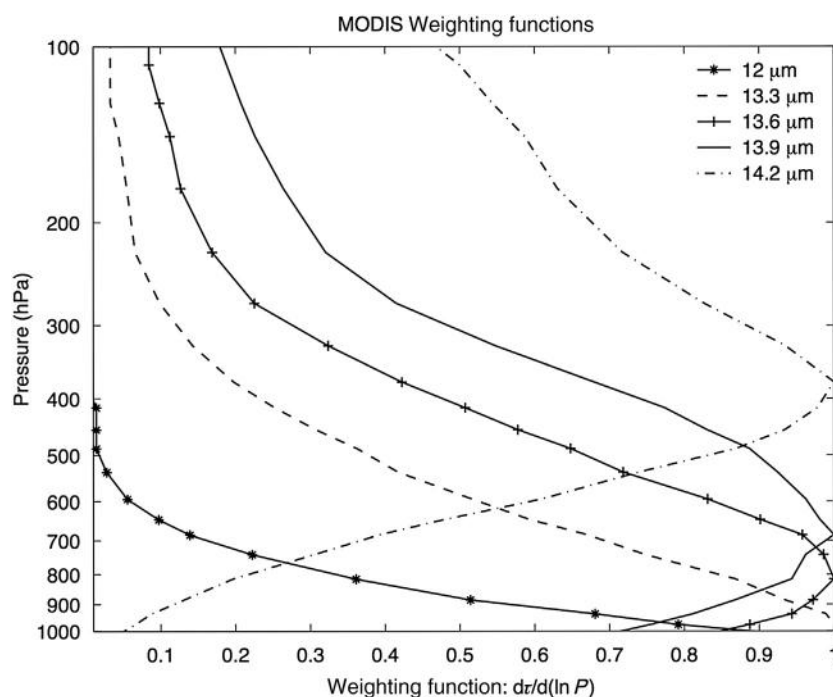


Figure 1 Weighting functions that are derived for MODIS wavelengths ranging from 12 to 14.2 μm . The weighting function is the derivative of the transmittance profile as a function of pressure. The peak in the weighting function provides an indication of what levels in the atmosphere provide most of the upwelling radiance that will be measured by a satellite.

following relation derived from the theory of radiative transfer:

$$\frac{R(v_1) - R_{\text{clear}}(v_1)}{R(v_2) - R_{\text{clear}}(v_2)} = \frac{N'_1 \int_{P_0}^{P_c} \Gamma(P, v_1) \{ \partial B[T(P), v_1] / \partial P \} dP}{N'_2 \int_{P_0}^{P_c} \Gamma(P, v_2) \{ \partial B[T(P), v_2] / \partial P \} dP}, \quad [1]$$

where $R(v_1)$ and $R(v_2)$ are the radiances measured at two channels centered at wave numbers v_1 and v_2 , whereas $R_{\text{clear}}(v_1)$ and $R_{\text{clear}}(v_2)$ are the corresponding clear-sky radiances. The terms Γ , T , and P indicate the transmissivity, temperature, and pressure, respectively. P_0 and P_c indicate the pressure values at the surface and cloud top, respectively. N' denotes the effective cloud amount that is the product of the cloud fraction and the cloud emissivity. If the two channels are selected to be sufficiently close in wave number, the corresponding effective cloud amount values are approximately the same. In this case, it is straightforward to find an appropriate value for the cloud-top pressure P_c by assuring the equality in eqn [1].

The pressure at cloud level is converted to cloud height and cloud temperature through the use of gridded meteorological products that provide temperature profiles at some nominal vertical resolution every 6 h. One benefit to this algorithm is that cloud properties are derived similarly for both daytime and nighttime conditions as the IR method is independent of solar illumination. This approach is very useful for the analysis of midlevel to high-level clouds and even optically thin clouds such as cirrus. The drawback to the use of the 15- μm channel is that the signal-to-noise ratio becomes small for clouds occurring in the lowest 3 km of the atmosphere, making retrievals problematic for low-level clouds. When low clouds are present, the 11- μm channel (also a window channel) is used to infer cloud height.

Cloud Thermodynamic Phase

While the cloud phase is extremely important in radiative transfer simulations of clouds and the retrieval of cloud properties, it is not always straightforward to determine a cloud's phase. If the cloud is located in the upper troposphere where the temperatures are extremely cold, it is assumed to be composed of ice. Conversely, if the cloud is located in the boundary layer over warm surfaces, it is assumed to be water. The difficulty lies in the inference of phase when the cloud-top temperature lies between 233 and 273 K. If the cloud temperature is below 233 K, the homogeneous nucleation temperature, it will be composed of ice. If the cloud temperature is above 273 K, it will be composed of water. If the cloud has a temperature between 233 and 273 K, it could be ice, water, or some mixture of both. In the high-latitude storm tracks in either hemisphere, large-scale stratiform cloud decks tend to form with cloud-top temperatures in the 250–265 K range, and cloud phase is quite difficult to discern.

At temperatures below 273 K, the supersaturation of ice is much higher than the supersaturation with respect to water. If water vapor is present in an atmospheric layer at a temperature in this range, say 260 K, and both water and ice particles are present in this layer, the water vapor will preferentially condense on the ice particles rather than the water particles. As the ice particles become larger, which occur over the course

of seconds to minutes, the growing ice particles will begin to fall through the cloud layer. In this situation, the top of the cloud layer tends to be populated primarily by very small water droplets, while ice particles fall through the cloud base. The cloud layer may contain both ice and water particles, so inference of the cloud phase from satellite data under these conditions is quite challenging.

Two simple approaches are discussed here to infer cloud phase from the radiometric observations made by a passive sensor. One method involves IR radiances measured at 8.5 and 11 μm . The radiances are converted to BTs through the Planck function, and the phase is inferred from the brightness temperature difference (BTD) between the 8.5 and 11 μm BTs (BTD[8.5–11]) as well as the 11 μm BT. Ice clouds exhibit positive BTD[8.5–11] values, whereas water clouds tend to exhibit highly negative values. There are three contributing factors to the behavior of the BTD[8.5–11] for ice and water clouds. First, the imaginary component of the index of refraction (m_i) differs for ice and water at these two wavelengths. Second, while the atmosphere is relatively transparent to gaseous absorption, absorption by water vapor in the atmospheric column above the cloud can still exert a considerable effect on the BTD values. As most of the atmospheric water vapor resides in the lower layers of the atmosphere near the surface, the BTD[8.5–11] values will be most affected in moist atmospheres rather than high-level clouds that reside above most of the water vapor. Third, while a small effect, cloud particles scatter radiation even at the IR wavelengths, and clouds with smaller particles will tend to scatter more radiation than those with larger particles. Multiple-scattering radiative transfer calculations show that for ice clouds, the BTD[8.5–11] values tend to be positive in sign, whereas for low-level water clouds, the BTD[8.5–11] values tend to be very negative (< -2 K).

This simple BTD approach with IR channels can be improved for optically thin ice cloud discrimination by calculating cloud emissivity ratios. In the simplest terms, the cloud emissivity for a channel is based on three numbers: the measured cloud radiance, the black cloud radiance, and the calculated clear-sky radiance. The term 'black' here means that the cloud radiates as a blackbody, which implies that it is opaque at the wavelength of the observation. This is more complicated than a simple BTD approach above because it requires the use of a radiative transfer model (RTM) to provide the clear-sky and black cloud radiances. However, what this approach provides is much more sensitive to optically thin ice clouds. The IR methods are not very useful when supercooled water clouds are present, however, since it is problematic to discriminate between water and ice as discussed previously.

One way to improve the discrimination between water and ice clouds is to analyze reflectances obtained at a VIS wavelength and a shortwave-infrared (SWIR) wavelength (e.g., 0.65 and 1.64 μm , respectively). At wavelengths less than about 0.7 μm , clouds composed of either liquid or ice tend to absorb very little solar radiation. However, at 1.64 μm (and 2.15 μm), the m_i values for both water and ice increase in comparison with those at the VIS wavelength and diverge, with m_i for ice being greater than the value of m_i for water. From this line of reasoning, one might expect that for two different clouds (one ice and one water) of similar particle size and habit (or particle

shape) distributions, the cloud reflectance at $0.65\ \mu\text{m}$ would not depend on thermodynamic phase, whereas the cloud reflectance at $1.64\ \mu\text{m}$ would. In theory and in practice, the $1.64\ \mu\text{m}$ (and $2.15\ \mu\text{m}$) reflectances are much lower for a cloud composed of ice than water particles.

The observations made by an active spaceborne sensor, for example, the Cloud–Aerosol Lidar with Orthogonal Polarization (CALIOP) on the Cloud–Aerosol Lidar and Infrared Pathfinder Satellite Observations (CALIPSO) platform can be used to effectively determine the cloud thermodynamic phase. The CALIOP 532-nm channel measurements offer polarization capabilities. Two quantities, the layer-integrated backscatter (γ') and the layer-integrated depolarization ratio (δ) can be employed to effectively discriminate cloud thermodynamic phase, which are defined as follows:

$$\gamma' = \int_{\text{cloud top}}^{\text{cloud base}} [\beta'_{\perp}(z) + \beta'_{\parallel}(z)] dz, \quad [2]$$

$$\delta = \frac{\int_{\text{cloud top}}^{\text{cloud base}} \beta'_{\perp}(z) dz}{\int_{\text{cloud top}}^{\text{cloud base}} \beta'_{\parallel}(z) dz}, \quad [3]$$

where $\beta'_{\perp}(z)$ and $\beta'_{\parallel}(z)$ indicate the vertical backscatter profiles associated with the perpendicular and parallel components, respectively. For a given cloudy scene, the γ' – δ relationship can be used to distinguish cloud phase. As illustrated in the right

panel of Figure 2 (the physical concept was originally developed by Yongxiang Hu at NASA Langley Research Center), water cloud pixels correspond to a γ' – δ relationship with a positive slope, whereas a γ' – δ relationship with a negative slope is related to ice cloud pixels. Furthermore, in the case of ice cloud pixels, the upper left branch of the γ' – δ curve corresponds to ice clouds containing horizontally oriented ice crystals, whereas the lower right branch of the γ' – δ curve is related to ice clouds composed of randomly oriented ice particles. The right panel of Figure 2 shows the frequency of occurrence of the γ' – δ relations of ice clouds based on the CALIOP data collected from July through December 2006.

Cloud Optical Thickness and Particle Size

The basic retrieval methodology for inferring the optical thickness and effective particle size is to (1) employ a RTM to develop a lookup table (LUT) for a wide range of assumed cloud properties and viewing geometries and subsequently (2) compare the measured radiances for selected wavelength channels to values in the LUT. The RTM requires a set of single-scattering properties for the cloud layer, which includes the single-scattering albedo, the scattering phase function, the scattering–absorption–extinction efficiencies, and the asymmetry factor. These parameters essentially determine how much incident radiation is reflected or absorbed by the cloud. The single-scattering albedo is defined as the ratio of the portion of energy scattered by a particle to the total extinction

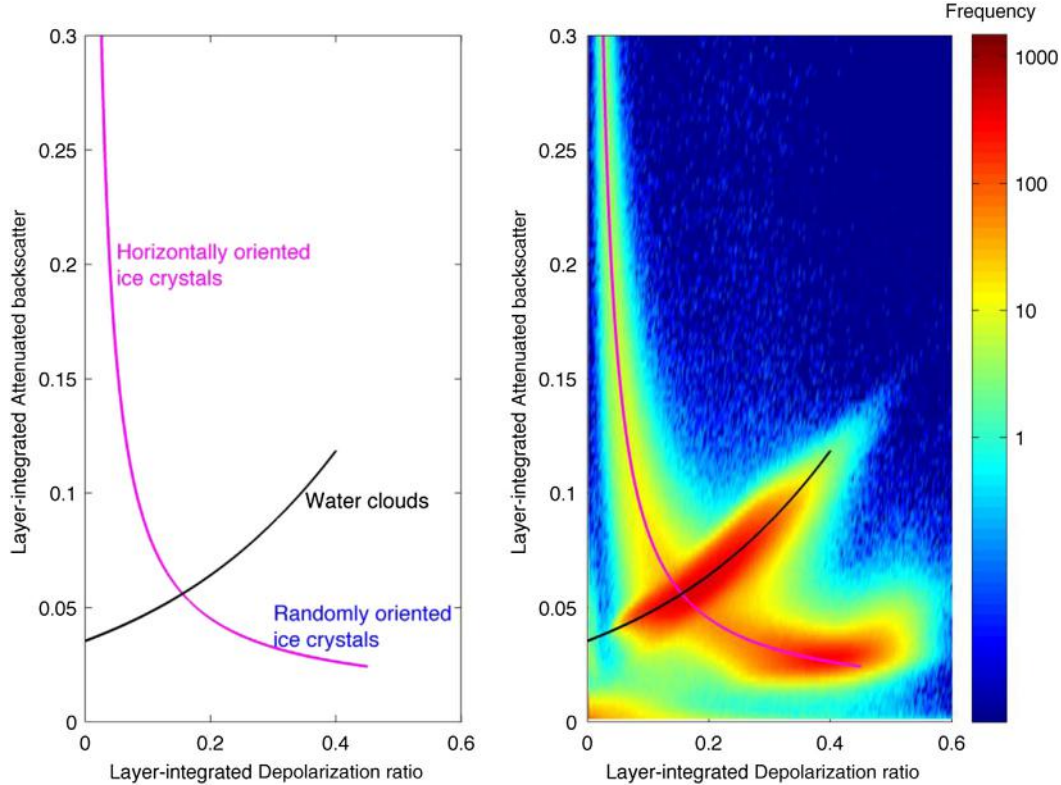


Figure 2 Left panel: schematic diagram showing the γ' – δ relationships for water and ice cloud pixels. Right panel: γ' – δ relationships based on the CALIOP measurements in the case when the lidar beam was pointed within 0.3° from the nadir.

(scattering + absorption) of energy by the particle. The phase function specifies the percentage of radiative energy that is not absorbed but is instead redistributed by the action of scattering by cloud particles when radiation impinges on clouds. The asymmetry factor describes the ratio of forward scattered to backscattered energy and is a quantity often used in radiative flux calculations. In practice, the single-scattering albedo and the asymmetry factor are parameterized in terms of analytical functions (normally polynomials) that depend on particle effective size for both water and ice clouds. In many RTMs, the radiative properties of clouds are described in terms of particle effective size and either liquid or ice water content (LWC or IWC), depending on the cloud phase. Cloud optical thickness and particle effective size are critically dependent on the accurate determination of the cloud bulk radiative properties, and a focus of recent research has been to improve the description of ice clouds in RTMs.

Various methods have been suggested to derive the optical thickness and effective particle size based on narrowband radiometer measurements by airborne- or satellite-based imagers. Operational methods tend to rely on IR bands or a combination of VIS and SWIR bands. The IR approach depends on the spectral information from thermal emission of clouds, whereas the VIS–SWIR approach is based on the reflection of solar radiation. Teruyuki Nakajima and Michael King were among the first to use reflected solar radiation to simultaneously retrieve cloud optical thickness and effective particle size for water clouds. The typical IR technique employs the BT or BTD values based on window channels at 8.5, 11, and 12 μm . Regardless of the detailed spectral information involved in these two methods, they are similar in that both depend on comparison of measured radiance data with simulated radiances derived for similar viewing and atmospheric conditions.

The first step in this process is to discuss the generation of reliable libraries of simulated cloud and clear-sky radiances. Single-scattering calculations must be carried out regarding how individual cloud particles interact with incident radiation. For water clouds, the liquid droplets can be well approximated as spheres for light scattering. The scattering properties of an individual liquid sphere can be calculated by using the well-known Lorenz-Mie theory that has been documented in many texts. James Hansen and Larry Travis have extensively discussed the effect of size distribution on single-scattering properties of spheres. Their work provides a theoretical framework for using and applying the bulk radiative properties of liquid droplet distributions which is briefly recaptured here.

Within a given water cloud, liquid water droplets span a range of sizes that may be represented mathematically in terms of the Gamma distribution, given by

$$n(r) = \frac{N_0 (r_{\text{eff}} V_{\text{eff}})^{(V_{\text{eff}}-1)/V_{\text{eff}}}}{\Gamma[(1-2V_{\text{eff}})/V_{\text{eff}}]} r^{(1-3V_{\text{eff}})/V_{\text{eff}}} \exp(-r/r_{\text{eff}} V_{\text{eff}}), \quad [4]$$

where N_0 is the total number of the droplets in a unit volume; r_{eff} and V_{eff} are the effective radius and effective variance that are defined, respectively, as follows:

$$r_{\text{eff}} = \frac{\int_{r_1}^{r_2} r^3 n(r) dr}{\int_{r_1}^{r_2} r^2 n(r) dr}, \quad [5]$$

$$V_{\text{eff}} = \frac{\int_{r_1}^{r_2} (r - r_{\text{eff}})^2 r^2 n(r) dr}{r_{\text{eff}}^2 \int_{r_1}^{r_2} r^2 n(r) dr}. \quad [6]$$

In a plot of the Gamma distribution, the peak of the distribution defines the r_{eff} , while V_{eff} affects the width of the distribution. Typical values of the effective variance for water clouds range from 0.05 to 0.1. For a given size distribution, the bulk-scattering properties of cloud droplets may be calculated. For example, the phase function averaged over a size distribution is given by

$$\langle P(\theta) \rangle = \frac{\int_{r_1}^{r_2} \sigma_s(r) P(\theta, r) n(r) dr}{\int_{r_1}^{r_2} \sigma_s(r) n(r) dr}, \quad [7]$$

where σ_s is scattering cross section of droplets and $P(\theta, r)$ is the phase function for droplets with radii of r , which describes the angular distribution of scattered radiation versus scattering angle θ .

Figure 3 shows the phase functions averaged for size distributions for water clouds at wavelengths 0.65, 1.63, and 11 μm . For the 0.65- μm wavelength, the phase function displays scattering maxima at 140 and 180°. Physically, the two maxima are due to mechanisms associated with the rainbow and glory, both characteristic features of Mie scattering. The phase functions at the SWIR wavelength (1.63 μm) are similar to those at 0.65 μm , but the rainbow and glory maxima are somewhat reduced by absorption within the particle. At the IR wavelength of 11 μm , the scattering maxima of the phase function are largely smoothed out due to absorption within the water droplets.

Another measure of the relative amounts of scattering versus absorption is provided by the single-scattering albedo. At 0.65 μm , the scattering of incident radiation by cloud droplets is conservative, meaning that energy may be scattered, but not absorbed, by the particles. Thus, the single-scattering albedo is unity at 0.65 μm but less than unity at 1.63 μm . The particle size also affects the single-scattering albedo at 1.63 μm . For example, for effective sizes 4 and 32 μm , the particle single-scattering albedo is unity at 0.65 μm , whereas the corresponding values at 1.63 μm are 0.9976 and 0.9824, respectively. Because of the difference in single-scattering albedo at the two wavelengths, reflection by an optically thick cloud at 0.65 μm is essentially a function of optical thickness. At 1.63 μm , however, cloud reflectance is sensitive to droplet effective size. This feature of cloud reflectance provides a mechanism to retrieve cloud optical thickness and particle sizes using two channels at VIS and SWIR wavelengths, as will be further explained later in this section.

Ice clouds are almost exclusively composed of nonspherical ice particles with various sizes and habits (i.e., shapes). Ice particles can consist of relatively simple shapes such as bullet rosettes, columns, and plates or more complex shapes such as aggregates of columns or plates. Most of the columnar particles can have hollow intrusions at the ends, which is caused by preferential molecular deposition onto a growing particle. In an environment where supercooled water droplets are present, the ice particles can also become rimed, which increases an individual particle's surface roughness. An increasing amount of research is showing that the consistency of inferred ice cloud properties improves between algorithms using solar, IR, or

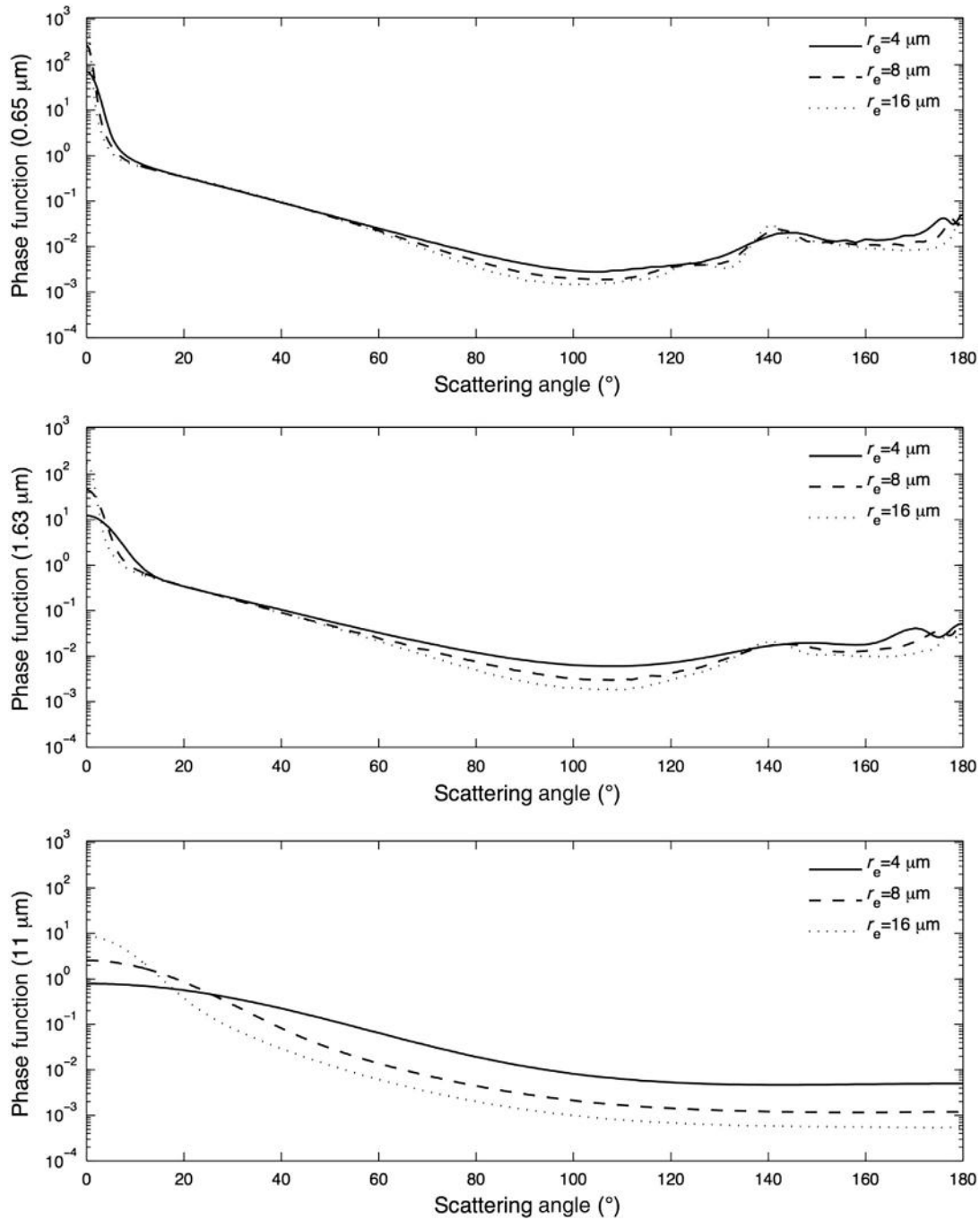


Figure 3 Scattering phase function of water droplets calculated at three wavelengths at 0.65, 1.63, and 11 μm for effective radii of 4, 8, and 16 μm .

polarized measurements if an assumption of ice particle severe surface roughening is adopted.

Research is underway to determine how to accurately calculate the single-scattering properties of a limited set of idealized ice habits. In practice, methods such as the discrete dipole approximation, finite-difference time domain technique, or the T-matrix method are used to calculate the scattering properties of a given habit for which the ratio of the particle circumference to the wavelength (also known as the size parameter) is small, i.e., less than 30. For ice particles

with larger size parameters, scattering calculations are performed using a ray-tracing technique based on the principles of geometric optics.

Figure 4 shows the phase matrices at 0.65- μm wavelength for two types of ice crystals: a solid column with smooth surfaces and aggregates of plates with rough surface. The phase function of smooth hexagonal columns displays a strong scattering peak at 22° and is produced by the hexagonal structure typical of ice crystals. In addition to the peak at 22° , the phase function of solid columns also displays a small peak

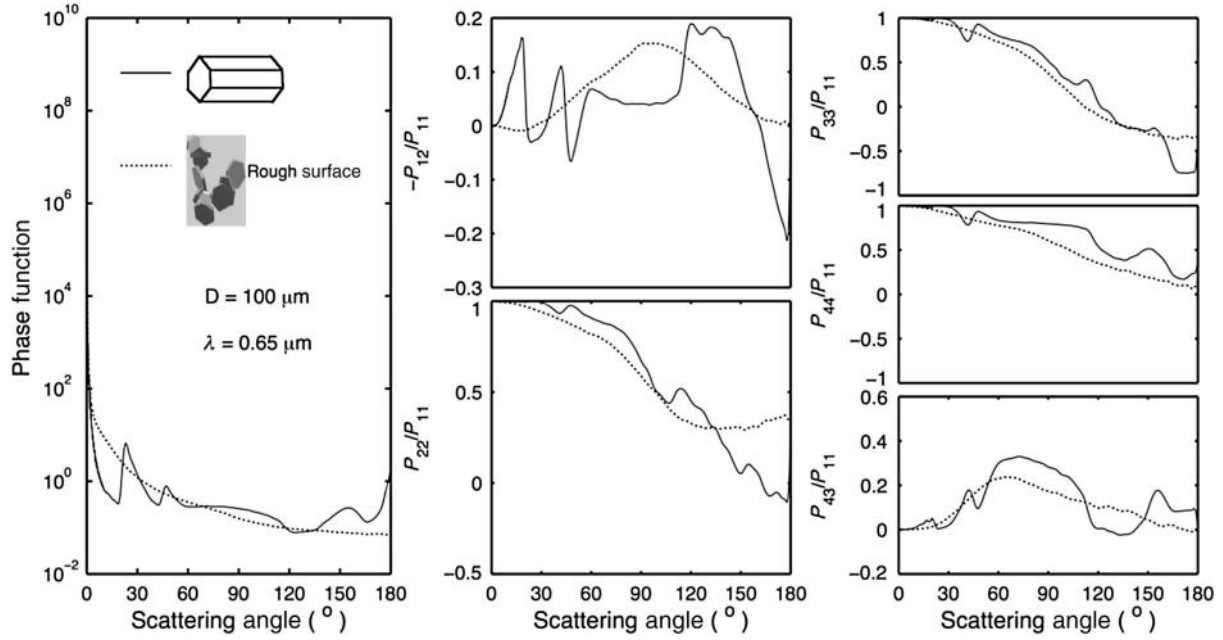


Figure 4 The scattering phase matrices of hexagonal ice crystals with smooth surface and aggregates of plates with rough surfaces.

corresponding to a 46° halo. Compared to the phase function for pristine crystal habits, the phase function for aggregates of plates is essentially featureless due to the severely roughened surface texture. The rougher the particle, the more featureless is the phase function. The other nonzero elements of the phase matrix are related to the polarization state of the scattered light. The impact of surface roughness on the polarization state is significant. Some recent studies have demonstrated that polarization measurements, for example, by the Polarization and Anisotropy of Reflectances for Atmospheric Sciences coupled with Observations from a Lidar (PARASOL) offer unprecedented capabilities to infer ice crystal habit and associated particle roughness. In particular, the comparison between the polarized reflectance observed by PARASOL and the relevant theoretical simulations illustrates that the closest match occurs when assuming the presence of ice crystals with severely roughened surfaces.

In reality, ice clouds are composed of many different crystal habits. To derive the bulk radiative properties of cirrus clouds, we need to consider not only a particle size distribution but also the percentages of the various particle habits that comprise the cloud. For this reason, the derivation of accurate radiative transfer simulations of ice clouds is considered more difficult than for water clouds. For a given size distribution, a number of definitions have been suggested for the effective size. If the effective size is defined as the ratio of total volume to total projected area, however, the bulk optical properties are insensitive to the detailed structure of the size distribution. The effective radius is then

$$r_{\text{eff}} = \frac{3 \int \sum_i f_i V_i(D) n(D) dD}{4 \int \sum_i f_i A_i(D) n(D) dD}, \quad [8]$$

where D is the maximum dimension of an ice particle, f_i is the habit fraction, V and A are the volume and projected area for

individual particle, and n is the particle number concentration. Based on *in situ* measurements within ice clouds, a modified gamma distribution is used most often to describe the particle size distribution.

In situ ice cloud measurements are now available from numerous field campaigns based at locations around the world. For example, Table 1 (data courtesy of Andrew Heymsfield, National Center for Atmospheric Research) lists a number of the particle size distributions obtained at various field campaigns and the instruments used for the microphysical measurements. This is by no means a complete list. A new generation of sensors is beginning to provide measurements of the smallest particles in a given particle population and even a sense of the particle roughening. *In situ* measurements indicate that the effective radius of ice crystals in cirrus clouds may range from about $5 \mu\text{m}$ (small ice particles near the tropopause) to more than $100 \mu\text{m}$ (deep convection). Larger particle radii might be expected for ice clouds formed in convective situations where the updraft velocity is much higher (m s^{-1}) than that found under conditions where optically thin cirrus tend to form (cm s^{-1}). The *in situ* measurements provide insight for the development of an appropriate ice cloud model in terms of the ice crystal habit and size distributions. As an example, the upper left panel of Figure 5 illustrates an ice model based on two habits (hexagonal columns and aggregates of plates) with surface roughness. The lower left panel of Figure 5 shows the comparisons of the computed medium mass diameter (where half the mass is in smaller particles and half in larger particles) versus *in situ* measurements, whereas the lower right panel shows the corresponding comparison for IWC. Apparently, the two-habit model can reasonably represent *in situ* microphysical measurements. The upper right panel of Figure 5 shows the phase function based on the two-habit model in comparison with the MODIS Collection 5 counterpart. Note that the asymmetry factors associated with the two

Table 1 Number of the ice particle size distributions obtained during various field campaigns and the instruments for the microphysical measurements

Field campaign	Year	Location	Probes
ARM-IOP	2000	Oklahoma, USA	2D-C, 2D-P, CPI
TRMM KWAJEX	1999	Kwajalein, Marshall Islands	2D-C, 2D-P, CPI
CRYSTAL-FACE	2004	Florida area, over ocean	CAPS, VIPS
Pre-AVE	2004	Houston, Texas	VIPS
MidCiX	2004	Oklahoma	CAPS, VIPS
ACTIVE-Hector	2005	Darwin	CAPS
ACTIVE-Monsoon	2005	Darwin	CAPS
ACTIVE-Squall Line	2005	Darwin	CAPS
SCOUT	2005	Darwin, Australia	FSSP, 2D-C
TC-4	2006	Costa Rica	CAPS, CPI
MPACE	2004	Alaska	2D-C, 2D-P, CPI

The data are filtered such that the *in situ* measurement occurs at a cloud temperature $T \leq 40^\circ\text{C}$.

Notes: (1) The table is from: http://www.ssec.wisc.edu/ice_models/microphysical_data.html. (2) The data sets currently include a total of 14 406 particle size distributions and the list will increase over time.

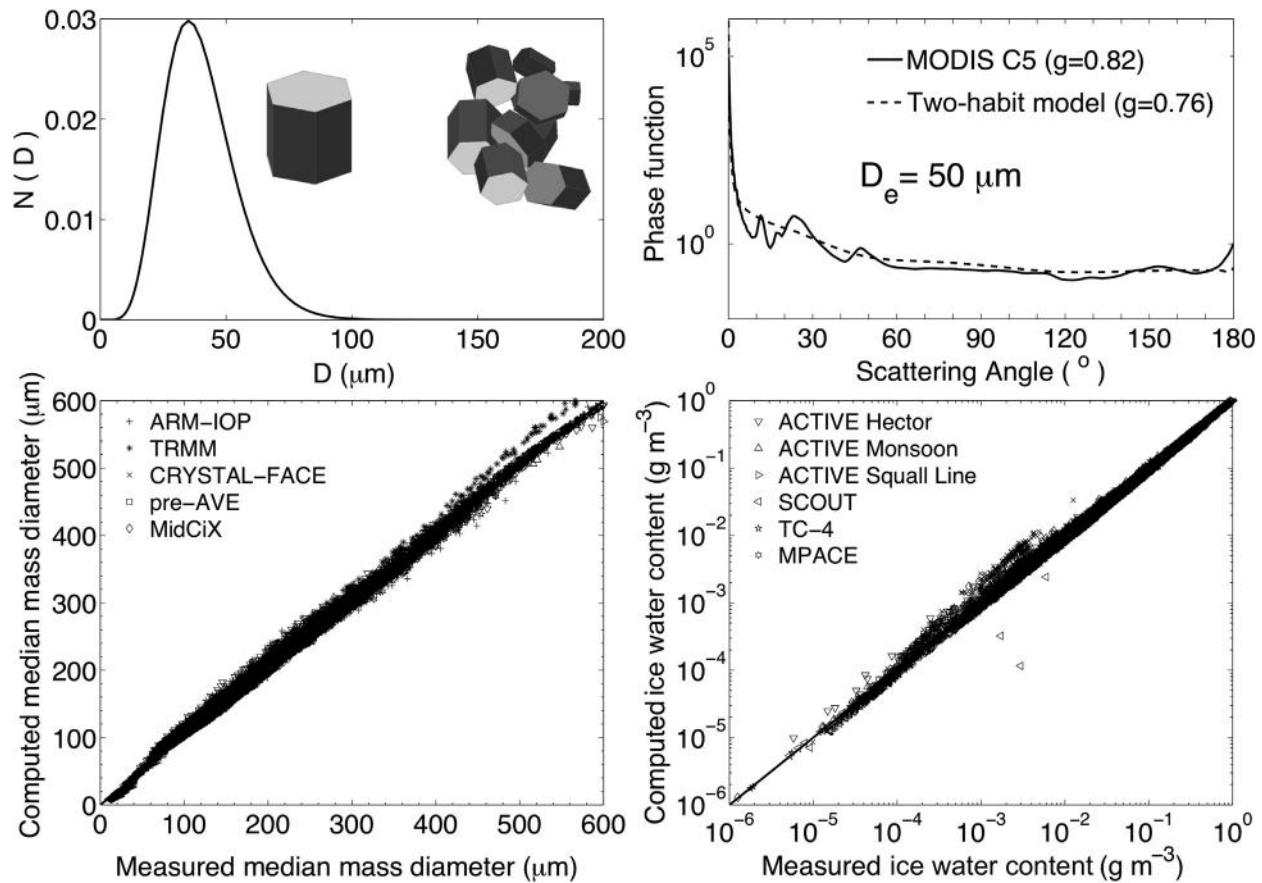


Figure 5 Upper left panel: a two-habit ice cloud model based on hexagonal columns and aggregate of plates in conjunction with the Gamma distribution. Lower left panel: comparison of the theoretical median mass diameter versus *in situ* measurements associated with the data sets listed in Table 1. Lower right panel: comparison of the theoretical IWC versus *in situ* measurements associated with the data sets listed in Table 1. Upper right panel: the phase function computed with the two-habit model in comparison with the MODIS Collection 5 phase function.

phase functions are quite different; particularly, the asymmetry factor for the two-habit model is approximately 0.76, whereas the MODIS Collection 5 counterpart is 0.82.

Given the single-scattering properties, radiative transfer computations can be carried out for various cloud optical

thickness and effective particle sizes for a number of solar and viewing geometry configurations. To calculate the bidirectional radiance of clouds, one can use well-established discrete ordinate or adding-doubling methods. Figure 6 shows the correlation of 2.13- μm reflectance and 0.86- μm reflectance of cirrus

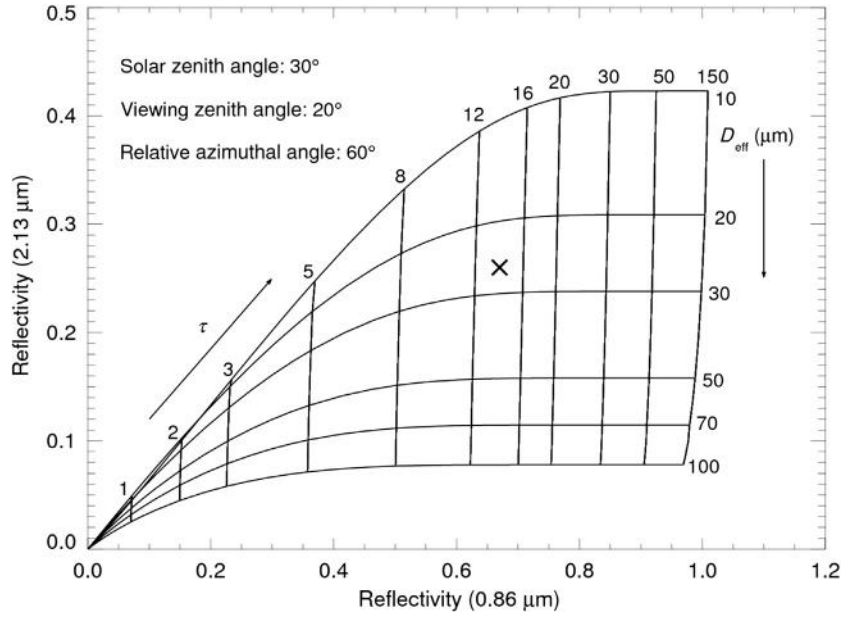


Figure 6 The theoretical relationship between the reflection function at 0.86 and 2.13 μm for various values of cloud optical thickness and effective particle size.

clouds for a number of optical thickness and effective sizes for a given incident-view geometry. At higher optical thicknesses (meaning the cloud is more opaque), there is a 'quasi-orthogonality' between the optical thickness and particle size curves. As we have mentioned previously, the cloud reflectance at 0.86 μm is primarily sensitive to cloud optical thickness, whereas the reflectance at 2.13 μm is sensitive to both the particle size and cloud optical thickness. This orthogonality forms the underlying principle for application of the two-channel correlation technique for retrieving cloud optical thickness and effective size. For example, assume the symbol 'X' in Figure 6 to represent the (0.86 and 2.13 μm) reflectivity pair. One may infer that the corresponding optical thickness is approximately 14, whereas the effective particle size is 25 μm . It should be pointed out that, in practice, the (0.86 and 2.13 μm) reflectivity combination is usually used for retrieval over ocean, the (0.64 and 2.13 μm) reflectivity combination is used for over land, and the (1.24 and 2.13 μm) reflectivity combination is used over snow or ice. In addition to the 2.13- μm channel, a channel located at 1.64 or 3.7 μm can be used as the SWIR or MWIR channel involved in the aforementioned bispectral method.

As an alternative or as a complement to the VIS-SWIR bispectral retrieval algorithm, IR channels in the window region (8–12 μm) may be used for retrieving cloud properties. The window region is an important part of the IR spectrum because terrestrial thermal emission peaks within this spectral region. IR-based methods are useful because a single approach may be used for both daytime and nighttime conditions, thereby simplifying the data reduction effort and also the comparison between daytime and nighttime cloud properties. IR methods are insensitive to sun glint over water that is often present in operational data. Interpretation of data over reflective surfaces is often performed more readily using IR methods rather than

those that involve VIS-SWIR wavelengths. The underlying principle for IR retrievals is based on the sensitivity of the BT or the cloud emissivity (related to blackbody or graybody emission) to cloud optical thickness and particle size. The BT is the temperature that, when applied to the calculation of Planck function for blackbody radiation, gives the same value as the satellite measured IR radiance. Figure 7 illustrates the sensitivity of the BT between the 11- and 12- μm channels as a function of the BT at the 11- μm channel for various cloud optical thickness and the effective particle size. Evidently, comparing the measurements of the BTD-BT relation with the theoretical computations permit simultaneous retrieval of cloud optical thickness and the effective particle size. However, the IR technique is more sensitive to the atmospheric profile (particularly, the temperature profile) and the surface emissivity than the VIS-SWIR technique.

In addition to the use of BDT and BT, a quantity known as the cloud emissivity has been widely used to infer cloud properties. In practice, the cloud emissivity can be calculated as follows:

$$\varepsilon(\lambda) = \frac{R(B) - R}{R(B) - R(C)}, \quad [9]$$

where R is the upwelling radiance at the cloud top, $R(B)$ is the upwelling radiance at the cloud bottom, and $R(C)$ is the upwelling blackbody radiance corresponding to the cloud temperature. In practice, for a given scene, the radiance at cloud base can be obtained by the noncloudy (i.e., clear sky) pixels.

Furthermore, the IR techniques for retrieving ice cloud properties are less sensitive than their VIS-SWIR counterparts to ice crystal habits assumed in the forward light-scattering and radiative transfer simulations. To illustrate this point, panels (a) and (b) of Figure 8 show the phase functions of two ice crystal habits (hexagonal columns and hollow bullet rosettes)

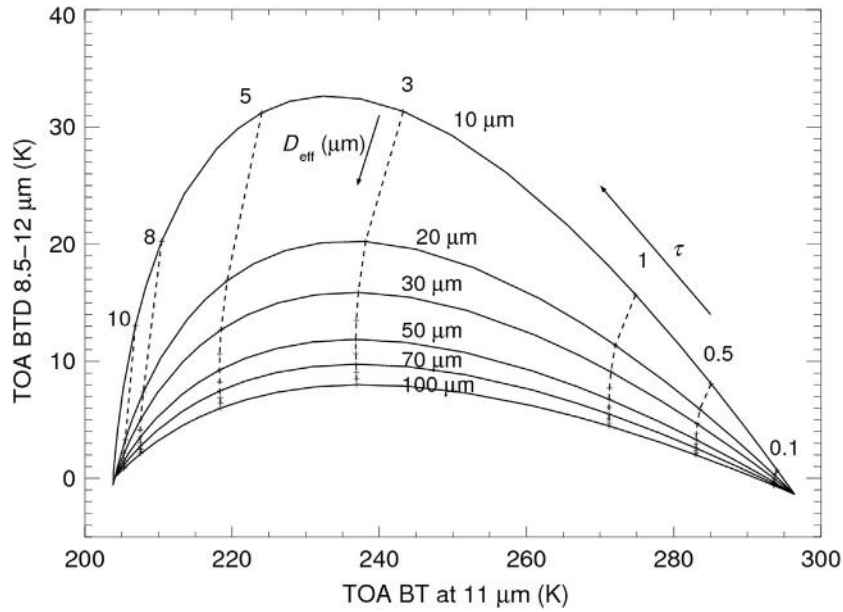


Figure 7 The variation of the BT between the 8.5- and 12- μm channels as a function of the BT at the 11- μm channel.

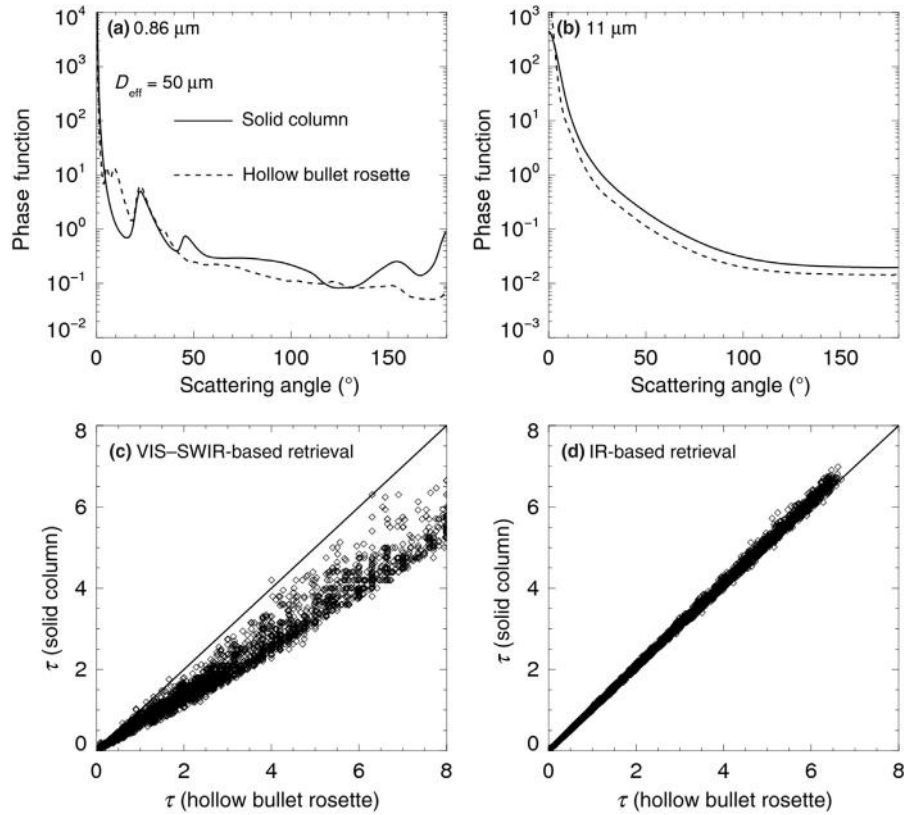


Figure 8 Panel (a): bulk phase functions of solid columns and hollow bullet rosette with an effective particle size of 50 μm at 0.86 μm . The gamma distribution is used to simulate the size distribution. Panel (b): similar to panel (a) except for a wavelength of 11 μm . Panel (c): comparison of cloud optical thickness retrievals based on the VIS-SWIR retrieval on the basis of a solid column habit model and a hollow bullet rosette habit model. Panel (d): similar to panel (c) except that an IR technique is used.

at wavelengths of 0.86 and 11 μm , respectively. Substantial differences are noticeable at the 0.86- μm near IR wavelength, whereas the two phase functions are quite similar at the 11- μm wavelength. Panel (c) of [Figure 8](#) compares the optical thickness values retrieved with the VIS–SWIR bispectral method based on MODIS Band 2 (0.86 μm) and Band 7 (2.13 μm) measurements. The impact of the assumed ice crystal habit on the retrieval is obvious from panel (c). The optical thickness values retrieved from the IR technique are shown in panel (d) based on the MODIS Band 29 (8.5 μm), Band 31 (11 μm), and Band 32 (12 μm) observations. The effect of ice crystal habit on the IR-based technique is negligible.

Cloud radiative and microphysical properties are cloud inherent properties that should be independent of a specific retrieval algorithm employed to infer the cloud properties. In this sense, the VIS–SWIR and IR retrievals should be consistent. The spectral consistency of cloud property retrievals is critical to some analyses, particularly, the study of the diurnal variations of cloud properties based on a VIS–SWIR algorithm for daytime and an IR algorithm for nighttime. In the case of ice cloud, recent studies have demonstrated that the ice cloud optical model involved in the forward radiative transfer simulation is essential for achieving spectral consistency.

Future Challenges in Cloud Property Retrieval

Current efforts to derive a global cloud climatology from satellite data generally do not account properly for multiple cloud layers in pixel-level imager data. To date, operational algorithms are designed to infer cloud properties for each imager pixel under the assumption that only one cloud layer is present. Climatologies of retrieved cloud properties do not address the effect of an optically thin upper cloud layer, such as cirrus, that may overlay a lower cloud layer such as a cumuli-form cloud deck. Surface observations show that clouds often occur in multiple layers simultaneously in a vertical column, i.e., cloud layers often overlap. Multiple cloud layers occur in about half of all cloud observations and are generally present in the vicinity of midlatitude fronts and in the tropics where cirrus anvils may spread hundreds of kilometers from the center of convective activity. When multilayered clouds are present, the retrieval algorithms will generally place the cloud layer at a height between the two (or more) actual layers present in the FOV. Currently, available satellite cloud climatologies provide a horizontal distribution of clouds but need improvement in the description of vertical distribution of clouds. At this point, a reliable method has not been developed for the retrieval of cloud properties (optical thickness, cloud thermodynamic phase, and effective particle size) when multilayered, overlapping clouds are present.

Even for a single-layered cloud, satellite retrieval algorithms do not account for the effect of a likely vertical variation of cloud microphysical properties, which in turn will decrease the ability of radiative transfer calculations to accurately simulate the cloud. It is unlikely that cloud particles are homogeneously distributed throughout any given cloud. For example, ice crystal size and habit are typically quite different for midlatitude cirrus at cloud top from at cloud base. A common assumption in satellite imager-based cirrus retrieval algorithms is that the radiative properties of a cirrus cloud may be

represented by those associated with a specific ice crystal shape (or habit) and a single particle size distribution. However, observations of synoptic cirrus clouds with low updraft velocities have shown that pristine small ice crystals with hexagonal shapes having an aspect ratio close to unity (length and width are approximately equal) are predominant in top layers. The middle layers of cirrus are often composed of well-defined columns and plates, while irregular polycrystals or aggregates are dominant near cloud base. This picture is quite different from ice particles that form in deep convection; in this case, the population of ice particles may be dominated by complex aggregates.

Another interesting area of complexity in satellite remote sensing is caused by mixed-phase clouds. Single-layered clouds composed of mixtures of supercooled water droplets and ice particles have been observed frequently during various field campaigns. Recent analyses of these data and MODIS satellite cloud property retrievals highlight the difficulty of ascertaining phase. If mixed-phase clouds are present in the data, one might expect larger errors in retrieved properties such as optical thickness and particle size than clouds that are primarily of a single phase. From the perspective of satellite remote sensing, the working assumption is that any imager pixel contains either ice or water but not a mixture. There is no rigorous method available for determining the single-scattering properties of mixed-phase clouds. From the microphysical cloud process perspective that is important for developing cloud model parameterizations, the presence of both ice particles and supercooled water droplets will affect cloud lifetime. Why? It is likely that the ice particles will grow much more quickly from vapor deposition than the water droplets as the environment may be supersaturated with respect to ice. The result of this process is that the ice particles will rime, grow quickly in size, and fall through the cloud, and the available water will be depleted quickly. The process of glaciation is very important for modelers because the water–ice conversion rates affect cloud lifetime. Details of cloud microphysics, such as cloud water amount, cloud ice amount, snow, graupel, and hail, are important for improving cloud retrieval.

While approaches exist to retrieve a variety of cloud properties from satellite imager data, it is not an easy problem to compare the satellite retrievals with ground-based measurements of the same cloud. Comparisons are often attempted between a surface-based measurement at a fixed location over a long temporal period and satellite measurements that provide an instantaneous measurement over a wide area. While difficult and often creative, confidence in retrievals is often gained through painstaking comparison between the two. For some cloud properties, it may be possible to compare properties derived from two or more different satellite instruments. This will be one of the more active areas in future research.

Acknowledgments

The authors are grateful to several individuals for their assistance in the preparation of the diagrams in this article, particularly, Lei Bi (for [Figure 4](#)), Chao Liu (for [Figure 5](#)), Chenxi Wang (for [Figures 6–8](#)), and Chen Zhou (for [Figure 2](#)).

See also: **Aerosols:** Aerosol–Cloud Interactions and Their Radiative Forcing. **Clouds and Fog:** Classification of Clouds; Climatology; Contrails; Measurement Techniques *In Situ*; **Lidar:** Backscatter. **Radiation Transfer in the Atmosphere:** Cloud-Radiative Processes; Scattering. **Satellites and Satellite Remote Sensing:** Research.

Mishchenko, M.I., Hovenier, J.W., Travis, L.D. (Eds.), 1999. Light Scattering by Nonspherical Particles: Theory, Measurements, and Geophysical Applications. Academic Press, San Diego.

Stephens, G.L., 1994. Remote Sensing of the Lower Atmosphere. Oxford University Press, Oxford.

Wendisch, M., Yang, P., 2012. Theory of Atmospheric Radiative Transfer – A Comprehensive Introduction. Wiley-VCH Verlag GmbH & Co., KGaA, Weinheim, Germany.

Further Reading

Kidder, S.Q., Vonder Haar, T.H., 1995. Satellite Meteorology: An Introduction. Academic Press.

Liou, K.N., 1992. Radiation and Cloud Processes in the Atmosphere. Oxford University Press, Oxford.

Research

MD King, University of Colorado, Boulder, CO, USA

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

Research satellites have been developed for measuring the properties of the Earth's atmosphere since 1960, with increasingly sophisticated instrumentation and attention to calibration and radiometric accuracy. Among the properties of the Earth's atmosphere that are routinely observed from space are (1) the Earth's radiant energy budget (the balance between the shortwave energy entering the Earth and the outgoing radiation leaving the Earth), (2) aerosol particles (dust, pollution, smoke), (3) cloud properties (frequency of occurrence, height, size of cloud particles), (4) chemical constituents (ozone, nitrogen dioxide, sulfur dioxide), (5) atmospheric temperature and water vapor, and (6) precipitation.

Introduction

The atmosphere changes chemically and physically on widely varying timescales – ranging from minutes to decades – and is therefore a challenge to measure precisely over the entire globe. But with the National Aeronautics and Space Administration's (NASA) 1960 launch of the Television Infrared Observation Satellite, Earth scientists began a new mission to observe large-scale weather patterns from space. In the late 1970s, their mission expanded to include global-scale measurements that would help them understand the causes and effects of longer term climate change. NASA and its affiliated agencies and research institutions collaborated to develop a series of research satellites that have enabled scientists to test new remote-sensing technologies that have advanced scientific understanding of both chemical and physical changes in the atmosphere. ('Remote sensing' involves the use of devices other than our eyes to observe or measure things from a distance without disturbing the intervening medium.) The goal is to examine our world comprehensively to determine what dynamics drive our planet's climate system and how climate change affects, and is affected by, our environment.

Depending on their measurement objectives, research satellites primarily fly in one of two orbits: (1) a near-polar, sun-synchronous orbit to allow their sensors to observe the entire globe at the same solar time each day, or (2) a mid-inclination, precessing orbit to focus their sensors on the equator and lower latitudes where the observations are made at different times of day to better sample time-varying phenomena such as clouds. Some polar-orbiting satellite sensors can observe any given place on the globe as often as every day, thus collecting data with high temporal (time) resolution. Other satellite sensors view any given place as infrequently as once every 16 days, thus having relatively low temporal resolution for a satellite sensor, but still far surpassing our ability to make these same measurements with surface-based or aircraft sensors. Satellite sensors with high spatial resolution (15 m per pixel) can discern objects in the atmosphere or on the surface as small as say a football field or farmland, thus providing high spatial resolution. Other satellite sensors that are designed to measure continental and global-scale dynamics typically have only moderate (500 m per pixel) to low (20 km per pixel) spatial resolution. Satellite sensors carry specially designed detectors that are particularly sensitive to certain wavelengths of the electromagnetic

spectrum, called spectral bands. The more precisely a remote sensor can measure narrow bands of radiant energy, and the greater the number of these discrete bands it can measure, the higher is its spectral resolution. The atmosphere interacts with solar radiation much like a venetian blind – selectively absorbing and reflecting certain wavelengths of solar energy while allowing others to pass through. Engineers design satellite remote sensors to be particularly sensitive to those wavelengths that can be reflected or emitted back up through the atmosphere to space, thus enabling them to make their measurements.

Earth-orbiting remote sensors provide the best means of collecting the data scientists need because they can measure things on scales of time and space that otherwise would not be possible. Moreover, satellite sensors not only observe wavelengths of visible light but also precisely measure wavelengths of radiant energy that our eyes cannot see, such as microwaves, ultraviolet rays, or infrared light. If scientists know how certain objects (like cirrus clouds or windblown dust) typically absorb, reflect, and emit particular wavelengths of radiant energy, then by using satellite sensors to precisely measure those specific bands of the electromagnetic spectrum, scientists can learn a lot about the Earth's atmosphere and surface. Remote sensors allow us to observe and quantify key climate and environmental vital signs such as temperature, ozone concentrations, carbon monoxide (CO) and other pollutants, water vapor and other greenhouse gases, cloud cover and their properties (height, optical thickness, phase, cloud particle size), aerosol concentrations, radiant energy fluxes, and many more.

Balancing Earth's Radiant Energy Budget

Climate is defined as the average state of the atmosphere, hydrosphere, and land over a given time period. Thus, measurements of radiant energy within the Earth's atmosphere are at the heart of the climate change discussion. How the climate changes is directly related to how our planet balances the amount of incoming sunlight with outgoing radiant energy. For scientists, to measure all incoming and outgoing energy is to have a bottom line ledger sheet on the sum total of all the physical motions and interactions of our world's climate system. Over the course of a year and over the entire globe, is the Earth's total energy budget in balance? If not, the Earth is either heating up or cooling down. So if scientists are to

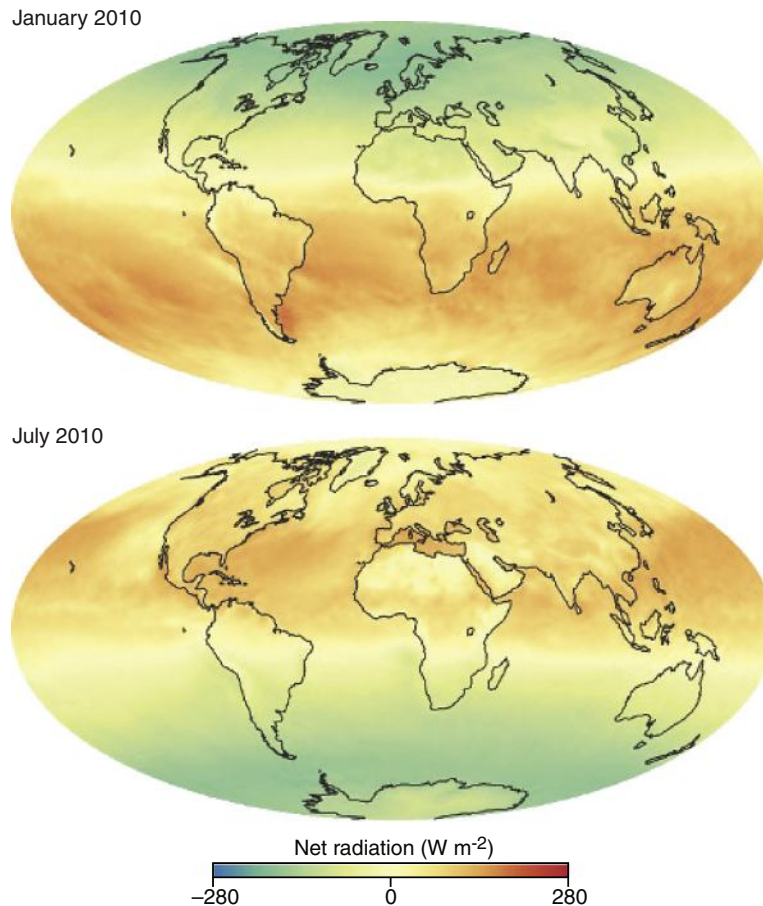


Figure 1 Global net radiation at the top of the Earth's atmosphere measured by CERES on Terra and Aqua on January 2010 (top) and July 2010 (bottom).

understand climate and accurately predict future climate change, then they must determine what drives the changes within the Earth's radiation balance.

In 1978, NASA launched its Nimbus-7 satellite carrying a new sensor, called the Earth Radiation Budget (ERB) experiment, designed to measure direct solar irradiance, reflected shortwave radiation (visible light), and emitted longwave radiation (heat) every day over the entire Earth. This was the first space-based sensor capable of self-calibrating so that its total solar irradiance measurements were accurate to within $\pm 0.5\%$. The Nimbus-7 ERB collected 9 years of global-scale data on which scientists began long-term climate studies. In the interest of extending the ERB data set and improving its measurement capabilities, NASA launched three more Earth Radiation Budget Experiments (renamed ERBE) in the 1980s.

In addition to total solar irradiance, ERBE measured the reflected solar and emitted thermal radiation from the Earth-atmosphere-ocean system. These observations revealed that over the course of a year, the global radiation budget is in balance – the Earth reflects and emits roughly the same amount of energy back into space that it receives from the sun. The data also showed that the average annual, global contribution by clouds is that they reflect 17 W m^{-2} more shortwave energy (visible light) than they trap as longwave energy (heat). Yet, due to calibration uncertainties, deficiencies in ERBE's sampling method, and the

limitations of existing angular dependence models, there still exists a significant uncertainty (about $\pm 5 \text{ W m}^{-2}$) regarding our understanding of Earth's radiation budget. Part of this uncertainty lies in our limited knowledge of the spatial distribution of clouds as well as the optical properties of these clouds over time. Moreover, we cannot be sure how the distribution and optical properties of clouds will change over time. The endeavor to address these issues began with the 1997 launch of the Clouds and the Earth's Radiant Energy System (CERES) sensor aboard the joint NASA/NASDA (National Association of State Departments of Agriculture) Tropical Rainfall Measuring Mission (TRMM) satellite. Twin CERES instruments were also launched aboard NASA's Terra satellite in December 1999 and NASA's Aqua satellite launched in May 2002. Many of the sampling and accuracy limitations on ERBE were addressed in the design of CERES so that it allows scientists to meet ERBE's same measurement objectives with better than twice the former sensor's accuracy.

CERES data permit the determination of net radiation (the balance between incoming and outgoing energy at the top of the atmosphere), which is the total energy available to influence the planet. Due to the seasonal shift of the sun, the excess energy into the system moves with the seasons (Figure 1), with excess energy into the Earth-atmosphere system in the Southern Hemisphere in January (Austral summer) and the Northern Hemisphere in July (Boreal summer). Places where

more energy was coming into the system than going out (positive net radiation) are shown in red-orange, whereas places where more energy was going out than coming in (negative net radiation) are shown in blue-green.

Dust in the Wind

Aerosols are tiny particles suspended in the air (mostly in the troposphere). Some aerosols come from natural sources, such as volcanic eruptions, dust storms, forest and grassland fires, living vegetation, and sea spray. About 10% of the total emitted aerosols in our atmosphere come from human activities, such as the burning of vegetation and fossil fuels and changing the natural land surface cover, which again leads to windblown dust. Most of that 10% is concentrated in the Northern Hemisphere, especially downwind of industrial sites. Yet human-produced aerosols account for about half of the total effect of all aerosols on incoming sunlight. From a satellite's perspective, aerosols raise the Earth's albedo, or make it appear brighter by scattering and reflecting sunlight back to space. The overall effect of these tiny particles is to cool the surface by absorbing and reflecting incoming solar radiation. They also serve as cloud condensation nuclei, or 'seeds' for cloud formation, which again helps to cool the surface. In terms of their net influence on global climate, aerosols represent scientists' greatest subject of uncertainty. Yet computer climate models estimate that over the last century, human-produced aerosols have offset some of the global warming due to greenhouse gases from what otherwise would have taken place.

Through the 1980s and most of the 1990s, the NOAA Advanced Very High-Resolution Radiometer (AVHRR) was the most frequently used satellite sensor for measuring aerosol optical thickness. (Aerosol optical thickness is a measure of how much sunlight airborne particles prevent from traveling through a column of atmosphere.) However, AVHRR can only make such measurements over the ocean, as the sensor requires a relatively uniform and dark-colored background. In April 1991, the European Space Agency (ESA) launched a new type of multiangle sensor, called the Along Track Scanning Radiometer (ATSR), aboard their first European Remote-Sensing Satellite (ERS-1). The ATSR makes aerosol optical thickness measurements by remotely sensing visible and near-infrared wavelengths at nadir and oblique forward scan angles (both within a 2-min interval). A modified version of the sensor, called the Advanced Along Track Scanning Radiometer, was launched in 1995 aboard ERS-2 and 2002 aboard Envisat.

In 1996, Japan launched the first in their series of Advanced Earth Observation Satellites (ADEOS) satellites, which carried a payload of two sensors—the Polarization and Directionality of the Earth's Reflectances (POLDER) sensor, contributed by the French Space Agency, and the Ocean Color and Temperature Scanner, provided by NASDA. Both sensors can retrieve aerosol measurements, but POLDER was the first satellite sensor specifically designed to measure aerosols and it can make its measurements over both land and ocean. The sensor observes Earth targets from 12 directions that enable measurements of the bidirectionality and polarization of solar radiation reflected from within the atmosphere. Unfortunately, due to its solar panel failing, the ADEOS mission ended

prematurely after only 8 months in orbit. A POLDER instrument on the French PARASOL (Polarization and Anisotropy of Reflectances for Atmospheric Sciences coupled with Observations from a Lidar) mission was launched in 2004 and is producing aerosol optical thickness retrievals over both land and ocean.

Three sensors aboard NASA's Terra satellite are particularly well-suited for studying the effects of aerosols on climate: CERES, Multiangle Imaging Spectroradiometer (MISR), and Moderate-Resolution Imaging Spectroradiometer (MODIS). Both the MODIS and the MISR have the capacity to measure both aerosol optical thickness and the sizes of aerosol particles over both ocean and land. Particle size is an indicator of the source of the aerosol particles and helps scientists distinguish aerosols of natural origin from those that are man-made. Moreover, with its nine different look angles, MISR is ideally designed to quantify the reflective properties. Again, CERES complements MODIS and MISR by providing measurements of the shortwave radiation that aerosols reflect back into space. Together, these sensors are providing new insights into the roles of aerosols in the Earth's total energy budget (Figure 2).

Abstract Art or Arbiters of Energy?

More than just the idle stuff of daydreams, clouds help control the flow of radiant energy around our world. Clouds are plentiful and widespread throughout Earth's atmosphere – covering between 66 and 69% of our planet at any given time – so they play a dominant role in determining how much sunlight reaches the surface, how much sunlight is reflected back into space, how and where warmth is spread around the globe, and how much heat escapes from the surface and atmosphere back into space. Clouds are also highly variable. Clouds' myriad variations through time and space make them one of the greatest areas of uncertainty in scientists' understanding and predictions of climate change. In short, they play a central role in our world's climate system. Whereas thick, low-level stratocumulus clouds cool the Earth's surface by reflecting incoming solar radiation, thin high-level cirrus clouds exert a warming influence by allowing sunlight to pass through but then trapping the heat emitted by the surface. Which of clouds' effects is greater over time, warming or cooling? Scientists did not know the answer to that question until relatively recently. Based on ERBE satellite data collected in mid to late 1980s, coupled with aircraft and surface-based measurements, scientists demonstrated that, globally, clouds' cooling effect on the surface is greater than their warming effect. So great is this cooling effect, it is as if clouds remove the heat of a 60-W light bulb from every $2 \times 2 \text{ m}^2$ of the Earth's surface. But will they continue to cool our planet over the next century if a greenhouse gas-driven global warming scenario comes to pass? Or even, could the type and distribution of clouds change so that they primarily exert a warming influence? (Figure 3).

Two sensors flying aboard NASA's Terra satellite, launched in December 1999, are designed to help scientists answer these questions. Both MODIS and MISR give scientists new capabilities for measuring the structure and composition of clouds. MODIS observes the entire Earth almost every day in

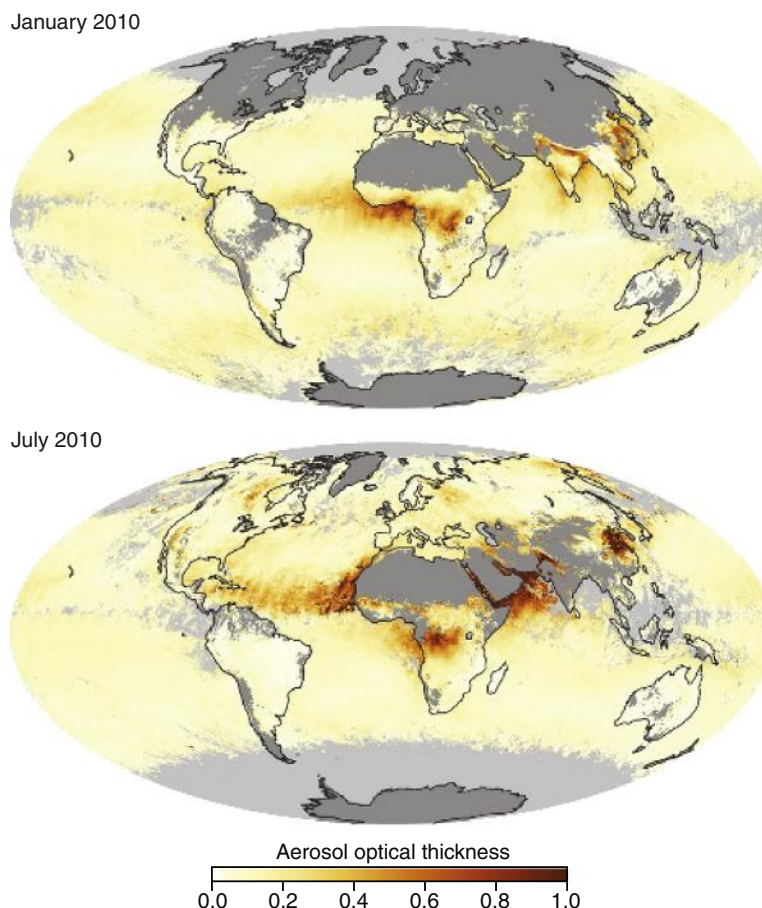


Figure 2 Global aerosol optical thickness measured by MODIS on Terra on January 2010 (top) and July 2010 (bottom).

36 spectral bands ranging from visible to thermal infrared wavelengths. With spectral and spatial resolutions superior to that of AVHRR (its heritage sensor), MODIS can measure a wide suite of clouds' physical and radiative properties. Specifically, MODIS can determine whether a cloud is composed of ice or liquid water particles (or some combination of the two), it can measure the effective radius of the particles within a cloud, it can observe how much (or little) sunlight passes through a cloud, and it can measure the temperature and altitude of cloud tops.

Complementing MODIS, the MISR instrument 'sees' the Earth simultaneously in red, green, blue, and near-infrared wavelengths at nine different angles – at four progressively more oblique angles ahead of Terra, four angles aft of the satellite, and one at nadir. Because it measures any given scene from multiple angles, MISR is ideally designed to help scientists better understand how clouds interact with radiant energy as both a function of their structure and type. CERES complements MODIS and MISR by providing measurements of the shortwave and longwave radiant energy that clouds reflect and emit back into space.

Starting in 2006, NASA launched two satellites carrying active sensors to monitor the vertical distribution of clouds and aerosols, and to monitor the water content within those clouds. CALIPSO (Cloud-Aerosol Lidar and Infrared Pathfinder

Satellite Observations) carried the CALIOP (Cloud-Aerosol Lidar with Orthogonal Polarization) instrument, which is a dual frequency monostatic lidar that sends its own laser pulse down to the Earth and measures the time delay from the return signal scattering by clouds, aerosols, and the Earth's surface, from which it determines the height and often thickness of clouds. This payload was complimentary to the CloudSat satellite that carried its Cloud Profiling Radar. These two nadir-viewing sensors are in the same orbit and only 60 s behind Aqua, which carries the wide-swath MODIS imager. These powerful active sensors are improving our assessment of cloud properties over complicated multilayered clouds and in Polar Regions where it is much more difficult to distinguish clouds from the bright snow-covered surfaces using passive remote-sensing techniques alone.

Serendipity and Stratospheric Ozone

In the early 1970s, as Earth scientists intensified their studies into the possible causes and effects of global warming, a suite of man-made gases in particular elicited the attention of scientists – chlorofluorocarbons (CFCs). Increasingly, CFCs were being used by industrial nations in the production of a variety of commercial products (e.g., refrigerants, aerosol

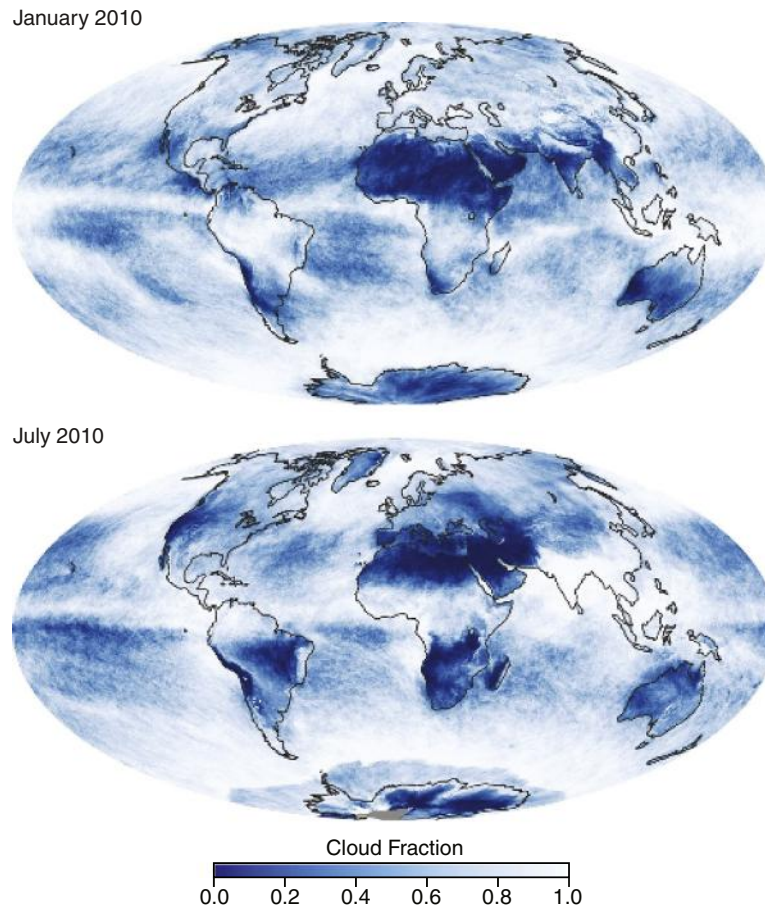


Figure 3 Global cloud fraction measured by MODIS on Terra on January 2010 (top) and July 2010 (bottom).

sprays). The concern is twofold: CFCs are up to 200 times more efficient than carbon dioxide at trapping heat in the Earth's atmosphere and the gas tends to remain in the atmosphere up to 120 years once released. Then, in 1974, two scientists wrote of a new concern that CFCs could potentially reduce levels of ozone in the stratosphere, the layer of atmosphere from 10 to 50 km in altitude. In 1975, the US Congress asked NASA to develop a 'comprehensive program of research, technology, and monitoring of phenomena of the upper atmosphere. 'In particular, Congress' intent was to ascertain the 'health' of the ozone layer (Figure 4).

So, in addition to ERB, in 1978 Nimbus-7 carried two other new NASA sensors designed to measure the total amount of ozone in a given column of atmosphere over the entire globe – called the Solar Backscatter Ultraviolet instrument and the Total Ozone Mapping Spectrometer (TOMS). Sensitive to radiant energy in the ultraviolet region of the spectrum, these sensors took advantage of the fact that molecules and aerosol particles reflect certain wavelengths of ultraviolet rays while ozone absorbs others at different levels in the atmosphere. By analyzing the amount of ultraviolet energy reflected back up to the spacecraft, researchers could produce profiles of how thick or thin the ozone was at different altitudes and locations. Ironically, it was not until October 1985 that a British team of scientists found a significant reduction in ozone over Halley Bay, Antarctica. Using a ground-based Dobson ozone

spectrophotometer, the team found that the amount of stratospheric ozone over Halley Bay was about 40% less than it had been the previous year. Their finding stunned the science community because they were expecting anthropogenic ozone depletion to occur first at upper levels in the stratosphere (30–50 km) and so they anticipated that the initial signal of depletion in a total column of ozone would be weak. NASA researchers hastily reviewed their TOMS data and found that it too had detected a dramatic loss of ozone over all of Antarctica. Why had not they discovered the phenomenon earlier? Unfortunately, the TOMS data analysis software had been programmed to flag and set aside data points that deviated greatly from expected measurements and so the initial measurements that should have set off alarms were simply overlooked. In short, the TOMS team failed to detect the ozone depletion years earlier because it was much more severe than scientists expected.

In the years following the discovery of the ozone hole, NASA and ESA satellites recorded ozone levels over Antarctica growing worse with each passing year. In response, in 1987, 43 nations signed the 'Montreal Protocol' in which they agreed to reduce the use of CFCs by 50% by the year 2000. This protocol was amended in 1990, 1992, and 1999 to eliminate all CFC emissions by 2000.

ESA's second European Remote-Sensing Satellite (ERS-2) carries a sensor called the Global Ozone Monitoring

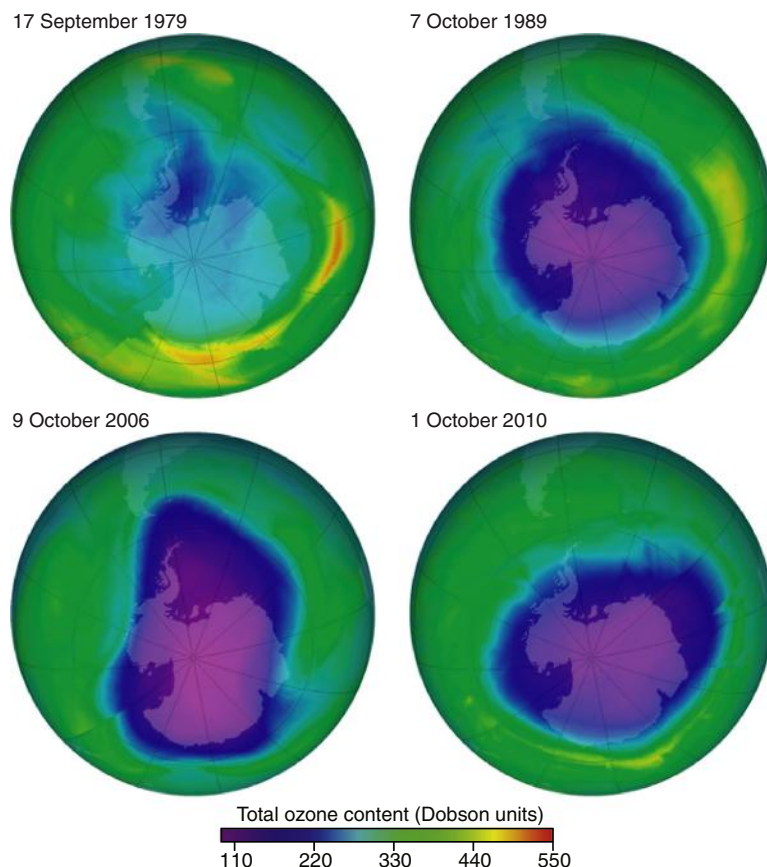


Figure 4 The minimum total ozone content of the Southern Hemisphere for selected years between 1979 and 2010, based on data from the TOMS on the Nimbus-7, Meteor 3, and Earth Probe satellites (1979–2003), and the OMI on the EOS Aura satellite (2004 to present). Dobson unit = 2.69×10^{16} molecules cm^{-2} .

Experiment (GOME). GOME is a nadir-looking sensor with four bands ranging from 240 to 790 nm for measuring backscattered visible and ultraviolet solar radiation. Since the summer of 1996, ESA has routinely produced 3-day global measurements of total ozone and nitrogen dioxide using GOME data, and later this was expanded with the launch of Scanning Imaging Absorption Spectrometer for Atmospheric Cartography (SCIAMACHY) on Envisat in 2002, which looks at reflected solar radiation not only at nadir but also at the Earth's limb.

The Ozone Monitoring Instrument (OMI) was developed in the Netherlands and launched aboard NASA's Aura satellite in 2004. This instrument now provides enhanced capability for monitoring the global distribution of total ozone on a daily basis, as well as the global distributions of other atmospheric trace gases such as sulfur dioxide and nitrogen dioxide.

As recently as 1998, both TOMS and GOME data show that at its Austral spring low, Antarctic ozone concentrations had worsened to 60% less than early 1970s levels. Today there is some evidence that levels of chlorine in the stratosphere are leveling off. Is this a scientific success story in the making? Will stratospheric ozone concentrations return to pre-1970s levels as the abundance of stratospheric chlorine stabilizes? Only time and continued monitoring will tell.

The Chemistry of Earth's Atmosphere

Some satellite sensors allow scientists to determine the chemical content of the Earth's upper atmosphere using a technique called 'solar occultation,' in which a sensor is pointed toward the horizon at sunrise and sunset to measure the profile of the stratosphere and mesosphere about 30 times per day. In this way sensors, such as the Stratospheric Aerosol and Gas Experiment (SAGE), can determine the presence and abundance of gases and particulates by measuring precisely the visible and ultraviolet wavelengths that are absorbed within the upper atmosphere. Since the spectra of ozone, nitrogen dioxide, sulfur dioxide, and certain aerosols are well-known, scientists can directly correlate SAGE's readings with the presence of these substances within the stratosphere. The solar occultation technique is particularly effective because the sensor is self-calibrating – each occultation event looks directly at the unattenuated sun outside the Earth's atmosphere just prior to sunset or just following sunrise. These observations are then compared to observations of the sun obtained by looking through the atmosphere. The direct sun observations establish an ongoing baseline of the sensor's performance. Adapted from the Stratospheric Aerosol Mission (SAM II) that flew aboard Nimbus-7, the SAGE sensor is essentially a modified sunphotometer. This kind of sensor first flew in 1979 aboard NASA's

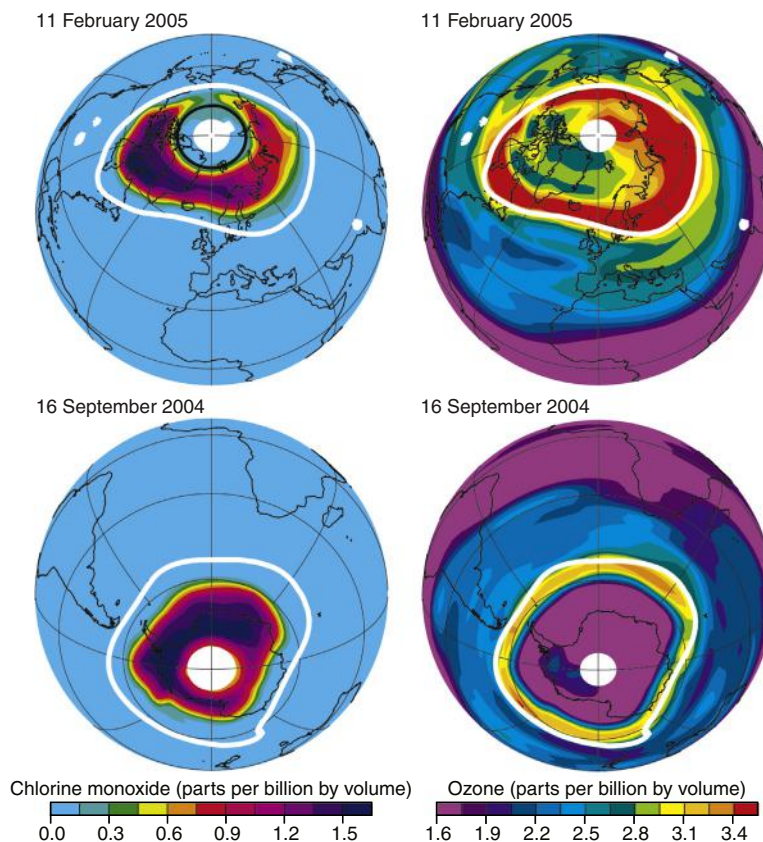


Figure 5 Chlorine monoxide (left) and ozone concentration (right) derived by MLS on Aura at approximately 18-km altitude on 11 February 2005 in the Northern Hemisphere (top) and 16 September 2004 in the Southern Hemisphere (bottom). The white lines demark the approximate edge of the winter polar vortex in each hemisphere. The high chlorine monoxide within the polar vortex in the left-hand figures (reds and dark purple shades) is directly associated with, and leads to, a reduced ozone concentration shown in the right-hand figures (which is especially striking in the Southern Hemisphere).

Applications Explorer Mission-2 (AEM-2). A subsequent version of SAGE (SAGE II) was launched aboard ERBS in 1984 and performed well throughout 2001, thus giving scientists a 17-year continuous dataset of upper atmosphere profile measurements. An enhanced version of SAGE (SAGE III) was launched aboard Russia's Meteor-3M spacecraft in 2001 and operated until 2006.

In 1991, NASA launched the Upper Atmosphere Research Satellite (UARS) with a payload of 10 sensors for measuring a wide array of chemical and physical phenomena in the stratosphere and mesosphere (the layers of the atmosphere from approximately 12–80 km in altitude). Not only did UARS extend scientists' ability to monitor stratospheric ozone concentrations into the 1990s but it also provided the first comprehensive picture of the photochemical processes involved in ozone destruction. The UARS Microwave Limb Sounder (MLS) demonstrated that there is a direct link between the presence of chlorine and the formation of chlorine monoxide during winter in the Northern and Southern Hemispheres, and the destruction of ozone. This capability has been further enhanced with the MLS onboard the Aura satellite, which was launched in 2004 (Figure 5).

Among the four instruments carried on the Aura satellite is the OMI instrument that is a nadir-viewing solar backscatter spectrometer used to derive column ozone, nitrogen dioxide, bromine monoxide, chlorine dioxide, sulfur dioxide, formaldehyde, and

aerosols. Nitrogen dioxide was first measured from space by GOME on ERS-2, and is also measured by SCIAMACHY on Envisat. Nitrogen oxides are formed from both natural (lightning, wildfires, and soil emissions) and anthropogenic (power plants, internal combustion engines, fertilizer applications, and agricultural burning) processes, and they have a relatively short lifetime (about a day) and are therefore concentrated near their sources. Because they are short-lived gases in the atmosphere, they show a markedly lower level of concentration during 'weekends' than 'weekdays,' where the 'weekend' is Sunday in Europe, North America, and Japan, Saturday in Israel, and Friday in the Islamic cities of the Middle East. Using data from OMI, GOME, and SCIAMACHY, it is also possible to see the lower level of NO₂ emissions in the Ohio River Valley of the United States due to greater controls on cars and power plants, in contrast to emissions from China that are continually increasing (Figure 6).

A Canadian instrument launched in 1999 aboard NASA's Terra satellite uses gas correlation spectroscopy to determine the abundance of methane and CO in the troposphere. The Measurements of Pollution in the Troposphere (MOPITT) sensor measures emitted and reflected radiation from the Earth in three spectral bands. As this light enters the sensor, it passes along two different paths through onboard containers of CO and methane. The different paths absorb different amounts of energy, leading to small differences in the resulting signals that

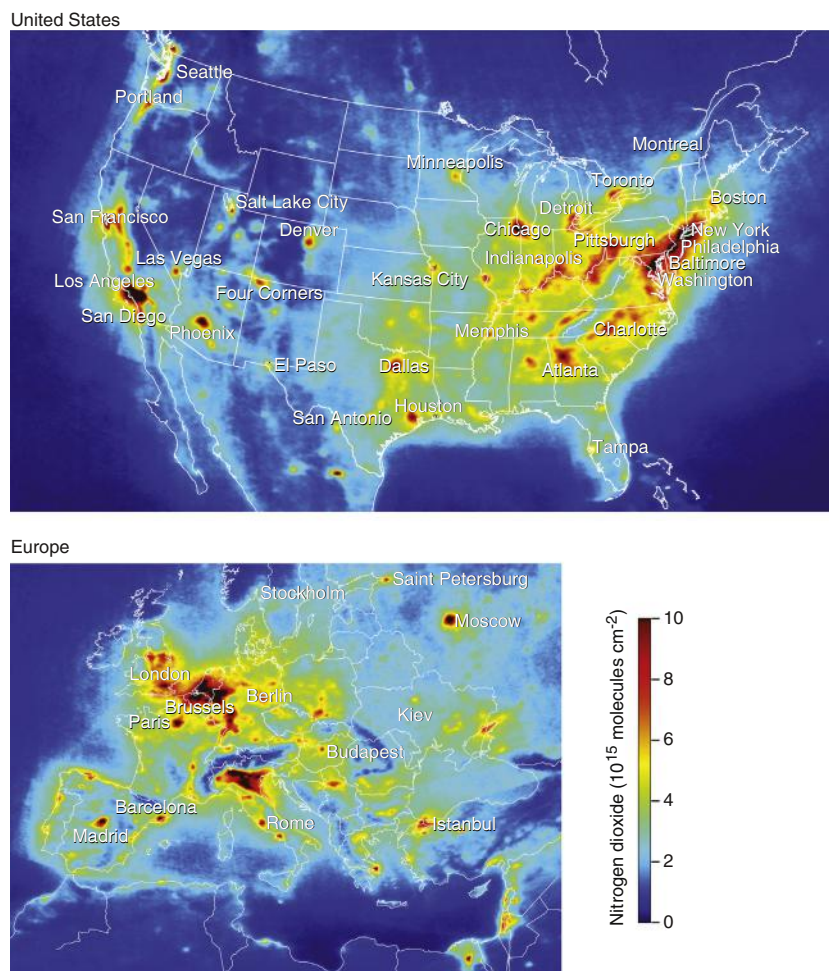


Figure 6 Nitrogen dioxide in the troposphere in the United States (top) and Europe (bottom) during 2006 derived by OMI on Aura. The principal sources of NO_2 occur over large populated regions, heavily industrialized areas, and power plants.

directly correlate with the presence of these gases in the atmosphere. Both methane and CO are by-products of burning vegetation as well as fossil fuels. Scientific interest in CO is twofold: the gas controls atmospheric concentrations of oxidants, thus affecting the ability of the atmosphere to clean itself from the ongoing generation of harmful tropospheric ozone from biomass burning and urban smog. Also, through chemical reactions within the lower atmosphere, CO contributes to the production of harmful ozone. MOPITT is helping scientists identify the main sources of these gases as well as improve four-dimensional models of their transport through the atmosphere.

NASA's Aqua satellite carries the Atmospheric Infrared Sounder that was originally designed to measure atmospheric temperature and water vapor throughout the troposphere and up to about 32 km in altitude, using a hyperspectral thermal emission spectrometer. The emission of radiation in the $15\text{-}\mu\text{m}$ wavelength region is strongly influenced by the concentration of carbon dioxide (CO_2) in the Earth's atmosphere, and this signature has been used to derive the concentration of CO_2 in the mid-troposphere. Results show that CO_2 is not as well-mixed as previously believed, and that its transport around the

globe is strongly influenced by mid-latitude jet streams as well as synoptic weather patterns, especially in the Southern Hemisphere (Figure 7).

Where Storm Clouds Gather

Rain clouds form when moisture-laden air is driven skyward by warm updrafts emitted from a sun-warmed land or ocean surface; or when mountain slopes push moist air aloft; or when a wedge of colder, denser air plows warmer, moist air upward to higher elevations. Because cold air cannot hold as much water vapor as warm air, and because the atmosphere cools at higher elevations, water vapor condenses readily into liquid droplet or ice crystal form, in the presence of seed aerosol particles. Were there no aerosol particles in the Earth's atmosphere, there would be no fog, no clouds, no mist, and probably no rain. When water evaporates at the surface, it absorbs energy from its surroundings and stores it as 'latent heat.' When water vapor condenses back into liquid or ice form, it releases its latent heat into its surroundings. Only about 25% of the energy contained within the atmosphere comes directly from

Mid-tropospheric (8 km) Carbon dioxide

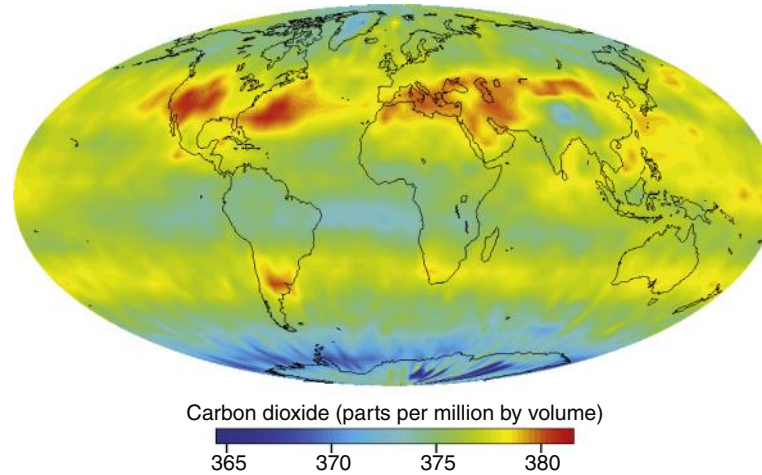
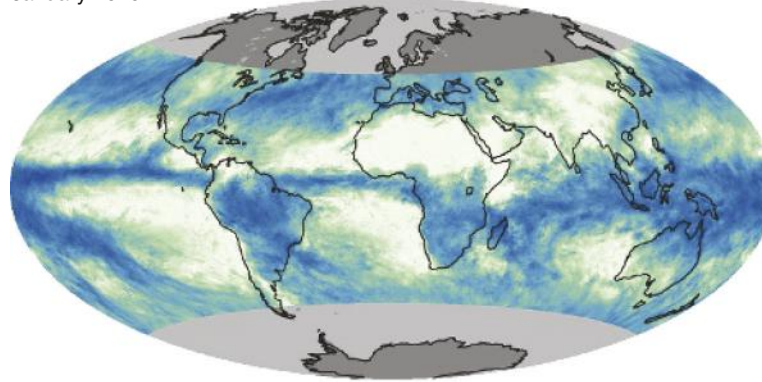


Figure 7 Carbon dioxide in the free troposphere (~ 8 km) during July 2003 derived by Atmospheric Infrared Sounder on Aqua. There is a belt of mid-tropospheric air enriched with carbon dioxide that encircles the globe in the mid-latitudes of both hemispheres, due in large part to the mid-latitude jet streams.

January 2010



July 2010

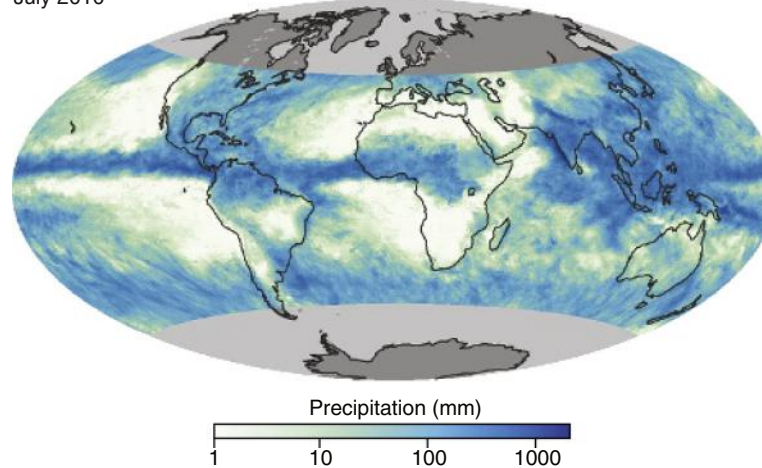


Figure 8 Global monthly averaged rainfall measured by TMI on TRMM on January 2010 (top) and July 2010 (bottom).

the sun's rays; the remaining 75% comes from the release of latent heat contained in water vapor. Most atmospheric water vapor originates from the tropical oceans, where it rises high into the atmosphere to form towering thunderheads. Encircling

the globe along the equator is an almost continuous band of thunderheads known as the Intertropical Convergence Zone, producing roughly three-quarters of the energy in our atmosphere that helps to drive its circulation patterns.

We cannot measure the latent heat contained within clouds. We can, however, measure tropical rainfall. If we are to more accurately determine how much energy our atmosphere receives from latent heat, then we must more accurately measure rainfall. In 1997, NASDA and NASA jointly developed and launched the TRMM into a mid-inclination (35°) precessing orbit. Scientists estimate about 60% of our world's precipitation falls within the band spanning between 30° north and south of the equator. TRMM carries two instruments designed to measure rainfall – the Precipitation Radar (PR) and the TRMM Microwave Imager (TMI). Designed and built by NASDA, the PR is the first satellite sensor to provide three-dimensional images of the internal structures of storm clouds. Its measurements show scientists the intensity and distribution of rain within a storm, the total height of a storm, and the elevation at which ice crystals melt into raindrops. Most importantly, the PR can measure rain rates as accurately as 0.7 mm h^{-1} . While scientists expected to use ground-based Doppler Radar stations to validate TRMM's PR measurements, much to their pleasant surprise they found that the latter exceeds most ground-based measurements in accuracy and spatial resolution.

The TMI is a 'passive' sensor designed to measure minute amounts of microwave energy emitted by the Earth's surface and from within its atmosphere. (Whereas 'active' sensors send pulses of energy and then measure how much gets absorbed and reflected by the target, 'passive' sensors receive and measure only energy originating from, or reflected by, external sources.) These measurements allow TMI to quantify the amount of water vapor, cloud water, and rainfall intensity within the atmosphere. Based on the design heritage of the Defense Meteorological Satellite Program's Special Sensor Microwave or Imager, the TMI has a wider viewing swath (780 km) and finer spectral resolution than its predecessors. TRMM shows that there is a band of heavy rain along the equator that moves north and south of the equator, leading to several months of near-daily rain followed by months of dryness. The Asian monsoon brings rain to China, Southeast Asia, and India between April and September, and South America goes through a rainy season from October through May (Figure 8).

Conclusion

As the preceding sections demonstrate, the Earth's atmosphere changes both physically and chemically over a range of scales of time and space. The atmosphere's chemical makeup affects its physical state, such as its radiative properties. Much like a venetian blind, the gases and particles in the atmosphere selectively absorb and reflect certain wavelengths of solar radiation while allowing others to pass through relatively unhindered. In turn, physical processes in the atmosphere also help determine its chemical makeup. There was growing consensus through the 1970s and 1980s among Earth scientists that we need to take a more holistic approach to global climate change studies. Scientists recognized that nature does not compartmentalize climate phenomena into discreet disciplines and therefore we need to examine the variables of change as integral parts of the vast, interconnected web of cause-and-effect that is Earth's climate system. In short, it is not enough to identify where and when changes occur; we need to understand how and why the mechanisms of change work. Satellite remote sensors offer the only viable means of conducting a comprehensive examination of our planet.

See also: **Aerosols:** Observations and Measurements. **Radiation Transfer in the Atmosphere:** Absorption and Thermal Emission; Scattering. **Satellites and Satellite Remote Sensing:** Earth's Radiation Budget; Measuring Ozone from Space – TOMS and SBUV; Orbits; Precipitation.

Further Reading

- King, M.D., Kaufman, Y.J., Tanré, D., Nakajima, T., 1999. Remote sensing of tropospheric aerosols from space: past, present, and future. *Bulletin of the American Meteorological Society* 80, 2229–2259.
- King, M.D., Parkinson, C.L., Partington, K.C., Williams, R.G. (Eds.), 2007. *Our Changing Planet: The View from Space*. Cambridge University Press, New York.
- Parkinson, C.L., 1997. *Earth from Above: Using Color-Coded Satellite Images to Examine the Global Environment*. University Science Books, Sausalito, CA.
- Ramanathan, V., Barkstrom, B.R., Harrison, E.F., 1989. Climate and the Earth's radiation budget *Physics Today* 42, 22–32.

Surface Wind and Stress

WT Liu, California Institute of Technology, Pasadena, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

The basics of scatterometry and air–sea turbulence transfer are discussed to bring out the capabilities of space sensors in measuring ocean surface stress and wind vectors. The scientific significance of wind and stress are described. Past knowledge of stress characteristics was based on those of wind, because of the lack of stress measurements. Using the unique measurement of stress by the scatterometer, the differences between stress and wind over oceanfronts and in hurricanes are revealed. Wind speed retrieval from measurements of radar altimeter, microwave radiometer, and synthetic aperture radar is summarized. Potential improvements in measuring wind and stress are suggested.

Introduction

Wind is air in motion, and it is a vector quantity with a magnitude (speed) and a direction. Sailors understand both the importance and the difficulty in getting information on wind over oceans. Textbooks still describe global ocean wind distribution in sailors' terms: the calms of the doldrums and horse latitudes, the steady trade winds, and the ferocity of the roaring forties. These features are clearly visible in [Figure 1](#), which is derived from one day of observations by a space-based scatterometer, QuikSCAT. Just a few decades ago, almost all ocean wind measurements came from merchant ships. However, the quality and geographical distribution of these wind reports were uneven. Today, operational numerical weather prediction (NWP) also gives us wind information, but NWP depends on models, which are limited by our knowledge of the physical processes and the availability of data.

The ocean interacts with the atmosphere in nonlinear ways; processes at one scale affect processes at other scales. Adequate wind coverage could be achieved only from the vantage point of space. Space-based microwave radars measure ocean surface roughness day and night, under clear and cloudy conditions. Ocean surface roughness is driven by stress, the turbulent transfer of momentum between the ocean and the atmosphere. Stress is another vector quantity closely related to wind. The microwave scatterometer is the best-established instrument to measure surface stress magnitude and direction, and has been promoted as a wind sensor. The scientific significance of wind and stress is introduced in the Scientific Significance section. The principles of the scatterometer are summarized in the Scatterometry section.

The primary functions of the radar altimeter, the synthetic aperture radar (SAR), and the microwave radiometer are not wind measurement, but they give wind speed as a secondary product. Wind speed, even without direction, is important, and wind speed from these sensors can be applied with directional information derived from other means. The space-based polarimetric radiometer also shows sensitivity to wind direction under favorable conditions. These instruments are described in the Other Sensors section.

While the general public knows and feels the wind, very few people know what stress is. Even for oceanographers, the concept of stress distribution is largely derived from that of wind, because there was no large-scale measurement of stress

over the ocean until the launch of the first scatterometer. The relation and the difference between wind and stress are discussed in the Relations between Wind and Stress section. Potential improvements in space-based measurements of wind and stress are given in the Potential Improvements and Conclusion section.

Scientific Significance

Ocean wind is strongly needed for marine weather forecasting and to avoid shipping hazards. Surface wind convergence brings moisture and latent heat that drive deep convection and fuel marine storms. The significance of wind measurement is clearly felt, for example, when a hurricane suddenly intensifies and changes course, or when the unexpected delay of monsoon brings drought. Detailed distribution of wind power is also needed for the optimal deployment of floating wind farms on the open sea that are enabled by new technology.

For oceanographers, it is stress more than wind that drives ocean circulation. The two-dimensional stress field is needed to compute the divergence and curl (vorticity) that control the vertical mixing. The mixing brings short-term momentum and heat trapped in the surface mixed layer into the deep ocean, where they are stored over time. It also brings nutrients and carbon stored in the deep ocean to the surface, where there is sufficient light for photosynthesis. The horizontal currents, driven in part by stress, distribute the stored heat and carbon in the ocean. Stress affects the turbulent transfer of heat, moisture, and gases between the ocean and the atmosphere and is critical in understanding and predicting weather and climate changes.

Scatterometry

During the Second World War, marine radar operators observed noises on their radar screens, which obscured small boats and low-flying aircraft. They termed this noise 'sea clutter'. This clutter was the backscatter of the radar pulses by the small waves on the ocean's surface. The radar operators at that time were quite annoyed by these noises, not knowing that, a few decades later, scientists would make important applications with them.

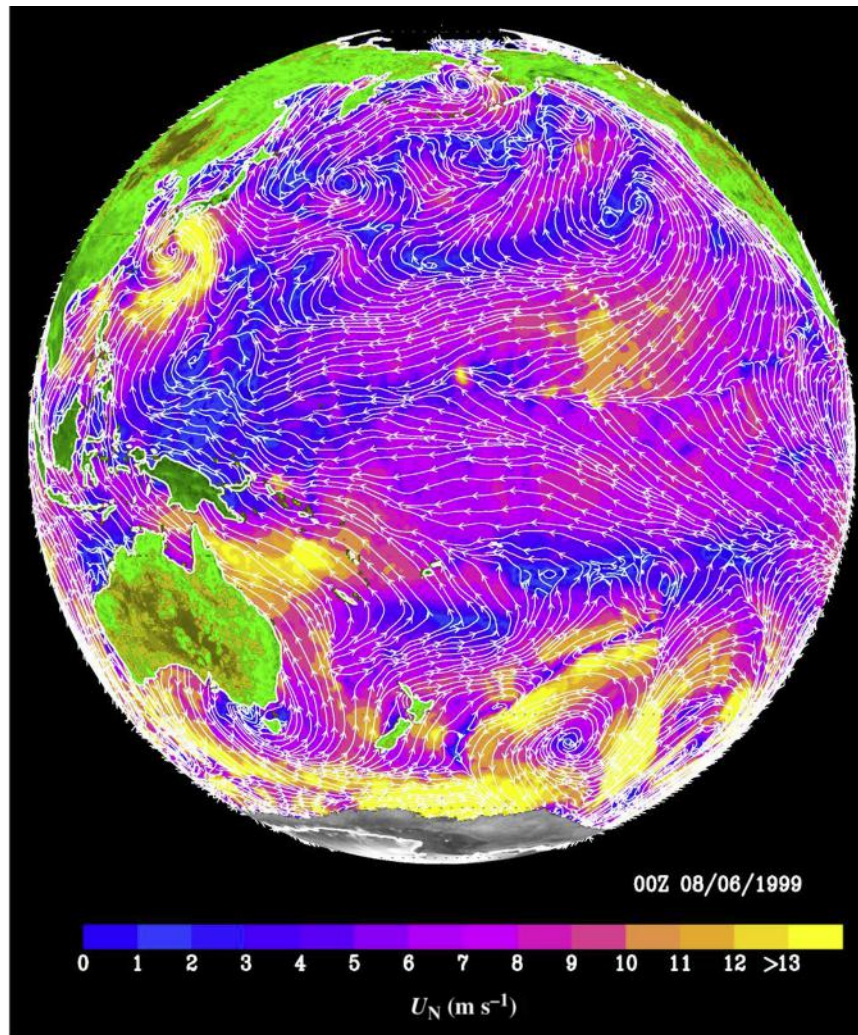


Figure 1 Over the ocean, white streamlines indicating wind direction are superimposed on the color image of U_N at 00Z on 6 August 1999, derived from objective interpolation of the observations by QuikSCAT. Normalized backscatter coefficients measured by the same instrument over land and Antarctica are also displayed.

The scatterometer sends microwave pulses to the Earth's surface and measures the power backscattered from the surface roughness. The roughness may describe the characteristics of polar ice or vegetation over land. Over the ocean, which covers nearly three-quarters of the earth's surface, the surface roughness is largely due to the small centimeter waves on the surface. These surface waves are believed to be in equilibrium with the local stress. The backscatter depends on not only the magnitude of the stress but also the stress direction relative to the direction of the radar beam (azimuth angle). The capability of measuring both stress magnitude and direction is the major unique characteristic of the scatterometer.

At incident angles greater than 20° , the radar return is governed by Bragg scattering, and the backscatter increases with stress. The backscatter is governed by the in-phase reflections from surface waves. The geophysical model functions (GMF), from which ocean surface wind and stress are retrieved from the observed backscatter, in the form of the normalized radar cross-section (σ_o), are largely based on empirical fits of

data. The symmetry of backscatter with wind direction requires observations at many azimuth angles to resolve the directional ambiguity. Because of uncertainties in the wind retrieval algorithm and noise in the backscatter measurements, the problem of directional ambiguity was not entirely eliminated even with three azimuthal looks in the scatterometers launched after SEASAT (in 1978). A median filter iteration technique has been commonly used to remove the directional ambiguity.

Because there is much more wind information than stress information and the public is more familiar with wind than stress, an equivalent neutral wind (U_N) is used as the geophysical product of the scatterometer. By definition, U_N is uniquely related to stress, while the relation between stress and the actual wind depends on atmospheric stability (see [Relations between Wind and Stress](#) section). Although scatterometers have been known to measure surface stress, they have been used and promoted as wind-measuring instruments, and U_N has been used as the actual wind, particularly in operational

weather applications. More explanation is given in the Relations between Wind and Stress section.

Over the large expanse of ocean, which is quasi-stationary and horizontally homogeneous under near-neutral conditions, and where ocean surface current is negligible, U_N may be the same as the actual wind. Over the sharp horizontal current shear and temperature gradients of oceanfronts, however, stress variation could be very different from that of winds. The observation of the rotation of scatterometer measurement in opposition to the surface current, in the meanders of the Kuroshio Extension current, is a clear characteristic of turbulent stress generated by shear. One would expect wind to be dragged in the same direction as the current, but stress is the vector difference between wind and current, and the direction would be deflected from the current. Figure 2 clearly shows that where the vorticity of U_N measured by QuikSCAT is positive, the vorticity of the surface current measured by the drifters is negative and vice versa, indicating opposite rotations.

The ubiquitous spatial coherence between sea surface temperature (T_s) and U_N , measured by the scatterometer and

found under a variety of atmospheric conditions, is also characteristic of turbulent stress generated by buoyancy. In the unstable region, atmospheric buoyancy generates turbulent momentum transport and increases the stress magnitude. Figure 3(a) shows the coherence over the Kuroshio Extension. Figure 3(b) shows similar coherence between T_s and U_N computed from a uniform wind field under similar stability conditions, demonstrating that the coherence is a characteristic of stress and not wind. Factors affecting larger-scale wind, such as the pressure gradient force, the Coriolis force, and baroclinicity, are not important at the small scales of turbulence, and that is the reason for ubiquitous coherence. The higher stress over warmer water affects atmospheric wind aloft, but the influence will be subjected to these large-scale factors. Ocean parameters, such as surface current and temperature, are needed to derive wind from stress in these frontal regions.

Retrieving strong winds from the scatterometer is also difficult. The problem is obvious in Figure 4, which is derived from NASA's scatterometer on QuikSCAT measuring at the Ku-band (14 GHz). Data for the 12 hurricanes in the North

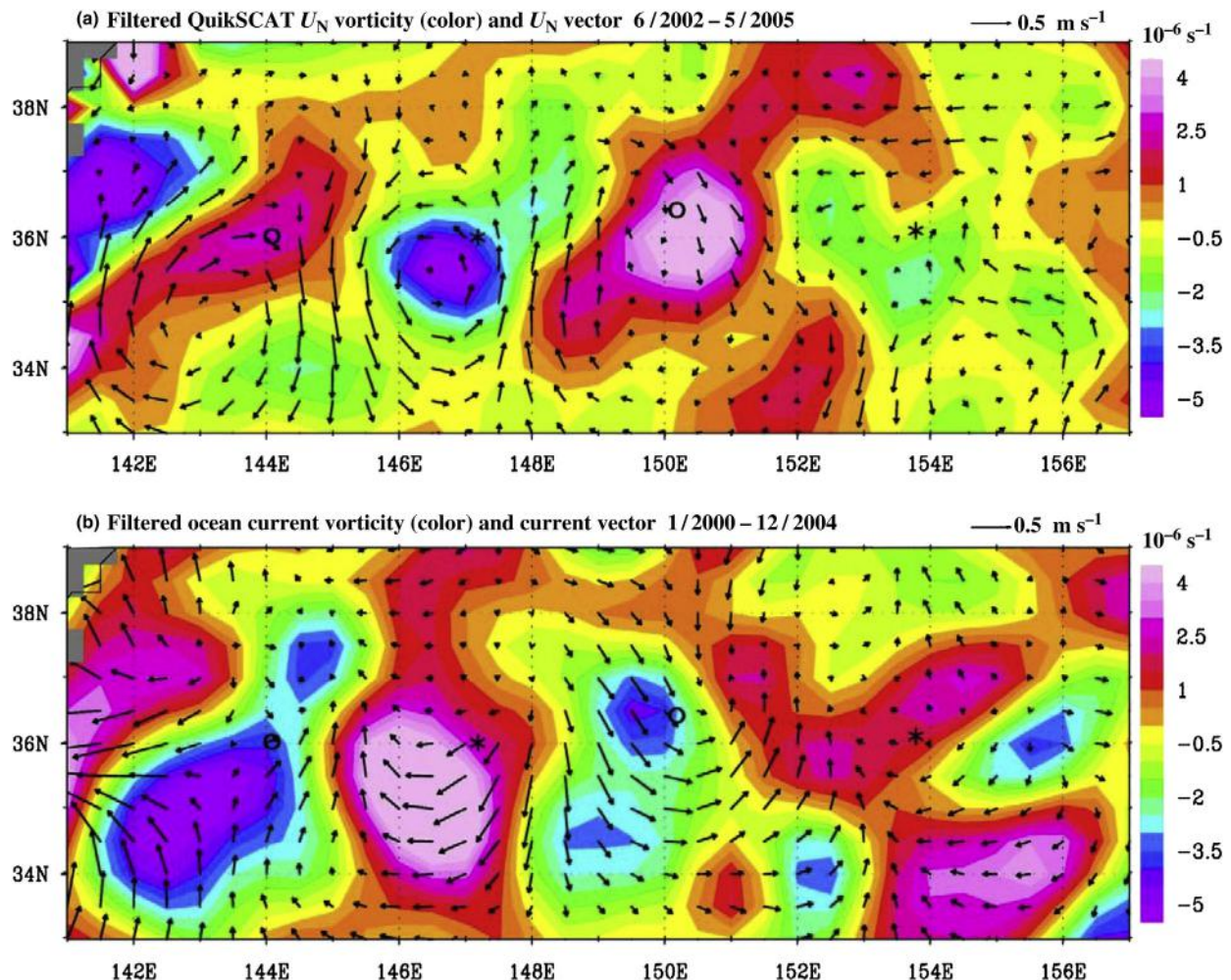


Figure 2 (a) Filtered vector (black arrows) superimposed on vorticity (color, 10^{-6} s^{-1}) of U_N observed by QuikSCAT, averaged from June 2002 to May 2005. (b) Filtered vector (black arrows) superimposed on vorticity (color, 10^{-6} s^{-1}) of the surface current measured by Lagrangian drifters, averaged from 2000 to 2004. The large-scale gradients are removed by a two-dimensional filter.

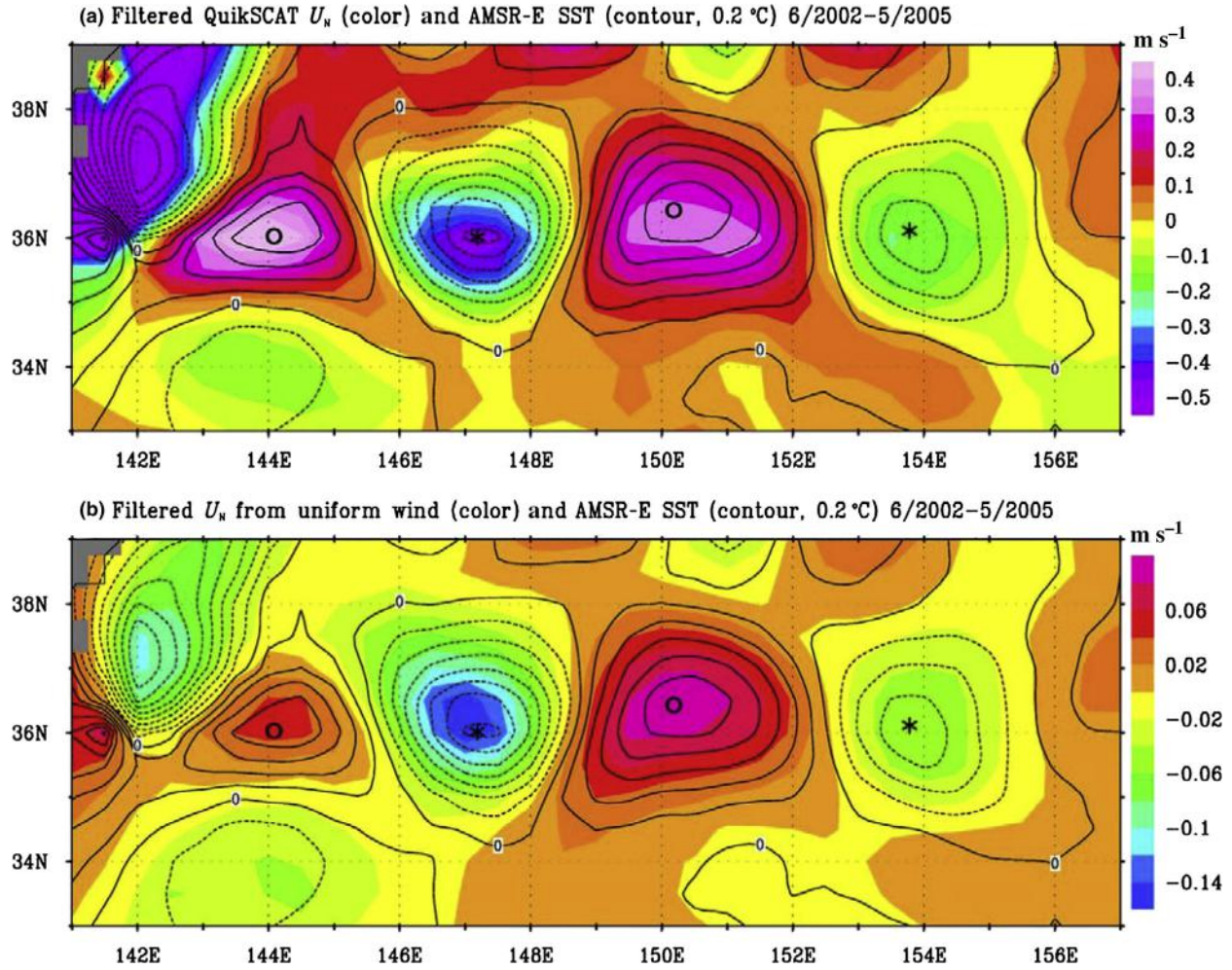


Figure 3 (a) Isotherms of filtered T_s measured by the Advanced Microwave Scanning Radiometer for EOS (AMSR-E; EOS is NASA's Earth Observing System) (0.2°C interval) superimposed on (a) filtered magnitude of QuikSCAT U_N (color, m s^{-1}); and (b) filtered U_N computed from a uniform wind field of $u = 7.5 \text{ m s}^{-1}$ (color, m s^{-1}), averaged from June 2002 to May 2005. Solid and broken lines represent positive and negative values, respectively. The same filter as in Figure 2 is applied.

Atlantic in the 2005 season, excluding those with over 10% chances of rain, were examined. Figure 4 shows that, in moderate winds ($U < 35 \text{ m s}^{-1}$), the logarithm of σ_o (in db) increases linearly with the logarithm of wind speed at both polarizations. At strong winds ($U > 35 \text{ m s}^{-1}$), however, σ_o increases at a much slower rate with increasing wind speed. Similar saturation is found in the European Advanced Scatterometer (ASCAT), measuring at the C-band (5 GHz). Such high wind saturation has also been observed from aircraft flying over hurricanes.

When the model function developed over the moderate wind range is applied to the strong winds, an underestimation of wind speed results. Strong efforts have been made to adjust the model function (slope in Figure 4) in strong winds and to find the right channel (a combination of polarization, frequency, and incident angle) that would be sensitive to the increase of strong winds. The success would be difficult if flow separation occurs at high winds and the surface roughness and stress do not increase with winds, as discussed in the Relations between Wind and Stress section.

Other Sensors

Both the microwave altimeter and SAR are similar to the scatterometer in the sense that they are active sensors that send microwave pulses to the Earth's surface and measure the backscattered power. The microwave radiometer is a passive sensor, observing the radiance from the Earth and its atmosphere.

While the scatterometer views at oblique angles, the altimeter views at nadir (very small incident angles). At nadir, the backscattered energy is a result of specular reflection (the wavelets serve as small mirrors), and the backscatter is not sensitive to the U_N direction. Because the instrument is not scanning, data are available only at very narrow (2 km) repeated ground tracks. The coverage of all the past altimeters is poor compared with the scatterometer and the microwave radiometers.

A SAR looks perpendicular to the aircraft path at only one azimuth angle, and cannot resolve the U_N direction like the scatterometer. SAR has spatial resolutions that are much better

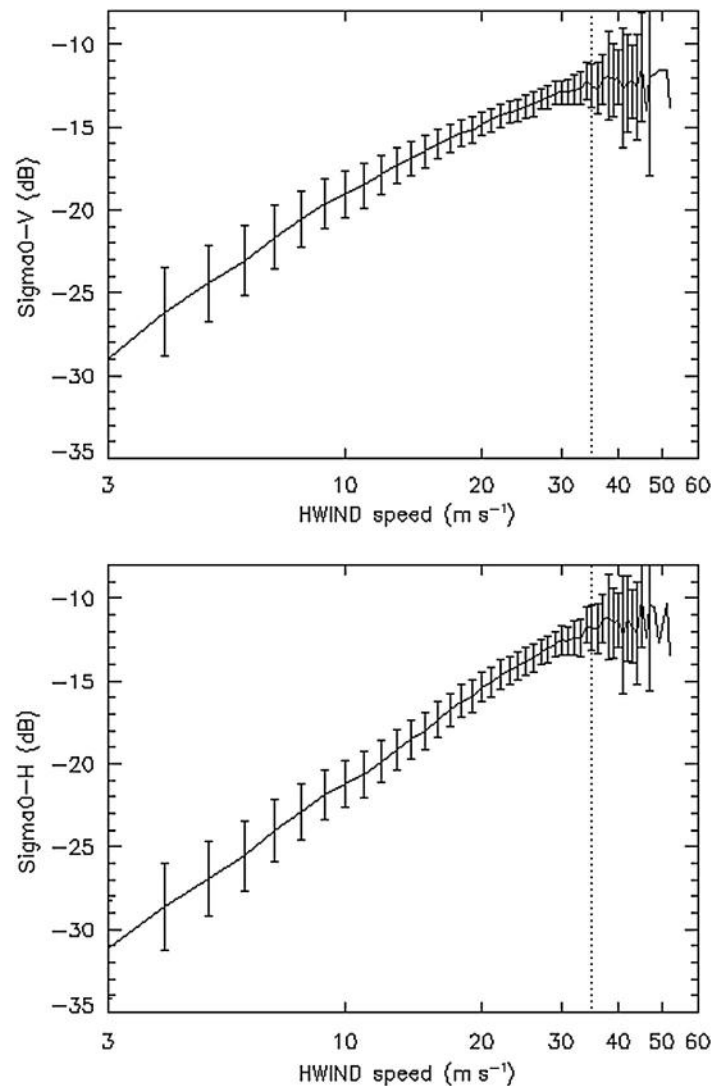


Figure 4 Normalized radar cross-section at two polarizations measured by QuikSCAT for 12 hurricanes as a function of collocated surface wind provided by the National Hurricane Center.

than those of scatterometers, but the high resolution also introduces higher uncertainties in accuracy caused by secondary effects that affect surface roughness. The instrument and the data-processing procedure are much more complicated than those of the scatterometer, and there have been serious calibration problems. The scatterometer GMF can be used to relate the σ_o measured by SAR to U_N . However, a particular value of σ_o may correspond to a range of U_N , depending on the azimuth angle. Hence, in order to retrieve U_N with the GMF, the U_N direction must first be specified. Whether the a priori direction information is derived from the orientation of kilometer-scale structure in the SAR image, or from operational NWP models, the spatial scales are much coarser than σ_o .

Ocean surface wind speed has also been derived from the radiance observed by a microwave radiometer. It is generally believed that wind speed affects the surface emissivity indirectly through the generation of ocean waves and foam. Radiometers are designed to observe how the ocean surface operates

primarily at window frequencies, where atmospheric absorption is low. Radiances at frequencies sensitive to sea surface temperature, atmospheric water vapor, and liquid water are also measured; they are used to correct for the slight interference by the atmosphere. It was demonstrated in several airborne experiments that the polarization properties of the sea surface emission vary not only as a function of the wind speed, but also as a function of wind direction. Wind direction-measuring capability has been evaluated for a polarimetric radiometer, WindSat, launched by the US Navy.

Relations between Wind and Stress

Ocean surface stress (τ) is the turbulent transfer of momentum generated by atmospheric instability caused by both wind shear (difference between wind and current) and buoyancy (vertical density stratification resulting from temperature and

humidity gradients). Direct τ measurement has been done in only a few field campaigns in the past. For all practical purposes, our knowledge of τ is derived from winds (U) at a reference height through a drag coefficient C_D , which is defined by

$$\tau = \rho C_D (U - U_s)^2 \quad [1]$$

where U_s is the surface current and ρ is the air density.

The drag coefficient C_D has been derived largely in field campaigns. **Figure 5** illustrates the behavior of C_D at neutral stability. At low wind speed ($U < 3 \text{ m s}^{-1}$), the flow is smooth; C_D increases with decreasing wind speed. And at moderate wind $3 < U < 20 \text{ m s}^{-1}$, C_D is an increasing function of wind speed for a rough sea with open fetch. Secondary factors, such as sea states and spray from breaking waves, whose data are not generally available, are not included in this parameterization scheme and should be part of the errors.

In a similar fashion, the turbulent fluxes of heat H and moisture E have been related to the mean parameters – wind speed U , potential temperature T , specific humidity Q at 10 m, sea surface temperature T_s , and the interfacial humidity Q_s (usually taken to be the saturation humidity at T_s), which are the measurements generally available from routine ship reports, through

$$H = \rho c_p C_H (T - T_s)(U - U_s) \quad [2]$$

$$E = \rho C_E (Q - Q_s)(U - U_s) \quad [3]$$

where c_p is the isobaric specific heat.

In the past, the transfer coefficients, C_H and C_E , were approximated with the same values as C_D . W.T. Liu first postulated in 1979 that, in a rough sea, under a moderate range of winds ($5\text{--}20 \text{ m s}^{-1}$), C_H and C_E do not increase with wind speed because of molecular constraint at the interface, while C_D may still increase because momentum is transported by form drag. Liu's hypothesis, as illustrated in **Figure 5**, was subsequently supported by measurements in field experiments. K. Emanuel argued in 1995, from theoretical and numerical model results, that Liu's hypothesis could not hold at the strong wind regime of a hurricane. To attain the wind strength of a hurricane, the energy dissipated by drag could not keep increasing while the energy fed by sensible and latent heat does not increase with wind speed. His argument puts limits on the increase of C_D as a function of wind speed. The postulation of

the level of the increase of C_D with wind speed at hurricane-scale winds was supported by the results of the laboratory studies and the aircraft experiments. Such flow separation may explain the saturation of scatterometer measurements at wind speeds higher than about 32 m s^{-1} , as shown in **Figure 4**.

The drag coefficient, or the bulk parameterization of stress, can be expressed as the nondimensional flux–profile relation (also called the similarity function) in the constant flux layer.

$$\frac{U - U_s}{U_*} = 2.5 \left(\ln \frac{z}{z_0} - \psi_U \right) = \frac{1}{\sqrt{C_D}} \quad [4]$$

where $U_* = (\tau/\rho)^{1/2}$ is the frictional velocity, z_0 is the roughness length, and ψ_U is a function of the stability parameter, which is the ratio of buoyancy to shear production of turbulence. Typical wind profiles at various stabilities are shown in **Figure 6**. From the zero intercept and the slope of the logarithm profile, z_0 and U_* can be determined. In general oceanographic applications, the surface current is assumed to be small compared with wind and the atmosphere is assumed to be nearly neutral. Neglecting U_s and ψ_U in eqn [4], U becomes U_N , and it is uniquely related to U_* (or τ). To compute U_N from conventional wind measurements of U (A on the blue curve in **Figure 6**), U_* and z_0 are computed as the slope and intercept at the surface of the curve in **Figure 6**. The neutral relation (straight line) defined by U_* and z_0 will give U_N (point B). This method has been used in the development and calibration of all scatterometers launched by NASA. At a given level, U_N is greater than the actual wind (U) under unstable conditions but lower under stable conditions. From eqn [4], $U_N - U = 2.5 U_* \psi_U$, assuming z_0 depends much stronger on wind shear than buoyancy and this difference is the inherent error of using scatterometer measurements as the actual wind. The formulation of ψ_U was largely based on experiment data on land, validated with only a small amount of measurement over ocean, and may have considerable uncertainties.

Potential Improvements and Conclusion

Historically, the European Space Agency used the C-band (5 GHz), but NASA prefers the Ku-band (14 GHz) in their scatterometers. The backscatter at higher frequencies is more sensitive to shorter ocean waves. The Ku-band is more sensitive

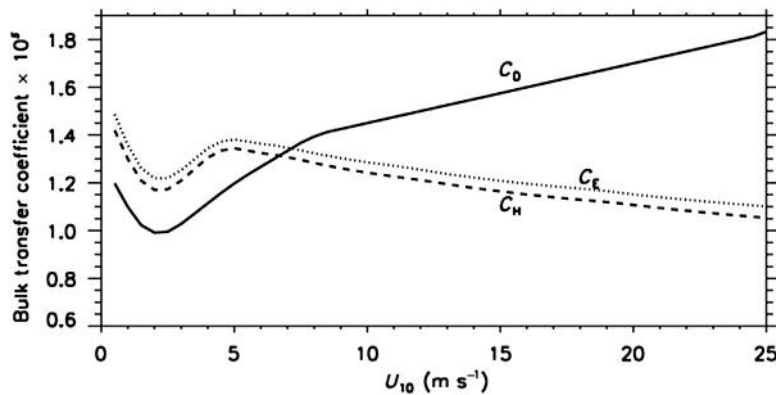


Figure 5 Variation of the bulk transfer coefficients of momentum (drag coefficient), heat, and moisture with wind speed by Liu et al. (1979).

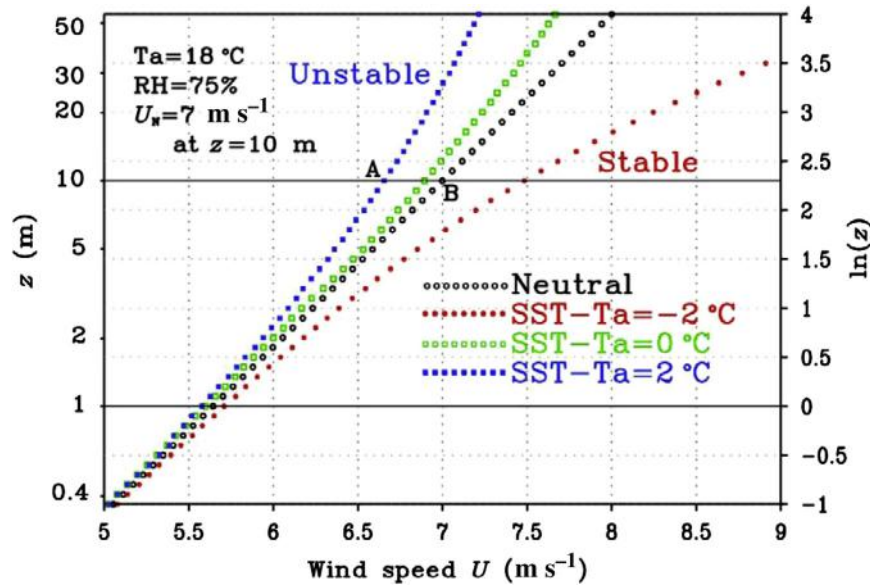


Figure 6 Typical wind profiles at various stability conditions derived from the flux–profile relation by Liu et al. (1979). B is the equivalent neutral wind corresponding to the actual wind measurement at A.

to weak wind–stress variations but is more subjective toward atmospheric effects and rain contamination. Wind retrieval at the L-band (1 GHz) has also been attempted because L-band backscatter is not sensitive to atmosphere and rain attenuation. There have been calls for a multifrequency scatterometer that is sensitive to various parts of the ocean surface wave spectrum and may reduce atmospheric and rain effects.

Present scatterometers are real-aperture systems, and the spatial resolution is limited by the antenna size. A larger antenna will, of course, enhance the spatial resolution. Another way to achieve higher resolution is to add synthetic aperture capability.

One of the drawbacks of present scatterometers is the ambiguity in retrieving wind–stress direction. The backscatter is a cosine function of the azimuth angle. In a recent experiment, it was demonstrated that the correlation between copolarized and cross-polarized backscatter of radiance is a sine function of the azimuth angle. By adding polarized measurement capabilities to the scatterometer, the directional ambiguity problem could be mitigated.

One polar-orbiting scatterometer at a low-altitude (e.g., 800 km) orbit can sample at a location on Earth not more than two times a day. Additional instrument flying in tandem will allow descriptions of higher temporal variability and the reduction of the aliasing (bias introduced by subsampling) of the mean wind–stress.

Not all space-based ocean surface wind and stress measurements are comparable in quality. Standardizing the technology requirements for observation accuracy with different research and operational applications and for international cooperation is very desirable. Many scientific reports have affirmed the need for high-quality, continuous, and consistent long time series of ocean surface vector winds and stress.

Acknowledgment

This article was prepared at the Jet Propulsion Laboratory, California Institute of Technology, under contract with the National Aeronautics and Space Administration (NASA). It was supported by the Ocean Vector Winds and the Physical Oceanography Programs of NASA. Xiaosu Xie and Wenqing Tang provided valuable assistance.

See also: Air Sea Interactions: Momentum, Heat, and Vapor Fluxes. **Boundary Layer (Atmospheric) and Air Pollution:** Surface Layer.

Further Reading

- Barale, V., Gower, J.F.R., Alberolanza, L., 2010. *Oceanography from Space*. Springer, Dordrecht. pp. 93–111 (Chapter 6).
- Emanuel, K., 1995. Sensitivity of tropical cyclones to surface exchange coefficients and a revised steady state model incorporating eye dynamics. *Journal of the Atmospheric Sciences* 52, 3969–3976.
- Gower, J., 2006. *Remote Sensing of the Marine Environment, Manual of Remote Sensing*, third ed. vol. 6. American Society for Photogrammetry and Remote Sensing. pp. 149–178 (Chapter 5).
- Liu, W.T., Katsaros, K.B., Businger, J.A., 1979. Bulk parameterization of air–sea exchanges in heat and water vapor including the molecular constraints at the interface. *Journal of the Atmospheric Sciences* 36, 1722–1735.
- Liu, W.T., Xie, X., Niiler, P.P., 2007. Ocean–atmosphere interaction over Agulhas extension meanders. *Journal of Climate* 20, 5784–5797.
- Muyeen, S.M., 2010. *Wind Energy Conversion Systems*. Springer-Verlag, London. pp. 367–383 (Chapter 15).

Temperature Soundings

A Dudhia, University of Oxford, Oxford, UK

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

Atmospheric temperature can be derived from satellite measurements using a variety of viewing geometries and wavelengths, exploiting emission, absorption or scattering processes which depend directly on temperature or indirectly via atmospheric density.

Operational temperature sounders, flown since the 1970s, are nadir-viewing instruments which sense thermally emitted radiation in the infrared or microwave regions and use the spectral variation of atmospheric absorption to sound different depths. Limb-viewing allows profile information to be obtained directly from the viewing geometry but also requires the simultaneous retrieval of pressure. Occultation instruments usually measure density profiles from which temperature can be inferred.

Introduction

Temperature plays a key role in radiative, dynamical, and chemical processes in the atmosphere. However, compared to most other parameters, atmospheric temperature has a relatively low variability: typically ± 20 K, or about $\pm 10\%$ of the absolute value, at any altitude (Figure 1). This low variability imposes correspondingly tight constraints on the useful accuracy of any measurements. Nevertheless, remote sounding has now developed to the point where temperature can be retrieved with accuracies of 2 K or better, comparable with the quality of measurements made *in situ* using radiosondes. The main impetus for this development has come from the meteorological community: although radiosondes provide good coverage over populated land areas, accurate weather forecasting requires global temperature fields, which can only be obtained from satellites.

'Operational' temperature sounders have been flown on the NOAA series of polar orbiting satellites since 1972 (augmented since 2006 by the European MetOp satellites and eventually to be superseded by the JPSS satellites). These are nadir viewing instruments, measuring emission in the infrared and microwave regions of the spectrum. By selecting channels sensitive to

emissions from different depths into the atmosphere the vertical temperature structure can be determined. Nowadays, the more sophisticated forecast models assimilate directly the satellite radiance measurements themselves, bypassing the need for any explicit temperature profile retrieval.

Temperature can also be retrieved using emission measurements from the atmospheric limb, i.e., viewing the atmosphere tangentially rather than vertically. Limb sounding allows temperature to be retrieved to higher altitudes and with improved vertical resolution compared to nadir sounding, but at the expense of reduced horizontal resolution and signal-to-noise ratio. A third possibility for temperature sounding is solar occultation: viewing the Sun as it rises or sets beyond the atmospheric limb and determining temperature through its effect on the atmospheric absorption.

Such measurements all rely on modeling molecular absorption spectra, which largely determine radiative transfer at infrared and microwave wavelengths. However, since temperature and density are linked via the hydrostatic equation, the temperature profile may also be inferred from measurements which are more directly related to the atmospheric density profile, such as scattering in visible and UV wavelengths, or refraction at radio frequencies. Molecular (Rayleigh) scattering is routinely used to determine atmospheric temperature from ground-based lidars, and atmospheric composition from space, but has had only limited application in temperature sounding from space. Radio occultation techniques for determining density via refraction were originally developed for sounding the atmospheres of other planets; however, the GPS network of navigational satellites is now also used to provide routine temperature soundings of the Earth's atmosphere.

Physical Mechanisms

Atmospheric temperature and density affect electromagnetic radiation through absorption, emission, refraction, and scattering. All of these mechanisms can be exploited in order to retrieve temperature using remote sensing techniques.

Thermal Radiation

Most techniques for temperature sounding rely on measurements of thermally emitted radiation in either the infrared or

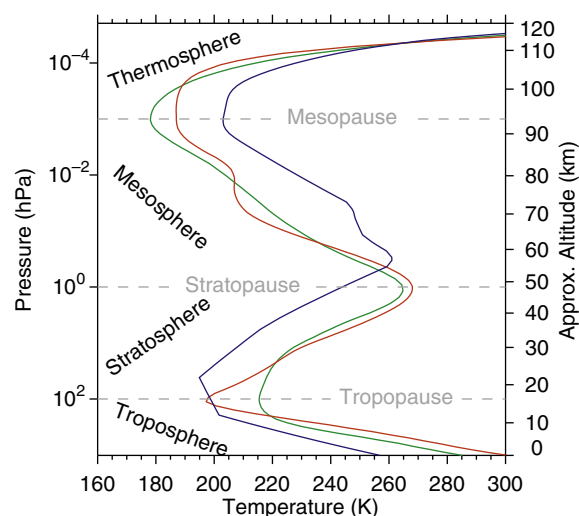


Figure 1 Typical atmospheric temperature profiles for midlatitude (green), equatorial (red), and polar winter (blue) conditions.

microwave region of the spectrum. The monochromatic radiance, I , from a line of sight through a nonscattering atmosphere in local thermodynamic equilibrium (LTE) can be represented by

$$I = \int_{\infty}^0 B \frac{d\tau}{ds} ds + I_{\infty} \tau_{\infty} \quad [1]$$

where B is the Planck function, τ the transmittance from space to a point at distance s along the path, and τ_{∞} represents the attenuation of any emission source I_{∞} beyond the atmosphere. The atmospheric contribution to this radiance is therefore a spatially weighted average of the Planck function along the path, $d\tau/ds$ being the 'weighting function.' Since B is a known function of temperature and wavelength, determining $B(T(s))$ from the above relationship is equivalent to retrieving the temperature profile. However, to use eqn [1] it is also necessary to know the transmittance $\tau(s)$ along the path, conveniently expressed in terms of an optical thickness χ :

$$\tau = \exp(-\chi) \quad [2]$$

$$\chi = \int_0^s \nu \sigma \rho ds \quad [3]$$

where ν is the absorber volume mixing ratio, ρ is the (molar) air density, and σ is the absorption coefficient ($\text{m}^2 \text{mol}^{-1}$).

At thermal wavelengths, σ is a function of the concentrations of various absorbing species, pressure and temperature. Composition ν can be eliminated as an unknown by selecting spectral regions where the absorption is primarily from a well-mixed species, usually the $15 \mu\text{m}$ CO_2 band in the infrared or the 60 GHz O_2 band in the microwave. Pressure is either implicit in the retrieval coordinates (nadir sounding) or retrieved simultaneously with temperature (limb sounding).

Planck Function

The Planck function (Figure 2) is given by:

$$B = \frac{2hc^2}{\lambda^5 (\exp(hc/\lambda kT) - 1)} \quad [4]$$

where h is Planck's constant, c is the speed of light, k is Boltzmann's constant, λ is the wavelength, and T is the temperature. This has a maximum at $\lambda_{\text{max}} = 2.9 \times 10^{-3}/T \text{ m}$ (Wien's displacement law), which for typical atmospheric temperatures lies between 10 and $15 \mu\text{m}$. The curve falls off as $1/\lambda^4$ at longer wavelengths, and as $\exp(-hc/\lambda kT)$ at shorter wavelengths. The rapid decay at short wavelengths means that thermal emission is negligible below a few microns and, in any case, below $4 \mu\text{m}$ scattered or reflected solar radiation becomes significant during the daytime.

The accuracy of a temperature retrieval depends not only on the radiance signal-to-noise ratio (SNR) but also on the sensitivity of radiance to temperature, dB/dT , which reduces with increasing wavelength. At $10 \mu\text{m}$, $B \approx T^6$ but in the microwave region ($1\text{--}10 \text{ mm}$ wavelengths) $B \propto T$ (in fact, 'radiance' and 'temperature' are often used interchangeably when referring to microwave measurements). For a given SNR, this means that temperature can be retrieved more accurately at shorter wavelengths. Conversely, this also means that at shorter wavelengths retrievals of other species are more sensitive to any given

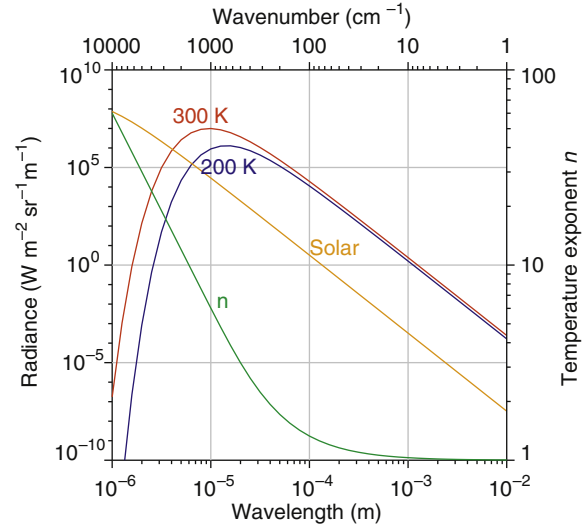


Figure 2 Red and blue lines show the Planck function evaluated at 300 and 200 K, representing the typical range of atmospheric temperatures. The orange line shows the Planck function for 5800 K, scaled to give an integrated irradiance of 1370 W m^{-2} (solar constant) representing the maximum diffuse solar contribution. The green line (right axis) shows the sensitivity n fitting $B \sim T^n$, evaluated at $T = 240 \text{ K}$. The wavenumber axis (top) is proportional to frequency, 1 cm^{-1} corresponding to approximately 30 GHz .

temperature error. For example, a number of constituents (H_2O , CH_4 , N_2O , NO_2) are often retrieved using bands in the $6\text{--}8 \mu\text{m}$ region. At these wavelengths, a 1 K temperature overestimate results in a 4% increase in predicted radiance, and therefore a roughly equivalent underestimate in retrieved concentrations. Hence the need for accurate temperature retrievals for all remote sensing experiments based on infrared emission.

Molecular Absorption

The molecules of many atmospheric species exhibit absorption bands at infrared wavelengths corresponding to transitions between quantized vibrational energy levels (molecules of nitrogen and oxygen being notable exceptions, having no permanent electric dipole moment). Superimposed on these vibrational states are rotational states which have a finer quantization. Changes in vibrational level are often accompanied by changes in the rotational quantum number J , giving the typical band structure shown in Figure 3. The central peak (Q-branch) corresponds to a pure vibrational transition ($\Delta J = 0$), and the envelope of lines at lower (P-branch) and higher (R-branch) wavenumbers correspond to $\Delta J = \pm 1$, the spread in these envelopes reflecting the rotational energy dependence $E_R \approx J(J+1)$. Pure rotational transitions also occur, leading to absorption features in the microwave spectrum (Figure 4).

The overall strength of a particular absorption band is determined by the population of the two vibrational levels involved. In LTE, populations of each vibrational energy level E_V follow the Boltzmann distribution $\sim \exp(-E_V/kT)$. For $T = 240 \text{ K}$, $kT \ll E_V$ (167 cm^{-1} compared to $\sim 1000 \text{ cm}^{-1}$) so usually only 'fundamental' transitions between the ground state and the first excited level are significant.

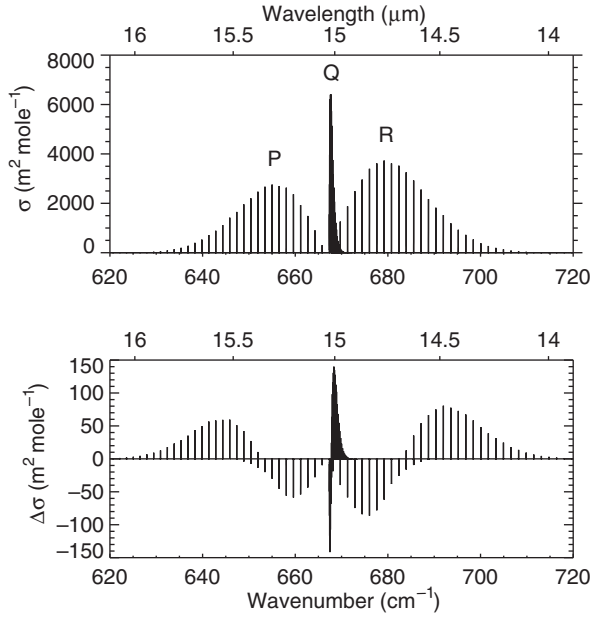


Figure 3 Top: The absorption coefficient σ of the CO_2 15 μm fundamental ν_2 vibration band at 10 hPa total pressure, 240 K, displaying the P, Q, and R branches. Bottom: The change $\Delta\sigma$ resulting from a 10 K increase in temperature. (The atmospheric CO_2 column amount is around 130 mol m^{-2} .)

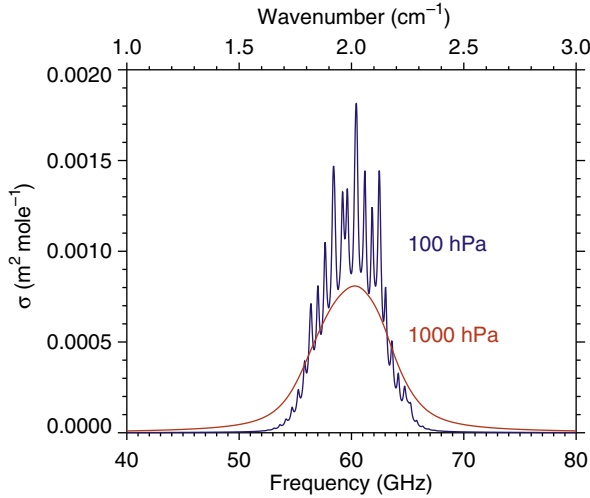


Figure 4 The absorption coefficient σ of the O_2 rotational band at 60 GHz, 1000 hPa (red), and 100 hPa (blue). (The atmospheric O_2 column amount is around $7.3 \times 10^4 \text{ mol m}^{-2}$.)

The shape of an absorption band, given by the envelope encompassing the P and R branches, depends on the population distribution over the rotational energy levels E_R . This also follows Boltzmann statistics but here $kT \gg E_R$. The additional degeneracy factor $(J + 1)$, representing the increased number of states available at higher rotational quantum numbers, ensures that the most probable rotational quantum number is not $J = 0$ but some higher number, and increases with increasing temperature (indicated by the outward displacement of the peaks of the P- and R-branches in Figure 3).

As well as the line strength and the shape of bands, the temperature–pressure profile also affects the width of individual lines through Doppler broadening (high altitudes) and pressure broadening (low altitudes, Figure 4).

The above assumes LTE, i.e., that the populations of the vibrational and rotational energy levels are characterized by the same temperature as the mean kinetic energy. In practice, this means that collisions between molecules occur sufficiently frequently to ensure that the internal energy levels are redistributed according to the local kinetic temperature. At high altitudes, other processes may dominate leading to nonthermal population distributions of the vibrational states (so-called ‘non-LTE’ effects). These often limit the upper altitude of practical infrared retrieval schemes. Non-LTE effects are usually negligible in the microwave region due to the small energy difference between rotational levels; instead, when sounding the mid-stratosphere or higher altitudes, complications are introduced by having to model the Zeeman splitting of lines in the Earth’s magnetic field.

Hydrostatic Equation

In most situations, the profiles of atmospheric pressure p , molar density ρ , and temperature T can be assumed to be in hydrostatic balance:

$$dp = -gM\rho dz \quad [5]$$

$$\frac{dp}{p} = -\frac{gM}{RT} dz \quad [6]$$

where g is the gravitational acceleration, M the molar weight of air, R the gas constant, and z the altitude.

In addition to retrieving temperature, it is usually necessary to know the pressure in order to evaluate terms in eqn [3], although absolute altitude is not critical. If $T(p)$ is retrieved directly (e.g., nadir sounding) this is not a problem and eqn [6] can be integrated to obtain layer thicknesses Δz . If $p(z)$ is retrieved (e.g., microwave limb sounding), eqn [6] can be used to obtain temperature. If $T(z)$ is retrieved (e.g., infrared limb sounding), it is also necessary to retrieve pressure at least at one altitude, $p(z_{\text{ref}})$, and obtain $p(z)$ by integrating eqn [6]. If $\rho(z)$ is retrieved (e.g., radio occultation, limb scattering), it may be adequate to assume some climatological pressure at high altitude and integrate eqn [5] downward to obtain $p(z)$, hence $T(z)$, with any error δp in the climatological assumption decreasing in significance further down the profile.

Refraction

The speed of electromagnetic radiation is reduced in air due to the polarizability of air molecules. The refractive index n is conveniently expressed in terms of a refractivity $N = n - 1$, which is usually assumed to be proportional to air density. For dry air at 15 °C and standard pressure, Edlén’s dispersion relation models the wavelength dependence of refractivity from 0.2 to 2 μm (Figure 5):

$$N \times 10^6 = 64.328 + \frac{29498.1}{146 - \mu^{-2}} + \frac{255.4}{41 - \mu^{-2}} \quad [7]$$

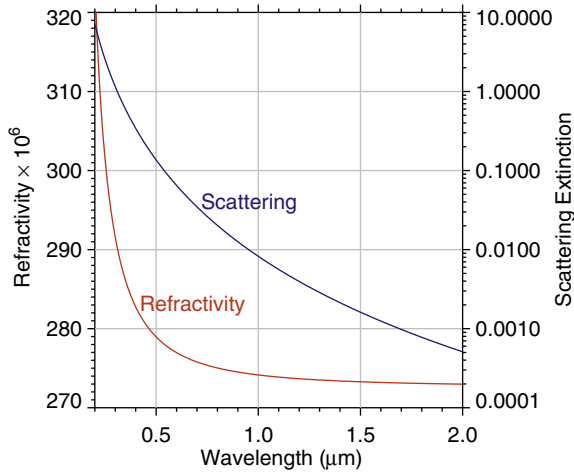


Figure 5 Variation of refractivity (red, left axis) and Rayleigh scattering cross section (blue, right axis) in the visible and near-infrared region of the spectrum. The scattering extinction is scaled to the number of molecules in a vertical column of atmosphere.

where μ is the wavelength expressed in μm . For wavelengths longer than $1 \mu\text{m}$, refractivity is essentially independent of wavelength and approaches a value of 2.73×10^{-4} . However, at radio frequencies ($<20 \text{ GHz}$) the dipole moment of water vapor has a significant effect, which can be modeled as:

$$N \times 10^6 = 77.6 \frac{(p - e)}{T} + 3.73 \times 10^5 \frac{e}{T^2} \quad [8]$$

where p and e are the total pressure and partial pressure of water vapor in hectopascals, respectively, and T is the temperature in Kelvin. At 15°C and standard pressure, this gives values of refractivity varying from $N = 2.73 \times 10^{-4}$ for dry air (i.e., the long wavelength limit of eqn [7]) to $N = 3.41 \times 10^{-4}$ for saturated air (1.6% water vapor by volume).

Refraction introduces a curvature into limb paths, lowering the tangent point and increasing the path-integrated air mass compared with the straight-line path. For a circularly symmetric atmosphere, it can be shown that the tangent height correction is:

$$\delta z \approx Na \quad [9]$$

where a is the radius of the Earth. For a tangent height of 25 km , $\delta z \approx 70 \text{ m}$ for infrared wavelengths, equivalent to a 1% increase in the tangent point pressure and a similar increase in integrated air mass. Since the effect is proportional to density ($\propto N$), it doubles for every $\sim 5 \text{ km}$ decrease in tangent height.

Scattering

For visible and shorter wavelengths, 'Rayleigh' scattering by air molecules becomes significant. The Rayleigh scattering cross section σ_{Ra} ($\text{m}^2 \text{mol}^{-1}$; cf. σ in eqn [3]) can be computed theoretically:

$$\sigma_{Ra} = \frac{32\pi^3}{3N_A \rho^2 \lambda^4} N^2 \quad [10]$$

where N_A is Avogadro's number. Since refractivity $N \propto \rho$, the Rayleigh scattering cross section per mole of air depends only on wavelength, through the $1/\lambda^4$ term (Figure 5).

For optically thin paths, single scattering can be assumed so that measurements of extinction or scattered radiation can be simply related to the air density. However, this assumption breaks down at higher pressures, when multiple scattering and/or Mie scattering (by particles of radius comparable to the wavelength, e.g., aerosols) have to be considered.

Nadir Sounding

The earliest satellite temperature sounders viewed downward, measuring the radiance emerging from the top of the atmosphere in a range of spectral bands. The different transmission characteristics of each band can be used to derive information on temperature from different optical depths into the atmosphere. This is still the basis of most of the operational temperature sounders used today.

Weighting Functions

Adapting eqn [1], the radiance I emerging from the top of the atmosphere above a nonreflective surface is given by:

$$I = B_0 \tau_0 + \int_0^\infty B \frac{d\tau}{dZ} dZ \quad [11]$$

where $Z = -\ln(p/p_0)$ is a heightlike coordinate, and subscript 0 indicates surface values. For nadir viewing, it is convenient to use a pressure-based coordinate such as Z since the transmittance, and therefore weighting functions are themselves mostly pressure dependent.

Using the hydrostatic equation (eqn [5]) to adapt eqn [3] to pressure coordinates, the optical depth χ from space to pressure level p for a well-mixed absorber with volume mixing ratio v and constant absorption coefficient σ is given by:

$$\chi = \int_p^0 \frac{v\sigma}{g} dp = \alpha p \quad [12]$$

where $\alpha = v\sigma/g$ is a constant. The weighting function is then given by:

$$\frac{d\tau}{dZ} = -p \frac{d\tau}{dp} = \alpha p \exp(-\alpha p) \quad [13]$$

It can be shown that this has a maximum where the optical depth $\chi = \alpha p = 1$ and a width at half maximum of $\Delta Z \approx 2.5$ scale heights ($\sim 15 \text{ km}$).

By suitable placement of filters within the band it is possible to select weighting functions peaking at different pressures (Figures 6 and 7).

Vertical Resolution

From eqn [13] and Figure 7, it can be seen that the width of nadir sounding weighting functions is comparable with the thickness of the entire troposphere. The weighting function width does not fundamentally limit the vertical resolution of the retrieval, but the large overlap means that in order to retrieve a profile at, say, seven levels corresponding to the weighting function peaks, most of the information will come from the difference between radiance measurements in adjacent channels rather than from the absolute values, hence reducing the effective SNR.

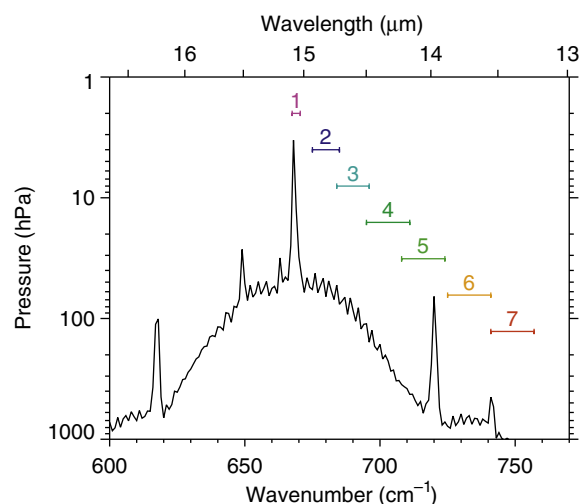


Figure 6 The CO₂ 15 μm absorption band showing the pressure level for which optical depth = 1 (i.e., from which transmittance to the top of the atmosphere is e^{-1}). The spectrum is averaged over 1 cm⁻¹ intervals. Also shown are the positions of HIRS/3 channels 1–7 (see Figure 7).

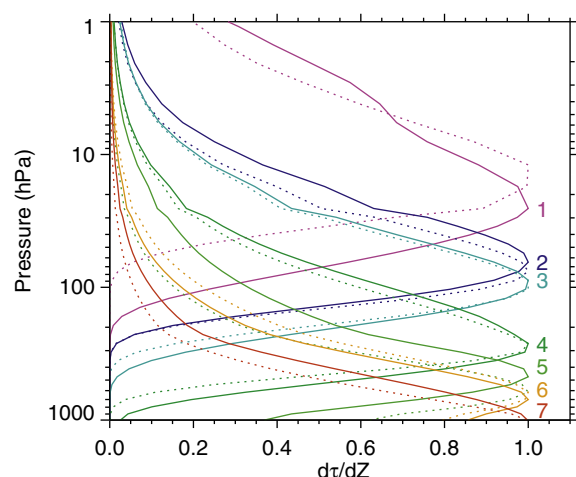


Figure 7 Solid lines: the weighting functions for HIRS/3 channels 1–7 (see Figure 6). These are normalized so that the maximum atmospheric contribution is 1. Also plotted (dotted lines) are the equivalent weighting functions for channels of width 1 cm⁻¹.

To produce narrower weighting functions requires finding spectral regions where the optical depth χ increases more rapidly with pressure than $\chi \propto p$ (eqn [12]). One method is to select spectral regions where emission is predominantly from the wings of pressure-broadened lines (giving $\sigma \propto p$, hence $\chi \propto p^2$). Another method is to target emission from a gas whose concentration increases with pressure, such as tropospheric water vapor. Since the absorber is no longer well mixed, additional channels are required in order to retrieve its concentration, and the weighting function peaks are no longer at fixed pressures.

A different method of improving the vertical resolution is to scan at an angle θ to nadir, increasing the optical path to approximately $\chi = \alpha p \sec \theta$, reducing both the peak pressure

and width by a factor $\cos \theta$. Taken to its extreme, this is, of course, the basis of limb sounding. Scanning to $\pm 50^\circ$ across the orbit track is commonly employed for nadir sounders, but this is done in order to cover the atmosphere between adjacent orbit tracks, and a ‘correction’ applied in order to remove the resulting variation of the weighting functions.

Nadir Sounding Instruments

Table 1 lists nadir viewing instruments that have been used for temperature sounding. The first such instruments were infrared filter radiometers targeting various parts of the CO₂ 15 μm band, a simple technique which is still in use on operational satellites (Figure 6). However, such filters are limited to a minimum width of several wavenumbers which does not allow much scope for improving vertical resolution or extending coverage to higher altitudes.

More recently, instruments have been developed to measure the full infrared spectrum at high resolution: 1 cm⁻¹ for AIRS (grating spectrometer), 0.5 cm⁻¹ for IASI (interferometer), and 0.1 cm⁻¹ for TES (interferometer, in nadir viewing mode). As illustrated in Figure 7, the weighting functions associated with these narrower channels are not significantly different in shape to those of High-Resolution Infrared Radiation Sounder (HIRS). In principle, the high resolution allows selection of spectral regions with pressure-broadened line wings and/or water vapor lines for a joint retrieval with temperature, but the main improvement in vertical resolution is achieved simply by combining a much larger ($\sim$ hundreds) number of channels.

Figure 6 suggests that 4 hPa is about the highest level that can be sounded using the 15 μm band with 1 cm⁻¹ resolution. However, Figure 8 demonstrates that emissions can be detected from higher levels, but in order to discriminate these it is necessary to resolve individual lines. Doppler-broadened line widths at the stratopause are of the order of 0.001 cm⁻¹, well beyond the resolution obtainable using spaceborne interferometry, and even were such a resolution attainable the reduced photon flux from such a narrow bandwidth would lead to SNR problems.

Gas correlation radiometry is one technique which has been used to extend the altitude range of infrared nadir sounding. By passing the signal through a pressure-modulated cell of CO₂ a synchronous component of the signal can be extracted corresponding to emission in just the modulated regions of the cell transmittance spectrum. The mean cell pressure determines the region of the atmosphere being sounded by this modulated component. Since this also integrates the signal over all lines within the filter band, it gives an improved SNR compared to a single narrow bandwidth measurement. This was the principle used in the Stratospheric Sounding Unit (SSU) that provided the stratospheric sounding channels for the TIROS Operational Vertical Sounder (TOVS) instruments.

The main advantage of sounding in the microwave rather than infrared region is that clouds are transparent at millimeter wavelengths. Spectral selection for microwave instruments is achieved by radio, rather than optical, techniques. Heterodyne mixing is used to combine the atmospheric microwave signal with a local oscillator (LO) at some central frequency ($\sim$ GHz). Since the mixing process is nonlinear, an ‘intermediate frequency’ ($\sim$ MHz) signal is produced

Table 1 Satellite nadir sounding temperature sensors (mid-2013)

Launch	Satellite	Instrument	Technique ^a
1969, 1970	Nimbus 3, 4	SIRS	Satellite Infrared Spectrometer
		IRIS	Infrared Interferometer Spectrometer
1970, 1972	Nimbus 4, 5	SCR	Selective Chopper Radiometer
1972–76	NOAA 2–5	VTPR	^b Vertical Temperature Profile Radiometer
1972	Nimbus 5	ITPR	Infrared Temperature Profile Radiometer
		NEMS	Nimbus E Microwave Spectrometer
1975	Nimbus 6	HIRS	High-Resolution Infrared Radiation Sounder
		PMR	Pressure Modulated Radiometer
		SCAMS	Scanning Microwave Spectrometer
1974–94	NOAA 6–14	TOVS	<i>TIROS^b Operational Vertical Sounder, comprising:</i>
		HIRS/2	High-Resolution Infrared Radiation Sounder/2
		SSU	Stratospheric Sounding Unit
		MSU	Microwave Sounding Unit
1976–82	DMSP F1–F6	SSH	Special Sensor H
1980–96	GOES 4–7	VAS	VISSR Atmospheric Sounder
1983–99	DMSP F7–15	SSM/T, T2	Special Sensor Microwave/Temperature
2003–09	DMSP F16–18	SSMIS	Special Sensor Microwave Imager/Sounder
1994–2010	GOES 8–15		GOES Sounder
1998–2009	NOAA 15–19	ATOVS	<i>Advanced TOVS^b, comprising:</i>
		HIRS/3, 4	High-Resolution Infrared Radiation Sounder/3, 4
		AMSU	Advanced Microwave Sounding Unit
2002	Aqua	AIRS	Atmospheric Infrared Sounder
2004	Aura	TES	Tropospheric Emission Spectrometer ^c
2006, 2012	Metop-A, B	IASI	Infrared Atmospheric Sounding Interferometer
		ATOVS	Advanced TOVS
2011	Suomi NPP	CrIS	Cross-Track Infrared Sounder
		ATMS	Advanced Technology Microwave Sounder

^aFR = Filter Radiometer, MI = Michelson Interferometer, GC = Gas Correlation Radiometer, MW = Microwave Radiometer, and GS = Grating Spectrometer.

^bSee also Table 2.

^cSee also Table 3.

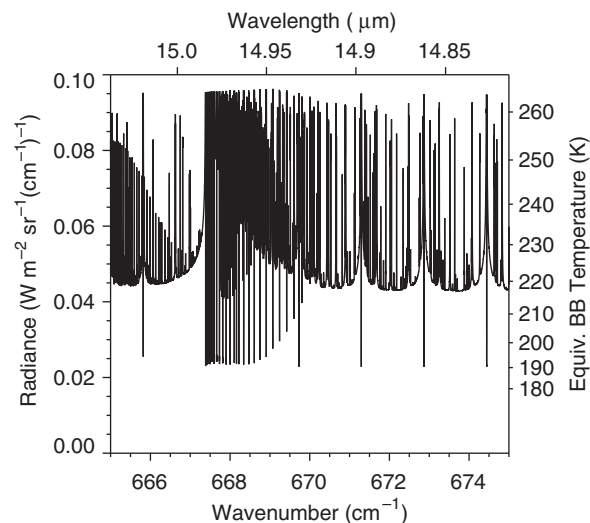


Figure 8 Atmospheric radiance spectrum (nadir view) calculated near the center of the CO₂ 15 μm band. The smooth envelope around 220 K corresponds to emission from the pressure-broadened line wings in the lower stratosphere (Figure 1), the upward spikes from Doppler-broadened lines at the stratopause (260 K), and the downward spikes from centers of strong lines near the mesopause (190 K).

corresponding to the difference between the two input signals. The result is to convert the atmospheric spectrum immediately above the LO frequency from microwave to radio frequencies, with the mirror image of the atmospheric spectrum below the LO frequency also superimposed. The spectral features can then be resolved with radio frequency filters. The technology has now developed to the point where it is possible to achieve adequate SNR in bandwidths comparable to a stratospheric line width, allowing sounding up to the stratopause (Figures 9 and 10).

Operational Temperature Sounders

The National Oceanic and Atmospheric Administration (NOAA) began routine atmospheric temperature sounding measurements (Table 2) with the Vertical Temperature Profile Radiometer (VTPR) instruments on board the NOAA 2–5 satellites which operated from 1972 to 1979. These were infrared radiometers with six temperature sounding channels from 13 to 15 μm, plus a water vapor channel at 18 μm, and another channel in the 11 μm atmospheric window.

The VTPR was superseded by the TOVS suite, first flown on the TIROS-N satellite in 1978 and subsequently on the NOAA 6–14 satellites. TOVS consisted of three instruments:

- HIRS/2, a development of the HIRS instrument originally flown on Nimbus 6. This was a 20-channel infrared radiometer with 12 temperature sounding channels covering

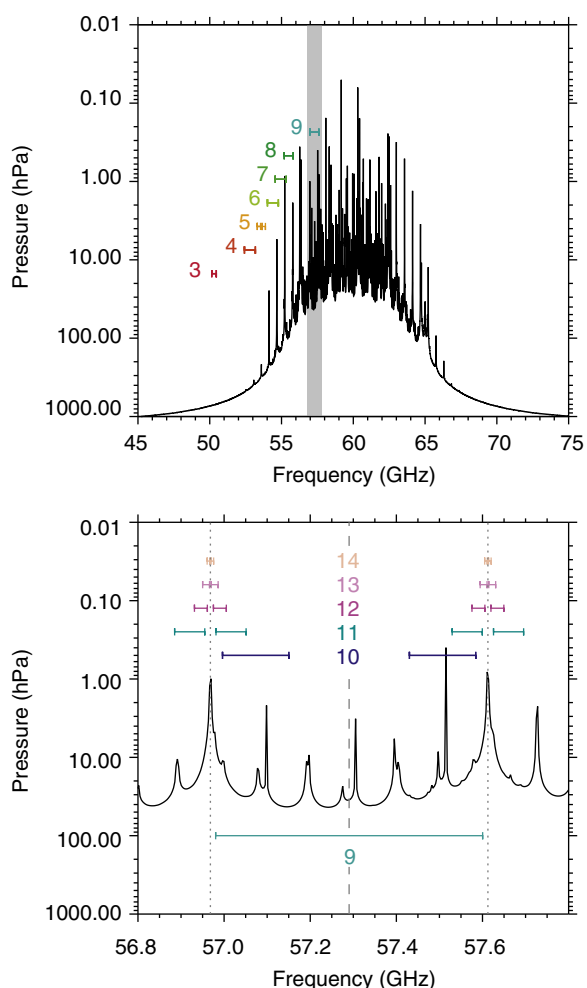


Figure 9 The O₂ 60 GHz absorption band (upper) showing the pressure at which the optical depth equals 1, and (lower) a close-up of the region around the AMSU/A 57.29 GHz LO frequency (marked by the dashed line). Also shown are the positions of AMSU/A channels 3–14. Note the use of both sidebands for channels 5 and 10, and four-sideband combinations for channels 11–14, distributed about combinations of two LO frequencies (57.29 ± 0.32 GHz, marked by dotted lines). This superimposes similar spectral features into the intermediate frequency signal in order to improve the SNR. The resulting weighting functions are shown in **Figure 10**.

both the 15 μm and the 4.3 μm CO₂ bands, in addition to water vapor, ozone, and atmospheric windows.

- SSU, a development of the Pressure Modulated Radiometer (PMR) instrument also flown on Nimbus 6. This measured CO₂ emission at 669 cm^{-1} using three different pressure modulator cells (at 1.5, 5, and 15 hPa) for stratospheric temperature sounding.
- Microwave Sounding Unit (MSU), a four-channel microwave radiometer sounding the O₂ band at 60 GHz.

Advanced TOVS (ATOVS) was first flown on NOAA-15, launched in 1998, and consists of two instruments:

- HIRS/3, a 20-channel infrared radiometer with similar spectral channels to HIRS/2 (**Figures 6 and 7**). HIRS/4, with improved spatial resolution, was flown on NOAA-18 and NOAA-19.

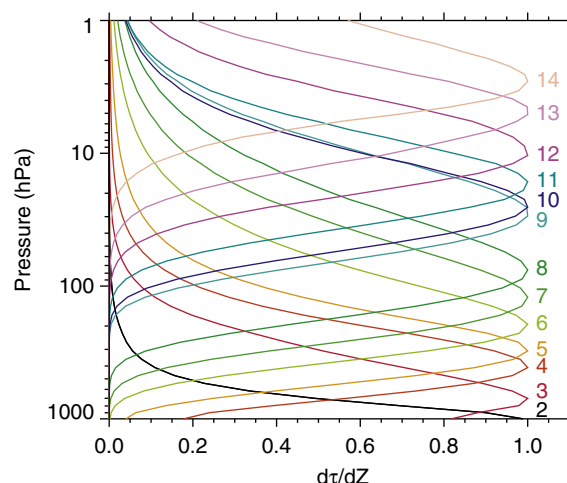


Figure 10 The AMSU/A weighting functions for channels 2–14 (channel 2 lies at 31.4 GHz, see **Figure 9** for the other channels). These are normalized so that the maximum atmospheric contribution is 1.

Table 2 The NOAA operational temperature sounders (mid-2013)

Satellite	Launch	Deactivated	Orbit
VTPR			
NOAA-2	Oct 72	Jan 75	AM
NOAA-3	Nov 73	Aug 76	AM
NOAA-4	Nov 74	Jun 86	AM
NOAA-5	Jul 76	Jul 79	AM
TOVS (HIRS/2, MSU, SSU)			
TIROS-N	Oct 78	Jan 80	PM
NOAA-6	Jun 79	Mar 87	AM
NOAA-7	Jun 81	Jun 86	PM
NOAA-8	Mar 83	Dec 85	AM
NOAA-9	Dec 84	Feb 98	PM
NOAA-10	Sep 86	Aug 01	AM
NOAA-11	Sep 88	Jun 04	PM
NOAA-12	May 91	Aug 07	AM
NOAA-13	Aug 93	Aug 93	PM
NOAA-14	Dec 94	May 07	PM
ATOVS (HIRS/3, AMSU)			
NOAA-15	May 98		AM
NOAA-16	Sep 00		PM
NOAA-17	Jun 02	Apr 13	AM
ATOVS (HIRS/4, AMSU)			
NOAA-18	May 05		PM
NOAA-19	Feb 09		PM

- Advanced Microwave Sounding Unit (AMSU), a 20-channel microwave radiometer designed for temperature and water vapor sounding (**Figures 9 and 10**). This replaced the MSU and SSU with a single microwave instrument.

Under the original Polar Operational Environmental Satellites (POES) program, NOAA aimed to maintain operational satellites in two different Sun-synchronous polar orbits: one with a southward Equator crossing at around 7.30 a.m. local time (AM orbit) and one with a northward Equator crossing at around 2.30 p.m. (PM orbit) so that coverage of any point is repeated every 6 h.

With the launch of Metop-A in 2006 (Table 1), the European Organisation for the Exploitation of Meteorological Satellites (EUMETSAT) took over responsibility for the AM orbit. This satellite, the first of three, contains both HIRS/4 and AMSU instruments (for continuity) as well as new instruments for temperature sounding: the infrared Fourier-transform spectrometer, IASI, and a GPS receiver GRAS.

After NOAA-19 (launched in 2009) the USA had planned to merge the NOAA and military DMSP programs into the National Polar orbiting Operational Environmental Satellite System (NPOESS). This led to the development of the interim NPP (NPOESS Preparatory Project) satellite, launched in October 2011 and renamed Suomi NPP. However, the NPOESS project has subsequently been abandoned and the NOAA component will continue with the Joint Polar Satellite System (JPSS) satellites in the PM orbits, while the military program continues with the Defence Weather Satellite Systems (DWSS) satellites in the AM orbits. The NPP and JPSS-1,2 satellites contain two new instruments for temperature sounding: Cross-track Infrared Sounder (CrIS), an infrared Fourier-transform spectrometer similar to IASI; and the Advanced Technology Microwave Sounder (ATMS), which is an improved version of AMSU.

Limb Sounding

Viewing tangentially through the atmospheric limb the background is cold space so that semitransparent optical paths can be used in preference to opaque paths, and consequently the weighting functions are determined by geometry rather than optical thickness. Compared with nadir viewing, limb viewing generally allows better vertical resolution and coverage to higher altitudes. However, a fundamental problem with the limb viewing geometry is that the ray paths traverse significant horizontal distances in the atmosphere (~ 200 km in the 1 km thick layer above the tangent point), which limits the scale of horizontal structures which can be resolved. Also, tropospheric limb views are more likely to be obscured by clouds than nadir views, restricting low altitude coverage using the infrared. For these reasons, limb viewing is particularly suited to temperature sounding in the stratosphere and mesosphere, while nadir sounding is preferred for the troposphere.

For a typical polar orbiting satellite at 700 km altitude, the tangent point is some 3000 km away so that 3 km at the tangent point subtends only 0.001 radians, or approximately 3' of arc. The narrow field of view reduces the radiance flux so that SNR becomes a significant problem. Diffraction is also a limiting factor for microwave instruments: angular resolution varies approximately as the ratio of antenna width to wavelength, so to resolve 0.001 radians at 2 cm (60 GHz) would require a 5 m antenna.

Weighting Functions

Neglecting the background term due to cold space, eqn [1] becomes:

$$I = \int_{-\infty}^{\infty} B \frac{d\tau}{dx} dx \quad [14]$$

where x is now the distance along the tangent path, with $x = 0$ at the tangent point and $x \approx \pm\infty$ at the satellite. Ignoring

refraction and assuming a spherical Earth of radius a , distance x , and altitude z ($\ll a$) along a path are related by

$$x^2 \approx 2a(z - z_t) \quad [15]$$

where z_t is the altitude of the tangent point. For an isothermal atmosphere at temperature T , the molar density of air, ρ , varies with altitude as:

$$\rho = \frac{p_t}{RT} \exp\left(-\frac{z - z_t}{H}\right) \quad [16]$$

where p_t is the pressure at the tangent point, and $H (=RT/gM)$, from eqn [6] the atmospheric scale height. Integrating eqn [3] along the path, assuming constant absorber volume mixing ratio ν and absorption coefficient σ , and converting to altitude coordinate $z' = z - z_t$:

$$\chi = \nu\sigma \int_{-\infty}^{\infty} \rho dx \quad [17]$$

$$= \nu\sigma \frac{p_t}{RT} \sqrt{2a} \int_0^{\infty} \frac{1}{\sqrt{z'}} \exp\left(-\frac{z'}{H}\right) dz' \quad [18]$$

$$= \nu\sigma \frac{p_t}{RT} \sqrt{2\pi aH} \quad [19]$$

If the absorption is weak then $I \approx B\chi$ and $d\tau/dz \sim -d\chi/dz$ so the characteristic shape of limb viewing weighting functions is given by:

$$\frac{d\tau}{dz} \propto \frac{1}{\sqrt{z'}} \exp\left(-\frac{z'}{H}\right) \quad [20]$$

In practice, the width of the peak is limited by the layering assumed for the retrieval. Examples, converted to temperature ($dI/dT(z)$ rather than $dI/dB(z)$), are plotted in Figure 11. Note

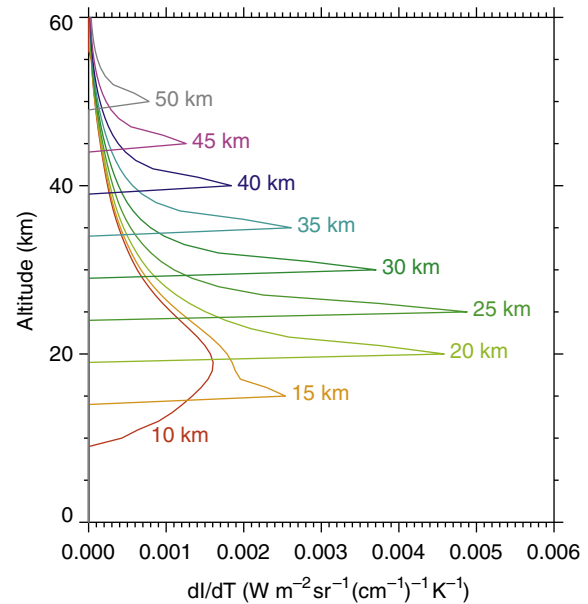


Figure 11 Temperature weighting functions for a limb viewing instrument using a 610–640 cm^{-1} filter (HIRDLS channel 3, see Figure 13). Curves show radiance response at altitudes indicated by the vertical axis to a 1 K temperature perturbation at the labeled tangent heights.

that at lower altitudes, where this spectral region becomes opaque, the weighting functions resemble those of a nadir sounder (Figure 7).

The vertical profile can be sounded either using a single detector and scanning in elevation, or by using a detector array to view the different elevation angles simultaneously. However, in practice, at least two spectral channels are required in order to retrieve pressure profile information as well.

Pressure Determination

To model the atmospheric transmittance (eqn [14]), it is also necessary to know the pressure, which is usually retrieved simultaneously with temperature (the problem does not arise in nadir sounding since the equivalent radiative transfer equation, eqn [11], is formulated in pressure coordinates). This requires radiance measurements in at least two spectral channels with different pressure–temperature characteristics.

Differences in temperature sensitivity arise from the spectral dependence of the Planck function or variations in the absorption coefficient, e.g., the changing shape of the rotational band structure (Figure 3). Some variation in pressure sensitivity may also arise from the absorption coefficient, but the biggest effect is the nonlinear variation of radiance with optical thickness: $I \approx B(1 - \exp(-\chi))$ (from eqn [14]), with $\chi \propto p$ (eqn [19]).

Figure 12 illustrates the problem graphically, with good pressure–temperature discrimination corresponding to conditions where radiance contours from any two channels intersect at a large angle. At high pressure, all radiances are independent of tangent point pressure ($I \approx B$, opaque limit). At low pressure, all radiances vary linearly with pressure ($I \approx \chi$, transparent limit). The best pressure–temperature discrimination occurs between 100 and 10 hPa (15–35 km in the atmosphere) as the channels undergo different transitions from the opaque to transparent limits.

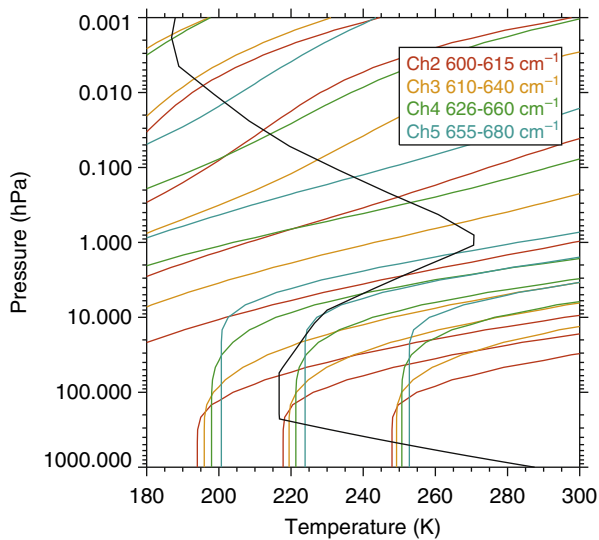


Figure 12 Radiance contours for the four HIRDLS CO₂ channels (Figure 13) for a range of pressures and temperatures simulating different tangent point conditions. The black line shows the $T(p)$ profile of the US Standard atmosphere.

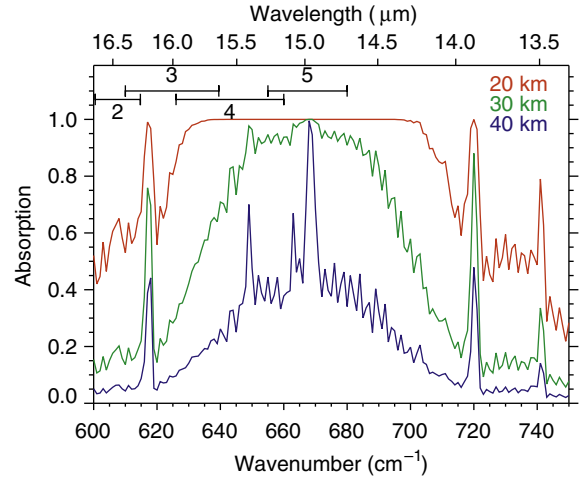


Figure 13 Absorption (1-transmittance) spectra for limb paths through different tangent heights across the CO₂ 15 μm band. Also shown are the spectral positions of the four HIRDLS temperature sounding channels (channels 2–5, see also Figure 12).

For spectrally resolving instruments such as MIPAS, it is usually possible to find optically thin spectral regions where the different temperature sensitivity of line strengths can be exploited to discriminate between tangent point temperature and pressure variations (Figure 14).

For microwave measurements, radiances can be almost independent of temperature. Assuming $I \approx B\chi$ (optically thin), and considering only the pressure- and temperature-dependent components of eqn [19]:

$$I \propto B\sigma \frac{p}{T} \sqrt{H} \quad [21]$$

For the wings of Lorentz-broadened lines, $\sigma \propto p/\sqrt{T}$, and since the scale height H and, in the microwave region, B are both proportional to T , the temperature dependences all cancel out, leaving $I \propto p^2$.

Since it is generally not possible to retrieve both temperature and pressure independently over all altitudes, it is usual to include some knowledge of the vertical distance between tangent points, either using the fixed geometry of a detector array or else pointing information from the elevation scan, and assume hydrostatic balance (eqn [6]) to constrain the problem.

Limb Emission Sounders

Table 3 lists limb emission instruments used for temperature retrievals. The various techniques for nadir viewing instruments all have their parallels in limb sounding, although with modifications driven by the different viewing geometry.

As with nadir sounding, the earliest measurements were made with infrared radiometers using the CO₂ 15 μm band. The good SNR performance achievable with broad filters means that this technique continues to be used for high spatial resolution measurements (Figure 13). Although the precise choice of filter position does not have as crucial an influence on the sounding characteristics as in the nadir viewing case, it is nevertheless desirable to choose spectral regions of intermediate absorption. Too opaque and little information is

Table 3 Satellite limb sounding temperature sensors (mid-2013)

Launch	Satellite	Instrument	Technique ^a
1976	Nimbus 6	LRIR	Limb Radiance Inversion Radiometer
1978	Nimbus 7	LIMS	Limb Infrared Monitoring of the Stratosphere
		SAMS	Stratospheric and Mesospheric Sounder
1981	SME		Solar Mesosphere Explorer
1991	UARS	ISAMS	Improved Stratospheric and Mesospheric Sounder
		CLAES	Cryogenic Limb Array Etalon Spectrometer
		HRDI	High Resolution Doppler Imager
		WINDII	Wind Imaging Interferometer
		MLS	Microwave Limb Sounder
1994, 1997	Shuttle	CRISTA	Cryogenic Infrared Spectrometers and Telescopes for the Atmosphere
2001	Odin	OSIRIS	Odin Spectrometer and IR Imaging System
		SMR	Submillimeter Radiometer
2001	TIMED	SABER	Sounding of the Atmosphere using Broadband Emission Radiometry
		TIDI	TIMED Doppler Interferometer
2002	Envisat	MIPAS	Michelson Interferometer for Passive Atmospheric Sounding
2004	Aura	HIRDLS	High Resolution Dynamics Limb Sounder
		MLS	Microwave Limb Sounder
		TES	Tropospheric Emission Spectrometer ^b

^aFR = Filter Radiometer, GC = Gas Correlation Radiometer, FP = Fabry–Perot Spectrometer, MI = Michelson Interferometer, MW = Microwave Radiometer, GS = Grating Spectrometer, and RS = Rayleigh Scattering.

^bSee also Table 1.

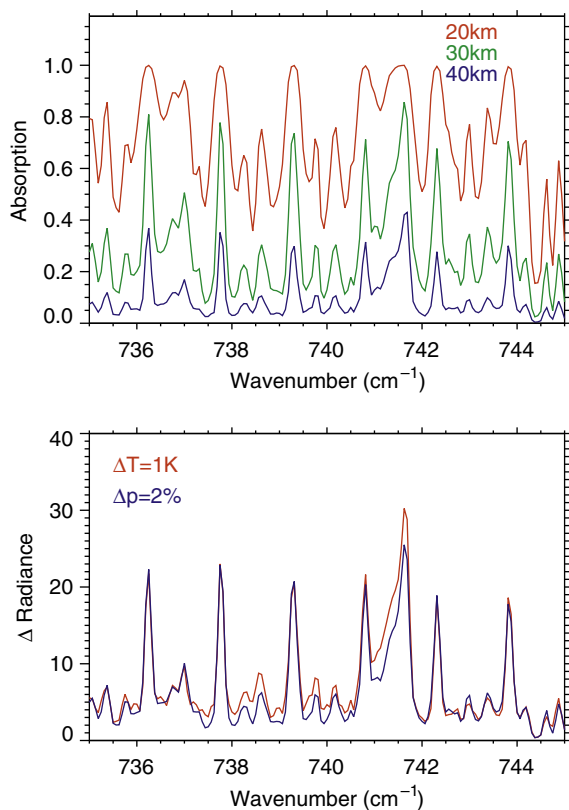


Figure 14 Upper: Absorption (1-transmittance) spectra for limb paths through different tangent heights within the CO₂ 15 μm band viewed with the MIPAS resolution of 0.0625 cm⁻¹ (the plotted region is around 13.5 μm in wavelength). Lower: The variation in the radiance (in units of nW cm⁻² sr⁻¹ (cm⁻¹)⁻¹) from 40 km tangent height with changes in the tangent point pressure and temperature. Note the change in response between 738 and 742 cm⁻¹, which can be used to discriminate between the pressure and temperature effects.

obtained from the tangent point, too transparent and the SNR is unnecessarily reduced. Apart from requiring at least two channels in order to retrieve both pressure and temperature, additional channels can be used in order to optimize the transmission characteristics for different altitude ranges.

Gas correlation and spectrally resolving instruments are usually employed in limb sounding in order to discriminate between lines of the target molecule and those of other species. However, foreign lines have only a small influence in the 15 μm CO₂ band so, for temperature sounding, the main advantage of these instruments is in improving the pressure–temperature discrimination. The main drawback is the reduced SNR associated with the narrower bandwidth and, in the case of gas correlation, the need to view perpendicularly to the orbital motion in order to avoid introducing Doppler shifts between the atmospheric lines and those in the onboard cell.

Microwave instruments have an advantage over infrared instruments in being insensitive to cloud at low altitude and to non-LTE effects at high altitude. However, vertical resolution is limited by diffraction, which can cause problems in resolving the tropopause and, at high altitudes it becomes necessary to model the Zeeman splitting of lines which varies with the Earth's magnetic field along the line of sight. Since the radiances are almost independent of temperature, these instruments effectively retrieve a pressure profile, with temperature information coming mostly from hydrostatic balance. Concomitantly, the impact of any temperature errors on the constituent retrievals is also reduced.

These measurements all rely on thermal emission from the atmosphere, but, at shorter wavelengths, Rayleigh-scattered sunlight can also be detected. Since the scattering is proportional to air density, measurements of the scattering profile can be used

to determine the temperature profile. The technique is only usable during daytime, and for relatively high altitudes where Rayleigh single scattering can be assumed, but has been used with SME and OSIRIS measurements (both grating spectrometers).

Occultation Measurements

Occultation measurements use the Sun, the Moon, or the Stars as the source of the detected radiation and monitor the change as the source rises or sets beyond the atmospheric limb (GPS occultation is discussed in a separate article). While the geometry is the same as that of limb emission measurements, the location of the tangent point is defined by ephemeris data (i.e., knowledge of the positions of the satellite, the Earth, and the source), which is usually more accurate than using the satellite attitude/pointing data which defines the tangent point for limb emission measurements. Potentially, occultation retrievals can therefore be performed on an absolute height scale. However, the relative motion of the source often means that the locus of tangent points is far from vertical, leading to 'slanted' profiles, extending over several hundred kilometers horizontally. Apart from the practical difficulties in using such profiles, the assumption of hydrostatic balance also becomes less valid.

Table 4 lists the instruments used for occultation measurements of temperature.

Solar Occultation

The Sun can be viewed through the atmospheric limb as a satellite passes between the day and night hemispheres, i.e., twice an orbit, or about 30 times in 24 h for a polar orbiting satellite.

Solar radiation, equivalent to that of a 5800 K blackbody, has a peak at visible wavelengths (Figure 2) but for temperature–pressure sounding it is necessary to use molecular features in the near-infrared such as the O₂ A-band at 0.76 μm or the CO₂ bands at 2.7 and 4.3 μm . With the Sun in the line of sight, thermal emissions from the atmosphere are negligible, so eqn [1] can be simply integrated:

$$I = B_{\text{sun}} \tau \quad [22]$$

Taking the ratio with the high-altitude radiance $I_{\text{sun}} = B_{\text{sun}}$ gives the atmospheric transmittance τ , with weighting functions $d\tau/dz$ as for limb sounding (eqn [20]).

The main advantage of solar occultation over limb emission measurements is the high SNR, allowing increased vertical and spectral resolutions, and sounding to higher altitudes. Also, since the measurements are of transmittance rather than emission, non-LTE effects are generally less significant and, by taking the ratio of radiances, the measurements are self-calibrating.

Since the absolute altitude of the tangent point is known (in practice this depends on the ability of the solar tracker to keep locked on to the same part of the solar disk), it is possible to retrieve temperature using only a single channel. Pressure information at the lowest altitude $p(z)$ obtained from meteorological fields and integrated upward.

Alternatively, transmittance spectra can be acquired at high spectral resolution using an interferometer (with the Sun as the source, SNR is less of a limitation on resolution than for emission measurements). As with limb emission measurements, spectral regions can usually be found where the pressure and temperature signatures can be distinguished, for example as shown in Figure 15.

Comparing this with the equivalent emission signatures in Figure 14, note that the temperature sensitivity is reduced relative to the pressure sensitivity since there is no significant atmospheric contribution from the Planck function (integral term in eqn [1]). In this particular example the pressure–temperature discrimination arises from the rotational structure of the R branch of a CO₂ transition, with the line strength (hence pressure sensitivity) decreasing toward higher wavenumbers, while the temperature sensitivity increases (Figure 3).

At shorter wavelengths, it is possible to determine the temperature profile (via air density) by measuring the attenuation due to Rayleigh scattering. This has been applied to SAGE II data, but measurements of molecular absorption are generally preferred.

Whichever technique is used, the main disadvantage of solar occultation is that only around 30 profiles a day can be obtained, with the sunrise and sunset profiles confined to two narrow latitude bands and no information obtainable during nighttime or the polar winter. In principle, the number of measurements could be doubled by including lunar occultation (which would also allow measurements other than at local sunrise–sunset times) but this has not been applied in practice. One instrument, GOMOS (primarily designed for high-altitude ozone measurements), uses stellar occultation: observing any

Table 4 Satellite occultation temperature sensors (mid-2013)

Launch	Satellite	Instrument	Technique ^a
1984	ERBS	SAGE II	Stratospheric Aerosol Gas Experiment II
1985	Spacelab 3	ATMOS	Atmospheric Trace Molecule Spectroscopy
1991	UARS	HALOE	Halogen Occultation Experiment
1993	SPOT-3	POAM II	Polar Ozone and Aerosol Measurement II
1996	ADEOS	ILAS	Improved Limb Atmospheric Spectrometer
1998	SPOT-4	POAM III	Polar Ozone and Aerosol Measurement III
2001	Meteor-3M	SAGE III	Stratospheric Aerosol Gas Experiment III
2002	Envisat	GOMOS	Global Ozone Monitoring by Occultation of Stars
2003	SCISAT	ACE	Atmospheric Chemistry Explorer

^aFR = Filter Radiometer, MI = Michelson Interferometer, GC = Gas Correlation Radiometer, and GS = Grating Spectrometer.

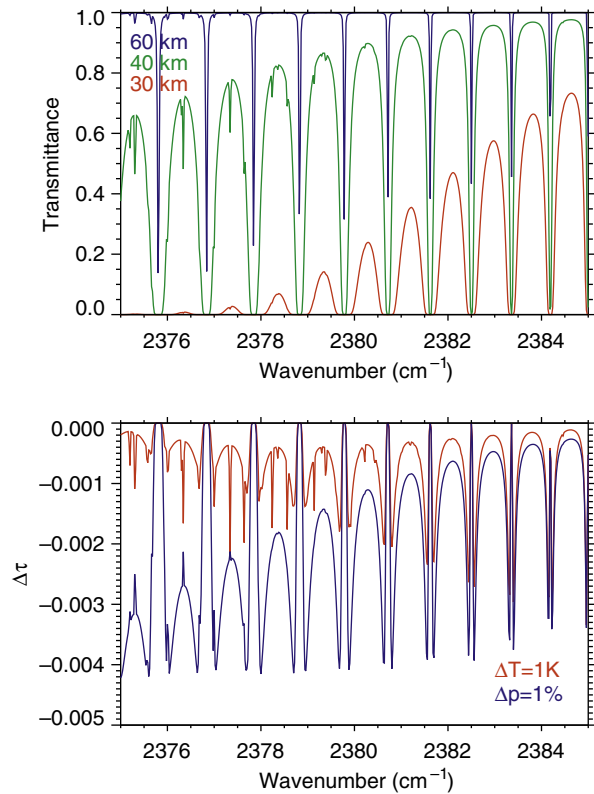


Figure 15 Upper: Transmittance spectra for limb paths through different tangent heights within the CO_2 $4.3\ \mu\text{m}$ band viewed with the ACE resolution of $0.01\ \text{cm}^{-1}$. Lower: The variation in transmittance at 40 km tangent height with changes in the tangent point pressure and temperature.

of ~ 100 bright stars as they rise or set through the atmosphere. This gives near global coverage in 24 h, although at the expense of much reduced SNR.

See also: **Chemistry of the Atmosphere:** Observations for Chemistry (Remote Sensing): IR/FIR (Satellite, Balloon and Ground); Observations for Chemistry (Remote Sensing): Microwave. **Radiation Transfer in the Atmosphere:** Absorption and Thermal Emission; Non-Local Thermodynamic Equilibrium; Scattering. **Satellites and Satellite Remote Sensing:** GPS Meteorology; Orbits; Research.

Further Reading

- Barnett, J.J., 1987. Satellite-borne measurements of middle-atmosphere temperature. *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences* 323, pp. 527–544.
- Houghton, J.T., Taylor, F.W., Rodgers, C.D., 1984. *Remote Sounding of Atmospheres*. Cambridge University Press, Cambridge, UK.
- Liou, K.N., 1992. *Radiation and Cloud Processes in the Atmosphere*. Oxford University Press, Oxford, UK.
- Rodgers, C.D., 2000. *Inverse Methods for Atmospheric Sounding: Theory and Practice*. World Scientific Publishing Co. Ltd., Singapore.
- Stephens, G.L., 1994. *Remote Sensing of the Lower Atmosphere, an Introduction*. Oxford University Press, Oxford, UK.

Water Vapor

JE Harries, Imperial College of Science, Technology and Medicine, Blackett Laboratory, London, UK

© 2003 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 5, pp 2005–2012, © 2003, Elsevier Ltd.

Introduction

The measurement of water vapor from space now has an extensive history going back to 1978, when the NASA Nimbus 7 spacecraft was launched carrying two sensors, the Limb Infrared Monitor of the Stratosphere (LIMS), and the Stratospheric and Mesospheric Sounder (SAMS). Both carried channels at 6 μm wavelength in the infrared, to detect thermal emissions from the atmosphere as the spacecraft orbited the Earth in a polar orbit. Shortly afterward, a new generation of operational meteorological sounders was launched by National Oceanic and Atmospheric Administration (NOAA), also in the United States, carrying the TIROS Operational Vertical Sounder (TOVS) package; part of TOVS was the High Resolution Infrared Sounder (HIRS), which made measurements of water vapor in the troposphere. Since that time satellite instruments operating in the infrared, the visible, and the microwave regions of the electromagnetic spectrum have operated in space, and a long series of measurements in both troposphere and stratosphere have been made. This article reviews these measurements.

Principles and Overview

In this section, we discuss some of the general principles involved in remote sensing of atmospheric constituents, before we present some specific examples of satellite missions which have measured atmospheric water vapor.

Emission or Absorption

Remote sensing operates by detection of electromagnetic (e.m.) radiation from the Earth by an instrument on an orbiting spacecraft. There are two principal sources of e.m. radiation that have been most widely used in measurements of water vapor to date. These are the thermal emission from water molecules, and absorption of the visible/near ultraviolet radiation from the Sun. There are other techniques that could be mentioned, such as measurement of solar scattering, or the use of artificial radiation sources, such as lasers, but since these have not been used very widely for water vapor measurements, we will not consider them further in this article.

Thermal Emission

The Planck radiation law describes the maximum thermal radiation that can be emitted as a function of wavelength from any object at a temperature, T . This law shows that the intensity of radiation is a smooth curve with a single maximum at a wavelength in the infrared, λ_m , defined by Wien's displacement law

$$\lambda_m T = 2898 \mu\text{m K} \quad [1]$$

where the wavelength is expressed in microns, μm . Therefore, as an example, at a temperature of $T = 250 \text{ K}$ (which is a typical temperature in the midtroposphere), the maximum intensity occurs at about 11.6 μm . In practice, the intensity of radiation is also dependent on the optical thickness of the atmospheric path under consideration: this, in turn, depends on the density or concentration of the absorbing molecules along that path. If this is written in terms of the Planck function, $B(\lambda, T)$, and the density of water vapor as a function of height, $\rho_w(z)$, then

$$I(\lambda, \theta) = \int_z^{z_t} B(\lambda, T(z)) \tau(z, z_t) \alpha_\lambda \rho_w(z) f(\theta) dz \quad [2]$$

where α_λ is the absorption coefficient expressing the strength of electromagnetic coupling to the radiation field at wavelength λ , $\tau(z, z_t) = \exp(-\int_z^{z_t} \alpha_\lambda \rho_w(z) f(\theta) dz)$ is the transmittance from z to z_t (the top of the atmosphere), and $f(\theta)$ allows for non-vertical transmission paths ($=1/\cos \theta$ for $\theta < \sim 60^\circ$). If the temperature of the atmosphere is known from other measurements, then the dependence of $I(\lambda, \theta)$ on $\tau(z, z_t)$ and $\rho_w(z)$ allows density to be determined.

Absorption

In the absorption case, the controlling equation, say for the case of looking at the Sun through the atmosphere at the limb of the atmosphere, is

$$I(\lambda, \theta) = I_0(\lambda) \tau(z_g, z_t) \quad [3]$$

where z_g is the grazing height, or the minimum height of the solar beam as it traverses the atmosphere, and $\tau(z_g, z_t)$ represents the transmittance of the path from the edge of the atmosphere on the sunward side, via the grazing height, and on to the edge of the atmosphere on the satellite side. Knowing the extraatmospheric solar intensity, $I_0(\lambda)$, allows the transmittance and therefore the density to be determined. The principal differences from emission sensing are that absorption does not depend (at least to first order) on atmospheric temperature, and that emission measurements may be made at any time of day or night, whereas solar-dependent methods can obviously be applied only when the Sun is in the right location.

Limb or Nadir Sounding

We have already indicated that the direction in which we view the atmosphere is important. There are two principal techniques in use. In the first, nadir sounding, widely used for sounding the lower atmosphere or troposphere, the satellite sensor is directed toward the nadir, i.e., downward

from the spacecraft. This is used most, often in emission mode, and the upwelling thermal radiation from the atmosphere is observed. It proves possible to detect emission from broad but distinguishable layers of the atmosphere because the term known as the ‘weighting function,’ $W(z)$, defined as

$$W(z) = \frac{d\tau}{dz} = \tau(z, z_t) \alpha_{\lambda} \rho_w(z) f(\theta) dz \quad [4]$$

displays a single peak at a height dependent on the values of the separate terms, which can be arranged to be at heights between the surface and the tropopause. Because of this property, the density of water vapor in the troposphere can be sounded as a function of height. The technique of nadir sounding using atmospheric emission is used widely in meteorological sounding, including the measurement of humidity.

In limb sounding, the limb of the atmosphere, just above the horizon, is viewed with a sensor with a narrow field of view. Either emission or solar absorption may be employed. By geometrically limiting the field of view, the vertical profile of emission or absorption, and therefore the vertical profile of water vapor density, may be scanned.

Choice of Wavelength

In principle, any wavelength at which there is thermal emission or solar absorption, and at which the water vapor in the atmosphere is spectrally active, may be employed. In practice, thermal emission from the infrared bands or microwave bands of water vapor have been used most frequently for the detection of atmospheric humidity.

Summary of Missions and Instruments

We have space here to only review a relatively few examples of satellite remote sensing of water vapor. Table 1 lists some of the more successful attempts to do so, organized into the three categories of infrared emission, microwave emission, and infrared and visible solar absorption (occultation).

Examples of Satellite Experiments to Measure Water Vapor in the Atmosphere

Infrared Thermal Emission Measurements

Troposphere

The principal satellite instrument for measuring tropospheric water vapor over the past few decades has been the HIRS, one of the suite of instruments making up the TOVS package. ‘TIROS’ in this nested acronym stands for the Television and Infra Red Operational Sounder, the original name for the NOAA weather monitoring system. HIRS is an infrared sounder, viewing the upwelling radiation from the atmosphere below the spacecraft. HIRS has 20 channels, formed using interference pass-band filters, throughout the thermal infrared, including three that measure around $6.3 \mu\text{m}$ wavelength in the ν_2 emission band of water vapor. The ‘footprint’ or spatial resolution of HIRS is about 25 km in the nadir, and about 40 km at each end of the sideways scan that it uses to maximize coverage. Table 2 gives some of the important parameters of channel 12 of HIRS, which detects upper-tropospheric water vapor.

The upwelling IR radiation comes from a restricted range of altitudes, defined by the strength of water vapor absorption at each wavelength. This idea is captured by the definition of the weighting function, $W(z)$, introduced above, which measures the relative amount of radiant energy reaching the spacecraft flying above the atmosphere from each layer within the atmosphere. $W(z)$ is actually equal to the vertical derivative of transmittance at the wavelength in question, as shown in eqn [4]. An example of the typical shape of the HIRS 12 weighting function is shown in Figure 1. At wavelengths where the spectral absorption coefficient is larger, the curve shown is higher in the atmosphere, and vice versa for a wavelength at which the absorption coefficient is smaller.

In order to develop a qualitative understanding of the shape of the weighting function, and how its height depends on absorption coefficient, consider the following. From the highest layers of the atmosphere, the density of water vapor is very low, so that the emission signal reaching our spacecraft instrument is very low from these layers. If we imagine moving deeper into the atmosphere, the emission signal seen at the

Table 1 Satellite instrument for measurement of water vapor: a sample

<i>Instrument</i>	<i>Latitude</i>	<i>Period of assessment</i>	<i>Vertical range</i>
1. IR emission			
TOVS (nadir sounder)	Global	1979–99 (daily)	200–500 hPa
LIMS (limb sounder)	64° S–84° N	October 1978–May 1979 (daily)	1–100 hPa
2. Microwave emission (limb)			
MLS stratospheric	34° N–80° S or 34° S–80° N	September 1991–April 1993	0.01–100 hPa
MLS UTH	See MLS stratospheric	September 1991–September 1994; intermittent thereafter	147, 215, 316, 464 hPa
MAS	Near-global range of latitudes per shuttle mission	March/April 1992, April 1993, November 1994	0.01–50 hPa
3. Solar occultation (limb)			
HALOE	70° S–70° N	October 1991–September 1999	0.01–200 hPa
ILAS	57° N–73° N, 64° S–88° S	November 1996–June 1997	0.1 hPa to cloud tops
POAM III (solar occultation)	54° N–71° N, 63° S–88° S	March 1998–September 1999	3 hPa to cloud tops

Table 2 Examples of weighting functions for HIRS 12 in the TOVS package. The HIRS 12 weighting function for two different profiles of water vapor, expressed as the precipitable water vapor in the layers 1000–500, 500–300, and 300–100 hPa, are shown, compared with weighting functions for two other instruments, SSM/T2, and GMS-5, a geostationary instrument

Parameter	Description
Method	Measure upwelling clear-sky IR emission near 6.3 μm
Accuracy	1.11 K (5 observations per month); 0.11 K (100 observations per month); 0.1 K (global/interannual)
Precision	0.01 K
Time/space resolution	Monthly average; 2.5° latitude and longitude
Altitude range	200–500 hPa (depends on water vapor amount)
Calibration	Onboard black body and cold space view

spacecraft increases, as the density of water vapor increases. At some point, defined by the magnitude of the absorption coefficient at the wavelength in question, another effect begins to come into play: because the emitting layers are now overlain by quite a lot of absorbing water vapor, the IR radiation emitted from these layers toward space is actually partly absorbed before it emerges from the top of the atmosphere. Moving deeper still into the atmosphere means that this reabsorption effect increases, until, for the deepest layers

(depending on the absorption coefficient, and therefore the wavelength), the energy actually reaching the top of the atmosphere can fall to near zero. Hence, the general shape of the weighting function. If the wavelength is changed to a more absorbing one, the height of the peak in $W(z)$ rises, and vice versa. In this way, though with modest vertical resolution, the water vapor density of the troposphere can be mapped in three dimensions as the satellite orbits over the greater part of the globe.

Data from HIRS/TOVS have been available since 1979, with minimal gaps, and provide an important source of information about the distribution and variability of water vapor in the troposphere, used by scientists interested in both short-term weather and long-term climate. Figure 2 illustrates some data from HIRS.

An instrument called MOPITT (Measurements of Pollution in the Troposphere) was launched on board the Terra satellite in December 1999, to provide measurements of CO in the troposphere, with relatively coarse vertical resolution, but with good horizontal sampling; these allow maps of CO distribution to be produced, and MOPITT techniques are also usable in principle to measure water vapor. In addition, in 2003 the EOS Aura satellite will carry the TES (Tropospheric Emission Spectrometer) instrument, which will make a variety of very valuable measurements of tropospheric constituents, including water vapor, with excellent spatial and temporal resolution. TES is based on the pedigree established by the Fourier

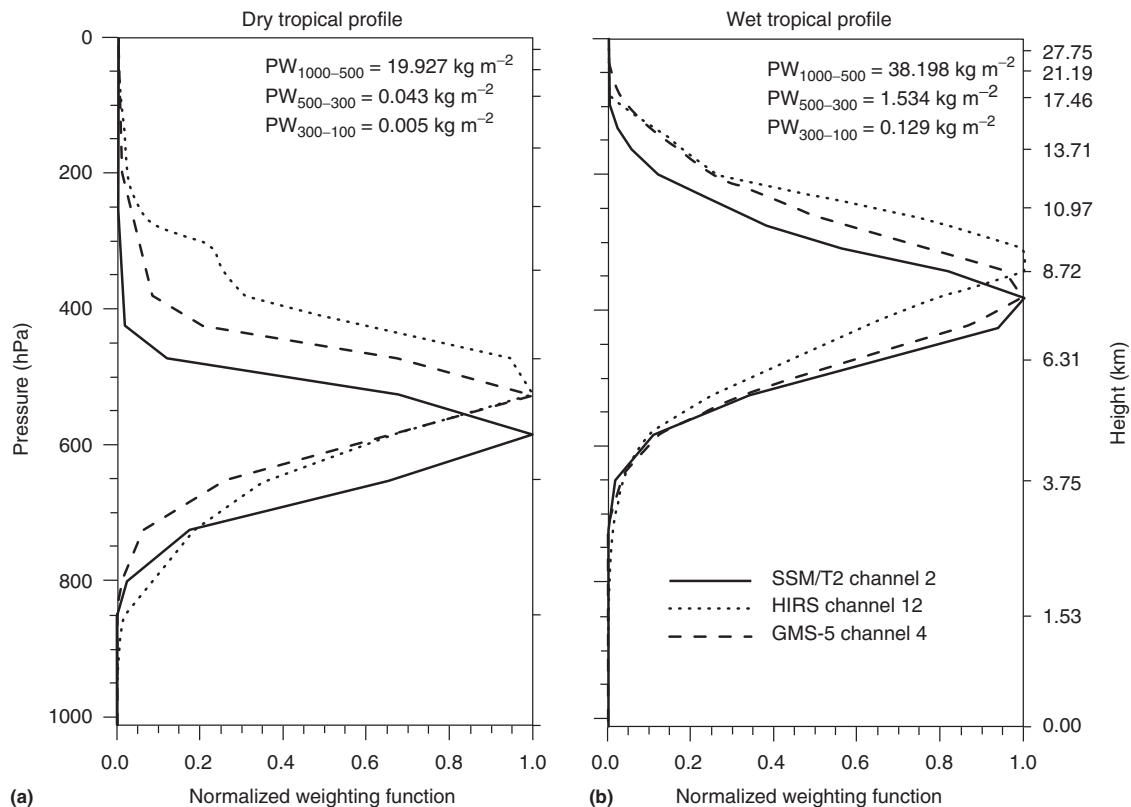


Figure 1 Examples of weighting functions for HIRS 12 in the TOVS package. The HIRS 12 weighting functions for two different profiles of water vapor, expressed as the precipitable water vapor in the layers 1000–500, 500–300, and 300–100 hPa, are shown, compared with weighting functions for two other instruments, SSM/T2, and GMS-5, a geostationary instrument.

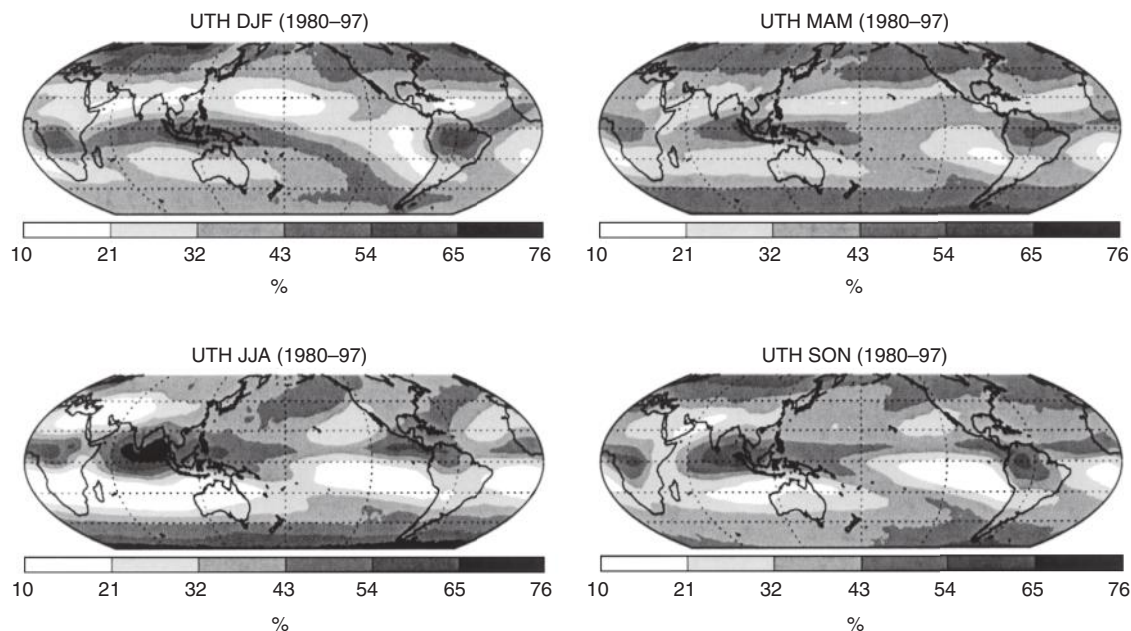


Figure 2 Measurements of upper-tropospheric humidity (UTH) in the layer 500–200 hPa measured by HIRS, expressed as maps for the four seasons DJF, MAM, JJA, and SON, averaged over the years 1980–97.

transform experiment called ATMOS, which flew on the Space Shuttle, to measure the high-resolution spectrum and composition of the stratosphere. Table 3 shows some of the basic characteristics of TES.

Stratosphere

In many ways, the remote sounding of the stratosphere is simpler than that of the troposphere. This is because in the stratosphere there are virtually no clouds to interrupt the line of sight and complicate the observed signal; the densities are lower, so that spectral lines are narrow and better separated than in the troposphere; and techniques like limb sounding may be used, in which the long path to the horizon maximizes the signal from trace molecules, and provides a cold background of space against which to make measurements. These advantages are considerable, and have led to the development of many stratospheric remote sounding techniques, while tropospheric sounding is still in its infancy.

As an example, we take the LIMS, which flew on the Nimbus 7 satellite, and which operated from 25 October 1978 to 28 May 1979. LIMS was a cryogenically cooled broadband filter

radiometer, which included a water vapor channel at $6.9 \mu\text{m}$. The 7-months life was limited, by design, by the lifetime of the solid cryogenics used.

LIMS provided the first comprehensive global view of stratospheric water vapor. The new data were used, along with measurements of CH_4 from the UK SAMS, to study the budget of water vapor and hydrogen in the stratosphere. Precision, or relative accuracy, varied from $\pm 10\%$ for pressures greater than about 10 hPa to 15% between 5 and 10 hPa. Much was learned about the operation and characteristics of an advanced IR sounder during this experiment, experience which was valuable in developing later generations of instruments.

The LIMS experiment has been followed by other IR sensors, employing a variety of techniques, too many to be reviewed fully here. The reader is referred to the survey contained in the book *Earthwatch* listed under Further Reading.

Microwave Thermal Emission Measurements of the Stratosphere

In the extreme far infrared, or microwave region, at wavelengths beyond about 1 mm or so, a fundamental change in techniques is necessary. Thermal radiation is still emitted by all objects, but the intensity is so weak that it may be detected with any certainty only by using the radio techniques of superheterodyne detection. Thus, systems involving mixers and local oscillators (LO) have been developed (the mixers/detectors are related to the old crystal set whistlers) and have been used to detect the sidebands formed by mixing of the incoming signal from the atmosphere with the LO. Such systems have been developed over many years to achieve the high sensitivity in brightness temperature that is necessary to detect emission lines from atmospheric gases such as H_2O , O_3 , CO , ClO , and others.

Table 3 TES characteristics

Maximum sampling time of 16 s with a signal-to-noise ratio of up to 600:1
Limb mode: altitude coverage = 0–34 km
Nadir and limb viewing (fully targetable)
Spectral region: $3.2\text{--}15.4 \mu\text{m}$
Swath: $5.3 \times 8.5 \text{ km}$
Spatial resolution: $0.53 \times 5.3 \text{ km}$
Mass: 385 kg (allocation)
Power: 334 W (allocation)
Date rate: 6.2 Mbps (peak); 4.9 Mbps (average)
Physical size: $140 \times 130 \times 135 \text{ cm}$ (stowed); $304 \times 130 \times 135 \text{ cm}$ (deployed)

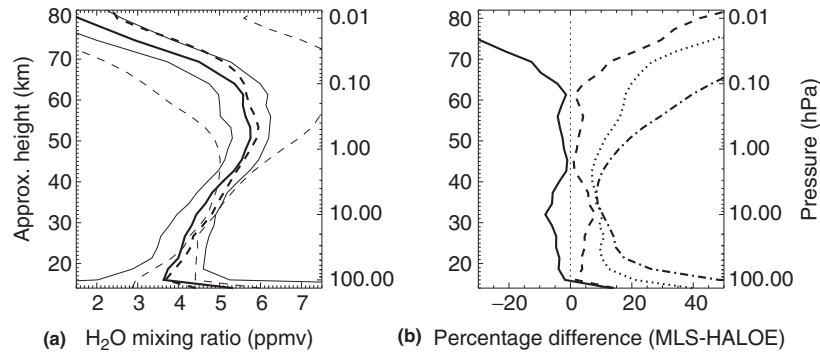


Figure 3 (a) Examples of midlatitude water vapor mixing ratios for the period October 1991–April 1993, measured by MLS (version 0104, solid lines) and HALOE (version 19, broken lines). In each case the darker line is the mean and the lighter lines the standard deviation. (b) Differences between MLS and HALOE: solid – mean difference; dotted – r.m.s. difference; dashed – absolute mean difference; dot-dashed – combined instrumental uncertainties.

The principal advantages of the microwave technique are that the spectral resolution of the system is extremely high, so that atmospheric line shapes can be resolved, while at these long wavelengths, scattering due to particles and droplets is small (see the λ^{-4} dependence of Rayleigh scattering on wavelength), so that clouds and dust are less of a problem to remote sounding. Problems can arise in the calibration of such systems, for example, because standing waves are set up in apparatus which can change if a component is physically rotated or moved, and which can give rise to a varying background signal. Also, because of the long wavelength, diffraction limits the field of view that can be achieved, unless very large antennae are used. Nevertheless, very useful measurements have been made, relatively uncontaminated by cloud or dust.

The accuracy quoted for latest version of MLS water vapor measurements is about ± 0.3 ppmv at 20 km, 0.2 ppmv at 40 km, and about 0.5 ppmv at 70 km.

The Microwave Limb Sounder (MLS) which flew on the Upper Atmosphere Research Satellite (UARS) demonstrated the true power of a limb sounding microwave radiometer for making global measurements for the first time. A more advanced instrument is being developed for the NASA EOS Aura mission, and is currently scheduled to be flown in 2003.

Some examples of MLS data, in a comparison with HALOE measurements, are shown in Figure 3.

Solar Occultation Measurements

Solar occultation, uses the Sun as a source of radiation and measures the change in signal as the Sun rises or sets behind the limb of the atmosphere. Solar occultation has been used with great success in a number of experiments. We note the valuable data produced by the Stratospheric Aerosol and Gas Experiment. However, we use as an example the most highly successful stratospheric experiment perhaps of recent times, the Halogen Occultation Experiment (HALOE). This is a thermal IR solar occultation device with a number of spectral channels, which uses gas correlation techniques as well as broadband filters. In gas correlation spectroscopy, a sample of the gas to be detected is carried in a cell on board, and provides a natural filter to atmospheric radiation specifically from that gas. HALOE has worked now for just on 10 years, and has provided

a quite unprecedented set of data on the stratosphere over this long period. Given that more and more scientific interest is coming to be centered on long-term changes in our climate, the continued operation of this extremely valuable instrument is most important.

Among the advantages of this experiment are the well-known ones due to limb sounding (see earlier). In addition, however, the solar occultation method allows an absolute calibration against the Sun outside the atmosphere at each limb scan. This has proven an extremely important advantage when the data have come to be used for long-term trend and variability studies. The main disadvantage of the solar occultation technique is that since the measurement is made only during local sunrise and sunset, only two measurements are made per orbit: for a 15 orbit day, that is 30 observations per day, with a quite widely spaced horizontal sampling (see Figure 4). This

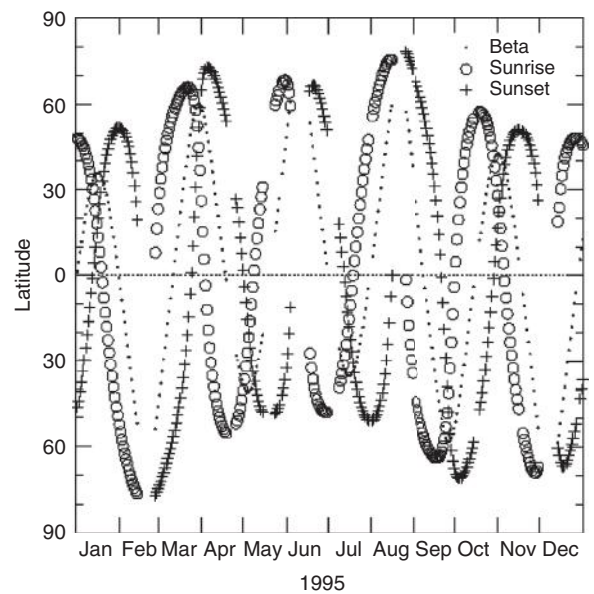


Figure 4 HALOE sampling pattern for the year of 1995. Circles mark the tangent points for sunrise observations, crosses those for sunset observations. The tracks are due to a combination of orbit, season, and the reversal of the spacecraft for thermal balance reasons.

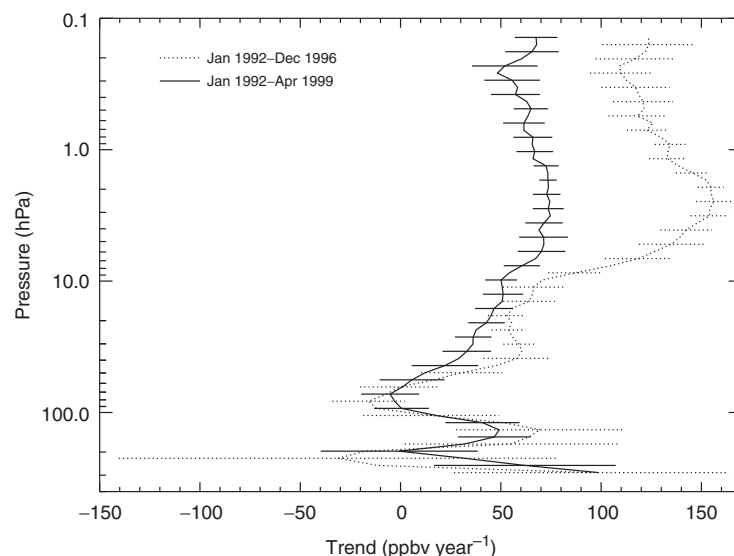


Figure 5 Trends in stratospheric water vapor at levels between 120 and 0.15 hPa for two periods, January 1992–December 1996 (dotted) and January 1992–April 1999 (solid). Bars show standard deviation at each level.

should be compared with the closely spaced, uniform sampling possible from a nadir sounder like HIRS.

HALOE measures not only H_2O but also CH_4 , O_3 , NO , NO_2 , and temperature. The combination of water vapor and methane measured simultaneously by one sensor has been exploited by a number of groups to study the hydrogen budget and water vapor trends of the stratosphere: trends of the order of $50\text{--}150\text{ ppbv year}^{-1}$ occurred during the first half of the 1990s, less in the second half. Such changes produce significant change in the water vapor amount in the stratosphere, which may influence the radiative balance due to this species.

Figure 5 shows an estimate from HALOE data of the trends over the periods January 1992–December 1996 and January 1992–April 1999, demonstrating the change in trends detected over these two periods. These trends are still not well understood, but it has been shown that they have a significant effect on the Earth's radiative energy balance.

See also: **Chemistry of the Atmosphere:** Methane; Observations for Chemistry (*In Situ*): Water Vapor Sondes; Observations for

Chemistry (Remote Sensing): IR/FIR (Satellite, Balloon and Ground); Observations for Chemistry (Remote Sensing): Microwave. **Climate and Climate Change:** Overview. **Global Change:** Upper Atmospheric Change. **Stratospheric Chemistry Topics:** Hydrogen Budget; Stratospheric Water Vapor.

Further Reading

For more comprehensive information about missions, instruments and data, the reader is particularly directed towards the following two publications:

Harries, J.E., 1995. *Earthwatch: The Climate from Space*. Wiley, Chichester.

SPARC, 2000. *Assessment of Upper Tropospheric and Stratospheric Water Vapour*. World Climate Research Programme Report No. 113. World Meteorological Organisation, Geneva. (The author is particularly indebted to the authors of this report, which has provided very valuable background in writing this article.)

Relevant Website

<http://www.earth.nasa.gov> – The NASA web site gives details of many instruments and missions.

SOLAR SYSTEM/SUN, ATMOSPHERES, EVOLUTION OF ATMOSPHERES

Contents

Evolution of Earth's Atmosphere

Planetary Atmospheres: Mars

Planetary Atmospheres: Venus

Solar Terrestrial Interactions: Climate Impact

Solar Winds

Meteors

Evolution of Earth's Atmosphere

YL Yung and ML Wong, California Institute of Technology, Pasadena, CA, USA

EJ Gaidos, University of Hawaii at Manoa, Honolulu, HI, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by E J Gaidos, Y L Yung, volume 2, pp 762–767, © 2003, Elsevier Ltd.

Synopsis

Earth's atmosphere has evolved over 4.5 Ga to reach its present state. The atmosphere is subject to myriad influences, both exogenous, like the solar wind and meteoritic impacts, as well as endogenous, such as outgassing and plate tectonics. By studying noble gas isotope ratios and the rock record, we can define how these processes shaped Earth's atmosphere. Life has also played a profound role in Earth's atmospheric history, affecting the content of key constituents such as oxygen, nitrogen, and various greenhouse gases. Based on life's role in the cycling of atmospheric gases, the Gaia hypothesis suggests that the biosphere regulates Earth's climate to sustain suitable conditions for living organisms.

The Diversity of Planetary Atmospheres

Although the planets Mercury, Venus, Earth, and Mars have masses within a single order of magnitude range, they possess atmospheres with extremely different properties (**Table 1**). These bodies may have initially possessed primordial atmospheres of solar composition whose dominant light gases (hydrogen and helium) were lost to space and replaced by

outgassed water, carbon dioxide, and nitrogen (and small amounts of other gases) during the final phase of accretion 4.5 billion years (Ga) ago. The divergence in atmospheric composition seen today may in part reflect differences in initial volatile abundance, but much of the diversity can be attributed to the individual evolutionary paths of these atmospheres over the age of the solar system. Rates of planetary atmospheric evolution have differed markedly. Whereas the other planets have suffered catastrophic atmospheric evolution (Mercury has experienced complete loss, Venus, a runaway greenhouse and devolatilization of surface rocks, and Mars has lost most of its atmosphere to space or its crust), the evolution of Earth's atmosphere has been comparatively mild. Atmospheric evolution is controlled by both external processes such as radiation and the corpuscular wind from the Sun and impacts, and internal processes such as volcanism and recycling of a planet's crust (e.g., plate tectonics). While some processes drive the exchange of compounds between the atmosphere and reservoirs in the surface, oceans, or interiors of planets, or the interconversion of different chemical species, others result in the secular, irreversible evolution of the atmosphere. Examples

Table 1 The atmospheres of the inner planets Mercury, Venus, Earth, and Mars

Planet	Mass (Earths)	Pressure (hPa)	Composition
Mercury	0.053	10^{-9}	H, He, Na
Venus	0.817	90 000	96% CO ₂ , 3% N ₂ , minor CO, SO ₂
Earth	1	1000	78% N ₂ , 21% O ₂ , 1% Ar, minor H ₂ O, CO ₂
Mars	0.107	7	95% CO ₂ , 3% N ₂ , 2% Ar, minor CO, O ₂

of the former include atmospheric photochemistry, volcanism, and plate tectonics. Examples of the latter include the accretion of new material (impacts of comets or meteorites), escape of hydrogen to space, and the sequestration of certain elements (siderophiles) into the metallic core.

Earth's atmosphere has been profoundly affected by another process called life. The modern atmosphere, containing abundant oxygen in gross chemical disequilibrium with surface organic carbon and gases such as methane, is a testament to life's ability to efficiently convert light energy into chemical energy, some of which is stored in the chemical disequilibrium between the atmosphere and surface. Significant disequilibrium is not present on the sterile worlds of Venus and Mars, and it has been suggested that the simultaneous presence of pairs of gases like O₂ and CH₄ in an atmosphere may serve as a planetary 'biosignature' that reveals the presence of life even at a distance. Some gases such as CO₂, the principal source of biological reduced carbon, are maintained at mixing ratios much lower than the level predicted in the absence of life. The current terrestrial atmosphere (Table 2) is far from the end state reached on Venus, where all of the surface volatiles are in the atmosphere. Also in contrast to neighboring planets, the terrestrial atmosphere maintains conditions suitable for life (providing a modest greenhouse effect and a shield against biologically harmful radiation), and has apparently done so for 3.5 Ga, despite a 40% increase in solar luminosity, giant impacts, and the changing tempo of plate tectonics.

Evolutionary Processes

The early atmosphere would have been subjected to frequent bombardment by planetesimals left over from early planetary accretion. These impacts may have brought in volatiles, but they may also have removed (to space) some of the existing atmosphere either by the momentum of the impact shock wave or heating of the upper atmosphere (see below). Impacts also turn over the crust and expose fresh surfaces, accelerating chemical interactions between the atmosphere and crustal rocks. Giant impact basins on the Moon date to 3.9 Gy of age,

indicating that a similar bombardment shaped the Earth's atmosphere for the first several hundred million years of its history.

A second process that may have profoundly (perhaps catastrophically) affected Earth's atmosphere was the formation of the iron–nickel core. This removal of metallic iron (and possible fractions of iron-soluble elements like sulfur, carbon, and hydrogen) left the Earth's surface and mantle in a considerably more oxidized chemical state, a condition that would have been reflected in the oxidation state of the compounds that were being outgassed by volcanoes. Thus, as is the case today, CO₂, N₂, and SO₂ were the dominant components of volcanic gases, rather than CH₄, H₂S, and NH₃.

Thermal escape of light atoms such as hydrogen and helium can occur from the exosphere (the uppermost layer of the atmosphere where the mean free path between collisions exceeds the scale height). Escape is efficient if the mean thermal speed is a significant fraction of the escape velocity such that a nonnegligible number of molecules in the 'tail' of the Maxwellian (thermal) distribution have speeds above the escape velocity. Extreme and vacuum ultraviolet (UV) radiation from the Sun (mostly in the 121.4 nm Lyman- α line of hydrogen) is absorbed and converted into heat in the Earth's thermosphere. The modern thermosphere lies above 400 km and has a temperature of 1200 K. Nonthermal escape mechanisms, driven by photochemical and other energetic processes, also operate. All these processes create energetic particles by photolytic, electron impact, or ionic reactions. In the absence of a magnetic field, direct momentum transfer may result in knocking atoms from the exosphere in a process known as 'sputtering.' Ions may also migrate to the polar region, where the magnetic field lines are open to the tail of the magnetosphere, and they can readily escape along these field lines. In charge exchange, H atoms lose an electron to a heavier ion, typically O⁺, and the resulting proton is accelerated by the electric field of the solar wind to exceed the escape velocity. The ionization potentials of O and H differ by only 0.02 eV. This difference is smaller than the thermal energies corresponding to typical upper atmosphere temperatures, and as a consequence the charge transfer process is very efficient.

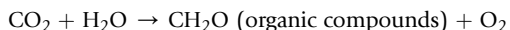
Hydrogen loss from the exosphere is very efficient, but the actual rate (3.0×10^8 atoms cm⁻² s⁻¹) of escape from the modern oxygen-containing atmosphere is limited by the transport of hydrogen from deeper in the atmosphere. Huntén's limiting flux theorem relates the maximum escape flux of hydrogen to the abundance of total hydrogen in the mesosphere, determined largely by chemical sources and transport. The escape rates of hydrogen in the past were probably much higher than the present rate. The primary reason is that the early atmosphere probably lacked oxygen: on the modern Earth, hydrogen is 'trapped' by recombination with oxygen to form water in the lower atmosphere before it can reach the upper atmosphere and escape. A second reason is that the Sun was more active in the past, providing a higher flux of UV radiation to power hydrogen escape. The flow of hydrogen may even have been sufficiently massive to cause the hydrodynamic escape of heavier gases by momentum transfer.

Biology contributes to atmospheric evolution by modulating the interconversion of compounds and the exchange of

Table 2 The reservoirs of the major volatiles on the Earth (mantle quantities are uncertain)

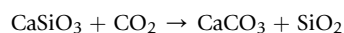
	Reservoir	Size (MPa)	Climate role
H ₂ O	Atmosphere	<0.001	Greenhouse gas, carbonate sink, weathering, biology, plate tectonics
	Ocean	26	
	Hydrated crust	10	
	Mantle	~100	
	Total	136	
CO ₂	Atmosphere	0.000 03	Greenhouse gas
	Ocean	0.0002	
	Hydrated crust	4	
	Mantle	~26	
	Total	30	
N ₂	Atmosphere	0.078	Buffer gas, enhances effect of greenhouse gases
	Crustal rocks	0.025	
	Mantle	~0.1	
	Total	0.2	

compounds between the atmosphere and other reservoirs. The paramount example of this is oxygenic photosynthesis, which in effect partitions CO_2 into organic carbon and oxygen:



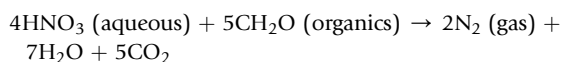
If the organic carbon is prevented from being immediately reoxidized (e.g., by combustion or by aerobic respiration in other organisms), the oxygen can enter the atmosphere. At some point in the past, photosynthetic production was able to overwhelm the sinks of oxygen, and the atmospheric mixing ratio of O_2 increased dramatically – an event that would never have occurred without the intervention of biology.

Biology is also responsible for repartitioning carbon dioxide between the atmosphere, oceans, and limestone rocks (kinetically favoring the last) by accelerating the process of silicate weathering and carbonate precipitation:



Without marine organisms, carbon dioxide levels would be much higher than current values: it has been suggested that the successive development of more sophisticated marine and terrestrial organisms (e.g., plants with vascular roots) has led to a secular decrease in atmospheric CO_2 over time.

A third process where life has played an important role is the recycling of oxides of nitrogen (specifically nitrates) back into the atmosphere by denitrification, e.g.:



In the absence of biology, the formation of nitrates by UV and lightning, and their rainout from the atmosphere, would have been unchecked, removing most of the N_2 (and thus most of the atmosphere) in about 1 Gy.

Aerosols are a minor component of the terrestrial atmosphere whose importance to climate is only now being more fully understood. Aerosols are micrometer-sized particles of sulfur, organic compounds, and salts that remain suspended in the atmosphere for years. They scatter sunlight, effectively cooling the surface, and serve as cloud condensation nuclei, i.e., sites of cloud water droplet formation. Clouds can either cool (by reflecting sunlight back to space) or warm (by blocking the escape of infrared radiation). The aerosol content of the atmosphere may have evolved because of changes in ocean salinity, the surface area of the continents, and biological production of sulfur compounds and organic compounds. The Archean Earth (3.8–2.5 Ga) was probably enshrouded by a photochemical haze composed of fractal aggregate hydrocarbon aerosols. These particles would be optically thick in the UV wavelengths while the remaining, relatively transparent in the midvisible wavelengths, thus providing a UV shield without causing cooling via the antigreenhouse effect.

The Record of Atmospheric Evolution

Because most of the atmosphere is recycled on a geologically short timescale, it has no 'memory' of anything but the most recent events in the history of the Earth. The exceptions to this rule are the noble gases Ne, Ar, Kr, and Xe, which are chemically inert and too heavy to escape easily from the Earth. The rare

gases in the terrestrial atmosphere are depleted with respect to other elements when compared with their expected values in the primordial nebula that formed the planets. They also exhibit strong elemental and isotopic fractionation: heavier elements and isotopes are more abundant (Figure 1). The most plausible explanation is that these gases were components of a primordial atmosphere that escaped to space early in the history of the Earth, perhaps via a 'hydrogen wind' that was driven by intense UV radiation from the young Sun or by the giant Moon-forming impact. (Terrestrial ^{40}Ar , in contrast to its sister isotopes ^{36}Ar and ^{38}Ar , is primarily a product of the radioactive decay of ^{40}K and therefore an exception.) Models of this escape also predict the loss of essentially all N- and C-containing compounds from such an atmosphere and imply that modern N and C have a separate, later origin – either from a long-term reservoir in the Earth's interior or a late 'veneer' of extraterrestrial material.

Our knowledge of the evolution of the other reactive gases is obtained by examination of the chemical and physical effects the paleoatmosphere has had on materials preserved in the rock record. The best record – and the one that has received the most attention – is that of atmospheric oxygen. Much of this work is based on the disparate aqueous solubility of compounds of certain elements, particularly Fe, Mn, and U, in the presence or absence of atmospheric oxygen. Under anoxic conditions, iron is in its highly soluble ferrous state, but insoluble ferric oxides and hydroxides form in the presence of

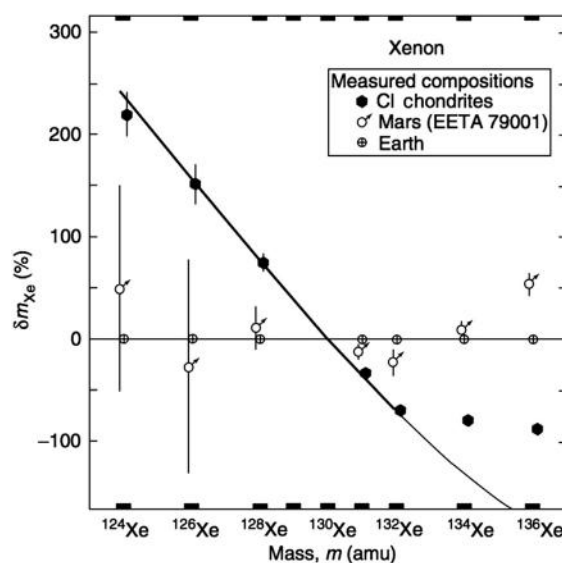


Figure 1 Evidence for loss of an early terrestrial atmosphere: isotopic composition of xenon in meteorites and in planetary atmospheres, plotted relative to terrestrial atmospheric xenon. Units are parts per thousand deviation from the terrestrial composition. The atmospheres of Earth and Mars are depleted in the light isotopes relative to meteorites, which are thought to reflect the composition of the primordial nebula. The lighter isotopes may have been carried away during the early escape of a hydrogen atmosphere. After Pepin, R.O., 1989. Atmospheric composition: key similarities and differences. In: Atreya, S.K., Pollack, J.B., Matthews, M.S. (Eds.), *Origin and Evolution of Planetary and Satellite Atmospheres*. University of Arizona Press, Tucson, AZ.

even small amounts of oxygen. Sedimentary units formed in the Archean (3.8–2.5 Ga) and early Proterozoic (2.5–0.54 Ga) are characterized by banded iron formations, finely laminated deposits of iron oxides and carbonates. The distribution of these deposits suggests that iron was relatively mobile and hence oxygen was absent in the deep ocean during these epochs. Whether the presence of oxygen in shallower water is not clear: iron is conspicuously rare in carbonate formations such as stromatolites from this time. Conversely, iron concentration in Archean paleosols (ancient soils) is low, indicating that Fe remained mobile (and hence removable) during the weathering process. If these soils were in chemical equilibrium with the atmosphere, then the stability of the ferrous state constrained pO_2 levels to 10^{-4} of the present atmospheric level. Oxidized soils and deposits (red beds) in which the (ferric) iron is retained do not appear in the geological record until after 2.4 Ga.

Reduced forms of mineral sulfur (pyrite) and uranium (uraninite) are insoluble and hence resistant to removal by aqueous process. These minerals are seen in Archean deposits, but not in later Proterozoic ones, again indicating an increase in atmospheric pO_2 . The ratio of thorium to uranium in mantle-derived rocks shows a marked increase from samples older than 2 Ga to younger samples, which can be explained if recycling of U into the mantle has become less efficient owing to increased retention of (soluble) U in an oxidizing hydrosphere. Manganese is oxidized only in the presence of abundant molecular oxygen, and on the modern Earth this occurs only through the intervention of Mn-oxidizing bacteria. Mn deposits appear only during certain intervals in the geological record. The early Proterozoic was a unique period of Mn deposit formation (some of the deposits are of great economic importance). If these deposits were formed as oxides (rather than carbonates), they would be unambiguous indicators of the appearance of high oxygen levels.

Although sulfur species are not a significant component of the present-day atmosphere, and the bulk of surface S is in the ocean, sulfur deposits may serve as indicators of past oxygen levels because the rate of oxidative weathering of sulfur minerals and the transfer of S as sulfate to the oceans are controlled by atmospheric O_2 . The other input of sulfur to the ocean-atmosphere system is SO_2 in volcanic gases. Variations in atmospheric oxygen will change the relative importance of these two sources. Possible evidence for such a change is the disappearance of iron formations after 1.8 Ga. Rather than being the result of an oxidation event, the immobility of Fe could be the indirect result of a rise in atmospheric oxygen and the initiation of a 'modern' sulfur cycle, in which oxidative weathering of sulfur-bearing minerals on the continents leads to transport of sulfate to the oceans. Sulfate reduction (by bacteria) would produce elevated concentrations of sulfide in deepwater and precipitation of Fe as pyrite. Sulfur from ancient marine sediments and evaporite deposits also records a shift in the distribution of relative abundance of the two major isotopes ^{32}S and ^{34}S , suggesting an increase in sulfate reduction by bacteria. Archean rocks also provide evidence for mass-independent fractionation among a three-sulfur isotope system that includes ^{33}S (Figure 2). This fractionation pattern is unlike those produced by the oxidation of sulfur by weathering or the reduction of sulfate by biology, where

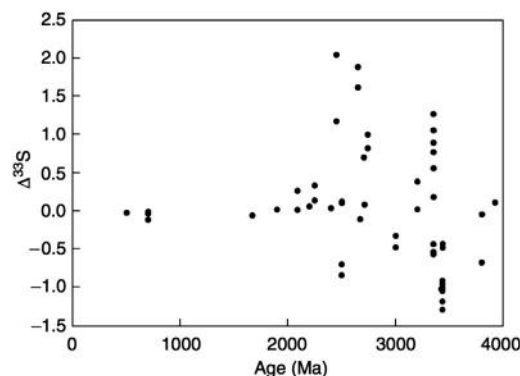


Figure 2 Evidence for a change in the sulfur cycle 2.5 Ga: the quantity $\Delta^{33}S$ is the deviation of $\delta^{33}S$ (the fractionation of the isotopes ^{33}S and ^{32}S with respect to a standard sample) away from a mass-dependent fractionation curve which relates $\delta^{33}S$ to $\delta^{34}S$ by a line with slope ~ 0.51 . All known biological and geochemical fractionation processes are mass dependent, and any samples experiencing only these effects would have $\Delta^{33}S = 0$. While all samples younger than 2 Gy have $\Delta^{33}S = 0$, older samples exhibit mass-independent fractionation which is suggestive of an important role for gas-phase (atmospheric) reactions. This fractionation may be a consequence of an anoxic Archean atmosphere and a lack of oxidative weathering of sulfide minerals. From Farquhar, J., Baoh, H., Thiemens, M., 2000. Atmospheric influence of Earth's earliest sulfur cycle. *Science* 289, 756–758.

fractionation is proportional to the mass differences between the isotopes. Mass-independent fractionation could indicate that in the absence of atmospheric oxygen during the Archean, certain gas-phase reactions were significant contributors to the sulfur cycle.

Carbon dioxide is a minor component (preindustrial $pCO_2 = 280$ ppm) of the modern atmosphere, but it is the principal incondensable greenhouse gas. Since the amounts of CO_2 in the oceans and carbonate rocks are equivalent to 0.023 and 40 atm, respectively, there is the potential for substantial evolution in atmospheric carbon dioxide levels. A higher concentration of atmospheric CO_2 is suspected as one cause for a warmer climate during the Jurassic and Cretaceous periods. It has been suggested that there was a secular decrease in atmospheric CO_2 over time, with levels as high as 300 hpa during the early Archean. The rationale for this hypothesis is the observation that the temperature-sensitive weathering of silicate rocks, which produces calcium and magnesium cations and alkalinity, could have regulated the formation of carbonate rocks and sequestration of CO_2 from the atmosphere. The weathering process acts as a global thermostat: lower carbon dioxide levels would bring about lower temperatures, which would slow weathering and carbonate formation, allowing CO_2 levels to rise. Likewise, lower solar luminosity in the past would have been compensated by higher pCO_2 . Although the geochemical logic is compelling, there is actually little evidence to confirm or refute the hypothesis. Under low oxygen conditions, high CO_2 concentrations would have favored the precipitation of minerals such as siderite (iron carbonate). Absence of siderite in some paleosols may place a constraint on atmospheric CO_2 .

The history of nitrogen, the primary but relatively inert component of the modern atmosphere, is even more poorly constrained by the geological record. The relative abundance of the isotopes ^{15}N and ^{14}N in kerogens (organic molecules that are the sedimentary product of the breakdown of living organisms) in Precambrian rocks indicates a shift in relative abundance from lighter to heavier nitrogen during the late Archean. One explanation for the shift is an increase in atmospheric oxygen content that led to an increased production in nitrate, either by atmospheric processes or by biology. Nitrate compounds are used by microorganisms, which either convert them to a reduced form (ammonia) to include in cellular material eventually, or use them as an oxidant in energy generation (denitrification). Reduction of nitrates by denitrifying bacteria yields isotopically light N_2 gas and isotopically heavy residual nitrates. As a result, surface nitrogen, including that in cells, becomes relatively heavy with respect to the atmosphere. The shift is small (about 10 ppm) and it is unlikely that there was a large change in the partitioning between the atmospheric reservoir (the largest) and surface nitrogen between the Archean and Proterozoic. Thus the nitrogen isotopes may be telling us more about the history of atmospheric oxygen than that of nitrogen.

Atmospheric Evolution and the Gaia Hypothesis

According to accepted theories of the evolution of main sequence stars, of which the Sun is a typical member, solar luminosity has steadily increased by about 40% since the Sun formed. The mean surface temperature of a planet like the Earth is determined by energy balance. If the composition of the atmosphere had remained unchanged, the Earth's mean surface temperature would have been below the freezing point of water before about 2 Ga. But the sedimentary record shows that liquid water has always been present on Earth. A plausible resolution of the 'faint young Sun' paradox is that the early atmosphere contained more greenhouse gases (e.g., CO_2). Two greenhouse gases other than CO_2 that have also been considered for this role are ammonia and methane. Both gases have strong absorption bands that fall within the spectral wavelength range or 'window,' where the Earth reradiates most of the solar energy it receives back to space, but outside the primary CO_2 absorption band at 15 mm. Their abundances (mixing ratios) in the atmosphere will depend on their rates of production and destruction. The largest present-day sources of methane and ammonia are biological, and that is likely to have been the case at early times as well. While ammonia is a very effective absorber of infrared radiation, it is easily destroyed by UV radiation and is unlikely to have been abundant enough in the Archean atmosphere. Methane, on the other hand, has a chemical lifetime of 10 years: CH_4 is stable with respect to molecular oxygen, but reaction with hydroxyl radicals (OH) from the photolysis of water vapor produces CH_3 , which does readily react with O_2 . In the absence of oxygen to remove the hydrogen also produced by photolysis, OH can recombine with H and is no longer an effective sink for methane. The modern production rate can maintain a CH_4 mixing ratio of 10^{-3} in an anoxic atmosphere, a level that can act in concert with a modest amount of

atmospheric CO_2 to maintain Archean temperatures above freezing. However, the dominant modern source of methane is the microbiota in the anoxic guts of termites and ruminants. It is not clear what the production rates would have been in the Precambrian before animals existed.

It has been suggested that the existence and evolution of life on this planet may have had a profound impact on the climate by regulating the amounts of CO_2 and other greenhouse molecules in the atmosphere, or by controlling the production of aerosols and thereby modulating cloud formation. Observations of the fundamental role of biology in the cycling of atmospheric gases is one basis for the 'Gaia' hypothesis in which the biosphere is conceived as regulating climate to maintain suitable conditions for biological activity. The ability of the biosphere to maintain a global environment that is optimal for life is known as homeorhesis. That the terrestrial atmosphere has been subject to homeorhesis by the biosphere throughout its long history is an intriguing hypothesis but difficult to prove. Investigations into the evolution of Earth's atmosphere not only grapple with profound questions about our planet's habitability but also address the possible existence and detection of habitable planets and life around other stars. Counter examples of the Gaia hypothesis include the Great Oxygenation Event (GOE) and Snowball Earth. The GOE around 2.4 Ga might have been triggered by cyanobacteria, which appeared 200 My before. The appearance of oxygen resulted in the loss of the methane greenhouse. In the subsequent Snowball Earth, peroxides could have accumulated in the ice, providing a potent source of oxidants that were lethal to the anaerobic biosphere.

See also: Solar System/Sun, Atmospheres, Evolution of Atmospheres: Planetary Atmospheres: Mars; Planetary Atmospheres: Venus.

Further Reading

- Holland, H.D., 1984. *The Chemical Evolution of the Atmosphere and Oceans*. Princeton University Press, Princeton, NJ.
- Hunten, D.M., Pepin, R.O., Walker, J.C.G., 1987. Mass fractionation in hydrodynamic escape. *Icarus* 69, 532–549.
- Kasting, J.F., 1993. Earth's early atmosphere. *Science* 259, 920–926.
- Kopp, R.E., Kirschvink, J.L., Hilburn, I.A., Nash, C.Z., 2005. The paleoproterozoic snowball Earth: a climatic disaster triggered by the evolution of oxygenic photosynthesis. *Proceedings of the National Academy of Sciences U.S.A.* 102, 11131–11136.
- Liang, M.C., Hartman, H., Kopp, R.E., Kirschvink, J.L., Yung, Y.L., 2006. Production of hydrogen peroxide in the atmosphere of a Snowball Earth and the origin of oxygenic photosynthesis. *Proceedings of the National Academy of Sciences U.S.A.* 103, 18896–18899.
- Lovelock, J.E., Margulis, L., 1974. Atmospheric homeostasis, by and for the biosphere: the Gaia hypothesis. *Tellus* 26, 1–9.
- Melosh, H.J., Vickery, A.M., 1989. Impact erosion of the primordial atmosphere of Mars. *Nature* 338, 487–489.
- Owen, T., Bar Nun, A., 1995. Comets, impacts, and atmospheres. *Icarus* 116, 215–226.
- Porcelli, D., Pepin, R.O., 2000. Rare gas constraints on early Earth history. In: Canup, R.M., Righter, K. (Eds.), *Origin of the Earth and Moon II*. University of Arizona Press, Tucson, AZ.
- Wolf, E.T., Toon, O.B., 2010. Fractal organic hazes provided an ultraviolet shield for early Earth. *Science* 328, 1266–1268.
- Yung, Y.L., DeMore, W.B., 1999. *Photochemistry of Planetary Atmospheres*. Oxford University Press, New York.

Planetary Atmospheres: Mars

RM Haberle, NASA/Ames Research Center, Moffett Field, Mountain View, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

The Martian atmosphere is mainly CO₂ with a mean surface pressure of 6.1 mb. Its thermal structure, wind systems, and photochemistry are similar to Earth's but with important differences. The present climate is characterized in terms of the seasonal cycles of dust, water, and CO₂. Past climates were different due to large changes in Mars' orbital characteristics, and a thicker atmosphere that existed very early in its history. Liquid water flowed on the surface at these early times, but is not stable on Mars today.

Introduction

The atmosphere of Mars is similar to Earth's; it is thin and relatively transparent to sunlight. Mars' spin rate and axial tilt are also Earthlike. Thus, it falls into the category of a rapidly rotating, differentially heated atmosphere with a solid lower boundary. However, there are important differences. The Martian atmosphere is primarily carbon dioxide with a much lower surface pressure, and Mars does not have oceans and an Earthlike hydrological cycle so latent heat release is not as important as it is for Earth. It does, however, contain suspended dust particles, which do provide significant diabatic heating.

Mars also appears to have experienced significant climate change. Today, Mars is cold and dry. Yet spacecraft images provide compelling evidence that the planet's climate was different in the past. Layered terrains in the polar regions may have been created by climate change associated with astronomical variations in Mars' orbit parameters. Valley networks and degraded craters in ancient terrains may be the result of a thicker atmosphere early in Mars' history. And there is some evidence that the planet may have had an ocean at some time in its past, perhaps on several occasions.

Composition and Mass

The composition of the Martian atmosphere was determined in the mid-1970s by the Viking Landers. The results of their measurements are given in Table 1. Carbon dioxide is the principal constituent, followed by nitrogen, argon, oxygen, and carbon monoxide. Trace amounts of the noble gases are also present. Additional minor and highly variable constituents include water vapor, ozone, and dust particles. Together, these gases exert a mean annual surface pressure of 6.1 mb, which corresponds to an average column mass loading of 164 kg m⁻². More recent measurements by the Mars Science Laboratory confirm the Viking measurements, but find equal abundances for N₂ and ⁴⁰Ar (both 1.9%). The change may be the result of different measurement techniques and/or an indication of time variable phenomena.

Temperatures

Temperatures depend critically on Mars' orbit parameters. These are listed in Table 2. The main points are that (1) Mars

receives about half as much annually averaged sunlight as Earth, (2) its orbit is much more eccentric than Earth's, and (3) its rotation rate and obliquity are similar to Earth's. Consequently, Mars is colder, experiences a much greater seasonal change in available insolation (40% compared to 6% for Earth), has Earthlike diurnal and seasonal changes, and has a similar Coriolis parameter.

Except during very dusty periods, the atmosphere of Mars is semitransparent to solar radiation. Consequently, its temperature structure is influenced by thermal emission from the surface. Models indicate that the globally averaged surface temperature on Mars is ~200 K. However, because Mars lacks oceans, its surface temperatures undergo considerable seasonal, diurnal, and latitudinal variation. The lowest surface temperatures (~150 K) occur in the polar regions during winter and are associated with the condensation of CO₂ on the surface. The highest surface temperatures (~300 K) occur in the southern subtropics, when Mars is closest to the Sun. In these same regions, diurnal variations can exceed 100 K.

Approximately 10–20% of the radiation emitted by the surface is absorbed in the atmosphere. Some of the absorbed radiation is reradiated back to the surface, producing a modest greenhouse effect. A convenient measure of the greenhouse effect is the difference between the average surface emission temperature T_{se} (~215 K) and the planet's effective temperature T_{eff} (~210 K). For Mars, this difference is about 5 K. By comparison, the Earth's atmosphere produces a much stronger greenhouse effect (~33 K) due to a much greater abundance of water vapor and other greenhouse gases.

Table 1 Composition of the Martian lower atmosphere (<120 km); the global mean annual surface pressure is ~6.1 hPa

Constituent	Abundance
CO ₂	95.32%
N ₂	2.7%
⁴⁰ Ar	1.6%
O ₂	0.13%
CO	0.07%
H ₂ O	0.03% (variable)
Ne	2.5 ppm
Kr	0.3 ppm
Xe	0.08 ppm
O ₃	0.04–0.2 ppm (variable)
Dust	0 to >> 5 (visible optical depth)

Table 2 Orbital parameters for Mars and Earth

Property	Mars	Earth
Mass, kg	6.46×10^{23}	5.98×10^{24}
Radius, m	3394	6369
Gravity at surface, m s^{-2}	3.72	9.81
Orbit eccentricity	0.093	0.017
Semimajor axis, AU	1.52	1.0
Solar flux, W m^{-2}	590	1360
Length of year, Earth days	687	365
Length of solar day, s	88 775	86 400
Spin-axis inclination, °	25.2	23.5
Longitude of perihelion, °	250	285

Dust particles in the atmosphere strongly influence the transmission of solar and infrared radiation. Based on lander and orbiter measurements, their mean particle radius is $\sim 1.5 \mu\text{m}$. Particles in this size range interact efficiently with sunlight and less so with thermal radiation. During the Viking mission, the daily mean temperature at the Viking Lander 1 site declined by several degrees Kelvin during the passage of a dust storm. This suggests that dust particles produce a modest antigreenhouse effect (i.e., they reflect more sunlight back to space than they emit in the infrared range to the surface).

Vertical Structure

Temperatures decrease with height in the Martian atmosphere, as they do on Earth. As illustrated in Figure 1, the variation of temperature with height on Mars gives rise to a troposphere,

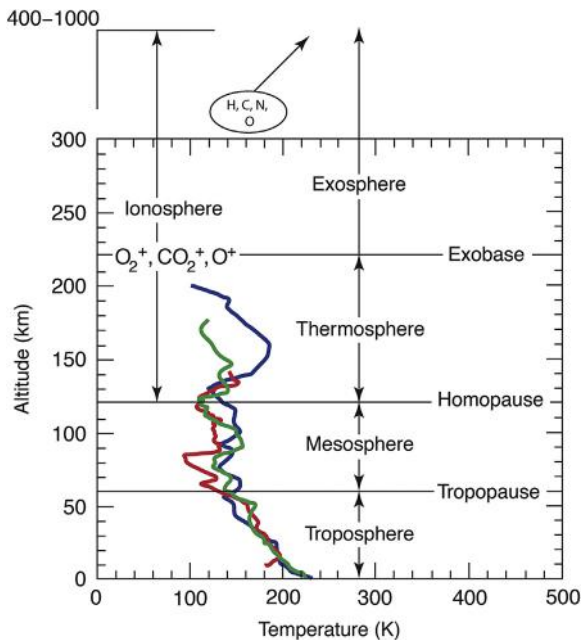


Figure 1 Vertical structure of the Martian atmosphere. Colored curves are temperatures inferred from deceleration measurements aboard the Viking 1 (blue), Viking 2 (green), and Pathfinder (red) Landers.

a mesosphere, and a thermosphere. Mars does not have a stratosphere because it lacks a significant ozone layer.

The troposphere on Mars is deep by comparison to Earth. Based on Viking and Pathfinder Lander entry measurements, the troposphere on Mars extends to almost 60 km with an average lapse rate of $\sim 2.5 \text{ K km}^{-1}$. On Earth, the troposphere is about 12 km deep, and the lapse rate is $\sim 6.5 \text{ K km}^{-1}$.

On both planets, the observed lapse rates are much less than the dry adiabatic lapse rate (Table 3). For Earth, this is due to latent heat release associated with the condensation of water vapor. For Mars, the additional heating comes from the absorption of solar radiation by suspended dust particles. On both planets, temperatures are further stabilized by vertical heat fluxes associated with large-scale circulation systems.

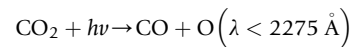
Theoretical studies indicate that daytime boundary layer convection could extend to very high altitudes ($\sim 10 \text{ km}$) on Mars. In such regions, the lapse rates should be close to the adiabatic value. Evidence for deep daytime convection on Mars was found in the Viking Lander 1 entry profile (Figure 1), which indicated a near-adiabatic lapse rate between the surface and 6 km. Above 15 km, temperatures continue to decrease with height, but are controlled almost entirely by radiation rather than convection.

In the Martian mesosphere, temperatures become nearly isothermal. Superimposed on this structure are oscillations due to the adiabatic heating and cooling associated with vertically propagating planetary waves. These waves are associated with a global system of thermal tides. As the tides propagate upward, their amplitude increases. Eventually, they produce superadiabatic lapse rates, at which point the waves 'break' and generate local mixing. There are several locations in the Viking entry profiles where wave breaking is indicated.

In the thermosphere, temperatures increase because of heating due to the absorption of radiation in the far and extreme ultraviolet ranges. This also occurs on Earth. The base of the thermosphere is about 80 km on Earth and about 100 km on Mars.

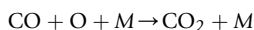
Photochemistry

Photochemical reactions occur throughout the Mars atmosphere. Carbon dioxide, the main constituent, is readily dissociated by ultraviolet radiation:


Table 3 General circulation parameters for Mars and Earth

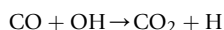
Property	Mars	Earth
Scale height, km	10.2	7.6
Mean temperature of lowest scale height, K	200	260
Dry adiabatic lapse rate, K km^{-1}	4.3	9.8
Mean lapse rate of lowest scale height, K km^{-1}	2.5	6.5
Brunt Väisälä frequency, 10^{-2} s^{-1}	~ 0.06	1.12
Radiative damping time, days	~ 2	> 20
Winter westerly jet speed, m s^{-1}	80	30
Planetary Rossby number	0.05	2.0
Rossby deformation radius, km	920	1150

However, the reverse reaction



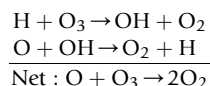
(where M is any nonreactive molecule) is very slow such that the oxygen atoms tend to form O_2 and O_3 rather than CO_2 . The time required to convert the present CO_2 atmosphere into one composed predominantly of CO and O_2 is only several thousand years. Yet CO_2 is the dominant constituent, while CO and O_2 are scarce. How CO_2 is stabilized in the Martian atmosphere is a major focus of Martian photochemical studies.

The prevailing view is that CO is oxidized by OH via



The OH itself is ultimately derived from the photolysis of water vapor. Thus, even though water vapor is a minor constituent, it may play a very important role in maintaining CO_2 as the dominant constituent. Support for the importance of this water chemistry comes from the detection of molecular hydrogen (H_2) in the atmosphere and the distribution and abundance of ozone.

Ozone abundances on Mars are much less than on Earth, and they range from below the threshold of detection in warm tropical regions to as high as 150×10^{15} molecules cm^{-2} in the cold polar regions. Ozone is produced when O and O_2 combine, and is destroyed by ultraviolet dissociation. However, in the absence of additional atmospheric sinks, ozone would be much more abundant than observed. The H and OH produced by water photolysis provide this additional sink by using the same catalytic cycle that operates in Earth's stratosphere, namely



Since the source of odd hydrogen is water vapor, ozone will be depleted in regions where it is abundant, and plentiful in region where it is absent. This anticorrelation between ozone and water vapor has been observed and provides support for the importance of water as a key chemical ingredient of the Martian atmosphere.

During the early 2000s, methane (CH_4) was detected in the Martian atmosphere at the 10 ppb level. The detection is significant since none is expected and its presence at these levels could be the result of a biological source. However, the reported seasonal and latitudinal variability has been difficult to explain since methane is expected to have a long lifetime (200–300 years). This variability implies there are strong sources and sinks for methane, which have yet to be identified and confirmed. Furthermore, the Mars Science Laboratory did not detect any significant methane in the Martian atmosphere during its first year of operations. Thus, the validity of the earlier measurements is controversial.

Escape Processes

Escape occurs in the exosphere, which begins on Mars at about 230 km. In the exosphere, the probability of collisions is so small that particles execute ballistic trajectories, some of which carry them away from the planet. The most important gases that can escape from Mars are hydrogen, oxygen, and nitrogen.

Molecular hydrogen (H_2) is one of the products of water vapor photolysis. Below the homopause, H_2 is well mixed and

has a long lifetime (10^3 years). Above the homopause, it is converted into atomic hydrogen and has enough thermal kinetic energy to escape to space. Ultraviolet spectrometers aboard the Mariner 9 spacecraft have detected atomic hydrogen escaping from Mars.

The escape of hydrogen implies that there must be a sink for O_2 . Otherwise, the amount of O_2 would double in about 2×10^5 years. Loss of oxygen can occur through the oxidation of surface materials and/or escape to space. Loss to the surface requires the continual exposure of surface materials and is not likely to be significant on 10^5 -year time scales. On the other hand, atomic oxygen is too heavy to escape on the basis of its thermal motion alone. However, it can escape when ionized oxygen molecules (O_2^+) in the ionosphere recombine with electrons. The recombination dissociates the molecule into its constituent atoms with enough kinetic energy to escape. This nonthermal escape mechanism – known as ‘dissociative recombination’ – yields an oxygen escape flux that is theoretically expected to adjust itself until it balances the hydrogen loss (i.e., for every oxygen atom that escapes, two hydrogen atoms also escape). In effect, water would be leaving the planet. However, observations suggest that the oxygen escape flux is much less than this prediction, indicating that the atmosphere is not in redox balance at the present time.

General Circulation

Although the meteorological database for Mars lacks the temporal and spatial coverage needed to fully characterize its general circulation, much can be inferred from it – particularly, in connection with general circulation models. Figure 2 is a schematic illustration of our present understanding of the general circulation. From the data and models, the main components of the general circulation are a zonally symmetric mean meridional circulation, stationary and propagating planetary waves, thermal tides, and a mass flow associated with the seasonal cycling of CO_2 into and out of the polar regions. The latter is a unique feature of Martian meteorology.

The mean meridional circulation dominates the lower latitudes and is characterized by a deep Hadley circulation, which undergoes significant seasonal variation in structure and intensity. At the equinoxes, two roughly symmetric Hadley cells develop that share a common rising branch centered at or near the equator. At the solstices, the two Hadley cells give way to a single cross-equatorial circulation. Models indicate that the intensity of the Hadley cell mass flux varies from 10^9 kg s^{-1} at the equinoxes to 10^{10} kg s^{-1} at the solstices.

To some extent, the Hadley cell is a mathematical construct in that the circulation is not in itself zonally uniform. In the rising branch, for instance, much of the upward motion may take place in narrow convective plumes embedded in a broader pattern of overall sinking motion. When averaged over longitude, the net result is upward flow. Thunderstorms play such a role in the Earth's tropics. While there are good theoretical reasons to expect that a similar situation exists on Mars, better observations are needed to confirm this.

The zonal wind component of the mean meridional circulation has been inferred from temperature data through the gradient wind relationship. An illustration of winds derived in

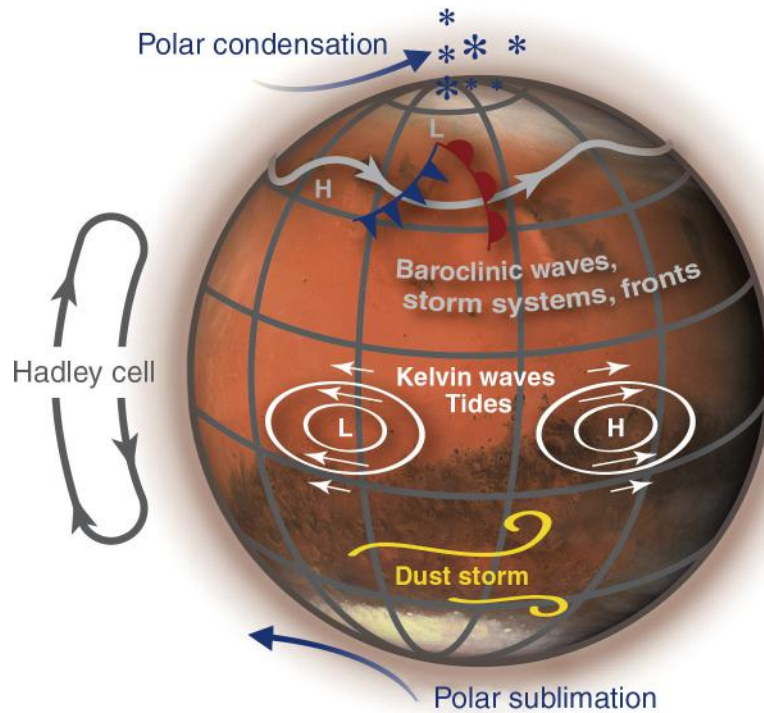


Figure 2 Schematic illustration of the general circulation on Mars.

this manner is shown in [Figure 3](#). Application of the gradient wind relationship to Mars indicates that easterly winds prevail in the tropics at all seasons, and in the summer hemisphere at the solstices. Westerlies prevail in the winter hemisphere at the solstices, and at middle and high latitudes during the equinoxes. If zonal winds at the surface are relatively weak, as was indicated by the Viking, Pathfinder, and Phoenix Landers, then

the thermal data indicate that the westerly jet stream in the winter hemisphere is typically on the order of 100 m s^{-1} .

At high northern latitudes during winter, the Viking Landers detected eastward propagating disturbances of high- and low-pressure systems. These traveling disturbances are very similar to terrestrial ‘weather’ systems in that southerly (northerly) winds are generally associated with falling (rising) pressures

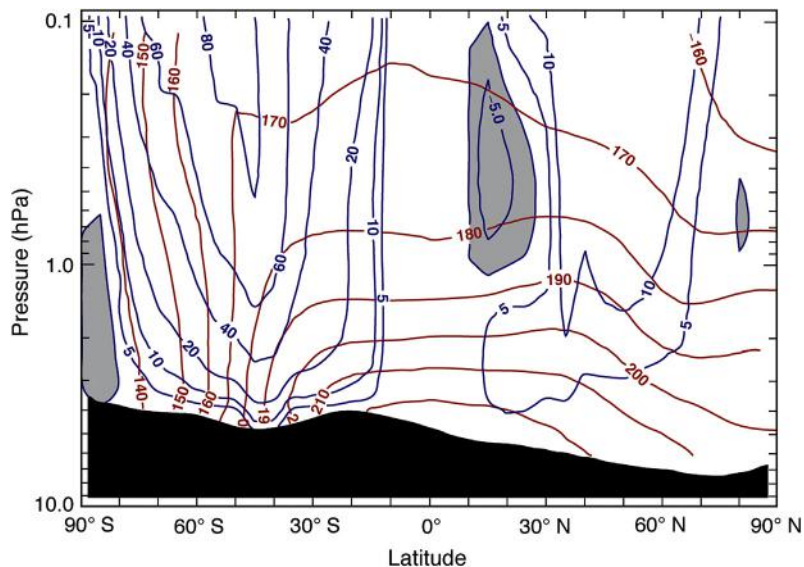


Figure 3 Time-averaged zonal mean winds as a function of latitude and pressure (altitude) for winter in the southern hemisphere. Winds are derived from thermal emission spectrometer (TES) data using the gradient wind relationship. Red-colored contours are isotherms (K); blue-colored contours are wind speed (m s^{-1}).

and warm (cold) air advection. Theory suggests that the transient eddies arise from baroclinic instability. Both theory and observations indicate that the dominant zonal wave number of the transient eddies varies between 1 and 4, and that they propagate around latitude circles with phase speeds between 10 and 20 m s⁻¹.

Dust Storms

The surface of Mars is mantled with a fine dusty material that is lifted into the atmosphere when surface winds become strong enough to initiate particle motion. Because of the low density of the Martian atmosphere, dust-raising winds must be quite strong. Viking Lander 1 measured surface winds gusting to 30 m s⁻¹ during the passage of a dust storm, suggesting that this is the minimum wind speed required to initiate lifting. However, the dust-raising process is complicated and the threshold for lifting can vary depending on surface properties and atmospheric stability.

Numerous dust storms occur each Martian year and are generally classified according to size. From the smallest to largest, they are dust devils (<10⁻¹ km²), local storms (~10³ km²), regional storms (~10⁶ km²), and planet-encircling storms (>10⁶ km²). In general, the smaller storms have shorter lifetimes and occur much more frequently than the larger storms. Dust devils, for example, occur daily and last from minutes to hours, whereas planet-encircling storms occur quasiannually and can last for months.

Dust devils are typically tens of meters in diameter and several kilometers tall. Some have been observed to heights of ~8 km. They generally form over smooth terrain within several hours of local noon. From orbit, these systems have been detected from the shadows they produce and from the trails they leave on the surface. From the surface, they have been detected in meteorological data and camera images. Dust devils are believed to be significant contributors to Martian atmospheric dust loading.

Local dust storms are also quite common. Based on Mars Global Surveyor (MGS) images, as many as 2000 local storms occur each Martian year. This gives a daily-averaged rate of 2–3 storms per Martian day. They have typical lifetimes of less than several days. Local dust storms tend to form along the edge of the polar caps and at the midlatitudes of both hemispheres. These systems often have a distinct convective morphology and can be quite optically thick.

Regional storms have been observed at nearly all seasons, but are most frequent during southern spring and summer. Most regional storms develop within ±30° of latitude, although there is a distinct bias toward the southern hemisphere. Regional storms can last from days to weeks. These storms can drift a significant distance from their original location, and new satellite storms can develop that are quite remote from the original center.

Planet-encircling storms are the least frequent but the most spectacular of Martian dust storms (Figure 4). They spread dust around all longitudes and most latitudes to heights in excess of 50 km. They generally begin in the southern hemisphere during southern spring and summer. None have been observed at other seasons. They start as regional storms and then expand in longitude, then in latitude. To date, eight planet-encircling storms have been confirmed. The most recent occurred in late April 2009 and was observed by the Mars Reconnaissance Orbiter (MRO).

The mechanisms responsible for the Martian dust storms are poorly understood. Dust devils are probably related to strong daytime convective heating, as they are on Earth. For the larger storms, however, radiative feedback effects must be important. Suspended Martian dust particles interact efficiently with sunlight. Thus, as dust loading increases, so does the direct in situ heating of the atmosphere. This then intensifies the circulation, which lifts additional dust. This positive feedback continues until the supply of mobile dust particles on the surface is exhausted, or the dust loading becomes so high that the atmosphere stabilizes and further lifting is suppressed.

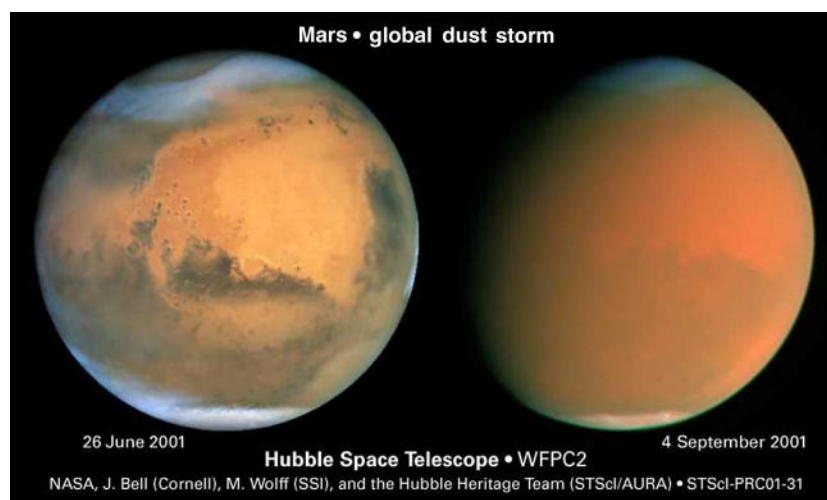


Figure 4 Hubble Space Telescope images of Mars before (left) and during (right) the 2001 global dust storm.

Clouds

Water vapor in the Martian atmosphere is controlled by saturation. Thus, water-ice clouds are fairly common. However, their water content is generally low (several $\text{pr-}\mu\text{m}$). Clouds are most prominent during northern summer in the tropics and during winter in the polar regions. The northern-summer tropical clouds are commonly referred to as the aphelion cloud belt (ACB) since Mars is near its aphelion at this season. These clouds form at altitudes above 15 km and are generally diffuse, although optically thick clouds can often be seen over the volcanoes in the Tharsis province (Figure 5). The winter polar clouds form in the lowest scale height and are called ‘polar hoods.’ These clouds are widespread, are optically thick, and often contain wave clouds forced by topographic obstacles such as craters. The north polar hood is more prominent than the south polar hood. In general, water ice clouds are more common in the northern hemisphere than in the southern hemisphere.

The size and shape of Martian water ice clouds are not well constrained. Retrievals from remote-sensing data indicate that their effective radii are typically in the 1–4 μm range, although light detection and ranging measurements at the Phoenix Lander site suggest sizes as large as 35 μm in clouds near the surface. Submicron-sized particles have also been inferred for thin hazes at high altitudes. These retrievals further suggest that there is very little dispersion in particle sizes and that their shapes are multifaceted and equidimensional.

While latent heat release is not important for Martian clouds, recent studies indicate that the radiative effect of clouds can significantly alter the thermal structure of the atmosphere. Clouds in the ACB warm the atmosphere by absorbing infrared

radiation upwelling from the warmer surface, while clouds in the polar hoods cool the atmosphere by emission to space. Thus, cloud radiative effects tighten the equator-to-pole temperature gradient and increase the baroclinicity of the atmosphere.

Clouds of CO_2 ice also form in the Martian atmosphere. In the winter polar regions, atmospheric temperatures can reach the frost point of CO_2 (<150 K). The presence of CO_2 ice clouds in this region has been inferred from thermal emission and laser altimeter data. CO_2 ice clouds also form at very high altitudes (>50 km) in tropical regions. Like water ice clouds, the physical properties of CO_2 ice clouds are uncertain, but retrievals from remote-sensing data and theoretical microphysical studies suggest they are larger than water ice particles with effective radii up to 60 μm . Unlike water ice clouds, however, latent heat release is important for CO_2 ice clouds and may result in buoyant convection.

Climate

The climate of Mars is characterized in terms of the seasonal cycles of CO_2 , water, and dust. Each of these cycles involves the exchange of material between surface and atmospheric reservoirs. The exchange itself is driven by daily and seasonal variations in insolation. The atmosphere plays a major role in these cycles by serving as the agent of transport from source to sink.

The condensation of CO_2 during winter and its subsequent sublimation during spring give rise to the familiar waxing and waning of the polar caps. Approximately 25% of the Martian atmosphere is cycled into and out of the polar regions each year by this process. The signature of this cycling can be seen in the measurements of surface pressure by the Viking Landers (Figure 6). Recent measurements by the Mars Science Laboratory show similar trends in surface pressure.

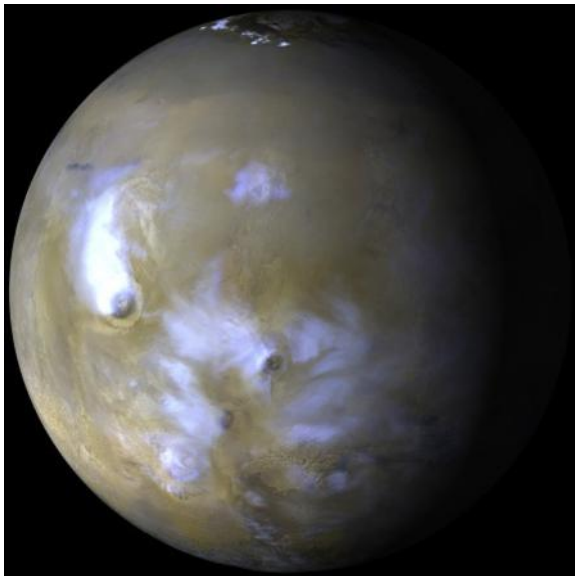


Figure 5 Clouds in the Martian atmosphere as observed by the Mars Color Imager (MARCI) on the Mars Reconnaissance Orbiter (MRO). Figure provided courtesy of Bruce Cantor of Malin Space Science Systems (MSSS): Cantor, B.A., 2007. MOC observations of the 2001 planet-encircling dust storm. *Icarus* 186, 60–96. <http://dx.doi.org/10.1016/j.icarus.2006.08.019>. Credit: NASA, JPL, and MSSS.

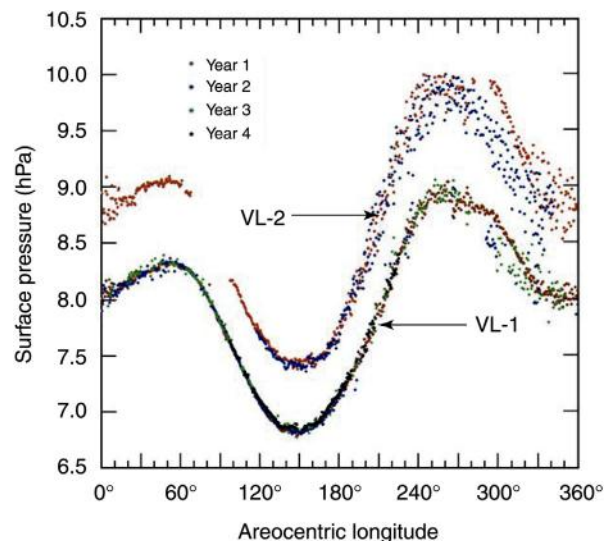


Figure 6 Seasonal variation of the daily-averaged surface pressure on Mars as measured by the Viking Landers. Season is expressed in terms of L_s , an angular measure of the planet's orbital position. $L_s = 0^\circ$ corresponds to the northern spring equinox, $L_s = 90^\circ$ is the summer solstice, $L_s = 180^\circ$ is the fall equinox, and $L_s = 270^\circ$ is the winter solstice.

At each site, there is a pronounced semiannual variation. The variation is semiannual rather than annual because while one cap is growing, the other cap is retreating. However, the variation is asymmetric, with a much deeper minimum occurring during southern winter than during northern winter. This asymmetry is a direct consequence of Mars' orbital eccentricity. Southern winters are much longer than northern winters such that much more CO₂ condenses out of the atmosphere. As a result, pressures are the lowest during the middle of southern winter and highest in late spring when the cap disappears.

At both poles, however, the caps never completely disappear during summer. At the North Pole, CO₂ frost completely sublimates by summer and leaves behind an underlying water ice cap. Near the South Pole, surface CO₂ frost appears to survive all summer long. Thus, the summer caps have different compositions: CO₂ frost in the south, and water ice in the north. The reason for this compositional asymmetry is not understood. Furthermore, the year-to-year changes in the morphology of the residual south cap suggest that it may be slowly eroding and possibly disappear completely within several Mars decades since it appears to be relatively thin (~5 m thick). Thermal data show that a thin layer of water ice underlies the visible CO₂ cap. Recent radar measurements indicate that as much as 5 mb global equivalent of CO₂ ice may be buried to depths of 500 m or more below that thin layer of water ice.

When water ice is exposed at the North Pole during summer, it sublimates into the atmosphere. Winds transport this water all around the planet. The Viking, MGS, and MRO orbiters have mapped the resulting seasonal and spatial variation of atmospheric water vapor. An example is shown in Figure 7. Maximum abundances of about 60 pr- μ m are found at high northern latitudes during summer. The northern-summer ice

cap is therefore an important source of atmospheric water vapor. Minimum observed abundances of less than several pr- μ m occurred in the polar regions during winter. Because of their low temperatures (~150 K), the seasonal CO₂ caps will act as a sink for any atmospheric water vapor that is brought in contact with them. For the current epoch, this implies that on an annual averaged basis, there is a net transfer of water from the north cap to the south cap. The Mars Express orbiter has observed water frost mixed with CO₂ frost at both poles, thus supporting this conjecture.

Climate Change

There is good evidence that Mars has experienced climate change in the recent geological past. Both polar regions are characterized by extensive layered terrains, which were likely formed by atmospheric sedimentation processes that were modulated in time. Furthermore, there are numerous nonpolar ice-related deposits that cannot be produced in today's climate such as tropical mountain glaciers, gullies, pedestal craters, polygons, and a latitude-dependent ice-rich mantle. The most widely accepted theory on the origin of these features is climate change forced by orbital variations. According to this theory, Mars' orbital parameters vary in a quasiperiodic fashion, and this alters the seasonal and latitudinal variation of zonally averaged insolation, which changes the general circulation and the dust, water, and CO₂ cycles.

The key orbital parameters are the obliquity and eccentricity. Because Mars does not have a large stabilizing moon, these parameters can vary significantly on time scales of tens to hundreds of thousands of years. Orbit models can accurately predict how these parameters have varied during the past 20 My (Figure 8). Beyond 20 My, however, the solutions

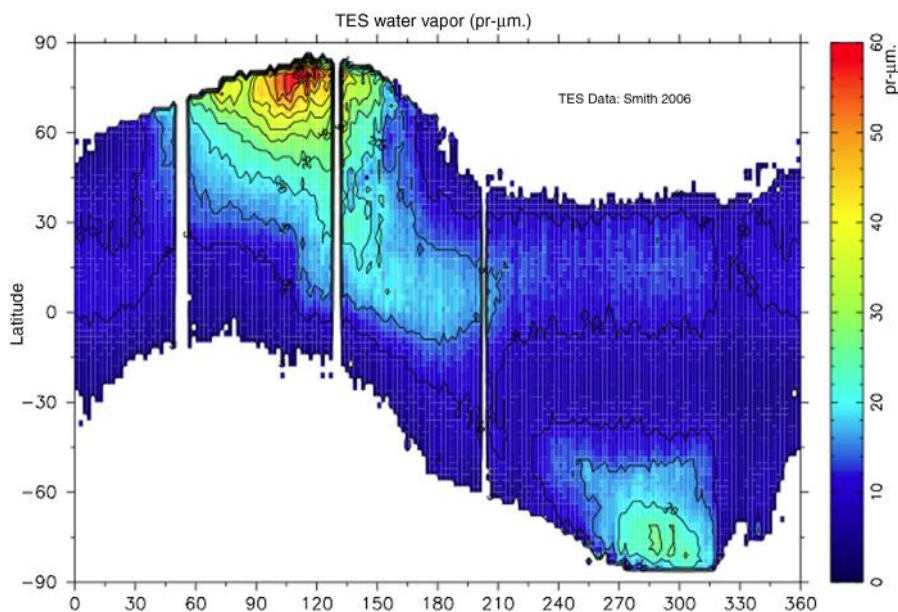


Figure 7 TES zonally averaged column water vapor abundances (pr- μ m) as a function of latitude and season (expressed in terms of L_s). Data provided courtesy of Michael Smith of NASA/GSFC: Smith, M.D., 2008. Spacecraft observations of the Martian atmosphere. *Annual Review of Earth and Planetary Sciences* 36, 191–219.

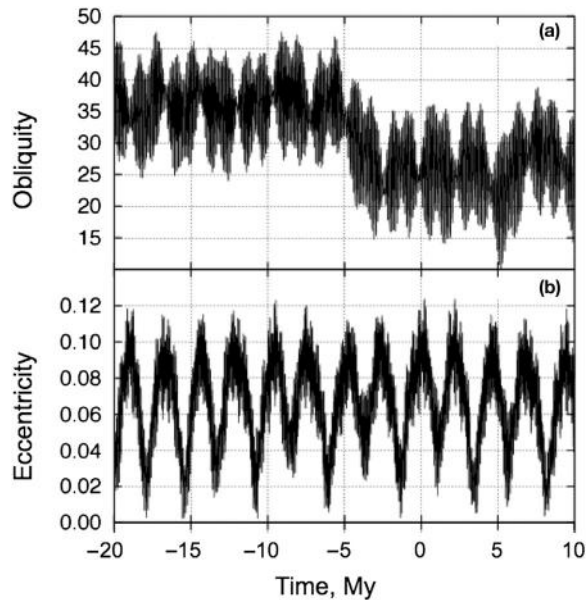


Figure 8 Variation of the (a) obliquity (top) and (b) eccentricity (bottom) of Mars for the past 20 My. Reproduced from Laskar, J., Correia, A.C.M., Gastineau, M., Joutel, F., Levrard, B., Robutel, P., 2004. Long term evolution and chaotic diffusion of the insolation quantities of Mars. *Icarus* 170, 343–364. <http://dx.doi.org/10.1016/j.icarus.2004.04.005>.

become chaotic and are predictable only in a statistical sense. The obliquity is the most important parameter. The current value is 25.2° , but it has varied from between 15 and 45° during the past 20 My, and between nearly 0 and 80° over the full history of the planet. The most likely value of the obliquity is $\sim 41^\circ$. Thus, today's obliquity is not representative of most of Mars' history.

As obliquity increases, annually averaged insolation increases at the poles and decreases at the equator. Consequently, the polar regions warm with respect to the equator. At such times, model studies indicate that water ice at the poles would be transferred to lower latitudes, which may explain some of the features mentioned in this article. Furthermore, surface pressures, and hence the atmospheric dust loading, may increase if the CO_2 ice buried at the South Pole is driven into the atmosphere. Surface pressures could also increase if any exchangeable adsorbed CO_2 is stored within the high-latitude regolith.

At low obliquities, polar regions cool and tropical regions warm. During these times, permanent CO_2 polar caps would form that would act as a sink for both CO_2 and water. Surface pressures would fall, possibly as low as 0.3 mb. Under these circumstances, dust storms would cease, the atmosphere would clear, and seasonal variations would subside. Thus, by modulating the latitudinal distribution of solar insolation, obliquity oscillations provide a mechanism to mobilize and redistribute dust, water, and CO_2 around the planet.

Early Mars

Three lines of evidence suggest that early in its history (>3.5 Gy ago), Mars had an atmosphere much different

from what it has today. The first are geochemical in nature and are based on the mineralogy of surface materials. The second are isotopic in nature and are based on measurements of isotopes in the atmosphere and in the SNC meteorites. The third are geological in nature and are based on the morphology of the surface and its implications for fluid flow.

An example of a geochemical argument is the presence of phyllosilicates (clays) on ancient Martian terrains. Phyllosilicates are the end products of the aqueous weathering of basaltic materials and therefore imply that liquid water was abundant when they formed. The liquid water could be the result of an active hydrological cycle with rainfall and runoff, which implies a warm climate system. Alternatively, it could have been part of a subsurface hydrothermal system driven by internal heating, which has no implications for climate conditions. Geochemical arguments tend to be somewhat ambiguous in their implications for the early Mars climate system.

An example of an isotopic argument is the observed enrichment of N^{15} relative to N^{14} . Like oxygen, nitrogen can escape from Mars nonthermally by dissociative recombination. Above the homopause (~ 120 km), the heavier isotope decreases in concentration more rapidly with height than the lighter one. Consequently, at the exobase (~ 200 km), there are fewer N^{15} atoms available for escape. Over the lifetime of the planet, this leads to an enrichment of N^{15} relative to N^{14} . The observed value is $\text{N}^{15}/\text{N}^{14} = 0.006$, or approximately 1.6 times the terrestrial value. This implies a higher abundance of nitrogen in the past, which also implies a higher abundance of CO_2 . Isotopic arguments like this generally indicate that a thicker atmosphere existed in the past.

Two features in the geologic record provide compelling evidence for a different environment on early Mars. They are the valley networks and the degraded crater rims. Both are found in terrains estimated from crater counts to be 3.5–3.8 Gy old. These terrains are found mostly in the southern hemisphere. An example is shown in Figure 9. The valley networks resemble terrestrial drainage systems and appear to have formed by erosion associated with rainfall and runoff. Such erosion can also explain the degraded crater rims. These features provide the best evidence for warmer, wetter conditions with an active hydrological cycle early in Mars' history.

The most likely way to achieve warmer and wetter conditions on early Mars is through a greenhouse effect, and the most likely greenhouse gases are CO_2 and water vapor. A strong greenhouse effect is required because the sun was 25% less luminous 3.5 Gy ago. At present, neither CO_2 nor water vapor is present in sufficient quantities to produce much greenhouse warming. However, these may have been present in ample quantities in the early Mars atmosphere. Unfortunately, climate models based on pure CO_2 – H_2O atmospheres have been unable to reproduce the conditions needed for sustained rainfall and runoff. Even when supplemented with additional greenhouse gases such as SO_2 or CH_4 , thicker CO_2 atmospheres are only marginally capable of significant warming. Thus, a sustained greenhouse effect capable of supporting liquid water on the surface has not yet been convincingly demonstrated for early Mars.

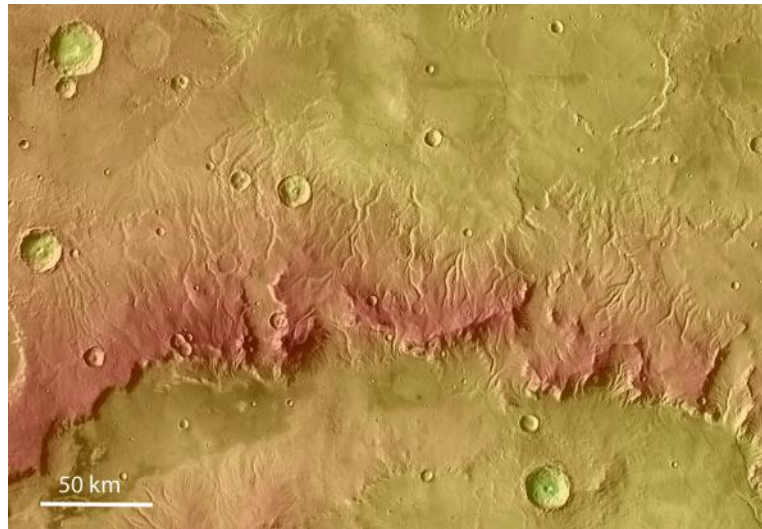


Figure 9 Huygens crater on Mars (10° S, 55° E). Valley networks drain toward the north into a depression where degraded crater rims can be seen (upper right and upper left). Figure provided courtesy of Carr, M.H., 1996. *Water on Mars*, Oxford University Press, New York, 229 pp. Credit: NASA, JPL.

An alternative hypothesis is that the valley networks were formed in multiple transient events associated with impacts. Meteors larger than 10 km in diameter deliver enough energy to temporarily generate a very warm and wet climate capable of providing significant rainfall for hundreds or perhaps thousands of years after the impact. This hypothesis is attractive because Mars obviously experienced a high impact rate early in its history, and we know that impacts can significantly alter the climate system on Earth. However, the predicted erosion rates of impact-generated climate change fall short of the required amount. Thus, a convincing explanation for the mineralogical, isotopic, and geological evidence about early Mars has yet to emerge.

The Future

Our knowledge of Martian atmosphere and climate has greatly benefited from a vigorous international Mars exploration program. Beginning with the MGS mission in 1996, eight spacecraft missions have successfully reached the red planet to

study its mysteries from orbit and the surface (Table 4). On 5 August 2012, the Mars Science Laboratory landed the *Curiosity* rover in Gale Crater with plans to operate it for at least one Mars year. The Rover Environmental Monitoring Station on *Curiosity* gathers hourly meteorological data to extend our record of the surface environment. In late 2013, NASA will launch the Mars Atmosphere and Volatile Evolution mission to study the upper atmosphere and measure escape rates of gases and ions. The European Space Agency and the Russian Space Agency (Roscosmos) are jointly planning several missions to Mars for the 2016 and 2018 launch opportunities that will deliver a lander and an orbiter in 2016 and a rover in 2018. The 2016 lander is expected to have some meteorological instruments, and the orbiter is expected to monitor trace gases in the atmosphere. Many countries are considering missions to Mars for the decade beginning in 2020 that have atmospheric science in their payloads. A sample return mission continues to be a high priority for many space agencies, and beyond that, human missions will ultimately be attempted. Atmospheric science will benefit from these missions as well.

Table 4 Successful spacecraft missions to Mars since 1996

<i>Mission</i>	<i>Launch date</i>	<i>Type</i>	<i>Status (as of Aug. 2013)</i>
Mars Global Surveyor	07 Nov. 1996	Orbiter	Completed
Pathfinder	04 Dec. 1996	Lander/Rover	Completed
Mars Odyssey	07 Apr. 2001	Orbiter	Ongoing
Mars Express	02 Jun. 2003	Orbiter	Ongoing
MER: Spirit	10 Jun. 2003	Rover	Completed
MER: Opportunity	08 Jul. 2003	Rover	Ongoing
Mars Reconnaissance Orbiter	12 Aug. 2005	Orbiter	Ongoing
Phoenix	04 Aug. 2007	Lander	Completed
Mars Science Laboratory	26 Nov. 2011	Rover	Ongoing

MER, Mars Exploration Rover.

See also: **Aerosols:** Role in Radiative Transfer. **Chemistry of the Atmosphere:** Principles of Chemical Change. **Climate and Climate Change:** Greenhouse Effect; Overview. **Dynamical Meteorology:** Baroclinic Instability; Overview. **General Circulation of the Atmosphere:** Overview. **Solar System/Sun, Atmospheres, Evolution of Atmospheres:** Evolution of Earth's Atmosphere; Planetary Atmospheres: Venus.

Further Reading

- Barlow, N.G., 2008. Mars: An Introduction to Its Interior, Surface and Atmosphere. Cambridge University Press, Cambridge, p. 264.
- Cantor, B.A., 2007. MOC observations of the 2001 planet-encircling dust storm. *Icarus* 186, 60–96. <http://dx.doi.org/10.1016/j.icarus.2006.08.019>.
- Carr, M.H., 1996. Water on Mars. Oxford University Press, New York, p. 229.
- Colaprete, A., Barnes, J.R., Haberle, R.M., Montmessin, F., 2008. CO₂ clouds, CAPE and convection on Mars: observations and general circulation modeling. *PSS* 56, 150–180.
- Forget, F., Haberle, R.M., Montmessin, F., Levrard, B., Head, J.W., 2006. Formation of glaciers on Mars by atmospheric precipitation at high obliquity. *Science* 311, 368–371. <http://dx.doi.org/10.1126/science.1120335>.
- Haberle, R.M., Kahre, M.A., 2010. Detecting secular climate change on Mars. *Mars* 5, 68–75. <http://dx.doi.org/10.1555/mars.2010.0003>.
- Kieffer, H.H., Jakosky, B.M., Synder, C.W., Matthews, M.S. (Eds.), 1992. Mars. University of Arizona Press, Tucson, p. 1498.
- Laskar, J., Correia, A.C.M., Gastineau, M., Joutel, F., Levrard, B., Robutel, P., 2004. Long term evolution and chaotic diffusion of the insolation quantities of Mars. *Icarus* 170, 343–364. <http://dx.doi.org/10.1016/j.icarus.2004.04.005>.
- Lefèvre, F., Lebonnois, S., Montmessin, F., Forget, F., 2004. Three-dimensional modeling of ozone on Mars. *Journal of Geophysical Research* 109, E07004. <http://dx.doi.org/10.1029/2004JE002268>.
- Leovy, C., 2001. Weather and climate on Mars. *Nature* 412, 245–249.
- Read, P.L., Lewis, S.R., 2004. The Martian Climate Revisited: Atmosphere and Environment of a Desert Planet. Springer, New York, p. 326.
- Smith, M.D., 2008. Spacecraft observations of the Martian atmosphere. *Annual Review of Earth and Planetary Sciences* 36, 191–219.
- Toon, O.B., Segura, T., Zahnle, K., 2010. The formation of Martian river valleys by impacts. *Annual Review of Earth and Planetary Sciences* 38, 303–322.
- Yung, Y.L., DeMore, W.B., 1999. Photochemistry of Planetary Atmospheres. Oxford University Press Inc., p. 456.
- Zahnle, K., Freedman, R.S., Catling, D.C., 2011. Is there methane on Mars? *Icarus* 212, 493–503. <http://dx.doi.org/10.1016/j.icarus.2010.11.027>.

Planetary Atmospheres: Venus

PJ Gierasch, Cornell University, Ithaca, NY, USA

YL Yung, California Institute of Technology, Pasadena, CA, USA

© 2003 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 4, pp 1755–1760, © 2003, Elsevier Ltd.

Introduction

Venus is Earth's sister planet, 0.723 times as far from the Sun as is the Earth. Its radius (6052 km) and mass (4.869×10^{24} kg) are 95% and 82% of the Earth's. Thus, in size and distance from the Sun it is quite similar to the Earth, but its atmosphere is strikingly different from the Earth's. The atmosphere of Venus is much more massive, with a mass per square meter of 10.8 kg m^{-2} compared to 0.1 kg m^{-2} on Earth. It is composed of carbon dioxide (96.5% mol fraction) and molecular nitrogen N_2 (3.5%) plus traces of other gases. A hydrosphere is absent on Venus. There is a very low abundance of water in any form. These compositional differences point to a drastically different evolution of the Venusian and terrestrial atmospheres through geological history.

The current state of the Venusian atmosphere also shows great differences from that of the Earth. There is a strong greenhouse effect, which raises the surface temperature to about 735 K (462 °C), despite the fact that cloud cover strongly reflects sunlight so that less heat flux is absorbed on Venus than on Earth. The cloud deck itself is extremely interesting. It is optically thick, and observed from Earth or from space. Venus appears featureless in visible light. The cloud is composed of droplets of a strong solution of sulfuric acid. Its maintenance depends on a cycling of sulfur compounds between the upper atmosphere, where sunlight produces H_2SO_4 by photochemistry, and the lower atmosphere, where high temperatures destroy the acid.

The solid body of Venus rotates very slowly, with a period of 243 days, in the opposite direction to the general rotation of the solar system. The general circulation of the Venus atmosphere is dominated by rotation of the atmosphere in the same direction as that of the solid planet but increasing in speed with height from the surface up to the cloud tops. It reaches maximum speed near the cloud tops, where the rotation period is about 4 days. Atmospheric dynamics are thus very different on Venus and on Earth. Coriolis forces are relatively weak because of slow planetary rotation. Instead, the atmosphere itself develops a rotation by internal mechanisms that are not yet fully understood.

History and Evolution of the Venusian Atmosphere

The total quantity of atmosphere on Venus is known from measurement of surface pressure by probes dropped through the atmosphere. Four of these were flown to Venus by the NASA *Pioneer Venus* spacecraft in 1979, and several were flown by the Russian *Venera* spacecraft program between 1965 and 1985. The surface pressure was found to be about 9.3 MPa.

When divided by the acceleration of gravity (8.87 m s^{-2} on Venus), the pressure yields the atmospheric mass per unit area. Most of this mass is carbon dioxide. On Earth, a similar quantity of carbon is buried in the crust, primarily in the form of carbonate minerals. On the other hand, Earth's oceans contain a total quantity of water equivalent to a layer with basal pressure of a few hundred kilopascals, compared to much smaller traces on Venus. Thus two questions are raised by compositional differences: Why is carbon dioxide on Venus in the atmosphere instead of in crustal minerals as on Earth? Why does Venus have so little water?

Venus and Earth are similar in size and density and are composed of approximately the same array of elements, and are thought to have formed in similar ways during the birth of the solar system. Thus the same inventories of volatile compounds were probably present on both planets at the time of planetary formation. The most abundant of these were water and carbon dioxide. Current differences are thought to be due to evolutionary effects rather than to initial conditions. In particular, Venus is closer to the Sun and the solar flux onto the top of the atmosphere is 1.9 times larger than on Earth. This is thought to have been large enough to push Venus into a runaway greenhouse state, in which water vapor feedback drove the temperature high enough that all water was forced into the atmosphere as vapor or cloud. Water in the upper atmosphere was then exposed to ultraviolet radiation and dissociated. The hydrogen escaped to space and the oxygen was chemically combined with crustal materials; the consequence was rapid loss of most of the initial water inventory.

In the presence of liquid water and relatively low temperatures, carbon dioxide over geological history on Earth has largely precipitated out in the form of carbonates such as calcite (CaCO_3). These Urey reactions remove carbon dioxide from the ocean–atmosphere system. On Venus, the high surface temperature pushes the equilibrium of the Urey reactions toward higher CO_2 pressures. The major portion of the initial inventory of CO_2 on Venus may still reside in the atmosphere.

These ideas regarding the compositional evolution of the Venusian atmosphere remain speculative because of the absence of detailed geological information from the surface of Venus. For example, detailed chemical analysis of surface minerals has not yet been performed, and sedimentary cores such as those used to investigate past epochs on Earth do not exist for Venus.

Thermal State of the Venusian Atmosphere

Figure 1 displays a temperature profile obtained from entry probes. The shape of the profile divides the atmosphere into

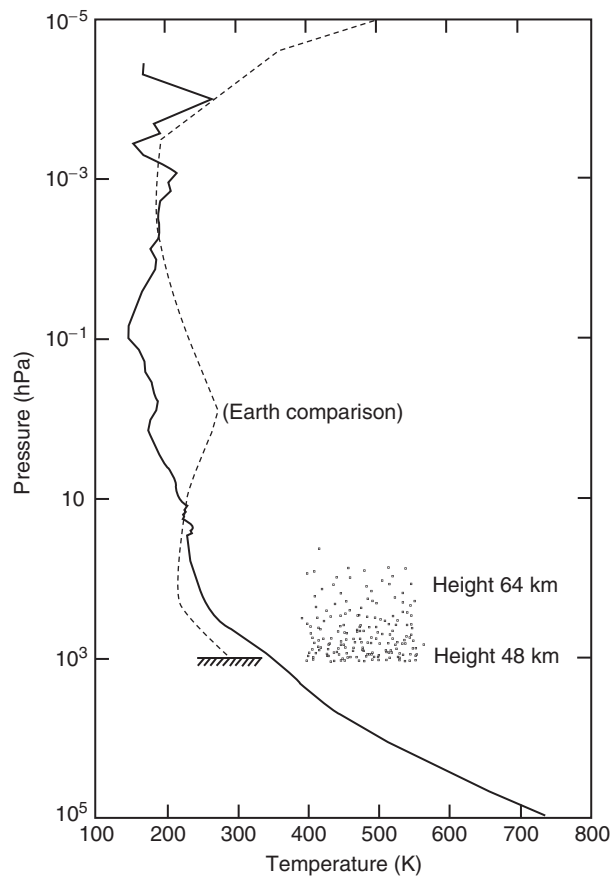


Figure 1 Temperature and pressure in the Venusian atmosphere. Measurements are from the *Pioneer Venus* sounder probe, which entered near the equator. The dotted line shows a mean profile for Earth's atmosphere and indicates Earth's surface pressure. Absorption of ultraviolet by ozone on Earth produces a stratospheric temperature maximum with no counterpart on Venus. The approximate location of the Venusian clouds is indicated. Reproduced from Seiff, A., Kirk, D.B., Young, R.E., et al. 1980. Measurements of thermal structure and thermal contrasts in the atmosphere of Venus and related dynamical observations: Results from the four *Pioneer Venus* probes. *Journal of Geophysical Research* 85, 7903–7933.

two regimes. The middle atmosphere is approximately isothermal and is analogous to Earth's middle atmosphere, but without the warm layer due to absorption of ultraviolet solar radiation by ozone on Earth. The lower atmosphere shows temperatures decreasing with height, and is analogous to Earth's troposphere. The heat balance in the middle atmosphere is dominated by heating due to absorption of sunlight in near infrared CO_2 bands and cooling to space from the $15\ \mu\text{m}$ CO_2 band. In the lower atmosphere the heat balance is more complicated and not fully understood. The solar heating is well constrained by direct measurements of solar flux by entry probes. About 76% of the insolation is reflected. The global average absorption is about $F = 157\ \text{W m}^{-2}$, which implies an effective temperature for emission to space of 229 K. Thus the net energy input is smaller than on Earth (where it is approximately $240\ \text{W m}^{-2}$, giving an effective emission temperature of 255 K) because of the high reflectivity of the Venusian cloud deck. More than half of the solar energy is

absorbed in the cloud deck, between heights of about 50 and 70 km. About 10% of the solar absorption is at the planetary surface. The remainder is smoothly distributed gas absorption. The large heat input in the cloud deck, about 60 km, or five scale heights, above the surface, is a very different driver for meteorology compared with that on Earth, where the bulk of the solar heating is near or at the surface.

Thermal radiation flux is less well understood than the solar flux. The high infrared opacity of the major constituent, CO_2 , is important in producing a strong greenhouse effect. But even a small spectral 'window' for leakage of radiation to space can greatly reduce the effectiveness of a greenhouse, and trace gases such as SO_2 and H_2O are important in filling spectral windows on both sides of the $15\ \mu\text{m}$ CO_2 band on Venus. SO_2 , H_2O , and other trace gases are spatially and temporally variable on Venus, and measurements of their distribution are sparse. Furthermore, at the high temperatures and pressures of the lower atmosphere of Venus, the opacities of many gases are not well-known. The spectral distribution of the thermal radiation flux has not been directly measured within the Venusian troposphere. Nevertheless, radiative modeling based on the best estimates of composition has successfully reproduced the Venus greenhouse, and the greenhouse explanation for high surface temperature on Venus is widely accepted.

Four *Pioneer Venus* entry probes measured temperature profiles in 1979, at different latitudes. The profile displayed in Figure 1 was near the equator. The troposphere temperature increases monotonically with depth but the gradient is not uniformly close to the adiabatic gradient. There is a stable layer near 0.3 MPa pressure, and another near 3 MPa. The middle atmosphere, where the pressure is less than 0.03 MPa, is of course also stably stratified. Thus the vertical structure comprises alternating layers, each a scale height or two thick, with nearly adiabatic, low-stability regions within the cloud deck near 0.1 MPa pressure, below the clouds near 2 MPa, and possibly also near the surface. All four of the profiles measured by *Pioneer Venus* showed similar stratification properties. A pair of Russian balloon probes in 1985 gave evidence for strong vertical motions, probably due to convection, in the low-stability layer within the cloud deck. The layered stability structure of the Venusian troposphere is probably important in controlling atmospheric dynamics, but wind measurements are too limited to reveal the effects.

The orbital period of Venus is 224.70 Earth days. Its spin about its axis is in the opposite direction, with a period of 243.02 days. In combination, these motions produce a solar day on Venus of 116.75 days. In spite of this long time, there is very little day–night temperature contrast in the lower atmosphere. A radiative time constant based on the average thermal cooling rate and the entire atmospheric mass is $t_R = c_p p T_M / (gF)$, where c_p is the heat capacity, p is the surface pressure, T_M is the mean atmospheric temperature, g is the acceleration of gravity, and F is the average thermal flux to space. With $c_p = 800\ \text{J kg}^{-1}$, $p = 9.2\ \text{MPa}$, $T_M = 600\ \text{K}$, $g = 8.87\ \text{m s}^{-2}$, and $F = 157\ \text{W m}^{-2}$, one obtains $t_R \sim 100$ years. On the other hand, a radiative time constant based on the atmospheric mass near and above the cloud deck is only a few days, and indeed measurements have shown that thermally driven tides are significant at these heights, with an amplitude of a few kelvins.

There is very little latitudinal temperature contrast in the lower atmosphere of Venus. Here again the large thermal time

constant is probably responsible. As long as dynamical transport times are shorter than the radiative time constant, heat should be well mixed laterally. The situation is probably analogous to that in the deep ocean basins on Earth. Within the clouds, however, there is a latitudinal temperature gradient, cooler toward the poles by about 20 K. At higher elevations the temperature gradient reverses.

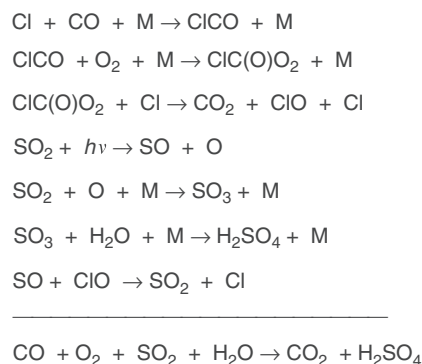
Chemical Stability of the Venusian Atmosphere

Above the cloud tops, absorption of solar UV radiation by CO_2 leads to its dissociation into CO and O. Most of the O atoms recombine to form O_2 . The restoration of CO, O, and O_2 back to CO_2 is effected by catalysis involving chlorine atoms derived from the photolysis of trace amounts of HCl in the Venusian atmosphere. The catalytic mechanism is similar to that proposed for the destruction of the terrestrial ozone layer by chlorine atoms derived from anthropogenic chlorofluorocarbons. Chlorine chemistry also catalyzes the oxidation of SO_2 to H_2SO_4 . The essence of the chemical scheme may be summarized as in **Scheme 1**. ClC(O)O_2 is the peroxychloroformyl radical; ClCO is the chloroformyl radical; and M is a third body (ambient atmosphere).

On decadal timescales, the composition of the atmosphere may be subject to perturbations by episodic volcanic eruptions. However, there is no proof that the reported secular variations in SO_2 abundance can be attributed to volcanic processes. It is conceivable that the observed changes are the result of a coupled upper atmosphere and lower atmosphere oscillation (the analog of El Niño in the terrestrial tropical atmosphere). Over geological timescales, both H_2O and SO_2 are unstable. The loss and resupply of these gases might have induced profound climatic changes and could provide an explanation of the tectonic deformations of the crust observed by the *Magellan* mission radar mapping.

The Venusian Clouds

The Venusian clouds are featureless and opaque in the visible light. Absorption in the blue gives the planet a slight overall yellowish tint to the eye. Absorption becomes strong at ultraviolet wavelengths, and also shows contrast that reveals cloud



Scheme 1

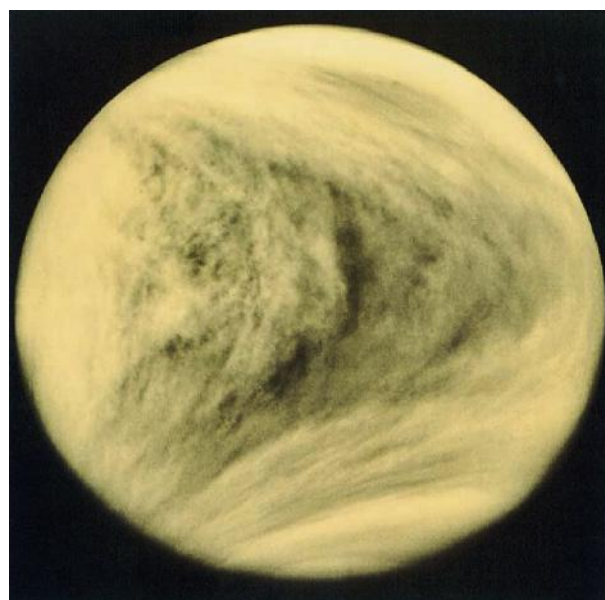


Figure 2 Ultraviolet image of Venus from the *Pioneer Venus* orbiter. North is at the top. The general circulation is from east to west (toward the left) and appears to spiral toward the poles. The large-scale dark pattern centered on the image is the 'Y' discussed in the text as a possible traveling wave. Reproduced from Rossow, W.B., Del Genio, A.D., Limaye, S.S., Travis, L.D., 1980. Cloud morphology and motions from *Pioneer Venus* images. *Journal of Geophysical Research* 85, 8107–8128.

patterns. **Figure 2** displays an ultraviolet image taken by the *Pioneer Venus* spacecraft in 1980. The identity of the ultraviolet absorber is unknown, but is most likely polysulfur. The principal component of the cloud is a haze of sulfuric acid droplets of radius about 1 mm. The gaseous precursor of the sulfate particles is SO_2 , but the ultimate source of sulfur is COS from the surface. Once transported to the upper atmosphere, COS is readily oxidized to SO_2 , using the oxygen derived from CO_2 photolysis. The oxidation of SO_2 to H_2SO_4 has been described in the previous section. However, in regions of the atmosphere where O_2 is deficient, COS may be reduced to S_x (polysulfur). Visibility within the haze is a few kilometers. The base of the cloud is at approximately 48 km elevation, where the pressure and temperature are about 0.13 Mpa and 365 K, respectively. Near the cloud base the *Pioneer Venus* probes detected a relatively thin layer, about 2 km deep, containing larger particles and higher number densities than in the main cloud. There is no sharp top to the cloud. The density tapers off gradually between 60 and 75 km elevation. At 65 km the pressure and temperature are about 0.01 Mpa and 245 K, respectively. In addition to their obvious impact on climate, the clouds of Venus have an important effect on atmospheric chemistry. The sulfate particles completely dehydrate the upper atmosphere. The process is analogous to the removal of water vapor by the cold trap at the tropopause of Earth's atmosphere.

The General Circulation

Figure 3 displays wind velocities measured by the four *Pioneer Venus* probes in 1979. Numerous Russian probes have

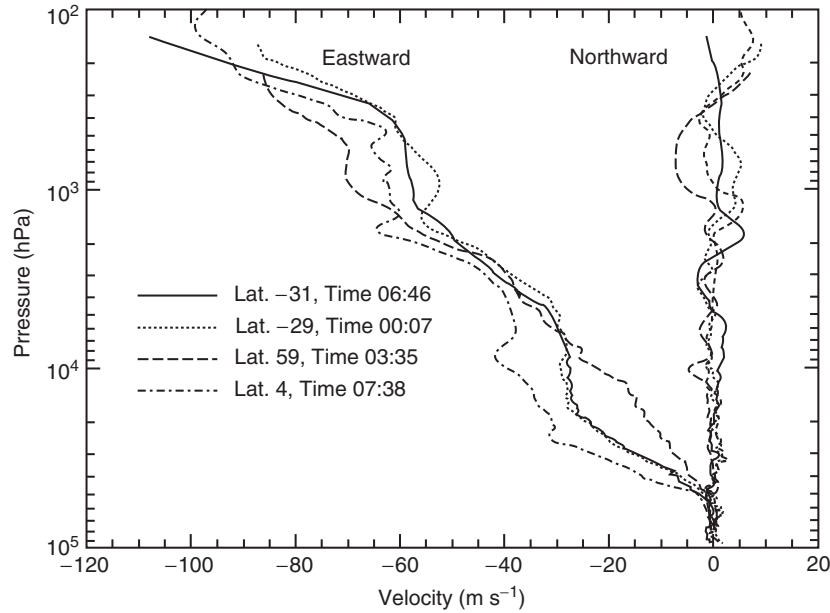


Figure 3 Wind measurements from drifts of *Pioneer Venus* probes. The east–west component of the velocity is far larger than the south–north component. Vertical shear is smallest at heights where the temperature gradient is closest to adiabatic. Reproduced from Counselman, C.C. III, Gourevich, S.A., King, R.W., Lorient, G.B., Ginsberg, E.S., 1980. Zonal and meridional circulation of the lower atmosphere of Venus determined by radio interferometry. *Journal of Geophysical Research* 85, 8026–8031.

produced similar results. The flow is dominated by an easterly zonal wind increasing with height up to approximately the top of the visible clouds, where it reaches about 100 m s^{-1} . This represents a rotation of the atmosphere in the same direction as that of the solid planet, but with a period of about 4 days within the cloud, about 50 times faster than the rotation of the solid planet. **Figure 3** shows that meridional (south–north) velocities are much smaller than zonal winds in the deep atmosphere.

The falling *Pioneer Venus* probes were only able to measure wind between the surface and about the 65 km level within the clouds. The ultraviolet features shown in **Figure 2** are slightly higher, near the cloud tops, probably at about 65–70 km elevation. Tracking of these features shows zonal motions consistent with probe measurements, but with the advantage that a complete latitudinal profile can be measured. These profiles vary over timescales of a year or so. Sometimes a well-defined high-latitude jet exists, and at other times the profile shows nearly constant angular velocity. The cloud top measurements also show poleward drift with a speed of roughly 10 m s^{-1} in both hemispheres. This may represent the upper branch of a Hadley circulation, and indeed the general appearance of **Figure 2** suggests a flow spiraling toward the poles. The Hadley interpretation may be premature, however, since cloud tracking is not possible on the dark side of the planet and a true zonal mean has not been measured. Solar-fixed tides are known to exist, and may make the dayside measurements unrepresentative of the mean.

The ultraviolet patterns in the Venusian clouds give evidence for large-scale traveling waves. One wave appears in the shape of a ‘Y’ lying over the equator with its base pointed eastward. The branches of the fork in the ‘Y’ merge into the poleward spiral at midlatitudes. The pattern remains coherent

over many rotations and travels with approximately the same velocity as the cloud top circulation. Another wave is occasionally observed at midlatitudes, traveling with a slightly slower velocity. The two waves are thought to be Kelvin and Rossby modes, respectively, but details of their structure and the nature of their excitation are not known.

The Venusian atmosphere is in *cyclostrophic* balance. Coriolis accelerations are weak because the planet rotates slowly. Cyclostrophic balance is the analog of Coriolis balance but with centrifugal replacing Coriolis accelerations. An ideal gas atmosphere in hydrostatic and cyclostrophic balance obeys an analog to the ordinary thermal wind equation of meteorology eqn [1],

$$H \frac{\partial}{\partial z} u_a^2 = -\cot \phi \left[\frac{\partial}{\partial \phi} (RT) \right]_{p \text{ constant}} \quad [1]$$

where the scale height $H = RT/g$, R is the gas constant, z is the height, u_a is the absolute zonal velocity in a nonrotating frame, ϕ is the latitude, and T is the temperature. The latitudinal derivative of temperature is taken at constant pressure. The rotation of the solid planet represents an equatorial velocity of $u_E = 1.8 \text{ m s}^{-1}$, and therefore u_a is the same as u displayed in **Figure 3**, but with a small correction equal to $u_E \cos \phi$. In the deep atmosphere this thermal wind equation can be used to infer temperature gradients from the measured vertical wind shear, and one finds that equator-to-pole contrasts of a few degrees are expected, with the high latitudes cooler. More probes at high latitudes would be necessary to measure this directly. Above the cloud tops, where no direct wind information exists, the measurements of temperature increasing toward the poles indicate that the atmospheric circulation decays with height in the middle atmosphere.

The remarkable spin of the Venusian atmosphere is one of the major puzzles in atmospheric science. In the absence of forcing mechanisms, friction and wave drag would be expected to bring the atmosphere into corotation with the solid planet. Theories fall into two classes. In one, vertically propagating waves transport momentum from the surface into the atmosphere. The waves would need to be selected or sorted by some process, so that those carrying momentum in the direction of the planetary spin are dominant. This explanation is similar to the accepted explanation for the quasibiennial oscillation (QBO) in Earth's stratosphere at low latitudes. The QBO reverses direction approximately every 2 years, however, and, over the 40 years that it has been observed, the direction of the Venus rotation has never reversed.

The other theory relies on a Hadley circulation plus large-scale eddies. The Hadley circulation on Venus is not well measured, but the general appearance of the cloud top circulation suggests that it extends from the equator to high latitudes. Rising motion at low latitudes transports angular momentum upward in such a circulation. Poleward drift in the upper branch leads to the formation of jets at middle and high latitudes. According to the theory, large-scale eddies cause erosion of the jets and act to transfer their angular momentum back toward low latitudes. The joint action of the Hadley circulation and the large-scale eddies can maintain the spin, or superrotation, of the atmosphere. This process has

been successfully simulated with numerical models. Observational test of the theory is not possible at present, however, because data on the Hadley circulation and on eddies are incomplete.

See also: **Solar System/Sun, Atmospheres, Evolution of Atmospheres:** Evolution of Earth's Atmosphere; Planetary Atmospheres: Mars.

Further Reading

- Bougher, S.W., Hunten, D.M., Phillips, R.J. (Eds.), 1997. *Venus II*. University of Arizona Press, Tucson, AZ.
- Counselman III, C.C., Gourevich, S.A., King, R.W., Loriot, G.B., Ginsberg, E.S., 1980. Zonal and meridional circulation of the lower atmosphere of Venus determined by radio interferometry. *Journal of Geophysical Research* 85, 8026–8031.
- DeMore, W.B., Yung, Y.L., 1982. Catalytic processes in the atmospheres of Earth and Venus. *Science* 217, 1209–1213.
- Rossow, W.B., Del Genio, A.D., Limaye, S.S., Travis, L.D., 1980. Cloud morphology and motions from Pioneer Venus images. *Journal of Geophysical Research* 85, 8107–8128.
- Seiff, A., Kirk, D.B., Young, R.E., et al., 1980. Measurements of thermal structure and thermal contrasts in the atmosphere of Venus and related dynamical observations: Results from the four Pioneer Venus probes. *Journal of Geophysical Research* 85, 7903–7933.
- Solomon, S.C., Bullock, M.A., Grinspoon, D.H., 1999. Climate change as a regulator of tectonics on Venus. *Science* 286, 87–90.

Solar Terrestrial Interactions: Climate Impact

JD Haigh, Blackett Laboratory, Imperial College London, London, UK

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

The evidence for influence of solar variability on the climate near the Earth's surface is discussed; the contribution of solar variability to 'global warming' of the past century is very small. Between the troughs and peaks of the '11-year' sunspot cycle the global average temperature varies by perhaps 0.1 °C. This signal is not, however, distributed uniformly across the planet, nor is it concentrated where solar irradiance is greatest – in the tropics, but found in particular regions. Some processes that might produce this pattern are outlined; changes in solar ultraviolet radiation may be the key.

Introduction

Many studies have shown an apparent response in weather or climate indicators to solar variability. At various locations, temperature, rainfall, surface pressure, cloud cover, storms, and droughts, among other meteorological parameters, have been found to correlate with measures of solar activity over the 11-year solar cycle as well as over periods extending from decades to centuries and longer. Some of these studies do not stand up to rigorous statistical analysis, and some appear to hold only over limited time periods, but there is mounting evidence of a solar influence on climate on many timescales. The response of the global average surface temperature is small, being approximately 0.15 °C warmer when the Sun is near maximum activity relative to minimum activity of the solar cycle. This warming does not, however, appear uniformly across the globe so, for example, greater warming is found in middle latitude air temperature while the eastern tropical Pacific Ocean appears to be cooler when the Sun is at peak activity.

The radiant energy output of the Sun varies by about 0.1% over the 11-year cycle which is consistent, using simple radiation balance estimates, with the global mean temperature response. Comparison of solar activity with estimates of global surface temperature back to the Maunder Minimum in sunspot numbers at the end of the seventeenth century suggests that the Sun has made a contribution to climate variability since that time but cannot alone account for the warming of the later half of the twentieth century.

The amount of solar radiation reaching the Earth is also modified by variations in the Earth's orbit around the Sun. These variations take place over periods of tens to hundreds of thousands of years and may be responsible for the occurrence of ice ages.

Variations in other factors than the total amount of solar radiant energy affect the atmosphere and possibly influence weather and climate. Solar ultraviolet (UV) emissions vary by several percent over the solar cycle and influence ozone production and heating of the stratosphere. The chemical structure of the stratosphere is also affected by high-energy protons and electrons emitted during solar flares and coronal mass ejections. Any change in the thermal structure of the stratosphere can influence the climate of the lower atmosphere through dynamical processes and these may contribute

to the nonuniform distribution of the solar signal in surface temperatures.

Alterations in the solar magnetic field affect the flux of galactic cosmic rays (GCRs) reaching the Earth and thus the strength of the Earth's electric field and ionization rates in the lower stratosphere. It is plausible, but unproven, that these result in changes in thunderstorm activity or cloud cover.

Observations

Data Analysis

Variations in the amount of solar radiation reaching the Earth produce the familiar diurnal and seasonal variations, but the effects of changing solar *activity* on weather and climate are much more difficult to observe. The detection of a solar influence on a meteorological record inevitably relies on statistical analysis and an unequivocal identification with the Sun of any derived signal must allow for both the huge natural variability of the climate system, such as associated for example with El Niño–Southern Oscillation (ENSO), and also the possible influence of other factors, such as increases in greenhouse gas concentrations or volcanic eruptions.

One commonly used approach is to investigate time series of data for periodicities associated with solar activity (e.g., around 11, 22, or 80 years). Another simple statistical approach is to estimate the degree of correlation between a meteorological parameter and some measure of solar activity. Again, the main difficulty with these methodologies is the possibility that the natural (unforced) variability of the climate system includes components of the same periodicity. More sophisticated methods explicitly take account of the effects of other influences and/or understanding of the innate statistics of the system (e.g., autocorrelation in time or space).

Influence on Solar Cycle Timescale

The clearest indication of changes in solar activity, noted since ancient times, is the number of sunspots on the solar surface. There are many other indicators, including solar radiation emissions across the electromagnetic spectrum and at specific wavelengths, particularly in the microwave and UV regions. Solar activity is also indicated by the strength of the Sun's magnetic field and in the way it influences the Earth's magnetic

field and the incidence of the GCRs reaching the Earth. The occurrence of solar storms, and the emission by the Sun of high-energy particles, also indicates high levels of activity. All these indicators have been observed to vary cyclically with a period ranging between 9 and 13 years, often referred to as the '11-year cycle'.

The observed amplitude of the solar cycle oscillation in global mean surface temperature is only about 0.15°C , but the geographical distribution is not uniform and larger signals have been observed in some regions of the globe. Cycles of 10–12 year periodicity have been isolated in records including sea surface temperature, surface temperature at many land stations, rainfall in the USA and Africa, surface pressure in the North Atlantic and North Pacific Oceans, North American forest fires, Atlantic tropical cyclones, tropical corals, and the Southern Oscillation.

An example can be seen in **Figure 1** which presents the difference in near-surface air temperature over land, and in sea level pressure, between winters of low and high solar activity in the North Atlantic region. When the Sun is less active, the

Northwest Europe and central Eurasia experience colder winters while Greenland is warmer. The patterns are consistent with the atmosphere tending to be in a more negative phase of the North Atlantic Oscillation at lower solar activity, a signal confirmed by other analyses/datasets.

Another regional response is found in sea surface temperatures where there is evidence that higher levels of solar activity are associated with a La Niña-like signal of cooler temperatures in the eastern tropical Pacific.

Away from the surface, solar cycle signals have been detected throughout the atmosphere, with very large changes, of over 500°C , in the thermosphere and smaller responses at lower altitudes. The tropical upper stratosphere, near 50 km altitude, is around $1\text{--}2^{\circ}\text{C}$ warmer when the Sun is more active. In the polar lower stratosphere in winter, the signal is enhanced significantly when the data are sorted according to the phase of the Quasibiennial Oscillation (QBO), although care has to be taken that the analysis remains statistically sound.

Figure 2 shows the amplitude of the solar cycle signal in temperature throughout the lower atmosphere. The largest

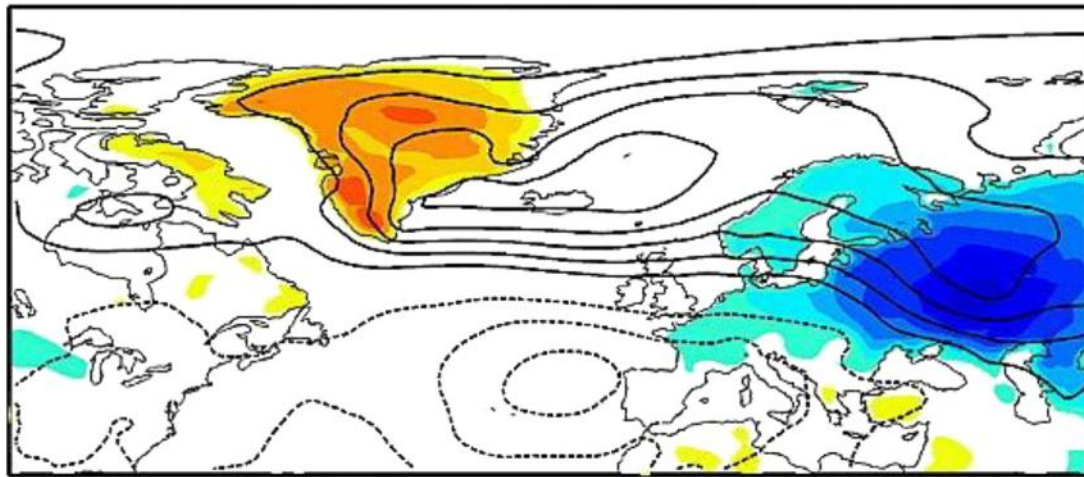


Figure 1 Difference between winters of low and high solar activity (as indicated by value of open magnetic flux) between 1957/1958 and 2000/2001. Colors show difference in near-surface air temperature over land ranging from $+2.0^{\circ}\text{C}$ (dark orange) to -3.5°C (dark blue). Contours show the difference in sea level pressure, the contour interval is 1 hPa and dashed contours represent negative values. Reproduced from Woollings, T., et al. 2010. Enhanced signature of solar variability in Eurasian winter climate. *Geophysical Research Letters* 37: L20805, (<http://dx.doi.org/10.1029/2010GL044601>).

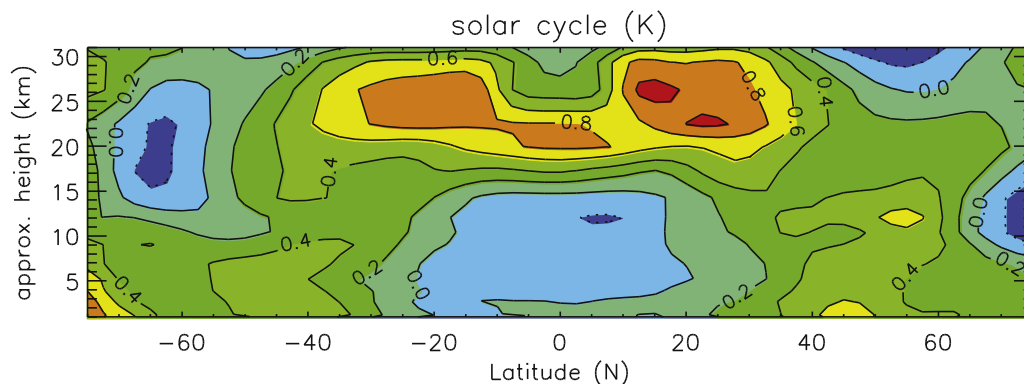


Figure 2 The signal in zonally averaged and deseasonalized temperature ($^{\circ}\text{C}$) found in response to the solar cycle (high minus low solar activity as indicated by the radio flux) over 1979–2002 with the effects of other influences (QBO, ENSO, volcanoes, and long-term climate change) having been removed. Reproduced from Haigh, J.D., 2003. The effects of solar variability on the Earth's climate. *Philosophical Transactions of the Royal Society A* 361: 95–111.

response, of up to 1 °C (solar maximum relative to minimum), is found in the subtropical lower stratosphere with bands of warming, of about 0.4 °C extending through the midlatitude troposphere. This pattern is different from what would be anticipated as a direct response to increased radiative heating and, along with the QBO modulation discussed above, suggests that changes in atmospheric motions play a role in determining the observed solar impact on temperatures.

Decadal-Centennial Scale Influence

The solar cycle fluctuates not only in length but also in amplitude. At 'solar minima' the sunspot number decays to near zero, while at 'solar maxima' it is very variable. At various times in the past, over periods of several decades, sunspot numbers during solar maxima have been observed to become very small – these periods are referred to as Grand Minima. For example, sightings of sunspots were rare over the entire second half of the seventeenth century, an episode now called the Maunder Minimum.

Variations associated with the 11-year cycle are useful for diagnosing signals in local meteorological records, and they can also indicate potential mechanisms whereby the solar impacts are produced, but unless the effects are somehow accumulative they do not imply any long-term influence on climate. On longer timescales, the apparent coincidences of the 'Little Ice Age' from the mid-sixteenth to mid-nineteenth centuries with the Spörer and Maunder solar minima, and of the 'Medieval Warm Period' between about 900 and 1300 with the Medieval solar maximum, have often been cited as inferring a solar influence on global temperatures. More detailed analysis of the temperature signal has revealed, however, that the responses in different geographical regions do not vary synchronously. This makes the link with solar activity ambiguous, and suggests that other factors, including inherent natural variability in the system, must play a part in determining the temporal variation. The 'Little Ice Age' is now viewed as a fairly modest average cooling of the Northern Hemisphere which may be partly ascribed to solar variability, but for which volcanic activity and changes in ocean circulation may also be important.

Long-Term Influence

Estimates of variation in atmospheric temperature over much longer periods may be derived from oxygen isotope ratios in glaciers and ice sheets, lake sediments, ocean sediments, and corals. Spectral analysis of these datasets shows that the dominant component in the record for the last 800 000 years has a period of around 100 000 years. This is of the same order as the periodicity of the eccentricity of the Earth's elliptical orbit around the Sun suggesting a direct solar influence on climate. However, the changes in solar irradiance associated with the variations in eccentricity are too small on their own to explain the observed temperature signal. They are also much smaller than the latitudinal and seasonal deviations in irradiance due to variations in other orbital parameters. These are the tilt of the Earth's axis, which demonstrates a periodicity of 41 000 years, and the precession of the equinoxes, with periodicities of 19 000 and 23 000 years. These periods are less evidenced in

the paleoclimate records indicating that other factors than simply the instantaneous total solar irradiance contribute to the long-term solar influence on climate.

Mechanisms for Solar–Climate Interaction

On the long timescales a key factor is the change in the seasonal intensity of solar radiation. For example, if increased summer irradiance is insufficient to melt the extended ice sheet resulting from a colder winter, then the climate system may be propelled into an ice age. Current General Circulation Models (GCMs) are not able to reproduce ice sheet growth suggesting that the representation of the processes may be incomplete (feedbacks involving vegetation cover and type may be an example) or that nonlinear interactions between the different processes are not properly included.

On solar cycle and decadal–centennial timescales, the observational analyses described above provide evidence for an influence of solar variability on climate but, without a clear understanding of the mechanism(s) through which such interactions take place, doubts may remain that the periodicities and correlations found are due to natural internal atmospheric variability or to other climate forcing factors. The most direct means whereby solar variability may affect climate is through modulation of the total solar radiative energy received by the Earth. Other, more indirect, means might be through modification of atmospheric chemical composition or of the Earth's electric field or of cloud formation.

Radiation Balance

The total radiative power emitted by the Sun crossing unit area at the Earth's mean orbital distance is approximately 1361 W m^{-2} , although differences in absolute calibration of the spaceborne radiometers mean this value is uncertain to a few watts per square meters. Historically, this was referred to as the 'solar constant' because it was believed not to vary but measurements made from satellites since the late 1970s have shown that total solar irradiance (TSI) changes by about 1.4 W m^{-2} , or 0.1%, over the 11-year cycle with higher values corresponding to periods of greater solar activity as indicated by sunspot numbers. Based on these measurements, and on historical observations of an indicator of solar activity such as sunspot number, it is possible to estimate values of TSI in the past. In doing this, however, it is also necessary to make some assumption about the slowly varying change in solar activity which underlies the 11-year cycles known and is known as the quiet Sun. There is considerable uncertainty in the magnitude of this component so that recent estimates of the difference in TSI between the late seventeenth century (during the Maunder Minimum in solar activity) and the present range from about 1.3 to 6.5 W m^{-2} .

The equilibrium change in global average surface temperature, T_s , due to a change in the Earth's radiation balance may be estimated from the expression $\Delta T_s = \lambda \Delta F$, where ΔF is the imbalance in global average radiative fluxes and λ a climate sensitivity parameter which represents the response of the surface temperature to applied radiative perturbations, taking into account atmospheric feedback mechanisms through, changes in humidity

or cloud cover. λ is estimated from atmosphere–ocean GCMs to be about $0.625 \text{ K W}^{-1} \text{ m}^{-2}$. An increase in TSI in the range of $1.3\text{--}6.5 \text{ W m}^{-2}$, since the Maunder Minimum, corresponds to a change in radiative flux ΔF of $0.23\text{--}1.14 \text{ W m}^{-2}$ taking account of global averaging and terrestrial *albedo*. Using the climate sensitivity value above this would suggest, using the expression above, a surface warming in the range of $0.14\text{--}0.71 \text{ K}$ due to the Sun since about 1700. The observed overall rise in temperature over that period is about 0.7 K , so the range in TSI might suggest that anything between one-fifth and all global warming is due to the Sun. The larger value is very unlikely, however, as it would pose two inconsistencies: (1) temperature variations in response to solar activity over earlier times should have been much larger, and (2) the evolution of the temperature variation since 1700 should correlate better with the TSI reconstruction. Clearly, more work is needed to establish the historical variations in TSI as well

as to advance understanding of the physical processes in the Sun that produces them.

Another problem associated with simply associating changes in temperature with differences in TSI is the neglect of other potential factors. One attempt to distinguish these in global mean surface temperature over the past century is portrayed in Figure 3. This shows, by the black curve in panel (a), the time series of the temperature measurements together with, in the other panels, those of the factors assumed to be contributing: ENSO, volcanic aerosols, solar irradiance, and human activities (mainly greenhouse gas emissions and industrial aerosol production). The amplitudes (as opposed to the temporal variations) of the contributions are not specified in advance but are determined by a statistical analysis which produces the best fit to the observed temperature data. The sum of the deduced contributions is portrayed by the orange curve in panel (a).

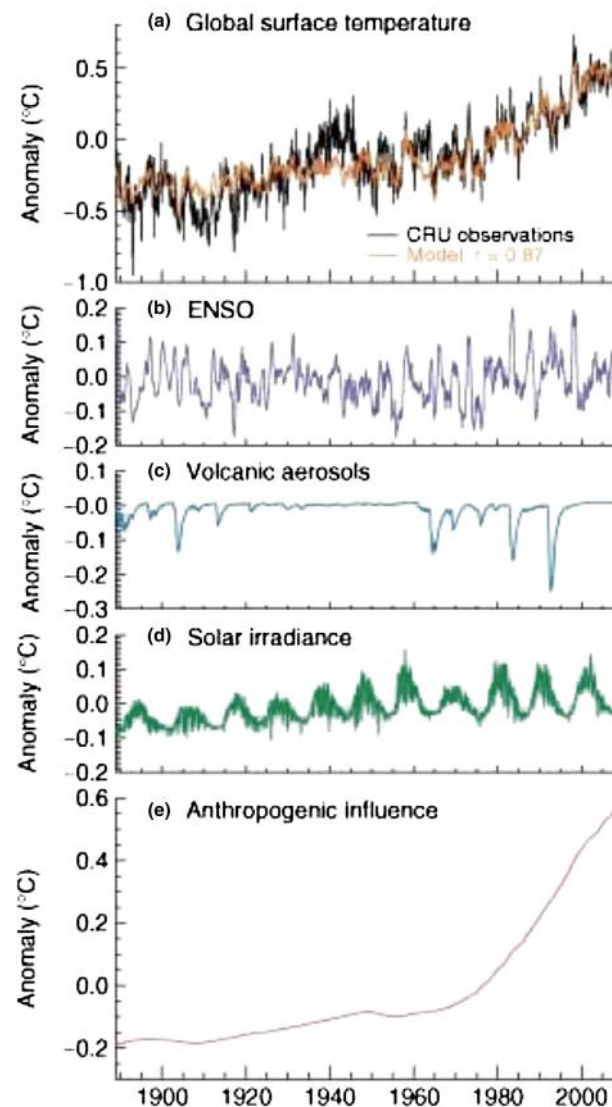


Figure 3 Global mean near-surface temperature 1889–2008 (a) from observational records (black) and reconstructed (orange) from the components associated with (b) ENSO, (c) volcanic aerosols, (d) solar irradiance, and (e) human activity. Reproduced from Lean, J.L., 2010. Cycles and trends in solar irradiance and climate. Wiley Interdisciplinary Reviews: Climate Change 1: 111–122, <http://dx.doi.org/10.1002/WCC.18>.

The 11-year cycle in irradiance makes a small contribution to the temperature variations while the underlying secular increase in irradiance over the first half of the century appears to contribute about one-third of the upward trend in temperature over that period. The Sun does not contribute to global warming over the later half of the century. The solar irradiance curve in panel (d) of Figure 3 is one which possesses an underlying trend (relative to the 11-year cycle amplitude) at the low end of current estimates but no reasonable reconstruction of solar irradiance over this period, however, can account for the significant warming of the past half century.

The La Niña-like signal, outlined above, of cooler sea surface temperatures in the tropical Pacific associated with higher levels of solar activity, may be associated with the variation in TSI. Differential heating of the sea surface between cloudy and clear regions might induce a change in surface wind stress and thus in the meridional overturning circulation of the upper ocean. Further work needs to be done, however, to definitively extract a solar signal, given the large innate variations in sea surface temperature related to ENSO.

UV

The variability in TSI, discussed above, of a few tenths of 1% represents the change integrated across the whole electromagnetic spectrum. Measurements made by satellite instruments, and also the results of theoretical models, show much larger fractional changes in the UV radiation which is absorbed higher in the atmosphere.

Most of the visible/near-infrared radiation passes through the atmosphere unhindered to the tropopause. Water vapor bands in the near-infrared cause some absorption in the lower troposphere, and some radiation is scattered by cloud, but most of this radiation reaches the surface. Shorter (UV) wavelengths, however, are absorbed higher in the atmosphere where they cause local heating and ozone production. The increased ozone tends to mask the lower atmosphere from the enhanced incident UV while the warmer stratosphere will cause increased emission of thermal infrared (TIR) radiation. Thus the changes in the UV and TIR radiations reaching the troposphere depend on the variations in ozone. However, the response of ozone to solar activity is not very well established. Multiple regression analysis of satellite ozone measurements suggests largest changes in the upper and lower stratosphere and a much smaller change in the middle stratosphere. However, the data are available only for about two solar cycles and have large uncertainties, especially in the lower stratosphere. Furthermore, it is difficult to extract a 'pure' solar signal separately from the influence of volcanic eruptions and ENSO. Coupled chemistry climate models have limited success in reproducing the signal extracted from the satellite record, and generally do better if variations in sea surface temperatures are included, possibly due to the mixing of the solar signal with ENSO.

Changes in stratospheric thermal structure may also affect the troposphere through dynamical interactions rather than through radiative forcing. Studies using computer models have shown that variations in solar UV radiation, which primarily heats the stratosphere, produce regions of preferred response to

solar activity in the troposphere and at the surface. These are produced by dynamical adjustments within the atmosphere, i.e., changes in winds and circulation patterns, rather than by direct radiative heating. Several such studies have reported a solar signal in the strength and extent of the tropical overturning circulations (Hadley and Walker cells) and in the positions of the middle latitude storm tracks. They also suggest that at higher solar activity, the circulation tends to favor more positive phases of the Northern and Southern Annular Modes. These responses are consistent with some of the observed signals at the surface, outlined earlier in this article, which shows a geographically nonuniform response to variations in solar input, with surface temperatures at some locations showing heating and some cooling during periods of higher solar activity.

Solar Energetic Particles and GCRs

Solar storms and flares occur more frequently during periods of higher solar activity. As a result of these events, high-energy particles (protons, electrons, and alpha particles) are ejected. These can enter the Earth's atmosphere along the open magnetic field lines near the polar caps. Particles with energies of $> \sim 10$ MeV penetrate down to the middle atmosphere where they cause an increase in the concentration of nitrogen oxides through ionization and dissociation of N_2 and O_2 . The NO produced can cause significant reduction in middle atmosphere ozone concentrations. Higher energy particles can penetrate deeper into the atmosphere and to lower latitudes. Although individual solar particle events only last on the order of a few hours the chemical perturbations may persist for several months, propagating downward and equatorward and possibly altering stratospheric dynamics. The climate impact of this is likely to be small but has not been studied in detail.

GCRs are particles that are formed outside of the solar system and bombard it from all directions. The flux of GCRs reaching the Earth is modulated by interaction with magnetic structures advected with the solar wind such that at times of higher solar activity the GCR flux is reduced by about 20% with respect to periods of lower solar activity. The flux into the atmosphere is also affected by the Earth's magnetic field such that it is greater at higher latitudes. The GCRs which do penetrate the atmosphere are a major source of ionization, particularly in the lower stratosphere.

There are two main theories advanced as to the means whereby variations in GCR flux might impact climate. The first concerns modulation of the Earth's electric field. Ionization of the atmosphere by GCRs, or indeed by solar energetic particles, will influence its conductivity and thus the current flow within the Earth's electric circuit. It is possible that the current flow into clouds affects cloud microphysics, ice formation, and hence thunderstorm activity, although details of this mechanism are not fully established. While GCRs are more prevalent during periods of minimum solar activity, solar energetic particles are more numerous at solar maximum so that their combined effect over the solar cycle is not clear.

The second means whereby it is proposed that GCRs may affect climate is through enhancing the production of cloud

condensation nuclei through growth of aerosol on ionized air molecules. Some evidence for the existence of this process has been obtained in observational and modeling studies, although not clearly in response to solar activity.

See also: **Climate and Climate Change:** Climate Variability: Decadal to Centennial Variability. **Global Change:** Climate Record: Surface Temperature Trends; Upper Atmospheric Change. **Middle Atmosphere:** Quasi-Biennial Oscillation. **Ozone Depletion and Related Topics:** Photochemistry of Ozone. **Solar System/Sun, Atmospheres, Evolution of Atmospheres:** Solar Winds. **Tropical Meteorology and Climate:** Hadley Circulation.

Further Reading

- Gray, L.J., Beer, J., Geller, M., et al., 2010. Solar influence on climate. *Reviews of Geophysics* 48, RG4001. <http://dx.doi.org/10.1029/2009RG000282>.
- Haigh, J.D., 2007. The Sun and the Earth's climate. *Living Reviews in Solar Physics* 4. <http://www.livingreviews.org/lrsp-2007-2>.
- Lean, J.L., 2010. Cycles and trends in solar irradiance and climate. *Wiley Interdisciplinary Reviews: Climate Change* 1, 111–122. 10.1002/WCC.18.

Solar Winds

ST Suess, NASA Marshall Space Flight Center, Huntsville, AL, USA
BT Tsurutani, Jet Propulsion Laboratory, Pasadena, CA, USA

© 2003 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 5, pp 2078–2095, © 2003, Elsevier Ltd.

Introduction

The Sun is losing mass in form of the solar wind, which has affected its evolution from its birth and will continue to do so until its death. This is not unusual, in that nearly all stars are losing mass through stellar winds throughout a major portion of their lives. As far as the Earth is concerned, the solar wind blows against the Earth's magnetosphere, causing auroras and geomagnetic storms.

The Corona

The surface of the Sun is called the photosphere, above which lies the Sun's atmosphere, known as the corona. The solar wind forms in the corona and is caused by high pressure in the corona relative to the low pressure far from the Sun in the interstellar medium. This pressure gradient exerts an outward force against gravity and accelerates the wind from low speeds in the low corona up to supersonic speeds at 2–10 solar radii ($R_{\odot}$). To give a sense of scale, the Earth is 1.5×10^8 km = $215R_{\odot}$ from the Sun (defined as 1 astronomical unit or 1 AU). Typical solar wind speeds beyond $10R_{\odot}$ are between 300 and 800 km s⁻¹ so it takes the solar wind 2.2 to 5.8 days to reach Earth. The existence of the solar wind was inferred prior to the space age from the existence of auroras, disturbances to the Earth's magnetosphere, and observations of comet tails. Today it is regularly observed with a number of spacecraft.

Coronal Expansion

Pressure in the corona is high because the temperature is high, more than 10^6 K, relative to the photospheric temperature of ~ 5800 K. This is a sufficiently high temperature that the corona emits copious X-rays. It is believed that the corona is heated to this high temperature as a by-product of magnetic field motions, interactions, and instabilities in the photosphere that directly transfer energy into the corona. This energy flux could be via direct heating, waves, jets of material, or other forms, but this is unknown and is the subject of research by several different observatories in space and a deep-space mission called Solar Probe that will travel to within $3R_{\odot}$ of the photosphere.

The corona is composed mainly of protons and electrons (ionized hydrogen), with minor amounts of silicon, carbon, iron, oxygen, and other elements. There is about 20% helium (by mass) that can be observed spectroscopically; this is how helium was first discovered – hence its name (after *Helios*, the Sun god of Greek mythology). All the components share in the expansion of the corona and can be measured *in situ* by spacecraft outside the Earth's magnetosphere.

The Sun's magnetic field makes the solar wind far from a simple spherical expansion of a hot gas. The magnetic field is dipole-like but undergoes a reversal during every 11-year solar sunspot cycle. At sunspot minimum the field is aligned with the solar rotation axis, while at solar maximum many sunspots appear and the dipole field weakens and becomes irregular. From solar maximum to minimum the field again becomes dipolar, but is inclined relative to the rotation axis. These changes are reflected in changes in coronal structure, which can be seen during solar eclipses such as that shown in [Figure 1](#). With the bright disk of the Sun being occulted by the moon, the faint corona becomes visible, primarily because of light coming from the photosphere being reflected off of electrons in the corona. The areas that are bright are regions of high density and are known as streamers. The dark regions at the top and bottom in [Figure 1](#) are coronal holes. The streamers in [Figure 1](#) lie over the magnetic equator and the density is higher because ions and electrons are trapped on closed loops of the dipolar magnetic field. The low-density coronal holes mark the locations of the north and south magnetic poles. [Figure 1](#) was taken in 1994, just prior to solar minimum, so that the magnetic axis was only slightly inclined to the rotational axis, which is vertical in this image.

[Figure 2](#) is a schematic of the stages of coronal evolution over the 11-year sunspot cycle, starting at solar maximum on the left. [Figure 1](#) is represented by the drawing in [Figure 2\(c\)](#). The magnetic field loops in streamers are shown here to help suggest why the ions and electrons are trapped, just as iron filings tend to align along magnetic field lines around a bar magnet. Coronal holes are shown by the dark areas on the solar disk. This figure also indicates that fast solar wind originates from coronal holes and slow solar wind from above streamers. Fast solar wind has speeds above ~ 600 km s⁻¹ at 1 AU and slow solar wind has speeds below ~ 500 km s⁻¹. This division

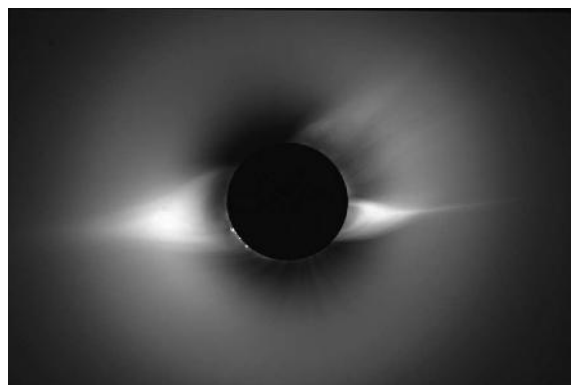


Figure 1 Total solar eclipse as seen from Putre, Chile on 3 November 1994. Photograph courtesy of High Altitude Observatory, National Center for Atmospheric Research.

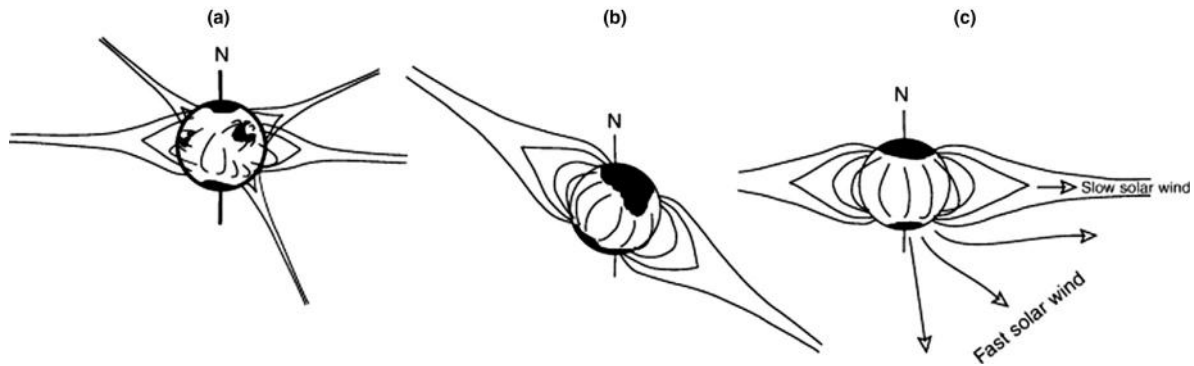


Figure 2 Schematic illustration of the three stages in the 11-year solar sunspot cycle. (a) Solar maximum, when the corona is filled with streamers and there are few or no coronal holes. There is no well-defined large-scale field. (b) Declining phases when the large-scale field is dipole-like and inclined to the heliographic equator. (c) Solar minimum when the field is dipolar, aligned with the rotation axis, and when the polar coronal holes are largest.

into fast and slow wind could be observed if one were able to pass around the Sun as shown in Figure 2(c) from south pole to north pole. It would then be possible to sample first fast wind from the south pole, slow wind from over the equatorial streamers, and then fast wind again from the north pole. The *Ulysses* spacecraft carried out this exercise in 1995–1996 at solar sunspot minimum and a plot of the observed solar wind speed is shown in Figure 3 using what is called a dial plot. The dial

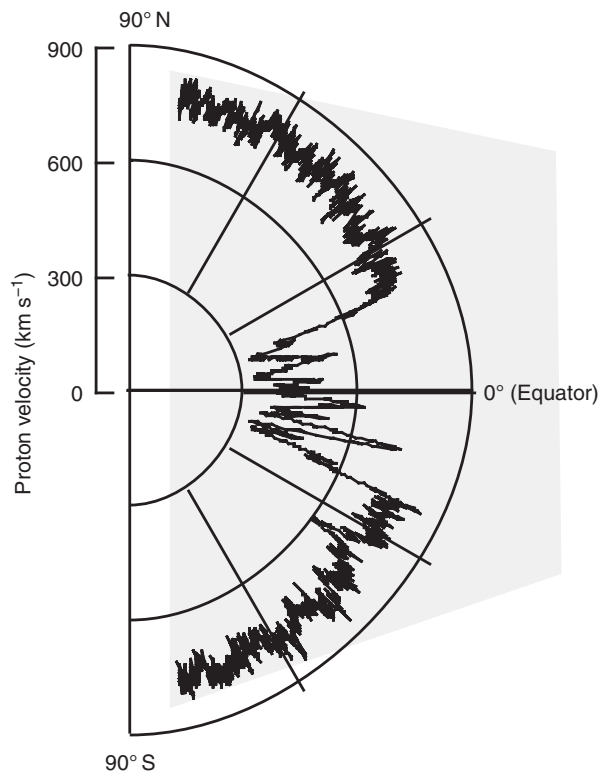


Figure 3 Dial plot of solar wind speed, indicated by radial distance from the origin, as a function of heliographic latitude, measured around the origin of the plot. Data were collected by the *Ulysses* solar wind plasma instrument between September 1994 and July 1995, during which time *Ulysses* swept from 80° south latitude to 80° north latitude.

plot indicates the solar latitude around the origin and the measured solar wind speed as distance from the origin. The fast wind in the north and south is very clearly divided from the slow wind above the equatorial streamers in this plot. This demonstrates one of the major discoveries in recent years — that fast and slow solar wind represent two distinct states between which there is no continuous change. Fast wind comes from coronal holes and is rather smooth and uniform at 1 AU. Conversely, slow wind is relatively irregular and comes either from thin boundaries around streamers or leaks somehow from within streamers. Figure 4 shows profiles of how fast and slow wind vary with distance from the Sun, illustrating not only that the speeds are different but also that there are characteristic densities and temperatures differences. T_e and T_p are the proton and electron temperatures in this plot. The distinct difference between the two solar wind states leads to important consequences because of solar rotation.

Solar Rotation and the Magnetic Field in the Solar Wind

Solar wind is an ionized gas made up primarily of protons and electrons with minor ions in amounts similar to those in the corona. The electrons and ions are very tightly bound to lines of magnetic flux, again like the coronal plasma in streamers. However, the magnetic field in the solar wind is relatively weak and thus is carried along by the solar wind. The rotation of the Sun results in the lines of magnetic flux in the solar wind being drawn into Archimedean spirals. This occurs because the Sun revolves once every ~ 25.5 days while, as mentioned above, it takes solar wind several days to reach 1 AU. Therefore, the Sun revolves through a significant angle during the time it takes the solar wind to reach the Earth. For example, taking a typical speed of 400 km s^{-1} , it takes solar wind 4.34 days to reach 1 AU. During the same time, the Sun will have revolved through about 60° , or about $1/6$ of a full rotation. The magnetic field in the solar wind, called the interplanetary magnetic field, or IMF, is attached to the Sun at the point where the solar wind began. Thus, the point on the field line attached to the Sun is turned through an angle of 60° relative to the point on the magnetic field line that is at 1 AU. The field line between the Sun and 1

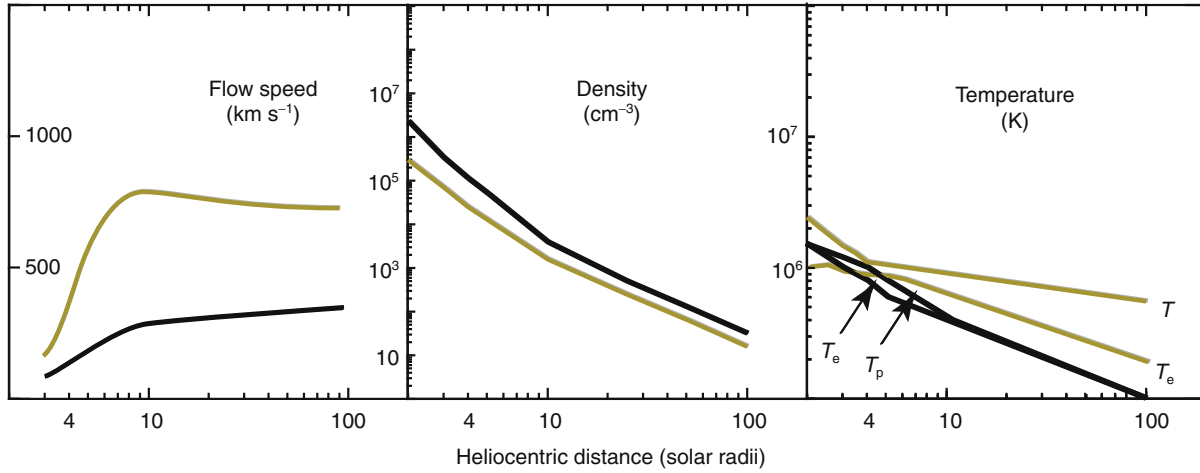


Figure 4 Solar wind flow speed, density, and temperature between $2R_S$ and $100R_S$, for coronal holes (yellow lines) and streamers (black lines). These are typical values, with the possible range around these values being quite large.

AU traces a continuous curve between these two points. Assuming the solar wind speed, v (km s^{-1}), is independent of distance from the Sun, this curve is described by eqn [1].

$$r - r_0 = \frac{-v}{\Omega \cos \theta} (\phi - \phi_0) \quad [1]$$

In eqn [1], r is the distance from the center of the Sun in km, $r_0 = 6.96 \times 10^5$ km is the radius of the Sun, $\Omega = 2.85 \times 10^{-6} \text{s}^{-1}$ is the angular velocity of the sun, and $(\phi - \phi_0)$ is the difference in longitude (in radians) at the two points on the field line. θ is solar latitude and the Earth lies in the range $-7.25 < \theta < 7.25$ degrees because the plane of the ecliptic is inclined to the solar equator by 7.25° . The angle $(\phi - \phi_0)$ is also the angle between the magnetic field line and the radial direction at 1 AU, or wherever eqn [1] is evaluated. This is called the spiral angle. The geometry of the curved field line is precisely an Archimedean spiral when v is constant and this is one of the important predictions made by E. Parker when he developed his theory for the solar wind in the 1950s and 1960s. Figure 5 illustrates two spirals computed using eqn [1]. The tighter spiral above results from low speeds, $< 500 \text{ km s}^{-1}$, and the spiral angle is $> 45^\circ$ at 1 AU. Conversely, the spiral angle at 1 AU is $< 45^\circ$ for speeds $> 500 \text{ km s}^{-1}$. Parker predicted that $(\phi - \phi_0) \sim 45^\circ$ (0.785 radians) at 1 AU and this is what has been measured for the average spiral angle by several different spacecraft.

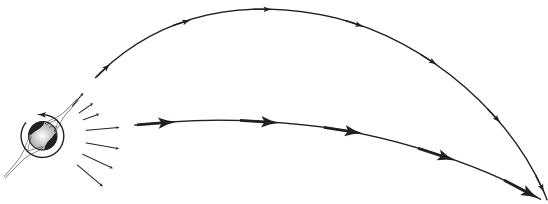


Figure 5 Diagram of spiraling interplanetary magnetic field (IMF) lines. The dependence on solar wind speed is illustrated by the more curved line at the top being for relatively slow wind and the less curved line at the bottom being for fast wind.

Corotating Interaction Regions

Solar rotation has an important effect on coronal expansion through the interaction of fast and slow wind. During the declining phases of the solar cycle, (Figure 2(b)), regions on the Sun producing slow wind will sometimes face the Earth and at other times regions producing fast wind will face the Earth. Thus it will often be the case, especially during declining phases of the solar cycle, that slow wind will be followed by fast wind. This is just the example diagrammed in Figure 5. When this happens, fast wind overtakes slow wind, the gas in between becomes compressed, and eventually shocks form with forward shocks moving away from the Sun and reverse shocks moving toward the Sun in the frame of reference moving with the solar wind. This is called a corotating interaction region (CIR) because it appears stationary in the frame of reference rotating with the Sun. As the plasma between the fast and slow wind becomes compressed, the velocity profile is dynamically altered and the CIR becomes stronger and stronger with increasing distance until the shock forms. A simple estimate for where the shocks will form can be made using eqn [2], where the same definitions are used as in eqn [1].

$$r - r_0 = \frac{v_1 v_2}{v_2 - v_1} \frac{(\phi_2 - \phi_1)}{\Omega \cos \theta} \quad [2]$$

The quantity $(\phi_2 - \phi_1)$ is the difference in longitude of the source regions of fast and slow wind, v_1 and v_2 are the slow and fast wind speeds, respectively, and r is the estimated distance for shock formation. Taking $(\phi_2 - \phi_1) = 0.53$ radians $= 30^\circ$, $v_1 = 400 \text{ km s}^{-1}$, and $v_2 = 800 \text{ km s}^{-1}$ gives $r = 1.5 \times 10^8 \text{ km} = 1 \text{ AU}$. During the declining phases of the solar cycle it is observed that shocks generally form around 2 AU, which is consistent with eqn [2] since $(\phi_2 - \phi_1)$ is more nearly 1 radian than 0.5 radians at those times. Forward shocks are rarely observed at 1 AU and reverse shocks are only observed in $\sim 20\%$ of CIRs at 1 AU. Equation [2] was derived simply by calculating when the two field lines shown in Figure 5 would cross. These field lines are the same as the streamlines in the frame of reference corotating with the Sun, and this is why eqn [2] looks closely related to eqn [1].

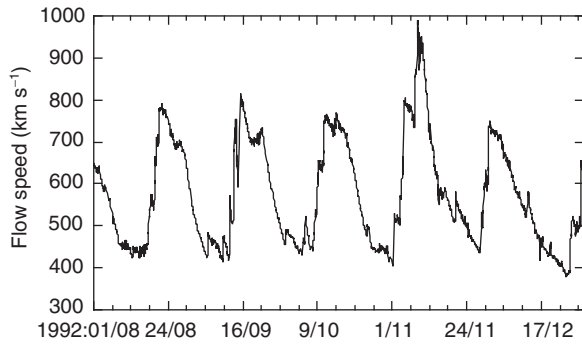


Figure 6 Solar wind speed at *Ulysses* during August–December 1992 when *Ulysses* was near the heliographic equator and at ~ 5 AU. Five corotating interaction regions (CIRs) are shown, occurring approximately every 25.5 days, or each solar rotation. Viewing the plot from left to right, each CIR is characterized by a sharp speed increases at forward and reverse shocks at the front of the CIR, followed by the speed maximum. The speed then decreases to a minimum before increasing in the next CIR. The very high speed on 10 November 1992 is due to a coronal mass ejection on top of the CIR.

CIRs have a very distinctive character, as seen in the long series of CIRs observed by *Ulysses* in 1992 when it was near the Sun's equator. About five solar rotations of the data are shown in [Figure 6](#). At the time *Ulysses* was at ~ 4 AU and fast wind had overtaken slow wind to form shocks where the speed is seen to abruptly jump upward as time progresses from left to right. CIRs have important consequences for the Earth since they can produce auroral activity and magnetic storms when they strike the Earth's magnetosphere if the IMF is also directed southward so that it can easily merge with the Earth's magnetic field. CIR-associated magnetic storms naturally tend to recur every solar rotation – 27 days as viewed from the Earth owing to the Earth's motion around the Sun. This activity also has a distinct solar cycle signature as the Sun moves through the phases diagrammed in [Figure 2](#). Thus, observation of coronal holes and streamers and the phase of the solar cycle provides a basic tool for the prediction of space weather and geomagnetic activity. A further consequence of CIRs is that the resulting shock waves produce large numbers of high-energy particles or cosmic rays. These particles affect the Earth's ionosphere and the radio communications that depend on the ionosphere.

Coronal Mass Ejections

Up to this point, a picture of the solar wind has been drawn that depicts it as quasi-steady, changing only slowly over the 11-year solar sunspot cycle. This is not an accurate picture at any time, especially near solar maximum. There are many forms of solar activity, including flares and erupting prominences, but the most dramatic is the release of a coronal mass ejection, or CME. A picture of a CME is shown in [Figure 7](#). This picture was taken from the SOHO spacecraft using a telescope called LASCO, which places an occulter over the solar disk so that the corona becomes visible, producing an artificial solar eclipse. The occulter is twice the size of the Sun, and the disk of the Sun is indicated by the white circle. Off to the lower right of the image is the CME. These are seen throughout the entire

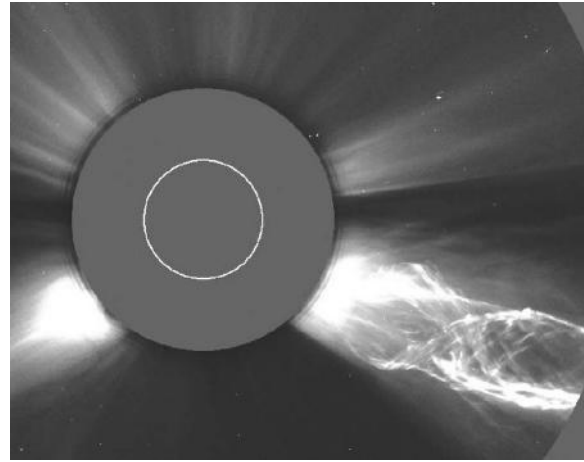


Figure 7 A coronal mass ejection is seen in the lower right quadrant in this image from the LASCO coronagraph on SOHO. The Sun, which is covered by an occulter that is $4R_{\odot}$ in diameter, is indicated by the white circle.

solar cycle, but they are 5–10 times more common near solar maximum, occurring at a rate of 3–4 per day. They occur in and near streamers, confined to low latitudes near solar minimum but reaching all latitudes at solar maximum.

When an interplanetary CME (ICME) strikes the Earth, the consequences are similar to those of a large CIR. The magnetosphere is compressed, auroral activity increases, and a magnetic storm or substorm may occur if the IMF in the ICME is directed southward. Ionospheric activity is also affected. This is therefore a phenomenon that is actively monitored in the context of space weather.

One CME is visible in the data shown in [Figure 6](#). At about 10 November 1992 the solar wind speed increased to ~ 1000 km s^{-1} . This is above any speed for simple fast solar wind. Instead, what is seen here is a fast ICME that has overridden a CIR. This could have a doubly strong impact on the magnetosphere owing to the large speed enhancement.

ICMEs are another phenomenon in the solar wind that is only partially understood. The propagation of an ICME can be modeled fairly well using computers and a numerical solution of the equations of motion. However, the basic mechanisms causing the initiation of a CME are not known. CMEs are related to solar magnetic activity such as flares and erupting prominences, but that relationship is not so simple that one can predict a CME for anything except the very largest of these events.

The Solar Wind over the Life of the Sun

The IMF is not completely passive in the solar wind. Because it is attached to the Sun, and has a small, but finite strength, the IMF tends to cause the solar wind to rotate with the Sun out to some distance above the photosphere. In doing this, the IMF causes angular momentum of the Sun to be transferred to the solar wind. Generally this is a small effect, with the corotation distance being $10\text{--}20R_{\odot}$ at most, or 0.1 AU. However, over the life of the Sun, the effect can be important. Calculations of the angular momentum transfer suggest that the present-day solar

wind and IMF could easily have doubled the rotation period of the Sun, from 12.25 to 25.5 days, over the 4.5 billion year life of the Sun.

Presently the solar wind carries away only a very small amount of mass from the Sun – so small that if assumed the same for 4.5 billion years it would have removed only $\sim 0.01\%$ of the total mass of the Sun. However, the Sun changes over its life, as do all stars. The Sun probably had a very vigorous wind early in its life when the solar convection zone extended throughout the entire volume of the Sun. Later in its life the Sun will go through a red giant phase, expanding outward to envelop the Earth, and the wind may again become quite strong. If the Sun undergoes a catastrophic collapse to form a white dwarf then there may be one or several episodes of impulsive mass ejection called novae. However, the Sun is a relatively small and inactive star; other stars can have quite different and often far more intense winds.

Winds from Other Stars

Stellar winds are, as indicated above, common. One means by which they are detected and analyzed is through Doppler shifts in spectral lines. Another is to infer the presence of the wind through analysis of properties of the associated star. Stellar winds found this way are all far stronger than the solar wind, but the reader should be cautioned that this is an observational selection effect. The Sun's wind would be invisible at stellar distances. If all stars were like the Sun, we would presently have no way to directly detect their winds. However, many stars are larger, hotter, denser, rotate faster, have stronger magnetic fields, are younger, or are older than the Sun and consequently have quite different kinds of winds. They fall into several categories that are in addition to winds like the solar wind that are primarily driven by a thermal pressure gradient.

Sound Wave-Driven Winds

In stars with a convection zone just below the photosphere, the convective motions can generate acoustic waves that propagate upwards through the photosphere. The waves produce a wave pressure in the atmosphere that results in an additional force working against the stellar gravity. Cool stars have convection zones of this type but the phenomenon is normally important only for very low-gravity stars. To make a massive wind requires something else in addition to sound waves because sound waves will normally dissipate low in the stellar atmosphere. The dissipation of sound waves heats the atmosphere so that there can be some crossover between thermally driven winds and sound wave-driven winds.

Dust-Driven Winds

The outer atmospheres of luminous cool giant stars and early type stars can be driven outward by radiation coming from the photosphere of the star. In the case of cool stars, dust can condense out of the atmosphere and absorb photons over a broad range of wavelengths. The radiation pressure forces the grains outward, dragging ions along by viscous drag if the atmosphere is dense, thus forming a dust-driven wind.

Alfvén Wave-Driven Winds

Alfvén waves are waves dominated by fluctuations transverse to the magnetic field direction. The restoring force is the resistance of the magnetic field to forming a kink, as opposed to the resistance of a gas to being compressed in sound waves. These waves are more important for stars with stronger magnetic fields. The dissipation of energy and momentum associated with Alfvén waves can lead to the acceleration of a wind, just as in sound wave-driven winds. The waves are formed by motions in the photosphere causing the magnetic field line to be moved. Alfvén waves have been suggested to be one source of the energy flux driving the solar wind. However, it is not yet known whether this is the dominant energy source. The dissipation of Alfvén waves will heat the atmosphere and increase the thermal pressure so that there is also some crossover between thermally driven winds and Alfvén wave-driven winds.

Radiation Pressure-Driven Winds

In these winds, atoms in the atmosphere of the star resonantly absorb radiation coming from the photosphere of the star. As might be expected, these winds exist for stars that are brighter and hotter than the Sun. Instead of 10^{-4} solar masses being lost over the life of the star, these stars can lose 10^{-6} solar masses in a single year. The flow speeds are typically $\sim 2000 \text{ km s}^{-1}$ and the density in these winds is many orders of magnitude higher than in the solar wind. The higher density means that the atmospheres of these stars are far more opaque than the solar corona. This is what enables them to absorb the radiation coming from the star. In this case the radiation pressure is the force that is working against the gravitational field of the star. The force ceases once the atmosphere becomes transparent as distance from the star increases. In red giant stars the radiation intensity is relatively weak, but the gravitational field is also weak and the stars are nevertheless observed to have radiatively driven winds. However, the strongest radiatively driven winds come from hot supergiants.

Magnetic Rotator Winds

In discussing the solar wind over the lifetime of the Sun we described how the magnetic field enhances the loss of angular momentum from the Sun by causing the ions and electrons to rotate together with the Sun as they move outward. At the same time, there is also a small outward centrifugal force, just as there is in a centrifuge. This force is completely negligible for the Sun, but one can imagine stars with stronger magnetic fields that might have centrifugally driven winds; these are called magnetic rotator winds.

The most obvious example of a magnetic rotator wind is that from a neutron star. These stars have very strong magnetic fields and centrifugal forces fill the neutron star magnetosphere with charged particles. At some distance from the star, the azimuthal velocity of the charged particles, as they are carried around the star, reaches the speed of light. The surface at this distance is the 'speed of light cylinder' and somewhere in this region around the star the particles force the field lines to open and they are released. This is how a pulsar is formed, an extreme example of a magnetic rotator.

Effects of Winds on Stellar Evolution and on the Surrounding Interstellar Medium

Winds from stars are one way in which matter that has been processed in stellar interiors reaches the interstellar medium and becomes available for new star formation; the other way is via novae and supernovae. The composition of the wind reflects, but may not be identical to, the composition of the star. Primordial material will be processed and enriched in heavy elements in this process.

The solar wind serves as an example of this process, even though the wind is relatively weak. The wind moves outward to interact with interstellar material that is always present in the galaxy. There is a contact surface that divides interstellar material from solar wind and the volume inside this surface is known as the heliosphere – that volume dominated by the Sun. The solar system is moving through the local interstellar medium at $\sim 25 \text{ km s}^{-1}$ – slow with respect to solar wind speeds – and the stand-off distance in the upstream direction is about 150 AU. Beyond this boundary lies pristine interstellar matter. In the downstream direction the solar wind flows into a heliotail that is analogous to the Earth's magnetotail and is the path by which the solar wind escapes and mixes with interstellar matter.

Acknowledgement

Portions of this work were performed at the Jet Propulsion Laboratory, California Institute of Technology, under contract with NASA.

See also: **Electricity in the Atmosphere:** Global Electrical Circuit. Magnetosphere. **Mesosphere:** Ionosphere. **Radiation Transfer in the Atmosphere:** Radiation, Solar. **Satellites and Satellite Remote Sensing:** Orbits. **Solar System/Sun, Atmospheres, Evolution of Atmospheres:** Solar Terrestrial Interactions: Climate Impact.

Further Reading

- Fleck, B., Noci, G., Poletto, G. (Eds.), 1994. Mass Supply and Flow in the Solar Corona. Kluwer, Dordrecht.
- Habbal, S.R., Esser, R., Hollweg, J.V., Isenberg, P.A. (Eds.), 1999. Solar Wind Nine, AIP Conference Proceedings 471. American Institute of Physics, New York.
- Hundhausen, A.J., 1972. Coronal Expansion and the Solar Wind. Springer-Verlag, New York.
- Kivelson, M.G., Russell, C.T., 1995. Introduction to Space Physics. Cambridge University Press, Cambridge.
- Lamers, H.J.G.L.M., Cassinelli, J.P., 1999. Introduction to Stellar Winds. Cambridge University Press, Cambridge.
- Marsden, R.G. (Ed.), 1986. The Sun and the Heliosphere in Three Dimensions. Reidel, Dordrecht.
- Parker, E.N., 1963. Interplanetary Dynamical Processes. Interscience/Wiley, New York.
- Sturrock, P.A., Holzer, T.E., Mihalas, D.M., Ulrich, R.K. (Eds.), 1980. Physics of the Sun, vols. I, II, and III. Reidel, Boston.
- Suess, S.T., Tsurutani, B.T. (Eds.), 1998. From the Sun: Auroras, Magnetic Storms, Solar Flares, Cosmic Rays. American Geophysical Union, Washington DC.
- Tsurutani, B.T., Gonzalez, W.D., Kamide, Y., Arballo, K.K. (Eds.), 1997. Magnetic Storms, Geophysical Monograph 98. American Geophysical Union, Washington DC.
- Ulmschneider, P., Priest, E.R., Rosner, R. (Eds.), 1991. Mechanisms of Chromospheric and Coronal Heating. Springer-Verlag, Berlin.
- Winterhalter, D., Gosling, J.T., Habbal, S.R., Kurth, W.S., 1996. Solar Wind Eight. In: Neugebauer, M. (Ed.), AIP Conference Proceedings 382. American Institute of Physics, New York.

Meteors

P Jenniskens, SETI Institute, Moffett Field, CA, USA

© 2003 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 3, pp 1279–1285, © 2003, Elsevier Ltd.

Nomenclature of Sources and Sinks

The Earth's upper atmosphere is continuously being bombarded by solid objects from interplanetary space. Their constituents may have provided prebiotic organic molecules for the origin of life, and even today this influx of solid particles is responsible for the continuous replenishment of a layer of metal atoms at an altitude of between 80 and 110 km. Even the smallest grains are precious momenta of the origin of the solar system and trace the dynamical processes of formation and disintegration of the first kilometer-sized planetesimals called comets and asteroids, from which they originate. The daily influx of meteoroids is the more relevant for most studies of atmospheric chemistry and physics. The much more dramatic, but rare, impacts of comets and asteroids will not be discussed here.

Meteoroids, also called interplanetary dust particles, carry much kinetic energy, about 1% of which is converted to visible light in collisions with air molecules, creating the transient phenomenon called a meteor. This energy and momentum transfer results in atmospheric chemistry and meteoroid ablation, and leaves behind a range of atomic, molecular, and solid particle products.

The following products have been identified in Earth's atmosphere. Meteor trains of enhanced densities of electrons, oxygen atoms, and meteoric metal atoms are routinely observed by radar, lidar, and other remote sensing techniques. Meteoric neutral metal atoms are the source of the sodium airglow emission and form a layer between about 80 and 110 km, called the neutral atom debris layer. These metal atoms can chemically react and condense into nanometer-sized solid particles called recondensed meteoric vapor. There is indirect evidence that nanometer-sized solid particles reside at altitudes of 70–85 km, where they act as nucleation sites for water vapor to form noctilucent clouds and are a source of weakly bound electrons. Larger solid particles, of size 10–100 μm , that survive the ablation process are called meteoric debris and have been collected in the stratosphere and at Earth's surface. Small, nanometer-sized meteoric matter with chondritic abundances has been detected as a contaminant in (or nucleation site of) larger sulfuric aerosols in the stratosphere.

If the meteoroids are sufficiently small ($<50 \mu\text{m}$) and come in slow enough ($<20 \text{ km s}^{-1}$), and if they are large enough to effectively radiate the heat ($>10 \mu\text{m}$), then they may survive the heating process almost intact. These particles are collected in the stratosphere and are called Brownlee particles, interplanetary dust particles, or micrometeorites. They are also collected in sediments on Earth's surface, where they are called exclusively micrometeorites. Micrometeorites tend to have an asteroidal origin, like the larger meteorites, because cometary material tends to be more fragile and enter Earth's atmosphere at higher speed and therefore tends to not survive the ablation process.

Annual and Daily Variation of Meteoroid Influx

The daily influx of extraterrestrial matter (Figure 1) is dominated by particles of size 100–200 μm , to the amount of $4 \times 10^7 \text{ kg yr}^{-1}$ (for sizes up to 300 μm). These meteoroids have a mean impact velocity of 25 km s^{-1} , but with a wide range from 11 to 72 km s^{-1} . Depending on the latitude of the observing site, most cometary matter in retrograde orbits and the dust particles of the zodiacal cloud that have nearly circular orbits are accreted in the morning hours, between 3.00 and 6.00 a.m. local time (the apex source), when the direction of Earth's motion is highest in the sky. For the same reason, the annual variation of meteor rates tends to peak in autumn on the Northern Hemisphere. On the other hand, the fall of meteorites and micrometeorites peaks in the afternoon hours because of the predominantly prograde orbit of the asteroidal matter and the low encounter speed needed for survival of this material (Figure 2).

The daily sporadic influx is interrupted by meteor showers, which are significant especially for relatively large meteoroids that cause visible meteors (10^{-5} g to 10^3 g). Most well-known

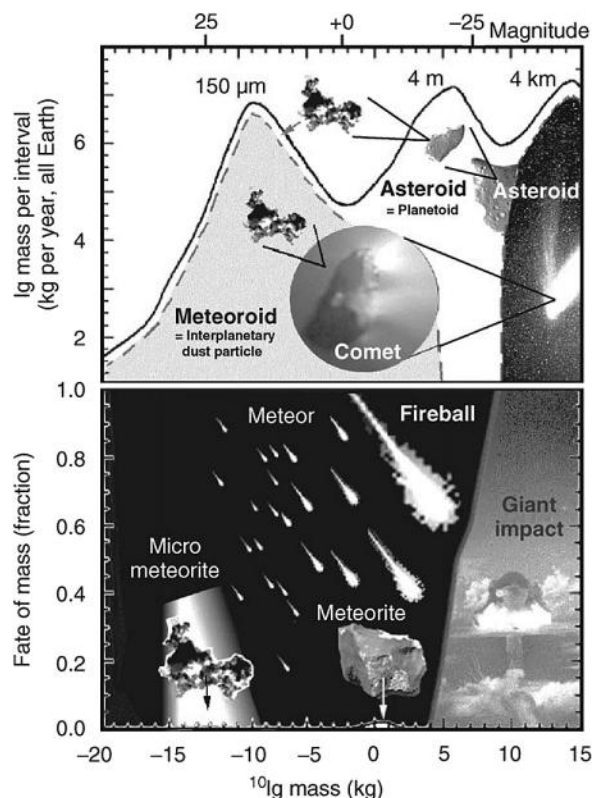


Figure 1 Overview of the influx of extraterrestrial matter in Earth's atmosphere and the nomenclature for sources, phenomena, and products. Strength of 4 m peak is uncertain.

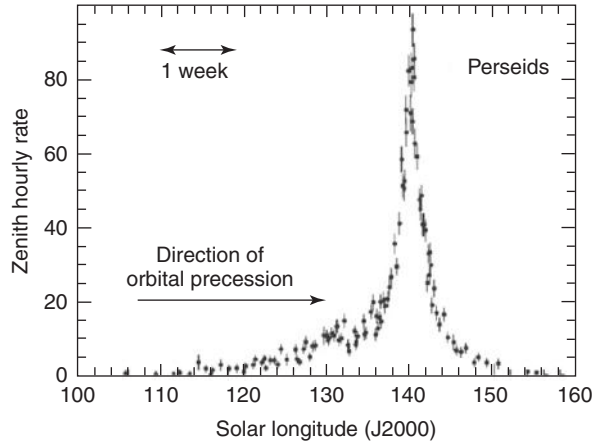


Figure 2 Typical variation of annual meteor shower activity. The asymmetry is the result of a gradually precessing orbit.

examples are the Perseid shower in summer and the Geminid and Quadrantid showers in winter (Table 1). Meteor showers are caused by dust grains released from comets, which all still approach Earth on nearly parallel orbits from a direction called the shower radiant (RA, DEC in Table 1). Unlike the sporadic background, the meteor shower size distribution is not collisionally equilibrated. Most mass is in the larger grains, which replenish the population of smaller meteoroids. The number distribution index $r = N(m+1)/N(m)$ tends to be lower than those of the sporadic background. r relates to the size distribution index: $s = 1 + 2.5 \lg r$.

During annual meteor showers, rates vary over time scales of days to weeks. The variation of rates is usually expressed in units of number influx called the zenith hourly rate (ZHR; the rate of visible meteors seen by a standard observer under ideal conditions, that is, radiant in the zenith and the star limiting magnitude = 6.5):

$$\text{ZHR} = \text{ZHR}_{\text{max}} \times 10^{-|B(\lambda_0 - \lambda_{0,\text{max}})|} \quad [1]$$

This defines the duration parameter B (degrees⁻¹). λ_0 (J2000) is the solar longitude, a measure of time and related to the Earth's position in its orbit.

Meteor outbursts represent even larger flux variations, that are observed when Earth crosses a recently formed comet dust trail. In that case, rates can increase up to 5 orders of magnitude over a period of 1–2 hours. If the meteor rate increases to above 1 per second (or $\text{ZHR} > 1000 \text{ h}^{-1}$), the event is called a meteor storm. More important than such arbitrary nomenclature are the dynamical processes that underlie these manifestations. Comet dust trails are a manifestation of ejected dust grains returning at different times from their orbit around the Sun because of slightly different orbital periods. The encounter conditions with fresh comet dust trails vary from year to year because the position of the trails relative to Earth's orbit is a function of the combined influence of planetary perturbations, which move the trails in a pattern that mirrors the Sun's reflex motion around the barycenter of the solar system. Over time, the natural waving motion of the trails fades into a broader stream called the filament, which is due to small perturbations of the orbital period of grains so that different sections of a dust trail catch up on each other. The filament

causes meteor outbursts with a duration of about a day. Such structures persist only when the grains become trapped in mean-motion resonances that prevent close encounters with the planets, because such close encounters will put the grains in the annual shower component.

Meteor Luminosity

When a meteoroid encounters the Earth's atmosphere, an initial phase of V-shaped luminosity is seen, possibly caused by the interaction of released electrons with the ambient ionosphere. Leonids 1 kg in size have been detected as high as 196 km.

Most meteors are not seen until around 135–120 km, where massive evaporation of meteoric silicates starts to occur. The process is one of sputtering, where each impact of an air molecule releases a cloud of atoms and molecules from the meteoroid. That builds a small vapor cloud traveling along with the meteoroid. Subsequent air collisions are with the vapor cloud, creating a warm ($T \approx 4400 \text{ K}$) plasma in the immediate wake of the meteoroid.

When the pressure of the vapor cloud exceeds that of the ambient air, it expands as into a vacuum. The density falls off rapidly with increasing distance from the meteoroid, while the temperature increases. This region may be the source of an emitting gas ($T \approx 10\,000 \text{ K}$) detected in bright fast meteors. The size of the vapor cloud is a function of particle mass and speed, and about 1–10 orders of magnitude larger than the size of the meteoroid (Figure 3).

The wake of the meteoroid is formed by the impacting air molecules and by the vapor cloud products that receive any amount of momentum transfer, that is after a single collision. This is the main source of light emission at visible wavelengths and reflections. Head echoes are the signature of a hard (unresolved) target crossing the radar beam. The decelerated atoms will expand into the ambient air until stopped by subsequent collisions. The mean free path determines the initial radius (r_i) of the wake, which is of the order of a few meters and has the following expected dependence on velocity and atmospheric mass density (ρ_a):

$$r_i \approx \rho_a^{-1} V^{0.8} \quad [2]$$

Observed dependencies are less steep ($r_i \approx \rho_a^{0.4-0.8}$), perhaps due to fragmentation. At high elevations, the attenuation due to this dispersion leads to low electron densities and an 'echo height ceiling' effect in backscatter radar. For a radar at any given wavelength there is a height beyond which no underdense echoes can be seen.

'Overdense echoes' occur when the electron density remains high enough to cause mirror-like reflection. Sometimes, there are also 'head echo' reflections, which are the signature of a hard (unresolved) target crossing the radar beam: a spherical region of ionization that travels along with the meteoroid and may be related to the spherical region of optical luminosity in the bottom part of Figure 3.

The observed light intensity (I) and electron line density (q) are proportional to the rate of loss of kinetic energy from the ablated atoms:

$$I = 0.5\tau \frac{dm}{dt} V^2 \quad [3]$$

Table 1 Meteor showers

Shower	$\lambda_0^{\max}$	$ZHR_{\max}$	B	RA, DEC		V_{inf}	r^b
Quadrantids	283.32	130	2.5230	+49	43	2.2	
γ -Velids	285.7	2.4	0.12	124	-47	35	3.0
ζ -Aurigids	285.8	(5)	0.20	077	+58	16	—
Corona-Borealis	291.1	(15)	0.20	230	+37	37	—
δ -Cancrids	298.9	(11)	0.20	132	+20	28	—
α -Leonids	309.0	(7)	0.10	157	+06	29	—
α -Carinids	311.2	2.3	0.16	091	-54	25	2.5
α -Centaurids ^a	319.4	7.3	0.18	209	-58	57	2.3
ω -Centaurids	323.4	2.2	0.15	175	-57	51	2.8
δ -Leonids	334.7	1.1	0.049	154	+19	23	3.0
π -Virginids	343.9	(2)	0.10	184	+01	32	3.0
S-Virginids	359.7	(4)	0.05	196	-01	28	3.0
N-Virginids	013.7	(3)	0.05	210	-08	30	3.0
γ -Normid	353.0	5.8	0.19	249	-51	56	2.4
δ -Pavonids	011.1	5.3	0.075	309	-63	60	2.6
April Lyrids ^a	032.4	12.8	0.22	272	+33	49	2.7
μ -Virginids	039.7	2.2	0.045	229	-06	30	3.0
η -Aquarids	046.5	36.7	0.080	339	-01	66	2.7
α -Scorpiids	055.9	3.2	0.13	251	-23	35	2.5
May Arietids	054.2	(7)	0.10	036	+18	27	—
σ -Cetids	064.7	(7)	0.08	028	+01	35	—
σ -Scorpiids	072.6	5.2	0.15	240	-20	21	3.0
Daytime Arietids	075.9	54	0.10	038	+24	38	2.7
ζ -Perseids	078.8	(17)	0.05	070	+27	27	—
γ -Sagittarids	089.2	2.4	0.037	285	-27	29	2.9
τ -Cetiids	095.7	3.6	0.18	096	+23	66	2.5
β -Taurids	096.7	(20)	0.08	086	+19	30	—
θ -Ophiuchids	097.7	(2.3)	(0.037)	251	-15	27	2.8
τ -Aquarids	098.0	7.1	0.24	342	-12	63	2.5
ν -Phoenicids	111.2	5.0	0.25	026	-41	48	3.0
σ -Cygnids	116.7	2.5	0.13	305	+47	37	2.7
α -Capricornids	122.4	2.2	0.041	301	-10	25	2.0
δ -Aquadrids N	124.1	1.0	0.063	324	-12	42	3.3
Pisc. Australids	124.4	(2.9)	(0.26)	337	-33	42	3.2
δ -Aquarids S	125.6	11.4	0.091	339	-17	43	3.3
ι -Aquaris Z	131.7	1.5	0.070	334	-15	36	3.3
Perseids ^a	140.19	84	0.20	046	+58	61	2.5
κ -Cygnids	146.7	2.3	0.069	290	+52	27	2.2
γ -Dorids	155.7	4.8	0.18	060	-50	41	2.8
Aurigids	158.2	(9)	0.19	072	+43	69	2.7
κ -Aquarids	177.2	2.7	0.11	338	-05	19	2.8
Sextantids	186.7	(9)	0.08	154	+00	32	—
Draconids ^a	197.2	2	0.20	262	+54	20	3.0
S-Arietids	201.7	(3)	0.03	035	+10	28	—
ε -Geminids	206.7	2.9	0.082	103	+28	71	3.0
Orionids ^a	208.6	25	0.12	095	+16	67	3.1
Leo-Minorids	209.7	1.9	0.14	159	+38	61	2.7
Taurids	223.6	7.3	0.026	049	+18	30	2.3
ζ -Puppids	232.2	3.2	0.13	117	-42	41	3.4
Leonids ^a	235.6	13	0.20	153	+22	71	3.0
Puppids/Velids	251.7	4.5	0.034	128	-42	40	2.9
Phoenicids	252.4	2.8	0.30	018	-58	18	2.8
Monocerotids	260.9	2.0	0.25	100	+14	43	3.5
Geminids	262.1	88	0.39	113	+32	36	2.6
σ -Hydrudids	265.5	2.5	0.10	133	-01	59	3.0
Ursids ^a	271.0	4.0	0.20	220	+75	35	3.0

^aOccasional outbursts.^b3.4 for sporadic meteors.

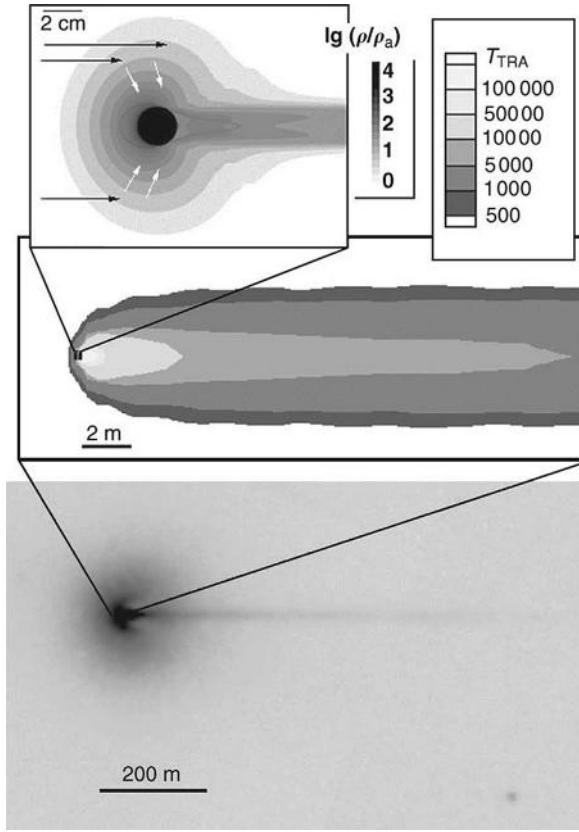


Figure 3 Meteor anatomy. Meteor vapor cloud in a model by Olga Popova, whereby gas densities are shown. Meteor wake in a model by Iain Boyd, showing translational temperatures. And a high frame-rate still image of a -3 magn. Leonid meteor by Hans Stenbaek-Nielsen.

with τ the dimensionless luminous-efficiency factor $\tau \approx 0.01$, depending on the particulars of the meteor emission spectrum (changing with speed, size, etc.) and the instrument response.

The blue and green part of the optical spectrum consists mainly of ablated metal atom emission lines with a $T \approx 4400$ K, mostly from iron and magnesium lines. Fast and bright meteors show the violet Ca^+ doublet. The orange and red part of the spectrum are dominated by neutral oxygen and nitrogen lines and by the first positive bands of molecular nitrogen. Those atmospheric lines and bands are also well described by $T \approx 4000$ K vibrational temperature, electronic excitation temperature, and chemical equilibrium. Many deviations from local thermodynamic equilibrium are observed, however, and are the focus of study. Emissions from small organic molecules such as CN have been reported, but with insufficient certainty to derive quantitative information. The thermal infrared spectrum remains unexplored, while the near-UV spectrum is dominated by lines of Mg and Mg^+ at 280 and 285 nm (Figure 4).

For the typical response of the photographic plates used in the Harvard Super Schmidt program (m_{ph}), the brightness as a function of mass is given by

$$\lg M(g) = 5.15 - 0.44 * m_{\text{ph}} - 3.89 \lg_{10} V_{\text{inf}} (\text{km s}^{-1}) - 0.67 \lg_{10} [\sin(h_r)] \quad [4]$$

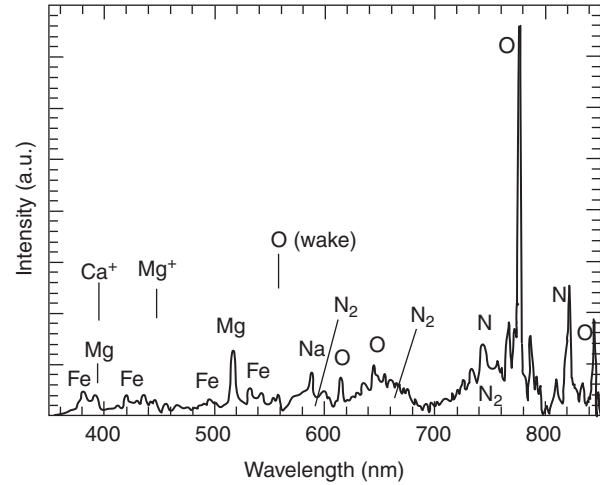


Figure 4 Optical spectrum of a Leonid meteor from data by Shinsuke Abe.

where h_r is the angle of incidence. The magnitude m_{ph} refers to an ‘absolute brightness’ at a distance of 100 km. On this scale, a zero-magnitude Leonid would weigh 0.07 g, and would emit visible light (400–700 nm) at a rate of about 500 W.

The relationship between visual meteor magnitude (m_v) and the number of electrons formed per meter of meteor path (the electron line density) is

$$m_v = 38 - 2.5 \lg_{10} q(\text{m}^{-1}) + 2.5 \lg V(\text{km s}^{-1}) \quad [5]$$

This density pertains to the wake shortly after it has formed. The density decays rapidly by ambipolar diffusion, with a coefficient $D \approx 4.2 \text{ m}^2 \text{ s}^{-1}$ at 93 km.

Altitude of Deposition

How is the ablated matter (and amount of ionization/chemistry) distributed as a function of altitude? The traditional approach is to assume the meteoroid mass loss to be proportional to the kinetic energy transferred to the intercepted air mass:

$$dm/dt = \frac{-\Lambda A}{2\zeta} \left(\frac{m}{\rho}\right)^{2/3} \rho_a V^3 \quad [6]$$

where ζ is the heat of ablation of the meteoroid material ($0.2\text{--}1.0 \times 10^3 \text{ Jg}^{-1}$); Λ is the dimensionless heat transfer coefficient that specifies the efficiency of the collision process in converting kinetic energy to heat (0.6, down to 0.1), A the effective surface area, and ρ the density of the meteoroid ($0.1\text{--}1.5 \text{ g cm}^{-3}$).

From this, the classical expression for the rise and fall of meteoric luminosity (or ionization) is derived. For given atmosphere density profile $\rho_a(H)$:

$$\frac{I}{I_{\text{max}}} = 9/4 * \frac{\rho_a}{\rho_a^{\text{max}}} * \left(1 - \frac{\rho_a}{3\rho_a^{\text{max}}}\right)^2 \quad [7]$$

where max denotes the altitude (H) of the peak brightness, which varies with entry velocity, zenith angle, and meteor brightness:

$$H_{\max}(\text{km}) = 21.2 + 44 \lg_{10} V(\text{km s}^{-1}) + 1.1M_{\max} \quad [8]$$

(and less for higher than typical meteoroid density).

Lightcurves from cometary meteoroids tend to be flatter than this and are better described by a sum of such classical lightcurves, the assumption being that a matrix material is evaporated or bonding sufficiently weakened for the meteoroid to break apart. Because there is not much momentum transfer to the grains, these fragments move along as a cloud of solid bodies. Because the larger fragments penetrate deeper than do the smaller ones, the size distribution determines the shape of the composite lightcurve. In the calculations, it is important to take into account the higher deceleration of the smaller fragments, with the resistance or drag equation:

$$\frac{dV}{dt} = -\left(\frac{\Gamma A}{m^{1/3} \rho^{2/3}}\right) \rho_a V^2 \quad [9]$$

with Γ ($= 0.5$ – 1.0) the drag coefficient. Evidence for such fragmentation comes from the early release of sodium relative to magnesium in optical spectroscopy, the detection of jets of debris fragments ejected by rapidly spinning meteoroids, and larger-than-expected emission volumes, among others.

The peak mass loss is also a function of the sublimation temperature of the mineral components of the meteoroid. Differential ablation is the term used to indicate the deposition of matter in successive order of volatility. The most volatile organic compounds with $T \approx 500$ K have their peak loss as high as 117 km in a fast 71 km s^{-1} Leonid, while the most refractory compounds with $T = 2400$ K are expected to have peak mass loss at 105 km altitude. For slower $T = 25 \text{ km s}^{-1}$ meteors, the organics are not lost until 90 km altitude in such models (Figure 5).

Evidence for differential ablation is much less abundant than expected. In general, optical emission spectra do not show evidence for differential ablation. Only in unusually fragile meteoroids do we see an earlier onset of the sodium than magnesium atoms, thought to be because the minerals involved are more efficiently exposed to the meteoric vapor plasma. Lidar observations of neutral atom debris trails, however, tend to

show only one of several metal atoms at a given altitude, which has been interpreted as evidence for differential ablation. Such trails are detected only minutes to hours after deposition, when they drift by the lidar beam (Figure 6).

Conditions in the Meteor Path after Deposition

Once the air is heated and meteoric matter has been deposited, the resulting pressure increase by more than a factor of 10 creates a shock wave which expands radially. Assuming that this expansion occurs adiabatically, then the pressure will equilibrate with the background atmosphere when the radius is about 70 m (example for a -12 -magnitude Leonid). Even at the very low pressures of the upper mesosphere/lower thermosphere ($<10^{-8}$ hPa), the size and velocity (Mach 270) of such a meteoroid would create a turbulent wake (Reynolds number >2000).

At this stage, the meteor is detected in several manners. First, a wake of the forbidden ‘green line’ emission at 557 nm from the O (1S – 1D) transition is seen in fast meteors, even relatively faint ones. The emission is thought to be produced by the Barth mechanism. Intensity estimates show that about 15% of the O_2 in the initial train is dissociated in a -12 magnitude Leonid.

Radar reflections have detected an extended wake out to 6 km and several seconds behind a meteoroid that contributes to nonspecular reflections at UHF and VHF frequencies. Plasma instabilities and turbulence are responsible for an anomalous cross-field diffusion of meteor trails in the Earth’s magnetic field that can be up to an order of magnitude faster than the rate expected from ambipolar diffusion. In bright fireballs, this wake is accompanied by an afterglow rich in metal atom line emissions, without the atmospheric emission lines that are typical for the meteor spectrum. The afterglow has been interpreted as due to secondary ablation from debris particles. Evidently, even a fast meteor can deposit solid debris in its path, if conditions

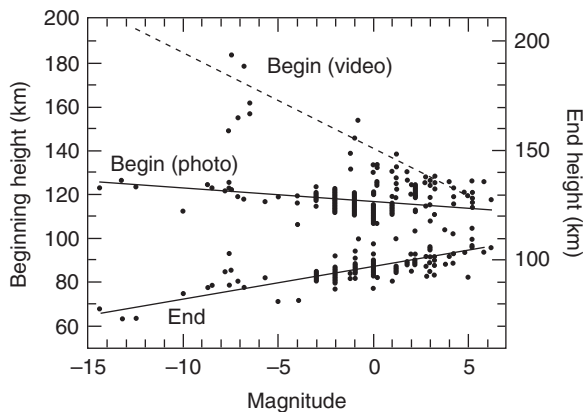


Figure 5 Beginning and end heights of Leonid meteors.

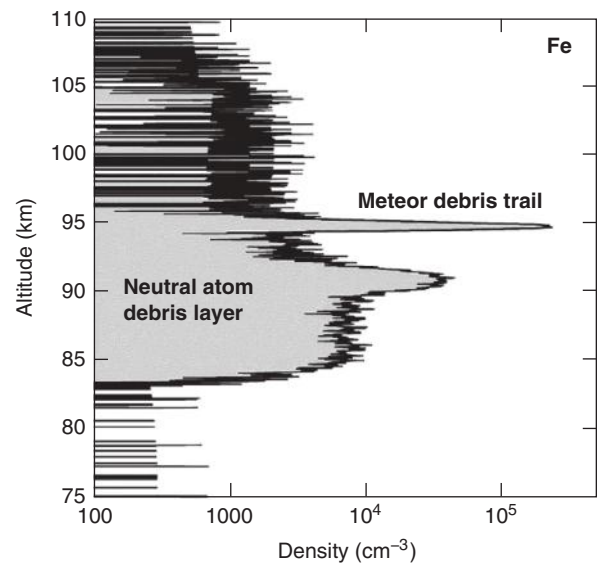
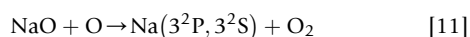
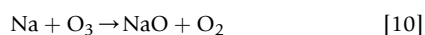


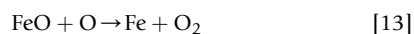
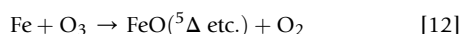
Figure 6 Lidar response to a neutral iron atom debris layer, measured by University of Illinois Fe boltzmann lidar.

are favorable. The ablation vapor cools from ~ 4400 K to ~ 1200 K in a few seconds. This afterglow is followed by a recombination emission phase that lasts tens of seconds. Mid-infrared emission remains detected for longer, even when the gas and dust has cooled to just 50 K over the ambient $T \approx 250$ K a few minutes after the fireball.

A persistent chemiluminescence remains for Leonids < -3 magnitude, which is called the persistent train (Figure 7). Those persistent trains can be visible at ~ 13 magn/arcsec² for many (tens of) minutes. The dynamical mechanism is not yet fully understood. Typically, two bright bands of light are seen with various levels of billowing. The train spectrum shows sodium emission lines and a broad molecular band continuum identified as the FeO orange arc band. The luminous mechanism is thought to be the catalytic recombination of ambient ozone with oxygen atoms in the trail through the Chapman airglow mechanism:



where the branching ratio of reaction [11] to produce the Na (3^2P) state (which then emits an orange photon at 589 nm) is about 10%. The FeO molecular emission band probably arises from



where reaction [12] is sufficiently exothermic to produce FeO in excited electronic states, leading to emission between 570 and 630 nm with about a 2% efficiency. Thus persistent trains serve as a model for natural airglow emission, in extreme conditions and with the reactive components separated. The trains also probe upper atmosphere winds, wind shear, and diffusion rates.

On the relatively short time scale of the train (minutes rather than hours), and particularly in the presence of elevated concentrations of atomic oxygen, it is very unlikely that the metallic species would be able to form more stable reservoir compounds such as NaHCO_3 or $\text{Fe}(\text{OH})_2$. Indeed, between 85 and 100 km the meteoric metals are overwhelmingly in the

atomic form in the background atmosphere. The postulated formation of nanometer-sized recondensed meteoric vapor particles in the warm meteor wake has not yet been demonstrated. However, when the metal atoms settle to lower elevations, they quickly react chemically and can condense to form particles. Most of this fine-grained material is expected to be transported to one of the poles, following seasonal winds in the upper mesosphere.

Impact of Meteors on the Atmosphere

Much remains unknown about the chemistry in the meteor wake plasma. Molecular abundances for equilibrium air plasmas show that the $T \approx 4400$ K temperature enables interesting organic chemistry in CO_2 -rich atmospheres, because not only CO_2 but also CO is dissociated. At slightly lower temperatures, much of the electron charge is balanced by NO^+ , while at higher temperatures much of the charge is carried by C^+ or O^+ . NO production rates may be affected by nonequilibrium processes and have not yet been calibrated by observations. The chemical changes of organic compounds in meteors are of particular interest, because meteors may have contributed more than two-thirds of all infalling organic matter on the early Earth, in a form determined uniquely by the chemistry in the meteor wake and subsequent terrestrial evolution.

Meteoric metals and ions contribute to a steady-state condition in the upper atmosphere. The influence of meteors may be measured directly during unusual disturbances of meteoric influx. Indeed, variations of OH airglow intensity with meteor shower rates have been reported in conjunction with the 1999 Leonid meteor shower. Tentative associations have also been made between meteor ionization and the occurrence of sprites and elves during lightning storms. Lightning has been observed to travel along the ionized path of a meteor. Much progress in these fields is expected in the near future.

See also: **Aerosols:** Aerosol–Cloud Interactions and Their Radiative Forcing. **Climate and Climate Change:** Carbon Dioxide. **Clouds and Fog:** Noctilucent Clouds. **Electricity in the Atmosphere:** Lightning.

Further Reading

- Bronshten, V.A., 1983. *Physics of Meteoric Phenomena*. Reidel, Dordrecht.
- Cepelcha, Z., Borovicka, J., Elford, W.G., et al., 1998. Meteor phenomena and bodies. *Space Science Reviews* 84, 327–471.
- Jenniskens, P., 2001. Meteors as a vehicle for the delivery of organic matter to the early Earth. ESA-SP 495. In: Warmbein, B. (Ed.), *Proceedings of the Meteoroids 2001 Conference*, Swedish Institute of Space Physics, Kiruna, Sweden, 6–10 August 2001. ESA, ESTEC, Noordwijk, pp. 247–254.
- Levin (Lewin), B.Y., 1961. *Physikalische Theorie der Meteore und die meteoritische Substanz im Sonnensystem*, vol. II. Scientia Astronomica 4. Akademie-Verlag, Berlin.
- McKinley, D.W.R., 1961. *Meteor Science and Engineering*. McGraw-Hill, New York.
- Rietmeijer, F.J.M., 2002. Interrelationships among meteoric metals, meteors, interplanetary dust, micrometeorites, and meteorites. *Meteoritics and Planetary Science* 35, 1025–1041.
- Warmbein, B. (Ed.), 2001. *Proceedings of the Meteoroids 2001 Conference*. ESA-SP 495. Swedish Institute of Space Physics, Kiruna, Sweden 6–10 August 2001. ESA, ESTEC, Noordwijk.

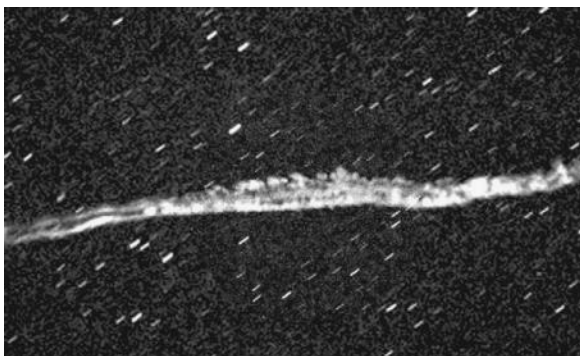


Figure 7 Persistent train of a Leonid fireball. Photo © R. Haas, Dutch Meteor Society.

STATISTICAL METHODS

Contents

Data Analysis: Empirical Orthogonal Functions and Singular Vectors

Data Analysis: Time Series Analysis

Data Analysis: Empirical Orthogonal Functions and Singular Vectors

CS Bretherton, University of Washington, Seattle, WA, USA

© 2003 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 2, pp 617–620, © 2003, Elsevier Ltd.

Introduction

One frequently needs to analyze atmospheric observations or model outputs that are time-varying fields of data at many points in space. Typically, one desires to isolate a few important and physically interpretable modes of variability in the data. Until the 1960s, this was usually done subjectively, by constructing an index (such as the Southern Oscillation Index of sea-level pressure difference between Tahiti and Darwin) that appears to correlate strongly with the variability in one or more data fields (sea-level pressure, sea surface temperature, winds, precipitation) at many locations. Regressing this index onto the data or binning the data corresponding to different ranges of the index yields composite spatial patterns of variability.

To avoid the trial and error of concocting such indices, it has become common to use techniques that directly find dominant modes of variability in the data. This article describes some popular techniques for doing this.

We first discuss principal component analysis (PCA) and some variants. PCA seeks spatial patterns that explain the maximal variance in a single data field. We then examine popular methods for isolating important spatial patterns of coupled variability between several data fields. Such methods include maximum covariance analysis or MCA (also widely called singular value decomposition or SVD – we prefer the name MCA, suggested by von Storch and Zwiers in 1999, since SVD is the name of a fundamental matrix decomposition whose applications are far broader than this specific technique of spatial data analysis); canonical correlation analysis (CCA); redundancy analysis; and variations. We discuss space–time data fitting with a linear autoregressive model, called principal oscillation pattern (POP) analysis. All of the above methods are thoroughly discussed and referenced by von Storch and Zwiers (see [Further Reading](#)). Lastly, we briefly mention some nonlinear techniques for space–time analysis.

Principal Component Analysis

PCA was pioneered in the social/biological sciences by Hotelling and Pearson in the 1930s, and first applied to atmospheric data by Lorenz in 1956. Consider a multivariate data set $y_i(t_n)$ consisting of several samples (indexed by n) of a set of variables (indexed by i). In a biological science application, the variables might be the sizes of different body parts (forearm length, waistline, height, foot size, etc.), sampled across many people. In an atmospheric context, the variables might be annually-averaged sea-level pressure anomaly $y_i(t_n)$ at I observing stations $i = 1, \dots, I$, sampled for $N = 50$ years t_n , $n = 1, \dots, N$. PCA seeks a single linear combination of the variables that explains the maximal fraction of the overall sample variance.

Mathematical Derivation

This can be phrased as the following mathematical optimization problem. Let $y(t_n)$ be the vector of observations at time t_n . Any linear combination of these observations can be written $a(t_n) = \mathbf{u} \cdot \mathbf{y}(t_n)$, where $\mathbf{u}$ is a vector of station weights u_i . It suffices to take $\mathbf{u}$ to be a vector of unit norm. Then $a(t_n)$ is the projection of the data onto the direction $\mathbf{u}$, and $\overline{a^2(t)}$ is the explained variance (eqn [1]).

$$\overline{a^2(t)} = \mathbf{u}^T \overline{\mathbf{y} \mathbf{y}^T} \mathbf{u} = \mathbf{u}^T \mathbf{C}_{yy} \mathbf{u} \quad [1]$$

Here, an overbar represents an average over the N sampling times, and the (i, j) th element of the matrix $\mathbf{C}_{yy}$ is the covariance of y_i with y_j . In this language, PCA seeks a $\mathbf{u}$ that maximizes eqn [1] subject to $|\mathbf{u}| = 1$.

Since $\mathbf{C}_{yy}$ is a symmetric positive definite matrix, standard linear algebra tells us that the desired $\mathbf{u} = \mathbf{e}_1$, the eigenvector of $\mathbf{C}_{yy}$ corresponding to its largest eigenvalue λ_1 . If the other eigenvalues are ordered from largest to smallest, the k th eigenvector is the direction that explains the maximal variance

in the data and is orthogonal to all previous eigenvectors. This eigenvector is called the k th empirical orthogonal function (EOF) of y and the corresponding time series $a_k(t)$ of the projection of the data onto this EOF is called the k th principal component (PC). The data field can be exactly rewritten in 'canonical form' in terms of its EOFs and PCs (eqn [2]).

$$y(t_n) = \sum_{k=1}^I a_k(t_n) \mathbf{e}_k \quad [2]$$

Furthermore, the $a_k(t_n)$ are temporally uncorrelated, so the total variance in y summed over all spatial locations is just the sum of the eigenvalues λ_k , and the fraction of the variance explained by the k th PC/EOF pair is given by eqn [3].

$$f_k = \frac{\lambda_k}{\sum_{l=1}^I \lambda_l} \quad [3]$$

Owing to these orthogonality properties, the EOFs can be interpreted as spatial patterns of variability as well as just weight functions. In particular, if $y_i(t)$ is regressed on the k th PC, the regression slope is just e_{ik} , the i th element of the k th EOF. Thus $\lambda_k^{1/2} \mathbf{e}_k$ can be interpreted as the characteristic change of $y(t)$ corresponding to a one standard deviation ($\lambda_k^{1/2}$) change in the k th PC.

Practical Interpretation of PCA

In practice, only a few leading EOFs explain most of the variance of the field. Since most atmospheric fields have long-range spatial correlations, these EOFs typically correspond to patterns of broad spatial scale, and are relatively robust even when the number of samples is low or the data are noisy. Thus, the leading PCs are often useful index time-series for compositing with respect to the leading modes of variability of a given field. Often, 1–5 PCs explain as much as 90% of the overall variance in the field, and these PC/EOF pairs form an efficient low-dimensional description of the entire field. In this case, projection onto these PCs is a useful space/time filter before other statistical methods such as canonical correlation analysis are employed.

If the variability of the field is dominated by propagating disturbances (e.g., mid-latitude sea-level pressure variations on synoptic time scales), the corresponding eigenvalues come in a nearly equal pair corresponding to two phases of the disturbances that are in quadrature.

Because higher EOFs are forced to be orthogonal to their predecessors, they may not be amenable to simple physical interpretation. Furthermore, since the covariance matrix was estimated on the basis of a finite number of physical realizations, its EOFs may contain sampling biases. This problem becomes more acute for higher-index EOFs, and can also mix EOFs with closely spaced eigenvalues. North's criterion states that if the covariance matrix is constructed on the basis of N independent samples, its eigenvalues have sampling uncertainties $\delta\lambda_k \approx 2\lambda_k/N$, and if the spacing between successive eigenvalues is comparable to or less than this sampling uncertainty, their EOFs will be heavily contaminated by sampling uncertainty. Another simple check on the robustness of the PCA is to divide the data set into randomly chosen

halves, then examine the similarity of the EOFs given by the two halves.

Rotated PCA

The spatial patterns produced by PCA can sometimes be difficult to interpret. For instance, two ocean basins might have characteristic patterns of sea surface temperature anomalies that explain a lot of the intrabasin variability, but the variability in the two basins might only be weakly coupled. To maximize the leading mode variance and maintain an orthogonal second mode, PCA would yield a leading mode in which the spatial patterns in the two basins are in-phase and a second mode of slightly smaller explained variance in which they are out of phase. This might direct attention to the weak coupling between the basins rather than the much more robust intrabasin variability.

Rotated PCA (in statistical parlance, a type of factor analysis) can be used to obtain localized modes that might have a simpler physical interpretation. It consists of choosing a set of spatial weights or directions that optimize some measure of localization. The data are typically filtered onto the leading EOFs before this procedure is carried out. Varimax rotation, which is most commonly used, maximizes a function of the weights that tends to force them to be closer to either zero or one than regular PCA. This favors modes that are weighted mainly to a few spatial points, inhibiting mode mixing of the kind discussed above. However, the modes are no longer temporally uncorrelated (hence they do not form a partitioning of the variance) and may not be spatially orthogonal for some rotation strategies. Furthermore, the leading mode will not explain so much of the variance as with regular PCA, and many spatial patterns do not have the localized character that rotated PCA selects for, so this method should be applied thoughtfully.

PCA for Propagating Disturbances

Several methods for extending PCA to better analyze spatially propagating disturbances have been proposed. One such technique is complex PCA (CPCA), in which a discrete Hilbert transform in time is applied to the field of data. A new field is created whose real and imaginary parts are the original field and its Hilbert transform. A PCA of the (now complex) covariance matrix of this field yields complex eigenvectors whose real and imaginary parts give the characteristic spatial patterns in both phases of the wave.

Singular Spectral Analysis (SSA)

Many modes of spatial variability have characteristic time evolutions, for instance, the life cycle of a baroclinic cycle. Singular spectral analysis is a method popular in some circles for characterizing the dominant patterns of space–time variability, and has been used as an alternative to power spectral analysis. Conventional PCA is based on the matrix of simultaneous (unlagged) covariances. SSA is based on PCA of an extended covariance matrix that also includes lagged covariances with a sequence of time lags. The EOFs are parsed into characteristic spatial patterns evolving through the sequence of lag times.

Methods for Isolating Patterns of Coupled Variability

Two linear methods, MCA and CCA, are widely used for isolating patterns of coupled variability in two fields. A third method, called 'redundancy analysis' (RA) by von Storch and Zwiers, is also useful for predicting the time-space structure of one field based on observations of another, correlated field.

Mathematically, these methods can be phrased as follows. We call the two fields the 'left' field $y_i(t_n)$ and the 'right' field $z_j(t_n)$. For example, we might look for coupling between monthly anomalies of tropical Pacific sea surface temperature at an array of buoys (left field) with monthly gridded Northern Hemisphere 500 hPa height anomalies (right field). The two fields must be given at the same time, but not necessarily at the same spatial locations.

Maximum Covariance Analysis (MCA)

MCA (also commonly known as SVD) was first applied to an atmospheric problem in 1976 by Prohaska, and was comprehensively compared with other methods of space-time analysis by Bretherton and colleagues. MCA finds an optimally coupled left spatial pattern $\mathbf{u}$ and a right spatial pattern $\mathbf{v}$, both unit vectors. They are chosen such that the projection $a_1(t) = \mathbf{u} \cdot \mathbf{y}(t)$ of the left-field values onto the $\mathbf{u}$ direction has maximal possible covariance with the corresponding projection $b_1(t) = \mathbf{v} \cdot \mathbf{z}(t)$ of the right-field values in the $\mathbf{v}$ direction. The solution is obtained by singular-value decomposition of the cross-covariance matrix C_{yz} between the gridpoint values of the left field and those of the right field. If its left and right singular vectors are $\mathbf{l}_k$ and $\mathbf{r}_k$, ordered from largest to smallest singular value σ_k , then the optimal $\mathbf{u}$ and $\mathbf{v}$ are the leading left and right singular vectors $\mathbf{l}_1$ and $\mathbf{r}_1$. Succeeding singular vectors define coupled spatial modes that maximize the covariance subject to orthogonality with the previous modes. Unlike with PCA, the corresponding modal amplitudes may be temporally correlated.

MCA is efficient to numerically implement, requires no prefiltering of the fields, and produces spatial patterns that exhibit broad coupling over as much as possible of the spatial domains being analyzed. Its optimality characteristics usually provide excellent results, but there are caveats. One is that even if two fields are perfectly linearly coupled (e.g. a two-dimensional streamfunction and corresponding vorticity field), MCA may only approximately identify this coupling. A second caveat is that as with PCA, the spatial orthogonality constraints on succeeding spatial patterns may mix physically distinct but spatially nonorthogonal modes of variability.

Canonical Correlation Analysis

CCA, first developed by Hotelling in 1936, seeks a linear combination $a(t) = \mathbf{u} \cdot \mathbf{y}(t)$ of the left-field values and a linear combination $b(t) = \mathbf{v} \cdot \mathbf{z}(t)$ of the right-field values that have maximal temporal correlation. This optimality criterion is similar to MCA (which maximizes covariance). In practice, the difference is that CCA can choose linear combinations that explain little of the variance in their field

but are highly correlated with the other field. MCA will select linear combinations that explain significant variance in their fields but may be slightly less well correlated with the other field.

The optimal solution can be found by first performing PCA on each field. This yields standardized left and right principal components $\alpha_k(t)$ and $\beta_l(t)$. Then a singular value decomposition is performed on the cross-covariance matrix between $\alpha_k(t)$ and $\beta_l(t)$. The left/right singular vectors are transformed back into gridpoint coordinates to get the weight vectors $\mathbf{u}_k/\mathbf{v}_k$. The singular values are the correlations between the corresponding time-series $a(t)$ and $b(t)$.

To obtain robust results with CCA if a limited number of times have been sampled, one should first prefilter the data onto a small set of leading principal components of each field (that collectively explain 70–90% of the variance) to avoid spurious large correlations associated with overfitting. In this case, CCA tends to give quite similar results to MCA for the leading pattern. Successive patterns are temporally uncorrelated but may be spatially nonorthogonal. CCA is slightly more numerically intensive than MCA, and the prefiltering step slightly complicates the analysis procedure. However, it can sometimes be preferable if strong coupling between the fields is geographically localized.

Redundancy Analysis

In redundancy analysis, an application of multiple regression first developed in the 1970s, one tries to explain the maximal fraction of the variance of the right (predictand) field using the left (predictor) field. This can be viewed as a collection of least-squares multiple-regression problems for the right-field values at each location. As with CCA, prefiltering is recommended to avoid overfitting of the data.

Aggregated PCA

The above methods look for coupling between two scalar fields. Coupling between more than two fields, or a vector field and a scalar field, can be handled by aggregating some of the fields after standardizing them to comparable variance, and analyzing coupling between two aggregated fields. Another approach in this case is to aggregate all of the fields into a single field, whose modes of variability are found using PCA. The resulting modes optimize a combination of the explained variance within individual fields and the covariance between fields, and may not pick out the patterns of coupling between the fields as effectively as the previous methods.

Principal Oscillation Pattern (POP) Analysis

Some physical systems can profitably be idealized as a damped linear response to stochastic forcing. POP analysis fits such a model to the space-time variability of a field or coupled fields. Given a vector $\mathbf{y}_n$ of observations at times t_n of one or more fields at one or more locations, the fitting model has the form of eqn [4].

$$\mathbf{y}_{n+1} = \mathbf{A}\mathbf{y}_n + \mathbf{f}_n \quad [4]$$

The stochastic forcing $\mathbf{f}_n$ is a noise vector uncorrelated between observation times, and $\mathbf{A}$ describes the evolution of the unforced system. The eigenvalues of $\mathbf{A}$ should all have magnitudes less than one, so that this model can describe a statistically stationary process. The corresponding eigenvectors of $\mathbf{A}$ (the POPs) describe the characteristic unforced patterns of spatial variability of the system; those with eigenvalues near to one in magnitude are slowly damped and will tend to dominate the observed response. Complex eigenvalues correspond to damped oscillations; the corresponding POPs are also complex, and the real and imaginary parts of the POP describe the oscillations in the spatial structure of the response.

To estimate $\mathbf{A}$ from the data, consider the expected value $E[\cdot]$ of an outer product of eqn [4] with $\mathbf{y}_n$ (eqn [5]).

$$E[\mathbf{y}_n \mathbf{y}_{n+1}^T] = \mathbf{A} E[\mathbf{y}_n \mathbf{y}_n^T] \quad [5]$$

From eqn [5], $\mathbf{A}$ can be estimated based on the empirical lag-0 and lag-1 covariance matrices.

Nonlinear Methods

The above 'linear' analysis techniques decompose an observed field into the sum of modes, each of which is the product of a spatial pattern and a time-varying amplitude. Some

phenomena, such as ENSO sea surface temperature variability, show a somewhat different spatial anomaly pattern in the positive phase from that in the negative phase, and may be more efficiently described as nonlinear modes whose pattern covaries with their amplitude. One standard check for this behavior is to composite the spatial anomalies corresponding to positive and negative values of an index time-series and see whether they are substantially different in structure. Neural net techniques have recently been developed that may prove to be useful for isolating such nonlinear modes of variability.

See also: Climate and Climate Change: Climate Variability: North Atlantic and Arctic Oscillation. **Tropical Meteorology and Climate:** El Niño and the Southern Oscillation: Theory.

Further Reading

- Bretherton, C.S., Smith, C., Wallace, J.M., 1992. An intercomparison of methods for finding coupled patterns in climate data. *Journal of Climate* 5, 541–560.
- North, G.R., Bell, T.L., Cahalan, R.F., Moeng, F.J., 1982. Sampling errors in the estimation of empirical orthogonal functions. *Monthly Weather Review* 110, 699–706.
- Strang, G., 1988. *Linear Algebra and Its Applications*, third ed. Harcourt-Brace-Jovanovich, New York.
- von Storch, H., Zwiers, F.W., 1999. *Statistical Analysis in Climate Research*. Cambridge University Press, Cambridge.

Data Analysis: Time Series Analysis

GR North, Texas A&M University, College Station, TX, USA

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

A sequence of data entries is called a time series. Most time series have a random character. A time series where there is no correlation from one member of the series to another is called *white noise*. Models of time series can be constructed from white noise by having each new entry a sum of a white noise contribution and a linear combination of previous entries. Special insights can be gained by examining the Fourier representation of the time series by developing it into a series of sinusoids of different frequency. A brief review of the technique is presented with attention to errors due to the limited sample or record length of a given time series.

Introduction

The study of the records of weather and climate over time is extremely valuable for practical as well as theoretical purposes. For example, the instantaneous temperature or humidity taken at hourly intervals constitutes a time series of measurements. Long records of data might be used for many purposes such as assembling a climatological summary for the site or for performing an analysis to identify the underlying physical processes responsible for certain interesting features of the data. In such an example, one might not be surprised to find a daily swing of the temperature and humidity as well as an annual oscillation. Were we to plot these hourly temperatures, a nearly repeating graph of period 24 h would result (the diurnal cycle). Examination over longer time spans leads to the identification of an annual cycle in the data. By simple inspection, one would rather quickly decide that there is some underlying physical agent responsible for these near repetitions or periodic statistics in the data stream. This is an example of the most primitive form of time series analysis.

Many decades of experience have led environmental and statistical scientists to devise very sophisticated methods of studying records of time series. The most powerful innovation is the idea of a mathematical model of the time series. A mathematical model involving random components is a very convenient way of representing a time series of data. Such models always employ simplifying assumptions, but such techniques work in a surprisingly large number of applications, especially in the geosciences. The idea of a random variable must be introduced before describing how such a time series can be constructed. A random variable can take on values drawn from a certain probability distribution. For example, the outcomes of flipping a coin are a random variable. Each flip constitutes a realization of the random variable – either heads (H) or tails (T). The probability distribution is 50–50 (probability 0.5 for H, 0.5 for T). More commonly encountered in geophysical and behavioral sciences is the case where the variable can take on a range of discrete or continuous values and its frequency distribution is the normal or bell-shaped curve distribution. An example is the distribution of heights of individuals. Drawing names from a hat and announcing the height of the individuals is equivalent to generating realizations of the height random variable. Most realizations of such a process yield values near the center of the bell-shaped curve, with large excursions from the

population average being very rare. The mean of a random variable is the arithmetic average of its values taken over many realizations (actually the limiting value derived from infinitely many). The variance is the average of the squares of the deviations from the mean. The deviation from the mean for two variables can be multiplied together and averaged over many realizations to form the covariance of the two random variables. As the name implies, the covariance is a measure of how two random variables covary. If the two variables are identical then their covariance is simply the variance. If the two variables are not statistically related to one another then the covariance is zero. The correlation between two random variables is the covariance divided by the square root of the product of the variances. Its magnitude is always less than or equal to unity. Anticorrelated variables have negative correlation (and covariance). In a time series, one is often interested in how an entry covaries with or is related to past (or future) entries. In particular, one would like to know if it is correlated with immediately past entries, a property known as serial correlation.

White Noise

The simplest model of a randomly based time series is one in which each new time step leads to a statistically independent (no serial correlation) drawing (like the heights of individuals above); but each new entry Z_n is drawn from an identical distribution. Each data entry is statistically independent of the previous ones. This model is called the statistically independent, identically distributed *white noise* model. It is the most important model in time series analysis since almost all other statistical models are derived from it. A few properties of the white noise model (it is hereafter assumed that the identically distributed property is included) are worth mentioning. First, the mean of the entries of such a time series can be estimated by adding the values of a long series of entries and dividing by the number of entries used. In doing this, an estimate of the underlying mean of the probability distribution of the individual entries is found. Another property of time series that is easily demonstrated with the white noise model time series is that an independent realization of a time series. Above the random variable was introduced. In time series, there is a random function or string of random variables, with individual realizations being individual graphs of the function. The property that is so evident is that

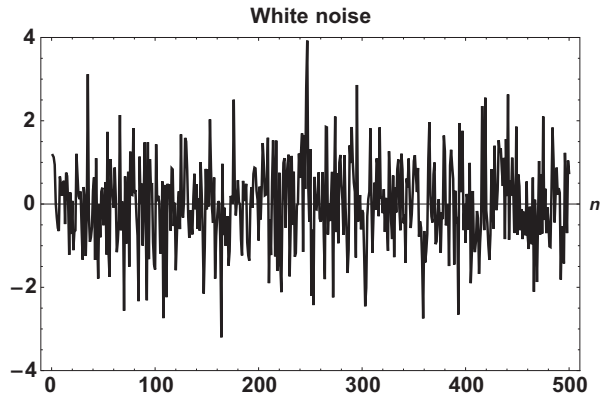


Figure 1 A white noise process with normally distributed entries having mean zero and standard deviation unity. The time series has length $N = 500$.

averaging along the time series is equivalent to fixing the time and averaging across realizations to find the mean. This property of *ensemble averaging* is a very powerful one, which enables many of the proofs and analyses of time series analysis. An example of a single realization of a white noise process is shown in [Figure 1](#). An example of a white noise time series is the heights of students standing in a cafeteria queue. On the other hand, consider the heights of succeeding first sons in an ancestral sequence. There is a genetically determined correlation between the height of a father and that of his son. Such a time series exhibits a positive serial correlation which diminishes to zero after a few generations, a phenomenon known as regression to the mean.

Stationary Time Series

Consider next a time series that is not necessarily white noise. That is to say, an entry may not necessarily be statistically independent of its predecessor – there may be serial correlation. Nevertheless, the time series may not have any preferred origin. That is, as seen along the time series, statistically speaking, each time is equivalent to each other time. In such a time series, the mean is independent of time and so is the variance. The covariance between an entry and that a certain number of steps, say n , earlier depends only on the temporal separation or lag, m . A time series having the above properties is called a stationary time series and these are common in nature. (Strictly speaking, this only defines a second moment stationary time series, since nonstationary properties of the probability distribution may still be present. If the random variables are normally distributed, these mean and covariance stationary properties suffice to determine the strong forms of stationarity.) Perhaps some examples of nonstationary time series will help clarify the concept. The diurnal or seasonal data mentioned above are examples of nonstationary time series since their means depend on local time of day or time of year. In addition, their variances might also have such a phase dependence; even their serial correlation structure may have a phase dependence for example, the serial correlation between entries may be greater in winter than in summer. At first glance, the sequence of heights of first sons across generations may

seem like a stationary time series, but there is known to be a secular trend of increasing heights over generations, probably because of better nutrition.

Despite the ability to enumerate many time series in nature that are nonstationary, the model of a stationary time series is very valuable in the geosciences. For example, annual averages of temperature at a location are likely to form a stationary time series at least to a good approximation. In some cases, the nonstationarity can be eliminated by removal of a trend in the mean or by some other simple modification. The statistics of such time series (mean, variance, serial correlation properties) make a good summary of the sequence and for many purposes may form an adequate substitute in practical applications. For example, an insurance company may want to know the likelihood of the temperature (or flood water level) exceeding a given threshold. The serial correlation structure is particularly important in drought, where sequences of dry years can be the most important indicator of consequences.

Autoregressive Processes

The most common type of time series encountered in the geosciences is the first-order autoregressive process (known as the AR1 process). In this process, each new entry can be written mathematically as the sum of two terms, the first proportional to the previous entry and the second an additive white noise term.

$$T_{n+1} = \lambda T_n + Z_n \quad [1]$$

where T_n is the random variable at time step n and Z_n is a white noise disturbance ordinarily taken to be normally distributed with mean zero and variance unity. λ is a parameter ($|\lambda| \leq 1$) that determines the lagged autocorrelation function, which can be shown to be given by λ^m , where m is the lag. Values of λ close to unity lead to long autocorrelation times, and small λ yields short autocorrelation times. Higher order autoregressive processes (ARp) model the next entry as a sum of $p + 1$ weighted terms. The weights are parameters similar to λ above. The final term is the additive white noise Z_n . Here it is concentrated on the AR1 process because of its central importance especially in geophysical applications. The parameters that describe the time series are its mean, its variance, and its so-called lag-one serial correlation. It is the job of the analyst to take the given data series and determine the parameter values that come closest to fitting the data. That is, one wants to know the mean, variance, and lag-one serial correlation in the data. If the lag-one serial correlation turns out to vanish, then one infers that the series can be modeled by a white noise time series (AR0). For the AR1 process, if the lag-one serial correlation is λ then the lag-two is λ^2 , and so on. In the limit of very small time steps in the series, this tends to an exponential falloff of serial correlation. The value of n for which the serial correlation falls to $1/e$ (e is the base of natural logarithms = 2.718 ...) is known as the autocorrelation time. The autocorrelation time is a measure of the memory of the system. It is often said that the system forgets its past values after a few autocorrelation times ([Figures 2 and 3](#)). The heavy line in [Figure 3](#) shows the ideal autocorrelation function λ^n for the process, while the thin lines indicate sample estimates based on a time series of length 500. The spread of these estimates suggests

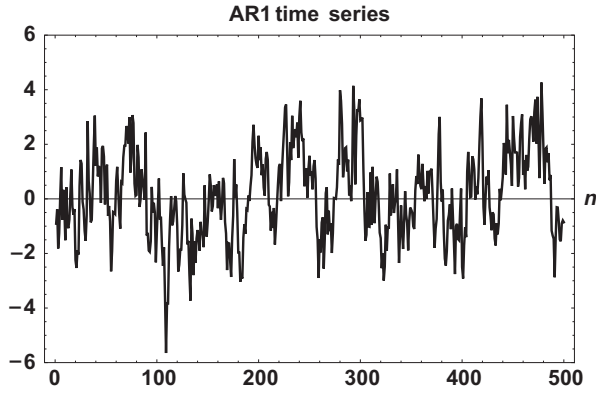


Figure 2 Realization of an AR1 time series with $\lambda = 0.8$. The time series has length $N = 500$. The discrete points (spaced 1 time unit apart along the time axis) along the curve have been connected with straight lines.

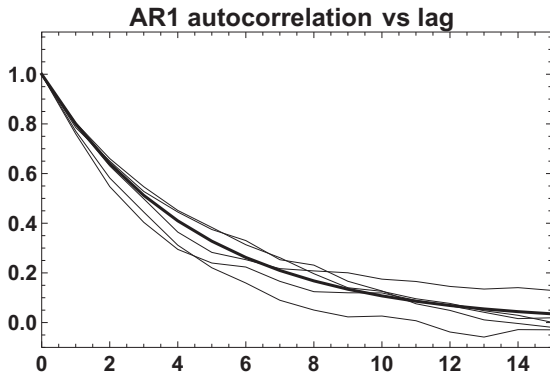


Figure 3 The lagged correlation function for the AR1 process with $\lambda = 0.8$. The discrete points (spaced one lag unit apart along the horizontal axis) along the curves have been connected with straight lines. The heavy line represents the ideal autocorrelation for $N \rightarrow \infty$, while the thin lines are sample estimates from five realizations of length $N = 500$.

size of the error involved in using a sample even as low as $N = 500$, especially when the value of λ is close to unity.

Fourier Analysis of Time Series

If there are physical reasons to think that a time series of data is stationary, then Fourier analysis of the data can lead to a number of powerful techniques useful in applications. One begins the analysis by taking the finite-length segment of data in the sequence and estimating the Fourier coefficients for representing the data as a Fourier series on the segment.

In this process, one is representing the data in terms of the Fourier coefficients instead of the temporal entries. The two are equivalent ways of expressing the content of the data. The discrete Fourier transform and its inverse are given by:

$$\tilde{T}(f_s) = \sum_{n=0}^{N-1} T_n e^{2\pi i f_s n} \quad \text{and} \quad T_n = \frac{1}{N} \sum_{s=0}^{N-1} \tilde{T}(f_s) e^{-2\pi i f_s n} \quad [2]$$

where $f_s = s/N$; $s = 0, 1, 2, \dots, N-1$; and N is the length of the sample time series. $\tilde{T}(f_s)$ is a component or (complex)

amplitude of a certain sinusoidal waveform in the data stream. From the point of view of time series modeling, the Fourier coefficients $\tilde{T}(f_s)$ are random variables, since from one realization of the process on the same time span to another realization the coefficients will differ. However, they will have certain statistical properties common across the ensemble of realizations. If the segment is sufficiently long and the series is stationary, it can be shown that the Fourier coefficients corresponding to different frequencies are uncorrelated – a very powerful statement. This permits us to perform an analysis of variance over the different frequency bands to examine how variance is distributed (additively) over frequencies. It is routine to plot a graph of the variance or sometimes known as power as a function of frequency. This is known as *spectral analysis*.

The most common example is the white noise time series. The white noise spectrum is flat; that is, every equal width frequency band has allotted to it the same variance. Hence, one way of determining whether a certain time series is white noise is to perform the Fourier analysis and plot the spectrum (variance or power versus frequency). The *periodogram* is the most elementary but direct characterization of the sample spectrum.

$$S^{(p)}(f_s) = \frac{1}{N} \left| \sum_{n=1}^N T_n e^{2\pi i f_s n} \right|^2 \quad [3]$$

An ensemble average across many equal length temporal segments of the same process leads to the *spectral density* if f_s is taken to be a continuous variable (smoothing across the discrete values f_s). If the spectrum is flat, one can infer that the time series is white noise. Of course, if the time series segment is short, there will be problems in estimating the spectrum of the underlying process because of sampling error. This is illustrated in **Figure 4** for a white noise sample of length 500.

The periodogram shown in **Figure 4** is anything but flat. It represents the magnitude squared of each of the discrete Fourier coefficients. Since each periodogram component is half the sum of the squares of the sine and cosine coefficients (each of which is normally distributed), the random variable representing each Fourier component of the periodogram can be shown

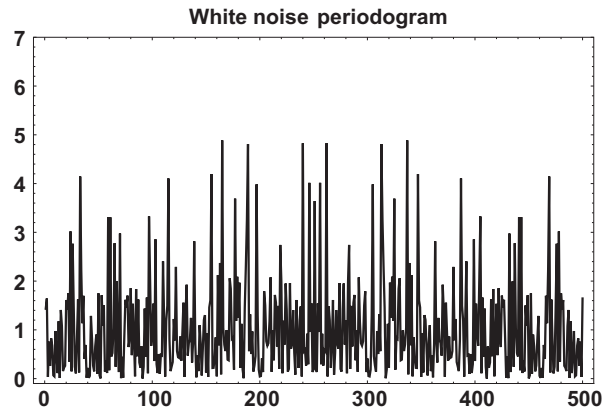


Figure 4 Periodogram for a white noise process with $N = 500$. The mean of this periodogram entries is 1.001 79, and the variance is 0.891 73; theoretically they should approach unity as N increases. Note that the periodogram is an even function about $s = 250$ or $f_s = 0.5$.

to obey an exponential distribution with mean and variance unity (actually this is one half the sum of two $\chi^2_{[1]}$ variables). Adding more terms to the time series (increasing N) does not reduce the error for an individual Fourier element in the periodogram. If one averages across many frequencies, one can smooth the periodogram and reduce the error variance of the spikes. The averaging works because the separate frequency variance components are statistically independent and this means that averaging over more lines smooths out the error proportionally to $1/\sqrt{w}$, where w is the width of the frequency band. This is one way to reduce the error for a band. The tradeoff is that by reducing error by widening a band, you give up spectral (frequency) resolution. The approach using the periodogram to estimate spectral density has many faults including bias (error in estimating the mean of a quantity even as $N \rightarrow \infty$) along with random sampling error. **Figure 5** shows the log of the ideal spectral density for $N \rightarrow \infty$ as the heavy black line. The colored thin lines are periodogram estimates based on four realizations of the AR1 process with $\lambda = 0.8$ and for $N = 500$. The variance about the heavy black line is large with many spurious peaks due to sampling error. In addition, there is a noticeable high bias that appears to be independent of frequency.

Errors are inevitably introduced because the sample has finite length, and attempts are being made to estimate the spectral density of an ideal infinite length time series from which the sample record was drawn. A raw finite length sample as in the time series shown in **Figure 1** begins abruptly at $n = 1$ and ends suddenly at $n = N$. This is equivalent to taking a segment of the infinitely long time series and multiplying it by a hard-edged rectangular function. Fourier transforms do not like sharp edges and they tend to introduce biases in the transformed variable. By way of a far too brief explanation, note that the absolute magnitude squared of the Fourier transform of a rectangular function symmetric about the origin has a central hump in frequency but with wiggly lobes on either side. The side lobes mix variance that should appear at one frequency band into frequency ranges nearby – peaks can turn into

multiple peaks or they can get displaced or smeared, etc. One way to mitigate this weird behavior is to apply a smooth taper on the time series segment at each end, but this is still unsatisfactory, although it is a step in the right direction. Some more satisfactory methods will be briefly discussed in the next section.

Introduction to Advanced Methods

The lagged covariance was introduced earlier. For a stationary time series T_n , it is given by

$$\tilde{\gamma}_m = \frac{1}{N-m} \sum_{n=1}^{N-m} T_n T_{n+m} \quad [4]$$

where $\tilde{\cdot}$ is used to indicate that the lagged covariance is a sample estimate. It is not difficult to show that eqn [3] is equivalent to

$$\tilde{S}(f_s) = \sum_{m=-M}^M \gamma_{|m|} e^{2\pi i f_s m} \quad [5]$$

where $M = \frac{1}{2}(N-1)$; N needs to be odd. If there is a periodic component of T_n , its lagged correlation will peak at lags equal to integral multiples of the periodic signal. The way to smooth the effects of the abrupt beginning and ending is to introduce a so-called lag window w_n into eqn [6].

$$\tilde{S}(f_s) = \sum_{m=-M}^M \gamma_{|m|} w_{|m|} e^{2\pi i f_s m} \quad [6]$$

The lag window $w_{|m|}$ can be a positive valued hump-shaped function (whose sum adds to unity) peaked $m = 0$. The hump needs to be wide enough to include roughly 10 oscillations of any important periodic signals that are being sought out in the spectrum. This gives us an idea of how long a time series segment needs to be in order to distinguish low-frequency phenomena. Numerous lag window shapes have been studied

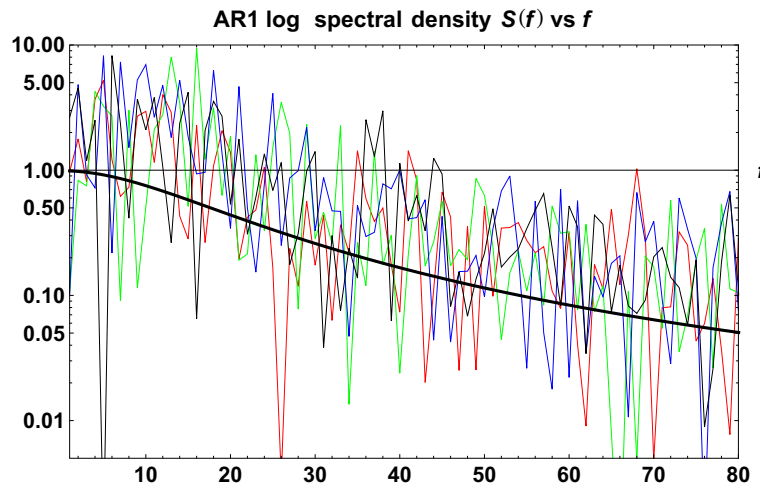


Figure 5 Heavy black line: the ideal log–log spectral density versus frequency for the AR1 process ($\lambda = 0.8$) shown in **Figure 4**. The values of the ordinate are connected with straight lines. The time series has length $N = 500$. Frequency is in units of $1/250$ cycle per unit time. The sampling rate is 1 per unit time for this process. The thin colored lines are periodograms for four separate realizations of the time series.

including a tent shape (Bartlett window) and various smoother ones including the Parzen window. They have the property of introducing an equivalent bandwidth to the spectral estimate similar to smoothing the periodogram, but both approaches are subject to biases. Figure 6 shows log of spectral density estimated with the Parzen window ($m = 200$) for the ideal spectrum (heavy black line, same as in Figure 5) with four realizations of the same time series, of length $N = 500$. While the Parzen window technique reduces the variance about the true spectral density, it shows a definite high bias at high frequencies.

The most favored methods are the multitaper techniques. Studies suggest that the bias at higher frequencies is somewhat reduced by the multitapering method. This approach introduces a series of data tapering functions. These functions are mutually orthogonal and have certain other optimal properties, related to maximizing the variance in a given spectral band for a given length of the time series segment available. The procedure is to obtain the spectral estimate from each of the tapered time series records, then estimate the spectrum from each, followed by taking a weighted average of the spectral estimates. The tapering functions turn out to be the discrete prolate spheroidal wave functions. This technique is discussed in the book by Percival and Walden (1993).

Applications of Time Series Analysis

Not only do time series analyses provide new insights into the underlying physical processes generating an empirical time series, but they are useful in a variety of practical applications.

A very common use of time series analysis is data smoothing. Often one wishes to smooth out the highly irregular short-term fluctuations in a time series to get a better view of longer term trends or undulations. This can be accomplished by running a smoother over the time series. For example, one might take as the value at a certain time the arithmetic average

over several future and past values. These past and future entries can be weighted in various ways to make the smoothing optimal for the particular application. This class of operations is known as moving average smoothing. Perhaps, the most important application is forecasting.

One might ask, given a segment of a data-derived time series, is it possible to use this information to forecast future entries. The answer in principle is simple. In the case of a white noise time series (or an empirical one which is indistinguishable from white noise), there can be no forecast skill, since the next entry is statistically independent of the past entries. But in the case of an AR1 process, there is correlation with past entries and this will permit some statistical estimate of future entries out to roughly one or two autocorrelation times. The estimate will not only provide a most probable value of the future entry, but some assessment of the uncertainty in the forecast perhaps even a theoretical frequency distribution of values that can be expected. Interpolation is a second application of time series modeling. Suppose there are missing values in an empirical time series and for some reason one wishes to insert values that are statistically consistent with the rest of the entries. First, one finds a model of the time series and then one can find the most probable entry with an associated theoretical frequency distribution. Depending on the application, one may wish to insert the most probable value or add to it a random number which is consistent in a statistical sense with the nearest neighbors. A third application is in the area of signal processing. One often finds a deterministic signal embedded in some kind of noise (white or colored). The object usually is to separate the noise from the signal and to estimate the amplitude of the signal. This is the problem in radio reception. The process is referred as detection. By time series modeling and knowing some characteristics of the signal waveform, one can find an optimal estimate of the signal strength. In electronics, one might want to clean the noise away from the signal and amplify the residual, while in other applications, one might want to find the amplitude of a periodic signal such as the diurnal or

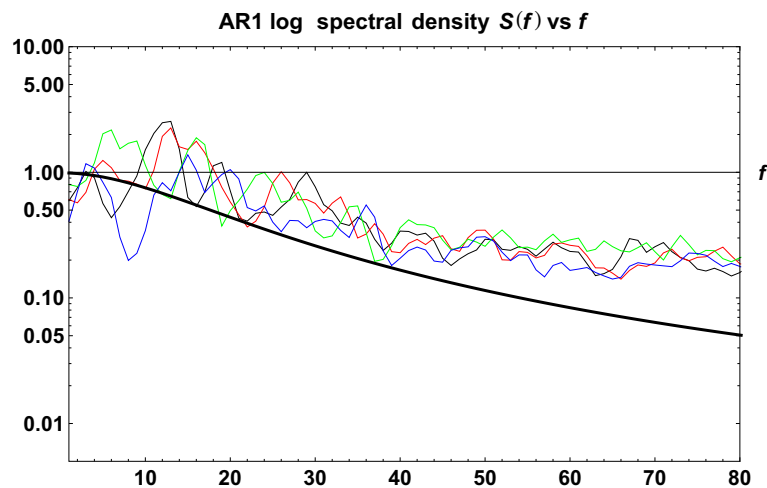


Figure 6 Heavy black line: the ideal log spectral density versus frequency for the AR1 process ($\lambda = 0.8$) shown in Figure 5. The values of the ordinate are connected with straight lines. The time series has length $N = 500$. Frequency is in units of 1/250 cycle per unit time, and the horizontal axis only extends to 80 of 250 cycles per unit time increment. The sampling rate is 1 per unit time for this process. The thin colored lines are periodograms for four separate realizations of the time series.

seasonal cycle. One of the most famous applications of this type in the geosciences is the detection of periodic signals in the record of changes in continental ice sheet volume. These excursions have been found to contain significant variance peaked at periods of 100 000, 43 000, and 20 000 years, and these just happen to coincide with the periods of the changes in the elliptical orbital parameters of planet Earth (eccentricity, obliquity, and precession of the equinoxes). Thus, time series analysis was able to show conclusively that the ice ages are linked to the changes in the Earth's orbital elements.

See also: **Statistical Methods:** Data Analysis: Empirical Orthogonal Functions and Singular Vectors.

Further Reading

- Bendat, J.S., Piersol, A.G., 1986. *Random Data: Analysis and Measurement Procedures*. Wiley, New York.
- Bloomfield, P., 1976. *Fourier Analysis of Time Series: An Introduction*. Wiley, New York.
- Brockwell, P.J., Davis, R.A., 1999. *Time Series: Theory and Methods*. In: second ed. (Ed.). Springer-Verlag, New York, 577 pp.
- Chatfield, C., 1992. *The Analysis of Time Series: An Introduction*. Chapman & Hall, New York.
- Percival, D.B., Walden, A.T., 1993. *Spectral Analysis for Physical Applications*. Cambridge University Press, Cambridge.
- Von Storch, H., Zwiers, F.W., 1999. *Statistical Analysis in Climate Research*. Cambridge University Press, Cambridge, UK, 484 pp.
- Wei, W.W.S., 1990. *Time Series Analysis*. Addison-Wesley, Redwood City, CA.

STRATOSPHERIC CHEMISTRY TOPICS

Contents

Overview

Halogens

Halogen Sources, Anthropogenic

Halogen Sources, Natural (Methyl Bromide and Related Gases)

HO_x

Hydrogen Budget

Reactive Nitrogen (NO_x and NO_y)

Stratospheric Water Vapor

Overview

JA Pyle, University of Cambridge, Cambridge, UK

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

Ozone is a key stratospheric constituent. Its basic variations with latitude, height, and season are presented. The main chemical cycles are introduced. These are the Chapman (oxygen-only) reactions and a variety of catalytic cycles, driven by radical species whose stratospheric concentrations are orders of magnitude lower than that of ozone.

Introduction and Background

Ozone is perhaps the most important stratospheric constituent. It absorbs solar ultraviolet radiation, particularly strongly at wavelengths below about 310 nm, where stratospheric ozone acts as a filter to protect life at the surface from these potentially harmful wavelengths. Absorption of solar radiation by ozone also results in heating of the stratosphere and leads to the observed stable temperature structure, where temperature increases with height throughout the stratosphere. Ozone is also infrared active and is an important gas for the climate system. For these reasons, the chemistry of the stratosphere is essentially the chemistry of ozone and the minor constituents involved in ozone chemistry.

In the troposphere, ozone is present in mixing ratios (the ratio of the concentration of ozone to that of air) of a few tens of parts per billion by volume (approximately $\text{few} \times 10^{-9}$, or a few ppbv) but its peak mixing ratio is much greater in the stratosphere, reaching almost 10 ppm (10×10^{-6} , or 10 ppmv) at just above 30 km (10 hPa) in low latitudes. **Figure 1** shows seasonally averaged mixing ratios of ozone in the stratosphere and mesosphere based on satellite and ozone sonde data. In contrast, the largest absolute concentrations of ozone are found

in high latitudes between 20 and 25 km and reach a few times 10^{12} molecules cm^{-3} . Another useful measure for ozone is its column abundance, the vertically integrated density of ozone above the surface (also often called 'total ozone'). For most species column densities are measured in molecules per square centimeter. For ozone, the traditional unit is the Dobson unit (DU), named after the Oxford scientist, who pioneered the routine measurement of column densities using spectrophotometers to measure the absorption by atmospheric ozone of the solar spectrum in the 1920s. A DU is a thickness of 1 millimeter at standard temperature and pressure. Typical column densities are 250 DU in the tropics, with little seasonal variation, and 400 DU in high latitudes in winter and spring. **Figure 2** shows the average variation of the ozone column, as a function of latitude and month, obtained from satellite measurements by the total ozone mapping spectrometer (TOMS) satellite instrument.

As we will see below, the ozone distribution is in part controlled by radical species, which themselves are present in even lower concentrations. Typical mixing ratios of the oxides of nitrogen are in the part per billion range, while for active chlorine species peak values are usually around or below a part per billion. Mixing ratios of odd-hydrogen species are even lower. The radicals themselves are produced from source gases,

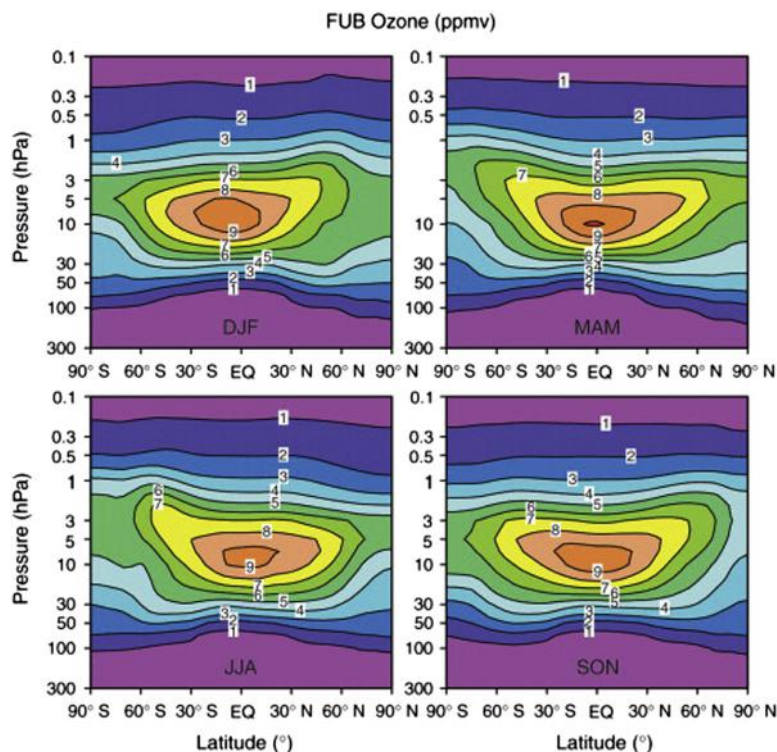


Figure 1 Seasonally averaged zonal mean cross section of ozone mixing ratio (ppmv) constructed from a combination of satellite and ozone sonde data by the Free University of Berlin. Courtesy of Dr Ulrike Langematz (FUB).

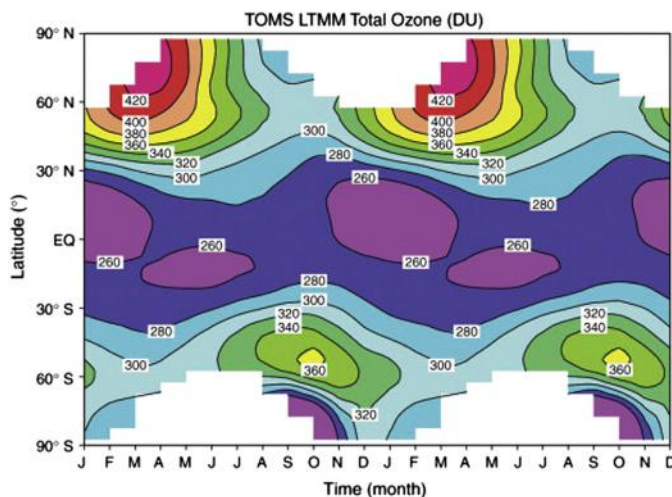


Figure 2 Long-term average latitude–time variation of the monthly and zonally averaged ozone column density (Dobson units) based on the available TOMS satellite data from November 1978 to December 1999. The values are repeated for 2 years to emphasize the annual cycle. The instrument measures reflected solar radiation, and the data gaps are in regions of darkness or twilight. Figure produced by Dr Peter Braesicke (University of Cambridge) from the original TOMS data (<http://toms.gsfc.nasa.gov/ozone/ozone.html>).

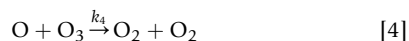
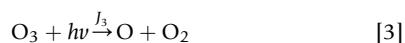
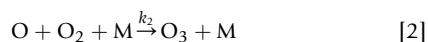
of both natural and anthropogenic origin and emitted in the troposphere. Thus, nitrous oxide is emitted from the Earth's surface and is relatively inert, and hence well mixed, in the troposphere with a present concentration of about 310 ppbv. It is oxidized in the stratosphere to produce NO (and hence NO₂). Similarly, water vapor (2–6 ppmv in the stratosphere) and methane (about 1.5 ppmv at the tropical tropopause) are

oxidized to yield the odd-hydrogen species H, OH, and HO₂. The halogen species, which have played an important role in ozone depletion during the last two decades, are mainly of recent anthropogenic origin. Their major source gases include CH₃Cl (with predominantly natural sources), CF₂Cl₂, and CFC₃. These latter species are the so-called freons, which were widely used in aerosol spray cans, refrigeration, and foam

blowing and are now regulated under the Montreal Protocol. Along with a number of other chlorinated species, these led to the present-day abundance of chlorine in the stratosphere of about 3.5 ppbv. Similarly, there is about 20 ppt of bromine in the stratosphere, arising from the degradation of methyl bromide (which has both natural and anthropogenic sources) and other industrially produced bromocarbons, used, for example, as fire retardants.

Odd Oxygen and the Chapman Reactions

In 1930, Sidney Chapman proposed a series of reactions to explain the distribution of stratospheric ozone of which the most important are:



(M represents any third body, usually N_2 or O_2 , required to conserve energy and momentum in a termolecular reaction).

Note that, in the troposphere and stratosphere, the photolysis of oxygen is much slower than the photolysis of ozone. Reactions [2] and [3] are rapid and have very short time constants for the conversion of O to O_3 and vice versa, and they establish a steady state much more rapidly than reactions [1] and [4].

$$J_3[\text{O}_3] = k_2[\text{O}][\text{O}_2][\text{M}] \quad [5]$$

([] represents concentration).

Note also that reactions [2] and [3] only interconvert the 'odd-oxygen' species O and O_3 ; i.e., they conserve odd oxygen ($[\text{O}] + [\text{O}_3]$). In contrast, odd oxygen is formed by reaction [1] and removed by reaction [4].

Thus, we can write for the rate of change of odd oxygen

$$d([\text{O}] + [\text{O}_3])/dt = 2J_1[\text{O}_2] - 2k_4[\text{O}][\text{O}_3] \quad [6]$$

The time scale for steady state between reactions [1] and [4] varies strongly with altitude, being on the order of hours at 40 km but on the order of many years at 20 km. Invoking steady state in the upper stratosphere (i.e., setting $d([\text{O}] + [\text{O}_3])/dt = 0$) is thus a good approximation. In the low stratosphere, it would clearly be a poor approximation since many external factors (the intensity of solar radiation, temperature, atmospheric transport, etc.) will all vary much more rapidly.

In the upper stratosphere, setting $d([\text{O}] + [\text{O}_3])/dt = 0$, we can calculate the steady-state distribution of ozone from eqns [5] and [6]:

$$[\text{O}_3] = (k_2J_1[\text{O}_2]^2[\text{M}]/k_4J_3)^{1/2} \quad [7]$$

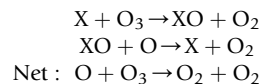
Equation [7] describes the steady-state ozone concentration in an oxygen-only atmosphere. The vertical profile derived from eqn [7] is consistent with the shape (but not the

magnitude) of the observed profile, especially in low latitudes. Thus, eqn [7] predicts a peak in the ozone mixing ratio at a little above 30 km. However, this equation also predicts that the ozone concentration should be very low in high latitudes, when, for example, the photolysis rate of molecular oxygen, J_1 , becomes very low. However, observations show large column amounts of ozone in high latitudes in winter and spring (see Figure 2), when photolysis will be at its slowest. The reason for the discrepancy lies in the long photochemical time constant for ozone at low altitudes. When the time constant is long, the transport of ozone must also be considered so that the continuity equation for odd oxygen (eqn [6]) must also include terms to describe the transport. In reality, ozone is produced in a source region in the low latitude middle stratosphere and moved to higher latitudes, where ozone is slowly destroyed, by the action of the stratospheric general circulation.

For many years, it was thought that Chapman's model could adequately explain the distribution of stratospheric ozone, at least in the middle and upper stratosphere. However, with improved measurements – both in the laboratory and in the atmosphere – it became apparent that reaction [4] only removes about 25% of the odd oxygen produced by oxygen photolysis. Calculations based on just the Chapman reactions will seriously overestimate stratospheric ozone concentrations, even when the photochemical time constant is short.

Catalytic Cycles

Reaction [4] has an unexpectedly high activation energy for such an exothermic reaction. It was realized that, at stratospheric temperatures (200–290 K), odd oxygen could be removed efficiently in catalytic cycles, which achieve the same result as reaction [4] without loss of the catalytic species X or XO:

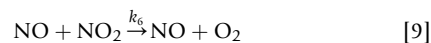
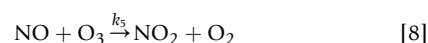


(i.e., the two reactions effectively catalyze reaction [4]).

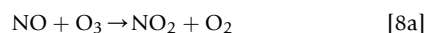
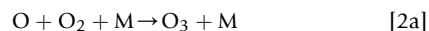
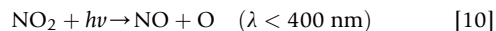
Cycles of this kind were discussed for mesospheric chemistry by David Bates and Marcel Nicolet in the 1950s. In the late 1960s and the early 1970s attention switched to their role in stratospheric chemistry, pioneered by, for example, Harold Johnston, Paul Crutzen, Mario Molina, and Sherry Rowland, who all highlighted an important potential role in ozone depletion for these cycles. They showed that if the concentration of X increases, the ozone concentration would fall: ozone would be depleted.

There are a number of candidates for X present in the stratosphere. These include NO, H, OH, Cl, and Br, all discussed in detail in separate articles. Here, we will take the cycle involving the oxides of nitrogen (NO and NO_2 , members of the odd-nitrogen family) as a single example of odd-oxygen destruction by these catalytic cycles.

So, substituting X = NO in the catalytic cycle,



This cycle is responsible for about 50% of odd-oxygen removal from the stratosphere, despite a number of competing reactions of which the most important is NO_2 photolysis, very rapid even at low altitudes. This produces a 'null cycle':



Assuming steady state between NO and NO_2 based on reactions [8]–[10] (a very good approximation) then, after some simple algebra, the rate of odd-oxygen change by the nitrogen oxides can be written,

$$d([\text{O}] + [\text{O}_3])/dt = -2k_6[\text{NO}_2][\text{O}] \quad [11]$$

and the total rate of odd-oxygen change for the combined Chapman and odd-nitrogen cycles would be given by adding eqns [6]–[11]. Similarly, other cycles, where XCl, OH, etc., have loss rates of the form given by eqn [11]; the rate-limiting step usually involves the reaction of XO with atomic oxygen, O. The concentration of O is low in the low stratosphere (since the rate at which O recombines to form O_3 , reaction [2], increases with increasing pressure) and thus the odd-oxygen loss rates are lower in the low stratosphere, leading to the longer photochemical time scales there.

These cycles dominate the middle atmosphere away from polar latitudes. In polar latitudes severe ozone depletion has been observed in recent years, forced by halogen chemistry and with the halogens turned into active form by reactions on polar stratospheric clouds, at the low temperatures found there. The cycles are again catalytic and involve both ClO and BrO.

One final general point is worth making, again to be discussed in more detail in the articles discussing the individual

chemical families. This is that in addition to the radical species involved in the catalytic cycles, other family members exist and can play important roles. For example, HNO_3 is an important reservoir species for odd nitrogen; i.e., a species which is a 'holding tank' for NO and NO_2 (and indeed OH and HO_2) but does not take part in ozone-destruction cycles. Similarly, HCl and ClONO₂, the reservoirs for odd chlorine, are usually the dominant form of chlorine in the lower stratosphere, a fact that limits chlorine-catalyzed ozone destruction, away from polar latitudes, mainly to the upper stratosphere.

See also: **Chemistry of the Atmosphere:** Chemical Kinetics; Observations for Chemistry (*In Situ*); Ozone Sondes. **Middle Atmosphere:** Transport Circulation. **Ozone Depletion and Related Topics:** Photochemistry of Ozone. **Stratospheric Chemistry Topics:** HO_x ; Halogen Sources, Anthropogenic; Halogen Sources, Natural (Methyl Bromide and Related Gases); Halogens; Hydrogen Budget; Reactive Nitrogen (NO_x and NO_y); Stratospheric Water Vapor.

Further Reading

An excellent series of reviews of stratospheric ozone have been published by the World Meteorological Organization as part of their 'Global Ozone Research and Monitoring Project' 1988. The most recent report is No. 44, Scientific Assessment of Ozone Depletion.

Brasseur, G., Solomon, S., 1986. *Aeronomy of the Middle Atmosphere*. Reidel, Dordrecht.

Finlayson-Pitts, B.J., Pitts Jr., J.N., 2000. *Chemistry of the Upper and Lower Atmosphere*. Academic Press, New York.

Wayne, R.P., 2000. *Chemistry of Atmospheres*. Oxford University Press, Oxford.

Halogens

D Toohey, University of Colorado Boulder, Boulder, CO, USA

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

Halogen-containing compounds from Earth's surface that reach the stratosphere release their constituent atoms by exposure to ultraviolet light and oxidation. The halogen atoms, mainly chlorine and bromine, undergo a series of photochemical reactions that catalytically destroy ozone. Consequently, very small abundances of halogens produce dramatic losses of ozone, such as those observed over the polar regions in springtime and in the exhaust plumes of rockets fueled by chlorine-containing propellants. This article summarizes the interactions between the chemical substances that determine the abundances of the highly reactive halogen free radicals in the stratosphere and the catalytic reactions that destroy ozone.

Introduction

The potential impact of halogen atoms (fluorine, chlorine, bromine, and iodine) on the chemistry of stratospheric ozone (O_3) was first recognized in the early 1970s, not long after researchers proposed that nitrogen oxides (NO_x) and hydrogen oxides (HO_x) could destroy ozone. These halogen atoms are produced by compounds that are relatively unreactive in the troposphere but that decompose photochemically in the presence of shortwave ultraviolet radiation in the stratosphere. Among such compounds are those known as halocarbons, which are predominantly industrial in origin.

For much of the second half of the twentieth century, a number of halocarbons were used for a variety of purposes, including refrigeration, manufacturing of foam products, extinguishing of fires, fumigation of crops, and production of polymers. Organisms in the upper ocean produce small, but significant, amounts of several halocarbons. There are only a few ways to destroy most halocarbons once they are released to the atmosphere, including reaction with hydroxyl (OH) (if the halocarbon contains a hydrogen atom), ultraviolet photolysis, and reaction with electronically excited oxygen atoms, $O(^1D)$. However, these processes also initiate the cycle of ozone destruction in the stratosphere.

Halogen atoms are examples of free radicals, species that typically (although not exclusively) possess an odd number of electrons and require an additional electron to fill a molecular orbital to become more stable. Upon collision, a free radical can acquire this additional electron by stripping it from another molecule (called electron transfer), by pulling an atom off the collision partner (called extraction), or by attaching to another free radical (called addition or recombination). With a variety of collision partners available, there are literally hundreds of possible reactions and dozens of inorganic halogen compounds that must be considered for an accurate description of halogen chemistry in the stratosphere. However, only halogen atoms react rapidly with ozone. Thus, atmospheric chemists refer to two types of inorganic halogen species in the stratosphere, free radicals, and reservoirs. Whereas the free radicals are directly involved in ozone destruction, the reservoirs are more stable compounds that do not react directly with ozone. However, the reservoirs can react with other free radicals or break down in sunlight to form free radicals, hence the origin of their name.

Laboratory studies show that the reactivities of the families of halogen compounds follow the general trend $I > Br > Cl > F$; however, stratospheric abundances follow the trend $[Cl] > [F] > [Br] > [I]$. Consequently, most of the ozone destruction by halogens in the stratosphere is due to chlorine and bromine species. Destruction of ozone has been quantitatively linked to chlorine and bromine free radicals, whereas inorganic fluorine species have little impact on ozone. The role of iodine appears to be limited owing to small concentrations in reactive forms.

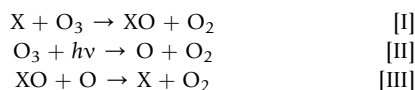
Gas-Phase Halogen Photochemistry

Halogen-containing organic compounds released at the Earth's surface, and that are relatively unreactive in the lower atmosphere, become mixed throughout the lower atmosphere, a process that takes about a year. When they reach the upper troposphere, these gases are slowly transported across the tropopause, primarily in the tropics. As air in the lower tropical stratosphere ascends, these source gases, which contain at least one halogen and one or more carbon atoms, are broken down by processes initiated by shortwave ultraviolet radiation. The halogen atoms that are released then react with ozone and other compounds to form inorganic compounds that contain only halogen, hydrogen, nitrogen, and oxygen atoms. If there is no selective separation (e.g., precipitation), the number-weighted sum of the mixing ratios of all forms of a particular halogen will be conserved. Thus, as the organic compounds break down in sunlight, the abundances of inorganic compounds increase concomitantly. Ultimately, the inorganic halogen compounds are removed from the stratosphere by slow downward transport into the upper troposphere at high latitudes. Because these compounds generally are acidic and water soluble (unlike the organic source gases), they are readily scavenged in the relatively wet troposphere, returning to the Earth's surface with precipitation.

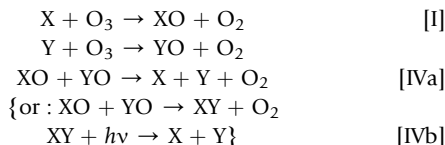
Halogen atoms destroy ozone by a series of catalytic reactions, so called because the halogen free radicals are cycling between various forms with no net change in abundance while the ozone is converted into diatomic oxygen, O_2 . The main catalytic cycles for ozone destruction can be written symbolically (where X and Y represent F, Cl, Br, or I, $h\nu$ represents

a photon of wavelength c/ν , and M is shorthand for an N_2 or O_2 molecule) as follows:

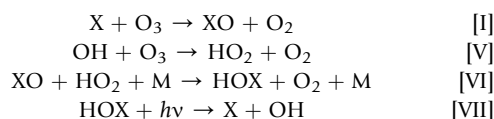
Cycle one



Cycle two



Cycle three



The net result of each of these cycles is loss of two ozone molecules with no change in radical abundance as in eqn [VIII]:

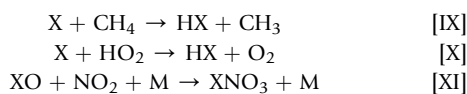


From the law of mass action, the rate of ozone destruction by a catalytic cycle is the product of the concentrations of the reactants and the rate constant for the rate-determining step. Additional reactions (shown below) shift the steady-state balance between the two forms of the halogen radicals in favor of XO, such that the rate-determining steps in each of the three cycles above involve XO radicals. Consequently, we can write the change of ozone vs time as

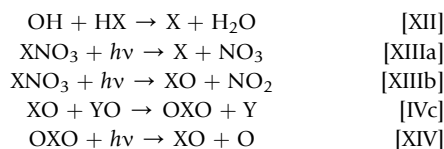
$$\begin{aligned} \frac{-d[O_3]}{dt} &= 2(k_{III}[XO][O] + k_{IV}[XO][YO] + k_{VII}[XO][HO_2] \\ &\quad + \text{minor cycles}) \end{aligned} \quad [1]$$

The factor of 2 in eqn [1] appears because two ozone molecules are destroyed for each pass through a given cycle.

If there were no other reactions to consider, a significant amount of ozone would be destroyed before stratospheric air mixed back into the troposphere. However, the halogen radicals are deactivated by reactions with other species that are present. The main reactions of importance are shown in eqns [IX]–[XI].



There are also important reactions that rerelease the radicals or that produce short-lived reservoirs, including those shown in eqns [XII]–[XIV], [IVc].



Halogen oxides rerelease halogen atoms by reacting rapidly with nitric oxide (NO) produced by photolysis of NO_2 and the reaction $O + NO_2$. Nitric oxide thereby strongly influences the

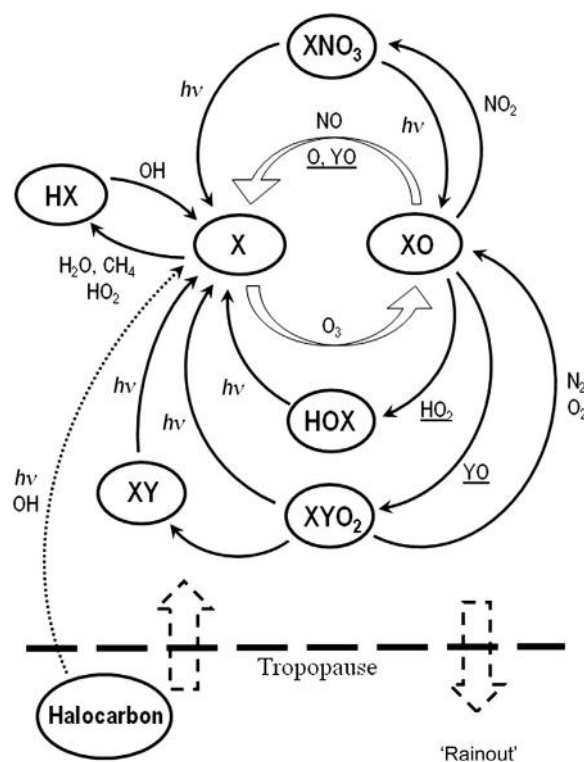


Figure 1 Schematic diagram of gas-phase halogen cycling in the Earth's stratosphere. Open arrows are used for fast exchange between the radical forms of inorganic chlorine. Large dashed arrows represent transport across the tropopause. X and Y are halogen atoms, Cl, Br, I, or F. Processes that are underlined result in catalytic ozone loss. See text for further discussion.

partitioning between the atomic and diatomic halogen radical forms (eqn [XV]).



Except at very high solar zenith angles and at very low altitudes in the stratosphere, most of these reactions occur rapidly, and a steady state is established as the various chemicals cycle from one form to another. This cycling is shown symbolically in Figure 1.

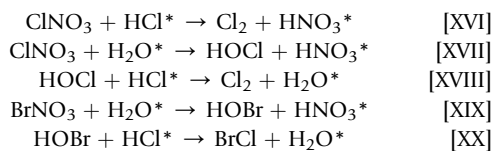
Because the rates of the analogous reactions vary among the different chemical families, the partitioning between the different chemical forms also varies. HF is the main form of inorganic fluorine; HCl and $ClNO_3$ account for more than 90% of inorganic chlorine, except in the polar regions in winter; BrO, $BrNO_3$, and HOBr are the primary inorganic bromine species; and it is believed that IO, I, and HOI are the primary iodine species.

Heterogeneous Halogen Chemistry

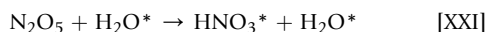
Early studies of stratospheric halogens focused primarily on reactions between gaseous species, or so-called gas-phase chemistry; however, a new class of reactions was necessary to explain the rapid appearance of the Antarctic ozone hole in the

1980s. These reactions, called heterogeneous because they involve the interactions of species in different phases (e.g., between gases and species dissolved in liquids or solids), are typically less rapid than gas-phase reactions because they require the additional processes of adsorption and dissolution. However, under dim sunlight, such as in the winter polar regions or near the bottom of the stratosphere, where most ultraviolet light has been removed by the column of ozone overhead, the rates of heterogeneous reactions can become competitive with those of gas-phase reactions. It is also in these regions that ozone production is slow, such that ozone destruction cycles can have a large local impact.

Extensive laboratory and modeling studies have shown that the following heterogeneous reactions of inorganic halogens have the greatest impact on stratospheric chemistry (* denotes a reactant dissolved in liquid or solid phase):



These reactions convert relatively long-lived reservoirs of chlorine and bromine into species that photolyze readily in dim sunlight to release ozone-destroying radicals, and simultaneously convert short-lived reservoirs of NO_x radicals into long-lived species. Because NO_x limits the reactivities of the halogen compounds to ozone (e.g., reaction [XI], or reaction [XV] followed by reactions [IX] and [X]), its removal results in further enhancements of the halogen oxides, and consequently more severe ozone loss. The reaction



also indirectly enhances the abundances of halogen oxides by converting nitrogen oxides into the long-lived reservoir nitric acid.

Several of these heterogeneous reactions also influence the budget of HO_x by either producing (e.g., eqns [XVII] and [XIX]) or removing (e.g., eqns [XVIII] and [XX]) its short-lived reservoirs. Many of these heterogeneous reactions depend strongly on temperature, and become important in the lower stratosphere only when temperatures drop below about 210 K. Because of this, and the interactions of the halogen radicals with NO_x and HO_x , the response of ozone to changes in temperature or changes in abundances of the halogen source gases can be quite complicated and sometimes counterintuitive. Therefore, detailed computer models are required for accurate assessments of the impact of halogen species on stratospheric ozone.

Observations of Halogen Species

Direct observations of inorganic halogen species form the basis for descriptions of present, and prediction of future, decreases of ozone in the stratosphere. Based on the rate-determining step, it is sufficient to measure the species that control ozone loss (i.e., the halogen oxides) in order to compute the consequent rate of ozone destruction. However, to develop a more definitive understanding of the mechanisms controlling the abundances of the free radicals it is necessary to measure as

many of the inorganic halogen species as possible. Since the mid-1970s, there have been many observations of a large number of inorganic halogen species in the stratosphere. By the year 2000, the following species had been quantified in the stratosphere: fluorine family, HF, CF_2O , and CFCIO ; chlorine family, Cl, ClO, HCl, ClNO_3 , HOCl, OClO, and Cl_2O_2 ; bromine family, BrO, HBr, and BrNO_3 ; and iodine family, IO. The remaining species are predicted to exist at abundances that represent significant challenges for current measurement techniques. Each of the families is examined separately below.

Chlorine

Emissions of industrially produced halocarbons such as CFC-11 (CFCl_3) and CFC-12 (CF_2Cl_2) have delivered about three to four parts per billion of chlorine to the stratosphere, more than all the other halogen families combined. Emissions by volcanoes and solid rocket motors can significantly enhance the local abundances of inorganic chlorine, but otherwise these sources have a small global impact following diffusion and mixing. Equations [IX] and [X] proceed rapidly for chlorine, so that in the tropics and middle latitudes ClO rarely exceeds 20% of the inorganic chlorine budget. The remainder of the budget consists primarily of HCl and ClNO_3 in roughly equal proportions, except at very high altitudes where HCl dominates. This partitioning is illustrated schematically in Figure 2(a).

At high latitudes in winter, where photolysis rates are small and particles are larger and more abundant than at lower latitudes, heterogeneous reactions can activate all of the inorganic chlorine into short-lived reservoirs that rapidly produce radicals. Under these low illumination conditions, reactions such as [IVa] and [IVb] proceed rapidly for weeks and months, destroying ozone at rates of a few percent per day. In regions where ozone production is very slow owing to the lack of short-wavelength ultraviolet light necessary to break the O_2 bond, significant losses of ozone occur. When solar illumination increases in springtime, ozone loss will cease if nitric acid is present to produce NO_2 , which rapidly ties up ClO into ClNO_3 . However, if nitric acid is removed (as occurs annually over the Antarctic and occasionally over the Arctic by sedimentation of cloud particles that contain nitric acid and water, called polar stratospheric clouds, or PSCs), ClNO_3 and HCl re-form at rates that are far too slow to avoid complete destruction of ozone. In such regions, measurements have identified a clear correlation of ozone loss with enhanced abundances of ClO.

More typically, ozone production and loss are in closer balance, and only a gradual year-by-year erosion of ozone has been detected as abundances of the halogen species increase. Consequently, the impact of halogens on stratospheric ozone at midlatitudes and in the tropics is assessed by long-term monitoring of as many chemical species as feasible; these observations are incorporated into detailed photochemical models of the stratosphere for interpretation. Such studies indicate that trends in industrial chlorine and bromine compounds can account for at least half of the downward trends in column ozone abundances that have been observed over the past several decades. A great deal of international cooperation was necessary to formulate regulations (e.g., the Montreal Protocol) that have only just recently begun to impact

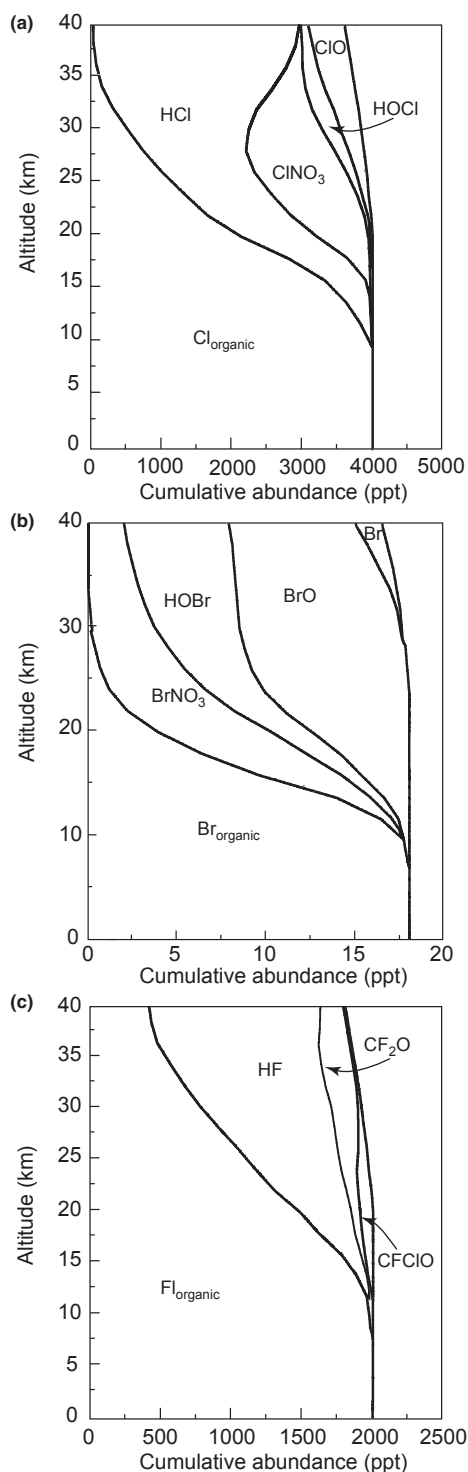


Figure 2 Midlatitude vertical profiles of the cumulative partitioning of (a) chlorine, (b) bromine, and (c) fluorine in the Earth's atmosphere from the ground to 40 km. Mixing ratios are typical of values in the 1990s. The slight falloff of total abundance with altitude in the middle to upper stratosphere reflects the lag time for air to reach these altitudes and the upward trends in source gas abundances in the 1990s. The partitioning of inorganic iodine is ignored because it is too uncertain at the present time.

the abundances of ozone-destroying forms of chlorine in the stratosphere.

Bromine

There are several sources of stratospheric inorganic bromine, including the halons and methyl bromide, the latter a compound that originates from both natural and industrial processes. Three inorganic bromine species, BrO, HBr, and BrNO₃, have been quantified in the stratosphere. BrO and BrNO₃ dominate the inorganic bromine budget. The abundances of the remaining species have been deduced by photochemical models and through the response of BrO to changing abundances of compounds with which it reacts. The partitioning of inorganic bromine is illustrated schematically in Figure 2(b).

It is believed that the sum of the abundances of all bromine-containing species in the stratosphere is around 20 parts per trillion (ppt), which is about a factor of 20 smaller than the sum of all the chlorine-containing species. However, the catalytic cycles involving bromine free radicals proceed much faster than their chlorine counterparts, and the percentage of bromine in free radical forms is larger than that of chlorine. Therefore, ozone destruction due to bromine is almost as significant as that due to chlorine, especially in the winter polar regions and in the lowermost stratosphere. However, because the natural sources of methyl bromide are not well characterized, and because there are still some important uncertainties in kinetics parameters, it is more difficult to attribute the anthropogenic contribution of bromine-catalyzed ozone losses. Consequently, international regulations for the industrial bromine compounds have taken longer to formulate than those for the chlorine compounds.

There have been some important developments in the study of atmospheric bromine based on new observations at the turn of the century that call into question some assumptions about the stratospheric bromine budget. First, by deploying grab samplers in the upper tropical troposphere, investigators have detected small, but significant, abundances of bromine-containing organic species that have fairly short lifetimes (days) in the troposphere. Presumably, these compounds are lofted to the upper troposphere by strong convective systems. Because abundances of these compounds vary widely at the surface, their contribution to the atmospheric bromine budget is hard to ascertain. These compounds have both natural and industrial origins, which further complicates assessments of anthropogenic ozone losses due to bromine. Second, observations have shown that large enhancements of bromine radicals occur sporadically near the Earth's surface in the polar regions by an uncertain, but probably heterogeneous, mechanism. It is possible that a similar mechanism operating in the lowermost stratosphere could alter the present understanding of bromine photochemistry. Third, elemental bromine has been detected in particles near the tropopause, raising questions about the sources and sinks of atmospheric bromine. New measurements in the atmosphere and laboratory will continue to improve our understanding of the impact of bromine on stratospheric ozone.

Iodine

Of all the halogens in the stratosphere, least is known about the iodine family. Laboratory measurements indicate that any organic iodine that is transported to the lower stratosphere will very quickly oxidize and release constituent atoms, and that these atoms will be even more destructive to ozone than chlorine and bromine. However, tropospheric measurements of potential source compounds suggest that the abundance of iodine in the stratosphere is on the order of 1 ppt or less, such that the iodine radicals will be at least an order of magnitude or smaller in abundance than BrO. Initial attempts to observe IO in the stratosphere have had mixed success, but generally indicate that there is no more than about 1 ppt of IO. However, little is known about the reactions that partition iodine into its various forms. It is likely that abundances of this species are highly variable.

There have been no attempts to measure other gas-phase inorganic iodine species whose concentrations are well below the detection limits of most modern instruments. However, the recent detection of elemental iodine in particles near the tropopause, an observation that is similar to the detection of elemental bromine, raises additional questions about the processes that control abundances of iodine in the atmosphere.

Even at low abundances, iodine could play an important role in ozone destruction in the lowermost stratosphere through its synergistic interactions with the bromine and chlorine free radicals. It has also recently been proposed that iodine could destroy ozone by reactions that involve OIO. Clearly, there is much to be learned about the role of iodine in the lower stratosphere.

Fluorine

Inorganic fluorine is produced by the photodecomposition of fluorocarbons, predominantly the chlorofluorocarbons (CFCs) CFCl_3 and CF_2Cl_2 . Measurements show that HF and the photodecomposition intermediates CF_2O and CFCIO can account for the entire inorganic fluorine budget, in agreement with models that incorporate laboratory measurements of fluorine reactions. Fluorine atoms react rapidly with hydrogen-containing species, especially H_2O and CH_4 that are present at part per million abundances in the stratosphere. In addition, there are no known ways to release fluorine atoms from HF, a very thermodynamically stable species. Consequently, immeasurably small abundances of fluorine radicals are present as extremely short-lived intermediates in the photodecomposition of fluorocarbons, and their contribution to ozone loss is negligibly small. Therefore, the primary role of fluorine in stratospheric chemistry is as a marker or tracer for other halogen species, in particular the CFCs. Ground-based and satellite measurements have shown that the rate of increase of HF in the stratosphere can be explained entirely by the buildup of CFCs in the troposphere followed by gradual transport into the stratosphere where they photodecompose in the presence of UV and chemical oxidants (primarily OH and $\text{O}(^1\text{D})$). The partitioning of inorganic fluorine is illustrated schematically in Figure 2(c).

Halogens and Ozone Loss

There are three distinct regimes in the stratosphere in which ozone losses have been attributed to halogens. These are the middle stratosphere year-round, in the Antarctic and Arctic polar vortices in springtime, and in rocket plumes within hours following launches. In all cases, two conditions are met that establish the link between halogens and ozone loss. First, the regions where ozone losses are detected are correlated with larger abundances of halogen radicals than in adjacent regions where there is less or no ozone loss. Second, the rates at which ozone losses occur are equivalent (within measurement uncertainties) to the corresponding rates predicted with models that include laboratory measurements of the rate constants for the rate-determining reactions. Examples of the anticorrelation between abundances of chlorine oxide and ozone are shown in Figure 3. The sources of enhancements in the chlorine radicals differ (direct local injection in the case of rockets, and heterogeneous reactions of HCl and ClNO_3 in the case of the Antarctic ozone hole). Consequently, the time-scales for ozone loss in these two cases are dramatically different, less than 1 h for the rocket plume and a month for the Antarctic ozone hole; however, in both cases the amounts of ozone destroyed over these periods are consistent with the known kinetics of the halogen radicals to within the uncertainties of the measurements.

In the two cases presented in Figure 3, the ozone destruction rates vastly exceeded the rates at which ozone could be replenished. Thus, regions of low ozone formed adjacent to regions of higher ozone where the abundances of the halogen free radicals were significantly lower. In the middle stratosphere, the situation is quite different, and ozone production and loss rates nearly match (i.e., ozone is in photochemical steady state). In addition, the spatial variability of the halogen radicals is small. Consequently, it is difficult to attribute an instantaneous ozone value to a particular abundance of halogen radicals. Rather, it is by correlating the long-term downward trend of ozone abundances with concomitant increases in halogen radical abundances that the link is deduced. Observations for the last 20 years of the twentieth century showed a decrease in ozone of approximately 10–15% between 35 and 50 km, an amount that agrees well with the decrease predicted as a result of the steady rise in abundances of stratospheric chlorine, the primary agent of halogen-induced ozone loss in the middle stratosphere.

Long-term reductions in ozone have been reported for other regions of the stratosphere, in particular the Arctic and the lowermost stratosphere at middle latitudes. In the first case, springtime ozone losses are consistent with calculations based on observed abundances of the radicals ClO and BrO. However, losses in the middle of winter appear to be significantly greater than expected, for reasons that are not yet clear. It is possible that transport between regions of differing ozone concentrations confounds efforts to attribute ozone loss to specific halogen radical abundances. In the midlatitude lower stratosphere, ozone losses are highly uncertain because they occur in a region where there is a strong vertical gradient in the ozone mixing ratio and because measurement techniques are not optimized for these altitudes. This is a region of the stratosphere where temperatures are very low ($\sim 200\text{ K}$) and

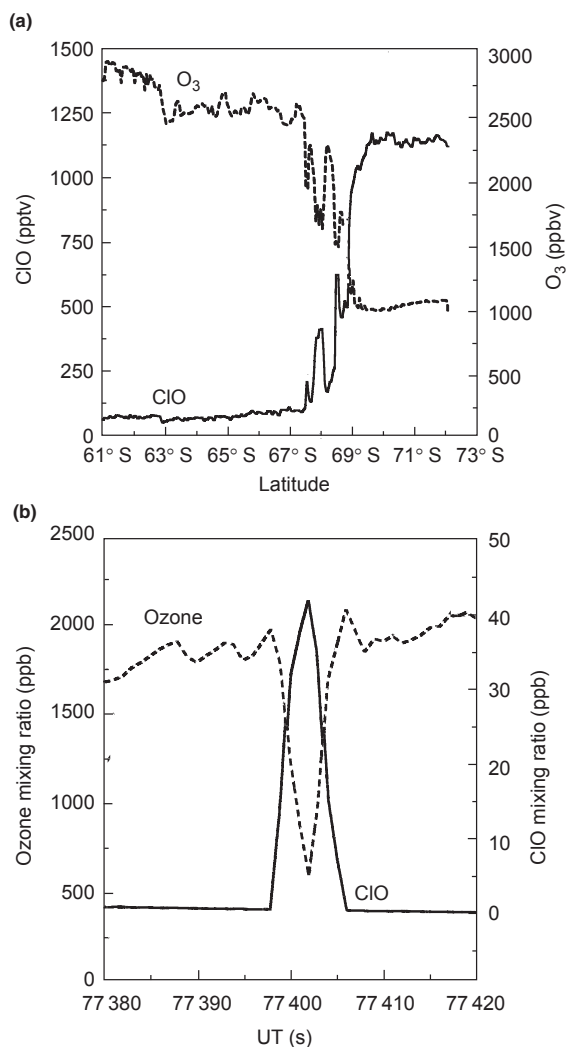


Figure 3 Examples of the measured correlation of ozone loss with ClO radical abundances in (a) the Antarctic ozone hole in 1987 (September 16) (excerpted with permission from Anderson, J.G., Toohey, D.W., Brune, W.H., 1991. Free radicals within the Antarctic vortex: the role of CFCs in Antarctic ozone loss. *Science* 251, 39–46. Copyright 1991 American Association for the Advancement of Science.) and (b) in a plume of a Delta II rocket. Adapted from Ross, M.N., et al., 2000. Observation of stratospheric ozone depletion associated with Delta II rocket emissions. *Geophysical Research Letters* 27, 2209–2212.

heterogeneous reactions on sulfuric acid aerosols or thin cirrus clouds could enhance halogen radical abundances. *In situ* observations in these regions have shown that chlorine radicals can be significantly enhanced to amounts that are sufficient to destroy ozone. However, the observations have been too infrequent and limited in geographical extent to link chlorine and bromine chemistry to ozone trends that have been inferred in the lowermost stratosphere. A renewed attention to halogen chemistry in the lowermost stratosphere promises to be very revealing.

See also: **Chemistry of the Atmosphere:** Chemical Kinetics; Principles of Chemical Change. **Middle Atmosphere:** Polar Vortex. **Numerical Models:** Chemistry Models. **Observations Platforms:** Balloons; Buoys; Rockets. **Ozone Depletion and Related Topics:** Ozone Depletion Potentials; Photochemistry of Ozone. **Stratospheric Chemistry Topics:** HO_x; Halogen Sources, Natural (Methyl Bromide and Related Gases); Overview.

Further Reading

- Anderson, J.G., Toohey, D.W., Brune, W.H., 1991. Free radicals within the Antarctic vortex: the role of CFCs in Antarctic ozone loss. *Science* 251, 39–46.
- Brune, W.H., 1998. Stratospheric chemistry – perspectives in environmental chemistry. In: Macalady, D.L. (Ed.), *Perspectives in Environmental Chemistry*. Oxford University Press, Oxford, pp. 292–324.
- Finlayson-Pitts, B.J., Pitts, J.N., 2000. *Chemistry of the Upper and Lower Atmosphere*. Academic Press, London.
- Halloway, A.M., Wayne, R.P., 2010. *Atmospheric Chemistry*. The Royal Society of Chemistry, Cambridge.
- Roan, S.L., 1989. *Ozone Crisis: The 15-Year Evolution of a Sudden Global Emergency*. Wiley, New York.
- Ross, M.N., et al., 2000. Observation of stratospheric ozone depletion associated with Delta II rocket emissions. *Geophysical Research Letters* 27, 2209–2212.
- Thornton, et al., 2007. Chlorine activation near the mid-latitude tropopause. *Journal of Geophysical Research: Atmospheres* 112, D18306. <http://dx.doi.org/10.1029/2006JD007640>.

Halogen Sources, Anthropogenic

A McCulloch, University of Bristol, Bristol, UK
PM Midgley, M & D Consulting, Leinfelden Musberg, Germany

© 2003 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 5, pp 2154–2162, © 2003, Elsevier Ltd.

Introduction

None of the anthropogenic carriers of halogens in the stratosphere is actually released there. They are emitted close to ground level and have to survive transport through the troposphere, requiring a lifetime in the atmosphere of at least a year. Thus soluble halogen-containing materials, such as hydrogen chloride (HCl), which are rained out of the atmosphere in a matter of days, do not provide a significant halogen input into the stratosphere; neither do the more reactive materials, such as trichloroethene or tetrafluoroethene, which are oxidized in the lower atmosphere within a similar time scale.

The bulk of the halogen input into the stratosphere is from anthropogenic gases that have atmospheric lifetimes significantly longer than 2 years. These are released from industrialized regions, principally in the Northern Hemisphere.

Chlorofluorocarbons (CFCs), with atmospheric lifetimes of 45–1700 years, were introduced in the 1930s as refrigerants that were safer than the toxic and flammable materials then used. Despite the fact that small amounts of CFCs have been measured in volcanic vents, the natural contribution is negligible compared with man-made sources. Carbon tetrachloride, which has an atmospheric lifetime of 35 years, had been used as a solvent until the middle of the twentieth century; subsequently it was mainly used as a raw material for CFC manufacture, and emissions into the atmosphere grew with CFC production. Hydrochlorofluorocarbons (HCFCs), with lifetimes between 1.4 and 19 years, were introduced in the 1940s for deep freezing applications otherwise served by ammonia. More recently, the HCFCs have become partial replacements for CFCs. Halons, fire-extinguishing chemicals with lifetimes of 11–65 years and containing bromine, were introduced in the 1960s. At the same time the use of methylchloroform (1,1,1-trichloroethane, atmospheric lifetime 4.8 years) as a precision cleaning solvent was expanding rapidly. Together with methyl chloride and methyl bromide, which have significant natural fluxes, these carry potentially reactive halogens (chlorine and bromine) into the stratosphere and, with the exception of methyl chloride, all are ozone-depleting substances controlled by the Montreal Protocol.

The history of anthropogenic emissions and the resulting atmospheric concentration is described here, together with the consequential rise of chlorine and bromine in the troposphere. CFCs, carbon tetrachloride, and most of the halons are removed from the atmosphere only by photolysis in the stratosphere, hence their relatively long atmospheric lifetimes. HCFCs, methylchloroform and methyl halides are oxidized in the troposphere and generally have shorter atmospheric lifetimes, but for all compounds the average time delay between

release in the lower atmosphere and decomposition in the ozone layer to generate stratospheric halogen is 3 years. Furthermore, the relative effectiveness in ozone depletion of each compound varies with the altitude at which its halogen is released, and this, together with the time delay, is taken into account when calculating the total Equivalent Effective Stratospheric Chlorine (EESC), which is a measure of the combined effect of all ozone-depleting substances.

Fluorine does not react with stratospheric ozone. Consequently, the hydrofluorocarbons (HFCs) that are designed to replace CFCs are not controlled under the Montreal Protocol. There is already a significant stratospheric fluorine concentration arising from decomposition of CFCs and this is starting to be augmented by HFCs, which have lifetimes between 1.5 and 240 years. The extent of this and the concentrations of the much less reactive perfluorocarbons (PFCs) are also discussed.

The Chlorine Flux

Chlorofluorocarbons

Most of the anthropogenic chlorine content of the atmosphere is a consequence of CFC emissions. Historically, the largest single source was aerosol spray cans from which the CFC propellant was released immediately the can was used. Currently, the principal release into the atmosphere is from a declining stock of CFC contained in refrigeration and air-conditioning systems and foam insulation; in these applications, release of the substance occurs some time after it was manufactured. The delay is variable. Automobile air conditioners can release all of their contents in a matter of a few years; on the other hand, a domestic refrigerator has a typical service life of 20 years and the CFCs it contains leak only when it is dismantled. Insulating foam can be in use for much longer, with only slow release until (and if) the foam is crushed. Uncertainties in estimates of the delay between manufacture and emission into the atmosphere, characteristic of such uses, contribute significantly to the uncertainty of the estimated emission.

For almost 30 years the manufacturers of CFCs have organized an annual collection of audited industrial production and sales data for CFC-11 and CFC-12. Historical production and sales records were also extracted by the manufacturers for the period back to first production in 1931 and the combined data provide the basis for calculation of emissions of these compounds. Annual emissions are estimated for each major category of application based on the quantities used, coupled with emission functions that take account of the rates of release of the materials during actual use and disposal (which are specific to the application). The survey procedure and

emissions estimation have been extended to most of the industrial halocarbons:

- CFC-11 (trichlorofluoromethane), principally used in aerosols and foam insulation
- CFC-12 (dichlorodifluoromethane), principally used in aerosols and refrigeration
- CFC-113 (1,1,2-trichlorotrifluoroethane), a solvent
- CFC-114 (1,1,2,2-tetrafluorodichloroethane), principally used in aerosols and refrigeration
- CFC-115 (chloropentafluoroethane), a refrigerant.

In much of the world, CFC production was carried out by subsidiaries of companies that reported their production and sales into the database and up to 15 years ago the only producing country not included was the USSR. Since then India, China, and Korea have become significant producers, although they too do not report into the industrial database. However, national aggregate CFC production now has to be reported to the Secretariat of the Montreal Protocol by all parties. The estimated historical quantities released, shown in **Table 1**, are based on a composite global estimate of annual production from the industrial and legislative databases.

For each sales category a characteristic pattern of emissions in time was established by market surveys carried out by the producers. This enabled estimates of annual emissions as

outlined above. These have been the subject of a sensitivity study that confirmed that the largest contributions to the uncertainties came from the fraction of production that was not reported in the industrial data and the rate of release of materials from closed-cell foams. The first of these has been addressed using the database from the Montreal Protocol, which now matches the industrial data reliably (to within 1% over the same set of countries). The second is a particular problem for CFC-11, where the range of emissions resulting from the lowest credible estimate of the release from closed-cell foams to the highest is 13.1%. For the period up to 1992, a mid-range estimate was used in **Table 1**. In recent years, as the degree of containment of CFCs in systems has improved, the historical emissions functions have tended to overestimate releases. This was allowed for in the estimates developed for recent Scientific Assessments and, from 1992 onwards, it is those values that are shown in **Table 1**.

In all cases the release estimates show substantial falls during the 1990s. The fall in consumption has actually been faster than that required under the Montreal Protocol; nevertheless, large quantities of CFCs remain in systems and may be released in the future: for example, over 700 Gg of CFC-11 and 250 Gg of CFC-12 are currently unreleased.

Figure 1 shows how the atmospheric concentrations of CFCs have grown. These were calculated using a simple single-box

Table 1 Emissions of CFCs (Gg y⁻¹)

Year	CFC-11 (CFC1 ₃ ; 45 y) ^a	CFC-12 (CF ₂ Cl ₂ ; 100 y)	CFC-113 (CF ₂ ClCFCl ₂ ; 85 y)	CFC-114 (CF ₂ ClCF ₂ Cl; 300 y)	CFC-115 (C ₂ F ₅ Cl; 1700 y)
1970	209.0	313.8	28.0	9.7	1.3
1971	229.8	338.0	32.1	10.1	1.6
1972	259.2	368.9	36.9	10.5	1.9
1973	296.6	408.7	42.2	10.9	2.2
1974	327.1	444.3	48.4	11.3	2.5
1975	318.4	435.5	55.5	11.7	3.0
1976	325.9	425.2	63.5	12.2	3.5
1977	314.0	406.1	72.8	12.7	4.0
1978	294.9	376.7	83.4	13.2	4.7
1979	276.1	375.7	95.5	13.7	5.4
1980	264.6	373.3	109.4	14.2	6.3
1981	263.8	385.3	119.4	14.2	7.2
1982	257.2	385.3	124.6	13.7	8.1
1983	273.8	394.4	138.3	14.1	8.9
1984	295.6	413.6	171.1	15.1	9.6
1985	308.6	426.1	201.7	16.2	10.1
1986	326.8	437.5	216.6	18.0	10.6
1987	345.7	451.0	236.4	18.2	11.0
1988	353.5	462.7	260.3	16.2	11.4
1989	304.7	436.9	271.6	14.5	11.9
1990	211.4	378.8	233.8	10.3	12.2
1991	213.2	335.7	181.5	6.3	12.6
1992	168.9	319.9	147.5	5.2	12.6
1993	146.8	302.8	80.5	4.6	12.6
1994	101.1	239.9	52.0	4.0	11.8
1995	100.7	239.3	43.2	3.1	10.9
1996	101.7	220.4	27.0	2.4	9.5
1997	99.2	185.0	9.5	2.3	7.8
1998	97.2	155.5	5.4	2.7	6.0

^aThe formula and atmospheric lifetime in years are given in parentheses.

Sources: AFEAS (Alternative Fluorocarbons Environmental Acceptability Study) (2000), Production, Sales and Atmospheric Release of Fluorocarbons through 1998. Washington, DC: AFEAS. Midgley PM and McCulloch A (1999) Production, sales and emissions of halocarbons from industrial sources. In: Fabian P and Singh ON (Eds) The Handbook of Environmental Chemistry, vol 4. Part E, Reactive Halogen Compounds in the Atmosphere, pp. 157–190. Heidelberg: Springer-Verlag.

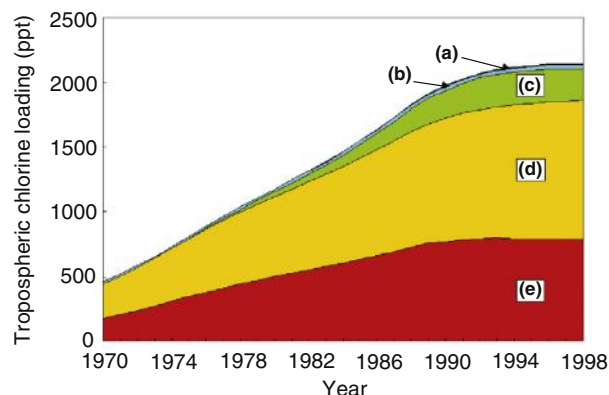


Figure 1 Contributions to tropospheric chlorine loading from chloro-fluorocarbons: (a) chloropentafluoroethane (CFC-115); (b) dichlorotetrafluoroethane (CFC-114); (c) trichlorotrifluoroethane (CFC-113); (d) dichlorodifluoromethane (CFC-12); (e) trichlorofluoromethane (CFC-11).

model of the atmosphere and current estimates of atmospheric lifetimes, according to eqn [1], where C_0 and C_y are the atmospheric concentrations in the starting year and in year y , S is the annual rate of release of the substance, and T is its atmospheric lifetime.

$$C_y = ST + (C_0 - ST) \exp(-y/T) \quad [1]$$

The units are parts per trillion (ppt, 1 in 10^{12}) of tropospheric chlorine loading, which is the calculated concentration of each CFC multiplied by the number of atoms of chlorine in its molecule. Thus, for CFC-11 (fluorotrichloromethane), the CFC concentration is multiplied by 3.

While the growth in chlorine loading arising from CFC-12 emissions has slowed in recent years, it is still the largest of the CFC contributors and its absolute concentration is still growing. The concentrations of CFC-11 and CFC-113 have fallen discernibly, and those from CFC-114 and CFC-115 are not large enough to matter. Overall, the CFC contribution to chlorine loading is now level in time.

Chlorocarbons

The next largest contribution to the chlorine loading of the atmosphere comes from carbon tetrachloride (CCl_4 , tetrachloromethane) and methylchloroform (CH_3CCl_3 , 1,1,1-trichloroethane). Estimates of their emissions are shown in Table 2.

Carbon tetrachloride is hepatotoxic at relatively low concentrations and so has not been used as a solvent in developed countries for many years. Its principal use is as raw material for the manufacture of CFC-11 and CFC-12 and it is thought that the accumulation in the atmosphere now results solely from process losses. It has not been possible to quantify these losses in the same way as for the CFC releases, consequently emissions into the atmosphere have been calculated from the change in atmospheric concentration over the period 1979 to 1996, using an inverted form of eqn [1].

For methyl chloroform, an audited production and sales database has been maintained from information collected by the producers in much the same way as for CFCs. With the

Table 2 Emissions of chlorocarbons (Gg y^{-1})

Year	Carbon tetrachloride (CCl_4 , 35 y) ^a	Methyl chloroform (CH_3CCl_3 , 4.8 y)
1970	69	133
1971	74	170
1972	84	214
1973	94	266
1974	99	305
1975	85	309
1976	92	382
1977	86	462
1978	83	513
1979	97	511
1980	95	538
1981	78	549
1982	78	523
1983	92	536
1984	74	585
1985	82	593
1986	100	602
1987	91	623
1988	89	666
1989	72	691
1990	64	718
1991	36	635
1992	44	593
1993	45	380
1994	36	283
1995	32	234
1996	10	84
1997	8	30
1998	8	16

^aThe formula and atmospheric lifetime in years are given in parentheses.

Sources: Midgley PM and McCulloch A (1999) Production, sales and emissions of halocarbons from industrial sources. In: Fabian P and Singh ON (eds) The Handbook of Environmental Chemistry, vol 4. Part E, Reactive Halogen Compounds in the Atmosphere, pp. 157–190. Heidelberg: Springer-Verlag. Simmonds PG et al. (1998) Global trends and emission estimates of CCl_4 from in situ background observations from July 1978 to June 1996. Journal of Geophysical Research, 103: 16017–16027.

exception of use as a chemical intermediate (in which it is totally converted and not released), methyl chloroform was used as an industrial solvent, with total emission into the atmosphere. Hence the emission function is relatively simple; 3/4 of annual sales are estimated to be emitted in that year and 1/4 in the following year. Long-term storage, over one or two years, was accommodated by a linear displacement of emissions in time. Prompt emissions, coupled with a relatively short atmospheric lifetime, have meant that the concentration of methyl chloroform shows the sharpest fall as a consequence of the Montreal Protocol. Figure 2 shows the contributions to chlorine loading from the individual chlorocarbons superimposed on that from the CFCs.

Hydrochlorofluorocarbons

Despite their potential to replace CFCs, hydrochlorofluorocarbons (HCFCs) have relatively little impact on atmospheric chlorine loading. The principal member of this group of substances, chlorodifluoromethane (HCFC-22), has been used as a refrigerant fluid since 1946; its low boiling point

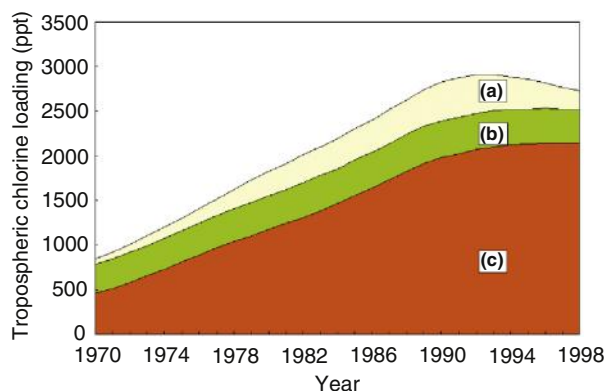


Figure 2 Contributions to tropospheric chlorine loading from CFCs and chlorohydrocarbons: (a) methyl chloroform (1,1,1-trichloroethane); (b) carbon tetrachloride (tetrachloromethane); (c) all CFCs combined, as shown in Figure 1.

makes it suitable for low-temperature duties and some air-conditioning. As shown in Table 3, HCFC-22 emissions have grown to about 250 Gg y^{-1} and are now stable. Emissions of the other HCFCs are one or two orders of magnitude lower. HCFC-124 (1,1,1,2-tetrafluorochloroethane), introduced comparatively recently, is an aerosol propellant and refrigerant

fluid that is produced in modest amounts. HCFC-141b (1,1-dichloro-1-fluoroethane), again a relative newcomer, is produced in much larger quantities for use either as a blowing agent for rigid plastic foams, such as those used for insulation, or as a solvent. HCFC-142b (1-chloro-1,1-difluoroethane) is also used to blow plastic foam. Over 94% of all HCFC production is in the developed world.

HCFCs were considered to be suitable temporary replacements for CFCs because of their low intrinsic potential to impact the ozone layer. In general, they contain less chlorine than CFCs, have shorter atmospheric lifetimes, so that they do not accumulate in the atmosphere to the same extent as CFCs, and are not photolyzed as effectively in the stratosphere, so that the chlorine they contain is not released directly into the ozone layer. Nevertheless, HCFCs are ozone-depleting substances and are to be phased out under the Montreal Protocol by 2020 in the developed world and 2040 elsewhere. The contribution of HCFCs to chlorine loading is shown in Figure 3; no allowance has been made for the relative effectiveness of their chlorine content.

Natural Source of Chlorine

Although the quantities released by human activities are small, methyl chloride (CH_3Cl , chloromethane) is produced

Table 3 Emissions of HCFCs (Gg y^{-1})

Year	HCFC-22 (CHF_2Cl , 11.8 y) ^a	HCFC-124 (CF_3CHFCl , 6.1 y)	HCFC-141b (CH_3CFCl_2 , 9.2 y)	HCFC-142b ($\text{CH}_3\text{CF}_2\text{Cl}$, 18.5 y)
1970	45.5			
1971	50.4			
1972	55.1			
1973	63.2			
1974	71.5			
1975	72.3			
1976	81.4			
1977	90.6			
1978	99.1			
1979	106.4			
1980	114.7			
1981	121.3			0.6
1982	123.1			0.5
1983	135.1			0.5
1984	145.0			0.6
1985	150.9			0.4
1986	160.2			1.6
1987	169.1			1.7
1988	188.5			2.0
1989	207.6			2.9
1990	214.7			5.8
1991	224.7		0.4	8.4
1992	236.3	0.1	3.9	10.8
1993	234.0	0.3	13.1	10.2
1994	240.8	0.4	24.8	10.7
1995	247.2	1.8	36.6	12.0
1996	266.2	3.5	39.5	11.7
1997	245.9	3.7	42.7	11.6
1998	255.8	3.2	49.8	10.6

^aThe formula and atmospheric lifetime in years are given in parentheses.

Sources: AFEAS (Alternative Fluorocarbons Environmental Acceptability Study) (2000) Production, Sales and Atmospheric Release of Fluorocarbons through 1998. Washington, DC: AFEAS. WMO (World Meteorological Organization) (1999) Scientific Assessment of Ozone Depletion: 1998, WMO Global Ozone Research and Monitoring Project Report No. 44. Geneva: WMO.

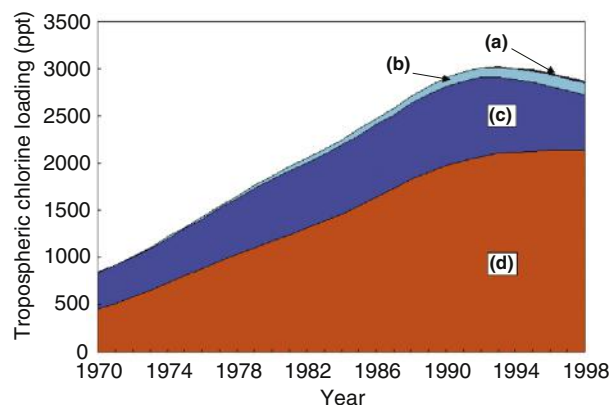


Figure 3 All contributions to tropospheric chlorine loading from (a) the combined concentrations of HCFC-124 (1,1,1,2-tetrafluorochloroethane), HCFC-141b (1,1-dichloro-1-fluoroethane), and HCFC-142b (1-chloro-1,1-difluoroethane); (b) HCFC-22 (chlorodifluoromethane); (c) all chlorohydrocarbons (as in Figure 2); (d) all CFCs (as in Figure 1).

by natural processes in sufficient amounts to contribute significantly to stratospheric chlorine. The lifetime of methyl chloride is only 1.3 years. However the flux of 4 Tg y^{-1} , mainly from the oceans, biomass burning, and terrestrial fungi, is large enough to maintain an atmospheric concentration of 550 ppt.

The Bromine Flux

Halons

The natural contribution to bromine in the stratosphere is similar to that from anthropogenic sources; of the total bromine loading of about 17 ppt, 9 ppt is attributable to man's activities and most of this comes from halon emissions. Halons were first produced as fire-extinguishing agents in 1963 and their use expanded to almost 20 Tg y^{-1} by the mid 1990s. Two substances predominated; Halon-1211 (bromochlorodifluoromethane), used mainly in portable extinguishers, and Halon-1301 (bromotrifluoromethane), an agent used in fixed systems. In addition, Halon-2402 (1,2-dibromotetrafluoroethane) was produced in somewhat smaller quantities and used in Eastern Europe. Halon-1202 (dibromodifluoromethane) has also been detected in small, but growing, amounts in the atmosphere.

Bromine is 60 times more potent in ozone depletion than chlorine in the current background stratospheric composition. This was recognized in the Montreal Protocol and halon production was phased out earlier than CFCs in the developed world (in 1994). However, production of Halon-1211 and Halon-1301 will continue in China, India, and Korea for the next few years and Russia has dispensation to continue the manufacture of Halon-2402.

In much the same way as for CFCs, audited production statistics are available from industry in the developed world and from the submissions required under the Montreal Protocol for the controlled halons, but the proportion of annual halon production that is unreleased is much higher than is the case for CFCs. Currently, halons should be released into the atmosphere only when they are used in earnest – on a fire or when the fire protection system is activated. Although

historically they were also released during training and system testing, there remains a considerable time delay between production and release and a stock of halon (known as the 'bank') has accumulated in systems and equipment. The emissions of Halon-1211 shown in Table 4 were based on a small part of the bank (currently 12%) being emitted each year. In the case of Halon-1301, calculated similarly, the emission factor is now 4% of the bank each year.

Production data for Halon-2402 do not exist in the same form and so the values for emissions in Table 4 were calculated by inverse modeling from measured atmospheric concentrations, using eqn [1]. The source of Halon-1202 has yet to be identified, although it is a well known by-product of the manufacture of Halon-1211. The emissions shown in Table 4 were calculated by inverse modeling.

Methyl Bromide

Methyl bromide contributes a total of 10 ppt to bromine loading. Of this, only about 1.9 ppt arises from human activities that are controlled under the Montreal Protocol, principally use of manufactured material for pest control in growing and harvested agricultural produce. Minor other anthropogenic sources that are not controlled add a further 0.4 ppt into the atmosphere; these include the exhausts of motor vehicles running on leaded gasoline and also chemical process emissions.

Although there is much uncertainty, the bulk of methyl bromide entering the atmosphere seems to come from natural processes. The role of the oceans is particularly difficult to untangle because they act as both sources and sinks. Methyl bromide is released into the atmosphere particularly from the polar oceans and is absorbed from the atmosphere into tropical waters where it is destroyed by bacteria.

A number of other bromine compounds are produced naturally: dibromomethane, bromochloromethane and dibromochloromethane together can add up to 6 ppt to bromine loading at ground level, particularly in the Arctic. However, these are very short-lived species and are not considered normally to be transported into the stratosphere.

Figure 4 shows the increase in anthropogenic bromine loading since 1970, subdivided into contributions from individual compounds. In the absence of better information, the contribution from methyl bromide has been shown as constant.

The Fluorine Flux

Neither F nor CF_3 radicals, nor their oxygenated derivatives, interact with stratospheric ozone; fluorine released into the stratosphere is converted into hydrogen fluoride (HF), which does not react further and is eventually removed when the stratospheric air circulates into the troposphere. However, it is a significant component of the stratospheric halogen budget. In much the same way as for chlorine and bromine, fluorine loading of the troposphere may be calculated from the atmospheric concentrations of CFCs, HCFCs, and halons, with the results shown in Figure 5. The contribution from hydrofluorocarbons (HFCs) is currently small but is increasing at

Table 4 Emissions of halons (Gg y^{-1})

Year	Halon-1211 (CF_2ClBr , 11 y) ^a	Halon-1301 (CF_3Br , 65 y)	Halon-2402 ($\text{CF}_2\text{BrCF}_2\text{Br}$, 25 y)	Halon-1202 (CF_2Br_2 , 3y)
1970	0.29	0.05	0.20	0.04
1971	0.41	0.11	0.25	0.04
1972	0.56	0.19	0.31	0.06
1973	0.74	0.30	0.39	0.07
1974	0.95	0.41	0.47	0.08
1975	1.18	0.57	0.54	0.10
1976	1.49	0.86	0.62	0.11
1977	1.74	1.10	0.70	0.13
1978	2.08	1.37	0.77	0.15
1979	2.44	1.70	0.85	0.16
1980	2.84	1.98	0.93	0.19
1981	3.18	2.35	1.00	0.21
1982	3.60	2.90	1.11	0.23
1983	4.18	3.27	1.18	0.25
1984	4.96	3.81	1.26	0.27
1985	5.96	4.39	1.36	0.29
1986	7.10	5.02	1.43	0.31
1987	8.41	5.61	1.50	0.33
1988	9.82	6.25	1.73	0.35
1989	10.08	6.01	1.73	0.37
1990	10.05	5.62	1.73	0.39
1991	11.49	3.56	1.74	0.41
1992	11.99	3.61	1.72	0.43
1993	12.49	1.16	1.70	0.51
1994	12.99	4.66	1.68	0.59
1995	13.49	3.47	1.30	0.67
1996	11.12	2.77	0.85	0.73
1997	11.46	2.80	0.70	0.79
1998	11.12	2.71	No data	No data

^aThe formula and atmospheric lifetime in years are given in parentheses.

Sources: Midgley PM and McCulloch A (1999) Production, sales and emissions of halocarbons from industrial sources. In: Fabian P and Singh ON (eds) The Handbook of Environmental Chemistry, vol 4. Part E, Reactive Halogen Compounds in the Atmosphere, pp. 157–190. Heidelberg: Springer-Verlag. Fraser PJ et al. (1999) Southern Hemispheric halon trends (1978–1998) and global halon emissions. Journal of Geophysical Research 104: 15985–15999.

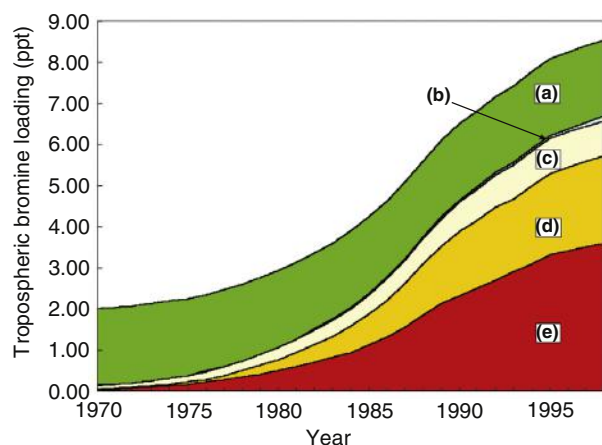


Figure 4 Contributions to tropospheric bromine loading from (a) that part of methyl bromide emissions that is controlled under the Montreal Protocol; (b) Halon-1202 (dibromodifluoromethane); (c) Halon-2402 (1,2-dibromotetrafluoroethane); (d) Halon-1301 (bromotrifluoromethane); (e) Halon-1211 (bromochlorodifluoromethane).

a significant rate. This is largely a consequence of releases of trifluoromethane (HFC-23, fluoroform), which is a by-product of the manufacture of HCFC-22, has a long atmospheric lifetime, and is decomposed in the stratosphere, so adding to the

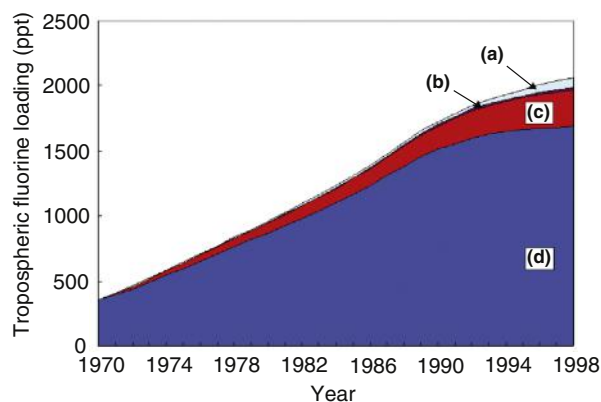


Figure 5 Contributions to tropospheric fluorine loading from (a) all halons; (b) all HCFCs; (c) all CFCs.

fluorine burden there. More recently this has also been augmented by releases of HFC-134a (1,1,1,2-tetrafluoroethane, $\text{CF}_3\text{CH}_2\text{F}$), which is manufactured for use as a refrigerant and now has a tropospheric concentration of 9 ppt.

Other fluorine-containing substances do not contribute significantly to fluorine loading either because the quantities released are currently too small to matter (the case with hydrofluorocarbons other than HFC-23 and HFC-134a) or

because they are so inert that they do not decompose to release fluorine in the stratosphere (the case with perfluorocarbons and sulfur hexafluoride).

Perfluorocarbons, in particular those that are formed as by-products of primary aluminum production, are much more abundant than hydrofluorocarbons. Tetrafluoromethane (CF_4 , PFC-14) has now reached a concentration of 80 ppt, half of which is due to aluminum production. The other 40 ppt is volcanic in origin and has accumulated in the atmosphere over many thousands of years. Hexafluoroethane (C_2F_6 , PFC-116), another aluminum by-product, has no natural source and its atmospheric concentration now stands at 3 ppt. These substances have atmospheric lifetimes over ten thousand years and are so inert that they do not contribute to the stratospheric loading of fluorine; indeed, the trend of their stratospheric concentrations with altitude is a good indicator of their historic tropospheric concentrations.

As for perfluorocarbons, the atmospheric lifetime of sulfur hexafluoride (SF_6) is long (3200 years) and it too does not contribute to the stratospheric loading of fluorine. Although there is a volcanic source, it is too small to be significant and most of the 4 ppt of sulfur hexafluoride that is now present in the atmosphere has been used in industrial applications such as electrical switchgear.

Equivalent Effective Stratospheric Chlorine and the Future

The concentrations so far discussed may be verified by direct measurement of the individual species in the troposphere, which is comparatively well mixed. Tropospheric loading describes the concentration of potentially active chlorine and bromine in the flux of air entering the stratosphere; it is not exactly equal to the loading of active halogen at the ozone layer. This is parameterized by equivalent effective stratospheric chlorine loading (EESC). To calculate the EESC, the tropospheric loadings of all compounds are adjusted by an overall factor to take account of the transport time between the troposphere and the stratospheric ozone layer and the contributions from individual chlorine- and bromine-containing compounds are adjusted by factors that accommodate their different effects on the ozone layer. The delay due to transport is set at 3 years. The effectiveness factor for the difference between chlorine and bromine is set at 60, as described above. The differences between individual chlorine compounds are much smaller than that; they range from 1.11 for HCFC-123 (CF_3CHCl_2) to 0.35 for HCFC-22 (CF_2HCl).

Figure 6 shows the way that EESC has developed in the past and the changes expected over the twenty-first century.

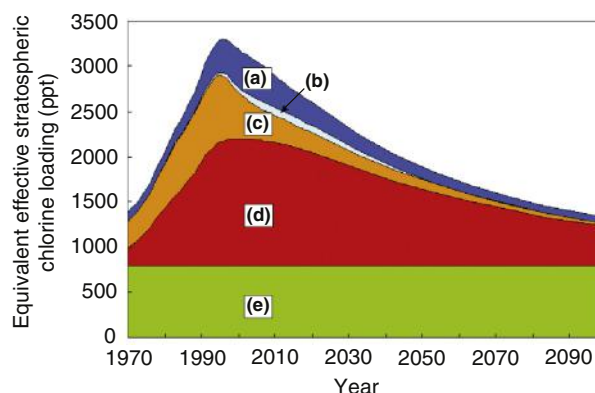


Figure 6 Contributions to the equivalent effective stratospheric halogen loading from (a) halons and controlled sources of methyl bromide; (b) all HCFCs; (c) controlled chlorocarbons; (d) all CFCs; and (e) methyl chloride and the natural (and other uncontrolled) sources of methyl bromide.

Although it is clear that the peak loading is past, it will take the whole of the twenty-first century for the stratospheric loading to return to 1970 levels and, all other things being equal, the return to loadings that predate the Antarctic ozone hole is expected to occur only towards the middle of the century.

See also: Ozone Depletion and Related Topics: Ozone Depletion Potentials; Photochemistry of Ozone. **Tropospheric Chemistry and Composition:** Oxidizing Capacity.

Further Reading

- AFEAS (Alternative Fluorocarbons Environmental Acceptability Study), 2001. Production, Sales and Atmospheric Release of Fluorocarbons through 2000. AFEAS, Washington, DC.
- Midgley, P.M., McCulloch, A., 1999. Production, sales and emissions of halocarbons from industrial sources. In: Fabian, P., Singh, O.N. (Eds.), 1999. The Handbook of Environmental Chemistry, vol. 4. Springer-Verlag, Heidelberg, pp. 157–190. part E, Reactive Halogen Compounds in the Atmosphere.
- SORG (United Kingdom Stratospheric Ozone Review Group), 1999. Stratospheric Ozone: 1999, Seventh Report of the UK SORG. Department of the Environment, Transport and the Regions, London.
- UNEP (United Nations Environment Programme), 1998. Production and Consumption of Ozone Depleting Substances 1986–1996. The Ozone Secretariat to the Vienna Convention and Montreal Protocol, Nairobi.
- WMO (World Meteorological Organization), 1999. Scientific Assessment of Ozone Depletion: 1998, WMO Global Ozone Research and Monitoring Project Report No. 44. WMO, Geneva.

Halogen Sources, Natural (Methyl Bromide and Related Gases)

S Yvon-Lewis, Texas A&M University, College Station, TX, USA

JH Butler, NOAA Earth System Research Laboratory, Boulder, CO, USA

© Published by Elsevier Ltd.

Synopsis

Anthropogenic halocarbons such as chlorofluorocarbons, CCl_4 , and CH_3CCl_3 are largely responsible for the observed depletion of stratospheric ozone, yet there is a contribution by gases that are both naturally and anthropogenically produced. Of the naturally produced halocarbons, CH_3Br and CH_3Cl are the largest contributors to stratospheric ozone depletion, accounting for about one-quarter of the equivalent chlorine in the atmosphere. Other shorter-lived halocarbons, such as CHBr_3 and CHCl_3 , may also be contributing, particularly as Earth's climate changes and alters their natural cycles. These shorter-lived gases also play an important role in tropospheric chemistry of the marine boundary layer.

Introduction

The depletion of stratospheric ozone (O_3) has been driven by long-lived, anthropogenic halocarbons emitted into the atmosphere during the past few decades. When these gases, which in large part resist degradation in the troposphere, reach the stratosphere, their halogen atoms are released as free radicals. Here, the radicals accelerate the removal of ozone through a series of catalytic reactions. Fluorine and iodine do not contribute significantly to stratospheric ozone depletion. Fluorine radicals are removed effectively in the stratosphere, and iodinated compounds react readily in the troposphere. Therefore, persistent halocarbons containing chlorine and bromine are the main halogenated compounds implicated in the destruction of stratospheric ozone, and chlorine and bromine radicals are the primary halogens of concern.

Not all halocarbons in the atmosphere are entirely anthropogenic. Although attention in atmospheric chemistry has centered on halocarbons resulting from human activities, the chlorofluorocarbons (CFCs), halons (CBrF_3 , CBrClF_2), chlorinated solvents (CH_3CCl_3 , CCl_4 , CH_2Cl_2 , CHCl_3), and their replacements, the hydrochlorofluorocarbons (HCFCs), and the methyl halides (methyl chloride (CH_3Cl) and methyl bromide

(CH_3Br)) are present in significant amounts in the troposphere (Figure 1). Other halogenated methanes, such as CHBr_3 , CH_2Br_2 , and CH_3I , can be locally high in atmospheric concentration, but their short tropospheric lifetimes significantly reduce their impact on stratospheric ozone. Nevertheless, halogen atoms from short-lived compounds do at times reach the stratosphere, likely through deep convection of these compounds. The contribution of these gases of lower concentration to ozone depletion is unknown, although considered by most to be small.

Of the naturally produced halocarbons, CH_3Br and CH_3Cl are the largest contributors to stratospheric ozone depletion, accounting for about one-quarter of the equivalent chlorine in the atmosphere (Figure 1). Methyl bromide is the single largest carrier of bromine to the stratosphere. Bromine, on a per-atom basis, is about 50 times more effective in depleting ozone than is chlorine. Although natural sources are the largest contributor to the methyl bromide budget, there was a sizable anthropogenic flux to the atmosphere through its use as a fumigant. By multilateral international agreement, its industrial production is being phased out largely because of its high ozone depletion potential. At this time, only critical uses such as quarantine and preshipment (QPS) fumigation are allowed (Figure 2). Methyl

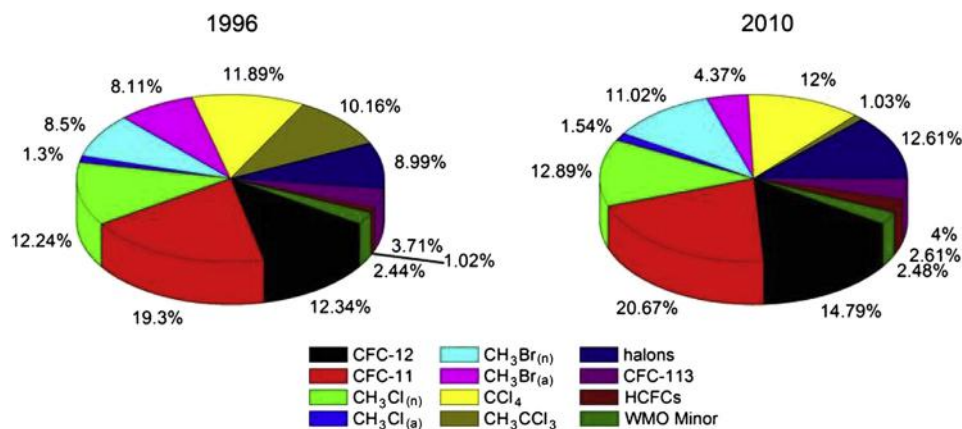


Figure 1 Relative contributions of selected source gases to the equivalent effective stratospheric chlorine loadings for 1996 and 2010 showing the increasing importance of the natural components of methyl bromide and, to a lesser degree, methyl chloride as the phaseout continues.

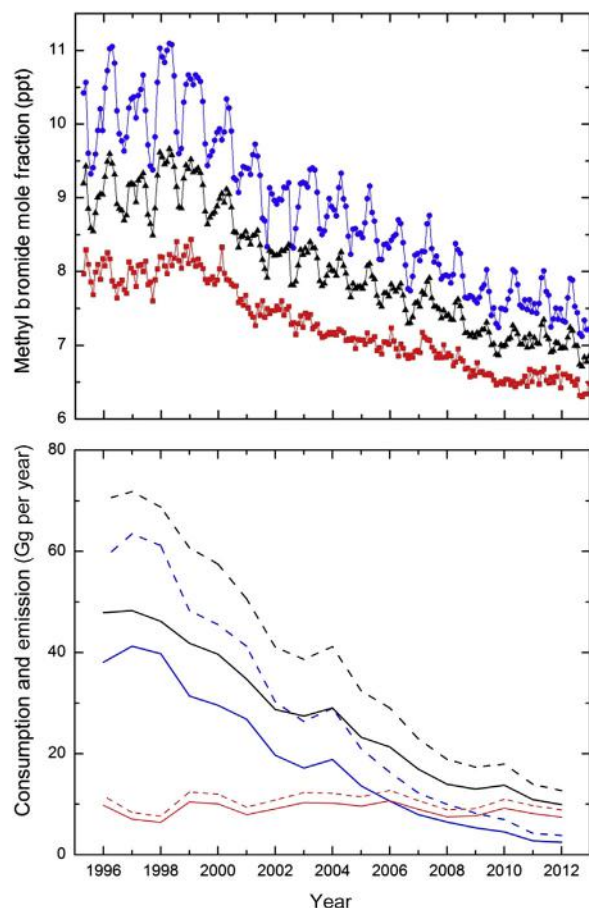


Figure 2 Trends since 1995 in methyl bromide (upper) mixing ratios with NH (●), SH (■), and global (▲) (NOAA data) and (lower) consumption (solid lines) as reported in the United Nations Environment Programme database for non-QPS uses (—), QPS uses (—) and total (—), and emission (dashed lines) from non-QPS uses (---), QPS uses (---), and total (---).

chloride, on the other hand, is the most abundant chlorine-containing compound in the atmosphere, contributing over 15% to the total tropospheric burden of organic chlorine. Its sources are believed to be largely natural and there is some evidence that it was present at over 90% of today's levels during the early twentieth century. Both of these methyl halides have lifetimes of around a year, making them much shorter-lived than the CFCs, solvents, and halons currently banned by international agreement. Nevertheless, their large fluxes into the atmosphere mean that some fraction of their emissions can reach the stratosphere, where they become involved in ozone depletion.

Methyl Bromide

Methyl bromide is present in the atmosphere at a mole fraction, or volume mixing ratio, of around 7 parts per trillion (1 ppt = 1×10^{-12} mol of specific gas per mole of air) (Figure 2) and its known natural sources include the

ocean, biomass burning, fungi, salt marshes, wetlands, rice paddies, mangroves, and tropical rain forests (Figure 3). Methyl bromide has an added anthropogenic source from fumigation uses (agricultural, structural, and QPS). While the non-QPS anthropogenic emissions have been phased out as a result of the Montreal Protocol and its Amendments, the QPS anthropogenic sources remain for CH_3Br (Figure 2). The removal processes for CH_3Br include reaction with hydroxyl radicals (OH) in the troposphere, degradation in the oceans and soils, and photolysis in the stratosphere.

Until the 1990s, little attention had been paid to this gas in the atmosphere, in part because of its low mixing ratio and short atmospheric lifetime. In the early 1990s, atmospheric methyl bromide was thought to emanate naturally from a large oceanic source and to be destroyed exclusively by reactions in the atmosphere. Anthropogenic emissions, mainly from disinfection of soils, commodities, and structures, were considered responsible for about 3 ppt of CH_3Br in the atmosphere. Biomass burning and emissions from burning of leaded gasoline were thought to be the possible contributors, but were not quantified at that time. Recognizing that there was a paucity of information on this important atmospheric gas, scientists began working to understand more completely its cycling and atmospheric budget. The results were surprising in a number of areas.

The first of these surprises was that the ocean was not the large source it was thought to be, but rather a small net sink for atmospheric CH_3Br . This net sink, however, results from rapid aquatic production and degradation working in opposition in the surface ocean, leaving it largely undersaturated. In some areas where production exceeds degradation, the ocean is supersaturated in methyl bromide, but in most of the surface ocean, most of the time, methyl bromide is undersaturated. Because the degradation rate of CH_3Br is so high in most of the surface ocean, it had to be included as a significant component of the atmospheric lifetime computation. Subsequent calculations of the atmospheric lifetime of CH_3Br yielded a rate that was almost equal to the loss due to reaction with OH in the troposphere. This alone lowered the atmospheric lifetime of CH_3Br from around 2 years to 1 year.

At about the same time, studies of the terrestrial environment revealed additional sinks and sources of atmospheric CH_3Br . The discovery that CH_3Br was degraded rapidly in a variety of soils, mainly by prokaryotic bacteria, lowered the estimates of atmospheric lifetime even further.

Our current understanding of atmospheric CH_3Br is that of a gas with numerous, diverse sources and significant sinks on land, in the ocean, and in the atmosphere. Its lifetime, including atmospheric, oceanic, and soil sinks, is now computed at 0.8 year, but its calculated atmospheric budget is largely out of balance, with sinks outweighing sources. New findings continue to reveal previously unidentified sources, which seem gradually to be closing the gap between calculated sources and sinks (Figure 3).

Anthropogenic emissions of CH_3Br were scheduled for phaseout by 2005 in developed countries and 2015 in developing countries. The atmospheric concentration of this gas has responded to these reductions in emissions, decreasing from the high of about 10 ppt in 1995 to about 8 ppt in 2012 (a 20% reduction). As the anthropogenic contribution decreases,

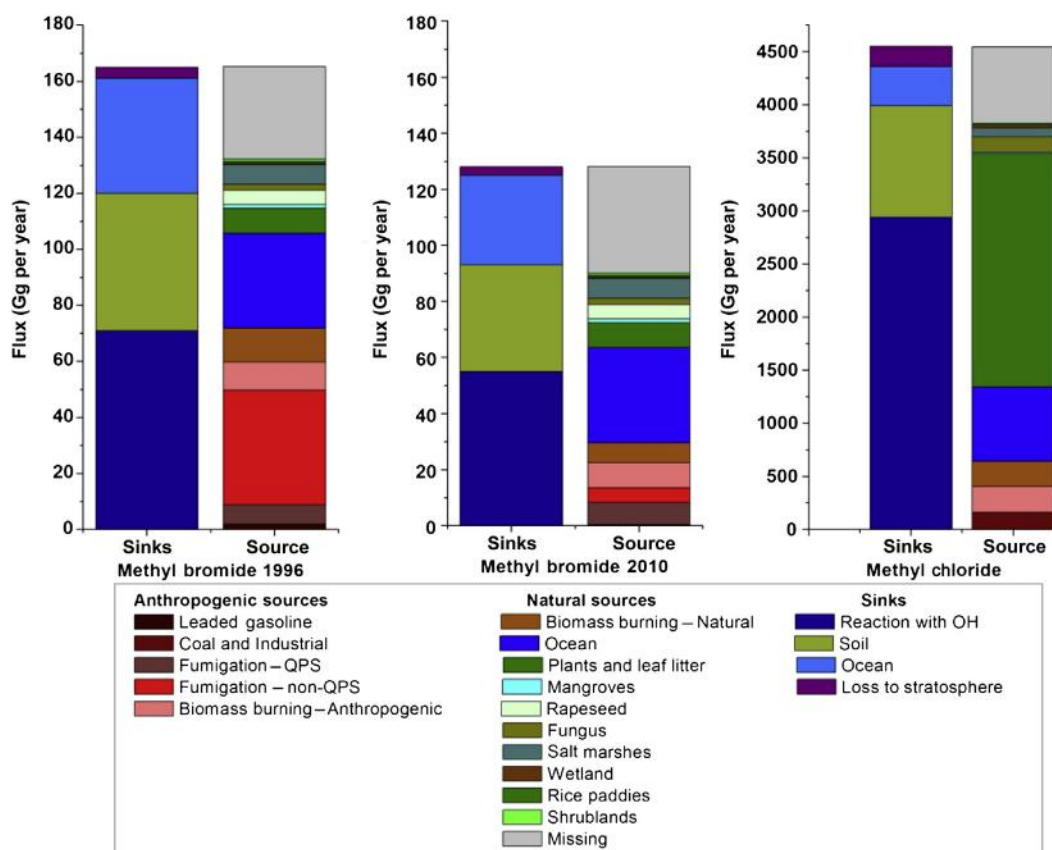


Figure 3 Source and sink fluxes of methyl bromide (1996 and 2010) and methyl chloride.

the natural emissions become relatively more influential in stratospheric ozone depletion (Figure 3).

Methyl Chloride

Like methyl bromide, methyl chloride, at roughly 540 ppt in the atmosphere, received little attention until the mid-1990s, as most research efforts were directed toward the rapidly increasing anthropogenic halocarbons. Methyl chloride also has a short atmospheric lifetime, ~ 1 year, relative to the anthropogenic halocarbons and its anthropogenic sources are very small (Figure 3). Until recently, it was thought that most of the global emissions of CH_3Cl came from the oceans. Although the oceans are still considered a major source of CH_3Cl , new and more detailed studies show that the oceanic source is responsible for at most 15–20% of the methyl chloride in the atmosphere. Similarly, wood-rotting fungi contribute only a small amount and anthropogenic emissions of CH_3Cl and constitute only about 1% of the total budget. While biomass burning is another substantial source of CH_3Cl , there remains a sizable deficit in the overall budget. Studies have suggested that emissions from tropical plants could potentially make up this deficit; however, total emissions from this source have not been quantified. The identified sinks for methyl chloride, mainly loss via reaction with OH in the troposphere, suggest that about half of the CH_3Cl in the atmosphere is unaccounted for.

Other Gases

Most of the remaining naturally produced organic halogens are of low concentration and short lifetime. They are therefore thought to pose only a small threat to stratospheric ozone. However, they have been observed at the tropopause (Table 1) and they can be convected rapidly from the Earth's surface into the upper troposphere and lower stratosphere. Of the purely chlorinated gases, chloroform (CHCl_3) and perchloroethylene (C_2Cl_4) appear to have significant natural sources, although their budgets have been little studied. The naturally occurring brominated species (e.g., CHBr_3 , CH_2Br_2 , CHBr_2Cl), although low in concentration, are of some concern because of the efficiency of bromine in depleting stratospheric ozone. These gases are produced in the ocean and are supersaturated throughout, by tens to hundreds of percent. Together, these lesser gases represent most of the total flux of organic bromine into the troposphere (Table 2).

Closing the Budgets

From recent research, it is clear that the missing sources of methyl bromide and methyl chloride are not oceanic. The saturations of these gases have now been mapped over most of the oceans. While the ocean has become a small net source of methyl bromide as the anthropogenic sources have been phased out, its emission and the lifetime relative to the

Table 1 Organic bromine in the troposphere

Compound	Compound mole fraction $\times 10^9$	Bromine mole fraction $\times 10^9$
CH ₃ Br	10	10
CBrF₃	2.3	2.3
CBrClF ₂	3.5	3.5
C₂Br₂F₄	0.45	0.9
CH ₂ Br ₂	0.75–1.5	1.5–3
CHBr ₂ Cl	0–0.5	0–1
CHBr ₃	0.5–4	1.5–12
CH ₂ BrCl	0–0.5	0–0.5
CHBrCl₂	0–0.5	0–0.5
Total	>17.5	>19.7

Purely anthropogenic compounds appear in bold. Compounds that are natural or have significant natural components to their budgets (e.g., CH₃Br) are shown in normal.

Table 2 Potential global bromine fluxes

Compound	Source	Flux (Gmol Br year ⁻¹)
CH ₃ Br	Anthropogenic	0.5
CH ₃ Br	Natural	1.0
CH ₂ Br ₂	Ocean	2.0
CHClBr ₂	Ocean	1.5
CHBr ₃	Ocean	2.0
CH ₂ BrCl	Ocean	0.5
CBrClF ₂	Anthropogenic	0.05
CBrF ₃	Anthropogenic	0.012
C ₂ Br ₂ F ₄	Anthropogenic	0.005
Total		7.5 (6.0)

ocean degradation have not changed over the course of the phaseout. The magnitude of the missing sources has also remained fairly constant from before the phaseout through today (2013). Together, these suggest that the missing sources are not from the ocean or an underestimate of the fraction of emission from the fumigation uses. The ocean also is insufficiently supersaturated in methyl chloride to explain more than a small percentage of its total flux to the atmosphere. Further, although the atmospheric mixing ratios of both gases show marked seasonal cycles, the cycles in the Northern Hemisphere are amplified over those in the Southern Hemisphere, particularly for methyl bromide. Although tropospheric OH is responsible in part for the seasonal cycling, the uneven match between hemispheres, especially with smaller amplitude in the Southern Hemisphere, speaks for a more complicated involvement of sources and sinks. Because fluxes from the ocean to the atmosphere are retarded significantly at the ocean surface, it is not possible for cycles in the oceanic flux to drive significant seasonal variations in the atmosphere.

This has led to several studies to determine whether methyl halides are released in significant amounts elsewhere, and it appears that they are. A number of investigations have shown that natural and cultivated terrestrial plants emit both of these gases, and others as well.

The emissions seem to be related to the amount of halide in the soil. Coastal plants, such as those in salt marshes and in tropical environments, emit large quantities of methyl halides and, although their global area of coverage is small, they seem to contribute significantly to the global budget. To date, only a few plants and a few ecotypes have been studied for emission of methyl halides. Additional investigations are likely to locate more sources from the terrestrial biosphere, and with this a possible closing of the atmospheric budgets of these gases.

Although the natural fluxes of these gases existed long before there were problems with stratospheric ozone depletion, this does not mean that they are not important. Contributions of bromine and chlorine from the anthropogenic gases are now declining and should continue to do so into the future; this should provide some relief, albeit slowly. If the fluxes of natural compounds remain the same and all countries abide by the Montreal Protocol and its Amendments, then 'preozone-hole conditions' could be reached by the mid-twenty-first century. However, everything may not remain the same. A question that will become more pressing with global change is, "How will the fluxes of methyl bromide, methyl chloride, and other halogenated gases between the Earth's surface and atmosphere change in the future?" A change in sea surface temperature or soil temperature will certainly affect the fluxes, as will changes in precipitation or land use patterns. It is possible that such alterations of natural fluxes could offset or delay the timing of recovery, but we cannot know until we more fully understand the natural cycles of these ozone-depleting gases.

See also: **Stratospheric Chemistry Topics: Halogens.**

Further Reading

- Blei, E., Hardacre, C.J., Mills, G.P., Heal, K.V., Heal, M.R., 2010. Identification and quantification of methyl halide sources in a lowland tropical rainforest. *Atmospheric Environment* 44, 1005–1010.
- Butler, J.H., King, D.B., Lobert, J.M., Montzka, S.A., Yvon-Lewis, S.A., Hall, B.D., Warwick, N.J., Mondeel, D.J., Aydin, M., Elkins, J.W., 2007. Oceanic distributions and emissions of short-lived halocarbons. *Global Biogeochemical Cycles* 21, GB1023. <http://dx.doi.org/10.1029/2006GB002732>.
- Hossaini, R., Mantle, H., Chipperfield, M.P., Montzka, S.A., Hamer, P., Ziska, F., Quack, B., Krüger, K., Tegtmeier, S., Atlas, E., Sala, S., Engel, A., Bönsch, H., Keber, T., Oram, D., Mills, G., Ordóñez, C., Saiz-Lopez, A., Warwick, N., Liang, Q., Feng, W., Moore, F., Miller, B.R., Maréchal, V., Richards, N.A.D., Dorf, M., Pfeilsticker, K., 2013. Evaluating global emission inventories of biogenic bromocarbons. *Atmospheric Chemistry and Physics Discussions* 13 (5), 12485–12539.
- Hu, L., Yvon-Lewis, S.A., Liu, Y., Bianchi, T.S., 2012. The ocean in near equilibrium with atmospheric CH₃Br. *Global Biogeochemical Cycles* 26 (3), GB3016. <http://dx.doi.org/10.1029/2011GB004272>.
- Hu, L., Yvon-Lewis, S.A., Butler, J.H., King, D.B., Lobert, J., Montzka, S.A., 2013. An improved oceanic budget for methyl chloride. *Journal of Geophysical Research* 118 (2), 715–725. <http://dx.doi.org/10.1029/2012JC008196>.
- Liu, Y., Yvon-Lewis, S.A., Thornton, D.C.O., Butler, J.H., Bianchi, T.S., Campbell, L., Hu, L., Smith, R.W., 2013. Spatial and temporal distributions of bromoform and dibromomethane in the Atlantic Ocean and their relationship with photosynthetic biomass. *Journal of Geophysical Research Oceans* 118, 3950–3965. <http://dx.doi.org/10.1002/jgrc.20299>.

Montzka, S.A., Reimann, S., 2011. Ozone-depleting substances (ODSs) and related chemicals. Chapter 1 in *Scientific Assessment of Ozone Depletion 2010, Global Ozone Research and Monitoring Project – Report No. 52*. World Meteorological Organization, Geneva.

World Meteorological Organization/United Nations Environment Programme, 2011. The 2010 Scientific Assessment of Ozone Depletion, Chapter 1, Ozone Depleting Substances and Related Chemicals. Ozone Secretariat, Nairobi, Kenya. http://ozone.unep.org/Assessment_Panels/SAP/Scientific_Assessment_2010/03-Chapter_1.pdf.

Ziska, F., Quack, B., Abrahamsson, K., Archer, S.D., Atlas, E., Bell, T., Butler, J.H., Carpenter, L.J., Jones, C.E., Harris, N.R.P., Hepach, H., Heumann, K.G., Hughes, C., Kuss, J., Krüger, K., Liss, P., Moore, R.M., Orlikowska, A., Raimund, S., Reeves, C.E., Reifenhäuser, W., Robinson, A.D., Schall, C., Tanhua, T., Tegtmeier, S., Turner, S., Wang, L., Wallace, D., Williams, J., Yamamoto, H., Yvon-Lewis, S., Yokouchi, Y., 2013. Global sea-to-air flux climatology for bromoform, dibromomethane and methyl iodide. *Atmospheric Chemistry and Physics* 13, 8915–8934. <http://dx.doi.org/10.5194/acp-13-8915-2013>.

Introduction

The hydrogen radical family (HO_x) consists of the hydrogen (H), hydroxyl (OH), and hydroperoxyl (HO₂) radicals. Concentrations of these highly reactive radicals are small, between 1 part per trillion (ppt) in the lower stratosphere and 400 ppt in the upper stratosphere. Despite this, HO_x is important because it participates in many reactions that control the photochemistry of stratospheric ozone. The hydrogen radicals are important in the removal of O₃ through direct reaction with O₃ and indirectly through reaction with the halogen oxides. HO_x also removes ozone through reactions with the nitrogen and halogen chemical families. The extensive coupling of HO_x to these chemical families leads to a particularly complex set of reactions that control HO_x photochemistry. Understanding these mechanisms is important in understanding both HO_x and, more broadly, the mechanisms that control ozone photochemistry.

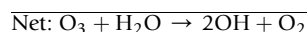
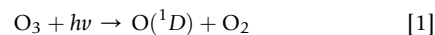
HO_x Sources

HO_x is produced from the oxidation of the stable hydrogen-containing species, water, methane, and molecular hydrogen. The relative strength of these sources is largely determined by their concentrations at the tropopause: H₂O ~4 parts per million (ppm), CH₄ ~1.5 ppm, and H₂ ~0.5 ppm. The numerous pathways that participate in this oxidation are diagrammed in Figure 1. The oxidation occurs mostly through gas phase reactions with highly reactive species, i.e. excited oxygen atoms, chlorine atoms, and OH. The oxidation of water also occurs via hydrolysis reactions catalyzed by acid aerosols

found in the lower stratosphere. In most cases, the production mechanisms include a large number of reactions, most of which have little direct effect on HO_x. These mechanisms are presented in terms of reaction sequences that can be simplified in terms of rate-determining reactions and net yields.

Gas Phase Processes

The largest single source of HO_x throughout the stratosphere is the oxidation of H₂O by the electronically excited oxygen atom O(¹D) which is generated from the photolysis of O₃ at wavelengths less than ~330 nm. The production from this mechanism is represented by a sequence of reactions that leads to a net conversion of H₂O to HO_x:



Only a small fraction of the O(¹D) produced from reaction [1] subsequently reacts via reaction [2]; most O(¹D) relaxes to ground state O(³P) after collisions with N₂ and O₂. The rate of this sequence is determined by the slowest or the 'rate-determining step', in this case reaction [2]. The rate of the production of HO_x from H₂O, which is equal to twice the rate of reaction [2], is shown in Figure 2. This and the following sequences are identified in the figures by the rate-determining steps.

The oxidation of methane requires a greater number of reactions to liberate all four hydrogen atoms. The sequence initiated by O(¹D) can produce as many as four HO_x:

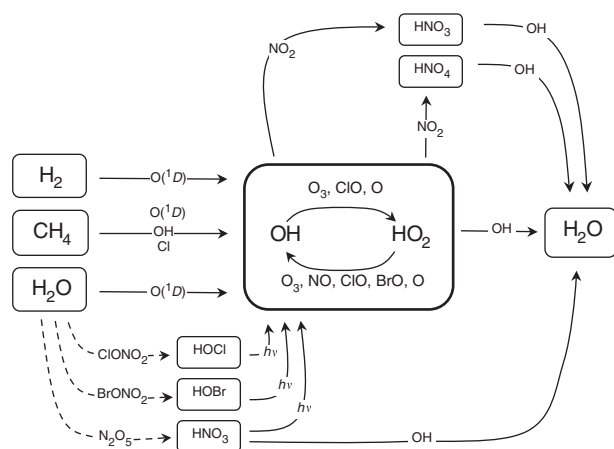
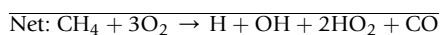
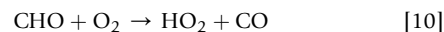
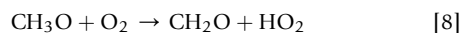
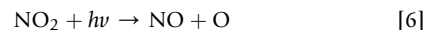
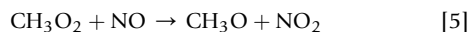
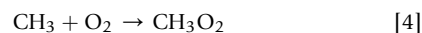
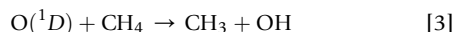
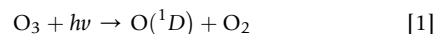


Figure 1 Primary sources and sinks of HO_x are shown. Gas phase reactions are denoted with solid lines and heterogeneous reactions with dashed lines.

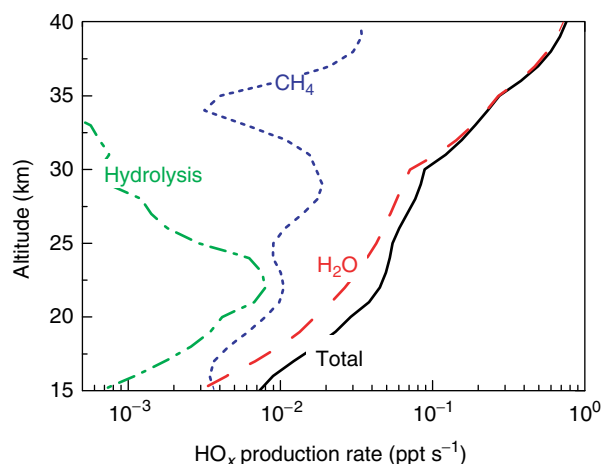


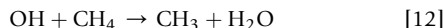
Figure 2 Typical mid-latitude daytime production rates of HO_x calculated from concentration profiles of the source gases are shown versus altitude. The CH₄ oxidation sequences initiated by O(¹D), OH, and Cl are grouped together. Likewise, the hydrolysis reactions involving N₂O₅, ClONO₂, and BrONO₂ are combined. The production from H₂ is less than 4% of the total at all altitudes.

An alternate pathway for the photolysis of CH₂O in reaction [9] is the production of H₂ and CO:



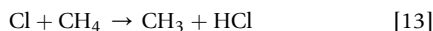
When the oxidation of methane follows this path, the net yield is only 2HO_x for every CH₄ consumed. Since reactions [9] and [11] compete, the yield of HO_x from CH₄ oxidation initiated by reaction [3] is somewhere between 2 and 4. The utility of the rate-determining step and net yields is particularly evident in this sequence, where nine reactions can be thought of in terms of one. In this case, the rate of production of HO_x from CH₄ is 2–4 times the rate of the rate-determining step, reaction [3]. The rate of this and the following CH₄ oxidation sequences decrease relative to the H₂O source at higher altitudes. This results from the net conversion of CH₄ to H₂O as the air in the lower stratosphere ages during the slow ascent into the mid to upper stratosphere.

The methane oxidation sequence that is initiated by OH is autocatalytic. OH is consumed at the initiation step,

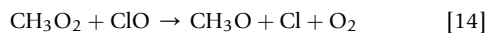


but the subsequent reactions [4], [5], [6], [7], [8], [9] and [10] can produce as many as 2HO_x. In the lower stratosphere where the oxidation of CH₄ is most important, the rate of reaction [12] is faster than that of reaction [3], so that the sequence initiated by OH is often more important than that initiated by O(¹D).

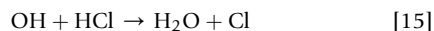
Oxidation of CH₄ is also initiated by reactions of CH₄ with Cl.



The methane oxidation sequence that is initiated by Cl can be the most important source of HO_x in the winter polar vortex when active chlorine levels are high and HCl concentrations are low. In these conditions, ClO controls the conversion of CH₃O₂:

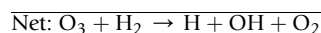
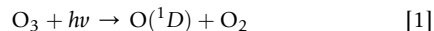


Outside the winter polar vortex this sequence is not a large source of HO_x, because Cl concentrations are lower and the reaction



is important, resulting in a lower yield for HO_x production. In these conditions [13] and [15] are the dominant production and loss terms of HCl.

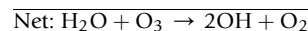
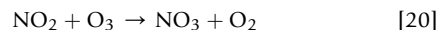
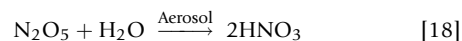
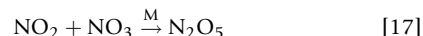
The production of HO_x from the oxidation of H₂ is analogous to that of H₂O:



This source is much smaller than the H₂O and CH₄ sources, owing to the much lower concentration of H₂ in the stratosphere and to a smaller rate constant. In the lower stratosphere where this sequence is most significant, H₂ accounts for only 4% of the total production of HO_x.

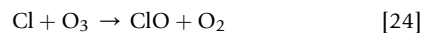
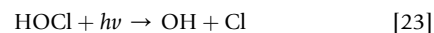
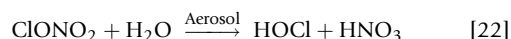
Heterogeneous Processes

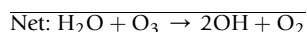
Heterogeneous reactions are important in the partitioning of the nitrogen and halogen families. These same reactions are important sources of HO_x in the lower stratosphere where aerosol concentrations are significant. The hydrolysis of N₂O₅ on sulfuric acid aerosols is a major source of HNO₃ and HO_x in the lower stratosphere:



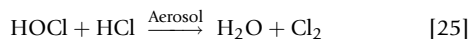
The first step in this sequence is significant only at night because NO₃ is easily photolyzed during the day. Thus, this sequence is not important in the high-latitude summer when continuous sunlit conditions occur. An alternate pathway for the removal of HNO₃ in reaction [19] is reaction with OH, discussed in the next section. This pathway removes HO_x so that the yield from the hydrolysis of N₂O₅ is less than 2HO_x.

In the winter and early spring polar vortex when ClO and aerosol concentrations are high, heterogeneous reactions of ClONO₂ are the dominant source of HNO₃. The hydrolysis of ClONO₂ can be a particularly strong source of HO_x:





The heterogeneous reaction [22] is strongly temperature-dependent, proceeding fastest at low temperatures (190–200 K). As in the case of N₂O₅ hydrolysis, the reaction of OH with HNO₃ competes with [19] to reduce the yield of this sequence. In addition, the heterogeneous reaction:

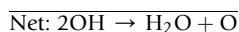
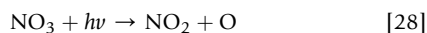
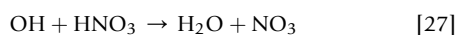
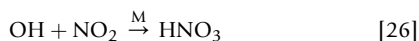


competes with reaction [23] and to further reduce the yield of this sequence. When ClONO₂ is absent, as in the polar winter lower stratosphere, reaction [25] can be a net sink of HO_x. The analogous hydrolysis of BrONO₂ is only weakly temperature-dependent, and it is only a small source throughout the lower stratosphere. This reaction produces a distinct increase in HO_x when HOBr produced is photolyzed at daybreak.

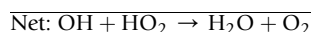
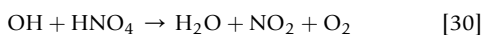
Sinks

The removal of HO_x in the stratosphere consists entirely of reactions that convert HO_x to H₂O. The numerous pathways involving NO_y and HO_x intermediates are shown in Figure 1. Unlike the nitrogen and halogen radical families, HO_x radicals are not sequestered in reservoirs. This is because these potential reservoirs (HNO₃ and HNO₄) react rapidly with OH, serving as sinks instead.

The primary removal mechanism for HO_x in the lower and middle stratosphere is the reaction of OH with NO₂, followed by reaction with HNO₃:



The similar sequence that produces and removes HNO₄ also removes HO_x:



The HNO₃ and HNO₄ sequences are most important in the lower stratosphere where the concentrations of HNO₃ and HNO₄ are high. The first-order loss rates of the removal sequences are shown in Figure 3. When concentrations of ClO are elevated, the reaction



is a significant sink of HO_x. Because concentrations of HCl are usually low when ClO concentrations are high, the reaction of OH with HCl [15] that completes the conversion of HO_x to H₂O is not significant.

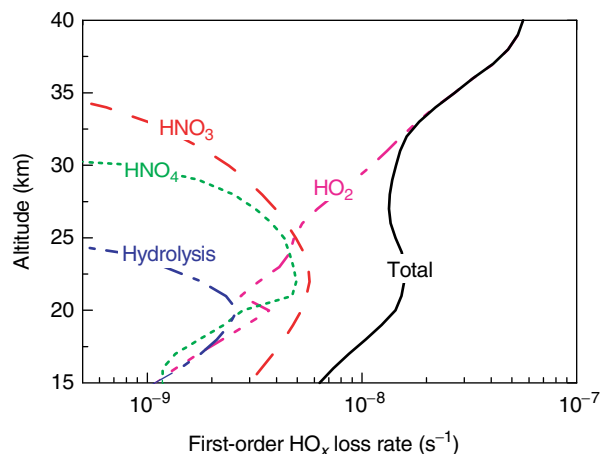
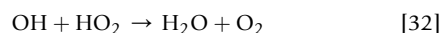


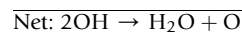
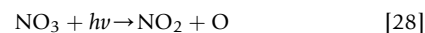
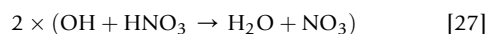
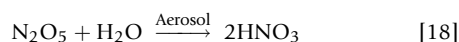
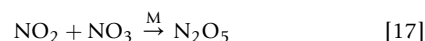
Figure 3 Typical mid-latitude first-order loss rates determined from profiles of the sinks of HO_x are shown versus altitude. The first-order loss rates, e.g. $k_{\text{OH}+\text{HNO}_3}[\text{HNO}_3]$, are the independent variables that control the removal of HO_x.

When the concentrations of HO_x are large the self-reaction of HO_x becomes significant:



In the mid to upper stratosphere reaction [32] is the dominant sink of HO_x. In this region of the stratosphere HO_x photochemistry is greatly simplified. Concentrations of HO_x are determined nearly entirely by production from H₂O and removal via the self-reaction.

The most significant heterogeneous removal mechanism of HO_x is the hydrolysis of N₂O₅:

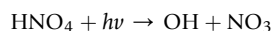
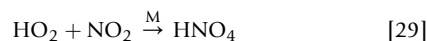


This reaction is important only in the lower stratosphere where aerosol concentrations are high. As mentioned in the previous section, an alternate pathway to reaction [27] is the photolysis of HNO₃[19], so that the occurrence of [18] leads to the removal of less than 2HO_x. The analogous removal mechanisms involving ClONO₂ and BrONO₂ are less significant because the reactions of OH with HOCl and HOBr are too slow to compete with the photolysis of these species.

Secondary Sources and Sinks

The production and loss mechanisms shown in Figure 1 are portrayed as part of a one-way flux from the primary sources (H₂, H₂O, and CH₄) through HO_x and back into H₂O. On average this flux is balanced, but there are certain situations

when other pathways are significant. These processes are part of null sequences. For example, the formation followed by photolytic destruction of HNO₄ is



This is a net null for HO_x averaged over a 24-hour diurnal cycle. However at any given instant within the diurnal cycle these reactions are not required to balance. In particular, at sunrise and sunset HNO₄ is not in steady state, i.e. the rate of [29] does not balance [30] and [33]. Under these twilight conditions, this sequence can be a source (sunrise) or a sink (sunset). Other null cycles that can be instantaneous sources or sinks of HO_x include the formation and photolysis of HONO, HOCl, HOBr, and H₂O₂. Conceptually, these terms should be considered secondary processes. That is, they do not influence the interpretation of the 24-hour average abundance of HO_x versus altitude or latitude, but they are significant in model calculations that attempt to reproduce HO_x at twilight conditions.

Diurnal Change

The production of HO_x is tied to the flux of ultraviolet (UV) radiation that drives the photolysis reactions and initiates the production sequences. This UV flux, which is strongly attenuated by O₃, depends on the O₃ slant column (the amount of O₃ between an air parcel and the Sun). During a diurnal cycle, this slant column changes with the angle of the Sun, resulting in large changes in UV flux and photolysis rates. The resulting change in the production rate of HO_x for conditions typical of the mid-latitude lower stratosphere is shown in Figure 4. The sharp increase in the production rate indicates the strong

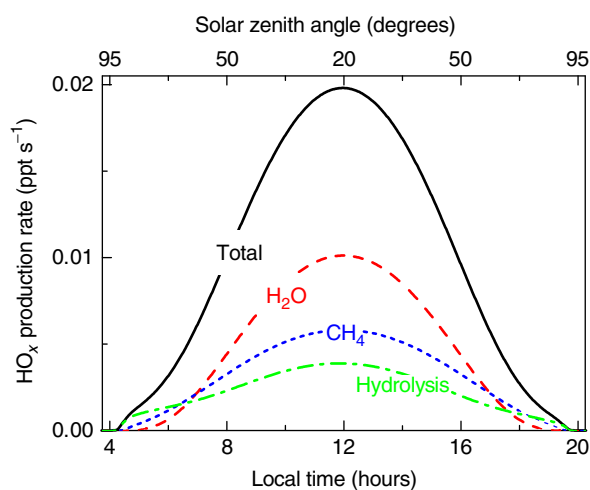


Figure 4 The production rate of HO_x is shown for a typical day in the mid-latitude lower stratosphere. The solar zenith angle is the angle between the Sun and the zenith (directly overhead). The sequences that oxidize CH₄ and the hydrolysis reactions are grouped together.

dependence on UV flux, hence solar zenith angle. The relative strengths of the sources and sinks of HO_x also depend on solar flux. For example, the oxidation of CH₄ is faster than that of H₂O at the highest solar zenith angles. This is because the oxidation of CH₄ can be initiated by OH and Cl atoms that are produced more easily than O(¹D) at twilight conditions.

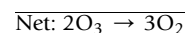
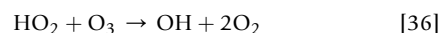
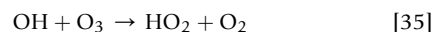
HO_x Cycling

The relative concentration of H, OH, and HO₂ is controlled by fast cycling reactions that do not produce or remove HO_x. H is converted to HO₂ via the extremely fast reaction with O₂.



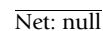
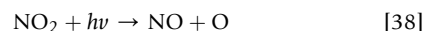
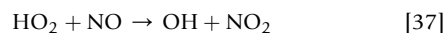
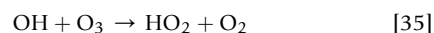
This reaction is significantly faster than the reactions that produce H (i.e. [9] and [16]), so that concentrations of H are negligible compared to the OH and HO₂. Because the conversion of H is so fast, H is often neglected and the production of H is considered equivalent to the production of HO₂.

The relative concentration of OH and HO₂ is controlled by reactions that interconvert OH and HO₂. The primary conversion mechanism in the lower and middle stratosphere is



This reaction sequence is a net loss for ozone, and accounts for a large fraction of the total ozone removal rate in the lower stratosphere. In this sequence, the rate constant for reaction [35] is much larger than that of reaction [36], so that concentrations of HO₂ are almost always greater than OH in the lower to mid stratosphere.

An important pathway for the conversion of HO₂ → OH is part of a null cycle:



This and the prior sequence illustrate the interaction between HO_x and NO_x on ozone loss rates. Reaction [36] removes O₃ and reaction [37] leads to the production of O₃. The relative rate of [36] compared with [37] determines the net O₃ removal rate following reaction [35].

The relative abundance of OH and HO₂ is proportional to the concentrations of the species and rate constants that interconvert OH and HO₂. In the mid-latitude lower to mid stratosphere, where reactions [35], [36] and [37] dominate the interconversion rate, the ratio HO₂/OH is controlled by O₃, NO, and the rate constants for these reactions. Figure 5 shows how the ratio HO₂/OH responds to the changes in O₃ and NO. At a fixed amount of NO, an increase in O₃ leads to greater

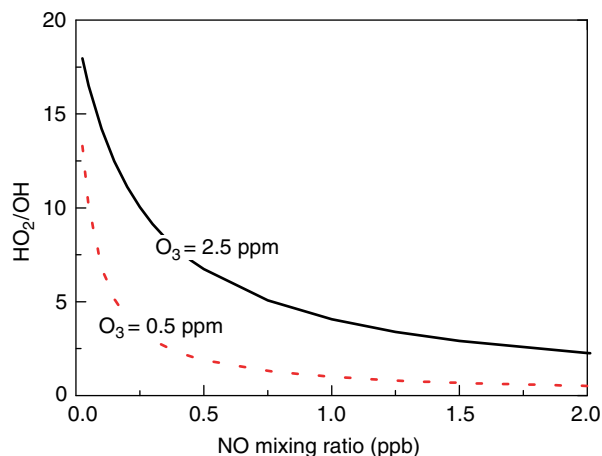
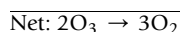
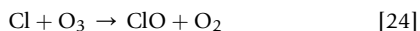
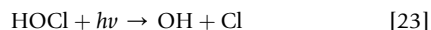
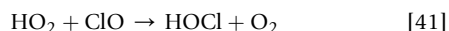
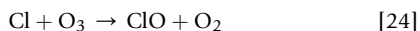
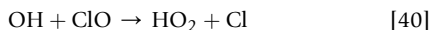


Figure 5 The ratio of HO₂/OH is controlled by NO antrun and O₃ in the mid-latitude lower stratosphere. The predicted ratio is shown versus NO for O₃ = 0.5 and 2.5 ppm.

HO₂/OH because the rate constant for reaction [35] is roughly 15 times faster than that of reaction [36]. At the limit of very high concentrations of O₃ (or low NO) the ratio of HO₂/OH is determined by the ratio of the rate constants for reactions [35] and [36], i.e. HO₂/OH $\sim k_{35}/k_{36}$. For a fixed amount of O₃, increases in NO lead to decreases in HO₂ because reaction [37] converts HO₂. At some limit of very high NO (or low O₃), HO₂/OH would approach zero.

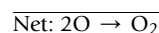
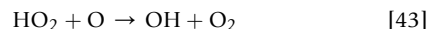
When the concentrations of ClO are high, such as the wintertime polar vortex, the reaction with ClO controls the balance between OH and HO₂:



The analogous sequence involving BrO also contributes to the interconversion of OH and HO₂ and to the removal of O₃.

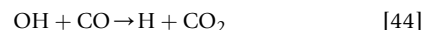
These sequences are almost always rate limited by reaction [41] and the analogous reaction of HO₂ with BrO.

In the upper stratosphere, the reactions with O atoms become important:



Above ~ 40 km this sequence dominates the conversion of OH and HO₂. The rate constant of reaction [43] is roughly twice as large as that for reaction [42], so that concentrations of OH are greater than [HO₂] in the highest part of the stratosphere.

The OH-initiated oxidation of CH₄, reactions [12] followed by [4], [5], [6], [7], [8], [9] and [10], converts OH to HO₂. In addition, the subsequent oxidation of CO converts OH $\rightarrow$ H, via



These reactions are important only in the lowest part of the stratosphere near the tropopause region, where they might account for a few percent of the total OH $\rightarrow$ HO₂ conversion rate.

See also: **Aerosols:** Aerosol Physics and Chemistry. **Chemistry of the Atmosphere:** Chemical Kinetics; Principles of Chemical Change. **Ozone Depletion and Related Topics:** Ozone Depletion Potentials. **Stratospheric Chemistry Topics:** Halogens; Hydrogen Budget; Reactive Nitrogen (NO_x and NO_y).

Further Reading

- Dessler, AE, 2000. *The Chemistry and Physics of Stratospheric Ozone*. Academic Press, San Diego.
- Jacob, DJ, 1999. *Introduction to Atmospheric Chemistry*. Princeton University Press, Princeton.
- Johnston, HS, Podolske, JR, 1978. Interpretations of Stratospheric Photochemistry. *Review of Geophysics and Space Physics* 16, 491–519.
- McElroy, MB, Salawitch, RJ, Minschwaner, K, 1992. The changing stratosphere. *Planetary and Space Science* 40, 373–401.
- Okabe, H, 1978. *Photochemistry of Small Molecules*. Wiley, New York.
- Wayne, RP, 2000. *Chemistry of Atmospheres: An Introduction to the Chemistry of the Atmospheres of Earth, the Planets, and Their Satellites*, third ed. Oxford University Press, Oxford.

Hydrogen Budget

JE Harries, Imperial College of Science, Technology and Medicine, Blackett Laboratory, London, UK

© 2003 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 5, pp 2180–2184, © 2003, Elsevier Ltd.

Introduction

The principal components of the hydrogen budget of the stratosphere are water vapor, H_2O (mixing ratio between 3 and 4 parts per million by volume, ppmv, in the lower stratosphere), methane, CH_4 (about 1.7 ppmv), and molecular hydrogen, H_2 (about 0.5 ppmv). The first two of these are greenhouse gases, with strong absorption bands in the infrared, making their concentration and evolution of interest in studies of the balance of climate, and how this might change with time. The hydrogen free radicals such as the hydroxyl radical OH, HO_2 , and atomic hydrogen, H, are present in much lower concentrations (mixing ratios of order 10^{-11} at 30 km), but are highly reactive, and are consequently of particular importance in atmospheric chemistry. Since this article is concerned with the budget of hydrogen in the stratosphere, we shall be concerned primarily with the three species that dominate the mass: the trace radicals will only be considered in so far as they enter into reactions which determine the concentrations of the three major constituents.

Because of the importance to radiative energy exchange and chemistry, a wide range of studies of both the major and minor hydrogen-bearing species have been undertaken over the years. In what follows, we shall review the state of knowledge on the concentrations, budget, and variability of the three principal hydrogen species H_2O , CH_4 , and H_2 .

Supply of Hydrogen Species to the Stratosphere

Water vapor is continuously supplied to the stratosphere from the ocean surface, via the troposphere. While concentrations in the troposphere are high, sometimes approaching saturation near the surface, the Brewer–Dobson circulation transports air upward over the tropics, and through the tropical tropopause, which is characteristically very cold and high. This causes the ‘freeze-out’ of water by the tropopause cold trap, with the consequence that the air moving into the stratosphere is extremely dry (mixing ratios of order a few parts in 10^{-6} (see **Stratospheric Chemistry Topics: Stratospheric Water Vapor**).

In the stratosphere, oxidation of both methane and H_2 takes place, adding to the concentration of water vapor, but simultaneously H_2 can be produced by oxidation of formaldehyde, CH_2O , which itself is derived from CH_4 . The overall net effect is to cause water vapor to increase with height, mainly at the expense of methane, which decreases in mixing ratio, while molecular hydrogen stays roughly constant in mixing ratio with height (since reactions which both add and delete H_2 occur with similar rates). Later we will examine some data for H_2O and CH_4 from both a satellite and a balloon experiment, which indicate that the sum of total hydrogen, which may be expressed as $\psi = 2 \times \text{CH}_4 + \text{H}_2\text{O} + \text{H}_2$, is, as far as can be measured, constant with height through most of the stratosphere.

Loss of Hydrogen Species from the Stratosphere

There are three principal mechanisms by which hydrogen species are lost from the stratosphere:

1. by being part of the overall global circulation of descending air at mid to high latitudes;
2. by loss, particularly of the lightest component, molecular hydrogen, from the top of the the atmosphere to space;
3. by removal of ice where condensation occurs and ice particles may be removed by movement into the troposphere and subsequent ice sublimation or ice melting and liquid evaporation: this can happen, for example, at the top of cumulonimbus clouds once they begin to decay, or in the Antarctic polar vortex, where extremely low temperatures and sinking motion prevail.

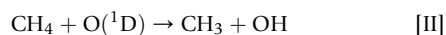
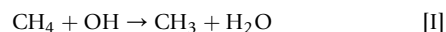
Photochemistry of Stratospheric Hydrogen

Since original work in 1950 by two major figures in atmospheric chemistry and aeronomy, David Bates and Marcel Nicolet, the photochemistry of the hydrogen family has been treated many times. A valuable treatment is to be found in the book by Brasseur and Solomon (see under [Further Reading](#)). The three principal components, H_2O , CH_4 , and H_2 all enter the stratosphere from the troposphere below, as part of the general circulation of the atmosphere. In the troposphere, water vapor derives, of course, ultimately from the large reservoir in the oceans; methane comes from anaerobic processes, and molecular hydrogen is thought to be produced from cars and biomass burning, as well as from natural sources at the surface. The approximate mean values of the mixing ratios of the three as they enter the stratosphere are:

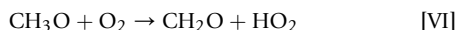
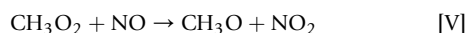
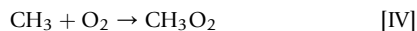
- H_2O : $\sim 3\text{--}4.2$ ppmv
- CH_4 : ~ 1.7 ppmv
- H_2 : ~ 0.5 ppmv

There has been some controversy over the so-called ‘mean entry level’ mixing ratio for water vapor, which will be discussed further below (under **Issues**).

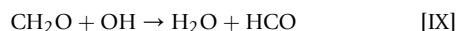
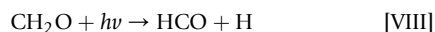
Once the source molecules are in the stratosphere, they are carried by the circulation to higher altitudes and latitudes. In the mid-stratosphere, the methyl radical, CH_3 , is formed by oxidation processes ([I], [II], and [III] below) and then may be converted to formaldehyde, CH_2O , by schemes [IV], [V], and [VI], in the presence of chlorine and nitrogen oxides.



Note that because of the distributions of OH, O, and Cl with height, these reactions are more effective in the middle and upper stratosphere than they are nearer the tropopause. This means that little methane is converted to water vapor or molecular hydrogen in the lower stratosphere. We will consider this further below.



The formaldehyde so formed may be photolysed, or may react with OH, as in reactions [VII], [VIII] and [IX] below.



The formaldehyde produced eventually forms H_2 and H_2O , and so the sequence from methane to water vapor (and to molecular hydrogen) is established. Also, if HO_x , HCl or HCO is produced from CH_4 , it is quickly converted to H_2O , much faster than it is produced from the methane, so that these particular processes are also essentially a means of producing H_2 and H_2O from CH_4 .

Observations of Hydrogen-Containing Constituents in the Stratosphere

Satellite Observations of H_2O and CH_4

The near-global perspective offered by satellites has been used in the study of the total hydrogen budget of the stratosphere. Not only are satellite instruments capable of near-global observations, but they are also, in principle, capable of making measurements of many atmospheric constituents simultaneously, and of the way they move around. In fact, water vapor and methane, being polyatomic molecules, have active infrared spectra that provide a mechanism for remote measurement of concentrations from space, either by measurement of thermal

emission or by solar absorption at the frequencies of the relevant vibration-rotation bands. Molecular hydrogen does not, however exhibit an allowed infrared spectrum, and has not yet been measured from space.

To date, three satellite projects have provided sufficient data to make a test of the hydrogen budget possible. These are the Nimbus 7 and the Upper Atmosphere Research Satellite projects, and the ATMOS Space Shuttle instrument, all from NASA in the USA (a useful web page listed under Further Reading will allow the viewer to investigate the details of past, present, and future NASA satellite missions). On Nimbus 7, launched in 1978, the Limb Infrared Monitor of the Stratosphere, LIMS, provided first infrared measurements of water vapor, while the Stratospheric and Mesospheric Sounder, SAMS, made measurements of methane. On UARS, the Halogen Occultation Experiment, HALOE, produced a decade of measurements of both water vapor and methane, simultaneously and precisely bore-sighted (i.e., colocated), which have provided copious quantities of data with which to test theory. Also, the ATMOS experiment, a high-resolution solar tracking Fourier transform spectrometer, has been flown on a number of Space Shuttle missions, and has provided complete spectral coverage in the infrared from about 3 to 16 microns, at high spectral resolution.

Each of these experiments has been used to study the distribution throughout the stratosphere of H_2O and CH_4 , and the ratio of the changes of mixing ratio of these two constituents with height, $R = \Delta\text{H}_2\text{O}/\Delta\text{CH}_4$ is a parameter which can provide a useful test of photochemical and dynamical processes that might determine the hydrogen budget of the stratosphere, as we will discuss further below. Values in the range $R = 1.5$ to -2.0 have been found.

The sum of 'total hydrogen' mixing ratio, $\psi = 2 \times \text{CH}_4 + \text{H}_2\text{O} + \text{H}_2$, has also been examined using these satellite data, or at least that part of ψ that can be measured, i.e., $\psi^* = 2 \times \text{CH}_4 + \text{H}_2\text{O}$, (so that $\psi = \psi^* + \text{H}_2$). It has usually been assumed in such studies that the mixing ratio of H_2 is constant at 0.5 ppmv. Figure 1 shows a result for ψ^* obtained from HALOE data, for the stratosphere and mesosphere above 10 hPa. This and other work (see later) indicate that this parameter is, indeed rather constant in the stratosphere, and took values in the 1990s in the range 6.0 to 7.5 ppmv, though there does

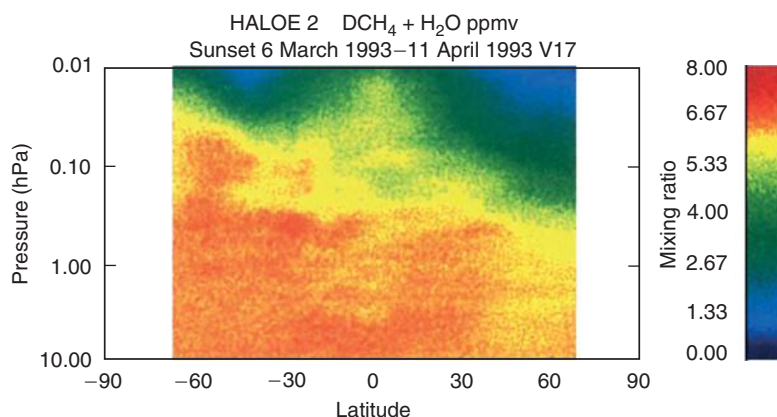


Figure 1 Height-latitude cross section of HALOE measurements of $\psi^* = 2 \times \text{CH}_4 + \text{H}_2\text{O}$, a proxy measure of total hydrogen. Units are ppmv.

seem to be a significant trend with time, according to the SPARC report on upper-tropospheric and stratospheric water vapor (see [Further Reading](#)). The value of ψ^* starts to fall significantly from a constant value above about 0.1 hPa, where rapid production of H_2 in preference to H_2O takes over. Work by the author of this article and his colleagues has shown that, on the assumptions that the total hydrogen budget is constant, and that water vapor, methane, and molecular hydrogen are the only significant components, such measurements of ψ^* may be used to derive the distribution of H_2 , particularly in the mesosphere, where it varies significantly (see [Further Reading](#)).

Aircraft and Balloon Measurements of H_2O , CH_4 and H_2

Very many observations of water vapor have been made from balloon and aircraft platforms, far too many to review here (see the SPARC assessment listed in [Further Reading](#)). Fewer, though still many, measurements have been reported of CH_4 , and still fewer of H_2 . Those measurements in which all three species have been measured together are very few! However, because of the near-constancy of the H_2 mixing ratio in the stratosphere, measurements of just water vapor and methane have proven valuable. The advantage of local measurements over satellite measurements is, of course, that they are sensitive to smaller spatial and time scales than are satellites, so that more detailed processes may be studied. Also, many of the sensors used on aircraft and balloons, especially some of the *in-situ* sampling sensors, are capable of higher relative and absolute accuracy than are satellite sensors. One example of the use of aircraft measurements of H_2O and CH_4 is shown in [Figure 2](#). This shows scatter plots for four different flights of

an aircraft at between 17 and 20 km, in the latitude range $15\text{--}40^\circ\text{N}$, in 1993. Water vapor was measured using a photo-fragment fluorescence sensor, methane by a tunable diode laser instrument. Both measurements are local. The data are compared with a line obtained by linearly fitting the data from all four flights, which has a gradient of $m = \Delta\text{H}_2\text{O}/\Delta\text{CH}_4 = -1.94 \pm 0.27$. This is close to the value of $R = -2.0$ expected if oxidation of methane goes completely to water vapor, and if the production and loss of molecular hydrogen is in balance. This will be discussed further below.

Other balloon and aircraft instruments have included infrared sensors, cryogenic trapping followed by laboratory analysis, gas chromatography, mass spectrometry, resonance fluorescence techniques, and frost-point hygrometry.

Some Issues about the Hydrogen Budget of the Stratosphere

Entry-Level Mixing Ratio of Water Vapor

There has been some controversy over the so-called mean 'entry level' of water vapor. Values based on Nimbus 7 data of 2.7 and 3.25 ppmv have been reported, while a value of 4.2 ± 0.5 ppmv has been reported from aircraft data. However, the picture is confused by the fact that the value may actually have changed with time (reports of increases of about 1% year $^{-1}$ have been published, i.e. about 10% decade $^{-1}$, a very significant change), and because only rather limited measurements have been reported near the tropopause. What is beyond disagreement now is that such a global mean value for the entry-level water vapor is merely an average over quite a range, probably from as

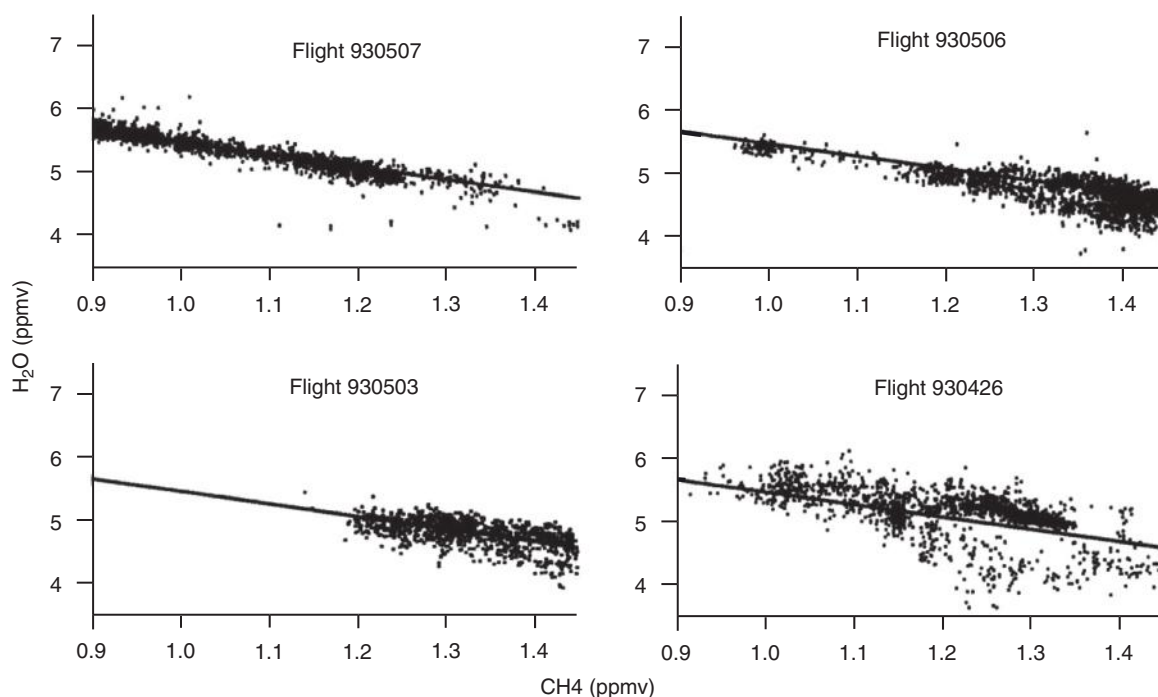


Figure 2 Airborne measurements (17–20 km altitude, $15\text{--}40^\circ\text{N}$) of H_2O and CH_4 made in 1993. The results of four different flights are shown, and the solid line is the result of a straight-line fit to all the data.

low as 2 ppmv to as high as 7 ppmv in different regions: moreover, there are probably a number of different processes that control the transport of water vapor in particular from the troposphere to the stratosphere (see **Stratospheric Chemistry Topics: Stratospheric Water Vapor**), so that the globally averaged view cannot resolve these important individual processes.

The Conversion of Methane to Water Vapor

An important question is that of the ratio of $R = \Delta\text{H}_2\text{O}/\Delta\text{CH}_4$, i.e. the increase in water vapor with respect to the decrease in methane with height. It is important in part because it is a ratio we may hope to test from observations. Is this ratio $R = -2.0$, which would imply that all the hydrogen in the methane, including that converted into intermediate compounds like CH_3 and CH_2O , has been converted to water vapor, and also that the H_2 mixing ratio has not changed significantly, at different heights? In other words, a value of $R = 2.0$ reflects a balance between large production and loss rates for molecular hydrogen. Reaction [VII] above is the principal source of additional H_2 in the stratosphere. However, H_2 can also be destroyed in the stratosphere by a number of reactions with OH , $\text{O}(^1\text{D})$, and Cl . The reaction rates for these are similar to those for reactions [I] and [III]. It is thought that the balance of these formation and destruction reactions for H_2 is such that the H_2 mixing ratio does not change much with height, at least from the tropopause to 40 or 50 km, which is the domain of our concerns here.

Thus, the ratio $R = -2.0$ in those regions where the oxidation of methane is rapid: it is less than $R = -2.0$ (as low as $R = -1.6$ has been suggested) in the lower stratosphere, but this is where methane oxidation is much slower, so that the precise value of the ratio is not so important. If, in addition, as seems likely, the H_2 profile is constant with height because production and loss mechanisms are roughly in balance, then in practice we would expect $R = -2.0$ to hold virtually everywhere where significant conversion from CH_4 to H_2O occurs. There appears to have been some disagreement in the literature about the value of R , but this has arisen largely over a misunderstanding of the relative insignificance of methane oxidation near the tropopause, where the ratio is formally less than 2.0, but where the effect on the measured profiles is small.

The Total Hydrogen Content of the Stratosphere

Finally, in this section, we ask what is the total amount of hydrogen in the stratosphere? Ignoring the minor species such as CH_2O , OH , and so on, we address the parameter $\psi = 2 \times \text{CH}_4 + \text{H}_2\text{O} + \text{H}_2$, which on the basis of current theory should be constant, unless some unknown significant sources or sinks exist. Estimates of this quantity based on observations vary from ~ 6.5 to ~ 7.0 ppmv from satellite observations from the Nimbus 7 (1979) and the ATMOS experiments (early 1980s), to 8.1 ± 0.6 ppmv from aircraft measurements between 15° and 40° N, made in 1993. These

observations were made at different epochs, and since there is known to have been a long-term upward trend in tropospheric methane, and an upward trend in stratospheric water vapor (of about $1\% \text{ year}^{-1}$) between about 1980 and the present, it is possible that these differences may be due to real changes. However, the uncertainties due both to experimental error and to variability and different sampling are probably also large enough to account for these differences. For the present we must adopt a value of $\psi = 2 \times \text{CH}_4 + \text{H}_2\text{O} + \text{H}_2$ in the range 6.5 to 8.0 ppmv. More accurate measurements are needed to distinguish a real change of the hydrogen budget.

See also: Chemistry of the Atmosphere: Methane; Observations for Chemistry (*In Situ*): Water Vapor Sondes; Observations for Chemistry (Remote Sensing): IR/FIR (Satellite, Balloon and Ground); Observations for Chemistry (Remote Sensing): Microwave. **Global Change:** Upper Atmospheric Change. **Lidar:** Resonance. **Mesoscale Meteorology:** Overview. **Middle Atmosphere:** Planetary Waves; Quasi-Biennial Oscillation. **Satellites and Satellite Remote Sensing:** Water Vapor. **Stratospheric Chemistry Topics:** Stratospheric Water Vapor.

Further Reading

- Bates, D.R., Nicolet, M., 1950. The photochemistry of atmospheric water vapor. *Journal of Geophysical Research* 55, 301–327.
- Brewer, A., 1949. Evidence for a world circulation provided by the measurements of helium and water vapor distributions in the stratosphere. *Quarterly Journal of the Royal Meteorological Society* 75, 351.
- Brasseur, G., Solomon, S., 1984. *Aeronomy of the Middle Atmosphere*. Reidel, Dordrecht.
- Dessler, A.E., Weinstock, E.M., Hints, E.J., et al., 1994. An examination of the total hydrogen budget of the lower stratosphere. *Geophysical Research Letters* 21, 2563–2566.
- Gunson, M.R., Farmer, C.B., Norton, R.H., et al., 1990. Measurements of CH_4 , N_2O , CO , H_2O and O_3 in the middle atmosphere by the ATMOS experiment of Spacelab 3. *Journal of Geophysical Research* 95, 13867–13882.
- Harries, J.E., Ruth, S., Russell, J.M., 1996. On the distribution of mesospheric molecular hydrogen inferred from HALOE measurements of H_2O and CH_4 . *Geophysical Research Letters* 23, 297–300.
- Holton, J., 1992. *An Introduction to Dynamic Meteorology*. Academic Press, San Diego.
- Jones, R.L., Pyle, J.A., Harries, J.E., et al., 1986. The water vapour budget of the stratosphere studied using LIMS and SAMS satellite data. *Quarterly Journal of the Royal Meteorological Society* 112, 1127–1143.
- Le Texier, H., Solomon, S., Garcia, R.R., 1988. The role of molecular hydrogen and methane oxidation in the water vapour budget of the stratosphere. *Quarterly Journal of the Royal Meteorological Society* 114, 281–295.
- NASA: Upper Atmosphere research Satellite web site: <http://uarsfot08.gsfc.nasa.gov/>
- Earth Science Enterprise Programme: <http://www.earth.nasa.gov/>
- Salby, M.L., 1996. *Fundamentals of Atmospheric Physics*. Academic Press, San Diego.
- SPARC Assessment of Upper Tropospheric and Stratospheric Water Vapor (2000) World Climate Research Programme Report No. 113. (The author is particularly indebted to the authors of this report, which has provided very valuable background in writing this article.)

Reactive Nitrogen (NO_x and NO_y)

Y Kondo, The University of Tokyo, Tokyo, Japan

© 2003 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 5, pp 2193-2202, © 2003, Elsevier Ltd.

Introduction

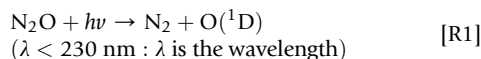
Reactive nitrogen (NO_y) plays important roles in controlling the abundance of stratospheric ozone. In this article, sources and sinks of NO_y are first described, together with the resulting NO_y distributions. Then, the role of NO_x, which is the most reactive form of NO_y, is explained. NO_x destroys catalytically stratospheric ozone and couples with other radical mechanisms. Photochemical processes controlling the relative abundance of component species of NO_y determine the NO_x abundance. The importance of heterogeneous reactions on sulfate aerosol is described in comparison with gas phase chemistry, typical for midlatitudes. In polar regions, sunlit conditions are very different from those at midlatitudes both in winter and summer. Behaviors of NO_x and NO_y under these extreme conditions are also explained.

NO_y

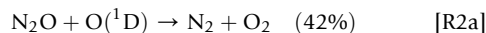
Sources and Sinks

Reactive nitrogen in the stratosphere is comprised of several component species: NO (nitric oxide), NO₂ (nitrogen dioxide), NO₃ (nitrogen trioxide), N₂O₅ (dinitrogen pentoxide), HNO₃ (nitric acid), HO₂NO₂ (peroxynitric acid), ClONO₂ (chlorine nitrate), and BrONO₂ (bromine nitrate). The sum of these species is defined as total reactive nitrogen NO_y. Namely, NO_y = NO + NO₂ + NO₃ + 2N₂O₅ + HNO₃ + HO₂NO₂ + ClONO₂ + BrONO₂. The reactions among NO_y component species do not lead to a net change in NO_y abundance.

N₂O, which is produced by bacteria in soil and released into the atmosphere, is the primary source of stratospheric NO_y. Since N₂O is very stable in the troposphere, it is transported into the stratosphere, mainly through the tropical tropopause. In the 1990s, the tropospheric concentration of N₂O was about 310 ppbv and increased at about 0.25% y⁻¹ over the 1978–96 period, due to imbalance between the global sources and sinks of N₂O. The stratosphere is a net photochemical source of NO_y and a net photochemical sink of N₂O. About 90 percent of the total loss of N₂O in the stratosphere occurs via its photolysis by ultraviolet radiation yielding N₂ and excited atomic oxygen (O(¹D)), which does not lead to NO production.



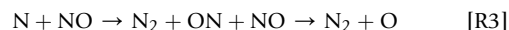
Reactions with O(¹D)) are responsible for 10% loss of N₂O.



where O(¹D)) is produced primarily by the photolysis of ozone at λ < 325. Reactions [R1], [R2a] and [R2b] occur in the middle

and upper stratosphere (at 35–45 km), mainly in the tropics where the intensity of the solar ultraviolet radiation is greatest, as shown in Figure 1.

NO_y is produced primarily via reaction [R2b]. Downward transport from the mesosphere and thermosphere, where NO is produced by solar radiation and auroral ionization, can provide an additional NO_y input to the upper stratosphere. Production of NO by lightning associated with upward transport through the tropical tropopause may be an additional NO_y source in the lower stratosphere at low latitudes, although its magnitude is poorly understood. Supersonic aircraft flying at much higher altitudes than subsonic aircraft injects NO molecules directly into the stratosphere. However, NO emissions from the currently operational Concorde are much smaller than the natural sources. There is also some emission, directly into the stratosphere, by long haul subsonic passenger aircraft, especially on flights routed over high latitudes. The net loss of NO_y occurs in the upper stratosphere and lower mesosphere, where NO is reduced into N₂ via the following reaction:



N is produced by the photolysis of NO by ultraviolet radiation. NO_y is also lost through transport of NO_y down to the troposphere where HNO₃ dissolves in water droplets and is removed from the atmosphere by precipitation.

Distribution of NO_y

Remote spectroscopic measurements by the Atmospheric Trace Molecule Spectroscopy (ATMOS) and Mk IV instruments from the Jet Propulsion Laboratory on board the space

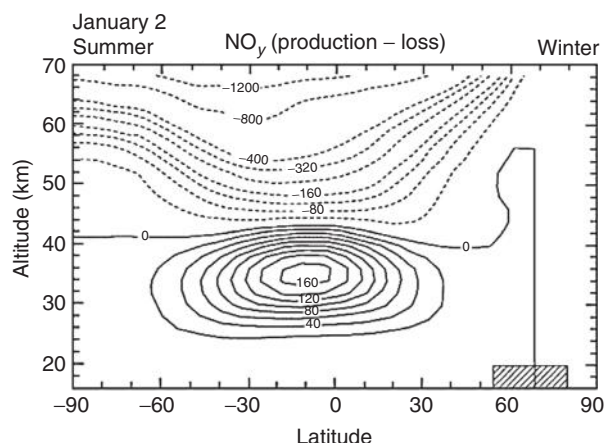


Figure 1 Contours of the local instantaneous value at noon of the net production rate of NO_y (NO_y (production–loss)), in units of 10^{–8} ppbv s^{–1}. The rates depend on the abundance of N₂O, O₃, and other species, rate coefficients of reactions [R1], [R2a], [R2b] and [R3], and solar radiation field. Reproduced with permission from Fahey et al. (1990).

shuttle and balloons, in situ measurements on board the NASA ER-2 aircraft, and balloon experiments, in addition to many other measurements, have provided extensive data on the distributions of NO_y species, together with N₂O. NO_y mixing ratios obtained by these measurements increase with altitude from the tropopause up to about 35 km, where values peak at about 18 ppbv (parts per billion by volume) at midlatitudes as shown in Figure 2. This increase is due to an increase in the NO_y production via reaction [R2b] and to NO_y loss through the tropopause. In contrast, N₂O mixing ratios decrease with altitude due to reactions [R1], [R2a] and [R2b] as shown in Figure 2. The NO_y mixing ratio decreases with altitude above 35 km due to NO_y loss by reaction [R3] and a decrease in the NO_y production rate at low N₂O mixing ratios. In the lower stratosphere, reaction [R3] is very slow and the lifetime of NO_y defined by this loss process is longer than 10 years.

Given the increase (decrease) of the NO_y (N₂O) mixing ratios with altitude up to 35 km, NO_y is anticorrelated with N₂O for N₂O values larger than 120 ppbv, as shown in Figure 3 and Figure 4. The relationship between NO_y and N₂O obtained by measurements on board the ER-2 in the lower stratosphere is expressed as

$$[\text{NO}_y] - [\text{NO}_y]_0 = 0.07([\text{N}_2\text{O}]_0 - [\text{N}_2\text{O}]) \quad [1]$$

Here [NO_y] and [N₂O] are the mixing ratios of NO_y and N₂O in ppbv. [NO_y]₀ and [N₂O]₀ are values at the tropical tropopause, which are 0.25 and 310 ppbv, respectively. This relationship indicates that the increase in NO_y is proportional to the decrease in N₂O with a constant slope of −0.07. Namely, 3.5% of N₂O molecules lost via reactions [R1], [R2a], [R2b] and [R3] are converted to NO_y. In contrast, the NO_y mixing ratio

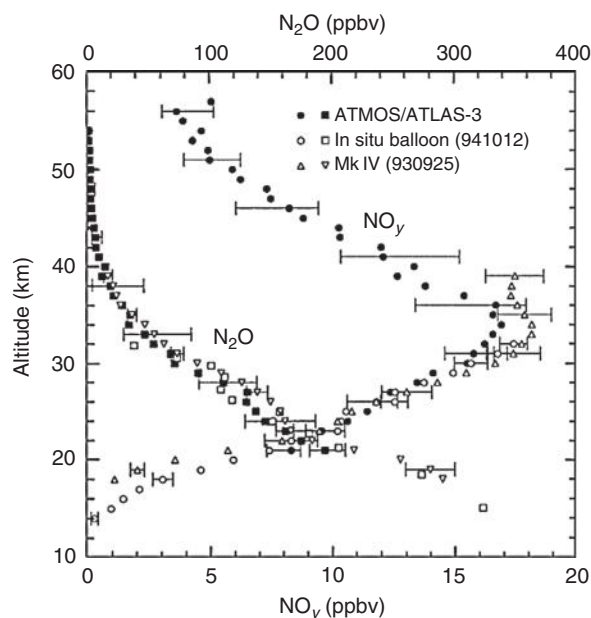


Figure 2 Profiles of NO_y and N₂O observed by ATMOS at 39–49° N in November 1994, Mk IV at 35° N in September 1993, and in situ balloon-borne measurements at 44° N in October 1994 at northern midlatitudes.

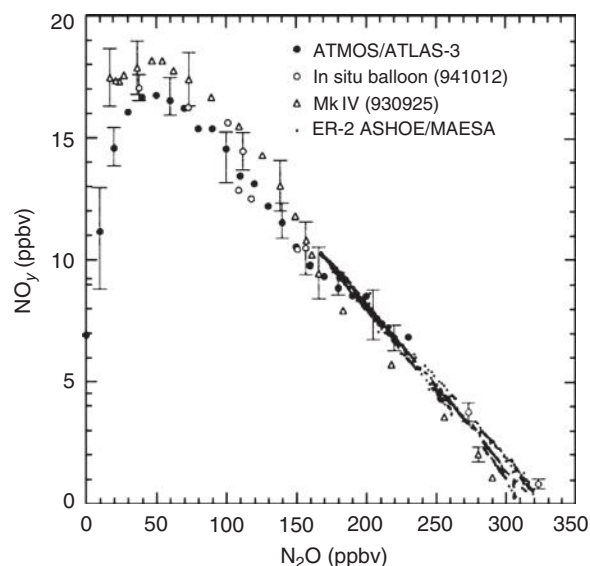


Figure 3 Correlation between NO_y and N₂O. The data used are the same as shown in Figure 2. In addition, the data obtained by the ER-2 measurements at 30–40° N in February and November 1994 are also shown.

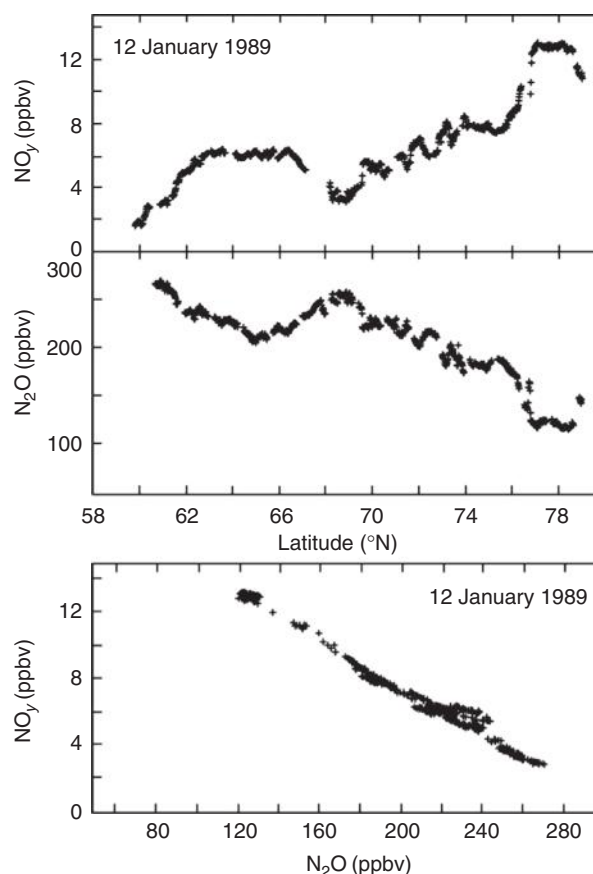


Figure 4 NO_y and N₂O mixing ratios obtained around 20 km from the north-bound leg of the ER-2 flight on 12 January 1989 in the lower Arctic stratosphere. Reproduced with permission from Fahey et al. (1990).

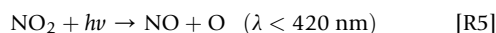
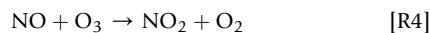
decreases along with the decrease in the N₂O mixing ratio above 35 km where the N₂O values are lower than 40 ppbv, leading to a positive correlation between NO_y and N₂O as shown in Figure 3.

Due to its long lifetime, NO_y produced mainly in the tropical upper stratosphere is transported to higher latitudes and lower altitudes. This large-scale transport process is known as the Brewer–Dobson circulation. All long-lived trace gases, including ozone, N₂O, and CH₄, are subject to this transport. As a consequence, the NO_y (N₂O) mixing ratios in the lower stratosphere are higher (lower) at higher latitudes, as can be seen from Figure 4. In general, compact correlations have been observed between long-lived species whose local photochemical lifetimes exceed the time scales for atmospheric transport as predicted theoretically. The NO_y–N₂O correlation shown in Figure 3 and Figure 4 is consistently compact, especially in the lower stratosphere, where the N₂O mixing ratios are higher than 120 ppbv. This correlation has proved to be very useful in predicting NO_y abundances from observed N₂O mixing ratios as described below.

Gas Phase Chemistry

Role of NO_x

A scheme showing important reactions controlling the level of each reactive nitrogen species is shown in Figure 5. Altitude profiles of NO_y species observed by Mk IV at 35° N in September 1993 are shown in Figure 6. The time constants for the photolysis of NO_y species for local noon at 44° N in October are shown in Figure 7). NO and NO₂ are the most reactive among NO_y species. NO is oxidized to NO₂ by ozone and NO₂ is photolyzed by visible sunlight to reform NO and atomic oxygen (O).



NO and NO₂ are often treated as a sum, defined as NO_x, because the time required for the exchange between NO and NO₂ during daytime is about 1 min (Figure 7).

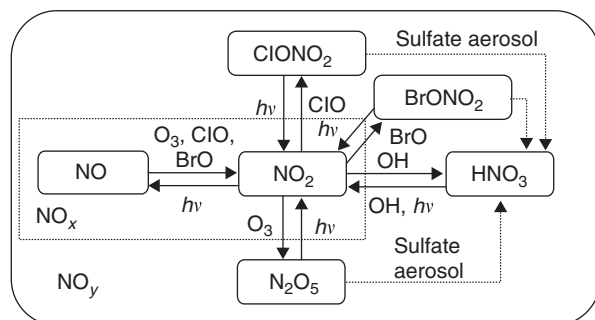


Figure 5 Schematic of the reaction pathways between the principal NO_y component species in the lower stratosphere. Photolysis reactions are indicated by $h\nu$. 'Sulfate aerosol' denotes heterogeneous reactions on sulfuric acid aerosol particles. Reproduced with permission from Gao et al. (1999).

NO_x plays important roles in controlling stratospheric ozone. First, NO_x destroys ozone catalytically via the following reactions:

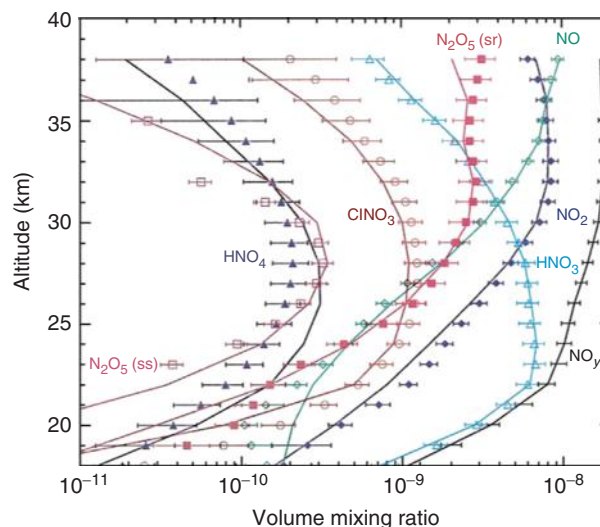
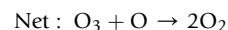
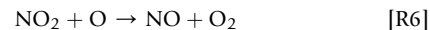
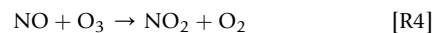


Figure 6 Observed (symbols) and calculated (lines) profiles of NO_y species, as indicated, for sunset at 35° N on 25 September 1993. Sunrise profiles for N₂O₅ are also shown. The NO_y profile represents the sum of nitrogen oxides measured by Mk IV and was used to constrain the model. The model calculation used JPL 2000 kinetic data. Reproduced with permission from Sen et al. (1998), modified for using updated model calculations by RJ Salawitch.

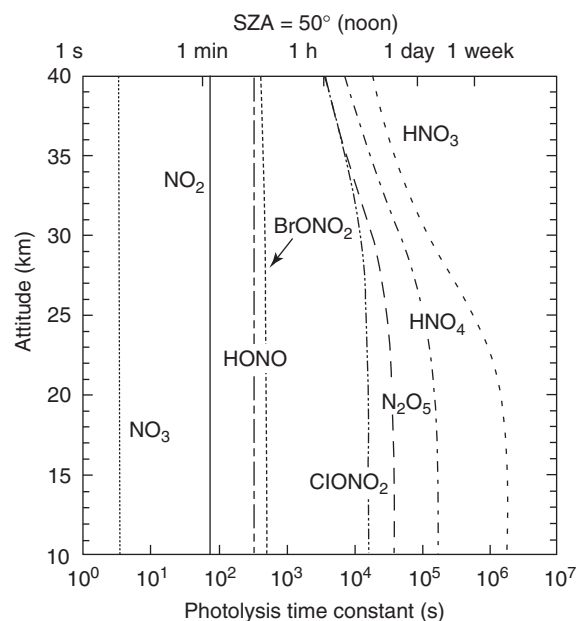
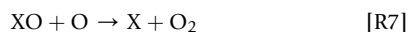


Figure 7 Time constant of the NO_y species due to the photolysis between 15 and 30 km altitude, for noon at 44° N on 12 October. Courtesy of MY Danilin.

The cycle is catalytic since NO_x is conserved: NO and NO₂ are simply interchanged. Reaction [R6] is rate determining for the catalytic cycle since reaction [R4] proceeds faster than reaction [R6] in the stratosphere. The ozone loss rate is therefore proportional to the product of the NO₂ and O concentrations ([NO₂][O]). Reactions analogous to [R4] and [R6] also represent catalytic ozone loss cycles by reactive hydrogen (HO_x), chlorine (ClO_x), and bromine (BrO_x) species if NO is replaced with OH, Cl, and Br, respectively. Similar to the NO_x catalytic cycle, the rate-determining reaction for the HO_x, ClO_x, and BrO_x cycles are



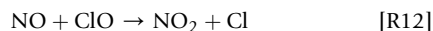
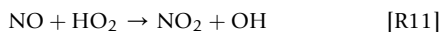
where X is OH, Cl, or Br, respectively. Ozone loss rates are therefore proportional to the product of the concentrations of XO and O ([XO][O]).

Secondly, NO_x buffers the ozone loss by HO_x, ClO_x, and BrO_x by converting OH, ClO, and BrO into HNO₃, ClONO₂, and BrONO₂, which do not destroy the ozone directly, via the following reactions:



Here, the third body M represents the major atmospheric molecules N₂ and O₂. Hence, the relative importance of catalytic loss cycles of ozone by HO_x, ClO_x, and BrO_x is strongly dependent upon the NO_x abundance.

NO_x buffers HO_x and ClO_x catalytic cycles also by the following interchange reactions:



Reactions [R11] and [R12] decrease the HO₂ and ClO levels, respectively, by shifting the HO₂/OH and ClO/Cl ratios. The reductions in HO₂ and ClO lead to decreases in ozone loss rates, which are proportional to [HO₂][O] and [ClO][O], as mentioned above. On the other hand, these reactions increase the NO₂/NO ratio, enhancing the ozone loss rate by the NO_x cycle. Profiles of the ozone loss rates by the NO_x, HO_x, ClO_x, and BrO_x cycles at 35° N in September are shown in Figure 8.

Oxidation of NO_x

NO_x levels are also controlled by chemical processes that lead to the production and loss of NO_x as detailed below. NO_x produced by reaction ([R2b]) is converted to higher oxides of nitrogen (N₂O₅, HNO₃, HO₂NO₂, ClONO₂, BrONO₂). Since these NO_y species do not react directly with ozone but produce NO_y by photolysis and reactions with OH, they are called reservoir NO_y species. N₂O₅ is produced through the following reactions:

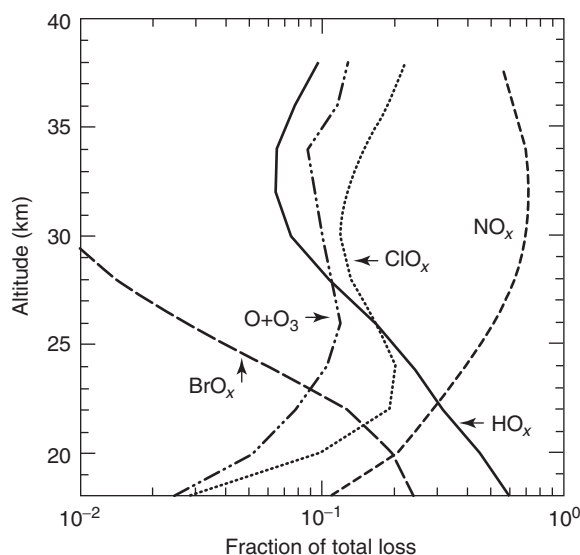
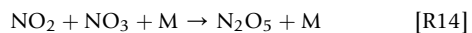
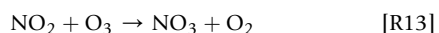
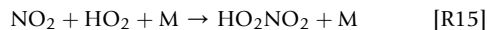


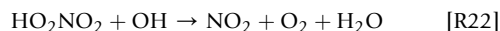
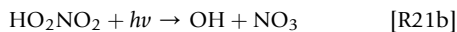
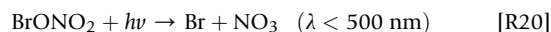
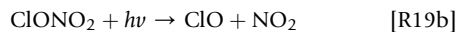
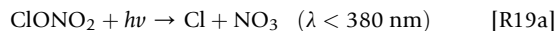
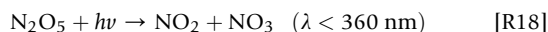
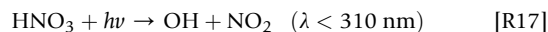
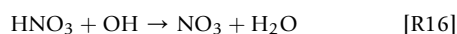
Figure 8 Odd oxygen sinks at 35° N on 25 September 1993. The fractional contribution of the dominant sinks and the diurnally averaged loss rate of odd nitrogen, computed using constraints imposed by the MK IV data. Losses due to each catalytic cycle are indicated as NO_x, HO_x, ClO_x, and BrO_x. O₃ + O denotes loss by recombination reaction of odd oxygen. Heterogeneous reactions included in the model calculations increase the contributions from HO_x, ClO_x, and BrO_x. Courtesy of RJ Salawitch and B Sen.

Reaction [R13] is rate determining for the formation of N₂O₅, which occurs only during nighttime since NO₃ is photolyzed within a few seconds by visible radiation ($\lambda < 670$) during daytime (Figure 7).

Other reservoir NO_y species are produced via reactions [R8], [R9] and [R10] and the following reaction:



Conversely, NO_x is produced by decomposition of reservoir species, such as



Here the photolysis wavelength thresholds are given for absorption cross-section limits of about $1 \times 10^{-21} \text{ cm}^2$.

Typical lifetimes of N₂O₅ and HNO₃ in the lower stratosphere as determined by the above decomposition processes are several hours and 1 week, respectively, for noontime

midlatitude fall conditions, as shown in Figure 7. Below 25 km, HNO₃ is the dominant NO_y species since NO_x is oxidized to form HNO₃ via the three-body reaction [R8], which proceeds faster at lower altitude (higher pressure) as shown in Figure 6. At higher altitudes, NO_x dominates among the NO_y species due to enhanced NO_x production by reactions [R17], [R18], [R19a] and [R19b]. The 2 × N₂O₅ mixing ratio at sunrise becomes comparable to NO_x in the lower stratosphere. The ClONO₂ mixing ratio shows a broad peak of about 1 ppbv centered around 25 km.

Diurnal and Seasonal Variations at Midlatitudes

NO_y species undergo temporal variations depending on their chemical lifetimes as shown in Figure 9. NO is oxidized to NO₂ within a few minutes after sunset and NO_x exists in the form of NO₂ during nighttime. Part of the NO₂ is photolyzed to produce NO soon after sunrise. Due to the formation of N₂O₅ during the nighttime and photolysis during the daytime, N₂O₅ mixing ratios reach maximum and minimum values at sunrise and sunset, respectively, as is partly shown by Mk IV observations (Figure 6). Corresponding to the diurnal variation of N₂O₅, NO_x shows a slow increase in the morning reaching a maximum value at sunset, when it starts to decrease again until sunrise. In contrast, HNO₃ does not undergo significant diurnal variation since the lifetime of HNO₃ is about a week as described above. Both N₂O₅ and HNO₃ levels respond to the seasonal variations of solar elevation and sunlit hours. The HNO₃/NO_y and N₂O₅/NO_y ratios at midlatitudes reach minimum values at summer solstice due to the largest rates of reactions [R17] and [R18]. Similarly, they reach maximum values near the winter solstice. The NO_x mixing ratio and NO_x/NO_y ratio show almost sinusoidal seasonal variations reaching maximum and minimum values at summer and winter solstices, respectively, as shown by ground-based and satellite remote sensing observations.

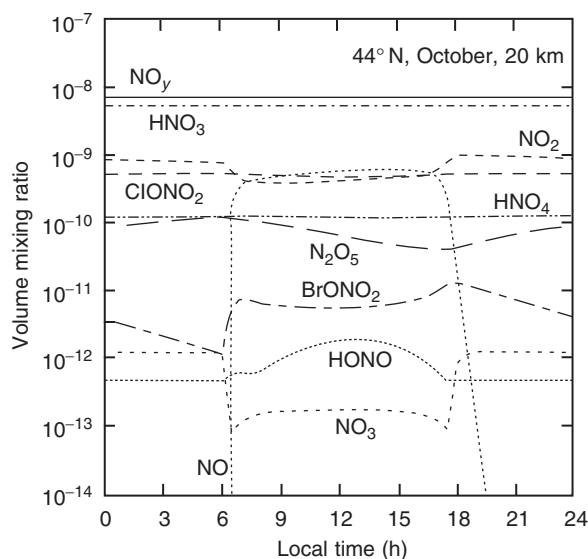
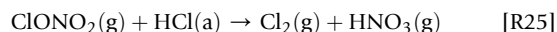
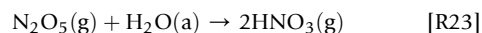


Figure 9 Diurnal variation of odd-nitrogen species in the stratosphere at 20 km altitude calculated at 44° N for October 1994 conditions. Courtesy of MY Danilin.

Heterogeneous Chemistry

Sulfate aerosol composed of liquid sulfuric acid (H₂SO₄) is ubiquitous in the lower stratosphere from the tropics to the polar regions. H₂SO₄ in the stratosphere is produced through oxidation of sulfur-containing gases transported from the troposphere mostly in the form of carbonyl sulfide (OCS). In addition to the gas phase chemistry described above, the partitioning of NO_y is also controlled by the heterogeneous reactions on sulfate aerosols listed below:



Here (g) and (a) denote species in gas phase and in aerosol, respectively.

Reaction [R23] occurs with a reaction probability of 0.1, weakly dependent on the composition of aerosols, temperature, and particle size. On the other hand, the effects of reactions [R24] and [R25] are important only at very low temperatures in high-latitude winter (temperature <210 K). Reaction [R26] is fast, but its impact on nitrogen species partitioning in the stratosphere is smaller than that due to reaction [R23] because of the smaller bromine content (~10 pptv (parts per trillion by volume)).

These reactions, especially reaction [R23], convert shorter-lived reservoir N₂O₅, ClONO₂, and BrONO₂ into longer-lived HNO₃, even at low and midlatitudes. They lengthen the time required to regenerate NO_x via reactions ([R18], [R19a], [R19b] and [R20]), resulting in an effective decrease in the NO_x levels (Figure 5). In addition, reaction [R23] oxidizes NO_x without consuming OH, resulting in a higher OH level. The higher OH abundance accelerates reaction [R8] causing further reduction in NO_x. The reduction of the NO_x/NO_y ratio in a stratospheric model due to the effect of heterogeneous reactions is shown in Figure 10.

The reduction of the NO_x level by heterogeneous reactions leads to increases in the HO_x, ClO_x, and BrO_x levels because of the slower reaction rates of reactions [R8], [R9] and [R10] at lower NO_x concentrations. The reduction of the NO_x level by heterogeneous reactions leads to increases in the HO_x, ClO_x, and BrO_x levels because of the slower reaction rates of reactions [R8], [R9] and [R10] at lower NO_x concentrations. The nonlinear dependence of ozone loss rates on HO_x, halogen (ClO_x and BrO_x), and NO_x abundances are shown in Figure 11. At the lowest NO_x levels (left dotted lines in Figure 11), ozone loss rates increase due to increased abundances of HO_x and halogen species that result from the lowering of NO_x in an air parcel. Typical values at midlatitudes (right dotted lines in Figure 11) are near the mid-range NO_x values where ozone loss rates have low sensitivity to the abundance of NO_x.

The effect of heterogeneous reactions on NO_x and ozone is enhanced by volcanic eruptions. Associated with large volcanic eruptions, significant amounts of SO₂ are injected into the stratosphere. The injected SO₂ is oxidized to H₂SO₄, which forms subsequently sulfuric acid aerosols within a short time. Mount Pinatubo in the Philippines erupted in June 1991 and

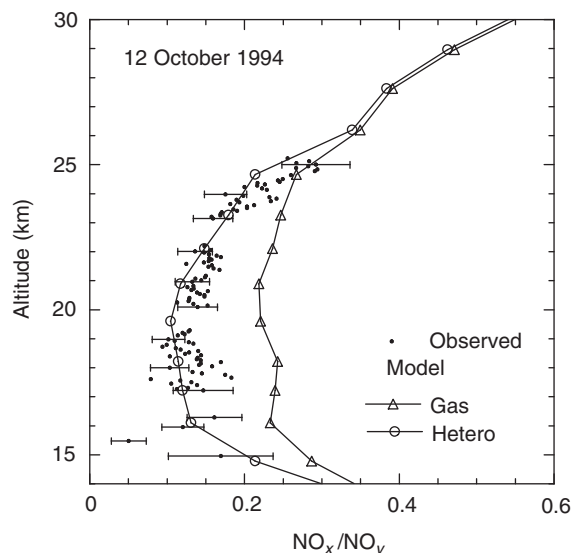


Figure 10 Vertical profiles of the NO_x/NO_y ratios (small solid circles) derived from the balloon observations made at 44°N on 12 October 1994. Bars show the total uncertainties in the NO_x and NO_y measurements. The calculated NO_x/NO_y ratios incorporating heterogeneous chemistry and gas phase chemistry only are compared. Reproduced with permission from Kondo et al. (2000), modified for using NO_x/NO_y ratio instead of NO/NO_y ratio.

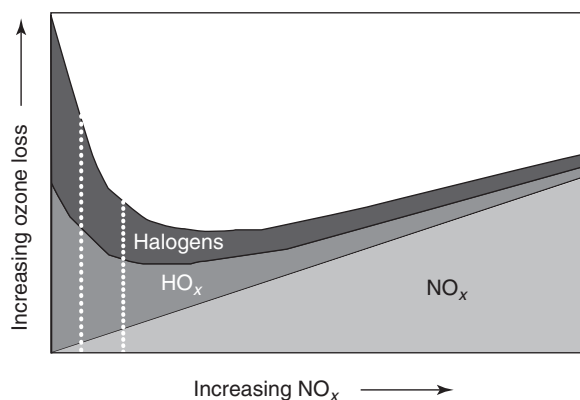


Figure 11 The O_3 removal rate versus NO_x levels. Because of the coupling that exists between the radical families, the response of the total O_3 removal rate to changes in NO_x abundance is highly nonlinear. At sufficiently low NO_x levels, such as observed at midlatitudes in May 1993, the removal rates are inversely correlated with NO_x abundance.

the aerosol loading increased by up to a factor of 100 over background values. The aerosol loading remained high for a few years with a gradual decrease with time. Corresponding to this enhanced aerosol loading, significant reductions in NO_x were observed as shown in Figure 10. It should be noted that the rate of reaction [R23] at large aerosol surface area is limited by the formation rate of N_2O_5 [R13] during the nighttime. Therefore the decrease in NO_x by reaction [R23] saturates eventually at a certain surface area, depending on photochemical conditions. The reduction in NO_x caused the decrease in the ozone levels as observed by a variety of in situ and remote sensing measurements, providing evidence of the effect of heterogeneous reactions on stratospheric chemistry.

Polar NO_x and NO_y Winter

Both NO_x mixing ratios and NO_x/NO_y ratios have been observed to decrease sharply at latitudes higher than $50\text{--}60^\circ$ in winter and early spring as shown in Figure 12. This large and sharp decrease in NO_x is caused by the large reduction of sunlight in high-latitude winter, which reduces greatly the formation of NO_x from NO_y reservoir species. N_2O_5 produced in the dark stratosphere is converted effectively to HNO_3 via reaction [R23] before being photolyzed. Therefore, HNO_3 dominates among the NO_y species in high-latitude winter.

The temperature in the Antarctic and Arctic lower stratosphere decreases to as low as 190–185 K by midwinter in the absence of solar heating. Under these very low temperatures, HNO_3 co-condenses with H_2O and/or H_2SO_4 to form polar stratospheric cloud (PSC) particles. These particles provide sites for heterogeneous reactions, such as reactions [R24] and [R25], which convert unreactive inorganic chlorine species into reactive chlorine very efficiently. The reactive chlorine destroys the ozone rapidly when exposed to sunlight. In addition, a polar vortex with westerly winds forms as the high-latitude stratosphere cools each winter season. The polar vortex isolates partially high-latitude stratospheric air from midlatitude air. Extensive ozone depletion inside the vortex in early spring over the Antarctic is well known as the Ozone Hole. Similar processes occur during the cold Arctic winters, although the temperatures in the Arctic are much warmer and show larger

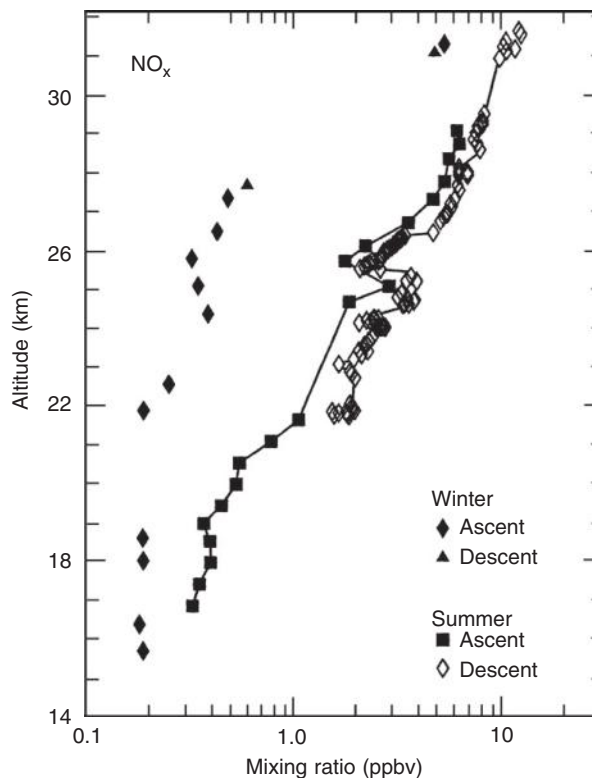


Figure 12 Contrast in the vertical distribution of NO_x in winter and summer. The balloon-borne in situ measurements were made at 51°N in August and December 1982. A strong reduction of NO_x occurred at 20–28 km. Reproduced with permission from Ridley et al. (1987).

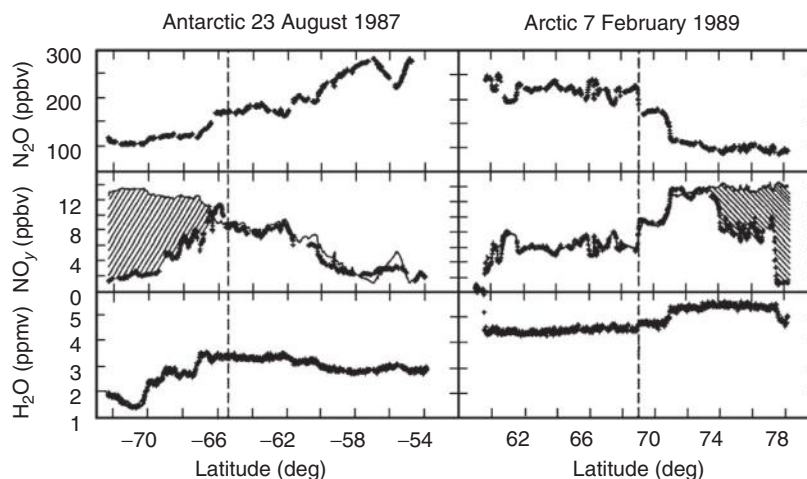


Figure 13 N_2O , NO_y , and H_2O mixing ratios observed by ER-2 at 16–19 km altitude from a portion of a flight in the Antarctic and Arctic missions. The average relation between NO_y and N_2O Equatorward of the vortex boundary (dashed line) is given by eqn [1]. The shaded areas, highlighting the difference between measured NO_y and NO_y calculated from eqn [1], represent denitrification in the sampled air masses. Reproduced with permission from Fahey et al. (1990).

year-to-year variations. In this way, HNO_3 contributes to ozone destruction through the formation of PSC particles.

With the reappearance of the Sun in early spring, NO_2 is produced from HNO_3 by reactions [R16] and [R17]. NO_2 deactivates reactive chlorine and bromine by reactions [R9] and [R10]. This process decelerates effectively the ozone destruction by halogen radicals after the formation of PSCs ceases in spring when the temperature rises above the PSC formation threshold. However, spring HNO_3 levels, and therefore NO_2 levels, are often much lower than in late fall or early winter for the following reason. HNO_3 -containing PSC particles in the crystalline form sometimes grow to larger than $10\ \mu\text{m}$ in radius under continued very low temperatures in midwinter. These particles fall out of the stratosphere in a few to several days, leading to permanent removal of NO_y . This NO_y loss process, called denitrification, lowers the level of HNO_3 , resulting in a delay in the deactivation of chlorine and extending the period of ozone depletion throughout the winter and early spring.

Denitrification has been detected as deviations of the NO_y values from those anticipated from the reference NO_y – N_2O correlation observed prior to denitrification in late fall as shown in Figure 13. Equation [1] represents a good reference for N_2O higher than 120 ppbv. Extensive denitrification occurs in the Antarctic winter, when temperatures fall persistently below even the ice saturation threshold. The temperature in the Arctic in winter is somewhat higher and much more variable than in the Antarctic, as described above, resulting in less extensive denitrification. Falling PSC particles evaporate if they experience temperatures higher than the HNO_3 – H_2O condensation threshold temperature. This leads to local enhancement in NO_y over the background value as has been observed at 12–15 km in the Arctic.

Summer

During the summer, large regions of the polar stratosphere receive uninterrupted sunlight for many weeks. Under these

conditions, daily N_2O_5 production via reactions [R13] and [R14] ceases abruptly with the onset of continuous photolysis in high-latitude air masses, because NO_3 , the intermediate in its formation, is photolyzed rapidly, thereby preventing N_2O_5 formation. Depletion of N_2O_5 shuts off the hydrolysis of N_2O_5 in the heterogeneous reaction [R23]. In addition, the photolysis of HNO_3 is augmented by continuous sunlight. The NO_y family simplifies to a near ‘gas-phase-only’ system in summer air masses because the NO_x/HNO_3 ratios become primarily controlled by reactions [R8], [R16] and [R17]. Due to these conditions, the NO_x/NO_y ratios at 18–20 km observed by the ER-2 aircraft in polar summer reach as high as 0.25, which are much higher than those at lower latitudes. It is noted that gas phase models predict NO_x/NO_y ratios close to those observed in polar summer even for midlatitude near equinox (Figure 10), thereby demonstrating the importance of reaction [R23] in determining the NO_x levels in the lower stratosphere. High NO_x abundances in polar summer have also been observed by satellite and ground-based spectroscopic measurements.

The measurements by the ER-2 aircraft of related radicals have shown the predominance of the NO_x catalytic ozone loss cycle over the HO_x , ClO_x , and BrO_x cycles in polar summer under high NO_x , as can be understood by the diagram of Figure 11. Total ozone loss rates calculated using aircraft data are as high as 10–20% per month at 18–20 km at 60–90° N in June. This ozone loss rate is consistent with that observed by satellites.

See also: Aerosols: Aerosol Physics and Chemistry. **Mesoscale Meteorology:** Overview. **Middle Atmosphere:** Polar Vortex; Transport Circulation. **Ozone Depletion and Related Topics:** Ozone Depletion Potentials; Photochemistry of Ozone. **Stratospheric Chemistry Topics:** HO_x ; Halogens. **Tropospheric Chemistry and Composition:** Hydroxyl Radical.

Further Reading

- Brasseur, G., Solomon, S., 1986. *Aeronomy of the Middle Atmosphere*. Reidel, Dordrecht.
- Dessler, A., 2000. *The Chemistry and Physics of Stratospheric Ozone*. Academic Press, London.
- Finlayson-Pitts, B.J., Pitt Jr, J.N., 2000. *Chemistry of the Upper and Lower Atmosphere*. Academic Press, London.
- Kaye, J.A., Jackman, C.H., 1994. Stratospheric ozone change. In: Hewitt, C.N., Sturges, W.T. (Eds.), *Global Atmospheric Chemical Change*. Chapman & Hall, London, pp. 123–168.
- Kolb, C.E., Worsnop, D.R., Zahniser, M.S., et al., 1995. Laboratory studies of atmospheric heterogeneous chemistry. In: Barker, J.R. (Ed.), *Progress and Problems in Atmospheric Chemistry*. World Scientific, Singapore, pp. 771–875.
- Ridley, B., Atlas, E., 1999. Nitrogen compounds. In: Brasseur, G.P., Orlando, J.J., Tyndall, G.S. (Eds.), *Atmospheric Chemistry and Global Change*. Oxford University Press, Oxford, pp. 235–287.
- Wayne, R.P., 1991. *Chemistry of Atmosphere*. Oxford University Press, Oxford.
- World Meteorological Organization (WMO), 1992. *Scientific Assessment of Ozone Depletion: 1991, Report 25*.
- World Meteorological Organization Global Ozone Research and Monitoring Project, World Meteorological Organization Global Ozone Research and Monitoring Project, Geneva.
- World Meteorological Organization (WMO), 1995. *World Meteorological Organization (WMO) (1995) Scientific Assessment of Ozone Depletion: 1994, Report 37*, World Meteorological Organization Global Ozone Research and Monitoring Project, Geneva.
- World Meteorological Organization (WMO) (1999) *Scientific Assessment of Ozone Depletion: 1998, Report 44*, World Meteorological Organization Global Ozone Research and Monitoring Project, Geneva.
- Zellner, R., 1999. Chemistry of the stratosphere. In: Zellner, R. (Ed.), *Global Aspects of Atmospheric Chemistry*. Springer, Darmstadt, pp. 181–254.

Stratospheric Water Vapor

KH Rosenlof, Earth System Research Laboratory, National Oceanic and Atmospheric Administration, Boulder, CO, USA

© Published by Elsevier Ltd.

Synopsis

Stratospheric water vapor plays a significant role in the climate system. In spite of the fact that its concentration is very small, it plays key roles in radiative and chemical processes. Additionally, its distribution can be used to assess zonally averaged transport processes as well as changes in stratospheric transport that are not directly measurable but important for understanding the evolution of the stratospheric ozone layer. Key mechanisms that contribute to the control of stratospheric water vapor concentrations are stratospheric photochemistry and transport, tropical tropopause dehydration, and polar dehydration. Transport from troposphere to stratosphere at midlatitudes also appears to impact stratospheric water vapor. Although global water vapor trend determination from the historical record is difficult due to the lack of long-term continuous measurements, there are indications that the stratospheric water vapor has increased over the past 50 years, which is to be expected given that methane emissions have increased, and oxidation of methane is the major production of water vapor in the stratosphere. However, we do not fully understand the stratospheric water vapor budget and the mechanisms that control entry into the stratosphere. It is therefore difficult to predict changes expected in a future climate, and hence the study of stratospheric water vapor remains an active area of research.

Introduction

The Earth is a planet ruled by water. Water exists in all three phases (solid, liquid, and gas) at commonly observed surface and atmospheric temperatures. It is of vital importance to the planet's ecosystem, with virtually all life that we know of depending on it. What we call weather is really the process of moving water through our ecosystem.

The water vapor content of the atmosphere is high near the surface, and drops down to extremely small values in the upper reaches of the troposphere. **Figure 1** shows a plot of the water vapor profile over Boulder, Colorado (40 °N). The concentration of water vapor in this plot is shown as a relative fraction of the total mass of air at any given altitude and is plotted as 'parts per million by volume' (ppmv). What this means is only a few molecules in every million (10^6) are water molecules. **Figure 1** is on logarithmic scale, so it shows the large dynamic range of

water from the surface to the stratosphere. The range goes from several thousand ppmv mixing ratio to less than ten ppmv at the upper reaches of the balloon flight, which is well into the stratosphere.

In the stratosphere, the water vapor concentration is very small. Nevertheless, this small concentration is profoundly important. This extreme dryness was first discovered during high-altitude research flights in Canberra aircraft over the United Kingdom, beginning as long ago as 1943 and continuing for many years thereafter. With a frost point hygrometer, a mirror surface is cooled until the ambient humidity causes a frosting of the surface: knowing the temperature at which this happens, and the local air pressure allows the humidity to be calculated. Using this device, scientists from the British Meteorological Office measured frost point temperatures at the tropopause (the boundary between the troposphere and stratosphere) of 215 K, which corresponds to a mixing ratio of about 55 ppmv. At altitudes about 2 km above the local tropopause, mixing ratios of 3 ppmv were observed. This proved to be a mystery initially, because, at the time of these measurements, the stratosphere was believed to be quiescent and in radiative equilibrium.

These extremely dry values measured in the stratosphere, coupled with observations of ozone and helium, led to the postulation that the overall circulation between the troposphere and the stratosphere was upward in the tropics across the tropopause and downward at middle-to-high latitudes. This circulation transports tropical air with tropospheric qualities into the stratosphere, and at higher latitudes, transports air with stratospheric qualities back into the troposphere. The classic 1949 paper describing this is by A. W. Brewer, and the circulation, known now as the Brewer–Dobson circulation, is shown in **Figure 2**. What Brewer realized that enabled him to deduce such a circulation existed with only vertical profiles of water over England, was that water vapor mixing ratios were extremely low in the stratosphere. How much water a parcel of air holds is a strong function of its temperature history, with colder values yielding lower values of water vapor. The only place that Brewer knew that was cold enough to achieve those

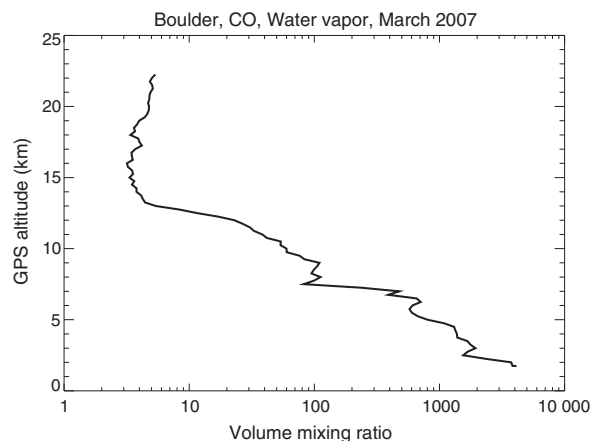


Figure 1 Water vapor profile from the balloon borne NOAA frost point hygrometer. This profile was measured over Boulder, Colorado, located at 40 °N. Data taken by the Global Monitoring Division of the NOAA Earth System Research Laboratory in Boulder, CO.

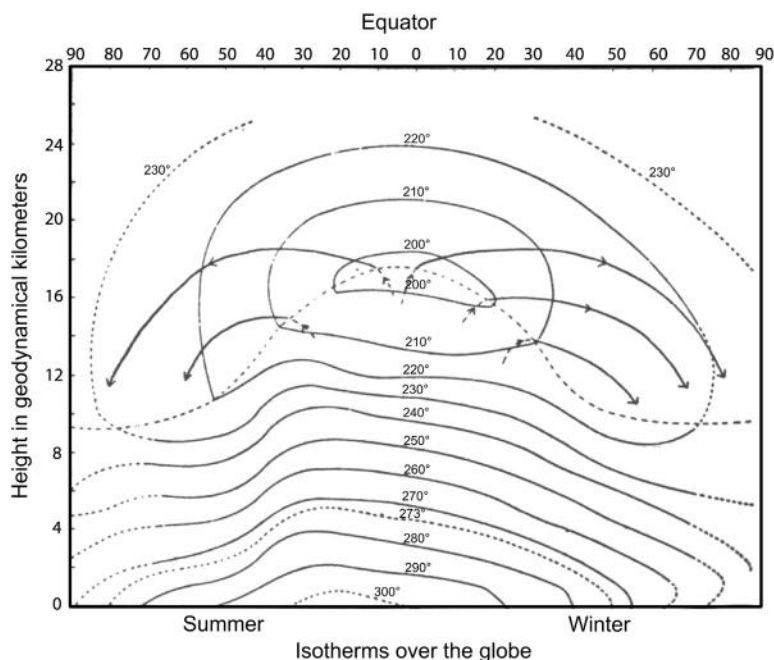


Figure 2 A contour plot of zonally averaged temperatures, taken from Brewer, A., 1949. Evidence for a world circulation provided by the measurements of helium and water vapor distributions in the stratosphere. Quarterly Journal of the Royal Meteorological Society 75, 351. The line with longer dashes indicates the tropopause. The solid lines with arrows indicate the direction of the circulation.

low mixing ratios was the tropical tropopause (see zonally averaged temperatures in Figure 2). Brewer also knew that stratospheric ozone was produced in the tropics, and that G. Dobson, in a 1929 paper, had found that the column of ozone is largest at high latitudes, hence the general flow in the stratosphere had to be poleward. Brewer's extremely clever detective work using a minimum of observations led to, what we know definitively today, the mean meridional stratospheric circulation.

It is well known that water vapor is the principal greenhouse gas and plays a key role in the tropospheric and stratospheric chemistry. Throughout the atmosphere, water vapor plays both a radiative and chemical role. An increase in stratospheric water vapor will radiatively cool the lower stratosphere and also affect the frequency of occurrence of polar stratospheric clouds, thereby impacting the stratospheric ozone chemistry. Enhanced levels of stratospheric water vapor strengthen ozone loss in the presence of ozone depleting substances (ODS). Hence, a climate with increased stratospheric water vapor will have a delayed ozone recovery even while ODSs are reduced. Changes in the stratospheric water vapor also can be a significant radiative forcing for surface climate. A $\sim 10\%$ drop in stratospheric water vapor levels near the tropopause at the end of 2000 has been estimated to have contributed a radiative forcing of -0.1 W m^{-2} ; this may have slowed the amount of warming from CO_2 increases over the 2000–09 period by 25%. Decadal variations of stratospheric water vapor concentrations also likely contributed to an enhanced rate of surface warming in the 1990s (Figure 3).

It is clear that the humidity of the stratosphere is important because the amount of water vapor determines important

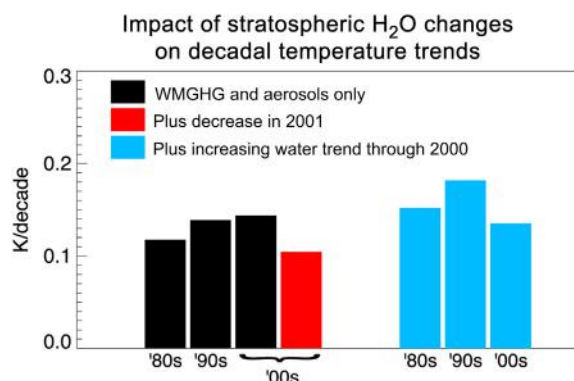


Figure 3 Decadal warming rates arising from well-mixed greenhouse gases and aerosols alone (black), as well as from including the stratospheric water decline after 2000 (red), and including both the stratospheric water vapor decline after 2000 and the water vapor increase in the 1980s and 1990s (cyan). The Bern intermediate complexity climate model was used for these estimates; it does not simulate internal variability from one year to another. Volcanoes were not been included in the radiative forcing. The climate sensitivity of the model used is 3°C for a doubling of atmospheric CO_2 , and the transient climate response is 1.7°C , slightly less than the mean of the range of models assessed by the Intergovernmental Panel on Climate Change (1). From Solomon, S., Rosenlof, K. H., Portmann, R., Daniel, J., Davis, S., Sanford, T., Plattner, G.-K., 2010. Contributions of stratospheric water vapor changes to decadal variations in the rate of global warming. Science 327, 1219–1223.

aspects of the planetary radiative energy balance through the strong cooling of space from water vapor. This contributes to determining the temperature of the stratosphere, which then affects the dynamical circulation of the upper atmosphere. Moreover, water vapor provides the source of the hydroxyl radical, OH, which takes part in a number of stratospheric chemical processes. Additionally, water vapor concentrations impact stratospheric temperature, which in turn can affect stratospheric chemistry through temperature dependent reaction rates. An understanding of how the distribution of water vapor is controlled, and of how this distribution might change in future, is therefore important for understanding certain aspects of climate change.

Water Vapor Observations

Because water exists in all three phases in the Earth's atmosphere, water vapor has proven difficult to measure at extremely high accuracy throughout the depth of the atmosphere. It is also difficult to determine how water vapor has changed in the industrial time period. After Brewer's pioneering measurements from aircraft, high-altitude balloons were used, and initially, these measurements indicated a much wetter stratosphere. However, it was soon realized that the balloon data were contaminated by moisture carried up by the balloon itself. A change in operating procedure, with measurements taken on descent and the inlet ahead of the instrument avoided the extraneous moisture from the balloon outgassing. Measurements taken by scientists at the Naval Research Laboratory in the United States and later from the National Oceanic and Atmospheric Administration (NOAA) showed that stratospheric humidity was indeed low, with small increases from the hygropause (level of minimum water) to 30 km in altitude. These balloon measurements, extending from the 1960s to present day, first over Washington DC and then Boulder, Colorado, form the backbone of our long-term record of stratospheric water vapor. In the recent years, regular frost point measurements have been made in the Southern Hemisphere (Lauder, New Zealand), the tropics and subtropics (San Jose, Costa Rica, and Hilo, Hawaii), and Europe (Lindenberg, Germany). Associated with various aircraft campaigns and intensive observation programs, there have been frost point balloons flown at other sites around the world. These sites include Indonesia, the Galapagos, and on ships as a part of the Japanese program known as SOWER (soundings of ozone and water in equatorial regions). Frost point balloon measurements have also been made associated with a number of aircraft campaigns and for satellite validation studies. The frost point balloon sonde will be used as a water vapor standard within the Global Climate Observing System Reference Upper Air Network.

Typically deployed on a campaign basis, there are instruments other than frost point hygrometers capable of measuring water at low stratospheric concentrations. A step up in sensitivity, which allowed measurements at much higher time and space resolution to be made, came with the invention of an alternative measurement technique using the resonance fluorescence hygrometer. This device uses the spectroscopic property of the water vapor molecule that if it is illuminated by

ultraviolet radiation from a lamp at a certain frequency, it will reemit radiation in measurable quantities and with an intensity that is proportional to the relative amount of water vapor to air molecules in the line of sight. This device has been widely deployed on aircraft and balloons around the world, and has given rise to a much more finely detailed knowledge of how water vapor is distributed around the globe. It was using this device that the 'hygropause,' a minimum in mixing ratio some 2–3 km above the local tropopause, was discovered in tropical regions. Since the first balloon borne Lyman-alpha hygrometers were flown, there have been versions built to also fly on high-altitude aircraft by several research and educational institutes around the world. The Lyman-alpha record forms a fairly long time record of water vapor measurements.

Tunable diode laser spectrometers are also used, both on aircraft and balloon, for *in situ* measurements of water vapor in both the troposphere and stratosphere. These use infrared absorption to determine water vapor concentration and there are both open and closed path versions of these instruments, flying on both balloons and aircraft. A technique newly developed for stratospheric water vapor is a chemical ionization mass spectrometer. For this instrument a portion of the sample flow passes through an ionization chamber where it is exposed to α -particle radiation. Air reacts with α -particles producing mostly O_2^+ and N_2^+ , which lead to a series of ion-molecule reactions resulting in the production of H_3O^+ ions from ambient H_2O . Ions are analyzed using a quadrupole mass spectrometer and the measured count rate of the H_3O^+ analyte ions is used to quantify ambient water vapor.

Operational radiosondes, typically using capacitive hygrometers, unfortunately are not able to measure the very low values of water vapor found in the stratosphere. So, there are limited long-term *in situ* records that can be used to look at long-term changes in stratospheric water vapor.

For a global view of the stratospheric water vapor distribution, space based techniques are essential. The first stratospheric water vapor measurements from satellite were made by the NASA Limb Infrared Monitor of the Stratosphere (LIMS) on the Nimbus 7 spacecraft launched in 1978. This device, which employed sensitive cooled detectors in space, detected emission from stratospheric water vapor from one of its infrared vibration-rotation bands (the ν_2 band centered at 6.3 μm). The intensity of this emission is proportional to the atmospheric temperature and the water vapor concentration. Knowing the temperature from separate measurements allowed the water vapor concentration to be determined globally as a function of altitude and position. This was a very exciting development, which employed the new technique of limb-sounding that increased the precision of stratospheric measurement by aiming the instrument sideways, toward the limb of the atmosphere, where the stratosphere is exposed against the cold and dark background of space. Nimbus 7 was notable for another reason: this program introduced the idea of assembling international 'Experiment' or 'Science' teams to assist in the development and scientific exploitation of the experiments onboard the satellite, a method adopted in almost all satellite experiments since.

There have been a large number of space-based water vapor instruments since LIMS. Also on Nimbus 7 was the British instrument called the stratospheric and mesospheric sounder,

which used another new technique, pressure modulation radiometry, to detect and measure water vapor (among other gases). This technique actually carried a sample of water vapor in a cell onboard, as a type of 'calibrator' of the detected infrared emissions. A Fourier transform spectrometer called ATMOS (atmospheric trace molecule spectroscopy) was flown several times on the space shuttle in the 1980s and 1990s, providing highly accurate spectral information about water vapor and many other stratospheric molecules. Long-term measurements were initiated using shorter-wavelength visible and near ultraviolet observations of water vapor absorption, using the Stratospheric Aerosol and Gas Experiment (SAGE). The Upper Atmosphere Research Satellite (UARS) launched in 1991 and operated until 2005, included the Halogen Occultation Experiment (HALOE). HALOE measured the absorption of infrared solar radiation by stratospheric water vapor, using the LIMS technique of looking through the limb of the atmosphere, in this case as the Sun rises or sets behind the atmosphere. The first Microwave Limb Sounder (MLS) also flew on UARS; it operated in the millimeter wave part of the spectrum. Currently flying on the NASA Aura satellite is another MLS; this one has operated since 2004. The Atmospheric Chemistry Experiment (ACE), a satellite mission onboard the Canadian satellite SCISAT-1 launched in 2003, includes a Fourier transform spectrometer performing measurements on several lines in the $1362\text{--}2137\text{ cm}^{-1}$ range; from these H_2O concentration profiles are retrieved up to 90 km in altitude. The Michelson Interferometer for Passive Atmospheric Sounding (MIPAS) onboard the Envisat research satellite provided stratospheric water vapor measurements from 2002 to 2012 when communications with Envisat were lost. MIPAS is a mid-infrared limb emission Fourier transform spectrometer designed for global vertical profile measurement of temperature and many atmospheric trace constituents relevant to atmospheric chemistry and climate change. ATMOS, ACE, and MIPAS are among instruments providing measurements of water vapor isotopologues, which have provided new insights on transport pathways from troposphere to stratosphere. This has not been a complete listing of remote sensors measuring stratospheric water vapor, but gives a sampling of the rich number of techniques available. There are also ground-based and balloon-based remote sensing instruments that provide water vapor measurements and have proved useful for monitoring at individual sites or for specific campaigns.

The *in situ* and remotely sensed water vapor measurements have helped us to understand much about what controls stratospheric humidity, and possible relations to tropospheric climate and stratospheric transport. Puzzles do remain, a few of which are discussed below.

Control of the Mean Distribution and Variability of Stratospheric Water Vapor

It is one of the fascinations of the study of stratospheric humidity that, while this extreme aridity and the overall mechanisms causing it have been known for more than half a century, the detailed understanding of precisely how this state is maintained remains elusive. This basic theory of the humidity of the stratosphere has survived and today forms

the basis of our understanding. However, we now realize that there are many important details that modify this model.

The key mechanisms that contribute to the control of stratospheric water vapor concentrations are stratospheric photochemistry and transport, tropical tropopause dehydration, and polar dehydration. Additionally, transport from troposphere to stratosphere at midlatitudes also appears to impact stratospheric water vapor.

Stratospheric Photochemistry

In the stratosphere, the high intensity of short-wave solar radiation means that methane and molecular hydrogen can be photolyzed, which releases active hydrogen compounds. When methane is oxidized it produces roughly two molecules of water vapor for every one molecule of methane destroyed. As a consequence, the quantity $2(\text{CH}_4) + (\text{H}_2\text{O})$ is quasi-conserved; this fact is useful for inferring motions and origins of air masses in the stratosphere. Lifetimes of both methane and molecular hydrogen are over 100 years at the tropopause, a few years at 30 km, and a few months at 40 km, owing to the increasing solar flux at short wavelength. Thus, air can be 'tagged': low values of (CH_4) and high values of (H_2O) in lower stratospheric air indicate that the air has been transported down from higher altitudes.

Tropical Tropopause Dehydration

The mechanisms responsible for stratospheric dehydration have been studied for more than 60 years; the concept was first introduced in a seminal paper by Alan Brewer, published in 1949. It is an amazing fact that, despite more than 60 years of research, we are still unsure about the precise mechanism(s) that cause the 'cold trap' phenomenon at the tropical tropopause. To first order, freeze-drying happens as air passes through the cold tropical tropopause when air cools, saturates, ice forms, and then falls out. Variations, such as an annual cycle in tropical lower stratospheric water, have been shown to be directly related to similar variations in tropical tropopause temperature; however, the zonally averaged tropical tropopause temperature is not low enough to produce the observed very low values in the stratosphere. This indicates that there are preferential longitudinal regions for the final dehydration of air entering the stratosphere, and trajectory model calculations using winds derived from data assimilation of wind and temperature fields support that conclusion. Current modeling and observational work is concentrating on teasing out the details of the microphysics involved in the final dehydration of air entering the stratosphere. Because models do not necessarily represent the physical processes involved with dehydration accurately in the tropical tropopause region, our predictions of changes in stratospheric humidity in the future are highly uncertain.

Polar Dehydration

Within the stratospheric vortices that form around each of the poles in winter, temperatures can fall to extremely low values (e.g., 180 K) and, of course, the air in the vortex is dehydrated under such conditions. In the north, the strength of the vortex, the amount of cooling within the vortex, and the consequent

degree of dehydration is not thought to be sufficient to affect the annual mean. However, in the south, the dessication is very significant and impacts the annual mean, even though it is only a seasonal effect. Below the 400 K potential temperature surface, in spring and summer, dehydration can affect midlatitudes, but the effect on the rest of the stratosphere is minimal.

Troposphere Stratosphere Exchange at Midlatitudes

In the midlatitude lower stratosphere, the extreme dryness of the stratosphere is maintained against the relatively very high humidity of the tropopause just a few kilometers away under the tropopause. At these latitudes there is not the strong vertical convection to maintain the cold trap mechanism and tropopause temperatures are warmer on average than at lower latitudes. What happens? First, the air in the stratosphere is, on the average, subsiding from higher up and so maintains an appropriate level of moisture, though there is also 'leakage' through the tropopause at midlatitudes. The most likely route is from the tropical high tropopause, along isentropes, passing through the break in the tropopause that often exists at midlatitudes (as a result of deformation of the tropopause into 'folds,' caused by various tropospheric dynamical features such as low-latitude troughs). The influence of high topography, such as the Himalayas and the Tibetan Plateau is also thought to influence isentropic flow from tropical upper troposphere into midlatitude stratosphere. There is also evidence that the summer monsoonal circulation, both for Asia and North America, result in water vapor transport into the lowermost stratosphere. Convection overshooting the tropopause may inject ice into the lower stratosphere, but that contribution has not been quantified at this point. It is clear that explaining the mean distribution of water vapor in the stratosphere involves understanding not only the global mean circulation, but also a range of detailed tropopause-level processes. A range of spatial and temporal scale processes are likely important.

Timescales

Each source of stratospheric water vapor is associated with a distinct timescale. The input of water vapor into the stratosphere by an individual air parcel in the tropics is largely a function of the lowest temperature a parcel encounters on its transit into the stratosphere, as originally noted by Brewer. The actual trajectory a parcel takes does need to be considered, but a simple model of horizontal processing of air by passage through the Western Pacific cold point tropopause reasonably reproduces observed stratospheric humidity. Variability in stratospheric water vapor on seasonal and interannual timescales has been well reproduced by climatological trajectory studies using saturation mixing ratios calculated from global temperature analyses, demonstrating that to first order, variability in the entry of water vapor into the stratosphere is controlled by variability in tropical cold point temperatures. Convection overshooting into the stratosphere has been observed on limited occasions, and when it occurs likely hydrates the stratosphere locally. However, evidence for a global impact of this phenomenon is lacking at this. Because

there is a relatively short turnover time for air in the lowermost stratosphere, on the order of months, changes in the entry value of water vapor to the stratosphere due to tropical cold point temperature changes will be seen almost immediately throughout the lowermost stratosphere. However, there will be a time lag before the signal reaches the middle and upper stratosphere on the order of years.

Accuracy of Stratospheric Water Vapor Measurements

Measuring low stratospheric water vapor concentrations is highly challenging and uncertain. There are significant discrepancies noted between coincident measurements of stratospheric water vapor concentrations using different *in situ* and satellite. These discrepancies range from 10 to 50% or even greater in some cases and preclude combining data sets for trend analysis without extreme care. This issue, of accuracy, is at the root of one of the current water vapor puzzles.

As moist air rises to colder regions in the atmosphere, in particular, in the region of the tropical tropopause, humidity rises above its equilibrium value over ice. As a consequence, the air releases its water vapor via ice cloud formation. These clouds form via ice nucleation in or on existing aerosol particles and then the ice particles grow through condensation of supersaturated water vapor onto the ice surfaces. Ice nucleation requires a supersaturation above a critical threshold value, which, for homogeneous freezing in the upper tropical troposphere, is $\sim 160\%$. Nucleation can also occur heterogeneously on particles known as ice nuclei. After nucleation, vapor molecules condense onto the ice particles, causing them to grow and the gas phase to become depleted in water until equilibrium is reached, and that equilibrium is assumed to be somewhere near 100% saturation. However, there are some observations of persistent clear air supersaturations with respect to ice greater than 200% and in cloudy air of 130%. This is a puzzle, because large supersaturations are expected to relax near 100% quite rapidly. The question remains as to whether these measurements, at quite low mixing ratios, are accurate. There has been a large international effort to attempt to address this accuracy issue; this effort has included instrument comparisons within a large cloud chamber (AquaVit), comparisons during aircraft missions, and consideration of calibration methods used at low mixing ratios. This effort is still ongoing.

As noted above, there are still concerns with absolute accuracy in measurements for stratospheric water vapor, however, data quality is sufficient to examine annual and interannual variations of water vapor in the tropical lower stratosphere as long as one considers long-term stable measurements. Many studies have shown that annual and interannual variations are to be in quantitative agreement with the idea that variations in tropical tropopause temperatures control the entry value of stratospheric water. The tropical annual variation, known as the water vapor tape recorder is an excellent example of the impact of tropical tropopause temperatures on water vapor. The level of coldest temperatures sets the water vapor mixing ratio; the correlation between temperature and water vapor is shown in Figure 4, with the dominant variation seen in the annual cycle. This annual variation in water vapor entering the stratosphere

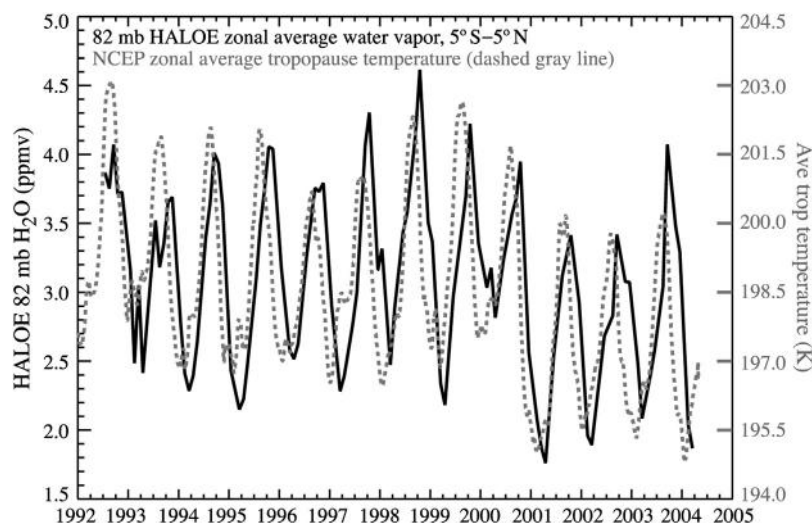


Figure 4 The 5°N–5°S water vapor mixing ratio from HALOE at the altitude of the average profile minimum in the tropics (black solid, scale on left) and NCEP/NCAR reanalysis zonal average tropopause temperatures (gray dashed, scale on right). The correlation maximizes with a 2-month shift, with water vapor lagging. Reproduced from Rosenlof, K. H., Reid, G. C., 2008. Trends in the temperature and water-vapor content of the tropical lower stratosphere: the sea-surface connection. *Journal of Geophysical Research* 113, D06107. doi:10.1029/2007JD009109.

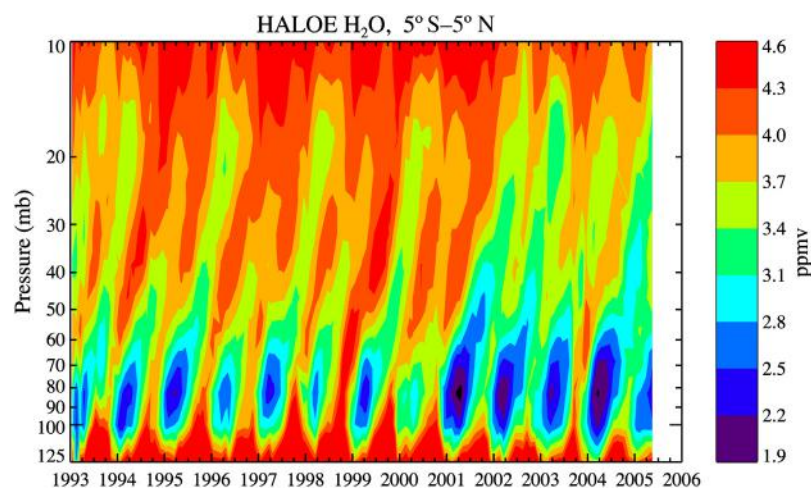


Figure 5 Tropical HALOE water vapor (tape recorder), 5°S–5°N, plotted versus time. Note the change to lower values of the hygropause at the end of 2000 and the upward propagation of those lower values in subsequent years. Reproduced from Rosenlof, K. H., Reid, G. C., 2008. Trends in the temperature and water-vapor content of the tropical lower stratosphere: the sea-surface connection. *Journal of Geophysical Research* 113, D06107. doi:10.1029/2007JD009109.

then propagates upward, as seen in Figure 5. The ability for a model to reproduce this feature is commonly used to assess quality of a simulation, as it tests both the model's stratospheric overturning circulation and the degree of mixing between low and middle latitudes. Comparisons of water vapor simulations with observations have been used as one means for assessing the quality of chemistry climate model simulations in recent studies.

Long-Term Trends

Global water vapor trend determination from the historical record is difficult. There are differences in trends noted between

measurement systems covering the same time period (e.g., the northern hemisphere frost point balloon as compared with the UARS HALOE satellite instrument). The multidecadal stratospheric water vapor record is limited to northern hemisphere midlatitudes. The longest continuous record of stratospheric water vapor data is from frost point balloon measurements taken at 40°N from Boulder, Colorado. At present, the longest satellite records are from SAGE II and HALOE instruments, however, the MLS instrument is now building up a significant time record.

There are indications that the stratospheric water vapor has increased over the past 50 years. This is to be expected given that methane emissions have increased, and oxidation of methane is the major production of water vapor in the

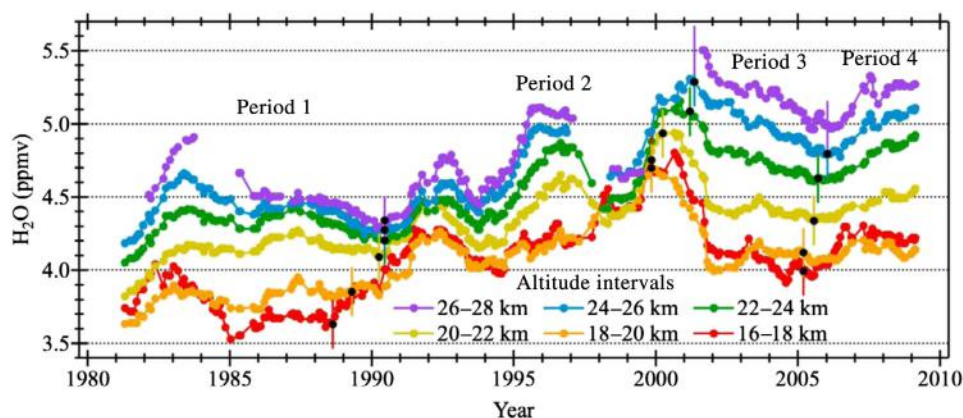


Figure 6 One-year moving averages of the 2-km water vapor mixing ratio averages in each of the six altitude levels denoted on the panel. From Hurst, D. F., Oltmans, S. J., Vömel, H., Rosenlof, K. H., Davis, S., Ray, E. A., Hall, E., Jordan, A., 2011. Stratospheric water vapor trends over Boulder, Colorado: analysis of the 30 year Boulder record. *Journal of Geophysical Research* 116, D02306. doi:10.1029/2010JD015065.

stratosphere. However, some studies indicate that the increase in methane emissions may significantly exceed, hence the need to consider mechanisms that change the transport of water vapor into the stratosphere. For example, if there is a change in the transport of air into the stratosphere, with the fraction of air entering the stratosphere bypassing the changing coldest tropical temperatures, there could be, conceivably, a change in the effective entry value of stratospheric water vapor without a change in the contribution due to methane oxidation. **Figure 6** shows interannual variability and the long-term trend for water vapor from 1980 to 2010 from frost point balloon data taken at one location in the northern midlatitudes. Note the increase over time at multiple altitudes, and the multiyear variability as well. Understanding this record and the possible feedbacks associated with changes in stratosphere water is the subject of current research, and encompasses studies of chemistry, dynamics, and radiation.

Summary

There is good understanding of the spatial distribution of water in the stratosphere as well as what drives the annual cycle of water vapor entering the stratosphere. The amplitude of the annual cycle is 50–60% of the mean and well explained by the known annual cycle in tropical tropopause temperatures. In contrast, the trend in stratospheric water vapor, which is a much smaller signal than the annual cycle, is not well understood. Over the period 1950–2000 there was an increase in the entry-level stratospheric water vapor on the order of 1% per year during a period of increasing tropospheric methane and decreasing tropopause temperatures. At the end of 2000, there was a decrease in the stratospheric entry-level water vapor coincident with a steplike drop in tropical tropopause temperatures. The observed long-term increase in stratospheric water vapor over the 1950–2000 period cannot be explained through tropical tropopause temperature trends, although many aspects of interannual variability can be. The more recent decrease in stratospheric water vapor can be explained by tropical tropopause temperature changes, although the mechanism driving that temperature change is

not well understood. Measurements are difficult to make with extremely high accuracy, and there are still questions as to whether high supersaturations that have been measured by *in situ* instruments are valid. This impacts our understanding of microphysical processes, and is a topic of current research. Because of the uncertainties in our understanding and modeling of past water vapor changes, it is difficult to predict changes expected in a future climate.

See also: Chemistry of the Atmosphere: Observations for Chemistry (*In Situ*): Water Vapor Sondes. Satellites and Satellite Remote Sensing: Water Vapor.

Further Reading

- SPARC Assessment of Upper Tropospheric and Stratospheric Water Vapour, 2000. World Climate Research Programme Report No. 113. World Meteorological Organisation, Geneva.
- Brewer, A., 1949. Evidence for a world circulation provided by the measurements of helium and water vapor distributions in the stratosphere. *Quarterly Journal of the Royal Meteorological Society* 75, 351.
- Mote, P.W., Rosenlof, K.H., Holton, J.R., Harwood, R.S., Waters, J.W., 1995. Seasonal variations of water vapor in the tropical lower stratosphere. *Geophysical Research Letters* 22, 1093–1096.
- Mote, P.W., Rosenlof, K.H., McIntyre, M.E., Carr, E.W., Gille, J.C., Holton, J.R., Kinnerson, J.S., Pumphrey, H.C., Russell III, J.M., Waters, J.W., 1996. An atmospheric tape recorder: the imprint of tropical tropopause temperatures on stratospheric water vapor. *Journal of Geophysical Research* 101, 3989–4006.
- Peter, T., Marcolli, C., Spichtinger, P., Corti, T., Baker, M.B., Koop, T., 2006. When dry air is too humid. *Science* 314, 1399–1402.
- Solomon, S., Rosenlof, K.H., Portmann, R., Daniel, J., Davis, S., Sanford, T., Plattner, G.-K., 2010. Contributions of stratospheric water vapor changes to decadal variations in the rate of global warming. *Science* 327, 1219–1223.
- Hurst, D.F., Oltmans, S.J., Vömel, H., Rosenlof, K.H., Davis, S., Ray, E.A., Hall, E., Jordan, A., 2011. Stratospheric water vapor trends over Boulder, Colorado: analysis of the 30 year Boulder record. *Journal of Geophysical Research* 116, D02306. <http://dx.doi.org/10.1029/2010JD015065>.
- Rosenlof, K.H., Reid, G.C., 2008. Trends in the temperature and water-vapor content of the tropical lower stratosphere: the sea-surface connection. *Journal of Geophysical Research* 113, D06107. <http://dx.doi.org/10.1029/2007JD009109>.
- Solomon, S., Rosenlof, K.H., Portmann, R., Daniel, J., Davis, S., Sanford, T., Plattner, G.-K., 2010. Contributions of stratospheric water vapor changes to decadal variations in the rate of global warming. *Science* 327, 1219–1223.

STRATOSPHERE/TROPOSPHERE EXCHANGE AND STRUCTURE

Contents

Global Aspects

Local Processes

Tropopause

Global Aspects

JR Holton, University of Washington, Seattle, WA, USA

© 2003 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 5, pp 2137–2143, © 2003, Elsevier Ltd.

Introduction

The troposphere and the stratosphere are separated by a boundary called the tropopause, whose altitude varies from about 16 km in the tropics to about 8 km near the poles. The troposphere is characterized by rapid vertical transport and mixing caused by weather disturbances; the stratosphere is characterized by very weak vertical transport and mixing. The tropopause thus represents a boundary between the troposphere, where chemical constituents tend to be well mixed; and the stratosphere, where chemical constituents tend to have strong vertical gradients. The two-way exchange of material that occurs across the tropopause is important for determining the climate and chemical composition of the upper troposphere and the lower stratosphere. This cross-tropopause transport is referred to as stratosphere–troposphere exchange. The upward transport of tropospheric constituents into the stratosphere occurs primarily in the tropics, and initiates much of the chemistry that is responsible for global ozone depletion. The downward transport of stratospheric constituents into the troposphere occurs mostly in the extratropics and not only serves as the major sink for some of the constituents involved in stratospheric ozone depletion, but also provides a source of upper tropospheric ozone.

This pattern of upward cross-tropopause transport in the tropics and downward cross-tropopause transport in the extratropics is part of a global mass circulation in the stratosphere that occurs as an indirect response to zonal (westward) forcing in the stratosphere, which is caused by the breaking of large-scale waves propagating from the troposphere. The magnitude and variability of this stratospheric mass circulation, and its consequences for atmospheric chemistry, are primary considerations in the study of stratosphere–troposphere exchange.

The Dynamics of Mean Mass Exchange

The Dynamical Definition of the Tropopause

The tropopause is traditionally defined by meteorologists as the lowest level at which the rate of decrease of temperature with respect to height (normally about 6 K km^{-1} in the troposphere) decreases to 2 K km^{-1} , and the average from this level to any level within the next 2 km does not exceed 2 K km^{-1} . This definition, however, does not always clearly mark the boundary between stratospheric and tropospheric air. The physical tropopause is better defined in terms of a specified critical value for a long-lived tracer such as ozone, which has distinctly different stratospheric and tropospheric values. Because global observations of ozone in the vicinity of the tropopause are very limited, it has become common to use as an alternative marker for the tropopause a dynamical field called the potential vorticity. Potential vorticity is somewhat analogous to spin angular momentum. For large-scale atmospheric motions potential vorticity is approximately given by $P = \rho^{-1} - (\zeta + f)(\partial\theta/\partial z)$, where ρ is the air density, ζ is the vertical component of relative vorticity, f is the Coriolis parameter (twice the local vertical component of the Earth's angular velocity), θ is the potential temperature (a measure of entropy, which increases rapidly with height in the stratosphere), and z is the height above sea level. Since $\partial z/\partial\theta$ may be regarded as a local measure of the depth of the layer between two potential temperature surfaces, an increase in $\partial z/\partial\theta$ implies stretching of vortex tubes and an increase in the absolute vorticity, while a decrease in $\partial z/\partial\theta$ implies shrinking of vortex tubes and a decrease in the absolute vorticity; this is somewhat like the spin angular momentum of a ballerina or figure skater.

Outside the tropics, potential vorticity is positively correlated with ozone in the extratropical lower stratosphere, and is

a particularly suitable tracer for defining the tropopause. Potential vorticity increases dramatically from troposphere to stratosphere, and it can be readily calculated from conventional wind and temperature data. As shown in Figure 1, the tropopause defined in terms of a critical value of potential vorticity does not coincide with an isentropic surface, but rather cuts across the isentropes as it slopes downward toward the poles in midlatitudes. The region of the stratosphere where the isentropes intersect the tropopause is called the 'lowermost stratosphere,' and must be clearly distinguished from the region above where the isentropes lie entirely in the stratosphere. This latter region is often referred to as the 'overworld.'

The Diabatic Circulation

Potential temperature is a function of specific entropy alone and is thus conserved by fluid parcels when the motion is adiabatic. Since diabatic processes operate on the timescale of weeks in the lower stratosphere, on shorter timescales parcels move approximately along constant potential temperature surfaces. Transfer of mass and chemical constituents from the troposphere into the stratospheric overworld, however, clearly requires motion across isentropic surfaces. This transport is accomplished by a mean meridional cross isentropic mass circulation. Such a circulation was first deduced by Brewer and Dobson, who showed that observations of the stratospheric distributions of water vapor and ozone were consistent with the notion that upward transport in the stratosphere is limited to the tropics, while downward transport occurs in the extratropics. Thus, the stratosphere is dehydrated by the freeze drying of air passing upward through the extremely cold tropical tropopause, and ozone accumulates at high latitudes in the lower stratosphere through poleward and downward transport from its source region in the upper tropical stratosphere.

This transport circulation is now commonly referred to as the Brewer–Dobson circulation. It is often called the 'diabatic circulation,' since it is associated with the diabatic processes of

radiative heating and upward motion across isentropes in the tropics, and with radiative cooling and downward motion across the isentropes in the extratropics. For long-lived trace constituents, such as methane and nitrous oxide, this pattern of meridional overturning, moving up in the tropics and down in the extratropics, tends to produce surfaces of constant mixing ratio that are elevated in the tropics and slope downward toward the poles, while mixing along the isentropes by planetary waves tends to flatten the slopes of surfaces of constant tracer mixing ratio.

Although it is associated with diabatic heating and cooling, the Brewer–Dobson circulation is not forced by radiative heating, nor is it forced directly from below by penetration of convection into the stratosphere. Rather, it is a nonlocal response to an extratropical wave-driven pumping action. This pumping is caused by the wave-induced westward force in the extratropical stratosphere. Because the Earth is rotating rapidly, pushing air westward produces a gyroscopic effect in which the air drifts poleward. By mass continuity a poleward drift in midlatitudes is compensated by upward motion accompanied by expansion and adiabatic cooling in the tropics and downward motion accompanied by compression and adiabatic warming in the extratropics (Figure 2). This distribution of adiabatic cooling and heating maintains the temperature below radiative equilibrium in the tropical upwelling region, and above radiative equilibrium in the extratropics. Thus, the distribution of radiative heating and cooling in the stratosphere does not drive the mean meridional mass flow, rather it is a response to the dynamically driven mass flow.

Rossby Waves

Wave driving in the extratropical stratosphere is caused primarily by Rossby wave breaking. Rossby waves owe their existence to the latitudinal gradient of potential vorticity along isentropic surfaces. Because of this gradient, a fluid parcel displaced poleward or equatorward (and materially conserving its

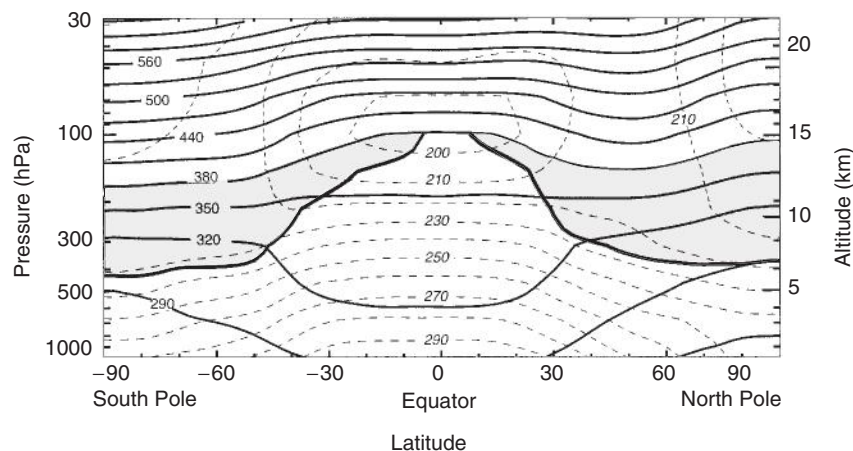


Figure 1 Latitude–altitude cross section for January 1993 showing longitudinally averaged potential temperature (solid contours) and temperature (dashed contours). The heavy solid contour (cutoff at the 380 K potential temperature surface) denotes a constant potential vorticity contour, which approximates the tropopause outside the tropics. Shaded areas denote the 'lowermost stratosphere' region whose potential temperature surfaces span the tropopause. Reproduced from Holton, J.R., Haynes, P.H., McIntyre, M.E., et al., 1995. Stratosphere–Troposphere exchange. *Reviews of Geophysics* 33, 403–439.

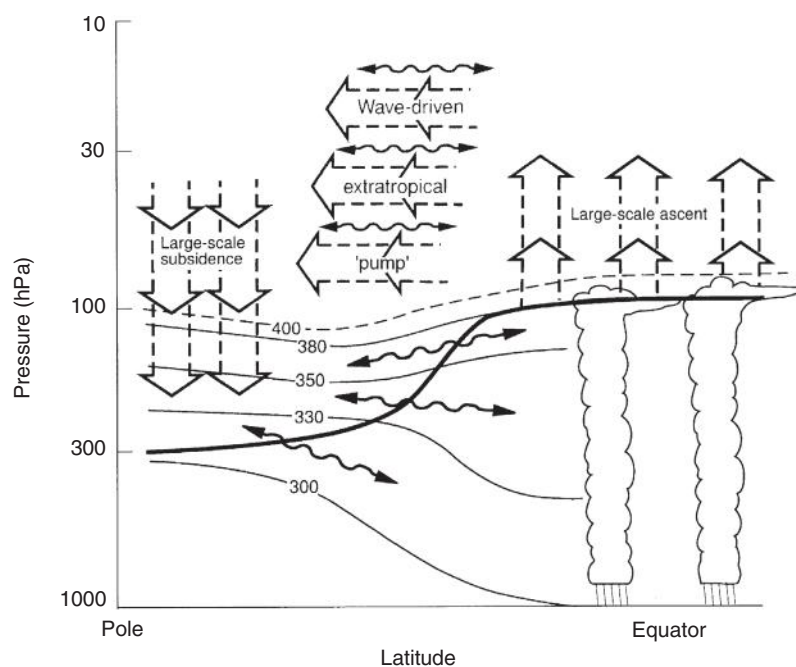


Figure 2 Dynamical aspects of stratosphere–troposphere exchange. The tropopause, defined by a specified constant potential vorticity in the extratropics and the 380 K potential temperature surface in the tropics, is shown by the thick line. Thin lines are potential temperature surfaces labeled in Kelvin. The shaded region is the ‘lowermost stratosphere’ where potential temperature surfaces intersect the tropopause, and isentropic exchange by tropopause folding occurs. Light shading in the stratosphere denotes the wave-induced westward zonal force. Wavy arrows indicate quasiisentropic transport and mixing by large-scale waves. The two-way exchange results in blocking of anticyclones, cyclone cutoff, and tropopause folds in the troposphere. The broad horizontal arrows are the meridional drift that balances the wave-induced zonal force, and the broad vertical arrows show the nonlocally driven equatorial upwelling and extratropical downwelling referred to here as the diabatic circulation. In the tropics some cumulonimbus clouds penetrate the stratosphere. Reproduced from Holton, J.R., Haynes, P.H., McIntyre, M.E., et al., 1995. Stratosphere–Troposphere exchange. *Reviews of Geophysics* 33, 403–439.

potential vorticity) will have potential vorticity different from that in the local environment and will induce a perturbation velocity disturbance. This will cause parcel displacements of the same sign to the west of the original displaced parcel, and of the opposite sign to the east. The result is a wave pattern in the potential vorticity field that propagates westward relative to the mean flow. When such a wave breaks in the stratosphere it produces a westward directed zonal force or wave drag (see *Dynamical Meteorology: Rossby Waves*).

Global Exchange: The Lowermost Stratosphere and the Overworld

Global mass exchange into and out of the chemically important region of the stratosphere is to a large degree controlled by the extratropical wave-driven pump discussed above. Air in the overworld, where isentropic surfaces lie entirely in the stratosphere, cannot reach the troposphere without first slowly descending across isentropic surfaces, a process that must be accompanied by diabatic cooling. The isentropic surface bounding the overworld and the lowermost stratosphere generally has a potential temperature around 380 K, depending on cloud top heights (see [Figure 2](#)).

The distinction between the overworld and the lowermost stratosphere implies that it is not always essential to measure

stratosphere–troposphere exchange by the transport across the tropopause.

For many purposes, transport into and out of the stratospheric overworld may be more relevant, and more effectively evaluated. For example, consider the case of a chemical species such as methane (CH_4) that has a tropospheric source and a stratospheric sink, with the sink being about 18 km or so, in the overworld. The transport across the 380 K potential temperature surface, which can largely be understood as part of the global-scale circulation of the overworld, is then an acceptable measure of exchange; indeed it is often more relevant because of the higher location of the photochemical sink. The same applies to a species that has a stratospheric source and a largely tropospheric sink. In this context, details of the transport across the tropopause are largely irrelevant.

For the understanding of mass and tracer transport into and out of the overworld, the replacement of the tropopause by a more convenient isobaric or isentropic control surface located in the lower stratosphere is useful. However, for some purposes, the mass transport across the actual tropopause is required. The net downward mass fluxes across the extratropical tropopause and across an isobaric or isentropic control surface that nearly coincides with the tropopause in the tropics should be equal for a sufficiently long time average. Such equality will not hold on seasonal or shorter time scales since the amount of mass in the layer between the extratropical

tropopause and an isobaric or isentropic surface coinciding with the tropical tropopause may vary with time.

To focus on the important aspects of global-scale exchange, it is useful to distinguish between the transport along isentropic surfaces, which can occur in rapid adiabatic motions (wavy arrows in Figure 2), and transport across isentropic surfaces, which requires diabatic processes. Since the tropopause intersects the isentropes, transport can occur in either way, and is likely to occur in both ways. In the region of the atmosphere called the lowermost stratosphere, where isentropic surface intersect the tropopause, air and chemical constituents can be irreversibly transported across the tropopause as adiabatic eddy motions lead to large latitudinal displacements of the tropopause, followed by irreversible mixing on small scales. The dark shading in Figure 2 shows the region within the lower stratosphere most directly affected by these eddy transport effects. The lowermost stratosphere must be distinguished from the rest of the stratosphere, being the only part of the stratosphere accessible from the troposphere via transport along isentropic surfaces.

Transport in the overworld must be clearly distinguished from transport in the lowermost stratosphere. Transport in the lowermost stratosphere requires consideration of the details of synoptic-scale and small-scale processes. Horizontal mixing can be especially significant in the lowermost stratosphere, especially during 'blocking' events, when meridional motions are enhanced. Thus exchange between the troposphere and the lowermost stratosphere can be significantly faster than exchange between the overworld and the lowermost stratosphere.

The Annual Cycle in Global Stratosphere–Troposphere Exchange

Diagnostics of the wave-driven zonal force in the extratropical stratosphere from analysis of conventional global meteorological data can be used to produce estimates of the vertical mass flux across a convenient control surface, such as the 100 hPa isobaric surface (~ 380 K potential temperature surface in the tropics), which can be regarded as approximating the lower boundary of the overworld. This technique works best for the solstice seasons, when the time change of zonal momentum is small compared to the wave-induced force. Results for the downward mass fluxes across the 100 hPa surface for the Northern and Southern Hemispheres (and by continuity the upward flux in the tropics) are shown in Table 1. The deduced

upward mass flux in the tropics is sufficient to completely replace the mass above the 100 hPa surface in about 2–2.5 years. The observations suggest that the mass transport across the tropical tropopause is twice as large in Northern Hemisphere winter as in Southern Hemisphere winter.

This deduced annual cycle in mass transport across the 100 hPa surface is consistent with the observed annual temperature cycle near the tropical tropopause, which is characterized by temperatures that are several degrees colder in January than in July throughout the tropics, and several degrees warmer in January than in July in the extratropics, as would be expected from the influence of the annual cycle in adiabatic cooling associated with the vertical motions. The annual cycle in mass transport is also consistent with the observed cycle in tropical total ozone, which is a minimum in February and a maximum in August, as would be expected from the enhanced vertical advection of ozone-poor tropospheric air into the tropical lower stratosphere during the Northern Hemisphere winter.

Isentropic Exchange in the Extratropics

As mentioned above, in midlatitudes the tropopause cuts across isentropic surfaces so that two-way stratosphere–troposphere exchange can occur through isentropic transport and mixing processes. Nevertheless, the boundary between stratospheric and tropospheric air remains very distinct in this region. This suggests that there must be a rather strong dynamical resistance to exchange along the isentropes. This resistance is supplied by the mechanism of Rossby wave propagation. Because of the strong isentropic gradient of potential vorticity that marks the tropopause in the lowermost stratosphere, there is a very strong Rossby wave restoring force in that region, which limits the extent of parcel displacements across the potential vorticity gradient. Hence, wave breaking only occurs for large-amplitude disturbances. Because there is a large store of available potential energy associated with the strong meridional temperature gradient in this region, large-amplitude weather disturbances quite frequently develop in this region, especially in wintertime. The vertical circulations associated with such disturbances create deep intrusions of stratospheric air into the troposphere, which may then become mixed with tropospheric air to produce irreversible transport into the troposphere. Much of the ozone transport from the lowermost stratosphere into the troposphere occurs in connection with such 'tropopause fold' events.

The quasiisentropic exchange initiated by tropopause folding could in theory occur in the absence of the diabatic circulation. But in that case there would necessarily be an equal quasiisentropic reverse transport of air from the troposphere into the stratosphere in order to maintain mass balance. The extremely low water vapor mixing ratios observed throughout the stratosphere indicates, however, that the quasiisentropic exchange in midlatitudes is mostly a one-way transport into the troposphere. Furthermore, the large-scale diabatic circulation is required to transport stratospheric constituents such as ozone downward from the overworld to the lowermost stratosphere. The average rate at which such a species can be transported into the troposphere is thus ultimately determined by the rate at

Table 1 Solstice season mass flux across the 100 hPa surface

Location	Mass flux ^a (10^6 kg s ⁻¹)		
	DJF	JJA	Annual mean
NH extratropics	–81	–26	–53
Tropics	114	56	85
SH extratropics	–33	–30	–32

^aNegative sign indicates downward flux. DJF, Dec, Jan, Feb; JJA, Jun, Jul, Aug; NH, Northern Hemisphere; and SH, Southern Hemisphere.

Reproduced from Rosenlof, K.H., Holton, J.R., 1993. Journal of Geophysical Research 98, 10465–10479.

which the dynamically controlled large-scale circulation transports mass into the lowermost stratosphere. For this reason, the details of mesoscale tropopause fold events may not be important for determining the global flux of ozone from the stratosphere, although they will certainly strongly influence the time and space distribution of such transport.

Tracer Exchange in the Lowermost Stratosphere

Exchange of trace constituents cannot be treated in the simple manner used above for the net mass flux because net tracer exchange can occur in the absence of mass exchange. For example, if one unit of air containing a high ozone mixing ratio flows into the troposphere and an equal unit with low ozone mixing ratio flows into the stratosphere, there will be a net downward ozone flux, but zero net mass flux. This sort of process could lead to tracer exchange at the lower edge of the lowermost stratosphere, where the tropopause cuts across isentropic surfaces.

Such exchange does not, however, occur on a continuous basis. As noted above, the boundary between stratospheric and tropospheric air along isentropes that span the tropopause is normally marked by strong isentropic potential vorticity gradients. The existence of this band of strong potential vorticity gradients, and indeed similarly strong gradients in mixing ratios of species such as ozone and water vapor, itself suggests that there must be rather strong dynamical resistance to cross-tropopause transport along the isentropes, since otherwise vigorous mixing of stratospheric and tropospheric air would destroy the band of strong gradients.

Nevertheless, stratosphere–troposphere exchange of tracers can occur by isentropic transport in the extratropical region. Development of strong upper-level weather disturbances can lead to displacement of the tropopause from its equilibrium position, followed by nonconservative processes such as diabatic heating or cooling or small-scale turbulent mixing. It is only in the presence of such vigorous eddy motions near the tropopause that the dynamical resistance to cross-tropopause exchange can be overcome, and deep intrusions of stratospheric air can penetrate into the troposphere. These intrusions may also to some extent be regarded as the result of the systematic effect of the large-scale ageostrophic circulations associated with the development of frontal structures near the tropopause. The stretching deformation that occurs during frontal development stretches stratospheric intrusions to ever finer scales and leads to irreversible transport, often speeded up by turbulence resulting from shear instabilities. Much of the ozone transport from the lowermost stratosphere into the troposphere is believed to occur in connection with such tropopause fold events. Many studies have confirmed that large

episodic stratosphere–troposphere exchange can occur in association with tropopause folding.

The Role of Tropical Convection in Stratosphere–Troposphere Exchange

Although convection does not control the rate at which the diabatic circulation moves mass into or out of the overworld, penetrative convection may influence a layer a few kilometers in depth in the region of the tropical tropopause. This tropopause layer plays an essential role in establishing the stratospheric water vapor budget. Aircraft and satellite observations of the water vapor distribution in the tropical lower stratosphere reveal the existence of a water vapor mixing ratio minimum (referred to as the ‘hygropause’). The hygropause often occurs well above the tropopause. This feature is now believed to result from dehydration to the saturation mixing ratio at the tropopause (the freeze drying process), followed by vertical advection of the resulting minimum in water vapor mixing ratio into the overworld by the diabatic circulation. Since, as pointed out above, the tropical tropopause temperature is colder in Northern winter than in Northern summer, the driest air enters the stratosphere in Northern winter (when the hygropause is observed to be just above the tropopause), and is advected upward to produce the observed elevated hygropause near 19 km in Northern summer.

See also: **Dynamical Meteorology:** Overview; Primitive Equations; Rossby Waves; Symmetric Stability; Waves. **Middle Atmosphere:** Transport Circulation. **Stratosphere/Troposphere Exchange and Structure:** Local Processes. **Tropical Cyclones and Hurricanes:** Hurricanes: Observation. **Tropical Meteorology and Climate:** El Niño and the Southern Oscillation: Observation; El Niño and the Southern Oscillation: Theory.

Further Reading

- Andrews, D.G., Holton, J.R., Leovy, C.B., 1987. *Middle Atmospheric Dynamics*. Academic Press, New York, NY.
- Brewer, A.M., 1949. Evidence for a world circulation provided by the measurements of helium and water vapor distribution in the stratosphere. *Quarterly Journal of the Royal Meteorological Society* 75, 351–363.
- Dobson, G.M.B., 1956. Origin and distribution of polyatomic molecules in the atmosphere. *Proceedings of the Royal Society of London A* 236, 187–193.
- Holton, J.R., Haynes, P.H., McIntyre, M.E., et al., 1995. Stratosphere–troposphere exchange. *Reviews of Geophysics* 33, 403–439.
- Salby, M.L., 1996. *Fundamentals of Atmospheric Physics*. Academic Press, New York, NY.

Local Processes

JF Lamarque and P Hess, National Center for Atmospheric Research, Boulder, CO, USA

© 2003 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 4, pp 2143–2150, © 2003, Elsevier Ltd.

The tropopause separates the stratosphere from the troposphere. It is located at the interface between two air masses with distinctly different characteristics in water vapor, ozone, potential vorticity, and other chemical or physical quantities. Stratosphere–troposphere exchange (STE) refers to the processes whereby mass and chemical species are transported between these two atmospheric regions across the tropopause. This exchange is important to the chemistry of both regions as it regulates the transport of species with tropospheric sources into the stratosphere (e.g., chlorofluorocarbons (CFCs), water vapor, and hydrocarbons) and

species with stratospheric sources into the troposphere (e.g., ozone and nitric acid). In this article we identify and describe the small-scale processes occurring in the vicinity of the tropopause that govern this exchange. We distinguish between these processes and the large-scale regulation of the exchange by the zonal mean–meridional Brewer–Dobson circulation (Figure 1).

The Brewer–Dobson circulation, a circulation largely forced by wave breaking remote from the tropopause, acts to drive air parcels up through isentropic surfaces in the tropics (corresponding to a mean heating of the air parcels) and down

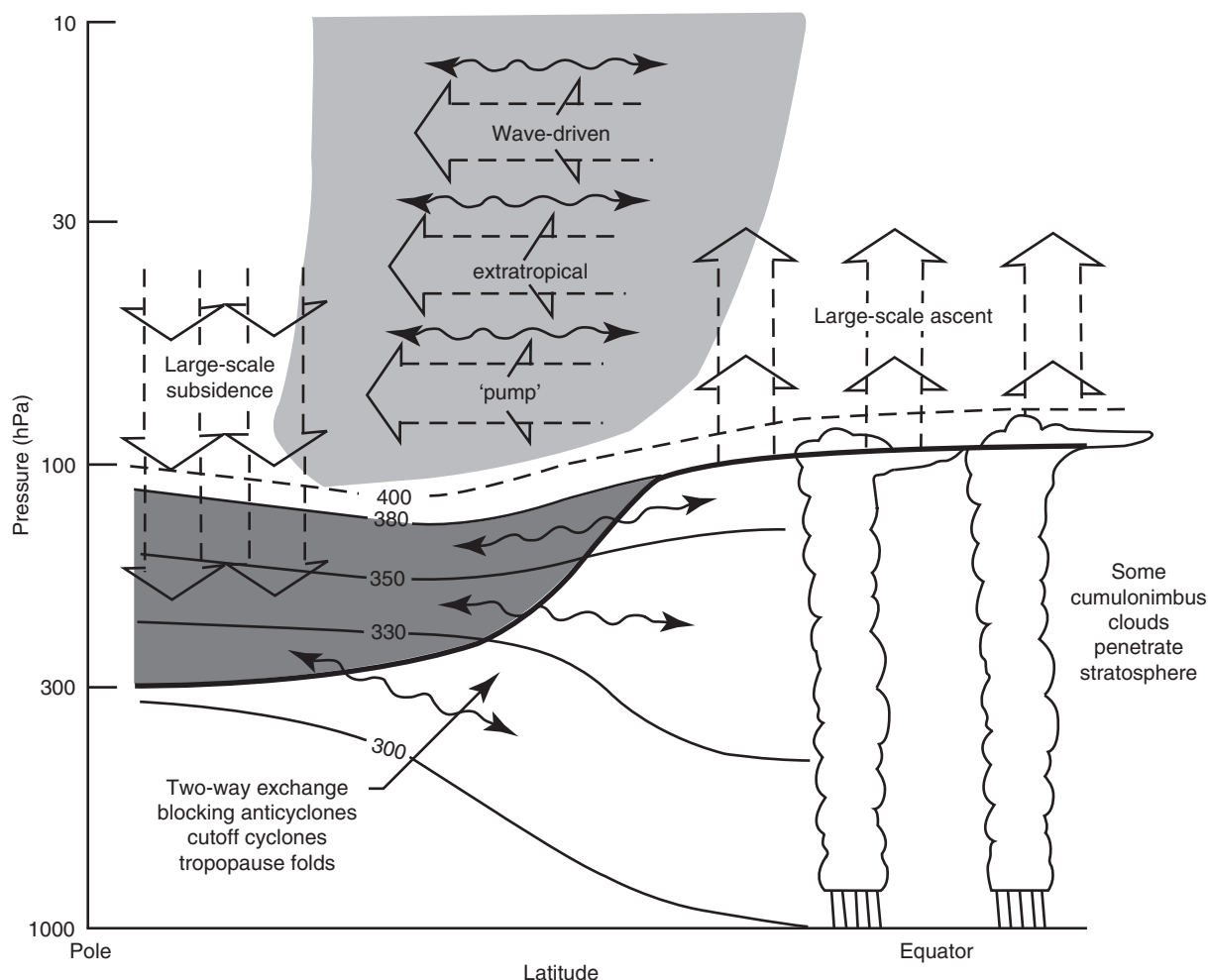


Figure 1 Dynamical aspects of stratosphere–troposphere exchange. The tropopause is shown by the thick line. Thin lines are isentropic surfaces labeled in Kelvin. The heavily shaded region is the 'lowermost stratosphere', where isentropic surfaces span the tropopause and isentropic exchange by tropopause folding occurs. The region above the 380 K surface is the 'overworld', in which isentropes lie entirely in the stratosphere. Light shading in the overworld denotes wave-induced forcing (the extratropical 'pump'). The broad arrows show transport by the global-scale circulation, which is driven by the extratropical pump. This global-scale circulation is the primary contribution to exchange across isentropic surfaces (e.g., the 400 K surface) that are entirely in the overworld. Reproduced with permission from Holton *et al.* (1995).

through the isentropic surfaces in the extratropics (corresponding to a mean cooling of the parcels). The dynamics of this circulation determines the net STE on an annual time scale. The upward branch of this circulation forces a net exchange from the troposphere to the stratosphere in the tropics and from the stratosphere to the troposphere in the extratropics. Small-scale processes influence precisely where and when STE of mass and chemical species occur. Modeling studies suggest the timing of the exchange, in particular, is important to tropospheric chemistry. Small-scale processes also influence the composition of the stratosphere in the vicinity of the tropopause. In particular, the composition of the lowermost extratropical stratosphere (the part of the stratosphere that shares isentropic surfaces with the troposphere) is strongly affected by small-scale processes.

The operational definition of the tropopause by the World Meteorological Organization is given in terms of the temperature lapse rate. However, in the extratropics a dynamically based definition of the tropopause is in terms of a potential vorticity (usually taken at a potential vorticity equal to $2 \times 10^{-6} \text{ m}^2 \text{ s}^{-1} \text{ K kg}^{-1}$). This definition cannot be extended to the tropics, where it is convenient to simply define the tropopause as the 380 K potential temperature surface. Regardless of the definition, the interface between stratospheric and tropospheric air masses forms a wavy surface with substantial geographic variations in height, latitude, and longitude. Significant displacements of the tropopause can occur without STE. The tropopause is a dynamic surface so that transport across it cannot be considered in the same manner as transport across a surface unaffected by transport (e.g., a constant altitude surface). For example, while the mean height of the Northern Hemisphere tropopause lowers during winter, the STE peaks in the spring months.

The advantage of using potential vorticity or potential temperature to mark the tropopause is that these quantities act as tracers of air mass motion, making them ideal to mark the interface between stratospheric and tropospheric air masses. Potential vorticity and potential temperature are conserved along trajectories except for the processes of diabatic heating (the vertical gradient of diabatic heating in the case of potential vorticity) and mixing. The extent to which these quantities are not conserved can be taken as a measure of the STE. Therefore, STE can be defined as the amount of mass or constituents transported across potential temperature surfaces in the tropics and potential vorticity surfaces in the extratropics. Diabatic heating and its vertical gradient are generally small in the upper troposphere and lower stratosphere. Mixing is also expected to be slow due to the high static stability of the stratosphere (which resists vertical displacements) and the high potential vorticity gradients (which resist horizontal displacements) of the lower extratropical stratosphere. However, as discussed later, under specific circumstances, these nonconservative processes are large enough to allow for significant STE.

Because of the different processes involved, the description of STE by small-scale processes is split between the tropics and extratropics. In each section we show how small-scale mixing and diabatic heating at the tropopause result in exchange between the troposphere and the stratosphere.

Although intensive research has taken place in the last 40 years, there are still a large number of uncertainties and unknowns in the small-scale processes involved in STE. In

particular, the precise mechanisms for the exchange across the tropical tropopause are still not completely understood.

Tropical Regions

The tropical tropopause (located at approximately 380 K) is located in the upwards branch of the Brewer–Dobson circulation (Figure 1) at a pressure of approximately 100 hPa and a temperature of approximately -70 to -80°C . Constituents lofted across the isentropic surface 400 K (approximately 90 hPa) subsequent to crossing the tropical tropopause are likely to be transported into the middle and upper stratosphere by the large-scale Brewer–Dobson circulation. There they can affect the composition of the stratosphere for years. Between the tropical tropopause and 400 K theoretical calculations and measurements of both water vapor and atomic bomb debris (from the 1950s and 1960s explosions) indicate considerable poleward transport of trace constituents. This suggests that a fraction of the constituents which cross the tropical tropopause are not transported much above 400 K, but are rapidly transported into the lowermost extratropical stratosphere, through mostly isentropic transport.

STE in the tropics is governed by a complex and poorly understood interplay between convection and the large-scale Brewer–Dobson circulation. Parcels that cross the tropopause are initially transported upwards in deep convective clouds. However, above some height, the Brewer–Dobson circulation will govern the subsequent uplift of the parcel. The transition height between convection and the large-scale circulation is not firmly fixed. At least the tropical tropopause is often not clearly demarcated. Instead it may be more accurate to regard the tropical tropopause as a rather deep transition region between the troposphere and the stratosphere.

It is still an open question whether the transition between convection and the large-scale circulation typically occurs above or below the defined tropical tropopause. Convective turrets do penetrate the tropopause on occasion, as observed in the Indonesian region, for example. However, there is some doubt as to whether these very deep convective events occur frequently enough to supply the requisite upward mass flux. In this case the upward motion across the tropical tropopause could be of large scale, in which case frequent high cloudiness near the tropopause would be expected. Subvisible cirrus clouds are observed over the warm pool of the western Pacific over 90% of the time during Northern Hemisphere winter, but the cause of this cloudiness is yet undetermined. On the other hand, if convection supplies more than the requisite mass flux above the tropopause, only the highest and coldest convective events may end up impacting the stratosphere. In this case, outside the convective updrafts the equatorial tropopause is in a subsident region.

The dryness of the air entering the equatorial stratosphere (approximately 3 ppm by volume during the Northern Hemisphere winter and 4.2 ppm by volume during Northern Hemisphere summer) tightly constrains the possible pathways through which tropical air can enter the stratosphere. As this is much drier than tropospheric air on average and typically drier than the saturation water vapor mixing ratio at the tropical tropopause, any theory of tropical STE must account for the dehydration of air parcels entering the stratosphere.

A possible mechanism for such low water vapor mixing ratio is that air that enters the stratosphere has been processed through a cloud. Indeed, as a parcel travels upward and cools, water in excess of the saturation vapor pressure condenses out. Efficient dehydration requires that the parcel remain at cold enough temperatures for ice crystals to grow to sufficient size for rapid sedimentation. Otherwise, as the parcel continues to rise into the stratosphere, the ice crystals may reevaporate. Air with low stratospheric mixing ratios of water vapor has sometimes been measured in association with deep convective clouds. However, processes other than convection may also play a role in dehydrating air. For example, gravity waves propagating near the tropopause may provide sufficient uplift to allow for additional condensation and loss of water vapor. Cloud processing will also affect the STE of chemical species through the attendant loss of soluble species.

Zonally averaged tropical tropopause temperatures are not consistent with the extreme dryness of the stratosphere. This suggests the hypothesis that there are preferred regions in which air enters the stratosphere; air passes locally upwards through the tropical tropopause only where the saturation vapor pressure is low enough (from the very cold temperatures) to allow for the sufficient dehydration of air parcels as described above. One such region occurs in the western Pacific (mostly in the vicinity of Indonesia) during Northern Hemisphere winter, in accord with the idea of a localized stratospheric 'fountain' through which air enters the stratosphere. However, during the Northern Hemisphere summer the temperature distribution from the large-scale meteorological analyses indicates no region with temperatures persistently cold enough to explain the water vapor record. At this time of year the cold temperatures and dehydration events must occur only sporadically in association with spatially and temporally restricted events not captured in the large-scale meteorological analyses. Another hypothesis, introduced recently and still being developed, is based on the existence of a deep tropopause transition layer. The dehydration of air occurs in convective systems but the transport of the dehydrated air into the stratosphere occurs in a slow ascent due to the overall net radiative heating in this part of the atmosphere. In this view, the dehydration and transport into the stratosphere occur at different times and locations. This view of tropical STE is more dynamic than the stratospheric 'fountain' and involves vertical and horizontal processes at very different scales. None of the hypotheses described above have yet been able to fully and consistently explain the observed distribution of water vapor in the tropical stratosphere.

Longitudinal variations in tropopause height and temperature, and therefore the preferred locations of equatorial STE, can be ascribed to an array of poorly understood local processes. The coldest tropopause heights are associated with the western Pacific warm pool and the Northern Hemisphere monsoon. This is consistent with convection playing an active role in shaping the morphology of the tropopause. However, the relationship between convection and the tropopause height is not straightforward. In particular, there is indication that minimum temperatures at the tropopause in January are centered on the Equator, while convection maximized slightly south. The radiative effects of convective clouds and the wave

motions forced by their diabatic heating obscure any straightforward relationship between convection, the height and temperature of the tropopause, and the location of STE.

Extratropics

Moving poleward from the Equator, the tropopause is conveniently defined in terms of a potential vorticity surface. STE occurs between the lowermost stratosphere and the troposphere through transport across this surface. While the transport can occur in either direction, it is predominantly from the stratosphere to the troposphere. The effect of transport in the opposite sense is short-lived due to the downwards large-scale mean meridional circulation, which acts to flush out the lowermost stratosphere within a relatively short period. In distinction to the tropical tropopause, the extratropical tropopause is usually clearly demarcated by strong gradients in potential vorticity and trace constituents.

The Subtropics

Stratosphere–troposphere exchange in this region occurs between the upper and mid-equatorial troposphere and the lowermost stratosphere. The subtropical tropopause drops rapidly near 30° from tropical heights to the level of the extratropical tropopause (approximately from 100 hPa to 300 hPa) (Figure 1). Trajectories from analyzed winds suggest very little STE occurs across this portion of the tropopause during the winter months but that considerable STE occurs during the summer months.

The subtropical tropopause cuts through the subtropical jet stream. This jet undergoes a substantial annual cycle in amplitude with the strongest winds occurring during the winter season. When the jet is strong, mixing across it between the troposphere and the stratosphere is inhibited; inhibited both through the large potential vorticity gradients associated with the jet, and the fact that breaking of tropospheric waves and the resulting mixing is unlikely to penetrate the jet core. Indeed, in the case of a strong jet the wind speeds are substantially larger than those associated with most tropospheric waves, implying that the critical layers (where the phase speed of the wave is equal to that of the large-scale flow field and therefore where the wave is unstable and breaks) will occur away from the jet core. During the summer months the subtropical jet weakens considerably, allowing mixing across the jet to be enhanced. Not only do the critical layers occur closer to the jet core during the summer months, but the smaller gradients of potential vorticity associated with the summer jet make for wider critical layers and weaker barriers to mixing.

The transport across the summertime subtropical jet is primarily associated with the Asian monsoon (Figure 2), and to a lesser extent the Mexican monsoon. While the monsoons of South America, Africa, and Australia probably play a similar role during the austral summer, their comparatively weak circulations are much less effective in transporting air across the tropopause. As indicated by the arrows in Figure 2, monsoon circulations are able to tap a particularly rich source of water vapor in the midlatitudes. The resulting STE is believed to be of primary importance to the seasonal cycle of water vapor in the

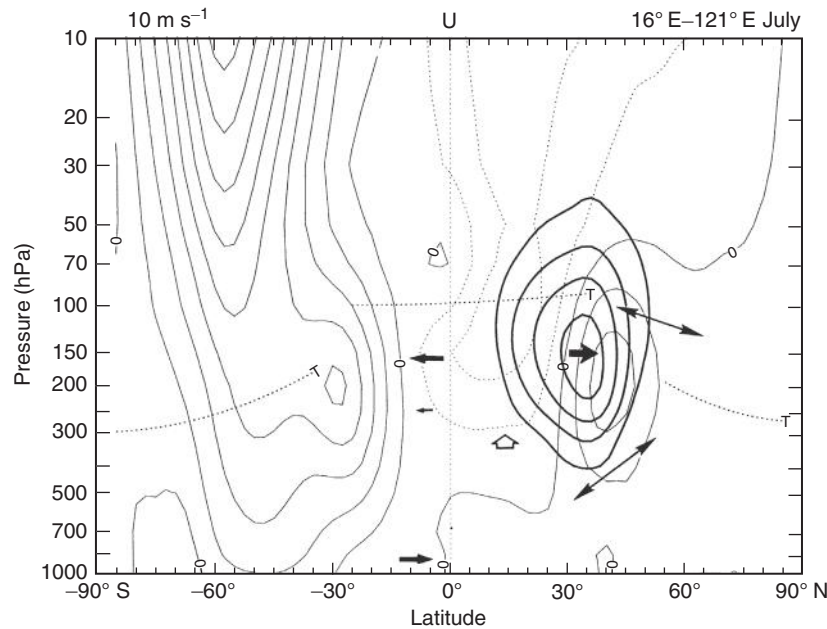


Figure 2 Schematic diagram of meridional transport and mixing adjacent to monsoon regions in northern summer, superposed on contours of zonal wind (interval 10 m s^{-1}). Heavy contours (interval 2 m s^{-1}) and one-way bold arrows indicate climatological meridional transport; two-way arrows illustrate mixing along isentropic surfaces. The large bold arrow at 30° N represents the western side of the Asian monsoon. Its direction should be reversed for the eastern side, where v (meridional velocity) is opposite and slightly smaller. The tropopause is shown (heavy dotted line, T) and the zero-wind line is labeled 0. Reproduced with permission from Dunkerton (1995).

extratropical lowermost stratosphere and does not involve the pronounced dehydration that occurs in the tropics.

The tropopause is elevated over monsoon regions with the associated anticyclonic circulation penetrating into the lowermost stratosphere. A steady state monsoon circulation will not in itself result in STE. However, due to the proximity of the monsoon circulation to the jet core, perturbations in the circulation are likely to be important, resulting in isentropic mixing between the troposphere and the stratosphere (Figure 2) and associated STE. Moreover, the interaction between monsoon and midlatitude synoptic disturbances or large-scale low-frequency transients will act to transport species across the tropopause. It has been demonstrated in the case of the Asian monsoon that the interaction can act to pull filaments (see the following section on extratropical STE) of moist tropospheric air into the stratosphere, and filaments of dry stratospheric air into the troposphere.

The Extratropics

In the extratropics a number of local processes result in STE. These include: stratospheric intrusions in the troposphere and their subsequent fragmentation; tropopause folds; cutoff lows; gravity waves; deep convection; radiative processes in the vicinity of the tropopause; and local dynamical instabilities. All the processes listed above are examined in more detail below.

The process of fragmentation (i.e., breaking into smaller and smaller structures such as filaments) of stratospheric intrusions is strongly related to isentropic mixing. Parcel advection calculations suggest that this mixing occurs vigorously throughout the year on isentropic surfaces below 330 K and is therefore responsible for most of the STE. The

fragmentation of stratospheric intrusions occurs as the large-scale velocity field causes tongues of stratospheric air to undergo large latitudinal excursions (Figure 3). Subsequently, these tongues can stretch and thin until they become mere filaments of stratospheric air embedded in the troposphere. This process can be viewed as the fragmentation of the tropopause itself. Once the filaments reach small enough scales they are rapidly and irreversibly mixed into the troposphere. This final mixing may occur due to dynamical instabilities growing at the interface of the filaments (e.g., Kelvin-Helmholtz shearing instabilities) (Figure 4) or through radiative decay. The associated potential vorticity anomalies become increasingly susceptible to radiative decay as they are stretched to small scales. Satellite measurements of ozone and water vapor suggest that fragmentation occurs continually in the vicinity of the tropopause.

The fragmentation of intrusions across the tropopause is similar to the fragmentation of the polar stratospheric vortex, creating the so-called stratospheric 'surf' zone. In both processes the associated irreversible mixing can be traced to the large meridional parcel displacements that occur in the vicinity of a Rossby wave's critical layer. As only large-scale waves can propagate into the stratosphere (due to the vertical structure of the zonal wind), the waves which break near the tropopause are of much smaller scale (generally wave number 4–7) than those that break higher up. Consequently the mixing regions are of smaller scale. The waves that break at the tropopause can often be linked to baroclinic instability. Depending on the horizontal shear of the flow, the mixing can occur on the equatorward side of the jet stream, in which case stratospheric air extrudes anticyclonically into the troposphere. In cases of enhanced horizontal shear, mixing can occur on the poleward

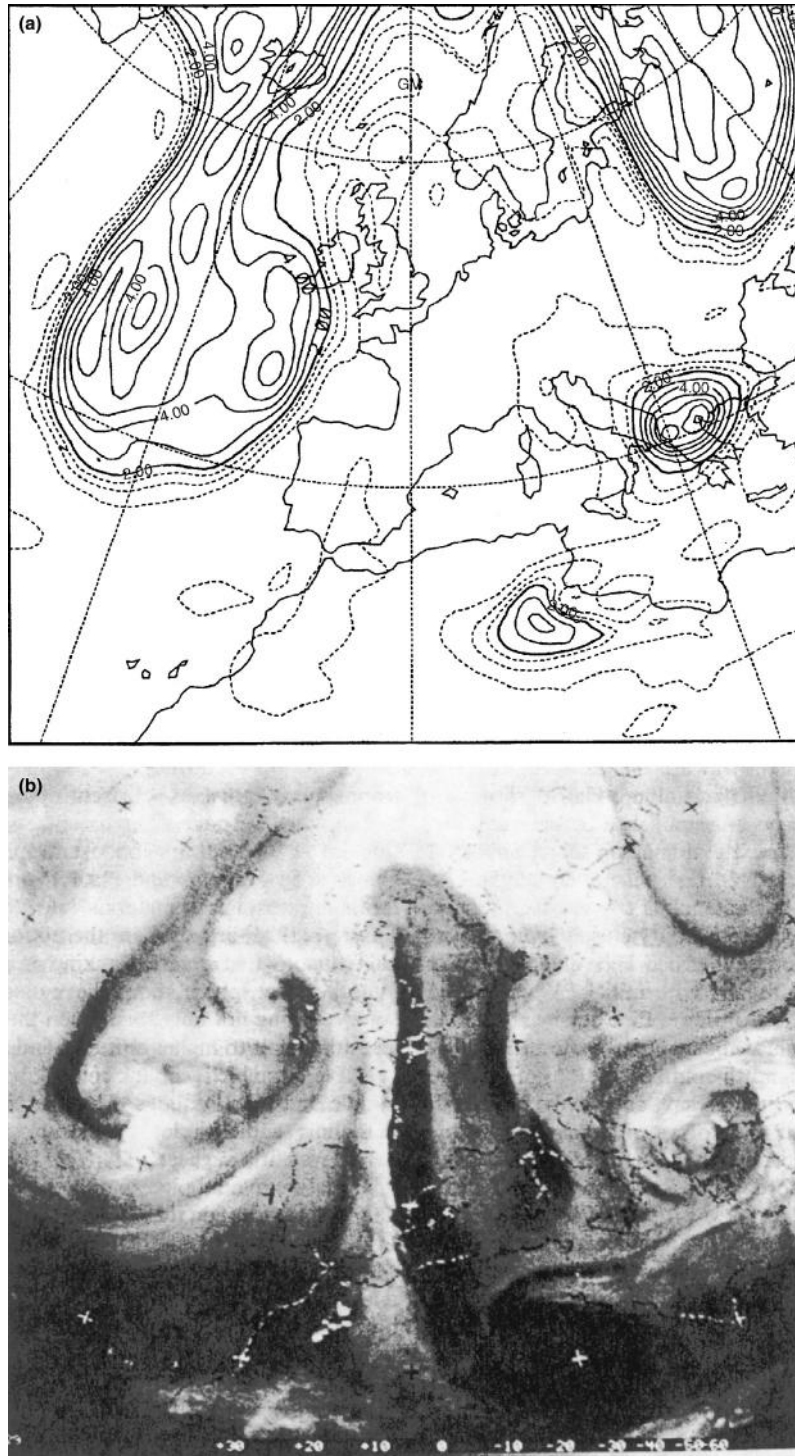


Figure 3 (a) Isentropic contours of potential vorticity on the 320 K surface for 14 May 1992, at 1200 UT, calculated from European Centre for Medium-Range Weather Forecasts (ECMWF) operational analyses. The instantaneous tropopause appears as the first solid contour (2 PVU); contours for 1, 1.5, 3, and 4 PVU are also shown. (b) Meteosat water vapor image for the same time as (a). The black structures in the upper-left and right are indicative of dry stratospheric air.

side of the jet, in which case tropospheric air is entrained into the stratosphere.

Owing to the large potential vorticity jump across the tropopause, strong ageostrophic circulations are often created

in association with baroclinic wave-breaking events. The ageostrophic circulations enhance the deformation fields due to the large-scale winds and drive stratospheric air deep into the troposphere along isentropic surfaces. During these tropopause

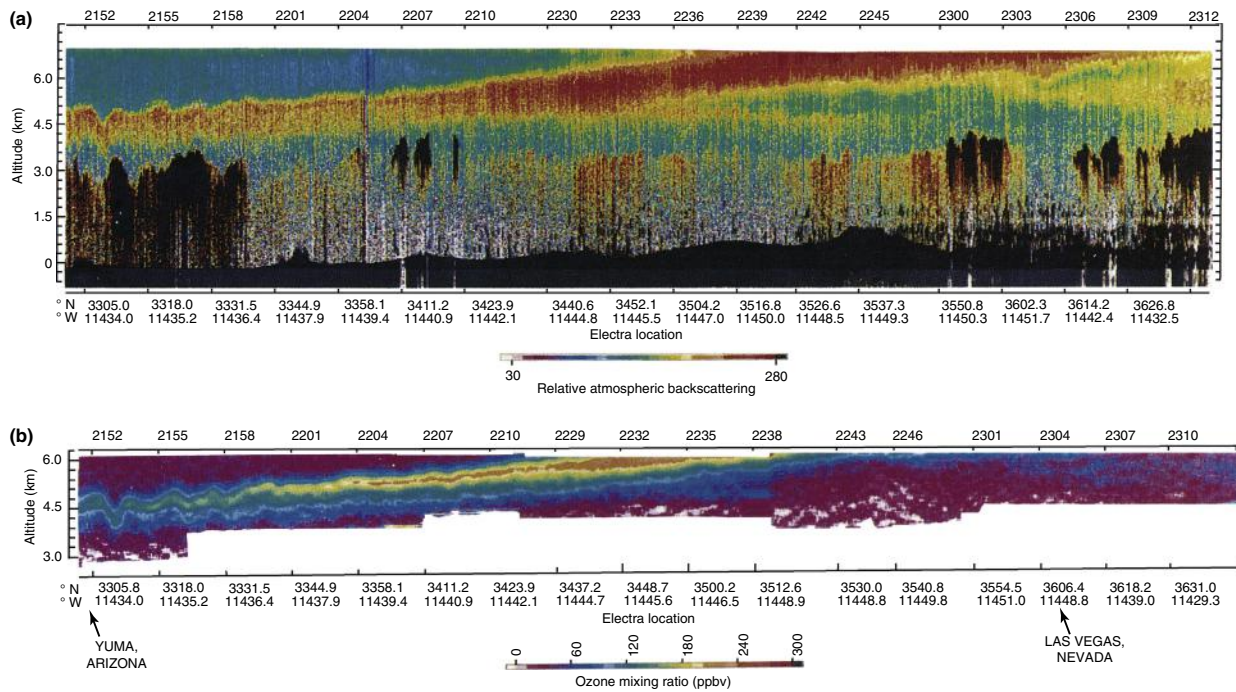


Figure 4 Cross-section of tropopause fold event on 20 April 1984. Color-scale displays of (a) relative aerosol distributions and (b) ozone mixing ratios. In each case, the higher values of the parameter are indicated by yellow and orange. The mixing of the fold by shearing instabilities can be seen at its leftmost edge.

folding events, sheets of stratospheric air with a small vertical to horizontal aspect ratio become embedded deep within the troposphere. The ageostrophic circulation associated with tropopause folding is transverse to the jet, with the strongest downward motion generally occurring in the northerly flow to the west of the upper level trough, near the jet entrance region (point A [Figure 5](#)). The exchange is associated with both mixing and diabatic processes. The mixing occurs mostly in areas of strong upward and downward motion, as shown by several high-resolution modeling studies. Diabatic effects (latent heat release in clouds and radiative heating/cooling in the vicinity of

clouds) seem to occur mostly in the center of the curvature of the jet stream. Significant STE occurs during this process. In fact, tropopause folding is considered the most evident form of STE.

Under some circumstances tropospheric or stratospheric filaments wind up so as to consist of interwoven regions of stratospheric and tropospheric air. This can create medium-scale potential vorticity anomalies; positive when stratospheric filaments wind up in the troposphere, and negative when tropospheric filaments wind up in the stratosphere. These anomalies will typically be associated with closed circulations – circulations that are temporarily resilient to deformation by the

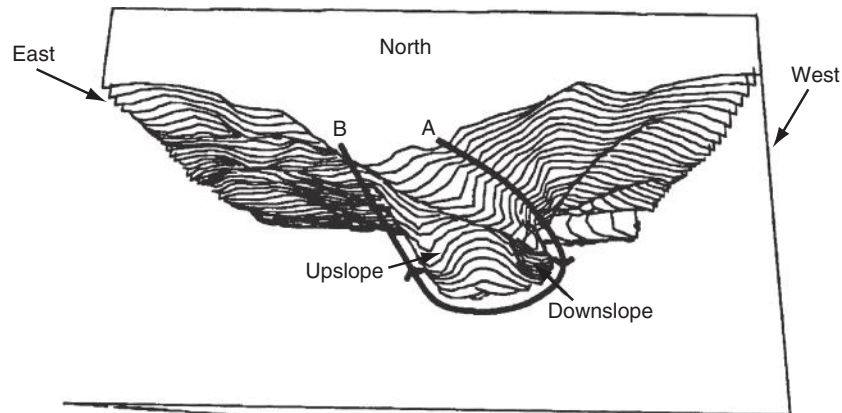


Figure 5 Three-dimensional view of the tropopause (PV2 PVU) during a tropopause folding. The jet entrance (exit) is indicated by A (B). Due to the ageostrophic circulation, the tropopause undergoes a strong downward motion at the western edge of the fold, followed by a strong upward motion at the eastern edge of the fold. Reproduced with permission from Lamarque and Hess (1994).

large-scale flow field. The corresponding cutoff cyclones (when a high potential vorticity region becomes trapped in the troposphere) and cutoff anticyclones (when a low potential vorticity region becomes trapped in the troposphere), are often rather long-lived, subject only to slow decay through mixing and turbulent processes, radiative processes, and convective mixing (in the cutoff cyclones). All the above processes will result in STE.

Other processes may also contribute to extratropical STE. Because potential vorticity is not conserved in the presence of a heating rate gradient, radiative heating in the vicinity of the tropopause is likely to be important, for example, the local heating induced by the high cirrus cloud shield associated with synoptic storms. Lidar (light detection and ranging) ozone measurements and high-resolution modeling suggest that gravity waves excited at the surface by strong winds over steep terrain may at times be responsible for mixing at the tropopause level. Strong thunderstorms in the extratropics occasionally penetrate the tropopause, presumably resulting in STE, although the effect of these storms has not been adequately documented. Both extratropical convection and topographic gravity waves will only result in STE under specific conditions: topographically forced gravity waves only occur under specific wind conditions and intense convection is most likely to during the summer months and over land. While it is difficult to extrapolate from local events to their global effects, the global importance of these processes is likely to be small, although their local effects may be significant.

See also: Dynamical Meteorology: Baroclinic Instability; Critical Layers. **Stratosphere/Troposphere Exchange and Structure:** Tropopause. **Tropical Meteorology and Climate:** Monsoon: Overview.

Further Reading

- Danielsen, E.F., 1968. Stratospheric–tropospheric exchange based upon radioactivity, ozone, and potential vorticity. *Journal of the Atmospheric Sciences* 35, 502–518.
- Dunkerton, T.J., 1995. Evidence of meridional motion in the summer lower stratosphere adjacent to monsoon regions. *Journal of Geophysical Research* 100, 16675–16688.
- Holton, J.R., Haynes, P.H., McIntyre, M.E., et al., 1995. Stratosphere–troposphere exchange. *Review of Geophysics* 33, 403–439.
- Hoskins, B.J., McIntyre, M.E., Robertson, A.W., 1985. On the use and significance of isentropic potential vorticity maps. *Quarterly Journal of the Royal Meteorological Society* 111, 877–946.
- Pierrehumbert, R.T., Yang, H., 1993. Global chaotic mixing on isentropic surfaces. *Journal of the Atmospheric Sciences* 50, 2464.
- Randel, W.J., Wu, F., Russell III, J.M., Zawodny, J.M., Oltmans, S.J., 2001. The seasonal variation of water vapor in the lower stratosphere observed in HALOE data. *Journal of Geophysical Research* 106, 14313–14325.
- Shapiro, M.A., 1980. Turbulent mixing within tropopause folds as a mechanism for the exchange of chemical constituents between the stratosphere and the troposphere. *Journal of the Atmospheric Sciences* 37, 994–1004.

Tropopause

M Dameris, Deutsches Zentrum für Luft- und Raumfahrt, Institut für Physik der Atmosphäre, Oberpfaffenhofen, Wessling, Germany

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

The tropopause is an important boundary layer in Earth's atmosphere dividing the lowermost atmospheric layer, the troposphere, from the stratosphere. Depending on the geographical region, it can be found at a height of between 6 and 18 km. The troposphere is characterized by the weather conditions. The stratosphere is stably stratified due to an increasing temperature with height that is caused by a pronounced ozone layer at a height of between about 15 and 30 km.

Introduction

The tropopause denotes the natural limit between the troposphere (Greek *tropos* = turn; troposphere = turning or mixing sphere) and the stratosphere (stratified as opposed to mixed). These two atmospheric regions differ in various dynamical and chemical parameters, such as the vertical temperature structure, the potential vorticity, and the concentration of climatically relevant chemical species (e.g., ozone and water vapor). The tropopause is often obvious in observations or analyses of these dynamical and chemical parameters. In particular, high-resolution radiosonde measurements of the thermal and wind structure of the extratropical tropopause region exhibit a strong increase of temperature just above a sharp local cold-point tropopause. In a tropopause-based coordinate system, a persistent inversion is clearly identified in a narrow layer above the tropopause. However, in order to compile quantitative statistics of tropopause properties, one needs an objective definition that picks out the tropopause as accurately as possible. The tropopause can be defined in terms of physical parameters. The tropopause can exist anywhere between about 70 hPa (~18 km) and 400 hPa (~6 km), and it is therefore not convenient to use a constant pressure level to describe the tropopause. In the tropics, the tropopause is generally located at higher altitudes than in polar regions; the height of the tropopause depends on the season and is affected by weather conditions in the troposphere. Long-term changes are observed, and it has been proposed that the tropopause height can be used as a robust indicator of climate change.

Thermal Definition of the Tropopause

The conventional ('classical') definition of the tropopause is based on differences in the temperature lapse rates between stratospheric and tropospheric air. In the turbulently mixed troposphere, the temperature generally decreases with height. In the stably stratified stratosphere, the temperature increases with height owing to the absorption of solar ultraviolet radiation in the ozone layer (*see Ozone Depletion and Related Topics: Ozone as a UV Filter*). In 1957, the World Meteorological Organization (WMO) defined the thermal tropopause as the lowest level at which the lapse rate decreases to 2 K km^{-1} or less, provided that the average lapse rate between this level and all higher levels within 2 km does not exceed 2 K km^{-1} . The latter condition eliminates stable layers in the planetary boundary

layer (*see Boundary Layer (Atmospheric) and Air Pollution: Stably Stratified Boundary Layer*), which occasionally occur over high-latitude continents during winter, leading to the erroneous assignment of a tropopause level. The thermal WMO criterion is operationally applied to determine the tropopause height from individual radiosonde temperature soundings.

The main advantage of the thermal tropopause definition can be seen in its simplicity. In order to determine the tropopause, the vertical profile of only the temperature needs to be provided, which is in most cases well known. The thermal definition can be globally applied. However, this definition of the tropopause may lead to ambiguities owing to the presence of multiple stable layers, particularly in the vicinity of the jet streams (*see Synoptic Meteorology: Jet Streaks*). A further disadvantage is that the temperature lapse rate is not conserved in adiabatic flow. This may lead to large displacements of the tropopause during changes in the static stability by convergence. Moreover, the thermal definition obscures the fact that the tropopause often behaves as if it were a material surface to varying degrees of approximation.

Dynamical Definition of the Tropopause

A second definition of the tropopause is based on dynamical features. This definition employs the isentropic potential vorticity P , which can be expressed mathematically as in eqn [1].

$$P = -g(f + \zeta_\theta)\partial\Omega/\partial p \quad [1]$$

Here, g denotes the acceleration of gravity, f is the Coriolis parameter ($=2\Omega \sin\phi$), ζ_θ is the relative vorticity along an isentropic surface, θ is the potential temperature, p is the atmospheric pressure, and ϕ is the geographical latitude. The location of the so-called dynamical tropopause is assigned to the position where the potential vorticity P reaches a critical value. The tropopause separates low values of P in the troposphere from high values in the stratosphere. The WMO defines the dynamical tropopause by the threshold value of $P = 1.6 \text{ PVU}$, where PVU stands for 'potential vorticity unit' ($1 \text{ PVU} = 1 \times 10^{-6} \text{ K m}^2 \text{ kg}^{-1} \text{ s}^{-1}$). The potential vorticity is determined by both the stability and the three-dimensional wind field and is conserved in adiabatic flow. Compared to the thermal definition, this guarantees a better spatial and temporal continuity of the dynamically defined tropopause. Therefore, the dynamical method is sometimes superior to the thermal approach, especially in the vicinity of developing

baroclinic disturbances. However, the potential vorticity approach fails in the tropics, since the vertical component of absolute vorticity ($f + \zeta_\theta$) changes the sign at the Equator. Hence, the dynamical method of defining the tropopause is applicable only poleward of 10° latitude. Similar limitations exist even in the extratropical regions, where P locally becomes small within strong anticyclonic flows.

In the literature, threshold values defining the dynamical tropopause range from 1 to 5 PVU. A systematic global analysis using P -values ranging from 2 to 5 PVU indicated that the tropopause pressure was generally overestimated for values less than 3 PVU and systematically underestimated for values greater than 4 PVU compared to the pressure of the thermal tropopause. Grewe and Dameris employed 10-year statistics based on global data of the European Centre for Medium-Range Weather Forecasts (ECMWF) and found that a value of 3 PVU represents an acceptable value for the tropopause, especially in midlatitudes of the winter hemisphere. Additionally, they showed that values of P calculated at the height of the thermally defined tropopause exceed at least 2 PVU, independently of the season. This indicates that, following the WMO criteria, the 1.6 PVU surface generally lies below the thermally defined tropopause. However, Hoinka and coworkers pointed out that the 1.6 PVU tropopause well describes variability in space and time in comparison with tropopauses determined from radiosonde measurements. **Figure 1** shows a comparison of calculated zonal mean values

for a single day that can be employed to estimate the height of the tropopause. It is obvious that the extratropical thermal tropopause corresponds roughly to a constant potential vorticity surface of about 3 PVU.

The Tropopause as a Chemical Transition Layer

There are many measurements of chemical species in the upper troposphere and the lower stratosphere indicating an abrupt transition of mixing ratios and concentrations of chemical compounds in the vicinity of the thermal-dynamical tropopause. The most prominent examples are ozone and water vapor.

Mean tropospheric ozone concentrations in northern midlatitudes are typically less than $1 \times 10^{18} \text{ m}^{-3}$, showing no clear vertical gradient. In the lower stratosphere, ozone concentrations steadily increase up to (mean) maximum ozone values of $5 \times 10^{18} \text{ m}^{-3}$ at around 22 km height. The transition layer of low tropospheric to high stratospheric ozone concentrations is near the tropopause. The 'ozone tropopause' (ozonopause) is mostly found below the thermal tropopause. For example, in midlatitudes of the Northern Hemisphere, the ozonopause is located about 800 m below the thermal tropopause. So far, no official definition of the ozonopause has been released by the WMO. Based on the analysis of about 600 ozone sonde profiles at different stations in Northern Europe

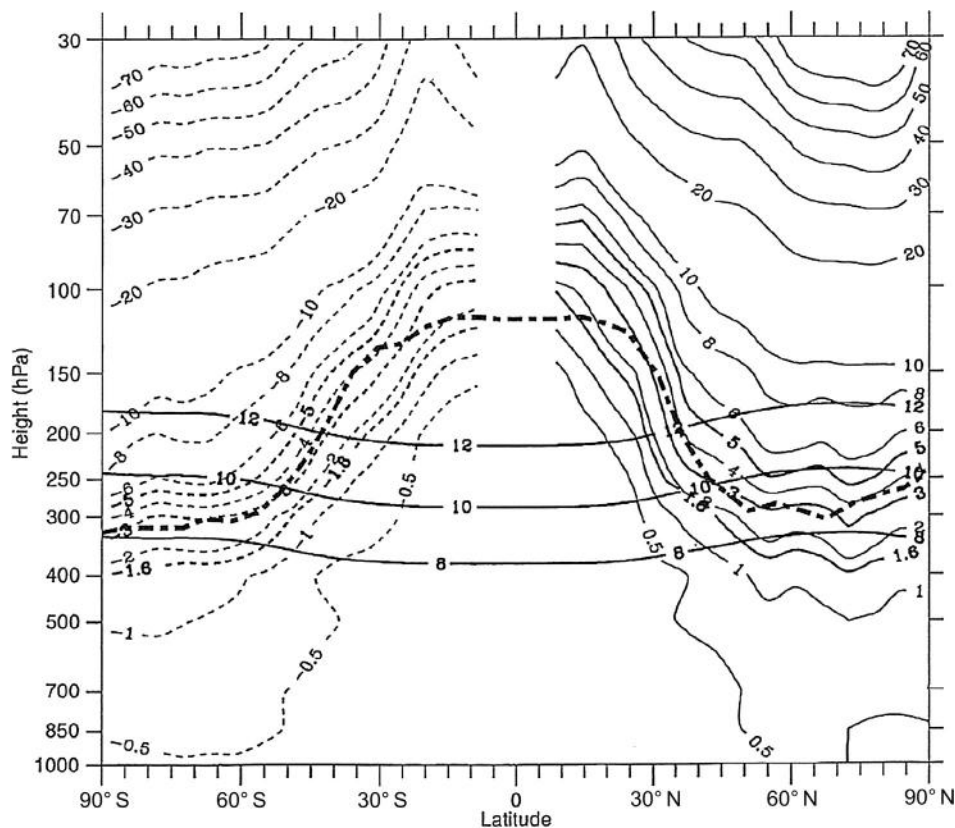


Figure 1 Zonally averaged values of potential vorticity P (thin dashed and thin solid lines, units in PVU; see text), geometric height (heavy lines, in km), and the thermally defined tropopause (heavy dashed line) for 1 January 1985. Reproduced from Grewe, V., Dameris, M., 1996. Calculating the global mass exchange between stratosphere and troposphere. *Annales de Geophysique* 14, 431–442.

between 1991 and 1994, Bethan and coworkers have suggested defining the 'ozone tropopause' as the lowest height at which the following three conditions are valid:

1. The vertical gradient of the ozone mixing ratio is greater than 60 ppbv (60 ppb (10^{-9}) by volume) per kilometer, calculated over a height of 200 m.
2. The ozone mixing ratio exceeds 80 ppbv.
3. The ozone mixing ratio directly above the tropopause is greater than 100 ppbv.

Water vapor mixing ratios decrease strongly with height within the troposphere. Lower tropospheric mixing ratios are typically greater than 1000 ppmv (1000 ppm (10^{-6}) by volume), depending on the geographical region and season. The strong vertical gradient in water vapor profiles is obvious up to heights near the tropopause, the so-called hygropause. The existence of a significant decrease in the water vapor mixing ratio above the hygropause in midlatitudes had been evident from the first measurements, which were reported in 1946. For example, in the Northern Hemisphere, lower-stratosphere mean water vapor mixing ratio values of 3–5 ppmv are measured. The hygropause was discovered in the tropics more than three decades later, in 1979. A detailed global analysis of tropopause water vapor mixing ratios (beside other tropopause parameters) was given by Hoinka in 1999, employing ECMWF analyses in combination with in situ and satellite data for water vapor. He found that the tropopause mixing ratio of water vapor reaches a minimum in the tropics above the Pacific and Indian Oceans.

Distribution of the Tropopause Height

Ignoring the various possibilities for defining the tropopause, which yield some differences in determination of the pressure (see [Figure 1](#)), the tropopause pressure depends strongly on the geographical latitude, season, and type of location (continental or oceanic). Hoinka carried out a detailed statistical analysis of global tropopause pressure, employing data provided by the ECMWF for the period between 1979 and 1993. His analysis of the frequency distribution of the tropopause pressure above tropical, midlatitude, and polar regions, classified according to continental and oceanic areas, yielded the following:

- Above the Southern Hemisphere, the tropopause is of more zonal character than above the Northern Hemisphere, where the continents generate zonal wave patterns.
- Above both hemispheres, the strongest meridional gradients in the tropopause pressure are found between 20° and 40° latitude.
- Strong standard deviations in tropopause pressure are observed in the regions of the Atlantic and Pacific storm tracks.

Conclusions

The explanations given so far assume implicitly that the tropopause is a two-dimensional surface. However, due to recent analyses and investigations, it may be more useful to think of a 'tropopause zone' or a 'tropopause layer' that is a few

kilometers thick and in which vertical and horizontal mixing are both significant. Highwood and Hoskins discussed the results of several studies that seem to indicate that the tropical tropopause is generally far from being a single well-defined level. Gettelman and Forster defined a 'Tropical Tropopause Layer' (TTL) as the layer between the level of maximum convective outflow and the cold-point tropopause. Another definition has been provided by Fueglistaler and coworkers as a shallower layer between the level of zero clear-sky radiative heating and the cold-point tropopause (15–19 km). A layer of finite depth is also defined for the extratropics (i.e., the extratropical tropopause transition layer (ExTL)). Shepherd interpreted the transition as the result of recurrent wave-breaking events, forced by synoptic-scale baroclinic disturbances stirring tropospheric and stratospheric air masses with different chemical characteristics. Hegglin and coworkers showed that the ExTL is a global feature with increasing depth toward higher latitudes, and has been found to be different for different chemical tracers.

Nevertheless, it is important to know the temporal and spatial structure of the tropopause and of meteorological and chemical parameters within the transition zone between the troposphere and the stratosphere. This is necessary in order to estimate the exchange of mass, water vapor, ozone, and other chemical compounds between these two atmospheric layers. The transport of mass through the tropopause is for many purposes a useful measure of global-scale stratosphere–troposphere exchange (see **Stratosphere/Troposphere Exchange and Structure: Global Aspects**), in particular on seasonal or longer time scales. Detailed knowledge of the air mass exchange between the stratosphere and the troposphere is required for reliable estimates of the effects of natural and anthropogenic emissions on atmospheric composition and therefore on climate; it must be understood for the prediction of global change.

See also: Boundary Layer (Atmospheric) and Air Pollution: Stably Stratified Boundary Layer. Ozone Depletion and Related Topics: Ozone as a UV Filter. Stratosphere/Troposphere Exchange and Structure: Global Aspects; Local Processes. Synoptic Meteorology: Jet Streaks.

Further Reading

- Bethan, S., Vaughan, G., Reid, S.J., 1996. A comparison of ozone and thermal tropopause heights and the impact of tropopause definition on quantifying the ozone content of the troposphere. *Quarterly Journal of the Royal Meteorological Society* 122, 929–944.
- Birner, T., 2006. Fine-scale structure of the extratropical tropopause region. *Journal of Geophysical Research* 111, D04104. <http://dx.doi.org/10.1029/2005JD006301>.
- Danielsen, E.F., Hipskind, R.S., 1980. Stratospheric–tropospheric exchange at polar latitudes in summer. *Journal of Geophysical Research* 85, 393–400.
- Dobson, G.M.B., Brewer, A.W., Cwilog, B.M., 1946. Meteorology of the lower stratosphere. *Proceedings of the Royal Society of London A* 185, 144–175.
- Fueglistaler, S., Dessler, A.E., Dunkerton, T.J., et al., 2009. Tropical tropopause layer. *Reviews of Geophysics* 47, RG1004. <http://dx.doi.org/10.1029/2008RG000267>.
- Gettelman, A., Forster, P.M.F., 2002. Climatology of the tropical tropopause layer. *Journal of the Meteorological Society of Japan* 80, 911–924.
- Grewe, V., Dameris, M., 1996. Calculating the global mass exchange between stratosphere and troposphere. *Annales de Geophysique* 14, 431–442.

- Hegglin, M., et al., 2009. A global view of the extratropical tropopause transition layer from Atmospheric Chemistry Experiment Fourier Transform Spectrometer O₃, H₂O, and CO. *Journal of Geophysical Research* 114, D00B11. <http://dx.doi.org/10.1029/2008JD009984>.
- Highwood, E.J., Hoskins, B.J., 1998. The tropical tropopause. *Quarterly Journal of the Royal Meteorological Society* 124, 1579–1604.
- Hoerling, M.P., Schaack, T.D., Lenzen, A.J., 1991. Global objective tropopause analysis. *Monthly Weather Review* 119, 1816–1831.
- Hoinka, K.P., 1997. The tropopause: discovery, definition and demarcation. *Meteorol. Z.* 6, 281–303.
- Hoinka, K.P., 1998. Statistics of the global tropopause pressure. *Monthly Weather Review* 126, 3303–3325.
- Hoinka, K.P., 1999. Temperature, humidity, and wind at the global tropopause. *Monthly Weather Review* 127, 2248–2265.
- Hoinka, K.P., Reinhardt, M.E., Metz, W., 1993. North Atlantic air traffic within the lower stratosphere: cruising times and corresponding emissions. *Journal of Geophysical Research* 98, 23113–23131.
- Holton, J.R., Haynes, P.H., McIntyre, M.E., et al., 1995. Stratosphere–troposphere exchange. *Reviews of Geophysics* 33, 403–439.
- Oltmans, S.J., Hofmann, D.J., 1995. Increase in lower stratospheric water vapour at a mid-latitude Northern Hemisphere site from 1981 to 1994. *Nature* 374, 146–149.
- Randel, W.J., Wu, F., Forster, P., 2007. The extratropical tropopause inversion layer: global observations with GPS data, and a radiative forcing mechanism. *Journal of Atmospheric Science* 64, 4489–4496.
- Reed, R.J., 1955. A study of a characteristic type of upper level frontogenesis. *Journal of Meteorology* 12, 226–237.
- Reiter, E.R., 1975. Stratospheric–tropospheric exchange processes. *Reviews of Geophysics and Space Physics* 13, 459–474.
- Santer, B.D., et al., 2003. Behavior of tropopause height and atmospheric temperature in models, reanalyses, and observations: decadal changes. *Journal of Geophysical Research* 108, 4002. <http://dx.doi.org/10.1029/2002JD002258>.
- Shepherd, T.G., 2007. Transport in the middle atmosphere. *Journal of the Meteorological Society of Japan* 85B, 165–191.
- Sherwood, S.C., Dessler, A.E., 2000. On the control of stratospheric humidity. *Geophysical Research Letters* 27, 2513–2516.
- Steinbrecht, W., Claude, H., Köhler, U., Hoinka, K.P., 1998. Correlations between tropopause height and total ozone: implications for long-term changes. *Journal of Geophysical Research* 103, 19183–19192.
- WMO, 1957. Definition of the Tropopause. Bulletin 6. World Meteorological Organization, Geneva, Switzerland.
- WMO, 2007. Scientific Assessment of Ozone Depletion: 2006. Global Ozone Research and Monitoring Project – Report No. 50. World Meteorological Organization, Geneva, Switzerland.

SYNOPTIC METEOROLOGY

Contents

Anticyclones

Forecasting

Weather Maps

Cyclogenesis

Extratropical Cyclones

Fronts

Fronts in the Lower Stratosphere

Frontogenesis

Jet Streaks

Lake-Effect Storms

Polar Lows

Thermal Low

Anticyclones

SJ Colucci, Cornell University, Ithaca, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

Anticyclones are horizontal circulations around relatively high sea-level pressure or high geopotential height on isobaric surfaces aloft. The structure, climatology, dynamics, and impacts of anticyclones are reviewed. Special kinds of anticyclones, such as blocking highs, subtropical highs, and stratospheric anticyclones associated with sudden stratospheric warmings are considered.

Introduction

Anticyclones are regions of relatively high pressure on horizontal surfaces, or high geopotential height on isobaric surfaces, around which air circulates clockwise in the Northern Hemisphere and counterclockwise in the Southern Hemisphere. Anticyclones are therefore characterized by negative relative vorticity and low but positive absolute vorticity in the Northern Hemisphere, while in the Southern Hemisphere they are distinguished by positive relative vorticity and low but negative absolute vorticity. On sea-level pressure or geopotential height analyses, they may be subjectively identified by closed isobars or height contours, whereas in vorticity analyses they may be objectively identified by relative vorticity minima in the Northern Hemisphere and maxima in the Southern Hemisphere.

At sea level, anticyclones typically originate as cold, shallow circulations that migrate equatorward and evolve into warm, subtropical high-pressure systems penetrating well into the troposphere. Aloft, anticyclones may appear at middle and high latitudes on isobaric surfaces. From hydrostatic considerations, these are relatively warm systems. Anticyclones aloft are often stationary or westward drifting and thus may block the eastward progress of other weather systems. Anticyclonic circulations at high latitudes may penetrate into the stratosphere where they may be associated with sudden stratospheric warmings (SSWs).

Although not as actively researched as cyclones, anticyclones are important because the clear, dry conditions usually associated with them may allow strong nighttime radiative cooling and cold surface temperatures. The convectively stable air of anticyclones may allow air pollutants to concentrate near

the earth's surface. Finally, the blocking action of anticyclones aloft may cause persistently anomalous weather conditions at the earth's surface.

Structure

Anticyclones may either be cold-core or warm-core systems. An example of each type is presented in **Figures 1** and **2**, respectively. Cold-core anticyclones are typically found on the poleward side of the midlatitude baroclinic zone. These are shallow systems with an anticyclonic circulation confined to the lower troposphere. The geostrophic relative vorticity (curvature in the isobars or geopotential height contours) is anticyclonic near the earth's surface but becomes cyclonic (or less anticyclonic) by the middle troposphere. The region of high sea-level pressure over northwestern North America in **Figure 1** is a cold-core anticyclone; notice its proximity to a local minimum in the 1000–500 mb thickness which, hydrostatically, is proportional to the vertically averaged temperature in the 1000–500 mb layer. This sea-level anticyclone is located between cyclonic and anticyclonic features at the 500-mb level (not shown).

Warm-core anticyclones are found equatorward of baroclinic zones and are characterized by circulations that remain or may become increasingly anticyclonic from sea level to the middle troposphere. The region of high sea-level pressure over the eastern Pacific Ocean in **Figure 2** is a warm-core anticyclone. Notice that it is located in a region of relatively high 1000–500 mb thickness (vertically averaged lower tropospheric temperatures) on the equatorward side of the midlatitude baroclinic zone (thickness gradient). This sea-level anticyclone is located directly beneath a closed, 500-mb anticyclone (not shown), indicating that it is a deep, warm system.

Both warm-core and cold-core anticyclones are characterized by gently subsiding vertical motion in the troposphere. This subsidence favors clear skies promoting strong nighttime radiative cooling of the earth's surface near the centers of these anticyclones. The adiabatic warming of the sinking air coupled with radiative cooling at the surface often produces an inversion in the vertical temperature profile; this inversion may be eroded or destroyed by daytime radiative heating and vertical mixing in the boundary layer. Regardless, anticyclones are distinguished by strong static stability.

Dynamics

Convergence of mass in the upper troposphere is the primary mechanism responsible for the relatively high sea-level pressure at anticyclone centers. From considerations of gradient wind balance, this mass convergence occurs downwind of anticyclonic circulations, or near regions of anticyclonic vorticity advection. The formation of new, cold-core anticyclones (or anticyclogenesis) is favored when this mass convergence occurs over a lower-tropospheric cold-air pool that, hydrostatically, would be associated with relatively high sea-level pressure. Warm anticyclogenesis may occur if mass convergence occurs over relatively high sea-level pressure at lower latitudes.

More commonly, cold anticyclones evolve into warm anticyclones as follows. The circulation around cold anticyclones draws cold air equatorward, forcing the sea-level pressure to locally rise. The anticyclone relocates toward rising sea-level pressure. Thus, cold anticyclones usually drift equatorward with time.

The mass convergence over the cold anticyclone forces air to sink through the troposphere and to adiabatically warm. The anticyclone thus becomes warmer over time and may eventually be located equatorward of the midlatitude baroclinic zone. Frictionally induced mass divergence at the earth's surface forces the sea-level pressure to fall at the anticyclone center, which then weakens. The anticyclone may reintensify as a warm system if mass convergence aloft exceeds in magnitude the lower tropospheric mass divergence near the anticyclone center. This is thought to be especially true of the quasistationary subtropical anticyclones, e.g., the Bermuda-Azores High of the Northern Hemisphere; mass convergence aloft associated with monsoonal circulations may help to maintain these systems.

Other mechanisms may contribute to anticyclone formation and intensification. While there is usually very little temperature advection in the lower troposphere over anticyclone centers, the advection of cold air in the upper troposphere and lower stratosphere over a sea-level anticyclone center may contribute to its intensification.

The convergence of mass in the upper troposphere over a developing or intensifying sea-level anticyclone is associated with sinking air and adiabatic warming in the troposphere, but rising air and adiabatic cooling in the stratosphere. Although vertical motions typically are an order of magnitude smaller in the stratosphere than in the troposphere, the static stability in the stratosphere is several orders of magnitude larger than that in the troposphere such that adiabatic temperature changes can be larger in the stratosphere than in the troposphere. Thus, while there is sinking motion and adiabatic warming in the troposphere over a developing or intensifying sea-level anticyclone center, there will be rising air and greater adiabatic cooling in the stratosphere, resulting in net cooling of the air column and, hydrostatically, sea-level pressure rises at the anticyclone center.

Clear conditions near cold anticyclone centers may result in the formation of ice fogs, and radiative heat loss from these fogs may contribute to sea-level pressure rises and anticyclone intensification. However, this effect is believed to be small. The sensible cooling of cold anticyclones by snow and ice-covered land surfaces may contribute to their formation and intensification. Similarly, the sensible cooling of warm anticyclones by relatively cool sea-surface temperatures over the eastern oceans favors these systems, especially during the warm seasons. Conversely, the sensible warming of cold anticyclones by relatively warm water in the cool seasons weakens these systems.

Climatology

Early climatologies of weather systems were constructed through manual inspection of sea-level pressure charts and subjective identification of centers of closed isobars.

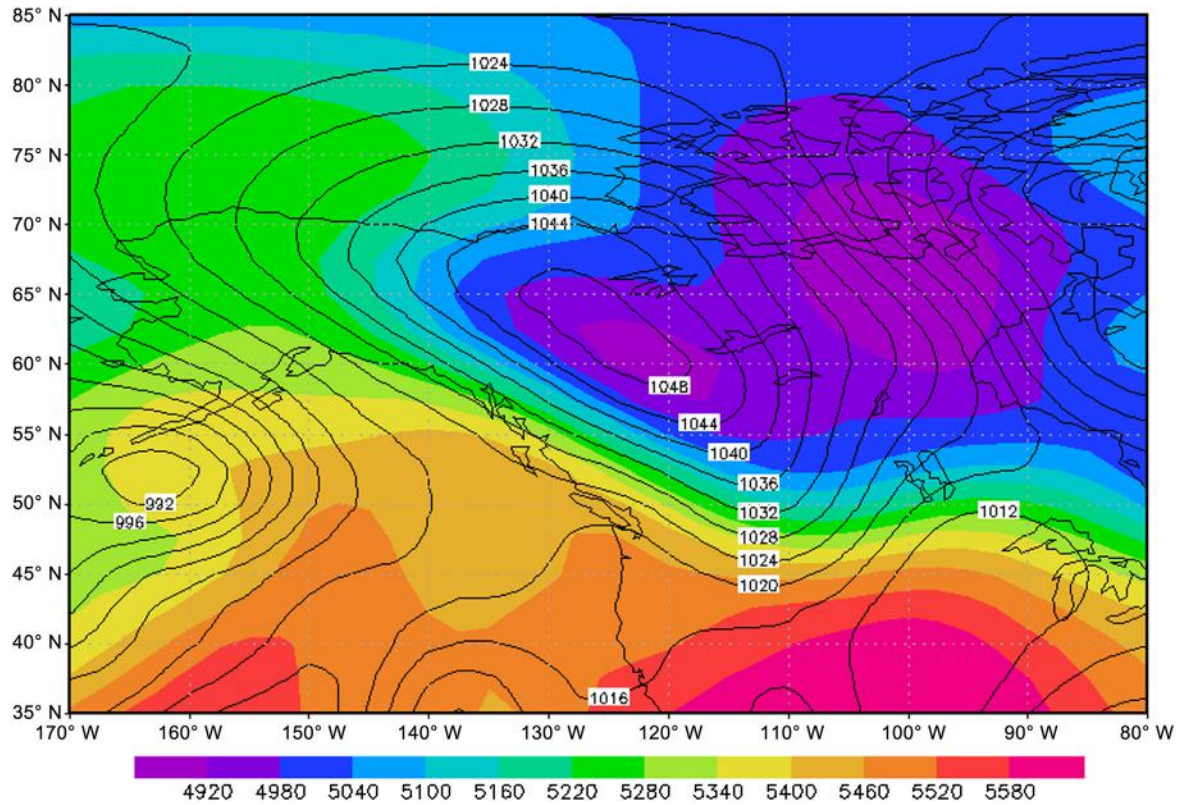


Figure 1 Sea-level pressure (solid contours in millibars at 4 mb intervals) and 1000–500 mb thickness (colors in dekameters, at 60 m intervals) at 0000 UTC 23 January 2009. Notice the sea-level anticyclone (closed isobars around relatively high sea-level pressure) near 120° W, 60° N, collocated with an area of low 1000–500 mb thickness (vertically averaged temperature in the 1000–500 mb layer).

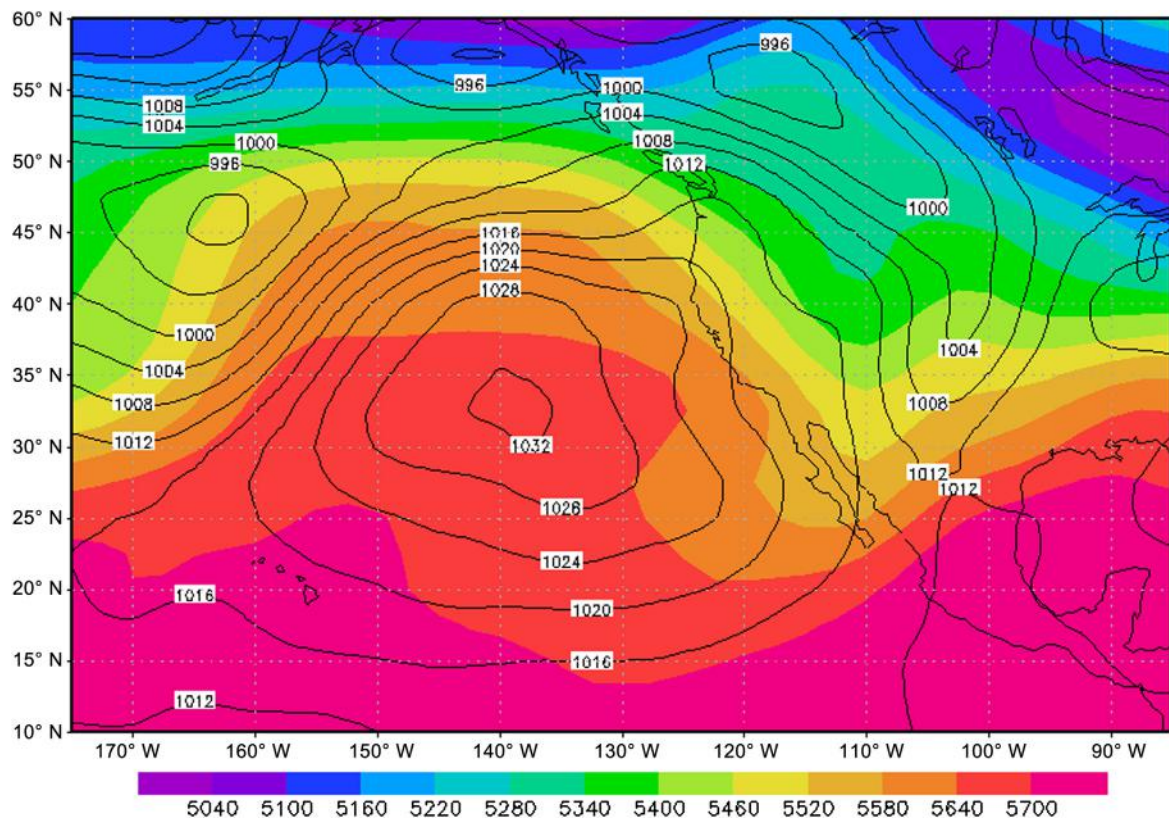


Figure 2 As in Figure 1 except at 0000 UTC 6 January 2009. Notice the sea-level anticyclone near 140° W, 30° N, collocated with an area of relatively high 1000–500 mb thickness.



Figure 3 Sea-level anticyclone track density, in number of centers per month, in the 1980–86 analyses of the European Center for Medium-Range Weather Forecasts. From Sinclair, M.R., Watterson, I.G., 1999. *Journal of Climate* 12, American Meteorological Society, Boston, pp. 3467–3485.



Figure 4 As in Figure 3 but for the Southern Hemisphere. From Sinclair, M.R., Watterson, I.G., 1999. *Journal of Climate* 12, American Meteorological Society, Boston, pp. 3467–3485.

Contemporary investigations employ automated procedures to objectively identify these systems from maxima in sea-level pressure analyses. Such investigations reveal that, over the Northern Hemisphere (Figure 3), sea-level anticyclones are most frequently analyzed in a band over the midlatitude Pacific Ocean, in a broad band near North America and centered on the Great Lakes, and near Mongolia. Over the Southern Hemisphere, similarly defined sea-level anticyclones are concentrated within a midlatitude (25° S– 45° S) band with little longitudinal variation (Figure 4).

In the middle troposphere over the Northern Hemisphere (Figure 5), anticyclones objectively identified from maxima in geopotential height fields are most frequently analyzed during the summer at low latitudes, especially over continental regions. During the winter these systems are rare but occasionally analyzed over high latitude oceans. These latter systems may be associated with blocking, as discussed below. No comparable results exist for the Southern Hemisphere to date, except for those anticyclones associated with blocking; these anticyclones are preferentially analyzed over the southern Pacific Ocean.

Blocking

Warm anticyclones that penetrate deep into the troposphere from the earth's surface in midlatitudes may, if stationary and persistent, block the normal eastward motion of other weather systems. Note that these are different from deep, warm subtropical anticyclones in that they are formed by the poleward breaking of midlatitude waves, resulting in a deep, warm

anticyclone that is detached from the warm air over the subtropics. As previously noted, subtropical anticyclones are formed by the equatorward drifting and warming of cold anticyclones, and are maintained by cold sea-surface temperatures and mass convergence aloft associated with monsoonal circulations.

Instead, a blocking anticyclonic circulation is maintained on its upstream flank by repeated fluxes of anticyclonic potential vorticity (lower tropospheric warm air and middle tropospheric anticyclonic vorticity). These fluxes are provided by smaller-scale waves approaching the blocking system, and help render the blocking anticyclone stationary and persistent. Additionally, the strong static stability of anticyclones helps maintain them against dissolution by convective mixing. Sinking motion in the statically stable, anticyclonic environment promotes adiabatic warming and maintenance of a deep, warm anticyclonic system.

Blocking anticyclones may be analyzed anywhere and at any time, but are favored during the cool season over the oceans, particularly the eastern oceans of the Northern Hemisphere and the eastern Pacific Ocean in the Southern Hemisphere. These locations are also downstream of the principal storm tracks and are also locations of climatologically preferred diffluent flow fields in the middle troposphere. The repeated interaction of small-scale systems with the diffluent flow enhances the diffidence (by making its poleward branch more anticyclonic) until an anticyclonic circulation is established aloft. Quasigeostrophically, this diffidence is enhanced by the local deposition of anticyclonic potential vorticity. Equatorward of the blocking anticyclone, the normal westerly flow may reverse to easterlies over a considerable longitudinal distance and for periods of a week or more. The blocking anticyclone

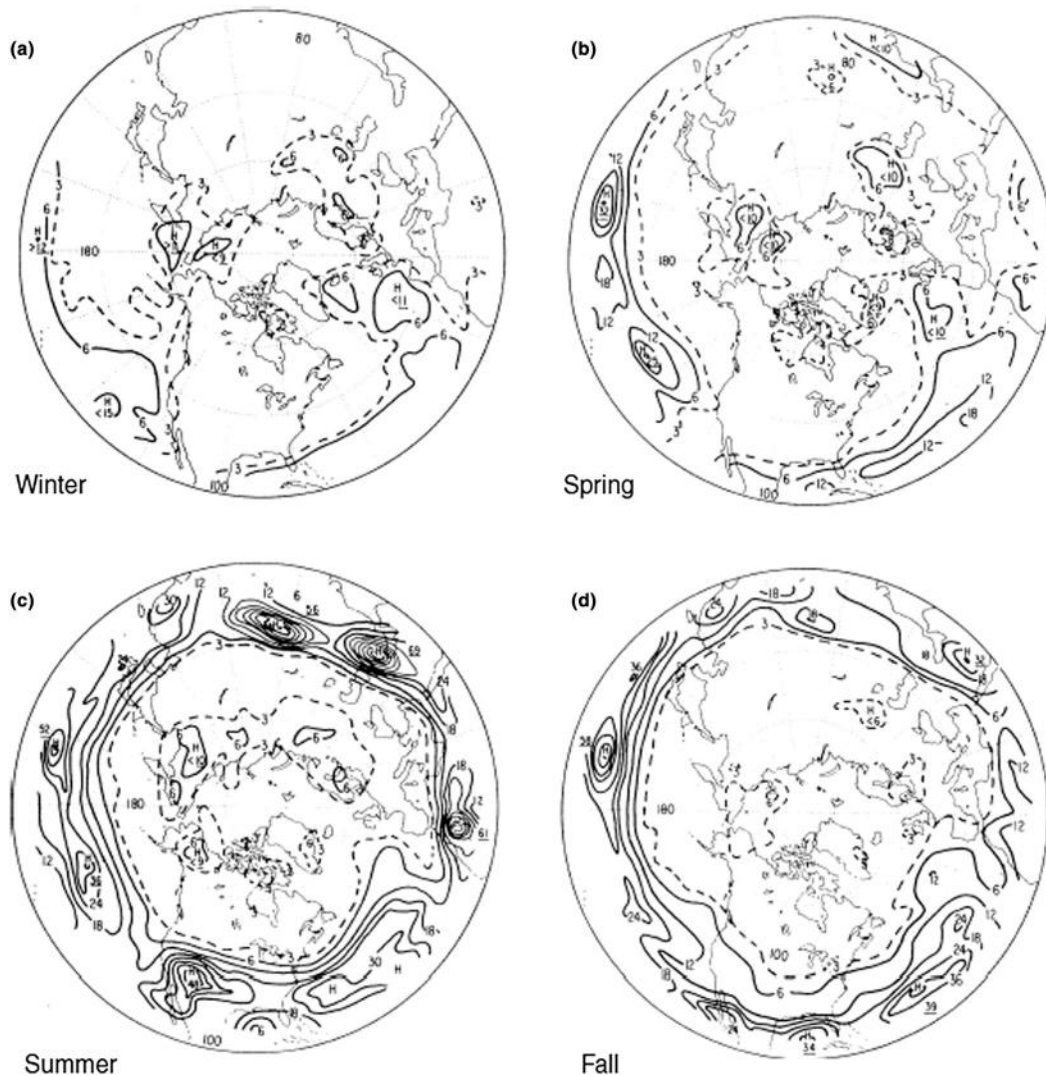


Figure 5 Number of twice-daily 500-mb geopotential height analyses during 1963–77 with anticyclones in 2° latitude by 5° grid boxes. From Bell, G.D., Bosart, L.F., 1989. *Monthly Weather Review*, vol. 117. American Meteorological Society, Boston, pp. 2142–2163.

may even penetrate into the stratosphere, causing a reversal of the zonal flow there from westerly to easterly and SSWs. In fact, almost all SSWs are preceded in time by recent tropospheric blocking events, but the converse is not true; blocking events are much more frequent than SSWs and thus are not necessarily each followed by an SSW.

An example of a blocking anticyclone is shown in [Figure 6](#). Notice the large 500-mb anticyclone over the eastern Atlantic Ocean, with highest 500-mb heights that are typical of much lower latitudes. An example of a stratospheric anticyclone is shown in [Figure 7](#). Notice the very large and broad anticyclone at 10 mb over northern North America. The zonally averaged zonal flow at 10 mb and 60° N reversed from westerly to easterly on the day of this analysis, defining an SSW. Although the analysis in [Figure 7](#) is 24 h after the analysis in [Figure 6](#), it is not known if the two phenomena are directly related or merely coincidental; the anticyclone in [Figure 7](#) may have resulted from the upward penetration of

a different, nonblocking anticyclonic circulation. Nevertheless, blocking anticyclones in general are of considerable scientific and practical importance.

Impact

Because they are characterized by clear skies and subsiding air, anticyclones are typically associated with fair weather. A stationary and persistent anticyclone may produce prolonged fair and dry weather conditions, depleting soil moisture, and stressing crops and water supplies. The strong stability of anticyclones may stagnate air near the earth's surface, leading to enhanced concentrations of pollutants. Clear skies near anticyclone centers favor strong nocturnal cooling near the earth's surface; these conditions during the growing season may damage crops. The equatorward circulation of cold air around anticyclones may cause sudden cold-air outbreaks over

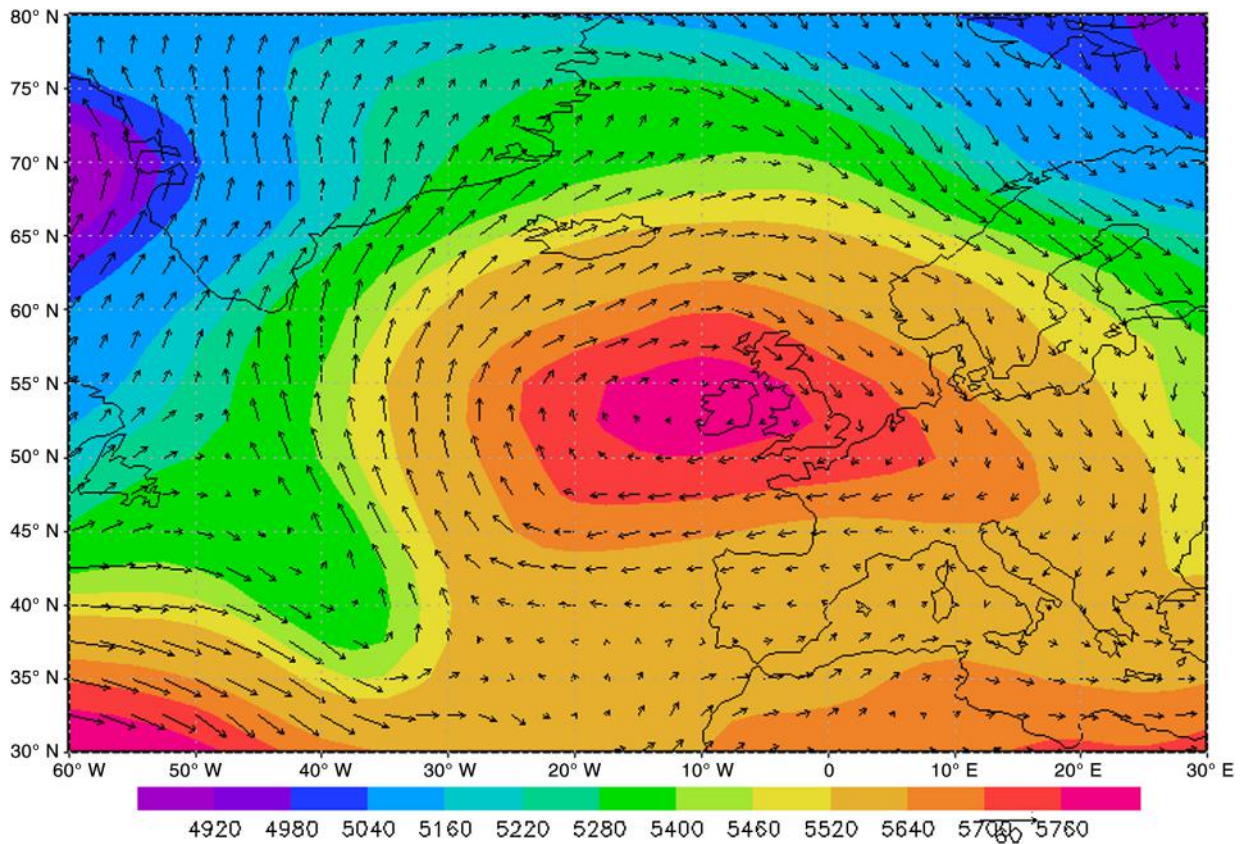


Figure 6 An example of a blocking anticyclone. Color contours are 500-mb geopotential height, in meters, and the arrows are 500-mb horizontal winds, analyzed at 0000 UTC 22 January 1987. Notice the closed contours around relatively high 500-mb heights near 10° W, 52.5° N, and the easterly winds along 45° N from 20° W to 10° E.

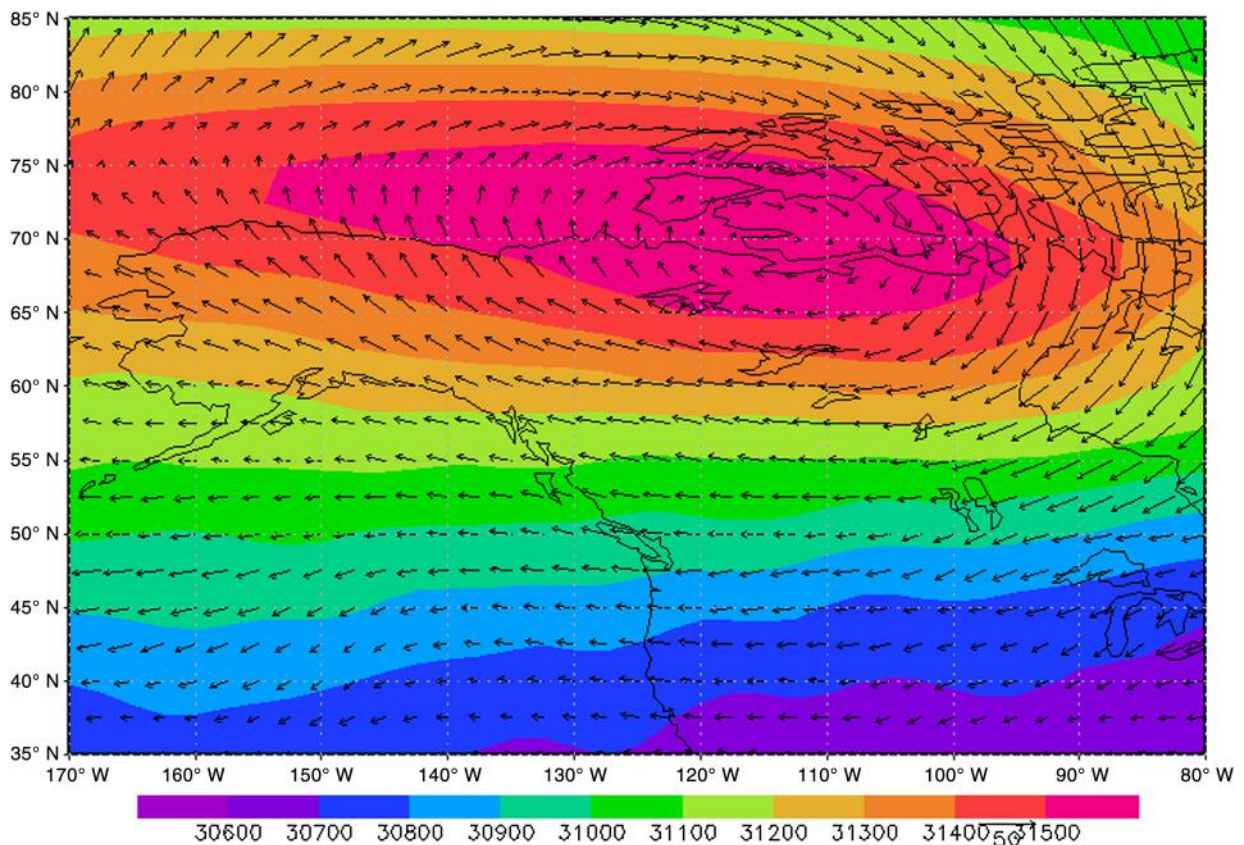


Figure 7 An example of a stratospheric anticyclone associated with a sudden stratospheric warming. As in Figure 6 except for 10 mb and 0000 UTC 23 January 1987. Notice the large area of easterly flow from 35° N to 65° N and from 170° W to 100° W.

midlatitudes. The cold air associated with anticyclones may become wedged or dammed against mountain ranges, leading to freezing rain or ice if warm moist air is circulated over the dammed, cold air. Thus, while perhaps not as dramatic as cyclones, anticyclones have their own unique and interesting features and impacts.

See also: **Middle Atmosphere:** Stratospheric Sudden Warmings. **Mountain Meteorology:** Cold Air Damming. **Synoptic Meteorology:** Weather Maps.

Further Reading

- Bluestein, H.B., 1992. Synoptic-Dynamic Meteorology in Midlatitudes. Oxford University Press, New York.
- Holton, J.R., Hakim, G.J., 2013. An Introduction to Dynamic Meteorology. Academic Press, New York.
- Palmen, E., Newton, C.W., 1969. Atmospheric Circulation Systems: Their Structure and Physical Interpretation. Academic Press, New York.

Forecasting

D Mansfield, National Meteorological Center, Bracknell, UK

© 2003 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 5, pp 2220–2230, © 2003, Elsevier Ltd.

Introduction

This article will consider the role of the human in forecasting for middle or high latitudes where the weather is dominated by synoptic-scale disturbances.

The role of the human in forecasting the weather, be it for the next few hours or for up to a week ahead, has changed enormously over the last 30 years. Long gone are the days when the forecaster relied on empirical rules and some very basic dynamics to predict the next day's sea-level pressure pattern and hence the weather. For some time this part of the task has been carried out for the forecaster, and, with increasing accuracy, by numerical weather prediction (NWP) models. For forecasts up to 36 h ahead serious errors in the predicted surface pressure and upper wind patterns are rare. One important advantage that human forecasters still have over the numerical model is their ability to interpret cloud or moisture patterns from satellites in terms of weather systems. Although the forecaster cannot normally expect to 'beat' the computer at predicting the pressure pattern over a large area, there is still scope for local adjustments based on an assessment of the accuracy of the initial conditions upon which the numerical forecast is based. NWP models are less accurate when predicting the actual weather elements such as precipitation amount and type, cloud amounts, fog, etc. The forecaster's role has increasingly become that of interpreting and refining raw NWP products, especially in terms of weather elements. However, there are still a few occasions when numerical guidance can go seriously wrong and the forecaster must continually monitor the NWP output for signs of this and be prepared to modify the whole forecast if necessary.

There are many different roles required of forecasters, depending on who their customers are. They may be providing central guidance on the synoptic-scale evolution to other (local) forecasters, or providing forecasts for the general public (most often via the media), to the military, to civil aviation or to other commercial customers and on a variety of time scales. One common aspect of all these roles is timeliness. A weather forecast, particularly a short-range forecast, is a very perishable commodity, and even forecasts for several days ahead may be subject to adjustment after 12 h, when the next set of NWP products are produced. It is normally 2–3 h after data time before NWP products become available to the forecaster and there is often a further chain of processing and briefing before the forecast reaches the customer.

As NWP models continue to improve and mesoscale and single-site models enable more accurate prediction of local weather, the ability of the forecaster to add value to the numerical guidance will continue to decrease, at least on average. The forecaster's role for most of the time will become that of interpretation. However, there will still be rare occasions when the NWP models produce large errors. Although not making much impact on skill scores such as rms errors of mean

sea-level pressure (MSLP), these are likely to be associated with rapidly developing systems that may produce life-threatening extreme weather events, and it is in recognizing these occasions that forecasters will continue to prove their worth.

Central Guidance

Most national weather services of developed countries have a central guidance center whose role is to provide an interpretation and assessment of the latest NWP products and to issue warnings of any expected severe weather likely to be a threat to life or property. In many cases this is extended to guidance on the actual weather details expected so as to ensure that all forecasts issued by different offices of the national weather service are consistent. Because of the time taken to disseminate the guidance, this is usually for the period from about 6 h ahead to perhaps 5–7 days ahead.

Analysis

The first step is for the forecaster to analyze the current situation. Up until a few years ago this would normally have involved hand drawing of surface-pressure maps and upper air height contours. Nowadays computer-drawn 'first guess' charts (usually a 3 or 6 h forecast from a previous model integration) are nearly always close enough to reality for the forecaster to use these along with surface and upper air observations and satellite and radar imagery to recognize the dominant weather systems and processes at the current time. Most NWP models are global in extent, but for short-period forecasts, the forecaster will normally restrict his or her interest to the forecast region and an area upstream, though this may be fairly large (typically the whole of the North Atlantic for European forecasters and most of the North Pacific for those in the United States or Canada). Using conceptual models of these processes and systems, the forecaster then compares satellite and radar imagery and surface and upper air observations with the computer-drawn charts in order to assess the accuracy of the NWP first-guess fields. If there is a discrepancy between the NWP field and the observations, the forecaster will be alerted to a possible problem with the subsequent forecast. In most cases, if the difference is small, it will be corrected by the numerical analysis scheme and the new analysis $t + 0$ h field of the next model run will be a closer fit to the observations. However, if the difference between the observations and the background field is large, the observations may be rejected by the quality control procedures. In some centers such as the UK Met Office, it is possible for the forecaster to intervene to assist the quality control scheme make the correct decisions, to add weight to crucial observations in the assimilation scheme, and even to invent 'bogus' data where satellite or radar imagery suggest the

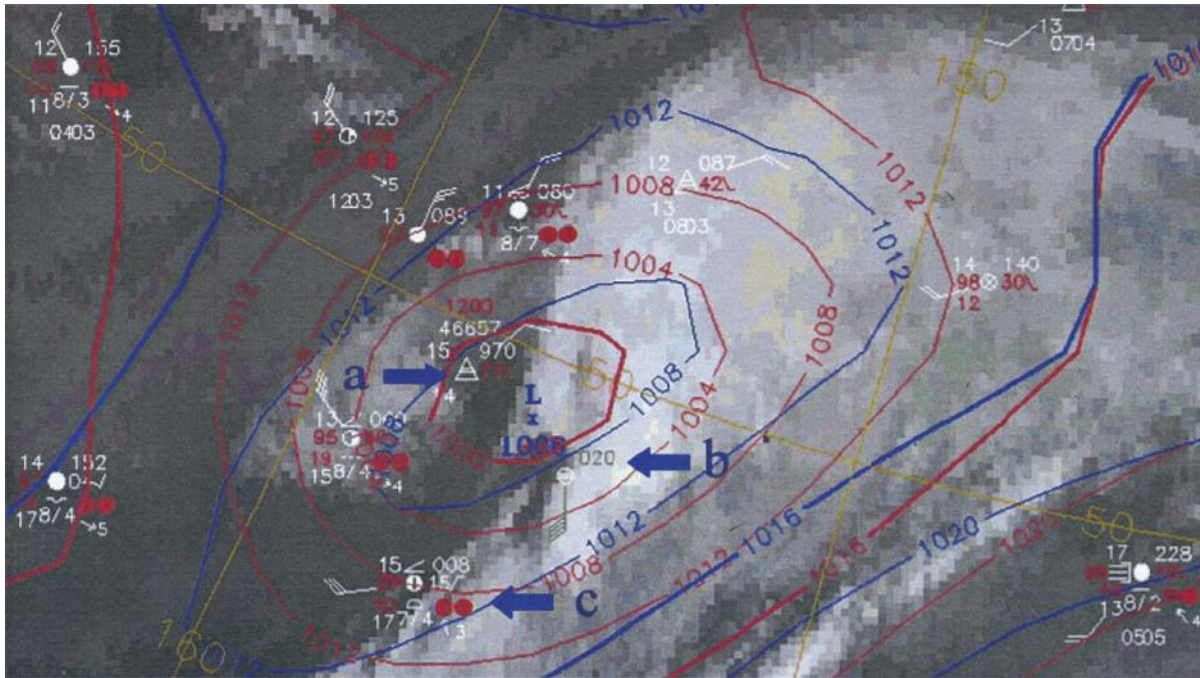


Figure 1 An example of a VDU display used to check NWP background field with observations and satellite imagery, in this case $t + 6$ h background field (blue contours) and $t + 0$ analysis (red) MSLP compared with surface observations and infrared satellite image over part of the central North Pacific. Important observations are arrowed. An observation from a drifting buoy (a) has a pressure of 997.0 hPa, nearly 13 hPa lower than the $t + 6$ background. The satellite image supports the idea of a more rapidly deepening low than suggested by the NWP field, but the ship observation (c) to the south looks unrealistically low, though the 30 knot (15 m s^{-1}) northwesterly wind supports the idea of a deeper low to the north. A pressure of 1008.0 hPa looks more likely than 1000.8; mistakes in coding frequently lead to this sort of error. To help the assimilation scheme a bogus observation (b) of 1002 hPa and a 40 knot (21 m s^{-1}) southerly wind has been inserted to the southeast of where the low center was estimated to be. As a result, the $t + 0$ NWP background pressure is a much better fit to the real and bogus observations and more in line with the forecaster's interpretation of the satellite imagery. Reproduced by permission of the Met Office.

NWP background is in error, but where there are no real observations in the area to correct this (see [Figure 1](#) for an example).

Diagnostics

Actual weather elements such as low cloud, fog, surface temperature, and some details of the precipitation, particularly showers, are less well forecast by NWP models than the basic pressure patterns. In order to be able to add value to the raw forecast in these areas the forecaster has to understand the dynamics of the large-scale environment in which the smaller-scale processes are embedded and the way in which the different scale processes interact. Forecasters have access to many diagnostic fields from the NWP models to help them in this task.

As well as surface-pressure maps and upper air contour charts, most commonly used are model relative humidity (as a proxy for cloud) for comparison with satellite imagery, and vorticity, vorticity advection ([Figure 2\(a\)](#)), thickness (a measure of the mean temperature between two levels in the atmosphere), and thickness advection, at various heights, to monitor the two most important aspects of large-scale flow. Wind strength ([Figure 2\(b\)](#)) or wind vectors or barbs are also useful in delineating model jet cores that can be compared with

satellite imagery. Jet cores are often apparent on infrared images as a linear contrast between bright areas of cold high cloud on the warm side of the jet and dry areas of subsided air on the cold side. These features are even more apparent in water vapor images. In differentiating between moist and dry regions of the middle troposphere, these images give information about the atmosphere in cloud-free regions and, through the associated changes in humidity, can indicate ascending and descending-motion associated with developing weather systems before this becomes apparent in other imagery.

Since potential vorticity (PV) has become available as a diagnostic from most NWP models, the strong relationship between water vapor imagery and the distribution of potential vorticity is becoming increasingly used as a tool to check the initial conditions and early stages of a forecast. Dark (dry) areas in the image are associated with high PV in the upper troposphere, while in dynamically active regions, particularly when cyclogenesis is taking place, contours of PV on a quasi-horizontal surface such as a pressure surface or isentropic surface curve anticyclonically over areas of ascent and developing cloud. One problem is picking a suitable surface on which to display the PV, as the area of interest is usually just below the tropopause and the associated pressure and potential temperature will vary with the season, current weather situation, and geographical location. Using the fact that PV increases sharply across the tropopause from around 1×10^{-6}

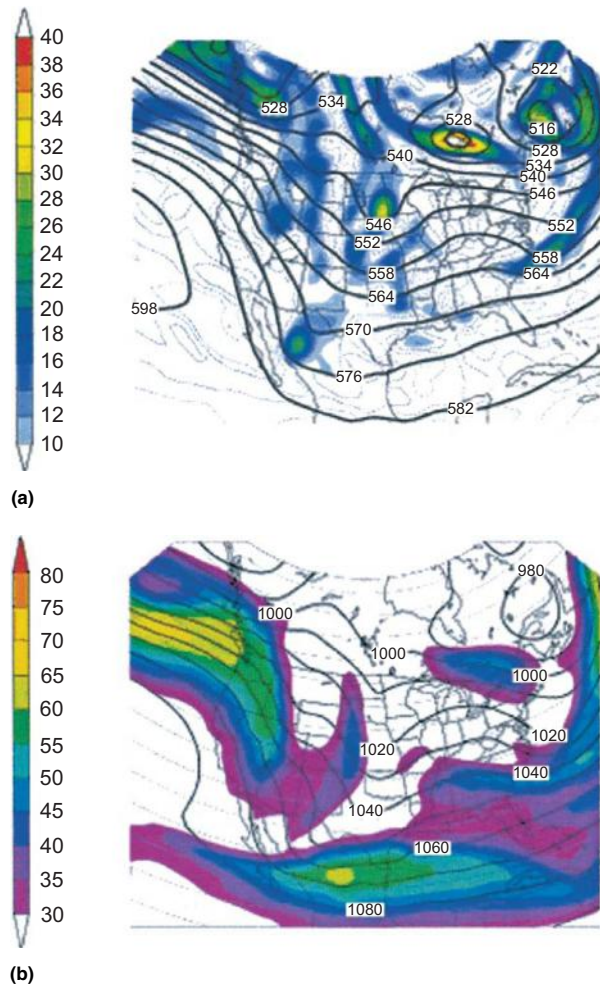


Figure 2 Examples of model diagnostics used to interpret and understand NWP output. (a) 500 hPa geopotential height contours and absolute vorticity (colors). The forecaster can quickly see where areas of large vorticity advection contribute to ascending and descending motion in the model. (b) 250 hPa geopotential height contours and wind strength (colors). Regions of maximum wind strength (jets) can be compared with indications of jet axes from satellite imagery and important dynamical regions at the entrances and exits to jets and their relative strengths quickly assessed. Images courtesy of the NOAA-CIRES Climate Diagnostic Center, Boulder, CO, USA from their website at <http://www.cdc.noaa.gov/>.

$\text{m}^2 \text{s}^{-1} \text{K kg}^{-1}$ (1 PV unit; PVU) to around 6 PVU in the lower stratosphere the problem can be avoided by plotting the height of a PV surface (usually 2 or 1.5 PVU) that is always close to the tropopause. An example is shown in Figure 3.

On some occasions the relationship between the PV and water vapor imagery can be confused or misleading, particularly in the very early stages of cyclone development. If the development is initially taking place in the low to middle troposphere, the image may show the pattern of ascent and descent before it influences the PV distribution at higher levels. However, another diagnostic, so-called ‘pseudo-imagery’, is becoming available to the forecaster to cope with these problems. The radiance at the top of the atmosphere in the water vapor channel is computed from the numerical model values of

temperatures and humidity and can be displayed either as an image or as contours of brightness for comparison with the real imagery. An example is shown in Figure 4. The comparison of the model PV and the real image (Figure 4(a)) shows a possible problem south-west of Portugal where the numerical model high PV is associated with low radiance in the water vapor image, but the pseudo-imagery (Figure 4(b)) also shows low radiance in this area, confirming that this is due to convective cloud penetrating into the otherwise dry upper troposphere. However, near the center of the image, to the south-west of the Azores, a small PV maximum also corresponds to a region of low radiance in the real image, but in this case the dry, dark area in the pseudo-image extends south to coincide with the PV maximum, suggesting a small error in the model in this area.

The pseudo-image gives a much closer comparison with the actual image, but still has to be used with caution and is best used in conjunction with the PV comparison. Apparent discrepancies between the two images may be due only to poor model simulation of the relative humidity, which may be unimportant in the subsequent developments. On the other hand, a close fit can also be misleading, as many NWP models assimilate water vapor radiances. Any adjustment to the NWP radiance is mostly through the humidity, so it is possible that this may mask an underlying problem with the dynamics.

Interpretation

Most global models will distinguish between and display different types and phases of precipitation, i.e., steady rain or snow from large-scale ascent and showers due to local convection. However, it is still necessary to refine the NWP output in these areas. For example, in most models, showers cannot be advected from their source region and therefore stop abruptly and unrealistically at windward coasts in winter as the air transfers from over the warm sea to over cold land. The extent to which showers penetrate inland will depend not only on local orography, not fully resolved by the numerical model, but also on the large-scale vertical motion. An important aspect of the precipitation in winter is the boundary between, rain, snow, or freezing rain (ice storms). There is a very fine balance between these different types of precipitation when the low level temperature is close to 0°C , depending on the initial vertical profiles of temperature and humidity, and the balance between thermal advection and the cooling of the air by evaporating or melting precipitation, which in turn will depend on the precipitation rate. Any of these physical processes may be inadequately parameterized, but the correct forecast of the type of precipitation is crucial in issuing timely warnings of severe weather. The forecaster must use his or her experience and knowledge of any weaknesses in the NWP models to try to add value to the forecast.

Although NWP models indicate the possibility of strong winds and heavy showers, the forecaster still has to distinguish those occasions with the potential for severe weather, such as violent thunderstorms, tornadoes, hail, and downslope winds, which are not directly forecast by numerical models, and issue advanced warnings. In regions where such severe weather is common, short-period detailed forecasts are issued locally using specialized models, radar, and other forecast aids.

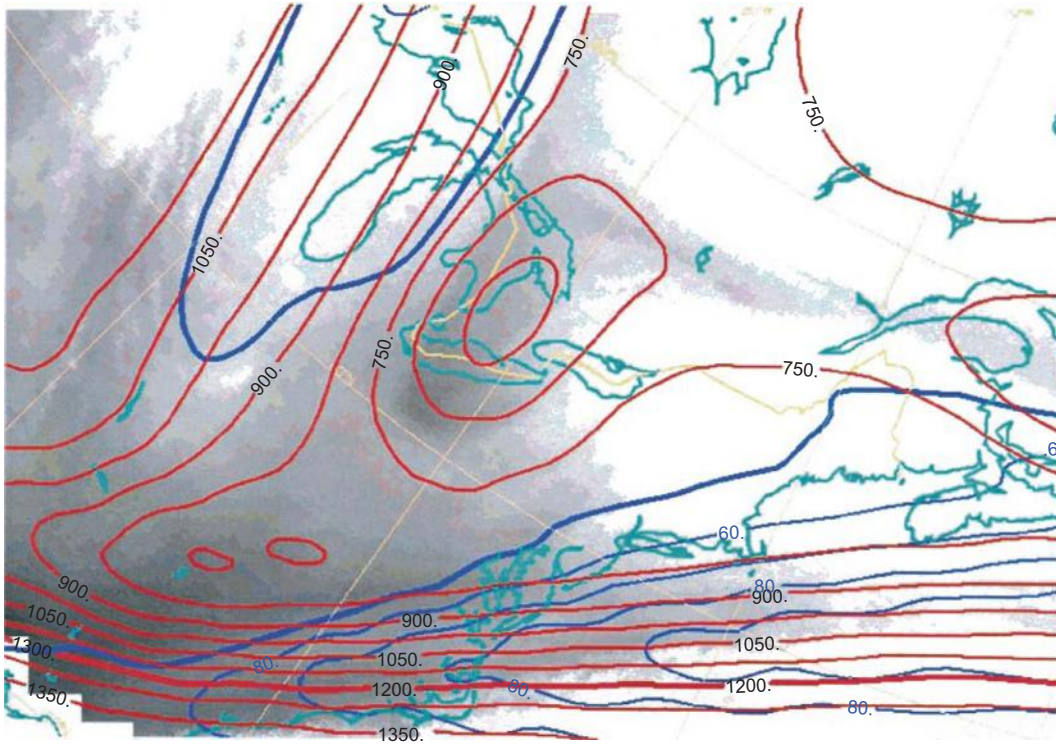


Figure 3 An example of a water vapor image for part of North America overlaid with contours of the height of the PV = 2 PVU surface (red) and wind strength on the surface (blue). The strong gradient of PV = 2 height at the foot of the figure and maximum wind speed corresponds to the edge of the bright area showing that the model jet is correctly positioned while the minimum in PV = 2 height just to the rear (north) of the dark area over the south-east of the Great Lakes shows that the cold trough (area of high upper level PV) in the model is also correctly positioned – reassuring for the forecaster as in the subsequent forecast these two features interact to form an intense depression over the Atlantic Ocean. Image and model fields from the French ARPEGE model supplied by Météo-France Forecast Laboratory.

Correction

When comparing model fields over the first few hours of a forecast with observations and satellite and radar imagery, the forecaster often finds small discrepancies such as not enough or too much rain in the NWP output, fronts or rainbands too fast or slow, or depressions not quite deep enough and hence winds not strong enough. The forecaster can then apply appropriate adjustments to the NWP forecast, assuming that these errors persist through the forecast period or decay or grow in a simple manner. This technique is effective up to 24 h or perhaps 36 h ahead. The adjustments are usually made in terms of written or verbal guidance or by the adjustment of single time forecast pressure charts, but techniques are becoming available to adjust electronically the NWP fields in a dynamically consistent manner at all time frames before the output is disseminated to other users.

Very rarely the NWP initial conditions may be so seriously in error that the forecaster has to disregard the model guidance and use his or her own synoptic and dynamic knowledge to make a new forecast. **Figure 5** shows a satellite image for the North Atlantic on 23 December 1997. Conceptual models of cyclogenesis suggest that the cloud area (a) indicates a rapidly developing depression, whereas the NWP field showed only a very weak circulation (b). Although not a bad fit to the available surface observations, the analysis and subsequent

forecast were considered completely inadequate, a theory confirmed by the development of the cloud area in the next 2–3 h. The forecaster overrode the NWP model to forecast a deep depression just west of Ireland 24 h later (**Figure 6**). The manual forecast depression was 16 hPa deeper than the unmodified forecast and only 3 hPa higher than the actual depth, though marginally displaced, and enabled the forecaster to give timely warning of damaging winds over parts of Ireland and the UK.

Medium Range

Even after 24 h, different models sometimes show significant differences in detail of weather patterns. An example is shown in **Figure 7**. Although the forecast pressure pattern (**Figure 7(a)**) is hardly different in the two models, apart from a deeper trough over Tennessee, there are large differences in the predicted rainfall over the west coast of the United States, over Mexico, and particularly over the south-eastern states (**Figure 7(b)**). The forecaster has to use his knowledge of the strengths and weaknesses of the two models to help decide which is more likely to be correct. Beyond about 36–48 h ahead, errors in the initial conditions or those due to imperfections in the numerical models have grown such that forecasts from different initial times or by different forecast centers normally

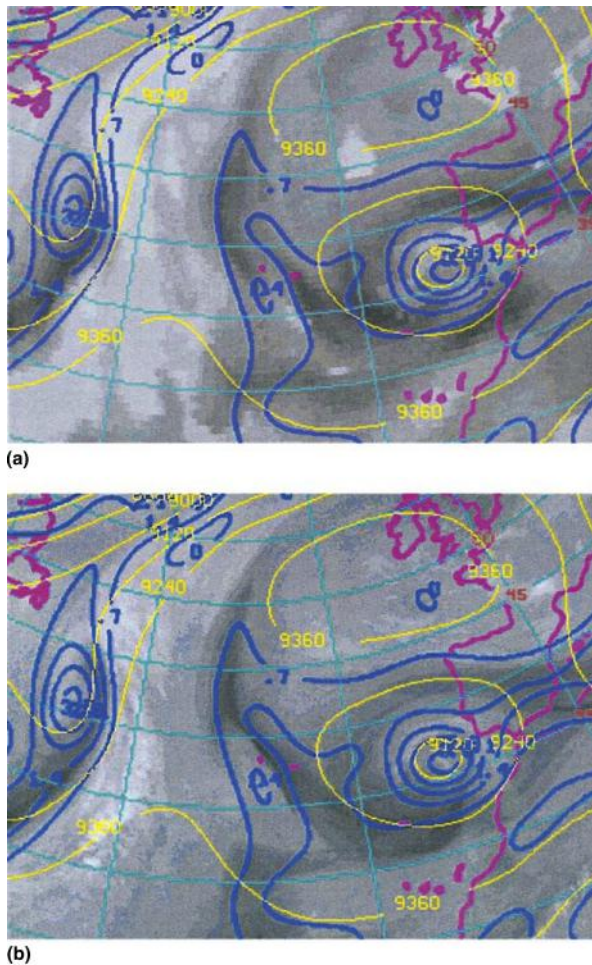


Figure 4 (a) An example of a water vapor image overlain with contours of PV (blue, interval 0.7 PVU) and geopotential height (yellow, interval 120 m) on the 300 hPa surface. (b) Corresponding pseudo water vapor image with the same fields superposed. Image and numerical model fields supplied by Servicio de Técnicas de Análisis y Predicción (STAP), Instituto Nacional de Meteorología (INM) Madrid, Spain.

begin to differ. It is no longer possible to predict with confidence details or exact timings of weather events, though the general evolution and type of weather systems likely to be experienced can usually be predicted out to 4 or 5 days, and sometimes beyond this.

At this range it is no longer possible to extrapolate errors in the initial conditions, nor is it possible to beat the models at a dynamical forecast, but there is still a role for the human forecaster. Most large forecast centers exchange raw NWP output with one or two other centers for use as backup, so that the forecaster can usually compare output from two to three or more different model integrations for their region of interest in the same format on screen. In addition to this, the output from many global models is available via the Internet, so that the forecaster may have available as many as 10 different models to choose from. At the same time, several centers around the world are addressing the problem of the uncertainty in the initial conditions and the subsequent error growth by running ensembles of forecast with slightly different initial conditions

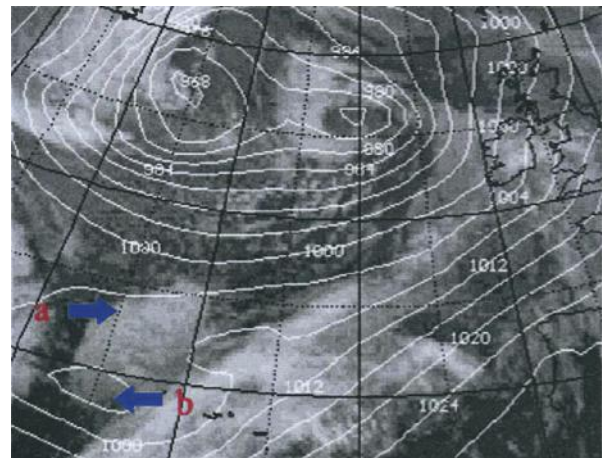


Figure 5 Infrared satellite image over part of the North Atlantic for 1200 UTC 23 December 1997 overlaid with first guess NWP $t + 6$ h MSLP field (contours every 4 hPa). The area of cloud (a) indicates a rapidly developing depression, whereas the model has only a weak depression (b). Reproduced by permission of the Met Office.

in an attempt to cover all the possible evolutions of the real atmosphere.

In spite of the inherent uncertainty, many customers still require a categorical forecast. The mean of the ensemble of different forecasts is on average more skillful, at least in rms terms, than an individual forecast because it averages out the less predictable smaller-scale features, but by its nature is very bland and does not give a good indication of the actual weather. The forecaster must use his or her judgment and synoptic experience to select most likely evolution or 'blend' elements from different models, the so called deterministic forecast.

However, it makes more sense to couch forecasts at this range in terms of probabilities. Even with the deterministic forecast, this is done to some extent by the confidence placed in the forecast. It is important to try and convey this in public service forecasts. If a large, slow-moving anticyclone covers the region, the forecaster may be almost 100% confident of dry weather, but in a more changeable spell of weather, even though the most probable forecast is for a transient ridge of high pressure to bring a dry day, possible errors in timing could mean that there is still a 50% chance of rain. Having decided on the most likely evolution, it is important to convey the degree of uncertainty associated with this, particularly when issuing guidance to other forecasters so that they can couch the forecast for their customers in suitable terms.

For forecasts of point probability, normally expressed in terms of the likelihood of a threshold being exceeded, such as wind speed of gale force or more, the ensemble forecast can give a direct estimate. However, ensembles based on a single model, in spite of perturbations to the physics within the model as well as to the initial conditions, still do not cover the whole spread of possible outcomes. An ensemble of different models such as accessed by the forecaster via the Internet often has greater spread (though it is not uncommon for all models to agree but still differ from reality). There is also still the problem that the models may not accurately predict the

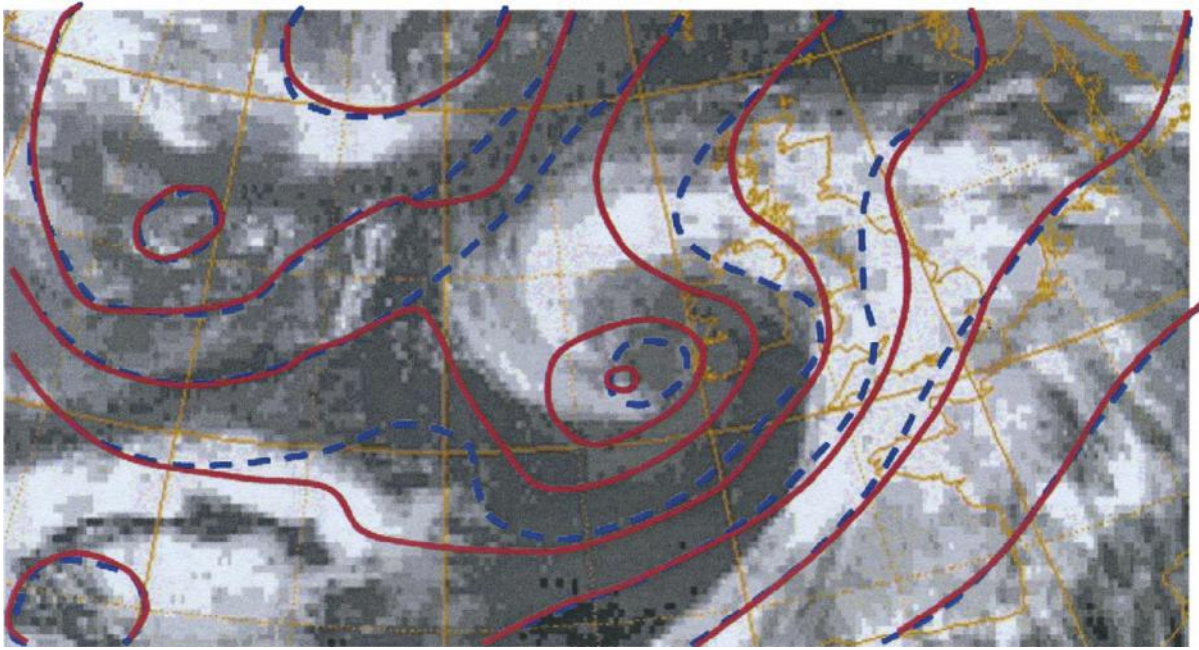


Figure 6 Infrared satellite image for area over and west of UK for 1200 UTC 24 December 1997 overlaid with 24 h numerical forecast MSLP (blue, dashed contours) and 24 h forecast modified by the forecaster (red contours), contour interval 8 hPa in each case. The satellite image clearly suggests a deep depression and surface observations confirmed that the modified forecast, of a depression of 976 hPa was much more accurate than the NWP version (990 hPa). Reproduced by permission of the Met Office.

weather elements, in spite of having the correct pressure pattern. There is therefore still scope for the forecaster to add value to probability forecasts of individual weather elements, though time will necessarily limit this to a few crucial parameters at only a few geographical locations. A logical conclusion is then for these corrections to the ensemble forecast to be applied to an appropriate degree at surrounding locations.

Specialist Forecasts

Aviation

Forecasts for aviation again rely very heavily on NWP guidance. They can be divided roughly into three types of forecast.

- *High-level significant weather forecasts.* These are forecasts of conditions near the tropopause where jet airliners fly.
- *Low-level significant weather forecasts.* These forecast conditions up to around 10 000 ft (3048 m), used by 'general' aviation, e.g., private pilots, small local airlines, military aircraft, couriers, etc.,
- *Terminal airfield forecasts (TAFs).* These are forecasts of surface wind and weather elements at specific airfields.

Upper-level significant weather charts are produced centrally by centers designated by the international Civil Aviation Authority to display jet streams, the level of the tropopause, and any high-level aviation hazards, in an agreed format, and are usually valid for fixed times 18–24 h ahead and are updated every 6 h. An example is shown in [Figure 8](#). At normal flight levels, the weather does not often present a serious hazard to modern airliners. The main concerns for airlines are the temperature

and wind speed, which will affect fuel consumption. These are generally forecast very well by NWP models and most companies take direct NWP forecasts of winds in digital form for use in flight planning. Only very rarely will a forecaster see the need to correct the NWP winds. The main hazards at these levels are thunderstorms and clear-air turbulence, though it is also the responsibility of weather services to track and warn of volcanic ash.

Areas of thunderstorm activity are reasonably well forecast by the NWP models, but it is still necessary for the forecaster to check and sometimes correct details such as cloud top height. Forecasters must also use their experience to decide whether thunderstorms are likely to be isolated, in which case they are not considered a hazard, or embedded in other cloud layers so that they cannot be easily detected, or difficult to avoid due to their spacing or due to being in a line. Occasionally the models may misplace or miss areas of thunderstorms altogether, especially in the tropics, where a series of recent satellite images may be a better guide to activity over the next 24 h.

Clear-air turbulence occurs in areas of strong wind shear, normally around jet streams. It is not associated with cloud and therefore cannot be detected in advance, and can be sufficiently violent to cause injury or even death to passengers or aircrew if not restrained by seat belts. However, it is a very intermittent phenomenon and impossible to forecast precisely at present. NWP models provide an indication of regions of strong vertical or horizontal shear where turbulence is likely to occur, but this is a necessary rather than a sufficient condition. Forecasters can add value by using conceptual models of the type of airflow most likely to lead to actual severe turbulence to refine the forecast. Areas where a risk of moderate or severe turbulence is

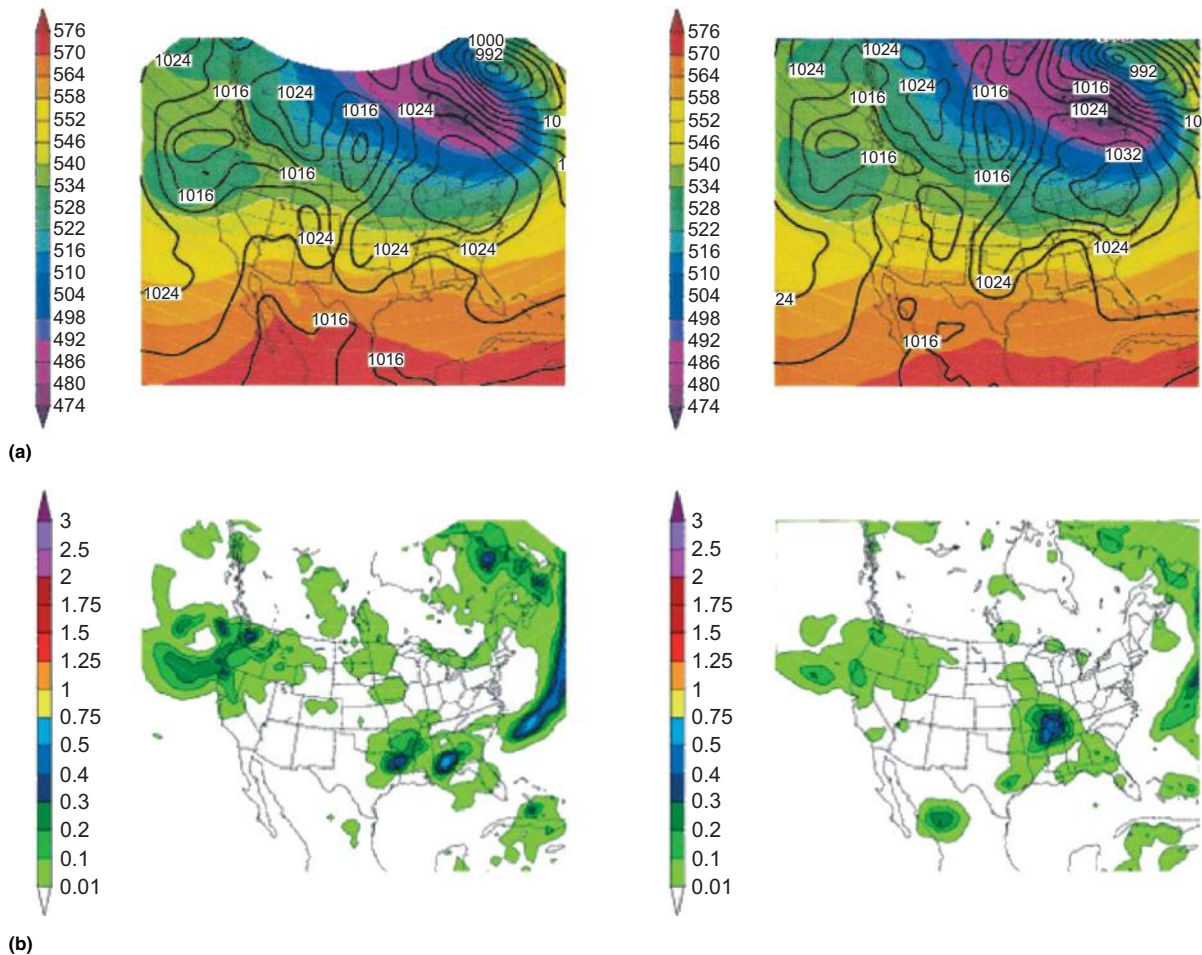


Figure 7 (a) Comparison of 24 h forecasts from two NCEP models, valid 0600 UTC 22 February 2001. Solid contours are of sea-level pressure, every 4 hPa. The colours represent the thickness layer between the 1000 and 500 hPa, a measure of the mean temperature of the lower troposphere. (b) Comparison of the forecast rainfall accumulations for the same forecasts as in (a) for the 12 h up to 0600 UTC. Images provided by the NOAA-CIRES Climate Diagnostic Center, Boulder, CO, USA from their website at <http://www.cdc.noaa.gov/>.

expected are marked on the significant weather charts along with the range of heights over which the hazard is expected to extend. However, most of the time, aircraft flying through these areas experience no serious problems. For this reason pilots encountering severe turbulence make an immediate report, which is relayed to the forecaster, who then issues a more definite forecast in the form of a SIGMET. This is a text forecast, which is disseminated with maximum priority to all aviation users, so that any following aircraft may take avoiding action or at least ensure seat belts are in use. If a forecaster is sufficiently convinced of a high risk of severe turbulence he or she may issue a SIGMET without any actual aircraft reports. SIGMETs are also issued when there is very high confidence in the forecast of other hazards such as embedded thunderstorms, line squalls, and severe low-level turbulence, or when these events are observed.

Low-level significant weather forecasts are also mostly produced centrally, but nationally rather than regionally, and are normally valid for shorter periods, typically up to 9–12 h ahead, though planning forecasts are produced from some centers for up to 36 h ahead. As well as the thunderstorms and

turbulence (though in this case normally low-level turbulence due to strongwinds flowing over the Earth's surface), low cloud (especially where it covers hills) and icing are the main forecast parameters. Although NWP provides a framework of the positions of frontal zones, areas of convection, etc., parameters such as the amounts and base of low cloud, visibility, and the likelihood of icing are not well forecast numerically, and the forecaster relies more on experience and extrapolation of present conditions, subject to any changes in the large-scale conditions indicated by the NWP models.

The third type of forecast, the TAF, is normally valid for 9 h ahead, although at major airports, forecasts of up to 24 h ahead are provided to give airlines an idea of likely risk of long-haul flights being diverted. Forecast parameters are wind speed and direction, cloud amount and height of base and visibility, plus any weather conditions that may be a hazard, such as thunderstorms, hail, snow, freezing rain, and mist or fog, though the latter are also implied by the visibility. These forecasts have traditionally been produced locally on site by forecasters who have a great deal of experience of the peculiarities of the particular airfield, and are based largely on extrapolation of

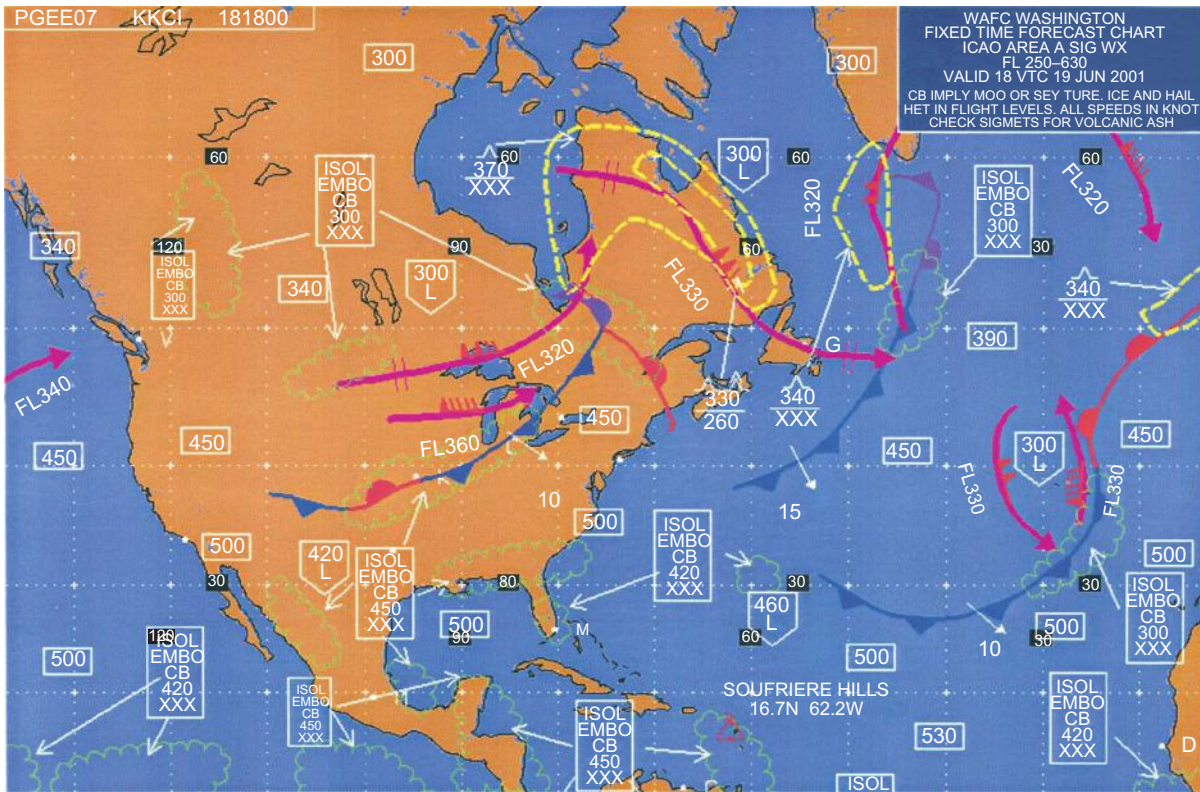


Figure 8 Part of a high-level aviation significant weather chart, produced by World Area Forecast Center Washington, valid 1800 UTC 19 June 2001. Surface fronts are shown by conventional symbols, and jet streams (wind speed >80 knot) by the magenta-colored arrows, with the maximum wind strengths in red (each barb represents 10 knot, and the solid triangles 50 knot). Yellow dashed lines outline areas of forecast moderate or severe clear air turbulence. The height range (in hundreds of feet) over which the turbulence may occur is indicated by the associated text. The green scalloped areas denote areas of significant thunderstorm activity, with details of flight levels affected in the associated boxes. Provided by the NOAA-CIRES Climate Diagnostic Center, Boulder, CO, USA from their website at <http://www.cdc.noaa.gov/>. ($1 \text{ knot} = 5.14444 \times 10^{-1} \text{ m s}^{-1}$.)

local or upstream conditions after allowance for diurnal changes, and effects of local topography and an idea of the synoptic-scale development. This is still true in many cases, especially at military airfields, but improved local detail and better estimation of actual weather parameters from mesoscale numerical models has meant that it is becoming possible for these forecasts also to be produced centrally with a single forecaster responsible for the TAFs for a dozen or more airfields.

Marine Forecasts

Detailed forecasts of surface wind speed and direction, visibility, and sea state are normally provided for up to 24–36 h ahead. For the high seas, well away from land, numerical models provide a good estimate of all but visibility, though it may be necessary for the forecaster to make some adjustments in accordance with any central guidance on the perceived accuracy of the latest NWP forecast. Visibility is estimated, within broad limits, from knowledge of the source of the air mass, the air temperature relative to the sea surface temperature (will the air be cooled by the sea to form mist or fog?), and consideration of recent ship reports in the same air mass.

Coastal forecasts rely slightly more on local knowledge and interpretation of numerical output around complex coastlines, as numerical models are not always detailed enough to

represent these effects on the wind speed and direction or sea state, nor do they adequately represent small-scale changes in sea surface temperature likely to have an important impact on mist or fog formation.

Local Forecasts

The local forecaster has to consider the numerical model output and any corrections that may be made to this in the central guidance, then adjust the forecast for any small-scale effects due to local topography that may not be fully resolved by the numerical model. He or she will be concerned with the combination of these effects on weather parameters such as rainfall rate, rain–snow boundaries, cloud amount and sunshine, temperature, fog, and how they vary across the region. The forecaster will also have to understand the synoptic-scale dynamical processes taking place in order to make sensible adjustments to the numerical forecast.

Accurate forecasts of temperature at individual sites are crucial for forecasts of fog, frost, and snow in winter, and for showers, and in particular thunderstorms, in summer. The central guidance may indicate that these weather elements are to be expected over a broad area, but the local forecaster must estimate the risk at individual locations. There are several semi-empirical models available to forecasters to estimate diurnal

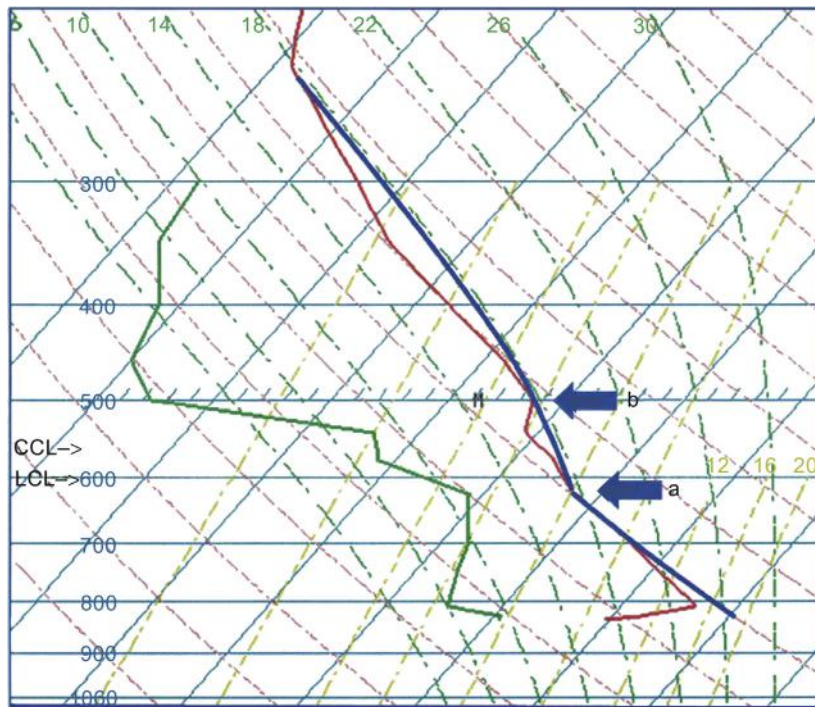


Figure 9 Radiosonde ascent from Denver, Colorado, at 1200 UTC 29 April 2001. The vertical scale is logarithmic in pressure and approximates to height. The diagonal pale blue lines are temperature in °C, every 10°. The solid red line shows the temperature profile and the green line the profile of humidity mixing ratio. The blue line represents the temperature curve of a parcel rising from the surface without any mixing with its environment and with an initial temperature of 20°C. The surface pressure is around 830 hPa because the station is over 5000 ft (1640 m) above sea level. Environmental curve provided by the NOAA-CIRES Climate Diagnostic Center, Boulder, CO, USA from their website at <http://www.cdc.noaa.gov/>. The parcel ascent curve was added by the author.

temperature changes that can be used to check and refine NWP output. The forecaster will use profiles of temperature and humidity from local radiosonde ascents, plotted on an aerological diagram, to diagnose the type and height of any cloud layers and assess the likely developments due to diurnal changes in temperature. The forecaster will then compare the basic profile and his or her analysis with those from the numerical model, and in particular assess the potential for shower or thunderstorm development. An example is shown in Figure 9. The solid red line shows the temperature profile and the green line the humidity mixing ratio, or dew point temperature. The temperature at the surface is colder than the air just above as this is a night time profile, but as the temperature at the surface rises during the day the temperature will become warmer than that just above the surface and the air will begin to rise. Unsaturated air will cool as it rises, following the red dashed lines. At the same time the humidity mixing ratio will remain constant and the dew point will therefore follow the yellow dashed lines. In the example the surface temperature must reach 20°C before the air can rise sufficiently to reach the condensation level and form convective (cumulus) clouds. The temperature will fall along the blue line until the air parcel becomes saturated at the point a, after which it will cool more slowly due to the release of latent heat as the water condenses in the cloud. In an unmixed parcel the temperature would now follow the dashed green lines. The continuing blue line therefore gives the maximum height to which a parcel could rise. In practice, except in the core of a large

cloud, mixing with drier air outside the cloud will lead to re-evaporation of some of the cloud water and the parcel will cool more quickly and air parcels would be unlikely to rise beyond the point b, as they would then be cooler than the surrounding air. However, the forecaster would have to consider if the air at this level is likely to be cooled by the large scale motion or if local conditions could lead to warmer or moister conditions near the surface. A small change may allow the air to rise all the way to the tropopause causing the formation of heavy showers or thunderstorms (unlikely though in the example, as the surrounding air is very dry and mixing would cool the parcel back towards the environment temperature). The forecaster would check his or her prediction of showers with the NWP output. If no showers were indicated, the forecaster would need to ask why. Does the model surface temperature reach the required value? Does the model represent the observed sounding adequately in its initial conditions? The formation or otherwise of even severe storms may hinge on the detail of a shallow layer of higher temperature, which may not be resolved by the numerical model.

See also: Aviation Meteorology: Aviation Weather Hazards. Data Assimilation and Predictability: Data Assimilation; Ensemble Prediction. Satellites and Satellite Remote Sensing: Precipitation. Synoptic Meteorology: Cyclogenesis; Weather Maps. Turbulence and Mixing: Overview.

Weather Maps

R Reynolds, University of Reading, Reading, UK

© 2003 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 5, pp 2230–2241, © 2003, Elsevier Ltd.

Introduction

Weather maps provide an irreplaceable, succinct summary of a wide variety of atmospheric phenomena and characteristics. They are a critically important working tool for the operational and research meteorologist, both of whom become very familiar with ways of illustrating features of interest that can be immediately appreciated by their peers. In a sense, weather satellite and weather radar images are maps of aspects of the weather. However, this section will not deal with remotely sensed fields, but with maps that portray weather-related features at the surface or in the upper air.

History

Surface Weather

Weather maps are nothing new. Synoptic meteorology is concerned with understanding relatively large-scale weather-producing disturbances like frontal depressions, tropical cyclones, and anticyclones – features that have a horizontal scale of many hundreds to a few thousand kilometers, and a lifetime counted in days rather than hours.

The earliest instrumental weather records began soon after the invention of the thermometer and barometer. Robert Hooke's manuscript from 1664, kept at the Royal Society in

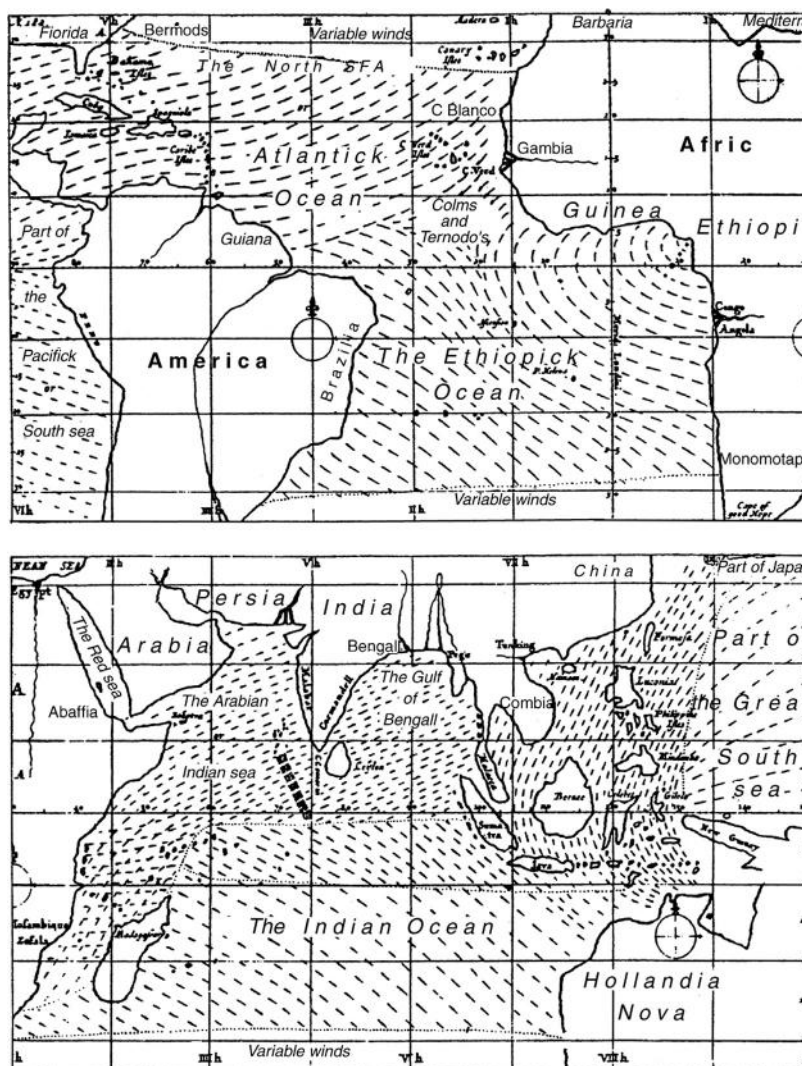


Figure 1 Halley's map of trade winds and monsoons. From Halley, E., 1686. An historical account of the trade winds, and monsoons, observable in the seas between and near the tropicks, with an attempt to assign the physical cause of the said winds. *Philosophical Transactions of the Royal Society* 16: 153–168.

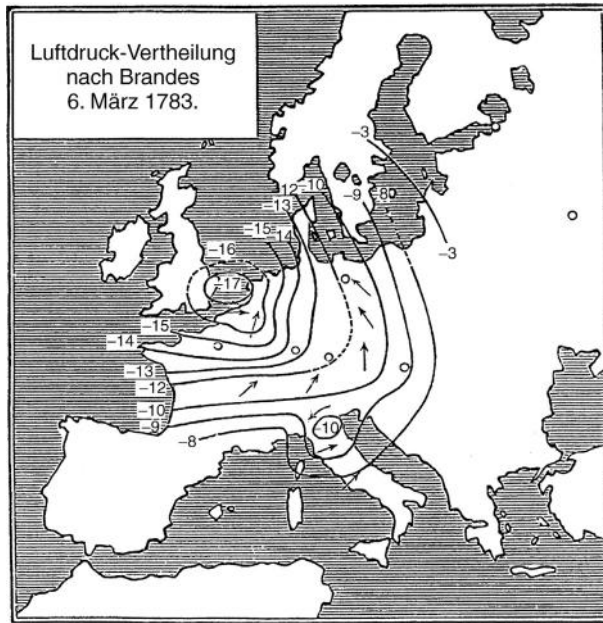


Figure 2 Brandes's weather map for 6 March 1783 (reconstruction). Reproduced with permission from Wilhelm Trabert, 1905. *Meteorologie und Klimatologie* Leipzig: Franz Deuticke.

London, summarized the keeping of such early observations. He did in fact appreciate the value of a network of weather observations, although it would be many decades before that became a reality. The late seventeenth and early eighteenth centuries saw the simultaneous use and gradual expansion of thermometry and barometry to many western European countries and America. This was also the age of sail, with many navies plying the Atlantic Ocean and other oceans.

It was the English natural philosopher Edmond Halley (1656–1742) who published the very first map to depict an

atmospheric variable, in the *Philosophical Transactions of the Royal Society* of 1686. This was strictly a climate map, to illustrate the mean surface winds over the tropical regions of the Atlantic and Indian Oceans (**Figure 1**).

The stimulus for producing the first weather map proper lies with a letter written to the *Annalen der Physik* in 1816 by Professor Heinrich W. Brandes of Breslau University (now Wrocław University). He suggested that a daily surface weather map could and should be produced for part of the period from 1781 to 1792: this was the era of the Meteorological Society of the Palatinate, based in Mannheim, Germany. The society fostered the science during this period when observations were taken three times a day, collected from 39 people across 18 mainly European countries.

It appears however, that this significant pioneering work was never summarized in map form – until Brandes's proposal that led him to plot 365 maps for 1783. These maps have never been found however; they were not a part of his article on 1783's weather in the *Annalen* of 1819. However, there are published reconstructions of his weather map for 6 March 1783 that illustrate both the simultaneous distribution of the departure of pressure from average and surface winds (**Figure 2**).

In 1831, the American James P. Espy (1785–1860) organized a committee from his base in the Franklin Institute in Philadelphia to collect weather data: in 1834 the Joint Committee on Meteorology was formed by the Franklin Institute and American Philosophical Society with Espy as chairman. The first American weather map based on widespread observations appeared in an 1837 issue of the *Journal of the Franklin Institute* (**Figure 3**).

The prime problem with all these early endeavors however was that the maps could be drawn only after the event. They did, though, offer meteorologists at least some insight into the scale of synoptic features, how pressure and wind appeared to be related, and how pressure features moved and evolved.

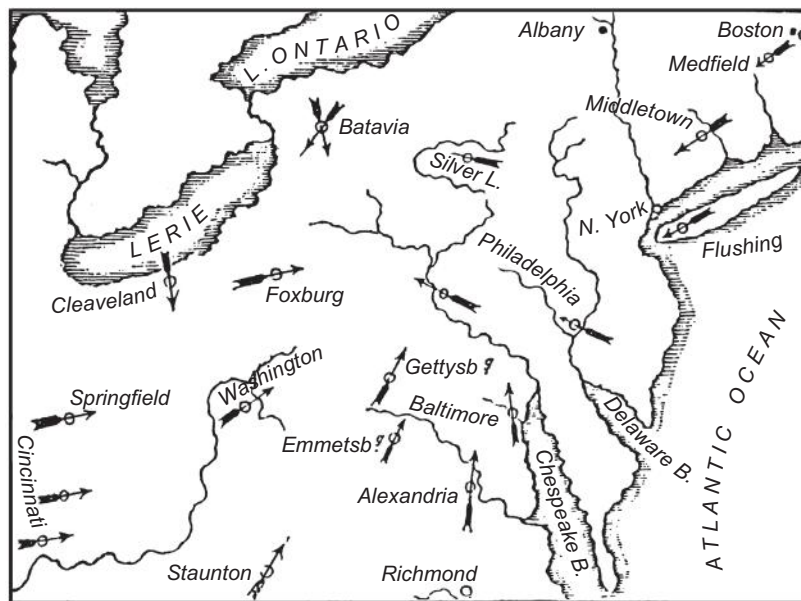


Figure 3 Espy's weather map for 20 June 1836. From Espy, J.P., 1841. *The Philosophy of Storms*. Boston: Charles C. Little and James Brown.

Register map indicating conditions of the atmosphere on the same day in several parts of Great Britain from observations collected at the great exhibition for 9 a.m. 24th Sept 1851 by the electric telegraph company

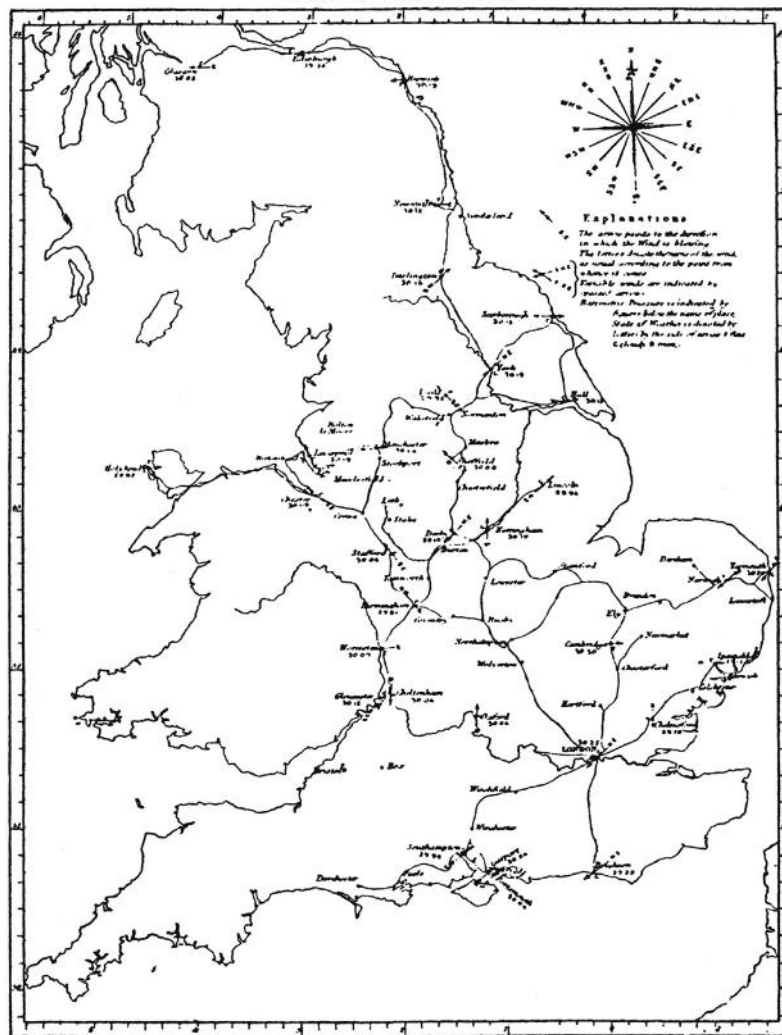


Figure 4 Daily weather map, Great Exhibition, London, 1851. Met Office Library.

The full utility of the weather map had to wait for a truly momentous event for the world as a whole, when Samuel Morse connected Washington and Baltimore by electric telegraph in 1844. This brilliant development paved the way for the ultimate 'live' mapping of weather observations and so to producing up-to-the-minute weather maps. By 1860, the Smithsonian Institution in Washington, DC had organized the electronic transmission and display of current weather reports from some 45 companies in the United States. The details were presented on a large map on public display in the institution.

In Europe, some few years earlier, the world's first same-day weather maps were being offered to the public gaze at the Great Exhibition of 1851. This was held in the Crystal Palace, situated at that time in Hyde Park in London. From 8 August to 11 October the public could purchase a copy of the day's weather map for the British Isles (**Figure 4**).

From about 1863, the Daily Weather Map Company of the Strand, London, offered monthly subscriptions to maps of

British and Irish weather (**Figure 5**). Across the Channel, Jean Joseph Le Verrier (1811–77), director of the Paris Observatory, founded a daily weather summary for mainly France in 1858. From September 1863, the bulletin of the Paris Observatory included a daily weather map.

On 1 April 1875, the London newspaper *The Times* initiated the presentation of a daily weather map to a much wider public. It published a chart of 8 a.m.'s weather for the previous day over the British Isles and parts of continental Europe that included plotted details of temperature, wind direction, 'weather', sea state, and analyzed isobars.

The first regularly published daily weather map in the United States appeared in the New York *Daily Graphic* on 9 May 1879, although this initiative lasted only a few years. By the final decade of that century however, many daily papers around the world had incorporated a weather map.

The gradual increase in the number of nations publishing 'government' weather maps led to the desirability of some sort of internationally accepted standard way to depict or symbolize

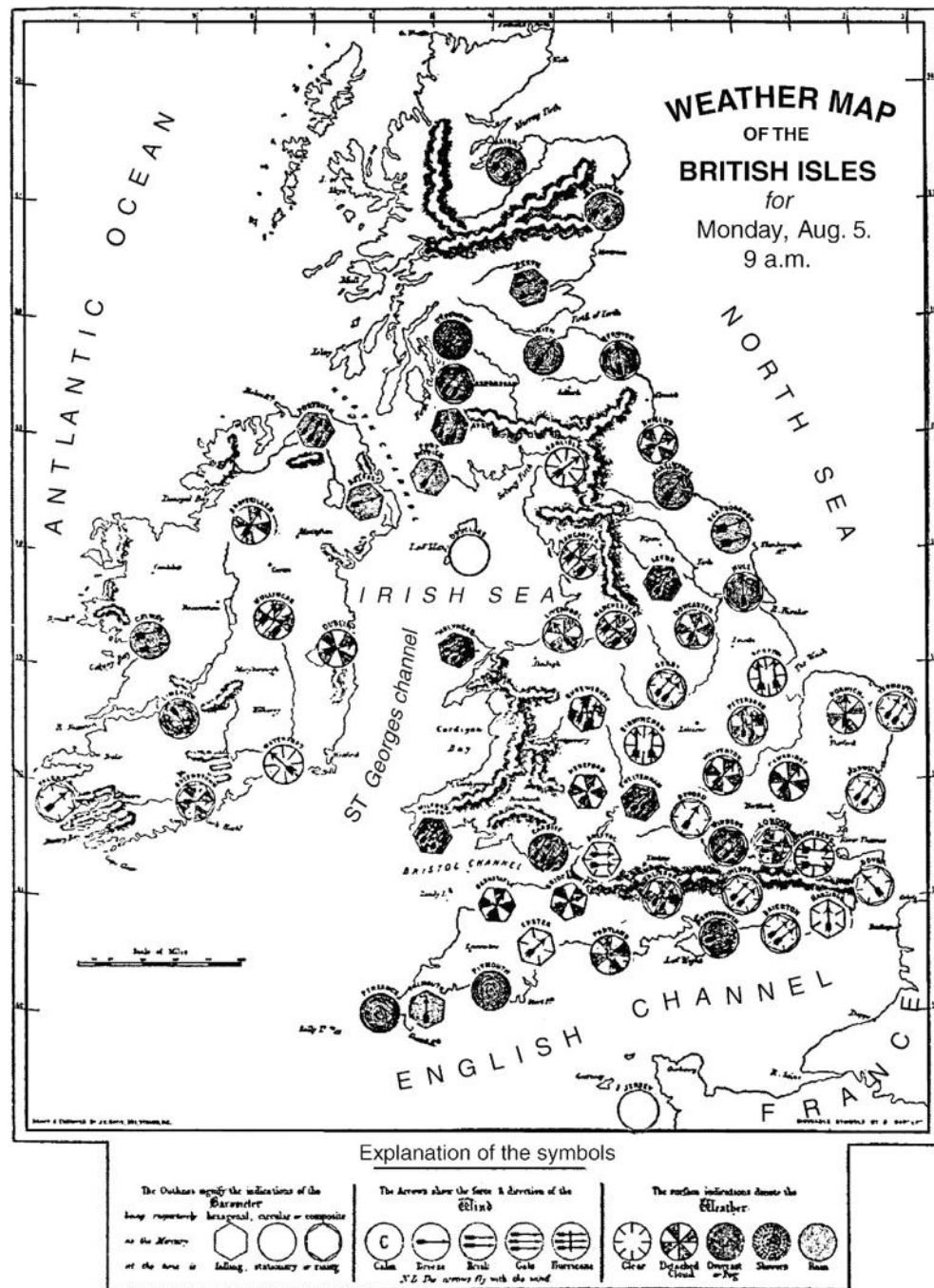


Figure 5 Daily Weather Map Company's map accompanying their promotional material (c. 1863).

the broad range of surface observations. Although such a standard was accepted at the International Meteorological Congress in Vienna in 1873, it was not to be globally accepted for some decades.

By 1891 at least, some 18 countries – mainly in Europe – were publishing government-service synoptic weather maps. As the surface weather network expanded, so did the area covered by weather maps. From 1 January 1914, the US Weather Bureau published surface weather maps for the entire Northern Hemisphere routinely. After World War I, many of the world's

weather services were producing their own analyses on this scale.

In postwar Norway a group of brilliant scientists led by Vilhelm Bjerknes (1862–1951), now known as the 'Bergen School', worked on the analysis of weather changes associated with the passage of traveling synoptic-scale disturbances in that region. The group refined some of the work from the previous century in which the notion of pulses of warm and cold air in the extratropics had been discussed. They developed the concept of warm and cold fronts, and the more

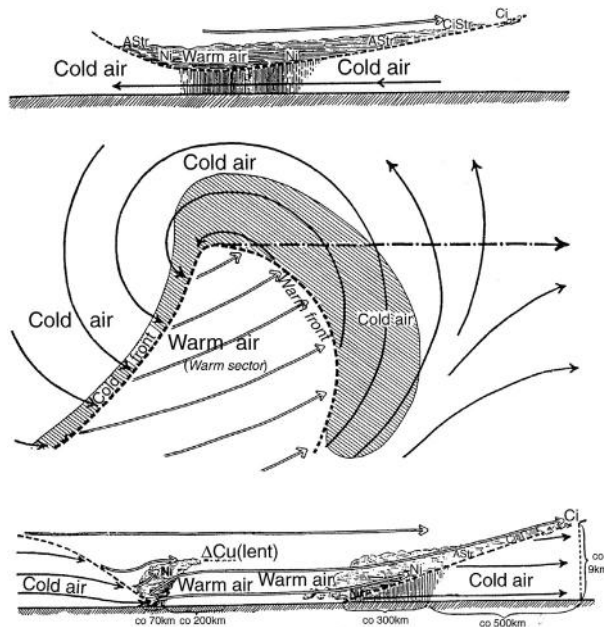


Figure 6 Vertical cross-sections and plan view of an open wave frontal depression. Reproduced with permission from Bjerknes, J., Solberg, H., 1922 Life cycle of cyclones and the polar front theory of atmospheric circulation. Geofisike Publikationer 3(1): 3–18.

general structure of the commonplace mobile frontal cyclones.

Jacob Bjerknes (1897–1975) and Halvor Solberg (1895–1974) published a highly significant ‘map’ of such a system in 1922 (Figure 6).

Another member of the school, Tor Bergeron (1891–1977), proposed the term ‘occlusion’ and the currently used symbols for the three types of surface fronts. In addition he suggested the used of slightly different symbols for upper fronts.

This exceptional work on the structure and evolution of extratropical frontal cyclones – and the location and representation of such fronts on surface weather maps – formed the basis for how all the world’s weather services located these critically important features. The methods have moved on, so that in the UK Met Office, for example, objective schemes are utilized for the automatic positioning of fronts – or at least for providing useful guidance to the analyst.

Upper-Air Weather

During the 1890s, Frank Bigelow of the then US Weather Bureau composed wind charts for three levels covering the contiguous United States. He did so by analyzing surface anemometer data along with lower and upper cloud drift observations supplied from 140 telegraphic stations. The cloud information was taken to represent the circulation at 3500 and 10 000 feet (1067 and 3048 m), so that early ‘upper air analyses’ were made available before the year 1900. It was as early as 1903 that Bigelow began the publication of daily barometric pressure charts for the two constant levels above and for that of mean sea level (Figure 7).

A supplement to these observations was provided by the use of kites. These sensed pressure, temperature, humidity, and wind speed up to a height of some 3 km but were replaced in

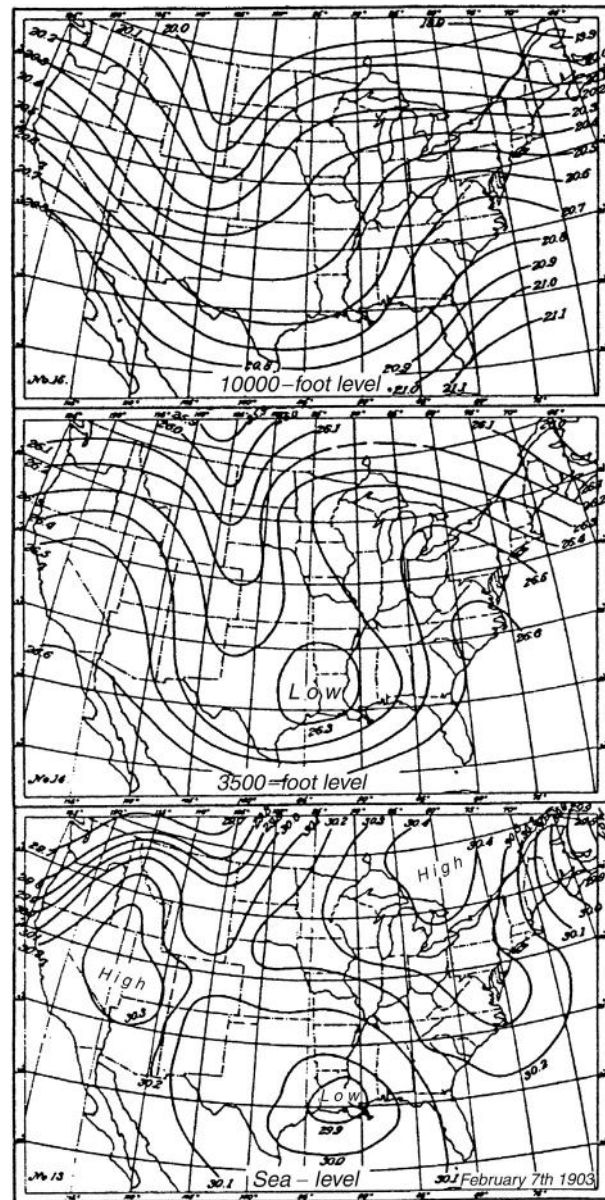


Figure 7 Barometric pressure analyses for mean sea level, 3500 feet (1067 m), and 10 000 feet (3038 m). Reproduced with permission from Bigelow, F.H., 1903 IV. The mechanism of countercurrents of different temperatures in cyclones and anticyclones. Monthly Weather Review 31: 26–29.

the 1930s by pilot balloons. Such balloons were tracked optically to enable mapping of wind direction and speed at a variety of heights – so long as cloud didn’t mask the observer’s view.

This significant problem was overcome during the next two decades with the advent and gradual spread of radio-tracked balloons known as radiosondes. They transmitted ‘live’ data to their ground launch station from their pressure, temperature, and humidity sensors, and were tracked by radar to derive wind information. Today there are some 600 to 650 unevenly scattered radiosonde stations globally that report pressure, wind, temperature, and humidity twice daily.

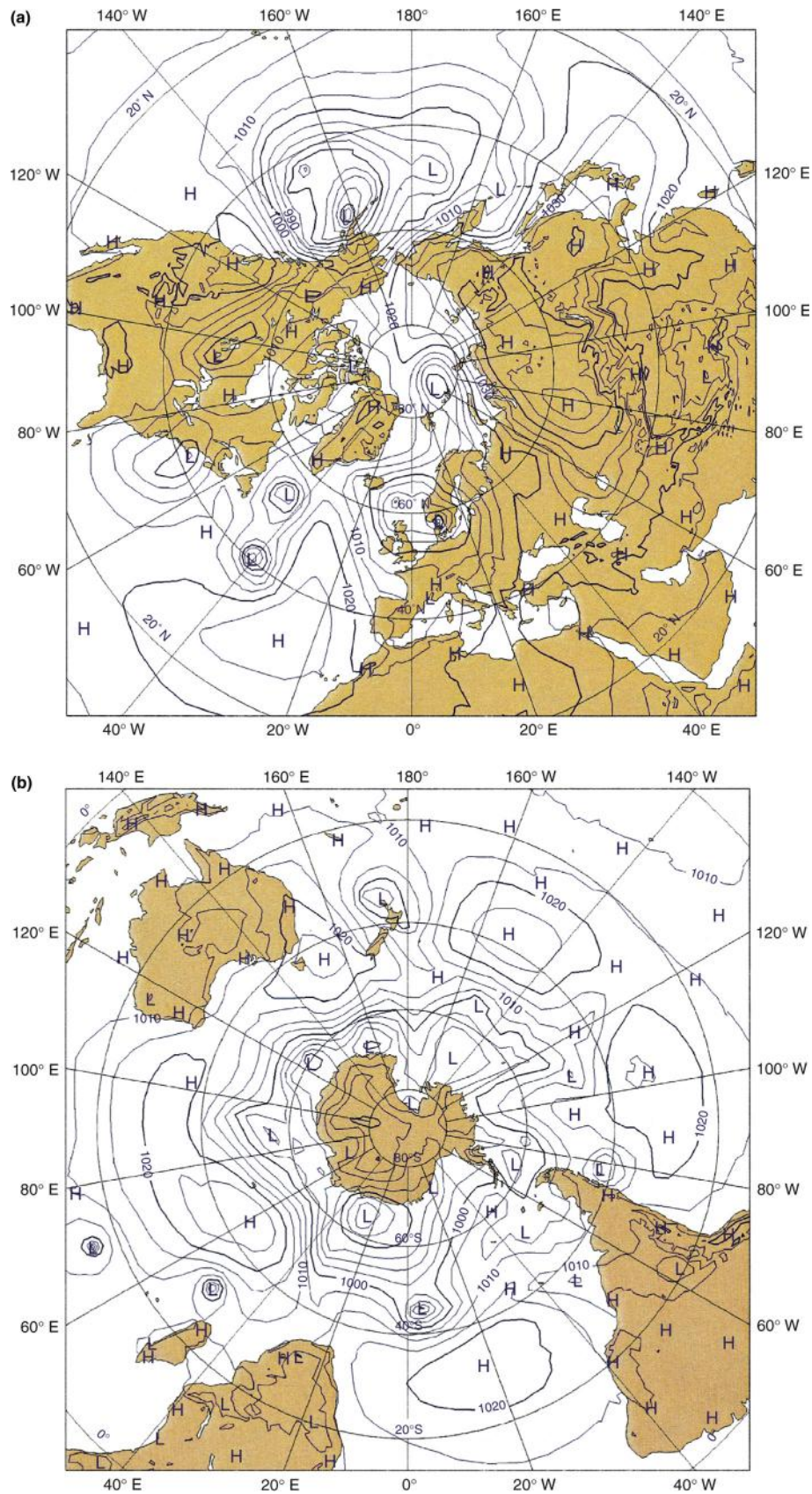


Figure 8 Northern and Southern Hemisphere mean sea-level pressure analyses. Reproduced with permission from the European Centre for Medium-Range Weather Forecasts.

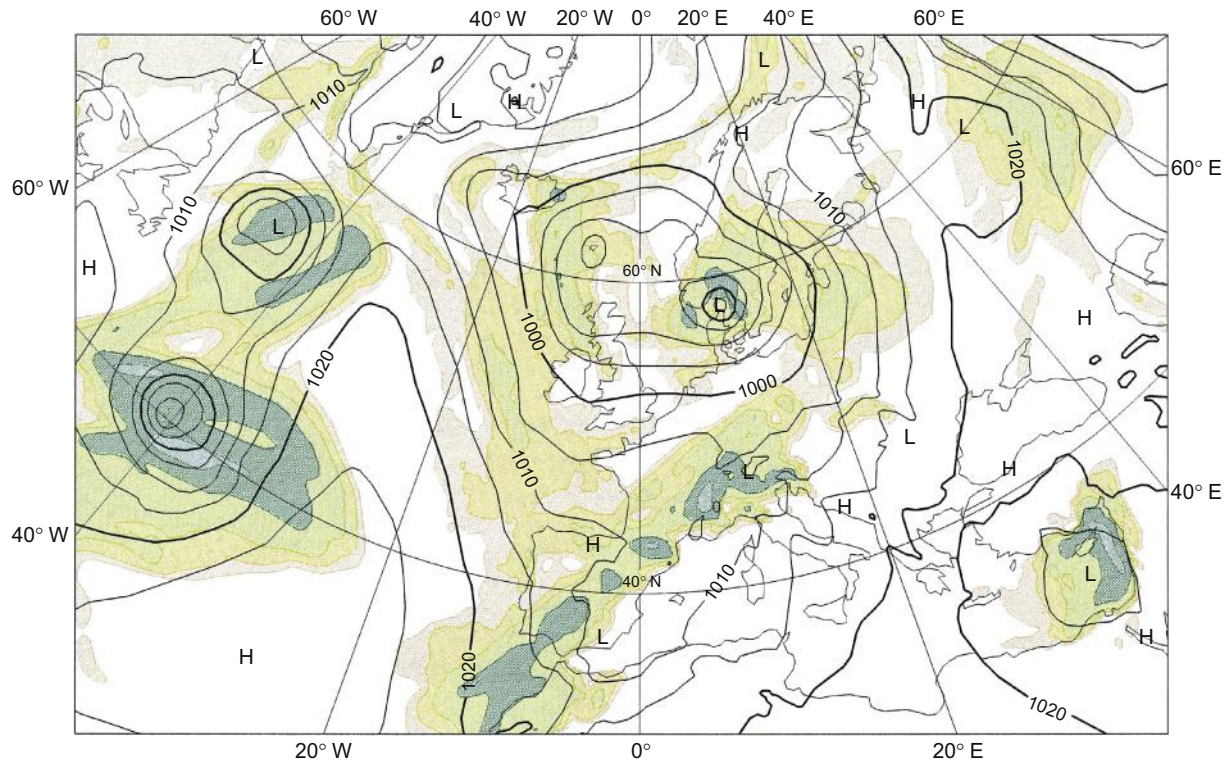


Figure 9 Predicted mean sea-level pressure and 24-h accumulated precipitation (shaded). Reproduced with permission from the European Centre for Medium-Range Weather Forecasts.

Modern Surface Weather Maps

It is true to say that the way in which the surface weather features like highs, lows, and fronts are represented on today's analysis charts is not very different from those of the interwar years of the twentieth century. There has been an extension of the symbolism to include frontogenesis and frontolysis, and greater appreciation by forecasters of the variety of, for example, cold fronts. The advent of weather satellites and radars has aided our knowledge of the variety of frontal structure – and this appreciation has led to the need to inform forecasters of the differences, for example, between 'ana' and 'split' cold fronts. The latter would be ideally represented on a weather map by a surface cold front symbol and a leading upper cold (humidity) front symbol.

The use of supercomputers in weather analysis and forecasting has opened up a massive array of surface weather representations that are mapped automatically. The 'classic' map of mean sea-level isobars is still produced as a global analysis. **Figure 8** exemplifies these for the larger part of the Northern and Southern Hemispheres, from the European Centre for Medium-Range Weather Forecasts (ECMWF). Larger-scale maps illustrate predictions of the mean sea-level pressure field, as well as a derived field. **Figure 9** is a map of the prediction of mean sea-level pressure, valid at 12 UTC at the end of a 48 hour forecast, and of total precipitation during the last of the two days (00 to 24 UTC). The isobars thus provide a snapshot of likely conditions at one instant, while the precipitation patterns are expressions of the rain, drizzle, or snow that is forecast to fall from disturbances over the whole 24-h period in question.

As an extra aid to operational meteorologists, it is possible also to indicate the likelihood that the 24-h precipitation total will exceed some critical value. **Figure 10** highlights three such criteria: the probability (5%, 35%, 65%, and 95%) that the total fall over 24 h will exceed 1 mm, 5 mm, and 10 mm. Similarly, **Figure 11** illustrates the same probability values for the 10 m wind speed to exceed 10 m s^{-1} and 15 m s^{-1} . The latter falls just inside the category of 'gale'.

Modern Upper-Air Maps

As with weather maps for the surface, the range of charts available for the representation of upper-air features has increased dramatically over the last decade or so.

There are still the 'traditional' synoptic maps. At ECMWF, for example, the predicted height field for the 500 hPa surface is charted as it has been for many decades (**Figure 12**). The contemporaneous thermal field at 850 hPa illustrates the large-scale waves of relatively warm and cold air in the lower troposphere. As with some surface phenomena, the model can provide indications of the extent of, for example, predicted 850 hPa anomalies that are greater than $\pm 4 \text{ K}$ and $\pm 8 \text{ K}$ (**Figure 13**).

An innovative synoptic map produced by ECMWF is that of the predicted cloud cover for low, medium, and high levels. Such maps are provided in daily time steps, valid at 1200 UTC, and in essence give a broad indication of what a satellite image might look like at each of these times. They act as another indicator for the forecaster – of whether a particular weather-producing system is composed of deep cloud or shallow cloud, for instance (**Figure 14**).

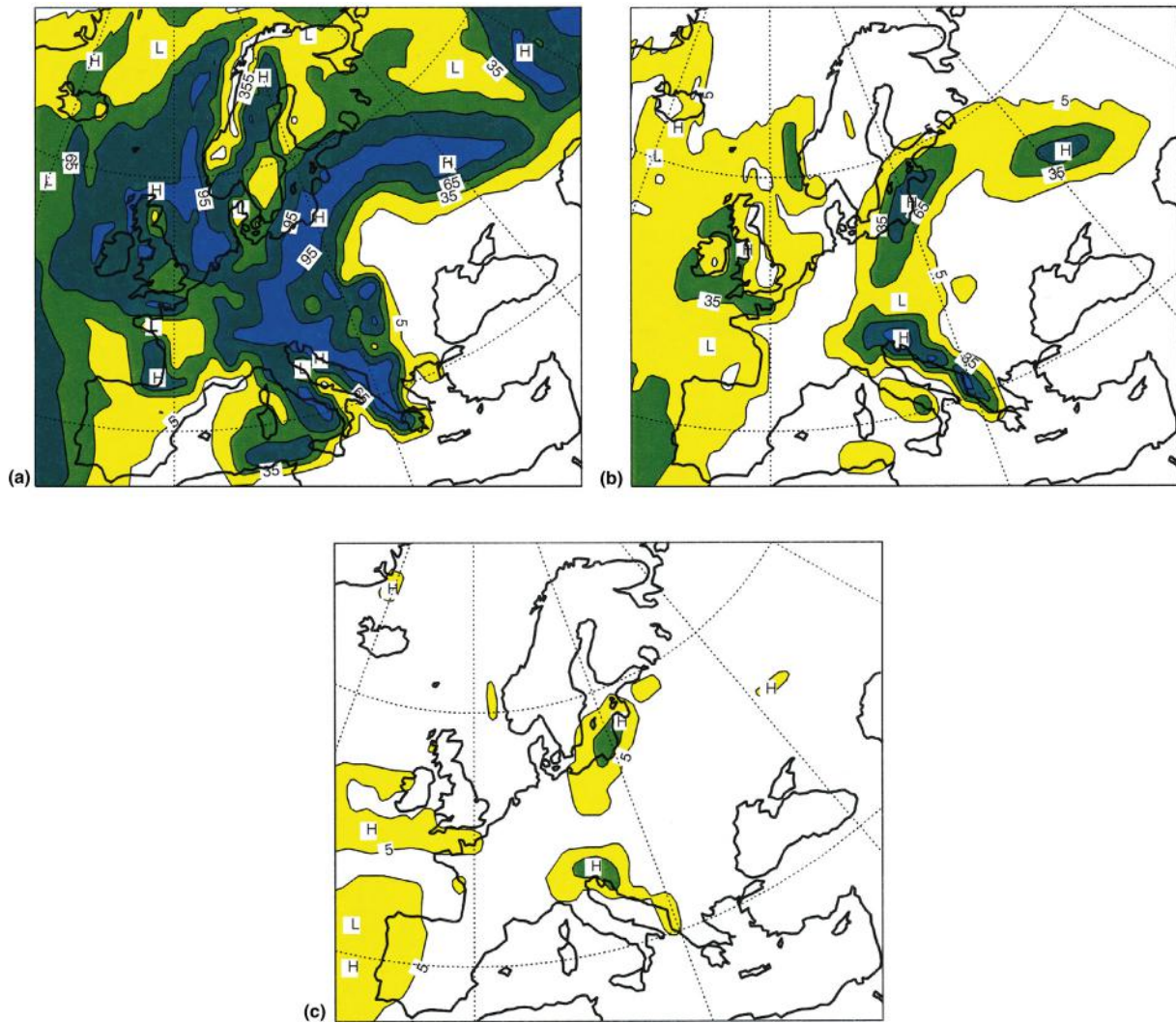


Figure 10 Probability of predicted 24-h precipitation exceeding (a) 1 mm, (b) 5 mm, and (c) 10 mm. Reproduced with permission from the European Centre for Medium-Range Weather Forecasts.

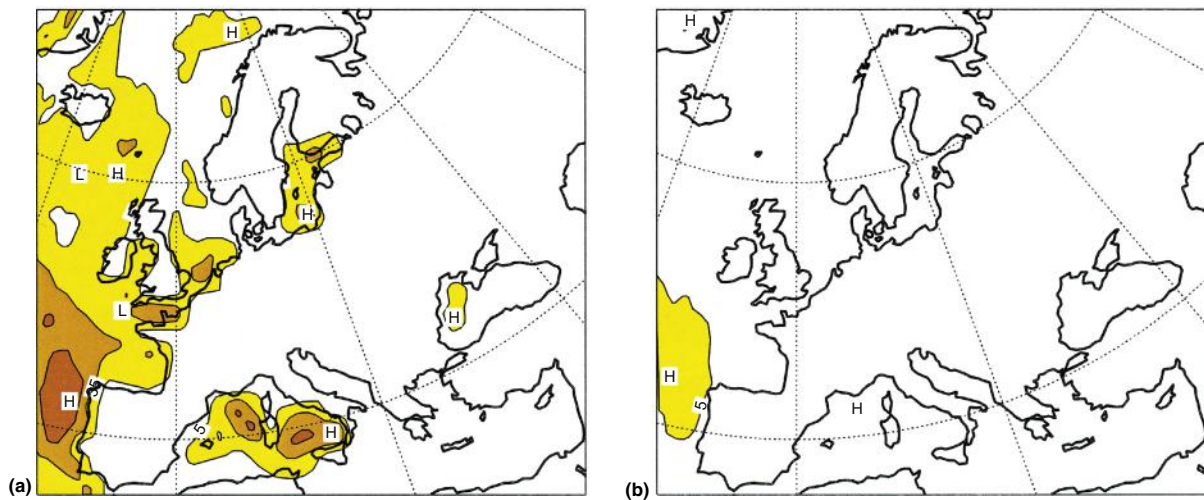


Figure 11 Probability of predicted 24-h 10 m wind speeds exceeding (a) 10 m s^{-1} and (b) 15 m s^{-1} . Reproduced with permission from the European Centre for Medium-Range Weather Forecasts.

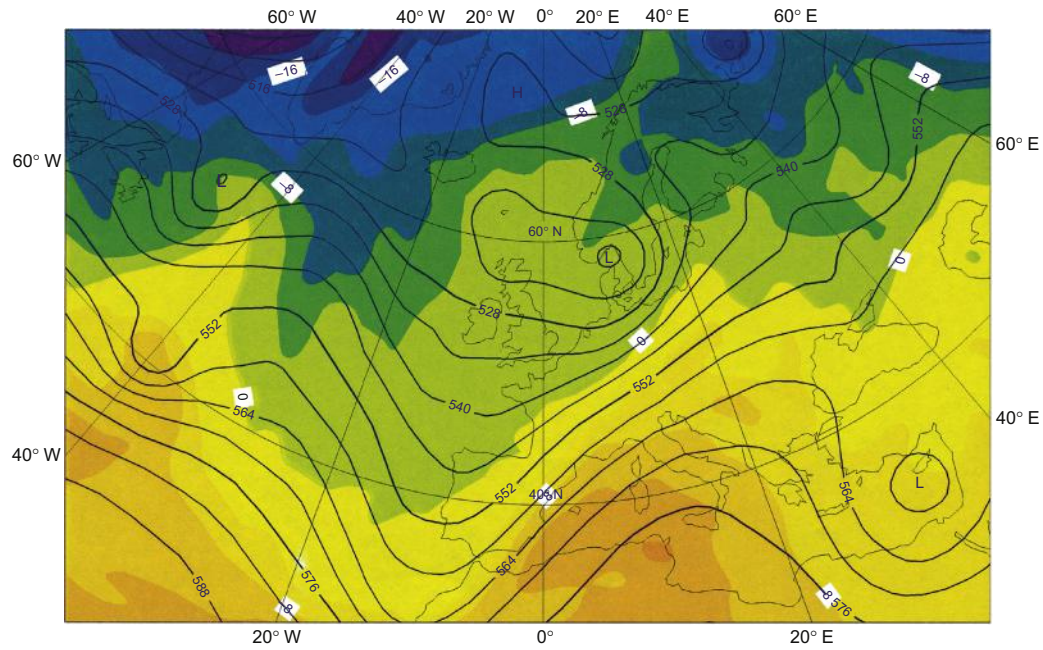


Figure 12 Predicted 850 hPa temperature (deg C) and 500 hPa height (dm). Reproduced with permission from the European Centre for Medium-Range Weather Forecasts.

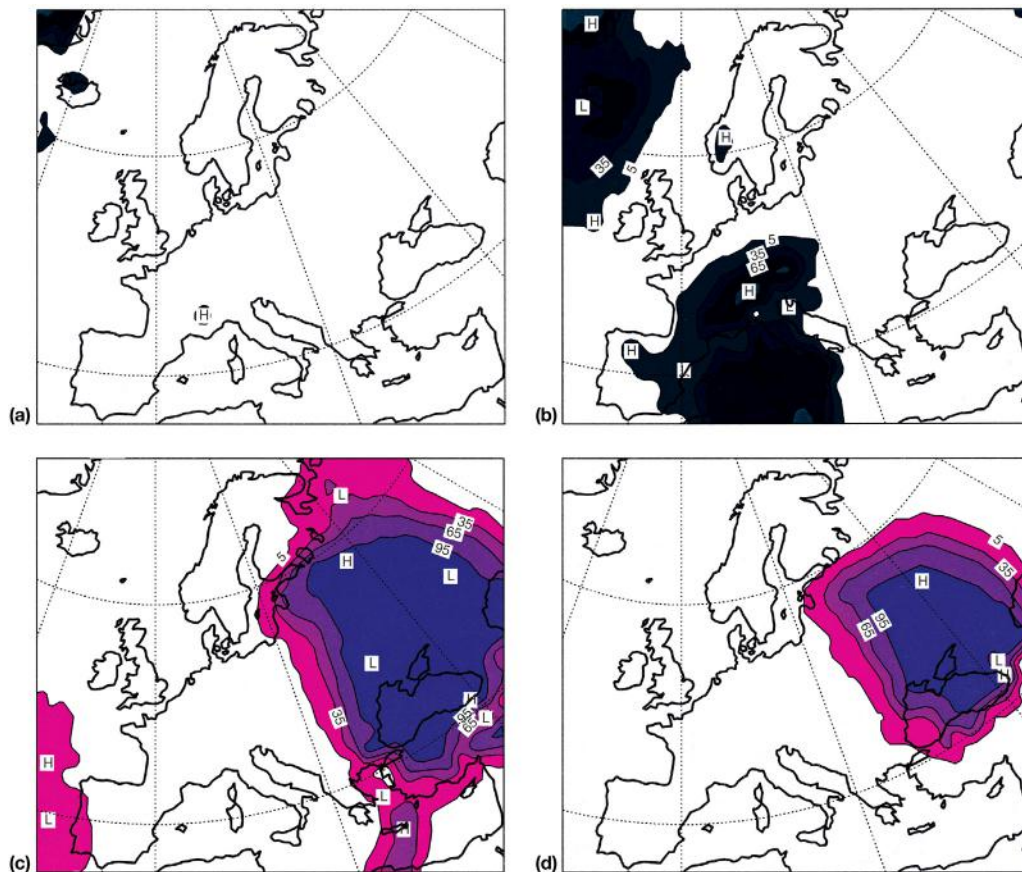


Figure 13 Probability of predicted 850 hPa temperature anomalies: (a) less than -8 K, (b) less than -4 K, (c) greater than 4 K, (d) greater than 8 K. Reproduced with permission from the European Centre for Medium-Range Weather Forecasts.

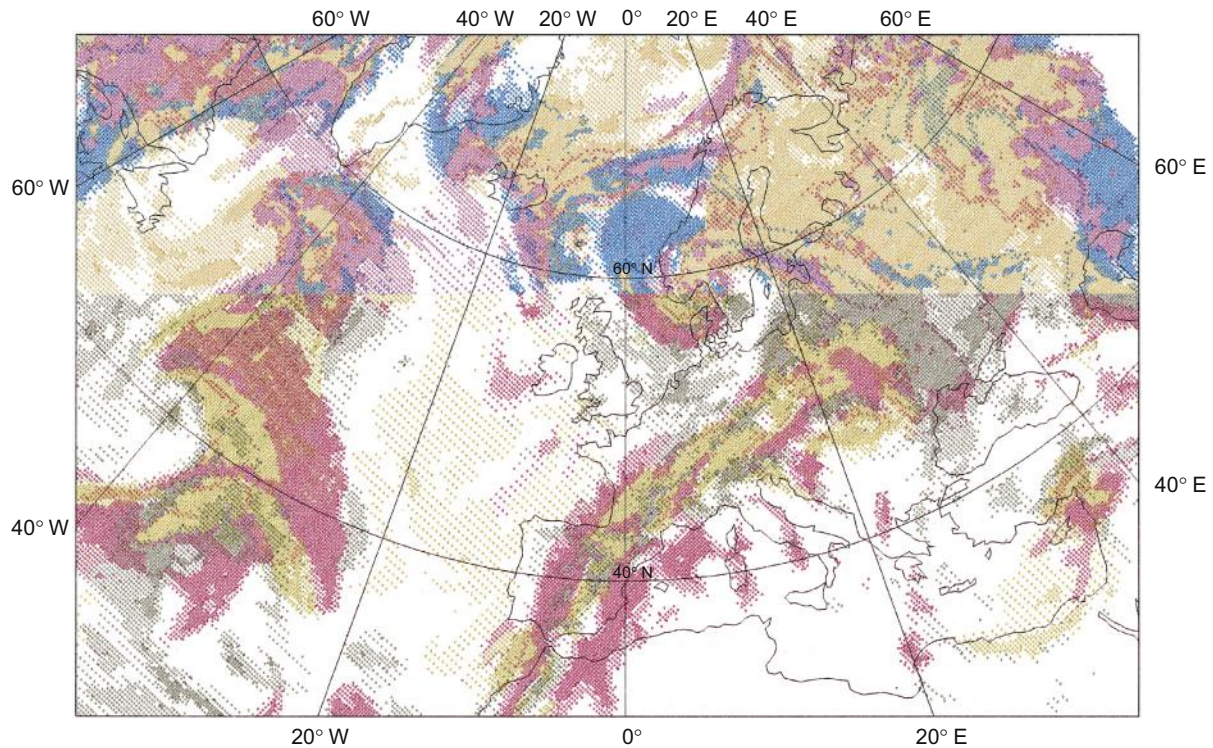


Figure 14 Predicted cloud cover: low (yellow), medium (red), and high (blue). Reproduced with permission from the European Centre for Medium-Range Weather Forecasts.

It is not possible to illustrate the truly enormous range of weather maps available to today's forecaster. The fact is that some charts have stayed the same over many decades, and for good reason. The mean sea-level isobaric and frontal analysis has stood the test of time: there have been extensions of the symbolism used as our knowledge of the variety of fronts has improved. There has not been the need to abandon these representations for something better. Similarly, the standard upper-air isobaric height analyses are still among the working charts that operational meteorologists use.

What is different nowadays, however, is that a whole host of analysed or forecast fields that are a particular forecaster's favorite can be called up at the press of a button. What might be chosen can depend on the situation at hand, and must of course be used profitably within the strict time confines of the operational forecast cycle.

This is the critical change to the production and utility of weather maps today. It is that they can be provided automatically, rapidly, for a larger range of basic or derived fields, and

can be overlain with satellite images, for example. This "richness" can not only aid the forecaster's day-to-day operations but also gradually improve their knowledge and understanding of the phenomena to hand.

See also: Data Assimilation and Predictability: Ensemble Prediction. **Observations Platforms:** Balloons; Radiosondes. **Radar:** Cloud Radar; Precipitation Radar. **Satellites and Satellite Remote Sensing:** Precipitation; Surface Wind and Stress; Temperature Soundings. **Synoptic Meteorology:** Forecasting; Frontogenesis; Fronts.

Further Reading

Monmonier, M., 1999. *Air Apparent: How Meteorologists Learned to Map, Predict and Dramatize Weather*. University of Chicago Press, Chicago and London.

Cyclogenesis

GJ Hakim, University of Washington, Seattle, WA, USA

© 2003 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 2, pp 589–594, © 2003, Elsevier Ltd.

Introduction

Cyclones are associated with horizontal winds circulating in the same sense as the local vertical component of planetary rotation; this circulation is clockwise in the Southern Hemisphere and counterclockwise in the Northern. Cyclogenesis is defined as the formation and amplification of cyclones. The emphasis here will be on cyclogenesis in the extratropical latitudes, near the westerly jet streams; i.e., poleward of about 30° latitude. Extratropical cyclones (hereafter, cyclones) are responsible for organizing significant short-term weather, such as cloud and precipitation patterns. Cyclones are also crucial components of the Earth's climate system because they regulate the pole–Equator temperature contrast, stabilize the density stratification, and maintain the westerly winds in midlatitudes against frictional dissipation.

Cyclones have horizontal length scales on the order of about 500–2500 km and may span the depth of the troposphere (~ 10 km). When compared with an undisturbed environment, cyclones are associated with relatively lower pressure (~ -10 hPa), circulating winds (~ 10 m s $^{-1}$), relatively warm air (~ 10 K), and rising air (~ 10 cm s $^{-1}$). The wind field is nearly in geostrophic balance, so that the wind flows parallel to lines of constant pressure and the wind speed is proportional to the magnitude of the horizontal pressure gradient (the spacing between the pressure contours). Cyclones also represent locally large values of other quantities that are derived from the wind, temperature, and pressure fields, such as: kinetic energy, vorticity, and potential vorticity. Vorticity provides a local measure for the rotation rate of the wind. Potential vorticity is approximately the product of vorticity and a measure of density stratification; it gives the vorticity that a sample of air would have (i.e., potentially) if taken to a reference latitude and rearranged adiabatically to a reference density stratification. Potential vorticity (PV) plays a central role in the modern understanding of cyclogenesis, and is reviewed below under 'Dynamics of Cyclone Development'.

Basic Facts and Definitions

Cyclones can be categorized with regard to the thermal structure of the atmosphere near their centers. Warm-core cyclones are strongest near the Earth's surface (hereafter, surface) and weaken with height. Cold-core cyclones are strongest near the tropopause and weaken toward lower height. Extratropical cyclones can be viewed as an amalgam of warm- and cold-core cyclones, with a warm cyclonic circulation at the surface and a cold cyclonic circulation near the tropopause. These circulations are not aligned vertically, but are displaced laterally by hundreds of kilometers in developing cyclones. This configuration allows the disturbance to extract energy from the jet stream.

Since the atmosphere near cyclones is in, to very good approximation, hydrostatic balance (vertical pressure-gradient

forces balance gravitational forces), cyclogenesis implies that there is a net loss of mass from an imaginary column of air over the surface cyclone. This loss of mass accounts for the drop in surface pressure, and is associated with net divergence of air from the air column. Frictional drag at the surface disrupts geostrophic balance, so that the near-surface winds converge toward low pressure. Therefore we can conclude that upper-level divergence must be important for cyclogenesis. Moreover, since the atmosphere is nearly incompressible on horizontal length scales typical of cyclones, lower-level convergence and upper-level divergence are linked by mass continuity to upward air motion in the troposphere.

It is reasonable to assume the existence of a constant-pressure surface, P_0 , in the stratosphere that is undisturbed by the cyclone. Decreasing surface pressure during cyclogenesis implies that the thickness between a constant-pressure surface near the ground and P_0 increases; i.e., the layer is warming with time. Falling pressure in the cyclone center also implies that the magnitude of the pressure difference between the cyclone center and the surrounding environment, and therefore the geostrophic wind speed, also increases with time.

These deductions on cyclogenesis, based on geostrophic and hydrostatic balance, are self-consistent; however, they do not reveal why or how these changes come about. Further analysis of the time variation of individual quantities, such as surface pressure, are incomplete, since, importantly, all quantities are dynamically related. An analogy for this difficulty considers analyzing a moving automobile to discover the process responsible for locomotion. One might consider the decreasing mass of gasoline in the fuel tank as a crucial aspect for locomotion, rather than as a diagnostic indicator for the action of the internal combustion engine. Fortunately, the dynamics of cyclogenesis can be concisely described in terms of a single quantity that implicitly incorporates all others: the potential vorticity. Before considering the dynamics of cyclones, it will prove useful to document the evolving structure of a typical cyclone.

Structure of Developing Extratropical Cyclones

Cyclones originate on zones of horizontal temperature contrast that are located in the vicinity of the extratropical jet streams (Figure 1(a)); occasionally, the temperature contrast is concentrated in a narrow frontal zone. The nascent cyclone appears as a region of low pressure downwind (following upper-level winds) from a preexisting upper-level disturbance called a short-wave trough. This horizontal displacement between the surface cyclone and the short-wave trough is required for the disturbance to extract energy from the jet stream. Viewed on a level surface, the upper-level disturbance appears as a trough of low pressure or, equivalently (as shown in Figure 1), as a trough of low geopotential height (or, simply, height) on a constant pressure surface. These features are

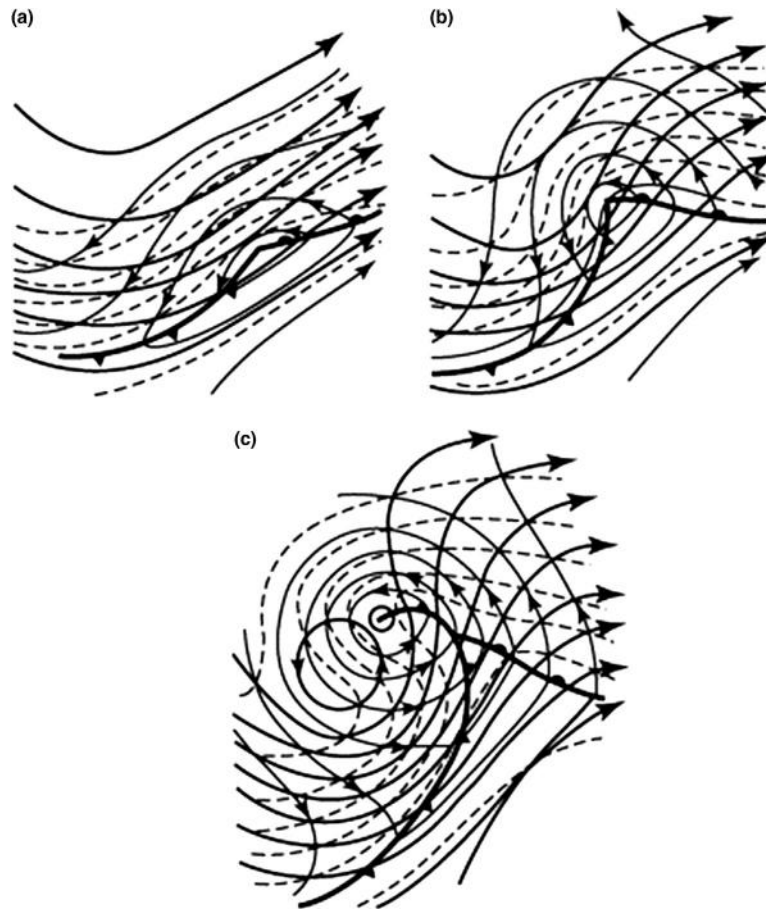


Figure 1 Schematic illustration of a developing extratropical cyclone. Heavy solid lines give contours of 500 hPa geopotential height; thin solid lines give contours of 1000 hPa geopotential height; dashed lines give contours of 500–1000 hPa thickness, which is proportional to the mean temperature in the 500–1000 hPa layer. Fronts are denoted by heavy solid lines with filled half-circles and triangles. (a) shows the early stages of development, (b) shows the mature stage, and (c) shows the occluded stage. Reproduced with permission from Palmén E and Newton CW (1969) *Atmospheric Circulation Systems: Their Structure and Physical Interpretation*. New York: Academic Press.

referred to as short-wave troughs because their 1000–2000 km horizontal length scale compares with greater than 5000 km for so-called long-wave troughs. The short-wave trough is actually a manifestation of a downward displacement of the tropopause, with large values of stratospheric PV extending to lower altitude. (PV is typically much smaller in the troposphere due to less-stable density stratification.) A region of large wind speed often can be found near the trough when the flow is not strongly curved – this feature is called a jet streak, since it represents a local region of strong winds within the jet stream.

As the cyclone continues to deepen, warm air advances poleward and cold air advances Equatorward – the leading edge of these advancing air masses marks the location of surface warm and cold fronts, respectively (Figure 1(b)). Warm, moist air flowing poleward rises at the warm front up and over the colder, denser, air at low levels, producing a widespread region of precipitation. Cold, dense air associated with the cold front displaces warmer air at the cold front, producing a narrow band of precipitation.

Eventually, the cyclone separates from the surface warm front and migrates toward the cold air forming an occluded front. The deepening rate of the cyclone slows during this

time, and eventually ceases when the surface cyclone and upper-level trough are nearly coincident in the vertical (Figure 1(c)). Viewed from an energy perspective, cyclogenesis represents a conversion of potential energy associated with horizontal temperature contrasts into kinetic energy. Cyclogenesis is a thermally direct circulation with warm air rising as it flows poleward, and cold air sinking as it flows Equatorward. A net result of this process is a reduction in the pole–Equator temperature contrast and more-stable density stratification (dense air underlying less-dense air). A typical duration for the cyclogenesis process depicted in Figure 1 ranges from about 12 to about 60 h.

Dynamics of Cyclone Development

Cyclogenesis is associated with the development of disturbances in the fields of wind, temperature, pressure, and density. Deriving a comprehensive dynamical understanding of cyclogenesis in terms of the temporal variation of each of these quantities is laborious and redundant, since these quantities are dynamically related. This difficulty is overcome through

two simplifications that engender concise and deep insight: approximating the dynamical equations (using the quasigeostrophic approximations) and casting the resulting equations in terms of a single variable, the potential vorticity (PV).

Quasigeostrophic Dynamics

The observations that cyclones are nearly in geostrophic and hydrostatic balance motivates adopting quasi-geostrophic (QG) dynamics, which represent a simplification of the actual governing equations. The important result for our purposes concerns the fact that the entire QG system can be formulated in terms of PV, assuming adiabatic and frictionless conditions:

$$D_g Q = 0 \quad [1]$$

where D_g is the time rate of change following the geostrophic motion. Q is the PV, which in the QG case is the sum of geostrophic vorticity and a measure of static stability, except at the surface where Q is proportional to temperature. Cyclonic Q has the same sign as the local vertical component of planetary rotation – positive in the Northern Hemisphere and negative in the Southern. At the surface, cyclonic Q is associated with relatively warm air.

Note that eqn [1] is a simplified version of a more general relation that applies to the unapproximated governing equations. Equation [1] indicates that PV is conserved (i.e., unchanging) following the geostrophic motion, and that QG dynamics are simply described by the rearrangement of PV. From this perspective, all other quantities (pressure, temperature, etc.) can be viewed as subservient to PV in that, it should be emphasized, they *evolve so as to conserve PV*. Since Q cannot be measured directly, it may seem less gratifying than quantities such as pressure and temperature. However, one may imagine that a discovery of ‘PV meters’ prior to barometers and thermometers would perhaps have rendered pressure and temperature as foreign and contrived quantities!

A second important fact about PV is that all other variables can be recovered from it through a process called inversion:

$$I(Q) \rightarrow \vec{V}_g, w, P, T \quad [2]$$

where $\vec{V}_g$ is the geostrophic wind, w the vertical component of the wind, P pressure, and T temperature. Inversion I can be performed analytically for simple distributions of Q , and otherwise can be performed numerically; the details are unimportant here. A useful illustration of inversion applies to a small sphere of cyclonic Q , which can be viewed as an elemental building block from which arbitrarily complex flow patterns can be constructed. Application of I to this small sphere shows that relatively low pressure is found in and around the sphere, a cyclonic geostrophic wind field is found around the sphere, and relatively warm (cold) air is found above (below) the sphere. These patterns have largest magnitude near the sphere, and weaken with distance in all directions. The fact that the fields of wind, temperature, etc., extend far from the small sphere is fundamental to understanding how atmospheric disturbances interact.

More complicated flow patterns can be recovered from this simple example by summing the contributions from PV in all

locations. This ‘superposition principle’ allows one to rigorously decompose the atmosphere into essential components, and to diagnose dynamical interactions among the components. For example, the following section shows how the dynamics of cyclogenesis are given by an interaction between the tropopause PV disturbance and the surface cyclone.

A Potential Vorticity Schematic of Cyclogenesis

Figure 2 gives idealized schematics of the dynamics of cyclogenesis in vertical cross-section and plan view. In the vertical cross-section view, a depression of the tropopause gives a local blob of cyclonic PV on level surfaces (Figure 2(a); note the plus sign). This PV disturbance is associated with the 500 hPa short-wave trough noted in Figure 1. There is low pressure near the PV disturbance and a region of cold air below it in accordance with the simple example given previously. (Note that the upward bulge in the isentropic surface is due to the fact that potential temperature increases with height in a stably stratified atmosphere.) Low pressure at the surface is located near relatively warm air downwind of the upper-level low pressure (cf. Figure 1). In contrast, the temperature disturbances tilt downwind with height; this

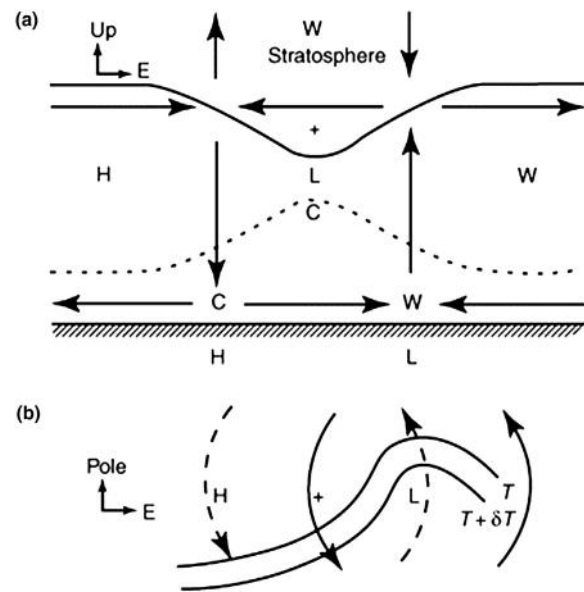


Figure 2 Schematic illustration of potential vorticity concepts applied to cyclogenesis. (a) East-west vertical section showing the tropopause (thick solid line), the surface (thick cross-hatched line), a lower-tropospheric isentropic surface (dashed line), and the air circulations required to conserve potential vorticity for the typical case where westerly winds are increasing from the surface to the tropopause. Regions of relatively warm and cold air are given by W and C, respectively, and regions of relatively low and high pressure are given by L and H, respectively. (b) Horizontal plan view showing the surface isotherms (solid lines), surface wind associated with the upper-level PV disturbance (dashed lines with arrows), and surface wind associated with the surface cyclone (solid lines with arrows). In both (a) and (b), the plus sign shows the location of the PV disturbance due to the lowered tropopause.

configuration is necessary to achieve a conversion of potential to kinetic energy.

The arrows in [Figure 2](#) show the air circulations necessary to conserve PV for the typical case where westerly winds are increasing from the surface to the tropopause. Air converges into the cyclone near the surface and then rises toward the tropopause where it diverges. Conversely, on the upwind side of the PV disturbance, air converges near the tropopause and then sinks toward the surface where it diverges near the surface anticyclone. Note that a reverse circulation exists in the lower stratosphere, and is only hinted at in [Figure 2](#). Downwind of the PV disturbance, these circulations raise (lower) the isentropic surfaces below (above) the level of the disturbance in order to conserve PV as the westerly winds transport the PV blob eastward.

In plan view, the horizontal wind field due to the upper-level PV disturbance extends downward to the surface and contributes to the poleward (Equatorward) transport of warm (cold) air near the surface cyclone (anticyclone) ([Figure 2\(b\)](#)); note that this illustration applies to the Northern Hemisphere). These transport patterns amplify the surface cyclone by warming the center of the cyclone (or, an increase in surface PV, since relatively warm air at the surface represents cyclonic PV). In contrast, the horizontal wind field due to the surface cyclone produces temperature transport patterns that are out of phase with respect to the surface cyclone. These transport patterns are responsible for propagating the disturbance toward the east. Similar interpretations apply near the tropopause, assuming that the PV increases horizontally toward the pole, as is typical. Near the tropopause, winds attributable to the surface cyclone amplify the upper-level PV disturbance, and winds attributable to the upper-level PV disturbance propagate the disturbance toward the west, relative to the westerly winds.

Finally, we note that diabatic heating can be accounted for within the PV framework. Diabatic heating contributes to the generation of the PV disturbances (non-conservation of PV, or a source term on the right-hand side of eqn [1]), with a cyclonic PV disturbance below, and an anticyclonic PV disturbance above, a region of maximum warming. Since the cyclonic PV disturbance below the heating is in close proximity to the surface, it contributes to a reduction in the cyclone surface pressure and enhances the winds at the cyclone center.

Theory of Cyclogenesis

The leading theory of cyclogenesis, baroclinic instability, posits that the westerly jet streams in the extratropics are unstable to infinitesimal disturbances that grow exponentially in time. The most rapidly growing disturbances in this theory are waves in the horizontal direction, and they have maximum amplitude at the tropopause and surface. For typical extratropical conditions, these disturbances have horizontal wavelengths around 4000 km, propagate at about 15 m s^{-1} , and double their amplitude over slightly less than one day. These characteristics are qualitatively in accord with observations, although the growth rate may appear to be too slow to account for the development of observed cyclones from infinitesimal perturbations over a 48 h time period.

Objections to Baroclinic Instability Theory

When surface friction is introduced to the baroclinic instability theory, the already small disturbance growth rates become even smaller. In fact, some research suggests that the jet streams are stable to baroclinic instability; in other words, the growing disturbances do not exist. Another objection points to the fact that observed cyclogenesis events proceed from preexisting large-amplitude disturbances on the tropopause, not infinitesimal noise. Moreover, cyclogenesis events tend to be highly localized in the horizontal, not plane waves as in baroclinic instability theory, and the vertical structure of observed cyclones tends to change with time as compared with the fixed structure of an unstable disturbance.

A second theory for cyclogenesis attempts to account for the noted deficiencies by employing disturbances having structures and growth rates that change in time. Although these disturbances decay over long time intervals, they can produce very large growth over short time intervals. The theory for optimal perturbations formalizes the search for disturbances that have the largest growth over short time intervals, such as 48 h. An amazing result of this theory is that even in cases where friction is sufficiently strong to stabilize a particular jet stream to baroclinic instability, optimal perturbations may still exhibit large growth over the time interval characteristic of observed cyclogenesis. Although the rapid growth and time-varying structure of optimal perturbations make them appealing candidates to explain observed cyclogenesis, there is at present no evidence to support such claims.

Reconciliation of Baroclinic Instability Theory with Observations

Careful examination of observations and the stability properties of observed flows together indicate that baroclinic instability theory can in fact account for observed cyclogenesis, once two factors are taken into consideration. First, observed precursor disturbances tend to be localized in the horizontal direction, which means that a sum of disturbances, rather than a single disturbance, should be considered; the sum allows for a localized disturbance through cancellation of the components away from the local maximum. Second, these localized disturbances have large, not infinitesimal, amplitude (cf. [Figure 1](#) and [Figure 2](#)). An initial disturbance of modest initial amplitude comprising an unstable component can grow over a few doubling times to account for the amplitude of observed disturbances.

It is clear, however, that other factors not emphasized here play a contributing role in cyclogenesis; for some cyclones this role may be essential. For example, diabatic heating appears to contribute positively to cyclogenesis in many cases. Another potentially important process involves the transfer of heat and moisture from the sea surface during oceanic cyclogenesis. The most intense and most rapidly developing cyclones are found over the western North Atlantic and North Pacific oceans, where strong ocean currents maintain regions of large horizontal temperature contrast in the lower troposphere. Another important effect

in these locations is due to cold continental air flowing over relatively warm water, which results in weak static stability. Weak static stability reduces the doubling time for baroclinic instability and may allow greater latent heat release through stronger upward air motion and precipitation. Orographic barriers are also important in initiating or enhancing cyclogenesis, owing to the strong adiabatic warming that occurs in the lee when air flows over the mountains. Cyclones often form in the lee of major mountain barriers and then move downwind.

Another energy source contributing to cyclogenesis involves a transfer of existing energy from upwind disturbances. Since the disturbances propagating along the jet stream are strongly dispersive Rossby waves (phase speeds vary with wavelength, so that a group of waves with similar wavelength does not move at the phase speed), the energy in these disturbances can radiate away from existing disturbances. This energy radiation can contribute to a sequence of cyclogenesis events through a process known as downstream development.

See also: Dynamical Meteorology: Balanced Flow; Baroclinic Instability; Overview; Quasigeostrophic Theory; Vorticity.
Stratosphere/Troposphere Exchange and Structure: Tropopause. **Synoptic Meteorology:** Extratropical Cyclones; Forecasting; Jet Streaks; Weather Maps.

Further Reading

- Bluestein, H.B., 1992. Synoptic-Dynamic Meteorology in Midlatitudes, vol. II: Observations and Theory of Weather Systems. Oxford University Press, New York.
- Carlson, T.N., 1991. Mid-Latitude Weather Systems. HarperCollins Academic, New York.
- Holton, J.R., 1992. An Introduction to Dynamic Meteorology. Academic Press, New York.
- Newton, C.W., Holopainen, E.O. (Eds.), 1990. Extra Tropical Cyclones: The Erik Palmén Memorial Volume. American Meteorological Society, Boston, MA.
- Shapiro, M.A., Grmås, S. (Eds.), 1999. The Life Cycles of Extratropical Cyclones. American Meteorological Society, Boston, MA.

Extratropical Cyclones

A Joly, Centre National de Recherches Météorologiques – Groupe d'étude de l'Atmosphère Météorologique, Météo-France and CNRS, Toulouse, France

© 2015 Elsevier Ltd. All rights reserved.

Glossary

Baroclinic interaction This expression is introduced in this article in order to name the main physical mechanism that explains the development of extratropical cyclones. It is useful to distinguish the actual mechanism that couples finite-amplitude vortices within a baroclinic zone from baroclinic instability, which applies to the spontaneous growth of small perturbations within such a zone.

Initiation of a cyclone The expression cyclogenesis is becoming ambiguous, since it does not necessarily applies to the birth of a new cyclone, but is also employed for its growth and amplification. This is acceptable in the perspective of very basic life cycles based on continuous growth under the same physical conditions. The climatological reality is somewhat different and calls for a more precise vocabulary. This article calls the initiation phase of a cyclone life cycle, the first detection of an individual vorticity maximum, that is the initial formation of a new cyclone.

Maturation phase in a cyclone life cycle This article refers to maturation as the 24-h period preceding the maximum amplitude reached by vorticity at the core of a cyclone in the course of its life cycle.

Weather regime This expression is employed in this article in the sense of a quasi-stationary large-scale flow pattern, 'large scale' meaning on the scale of an oceanic basin like the North Atlantic or North Pacific.

Units

In general, SI units are used for contour intervals in particular for vorticity (s^{-1}). The following exceptions can be found:

Geopotential The SI unit for geopotential ϕ is the J kg^{-1} . However, maps of geopotential are rescaled to the so-called dynamic height ϕ/g , where g is a constant conventional value of the acceleration of gravity internationally agreed upon by the World Meteorological Organization ($g = 9.80665 \text{ m s}^{-2}$). The unit of dynamic height employed is denoted mgp for meter geopotential: other symbols may be used.

Potential vorticity A convenient nonconventional unit is employed, the pvu (potential vorticity unit).
 $1 \text{ pvu} = 10^{-6} \text{ K m}^2 \text{ kg}^{-1} \text{ s}^{-1}$.

Wind speed In addition to the SI unit (m s^{-1}), the practical unit knot (kt) is also employed. $1 \text{ kt} = 1852 \text{ m}/3600 \text{ s}$.

Synopsis

This article opens with an example of extratropical cyclone introducing life cycles and a few other key concepts. The next section documents the distribution of cyclones. This is followed by a brief summary of how cyclones have been conceptualized in the past and how some of them can be seen using data with very little preconceived ideas. The main mechanism involved in cyclone development and some recent views on how this mechanism may be implemented in the atmosphere is presented. The article is then completed by looking at cyclones at two opposed scales: first from a small-scale perspective presenting some key cyclone substructures and then from a large-scale one, with some ideas on the feedback of extratropical cyclones on the general circulation.

Introduction

Extratropical cyclones are the ubiquitous sources of weather variability at mid- and high latitudes on the timescale of a day or so. A full day of more or less continued snow or rainfall, with a low dark gray sky overcast with clouds followed by a colder showery spell is a sequence of events generally organized by an extratropical cyclone moving over a given location.

Extratropical cyclones are key actors of the climate of these latitudes, particularly in the autumn and winter. Their presence during the cold seasons means mild air temperature and water supply, albeit unpleasant at the time it happens, especially

when it takes the form of tons of snow. However, their absence often implies spells of extreme temperatures and, if it lasts, it may be the precursor of drought in the warmer seasons.

Extratropical cyclones are also known as storms, gales, and other tempests when they reach extreme intensities. They are, in fact, the main causes of weather-related disruptions and destructions over areas the size of a state or a country at mid- and high latitudes. These result primarily from extreme winds but slow and extended floods are another hindrance. Indeed, if such cyclones show up every other day over the same places for weeks or even months in a row, the water supply turns into major flooding. Along the coasts, storm surges, namely the sea

level reaching unusual heights, also are dangers related to extratropical cyclones. They can drown dikes and the lowlands that it was supposed to protect.

Because of their large impact on navigation and circulation in general, extratropical cyclones have long been the focus of weather forecast and weather understanding. For some time when the first attempts of applying scientific reasoning to weather were undertaken, extratropical and tropical cyclones have been combined into the number one marine and coastal weather forecast problem. However, although their name somehow maintains this confusion within the language, they are quite distinct phenomena. One characteristic they do share, however, seen from the surface, is that the most intense ones seem to happen quite suddenly as a bad surprise. Predicting the occurrence of a midlatitude storm with enough lead time has challenged forecasters for decades if not centuries. About every forecasting technique that has been devised ever since weather forecast is attempted has been applied to storm alert and has failed. It is only very recently that the best global prediction systems seem to have predicted most major events. As will become apparent in the review that follows, everything might be important for achieving such a result. But this may primarily be attributed to unprecedented observation and data assimilation capabilities as well as stepping back from making deterministic predictions. However, the focus of this article is on the observation and the understanding gained from combining observations and theoretical ideas applied to extratropical cyclones. Although the latter are often integral parts of ‘understanding,’ one ambition of this article is to try to keep a link with real world extratropical cyclones as seen in real data.

It begins with an example of extratropical cyclone from which key aspects can be identified, such as life cycles or relevant key fields for representing them. The next section provides some statistical basis to these aspects; it also documents the distribution of cyclones. This is followed by a brief summary of how cyclones have been conceptualized in the past and how some of them can be seen using data with as least as possible preconceived ideas nowadays. It provides a good introduction to reviewing the main mechanism involved in cyclone development and some recent views on how this mechanism may be implemented in the atmosphere. The article is then completed by looking at cyclones at two opposed scales: first from

a small-scale perspective presenting some key cyclone substructures and then from a large-scale one, with some ideas on the feedback of extratropical cyclones on the general circulation.

An Introductory Case Study

Now that we are well into the weather satellite era, the look of an extratropical cyclone seen from outside the Earth is quite familiar: it is a large roughly λ -shaped cloud system, with the upper part of the λ more or less hooking back about a nearly cloud-free area. An example of this organization is shown in [Figure 1](#). One panel displays a large cyclone on the Earth ‘seen from infinity’ over the northeastern Pacific south of the Aleutian Islands. This image has been built by combining three channels of the US NOAA GOES (National Oceanographic and Atmospheric Administration Geostationary Operational Environmental Satellites) west geostationary satellite. Before the satellite era, the field that best gave the idea of the extent of an extratropical cyclone is the mean sea-level pressure. It is shown in the other panel together with an alternative representation of the cloud system, using an angle preserving plane view. In this framework, an extratropical cyclone is a ‘low,’ that is a broadly circular area of lower pressure. These are relatively large-scale views, with various sources of smoothing, partly due to the large size of that particular case. [Figure 2](#) provides some details of the same case, thanks to zooming on a part of it and of making use of a high-resolution instrument on a low-orbiting satellite. There, the main cloud band appears to be filamented in its highest part transversal to its overall orientation. While this band nonetheless remains a continuous cloudy area seen at that level of detail, this is quite different with the multiple forms of organizations that can be seen within or near the center of the system or in its northwestern part, with many individual clouds.

This weather system has been chosen somewhat randomly to a large extent because at that stage, it looked ‘typical,’ typical in particular of the storms that hit the western coast of North America. Its extent and some of its structure is revealed by three different ways. One uses measurements of the radiation emitted by the atmosphere in the infrared band or infrared wavelength, then collected by a radiometer orbiting at the same speed as the

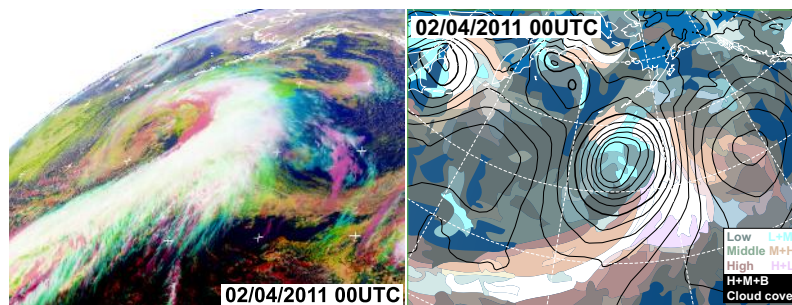


Figure 1 Example of extratropical cyclone in the Northeastern Pacific Ocean. Left: false colors infrared multichannel composition from NOAA GOES west geostationary satellite. Right: analysis from the operational global data assimilation system of Météo-France. Black contours are isobars, interval 4 hPa, thicker reference 1000 hPa. Colored surfaces represent cloudiness derived from explicit cloud water content, combined from 70 levels to three layers (low-level (L), midlevel (M), and high-level (H) clouds). The main colors are given in the figure. The more cloudiness, the lighter the shade, shades correspond to a layer or a combination of these three layers, so that white corresponds to an area overcast by the three overlapping layers. The characteristic cloud system is closely associated with an area of low pressure.

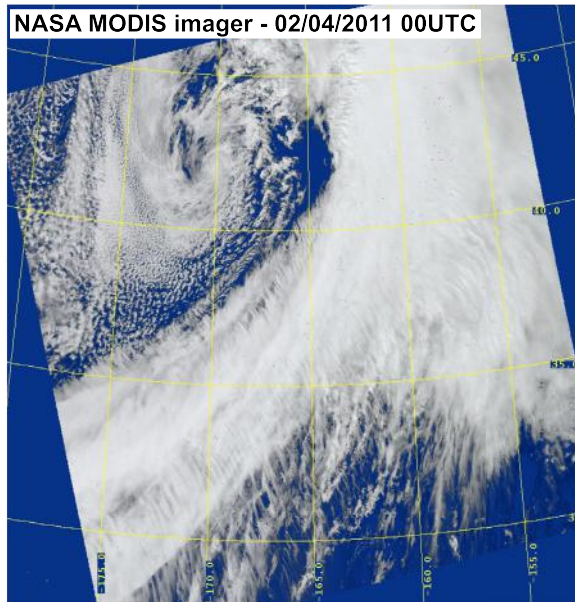


Figure 2 The same extratropical cyclone as observed by the high-resolution visible imager Moderate Resolution Imaging Spectroradiometer (MODIS) from NASA, at about the same time. This image provides details of the different cloud types and their organizations.

Earth, nearly 36 000 km away from it. Another one is, like the human eye, sensitive to radiation reflected by the atmosphere, especially the clouds: the instrument is more like a kind of digital camera that requires an external source of light, here, the Sun. The third representation combines part of these data with many other sources and with a numerical model of the atmosphere: it results from a continuous data assimilation process.

Sources of Data

This leads to presenting briefly what are the various sources of data that can be used to study extratropical cyclones. There are

three main sources of combined, processed informations suitable for dynamical analyses of systems of that scale:

- the products of operational, routine observation, and data assimilation systems. The example case is illustrated using such a source, which is now a mature system processing a very large amount of data.
- the result of dedicated observational programmes or field experiments. This has long been very helpful, even mandatory, because the routine observing system has, for a long time, suffered from many weaknesses and blind spots. To some extent, such programmes remain useful today, thanks to the level of *in situ* data they enable to collect, while the routine network is completely dominated by remotely sensed data. **Figure 3** illustrates such a reasonably recent field project, called the Fronts and Atlantic Storm-Track Experiment (FASTEX). It was ambitiously innovative in that it aimed at tracking cyclones at several times and where they were. Up to then, it had mostly been possible to wait for them at some place, even a wide place. Depending on the programme design, not the same aspects could then be studied.
- both those approaches enable excellent description of some cases, but only a limited number of them. In the case of dedicated observation programmes, there are special efforts to collect significantly more data than the routine network does, but this effort is clearly limited in time. In the case of operational analyses resulting from data assimilation, a large collection of cases is prevented by the lack of homogeneity of the data: such operational systems change quite often, which allows for improving them in many ways. When it comes to ascertain what is 'typical' and what is special, a large, homogeneous set of cases is required. This is provided by reanalyses. Reanalyses are based on the same approach as the first source data assimilation. However, they benefit from two specific features. One is the recovery or reprocessing of past data, some of which was not used in operational analyses. Or, simply, they are observations older than the first operational analysis system. Another

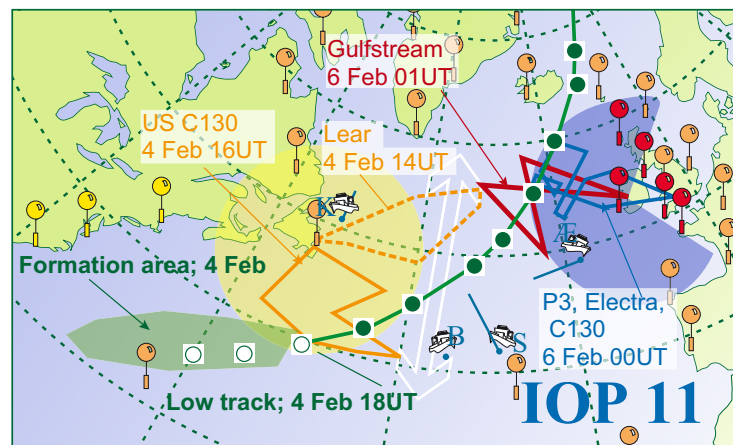


Figure 3 An example of special deployment of observational platforms to document the life cycle of extratropical cyclones, here during the North Atlantic wide FASTEX project in 1997. The balloon symbols mark the location of radiosounding stations. The darker they are colored, the most frequent were the soundings, from 12 hourly to 3 hourly. The ships are also intensive radiosounding locations. Beside the cyclone track, marked 6 hourly by heavy dots, some of the flight tracks are also shown.

feature is that they are performed over rather long periods (half a century, now aiming to cover a century) with the same but recent data assimilation system. This provides both the statistical depth and the homogeneity.

The availability and the growing use of reanalyses to study a number of atmospheric phenomena such as extratropical cyclones from a well organized yet observation-rooted dataset is probably one of the most notable recent progress. Several aspects of cyclones are presented here using several generations of reanalyses.

The data assimilation technique that is behind the description of the example case or behind the most recent reanalysis uses more than 10^6 individual measurements in each 6-h time windows. More than two-thirds of these measurements are satellite derived. Three main sources of data are combined: measurements of many different kinds, now covering the whole globe; a first guess estimate of what the weather is at the time the measurements were made, provided by a model time integration and the previous assimilation window; this embodies the memory of the weather system; and the third source is a mixture of dynamical and statistical constraints, such as an estimate of the uncertainty, which is now changing as the situation evolves. These sources are combined using optimal control theory.

The result of most of these approaches takes the form of collections of four-dimensional (4D) gridded fields. Their most frequent use is for visualization or diagnostic calculations, in the spirit of an essentially observational science. However, more experimental approaches are becoming available as will be shown later. This is made possible because the data organized in this way can also be used for running models, some of them highly realistic.

Life Cycle Description

The cloud and low-pressure system seen in [Figure 1](#) can be recovered at other times, both in the images and in the sequence of analyses. A summary is provided by [Figures 4](#) and [5](#). They are built like [Figure 1](#), with a multichannel infrared image on one column and the clouds and mean sea-level pressure from the assimilations on the other. The first fact to point out is that a cyclone such as this travels and changes shape. [Figure 4](#) begins 500 km south of mainland Japan and continues 4.5 days later with a stormy day at Anchorage, Alaska. It is a distance of more than 7000 km covered at an average speed of roughly 18 m s^{-1} or 35 kt. In terms of intensity measured by pressure deepening, it keeps amplifying throughout this period, reaching a minimum of about 970 hPa, starting at around 1013 hPa. However, this is not a proper measure of the intrinsic intensity of the system because the pressure field also has a strong large-scale component: part of the apparent pressure fall simply results from the system moving northward, where the large-scale pressure is lower. The strongest low-level winds, another indication of intensity related to the horizontal gradient of the pressure field, undergo a different, much more complex history as discussed below. These apparent discrepancies in measures of intensity also are typical. The look of the cloud system, supported by the pressure field, brings, on the other hand, a nontypical aspect of this case: prior to the time of [Figure 1](#), the

system appears to be made up with two systems. Notice that the double-cloud system that can clearly be seen at times 31/03 00UTC and 01/04 00UTC in the images is also well represented by the assimilation fields, in spite of the fact of this event taking place in the middle of the Pacific Ocean. A third system further to the west is also present and vanishes both in images and in assimilation fields.

The underlying representation of the layering of the cloud systems in both series of panels is somewhat different. In the images, there is no signal that can come from below the top of the highest thick (nontransparent) cloud layer. The temperature of this layer is indicative of its height. Combining several channels conveys some idea of the spatial variation of this cloud top height. In the assimilation, cloud cover, here consistent with cloud water content, redistributed from 70 to 3 layers has been used. The color scheme explicitly represents overlapping layers. Hence, white in the images indicates the coldest clouds in all channels, meaning high continuous clouds: this information is conceptually turned into meaning that this is also the location of the deepest, snow-providing clouds. In the assimilation, deep clouds are actually indicated.

The main low is called hereafter Low M, while the other system is denoted Low P, for precursor low because it is the first to form. From this first representation, it could be said that these two systems merge into one. This is only a first impression.

Surface pressure and cloud cover are familiar representations of an extratropical cyclones but more information is needed to start develop an understanding of how all this works. [Figure 6](#) proposes some of this information. Instead of pressure, the low-level part of the systems is now represented by the 850 hPa relative vorticity. This level is located at a height of about 1.5 km. It is the bottom of a troposphere less turbulent than the boundary layer located below. Relative vorticity is the curl of horizontal wind. It is large in the presence of cyclonic rotation induced by the wind field, or in the presence of large cyclonic horizontal shear. The latter is better seen in a high-resolution vorticity field, while the former is most apparent, with smaller amplitudes, in a low-resolution vorticity field. At the times shown in [Figure 6](#), the two systems M and P are closely coupled, possibly with at least another transient one in between on 30/03 12UTC. While the amplitude of Low P increases, that of Low M decreases.

The other fields, like pressure in previous figures, combine the signature of the systems of interest with a larger scale component more representative of their environment. Such a one is the maximum wind field near the tropopause, the surface that can be considered, very loosely, as the top of the weather-active atmosphere. The representation used follows the World Meteorological Organization (WMO) conventions: the direction is the direction from which the wind blows, the speed is given by the barbs and triangles: a full barb is 10 kt ($\sim 5 \text{ m s}^{-1}$); a triangle is 50 kt ($\sim 26 \text{ m s}^{-1}$). The wind is plotted only above a minimum speed, so that provides the shape of the areas of the strongest upper-level wind. It is clear from these figures that the evolution of the cyclones is closely related to the confined maximum of upper-level wind. Because it is of limited meridional extension, this characteristic organization of the wind field is called a jet stream. Notice also that the jet stream changes significantly in the zonal direction: at the

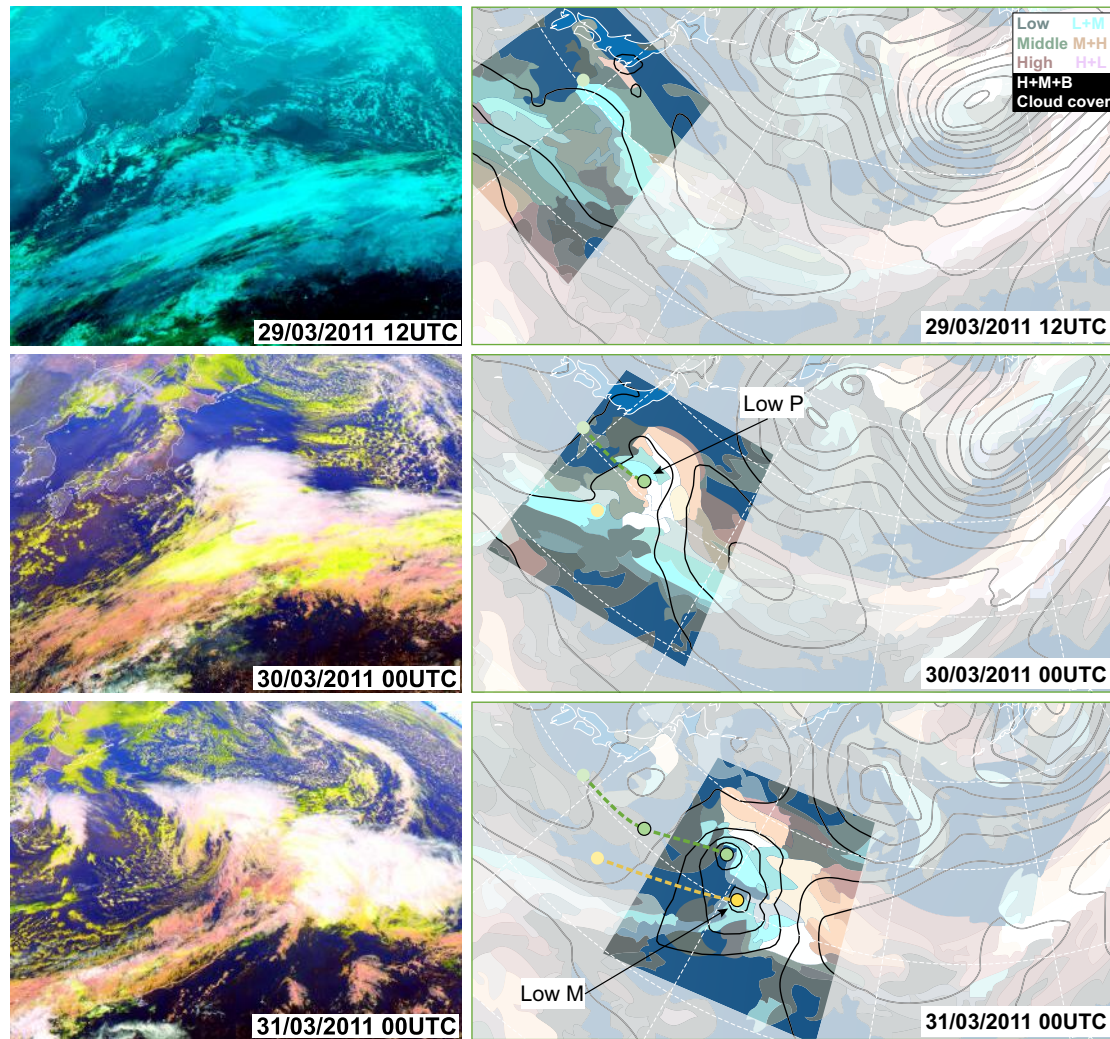


Figure 4 Left column: multichannel compositions from two geostationary satellites over the Pacific Ocean. The channels used are near-infrared ($\sim 0.6 \mu\text{m}$), thermal infrared ($\sim 10.8 \mu\text{m}$), and mid-infrared ($\sim 3.8 \mu\text{m}$). The main cloud systems reaching the highest levels appear white, except in the first image from which the near-infrared is absent because it is a night scene. Images are from the Japan Meteorological Agency Multi-Functional Transport Satellite (MTSAT). Right column: analysis from the operational global data assimilation system of Météo-France, as in Figure 1. Images are centered on the same extratropical cyclones, while maps show them windowed on a single projection. This highlights that extratropical cyclones travel at the same time as they form and grow. In this case, they appear south of Japan and begin to cross the Pacific Ocean.

first time shown (30/03 12UTC), the systems are just west of an area of expansion of the jet, a confluence or entrance zone. More to the east, there is an area of minimum expansion and speed.

The description of the upper level also includes one representation of the tropopause surface, namely its height. Here, the tropopause is given a dynamical representation: it is a surface of constant potential vorticity, instead of being given solely by the change of the vertical temperature gradient. Potential vorticity combines this gradient with others resulting in a quasi-conserved property. The height of the tropopause can roughly be read like a mean sea-level pressure map is. It can be seen that the tropopause is higher on the warm side of the jet stream than on the cold side: the jet stream is also the location of a kind of meridional step of the tropopause height. It can also be seen that the tropopause is especially low on the cold side of the jet, just

west of the systems of interest. Indeed, cyclonelike anomalies can be noted in the height field aloft Low P.

To document further the atmosphere at low levels, contours of moist potential temperature θ'_w are displayed. Like the tropopause height, this field has a strong zonally large-scale meridional gradient. It is enough to keep in mind that there is warm and moist air on the equatorward side and cold and relatively dry air on the poleward side. Values are not given because there are several definitions of moist potential temperature, some giving very different but not relevant magnitudes: magnitudes can be considered to be somewhat arbitrary. Away from the systems of interest, this gradient appears to be located on the warm side of the jet stream aloft. This field is also closely related to the systems of interest: contours are moving into the jet area with them, the gradients also appear to be modified in their vicinity.

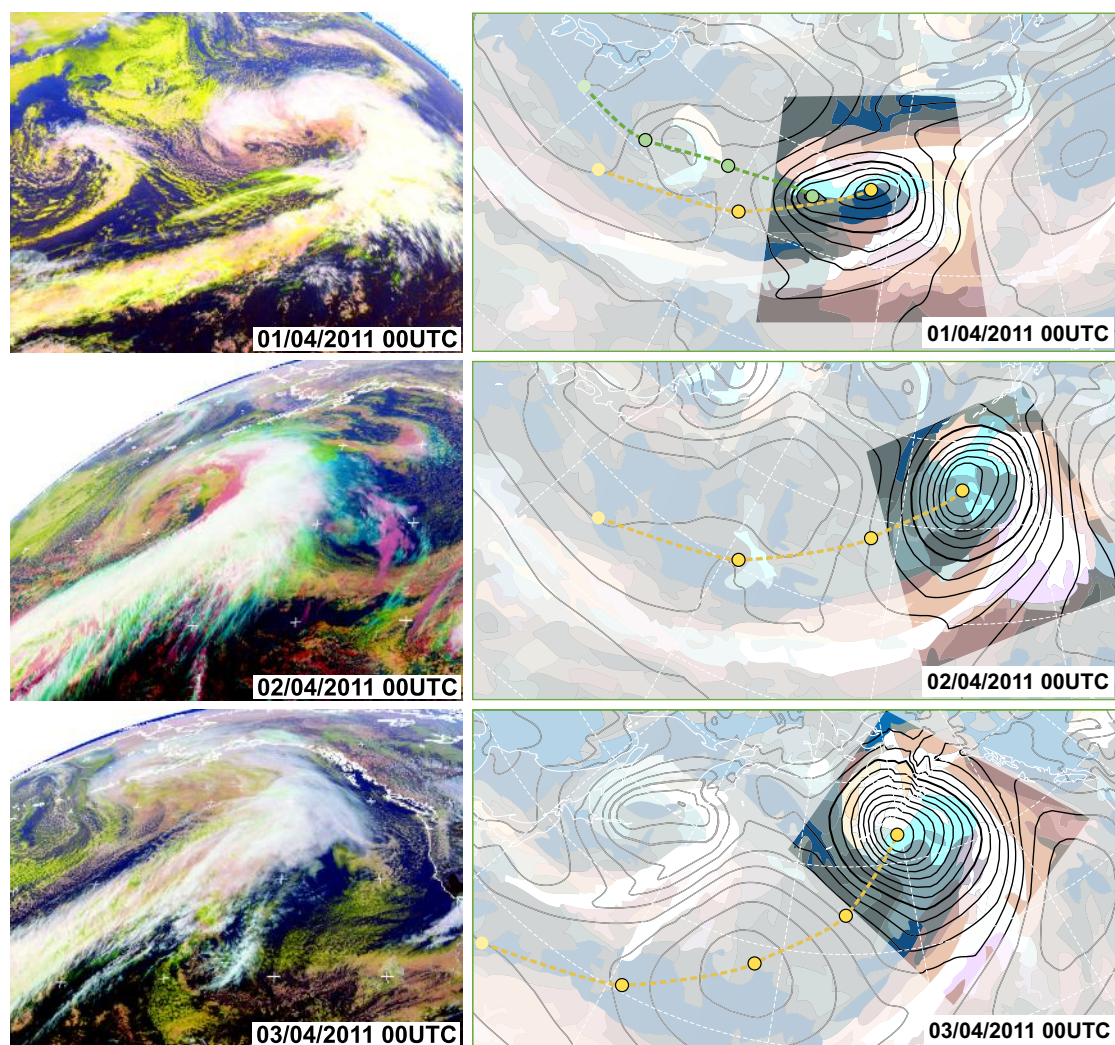


Figure 5 Continuation of **Figure 4**, showing the systems apparently merging into the single cyclone of **Figure 4**, while closing the northeastern Pacific and the coasts of Alaska and Canada. The first image is from MTSAT while the others are from NOAA GOES west satellite.

The picture is completed by the distribution of relative humidity at 700 hPa (about 3 km altitude). This can be seen as an indication of a single layer cloud cover. Nearly saturated areas are also places where precipitation processes are active. This field can also be read as a smoothed version of the vertical velocity. There is a distinct maximum accompanying the systems of interest, but note also the bandlike structure that runs along the warm side of the jet stream.

Figure 7 shows two later times. It appears that between time 31/03 12UTC and 01/04 12UTC, Low M has intensified but it decays somewhat later on. It continues to move northeastward. In fact, and quite remarkably, it has moved from the warm to the cold side of the jet stream in an area where the jet weakens. The upper-level tropopause height anomaly catches up with Low M, later to find itself equatorward of the low-level system at the same time as being heavily stretched along the poleward jet side. The gradients of low-level moist potential temperature have become very strong in some places, but this also is transient and reduces at the last time shown. Finally, Low P, seen through its vorticity signature, does not merge into Low M.

Instead, it seems to orbit cyclonically around it while decaying rapidly and crossing the jet stream from cold to warm side.

After having described the systems and some aspects of their environment when they grow most and reach a mature stage, consider now the conditions at the time when some early sign of their formation can be seen, starting with Low M (**Figure 8**). It may be worth noticing that both these systems, as well-defined low-level vorticity maxima or surface pressure minima, cannot be traced back indefinitely: they both have a genesis or initiation time. The previously described fields are used here as well, although organized slightly differently. The upper-level jet structure is reinforced by showing contours of its speed: the presence of confluence aloft the area of formation of Low M is quite clear. But the picture is dominated by the presence of Low P, which at that time is already quite developed. It strongly influences the low-level flow, as can be seen in both the moist potential temperature contours and the 850 hPa wind field. Actually, the low-level flow, where Low M forms, is quasi normal to the θ'_w contours. At the same location, there is a relative humidity maximum and ascent. A further noticeable

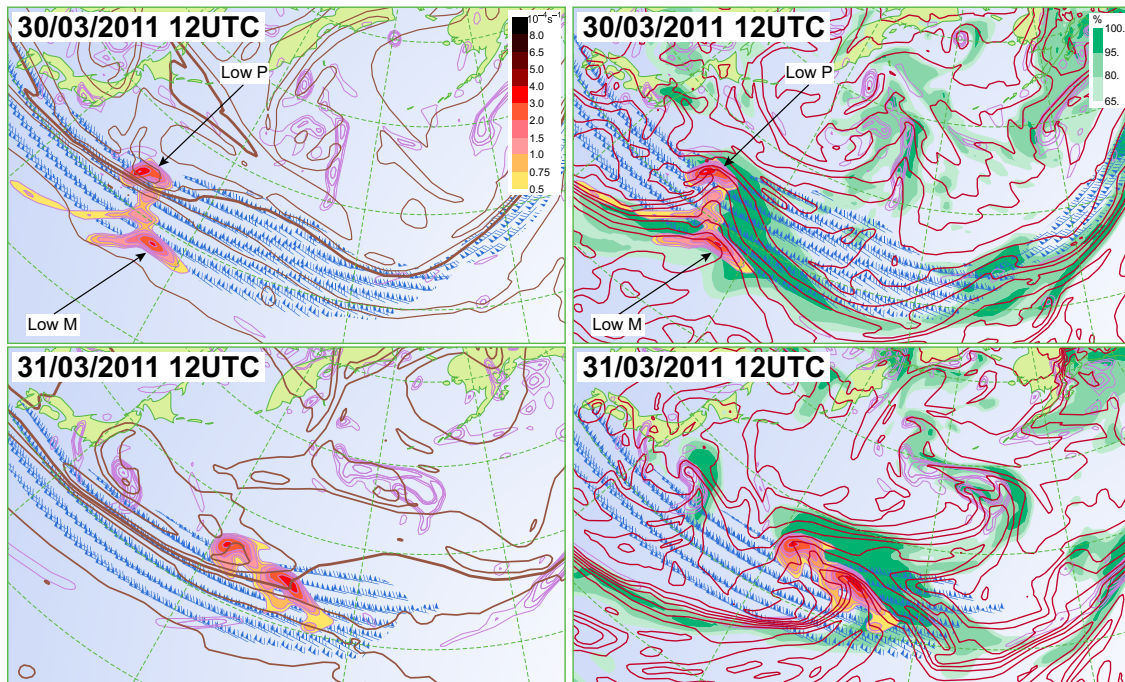


Figure 6 An alternative representation of the life cycles of Low P and Low M, more suitable for a dynamical analysis of their evolution. Instead of mean sea-level pressure, the low-level part of the structure is shown by the positive part of the 850 hPa relative vorticity field (for legibility) (purple thin contours, in units of f , the Coriolis parameter, with yellow to red shading for the systems of interest, using the color scale given in the first panel). The flow at the tropopause is shown by the upper-level wind, concentrated in a westerly jet stream, shown at 200 hPa by the blue barbs drawn when the wind speed is larger than 60 m s^{-1} . This enables to view the flow directions as well as areas of confluence and diffluence, as the major one in the central Pacific. This sequence corresponds to the development phase of the life cycle of Low M. The left column displays another key upper-air field, the tropopause height, defined as the geopotential height of the surface of constant potential vorticity with a value of 1.5 pvu or $1.5 \times 10^{-6} \text{ km}^2 \text{ kg}^{-1} \text{ s}^{-1}$ (brown contours, interval 1000 mgp, thicker contour 10 000 mgp). The right column concentrates on the lower troposphere, with the 850 hPa moist potential temperature (dark red contours, interval 2 K) and the 700 hPa relative humidity (green shading, color scale in the top right panel). All fields are from the Météo-France global data assimilation system.

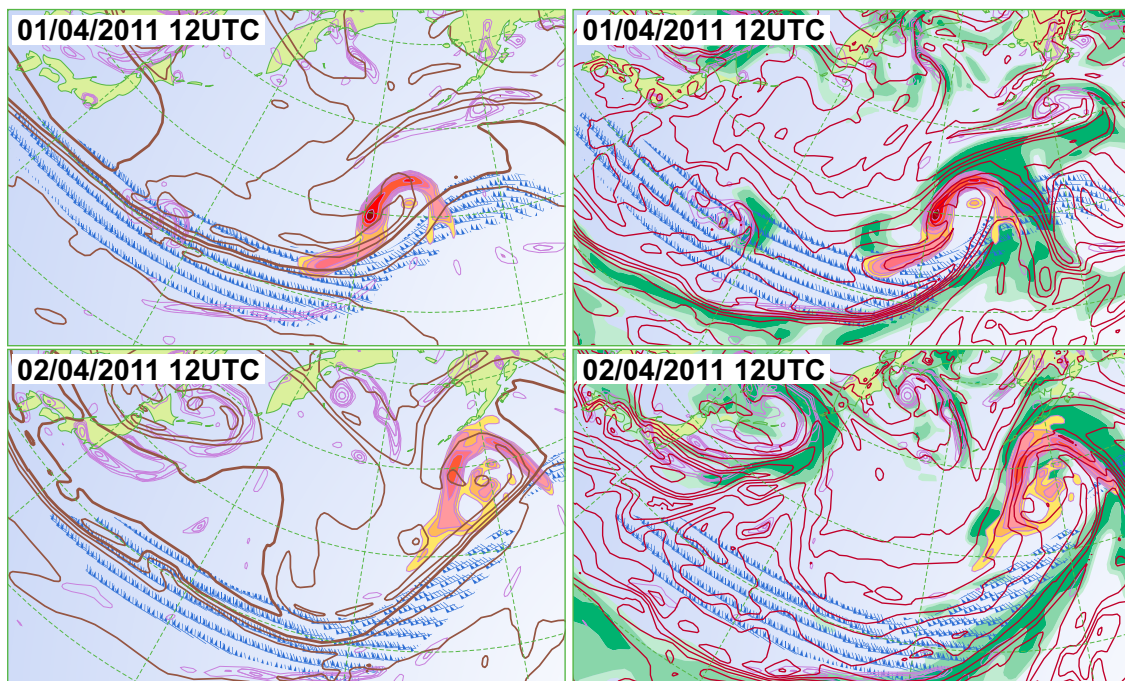


Figure 7 Same as Figure 6 but for the mature phase of Low M.

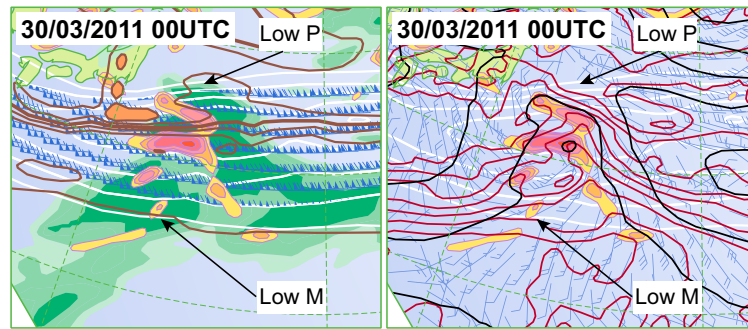


Figure 8 The situation at the time when the incipient Low M can first be hinted in the 850 hPa vorticity field. Left panel: mid- and upper-troposphere fields, superimposed on the low-level vorticity (purple contours, yellow to red shading with scale as in Figure 6). The fields are the 700 hPa relative humidity as in Figure 6, the geopotential height of the tropopause (brown contours, -2000 m anomalies highlighted with orange shading) and the 200 hPa wind field above 60 m s^{-1} (blue barbs and white contours every 10 m s^{-1} above 30 m s^{-1}) also as in Figure 6. Right panel: 850 hPa relative vorticity is related to mean sea-level pressure (black contours, interval 4 hPa, thicker reference 1000 hPa), 850 hPa moist potential temperature (dark red contours, interval 2 K), and 850 hPa wind field (light blue barbs).

feature is the upper-level anomaly of tropopause height, with a well-defined minimum slightly to the west of Low P. Note that the vorticity maximum indicated as the incipient Low M is not the strongest one in the area. This specific attribution may indeed be disputed: it remains that these conditions are creating a number of new systems that, in a way or another, will turn into Low M.

The conditions of formation of Low P share one or two aspects with those of the formation of Low M: the presence of confluence of the jet stream aloft as well as the preexistence of an upper-level cyclonic anomaly seen here in the tropopause height (Figure 9). Unlike the previous situation, the low level appears to be quite unorganized aside from the meridional gradient of moist potential temperature. One sign of this is the lack of connections between surface pressure and vorticity, unlike what was seen at later stages. Moist and ascending areas can be related to the jet (present along the warm side) or to the upper-air anomaly (on its eastward side) but not where Low P begins to form. In short, Low P as a low-level system is a new one, but it seems to have a precursor itself, in the form of an upper-level only system.

Case Study Summary

This real world extratropical cyclone case provides a number of significant facts that will be gathered together here.

Consider first the weather system itself, or rather the two systems Lows P and M. They are well materialized by fields such

as relative vorticity or by the finite-amplitude anomalies in fields such as surface pressure, various temperatures, wind, and of course clouds. One way to summarize their evolution is to plot the maximum vorticity amplitude as a function of time (Figure 10). Both systems have a genesis time in the sense that they can be tracked back up to a certain time before which they cannot be identified any more. Low P undergoes a nearly continuous growth phase that lasts for about 18 h. This is followed by quasi-continuous decay, to the point this system, as a vorticity maximum accompanied by anomalies in other fields, vanishes. It takes about 3 days. Initiation or genesis precedes growth to a mature stage then decay to an end also called cyclolysis. This sequence of events, called a life cycle, seems quite close to the ideal situation most often considered in theoretical views: a single, continuous growth phase follows the first appearance of the system. When this growth phase ends, it is followed by continuous although possibly slow decay. There is, however, a huge difference with the ideal life cycle: the important influence Low P and Low M had on one another.

Low M offers a somewhat different profile. Low M starts with an initial, very strong impulse growth during 12 h like Low P. However, this is interrupted for the next 12-h period. Then a new growth phase begins, less rapid but lasting about 24 h, leading to its peak amplitude. During the next 24 h, it loses quite a lot of intensity. A new, third, slow growth phase takes place. The end of this system is not included in the time slot of this study, and in fact, is not easy to define. This is what

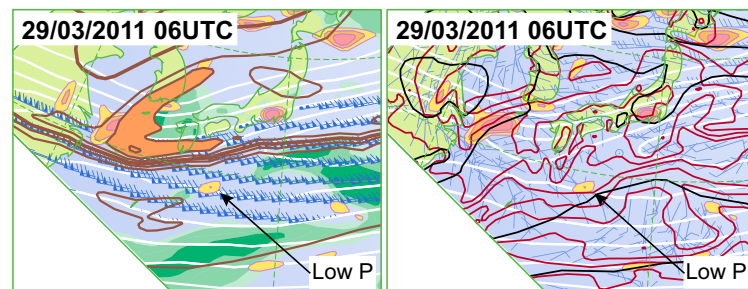


Figure 9 Same as Figure 8, but showing the situation at the time the incipient Low P can first be detected in the 850 hPa vorticity field.

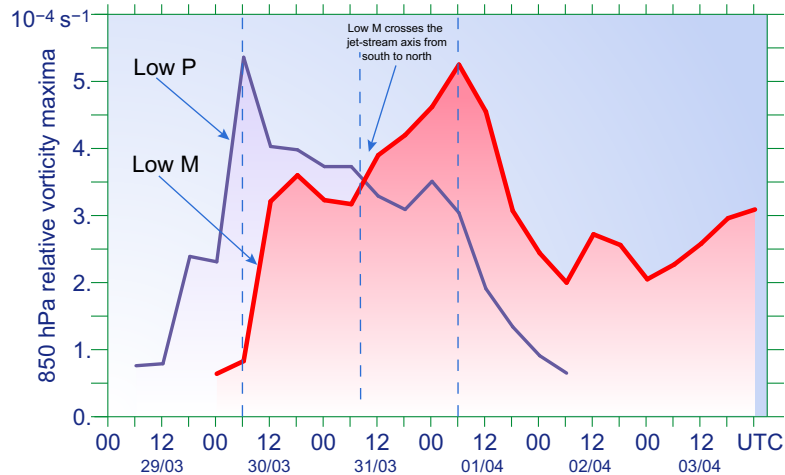


Figure 10 Time evolution of the amplitude of Low P and Low M measured by the maximum relative vorticity at 850 hPa.

is called a complex life cycle which actually is quite frequent, as the FASTEX programme mentioned above particularly pointed out. The fact that there are several successive phases of growth is also apparent in other fields, although their beginning and end might seem a bit different.

In vorticity and some other fields, it is quite easy to identify the systems on the vertical in the lower half of the troposphere and somewhat higher. Their signature reaches the upper troposphere in clouds and for Low M, in the wind field, in the form a secondary, transient maximum sometimes called the outflow jet. When looking at the tropopause, however, it is not really possible to spot a structure with a one to one correspondence with Low P and another one with Low M. What can be seen, however, in their close vicinity is indeed a finite-amplitude anomaly that can be tracked as well. But very much like the end of Low M is beyond the time slot considered, both the beginning and the end of this upper-air anomaly happened elsewhere at other times. In particular, this anomaly exists before the low-level systems do. For this reason, the upper-air anomaly can be, must be considered to be an independent system, although like Low P and Low M, there may be some interactions.

This is more apparent when considering the respective tracks of these three structures, as in [Figure 11](#). This figure also summarizes the kind of environment within which these life cycles took place. Quite crudely, this environment is obtained by averaging a number of fields over the period considered: this is a simple way to bring out some key aspects of the large-scale flow. More rigorous time filtering techniques are applied to obtain the research results shown later on. The trajectories, and therefore the life cycles are closely related to the jet stream, which is also an area of stronger meridional gradient of moist potential temperature in the troposphere. The upper-level anomaly and Low P remained on the cold, poleward side of the large-scale jet stream. As this figure shows particularly well, Low M, on the other hand, started on the warm side of the jet and then crossed it. It then continued to move toward Alaska remaining on the poleward side. The crudeness of the way employed to display some characteristics of the environment is probably responsible for the extension of the jet northeastward

from Central Pacific to North America: this is more likely to be the trace of the jet outflow of the then mature Low M.

It is of interest to pay attention to the positions of the lows at several key phases of their life cycles. This figure confirms that both geneses took place while the large-scale jet accelerates. It also shows that the first growth phase of Low M ends while it reaches the longitude of the jet maximum. The second growth phase, on the other hand, corresponds to Low M crossing the jet and also reaching the diffluence area east of the core. Finally, during the growth phases of Low P and the first two of Low M, the upper-level system is located to their west or northwest.

[Figure 12](#) concludes this first description of the key features of the life cycle of extratropical cyclones by focusing on the growth phase of Low M that led it to its mature stage. Maps are provided 12 hourly and, most important, they are kept centered on Low M maximum vorticity, in order to better see the relative positions of the other features as it amplifies. The pressure falls by about 20 hPa in 24 h: this is a quite rapid deepening and another measure of development. The jet outflow is clearly visible in the last panel at least as a secondary maximum at the east of the low. The position of the upper-level anomaly is indicated by highlighting areas where the tropopause is more than 2 km down into the troposphere: it is right above Low P, northwest and west of Low M.

Some of these facts are specific of these cases and rather unusual. But some others are really characteristic of extratropical cyclones. Prior to review what sense is currently made of some of the features seen in this section, their representativeness needs to be documented.

Overview of the Global Distribution of Cyclones

The reanalyses that have been produced these past years are invaluable when it comes to shape what is a 'typical' extratropical cyclone in its 'typical' environment. As explained above, reanalyses are collections of consistent time series of gridded fields, a form that enables many new possibilities with respect to working directly on raw data or even on operational datasets. Very large homogeneous samples are available

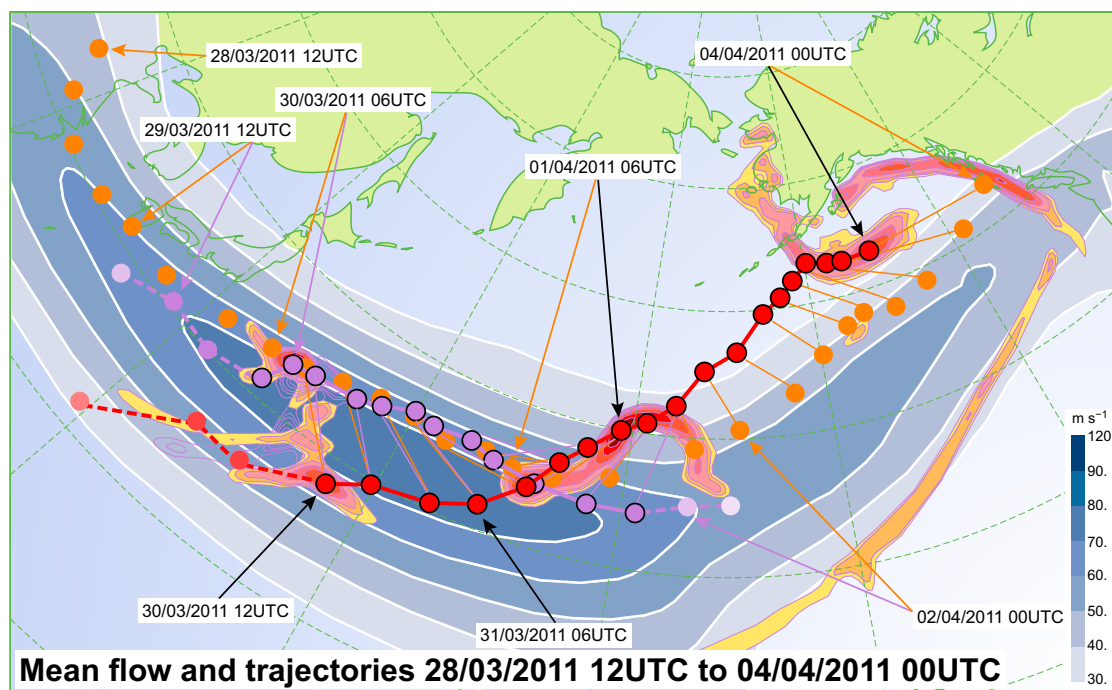


Figure 11 Summary of the life cycles of Low P and Low M through their trajectories relative to the mean westerly flow during their lifetime, representative of their large-scale environment. Locations are shown every 6 h. Red circles and line: Low M trajectory. Purple circles and line: Low P trajectory. When these systems can be seen as closed lows in the pressure field, circles are contoured in black. When the detection is not fully definite, a dashed line is employed and the circles have lighter color: this is particularly true for the very first circle of each trajectory. The 850 hPa vorticity field directly related to the systems is shown at four times (purple contours, yellow to red shading, see Figure 6 for the color scale). The mean 200 hPa wind speed is in the background, white contours and blue shading, interval 10 m s^{-1} above 30 m s^{-1} , color scale in the figure. The series of orange circles tracks the minimum tropopause height that can be followed during the same period. Thin colored lines link positions at the same time, but this link does not imply necessarily an actual physical connection.

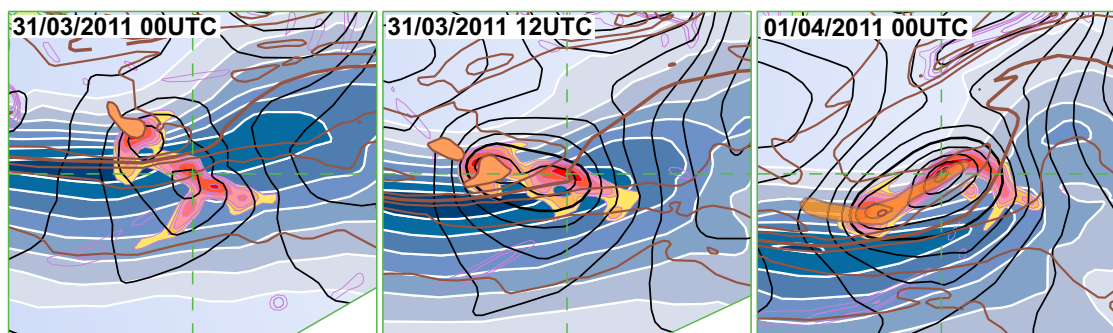


Figure 12 The main growth phase of Low M shown in a kind of system-relative frame of reference. Namely, each panel is centered on the 850 hPa vorticity maximum of Low M and oriented so that the large-scale flow and baroclinity is oriented from left to right. Fields are 850 hPa relative vorticity (purple contours and yellow to red shading as in Figure 6), mean sea-level pressure (black contours as in Figure 1), tropopause geopotential height (brown contours with orange highlighted low height anomaly as in Figure 8), and 200 hPa wind speed (white contours and blue shading, interval 10 m s^{-1} above 30 m s^{-1} , color scale shown in Figure 11).

through reanalyses, complex computations may be performed on them. A further novelty has been introduced soon after the reanalyses. Synoptic meteorology concepts or objects, such as an extratropical cyclone, a front or a jet streak have begun to receive definitions amenable to algorithmic, and therefore automated, identification in a given set of maps at a given time. Then, comparing successive maps, these structures can be tracked.

Previously, analysis fields were used as gridded times series. Various purely statistical approaches have been applied to them, including time filtering, spatial correlation structures of time filtered fields and the like. In this case, there is no actual meteorological 'object': they are implicitly represented by the choice of a characteristic time period. On the other hand, when studies were undertaken specifically on such 'objects,' most of the work had to be done by eye and hand. As a result, the

samples studied are quite small. Of course, both approaches have provided useful, indeed, essential results. But reanalyses enable to remove the drawbacks of both approaches and allow to reach a further level of description.

One popular reanalysis is known as ERA40 (European centre for medium-range weather forecast (ECMWF), reanalysis covering roughly the last 40 years of the twentieth century). Moving from the North Pacific to the North Atlantic, consider now some results obtained from extracting north-hemispheric fields at a resolution of 1.5° horizontally, 13 pressure levels vertically, every 6 h and for a cold season defined from 15 October to 15 April from 1958 to 2000.

An elementary processing, a simple average of these 42 seasons, provides first a statistically more robust description of what the 'typical' environment of extratropical cyclones is (Figure 13). As in the Pacific case study, a meridionally confined, essentially zonal jet flow dominates the scene. Actually, the oceanic basin is the location of its core, located aloft the eastern coast of the United States and of a vast zone of diffuence. Contours of constant tropopause height or 850 hPa moist potential temperature θ'_w are mostly oriented from west to east, with large values equatorward decreasing poleward. Notice that the stronger the upper-level wind, the larger the meridional gradients of both these fields.

To see that this is indeed the large-scale environment of extratropical cyclones, one of the new approaches enabled by reanalyses has to be employed. It is the automatic tracking of cyclones. To obtain the first results shown here, the following definitions are employed. A cyclone is a well-defined relative vorticity maximum at 850 hPa, that is vorticity is larger than at other grid points within 380 km of the location considered. Certainly only the synoptic scale component of vorticity is needed here. Tracks are constructed first by pairing individual cyclonic events every 6 h. This is performed by applying several criteria taking the wind at the level of the vorticity maximum and the wind at 700 hPa into account and maximizing the likelihood to find the most reasonable association. Once a first

set of trajectories has been obtained, they are also evaluated as a whole by considering some regularity criteria. As this leads to question some of the pairings, the algorithm is iterative and performed backward and forward in time, until the trajectory set does not change any more. One aspect of this algorithm worth pointing out is that it uses very little preconceived ideas. The main one is the use of the 700 hPa wind for estimating a priori a future propagation location based on a kind of wave specific of these environments in a rotating frame called Rossby waves, discussed below. In particular, this technique does not impose any vertical constraint on vorticity. One result from this weakly selective approach is that the number of events found is very large, with many weak and short-lived systems. Some further selection is performed, so that an extratropical cyclone in this paragraph is now a vorticity maximum at low level with a characteristic scale of at least 380 km, a lifetime of at least 24 h and reaching an amplitude of at least $1.5 \times 10^{-4} \text{ s}^{-1}$.

The density of the trajectories of the extratropical cyclone obtained in this way appears in Figure 14. The density of trajectories indicates how many trajectories move over a given point on a 380-km resolution grid during a given reference time, namely an entire cold season. In this approach, one event represented by its full trajectory is counted once per point it moves over: this is different from counting the events themselves. The latter will count the same event several times at a point if it becomes stationary. The displacement speed of the systems can also be computed, and its averaged value at each point is indicated in the figure. The connection of the tracks of extratropical cyclones with the jet flow becomes clearer: they tend to begin close to the mean jet core and continue toward the northeast more on the poleward side of the jet. They also move faster near the jet and on its equatorward side, with velocities of about 30 kt (15.4 m s^{-1}) than on the polewardmost cold side. The selection criteria employed select only significant events, of the same kind as those in the case study. A bit less than 7000 trajectories are included: that is roughly such an event each day of each season somewhere in the area covered by the map in the

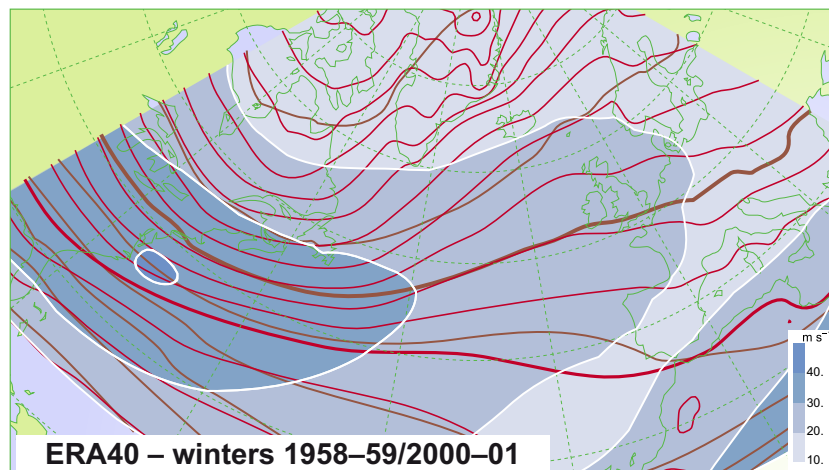


Figure 13 Averaged conditions over the North Atlantic Ocean basin derived from the ECMWF ERA40 reanalysis for cold seasons between 15 October and 15 April. Brown contours, the tropopause geopotential height (interval 1000 mgp, thicker reference 10 000 mgp). Dark red contours, 850 hPa moist potential temperature θ'_w (interval 2 K, thicker reference 283.15 K or 10°C). White contours and blue shading, tropopause wind speed (interval 10 m s^{-1} , color scale in the figure).

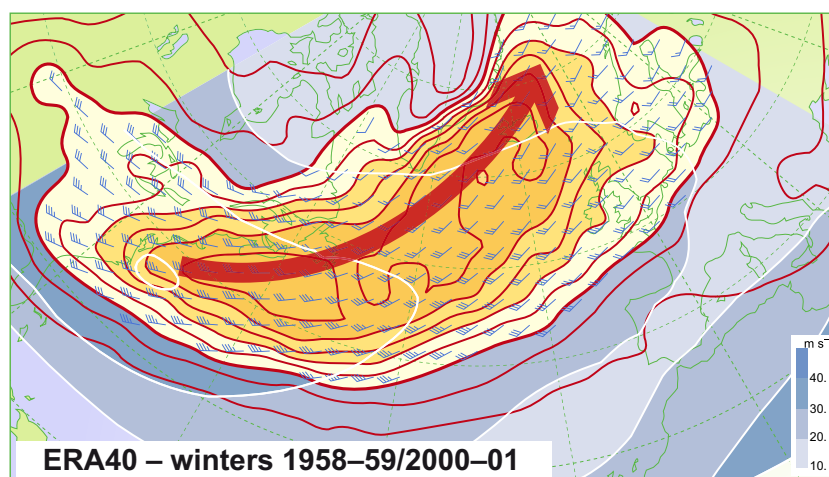


Figure 14 The winter North Atlantic stormtrack as derived from the ECMWF ERA40 reanalysis. Red contours and shading are trajectory density: the number of cyclones passing close to each of the grid points on average during each cold season. Contour interval three cases per season, thicker reference line nine cases. Blue bars indicate the average speed and direction along trajectories at each point. Only a subset of systems reaching a threshold amplitude and lasting a threshold time are included in this map. The large red arrow materializes the most frequent path taken by extratropical cyclones. The white contours and blue shading are the upper-level mean wind during that same period of 42 cold seasons, with the scale given in the figure. Figure constructed with the tracks obtained by Joly, B., Ayrault, F. from Météo-France, from the ECMWF ERA40 reanalysis 6-hourly fields.

figure. Actually, their distribution within the 6-month period considered can be computed and they turn out to be most frequent in January and February. The mean lifetime is 5 days but the median lifetime is 3 days: there is a large dispersion.

The combination of the large-scale jet stream, with its accompanying meridional gradients of temperature and tropopause height together with the main tracks of cyclones forms a large-scale dynamical system called a stormtrack.

Consider now the whole of the Northern Hemisphere. The data employed to construct **Figure 15** is the first generation European Centre for Medium-Range Weather Forecast (ECMWF) reanalysis, called ERA15 (ECMWF Reanalysis, 15 years, 1979–94), supplemented with operational analyses and the equivalent National Center for Atmospheric Research (NCAR) reanalysis. The cold season is also reduced to the more traditional months of December, January, and February (the so-called DJF season). The tracking technique is more general than the one just described in that it can be applied to a variety of features: vorticity maxima as above but pressure anomaly minima as well (which forces to define what an anomaly is) or vorticity minima (anticyclone tracking) or ascending velocity maxima. Nonetheless, the figure shown is for vorticity maxima tracks. Densities and mean speed compare reasonably well over the North Atlantic with the more recent one in **Figure 14**. This is worth mentioning, as it shows results are not too sensitive to the remaining arbitrariness of the approaches as well as to the dataset. Beside the North Atlantic, there are other areas where extratropical cyclones are to be observed. One is the North Pacific: the density map indicates that the trajectories followed by the systems in the case study above are typical of that region. The western Pacific off Japan appears to be an area where cyclones are even more frequent than in the Atlantic. There are also very frequent cyclones just east of the Rockies and over the continental United States, the Great Lakes especially. Other favored locations are Siberia at high enough latitude

and the Mediterranean. This spatial distribution suggests that extratropical cyclones, while not absent from some continental areas, are primarily a maritime atmospheric phenomenon.

Many properties can then be computed to provide sound statistical informations on cyclones, using the trajectories. One key concept introduced with the case study is that of life cycle. The density of appearance of a new cyclone can be mapped, as well as the density of disappearance. One such map is included in **Figure 15**, or genesis. This map confirms the existence of maxima of initiation of new systems off the eastern coasts of the United States and of Japan, that is on the westernmost part of the stormtracks. However, there are also more localized and more active areas of genesis over Alberta in Canada and Colorado in the United States, always just east of the Rockies and near Mongolia, northeast of the Tibetan Plateau. There is a weaker maximum of genesis east of that other large-scale mountainous area in China. The Gulf of Genoa in the Mediterranean also shows up here as a place of frequent genesis. Cyclones disappear in places traditionally known as the Icelandic Low or the Aleutian Low.

Several properties relating to the growth of the cyclones can also be derived. **Figure 15** includes a map of mean growth rate defined as $\bar{\sigma} = \zeta^{-1} \Delta \zeta / \Delta t$, where ζ is the vorticity amplitude along the track. The largest growth is observed over the Rockies and at the western beginnings of the main oceanic sets of tracks in the Atlantic and the Pacific. Their average values are about $0.4\text{--}0.6 \text{ day}^{-1}$ ($\sim 5\text{--}7 \times 10^{-6} \text{ s}^{-1}$). The number of growth phases along the tracks can also be monitored. For example, in the ERA40 sample in the North Atlantic, it is found that about 17% of the cases decay after they have first been detected, but only 4% never grow again at later stages. About 40% undergo a 'standard' life cycle, that is a single growth phase peaking some time between genesis and lysis. Complex life cycles happen to the others, in a form or another, with 28% having two growth phases. The mean lifetime when

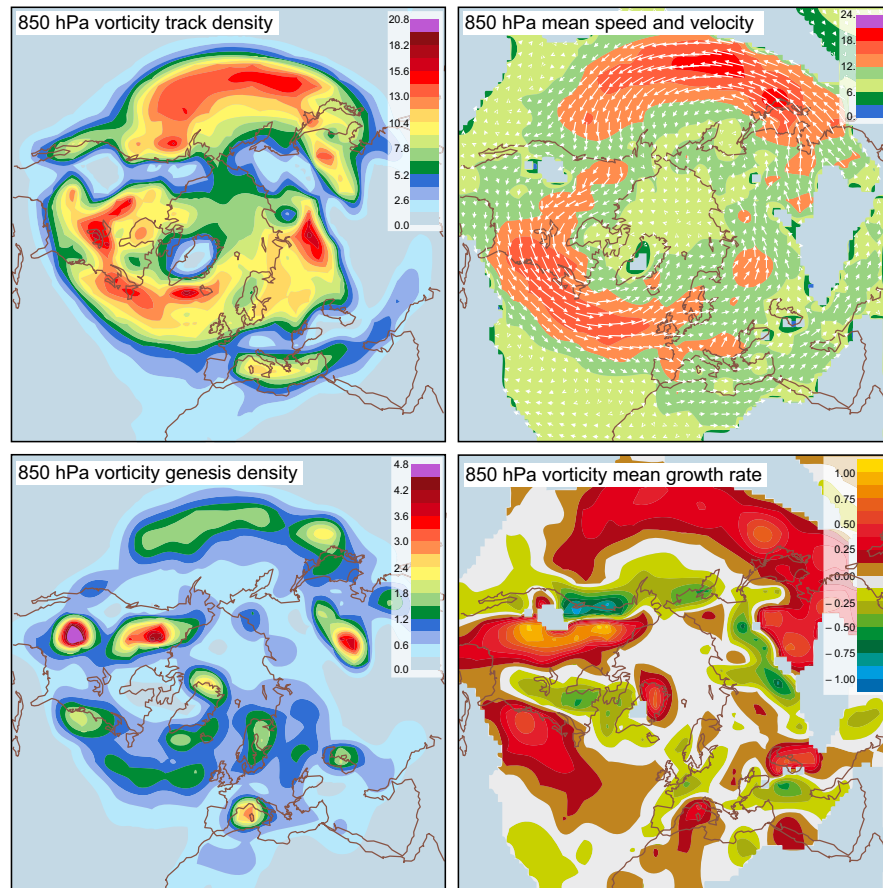


Figure 15 Results from an automatic 'feature' tracking algorithm covering the Northern Hemisphere wintertime (DJF), based on a nonhomogeneous mixture of ECMWF ERA15 reanalysis, ECMWF operational analyses, National Center for Environmental Prediction (NCEP) reanalysis. The feature selected here is 850 hPa relative vorticity maxima. Upper left: track density over areas of 5° spherical caps per month. Upper right: mean speed and velocity. Lower left: genesis density, or density of first detection of new systems, same convention as for tracks. Lower right: mean growth rate, that is average value of $a^{-1}da/dt$, where a is the amplitude, in units of day^{-1} . From Hoskins, B.J., Hodges, K.I., 2002. A new perspective on northern hemisphere winter storm tracks. *Journal of Atmospheric Sciences* 59, 1041–1061.

maximum amplitude is reached is 1.3 days, the median 1.8 days. Comparing with the total lifetimes, this suggests, as seen with the Low M case above, that growth rather takes place at the beginning, the first 2 days in general and it is followed by a rather long decay phase of several days, with a large case-to-case variability.

It is also possible to document extratropical cyclones in the Southern Hemisphere using the same data and techniques. Some results obtained with the ERA40 reanalysis are shown in [Figure 16](#) and for south hemisphere winter: the months of June, July, and August. This figure offers a comparison of vorticity maxima tracking against pressure anomaly minima. Both indicate a simpler distribution in the Southern Hemisphere than in the northern one. There is a single semicircular maximum of track densities extending a bit more than 180° from the central southern Atlantic Ocean to the southern western Pacific, with relative maxima in the Indian Ocean to the southwest of Australia and south of New Zealand. Genesis preferably takes place just east of the Andes in Argentina and on the coast of Antarctica, south of New Zealand. Growth rates have the same magnitude as in the Northern Hemisphere.

Composite Types of North Atlantic Extratropical Cyclones

The use of reanalysis data enables to make a further step toward portraying the generic or 'typical' extratropical cyclone. In this section, the first generation ECMWF reanalysis, ERA15, is employed in order to build composites representative of hundreds of individual cases, described entirely numerically with a set of consistent fields and of known representativeness.

ERA15 covers the period 1979–93. It is an extraction of the extended winter season previously mentioned (15 October–15 April) over a domain centered over the North Atlantic.

The trajectories of 850 mbar vorticity maxima have been constructed as in the previous section. In order to distribute events into classes, the large state vectors that result from extracting fields in boxes of size of about 2500 km centered on the low-level vorticity maxima along the trajectories are reduced by a principal component analysis. The remaining components are distributed into classes using ascending hierarchical classification. Paradigms of situation conducive

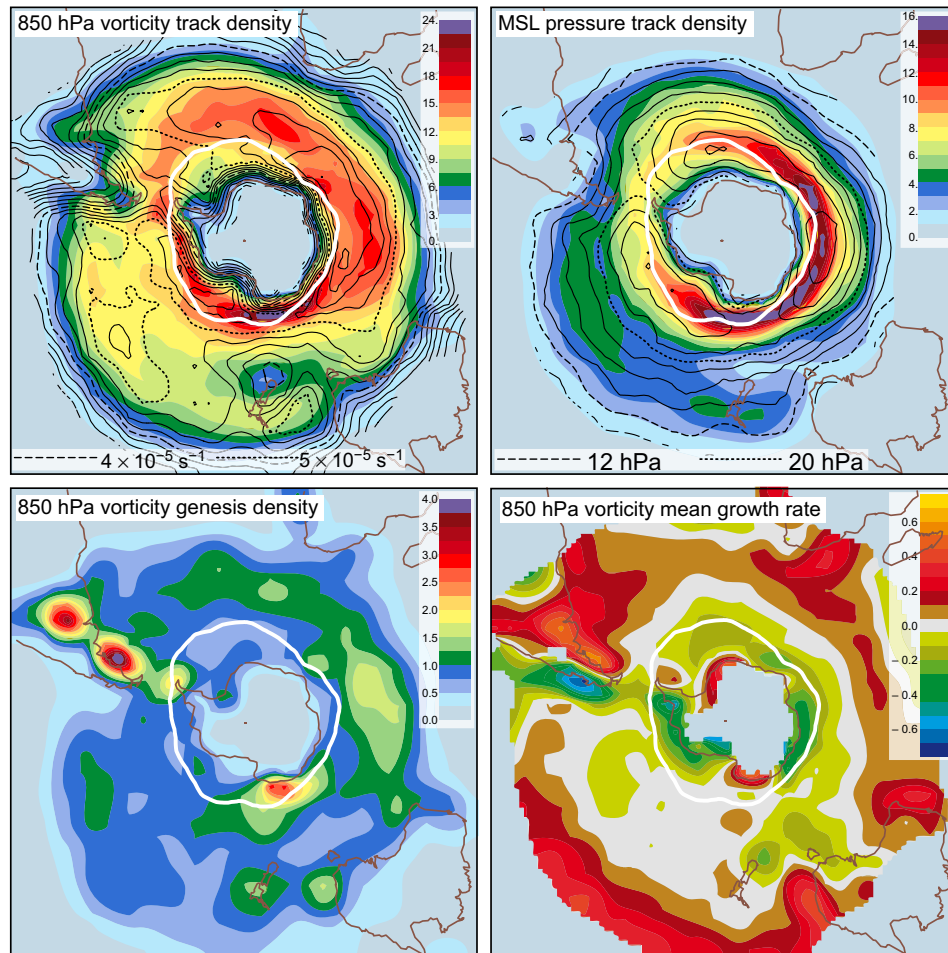


Figure 16 Same as Figure 15 but covering the Southern Hemisphere wintertime (JJA) and based on the homogeneous ECMWF ERA40 reanalysis. Upper left: 850 hPa vorticity maxima track density over areas of 5° spherical caps per month. Contours are mean amplitude, contour interval $5 \times 10^{-6} \text{ s}^{-1}$, two reference contours are shown below the panel. Upper right: track density resulting from tracking mean sea-level pressure minima (shading) and mean amplitude (contours, contour interval 2 hPa, two contour references provided in the panel). Lower left: genesis density of the tracks of 850 hPa vorticity maxima. Lower right: mean growth rate of maximum vorticity amplitude in units of day^{-1} . From Hoskins, B.J., Hodges, K.I., 2005. A new perspective on southern hemisphere storm tracks. *Journal of Climate* 18 (20), 4108–4129.

to cyclogenesis are then obtained by averaging the original fields within each class.

Study of the distributions of amplitudes, of durations, and of the various growth phases discourages the automatic classification of full life cycles. Two particular stages have been isolated. One is the 12-h period centered on the 0 h time of each case, extrapolating a location for the field extraction at -6 h. This phase is called the initiation of a new cyclone. The other is the 24-h period preceding the time of maximum amplitude; this is called the maturation phase and it generally includes the period of maximum vorticity growth undergone by each case.

Summary of Genesis Situations

The classification of initiation has 12 classes. The best previous studies dealt with two dozens of cases, while this one uses nearly 6000 cases. As it is impossible to fully account for all the classes here, composites are gathered into ‘families’ of which there are five.

The most frequent family corresponds to the genesis of a cyclone on the northern side of the baroclinic zone, in the cold air (28% of cases). Nearly as frequent, there is a family that comes close to the extensively studied frontal instability problem (24%, Figure 17). It departs from the idealized situation on one critical aspect: the flow is three-dimensional (3D) and the new cyclone forms at the end of the gradient zone, below a jet entrance. This generic situation shares some aspects with the initiation of Low M in the case study above, except for the apparent absence of an equivalent of Low P. Neither of these families seems to feature an upper-level precursor of the cyclogenesis. The next two, on the other hand, do have such a preexisting structure of cyclone scale: the class that corresponds to a surface development induced by an upper-level disturbance and a new family where the precursor is a full-scale cyclone that splits into the old and a new system. The former seems to account for the initiation of Low P. As will be seen in the next section, it is also a well-known situation, indeed a kind of

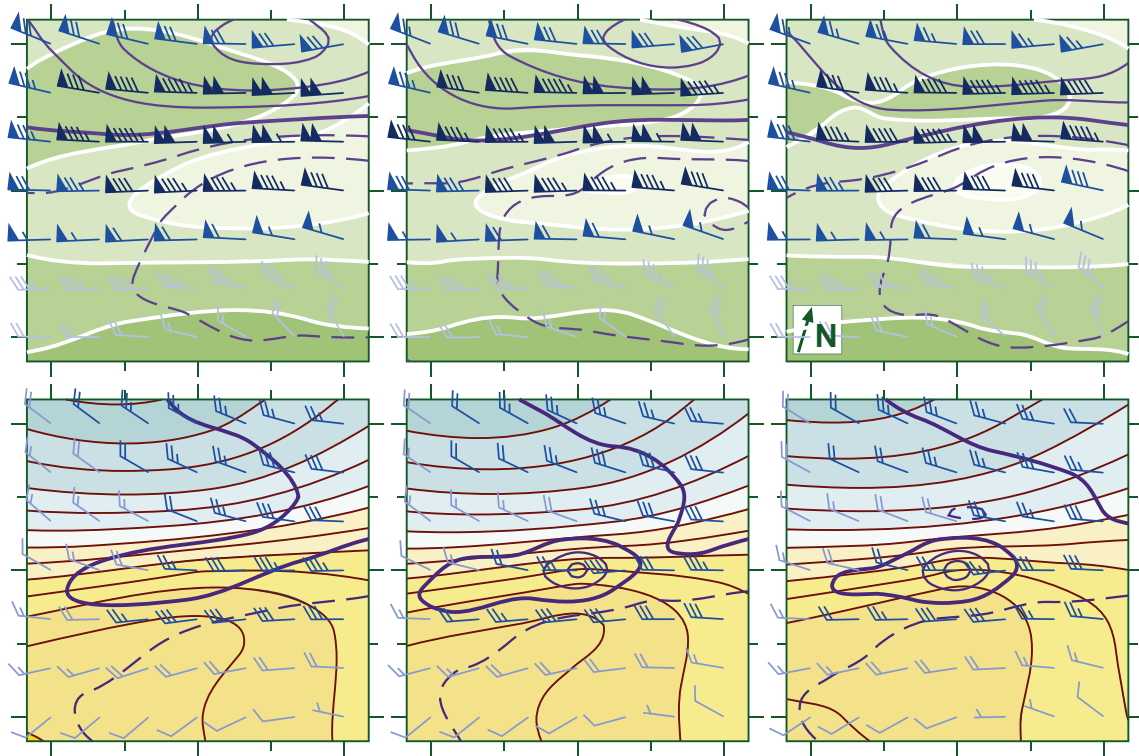


Figure 17 One of the initiation Class W composite at times (from left to right) -6 h, 0 h (first detection time of an individual vorticity maximum), and $+6$ h, at 300 hPa (top) and at 850 hPa (bottom). The fields on the top row are wind vectors (blue barbs, kt) colored according to speed so as to highlight the jet-flow structure (the darkest blue above 40 m s^{-1}) and relative vorticity ζ_r (light purple lines, contour interval $2 \times 10^{-5} \text{ s}^{-1}$, negative contours dashed). The green-shaded background field is the 600 mbar relative humidity (contour interval 10% , white for values larger than 90% , color darker for smaller values). Bottom fields are relative vorticity (as on the top, with darker contours) and equivalent potential temperature θ_e (brick-red thin contours and shading, interval 3 K , limit between blue-green shade and light orange 300 K). The 850 mbar wind is also shown as above, the darkest blue threshold is changed to 10 m s^{-1} . The boxed arrow is a compass indicating the average direction of the geographical north, since the area shown is rotated. Adapted from Ayrault, F., Joly, A., 2000. Une nouvelle typologie des dépressions météorologiques: classification des phases de maturation. *Comptes Rendus de l'Académie des Sciences Paris, Sciences de la Terre et des planètes* 330, 173–178.

paradigm. A single picture summary of each family is shown in [Figure 18](#).

Summary of Development Situation

The maturation phase composites show seven classes. [Figure 19](#) shows some of the morphological characteristics of Class A, the one that undergoes the largest deepening ($-25 \text{ hPa}/24 \text{ h}$) and happens in 12% of the cases: it is what some authors have termed the meteorological ‘bomb.’ Centered on the low-level component of the system, the figure clearly depicts the phase change between this and the upper-level component: as there are independent studies of these strong cases, this class validates the approach. Note, in particular, the similarity between this evolution that combines together about a 100 cases and the growth phase of Low M in [Figure 12](#).

Class B, although not deepening, remains in warm air and yields significant precipitations. An interesting characteristic of this type is that it occurs in rather strong baroclinic zones, like Class A and with the same frequency. Yet, it does not develop strongly ([Figure 20](#)). The magnitude of the baroclinity does not explain the two different behaviors, and this questions explanations in terms of simple instabilities. Class B does not have the upstream upper-level signature that Class A shows well.

The other types correspond to cyclones deepening between 9 and 15 hPa in 24 h . These classes show, in common with Class A, the presence of a well-defined upper-level signature. One feature that discriminates these types is the shape of the upper-level jet and the position of the cyclone relative to it ([Figure 21](#)). Using time filtering to separate the cyclones and their environment, it is possible to estimate the energy transfers. The preferred energy conversions are similar for all the developing classes and correspond to the early twentieth century idea that cyclones develop by converting the energy contained in horizontal gradients of temperature into wind. This is called baroclinic interaction. In that sense, the core physical mechanism for growth is the same for all these types.

The method employed yields so far two sets of composites. A fundamental result comes from the clear separation between these two sets: the subsequent amplification of a cyclone does not necessarily involve the same processes as its genesis. Indeed, while baroclinic interaction is the only significant process leading to large amplitudes, it is the dominant mechanism in 20% of the initiations at best.

Altogether, case studies such as the well-documented FASTEX cases and many others and the climatological composites enabled by the reanalyses indicate (1) that there are distinctive structural features between rapid deepeners or ‘bombs’ on the

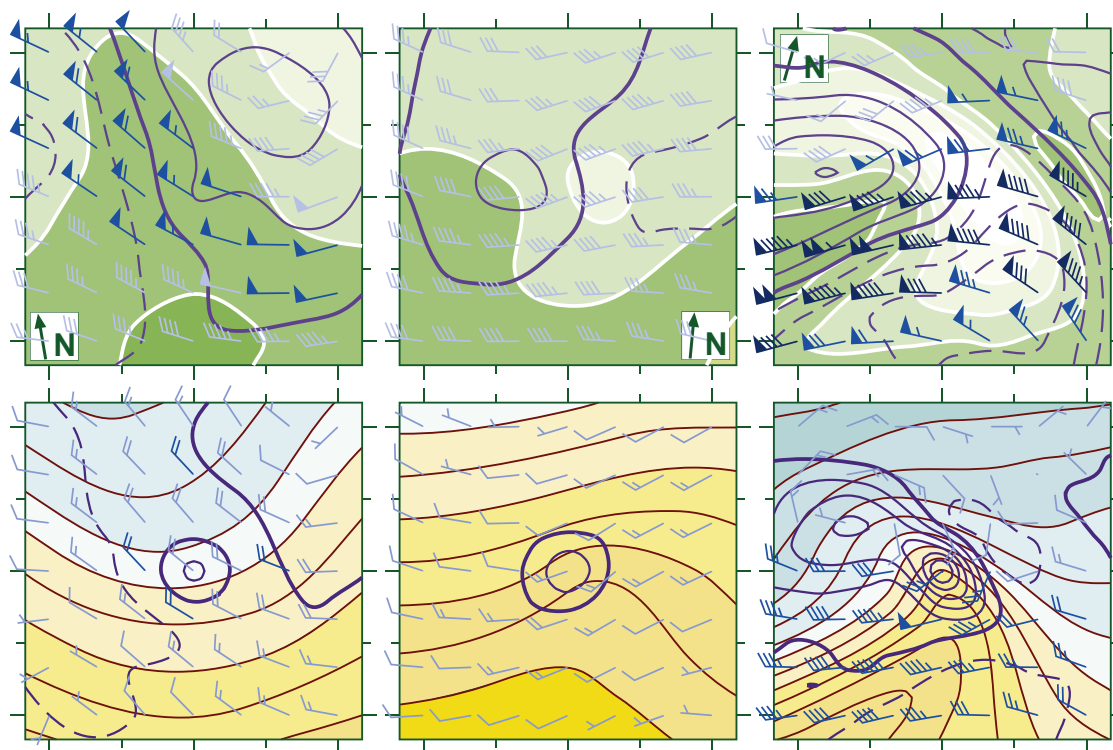


Figure 18 A single-time summary of three of the other families of initiation composites. The structures are shown at the time of the first detection of an individual vorticity maximum, that is the appearance of a new cyclone. From left to right: a class of cold air cyclogenesis; cyclogenesis induced by a preexisting upper-level precursor; and a class of cyclogenesis resulting from the apparent aft-splitting of an older cyclone. The fields are displayed as in [Figure 17](#). Adapted from Ayraut, F., Joly, A., 2000. Une nouvelle typologie des dépressions météorologiques: classification des phases de maturation. *Comptes Rendus de l'Académie des Sciences Paris, Sciences de la Terre et des planètes* 330, 173–178.

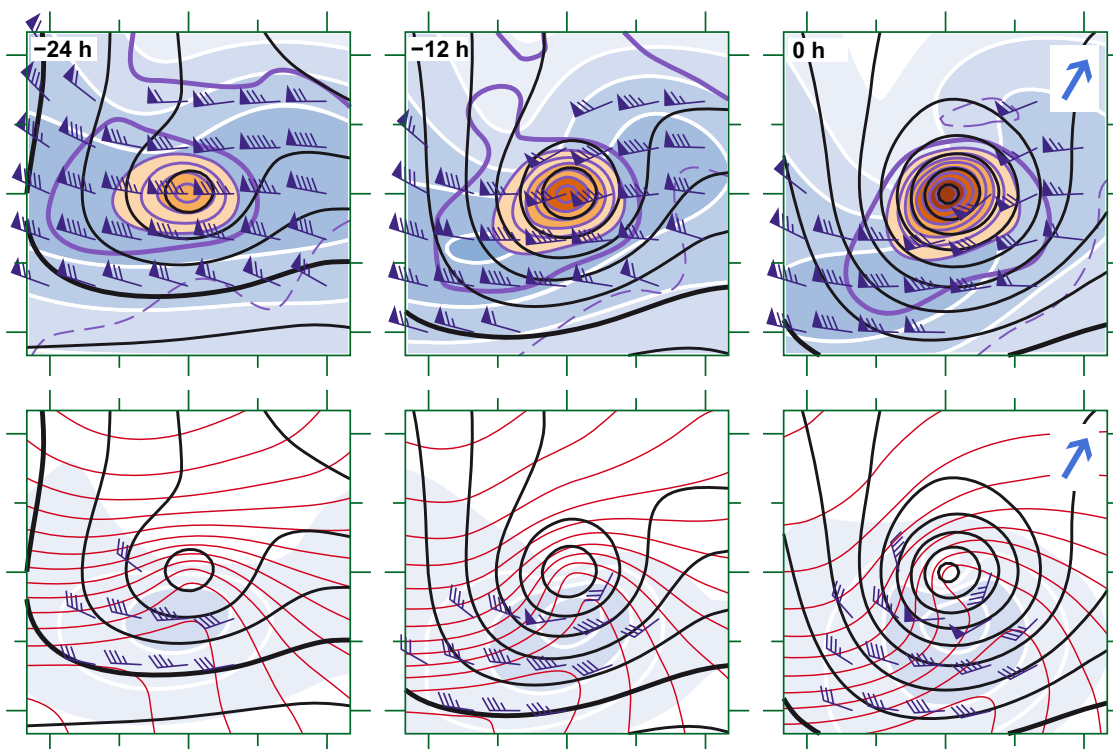


Figure 19 Example of one of the types of composite mature extratropical cyclone, the most intense ones (Class A). The composite structure is shown at times -24 h, -12 h and at the time of maximum amplitude (from left to right), at 300 mbar (top) and at 850 mbar (bottom). The fields are mean sea-level pressure (black contours, interval 5 hPa, thicker reference 1015 hPa), 850 hPa moist potential temperature (red contours, interval 3 K), 850 hPa relative vorticity (purple contours and orange shading, interval $2 \times 10^{-5} \text{ s}^{-1}$, negative contours dashed), as well as in the top panels the 300 hPa wind (barbs) and wind speed (white contours, blue shading, interval 10 m s^{-1}) and, similarly represented, the 850 hPa wind and wind speed in the bottom panels. Derived from Ayraut, F., Joly, A., 2000. Une nouvelle typologie des dépressions météorologiques: classification des phases de maturation. *Comptes Rendus de l'Académie des Sciences Paris, Sciences de la Terre et des planètes* 330, 167–172.

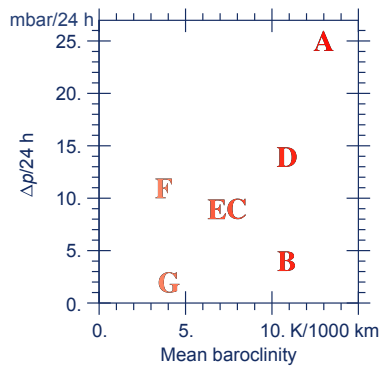


Figure 20 The mean 24-h pressure deepening $\Delta p/24 \text{ h}$ of the seven classes of maturation phases are plotted against the low-pass average baroclinity, the temperature gradient of their environment. From Ayrault, F., Joly, A., 2000. Une nouvelle typologie des dépressions météorologiques: classification des phases de maturation. *Comptes Rendus de l'Académie des Sciences Paris, Sciences de la Terre et des planètes* 330, 167–172.

entrance side of a stormtrack and on the exit side and (2) that the maturation of a cyclone definitely involves finite-amplitude interactions between low level and upper level. The basic mechanism, that can be called baroclinic interaction, is understood in broad terms (see below), but many of its critical properties (existence of amplitude thresholds, critical distances, and the likes) are not yet known.

Other Types of Cyclones

So far, extratropical cyclones that are closely associated with a baroclinic area or jet flow have been discussed. The baroclinic area provides their energy and leads to a single broad mechanism for development, with several variants. They are the most common ones. However, there exist other kinds of extratropical

cyclones that depend on profoundly different mechanisms. Three examples deserve to be mentioned.

One kind is termed polar lows. As their name suggest, they are cyclones that are observed at high latitudes, in cold air. They appear to start as quite small-scale systems (compared to the more standard 'baroclinic' systems) but they are able to develop very strong low-level winds. The mechanism that makes polar lows cousin to tropical cyclones is they are dependent on air–sea interactions. While tropical cyclones subsist on latent heat fluxes (or the evaporation of water from the sea), polar lows depend on the temperature difference between very cold air and a relatively warm sea. Some polar lows grow in scale and may turn into more classical weather systems interacting with some upper-level flow. It must be noted that systems of that kind, more symmetrical than standard extratropical cyclones, sometimes with an eye, have been observed in the Mediterranean Sea, although they cannot be termed either tropical or polar.

Another kind, also well represented in the Mediterranean area, is lee cyclones. These cyclones form in the lee of a large-scale orographic barrier, such as the Alps. The cyclogenesis is triggered by the interaction of a preexisting weather feature (such as a front or a trough) with the orography. The global view given by the reanalysis shows that most large mountainous barriers induce a cyclone initiation area in their lee.

A final kind of low is that associated with a high-pressure center located to its poleward side. The low and high are coupled: they form what is called a block. Blocks are persisting large-scale structures; their lifetime is longer than that of baroclinic cyclones and they are larger than the baroclinic cyclones. Their vertical structure is also simpler. Blocks form at the end of stormtracks and persist by deriving their energy from the baroclinic activity within the stormtrack upstream of them. The lows that are part of a block shows bursts of activity, with transient precipitating features that tend to rotate around the

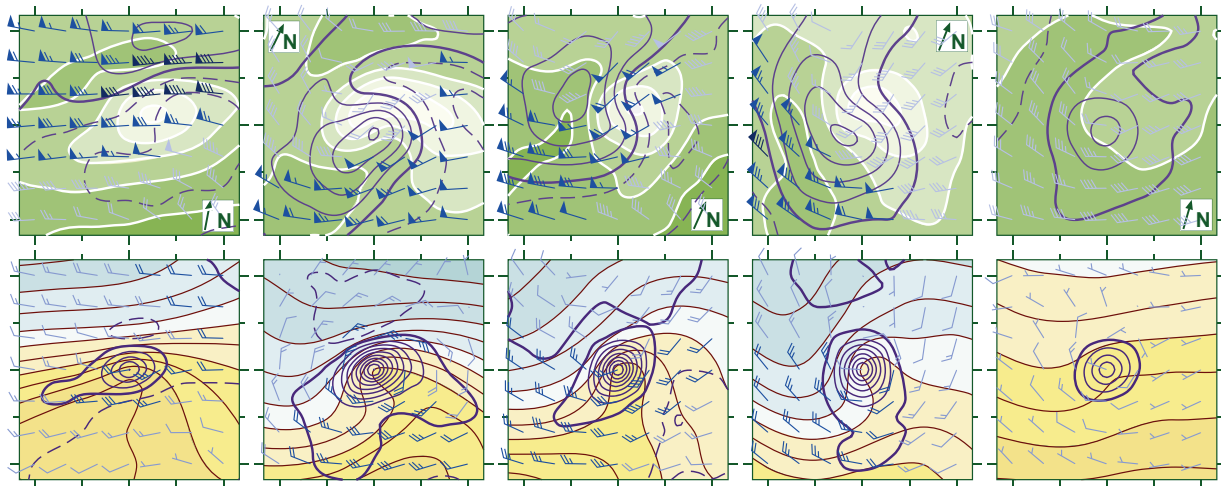


Figure 21 A single-time summary of the other classes of mature stage composites. Structures are shown at the time of maximum amplitude. From left to right: the weak wave class (Class B); the class of cyclone maturing between two upper-level jet streaks (Class C); the class of cyclones developing when reaching a diffluent zone while coming from the warm side (Class D); an example of a class of cyclones linked to a marked upper-level trough (Class E); and the most populated class of nondeveloping systems (Class G). The fields are displayed as in Figure 17 except for the wind shown above 50 kt. Adapted from Ayrault, F., Joly, A., 2000. Une nouvelle typologie des dépressions météorologiques: classification des phases de maturation. *Comptes Rendus de l'Académie des Sciences Paris, Sciences de la Terre et des planètes* 330, 167–172.

low center (quite unlike the fronts of a baroclinic system which remain fixed relative to the center).

Previous Recent Conceptualizations of Extratropical Cyclones

The evolution of the understanding of extratropical cyclones is tied to the progress of two areas: better observations on the one hand, and on the other the introduction into meteorology of new mathematical tools for physics, allowing assumptions and their consequences to be quantified and checked. From that point of view, early attempts using concepts like the thermal cyclone or the polar front cyclone, although seminal in some respects, have not been very successful.

Upper-Level Induced Development Model

The emergence of operational meteorology accompanying the development of aviation led to better coordinated observational networks. The surface measurements benefited from the first ships able to hold a fixed location in the ocean. Most importantly, upper-air observations over continents became more frequent; they used sondes attached to balloons. Radiosondes transmitting their data while ascending were introduced, removing the need to recover the recorded measurements after the balloon burst and the sonde fell in some unlikely place. In 1935, a 'coordinated ascents' experiment on the scale of western and northern Europe took place. For the first time, a few mature cyclones were sampled in three dimensions.

The upper-level structure of the flow about extratropical cyclones was then quite clearly obtained. The wind increased with height much more rapidly near the cyclone center than away from it. A very clear vertical tilt with height of the minimum pressure location was definitely shown to exist. It took some time to make something of the wind distribution: ultimately, and with much further observations, it was recognized to be organized in meridionally localized streams of air, called jet streams. But the importance of the vertical tilt and the related upper-level mass distribution was grasped very quickly, partly because it had been anticipated both with observations and theoretically on several occasions before. Several approaches have since been developed. One is called 'development.' This had been outlined shortly before the 1940s by the British scientist Sutcliffe, who proposed that the emergence and amplification of a surface cyclone resulted from the influence of a preexisting upper-level trough in a broad baroclinic zone (rather than an extreme front). This view is supported by simplifying fundamental equations, rather than linearizing them, and concentrating, in particular, on vorticity dynamics. In problems where rotation is important, looking at vorticity is a good tactic.

Along this line of thought, the German meteorologist Kleinschmidt put some flesh on the concept of the upper-level component of an extratropical cyclone and its relation to vorticity. Using the homogeneous and extended network of radiosoundings available over Europe in 1943, he built a striking 3D picture of this structure. Furthermore, he gave the description in terms of a quantity akin to vorticity and called

potential vorticity. He particularly undertook to show that, under certain assumptions, it is possible to derive the wind and temperature distributions from that of potential vorticity, a process called inversion.

Sutcliffe and Kleinschmidt are, together with the Swede Petterssen and a few others, representatives of a school of thought that made the most of available observations to conclude that cyclones are largely caused by finite-amplitude interactions of some preexisting 'structures': vertically confined vortices, baroclinic zones or jet streams, fronts, and the like. However, until recently, they represented a minority among those attempting to understand cyclones.

Upper-Level Jet, Baroclinic Zone, and Baroclinic Instability

The school of Sutcliffe, Kleinschmidt, Petterssen, and others was in a minority because an elegant mathematical framework was set up in parallel and attracted most of the attention. Discarding the unrealistic extreme front model, Charney in California and Eady in England applied, in the early 1940s, the linear stability technique to the study of the properties of a simple broad baroclinic zone. This is an area of continuous and moderate transition from warm to cold air, which, as a consequence of rotation, is accompanied by a zonal wind increasing with height – an idealized form of the upper-level jet stream. This basic or initial flow indeed fits the observed average flow at midlatitudes much more reasonably than the extreme front model. This problem can be idealized to a point that allows analytical solutions to be derived that still share some properties with real cyclones. Most notably, the vertical tilt with height is part of the most unstable solution, together with reasonable spatial scales. This is the baroclinic instability theory of cyclogenesis.

These similarities made linear studies the preferred path to explaining natural phenomenon like the growth of cyclones. Most textbooks on dynamical meteorology depend heavily on linearized equations and stability problems. Very elegant general existence theorems have been derived.

Because it is so elegant and appealing to mathematically oriented minds, the instability model dominated the theoretical scene throughout the 1950s and 1960s. But because part of its elegance comes from mathematical tractability, only extremely idealized problems can be studied and the relationship with observations is the weak spot of all of this work. As a consequence (or measure) of this, the results have had little impact on practical forecasting aside from some guidelines for interpreting numerical model output.

Quasi-Geostrophic Model and Theoretical Results

In the early version of the instability approach, subtle simplifications were introduced at key points in the derivation, limiting the number of solutions. This practice is in fact very powerful as far as gaining an understanding is concerned. It was soon reconsidered more formally by Charney in terms of a scale analysis of the basic equations; this is called filtering on the basis of a balance assumption. A simplified set of equations specifically oriented toward understanding extratropical cyclones and their large-scale environment can be derived at the outset. This is the quasi-geostrophic model.

These equations represent cyclones and some of their behavior, but they exclude, for example, gravity waves. They include well-defined cause-and-effect relationships. In this framework, for example, vertical motion is the computable effect of the interaction between the thermal and horizontal wind fields.

It is important to realize that this model is the common ground of both the intrinsic instability approach and the finite-amplitude interactions approach. With hindsight, it is easy to show that Sutcliffe's calculations are based on an early form of the quasi-geostrophic filtering and so are Eady's. Potential vorticity inversion, as outlined by Kleinschmidt, is a well-posed classical Poisson problem in this framework. The current theoretical understanding of cyclones undoubtedly rests on the quasi-geostrophic framework.

The case study and most of all, the study of reanalyses introduces the relevance of the baroclinic zone with a jet-stream flow as the environment of extratropical cyclones. It has also shown the diversity of types and life cycles only part of which are understood.

The Mechanisms of Cyclone Development

This section aims at providing a simple dynamical understanding of the main mechanism behind the development of most extratropical cyclones called baroclinic interaction and continues to discuss various ways it can take in the atmosphere.

Thermal Wind Balance

The various observational studies mentioned above allow retrieval of a number of properties of extratropical cyclones. Thus, on the scale of such systems, the vertical profile of temperature barely departs from that given by assuming that the atmosphere is in hydrostatic equilibrium. The horizontal wind is also quite close to the wind deduced from the distribution of mass assuming geostrophic equilibrium. The hydrostatic equilibrium relates the vertical distribution of mass to temperature; the geostrophic equilibrium relates the horizontal distribution of mass to the wind. Combining the two means that a cyclone evolves in a state that preserves some kind of equilibrium between wind and temperature, called thermal wind balance (Figure 22). The evolution of an atmosphere remaining close to thermal wind balance is described by the quasi-geostrophic system of equations. With a proper choice of vertical coordinate z (pressure rescaled as a height), this can be expressed as:

$$f \frac{\partial u}{\partial z} \sim -\frac{g}{\theta_0} \frac{\partial \theta}{\partial y}, \quad f \frac{\partial v}{\partial z} \sim \frac{g}{\theta_0} \frac{\partial \theta}{\partial x}, \quad [1]$$

where x is oriented eastward, y is oriented northward; g/θ_0 is related to the scale height of the atmosphere $H_s = R_a \theta_0 / g$, with R_a the perfect gas constant for air; u is the zonal wind, v the meridional wind; and θ is the potential temperature, a temperature conserved in adiabatic displacements. The wind and temperature fields are therefore not independent: they can be described in terms of just one field, either the geopotential ϕ or the quasi-geostrophic quasi-potential vorticity q_g . The latter is conserved by geostrophic displacements, whereas the

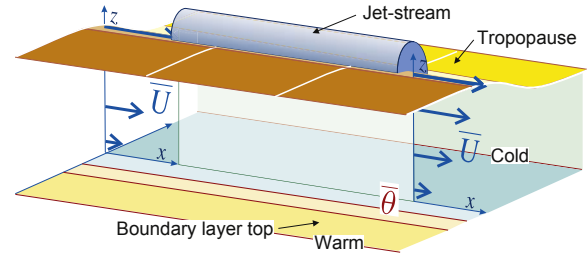


Figure 22 Schematic showing an idealized baroclinic zone, the environment within which extratropical cyclones develop. It is an area of reinforced but continuous horizontal thermal gradient in thermal wind balance with an upper-level jet flow. In the simplest case, the baroclinic zone is assumed to be homogeneous in the zonal (x) direction, so the scheme is isolating a segment. Adapted from Malardel, S., Joly, A., 2000. EAO Anasyg/Presyg. Météo-France, Toulouse.

former is not. Both are related to one another by a Poisson-like equation:

$$q_g = \nabla_{qg}^2(\phi), \quad [2]$$

where ∇_{qg}^2 is a scaled Laplace operator.

The wind field, in particular, must be separated into two components: a geostrophic one ($u_g, v_g, 0$) such that:

$$u_g = -\frac{1}{f} \frac{\partial \phi}{\partial y}, \quad v_g = \frac{1}{f} \frac{\partial \phi}{\partial x}, \quad \text{with also } \theta = \frac{\theta_0}{g} \frac{\partial \phi}{\partial z}. \quad [3]$$

The other is a departure from geostrophy (u_{ag}, v_{ag}, w). This includes in particular the vertical motion w , which is directly responsible for the formation of clouds and rain.

Thus, the wind is not an independent variable, and in particular, the nongeostrophic component is a nontrivial function of the geostrophic wind and temperature. The physical idea is that the nongeostrophic component, which is zero when the thermal wind equilibrium is strict, appears as a response to geostrophic tendencies that, alone, would lead the system away from equilibrium. The sense of this response is to restore a state close to equilibrium. The nongeostrophic motion is caused by interactions between the geostrophic wind and the temperature that tend to break the thermal wind balance and is such that it has effects that oppose the cause that originally led to it (this is called negative feedback).

This can be written formally in terms of an equation for the vertical motion. It takes the form:

$$\nabla_{wqg}^2(w) = \mathcal{F}(V_g, \theta), \quad [4]$$

where ∇_{wqg}^2 is yet another scaled Laplace operator and $\mathcal{F}(V_g, \theta)$ describes how the geostrophic component of the flow forces vertical motion. The latter term is nonzero whenever the geostrophic wind tends to displace isotherms horizontally, or when isolines of the geostrophic stream function (close to ϕ) cross isotherms. Assuming zero vertical motion at the surface and the tropopause yields:

$$\frac{1}{\mathcal{V}} \int_{\mathcal{V}} w \mathcal{F} d\mathcal{V} < 0, \quad [5]$$

that provides a broad idea of the sense of the vertical motion, negatively correlated with the forcing $\mathcal{F}$ within an arbitrary volume.

Baroclinic Interaction

The vertical motion w is not only important because it is the direct cause of the formation of clouds in ascent zones. It is a key dynamical parameter. In particular, it controls the evolution of vorticity. The simplest form of the quasi-geostrophic equation of evolution of geostrophic relative vorticity $\zeta_g = \partial v_g / \partial x - \partial u_g / \partial y$ reads:

$$\mathcal{D}_g \zeta_g = f \frac{\partial w}{\partial z}, \quad [6]$$

where $\mathcal{D}_g = \partial / \partial t + u_g \partial / \partial x + v_g \partial / \partial y$ is the geostrophic material derivative or the change of ζ_g following the geostrophic motion. This equation shows that the only remaining source of vorticity changes is the term called stretching. When the vertical velocity w increases with height, vortex tubes are stretched and this increases their vorticity.

At any given time, the extratropical troposphere shows a planetary-scale wavy pattern with a broad area of increased meridional gradients (the main baroclinic zone), together with a number of vortex-like features of various scales. Some of these vortex structures are like filaments coming out of the planetary polar vortex, others are cut off from this and look like isolated or sometimes coupled maxima or minima drifting in the vorticity field.

Consider therefore the following idealized situation: a single vorticity maxima (say) trapped near the tropopause close to a homogeneous baroclinic zone (Figure 23). Because of thermal wind balance, the latter can be seen either as a deep area of reinforced meridional temperature contrast or as where the upper-level jet stream is. The vorticity maximum and the baroclinic zone interact with each other in the sense that the stream function signature of the vorticity maximum crosses the isotherms of the baroclinic zone (and vice versa). This interaction leads to some response in the nongeostrophic wind field; in particular, it causes vertical motion. The vertical motion is strongest where the geostrophic cross-isotherm flow is the strongest (to first order), namely in two locations: just downstream (eastward) and just upstream (westward) of the vorticity maximum in the flow oriented by the jet stream. The rule of thumb is that the vertical motion opposes the effects of the geostrophic wind on the isotherms. On the downstream side of the vorticity maximum in the troposphere, the cyclonic flow leads to warming and cooling on the upstream side. This is opposed by tropospheric ascent downstream and descent upstream.

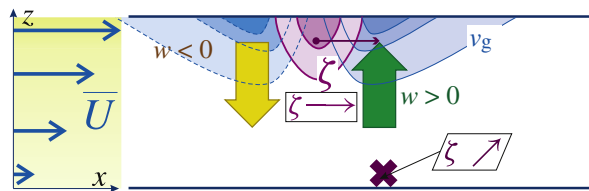


Figure 23 Vertical cross section in the same framework as in Figure 22 showing some aspects of an upper-level vorticity anomaly in the jet flow. Such a structure propagates but does not develop or, at least, not fast. However, it is a simple model of upper-level precursor. Adapted from Malardel, S., Joly, A., 2000. EAO Anasyg/Presyg. Météo-France, Toulouse.

Consider now the effect of this vertical velocity field on the vorticity maximum. There is stretching on the westward side, nothing at the level of the maximum and squeezing on the eastward side. This means that the vorticity maximum will propagate upstream (toward the west), in other words it will move more slowly than the wind in the jet, but will not amplify since no vorticity is created at its core. Here, we have a simple model of a structure that can persist for some time, move relative to main flow but will remain barely visible in terms of 'weather.' This is one simple example of a precursor structure. Away from the surface, as assumed here, it is little affected by friction. However, this kind of vorticity monopole is generally dispersive, which sets a limit to its 'useful' lifetime.

Consider now the effect of that vertical velocity field on the other boundary, in this case, near the surface. The ascent is a source of stretching for surface-based vortex tubes. So near-surface vorticity is generated below the ascent, slightly downstream of the upper-level precursor. This is the source of cyclone development contemplated by Sutcliffe and Petterssen.

Similarly, an isolated surface-based vortex is accompanied by the same kind of vertical velocity field as an upper-level one when interacting with a baroclinic zone (Figure 24). This vortex will propagate downstream (toward the east) but, similarly, it will not develop. Not only these low-level precursors are also dispersive, but also they are affected by friction. Nonetheless, they may exist for short periods. They also have an effect on the vorticity dynamics on the other boundary, the tropopause.

Rapid mutual development of such vortices in a baroclinic zone takes place when they are located relative to one another in a way such that their associated vertical velocity zones stretch the vorticity maximum at the other boundary. In this case, both vortices amplify each other: this is the essence of baroclinic interaction. This happens as long as the low-level vortex is located downstream of the upper-level one (Figure 25). If the

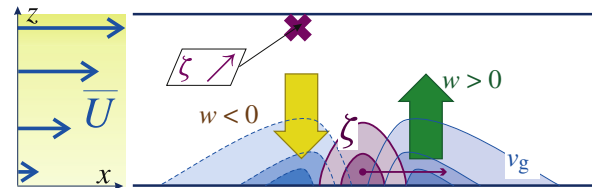


Figure 24 Same as Figure 23, but for a low-level vorticity anomaly. Its behavior is the same as the upper-level anomaly. Adapted from Malardel, S., Joly, A., 2000. EAO Anasyg/Presyg. Météo-France, Toulouse.

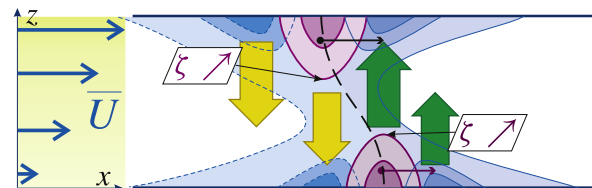


Figure 25 Constructive baroclinic interaction takes place when two structures like those shown in Figures 23 and 24 influence each other as shown in this cross-section, amplifying each other. Adapted from Malardel, S., Joly, A., 2000. EAO Anasyg/Presyg. Météo-France, Toulouse.

low-level vortex is found upstream of the upper-level one, then the interaction is destructive and the composite system collapses rather than amplifies.

This shows how the ideas of balance lead to a characteristic constraint on the vertical structure of a developing cyclone. The baroclinic zone provides the thermal energy available for conversion into cyclone wind. The low-level vorticity signature of the cyclone must be located downstream of the upper-level part, as shown in the climatological composites, in order for a balanced circulation to amplify as this energy is converted.

From this elementary description of the mechanism of baroclinic interaction, one can spot a number of critical parameters. The process depends on the influence of a vorticity structure on the other horizontal boundary; the stronger this influence, the more efficient the process. This is clearly related to the vertical scale H_ζ of the initial structures and to the way the medium (the baroclinic zone) opposes or aids this influence. Concerning the first point, taking into account the fact that potential vorticity is quasi uniform in the troposphere makes the vertical scale a simple function of the horizontal one L_ζ with $H_\zeta \sim L_\zeta \times f/N$. Here, N is the vertical static stability of the troposphere. This vertical scale must be of the same order as the depth of the troposphere H . In short, only anomalies with a horizontal scale larger than some threshold may amplify efficiently through baroclinic interaction. This threshold is of the order of the Rossby radius of deformation $L_R = HN/f$. Below this scale, they can only propagate (or disperse). The critical property of the medium is its potential vorticity, which, at the level of quasi-geostrophic theory, is mostly related to the static stability. The smaller the potential vorticity is, the easier the interaction is. This is included in the definition of the Rossby radius.

The other parameters include the vertical phase between the vorticity components. For systems that are locked in phase, there is an optimum phase tilt that allows for the fastest sustained growth over a long time. However, the composites do not show vertical phase locking and the problem must be seen differently. The climatological study reveals that the observed timescale of development is about 24 h: it can then be shown that systems that are not phase locked (with a tilt that changes with time) amplify significantly more rapidly over a wider range of scales during this timescale. A further parameter is the initial distance between the structures, when two potential precursors exist at both boundaries. Below some value, a single powerful cyclone will result from the interaction. Above some value, each precursor will tend to generate its cyclone, but as this process is building up slowly, one will notice two small amplitude systems. The duration of the interactions is another parameter. Finally, the nature of the underlying surface is also important in determining when an amplifying interaction may start: it has to create more kinetic energy than is dissipated by friction. Although the finite-amplitude maturation of an extratropical cyclone may take place over land (unlike a tropical cyclone), its early growth is easier over sea because of the reduced friction.

The Two Sources of Cyclone Growth in a Balanced Flow

The preceding paragraphs discuss how the shape of the composite system temporarily or permanently made up of two vorticity maxima at low level and upper level enables it to grow

from tapping energy from the existence of a horizontal temperature gradient in its environment. The critical aspect of the shape is that it must be elongated against the thermal wind that accompanies, in a rotating frame, the horizontal temperature gradient.

It is useful for future reference to provide some formalism to this type of growth through balanced interactions. The approach is to write how the energy of the perturbation cyclone can grow. At the level of the 'dry' adiabatic quasi-geostrophic theory, there are only two sources that can feed the growth of a perturbation in a quasi nondivergent environmental flow. One is the baroclinic interaction just explained. In a more general situation, the baroclinic interaction as an energy source term can be written as $S_{bc} = \mathbf{F}' \cdot \mathbf{\bar{B}}$ with:

$$\mathbf{F}' = \left(\frac{g}{\theta_0} \right) \left(\frac{v'_g \theta'}{N}, -\frac{u'_g \theta'}{N} \right), \quad [7]$$

$$\mathbf{\bar{B}} = \left(\frac{g}{\theta_0} \right) \left(-\frac{\partial \bar{\theta}}{\partial y}, \frac{\partial \bar{\theta}}{\partial x} \right) = \left(\frac{f \partial \bar{u}_g / \partial z}{N}, \frac{f \partial \bar{v}_g / \partial z}{N} \right). \quad [8]$$

In these equations, the properties of the environment are indicated with an overbar $\bar{\cdot}$, those of the cyclone or any other structure embedded in it are indicated with $\cdot'$. This implies that a given situation has been split into at least two parts, but it does not imply anything about amplitudes. The $\mathbf{\bar{B}}$ vector represents the baroclinity of the environment as a vector aligned with the thermal wind, that is oriented along the jet-stream flow. The vector $\mathbf{F}'$ represents on first inspection the heat fluxes associated to the perturbation. However, in a balanced flow, both u'_g and θ' depend on the perturbation geopotential ϕ' . Using this property, these fluxes turn out to be related to the shape of the perturbation in each vertical plane. It is another way to recover the need for the westward vertical tilt to grow in a balanced zonal jet.

The other source of growth is called the barotropic interaction: in this case, the perturbation extracts its own kinetic energy from the kinetic energy of the environment. The barotropic interaction as an energy source reads $S_{bt} = \mathbf{E}' \cdot \mathbf{\bar{D}}_f$, where:

$$\mathbf{E}' = \left(\frac{1}{2} (v_g'^2 - u_g'^2), -u'_g v'_g \right), \quad [9]$$

$$\mathbf{\bar{D}}_f = \left(\frac{\partial \bar{u}_g}{\partial x} - \frac{\partial \bar{v}_g}{\partial y}, \frac{\partial \bar{v}_g}{\partial x} + \frac{\partial \bar{u}_g}{\partial y} \right). \quad [10]$$

The vector $\mathbf{\bar{D}}_f$ is the deformation field of the environment. The $\mathbf{E}'$ vector can be related to the horizontal anisotropy of the cyclone perturbation, that is to its shape in the horizontal plane. At the level of the quasi-geostrophic assumption, again the terms of $\mathbf{E}'$ depend only on the geopotential acting as a stream function. In this case, $\mathbf{E}'$ is related to the horizontal tilts of the cyclone perturbation. It can be shown that growth takes place in a simple situation when the perturbation tilts against the shear (Figure 26). In more general situations, several angles come into play, describing the shape of the perturbation or principal axes of deformation.

Of course, these same terms can also account for energy transfers from the perturbation to its environment. This is why they will be encountered again.

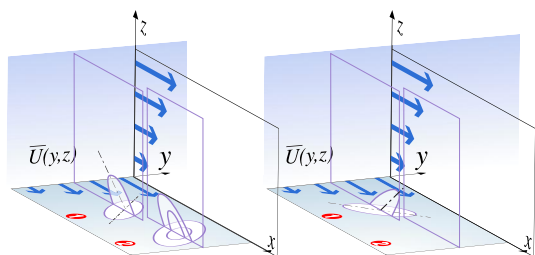


Figure 26 Schematic illustrating how the shape of a perturbation (purple contours and black dash-dotted axes) leads it to grow or to vanish in a situation where shear is present both horizontally and vertically. Perturbation structures are shown at two times denoted 1 and 2. The vertical shear implies the presence of a horizontal gradient of temperature. Left side: the perturbation grows from both the baroclinic (vertical tilt) and barotropic (horizontal tilt) interactions. Right side: the perturbation vanishes because it loses energy by both mechanisms.

Conditions Favoring Cyclone Development

Various aspects have been studied, either using linear models or by performing sensitivity experiments in nonlinear runs. A first kind of experiment is to change the physics of the model, such as by removing cloud processes. More recently, the initial conditions are also used to test ideas.

To summarize, it is found that low static stability, weak environment deformation, small horizontal shear in the jet stream, and latent heat release contribute to accelerating and strengthening the development of extratropical cyclones. The role of turbulent fluxes is more complex. The importance of low friction at the beginning of life cycles has already been stressed. Heat fluxes have contradictory effects within a growing cyclone. However, large heat fluxes ahead of or prior to the coming of a cyclone may reduce static stability by mixing and create a medium that will favor rapid growth; in other words, they reduce the Rossby radius L_R . This is called preconditioning.

Brief Review of Linear Stability Problems Relating to Cyclogenesis

Originally, the basic mechanism was called baroclinic instability. It was first derived by solving the linear stability problem that results from linearizing balanced equations with respect to a simple baroclinic flow. One solution is emphasized and somehow associated with a developing cyclone, the most unstable normal mode. The necessary vertical tilt is nicely predicted by this approach. However, the timescale is not properly handled. Introducing such unstable solutions into nonlinear models, while it allows the description of some of the structure changes, rather emphasize this too slow timescale problem. This raises enough doubts to lead to cast the problem differently, leading to look for more general solutions to obtain considerably larger amplification rates together with rapidly changing structures in the linear framework. In both these approaches, the view still persists that cyclones result from amplifying a barely measurable initial perturbation in a smooth large-scale baroclinic flow.

Other aspects that have been analyzed in terms of linear instabilities, either standard or generalized are slightly more complex basic flows such as two-dimensional fronts,

characterized by a large low-level relative vorticity maximum. Strong normal modes can be found when the vorticity maximum corresponds to a large amplitude, narrow potential vorticity anomaly, such as can be generated by latent heat release in a real frontogenetic process. More recently, the possibility of instabilities in time-dependent flows has been examined. The aim is to determine in particular the role of the background deformation field that may act on the basic flow, making a front stronger or creating areas of diffluence or confluence in a jet flow. The influence of the horizontal shear on the stability properties of a jet flow has also been examined in this way: a large horizontal shear makes the flow less unstable.

Potential Vorticity Attribution and Its Use in Theory Verification and Mechanism Finding

The framework outlined at the beginning of this section can be employed in other ways than finding the kinds of wavelike, possibly unstable solutions that it can support. This has brought out new possibilities for understanding less idealized flow configurations. Furthermore, these new approaches initially born from properties of the quasi-geostrophic representation of the synoptic scale atmosphere can be extended and practically implemented with far less constraining balance conditions. This has further enabled to start experimenting with realistic situations.

A first key idea is inversion of potential vorticity distribution. It is embodied in eqn [2], where the quasi-potential quasi-geostrophic vorticity q_g can be related to the geopotential ϕ following $\nabla_{qg}^2(\phi) = q_g(x, y, z)$. The situation can then be compactly summarized by the single distribution of potential vorticity $q_g(x, y, z)$ complete with boundary conditions. The vertical boundary conditions are important, they can take the form of the distribution of either the geopotential $\phi(z=0)$ at the surface and possibly at some upper boundary or the potential temperature $\theta(z=0) \propto \partial\phi/\partial z$ at the surface and at an upper boundary. From this small amount of information, the inversion of the operator ∇_{qg} yields the complete distribution of geopotential $\phi(x, y, z)$. From that, first the geostrophic components of the flow (u_g, v_g, θ) and then the ageostrophic ones, such as the vertical velocity w , can be recovered (using eqns [3] and [4]).

The next key idea is to see the distribution of potential vorticity (including the boundary conditions) as the sum of several individual parts, each attached to some kind of 'object' in the flow. Typically, there will be a large-scale part representing a baroclinic zone and a number of anomalies. Each of these parts should be represented by a piece of potential vorticity distribution δq_{gn} together with the corresponding boundary conditions, $\delta\theta_n(z=0)$, say. The situation is therefore now seen as the sum of these parts: $q_g(x, y, z) = \sum_N \delta q_{gn}$, similarly with the boundary conditions. The flow corresponding to one of these parts can be inverted from its own potential vorticity distribution δq_{gn} and its associated boundary conditions: this will provide anomalies of the geostrophic components ($\delta u_{gn}, \delta v_{gn}, \delta\theta_n$) and other fields. To the extent that the addition theorem applies, the total flow previously represented as a whole has then been decomposed into distinct anomalies, each related to a particular feature of the potential vorticity distribution including boundaries. The wind or the temperature at some place can now be seen as the sum of the anomalies

each attached to some aspect of the potential vorticity distribution: this is called attribution. In principle, the total flow is exactly the same as the one resulting from inverting the total potential vorticity distribution q_g . But having decomposed it into several consistent anomalies opens a completely new avenue: now, the dynamical interactions between the different parts of the flow can be computed and should provide a direct proof of, say, the explanation of the appearance of some new system. Conversely, a given potential vorticity structure and its boundary conditions can be removed from some situation, together with that part of the flow that has been attributed to it. Then, it is possible to see exactly what happens in the absence of this removed structure.

One application of these ideas is to linear instability theory verification in a real case. Linear theory has it that extratropical cyclones can be identified, at least initially to a linear perturbation that can be computed knowing the large-scale flow alone. The theory provides the structure the new cyclone should have as well as key properties such as a phase speed and a rate of growth: verifying the theory means comparing such predictions with the observed cyclone. In the real atmosphere, the flow is one, while in the theory, the flow is by construction split into the environment and the linear perturbation. To verify the theory, the real flow must be split as well, this is known as the separation problem. Potential vorticity attribution and inversion enables to solve that problem by working on a single field, potential vorticity, from which the other dynamically consistent anomalies of wind, temperature, and others can be deduced. This provides a representation of a real cyclone in perturbation form quite comparable to a predicted vector, it also gives a chance to model the large-scale flow without the event of interest that is required to compute the relevant linear perturbation.

Figure 27 is a result of actually using these ideas extended to a very realistic primitive equation model to the 26 December 1999 intense storm that hit western Europe. Two diabatic primitive equation solutions have been obtained by extracting a small structure at initial time, one that has the event and the other that has most of the characteristics of the exceptional baroclinic environment of that case, except the storm itself. It is then possible to employ the theoretical framework without further approximation and to compare the predicted unstable modes with the storm represented itself as a perturbation. Two aspects of the theory are especially studied. One is a comparison of the properties of the real and predicted systems, focusing on their structures. One aspect of this comparison is illustrated in the figure. The other deals with the idea that precursor structures, although very different from the theoretical modes, trigger the cyclogenesis by exciting these modes. It appears that the classical predictions (scales, etc.) of such a theory are, for most of them, far away from the observed properties. It is clear that the structure of a generalized unstable vector has little to share with that of a real cyclone. Yet, a weaker, slower storm does occur as a result of applying the theory to the stormless trajectory.

Aside from this attempt, the verification of the theory of baroclinic instability has always been performed with poor standards compared with, for example, the quantitative verification of models of the boundary layer fluxes and profiles and some other processes. This is to a large extent due to the difficulty of splitting the observed flow without introducing imbalances that will spoil any conclusion nor without introducing too much subjectiveness. The theory is successful only in very, very broad terms.

Linear results may not be readily associated to real cyclones. However, they are a useful model for describing how forecast

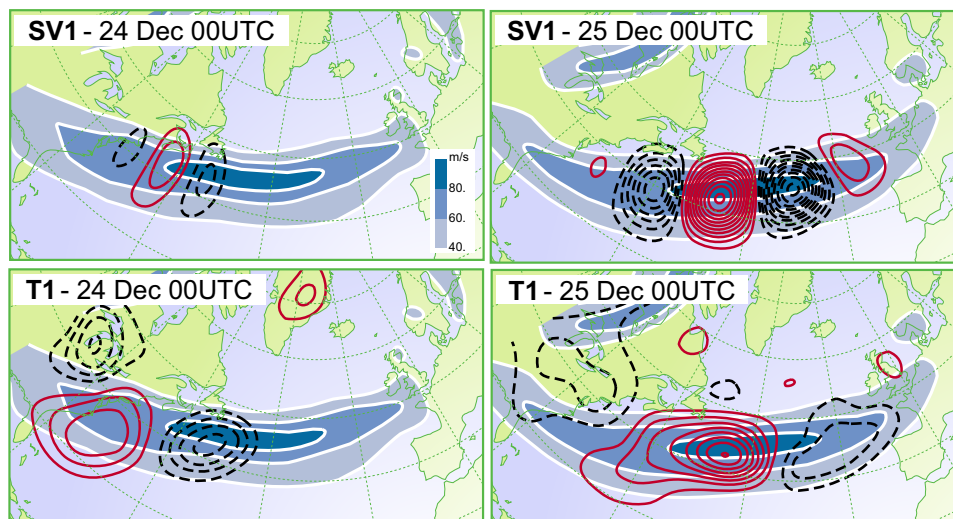


Figure 27 The use of potential vorticity inversion enables a comparison between a real case of extratropical cyclone extracted from an analysis, denoted T1, and the perturbation obtained from an instability calculation of the flow without T1, denoted SV1. The figure shows one aspect of the structure focusing on the horizontal plane. It shows the 850 hPa potential temperature perturbation θ' for SV1 at (a) 0000 UTC 24 December 1999 (contour interval is 0.01 K) and (b) 0000 UTC 25 December 1999 (contour interval is 0.02 K). 850 hPa potential temperature perturbation θ' for T1 at (c) 0000 UTC 24 December 1999 and (d) 0000 UTC 25 December 1999 – note enlarged contour interval of 1 K. Also shown is the wind speed on the 1.5 pvu surface of the trajectory (shading interval is 20 m s⁻¹ above 40 m s⁻¹). From Descamps, L., Ricard, D., Joly, A., Arbogast, Ph., 2007. Is a real cyclogenesis case explained by generalized linear baroclinic instability? *Journal of Atmospheric Sciences* 64, 4287–4308.

errors amplify during the short ranges (the first day or so) in a numerical representation of the atmosphere.

As shown by the diversity of composite cyclones, the initiation from an upper-air disturbance envisioned by Sutcliffe and Petterssen is far from being the only mechanism. Other has to be explicated. Another use of potential vorticity attribution and inversion is to prove the importance of some supposed interaction in a realistic circumstance. Recall the case study in the initial section: it has been assumed that the genesis of Low M involves some interaction between Low P and the large-scale jet flow confluence zone. In fact, this assumption is based on experiments performed on a rather similar case especially well observed during the FASTEX project, a situation which is also close to the initiation Class W of composite cyclone formation. The 'cause' preexisting low can be isolated and part of the flow attributed to it. The wind attributed to a preexisting low-level decaying system approaching the western North Atlantic confluence zone from

the northwest is explicitly shown in two panels of Figure 28. Note that an upper-level anomaly is similarly isolated and the corresponding flow also explicated: again, this is quite like the case study above. So there are three components, the flow attributed to the decaying low-level cyclone that attributed to the upper-level anomaly and the rest. The sum of these components is the analyzed flow, it mostly features the jet inflow. The baroclinic interaction of the upper-level structure weakly sustains the decaying low but otherwise, at the time the new system forms, it acts away from where it appears. The barotropic interaction previously introduced between the decaying low and the lower part of the jet inflow, on the other hand, does happen exactly where the new low appears. Simply removing that decaying low-level system suppresses the appearance of the new system. In the case study above, the total flow bears the signature of Low P, quite strong and not yet decaying. This intensity allows a barotropic interaction with the low-level large-scale jet inflow (represented by the

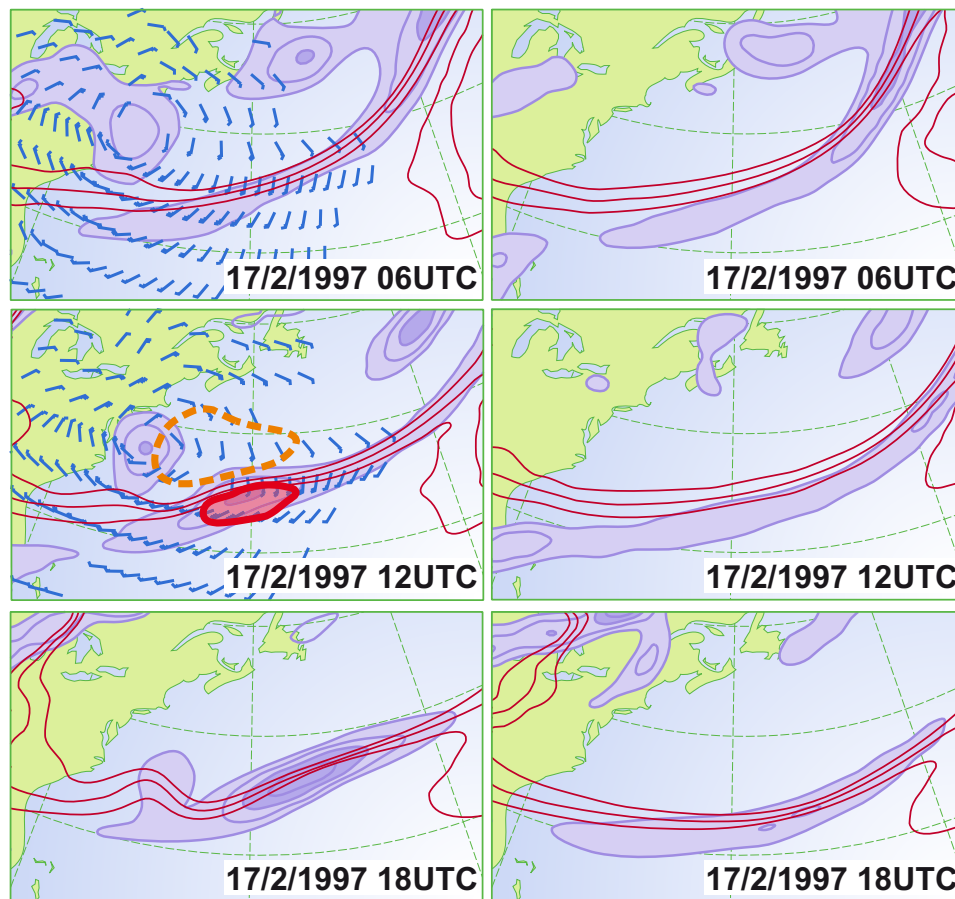


Figure 28 Demonstration of an extratropical cyclone initiation mechanism using initial conditions modifications with potential vorticity inversion. Each column of three panels is the time sequence of an experiment with Météo-France global model with modified initial states. The left column features a preexisting low-level cyclone, showing the wind field that can be attributed to it (blue barbs). This wind field interacts with the reinforced baroclinic zone accompanying the jet-stream confluence aloft, and this creates a new system. The interaction is materialized by the single red contour with transparent red shading, a maximum of positive barotropic energy conversion. The right column shows the same sequence in the absence of that preexisting decaying cyclone: no new low forms, in spite of the fact that in both cases, an upper-level structure reaches the confluent zone. Its interaction with the low-level flow is indicated by the orange contour, the baroclinic conversion term. The fields are 850 hPa relative vorticity (purple contours and shading) and selected isotherms of moist potential temperature (red contours). Adapted from Arbogast, Ph., 2004. Frontal wave development by interaction between a front and a cyclone: application to the FASTEX IOP 17. Quarterly Journal of the Royal Meteorological Society 130, 1675–1696, published by John Wiley & Sons Ltd.

iso- θ'_w) to be intense as well, which accounts for the extremely rapid initial growth of Low M.

This way, the understanding of real cases makes new ground, with the possibility to actually demonstrate growth mechanisms.

Possible Mechanisms for Rapid Growth in Jet Stream Diffuence Zones

The latter subsections have mostly considered mechanisms for initiation. The general idea of baroclinic interaction would suggest that, when it takes place, the larger the baroclinity, the most efficient it will be. But that is not what is observed. What is observed in the initial case of Low M, in the maturation phase of the composite cyclone A or in the ERA40 North Atlantic trajectory set, and from other sources as well is that growth takes place somewhat downstream of the largest baroclinity, furthermore in a large-scale diffuent environment where consequently baroclinity weakens. There is also the fact that part of the lows cross the jet axis during this growth phase in a diffuent zone.

This is the problem of understanding the localization of the baroclinic interaction in the real atmosphere. Some mechanisms have been brought out by revisiting the barotropic and baroclinic energy source terms presented earlier. These are well-known expressions, but they have many sides. The following results from interpreting those terms first as necessarily dominant ones when a balanced perturbation grows in a balanced environment and second, locally, without making any assumption on the representation of the perturbation, whether it is a wave or something else. Starting with the barotropic interaction in a weakly curved flow, the contribution of the environment is shown to depend on a key property called the effective deformation $\bar{\Delta}$. It is defined as:

$$\bar{\Delta} = \bar{D}_f^2 - \bar{\zeta}^2, \quad [11]$$

where $\bar{\zeta}$ is the vorticity of the environment and $\bar{D}_f$ is the magnitude of the deformation field of the environment.

Upper-level anomalies are being tracked along the poleward side of the jet stream, as seen for example in the case study. However, this is an area of large deformation. Applying basic reasoning, this should rapidly stretch any structure in this area to the point it cannot be seen any more. Yet, these also are regions of large vorticity due to the shear. The fact that it is the difference between these two properties that is dynamically relevant, both being large on both sides of the jet, explains why structures can be followed.

As seen above, the other key aspect of barotropic interaction is the horizontal shape of the perturbation, through the E' vector. So this subsection sets on the front scene the deformation as a key property of the environment and the shape of the perturbation interacting with it. The deformation is important in regions of diffuence, confluence, and shear. It plays an important role in the quasi-geostrophic framework in the forcing of vertical velocity. What is meant by shape is the existence of anisotropy or aspect ratio. Both the deformation and the shape can be characterized by directions of dilatation or elongation normal to directions of contraction or reduction: these special directions are the principal axes.

Positive values of $\bar{\Delta}$ correspond to regions where cyclones can be strongly stretched because of the deformation action, and where barotropic processes can be very important.

In regions of negative $\bar{\Delta}$, barotropic processes are quite weak. It is observed that cyclones tend to be trapped in areas of positive values of $\bar{\Delta}$.

The baroclinic interaction, on the other hand, is locally proportional to $\bar{B}F' \cos(\mathbf{F}' \bar{\mathbf{B}})$, where $\bar{B}$ and F' are the magnitudes of $\bar{\mathbf{B}}$ and $\mathbf{F}'$ given by eqns [8] and [7].

The idea is that a situation that leads to the shape of the cyclone perturbation to change or to change the way it is aligned with either the environment effective deformation or the environment baroclinity can more than compensate for a decreasing baroclinity and generate growth. Two mechanisms at least have been proposed along this line. Both ultimately lead to improve the baroclinic interaction. One does this by first allowing for a strong phase of transient barotropic growth that recreates a vertical tilt with height. It nicely explains the main growth phase of the FASTEX case the initiation of which has been explained in Figure 28. The other is summarized schematically in Figure 29. The presence of the diffuence zone reduces the horizontal distance between the upper and lower parts of the structure of the environment baroclinic zone (the slope of the iso- θ and other fields become more vertical). As the cyclone tends to cling to the lower baroclinic area, on the side where effective deformation is positive, it is also brought closer to the jet axis. This improves the alignment of the perturbation with the baroclinity, which explains the baroclinic growth. Such a configuration is called a baroclinic critical region. It seems to apply to the Low M case study.

The crossing of the jet stream by some of the cyclones has furthermore be the focus of recent idealized simulations. The main result is that this is a deeply nonlinear process. Nonlinear interaction terms seem to be more important for the crossing to take place than the presence of diffuence.

Upstream–Downstream Influence of the Flow on Cyclone Development

So far, the quasi-local influence of coherent anomalies in a baroclinic flow has been stressed. However, most of the structures considered are dispersive. This means that the energy converted when a baroclinic interaction takes place may propagate at a group velocity, which is different from the average phase speed of the structures involved. This is called the upstream–downstream influence on cyclone development.

For an idealized situation – a purely zonal flow, for example – consider an amplifying isolated cyclone. Upstream and downstream development takes the form of a train of anticyclones and cyclones that begin to form and develop both downstream and upstream of the original system.

In the real atmosphere, this means that the growth of a cyclone may not only be controlled by its local conditions; the presence of another system may be important as well. The easiest mechanism to diagnose is the downstream influence of a rapidly growing cyclone near the east coast of America (say) on a cyclone close to Europe. Examples of upstream influence are still to be found.

Cyclone Substructures and Air Streams

So far, the presentation focused on the cyclones as a whole and on their links with their larger scale environment. However,

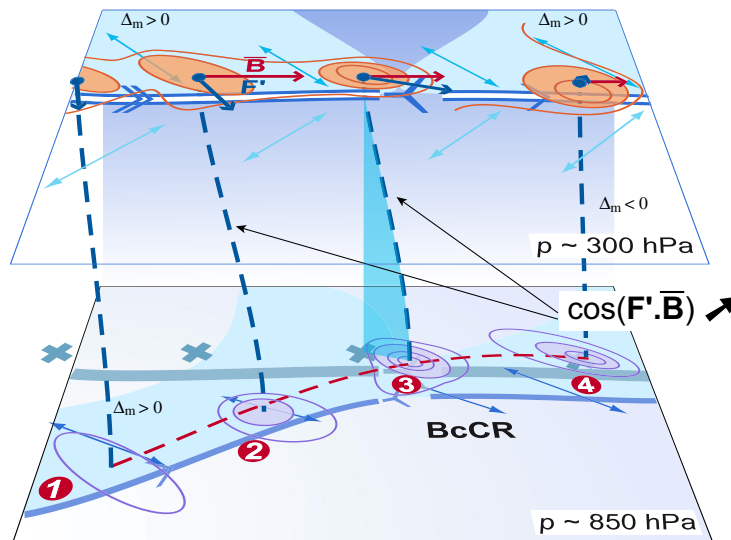


Figure 29 One possible mechanism accounting for the development of an extratropical cyclone in a region of diffluence (or exit) of the environment jet stream, when it is organized as a *baroclinic critical region* (BcCR). The large-scale flow making a BcCR is such that reducing the speed of the jet as one nears the exit implies, in a balanced atmosphere, that the vertical slope of a number of properties such as isentropes or absolute momentum surfaces increases: the low-level jet closes on the location of the upper-level one. A cyclone moves through it, shown by the four-step sequence of heavy solid contours of vorticity or geopotential. This implies that, as the system nears the exit area, its low-level component tends to close on the upper-level one and to align with it in the direction of the large-scale baroclinicity (step 2 and step 3). At step 3, provided the upper-level component remains upstream, the vertical structure is in line with BcCR so that it is optimal at least for that component of the baroclinic source term, and growth is increased, possibly significantly, depending on the other aspects. Adapted from Rivière, G., Joly, A., 2006. Role of the low-frequency deformation field on the explosive growth of extratropical cyclones at the jet exit. Part II: Barotropic critical region. *Journal of Atmospheric Sciences* 63, 1982–1995.

cyclones, especially mature ones, are in themselves a world inhabited by different kinds of smaller scale organizations of the atmospheric flow. A presentation of extratropical cyclones would not be complete without mentioning some of these substructures.

Fronts

One kind of such airflow organization that is intimately linked to extratropical cyclones is atmospheric front. The name means that, when one is overpassed by an atmospheric front, one observes very rapid changes of the atmospheric conditions. In the case of a cold front, for example, the temperature drops, the pressure rises, the sky changes from overcast and low to quite clear, and the wind turns and its speed either rises or falls. All these events happen suddenly, quasi simultaneously.

The concept of front replaces with one word and one picture this list of closely related evolutions of many individual fields. It is this concentration of rapid changes or large gradients that really is the hallmark of a front. It is a well-observed feature. Returning to Low M in the Pacific Ocean provides some examples of fronts. **Figure 30** pictures several aspects of the low-level structure of that cyclone at the time its core vorticity is largest. Except in the cores of Low M and the remnant of Low P, vorticity is organized in narrow bands (less than 200 km) reaching values of $2f$ and more. Having in mind that the fields are truncated, this is an underestimate. The moist potential temperature gradients, averaged over the whole period of the case study reach a value of $6 \text{ K}/500 \text{ km}$ below the mean jet-stream core ($1.1 \times 10^{-5} \text{ K m}^{-1}$). In two places of Low M

at that time, these same gradients reach $10 \text{ K}/150 \text{ km}$ ($6.5 \times 10^{-5} \text{ K m}^{-1}$). Notice also that these areas of very large gradients and of vorticity are superimposed, they also coincide with very neat boundaries of saturated air. The wind is also shown. Where the θ'_w horizontal gradient is strongest, it reaches 55 kt (28 m s^{-1}) in a narrow southwestern extension close to the vorticity band. Closer to the core, a kind of low-level jet is present with southwesterly wind reaching 75 kt (39 m s^{-1}). These confined areas are also places of strong turbulence. All these aspects are graphically translated by frontal and other symbols, a way to compact quite a lot of information.

However, it is also important to note that these fronts are quite limited both in space and in time. They really are embedded in the cyclone and are present only in portions of it. Their presence is also temporary, as can be seen, for example, in **Figure 7**. It shows that the cold front is still strong and has gained extension on 01/04 at 06UTC but then weakens significantly during the following 24 h. It is important to point this out because during most of the twentieth century, fronts have been depicted as planetary scale semipermanent organizations, indeed the large-scale environment from which extratropical cyclones emanated. To some extent, there is a planetary scale but quite weak front, the low-level signature of the jet-stream flow on the meridional temperature gradient. But it is better to draw a line between this view and the current one for the following reasons. First, the locally well-observed rapid changes that mark a front are definitely not observed on the planetary scale: using the same word for very active and quite peaceful environments simply weakens the concept to the point of making it useless. Second, the growth mechanism that could sustain cyclones

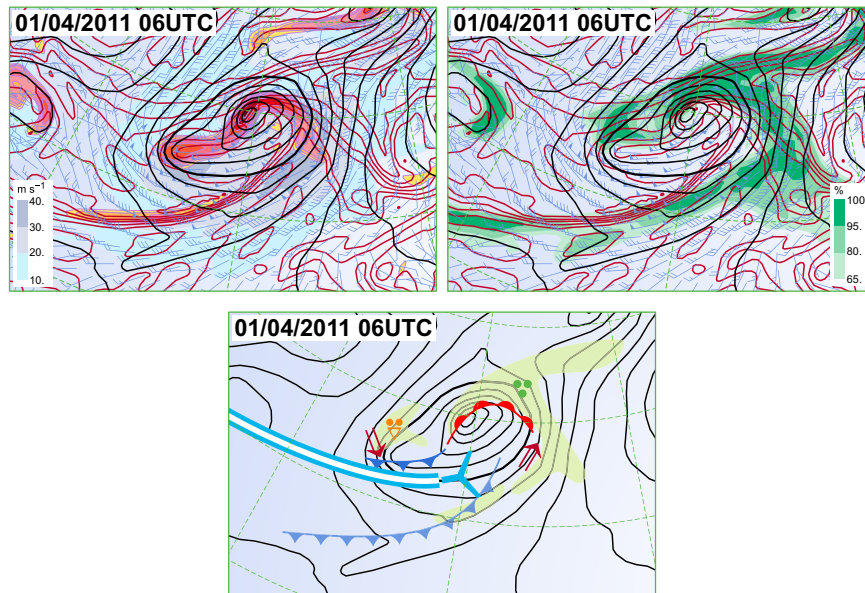


Figure 30 Details of the low-level organization of Low M at the time it reaches its maximum development, showing the presence and the structure of atmospheric fronts. Top panel: 850 hPa relative vorticity (purple contours and yellow to red shading, as in [Figure 6](#)), mean sea-level pressure (black contours as in [Figure 1](#)), 850 hPa moist potential temperature (dark red contours, interval 2 K), and wind at the same level (light blue barbs and shading with scale in the panel). Middle panel: pressure, 850 hPa moist potential temperature and wind superimposed on relative humidity at 700 hPa (green shading, scale in the panel). Bottom panel: the characteristic features that can be seen in these fields are graphically translated in this last map. The most important ones are the fronts, most intense when combining large temperature gradients, vorticity, and strong winds. The large arrows indicate low-level jets. The double blue line stands for the upper-level jet, with indication of a diffluent zone. The graphical code is from Santurette, P., Joly, A., 2002. ANASYG/PRESYG, Météo-France's new graphical summary of the synoptic situation. *Journal of Applied Meteorology* 9, 129–154.

growing from a large-scale front is primarily barotropic, and this, among other many things, cannot lead to large pressure deepening. Only baroclinic conversion taking place over the depth of the troposphere can allow for this.

So atmospheric fronts are transient features that result from the development of extratropical cyclones. The mechanism that leads to this is now well understood. It is an inherently nonlinear process that is to a large extent explained by an elegant extension of the quasi-geostrophic framework, the semi-geostrophic theory. The main aspect that needs to be added to this theory in its simplest form in order to obtain idealized fronts quite close to observed ones is latent heat in the ascents.

Fronts result from a positive feedback between the geostrophic forcing of the ageostrophic circulation and that circulation in the presence of a boundary. Recall that the ageostrophic circulation sets up in the sense of limiting the geostrophic flow tendency to weaken thermal wind balance: an ascent (say) reduces the geostrophic warming resulting from a poleward flow in the presence of a meridional temperature gradient. However, close to a boundary such as the surface, vertical motion is geometrically limited, so the countering action is not efficient. At the same location, however, a growing horizontal ageostrophic convergent flow is required to provide air to the ascent base as requested by mass conservation. The action of the deformation is poorly controlled, so its effect increases, so the response in vertical motion also increases. This implies that the horizontal convergence also grows, and reaches amplitudes such that its own advective effect on the thermal field cannot be ignored any more, as it is in quasi-geostrophic balance. It brings the extra contribution to

increasing the gradients that lead, in an idealized flow, to infinite gradients in a finite time. This theory indicates that fronts initially form near boundaries such as the surface or the tropopause, and possibly extend into the troposphere.

System-Relative Air Streams

So far, the description of extratropical cyclones and of their growth mechanism has been given in terms of coherent flow organizations such as vorticity extrema, wind field, and the likes. These structures can be represented mathematically by waves or, rather, by wave packets. They can be tracked and their interactions can be evaluated. This is a useful framework for understanding the basic adiabatic dynamics operating in cyclones.

However, some aspects of the description of the associated weather, namely the cloud and precipitation patterns, the origin of the air involved in a cyclone circulation does not fit well into this framework. They seem to come last in a long chain of dynamical causes and effects, while these properties come first in terms of noticing the weather.

In other words, most of this article describes cyclones in terms of Eulerian-based fields. However, the associated motion displaces air parcels that carry with them a number of properties such as their water content, initial temperature, and location. The fact is the motion of the Eulerian coherent structures in fields, that can be described by dispersion equations, has little or nothing to do with the actual motion of air parcels relative to the system. Yet, this is of interest to complete the picture.

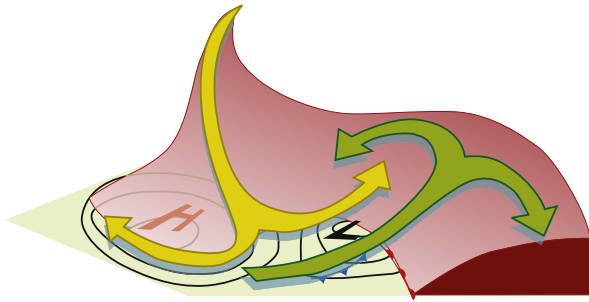


Figure 31 Overview of the main air streams that can be observed within a baroclinic wave, relative to the system (large arrows). The sloping surface that supports them is a surface of constant potential temperature θ , or rather, of constant moist potential temperature θ'_w . From Thorncroft, C.D., Hoskins, B.J., McIntyre, M.E., 1993. Two paradigms of baroclinic-wave life-cycle behaviour. *Quarterly Journal of the Royal Meteorological Society* 119, 17–55, published by John Wiley & Sons Ltd.

Taking the large-scale view, it has been thought that a cyclone involved two basic air streams relative to its overall structure (Figure 31). One involves the warm air, which comes from low latitudes and low levels, ascends while going through the cyclone core and escapes the cyclone either westward (moving more slowly than the cyclone) or eastward (when moving more rapidly than the cyclone) in an outflow jet. The other air stream has a poleward, stratospheric origin and descends toward the westward side of the cyclone where it splits into two branches, one reaching for low latitudes and low levels, the other moving toward the cyclone core, possibly reascending. Because of its origin, this air is very dry. The splitting pattern can be seen on satellite images in the form of a dry hammer-shaped feature. The scale of these streams is of the order of L_R .

A more general approach has been developed in order to explore the existence of actual coherent air streams relative to real cyclones. A series of Eulerian analyses are used to compute

a large number of air parcel trajectories, emanating from almost everywhere, over the lifetime of a given cyclone of interest. A number of properties are then interpolated to these individual trajectories: their location in 4D space, of course, but also their potential temperature, water vapor content, potential vorticity, etc. Then, a given objective criterion is selected, such as 'trajectories that ascended the most' during the selected period, or 'trajectories with parcels that lost the most water vapor' or those that 'underwent the largest descent.'

The extraordinary thing is that, when trajectories are selected in this way, they appear not to be randomly distributed but to be concentrated into coherent ensembles of trajectories for a lapse of time of the same order as the time of a cyclone development. This strongly supports the idea of the existence of coherent air streams relative to cyclones. Some actual examples are shown in Figure 32.

Consider the warm air stream of the broad view. The 'coherent ensemble of trajectory' view reveals a very large relative change of location between the stream and the developing cyclone. The warm air stream that reaches the core of the cyclone at the mature stage is made of air that comes from low latitudes and low levels far away to the east of the incipient cyclone when it starts to form.

The broad picture, however, is confirmed. The main idea to keep is that, while the shape of a cyclone is made up from coherent structures of the flow that propagate in a roughly zonal direction, the air that is mixed within its core is brought by parcels that move mostly meridionally, which is not completely intuitive. The stronger the cyclone, the more directly meridional air parcels converge toward its core at the mature stage.

Example of Recent Concept: The Sting Jet

The detailed study of high-resolution satellite image of cyclones or of radar measurements reveals a wealth of meso- and small-scale features within these air streams. Some of these can

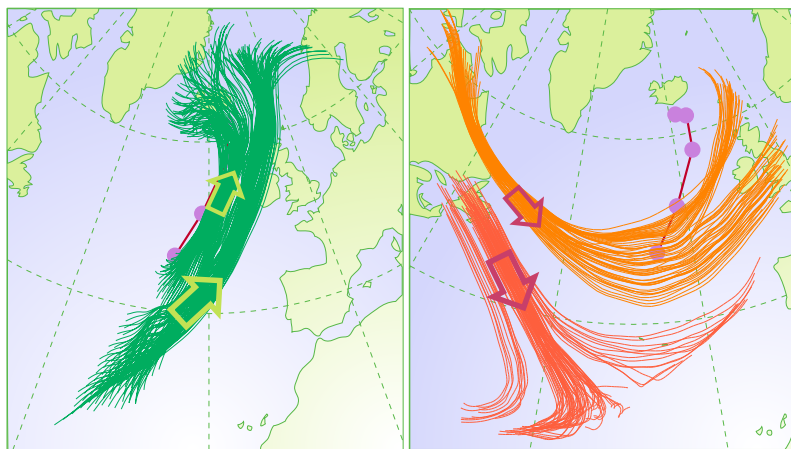


Figure 32 Objectively computed coherent ensembles of trajectories between 22 November 00UTC and 24 November 1992 00UTC. At the final time, there is a cyclone south of Iceland; the trajectory of the minimum pressure during the same 48 h is shown by the linked circles. The air trajectories are absolute, not relative to the system, they are derived from a series of three-dimensional analyzed fields. Left: trajectories undergoing an ascent larger than 620 hPa. Right: trajectories descending more than 380 hPa (southernmost ones in brown) and descending more than 350 hPa with initial potential vorticity larger than 2 pvu together with a third ensemble of descending trajectories keeping a large potential vorticity (northern ensembles in orange). Adapted from Wernli, H., 1997. A Lagrangian-based analysis of extratropical cyclones. II: A detailed case-study. *Quarterly Journal of the Royal Meteorological Society* 123, 1677–1706, published by John Wiley & Sons Ltd.

be seen in the MODIS image of Low M (Figure 2). Precipitations, for example, captured by the radars look like bands, continuous or dashed, much more narrow than the clouds within which they are. It is not possible to detail here the numerous ideas that have been proposed to account for these facts.

Rather, it is interesting to focus on a small-scale aspect of extratropical cyclones that has received attention in recent years. This is because this aspect is significant from the point of view of the destructions caused by stormy cyclones. Indeed, it has been noted that the most destructive part of an extratropical cyclone is not related to the low-level jets associated to its cold or warm front. It happens in the clearer, colder air at the rear of a passing cyclone. Seeing atmospheric conditions improving, one would think that the event is closing to an end (locally speaking), yet the wind brutally increases to gale force, with intense gusts. It results in many damages, more important ones than those caused by other parts of such storms, that form a finite track of wrecks along the path of the cyclone.

The largest scale signature of such a feature is a local extra increase of the pressure gradient to the west or southwest of the minimum pressure of a cyclone moving east. Because it hits so to speak from the rear, it is called a sting jet like 'the sting at the end of the tail.'

Figure 33 is a schematic showing how this feature is currently understood. The extra intensity of the wind would result from rapidly descending air hitting the surface, as if falling and then splashing on the ground. The reason why the subsidence would have been turned into rapid descent is that this air would have moved below the hooking end of the ascending core, which in that area is well away from the ground. This

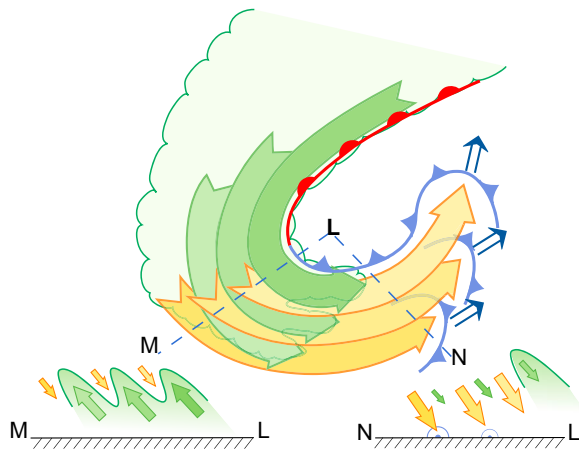


Figure 33 Schematic showing how the small-scale organization of precipitations within an extratropical cyclone may lead to an area of very strong surface wind at the rear of the system that can be destructive, termed *sting jet*. The large green arrows represent moist ascending air, the ascent zone being subdivided into narrower bands. Interspaced with those are streams of descending air that can be cooled by evaporating rain or snow falling from the moist bands (large orange arrows): this can strongly accelerate the descending motion and, when reaching the ground, suddenly reinforces the surface wind as this very cold air spills on the ground (blue strong wind arrow symbols). Also shown are two vertical cross sections in the system. Adapted from Browning, K.A., 2004. The sting at the end of the tail: damaging winds associated with extratropical cyclones. *Quarterly Journal of the Royal Meteorological Society* 130, 375–399, published by John Wiley & Sons Ltd.

ascending air is precipitating, and it is the evaporation of those precipitations in the dry descending air that would cool it, create negative buoyancy and push it downward.

Extratropical Cyclones and the General Circulation

This final section addresses questions such as why the weather takes the form of cyclones at middle and high latitudes. It is then useful to attempt to understand the evolution of weather on the slightly longer timescale of the order of the week to the month.

Why Extratropical Cyclones?

Atmospheric motions result from and redistribute the inhomogeneous input of energy from the sun, which is in excess at low latitudes and in deficit at high ones. A planetary scale, purely meridional overturning of air, with ascent about the Equator and descent near both poles could both perform the desired redistribution and be energetically self-consistent. However, the spinning of the Earth precludes such a simple meteorology and introduces strong mechanical constraints. These can be understood in terms of angular momentum of air parcels about the axis of rotation of the Earth. To first order, neglecting friction, the only source of absolute motion is the pressure difference between two regions. In the simple overturning model assuming zonal homogeneity, the pressure force is directed toward or away from the axis of rotation; its moment is zero and therefore air parcels moving toward or from the axis conserve their angular momentum. The angular momentum has two components: one relates to the local value of the planetary rotation and the other to the angular momentum of the motion of the parcel relative to the planet normal to the axis (zonal, that is). Thus, the conservation of angular momentum implies a characteristic zonal wind distribution: it turns out to be completely unrealistic with unobserved values at latitudes higher than about 25°. In other words, angular momentum cannot be conserved and atmospheric motions redistribute both energy and angular momentum.

Angular momentum exchanges between air parcels moving away from and toward the axis of rotation require a component of a body force, the pressure force, that is not directed toward the axis – in other words, that has a zonal component. The pressure and other related fields must have zonal variations. Together with the meridional distribution, this leads to alternate cells of high and low pressure at midlatitudes and high latitudes. Hence the vortex-like overall shape of the structures where these exchanges of heat and angular momentum take place.

As has been discussed above, the scale of extratropical cyclones and anticyclones implies that they evolve close to some balanced state of the atmosphere. Among other consequences, low-pressure zones turn out to be rather associated with ascending motion, and therefore clouds and rain. This makes them more 'weatherly' than anticyclones.

Feedback of Cyclones on the Large-Scale Flow

Extratropical cyclones can be seen as synoptic and subsynoptic scale perturbations of the large-scale flow. A short summary

of the main exchange terms has been provided above, in eqns [7]–[10]. Those same terms appear on the large-scale evolution side, except the perturbation components, the fluxes, is averaged and the signs are reversed. The fact that cyclones transport heat upward and poleward at mid and higher latitudes can be quantitatively verified by computing such fluxes. While this is quite frequent and generally easy, in idealized studies, finding characteristic results in case studies is less easy. Figure 34 is such an example. The case, or the period rather, illustrated by this figure pertains to the Southern Hemisphere as its circulation seems closer to a simple flow over a mountainless planet: comparison with idealized studies may be easier. The expected poleward heat flux is indeed present when the main system develops. There is also a large poleward flux of zonal momentum, a component that may modify the jet flow within which the cyclone evolves. This is the main aspect addressed in this section.

The emergence of long homogeneous reanalyses has enabled, on this topic as well, to provide a much-improved view of the behavior of real cyclones. The question has been originally considered in the first consistent framework for understanding cyclones mentioned in an earlier section above: how cyclones may influence the zonal flow within the quasi-geostrophic approximation. In this case, the averaging operator is essentially a zonal mean. The zonal (large scale) wind and temperature distributions can be influenced by

cyclones through spatial operators acting on the Eliassen–Palm flux $\mathcal{F}_{EP}$:

$$\mathcal{F}_{EP} = \left(-[u'v'], \frac{g}{\theta_0 N^2} [v'\theta'] \right), \quad [12]$$

where the operator $[\cdot]$ is a zonal mean. The two ‘components’ of this pseudovector are defined in a meridional vertical cross-section oriented northward and upward. The relationship between the divergence (in the meridional plane) of that vector and the evolution of the zonal large-scale flow is in fact nontrivial at all. This is further complicated by the existence of imposed cancellations between perturbation contributions through fluxes and large-scale flow only processes.

The focus now is less on the zonal flow than on critical zonal inhomogeneities of the jet flow, as discussed previously. In order to study 3D or 4D large-scale flows, the relevant averaging operator locally selects chosen time and/or spatial scales. A popular way to separate the flow into a large-scale environment and perturbations using time series of gridded fields such as reanalyses is by using time filters on each grid point time series. This kind of separation enabled to unveil characteristic patterns of the large-scale flow. The question of the feedback of the cyclones becomes that of understanding whether cyclones are contributing to the maintenance of these large-scale patterns or whether they can explain their changes. This turns out to be a very complex question that remains open in several aspects.

The large-scale flow patterns have been extracted in various ways. One statistical approach is to perform a classification or clustering of the day-to-day low-pass time-filtered maps of the large-scale flow. Some of these organizations can be given some dynamical meaning either a posteriori by applying a proper forcing to simplified atmospheric models and studying the response. Alternatively, characteristic large-scale patterns have also been obtained directly from dynamical constraints, such as finding structures that are quasi stationary by minimizing their time-evolution tendency. It is convenient here to distinguish between very large-scale patterns, the dominant shape of which will characterize the large-scale flow of, say, a season and other such patterns that are reasonably steady or propagate slowly with a time constant between a week and 10 days. The former have received names such as the North Atlantic oscillation or the Pacific North American oscillation. The latter are sometimes called weather regimes. As their name suggest, large-scale oscillations correspond to large-scale anomalies with similar shapes but opposite signs, a kind of two-state oscillation. Weather regimes have richer structures. For example, many different techniques, statistical, dynamical, or mixed yield from different data sources four quite robust structures preferred by the large-scale flow over the North Atlantic basin. The large-scale flow in that area can be seen as periods dominated by one of these four flows or regimes that seems steady for several consecutive days, sometimes weeks followed by relatively rapid transitions from one regime to another. Here, ‘rapid’ means that the change occurs in a matter of 2 or 3 days. The large size of the North Pacific basin requires extended techniques for regimes to be found: while regimes are indeed stationary in the North Atlantic, they propagate as large-scale Rossby waves in the North Pacific. The separation between oscillations and

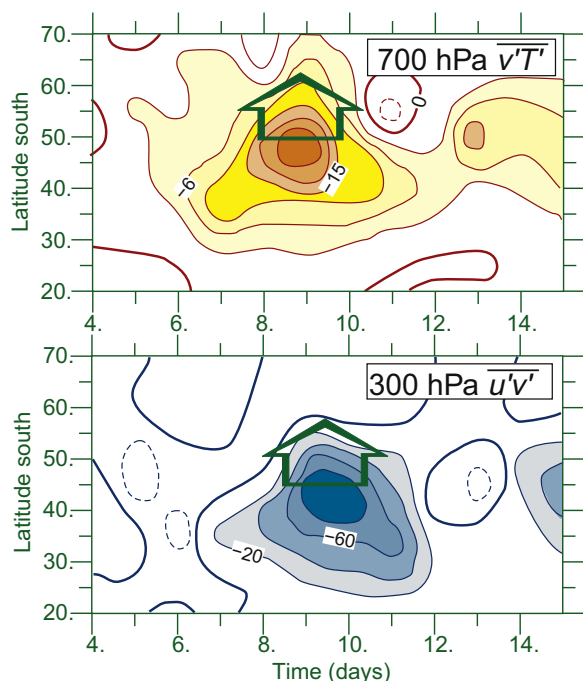


Figure 34 Latitude-time sections of the zonally averaged northward heat flux at 700 hPa in K m s^{-1} (top panel) and northward momentum flux at 300 hPa in $\text{m}^2 \text{s}^{-2}$ (bottom panel) for zonal wavenumbers 5–7. Negative fluxes are directed poleward, as indicated by the large arrow, since this is a Southern Hemisphere case study. The day of maximum activity, day 9 on the panels, is 13 December 1979. Adapted from Randel, W.J., Stanford, J.L., 1985. The observed life cycle of a baroclinic instability. *Journal of Atmospheric Sciences* 42, 1364–1373.

regimes is not as clearcut as presented here for simplicity: some authors choose to ignore the distinction and characterize all large-scale variability with the oscillations patterns.

These large-scale patterns determine the overall configuration of the jet stream over a given stormtrack area: location of main jet entrance and exit zones, therefore extent as well as orientation and curvature of the jet. This will considerably influence the trajectories and main phases of the synoptic scale cyclones, following the mechanisms outlined in previous sections.

The adiabatic part of the feedback of the cyclones on these 3D patterns, on the other hand, depends on the pseudovector $\mathbf{E}$:

$$\mathbf{E} = \left(\mathbf{E}', \frac{g}{\theta_0 N^2} \overline{v'\theta'} \right),$$

where $\mathbf{E}'$ is the horizontal vector defined by eqn [9]. Here, the $\overline{}$ operator is a time average or a scale average that keeps some of the zonal variability. To be more precise, the meridional component of the gradient of the divergence of $\mathbf{E}$ is one of the two major terms forcing the evolution of the large-scale potential vorticity field. It is indeed the major contribution to the divergence of the meridional flux of the perturbation potential vorticity. These results approximately hold at the level of the quasi-geostrophic approximation. At a similar level of approximation, the $\mathbf{E}$ pseudovector has a number of interesting heuristic properties that have been hinted previously. In a balanced flow, $\mathbf{E}$ depends on the horizontal and vertical tilts of the perturbations. When the $\mathbf{E}$ vector is oriented westward (or in the direction of the large-scale flow), the associated perturbations tend to be stretched in directions normal to $\mathbf{E}$. Namely, they are elongated meridionally. The trough associated to the perturbation tilts horizontally normal to a westward $\mathbf{E}$. The group velocity vector of the perturbations, in the same context, is close to $\mathbf{E}$, given a known acute angle. Recall also that the 'vertical' component of $\mathbf{E}$, the meridional heat flux, is related to the vertical tilt of the perturbations.

Figure 35 shows the distributions of momentum fluxes and of $\mathbf{E}$ pseudovectors at the eastern end of the Pacific stormtrack

and over the North Atlantic stormtrack. The perturbations highlighted by this figure have characteristic timescales between 2 and 5 days. At these scales, the momentum fluxes are strongly poleward at the southeastward end of the stormtracks, but equatorward fluxes in the northwestern and central part of the stormtrack. The $\mathbf{E}$ vectors are clearly pointing eastward, northeastward in the northern part of the stormtrack, southeastward in the southern end. This corresponds to perturbations generally taking a meridional extension. This configuration is in the sense of increasing the vertically averaged zonal flow at the entrance and heart of the North Atlantic stormtrack.

There are, however, limitations to these approaches that can be perceived even at the level of this very short account. Just as with the Eliassen–Palm flux and the zonal mean circulation, it is not obvious to go from the distribution of $\mathbf{E}$ to the change in the jet flow. Indeed, the large-scale kinetic energy evolves depending on the correlation between the large-scale zonal flow $\bar{u}$ and $\nabla \cdot \mathbf{E}$, so that a given distribution of $\mathbf{E}$ does not have the same effect on its environment if the latter is different. Furthermore, $\mathbf{E}$ vectors can be constructed for different scales, depending on the averaging operator chosen. These lead to another limitation: it happens that varying the scales of perturbations analyzed in this way yields quite different distributions of $\mathbf{E}$. The difficulty comes from the fact that the scale of cyclone perturbations changes significantly: as it moves through the stormtrack and grows, it tends to gain a larger scale.

These are some of the reasons why, recently, a heuristic understanding of the feedback of cyclones on the jet flow has favored a complementary approach allowing this feedback to be more easily anticipated at case study level. The idea is that cyclones near the end of the stormtrack reach their maximum amplitude and interact in a nonlinear way with their environment. They also tend to lose their baroclinic tilt and behave in a more barotropic way on the vertical, with maximum amplitudes in the upper troposphere. Looking at their evolution on a semihorizontal surface, such as a constant potential vorticity surface or the poleward part of a surface of constant potential

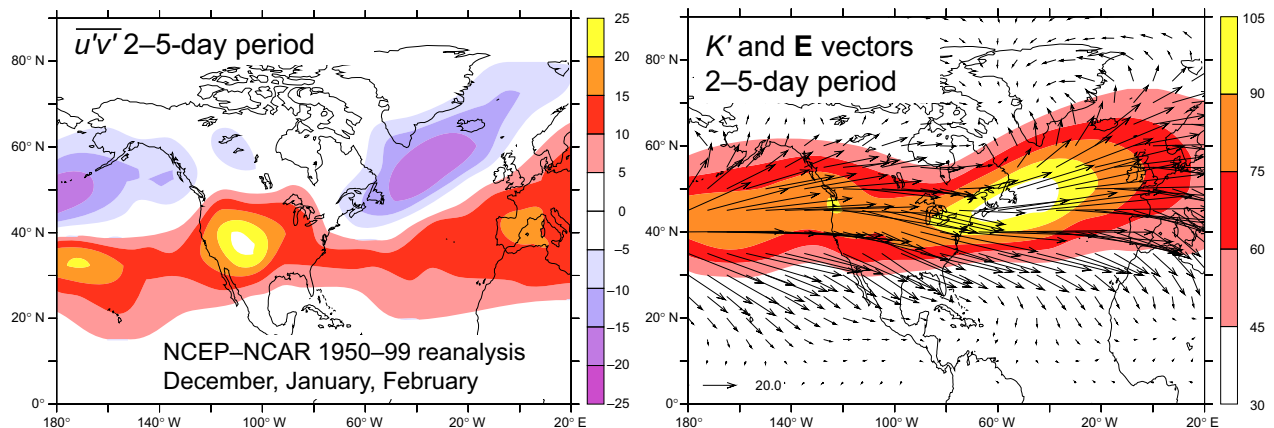


Figure 35 Properties of perturbations resulting from selecting the 2–5-day period when time-filtering the NCEP–NCAR 1950–99 reanalysis fields at 00UTC and 12UTC, December–February. Top panel: meridional momentum fluxes in $\text{m}^2 \text{s}^{-2}$. Bottom panel: shading, kinetic energy in $\text{m}^2 \text{s}^{-2}$, arrows, $\mathbf{E}$ -vectors. Adapted from Rivière, G., Orlanski, I., 2007. Characteristics of the Atlantic storm-track eddy activity and its relation with the North Atlantic oscillation. *Journal of Atmospheric Sciences* 64, 241–266.

temperature, this late part of their evolution looks like they are breaking waves.

It turns out that there are essentially two ways for mature cyclones to break as waves and strongly interact with the jet flow. This has been initially uncovered in an idealized study of cyclone life cycle, but it appears to be representative of observed behaviors.

Figure 36 summarizes with a single contour these two kinds of behavior and their main effect on the jet flow. One was originally called life cycle 1 (LC1) and it evolves into an anticyclonic wave breaking. It is characterized by the development of an upper-air trough that is significantly stretched with a northeast–southwest orientation. The sketch in the figure is based on a single contour of a quasi-conservative parameter (such as potential vorticity or absolute vorticity) on a quasi-conservative surface that summarizes the overall shape of the total flow locally dominated by the cyclone perturbation. Superimposed is the corresponding cyclonic upper-tropospheric vorticity anomaly, which is also stretched. From this, assuming that these perturbations are essentially nondivergent (consistent with quasi-geostrophic ideas), it is easy to recover the associated momentum flux that appears to be consistent with the jet displacement relative to the system. An anticyclonic wave breaking is associated with momentum fluxes converging on the poleward side of the jet (and diverging on the equatorward side). This results in shifting the jet poleward. In this situation, the amplitude of the surface cyclone tends to decrease rapidly as the upper-air part cuts off from the main flow, prior to being stretched out.

The other life cycle, called life cycle 2 (LC2), leads to cyclonic wave breaking. In the upper troposphere, the cyclonic perturbation remains tilted northwest–southeast while expanding in scale. In this case, momentum fluxes diverge on the poleward side, converge on the equatorward side: the jet is shifted equatorward. During this kind of breaking, the cyclone kinetic energy peaks at larger values than those reached by anticyclonic wave breaking. The resulting large-scale cyclonic trough also tends to persist, while moving poleward.

As the schematic shows, to some extent, the views provided by E vectors and contour dynamics are complementing each other. However, because the latter directly shows such things as a local change of sign of the gradient of a conservative variable such as potential vorticity, it directly implies a change in the background flow, effortlessly selecting only the events the amplitude and scale of which do lead to serious feedback.

Recently, automatic detection of wave breaking using contour tracking in reanalyses has enabled to build composites of such situations. In the North Atlantic basin, consistently with **Figure 35** above, cyclonic wave breaking is favored in the northwestern part of the winter stormtrack, while anticyclonic wave breaking is frequent in the southeastern part, close to or aloft southwestern Europe and the western Mediterranean. On average, anticyclonic wave breaking is the preferred evolution, implying a poleward shift of the jet at the end of stormtracks that is indeed noted. Indeed, returning again to the example of Low M in the northern Pacific, **Figure 37** shows it as a realization of an anticyclonic wave-breaking event.

There are many aspects of the dynamics that influence the type of wave breaking. It must be noted, for example, that from

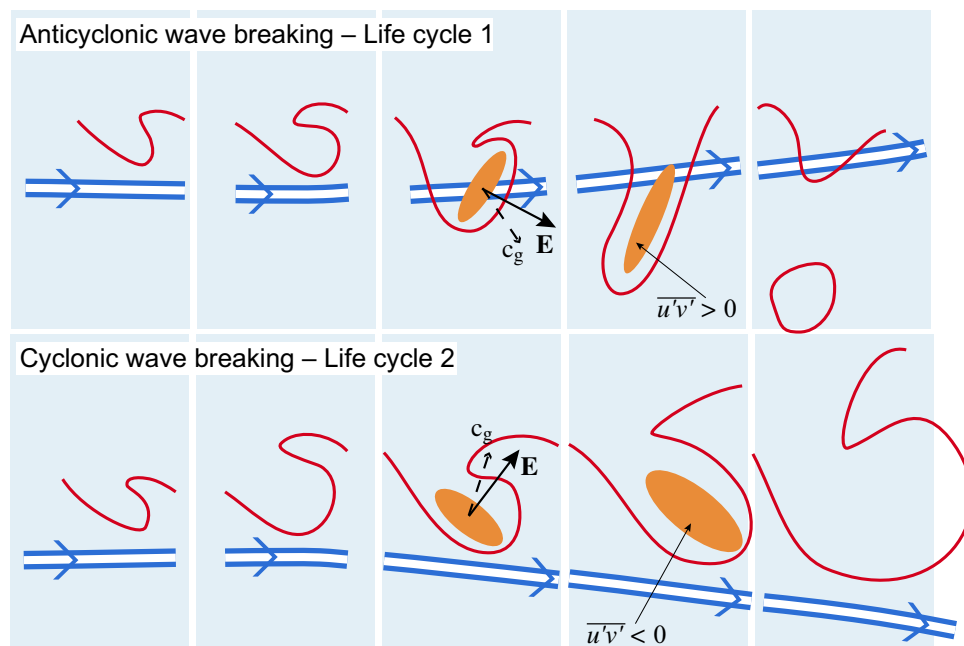


Figure 36 Time sequences summarizing with a single contour representative of the upper-tropospheric flow configuration two paradigms of the nonlinear phase of extratropical cyclones. The shaded surface stands for the relative vorticity perturbation. The double line represents the upper-level jet. Derived from Thorncroft, C.D., Hoskins, B.J., McIntyre, M.E., 1993. Two paradigms of baroclinic-wave life-cycle behaviour. *Quarterly Journal of the Royal Meteorological Society* 119, 17–55 and from Hoskins, B.J., James, I.N., White, G.H., 1983. The shape, propagation and meanflow interaction of large-scale weather systems. *Journal of Atmospheric Sciences* 40, 1595–1612.

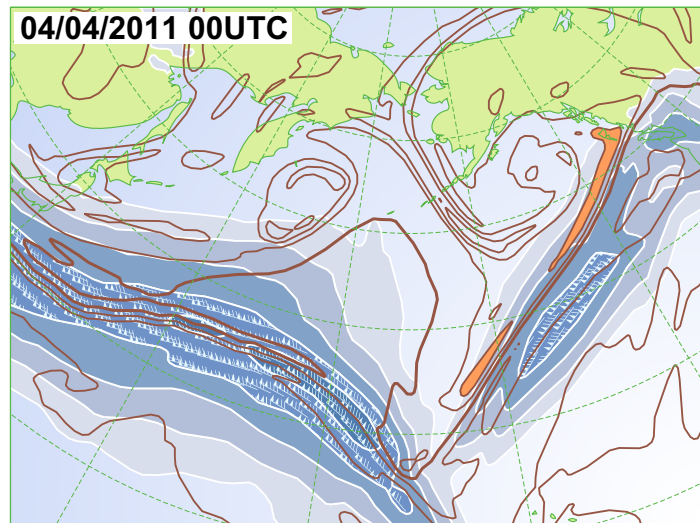


Figure 37 The upper troposphere flow configuration at a late stage of the evolution of the example of extratropical cyclone shown in Figures 1–12. It appears to provide an example of the more frequent anticyclonic wave breaking. Fields from Météo-France global assimilation system. Wind barbs, white contours, and blue shading: 200 hPa wind field, see Figure 11 for the color scale. Brown contours: tropopause geopotential height, interval 1000 m, thicker contour 10 000 m.

that perspective, the quasi-geostrophic approximation has clear shortcomings with respect to the more precise primitive equation dynamics (with hydrostatic balance as the only scale assumption). The latter introduces an initial bias of the incipient cyclones toward a cyclonic horizontal tilt that the former is not able to provide. Sphericity and latitude also may influence the type of wave breaking through, for example, the so-called stretching term in the evolution of perturbation vorticity.

Initially, however, the key parameter appeared to be the presence of a barotropic horizontal shear component in the jet flow. That is what led idealized simulations to yield either an LC1 or an LC2. The favored breaking is determined by the sign of the shear: cyclonic shear leads to cyclonic wave breaking. However, the amplitude required is not symmetric at all. A small anticyclonic shear is sufficient to lead to anticyclonic breaking, while a large cyclonic shear is needed to reverse that preferred behavior. However, the latitude of the jet, its intensity, its width are also critical. For example, a more poleward jet further favors anticyclonic wave breaking.

These ideas are useful to understand some aspects of the changes of the climate in relation to stormtracks. Using the conceptual tools summarized in this section, it has been shown in both idealized and observation-based studies that small and moderate amplitude cyclones forming within a given large-scale pattern such as a weather regime are shaped in such a way that they have a positive feedback on the maintenance of that regime. In other words, the average cyclone contributes through its heat and momentum fluxes to the persistence of the regime. Whether large amplitude systems behave similarly or conversely trigger regime transitions remains an open question, in spite of a number of single-case based reports.

See also: **Dynamical Meteorology:** Balanced Flow; Baroclinic Instability; Overview; Potential Vorticity; Quasigeostrophic Theory; Rossby Waves; Vorticity; Wave Mean-Flow Interaction; Waves. **Observations Platforms:** Radiosondes. **Stratosphere/Troposphere Exchange and Structure:** Tropopause. **Synoptic Meteorology:** Cyclogenesis; Forecasting; Frontogenesis; Fronts; Jet Streaks.

Further Reading

- Ayrault, F., 1998. *Environnement, Structure et Évolution des Dépressions Météorologiques: Réalité Climatologique et Modèles Types*. Doctorat de. Université P. Sabatier, Toulouse, France.
- Bader, M.J., Forbes, G.S., Grant, J.R., Lilley, R.B.E., Waters, A.J., 1995. *Images in Weather Forecasting, a Practical Guide for Interpreting Satellite and Radar Imagery*. Cambridge University Press, Cambridge.
- Bishop, C.H., Thorpe, A.J., 1994. Potential vorticity and the electrostatics analogy: quasi-geostrophic theory. *Quarterly Journal of the Royal Meteorological Society* 120, 713–731.
- Grønås, S., Shapiro, M.A. (Eds.), 1999. *The Life Cycles of Extratropical Cyclones*. American Meteorological Society, Boston.
- Hoskins, B.J., 1982. The mathematical theory of frontogenesis. *Annual Review of Fluid Mechanics* 14, 131–151.
- James, I.N., 1994. *Introduction to Circulating Atmospheres*. Cambridge University Press, Cambridge.
- Newton, C.W., Holopainen, E.O. (Eds.), 1990. *Extratropical Cyclones, the Erik Palmén Memorial Volume*. American Meteorological Society, Boston.
- Rivière, G., 2011. A dynamical interpretation of the poleward shift of the jet streams in global warming scenarios. *Journal of Atmospheric Sciences* 68, 1253–1272.
- Rivière, G., Hua, B.L., Klein, P., 2003. Perturbation growth in terms of barotropic alignment properties. *Quarterly Journal of the Royal Meteorological Society* 129, 2613–2635.
- Vallis, G.K., 2006. *Atmospheric and Oceanic Fluid Dynamics*. Cambridge University Press, Cambridge.

Fronts

DM (David) Schultz, University of Manchester, Manchester, UK
W Blumen[†], University of Colorado Boulder, Boulder, CO, USA

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

Traditionally, fronts have been defined as boundaries between air masses. Because this definition can be problematic in a modern context, a new definition is proposed. A front is a region characterized by frontogenesis containing both a horizontal potential temperature gradient and vorticity maximum. Fronts are classified as surface based or upper level. Surface-based fronts are further classified as those associated with synoptic-scale phenomena (cold, warm, stationary, or occluded) or those associated with mesoscale phenomena (sea breeze, gust, or drainage). Front-like features such as drylines also exist.

Introduction

The traditional definition of a front is the boundary between two different air masses, which are large bodies of near-surface air extending for hundreds or thousands of kilometers with nearly uniform temperature, moisture content, and static stability. Air masses are classified by where they originate from and the surface underlying their origin (i.e., polar vs tropical, continental vs maritime). Being the boundary between two different air masses means that fronts are often regions where changes in temperature, pressure, moisture content, and wind occur over small distances. Fronts can also be associated with precipitation. Thus, for these reasons, identification and tracking of fronts are crucial for weather forecasting.

This traditional definition of fronts was developed by Norwegian meteorologists in the 1910s and 1920s studying the structure and evolution of extratropical cyclones (Bjerknes, 1919; Bjerknes and Solberg, 1922). Using the terminology of the recent First World War, fronts were lines on a map drawn where the clash of cold and warm air masses occurred. Although abrupt changes in wind and temperature had been previously recognized before the Norwegians (e.g., historical reviews by Bergeron, 1959; Kutzbach, 1979, Section 6.7; Davies, 1997; Newton and Rodebush Newton, 1999; Volkert, 1999), the Norwegians were the first to embed fronts as the unifying concept within the three-dimensional structure and evolution of extratropical cyclones.

Toward a New Definition

In a modern scientific context, however, this traditional definition can be problematic. First, fronts do not always form along the edge of air masses, but may form within air masses. Because the air near the surface may not be easily classified into large regions of nearly homogenous properties, even within well-defined air masses, gradients of temperature and wind may be present and could be considered to be fronts. Thus, we seek a definition of a front that could allow for the existence of fronts within an air mass.

Second, although boundaries between air masses may be clear in some cases, what constitutes an air mass or its boundary may not be clear in all cases. As an attempt to address this weakness, some investigators have defined a front only if the horizontal temperature gradient or other higher derivatives of the temperature gradient exceeds strict thresholds (e.g., Hewson, 1998; Sanders, 1999; Sanders and Hoffman, 2002). Although such approaches automate the analysis and detection of fronts, they do not help better understand the processes affecting fronts and their intensity or why fronts form where they do.

Third, the large horizontal scale of air masses implies that fronts – as the boundaries of these air masses – must possess the same characteristic scale: large mesoscale or synoptic scale in the along-front direction (200–2000 km) and small mesoscale in the across-front direction (20–200 km; Keyser, 1986). Not all fronts, however, may be this large; some fronts may be small mesoscale in the along-front direction and microscale (2–20 km) in the across-front direction. Thus, the traditional definition may exclude smaller scale features that share the same physical processes and deserve to be considered fronts, as well.

Finally, and perhaps most significantly, the traditional definition of a front as the boundary between air masses forced meteorologists into an unproductive game of arguing over where to draw the line on the map representing the front, rather than focusing on the relevant physical processes creating and maintaining the front and associated sensible weather (e.g., wind shifts, temperature changes, precipitation). Part of the dilemma with frontal analysis is that the characteristics of fronts used for analysis are not clearly defined, allowing any number of plausible analyses (e.g., Uccellini et al., 1992; Sanders and Doswell, 1995; Lackmann, 2011, Section 6.1). Some have raised issues with synoptic analysis (e.g., Mass, 1991; Sanders and Doswell, 1995; Sanders, 2005), and others have proposed alternative analysis schemes, particularly to deal with fronts associated with mesoscale phenomena (e.g., Colby and Seitter, 1987; Young and Fritsch, 1989). Manual analysis of weather maps is a crucial step to producing a weather forecast, but it is ultimately subjective and its overemphasis has probably inhibited a more modern approach to frontal analysis (e.g., Sutcliffe, 1952; Mass, 1991; Schultz, 2008).

[†] Deceased.

An alternative approach is a mathematical one. Petterssen (1936) defined frontogenesis as the Lagrangian rate of change of the magnitude of the horizontal potential temperature (θ) gradient due to the horizontal wind ($\mathbf{V}_2 = u\mathbf{i} + v\mathbf{j}$):

$$F = \frac{d}{dt} |\nabla_2 \theta|, \quad [1]$$

where

$$\frac{d}{dt} = \frac{\partial}{\partial t} + u \frac{\partial}{\partial x} + v \frac{\partial}{\partial y},$$

$$\nabla_2 = \mathbf{i} \frac{\partial}{\partial x} + \mathbf{j} \frac{\partial}{\partial y}.$$

This mathematical definition of a front emphasizes that both the temperature field and the wind field are responsible for producing frontogenesis.

For these reasons and to update the traditional definition, a modern definition of a front is proposed that combines the physical and mathematical approaches (e.g., Keyser, 1986; Sanders, 1999; Lackmann, 2011). A front is a region characterized by frontogenesis containing both a horizontal potential temperature gradient and vorticity maximum. Sometimes, the vorticity is not coincident with the temperature gradient, but often is. The benefit of using frontogenesis in the definition is that it treats fronts as processes, not rigid objects being advected around by the flow, which is an unfortunate result of treating fronts as the boundaries between air masses. If a region contains no temperature gradient or vorticity, then it cannot be a front (Sanders, 1999). If the gradient did not arise through or possess frontogenesis, then it is not a front. This definition also allows for the possibility of fronts forming within air masses and is independent of scale. Finally, this definition alleviates the problem of where to draw the line on the map as regions of frontogenesis are calculated explicitly.

Fronts Associated with Synoptic-Scale Phenomena

The intensity of a front, measured by changes in the temperature and wind fields that occur across the frontal transition zone, is most pronounced at or near the Earth's surface and the tropopause. These fronts occur on the synoptic scale and may retain their individual identities for many hours or even days. The physical processes that give rise to fronts are controlled by four principal forces: the buoyancy force, the pressure gradient force, the Coriolis force, and the friction force, whose influence is primarily restricted to the lowest 1–1.5 km of the atmosphere, the planetary boundary layer. The buoyancy force results from a small imbalance between gravity, which acts downward toward the Earth's surface, and an upward vertical pressure gradient force. When the two are equal in magnitude, hydrostatic balance occurs. The direction of the buoyancy force controls the direction of vertical motions associated with fronts. The horizontal component of the pressure gradient force, which is directed toward low pressure, is largely opposed by the Coriolis force. The Coriolis force is directly proportional to the Earth's rotation rate and to the magnitude of the horizontal wind vector, directed to the right of the horizontal wind in the Northern Hemisphere. These forces tend to be in

approximate geostrophic balance above the planetary boundary layer, providing a counterclockwise circulation around a low-pressure center. The friction force reduces the wind speed within the planetary boundary layer, and lessens the magnitude of the Coriolis force, which leads to cross-isobaric flow toward low pressure.

The language and notation for synoptic-scale surface fronts were created by the Norwegian meteorologists (e.g., Jewell, 1981; Friedman, 1989). The strength of the Norwegian cyclone model was that it created a holistic conceptual model that related the structure and evolution of an extratropical cyclone (low-pressure system) and anticyclone (high-pressure system) with their attendant fronts and sensible weather. In the Norwegian cyclone model, an initial cyclone initiated on a broad temperature gradient (Figure 1). As the cyclone deepens, the circulation around the cyclone increases, bringing cold air equatorward to form the cold front and bringing warm air poleward to form the warm front.

Cold Fronts

The passage of a cold front typically is indicated by a drop in temperature as the advancing cold air replaces warm air (e.g., Sanders 1955; Schultz 2008; Schultz and Roebber 2008). The leading edge of the cold front is delineated by the triangles that point in the direction of movement of the cold air (Figure 1). Conceptual models typically depict traditional cold frontal passages, not only with a temperature decrease, but also with a cyclonic wind change, pressure minimum, and decrease in dew point temperature, all coincident with a line of deep,

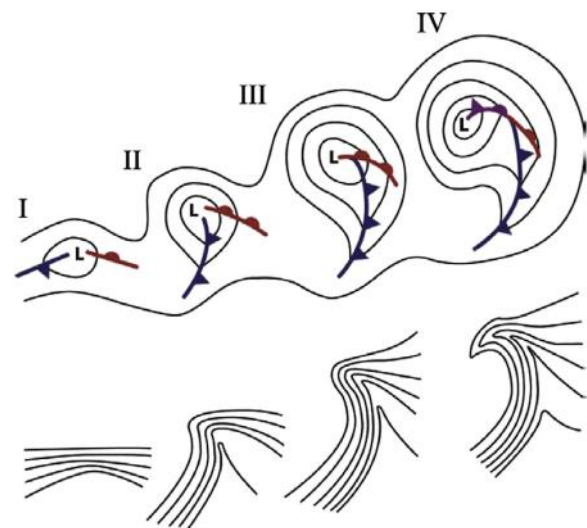


Figure 1 Conceptual model of a Norwegian cyclone showing (top) lower-tropospheric (e.g., 850 mb) geopotential height and fronts, and (bottom) lower-tropospheric potential temperature. The stages in the respective cyclone evolutions are separated by approximately 6–24 h and the frontal symbols are conventional. The characteristic scale of the cyclones based on the distance from the geopotential height minimum, denoted by I, to the outermost geopotential height contour in stage IV is 1000 km. Adapted with permission from Schultz, D.M., Vaughan, G., 2010. Occluded fronts and the occlusion process: a fresh look at conventional wisdom. *Bulletin of the American Meteorological Society* 92, 443–466.

convective clouds. In reality, however, some cold fronts do not have coincident wind and temperature changes (e.g., Schultz, 2004, 2005), are not coincident with clouds and precipitation (e.g., Mass and Schultz, 1993), or may not even be associated with temperature drops because cold air pooled in valleys ahead of the front may be colder than the postcold frontal air (e.g., Sanders and Kessler, 1999; Doswell and Haugland, 2007). Thus, the representation of cold fronts should be modified to encompass these cold fronts that do not meet the traditional criteria.

Different types of cold fronts exist, classified by their direction of motion (e.g., backdoor cold fronts), location (e.g., southerly buster), or a combination of both (e.g., arctic fronts). For example, a backdoor cold front is one that moves westward (against the prevailing westerlies). In the United States, a New England backdoor cold front is associated with the westward expansion of a cold surface high-pressure system situated near the North Atlantic coast during winter (Bosart et al., 1973; Hakim, 1992). The thermal gradient is reversed in this case, and the frontal zone is characterized by fog and low stratiform clouds formed from evaporation of moisture from the ocean. This cold front moves southward along the eastern slope of the Appalachian Mountains, which serve as a barrier to inland penetration.

The southerly buster or southerly burster is an intense summertime cold front that arrives at the southeastern tip of Australia from the Southern Ocean (Baines, 1980). Arrival of this front in the afternoon can be accompanied by temperature changes of 10–15°C over a period of a few minutes, but precipitation is not usually associated with a southerly buster. The front travels equatorward, acquiring a characteristic S-shape as its movement is inhibited by the east coast mountain chain, but movement inland and along coastal waters is less restrained.

Another subclass of cold fronts associated with the advection of arctic air equatorward is an arctic front (e.g., Wang et al., 1995). Dramatic examples of these kinds of fronts occur in the central United States, with the cold air traveling as far equatorward as Central America (e.g., Schultz et al., 1997, 1998). The bitter cold air behind the front is formed from strong radiational cooling in the arctic region (Emanuel, 2008).

Warm Fronts

Warm fronts are characterized by the advection of warm air into cold air. The leading edge of the warm front is delineated by the semicircles that point in the direction of movement of the warm air (Figure 1). Traditionally, a vertical cross section through a warm front shows a gently sloping zone over the cold air. Above the strongly sheared zone, warm air gently rises over the warm front. At the surface, a warm frontal passage would be characterized by increasing temperature and dew point, and a cyclonic wind shift. A sequence of clouds would progress from cirrus to altocumulus to nimbostratus before the frontal passage, followed by clearing skies after frontal passage.

A more modern characterization of warm fronts indicates that the flow over the warm front may not be characterized by gentle ascent. Instead, the flow may consist of convective elements embedded within the ascent (the so-called escalator–elevator paradigm of Neiman et al., 1993) or may be

characterized by banded convective precipitation (e.g., Novak et al., 2004).

Stationary Fronts

As the name implies, a stationary front has historically been classified as a front that is not moving or moving very slowly. The stationary front is denoted on the surface map by alternating triangles that point away from the cold air side of the front and semicircles that point away from the warm air side of the front, indicating the standoff between the advance of the cold air and the advance of the warm air. This type of front may reflect either a change in the synoptic-scale circulation pattern that halts the translation of the frontal zone or the heterogeneity of the Earth's surface that provides conditions to fix a frontal transition zone to a preferred location.

Schematics of stationary fronts are similar to warm fronts with their gentle slope. Stationary fronts have since been recognized to have either the slope of warm fronts or cold fronts. In fact, some stationary fronts in the central United States transition rapidly from moving equatorward as cold fronts, to becoming stationary, to moving poleward as warm fronts (Bosart et al., 2008). Consequently, a modern interpretation of a stationary front would be to continue to analyze fronts as stationary fronts, but not to apply any given conceptual model to it.

Another example of a stationary front is a coastal front. Coastal fronts are frontal zones separating relatively warm oceanic air from colder continental air, such as along the East Coast of the United States in winter. Such environments promote the formation of a stationary front (Bosart et al., 1972; Bosart, 1975). Although these fronts exhibit transition zones that are comparable to those of a cold front, they are of limited extent, ranging from 200 to 600 km, and of limited duration, lasting up to a day or less. Coastal fronts often result in bands of precipitation, parallel to the front, with the maximum precipitation occurring on the cold side, and possibly a transition from snow to sleet to freezing rain to rain on the warm side.

Occluded Fronts

Traditionally, occluded fronts were considered to be the end products of the evolution of extratropical cyclones, formed by the catch-up of a slower-moving warm front by a faster-moving cold front. The occluded front is denoted on the surface map by alternating triangles and semicircles that point in the direction of movement (Figure 1). Two kinds of occluded fronts were hypothesized to exist: warm-type occlusions and cold-type occlusions. Warm-type occlusions were hypothesized to form if the air ahead of the warm front was colder than the air behind the cold front, lifting the merging cold front over the warm front. In contrast, cold-type occlusions were hypothesized to form if the air ahead of the warm front was warmer than the air behind the cold front, lifting the merging warm front over the cold front.

This traditional explanation does not adequately explain the structure and evolution of occluded fronts. First, much of the length of some occluded fronts is formed, not by the catch up of fronts, but by the lengthening of the front by deformation

due to differential rotation around the cyclone. Second, many cyclones continue to deepen after the formation of the occluded front. In fact, the heaviest precipitation and the strongest winds in the cyclone typically occur after the occluded front has formed. Third, few well-documented examples of cold-type occlusions exist; yet this is not explained by the traditional model.

To address these inadequacies, Schultz and Vaughan (2011) presented a more modern interpretation that ascribes occluded fronts to the wrap up of the cold and warm fronts around an extratropical cyclone. The extreme length of some occluded fronts is better characterized as the wrap up of the warm sector and occluded front than by the catch up of one front by another. In the wrap-up paradigm, the formation of an occluded front no longer determines the end of deepening of the cyclone, but is a result of the deepening. The deeper the cyclone, the more likely a cyclone will form an occluded front. The wrap-up paradigm also explains why many weak cyclones do not form occluded fronts.

Also in the wrap-up paradigm, the two kinds of occluded fronts are not determined by the relative temperatures across the occluded front, but by their relative static stabilities (Stoeckling et al., 2002). Warm-type occlusions form if the warm frontal zone is more stable than the cold frontal zone, and cold-type occlusions form if the cold frontal zone is more stable than the warm frontal zone. Because warm frontal zones tend to be much more stable than warm frontal zones, warm-type occluded fronts would be more common, explaining the relative dearth of cold-type occlusions.

Closely related to an occluded front is a back-bent front (e.g., Bjerknes, 1930; Bergeron, 1937; Shapiro and Keyser, 1990). This front is a warm or occluded front that has been wrapped around the cyclone, enclosing a region of relatively warm postcold frontal air. A region of strong winds known as the sting jet is sometimes found along a back-bent front (Grønås, 1995; Browning, 2004; Schultz and Sienkiewicz, 2013).

Dryline

The dryline is a zone of strong dew-point temperature gradient that forms in the southern United States during the warm season, separating moist air originating from the Gulf of Mexico from drier air originating from the southwest United States (Schaefer, 1974, 1986; Hoch and Markowski, 2005). Although not a front per se, the dryline is a boundary between two synoptic-scale air masses and can sometimes possess frontal characteristics of a temperature gradient and wind shift (Ziegler and Hane, 1993; Buban et al., 2007). Like fronts, the strength of the dryline (magnitude of the gradient of the dew point temperature) can be regulated by deformation and convergence (Schultz et al., 2007). Temperature differences across the dryline can be enhanced by the differing moisture contents in the air (i.e., moist air cools down more slowly than dry air because of the greenhouse effect of the water vapor).

Upper-Level Fronts

An upper-level front is a transition zone exhibiting a sharp thermal contrast and wind shear that may extend from the tropopause down to as low as 2–3 km above the ground. The mature upper-level frontal structure displayed in Figure 2 shows a frontal zone depicted by the concentration of isentropes (lines of constant potential temperature). Upper-level fronts are approximately 50–200 km wide and hundreds of kilometers long. These fronts are not described as cold or warm fronts, although cold-air and warm-air advections often occur along its length (Keyser and Shapiro, 1986; Schultz and Doswell, 1999). Upper-level fronts are not usually associated with the type of weather that characterizes surface fronts. Clear-air turbulence and stratospheric–tropospheric exchange of air are often defining characteristics that distinguish upper-level fronts.

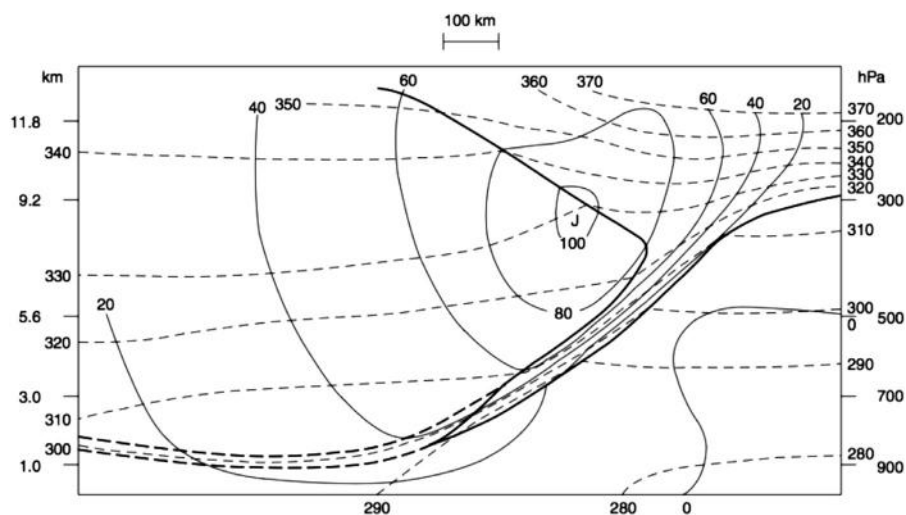


Figure 2 Cross section of an upper-level front. The tropopause is denoted by thick solid lines. The thin dashed lines are isentropes (K), and the thin solid lines are isotachs (m s^{-1}). J denotes the axis of the jet stream, directed out of the section. Pressure levels (hPa) and standard heights (km) are shown on the abscissa. Adapted with permission from Reed, R.J., 1955. A study of a characteristic type of upper-level frontogenesis. *Journal of Meteorology* 12, 226–237.

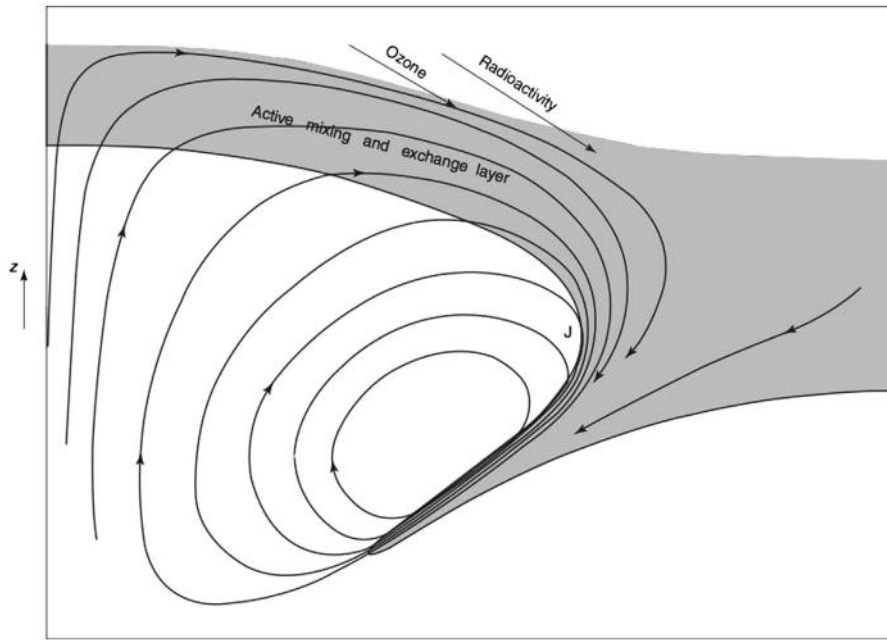


Figure 3 Schematic illustration of the transverse secondary circulation associated with an upper-level front as shown in [Figure 2](#). Adapted with permission from Danielson, E.F., 1968. Stratospheric–tropospheric exchange based on radioactivity, ozone and potential vorticity. *Journal of the Atmospheric Sciences* 25, 502–518.

Superposed on this front and jet-stream system is a transverse or cross-front circulation ([Eliassen, 1990](#); [Figure 3](#)). This circulation provides a consistent explanation of how the upper-level front is maintained and why stratospheric constituents, such as ozone and high-level radioactivity from nuclear explosions, can be observed in the lower troposphere. The subsiding branch maintains the most prominent characteristics of the upper-level front. In particular, the descent of stratospheric air within a narrow pocket can deform the tropopause into what is called a tropopause fold.

Thermal wind balance provides the explanation for the high-speed jet flow above the frontal zone. In this example, the sharp thermal gradient across the front produces a relatively strong jet that flows southward. Two characteristic features of an upper-level front are cyclonic shear and the temperature gradient. These two features are always associated with the development and enhancement of a prominent upper-level jet and frontal system, and with surface-based fronts that extend to the upper troposphere.

Surface Fronts Associated with Mesoscale Phenomena

Extratropical cyclones are not the only weather systems that organize temperature gradients and wind shifts into fronts; mesoscale phenomena can also do so. Small-scale fronts occupy a relatively limited horizontal domain, have relatively short lifetimes, and are surface-based phenomena. Because they exist only for a few minutes to a few hours, the Coriolis force is usually not dominant, but the nonhydrostatic acceleration may be important in the dynamics of these fronts.

Consequently, these fronts may take the form of a density current. In this section, three types of fronts associated with mesoscale phenomena are discussed: sea-breeze fronts, gust fronts associated with convective storms, and drainage fronts.

Sea-Breeze Fronts

Coastlines are prime regions for strong temperature differences to develop and take the form of fronts. Like the dryline, temperature gradients across the coast may develop in a diurnal cycle. Air masses over the water tend to heat up less during the day and cool down less during the night than adjacent air masses over the land. A pressure gradient, which develops in response to this differential heating, drives an onshore flow at low levels, with a return flow at about 1–2 km above the surface. The leading edge of the cooler air may develop frontal characteristics. The sea-breeze front is characterized by both a sharp temperature drop of several degrees centigrade or more and a marked increase in the humidity that can occur over a horizontal distance of a kilometer or less. Sea-breeze fronts are most prominent in the warm part of the year under relatively benign synoptic-scale flow. Inland penetration of the sea-breeze front by 10 km or more may be opposed by an offshore wind ahead of the front and by turbulent convective mixing over land, which tends to weaken the temperature and humidity gradient across the frontal zone.

Gust Fronts

Convective storms are associated with the ascent of warm, moist, unstable air. To maintain conservation of mass, descent must accompany the storms. The cooler descending air reaches

the surface and spreads out. The leading edge of this air is called a gust front and can frequently possess the characteristics of a front. The gust front develops from evaporative cooling associated with precipitation below large convective clouds. Its vertical extent is limited by the height of the cloud base, usually below 2 km. Temperature changes as high as 10° C over a few tens of meters can occur, and the wind gusts may represent a danger to aircraft that attempt to land on runways where gust fronts are evident. A gust front will last only minutes or tens of minutes if the cold air moves away from its source. It may, however, persist for a few hours or more if the cold air below the cloud base moves with the convective system. Relatively warm, moist air ahead of the front moves up and along the frontal surface to the cloud base in order to maintain the convection and precipitation that ultimately drives the gust front.

Drainage Front

A drainage front is the leading edge of downslope drainage of cold air from high elevations. Drainage currents develop at night, and the most favorable conditions for occurrence are usually met during the fall under clear skies and light winds. A pressure differential develops between the air over the slope and air at the same level over the level terrain. This pressure gradient force, together with gravity, drives a downslope cold current of air with a temperature differential at its leading edge. The distinguishing characteristics of the topography and the depth of radiatively cooled air along the slope will determine the frontal characteristics. Compressional heating during descent will also modify the cold temperatures behind the front, but temperature drops of 5° C are not uncommon after frontal passage. Humidity changes are usually not considered a significant factor in this type of front, but relatively light and gusty winds of about 5 m s⁻¹ are often encountered.

See also: Mesoscale Meteorology: Density Currents. Synoptic Meteorology: Extratropical Cyclones; Frontogenesis.

References

- Baines, P.G., 1980. The dynamics of the southerly buster. *Australian Meteorological Magazine* 28, 175–200.
- Bergeron, T., 1937. On the physics of fronts. *Bulletin of the American Meteorological Society* 18, 265–275.
- Bergeron, T., 1959. Methods in scientific weather analysis and forecasting: an outline in the history of ideas and hints at a program. In: Bolin, B. (Ed.), *The Atmosphere and Sea in Motion: Scientific Contributions to the Rossby Memorial Volume*. Rockefeller Institute Press, pp. 440–474.
- Bjerknes, J., 1919. On the structure of moving cyclones. *Geofysiske Publikasjoner* 1 (2), 1–8.
- Bjerknes, J., 1930. Practical examples of polar-front analysis over the British Isles in 1925–6. *Geophysical Memoirs* 5 (10), 1–21 and 28 pp. of figs.
- Bjerknes, J., Solberg, H., 1922. Life cycle of cyclones and the polar front theory of atmospheric circulation. *Geofysiske Publikasjoner* 3 (1), 3–18.
- Bosart, L.F., 1975. New England coastal frontogenesis. *Quarterly Journal of the Royal Meteorological Society* 101, 957–978.
- Bosart, L.F., Pagnotti, V., Lettau, B., 1973. Climatological aspects of eastern United States back-door cold frontal passages. *Monthly Weather Review* 101, 627–635.
- Bosart, L.F., Vaudo, C.J., Helsdon Jr., J.H., 1972. Coastal frontogenesis. *Journal of Applied Meteorology* 11, 1236–1258.
- Bosart, L.F., Wasula, A.C., Drag, W.H., Meier, K.W., 2008. Strong surface fronts over sloping terrain and coastal plains. *Synoptic-Dynamic Meteorology and Weather Analysis and Forecasting: A Tribute to Fred Sanders. Meteorologist Monographs*, No. 55. American Meteorological Society, pp. 35–85.
- Browning, K.A., 2004. The sting at the end of the tail: damaging winds associated with extra-tropical cyclones. *Quarterly Journal of the Royal Meteorological Society* 130, 375–399.
- Buban, M.S., Ziegler, C.L., Rasmussen, E.N., Richardson, Y.P., 2007. The dryline on 22 May 2002 during IHOP: ground-radar and in situ data analyses of the dryline and boundary layer evolution. *Monthly Weather Review* 135, 2473–2505.
- Colby Jr., F.P., Seitter, K.L., 1987. A new analysis technique for fronts. *Extended Abstracts, Third Conf. on Mesoscale Meteorology*. American Meteorological Society, Vancouver, BC, Canada, pp. 156–157.
- Davies, H.C., 1997. Emergence of the mainstream cyclogenesis theories. *Meteorologische Zeitschrift* 6, 261–274.
- Doswell III, C.A., Haugland, M.J., 2007. A comparison of two cold fronts—Effects of the planetary boundary layer on the mesoscale. *Electronic Journal Severe Storms Meteorology* 2 (4), 1–12.
- Eliassen, A., 1990. Transverse circulations in frontal zones. In: Newton, C., Holopainen, E.O. (Eds.), *Extratropical Cyclones: The Erik Palmén Memorial Volume*. American Meteorological Society, pp. 155–165.
- Emanuel, K.A., 2008. Back to Norway: An essay. *Synoptic-Dynamic Meteorology and Weather Analysis and Forecasting: A Tribute to Fred Sanders. Meteorologist Monographs*, No. 55. American Meteorological Society, pp. 87–96.
- Friedman, R.M., 1989. Appropriating the Weather: Vilhelm Bjerknes and the Construction of a Modern Meteorology. Cornell University Press, 251 pp.
- Grønås, S., 1995. The seclusion intensification of the New Year's day storm 1992. *Tellus* 47A, 733–746.
- Hakim, G.J., 1992. The eastern United States side-door cold front of 22 April 1987: a case study of an intense atmospheric density current. *Monthly Weather Review* 120, 2738–2762.
- Hewson, T.D., 1998. Objective fronts. *Meteorological Applications* 5, 37–65.
- Hoch, J., Markowski, P.M., 2005. A climatology of springtime dryline position in the United States Great Plains region. *Journal of Climate* 18, 2132–2137.
- Jewell, R., 1981. Tor Bergeron's first year in the Bergen school: towards an historical appreciation. In: Liljequist, G.H. (Ed.), *Contributions to Current Research in Geophysics, Weather and Weather Maps: A Volume Dedicated to the Memory of Tor Bergeron*, vol. 10. Birkhäuser Verlag, pp. 474–490.
- Keyser, D., 1986. Atmospheric fronts: An observational perspective. *Mesoscale Meteorology and Forecasting*. P. S. Ray, Ed. American Meteorological Society, 216–258.
- Keyser, D., Shapiro, M.A., 1986. A review of the structure and dynamics of upper-level frontal zones. *Monthly Weather Review* 114, 452–499.
- Kutzbach, G., 1979. The Thermal Theory of Cyclones. American Meteorological Society, 255 pp.
- Lackmann, G., 2011. Midlatitude Synoptic Meteorology: Dynamics, Analysis & Forecasting. American Meteorological Society, 345.
- Mass, C.F., 1991. Synoptic frontal analysis: time for a reassessment? *Bulletin of the American Meteorological Society* 72, 348–363.
- Mass, C.F., Schultz, D.M., 1993. The structure and evolution of a simulated midlatitude cyclone over land. *Monthly Weather Review* 121, 889–917.
- Neiman, P.J., Shapiro, M.A., Fedor, L.S., 1993. The life cycle of an extratropical marine cyclone. Part II: Mesoscale structure and diagnostics. *Monthly Weather Review* 121, 2177–2199.
- Newton, C.W., Newton, H.R., 1999. The Bergen school concepts come to America. In: Shapiro, M.A., Grønås, S. (Eds.), *The Life Cycles of Extratropical Cyclones*. American Meteorological Society, pp. 41–59.
- Novak, D.R., Bosart, L.F., Keyser, D., Waldstreicher, J.S., 2004. An observational study of cold season-banded precipitation in northeast U.S. cyclones. *Weather Forecasting* 19, 993–1010.
- Pettersen, S., 1936. Contribution to the theory of frontogenesis. *Geofysiske Publikasjoner* 11 (6), 1–27.
- Sanders, F., 1955. An investigation of the structure and dynamics of an intense surface frontal zone. *Journal of Meteorology* 12, 542–552.
- Sanders, F., 1999. A proposed method of surface map analysis. *Monthly Weather Review* 127, 945–955.
- Sanders, F., 2005. Real front or baroclinic trough? *Weather Forecasting* 20, 647–651.
- Sanders, F., Doswell III, C.A., 1995. A case for detailed surface analysis. *Bulletin of the American Meteorological Society* 76, 505–521.
- Sanders, F., Hoffman, E.G., 2002. A climatology of surface baroclinic zones. *Weather Forecasting* 17, 774–782.

- Sanders, F., Kessler, E., 1999. Frontal analysis in the light of abrupt temperature changes in a shallow valley. *Monthly Weather Review* 127, 1125–1133.
- Schaefer, J.T., 1974. The life cycle of the dryline. *Journal of Applied Meteorology* 13, 444–449.
- Schaefer, J.T., 1986. The dryline. In: Ray, P.S. (Ed.), *Mesoscale Meteorology and Forecasting*. American Meteorological Society, pp. 549–572.
- Schultz, D.M., 2004. Cold fronts with and without prefrontal wind shifts in the central United States. *Monthly Weather Review* 132, 2040–2053.
- Schultz, D.M., 2005. A review of cold fronts with prefrontal troughs and wind shifts. *Monthly Weather Review* 133, 2449–2472.
- Schultz, D.M., 2008. Perspectives on Fred Sanders' research on cold fronts. *Synoptic-Dynamic Meteorology and Weather Analysis and Forecasting: A Tribute to Fred Sanders*. Meteor. Monogr., No. 55. American Meteorological Society, pp. 109–126.
- Schultz, D.M., Doswell III, C.A., 1999. Conceptual models of upper-level frontogenesis in south-westerly and north-westerly flow. *Quarterly Journal of the Royal Meteorological Society* 125, 2535–2562.
- Schultz, D.M., Roebber, P.J., 2008. The fiftieth anniversary of sanders (1955): A mesoscale model simulation of the cold front of 17–18 April 1953. *Synoptic-Dynamic Meteorology and Weather Analysis and Forecasting: A Tribute to Fred Sanders*. Meteor. Monogr., No. 55. American Meteorological Society, pp. 126–143.
- Schultz, D.M., Sienkiewicz, J.M., 2013. Using frontogenesis to identify sting jets in extratropical cyclones. *Weather Forecasting* 28, 603–613.
- Schultz, D.M., Vaughan, G., 2011. Occluded fronts and the occlusion process: A fresh look at conventional wisdom. *Bulletin of the American Meteorological Society* 92, 443–466. ES19–ES20.
- Schultz, D.M., Bracken, W.E., Bosart, L.F., Hakim, G.J., Bedrick, M.A., Dickinson, M.J., Tyle, K.R., 1997. The 1993 superstorm cold surge: Frontal structure, gap flow, and tropical impact. *Monthly Weather Review* 125, 5–39. Corrigenda, 125, 662.
- Schultz, D.M., Bracken, W.E., Bosart, L.F., 1998. Planetary-and synoptic-scale signals associated with Central American cold surges. *Monthly Weather Review* 126, 5–27.
- Schultz, D.M., Weiss, C.C., Hoffman, P.M., 2007. The synoptic regulation of dryline intensity. *Monthly Weather Review* 135, 1699–1709.
- Shapiro, M.A., Keyser, D., 1990. Fronts, jet streams and the tropopause. In: Newton, C.W., Holopainen, E.O. (Eds.), *Extratropical Cyclones, the Erik Palmén Memorial Volume*. American Meteorological Society, pp. 167–191.
- Stoelinga, M.T., Locatelli, J.D., Hobbs, P.V., 2002. Warm occlusions, cold occlusions, and forward-tilting cold fronts. *Bulletin of the American Meteorological Society* 83, 709–721.
- Sutcliffe, R.C., 1952. Principles of synoptic weather forecasting. *Quarterly Journal of the Royal Meteorological Society* 78, 291–320.
- Uccellini, L.W., Corfidi, S.F., Junker, N.W., Kocin, P.J., Olson, D.A., 1992. Report on the surface analysis workshop held at the National Meteorological Center 25–28 March 1991. *Bulletin of the American Meteorological Society* 73, 459–472.
- Volkert, H., 1999. Components of the Norwegian cyclone model: Observations and theoretical ideas in Europe prior to 1920. In: Shapiro, M.A., Grønås, S. (Eds.), *The Life Cycles of Extratropical Cyclones*. American Meteorological Society, pp. 15–28.
- Wang, P.-Y., Martin, J.E., Locatelli, J.D., Hobbs, P.V., 1995. Structure and evolution of winter cyclones in the central United States and their effects on the distribution of precipitation. Part II: Arctic fronts. *Monthly Weather Review* 123, 1328–1344.
- Young, G.S., Fritsch, J.M., 1989. A proposal for general conventions in analyses of mesoscale boundaries. *Bulletin of the American Meteorological Society* 70, 1412–1421.
- Ziegler, C.L., Hane, C.E., 1993. An observational study of the dryline. *Monthly Weather Review* 121, 1134–1151.

Fronts in the Lower Stratosphere

AL Lang, University of Albany – State University of New York, Albany, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

A jet streak located along the sloping midlatitude tropopause has mesoscale structure in both the upper troposphere and lower stratosphere. This article explores the historical to modern day observation, theory, and dynamical insight into the structure and evolution of the lower stratospheric portion of the tropopause jet, known as the lower stratospheric front.

Introduction

Tropopause jet-front systems are transient yet pervasive dynamic and thermodynamic features are found in the vicinity of the sloping midlatitude tropopause. These systems, also known as upper-level jet-front systems, are characterized by a local wind speed maximum (i.e., the jet core) along the tropopause surface. Through the thermal wind relationship, the regions of vertical shear located above and below the jet core are directly associated with attendant baroclinic zones (i.e., the fronts). A characteristic example of the structure of a tropopause jet-front system is represented by the vertical cross section taken across the direction of the flow depicted in [Figure 1](#). As deep, three-dimensional, features centered about the tropopause, tropopause jet-front systems have significant vorticity and thermal structures residing both in the upper troposphere (upper tropospheric front, labeled 'UT Front' in [Figure 1](#)) and in the lower stratosphere (lower stratospheric front, labeled 'LS Front' in [Figure 1](#)). The majority of research attention regarding tropopause jet-front systems has been focused on the upper tropospheric portions of these systems, while a disproportionally small fraction of research attention has been given to the portions of these structures that reside above the jet core, within the lower stratosphere. Since the mid-twentieth century, the structure, evolution, and life cycle of these tropopause jet-front systems have been studied in a variety of contexts, including clear air turbulence, stratosphere–troposphere exchange, and chemical transport, as well as the extratropical cyclone life cycle.

This article will provide a historical perspective on, as well as our current understanding of, lower stratospheric portion of tropopause jet-front systems. This article is structured in the following manner. In order to fully explore the relevance of the lower stratospheric portion of tropopause jet-front systems, a background in the upper tropospheric portion of these systems is provided in [Section Background](#). A historical perspective of research on the lower stratospheric portion of the midlatitude jet is provided in [Section Historical Perspective](#). An examination of the typical structure and evolution of a lower stratospheric front within a baroclinic wave is presented in [Section Structure and Evolution](#). [Section The Role of Tropospheric Latent Heating in Lower Stratospheric Frontogenesis](#) provides details on the role of tropospheric latent heating on lower stratospheric fronts. Finally, [Section The Role of the Midlatitude Lower Stratospheric Structure in Large-Scale Circulations](#) explores a large-scale view of the role of the

lower stratospheric front by considering them as they are represented in a zonally averaged perspective.

Background

Much of what is currently known about the role of the tropopause jet-front system in weather and climate has centered on the life cycle and associated implications of the upper tropospheric component of these systems, known as upper-level fronts or upper tropospheric fronts. The process of upper tropospheric frontogenesis has been shown to be an important process in the development and life cycle of surface cyclones. The ageostrophic transverse vertical circulation associated with a dynamically active tropopause jet-front system not only provides the cross-stream differential subsidence required to intensify the upper tropospheric front, but also simultaneously steepens and lowers the dynamic tropopause below the jet core. This process forces a thin wedge of stratospheric air, characterized by large (small) values of potential vorticity (PV, water vapor), downward into the upper and middle troposphere. Within the upper troposphere, the large values of PV are manifested as a local maximum in absolute vorticity. The local maximum in absolute vorticity within the developing upper tropospheric front can subsequently provide a precursor disturbance to a wave scale surface cyclogenesis event. Thus, the development of an upper tropospheric vorticity maximum in the upper tropospheric frontogenesis process implicates the tropopause jet-front systems in the incipient stages of the extratropical cyclone life cycle.

Additionally, the folding of the dynamic tropopause in association with intense upper tropospheric frontogenesis is one of the most efficient and dominant forms of stratosphere–troposphere exchange in the midlatitudes. In the 1960s, the US Defense Atomic Support Agency and the Atomic Energy Commission supported a field campaign to document the pathway of radioactive fallout from the stratospheric atomic weapons testing. The campaign, known as Project Springfield, confirmed the theoretical role of individual tropopause jet-front systems, specifically the upper tropospheric front, in stratosphere–troposphere exchange. The substantial displacements of the dynamic tropopause on isentropic surfaces that occur within tropopause folds facilitate the exchange of mass across the tropopause. As a result of the synoptic and mesoscale processes in the vicinity of a tropopause fold, dry, ozone-rich stratospheric air descends into the troposphere and can

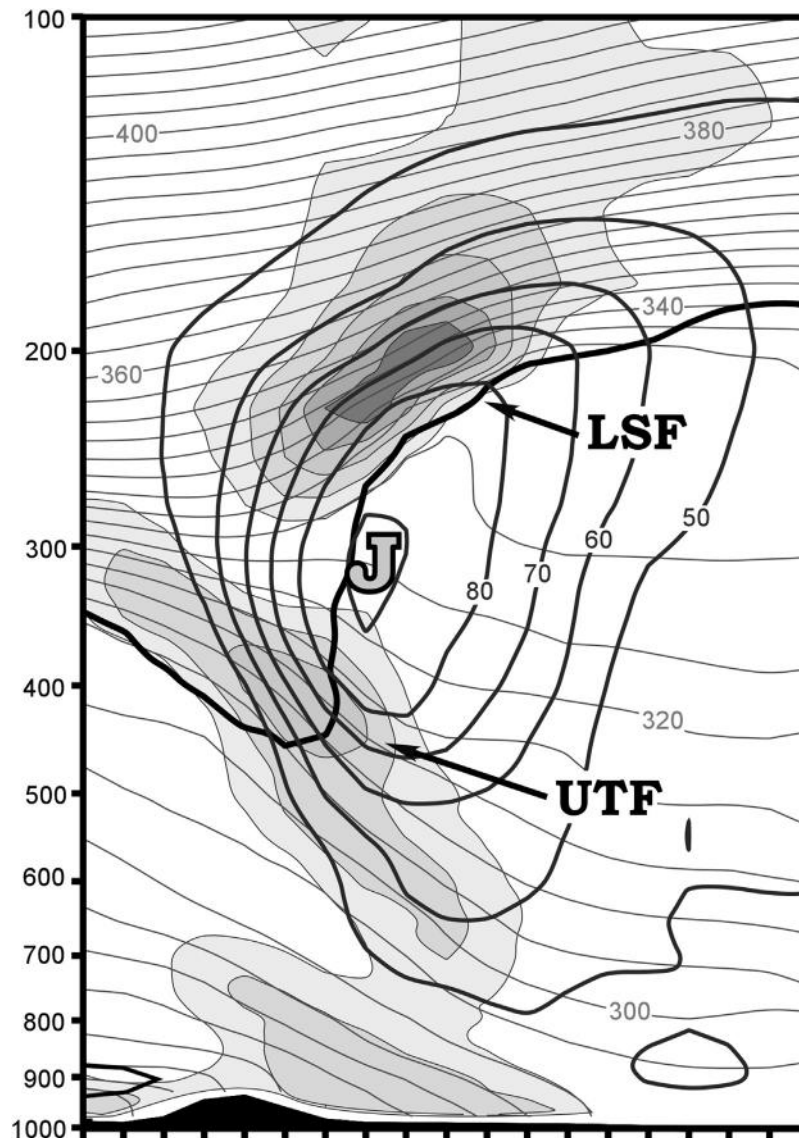


Figure 1 Cross section the North Atlantic jet off the east coast of North America at 1200 UTC 27 February 2008, highlighting the jet core (marked with a 'J') and associated frontal structures. The upper tropospheric front is labeled 'UTF' and the lower stratospheric front is labeled 'LSF.' Isotachs are thick lines every 10 m s^{-1} beginning at 40 m s^{-1} , isentropes are thin lines every 4 K , and the magnitude of the horizontal potential temperature gradient is filled every $1 \text{ K (100 km)}^{-1}$ beginning at $2 \text{ K (100 km)}^{-1}$. Reproduced with permission from Lang, A.A., Martin, J.E., 2013 The structure and evolution of lower stratospheric frontal zones. Part II: The influence of tropospheric convection on lower stratospheric frontal development. Quarterly Journal of the Royal Meteorological Society 139, 1798–1809.

irreversibly mix with tropospheric air. It is hypothesized that at least 10% of global tropospheric ozone originates from the stratosphere via intense upper tropospheric frontogenesis and tropopause folding.

Historical Perspective

Observation

In one of the first conceptual models of a jet-front system, Swedish meteorologist Roy Berggren proposed that a continuous frontal zone (at the time, known as the polar front)

stretched upward from the surface, through the upper troposphere and into the lower stratosphere (Figure 2). The controversial model applied the concept of a 'front' to upper air observations from the high spatial and temporal resolution European sounding network. The conceptual model suggested that in the upper troposphere, the 'front' was characterized by mesoscale gradients of temperature as well as cyclonic shear. At the level of maximum wind the 'front' was manifested solely as a mesoscale zone of cyclonic shear, while in the lower stratosphere, where the gradient of temperature reversed above the jet core, the 'front' once again was characterized by both temperature gradients and cyclonic shear. This proposed

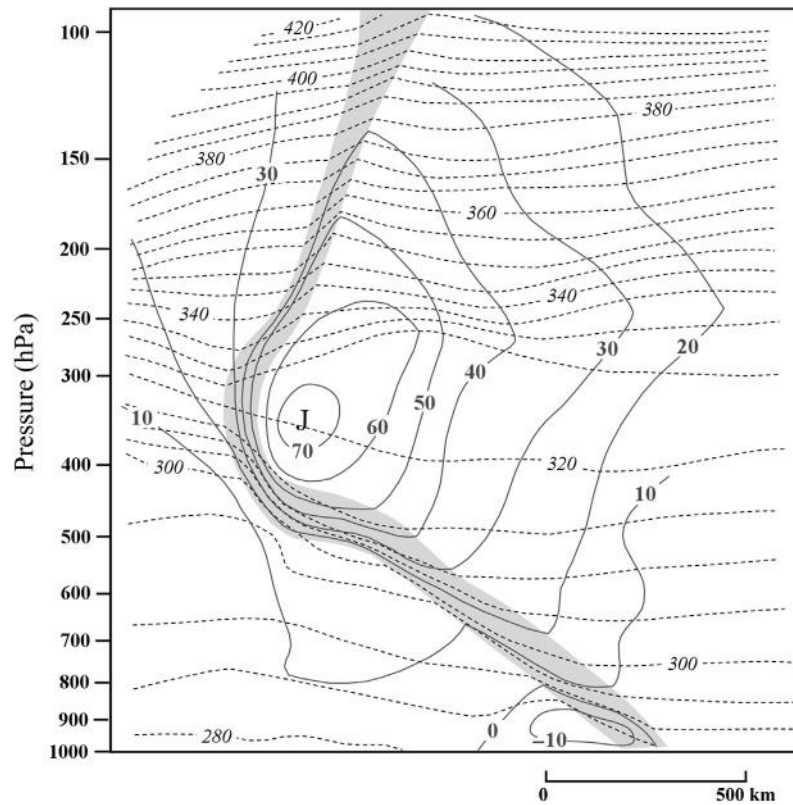


Figure 2 Vertical cross section of isotachs (m s^{-1} , solid) and isentropes (K, dashed) from Valentia, Ireland to Hannover, Germany at 0300 UTC 9 November 1949. The shaded area represents a continuous frontal zone as described by Berggren (1952). Adapted from Berggren, R., 1952. The distribution of temperature and wind connected with active tropical air in the higher troposphere and some remarks concerning clear air turbulence at high altitude. *Tellus* 4, 43–53 under the terms of the Creative Commons Attribution-Noncommercial 3.0 Unported License (<http://creativecommons.org/licenses/by-nc/3.0/>).

model, which included mesoscale structures at and above the jet core, was largely dismissed as lack of consistent balloon observation of the lower stratosphere. Berggren's conceptual model subsequently led researchers to question the scale of the vorticity and baroclinic structures in the lower stratosphere. Utilizing the (lower resolution) US sounding network, researchers argued that the lower stratospheric structure was not confined to the mesoscale, but rather was on the larger synoptic scale and thus did not fit the 'frontal' model. Subsequent research attention remain focused on the more consistently and conveniently observable upper tropospheric half of these systems. The focus on the upper tropospheric portion of the tropopause jet-front system led to a fruitful avenue of research explaining linkages of the jet with surface cyclogenesis, stratosphere–troposphere exchange, and clear air turbulence.

It was not until the 1960s, when instrumented research aircraft observations in field campaigns became a routinely available supplement to conventional radiosonde data, that a much-improved resolution of tropopause jet-front system structures was afforded. By the 1970s, high-resolution aircraft measurements confirmed that there was indeed a mesoscale (~ 100 km) confinement of cyclonic wind shear below, at, and above the jet core. This observationally based analysis

supported the Berggren conceptual model of a mesoscale 'frontal' structure extending from the upper troposphere through the level of maximum wind, to the lower stratosphere.

Theory

Though the notion of one continuous mesoscale frontal structure extending through the tropopause was suggested by observational studies, the theoretical and idealized studies of tropopause jet-front systems support the notion that the jet core vertically separates two separate centers of frontogenetic forcing. This idea of two separate centers of frontogenetic forcing was supported by the fact that true fronts are regions characterized by both vertical and horizontal shears as well as locally enhanced static stability. Thus, the baroclinic zones in the upper troposphere and lower stratosphere can be characterized as two separate fronts, while the cyclonic shear region at the jet level characterized as frontal in the Berggren model lacks the characteristics of a true frontal zone.

Early analyses of lower stratospheric thermal structure in the midlatitudes revealed that above the jet core the warmest and coldest air were arranged in elongated parallel bands, horizontally separated by the local wind speed maximum. This configuration of the environment above the level of maximum

wind led several researchers in the 1940s to hypothesize that the wind and thermal fields above the jet were connected with a cross-stream vertical circulation. By virtue of the strong static stability of the lower stratosphere, the vertical motion in the vicinity of the jet was hypothesized to have a rather substantial effect on the horizontal thermal field via adiabatic warming and cooling. Investigations into the characteristics of the thermal structure above the jet core subsequently led to research inquiries aimed at describing the nature of the vertical circulations in the vicinity of the jet. By the late 1940s, a hypothesis was presented which suggested that superimposed on the wave scale vertical motion between a ridge and downstream trough were two separate, cross-stream vertical circulations associated with the midlatitude jet; one in the lower stratosphere above the jet core and another in the upper troposphere below the jet core (Figure 3). The upward and downward branches of these circulations, both in the upper troposphere and in the lower stratosphere, could independently act to enhance or weaken the wave-scale vertical motion associated with an upper-level trough. These early studies suggested that a combination of the jet circulations and the wave scale vertical motions would have a substantial effect on the production of sensible weather.

The pioneering work of the US synoptic meteorologists Richard Reed and Fred Sanders in the mid-1950s established that differential vertical motions near the jet were at the heart of upper tropospheric frontal development. In the vicinity of the tropopause, the differential tilting of isentropes associated with cross-stream gradients of vertical motion could create sloping stable layers, with roots in the lower stratosphere, that were characterized by large horizontal temperature gradients. The differential vertical motion would also act to tilt horizontally oriented vortex tubes (associated with the vertical shear) into a more vertical orientation, thus enhancing the vertical vorticity within the developing upper-level front.

In his extension of John S. Sawyer's study on jet circulation dynamics, Arnt Eliassen derived a mathematical formulation to calculate the important vertical circulation in the vicinity of the fronts. This equation, which has since become known as the Sawyer–Eliassen equation, was the first to make the mathematical connection between the primary forcing captured by geostrophic frontogenesis and the creation of a secondary

ageostrophic circulation in the vicinity of a frontal zone. Their equation,

$$2\gamma \left(-\frac{\partial u_g}{\partial x} \frac{\partial \theta}{\partial y} + \frac{\partial u_g}{\partial y} \frac{\partial \theta}{\partial x} \right) = -\gamma \frac{\partial \theta}{\partial x} \frac{\partial^2 \psi}{\partial y^2} + 2 \frac{\partial m}{\partial p} \frac{\partial^2 \psi}{\partial y \partial p} - \frac{\partial m}{\partial y} \frac{\partial^2 \psi}{\partial p^2}, \quad [1]$$

shows that in a coordinate system where the x -direction is oriented along a baroclinic zone, the geostrophic (subscript g) shearing and stretching deformation acting on a baroclinic zone (left-hand side) would force a secondary ageostrophic cross-stream circulation (stream function, ψ , in the right-hand side). In their equation, m is the geostrophic momentum ($u_g - fy$), $\gamma = \frac{R}{fP_0} \left(\frac{P_0}{P} \right)^{c_p/c_p}$ and is constant on a pressure (P) surface, $\frac{\partial \psi}{\partial y} = \omega$ and $-\frac{\partial \psi}{\partial p} = v_{ag}$. The application of the Sawyer–Eliassen equation suggested that frontogenesis was a two-step process in which (1) the deformation in the primary geostrophic flow provides forcing to tighten the temperature gradient, resulting in a secondary ageostrophic circulation, then (2) that secondary ageostrophic circulation potentially acts to further intensify the temperature gradient. The original analysis by Eliassen explored the idea that the geostrophic deformation in the vicinity of a tropopause jet-front system resulted in forcing for two separate secondary circulation cells, one above and one below the jet core. These two ageostrophic circulations were centered about the baroclinic zones lying above and below the level of maximum wind.

The secondary circulations associated with idealized variations of the geostrophic shearing and stretching deformation forcing in the vicinity of the tropospheric half of the tropopause jet-front system have also been considered in conceptual analyses. In the absence of temperature advection (e.g., a case of pure stretching deformation), the Sawyer–Eliassen circulations in the jet entrance and exit regions result in the traditional four-quadrant model; a thermally direct (indirect) circulation in the entrance (exit) region (Figure 4(a)). When geostrophic cold air advection was present through the jet core, the thermally direct (indirect) circulation in the jet entrance (exit) region was shifted toward the anticyclonic (cyclonic) side of the jet, so that subsidence characterized the jet core (Figure 4(b)). Such

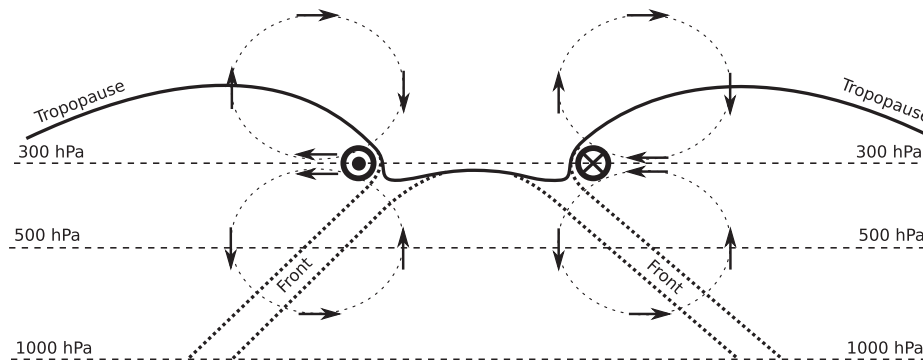


Figure 3 An idealized cross section through a synoptic-scale trough with the cross-stream vertical circulations (arrows) superimposed on the idealized locations of the polar front (dashed thick lines) and tropopause (thick line). The schematic cross section represents a composite of mid-twentieth century theories on jet circulations in the vicinity of the tropopause.

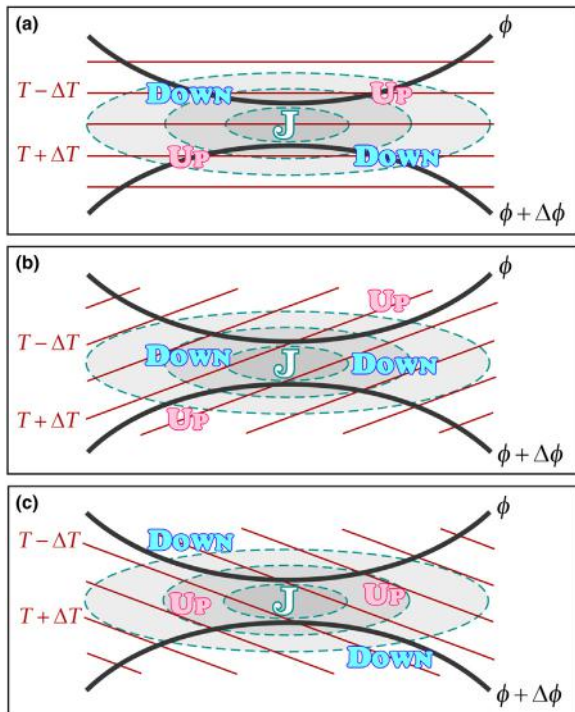


Figure 4 Schematic illustration of idealized configurations of potential temperature along a straight jet streak maximum on an upper-tropospheric isobaric surface. Geopotential height (thick solid), potential temperature (dashed), isotachs (thin solid filled) with the jet maximum represented by 'J,' and a sense of the midtropospheric Sawyer–Eliassen vertical motions (up or down) for (a) no thermal advection along the jet, (b) geostrophic cold air advection along the jet, and (c) geostrophic warm air advection along the jet. Reproduced with permission from Lang, A.A., Martin, J.E., 2012. The structure and evolution of lower stratospheric frontal zones. Part I: Examples in northwesterly and southwesterly flow. *Quarterly Journal of the Royal Meteorological Society* 138, 1350–1365.

a distribution of subsidence is consistent with that attributed to negative vorticity advection by the thermal wind, which exists along the jet axis when geostrophic cold air advection is present along a tropopause jet-front system. Subsidence through the jet core is favorable for upper-tropospheric frontogenesis, thus locally enhanced subsidence resulting from the establishment of geostrophic cold air advection along the jet (termed the Shapiro Effect) has been described as an important aspect of the upper-front life cycle. In contrast, an environment characterized by geostrophic warm air advection through the jet core shifts the thermally direct (indirect) circulation in the jet entrance (exit) region toward the cyclonic (anticyclonic) side, so that the jet core is conversely characterized by ascent (Figure 4(c)).

Figure 5 extends the Figure 4, a conceptual model to the stratospheric half of a tropopause jet-front system, where the geostrophic vertical shear and horizontal temperature gradient are reversed above the jet core, and illustrates the vertical motions forced by the idealized geostrophic stretching and shearing deformation above the level of maximum winds. An environment of pure geostrophic stretching deformation, illustrated as a straight jet streak in the absence of thermal advection, is shown in Figure 5(a). Resembling the traditional tropospheric

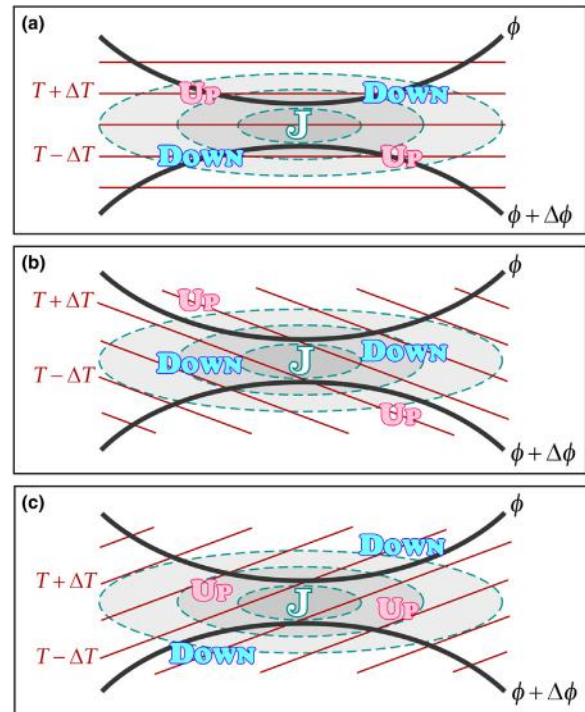


Figure 5 Schematic illustration of idealized configurations of potential temperature along a straight jet streak maximum on a lower stratospheric isobaric surface. Geopotential height (thick solid), potential temperature (dashed), isotachs (thin solid filled) with the jet maximum represented by 'J,' and a sense of the Sawyer–Eliassen vertical motions (up or down) above the jet core for (a) no thermal advection along the jet, (b) geostrophic cold air advection along the jet, and (c) geostrophic warm air advection along the jet. Reproduced with permission from Lang, A.A., Martin, J.E., 2012. The structure and evolution of lower stratospheric frontal zones. Part I: Examples in northwesterly and southwesterly flow. *Quarterly Journal of the Royal Meteorological Society* 138, 1350–1365.

four-quadrant model, the thermally direct entrance region and thermally indirect exit region are evident. However, in this lower stratospheric version of the four-quadrant model, the jet entrance (exit) region ascent is located on the cyclonic (anticyclonic) shear side of the jet, opposite to its location in the corresponding environment below the jet core. Figure 5(b) shows the resulting circulations forced by geostrophic cold air advection along an idealized jet. In this case, the thermally direct (indirect) circulation shifts toward the cyclonic (anticyclonic) shear side of the entrance (exit) region. Despite this altered distribution, the same outcome is produced in the lower stratosphere as in the upper troposphere, subsidence through the local jet maximum. Finally, the case of geostrophic warm air advection along the jet is illustrated in Figure 5(c). As in the cold air advection case, though the entrance and exit region circulations shift in opposition to their counterparts below the jet core, the net result is a band of ascent through the local jet axis.

Structure and Evolution

Consideration of the lower stratospheric portion of a tropopause jet-front system as a distinct front encourages the

extension of many of the concepts of upper tropospheric frontal dynamics to regions above the level of maximum wind. The baroclinic zone located in the lower stratosphere above the jet core is best considered a frontal structure, with its own distinct frontal circulation, that separates stratospheric air from tropospheric air. Above the level of maximum winds, where the pole to equator temperature gradient reverses in accordance with thermal wind balance, the lower stratospheric front represents a boundary separating cold midlatitude upper tropospheric air from warmer midlatitude lower stratospheric air. Frontogenetic tilting across a lower stratospheric frontal zone results from a combination of subsidence maximized in the warm lower stratosphere and/or ascent maximized in the cold upper troposphere (e.g., a thermally indirect circulation). Analogous to the situation in the upper troposphere, these differential vertical motions will result in the deformation of the tropopause above the level of maximum wind, where positive tilting frontogenesis is associated with a more steeply sloped tropopause. Though the governing dynamics for their developments (tilting frontogenesis) are the same, the upper tropospheric and lower stratospheric fronts can develop asynchronously within the evolution of a tropopause jet-front system through the northwesterly and southwesterly flow portions of a baroclinic wave.

During the winter season, as a tropopause jet-front system emerges in northwesterly flow, there is a tendency for it to be characterized by a vigorous lower stratospheric frontal zone and a relatively weaker upper tropospheric frontal zone, as measured by the magnitude of the potential temperature gradients at 200 hPa (i.e., a level above the jet core) and 500 hPa (i.e., a level below the jet core), respectively. As a dynamically active tropopause jet-front system progressed through northwesterly flow, it tends to be associated with a lower stratospheric front characterized by geostrophic cold air advection in cyclonic shear (particularly in the jet entrance region). This circumstance is associated with subsidence on the cold side of the lower stratospheric frontal zone, which, in turn, decreases the intensity of the associated baroclinicity above, via tilting. This subsidence, associated with lower stratospheric frontal processes, protruded downward to sufficient depth to reach the warm side of the weak upper tropospheric front, contributing to the intensifying frontal characteristics of the attendant upper tropospheric front. Thus, forcing for subsidence within the decaying lower stratospheric frontal environment is hypothesized to be a contributing factor in an already favorable environment for the initial development of the upper tropospheric frontal portion of the tropopause jet-front system. Furthermore, as the lower stratospheric front weakens, the upper tropospheric front intensifies demonstrating a clearly asynchronous, yet dynamically interactive developmental relationship between the two component frontal zones of the tropopause jet-front system.

The evolution of lower stratospheric fronts in southwesterly flow can be notably different than their northwesterly flow counterparts. Case analyses have revealed that southwesterly flow lower stratospheric fronts tend to experience a period of intensification coincident with the weakening of their upper tropospheric counterpart. An intensifying lower stratospheric front in southwesterly flow is characterized by geostrophic warm air advection on the cyclonic shear side of the jet. Such

a circumstance promotes quasi-geostrophic conditions favorable for ascent through the jet core, which is located on the cold side of the lower stratospheric front. Such ascent is frontogenetic above the level of maximum winds, resulting in an increase of both the slope of the isentropes and the slope of the tropopause above the jet core. Ascent coincident with the lower stratospheric frontal processes is typically discernibly separate from the ascent associated with lower tropospheric frontal forcing. In special circumstances, the near-vertical superposition of a lower stratospheric front and a lower tropospheric front can create a tropospheric deep column of ascent that acts frontolytically in the middle and upper troposphere but frontogenetically in the lower stratosphere.

A conceptual model highlighting some common characteristics of northwesterly flow cases is presented in [Figure 6\(a\)](#). In northwesterly flow, the lower stratospheric frontal zone experiences frontolysis while its upper tropospheric counterpart experiences frontogenesis when the lower stratosphere in the vicinity of the tropopause jet-front system is characterized by geostrophic cold air advection in cyclonic shear. Attendant with the upper tropospheric frontogenesis is an extrusion of stratospheric PV and O₃ into the upper troposphere. The extruded stratospheric PV is manifest as an upper tropospheric absolute vorticity maximum that can serve as a precursor to surface cyclogenesis. Conversely, in southwesterly flow characterized by lower stratospheric geostrophic warm air advection in cyclonic shear, the upper tropospheric frontal zone weakens while its lower stratospheric counterpart intensifies via quasi-geostrophic ascent through the jet core ([Figure 6\(b\)](#)). As a result, it is hypothesized that tropospheric water vapor can be exported into the midlatitude lower stratosphere and low values of PV, from the lower troposphere, flood the upper troposphere/lower stratosphere on the anticyclonic shear side of a strengthening southwesterly jet. This process can lead to the amplification of the ridge at upper tropospheric levels and to anticyclogenesis in the lower troposphere downstream of the intensified jet.

It strongly suggests that consideration of the separate evolution of the lower stratospheric and upper tropospheric frontal zones associated with a tropopause jet-front system can lead to a better understanding of the comprehensive life cycle of a tropopause jet-front system through a baroclinic wave. It is left to future research to investigate how lower stratospheric frontal processes are explicitly connected to upstream or downstream sensible weather. For example, it has been hypothesized that lower stratospheric geostrophic warm air advection in cyclonic shear consistently tied to ridge building and blocking, in the way that geostrophic cold air advection in the cyclonic shear within an upper tropospheric front (e.g., the Shapiro Effect) is tied to upper-level trough and cyclone development.

The Role of Tropospheric Latent Heating in Lower Stratospheric Frontogenesis

Associated with their position in a baroclinic wave, there is a tendency for the near-superposition of southwesterly flow lower stratospheric fronts and the lower tropospheric/surface fronts associated with mature or occluded midlatitude cyclones. When the superposition of these two distinctly

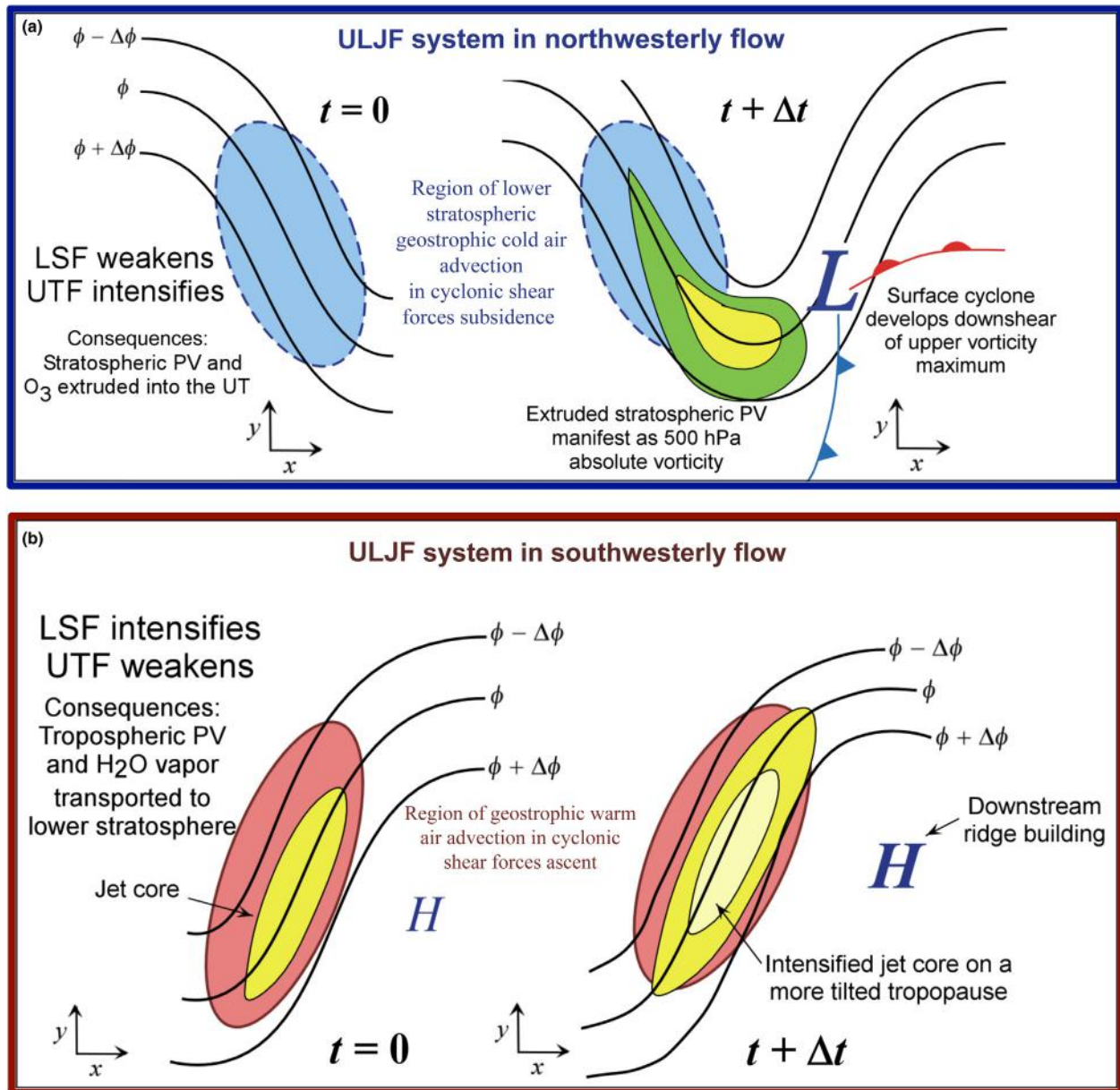


Figure 6 A conceptual model of the evolution of a tropopause jet-front system in the northwesterly (a) and southwesterly (b) flow portions of a baroclinic wave, see text for details. Reproduced with permission from Lang, A.A., Martin, J.E., 2012. The structure and evolution of lower stratospheric frontal zones. Part I: Examples in northwesterly and southwesterly flow. Quarterly Journal of the Royal Meteorological Society 138, 1350–1365.

separate fronts occurs, the lower tropospheric/surface front has the potential to impact lower stratospheric frontogenesis via diabatic processes. In such a situation, the diabatic impact resulting from tropospheric frontal ascent and/or convection ahead of a surface front (e.g., reduced upper tropospheric static stability) is positioned so as to favor a robust response to lower stratospheric frontogenetic processes promoting ascent on the cold side of the lower stratospheric frontal zone.

Similarly, to the extent that lower stratospheric and lower tropospheric frontal circulations are vertically aligned, the kinetic energy released in any convective or nonconvective ascent along the tropospheric front can enhance the effect of

the frontogenetic tilting already present in the lower stratospheric frontal circulation. Latent heat release can further increase the slope of the dynamic tropopause associated with an already intensifying lower stratospheric front. In such cases, the diabatically generated outflow from latent heat release in regions of vigorous or persistent ascent, proximate to a tropopause jet-front system, produces a reduction in static stability in the vicinity of the jet core. This reduction in static stability is associated with a robust response from the dynamically forced frontal circulations in the vicinity of the extratropical tropopause and the lower stratospheric front. Cases of this sort tend to occur upstream of intensifying surface high-pressure system

and within southwesterly flow at upper levels. Consistent with a steepening tropopause, the horizontal PV gradient in the vicinity of the jet also intensified, having substantial implications on the strength of the midlatitude Rossby waveguide. Thus, the dynamics of the lower stratospheric front are influential in the sensible weather experienced at the surface, as they are intimately involved in shaping the structure of the tropopause and the midlatitude Rossby waveguide.

This coupling between the lower tropospheric and lower stratospheric processes is conceptually illustrated in Figure 7 and summarized below. As latent heat release associated with ascent linked to a lower tropospheric/surface frontal boundary reconfigures the potential temperature field within the upper troposphere, the static stability of the near-tropopause upper troposphere decreases. Such a distribution of the diabatic static stability tendency directly influences the magnitude of the quasi-geostrophic ascent on the cold side of the lower stratospheric front. The ascent that responds to the quasi-geostrophic processes, subsequently contributes to the intensification of a robust lower stratospheric front, and a notable steepening of the tropopause above the jet core, via tilting frontogenesis. In the southwesterly flow portion of a baroclinic wave, the interaction of lower tropospheric based ascent and upper tropospheric static stability is likely able to influence the development of lower stratospheric fronts and substantially alter the structure of a midlatitude jet-front system by increasing the response to lower stratospheric processes and associated ascent. Subsequently, the robust ascent maximum above the jet core can lead to an increase in the slope of not only the isentropes within the lower stratospheric front, but also the tropopause above the jet core, via tilting. When latent

heat release occurs in the vicinity of a tropopause jet-front with a dynamically active lower stratospheric front, the lower stratospheric frontal circulation may be able to enhance the mesoscale structure of the horizontal PV gradient (i.e., increase the tropopause slope) above the jet core.

The interactions between lower stratospheric frontal development and the reduction of static stability produced by *in situ* latent heating along a lower tropospheric/surface frontal zones suggest that there are a variety of mechanisms whereby poorly stratified air may be delivered to the upper troposphere to similarly influence the lower stratospheric development. A particularly intriguing possibility for such delivery lies in the outflow from organized tropical convection, manifest either as tropical convective clusters, tropical plumes or tropical cyclones (TCs), which processes low PV, low static stability boundary layer air in convective updrafts and exhausts such air into the upper troposphere. If a preexisting midlatitude jet becomes juxtaposed with such outflow, the convective outflow may be linked to a period of lower stratospheric frontogenesis associated with changes in the slope of the tropopause above the jet core. Such changes, tied as they are to increases in lower stratospheric baroclinicity via tilting from a thermally indirect circulation, must also be potentially energy generating and, thus, may provide a means by which energy released in the tropical convection might be stored for subsequent release downstream in space and time. A number of studies have recently highlighted a link between the convective TC environment and associated changes in the structure of the midlatitude jet. A recent analysis illustrated that a region of diabatically reduced PV and static stability is introduced into the vicinity of the tropopause jet-front system during

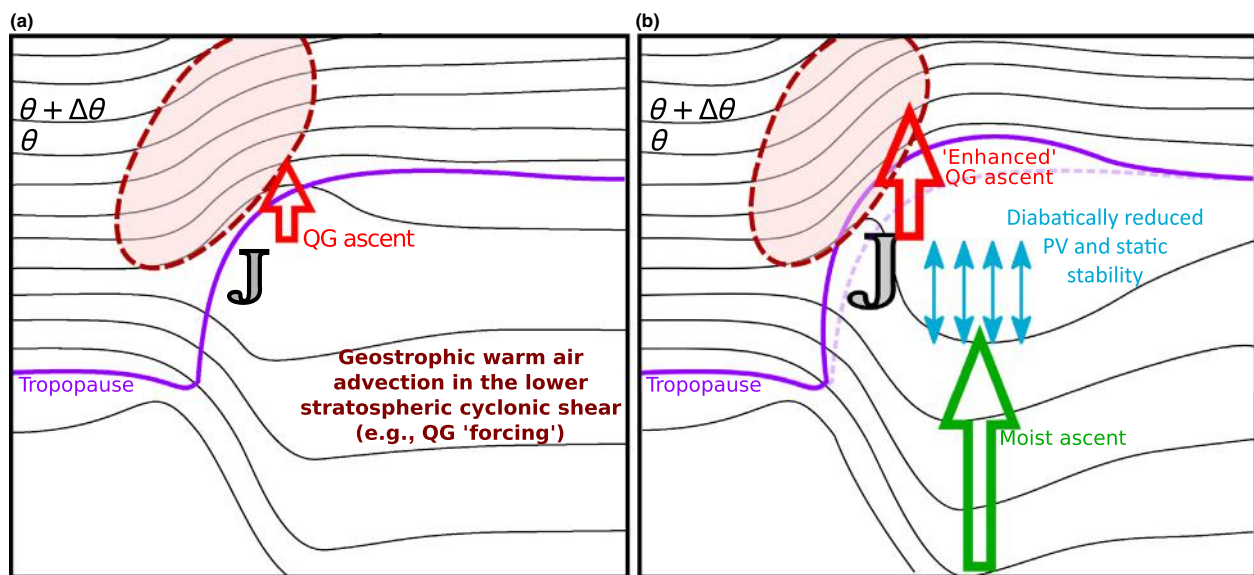


Figure 7 A schematic illustration of the interaction between diabatic heating and quasi-geostrophic processes in the vicinity of a lower stratospheric front. (a) A region of geostrophic warm air advection in the vicinity of the lower stratospheric front (shaded) is associated with modest QG ascent (arrow) above the jet core ('J'). (b) When moist ascent (large arrow) along a lower tropospheric front results in a diabatic reduction in static stability in the upper troposphere (filled arrows), the response to the geostrophic warm air advection in the vicinity of the lower stratospheric front (shaded) is a region of 'enhanced' QG ascent. The robust ascent is associated with a more steeply sloped tropopause (thick line) above the jet core and an increase in wind speed within the jet. Reproduced with permission from Lang, A.A., Martin, J.E., 2013. The structure and evolution of lower stratospheric frontal zones. Part II: The influence of tropospheric convection on lower stratospheric frontal development. Quarterly Journal of the Royal Meteorological Society 139, 1798–1809.

extratropical transition (ET) of a TC and can be associated with robust vertical circulations resulting in tilting frontogenesis above the jet core, along the lower stratospheric front. In addition, the vertical circulations provide a mechanism to tilt the tropopause into a more upright orientation (e.g., enhancing the horizontal PV gradient) that subsequently altered the mesoscale structure of the extratropical tropopause jet-front system during ET. Through the lower stratospheric front pathway, a TC undergoing ET can have substantial impacts on the intensity of the midlatitude jet circulations and the mesoscale structure and slope of the tropopause above the jet core. This highlights the lower stratospheric mechanism in which a recurving TC can have implications on the large scale by restructuring the midlatitude waveguide.

The Role of the Midlatitude Lower Stratospheric Structure in Large-Scale Circulations

In the scientific literature, the importance of the lower stratospheric portions of midlatitude tropopause jet-front systems has received the majority of its research attention in so far as it is manifested in a zonally averaged sense and on seasonal timescales. In a zonally averaged perspective, the zonally averaged meridional temperature gradient in the lower stratosphere, associated with the sloping midlatitude tropopause, is coupled (via thermal wind) to the zonal mean zonal wind field at and above the tropopause. The configuration of zonal mean zonal wind field above the tropopause determines the propagation characteristics of atmospheric waves into the stratosphere. Thus, the structure of the thermal and momentum fields above the tropopause governs the large-scale stratospheric circulation and the seasonal scale stratosphere–troposphere exchange accomplished via the wave-driven Brewer–Dobson circulation. From a zonally averaged perspective, the midlatitude lower stratospheric baroclinic structure plays a large role in the general circulation.

On shorter synoptic timescales the details of the structure, evolution, and dynamics of the lower stratospheric frontal portions of individual tropopause jet-front system systems are less understood. The changes to the thermal field at and above the tropopause associated with lower stratospheric frontogenesis are also associated with variation in the

structure of the vertical shear within the lower stratosphere. Thus, fully understanding lower stratospheric frontogenesis has further implications for the coupling between the stratosphere and troposphere by potentially explaining some of the variation in the environment governing the vertical propagation of atmosphere waves. It is suggested that consideration of lower stratospheric frontal processes will provide further insight into the not yet completely understood role of synoptic-scale baroclinic waves in transmitting the downward propagating anomalies, associated with the Northern Annular Mode, from the stratosphere to the troposphere and vice versa.

See also: Stratosphere/Troposphere Exchange and Structure: Tropopause. Synoptic Meteorology: Fronts; Jet Streaks.

Further Reading

- Berggren, R., 1952. The distribution of temperature and wind connected with active tropical air in the higher troposphere and some remarks concerning clear air turbulence at high altitude. *Tellus* 4, 43–53.
- Eliassen, A., 1962. On the vertical circulation in frontal zones. *Geofysiske Publikasjoner* 24, 147–160.
- Keyser, D., Shapiro, M.A., 1986. A review of the structure and dynamics of upper-level frontal zones. *Monthly Weather Review* 114, 452–499.
- Lang, A.A., Martin, J.E., 2012. The structure and evolution of lower stratospheric frontal zones. Part I: Examples in northwesterly and southwesterly flow. *Quarterly Journal of the Royal Meteorological Society* 138, 1350–1365.
- Lang, A.A., Martin, J.E., 2013. The structure and evolution of lower stratospheric frontal zones. Part II: The influence of tropospheric convection on lower stratospheric frontal development. *Quarterly Journal of the Royal Meteorological Society* 139, 1798–1809.
- Palmén, E., Nagler, K.M., 1949. The formation and structure of a large scale disturbance in the Westerlies. *Journal of Meteorology* 6, 228–242.
- Reed, R.J., 1955. A study of a characteristic type of upper-level frontogenesis. *Journal of Meteorology* 12, 226–237.
- Reed, R.J., Sanders, F., 1953. An investigation of the development of a mid-tropospheric frontal zone and its associated vorticity field. *Journal of Meteorology* 10, 338–349.
- Sawyer, J.S., 1956. The vertical circulation at meteorological fronts and its relation to frontogenesis. *Proceedings of the Royal Society of London A234*, 346–362.
- Shepherd, T.G., 2002. Issues in stratosphere-troposphere coupling. *Journal of the Meteorological Society of Japan* 80, 769–792.

Frontogenesis

DM (David) Schultz, University of Manchester, Manchester, UK

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

The expression 'frontogenesis' can be used in three ways. The first usage is colloquial, with frontogenesis used to describe the formation or intensification of a frontal zone. The second usage is mathematical, where the frontogenesis function is the Lagrangian rate of change of the potential temperature gradient. Various mathematical forms of the frontogenesis function have been developed, ranging from the simplest scalar form of the Petterssen frontogenesis to the complete scalar form of the Miller frontogenesis. Two vector forms have also been developed, those pertinent to two-dimensional flow and three-dimensional flow. Each form of the scalar or vector frontogenesis function has a different purpose. The third usage is physical, a description of the physical processes involved in creating and intensifying a front. Specifically, frontogenesis occurs as an initial large-scale geostrophic confluence or shear acts to increase the magnitude of a quasi-horizontal temperature gradient, eventually resulting in its nonlinear collapse due to the resulting ageostrophic secondary circulation.

Introduction

Frontogenesis has three common usages. The first usage is colloquial, where frontogenesis is used to describe the formation or intensification of a frontal zone. Because the primary measure of the intensity of a front is the horizontal temperature gradient, intensifying fronts are defined as those with increasing temperature gradients. The second usage is mathematical, where the frontogenesis function calculates the Lagrangian rate of change of the potential temperature gradient. The third usage is physical, where frontogenesis describes the process by which fronts intensify. Which definition of frontogenesis is being employed at any given time is usually understood from its context. In all three usages, the opposite of frontogenesis is called frontolysis, the weakening or dissipation of a front.

Colloquial Usage

The term frontogenesis, when used in the vernacular, describes the formation or intensification of a front or a frontal zone, usually measured by an increase in the horizontal temperature gradient across the frontal zone. For example, 'frontogenesis occurred along the leading edge of the arctic air mass over southern Kansas during the late afternoon on 25 October.'

To calculate the intensification of a frontal zone, the horizontal temperature gradient across the front must be measured in a reference frame moving with the front. For this reason, the mathematical form of the colloquial usage can be easily expressed, but it is difficult to calculate because a field representing the vector motion of the front is needed. Thus, developing an automated approach to measuring frontogenesis in this quasi-Lagrangian framework is rarely performed in practice (Schultz, 2007; Markowski and Stonitsch, 2007). Instead, point measurements of the temperature gradient across the front are usually computed and compared from one time to the next.

The second usage of the term frontogenesis avoids these practical difficulties, but at the expense of a direct measure of

frontal intensification. Instead, the mathematical formulation of frontogenesis in the next section becomes a kinematic description of flow patterns favorable for fronts to intensify.

Mathematical Usage

Petterssen (1936) first defined a mathematical quantity called frontogenesis involving the potential temperature gradient and the wind field. As a kinematic quantity, it is not a predictive tool to say whether a front will intensify, but merely a diagnostic one showing regions where the wind field is acting to intensify the gradient of potential temperature. Later researchers developed other forms of frontogenesis for various specialized purposes. The history and development of mathematical expressions for frontogenesis are discussed in this section.

Petterssen Frontogenesis

Petterssen (1936) defined frontogenesis F as the Lagrangian rate of change of the magnitude of the horizontal potential temperature (θ) gradient due to the horizontal wind ($\mathbf{V}_2 = u\mathbf{i} + v\mathbf{j}$):

$$F_{\text{Pet2}} = \frac{d}{dt} |\nabla_2 \theta|, \quad [1]$$

where

$$\begin{aligned} \frac{d}{dt} &= \frac{\partial}{\partial t} + u \frac{\partial}{\partial x} + v \frac{\partial}{\partial y}, \\ \nabla_2 &= \mathbf{i} \frac{\partial}{\partial x} + \mathbf{j} \frac{\partial}{\partial y}. \end{aligned}$$

Equation [1] can be written as

$$F_{\text{Pet2}} = -\frac{1}{|\nabla_2 \theta|} \left[\underbrace{\frac{\partial \theta}{\partial x} \left(\frac{\partial u}{\partial x} \frac{\partial \theta}{\partial x} + \frac{\partial u}{\partial y} \frac{\partial \theta}{\partial y} \right)}_{p_1} + \underbrace{\frac{\partial \theta}{\partial y} \left(\frac{\partial v}{\partial x} \frac{\partial \theta}{\partial x} + \frac{\partial v}{\partial y} \frac{\partial \theta}{\partial y} \right)}_{p_2} \right], \quad [2]$$

where terms p_1 and p_2 represent frontogenesis due to horizontal deformation/divergence. Petterssen (1936) showed that, if potential temperature is conserved, then the terms p_1 and p_2 can be rewritten as

$$F_{\text{Pet}2} = \frac{1}{2} |\nabla_2 \theta| \underbrace{(E \cos 2\beta)}_{\text{deformation}} - \underbrace{(\nabla_2 \cdot \mathbf{V}_2)}_{\text{divergence}}, \quad [3]$$

where β is the local angle between an isentrope and the axis of dilatation. The resultant deformation E is $(E_{\text{st}}^2 + E_{\text{sh}}^2)^{1/2}$, where $E_{\text{st}} = \partial u/\partial x - \partial v/\partial y$ is the stretching deformation and $E_{\text{sh}} = \partial v/\partial x + \partial u/\partial y$ is the shearing deformation. Frontogenesis in eqn [3] consists of two terms. The first term is proportional to the resultant deformation, and the second term is proportional to the horizontal divergence. Deformation acting on a horizontal temperature gradient will increase the gradient if the angle between the isentrope and the axis of dilatation is less than 45° . Deformation acting on a horizontal temperature gradient will decrease the gradient if the angle between the isentrope and the axis of dilatation is greater than 45° . Convergence acting on a horizontal temperature gradient will increase the gradient, and divergence will decrease the gradient.

Equation [3] is referred to as the *Petterssen frontogenesis function*, or simply *Petterssen frontogenesis*. Petterssen frontogenesis is the most commonly used mathematical form of frontogenesis. If the flow is geostrophic, then the divergence term in eqn [3] is zero because the geostrophic wind on an f -plane is nondivergent. Consequently, the frontogenesis due to the geostrophic wind is a result of the deformation only. If the geostrophic wind is used in eqn [3], then Keyser et al. (1988) showed that Petterssen frontogenesis is related to the forcing for quasigeostrophic vertical velocity associated with the divergence of the component of the $\mathbf{Q}$ -vector normal to the isentropes. Ascent will occur on the warm side of a region of positive Petterssen frontogenesis and descent will occur on the cool side of a region of positive Petterssen frontogenesis. Because Petterssen frontogenesis does not consider vertical velocities, it is most appropriate for use at the surface where the vertical velocity is zero. Consequently, the most common uses of Petterssen frontogenesis are to (1) examine the contributions of the total horizontal wind to changing the thermal gradient and (2) estimate the location of the secondary circulation (more on this later). Forms of the Petterssen frontogenesis function have been used to link frontogenesis to heavy rains (Sanders, 2000), elevated convective winter storms (Trapp et al., 2001), and fronts discretely propagating through the Intermountain West (Steenburgh et al., 2009). Other practical uses include an objective method for identifying the locations of fronts (Schultz, 2004; Kemppi and Sinclair, 2011) or determining the time of extratropical transition of a tropical cyclone (Harr and Elsberry, 2000).

Petterssen Frontogenesis for Three-Dimensional Flow

Because Petterssen frontogenesis does not include vertical velocity, this form of frontogenesis is generally inappropriate for application to fronts above the surface. Horizontal gradients in vertical velocity can be crucial to frontogenesis above

the surface, particularly upper-level frontogenesis occurring at the tropopause. To account for this effect, Petterssen frontogenesis can be reformulated to account for horizontal gradients of vertical motion that lead to tilting of isentropes in the vertical and that produce horizontal temperature gradients.

$$F_{\text{Pet}3} = \frac{d}{dt} |\nabla_2 \theta|, \quad [4]$$

where $\mathbf{V}_3 = (u, v, w)$ is the total velocity and

$$\frac{d}{dt} = \frac{\partial}{\partial t} + u \frac{\partial}{\partial x} + v \frac{\partial}{\partial y} + w \frac{\partial}{\partial z}.$$

Equation [4] can be written as

$$F_{\text{Pet}3} = -\frac{1}{|\nabla_2 \theta|} \left[\underbrace{\frac{\partial \theta}{\partial x} \left(\frac{\partial u}{\partial x} \frac{\partial \theta}{\partial x} + \frac{\partial u}{\partial y} \frac{\partial \theta}{\partial y} \right)}_{p_1} + \underbrace{\frac{\partial \theta}{\partial y} \left(\frac{\partial v}{\partial x} \frac{\partial \theta}{\partial x} + \frac{\partial v}{\partial y} \frac{\partial \theta}{\partial y} \right)}_{p_2} + \underbrace{\frac{\partial \theta}{\partial z} \left(\frac{\partial w}{\partial x} \frac{\partial \theta}{\partial x} + \frac{\partial w}{\partial y} \frac{\partial \theta}{\partial y} \right)}_{p_3} \right], \quad [5]$$

where terms p_1 and p_2 represent frontogenesis due to horizontal deformation/divergence and the term p_3 represents frontogenesis due to tilting of isentropes by horizontal gradients in vertical velocity. The sum of the horizontal frontogenesis terms p_1 and p_2 in eqn [5] also can be expressed as Petterssen frontogenesis.

The term p_3 represents tilting frontogenesis due to vertical-velocity gradients acting on the quasi-horizontal potential temperature gradient. The term p_3 can also be expressed as

$$p_3 = -\frac{\partial \theta}{\partial z} (\nabla_2 w \cdot \nabla_2 \theta) = -\frac{\partial \theta}{\partial z} \left(\frac{\partial w}{\partial n} \right), \quad [6]$$

such that

$$F = \frac{1}{2} |\nabla_2 \theta| (E \cos 2\beta - \nabla_2 \cdot \mathbf{V}_2) - \frac{1}{|\nabla_2 \theta|} \frac{\partial \theta}{\partial z} (\nabla_2 w \cdot \nabla_2 \theta). \quad [7]$$

This form was developed by Schultz and Steenburgh (1999) to explore the importance of the vertical motion toward changing the horizontal thermal gradient in a case of a forward tilted front during the Superstorm of March 1993. Because this form of frontogenesis already includes the vertical velocity, this form should not be used to examine the forcing for the secondary circulation.

Early forms of the three-dimensional Petterssen frontogenesis include Reed and Sanders (1953), Sanders (1955), and Ogura and Portis (1982) who first showed the kinematic processes acting on fronts. Using a model simulation of the Sanders (1955) cold front, Schultz and Roebber (2008) calculated the terms in this form of frontogenesis to reexamine the processes favoring frontogenesis. Because the isentropes were steeply sloped at the leading edge of the front, Schultz and Roebber (2008) concluded that tilting frontogenesis was negligible, contrary to the original results of Sanders (1955).

Miller (1948) Three-Dimensional Frontogenesis

A further generalization is to include the magnitude of the three-dimensional potential temperature gradient, leading to the Miller (1948) form of frontogenesis. Miller frontogenesis explains the time rate of change of the three-dimensional potential temperature gradient following air parcel motion in three dimensions. Miller (1948) extended Petterssen's definition to fronts in the free atmosphere by developing an equation for the changes of the *three-dimensional* gradient of the potential temperature due to the three-dimensional wind:

$$F_{\text{Miller}} = \frac{d}{dt} |\nabla_3 \theta|, \quad [8]$$

where

$$\begin{aligned} \frac{d}{dt} &= \frac{\partial}{\partial t} + u \frac{\partial}{\partial x} + v \frac{\partial}{\partial y} + w \frac{\partial}{\partial z}, \\ \nabla_3 &= \mathbf{i} \frac{\partial}{\partial x} + \mathbf{j} \frac{\partial}{\partial y} + \mathbf{k} \frac{\partial}{\partial z}. \end{aligned}$$

$$F_{\text{Miller}} = -\frac{1}{|\nabla_3 \theta|} \times \left[\frac{\partial \theta}{\partial x} \left(\underbrace{\frac{\partial u}{\partial x} \frac{\partial \theta}{\partial x}}_1 + \underbrace{\frac{\partial v}{\partial x} \frac{\partial \theta}{\partial y}}_2 + \underbrace{\frac{\partial w}{\partial x} \frac{\partial \theta}{\partial z}}_3 \right) + \frac{\partial \theta}{\partial y} \left(\underbrace{\frac{\partial u}{\partial y} \frac{\partial \theta}{\partial x}}_4 + \underbrace{\frac{\partial v}{\partial y} \frac{\partial \theta}{\partial y}}_5 + \underbrace{\frac{\partial w}{\partial y} \frac{\partial \theta}{\partial z}}_6 \right) + \frac{\partial \theta}{\partial z} \left(\underbrace{\frac{\partial u}{\partial z} \frac{\partial \theta}{\partial x}}_7 + \underbrace{\frac{\partial v}{\partial z} \frac{\partial \theta}{\partial y}}_8 + \underbrace{\frac{\partial w}{\partial z} \frac{\partial \theta}{\partial z}}_9 \right) \right].$$

Terms 1, 2, 4, and 5 are the horizontal-deformation terms; 7 and 8 are the vertical-deformation terms; 3 and 6 are the tilting terms (as shown in the previous section); term 9 is the vertical divergence term (recall that $\partial w / \partial z = \nabla_2 \cdot \mathbf{V}_2$).

The terms in the Miller frontogenesis comprise two groups: one for the horizontal potential temperature gradient (terms 1–6), comparable to the Petterssen frontogenesis with tilting (5), and one for the vertical potential temperature gradient (terms 7–9). Given that the vertical potential temperature gradient is the dry static stability $\partial \theta / \partial z$, the second group represents the physical processes responsible for static stability increasing or decreasing (Bosart, 1970).

$$\frac{d}{dt} \left(\frac{\partial \theta}{\partial z} \right) = -\frac{1}{|\nabla_3 \theta|} \frac{\partial \theta}{\partial z} \left(\underbrace{\frac{\partial u}{\partial z} \frac{\partial \theta}{\partial x}}_7 + \underbrace{\frac{\partial v}{\partial z} \frac{\partial \theta}{\partial y}}_8 + \underbrace{\frac{\partial w}{\partial z} \frac{\partial \theta}{\partial z}}_9 \right). \quad [9]$$

Because vertical gradients of potential temperature are an order or two larger than horizontal gradients of potential temperature, these terms dominate the frontogenesis calculation. Consequently, for some applications, separating the horizontal terms (Petterssen frontogenesis with tilting) from the vertical terms (stability frontogenesis) may be a fruitful way to analyze the

frontogenesis, neglecting the difference between $\nabla_2 \theta$ and $\nabla_3 \theta$ in the denominator. For this reason, the complete form of Miller frontogenesis is rarely used in the meteorological literature.

Vector Frontogenesis: Keyser et al. (1988)

Until this point, frontogenesis has been considered to be a scalar quantity related to the magnitude of the potential temperature gradient. Yet, the potential temperature gradient is a vector, so its direction may also be relevant. Consequently, a form of frontogenesis can be written to treat frontogenesis as a vector. Written in this manner, frontogenesis is a vector with a component normal to the isentropes and a component along the isentropes. Keyser et al. (1988) showed that the Petterssen form of frontogenesis could be generalized to a vector form where

$$\mathbf{F}_2 = \frac{d}{dt} \nabla_2 \theta, \quad [10]$$

where

$$\frac{d}{dt} = \frac{\partial}{\partial t} + u \frac{\partial}{\partial x} + v \frac{\partial}{\partial y}.$$

Resolving $\mathbf{F}_2$ into natural coordinates (s, n) such that the s axis is locally tangent to an isentrope and the n axis points toward colder air on a constant pressure surface (i.e., $\mathbf{s}$ points in the same direction as the thermal wind):

$$\frac{d}{dt} \nabla_2 \theta = \mathbf{n} \left(\mathbf{n} \cdot \frac{d}{dt} \nabla_2 \theta \right) + \mathbf{s} \left(\mathbf{s} \cdot \frac{d}{dt} \nabla_2 \theta \right), \quad [11]$$

where

$$\mathbf{n} = -\frac{\nabla_2 \theta}{|\nabla_2 \theta|}, \quad [12a]$$

$$\mathbf{s} = \mathbf{n} \times \mathbf{k}. \quad [12b]$$

Equation [10] becomes

$$\mathbf{F}_2 = F_{2n} \mathbf{n} + F_{2s} \mathbf{s}, \quad [13]$$

where F_{2n} is referred to as the scalar frontogenesis and F_{2s} is referred to as the rotational frontogenesis (Keyser et al., 1988, eqns [2.3a] and [2.3b]):

$$F_{2n} = -\frac{d}{dt} |\nabla_2 \theta|, \quad [14a]$$

$$F_{2s} = \mathbf{n} \cdot \left(\mathbf{k} \times \frac{d}{dt} \nabla_2 \theta \right). \quad [14b]$$

Therefore, positive F_{2n} implies frontolysis (note the negative sign), whereas positive F_{2s} implies cyclonic rotation of the isentropes.

Development of these expressions results in

$$F_{2n} = -\frac{1}{|\nabla_2 \theta|} \left\{ \underbrace{\frac{\partial \theta}{\partial x} \left(-\frac{\partial u}{\partial x} \frac{\partial \theta}{\partial x} - \frac{\partial v}{\partial x} \frac{\partial \theta}{\partial y} \right)}_{n_1} + \underbrace{\frac{\partial \theta}{\partial y} \left(-\frac{\partial u}{\partial y} \frac{\partial \theta}{\partial x} - \frac{\partial v}{\partial y} \frac{\partial \theta}{\partial y} \right)}_{n_2} \right\}, \quad [15a]$$

$$F_{2s} = -\frac{1}{|\nabla_2\theta|} \left\{ \underbrace{\frac{\partial\theta}{\partial y} \left(-\frac{\partial u}{\partial x} \frac{\partial\theta}{\partial x} - \frac{\partial v}{\partial x} \frac{\partial\theta}{\partial y} \right)}_{s_1} - \underbrace{\frac{\partial\theta}{\partial x} \left(-\frac{\partial u}{\partial y} \frac{\partial\theta}{\partial x} - \frac{\partial v}{\partial y} \frac{\partial\theta}{\partial y} \right)}_{s_2} \right\}. \quad [15b]$$

Following Keyser et al. (1988: pp. 764–765), expressions similar to their eqns [2.10a] and [2.10b] result in

$$F_{2n} = \frac{1}{2} |\nabla_2\theta| (\nabla_2 \cdot \mathbf{V}_2 - E \cos 2\beta) \quad [16a]$$

$$F_{2s} = \frac{1}{2} |\nabla_2\theta| (\mathbf{k} \cdot \nabla_2 \times \mathbf{V}_2 + E \sin 2\beta). \quad [16b]$$

This expression for F_{2n} (eqn [16a]) comprises two terms related to divergence and deformation (as before), whereas the expression for F_{2s} (eqn [16b]) comprises two terms related to relative vorticity and deformation.

Vector Frontogenesis, Including Vertical Motion Terms

Since contributions from vertical motion are typically small for surface fronts, but not necessarily so for upper-level fronts where tilting may be important, Keyser et al.'s (1988) methodology can be extended to include the frontogenetical effects due to vertical motion (Schultz and Doswell, 1999). Therefore, in this section, we define vector frontogenesis, F_3 , as the Lagrangian rate of change of the magnitude and direction of the horizontal potential temperature gradient due to the three-dimensional wind:

$$\mathbf{F}_3 = \frac{d}{dt} \nabla_2\theta, \quad [17]$$

where

$$\frac{d}{dt} = \frac{\partial}{\partial t} + u \frac{\partial}{\partial x} + v \frac{\partial}{\partial y} + w \frac{\partial}{\partial z}.$$

This results in the following expressions:

$$F_{3n} = -\frac{1}{|\nabla_2\theta|} \left\{ \underbrace{\frac{\partial\theta}{\partial x} \left(-\frac{\partial u}{\partial x} \frac{\partial\theta}{\partial x} - \frac{\partial v}{\partial x} \frac{\partial\theta}{\partial y} \right)}_{n_1} + \underbrace{\frac{\partial\theta}{\partial y} \left(-\frac{\partial u}{\partial y} \frac{\partial\theta}{\partial x} - \frac{\partial v}{\partial y} \frac{\partial\theta}{\partial y} \right)}_{n_2} + \underbrace{\frac{\partial\theta}{\partial z} \left(-\frac{\partial w}{\partial x} \frac{\partial\theta}{\partial x} - \frac{\partial w}{\partial y} \frac{\partial\theta}{\partial y} \right)}_{n_3} \right\}, \quad [18a]$$

$$F_{3s} = -\frac{1}{|\nabla_2\theta|} \left\{ \underbrace{\frac{\partial\theta}{\partial y} \left(-\frac{\partial u}{\partial x} \frac{\partial\theta}{\partial x} - \frac{\partial v}{\partial x} \frac{\partial\theta}{\partial y} \right)}_{s_1} - \underbrace{\frac{\partial\theta}{\partial x} \left(-\frac{\partial u}{\partial y} \frac{\partial\theta}{\partial x} - \frac{\partial v}{\partial y} \frac{\partial\theta}{\partial y} \right)}_{s_2} + \underbrace{\frac{\partial\theta}{\partial z} \left(-\frac{\partial w}{\partial y} \frac{\partial\theta}{\partial x} - \frac{\partial w}{\partial x} \frac{\partial\theta}{\partial y} \right)}_{s_3} \right\}. \quad [18b]$$

Terms n_1 , n_2 , s_1 , and s_2 represent similar terms in eqns [15a] and [15b]. The term n_3 represents tilting scalar frontogenesis due to vertical-velocity gradients across the horizontal potential temperature gradient and the term s_3 represents tilting rotational frontogenesis due to vertical-velocity gradients across the horizontal potential temperature gradient.

Following Keyser et al. (1988: pp. 764–765), expressions similar to their eqns [2.10a] and [2.10b] can be derived:

$$F_{3n} = \frac{1}{2} |\nabla_H\theta| (\nabla_H \cdot \mathbf{V}_H - E \cos 2\beta) + \frac{1}{|\nabla_H\theta|} \frac{\partial\theta}{\partial p} (\nabla_H\omega \cdot \nabla_H\theta), \quad [19a]$$

$$F_{3s} = \frac{1}{2} |\nabla_H\theta| (\mathbf{k} \cdot \nabla_H \times \mathbf{V}_H + E \sin 2\beta) + \frac{1}{|\nabla_H\theta|} \frac{\partial\theta}{\partial p} \mathbf{k} \cdot (\nabla_H\omega \times \nabla_H\theta). \quad [19b]$$

This expression for F_{3n} (eqn [19a]) comprises three terms related to divergence, deformation, and tilting, whereas (eqn [19b]) for F_{3s} comprises three terms related to relative vorticity, deformation, and tilting.

Equations [19a] and [19b] correct typographical errors in eqns [11a] and [11b] in Schultz and Doswell (1999), as documented in Schultz (2013).

Equations [19a] and [19b] were used by Schultz and Doswell (1999) to investigate the physical processes that lead to upper-level frontogenesis. They found that the onset of cold advection often observed at the leading edge of upper-level fronts was associated with the rotational frontogenesis in that region.

Other Forms

The most general form of the frontogenesis function would be an expression for vector frontogenesis using the full three-dimensional potential temperature gradient. What value such an expression would have for understanding fronts or calculating frontogenesis is not immediately apparent. Specifically, as of the writing of this article, vector frontogenesis expansions for three-dimensional Miller frontogenesis have not been constructed. Even if they were, how they would be used in operational weather forecasting or research is not clear.

Physical Usage

The physical usage of the term frontogenesis refers to the physical processes responsible for increasing the temperature gradient associated with a frontal zone. Specifically, many fronts are formed in a region where large-scale deformation occurs, differentially advecting isentropes and increasing the horizontal temperature gradient. Hoskins and Bretherton (1972) demonstrated that frontogenesis proceeds in two steps. In the first step, large-scale geostrophic deformation increases the horizontal temperature gradient. If that were the only frontogenetical process, having a front reach a first-order discontinuity in temperature gradient (discontinuous in potential temperature gradient), as is observed in many well-developed fronts, would take an infinite time. Sanders (2000) showed this by simplifying the Miller form of the frontogenesis function:

$$\frac{d}{dt} \{\ln(-\partial\theta/\partial y)\} = -\partial v/\partial y. \quad [20]$$

The logarithmic relationship between the deformation and the temperature gradient implies that it will take an infinitely long time for a first-order discontinuity to be reached. For this reason, geostrophic deformation by itself is insufficient to produce realistic fronts.

To produce realistic fronts, the second step is required. As the temperature gradient increases, a secondary ageostrophic circulation develops to maintain thermal wind balance. Ascending air on the warm side of the front cools adiabatically and descending air on the cold side of the front warms adiabatically, reducing the temperature gradient across the front. The resulting horizontal branches of the circulation, required to maintain continuity, are torqued by the Coriolis force, increasing the vertical wind shear across the front and leading to contraction of the horizontal temperature gradient in a finite time. These two components of the secondary circulation help to maintain thermal wind balance across the front, in the face of increasing thermal gradients. Mathematically, this secondary circulation can be expressed through the Sawyer–Eliassen equation (Sawyer, 1956; Eliassen, 1962, 1990).

Sawyer–Eliassen Equation

If the atmosphere is hydrostatic, then the adiabatic Sawyer–Eliassen equation can be derived for the resulting circulation in the cross-frontal (y - z) plane:

$$N^2 \frac{\partial^2 \psi}{\partial y^2} + 2S^2 \frac{\partial^2 \psi}{\partial y \partial z} + F^2 \frac{\partial^2 \psi}{\partial z^2} = 2Q_2, \quad [21]$$

where ψ is the ageostrophic streamfunction in the y - z plane

$$v_{ag} = -\frac{\partial \psi}{\partial z},$$

$$w = \frac{\partial \psi}{\partial y},$$

and terms

$$N^2 = \frac{\partial b}{\partial z},$$

$$S^2 = -\frac{\partial b}{\partial y} = f \frac{\partial u_g}{\partial z},$$

and

$$F^2 = f \left(f - \frac{\partial u_g}{\partial y} \right)$$

represent the static stability, baroclinic stability, and inertial stability. The two terms in Q_2 (often called the forcing terms) represent the geostrophic stretching and geostrophic shearing contributions to frontogenesis.

$$Q_2 = f \left(\frac{\partial v_g}{\partial y} \frac{\partial u_g}{\partial z} - \frac{\partial u_g}{\partial y} \frac{\partial v_g}{\partial z} \right). \quad [22]$$

When these terms are negative, frontogenesis occurs and is associated with a balanced direct secondary circulation. The shape of the circulation depends upon the static, baroclinic and inertial stabilities (Hakim and Keyser, 2001).

The Sawyer–Eliassen equation can be solved if the potential vorticity q is positive, producing an elliptic equation, and appropriate boundary conditions are specified.

$$q = N^2 F^2 - S^4 > 0. \quad [23]$$

The condition for positive potential vorticity is equivalent to the criterion for symmetric stability.

Summary

Three definitions of frontogenesis are in common use today. The first is a vernacular expression to describe the intensification of a front. Although a mathematical expression can be derived, in principle, its calculation is not so simple.

The second definition is a mathematical one, related specifically to the Lagrangian time rate of change of the potential temperature gradient. Although various forms of the mathematical expression for frontogenesis can be derived, which form is needed will depend upon a given circumstance. To qualitatively represent the region of frontal forcing for ascent for a surface front or to produce a surface field indicating where horizontal temperature gradients are strengthening, Petterssen frontogenesis will suffice. Above the surface, three-dimensional terms must be included. Finally, because the temperature gradient is a vector, an expression for vector frontogenesis can also be derived.

The third definition is a process-based definition: the physical processes producing an intensifying frontal zone. The process of frontogenesis proceeds in two steps: the large-scale geostrophic deformation, followed by the secondary circulation further contracting the frontal zone and acting to maintain thermal wind balance.

See also: Synoptic Meteorology: Extratropical Cyclones; Fronts.

Further Reading

- Bosart, L.F., 1970. Mid-tropospheric frontogenesis. *Quarterly Journal of the Royal Meteorological Society* 96, 442–471.
- Eliassen, A., 1962. On the vertical circulation in frontal zones. *Geophys. Publ.* 24 (4), 147–160.
- Eliassen, A., 1990. Transverse circulations in frontal zones. In: Newton, C., Holopainen, E.O. (Eds.), *Extratropical Cyclones: The Erik Palmén Memorial Volume*. American Meteorological Society pp. 155–165.

- Hakim, G.J., Keyser, D., 2001. Canonical frontal circulation patterns in terms of Green's functions for the Sawyer-Eliassen equation. *Quarterly Journal of the Royal Meteorological Society* 127, 1795–1814.
- Harr, P.A., Elsberry, R.L., 2000. Extratropical transition of tropical cyclones over the western North Pacific. Part I: Evolution of structural characteristics during the transition process. *Monthly Weather Review* 128, 2613–2633.
- Hoskins, B.J., Bretherton, F.P., 1972. Atmospheric frontogenesis models: Mathematical formulation and solution. *Journal of Atmospheric Sciences* 29, 11–37.
- Kemppi, M.L., Sinclair, V.A., 2011. Structure of a warm front: Helsinki Testbed observations and model simulation. *Monthly Weather Review* 139, 2876–2900.
- Keyser, D., Reeder, M.J., Reed, R.J., 1988. A generalization of Petterssen's frontogenesis function and its relation to the forcing of vertical motion. *Monthly Weather Review* 116, 762–780.
- Markowski, P.M., Stonitsch, J.R., 2007. Reply. *Monthly Weather Review* 135, 4240–4246.
- Miller, J.E., 1948. On the concept of frontogenesis. *Journal of Meteorology* 5, 169–171.
- Ogura, Y., Portis, D., 1982. Structure of the cold front observed in SESAME-AVE III and its comparison with the Hoskins–Bretherton frontogenesis model. *Journal of Atmospheric Sciences* 39, 2773–2792.
- Petterssen, S., 1936. Contribution to the theory of frontogenesis. *Geophys. Publ.* 11 (6), 1–27.
- Reed, R.J., Sanders, F., 1953. An investigation of the development of a mid-tropospheric frontal zone and its associated vorticity field. *Journal of Meteorology* 10, 338–349.
- Sanders, F., 1955. An investigation of the structure and dynamics of an intense surface frontal zone. *Journal of Meteorology* 12, 542–552.
- Sanders, F., 2000. Frontal focusing of a flooding rainstorm. *Monthly Weather Review* 128, 4155–4159.
- Sawyer, J.S., 1956. The vertical circulation at meteorological fronts and its relation to frontogenesis. *Proceedings of the Royal Society of London A* 234, 346–362.
- Schultz, D.M., 2004. Cold fronts with and without prefrontal wind shifts in the central United States. *Monthly Weather Review* 132, 2040–2053.
- Schultz, D.M., 2007. Comments on “Unusually long duration, multiple-Doppler radar observations of front in a convective boundary layer.” *Monthly Weather Review* 135, 4237–4239.
- Schultz, D.M., 2013. Comments on ‘The influence of rotational frontogenesis and its associated shearwise vertical motions on the development of an upper-level front’, A.A. Lang, J.E. Martin, (January A, 2010, 136, 239–252). *Quarterly Journal of the Royal Meteorological Society* 139, 269–272.
- Schultz, D.M., Doswell III, C.A., 1999. Conceptual models of upper-level frontogenesis in south-westerly and north-westerly flow. *Quarterly Journal of the Royal Meteorological Society* 125, 2535–2562.
- Schultz, D.M., Roebber, P.J., 2008. The fiftieth anniversary of Sanders (1955): a mesoscale model simulation of the cold front of 17–18 April 1953. *Synoptic-dynamic meteorology and weather analysis and forecasting: a tribute to Fred Sanders. Meteor. Monogr.*, No. 55. American Meteorological Society pp. 126–143.
- Schultz, D.M., Steenburgh, W.J., 1999. The formation of a forward-tilting cold front with multiple cloud bands during Superstorm 1993. *Monthly Weather Review* 127, 1108–1124.
- Steenburgh, W.J., Neuman, C.R., West, G.L., Bosart, L.F., 2009. Discrete frontal propagation over the Sierra–Cascade Mountains and Intermountain West. *Monthly Weather Review* 137, 2000–2020.
- Trapp, R.J., Schultz, D.M., Ryzhkov, A.V., Holle, R.L., 2001. Multiscale structure and evolution of an Oklahoma winter precipitation event. *Monthly Weather Review* 129, 486–501.

Jet Streaks

P Cunningham, Florida State University, Tallahassee, FL, USA

D Keyser, University at Albany, State University of New York, Albany, NY, USA

© 2003 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 3, pp 1043–1057, © 2003 Elsevier Ltd.

Introduction

Jet streaks, defined as localized wind speed maxima along the axis of a jet stream, are generally found in association with frontal zones in the upper troposphere, and together these features are often referred to as 'jet-front systems.' In many cases, jet streaks are also accompanied by a localized downward intrusion of air from the stratosphere into the troposphere known as a 'tropopause fold.' The interest that jet streaks have attracted in the fields of synoptic and dynamic meteorology stems from their importance in the development and behavior of synoptic-scale and mesoscale weather systems, including extratropical cyclones and mesoscale convective systems. In addition to their association with these weather systems, jet streaks are often implicated with other phenomena and processes of interest to atmospheric science, such as clear air turbulence, stratosphere–troposphere exchange, and a class of motions referred to as 'unbalanced' (e.g., inertial and buoyancy oscillations). The basic structure of jet streaks and the role that these features play in the evolution of extratropical cyclones and of mesoscale convective systems are often portrayed schematically in terms of conceptual models that relate jet streaks in various flow configurations to characteristic patterns of horizontal divergence and vertical motion. Dynamical interpretation of jet streaks has proceeded along two lines of thinking: in the first, referred to as 'balanced,' it is assumed that a dynamical relationship constrains the wind and geopotential fields; in the second, referred to as 'unbalanced,' no such constraint exists. In the balanced case, jet streaks may be interpreted either as an integral part of upper-tropospheric baroclinic waves, in which case their evolution is controlled by energy dispersion associated with the wave, or in terms of the interaction between coherent vortices and the jet stream, in which case their evolution occurs in response to the motion and evolution of the constituent vortices. In the unbalanced case, the dynamics are strongly coupled to a category of buoyancy oscillations referred to as inertia-gravity waves, which may strongly modulate weather systems on short time scales.

General Characteristics of Jet Streaks

Jet streaks are observed along the polar-front and subtropical jet streams in both the Northern and Southern Hemispheres and can be either mobile or stationary with respect to the stream in which they are embedded. Like atmospheric fronts, to which they are closely related, jet streaks are often significantly anisotropic, exhibiting disparate length scales in the along-jet and cross-jet directions, with the former being up to an order of magnitude larger than the latter. In the case of the polar-front jet stream, jet streaks typically have length scales of 2000 km or less in the along-jet direction and progress

eastward through the meanders of the stream. Some polar-front jet streaks are observed to translate continuously along the jet stream, whereas others travel discretely and appear to jump between the north-westerly and south-westerly flow inflections of the wave pattern in the stream. In the case of the subtropical jet stream, jet streaks possess along-jet length scales of 4000–8000 km and tend to be approximately stationary and located in the ridges of the wavelike meanders in the stream.

Jet streaks often play a critical role in the development of extratropical cyclones and in the organization of mesoscale convective systems, and as such have become an important part of mainstream synoptic meteorology and weather forecasting. The association of jet streaks with these phenomena is related to the observation that the divergence of the horizontal wind (hereafter, this quantity will be referred to as the horizontal divergence, while the terms divergence and convergence will be used to distinguish between positive and negative values of this quantity, respectively) is frequently large in the vicinity of jet streaks. The relationship of horizontal divergence in the upper troposphere to changes in sea level pressure and to patterns of vertical motion in the middle troposphere is well known. Indeed, a fundamental principle of early synoptic meteorology states that the development of surface cyclones requires upper-level divergence slightly in excess of lower-level convergence, which results in a net reduction in mass in the column and a decrease in the sea level pressure. The vertical circulations associated with patterns of horizontal divergence in the upper troposphere may be inferred using the equation of mass continuity formulated in pressure coordinates, given by $\nabla \cdot \mathbf{V} = -\partial\omega/\partial p$, where ∇ is the horizontal gradient operator evaluated on an isobaric surface, $\mathbf{V}$ is the horizontal wind velocity, and ω is the vertical velocity in pressure coordinates. The pressure-coordinate vertical velocity is related approximately to the height-coordinate vertical velocity, w , via $\omega \approx -\rho gw$, where ρ is air density, g is the acceleration due to gravity, and positive (negative) ω corresponds to downward (upward) motion. The pressure-coordinate form of the vertical velocity is adopted for the remainder of this article. Since the vertical velocity is generally close to zero near the tropopause, divergence (convergence) in the upper troposphere will be associated with ascending (descending) motion in the middle troposphere. This relationship between upper-tropospheric horizontal divergence and middle-tropospheric vertical motion will be alluded to frequently in the remainder of this article.

Although the observed structure and behavior of jet streaks can vary considerably from case to case, several properties are common to these features and are illustrated through the analysis of a jet streak observed over the north central United States on 3 November 1995. In [Figure 1](#), which depicts the wind speed and geopotential height fields at 300 hPa, the jet streak is located in the south-westerly flow inflection of a synoptic-scale trough–ridge pattern. In this jet streak, the

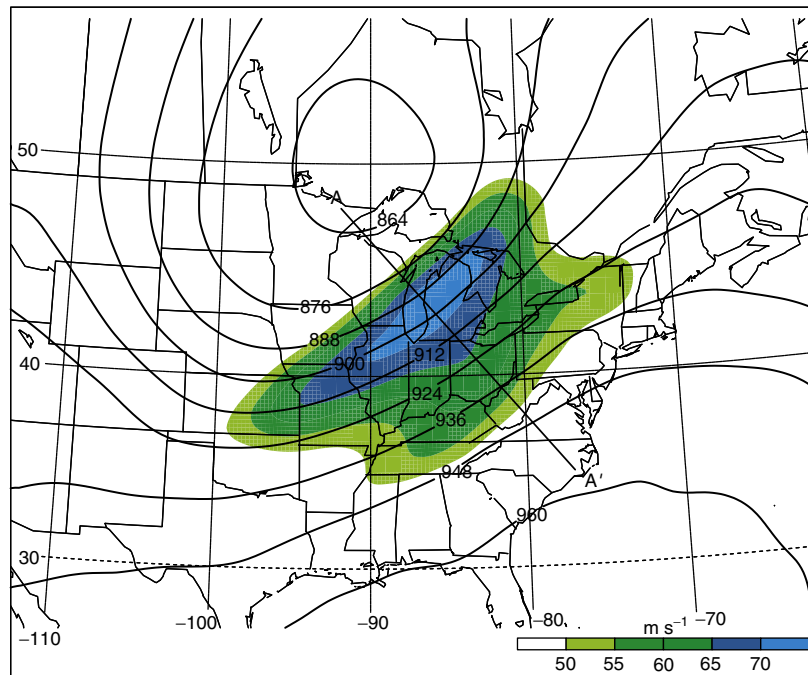


Figure 1 European Centre of Medium-Range Weather Forecasts (ECMWF) analysis on the 300 hPa pressure surface, valid at 1200 UTC 3 November 1995, depicting wind speed (values greater than 50 m s^{-1} shaded as indicated) and geopotential height (contour interval 120 m; solid lines).

maximum wind speed is approximately 75 m s^{-1} , and the aspect ratio expressed in terms of along-jet to cross-jet length scales is approximately 4:1 as determined by the configuration of the 60 m s^{-1} isotach.

A vertical cross-section through the jet streak under consideration in the cross-jet direction is shown in **Figure 2**, which illustrates the components of a jet-front system. The tropopause, depicted as the thick solid line, is defined 'dynamically' as the 2 PVU potential vorticity surface ($1 \text{ PVU} = 10^{-6} \text{ m}^2 \text{ s}^{-1} \text{ K kg}^{-1}$). Here potential vorticity, P , is given by $P = -g(\partial\theta/\partial p)(f + \zeta_\theta)$, where f is the Coriolis parameter, θ is the potential temperature, and ζ_θ is the vertical component of the relative vorticity evaluated on an isentropic surface. In the vicinity of the jet streak, the tropopause transects the core of maximum wind speed (i.e., the 'jet core') and slopes steeply, being considerably lower on the poleward side than on the equatorward side. The downward and equatorward protrusion of the tropopause into the middle troposphere beneath the jet streak corresponds to an incipient tropopause fold. In a well-developed fold, the tropopause generally exhibits an S-shaped pattern, such that there is a significant region in which the tropopause height is multivalued. Above the jet core, the contours of constant potential temperature (i.e., isentropes) slope upward. Below the jet core, the isentropes slope downward, such that the potential temperature gradient evaluated on an isobaric surface in this region is locally large. This area of locally enhanced horizontal potential temperature gradient defines the upper-tropospheric frontal zone, which in this case extends well into the middle troposphere and joins with the upward extension of a surface frontal zone. The fan-shaped pattern of the isentropes observed in conjunction with this jet streak is a consequence of the thermal

wind relation, which links the vertical shear of the geostrophic wind to the horizontal potential temperature gradient.

Jet streaks in the upper troposphere and their associated frontal zones can also be preferred locations for phenomena not normally thought of as related to weather. In the regions of large vertical shear above and below the jet streak, the Richardson number, Ri , defined as $Ri = (g/\theta)(\partial\theta/\partial z)/|\partial\mathbf{V}/\partial z|^2$, can become small. When this is the case, the flow can break down into turbulent eddies. Since this breakdown often occurs in regions of background subsidence and relatively cloud-free air, this phenomenon is generally referred to as 'clear air turbulence,' which can be hazardous to aircraft. Moreover, the region of turbulent mixing tends to be localized in the vicinity of the tropopause fold, which contains stratospheric air protruding downwards into the troposphere. In the region of mixing, the tropopause, which ordinarily corresponds to a material surface separating stratospheric from tropospheric air, is highly porous, allowing for a vigorous two-way exchange of air and trace constituents between the stratosphere and troposphere.

The applicability of the thermal wind relation referred to above to jet-front systems implies the validity of geostrophic balance, which will now be employed in considering the dynamics of jet streaks. Geostrophic balance is valid to a first approximation in many jet streaks, such that the so-called ageostrophic wind (defined as $\mathbf{V}_{ag} = \mathbf{V} - \mathbf{V}_g$, the vector difference between the actual horizontal wind and the geostrophic wind) is relatively small in comparison to the geostrophic wind. The magnitude of the departure from geostrophic balance is quantified by the Rossby number, Ro , which is defined as the ratio of the characteristic scales for the velocity acceleration and the Coriolis force, and is given by $Ro = V/fL$, where L and V

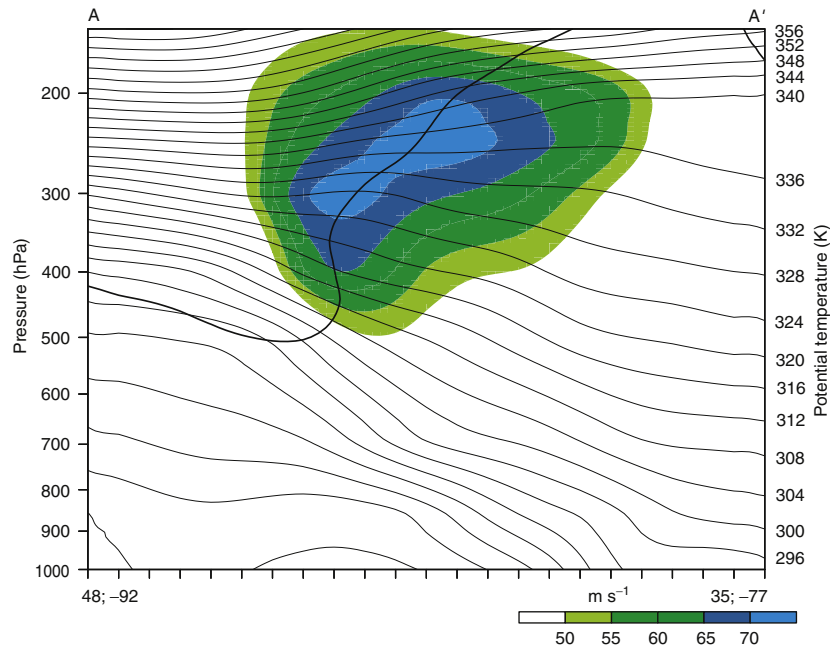


Figure 2 Cross-section along A–A' indicated in [Figure 1](#), valid at 1200 UTC 3 November 1995, depicting wind speed normal to the plane of the section (values greater than 50 m s^{-1} shaded as indicated), potential temperature (contour interval 4 K; thin solid lines) and the dynamic tropopause (thick solid line).

may be taken to be the along-jet length scale and the maximum wind speed of the jet streak, respectively. If Ro is small with respect to unity, the jet streak will be in approximate geostrophic balance. In such cases the so-called quasi-geostrophic system, an approximation to the full equations of motion valid for small Ro , provides a useful foundation for understanding the structure and behavior of jet streaks. In fact, the quasi-geostrophic system often is applicable under circumstances beyond the limits of its strict validity, so that it may be employed qualitatively to interpret the structure and behavior of jet streaks in which Ro is not particularly small. Many of the existing dynamical concepts regarding jet streaks are based on deductions derived from the quasi-geostrophic system. Nevertheless, sometimes Ro can be so large in the vicinity of jet streaks that the quasi-geostrophic system is not even qualitatively useful. In such jet streaks, unbalanced motions may be important to the evolution of the streak and the full equations of motion, also known as the primitive equations, are required to elucidate the properties of the flow and the behavior of the streak.

The ageostrophic wind at 300 hPa for the jet streak under consideration, shown in [Figure 3](#), displays a four-gyre circulation pattern that is cyclonic upstream and downstream of the jet core and anticyclonic on the flanks of the streak. Hence, in the region upstream of the jet core, known as the entrance region, the ageostrophic wind is directed toward the poleward (i.e., cyclonic-shear) side of the streak, corresponding to lower geopotential height; in the region downstream of the jet core, known as the exit region, the ageostrophic wind is directed toward the equatorward (i.e., anticyclonic-shear) side of the streak, corresponding to higher geopotential height. There is also a significant along-jet component of the ageostrophic wind directed upstream, indicating that the actual wind in this

jet streak is sub-geostrophic (i.e., the actual wind speed in the core of the jet is smaller in magnitude than the geostrophic wind speed), a property that may be related to the cyclonically curved orientation of the jet streak. Nevertheless, this particular property is not generic; some jet streaks are approximately geostrophic in the along-jet direction, whereas others, particularly those that are anticyclonically curved, may be super-geostrophic. Explanations for employing the terminology of 'entrance' and 'exit' with regard to jet streaks and for the presence of sub-geostrophic and super-geostrophic along-jet flow will be provided subsequently.

Conceptual Models of Jet Streaks

Since the horizontal divergence is often a small residual between two larger kinematic quantities that are nearly equal in magnitude but opposite in sign, uncertainties arise in calculating the upper-tropospheric horizontal divergence directly from conventional upper-air data. These uncertainties may be particularly significant if the jet streak is located in a data-sparse region, or if the streak is so narrow that the observational network is unable to resolve accurately the structure of the streak. Consequently, the horizontal divergence has been inferred using air parcel arguments applied to the vector momentum equation and to the vorticity equation, the results of which have been portrayed schematically in terms of conceptual models relating horizontal divergence patterns to various aspects of the structure of jet streaks. Several well-known conceptual models of jet streaks are reviewed below: (i) models of straight and curved jet streaks; (ii) models of jet streaks coupled with other upper-tropospheric jet streaks or with lower-tropospheric frontal zones; and (iii) a model

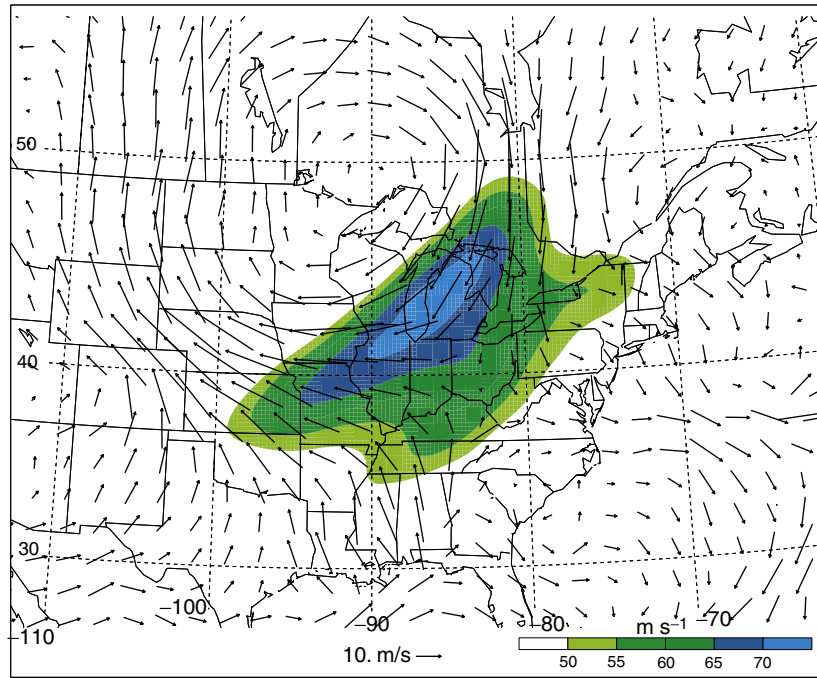


Figure 3 ECMWF analysis on the 300 hPa pressure surface, valid at 1200 UTC 3 November 1995, depicting wind speed (values greater than 50 m s^{-1} shaded as indicated) and ageostrophic wind (arrows; vector scale indicated at bottom of diagram).

describing the life cycle of a jet streak embedded within an evolving synoptic-scale baroclinic wave.

Straight Jet Streaks

Perhaps the best known conceptual model of jet streaks, illustrated in [Figure 4](#), depicts a straight isotach maximum in which the isentropes are assumed to be oriented parallel to the jet axis. This conceptual model is often referred to as the ‘four-quadrant model’ because the upstream (entrance) and downstream (exit) regions of the streak are divided into left and right regions by the jet axis (with left and right defined facing downstream). To infer the nature of the ageostrophic flow and the horizontal divergence in this and subsequent conceptual models, it is convenient to adopt a natural coordinate representation of the frictionless vector momentum equation (eqns [1] and [2]).

$$v_{\text{agn}} = \frac{1}{f} \frac{DV}{Dt} \quad [1]$$

$$v_{\text{ags}} = -\frac{K_t V^2}{f} \quad [2]$$

Here v_{agn} and v_{ags} are the cross-stream and along-stream components of the ageostrophic wind in a ‘right-handed’ natural coordinate system such that $\mathbf{n} = \mathbf{k} \times \mathbf{s}$, $\mathbf{s} = \mathbf{V}/V$, V is the horizontal wind speed, and K_t is the parcel trajectory curvature. At the level of the jet core, the vertical velocity is assumed to be small and may be neglected; hence the Lagrangian rate of change $D/Dt (= \partial/\partial t + \mathbf{V} \cdot \nabla)$ may be defined following the horizontal flow.

In general, jet streaks travel at a speed that is slower than the maximum wind speed in the jet core. Consequently, air parcels travel through the streak, entering upstream and exiting

downstream – hence the terms entrance and exit regions introduced previously. An air parcel in the entrance region will experience increasing wind speed ($DV/Dt > 0$), and an air parcel in the exit region will experience decreasing wind speed ($DV/Dt < 0$). Since the curvature is negligible for the straight jet streak depicted in [Figure 4\(a\)](#), the along-stream component of the ageostrophic wind is small (see eqn [2]); the cross-stream component of the ageostrophic wind is directed from higher to lower values of geopotential height in the entrance region and from lower to higher values in the exit region (see eqn [1]). Away from the jet axis, where speed accelerations are weaker, the ageostrophic flow decreases in magnitude and thus there is convergence in the left entrance and right exit regions, and there is divergence in the right entrance and left exit regions. The vertical circulations transverse to the jet streak, depicted in [Figure 4\(b\)](#), are thermally direct in the entrance region and thermally indirect in the exit region. The sense of these circulations is consistent with the conversion from potential to kinetic energy required for an air parcel to increase its speed in the entrance region of the streak, and with the conversion from kinetic to potential energy required for an air parcel to decrease its speed in the exit region.

It is also possible to deduce the patterns of horizontal divergence illustrated in [Figure 4\(a\)](#) using the equation governing the evolution of the vertical component of the relative vorticity evaluated on an isobaric surface, ζ , again at the level of the jet core, given by eqn [3].

$$\nabla \cdot \mathbf{V} = -\frac{1}{(f + \zeta)} \frac{D\zeta}{Dt} \quad [3]$$

In addition to the neglect of the tilting, vertical advection, and friction terms, the Lagrangian rate of change of planetary vorticity is here assumed to be small in comparison to the rate

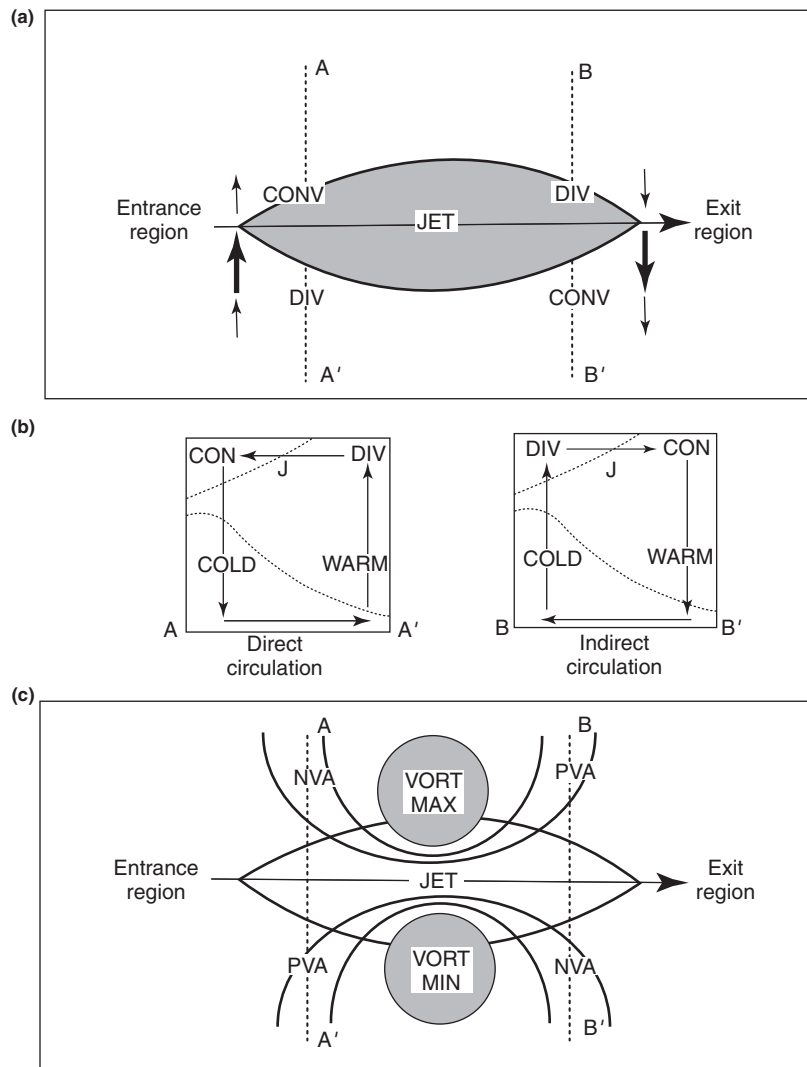


Figure 4 Schematic illustration, applicable to the Northern Hemisphere, of ageostrophic circulations and vorticity patterns for a straight jet streak. (a) Transverse ageostrophic wind components and associated patterns of convergence (CONV) and divergence (DIV) in the entrance and exit regions at the level of maximum wind. (b) Transverse ageostrophic circulations in entrance (cross-section A–A') and exit (cross-section B–B') regions of jet streak depicted in (a), along with schematic isentropes (dotted lines) and location of jet core (J). (c) Relative vorticity and associated advection patterns, with NVA and PVA indicating negative (anticyclonic) and positive (cyclonic) vorticity advection, respectively. From Uccellini LW and Kocin PJ (1987). The interaction of jet streak circulations during heavy snow events along the East Coast of the United States. *Weather and Forecasting* 2: 289–308. American Meteorological Society, Boston.

of change of relative vorticity. As illustrated in Figure 4(c), a straight jet streak exhibits cyclonic relative vorticity on its poleward side and anticyclonic relative vorticity on its equatorward side. Hence, an air parcel traveling downstream through the left entrance or right exit regions will experience an increase in cyclonic relative vorticity, and an air parcel traveling downstream through the right entrance or left exit regions will experience an increase in anticyclonic relative vorticity. Equation [3] indicates that convergence is expected in the left entrance and right exit regions, whereas divergence is expected in the right entrance and left exit regions.

The conceptual model shown in Figure 4 assumes that isentropes are parallel to the jet axis. For many observed jet streaks, however, this assumption is overly restrictive and the

isentropes are generally aligned at an angle to the jet axis. The effect of the resultant along-jet temperature advection on the patterns of horizontal divergence and vertical motion accompanying a straight jet streak is illustrated in Figure 5 for two different alignments of the isentropes relative to the jet axis. When there is cold advection along the jet (Figure 5(a)), the transverse vertical circulations are shifted toward the equatorward (anticyclonic-shear) side of the jet streak in the entrance region and to the poleward (cyclonic-shear) side in the exit region. Conversely, along-jet warm advection (Figure 5(b)) results in a shift of the transverse circulations to the poleward (cyclonic-shear) side in the entrance region and to the equatorward (anticyclonic-shear) side in the exit region. As depicted in these illustrations, these shifts in the vertical circulations can

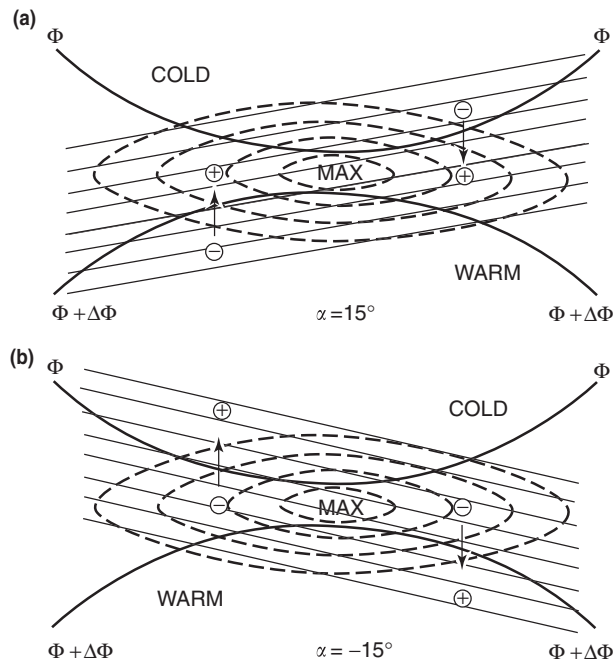


Figure 5 Schematic illustration of geopotential height (heavy solid lines), wind speed (heavy dashed lines), and potential temperature (thin solid lines) associated with a straight jet streak for isentropes aligned at a small angle to jet axis: (a) along-jet cold advection; (b) along-jet warm advection. Arrows indicate the sense of the cross-front ageostrophic wind component, and the plus and minus signs indicate maxima and minima in the middle-tropospheric pressure-coordinate vertical velocity. Recall that positive (negative) values of this quantity correspond to downward (upward) motion. Adapted from Shapiro (1982).

lead to the maximum vertical motions being located beneath the jet axis, rather than to either side as in Figure 4. In the cold advection case, such a configuration may be important to the development of tropopause folds, which requires the presence of significant subsidence localized beneath the jet axis.

Curved Jet Streaks

Although evidence exists supporting the idealized structure of isolated straight jet streaks such as those depicted in Figure 4 and Figure 5, jet streaks exhibiting a four-quadrant pattern of horizontal divergence and vertical motion are the exception rather than the rule. The rarity of the four-quadrant pattern is a consequence of the fact that few jet streaks are immune to the effects of curvature: many jet streaks have curved axes, and even jet streaks that have straight axes tend to be embedded in a large-scale jet stream that may itself be curved. The influence of curvature on the horizontal divergence patterns associated with jet streaks may be considered by appealing to the schematic depiction of a steady curved jet stream shown in Figure 6, which exhibits uniform wind speed in the along-stream direction such that $DV/Dt \equiv 0$. Assuming that parcel trajectory curvature may be approximated by streamline curvature, eqn [2] implies that in the trough, where the curvature is cyclonic ($K_t > 0$ in the Northern Hemisphere), the ageostrophic wind is directed upstream ($v_{ags} < 0$) and the actual wind is

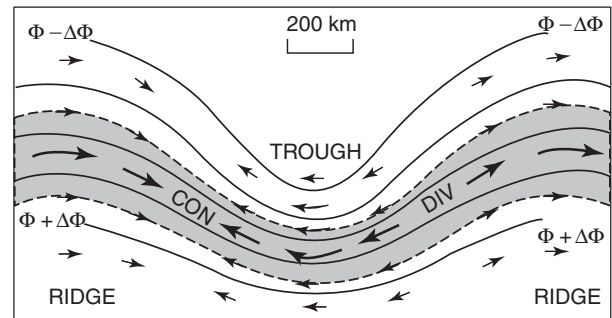


Figure 6 Schematic illustration of geopotential height (solid lines), wind speed (dashed lines; region of maximum wind speed shaded), ageostrophic wind (arrows) and associated patterns of convergence (CON) and divergence (DIV) in the vicinity of a curved jet stream exhibiting uniform wind speed in the along-stream direction. Adapted from Shapiro MA and Kennedy PJ (1981) Research aircraft measurements of jet stream geostrophic and ageostrophic winds. *Journal of the Atmospheric Sciences* 38: 2642–2652. American Meteorological Society, Boston.

sub-geostrophic, whereas in the ridge, where the curvature is anticyclonic ($K_t < 0$ in the Northern Hemisphere), the ageostrophic wind is directed downstream ($v_{ags} > 0$) and the actual wind is super-geostrophic. In the inflections between the trough and ridge, the along-stream ageostrophic wind is zero; hence, divergence is expected between the trough and ridge, whereas convergence is expected between the ridge and trough. Consequently, the pattern of horizontal divergence for a jet streak embedded in such a curved jet stream may be modified significantly from that shown in Figure 4(a), depending on where the streak is located with respect to the trough or the ridge. For example, for a jet streak located near the base of the trough, the convergence and divergence in the left entrance and left exit regions of the streak will be enhanced by the convergence upstream and the divergence downstream associated with the jet stream. As a result, the horizontal divergence may exhibit a two-cell pattern, rather than the four-cell pattern associated with a straight jet streak. The two-cell pattern appears to be more common, given the frequent association of jet streaks with so-called ‘short-wave troughs’ in the upper troposphere, which are characterized by cyclonic curvature.

Coupled Jet Streaks

Sometimes two separate jet streaks in the upper troposphere may come into such close proximity that the vertical circulations associated with these features become coupled. Such a configuration, depicted schematically in Figure 7, is common to cyclogenetic events that produce heavy snow along the East Coast of the United States. In this configuration, the left exit region of the equatorward jet streak and the right entrance region of the poleward jet streak become colocated, such that the divergence in the upper troposphere and the ascent below are enhanced. This coupling of the vertical circulations can lead to the organization of heavy precipitation and to the rapid development of the surface cyclone.

The vertical circulations associated with upper-tropospheric jet streaks can also interact with the corresponding circulations associated with lower-tropospheric jets and frontal zones.

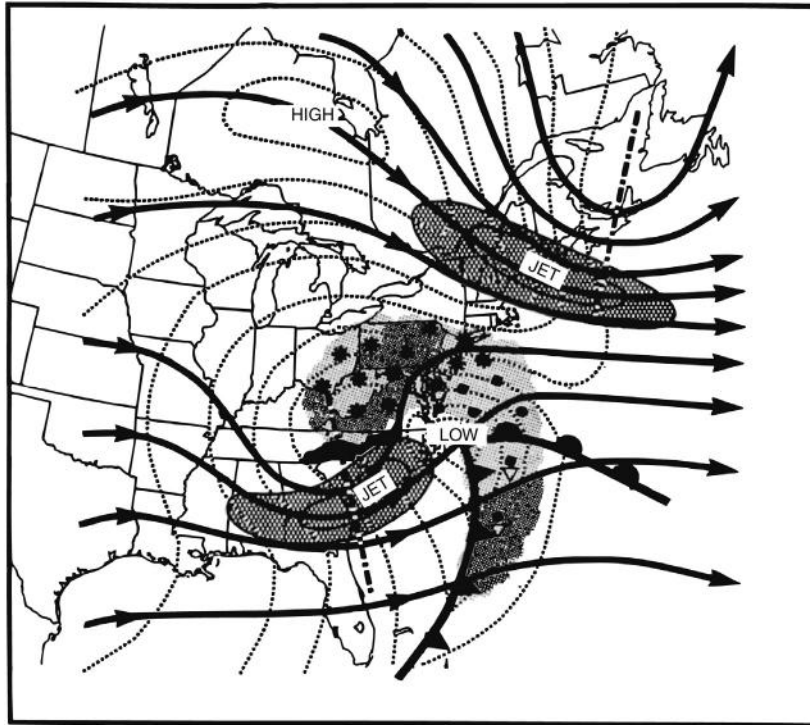


Figure 7 Schematic illustration of surface cold and warm fronts, high and low pressure centers, sea level isobars (dotted lines), precipitation (shaded, with asterisks representing snow and dots representing rain), upper-level flow (arrows), upper-level trough axes (dot-dashed lines), and jet streaks (cross-hatched) associated with a 'typical' heavy snow event along the East Coast of the United States. From Uccellini LW and Kocin PJ (1987) *The interaction of jet streak circulations during heavy snow events along the East Coast of the United States*. *Weather and Forecasting* 2: 289–308. American Meteorological Society, Boston.

A hypothetical example depicting unfavorable and favorable configurations of upper-level and lower-level systems is provided in (Figure 8). In the unfavorable configuration (Figure 8(a,b)), the right exit region of an upper-tropospheric jet streak is located above a surface frontal zone. The subsidence associated with the jet streak in this region overlies the ascent associated with the surface frontal zone, acting to stabilize the air above the front and to suppress convection. In the favorable configuration (Figure 8(c,d)), the left exit region of the jet streak is located above the surface frontal zone, and the ascent associated with the jet streak can couple with the ascent accompanying the surface front, leading to the release of convective instability and to the outbreak of severe convective storms.

Jet Streak Life Cycles

The foregoing conceptual models treat jet streaks as static entities in the sense that their structure and amplitude remain steady, although they can translate or propagate. In reality, jet streaks undergo life cycles while interacting with the background-flow environment. A hypothetical jet streak life cycle is illustrated in Figure 9, in which the jet streak is traveling through and interacting with an evolving synoptic-scale baroclinic wave. In this scenario, the jet streak forms in the confluence zone situated within a trough-over-ridge pattern (Figure 9(a)). As the jet streak migrates into the north-westerly flow inflection between the

ridge and the trough (Figure 9(b)), the baroclinic wave amplifies. The jet streak subsequently advances toward the base of the trough (Figure 9(c)), at which time the wave is fully developed and after which time the wave decays as the streak travels through the south-westerly flow downstream of the trough (Figure 9(d)). Although this scenario is seldom observed in its entirety, it is common for an evolving jet streak to progress through at least one or several of the stages depicted in (Figure 9) during some portion of its life cycle.

Dynamics of Jet Streaks

The conceptual models of jet streaks presented above are based on hypothetical configurations of wind, geopotential height, and potential temperature fields, and on deductions based on air parcel arguments. However, it is not immediately obvious that these configurations correspond to realizable solutions, obtained using either analytical or numerical methods, to consistent sets of dynamical equations governing the evolution of the flow. To describe and understand the processes by which jet streaks evolve, it is desirable to refer to such solutions. In the balanced case, the dynamics of jet streaks may be described in terms of linear waves or nonlinear coherent vortices using the quasi-geostrophic system, whereas in the unbalanced case, the dynamics of jet streaks are strongly coupled to inertia-gravity waves.

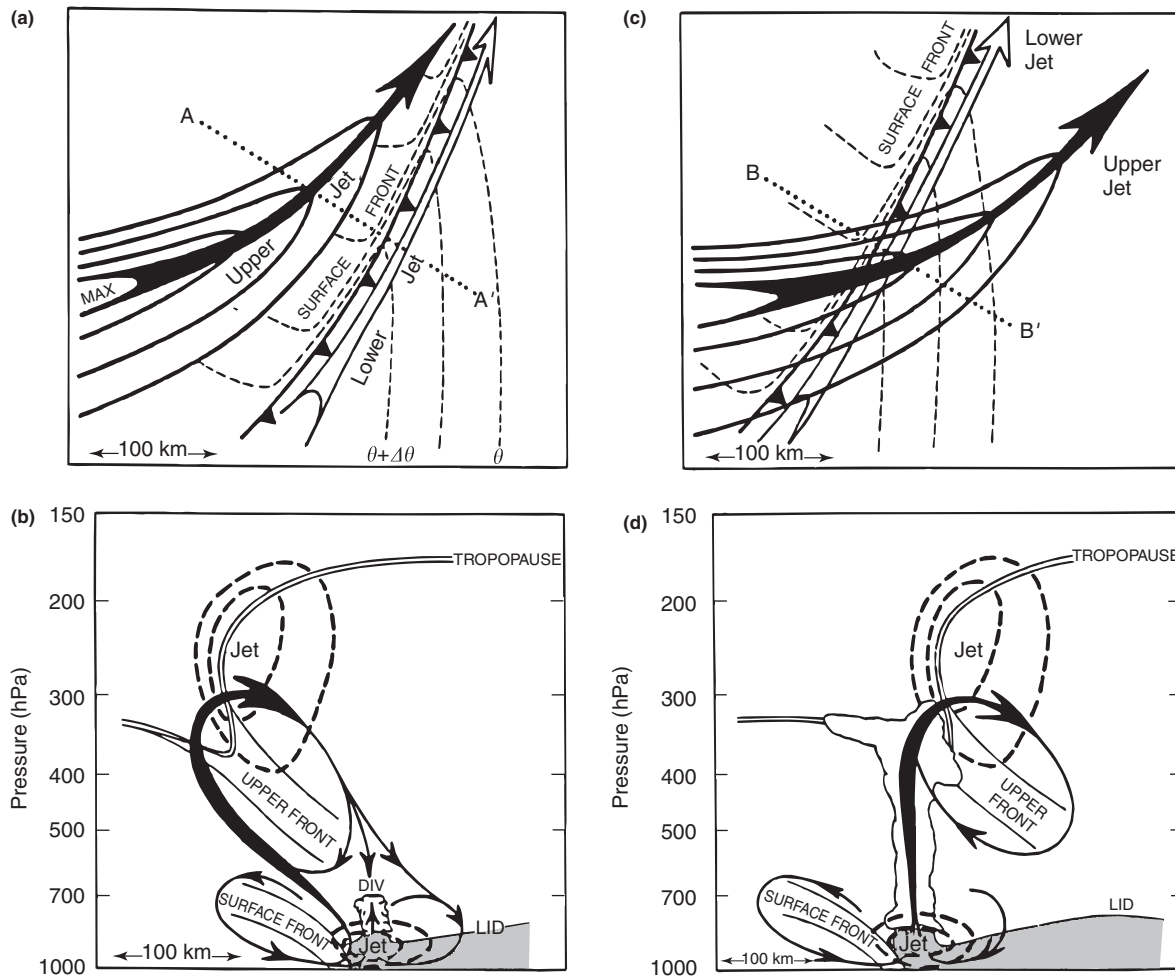


Figure 8 Schematic illustration of vertically uncoupled (a,b) and vertically coupled (c,d) upper- and lower-tropospheric jet-front systems. (a) Upper-tropospheric jet exit region (isentachs, heavy solid lines; jet axis, solid arrow) aligned along and displaced toward cold side of surface frontal zone (isentachs, dashed lines; cold front marked conventionally) and lower-tropospheric jet (jet axis, open arrow). (b) Cross-section along A-A' indicated in (a), depicting upper- and lower-tropospheric jets (isentachs, thick dashed lines), upper- and lower-tropospheric frontal zones (bounded by thin solid lines), tropopause (double solid lines), moist boundary layer (shading) capped by lid, and streamlines of transverse ageostrophic circulation (solid arrows, strength of circulation proportional to width). (c) As in (a) except for upper-tropospheric jet exit region aligned across surface frontal zone. (d) As in (b) except for cross-section along B-B' indicated in (c). Adapted from Shapiro (1982).

Jet Streaks and Baroclinic Waves

It was noted previously that jet streaks often play a significant role in extratropical cyclogenesis. In fact, there is evidence suggesting that jet streaks are an integral part of the upper-tropospheric baroclinic wave associated with cyclogenesis and of a dynamical process known as 'downstream development', which accounts for the growth and decay of many such waves. In the downstream development process, kinetic energy propagates from an upstream trough that is decaying toward a new trough that is growing downstream. This energy is maximized in the inflections between the troughs and ridges, and these kinetic energy maxima correspond to jet streaks. An idealized picture of the structure of a baroclinic wave undergoing downstream development is shown in [Figure 10](#). The evolution of the jet streaks in this process is consistent with the observation that many streaks appear to

'jump' between the north-westerly and south-westerly flow inflections of the wave, rather than translate continuously through the wave pattern, as is the case for the conceptual life cycle shown in [Figure 9](#).

Jet Streaks and Coherent Vortices

Observations suggest that, in some cases, jet streaks are associated with monopolar and dipolar vortices of mesoscale dimensions (length scales of approximately 500 km) that are embedded within a larger-scale jet stream. These features differ from the synoptic-scale baroclinic waves described above in that they do not disperse their energy; instead they travel coherently, maintaining their identities for periods of up to several weeks. A straight jet streak represented in terms of a symmetric vortex dipole solution to the quasi-geostrophic

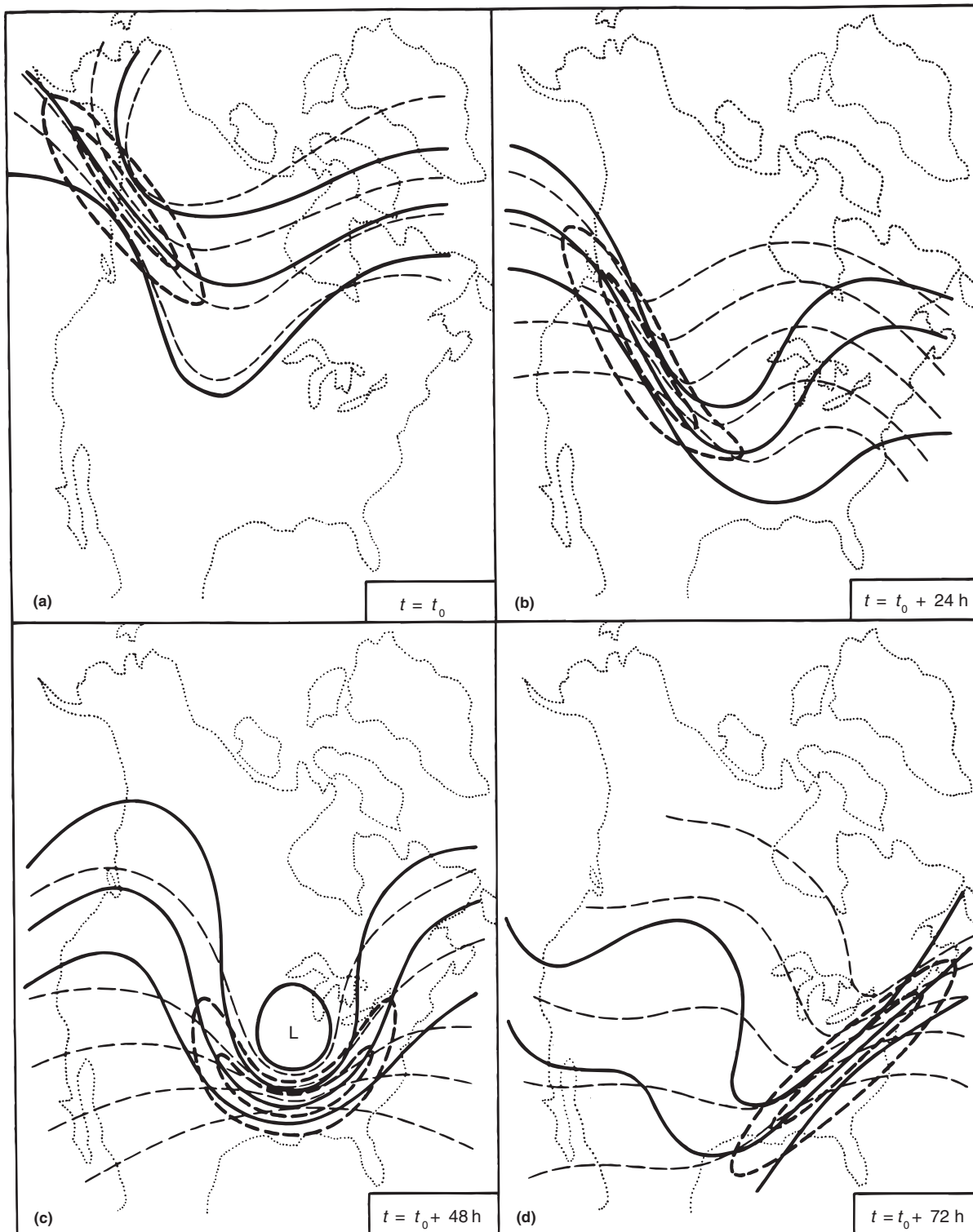


Figure 9 Schematic illustration of geopotential height (heavy solid lines), wind speed (heavy dashed lines), and potential temperature (thin dashed lines) associated with the migration of an upper-tropospheric jet-front system through a synoptic-scale baroclinic wave over a 72 h period. (a) Jet-front forming in the confluence zone between middle- and polar-latitude currents. (b) Jet-front situated in the north-westerly flow inflection of amplifying wave. (c) Jet-front at the base of the trough of fully developed wave. (d) Jet-front situated in the south-westerly flow inflection of decaying wave. From Shapiro (1982).

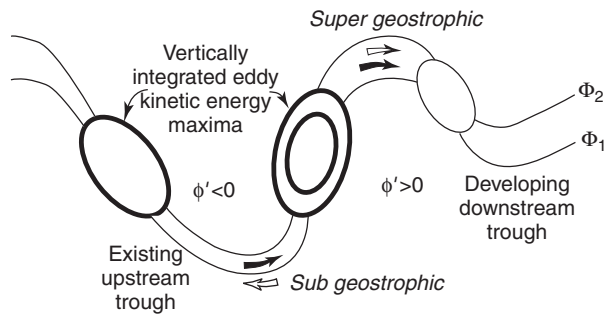


Figure 10 Schematic illustration of the relationship between the components of a synoptic-scale baroclinic wave. The upper-level geopotential field is indicated by two contours labeled ϕ_1 and ϕ_2 . The geopotential anomaly, ϕ' , relative to the time mean is positive/negative in the ridge/trough. Wave-relative air flow is indicated by heavy solid arrows and the ageostrophic wind is indicated by open arrows. Centers of maximum vertically integrated eddy kinetic energy are shown as ellipses. From Orlanski I and Sheldon JP (1995) Stages in the energetics of baroclinic systems. *Tellus 47A*: 605–628. Munksgaard International Publishers, Copenhagen.

system is depicted schematically in **Figure 11**; this schematic may be compared with its observationally based counterpart in **Figure 4**. In the representation depicted in **Figure 11**, the flow associated with the cyclonic vortex advects the anticyclonic vortex and vice versa, while the combined flow of both vortices results in a localized wind speed maximum (i.e., a jet streak) in the region between the vortices (**Figure 11(a)**). Hence the jet streak moves downstream in conjunction with the dipole, but at a slower speed than the maximum wind speed in the jet core, consistent with observations. The rotational component of the ageostrophic flow (**Figure 11(b)**), which dominates the total ageostrophic flow in the vortex dipole solution, exhibits a four-gyre circulation pattern consistent with the total ageostrophic flow pattern depicted in **Figure 3**. The divergent component of the ageostrophic flow (**Figure 11(c)**) is associated with a four-quadrant pattern of horizontal divergence consistent with that depicted in **Figure 4(a)**.

Unbalanced Flow in Jet Streaks

The foregoing discussion of the dynamics of jet streaks is predicated on the assumption that the flow satisfies some balance relationship connecting the wind and geopotential fields. In some cases, however, jet streaks do not satisfy any particular balance relationship; jet streaks are often characterized by Rossby numbers that are large compared with unity, and by large values of horizontal divergence and Lagrangian rates of change of this quantity. A well-documented example of an unbalanced flow configuration is that of an upper-tropospheric jet streak located in the south-westerly flow inflection between a trough and ridge in the circumstance where the distance between the trough and ridge is contracting. It has been hypothesized that this upper-tropospheric configuration is conducive to the generation of large-amplitude inertia-gravity waves in the lower troposphere, which may be accompanied by hazardous weather. Moreover, inertia-gravity waves associated with this configuration may be important in

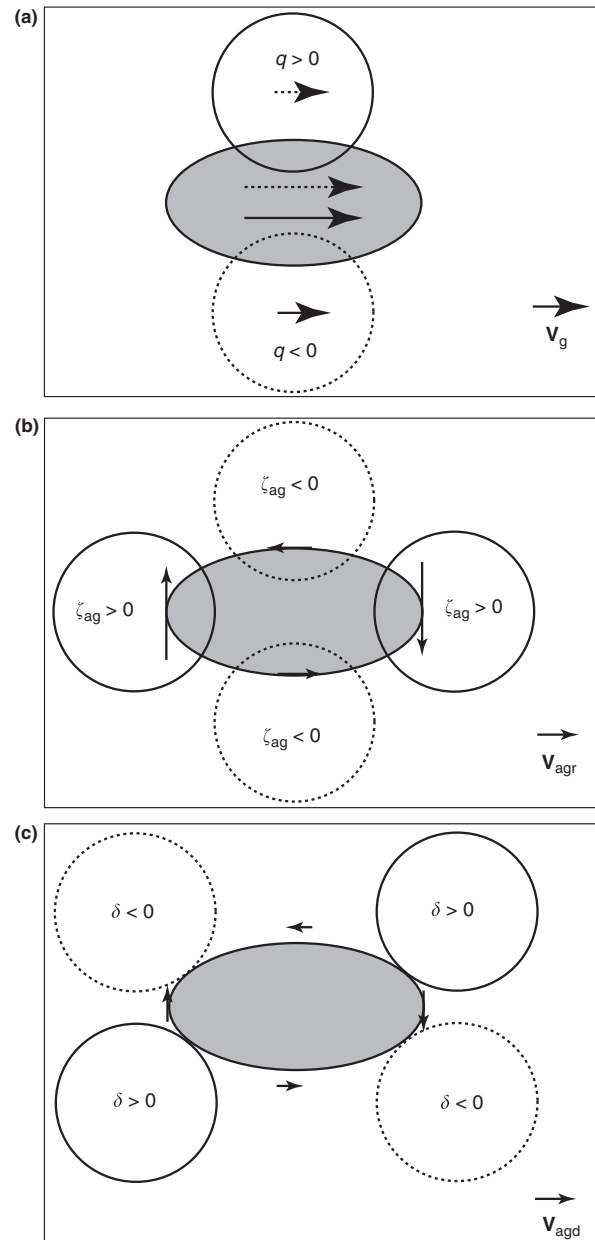


Figure 11 Schematic illustration of a straight jet streak (region of maximum wind speed shaded) represented in terms of a quasi-geostrophic vortex dipole in a uniform zonal background flow. (a) Quasi-geostrophic potential vorticity, q , and geostrophic velocities associated with respective constituent vortices, such that solid arrows indicate flow induced by cyclonic vortex ($q > 0$, solid line) and dotted arrows indicate flow induced by anticyclonic vortex ($q < 0$, dotted line). (b) Ageostrophic vorticity, ζ_{ag} ($\zeta_{ag} > 0$, solid line; $\zeta_{ag} < 0$, dotted line), and induced rotational ageostrophic velocity given by solid arrows. (c) Horizontal divergence δ ($\delta > 0$, solid line; $\delta < 0$, dotted line) and induced divergent ageostrophic velocity given by solid arrows. Vector scales of arrows are equivalent in (b) and (c), and an order of magnitude smaller than in (a).

the redistribution of momentum and energy between the troposphere and the stratosphere and mesosphere above, and as such may play a significant role in the general circulation of the middle atmosphere.

See also: **Aviation Meteorology:** Clear Air Turbulence. **Dynamical Meteorology:** Overview; Quasigeostrophic Theory. **Gravity Waves:** Buoyancy and Buoyancy Waves: Theory. **Mesoscale Meteorology:** Mesoscale Convective Systems; Severe Storms. **Numerical Models:** Mesoscale Atmospheric Modeling. **Stratosphere/Troposphere Exchange and Structure:** Global Aspects; Local Processes; Tropopause. **Synoptic Meteorology:** Cyclogenesis; Extratropical Cyclones; Forecasting; Fronts; Weather Maps.

Further Reading

- Bjerknes, J., 1951. Extratropical cyclones. In: Malone, T.F. (Ed.), *Compendium of Meteorology*. American Meteorological Society, Boston, pp. 577–598.
- Bluestein, H.B., 1986. Fronts and jet streaks: a theoretical perspective. In: Ray, P.S. (Ed.), *Mesoscale Meteorology and Forecasting*. American Meteorological Society, Boston, pp. 173–215.
- Bluestein, H.B., 1993. Synoptic–Dynamic Meteorology in Midlatitudes. In: *Observations and Theory of Weather Systems, Volume II*. Oxford University Press, New York.
- Carlson, T.N., 1991. *Mid-Latitude Weather Systems*. HarperCollins Academic, London.
- Holton, J.R., Haynes, P.H., McIntyre, M.E., et al., 1995. Stratosphere–troposphere exchange. *Reviews of Geophysics* 33 (4), 403–439.
- Keyser, D., 1986. Atmospheric fronts: an observational perspective. In: Ray, P.S. (Ed.), *Mesoscale Meteorology and Forecasting*. American Meteorological Society, Boston, pp. 216–258.
- Keyser, D., Shapiro, M.A., 1986. A review of the structure and dynamics of upper-level frontal zones. *Monthly Weather Review* 114, 452–499.
- Kocin, P.J., Uccellini, L.W., 1990. Snowstorms along the northeastern coast of the United States: 1955 to 1985. *Meteorological Monographs*, vol. 22. American Meteorological Society, Boston no. 44.
- Palmén, E., Newton, C.W., 1969. *Atmospheric Circulation Systems: Their Structure and Physical Interpretation*. Academic Press, New York.
- Riehl, H., collaborators, 1952. Forecasting in middle latitudes. In: *Meteorological Monographs*, vol. 1. American Meteorological Society, Boston no. 5.
- Shapiro, M.A., 1982. *Mesoscale Weather Systems of the Central United States*. Cooperative Institute for Research in Environmental Sciences (CIRES)/National Oceanographic and Atmospheric Administration (NOAA), University of Colorado, Boulder.
- Shapiro, M.A., Keyser, D., 1990. Fronts, jet streams and the tropopause. In: Newton, C.W., Holopainen, E.O. (Eds.), *Extratropical Cyclones, The Erik Palmén Memorial Volume*. American Meteorological Society, Boston, pp. 167–191.
- Uccellini, L.W., 1990. Processes contributing to the rapid development of extratropical cyclones. In: Newton, C.W., Holopainen, E.O. (Eds.), *Extratropical Cyclones, The Erik Palmén Memorial Volume*. American Meteorological Society, Boston, pp. 81–105.
- Uccellini, L.W., Koch, S.E., 1987. The synoptic setting and possible energy sources for mesoscale wave disturbances. *Monthly Weather Review* 115, 721–729.

Lake-Effect Storms

PJ Sousounis, AIR Worldwide, Boston, MA, USA

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

Lake-effect snow develops when sufficiently cold air passes over warm water. It is an integral part of the snowfall climatology for the Great Lakes Region of North America. The most intense snowfall situations occur when wind blows parallel to the long axis of a lake. Because the snow bands are only several kilometers wide but sometimes several hundred kilometers long, small changes in synoptic conditions can have big impacts on where the heaviest snow falls. Basic dynamics are well understood, but the specifics of which band type will form and where, are not. Forecasting lake-effect snow is still challenging. Lake-effect snow occurs in other parts of the world.

Introduction

Each winter, lake-effect storms develop on the downwind shores of the Great Lakes, as arctic winds blow across the relatively warm water. The associated clouds and snow (or rain) showers tend to organize in narrow bands, usually only a few kilometers wide but sometimes more than 200 km long. There may be one band, or there may be as many as 10 or 20, each separated by only a few kilometers of clear sky. These bands may remain stationary over a region or they may oscillate in a snakelike fashion. They may produce nothing more than 1 or 2 cm of snow, or they may dump more than 120 cm of snow in a single storm. These lake-effect storms are primarily a product of relatively simple cold air mass modification by warm water, complicated lakeshore geometry, and the prevailing synoptic situation.

Lake-effect storms develop in other parts of the United States, Canada, and the world, but nowhere else do they occur as frequently or with such intensity as they do in the Great Lakes region. The reasons for the unique weather in the Great Lakes region can be traced to several geographic aspects. The fact that the Great Lakes are the largest single source of fresh water in the world (except for the polar ice caps), the fact that the Great Lakes are situated approximately half way between the equator and pole, the fact that Great Lakes are located in the interior of a large continent, the fact that each of the lakes is approximately the size of a small inland sea, and the fact that there are several lakes – each separated by distances less than their own size, makes for some very unique weather in the region. Additionally, because of their depth, the lakes rarely freeze over completely, even in the coldest of winters, and thus remain a nearly continuous and very large source of heat and moisture for the atmosphere. Lake-effect storms continue to be a forecast challenge despite improvements in numerical weather prediction models because small changes in the prevailing synoptic conditions coupled with their meso- γ /meso- β scale size can lead to dramatic changes in snowfall totals for a given location.

Climatology

Lake-effect snow accounts for 25–50% of the total annual snowfall in many lakeshore regions (see [Figure 1](#)). The

snowbelts (areas of heavier snow) that shoulder the southern and eastern shores of the Great Lakes reflect the direction of the prevailing northwesterly flow relative to the orientation of the lakes, the sharp contrast in surface friction between the relatively smooth lake surface and the rough land, and terrain effects. The largest snowfall totals exist in the upper peninsula of Michigan, where northwesterly flow across Lake Superior is forced upward abruptly over steep terrain upon reaching the northern coast of the upper peninsula – especially around the Keweenaw Peninsula, and the Tug Hill Plateau in western New York, where west-southwesterly flow across the lower lakes provides a long fetch and ample opportunity for the air to be moistened and destabilized. In both of these locations, long fetches and orography are key aspects. Terrain can enhance individual snowstorm totals by about 5 cm for every 100 m of rise. Additionally, portions of the lakeshore with enhanced concavity promote convergence zones that can further enhance snowfall totals. Heavier lake-effect amounts typically fall during cold winters, when the lake-air temperature differences are enhanced.

Lake-effect snowfalls almost exclusively during the *unstable season* – that portion of the year when the lakes are climatologically warmer than the ambient air and thus provide heat and moisture to the lower atmosphere to destabilize it. Enhanced cloudiness and precipitation exist across much of the lakeshore regions and far inland as well. The percentage of cloudy days peaks in November for many places of the Great Lakes region – due in part to significant lake-enhanced cloudiness. Precipitation during the unstable season typically begins with episodes of nocturnal rain showers during cool nights in late August. As the mean air temperature drops through the fall months, lake-effect rain showers change to lake-effect snow showers. Much of the lake-effect snow typically falls between November and February, which constitutes the heart of the unstable season, when lake-air temperature differences tend to be greatest. Climatological lake-air temperature differences may be around 7–8 °C, but may exceed 30 °C during intense cold air outbreaks. Coupled with winds sometimes in excess of 20 m s⁻¹, combined surface sensible and latent heat fluxes can typically exceed 1000 W m⁻² – comparable to that found in a category 1 hurricane.

Lake-effect clouds and snow can occur locally on ~30–40% of the days in winter under a variety of synoptic

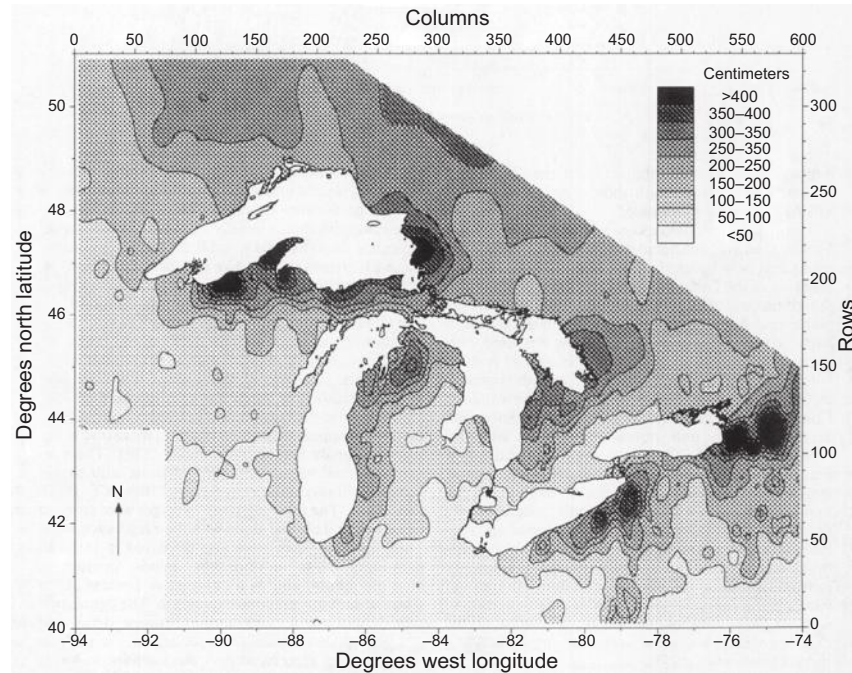


Figure 1 Average 1951–80 Great Lakes seasonal snowfall total. From Figure 2 in Norton, D.C., Bolsenga, S.J., 1993. Spatiotemporal trends in lake-effect and continental snowfall in the Laurentian Great Lakes, 1951–1980. *Journal of Climate* 6, 1943–1956. Adapted with permission from the American Meteorological Society.

patterns – whenever there is an onshore fetch and the lake-air temperature difference allows the lowest layers of the air to destabilize. However, certain synoptic patterns are more favorable than others for allowing lake-effect snow to develop. A typical sequence of events begins with a synoptic-scale low moving across the Great Lakes region from southwest to northeast (see [Figure 2](#)). Additionally, in late autumn especially, these lows are deepening as they cross the region because of baroclinic forcing and aggregate heating from all the lakes. Strong northwesterly winds on the back side of the low bring progressively colder polar or arctic air across the warm lakes. Subfreezing temperatures may reach as far south as the Gulf Coast and northern Florida, with -20°C readings just north of the lakes. The strong winds and cold air generate strong surface fluxes over the lakes that moisten and destabilize the air, leading to snow showers along the downwind lakeshores of the Great Lakes. The deepening of the low and the destabilization both allow stronger winds from above to mix down to the surface and further increase the heat and moisture fluxes.

Depending on the wind speed, the orientation of the wind flow relative to the long lake axis, stability, moisture, and upper-level forcing, different types of lake-effect storms can develop. Basically, when the prevailing flow is more parallel to the short axis than to the long axis of a lake (e.g., strong short-axis winds), multiple wind parallel bands (Type II) develop (see [Figure 3](#) middle panel). These bands are typically 2–4 km wide and spaced 5–8 km apart. Snowfall is usually spread over a large area of the downwind lakeshore and amounts are usually light (<4 mm liquid precipitation per day). When the prevailing short-axis winds are weak, midlake (Type I) or shore-parallel (Type IV) bands can

develop – even when the long-axis prevailing winds are strong (see [Figure 3](#) upper panel). These bands can be 10–20 km wide and generate copious amounts of snow. If the short-axis wind is essentially not present, then the band will be located near the middle of the lake. If the short-axis prevailing wind is present but weak, then the band will be located closer to the downwind shore. Sometimes, especially over Lakes Erie or Ontario, a midlake band will develop at an obtuse angle to the long axis of the lake. These bands (Type III) actually develop as midlake bands in northwesterly flow over Lake Huron, farther upwind (see [Figure 3](#) lower panel). They may lose their visible cloud characteristics over southern Ontario but maintain the convergence zone so they redevelop once they reach the lower lakes. If the wind is very light against areas of enhanced concavity, then lake vortices (Type V) may develop (see [Figure 3](#) upper panel). Stretching and tilting of vorticity from low-level convergence and vertical wind shear; differential diabatic heating; and synoptic-scale vorticity and temperature advection are thought to play roles in this type of lake-effect storm. [Figure 4](#) summarizes the relationship of storm morphology to lake-air temperature difference and wind speed.

The most common type of lake-effect event over the northern lakes (Superior, Huron, and Michigan) is the multiple band variety, which occurs 50% of the time. The next most common is the shore-parallel or midlake variety, which occurs 25% of the time. The remaining 25% includes meso-scale vortices, hybrid combinations, and undeterminable forms. The most common type of lake-effect event over the southern lakes (Erie and Ontario) is the shore-parallel or midlake variety, because of the different orientation of these lakes (see [Figure 5](#)). In general, the frequency of multiple

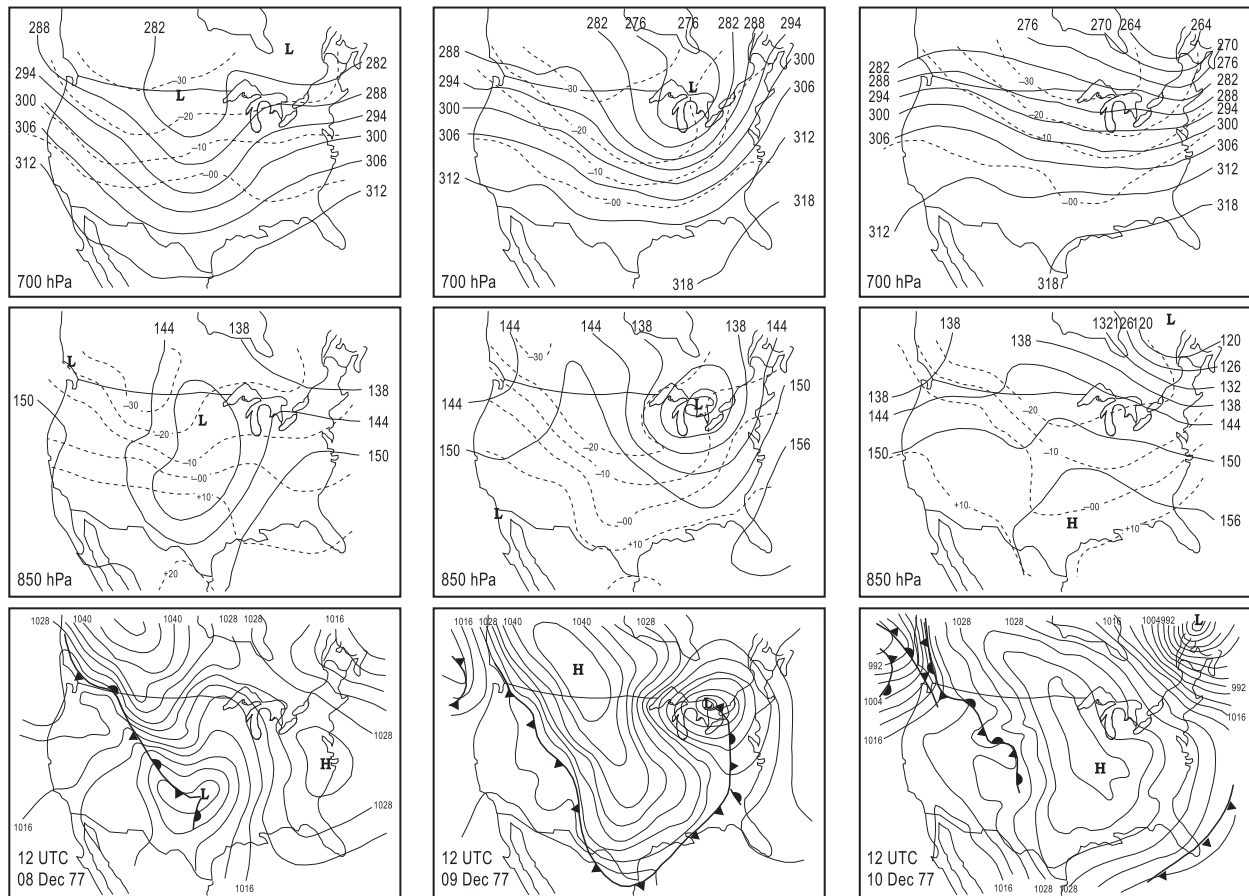


Figure 2 Surface and upper air analyses depicting typical synoptic setting for lake-effect snow in the Great Lakes. (upper row) 700 hPa heights (solid, dm) and temperatures (dashed, °C); (middle row) 850 hPa heights (solid, dm) and temperatures (dashed, °C); (bottom row) sea-level pressure (solid, hPa) and locations of highs, lows, and fronts for times shown. From Figure 2 in Braham Jr., R.R., 1983. The midwest snow storm of 8–11 December 1977. Monthly Weather Review 111, 253–272. Adapted with permission from the American Meteorological Society.

bands decreases from west to east and the frequency of shore-parallel and midlake bands increases from west to east across the region.

Boundary Layer Dynamics

When cold air flows across the warm waters of the Great Lakes, strong sensible and latent heat fluxes warm and moisten the air closest to the surface first, causing the lowest levels of the atmosphere to destabilize. Strong turbulent motions mix upward the warmed and moistened air in convective fashion. Steam fog typically develops and steam devils may be visible near the surface, especially within a few tens of kilometers of fetch. A very unstable convective internal boundary layer (CIBL) forms near the surface and grows rapidly upward in the downwind direction (see Figure 6). The upwind temperature, humidity, and flow characteristics of the air before it reaches the lake determine how the air will be modified. As the air crosses the downwind lakeshore, frictional convergence enhances ascent.

Convective updrafts can exceed $4\text{--}5\text{ m s}^{-1}$ in narrow, 100-m wide cores. After only a few kilometers of fetch, the depth of the internal boundary layer may reach or exceed the depth of

the planetary boundary layer. At some point, the air within the thermal internal boundary layer becomes moist enough and deep enough so that the lifting condensation level is below the top of the boundary layer and cloud forms (Type I boundary layer). Subsequent latent heat release and radiative heat transfer become important and increase the rate of entrainment. The clouds develop initially as two-dimensional bands but in time, with sufficient fetch and heating, these bands can develop into chains of three-dimensional cells and may eventually develop into a regular array of three-dimensional mesoscale cellular convection. Maximum cloud drop concentration and liquid water content (e.g., 0.25 g m^{-3}) occur near cloud-top and increase over the first two-thirds of fetch. Near the downwind lakeshore, snow particle production increases (e.g., 5 l^{-1}) to reduce drop concentration and increase the height of cloud base.

The most common type of convection associated with cold air outbreaks over the Great Lakes is longitudinal rolls. This type occurs in other parts of the world – wherever cold air crosses warm water. Over the Great Lakes, the rolls are usually oriented parallel to the wind at the base of the inversion and also to the low-level wind shear. Classical rolls are typically 2–5 km wide and 0.5–2 km deep so the aspect ratio is about

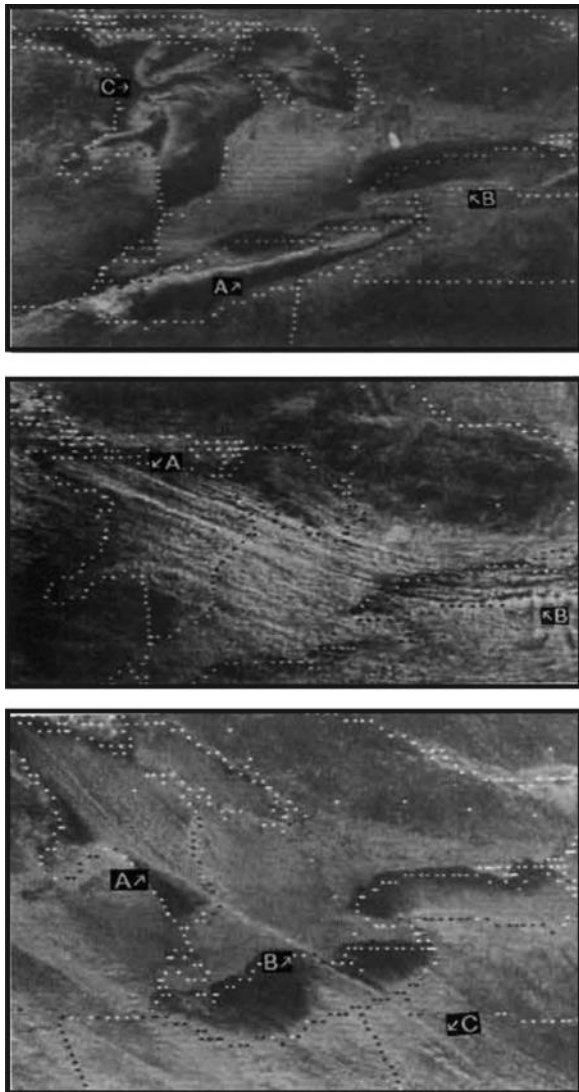


Figure 3 Lake-effect snow band structures. Upper panel shows a lake vortex (Type V) over Lake Huron (C); midlake band (Type I) over Lake Erie (A); shore-parallel band (Type IV) over Lake Ontario (B). Middle panel: multiple bands (Type II) over Lake Huron (A) and shore-parallel band (Type IV) over Lake Ontario (B). Lower panel: midlake band (Type I) over Lake Huron (A); hybrid band (Type III) over Lake Erie (B and C). Top panel from Figure 6 and middle and bottom panels from Figure 5 in Niziol, T.A., Snyder, W.R., Waldstreicher, J.S., 1995. Winter weather forecasting throughout the Eastern United States. Part IV: Lake-effect snow. *Weather and Forecasting* 10, 61–77. Adapted with permission from the American Meteorological Society.

2–5. Rolls over the Great Lakes are 10–20 km wide and 1–2 km deep so the aspect ratio is about 10. Some roll circulations exhibit a multiscale configuration so that the cloud streets are best developed when the wavelengths of the convection and rolls are in phase. Such multiscale phasing can generate cloud streets with aspect ratios of 10–20. In many instances, other cloud structures are simultaneously present in addition to the longitudinal rolls (see Figure 7).

Linear analytic studies indicate that three different types of instability mechanisms may be responsible for rolls. First,

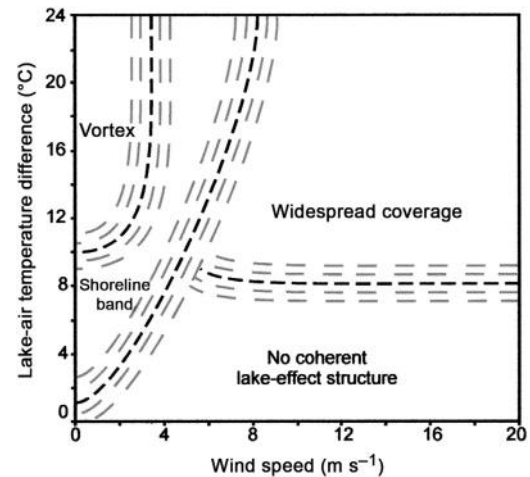


Figure 4 Diagram showing general conditions favorable for different lake-effect morphological regimes using wind speed vs lake-air temperature difference parameter space. The diagram is a composite based on the existing lake-effect literature. The dashed lines represent nondistinct boundaries between regimes. From Figure 5 in Laird, N.F., Kristovich, D.A.R., John E.W., 2003. Idealized model simulations examining the mesoscale structure of winter lake-effect circulations. *Monthly Weather Review* 131, 206–221. Adapted with permission from the American Meteorological Society.

inflection point instability theory suggests that a dynamical instability can develop in a (rotating) Ekman layer that is neutrally stratified if the cross-geostrophic wind component exhibits an inflection point. The most unstable wind profiles lead to rolls with an aspect ratio of 3. Rolls are best developed when they are oriented about 14° to the left of the geostrophic wind (in neutral conditions). Second, parallel instability theory involves the curvature of wind speed profiles parallel to the roll axes and the boundary layer mean wind shear. This type of instability mechanism generates rolls with aspect ratios that are twice as large as those from inflection point instability and closer to those observed during lake-effect conditions. Third, convective instability (i.e., an unstable boundary layer) in the presence of wind shear also leads to the formation of rolls that are parallel to the mean wind shear but with aspect ratios that are smaller than those from inflection point instability (e.g., around 2).

Understanding roll development during lake-effect situations is further complicated by the fact that large-eddy numerical simulations suggest that only convective instability is capable of generating rolls. Gravity waves generated within the stable air above the convective boundary layer by wind shear near the boundary layer top have also been suggested as a possible mechanism. If the wind shear at the top of the boundary layer is roughly perpendicular to the boundary layer wind direction, then bands of convection that are parallel to the mean boundary layer wind can be induced by gravity waves. Latent heat release, cloud-microphysical processes, and low-level wind shear (e.g., below 200 m) may also influence the development of rolls. The partial agreement between observations and theoretical studies thus far suggests that a combination of mechanisms may be responsible for different aspects of roll development or during certain conditions.

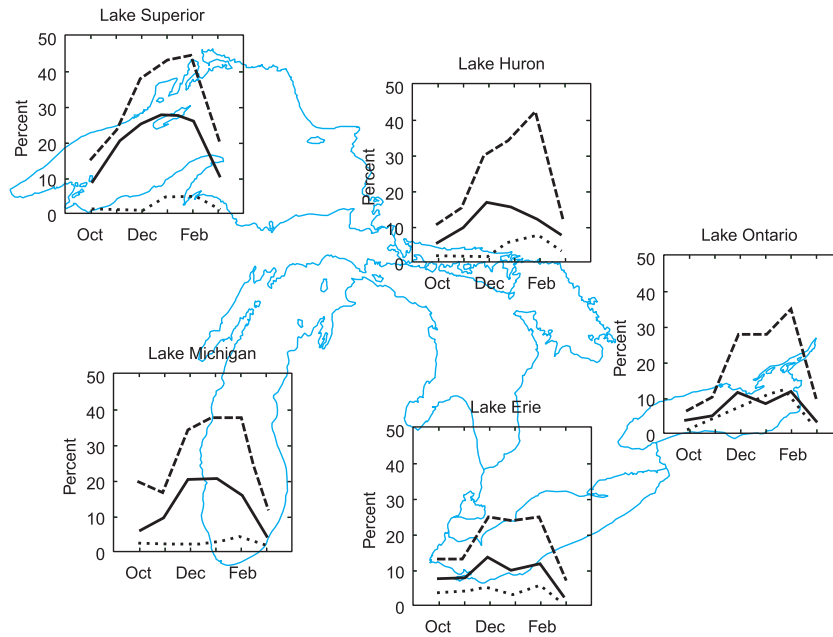


Figure 5 Percentages of all days categorized as lake-effect bands (dashed), wind parallel bands (solid), and shore-parallel bands (dotted). Results based on visible satellite imagery from 1988 to 1993 (October–March). From Figure 2 in Kristovich, D.A.R., Steve III, R.A., 1995. A satellite study of cloud-band frequencies over the Great Lakes. *Journal of Applied Meteorology* 34, 2083–2090. Adapted with permission from the American Meteorological Society.

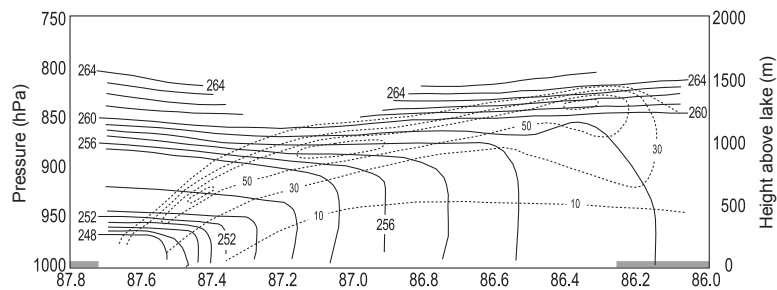


Figure 6 Vertical cross-section of equivalent potential temperature (solid 1 K contour interval) and cloud frequency (dashed 10% contour interval) across southern Lake Michigan (from Sheboygan, WI, to Benton Harbor, MI) during a cold air outbreak (on 20 January 1984). Presence of cloud determined by cloud droplet concentrations greater than 10 cm^{-3} (not shown). Wisconsin and Michigan land surfaces are indicated by thick horizontal gray bars. From Figure 5 in Chang, S.S., Braham Jr., R.R., 1991. Observational study of a CIBL over Lake Michigan. *Journal of Atmospheric Sciences* 48, 2265–2279. Adapted with permission from the American Meteorological Society.

Sensitivity to Synoptic Conditions

The thermal modification of air over relatively cold land as it crosses a relatively warm lake results in a horizontal temperature gradient across the lakeshore, which can generate a thermally direct solenoidal circulation similar to that of a sea/land breeze. In that sense, it can be seen that the lake-air temperature difference, speed and direction of the prevailing wind, and height and strength of the capping inversion are perhaps the most crucial parameters for determining not only how much lake-effect snow will fall but also how it will fall. For example, for conditions characterized by moderate lake-air temperature differences ($>10^\circ\text{C}$) and strong winds ($>10 \text{ m s}^{-1}$) blowing across the short axis of a lake, multiple snow bands will usually develop. If the winds across the short axis are lighter and/or the temperature difference is greater, then some form of shore-parallel band will develop. If the wind is very light against

areas of enhanced concavity, then lake vortices can develop (see Figure 4).

The impact of the wind speed on the lake-effect response characteristics can be analyzed two-dimensionally in terms of the Froude Number $Fr = \pi U/NH$, where U is the mean wind speed across the short axis of the lake, and N and H are the Brunt–Väisälä frequency and depth of the planetary boundary layer respectively. The values for N and H depend on the lake-air temperature difference and stability of the prelake-modified air. The Froude number may be interpreted as the ratio of the mean wind speed U to the gravity wave speed $c_g = NH/\pi$ for a boundary layer of depth H . Three regimes are important to consider. When $Fr < 1$, the gravity wave speed exceeds the mean flow speed ($c_g > U$). Opposing sea-breeze type circulations develop with respect to the short axis of the lake and the heaviest precipitation falls over the lake (see upper panel in Figure 8). This regime corresponds to the midlake band

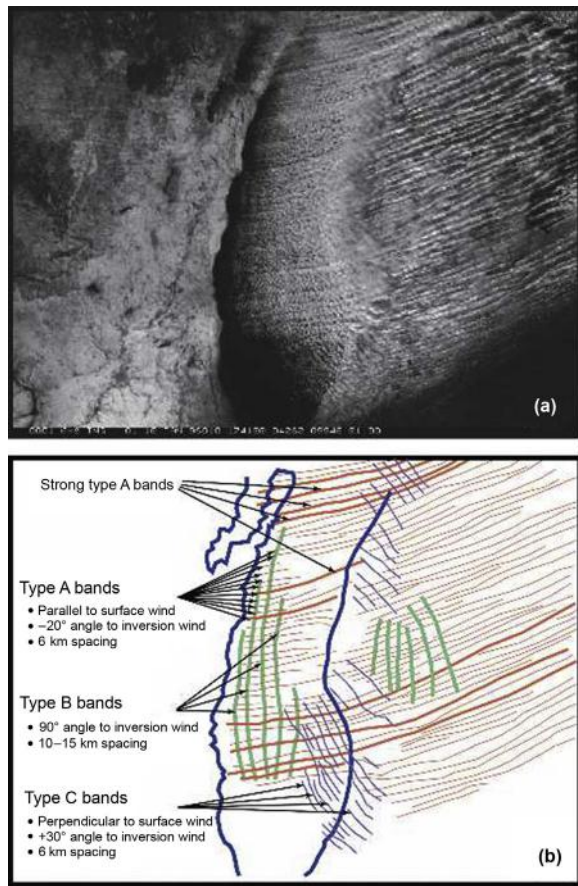


Figure 7 Superrapid scan Geostationary Operational Environmental Satellite (GOES) visible imagery (1-km resolution) for 1743 UTC on 10 Jan 1998. (a) Visible image over Lake Michigan region; (b) analysis of cloud structures in image. Major categories of cloud structures are labeled as type A–C bands with their approximate spacing and orientation relative to winds at the inversion level. From Figure in Tripoli, G.J., 2005. Numerical study of the 10 January 1998 lake-effect bands observed during lake-ICE. *Journal of Atmospheric Sciences* 62, 3232–3249. Adapted with permission from the American Meteorological Society.

type of event. When $Fr > 1$, the gravity wave speed is less than the mean flow speed ($c_g < U$). The response is characterized by alternating regions of ascent and descent that propagate downwind from the leeward shore, and precipitation is diffuse and weak. This regime corresponds to the multiple band type of event (see middle panel in Figure 8). When $Fr = 1$, the gravity wave speed equals the mean flow speed ($c_g = U$). A resonance condition develops where gravity waves that are generated at the downwind shore cannot propagate upwind. This regime corresponds to the shore-parallel band type of event that can generate significant precipitation at the downwind lakeshore (see lower panel in Figure 8). Setting $Fr = 1$, and using typical values of $N = 10^{-2} \text{ s}^{-1}$ and $H = 2 \text{ km}$, suggests that a value of $U \sim 6 \text{ m s}^{-1}$ maximizes heavy snow along the downwind lakeshore, which is consistent with observations.

The impact of fetch on lake-effect storm development has been known since the early 1900s. Long fetches usually result in heavy snowfalls. Short fetches also may produce significant

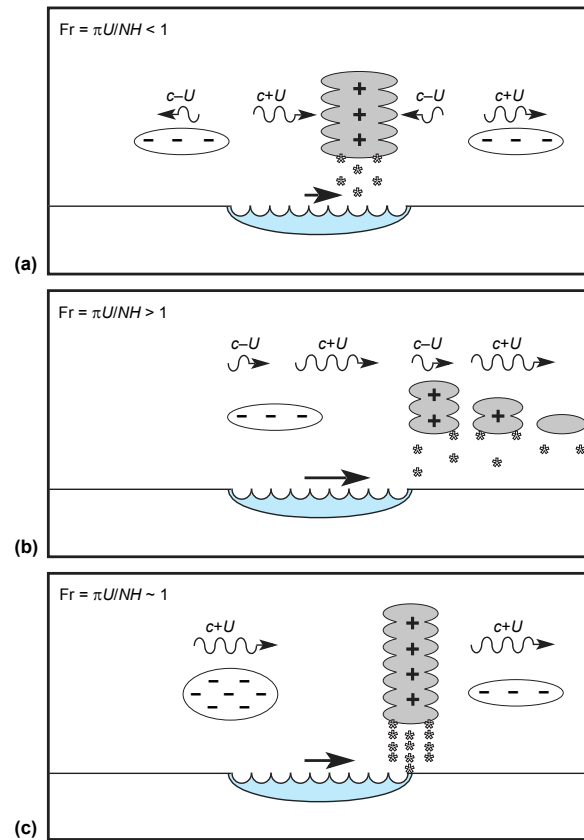


Figure 8 Gravity wave interpretation of lake-effect morphology dependence on wind speed. (a) Weak wind speeds create subcritical ($Fr < 1$) regime, which allows gravity waves to propagate upwind and downwind and a midlake band and moderate snow to develop over the lake; (b) strong wind speeds create supercritical ($Fr > 1$) regime, which allows gravity waves to propagate only downwind and multiple bands and light snow to develop beyond the downwind lakeshore; (c) moderate wind speeds create near-critical ($Fr \sim 1$) regime, which allows gravity waves to travel downwind but traps gravity waves trying to propagate upwind – resulting in an intense shore-parallel band and heavy snow to develop along the downwind lakeshore. Heavy arrows indicate wind speed and wavy arrows indicate gravity wave propagation. Plus (+) signs and shaded columns indicate ascent; minus (−) signs and open ovals indicate descent. Asterisks indicate snowfall.

snowfalls if the prelake-modified air is relatively unstable, if the lakeshore geometry enhances radial convergence, or if the nearby orography enhances lifting. These observed impacts of fetch have only been recently confirmed using analytic and numerical models. For example, it has been shown analytically for Lake Michigan that three convergence centers develop near the eastern shore when a westerly wind prevails, two cells or snow bands develop when a northwesterly wind prevails, and one midlake band develops when a northerly wind prevails. In some instances, the Rossby Number Ro , another nondimensional parameter, defined as U/fL , where f is the Coriolis parameter and L is the characteristic length scale of the lake (fetch), is useful for understanding lake-effect storm morphology. Ro can be thought of as the ratio of the time for Earth's rotation to become effective to the advective time it takes for an air parcel to cross the lake. Low values correspond

to a vortex structure, medium values correspond to a shoreline band, and high values yield multiple bands.

The vertical structure of the environmental wind also affects lake-effect storms. For example, when the prevailing wind is parallel to the long axis of Lake Erie, moderate directional shear (e.g., between 30 and 60°) from the surface to 700 hPa causes weakly to moderately precipitating multiple snow bands rather than a single intensely precipitating snow band to occur. Stronger shear (e.g., greater than 60°) over Lake Erie causes the breakdown of precipitating snow bands altogether – allowing only instead the development of a nonprecipitating stratocumulus deck. While wind shear can understandably inhibit the organization of rolls or wider bands, it is not clear whether some of the observed effects from shear are simply a manifestation of a shallow boundary layer.

The height and strength of the (capping) inversion are significant limiting factors to cloud depth and therefore to precipitation. Typically, the boundary layer must have a depth greater than 1 km in order for lake-effect snow to develop. The most convectively active lake-effect storms have inversion heights exceeding 3 km. Sometimes the capping inversion is entirely absent. During such cases, thunder and lightning typically accompany copious (e.g., exceeding 10 cm h^{-1}) snowfall rates.

The vertical temperature and moisture distributions within the boundary layer also play a role. For example, it has long been known that a minimum temperature difference of 13°C between the lake surface and the upstream airflow at 850 hPa is required for lake-effect storms to develop. This temperature difference criteria means that the lapse rate should be unstable with respect to unsaturated ascent. A dry boundary layer is less conducive for lake-effect snow than a moist one, although a long fetch can compensate for very dry boundary layers. The impact of moisture is greater for low-stability profiles than for high-stability ones.

The presence of large-scale forcing can also influence lake-effect storm development. Typically, the coldest air passes over the lakes as high pressure at the surface moves eastward across them, accompanied by negative vorticity advection at upper levels and cold advection near the surface. Thus, the impacts of synoptic-scale forcing typically act to suppress lake-effect storm development (see Figure 2). There are however instances when cold air, positive vorticity advection and even warm advection exist simultaneously over the region. Such situations usually come in the form of Alberta Clippers (short waves) that develop in cold air masses and move southeastward across the region. The synoptic forcing coupled with the cold air that is already established over the region can combine to generate intense snowfall.

Sensible and latent heating from all the Great Lakes (e.g., the lake aggregate) can also influence lake-effect storms over individual lakes. Basically, if warming (and moistening) occurs over all the Great Lakes for at least a day, then surface pressures and stability can drop over a broad region and cause a perturbation aggregate-scale, low-level cyclonic circulation to develop. The position, the size, and the warmth and moisture from this aggregate circulation can modify lake-effect precipitation throughout the region. Specifically, when the synoptic-scale flow is northwesterly, aggregate effects can augment snowfall along the northwestern shores of lower Michigan, and reduce snowfall along the southwestern shores (cf Figure 9).

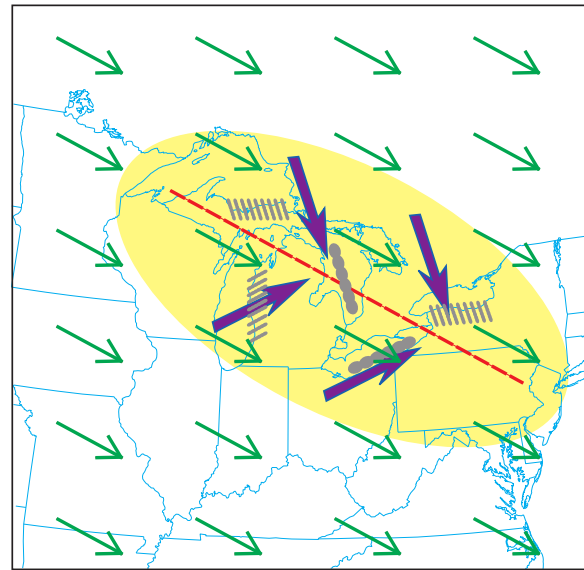


Figure 9 Illustration of lake-aggregate effect on prevailing winds and lake-effect snowstorms. On the south side of the developing warm plume (shaded oval), northwest winds respond in sea-breeze fashion to become southwest winds with increased fetch and heavy snow across Lake Erie. Lake-effect snows across portions of western Michigan (lower peninsula) may or may not change characteristics. On the north side, northwest winds respond to become north-northwesterly winds with reduced fetch and light snow across Lakes Superior and Ontario and increased fetch heavy snow across Lake Huron. Heavy and light snow bands indicated by solid gray-shaded scalloped and gray-striped cloud strips, respectively.

Shore-parallel bands located offshore can migrate eastward (e.g., onshore) or evolve into multiple bands. These aggregate affects over Lake Michigan include enhanced westerly flow, increased heat and moisture, and lower stability. The lake aggregate can also influence lake-effect precipitation in the lower lakes region. For example, as the lake-aggregate-induced plume of heat and moisture extends southeastward, surface winds across Lake Ontario (north of the aggregate-induced plume) can become more northerly. In contrast, surface winds across Lake Erie (south of the aggregate-induced plume) can become more westerly. The aggregate-altered winds can cause a longer fetch across Lake Erie and a shorter fetch across Lake Ontario, and can shift the regions of lake-effect convective bands, so that less and less-intense lake-effect precipitation can fall along the lakeshores downwind (east) of Lake Ontario and more lake-effect precipitation can fall along the eastern shores of Lake Erie (cf Figure 9).

Forecasting

The mesoscale nature of lake-effect storms, their intensity, and the short development times continue to challenge forecasters. Highly variable snow-to-liquid ratios ($10:1$ – $50:1$) and terrain effects, especially near Lakes Erie and Ontario, can enhance the inherently large spatial variability of lake-effect snow and hence the forecast challenge. While the problems of forecasting when lake-effect snow is going to occur have essentially been solved, the equally significant problems of exactly where lake-effect

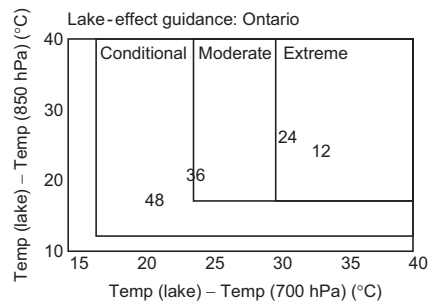
snow will occur, what form(s) it will take, how intense it will be, and how long it will last remain outstanding forecast issues. A combination of high-resolution numerical weather prediction models, statistical methods, Doppler Radar, and forecaster savvy are the basic forecast tools.

Numerical models have come a long way since the use of the Limited Fine Mesh (LFM) model. The coarse resolution (180 km) and the exclusion of the lakes in terms of their heat, moisture, and momentum characteristics in that model precluded any explicit model development of lake-effect precipitation. Regardless, operational forecasters relied on this model because of its ability to forecast the large-scale conditions, to which lake-effect snowstorm development is very sensitive. The LFM, in conjunction with forecaster decision trees based on key large-scale parameters, and experience, allowed forecasters to at least be able to issue general forecasts of when lake-effect snow was going to occur. Two operational models currently being used include the North American Model with 4 km resolution over the continental United States that runs four times daily out to 84 h, and the Rapid Refresh Weather Research and Forecast (WRF) Model with 13 km resolution that runs hourly out to 18 h.

Despite significant improvements in numerical weather prediction and the high-resolution model output provided by the National Centers for Environmental Prediction (NCEP), lake-effect snow continues to challenge the abilities of even

the most sophisticated numerical forecast models because of several inadequacies. These inadequacies include horizontal resolution that is still too coarse for resolving the 2–4 km wide bands, convective schemes tuned originally for deep (tropical) convection that are inappropriate to simulate intense shallow precipitating convection, and boundary layer schemes that are too simplistic to develop the low-level temperature, moisture, and cloud-microphysical structures that exist within lake-effect snow environments near the surface.

To supplement the NCEP information, forecasters at several National Weather Service Forecast Offices (WSFOs) in the Great Lakes region run their own versions of WRF, based on the WRF Environmental Modeling System. These regionally run versions include better-defined Lake Surface Temperatures and ice presence through NASA-SPoRT, and the ability for local forecasters to run simulations at very high (1 km) resolution and to generate ensembles that provide probabilities for different forecast scenarios. Additionally, forecasters still use various statistical methods. These methods were used almost exclusively prior to the existence of numerical models. As early as the middle of last century, various investigators had outlined conditions necessary for prolonged lake-effect storms to occur at the eastern end of Lake Erie (cf Figure 10). Highly sophisticated statistical methods, such as BUFKIT, are still used today in conjunction with numerical forecast model output to optimize the accuracy of lake-effect snow forecasts.



12 h extreme instability...030 degree shear from 700 hPa to SFC...and 090-mile fetch at 850 hPa.
 24 h moderate instability...020 degree shear from 700 hPa to SFC...and 070-mile fetch at 850 hPa.
 36 h moderate instability...000 degree shear from 700 hPa to SFC...and 090-mile fetch at 850 hPa.
 48 h moderate instability...090 degree shear from 700 hPa to SFC...and 150-mile fetch at 850 hPa.

Output from NGM, 12Z 18 Feb 1993: Forecast parameters				
Variable	12 h	24 h	36 h	48 h
Wind direction (degrees)				
700 hPa	260	310	290	270
850 hPa	290	300	290	240
B.L.	290	290	290	180
Change in wind direction with height				
850/700	030	010	000	030
B.L./700	030	020	000	090
Fetch (miles)				
850 hPa	090	070	090	150
Temp (°C)				
700 hPa	-30	-28	-21	-18
850 hPa	-21	-23	-17	-14
B.L.	-10	-14	-10	-09
Lake	02	02	02	02
TS/T3 layer inversion intensity				
TS-T3	-08	-08	-03	-05
Vertical velocity (microbar/s)				
700 hPa	+02	-02	-02	+00

Figure 10 Lake-effect snow guidance product for 1200 UTC on 18 February 1993 generated for Lake Ontario at WSFO Buffalo. TS and T3 correspond to the surface and 900 hPa temperatures, respectively and BL corresponds to the boundary layer. From Figure 10 in Niziol, T.A., Snyder, W.R., Waldstreicher, J.S., 1995. Winter weather forecasting throughout the eastern United States. Part IV: Lake effect snow. *Weather and Forecasting* 10, 61–77. Adapted with permission from the American Meteorological Society.

See also: **Boundary Layer (Atmospheric) and Air Pollution:** Convective Boundary Layer. **Mesoscale Meteorology:** Cloud and Precipitation Bands; Convective Storms: Overview. **Mountain Meteorology:** Land and Sea Breezes. **Numerical Models:** Mesoscale Atmospheric Modeling. **Synoptic Meteorology:** Polar Lows; Weather Maps.

Further Reading

- Braham Jr., R.R., 1983. The midwest snow storm of 8–11 December 1977. *Monthly Weather Review* 111, 253–271.
- Chang, S.S., Braham Jr., R.R., 1991. Observational study of a convective internal boundary layer over Lake Michigan. *Monthly Weather Review* 48, 2265–2279.
- Kristovich, D.A.R., Steve III, R.A., 1995. A satellite study of cloud band frequencies over the Great Lakes. *Journal of Applied Meteorology* 34, 2083–2090.
- Laird, N.F., Kristovich, D.A.R., Walsh, J.E., 2003. Idealized model simulations examining the mesoscale structure of Winter Lake-effect circulations. *Monthly Weather Review* 131, 206–221.
- Niziol, T.A., Snyder, W.R., Waldstreicher, J.S., 1995. Winter weather forecasting throughout the United States. Part IV: Lake-effect snow. *Weather Forecasting* 10, 61–77.
- Norton, D.C., Bolsenga, S.J., 1993. Spatiotemporal trends in lake-effect and continental snowfall in the Laurentian Great Lakes, 1951–1980. *Journal of Climate* 6, 1943–1956.
- Sousounis, P.J., Mann, G.E., 2000. Lake-aggregate mesoscale disturbances. Part V: Impacts on lake-effect precipitation. *Monthly Weather Review* 128, 728–743.
- Tripoli, G.J., 2005. Numerical study of the 10 January 1998 lake-effect bands observed during lake-ice. *Journal of Atmospheric Sciences* 62, 3232–3249.

Polar Lows

IA Renfrew, University of East Anglia, Norwich, UK

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

Polar lows are subsynoptic-scale cyclones that develop poleward of the main polar front over the ocean. Polar lows are the most intense variety of a general classification of polar mesoscale cyclones. They are typically less than 1000 km in scale and last for just 12–48 h, although they can have associated surface wind speeds of over 30 m s^{-1} (60 knots). Their scale and brief life cycle make them difficult to observe and forecast, therefore a genuine maritime hazard. Their role in the climate system and their outlook in a changing climate are topics of current research.

Introduction

Through the years, mariners of the Nordic Seas have told tales of unexpected encounters with fierce storms that appeared out of nowhere to wreak havoc on the seas. These storms were smaller and more transient than the more predictable synoptic-scale weather systems that they were used to; yet with the advent of the satellite era, it appeared these smaller scale vortices were ubiquitous over the polar seas. The high latitude location of these storms, and their identification as low pressure centers, led to them becoming known as polar lows; although a host of alternative names have been used in reference to such weather systems, for example, Arctic instability lows, comma clouds, and, in the Southern Hemisphere, Antarctic coastal vortices. Polar lows are now defined as intense maritime cyclones with scales less than 1000 km and near-surface wind

speeds in excess of 15 m s^{-1} . Polar lows are the most intense example of the family of mesoscale cyclonic vortices, poleward of the main polar front, known generically as polar mesoscale cyclones. This article will discuss both, but will focus on those more energetic systems known as polar lows.

Polar lows are characterized as subsynoptic in scale, typically, 100–500 km in diameter, and short lived, lasting only 6–48 h. They develop over water, but often move over land or ice where they tend to fade rapidly. In terms of dynamics, polar lows are fundamentally baroclinic or convective in nature. In most cases, some element of convection is necessary for rapid development to an intense polar low. Purely baroclinic, or topographically forced, systems tend to remain as weaker polar mesoscale cyclones. One of the most prominent hallmarks of a polar low is the spiral of cloud associated with the vortical flow (**Figure 1**).

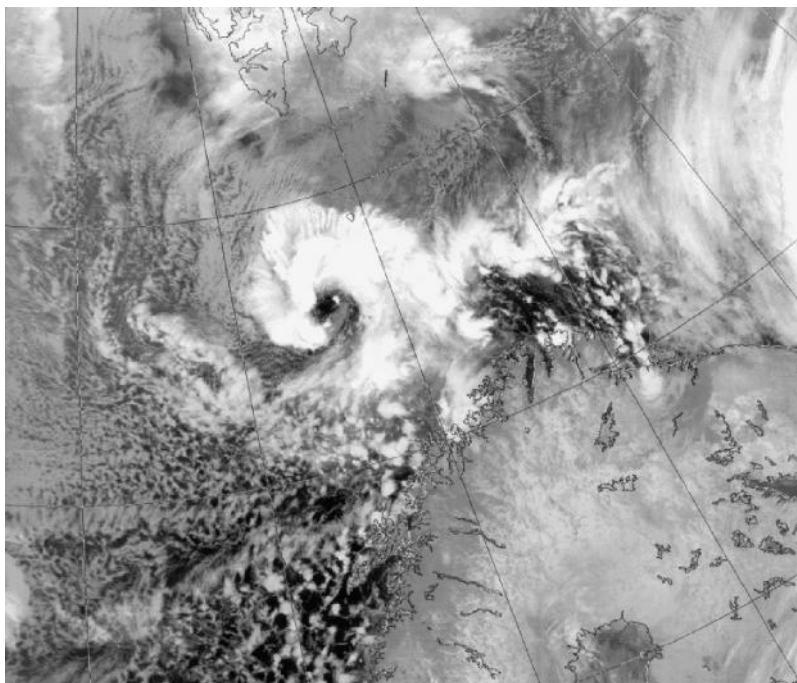


Figure 1 Infrared satellite image of the Barents Sea area at 02:40 UTC 13 December 1982. Light colors represent cold brightness temperatures. A polar low is well defined by a cloud vortex with a clear central eye. Image courtesy of the NERC Satellite Receiving Station, University of Dundee.

Occasionally, there is a clear eye at the center of the cloud vortex suggesting an analogy with tropical cyclones, indeed such systems have been described as Arctic hurricanes. More often than not they are warm-cored vortices, and many have well-defined fronts, suggesting parallels with synoptic-scale extratropical cyclones.

The distribution around the globe of the landmasses, sea ice, and sea surface temperatures leads to some favored locations for the development of intense polar lows. These are regions prone to cold-air outbreaks, where relatively cold continental air is advected over relatively warm ice-free waters. For example, in the Nordic Seas (the Greenland, Iceland, Norwegian, and Barents seas), the Irminger Sea, the Labrador Sea, the Bering Sea, the Gulf of Alaska, and the Sea of Japan. In the Southern Hemisphere, the eastern Weddell, the Bellingshausen, and the Ross seas are favored locations. Satellite- and model-based climatological studies show there are typically hundreds of polar mesoscale cyclones per year in each of these regions, with perhaps 10–20% being more intense polar lows. In the Northern Hemisphere, polar mesoscale cyclone development is most common during winter, followed by the autumn and spring seasons. In the Southern Hemisphere, there appears to be no significant seasonal variability, but there are preferences in location that tend to be dictated by variations in the sea ice distribution.

Polar lows are extremely important weather systems for certain regions as their strong winds and severe weather are potentially hazardous. However, it is also likely that polar mesoscale cyclones play an important, but as yet unclear role in the climate system at high latitudes; for example, through a strong coupling of the polar atmosphere and ocean through air–sea heat exchange. It has been suggested that the distribution of polar lows will move poleward under current climate

change projections, leading to fewer impacting Norway for example.

Observations of Polar Low Structure

Polar mesoscale cyclones are relatively small scale and short lived, which make them difficult to observe by conventional means as they tend to fall between the spatial and temporal gaps in the observing network. Their location in sparsely populated or maritime regions compound this difficulty. With the advent of the satellite era, and the development of satellite remote sensing as a tool for meteorological observation, our knowledge has vastly improved; not least in simply providing visible and infrared images showing the cloud structures associated with the systems (e.g., [Figures 1 and 2](#)).

Instrumented Aircraft Data

The availability of satellite imagery in real time has allowed the targeting of incipient polar lows for further investigation by instrumented aircraft. Only a very small number of polar lows have been investigated by aircraft, but these few cases provide otherwise unobtainable details of their structure. One case from over the Norwegian Sea in February 1984 is discussed as an illustration. [Figure 3](#) shows an infrared satellite image: light colors represent colder brightness temperatures and therefore generally higher clouds. There is a high comma-shaped cloud spiraling into the center of a mature polar low and contrasting with the warmer brightness temperatures of the sea surface beneath. To the south and east there are plumes of cloud associated with shallow convection. Further

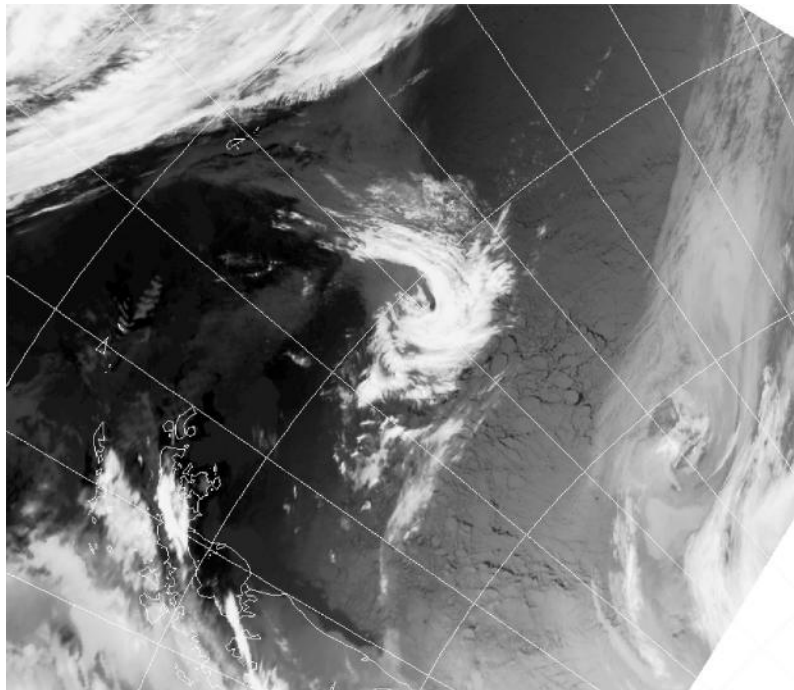


Figure 2 Infrared satellite image of the sea ice covered Weddell Sea area at 17:30 UTC 6 October 1995. A polar mesoscale cyclone is clearly defined by the comma-shaped cloud vortex. Image courtesy of the British Antarctic Survey.

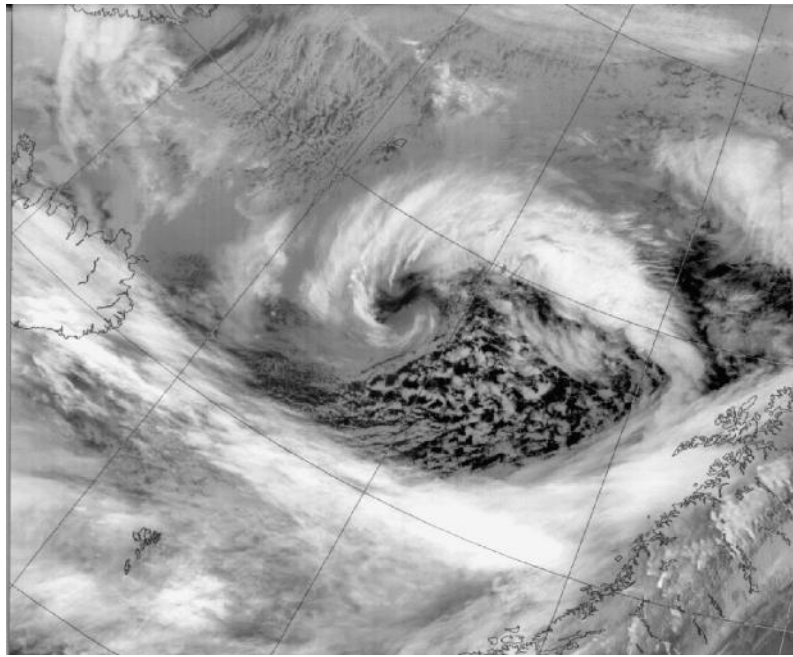


Figure 3 Infrared satellite image of the Norwegian Sea area at 13:40 UTC 27 February 1984. A polar low with a well-defined clear central eye is visible. Image courtesy of the NERC Satellite Receiving Station, University of Dundee.

to the south is an elongated band of high cloud associated with a synoptic-scale weather system.

An instrumented aircraft made a number of passes through the polar low over a period of a few hours centered at this time. Atmospheric soundings from dropsonde releases and nearby radiosonde ascents, along with the flight-level data, have been synthesized together to produce a highly detailed picture of this particular polar low. A surface pressure analysis, along with part of the aircraft track and wind observations at 300 m above sea level, is shown in [Figure 4](#). Note the analysis has been time-adjusted for the motion of the low during the aircraft sampling. The low was about 400 km across with 12 mb of closed isobars. The lowest sea-level pressure recorded was 979 mb at the point of calm winds. The low was not axisymmetric, rather somewhat elongated on a northeast–southwest axis. Mesoscale fronts in these directions (dashed lines) mark a strong horizontal wind shear, where the flight-level wind barbs show abrupt changes in direction over tens of kilometers. Wind speeds were over 30 m s^{-1} in the western and southern parts of the low. The relative vorticity of the flight-level winds was an S-shape, following the mesoscale fronts, with typical values over $10 \times 10^{-4} \text{ s}^{-1}$ and peak values of twice that. The relative vorticity values were on a par with those observed in synoptic-scale extratropical cyclones. The 300-m temperature field showed a secluded warm core, 2–3 °C warmer than surrounding temperatures, and of a similar elongated shape to the pressure field. The low was relatively shallow with only a weak circulation signature in the midtroposphere.

The low's vertical structure is illustrated in a cross section of potential temperature contours and storm-relative transverse velocity vectors in [Figure 5](#). The cross section is oriented approximately north–south at about 4° W. To the north (left) was a 1-km deep near-neutral marine boundary layer capped by a stable layer (bold lines). To the south (right) the marine

boundary layer was about 2-km deep. In the center of the cross section was a warm front, delineated by a strong horizontal gradient in potential temperature. There was low-level (storm-relative) convergence into the frontal region and vigorous ascent – the upward vertical velocities reaching 1 m s^{-1} . There was weaker descent on the warmer side of the front, completing a cross-frontal circulation. The warm-core structure of the system is depicted by the downwelling potential temperature contours above the frontal region. Cross sections of equivalent potential temperature showed a decrease in height up to about 2 km in the convergence region: in other words, the frontal region was unstable to moist convection. Convective precipitation was observed in this region, as radar reflectivity echoes, in a shallow layer below 3 km. Surface turbulent heat fluxes were calculated to be extremely high, up to 500 W m^{-2} for both surface sensible and latent heat fluxes. The total surface turbulent heat fluxes in this case were on a par with those observed in tropical cyclones (hurricanes). In some respects, for example, the fronts and the pressure distribution, this polar low was similar to archetypal synoptic-scale extratropical cyclones (see *Synoptic Meteorology: Extratropical Cyclones*). However, the large surface turbulent heat fluxes and clear eye ([Figure 3](#)) are reminiscent of tropical cyclones (see *Tropical Cyclones and Hurricanes: Hurricanes: Observation*).

The observations summarized above paint a vivid picture of a small-scale vortex, with remarkably strong winds, well-defined mesoscale fronts, and a number of convectively unstable areas where precipitation was falling. This structure is typical of many intense baroclinic-convective types of polar low. However, other case studies have shown more axisymmetric systems, with less well-defined fronts than illustrated above (e.g., [Figure 1](#)); or systems without areas of conditional instability, where convection is less important; or systems where upper-level forcing is vital. Clearly there is

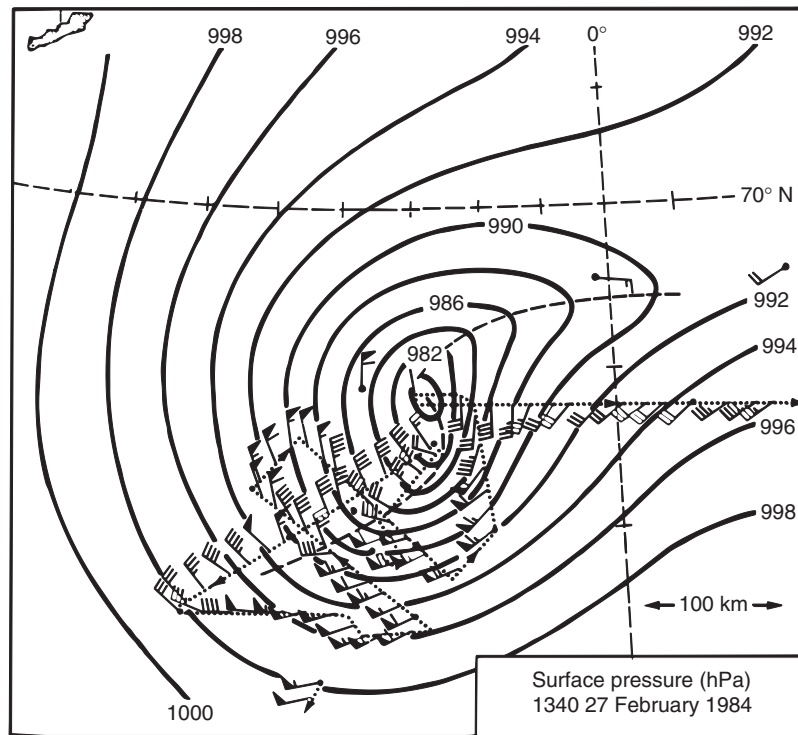


Figure 4 A surface pressure analysis for 13:40 27 February 1984. Overlaid is the track of the instrumented aircraft and the observed flight-level winds. Each full-tick on the wind barbs is 5 m s^{-1} . S-shaped mesoscale fronts curving to the northeast and the southwest are depicted as dashed lines. Adapted from Shapiro, M.A., Fedor, L.S., Hampel, T., 1987. Research aircraft measurements of a polar low over the Norwegian Sea. Tellus 39a, 272–306 with permission.

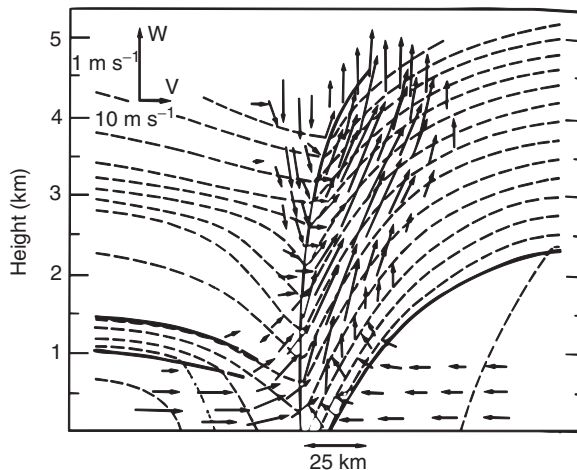


Figure 5 A cross section through the polar low of 27 February 1984, running approximately north-south at about 4° W . The dashed lines show potential temperature (every 1 K). The solid lines delineate the capping of the near-neutral marine boundary layer and the frontal boundary. The vectors show the storm-relative transverse circulation. Adapted from Shapiro, M.A., Fedor, L.S., Hampel, T., 1987. Research aircraft measurements of a polar low over the Norwegian Sea. Tellus 39a, 272–306 with permission.

a spectrum of characteristics within the polar low category of systems. Of course this amount of structural detail is not routinely available, instead in recent years we have come to rely on satellite remote sensing for much information.

Satellite Remote Sensing Data

The most vivid satellite observations of polar lows, and polar mesoscale cyclones, are the high-resolution infrared and visible images such as those in Figures 1–3. These images are simply pictures taken from space at certain wavelengths in the infrared or visible parts of the spectrum. The highest resolution (250–1000 m) images are from sensors on board polar orbiting satellites, for example, the Advanced Very High Resolution Radiometer (AVHRR) or the Moderate Resolution Imaging Spectroradiometer. These provide a fantastically detailed picture of cloud structure and height, often hinting at streamlines through the shape of the clouds. Also flown on many polar orbiting satellites are vertical sounders, which combine a variety of infrared and microwave channels to obtain atmospheric emission temperatures from several heights and thus temperature or humidity profiles for an atmospheric column. Such data have been used to obtain the synoptic backdrop to polar low development and, for example, to corroborate the warm-cored nature of the polar low described above. One limitation of these sounders is their poor functionality in areas of very thick clouds, where the humidity retrievals in particular are affected. This problem is largely bypassed when a more comprehensive passive microwave sounder is used, such as the

Special Sensor Microwave/Imager (SSM/I). This instrument can provide contemporaneous fields of precipitation, cloud liquid water content, and surface wind speed, at a resolution of around 100 km, although again heavy rain can interfere with the sensors. Observations from the SSM/I have helped to determine whether some mesoscale vortices are convective, or not, as convective precipitation gives a strong backscatter at certain wavelengths. Surface wind speed can also be measured by radar altimeters (such as Geosat), again through the influence of the winds on the sea state, and thus the radar backscatter. But due to their narrow swaths these sensors only provide cross sections through weather systems. To obtain both wind speed and wind direction from space, one has to revert to scatterometers (such as QuikScat or ASCAT): these are able to measure capillary waves on the ocean's surface and so infer wind speed and wind direction. They have a resolution of typically 25 km, but a relatively narrow swath or gaps within the swath. In addition, there is an ambiguity problem in determining the wind direction. To rectify this, external information such as wind data from numerical weather prediction models must be supplied. Surface wind data from scatterometers are at a suitable resolution for polar mesoscale cyclone investigations and have assisted in a number of case studies and climatologies. It has been shown that around 75% of cloud vortices (seen in AVHRR imagery) have a distinct surface circulation signature in scatterometer winds.

A large amount of operational satellite remote sensing data are now incorporated, via the data assimilation processes, directly into operational numerical weather prediction models and are thus implicit in meteorological analyses or reanalyses products. This includes vertical sounder temperatures, pressures, or thicknesses (etc.), passive microwave radiances, cloud-track winds, and scatterometer winds. Depending on the resolution and quality of the data assimilation and modeling system, this has resulted in many more polar mesoscale cyclones being represented in numerical weather prediction analyses, reanalyses, and forecasts.

Dynamical Theories of Polar Low Development

A number of dynamical mechanisms have been proposed for the initiation and growth of polar lows and polar mesoscale cyclones. The two most widely accepted theories being that polar lows are fundamentally baroclinic disturbances or fundamentally convective disturbances. A wealth of observational and numerical modeling studies would now suggest there is a continuous spectrum of polar mesoscale cyclones that spans these, and other, development mechanisms.

Baroclinic Dynamics

The classic baroclinic instability mechanism, as first developed in the 1940s and the 1950s, plays a primary role in the development of synoptic-scale extratropical cyclones (see **Dynamical Meteorology: Baroclinic Instability**; **Synoptic Meteorology: Extratropical Cyclones**). Essentially a continuously stratified fluid is unstable to small amplitude perturbations, if the basic state encompasses a reversal in sign of the potential vorticity (PV) gradient. The simplest archetype of

baroclinic instability is that of a zonally symmetric atmosphere, with a rigid lid of height H , a meridional potential temperature gradient Δ (equivalent to a zonal vertical shear – assuming thermal wind balance), a Coriolis force f , and a static stability N . Here the PV gradient reversal is at the boundaries. The growth rate is a linear function of $f\Delta/N$, and the scale of the unstable waves is a linear function of NH/f . For typical midlatitude values this yields e-folding times of about a day and fastest growing modes on the scale of several thousand kilometers. This is an order of magnitude larger than that of observed polar mesoscale cyclones; so to use this theory to explain mesoscale cyclones, it requires a scale reduction through a modification of the basic-state atmosphere. For a start, at high latitudes f is larger, but more importantly both H and N can be much smaller than is typical at midlatitudes. For example, during cold-air outbreaks the atmosphere can be represented as a near-neutral marine boundary layer capped by a strongly stable free atmosphere (e.g., **Figure 5**). In this case, if the rigid lid is taken as the strongly stable capping layer this means a much reduced H , and a reduced N in the near-neutral layer, leading to fastest growing modes on about the scale of observed polar lows (**Figures 3 and 4**). A different basic state, also leading to a scale reduction, is that of a reverse shear flow; so called because the baroclinic waves propagate in the opposite direction to the thermal wind, i.e., the shallow extent of the waves means their steering level is within the reverse flow. Here weak static stabilities and strong vertical shears can combine to yield unstable modes of the order 500 km.

A different paradigm of baroclinic instability is the initial value problem: where an upper-level precursor, such as an upper-level PV anomaly, induces growth at low levels. This scenario can also be envisaged through cyclonic vorticity advection and omega equation arguments, but is perhaps more transparent within an isentropic PV framework (see **Dynamical Meteorology: Potential Vorticity**). To achieve rapid growth, an upper-level PV anomaly must act in synergy with a low-level baroclinic zone, i.e., a potential temperature gradient (see **Figure 6**). The induced circulation from the upper-level PV anomaly deforms the low-level temperature field into a wave, which then induces its own cyclonic circulation, as potential temperature anomalies at a boundary are equivalent to PV anomalies. The upper- and lower-level anomalies reinforce each other's circulation and thus mutual growth occurs. In this scenario, the scale of the disturbance is again a function of NH/f . The presence of an upper-level PV anomaly is often associated with the intrusion of the stratosphere down into the troposphere, i.e., a low dynamical tropopause. In this case, the scale height H is reduced and (as before) when this is combined with low static stabilities, growth on the scale of polar lows is induced. This type of PV thinking has helped to explain a number of observed and modeled polar lows. It has also enabled another form of satellite remotely sensed data to be used, as stratospheric intrusions (i.e., PV anomalies) can be detected from their large total column ozone amounts.

Baroclinic instability mechanisms are significantly modified by the presence of moisture: latent heat is released through condensation of water vapor during ascent, and this additional heating reinforces the baroclinicity of the system and thus enhances growth. The most intense polar mesoscale cyclones

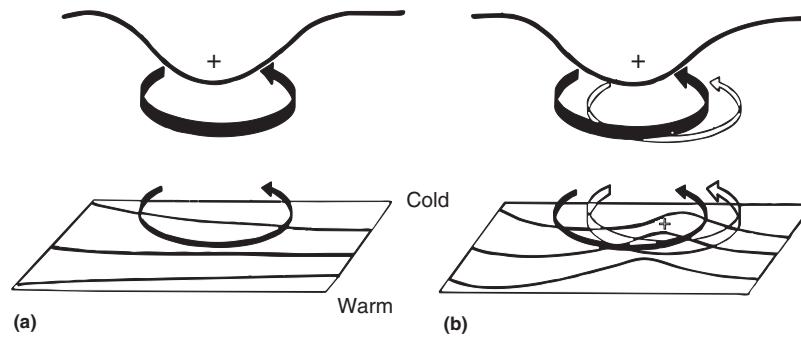


Figure 6 The interaction of an upper-level potential vorticity (PV) anomaly (+ sign) with a low-level baroclinic zone. The cyclonic circulation induced by the upper-level PV anomaly is illustrated through the solid curves. Panel (a) shows the initial state. Panel (b) shows the advection of the low-level potential temperature field into a warm anomaly. At the surface this is equivalent to a positive low-level PV anomaly (open + sign) and thus induces a cyclonic circulation illustrated through the open curves. The interaction is self-reinforcing and leads to rapid cyclogenesis. Adapted from Hoskins, B.J., McIntyre, M.E., Robertson, A.W., 1985. On the use and significance of isentropic potential vorticity maps. *Quarterly Journal of the Royal Meteorological Society* 111, 877–946 with permission.

tend to form as polar lows in maritime environments, where there is a ready supply of moisture, which suggests that any baroclinic dynamics are likely to be enhanced by latent heating. Indeed a number of numerical modeling case studies have found that without moist processes the modeled polar low is unable to grow to the strength observed.

Convective Dynamics

The axisymmetric cloud and the presence of a clear central eye in some polar lows (e.g., [Figure 1](#)) have suggested that these lows may be akin to small tropical cyclones. This idea has been pursued through the modification of some tropical cyclone development theories to polar environments. One such mechanism is conditional instability of the second kind (CISK – see **Dynamical Meteorology: Wave-CISK**). This mechanism involves the organization of cumulonimbus convection into a coherent weather system. An initial disturbance causes low-level convergence and ascent, which, in a conditionally unstable atmosphere, gives rise to latent heat release. This generates cyclonic relative vorticity that forces further low-level convergence and, by continuity, high-level divergence. The low-level convergence provides a continued source of moisture for further latent heat release and thus a positive feedback is established. The scale of CISK-driven disturbances is determined by the vertical distribution of diabatic heating during development. The growth rate is directly dependent upon the convective available potential energy (CAPE) of the atmosphere. In the polar regions, during cold-air outbreaks, extremely deep conditionally unstable layers can develop, thus providing enough CAPE for polar low development. However, it has also been argued, and observed from soundings, that the atmosphere does not sustain the required amounts of CAPE for long enough to allow the CISK mechanism to act as described.

An alternative convective theory is that of wind-induced surface heat exchange (WISHE) – also known as air–sea interaction instability – see **Tropical Cyclones and Hurricanes: Hurricanes: Observation**. This proposes that surface sensible and latent heat fluxes are responsible for the growth of tropical cyclones and polar lows. Moist convection mixes this additional heat and moisture through the troposphere on the scale of the growing vortex. Latent heat release within the convective

clouds contributes to vortex growth, and as the surface turbulent heat fluxes are proportional to the surface winds, a positive feedback is established. The theory allows an estimate of the central pressure drop within a mature system from Carnot energy-cycle arguments and known environmental conditions; for example, the observed pressure drop of the polar low in [Figure 1](#) is consistent with this theory.

In some sense the above two mechanisms are variations of convective closure rather than being fundamentally different. CISK relies on low-level convergence from friction, where WISHE does not, but on the other hand, it could be argued that CISK implicitly includes surface sensible and latent heat fluxes as part of the dynamics. Recent idealized numerical modeling studies have found that polar low growth is more sensitive to the parameterization of surface heat fluxes than to the parameterization of surface friction, which provides some evidence for the WISHE mechanism, but in the real world the two mechanisms are more difficult to distinguish. A common requirement for the convective theories is an initial disturbance of the basic state about which the convection can become organized. This begs the question, what causes the initial disturbance? In many observational and numerical studies, it would appear that some sort of baroclinic forcing causes the initial disturbance. As described in the previous section, this may be a linear instability of a baroclinic atmosphere or an upper-level precursor inducing flow at low levels. The initial polar low can then evolve into a convectively driven phenomenon, if the conditions are right. In other cases perhaps a topographic forcing, or a barotropic forcing, causes the initial disturbance. What is clear, is that moist convection plays a central role in the development of many polar lows in a way that it does not in other extratropical cyclones.

It has been found that long-wave radiative cooling generally increases the growth rate and maximum intensity of polar lows, due to enhanced cooling of the cyclones environment relative to its warm core.

Numerical Modeling and Forecasting of Polar Lows

In a research context, numerous case studies have shown that state-of-the-art numerical models are able to simulate polar

lows, and polar mesoscale cyclones, of all varieties. In fact, numerical modeling studies have been instrumental in elucidating the different roles of the dynamical mechanisms discussed above. To capture the growth and mesoscale structure to a reasonable accuracy, it typically requires the parameterization of microphysical processes (including latent heat release, cloud microphysics, and multiphase precipitation); radiative and turbulent heat fluxes (including surface turbulent heat fluxes); and moist convection. If any of these physical processes is switched off, the simulated polar lows tend to be weak or nonexistent. It is also essential that the numerical model's grid resolution is suitable for the scale of the cyclone. Typically horizontal grid squares of 20–50 km are required for a cyclone of a few hundred kilometers and higher resolutions (less than 10 km), if details such as mesoscale fronts are to be simulated. On the other hand, it is also necessary to simulate the evolving synoptic-scale environment as this affects both polar low development and movement. In general, the synoptic-scale flow plays a primary role in steering mesoscale cyclones in a passive manner. However, if two or more polar lows are active within close proximity, it is possible that they will interact and this also affects their movement. Hence forecasting polar low tracks requires both relatively high resolution and large domain sizes.

In an operational context, the simulation and forecasting of polar lows is still problematic; indeed there are examples of misforecasts leading to tragedies as recently as the 2000s. Some state-of-the-art global models, and certainly most regional models, now have sufficient grid resolution for simulation. It is their correct initialization that is the problem. Data assimilation grids tend to be coarser and the observations filtered, to remove what are assumed to be small-scale anomalies, thus many polar lows are not in the initial conditions or the analysis increments when they should be. Furthermore, their relatively rapid development, often from a somewhat random location through convection, means the forecast model may well develop a polar low at the correct time but in the wrong location. There is no doubt that modern data assimilation and forecasting systems are capable of accurate forecasts of polar lows, but they are at the limit of what is possible and misforecasts of location or timing are not infrequent. Regional ensemble prediction systems are now being utilised (e.g. for the Norwegian Sea) which will allow best-track forecasts and provide some statistical forecast information.

Polar Lows and Climate

Their relatively small scale and short lifetime makes polar lows subgrid scale for the vast majority of climate models. Yet their location in the polar and subpolar seas, often close to the sea ice edge, means they are located in a sensitive and rapidly changing part of the climate system. In recent years attention has been focused on the role polar lows play in the climate system and how this may alter in a changing climate.

Case studies of polar lows have often associated them with large air–sea turbulent heat fluxes (as described above) and, given their location in the subpolar seas, there is the potential for this to have a significant impact on the ocean. Recent studies have quantified this impact through ocean modeling experiments using atmospheric forcing fields with and without polar lows. The results demonstrate a dramatic local impact on the

ocean when polar lows are included – a regional surface-layer cooling, an increase in mixed layer depth, a spin-up of the Nordic Seas gyre, and an increase in deep-water formation. The broader-scale impact of polar lows on the ocean is not yet clear.

Complementing earlier satellite-based climatologies, a number of techniques have been developed to quantify the numbers of polar mesoscale cyclones in atmospheric model-based data sets. These include tuning cyclone-tracking algorithms to focus on subsynoptic-scale systems; using proxies such as marine cold-air outbreaks; using a lower–upper temperature threshold to discriminate vortices in cold-air outbreaks; or using a full dynamical downscaling approach (i.e., running a higher resolution nested regional model driven by a global climate model). These approaches have largely corroborated the satellite-based studies, replicating the geographical and seasonal patterns discussed earlier. In addition they have also allowed projections of polar low activity in the future. These projections suggest a poleward migration of polar low activity, following the projected retreat of the sea ice edge, thus a decrease in polar low numbers in the northern midlatitudes and an increase in the sub-Arctic. At this stage, the details of future polar low activity are relatively uncertain.

See also: **Air Sea Interactions:** Momentum, Heat, and Vapor Fluxes. **Dynamical Meteorology:** Baroclinic Instability; Potential Vorticity; Wave-CISK. **Numerical Models:** Mesoscale Atmospheric Modeling. **Synoptic Meteorology:** Extratropical Cyclones. **Tropical Cyclones and Hurricanes:** Hurricane Dynamics; Hurricane Predictability; Hurricanes: Observation; Tropical Cyclogenesis.

Further Reading

- Condon, A., Renfrew, I.A., 2013. The impact of polar mesoscale storms on northeast Atlantic Ocean circulation, *Nature Geoscience* 6, 34–37, <http://dx.doi.org/10.1038/ngeo1661>.
- Craig, G.C., Gray, S.L., 1996. CISK or WISHE as the mechanism for tropical cyclone intensification. *Journal of the Atmospheric Sciences* 53, 3528–3540.
- Emanuel, K.A., Rotunno, R., 1989. Polar lows as arctic hurricanes. *Tellus* 41A, 1–17.
- Gronas, S., Kvamsto, N.G., 1995. Numerical simulations of the synoptic conditions and development of arctic outbreak polar lows. *Tellus* 47A, 797–814.
- Hoskins, B.J., McIntyre, M.E., Robertson, A.W., 1985. On the use and significance of isentropic potential vorticity maps. *Quarterly Journal of the Royal Meteorological Society* 111, 877–946.
- Kolstad, E.W., Bracegirdle, T.J., 2008. Marine cold-air outbreaks in the future: an assessment of IPCC AR4 model results for the Northern Hemisphere. *Climate Dynamics* 30, 871–885.
- Kristiansen, J., Sørland, S.L., Iversen, T., Bjørge, D., Koltzow, M.O., 2011. High-resolution ensemble prediction of a polar low development. *Tellus* A 63, 585–604.
- Rasmussen, E., 1979. The polar low as an extratropical CISK disturbance. *Quarterly Journal of the Royal Meteorological Society* 105, 531–549.
- Rasmussen, E., Turner, J., 2003. *Polar Lows: Mesoscale Weather Systems at High Latitudes*. Cambridge University Press.
- Shapiro, M.A., Fedor, L.S., Hampel, T., 1987. Research aircraft measurements of a polar low over the Norwegian Sea. *Tellus* 39a, 272–306.
- Turner, J., Lachlan-Cope, T., Thomas, J., 1993. A comparison of Arctic and Antarctic mesoscale vortices. *Journal of Geophysical Research* 98, 13019–13034.
- Zahn, M., von Storch, H., 2010. Decreased frequency of North Atlantic polar lows associated with future climate warming. *Nature* 467, 309–312.

Relevant Websites

<http://polarlow.met.no/index.shtml>
<http://polarlows.wordpress.com/>

Thermal Low

RH Johnson, Colorado State University, Fort Collins, CO, USA

© 2003 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 5, pp 2269–2273, © 2003, Elsevier Ltd.

Introduction

A thermal low (sometimes referred to as a heat low) is a low-pressure area resulting from high temperatures in the lower troposphere caused by a localized area of intense heating at the Earth's surface. Thermal lows occur typically during the summer over subtropical continental areas and are most intense in the desert regions of the world. These regions are characterized by clear skies and a lack of vegetation. Consequently, there is a large diurnal cycle of surface heating, which in turn creates a pronounced diurnal cycle in the intensity of thermal lows, with a maximum intensity (i.e., minimum surface pressure) during the afternoon. Because thermal lows are linked directly to surface properties, they are nonmigratory in nature. Moreover, they do not exhibit any frontal characteristics, nor are clouds or precipitation associated with them.

Since thermal lows arise from surface heating, the maximum amplitudes of temperature and circulation anomalies associated with them are confined to the lower troposphere (i.e., below 5 km or 500 hPa). Most of the desert areas of the world that exhibit strong thermal lows are surrounded by bodies of water. As a result, horizontal heating gradients that develop generate sea and land breezes that influence the thermal-low circulations. In regions away from the equator, the low-level inflow associated with the daytime sea breeze produces, through the action of the Earth's rotation (the Coriolis force), a cyclonic vorticity anomaly in the lower troposphere. The converging air at low levels in the thermal low also produces upward motion in the lower troposphere, but sinking motion occurs aloft. The cyclonic circulation often persists into the nighttime hours above the surface following the development of a shallow, nocturnal temperature inversion. By the morning, subsidence extends all the way down to the surface.

Because of the clear-sky conditions and high surface reflectance (albedo) of the desertlike areas where thermal lows persist, these regions have often been regarded as large-scale radiative energy sinks. However, in some thermal low regions dust is raised from the surface by the intense heating which results in short-wave absorption that makes at least a portion of the thermal low region a radiative energy source.

Geographical Distribution

The strongest thermal lows are located over the great deserts of the world, e.g., the Sahara, Arabian, Kalahari, Australian Great Western Desert, and Mojave/Sonoran Deserts. In these regions the thermal lows are so strong that they appear as closed lows or troughs in the mean-sea-level pressure maps of the summer hemispheres (Figure 1). For example, note the low-pressure troughs over the African Sahara, the Indian subcontinent, and the Southwest US in July and over northern Australia in January. The trough over southern Asia near India in July is

closely associated with cross-equatorial flow over the western Indian Ocean and the Asian summer monsoon.

The locations of thermal lows also correspond to regions of intense surface heating, as can be seen from a map of the global distribution of surface sensible heat flux (Figure 2). The lack of cloudiness, vegetation, and surface moisture in these regions accounts for the large values of surface sensible heat flux. Maximum values upwards of $60\text{--}70\text{ W m}^{-2}$ can be seen in the regions of the major deserts. The close relationship between the thermal lows and surface sensible heat flux patterns indicates a close physical linkage between these two features. Specifically, intense, localized heating warms the lower atmosphere and hydrostatically reduces the surface pressure in that region.

The wind vectors illustrated in Figure 1 show generally confluent surface flow into the thermal low-pressure areas. This flow arises from strong horizontal gradients in the surface sensible heat flux between the desert areas and the surrounding oceans (sensible heat fluxes over the ocean are typically 10 W m^{-2} or less), which drives sea breeze circulations during the daytime hours. At night, a land breeze or offshore surface flow develops, but the intense heating causes the sea breeze to dominate over the land breeze circulation.

Vertical Structure and Energetics

Thermal lows generally develop over arid lands or deserts in the summertime. Figure 3 depicts the thermodynamic structure of the troposphere within a thermal low. The case shown represents a synthesis of data from research aircraft flights over the Arabian desert, although it is typical of other desert regions. To illustrate mixing processes in the atmosphere, vertical profiles of two conserved thermodynamic quantities for dry atmospheric motions, the specific humidity q and the potential temperature θ , are plotted in Figure 3. From the surface to about 5 km or 550 hPa, θ is approximately constant and q decreases with height. This structure is characteristic of deep, continental atmospheric boundary layers containing vigorous, dry convective plumes or thermals. The sharp vertical gradients of q and θ near 5 km represent a transition zone between turbulent air in the atmospheric mixed layer below and laminar flow in the free atmosphere above. The vertical gradients in q and θ in the upper part of the mixed layer are a result of entrainment of drier (lower q) and potentially warmer (higher- θ) air from above into the mixed layer. At night, radiational cooling acts to decrease θ in the lowest kilometer, creating a nocturnal inversion.

Thermal lows typically exist in regions of large-scale subsidence. However, owing to the development of sea breezes during the daytime, the vertical motion in the lowest levels can become upward during the day. This behavior can be seen in Figure 4, which contains profiles of vertical motion at day and night over the Arabian desert. Above 800 hPa there is sinking

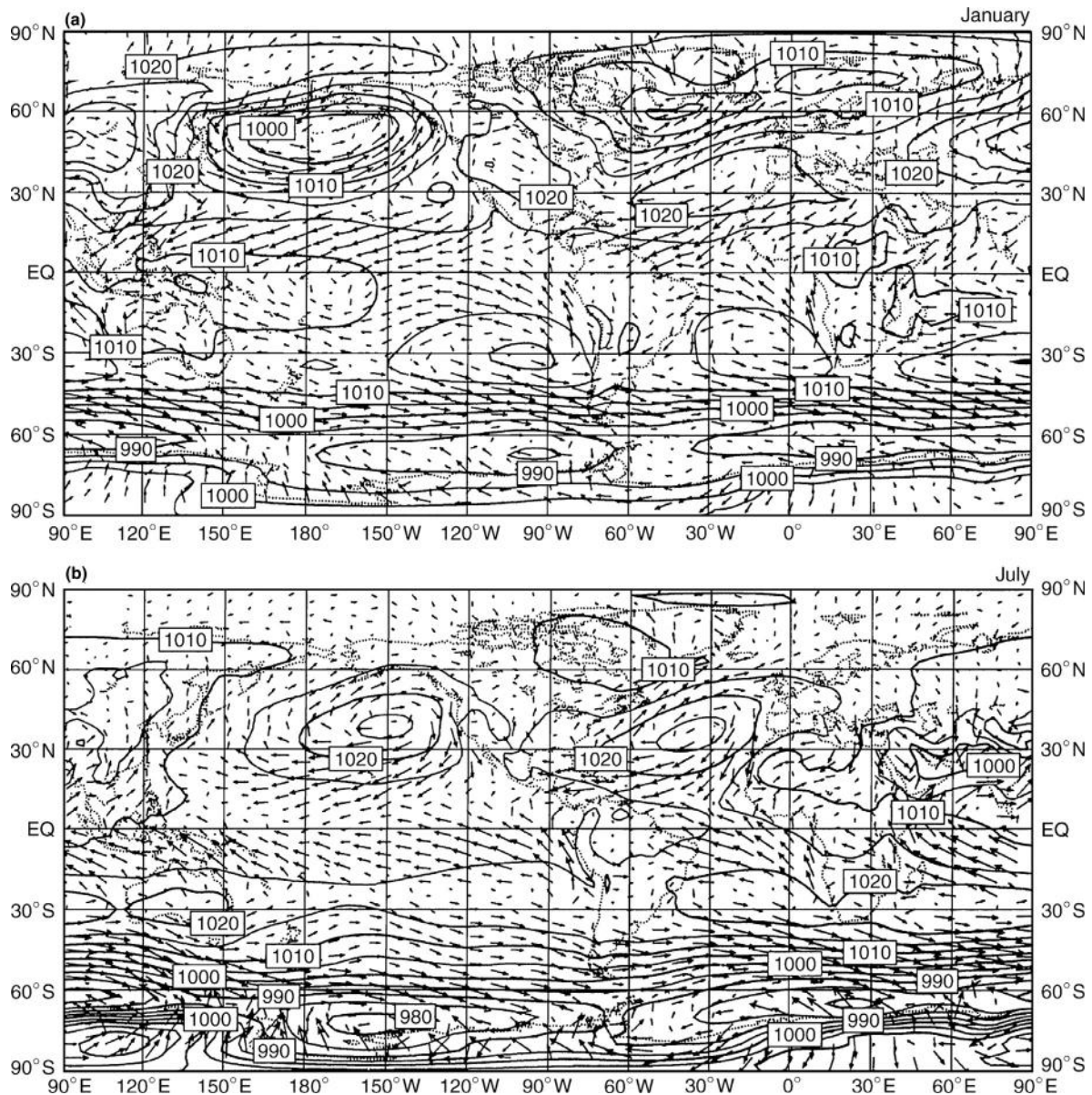


Figure 1 Maps of mean sea-level pressure for January and July. Wind vectors for the 1000 hPa level are superimposed. Data are 1980–87 analyses from a forecast model. The contour interval is 5 hPa and the largest vector represents a wind speed of 12 m s^{-1} . Reproduced with permission from Hartmann, D. L., 1994. *Global Physical Climatology*. Academic Press, San Diego, CA.

motion during both day and nighttime hours, although it is weaker during the day owing to short-wave radiative heating of the surface and atmosphere.

The vertical profiles of vertical motion in [Figure 4](#) imply horizontal divergence in the lower to middle troposphere. This divergence is consistent with heating in the lower troposphere that lifts isobaric surfaces above the heat source and produces a midlevel high and coincident outflow of air. This outflow in turn reduces the surface pressure and assists in the inflow of air into the surface low. Part of the outflowing air aloft during the daytime represents the return flow of the sea breeze circulation.

The vertical structure of thermal lows can also be considered from an energy budget perspective. Field studies of thermal

lows show that the raising of dust by daytime heating makes their energetics rather complex. To illustrate this complexity, the vertical structure of the energy balance of the Arabian heat low is shown in [Figure 5](#). This heat low can be characterized as a three-layer system. The lower atmosphere forms a deep mixed layer from the surface to 550 hPa. An upper layer from 550 hPa to the upper boundary is a region where radiative cooling is approximately balanced by adiabatic heating (subsidence warming). A middle layer from 550 to 800 hPa undergoes both sensible and radiative heating. The sensible heating arises from the convergence of eddy heat flux owing to mixed-layer turbulence. The radiative heating results from enhanced daytime shortwave absorption due to a substantial aerosol or

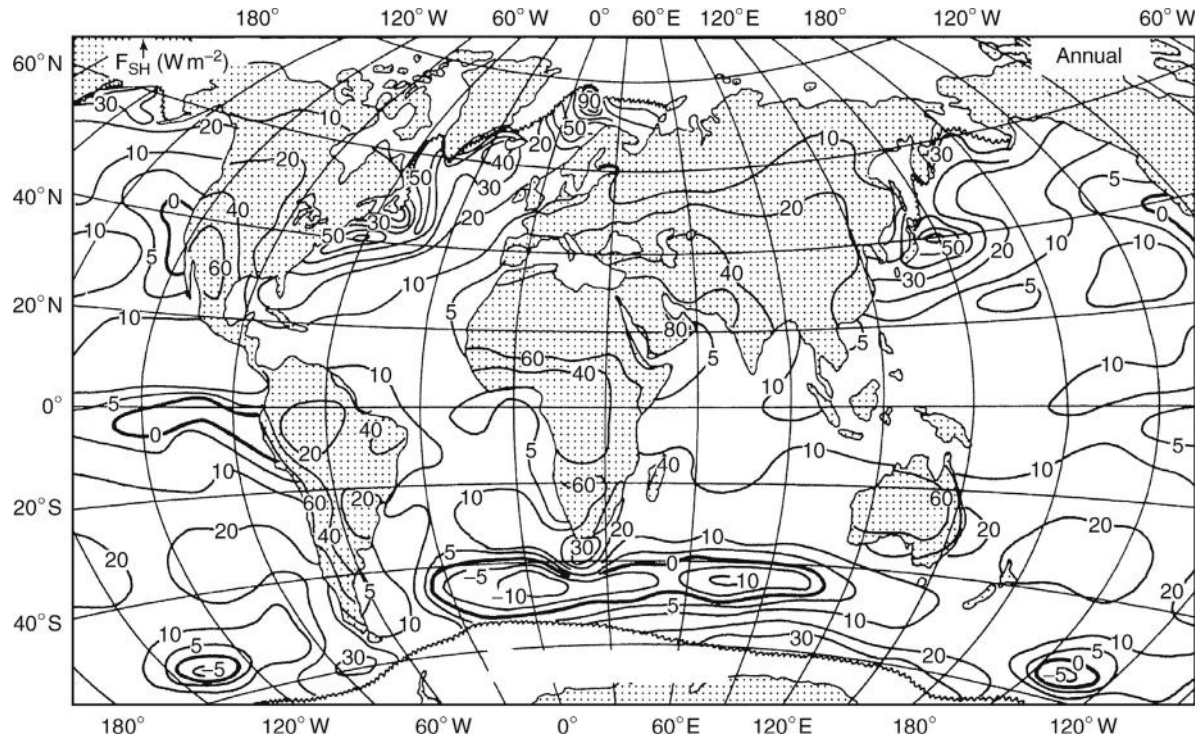


Figure 2 Global distribution of the sensible heat flux from the earth's surface into the atmosphere in W m^{-2} for annual-mean conditions. Reproduced with permission from Peixoto, J. P., Oort, A. H., 1992. *Physics of Climate*. American Institute of Physics, New York.

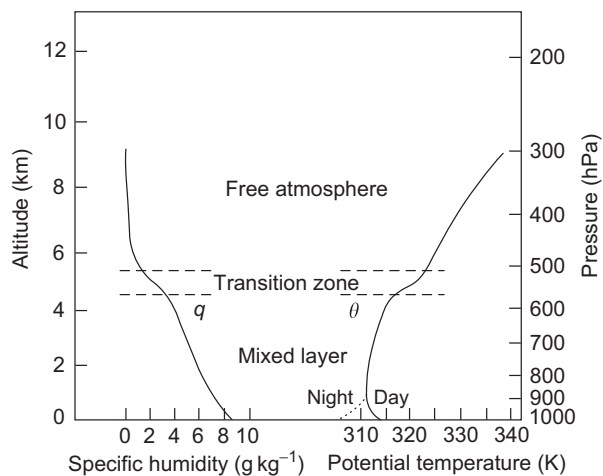


Figure 3 Typical vertical profiles of specific humidity q and potential temperature θ within a thermal low extending from the surface through the mixed layer, the transition zone, and into the free atmosphere. Adapted with permission from Smith, E. A., 1986. The structure of the Arabian heat low part II: Bulk tropospheric heat budget and implications. *Monthly Weather Review* 114: 1084–1102.

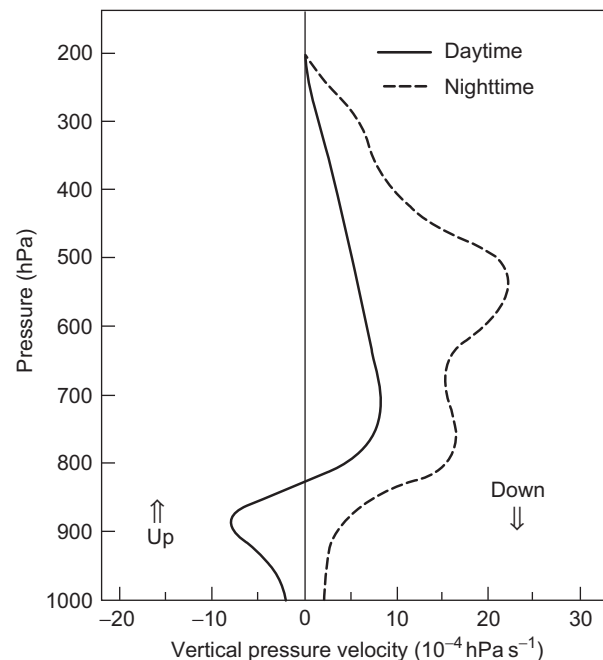


Figure 4 Daytime and nighttime vertical motion profiles (vertical p -velocity) over the Arabian heat low. The positive values indicate downward motion. Adapted with permission from Smith, E. A., 1986. The structure of the Arabian heat low, part II: Bulk tropospheric heat budget and implications. *Monthly Weather Review* 114: 1084–1102.

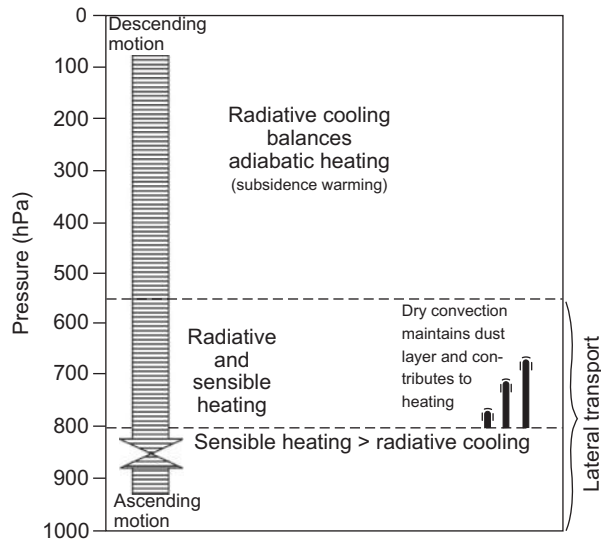


Figure 5 Conceptual three-layer structure of the daytime Arabian heat low. Adapted with permission from Smith, E. A., 1986. The structure of the Arabian heat low, part II: Bulk tropospheric heat budget and implications. *Monthly Weather Review* 114: 1084–1102.

dust loading. The dust is generated locally by daytime vigorous boundary layer thermals and associated gusty surface winds. In the lower layer (surface to 850 hPa) the convergence of eddy heat fluxes dominates radiative cooling, resulting in a net lower-tropospheric warming.

The processes illustrated in Figure 5 indicate that the middle and lower troposphere undergoes a net energy gain due to diabatic processes. The only mechanism to balance this energy gain is a lateral transport of energy out of the region. For the Arabian heat low, the main lateral transport is over the Arabian Sea, where the warm air serves to maintain the West Arabian Sea inversion. This inversion acts to trap water vapor in the lower troposphere as it is carried toward India in the strong south-west summer monsoon flow. This is one mechanism by which thermal lows can have an important influence on surrounding weather patterns.

Dynamics and the Diurnal Cycle

Direct observations indicate that a low-level cyclonic circulation develops typically within thermal lows with a maximum amplitude during the daytime hours. Observations are rather sparse, however, and details of the diurnal cycle of this circulation have not been well documented. Therefore, numerical modeling studies have been used to provide further insight into the dynamics of thermal lows.

Numerical simulations of thermal lows over land surrounded by an ocean show a pronounced diurnal cycle in the circulation patterns. In fact, the circulation cannot be understood without consideration of the diurnal cycle. Simulations show that while the thermal low has a surface pressure minimum in the late afternoon following strong solar heating

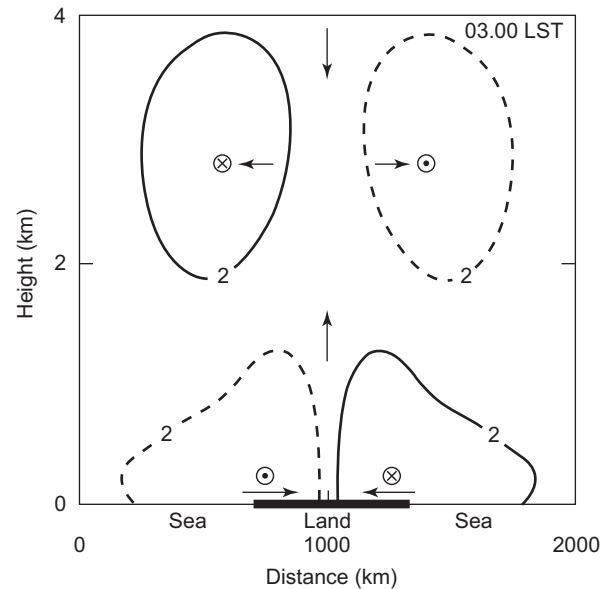


Figure 6 Vertical west-east cross-section at 03.00 LST through a thermal low that has formed over a 600×600 km square land area surrounded by ocean in the Northern Hemisphere. Contours indicate meridional (N–S) wind (m s^{-1}). $\otimes$ indicates the center of the maximum southerly flow, $\odot$ the center of the maximum northerly flow. The horizontal arrows denote the centers of maximum zonal (E–W) flow and the vertical arrows indicate the centers of maximum vertical motion. The heavy line indicates the land area. Adapted with permission from Rácz, Z., Smith, R. K., 1999. The dynamics of heat lows. *Quarterly Journal of the Royal Meteorological Society* 125: 225–252.

of the land, the relative vorticity is strongest in the early morning hours as a result of a prolonged period of low-level convergence. Thus the thermal low is not approximately in quasi-geostrophic balance.

A depiction of the zonal, meridional, and vertical velocity components at 03.00 LST for an idealized simulation of a thermal low over a $600 \text{ km} \times 600 \text{ km}$ land area surrounded by ocean at 20°N is shown in Figure 6. Low-level convergence, initially developed in connection with a daytime sea breeze, is seen in the early morning hours to evolve into a nocturnal low-level jet. This jet develops as a result of strong nocturnal surface cooling over land. The earth's rotation acting on this circulation generates a cyclonic circulation at low levels, evident in the meridional wind field. The sea breeze return flow aloft is horizontally divergent and it generates an anticyclonic circulation at upper levels, again evident in the meridional wind field. Rising motion occurs at low levels and sinking motion aloft, consistent with the daytime pattern illustrated in Figure 4. The vertical motion over the land eventually reverts to sinking at all levels by 06.00 LST, as in the nighttime curve in Figure 4.

The dynamics of thermal lows can also be considered from a potential vorticity perspective. Observations of the summertime thermal low over the Iberian Peninsula (Spain) show that as an unstable lapse rate forms in the afternoon, a negative potential vorticity anomaly develops over the peninsula. This negative anomaly exists within a large-scale environment of positive

potential vorticity. At night this negative anomaly disappears as stable air develops near the surface due to nocturnal cooling. It has been postulated that the dome of negative potential vorticity associated with the thermal low over the Iberian Peninsula can act to inhibit the development of nearby Algerian lows by increasing the effective interaction distance between upper-level and low-level potential vorticity anomalies.

See also: **Aerosols:** Dust; Role in Radiative Transfer. **Boundary Layer (Atmospheric) and Air Pollution:** Diurnal Cycle; Surface Layer. **Climate and Climate Change:** Energy Balance Climate Models. **Dynamical Meteorology:** Coriolis Force. **Hydrology, Floods and Droughts:** Deserts and Desertification. **Mountain Meteorology:** Land and Sea Breezes.

Further Reading

- Griffiths, J. F., Soliman, K. H., 1972. The northern desert. In: Griffiths, J. F. (Ed.), *World Survey of Climatology. Climates of Africa*, vol. 10. Elsevier, New York, pp. 75–111.
- Hartmann, D. L., 1994. *Global Physical Climatology*. Academic Press, San Diego, CA.
- Peixoto, J. P., Oort, A. H., 1992. *Physics of Climate*. American Institute of Physics, New York.
- Rácz, Z., Smith, R. K., 1999. The dynamics of heat lows. *Quarterly Journal of the Royal Meteorological Society* 125, 225–252.
- Ramage, C. S., 1971. *Monsoon Meteorology*. Academic Press, New York.
- Smith, E. A., 1986. The structure of the Arabian heat low, pt II. Bulk tropospheric heat budget and implications. *Monthly Weather Review* 114, 1084–1102.

Contents

Humidity Variables

Moist (Unsaturated) Air

Saturated Adiabatic Processes

Humidity Variables

JA Curry, Georgia Institute of Technology, Atlanta, GA, USA

© 2003 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 3, pp 930–941, © 2003, Elsevier Ltd.

Introduction

Atmospheric humidity is the amount of water vapor present in the atmosphere. Both the rate of evaporation and the time and place that condensation of cloud water occurs are controlled by humidity. Determination of the atmospheric humidity is important for determining surface evaporation, atmospheric radiative transfer, and certain chemical reactions in the atmosphere.

Vapor Pressure

The partial pressure of water vapor, e , is used as the fundamental measure of water vapor content in the atmosphere. Using Dalton's law of partial pressures, we can use the ideal gas law to write the following expression for water vapor pressure:

$$e = \rho_v R_v T \quad [1]$$

where ρ_v is the water vapor density (often referred to as the absolute humidity), R_v is the specific gas constant for water, and T is the atmospheric temperature. The atmospheric relative humidity is defined as the ratio of the atmospheric vapor pressure to the equilibrium (or saturation) vapor pressure at the temperature of the air, e_s .

To understand what is meant by the saturation vapor pressure, we must first examine the equilibria between phases. Consider a closed container half full of pure water and overlain by dry air. As the water begins to evaporate from the water surface, a small increase in pressure is detected in the air above, resulting from the motion of the water vapor molecules that are added to the air through evaporation. As more and more molecules escape from the water surface in the closed container, the steadily increasing vapor pressure in the air above forces more and more of these molecules to return to the

liquid. Eventually an equilibrium is reached where the number of water vapor molecules returning to the surface balances the number leaving. At that point the air is said to be saturated with water vapor, and the partial pressure of the water vapor is equal to the saturation vapor pressure. The relative humidity is equal to unity (or 100%). If the temperature of the water in the closed container were increased, more water would evaporate before a balance was reached.

A rigorous derivation of the saturation vapor pressure can be accomplished from the second law of thermodynamics. For the liquid–vapor equilibrium, we can write the Clausius–Clapeyron equation as follows,

$$\frac{de_s}{dT} = \frac{L_{lv} e_s}{R_v T^2} \quad [2]$$

where L_{lv} is the latent heat of vaporization (energy per unit mass). At any given temperature, eqn [2] states that there is one and only one pressure at which water vapor is in equilibrium with liquid water, with the saturation vapor pressure increasing approximately exponentially with increasing temperature.

Integration of eqn [2] is made difficult owing to the variation of the latent heat of vaporization with temperature. Additionally, application of the Clausius–Clapeyron equation to determining the saturation vapor pressure in the atmosphere is not strictly valid because of the presence of other gases. Hence empirical values of the saturation vapor pressure are typically used (shown in Table 1).

These empirical values of saturation vapor pressure can be represented by a sixth-order polynomial:

$$e_s = a_1 + \sum_{n=2}^7 a_n (T - T_{tr})^{n-1} \quad [3]$$

where T_{tr} , the so-called 'triple point' of water, is 273.15 K, and the coefficients for the saturation vapor pressure over water and

Table 1 Saturation pressures over pure liquid water and pure ice as a function of temperature

$T(^{\circ}\text{C})$	e_w (hPa)	e_i (hPa)	$T(^{\circ}\text{C})$	e_w (hPa)	e_i (hPa)	$T(^{\circ}\text{C})$	e_w (hPa)	$T(^{\circ}\text{C})$	e_w (hPa)
-50	0.0635	0.0393	-24	0.8826	0.6983	1	6.565	26	33.606
-49	0.0712	0.0445	-23	0.9647	0.7708	2	7.054	27	35.646
-48	0.0797	0.0502	-22	1.0536	0.8501	3	7.574	28	37.793
-47	0.0892	0.0567	-21	1.1498	0.9366	4	8.128	29	40.052
-46	0.0996	0.0639	-20	1.2538	1.032	5	8.718	30	42.427
-45	0.1111	0.0720	-19	1.3661	1.135	6	9.345	31	44.924
-44	0.1230	0.0810	-18	1.4874	1.248	7	10.012	32	47.548
-43	0.1379	0.0910	-17	1.6183	1.371	8	10.720	33	50.303
-42	0.1533	0.1021	-16	1.7594	1.505	9	11.473	34	53.197
-41	0.1704	0.1145	-15	1.9114	1.651	10	12.271	35	56.233
-40	0.1891	0.1283	-14	2.0751	1.810	11	13.118	36	59.418
-39	0.2097	0.1436	-13	2.2512	1.983	12	14.016	37	62.759
-38	0.2322	0.1606	-12	2.4405	2.171	13	14.967	38	66.260
-37	0.2570	0.1794	-11	2.6438	2.375	14	15.975	39	69.930
-36	0.2841	0.2002	-10	2.8622	2.597	15	17.042	40	73.773
-35	0.3138	0.2232	-9	3.0965	2.837	16	18.171	41	77.798
-34	0.3463	0.2487	-8	3.3478	3.097	17	19.365	42	82.011
-33	0.3817	0.2768	-7	3.6171	3.379	18	20.628	43	86.419
-32	0.4204	0.3078	-6	3.9055	3.684	19	21.962	44	91.029
-31	0.4627	0.3420	-5	4.2142	4.014	20	23.371	45	95.850
-30	0.5087	0.3797	-4	4.5444	4.371	21	24.858	46	100.89
-29	0.5588	0.4212	-3	4.8974	4.756	22	26.428	47	106.15
-28	0.6133	0.4668	-2	5.2745	5.173	23	28.083	48	111.65
-27	0.6726	0.5169	-1	5.6772	5.622	24	29.829	49	117.40
-26	0.7369	0.5719	-0	6.1070	6.106	25	31.668	50	123.39
-25	0.8068	0.6322	-	-	-	-	-	-	-

Table 2 Coefficients of the sixth-order polynomial fits to saturation vapor pressure for the temperature range -50° to 50°C for both liquid water and ice

Coefficient	Liquid Water	Ice
a_1	6.111 767 50	6.109 526 65
a_2	0.443 986 062	0.501 948 336
a_3	$0.143\,053\,301 \times 10^{-1}$	$0.186\,288\,989 \times 10^{-1}$
a_4	$0.265\,027\,242 \times 10^{-3}$	$0.403\,488\,906 \times 10^{-3}$
a_5	$0.302\,246\,994 \times 10^{-5}$	$0.539\,797\,852 \times 10^{-5}$
a_6	$0.203\,886\,313 \times 10^{-7}$	$0.420\,713\,632 \times 10^{-7}$
a_7	$0.638\,780\,966 \times 10^{-10}$	$0.147\,271\,071 \times 10^{-9}$

After Flatau, P.J., Walko, R.L., Cotton, W.R., 1992. Polynomial fits to saturation vapor pressure. *Journal of Applied Meteorology* 31: 1507–1513.

over ice are as given in [Table 2](#). Note that, at a given temperature, the saturation vapor pressure over ice is smaller than the saturation vapor pressure over liquid water; this is because the latent heat of sublimation (associated with the phase transition between vapor and ice) is larger than the latent heat of vaporization.

Humidity Variables

Values of the saturation vapor pressure are used in the determination of some of the commonly used atmospheric

humidity variables. A wide variety of different humidity variables are used in atmospheric science. The reasons for this are partly historical and partly related to the different methods by which atmospheric humidity is measured. Given information about the ambient atmospheric temperature and pressure, one humidity variable can be used to determine each of the other humidity variables.

The relative humidity, H , is defined as

$$H = \frac{e}{e_s} \quad [4]$$

and H_i , the relative humidity with respect to ice saturation, is defined as

$$H_i = \frac{e}{e_{si}} \quad [5]$$

where e_{si} is the saturation vapor pressure over ice. The relative humidity is the ratio of the actual partial pressure of water vapor in the air to the saturation vapor pressure, and is a function only of e and T . It is commonly multiplied by 100 and expressed as a percentage. Relative humidity can be changed if moisture is added or removed (which changes e), or if the air temperature changes (which changes e_s). At temperatures below 0°C , it is necessary to specify whether the relative humidity is being evaluated relative to the saturation vapor pressure over liquid water or over ice. [Table 3](#) shows that an atmosphere saturated with respect to liquid water is supersaturated with respect to ice, and that the degree of supersaturation increases with the supercooling.

Table 3 Variation of H_i with T for constant $H = 1$

$T(^{\circ}\text{C})$	H	H_i
0	1.0	1.0
-10	1.0	1.10
-20	1.0	1.22
-30	1.0	1.34
-40	1.0	1.47

The water vapor mass mixing ratio, w_v , is the ratio of the mass of water vapor (m_v) present to the mass of dry air (m_d). It is thus defined, after substituting from the ideal gas law [1], as

$$w_v = \frac{m_v}{m_d} = \frac{\rho_v}{\rho_d} = \varepsilon \frac{e}{p - e} \quad [6]$$

where $\varepsilon = M_v/M_d = 0.622$ is the ratio of the molecular weight of water to the molecular weight of the mixture of dry air gases, ρ_d is the density of dry air, and p is total atmospheric pressure.

A value of the saturation mass mixing ratio, w_s , is given by

$$w_s = \varepsilon \frac{e_s}{p - e_s} \quad [7]$$

Expressions for e and e_s from eqns [6] and [7] can be substituted into eqn [4]. Since $p \gg e$ and $p \ll e_s$,

$$H \approx \frac{w_v}{w_s} \quad [8]$$

is an approximate definition of the relative humidity.

The specific humidity is defined as the ratio of the mass of water vapor to the total mass of atmospheric gases. The water vapor mixing ratio can be related to the specific humidity, q_v , as

$$q_v = \frac{m_v}{m_d + m_v} = \varepsilon \frac{e}{p - (1 - \varepsilon)e} = \frac{w_v}{1 + w_v} \quad [9]$$

Since both w_v and q_v are always smaller than 0.04 in the Earth's atmosphere, $q_v \approx w_v$.

The total mass of water vapor in a column of unit cross-sectional area extending from the surface to the top of the atmosphere is called the precipitable water, W_v :

$$W_v = \int_0^{\infty} \rho_v dz \quad [10]$$

where z is height above the Earth's surface. The term 'precipitable water' is used because, if all the vapor in the column were to be condensed into a pool of liquid at the base of the column, the depth of the pool would be equal to W_v/ρ_l , where ρ_l is the density of liquid water. To obtain a relationship between precipitable water and specific humidity, we can write eqn [10] in terms of pressure by incorporating the hydrostatic equation:

$$W_v = \frac{1}{g} \int_p^{p_0} \frac{\rho_v}{\rho_a} dp = \frac{1}{g} \int_p^{p_0} q_v dp \quad [11]$$

where p_0 is the surface pressure, corresponding to $z = 0$ and ρ_a is the air density.

The dew point temperature, denoted by T_D , is defined as the temperature to which moist air must be cooled, the atmospheric pressure and water vapor mixing ratio remaining constant, such that it becomes just saturated with respect to water. Obviously, at T_D the mixing ratio of the air becomes its saturation mixing ratio

$$w_v = w_s(T_D) \quad [12]$$

or equivalently

$$e = e_s(T_D) \quad [13]$$

We can determine the dew point temperature by inverting either eqn [12] or eqn [13], which can be done using eqns [3] and [6]. Although the dew point temperature is expressed in Kelvin, the dew point temperature is a measure not of temperature, but of atmospheric humidity. The term $T - T_D$ is called the dew point depression. Dew point depression is inversely proportional to relative humidity; a relative humidity of 100% corresponds to a dew point depression of zero.

Analogously to the dew point temperature, we define the frost point temperature as the temperature at which ice saturation occurs. The frost point temperature, T_F , is thus defined as

$$e = e_{si}(T_F) \quad [14]$$

or equivalently as

$$w_v = w_{si}(T_F) \quad [15]$$

where w_{si} is the saturation mass mixing ratio over ice. Note that for a given atmospheric mixing ratio, $T_F < T_D$, since $e_{si}(T) < e_s(T)$.

Example Calculations

Air has temperature of 30°C, a relative humidity of 50%, and a pressure of 1000 hPa. This information can be used to determine each of the humidity variables for these conditions:

partial pressure of water vapor, $e = 21.2$ hPa

water vapor mixing ratio, $w_v = 0.0135$

specific humidity, $q_v = 0.0133$

dew point temperature, $T_D = 291.8$ K

See also: Satellites and Satellite Remote Sensing: Water Vapor. Thermodynamics: Moist (Unsaturated) Air; Saturated Adiabatic Processes.

Further Reading

Curry, J.A., Webster, P.J., 1999. *Thermodynamics of Atmospheres and Oceans*. Academic Press, San Diego.
 Flatau, P.J., Walko, R.L., Cotton, W.R., 1992. Polynomial fits to saturation vapor pressure. *Journal of Applied Meteorology* 31, 1507–1513.

Moist (Unsaturated) Air

JA Curry, Georgia Institute of Technology, Atlanta, GA, USA

© 2003 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 5, pp 2274–2278, © 2003, Elsevier Ltd.

Introduction

The thermodynamics of air is complicated by the presence of a mixture of gases, with water vapor having a variable concentration in the atmosphere. The term 'moist air' refers to the dry air gases (predominantly nitrogen and oxygen) plus water vapor. The basic thermodynamics of moist air includes the equation of state, and applications of the first and second laws of thermodynamics to the moist atmosphere (but excludes clouds and the condensation process). These principles of thermodynamics are used to determine how moist air responds to heating and cooling, and how its temperature changes in response to rising and sinking motions. Thermodynamic processes related to condensation, clouds, and precipitation are considered elsewhere in this encyclopedia (see **Thermodynamics: Saturated Adiabatic Processes**).

Equation of State for Air

Except when water vapor is near condensation, air is observed to obey the ideal gas law

$$pV = nR^*T \quad [1]$$

where p is the pressure, V is the volume, n is the number of moles, R^* is the universal gas constant, and T is the temperature. Since the volume and number of moles are not easily measured in the atmosphere, a more useful form of the ideal gas law can be obtained by dividing both sides of eqn [1] by mass, m , yielding

$$p \frac{V}{m} = \frac{n}{m} R^* T \quad [2]$$

Using the definitions of molecular weight, $M = m/n$, and specific volume, $\nu = V/m$, eqn [2] can be written as

$$p\nu = \frac{R^*}{M} T \quad [3]$$

A specific gas constant, R , may be defined as $R = R^*/M$, so that eqn [3] becomes

$$p\nu = RT \quad [4]$$

Strictly speaking, air does not have a molecular weight, since it is a mixture of gases and there is no such thing as an 'air molecule.' However, it is possible to assign an apparent molecular weight to air, since air as a mixture is observed to behave like an ideal gas. By applying Dalton's law of partial pressures, which states that the total pressure exerted by a mixture of gases is equal to the sum of the partial pressures that would be exerted by each constituent alone, it can be determined that each gas individually obeys the ideal gas law

and that the ideal gas law for a mixture of gases can be written as

$$V \sum_j p_j = T \sum_j m_j R_j$$

where p_j , m_j , and R_j are the partial pressure, mass, and specific gas constant, respectively, for the j th constituent, and $\nu = V/m$ has been used. A mean specific gas constant can now be defined as

$$\bar{R} = \frac{\sum_j m_j R_j}{m} \quad [5]$$

The equation of state for the mixture of dry air gases can therefore be written as

$$p\nu = R_d T \quad [6]$$

where R_d is the specific gas constant for dry air. For the gaseous composition of the dry air (which excludes water vapor), a value for R_d is determined to be $287.104 \text{ J K}^{-1} \text{ kg}^{-1}$. The mean molecular weight of the mixture is

$$\bar{M} = \frac{\sum_i n_i M_i}{n} = \frac{m}{n}$$

The mean molecular weight for dry air gases, M_d , is determined to be 28.96 g mol^{-1} .

The equation of state for air is complicated by the presence of water vapor, which has a variable amount in the atmosphere. Assuming that the water vapor is not near condensation, the ideal gas law may be used and is obtained

$$e = \rho_v R_v T \quad [7]$$

where the notation e is commonly used to denote the partial pressure of water vapor and the subscript v denotes the vapor. The specific gas constant for water vapor is $R_v = R^*/M_v = 461.51 \text{ J K}^{-1} \text{ kg}^{-1}$. In a mixture of dry air and water vapor (moist air), the equation of state is

$$p = p_d + e = (\rho_d R_d + \rho_v R_v) T \quad [8]$$

The subscript d denotes the dry air value, and the absence of a subscript denotes the value for the mixture of dry air plus water vapor.

The specific gas constant for moist air is determined from eqn [5] to be

$$R = \frac{(m_d R_d + m_v R_v)}{(m_d + m_v)} \quad [9]$$

where m_d and m_v are the mass of dry air and water vapor, respectively, and $m = m_d + m_v$. The specific humidity, q_v , is defined as

$$q_v = \frac{m_v}{m_v + m_d} \quad [10]$$

so that the specific gas constant for moist air can be written as

$$R = (1 - q_v)R_d + q_vR_v = R_d(1 + 0.608q_v) \quad [11]$$

Incorporating eqn [11] into eqn [8], the equation of state for moist air becomes

$$p_v = R_d(1 + 0.608q_v)T \quad [12]$$

It is awkward to have a variable gas constant, so it is the convention among meteorologists to make the humidity adjustment to the temperature rather than to the gas constant. Thus a virtual temperature, T_v , is defined as:

$$T_v = (1 + 0.608q_v)T \quad [13]$$

so that the ideal gas law for moist air becomes

$$p_v = R_dT_v \quad [14]$$

The virtual temperature may be interpreted as the temperature of dry air having the same values of p and v as the moist air under consideration. Since q_v seldom exceeds 0.02, the virtual temperature correction rarely exceeds more than 2 or 3 °C; however, the small virtual temperature correction has an important effect on buoyancy and hence vertical motions in the atmosphere.

First and Second Laws of Thermodynamics

The first law of thermodynamics is an extension of the principle of conservation of mechanical energy. The conservation principle can be used to define a function U called the internal energy. When an increment of heat dQ is added to moist air, the energy is used to increase the speed of the molecules (i.e., to increase the temperature of the system). The internal energy of a system can increase when heat enters into the system from the surroundings, and/or when work, dW , is done on the system by the surroundings. If dU is taken to denote an increment of internal energy, then

$$dU = dQ + dW \quad [15]$$

This statement is the differential form of the first law of thermodynamics.

The type of work of primary importance in the atmosphere is expansion work, which is defined as

$$dW = -pdV \quad [16]$$

where dV is the differential volume change associated with the work done against the external pressure, p . There are numerous examples of expansion work in the atmosphere, wherein a parcel of air rises in the atmosphere and its pressure decreases and volume increases. Some processes that cause air to rise are (1) orographic lifting, (2) frontal lifting, (3) low-level convergence, (4) buoyant rising of warm air, and (5) mechanical mixing.

Under the conditions of expansion work, the first law of thermodynamics is written as

$$dU = dQ - pdV \quad [17]$$

The first law of thermodynamics can also be written in terms of enthalpy, where enthalpy, H , is defined as $H = U + pV$:

$$dH = dQ + Vdp \quad [18]$$

Equations [17] and [18] are equivalent forms of the first law of thermodynamics. The enthalpy form of the first law is advantageous when considering constant pressure processes, and is more often used in atmospheric science.

A change in heat is related to temperature by

$$dQ_v = mc_vdT$$

for a constant volume process and

$$dQ_p = mc_pdT$$

for a constant pressure process, where c_v is the specific heat at constant volume and c_p is the specific heat at constant pressure. This implies, from eqns [17] and [18], that for ideal gases

$$dU = mc_vdT$$

$$dH = mc_pdT$$

For an ideal gas, it can be shown that $c_p - c_v = R$, where R is the specific gas constant. The magnitude of c_p is greater than c_v because in a constant pressure expansion part of the heat is used for expansion work, while in constant volume heating all of the heat is used to increase the temperature.

The second law of thermodynamics limits both the amount and the direction of heat transfer. According to the second law, a given amount of heat cannot be totally converted into work, thus limiting the amount of heat transfer, and the spontaneous flow of heat must be from a body or substance with a higher temperature to one with a lower temperature, thus stipulating the direction of heat transfer.

Central to understanding the second law of thermodynamics is the concept of reversible and irreversible processes. A reversible process is one in which the system is in an equilibrium state throughout the process. Thus the system passes at an infinitesimal rate through a continuous succession of balanced states that are infinitesimally different from each other. In such a scenario, the process can be reversed, and the system and its environment will return to the initial state. Irreversible processes proceed at finite rates: if the system is restored to its initial state, the environment will have changed from its initial state. The term 'irreversible' does not mean that a system cannot return to its original state, but that the system plus its environment cannot be thus restored. Examples of irreversible processes in the atmosphere are radiative transfer and precipitation.

According to the second law of thermodynamics, there exists an additive function of state known as the equilibrium entropy, which can never decrease in a thermally isolated system. In other words, a thermally isolated system cannot spontaneously regain order which has been lost. The second law may be applied to a system and its surroundings to determine the total entropy change $\Delta\eta_{\text{tot}}$

$$\Delta\eta_{\text{tot}} \geq 0$$

which is known as Clausius' inequality. For the special case of a reversible adiabatic process, the entropy change will be zero in the system alone, $\Delta\eta_{\text{sys}} = 0$, for all reversible adiabatic changes. Reversible adiabatic processes are therefore isentropic.

Consider the first law of thermodynamics in enthalpy form (eqn [18]) for a reversible process:

$$\frac{1}{m}dQ = c_p dT - v dp$$

Reversible heating is an abstract concept, whereby heating of a system occurs infinitesimally slowly through contact with an infinite heat reservoir. For the reversible expansion of an ideal gas, the specific volume may be substituted from the equation of state and divided by temperature:

$$\frac{1}{m} \frac{dQ}{T} = c_p \frac{dT}{T} - R \frac{dp}{p} = c_p d(\ln T) - R d(\ln p) \quad [19]$$

Now, a new thermodynamic state function, the entropy, η , can be defined with units Joules per Kelvin per kilogram, to be

$$d\eta = \left(\frac{dQ}{T} \right)_{\text{rev}} \quad [20]$$

Entropy changes for an ideal gas can be determined from eqns [19] and [20]:

$$\frac{1}{m} d\eta = c_p d(\ln T) - R d(\ln p) \quad [21]$$

Adiabatic Processes in the Moist (but Unsaturated) Atmosphere

The first law of thermodynamics is applied to dry air. The thermodynamic characteristics of dry air have been shown to be:

1. the equation of state is $p\nu = R_d T$;
2. the internal energy is a function of its temperature alone ($dU = m c_v dT$; $dH = m c_p dT$); and
3. the specific heats are related by $c_p - c_v = R_d$.

The first law of thermodynamics for dry air is thus written as

$$m c_v dT = dQ - p dV \quad [22a]$$

$$m c_p dT = dQ + V dp \quad [22b]$$

in internal energy (eqn [22a]) and enthalpy (eqn [22b]) forms.

An adiabatic process is one in which no heat is exchanged between the system and its environment, so that $dQ = 0$. The first law in enthalpy form for an adiabatic expansion of an ideal gas is thus written as

$$c_p dT = v dp \quad [23a]$$

where eqn [23a] has been divided by mass (m) and $v = V/m$ is the specific volume. Considering a reversible adiabatic expansion for an ideal gas, from eqn [23a] and the equation of state eqn [6] is obtained

$$c_p \frac{dT}{T} = R_d \frac{dp}{p} \quad [23b]$$

which may be integrated between an initial and final state to give

$$c_p \ln \left(\frac{T_2}{T_1} \right) = R_d \ln \frac{p_2}{p_1}$$

so that

$$\frac{T_2}{T_1} = \left(\frac{p_2}{p_1} \right)^{R_d/c_p} \quad [24]$$

This relationship holds for reversible, adiabatic processes.

The lifting of air parcels by processes such as orographic lifting, frontal lifting, low-level convergence, and vertical mixing causes pressure to decrease, with a corresponding temperature decrease that is specified by eqn [24]. The lifting of air parcels can be considered a dry adiabatic process as long as condensation does not occur.

If $p_0 = 1000$ hPa is chosen to correspond to a temperature θ , eqn [24] becomes

$$\theta = T \left(\frac{p_0}{p} \right)^{R_d/c_p} \quad [25]$$

where $R_d/c_d = 0.286$ for dry air. The temperature θ is called the potential temperature. It is the temperature a sample of gas would have if it were compressed (or expanded) in an adiabatic reversible process from a given state, p and T , to a pressure of 1000 hPa. θ is a conservative quantity for reversible adiabatic processes in the atmosphere. Consider an atmospheric temperature profile with a lapse rate $\Gamma = 6^\circ \text{C km}^{-1}$. For atmospheric pressures less than 1000 hPa, the potential temperature of a sample of air is greater than the physical temperature, since adiabatic compression must be done to lower the parcel to 1000 hPa. Conversely, the potential temperature of a sample of air with pressure greater than 1000 hPa will be less than the physical temperature. At a pressure level of 1000 hPa, $\theta = T$.

A relationship between entropy and potential temperature for the dry atmosphere is derived by logarithmically differentiating eqn [25]

$$d(\ln \theta) = d(\ln T) - \frac{R_d}{c_p} d(\ln p). \quad [26]$$

Comparison of eqn [26] with eqn [23b] shows that

$$\frac{1}{m} d\eta = c_p d(\ln \theta) \quad [27]$$

This means that for reversible processes in an ideal gas, potential temperature may be considered an alternative variable for entropy.

Equation [25] does not account for water vapor. The specific heat of moist air is

$$c_p = (1 - q_v) c_{pd} + q_v c_{pv} \approx c_{pd} (1 + 0.87 q_v) \quad [28]$$

where the subscripts d and v refer to dry air and water vapor, respectively. The ratio R/c_p for moist air can then be determined from eqn [11] to be

$$\frac{R}{c_p} = \frac{R_d}{c_{pd}} \left(\frac{1 + 0.608 q_v}{1 + 0.87 q_v} \right) \approx \frac{R_d}{c_{pd}} (1 - 0.26 q_v) \quad [29]$$

The potential temperature of moist air then becomes

$$\theta = T \left(\frac{p_0}{p} \right)^{R_d(1 - 0.26 q_v)/c_{pd}} \quad [30]$$

The difference between the dry air and moist air values of θ is generally less than 0.1°C , so that adiabatic expansion or compression of moist air can be treated as if it were dry air. Note that θ is not conserved if a phase change of water occurs.

A virtual potential temperature, θ_v , can also be defined by neglecting the water vapor dependence of the exponent of [eqn \[30\]](#) and replacing the temperature by the virtual temperature

$$\theta_v = T_v \left(\frac{p_0}{p} \right)^{R_d/c_{pd}} \quad [31]$$

If the adiabatic ascent of a parcel of air is considered in the atmosphere, the temperature of the parcel will decrease and the potential temperature will remain the same. The rate of decrease of temperature with height in an adiabatic ascent can be determined by considering the first law in enthalpy form for an adiabatic process ([eqn \[23a\]](#))

$$c_p dT = v dp$$

If it is assumed that the ascent of the parcel does not involve any large vertical accelerations and the hydrostatic relation applies, the hydrostatic relation can be substituted

$$g = \frac{1}{\rho} \frac{\partial p}{\partial z}$$

where ρ is the density of air, into [eqn \[23a\]](#) to give

$$c_p dT = -g dz$$

From the definition of lapse rate, $\Gamma = -dT/dz$, an expression for the dry adiabatic lapse rate, Γ_d , can be written as

$$\Gamma_d = \frac{g}{c_p} \quad [32]$$

which has a value of approximately 9.8 K km^{-1} . Both [eqns \[25\] and \[32\]](#) describe the temperature evolution of a parcel of air in dry adiabatic ascent, but [eqn \[32\]](#) is slightly more restrictive than [eqn \[25\]](#) in that it applies only to a hydrostatic process. The adiabatic lapse rate for moist air differs only slightly from [eqn \[32\]](#) and can be expressed as

$$\Gamma = \frac{g}{c_p(1 + 0.87q_v)}$$

Outside of clouds, diabatic processes (e.g., radiative heating) operate on much longer timescales than the characteristic timescale of vertical displacement of the air parcel. Therefore, the lifting of air parcels by processes such as orographic lifting, frontal lifting, low-level convergence, and vertical mixing can be considered dry adiabatic processes as long as condensation does not occur.

See also: Thermodynamics: Humidity Variables; Saturated Adiabatic Processes.

Further Reading

Curry, J.A., Webster, P.J., 1999. *Thermodynamics of Atmospheres and Oceans*. Academic Press, San Diego, CA.

Saturated Adiabatic Processes

JA Curry, Georgia Institute of Technology, Atlanta, GA, USA

© 2003 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 6, pp 2278–2282, © 2003, Elsevier Ltd.

Introduction

Adiabatic processes of moist (but unsaturated) air are described elsewhere in this encyclopedia (*see Thermodynamics: Moist (Unsaturated) Air*), and it is shown that potential temperature remains constant during such processes as long as condensation does not occur. Once condensation occurs in adiabatic cooling associated with ascent, the latent heat of condensation is released. The rate at which saturated air cools as it expands adiabatically is smaller than the rate at which unsaturated air cools adiabatically, because part of the cooling is canceled by the latent heat released during condensation.

The term ‘saturation’ indicates that the atmosphere has a relative humidity (RH) of 100% or greater, with respect to either liquid water or ice (*see Thermodynamics: Humidity Variables*). In saturated air, the thermodynamic system under consideration consists of dry air gases (primarily nitrogen and oxygen), water vapor, and water in a condensed phase (liquid and/or ice particles).

The RH is defined in eqn [1].

$$H = \frac{w_v}{w_s(T)} \quad [1]$$

where w_v is the water vapor mixing ratio (defined as the ratio of the water vapor mass to the dry air mass) and w_s is the saturation mixing ratio (defined as the ratio of the water vapor mass at saturation to the dry air mass). For initially unsaturated air to become saturated, the RH must increase. An increase in RH can be accomplished by increasing the amount of water vapor in the air (i.e., increasing w_v) and/or by cooling the air, which decreases $w_s(T)$. Here, the focus is on adiabatic cooling in rising air as a mechanism for reaching saturation and the subsequent thermodynamic processes associated with the adiabatic cooling of saturated (and cloudy) air.

Adiabatic Processes

Expansion in the atmosphere occurs when air rises due to mechanical lifting (e.g., orographic and frontal), large-scale low-level convergence, turbulent mixing, and buoyancy effects. The expansion is adiabatic if no heat is exchanged between the air and the environment (e.g., radiative transfer).

The entropy equation for an adiabatic process for moist (but unsaturated) air in the absence of condensation is written as eqn [2] (*see Thermodynamics: Moist (Unsaturated) Air*).

$$0 = c_p d(\ln T) - R d(\ln p) \quad [2]$$

where c_p is the specific heat at constant pressure of air, T is temperature, R is the specific gas constant for air, and p is

pressure. From this equation, an expression for the potential temperature, θ (eqn [3]), can be derived.

$$\theta = T \left(\frac{p_0}{p} \right)^{R/c_p} \quad [3]$$

The potential temperature is the temperature air would have if it were compressed (or expanded) in an adiabatic reversible process from a given state, p and T , to a pressure of 1000 hPa. θ is a conservative quantity for reversible adiabatic processes in the atmosphere in the absence of phase changes associated with condensation. From the definition of lapse rate, $\Gamma = -dT/dz$, [1], and the hydrostatic equation, an expression for the dry adiabatic lapse rate, Γ_d , (eqn [4]), can be written.

$$\Gamma_d = \frac{g}{c_p} \approx 9.8^\circ \text{C km}^{-1} \quad [4]$$

As air expands adiabatically and cools, the RH increases as the temperature and saturation mixing ratio decreases. The water vapor mixing ratio remains constant during adiabatic ascent. At some point, the RH reaches 100%, and further cooling results in saturation. (Note: Condensation is initiated typically at relative humidities that slightly exceed 100%.) The temperature at which saturation is reached can be approximated using eqn [5].

$$T_s = \frac{1}{\frac{1}{T - 55} - \frac{\ln H}{2840}} + 55 \quad [5]$$

for initial values of T (in Kelvin) and H . From eqn [3], the saturation pressure, p_s , can be determined as in eqn [6].

$$p_s = p \left(\frac{T_s}{T} \right)^{c_p/R_d} \quad [6]$$

The coordinate (T_s, p_s) is known as the saturation point of the air mass.

During ascent, the water vapor mixing ratio, w_v , remains constant until saturation occurs. The dew point temperature (*see Thermodynamics: Humidity Variables*), however, decreases slightly during the ascent as pressure decreases. The lifting condensation level, z_s , corresponds to the level of the saturation pressure, p_s , which can be approximated as in eqn [7].

$$z_s = 0.12(T_0 - T_{D0}) \quad (\text{km}) \quad [7]$$

This relation is an approximate expression of the height of the lifting condensation level achieved in an adiabatic ascent where T_0 and T_{D0} represent the initial temperature and dew point temperature of the air mass that is being lifted. Calculation of the lifting condensation level provides a good estimate of the cloud base height for clouds that form by adiabatic ascent.

Once saturation occurs, further lifting of the air mass results in condensation. Because of the latent heat released during condensation, the decrease of temperature with height will be smaller than that in dry adiabatic ascent. In addition, the potential temperature, θ , which was conserved in a reversible dry adiabatic ascent, is no longer conserved once condensation occurs.

The adiabatic entropy equation for air with changes of phase between water vapor, liquid, and ice is written as eqn [8].

$$0 = (c_{pd} + w_l c_l) d(\ln T) - R_d d(\ln p_d) - w_v d\left(\frac{A_{lv}}{T}\right) + d\left(\frac{L_{lv} w_v}{T}\right) + w_i d\left(\frac{A_{il}}{T}\right) - d\left(\frac{L_{il} w_i}{T}\right) \quad [8]$$

Here, c_{pd} is the specific heat of dry air, c_l is the specific heat of liquid water, R_d is the specific gas constant for dry air, w_l is the liquid water mixing ratio, w_i is the ice water mixing ratio, w_t is the total water mixing ratio, L_{lv} is the latent heat of vaporization, L_{il} is the latent heat of sublimation, A_{lv} is the affinity for vaporization, and A_{il} is the affinity for freezing. An approximate form of the entropy equation that has no ice phase assumes that condensation occurs at 100% RH, and neglects the specific heats of water relative to dry air, is written as eqn [9].

$$0 = c_{pd} d(\ln T) - R_d d(\ln p) + \frac{L_{lv}}{T} dw_s \quad [9]$$

The saturated adiabatic lapse rate, Γ_s , can be determined from the adiabatic entropy eqn [9], the hydrostatic equation, ideal gas law, the Clausius–Clapeyron equation, and dry adiabatic lapse rate (eqn [10]).

$$\Gamma_s = \Gamma_d \left[\frac{1 + (L_{lv} w_s / R_d T)}{1 + (\varepsilon L_{lv}^2 w_s / c_{pd} R_d T^2)} \right] \quad [10]$$

where $\varepsilon = 0.622$ (the ratio of the molecular weights of water to dry air). The denominator of eqn [10] is larger than the numerator, and thus $\Gamma_s < \Gamma_d$. Table 1 shows the values of Γ_s for selected values of T and p . It is seen that the temperature variation of Γ_s exceeds the pressure variation. At low temperatures and high pressures, Γ_s approaches Γ_d . Values of Γ_s determined from eqn [10] are within about 0.5% of the values determined from a more exact form of the entropy eqn [8]. Because of the approximate nature of eqn [9], Γ_s is sometimes called the pseudoadiabatic lapse rate.

The amount of water condensed in saturated adiabatic ascent, called the adiabatic liquid water content, can be

determined from the adiabatic entropy eqn [9] and the hydrostatic equation (eqn [11]).

$$dw_l = \frac{c_p}{L_{lv}} \left[\frac{dT}{dz} + \frac{g}{c_p} \right] dz \quad [11]$$

Substituting $\Gamma_d = g/c_p$ and $\Gamma_s = -dT/dz$ yields eqn [12]

$$dw_l = \frac{c_p}{L_{lv}} [\Gamma_d - \Gamma_s] dz \quad [12]$$

Integrating eqn [12] from cloud base to height z gives the liquid water content at height z . Because of the complicated form of Γ_s , this equation must be integrated numerically. Integration of eqn [12] shows that the adiabatic liquid water content increases with height above the cloud base and increasing cloud base temperature. Because of the variation of Γ_s with temperature, clouds with warmer bases have larger values of $\Gamma_d - \Gamma_s$ and thus larger values of the adiabatic liquid water content. The adiabatic liquid water content represents an upper bound on the liquid water that can be produced in a cloud by rising motion. Processes such as precipitation and mixing with dry air reduce the cloud liquid water content relative to the adiabatic value.

As adiabatic cooling proceeds, the cloud may eventually cool to the point where ice crystals form. Assuming that a water cloud is present initially, then the formation of ice crystals releases latent heat during fusion. Once the cloud glaciates, it is supersaturated with respect to ice, and deposition occurs on the ice crystals, releasing the latent heat of sublimation, until the ambient RH is at ice saturation. Further cooling will result in the increase of ice water content in the cloud and the release of the latent heat of sublimation into the atmosphere.

Assuming that the thermodynamic system consists of moist air plus liquid water, and that the freezing and subsequent deposition occur isobarically and adiabatically, then the enthalpy of the system will not change during this transformation. This process can be idealized by assuming that first the water freezes at constant temperature and latent heat of freezing is released, then water vapor is deposited on the ice and latent heat of sublimation is released. The temperature change associated with the freezing and subsequent deposition can be approximated from eqn [8] according to eqn [13].

$$\Delta T = \frac{L_{il} w_l + L_{iv} w_s [1 - (e_{si}/e_s)]}{c_p + (\varepsilon w_l L_{iv}^2 / R_d T^2)} \quad [13]$$

This expression gives the increase in temperature due to the freezing of cloud water and the subsequent deposition of water vapor onto the ice crystals. In clouds that cool by adiabatic ascent, the freezing does not occur isobarically, but gradually over a temperature interval.

Once the cloud has glaciated, further adiabatic ascent results in deposition of water vapor onto the ice crystals. Analogously to eqn [10], the ice-saturation adiabatic lapse rate is given by eqn [14], where w_{si} is the saturation mixing ratio with respect to ice and L_{iv} is the latent heat of sublimation.

$$\Gamma_{si} = \Gamma_d \left[\frac{1 + (L_{iv} w_{si} / R_d T)}{1 + (\varepsilon L_{iv}^2 w_{si} / c_{pd} R_d T^2)} \right] \quad [14]$$

The melting process is distinctly different from the freezing process. Melting may occur as ice particles fall to temperatures

Table 1 Γ_s for selected values of temperature and pressure ($K \text{ km}^{-1}$)

T ($^{\circ}C$)	Pressure (hPa)		
	1000	700	500
−30	9.2	9.0	8.7
−20	8.6	8.2	7.8
−10	7.7	7.1	6.4
0	6.5	5.8	5.1
10	5.3	4.6	4.0
20	4.3	3.7	3.3

that are above the melting point. In contrast to freezing, which may be distributed through a considerable vertical depth, melting of ice particles can be quite localized, occurring in a very narrow layer around the freezing point. Cooling of the atmosphere from the melting can result in an isothermal layer near 0 °C. Because of their large size and density, hailstones do not melt at the freezing level in the same manner as a small ice crystal or a snowflake with a low density, but melt over a deeper layer. If atmospheric relative humidities are low in the atmosphere below the melting level, then the melting water will evaporate, cooling the hailstone and retarding the melting.

Conserved Thermodynamic Variables under Saturated Conditions

Potential temperature is a conserved variable in reversible adiabatic processes. The concept of potential temperature becomes less useful when applied to a saturated air, since potential temperature is not conserved during phase changes of water. Derivation of a potential temperature that is conserved in saturated adiabatic ascent eliminates the need to include latent heat source terms in the time-dependent thermodynamic equation.

An analogous variable that is conserved for a cloud in adiabatic ascent can be determined that relates temperature and pressure in a saturated adiabatic process. A conserved temperature for cloud in adiabatic ascent can be derived from eqn [8]. A conserved potential temperature for clouds will obviously be far more complex than the potential temperature derived for a dry adiabatic process, since eqn [8] is considerably more complex than eqn [1]. A number of different conserved potential temperatures have been used for clouds that employ various approximate forms of eqn [8].

The simplest possible case is that in which saturation conditions are maintained, ice is not present, and the heat capacities of the water vapor and condensed water are neglected relative to that of dry air. Using these approximations, the entropy eqn [8] becomes eqn [15].

$$0 = c_{pd}d(\ln T) - R_d d(\ln p) + d\left(\frac{L_{lv}w_s}{T}\right) \quad [15]$$

For a dry adiabatic process, eqn [16] is obtained from eqn [3].

$$c_{pd}d(\ln \theta) = c_{pd}d(\ln T) - R_d d(\ln p) \quad [16]$$

Equating eqn [3] with eqn [16] yields eqn [17].

$$-d\left(\frac{L_{lv}w_s}{T}\right) = c_{pd}d(\ln \theta) \quad [17]$$

This expression is integrated to a height in the atmosphere where all of the water vapor has been condensed out by adiabatic cooling. The corresponding temperature is called the equivalent potential temperature, θ_e , given in eqn [18].

$$\theta_e = \theta \exp\left(\frac{L_{lv}w_s}{c_{pd}T}\right) \quad [18]$$

It is easily determined that $\theta_e > \theta$, which arises from the latent heat released from the condensation of water vapor. Because of the approximations made in eqn [16], the

equivalent potential temperature is only approximately conserved in a saturated adiabatic process. Although approximate, eqn [18] retains the essential physics of the process, whereby the condensation of water vapor provides energy to the moist air and increases its temperature relative to what the temperature would have been in dry adiabatic ascent.

An alternative but analogous potential temperature, the liquid water potential temperature, θ_l , is derived as follows. Equation [16] is written as in eqn [19], where $dw_s = -dw_l$

$$0 = c_{pd}d(\ln T) - R_d d(\ln p) - d\left(\frac{L_{lv}w_l}{T}\right) \quad [19]$$

Then a procedure analogous to the derivation of θ_e write θ_l as in eqn [20] can be followed.

$$\theta_l = \theta \exp\left(-\frac{L_{lv}w_l}{c_{pd}T}\right) \quad [20]$$

One advantage of θ_l over θ_e is that θ_l reverts to θ , the dry potential temperature, in the absence of liquid water.

In the presence of ice, an ice-liquid water potential temperature, θ_{il} , can be derived from the approximate form of eqn [8] given as eqn [21].

$$0 = c_{pd}d(\ln T) - R_d d(\ln p) - d\left(\frac{L_{lv}w_l}{T}\right) - d\left(\frac{L_{iv}w_i}{T}\right) \quad [21]$$

θ_{il} can then be written as in eqn [22].

$$\theta_{il} = \theta \exp\left(-\frac{L_{lv}w_l}{c_{pd}T} - \frac{L_{iv}w_i}{c_{pd}T}\right) \quad [22]$$

The derivation of the ice-liquid water potential temperature implies that it is applicable only under conditions of equilibrium, since the affinity terms were not included. Since ice and liquid are both at equilibrium only at the triple point, use of the ice-liquid water potential temperature is inconsistent physically at temperatures away from the triple point. Nevertheless, the ice-liquid water potential is an economical and not too inaccurate way to treat ice processes in a numerical cloud model.

The entropy potential temperature, θ_η , includes ice processes and is derived from the complete form of the adiabatic entropy eqn [8] as given in eqn [23].

$$\theta_\eta = T\left(\frac{p_0}{p}\right)^{R_d/(c_{pd}+w_l c_l)} \times \exp\left(\frac{(L_{iv} + A_{iv})w_l}{(c_{pd} + w_l c_l)T} - \frac{(L_{il} + A_{il})w_i}{(c_{pd} + w_l c_l)T}\right) \quad [23]$$

The entropy potential temperature is thus the most general potential temperature considered here. Unlike θ_l and θ_{il} , θ_η is applicable to nonequilibrium conditions such as subsaturated or supersaturated environments.

Another moist thermodynamical variable that is often used is the moist static energy, h (eqn [24]).

$$h = (c_{pd} + w_l c_l)T + L_{lv}w_v + (1 + w_l)gz \quad [24]$$

The moist static energy is conserved for adiabatic, saturated, or unsaturated transformations for a closed system in which the pressure change is hydrostatic.

It is important to note the conditions under which θ_e and the other conserved thermodynamic variables are not conserved. Examples include cases where external radiative

heating or conduction takes place, since these alter the entropy. Other examples include atmospheric conditions in which latent heating occurs externally, such as the evaporation of water into air from the ocean or when precipitation falls out.

See also: **Thermodynamics:** Humidity Variables; Moist (Unsaturated) Air.

Further Reading

- Curry, J.A., Webster, P.J., 1999. *Thermodynamics of Atmospheres and Oceans*. Academic Press, London, UK.
- Dutton, J.A., 1986. *The Ceaseless Wind: An Introduction to the Theory of Atmospheric Motion*. Dover Publications, Mineola, NY.
- Emanuel, K.A., 1994. *Atmospheric Convection*. Oxford University Press, New York, NY.
- Iribarne, J.V., Godson, W.L., 1981. *Atmospheric Thermodynamics*. Kluwer, Boston, MA.

Thermosphere

SC Solomon and RG Roble, National Center for Atmospheric Research, Boulder, CO, USA

© Published by Elsevier Ltd.

This article is a revision of the previous edition article by R G Roble, volume 6, pp 2282–2290, © 2003, Elsevier Ltd.

Synopsis

The thermosphere is the atmospheric region from ~ 85 to ~ 500 km altitude, containing the ionosphere. It is characterized by high temperature and large variability, in response to changes in solar ultraviolet radiation and solar-driven geomagnetic activity. The composition of the lower thermosphere is primarily N_2 and O_2 , similar to the lower and middle atmosphere, but in the upper thermosphere atomic oxygen (O) becomes the dominant gas, due to the importance of photodissociation and molecular diffusion at high altitude. Thermospheric dynamical motions result from complex interactions between solar heating, auroral processes, the resulting pressure gradients, and dissipative processes.

Introduction

The vertical structure of the Earth's atmosphere is categorized by region, based on the vertical structure of the temperature profile. These regions are the troposphere, the stratosphere, the mesosphere, the thermosphere, and the exosphere, and they are separated by boundaries known as the tropopause, the stratopause, the mesopause, and the exobase. The troposphere extends between the ground and about 10 km, where most of the weather that we experience exists. The stratosphere lies between 10 and 50 km, where the ozone layer resides. The mesosphere lies between 50 and 85 km, and the thermosphere begins at ~ 85 km and extends to about 500 km. The exosphere begins near 500 km and extends far out into space. The number densities of the gases in the exosphere are so low that continuum fluid dynamics no longer apply, and gas particles follow ballistic trajectories and orbits.

The thermosphere is the rarefied region of our atmosphere where the temperature increases dramatically with altitude. Many satellites orbit in the thermosphere, experiencing a small drag force from collisions with atoms and molecules that gradually lowers their orbits and eventually causes reentry into the atmosphere. Embedded within the thermosphere is the ionosphere, a weakly ionized plasma that has strong dynamic and electrodynamic interactions with the neutral gases in the thermosphere. Both the thermosphere and ionosphere are strongly influenced by the absorption of solar ultraviolet (UV) radiation. In the polar regions, the thermosphere and ionosphere are affected by auroral processes, that result from the interaction of the solar wind with the Earth's magnetic field. They are also influenced by dynamic processes propagating upward from the lower atmosphere, such as gravity waves, tides, and planetary waves. The combined effect of all of these various forcings gives rise to the great observed variability that exists in the thermosphere/ionosphere system.

Global Mean Structure

The physical and chemical processes that establish the global mean structure of the thermosphere are considerably different from those in the lower atmosphere. The basic structure is

established primarily by the absorption of solar extreme ultraviolet (EUV) radiation at wavelengths shorter than 103 nm, which can ionize and dissociate the gases in the thermosphere, and by the absorption of solar UV radiation at wavelengths between 130 and 175 nm, which can dissociate molecular oxygen. The energetic EUV wavelengths that ionize the gases in the thermosphere create the ionosphere, a weakly ionized plasma that is embedded within and interacts with the neutral gases in the thermosphere. The radiation also dissociates the molecular species of O_2 and N_2 into atomic species of O and N, thus changing the composition of the thermosphere from a molecular atmosphere near the mesopause around 85 km to an atomic atmosphere at the exobase near 500 km. Virtually, all solar photons at wavelengths less than 200 nm are absorbed by the major constituents of the thermosphere, O, O_2 , and N_2 . The neutral gases of the thermosphere are heated locally by the absorption of solar EUV and UV radiation with a heating efficiency of about 33%. The remaining 67% can be either radiated upward to space or downward where it is absorbed in the lower atmosphere. It can also end up as chemical energy from the dissociation of molecular species into atomic species that are transported away from the site of photodissociation. Chemical energy can be transported long distances in the thermosphere because recombination, with the release of chemical energy, is a three-body reaction that requires a high density of neutral particles to proceed rapidly. This occurs mainly in the lower thermosphere near 90 km, where the atomic oxygen density peaks, and therefore the atomic species must be transferred to the lower thermosphere in order to recombine.

The infrared (IR) cooling mechanisms in the thermosphere (5.3 μm emission from NO, 15 μm emission from CO_2 , and 63 μm emission from the fine structure of the ground level of O) are all relatively weak and they are not sufficient to balance the solar heating and maintain radiative equilibrium. The thermosphere is therefore cooled primarily by downward molecular heat conduction above about 120 km and by downward eddy heat conduction to the vicinity of the mesopause where IR cooling becomes sufficiently large to radiate the excess energy to space. Downward thermal conduction implies that the temperature increases with altitude, thus giving a positive temperature gradient with increasing altitude in the

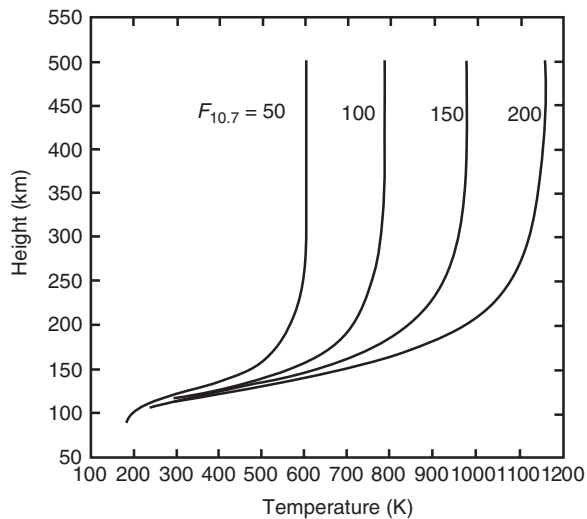


Figure 1 Global mean neutral gas temperature vertical profiles as a function of the solar decimeter radio emission at 10.7 cm wavelength ($F_{10.7}$). The solar decimeter radio emission $F_{10.7}$ (in units of $10^{-22} \text{ W m}^{-2} \text{ Hz}^{-1}$) is an indicator of the solar EUV radiative output.

lower thermosphere. In the upper thermosphere, the molecular thermal conduction coefficient becomes sufficiently large because of the rarefied atmosphere that it dominates the local heating rate. Therefore, the temperature becomes isothermal with altitude at a value called the exospheric temperature.

There is a large variation of temperature and density in the thermosphere because of a large variation of solar EUV and UV radiation with time. In addition to the overall change in solar radiation between solar minimum and maximum, due to the ~ 11 -year solar cycle, there are also shorter term variations on 27-day solar rotation timescales, daily and hourly variations, and impulsive solar flares. Solar spectral irradiance changes influence the thermal, compositional, and dynamic structure of the thermosphere and ionosphere. The global mean temperature profile as a function of the solar activity (represented by the solar 10.7 cm radio emission, which correlates well with the solar EUV flux) is shown in Figure 1. There is an exospheric temperature increase of about 600 K between solar minimum and solar maximum conditions, with a corresponding density increase of about an order of magnitude at 500 km altitude. At short solar EUV wavelengths, less than 30 nm, the EUV emissions that emanate from the Sun's corona are highly variable, increasing by factors of 10–100 between solar minimum and maximum. For wavelengths between 30 and 120 nm, the emissions that emanate from the chromosphere, a more stable region below the corona, are much less variable, showing factors of 2–3. At larger wavelengths, the variability then decreases from a factor of 2 near 100 nm to a factor of 1.1 near 200 nm.

The thermosphere is where the aurora occurs, in the polar regions (Figure 2). The energy and momentum inputs into the ionosphere and thermosphere are governed by the interaction of the highly variable solar wind with the Earth's geomagnetic field. The energetic auroral electrons and protons bombard the high-latitude thermosphere and ionize the neutral gas, dissociate molecular species, and excite various atomic and

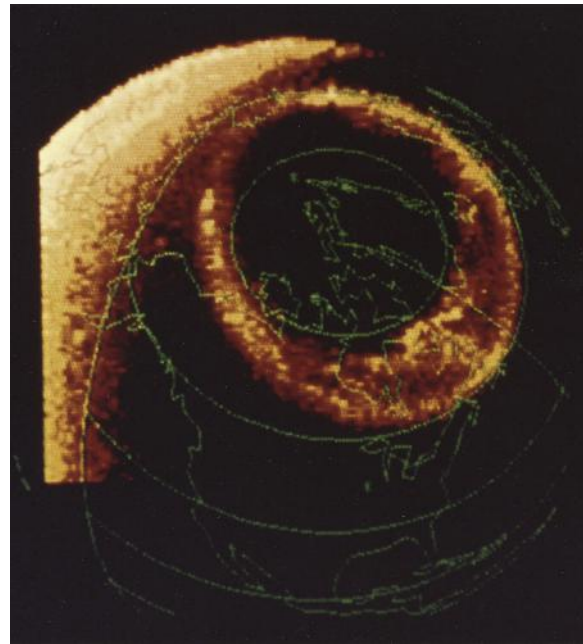


Figure 2 This auroral image, displayed as a geographic polar projection covering latitudes greater than 40° , illustrates the large spatial extent of the aurora during the later stages of a magnetic storm. This image was obtained from the University of Iowa's auroral imaging instrumentation on board the NASA satellite Dynamics Explorer 1 using an UV photometer in the 123–165 nm wavelength region.

molecular energy levels that produce the airglow seen visually from the ground and space. There is also a current system of about a million amperes associated with the aurora; the Joule dissipation of these currents strongly heats the thermosphere in the polar regions. The $\mathbf{J} \times \mathbf{B}$ momentum force of this current system can also accelerate the gases in the thermosphere through collisions with the drifting ions. The auroral particle and Joule heating rates are highly variable, on the order of 10^{10} , 10^{11} , and 10^{12} W for quiet, moderate, and highly disturbed auroral conditions, respectively.

Thermosphere Heating and Cooling Processes

To calculate the global average structure of the thermosphere, the thermodynamic and multispecies composition vertical structure equations must be solved simultaneously. The time rate of change of temperature is determined by a balance between downward molecular and eddy heat transport and radiative heat sources and sinks. In the thermosphere, the neutral gas heating rate above 100 km consists of the following 10 component processes: (1) absorption of solar UV radiation in the O_2 Schumann–Runge continuum region (130–175 nm); (2) likewise for the Schumann–Runge bands (175–210 nm); (3) heating by exothermic ion–neutral chemical reactions; (4) heating by exothermic neutral–neutral chemical reactions; (5) heating by elastic and inelastic collisions between ambient electrons, ions, and neutrals; (6) quenching of metastable species, such as $\text{O}(^1\text{D})$, by N_2 and O_2 ; (7) atomic oxygen

recombination; (8) heating by energetic photoelectrons and auroral electrons; (9) Joule dissipation of ionospheric currents in the aurora and dynamo region; and (10) heating by the molecular dissipation of tides, planetary waves, and gravity waves excited in the lower atmosphere that propagate to thermospheric heights. The main cooling processes for the thermosphere are (1) molecular heat conduction; (2) eddy heat conduction; (3) CO_2 $15\ \mu\text{m}$ radiation; (4) NO $5.3\ \mu\text{m}$ radiation; and (5) O $63\ \mu\text{m}$ radiation.

The vertical structure equation for composition includes (1) molecular diffusion; (2) eddy diffusion; (3) composition sources; and (4) composition sinks. A global average vertical structure of temperature and major species number density for solar cycle medium conditions, solar $F_{10.7}$ index ~ 150 , are shown in Figure 3.

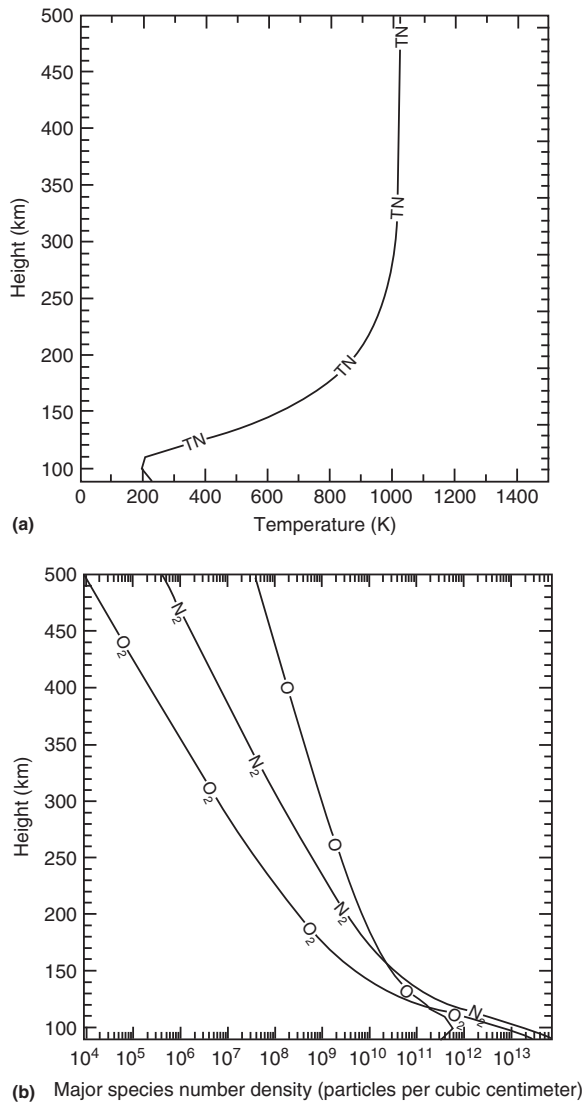


Figure 3 (a) A calculated global mean vertical neutral gas temperature profile for solar medium conditions ($F_{10.7} = 150$). (b) The corresponding vertical distribution of the major thermospheric species O , O_2 , and N_2 .

Equations of Motion

The equations of motion governing the dynamics of the thermosphere are the same as those used for predicting weather systems in the lower atmosphere; however, they need to be modified to consider three additional processes that are important at thermospheric heights. These are the viscous force, ion drag force, and pressure forces generated by differences in composition or mean molecular mass.

Kinematic molecular viscosity increases exponentially with altitude by several orders of magnitude in the thermosphere. Its main effect is to transfer momentum between the various altitude regions and thus to smooth out vertical gradients in wind velocity. This viscous force becomes so strong in the upper thermosphere that the winds are uniform with altitude above 300 km. It is also large enough to prevent the development of large horizontal shears and it effectively dissipates any turbulent structures. The viscous force depends upon the vertical wind shear and a characteristic time rate of change associated with viscosity, which is $\sim 10\ \text{s}$ near 200 km, varying exponentially with altitude as the inverse of air density for heights above and below.

The ion drag force is associated with the Ampere acceleration $\mathbf{J} \times \mathbf{B}$ that is important in the thermosphere. It is a collisional interaction of the neutral particles in the thermosphere with the charged particles of the ionosphere. It can be expressed as drag terms that are added to the zonal and meridional momentum equations (eqns [1] and [2]):

$$D_x = \lambda_{xx}(U_I - U_N) + \lambda_{xy}(V_I - V_N) \quad [1]$$

$$D_y = \lambda_{yy}(V_I - V_N) + \lambda_{yx}(U_I - U_N) \quad [2]$$

In these equations, the λ terms are given by eqns [3]–[5]:

$$\lambda_{xx} = \sigma_P B^2 \rho^{-1} \quad [3]$$

$$\lambda_{xy} = \lambda_{yx} = \sigma_H B^2 \rho^{-1} \sin I \quad [4]$$

$$\lambda_{yy} = \sigma_P B^2 \rho^{-1} \sin^2 I \quad [5]$$

In eqns [1]–[5], σ_P and σ_H are the Pedersen and Hall electrical conductivities; B is the magnitude of the magnetic field strength; I is the magnetic dip angle; ρ is the neutral gas density; U_I and V_I are the meridional and zonal ion drift velocities, respectively, and U_N and V_N are the zonal and meridional neutral wind components, respectively; and D_x and D_y are the ion drag terms to be added to the zonal and meridional momentum equations.

The coefficients (λ) are called the ion drag coefficients because the acceleration is a frictional force resulting from ion–neutral collisions. At thermospheric heights above 120 km, the ion–neutral collision frequency is much smaller than the ion gyro frequency (the frequency at which a charged particle spirals around a magnetic field line), and therefore ions are locked to geomagnetic field lines and can only move with them when driven by an electric field. Outside of the auroral zone, electric fields are small and, to a good approximation, the ions can be considered to simply corotate with the magnetic field of the Earth. A neutral wind flowing through the corotating ions experiences a collisional drag that becomes a maximum at the

peak of the ionospheric density layer near 300 km, and this drag provides the main resistance that balances the pressure force. During the day, when the ionization and hence ion drag are large, the wind flows across constant pressure surfaces from the subsolar high-temperature, high-pressure region to the antisolar low-temperature, low-pressure region of night. At night, the density of electrons and ions decays, and the thermospheric pressure forces drive much larger winds because of the reduced ion drag. In the daytime, F-region ionosphere near 300 km, the Pedersen ion drag characteristic inverse time constant can exceed 10^{-3} s^{-1} , which is much greater than the characteristic inverse time constant associated with the inertial and Coriolis terms in the momentum equation. During the day, the Hall ion drag coefficient peaks near 125 km, with a value that can be less than 10^{-4} s^{-1} , roughly comparable with the Coriolis parameter. This acceleration acts perpendicular to the wind vector and is usually opposite in sign to the Coriolis acceleration.

There is also a Joule heating term [6] that is added to the thermodynamic equation:

$$Q_J = \lambda_{xx}(U_I - U_N)^2 + \lambda_{xx}(V_I - V_N)^2 \quad [6]$$

At high magnetic latitudes, large electric fields associated with ion convection can drive large ion drift velocities ($\sim 100\text{--}2000 \text{ m s}^{-1}$) and thus there can be strong acceleration of the neutral winds in the direction of the ion convection pattern. However, the atmosphere will adjust hydrodynamically to this forcing in a complex fashion, either flowing with the drifting ions or building up a back pressure to resist the momentum forcing. Thus the electric fields are both a heat source for the neutral atmosphere, converting the ordered motion of the $\mathbf{E} \times \mathbf{B}$ charged particle drift into increased random thermal motion through collisional processes, and a momentum source ($\mathbf{J} \times \mathbf{B}$) because of the transfer of charged particle momentum to the neutral gas through collisional processes.

An important contrast between the atmosphere above 100 km and the lower atmosphere is that above 100 km the major gaseous atmospheric constituents are no longer uniformly mixed. Above about 160 km, each gas is separately in hydrostatic balance with its own local pressure scale height defined by R^*T/m , where m is the molecular mass of the gas, T is temperature, and R^* is the universal gas constant. This state is referred to as diffusive equilibrium. Between 100 and 160 km, the atmosphere is neither mixed nor in diffusive equilibrium. In particular, molecular oxygen (O_2) is dissociated into atomic oxygen (O) at these levels and the O must be transported downward to below 100 km before the atmosphere is dense enough for chemical recombination to occur. The photoproduction of O from O_2 occurs slowly, with maximum daytime timescales of several days or more, so it cannot drive large diurnal variations in composition. However, transport by large-scale atmospheric motions can be large enough to force significant departures from diffusive equilibrium on both diurnal and longer timescales. Indeed, the mean motion in the thermosphere is a circulation from the summer pole to the winter pole and this circulation transports atomic oxygen from the summer hemisphere, where it is dissociated, to the winter hemisphere, where the maximum densities occur. The simplest feedback of composition on the dynamics is the 'virtual

temperature' effect. That is, the sum of the partial pressures of all the gases taken together is in hydrostatic equilibrium, the vertical variation of geopotential height is proportional to T/M , where M is the mean molecular mass determined by local composition. Dynamic processes tend to eliminate horizontal pressure gradients, or, equivalently, gradients of geopotential height on a constant pressure surface. Hence, motions tend to weaken gradients in T/M with the result that large temperature anomalies are expected to accompany large anomalies in M . Thus, composition – through variations in the mean molecular mass – has an important influence on the pressure force in the thermosphere.

There are other heat and momentum sources acting in the thermosphere. These include heat conduction in the plasma from the magnetosphere to the ionosphere and eventually to the neutral atmosphere, dissipation by thermal conduction, molecular viscosity, and compositional damping of tidal, planetary, and gravity waves that are excited in the lower atmosphere and propagate upward; Joule heating by tide-driven ionospheric current systems; and all sorts of plasma energy and momentum interactions with the neutral atmosphere. The global impact of all of these interactive processes and the coupling of the small-scale phenomena with global scale processes is not completely known.

A schematic of the various physical and chemical processes acting in the thermosphere is shown in Figure 4. The magnetospheric electric field is generated by the interaction of the solar wind with the Earth's magnetic field, which produces a large potential drop across the polar cap. The resulting dawn-to-dusk electric field maps down to thermospheric heights, where it causes the ionospheric ions to drift in an $\mathbf{E} \times \mathbf{B}$ direction and, through collisions with the neutral gas, affects neutral dynamics. To model the thermosphere one needs to consider the complex interactions between neutral and plasma dynamics and electrodynamics. Thus, a general circulation model of the thermosphere must include the interactions with the ionosphere, energy and momentum inputs from the magnetosphere and lower atmosphere, and the electrodynamics interactions throughout the system.

Global Geomagnetic Quiet Time Circulation

The dominant effect driving the winds in the thermosphere is the diurnal variation in the absorption of solar EUV and UV radiation, which heats and expands the dayside thermosphere, creating day-to-night horizontal pressure gradients. In the optically thin upper thermosphere, above 200 km, the solar heating distribution is rather uniform over the dayside, whereas below 200 km the optical depth increases and the solar heating varies with the solar zenith angle. Fourier analysis of the horizontal heating rate distribution shows that the amplitude of wave numbers 1 and 2 are comparable and are generally in phase in the lower thermosphere, but in the upper thermosphere wave number 1 dominates over wave number 2 and they are out of phase. Thus, the upper thermosphere has a diurnal temperature, density, and wind response, whereas the lower thermosphere has a semidiurnal response, as shown in Figure 5.

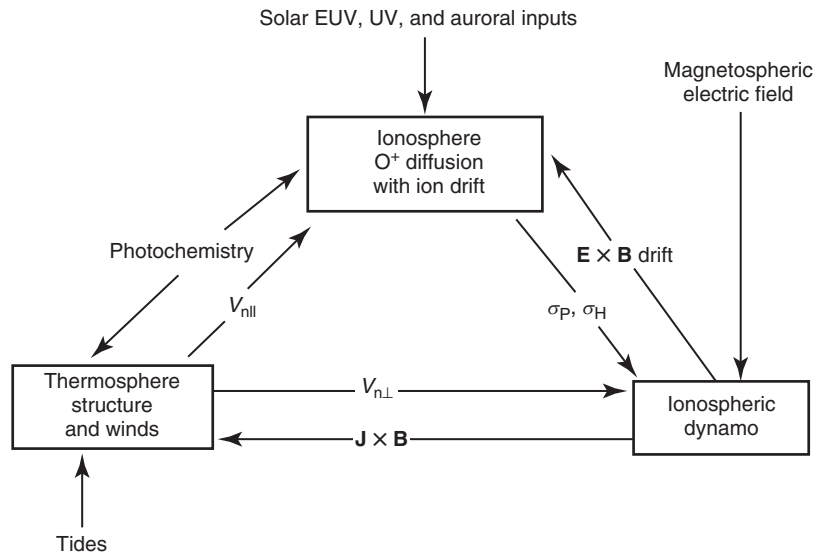


Figure 4 Schematic of various physical and chemical processes operating in the thermosphere and ionosphere system.

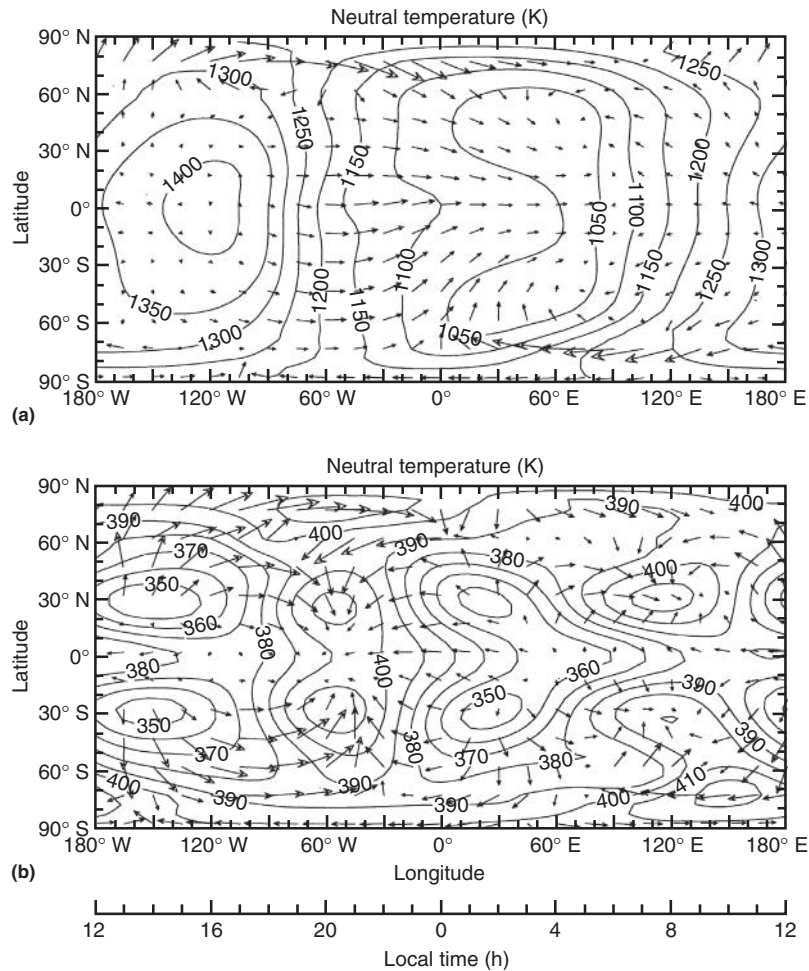


Figure 5 (a) A calculation of temperature in degrees Kelvin (contours) and wind vectors (arrows) in the upper thermosphere near 300 km altitude, for solar medium conditions ($F_{10.7} = 150$). The temperature and wind structure are shown for 0 UTC. The length of the maximum wind vectors represents 450 m s^{-1} . (b) A calculation of temperature and winds in the lower thermosphere near 120 km altitude for the same conditions and time. The length of the maximum wind vector represents 110 m s^{-1} .

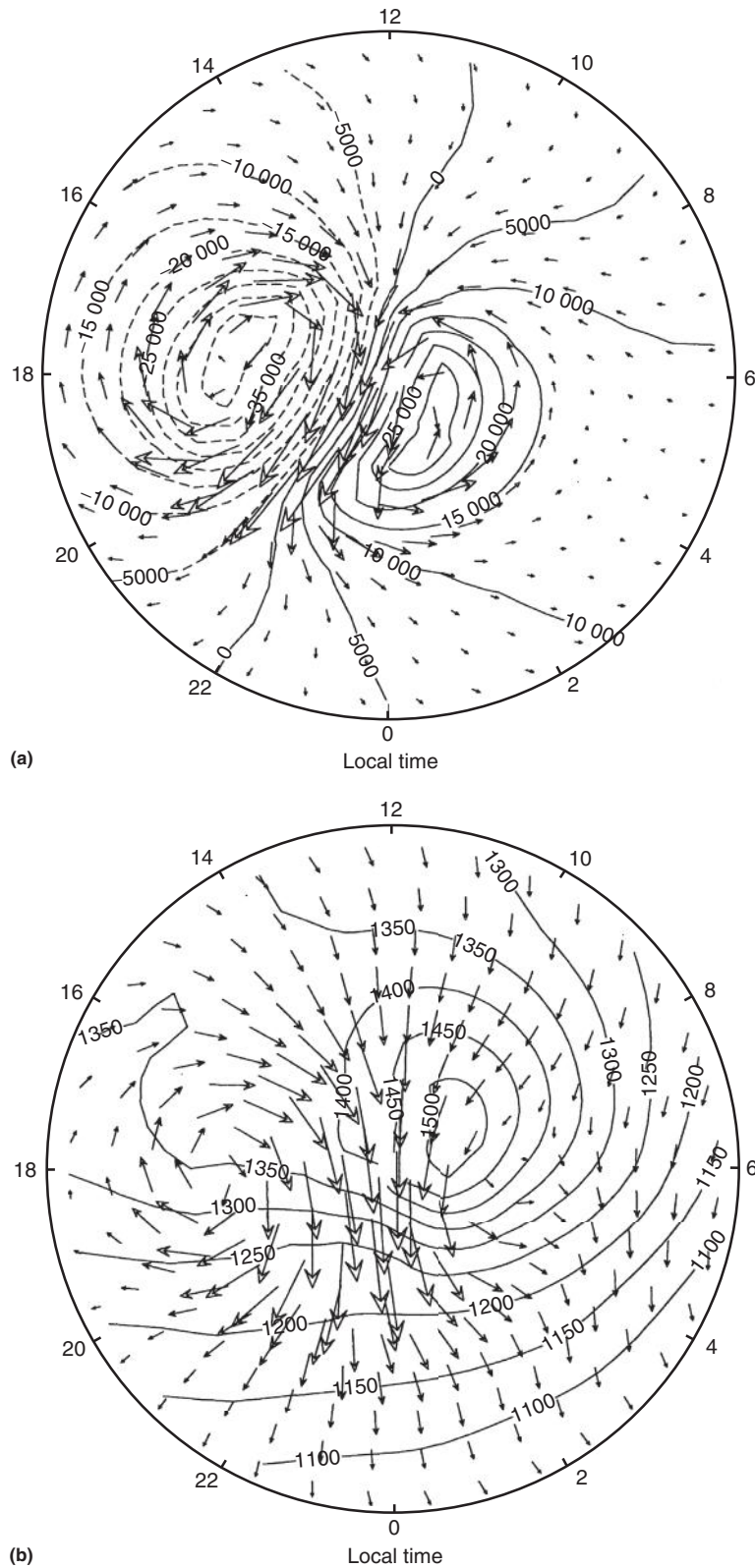


Figure 6 (a) A calculation of the ionospheric electric potential in volts over the Northern Hemisphere polar cap. The vectors show the $\mathbf{E} \times \mathbf{B}$ ion drift for moderate levels of auroral activity. The potential drop across the polar cap is 95 kV and the length of the maximum arrow represents a drift speed of 650 m s^{-1} . The electric potential and ion drift pattern are shown for 0 UTC. (b) A calculation of temperature in degrees Kelvin, (contours) and wind vectors (arrows) in the upper thermosphere near 300 km altitude responding to the ion drag forcing shown in (a). The length of the maximum wind vector represents 600 m s^{-1} .

Solar EUV radiation also ionizes the constituents of the thermosphere and produces the ionosphere. Fast photochemical equilibrium exists in the lower ionosphere and produces an electron density distribution that varies with the solar zenith angle. In the upper ionosphere, near 300 km, the ionospheric chemical time constants are much longer, and photochemical equilibrium gives way to complex transport processes that are driven by neutral winds, and by electric fields that are generated in the dynamo region of the thermosphere.

At low- to midlatitudes in the upper thermosphere, the wind flow is, to a first approximation, governed by a balance between the pressure force generated by solar heating and the ion drag and viscous force. The Coriolis force is much smaller, and thus the wind flow is from the high-temperature, high-density region on the dayside to the low-temperature, low-density region of night. This diurnal counter-gradient flow is characteristic of the thermosphere above about 200 km. The magnitude of the wind speed is less on the dayside ($50\text{--}100\text{ m s}^{-1}$) because of increased electron density, and hence ion drag, than at night ($100\text{--}300\text{ m s}^{-1}$), where a significant decrease in electron density occurs because of recombination, significantly reducing the ion drag.

Below 150 km, the ion drag force becomes comparable to the Coriolis force, and the semidiurnal component of thermospheric heating drives a complex semidiurnal oscillation. Below about 100 km, the ion drag force and viscous and compositional forces become small and the more familiar lower atmosphere dynamic equations govern the dynamics of the atmosphere.

Auroral Influences

The basic thermospheric circulation driven by solar heating, shown in [Figure 5](#), is for geomagnetic quiet times when the auroral inputs into the high-latitude thermosphere are small. However, the aurora is seldom quiet for long periods, and in general the interaction of the solar wind with the magnetosphere is highly variable. The aurora particle inputs into the high-latitude thermosphere are electron and proton precipitation, primarily in the auroral oval that surrounds the Earth's magnetic polar cap at about $60\text{--}70^\circ$ geomagnetic latitude in both the Northern and Southern Hemispheres. In addition, the electric fields caused by the interaction of the solar wind with the terrestrial magnetic field drive a two-cell ion drift circulation pattern that is Sun aligned and rotates with the geomagnetic pole around the geographic pole, as shown in [Figure 6](#) for moderate levels of geomagnetic activity. The ion drift imparts a momentum source that causes the neutral gas to follow, but generally lag, the ion drift motion, as also shown in [Figure 6](#). Joule heating causes the temperature to increase in the polar cap. The variability associated with auroral interaction produces large-scale changes in thermospheric circulation, launches large-scale waves that propagate globally, generates major changes in neutral composition, alters the ionospheric dynamic and electrodynamic structure, and produces chemical species such as nitric oxide that alter the radiative balance of the thermosphere. The thermosphere is, thus, in a constant state of agitation, depending upon the magnitude of the auroral inputs, which vary on daily, hourly, and sometimes even faster timescales.

Coupling with the Lower Atmosphere

In addition to solar and auroral forcing, the thermosphere is also affected by dynamics propagating up from the lower atmosphere. The main influence from below is due to atmospheric tides that grow considerably in amplitude as they propagate to high altitudes. The diurnal tide is observed to propagate up to about 110 km in the low-latitude thermosphere before being dissipated by molecular viscosity, thermal conductivity, and ion drag. The semidiurnal tide, with its much longer vertical wavelength, can extend to higher altitudes, reaching about 300 km. Other tidal components are either generated *in situ* or propagate up from the lower atmosphere. Planetary waves have also been observed in the lower thermosphere, especially in the winter high-latitude region, and can have significant effects on ionospheric variability.

Gravity waves generated in the lower atmosphere can propagate up to thermospheric heights and deposit their momentum in the upper mesosphere and lower thermosphere. These waves are filtered by the underlying mean winds, and interact with tides, planetary waves, and other waves generated by processes such as auroral heating. Gravity waves are also damped by thermospheric dissipative processes.

These disturbances or waves from the lower atmosphere greatly influence the lower thermosphere between 90 and 200 km. Above 200 km their influence becomes much smaller because the dynamics are controlled by the solar EUV forcing. Thus, the upper thermosphere has a strong diurnal variation, whereas the lower thermosphere below 200 km experiences semidiurnal variation strongly perturbed by processes propagating upward from the lower atmosphere.

See also: **Chemistry of the Atmosphere:** Ion Chemistry. **Dynamical Meteorology:** Atmospheric Tides; Overview; Primitive Equations. **Gravity Waves:** Overview. **Magnetosphere.** **Mesosphere:** Ionosphere. **Middle Atmosphere:** Planetary Waves. **Radiation Transfer in the Atmosphere:** Ultraviolet Radiation. **Solar System/Sun, Atmospheres, Evolution of Atmospheres:** Solar Winds.

Further Reading

- Banks, P.M., Kocharts, G., 1973. *Aeronomy*. Academic Press, New York.
 Brekke, A., 1997. *Physics of the Upper Polar Atmosphere*. Wiley-Praxis, New York.
 Chamberlain, J.W., Hunten, D.M., 1987. *Theory of Planetary Atmospheres*, second ed. Academic Press, New York.
 Chapman, S., Lindzen, R.S., 1970. *Atmospheric Tides*. Reidel, Dordrecht.
 Hines, C.O., Paghis, I., Hartz, T.R., Fejer, J.A., 1965. *Physics of the Earth's Upper Atmosphere*. Prentice-Hall, Englewood Cliffs.
 Kato, S., 1980. *Dynamics of the Upper Atmosphere*. Reidel, Tokyo.
 Kelley, M.C., 1989. *The Earth's Ionosphere*. Academic Press, New York.
 Rees, M.H., 1989. *Physics and Chemistry of the Upper Atmosphere*. Cambridge University Press, Cambridge.
 Schunk, R.W., Nagy, A.F., 2009. *Ionospheres: Physics, Plasma Physics, and Chemistry*, second ed. Cambridge University Press.
 Volland, H., 1995. *Handbook of Atmospheric Electrodynamics*. CRC Press, Boca Raton.

Second Edition

ENCYCLOPEDIA OF ATMOSPHERIC SCIENCES

Edited by
Gerald R. North **Editor-in-Chief**
John Pyle and Fuqing Zhang



6 VOLUME SET

ENCYCLOPEDIA OF **ATMOSPHERIC SCIENCES**

SECOND EDITION

VOLUME 6

This page intentionally left blank

ENCYCLOPEDIA OF **ATMOSPHERIC SCIENCES**

SECOND EDITION

EDITOR-IN-CHIEF

GERALD R NORTH

Texas A&M University, College Station, TX, USA

EDITORS

JOHN PYLE

Cambridge University, Cambridge, UK

FUQING ZHANG

Pennsylvania State University, University Park, PA, USA

VOLUME 6



Amsterdam • Boston • Heidelberg • London • New York • Oxford
Paris • San Diego • San Francisco • Singapore • Sydney • Tokyo
Academic Press is an imprint of Elsevier



Academic Press is an imprint of Elsevier
32 Jamestown Road, London NW1 7BY, UK
225 Wyman Street, Waltham, MA 02451, USA
525 B Street, Suite 1800, San Diego, CA 92101-4495, USA

Copyright © 2015 Elsevier Ltd. unless otherwise stated. All rights reserved.

No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means electronic, mechanical, photocopying, recording or otherwise without the prior written permission of the publisher

Permissions may be sought directly from Elsevier's Science & Technology Rights Department in Oxford, UK: phone (+44) (0) 1865 843830; fax (+44) (0) 1865 853333; email: permissions@elsevier.com. Alternatively you can submit your request online by visiting the Elsevier web site at <http://elsevier.com/locate/permissions>, and selecting *Obtaining permission to use Elsevier material*

Notice

No responsibility is assumed by the publisher for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions or ideas contained in the material herein, Because of rapid advances in the medical sciences, in particular, independent verification of diagnoses and drug dosages should be made

British Library Cataloguing in Publication Data

A catalog record for this book is available from the British Library

Library of Congress Catalog Number:

A catalog record for this book is available from the Library of Congress

ISBN (print): 978-0-12-382225-3

For information on all Elsevier publications
visit our website at store.elsevier.com

Printed and bound in the United Kingdom

15 16 17 18 19 10 9 8 7 6 5 4 3 2 1

		Working together to grow libraries in developing countries
www.elsevier.com • www.bookaid.org		

<i>Acquisitions Editor:</i> Simon Holt <i>Project Manager:</i> Michael Nicholls <i>Associate Project Manager:</i> Marise Willis <i>Designer:</i> Matthew Limbert

DEDICATION

This second edition of the *Encyclopedia of Atmospheric Sciences* is dedicated to the memory of James Holton who was editor-in-chief of the first edition. He was a great researcher and colleague inspiring an entire generation of atmospheric scientists.

This page intentionally left blank

CONTENTS

<i>List of Contributors</i>	xxvii
<i>Preface to the First Edition</i>	xxxix
<i>Preface to the Second Edition</i>	xli
<i>Editor Biographies</i>	xliii
<i>Guide to Using the Encyclopedia</i>	xlvi

VOLUME 1

BASIC ATMOSPHERIC STRUCTURE AND CONCEPTS	1
Beaufort Wind Scale	1
<i>L Hasse</i>	
Wind Chill	7
<i>M Bluestein</i>	
Standard Atmosphere	12
<i>W W Vaughan</i>	
AEROSOLS	17
Aerosol–Cloud Interactions and Their Radiative Forcing	17
<i>U Lohmann</i>	
Aerosol Physics and Chemistry	23
<i>M Kalberer</i>	
Climatology of Stratospheric Aerosols	32
<i>L W Thomason and J-P Vernier</i>	
Climatology of Tropospheric Aerosols	40
<i>N Bellouin and J Haywood</i>	
Dust	48
<i>I N Sokolik</i>	
Observations and Measurements	53
<i>P H McMurry</i>	
Role in Radiative Transfer	66
<i>G A Ban-Weiss, and W D Collins</i>	

Role in Climate Change <i>N Bellouin</i>	76
Soot <i>P Chylek, S G Jennings, and R Pinnick</i>	86
Agricultural Meteorology and Climatology <i>E S Takle</i>	92
ARCTIC AND ANTARCTIC	98
Antarctic Climate <i>J Turner</i>	98
Arctic Climate <i>M C Serreze</i>	107
Arctic Haze <i>L M Russell and G E Shaw</i>	116
AIR SEA INTERACTIONS	
Freshwater Flux <i>J Schulz</i>	122
Momentum, Heat, and Vapor Fluxes <i>P K Taylor</i>	129
Sea Surface Temperature <i>W J Emery</i>	136
Surface Waves <i>A Benilov</i>	144
AVIATION METEOROLOGY	153
Aircraft Emissions <i>R R Friedl</i>	153
Aircraft Icing <i>M K Politovich</i>	160
Aviation Weather Hazards <i>A J Bedard, Jr</i>	166
Clear Air Turbulence <i>G P Ellrod (Retired), J A Knox, P F Lester, and L J Ehernberger (Retired)</i>	177
BIOGEOCHEMICAL CYCLES	187
Sulfur Cycle <i>P Brimblecombe</i>	187
Bromine <i>R von Glasow and C Hughes</i>	194
Heavy Metals <i>T D Jickells and A R Baker</i>	201

Iodine	205
<i>L J Carpenter</i>	
BOUNDARY LAYER (ATMOSPHERIC) AND AIR POLLUTION	220
Overview	220
<i>P J Mason and D J Thomson</i>	
Air Pollution Meteorology	227
<i>X-M Hu</i>	
Coherent Structures	237
<i>F T M Nieuwstadt and J C R Hunt</i>	
Complex Terrain	242
<i>J J Finnigan</i>	
Convective Boundary Layer	250
<i>M A LeMone</i>	
Microclimate	258
<i>M W Rotach and P Calanca</i>	
Modeling and Parameterization	265
<i>A A M Holtslag</i>	
Observational Techniques <i>In Situ</i>	274
<i>E F Bradley</i>	
Observational Techniques: Remote	284
<i>W M Angevine and C J Senff</i>	
Ocean Mixed Layer	290
<i>L Kantha and C A Clayson</i>	
Stably Stratified Boundary Layer	299
<i>L Mahrt</i>	
Surface Layer	305
<i>G L Geernaert</i>	
Urban Heat Islands	310
<i>J C Luvall, D A Quattrochi, D L Rickman, and M G Estes, Jr</i>	
Diurnal Cycle	319
<i>A Betts</i>	
CHEMISTRY OF THE ATMOSPHERE	324
Chemical Kinetics	324
<i>R P Wayne</i>	
Ion Chemistry	333
<i>J L Fox</i>	
Isotopes, Stable	348
<i>C A M Brenninkmeijer</i>	
Laboratory Kinetics	356
<i>D J Donaldson and S N Wren</i>	

Methane <i>E Dlugokencky, and S Houweling</i>	363
Observations for Chemistry (<i>In Situ</i>): Ozone Sondes <i>H G J Smit</i>	372
Observations for Chemistry (<i>In Situ</i>): Particles <i>T Deshler</i>	379
Observations for Chemistry (<i>In Situ</i>): Water Vapor Sondes <i>J B Smith</i>	387
Observations for Chemistry (Remote Sensing): IR/FIR (Satellite, Balloon and Ground) <i>H Fischer and F Hase</i>	401
Observations for Chemistry (Remote Sensing): Lidar <i>G Vaughan</i>	411
Observations for Chemistry (Remote Sensing): Microwave <i>J Waters</i>	418
Principles of Chemical Change <i>R P Wayne</i>	429
Radioactivity: Cosmogenic Radionuclides <i>D Lal</i>	437
Volcanoes: Composition of Emissions <i>M T Coffey and J W Hannigan</i>	446
Tracers <i>K A Boering</i>	450

VOLUME 2

CLIMATE AND CLIMATE CHANGE	1
Overview <i>D L Hartmann</i>	1
Carbon Dioxide <i>C L Sabine and R A Feely</i>	10
Climate Feedbacks <i>A E Dessler and M D Zelinka</i>	18
Climate Prediction: Empirical and Numerical <i>S Hastenrath</i>	26
Climate Variability: Decadal to Centennial Variability <i>D G Martinson</i>	33
Climate Variability: Nonlinear and Random Effects <i>M Ghil</i>	38
Climate Variability: North Atlantic and Arctic Oscillation <i>J W Hurrell</i>	47
Climate Variability: Seasonal and Interannual Variability <i>D S Gutzler</i>	61

Energy Balance Climate Models <i>G R North and K-Y Kim</i>	69
Global Impacts of the Madden–Julian Oscillation <i>C Zhang</i>	73
Greenhouse Effect <i>G R North</i>	80
History of Scientific Work on Climate Change <i>S Weart</i>	87
Intergovernmental Panel on Climate Change <i>K E Trenberth</i>	90
Nuclear Winter <i>A Robock</i>	95
Radiative–Convective Equilibrium Climate Models <i>N O Renno and X Huang</i>	102
Volcanoes: Role in Climate <i>A Robock</i>	105
CLOUDS AND FOG	112
Cloud Modeling <i>W-K Tao and M Moncrieff</i>	112
Contrails <i>P Minnis</i>	121
Cloud Microphysics <i>D Lamb</i>	133
Classification of Clouds <i>A L Rangno (Retiree)</i>	141
Climatology <i>S Warren, R Eastman, and C J Hahn</i>	161
Measurement Techniques <i>In situ</i> <i>D Baumgardner, J-F Gayet, A Korolev, C Twohy, and J Fugal</i>	170
Fog <i>P J Croft and B Ward</i>	180
Noctilucent Clouds <i>G E Thomas</i>	189
Stratus and Stratocumulus <i>R Wood</i>	196
CRYOSPHERE	201
Glaciers, Topography, and Climate <i>A B G Bush and M P Bishop</i>	201
Permafrost <i>T E Osterkamp and C R Burn</i>	208

Sea Ice	217
<i>M C Serreze, F Fetterer, and W F Weeks (Retired)</i>	
Snow (Surface)	227
<i>M Sturm</i>	
DATA ASSIMILATION AND PREDICTABILITY	237
Data Assimilation	237
<i>A C Lorenc</i>	
Ensemble-Based Data Assimilation	241
<i>Z Meng and F Zhang</i>	
Ensemble Prediction	248
<i>R Buizza</i>	
Predictability and Chaos	258
<i>L A Smith</i>	
DYNAMICAL METEOROLOGY	265
Overview	265
<i>J R Holton</i>	
Acoustic Waves	272
<i>K E Gilbert</i>	
Atmospheric Tides	287
<i>J Oberheide, M E Hagan, A D Richmond, and J M Forbes</i>	
Balanced Flow	298
<i>M E McIntyre</i>	
Baroclinic Instability	304
<i>R Grotjahn</i>	
Coriolis Force	313
<i>D W Moore</i>	
Critical Layers	317
<i>P Haynes</i>	
Hamiltonian Dynamics	324
<i>T G Shepherd</i>	
Hydraulic Flow	332
<i>R B Smith</i>	
Inertial Instability	334
<i>J A Knox</i>	
Kelvin–Helmholtz Instability	343
<i>P G Drazin</i>	
Kelvin Waves	347
<i>B Wang</i>	
Kinematics	353
<i>D D Houghton</i>	

Laboratory Geophysical Fluid Dynamics <i>R L Pfeffer</i>	360
Lagrangian Dynamics <i>I Roulstone</i>	369
Potential Vorticity <i>M E McIntyre</i>	375
Primitive Equations <i>A A White and N Wood</i>	384
Quasigeostrophic Theory <i>H C Davies and H Wernli</i>	393
Rossby Waves <i>P B Rhines</i>	404
Solitary Waves <i>J P Boyd</i>	417
Static Stability <i>J A Young</i>	423
Stationary Waves (Orographic and Thermally Forced) <i>S Nigam and E DeWeaver</i>	431
Symmetric Stability <i>H B Bluestein</i>	446
Vorticity <i>J R Holton</i>	451
Wave-CISK <i>C S Bretherton</i>	455
Wave Mean-Flow Interaction <i>M Juckes</i>	458
Waves <i>J R Holton</i>	464

VOLUME 3

ELECTRICITY IN THE ATMOSPHERE	1
Global Electrical Circuit <i>E R Williams</i>	1
Ions in the Atmosphere <i>K L Aplin and R G Harrison</i>	9
Lightning <i>M B Baker</i>	14
Sprites <i>W A Lyons</i>	20
Forensic Meteorology <i>L E Branscome</i>	28

GENERAL CIRCULATION OF THE ATMOSPHERE	33
Overview	33
<i>J M Wallace, D W J Thompson, and P Beresford</i>	
Angular Momentum of the Atmosphere	43
<i>D A Salstein</i>	
Energy Cycle	51
<i>R Grotjahn</i>	
Weather Regimes and Multiple Equilibria	65
<i>F Molteni</i>	
Mean Characteristics	73
<i>R Grotjahn</i>	
Teleconnections	90
<i>S Nigam and S Baxter</i>	
 GLOBAL CHANGE	 110
Climate Record: Surface Temperature Trends	110
<i>P D Jones</i>	
Sea Level Change	121
<i>R S Nerem</i>	
Upper Atmospheric Change	128
<i>R G Roble</i>	
Biospheric Impacts and Feedbacks	132
<i>B A Hungate and G W Koch</i>	
 GRAVITY WAVES	 141
Overview	141
<i>D C Fritts</i>	
Buoyancy and Buoyancy Waves: Optical Observations	153
<i>M J Taylor and W R Pendleton, Jr</i>	
Buoyancy and Buoyancy Waves: Theory	160
<i>T J Dunkerton</i>	
Gravity Waves Excited by Jets and Fronts	164
<i>R Plougonven and F Zhang</i>	
Convectively Generated Gravity Waves	171
<i>T P Lane</i>	
 HYDROLOGY, FLOODS AND DROUGHTS	 180
Overview	180
<i>R C Bales</i>	
Deserts and Desertification	185
<i>V P Tchakerian</i>	
Drought	193
<i>S Quiring</i>	

Flooding	201
<i>C A Doswell III</i>	
Groundwater and Surface Water	209
<i>S Ge and S M Gorelick</i>	
Modeling and Prediction	217
<i>Z Yu</i>	
Palmer Drought Severity Index	224
<i>L Nkemdirim</i>	
Soil Moisture	232
<i>A Robock</i>	
LAND-ATMOSPHERE INTERACTIONS	240
Overview	240
<i>R E Dickinson</i>	
Canopy Processes	244
<i>P D Blanken</i>	
Trace Gas Exchange	256
<i>J N Cape and D Fowler</i>	
LIDAR	262
Atmospheric Sounding Introduction	262
<i>P S Argall and R Sica</i>	
Backscatter	270
<i>C M R Platt and R L Collins</i>	
Differential Absorption Lidar	277
<i>S Ismail and E V Browell</i>	
Doppler	289
<i>R M Hardesty</i>	
Raman	296
<i>D N Whiteman</i>	
Resonance	305
<i>C S Gardner and R L Collins</i>	
Magnetosphere	309
<i>G K Parks</i>	
MESOSCALE METEOROLOGY	316
Overview	316
<i>D J Parker</i>	
Cloud and Precipitation Bands	323
<i>R M Rauber and M Ramamurthy</i>	
Gust Fronts	331
<i>R Rotunno</i>	

Hail and Hailstorms	334
<i>C Knight, N Knight, and H E Brooks</i>	
Mesoscale Convective Systems	339
<i>A Laing</i>	
Microbursts	335
<i>R M Wakimoto</i>	
Severe Storms	361
<i>C A Doswell III</i>	
Waterspouts	369
<i>J H Golden</i>	
Bow Echoes and Derecho	384
<i>M L Weisman</i>	
Density Currents	395
<i>P G Baines</i>	
Convective Storms: Overview	401
<i>M L Weisman</i>	
MESOSPHERE	411
Atomic Species in the Mesopause Region	411
<i>M G Mlynczak and L A Hunt</i>	
Ionosphere	422
<i>M C Kelley</i>	
Metal Layers	430
<i>J M C Plane</i>	
Polar Summer Mesopause	436
<i>R H Varney and M C Kelley</i>	
VOLUME 4	
MIDDLE ATMOSPHERE	1
Planetary Waves	1
<i>A K Smith and J Perlwitz</i>	
Polar Vortex	12
<i>M R Schoeberl and P A Newman</i>	
Quasi-Biennial Oscillation	18
<i>T J Dunkerton, J A Anstey, and L J Gray</i>	
Semiannual Oscillation	26
<i>K Hamilton</i>	
Stratospheric Sudden Warmings	30
<i>A O'Neill, A J Charlton-Perez, and L M Polvani</i>	
Transport Circulation	41
<i>S E Strahan</i>	

Zonal Mean Climatology <i>P Braesicke</i>	50
MOUNTAIN METEOROLOGY	57
Overview <i>R B Smith</i>	57
Cold Air Damming <i>B A Colle</i>	62
Downslope Winds <i>D R Durran</i>	69
Katabatic Winds <i>T R Parish</i>	75
Land and Sea Breezes <i>R A Pielke, Sr</i>	80
Lee Vortices <i>C C Epifanio</i>	84
Lee Waves and Mountain Waves <i>D R Durran</i>	95
Orographic Effects: Lee Cyclogenesis <i>C Schär</i>	103
Valley Winds <i>D Zardi</i>	114
NUMERICAL MODELS	135
Chemistry Models <i>M P Chipperfield and S R Arnold</i>	135
Coupled Ocean-Atmosphere Models: Physical Processes <i>M Zhang</i>	144
General Circulation Models <i>C R Mechoso and A Arakawa</i>	153
Methods <i>J Thuburn</i>	161
Model Physics Parameterization <i>D J Stensrud, M C Coniglio, K H Knopfmeier, and A J Clark</i>	167
Parameter Estimation <i>A Aksoy</i>	181
Parameterization of Physical Processes: Clouds <i>R Forbes, C Jakob, and M Miller</i>	187
Parameterization of Physical Processes: Gravity Wave Fluxes <i>M J Alexander</i>	194
Parameterization of Physical Processes: Turbulence and Mixing <i>A Beljaars</i>	200

Spectral Models <i>F Baer</i>	212
Mesoscale Atmospheric Modeling <i>R A Pielke, Sr</i>	219
Cloud-System Resolving Modeling and Aerosols <i>W-K Tao and T Matsui</i>	222
Large-Eddy Simulation <i>C-H Moeng and P P Sullivan</i>	232
Regional Prediction Models <i>B W Golding</i>	241
Convective Storm Modeling <i>M D Parker</i>	246
OBSERVATIONS PLATFORMS	255
Balloons <i>J-P Pommereau</i>	255
Buoys <i>J M Hemsley</i>	264
Kites <i>B B Balsley</i>	268
Radiosondes <i>W F Dabberdt and H Turtiainen</i>	273
Rockets <i>M F Larsen</i>	285
OCEANOGRAPHIC TOPICS	290
General Processes <i>N C Wells</i>	290
Surface/Wind Driven Circulation <i>R X Huang</i>	301
Thermohaline Circulation <i>R X Huang</i>	315
Water Types and Water Masses <i>W J Emery</i>	329
OPTICS, ATMOSPHERIC	338
Optical Remote Sensing Instruments <i>G G Shepherd</i>	338
Airglow Instrumentation <i>M Conde</i>	346

OZONE DEPLETION AND RELATED TOPICS	353
Long-Term Ozone Changes <i>N R P Harris</i>	353
Ozone as a UV Filter <i>J E Frederick</i>	359
Ozone Depletion Potentials <i>D J Wuebbles</i>	364
Photochemistry of Ozone <i>G K Moortgat and A R Ravishankara</i>	370
Stratospheric Ozone Recovery <i>D J Hofmann and R Müller</i>	380
Surface Ozone Effects on Vegetation <i>M Ashmore</i>	389
Surface Ozone (Human Health) <i>M Lippmann</i>	397
PALEOCLIMATOLOGY	404
Ice Cores <i>E J Steig</i>	404
Varves <i>R Gilbert</i>	411
RADAR	415
Cloud Radar <i>T Uttal</i>	415
Incoherent Scatter Radar <i>M P Sulzer</i>	422
Mesosphere—Stratosphere—Troposphere and Stratosphere—Troposphere Radars and Wind Profilers <i>G Vaughan and D Hooper</i>	429
Meteor Radar <i>N J Mitchell</i>	438
Polarimetric Doppler Weather Radar <i>R J Doviak and R D Palmer</i>	444
Precipitation Radar <i>S E Yuter</i>	455
Synthetic Aperture Radar (Land Surface Applications) <i>R K Vincent</i>	470
VOLUME 5	
RADIATION TRANSFER IN THE ATMOSPHERE	1
Radiation, Solar <i>Q Fu</i>	1

Absorption and Thermal Emission <i>R M Goody and X Huang</i>	5
Cloud-Radiative Processes <i>Q Fu</i>	13
Non-local Thermodynamic Equilibrium <i>M López-Puertas and B Funke</i>	16
Scattering <i>M Mishchenko, L Travis, and A Lacis</i>	27
Ultraviolet Radiation <i>K Stamnes</i>	37
Ultraviolet, Surface <i>R McKenzie and S Madronich</i>	45
SATELLITES AND SATELLITE REMOTE SENSING	51
Aerosol Measurements <i>R A Kahn</i>	51
Earth's Radiation Budget <i>N G Loeb and B A Wielicki</i>	67
GPS Meteorology <i>S S Leroy</i>	77
Measuring Ozone from Space — TOMS and SBUV <i>R D McPeters and R S Stolarski</i>	87
Orbits <i>S Q Kidder</i>	95
Precipitation <i>G Liu</i>	107
Remote Sensing: Cloud Properties <i>P Yang and B A Baum</i>	116
Research <i>M D King</i>	128
Surface Wind and Stress <i>W T Liu</i>	138
Temperature Soundings <i>A Dudhia</i>	145
Water Vapor <i>J E Harries</i>	157
SOLAR SYSTEM/SUN, ATMOSPHERES, EVOLUTION OF ATMOSPHERES	163
Evolution of Earth's Atmosphere <i>Y L Yung, M L Wong, and E J Gaidos</i>	163
Planetary Atmospheres: Mars <i>R M Haberle</i>	168

Planetary Atmospheres: Venus <i>P J Gierasch and Y L Yung</i>	178
Solar Terrestrial Interactions: Climate Impact <i>J D Haigh</i>	183
Solar Winds <i>S T Suess and B T Tsurutani</i>	189
Meteors <i>P Jenniskens</i>	195
STATISTICAL METHODS	201
Data Analysis: Empirical Orthogonal Functions and Singular Vectors <i>C S Bretherton</i>	201
Data Analysis: Time Series Analysis <i>G R North</i>	205
STRATOSPHERIC CHEMISTRY TOPICS	211
Overview <i>J A Pyle</i>	211
Halogens <i>D Toohey</i>	215
Halogen Sources, Anthropogenic <i>A McCulloch and P M Midgley</i>	221
Halogen Sources, Natural (Methyl Bromide and Related Gases) <i>S Yvon-Lewis and J H Butler</i>	228
HO _x <i>T F Hanisco</i>	233
Hydrogen Budget <i>J E Harries</i>	238
Reactive Nitrogen (NO _x and NO _y) <i>Y Kondo</i>	242
Stratospheric Water Vapor <i>K H Rosenlof</i>	250
STRATOSPHERE/TROPOSPHERE EXCHANGE AND STRUCTURE	257
Global Aspects <i>J R Holton</i>	257
Local Processes <i>J F Lamarque and P Hess</i>	262
Tropopause <i>M Dameris</i>	269

SYNOPTIC METEOROLOGY	273
Anticyclones <i>S J Colucci</i>	273
Forecasting <i>D Mansfield</i>	280
Weather Maps <i>R Reynolds</i>	289
Cyclogenesis <i>G J Hakim</i>	299
Extratropical Cyclones <i>A Joly</i>	304
Fronts <i>D M (David) Schultz and W Blumen</i>	337
Fronts in the Lower Stratosphere <i>A L Lang</i>	344
Frontogenesis <i>D M (David) Schultz</i>	353
Jet Streaks <i>P Cunningham and D Keyser</i>	359
Lake-Effect Storms <i>P J Sousounis</i>	370
Polar Lows <i>I A Renfrew</i>	379
Thermal Low <i>R H Johnson</i>	386
THERMODYNAMICS	391
Humidity Variables <i>J A Curry</i>	391
Moist (Unsaturated) Air <i>J A Curry</i>	394
Saturated Adiabatic Processes <i>J A Curry</i>	398
Thermosphere <i>S C Solomon and R G Roble</i>	402
VOLUME 6	
TROPICAL CYCLONES AND HURRICANES	1
Overview and Theory <i>R A Tomas and P J Webster</i>	1

Hurricane Dynamics <i>Y Wang</i>	8
Hurricane Predictability <i>J A Sippel</i>	30
Hurricanes: Observation <i>F D Marks</i>	35
Tropical Cyclogenesis <i>Z Wang</i>	57
Tropical Cyclones and Climate Change <i>T R Knutson</i>	65
Tropical Cyclones in the Western North Pacific <i>J C L Chan</i>	77
Tropical Cyclones: Secondary Eyewall Formation <i>C-C Wu and Y-H Huang</i>	85
TROPICAL METEOROLOGY AND CLIMATE	91
El Niño and the Southern Oscillation: Observation <i>N Nicholls</i>	91
El Niño and the Southern Oscillation: Theory <i>P Chang and S E Zebiak</i>	97
Equatorial Waves <i>M C Wheeler and H Nguyen</i>	102
Hadley Circulation <i>J Lu and G A Vecchi</i>	113
Intertropical Convergence Zone <i>D E Waliser and X Jiang</i>	121
Intraseasonal Oscillation (Madden–Julian Oscillation) <i>R A Madden</i>	132
Madden–Julian Oscillation: Skeleton and Conceptual Models <i>A J Majda and S N Stechmann</i>	137
Monsoon: Overview <i>J Slingo</i>	146
Monsoon: Dynamical Theory <i>P J Webster and J Fasullo</i>	151
Monsoon: ENSO–Monsoon Interactions <i>K-M Lau</i>	165
Tropical Climates <i>S Hastenrath</i>	170
Walker Circulation <i>K-M Lau and S Yang</i>	177

TROPOSPHERIC CHEMISTRY AND COMPOSITION	182
Aerosols/Particles <i>J H Seinfeld</i>	182
Aliphatic Hydrocarbons <i>J Rudolph and O Stein</i>	188
Aromatic Hydrocarbons <i>I Barnes</i>	204
Biogenic Hydrocarbons <i>A Guenther</i>	214
Cloud Chemistry <i>P Herckes and J L Collett, Jr</i>	218
H ₂ <i>U Schmidt and T Wetter</i>	226
Hydroxyl Radical <i>K C Clemitshaw</i>	232
Mercury <i>J Munthe and J Sommar</i>	239
Oxidizing Capacity <i>D H Ehhalt, F Rohrer, and A Wahner</i>	243
Peroxyacetyl Nitrate <i>H B Singh</i>	251
Sulfur Chemistry, Organic <i>I Barnes</i>	255
Volatile Organic Compounds Overview: Anthropogenic <i>R G Derwent</i>	265
TURBULENCE AND MIXING	268
Overview <i>P Haynes</i>	268
Turbulence, Two Dimensional <i>P Bartello</i>	273
Turbulent Diffusion <i>A Venkatram and S Du</i>	277
WEATHER FORECASTING	287
Marine Meteorology <i>L Xie and B Liu</i>	287
Operational Meteorology <i>D R Novak</i>	293
Seasonal and Interannual Weather Prediction <i>J P Li and R Q Ding</i>	303
Severe Weather Forecasting <i>D J Stensrud, H E Brooks, and S J Weiss</i>	313

Wildfire Weather <i>J Coen</i>	323
Inadvertant Weather Modification <i>S A Changnon</i>	332
<i>Appendix 1: Physical Constants</i>	337
<i>Appendix 2: Units and their SI Equivalents</i>	339
<i>Appendix 3: Periodic Table of the Elements</i>	340
<i>Appendix 4: The Geologic Time Scale</i>	341
<i>Index</i>	343

This page intentionally left blank

LIST OF CONTRIBUTORS

A. Aksoy
University of Miami, Miami, FL, USA; and NOAA Hurricane Research Division, Miami, FL, USA

M.J. Alexander
NorthWest Research Associates (NWRA), Boulder, CO, USA

W.M. Angevine
CIRES, University of Colorado; and NOAA Earth System Research Laboratory, Boulder, CO, USA

J.A. Anstey
University of Oxford, Oxford, UK

K.L. Aplin
University of Oxford, Oxford, UK

A. Arakawa
University of California, Los Angeles, CA, USA

P.S. Argall
The University of Western Ontario, London, ON, Canada

S.R. Arnold
University of Leeds, Leeds, UK

M. Ashmore
University of York, York, UK

F. Baer
University of Maryland, College Park, MD, USA

P.G. Baines
University of Melbourne, Melbourne, VIC, Australia

A.R. Baker
University of East Anglia, Norwich, UK

M.B. Baker
University of Washington, Seattle, WA, USA

R.C. Bales
University of Arizona, Tucson, AZ, USA

B.B. Balsley
University of Colorado, Boulder, CO, USA

G.A. Ban-Weiss
Lawrence Berkeley National Laboratory, Berkeley, CA, USA; and University of Southern California, Los Angeles, CA, USA

I. Barnes
University of Wuppertal, Wuppertal, Germany

P. Bartello
McGill University, Montréal, QC, Canada

B.A. Baum
University of Wisconsin–Madison, Madison, WI, USA

D. Baumgardner
Universidad Nacional Autónoma de México, Mexico City, D.F., Mexico

S. Baxter
University of Maryland, College Park, MD, USA

A.J. Bedard, Jr.
National Oceanic and Atmospheric Administration, Boulder, CO, USA

A. Beljaars
European Centre for Medium-Range Weather Forecasts, Reading, England

N. Bellouin
University of Reading, Reading, UK

A. Benilov
Acute Solutions, Highlands, NJ, USA

P. Beresford
European Centre for Medium-Range Weather Forecasts, Reading, UK

A. Betts
Atmospheric Research, Pittsford, VT, USA

M.P. Bishop
Texas A&M University, College Station, TX, USA

P.D. Blanken
University of Colorado at Boulder, Boulder, CO, USA

H.B. Bluestein
University of Oklahoma, Norman, OK, USA

M. Bluestein
Indiana University – Purdue University, Indianapolis, IN, USA

W. Blumen[†]
University of Colorado Boulder, Boulder, CO, USA

K.A. Boering
University of California – Berkeley, Berkeley, CA, USA

J.P. Boyd
University of Michigan, Ann Arbor, MI, USA

E.F. Bradley
CSIRO Land and Water, Canberra, ACT, Australia

P. Braesicke
Karlsruhe Institute of Technology, Karlsruhe, Germany

L.E. Branscome
Climatological Consulting Corporation, FL, USA

C.A.M. Brenninkmeijer
Max Planck Institute for Chemistry, Mainz, Germany

C.S. Bretherton
University of Washington, Seattle, WA, USA

P. Brimblecombe
University of East Anglia, Norwich, UK

H.E. Brooks
National Oceanic and Atmospheric Administration, Norman, OK, USA

E.V. Browell
STARS II Affiliate, NASA Langley Research Center, Hampton, VA, USA

R. Buizza
ECMWF, Reading, UK

C.R. Burn
Carleton University, Ottawa, ON, Canada

A.B.G. Bush
University of Alberta, Edmonton, AB, Canada

J.H. Butler
NOAA Earth System Research Laboratory, Boulder, CO, USA

P. Calanca
Agroscope Reckenholz-Taenikon, Zurich, Switzerland

J.N. Cape
Edinburgh Research Station, Midlothian, UK

L.J. Carpenter
University of York, York, UK

J.C.L. Chan
City University of Hong Kong, Hong Kong

P. Chang
Texas A&M University, College Station, TX, USA

S.A. Changnon
University of Illinois, IL, USA

A.J. Charlton-Perez
University of Reading, Earley Gate, Reading, UK

M.P. Chipperfield
University of Leeds, Leeds, UK

P. Chylek
Dalhousie University, NS, Canada

A.J. Clark
University of Oklahoma and National Oceanic and Atmospheric Administration, Norman, OK, USA

C.A. Clayson
Woods Hole Oceanographic Institution, Woods Hole, MA, USA

K.C. Clemitshaw
Imperial College of Science, Technology, and Medicine, Ascot, UK

J. Coen
National Center for Atmospheric Research, Boulder, CO, USA

M.T. Coffey
National Center for Atmospheric Research, Boulder, CO, USA

B.A. Colle
Stony Brook University – SUNY, Stony Brook, NY, USA

J.L. Collett, Jr.
Colorado State University, Fort Collins, CO, USA

R.L. Collins
University of Alaska Fairbanks, Fairbanks, AK, USA

W.D. Collins
Lawrence Berkeley National Laboratory, Berkeley, CA, USA

S.J. Colucci
Cornell University, Ithaca, NY, USA

M. Conde
University of Alaska Fairbanks, Fairbanks, AK, USA

M.C. Coniglio
National Oceanic and Atmospheric Administration, Norman, OK, USA

[†] Deceased.

- P.J. Croft
Kean University, Union, NJ, USA
- P. Cunningham
Florida State University, Tallahassee, FL, USA
- J.A. Curry
Georgia Institute of Technology, Atlanta, GA, USA
- W.F. Dabberdt
Vaisala Company, Boulder, CO, USA
- M. Dameris
Deutsches Zentrum für Luft- und Raumfahrt, Institut für Physik der Atmosphäre, Oberpfaffenhofen, Wessling, Germany
- H.C. Davies
Institute for Atmospheric and Climate Science, Swiss Federal Institute of Technology (ETH), Zürich, Switzerland
- R.G. Derwent
rdscientific, Newbury, UK
- T. Deshler
University of Wyoming, Laramie, WY, USA
- A.E. Dessler
Texas A&M University, College Station, TX, USA
- E. DeWeaver
University of Wisconsin, Madison, WI, USA
- R.E. Dickinson
University of Texas at Austin, Austin, TX, USA
- R.Q. Ding
Institute of Atmospheric Physics, Chinese Academy of Sciences, Beijing, China
- E. Dlugokencky
NOAA Earth System Research Laboratory, Boulder, CO, USA
- D.J. Donaldson
University of Toronto, Toronto, ON, Canada
- C.A. Doswell, III
University of Oklahoma, Norman, OK, USA
- R.J. Doviak
National Severe Storms Laboratory, Norman, OK, USA
- P.G. Drazin[†]
University of Bath, England, UK
- S. Du
California Air Resources Board, Sacramento, CA, USA
- A. Dudhia
University of Oxford, Oxford, UK
- T.J. Dunkerton
Northwest Research Associates, Bellevue, WA, USA
- D.R. Durran
University of Washington, Seattle, WA, USA
- R. Eastman
University of Washington, Seattle, WA, USA
- L.J. Ehernberger
National Aeronautics and Space Administration, Dryden Flight Research Center, Edwards, CA, USA
- D.H. Ehhalt
Forschungszentrum Jülich, Jülich, Germany
- G.P. Ellrod
National Oceanographic and Atmospheric Administration/National Environmental Satellite, Data, and Information Service, Granby, CT, USA
- W.J. Emery
University of Colorado, Boulder, CO, USA
- C.C. Epifanio
Texas A&M University, College Station, TX, USA
- M.G. Estes
Universities Space Research Association, Huntsville, AL, USA
- J. Fasullo
University of Colorado – Boulder, Boulder, CO, USA
- R.A. Feely
NOAA Pacific Marine Environmental Laboratory, Seattle, WA, USA
- F. Fetterer
University of Colorado, Boulder, CO, USA
- J.J. Finnigan
CSIRO Atmospheric Research, Black Mountain, ACT, Australia
- H. Fischer
Institute of Meteorology and Climate Research (IMK), Karlsruhe Institute of Technology, Karlsruhe, Germany
- J.M. Forbes
University of Colorado, Boulder, CO, USA
- R. Forbes
European Centre for Medium-Range Weather Forecasts, Reading, UK
- D. Fowler
Edinburgh Research Station, Midlothian, UK

[†] Deceased.

J.L. Fox
Wright State University, Dayton, OH, USA

J.E. Frederick
The University of Chicago, Chicago, IL, USA

R.R. Friedl
California Institute of Technology, Pasadena, CA, USA

D.C. Fritts
GATS Inc., Boulder, CO, USA

Q. Fu
University of Washington, Seattle, WA, USA

J. Fugal
Max Planck Institute of Chemistry, Mainz, Germany

B. Funke
Instituto de Astrofísica de Andalucía, CSIC, Granada, Spain

E.J. Gaidos
University of Hawaii at Manoa, Honolulu, HI, USA

C.S. Gardner
University of Illinois at Urbana-Champaign, Urbana, IL, USA

J.-F. Gayet
Université Blaise Pascal, Clermont Ferrand, France

S. Ge
University of Colorado, Boulder, CO, USA

G.L. Geernaert
US Department of Energy, Washington, DC, USA

M. Ghil
Ecole Normale Supérieure, Paris, France; and University of California, Los Angeles, CA, USA

P.J. Gierasch
Cornell University, Ithaca, NY, USA

K.E. Gilbert
University of Mississippi, University, MS, USA

R. Gilbert
Queen's University, Kingston, ON, Canada

J.H. Golden
Forecast Systems Laboratory, NOAA, Boulder, CO, USA

B.W. Golding
Met Office, Exeter, UK

R.M. Goody
Harvard University (Emeritus), Cambridge, MA, USA

S.M. Gorelick
Stanford University, Stanford, CA, USA

L.J. Gray
University of Oxford, Oxford, UK

R. Grotjahn
University of California, Davis, CA, USA

A. Guenther
Pacific Northwest National Laboratory, Richland, WA, USA

D.S. Gutzler
University of New Mexico, Albuquerque, NM, USA

R.M. Haberle
NASA/Ames Research Center, Moffett Field, Mountain View, CA, USA

M.E. Hagan
National Center for Atmospheric Research, Boulder, CO, USA

C.J. Hahn
University of Arizona, Tucson, AZ, USA

J.D. Haigh
Blackett Laboratory, Imperial College London, London, UK

G.J. Hakim
University of Washington, Seattle, WA, USA

K. Hamilton
University of Hawaii, Honolulu, HI, USA

T.F. Hanisco
Harvard University, Cambridge, MA, USA

J.W. Hannigan
National Center for Atmospheric Research, Boulder, CO, USA

R.M. Hardesty
NOAA Environmental Technology Laboratory, Boulder, CO, USA

J.E. Harries
Imperial College of Science, Technology and Medicine, Blackett Laboratory, London, UK

N.R.P. Harris
University of Cambridge, Cambridge, UK

R.G. Harrison
The University of Reading, Reading, UK

D.L. Hartmann
University of Washington, Seattle, WA, USA

F. Hase
Institute of Meteorology and Climate Research (IMK), Karlsruhe Institute of Technology, Karlsruhe, Germany

- L. Hasse
Universität Kiel, Kiel, Germany
- S. Hastenrath
University of Wisconsin, Madison, WI, USA
- P. Haynes
University of Cambridge, Cambridge, UK
- J. Haywood
Met Office, Exeter, UK
- J.M. Hemsley
National Data Buoy Center, Stennis Space Center, MS, USA
- P. Herckes
Arizona State University, Tempe, AZ, USA
- P. Hess
National Center for Atmospheric Research, Boulder, CO, USA
- D.J. Hofmann[†]
NOAA Climate Monitoring and Diagnostics Laboratory, Boulder, CO, USA
- J.R. Holton
University of Washington, Seattle, WA, USA
- A.A.M. Holtslag
Wageningen University, Wageningen, The Netherlands
- D. Hooper
Science & Technology Facilities Council (STFC), Didcot, UK
- D.D. Houghton
University of Wisconsin-Madison, Madison, WI, USA
- S. Houweling
SRON Netherlands Institute for Space Research, Utrecht, The Netherlands
- X.-M. Hu
University of Oklahoma, Norman, OK, USA
- R.X. Huang
Woods Hole Oceanographic Institution, Woods Hole, MA, USA
- X. Huang
University of Michigan, Ann Arbor, MI, USA
- Y.-H. Huang
National Taiwan University, Taipei, Taiwan
- C. Hughes
University of York, York, UK
- B.A. Hungate
Northern Arizona University, Flagstaff, AZ, USA
- J.C.R. Hunt
University College London, London, UK
- L.A. Hunt
Science Systems and Applications Incorporated, Hampton, VA, USA
- J.W. Hurrell
National Center for Atmospheric Research, Boulder, CO, USA
- S. Ismail
Science Directorate, NASA Langley Research Center, Hampton, VA, USA
- C. Jakob
Monash University, VIC, Australia
- S.G. Jennings
National University of Ireland, Galway, Ireland
- P. Jenniskens
SETI Institute, Moffett Field, CA, USA
- X. Jiang
University of California, Los Angeles, CA, USA
- T.D. Jickells
University of East Anglia, Norwich, UK
- R.H. Johnson
Colorado State University, Fort Collins, CO, USA
- A. Joly
Centre National de Recherches Météorologiques – Groupe d'étude de l'Atmosphère Météorologique, Météo-France and CNRS, Toulouse, France
- P.D. Jones
Climatic Research Unit, University of East Anglia, Norwich, UK
- M. Jukes
University of Oxford, Oxford, UK
- R.A. Kahn
NASA Goddard Space Flight Center, Greenbelt, MD, USA
- M. Kalberer
University of Cambridge, Cambridge, UK
- L. Kantha
University of Colorado, Boulder, CO, USA
- M.C. Kelley
Cornell University, Ithaca, NY, USA

[†] Deceased.

D. Keyser
*University at Albany, State University of New York,
Albany, NY, USA*

S.Q. Kidder
*Colorado State University, Fort Collins,
CO, USA*

K.-Y. Kim
Seoul National University, Seoul, Korea

M.D. King
University of Colorado, Boulder, CO, USA

C. Knight
*National Center for Atmospheric Research, Boulder,
CO, USA*

N. Knight
*National Center for Atmospheric Research, Boulder,
CO, USA*

K.H. Knopfmeier
*University of Oklahoma; and National Oceanic and
Atmospheric Administration, Norman, OK, USA*

J.A. Knox
University of Georgia, Athens, GA, USA

T.R. Knutson
*NOAA Geophysical Fluid Dynamics Laboratory,
Princeton, NJ, USA*

G.W. Koch
Northern Arizona University, Flagstaff, AZ, USA

Y. Kondo
The University of Tokyo, Tokyo, Japan

A. Korolev
*Meteorological Service of Canada, Toronto, ON,
Canada*

A. Lacis
*Goddard Institute for Space Studies, New York,
NY, USA*

A. Laing
*National Center for Atmospheric Research, Boulder,
CO, USA*

D. Lal
*Scripps Institution of Oceanography, La Jolla,
CA, USA*

J.F. Lamarque
*National Center for Atmospheric Research, Boulder,
CO, USA*

D. Lamb
*The Pennsylvania State University, University Park,
PA, USA*

T.P. Lane
*The University of Melbourne, Melbourne, VIC,
Australia*

A.L. Lang
*University of Albany – State University of New York,
Albany, NY, USA*

M.F. Larsen
Clemson University, Clemson, SC, USA

K.-M. Lau
*NASA/Goddard Space Flight Center, Greenbelt, MD,
USA*

M.A. LeMone
*National Center for Atmospheric Research, Boulder,
CO, USA*

S.S. Leroy
*Harvard School of Engineering and Applied Sciences,
Cambridge, MA, USA*

P.F. Lester
San Jose State University, San Jose, CA, USA

J.P. Li
*Institute of Atmospheric Physics, Chinese Academy of
Sciences, Beijing, China*

M. Lippmann
New York University, Tuxedo, NY, USA

B. Liu
North Carolina State University, Raleigh, NC, USA

G. Liu
Florida State University, Tallahassee, FL, USA

W.T. Liu
*California Institute of Technology, Pasadena,
CA, USA*

N.G. Loeb
NASA Langley Research Center, Hampton, VA, USA

U. Lohmann
ETH Zurich, Zürich, Switzerland

M. López-Puertas
*Instituto de Astrofísica de Andalucía, CSIC, Granada,
Spain*

A.C. Lorenc
The Met Office, Bracknell, Berkshire, UK

J. Lu
*Pacific Northwest National Laboratory, Richland, WA,
USA*

J.C. Luvall
*National Space Science and Technology Center,
Marshall Space Flight Center, Huntsville, AL, USA*

- W.A. Lyons
FMA Research Inc., Fort Collins, CO, USA
- R.A. Madden
National Center for Atmospheric Research, Boulder, CO, USA
- S. Madronich
National Center for Atmospheric Research, Boulder, CO, USA
- L. Mahrt
Oregon State University, Corvallis, OR, USA
- A.J. Majda
New York University, New York, NY, USA
- D. Mansfield
National Meteorological Center, Bracknell, UK
- F.D. Marks
Hurricane Research Division, Miami, FL, USA
- D.G. Martinson
Columbia University, Palisades, NY, USA
- P.J. Mason
Met Office, Bracknell, UK
- T. Matsui
NASA/Goddard Space Flight Center, Greenbelt, MD, USA; and University of Maryland, College Park, MD, USA
- A. McCulloch
University of Bristol, Bristol, UK
- M.E. McIntyre
University of Cambridge, Cambridge, UK
- R. McKenzie
National Institute for Water and Atmospheric Research, Lauder, Central Otago, New Zealand
- P.H. McMurtry
University of Minnesota, Minneapolis, MN, USA
- R.D. McPeters
NASA Goddard Space Flight Center, Greenbelt, MD, USA
- C.R. Mechoso
University of California, Los Angeles, CA, USA
- Z. Meng
Peking University, Beijing, China
- P.M. Midgley
M & D Consulting, Leinfelden Metter, Germany
- M. Miller
European Centre for Medium-Range Weather Forecasts, Reading, UK
- P. Minnis
Science Directorate, NASA Langley Research Center, Hampton, VA, USA
- M. Mishchenko
Goddard Institute for Space Studies, New York, NY, USA
- N.J. Mitchell
The University of Bath, Bath, UK
- M.G. Mlynarczyk
NASA Langley Research Center, Hampton, VA, USA
- C.-H. Moeng
National Center for Atmospheric Research, Boulder, CO, USA
- F. Molteni
Abdus Salam International Centre for Theoretical Physics, Trieste, Italy
- M. Moncrieff
National Center for Atmospheric Research, Boulder, CO, USA
- D.W. Moore
Pacific Marine Environmental Laboratory, Seattle, WA, USA
- G.K. Moortgat
Max-Planck-Institute for Chemistry, Mainz, Germany
- R. Müller
Institute for Energy and Climate Research (IEK-7), Forschungszentrum Jülich, Jülich, Germany
- J. Munthe
IVL Swedish Environmental Research Institute, Göteborg, Sweden
- R.S. Nerem
University of Colorado, Boulder, CO, USA
- P.A. Newman
NASA Goddard, Space Flight Center, Greenbelt, MD, USA
- H. Nguyen
Centre for Australian Weather and Climate Research, Bureau of Meteorology, Melbourne, VIC, Australia
- N. Nicholls
Bureau of Meteorology Research Centre, Melbourne, VIC, Australia
- F.T.M. Nieuwstadt
Delft University of Technology, Delft, The Netherlands
- S. Nigam
University of Maryland, College Park, MD, USA

L. Nkemdirim
University of Calgary, Calgary, AB, Canada

G.R. North
Texas A&M University, College Station, TX, USA

D.R. Novak
Weather Prediction Center, College Park, MD, USA

A. O'Neill
University of Reading, Earley Gate, Reading, UK

J. Oberheide
Clemson University, Clemson, SC, USA

T.E. Osterkamp
University of Alaska, Fairbanks, AK, USA

R.D. Palmer
University of Oklahoma, Oklahoma, OK, USA

T.R. Parish
University of Wyoming, Laramie, WY, USA

D.J. Parker
University of Leeds, Leeds, UK

M.D. Parker
North Carolina State University, Raleigh, NC, USA

G.K. Parks
University of Washington, Seattle, WA, USA

W.R. Pendleton
Utah State University, Logan, UT, USA

J. Perlwitz
University of Colorado, Boulder, CO, USA

R.L. Pfeffer
Florida State University, Tallahassee, FL, USA

R.A. Pielke, Sr.
University of Colorado at Boulder, CO, USA

R. Pinnick
US Army Research Laboratory, Adelphi, MD, USA

J.M.C. Plane
University of Leeds, Leeds, UK

C.M.R. Platt
Colorado State University, Fort Collins, CO, USA

R. Plougonven
Ecole Polytechnique, Palaiseau, France

M.K. Politovich
National Center for Atmospheric Research, Boulder, CO, USA

L.M. Polvani
Columbia University, New York, NY, USA

J.-P. Pommereau
LATMOS, CNRS, Guyancourt, France

J.A. Pyle
University of Cambridge, Cambridge, UK

D.A. Quattrochi
National Space Science and Technology Center, Marshall Space Flight Center, Huntsville, AL, USA

S. Quiring
Texas A&M University, College Station, TX, USA

M. Ramamurthy
University Corporation for Atmospheric Research, Boulder, CO, USA

A.L. Rangno (Retiree)
University of Washington, Seattle, WA, USA

R.M. Rauber
University of Illinois at Urbana-Champaign, Urbana, IL, USA

A.R. Ravishankara
Colorado State University, Fort Collins, CO, USA

I.A. Renfrew
University of East Anglia, Norwich, UK

N.O. Renno
University of Michigan, Ann Arbor, MI, USA

R. Reynolds
University of Reading, Reading, UK

P.B. Rhines
University of Washington, Seattle, WA, USA

A.D. Richmond
National Center for Atmospheric Research, Boulder, CO, USA

D.L. Rickman
National Space Science and Technology Center, Marshall Space Flight Center, Huntsville, AL, USA

R.G. Roble
National Center for Atmospheric Research, Boulder, CO, USA

A. Robock
Rutgers University, New Brunswick, NJ, USA

F. Rohrer
Forschungszentrum Jülich, Jülich, Germany

K.H. Rosenlof
Earth System Research Laboratory, National Oceanic and Atmospheric Administration, Boulder, CO, USA

- M.W. Rotach
University of Innsbruck, Innsbruck, Austria
- R. Rotunno
National Center for Atmospheric Research, Boulder, CO, USA
- I. Roulstone
University of Surrey, Guildford, UK
- J. Rudolph
York University, Toronto, ON, Canada
- L.M. Russell
Scripps Institution of Oceanography, University of California, San Diego, La Jolla, CA, USA
- C.L. Sabine
NOAA Pacific Marine Environmental Laboratory, Seattle, WA, USA
- D.A. Salstein
Atmospheric and Environmental Research, Inc., Lexington, MA, USA
- C. Schär
Atmospheric and Climatic Science ETH, Zürich, Switzerland
- U. Schmidt
Johann Wolfgang Goethe-University, Frankfurt am Main, Institut für Meteorologie und Geophysik, Frankfurt am Main, Germany
- M.R. Schoeberl
Science and Technology Corporation, Lanham, MD, USA
- D.M. (David) Schultz
University of Manchester, Manchester, UK
- J. Schulz
Meteorological Institute, University of Bonn, Bonn, Germany
- J.H. Seinfeld
California Institute of Technology, Pasadena, CA, USA
- C.J. Senff
CIRES, University of Colorado; and NOAA Earth System Research Laboratory, Boulder, CO, USA
- M.C. Serreze
University of Colorado, Boulder, CO, USA
- G.E. Shaw
Geophysical Institute, University of Alaska, Fairbanks, AK, USA
- G.G. Shepherd
York University, Toronto, ON, Canada
- T.G. Shepherd
University of Toronto, Toronto, ON, Canada
- R. Sica
The University of Western Ontario, London, ON, Canada
- H.B. Singh
NASA Ames Research Center, Mountain View, CA, USA
- J.A. Sippel
National Aeronautics and Space Administration (NASA), Greenbelt, MD, USA
- J. Slingo
University of Reading, Reading, UK
- H.G.J. Smit
Research Centre Jülich, Jülich, Germany
- A.K. Smith
National Center for Atmospheric Research, Boulder, CO, USA
- J.B. Smith
Harvard University, Cambridge, MA, USA
- L.A. Smith
London School of Economics, Centre for the Analysis of Time Series, London, UK
- R.B. Smith
Yale University, New Haven, CT, USA
- I.N. Sokolik
Georgia Institute of Technology, Atlanta, GA, USA
- S.C. Solomon
National Center for Atmospheric Research, Boulder, CO, USA
- J. Sommar
Göteborg University, Göteborg, Sweden
- P.J. Sousounis
AIR Worldwide, Boston, MA, USA
- K. Stamnes
Stevens Institute of Technology, Hoboken, NJ, USA
- S.N. Stechmann
University of Wisconsin–Madison, Madison, WI, USA
- E.J. Steig
University of Washington, Seattle, WA, USA
- O. Stein
IEK 8: Troposphere, Research Center Juelich, Juelich, Germany

D.J. Stensrud
*National Oceanic and Atmospheric Administration,
Norman, OK, USA*

R.S. Stolarski
Johns Hopkins University, Baltimore, MD, USA

S.E. Strahan
*Universities Space Research Association, NASA Goddard
Space Flight Center, Greenbelt, MD, USA*

M. Sturm
*US Army Cold Regions Research & Engineering
Laboratory-Alaska, Fort Wainwright, AL, USA*

S.T. Suess
*NASA Marshall Space Flight Center, Huntsville, AL,
USA*

P.P. Sullivan
*National Center for Atmospheric Research, Boulder,
CO, USA*

M.P. Sulzer
Arecibo Observatory, Arecibo, PR, USA

E.S. Takle
Iowa State University, Ames, IA, USA

W.-K. Tao
*NASA/Goddard Space Flight Center, Greenbelt, MD,
USA*

M.J. Taylor
Utah State University, Logan, UT, USA

P.K. Taylor
*Southampton Oceanography Centre, Southampton,
UK*

V.P. Tchakerian
Texas A&M University, College Station, TX, USA

G.E. Thomas
University of Colorado, Boulder, CO, USA

L.W. Thomason
NASA Langley Research Center, Hampton, VA, USA

D.W.J. Thompson
Colorado State University, Fort Collins, CO, USA

D.J. Thomson
Met Office, Bracknell, UK

J. Thuburn
University of Exeter, Exeter, UK

R.A. Tomas
University of Colorado – Boulder, Boulder, CO, USA

D. Toohey
University of Colorado Boulder, Boulder, CO, USA

L. Travis
*Goddard Institute for Space Studies, New York,
NY, USA*

K.E. Trenberth
*National Center for Atmospheric Research, Boulder,
CO, USA*

B.T. Tsurutani
Jet Propulsion Laboratory, Pasadena, CA, USA

J. Turner
British Antarctic Survey, Cambridge, UK

H. Turtiainen
Vaisala Company, Helsinki, Finland

C. Twohy
Oregon State University, Corvallis, OR, USA

T. Uttal
NOAA, Boulder, CO, USA

R.H. Varney
Cornell University, Ithaca, NY, USA

G. Vaughan
University of Manchester, Manchester, UK

W.W. Vaughan
*University of Alabama in Huntsville, Huntsville,
AL, USA*

G.A. Vecchi
GFDL/NOAA, Princeton, NJ, USA

A. Venkatram
*University of California – Riverside, Riverside,
CA, USA*

J.-P. Vernier
*Science Systems and Applications, Inc., Hampton,
VA, USA*

R.K. Vincent
*Bowling Green State University, Bowling Green,
OH, USA*

R. von Glasow
University of East Anglia, Norwich, UK

A. Wahner
Forschungszentrum Jülich, Jülich, Germany

R.M. Wakimoto
*National Center for Atmospheric Research, Boulder,
CO, USA*

D.E. Waliser
California Institute of Technology, Pasadena, CA, USA

J.M. Wallace
University of Washington, Seattle, WA, USA

- B. Wang
University of Hawaii, Honolulu, HI, USA
- Y. Wang
University of Hawaii at Manoa, Honolulu, HI, USA
- Z. Wang
University of Illinois at Urbana-Champaign, Urbana, IL, USA
- B. Ward
Public Works and Natural Resources, Longmont, CO, USA
- S. Warren
University of Washington, Seattle, WA, USA
- J. Waters
California Institute of Technology, Jet Propulsion Laboratory, Pasadena, CA, USA
- R.P. Wayne
University of Oxford, Oxford, UK
- S. Weart
Center for History of Physics, American Institute of Physics, College Park, MD, USA
- P.J. Webster
Georgia Institute of Technology, Atlanta, GA, USA
- P.J. Webster
University of Colorado – Boulder, Boulder, CO, USA
- W.F. Weeks
University of Alaska Fairbanks, Fairbanks, AK, USA
- M.L. Weisman
National Center for Atmospheric Research, Boulder, CO, USA
- S.J. Weiss
National Oceanic and Atmospheric Administration, Norman, OK, USA
- N.C. Wells
University of Southampton, Southampton, UK
- H. Wernli
Institute for Atmospheric and Climate Science, Swiss Federal Institute of Technology (ETH), Zürich, Switzerland
- T. Wetter
Johann Wolfgang Goethe-University, Frankfurt am Main, Institut für Meteorologie und Geophysik, Frankfurt am Main, Germany
- M.C. Wheeler
Centre for Australian Weather and Climate Research, Bureau of Meteorology, Melbourne, VIC, Australia
- A.A. White
University of Surrey, Guildford, UK
- D.N. Whiteman
NASA Goddard Space Flight Center, Greenbelt, MD, USA
- B.A. Wielicki
NASA Langley Research Center, Hampton, VA, USA
- E.R. Williams
Massachusetts Institute of Technology, Cambridge, MA, USA
- M.L. Wong
California Institute of Technology, Pasadena, CA, USA
- N. Wood
Met Office, Exeter, UK
- R. Wood
University of Washington, Seattle, WA, USA
- S.N. Wren
University of Toronto, Toronto, ON, Canada
- C.-C. Wu
National Taiwan University, Taipei, Taiwan
- D.J. Wuebbles
University of Illinois, Urbana, IL, USA
- L. Xie
North Carolina State University, Raleigh, NC, USA
- P. Yang
Texas A&M University, College Station, TX, USA
- S. Yang
NOAA/NWS/NCEP, Climate Prediction Center, Camp Springs, MD, USA
- J.A. Young
University of Wisconsin, Madison, WI, USA
- Z. Yu
College of Hydrology and Water Resources, Hohai University, Nanjing, China; and University of Nevada Las Vegas, Las Vegas, NV, USA
- Y.L. Yung
California Institute of Technology, Pasadena, CA, USA

S.E. Yuter
*North Carolina State University, Raleigh,
NC, USA*

S. Yvon-Lewis
*Texas A&M University, College Station,
TX, USA*

D. Zardi
University of Trento, Trento, Italy

S.E. Zebiak
*International Research Institute for Climate Prediction,
Palisades, NY, USA*

M.D. Zelinka
*Lawrence Livermore National Laboratory, Livermore,
CA, USA*

C. Zhang
University of Miami, Miami, FL, USA

F. Zhang
*Pennsylvania State University, University Park,
PA, USA*

M. Zhang
*Stony Brook University, Stony Brook,
NY, USA*

PREFACE TO THE FIRST EDITION

A half century ago the American Meteorological Society published the *Compendium of Meteorology*, which in a single volume of 1334 pages summarized the state of understanding of the atmosphere at that time. A perusal of the contents of that volume indicates that although a broad range of topics was covered, the vast bulk of the volume was devoted to traditional meteorological topics such as atmospheric dynamics, cloud physics, and weather forecasting. Barely 4 percent of the volume was devoted to articles related to atmospheric chemistry or air pollution and, of course, none of the volume was devoted to techniques such as satellites and remote sensing.

As Sir John Mason aptly notes in his foreword to the present work, the atmospheric sciences have expanded in scope enormously over the past 50 years. Topics such as atmospheric chemistry and global climate change, of only marginal interest 50 years ago, are now central disciplines within the atmospheric sciences. Increasingly, developing areas within the atmospheric sciences require students, teachers, and researchers to familiarize themselves with areas far outside their own specialties. This work is intended to satisfy the need for a convenient and accessible references source covering all aspects of atmospheric sciences. It is written at a level that allows undergraduate science and engineering students to understand the material, while providing active researchers with the latest information in the field.

More than 400 scientists, from academia, government, and industry have contributed to the 330 articles in this work. We are very grateful to these authors for their success in providing concise and authoritative summaries of complex subjects. As editors, we have benefited from the chance to learn from these articles, and we believe that all students and active scientists who want to increase their knowledge of the atmosphere will benefit enormously from access to this work.

We are also grateful to the 31 members of the Editorial Advisory Board who have guided us in our coverage of the very broad range of topics represented in this encyclopedia. Their willingness to suggest topics and authors, and to carefully review draft articles has contributed significantly to our success.

The production of this multivolume encyclopedia would not have been possible without the dedicated work of the staff of the Major Reference Works group at Academic Press. We are especially grateful to the Major Reference Work Development Manager, Colin McNeil, who has worked closely with us during the entire process. Finally, we appreciate the liberal use of color figures in the printed encyclopedia.

James R Holton, Judith A Curry, and John Pyle

This page intentionally left blank

PREFACE TO THE SECOND EDITION

Since the publication of the first edition of the *Encyclopedia of Atmospheric Sciences*, significant advances in research have been achieved all across the broad and expanding spectrum of the field and related disciplines. In particular, climate science with primary input from the atmospheric research emerges as a new field and integrator of interlocking peripheral disciplines over the last decade. These events have demanded the solicitation of new and updated articles for the 2003 edition. Some articles from the earlier publication were judged to be of such a fundamental and enduring nature that they did not require modification. But huge amounts of new information from Earth-orbiting satellite observatories have brought much new insight to the field. In addition there are new findings in many areas such as the latest simulations of meteorological and climatic processes of interest as well as simulations and observations of the composition and interaction of the field's chemical constituents. While interest in the ozone hole and its ramifications may have reached a plateau, ever more understanding of the stratosphere and its role in climate change emerges. The study of past climates provides new means of testing climate models and theories. In weather prediction we see new progress on how data are to be better assimilated for much improved initialization of the forecast model leading to the promise of more accurate predictions of severe weather and tropical cyclones over longer lead times. These are just a few of the new features of the second edition.

The editors of the second edition are greatly indebted to our predecessors in the first edition. They set the outline of topics and solicited the original authors, while establishing a high standard for the content of this publication. In many cases we decided to reprint those articles or request only minor updates. Nevertheless, many articles in this edition are entirely original, based on which we also made significant reorganization of the content. We are proud of our product and hope it provides the same assistance to students, researchers, and practitioners throughout the science and engineering communities.

Editors of the second edition
Gerald R North
Fuqing Zhang
John Pyle

This page intentionally left blank

EDITOR BIOGRAPHIES



Gerald R North received his PhD in theoretical physics from the University of Wisconsin in 1966. After postdoctoral research at the University of Pennsylvania he became a faculty member in physics at the University of Missouri–St. Louis. He shifted his research focus to climate science research during his sabbatical year at the National Center for Atmospheric Research, where he won the Outstanding Paper Award in 1975. He moved to NASA Goddard Space Flight Center (GSFC) in 1978 where he was awarded the NASA Medal for Research Excellence. During his stay at GSFC, he was the proposer and first study scientist for the Tropical Rainfall Measuring Mission, which was launched in 1997 and is still orbiting in 2014.

He moved to Texas A&M University in 1986 as a university distinguished professor of atmospheric sciences where he served as department head from 1995 to 2003. He has served as editor-in-chief of the *Reviews of Geophysics* and is recognized as one of the most cited authors in geosciences (Web of Science). He has chaired and/or served on a number of national committees and is a Fellow of the American Geophysical Union, American Meteorological Society (AMS) and the American Association for the Advancement of Science, and winner of the Jule Charney Award for Research (AMS).

He has published about 150 refereed papers not including many book chapters and reviews. His books include *Paleoclimatology*, co-authored with Thomas Crowley, and *An Introduction to Atmospheric Thermodynamics* co-authored with Tatiana Erikhimova.

North's interests are focused on the use of mathematical and statistical tools to solve climate problems over a wide range of issues including: analytical solutions of simplified energy balance climate models, use of random field techniques in representing and interpreting climate data and model simulations, detection of deterministic signals in climate change, statistical analysis satellite remote sensing for mission planning and analysis of data, paleoclimate problems using simplified climate models.



John Pyle obtained a BSc in Physics at Durham University before moving to Oxford where he completed a DPhil in Atmospheric Physics, helping to develop a numerical model for stratospheric ozone studies. After a short period at the Rutherford Appleton Laboratory he moved to a lectureship at Cambridge University in 1985. In 2000 he was appointed professor of atmospheric science and since 2007 has been the 1920 professor of physical chemistry. He is a Professorial Fellow at St Catharine's College. He has been a codirector of Natural Environment Research Council's National Centre for Atmospheric Science, where he is currently Chief Scientist.

His research focuses on the numerical modelling of atmospheric chemistry. Problems involving the interaction between chemistry and climate have been addressed; these range from stratospheric ozone depletion to the changing tropospheric oxidizing capacity and have included the environmental impact of aviation, land use change, biofuel technologies, and the hydrogen economy. He has studied palaeochemistry problems as well as the projected atmospheric composition changes during the current century. He has published more than 250 peer reviewed papers.

He played a major role in building an EU stratospheric research programme in the 1990s, coordinating several major field campaigns. He has contributed to all the WMO/UNEP assessments on stratospheric ozone since the early 1980s and is now one of the four international cochairmen

of the Scientific Assessment Panel, responsible for these assessments. He was a convening lead author in the IPCC Special report "Safeguarding the ozone layer and the global climate system," published in 2006.

He was elected Fellow of the Royal Society in 2004 and an American Geophysical Union Fellow in 2011. He was awarded the Cambridge ScD degree in 2012. Other honours and awards include membership of Academia Europaea (1993), Royal Society of Chemistry (Interdisciplinary award, 1991, and John Jeyes lectureship, 2008), and the Royal Meteorological Society Adrian Gill Prize, in 2004.



Fuqing Zhang is a professor with tenure in the Department of Meteorology at the Pennsylvania State University, with a joint appointment in the Department of Statistics, along with an endowed position as the E Willard & Ruby S Miller Faculty Fellow at the College of Earth and Mineral Sciences at the Pennsylvania State University. His research interests include atmospheric dynamics and predictability, data assimilation, ensemble forecasting, tropical cyclones, gravity waves, mountain plains and sea-breeze circulations, warm-season convection, and regional-scale climate. He earned his BS and MS in meteorology from Nanjing University, China in 1991 and 1994, respectively, and his PhD in atmospheric science in 2000 from North Carolina State University.

He spent seven years as an assistant and then associate professor at Texas A&M University before coming to Penn State University as a full professor in 2008. In 2000, he spent a year and a half as a postdoctoral fellow at the National Center for Atmospheric Research. He also held various visiting scholarship appointments at various academic and research institutions including the National Center for Atmospheric Research in Boulder, Colorado; the Navy Research Laboratory in Monterey, California; NOAA/AOML Hurricane Research Division, Miami, Florida; Peking University and Nanjing University, China; the Chinese State Key Laboratory of Severe Weather in Beijing, China; and Laboratoire de Meteorologie Dynamique, École Normale Supérieure in Paris, France.

He has authored/co-authored about 130 peer reviewed journal publications and has given more than 160 keynote speeches or invited talks at various institutions and meetings. He has served as principal investigator/co-principal investigator for 30 federal or state-sponsored research grants. He has chaired/cochaired more than 10 scientific meetings or workshops. He also served on various review or advisory panels for numerous organizations that include National Science Foundation, Office of Naval Research, NASA, NOAA, and National Academies. He has also served as editor of several professional journals including *Monthly Weather Review*, *Science China*, *Atmospheric Science Letter*, *Acta Meteorologica Sinica*, and *Computing in Science & Engineering*.

He has also received numerous awards for his research and service. Notably, in 2007 he received the **Outstanding Publication Award** from the National Center for Atmospheric Research. In 2009, was the sole recipient of the American Meteorological Society's 2009 **Clarence Leroy Meisinger Award** "for outstanding contributions to mesoscale dynamics, predictability, and ensemble data assimilation." Most recently, he received the 2014 American Meteorological Society's **Banner Miller Award** "for valuable insights into incorporating real-time airborne Doppler radar measurements via ensemble data assimilation, leading to improvements in forecasts of tropical cyclone track and intensity."

GUIDE TO USING THE ENCYCLOPEDIA

Structure of the Encyclopedia

The material in the encyclopedia is not arranged by ordinary alphabetical order, but by alphabetical order according to 49 principal topic areas taken to allow all papers belonging to each principal topic to appear together in the same volume. Within each principal subject, article headings are also arranged alphabetically, except where logic dictates otherwise. For example, overview articles appear at the beginning of a section.

There are four features that help you find the topic in which you are interested:

- i. the contents list
- ii. cross-references to other relevant articles within each article
- iii. a full subject index
- iv. contributors

i. Contents List

The contents list, which appears at the front of each volume, lists the entries in the order that they appear in the encyclopedia. It includes both the volume number and the page number of each entry.

ii. Cross-references

All of the entries in the encyclopedia have been cross-referenced. The cross-references, which appear at the end of an article as a See also list, serve four different functions:

- i. To draw the reader's attention to related material in other entries

- ii. To indicate material that broadens and extends the scope of the article
- iii. To indicate material that covers a topic in more depth
- iv. To direct readers to other articles by the same author(s)

Example

The following list of cross-references appears at the end of the article.

See also: **Biogeochemical Cycles:** Biogeochemistry of Iodine. **Stratospheric Chemistry Topics:** HO_x; Halogen Sources, Natural (Methyl Bromide and Related Gases); Halogens; Hydrogen Budget; Overview; Reactive Nitrogen (NO_x and NO_y).

iii. Index

The index includes page numbers for quick reference to the information you are looking for. The index entries differentiate between references to a whole article, a part of an article, and a table or figure.

iv. Contributors

At the start of each volume there is list of the authors who contributed to that volume.

This page intentionally left blank

TROPICAL CYCLONES AND HURRICANES

Contents

Overview and Theory
Hurricane Dynamics
Hurricane Predictability
Hurricanes: Observation
Tropical Cyclogenesis
Tropical Cyclones and Climate Change
Tropical Cyclones in the Western North Pacific
Tropical Cyclones: Secondary Eyewall Formation

Overview and Theory

RA Tomas, University of Colorado – Boulder, Boulder, CO, USA

PJ Webster, Georgia Institute of Technology, Atlanta, GA, USA

© 2003 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 6, pp 2306–2313, © 2003, Elsevier Ltd.

Introduction

The tropics comprise roughly half the area of the globe and for this reason alone, the motion of the tropical atmosphere represents an important component of the dynamics of the planetary atmosphere. Even if one is primarily interested in the midlatitudes, it is also important to understand the dynamics of the tropics because the circulations at low latitudes can significantly influence the circulations at higher latitudes through various teleconnections. A well-known example of this is the effect that the El Niño Southern Oscillation (ENSO) has on the climate over North America. The influence of one region upon the other can also extend in the opposite direction, with the midlatitudes influencing the tropics. For example, planetary waves sometimes propagate into the tropics from the midlatitudes and excite equatorially trapped waves through lateral forcing and by inducing heating anomalies through enhanced convection.

In the midlatitudes, the rotation of the Earth plays an important role in determining the atmospheric response to forcing. This is also true for the tropics, but the role can be considerably different depending upon the proximity to the Equator. As one approaches the Equator from either pole, the Coriolis parameter $f = 2\Omega \sin(\phi)$ decreases. Here Ω is the rotation rate of the Earth and ϕ is the latitude, and thus the Coriolis parameter is the local vertical component of the Earth's rotation. At 4° latitude f is approximately only 10% as large as it is at 45° latitude. The meridional gradient of f , $\beta = \partial f / \partial y$, is largest at the Equator, and at 10° latitude f increases to be approximately 25% the value at 45° latitude. The Coriolis parameter is zero at the

Equator and it changes sign as one moves across the Equator, from one hemisphere into the other. Many of the differences between the dynamics of the midlatitudes and the dynamics of the tropics can be accounted for by these aspects of behavior of the vertical component of the Earth's rotation as one approaches the Equator.

The diminished size of the Coriolis parameter at low latitudes is responsible for the relatively small geopotential fluctuations – and, through the hydrostatic relation, the temperature fluctuations as well – compared to those in the midlatitudes. Temperature fluctuations are associated with available potential energy, and thus there is considerably less mean and eddy available potential energy in the tropics than in the midlatitudes. As a result, disturbances that grow in the tropics must do so by other means than baroclinic energy conversions, which dominate at higher latitudes. The possible energy sources for tropical disturbances include lateral forcing from the midlatitudes, barotropic energy conversions, and systematic interaction with convective heating such that temperature and heating anomalies are positively correlated. In the next section, we perform a scale analysis to examine the consequences of the magnitude of f being small in the tropics.

A Scale Analysis

Some information about the character of the synoptic-scale motions in the tropics can be obtained by performing a scale analysis of the governing equations. To do this, characteristic

scales are determined for the temporal and spatial variations of the motion based upon observations.

It can be somewhat more intuitive to use a form of the governing equations in which the vertical coordinate has units of length rather than pressure. However, if height is used as the vertical coordinate, density appears explicitly in the equations and this complicates matters rather than simplifying them. A way to avoid this complication is to specify the vertical coordinate as being proportional to the log of the pressure as in eqn [1].

$$z^* \equiv -H \ln \frac{p}{p_s} \quad [1]$$

In eqn [1], p is the pressure, p_s is a standard reference pressure (taken to be 1000 hPa), H is a standard scale height, $H \equiv RT_s/g$, R is the gas constant for dry air, T_s is an average temperature, and g is the gravity force. The units of z^* are the same as those of H , and z^* is approximately equal to the actual height from the surface in the troposphere.

The vertical velocity in this coordinate system is given by eqn [2].

$$w^* \equiv \frac{Dz^*}{Dt} \quad [2]$$

The operator D/Dt is defined in eqn [3], where $\mathbf{V} = iu + jv$, u and v are the zonal and meridional wind components, respectively, and $\nabla = i\partial/\partial x + j\partial/\partial y$.

$$\frac{D}{Dt} \equiv \frac{\partial}{\partial t} + \mathbf{V} \cdot \nabla + \frac{w^* \partial}{\partial z^*} \quad [3]$$

The governing equations are eqns [4], [5], [6], and [7].

$$\left(\frac{\partial}{\partial t} + \mathbf{V} \cdot \nabla + w^* \frac{\partial}{\partial z^*} \right) \mathbf{V} + f \mathbf{k} \times \mathbf{V} = -\nabla \Phi \quad [4]$$

$$\frac{\partial \Phi}{\partial z^*} = \frac{RT}{H} \quad [5]$$

$$\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} + \frac{\partial w^*}{\partial z^*} - \frac{w^*}{H} = 0 \quad [6]$$

$$\left(\frac{\partial}{\partial t} + \mathbf{V} \cdot \nabla \right) T + \frac{w^* N^2 H}{R} = \frac{J}{C_p} \quad [7]$$

These equations represent the horizontal momentum equation, the hydrostatic equation, the conservation of mass, and the conservation of energy, respectively. Here, Φ is the geopotential, J is the heating rate, $N^2 = (R/H)(\partial T/\partial z^* + \kappa T/H)$ is the buoyancy frequency squared, $\kappa \equiv R/C_p$, and C_p is the specific heat for dry air at constant pressure.

For the case of circulations with wavelike structure in space and time, the length and temporal scales are a quarter wavelength (Figure 1). Over a quarter-wavelength interval, a quantity initially with zero amplitude attains (absolute) peak amplitude. In this analysis, we do not provide estimates for the amplitude fluctuations of all quantities appearing in the governing equations. The unspecified quantities are diagnosed from the values provided using the governing equations. We do this because it is instructive to see why, for example, temperature perturbations are relatively small in the tropics. The magnitude of the temperature and vertical velocity perturbations that are diagnosed in the following analysis are in

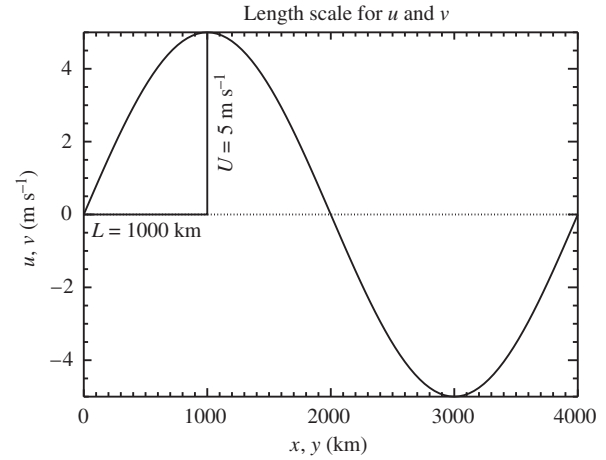


Figure 1 Schematic diagram showing an example of how the length and amplitude values were chosen for the scale analysis based on disturbances with wavelike structures in space. Time scales were chosen in a similar way, with time replacing distance along the abscissa.

agreement with observations of the type of motions used to determine the values for the scales.

We define characteristic scales for the field variables as:

$D \sim 2.5 \times 10^3 \text{ m}$ vertical depth scale

$L \sim 1.0 \times 10^6 \text{ m}$ horizontal length scale

$U \sim 5.0 \text{ ms}^{-2}$ horizontal velocity scale

$t \sim 1.0 \times 10^5 \text{ s}$ time scale

$f \sim 2.0 \times 10^{-5} \text{ s}^{-1}$ Coriolis parameter at approximately 8°

$\beta = \partial f / \partial y \sim 2.0 \times 10^{-11} \text{ m}^{-1} \text{ s}^{-1}$

$H \sim 8.0 \times 10^3 \text{ m}$ scale height

$N \sim 1.2 \times 10^{-2} \text{ s}^{-1}$ buoyancy frequency

Note that in this analysis, $D < H$.

One consequence of $f = 0$ at the Equator is the existence of disturbances that propagate zonally and whose amplitude decays with distance from the Equator. These are known as equatorially trapped waves. The scale values listed above are representative of equatorially trapped waves with a horizontal wavelength of $4 \times 10^3 \text{ km}$ and vertical wavelength of about 10 km . Numerous theoretical and observational studies confirm their existence in the tropical atmosphere and they are described in detail elsewhere in this volume (see **Tropical Meteorology and Climate: Equatorial Waves**).

We now use these values to estimate the magnitude of the terms appearing in eqn [4], as given by eqns [8] and [11].

$$\left| \frac{\partial \mathbf{V}}{\partial t} \right| \sim \frac{U}{t} \sim 5.0 \times 10^{-5} \text{ ms}^{-2} \quad [8]$$

$$|(\mathbf{V} \cdot \nabla) \mathbf{V}| \sim \frac{U^2}{L} \sim 2.5 \times 10^{-5} \text{ ms}^{-2} \quad [9]$$

$$\left| w^* \frac{\partial \mathbf{V}}{\partial z^*} \right| \sim w^* \frac{U}{D} \sim w^* \times 2.0 \times 10^{-3} \text{ ms}^{-2} \quad [10]$$

$$|f \mathbf{k} \times \mathbf{V}| \sim \frac{U}{f} \sim 1.0 \times 10^{-4} \text{ ms}^{-2} \quad [11]$$

We do not introduce an estimate for the size of w^* from observations and for now we assume that it is small enough so that the vertical advection is considerably less than the other terms. This will turn out to be the case, as will be shown later. The Coriolis term is larger than the other terms on the right-hand side of eqn [4] by a factor of 2 compared to the local time rate of change, and by a factor of 4 compared to the horizontal advection term. For a disturbance with a wavelike horizontal structure, the velocity and the gradient of the velocity are out of phase. Thus, the advection term is likely smaller than estimated in eqn [9]. We assume, then, that the geopotential gradient term is of the same order as the Coriolis term, although it could be somewhat smaller, particularly closer to the Equator. We then have relation [12].

$$\begin{aligned}\nabla\Phi &\sim \frac{\delta\Phi}{L} \sim 1.0 \times 10^4 \text{ m s}^{-2} \\ \Rightarrow \delta\Phi &\sim 100 \text{ m}^2 \text{ s}^{-2}\end{aligned}\quad [12]$$

From eqn [5], we can estimate the temperature perturbation associated with a geopotential perturbation of this size (eqn [13]).

$$T = \frac{H}{R} \frac{\partial\Phi}{\partial z^*} \sim \frac{H}{R} \frac{\delta\Phi}{D} \sim 1.0 \text{ K} \quad [13]$$

The geopotential and the associated temperature perturbations are quite small, about an order of magnitude less than those with the same scale in the midlatitudes. This is a consequence of the Coriolis parameter being an order of magnitude smaller in the tropics than in the midlatitudes.

The terms in the thermodynamic energy equation [7] can now be evaluated as in eqns [14] and [15].

$$\frac{\partial T}{\partial t} \sim \frac{T}{t} \sim 1.0 \times 10^{-5} \text{ K s}^{-1} \quad [14]$$

$$(\mathbf{V} \cdot \nabla)T \sim \frac{UT}{L} \sim 5.5 \times 10^{-6} \text{ K s}^{-1} \quad [15]$$

The local time rate of change is larger than the horizontal advection term by approximately a factor of 2. For the case of baroclinic motions, the temperature perturbations are in quadrature (in the vertical direction) with the horizontal motions, so the horizontal advection of temperature is smaller than the estimate in eqn [15]. We assume that the flow is adiabatic ($J = 0$) and the local time rate of change of temperature is balanced by vertical motion. The balance between the local time rate of change of temperature and vertical motion may then be used to estimate the vertical motion (eqn [16]).

$$\begin{aligned}\frac{\partial T}{\partial t} &\sim \frac{w^* N^2 H}{R} \\ \Rightarrow w^* &\sim \frac{R}{HN^2} \frac{\partial T}{\partial t} \sim 2.5 \times 10^{-3} \text{ m s}^{-1}\end{aligned}\quad [16]$$

Substituting this value into eqn [10], we see that the vertical advection term is small compared to the other terms in eqn [4], as we assumed earlier.

We can now evaluate the magnitude of all of the terms appearing in the continuity equation [6] as in eqns [17], [18], and [19].

$$\frac{\partial u}{\partial x} \sim \frac{\partial v}{\partial y} \sim \frac{U}{D} \sim 5.0 \times 10^{-6} \text{ s}^{-1} \quad [17]$$

$$\frac{\partial w^*}{\partial z^*} \sim \frac{w^*}{D} \sim 1.0 \times 10^{-6} \text{ s}^{-1} \quad [18]$$

$$\frac{w^*}{H} \sim 3.4 \times 10^{-7} \text{ s}^{-1} \quad [19]$$

Comparing eqns [17] and [18], it is clear that there is considerable cancellation between the two horizontal divergence terms. Nevertheless, the horizontal divergence given by their sum is crucial for the existence of equatorially trapped waves. The associated vertical motion and adiabatic temperature changes are necessary to keep the geopotential and temperature fields in hydrostatic balance.

The higher-frequency modes, i.e., the eastward and westward inertio-gravity waves, and the Kelvin waves of all frequencies are highly divergent. To first order, they propagate because of the geopotential changes that result from divergence and convergence. These waves behave like gravity waves in this respect.

The wind fields associated with the lower-frequency modes – the Rossby waves – are much less divergent much more rotational. To first order, they propagate because of vorticity changes that result from the advection of planetary vorticity. This may be illustrated using the equation describing the time rate of change of the vertical component of vorticity. A vorticity equation may be obtained by taking $\mathbf{k} \cdot \nabla$ [4]. Neglecting the terms that are small for synoptic scale motions (vertical advection, tilting, and solenoidal terms) the result is given by eqn [20], where $\zeta = \partial v / \partial x - \partial u / \partial y$.

$$\frac{\partial \zeta}{\partial t} = (\mathbf{V} \cdot \nabla)(\zeta + f) + (\zeta + f)(\nabla \cdot \mathbf{V}) \quad [20]$$

Substituting in the scaling values for the two terms on the right-hand side, we obtain eqns [21] and [22].

$$\begin{aligned}(\mathbf{V} \cdot \nabla)(\zeta + f) &\sim U \left(\frac{U}{L^2}, \beta \right) : \\ \frac{U^2}{L^2} &\sim 2.5 \times 10^{-11} \text{ s}^{-2} \\ U\beta &\sim 1.0 \times 10^{-10} \text{ s}^{-2}\end{aligned}\quad [21]$$

and

$$\begin{aligned}(\zeta + f)(\nabla \cdot \mathbf{V}) &\sim \left(\frac{U}{L}, f \right) \frac{w^*}{D} : \\ \frac{U w^*}{L D} &\sim 5.0 \times 10^{-12} \text{ s}^{-2} \\ f \frac{w^*}{D} &\sim 2.0 \times 10^{-11} \text{ s}^{-2}\end{aligned}\quad [22]$$

Two values are given for each term. The first represents the value obtained using the relative vorticity, ζ , and the second represents the value obtained using the planetary vorticity, f . The second value listed for eqn [21] is appropriate for the entire tropical region because β does not change very much between 30° S and 30° N. In contrast the Coriolis parameter, f , varies considerably within this region. The second value listed for eqn [22] is evaluated at 8° latitude and it approaches zero as $f = \beta y \rightarrow 0$.

For $v \neq 0$ the meridional advection of planetary vorticity is an order of magnitude larger than any of the other terms.

It must be balanced by the time rate of change of vorticity, resulting in wave propagation. This is also true for Rossby waves in the midlatitudes. For the Kelvin wave $v \equiv 0$, the advection of planetary vorticity is zero, and associated vorticity changes, which are small compared to those for the Rossby wave, result only from the divergence term.

The theoretical aspects of equatorially trapped waves are well understood and a number of observational studies have documented and described them (*see Tropical Meteorology and Climate: Equatorial Waves*). They are sometimes systematically associated with convection and, at least in some cases, convective heating acts as a substantial source of energy for the circulations.

The results from this scale analysis depend critically on the value of the depth scale, D . This is because the divergence is proportional to $\delta\Phi/D^2$ and therefore the magnitude of the divergence decreases markedly as D increases. An alternate approach to the one taken in this section is to choose the depth scale to be the same as the scale height. If this is done, one concludes that synoptic-scale motion in the tropics is essentially nondivergent. These circulations cannot convert potential energy to kinetic energy and thus they must be forced laterally from the midlatitudes or locally by barotropic energy conversions from the basic state flow. There is some observational support for the existence of such motion, but not nearly as much as there is for equatorially trapped modes. Perhaps this is because they have not attracted much attention because non-divergent motions would not be associated with precipitation.

The Large-Scale Circulation in the Tropics

What processes determine the structure of the observed large-scale flow in the tropics? Even if one ignores the fact that the tropical circulation and the circulation at higher latitudes influence each other in ways that are sometimes very important, the question posed is still very complex because it depends on processes that have vastly different time and length scales. Also, the forcing functions are not independent of the motion itself.

Latent heating is the dominant source of energy for the tropical circulation, with radiative heating playing an important but secondary role. Therefore, motions on the scale of individual convective clouds and the mesoscale and synoptic-scale systems in which they are sometimes organized are an important part in determining the strength and location of convection that drives the large-scale flow. In addition, the distributions of radiative heating and cooling are strongly influenced by the distribution and the physical characteristics of clouds. The processes that determine the cloud radiative forcing are also significant parts of the answer to the question posed. Inclusion of this level of resolution in a modeling or theoretical study is not very practical, however, and the average effect of convective and radiative processes must be modeled or parameterized in terms of larger-scale variables if they are to be included in studies of the large-scale flow.

The Strength and Location of Convection

As discussed above, tropical motions are driven principally by the latent heat released during convection. The strength and the

location of convective activity, on the other hand, depends strongly on the state of the atmosphere. Different properties of the atmosphere influence convective activity for different reasons. In this section, we consider various processes that influence the strength and location of convection starting at the lower boundary and progressing upward into the free atmosphere.

At the surface, the state of the atmosphere and the state of the ocean and land surface both influence convective activity through their effect on the surface fluxes of latent and sensible heat. The flux of latent heat from the ocean surface supplies the moisture for convective activity. In some instances, this moisture is transported horizontally for large distances before it moves vertically and condenses out of the atmosphere as precipitation. The flux of sensible heat may alter the surface temperature, which in turn may increase the instability of the atmosphere with respect to convection. The physical processes responsible for these fluxes have too small a scale to be explicitly resolved in large-scale modeling and theoretical studies. Data with the resolution necessary to explicitly resolve these processes is rarely available for large-scale analytical studies. In these cases, when knowledge of the fluxes is required, the fluxes must be parameterized in terms of the resolved scale variables. The sensible and latent heat fluxes may be parameterized using a so-called bulk formulation, as in eqns [23] and [24], for example.

$$\frac{H}{\rho C_p} = \overline{w'T'} = C_h U (T_{\text{surface}} - T) \quad [23]$$

$$\frac{E}{\rho L_v} = \overline{w'q'} = C_q U (q_{\text{surface}} - q) \quad [24]$$

In these equations, H is the sensible heat flux, E is the latent heat flux, and, w' , T' , and q' are the sub-grid-scale vertical velocity, temperature, and specific humidity, respectively. C_h and C_q are the transfer coefficients for the sensible and latent heat, which are determined empirically, and $U = (u^2 + v^2)^{1/2}$. T_{surface} is the sea surface temperature (SST) or the land surface temperature and q_{surface} is the saturation mixing ratio calculated using the SST. In the case of land, q_{surface} is calculated in the same way but, depending on the amount of water in the surface layer, it may not be saturated.

In this formulation, the sensible heat flux is proportional to the difference between the surface layer temperature and the atmospheric temperature multiplied by the wind speed. In similar way, the latent heat flux is proportional to the difference between the specific humidity at the surface (which in the case of an ocean surface, is assumed to be saturated at the temperature of the sea surface) and the specific humidity in the atmosphere, multiplied by the wind speed. Feedbacks between the atmosphere and the ocean exist for several reasons. One of the simplest results from the fluxes of latent and sensible heat that cool the ocean near the surface. The cooling reduces the temperature difference at the interface and hence reduces the flux, and there exists a negative feedback.

Another mechanism whereby the atmospheric circulation and the sea surface temperature interact may be explained using the observation that the virtual potential temperature ($\theta_v = T_v(p_0/p)^{\kappa}$), where T_v is the virtual temperature, is approximately constant with height in a well-mixed marine

boundary layer. In addition, observations show that the horizontal variations of θ_v in the atmosphere closely resemble those in the SST. This occurs because shallow convection within the boundary layer mixes air near the surface upward. This temperature and moisture distribution then influences the pressure distribution through the hydrostatic relation. This pressure distribution then determines the boundary layer winds, which may act to transport moisture for some distance. The wind patterns may also result in regions of moisture convergence throughout the depth of the boundary layer and this can have a strong influence on the strength and the location of convection. The convergence does not influence the convection on the scale of individual cumulus clouds but rather on the larger scale, by increasing the low-level moisture over a large region and making it more likely that individual clouds reach the level of free convection.

As evidence in support of this mechanism, observations show that in some locations, such as the western Pacific, warmer SSTs generally associated with stronger convection and the strongest convection is collocated with the warmest SSTs. Such relationships might be expected based upon consideration of eqns [23] and [24] alone. In locations where there exists a moderate pressure gradient across the Equator, however, these relationships break down. These are the

intertropical convergence zones (ITCZs). The SSTs in the ITCZs are observed to be warm but are several degrees cooler than in the west Pacific. In spite of this, the outgoing longwave radiation (OLR) values are comparable in both regions. Also, the convection in the ITCZs is not collocated with the local SST maximum. It has been suggested that this is the result of dynamically driven convergence in the boundary layer that results from the cross-Equator pressure gradient force. A study using a simple model shows that the nonlinear dynamical response to an imposed pressure gradient force directed across the Equator results in a single meridional wind maximum located in the hemisphere with the low-pressure center. The maximum has a narrow meridional extent and is thus associated with a dipole in the convergence field, with the convergence center located between the Equator and the low-pressure center. Some of the results from this study are shown in Figure 2. The solid lines indicate the solution obtained using the full model equations. In this study, the presence of locally anticyclonic absolute vorticity was shown to be important for there to be a single maximum in the meridional wind. When the model was altered to keep the absolute vorticity cyclonic everywhere, the single meridional wind maximum was replaced with two weaker maxima, one in each hemisphere. These results are shown in Figure 2 as dashed lines.

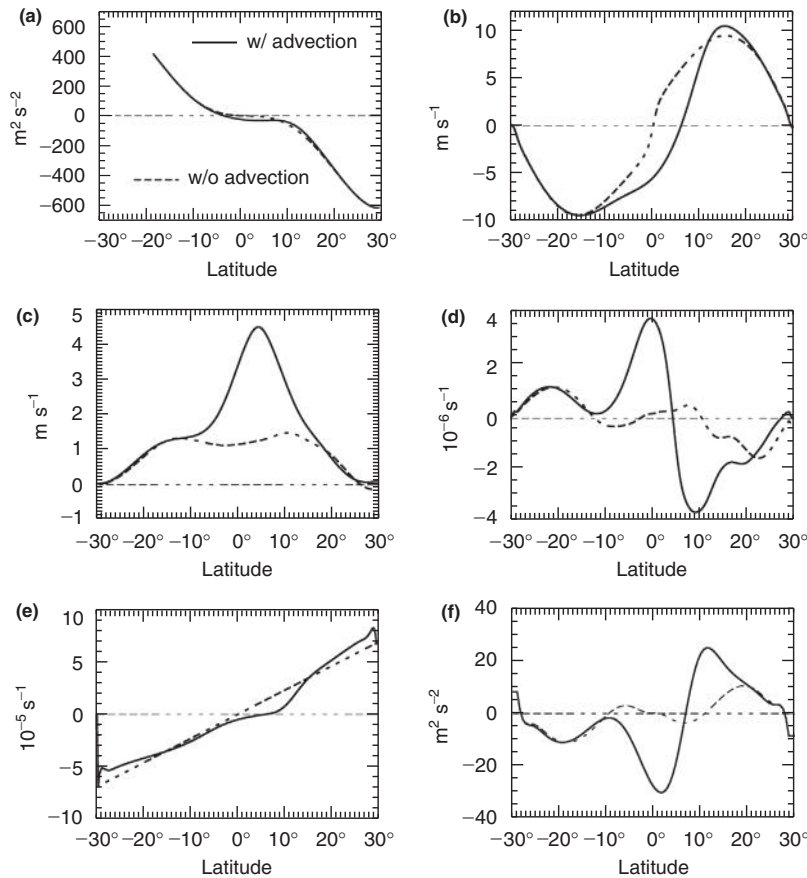


Figure 2 Latitudinal profiles of simulated steady-state parameters obtained using the idealized forcing: (a) geopotential, (b) zonal wind (c) meridional wind, (d) divergence, (e) absolute vorticity, (f) difference between simulated and basic state geopotential. Solid lines indicate the case with the term $v \, du/dy$ included; dashed lines indicate the case with the term omitted. After Tomas R, Holton J, and Webster P (1999).

Finally, divergence in the upper levels of the atmosphere can influence convection. This was recognized early in the modern history of tropical meteorology. The continuity equation requires that vertical motion must exist below and/or above the divergence centers. When the divergence is located in the upper troposphere, the air above is much more statically stable than the air below; thus upward motion from below dominates in the conservation of mass. This upward motion must be balanced by adiabatic cooling for the thermodynamic energy equation to hold. This cooling increases the convective available potential energy in the column. If this cooling occurs low enough in the troposphere, it may also lower the level of free convection.

So far we have discussed various processes whereby the atmosphere and ocean and land exert their influence on the strength and the location of convection and some feedbacks that occur in connection with these processes. This discussion was motivated by the observations that the tropical atmosphere is driven principally by latent heating and that the distribution of latent heating depends strongly on the atmospheric flow. An important question at this point is what the circulation develops in response to convective heating.

Modeling the Response to Imposed Heating

A full account of the large-scale flow in the tropics involves many elements of a complex coupled climate system that includes a wide range of important temporal and spatial scales. Clearly, an approach that involves looking at certain aspects of the tropical circulation in isolation from others can simplify the problem to the point that allows the identification of fundamentally important mechanisms. This strategy has been used successfully to examine the tropical atmosphere's response to imposed heat sources. The approach used in several studies is to construct a simple model of the steady-state atmosphere using the shallow water equations and assuming that the heating is sufficiently small that the motion may be accurately modeled using linear dynamics. The resulting equations after non-dimensionalizing are given as eqns [25], [26], [27], and [28].

$$\varepsilon u - \frac{1}{2}\gamma v = -\frac{\partial p}{\partial x} \quad [25]$$

$$\varepsilon v + \frac{1}{2}\gamma u = -\frac{\partial p}{\partial y} \quad [26]$$

$$\varepsilon p + \frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} = -Q \quad [27]$$

$$w = -\varepsilon p + Q \quad [28]$$

Here (x, y) is the nondimensional distance in the x (eastward) and y (northward) directions, (u, v) is proportional to the horizontal velocity, p is proportional to the pressure perturbation, and Q is the heating rate, such that if the sign of Q is positive, the signs of u, v, p correspond to those quantities at the surface. The vertical velocity w is given by eqn [28], which is derived from the buoyancy equation. Rayleigh friction (dissipation proportional to speed) and Newtonian cooling (heating or cooling proportional to temperature anomalies relative to the mean temperature) are including using the same coefficient, ε .

The distribution of heating may be specified in terms of the normal modes to these equations, making it possible to obtain analytical solutions. The solution to eqns [25], [26], and [27] for the case of heating centered on the Equator is shown in Figure 3.

One of most striking aspects of this solution is that the influence of the heating on the circulation extends farther eastward than it does westward. Also, the solution to the east of the heating is confined more closely to the Equator than the solution to the west of the heating. Having the analytical solutions to eqns [25], [26], and [27] allows one to interpret this response in terms of the propagation of equatorially trapped waves. The low-level easterlies to the east of the heat source are the result of a Kelvin wave propagating eastward, under the influence of dissipation. The low-level westerlies lying to the west of the heating are the result of a $(n = 1)$ Rossby wave propagating westward under the influence of dissipation. Since the Kelvin wave travels three times faster than the Rossby wave, the easterlies extend farther from the heating than do the westerlies. The Kelvin wave has $v \equiv 0$ and the zonal flow is confined close to the Equator. The $(n = 1)$ Rossby wave

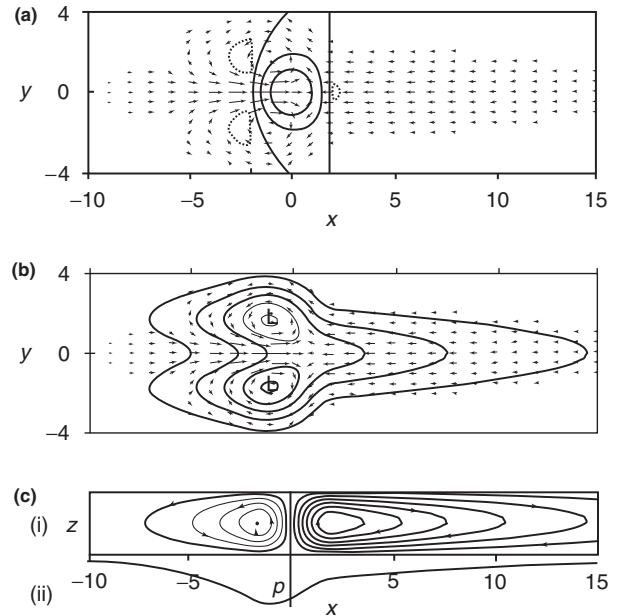


Figure 3 Solution for heating that is symmetric about the Equator in the region $|x| < 2$ for decay factor $\varepsilon = 0.1$. (a) Contours of vertical velocity w (solid contours are 0, 0.3, 0.6, broken contour is -0.1) superimposed on the velocity field for the lower layer. The field is dominated by the upward motion in the heating region, where it has approximately the same shape as the heating function. Elsewhere there is subsidence with the same pattern as the pressure field. (b) Contours of perturbation pressure p (contour interval is 0.3), which is everywhere negative. There is a trough at the Equator in the easterly regime to the east of the forcing region. On the other hand, the pressure in the westerlies to the west of the forcing region, though depressed, is high relative to its value of the Equator. Two cyclones are found to the north-west and south-west flanks of the forcing region. (c) The meridionally integrated flow showing (i) stream function contours, and (ii) perturbation pressure. Note the rising motion in the heating region (where there is a trough) and subsidence elsewhere. After Gill (1980).

horizontal structure consists of two gyres straddling the Equator and rotating in opposite senses.

The distributions of heating and low-level winds are similar to those observed in the Indian and Pacific ocean regions. That is, there is strong convection over the Indonesian region. To the west, over the Indian Ocean, there are westerlies. To the east, over the Pacific Ocean, there are easterlies and the easterlies extend over a much larger distance. The greatest limitation to applying these results to the atmosphere is that they were obtained by specifying the forcing, which is assumed to represent convective heating. As discussed in the previous subsection, the heating depends strongly on the atmospheric circulation. A consistent model of the atmospheric circulation would predict the distribution of heating rather than having it prescribed. Nevertheless, studies of the steady atmospheric response to imposed forcing provide valuable insights to important physical mechanisms that determine horizontal flow patterns and show reasonably good agreement with the broad aspects of the observed circulation.

See also: **Dynamical Meteorology:** Inertial Instability; Overview; Quasigeostrophic Theory. **Tropical Cyclones and Hurricanes:** Hurricanes: Observation. **Tropical Meteorology and Climate:** El Niño and the Southern Oscillation: Observation; El Niño and the Southern Oscillation: Theory; Equatorial Waves; Intertropical Convergence Zone; Intraseasonal Oscillation (Madden–Julian Oscillation);

Monsoon: Dynamical Theory; Monsoon: ENSO–Monsoon Interactions; Monsoon: Overview; Tropical Climates.

Further Reading

- Battisti, D., Sarachik, E., Hirst, A., 1999. A consistent model for the large scale steady surface atmospheric circulation in the tropics. *Journal of Climate* 12, 2956–2964.
- Charney, J., 1963. A note on large-scale motion in the tropics. *Journal of the Atmospheric Sciences* 20, 607–609.
- Gill, A.E., 1980. Some simple solutions for heat-induced tropical circulation. *Quarterly Journal of the Royal Meteorological Society* 106, 447–462.
- Holton, J., 1992. *An Introduction to Dynamic Meteorology*, 3rd edn. Academic Press, New York.
- Lindzen, R., Nigam, S., 1987. On the role of the sea surface temperature gradients in forcing low level winds and convergence in the tropics. *Journal of the Atmospheric Sciences* 44, 2418–2436.
- Matsuno, T., 1966. Quasi-geostrophic motions in the equatorial area. *Journal of the Meteorological Society of Japan* 44, 25–43.
- Riehl, H., 1954. *Tropical Meteorology*. McGraw-Hill, New York.
- Tomas, R., Holton, J., Webster, P., 1999. The influence of cross-equatorial pressure gradients on the location of near-equatorial convection. *Quarterly Journal of the Royal Meteorological Society* 125, 1107–1127.
- Tomas, R., Webster, P., 1997. The role of inertial instability in determining the location and strength of near-equatorial convection. *Quarterly Journal of the Royal Meteorological Society* 123, 1445–1481.
- Webster, P., 1972. Response of the tropical atmosphere to local steady forcing. *Monthly Weather Review* 100, 518–541.

Hurricane Dynamics

Y Wang, University of Hawaii at Manoa, Honolulu, HI, USA

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

Hurricanes are strong cyclonically rotating systems in the troposphere. A mature hurricane is always well organized with rich structures and unique dynamics. In this article, some key dynamics of hurricanes will be covered, including the basic dynamic structure of a mature hurricane; dynamics of hurricane genesis; dynamics of intensification, maximum potential intensity; and dynamics of hurricane motion. Some other aspects of hurricanes can be found in other relevant articles.

Basic Dynamic Structure

There are different dynamical features in different regions of a mature hurricane (Figure 1). Radially, a hurricane has an eye at its center where little precipitation with almost free or weak radar echoes. The eye is surrounded by outwardly sloped deep cloud wall, named eyewall with heavy precipitation and strongest swirling winds. Immediately outside the eyewall, there are often well-organized narrow rainbands, called inner spiral rainbands where radial shear of swirling wind is extremely large and high wave number asymmetries are often damped effectively. Further outside, namely outside of about 2–3 times of the radius of maximum wind (RMW) are loosely organized rainbands with relatively wide scale in the radial direction and embedded strong convective cells. Those rainbands are often referred to as outer spiral rainbands.

In the vertical, a hurricane consists of three regions. In the lower troposphere, there is a layer immediately above the surface where strong inflow exists, often referred to as the inflow boundary layer. The inflow brings absolute angular momentum (AAM, $M = V_T r + \frac{fr^2}{2}$, where V_T is the tangential wind velocity, r is the radius from the hurricane center, and f is the Coriolis parameter) inward to intensify the tangential wind against the loss of angular momentum to surface friction. The inflow then converges near the RMW and turns upward in the eyewall to support eyewall deep convection. After air in the strong eyewall updrafts reaches the upper troposphere, a small part of the air is

entrained into the eyewall region and is forced to subside and the associated adiabatic warming leads to the formation of a warm-core structure in the mid-upper troposphere. A large part of the air from the eyewall updrafts turns radially outward, forming a relatively deep outflow layer. At some distance away from the eyewall, the outflow will become anticyclonic to conserve AAM and form a large divergent anticyclonic circulation in the upper troposphere. Because of the warm-core structure, the strength of cyclonic circulation decreases with height, leading to the outward tilt of the AAM surface. Since upward motion above the inflow boundary layer follows the AAM surface, the eyewall slopes radially outward with height.

Axisymmetric Structure

To the first order, hurricanes can be viewed as a horizontally quasi-symmetric, primary circulation superimposed by a thermally direct vertical radial, secondary circulation. The primary circulation is nearly in gradient wind balance, namely, centrifugal force and Coriolis force balance the pressure gradient force except for in the frictional boundary layer (Willoughby, 1990):

$$\frac{V_T^2}{r} + fV_T = \frac{1}{\rho} \frac{\partial p}{\partial r} \quad [1]$$

where ρ is the air density and p is the air pressure. The gradient wind (V_G) can be written as

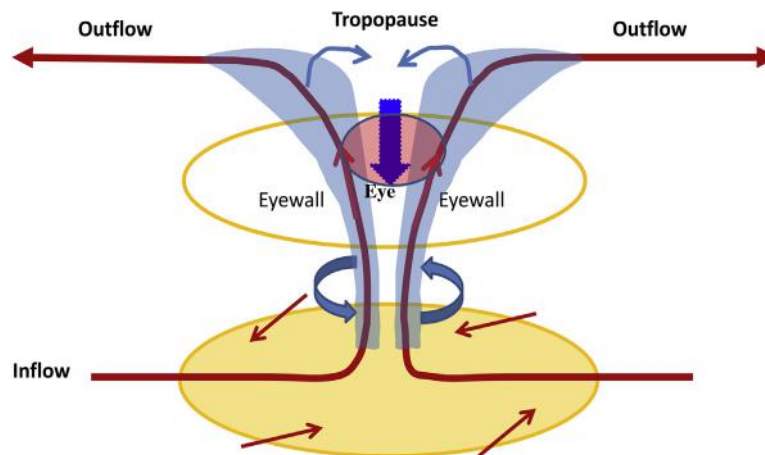


Figure 1 Schematic diagram showing the overall structure of a mature hurricane.

$$V_G = -\frac{fr}{2} + \sqrt{\frac{f^2 r^2}{4} + \frac{1}{\rho} \frac{\partial p}{\partial r}} \quad [2]$$

Observations in hurricanes indicate that gradient wind balance is a good approximation to the actual winds above the boundary layer.

Hurricanes are warm-cored vortices. The radial temperature gradient implies the decrease of tangential wind with height. The vertical shear of tangential wind and the radial temperature gradient satisfy the thermal wind balance. If we write the gradient wind balance in pressure coordinate and use the hydrostatic approximation, we can get

$$\left(f + \frac{2V_T}{r}\right) \frac{\partial V_T}{\partial p} = -\frac{R}{p} \frac{\partial T}{\partial r} \quad [3]$$

where T is the air temperature. Since the vertical shear in tangential wind is not only a function of radial temperature gradient but also the pressure. For the same radial temperature gradient, vertical shear is larger in the upper troposphere than in the lower troposphere and is smaller in the core region than in the region outside the core because of the small radius in the core. This suggests that small hurricanes have a smaller outward slope of the eyewall than the large hurricanes. Thermal wind balance is important to understanding many aspects of hurricane dynamics.

Dynamically, the strong cyclonically swirling flow indicates strong inertial stability (I^2) in the inner core region. In an axisymmetric system, I^2 can be expressed as

$$I^2 = \left(f + \frac{2V_T}{r}\right) \left(f + \frac{1}{r} \frac{\partial r V_T}{\partial r}\right) \quad [4]$$

The high inertial stability implies the resistance to the radial inflow in the boundary layer. As a result, the inertial stability and the strength of the inflow determine where the eyewall upwards would occur and thus the size of the eyewall. Inertial stability also determines the efficiency of warming in the eye forced by diabatic heating in the eyewall (Schubert and Hack,

1982). Since the inertial stability is maximum inside the RMW (Figure 2), diabatic heating inside of the RMW contributes more to the rapid intensification (RI) of a hurricane (Vigh and Schubert, 2009).

Another distinct feature of the strong differential rotation in the inner core of the hurricane vortex is the existence of an annular region immediately outside the RMW. In this annular region, most of the high azimuthal wave number asymmetries would be filamented and thus effectively damped and axisymmetrized. This region is characterized by its strong shear deformation and is often referred to as the filamentation zone. Under the assumption of the nearly circular particle trajectories in a quasi-axisymmetric hurricane vortex, the quantity

$$Q = \frac{1}{4}(S_1^2 + S_2^2 - \zeta^2) \quad [5]$$

measures the relative magnitudes of strain and vorticity and is often called Okubo-Weiss criterion (Guinn and Schubert, 1993), where

$$S_1^2 + S_2^2 = \left(\frac{\partial V_r}{\partial r} - \frac{V_r}{r} - \frac{\partial V_T}{r \partial \lambda}\right)^2 + \left(\frac{\partial V_T}{\partial r} - \frac{V_T}{r} + \frac{\partial V_r}{r \partial \lambda}\right)^2 \quad [6]$$

$$\zeta^2 = \left(\frac{\partial V_T}{\partial r} + \frac{V_T}{r} - \frac{\partial V_r}{r \partial \lambda}\right)^2 \quad [7]$$

where V_r is the radial wind and λ is the azimuth, S_1 is the stretching deformation and S_2 is the shearing deformation, and ζ is the vertical component of relative vorticity.

In regions where $Q < 0$, vorticity dominates strain, trajectories of two neighboring particles do not separate exponentially in time, coherent structures such as mesovortices can survive. This is the case in the eyewall region within the RMW. Conversely, in regions where strain dominates vorticity $Q > 0$, the vorticity gradient intensifies, forming long, thin vorticity gradient sheets across which vorticity changes rapidly along the trajectories. As a result, any asymmetries, particularly those with high azimuthal wave numbers, would be highly distorted

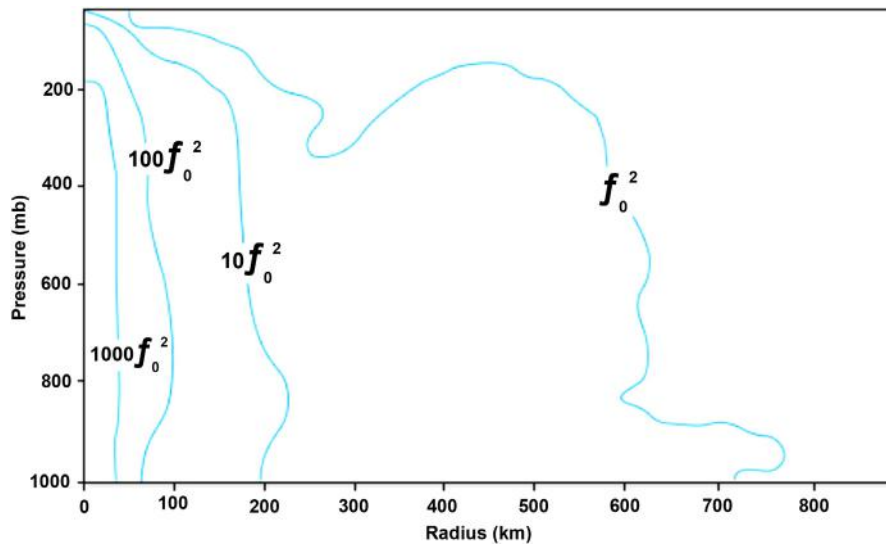


Figure 2 Vertical-radial distribution of inertial stability in a typical hurricane. From Holland, G.J., Merrill, R.T., 1984. On the dynamics of tropical cyclone structural changes. Quarterly Journal of the Royal Meteorological Society 110, 723–745.

(filamented) and axisymmetrized. The axisymmetrization can be measured by the so-called filamentation time that is defined as the e -folding time of the exponential solution in the strain-dominated regions

$$\tau_{\text{fil}} = 2(S_1^2 + S_2^2 - \zeta^2)^{-1/2} \quad \text{for } S_1^2 + S_2^2 > \zeta^2 \quad [8]$$

Since deep, moist convection in the hurricane core region has a typical overturning timescale of about 30 min, the region in which $\tau_{\text{fil}} < 30$ min can be defined as the rapid filamentation zone (RFZ; Rozoff et al., 2006). In the near-core environment of intense tropical cyclones (TCs), the primary flow is quasi-axisymmetric and rapidly rotating. As a result, the shearing deformation generally dominates the stretching deformation in the mid-lower troposphere. Essentially, all fields in the RFZ are generally filamented and deep convection could be highly distorted and even suppressed. Nevertheless, the strain flow in the RFZ outside the RMW provides a favorable environment for

the organized inner spiral rainbands, while the filamentation time provides a quantitative measure of the stabilization/axisymmetrization of high azimuthal wave number asymmetries in the inner core by shearing deformation and filamentation (Wang, 2008). An example for the asymmetric eddy kinetic energy (EKE) in a simulation of an idealized hurricane is shown in Figure 3. In this case, the hurricane has an RMW of about 20 km and RFZ is roughly between radii of 30 and 60 km where relatively small EKE occurs, indicating suppressed asymmetries in the RFZ of the simulated hurricane.

Asymmetric Structure

Although a strong mature hurricane is highly axisymmetric, considerable asymmetric structure can develop as a result of either internal dynamical instability or external forcing, such as environmental vertical wind shear, or both. The asymmetries

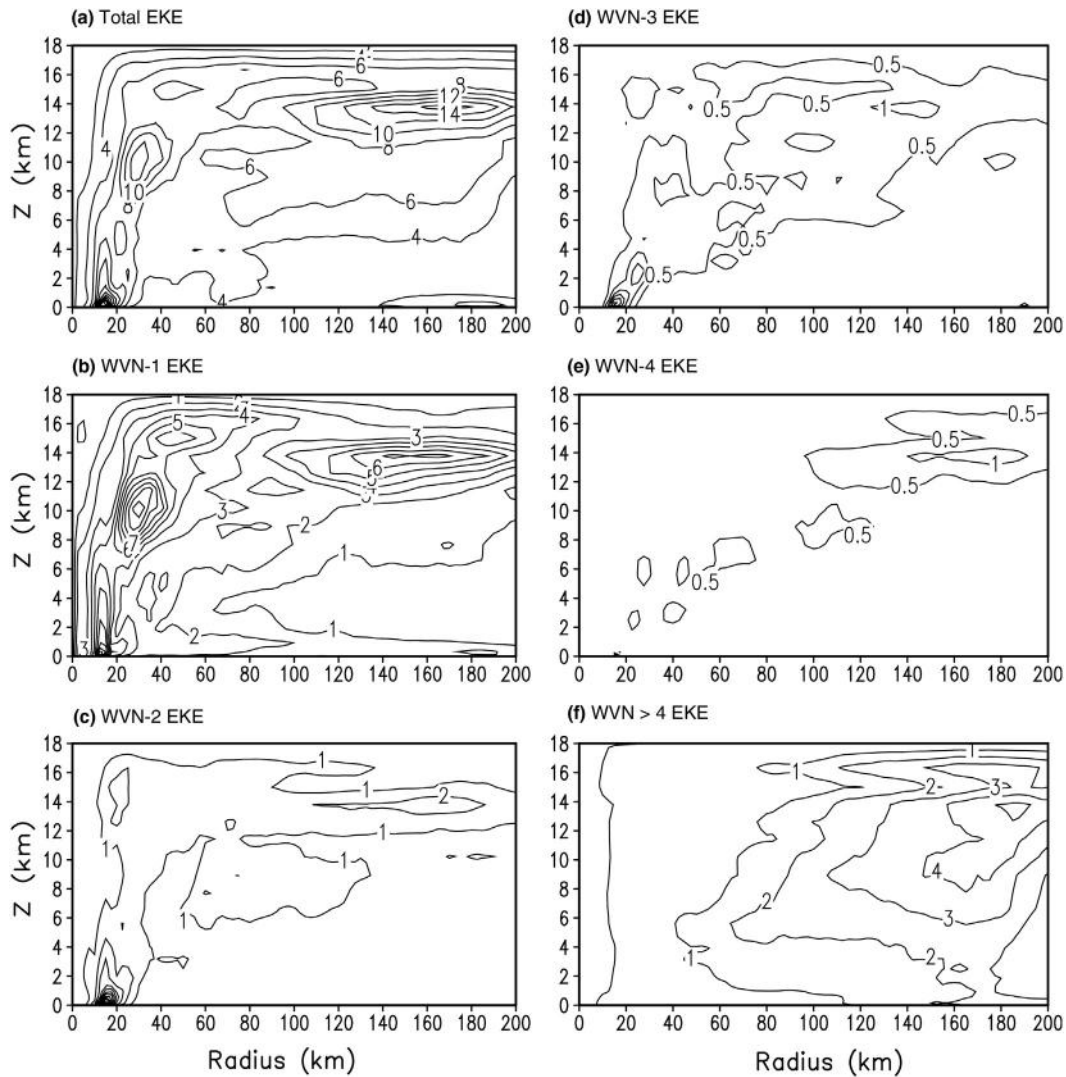


Figure 3 Vertical-radial distributions of (a) the azimuthal mean total, (b) wave number-1, (c) wave number-2, (d) wave number-3, (e) wave number-4, and (f) wave number > 4, eddy kinetic energy (EKE, in $\text{m}^2 \text{s}^{-2}$) averaged between 96 and 144 h of a simulation for an idealized hurricane. Note that the contour intervals are $2 \text{ m}^2 \text{s}^{-2}$ in (a), $1 \text{ m}^2 \text{s}^{-2}$ in (b, c, f), and $0.5 \text{ m}^2 \text{s}^{-2}$ in (d, e). WVN, wave number. From Wang, Y., 2008. Rapid filamentation zone in a numerically simulated tropical cyclone. *Journal of Atmospheric Sciences* 65, 1158–1181.

are characterized by the so-called vortex Rossby waves (VRWs) in the inner core region and inertia-gravity waves further outside, mesovortices in the eyewall, and polygonal eyewall (Figure 4). In particular, the dynamics of VRWs has been studied comprehensively in the past decade or so. The dynamics of VRWs is very similar to the planetary Rossby waves (MacDonald, 1968; Montgomery and Kallenbach, 1997). The restoring force for the former is the radial gradient of vorticity (or potential vorticity (PV)) while that for the latter is the meridional gradient of Coriolis parameter or equivalently the differential rotation of the earth surface.

Hurricanes are localized vortices with elevated cyclonic PV concentrated in the inner core region near the RMW with large radial gradients. Any radial perturbation of air parcel would experience restoring force due to the presence of PV gradients, and generate edge PV waves. Because their resemblance to the Rossby waves in the mid-latitude large-scale motion, these PV waves are usually termed as VRWs, namely, Rossby-type waves in a circular vortex. The radially outward-propagating VRWs are responsible for initiation of the inner spiral rainbands, and

can affect the structure and intensity of the mean vortex by wave-mean flow interaction. One of the distinct features of Rossby waves in the middle latitude is its quasi-geostrophic balance between mass and circulation fields. The VRWs in hurricanes are dominated by low azimuthal wave numbers in the eyewall with their maximum amplitudes near the RMW and a spiral structure in PV fields. These waves propagate upwind around the eyewall relative to the tangential flow of the azimuthal mean cyclone. The VRWs have a coherent structure in the mid-lower troposphere where the radial PV gradients are large (Chen and Yau, 2001; Wang, 2002a).

The simplest way to understand the basic dynamics of VRWs is based on the linearized nondivergent barotropic vorticity equation below

$$\left(\frac{\partial}{\partial t} + \frac{\bar{V}_T}{r} \frac{\partial}{\partial \lambda} \right) \zeta' - \frac{\partial \psi'}{r \partial \lambda} \frac{d\bar{\eta}}{dr} = 0 \quad [9]$$

where ψ' denotes the perturbation streamfunction, $\zeta' = \nabla^2 \psi'$ the perturbation vorticity, $\bar{v}$ the basic-state tangential velocity at

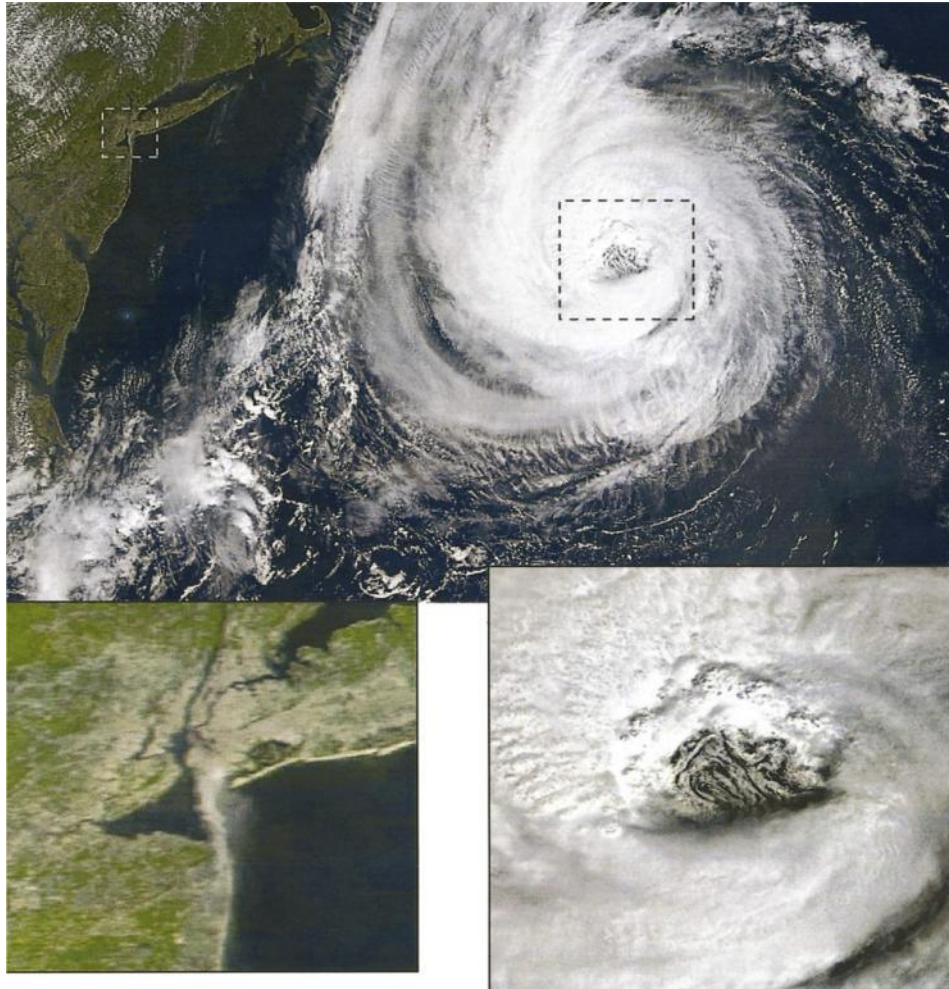


Figure 4 Moderate Resolution Imaging Spectroradiometer satellite image of Atlantic Hurricane Erin at 1515 UTC 11 September 2001. The right inset shows a magnified image of convoluted clouds in the eye. The left inset shows a smoke and dust plume being carried southward from the collapsed World Trade Center buildings, away from the populated island of Manhattan, in part by Erin's surface flow. From Kossin, J.P., McNoldy, B.D., Schubert, W.H., 2002. Vortical swirls in hurricane eye clouds. *Monthly Weather Review* 130, 3144–3149.

radius r , θ the azimuthal angle, t time, ∇^2 the horizontal Laplacian operator, $\bar{\eta} = f + \frac{1}{r} \frac{d(r\bar{V}_T)}{dr}$ the basic-state absolute vertical vorticity, $\bar{V}_T$ the tangential wind of the basic vortex, and f the Coriolis parameter. Given the fact that the outward-propagating waves of eqn [9] often possess radial length scale (l) that is small compared to the characteristic radial scale (L) of the vortex. It is thus reasonable to seek approximate sheared wave solutions near $r = R$ in the form (Montgomery and Kalenbach, 1997)

$$\psi' \approx \Psi(t) \exp\{i[n\lambda + k(t)(r - R) - \Lambda(t)]\} \quad [10]$$

where $\Psi(t)$ is a time-dependent amplitude, $k(t)$ is a time-dependent radial wave number, and $\Lambda(t)$ is a time-dependent phase. In the tightly wound limit, we can assume $kR \gg 1$ without loss of generality. With the Wentzel-Kramers-Brillouin (WKB) approximation for slowly evolving basic state, we can get the following solution for time change of radial wave number

$$\dot{k} = -n\bar{\Omega}' \quad [11]$$

whose integration yields

$$k(t) = k_0 - nt\bar{\Omega}' \quad [12]$$

where k_0 is an initial radial wave number, n is the wave number in the azimuth, and $\bar{\Omega}'(R) = \bar{V}_T'(R)/R$, prime here indicates the derivative respect to radius r , and the solution for the local dispersion relation can be expressed as

$$\omega = n\bar{\Omega} + \frac{n}{R} \frac{\bar{\zeta}'}{(k^2 + n^2/R^2)} \quad [13]$$

for a spectrally localized wave packet whose initial central wave numbers are k_0 and n , respectively. This dispersion relation is analogous to the dispersion relation for nondivergent Rossby waves on a β -plane in a uniform zonal wind. The meridional derivative of planetary vorticity is replaced by the radial derivative of basic-state storm vorticity. For a cyclonic monopole whose basic-state radial vorticity gradient is negative, the dispersion relation predicts that individual waves retrogress relative to the local angular velocity. Unlike unsheared Rossby waves on β -plane, the radial wave number is everchanging for VRWs because of the symmetrizing effect of the vortex. This has fundamental consequences on the kinematics and dynamics of vortex wave packets. Radial and azimuthal phase speeds are given by

$$C_{pr} = \frac{\omega}{k} = \frac{n}{k} \bar{\Omega} + \frac{n}{kR} \frac{\bar{\zeta}'}{(k^2 + n^2/R^2)} \quad [14]$$

$$C_{p\lambda} = \frac{\omega R}{n} = R\bar{\Omega} + \frac{\bar{\zeta}'}{(k^2 + n^2/R^2)} \quad [15]$$

The speed and direction of energy propagation is controlled by the group velocity whose radial and azimuthal components are given by

$$C_{gr} = \frac{\partial \omega}{\partial k} = \frac{-2kn\bar{\zeta}'}{R(k^2 + n^2/R^2)^2} \quad [16]$$

$$\begin{aligned} C_{g\lambda} &= \frac{\partial \omega}{\partial (n/R)} \\ &= R\bar{\Omega} + \frac{\bar{\zeta}'}{(k^2 + n^2/R^2)^2} \left[k_0^2 - \frac{n^2}{R^2} \left(1 + t^2 R^2 \bar{\Omega}'^2 \right) \right] \end{aligned} \quad [17]$$

Note that in eqns [13]–[17], k is given by eqn [12]. Since a trailing spiral wave is associated with positive k , it corresponds to positive radial group velocity and thus an outward propagation of the wave energy. As the packet propagates outward it is continuously slowed by the shearing effect that increases its radial wave number. Since for large k the radial group velocity goes as $O(k^{-3})$, the shearing effect eventually dominates and radial propagation ceases. Such a location is referred to as stagnation radius. For the instantaneous position of a wave packet, the stagnation radius r_s , on letting $t \rightarrow \infty$, can be given by

$$r_s = R + \frac{\bar{\zeta}'}{R\bar{\Omega}'} \frac{1}{(k_0 + n^2/R^2)} \quad [18]$$

The existence of a stagnation radius implies that these waves are confined to the near-vortex region and cannot radiate to infinity. Their dynamics is thus distinct from that of gravity-inertia waves that, in the absence of critical levels, radiate to infinity.

On repeating the above WKB analysis, the dispersion relation for an isolated wave packet in a divergent barotropic flow is only slightly modified by the variable deformation radius. In that case, the instantaneous wave frequency becomes

$$\omega = n\bar{\Omega} + \frac{n}{R} \frac{\bar{\xi}}{\bar{q}} \frac{\bar{q}'}{(k^2 + n^2/R^2 + \gamma^2)} \quad [19]$$

where $\bar{\xi} = f + 2\bar{V}_T/r$ is the inertia parameter, $\bar{q} = \bar{\eta}/(g\bar{H})$ is the basic-state PV, with $\bar{\eta} = f + (1/r)d(r\bar{V}_T)/dr$ being the absolute vorticity, $\gamma^2 = \bar{\eta}\bar{\xi}/(g\bar{H})$ is the square of the inverse of the local Rossby radius of deformation; the subscript zero has the same meaning as those in eqn [13]. The corresponding expressions for radial and azimuthal phase speeds are

$$C_{pr} = \frac{\omega}{k} = \frac{n}{k} \bar{\Omega} + \frac{n}{kR} \frac{\bar{\xi}}{\bar{q}} \frac{\bar{q}'}{(k^2 + n^2/R^2 + \gamma^2)} \quad [20]$$

$$C_{p\lambda} = \frac{\omega}{(n/R)} = R\bar{\Omega} + \frac{\bar{\xi}}{\bar{q}} \frac{\bar{q}'}{(k^2 + n^2/R^2 + \gamma^2)} \quad [21]$$

Corresponding expressions for the radial and azimuthal group velocities are

$$C_{gr} = \frac{\partial \omega}{\partial k} = \frac{-2kn}{R} \frac{\bar{\xi}}{\bar{q}} \frac{\bar{q}'}{(k^2 + n^2/R^2 + \gamma^2)^2} \quad [22]$$

$$\begin{aligned} C_{g\lambda} &= \frac{\partial \omega}{\partial (n/R)} = R\bar{\Omega} + \frac{\bar{\xi}}{\bar{q}} \frac{\bar{q}'}{(k^2 + n^2/R^2 + \gamma^2)^2} \\ &\quad \times \left[k_0^2 + \gamma^2 - \frac{n^2}{R^2} \left(1 + t^2 R^2 \bar{\Omega}'^2 \right) \right] \end{aligned} \quad [23]$$

The stagnation radius follows upon integrating C_{gr} with respect to time and letting $t \rightarrow \infty$:

$$r_s = R + \frac{\bar{\xi}\bar{q}'}{\bar{q}R\bar{\Omega}'} \frac{1}{(k_0 + n^2/R^2 + \gamma^2)} \quad [24]$$

The dependence of r_s on γ^2 indicates a tendency for the vorticity bands to remain closer to the core with increasing γ^2 .

These can be understood from the local dispersion relation for VRWs modified to include the baroclinic effect as given below (see Möller and Montgomery, 2000 for details).

$$\omega = n\bar{\Omega} + \frac{n}{R} \frac{\bar{\xi}}{\bar{q}} \frac{\bar{q}'(r)}{(k^2 + n^2/R^2 + m^2\gamma^2)} \quad [25]$$

Here, m is the vertical wave number, and

$$\gamma^2 = \frac{\bar{\eta}\bar{\xi}}{N^2} = \frac{I^2}{N^2} \quad [26]$$

where I is the inertial frequency, and N is the Brunt–Väisälä frequency of the azimuthal mean barotropic vortex.

The azimuthal phase speed of the waves can be written as

$$C_{p\lambda} = \frac{\omega}{n/R} = R\bar{\Omega} + \frac{\bar{\xi}}{\bar{q}} \frac{\bar{q}'(R)}{[k^2 + n^2/R^2 + m^2\gamma^2]} \quad [27]$$

The radial group velocity is

$$C_{gr} = \frac{\partial\omega}{\partial k} = -\frac{n}{R} \frac{\bar{\xi}}{\bar{q}} \frac{2k\bar{q}'(R)}{[k^2 + n^2/R^2 + m^2\gamma^2]^2} \quad [28]$$

Now the stagnation radius becomes

$$r_s = R + \frac{\bar{\xi}\bar{q}'}{\bar{q}R\bar{\Omega}'} \frac{1}{[k_0^2 + n^2/R^2 + m^2\gamma^2]} \quad [29]$$

These new formulae for baroclinic motion are very useful but caution needs to be paid to the condition in obtaining them. For example, this is a linear WKB solution. The basic state is assumed slowly varying in radial direction and barotropic in the vertical.

VRWs can be generated by various physical processes in hurricanes. In the lower troposphere, air parcels flowing inward and spiraling upward in the convective eyewall usually experience a material increase in PV. Since this material increase in PV can be especially rapid near the eyewall, an annular tower of high PV with low PV in the central eye region could be possible if there are no asymmetries in eyewall convection at the beginning, and if the material PV concentration could continue without any other processes to mix the PV into the eye region. Thus, barotropic instability near the eyewall can be expected. This type of instability can lead to eyewall breakdown and polygonal eyewall structure. When the instabilities grow to finite amplitude, the vorticity in the eyewall region pools into discrete areas, creating the appearance of polygonal eyewalls. The circulations associated with these vorticity pools exhibit a connection to mesovortices in the eyewall. At later times, the vorticity is redistributed into a nearly monopole circular vortex through eye contraction and PV mixing. In addition to barotropic instability, VRWs can be forced and driven by convective asymmetries in the eyewall. Such convective asymmetries in the eyewall can be generated by barotropic instabilities in the upper-tropospheric outflow layer, the beta effect, or environmental flow and flow shears (Wang and Wu, 2004). Neutral VRWs could also be present near the eyewall region where the radial PV gradients are large. Propagation of these forced or neutral waves along the eyewall can produce an apparent cyclonic rotation of the polygonal eyewall in some hurricanes (Kuo et al., 1999; Wang, 2002b).

The barotropic instability in the hurricane eyewall can be understood in the nondivergent barotropic model (Schubert

et al., 1999). The linearized nondivergent barotropic flow on an f -plane in cylindrical coordinates is governed by

$$\left(\frac{\partial}{\partial t} + \bar{\Omega} \frac{\partial}{\partial \lambda}\right) \nabla^2 \psi' - r \frac{\partial \psi'}{\partial \lambda} \frac{\partial \bar{\xi}}{\partial r} = 0 \quad [30]$$

where $\bar{\Omega}(r) = \frac{\bar{V}_\theta(r)}{r}$ is basic-state angular velocity and $\bar{\xi}(r) = \frac{d(r\bar{V}_\theta)}{rdr}$ is basic-state relative vorticity. Assuming a three-region distribution of vorticity

$$\bar{\xi}(r) = \begin{cases} \xi_1, & 0 \leq r \leq r_1 \\ \xi_1 + \xi_2, & r_1 < r \leq r_2 \\ 0, & r_2 < r < \infty \end{cases} \quad [31]$$

Search the solutions of the form

$$\psi'(r, \lambda, t) = \hat{\psi}(r) e^{i(m\lambda - \omega t)} \quad [32]$$

Substituting eqn [32] into eqn [30], and noting eqn [31] and if r does not cross the boundaries of the three regions, we have

$$r \frac{d}{dr} \left(r \frac{d\hat{\psi}}{dr} \right) - m^2 \hat{\psi} = 0, \quad \text{for } r \neq r_1, r_2 \quad [33]$$

Solutions of eqn [33] are first and second Bessel functions. Using the continuity conditions at r_1 and r_2 , finally we can obtain the growth rate formula

$$\frac{\nu}{\zeta_{ave}} = \frac{1}{4} \left\{ m + (m-1)\gamma \pm \left[\left(m - (m-1)\gamma - 2 \left(\frac{1-\gamma\delta^2}{1-\delta^2} \right)^2 + 4 \left(\frac{1-\gamma\delta^2}{1-\delta^2} \right) \left(\gamma - \frac{1-\gamma\delta^2}{1-\delta^2} \right) \delta^{2m} \right]^{1/2} \right\} \quad [34]$$

where $\delta = \frac{r_1}{r_2}$ is the ring thinness parameter ($\delta \rightarrow 1$ for very

thin rings), $\gamma = \frac{\xi_1}{\zeta_{ave}}$ is the hollowness parameter ($\gamma \rightarrow 0$ for very hollow rings), and ζ_{ave} is the average vorticity inside $r = r_2$

Figure 5 shows clearly the preference for higher wave numbers as the annular ring becomes thinner. Thinner annular regions would produce the highest growth rates but at much higher azimuthal wave numbers. Note that all basic states with $\gamma < 1$ satisfy the Rayleigh necessary condition for barotropic instability but that most of the region $\gamma < 1$, $\delta < 1/2$ is in fact stable. Note also the overlap in the unstable regions of γ - δ plane for different azimuthal wave numbers. For example, the lower-right area of the plane is unstable to all the azimuthal wave numbers $m = 3, 4, \dots, 8$.

The linear stability analysis is applicable to initial development of the instability to small perturbations. Once the amplitude of the disturbances increases exponentially with time initially, the subsequent evolution is nonlinear and involves wave breaking and PV mixing and pooling, appearance of mesovortices, merging and axisymmetrization, and eventually formation of monopole structure (Figure 6) or relatively stable discrete vortices (Figure 7). The appearance of wave breaking during the evolution toward a monopole vorticity distribution is hypothesized as one of the processes contributing to the polygonal eyewall structure in real hurricanes. The existence of the multiple mesovortices in the final stage of the nonlinear evolution as shown in Figure 7 can explain the formation of eyewall mesovortices in observations.

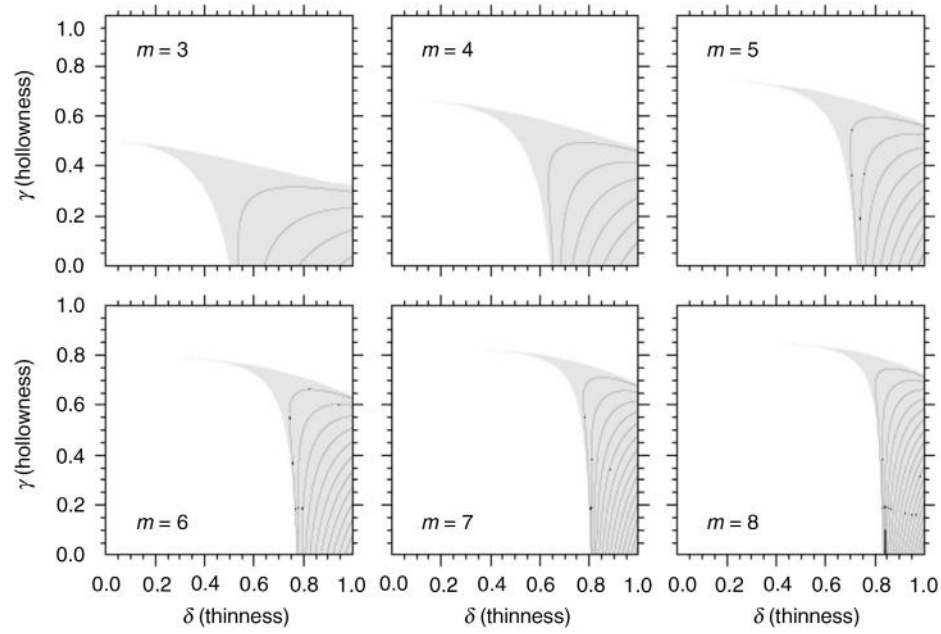


Figure 5 Isolines of the dimensionless growth rate, computed from eqn [34], as a function of δ and γ for azimuthal wave number $m = 3, 4, \dots, 8$. Positive growth rates occur only in the shaded regions. From Schubert, W.H., Montgomery, M.T., Taft, R.K., Guinn, T.A., Fulton, S.R., Kossin, J.P., Edwards, J.P., 1999. Polygonal eyewalls, asymmetric eye contraction, and potential vorticity mixing in hurricanes. *Journal of Atmospheric Sciences* 56, 1197–1223.

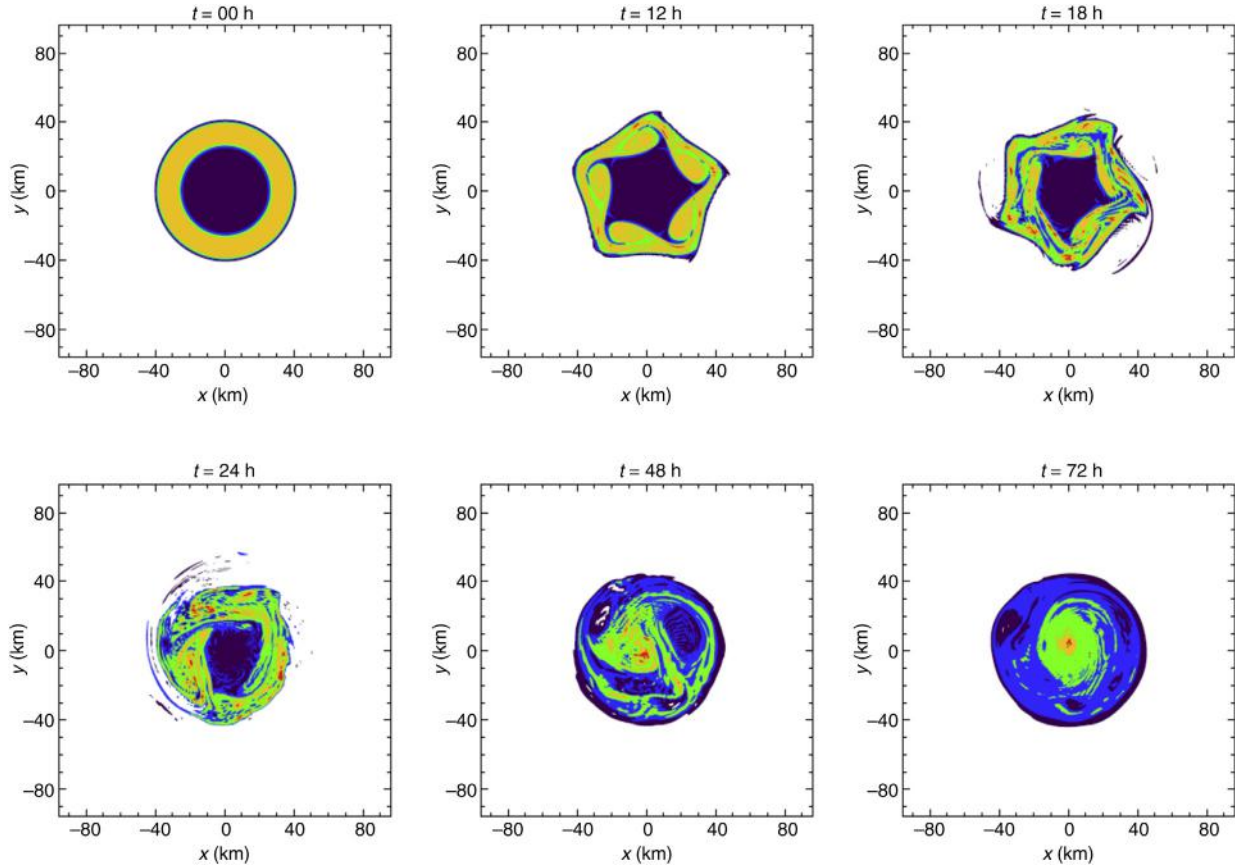


Figure 6 Vorticity contour plots for the representative nonlinear evolution of an annular vorticity ring involving into a monopole structure. Note that the initially most unstable wave number of the asymmetries follows the prediction by the linear stability (Taft, R.K., Schubert, W.H., 2013).

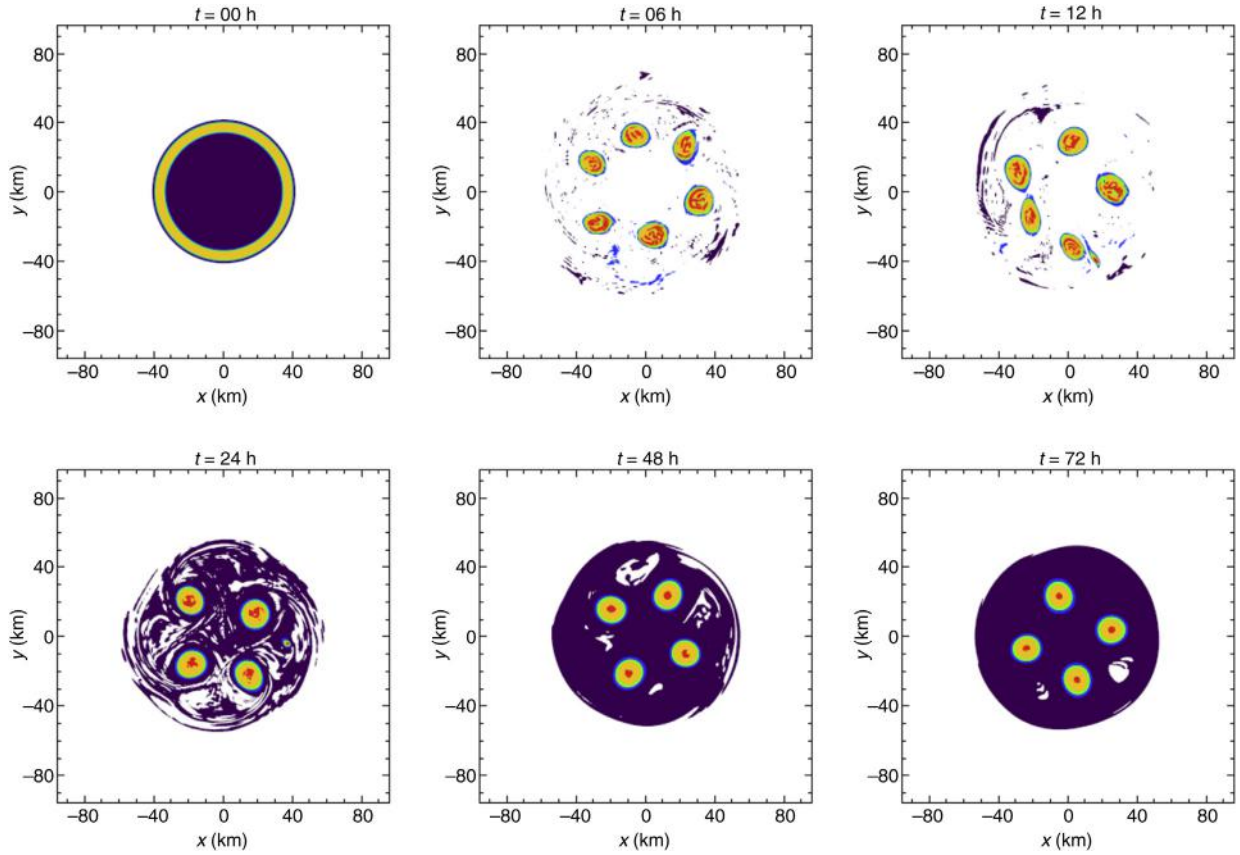


Figure 7 Vorticity contour plots for the representative nonlinear evolution of an annular vorticity ring involving into discrete mesovortices. Note that the initially most unstable wave number of the asymmetries follows the prediction by the linear stability (Taft, R.K., Schubert, W.H, 2013).

In addition to the internal dynamical instability, vertical shear of horizontal wind is considered to be one of the major environmental forcings that can cause robust asymmetric structure in hurricanes (e.g., Reasor et al., 2000; Black et al., 2002). The vertical shear-induced convective asymmetry in the inner core region is considered to be negative to hurricane intensification and maximum potential intensity (MPI). The development of convective asymmetries in the inner core region of a hurricane embedded in a vertically sheared environmental flow is often attributed to the relative flow across the elevated cyclonic vorticity or PV core (Wang and Holland, 1996a; Frank and Ritchie, 2001; Black et al., 2002; Zhang and Kieu, 2005). Both observations and numerical simulations show enhanced upward motion and convection downshear left of the hurricane center and suppressed upshear right when facing downshear (Figure 8).

Vertical wind shear can often cause the vertical tilt of the hurricane vortex. However, strong hurricanes can still maintain entity in weak and moderate vertical wind shear. This vertical alignment can be understood as an initially barotropic vortex embedded in a vertical shear on an f -plane (Figure 9). A vertically tilted vortex can be considered to consist of a core of positive vorticity surrounded by a skirt of lesser positive vorticity align through projection of the tilt asymmetry onto VRWs and their subsequent damping (Reasor et al., 2004). The VRW damping mechanism provides

a direct means of reducing the tilt of intense hurricane vortices in unidirectional vertical shear (Figure 9). For initially upright, hurricane-like vortices in vertical shear, a ‘downshear-left’ tilt equilibrium is achieved when the VRW damping is effective. Therefore, the inviscid damping mechanism intrinsic to the dry adiabatic dynamics of a hurricane vortex can suppress departure from the upright state of a strong hurricane. Figure 10 shows examples of the evolution of two vortices embedded in a westerly shear. For a broad vortex that has vorticity skirt, the VRW damping mechanism is quite effective and results in a quasi-left tilt to the vertical shear (Figure 10(a)). For a narrow vortex with little vorticity skirt, an oscillation between the upper- and lower-level centers occurs (Figure 10(b)). Therefore, the vertical alignment of a hurricane-like vortex in vertical shear is sensitive to the structure of the hurricane vortex itself. Nevertheless, if the vertical shear is too strong, the alignment would be impossible and an up-down weakening of the hurricane vortex will occur even though the moist processes are included (Frank and Ritchie, 2001; Xu and Wang, 2013).

Concentric Eyewall Cycles

In some strong hurricanes, some of the outer rainbands may strengthen and organize into a ring of convection, named an outer eyewall or a secondary eyewall, that slowly moves inward

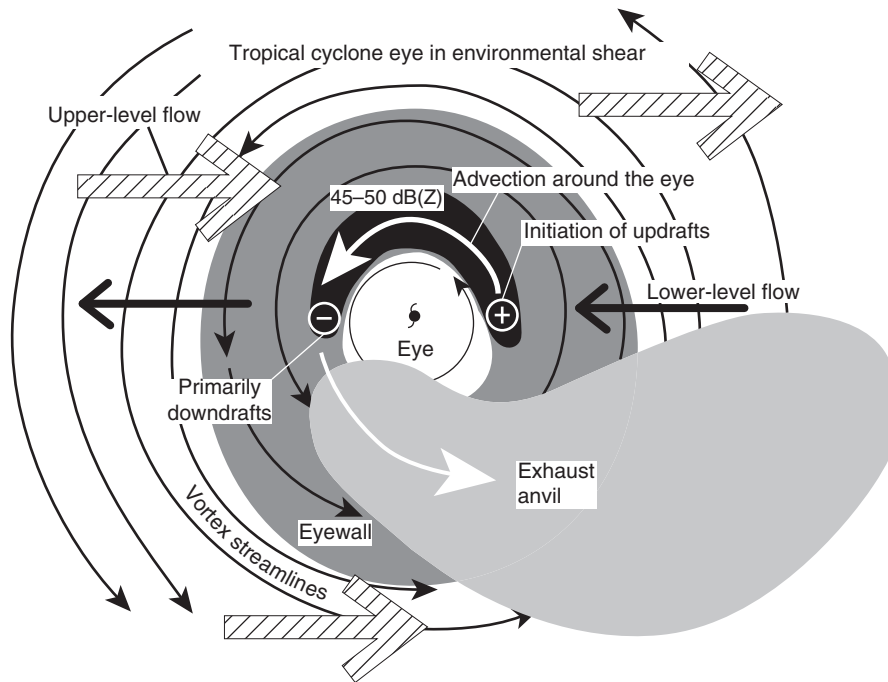


Figure 8 Schematic illustration of the shear-induced convective asymmetry based upon observations of Hurricanes Jimena and Olivia. The low-level environmental flow is indicated by the two solid black arrows. Upper-level flow is indicated by the three stippled arrows. Convective cells form somewhat upwind of the downshear side of the eyewall. They advect around the eye into the semicircle to the left of the shear vector where warm rain processes generate hydrometeors large enough to reflect radar effectively. Precipitation-driven downdrafts begin about 90° to the left of the shear vector. By the time, the cells reach the upshear side of the eyewall they have ascended through the 0 °C isotherm and downdrafts predominate below 6 km. As the cells move into the semicircle on the right of the shear vector, most condensate freezes or falls out of the active updrafts. The unloaded updrafts accelerate upward. They detach themselves from the eyewall and approach the tropopause as they rotate through the semicircle to the right of the shear. From Black, M.L., Gamache, J., Dodge, P., Barnes, G., Marks, F., Hudson, J., Castells, T., 2002. Eastern Pacific Hurricanes Jimena of 1991 and Olivia of 1994: the effect of vertical shear on structure and intensity. *Monthly Weather Review* 130, 2291–2312.

and robs the original eyewall or the inner eyewall of its needed moisture and angular momentum. This usually causes the weakening of the inner eyewall and the storm during this phase. Eventually, the outer eyewall replaces the inner one and the storm may re-intensify (an example is shown in [Figure 11](#) for Hurricane Ivan (2004), [Terwey and Montgomery, 2008](#)). As such, the development of a concentric eyewall and the subsequent eyewall replacement (or the concentric eyewall cycle) are one of the major mechanisms that often result in rapid hurricane structure and intensity changes ([Willoughby et al., 1982](#); [Houze et al., 2007](#); [Stikowski et al., 2012](#); [Kossin and Stikowski, 2012](#)).

Although the concentric eyewall cycle is important to hurricane structure and intensity changes, there have been no any complete theories that can explain all aspects of the concentric eyewall cycle so far. Several hypotheses have been proposed to explain the dynamics of the secondary (concentric) eyewall formation (SEF). Axisymmetrization of spiral rainbands is considered to be a plausible process causing the SEF ([Willoughby et al., 1982](#); [Black and Willoughby, 1992](#); [Kuo et al., 2004, 2008](#); [Wang, 2009](#); [Judt and Chen, 2010](#)). Since the spiral rainbands can be triggered by either internal dynamics or external forcing, the rainband axisymmetrization mechanism suggests that the SEF could be either an internal process or triggered by external forcing or their combination. For example,

the eddy angular momentum flux associated with an upper-tropospheric trough may initiate deep convection outside the eyewall and trigger major outer spiral rainbands and the SEF ([Molinari and Vollaro, 1990](#); [Nong and Emanuel, 2003](#); [Zhu et al., 2004](#)). Spiral rainbands can also be triggered by internal dynamics, such as VRWs, convective instability outside the eyewall, and frictional boundary layer dynamics. These are proposed to explain the dynamics of the SEF or part of the concentric eyewall cycle.

Although the secondary eyewall involves moist processes, the dynamics of its initiation has also been explained by dry adiabatic processes. The first mechanism is that associated with the activity of VRWs. The stagnation radius of outward energy dispersion of VRWs is suggested to be an ideal location for strongest wave-mean flow interaction and could trigger the formation of a local (secondary) wind maximum, which can supposedly enhance the surface heat and moist flux, helps to initiate a ring of convection, and thus the SEF ([Montgomery and Kallenbach, 1997](#); [Qiu et al., 2010](#); [Abarca and Corbosiero, 2011](#); [Martinez et al., 2011](#)).

The second mechanism is the so-called beta skirt axisymmetrization (BSA) mechanism ([Terwey and Montgomery, 2008](#)) that describes an upscale cascade of EKE in the presence of the differential rotation associated with the radial gradient of the azimuthal mean vorticity (beta skirt). By the BSA

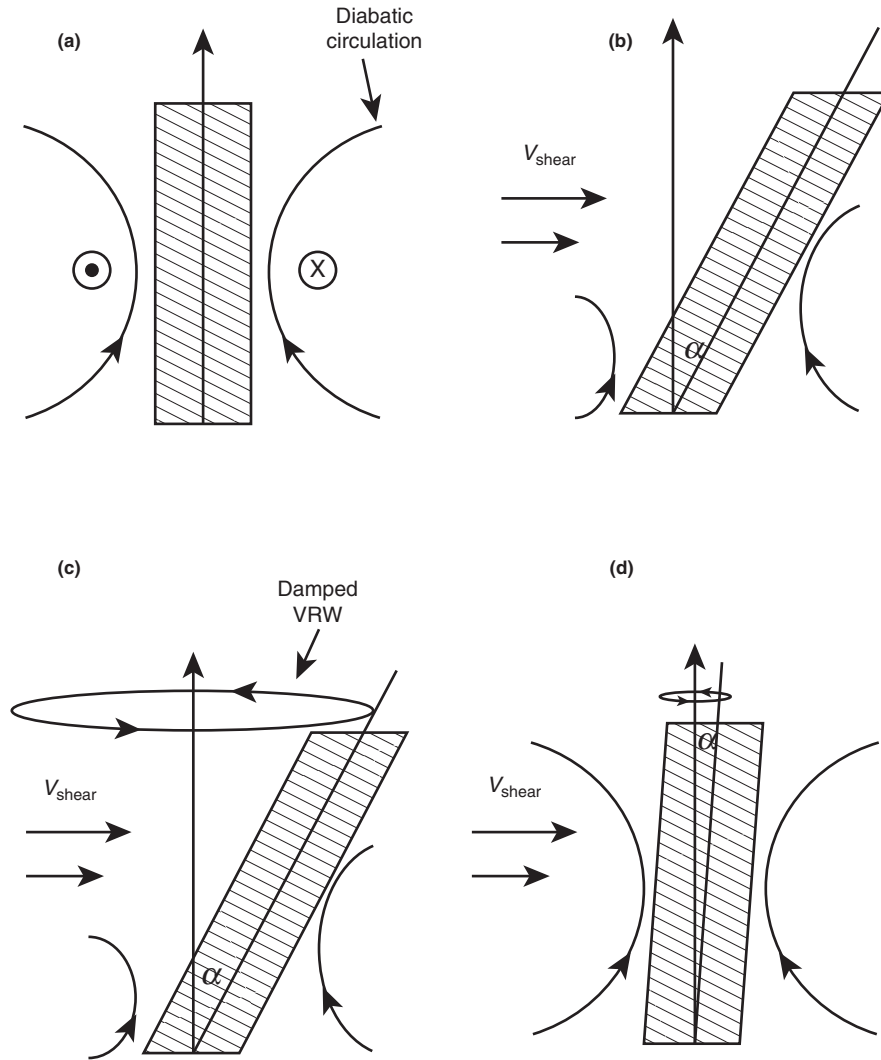


Figure 9 Schematic of the tropical cyclone alignment mechanism when the tropical cyclone is tilted by vertical shear. (a) The primary tropical cyclone circulation (into and out of the page) is maintained in the presence of frictional drag through stretching of mean tropical cyclone vorticity by the diabatically driven axisymmetric secondary circulation. (b) Vertical shear causes the tropical cyclone to tilt (with vertical tilt angle α). Eyewall convection responds by becoming increasingly asymmetric, with a maximum in the downshear quadrant of the storm. The asymmetric component of convection increases at the expense of the symmetric component shown in (a). No longer able to offset the frictional drag, the mean tropical cyclone weakens. In the absence of an intrinsic mechanism to realism, the tropical cyclone will shear apart. (c) Recent work has identified an intrinsic alignment mechanism called vortex Rossby wave (VRW) damping. VRW damping counters differential advection of the tropical cyclone by the vertical shear flow. (d) For sufficiently strong VRW damping, a quasi-aligned vortex is possible with a secondary circulation similar to the axisymmetric configuration of (a). From Reasor, P.D., Montgomery M.T., Grasso, L.D., 2004. A new look at the problem of tropical cyclones in vertical shear flow: vortex resiliency. *Journal of Atmospheric Sciences* 61, 3–22.

mechanism, in a region with a substantial overlap between the beta skirt and an area of strong convective potential outside the eyewall, cumulus convective activity acts as a source of eddy vorticity and aids in the formation of a low-level jet by forcing additional low-level inflow into a narrow annular region. Upon coupling with the boundary layer via surface enthalpy flux, the induced jet can then amplify and enforce additional cumulus convection. This positive feedback and the continuous axisymmetrization lead to the SEF.

Different from the wave-mean flow and the eddy-mean flow interaction mechanisms that focus on asymmetric processes and the subsequent axisymmetrization for the SEF, the third

mechanism is an axisymmetric view associated with the forcing in the inflow boundary layer (Wu et al., 2012; Huang et al., 2012; Bell et al., 2012; Wang et al., 2013; Abarca and Montgomery, 2013; Kepert, 2013). By this mechanism, the SEF is described as a sequence of structure changes in the outer core region starting from the broadening of the tangential winds above the inflow boundary layer. This is followed by an increase of the radial inflow in the boundary layer and super-gradient winds in and just above the boundary layer, leading to an enhancement of convergence in the boundary layer where convective ring is triggered and the secondary eyewall forms. The activity of spiral rainbands contributes to the broadening

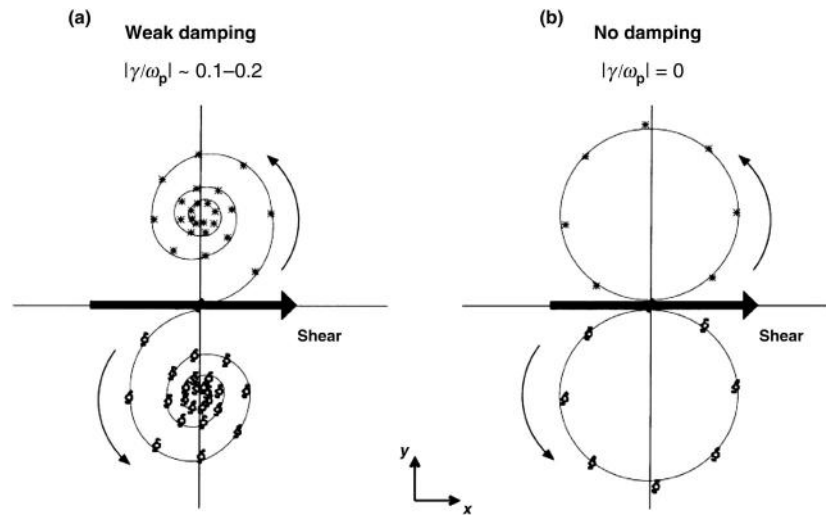


Figure 10 Schematic illustration of the vortex tilt evolution for the case where (a) the vortex Rossby wave (VRW) damping rate is a nonnegligible fraction of the precession frequency and (b) the VRW damping rate is zero. The initially aligned vortex is tilted by westerly vertical shear. The vortex center at upper levels is denoted by the stars and at lower levels by the hurricane symbols. The thin arrows indicate the direction of motion of the upper- and lower-level centers. The vortex in (a) achieves a steady-state tilt to the left of the vertical shear vector. In (b) the vortex tilts downshear, precesses cyclonically upshear, realigns, and then repeats this evolution. From Reasor, P.D., Montgomery M.T., Grasso, L.D., 2004. A new look at the problem of tropical cyclones in vertical shear flow: vortex resiliency. *Journal of Atmospheric Sciences* 61, 3–22.

of tangential wind above the boundary layer (Fudeyasu and Wang, 2011; Rozoff et al., 2012; Sun et al., 2013) and the hurricane wind structure determines the radial location of the boundary layer convergence and thus convective ring and the SEF (Abarca and Montgomery, 2013; Kepert, 2013). Note that diabatic heating in the stratiform precipitation region in spiral rainbands plays much more important role in the broadening and thus the SEF (Moon and Nolan, 2010; Moon et al., 2010; Fudeyasu and Wang, 2011).

In summary, the SEF is initially triggered by active spiral rainbands, in particular outer spiral rainbands. These rainbands play two important roles. On one hand, the balance response to diabatic heating in spiral rainbands is a broadening of tangential wind in the mid-lower troposphere above the boundary layer (Fudeyasu and Wang, 2011; Rozoff et al., 2012). On the other hand, outer spiral rainbands act as a barrier to the boundary layer inflow and produce downdrafts in the inflow boundary layer inside the rainbands, both suppressing eyewall convection (Wang, 2009) and terminating the intensification of the hurricane or weakening the hurricane. The reduction of diabatic heating associated with eyewall convection would reduce the boundary layer inflow into the eyewall, also favoring the broadening or outward expansion of tangential wind. The increase of the radial inflow in the boundary layer and the associated supergradient winds in and just above the boundary layer leads to an enhanced convergence in the boundary layer where the spiral rainbands are axisymmetrized to form a convective ring, leading to the SEF (Sun et al., 2013). Since the SEF is preconditioned by active spiral rainbands, any weak asymmetric forcing may be helpful for the SEF. These asymmetric forcings may be beta effect (Wang, 2006; Fang and Zhang, 2012), weak vertical wind shear (Zhu et al., 2004), and the upper-tropospheric trough (Molinari and Vollaro, 1990; Nong and Emanuel, 2003).

Dynamics of Genesis

A fundamental concept in understanding the atmospheric phenomena in the tropics is the so-called scale interaction that describes the positive/negative feedback processes among circulations at different time and spatial scales. In particular, cumulus convection and its synoptic-scale environmental flow are generally interacting in some cooperative ways. Some of these interactions are proposed to explain hurricane genesis. One of these feedbacks is the so-called conditional instability of the second kind (CISK) that describes the cooperative development of an existing synoptic-scale disturbance and the embedded convection (Charney and Eliassen, 1964; Ooyama, 1964, 1969). In this viewpoint, diabatic heating due to latent heat released by cumulus clouds enhances a preexisting synoptic cyclonic disturbance; this disturbance, in turn, through boundary layer pumping in the presence of surface friction, drives the low-level moisture convergence necessary to maintain an environment favorable for the development of cumulus convection. Since the boundary layer friction and Ekman pumping play a central role in enhancing the large-scale moisture convergence, this feedback is termed Ekman-CISK. This positive feedback can be schematically illustrated in Figure 12. In addition to the Ekman-CISK, moist convection can be organized and enhanced by waves in the tropics as well. In this case, upward motion and large-scale moisture convergence in the wave trough may trigger deep convection; latent heating due to cumulus convection in turn can then amplify the wave, and thus a positive feedback. This positive feedback is termed wave-CISK. Both Ekman-CISK and wave-CISK are frequently used to explain the development of synoptic-scale disturbances in the tropics, including the genesis and intensification of hurricanes. The CISK needs the conditionally unstable conditions in the large-scale

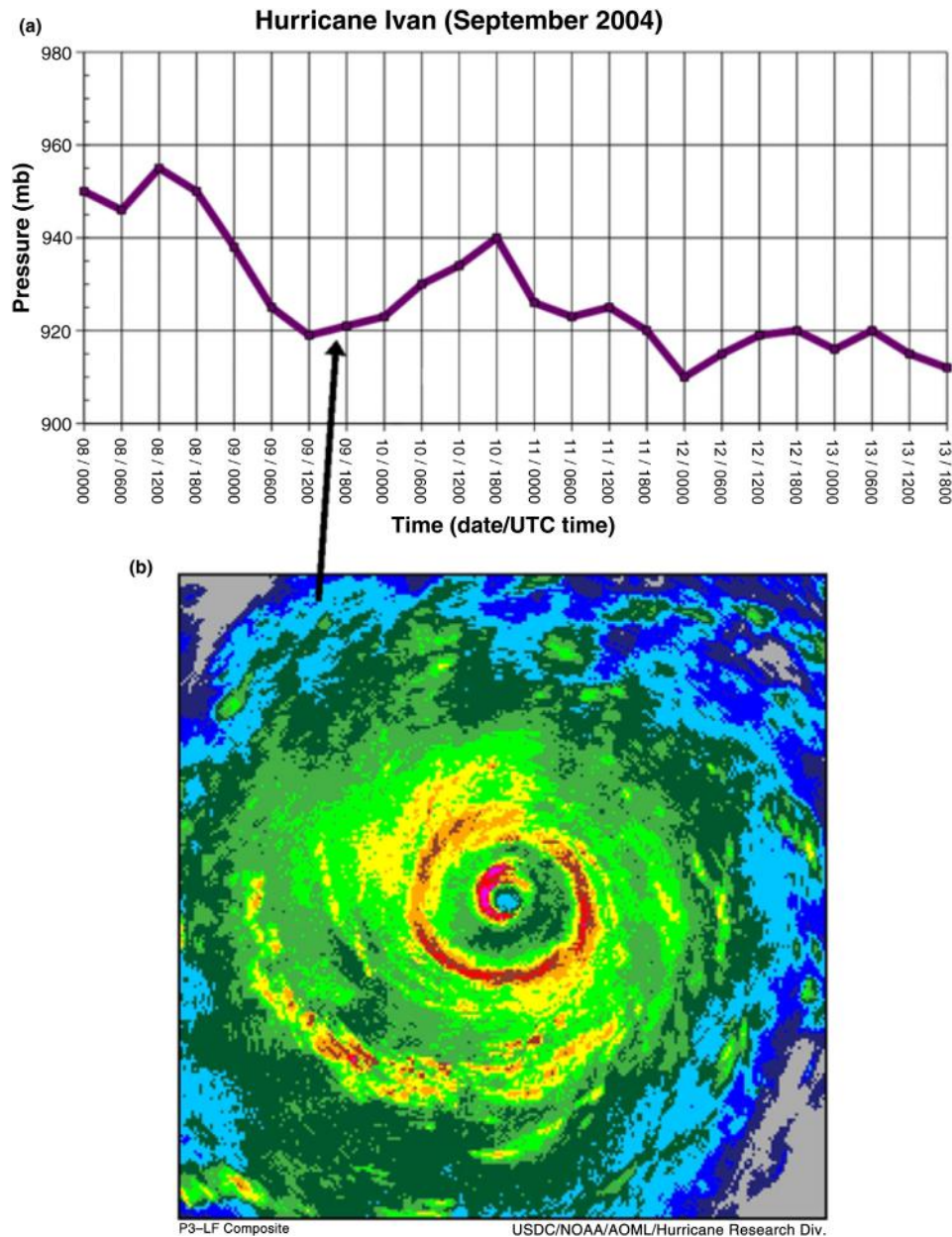


Figure 11 (a) Best track sea surface pressure trace from the National Hurricane Center for Hurricane Ivan (2004). (b) Research aircraft radar reflectivity composite of Ivan with two eyewalls. At this time, the central pressure is about to begin increasing, indicating a temporary decrease in the intensity of the storm correlated with the secondary eyewall cycle. Terwey, W.D., Montgomery, M.T., 2008. Secondary eyewall formation in two idealized, full-physics modeled hurricanes. *Journal of Geophysical Research* 113, D12112. <http://dx.doi.org/10.1029/2007JD008897>.

environment and convection is considered as an internal heating source with fundamental forcing from boundary layer moisture convergence.

Another distinct view of hurricane genesis is the air-sea interaction that describes the feedback between the surface cyclonic winds in the inner core of a prehurricane synoptic disturbance and the wind-induced surface heat exchange (WISHE). According to this viewpoint, the potential energy for the synoptic disturbance arises from the thermodynamic disequilibrium between the atmosphere and the underlying ocean (Emanuel, 1986, 2003). The efficacy of air-sea

interaction in providing potential energy to balance frictional dissipation depends on the rate of transfer of latent heat from the ocean to the atmosphere. This is a function of surface wind speed; strong winds, which produce a rough sea surface, can increase the ocean surface evaporation. Thus the system development depends on the presence of a finite-amplitude preexisting synoptic disturbance to provide the winds required to produce strong evaporation. Given a suitable initial disturbance, a feedback may occur in which an increase in inward spiraling surface winds increases moisture flux from the ocean, which by bringing the boundary layer toward saturation

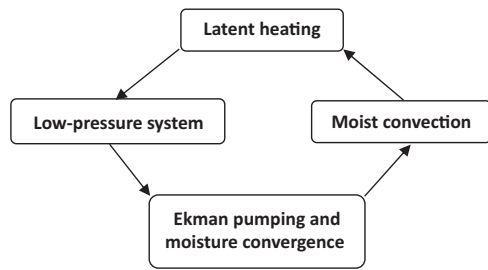


Figure 12 Schematic of a positive feedback between moist convection and synoptic-scale disturbance as described in the Ekman-CISK paradigm of hurricane genesis. Frictionally induced boundary layer convergence moistens the synoptic disturbance and destabilizes it through Ekman pumping and layer ascent. This enables small-scale plumes to reach their levels of free convection easily and to produce cumulonimbus clouds. Diabatic heating due to the resulting precipitation lowers the low-pressure system and drives the large-scale circulation and thus maintains/enhances the large-scale convergence.

increases boundary layer equivalent potential temperature (θ_e) and enhances convection. Convection plays a role in redistributing θ_e vertically along the outwardly tilted AAM surface near the RMW. This vertical redistribution of θ_e lowers the central pressure and increases radial pressure gradient and the surface wind and thus surface heat flux and boundary layer θ_e again – a positive feedback (Figure 13). WISHE theory is widely accepted to explain both genesis and the subsequent intensification of hurricanes and is better supported by observations.

Both CISK and WISHE emphasize the cooperative intensification of storm scale circulation and either cumulus convection or WISHE near the RMW under the eyewall. However, observations show the importance of extremely strong swirling convective cores, which are termed vortical hot towers (VHTs), in hurricane genesis. The VHT route to hurricane genesis describes an upscale organization process (Hendricks et al., 2004; Montgomery et al., 2006). VHTs are referred to cumulonimbus towers possessing preferred coherent structures with intense cyclonic vorticity in their cores. VHTs acquire their vertical vorticity through a combination of tilting of

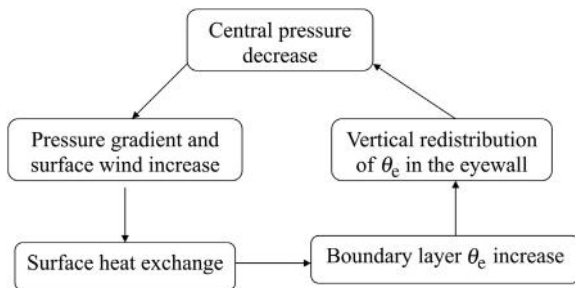


Figure 13 Schematic of a positive feedback between surface enthalpy flux and winds in a preexisting synoptic-scale disturbance as described in the WISHE paradigm of hurricane genesis. Surface heat exchange increases boundary layer θ_e , high θ_e is redistributed vertically in the outwardly tilted eyewall along the AAM near the radius of maximum wind (RMW), lowers the central pressure and increases the pressure gradient across the RMW and thus surface wind, which in turn increases the surface heat exchange.

horizontal vorticity, stretching, and diabatic heating. VHTs often have convective lifetimes in the order of an hour or so. They can overcome the adverse effect of downdrafts by consuming convective potential energy in their local environment, humidifying the mid-upper troposphere, and undergoing diabatic vortex merger with neighboring towers to increase both the strength and size of the initial mesoscale convective vortex (MCV). A quasi-balanced transverse circulation develops on the storm scale as a response to the collective diabatic heating from the ensemble or aggregate of VHTs. The forced low-level inflow can pool cyclonic vorticity of the initial MCV and small-scale vorticity anomalies generated by subsequent VHTs. This can accelerate the spinup of near-surface mean tangential winds and serves as an upscale growth mechanism for the development of a preexisting synoptic disturbance on timescales in the order of 1–2 days. In addition, VHTs themselves can be agents to largely enhance the surface entropy flux and thus the energy gain from the ocean for the parent MCV. Since WISHE describes the positive feedback between system-scale surface winds and surface entropy flux, the VHT route thus does not critically depend on the WISHE mechanism.

The VHT route above can be generalized to take into account multiscale vortices or given a term ‘multiscale vortex route’ to hurricane genesis (Sippel et al., 2006; Fudeyasu et al., 2010; Fang and Zhang, 2010, 2011). By multiscale vortex route, various scale vortices might coexist in a system-scale vortex (the preexisting synoptic-scale disturbance or an embedded incipient synoptic vortex), including MCVs (meso- α , - β , - γ scale), VHTs, and also convective bursts and convective cells or cloud clusters. Most of these vortices are vorticity-rich and show apparent self-similarities in their structures and behaviors. Each of these vortices may undergo several cycles before it merged or aggregated with its stronger neighbor. The vorticity and energy produced by these individual vorticity-rich convective vortices first saturate at convective scales that are then transferred to larger scales. As in the VHT route, the ensemble of diabatic heating released from these convective vortices acts as a persistent forcing to the quasi-balanced system-scale vortex and drives the transverse secondary circulation. The system-scale secondary circulation converges the cluster- and convective-scale vorticity anomalies into the core region of the system-scale vortex. As a result, convergence and projections of the smaller-scale vorticity to the larger scales eventually lead to the spinup of the system-scale vortex and thus the hurricane genesis (Figure 14).

Another paradigm of hurricane genesis is the so-called marsupial paradigm that emphasizes the importance of the critical layer of a tropical easterly wave to hurricane formation (Dunkerton et al., 2009; Wang et al., 2010). The theory is based on the following three hypotheses: (1) Wave breaking or roll-up of the cyclonic vorticity near the critical surface in the lower-troposphere provides a favored region for the aggregation of vorticity seedlings and TC formation; (2) The wave critical layer is a region of closed circulation, where air is repeatedly moistened by convection and protected from dry air intrusion; (3) The parent wave is maintained and possibly enhanced by diabatically amplified mesoscale vortices within the wave. Genesis tends to occur near the intersection of the critical surface and the trough axis of the precursor parent wave,

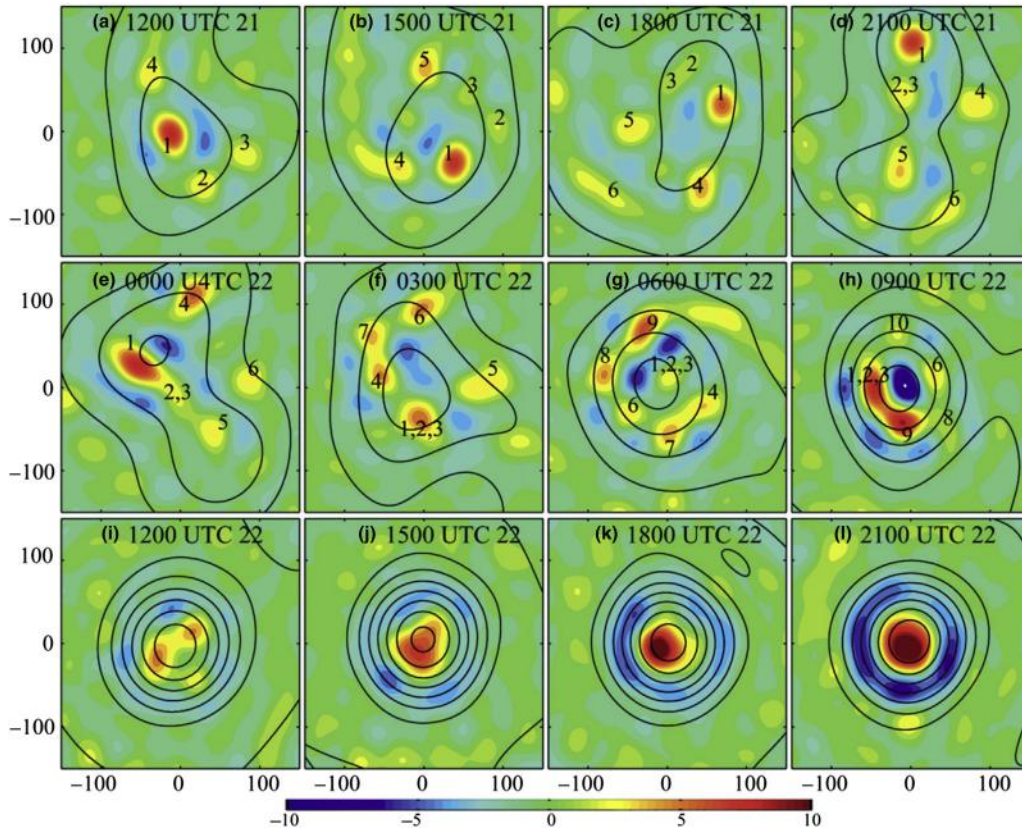


Figure 14 Evolution of the vorticity anomalies at $z = 188$ m centered on the incipient Hurricane Dolly (2006) every 3 h from 1200 UTC 21 to 2100 UTC 22 July 2006 in a cloud-resolving simulation. Color shadings represent vorticity of the anomalies with the horizontal scales larger than 50 km but smaller than 150 km. Contours represent vorticity of the anomalies with the horizontal scales larger than 150 km (every $1.5 \times 10^{-4} \text{ s}^{-1}$) and axes are in kilometers. Note that all vorticity anomalies with horizontal scales less than 50 km have not been shown for clarity. From Fang, J., Zhang, F., 2010. Initial development and genesis of Hurricane Dolly (2008). *Journal of Atmospheric Sciences* 67, 655–672.

which is the center of the pouch (see **Tropical Cyclones and Hurricanes: Tropical Cyclogenesis** for more details about this paradigm). The marsupial paradigm thus emphasizes the roles of the preexisting synoptic disturbance and the subsequent mesoscale and convective-scale processes in hurricane genesis.

Hurricane genesis is a highly nonlinear process and should be considered as a finite-amplitude amplification process (Ooyama, 1982; Emanuel, 1989). The ability of the initial convection to survive for several days and to interact with the synoptic disturbance depends on the Rossby radius of deformation that is defined as

$$L_R = \frac{NH}{f} \quad [35]$$

where N is the Brunt–Vaisala frequency, H is the scale height, and f is the Coriolis parameter. The Rossby radius of deformation is the length scale at which rotational effect becomes as important as gravity wave effect in the evolution of the flow in a disturbance. L_R is inversely proportional to latitude so it is very large in the tropics. However, in a finite-amplitude disturbance with considerable vorticity (e.g., a circular vortex or a hurricane), the Rossby radius of deformation can be largely reduced locally. In this case, the Coriolis parameter will be replaced by the inertial stability parameter and the local Rossby radius of deformation becomes

$$L_R = \frac{NH}{I} \quad [36]$$

The small local Rossby radius of deformation makes the convective heating locally to allow the low-pressure system to develop. Both CISK and WISHE require a preexisting synoptic-scale disturbance. This disturbance needs to be finite-amplitude to allow the feedback to occur. Furthermore, although we have several mechanisms that may explain the genesis processes, their relative importance is still hard to measure because of the nonlinear nature of the processes. Nevertheless, these mechanisms may function cooperatively in hurricane genesis.

Dynamics of Intensification

In the dynamics of genesis, both CISK and WISHE may explain the intensification of hurricanes as well. The dynamics of intensification of a hurricane can be further understood based on the balanced dynamics. The transverse circulation in an axisymmetric hurricane vortex, which is in both the gradient wind and hydrostatic balances, can be described using the Sawyer–Eliassen equation in balanced dynamics (Shapiro and Willoughby, 1982)

$$\frac{\partial}{\partial r} \left(\frac{A}{r\rho_0} \frac{\partial \psi}{\partial r} + \frac{B}{r\rho_0} \frac{\partial \psi}{\partial z} \right) + \frac{\partial}{\partial z} \left(\frac{C}{r\rho_0} \frac{\partial \psi}{\partial z} + \frac{B}{r\rho_0} \frac{\partial \psi}{\partial r} \right) = -\frac{\partial(\xi F_v)}{\partial z} + \frac{g}{C_p T_0} \frac{\partial Q}{\partial r} \quad [37]$$

where A , B , and C are, respectively, static stability, baroclinity, and inertial stability, and are given by $A = N^2$, $B = -\xi S = -\left(f_0 + \frac{2V_T}{r}\right) \frac{\partial V_T}{\partial z} = -\frac{g}{\theta_0} \frac{\partial \theta}{\partial r}$, and $C = \xi \eta = \left(f_0 + \frac{2V_T}{r}\right) \left(f_0 + \frac{1}{r} \frac{\partial r V_T}{\partial r}\right) = I^2$. Given the structure of the vortex, thus A , B , and C , and the distribution of heating (Q) and momentum (F_v) sources, solution of this generalized Poisson equation yields the secondary circulation when the elliptic condition $AC - B^2 > 0$ is satisfied everywhere in the vortex and proper boundary conditions are specified. The natural boundary conditions for eqn [37] are ψ vanishes at $r = 0$ and $z = 0, z_T$, and $r\psi \rightarrow 0$ as $r \rightarrow \infty$, where z_T is the top of the vortex, such as the tropopause. The above balance relation can be used to deduce the secondary (transverse) circulation from the distribution of heat and momentum sources given the axisymmetric structure of the basic hurricane vortex. Since the mass is nondivergent in the r - z plane, it may be represented with the mass streamfunction such that

$$V_r = -\frac{1}{r\rho_0} \frac{\partial \psi}{\partial z}, \quad w = \frac{1}{r\rho_0} \frac{\partial \psi}{\partial r} \quad [38]$$

Since the right-hand side of eqn [37] contains derivatives of momentum and heat sources with respect to z and r , respectively, the streamfunction field due to an isolated patch of forcing would be a dipole, namely, two adjacent counter-rotating elliptical gyres with flow between them passing through the source. For a heat source, the dipole axis that joins the foci of the ellipses is nearly horizontal. An updraft passes through the source following a (nearly vertical) surface of constant angular momentum and is flanked by compensating descent on both sides (Figure 15(a)). For a cyclonic momentum source, the dipole axis is nearly vertical with radial outflow passing through the source following an (nearly horizontal) isentropic surface and compensating inflow above and below (Figure 15(b)).

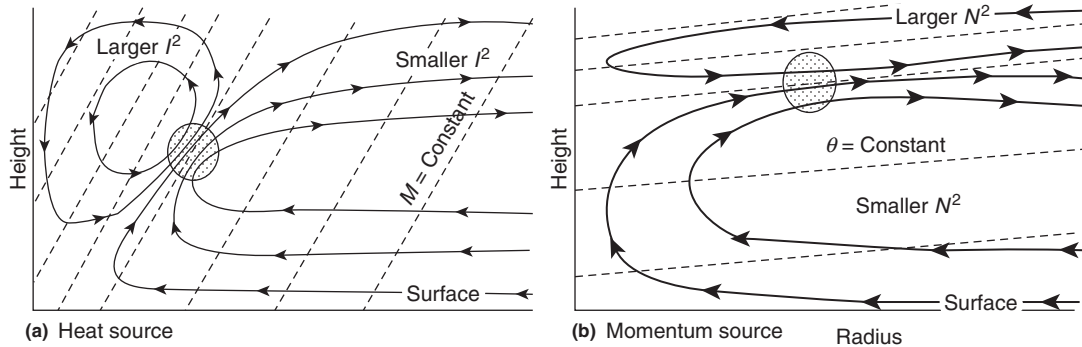


Figure 15 Secondary circulation induced in a balanced warm-core vortex by (a) a heat source and (b) a cyclonic momentum source, showing the distortion induced by variation in inertial stability I^2 , thermodynamic stability N^2 , and baroclinity B . The strong motions through the source follow lines of constant angular momentum for a heat source and of constant potential temperature for a momentum source. From Willoughby, H.E., 1995. Mature structure and evolution. In: Elsberry, R.L. (Ed.), *Global Perspectives of Tropical Cyclones*. pp. 21–62. WMO/TD-No 693. World Meteorological Organization, Geneva, 289pp.

Substitution of V_r and w into tangential momentum equation and thermodynamic equation allows calculation of $\partial V_T/\partial t$ and $\partial \theta/\partial t$ to march the solution forward in a prognostic model below

$$\frac{\partial V_T}{\partial t} + V_r \zeta_a + w \frac{\partial V_T}{\partial z} = F_v \quad [39]$$

$$\frac{g}{\theta_0} \frac{\partial \theta}{\partial t} + \frac{g}{\theta_0} V_r \frac{\partial \theta}{\partial r} + N^2 w = \frac{gQ}{C_p T_0} \quad [40]$$

where $\zeta_a = f + \frac{\partial r V_T}{\partial r}$ is the absolute vorticity. We can get the evolution of the balanced vortex provided the heat and momentum sources. Since θ is connected to pressure field through hydrostatic equation, once we get the solution for the secondary circulation (V_r, w) through the streamfunction given either heat or momentum source or both, we can readily get the tendencies of pressure and tangential wind speed. This can be used as a diagnostic approach to understand the intensification dynamics of hurricanes.

Figure 16 gives examples for the response of surface tangential wind to a specified mid-tropospheric heating source (left) and a cyclonic momentum source (right) in a balanced barotropic hurricane-like vortex. In this case, heating source centered at the RMW ($r^* = 1$) spins down the tangential winds in the eye region but spins up tangential winds near and outside the RMW (left in Figure 16). In contrast, any cyclonic momentum source imposed in the mid-troposphere near the RMW would spin up the tangential winds mainly within the RMW (right in Figure 16). In both cases, the increase in tangential winds is the maximum inside the RMW, indicating a contraction of the RMW and an intensification of the storm.

The response is sensitive to the radial position of the sources and is the strongest when the sources are slightly inside of the RMW but becomes weaker when the sources are located far away from the RMW. For the lower-tropospheric heat or momentum source at somewhat larger radius, negative tangential wind tendencies would be induced near the center or even beyond the RMW. This mechanism may explain how outer concentric eyewalls replace a preexisting inner eyewall. This may also explain the increase in storm inner-core size in

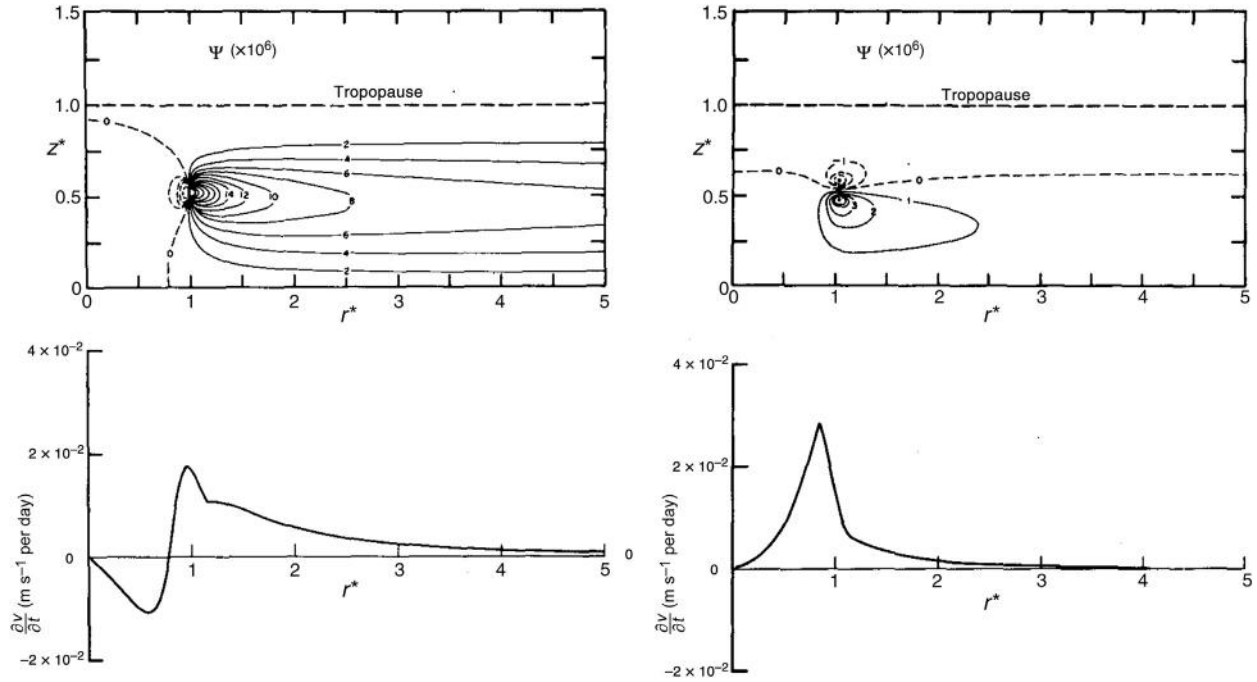


Figure 16 Responses of the transverse circulation (upper panels) and tangential wind at the surface (lower panels) to the mid-tropospheric heat (left) and cyclonic momentum (right) sources at the RMW in a balance barotropic vortex. From Shapiro, L.J., Willoughby, H.E., 1982. The response of balanced hurricanes to local sources of heat and momentum. *Journal of Atmospheric Sciences* 39, 378–394.

real hurricanes. Although the details in a baroclinic vortex could be different, the response is qualitatively similar to those shown in Figure 16 in a barotropic vortex.

The response efficiency to heat and momentum sources also depends on the intensity of the vortex itself (Figure 17). It is clear that the surface wind tendency response to heat source near the RMW increases with the intensity of the balanced vortex or equivalently with the decrease in the Rossby radius of deformation that is inversely proportional to the inertial stability. This is mainly because the heating efficiency depends on the inertial stability in the neighborhood of the heat source. As we can see from Figure 18, the warming in the inner core region of a hurricane vortex in response to heating in the eyewall is proportional to the inertial stability. This is consistent with the surface tangential tendency response to heat source. For a fixed heating rate, the surface wind tendency is greatest for heating inside the eye in vortices having the largest vorticity just inside the RMW; whereas the greatest tendency occurs for heating outside the eye for vortices with the largest vorticity outside the RMW (Pendergrass and Willoughby, 2009; Vigh and Schubert, 2009).

Since the inertial stability is a function of the maximum tangential wind speed and the RMW and in general the maximum tangential wind increases and the RMW decreases in intensifying hurricane vortex, the intensification rate for fixed heating thus would increase with increasing intensity if the vortex shape remains the same. This implies a positive feedback during intensification of a hurricane. Eyewall heating causes strengthening of the primary vortex and contraction of the eye that make further intensification more efficient. In nature,

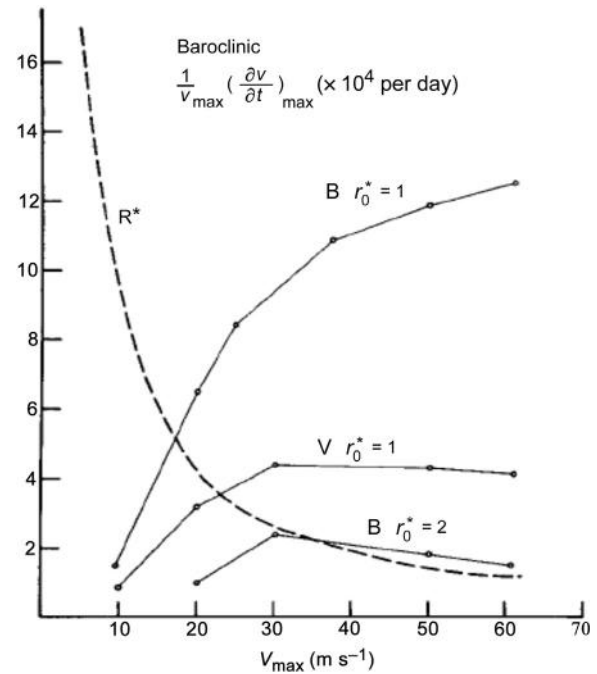


Figure 17 Normalized tendency of surface tangential wind for baroclinic vortices of different strengths. Response to heat source (B) at $r = r_m$ and $2r_m$ and to momentum source (V) at $r = r_m$ in the mid-troposphere. The dashed line shows the local nondimensional Rossby radius of deformation at the RMW. From Shapiro, L.J., Willoughby, H.E., 1982. The response of balanced hurricanes to local sources of heat and momentum. *Journal of Atmospheric Sciences* 39, 378–394.

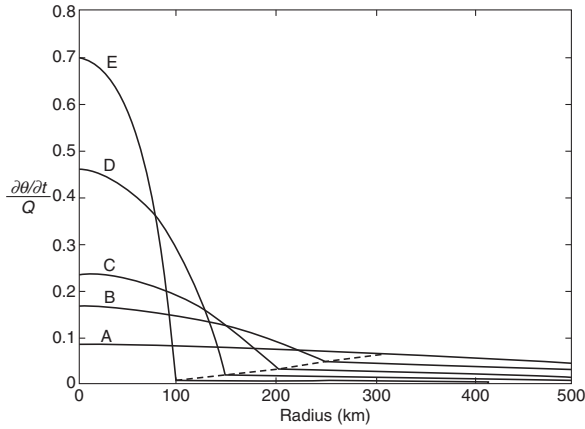


Figure 18 The radial distribution of $\partial\theta/\partial t/Q$ as a function of increasing inertial stability of the balanced vortex (A–E). From Schubert, W.H., Hack, J.J. 1982. Inertial stability and tropical cyclone development. *Journal of Atmospheric Sciences* 39, 1687–1697.

however, as a hurricane intensifies to approach its MPI, the role of heating will be balanced by the effect of surface frictional dissipation, causing the intensity to asymptote to the MPI despite greater efficiency.

The above insights into the dynamics of hurricane intensification are gained based on balanced dynamics. The unbalanced dynamics due to the presence of the frictional boundary layer also plays important roles in hurricane intensification. On the one hand, surface friction is negative to the boundary layer tangential wind tendency. On the other hand, it reduces the tangential wind speed and thus the centrifugal force, largely enhancing the boundary layer inflow and thus the inward transport of AAM and positive to the tangential wind tendency. As a result, the net effect of surface friction to hurricane intensification depends on the magnitudes of the two contributions to the tangential wind tendency. If the radial pressure gradient, thus the gradient wind, above the boundary layer is determined by the eyewall heating, the boundary layer inflow enhanced by surface friction can dominantly increase the tangential wind tendency in the inner core region of a hurricane vortex where the absolute vorticity is elevated (see eqn [39]). This process is often used to explain the presence of the supergradient wind in the upper part of the inflow boundary layer under the eyewall and a low-level jet near the top of the inflow boundary layer in the inner core region (Kepert, 2001; Kepert and Wang, 2001; Smith, 2003; Smith and Montgomery, 2010). This is also considered a second mechanism for the inner core spinup of an intensifying hurricane vortex (Bui et al., 2009; Smith et al., 2009).

Surface friction also significantly enhances the inward penetration of boundary layer inflow into the eye region and leads to the maximum in eyewall updrafts slightly insight the RMW at the top of the boundary layer (Wang and Wang, 2014), thus increasing the heating efficiency as discussed above. On the other hand, this inward penetration also contributes to the contraction of the eyewall in addition to the formation of supergradient wind and low-level jet near the top of the inflow boundary layer in the inner core region. Therefore, surface friction is important to the spinup of the inner core and

intensification and intensity of hurricanes (Bui et al., 2009; Smith et al., 2009; Smith and Montgomery, 2010).

In addition to the axisymmetric dynamics, hurricane genesis and intensification also involve asymmetric dynamics (Nolan and Farrell, 1999; Nolan et al., 2007). This should be more realistic but the asymmetric dynamics is neither necessary nor sufficient for hurricane genesis and intensification. Actually, the asymmetries in the inner core region of a hurricane vortex may play different roles in the different stage of the development of the hurricane (Möller and Montgomery, 2000; Wu and Braun, 2004; Yang et al., 2007; Persing et al., 2013). It seems that asymmetries in the core could contribute positively to the genesis and early intensification but negatively to the subsequent RI and the final maximum intensity.

Dynamics of MPI

A hurricane can be viewed as a Carnot heat engine with its energy cycle schematically shown in Figure 19 (Emanuel, 1986, 1991). The classical Carnot cycle in thermodynamics can be summarized based on the four legs. Leg 1 represents the isothermal expansion, namely gas is been heated and pressure is reduced, namely θ_e increases. Leg 2 represents an adiabatic expansion where there is no heat exchange and thus pressure is reduced but temperature is decreasing, keeping θ_e constant. Leg 3 represents an isothermal compression where gas is cooled, with pressure increased and thus θ_e decreases. Leg 4 represents an adiabatic compression where there is no heat exchange and pressure is increased while θ_e keeps constant. The mechanical work from the Carnot engine can be determined by

$$W = Q \left(\frac{SST - T_{out}}{SST} \right) \quad [41]$$

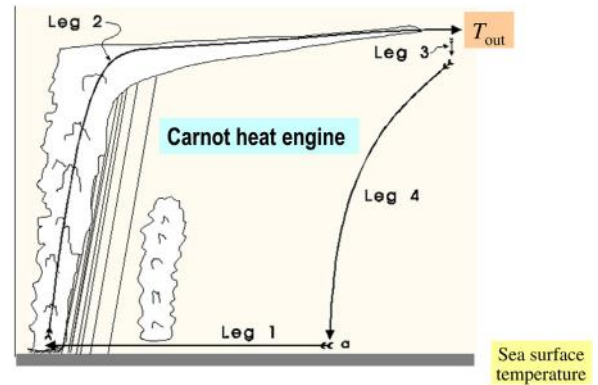


Figure 19 Schematic of the hurricane viewed as a Carnot engine of the classical Carnot cycle in thermodynamics. Leg 1 represents the isothermal expansion, namely gas is been heated and pressure is reduced, namely θ_e increases. Leg 2 represents an adiabatic expansion where there is no heat exchange and thus pressure is reduced but temperature is decreasing, keeping θ_e constant. Leg 3 represents an isothermal compression where gas is cooled, with pressure increased and thus θ_e decreases. Leg 4 represents an adiabatic compression where there is no heat exchange and pressure is increased while θ_e keeps constant. From Emanuel, K.A., 1991. The theory of hurricanes. *Annual Review of Fluid Mechanics* 23, 179–196.

where Q is the heat energy input from leg 1 and $\varepsilon = \frac{SST - T_{out}}{SST}$ is the thermodynamic efficiency of the heat engine, where SST is sea surface temperature and T_{out} is the temperature in the outflow layer.

For the hurricane case, the energy imports from the underlying ocean and exports in the outflow layer, giving the efficiency of the heat engine, and system loses mechanical energy to the surface due to friction (Emanuel, 1997). The net energy input gained for mechanical work thus can be written as

$$\text{Energy}_{\text{input}} = \int \varepsilon C_k \rho |\vec{V}| (k_o^* - k_a) r dr \quad [42]$$

where C_k is the enthalpy transfer coefficient at the surface, k_o^* is the saturation enthalpy at the SST, k_a is the enthalpy of the air in the well-mixed boundary layer,

$$k = [C_p(1 - q_v) + C_l q_v] T + L_v q \quad [43]$$

C_p and C_l are the heat capacities of dry air at constant pressure and of liquid water, respectively; $\vec{V}$ is the total near-surface wind speed, and ρ is the air density.

The mechanical energy loss to the underlying ocean from the system can be expressed as

$$\text{Dissipation} = \int C_D \rho |\vec{V}|^3 r dr \quad [44]$$

The changes of dissipation and energy input with wind speed are schematically given in Figure 20. We can see that the energy input increases with wind speed linearly while the dissipation increases with the cube of wind speed. The point where the two curves meet is the MPI the hurricane can achieve. Since the balance is mainly dominated by the winds under the eyewall, the balance can be assumed near the RMW. As a result, the MPI can be approximated by

$$V_m = \sqrt{\frac{C_k}{C_D} \varepsilon (k_o^* - k_a)} \quad [45]$$

Note that the dissipative heating due to surface friction is not included in eqns [42] and [45]. With dissipative heating considered, the factor SST/T_{out} should be multiplied to the right-hand side of eqn [42]. Therefore, compared with the result without the dissipative heating, the effect of including dissipative heating is identical to that of increasing the ratio of

surface exchange coefficient, C_k/C_d , by the factor SST/T_{out} . In hurricane environments, $SST/T_{out} \approx 3/2$, so including dissipative heating is equivalent to increasing the enthalpy transfer coefficient by 50% or to increasing the MPI by 22% (Bister and Emanuel, 1998).

The theoretical MPI discussed above considers the processes under the eyewall with a passive eye. However, the lower surface pressure in the eye provides a potential source of higher entropy air, which could be entrained into the eyewall, increasing the eyewall entropy and thus the MPI. This is proposed to explain the superintensity phenomenon in both observed and numerically simulated hurricanes whose intensity exceeds the MPI (Persing and Montgomery, 2003). It is suggested that the introduction of heat from the eye could lead to a modified Carnot cycle that allows for a stronger storm than that calculated from eqn [45]. The superintensity concept challenges the existing MPI theory. Nevertheless, the explanation by the excess energy in the eye region is found not to be sufficient since some studies indicate that the removal of high entropy anomaly in the eye only results in a reduction of about 4% in the maximum tangential wind speed, far too small to explain over 50% of the maximum tangential wind speed in observations and numerical simulations. Indeed, entropy budget analysis indicates that only less than 3% of the total entropy input to the TC comes from the eye region because of the small volume of the eye. As a result, the total magnitude of entropy transport from the eye to the eyewall is negligible to the entropy budget and the intensity of the hurricane in general (Bryan and Rotunno, 2009a).

The superintensity can be plausibly explained by energy input from outside of the eyewall region as later demonstrated in a numerical study, which shows that the surface frictional dissipation rate is about 25% higher than the energy production rate near the RMW (Wang and Xu, 2010). This indicates that the local balance hypothesis used to obtain eqn [45] would underestimate the maximum intensity of a hurricane. In contrast, the surface frictional dissipation rate is much lower than the energy production rate outside the eyewall. This implies that the excess frictional dissipation under the eyewall should be partially balanced by the energy production outside the eyewall. It is found that the energy production within about 2–2.5 times of the RMW outside the eyewall is also critical to the maximum intensity of a hurricane. This nonlocal effect is mainly accomplished by the unbalanced flow in the inflow boundary layer (Smith et al., 2008; Smith and Montgomery, 2010; Bryan and Rotunno, 2009b). Nevertheless, a complete theory including this closed energy budget for the MPI of a hurricane is yet to be developed.

Dynamics of Hurricane Motion

The motion of a hurricane is mainly controlled by the deep-layer mean environmental flow, or simply the steering flow. However, a hurricane often moves with considerable deviation from the steering flow. This deviation from the steering flow results mainly from the interaction between the hurricane vortex and the earth's vorticity gradient, namely the beta effect (variation of Coriolis parameter with latitude). Such a motion deviated from the steering flow is often termed beta drift and

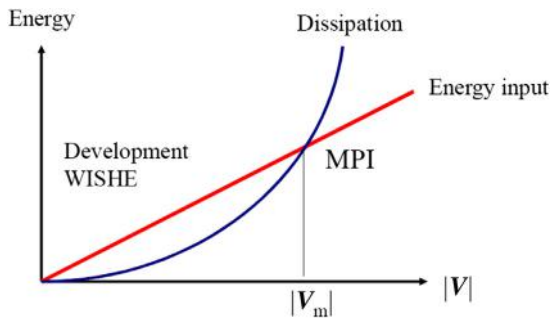


Figure 20 Schematic of changes of dissipation and energy input with wind speed in a typical hurricane. WISHE, wind-induced surface heat exchange; MPI, maximum potential intensity. Wang, Y., 2012. Recent research progress on tropical cyclone structure and intensity. Tropical Cyclone Research and Review 1, 254–275.

can be understood based on the nondivergent barotropic model that describes the conservation of absolute vorticity (Chan and Williams, 1987). The governing equation in Cartesian coordinate can be written as

$$\frac{\partial \nabla^2 \psi}{\partial t} + J(\psi, \nabla^2 \psi + f) = 0 \quad [46]$$

where ψ is streamfunction and f is the Coriolis parameter. The linearized equation with no mean flow on an equatorial beta plane can then be written

$$\frac{\partial}{\partial t} \left(\frac{\partial^2 \psi}{\partial x^2} + \frac{\partial^2 \psi}{\partial y^2} \right) + \beta \frac{\partial \psi}{\partial x} = 0 \quad [47]$$

where β is the latitudinal variation of the Coriolis parameter, and is assumed to be constant. Assuming periodic boundary conditions in both x and y directions, the Fourier transform of eqn [47] gives

$$\frac{\partial \Psi}{\partial t} - \frac{ik\beta}{k^2 + l^2} \Psi = 0 \quad [48]$$

where

$$\Psi(k, l, t) = \int_0^{2\pi} \int_0^{2\pi} \psi(x, y, t) e^{i(kx + ly)} dx dy \quad [49]$$

and, k and l are the wave numbers in x and y directions, respectively. The solution for eqn [48] is oscillatory

$$\Psi(k, l, t) = A(k, l) e^{-i\omega t} \quad [50]$$

Substituting eqn [50] into eqn [48] yields $A(k, l) = \Psi(k, l, 0)$, and

$$\omega = -\frac{k\beta}{k^2 + l^2} \quad [51]$$

The streamfunction at any time t is then given by the inverse transform of $\Psi(k, l, t)$ as

$$\psi(x, y, t) = \frac{1}{2\pi} \sum_k \sum_l A(k, l) e^{-i(kx + ly + \omega t)} \quad [52]$$

Given the initial condition $\psi(x, y, 0)$, we can then solve for $\psi(x, y, t)$ using eqns [49]–[52]. Note that since the wave number l in the y direction only appears as l^2 in eqn [51], the solution will be symmetric about the x -axis.

Given an initial distribution of the streamfunction for an initially axisymmetric vortex, the evolution is shown in Figure 21(a). The vortex is elongated westward with time as a result of the dispersive effect of Rossby waves. The dispersion relation given by eqn [51] implies that longer waves have larger westward phase speeds than shorter waves. Since the waves

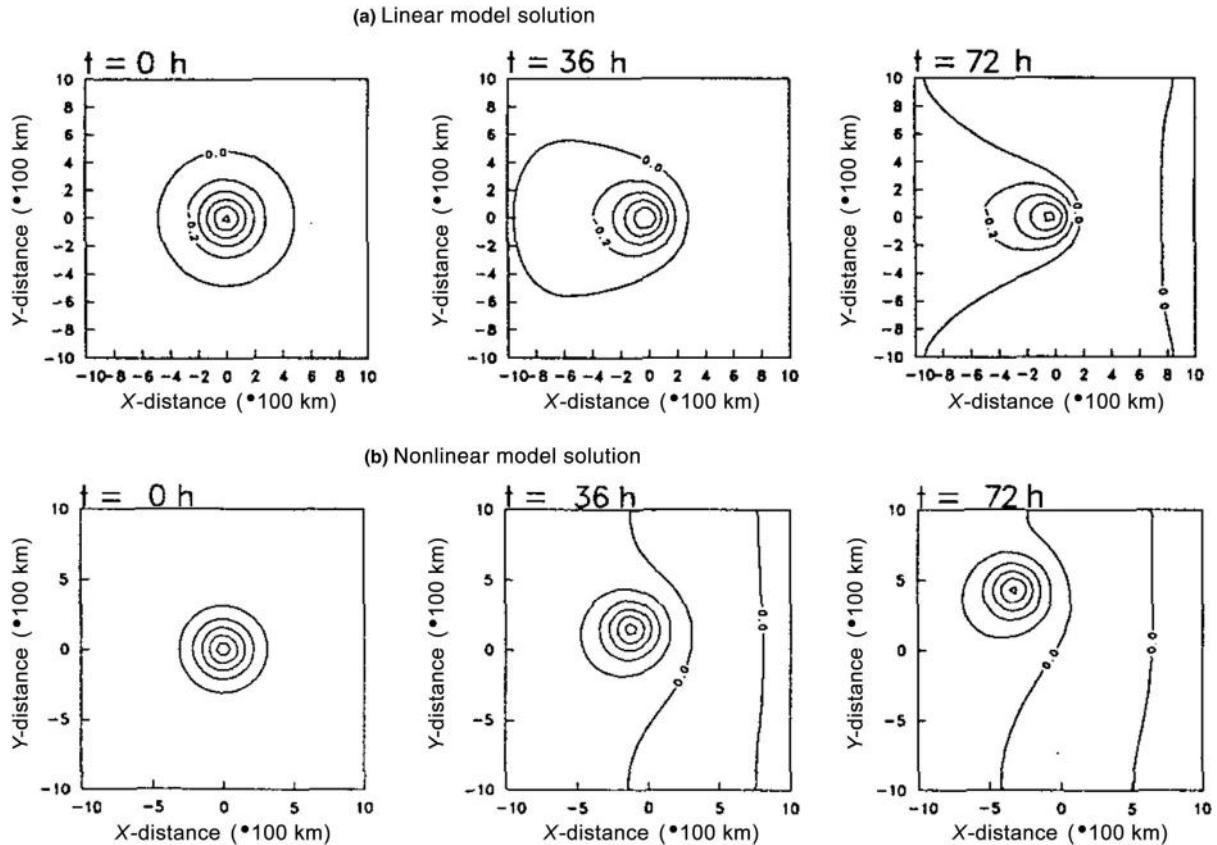


Figure 21 Streamfunction fields at 0, 36, and 72 h for the analytical (a) linear model and (b) the nonlinear numerical solution with an initially axisymmetric cyclonic vortex. Chan, J.C.L., Williams, R.T., 1987. Analytical and numerical studies of the beta-effect in tropical cyclone motion. Part I: Zero mean flow. *Journal of Atmospheric Sciences* 44, 1257–1265.

representing the outer circulation have lower-wave numbers than those representing the inner circulation, the outer part of the vortex propagates westward faster than the inner part, resulting in a stretching of the vortex westward.

The asymmetry in the streamfunction also implies an asymmetry in the wind fields. When the nonlinear advection is included, vortex will propagate poleward and westward (Figure 21(b)). This propagation is called beta drift since it is caused purely by the beta effect. The beta drift is associated with the development of the so-called ventilation flow over the vortex center as we can see from the development of asymmetric streamfunction shown in Figure 22, where the asymmetric ventilation flow advects the vortex toward the northwest (southwest) in the Northern (Southern) Hemisphere.

The developed asymmetric circulation, anticyclonic to the east and cyclonic to the west, is called the beta gyres. Its development can be understood through the conservation of

the absolute vorticity. To the east of the vortex center, the southerly brings low earth vorticity to the north, generating a negative vorticity anomaly and thus anticyclonic circulation. To the west, on the other hand, the northerly brings high earth vorticity to the south, generating positive vorticity anomaly and thus cyclonic circulation. The northerly flow between the two beta gyres advects the vorticity of the vortex poleward. On the other hand, the primary cyclonic circulation of the vortex may rotate the counter-rotating gyres cyclonically. In general, a quasi-steady solution can be reached as seen in Figure 22. As a result, the counter-rotating gyres are oriented in northeast-southwest direction and the ventilation flow is a southeasterly over the vortex center and thus a northwestward propagation of the hurricane in the Northern Hemisphere (Fiorino and Elsberry, 1989).

Since the phase speed of Rossby waves is a function of wavelength as discussed previously, it is expected that the beta

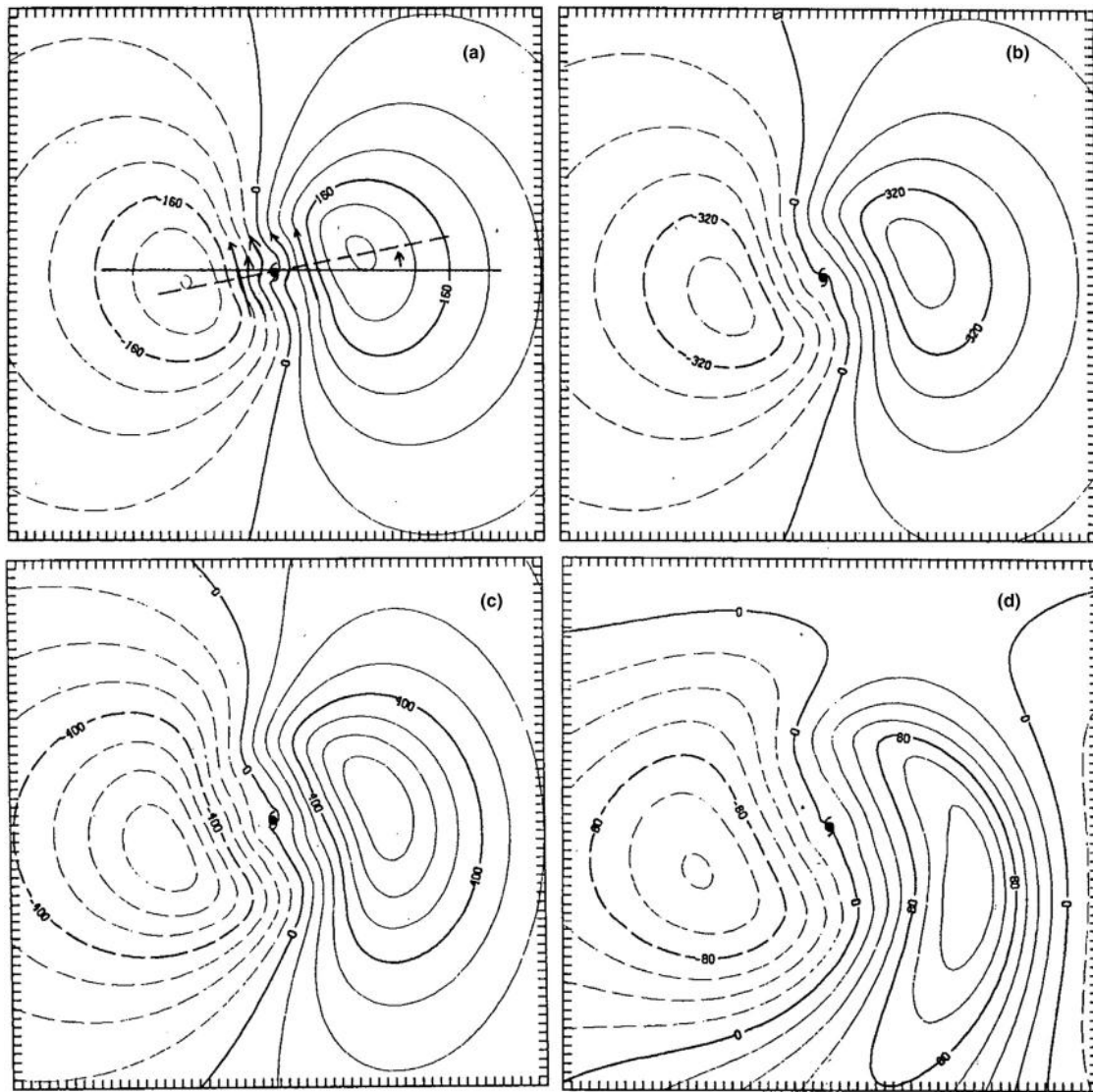


Figure 22 Temporal evolution of the asymmetric streamfunction at (a) 6 h, (b) 12 h, (c) 24 h, and (d) 72 h. The domain in each panel is 2400×2400 km centered at the vortex. Fiorino, M., Elsberry, R.L., 1989. Some aspects of vortex structure related to tropical cyclone motion. *Journal of Atmospheric Sciences* 46, 975–990.

drift would be sensitive to the outer wind profile of the initial vortex that reflects larger wavelengths. But it would not be sensitive to the inner core intensity that composes mostly the shorter wavelengths. In addition, the beta drift of a hurricane vortex described so far is based on unforced barotropic dynamics. In real atmosphere, the beta drift can be considerably affected by both the baroclinic and diabatic effects. For example, the westward motion component of the beta drift of a diabatic baroclinic hurricane vortex would be reduced by the downward penetrative effect of the upper-level anticyclonic circulation (Wang and Holland, 1996b,c). In addition to the beta drift caused by the earth's vorticity gradient, the PV gradient in the hurricane environment can lead to a similar drift of the hurricane vortex from the deep-layer mean steering flow (Shapiro, 1992).

In addition to the beta effect, vertical wind shear may also cause a hurricane vortex to deviate from the deep-layer mean flow (Wu and Emanuel, 1993). Divergent circulation driven by diabatic heating can lead to the development of an anticyclonic circulation above the deep cyclonic hurricane vortex. In a vertical wind shear, the upper-level anticyclonic anomalies would be displaced downshear relative to the lower part of the hurricane vortex. The downward penetration flow of the upper-level anticyclonic PV anomaly associated with the upper-level anticyclone will steer the lower part of the hurricane vortex to the left (right) of the shear in the Northern (Southern) Hemisphere. The magnitude of such a leftward (rightward in the Southern Hemisphere) propagation can be comparable to the beta drift and is determined by static stability, latitude, storm size and intensity, and vertical shear vector (Wu and Emanuel, 1993).

See also: Dynamical Meteorology: Rossby Waves. Tropical Cyclones and Hurricanes: Hurricane Predictability; Hurricanes: Observation; Tropical Cyclogenesis.

References

- Abarca, S.F., Corbosiero, K.L., 2011. Secondary eyewall formation in WRF simulation of Hurricanes Rita and Katrina (2005). *Geophys. Res. Lett.* 38, L07802. <http://dx.doi.org/10.1029/2011GL047015>.
- Abarca, S.F., Montgomery, M.T., 2013. Essential dynamics of secondary eyewall formation. *J. Atmos. Sci.* 70, 3216–3230.
- Bell, M.M., Montgomery, M.T., Lee, W.-C., 2012. An axisymmetric view of concentric eyewall evolution in Hurricane Rita (2005). *J. Atmos. Sci.* 69, 2414–2432.
- Bister, M., Emanuel, K.A., 1998. Dissipative heating and hurricane intensity. *Meteor. Atmos. Phys.* 50, 233–240.
- Black, M.L., Willoughby, H.E., 1992. The concentric eyewall cycle of Hurricane Gilbert. *Mon. Wea. Rev.* 120, 947–957.
- Black, M.L., Gamache, J., Dodge, P., Barnes, G., Marks, F., Hudson, J., Castells, T., 2002. Eastern Pacific Hurricanes Jimena of 1991 and Olivia of 1994: the effect of vertical shear on structure and intensity. *Mon. Wea. Rev.* 130, 2291–2312.
- Bryan, G.H., Rotunno, R., 2009a. The influence of near-surface, high-entropy air in hurricane eyes on maximum hurricane intensity. *J. Atmos. Sci.* 66, 148–158.
- Bryan, G.H., Rotunno, R., 2009b. The maximum intensity of tropical cyclones in axisymmetric numerical model simulations. *Mon. Wea. Rev.* 137, 1770–1789.
- Bui, H.H., Smith, R.K., Montgomery, M.T., Peng, J., 2009. Balanced and unbalanced aspects of tropical cyclone intensification. *Q. J. R. Meteorol. Soc.* 135, 1715–1731.
- Chan, J.C.L., Williams, R.T., 1987. Analytical and numerical studies of the beta-effect in tropical cyclone motion. Part I: Zero mean flow. *J. Atmos. Sci.* 44, 1257–1265.
- Charney, J.G., Eliassen, A., 1964. On the growth of the hurricane depression. *J. Atmos. Sci.* 21, 68–75.
- Chen, Y., Yau, M.Y., 2001. Spiral bands in a simulated hurricane. Part I: Vortex Rossby wave verification. *J. Atmos. Sci.* 58, 2128–2145.
- Dunkerton, T.J., Montgomery, M.T., Wang, Z., 2009. Tropical cyclogenesis in a tropical wave critical layer: easterly waves. *Atmos. Chem. Phys.* 9, 5587–5646.
- Emanuel, K.A., 1986. An air-sea interaction theory for tropical cyclones. Part I: Steady state maintenance. *J. Atmos. Sci.* 43, 585–604.
- Emanuel, K.A., 1989. The finite-amplitude nature of tropical cyclogenesis. *J. Atmos. Sci.* 46, 3431–3456.
- Emanuel, K.A., 1991. The theory of hurricanes. *Annu. Rev. Fluid Mech.* 23, 179–196.
- Emanuel, K.A., 1997. Some aspects of hurricane inner-core dynamics and energetics. *J. Atmos. Sci.* 54, 1014–1026.
- Emanuel, K.E., 2003. Tropical cyclones. *Annu. Rev. Earth Planet. Sci.* 31, 75–104.
- Fang, J., Zhang, F., 2010. Initial development and genesis of Hurricane Dolly (2008). *J. Atmos. Sci.* 67, 655–672.
- Fang, J., Zhang, F., 2011. Evolution of multiscale vortices in the development of Hurricane Dolly (2008). *J. Atmos. Sci.* 68, 103–122.
- Fang, J., Zhang, F., 2012. Effect of beta shear on simulated tropical cyclones. *Mon. Wea. Rev.* 140, 3327–3346.
- Fiorino, M., Elsberry, R.L., 1989. Some aspects of vortex structure related to tropical cyclone motion. *J. Atmos. Sci.* 46, 975–990.
- Frank, W.M., Ritchie, E.A., 2001. Effects of vertical wind shear on the intensity and structure of numerically simulated hurricanes. *Mon. Wea. Rev.* 129, 2249–2269.
- Fudeyasu, H., Wang, Y., Satoh, M., Nasuno, T., Miura, H., Yanase, W., 2010. Multiscale interactions in the lifecycle of a tropical cyclone simulated in a global cloud-system resolving model. Part II: Mesoscale and storm-scale processes. *Mon. Wea. Rev.* 138, 4305–4327.
- Fudeyasu, H., Wang, Y., 2011. Balanced contribution to the intensification of a tropical cyclone simulated in TCM4: outer core spin-up process. *J. Atmos. Sci.* 68, 430–449.
- Guinn, T.A., Schubert, W.H., 1993. Hurricane spiral bands. *J. Atmos. Sci.* 50, 3380–3403.
- Hendricks, E.A., Montgomery, M.T., Davis, C.A., 2004. The role of “vortical” hot towers in the formation of tropical cyclone Diana (1984). *J. Atmos. Sci.* 61, 1209–1232.
- Holland, G.J., Merrill, R.T., 1984. On the dynamics of tropical cyclone structural changes. *Q. J. R. Meteorol. Soc.* 110, 723–745.
- Houze, R.A., Chen, S.S., Smull, B.F., Lee, W.-C., Bell, M.M., 2007. Hurricane intensity and eyewall replacement. *Science* 315, 1235–1239.
- Huang, Y.H., Montgomery, M.T., Wu, C.-C., 2012. Concentric eyewall formation in typhoon Sinlaku (2008). Part II: Axisymmetric dynamical processes. *J. Atmos. Sci.* 69, 662–674.
- Judt, F., Chen, S.S., 2010. Convectively generated potential vorticity in rainbands and formation of the secondary eyewall in Hurricane Rita of 2005. *J. Atmos. Sci.* 67, 3581–3599.
- Keptert, J.D., 2001. The dynamics of boundary layer jets within the tropical cyclone core. Part II: Linear theory. *J. Atmos. Sci.* 58, 2469–2484.
- Keptert, J.D., Wang, Y., 2001. The dynamics of boundary layer jets within the tropical cyclone core. Part II: Nonlinear enhancement. *J. Atmos. Sci.* 58, 2485–2501.
- Keptert, J.D., 2013. How does the boundary layer contribute to eyewall replacement cycles in axisymmetric tropical cyclones? *J. Atmos. Sci.* 70, 2808–2830.
- Kossin, J.P., McNoldy, B.D., Schubert, W.H., 2002. Vortical swirls in hurricane eye clouds. *Mon. Wea. Rev.* 130, 3144–3149.
- Kossin, J.P., Stikowski, M., 2012. Predicting hurricane intensity and structure changes associated with eyewall replacement cycles. *Wea. Forecasting* 27, 484–488.
- Kuo, H.-C., Williams, R.T., Chen, J.-H., 1999. A possible mechanism for the eye rotation of typhoon Herb. *J. Atmos. Sci.* 56, 1659–1673.
- Kuo, H.-C., Lin, L.-Y., Chang, C.-P., Williams, R.T., 2004. The formation of concentric vorticity structures in typhoons. *J. Atmos. Sci.* 61, 2722–2734.
- Kuo, H.-C., Schubert, W.H., Tsai, C.-L., Kuo, Y.-F., 2008. Vortex interactions and barotropic aspects of concentric eyewall formation. *Mon. Wea. Rev.* 136, 5183–5198.
- MacDonald, N.J., 1968. The evidence for the existence of Rossby-like waves in the hurricane vortex. *Tellus* 20, 138–150.
- Martinez, Y., Brunet, G., Yau, M.K., Wang, X., 2011. On the dynamics of concentric eyewall genesis: space-time empirical normal modes diagnosis. *J. Atmos. Sci.* 68, 457–476.
- Molinari, J., Vollaro, D., 1990. External influences on hurricane intensity. Part II: Vertical structure and response of the hurricane vortex. *J. Atmos. Sci.* 47, 1902–1918.
- Möller, J.D., Montgomery, M.T., 2000. Tropical cyclone evolution via potential vorticity anomalies in a three-dimensional balance model. *J. Atmos. Sci.* 57, 3366–3387.

- Montgomery, M.T., Kallenbach, R.J., 1997. A theory for vortex Rossby waves and its application to spiral bands and intensity changes in hurricanes. *Q. J. R. Meteorol. Soc.* 123, 435–465.
- Montgomery, M.T., Nicholls, M.E., Cram, T.A., Saunderson, A.B., 2006. A vertical hot tower route to tropical cyclogenesis. *J. Atmos. Sci.* 63, 355–386.
- Moon, Y., Nolan, D.S., 2010. The dynamic response of the hurricane wind field to spiral rainband heating. *J. Atmos. Sci.* 67, 1779–1805.
- Moon, Y., Nolan, D.S., Iskandarani, M., 2010. On the use of two-dimensional incompressible flow to study secondary eyewall formation in tropical cyclones. *J. Atmos. Sci.* 67, 3765–3773.
- Nolan, D.S., Farrell, B.F., 1999. The intensification of two-dimensional swirling flows by stochastic asymmetric forcing. *J. Atmos. Sci.* 56, 3937–3962.
- Nolan, D.S., Moon, Y., Stern, D.P., 2007. Tropical cyclone intensification from asymmetric convection: energetics and efficiency. *J. Atmos. Sci.* 64, 3377–3405.
- Nong, S., Emanuel, K.A., 2003. A numerical study of the genesis of concentric eyewalls in hurricanes. *Q. J. R. Meteorol. Soc.* 129, 3323–3339.
- Ooyama, K.V., 1964. A dynamical model for the study of tropical cyclone development. *Geophys. J. Int.* 4, 187–198.
- Ooyama, K.V., 1969. Numerical simulation of the life cycle of tropical cyclones. *J. Atmos. Sci.* 26, 3–40.
- Ooyama, K.V., 1982. Conceptual evolution of the theory and modeling of the tropical cyclone. *J. Meteorol. Soc. Japan* 60, 369–379.
- Pendergrass, A.G., Willoughby, H.E., 2009. Diabatically induced secondary flows in tropical cyclones. Part I: Quasi-steady forcing. *Mon. Wea. Rev.* 137, 805–821.
- Persing, J., Montgomery, M.T., 2003. Hurricane superintensity. *J. Atmos. Sci.* 60, 2349–2371.
- Persing, J., Montgomery, M.T., McWilliams, J., Smith, R.K., 2013. Asymmetric and axisymmetric dynamics of tropical cyclones. *Atmos. Chem. Phys.* 13, 12299–12341.
- Qiu, X., Tan, Z.-M., Xiao, Q., 2010. The roles of vortex Rossby waves in hurricane secondary eyewall formation. *Mon. Wea. Rev.* 138, 2092–2109.
- Reasor, P.D., Montgomery, M.T., Marks Jr., F.D., Gamache, J.F., 2000. Low-wave-number structure and evolution of the hurricane inner core observed by airborne dual-doppler radar. *Mon. Wea. Rev.* 128, 1653–1680.
- Reasor, P.D., Montgomery, M.T., Grasso, L.D., 2004. A new look at the problem of tropical cyclones in vertical shear flow: vortex resiliency. *J. Atmos. Sci.* 61, 3–22.
- Rozoff, C.M., Schubert, W.H., McNoldy, B.D., Kossin, J.P., 2006. Rapid filamentation zones in intense tropical cyclones. *J. Atmos. Sci.* 63, 325–340.
- Rozoff, C.M., Nolan, D.S., Kossin, J.P., Zhang, F., Fang, J., 2012. The roles of an expanding wind field and inertial stability in tropical cyclone secondary eyewall formation. *J. Atmos. Sci.* 69, 2621–2643.
- Schubert, W.H., Hack, J.J., 1982. Inertial stability and tropical cyclone development. *J. Atmos. Sci.* 39, 1687–1697.
- Schubert, W.H., Montgomery, M.T., Taft, R.K., Guinn, T.A., Fulton, S.R., Kossin, J.P., Edwards, J.P., 1999. Polygonal eyewalls, asymmetric eye contraction, and potential vorticity mixing in hurricanes. *J. Atmos. Sci.* 56, 1197–1223.
- Shapiro, L.J., 1992. Hurricane vortex motion and evolution in a three-layer model. *J. Atmos. Sci.* 49, 140–153.
- Shapiro, L.J., Willoughby, H.E., 1982. The response of balanced hurricanes to local sources of heat and momentum. *J. Atmos. Sci.* 39, 378–394.
- Sippel, J.A., Nielsen-Gammon, J.W., Allen, S.E., 2006. The multiple-vortex nature of tropical cyclogenesis. *Mon. Wea. Rev.* 134, 1796–1814.
- Smith, R.K., 2003. A simple model of the hurricane boundary layer. *Q. J. R. Meteorol. Soc.* 129, 1007–1027.
- Smith, R.K., Montgomery, M.T., Vogl, S., 2008. A critique of Emanuel's hurricane model and potential intensity theory. *Q. J. R. Meteorol. Soc.* 134, 551–561.
- Smith, R.K., Montgomery, M.T., Nguyen, S.V., 2009. Tropical cyclone spin up revisited. *Q. J. R. Meteorol. Soc.* 135, 1321–1335.
- Smith, R.K., Montgomery, M.T., 2010. Hurricane boundary layer theory. *Q. J. R. Meteorol. Soc.* 136, 1665–1670.
- Stikowski, M., Kossin, J.P., Rozoff, C.M., 2012. Intensity and structure changes during hurricane eyewall replacement cycles. *Mon. Wea. Rev.* 139, 3829–3847.
- Sun, Y.Q., Jiang, Y., Tan, B., Zhang, F., 2013. The governing dynamics of the secondary eyewall formation of typhoon Sinlaku (2008). *J. Atmos. Sci.* 70, 3818–3837.
- Tervey, W.D., Montgomery, M.T., 2008. Secondary eyewall formation in two idealized, full-physics modeled hurricanes. *J. Geophys. Res.* 113, D12112. <http://dx.doi.org/10.1029/2007JD008897>.
- Vigh, J.L., Schubert, W.H., 2009. Rapid development of the tropical cyclone warm core. *J. Atmos. Sci.* 66, 3335–3350.
- Wang, H., Wang, Y., 2014. A numerical study of typhoon Megi (2010). Part I: Rapid intensification. *Mon. Wea. Rev.* 124, 29–48.
- Wang, Y., 2002a. Vortex Rossby waves in a numerically simulated tropical cyclone. Part I: Overall structure, potential vorticity, and kinetic energy budgets. *J. Atmos. Sci.* 59, 1213–1238.
- Wang, Y., 2002b. Vortex Rossby waves in a numerically simulated tropical cyclone. Part II: The role in tropical cyclones structure and intensity changes. *J. Atmos. Sci.* 59, 1239–1262.
- Wang, Y., 2006. Concentric eyewall simulated in a fully compressible, nonhydrostatic, multiply nested, movable mesh tropical cyclone model (TCM4). Preprints, 27th Conf. on Hurricanes and Tropical Meteorology, Monterey, CA. Amer. Meteor. Soc. 6B.6. Available online at https://ams.confex.com/ams/27Hurricanes/techprogram/paper_107727.htm.
- Wang, Y., 2008. Rapid filamentation zone in a numerically simulated tropical cyclone. *J. Atmos. Sci.* 65, 1158–1181.
- Wang, Y., 2009. How do outer spiral rainbands affect tropical cyclone structure and intensity? *J. Atmos. Sci.* 66, 1250–1273.
- Wang, Y., 2012. Recent research progress on tropical cyclone structure and intensity. *Tropical Cyclone Research and Review* 1, 254–275.
- Wang, Y., Holland, G.J., 1996a. Tropical cyclone motion and evolution in vertical shear. *J. Atmos. Sci.* 53, 3313–3332.
- Wang, Y., Holland, G.J., 1996b. The beta drift of baroclinic vortices. Part I: Adiabatic vortices. *J. Atmos. Sci.* 53, 411–427.
- Wang, Y., Holland, G.J., 1996c. The beta drift of baroclinic vortices. Part II: Diabatic vortices. *J. Atmos. Sci.* 53, 3737–3756.
- Wang, Y., Wu, C.-C., 2004. Current understanding of tropical cyclone structure and intensity changes—a review. *Meteor. Atmos. Phys.* 87, 257–278.
- Wang, Y., Xu, J., 2010. Energy production, frictional dissipation, and maximum intensity of a numerically simulated tropical cyclone. *J. Atmos. Sci.* 67, 97–116.
- Wang, X., Ma, Y., Davidson, N., 2013. Secondary eyewall formation and eyewall replacement cycles in a simulated hurricane: effect of the net radial force in the hurricane boundary layer. *J. Atmos. Sci.* 70, 1317–1341.
- Wang, Z., Montgomery, M.T.M., Dunkerton, T.J., 2010. Genesis of pre-Hurricane Felix (2007). Part I: The role of the easterly wave critical layer. *J. Atmos. Sci.* 67, 1711–1729.
- Willoughby, H.E., 1990. Gradient balance in tropical cyclones. *J. Atmos. Sci.* 47, 265–274.
- Willoughby, H.E., 1995. Mature structure and evolution. In: Elsberry, R.L. (Ed.), *Global Perspectives of Tropical Cyclones*. WMO/TD-No 693. World Meteorological Organization, Geneva, pp. 21–62, 289pp.
- Willoughby, H.E., Clos, J.A., Schoreibah, M.G., 1982. Concentric eyewalls, secondary wind maxima, and the evolution of the hurricane vortex. *J. Atmos. Sci.* 39, 395–411.
- Wu, C.-C., Emanuel, K.A., 1993. Interaction of a baroclinic vortex with background shear: application to hurricane movement. *J. Atmos. Sci.* 50, 62–76.
- Wu, C.-C., Huang, Y.H., Lien, G.Y., 2012. Concentric eyewall formation in typhoon Sinlaku (2008). Part I: Assimilation of T-PARC data based on the ensemble Kalman filter (EnKF). *Mon. Wea. Rev.* 140, 506–527.
- Wu, L., Braun, S.A., 2004. Effect of convective asymmetries on hurricane intensity: a numerical study. *J. Atmos. Sci.* 61, 3065–3081.
- Xu, Y.-M., Wang, Y., 2013. On the initial development of asymmetric vertical motion and horizontal relative flow in a mature tropical cyclone embedded in environmental vertical shear. *J. Atmos. Sci.* 70, 3471–3491.
- Yang, B., Wang, Y., Wang, B., 2007. The effects of internally generated asymmetries on tropical cyclone potential intensity. *J. Atmos. Sci.* 64, 1165–1188.
- Zhang, D.-L., Kieu, C.Q., 2005. Shear-forced vertical circulations in tropical cyclones. *Geophys. Res. Lett.* 32, L13822. <http://dx.doi.org/10.1029/2005GL023146>.
- Zhu, T., Zhang, D.-L., Weng, F., 2004. Numerical simulation of Hurricane Bonnie (1998). Part II: Eyewall evolution and intensity changes. *Mon. Wea. Rev.* 132, 225–241.

Hurricane Predictability

JA Sippel, National Aeronautics and Space Administration (NASA), Greenbelt, MD, USA

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

This article reviews the current understanding of tropical cyclone predictability. General concepts regarding atmospheric predictability are first presented, and are then related to tropical cyclones. A central theme is that error growth associated with moist convection limits the intrinsic predictability of tropical cyclone intensity, which is similar to what limits predictability in the midlatitudes. Meanwhile, practical predictability is limited by convective amplification of errors in specifying the initial cyclone vortex, the background state, particularly near gradients, and in the aerosol concentration. Predictability is case dependent and appears to be lower in environments with higher convective instability and wind shear.

Introduction

Understanding sources of tropical cyclone forecast uncertainty and error growth is essential for future improvement of hurricane forecasts. Although some aspects of tropical cyclone forecasts have been improved considerably during the past decade, they are still plagued by significant errors. While today's 48-h position forecast is as accurate as the 24-h forecast 10 years ago, there has been little improvement in intensity forecasts (Figure 1). In addition, forecasting cyclogenesis remains a challenge, and predictions of rapid intensification (RI) and decay remain particularly problematic. The intent of this article is to serve as a review of the current understanding of tropical cyclone predictability.

To better understand tropical cyclone predictability, one must first understand some fundamental concepts of general atmospheric predictability. Previous research has elucidated two types of predictability: practical predictability is the ability to predict based on currently available procedures, and intrinsic predictability is the extent to which prediction is possible if an optimum procedure is used. Practical predictability is limited by uncertainties in the forecast model and initial conditions, which are presently quite large. Constraints on practical predictability include the sufficiency of observations (including accuracy, quantity, and usability), data assimilation, and modeling. Meanwhile, intrinsic predictability is that which can be attained

with nearly perfect knowledge of both the atmospheric state and physics.

Both practical and intrinsic predictability are flow-dependent and are limited during regime transition. For example, the large-scale atmosphere can be likened to the Lorenz model, as different flow regimes are similar to the famous Lorenz attractors (e.g., Figure 2). Most of the unpredictability in the model arises from unsteady behavior between the two attractors, and the atmosphere likewise lacks predictability between regimes.

These ideas appear to be applicable to mesoscale flows as well. For example, recent work has shown that small variations in initial conditions can produce very different realizations of a midlatitude squall line; a generalization of this is shown in Figure 3(a). Though initial differences might be far less than typical analysis error, some realizations develop a severe squall line after 12–18 h, while others generally demonstrated little or no linear development. Particularly large differences occurred over a very small spectrum of initial condition perturbations, similar to the unstable region in Figure 2. In summary, practically immeasurable error can grow rapidly when the background state straddles very different regimes, particularly in the presence of moist convection.

The growth of small-scale error has been a subject of a number of other recent studies. Generally speaking, small-scale and small-amplitude initial condition perturbations may lead to substantial mesoscale differences from 24 to 36 h

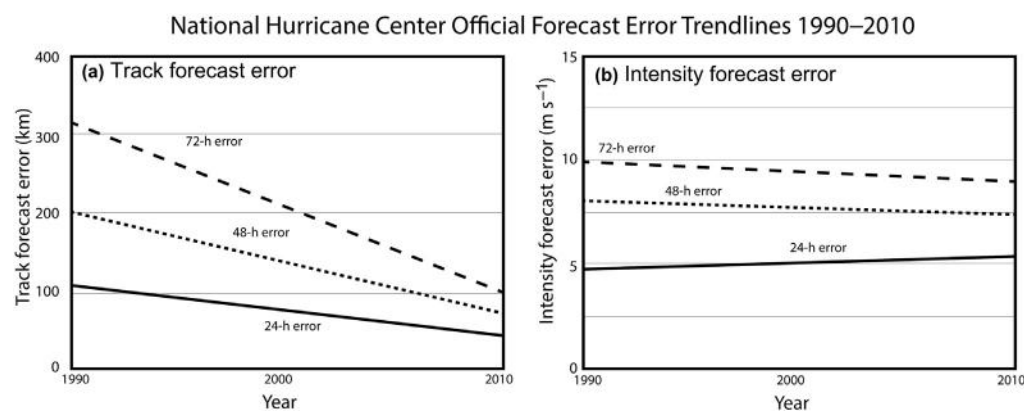


Figure 1 The evolution of National Hurricane Center forecast error in terms of tropical cyclone (a) track and (b) maximum 10-m winds.

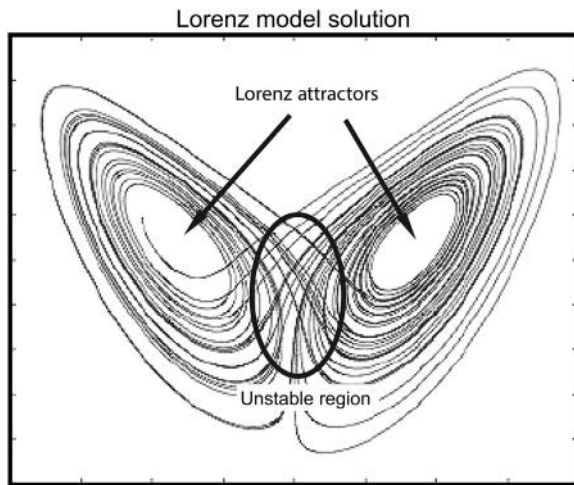


Figure 2 A solution to the Lorenz 63 model with the unstable region and attractors denoted.

into a forecast. These errors first grow from convective instability in regions of moist convection and saturate convective scales several hours into the forecast. Subsequent error growth occurs through geostrophic adjustment and/or cold pool dynamics, leading to mesoscale error saturation on a time scale of order $1/f$. The balanced component of the mesoscale error grows further in the presence of large-scale baroclinicity, which eventually limits large-scale predictability as well. This type of error growth, which can be found in a variety of midlatitude systems, in general explains why warm-season precipitation remains the least-accurate forecast element at all scales.

Despite the lack of predictability associated with warm-season convection, cool-season precipitation prediction fares much better. This is due to the fact that small-scale convective processes dominate warm-season precipitation, while synoptic-scale systems tend to force cool-season precipitation. This is in agreement with the finding that the rate of error growth is likely positively related to the magnitude and areal coverage of convective available potential energy (CAPE).

Predictability Concepts Applied to Tropical Cyclones

Before examining how general atmospheric predictability concepts apply to tropical cyclones, we take a brief detour to explain the general mechanism by which they intensify. Though there is no shortage of studies that have investigated factors influencing tropical cyclone intensification, any increase in tropical cyclone strength must occur through an inward radial advective flux of vorticity, which itself is driven by diabatic heating associated with moist convection. In terms of vertical vorticity tendency, stretching is the primary 'driver' of this process. Viewed through this lens, any process that enhances or disrupts latent heating and vortex stretching can potentially impact tropical intensity.

The sensitivity of hurricane intensity change to latent heating helps to clarify reasons for the disparity between hurricane track and intensity forecast improvement. Hurricane movement is generally controlled by the large-scale horizontal wind, of which forecasts have consistently improved during recent decades. However, tropical cyclone intensity is more strongly related to difficult-to-predict warm-season convection. The remainder of this section will examine how current ideas regarding predictability apply to tropical cyclones.

Intrinsic Predictability in Quiescent Environments

As a first step in examining the predictability of tropical cyclones, it is instructive to quantify intrinsic intensity predictability in environments with no mean flow or background asymmetries. A number of recent studies have used ensemble forecasts with idealized models to investigate tropical cyclone forecast differences that arise as a result of tiny initial condition perturbations. Though there are some differences in the methods and results, all of these studies have found for steady-state intense cyclones that the largest differences in maximum wind between any two ensemble members were typically $5\text{--}10\text{ m s}^{-1}$. In addition, ensemble standard deviation appears to be about $3\text{--}4\text{ m s}^{-1}$. Such differences are commensurate with discrepancies that arise from integrating exactly the same initial conditions on different supercomputer nodes.

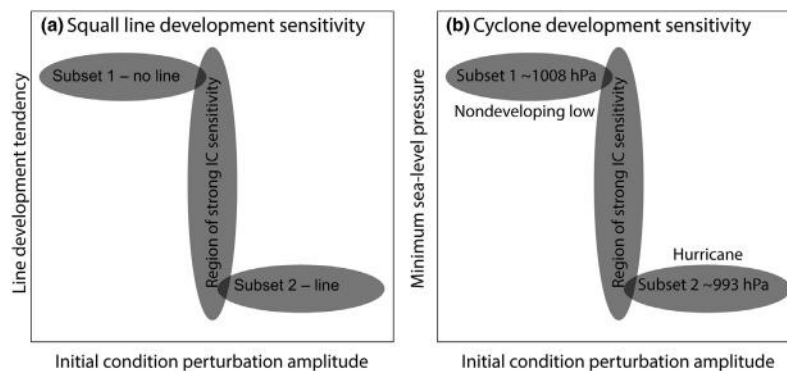


Figure 3 The sensitivity to small changes in initial conditions from recent studies simulating the development of (a) a squall line (source data is presented in Melhauser, C., Zhang, F., 2012. Practical and intrinsic predictability of severe and convective weather at the mesoscales. *Journal of Atmospheric Sciences* 69, 3350–3371) and (b) a tropical cyclone (source data is presented in Zhang and Sippel, 1999).

Rapid Intensification

RI, which is roughly defined as an increase in maximum surface wind speed of about 15 m s^{-1} over a 24-h period, is an important aspect of tropical cyclone intensification. About 20–30% of all tropical cyclones (the number varies by ocean basin) undergo RI at least once during their lifetimes. More importantly, RI occurs in a great majority $\sim 70\%$ of major tropical cyclones ($>51 \text{ m s}^{-1}$). Because of the frequent occurrence and potentially severe implications of a tropical cyclone undergoing RI, accurate RI forecasts could be very valuable.

Unfortunately, RI is the most problematic and error-prone facet of tropical cyclone intensity forecasting. It appears that environmental parameters of rapidly intensifying tropical cyclones are not much different than those of cyclones intensifying more slowly. This indicates that the rate of intensification is only weakly dependent on the environmental conditions, and in favorable intensification environments RI is more strongly controlled by internal dynamical processes. This is consistent with knowledge that local moist convection plays a critical role in RI, which should intrinsically limit its predictability. Thus, deterministic forecasts may never predict RI very well, suggesting the need for probabilistic forecasts.

The positive relationship between error growth and convective instability might also apply to RI. Recent work has shown that RI in at least a few cyclones commences when reservoirs of high convective instability are released, allowing for a significant expansion in central-core convection. For example, one recent study examined tropical storm Humberto (2001) immediately before it commenced RI to a hurricane. They found that high- θ_e air in the storm's inflow contributed to a buildup in CAPE, which was not immediately consumed due to a mesoscale inversion. The inversion, which was associated with a stratiform rain region, allowed CAPE to reach at least $2500 \text{ m}^2 \text{ s}^{-2}$ before convection erupted. The convection that ensued became the hurricane's eyewall, and RI immediately followed. Meanwhile, another study examined Hurricane Lili (2002) and found that RI commenced when a reservoir of high- θ_e air was released from the eye into the eyewall. They suggested that the reservoir was insufficiently large to sustain the entire RI cycle, but instead it acted to trigger the process. These findings are consistent with the general trend of rapidly intensifying storms to be in more unstable environments than steady-state storms (e.g., Figure 4).

Tropical Cyclone Structure and Size

Much of the uncertainty in tropical cyclone forecasts likely arises in the specification of the initial vortex. For example, the intensification of Hurricanes Danielle and Karl (2010) has been shown to be most sensitive to the initial circulation and thermodynamic fields associated with the pregenesis systems. The largest sensitivity was in the lower troposphere, close to the circulation maximum, and sensitivity to the surrounding environment was smaller.

Though no published work has directly examined the impact of a tropical cyclone's size or structure upon its intrinsic predictability, recent studies suggest that smaller tropical cyclones might be fundamentally less predictable

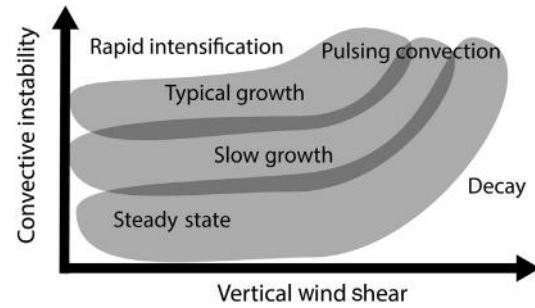


Figure 4 Hurricane intensity behavior as a function of environmental vertical wind shear and convective instability. Source data for this figure is presented in DeMaria, M., 2009. A simplified dynamical system for tropical cyclone intensity prediction. *Monthly Weather Review* 137, 68–82.

than larger ones. For example, new research has revealed that tropical cyclones with compact inner cores rapidly intensify more often than their larger counterparts. In addition, these compact cyclones are driven more strongly by inner-core convection and dynamics, whereas storms with larger inner cores are more strongly influenced by external forcing. Thus, the propensity for RI in small tropical cyclones appears to be a reflection of their sensitivity to unpredictable, small-scale convective processes. Nevertheless, much more work is needed to precisely assess the impact of tropical cyclone size upon predictability.

Cold Pool Dynamics and Cyclogenesis

As with midlatitude systems, convection and cold pool dynamics can rapidly amplify small analysis errors in short-term tropical cyclone forecasts. Similar to the midlatitude squall line case mentioned in the Section [Introduction](#), very small subsets of initial condition perturbations can make the difference between strongly developing and nondeveloping realizations of a tropical cyclone (Figure 3(b)). Also, small differences in moisture and instability can strongly impact the location, extent, and strength of cold pools, which affects the incipient low-level mesoscale vortex. If placed correctly, spreading cold pools can potentially destroy the initial low-level vortex, affecting both maximum intensity and size of the subsequent cyclone.

This demonstrates a particular difficulty in forecasting tropical cyclone formation. Significant low-level vortex amplification generally does not begin until sustained convection occurs very near or over the vortex center (i.e., vortex stretching), so the location of a cold pool can have a profound effect on an incipient tropical cyclone. If a cold pool develops but does not disrupt near-center convection, then genesis may proceed quickly, perhaps leading to RI. However, any disruption of convection requires time for the near-center planetary boundary layer to recover, which can take at least 12 h. If subsequent cold pools further inhibit convection over the center, then genesis could take even longer to commence. The lack of predictability associated with cold pools and mesoscale convection suggests the need for probabilistic forecasts of tropical cyclone formation.

Nonhomogenous Environments, Background Uncertainty, and Cyclogenesis

Nonhomogenous backgrounds and background forecast uncertainty can also strongly modulate the predictability of tropical cyclones. For example, a recent study used an ensemble forecast initialized with an ensemble Kalman filter (EnKF) to investigate reasons for unusually large forecast uncertainty associated with Hurricane Humberto (2007). Humberto rapidly developed off the upper Texas coast, and operational models completely failed to predict the storm's genesis, which resulted in extremely poor official forecasts. Though EnKF analysis uncertainty was fairly low, subsequent ensemble forecast spread grew rapidly. Intensification of the incipient cyclone was particularly affected by varying surface moisture and convective instability, which was related to the strength of a nearby front and proximity of the cyclone to the front. These discrepancies directly impacted the ability of convection to amplify the incipient vortex so that RI began immediately in some ensemble members, while others exhibited little intensity change.

Another recent study examined the effect of dry midlevel air on tropical cyclone formation by varying the distance between a relative humidity gradient and the center of an incipient vortex. They found that cyclone strength and structure in multiday forecasts was fairly sensitive to the initial location of the dry-air boundary when the boundary closed to within 150 km of the cyclone center. This is most relevant to cyclogenesis since the centers of developing systems can move discretely, jumping between convective complexes, and since weaker cyclones are more susceptible to infiltration of nearby dry air than are stronger systems.

Vertical Wind Shear and RI

Vertical wind shear can have profound impacts on the predictability of tropical cyclones. In idealized ensemble simulations with only very tiny initial condition perturbations, ensemble spread increases dramatically for environments of weak to moderate vertical wind shear. This is particularly true for the genesis and RI periods, where differences in maximum wind speed between different members can exceed 40 m s^{-1} . Ultimately, in the presence of vertical wind shear, random initial noise may change the onset and ending of RI by as much as 1–2 days.

The reason for these large differences is the randomness and chaotic nature of moist convection, which is similar to a number of aforementioned studies. Differences in moist convection first alter the tilt amplitude and angle of the vortex, which later significantly changes the precession and vortex alignment. Tropical cyclones rapidly intensify immediately after the tilt and local shear reach their minima, but the time at which this occurs is different in each member.

Some recent observational studies likewise suggest decreased predictability of tropical cyclones in the presence of strong vertical wind shear. A number of papers have investigated cases where tropical cyclone maximum intensity vacillated considerably over periods of less than a day. In the case of 6-h Hurricane Claudette (2003), vertical wind shear was directly ascribed to the quick weakening of the system, though there were insufficient observations to draw conclusions regarding the cause of the preceding intensification. Meanwhile, strong shear was also

responsible for both quick intensification and subsequent weakening of Tropical Storms Gabrielle (2004) and Edouard (2002) in the Gulf of Mexico.

The details of convection and cold pool evolution appear to be the mechanisms through which shear caused large intensity fluctuations in the observed cases. Vertical wind shear induces stronger, deeper inflow and much larger CAPE and helicity in the downshear region of a sheared storm, a combination that favors potentially explosive rotating convection. If that convection is sufficiently close to the cyclone core of high vorticity, then extreme RI can ensue (e.g., Tropical Storm Gabrielle intensified by 22 hPa in less than 3 h). Just as the downshear region is favorable for convection, the area upshear is quite unfavorable, largely because of shear-induced subsidence. This subsidence can dry a deep layer, which occasionally entrains into the downshear convection, resulting in cool downdrafts and cold pools. Such cold pools have been described as an 'antifuel' for tropical cyclones because of their ability to inhibit convection and quickly diminish the tropical cyclone vortex.

There are of course, several differences between the above observed and idealized ensemble cases. First, there has been no explicit examination of the predictability in the observed cases, though they were associated with a large degree of forecast error. In addition, such erratic intensity evolution and strong dependence upon convection is highly symptomatic of a lack of intrinsic predictability. Another difference is that all the idealized ensemble members eventually reached steady-state hurricane intensity, whereas the observed cases weakened soon after they intensified. This difference is probably due to the difference in vertical wind shear between the modeled and the observational studies. In the idealized ensemble, the strongest shear imposed was 5 m s^{-1} , whereas the observed storms all had greater than 10 m s^{-1} of vertical wind shear. The difference in development is consistent with what is shown in Figure 4.

Atmospheric Aerosols and Convection

New research also suggests that small variations in atmospheric aerosol concentrations can impact the predictability of tropical cyclones. Ensemble experiments with a model that accounts for microphysics–aerosol interactions have shown a nonlinear (albeit monotonically negative) relationship between the initial aerosol concentration and subsequent cyclone maximum intensity. When the initial number concentration increased from 100 to 1000 cm^{-3} , mean intensity decreased by roughly 15 m s^{-1} after 24 h. However, roughly half this intensity reduction was obtained by increasing the initial concentration to only 101 cm^{-3} . While more work is needed to fully assess the impact of aerosols on tropical cyclones, these results suggest that error growth associated with moist convection is amplified by uncertainty in background aerosol concentration.

Summary and Discussion

An underlying finding in most recent research is that error growth associated with moist convection limits the intrinsic predictability of tropical cyclones. As with midlatitude convection, practically immeasurable initial errors can grow upscale to introduce large forecast errors, particularly during RI.

Aside from tiny initial perturbations, typical error sources include uncertainty in specifying the initial cyclone vortex, uncertainty in the background state, particularly near gradients, and uncertainty in aerosol concentration.

Though forecast error can explosively grow in some tropical cyclone forecasts, other cases are relatively predictable implying that ensemble spread is case dependent. Thus, some tropical cyclone forecasts are particularly sensitive to initial condition errors, while others are not. In terms of intensity forecasts, the least predictable cases are likely to be those with high error growth potential of near-center convection, possibly associated with the consumption of CAPE. This most obviously includes environments that are favorable for intensification, since RI is usually a possibility in favorable environments. In addition, otherwise favorable environments with moderate vertical wind shear can also lead to very large error growth due to the buildup of CAPE and the influence of cold pools in sheared tropical cyclones. Particularly low predictability also appears to accompany storms that are in locally favorable environments but with less favorable conditions nearby.

The fact that some tropical cyclones have higher predictability suggests that small-scale convective processes do not always control cyclone strength. In these cases, there could be strong modulation of convection by the more predictable large-scale environment, or perhaps forcing associated with the mean secondary circulation of the parent vortex becomes more important than smaller scale processes. This is supported by the fact that global models demonstrate skill at forecasting tropical cyclone genesis even though they cannot adequately resolve mesoscale processes. For developed tropical cyclones, enhanced predictability is likely to be found in systems with large inner cores, which are known to have slower intensity changes than those with more compact cores. The midlatitude corollary to cyclones with higher predictability could be that strongly forced cold season precipitation is more predictable than warm-season convection.

The change in predictability from one system to the next illustrates the utility of advanced ensemble prediction systems that can provide event-dependent probabilistic forecasts. Despite the demonstrated benefits of using ensembles in hurricane forecasting, the uncertainty involved with today's operational hurricane forecasts is still mostly based on averaged climatological errors and thus is not case dependent. The body of work discussed herein demonstrates that some events can be associated with very large uncertainty, and operational use of this information would more accurately convey the risks associated with individual systems.

The results of recent experiments that have examined tropical cyclone intrinsic predictability (see Section [Intrinsic Predictability in Quiescent Environments](#)) give some insight into recent intensity error trends. A lower bound of intrinsic predictability, which can be estimated by the ensemble spread in recent intrinsic predictability investigations, is about $3\text{--}4\text{ m s}^{-1}$. This is only slightly higher than current 24-h intensity error, which might mean that average short-term intensity forecast error is already near the limit of intrinsic predictability. If this is true, then even the 36-h forecast error may not improve much more, though there is still room for improvement in longer range intensity forecasts. This hypothesis is supported by the results of a recent multiyear study that examined intensity error in forecasts initialized with EnKF analyses of high-resolution hurricane

data. For all forecast times at and beyond 24 h, bias-corrected intensity error was $4\text{--}5\text{ m s}^{-1}$ on average.

Though assimilation of high-resolution tropical cyclone observations can significantly improve track and intensity forecasts, many aspects of the intrinsic limit of hurricane predictability remain unclear. Even if we are approaching the limit of predictability in short-term maximum intensity forecasts, the scientific community has only recently begun to assess forecasts of hurricane size. This forecast metric is particularly important since it is strongly related to storm surge potential, and storm surge can be the most catastrophic and deadly aspect of a hurricane. Meanwhile, practical predictability associated with realistic initial condition and model errors may be increased through improving our understanding of tropical cyclones, developing better numerical models, and improving data coverage and assimilation techniques. However, as discussed above, there always will be forecast errors due to the inherent limit of predictability arising from initial errors with amplitudes far smaller than any observation or analysis system.

See also: Data Assimilation and Predictability: Ensemble-Based Data Assimilation; Predictability and Chaos. **Tropical Cyclones and Hurricanes:** Hurricane Dynamics; Hurricanes: Observation; Tropical Cyclogenesis; Tropical Cyclones: Secondary Eyewall Formation.

Further Reading

- Aksoy, A., 2013. Assimilation of high-resolution tropical cyclone observations with an ensemble Kalman filter using NOAA/AOML/HRD's HEDAS: evaluation of the 2008–2011 vortex-scale analyses. *Monthly Weather Review* 141, 1842–1865.
- DeMaria, M., 2009. A simplified dynamical system for tropical cyclone intensity prediction. *Monthly Weather Review* 137, 68–82.
- Hendricks, E.A., Peng, M.S., Fu, B., Li, T., 2010. Quantifying environmental control on tropical cyclone intensity change. *Monthly Weather Review* 138, 3243–3271.
- Krishnamurti, T.N., et al., 1999. Improved weather and seasonal climate forecasts from multimodel superensemble. *Science* 285, 1548–1550.
- Lorenz, E.N., 1969. The predictability of a flow which possesses many scales of motion. *Tellus* 21, 289–307.
- Melhauser, C., Zhang, F., 2012. Practical and intrinsic predictability of severe and convective weather at the mesoscales. *Journal of Atmospheric Sciences* 69, 3350–3371.
- Munsell, E.B., Zhang, F., Stern, D.P., 2013. Predictability and dynamics of a nonintensifying tropical storm: Erika (2009). *Journal of Atmospheric Sciences* 70, 2505–2524.
- Olson, D.A., Junker, N.W., Korty, B., 1995. Evaluation of 33 years of quantitative precipitation forecasting at the NMC. *Weather Forecasting* 10, 498–511.
- Palmer, T.N., 1993. Extended-range atmospheric prediction and the Lorenz model. *Bulletin of the American Meteorological Society* 74, 49–65.
- Sippel, J.A., Zhang, F., 2010. Factors affecting the predictability of Hurricane Humberto (2007). *Journal of Atmospheric Sciences* 67, 1759–1778.
- Weng, Y., Zhang, F., 2012. Assimilating airborne Doppler radar observations with an ensemble Kalman filter for convection-permitting hurricane initialization and prediction: Katrina (2005). *Monthly Weather Review* 140, 841–859.
- Zhang, F., Sippel, J.A., 2009. Effects of moist convection on hurricane predictability. *Journal of Atmospheric Sciences* 66, 1944–1961.
- Zhang, F., Tao, D., 2009. Effects of vertical wind shear on the predictability of tropical cyclones. *Journal of Atmospheric Sciences* 70, 975–983.
- Zhang, F., Bei, N., Rotunno, R., Snyder, C., Epifanio, C.C., 2007. Mesoscale predictability of moist baroclinic waves: convection-permitting experiments and multistage error growth dynamics. *Journal of Atmospheric Sciences* 64, 3579–3594.
- Zhang, F., Weng, Y., Sippel, J.A., Meng, Z., Bishop, C.H., 2009. Cloud-resolving hurricane initialization and prediction through assimilation of Doppler radar observations with an ensemble Kalman filter: Humberto (2007). *Monthly Weather Review* 137, 2105–2125.
- Zhang, F., Weng, Y., Gamache, J.F., Marks, F.D., 2011. Performance of convection-permitting hurricane initialization and prediction during 2008–2010 with ensemble data assimilation of inner-core airborne Doppler radar observations. *Geophysical Research Letters* 38, L15810.

Hurricanes: Observation

FD Marks, Hurricane Research Division, Miami, FL, USA

© 2003 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 3, pp 942–966, © 2003, Elsevier Ltd.

Introduction

‘Hurricane’ is the term used in the Western Hemisphere for one of the general class of strong tropical cyclones, including western Pacific typhoons and similar systems, that are known simply as cyclones in the Indian and southern Pacific Oceans. A tropical cyclone is a low-pressure system which derives its energy primarily from evaporation from the sea in the presence of 1-minute sustained surface wind speeds $>17 \text{ m s}^{-1}$ and the associated condensation in convective clouds concentrated near its center. In contrast, midlatitude storms (low-pressure systems with associated fronts) get their energy primarily from the horizontal temperature gradients that exist in the atmosphere. Structurally, the strongest winds in tropical cyclones are near the Earth’s surface (a consequence of being ‘warm-core’ in the troposphere), while the strongest winds in midlatitude storms are near the tropopause (a consequence of being ‘warm-core’ in the stratosphere and ‘cold-core’ in the troposphere). ‘Warm-core’ refers to being warmer than the environment at the same pressure surface.

A tropical cyclone with the highest sustained wind speeds between 17 and 32 m s^{-1} is referred to as a tropical storm, whereas a tropical cyclone with sustained wind speeds $\geq 33 \text{ m s}^{-1}$ is referred to as a hurricane or typhoon. Once a tropical cyclone has sustained winds $\geq 50 \text{ m s}^{-1}$ it is referred to as a major hurricane or super typhoon. In the Atlantic and eastern Pacific Oceans hurricanes are also classified by the damage they can cause using the Saffir–Simpson scale (Table 1).

The Saffir–Simpson scale categorizes hurricanes on a scale from 1 to 5, with 1 the weakest and 5 the most intense. Major hurricanes correspond to categories 3 and higher. The reasons that some disturbances intensify to a hurricane, while others do not, are not well understood. Neither is it clear why some tropical cyclones become major hurricanes, while others do not. Major hurricanes produce 80–90% of the United States hurricane-caused damage despite accounting for only one-fifth of all landfalling tropical cyclones. Only two category 5 hurricanes made landfall on the mainland United States (Florida Keys 1935 and Camille 1969). Recent major hurricanes to make landfall on the United States were Hurricanes Bonnie and Georges in 1998, and Bret and Floyd in 1999.

As with large-scale extratropical weather systems, the structure and evolution of a tropical cyclone is dominated by the fundamental contradiction that while the airflow within a tropical cyclone represents an approximate balance among forces affecting each air parcel, slight departures from balance are essential for vertical motions and resulting clouds and precipitation, as well as changes in tropical cyclone intensity. As in extratropical weather systems, the basic vertical balance of forces in a tropical cyclone is hydrostatic except in the eyewall, where convection is superimposed on the hydrostatic motions. However, unlike in extratropical weather systems, the basic horizontal balance in a tropical cyclone above the boundary layer is between the sum of the Coriolis ‘acceleration’ and the centripetal ‘acceleration’, balanced by the horizontal pressure gradient force. This balance is referred to as gradient balance, where the Coriolis ‘acceleration’ is defined as the horizontal velocity of an air parcel, v , times the Coriolis parameter, f . (f is the Coriolis parameter ($f = 2\Omega \sin \phi$), where Ω is the angular velocity of the Earth ($7.292 \times 10^{-5} \text{ s}^{-1}$) and ϕ is latitude. The Coriolis parameter is zero at the equator and 2Ω at the pole.) Centripetal ‘force’ is defined as the acceleration on a parcel of air moving in a curved path, directed toward the center of curvature of the path, with magnitude v^2/r , where v is the horizontal velocity of the parcel and r the radius of curvature of the path. The centripetal force alters the original two-force geostrophic balance and creates a nongeostrophic gradient wind.

The inner region of the tropical storm, termed the cyclone ‘core’, contains the spiral bands of precipitation, the eyewall, and the eye that characterize tropical cyclones in radar and satellite imagery (Figure 1). The primary circulation – the tangential or swirling wind – in the core becomes strongly axisymmetric as the cyclone matures. The strong winds in the core, which occupies only 1–4% of the cyclone’s area, threaten human activities and make the cyclone’s dynamics unique. In the core, the local Rossby number is always >1 and may be as high as 100. The Rossby number indicates the relative magnitude of centrifugal (v/r) and Coriolis (f) accelerations, $Ro = V/fr$, where V is the axial wind velocity, f the Coriolis parameter, and r the radius from the storm center. An approximate breakdown of regimes is: $Ro < 1$, geostrophic

Table 1 Saffir–Simpson scale of hurricane intensity

Category	Pressure (hPa)	Wind (ms^{-1})	Storm surge (m)	Damage
1	>980	33–42	1.0–1.7	Minimal
2	979–965	43–49	1.8–2.6	Moderate
3	964–945	50–58	2.7–3.8	Extensive
4	944–920	59–69	3.9–5.6	Extreme
5	<920	≥ 70	≥ 5.7	Catastrophic

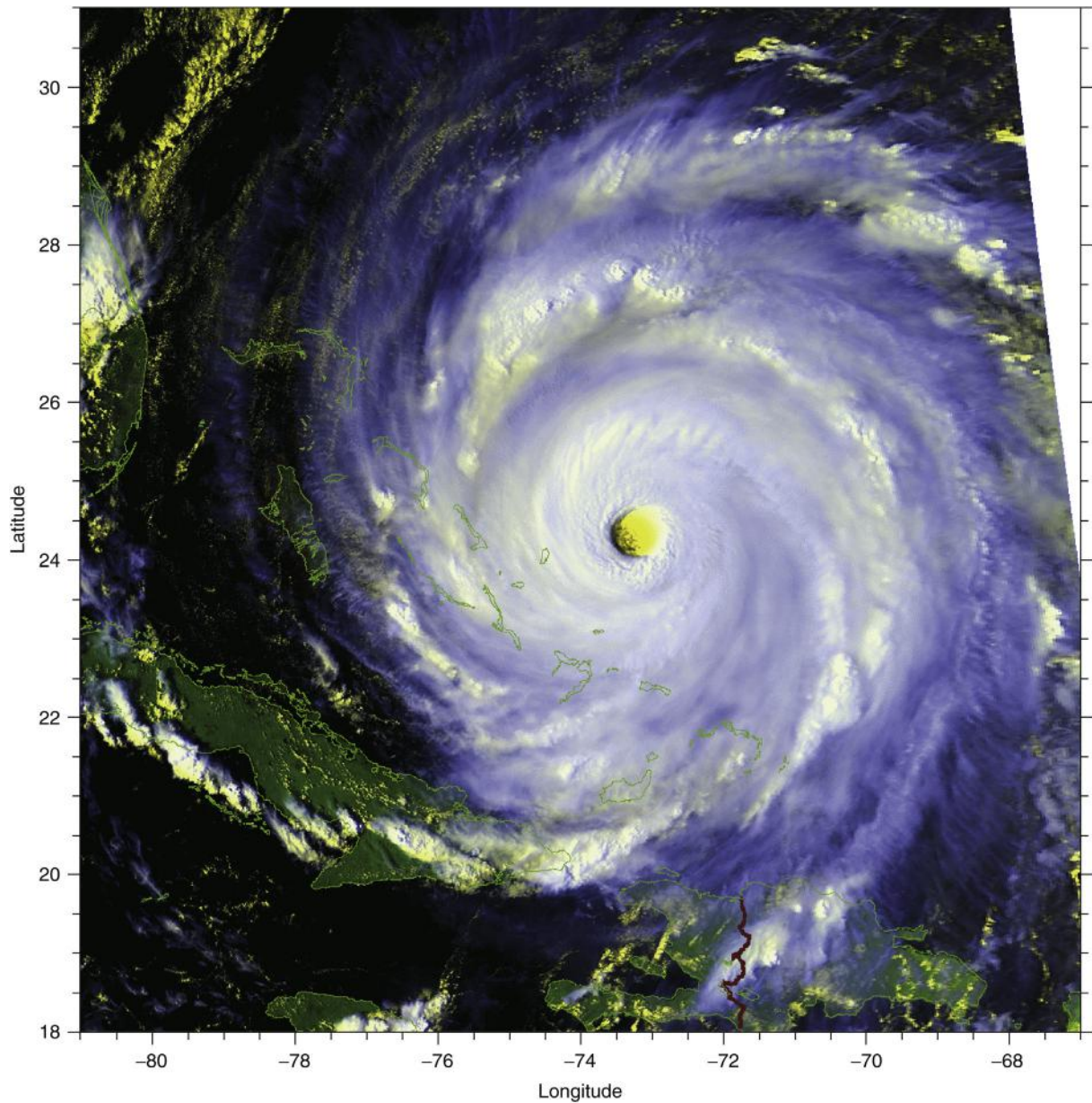


Figure 1 NOAA-14 AVHRR multispectral false color image of Hurricane Floyd at 2041 UTC, 13 September 1999 about 800 km east of southern Florida. Photo courtesy of NOAA Operationally Significant Event Imagery website: <http://www.osei.noaa.gov/>.

flow; $Ro > 1$, gradient flow; and $Ro > 50$, cyclostrophic flow. When the Rossby number significantly exceeds unity, the balance in the core becomes more cyclostrophic, where the pressure gradient force is almost completely balanced by the centrifugal ‘force’. The time scales are such that air swirling around the center completes an orbit in much less than a pendulum day (defined as $1/f$).

When the atmosphere is in approximate horizontal and vertical balance, the wind and mass fields are tightly interconnected. The distribution of a single mass or momentum variable may be used as a starting point to infer the distribution of all other such variables. One such variable is potential vorticity (PV), approximately equal to the vorticity times the

thermal stratification, which is related to the three-dimensional mass and momentum fields through an inverse second-order Laplacian-like operator. The benefit of such a relationship is that PV variations in a single location are diagnostically related to variations in mass and wind fields at a distance. Areas of high PV correspond locally to low mass, or cyclones, while areas of low PV correspond to anticyclones.

Typical extratropical weather systems contain high PV values around $0.5 \times 10^{-6} \text{ m}^2 \text{ s}^{-1} \text{ K kg}^{-1}$ (0.5 PVU) to 5 PVU, whereas typical values in the tropical cyclone core are ≥ 10 PVU. **Figure 2** shows the wind and mass fields associated with an idealized axially symmetric tropical cyclone PV anomaly with the PV concentrated near the surface rather than in

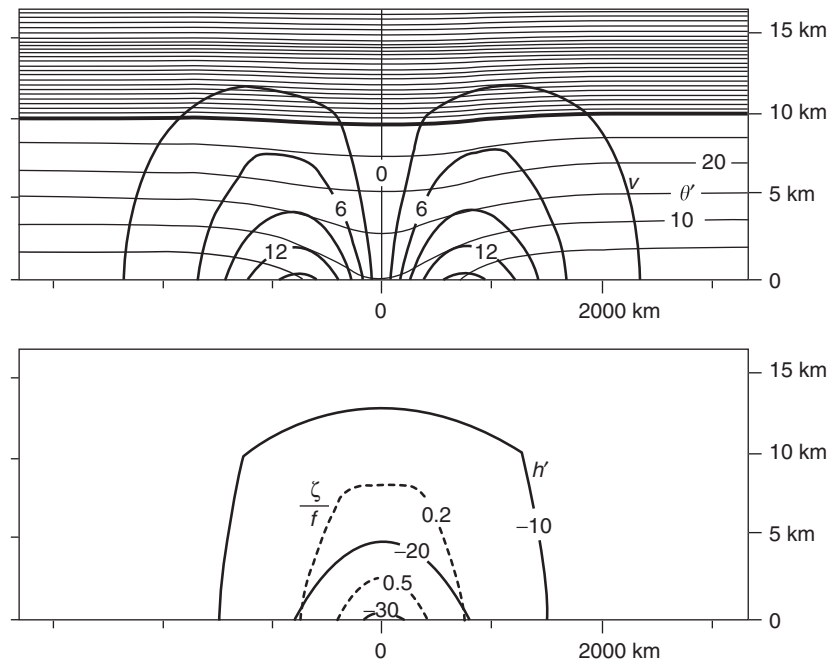


Figure 2 v (gradient wind) (ms^{-1}) and θ' (perturbation potential temperature) (K, top panel); and h' (geopotential height perturbation) (dm) and ζ/f (bottom panel) for a warm core, lower cyclone. The tropopause location is denoted by the bold solid line, and the label 0 on the horizontal axis indicates the core (and axis of symmetry) of the disturbance. The equivalent pressure deviation at the surface in the center of the vortex is -31 hPa. Reproduced with permission from Thorpe AJ (1986) Synoptic scale disturbances with circular symmetry. *Monthly Weather Review* 114: 1384–1389; © American Meteorological Society.

a vertical column. The cyclonic anomaly (positive in the Northern Hemisphere) is associated with a cyclonic circulation that is strongest at the level of the PV anomaly near the surface, and decreases upward. Temperatures are anomalously warm above the PV anomaly (isentropic surfaces are deflected downward). While consistent with the simple PV distribution, the wind and mass fields are also in horizontal and vertical balance. The tropical cyclone being a warm-core vortex, the PV inversion dictates that the winds that swirl about the center decrease with increasing height, but they typically fill the depth of the troposphere. If the PV reaches values ≥ 10 PVU, the inner region winds can become intense, as in Hurricane Gloria (Figure 3). Gloria had PV values exceeding 500 PVU just inside the radius of maximum winds of 15 km where the axisymmetric mean tangential winds exceeded 65 m s^{-1} .

Many features in the core, however, persist with little change for (pendulum) days (mean life span of a tropical cyclone is about 5–10 days). Because these long lifetimes represent tens or hundreds of orbital periods (~ 1 h), the flow is nearly balanced. Moreover, at winds $> 35 \text{ m s}^{-1}$, the local Rossby radius of deformation is reduced from its normal $\sim 10^3$ km to a value comparable with the eye radius. The Rossby radius of deformation is the ratio of the speed of the relevant gravity wave mode and the local vorticity, or, equivalently, the ratios of the Brunt–Väisälä and inertial frequencies. This scale indicates the amount of energy that goes into gravity waves compared with inertial acceleration of the wind. In very intense tropical cyclones, the eye radius may approach the depth of the troposphere (15 km), making the aspect ratio unity. Thus, the dynamics near the center of a tropical cyclone are so exotic

that conditions in the core differ from the Earth's day-to-day weather as much as the atmosphere of another planet does.

Climatology

There are 80–90 tropical cyclones worldwide per year, with the Northern Hemisphere having more tropical cyclones than the Southern Hemisphere. Table 2 shows that of the 80–90 tropical cyclones, 45–50 reach hurricane or typhoon strength and 20 reach major hurricane or super typhoon strength. The western North Pacific (27 tropical cyclones), eastern North Pacific (17 tropical cyclones), south-west Indian Ocean (10 tropical cyclones), Australia/south-west Pacific (10 tropical cyclones), and North Atlantic (10 tropical cyclones) are the major tropical cyclone regions. There are also regional differences in the tropical cyclone activity by month with the majority of the activity in the summer season for each basin. Hence, in the Pacific, Atlantic, and North Indian Ocean the maximum numbers of tropical cyclones occur in August through October, while in the South Pacific and Australia regions the maxima are in February and March. In the South Indian Ocean, the peak activity occurs in June. In the western North Pacific, Bay of Bengal, and South Indian Ocean regions tropical cyclones may occur in any month, while the other regions at least one tropical cyclone-free month occurs per year. For example, in the North Atlantic, there has never been tropical cyclone activity in January.

Some general conclusions can be drawn from the global distribution of tropical cyclone locations (Figure 4(a)). Tropical

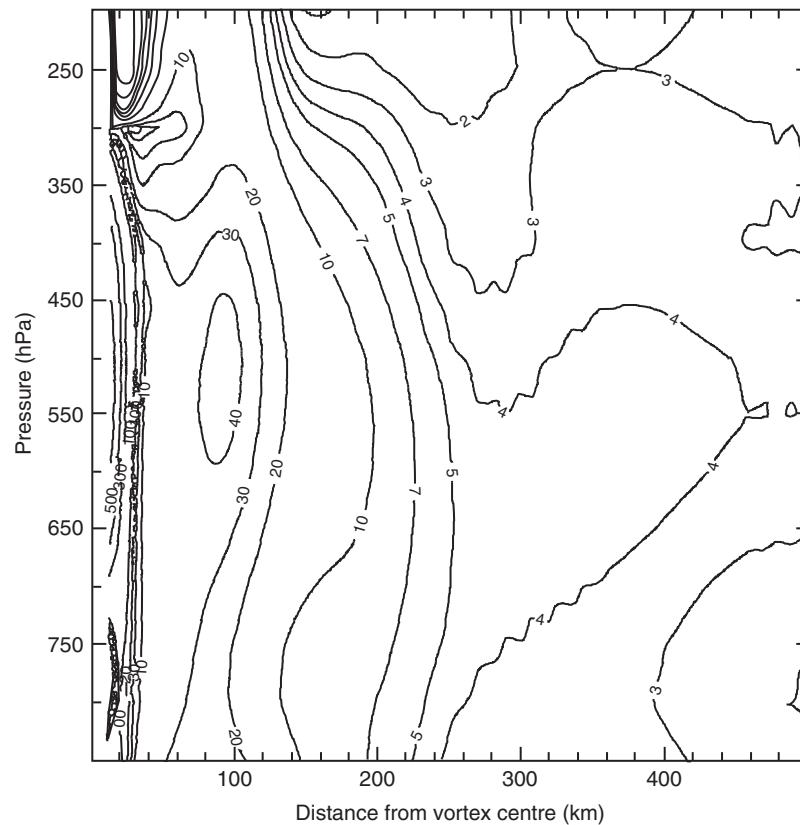


Figure 3 Radial-height cross-section of symmetric potential vorticity for Hurricane Gloria, 24 September 1985. Contours are 0.1 PVU. Values in the data-sparse region, within 13 km of vortex center, are not displayed. Reproduced with permission from Shapiro LJ and Franklin JL (1995) Potential vorticity in Hurricane Gloria. Monthly Weather Review 123: 1465–1475; © American Meteorological Society.

Table 2 Mean annual frequency, standard deviation (σ) and percentage of global total of the number of tropical storms (winds $\geq 17 \text{ m s}^{-1}$), hurricane-force tropical cyclone (winds $\geq 33 \text{ m s}^{-1}$), and major hurricane-force tropical cyclone (winds $\geq 50 \text{ m s}^{-1}$). Dates in parentheses provide the nominal years for which accurate records are currently available

Tropical cyclone basin	Tropical storm annual frequency (σ)	% of total	Hurricane annual frequency (σ)	% of total	Major hurricane annual frequency (σ)	% of total
Atlantic (1944–00)	9.8 (3.0)	11.4	5.7 (2.2)	12.1	2.2 (1.5)	10.9
NE Pacific (1970–00)	17.0 (4.4)	19.7	9.8 (3.1)	20.7	4.6 (2.5)	22.9
NW Pacific (1970–00)	26.9 (4.1)	32.1	16.8 (3.6)	35.5	8.3 (3.2)	41.3
N Indian (1970–00)	5.4 (2.2)	6.3	2.2 (1.8)	4.6	0.3 (0.5)	1.5
SW Indian (30–100° E) (1969–00)	10.3 (2.9)	12.0	4.9 (2.4)	10.4	1.8 (1.9)	9.0
Australian/SE Indian (100–142° E) (1969–00)	6.5 (2.6)	7.5	3.3 (1.9)	7.0	1.2 (1.4)	6.0
Australian/SW Pacific (142° E) (1969–00)	10.2 (3.1)	11.8	4.6 (2.4)	9.7	1.7 (1.9)	8.5
Global (1970–00)	86.1 (8.0)		47.3 (6.5)		20.1 (5.7)	

cyclone formation is confined to a region approximately 30° N and 30° S, with 87% of them located within 20° of the Equator. There is a lack of tropical cyclones near the Equator, as well as in the eastern South Pacific and South Atlantic basins. From these observations there appear to be at least five necessary conditions for tropical cyclone development.

- Warm sea surface temperature (SST) and large mixed-layer depth (i.e., the thickness of the mixed layer, defined as the depth of

the sharp temperature inversion (also referred to as the thermocline) between the cooler bottom water and the warmer near surface water). Numerous studies suggest a minimum SST criterion of 26 °C for development. The warm water must also have sufficient depth (i.e., 50 m). Comparison of Figures 4(a) and 4(b), the annual mean global SST, shows the strong correlation between regions where the SST is $> 26^\circ \text{C}$ and annual tropical cyclone activity. An SST of $> 26^\circ \text{C}$ is sufficient but not necessary for tropical

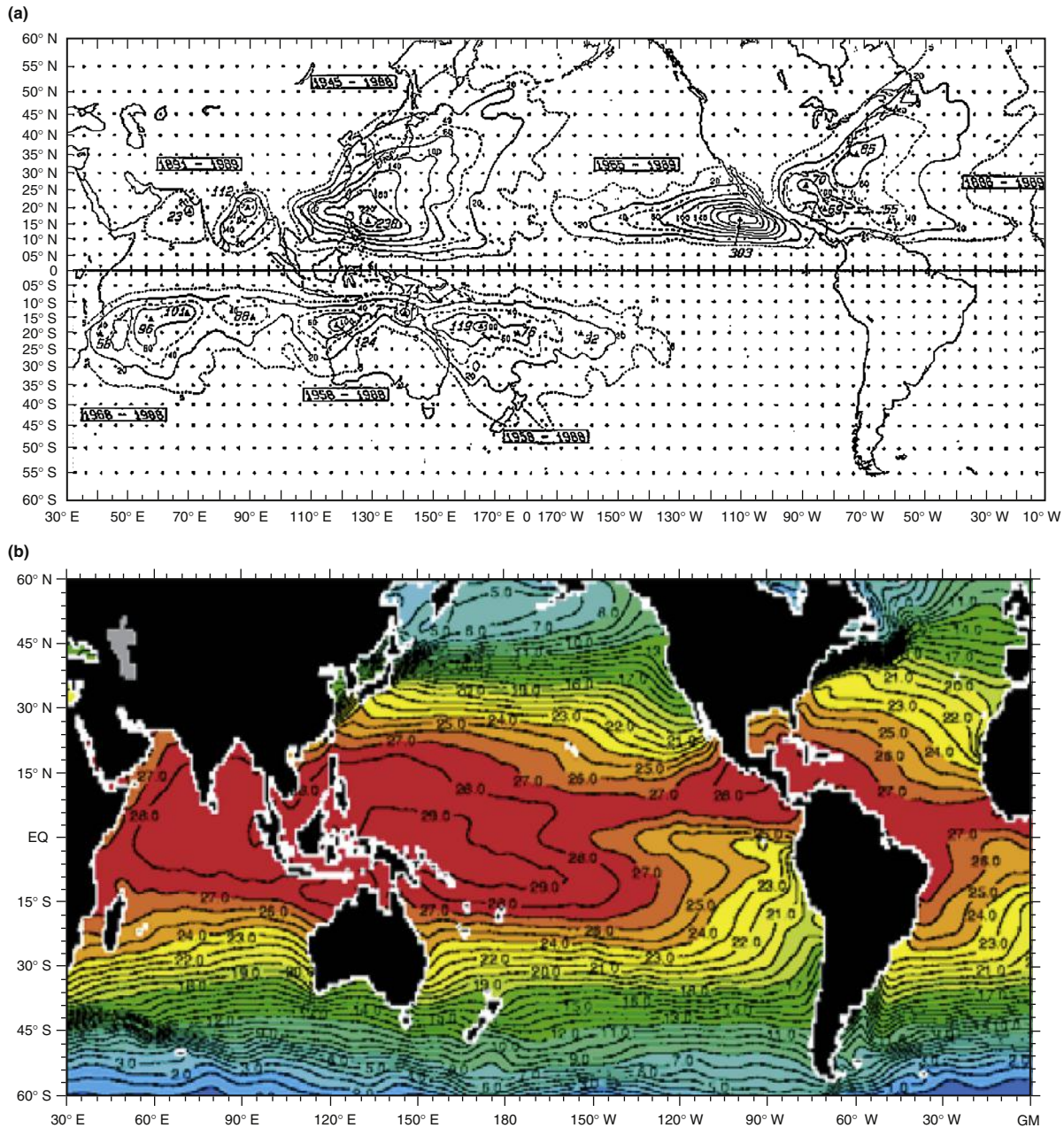


Figure 4 (a) Frequency of tropical cyclones per 100 years within 140 km of any point. Solid triangles indicate maxima, with values shown. Period of record is shown in boxes for each basin. (b) Annual sea surface temperature distribution ($^{\circ}\text{C}$).

cyclone activity, as is evidenced by the regions with tropical cyclone activity when the SST $< 26^{\circ}\text{C}$. Some of the discrepancy exists because storms that form over regions where the SST is $> 26^{\circ}\text{C}$ are advected poleward during their life cycle. However, tropical cyclones are observed to originate over regions where the SST $< 26^{\circ}\text{C}$. These occurrences are not many, but the fact that they exist suggests that other factors are important.

- **Background Earth vorticity.** Tropical cyclones do not form within 3° of the Equator. The Coriolis parameter vanishes at the Equator and increases to extremes at the poles. Hence, a threshold value of Earth vorticity (f) must exist for

a tropical cyclone to form. However, the likelihood of formation does not increase with increasing f . Thus, nonzero Earth vorticity is necessary, but not sufficient to produce tropical cyclones.

- **Low vertical shear of the horizontal wind.** In order for tropical cyclones to develop, the latent heat generated by the convection must be kept near the center of the storm. Historically, shear was thought to ‘ventilate’ the core of the cyclone by advecting the warm anomaly away. The ventilation argument suggests that if the storm travels at nearly the same speed as the environmental flow in which it is embedded then its heating remains over the disturbance

center. However, if it is moving slower than the mean wind at upper levels then the heating in the upper troposphere is carried away by the mean flow. Recent analysis suggests that the effect of shear is to force the convection into an asymmetric pattern such that the convective latent heat release forces flow asymmetry and irregular motion rather than intensification of the symmetric vortex. Thus, if the vertical shear is too strong ($>16 \text{ m s}^{-1}$) then existing tropical cyclones are ripped apart and new ones cannot form.

- **Low atmospheric static stability.** Static stability is the ability of a fluid to become turbulent (unstable) or laminar (stable) due to the effects of buoyancy neglecting all other inertial effects of motion. The troposphere must be potentially unstable to sustain convection for an extended period. Typically measured as the difference between the equivalent potential temperature, θ_e , at the surface and 500 hPa, instability must typically be $>10 \text{ K}$ for convection to occur. This value is usually satisfied over tropical oceans.
- **Tropospheric humidity.** The higher the midlevel humidity, the longer a parcel of air can remain saturated as it entrains the surrounding air during its ascent. Vigorous convection occurs if the parcel remains saturated throughout its ascent. A relative humidity of 50–60% at lower to midlevels (700–500 hPa) is often sufficient to keep a parcel saturated during ascent. This condition is regularly evident over tropical oceans.

These conditions are usually satisfied in the summer and fall seasons for each tropical cyclone basin. However, even when all of the above conditions are favorable, tropical cyclones don't necessarily form. In fact, there is growing evidence for significant interannual variability in tropical cyclone activity, where numerous tropical cyclones form in a given basin over a week to 10 days, followed by 2–3 weeks with little or no tropical cyclone activity. **Figure 5** shows just such an active period in the Atlantic basin in mid-September 1999, where two hurricanes (Floyd and Gert), both major, and an unnamed tropical depression formed within a few days of each other. During these active phases almost every disturbance makes at least tropical storm strength, whereas in the inactive phase practically none of the disturbances intensify. The two hurricanes and unnamed depression in **Figure 5** represented the second 10-day active period during the summer of 1999. An earlier period in mid-August also resulted in the development of three hurricanes (Brett, Cindy, and Dennis), two of which were major, as well as a tropical storm (Emily). There is speculation that the variability is related to the propagation of a global wave. Because the SST, static stability, and Earth vorticity don't vary that much during the season, the interannual variability is most likely related to variations in tropospheric relative humidity and vertical wind shear.

It has long been recognized that the number of tropical cyclones in a given region varies from year to year. The exact causes of this remain largely speculative. The large-scale global

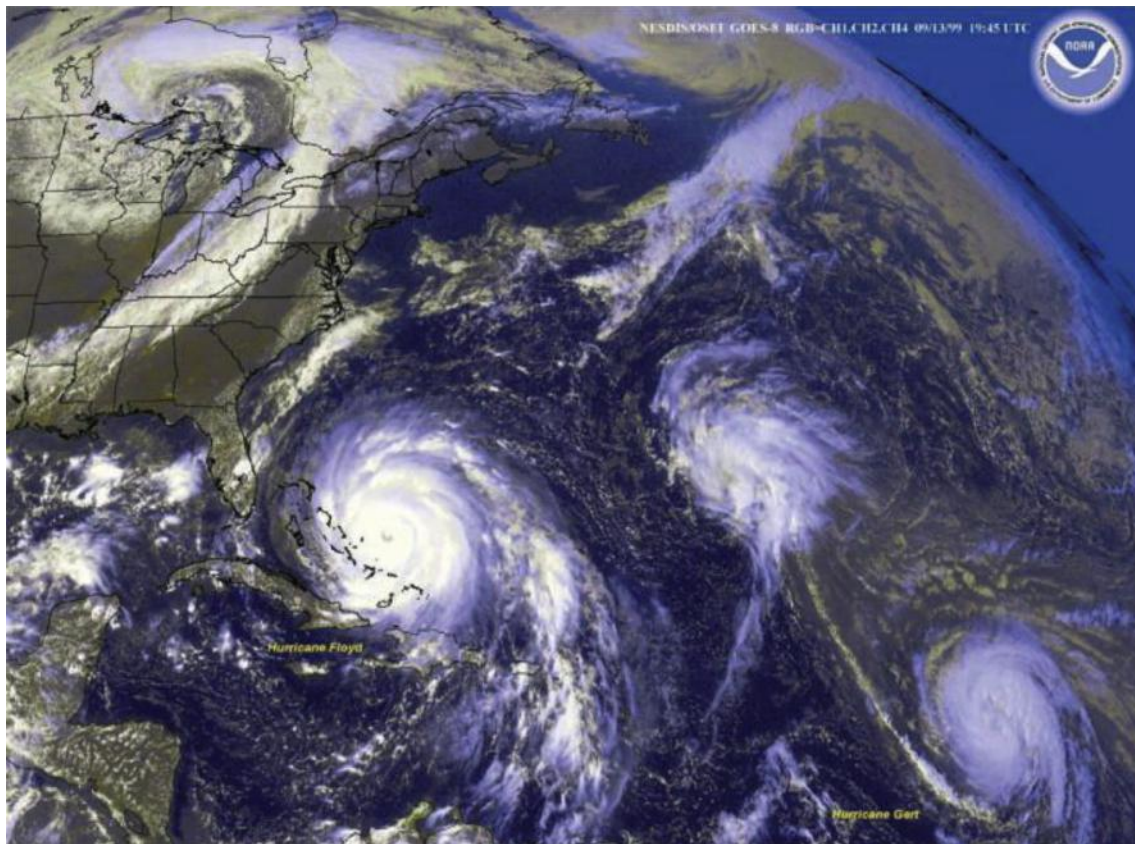


Figure 5 GOES multispectral false color image of Hurricanes Floyd and Gert and an unnamed tropical depression at 1935 UTC, 13 September 1999. Photo courtesy of NOAA Operationally Significant Event Imagery website: <http://www.osei.noaa.gov/>.

variations in atmospheric phenomena such as the El Niño Southern Oscillation (ENSO) and the Quasi-Biennial Oscillation (QBO) appear to be related to annual changes in the frequency of tropical cyclone formation, particularly in the Atlantic Ocean. The ENSO phenomenon is characterized by warmer SSTs in the eastern South Pacific and anomalous winds over much of the equatorial Pacific. It influences tropical cyclone formation in the western North Pacific, South Pacific, and even the North Atlantic.

During the peak phase of the ENSO, often referred to as El Niño (which usually occurs during the months July–October), anomalous westerly winds near the Equator extend to the dateline in the western North Pacific acting to enhance the intertropical convergence zone (ITCZ) in this area, making it more favorable for formation of tropical cyclones. Another effect of the El Niño circulation is warmer SST in the eastern South Pacific. During such years, tropical cyclones form closer to the Equator and farther east. Regions such as French Polynesia, which are typically unfavorable for tropical cyclones owing to a strong upper-level trough, experience numerous tropical cyclones. The eastern North Pacific is also affected by the El Niño through a displacement of the ITCZ south to near 5° N. Additionally, the warm ocean anomaly of El Niño extends to near 20° N, which enhances the possibility of tropical cyclone formation. The result is an average increase of two tropical cyclones during El Niño years. Cyclones also develop closer to the Equator and farther west than during a normal year.

The QBO is a roughly 2-year oscillation of the equatorial stratosphere (30–50 hPa) winds from easterly to westerly and back. The phase and magnitude of QBO are associated with the frequency of tropical cyclones in the Atlantic. Hurricane activity is more frequent when the 30-hPa stratospheric winds are westerly. The exact mechanism by which the QBO affects tropical cyclones in the troposphere is not clear; however, there are more North Atlantic tropical cyclones when the QBO is in the westerly phase than when it is in the easterly.

Tropical Cyclogenesis

Enthalpy is a thermodynamic state function defined for an ideal gas as the temperature times the specific heat at constant pressure plus a constant. For a system like the atmosphere which consists of a mixture of components the total enthalpy is the mass-weighted sum of the enthalpies of each component. Thus, the total enthalpy for a system consisting of a mixture of dry air, water vapor, and liquid water is defined as a constant plus the temperature times the sum of the specific heats at constant pressure for each component times the masses of each component, respectively. In an adiabatic, reversible process, the total enthalpy is conserved, although the component enthalpies may not be due to the exchange of enthalpy between components in phase changes. Most of the energy needed for tropical cyclones to form and maintain themselves is realized through the difference in enthalpy between the warm near-surface waters of the tropical ocean and the tropospheric column. The process of bringing the late-summer tropical troposphere into thermodynamic equilibrium with the sea surface at 28–30 °C, mainly through the irreversible energy

transfer from the ocean to the air by evaporation, can produce hydrostatic pressures as low as the minimum sea-level pressures of the most intense tropical cyclones. Thus, much of the tropical oceans contain enough moist enthalpy to support a major hurricane.

Throughout most of the Trade Wind regions, gradual subsidence causes an inversion that traps water vapor in the lowest kilometer. Sporadic convection (often in squall lines) that breaks through the inversion exhausts the moist enthalpy stored in the near-surface boundary layer quickly, leaving a wake of cool, relatively dry air. This air comes from just above the inversion and is brought to the surface by downdrafts driven by the weight of hydrometeors and cooling due to their evaporation. If the squall line does not keep moving it quickly runs out of energy. A day, or even several days, may pass before normal fair-weather evaporation can restore the preexistent moist enthalpy behind the squall. The reasons why squall line convection generally fails to produce hurricanes lie in the limited amount of enthalpy that can be stored in the sub-inversion layer and the slow rate of evaporation under normal wind speeds in the trades.

For a tropical cyclone to occur, evaporation must speed up and the equilibrium enthalpy at the sea surface temperature must rise through a lowering of the surface pressure. Tropical cyclones are thus finite-amplitude phenomena. They do not grow by some linear process from infinitesimal amplitude. The normal paradigm of searching for the most rapidly growing unstable linear mode used to study midlatitude cyclogenesis through baroclinic instability fails here. The surface wind has to exceed roughly 20 m s⁻¹ before evaporation can prevail against downdraft cooling.

How then do tropical cyclones reach the required finite amplitude? The answer seems to lie in the structure of tropical convection. As explained previously, behind a squall line the lower troposphere (below the 0 °C isotherm at ~5 km) is dominated by precipitation-driven downdrafts which lie under the ‘anvil’ of nimbostratus and cirrostratus that spreads behind the active convection. Above 5 km, a combination of differential radiative fluxes at the top and bottom of the anvil and residual condensational heating from the main updraft maintains weak rising motion. This updrafts-over-downdrafts arrangement requires horizontal convergence centered near 5 km altitude to maintain mass continuity. The important kinematic consequence is formation of patchy shallow vortices near the altitude of the 0 °C isotherm. The typical horizontal scales of these ‘mesovortices’ are tens to hundreds of kilometers. If they were at the surface or if their influence could be extended downward to the surface then they would be the means to get the system to the required finite amplitude.

The foregoing reasoning defines the important unanswered questions: (1) how do the midlevel mesovortices extend their influence to the surface, and (2) what are the detailed thermodynamics at the air–sea interface during this process? Leading hypotheses for (1) are related to processes that can increase the surface vorticity through changes in static stability and momentum mixing, both horizontally and vertically. However, the answers to these questions await new measurements that are just becoming available through improved observational tools.

Basic Structure

Primary and Secondary Circulations

Inner core dynamics have received a lot of attention over the last 40 years through aircraft observations of the inner core structure. These observations show that the tropical cyclone inner core dynamics are dominated by interactions between ‘primary’ (horizontal axisymmetric), ‘secondary’ (radial and vertical) circulations, and a wavenumber one asymmetry caused by the storm motion. The primary circulation is so strong in the cyclone core that it is possible to consider axisymmetric motions separately, if account is taken of forcing by the asymmetric motions. The primary circulation is in near-gradient balance, and evolves when heat and angular momentum sources (often due to asymmetric motions) force secondary circulations, which in turn redistribute heat and angular momentum.

Figure 6 shows that the primary circulation is sustained by the secondary circulation that consists of frictional inflow that loses angular momentum to the sea as it gains moist enthalpy. (Angular momentum $M = Vr + fr^2/2$, where V is the tangential wind velocity, f the Coriolis parameter, and r the radius from the storm center.) The inflow picks up latent heat through evaporation, and exchanges sensible heat with the underlying ocean, as it spirals into lower levels of the storm under influence of friction. The evaporation of sea spray adds moisture to the air, while at the same time cooling it. This process is important in determining the intensity of a tropical cyclone. Near the vortex center, the inflow turns upward and brings the latent heat it acquires in the boundary layer into the free atmosphere. Across the top of the boundary layer, turbulent eddies cause significant downward flux of sensible heat from the free atmosphere to the boundary layer. The energy source for the turbulent eddies is mechanical mixing caused by the strong winds. The eddies are also responsible for downward mixing of angular momentum. Hence, these turbulent eddy fluxes fuel the storm.

As the air converges towards the eye and is lifted in convective clouds that surround the clear eye, it ascends to the tropopause (the top of troposphere, where temperature stops

decreasing with height). As shown in Figure 6, the convective updrafts in the eyewall turn the latent heat into sensible heat through the latent heat of condensation to provide the buoyancy needed to loft air from the surface to tropopause level. The updraft entrains midlevel air, promoting mass and angular momentum convergence into the core. It is the midlevel inflow that supplies the excess angular momentum needed to spin up the vortex. The thermodynamics of a storm can be modeled as an idealized heat engine, running between a warm heat reservoir, the sea, at around 300 K, and a cold reservoir, 15–18 km up in the tropical troposphere, at about 200 K. The net energy realized in the whole process is proportional to the difference in temperature between the ocean and the upper troposphere. Storm-induced upwelling of cooler water reduces ocean SST by a few degrees, which has a considerable effect on the storm’s intensity.

As shown in Figure 7, the secondary circulation also controls the distribution of hydrometeors and radar reflectivity. It is much weaker than the primary circulation except in the anticyclonic outflow, where the vortex is also much more asymmetric. Precipitation-driven convective updrafts form as hydrometeors fall from the outward sloping updraft. Condensation in the anvil causes a mesoscale updraft above the 0 °C isotherm and precipitation loading by snow falling from the overhanging anvil causes a mesoscale downdraft below 0 °C isotherm. The melting level itself marks the height of maximum mass convergence. Inside the eye, dynamically driven descent and momentum mixing leads to substantial pressure falls.

In order for the primary circulation to intensify, the flow cannot be in exact balance. Vertical gradients of angular momentum due to vertical shears of the primary circulation cause updrafts to pass through the convective heat sources, because the path of least resistance for the warmed air lies along constant angular momentum surfaces. Similarly, horizontal temperature gradients due to vertical shears cause the horizontal flow to pass through momentum sources, because the path of least resistance lies along isentropes (potential temperature or θ surfaces). Although the flow lies generally along the angular

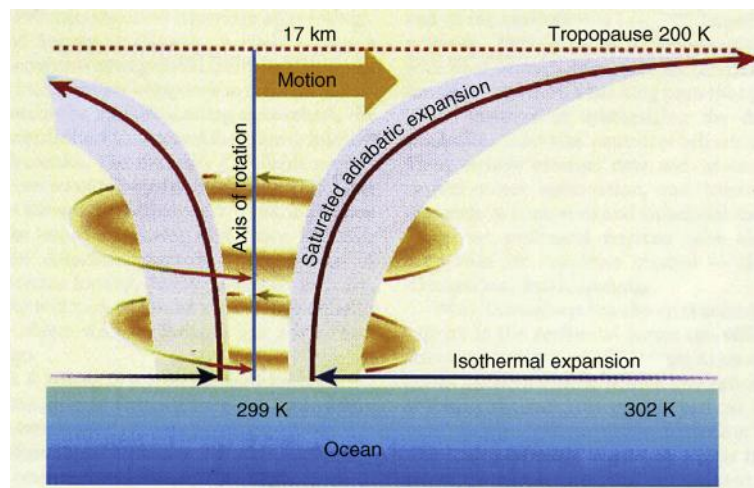


Figure 6 Schematic of the secondary circulation thermodynamics. Reproduced with permission from Willoughby HE (1999) Hurricane heat engines. *Nature* 401: 649–650; © Macmillan Magazines Ltd.

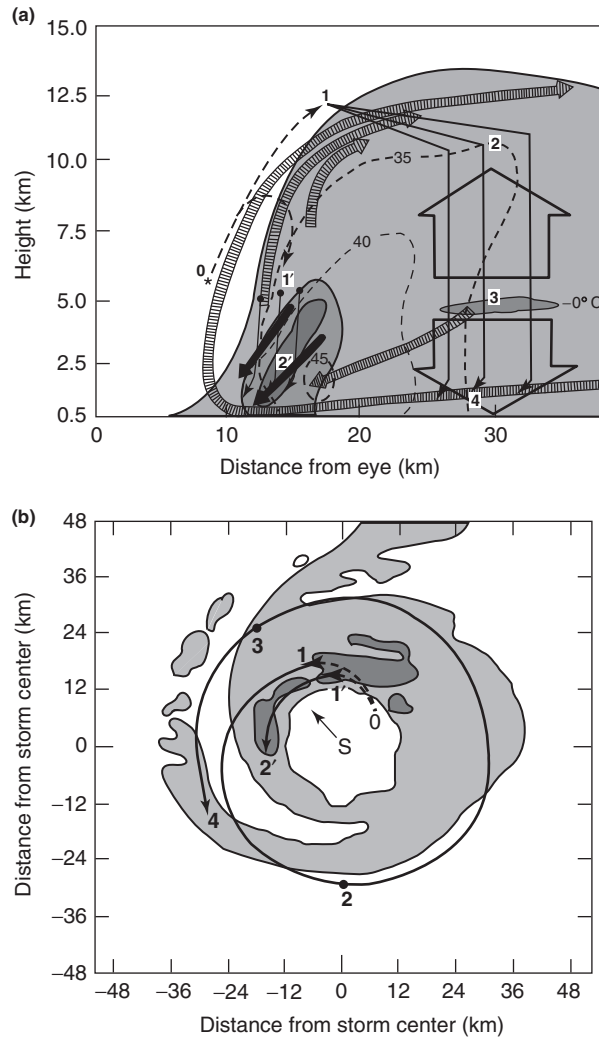


Figure 7 (a) Schematic of the radius–height circulation of the inner core of Hurricane Alicia (1983). Shading depicts the reflectivity field, with contours of 5, 30, and 35 dBZ. The primary circulation (m s^{-1}) is depicted by dashed lines and the secondary circulation by the wide hatched streamlines. The convective downdrafts are denoted by the thick solid arrows, while the mesoscale up- and downdrafts are shown by the broad arrows. (b) Schematic plan view of the low-level reflectivity field in the inner core of Hurricane Alicia superimposed with the middle of the three hydrometeor trajectories in (a). Reflectivity contours in (b) are 20 and 35 dBZ. The storm center and direction are also shown. In (a) and (b) the hydrometeor trajectories are denoted by dashed and solid lines labeled 0-1-2-3-4 and 0'-1'-2'. Reproduced with permission from Marks FD and Houze RA (1987) Inner core structure of Hurricane Alicia from airborne Doppler radar observations. *Journal of the Atmospheric Sciences* 44: 1296–1317; © American Meteorological Society.

momentum or isentropic surfaces, it has a small component across them. The advection by this component, not the direct forcing, is the mechanism by which the primary circulation evolves.

Some of the most intense tropical cyclones exhibit ‘concentric’ eyewalls, i.e., two or more eyewall structures centered at the circulation center of the storm. In much the same way as the

inner eyewall forms, convection surrounding the eyewall can become organized into distinct rings. Eventually, the inner eye begins to feel the effects of the subsidence resulting from the outer eyewall, and the inner eyewall weakens, to be replaced by the outer eyewall. The pressure rises due to the destruction of the inner eyewall are usually more rapid than the pressure falls, due to the intensification of the outer eyewall, and the cyclone itself weakens for a short period of time. This mechanism, referred to as the eyewall replacement cycle, often accompanies dramatic changes in storm intensity. The intensity changes are often associated with the development of secondary wind maxima outside the storm core.

A good example of contracting rings of convection effecting the intensification of a hurricane is shown in **Figure 8** for Hurricane Gilbert on 14 September 1988. Two convective rings, denoted by intense radar reflectivity, are evident in **Figure 8(a)**. The outer ring is located near 80–90 km radius and the inner one at 10–12 km radius. **Figure 8(b)** shows that both are associated with maxima in tangential wind and vorticity. **Figure 9** shows that in the ensuing 12–24 h the storm filled

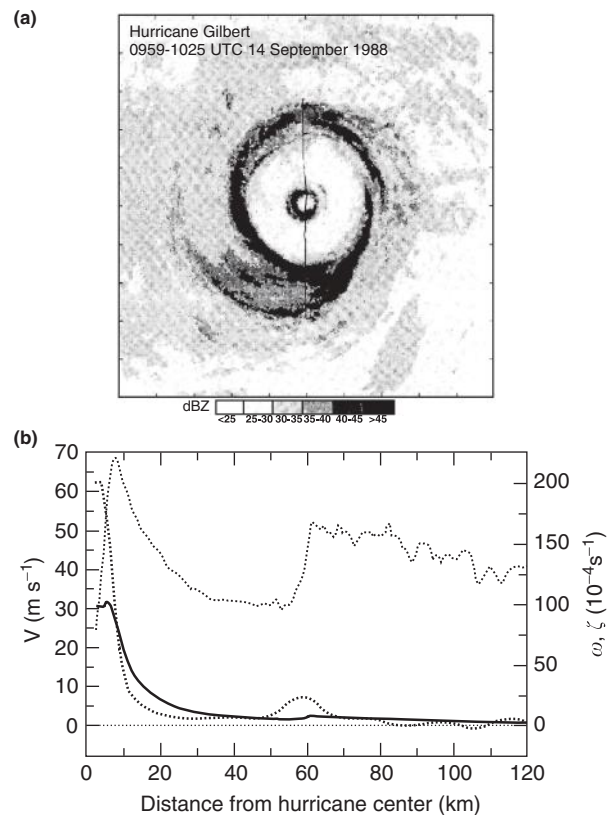


Figure 8 (a) Composite horizontal radar reflectivity of Hurricane Gilbert for 0959–1025 UTC, 14 September 1988; the domain is $360 \text{ km} \times 360 \text{ km}$, marked every 36 km. The line through the center is the WP-3D aircraft flight track. (b) Profiles of flight-level angular velocity (ω , solid), tangential wind (short dash), and smoothed relative vorticity (ζ , long dash) along the southern leg of the flight track shown in (a). Reproduced with permission from Kossin JP, Schubert WH, and Montgomery MT (2000) Unstable interactions between a hurricane’s primary eyewall and a secondary ring of enhanced vorticity. *Journal of the Atmospheric Sciences* 57: 3893–3917; © American Meteorological Society.

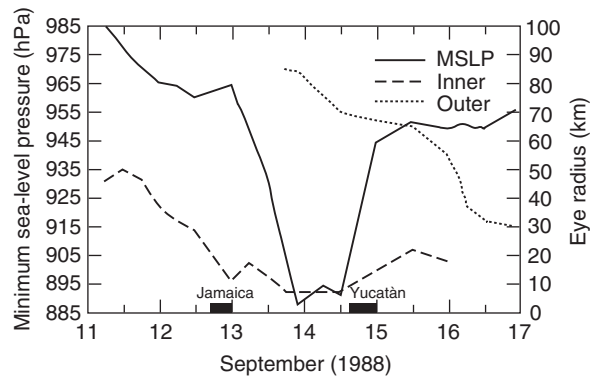


Figure 9 Hurricane Gilbert's minimum sea-level pressure (MSLP) and radii of the inner and outer eyewalls as a function of time, September 1988. Solid blocks at bottom indicate times over land. Reproduced with permission from Black ML and Willoughby HE (1999) The concentric eyewall cycle of Hurricane Gilbert. *Monthly Weather Review* 120: 947–957; © American Meteorological Society.

dramatically. However, it is not clear how much of the filling was caused by the storm moving over land and how much by the contracting outer ring and decaying inner ring of convective activity.

A process has been proposed whereby: (1) nonlinear balanced adjustment of the vortex to eddy heat and angular momentum sources generated by some environmental interaction in the storm's periphery produces an enhanced secondary circulation; (2) a secondary wind maximum develops in response; and (3) the wind maximum contracts as a result of differential adiabatic warming associated with the convective diabatic heating in the presence of an inward radial gradient of inertial stability. Under these circumstances, understanding the intensification of the tropical cyclone

reduces to determining what mechanisms can produce an enhanced secondary wind maximum.

Inner core – eyewall and eye

The most recognizable feature found within a hurricane is the eye (**Figure 10**). It is found at the center and is typically between 20–50 km in diameter. The eye is the focus of the hurricane, the point about which the primary circulation rotates and where the lowest surface pressures are found in the storm. The eye is a roughly circular area of comparatively light winds and fair weather found at the center of strong tropical cyclones. Although the winds are calm at the axis of rotation, strong winds may extend well into the eye. As seen in **Figure 10**, there is little or no precipitation and sometimes blue sky or stars can be seen. The eye is the region of warmest temperatures aloft – the eye temperature may be $\geq 10^{\circ}\text{C}$ warmer at an altitude of 12 km than the surrounding environment, but only $0\text{--}2^{\circ}\text{C}$ warmer at the surface.

The eye is surrounded by the eyewall, the roughly circular area of deep convection associated with the up-branch of the secondary circulation and the highest surface winds. The eye is composed of air that is slowly sinking and the eyewall has a net upward flow because of many moderate – occasionally strong – updrafts and downdrafts. The eye's warm temperatures are due to warming by compression of the subsiding air. Most soundings taken within the eye are similar to that for Hurricane Hugo in **Figure 11**. They show a low-level layer which is relatively moist, with an inversion above, suggesting that the sinking in the eye typically does not reach the ocean surface, but instead gets only within 1–3 km of the surface. An eye is usually present only in hurricane-strength tropical cyclones.

The general mechanisms by which the eye and eyewall are formed are not fully understood, although observations shed some light on the problem. The calm eye of the tropical

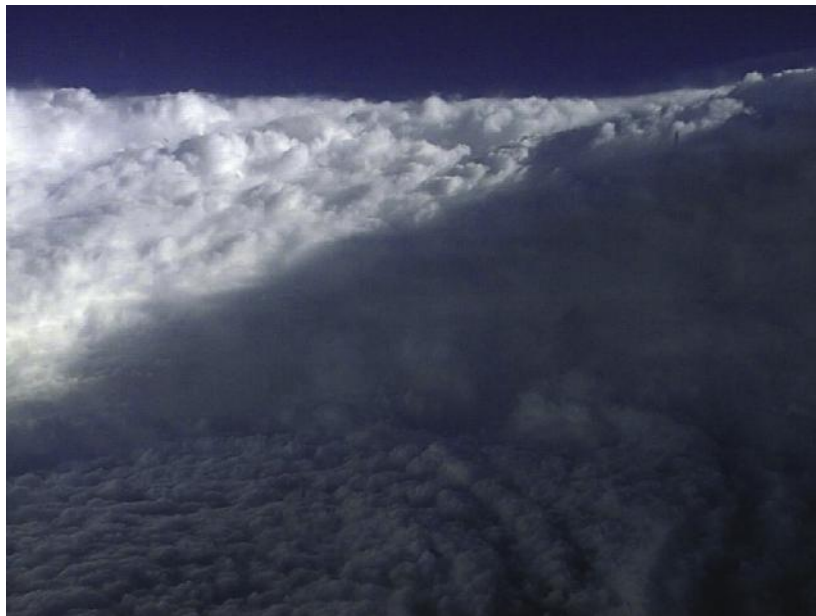


Figure 10 Eyewall of Hurricane Georges, 1945 UTC, 19 September 1998. Photo courtesy of M. Black, NOAA/OAR/AOML Hurricane Research Division.

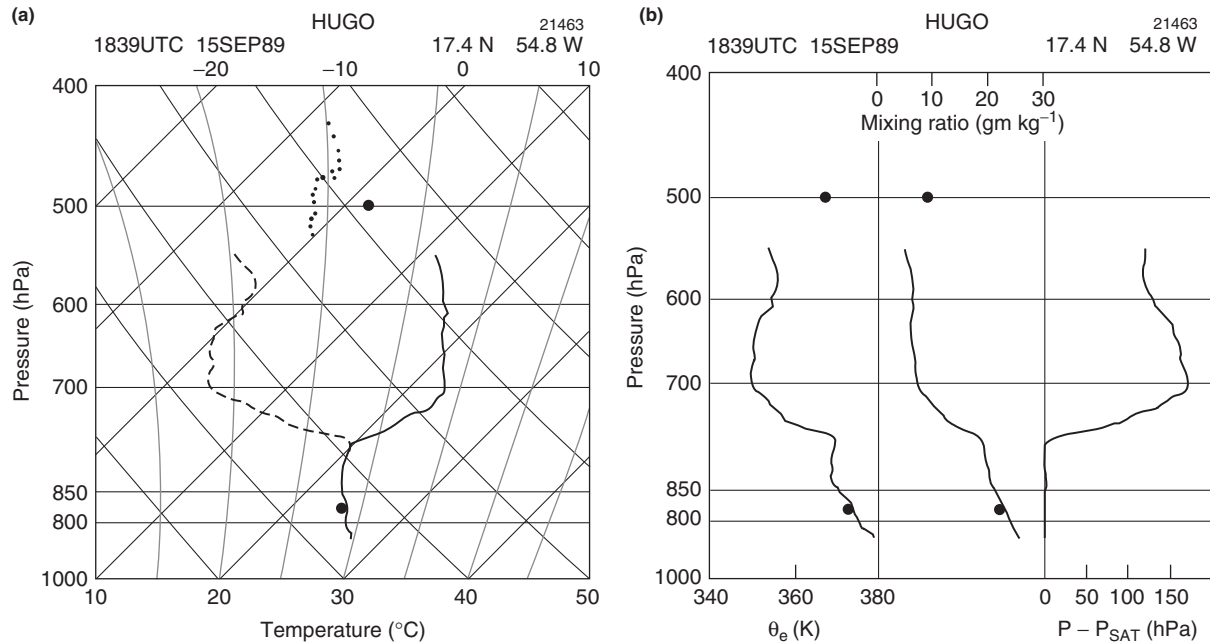


Figure 11 (a) Skew $T \lg p$ diagram of the eye sounding in Hurricane Hugo at 1839 UTC, 15 September 1989, 17.4° N, 54.8° W. Isotherms slope upward to the right; dry adiabats slope upward to the left; moist adiabats are nearly vertical curving to the left. Solid and dashed curves denote temperature and dew point, respectively. The smaller dots denote saturation points computed for the dry air above the inversion, and the two larger dots temperature observed at the innermost saturated point as the aircraft passed through the eyewall. (b) θ_e , water vapor mixing ratio, and saturation pressure difference, $P - P_{SAT}$, as functions of pressure at 2123 UTC. Reproduced with permission from Willoughby HE (1998) Tropical cyclone eye thermodynamics. *Monthly Weather Review* 126: 3189–3211; © American Meteorological Society.

cyclone shares many qualitative characteristics with other vortical systems such as tornadoes, waterspouts, dust devils, and whirlpools. Given that many of these lack a change of phase of water (i.e., no clouds and diabatic heating are involved), it may be that the eye feature is a fundamental component to all rotating fluids. It has been hypothesized that supergradient wind flow (i.e., swirling winds generating stronger centrifugal ‘force’ than the local pressure gradient can support) present near the radius of maximum winds causes air to be centrifuged out of the eye into the eyewall, thus accounting for the subsidence in the eye. However, others found that the swirling winds within several tropical cyclones were within 1–4% of gradient balance. It may be thought that the amount of supergradient flow needed to cause such centrifuging of air is only on the order of a couple of percent and thus difficult to measure.

Another feature of tropical cyclones that probably plays a role in forming and maintaining the eye is the eyewall convection. As shown in Figure 12, convection in developing tropical cyclones is organized into long, narrow rainbands that are oriented in the same direction as the horizontal wind. Because these bands seem to spiral into the center of a tropical cyclone, they are sometimes called spiral bands. The earliest radar observations of tropical cyclones detected these bands, which are typically 5–50 km wide and 100–300 km long. Along these bands, low-level convergence is a maximum, and therefore upper-level divergence is most pronounced. A direct circulation develops in which warm, moist air converges at the surface, ascends through these bands, diverges aloft, and descends on both sides of the bands. Subsidence is distributed over a wide

area outside of the rainband, but is concentrated in the small inside area. As the air subsides, adiabatic warming takes place, and the air dries. Because subsidence is often concentrated on the inside of the band, the adiabatic warming is stronger inward from the band, causing a sharp contrast in pressure falls across the band since warm air is lighter than cold air. Because of the pressure falls on the inside, the tangential winds around the tropical cyclone increase, owing to an increased pressure gradient. Eventually, the band moves toward the center and encircles it and the eye and eyewall form.

The circulation in the eye is comparatively weak and, at least in the mature stage, thermally indirect (warm air descending), so it cannot play a direct role in the storm energy production. On the other hand, the temperature in the eye of many hurricanes exceeds that which can be attained by any conceivable moist adiabatic ascent from the sea surface, even accounting for the additional entropy (positive potential temperature, θ , anomaly) owing to the low surface pressure in the eye (the lower the pressure, the higher the θ at a given altitude and temperature). Thus, the observed low central pressure of the storm is not consistent with that calculated hydrostatically from the temperature distribution created when a sample of air is lifted from a state of saturation at sea surface temperature and pressure. The thermal wind balance restricts the amount of warming that can take place. In essence, the rotation of the eye at each level is imparted by the eyewall, and the pressure drop from the outer to the inner edge of the eye is simply that required by gradient balance.

Because the eyewall azimuthal velocity decreases with height, the radial pressure drop decreases with altitude, requiring,

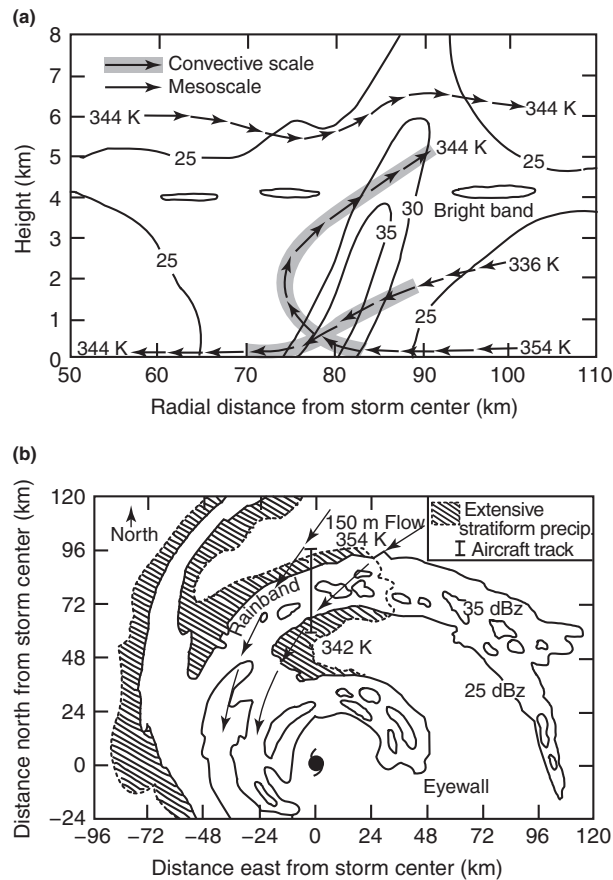


Figure 12 (a) Schematic of the rainband in radius–height coordinates. Reflectivity, θ_e , mesoscale (arrows), and convective scale motions are shown. (b) Plan view. Aircraft track, reflectivities, cells, stratiform precipitation, 150 m flow, and θ_e values are shown. Reproduced with permission from Barnes GM, Zipser EJ, Jorgensen DP, and Marks FD (1983) Mesoscale and convective structure of a hurricane rainband. *Journal of the Atmospheric Sciences* 40: 2125–2137; © American Meteorological Society.

through the hydrostatic equation, a temperature maximum at the storm center. Thus, given the swirling velocity of the eyewall, the steady-state eye structure is largely determined. The central pressure, which is estimated by integrating the gradient balance equation inward from the radius of maximum winds, depends on the assumed radial profile of azimuthal wind in the eye.

In contrast, the eyewall is a region of rapid variation of thermodynamic variables. As shown in Figure 13, the transition from the eyewall cloud to the nearly cloud-free eye is often so abrupt that it has been described as a form of atmospheric front. Early studies were the first to recognize that the flow under the eyewall cloud is inherently frontogenetic. The eyewall is the upward branch of the secondary circulation and a region of rapid ascent that, together with slantwise convection, leads to the congruence of angular momentum and moist entropy (θ_e) surfaces. Hence, the three-dimensional vorticity vectors lie on θ_e surfaces, so that the moist PV vanishes. As the air is saturated, this in turn implies, through the invertibility principle applied to flow in gradient and hydrostatic balance, that the entire primary circulation may be deduced from the radial distribution of θ_e in the boundary layer and the distribution of vorticity at the tropopause.

In the classic semigeostrophic theory of deformation-induced frontogenesis, the background geostrophic deformation flow provides the advection of temperature across surfaces of absolute momentum that drives the frontogenesis, whereas in the hurricane eyewall, surface friction provides the radial advection of entropy across angular momentum surfaces. Also note that the hurricane eyewall is not necessarily a front in surface temperature, but instead involves the θ_e distribution, which is related directly to density in saturated air.

There is likely a two-stage process in eye formation. The amplification of the primary circulation is strongly frontogenetic and results, in a comparatively short time, in frontal collapse at the inner edge of the eyewall. (Frontal collapse is an increase in the horizontal gradient of an airmass property, principally density, and the development of the accompanying features of the wind field through the secondary circulation that typify a front.) The frontal collapse leads to a dramatic transition in the storm dynamics. While the tropical cyclone inner core is dominated by axisymmetric motions, hydrodynamic instabilities are potential sources of asymmetric motions within the

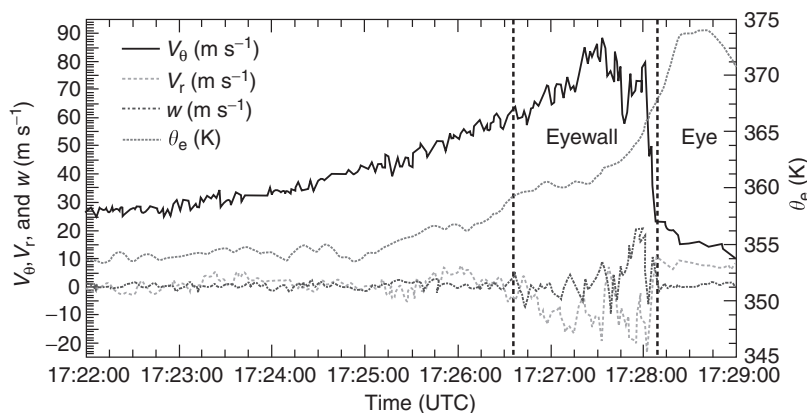


Figure 13 Time series plots of tangential wind (V_θ), radial wind (V_r), vertical velocity (w), and θ_e in Hurricane Hugo at 1721–1730 UTC, 15 September 1989. The aircraft flight track was at 450 m. Thick dashed vertical lines denote the width of the eyewall reflectivity maximum at low levels.

core. In intense tropical cyclones the wind profile inside the eye is often U-shaped, in the sense that the wind increases outwards more rapidly than linearly with radius (Figure 13). The strong cyclonic shear just inside the eyewall may result in a local maximum of absolute vorticity or angular momentum, so that the profile may actually become barotropically unstable. (This refers to the hydrodynamic instability arising from certain distributions of vorticity in a two-dimensional nondivergent flow. It is an inertial instability in that kinetic energy is the only form of energy transferred between the current and perturbation. A well-known necessary condition for barotropic instability is that the basic state vorticity gradient must have both signs in the domain of interest.) This instability leads to frontal collapse as a result of radial diffusion of momentum into the eye, and also may explain the ‘polygonal eyewalls’ where the eyewall appear on radar to be made up of a series of line segments rather than as a circle. It may also explain intense mesoscale vortices observed in the eyewalls of Hurricanes Hugo of 1989 and Andrew of 1992.

Once the radial turbulent diffusion of momentum driven by the instability of the primary circulation becomes important, it results in a mechanically induced, thermally indirect (warm air sinking) component of the secondary circulation in the eye and eyewall. Such a circulation raises the vertically averaged temperature of the eye beyond its value in the eyewall and allows for an amplification of the entropy distribution. Feedbacks with the surface fluxes then allow the boundary layer entropy to increase and result in a more rapid intensification of the swirling wind. Thus, the frontal collapse of the eyewall is an essential process in the evolution of tropical cyclones. Without it, amplification of the temperature distribution relies on external influences, and intensification of the wind field is slow. Once it has taken place, the mechanical spinup of the eye allows the temperature distribution to amplify without external influences and, through positive feedback with surface fluxes, allows the entropy field to amplify and the swirling velocity to increase somewhat more rapidly.

Outer structure and rainbands

The axisymmetric core is characteristically surrounded by a less symmetric outer vortex that diminishes into the synoptic ‘environment’. In the lower troposphere, the cyclonic circulation may extend more than 1000 km from the center. As evident in Figure 14 the boundary between cyclonic and anticyclonic circulation slopes inward with increasing height, so that the circulation in the upper troposphere is primarily anticyclonic, except near the core. In the outer vortex, there are no scale separations between the primary and the secondary circulations, the asymmetric motions, or the vortex translation, as they are roughly all of the same magnitude. The asymmetric flows in this region control the vortex motion and sustain an eddy convergence of angular momentum and moisture toward the center. Interactions between the symmetric motions of the inner core with the more asymmetric motions in the outer portion of the storm are the key to improved forecasts of tropical cyclone track and intensity.

Spiral bands of precipitation characterize radar and satellite images in this region of the storm (Figure 1 and Figure 15). As seen in Figure 8, Figure 12 and Figure 15, radar reflectivity

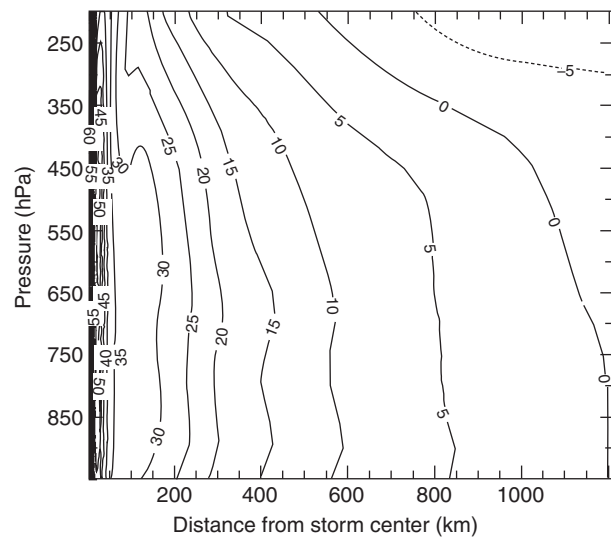


Figure 14 Vertical cross-section of the azimuthal mean tangential wind for Hurricane Gloria on 24 September 1985. Anticyclonic contours are dashed. Reproduced with permission from Franklin JL, Lord SJ, Feuer SE, and Marks FD (1993) The kinematic structure of Hurricane Gloria (1985) determined from nested analyses of dropwindsonde and Doppler radar data. *Monthly Weather Review* 121: 2433–2451; © American Meteorological Society.

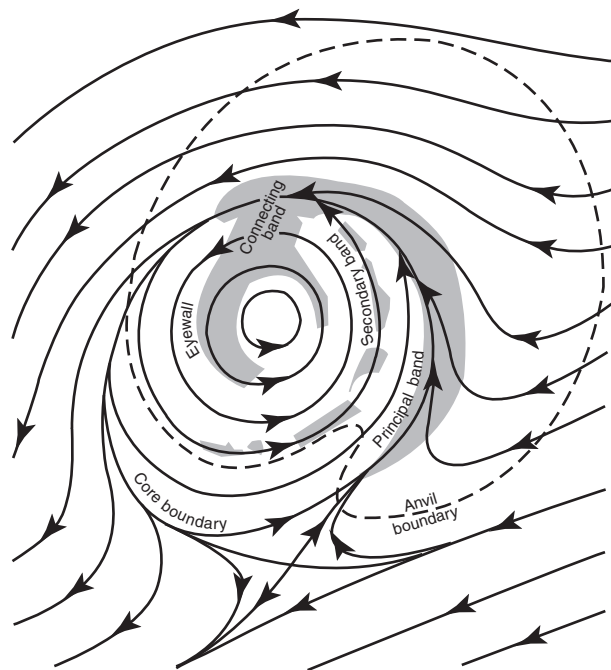


Figure 15 Schematic representation of the stationary band complex, the entities that compose it, and the flow in which it is embedded. Reproduced with permission from Willoughby HE, Marks FD, and Feinberg RJ (1984) Stationary and moving convective bands in hurricanes. *Journal of the Atmospheric Sciences* 41: 3189–3211; © American Meteorological Society.

patterns in tropical cyclones provide a good means for flow visualization, although they represent precipitation, not winds. Descending motion occupies precipitation-free areas, such as the eye. The axis of the cyclone's rotation lies near the center of the eye. The eyewall surrounds the eye. In intense hurricanes, it may contain reflectivities as high as 50 dBZ equivalent to rainfall rates of 74 mm h^{-1} . (1 dBZ $10 \lg Z$, where Z is equivalent radar reflectivity factor ($\text{mm}^6 \text{ m}^{-3}$).) Less extreme reflectivities, 40 dBZ (13 mm h^{-1}), characterize most convective rainfall in the eyewall and spiral bands. The vertical velocities (both up and downdrafts) in convection with highest reflectivity may reach 25 m s^{-1} , but typical vertical velocities are $<5 \text{ m s}^{-1}$. Such intense convection occupies less than 10% of the tropical cyclone's area. Outside convection, reflectivities are still weaker, 30 dBZ, equivalent to a 2.4 mm h^{-1} rain rate. This 'stratiform rain', denoted by a distinct reflectivity maximum or 'bright band' at the altitude of the 0°C isotherm, falls out of the anvil cloud that grows from the convection. The spiral bands tend to lie along the friction-layer wind that spirals inward toward the eyewall (Figure 12).

Many aspects of rainband formation, dynamics, and interaction with the symmetric vortex are still unresolved. The trailing spiral shape of bands and lanes arises because the angular velocity of the vortex increases inward and deforms them into equiangular spirals. In the vortex core, air remains in the circulation for many orbits of the center, while outside the core, the air passes through the circulation in less than the time required for a single orbit. As the tropical cyclone becomes more intense, the inward ends of the bands approach the center less steeply approximating arcs of circles. Some bands appear to move outward, while others maintain a fixed location relative to the translating center.

As shown in Figure 15, motion of the vortex through its surroundings may cause one stationary band, called the principal band, to lay along a convergent streamline asymptote that spirals into the core. A tropical cyclone advected by midlevel steering with westerly shear moves eastward through surrounding air at low levels. Thus, the principal band may be like a bow wave, caused by the displacement of the environmental air on the eastern side of the vortex. Its predominant azimuthal wavenumber is one.

Moving bands, and other convective features, are frequently associated with cycloidal motion of the tropical cyclone center, and intense asymmetric outbursts of convection are observed to displace the tropical cyclone center by tens of kilometers. The bands observed by radar are often considered manifestations of internal gravity waves, but these waves can exist only in a band of Doppler-shifted frequencies between the local inertia frequency (defined as the sum of the vertical component of the earth's inertial frequency, f , and the local angular velocity of the circulation, V/r) and the Brunt-Väisälä frequency (i.e., the natural gravity wave frequency, the square root of the static stability defined as $(g/\theta) \partial\theta/\partial z$). Only two classes of trailing-spiral, gravity wave solutions lie within this frequency band: (1) waves with any tangential wavenumber that move faster than the swirling wind, and (2) waves with tangential wavenumber ≥ 2 that move slower than the swirling wind.

Bands moving faster than the swirling wind with outward phase propagation are observed by radar. They are more like squall lines than linear gravity waves. Waves moving slower

than the swirling wind propagate wave energy and anticyclonic angular momentum inward, grow at the expense of the mean-flow kinetic energy, and reach appreciable amplitude if they are excited at the periphery of the tropical cyclone. Alternate explanations for these inward-propagating bands involve filamentation of vorticity from the tropical cyclone environment, asymmetries in the radially shearing flow of the vortex, and high-order vortex Rossby waves. Detailed observations of the vortex-scale rainband structure and windfield are necessary to determine which mechanisms play a role in rainband development and maintenance.

While the evolution of the inner core is dominated by interactions between the primary, secondary, and track-induced wavenumber-one circulation, there is some indication that the local convective circulations in the rainbands may impact on intensity change. Although precipitation in some bands is largely stratiform, condensation in most bands tends to be concentrated in convective cells rather than spread over wide mesoscale areas. As shown in Figure 12, convective elements form, move through the bands, and dissipate as they move downwind. Doppler radar observations indicate that the roots of the updrafts lie in convergence between the low-level radial inflow and gust fronts produced by convective downdrafts. This convergence may occur on either side of the band. A 20 K decrease in low-level θ_e was observed in a rainband downdraft, suggesting that the draft acts as a barrier to inflow. This reduction in boundary layer energy may be advected near the center, inhibit convection, and thereby alter storm intensity.

Motion

Tropical cyclone motion is the result of a complex interaction between a number of internal and external influences. Environmental steering is typically the most prominent external influence on a tropical cyclone, accounting for as much as 70–90% of the motion. Theoretical studies show that in the absence of environmental steering, tropical cyclones move poleward and westward owing to internal influences.

Accurate determination of tropical cyclone motion requires accurate representation of interactions that occur throughout the depth of the troposphere on a variety of scales. Observations spurred improved understanding of how tropical cyclones move using simple barotropic and more complex baroclinic models. To first order, the storm moves with some layer average of the lower-tropospheric environmental flow: the translation of the vortex is roughly equal to the speed and direction of the basic 'steering' current. However, the observations show that tropical cyclone tracks deviate from this simple steering flow concept in a subtle and important manner. Several physical processes may cause such deviations. The approach in theoretical and modeling of tropical cyclones has been to isolate each process in a systematic manner to understand the magnitude and direction of the track deviation caused by each effect. The β effect opposes the advection of relative vorticity through the differential advection of the Earth's vorticity, f , that slows the advection of the disturbance. (The β effect is the asymmetric vorticity advection around the vortex caused by the latitudinal gradient of f ; $\beta = 2\Omega \cos \phi$. β has a maximum value at the Equator (i.e., $2.289 \times 10^{-11} \text{ s}^{-1}$)

and becomes zero at the pole.) Models that are more complete describe not only the movement of the vortex but also the accompanying wavenumber-one asymmetries that develop owing to the differential advection of f on the east and west side of the vortex. It was also discovered that the role of meridional and zonal gradients of the environmental flow could add greatly to the complexity even in the barotropic evolution of a vortex. Hence, the evolution of the movement depends on not only the relative vorticity gradient and on shear of the environment but also the structure of the vortex itself.

Generally, the propagation vector of these model baroclinic vortices is very close to that expected from a barotropic model initialized with the vertically integrated environmental wind. An essential feature in baroclinic systems is the relative vorticity advection through the storm center, where the vertical structure of the tropical cyclone produces a tendency for the low-level vortex to move slower than the simple propagation of the vortex due to β . Vertical shear plays an important factor in determining what the relative flow is, though there is no unique relation between the shear and storm motion. Diabatic heating effects also alter this flow and change the propagation velocity. Thus, tropical cyclone motion is primarily governed by the dynamics of the low-level cyclonic circulation; however, the addition of observations of the upper-level structure may alter this finding.

Interaction with the Atmospheric Environment

A consensus exists that small vertical shear of the environmental wind and lateral eddy imports of angular momentum are favorable to tropical cyclone intensification. The inhibiting effect of vertical shear in the environment on tropical cyclone intensification is well known from climatology and forecasting experience. Exposure of the low-level circulation from under the tropical cyclones large area of cirrus (central dense overcast) in satellite imagery is universally recognized as a symptom of shear, and as an indication of a nonintensification or weakening. Nevertheless, the detailed dynamics of a vortex in shear has been the topic of surprisingly little study, probably because while the effect is a reliable basis for practical forecasting, it is difficult to measure and model.

In contrast, the positive effect of eddy momentum imports at upper levels has received extensive study. Modeling studies with composite initial conditions show that eddy momentum fluxes can intensify a tropical cyclone even when other conditions are neutral or unfavorable. It has been shown theoretically that momentum imports can form a tropical cyclone in an atmosphere with no buoyancy. Statistical analysis of tropical cyclones reveals a clear relationship between angular momentum convergence and intensification, but only after the effects of shear and SST variations are accounted for. Such interactions occur frequently (35% of the time, defined by eddy angular momentum flux convergence exceeding $10 \text{ m s}^{-1} \text{ day}^{-1}$), and likely represent the more common upper-boundary interaction for tropical cyclones. Frequently they are accompanied by eyewall cycles and dramatic intensity changes.

The environmental flows that favor intensification, and presumably inward eddy momentum fluxes, usually involve interaction with a synoptic-scale cyclonic feature, such as a midlatitude upper-level trough or PV anomaly. Given the

interaction of an upper-level trough and the tropical cyclone, the exact mechanism for intensification is still uncertain. The secondary circulation response to momentum and heat sources is very different. Upper-tropospheric momentum sources can influence the core directly. As can be seen in Figure 16, large inertial stability in the lower troposphere protects the mature tropical cyclone core from direct influence by momentum sources (inertial stability is a measure of the resistance to horizontal displacements, based on the conservation of angular momentum for a vortex in gradient balance, and is defined as $(f + \zeta)(f + V^2/r)$, where ζ is the relative vorticity, V the axial wind velocity, f the Coriolis parameter, and r the radius from the storm center); however, the inertial stability in the upper troposphere is smaller and a momentum source can induce an outflow jet with large radial extent just below the tropopause. If the eyewall updraft links to the direct circulation at the entrance region of the jet, as shown in Figure 17(c) and 17(d), the exhaust outflow is unrestricted. The important difference between heat and momentum sources is that the roots of the diabatically induced updraft must be in the inertially stiff lower troposphere, but the outflow jet due to a momentum-flux convergence can be confined to the inertially stable upper troposphere. Momentum forcing does not spin the vortex up directly. It makes the exhaust flow stronger and reduces local compensating subsidence in the core, thus cooling the upper troposphere and destabilizing the sounding. The cooler upper troposphere leads to less thermal-wind shear and a weaker upper anticyclone.

A two-dimensional balanced approach provides reasonable insight into the nature of the tropical cyclone intensification as a trough approaches. Isentropic PV analysis (Figure 17)), which express the problem in terms of a quasi-conserved variable in three dimensions, are used to describe various processes in

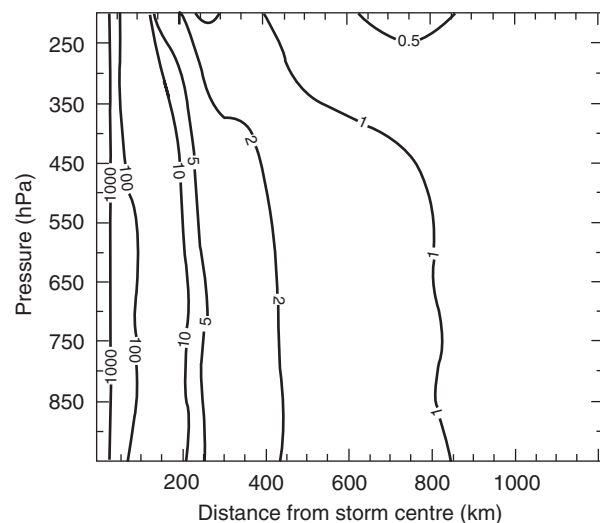


Figure 16 Axisymmetric mean inertial stability for Hurricane Gloria on 24 September 1985. Contours are shown as multiples of f^2 at the latitude of Gloria's center. Reproduced with permission from Franklin JL, Lord SJ, Feuer SE, and Marks FD (1993) The kinematic structure of Hurricane Gloria (1985) determined from nested analyses of dropwind-sonde and Doppler radar data. Monthly Weather Review 121: 2433–2451; © American Meteorological Society.

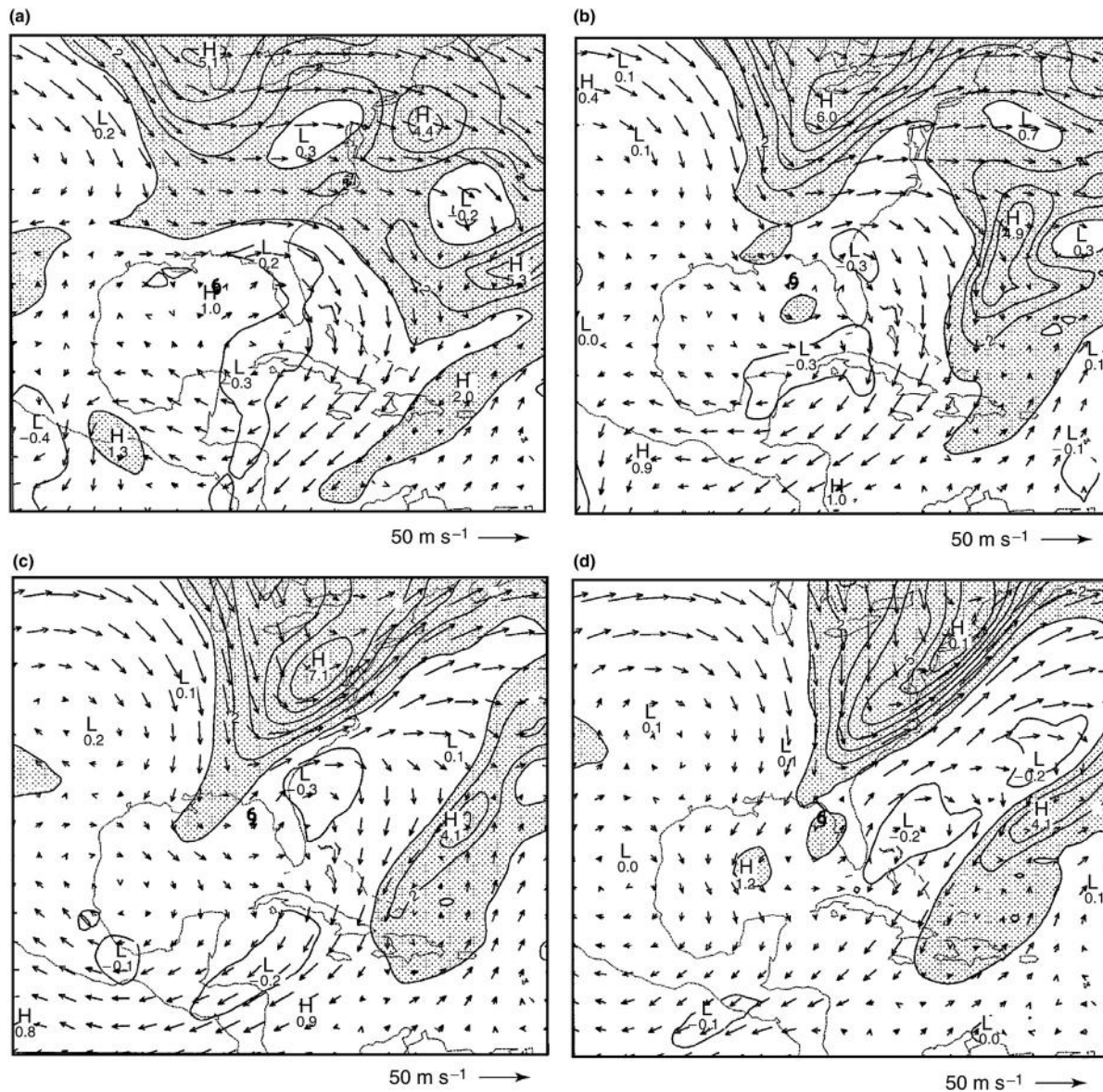


Figure 17 Wind vectors and potential vorticity on the 345 K isentropic surface at (a) 1200 UTC 30 August; (b) 0000 UTC 31 August; (c) 1200 UTC 31 August; and (d) 0000 UTC 1 September 1985. PV increments are 1 PVU and values >1 PVU are shaded. Wind vectors are plotted at 2.25° intervals. The 345 K surface is approximately 200 hPa in the hurricane environment and ranges from 240 to 280 hPa at the storm center. The hurricane symbol denotes the location of Hurricane Elena. Reproduced with permission from Molinari J, Skubis S, and Vollaro D (1995) External influences on hurricane intensity. 3. Potential vorticity structure. *Journal of the Atmospheric Sciences* 52: 3593–3606; © American Meteorological Society.

idealized tropical cyclones with considerable success. The eddy heat and angular momentum fluxes are related to changes in the isentropic PV through their contribution to the eddy flux of PV.

It has been suggested that outflow-layer asymmetries, as in [Figure 17](#), and their associated circulations could create a mid- or lower-tropospheric PV maximum outside the storm core, either by creating breaking PV waves on the mid-tropospheric radial PV gradient ([Figure 3](#)) or by diabatic heating. It has been shown that filamentation of any such PV maximum in the ‘surf zone’ outside the tropical cyclone core (the sharp radial PV gradient near 100 km radius) produces a feature much like a secondary wind maximum, which was apparent in the PV

fields of Hurricane Gloria in [Figure 3](#). These studies thus provide mechanisms by which outflow-layer asymmetries could bring about a secondary wind maximum.

An alternative argument has been proposed for storm reintensification as a ‘constructive interference without phase locking’, as shown in [Figure 18](#). As the PV anomalies come within the Rossby radius of deformation, the pressure and wind perturbations associated with the combined anomalies are greater than when the anomalies are apart, even though the PV magnitudes are unchanged. The perturbation energy comes from the basic-state shear that brought the anomalies together. However, constructive interference without some additional

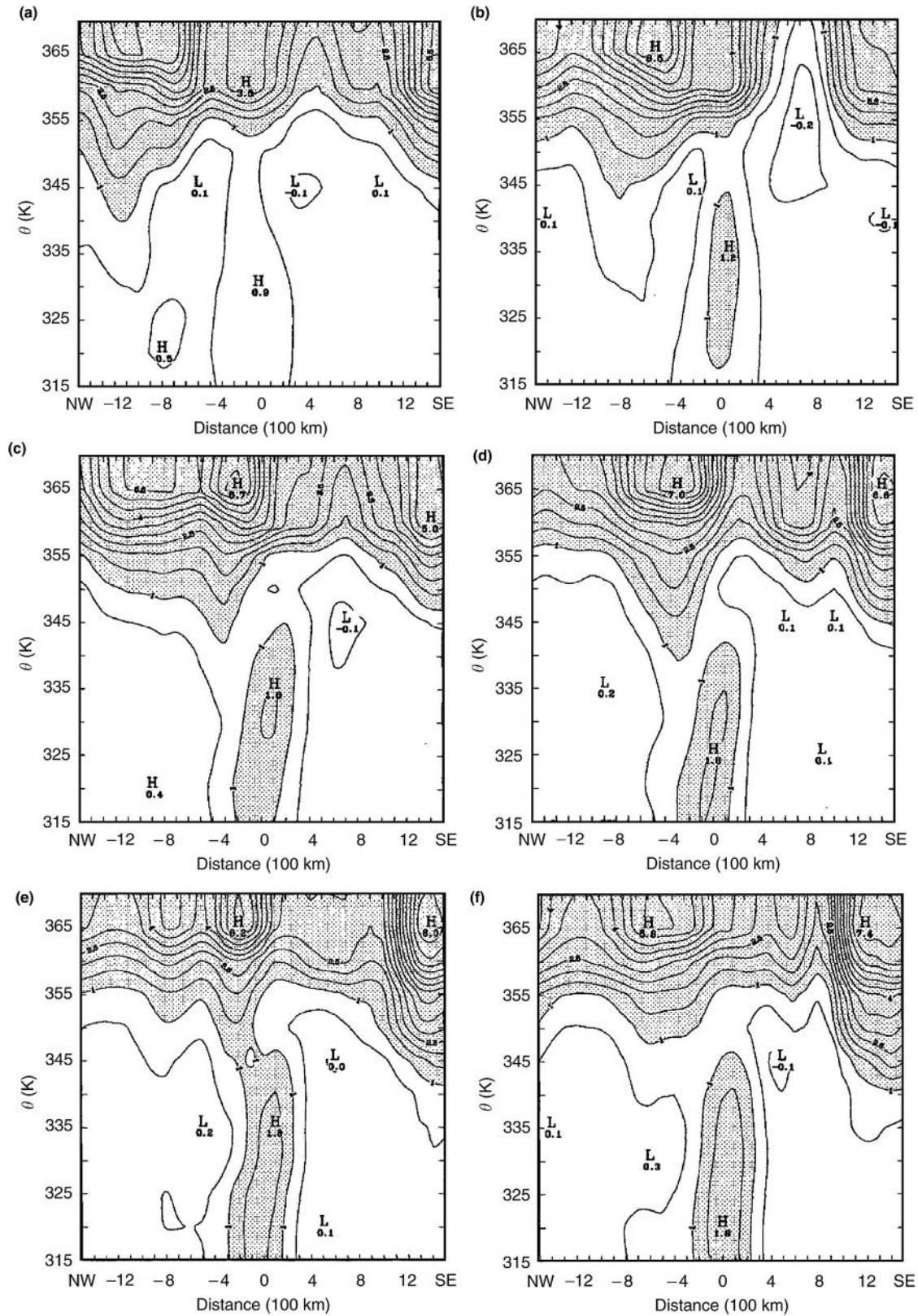


Figure 18 Cross-sections of potential vorticity from north-west (left) to south-east (right) through the observed center of Hurricane Elena at the same times as in Figure 17, plus (A) at 0000 UTC 30 August and (F) 1200 UTC 1 September. Increment is 0.5 PVU and shading above 1 PVU. Reproduced with permission from Molinari J, Skubis S, and Vollaro D (1995) External influences on hurricane intensity. 3. Potential vorticity structure. *Journal of the Atmospheric Sciences* 52: 3593–3606; © American Meteorological Society.

diabatic component cannot account for intensification. It is possible that intensification represents a baroclinic initiation of a wind-induced surface heat exchange. By this mechanism, the constructive interference induces stronger surface wind anomalies, which produce larger surface moisture fluxes and thus higher surface moist enthalpy. This feeds back through the associated convective heating to produce a stronger secondary circulation and thus stronger surface winds. The small effective static stability in the saturated, nearly moist neutral storm core ensures a deep response, so that even a rather narrow upper trough can initiate this feedback process. The key to this mechanism is the direct influence of the constructive interference on the surface wind field, as that controls the surface flux of moist enthalpy.

Interaction with the Ocean

As pointed out in the climatology section, preexisting SSTs $>26^{\circ}\text{C}$ are a necessary but insufficient condition for tropical cyclogenesis. Once the tropical cyclone develops and translates over the tropical oceans, statistical models suggest that warm SSTs describe a large fraction of the variance (40–70%) associated with the intensification phase of the storm. However, these predictive models do not account either for the oceanic mixed layers having temperatures of $0.5\text{--}1^{\circ}\text{C}$ cooler than the relatively thin SST over the upper meter of the ocean or horizontal advective tendencies by basic-state ocean currents such as the Gulf Stream and warm core eddies. Thin layers of warm SST are well mixed with the underlying cooler mixed layer water well in advance of the storm where winds are a few meters per second, reducing SST as the storm approaches. However, strong oceanic baroclinic features advecting deep, warm oceanic mixed layers represent moving reservoirs of high-heat-content water available for the continued development and intensification phases of the tropical cyclone. Beyond a first-order description of the lower boundary providing the heat and moisture fluxes derived from low-level convergence, little is known about the complex boundary layer interactions between the two geophysical fluids.

One of the more apparent aspects of the atmospheric–oceanic interactions during tropical cyclone passage is the upper-ocean cooling as manifested in the SST (and mixed-layer temperature) decrease starting just in back of the eye. As seen in Figure 19, ocean mixed-layer temperature profiles acquired during the passage of several tropical cyclones revealed a crescent-shaped pattern of upper-ocean cooling and mixed-layer depth changes, which indicated a rightward bias in the mixed-layer temperature response with cooling by $1\text{--}5^{\circ}\text{C}$ extending from the right rear quadrant of the storm into the wake regime. These SST decreases are observed through satellite-derived SST images, such as that of the post-Hurricane Bonnie SST (Figure 20), which are indicative of mixed-layer depth changes due to stress-induced turbulent mixing in the front of the storm and shear-induced mixing between the layer and thermocline in the rear half of the storm. The mixed-layer cooling represents the thermodynamic and dynamic response to the strong wind that typically accounts for 75–85% of the ocean heat loss, compared with the 15–25% caused by surface latent and sensible heat fluxes from the ocean to the atmosphere. Thus, the upper ocean's heat content for tropical

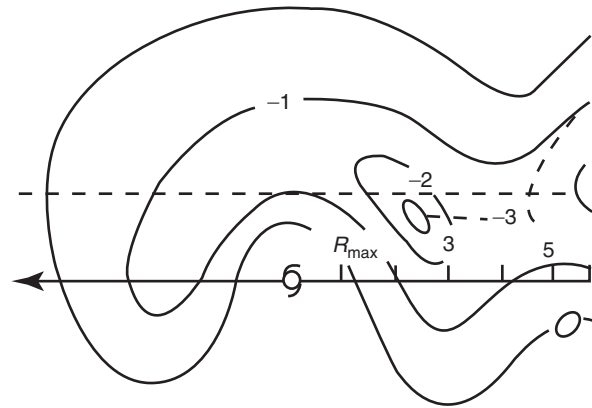


Figure 19 Schematic SST change ($^{\circ}\text{C}$) induced by a hurricane. The distance scale is indicated in multiples of the radius of maximum wind. Storm motion is to the left. The horizontal dashed line is at 1.5 times the radius of maximum wind. Reproduced with permission from Black PG, Elsberry RL, and Shay LK, (1988) Airborne surveys of ocean current and temperature perturbations induced by hurricanes. *Advances in Underwater Technology, Ocean Science and Offshore Engineering* 16: 51–58; © Society for Underwater Technology (Graham Trotman).

cyclones is governed not solely by SST; rather, it is the mixed-layer depths and temperatures that are significantly affected along the lower boundary by the basic state and transient currents.

Recent observational data have shown that the horizontal advection of temperature gradients by basic state currents in a warm core ring affected the mixed-layer heat and mass balance, suggesting the importance of these warm oceanic baroclinic features. In addition to enhanced air–sea fluxes, warm temperatures ($>26^{\circ}\text{C}$) may extend to 80–100 m in warm core rings, significantly impacting the mixed-layer heat and momentum balance. That is, strong current regimes ($1\text{--}2\text{ m s}^{-1}$) advecting deep, warm upper-ocean layers not only represent deep reservoirs of high-heat-content water with an upward heat flux, but transport heat from the tropics to the subtropical and polar regions as part of the annual cycle. Thus, the basic state of the mixed layer and the subsequent response represent an evolving three-dimensional process with surface fluxes, vertical shear across the entrainment zone, and horizontal advection. Simultaneous observations in both fluids are lacking over these baroclinic features prior, during, and subsequent to tropical cyclone passage, and are crucially needed to improve our understanding of the role of lower boundary in intensity and structural changes to intensity change.

In addition, wave height measurements and current profiles revealed the highest waves and largest fetches to the right side of the storm where the maximum mixed-layer changes occurred. Mean wave-induced currents were in the same direction as the steady mixed-layer currents, modulating vertical current shears and mixed-layer turbulence. These processes feed back to the atmospheric boundary layer by altering the surface roughness and hence the drag coefficient. However, little is known about the role of strong surface waves on the mixed-layer dynamics, and their feedback to the atmospheric boundary layer under tropical cyclone force winds by altering the drag coefficient.

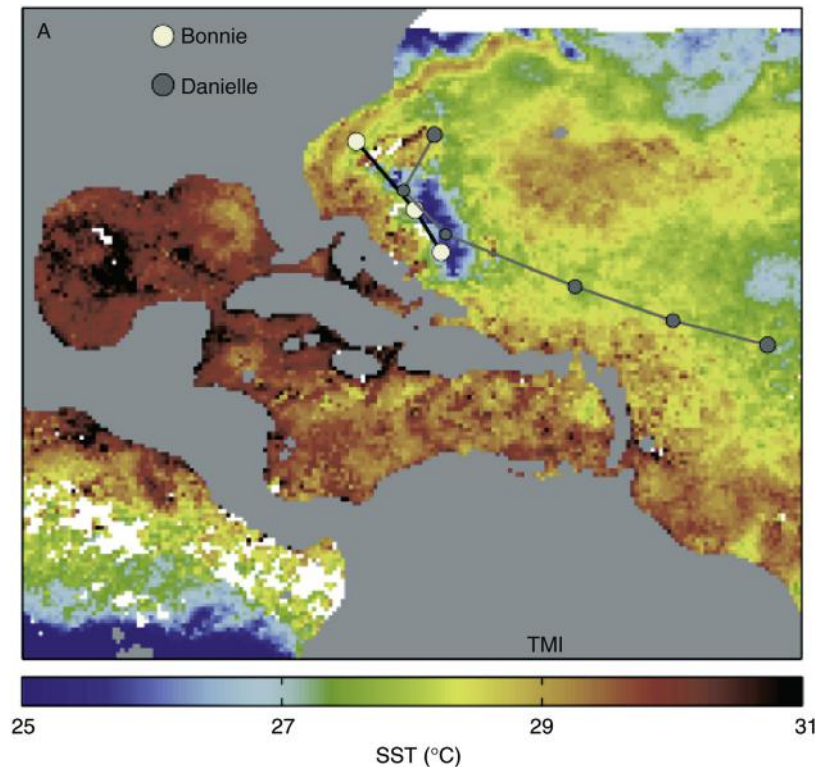


Figure 20 Cold wake produced by Hurricane Bonnie for 24–26 August 1998, as seen by the NASA TRMM satellite Microwave Imager (TMI). Small white patches are areas of persistent rain over the 3-day period. White dots show Hurricane Bonnie's daily position from 24 to 26 August. Gray dots show the later passage of Hurricane Danielle from 27 August to 1 September. Danielle crossed Bonnie's wake on 29 August and its intensity dropped. Reproduced with permission from Wentz FJ, Gentemann C, Smith D, and Chelton D (2000) Satellite measurements of sea surface temperature through clouds. *Science* 288: 847–850; © owned by the American Geophysical Union (<http://www.sciencemag.org>).

Tropical Cyclone Rainfall

Precipitation in tropical cyclones can be separated into either convective or stratiform regimes. Convective precipitation occurs primarily in the eyewall and rainbands, producing rains $>25 \text{ mm h}^{-1}$ over small areas. However, observations suggest that only 10% of the total rain area is comprised of these convective rain cores. The average core is 4 km in radius (area of 50 km^2), and has a relatively short lifetime, with only 10% lasting longer than 8 min (roughly the time a 1 mm diameter raindrop takes to fall from the mean height of the 0°C isotherm at terminal velocity). The short life cycle of the cores and the strong horizontal advection produce a well-mixed and less asymmetric precipitation pattern in time and space. Thus, over 24 h the inner core of a tropical cyclone as a whole produces 1–2 cm of precipitation over a relatively large area, and 10–20 cm in the core. After landfall, orographic forcing can anchor heavy precipitation to a local area for an extended time. Additionally, midlatitude interaction with a front or upper-level trough can enhance precipitation, producing a distortion of the typical azimuthally uniform precipitation distribution.

Energetics

Energetically, a tropical cyclone can be thought of, to a first approximation, as a heat engine; obtaining its heat input from the warm, humid air over the tropical ocean, and

releasing this heat through the condensation of water vapor into water droplets in deep thunderstorms of the eyewall and rainbands, then giving off a cold exhaust in the upper levels of the troposphere ($\sim 12 \text{ km}$ up). One can look at the energetics of a tropical cyclone in two ways: (1) the total amount of energy released by the condensation of water droplets or (2) the amount of kinetic energy generated to maintain the strong swirling winds of the hurricane. It turns out that the vast majority of the heat released in the condensation process is used to cause rising motions in the convection, and only a small portion drives the storm's horizontal winds.

Using the first approach, we assume an average tropical cyclone produces 1.5 cm day^{-1} of rain inside a circle of radius 665 km. Converting this to a volume of rain gives $2.1 \times 10^{16} \text{ cm}^3 \text{ day}^{-1}$ (1 cm^3 of rain weighs 1 g). The energy released through the latent heat of condensation to produce this amount of rain is $5.2 \times 10^{19} \text{ J day}^{-1}$ or $6.0 \times 10^{14} \text{ W}$, which is equivalent to 200 times the worldwide electrical generating capacity.

Under the second approach we assume that for a mature hurricane, the amount of kinetic energy generated is equal to that being dissipated due to friction. The dissipation rate per unit area is air density times the drag coefficient times the wind speed cubed. Assuming an average wind speed for the inner core of the hurricane of 40 m s^{-1} winds over a 60-km radius, the wind dissipation rate (wind generation rate) would be $1.5 \times 10^{12} \text{ W}$. This is

equivalent to about half the worldwide electrical generating capacity.

Either method suggests hurricanes generate an enormous amount of energy. However, they also imply that only about 2.5% of the energy released in a hurricane by latent heat released in clouds actually goes to maintaining the hurricane's spiraling winds.

Tropical Cyclone-Related Hazards

In the coastal zone, extensive damage and loss of life are caused by the storm surge (a rapid, local rise in sea level associated with storm landfall), heavy rains, strong winds, and tropical cyclone-spawned severe weather (e.g., tornadoes). The continental United States currently averages nearly \$5 billion (in 1998 dollars) annually in tropical-cyclone-caused damage, and this is increasing, owing to growing population and wealth in the vulnerable coastal zones.

Before 1970, large loss of life stemmed mostly from storm surges. The height of storm surges varies from 1 to 2 m in weak systems to more than 6 m in major hurricanes that strike coastlines with shallow water offshore. The storm surge associated with Hurricane Andrew (1992) reached a height of about 5 m, the highest level recorded in south-east Florida. Hurricane Hugo's (1989) surge reached a peak height of nearly 6 m about 20 miles north-east of Charleston, South Carolina, and exceeded 3 m over a length of nearly 180 km of coastline. In recent decades, large loss of life due to storm surges in the United States has become less frequent because of improved forecasts, fast and reliable communications, timely evacuations, a better-educated public, and a close working relationship between the National Hurricane Center (NHC), local weather forecast offices, emergency managers, and the media. Luck has also played a role, as there have been comparatively few landfalls of intense storms in populous regions in the last few decades. The rapid growth of coastal populations and the complexity of evacuation raises concerns that another large storm surge disaster might occur along the eastern or Gulf Coast shores of the United States.

In regions with effectively enforced building codes designed for hurricane conditions, wind damage is typically not so lethal as the storm surge, but it affects a much larger area and can lead to large economic loss. For instance, Hurricane Andrew's winds produced over \$25 billion in damage over southern Florida and Louisiana. Tornadoes, although they occur in many hurricanes that strike the United States, generally account for only a small part of the total storm damage.

While tropical cyclones are most hazardous in coastal regions, the weakening, moisture-laden circulation can produce extensive, damaging floods hundreds of miles inland long after the winds have subsided below hurricane strength. In recent decades, many more fatalities in North America have occurred from tropical-cyclone-induced inland flash flooding than from the combination of storm surge and wind. For example, although in 1972 the deaths from storm surge and wind along the Florida coast from Hurricane Agnes were minimal, inland flash flooding caused more than 100 deaths over the north-eastern United States. More recently, rains from Hurricane Mitch in 1998 killed at least 10 000 people in Central America, the majority after the storm had weakened to tropical storm strength. An essential difference in the threat

from flooding rains, compared with that from wind and surge, is that the rain amount is not tied to the strength of the storm's winds. Hence, any tropical disturbance, from depression to major hurricane, is a major rain threat.

Over each ocean basin that experiences tropical cyclones, the poleward movement of a tropical cyclone into the midlatitudes is normally associated with the decay stage of its life cycle. However, these systems can develop into fast-moving and rapidly developing extratropical cyclones that contain tropical cyclone-force winds. These transforming tropical cyclones may accelerate from forward speeds of 5 m s^{-1} in the tropics to $>20 \text{ m s}^{-1}$ in the midlatitudes. They pose a serious threat and forecast problem to maritime activities and shore locations over wide geographic regions that do not normally experience such conditions. A common problem in all these regions is the difficult challenge of predicting accurately the track, intensity, and impacts of these rapidly changing systems after advisories have been discontinued by the tropical cyclone forecast center. Forecasters responsible for producing warnings and advisories during extratropical transition are faced with the potential for large amounts of precipitation, continued high wind speed, and generation of large ocean waves and swell common in tropical cyclones. However, the increased translation speed decreases the warning time. Over land, the impacts are related to the intensity of surface winds and precipitation as in Hurricane Hazel in 1954, which resulted in a rapid intensification and precipitation amounts $>200 \text{ mm}$, leading to 83 deaths in the Toronto area of southern Ontario. Over the open ocean the increased storm motion combined with the continued high wind speeds produces extremely large surface waves, as in Hurricane Luis in 1995, which produced waves $>30 \text{ m}$, causing extensive damage to the luxury liner *Queen Elizabeth II*.

Tropical Cyclone Forecasting

When a tropical cyclone threatens there are four questions that must be answered: (1) where will it hit, (2) when will it hit, (3) how strong will it be, and (4) what type of threat should be expected (i.e., wind, storm surge, heavy rain, severe weather). The answers to these four questions are the goal of tropical cyclone forecasters. Track is the most important forecast, as it determines the answers to the first two questions. Operational tropical cyclone track forecasting is a semiobjective process that combines conventional, satellite, and reconnaissance observations with input from objective prediction models. At 12–24-h forecast intervals, persistence of storm motion is a major component of the forecast. However, an error in the initial motion of 1 m s^{-1} will yield an 84.6 km error in the 24 h forecast position. Since the tropical cyclone motion processes are complex and nonlinear, track uncertainty increases with time. For example, in the United States, where track errors are the lowest globally, the mean 24-h track error over the last 10 years is 170 km. However, 5% of the 24-h track errors over the last 10 years are $>370 \text{ km}$. To minimize the possibility that a coastal area may be struck without time to prepare, much larger areas are warned than will actually experience damaging winds. While specific track models have indicated up to 15% improvement over the past 2–3 years, the average length of coastline warned, 730 km (roughly a 4:1 ratio to the track error) has not decreased over the past decade. In fact, it has increased

over the 30-year mean of 556 km in response to the emergency manager's desire for longer lead time.

In the United States, emergency managers require communities with limited escape routes to complete preparation and evacuation before 17 m s⁻¹ winds arrive on the coast. Hence, the length of coastline warned is a combination of the forecast uncertainty in the track forecast and the uncertainty in the forecast of the radius of the 17 m s⁻¹ winds. The mean 24-h error of the forecast 17 m s⁻¹ wind radii is 71 km, which is about 30–35% of the actual radii, and represents a 4-h error in lead time for a typical storm motion of 5 m s⁻¹. However, 5% of the 17 m s⁻¹ wind radii forecasts exceed 255 km. The warning of 730 km of the coastline is justified as the sum of the 370 km uncertainty in the track forecast plus the 255 km uncertainty in the 17 m s⁻¹ wind radii forecast to be more than 95% confident that a coastal region will not be struck without warning. Such overwarning is costly! Our best estimates of average preparation costs for the warned coastline have increased roughly sixfold from \$50 million in 1989 to \$300 million in 1996 (\$410 000 km⁻¹ warned).

Another factor in the warning equation is storm intensity, which addresses the third question. Preparations differ considerably for a major hurricane or super typhoon than for a lesser storm (e.g., storm surge inundation varies greatly with storm intensity). In the United States the mean 24-h intensity forecast error over the last 10 years is 5 m s⁻¹ (half a category on the Saffir–Simpson scale; Table 1), and 5% of the intensity errors are >12.5 m s⁻¹ (over 1 category on the Saffir–Simpson scale). At 48 h, the mean intensity error is 8 m s⁻¹, and 5% are >20 m s⁻¹ (two categories on the Saffir–Simpson scale). Because the forecast intensity of the storm at landfall is a key factor in who evacuates, greater accuracy can lead to increased public safety and reduced costs, although these savings have yet to be quantified.

Understanding and predicting intensity change is more complex than that for track as it requires knowledge of interactions throughout the depth of the troposphere over a broad spectrum of scales. Observations are sparse in upper troposphere, atmospheric boundary layer, and upper ocean, limiting knowledge of environmental interactions, angular momentum imports, boundary layer stress, and air–sea interactions. Moreover, sea surface temperature remains an important but incomplete measure of the ocean's influence on tropical cyclone intensity change. Crucial unanswered questions concerning tropical cyclone intensity change lie in the relative impact and interactions of three major components: (1) the structure of the upper-ocean circulations that control the oceanic mixed-layer heat content, (2) the storm's inner core dynamics, and (3) the structure of the synoptic-scale upper-tropospheric environment. A successful intensity forecast requires knowledge of the mechanisms that modulate tropical cyclone intensity within the envelope defined by these three components.

Once the 17 m s⁻¹ wind radius crosses the coast the important forecast issues relate to determining how much damage is likely from wind, ensuing surge, rain, and severe weather, the answer to the fourth question. These factors will determine the type and level of emergency management response, and dictate where the most resources for recovery are needed. To a first order, the areas impacted by wind,

ensuing surge, rain, and severe weather are determined by the horizontal and vertical wind distribution in the boundary layer and the interaction of the wind field with coast and inland topography. Currently, only a little information is known about the wind structure and its distribution with height and radial distance near landfall because of the paucity of observations (research quality or otherwise). Hence, a concerted effort is needed to better understand what determines the distribution of wind with height and radial distance in a high-wind boundary layer. As with the intensity forecasts, greater accuracy in the specification of the three-dimensional wind structure at landfall should lead to increased public safety and reduced costs, although these savings have yet to be quantified.

Summary

'The eye of the storm' is a metaphor for calm within chaos. The core of a tropical cyclone, encompassing the eye and the inner 100–200 km of the cyclone's 1000–1500 km radial extent, is hardly tranquil. However, the rotational inertia of the swirling wind makes it a region of orderly, but intense, motion. It is dominated by a cyclonic primary circulation in balance with a nearly axisymmetric, warm-core, low-pressure anomaly. Superimposed on the primary circulation are weaker asymmetric motions and an axisymmetric secondary circulation. The asymmetries modulate precipitation and cloud into trailing spirals. Because of their semibalanced dynamics, the primary and secondary circulations are relatively simple and well understood. These dynamics are not valid in the upper troposphere, where the outflow is comparable to the swirling flow, nor do they apply to the asymmetric motions. Since the synoptic-scale environment appears to interact with the vortex core in the upper troposphere by means of the asymmetric motions, future research should emphasize this aspect of the tropical cyclone dynamics and their influence on the track and intensity of the storm.

Improved track forecasts, particularly the location and time when a tropical cyclone crosses the coast, are achievable with more accurate specification of the initial conditions of the large-scale environment and the tropical cyclone wind fields. Unfortunately, observations are sparse in the upper troposphere, atmospheric boundary layer, and upper ocean, limiting knowledge of environmental interactions, angular momentum imports, boundary layer stress, and air–sea interactions. In addition to the track, an accurate forecast of the storm intensity is needed because it is the primary determinant of localized wind damage, severe weather, storm surge, ocean wave runup, and even precipitation during landfall. A successful intensity forecast requires knowledge of the mechanisms that modulate tropical cyclone intensity through the relative impact and interactions of three major components: (1) the structure of the upper ocean circulations that control the mixed-layer heat content, (2) the storm's inner core dynamics, and (3) the structure of the synoptic-scale upper-tropospheric environment. Even if we could make a good forecast of the landfall position and intensity, our knowledge of how a tropical cyclone's structure changes as it makes landfall is in its infancy, because few hard data survive the harsh condition. To improve

forecasts, developments to improve our understanding through observations, theory, and modeling need to be advanced together.

See also: **Dynamical Meteorology:** Balanced Flow; Overview; Potential Vorticity. **Mesoscale Meteorology:** Convective Storms: Overview; Severe Storms. **Middle Atmosphere:** Quasi-Biennial Oscillation. **Synoptic Meteorology:** Cyclogenesis. **Tropical Meteorology and Climate:** El Niño and the Southern Oscillation: Observation; El Niño and the Southern Oscillation: Theory; Intertropical Convergence Zone.

Further Reading

- World Meteorological Organization Report No. TCP-38. In: Elsberry, R. (Ed.), 1995. Global Perspectives of Tropical Cyclones. WMO, Geneva.
- Emanuel, K.A., 1986. An air–sea interaction theory for tropical cyclones. 1. Steady-state maintenance. *Journal of the Atmospheric Sciences* 43, 585–604.
- Ooyama, K.V., 1982. Conceptual evolution of the theory and modeling of the tropical cyclone. *Journal of the Meteorological Society of Japan* 60, 369–380.

Tropical Cyclogenesis

Z Wang, University of Illinois at Urbana-Champaign, Urbana, IL, USA

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

Tropical cyclogenesis is the least understood phase of the tropical cyclone life cycle. This article provides an overview of the necessary large-scale conditions for tropical cyclone formation, the climatology and seasonality of tropical cyclone formation over different oceanic basins, as well as the theories and predictability of tropical cyclone formation.

Introduction

Tropical cyclones are the most severe storm systems in the tropics. Tropical cyclogenesis, or the formation of a tropical cyclone, is the least understood phase of the tropical cyclone life cycle and one of the great mysteries in tropical meteorology. The challenge can be attributed to the lack of *in situ* observations that capture the genesis processes over the remote open ocean, as well as the multiscale nature of tropical cyclogenesis, which involves interaction of processes from the convective scale to the planetary scale.

According to the National Hurricane Center, “a tropical cyclone is a rotating, organized system of clouds and thunderstorms that originates over tropical or subtropical waters and has a closed low-level circulation.” Based on the maximum sustained wind speed (i.e., the Saffir–Simpson scale), tropical cyclones are categorized into tropical depressions, tropical storms, and category 1 to category 5 hurricanes. A tropical depression is a relatively weak tropical cyclone with maximum sustained winds of 38 mph (33 knots) or less, and a tropical storm has maximum sustained winds of 39–73 mph (34–63 knots). In tropical cyclone forecasting, genesis is typically considered to have occurred with the formation of a tropical depression or a tropical storm. In agreement with the definition of a tropical cyclone, one of the operational criteria for genesis is the presence of a closed low-level circulation. This criterion, however, introduces some ambiguity when a vortex is embedded in a strong environmental flow, since a vortex may appear open in the Earth-relative frame of reference. In tropical cyclone research, a loose but more physically based definition is often used, in which genesis is defined as the formation of a self-sustaining system.

Necessary Conditions for Tropical Cyclogenesis

Tropical cyclone formation is sensitive to some large-scale conditions. The necessary, but not sufficient, conditions for tropical cyclone formation were identified more than four decades ago and are summarized below:

1. warm ocean waters throughout a sufficient depth (sea surface temperature (SST) > 26 °C to a depth of 60 m);
2. enhanced midtroposphere (~700 hPa) humidity;
3. conditional instability;
4. a preexisting low-level cyclonic disturbance;

5. weak vertical shear of the horizontal winds at the genesis site; and
6. 5° minimum latitudinal displacement from the equator.

The warm ocean is the energy source for tropical cyclone formation and intensification. A deep layer of warm water ensures sufficient thermal energy from the ocean. As shown in [Figure 1](#), nearly all tropical cyclones form in oceanic regions with SST greater than 26 °C. Recent studies suggest that this empirical SST threshold increases with the tropical mean SST. As reported in the IPCC WGI Fifth Assessment Report, the global frequency of tropical cyclones is likely to either decrease or remain essentially unchanged in the twenty-first century while the global mean tropical cyclone maximum wind speed and rain rates are likely to increase.

The majority of tropical cyclones form between 20° S and 20° N ([Figure 1](#)), but genesis is rare within 5° latitude from the equator due to the weak planetary vorticity. One of the outliers is Typhoon Vamei in 2001, which was formed at 1.4° N in the South China Sea. The development was facilitated by the interaction between cold surges originating from higher latitudes and a low-level cyclonic circulation, which provided enhanced background vorticity.

Strong vertical wind shear is detrimental for tropical cyclogenesis as it displaces the midlevel warm core from the surface circulation. Strong vertical shear can also advect dry air into the storm and weaken moist convection. It is generally believed that tropical cyclogenesis is extremely unlikely when the vertical shear between 200 and 850 hPa exceeds 15 m s⁻¹.

Although numerical model simulations suggest that spontaneous tropical cyclogenesis is possible if favorable environmental conditions are available for a sufficiently long time period, a preexisting low-level cyclonic disturbance is one of the necessary conditions for tropical cyclone formation in the present climate. A cyclonic precursor disturbance not only enhances the absolute vorticity, but also promotes vorticity aggregation and protects the incipient tropical cyclone vortex from a hostile environment, as discussed in [Section The Marsupial Paradigm](#).

The SST exceeds the empirical threshold over vast areas in the tropics and varies slowly in time, but other environmental conditions (e.g., midlevel moisture, low-level vorticity, vertical shear, and convective instability) can be modified significantly on the synoptic to subseasonal timescales, which contributes to variations of tropical cyclone activity.

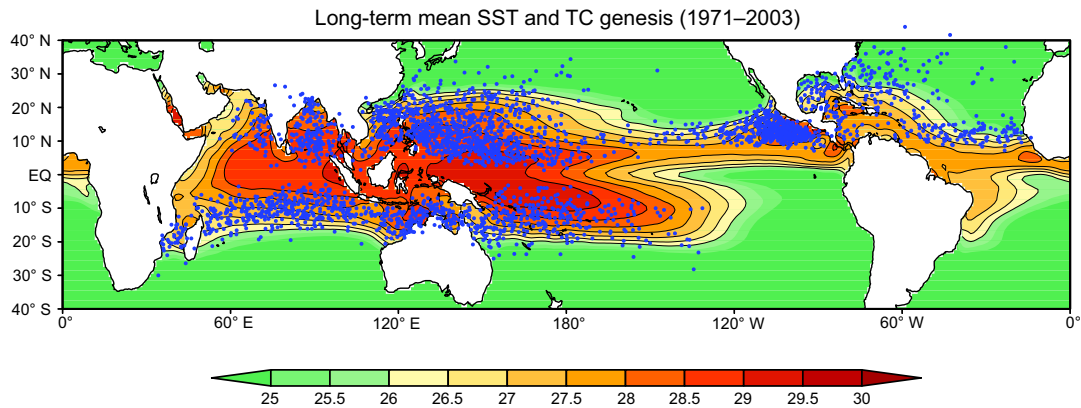


Figure 1 Long-term annual mean sea surface temperature (SST) and tropical cyclogenesis locations during 1971–2003. Note that although the long-term annual mean SST is less than 26 °C in some regions SST may exceed the threshold in the hurricane season.

Climatology of Tropical Cyclone Formation

The average annual number of global tropical cyclones with maximum sustained winds of 17 m s^{-1} or more is between 80 and 90, among which 40–50 intensify into hurricanes. Tropical cyclone frequency varies from basin to basin and from season to season. The climatology in different tropical basins is briefly described in this section. The statistics are derived from the International Best Track Archive for Climate Stewardship dataset (IBTrACS) for the time period 1974–2011. The numbers vary slightly depending on the time period examined or the dataset used.

Over the Atlantic, there are 12 tropical cyclones each year on average, which account for about 12% of the global storms. The Atlantic hurricane season runs from 1 June to 30 November, but tropical cyclones have been recorded in each calendar month except February and March. August and September are the peak season for tropical cyclone formation (**Figure 2(a)**), with about seven storms forming in these 2 months on average and 2–3 storms forming before August or after September. Tropical cyclogenesis location also has strong seasonal variations (see the genesis locations by 10-day period at <http://www.nhc.noaa.gov/climo>). Early in the hurricane season, tropical cyclones tend to form over the Gulf of Mexico and the Caribbean Sea. In the peak season, a large number of storms form over the Central and East Atlantic. These storms spend more time over the warm ocean and have a better chance to intensify into major hurricanes before making landfall or recurving poleward. In October and November, the genesis center shifts westward, and tropical cyclone formation is increasingly subject to extratropical impacts.

The East Pacific has 16 tropical cyclones per year on average. The hurricane season lasts from 15 May to 30 November, with a peak frequency in late August (**Figure 2(b)**). Most tropical cyclones form over a narrow band of warm SST and weak vertical wind shear north of the inter-tropical convergence zone (ITCZ) (**Figure 1**), a region of the highest genesis frequency per unit area over the globe. Tropical cyclones over the East Pacific may develop from tropical easterly waves originating over the Atlantic (or even West Africa) or from disturbances resulting

from the ITCZ breakdown. The latter sometimes leads to the simultaneous development of multiple cyclones.

The western North Pacific is the only basin where tropical cyclones form in all months of the year (**Figure 2(c)**). The genesis frequency peaks in August and September, with a minimum from January to April. On average, 26 tropical cyclones develop per year in this basin, which account for about 30% of the global storms. Tropical cyclogenesis in this basin is closely related to the monsoon circulation. Tropical cyclones frequently form in a monsoon trough or a monsoon confluence zone. Genesis may be associated with tropical easterly waves, equatorial Rossby waves, northwestward propagating tropical-depression disturbances, or the tropical upper tropospheric trough. Energy dispersion from a preexisting tropical cyclone may also spawn a new storm to its southeast. A giant typhoon occasionally develops near the center of a large-scale (radius of $\sim 1000 \text{ km}$) monsoon gyre.

There are only five tropical cyclones per year on average over the North Indian Ocean, with more storms forming over the Bay of Bengal than over the Arabian Sea (**Figure 1**). The seasonal cycle of the tropical cyclone formation has a bimodal distribution (**Figure 2(d)**), with a primary peak in November after the monsoon season and a secondary peak in May prior to the monsoon season. The reduced tropical cyclone (TC) activity during the monsoon season can be attributed to the strong vertical wind shear and the location of the monsoon trough. In the monsoon season, the monsoon trough shifts further northward and is located closer to land. Monsoon depressions, the primary precursor disturbances for tropical cyclone formation in this basin, form in the vicinity of the monsoon trough and tend to track northward over land, instead of moving westward over the warm ocean. Despite the low storm number, the deadliest storms have occurred in this basin. A tropical cyclone of moderate intensity may produce destructive storm surges in the Bay of Bengal due to the low and flat coastal plains and funnel-shaped coastlines. The 1970 Bhola cyclone killed more than 200 000 people in Bangladesh, and Cyclone Nargis in 2008 killed more than 100 000 in Myanmar.

In the Southern Hemisphere, tropical cyclones form over the South Indian Ocean and the western South Pacific. The

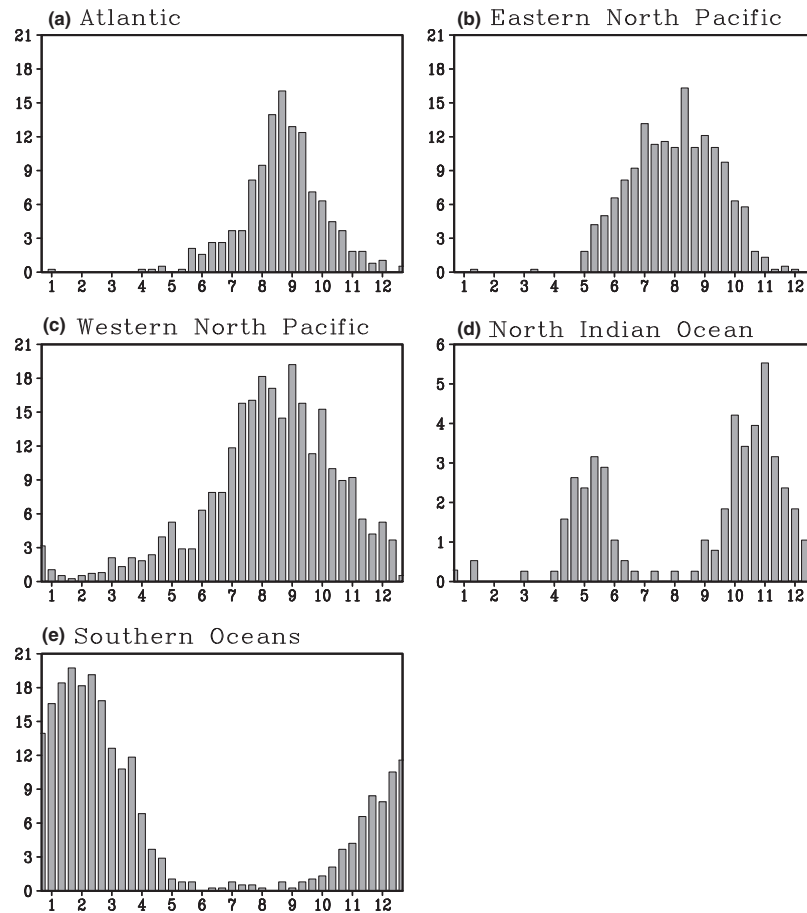


Figure 2 Seasonal cycle of 10-day mean tropical cyclone genesis frequency over different basins. The abscissa indicates the calendar months, and the ordinate is the number of storms per day multiplied by 100.

absence of tropical cyclones over the tropical South Pacific can be attributed to cold SSTs. Despite warm SSTs, tropical cyclones rarely develop over the South Atlantic because of the presence of the strong tropospheric (near surface to 200 mb) vertical wind shear and the absence of the ITCZ over the ocean. There are about 23 tropical cyclones per year on average over the southern Indian Ocean–Australia–the southwestern Pacific, and most of them form from November to April (**Figure 2(e)**). A large number of tropical cyclones develop near the northern coast of Australia (**Figure 1**), posing a great forecast challenge.

Variations of Tropical Cyclone Activity

Tropical cyclone frequency has strong variations on both the intraseasonal and longer timescales. On the intraseasonal timescale, tropical cyclone formation is strongly modulated by the Madden–Julian Oscillation (MJO). The MJO is the dominant mode of intraseasonal variability in the tropics. Its period varies between 30 and 60 days, and the local wavelength is roughly 12 000–20 000 km. The MJO is characterized by the eastward propagation of enhanced (and suppressed) convection in the tropics. The MJO affects tropical cyclone formation

through the modulation of the vertical wind shear, low-level vorticity, and midlevel moisture. As the active center of the MJO progresses eastward, the favored regions for tropical cyclone formation shift from the West Pacific to the East Pacific and eventually to the Atlantic (**Figure 3**). The hurricane genesis frequency in the Gulf of Mexico in the westerly (active) phase of the MJO can be four times as high as in the easterly (inactive) phase.

On the interannual timescale, tropical cyclone activity is strongly impacted by the El Niño–Southern Oscillation (ENSO). In the El Niño years (the positive phase of the ENSO), the ascending branch of the Walker circulation shifts eastward from the Maritime Continent–western Pacific sector to the Central Pacific due to SST warming over the Central and East Pacific. The change of the Walker circulation is accompanied by variations of the vertical wind shear and SST across the tropics. Cooler SST and stronger wind shear suppress tropical cyclone activity over the Atlantic and the western North Pacific, while warmer SST and weaker vertical shear contribute to enhanced tropical cyclone activity over the Central Indian Ocean, the East and Central Pacific. Conversely, tropical cyclone activity is enhanced over the Atlantic and the western North Pacific and reduced over the Central and East Pacific in the La Niña years.

Composite evolution of 200-hPa velocity potential anomalies ($10^6 \text{ m}^2 \text{ s}^{-1}$) and points of origin of tropical systems that developed into hurricanes/typhoons

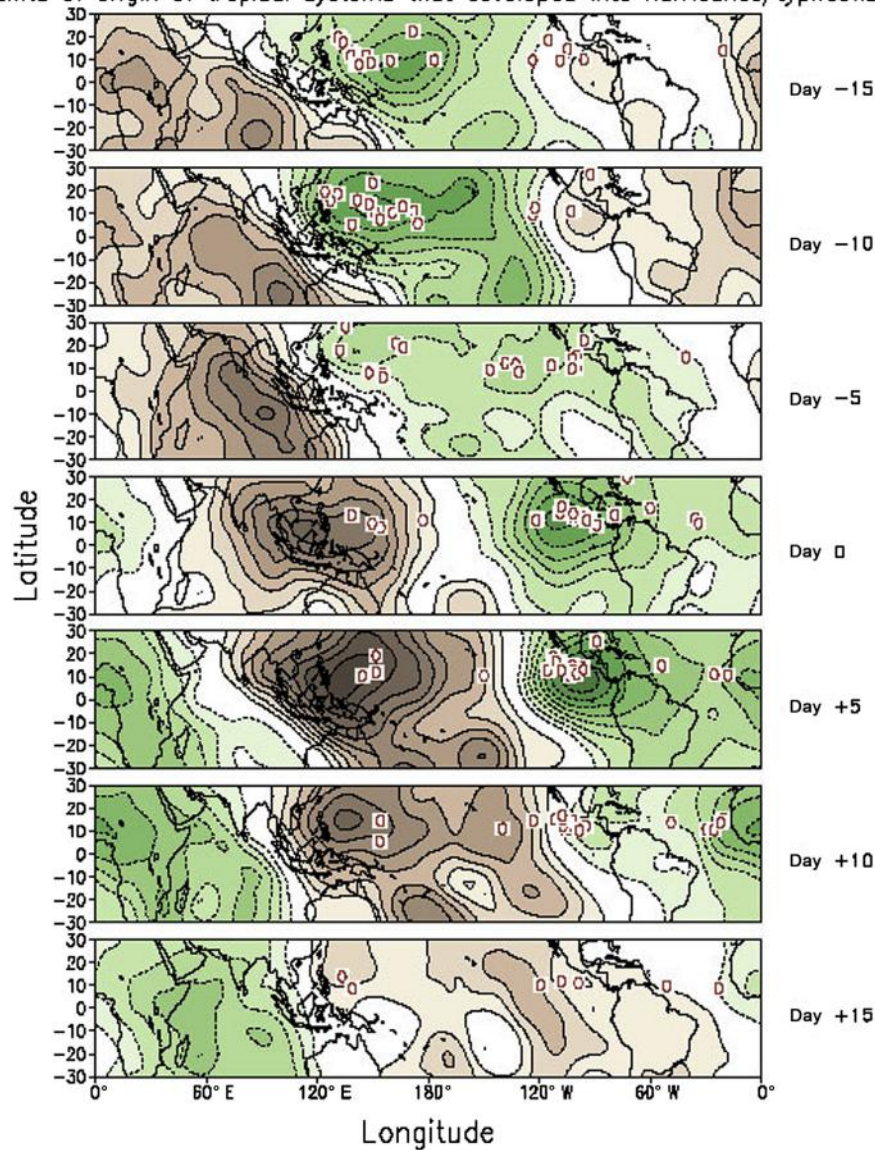


Figure 3 Tropical cyclone genesis locations (circles) in different phases of the Madden–Julian Oscillation (MJO). The MJO phases are identified by the 200-hPa velocity potential anomalies. Positive values (brown shading) indicate where convection is suppressed, and negative values (green shading) indicate where convection is enhanced. Figure from the National Weather Service, Climate Prediction Center: http://www.cpc.ncep.noaa.gov/products/intraseasonal/intraseasonal_fa.html.

Besides the ENSO, tropical cyclone activity is also affected by other climate factors. For example, the Atlantic Meridional Mode modulates the Atlantic ITCZ and strongly impacts the tropical cyclone genesis frequency and location as well as the storm track and intensity. On the decadal to multidecadal timescales, Atlantic tropical cyclone activity is modulated by the Atlantic Multidecadal Oscillation.

Theories for Tropical Cyclogenesis

It is instructive to consider tropical cyclone formation as a two-stage process. The first stage is the preconditioning of

the synoptic- and subsynoptic-scale environment, and the second stage is the construction and intensification of the tropical cyclone proto-vortex. The two stages distinguish processes at different spatial scales but may overlap in time. Tropical cyclones can develop in different synoptic-scale environments. Five genesis pathways have been identified based on the upper-level forcing and the low-level baroclinicity: nonbaroclinic, low-level baroclinic, trough-induced, weak tropical transition, and strong tropical transition. The first two pathways are not involved with significant upper-level forcing and account for about 70 and 9% of the global tropical cyclone developments, respectively. Over the Atlantic and the East Pacific, these two genesis pathways are generally associated

with tropical easterly waves. A recently proposed theory for tropical cyclone formation within tropical easterly waves is summarized in Section [The Marsupial Paradigm](#).

Despite the variety of synoptic-scale environments or precursor disturbances, it is generally believed that tropical cyclone formation proceeds through essentially the same mesoscale evolution processes at the second stage. Theories for the construction and intensification of the tropical cyclone proto-vortex are highlighted in Section [Development of the Tropical Cyclone Proto-Vortex](#).

The Marsupial Paradigm

Over the Atlantic, about 60% of tropical cyclones and 85% of intense hurricanes originate from tropical easterly waves. These waves have a typical wavelength of 2500–4000 km, and propagate westward at the speeds of 5–10 m s⁻¹. The waves generally peak at the same altitude as the African easterly jet (600–700 hPa over Africa and the East Atlantic) and have a cold-core structure in the lower troposphere. How an intense, warm-core tropical cyclone develops from such a synoptic-scale wave is an intriguing question. Recent studies show that a region of approximately closed Lagrangian circulation within the wave critical layer (i.e., the Kelvin cat's eye) plays an important role in tropical cyclone formation. The wave critical layer is the region surrounding the wave's critical latitude in a latitudinally sheared flow. In the enclosed Kelvin cat's eye, particles are trapped and recirculate rather than being swept one way or the other by the surrounding shear flow ([Figure 4](#)). The cat's eye is hypothesized to be important for tropical cyclone formation in three ways. First, that wave breaking or roll-up of the cyclonic vorticity and moisture near the critical latitude in the lower troposphere provides a favorable environment for vorticity aggregation leading to tropical cyclone formation. Second, as a region of quasiclosed Lagrangian circulation, the cat's eye protects moist convection inside from dry air intrusion to some degree. Third, there is positive feedback from the moist convective vortices to the synoptic-scale wave. The wave is maintained and possibly enhanced by diabatically amplified mesoscale vortices within the cat's eye. The combination of the three hypotheses is labeled the marsupial paradigm as the cyclogenesis sequence is likened to the development of a marsupial infant in its mother's pouch. The cat's eye within the wave critical layer is dubbed the 'wave's pouch' or simply 'pouch.'

Cloud-resolving numerical model simulations and observational analyses further reveal the dynamic and thermodynamic properties of the wave pouch. In the inner wave pouch region (a meso- β -scale area around the pouch center), the flow is characterized by strong rotation and weak shear/strain deformation. Moisture lofted by convection from the boundary layer can accumulate in this region and effectively moisten the column, while it is spread out or distorted into filaments away from the pouch center because of the strong strain rate. The wave pouch center is thus the preferred location for deep convection. The organized convection near the pouch center drives the transverse circulation and spins up the system-scale circulation. Numerical model simulations and *in situ* observational data show that the midlevel equivalent potential temperature becomes several degrees higher in the

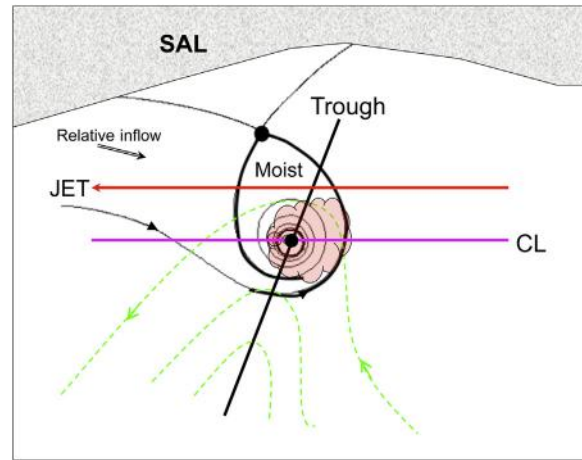


Figure 4 Formation of a tropical storm within a wave pouch. The dashed contours represent the streamlines in the ground-based frame of reference, which usually have an inverted-V pattern. The solid streamlines delineate the wave pouch as viewed in the frame of reference moving at the same speed with the wave. The pouch can protect the mesoscale vortices inside from the hostile environment, such as dry air associated with the Saharan air layer (SAL). Deep convection (gray shading) is sustained within the pouch. Owing to convergent flow, the pouch may have an opening that allows the influx of environmental air and vorticity. The easterly jet (JET), the critical latitude (CL), and the wave trough axis (Trough) are also indicated in the schematic. The intersection of the critical latitude and the trough axis pinpoints the pouch center as the preferred location for tropical cyclogenesis. Adapted figure from Wang, Z., Montgomery, M.T., Dunkerton, T.J., 2010. Genesis of Pre-Hurricane Felix 2007. Part I: The role of the easterly wave critical layer. *Journal of the Atmospheric Sciences* 67, 1711–1729. © American Meteorological Society. Used with permission.

inner pouch region than in the outer pouch region 1–2 days prior to genesis ([Figure 5](#)). The increase in the midlevel equivalent potential temperature in the inner pouch region is due to the accumulative effect of moist convection, and may serve as an indicator for organized convection and the impending tropical cyclogenesis.

The marsupial paradigm was originally proposed for tropical easterly waves. Subsequent studies show that it is also valid for tropical cyclone formation associated with other types of waves. Based on the marsupial paradigm, real-time forecast products have been developed to predict the genesis location using global model products. These products have provided useful guidance for flight planning in some recent tropical cyclone field experiments, including the Tropical Cyclone Structure field experiment in 2008 over the western North Pacific and the PREDICT field experiment in 2010 over the Atlantic.

Development of the Tropical Cyclone Proto-Vortex

There are two groups of theories regarding the development of the meso- β -scale tropical cyclone proto-vortex near the surface: the top-down development and the bottom-up development. The top-down theory suggests that the merger of multiple midlevel vortices in the stratiform precipitation

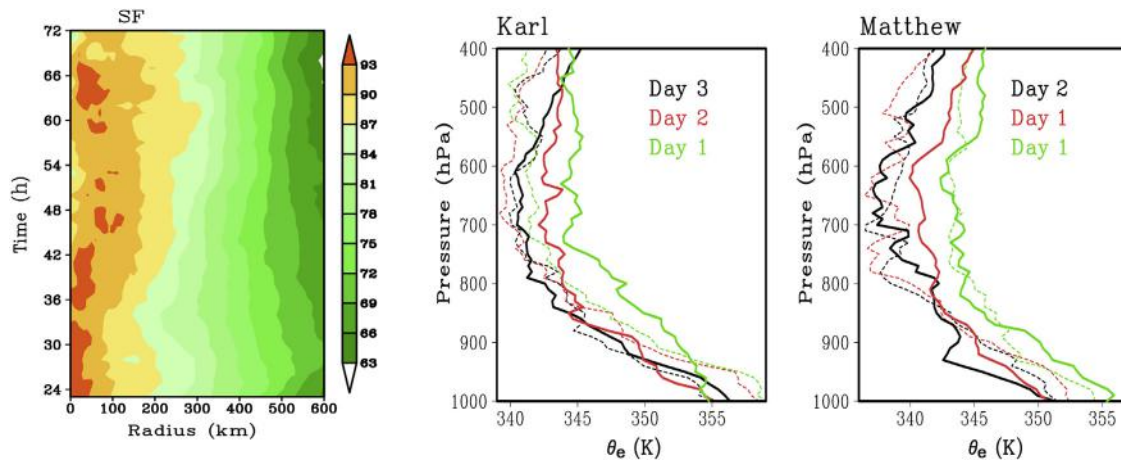


Figure 5 Left panel: Time–radius plots of saturation fraction (SF; %) derived from the numerical model simulation of Felix. Right two panels: vertical profiles of equivalent potential temperature (θ_e) for Karl and Matthew derived from the PREDICT dropsonde data. Solid lines and dashed lines represent the averages over the inner and outer pouch regions, respectively. Different colors represent different days prior to or after genesis. The plots show that θ_e increases up to 6 K prior to genesis in the inner pouch region but does not have any systematic increase in the outer pouch region. Adapted figure from Wang, Z., 2012. Thermodynamic aspects of tropical cyclone formation. *Journal of the Atmospheric Sciences* 69, 2433–2451. © American Meteorological Society. Used with permission.

region can create a vortex of a stronger intensity and a larger spatial scale. The vortex extends downward and engenders a cyclonic circulation near the surface, which is assumed to further intensify through the wind-induced surface heat exchange mechanism. It is also suggested that a midlevel vortex, with its associated cold core in the lower troposphere and warm core in the upper troposphere, modifies the thermodynamic environment and makes the vertical mass flux profile more bottom-heavy, which is favorable for the intensification of the low-level circulation.

The bottom-up theory emphasizes the role of vortical hot towers (VHTs), or rotating cumulonimbus clouds, in tropical cyclone formation (Figure 6). Although the lifetime of convective updrafts is of the order of 1 h, convective vorticity anomalies last much longer. Merger and axis-symmetrization of VHTs and their vortical remnants are hypothesized to lead to the formation of the proto-vortex. Furthermore, it is suggested that VHTs collectively act as a quasisteady heat source for driving the transverse circulation. The associated low-level convergence concentrates vorticity and intensifies the system-scale circulation.

Briefly speaking, the top-down development emphasizes the importance of stratiform processes and a midlevel vortex in initiating the surface cyclonic circulation, while the bottom-up development emphasizes the critical role of deep convection and the associated low-level convergence. While low-level convergence is the most effective way to intensify the surface circulation, it is worth noting that the bottom-up and top-down development routes may not be mutually exclusive. Different theories may be more relevant in different regions, and vorticity evolution at different spatial scales may take different vertical development routes. Besides, recent studies also suggest that other modes of moisture convection, besides VHTs, produce a wide range of vorticity anomalies and contribute to genesis. In

particular, cumulus congestus is proposed to play a dominant role in moistening the lower and middle troposphere and preconditioning the atmosphere for the transition to sustained deep convection and genesis.

Predictability of Tropical Cyclogenesis

A tropical cyclone is a meso- β -scale vortex, residing between the synoptic-scale waves and the meso- γ -scale convective processes. There is a downscale enstrophy cascade from the synoptic scale to the meso- β -scale and an upscale energy cascade from the convective scale (meso- γ -scale) to the meso- β -scale. Tropical cyclogenesis thus can be regarded as a combination of the large-scale downscaling and convective-scale upscaling processes. The large-scale environment to a large extent determines the track of the preferred genesis locations for storms originating from tropical waves, but moist convective processes limit the predictability of tropical cyclone formation, in particular, the genesis time. Predictability of tropical cyclogenesis can be evaluated using ensemble sensitivity experiments or adjoint modeling systems. Among various environmental parameters, tropical cyclone intensity forecasts are most sensitive to variations in midlevel moisture and low-level convective instability. Small errors in initial conditions can grow rapidly at the convective scale and then lead to strong divergence of the storm evolution in forecasts, similar to how moist convection limits the predictability of midlatitude winter storms. The predictability of tropical cyclone formation is also significantly influenced by vertical shear, and the forecast uncertainty increases with increasing vertical shear. The limited intrinsic predictability of tropical cyclone formation highlights the need to develop advanced ensemble prediction systems to provide probabilistic forecasts and risk assessment.

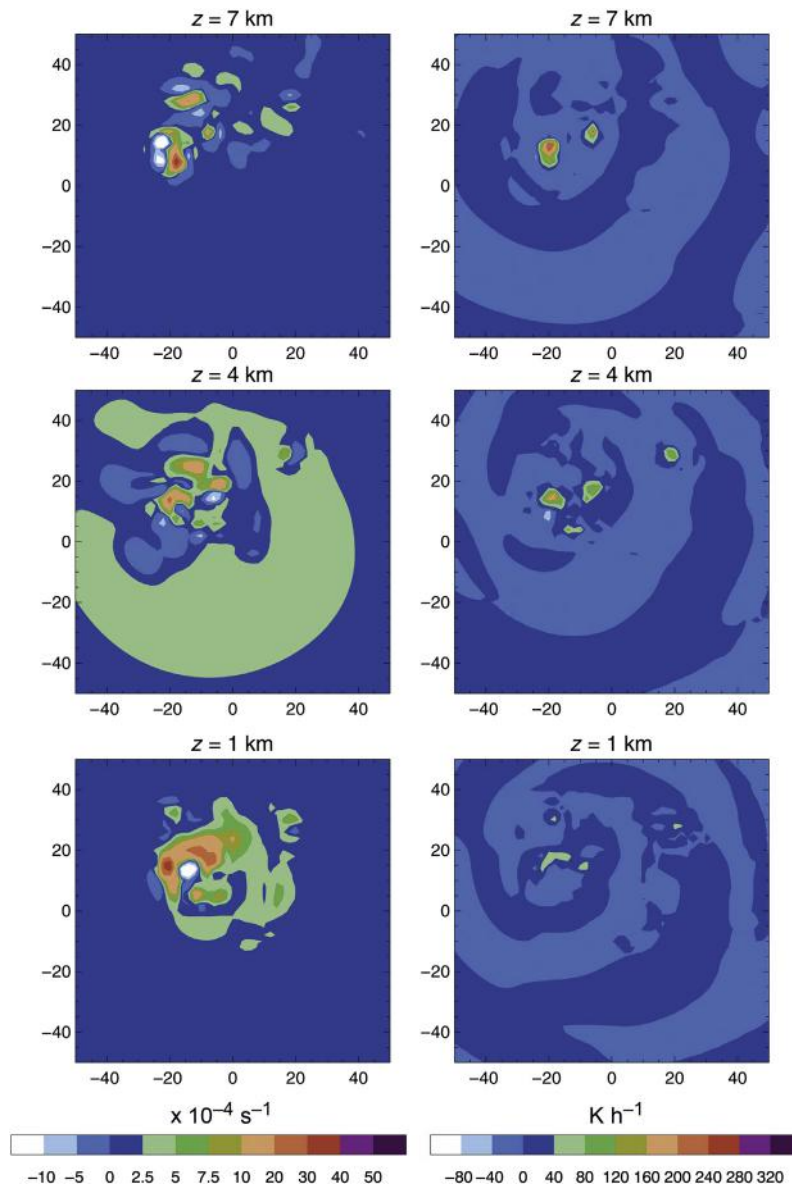


Figure 6 Absolute vorticity (units: 10^{-4} s^{-1}) (left) and diabatic heating rate (units: K h^{-1}) (right) on horizontal surfaces $z = 1, 4$, and 7 km in a numerical model simulation. Axes are in kilometers. Montgomery, M.T., Nicholls, M.E., Cram, T.A., Saunders, A.B., 2006. A vortical hot tower route to tropical cyclogenesis. *Journal of the Atmospheric Sciences* 63, 355–386. © American Meteorological Society. Used with permission.

See also: **Climate and Climate Change:** Global Impacts of the Madden–Julian Oscillation. **Tropical Cyclones and Hurricanes:** Tropical Cyclones and Climate Change; Tropical Cyclones in the Western North Pacific. **Tropical Meteorology and Climate:** Intertropical Convergence Zone; Monsoon: Overview.

Further Reading

- Davis, C., Bosart, L.F., 2004. The TT problem: forecasting the tropical transition of cyclones. *Bulletin of the American Meteorological Society* 85, 1657–1662.
- Dunkerton, T.J., Montgomery, M.T., Wang, Z., 2009. Tropical cyclogenesis in a tropical wave critical layer: easterly waves. *Atmos. Chem. Phys.* 9, 5587–5646.
- Fang, J., Zhang, F., 2011. Evolution of multiscale vortices in the development of Hurricane Dolly (2008). *Journal of the Atmospheric Sciences* 68, 103–122.
- Frank, W.M., Roundy, P.E., 2006. The role of tropical waves in tropical cyclogenesis. *Monthly Weather Review* 134, 2397–2417.
- Gray, W.M., 1968. Global view of the origin of tropical disturbances and storms. *Monthly Weather Review* 96, 662–700.
- Kossin, J.P., Vimont, D.J., 2007. A more general framework for understanding Atlantic hurricane variability and trends. *Bulletin of the American Meteorological Society* 88, 1767–1781.
- Maloney, E.D., Hartmann, D.L., 2000. Modulation of hurricane activity in the Gulf of Mexico by the Madden-Julian oscillation. *Science* 287, 2002–2004.
- McBride, J.L., 1995. Tropical cyclone formation. Global perspectives on tropical cyclones. In: Elsberry, R.L. (Ed.), WMO/TD-No. 693. World Meteorological Organization, Geneva, pp. 63–105.
- McTaggart-Cowan, R., Galarneau, T.J., Bosart, L.F., Moore, R.W., Martius, O., 2013. A global climatology of baroclinically influenced tropical cyclogenesis. *Monthly Weather Review* 141, 1963–1989.

- Montgomery, M.T., Nicholls, M.E., Cram, T.A., Saunders, A.B., 2006. A vortical hot tower route to tropical cyclogenesis. *Journal of the Atmospheric Sciences* 63, 355–386.
- Sippel, J.A., Zhang, F., 2008. A probabilistic analysis of the dynamics and predictability of tropical cyclogenesis. *Journal of the Atmospheric Sciences* 65, 3440–3459.
- Tory, K.J., Frank, W.M., 2010. Tropical cyclone formation. In: Chan, J., Kepert, J.D. (Eds.), *Global Perspectives on Tropical Cyclones*, second ed. World Scientific, pp. 55–92.
- Wang, Z., 2012. Thermodynamic aspects of tropical cyclone formation. *Journal of the Atmospheric Sciences* 69, 2433–2451.
- Wang, Z., 2014. Role of Cumulus Congestus in Tropical Cyclone Formation in a High-Resolution Numerical Model Simulation. *Journal of the Atmospheric Sciences* 71, 1681–1700.
- Wang, Z., Montgomery, M.T., Dunkerton, T.J., 2010. Genesis of Pre-Hurricane Felix (2007). Part I: The role of the easterly wave critical layer. *Journal of the Atmospheric Sciences* 67, 1711–1729.
- Zhang, F., Sippel, J.A., 2009. Effects of moist convection on hurricane predictability. *Journal of the Atmospheric Sciences* 66, 1944–1961.

Relevant Websites

- http://www.meted.ucar.edu/tropical/textbook_2nd_edition/ – Introduction to Tropical Meteorology 2nd Edition.
- <http://www.nhc.noaa.gov/climo/> – Tropical Cyclone Climatology.

Tropical Cyclones and Climate Change

TR Knutson, NOAA Geophysical Fluid Dynamics Laboratory, Princeton, NJ, USA

© Published by Elsevier Ltd.

Synopsis

It remains uncertain whether any past changes in tropical cyclone activity measures exceed the levels expected from natural variability alone. Using statistical and dynamical simulations, the late twenty-first century projections of tropical cyclone activity have been constructed. Although the confidence in these projections is limited, the general picture that emerges is a warmer climate with roughly 20% fewer tropical storms globally, about a 5% increase in the average intensity of tropical cyclones, and possibly more of the most intense (Category 4 and 5) systems. An increase in the precipitation rates near tropical cyclones, by about 20%, is also projected. Rising sea levels are expected to exacerbate surge flooding risk, although the net change in surge risk will depend on changes in landfalling tropical cyclone climate.

Introduction

The term tropical cyclone refers collectively to hurricanes, typhoons, tropical storms, and weaker tropical depressions, all of which are cyclones that form over tropical ocean regions. Here we will use the term more strictly to refer to hurricanes, typhoons, and tropical storms. An important question regarding these storms is whether they may change in character in response to climate change. Here we use the term climate change to refer to long-term (multidecadal to century scale) changes in climate caused by human activities such as burning of fossil fuels. Thus we are concerned with whether tropical cyclones have already changed due to the global warming that has already occurred since the late 1800s, and how tropical cyclones would behave if the tropical climate warms substantially ($\sim 2^\circ\text{C}$) during the twenty-first century as projected by the Intergovernmental Panel on Climate Change (IPCC) Fourth Assessment Report (IPCC, 2007) for some typical anthropogenic emission scenarios. The science of tropical cyclones and climate change is interdisciplinary, involving both climate science and hurricane science. As such, scientists from these disciplines collaborate to address this research problem. Using tools such as long observational records of tropical cyclone activity and related climate records, they have sought clues to the past behavior of tropical cyclones and an understanding of how this behavior is linked to other aspects of climate change – both natural and anthropogenic. Scientists have also used theory and models to better understand the relationship between large-scale climate and tropical cyclone activity. Based on these combined activities – observations, theory, and modeling – scientific assessments have emerged, which will be reviewed here. It is emphasized that the science of tropical cyclones and climate change continues to evolve as new or improved observational evidence, theories, or models become available that can further enhance our understanding of the tropical cyclone–climate change connection.

In this article, we first review the issue from a historical perspective, describing some initial findings that pointed toward a possibly high sensitivity of tropical cyclone activity to the past climate warming. This led to concern about a possible dramatic future increase in tropical cyclone activity in response to projected twenty-first century climate warming. We then

review subsequent developments in the science which found that, after accounting for observing system changes, the estimated past changes in tropical cyclone activity were far less dramatic than the initial analyses had reported. These new findings then implied considerably lower estimates of the sensitivity of tropical cyclones to climate warming. As of this article, it remains uncertain whether any trends or changes seen in the past records of tropical cyclones are outside the range of natural variability.

Why Is There a Concern about Increased Tropical Cyclone Activity in a Warmer Climate?

Concern about a possible influence of global warming on tropical cyclone activity arose from several research findings that suggested a possible strong sensitivity of tropical cyclone activity to climate warming. First, scientists had long ago recognized that tropical cyclones are generally formed and intensified in relatively warm tropical and subtropical oceanic regions. In the context of global warming, this led to a speculation that tropical cyclones would become more prevalent or more intense as the tropical oceans became warmer. Second, scatter plots of the maximum intensities of tropical cyclones vs sea surface temperature (SST) (Figure 1) showed that upper limit intensity of hurricanes tended to increase over relatively warm waters and that in fact the most intense hurricanes do not occur over cool sea surfaces – further reinforcing the view of a strong climate warming/hurricane intensity link. Theoretical models of hurricane intensity suggested that the maximum intensity of tropical cyclones would increase in a warming climate, though not as strongly as the apparent relationship in scatter plots such as Figure 1. Third, multiyear fluctuations of tropical cyclone power dissipation, an aggregate measure of hurricane activity, were shown to be well correlated in the Atlantic basin with tropical Atlantic SST, especially on decadal timescales and extending back as far as 1950 (Figure 2). Fourth, initial studies of Atlantic tropical storm frequency trends extending back to the late 1800s and early 1900s revealed a doubling in the annual frequency of these storms over the twentieth century, with similar increases in hurricanes since 1900 as well (e.g., Figure 3, blue curve). Analyses of tropical

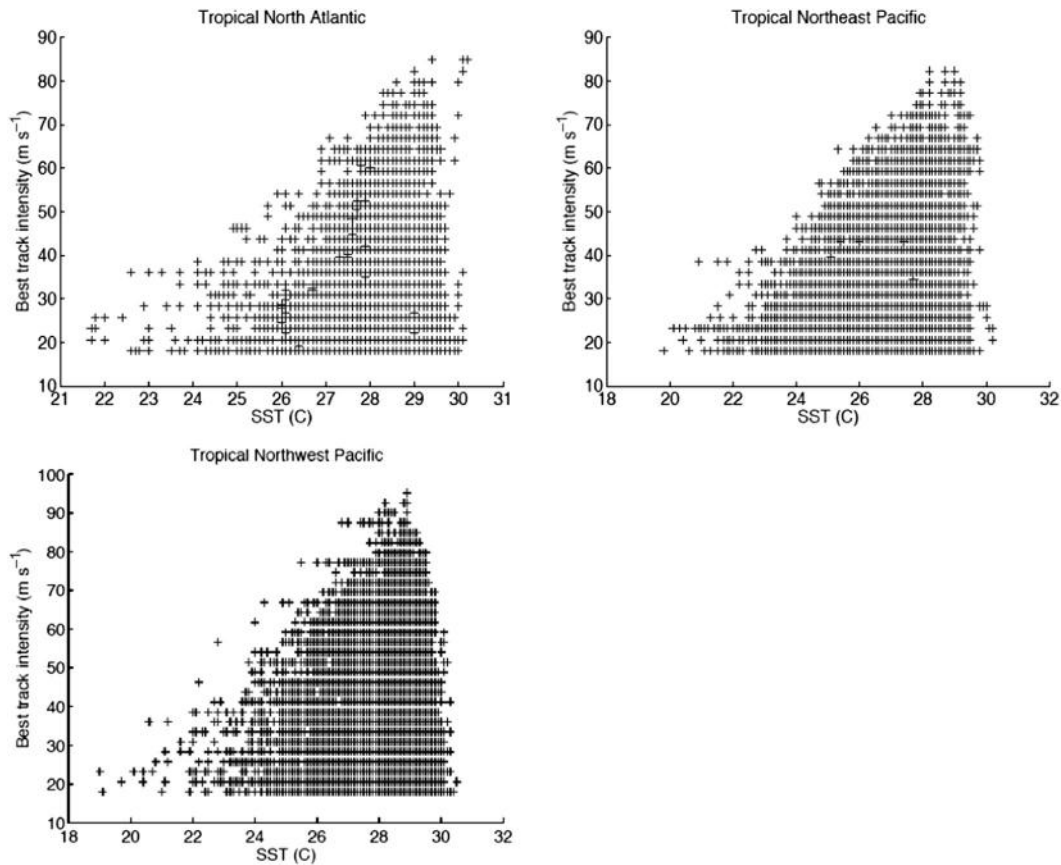


Figure 1 Scatter plot of observed tropical cyclone intensities vs SST in the environment of the storm. The most intense tropical cyclones occur only at relatively warm SSTs. Figure courtesy of Jim Kossin, University of Wisconsin/NOAA/National Climatic Data Center (NCDC).

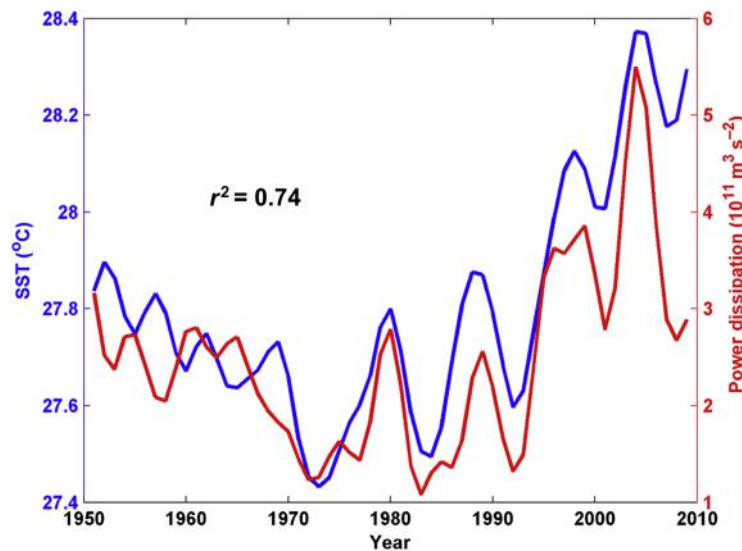


Figure 2 Low-pass filtered tropical Atlantic SST (blue) vs the power dissipation index (PDI) (red) for North Atlantic hurricanes based on data through 2011. SST is averaged over 6–18° N, 20–60° W, August–October. The data for 1949–69 are bias-adjusted to correct for a high bias in the North Atlantic hurricane database (HURDAT) winds. No adjustments are made to account for incomplete sampling of tropical cyclone frequency, duration or intensity, which may adjust the earlier part of the record toward higher values of PDI. Source: Updated from Emanuel, K., 2007. Environmental factors affecting tropical cyclone power dissipation. *Journal of Climate* 20, 5497–5509.

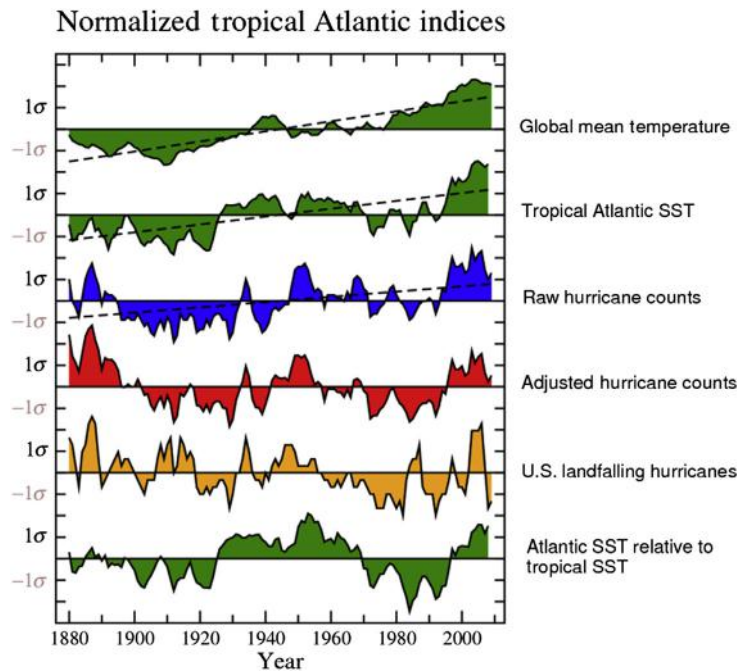


Figure 3 Five-year running means of tropical Atlantic indices. Green curves depict global annual-mean temperature anomalies (top) and August–October main development region (MDR) SST anomalies (second from top). Blue curve shows unadjusted Atlantic hurricane counts. Red curve shows adjusted Atlantic hurricane counts that include an estimate of ‘missed’ hurricanes in the presatellite era. Orange curve depicts annual US landfalling hurricane counts. Bottom green curve depicts August–October anomaly of MDR SST minus tropical mean SST. Hyphens along the vertical axis denote one standard deviation intervals (shown by the σ symbol). Dashed lines show linear trends. Only the top three curves have statistically significant trends. Source: Updated from Vecchi, G.A., Knutson, T.R., 2011. Estimating annual numbers of Atlantic hurricanes missing from the HURDAT database (1878–1965) using ship track density. *Journal of Climate* 24 (6). <http://dx.doi.org/10.1175/2010JCLI3810.1>.

Atlantic SST (the second curve from the top in Figure 3) found a likely detectable anthropogenic component, which, assuming that the close correlation between the tropical Atlantic SST and hurricane activity shown in Figure 2 was causal, implied that a detectable human influence on hurricane activity could be indirectly inferred. Finally, an analysis of ‘Best Track’ data sets of tropical cyclone activity in multiple basins around the globe extending back to the mid-1970s revealed a pronounced widespread increase over time in the number and proportion of the most intense tropical cyclones.

Adding to this concern were the very active 2004 and 2005 Atlantic hurricane seasons, with 2005 being a particularly active year for Atlantic hurricanes, featuring an extremely costly US landfalling event – Hurricane Katrina in the New Orleans/Mississippi coast vicinity. Based on this confluence of research findings and notable storm events, the climate change/hurricane community seemed on the brink of declaring that a pronounced or even dramatic anthropogenic increase in Atlantic hurricane activity, and possibly global tropical cyclone activity as well, was already underway. The existing counterarguments to this view at the time focused on the lack of a trend in US landfalling hurricane activity and on the much smaller sensitivity of tropical cyclones to climate change being simulated by dynamical models, or being inferred from the theories of hurricane intensity. In particular, the model-projected changes were much smaller than the various hurricane change projections being inferred from trends in the historical observations.

At this point, however, a new series of studies began to challenge the view of a pronounced detectable anthropogenic influence on twentieth century hurricane and tropical cyclone activity. The new studies began to elaborate issues with several of the previous findings, with implications for the question of a detectable anthropogenic influence on hurricanes. These new studies and the conclusions drawn from them are discussed in the following section.

Has Anthropogenic Climate Change Altered Tropical Cyclone Activity?

Detection of a climate change, as used here, refers to demonstrating that an observed change is highly unusual compared to the changes expected from natural processes alone, so that the change is likely to have been caused at least in part by anthropogenic forcing agents, such as increased greenhouse gases. An example of such a detectable climate change is the global mean temperature record since the late 1800s shown at the top curve in Figure 3. The increase in global mean temperature is broadly considered detectable because, among other findings, the increase is similar to that simulated in climate models that incorporate changes in greenhouse gas forcing and aerosols, while it cannot be simulated in models that do not include the warming influence of increased greenhouse gases.

Since credible century-long simulations of hurricane activity, including multicentury pre-industrial simulations, were

generally not available from current climate models, owing to their limited horizontal resolution, a simplified form of climate change detection with regard to Atlantic hurricanes was initially undertaken. The approach was to test whether a significant long-term trend, similar to that seen for global mean temperature or tropical Atlantic SST (top two curves) in [Figure 3](#), was also present for Atlantic hurricane frequency. While such an increase was indeed found for the hurricane counts (blue curve in [Figure 3](#)), independent research teams began to question the completeness of the observed Atlantic hurricane count record, particularly for hurricanes that remained at sea and never made landfall. Prior to the aircraft reconnaissance in the 1940s and especially prior to the satellite observing era in the mid-1960s, vast regions of the Atlantic were not systematically monitored for hurricane or tropical storm occurrence, and thus historical frequency estimates relied on essentially chance encounters between the storms and ships at sea. Although meteorologists have made a strong effort to identify missing storm cases based on available records, the question remained whether there was a dense enough network of ships available during the early part of the record to detect the non-landfalling storms that existed during that period. At least two independent groups concluded that there very likely was not a dense enough network of shipping traffic to adequately detect all the storms of the late 1800s and early 1900s. The red curve in [Figure 3](#) (fourth from the top) labeled ‘adjusted hurricane counts’ shows an estimate of what the actual history of Atlantic basin-wide hurricane counts would have been if the basin had adequate ship track density over the entire period so that no storms would have been missed. The adjusted curve has a negligible long-term trend. Interestingly, its time evolution bears a strong resemblance to that of the US landfalling hurricanes (orange curve in [Figure 3](#)), which suggests that the adjustments made to the basin-wide record are at least plausible. The two curves would be expected to be similar if the fraction of hurricanes that made landfall were unchanging over time; however, the assumption of constant landfalling fraction is not particularly compelling, necessitating the ship track density analyses. In summary, the long-term records of Atlantic hurricanes and tropical storm counts, after being assessed for potential problems with missing storms early in the record, do not provide compelling evidence of an anthropogenic influence in the form of a long-term rising trend.

The close correlation between multiyear fluctuations of Atlantic hurricane power dissipation and tropical Atlantic SSTs ([Figure 2](#)) is apparently not as influenced by ship track density issues as the century-long hurricane basin-wide frequency record, since the analysis in [Figure 2](#) begins around 1950, when aircraft reconnaissance was already available for monitoring hurricanes that were threatening landfall. A critical aspect of the relation shown in [Figure 2](#) is that it is a statistical correlation, which does not necessarily imply causation. In this case, independent investigators examining these data sets found that an alternative SST index to the local tropical Atlantic SST used in [Figure 2](#), was also strongly correlated to the power dissipation series. [Figure 4](#) shows a comparison of the statistical fit between historical hurricane activity and historical SST using two different statistical indices. The upper panel in [Figure 4](#) shows the relationship between hurricane activity and local tropical Atlantic SST, while the bottom panel shows the same, but using a ‘relative sea surface temperature’ index, defined as

the tropical Atlantic SST minus the tropical average SST. (An even longer record of relative SST fluctuations, which has no long-term rising trend, is shown at the bottom of [Figure 3](#).)

The findings illustrated in [Figure 4](#) had two important implications. First, they showed that projections of the late twenty-first century Atlantic hurricane power dissipation based on statistical models could vary from a slight increase to a dramatic 300% increase depending on whether the relative SST or the local tropical Atlantic SST was used as a predictor. Second, the analysis implied that a detectable human influence on past Atlantic hurricane could be indirectly inferred from the SST data, but only if one used the local SST as the statistical model of hurricane activity, and not if one used the relative SST as the statistical model. Clearly, the choice of statistical model had a dramatic impact on both future projections of hurricane activity and on the interpretation of past hurricane activity. The analysis in [Figure 4](#) also showed that if one used the dynamical model projections of the late twenty-first century hurricane activity (rather than statistical models) the results (depicted by the symbols at the right side of the diagrams in [Figure 4](#)) strongly favored a lower sensitivity scenario (lower panel) with only relatively small changes in power dissipation projected on average for the late twenty-first century – smaller in magnitude than the twentieth century multidecadal variations.

Another study had found – through an analysis of ‘Best Track’ data sets of tropical cyclone activity in multiple basins extending back to the mid-1970s – a pronounced increase over time in the number and proportion of the most intense tropical cyclones at the global scale, and over multiple independent basins. In response to this study, and to concerns expressed by other hurricane scientists of likely problems with the tropical cyclone historical data sets, a highly specialized satellite-based data set of tropical cyclone intensities was developed that encompassed all tropical cyclone basins and attempted to remove any artificial trends or jumps in the data series. The strategy used was to keep the tropical cyclone intensity estimation procedure as constant over time as possible through the use of carefully screened multi-basin satellite data. This involved, for example, degrading the quality of the satellite data in later years of the record so that they would be more compatible with data from the earlier years. This approach supported earlier findings of increasing storm intensities of strong storms in the Atlantic basin since the early 1980s, but revised downward the trends estimated from the ‘Best Track’ tropical cyclone data in several other basins. Using this method, a large increase in intensities of the strongest Atlantic hurricanes since 1983 was assessed to be statistically significant. This change has not been generally regarded as indicating a detectable anthropogenic change because the record length was relatively short, given the existence of pronounced multidecadal variability over the twentieth century in the basin. The notion that a trend since 1983 should not be taken as necessarily representative of the longer term trend behavior in the basin can be seen by inspection of the various century-scale Atlantic hurricane-related time series in [Figure 3](#).

In short, while initial studies had pointed to the likely emergence of an anthropogenic climate change signal in the Atlantic hurricane and global tropical cyclone data, subsequent studies identified substantial problems with the underlying data sets and proposed alternative statistical interpretations of

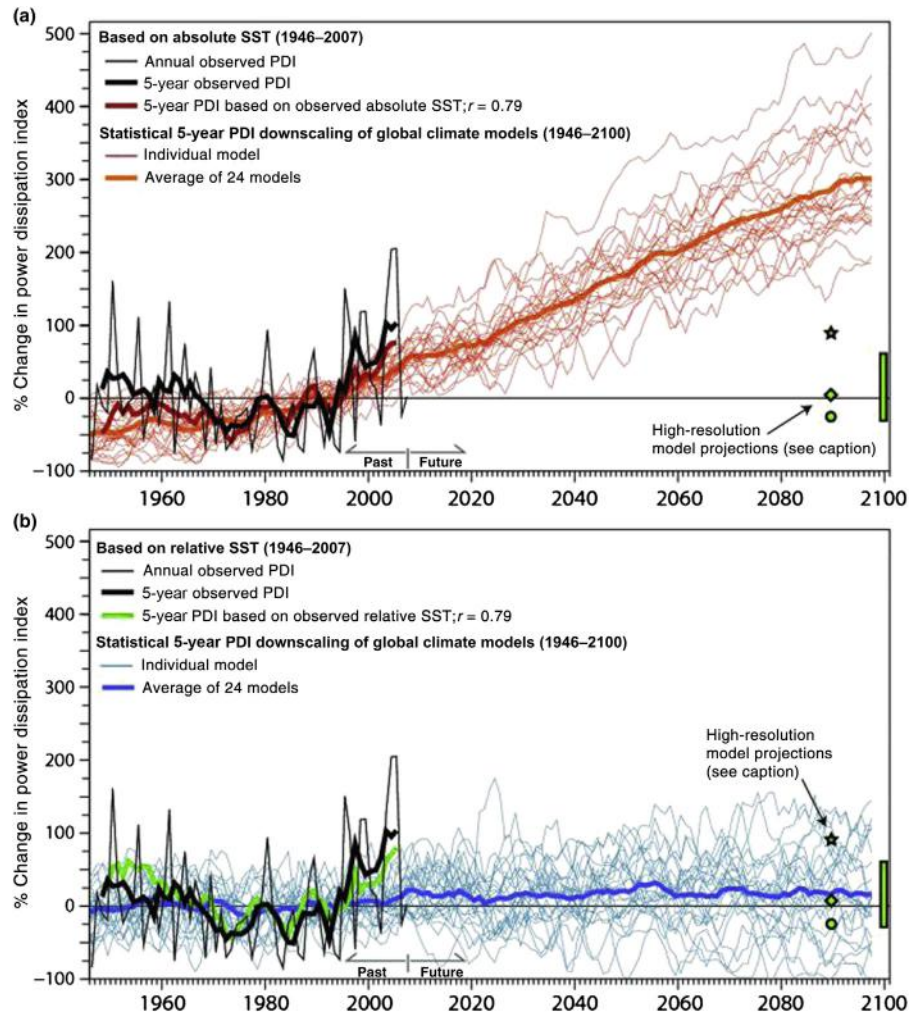


Figure 4 Past and extrapolated future changes in Atlantic hurricane PDI from statistical models. Anomalies are regressed onto (a) tropical Atlantic SST or (b) tropical Atlantic SST relative to tropical mean SST (1946–2007), and these regression models are then used to statistically estimate the PDI from several climate models. Anomalies are percent change relative to 1981–2000 average ($2.13 \times 10^{11} \text{ m}^3 \text{ s}^{-2}$). The green bar denotes approximate range of the PDI anomaly predicted by statistical–dynamical calculations. The green circle, star, and diamond denote approximate values suggested by several high-resolution dynamical models. SST region is $20\text{--}70^\circ \text{ W}$, $7.5\text{--}22.5^\circ \text{ N}$. Source: Vecchi, G.A., Swanson, K.L., Soden, B.J., 2008. Whither hurricane activity. *Science* 322 (5902). <http://dx.doi.org/10.1126/science.1164396>.

the data. These developments led several recent tropical cyclone–climate change assessments (see [Further Reading](#)), to conclude that it remained uncertain whether any changes in observed tropical cyclone metrics were outside the range of natural variability.

What Are the Near-Term Prospects for Identification of a Human Influence on Tropical Cyclones?

Although no detectable human influence on tropical cyclone activity has been identified, according to recent assessments, one can ask what the prospects are for such a detection in the near future. Here we discuss several tropical cyclone changes that have been reported that may be good candidates in the near term for identifying a detectable human influence on hurricanes.

As one possible candidate, a century-scale record of land-falling intense tropical cyclones in northeastern Australia has been found to have a statistically significant decreasing trend over time. This is a reasonable candidate for a climate change detection given both the length of record involved and the finding that the trend is significant by conventional statistical tests. One concern about a climate change detection claim in this case is the distinction between the observed changes (decreasing frequency trend) and trends that could occur due to natural processes alone. The Pacific basin is characterized by pronounced internal (natural) variability on interannual (2–7 years) timescales associated with El Niño/Southern Oscillation (ENSO) and there is also evidence of decadal to multidecadal scale internal variability, often referred to as the Interdecadal Pacific Oscillation. In order to make a more robust climate change detection, a long (multicentury) simulation of natural variability of tropical cyclone activity in the

region would be desirable, so that an analysis similar to that used for global mean temperature could be undertaken. Features of the model-simulated variability could be compared to observations on various timescales to provide further information on the capability of the model to simulate variability representative of that in the real world. New simulations of tropical cyclone activity with higher resolution coupled climate models are becoming available, and improving in quality, which may facilitate such an analysis.

A second candidate for possible near-term tropical cyclone climate change detection is the temporary lull in hurricane activity that occurred in the North Atlantic basin in the late twentieth century. This epoch, clearly apparent in [Figures 2 and 3](#), was marked by relatively low levels of hurricane activity (particularly for major hurricanes) during the 1970s and 1980s, compared with the higher levels of the 1950s and 1960s, and from 1995 through at least 2012. Two leading hypotheses to explain this prominent multidecadal variation focus on explaining the causes of the similar multidecadal variations in SST that apparently caused the changes in hurricane activity. The two hypothesized causes of the SST variations include natural multidecadal variability (often termed the Atlantic Multidecadal Oscillation (AMO)) and changes in anthropogenic climate forcing (primarily from aerosols). The aerosol forcing hypothesis concerns the strong increases in aerosol forcing from North American sources from the 1940s through the 1970s with a substantial reduction in this aerosol forcing through the 1980s as stronger policies to reduce air pollution came into force in North America. There is no consensus at this time regarding the relative contribution of these two processes to the observed SST or hurricane changes. For example, some models simulate a strong response in the Atlantic to the changes in aerosol forcing, while other studies suggest that the vertical structure of the ocean temperature changes during the multidecadal event seem inconsistent with a surface forcing such as aerosols and more consistent with an internal climate variation such as the AMO, as simulated in climate models.

Also in the Atlantic basin, one study has attempted to reconstruct the twentieth century US landfalling tropical cyclone activity using a network of tide gauges. Although the authors find a statistically significant increase in the frequency of large surge events, questions remain about precisely how their index relates to hurricane activity and whether the changes they find in the surge index are highly unusual compared with natural climate variability.

A complementary approach is to address the question of *when* a detectable human impact on hurricane activity might be identified. This approach uses model projections of future tropical cyclone changes, under certain assumed emission scenarios, to estimate a detection timescale, or likely time of detection. One study has estimated that even assuming a 10% increase per decade in Category 4 and 5 hurricane frequency in the Atlantic basin, a time period of about six decades would still be required for the signal to be detectable above the background of natural variability. This long timescale for detection is related to the apparently large levels of natural variability of tropical cyclones and related metrics in the Atlantic basin; if most of the observed multidecadal variability turned out to be forced by anthropogenic activities, rather than due to natural variability, the expected detection timescale would be shorter.

Model Simulations of Present-Day Tropical Cyclone Activity

To explore how tropical cyclone activity may respond to future climate change will require the use of models to simulate the tropical cyclone activity. It is therefore necessary to first assess how well models can simulate the present-day climate of tropical cyclones. An example of a simulation of tropical cyclone genesis and tracks using an atmospheric model with prescribed SSTs is shown in [Figure 5](#), which compares observed tracks for the period 1981–2005 with the model-simulated tracks for those same years. The model generates its own storms based on the interactions within the model. These do not correspond to individual storms in the real world but have a similar distribution in a statistical sense. As another measure of model skill, [Figure 6](#) shows how the annual number of Atlantic hurricanes generated by the model in [Figure 5](#) varies from year to year compared with the observed record of hurricanes. The model is able to simulate the year-to-year variability of hurricane counts remarkably well using the observed SSTs as input. Several other atmospheric models show a roughly similar skill for this type of simulation test. The main point is that current models can translate large-scale information, such as the global SST distribution, into fairly realistic information about aggregate Atlantic hurricane activity. The simulations of year-to-year tropical cyclone variability in other basins are typically not as impressive as for the Atlantic basin, but models show some skill in several other basins. These results suggest that the current models in principle can be used to simulate changes in tropical cyclone activity in response to climate change such as global warming; however, the models rely on the SST patterns used as boundary conditions to be realistic, and these must be obtained elsewhere.

An Introduction to Tropical Cyclone Activity Projections for the Late Twenty-First Century

In the previous sections, we considered the issue of detecting an anthropogenic climate change influence on tropical cyclone activity, using past observations together with various statistical or dynamical models. Here we consider future (late twenty-first century) anthropogenic climate change, in which the global warming is expected to be considerably larger in magnitude than the changes experienced over the twentieth century, and ask what impact such changes could have on tropical cyclone activity.

The general approach is to use the projected climate changes for the late twenty-first century, as projected by various climate models (such as those used for the Coupled Model Intercomparison Project 5 (CMIP5)) and either track the changes in tropical cyclones simulated directly in those climate models, or use higher resolution atmospheric models or statistical models to ‘downscale’ the tropical cyclone activity based on large-scale fields obtained from the global climate models. (‘Downscaling’ here refers to the use of a statistical model or dynamical model to try to simulate the behavior of certain smaller scale weather or climate phenomena, such as hurricanes, that are poorly resolved in the global climate models due to the large spacing between grid points in those models. For example, the SSTs from a global climate model may be used as boundary forcing

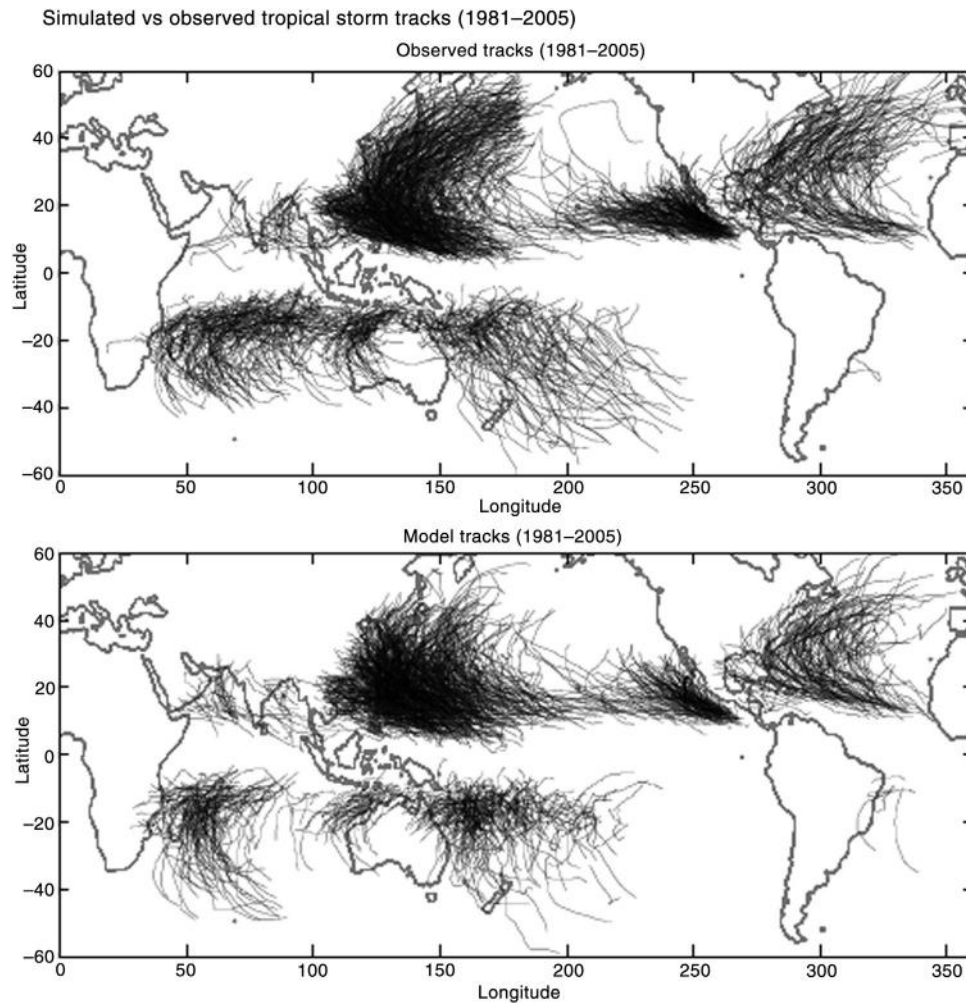


Figure 5 Comparison of tracks for tropical cyclones that reached hurricane force for 1981–2005 from observations (top) and from a 50-km grid global atmospheric model running over observed SSTs for the same time period (bottom). From Zhao, M., Held, I., Lin, S.-J., Vecchi, G.A., 2009. Simulations of global hurricane climatology, interannual variability, and response to global warming using a 50 km resolution GCM. *Journal of Climate* 22, 6653–6678.

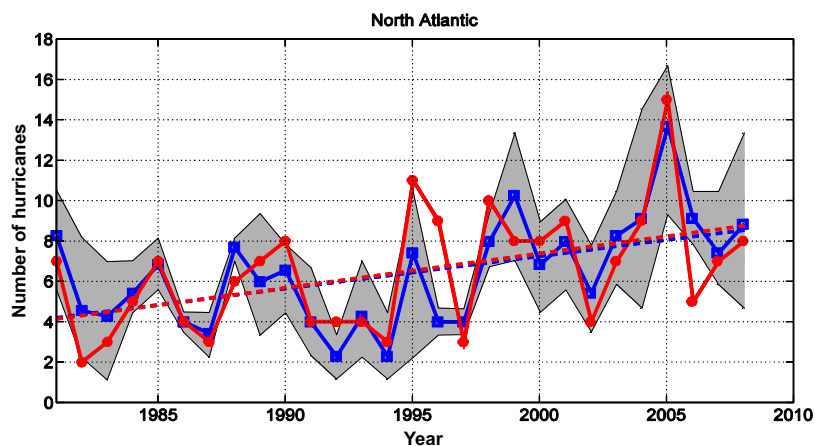


Figure 6 Annual number of observed (red) and simulated (blue) hurricanes in the Atlantic basin (1981–2008). The simulation uses an atmospheric model running over observed SSTs. The gray shading is the range of results across a five-member ensemble, with the blue line being the ensemble mean of these members. Dashed lines are least square linear trend lines. Updated from Zhao, M., Held, I., Lin, S.-J., Vecchi, G.A., 2009. Simulations of global hurricane climatology, interannual variability, and response to global warming using a 50 km resolution GCM. *Journal of Climate* 22, 6653–6678.

for an atmosphere-only model that has higher spatial resolution for resolving hurricane-like features more realistically.)

Before discussing the model projections, a brief discussion of confidence levels for these projections is needed. As discussed in the previous sections, there is as yet no clear detectable anthropogenic influence on tropical cyclone activity. This does not necessarily mean that no such detectable change is expected in the coming century, or that anthropogenic forcing has not yet affected tropical cyclones at all. First, the global mean temperature change over the twenty-first century in the IPCC Fourth Assessment Report (AR4) future emission scenarios is projected to be much larger than the temperature change over the twentieth century ($\sim 0.6^\circ\text{C}$), with a range of 'best estimate' global warming of about $1.8\text{--}4.0^\circ\text{C}$ depending on the IPCC Special Report on Emission Scenarios (SRESs) used (ranging from an aggressive emission reduction scenario B1 to fossil fuel-intensive scenario A1FI). Second, the observing system for tropical cyclones over the next century will undoubtedly be better than that over the twentieth century as a whole, making climate change detection studies more reliable. Third, the longer records available with additional decades of data make climate change detection more viable. However, given that no clearly detectable anthropogenic change in tropical cyclones has been identified yet, the confidence levels for projections of changes in tropical cyclone activity over the twenty-first century must necessarily be substantially lower than those for certain other climate variables (e.g., global mean surface temperature) that have strong evidence for a detectable anthropogenic signal over at least the past half century. The term 'likely' here is used to mean that the probability that the projected twenty-first century change will occur (assuming a twenty-first century emission scenario similar to IPCC SRES A1B – roughly a 'middle of the road' scenario) is subjectively assessed as about two chances in three or greater (but not 90% confidence, for example, which would be characterized as 'very likely').

Tropical Cyclone Frequency Projections

More than two dozen different studies have been done with various types of dynamical models of the atmosphere or atmosphere–ocean system aimed at simulating tropical cyclone frequency under warmer climate conditions. A common feature across 10 of these studies that explored global tropical cyclone activity is a decrease, ranging from a few percent to about one-third, in the total global annual count of these storms. The decrease is not seen in all individual basins in these models and is more pronounced in the Southern Hemisphere tropical cyclone basins than in the Northern Hemisphere.

The reasons for the decrease in global tropical cyclone frequency in the models are not well understood, partly because a full generalized theory of tropical cyclone genesis has not been developed. For example, although we know that about 90 tropical cyclones form each year in the current climate, we do not yet have a theory for why this number is 90 and not half or twice this number. Since atmospheric models can be made to produce a wide range in annual number of tropical cyclones through adjusting certain model parameters, they cannot offer an explanation directly. (Any such model

adjustment parameters are held fixed for climate change perturbation experiments such as those described above.) These fundamental questions remain a challenge for the tropical cyclone and climate research communities, although clearly progress has been made in our ability to model aspects of tropical cyclone climate, and even toward providing improved seasonal dynamical predictions of tropical cyclone activity, especially in the Atlantic basin.

Based on consistent projections from models, but tempered by the lack of a detectable decrease in the past data and by absence of a clear understanding of the physical mechanism producing the simulated decrease in the models, the projection of a global decrease in tropical cyclone frequency has been assessed as 'likely.'

An example of four maps of tropical cyclone frequency change projections for the late twenty-first century is shown in [Figure 7](#). The projections are from a single atmospheric model, but use projected SST changes from several different climate models. The fact that the detailed regional tropical cyclone frequency changes differ so much across the different projections suggests the following implication: even if climate scientists had a perfect model for translating future SST projections into tropical cyclone frequency projections, they would still have difficulty making confident tropical cyclone projections at the regional scale because different climate models disagree on important details of SST projections. That is, even though climate models all agree that global mean temperature will increase over the twenty-first century under the IPCC SRES scenarios, the detailed patterns in those warming projections differ between the models, and those differences apparently have important consequences for regional tropical cyclone activity.

These results suggest a long-term path forward toward more confident future projections of regional tropical cyclone activity: that the climate change research community eventually reduces the uncertainty of regional surface temperature projections, at least for a given emission scenario. Meanwhile, researchers can at least identify the more robust subset of projection signals for certain geographical domains (e.g., global, Southern Hemisphere) based on the existing models.

Tropical Cyclone Intensity Projections

In contrast to the decrease in global tropical cyclone frequency projected under climate warming by current models, the average intensity of tropical cyclones and the intensities of the strongest storms are projected to increase over the twenty-first century by roughly 5% in terms of maximum surface wind speed.

The maximum intensity of tropical cyclones has a more well-established theoretical basis and the ability to simulate tropical cyclone intensities with models has improved over time. Although the short-term (1–5 days) prediction of tropical cyclone intensity for individual storms remains a challenge and is much more difficult than storm track prediction, current theories and models at least produce credible climatological statistics of tropical cyclone intensities for the present-day climate, encouraging their use for climate change projections.

An early study in 1987 suggested that the maximum intensity of tropical cyclones could be successfully reproduced

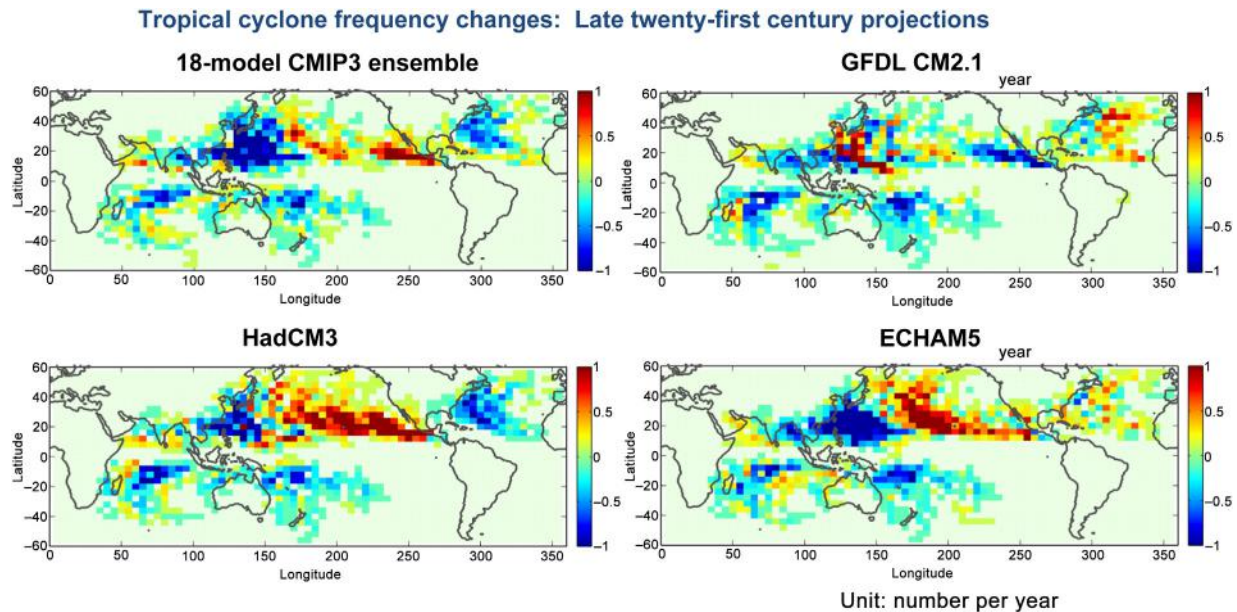


Figure 7 Model-projected changes in tropical cyclone occurrence frequency for the late twenty-first century obtained using three different climate model projections of SST changes, or using an 18-model average change. Red colors indicate increases and blue colors decreases. Adapted from Zhao, M., Held, I., Lin, S.-J., Vecchi, G.A., 2009. Simulations of global hurricane climatology, interannual variability, and response to global warming using a 50 km resolution GCM. *Journal of Climate* 22, 6653–6678.

from a theory linking the intensity to the SST and vertical thermodynamic structure of the large-scale environment around the storm. The theory, developed by Kerry Emanuel at MIT, proposed that the environmental changes for CO₂ warming simulated by climate models implied that tropical cyclones would become more intense in a warmer climate. Dynamical modelers later found a similar sensitivity of tropical cyclone intensity to greenhouse warming using hurricane simulation models. Since then, a growing number of studies employing increasingly realistic higher resolution models (grid spacing from about 50 km down to about 10 km or less) have found a roughly similar sensitivity, at least globally. The range of values for global mean changes in these studies varies from a few percent to about 10%. The increased intensity is often not found using relatively coarse resolution models, which may not be too surprising, since these models only very crudely resolve a hurricane-like structure and often do not simulate such distinct features of strong tropical cyclones as the ‘eye’ at the center of hurricanes and typhoons. While a global mean increase of tropical cyclone intensity with climate warming is assessed as ‘likely,’ such increases are not projected to occur in all regions due to various regional patterns in projected surface warming and other factors.

Projected Frequency of Very Intense Tropical Cyclones

An increase of average tropical cyclone intensity would generally be expected to lead to an increase in the number of very intense storms. However, since model projections also indicate a likely decrease in the overall number of tropical cyclones,

there remains the question of whether the number of intense storms will increase or not. Furthermore, very intense (Category 4 and 5) tropical cyclones are typically not produced by atmospheric models unless the resolution is quite high (grid spacing of about 10–20 km or less, depending on the model). At least two studies using such models are currently available – one for the Atlantic basin using a regional hurricane model coupled to an ocean model, and the other using a global atmospheric model but with no interactive ocean model component. These studies both indicate that despite the reduced overall frequency of tropical cyclones in either the Atlantic basin or global domain they studied, the frequency of the most intense tropical cyclones tended to increase (e.g., Figure 8). However, there are only relatively few studies to date with sufficient resolution to simulate the frequency of very intense storms, and the statistical robustness of the projected changes even in these available experiments is marginal. Considering such factors, a recent assessment concluded that there were greater than even odds (>50% chance) of an increase in the frequency of very intense tropical cyclones over the twenty-first century.

While the confidence in such a projection is limited, the importance of such a future change could be considerable owing to the fact that very intense storms have caused a disproportionate fraction of tropical cyclone damage, according to historical analyses. It is also worth noting that such historical damage-based analyses, when adjusted for changes in demographics such as wealth and population, do not show strong evidence for a climate-change-related increase in tropical cyclone activity over time. Instead they indicate a prominent role for demographic changes in explaining historical increases in tropical cyclone damage observed over the twentieth century.

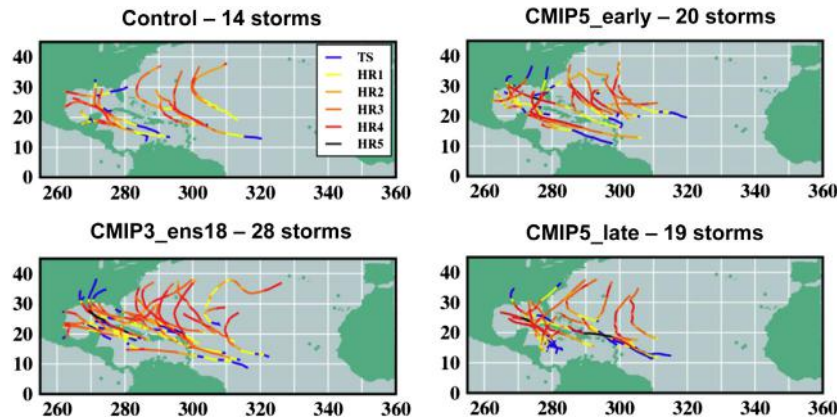


Figure 8 Tracks and intensities of all storms reaching Category 4 or 5 intensity (maximum surface winds $\geq 59 \text{ m s}^{-1}$) in a series of Geophysical Fluid Dynamics Laboratory (GFDL) hurricane model downscaling experiments (27 seasons). Results are shown for the control climate (upper left) and for downscaling experiments using the CMIP3/A1B 18-model ensemble climate change for the late twenty-first century (bottom left); CMIP5/Representative Concentration Pathways (RCP4.5) models early twenty-first century ensemble (upper right); and CMIP5/RCP4.5 late twenty-first century ensemble (lower right). The track colors (legend) depict the Saffir–Simpson category of the storms along their trajectories. CMIP3/A1B and CMIP5/RCP4.5 are two commonly used, representative future emission scenarios. From Knutson, T.R., Sirutis, J.J., Vecchi, G.A., Garner, S.T., Zhao, M., Kim, H.-S., Bender, M.A., Tuleya, R.E., Held, I.M., Villarini, G., 2013. Dynamical downscaling projections of twenty-first-century Atlantic hurricane activity: CMIP3 and CMIP5 model-based scenarios. *Journal of Climate* 22, 6591–6617. doi:10.1175/JCLI-D-12-00539.1.

Tropical Cyclone Rainfall Rate Projections

Perhaps the most robust projection from models concerning future tropical cyclone activity, in terms of both model consistency and physical understanding, is that the average precipitation rate in the vicinity of tropical cyclones is expected to increase. This could be important for future damage because inland flooding from tropical cyclone rainfall has

caused considerable damage and loss of life in the historical period. The projected fractional increase in rain rate obtained from models seems to depend on the averaging radius considered. In a recent study, the rate increased by about 10% for an averaging radius ranging from 200 to 400 km. The rainfall rate increased by as much as 30% or more (Figure 9) when averaged over a smaller region (50 km radius about the storm center).

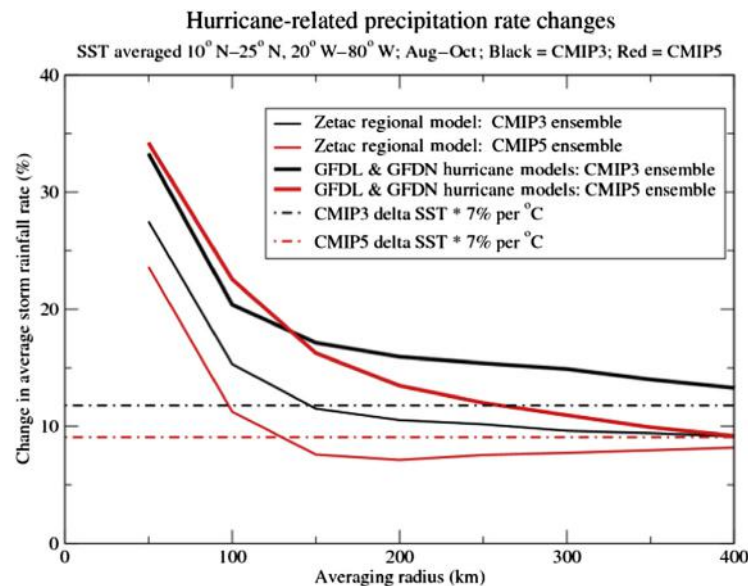


Figure 9 Hurricane-related precipitation rate changes in percent for two different hurricane simulation models (Zetac and GFDL/GFDL-Navy version (GFDN)) and for two different sets of climate model input (CMIP3 and CMIP5). The dashed lines show the approximate increase in water vapor content in the environment of the storms. The rain rate increases by about the same percentage as the water vapor content, except near the storm ($< 100 \text{ km}$ from center) where the percent increase in rainfall rate is even larger than that for the environmental water vapor. From Knutson, T.R., Sirutis, J.J., Vecchi, G.A., Garner, S.T., Zhao, M., Kim, H.-S., Bender, M.A., Tuleya, R.E., Held, I.M., Villarini, G., 2013. Dynamical downscaling projections of twenty-first-century Atlantic hurricane activity: CMIP3 and CMIP5 model-based scenarios. *Journal of Climate* 26, 6591–6617. doi:10.1175/JCLI-D-12-00539.1.

The physical mechanism for the rainfall increase involves an increase in moisture-holding capacity of the atmosphere at higher temperatures. As the climate warms in models due to increasing greenhouse gas concentrations, a robust response is that the relative humidity in the lower troposphere does not change very much. Since warmer air at a given relative humidity holds more water vapor than cooler air, the tropical atmosphere in the warmer climate is projected to contain significantly more water vapor (about 7% more for every 1 °C rise in tropical SST, shown schematically by dashed lines in Figure 9). For a tropical cyclone, most of the moisture being rained out of the storm (at least within 400 km or less from the center) is fed by water vapor converged into the storm by the winds, rather than being evaporated from the ocean within the 400 km radius region. Thus, with climate warming the atmosphere holds more water vapor, so this moisture convergence is enhanced, and the precipitation rate consequently increases. Increased intensity of the cyclone circulation can further enhance the moisture convergence.

Influence of Sea Level Rise on Tropical Cyclone Storm Surge

Storm surge, which is a surge of ocean water onto the land associated with a tropical cyclone, is responsible for most of the deaths globally in tropical cyclones, and for a considerable fraction of the physical damage. The surge is caused mainly by the winds associated with a storm, although the low surface pressure also contributes. The magnitude of the surge depends on a number of factors, such as the near-shore bathymetry, the angle of 'attack' of the storm at the coastline, the storm size and intensity, and the timing and magnitude of the tide.

A robust projection of global climate models is a rise of global mean sea level over the next century. According to the IPCC AR4 report, the average rate of global mean sea level rise over the twenty-first century will very likely exceed that observed during 1961–2003 for a range of future emission scenarios. Projections of sea level rise over the twenty-first century remain uncertain due to a limited understanding of the likely contributions to sea level rise from the Greenland and Antarctic ice sheets.

Although the magnitude and regional distribution of the sea level rise have considerable uncertainty, an increase in sea level would be expected to exacerbate storm surge risk, all other factors being equal. In fact, much of the impact of sea level rise may be experienced through the interaction of sea level rise with extreme events such as tropical cyclones or other large coastal storms. In order to project whether storm surge damage will increase at a location and if so by how much, one needs to know not just the regional sea level change, but also the future changes in storm climate, such as landfalling frequency, tracks, size, intensity, and so forth at a location. This makes the projection of future storm surge damage a challenging research topic that is now beginning to be explored.

Summary and Conclusions

It remains uncertain whether any past changes in various tropical cyclone metrics (frequency, intensity, rainfall rates, and so

forth) exceed the levels expected from natural variability alone. Early analyses that appeared to show strong rising trends in several tropical cyclone indices and strong statistical relationships between storm activity and rising SSTs in the tropical Atlantic basin, were followed by other studies showing that observing system changes over time were likely responsible for the most pronounced trends. The lack of a clearly detectable climate change signal in past data limits our confidence in future projections of tropical cyclone activity for the twenty-first century, especially in comparison to other climate variables such as global mean surface temperature, where a much more robust climate change detection and attribution to human activities has been established and future projections are more confident.

Owing to the potential importance of the issue of tropical cyclones and climate change, attempts to infer future changes in tropical cyclones, using projected large-scale environmental changes from climate models, have received considerable attention. Relevant environmental changes include features such as oceanic warming, tropospheric warming, increased water vapor, sea level rise, atmospheric circulation changes, and even changes in the CO₂ concentration alone. Using various statistical modeling and dynamical simulation methods, the future (late twenty-first century) picture that emerges is a warmer climate with roughly 20% fewer tropical storms globally, about a 5% increase in the average intensity of tropical cyclones, and possibly more of the most intense (Category 4 and 5) systems. Owing to the higher atmospheric water vapor content, an increase in the precipitation rates near tropical cyclones, by about 20%, is also projected. These changes are projected to occur against a backdrop of rising sea levels, which, at least without considering changes in tropical cyclone climate, would lead to a greater risk of storm surge flooding.

Further Reading

- Callaghan, J., Power, S.B., 2011. Variability and decline in the number of severe tropical cyclones making land-fall over eastern Australia since the late nineteenth century. *Climate Dynamics* 37, 647–662.
- Elsner, J.B., Kossin, J.P., Jagger, T.H., 2008. The increasing intensity of the strongest tropical cyclones. *Nature* 455, 92–95. <http://dx.doi.org/10.1038/nature07234>.
- Emanuel, K.A., 1987. The dependence of hurricane intensity on climate. *Nature* 326, 483–485.
- Emanuel, K.A., 2007. Environmental factors affecting tropical cyclone power dissipation. *Journal of Climate* 20, 5497–5509.
- Grinsted, A., Moore, J.C., Jevrejeva, S., 2012. A homogeneous record of Atlantic hurricane surge threat since 1923. *Proceedings of the National Academy of Sciences USA* 109 (48), 19601–19605.
- IPCC, 2007. *Climate change 2007: the physical science basis*. In: Solomon, S., Qin, D., Manning, M., Chen, Z., Marquis, M., Averyt, K.B., Tignor, M., Miller, H.L. (Eds.), *Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, Cambridge, UK, and New York, NY, USA, 996 pp.
- Knutson, T.R., McBride, J., Chan, J., Emanuel, K.A., Holland, G., Landsea, C., Held, I.M., Kossin, J., Srivastava, A.K., Sugi, M., 2010. Tropical cyclones and climate change. *Nature Geoscience* 3, 157–163. <http://dx.doi.org/10.1038/ngeo779>.
- Knutson, T.R., Sirutis, J.J., Vecchi, G.A., Garner, S.T., Zhao, M., Kim, H.-S., Bender, M.A., Tuleya, R.E., Held, I.M., Villarini, G., 2013. Dynamical downscaling projections of twenty-first-century Atlantic hurricane activity: CMIP3 and CMIP5 model-based scenarios. *Journal of Climate* 26, 6591–6617. doi:10.1175/JCLI-D-12-00539.1.
- Lee, T.-C., Knutson, T.R., et al., 2012. Impacts of climate change on tropical cyclones in the western North Pacific basin, Part I: Past observations. *Tropical Cyclone Research and Review* 1 (2), 213–230.

- Lin, N., Emanuel, K., Oppenheimer, M., Vanmarcke, E., 2012. Physically based assessment of hurricane surge threat under climate change. *Nature Climate Change*. <http://dx.doi.org/10.1038/nclimate1389>.
- Murakami, H., Wang, Y., Yoshimura, H., Mizuta, R., Sugi, M., Shindo, E., Adachi, Y., Yukimoto, S., Hosaka, M., Kusunoki, S., Ose, T., Kitoh, A., 2012. Future changes in tropical cyclone activity projected by the new high-resolution MRI-AGCM. *Journal of Climate* 25, 3237–3260.
- Seneviratne, S.I., Nicholls, N., Easterling, D., Goodess, C.M., Kanae, S., Kossin, J., Luo, Y., Marengo, J., McInnes, K., Rahimi, M., Reichstein, M., Sorteberg, A., Vera, C., Zhang, X., 2012. Changes in climate extremes and their impacts on the natural physical environment. In: Field, C.B., Barros, V., Stocker, T.F., Qin, D., Dokken, D.J., Ebi, K.L., Mastrandrea, M.D., Mach, K.J., Plattner, G.-K., Allen, S.K., Tignor, M., Midgley, P.M. (Eds.), *Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation*. Cambridge University Press, Cambridge, UK, and New York, NY, USA, pp. 109–123. A Special Report of Working Groups I and II of the Intergovernmental Panel on Climate Change (IPCC).
- Vecchi, G.A., Knutson, T.R., 2011. Estimating annual numbers of Atlantic hurricanes missing from the HURDAT database (1878–1965) using ship track density. *Journal of Climate* 24 (6). <http://dx.doi.org/10.1175/2010JCLI3810.1>.
- Vecchi, G.A., Swanson, K.L., Soden, B.J., 2008. Whither hurricane activity. *Science* 322 (5902). <http://dx.doi.org/10.1126/science.1164396>.
- Zhao, M., Held, I., Lin, S.-J., Vecchi, G.A., 2009. Simulations of global hurricane climatology, interannual variability, and response to global warming using a 50 km resolution GCM. *Journal of Climate* 22, 6653–6678.
- Zhang, R., Delworth, T.L., Sutton, R., Hodson, D., Dixon, K.W., Held, I.M., Kushnir, Y., Marshall, D., Ming, Y., Msadek, R., Robson, J., Rosati, A., Ting, M., Vecchi, G.A., 2013. Have aerosols caused the observed Atlantic multidecadal variability? *Journal of Atmospheric Sciences* 70 (4). <http://dx.doi.org/10.1175/JAS-D-12-0331.1>.

Tropical Cyclones in the Western North Pacific

JCL Chan, City University of Hong Kong, Hong Kong

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

This article presents a climatology of tropical cyclones in the western North Pacific ocean basin, as well as the mechanisms of genesis of these cyclones in this ocean basin. Some unique features of the cyclone tracks and the variations of tropical cyclone activity in this ocean basin on different temporal scales are also described.

Introduction

Marks (2003) presented an overview of the climatology, structure, formation, and motion of a tropical cyclone (TC) as well as the interaction of the TC with the ocean, the hazards associated with a TC, and the state of art in the forecasting of TCs. In this article, we will focus on the TCs in one particular ocean basin – the western North Pacific (WNP). As will be seen below, WNP is the ocean with the highest annual number of TCs almost every year as well as the highest number of intense TCs. It is thus of interest to study in greater detail TCs in this ocean basin.

In this article, we will first present the climatology of TCs in this ocean basin, followed by the mechanisms of genesis. Some unique features of the TC tracks in the WNP will then be described. The large variations of TC activity in this ocean basin on various temporal scales will also be discussed in relation to various atmospheric phenomena. The structure of TCs in the WNP will not be included because it is the same as those in other basins and has been presented in detail in Marks (2003). The mechanisms of intensification are also not unique in the WNP and will also not be discussed. Interested readers should refer to Marks (2003).

Climatology

As can be seen from Figure 1, TCs can form throughout the 12 months of the year, with the maximum occurrence of six or

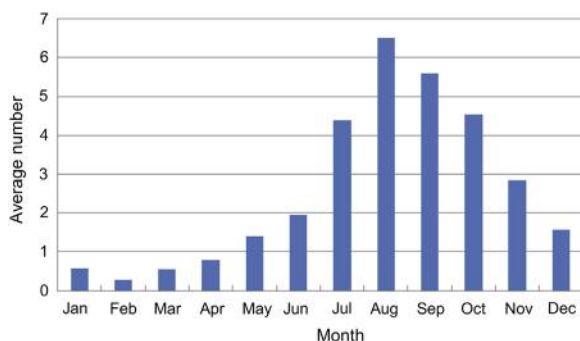


Figure 1 Average number of TCs in the western North Pacific in each month during the period 1959–2011. Data source: US Joint Typhoon Warning Center.

more in August on the average. During the period 1959–2011, the mean annual total number of TCs is close to 31, which is the highest among all tropical ocean basins. Of these, about 17 reach typhoon intensity (maximum sustained winds (MSWs) near the earth's surface $\geq 118 \text{ km h}^{-1}$), close to 10 could only attain either severe tropical storm (MSW: $88\text{--}117 \text{ km h}^{-1}$) or tropical storm intensity (MSW: $63\text{--}87 \text{ km h}^{-1}$), and the rest are tropical depressions (MSW $< 63 \text{ km h}^{-1}$).

A large year-to-year variation of TC activity in the WNP exists, as can be seen from the annual number of TCs during the period 1959–2011 (Figure 2). The annual number of TCs varies from a high of 44 in 1964 and 1996 to a low of 19 in 2010, and apparently goes through a multidecadal cycle, with maximum in the 1960s and 1990s, and a minimum in the late 1970s, and in the recent decade. There also appears to be a downward trend, with a decrease of around one TC per decade. A similar multidecadal cycle can also be found in the annual number of typhoons (blue curve in Figure 2). A correlation of ~ 0.7 exists between the two time series. A more detailed discussion of these variations will be presented later in this article on the climatic variations.

Genesis

The WNP has the largest area of high sea-surface temperature (SST) among all tropical oceans (see Figure 3 in Marks, 2003), with the area of over 27°C covering almost the entire ocean basin. As a result, the thermodynamic conditions necessary for genesis (see Marks, 2003) are satisfied throughout the year. Therefore, if the dynamic conditions can also be satisfied, TC genesis can occur.

In the WNP, a monsoon trough is present from about May to October each year, which can be seen from the flow pattern in the lower troposphere (Figure 3). In the tropical region, the west to southwesterly flow extends from the Indochina region into the South China Sea. At the same time, two cross-equatorial flows, one between 110 and 130°E and another between 140 and 160°E converge east of the Philippines. These three flows then form the monsoon trough that extends from northern Vietnam through Hainan Island and south-eastward through the Philippines, and ends at around 140°E . The relative vorticity distribution over the ocean basin also shows a similar feature (Figure 4). The highest vorticity patch stretches from the South China Sea all the way to the dateline. Thus, convective clusters that form within this strip of high

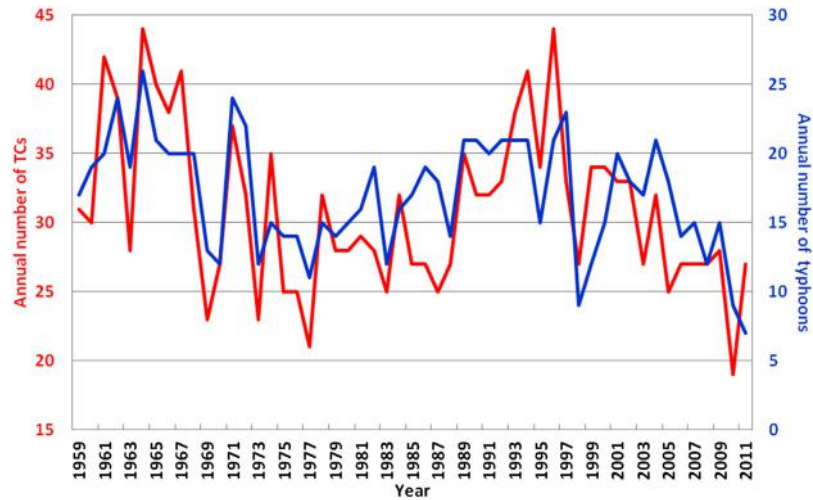


Figure 2 Annual number of TCs (red curve, scale on left axis) and of typhoons (blue curve, scale on right axis) in the WNP during the period 1959–2011. Data source: US Joint Typhoon Warning Center.

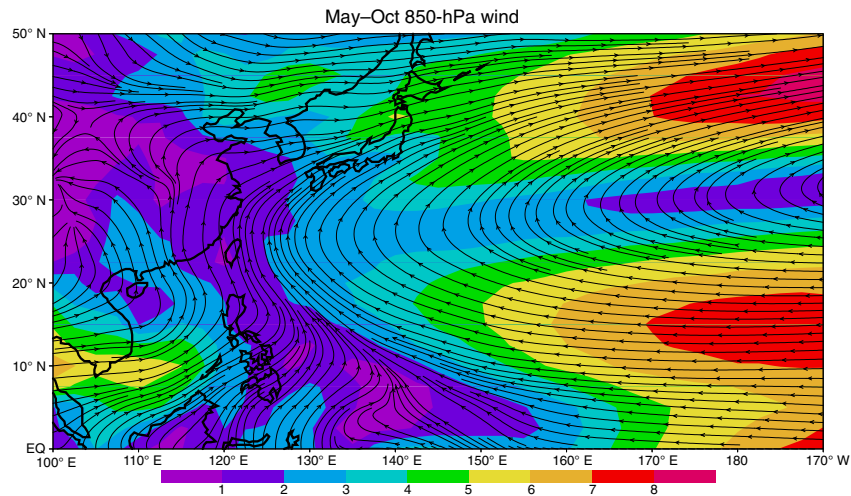


Figure 3 850-hPa flow in the WNP averaged between May and October. Color shading indicates the flow speed in m s^{-1} . Plotted using the US/NOAA ESRL webpage.

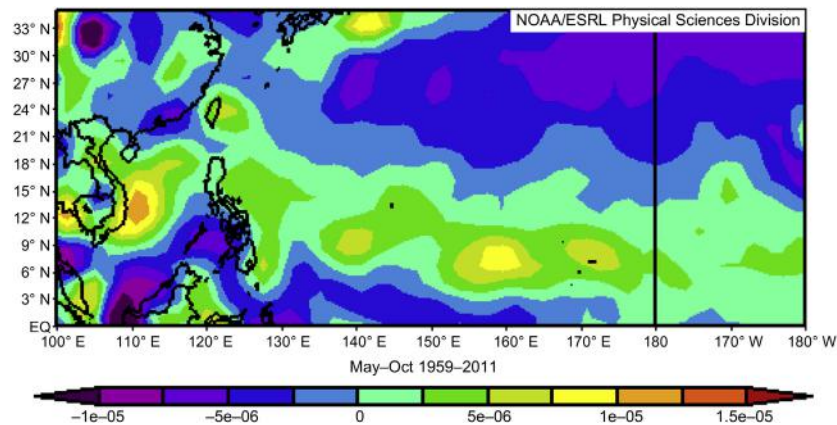


Figure 4 Relative vorticity (s^{-1}) over the WNP near the earth's surface averaged between May and October during the period 1959–2011. Plotted using the US/NOAA ESRL webpage.

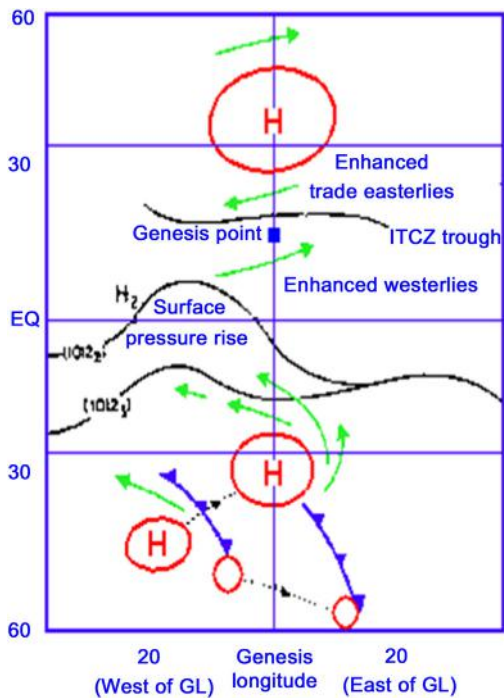


Figure 5 Schematic of the triggering of TC genesis through a surge from the Southern Hemisphere. The green arrows indicate the directions of the low-level flow. Adapted from Love, G., 1985. Cross-equatorial interactions during tropical cyclogenesis. *Monthly Weather Review* 113, 1499–1509.

relative vorticity already have the background vorticity condition satisfied. What it takes is a trigger to spin up the cluster.

In a majority of cases, the trigger comes from the Southern Hemisphere in the form of a monsoon surge (Figure 5). As a cold front in the Southern Hemisphere advances northward, the subtropical high is pushed toward the equator. This causes a pressure rise across the equator, triggering a southwesterly flow. If this is coupled with an enhancement of easterlies in the

Northern Hemisphere due to an intensification of the subtropical high, additional positive vorticity will be injected into a preexisting cluster, providing it with enough spin to become a TC.

Another trigger occurs in May and June over the South China Sea associated with a Mei-yu front (Figure 6). The southwesterly flow associated with the summer monsoon and a continental anticyclone over Asia maintain a Mei-yu front along the South China coast (Figure 6, left panel). As the continental high moves eastward, the front migrates southward (Figure 6, middle panel). When the high moves out of the Asia mainland, a surge of northeasterlies then couples with southwesterlies to form a tropical depression (Figure 6, right panel).

Trade wind surges can also serve as a trigger, especially in the late season (Figure 7). At first, the trade wind maintains an area of cyclonic vorticity to its south (Figure 7(a) and 7(b)). A continental anticyclone moving out of the Asia mainland enhances the northeasterlies north of the area of cyclonic vorticity and increases the magnitude of the latter (Figure 7(c)) and finally spins up a TC (Figure 7(d)). A similar mechanism exists for genesis in the South China Sea associated with a surge of the northeast monsoon, again in the late season (Figure 8).

Another mechanism, though not very common, of TC genesis in the WNP is associated with the Tropical Upper Tropospheric Trough (TUTT) (Figure 9). If a cloud cluster and an associated low-pressure system (D) is positioned in such a way that its outflow can be enhanced by a cell (C) in the TUTT, this system can develop into a TC (Sadler, 1976).

A unique feature in the WNP is the monsoon gyre, which is a cyclonic circulation with a radius of over 1000 km. Occasionally, cloud clusters embedded in the periphery of this gyre can become TCs that rotate cyclonically around the center of the gyre (Figure 10(a)). In some instances, the gyre eventually becomes a large TC itself (Figure 10(b)).

The different environmental flow patterns for genesis have also been classified by Ritchie and Holland (1999), and more recently by Yoshida and Ishikawa (2013), into five types: monsoon shear line, monsoon confluence region, monsoon gyre, easterly wave, and preexisting TC. The first three types have been discussed in detail above. The easterly wave type was

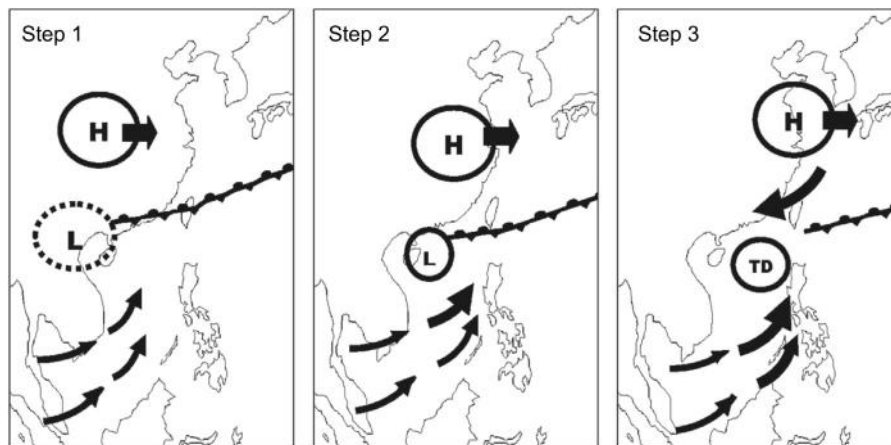


Figure 6 Three steps in the conceptual model of TC formation associated with a Mei-yu front. From Lee, C.-S., Lin, Y.-L., Cheung, K.K.W., 2006. Tropical cyclone formations in the South China Sea associated with the Mei-Yu front. *Monthly Weather Review* 134, 2670–2687.

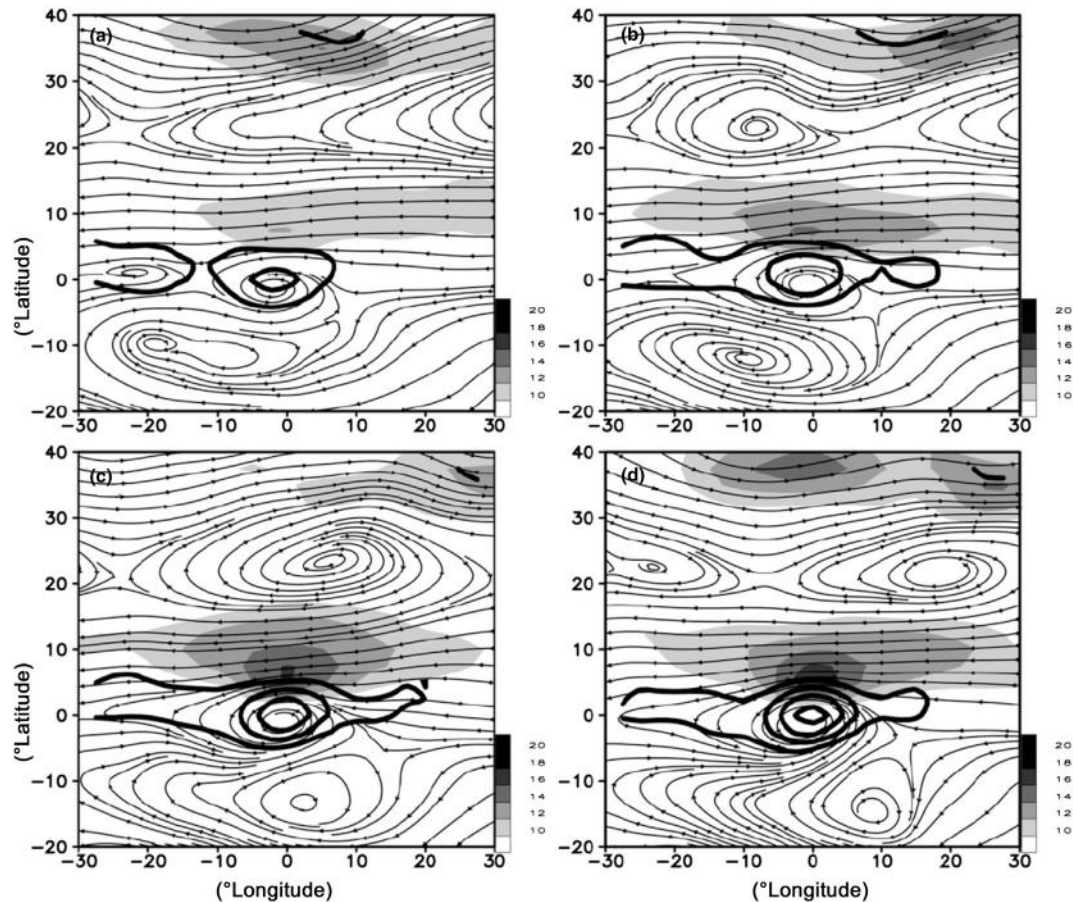


Figure 7 Composite of 850- to 925-hPa flow during TC formation associated with a trade wind surge. (a)–(d) are at –72, –48, –24, and time of TC formation alert, respectively. The system center is located at the coordinate (0,0). The heavy contours indicate positive relative vorticity (interval $1 \times 10^{-5} \text{ s}^{-1}$) and the shaded regions are wind speeds $>10 \text{ m s}^{-1}$ (interval: 2 m s^{-1}). From Chang, L.-Y., Cheung, K.K.W., Lee, C.-S., 2010. The role of trade wind surges in tropical cyclone formations in the western North Pacific. *Monthly Weather Review* 138, 4120–4134.

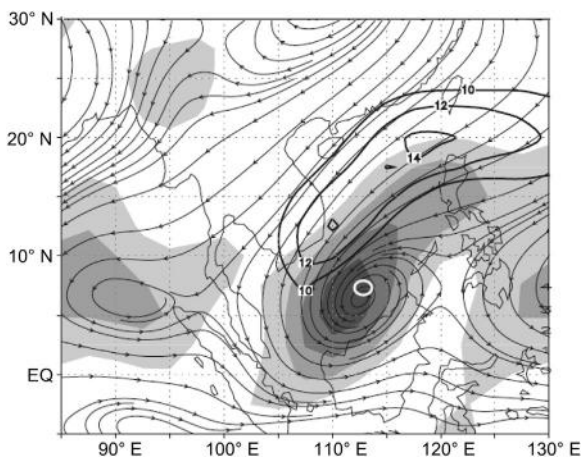


Figure 8 Composite streamlines and isotachs at 925 hPa associated with TC formation over the South China Sea at 42 h after a closed isobar is first identified on surface weather maps. The shading is relative vorticity (10^{-5} s^{-1}) and the white circle is the system center. Adapted from Lin, Y.-L., Lee, C.-S., 2011. An analysis of tropical cyclone formations in the South China Sea during the late season. *Monthly Weather Review* 139, 2748–2760.

considered by some researchers as part of the monsoon trough disturbance (see [Figure 3](#)) or a local vorticity maximum within the trade wind surge (see [Figure 7](#)). Indeed, [Ritchie and Holland \(1995\)](#) concluded that there is not enough evidence to support this type of genesis mechanism.

[Carr and Elsberry \(1997\)](#) first noted from barotropic model simulations that an anticyclone and cyclone pair could develop to the southeast of a relatively large TC through Rossby wave dispersion ([Figure 11](#)). [Ritchie and Holland \(1999\)](#), and later [Li and Fu \(2006\)](#), have shown that some TCs could indeed develop from a preexisting TC ([Figure 12](#)), although such formation does not occur very frequently as the TC has to have a certain size and the environmental flow also needs to have a good background relative vorticity ([Li et al., 2006](#)).

The prevalence of the monsoon trough throughout the peak season and the occurrence of the northeast monsoon explain why TCs can occur throughout the year in the WNP. The climatological distribution of the genesis locations ([Figure 13](#)) shows two maxima, one over the South China Sea and the other east of the Philippines. The former is mostly associated with the Mei-Yu front and northeast monsoon surges, while the latter with the monsoon trough and the trade wind surges. The monsoon trough shown in [Figure 4](#) and the

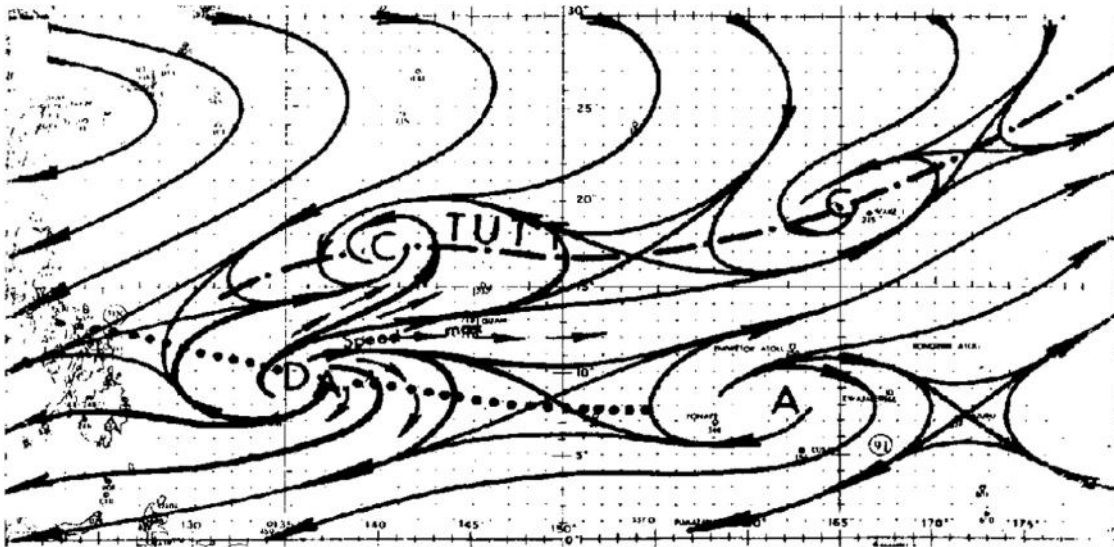


Figure 9 A model of how a cell (labeled C to the northeast of D) in the TUTT can enhance the formation of a TC (located at position D). From Sadler, J.C., 1976. A role of the tropical upper tropospheric trough in early season typhoon development. *Monthly Weather Review* 104, 1266–1278.

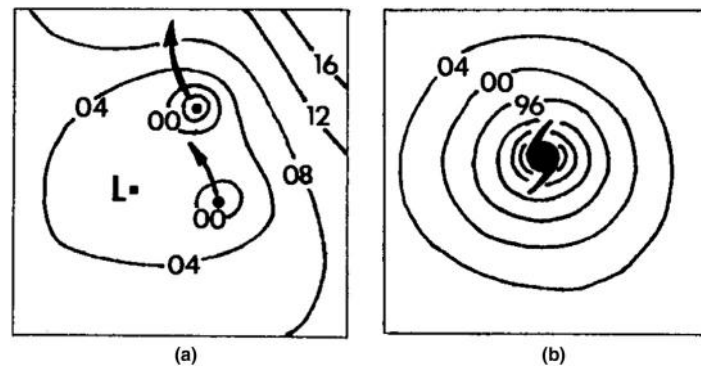


Figure 10 Schematic of TC genesis associated with a monsoon gyre. (a) Two TCs form in the eastern periphery of the gyre, with the center indicated by the letter 'L'. (b) The monsoon gyre becomes a TC itself. The contours are surface isobars. Adapted from Lander, M.A., 1994. Description of a monsoon gyre and its effects on tropical cyclones in the western North Pacific during August 1991. *Weather Forecasting* 9, 640–654.

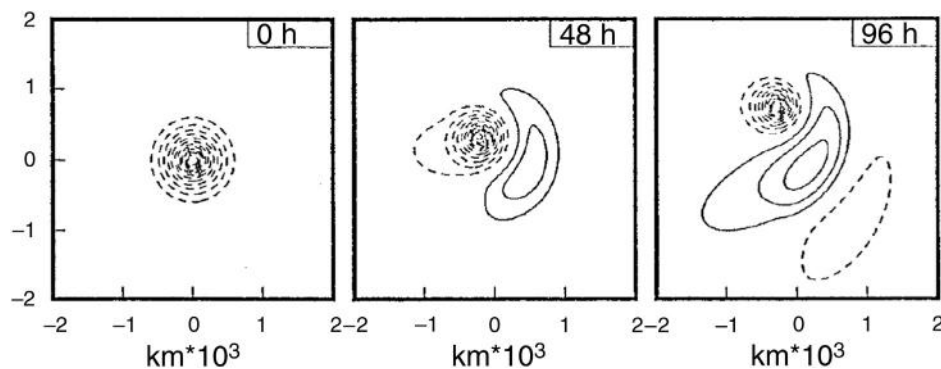


Figure 11 An example of the formation of an anticyclone–cyclone pair to the southeast of an existing cyclonic vortex during the first 96 h of integration in a barotropic model. Dashed (solid) lines indicate negative (positive) stream function fields. This example is for a vortex with vanishing winds at 800 km from the vortex center located at 20° N. Adapted from Carr and Elsberry (1997).

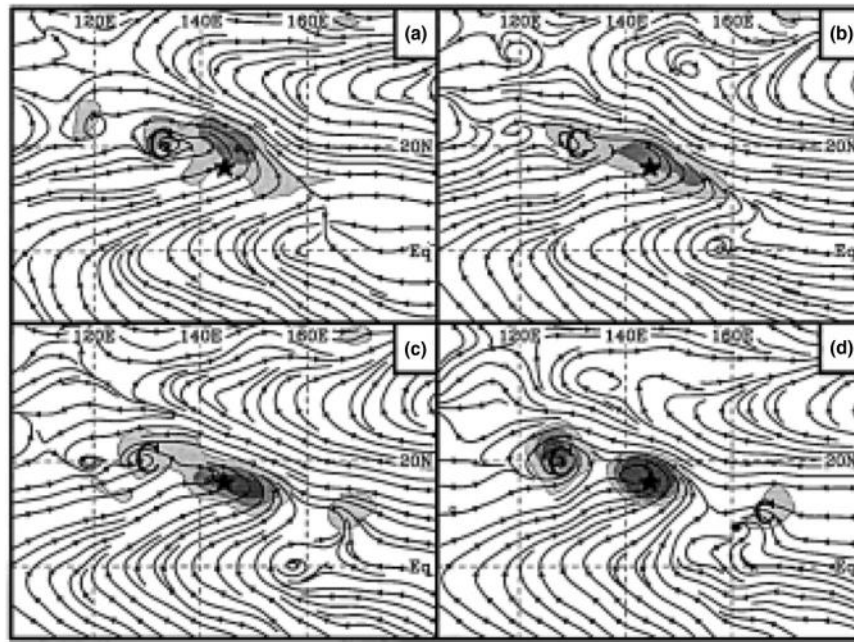


Figure 12 Composite 850-hPa streamlines and values of cyclonic relative vorticity (shaded, only values $>8 \times 10^{-6} \text{ s}^{-1}$, contour interval: $4 \times 10^{-6} \text{ s}^{-1}$) associated with TC genesis from a preexisting TC at (a) -72, (b) -48, (c) -24 h, and (d) genesis time. 'C' in (a)–(c) indicates the precedent TC and in (d) indicates the new TC. The star symbol indicates the mean genesis location in the composite of Ritchie and Holland (1995). Adapted from Ritchie, E.A., Holland, G.J., 1999. Large-scale patterns associated with tropical cyclogenesis in the Western Pacific. *Monthly Weather Review* 127, 2027–2043.

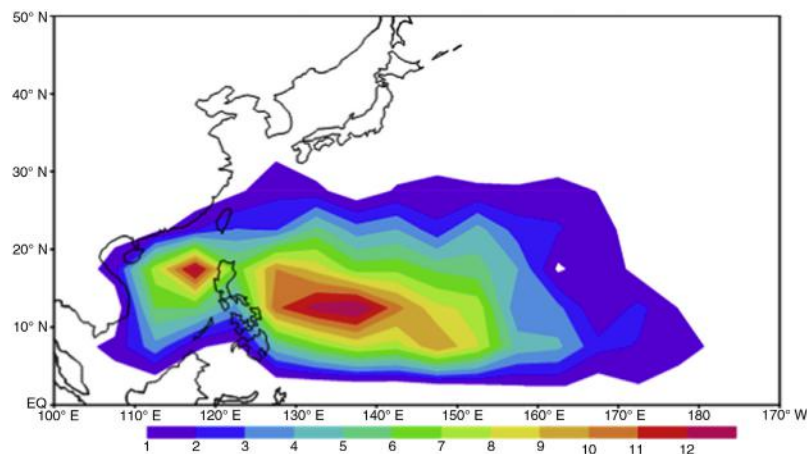


Figure 13 Frequency of TC genesis in the western North Pacific (number per 10 years). Data source: US Joint Typhoon Warning Center.

area of highest relative vorticity in [Figure 3](#) match very well with the locations of the peak values in genesis frequencies in [Figure 13](#), which suggests that in the WNP, the dynamic conditions are the determining factors in controlling TC genesis.

It should also be noted that the contours in [Figure 13](#) only go down to 1 (per 10 years). Some TCs actually form outside these contours, especially for very low-latitude genesis. For example, Typhoon Vamei (2001) formed at only 1.4° N , which is probably the lowest latitude at which a TC has ever formed ([Chang et al., 2003](#)).

The Fujiwhara Effect

During the peak season in the WNP (May to October), it is likely that more than one TC is present at any given time. Sometimes, they are in quite close proximity to each other and the circulation of one TC can affect the movement of the other. This is known as the Fujiwhara effect. The two TCs tend to rotate about a common 'center of mass.' [Lander and Holland \(1993\)](#) proposed a schematic in which the two TCs first approach each other, then go through a capture process and finally escape from each other, with each moving away from

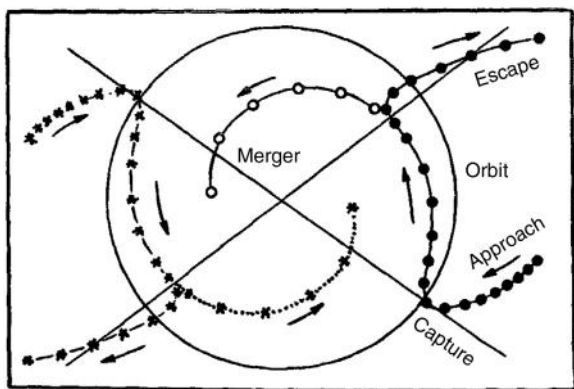


Figure 14 Schematic showing the relative tracks of two TCs that are in close proximity to each other. The long-dashed track of the TC represented by the crosses and the continuous line connecting the filled circles representing the other TC show the 'escape routes' of the two TCs. The dotted track and the open circles represent the merger process. From Lander, M.A., Holland, G.J., 1993. On the interaction of tropical cyclone scale vortices. I: Observations. *Quarterly Journal of the Royal Meteorological Society* 119, 1347–1361.

the other (Figure 14). In some instances, the TCs can merge during the capture process and become one TC. Whether or not the two TCs will merge or escape depends on the structure of the vortices (Chan and Law, 1995). While the tracks portrayed in Figure 12 are the relative ones, the actual tracks can appear to be very erratic.

Temporal Variations

As shown in Figure 2, the annual number of TCs in the WNP can have a large variation. An important factor in driving the variation on an interannual timescale is the occurrence of an El Niño or La Niña. In general, during an El Niño year, there tends to be more TCs and they form further to the southeast of the ocean basin. In fact, the area of TC genesis shown in Figure 11 east of 160° E is mainly contributed by those during El Niño years. On the other hand, TC genesis during La Niña years tends to occur in the northwest of the ocean basin. During El Niño years, because more TCs form further to the southeast, they have longer tracks over the ocean. As a result, they have a higher chance of becoming more intense. Therefore, the accumulated cyclone energy, defined as the sum of the squares of the MSW every 6 h of all TCs during a year, in an El Niño year tends to be higher than that in a La Niña year. It is useful to point out that the SST in the WNP during El Niño years is actually slightly lower than that during La Niña years (Chan and Liu, 2004), although it is still $>27^{\circ}\text{C}$. Thus, as discussed above, variations in the SST in the WNP do not control TC genesis there. Such a change in SST is actually in response to the SST conditions in the El Niño region.

On a longer time scale, there has been some evidence of the relationship between the stratospheric quasi-biennial oscillation (QBO) and TC activity variations. This relationship exists because of the change in the vertical wind shear during the different phases of the QBO, and hence the conditions for TC

genesis (Chan 1995). Both the number of TCs and the number of typhoons also exhibit multidecadal variations (Figure 2). These have been shown to be related to similar variations in the atmospheric flow patterns (Chan, 2008).

Summary

This article presents some particular features of TCs in the WNP, especially on the climatology, genesis mechanisms, the occurrence and movement of binary cyclones, and the relationship between TC occurrence and other atmospheric phenomena. No discussion is given on other aspects as they are not unique to this ocean basin and have been largely covered in Marks (2003). Although some statements on the possible variation in TC activity in relation to global warming have been made (Knutson et al., 2010), this topic is also not considered because studies on this topic are still ongoing.

Recently, some topics have received more attention such as changes in intensity as a result of the so-called eyewall replacement cycles that are more prevalent in the WNP, and the enhancement of rainfall around landfall as a result of the interaction between the TC and the environment such as the monsoon, especially when steep topography is present as in the case of Taiwan or the Philippines. In addition, the effect of warm ocean eddies in the WNP associated with the Kuroshio Current on the intensification of TC is also an important topic. All these are still under investigation.

See also: Tropical Cyclones and Hurricanes: Hurricanes: Observation.

References

- Carr III, L.E., Elsberry, R.L., 1997. Models of tropical cyclone wind distribution and beta-effect propagation for application to tropical cyclone track forecasting. *Monthly Weather Review* 125, 3190–3209.
- Chan III, J.C.L., 1995. Tropical cyclone activity in the western North Pacific in relation to the stratospheric quasi-biennial oscillation. *Monthly Weather Review* 123, 2567–2571.
- Chan, J.C.L., 2008. Decadal variations of intense typhoon occurrence in the western North Pacific. *Proceedings of the Royal Society A* 464, 249–272.
- Chan, J.C.L., Law, A.C.K., 1995. The interaction of binary vortices in a barotropic model. *Meteorology and Atmospheric Physics* 56, 135–155.
- Chan, J.C.L., Liu, K.S., 2004. Global warming and western North Pacific typhoon activity from an observational perspective. *Journal of Climate* 17, 4590–4602.
- Chang, C.-P., Liu, C.-H., Kuo, H.-C., 2003. Typhoon Vamei: an equatorial tropical cyclone formation. *Geophysical Research Letters* 30, 1150. <http://dx.doi.org/10.1029/2002GL016365>.
- Chang, L.-Y., Cheung, K.K.W., Lee, C.-S., 2010. The role of trade wind surges in tropical cyclone formations in the western North Pacific. *Monthly Weather Review* 138, 4120–4134.
- Knutson, T.R., McBride, J.L., Chan, J.C.L., Emanuel, K., Holland, G., Landsea, C., Held, I., Kossin, J.P., Srivastava, A.K., Sugi, M., 2010. Tropical cyclones and climate change. *Nature Geoscience* 3, 157–163.
- Lander, M.A., 1994. Description of a monsoon gyre and its effects on tropical cyclones in the western North Pacific during August 1991. *Weather Forecasting* 9, 640–654.
- Lander, M.A., Holland, G.J., 1993. On the interaction of tropical cyclone scale vortices. I: Observations. *Quarterly Journal of the Royal Meteorological Society* 119, 1347–1361.
- Lee, C.-S., Lin, Y.-L., Cheung, K.K.W., 2006. Tropical cyclone formations in the South China Sea associated with the Mei-Yu front. *Monthly Weather Review* 134, 2670–2687.

- Li, T., Fu, B., 2006. Tropical cyclogenesis associated with Rossby wave energy dispersion of a preexisting typhoon. Part I: Satellite data analyses. *Journal of Atmospheric Sciences* 63, 1377–1389.
- Li, T., Ge, X., Wang, B., Zhu, Y., 2006. Tropical cyclogenesis associated with Rossby wave energy dispersion of a preexisting typhoon. Part II: Numerical simulations. *Journal of Atmospheric Sciences* 63, 1390–1409.
- Lin, Y.-L., Lee, C.-S., 2011. An analysis of tropical cyclone formations in the South China Sea during the late season. *Monthly Weather Review* 139, 2748–2760.
- Love, G., 1985. Cross-equatorial interactions during tropical cyclogenesis. *Monthly Weather Review* 113, 1499–1509.
- Marks, F.D., 2003. Hurricanes. In: *Encyclopedia on Atmospheric Sciences*.
- Ritchie, E.A., Holland, G.J., 1999. Large-scale patterns associated with tropical cyclogenesis in the Western Pacific. *Monthly Weather Review* 127, 2027–2043.
- Sadler, J.C., 1976. A role of the tropical upper tropospheric trough in early season typhoon development. *Monthly Weather Review* 104, 1266–1278.
- Yoshida, R., Ishikawa, H., 2013. Environmental factors contributing to tropical cyclone genesis over the western North Pacific. *Monthly Weather Review* 141, 451–467.

Tropical Cyclones: Secondary Eyewall Formation

C-C Wu and Y-H Huang, National Taiwan University, Taipei, Taiwan

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

Secondary eyewall formation (SEF), a phenomenon often observed in intense tropical cyclones (TCs), is one of the key issues for TC research and forecasting as indicated by the widely documented findings related to the increase in storm size and short-term changes in TC intensity. This review article provides a comprehensive summary and discussions of current understandings of the plausible physical mechanisms for SEF in the literature. Provided with a favorable moist environment, a variety of internal dynamical processes for SEF have been proposed, such as the axisymmetrization process, energy accumulation through vortex Rossby wave activities, beta-skirt-induced energy cascade, unbalanced responses to boundary layer dynamics, and the balanced response to convective heating. The merits and caveats among different dynamical interpretations are discussed, while prominent unsolved SEF issues are also addressed to provide valuable guidance for future SEF research.

Background

Secondary eyewall formation (SEF) and the eyewall replacement cycle in tropical cyclones (TCs) have been widely documented by aircraft observations and high-resolution satellite imagery. Two concentric quasi-circular deep convective rings (inner and outer TC eyewalls) and a nearly cloud-free region (moat) located in between can be easily identified during the double-eyewall episode in TCs (Figure 1). For most such cases, the outer eyewall is established later, with characteristics similar to the inner eyewall. A local maximum swirling wind is often present in the outer eyewall, with the strong swirling wind region confined to the lower- or midtroposphere. Statistical analyses based on a 10-year data set (1997–2006) showed

that on average 70% of the Atlantic, 50% of the eastern Pacific, 40% of the Southern Hemisphere, and 80% of the western Pacific intense storms (maximum wind >120 kts) underwent at least one eyewall replacement cycle during their lifetime. Consistent statistical results were also obtained in recent studies, showing that SEF is a relatively common phenomenon in intense TCs. Because SEF, along with the subsequent eyewall replacement cycle, is often associated with temporary weakening of storms and concomitant increase in the extent of damaging gale-force winds, it remains as an important forecast priority for populated coastal communities and seagoing vessels over the open ocean.

Numerical simulations have been extensively applied to investigate SEF in many previous studies. Based on physical

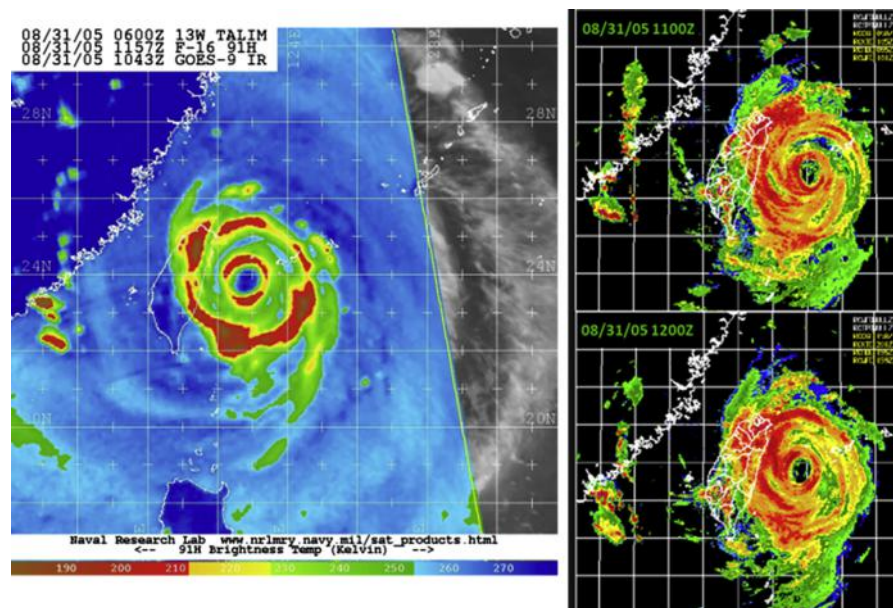


Figure 1 Concentric eyewalls in Typhoon Talim (2005) near Taiwan. The figure on the left is the image retrieved from the Special Sensor Microwave Imager/Sounder sensor on the Defense Meteorological Satellite Program satellite at 1157 UTC 31 August 2005, available at the NRL (Naval Research Laboratory) Tropical Cyclone Page (www.nrlmry.navy.mil/sat_products.html). The panel on the right contains radar images of composite reflectivity (dBZ) at 1100 and 1200 UTC 31 August 2005, provided by the Central Weather Bureau (CWB) of Taiwan.

analyses and sensitivity experiments among numerical studies, a number of factors and mechanisms have been suggested as contributors to SEF. Section [Environmental Conditions](#) introduces the possible roles of various environmental conditions in SEF. As for the internal mechanisms responsible for SEF, there are a number of hypotheses proposed in the literature. Recent studies have found that the physical constraints applied to some hypotheses are not able to fully capture the physical processes involved in the life cycle of a real SEF event. Numerical simulations with more sophisticated physical processes, such as diabatic heating and boundary layer processes, suggested that some of the proposed hypotheses could not adequately address the key internal mechanisms of SEF in a real TC. The details of the aforementioned hypotheses are introduced and discussed in Section [Internal Mechanisms of SEF](#). Finally, Section [Vortex Rossby Waves](#) contains the concluding remarks upon the current understanding of SEF and proposes key issues to be further investigated.

Environmental Conditions

Environmental moisture is one of the major factors affecting TC intensity and structure. Numerical models with different designs have been adopted to study the impact of ambient relative humidity on SEF. In the axisymmetric model framework, high relative humidity is critical to SEF only if the hydrostatic assumption is applied. In contrast, full-physics model simulations showed that high environmental relative humidity is crucial to the formation of an outer eyewall in a TC. A plausible explanation is that high ambient humidity can provide more moisture and thus can enhance latent heat release in the vortex's outer core region, which likely provides a favorable environment for the expansion of storm size, the formation/intensification of rainbands, and even SEF. Another elucidation for the environmental influence on SEF is the sustained eddy angular momentum fluxes caused by interactions between a mature TC and its environment. If this TC–environment interaction is sufficiently influential in strength and space that gives rise to the Wind-Induced Air-Sea Heat Exchange (WISHE) process, a secondary eyewall likely establishes outside the primary eyewall.

Internal Mechanisms of SEF

Five internal mechanisms that have been frequently discussed in existing SEF studies are introduced and discussed in the subsections below, which are titled vortex Rossby waves (VRWs), axisymmetrization process, beta-skirt axisymmetrization (BSA) formation hypothesis, unbalanced boundary layer dynamics near the top of TC boundary layer, and balanced response to diabatic heating in a region of enhanced inertial stability.

Vortex Rossby Waves

Based on radar and satellite images, it has been found that disturbances with similar characteristics to Rossby waves are

often associated with rainbands in TCs. VRWs are thus coined for these eddy activities in TCs. Similar to Rossby waves in a planetary system, the dispersion relation for VRWs is closely related to the vorticity gradient of TCs. In a dry and barotropic framework, Montgomery and Kallenbach obtained an analytic solution of the stagnation radius for VRWs that propagate radially outward from the eyewall. It has then been proposed that the accumulation of energy near that stagnation radius can have impacts on the outer rainband structure and perhaps on the formation of secondary wind maximum. Houze and colleagues found that the model failed to simulate the eyewall replacement cycle of Hurricane Rita (2005), when the model resolution was reduced from 1.67 to 5 km. Meanwhile, they noticed that during Rita's eyewall replacement cycle, the small-scale features captured by the radar in the inner core region may be related to the VRWs' dynamics. The importance of high model resolution and the value of targeting small-scale structures in TCs are thus suggested for a better understanding of SEF and the subsequent eyewall replacement cycle. In contrast to consensus on the role of VRWs in TC rainbands, some recent studies argued the role of VRWs in SEF by using the model results with more sophisticated physical processes. Judt and Chen indicated that the near-zero potential vorticity (PV) gradient, subsidence, and straining effect which are already present prior to SEF are not conducive for VRWs' outward propagation, and thus provided reasonable doubt against the essential role of VRWs in SEF. Noting the ambiguous role of the eddy fluxes associated with VRWs in speeding up the tangential wind in the SEF region, Corbosiero and colleagues inferred that the outward propagation of the convectively coupled VRWs from the inner eyewall may act to redistribute PV and dump out moisture at the stagnation radius. It was thus suggested that VRWs make the active convection more prominent, but not to the extent that would directly cause SEF.

Axisymmetrization Process

Drawing from studies of vortex dynamics, the axisymmetrization process has been used to interpret the formation of vorticity ring outside the parent vortex in a two-dimensional barotropic model. Kuo and colleagues suggested that the primary vortex can axisymmetrize weak vorticity patches into a vorticity ring, provided that the primary vortex is strong enough and the two vortices are sufficiently close to each other. Nevertheless, recent studies noted that PV patches outside the eyewall can be of comparable magnitude to that in the eyewall region and have dipole structures in the real TC environment as simulated with more realistic physical processes (such as the moist convection). In particular, Moon and colleagues demonstrated that the interaction between the TC core vortex and the convection-induced small vorticity dipoles of considerable strength in two-dimensional flows does not lead to the formation of a coherent concentric vorticity ring. Thus the axisymmetrization process under a simplified two-dimensional incompressible flow appears insufficient for describing SEF in the real atmosphere. The critical role of the three-dimensional moist process in the maintenance of a vorticity ring has also been shown in recent studies.

Beta-Skirt Axisymmetrization Formation Hypothesis

Terwey and Montgomery presented a new moist-based BSA formation hypothesis as an intrinsic SEF mechanism. This hypothesis requires a region with sufficiently long filamentation time and moist convective potential, a sufficient low-level radial PV gradient (i.e., a beta skirt) associated with the primary swirling flow and the follow-up WISHE process. The long filamentation time and sufficient moist convective potential set the scene for a convectively favorable environment. The beta-skirt structure and WISHE process provide a dynamical pathway to SEF. Applying the two-dimensional turbulence theory to the problem of SEF in a rotating TC environment, the theme of the BSA hypothesis is that the upscale energy cascade tends to occur on the beta skirt. Following this pathway, eddy kinetic energy associated with the sporadic convective cells outside the primary eyewall may be injected into the tangential direction and enhance local low-level jets (axisymmetrized into the mean tangential flow) on the skirt of the vortex's PV profile. Once the low-level jet strengthens substantially, it could further intensify by coupling with the boundary layer through a wind-induced moisture feedback process such as WISHE and may ultimately lead to SEF. The timescale of this energy cascade process and the width of the corresponding jet can be evaluated by the values of its PV gradient. Though simulated results in relevant studies showed consistency between the evaluated and simulated jet width, direct supporting evidence for the described energy upscale process needs to be further investigated.

Unbalanced Boundary Layer Dynamics near the Top of TC Boundary Layer

A deeper understanding of the underlying dynamics of SEF has been proposed in two recent companion studies, based on the two mechanisms for the spin-up of mean tangential winds in single-eyewall TCs. Both mechanisms are associated with the radial advection of absolute angular momentum ($M = fr^2/2 + rv$). The first mechanism is for the spin-up above the boundary layer where M is materially conserved. The convergence of M is enhanced by the negative radial gradient of a diabatic heating rate associated with convective structures in a TC. This mechanism has been addressed in many extant studies. It explains why the vortex expands in size in terms of axisymmetric balanced dynamics, wherein the vortex is well approximated by gradient wind and hydrostatic balance. The second mechanism is related to the spin-up process within the boundary layer and is considered important in the inner core region of the storm. Although M is not materially conserved here, tangential winds can still be enhanced if the boundary layer inflow is sufficiently large to bring the air parcels to the small radii with minimal loss of M to friction. The boundary layer flow is coupled to the interior flow via the radial pressure gradient at the top of the boundary layer, but the spin-up of a vortex is ultimately tied to the dynamics of the boundary layer where inflow is prevailing and the swirling wind is not in gradient wind balance over a substantial radial span.

By assimilating T-PARC (THORPEX Pacific Asian Regional Campaign) data (particularly aircraft reconnaissance and

surveillance observations) into the Weather Research and Forecasting (WRF) model based on ensemble Kalman filter data assimilation, Wu and colleagues constructed a model/observation-consistent and high-spatial/temporal resolution data set for Typhoon Sinlaku (2008) in the first part of the two companion works. The key features related to the horizontal broadening of low-level troposphere swirling flow and intensification of boundary layer inflow over the outer region are identified before SEF (Figure 2). These two important features are consistent with the two mechanisms highlighted for the spin-up of single-eyewall TCs and set the scene for a progressive boundary layer control pathway to SEF.

As the second part of the two companion papers, Huang and colleagues addressed the association between increases in storm size and SEF from the axisymmetric aspect. The findings point to collective structural changes in the outer core region of a mature TC, which ultimately culminates in the formation of a secondary eyewall. The sequence begins with a broadening of the low-level tangential wind field associated with the intensification of the eyewall that can be demonstrated by the balanced response above the boundary layer (the first mechanism mentioned above). Due to the presence of surface friction, boundary layer inflow increases underneath the broadened swirling wind, and becomes large enough to enhance the swirling circulation within the boundary layer (the second spin-up mechanism). This rapid increase in tangential winds near the top of the boundary layer breaks the gradient wind balance, leading to the local development of supergradient winds, which decelerate the inflow air parcels and impede them from moving inward (Figure 3). This process leads to the transition outside the primary eyewall from sporadic and/or weak convergence in the lower troposphere to a well-defined convergence zone concentrated within and just above the boundary layer. The progressive increase of supergradient forces thus continuously provides a mechanical mean for high enthalpy air to erupt from the boundary layer. Given the dynamically and thermodynamically favorable environment for convective activities, the progressive response of the unbalanced boundary layer flow to an expanding swirling wind field and the positive feedback loop in between are demonstrated to be an important mechanism for concentrating and sustaining deep convection in a narrow supergradient-wind zone collocated with the SEF region. The pathway to SEF can be summarized in a schematic diagram (Figure 4), elucidating the chain of TC structure changes and associated physical processes and feedbacks.

The SEF paradigm advanced in these two companion works is attractive on physical grounds because of its simplicity and consistency with the three-dimensional numerical simulations presented. While understanding the importance of the balanced response (cf [Balanced Response to Diabatic Heating in a Region of Enhanced Inertial Stability](#) subsection), this paradigm peculiarly highlights the critical role of unbalanced dynamics in SEF. Two recently published studies also investigated the impact of boundary layer dynamics on SEF from different perspectives. Concerning the asymmetry associated with rainbands that prevail prior to SEF, Qiu and Tan reexamined this SEF paradigm in an asymmetric framework. The sequence of structure changes

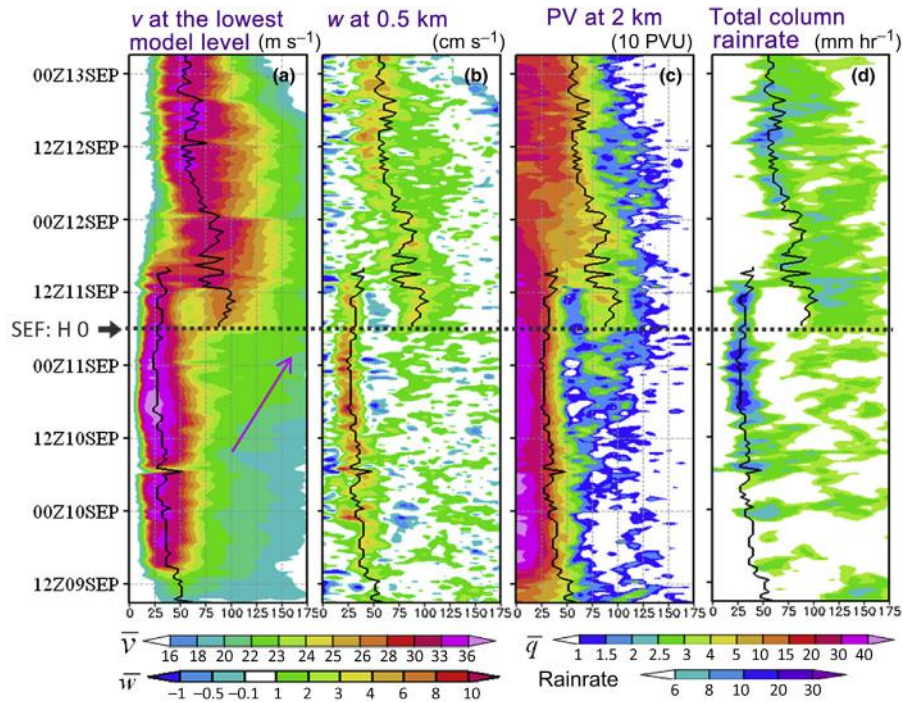


Figure 2 Time-radius diagrams of the azimuthally mean indicating (a) tangential wind (m s^{-1}) at the lowest model level, (b) vertical velocity (m s^{-1}) at 0.5 km height, (c) potential vorticity (10 (potential vorticity unit) PVU) at 2 km height, and (d) total column rainrate (mm h^{-1}) for the ensemble mean (average of 28 ensemble members). SEF time is indicated by a dotted line and an arrow on the y axis. The black solid lines are the radii of the local maxima in the surface tangential wind. Taken from Wu, C.-C., Huang, Y.-H., Lien, G.-Y., 2012. Concentric eyewall formation in Typhoon Sinlaku (2008). Part I: Assimilation of T-PARC data based on the ensemble Kalman filter (EnKF). *Monthly Weather Review* 140, 506–527.

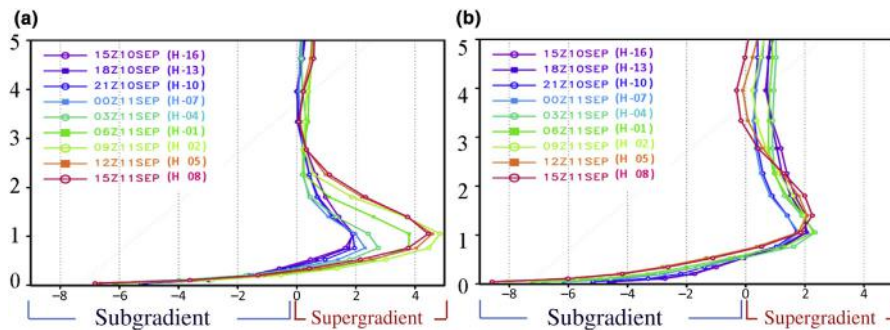


Figure 3 Azimuthally, area- and temporally averaged values over $(t - 3 \text{ h}, t + 3 \text{ h})$ for the gradient wind (unit: m s^{-1}). Analyses from 1500 UTC 10 Sep to 1500 UTC 11 Sep are displayed with a 3-h interval. The green line represents 1 h prior to SEF, while the light-green line represents 2 h after SEF. (a) SEF region ($r = 75 \sim 125 \text{ km}$) and (b) outside SEF region ($r = 125 \sim 180 \text{ km}$). Taken from Huang, Y.-H., Montgomery, M.T., Wu, C.-C., 2012. Concentric eyewall formation in Typhoon Sinlaku (2008). Part II: Axisymmetric dynamical processes. *Journal of Atmospheric Sciences* 69, 662–674.

within and just above the boundary layer preceding SEF and the corresponding dynamical pathway to SEF found in their simulation showed supporting results for the presented paradigm. Meanwhile, Wang and colleagues investigated the depth-integrated boundary layer flow, and demonstrated results that generally agree with the SEF pathway proposed in this paradigm.

The presented progressive boundary layer control on SEF also implies that the boundary layer scheme and its coupling to the above atmosphere need to be adequately represented in numerical models to improve the understanding of SEF, as well

as the accuracy of SEF forecasts, including the timing and preferred radial intervals.

Balanced Response to Diabatic Heating in a Region of Enhanced Inertial Stability

As discussed in the previous subsection, responses of transverse circulation to a heating/momentum source/sink framed in the thermal wind balance relationship known as Sawyer–Eliassen equation, has been applied to understand the evolution of the mean swirling circulation in idealized vortices in many extant

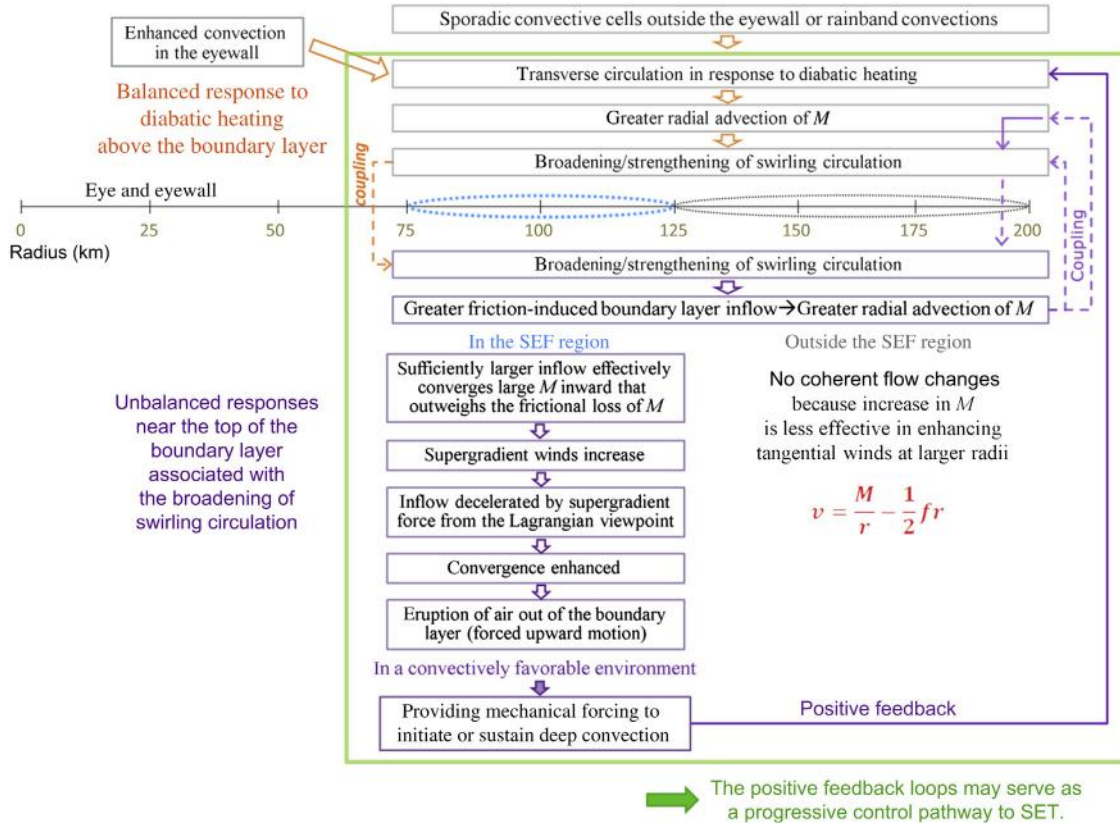


Figure 4 The schematic diagram describing the SEF paradigm introduced in Section [Unbalanced Boundary Layer Dynamics near the Top of TC Boundary Layer](#). Orange arrows refer the balanced response to diabatic heating, which is constrained by the thermal wind relationship. Purple arrows indicate the unbalanced dynamics, characterized by the development of supergradient wind near the top of boundary layer and a chain of corresponding structure changes. Dashed lines stand for the coupling relationship between the boundary layer and the free atmosphere aloft. The positive feedback loop enclosed by the green box is shown as a progressive control pathway to SEF.

studies. Using an idealized, cloud-resolving WRF simulation, Rozoff and colleagues revisited the balanced dynamics for SEF from the axisymmetric aspects. An expansion of kinetic energy (or enhanced inertial stability) is found prior to the SEF in their WRF simulation, a feature consistent with the presence of beta-skirt structure and the expansion of tangential winds in previous studies. The impact of this enhanced kinetic energy and diabatic heating on SEF was further investigated using an axisymmetric linearized, nonhydrostatic model (a balanced vortex model quite similar to the Sawyer–Eliassen model). Given the axisymmetric tangential wind and temperature profiles from the WRF model output, this simple model depicts how the transverse circulation responds to diabatic heating and surface friction prescribed also from WRF. The diagnosed results share similarity with the mean vortex structure in the WRF simulation in a number of ways, and suggest that the sustained diabatic heating along with the broadening wind outside the primary eyewall contributes the most to the enhancement of tangential winds in the SEF region.

Concluding Remarks

This article revisits favorable conditions and mechanisms that have been suggested for SEF in the existing studies.

Regarding the environmental control of SEF, model initial relative humidity has been identified critical for the increase of storm size and the subsequent SEF in recent studies using sophisticated numerical models. On the other hand, a variety of different approaches have been proposed as the internal mechanisms of SEF, including (1) axisymmetrization of prescribed/present outer vorticity patches, (2) the accumulation of eddy kinetic energy associated with VRWs near their stagnation radii, (3) the energy cascade process over the beta skirt of the TC vortex in a convectively favorable condition and subsequent positive feedback provided by WISHE (BSA hypothesis), (4) unbalanced response (i.e., the generation of supergradient wind, and its impact on the transverse circulation) to the expanding winds, and (5) balanced response of transverse circulation to diabatic heating over the area with enhanced inertial stability. Particularly noteworthy is that the broadening tangential wind (the beta-skirt structure and enhanced kinetic energy basically indicate a similar structure as well) furnishing the pathway to SEF is a vital process among different internal dynamical interpretations of SEF. Various conditions and mechanisms have been suggested for the establishment of such a skirted vortex structure, including higher environmental relative humidity or diabatic heating associated with outer rainbands, the initial vortex size/shape, concurrent

storm intensity, convective heating in an intensifying storm, and radially outward propagating VRWs.

Although recent advances of the unbalanced and balanced dynamics in TCs with double eyewalls appear promising in interpreting SEF from the axisymmetric perspective, the quantitative impacts of these aspects on SEF and their mutual feedback remain to be further investigated. While the symmetric dynamics has been shown to play a critical role in SEF, how the asymmetric components (e.g., spiral rainbands and sporadic convective cells in the vortex's outer core region) influence SEF also needs further studies to gain more comprehensive understanding. While the environmental conditions are relatively well understood and better presented in the observations and numerical models, discoveries are also being made on more uncertainties associated with the evolution of a TC vortex and the accompanied convective-scale features. More TC observations and further investigation on the internal vortex dynamics are thus required to better present the corresponding physical processes (e.g., microphysics, boundary layer dynamics, etc.) in numerical simulations of SEF, as well as the whole TC life cycle.

See also: Mesoscale Meteorology: Severe Storms. **Satellites and Satellite Remote Sensing:** Precipitation. **Tropical Cyclones and Hurricanes:** Hurricanes: Observation.

Further Reading

- Corbosiero, K.L., Abarca, S., Montgomery, M.T., 2012. Vortex Rossby waves and secondary eyewall formation in a high-resolution simulation of Hurricane Katrina (2005). *Proceedings of 30th Conference on Hurricanes and Tropical Meteorology*. American Meteorological Society, Jacksonville, FL. 1A.6.
- Hill, K.A., Lackmann, G.M., 2009. Influence of environmental humidity on tropical cyclone size. *Monthly Weather Review* 137, 3294–3315.
- Houze Jr., R.A., Chen, S.S., Smull, B.F., Lee, W.-C., Bell, M.M., 2007. Hurricane intensity and eyewall replacement. *Science* 315, 1235–1239.
- Huang, Y.-H., Montgomery, M.T., Wu, C.-C., 2012. Concentric eyewall formation in Typhoon Sinlaku (2008). Part II: Axisymmetric dynamical processes. *Journal of Atmospheric Sciences* 69, 662–674.
- Judt, F., Chen, S.S., 2010. Convectively generated potential vorticity in rainbands and formation of the secondary eyewall in Hurricane Katrina of 2005. *Journal of the Atmospheric Sciences* 67, 3581–3599.
- Kuo, H.-C., Schubert, W.H., Tsai, C.-L., Kuo, Y.-F., 2008. Vortex interaction and barotropic aspects of concentric eyewall formation. *Monthly Weather Review* 137, 5182–5198.
- Montgomery, M.T., Kallenbach, R.J., 1997. A theory for vortex Rossby waves and its application to spiral bands and intensity changes in hurricanes. *Quarterly Journal of the Royal Meteorological Society* 123, 435–465.
- Moon, Y., Nolan, D.S., Iskandarani, M., 2010. On the use of two-dimensional incompressible flow to study secondary eyewall formation in tropical cyclones. *Journal of Atmospheric Sciences* 67, 3765–3773.
- Nong, S., Emanuel, K.A., 2003. A numerical study of the genesis of concentric eyewalls in hurricane. *Quarterly Journal of the Royal Meteorological Society* 129, 3323–3338.
- Qiu, X., Tan, Z.-M., 2013. The roles of asymmetric inflow forcing induced by outer rainbands in tropical cyclone secondary eyewall formation. *Journal of Atmospheric Sciences* 70, 953–974.
- Rozoff, C.M., Nolan, D.S., Kossin, J.P., Zhang, F., Fang, J., 2012. The roles of an expanding wind field and inertial stability in tropical cyclone secondary eyewall formation. *Journal of Atmospheric Sciences* 69, 2621–2643.
- Tervey, W.D., Montgomery, M.T., 2008. Secondary eyewall formation in two idealized, full-physics modeled hurricanes. *Journal of Geophysical Research* 113, D12112.
- Wang, Y., 2009. How do outer spiral rainbands affect tropical cyclone structure and intensity? *Journal of Atmospheric Sciences* 66, 1250–1273.
- Wang, X., Ma, Y., Davidson, N.E., 2013. Secondary eyewall formation and eyewall replacement cycles in a simulated hurricane: effect of the net radial force in the hurricane boundary layer. *Journal of Atmospheric Sciences* 70, 1317–1341.
- Wu, C.-C., Huang, Y.-H., Lien, G.-Y., 2012. Concentric eyewall formation in Typhoon Sinlaku (2008). Part I: Assimilation of T-PARC data based on the ensemble Kalman filter (EnKF). *Monthly Weather Review* 140, 506–527.

TROPICAL METEOROLOGY AND CLIMATE

Contents

- El Niño and the Southern Oscillation: Observation**
- El Niño and the Southern Oscillation: Theory**
- Equatorial Waves**
- Hadley Circulation**
- Intertropical Convergence Zone**
- Intraseasonal Oscillation (Madden–Julian Oscillation)**
- Madden–Julian Oscillation: Skeleton and Conceptual Models**
- Monsoon: Overview**
- Monsoon: Dynamical Theory**
- Monsoon: ENSO–Monsoon Interactions**
- Tropical Climates**
- Walker Circulation**

El Niño and the Southern Oscillation: Observation

N Nicholls, Bureau of Meteorology Research Centre, Melbourne, VIC, Australia

© 2003 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 2, pp 713–719, © 2003, Elsevier Ltd.

Introduction

During 1877 and 1878 much of China was struck by famine, due to a severe drought. More than nine million people perished. In India, at the same time, more than eight million deaths were attributed to famine also caused by a drought. In many districts, a quarter of the population died. Drought in the same period also caused crop failures, scarcity of food or even famine, in north-eastern Brazil, Egypt, Indonesia, Fiji, Australia, and southern Africa. In other parts of the world, including Ceylon, the Pacific coast of South America, and Tahiti, many lives were lost from unusual storms or extended periods of heavy rain and flood. The El Niño Southern Oscillation, a major pattern of climate variation, links these climatic extremes in different parts of the world; the first major El Niño event for which good records exist was the 1877 event. Subsequent El Niño events have often reproduced the pattern of climate extremes and societal impacts observed in the 1877 event. We now use the phenomenon to make predictions of seasonal climate variations in many parts of the world.

How Was the El Niño Southern Oscillation Discovered?

The famine in India in 1877 led to the first scientific attempts to understand and predict monsoon failures and drought, and eventually to the mapping of the El Niño Southern Oscillation.

Henry Blanford, then the head of the India Meteorological Department, noticed that atmospheric pressures were higher than usual over India during the drought. He advised meteorologists in other parts of the British Empire of this and asked them about atmospheric pressures in their colonies.

Blanford's message reached the South Australian Government Astronomer and Meteorologist, Charles Todd, who noticed that Australian atmospheric pressures were also high, and that the country had been experiencing a drought at the same time as India. When another drought struck Australia in 1888, Todd realised that India and Australia often experienced drought at the same time. This synchronism of drought in the two countries is part of the suite of long-range connections (teleconnections) between climate fluctuations in different parts of the world that we now call the Southern Oscillation. For the next few decades, several meteorologists around the world were occupied in mapping these teleconnections into a coherent pattern. Sir Gilbert Walker was the most prominent among these mappers, and it was he who named the teleconnection patterns the Southern Oscillation. Walker used these teleconnections to develop statistical systems for forecasting climate anomalies in many parts of the world.

In the middle of the twentieth century, interest in the Southern Oscillation declined. This was partly because the focus of atmospheric scientists shifted to shorter time scales, as computer models exhibited their ability to forecast weather. A second reason for the decline in interest was the absence of

any theory explaining the teleconnections or the long time scale of the phenomenon. In the early 1960s, Hendrik Berlage and Jacob Bjerknes separately demonstrated that the El Niño and the Southern Oscillation were related. The term El Niño originally (at the end of the nineteenth century) referred to the annual weak warm current that runs southward along the coast of Peru and Ecuador at the end of the year. Subsequently, scientists applied the term to denote the occasional large warmings that occur every few years and result in major disruptions to the region. Bjerknes developed a theory for how this essentially tropical phenomenon could affect climate at higher latitudes. This step, along with the severe ecological and human consequences of the major El Niño episodes of 1972 and 1982, revived scientific interest in the study of interannual climate variations and their prediction. The phenomena are now jointly referred to as the El Niño Southern Oscillation, reflecting their close relationship.

What Causes the El Niño Southern Oscillation?

The joint name El Niño Southern Oscillation, is appropriate because ocean–atmosphere interaction is the cause of the phenomenon. Easterly winds over the eastern and central equatorial Pacific cause oceanic ‘upwelling’ (cooler subsurface waters being lifted to the surface) along the Equator. Southerly winds in the eastern Pacific also cause upwelling along the South American coast. As a result, the Pacific Ocean is usually cooler in the east than in the west by several degrees.

At tropical latitudes, heavy rains accompany warm oceans, so the warm west Pacific (including Indonesia and New Guinea) is a heavy rainfall region, while the cooler east Pacific receives little rainfall. [Figure 1](#) shows the mean sea surface

temperatures for December. The relative coolness of the east Pacific, compared to the west equatorial Pacific is evident.

This is the ‘average’ situation, but during an El Niño the ocean temperature gradient from one side of the Pacific to the other weakens, and the easterly winds weaken. Droughts occur in the west (around the Indian Ocean and the west Pacific) associated with cooler than normal ocean temperatures, while the unusually warm waters in the east bring heavy rains and floods to the normally arid Pacific coast of South America. [Figures 2 and 3](#) show the strong warming of the east equatorial Pacific that took place during the 1997/98 El Niño. [Figure 2](#) shows the sea surface temperatures during December 1997, at the peak of the El Niño. Warming in the east Pacific at that time had almost completely removed the east–west temperature gradient. The east equatorial Pacific warming of about 5°C is shown in [Figure 3](#), which exhibits the anomalies (deviations from climatology).

How does an El Niño start? A small change in the usual sea surface temperature pattern can produce a change in the winds along the Equator. In turn, these wind changes affect the currents that change the pattern of sea surface temperatures even more. This process continues, with ocean temperatures affecting winds that affect currents that, in turn, affect ocean temperatures. One important change is related to bursts of westerly winds in the western Pacific. These can trigger eastward-moving ocean disturbances that cause the thermocline (the transition layer between warm surface water and cooler, lower waters) to deepen in the east Pacific. This means that it is harder for the upwelling to cool the surface (because the upwelled water is now coming from the upper, warmer layer), so the east Pacific warms. Eventually, in the biggest El Niño events, the difference in temperature between the west and east equatorial Pacific Ocean can disappear altogether. As a result of

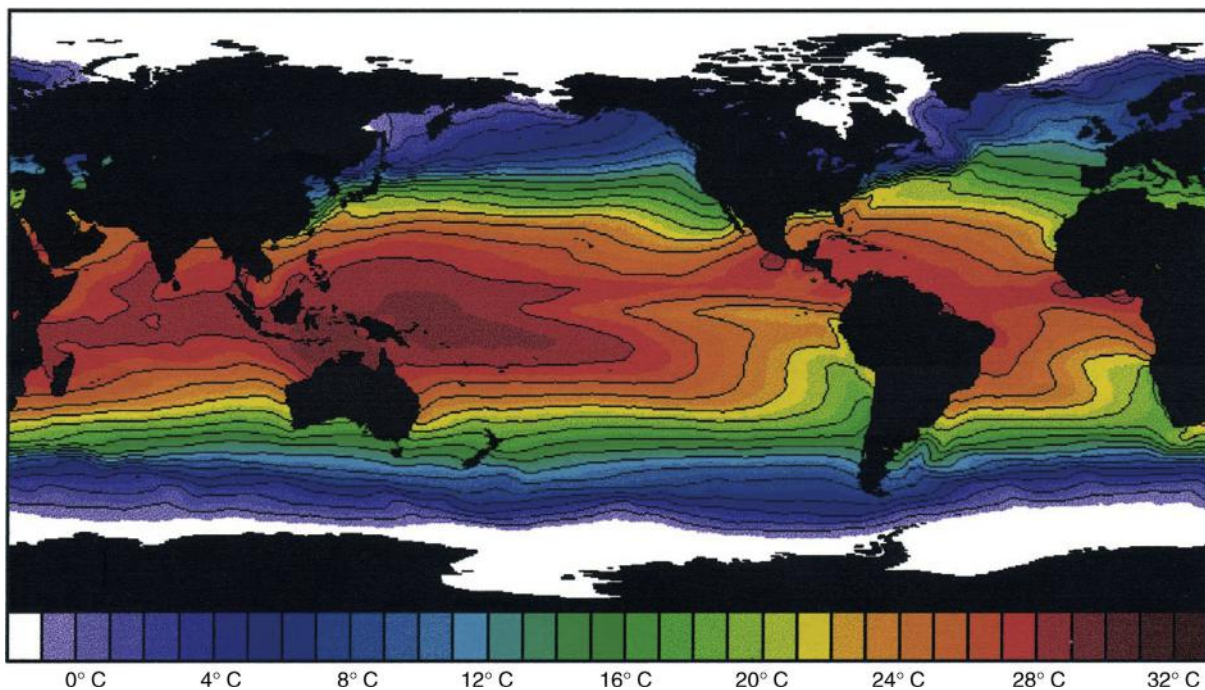


Figure 1 Climatological sea surface temperature for December. Analysis from Bureau of Meteorology, Australia.

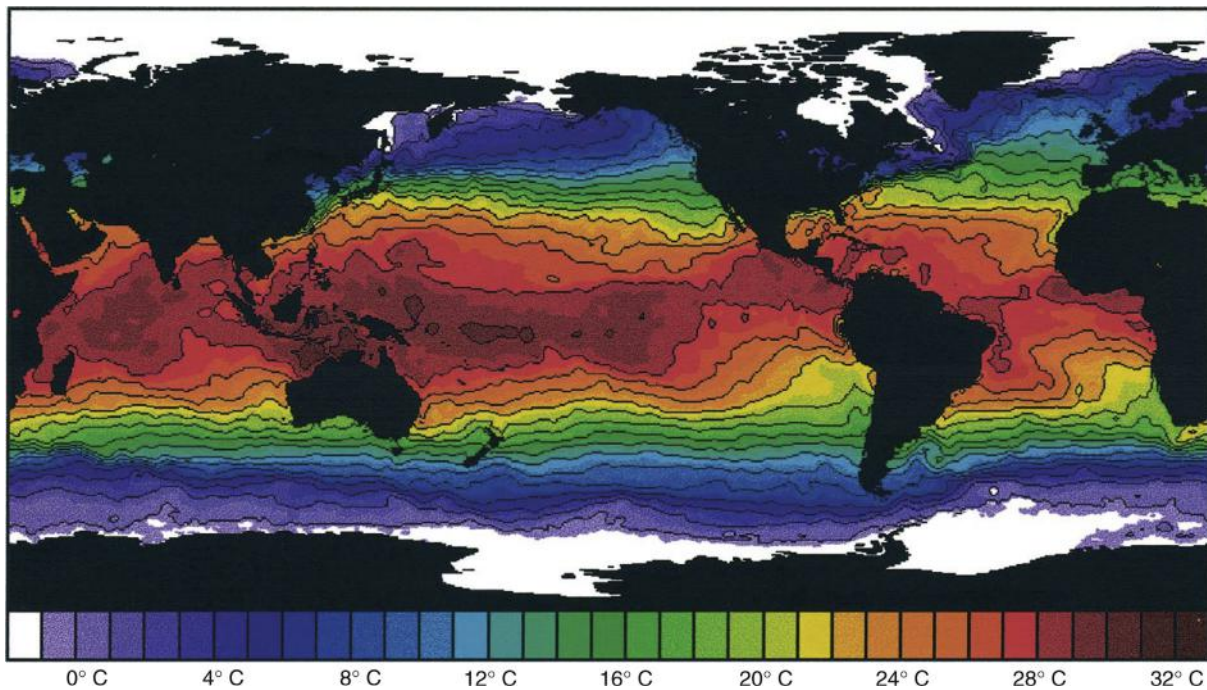


Figure 2 Sea surface temperature for 15–21 December 1997. Analysis from Bureau of Meteorology, Australia.

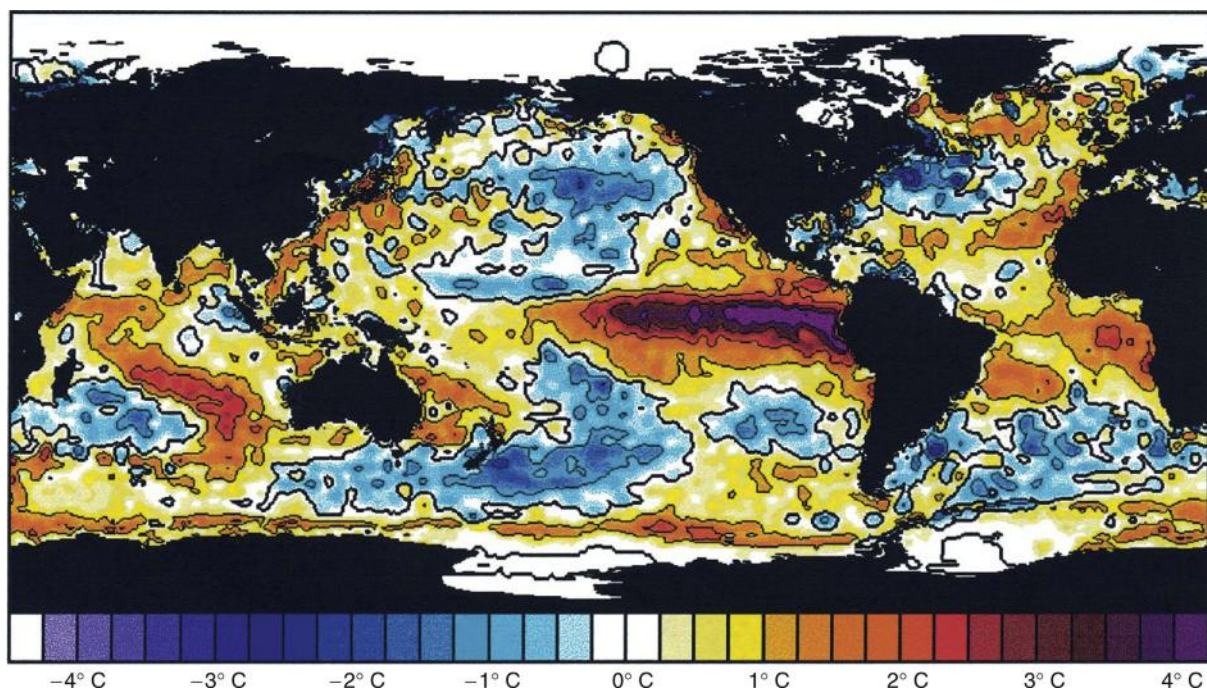


Figure 3 Sea surface temperature anomalies (deviations from climatology) for 15–21 December 1997. Analysis from Bureau of Meteorology, Australia.

these major changes in sea surface temperature and the surface winds, the whole pattern of climate and atmospheric circulation across the Pacific and Indian Oceans, and the surrounding continents, is disrupted, with droughts in normally wet areas and heavy rains over normally arid regions.

The changes associated with the El Niño often persist for about a year and then usually collapse quite quickly.

Sometimes a mirror-image pattern of climate disturbances, with flooding in Australia, India, Indonesia, northeast Brazil, and dry conditions on the Pacific coast of South America, follows. This set of conditions is called La Niña. La Niña episodes also usually last about a year or so.

As alluded to earlier, the atmospheric variations associated with El Niño and La Niña events are called the Southern

Oscillation. This name derives from the observation (dating back to the time of Blanford and Todd) that, during an El Niño, atmospheric pressures are usually higher than normal over Australian and the Indian Ocean and lower than normal in the southeast Pacific. During the opposite phase, the La Niña, the pressure anomalies are reversed. So, in a sense, the atmosphere acts like a seesaw, with high or low pressures on either side of the Pacific.

We can monitor this seesaw in atmospheric pressure with the Southern Oscillation Index or SOI. This is the standardized difference in pressure between Tahiti and Darwin. When the SOI is negative, pressures are high over the Australian region and relatively low in the southeast Pacific. This is an indication that the Trade Winds are weak across the Pacific, and these weaker winds result in warm east equatorial Pacific sea surface temperatures – an El Niño. Figure 4 shows time-series of the Darwin mean sea level pressure and sea surface temperatures in the ‘cold tongue’ of the east equatorial Pacific (180°–90°W, 6° N–6°S). The close relationship between the atmospheric pressure on one side of the Pacific and sea surface temperatures on the other side is clear, as is the tendency for El Niño and La Niña events to last about 12 months.

This tendency to last about 12 months means that the climate effects related to the El Niño Southern Oscillation are strongly persistent and thus predictable. This persistence is

greater during the second half of the calendar year, because El Niño episodes tend to start around March–May and finish around the same time a year later. Thus, if an event is under way by midyear it is likely to persist through the second half of the year. This means that climate anomalies usually associated with the presence of an El Niño at this time can often be predicted well in advance. The tendency for El Niño events to start around March–May is illustrated in Figure 5, which shows east equatorial Pacific sea surface temperature anomalies during the major events of the second half of the twentieth century. In each of the five events, sea surface temperature anomalies in the east equatorial Pacific were relatively low at the start of the year, and then increased rapidly from about March, reaching a peak near the end of the calendar year. The temperature anomalies subsequently weakened over the next few months.

What Areas Does the El Niño Southern Oscillation Affect?

The pattern of climate anomalies seen in the 1877 El Niño tends to be repeated each time an El Niño occurs. The typical pattern of rainfall anomalies associated with an El Niño is shown in Figure 6. The figure indicates, for each area

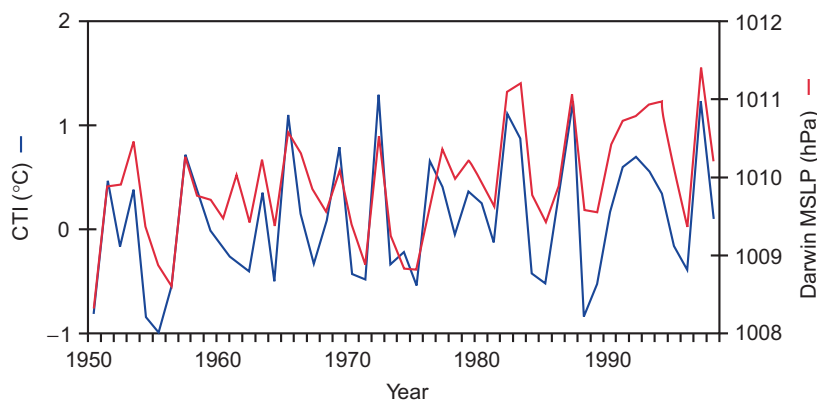


Figure 4 Time series of annual mean Darwin mean sea level pressure (MSLP) and sea surface temperature in the area 180°–90° W, 6° N–6° S ('Cold Tongue Index', CTI). Darwin data from Bureau of Meteorology, Australia. CTI data from Todd Mitchell, JISAO, University of Washington.

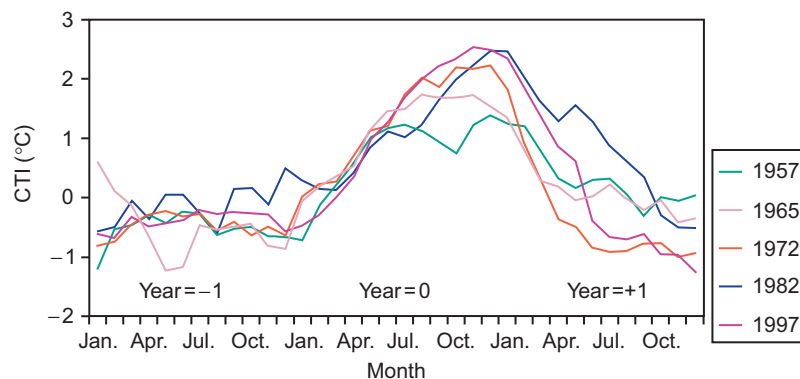


Figure 5 Time-series of monthly CTI during the major El Niño events of the second half of the twentieth century (1957, 1965, 1972, 1982, 1997). CTI data from Todd Mitchell, JISAO, University of Washington.

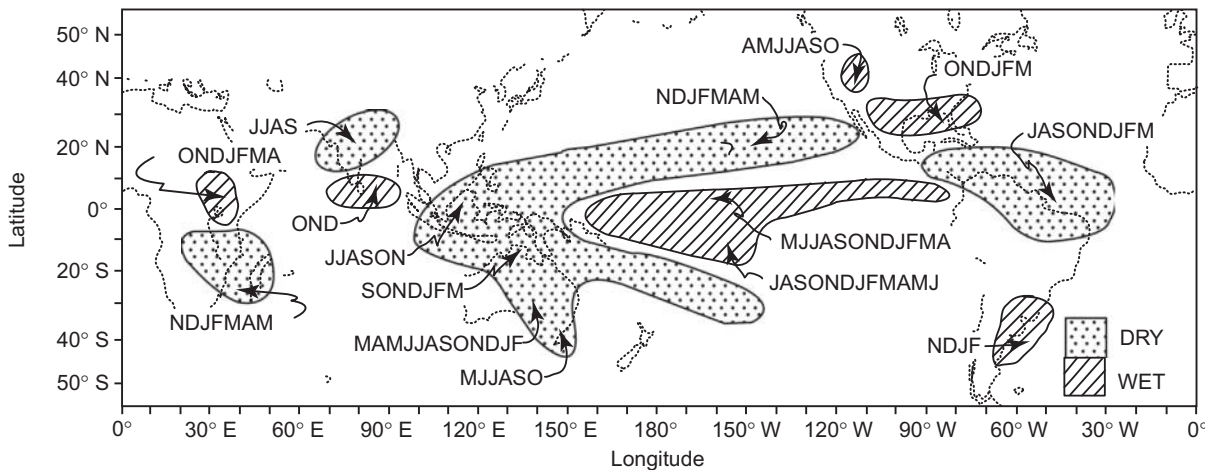


Figure 6 Schematic of areas with a consistent precipitation signal associated with El Niño events. For each region the months are shown during which it is consistently wetter or drier than normal. In each region the list of months begins in the initial year of the El Niño (year=0). Reprinted with permission from Cambridge University Press of Trenberth, K.E., 1991. in Glantz, et al., 1991.

consistently affected by the El Niño, the months in which the anomalies are most consistent. The pattern of precipitation anomalies associated with the other extreme of the El Niño Southern Oscillation, the La Niña, is essentially the opposite of that depicted in Figure 6 (i.e., where drier than normal conditions are usually experienced during an El Niño, then wetter than normal conditions can be anticipated during a La Niña episode).

The El Niño Southern Oscillation also affects temperatures in some parts of the world. Thus in December–February at the peak of an El Niño, temperatures are usually above average throughout central and southern Africa, southern Asia, and the western Pacific, Canada, and the Pacific coasts of North and South America. The south-east of the United States tends to be cooler than average. Severe frosts can occur in places where drought accompanies an El Niño episode, such as the highlands of Papua New Guinea and inland eastern Australia.

The El Niño Southern Oscillation also affects tropical cyclones and some other weather and climate extremes. Figure 7 is a time-series of the SOI and of the number of tropical cyclones around Australia. When an El Niño is under way (i.e., when the SOI is strongly negative), fewer than normal tropical cyclones are observed around Australia. Similarly, Atlantic hurricane activity is reduced during El Niño episodes. On the other hand, tropical cyclones are more frequent than usual in the east Pacific during these episodes.

Prehistoric Behavior of the El Niño Southern Oscillation

Instrumental records relevant to the study of the El Niño Southern Oscillation are available back into the late nineteenth century. The study of El Niño episodes prior to this depends on

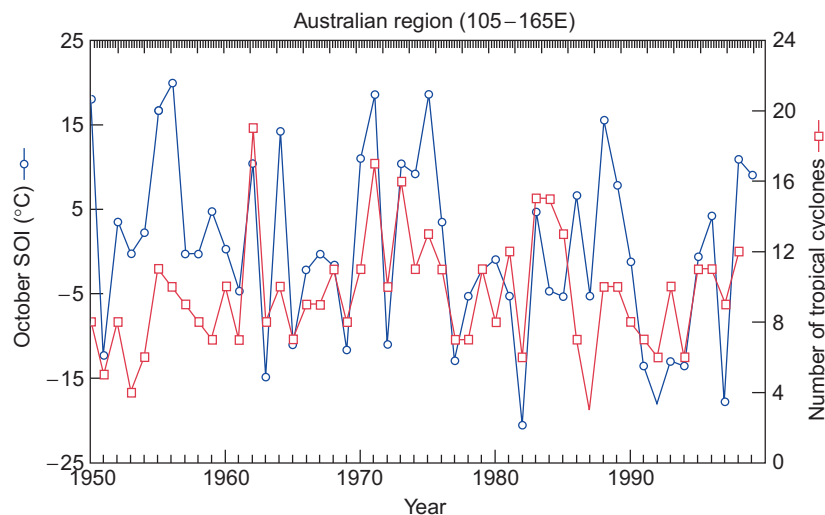


Figure 7 Time-series of the Southern Oscillation Index (SOI) (○) and the number of tropical cyclones in the Australian region (0°–15° S, 105°–165° E) (□). Data from the Bureau of Meteorology, Melbourne, Australia.

documentary records, and paleoclimatic (proxy) records. Documentary evidence of heavy rains and floods on the Pacific coast of South America (always associated with El Niño episodes during the instrumental era) are available from the sixteenth century. Comparisons of the dates of heavy rains and floods in South America with dates of droughts in other parts of the world have confirmed that the El Niño Southern Oscillation has been operating for at least hundreds of years. The teleconnections between droughts and floods in these various parts of the world have been similar throughout these five centuries, reflecting the effects of the El Niño Southern Oscillation throughout this period.

Paleoclimatic data, from corals, ice cores in glaciers, tree rings, and marine and lacustrine sediments also provide information regarding the occurrence of El Niño episodes prior to the instrumental period. This evidence, although not conclusive, suggests that El Niño episodes have been occurring for at least several thousand years.

The El Niño Southern Oscillation in the Recent Past

The prominence of the El Niño Southern Oscillation has varied through the instrumental period. Very strong El Niño episodes occurred in the first quarter of the twentieth century, with only relatively infrequent, and weak, events in the period 1925–1950. After 1950, more intense El Niño and La Niña events were observed. Since the mid-1970s, there appears to have been a shift toward more frequent, or stronger, El Niño episodes, with La Niña episodes becoming relatively infrequent. Some analyses suggest that this behavior is very unusual, given the (admittedly short) historical record.

Future Observations of the El Niño Southern Oscillation

For most of the period during which the El Niño Southern Oscillation has been monitored and studied, observations

originally intended for other purposes have been the main source of information. Atmospheric pressure, rainfall, and temperature observations originally taken for the purposes of weather recording and forecasting, or to determine the ‘average’ climate, have been used in studies of how the phenomenon affects climate variations around the globe. Sea surface temperatures recorded by merchant and other ships have been the main source of information about the ocean variations associated with the El Niño. In recent decades, however, new and improved observations, specifically designed for climate studies, have been initiated. These include satellite observations of rainfall and sea surface temperature and sea-level, moored buoys monitoring the ocean and atmosphere in critical parts of the ocean, and subsurface analyses of the ocean thermal structure. The analysis of these new data is in its infancy, but the data have already enhanced our ability to monitor, understand, and predict the El Niño Southern Oscillation.

See also: Tropical Meteorology and Climate: El Niño and the Southern Oscillation: Theory; Monsoon: ENSO–Monsoon Interactions; Walker Circulation.

Further Reading

- Allan, R., Lindesay, J., Parker, D., 1996. *El Niño Southern Oscillation and Climatic Variability*. CSIRO Publishing, Collingwood.
- Diaz, H.F., Markgraf, V. (Eds.), 1992. *El Niño. Historical and Paleoclimatic Aspects of the Southern Oscillation*. Cambridge University Press, Cambridge.
- Glantz, M.H., 1996. *Currents of Change. El Niño's Impact on Climate and Society*. Cambridge University Press, Cambridge.
- Glantz, M.H., Katz, R.W., Nicholls, N. (Eds.), 1991. *Teleconnections Linking Worldwide Climate Anomalies. Scientific Basis and Societal Impact*. Cambridge University Press, Cambridge.
- Philander, S.G.H., 1990. *El Niño, La Niña, and the Southern Oscillation*. Academic Press, New York.

El Niño and the Southern Oscillation: Theory

P Chang, Texas A&M University, College Station, TX, USA

SE Zebiak, International Research Institute for Climate Prediction, Palisades, NY, USA

© 2003 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 2, pp 719–724, © 2003, Elsevier Ltd.

Introduction

The El Niño Southern Oscillation (ENSO) is a spectacular, planetary-scale climate phenomenon that is inherently caused by interactions between the atmosphere and the ocean. Historically, El Niño refers to unusually warm ocean temperatures that occur every 2–7 years around Christmas time along Peruvian coast, extending into equatorial eastern and central Pacific Ocean. The Southern Oscillation, named by its discoverer – Sir Gilbert Walker – on the other hand, refers to a ‘seesaw’ of the atmospheric pressure between the Pacific and Indian Oceans. It was not until the seminal work of Jacob Bjerknes in the late 1960s that scientists realized that these two phenomena are intimately linked. The acronym ENSO has now been widely used to describe this fascinating interannual climate fluctuation, emphasizing the inherent ocean–atmosphere coupling.

Although the origins of ENSO lie in the tropical Pacific, the impact of ENSO is global, owing to planetary waves of the atmosphere that redistribute vorticity from tropics to extra-tropics. The ‘teleconnection’ of ENSO can disrupt weather patterns around the globe. For this reason, ENSO has been recognized as the most important climate phenomenon at interannual timescales.

Theoretical understanding of the development and evolution of ENSO, and of underlying dynamical mechanisms for its irregular oscillation at interannual timescales, goes beyond the boundary of traditional dynamical meteorology and oceanography, because it requires knowledge about how the tropical atmosphere responds to sea surface temperature changes; how the equatorial ocean adjusts to changes in winds; and how various feedback loops between the atmosphere and ocean operate and interplay. This understanding provides the theoretical basis for the development of ENSO prediction systems, which are critical for operational seasonal-to-interannual climate forecasting.

The Southern Oscillation and Walker Circulation

From an atmospheric perspective, the Southern Oscillation can be viewed as a perturbation about a thermally driven east-to-west circulation of the tropical atmosphere across the Pacific Ocean. This circulation, known as the Walker Circulation, is caused by the sharp contrast in sea surface temperature across the tropical Pacific Ocean. The western tropical Pacific contains the warmest regions of the world’s ocean, known as the Western Pacific Warm Pool, where the sea surface temperature is above 28°C. In contrast, the eastern equatorial Pacific features relatively cold ocean surface waters, extending from

South America coasts westward along the Equator. This is known as the Eastern Equatorial Pacific Cold Tongue, where the sea surface temperature is 5–10°C colder than the surface water of the warm pool. The warm water in the western Pacific creates low surface pressure, which causes moisture-laden air to converge into the region. The air rises and the moisture condenses in strong convective events, resulting in widespread cloudiness and heavy precipitation. The rising air descends from the upper troposphere to the surface in the Eastern Equatorial Pacific Cold Tongue as dry air. Cool temperatures result in relatively high surface pressure, divergent flow, and little rainfall. These motions – rising in the west, sinking in the east – are connected through easterly Trade Winds near the surface and a westerly wind aloft, forming the Walker Circulation.

Fluctuations in the position and intensity of the Walker Circulation cause the Southern Oscillation. When the sea surface temperature in the eastern Pacific is warmer than normal, such as during El Niño years, the low atmospheric pressure center normally situated in the Western Pacific Warm Pool moves eastward, bringing along with it the rising moist air and heavy precipitation. As a result, the east–west pressure difference across the Pacific is reduced and the easterly Trade Winds are weakened. This produces a weak Southern Oscillation (a negative phase). By the same token, the strength of the Southern Oscillation is enhanced (a positive phase) when the sea surface temperature in the eastern Pacific falls below normal. This sensitivity of the tropical atmosphere circulation to sea surface temperature fluctuations is one of the key elements of ENSO physics.

A secondary process that contributes to equatorial trade wind fluctuations involves the so-called Hadley circulation, a meridional overturning cell spanning the tropical Pacific (and the global tropics). This circulation consists of a rising branch, concentrated in a narrow zone generally north of the Equator, known as the intertropical convergence zone (ITCZ), and sinking motion, with increasingly strong surface easterlies both north and south of the ITCZ. During El Niño, as the equatorial surface temperature warms, the ITCZ and attendant Hadley circulation shift equatorward, leading to a reduction of equatorial easterly winds, beyond that associated with the Walker Circulation.

Many of the essential features of the Walker Circulation (and Hadley circulation) can be captured by a simple physical model in which the tropical atmosphere is assumed to be forced by a diabatic heating source and subject to simple dissipation of momentum and heat (with common decay rate r of order of 1–2 days). The diabatic heating is largely induced by latent heat released by the rising moist air over the warm ocean, and this can be approximated as a function of sea surface

temperature. Assuming that the vertical structure of the diabatic heating is fixed and has a simple structure with a single maximum at midlevels decreasing to near zero at the surface and upper levels (which approximates to the heating field produced by cumulus convection in the tropics), then the entire circulation pattern in the vertical projects primarily onto the so-called first baroclinic mode of the tropical atmosphere, and the horizontal motions at each level obey a two-dimensional set of equations (the so-called shallow water equations). Therefore, the forced solution of the shallow water equation can give the first-order approximation of the Walker Circulation. In the region of heating, for example, in the Western Pacific Warm Pool, the diabatic heating gives rise to a vertical velocity which causes the lower atmosphere to expand and the vortex to stretch. To conserve potential vorticity, the surface air parcels must move poleward to generate cyclonic vorticity in order to balance the ambient vorticity change. The diabatic heating also excites planetary waves which induce remote response outside the region of heating. Of particular interest is a Kelvin wave propagating eastward at a speed c . It gives an easterly wind symmetric about the Equator which decays at a rate of r/c per unit distance. This easterly wind is reminiscent of the Trade Winds along the equatorial Pacific as a part of the Walker Circulation. To the west of the forcing region, there are Rossby waves which cause the cyclonic flow to expand westward. The simple atmospheric model gives an analytical expression linking changes in the equatorial Trade Winds to changes in sea surface temperature through its effect on the diabatic heating of the atmosphere.

Equatorial Ocean Adjustment

From an oceanic perspective, changes in sea surface temperature associated with El Niño can be understood in terms of an equatorial ocean response to changes in the Trade Winds. Under normal conditions, an easterly wind stress supplied by the Trade Winds in the central and eastern Pacific acts on the ocean surface. This stress is balanced by friction and the Coriolis force, resulting in poleward surface flow in either hemisphere and thus upwelling in the eastern equatorial Pacific and South American coast. The westward winds also 'push' the relatively warm surface water westward, bringing cold subsurface water to the ocean surface and lifting the so-called thermocline of the ocean in the east. The thermocline represents a band of water within which the temperature changes rapidly with depth, thereby separating the warm upper ocean from the cold deep ocean. The Trade Winds thus cause the thermocline to shoal from west to east across the equatorial Pacific Ocean, maintaining the warm pool in the west and the cold tongue in the east. During a negative phase of the Southern Oscillation, the easterly Trade Winds are weaker than normal, which reduces the upwelling in the east, deepens the thermocline, and causes the sea surface temperature to rise. An El Niño is produced! The opposite occurs during a positive phase of the Southern Oscillation, when the Trade Winds are strengthened.

The close relationship among Trade Winds, thermocline, and sea surface temperature results from the rapid adjustment of the equatorial ocean. Because the Coriolis force vanishes at the Equator, there is a wave guide along the Equator, where

a variety of waves are trapped to within a few degrees to either side of it in the ocean (and a few tens of degrees in the atmosphere). Two types of waves, the equatorial Kelvin and Rossby waves, are of particular importance. Kelvin waves are special gravity waves that propagate eastward with a speed of approximately $2\text{--}3\text{ m s}^{-1}$, and can travel across the Pacific Ocean in 2 months or so. Rossby waves are planetary vorticity waves that propagate westward at a rate of about $0.6\text{--}0.8\text{ m s}^{-1}$, and can travel across the Pacific in 6–7 months. Both these waves propagate at a rate that is faster by an order of magnitude or more than the planetary waves in extratropical oceans. For this reason, equatorial oceans adjust much more rapidly than extratropical oceans in response to changes in the wind stress forcing. Therefore, at interannual timescales, the zonal gradient of the equatorial thermocline and overlying Trade Wind stress are approximately at balance.

The response of the equatorial ocean to changes in the zonal wind stress can again be modeled in terms of a shallow water model. Here, the ocean is approximated as a two-layer fluid system with a thin, warm layer on top of a deep, cold layer. The interface between the two layers represents the ocean thermocline which has an average depth of approximately 150 m in the equatorial Pacific. Motions of upper layer and the interface obey shallow water equation subject to wind stress forcing. The steady-state solution is very nearly a balance between the zonal wind stress and zonal gradient of thermocline. When a change in the trade wind occurs, the ocean is subject to an anomalous zonal wind stress forcing. A typical wind stress anomaly associated with the Southern Oscillation has a spatial structure that has the largest amplitude in the western central equatorial Pacific and decays away from the Equator. Because of the nonuniform spatial structure, off-equatorial wind stress curl is generated. The oceanic response to such an anomalous wind stress forcing in the shallow water system takes place in two stages: first, an equatorial Kelvin wave is excited by the strong wind stress anomaly at the Equator, propagating eastward and causing changes in depth of the thermocline; and sea surface temperature in the eastern equatorial Pacific upon its arrival. At the time when the Kelvin wave is excited, Rossby waves of opposite sign to the Kelvin wave are also generated by the off-equatorial wind stress curl associated with the wind stress anomaly. These Rossby waves propagate westward and reflect at the western boundary as a second Kelvin wave – now with the opposite sign to the first Kelvin wave – and thus work against the effect brought by the first Kelvin wave in the eastern equatorial Pacific with a time delay. It is this delayed response in the eastern equatorial Pacific produced by the same wind stress forcing that provides the oceanic 'memory', another key element of ENSO physics.

Coupled Dynamics

If the atmosphere and ocean were decoupled, then small perturbations in either the sea surface temperature or the winds would fade quickly away because of dissipation in the oceans and atmosphere. In reality, the tropical atmosphere and equatorial ocean in the Pacific Ocean are tightly coupled, because of the sensitivity of the atmospheric response to sea surface temperature changes and the rapid adjustment of the

equatorial ocean to changes in winds. Therefore, a modest change in either the equatorial sea surface temperature or the Trade Winds can trigger a chain reaction in the coupled ocean–atmosphere system, involving a positive feedback between the atmosphere and the ocean.

Bjerknes Feedback Mechanism

The key ingredients of this positive feedback were first pointed out by Bjerknes. Consequently, the mechanism has become known as the Bjerknes hypothesis. Imaging that there is initially a weak westerly wind anomaly along the Equator that causes the Trade Winds to weaken. From the equatorial ocean adjustment discussed above, a weak warm sea surface temperature anomaly is expected to occur in the eastern equatorial Pacific, owing to deepening in the thermocline depth and weakening in equatorial upwelling. Because the tropical atmosphere is sensitive to changes in sea surface temperature, the small increase in sea surface temperature in the east will tend to move atmospheric convection eastward, reduce the diabatic heating in the west, and weaken the Walker Circulation. This causes a further weakening of the Trade Winds, which in turn leads to a further warming in the eastern equatorial Pacific, and so on. Key elements of this feedback loop are illustrated in Figure 1. The Bjerknes hypothesis marks the beginning of the formation of modern ENSO theory.

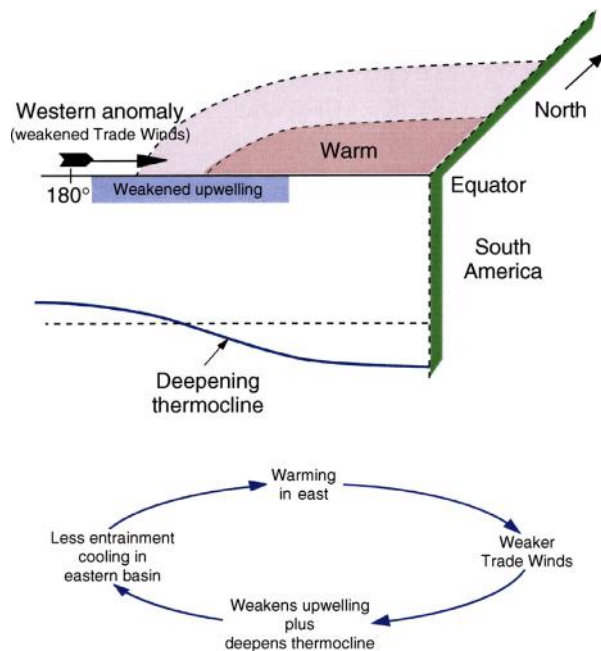


Figure 1 The Bjerknes hypothesis describes a positive feedback between the ocean and atmosphere in the equatorial Pacific: a weakening in the Trade Winds causes a warming in the eastern Pacific through a deepening in the thermocline and a weakening in the upwelling, which in turn leads to a further weakening in the Trade Winds. Reproduced from Chang, P., Battisti, D.S., 1998. The physics of El Niño. *Physics World* 11, 41–47.

Coupled Modes

A quantitative understanding of the coupled dynamics can be gained by formulating a simple coupled ocean–atmosphere model, such as the two shallow water systems described above. A stability analysis can then be applied to the coupled system to obtain a set of modes which describe how the coupling between the ocean and atmosphere can modify free oceanic and atmospheric wave modes in a wide range of dynamical parameter space. Such an analysis leads to the discovery of a new breed of ‘wave modes’, with characteristics that depend on the strength of the air–sea feedback and the adjustment timescale of the ocean. For a weak air–sea feedback, the adjustment time of the ocean is much faster than that imposed by the weak air–sea feedback. Therefore, the free waves are not significantly influenced by the air–sea feedback and the coupled modes behavior similar to those of free waves. For a strong air–sea feedback, the adjustment time of the ocean is much slower than that imposed by the strong air–sea feedback. The coupled modes depend critically on the air–sea feedback, but are less influenced by the oceanic adjustment. In this limit, the coupled modes behave quite differently from the free ocean waves and tend to be more unstable because of the strong influence of the positive air–sea feedback.

In between these two extreme limits lies the most interesting dynamic regime, where the adjustment time of the ocean and of air–sea feedback are comparable to each other. This is the dynamic regime, where ENSO is believed to reside in nature. The coupled modes in this regime have a mixed behavior of free oceanic waves and of the mode that depends primarily on air–sea feedbacks, and can be either stable or unstable depending on the strength of the feedback. These modes can be characterized best by a delayed action oscillator – a prototype model for ENSO. Let T denote sea surface temperature anomaly in the eastern equatorial Pacific, then the time evolution of T obeys a differential delay equation $dT/dt = cT(t) - bT(t - \tau)$, where t is time and τ is a time delay associated with the adjustment of the equatorial ocean. The behavior of T in this model is determined by two competing processes: (1) cT represents the Bjerknes positive feedback, contributing to a growth of sea surface temperature; and (2) $bT(t - \tau)$ represents the delayed oceanic adjustment discussed in the previous section, constituting a negative feedback. This negative feedback due to the oceanic ‘memory’ effect counteracts the positive feedback in the eastern equatorial Pacific with the time delay τ . Physically, this is achieved through excitation of the Rossby waves by off-equatorial wind stress curl and the reflection at the western boundary. Therefore, the time delay τ is essentially determined by the propagation of the Rossby waves from the western central Pacific to the western boundary and of the reflected Kelvin wave from the western boundary to the eastern equatorial Pacific. This negative feedback process is illustrated in Figure 2.

Because both the positive and negative feedbacks are comparable in strength, the differential delay equation can support oscillations with periods longer than the wave adjustment timescale τ . In fact, when realistic parameters are used, the differential delay equation produces an oscillation with a period

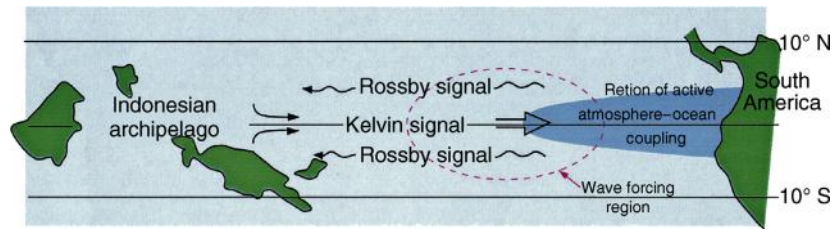


Figure 2 Oceanic adjustment associated with the delayed oscillator type of coupled ocean–atmosphere mode. Rossby waves generated by the relaxation of the Trade Winds in the western central equatorial Pacific propagate westward and are reflected into Kelvin waves at the western boundary near the Indonesian archipelago. The Kelvin waves reach the eastern equatorial Pacific with a time delay τ after the relaxation of the Trade Winds and bring cold water to the region, which ultimately shuts down the warm ENSO event. Together with the Bjerknes positive feedback, this negative oceanic feedback process forms the delayed oscillator mechanism for ENSO. Reproduced from Chang, P., Battisti, D.S., 1998. The physics of El Niño. *Physics World* 11, 41–47.

of 3–4 years, which agrees with the average occurrence of El Niño in reality. Therefore, the coupled modes of delayed oscillator type are believed to capture the essential physics of ENSO.

Nonlinear vs Stochastic ENSO Theory

The delayed oscillator theory offers an explanation for the oscillatory behavior of ENSO, but cannot explain the irregularity of ENSO. In nature, an ENSO occurs every 2–7 years with considerable variations in its strength. Where does this irregularity come from? Although answers to this question are not entirely clear, there are two competing theories that offer different views on this issue.

One theory relies on inherent nonlinear interactions within the coupled ocean–atmosphere system. It assumes that the Bjerknes positive feedback is strong enough for the coupled system to reside in an unstable dynamic regime and maintain a self-sustained oscillation at an interannual timescale. This self-sustained oscillation interacts nonlinearly with the annual cycle, which is driven by the seasonal variation of the solar radiation. Because the intrinsic mode of the coupled system oscillates at a different frequency from the driving frequency, nonlinear interaction between the two cycles can give rise to complicated behavior of the coupled system response. Depending upon the relative strength of the driving cycle (the annual cycle) and the intrinsic cycle (the ENSO cycle), the response can be either locked into a periodic cycle (with a period equal to a rational number between the period of annual cycle and that of intrinsic coupled mode; this phenomenon is known as frequency locking in nonlinear dynamics) or chaotic. It is hypothesized that the reality resides in the chaotic regime, and thus the irregularity of ENSO can be attributed partially to the chaos generated by the nonlinear interaction between the annual cycle and the intrinsic oscillating mode of the coupled system.

The other, competing theory puts ENSO in a weak feedback regime, so that the coupled system does not support a self-sustained oscillation, but is forced externally by ‘weather noise’. Here, ‘weather noise’ refers to the high frequency variability that is not generated directly by ocean–atmosphere interactions, but is produced by hydrodynamical instability processes of the atmosphere. Although these high frequency fluctuations of the atmosphere have coherent spatial structure, in time they can be represented approximately as a normally distributed white

noise process. Under these approximations, ENSO can be modeled as a multivariate linear stochastic system, i.e., $(d/dt)s = As + F\xi$, where s is a state vector comprised of sea surface temperature anomalies throughout the tropical Pacific basin; A , a system matrix governing the deterministic dynamics of ENSO; and F represents the spatial distribution of the weather noise whose temporal fluctuations ξ are represented as a normally distributed white noise processes. The sea surface temperature evolution is then determined by the properties of the system matrix A and the noise forcing structure matrix F . In particular, the leading eigenvectors of A give the dominant coupled modes in ENSO system, including the delayed oscillator type of modes. However, since the system is linear, all eigenvectors must be stable, i.e., decaying with time (otherwise the variance would not be bounded). Therefore, they cannot support a self-sustained oscillation and the variability of the system must be maintained by the noise forcing. One important distinction between this view of ENSO physics and the nonlinear theory is that the evolution of ENSO is not necessarily dominated by a single mode (the most unstable mode, according to the nonlinear theory), but rather determined by the interference among many stable modes. Constructive interferences cause sea surface temperature anomalies to grow, whereas destructive interferences cause them to decay. In this theory the irregularity of the ENSO cycle comes naturally because the variability of the system is maintained by a random forcing.

Understanding the cause of ENSO irregularity has important implications for the predictability of this phenomenon. If ENSO evolution is governed by a low-order chaos, then its predictability limit is determined by the inherent nonlinear dynamics of the coupled system. On the other hand, if stochastic processes in the atmosphere are the main cause of ENSO irregularity then its predictability depends not only on deterministic dynamics in the coupled system but also the nature of noise forcing, which is determined largely by the internal dynamics of the atmosphere. These remain topics of active research.

See also: Dynamical Meteorology: Coriolis Force; Kelvin Waves; Rossby Waves. **General Circulation of the Atmosphere:** Overview. **Numerical Models:** General Circulation Models. **Oceanographic Topics:** General Processes. **Tropical Meteorology and Climate:** El Niño and the Southern Oscillation: Observation; Hadley Circulation; Walker Circulation.

Further Reading

- Gill, A.E., 1982. Atmosphere-Ocean Dynamics. Academic Press, New York, NY.
- Philander, S.G.H., 1990. El Niño, La Niña and the Southern Oscillation. Academic Press, New York, NY.
- Neelin, J.D., Battisti, D.S., Hirst, A.C., et al., 1998. ENSO theory. Journal of Geophysical Research 103, 14261–14290.
- Chang, P., Battisti, D.S., 1998. The physics of El Niño. Physics World 11, 41–47.

This page intentionally left blank

Equatorial Waves

MC Wheeler and H Nguyen, Centre for Australian Weather and Climate Research, Bureau of Meteorology, Melbourne, VIC, Australia

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

Horizontal structures and dispersion relations of observed equatorial waves are analogous to solutions of the shallow water equations on an equatorial β -plane. In the past few decades, the significance of equatorial waves for tropical meteorology has become increasingly apparent. Equatorial waves exist in the atmosphere and ocean and cause perturbations in dynamical and thermodynamical variables large enough to be crucial for many aspects of tropical weather and climate variability. They form the basic building blocks of tropical circulations. Some aspects of their dynamics remain poorly understood.

Introduction

Equatorial waves serve as the fundamental building blocks of many aspects of tropical weather and climate. They are geophysical fluid waves, existing for a range of spatial and temporal scales, which are trapped near the Equator. They propagate in the zonal and vertical directions, and may exist at any altitude level in the atmosphere or depth in the ocean. They cause oscillations in the pressure, temperature, and winds or currents, with the magnitude of such oscillations being large enough to cause changes in the weather. Conversely, equatorial waves may be excited by energetic weather events, such as by the latent heating of the atmosphere by organized tropical convection. Because equatorial waves transmit energy in the longitudinal and vertical directions, they are a means by which a location in the tropical atmosphere (or ocean) may be influenced by a remote energetic disturbance. This influence can sometimes extend around the entire circumference of the Earth at the Equator.

Like many meteorologically important phenomena, equatorial waves owe their existence to the restoring forces of gravity, the pressure gradient force, and the apparent Coriolis force that applies to displacements of parcels of air or water. The change in sign of the Coriolis force at the Equator allows these waves to take on a discrete set of meridional patterns and causes their trapping near the Equator. Thus, although the Coriolis force is small near the Equator, and consequently geostrophic balance is no longer expected, the rotation of the planet still plays an essential role for these waves.

This article concentrates on equatorial waves that occur in the troposphere, that is, those that most readily influence the weather. The article begins with the theory needed to give a basic understanding of their existence and appearance. In many respects, the theory is equally applicable to equatorial waves existing in the ocean or middle atmosphere. The analysis continues with a simulation of equatorial wave dispersion in a numerical model of the atmosphere. The rest of the article is devoted to observed equatorial waves and their significance for the weather.

Theory

The theory of equatorial waves may start with the (hydrostatic) primitive equations, that is, the full set of basic equations that

govern the time evolution of the three-dimensional large-scale thermodynamic fields in the stratified atmosphere. These equations are often written as six equations in six unknowns: the zonal velocity u , meridional velocity v , vertical velocity w , density ρ , pressure p , and temperature T .

Mathematically, it can be shown that when linearized about a basic state with no (or vertically constant) mean flow, the primitive equations may be reduced, or separated, into the 'shallow water' equations and a vertical structure equation. The shallow water equations govern the horizontal and time-varying behavior of each 'normal mode' of the atmosphere, and are written as three equations in three unknowns only. The vertical structure equation, on the other hand, may be solved to give the vertical structures of each normal mode, as well as the so-called 'equivalent depth' of each of the modes. The equivalent depth also appears in the shallow water equations, and is the depth of the shallow layer of fluid required to give the correct horizontal and time-varying structure of each mode. It provides the link between the vertical structure equation and the shallow water equations, but it also provides an important link between the observed waves and the shallow water theory. Mathematically, the equivalent depth is the separation constant.

This separation of the primitive equations into horizontal and time versus vertical components is very convenient for the derivation of the equatorial waves. As will be shown, the equatorial waves are the solutions, or normal modes, of the equations that are trapped near the Equator.

Horizontal and Time-Varying Structure

The shallow water equations on the rotating Earth in Cartesian geometry may be written as eqns [1]–[3]:

$$\frac{\partial u'_l}{\partial t} - f v'_l = -\frac{\partial \Phi'_l}{\partial x} \quad [1]$$

$$\frac{\partial v'_l}{\partial t} + f u'_l = -\frac{\partial \Phi'_l}{\partial y} \quad [2]$$

$$\frac{\partial \Phi'_l}{\partial t} + g h_l \left(\frac{\partial u'_l}{\partial x} + \frac{\partial v'_l}{\partial y} \right) = 0 \quad [3]$$

where f is the Coriolis parameter ($f \equiv 2\Omega \sin \phi$ where Ω is the angular velocity of the rotation of the Earth and ϕ is the latitude), g is the acceleration due to gravity, h_l is the equivalent

depth, u'_l and v'_l are the (horizontal) velocities in the x and y directions, respectively, and Φ'_l is the geopotential. The primes denote the fact that these are perturbation fields relative to the basic state of no mean flow about which the equations are linearized. The subscript l signifies that these equations govern a particular vertical normal mode only. In the shallow water equations, the unknown variables are a function of horizontal position (x, y) and time (t) only.

As written, these equations are valid for perturbations to the free surface of a shallow layer of incompressible, homogeneous density fluid of mean depth h_l (also called the divergent barotropic model). However, these equations are equally valid for internal modes of the stratified atmosphere, as discussed above, provided the correct choice of h_l is made. Then, to obtain the complete four-dimensional (in x, y, t , and z) structure of the internal normal mode, one must multiply the solutions for u'_l, v'_l , and Φ'_l by the particular internal mode's vertical structure function $G_l(z)$, as in eqn [4].

$$u'(x, y, z, t) = u'_l(x, y, t)G_l(z) \quad [4]$$

Returning to eqns [1]–[3], the approximation that the Coriolis parameter, f , is linearly proportional to the latitude is now made. That is, $f = \beta y$, to give what is called the equatorial β -plane. Equations [1]–[3] then become eqns [5]–[7].

$$\frac{\partial u'_l}{\partial t} - \beta y v'_l = -\frac{\partial \Phi'_l}{\partial x} \quad [5]$$

$$\frac{\partial v'_l}{\partial t} + \beta y u'_l = -\frac{\partial \Phi'_l}{\partial y} \quad [6]$$

$$\frac{\partial \Phi'_l}{\partial t} + g h_l \left(\frac{\partial u'_l}{\partial x} + \frac{\partial v'_l}{\partial y} \right) = 0 \quad [7]$$

Now solutions are sought in the form of zonally propagating waves, eqn [8], where k is the zonal wave number and ω is the frequency.

$$(u'_l, v'_l, \Phi'_l) = \Re\{[\hat{u}(y), \hat{v}(y), \hat{\Phi}(y)] \exp[i(kx - \omega t)]\} \quad [8]$$

Substitution of eqn [8] into eqns [5]–[7] then yields a set of ordinary differential equations in y for the meridional structure functions $\hat{u}, \hat{v}, \hat{\Phi}$, eqns [9]–[11].

$$-i\omega \hat{u} - \beta y \hat{v} = -ik \hat{\Phi} \quad [9]$$

$$-i\omega \hat{v} + \beta y \hat{u} = -\frac{d\hat{\Phi}}{dy} \quad [10]$$

$$-i\omega \hat{\Phi} + g h_l \left(ik \hat{u} + \frac{d\hat{v}}{dy} \right) = 0 \quad [11]$$

This set of equations may be arranged to eliminate both $\hat{u}$ and $\hat{\Phi}$, yielding a second-order differential equation in $\hat{v}$ only (eqn [12]).

$$\frac{d^2 \hat{v}}{dy^2} + \left(\frac{\omega^2}{g h_l} - k^2 - \frac{k}{\omega} \beta - \frac{\beta^2 y^2}{g h_l} \right) \hat{v} = 0 \quad [12]$$

Since wave solutions near the Equator are being considered, the boundary condition for this equation that the solutions will decay at large $|y|$ is required. This boundary

condition is also required so that the equatorial β -plane approximation remains valid. It is the case that eqn [12] with the said boundary condition is of the same form as one encountered in physics, called the Schrödinger equation for a simple harmonic oscillator. Decay of solutions away from the Equator is only satisfied when the constant part of the coefficient in parentheses satisfies the relationship in eqn [13].

$$\frac{\sqrt{g h_l}}{\beta} \left(\frac{\omega^2}{g h_l} - k^2 - \frac{k}{\omega} \beta \right) = 2n + 1; \quad n = 0, 1, 2, \dots \quad [13]$$

This equation gives a relation between the frequency, ω , and the longitudinal wave number, k , for some positive integer number n . That is, it defines the (horizontal) dispersion relations of the equatorial waves. Since it is a cubic equation in ω , there are three roots for ω when n and k are specified. As will be discussed later, n corresponds to the number of nodes in the meridional profile of the meridional velocity, and is thus called the meridional mode number. Also of note is the symmetry of eqn [13] such that the quadrant of wave numbers and frequencies defined by $k < 0, \omega < 0$ is identical to that defined by $k > 0, \omega > 0$. Similarly, the quadrant defined by $k > 0, \omega < 0$ is identical to that defined by $k < 0, \omega > 0$. By eqn [8], these quadrants of wave number–frequency space are separately occupied by eastward- or westward-propagating waves.

Solutions to eqn [13] may be investigated by making various approximations. For example, at low frequencies, the term $\omega^2/g h_l$ can be neglected. This gives eqn [14] as an approximation of one of the three roots.

$$\omega_{\text{Rossby}} \approx \frac{-\beta k}{k^2 + (2n + 1)\beta/\sqrt{g h_l}} \quad [14]$$

The wave class corresponding to this root is called the equatorial Rossby waves. They are westward propagating only, as ω is of the opposite sign to k . For $h_l \rightarrow \infty$, this dispersion relation becomes like that for midlatitude Rossby waves.

At high frequencies, the term $-k\beta/\omega$ in eqn [13] can be neglected. This gives eqn [15] as an approximation for the other two roots of eqn [13].

$$\omega_{\text{IG}} \approx \pm \left[(2n + 1)\beta\sqrt{g h_l} + k^2 g h_l \right]^{1/2} \quad [15]$$

The wave classes corresponding to these roots are called the eastward inertio-gravity (EIG) (for the positive root) and westward inertio-gravity (for the negative root) waves, respectively. For large k and small n , the dispersion of these waves approximates that of pure gravity waves.

All the solutions of eqn [13], computed with no approximation, are shown in Figure 1. The wave number and frequency have been nondimensionalized by taking the units of time and length as given in eqn [16].

$$[T] = (1/\beta\sqrt{g h_l})^{1/2} \quad \text{and} \quad [L] = (\sqrt{g h_l}/\beta)^{1/2} \quad [16]$$

This scaling allows the curves to be plotted irrespective of the value of h_l, β , or g . The three classes of waves discussed so far appear in the bottom left, upper right, and upper left portions of the diagram, respectively. Also shown are the dispersion

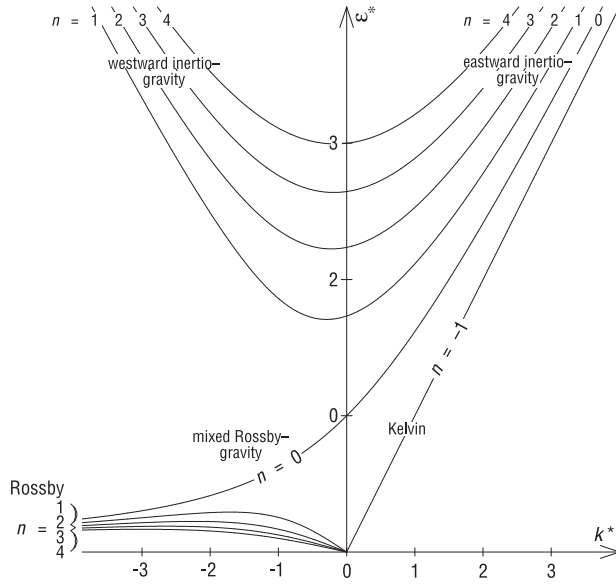


Figure 1 Dispersion curves for equatorial waves (up to $n = 4$) as a function of the nondimensional frequency, ω^* , and zonal wave number, k^* , where $\omega^* \equiv \omega / (\beta \sqrt{gh_l})^{1/2}$, and $k^* \equiv k (\sqrt{gh_l} / \beta)^{1/2}$. For all but the Kelvin wave, these dispersion curves are solutions of eqn [13]. Eastward-propagating waves (relative to the zero basic state employed) appear on the right-hand side (i.e., for $k^* > 0$), and westward-propagating waves appear on the left (i.e., for $k^* < 0$).

curves of two additional wave types that require special consideration. They are the $n = 0$ solutions of eqn [13], and the Kelvin wave, for which $\hat{v} = 0$ and which is thus not covered by the solutions of eqn [12] or [13].

The $n = 0$ solution can be obtained directly from eqn [13]. The allowed roots are given exactly by eqn [17].

$$\omega_{n=0} = k \sqrt{gh_l} \left[\frac{1}{2} \pm \frac{1}{2} \left(1 + \frac{4\beta}{k^2 \sqrt{gh_l}} \right)^{1/2} \right] \quad [17]$$

The positive root corresponds to an EIG wave, while the negative root corresponds to a westward-propagating wave. This westward wave is generally called the mixed Rossby-gravity wave, as it shares properties of both the Rossby and inertio-gravity waves. It is also sometimes called the Yanai wave. For $|k| \rightarrow 0$, the positive and negative roots coincide to produce a continuous curve, as can be seen in Figure 1.

The Kelvin wave dispersion curve and horizontal structure are obtained from eqns [9]–[11] by setting $\hat{v} = 0$. Combining the equations to eliminate $\hat{\Phi}$, one is able to show that the Kelvin wave dispersion relation is given by $\omega_{\text{Kelvin}} = \sqrt{gh_l} k$, with the meridional structure of u given by $u_{\text{Kelvin}} = \exp(-\beta y^2 / 2 \sqrt{gh_l})$. This wave is often labeled as the $n = -1$ wave, as its dispersion relation can also be obtained by setting $n = -1$ in eqn [13].

For the horizontal structures of the rest of the waves (i.e., for all waves but the Kelvin wave), one must return to eqn [12]. From the study of the similar Schrödinger equation in physics, it is known that the solutions of eqn [12] are given

exactly by eqn [18], where H_n is the Hermite polynomial of the n th order.

$$\hat{v}(y) = \exp(-\beta y^2 / 2 \sqrt{gh_l}) H_n((\beta / gh_l)^{1/2} y) \quad [18]$$

As required, these solutions for $\hat{v}$ decay for large $|y|$. The first few of the Hermite polynomials are as listed in eqn [19].

$$\begin{aligned} H_0 &= 1, & H_1(\varepsilon) &= 2\varepsilon, & H_2(\varepsilon) &= 4\varepsilon^2 - 2, \\ H_3(\varepsilon) &= 8\varepsilon^3 - 12\varepsilon \end{aligned} \quad [19]$$

As already mentioned, the index n corresponds to the number of nodes in $\hat{v}$ in the domain $|y| < \infty$. These solutions for $\hat{v}$, along with the dispersion relations from eqn [13], may be substituted back into eqns [9]–[11] to obtain the solutions for $\hat{u}$ and $\hat{\Phi}$. The corresponding full horizontal structures, nondimensionalized using the same units of length and time as given in eqn [16], are as displayed in Figure 2. Also shown (in Figure 2(h)) is the full horizontal structure of the Kelvin wave which was obtained from eqns [9]–[11] with $\hat{v} = 0$.

Through inspection of Figures 1 and 2, further properties of these waves can be demonstrated. Note that to obtain a dimensioned value from these figures, the variables must be multiplied and divided by the correct combination of time and length scales; for example, $\Phi = \Phi^* \times [L]^2 / [T]^2$, where Φ^* is the plotted nondimensional geopotential. It is also important to note that because eqns [1]–[3] are linear, any linear combination of the waves described by the figures is also a solution to the equations.

Concentrating for the moment on the dispersion curves for the waves (Figure 1), many of their important properties can be discerned by recalling the equations for the zonal phase speed and the zonal component of the group velocity as $c_p^{(x)} \equiv \omega / k$ and $c_g^{(x)} \equiv \partial \omega / \partial k$, respectively. That is, the group velocity is the local slope of the curves, while the phase speed is determined by the position on the diagram relative to the origin. Thus it can be determined that equatorial Rossby waves only propagate to the west, while their energy (as inferred from the group velocity) may propagate to the east or west, depending on their zonal scale. Mixed Rossby-gravity waves, on the other hand, always have westward phase and eastward energy propagation. Kelvin waves are nondispersive waves, with their phase propagating relatively quickly to the east at the same speed as their group velocity. EIG waves, as their name implies, always have phase propagation to the east. Their group propagation is also always to the east. Finally, westward inertio-gravity waves have phase propagation to the west, and their group propagation is also to the west, except for very low zonal wave numbers. Obviously, the inertio-gravity waves propagate much more quickly than the Rossby waves, while the Kelvin wave has a phase speed of intermediate magnitude.

Typical values of the Kelvin wave phase speed, also known as the gravity wave speed, are in the range of $c_l = \sqrt{gh_l} \approx 10 - 50 \text{ m s}^{-1}$ in the troposphere (corresponding to h_l in the range of 10–250 m), and skewed to higher values in the middle atmosphere. For equatorial waves that propagate along the thermocline (i.e., internally in the ocean), appropriate values of c_l are in the range of $0.5 - 3.0 \text{ m s}^{-1}$ (corresponding to h_l in the range of 0.025–1.0 m).

Now attention may be turned to the wave's horizontal structures as presented in Figure 2, with each wave displayed in

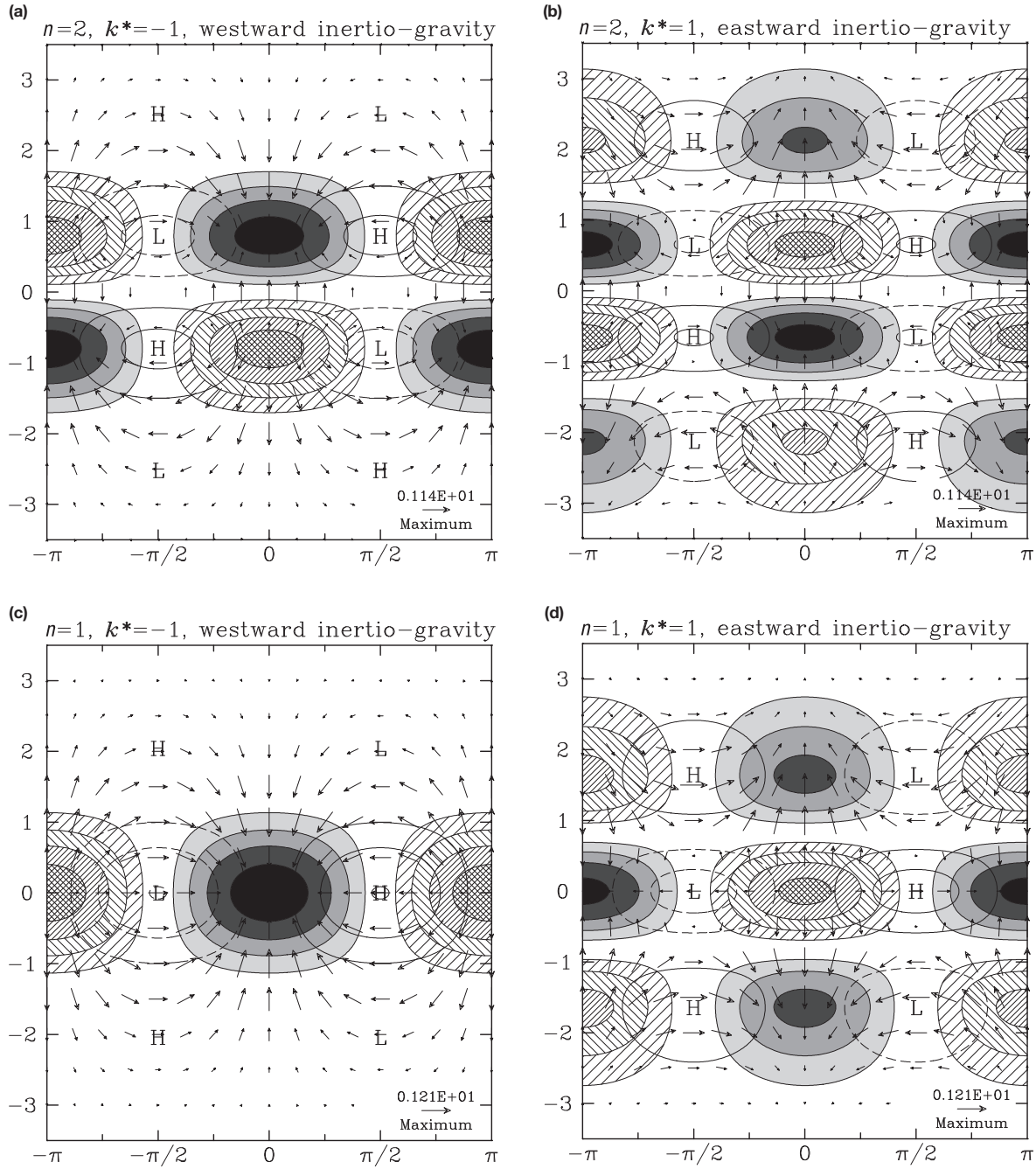


Figure 2 Horizontal structures of the equatorial wave solutions to the shallow water equations on an equatorial β -plane (eqns [5]–[7]), for nondimensional zonal wave number $|k^*| = 1$. All scales and fields have been nondimensionalized by taking the units of time and length as $[T] = (1/\beta\sqrt{gh_i})^{1/2}$ and $[L] = (\sqrt{gh_i}/\beta)^{1/2}$, respectively. The Equator runs through the center of each diagram. Hatching is for divergence and shading for convergence, with a 0.6 unit interval between successive levels. Unshaded contours are for geopotential, with a contour interval of 0.5 units. Negative contours are dashed and the zero contour is omitted. The largest wind vectors are as specified in the bottom right corner. (a)–(d) are for the $n=1$ and $n=2$ inertio-gravity waves; (e) and (f) for the $n=0$ waves; (g) for the $n=1$ Rossby wave; (h) for the Kelvin wave; (i) For the $n=2$ Rossby wave; and (j) for the $n=3$ Rossby wave.

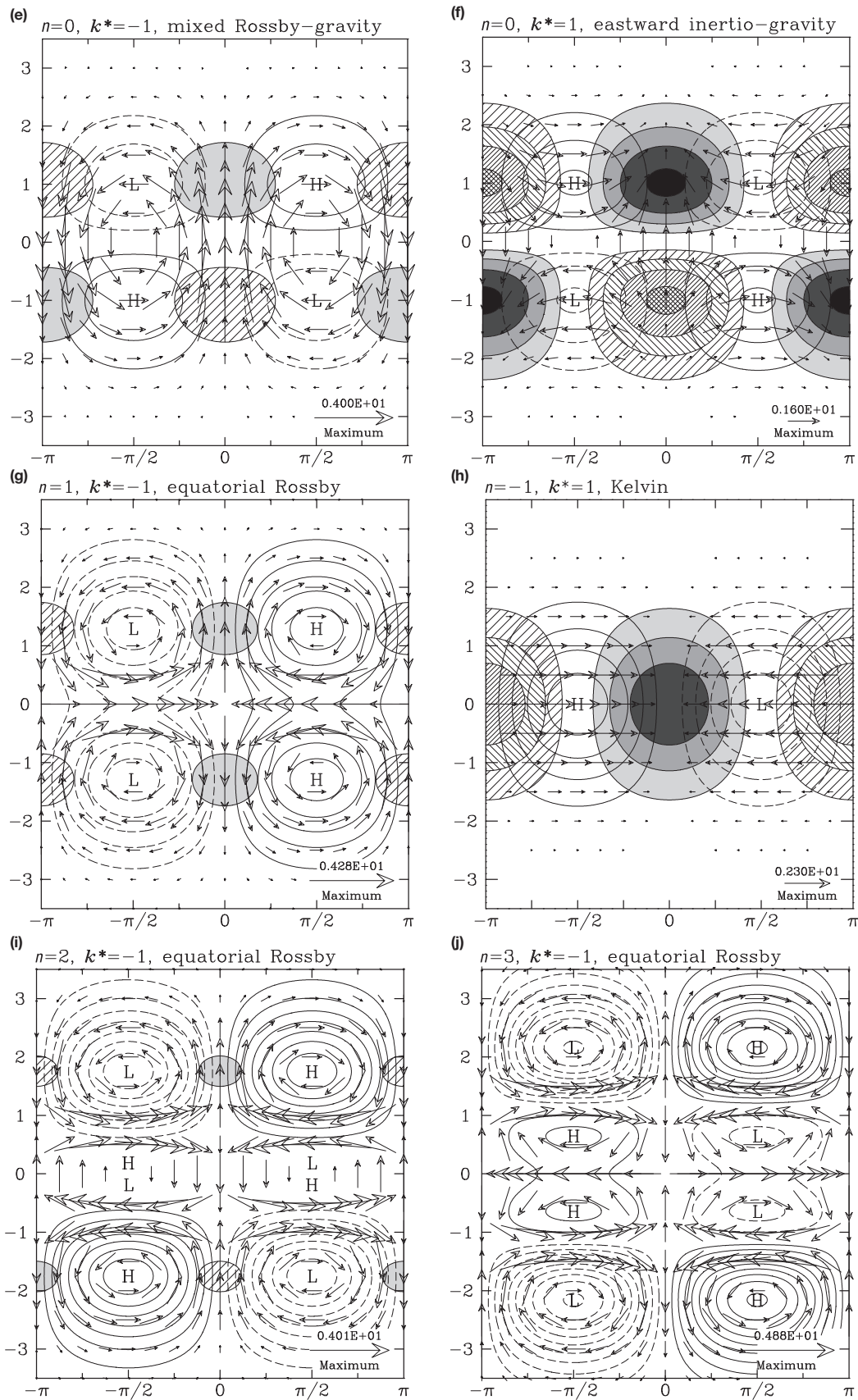


Figure 2 (Continued)

a different panel for a nondimensional zonal wave number of $|k^*| = 1$. The horizontal scale of the waves may be determined by computing the unit of length by which the solutions have been scaled, $[L]$. This scale is also known as the equatorial Rossby radius. It determines the meridional scale of the waves, and is a function of the equivalent depth. For the troposphere, with $h_l \approx 10$ –250 m, $[L] \approx 6$ –13 degrees of latitude. For internal modes in the ocean, however, $[L] \approx 1.3$ –3.3 degrees of latitude. Hence, equatorial waves in the ocean are more closely trapped to the Equator than those in the troposphere.

Fundamental differences in the general structures of the waves can also be seen. For the inertio-gravity waves, the signal of their divergence is very strong, while the magnitude of their winds and geopotential are relatively weak. For Rossby waves the opposite is true. That is, Rossby waves are much more rotational in character than the more divergent inertio-gravity waves.

Vertical Structure

As mentioned in the previous section, to obtain the complete structure of each equatorial wave mode, one must multiply the solutions from the shallow water equations (as presented at an instant in time in Figure 2) by the vertical structure function $G_l(z)$, as in eqn [20], where $G_l(z)$ is governed by the vertical structure equation.

$$u'(x, y, z, t) = u'_l(x, y, t)G_l(z) = \Re\{\hat{u}(y)\exp(ikx - \omega t)\} \times G_l(z) \quad [20]$$

In a dry atmosphere of constant buoyancy frequency N , the vertical structure equation may be expressed as eqn [21].

$$\frac{1}{\rho_0} \frac{d}{dz} \left(\rho_0 \frac{dG_l}{dz} \right) + \frac{N^2}{gh_l} G_l = 0 \quad [21]$$

This equation is a consequence of the mathematical separation. Given the basic state density distribution $\rho_0 \propto \exp(-z/H_s)$, where H_s is the scale height, solutions to this equation may be sought of the form in eqn [22].

$$G_l = \exp(z/2H_s) \exp(\pm imz) \quad [22]$$

Substitution of eqn [22] into eqn [21] then yields eqn [23].

$$m \equiv \frac{2\pi}{L_z} = \left(\frac{N^2}{gh_l} - \frac{1}{4H_s^2} \right)^{1/2} \quad [23]$$

In eqn [23], L_z is the vertical wavelength of the l th normal mode, m is the vertical wave number, $N^2 = ((R/H_s)(dT_0/dz + g/C_p))$ is the buoyancy frequency squared, dT_0/dz is an average lapse rate, and R and C_p are the gas constant and specific heat for dry air, respectively. Equation [23] gives a relationship between the vertical wavelength of a normal mode in a constant N atmosphere, and its equivalent depth, h_l .

Equations [22] and [23] reveal that in a constant N atmosphere, vertical normal modes are sinusoidal in z , with corrections for the density variation with height. In a realistic stratification, however, the modes are only quasisinusoidal, with a local wavelength that is small in regions of large N (such as in the stratosphere) and larger in regions of low N (such as in the troposphere). Example values of the local vertical wavelength computed using eqn [23], given h_l , are presented in

Table 1 Vertical structure values for dry waves in constant N atmosphere^a

h_l (m)	$\sqrt{gh_l}$ (m s ⁻¹)	L_z (km)
$H_s = 7.3$ km, $dT_0/dz = -7.0$ km ⁻¹ (Troposphere)		
10	9.9	6.0
20	14.0	8.5
50	22.1	13.4
100	31.3	19.2
200	44.3	27.9
500	70.0	47.5
$H_s = 6.1$ km, $dT_0/dz = +2.5$ km ⁻¹ (Lower stratosphere)		
10	9.9	2.6
20	14.0	3.7
50	22.1	5.8
100	31.3	8.3
200	44.3	11.8
500	70.0	18.9

^aAs calculated from eqn [23].

Table 1, separately for stratifications representative of the troposphere and lower stratosphere.

The information in Table 1 may be used to estimate which vertical modes are excited by a particular forcing of the atmosphere. For example, if the equatorial atmosphere is forced by the latent heating that occurs in conjunction with atmospheric moist convection, and this heating extends up to the upper troposphere (~ 14 km) with greatest magnitude around the mid-troposphere (~ 7 km), then this heating projects most strongly onto the vertical mode with a vertical wavelength of about 28 km. In the troposphere, this corresponds to an equivalent depth of about 200 m. Such a heating, however, also projects onto other vertical modes (but with reduced strength) because it takes more than one sinusoid to reconstruct a half wavelength in the troposphere. Hence, the resulting complete structure must be computed by a summation of the form in eqn [24], where each u'_l is calculated from the shallow water equations with a different h_l .

$$u'(x, y, z, t) = \sum_l u'_l(x, y, t)G_l(z) \quad [24]$$

Combining vertical modes in this way allows for vertical propagation of a wave packet. Then, phase lines of the wave packet may slope in the vertical, perpendicular to the total wave number vector given by $\vec{k} \equiv (k, m)$, where m is defined to be positive for upward phase propagation and negative for downward phase propagation, assuming that the frequency (ω) is defined to be positive. The total phase of the wave is then given by $\varphi = kx + mz - \omega t$.

Finally, it is important to note that the values in Table 1 apply only when there is no interaction between the waves and moist convection. Theoretical and observational studies have shown that the presence of interacting convection tends to slow down the zonal propagation of waves, or effectively reduce the equivalent depth of the most energetic waves that are seen. For many purposes, however, this theory for the dry waves is very useful for interpretation and understanding, especially in regions such as the stratosphere and over cool sea surface temperatures, where moist convection does not occur.

Model Experiment

Many of the theoretical properties of equatorial waves, as derived in the previous section, may be conveniently illustrated by showing results from a numerical model of the global atmosphere. Here a numerical experiment in which the primitive equations are forced by an imposed heating over 2 days is shown. The heating is representative of latent heating from condensation that occurs within a large aggregation of precipitating convective clouds. The heating is switched on slowly over 1 day, and switched off over the course of the next day. In the vertical, the heating extends from near the surface up to a level of about 14 km, peaking in the mid-troposphere. At its peak, the heating forces the model's equation for potential temperature by an amount equivalent to 10 K day^{-1} , corresponding to a peak precipitation rate of about 40 mm day^{-1} . In the horizontal, the heating spreads over 30 degrees of latitude and 40 degrees of longitude, and is centered on the Equator. The model is integrated forward in time from an initial state of rest with a realistic tropical temperature stratification specified globally, i.e., with tropospheric and stratospheric lapse rates like those specified in Table 1.

Figure 3 shows a horizontal cross section of the model experiment at a level in the upper troposphere. Given the symmetric (with respect to the Equator) nature of the forcing, only equatorial modes that are symmetric about the Equator

are forced. In particular, a Kelvin wave packet propagates to the east from the region of forcing, and a packet of (primarily $n = 1$) Rossby waves slowly develops to the west. Both of these packets are comprised of waves with a number of different wave numbers and frequencies. By the fifth day of the model run, the energy of the Rossby wave packet has dispersed such that a new pair of circulation cells have developed not far from the location of the initial forcing. In contrast, the main Kelvin wave packet has shifted, with little dispersion, half way around the globe.

Figure 4 shows a vertical cross section of the experiment. As explained in the previous section, a heating of this vertical extent projects most strongly onto the vertical mode with an equivalent depth of around 200 m. The main Kelvin wave front in the troposphere, with a vertical half wavelength filling the depth of the troposphere, thus propagates to the east at around 40 m s^{-1} ($\approx \sqrt{9.8 \times 200}$). The projection in the vertical, however, is not restricted to this vertical mode. Therefore, the combination of vertical modes allows vertical energy propagation into the stratosphere via vertically propagating waves, as indicated by the tilted contours of temperature there. In addition, for the same horizontal propagation speed, the vertical wavelength in the stratosphere tends to be about half of that in the troposphere, consistent with Table 1. Indeed, the position of the tropopause at around 100 hPa seems to play an important role in the appearance of the waves.

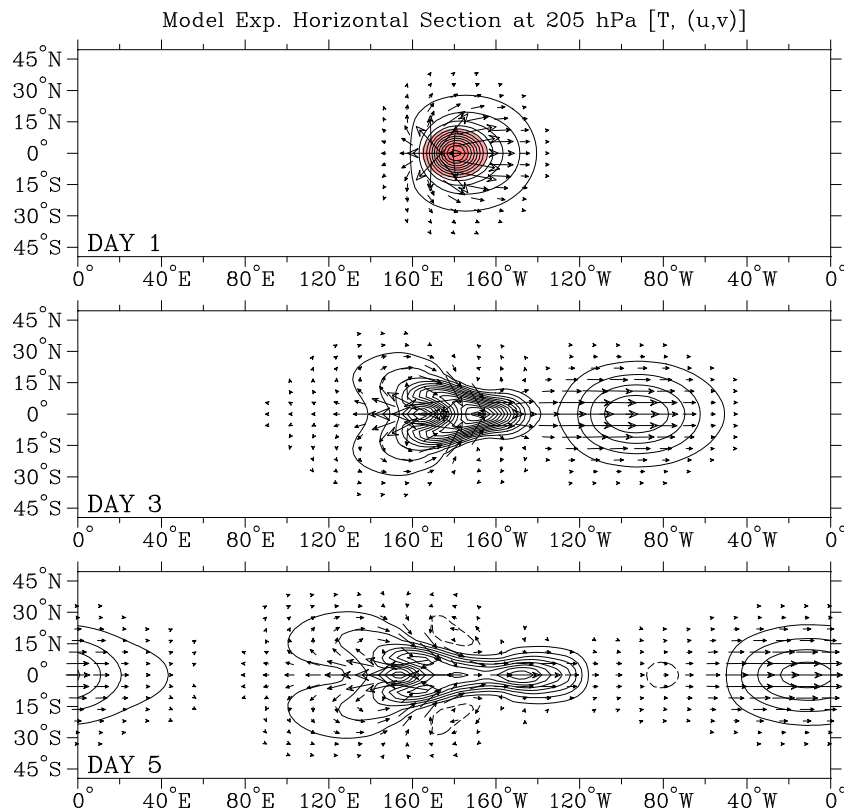


Figure 3 Horizontal cross section of model experiment at a level of 205 hPa (i.e., in the upper troposphere). Vectors show the horizontal winds, and contours show the temperature perturbations. Contour interval is 0.05 K with the zero contour omitted and negative contours dashed. The maximum vector is 3.8 m s^{-1} on day 1, 6.2 m s^{-1} on day 3, and 6.2 m s^{-1} on day 5. The red coloring in the first panel shows the imposed heating. The axes are labeled with latitude and longitude.

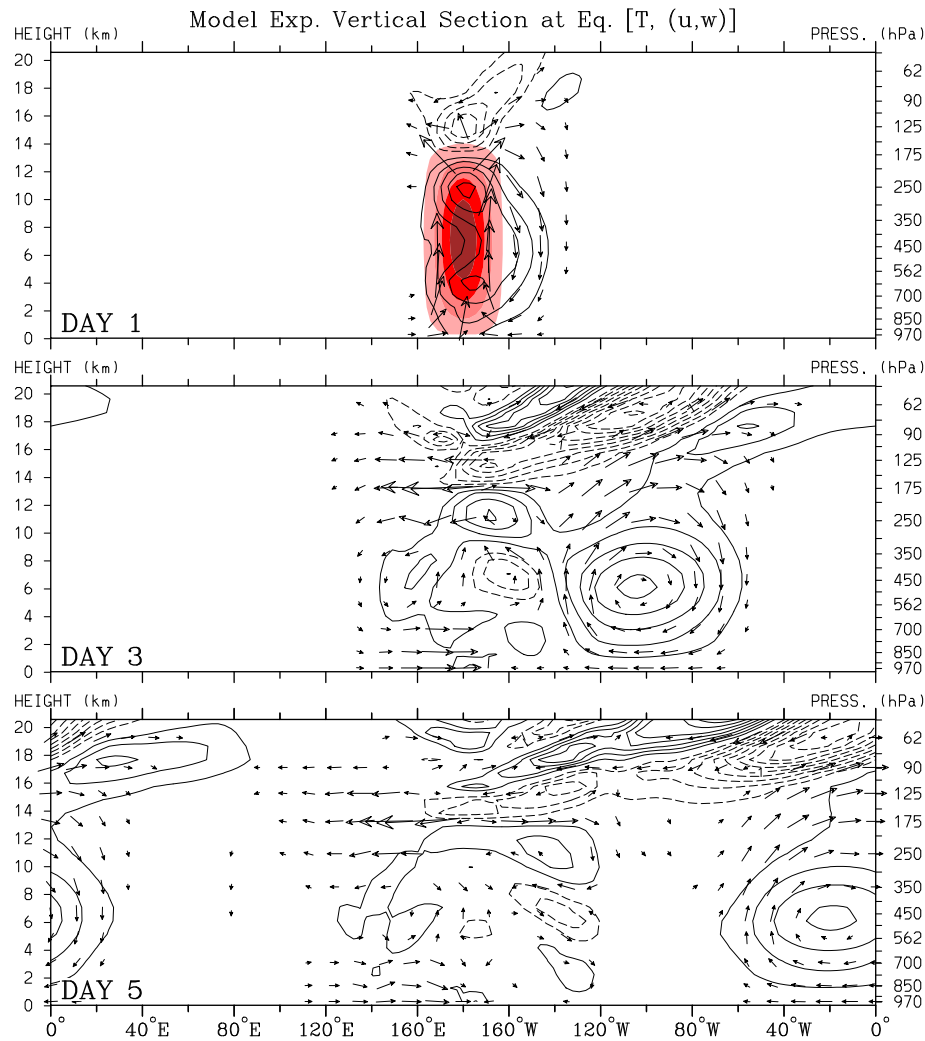


Figure 4 Vertical cross section of model experiment at the Equator. Vectors show the zonal-vertical components of the wind, where the vertical component has been scaled by the same amount that the vertical axis has been stretched. The largest vector on day 5 is 8.1 m s^{-1} . Contours show the temperature perturbations with a contour interval of 0.2 K . Negative contours are dashed. The red coloring in the first panel shows the imposed heating at the levels of $0.5, 3.0, 5.5$, and 8.0 K day^{-1} .

Further consistency with the theory is the appearance of a second Kelvin wave front that is located at about 130°W by day 5 and with a vertical wavelength about three times shorter than the main front. Its speed is also about three times slower than that of the main front. In summary, the linear shallow water theory can be used to understand most aspects of what happens in this model of the atmosphere. In the real world, however, with many other processes and events going on at the same time, equatorial waves are not always so easy to identify.

Observations and Current Understanding

Conclusive observations of the existence of equatorial waves in the atmosphere have been made since the mid-1960s. The first observations were those of mixed Rossby-gravity waves and Kelvin waves. The waves were identified through the analysis of wind measurements from balloon soundings by the researchers

Yanai, Maruyama, Wallace, and Kousky. They were found to be propagating vertically into the stratosphere from a tropospheric energy source, much like the Kelvin waves in the model experiment of the previous section. The waves were also found to be important sources of momentum for the equatorial stratosphere, providing some of the necessary alternating zonal wind accelerations of the stratospheric quasi-biennial oscillation.

A more recent example of such balloon sounding observations of equatorial waves is presented in Figure 5. The soundings were taken from the island of Nauru near the Equator in the Western Pacific. The meridional component of the wind shows a prominent oscillation with a period of around 4–5 days. This oscillation can be identified to be that of a mixed Rossby-gravity wave, because these are the only waves that have a period in this range, for the appropriate range of vertical structures, and have strong meridional winds on the Equator.

This particular event of strong mixed Rossby-gravity waves lasted for about 2 weeks, and was restricted in its location to the Western to Central Pacific. Such events do not

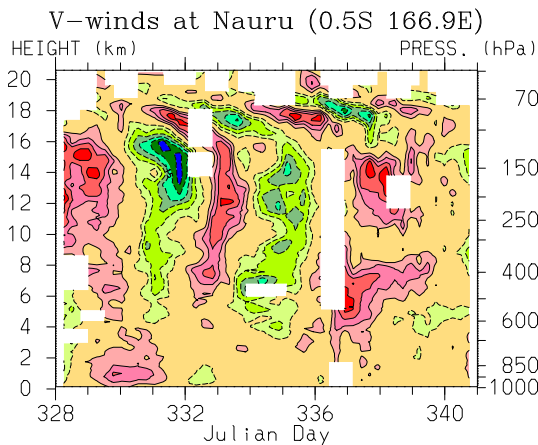


Figure 5 Time-height plot of the meridional component of the wind recorded by balloon soundings taken at the island of Nauru (0.5°S , 166.9°E) from 23 November to the end of 5 December 1992. Contour interval is 3 m s^{-1} , with the zero contour omitted. Southerly winds stronger than 3 m s^{-1} are colored red, northerly winds stronger than 3 m s^{-1} are colored green. Blank areas indicate missing data.

occur all the time, and may occur at different longitudes at different times. To get an appreciation of the global occurrence of equatorial waves in space and time, researchers have used the technique of space-time Fourier spectral analysis. Such analysis gives information on the strength (or 'power') of oscillations in winds or other variables at each wave number and frequency. The technique is particularly useful for the analysis of equatorial waves because it can distinguish between oscillations resulting from eastward versus westward propagation.

Figure 6 shows a particular presentation of the power spectral peaks of a 22-year record of a satellite-observed proxy for tropical moist convection. The proxy for convection is the infrared brightness temperature (T_b) as recorded by multiple satellites and combined into a single dataset known by the Cloud Archive User Services (CLAUS). Infrared T_b data are useful as a measure of tropical moist convection because the tops of deep convective clouds emit much less infrared radiation than the comparatively warm sea or land surface. Before computation of the spectra, the T_b data are decomposed into symmetric and antisymmetric components about the Equator to distinguish between the spectral power that may be contributed by either symmetric or antisymmetric equatorial waves (see **Figure 2**). Once the raw power is computed and summed over the latitudes from 15°S to 15°N , it is divided by an estimate of the red noise background power. It is this ratio that is presented by the contours in **Figure 6**. Spectral regions having a ratio greater than 1.1, that is, with power more than 10% above the background, are shaded. Also shown are the theoretical dispersion curves of the equatorial waves, as in **Figure 1**, except scaled for the three equivalent depths of 12, 25, and 50 m. The match between the spectral peaks of the observed dataset and the theoretical dispersion curves is readily apparent. This match indicates that equatorial waves play an important role in shaping the space and time distribution of tropical convection. That is, equatorial waves have some role in determining when and where tropical convective weather will occur.

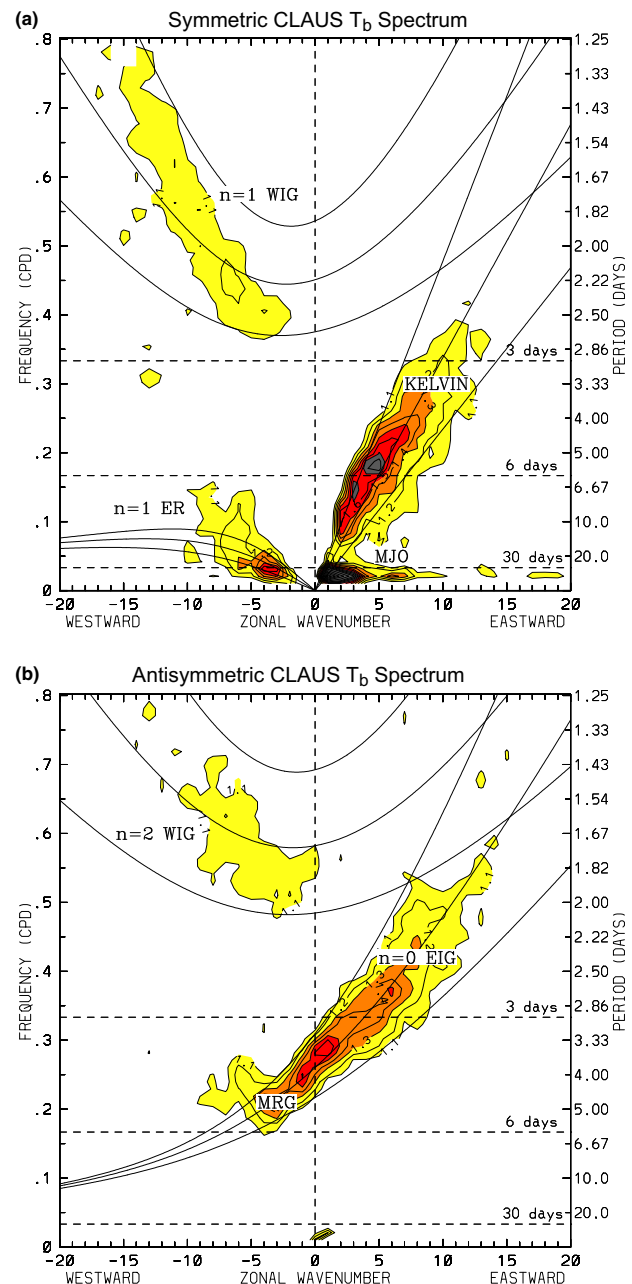


Figure 6 A presentation of the wave number–frequency spectral peaks of a long record of CLAUS T_b data, separately for the (a) antisymmetric and (b) symmetric components of the T_b with respect to the Equator, using data for latitudes from 15°S to 15°N . To represent the peaks, the contours show a ratio of the actual power with an estimate of the red noise background power. A ratio of greater than 1.1 is a statistically significant spectral peak (95% level). Also shown are the dispersion curves for the equatorial waves for equivalent depths of 12, 25, and 50 m. The (planetary) zonal wave number is the number of wavelengths that fit around the globe at the Equator. Unit of frequency is cycles per day (cpd). Reproduced from Kiladis, G.N., Wheeler, M.C., Haertel, P.T., Straub, K.H., Roundy, P.E., 2009. Convectively coupled equatorial waves. *Review of Geophysics* 47: RG2003, doi:10.1029/2008RG000266.

The phenomena contributing to these spectral peaks have come to be known as the ‘convectively coupled equatorial waves’. They are described as ‘convectively coupled’ because they involve an interaction between tropical convection and the dynamics of equatorial waves. The convection tends to occur where the waves generate deep upward motions in the troposphere, with the resulting moist convection reinforcing

this upward motion through the associated latent heating of condensation and consequent further generation of upward motion. This interaction modifies the wave’s characteristics from those that occur in a dry atmosphere. In particular, the dispersion of convectively coupled equatorial waves corresponds to shallower equivalent depths than what would occur for the most energetic waves in a dry atmosphere. Not all of the

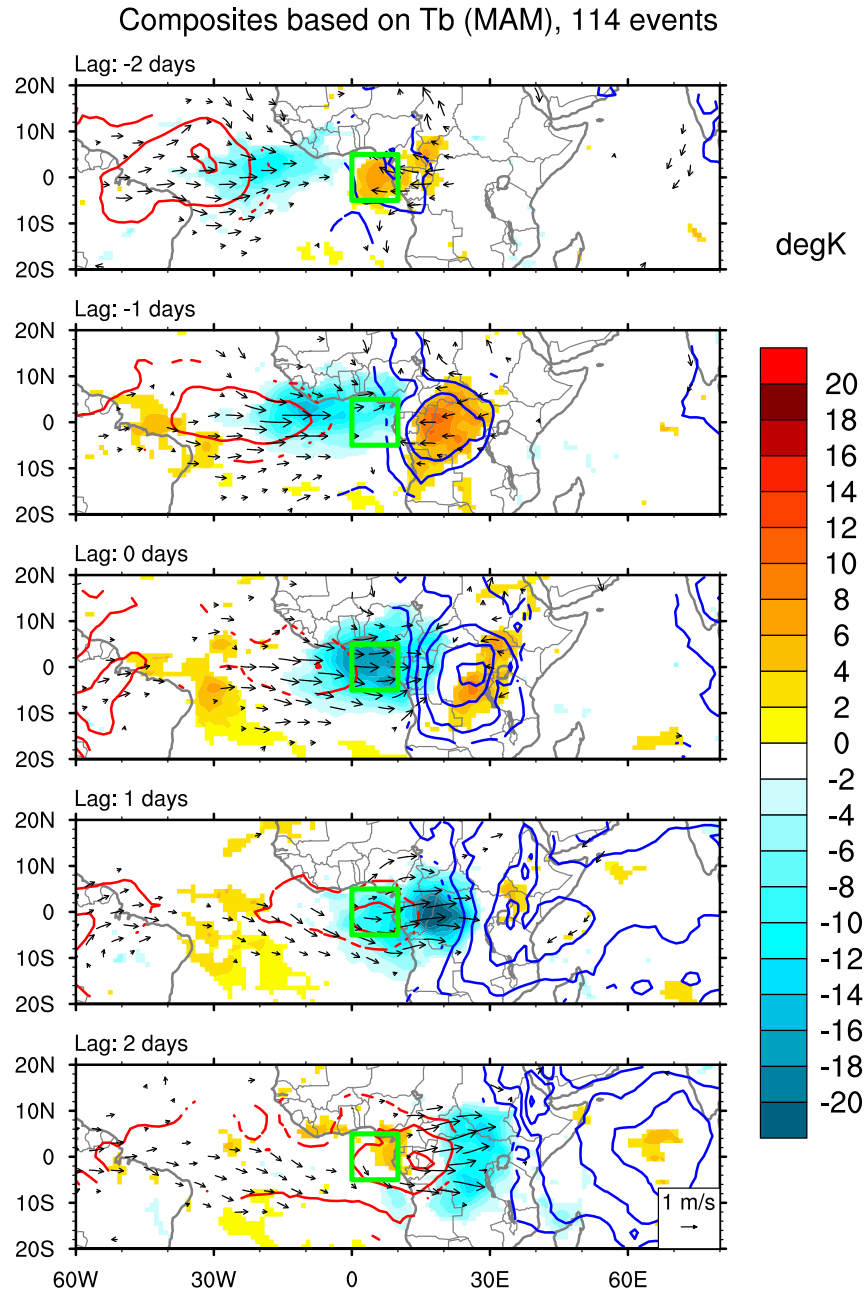


Figure 7 Composite horizontal structure of a convectively coupled Kelvin wave over the Atlantic/Africa region. Each panel represents a snapshot at a particular point in time, with each panel being separated by 1 day. The green box shows the reference region used for the computation, with the composite being computed relative to a minimum in T_b in the green box at a lag of 0 days. Cold/warm color shadings show negative/positive satellite-observed T_b anomalies (see color scale at right in degree Kelvin). Red/blue contours are for positive/negative 850-hPa geopotential height anomalies (contour interval of 1 m with the zero contour omitted) and vectors are 850-hPa wind anomalies (see scale of vector in bottom panel). Blank areas indicate where the signal of the Kelvin wave is not statistically significant (at the 90% level).

equatorial waves appear as ‘convectively coupled’ signals, such as the $n = 1$ EIG wave that shows no spectral peak in the convection (Figure 6). Why some waves appear as convectively coupled signals and others do not is still not fully understood. The exact nature of the interaction between the convection and dynamics that causes the convectively coupled equatorial waves to occur for the select range of equivalent depths is also not fully understood. Besides these questions, the ubiquitous nature of equatorial waves in the tropical atmosphere is now firmly established. Of further note is the existence of the Madden–Julian oscillation (MJO) in the symmetric spectrum of Figure 6, for eastward planetary zonal wave numbers around 1–4 and periods between 30 and 60 days. The MJO is unlike the convectively coupled equatorial waves in that its spectrum extends across wave numbers and frequencies away from the equatorial wave dispersion curves.

One useful application of the spectral results of Figure 6 is that regions of wave number and frequency can be defined for each wave and used for filtering for that wave. Usually, a box is drawn around each of the spectral peaks to define the wave numbers and frequencies that one will filter for. The filtering is performed using inverse Fourier transforms to retain only the variability that may be associated with each wave. For example, a ‘Kelvin wave-filtered’ time series of T_b data at each location this way may be defined.

A more detailed view of observed equatorial waves may be obtained by averaging together observations of many separate events to produce an image of a typical wave event, or ‘composite’. This can be conveniently done by using observations that have been input into a global weather prediction model and output onto a grid as an observed ‘analysis’. A composite can then be made of all of the global atmospheric data that occur during each phase of the wave. An example of such a composite for the convectively coupled Kelvin wave is given in Figure 7. This figure shows the composite structure of lower tropospheric wind and geopotential height anomalies of a Kelvin wave over a sequence of 5 days, together with the typical signal in the convection, as indicated by the T_b . The composite was computed based on the signal of the Kelvin wave-filtered T_b averaged over the area in the green box in Figure 7. Lag day 0 corresponds to the maximum convection (i.e., minimum of T_b) of the Kelvin wave in this box. In the first panel (i.e., 2 days before the maximum in convection in the reference region), suppressed convection associated with the suppressed phase of the Kelvin wave is occurring in the Gulf of Guinea at longitudes of 0–15° E. At the same time, easterly wind anomalies and negative geopotential height anomalies are occurring slightly to the east over the west coast of Equatorial Africa, and westerly anomalies and positive geopotential height anomalies to the west over the Equatorial Atlantic. In the second panel (i.e., 1 day before the peak in convection in the green box), the whole system has moved eastward; suppressed convection has moved over the Congo Basin about a quarter

wavelength and enhanced convection has moved to be situated near the Equator for the longitudes of 15° W–10° E. Maximum convection in the green box occurs the following day (third panel), and once again there has been a shift of the whole pattern of winds and geopotential to the east. In the last two panels, the enhanced convection associated with the Kelvin wave has shifted into the Congo Basin region, with little dispersion, consistent with the nondispersive character of a theoretical equatorial Kelvin wave. These Kelvin wave disturbances are a robust feature of synoptic timescale convective activity over the Guinea coast and Central Africa during spring. More importantly they can potentially be used to forecast rainfall events in Central Africa with several days ahead, by monitoring their initiation and propagation. Elsewhere in the world, equatorial waves have been linked to the development of tropical cyclones and other severe tropical weather.

See also: **Dynamical Meteorology:** Kelvin Waves; Primitive Equations; Rossby Waves; Waves. **Middle Atmosphere:** Quasi-Biennial Oscillation. **Tropical Cyclones and Hurricanes:** Overview and Theory. **Tropical Meteorology and Climate:** Intraseasonal Oscillation (Madden–Julian Oscillation).

Further Reading

- Andrews, D.G., Holton, J.R., Leovy, C.B., 1987. *Middle Atmosphere Dynamics*. Academic Press, New York, NY.
- Bessafi, M., Wheeler, M.C., 2006. Modulation of South Indian ocean tropical cyclones by the Madden–Julian oscillation and convectively-coupled equatorial waves. *Monthly Weather Review* 134, 638–656.
- Gill, A.E., 1982. *Atmosphere–Ocean Dynamics*. Academic Press, New York, NY.
- Holton, J.R., 2004. *Introduction to Dynamic Meteorology*, fourth ed. Academic Press, New York, NY.
- Kiladis, G.N., Wheeler, M.C., Haertel, P.T., Straub, K.H., Roundy, P.E., 2009. Convectively coupled equatorial waves. *Review of Geophysics* 47, RG2003. doi: 10.1029/2008RG000266.
- Matsuno, T., 1966. Quasi-geostrophic motions in the equatorial area. *Journal of the Meteorological Society of Japan* 44, 25–43.
- Nguyen, H., Duvel, J.-P., 2008. Synoptic wave perturbations and convective systems over equatorial Africa. *Journal of Climate* 21, 6372–6388.
- Pedlosky, J., 1987. *Geophysical Fluid Dynamics*. Springer-Verlag, Berlin.
- Roundy, P.E., Frank, W.M., 2004. A climatology of waves in the equatorial region. *Journal of the Atmospheric Sciences* 61, 2105–2132.
- Straub, K.H., Kiladis, G.N., 2002. Observations of a convectively coupled Kelvin wave in the eastern Pacific ITCZ. *Journal of the Atmospheric Sciences* 59, 30–53.
- Wallace, J.M., Kousky, V.E., 1968. Observational evidence of Kelvin waves in the tropical stratosphere. *Journal of the Atmospheric Sciences* 25, 900–907.
- Wheeler, M.C., Kiladis, G.N., 1999. Convectively coupled equatorial waves: Analysis of clouds and temperature in the wavenumber-frequency domain. *Journal of the Atmospheric Sciences* 56, 374–399.
- Wheeler, M.C., Kiladis, G.N., Webster, P.J., 2000. Large-scale dynamical fields associated with convectively coupled equatorial waves. *Journal of the Atmospheric Sciences* 57, 613–640.
- Yanai, M., Maruyama, T., 1966. Stratospheric wave disturbances propagating over the equatorial Pacific. *Journal of the Meteorological Society of Japan* 44, 291–294.

Hadley Circulation

J Lu, Pacific Northwest National Laboratory, Richland, WA, USA

GA Vecchi, GFDL/NOAA, Princeton, NJ, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by I N James, volume 3, pp 919–924, © 2003, Elsevier Ltd.

Synopsis

The Hadley circulation, a prominent circulation feature characterized by rising air near the Equator and sinking air in the subtropics, defines the position of dry subtropical areas and is a fundamental regulator of the earth's energy and momentum budgets. The character of the Hadley circulation, and its related precipitation regimes, exhibits variation and change in response to both climate variability and radiative forcing changes. The strength and position of the Hadley circulation change from year to year paced by El Niño and La Niña events. Over the last few decades of the twentieth century, the Hadley cell has expanded poleward in both hemispheres, with changes in atmospheric composition (including stratospheric ozone depletion and greenhouse gas increases) thought to have contributed to its expansion. This article introduces the basic phenomenology and driving mechanism of the Hadley circulation and discusses its variations under both natural and anthropogenic climate forcings.

Introduction

The 'Hadley circulation,' named after George Hadley, an English Lawyer and amateur meteorologist, is perhaps the earliest attempt to account for the global-scale distribution of winds in the Earth's atmosphere in terms of basic physical processes. Halley in 1685 and Hadley in 1735 both proposed that the 'trade winds' that blow toward the Equator at low latitudes could be understood as the lower branch of an axially symmetric convection cell driven by the temperature difference between the Equator and poles of the Earth. Their ideas were ahead of their time, especially as there was then no prospect of determining winds at upper levels of the atmosphere and thus verifying their hypothesis. When routine upper-air observations became available in the mid-twentieth century, the ideas of Halley and Hadley were essentially confirmed. Today, the term 'Hadley circulation' refers to the zonally averaged meridional overturning motions in the low-latitude troposphere.

Figure 1 provides a schematic view of the traditional global atmospheric circulation dividing the Earth into a set of climate zones, with the trade wind regime confined to the tropics. The trade winds are simply the low-level part of the overturning 'Hadley cells,' with ascent near the Intertropical Convergence Zone (ITCZ), descent in the subtropics and a poleward return flow at upper levels. The Hadley circulation plays a key role in transporting heat, moisture, momentum meridionally and is an indispensable component for understanding the global climate system and its variability. The descending branch of the Hadley circulation sets the position of the dry subtropical regions (in which most of the world's deserts are found) and is bounded on its poleward flank by the extratropical storm tracks, with mean surface westerly flow. The transition from easterly to westerly flow through the subtropics results in anticyclonic wind stress curl, which drives ocean surface convergence and results in large regions of suppressed ocean biological productivity. Thus, the poleward branches of the Hadley circulation set the location of the main land and ocean deserts, the former through reduced moisture supply and the

latter through reduced nutrient supply. The more disturbed midlatitudes are characterized by generally westerly winds, with irregular growing and decaying eddies, the cyclonic and anticyclonic weather systems generated by baroclinic instabilities. When averaged around entire latitude circles, this turbulent midlatitude flow averages to a weak 'Ferrel circulation,' in which warmer air at lower latitudes sinks and colder air at high latitude rises. There is some evidence of a very weak 'polar cell' at high latitudes.

The energy that drives the Hadley circulation comes from the conversion of heat energy to mechanical energy in the tropical atmosphere: the Hadley circulation is a classic example of a thermodynamic 'heat engine.' Such heat engines are ultimately responsible for maintaining all motions in the atmosphere against the dissipative effects of friction. The operation of the atmospheric heat engine is shown in Figure 2, which is a classic thermodynamic diagram in which temperature is plotted against specific entropy. The thermodynamic state of an air parcel – that is, its temperature, pressure, density, and so on – are represented by a point on the thermodynamic diagram, and any change of its thermodynamic state by a curve on the diagram. The area under the process curve is proportional to the heat energy entering an air parcel. The diagram also shows two different lines of constant atmospheric pressure, one near the Earth's surface and one in the tropical upper troposphere. Near the surface, air flows toward the Equator, along the segment marked AB, gaining heat from the surface (in a manner not so differently from the boundary layer inflow into the eye of a hurricane). Near the Equator, it rises almost adiabatically (i.e., with little heat entering or leaving the air) along the segment BC. It then moves poleward along segment CD, largely maintaining its temperature and angular momentum. The overturning cell is closed by a subtropical descent, as denoted by segment DA, during which the air parcel loses entropy by emitting infrared radiation to space. During this cyclic process, more heat is added to the air along AB and BC than is removed along CD and DA, with the net gain of heat being proportional to the

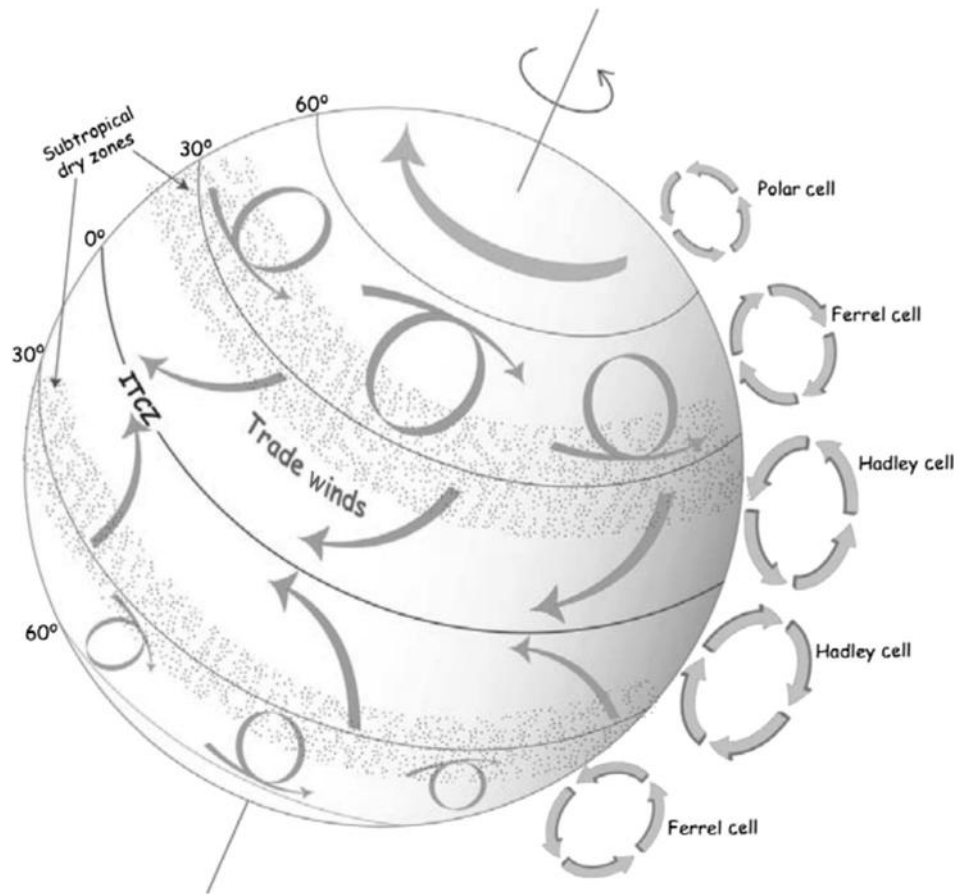


Figure 1 A schematic view of the mean circulation of the troposphere. The arrows on the globe show the winds near the Earth's surface. The cells at the side show the zonal mean circulation cells at various latitudes.

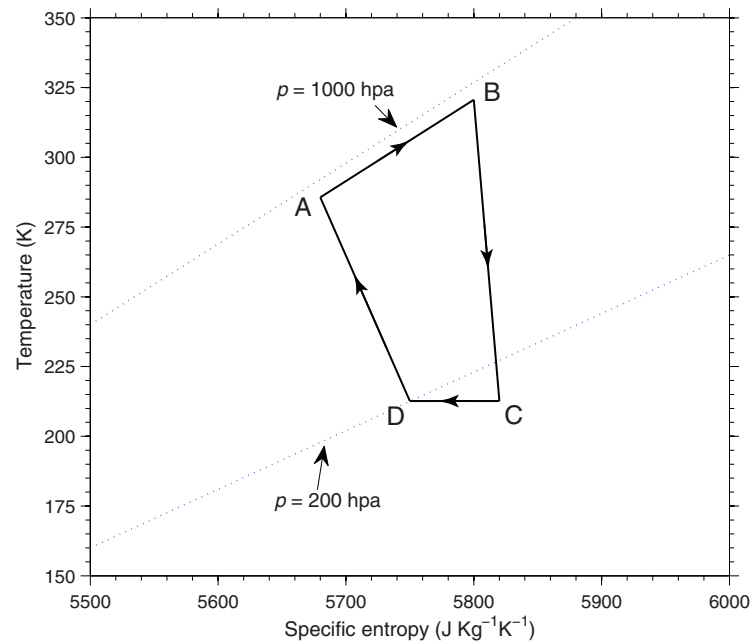


Figure 2 A schematic thermodynamic diagram for the Hadley circulation.

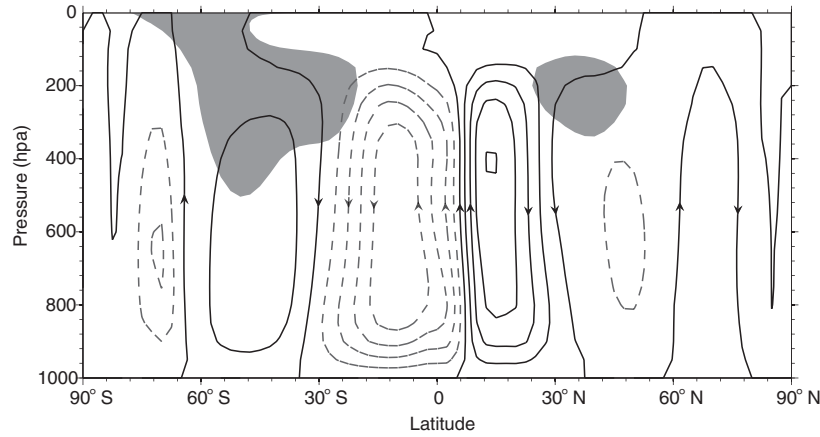


Figure 3 The annual mean meridional stream function. Contour interval is $2 \times 10^{10} \text{ kg s}^{-1}$. Shading shows zonal winds greater than 20 m s^{-1} . Based on an analysis of 44 years of ERA40 reanalysis.

area enclosed by the loop ABCD in the diagram. The excess heat is converted to mechanical energy associated with the circulation of the tropical air.

The condition for such an energy conversion to take place is that heat should on an average be added at higher pressure than it is removed. Equivalently, one can say that air must rise on average when it is warmer, and descend when it is cooler. A circulation with these properties is called a 'thermally direct circulation.' A thermally indirect circulation, in contrast, must be driven by a source of mechanical energy; a refrigerator cycle is an example of such a thermally indirect circulation. In the schematic diagram of Figure 2, the Hadley circulation is thermally direct, and therefore generates mechanical energy. In contrast, the Ferrel circulation of midlatitudes is thermally indirect and consumes mechanical energy.

The observed annual mean meridional circulation is shown in Figure 3. The contours are parallel to the northward and upward winds averaged around latitude circles and in time. The contour values have been scaled to have units of kg s^{-1} . They may be thought of as denoting the mass flux across a line from the edge of the plot to that point. The most striking feature is the strong rising motion near the Equator, and sinking motion at latitudes of about 25° N and 25° S , defining two overturning cells, the 'Hadley cells,' one in each hemisphere. However, the actual winds associated with these circulations are not particularly strong: they barely exceed 5 m s^{-1} . The diagram also reveals that there is a close relationship between the westerly component of the wind, shown by the gray shading, and the meridional flow. The westerly component is much stronger, with values up to 40 m s^{-1} . These maximum winds, which are the so-called 'subtropical jet,' are found in the upper troposphere, just where the circulations associated with the Hadley cells meet those associated with the Ferrel cells. There is also a close relationship between the zonal winds and the temperature fields: they are linked, to a very good approximation by the thermal wind relationship, which can be written as eqn [1], where u and T denote the zonal wind and temperature averaged around latitude circles, respectively; p is pressure, which decreases with height, and is often used as the vertical

coordinate for the governing equations for the atmosphere; and $R = 287 \text{ J kg}^{-1} \text{ K}^{-1}$ is the specific gas constant for dry air.

$$\frac{\partial u}{\partial p} = \frac{R}{f} \frac{\partial T}{\partial y} \quad [1]$$

That is, a strong vertical wind shear is associated with a strong poleward temperature gradient. In the deep tropics where the Coriolis parameter f is small, this relationship indicates that the temperature gradients must be small, whatever the wind field is. But in the subtropics and midlatitudes, the increasing westerly wind with height is associated with the fall of temperature toward the poles.

The Held–Hou Model

An elegant model due to Held and Hou (1980) gives considerable insight into the Hadley circulation and the factors that determine its extent. Figure 4 illustrates a two-layer representation of the model. The lower layer is affected by friction at the

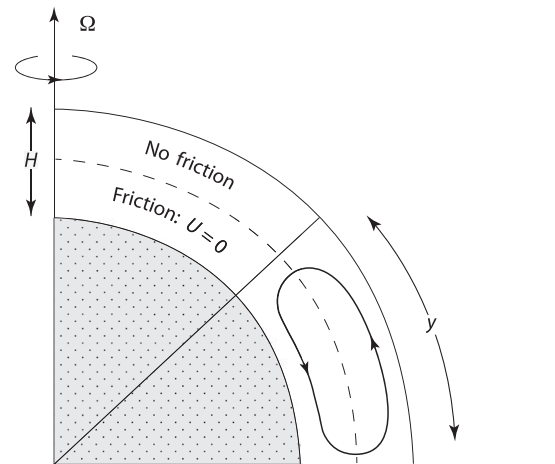


Figure 4 The configuration of the Held–Hou model.

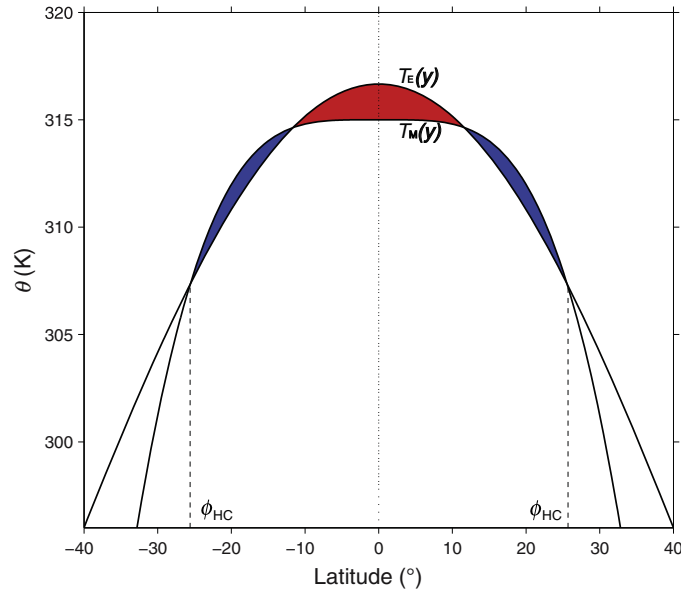


Figure 5 Solution of Held-Hou model.

ground, and flow within it is supposed to be generally small. Friction is effectively zero in the upper layer and so at this level rings of air conserve their angular momentum as they move poleward. Assuming that such rings start at the Equator with zero zonal wind relative to the solid Earth, the wind at higher latitudes in the upper layer is given by eqn [2], where Ω is the rotation rate of the Earth, a is the radius of the Earth, and y is the distance from the Equator, proportional to the latitude.

$$u_M = \frac{\Omega}{a} y^2 \quad [2]$$

Equation [2] implies that the zonal wind of an air parcel that conserves angular momentum increases rapidly with latitude as it moves poleward. Using the principle of thermal wind balance in the form of eqn [3], the formula for u_M can be used to predict the variation of temperature with latitude, $T_M(y)$.

$$\frac{\partial u_M}{\partial z} = \frac{ga}{2\Omega y} \frac{\partial T_M}{\partial y} \quad [3]$$

Substituting eqn [2] into eqn [3] and solving for T_M , one can see that the upper layer air temperature in conformity with the constraint of angular momentum conservation is proportional to the fourth power of the latitude. That is, T_M is flat near the equator and drops rapidly at subtropics. This is to be compared with the hypothetical 'radiative equilibrium' temperature distribution $T_E(y)$ of an atmosphere that is not permitted to circulate. Where the actual temperature is less than radiative equilibrium there is net heating, and vice versa. In a steady state, this heating and cooling should exactly balance in the Hadley circulation and this requirement fixes the meridional extent and strength of the Hadley circulation.

Figure 5 illustrates a graphical solution of the Held-Hou model. The actual temperature varies very little with latitude in the tropics but drops rapidly in the subtropics and midlatitudes. The radiative equilibrium temperature is maximum at the Equator. The temperature on the equator is set by requiring

that there be no net heating of air parcels as they circulate, that is, that the area shaded in red must be equal to the areas shaded in blue. The poleward limit of the Hadley circulation is at the latitude where these curves cross for the second time. A formula for the distance of the poleward edge of the Hadley cell from the equator results from this solution:

$$Y_H = a\phi_H = \sqrt{\frac{5gH\Delta T}{3\Omega^2 T_0}}, \quad [4]$$

where ΔT is equator-to-pole temperature difference, T_0 the global mean temperature, and H the vertical extent of the Hadley cell. This formula suggests a value for Y_H of about 2500 km, in good agreement with observations considering the simplicity of the model. Elaborating on the Held-Hou model, vertical motions can be estimated, which are proportional to the heating in the regions of ascent and descent. Although the model predicts a vertical circulation that is weaker than that observed, the model can be modified to incorporate the effect of latent heat release in cumulonimbus clouds, which leads to intensified but narrower ascent and broad regions of descent. The basic physical processes in the model, which predicts that the Hadley circulations are confined to 2500 km or so of the equator, remain relevant.

Seasonal Effects

The annual mean circulation shown in Figure 3 is in fact the average of two quite different circulation regimes that persist around the solstices. Figure 6 shows the circulation for the mean Northern Hemisphere winter and summer seasons. In both cases, there is a single strong thermally direct winter Hadley cell with rising motion in the summer hemisphere and descent in the winter hemisphere, while the strength of the summer cell is very much suppressed. Weaker, thermally

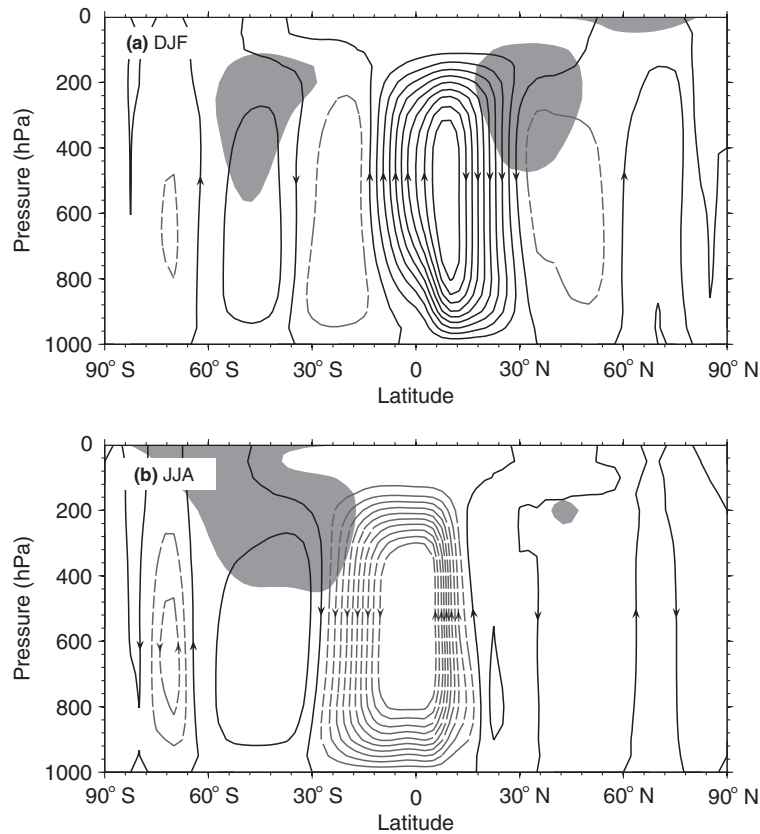


Figure 6 The mean meridional circulation for (a) the December–January–February season and (b) the June–July–August season. Other details are as for Figure 3.

indirect Ferrel cells are seen at middle latitudes in both hemispheres. Looking at the mean circulation for shorter periods reveals that the transition between a circulation like that of Figure 6(a) and one like that of Figure 6(b) is quite abrupt. At most times, there is just a single tropical Hadley cell whose circulation links the two hemispheres: at some point in the spring and autumn, its direction of circulation switches abruptly as the insolation (influx of solar energy) maximum crosses the Equator.

The Held–Hou model can be adapted to the situation where the heating is not symmetric about the Equator. Assume that the maximum radiative equilibrium temperature is no longer at the Equator, but at some latitude ϕ_0 . As well as the latitude of the northern and southern edges of the Hadley cells, the latitude ϕ_C of the streamline that divides circulation into the summer and winter hemispheres, and which is not the same as ϕ_0 , must be determined. The algebra is a little more complicated, but the steps in the argument are just the same as for symmetric Hadley cell described in the previous section.

Figure 7 shows the results. For even small ϕ_0 , the summer cell shrinks drastically and the winter cell intensifies. Almost all the circulation is associated with ascent in the summer hemisphere and with descent in the winter. The strength of the circulation is indicated by the area between the temperature curve and the radiation equilibrium curve. For ϕ_0 of only 5° , the winter cell has intensified by a factor of about 10 compared to the symmetric case, while the summer cell has weakened by

a similar factor. The winter cell is therefore some 100 times as intense as the summer cell. Such a highly nonlinear response to the latitude of the heating maximum means that the annual mean meridional circulation is much more intense than the circulation derived from the annual mean heating. This is a particularly pointed example of the problem of ‘nonlinear averaging,’ which is ubiquitous in the study of climate. This result also reconciles the weak circulations of the Held–Hou model with the stronger observed circulation: we should interpret the annual mean circulation as the average of the two solstitial circulations, not as the response to the annual mean thermal forcing.

The Eddy Mediating Effect on the Hadley Circulation

The structure of the Hadley cell is not entirely determined by the tropical heating; fluctuations in the flow (often termed ‘eddies’) also play a significant role in shaping the intensity and structure of the Hadley circulation. The momentum and heat transport by eddies acts to amplify the subtropical portion of the Hadley cell. Evidence suggests that eddies, rather than the energetic closure, have more direct relevance to the terminus of the Hadley cell. An alternative view of how the width of the Hadley cell is determined is that angular-momentum conservation of the upper tropospheric zonal wind continues poleward until the resulting vertical shears become baroclinically

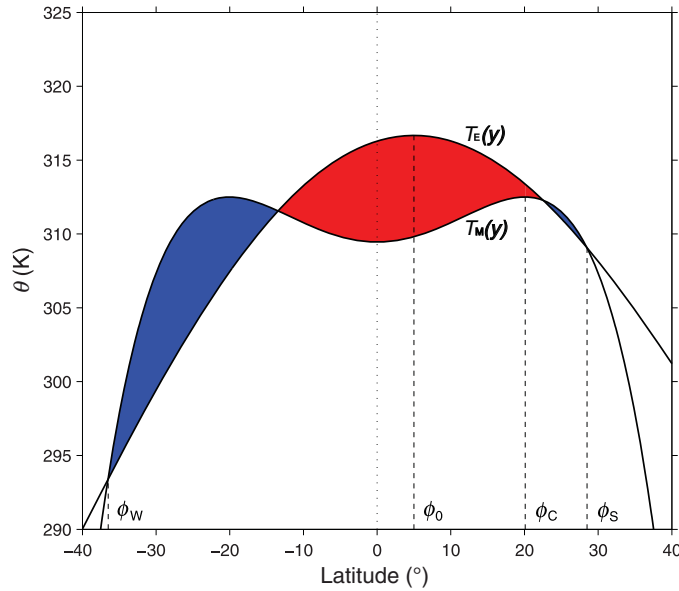


Figure 7 As Figure 5, but for a situation in which the heating maximum is located away from the Equator. ϕ_0 is the latitude of maximum radiative equilibrium temperature, ϕ_C is the latitude dividing the winter and summer Hadley cells, ϕ_W and ϕ_S designate the limits of the winter and summer Hadley cells, respectively.

unstable (Palmén and Newton, 1969; Held, 2000). If one uses the two-layer model's criterion for instability, the terminus of the Hadley cell occurs at the latitude where

$$\frac{\partial H}{\partial y} = \frac{\beta H_0}{f}, \quad [5]$$

where H is the thickness of the upper layer, indicating the tropospheric mean temperature, H_0 is the mean thickness. From thermal wind balance,

$$g^* \frac{\partial H}{\partial y} = f(u_1 - u_2), \quad [6]$$

with the low layer wind $u_2 \approx 0$, then, in the upper layer the criterion can be expressed as

$$u_1 \approx \beta \frac{g^* H_0}{f^2}, \quad [7]$$

where $g^* \equiv \frac{\rho_2 - \rho_1}{\rho_1} g$ is reduced gravity. Invoking the small angle approximation for algebraic simplicity, the terminus of the Hadley cell can be obtained by equating the baroclinically unstable wind with the angular-momentum conserving wind, that is, $u_1 = u_M$. This leads to alternative formula for the width of the Hadley cell:

$$Y_{BC} = a\phi_{BC} = \left(\frac{a^2 g H_0}{\Omega^2 \Delta_v} \right)^{\frac{1}{4}}, \quad [8]$$

where Δ_v is the fractional change in potential temperature in the vertical, representing the bulk static stability of the air column. Substituting into eqn [8] the parameters typically observed for the Earth's climate would give a Hadley cell width of 2500–3000 km, also in good agreement with the observations.

One could combine the two scaling theories for the Hadley cell width by saying that the Hadley cell stops at the smaller of

Y_{BC} and Y_H . If $Y_{BC} < Y_H$, the flow would become unstable before reaching the axisymmetric limit. If $Y_H < Y_{BC}$, on the other hand, the Hadley cell would terminate before becoming baroclinically unstable. Unfortunately, $Y_H \approx Y_{BC}$ based on the observed parameters, rendering it difficult to discern which mechanism is actually operating in the atmosphere. Evidence from analyzing observational data and data from climate models suggests that the instability-based scaling theory is more relevant to the terminus of the Hadley cell in reality. In addition, scaling relation [8] seems to capture the sensitivity of the Hadley cell width to a range of climate change perturbations. The scaling relation [8] can be easily generalized for interhemispherically asymmetric situations. See the schematic in Figure 8 for details.

Near the subtropical edge of the Hadley cell, the vertical motion is primarily maintained by the eddy momentum drag via the so-called 'eddy-pump' effect, and the edge of the Hadley cell tends to align with the transition point between divergence and convergence of the eddy momentum fluxes. Therefore, a third view interprets the Hadley cell terminus as the latitude poleward of which vertical eddy activity fluxes are sufficiently deep to reach the upper troposphere and begin to propagate meridionally, leading to the eddy momentum flux convergence/divergence. However, this midlatitude-centric view has yet to be verified against observation and more realistic atmospheric general circulation model simulations.

Hadley Circulation under Climate Change

Hadley circulation plays a key role in transporting moisture, momentum, and energy in the tropical atmosphere. It also serves as an important conduit for interhemispheric exchange and tropical–extratropical interaction between the thermally direct circulation and the eddy-driven circulation. Both the

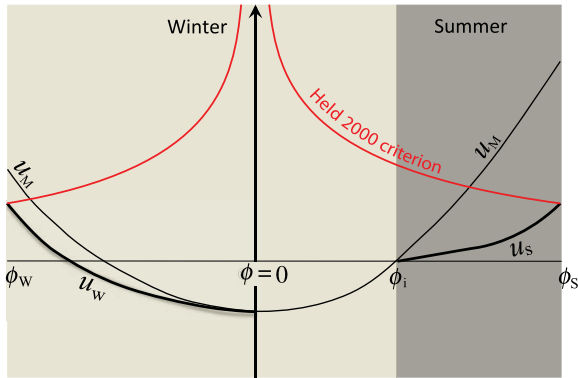


Figure 8 Schematic for a scaling relation that distinguishes between winter and summer Hadley cells. The thin solid line indicates the angular-momentum conserving zonal wind profile (u_M), obtained under the assumption of $u = 0$ at the ITCZ ϕ_i . The red lines represent the baroclinic instability criterion for two-layer model. Their intersections with the zonal wind profiles (thick lines) determine the edges of the Hadley cell. The zonal wind in summer cell deviates more from u_M compared to the zonal wind in winter cell due to the greater damping effect from the midlatitude eddies.

intensity and structure of the Hadley circulation can vary due to natural climate variability as well as external climate forcings. For example, El Niño, a climate event that occurs every few years with an anomalous sea surface temperature (SST) warming in the central and eastern equatorial Pacific, can drive an intensification and contraction of the Hadley cell. In accordance to the Hadley circulation response, the midlatitude westerly jet and storm track also shift equatorward. And vice versa for a La Niña. Climate change forcing agents, including greenhouse gases, ozone, aerosols, all can impact on the Hadley circulation.

Increasing the concentration of greenhouse gases can lead to the expansion of the Hadley cell. Since the majority of Earth's driest and arid regions are located in the areas underneath the descending branches of the Hadley circulation around 30° latitude, the expansion of the Hadley circulation can lead to significant changes in precipitation in the latitudes near the edge of the cells. As the areas around the latitudes of the Hadley cell edge become drier, those inhabiting those regions will see less rainfall than traditionally expected, which could cause major problems with water supplies and livability. The mechanisms for the global warming-induced Hadley cell expansion can be complicated. Studies suggested that the amount of the expansion seems to match the scaling relation [8] and attributed the expansion to the increase of the subtropical atmospheric stratification (i.e., Δ_v) under greenhouse warming of the climate.

In fact, we have just witnessed an expansion of the Hadley cell in both hemispheres during the last few decades of the twentieth century. The Southern Hemispheric expansion has been largely blamed on the depleting trend of the stratospheric ozone over Antarctica since the late 1970s. The stratospheric 'ozone hole' in the austral spring to summer not only shifts the jet stream and the storm tracks but also causes the austral summer Hadley cell to expand. These circulation changes result in decrease of precipitation around 45° S and increase of

precipitation around 60° S associated with the poleward-shifted storm track, as well as moistening in the subtropics over the southwestern Indian Ocean, eastern Australia, and southern flank of the Southern Pacific Convergence Zone. The causes for the Northern Hemisphere Hadley cell expansion are more complicated. Aerosols (especially black carbon), tropospheric ozone (as a greenhouse gas), and multidecadal trend of SST might all have contributed to the expansion in the Northern Hemisphere.

Since industrial revolution, the atmosphere has received aerosol pollution more strongly in the Northern Hemisphere more than the Southern Hemisphere, thus imposing an inter-hemispherically asymmetric forcing to the Earth's climate due to the dimming effect of the pollutants. In balancing this radiative perturbation, the 'thermal equator,' where the meridional energy transport of the atmosphere crosses zero, should shift toward the relatively warmed hemisphere since it now does not need as much energy input through transport as it otherwise would. As a consequence, the ITCZ as well as the Hadley cells shifted southward toward the relatively warmed hemisphere. This response, it is believed, has contributed to the devastating drought and famine throughout the Sahel nations in the 1960s–1980s. In fact, any interhemispheric asymmetric forcing for the Earth climate system, including the seasonality in the solar irradiance, orbital forcing, melting of polar ice, oceanic heat flux associated with the Atlantic meridional overturning circulation, etc., can potentially shift the Hadley cell and ITCZ.

A Lagrangian View

The diagrams of the meridional circulation shown so far have all been based on the so-called 'Eulerian averages.' That is, the winds have been averaged at fixed points in space to produce the time-mean, zonal-mean circulation. At all points in space, the winds and temperatures fluctuate to some degree as weather systems pass across the observing site. An alternative is to follow individual elements of fluid as they move around in the atmosphere, and average their properties to define a mean circulation. Such a mean is called the 'Lagrangian mean,' and in many ways is a much preferable way to describe the circulation. For example, the laws of physics applied to the atmosphere all refer to the properties of discrete, identifiable lumps of fluid. However, the Lagrangian mean is very difficult to calculate in practice, not least because individual elements of fluid rapidly become distorted and eventually thoroughly mixed with neighboring elements.

An approximation to the Lagrangian meridional mean circulation can easily be calculated, and is shown in Figure 8. In constructing this diagram, the wind data were averaged not on surfaces of constant pressure (as in Figures 3 and 6) but on surfaces of constant 'potential temperature.' The potential temperature of an air parcel generally remains more or less constant for periods of less than a few days. It follows that surfaces of constant potential temperature move up and down in response to the movement of the air. Averaging on potential temperature surfaces is equivalent, to the degree that potential temperature is indeed conserved, to taking the Lagrangian average.

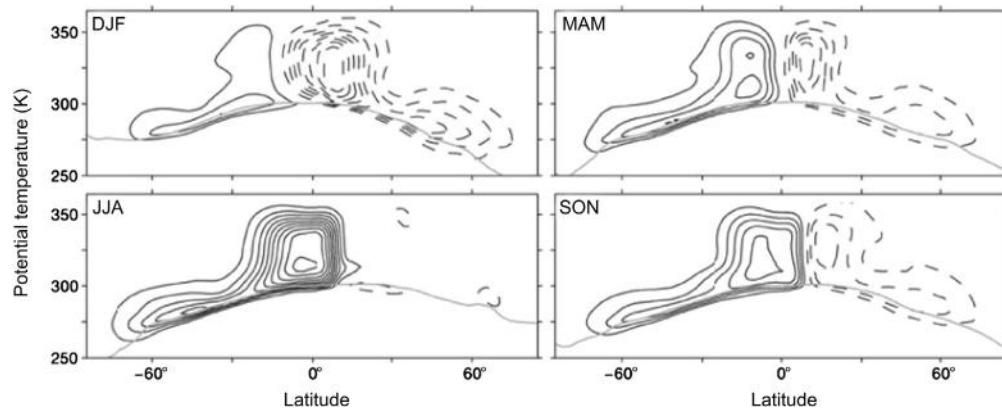


Figure 9 The mean meridional circulation for December–January–February (DJF), March–April–May season (MAM), June–July–August (JJA), and September–October–November (SON), but with the data zonally averaged on surfaces of constant potential temperature rather than on surfaces of constant pressure. Gray lines show median surface potential temperature. Contour interval is $2.5 \times 10^{10} \text{ kg s}^{-1}$. Computed from ERA40 reanalysis. Adapted from Walker, C., Schneider, T., 2006. Eddy influence on Hadley circulation: simulations with an idealized GCM. *Journal of Atmospheric Sciences* 63, 3333–3350.

Figure 9 differs dramatically from the corresponding Eulerian mean circulation shown in Figure 6. The tropical Hadley cell is still present, but the midlatitude, thermally indirect Ferrel is largely eradicated. Instead, a thermally direct circulation extends all the way from the tropics to the pole in the winter hemisphere. The original picture of the global circulation suggested by Halley and Hadley is largely vindicated if one views the circulation in Lagrangian terms.

The thermally indirect Ferrel cell actually transports heat against the temperature gradient, from high latitudes to low latitudes. At the same time, eddies more than compensate by transporting heat down the temperature gradient, from low latitudes to high latitudes. In fact, the partitioning of the flow into mean and eddy parts is arbitrary. The Lagrangian circulation, dominated by thermally direct circulations at nearly all latitudes, is a more natural and less arbitrary description. However, the Eulerian depiction of the Hadley circulation remains relevant and a useful framework to interpret some of its dynamical impacts on climate.

See also: Boundary Layer (Atmospheric) and Air Pollution: Convective Boundary Layer. **Dynamical Meteorology:** Baroclinic Instability; Coriolis Force; Lagrangian Dynamics. **General Circulation of the Atmosphere:** Energy Cycle. **Tropical Meteorology and Climate:** Intertropical Convergence Zone.

- Hadley, G., 1735. Concerning the cause of the general trade winds. *Philosophical Transactions of the Royal Society of London* 39, 58–62.
- Hadley, E., 1686. A historical account of the trade winds, and monsoons, observable in the seas between and near the tropics, with an attempt to assign the physical cause of the said winds. *Philosophical Transactions of the Royal Society of London* 16, 153–168.
- Held, I.M., 2000. The General Circulation of the Atmosphere. In: *Proceedings of the 2000 Woods Hole Oceanographic Institute Geophysical Fluid Dynamics Program*. Available from: <http://www.whoi.edu/fileserver.do?id=21464&pt=10&p=17332>.
- Held, I.M., Hou, A.Y., 1980. Nonlinear axially symmetric circulations in a nearly inviscid atmosphere. *Journal of Atmospheric Sciences* 37, 515–533.
- James, I.N., 1995. *Introduction to Circulating Atmosphere*. Cambridge University Press, Cambridge.
- Kang, S., Held, I.M., Frierson, D.M.W., Zhao, M., 2008. The response of the ITCZ to extratropical thermal forcing: idealized slab-ocean experiments with a GCM. *Journal of Climate* 21, 3521–3532.
- Kang, S., Lu, J., 2012. Expansion of the Hadley cell under global warming: winter versus summer. *Journal of Climate* 25, 8387–8393.
- Korty, R.L., Schneider, T., 2008. Extent of Hadley circulation in dry atmospheres. 35. <http://dx.doi.org/10.1029/2008GL035847>.
- Lindzen, R.S., Hou, A.Y., 1988. Hadley circulations for zonally averaged heating centered off the equator. *Journal of Atmospheric Sciences* 45, 2417–2427.
- Lu, J., Vecchi, G.A., Reichler, T., 2007. Expansion of the Hadley cell under global warming. *Geophysical Research Letters* 34. <http://dx.doi.org/10.1029/2006GL028443>.
- Palmén, E., Newton, C.W., 1969. *Atmospheric Circulation Systems*. Academic Press, New York, 603.
- Pixóto, J.P., Oort, A.H., 1992. *Physics of Climate*. American Physical Society, New York.
- Walker, C., Schneider, T., 2006. Eddy influence on Hadley circulation: simulations with an idealized GCM. *Journal of Atmospheric Sciences* 63, 3333–3350.

Further Reading

- Fang, M., Tung, K., 1999. Time dependent nonlinear Hadley circulation. *Journal of Atmospheric Sciences* 56, 1797–1807.
- Frierson, D.M.W., Lu, J., Chen, G., 2007. The width of the Hadley cell in simple and comprehensive general circulation models. *Geophysical Research Letters* 34. <http://dx.doi.org/10.1029/2007GL031115>.

Intertropical Convergence Zone

DE Waliser, California Institute of Technology, Pasadena, CA, USA

X Jiang, University of California, Los Angeles, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

The Intertropical Convergence Zone (ITCZ) lies in the equatorial trough, a permanent low-pressure feature where surface trade winds, laden with heat and moisture, converge to form a zone of increased convection, cloudiness, and precipitation. The latent heat released in the ITCZ is critical to the atmospheric energy budget and ITCZ cloudiness provides an important contribution to the planetary albedo. The ITCZ's position, structure, and migration influence ocean–atmosphere and land–atmosphere interactions on a local scale, the circulation of the tropical oceans on a basin scale, and a number of features of the Earth's climate on a global scale.

Introduction

One of the features that is most readily identified with the tropical atmosphere is the Intertropical Convergence Zone (ITCZ). The ITCZ lies in the equatorial trough, a permanent low-pressure feature that marks the meteorological equator where surface trade winds, laden with heat and moisture from surface evaporation and sensible heating, converge to form a zone of increased mean convection, cloudiness, and precipitation. The latent heat released in the convective cloud systems of the ITCZ is a critical component of the atmospheric energy balance, and the enhanced cloudiness associated with these cloud systems provides an important contribution to the planetary albedo. The fluxes of heat, moisture, momentum, and radiation between the atmosphere and the surface differ dramatically between the ITCZ region and the regions to the north and south of the ITCZ. Thus, the position, structure, and migration of the ITCZ play an important role in determining the characteristics of ocean–atmosphere and land–atmosphere interactions on a local scale, the circulation of the tropical oceans on a basin scale, and a number of features of the Earth's climate on a global scale.

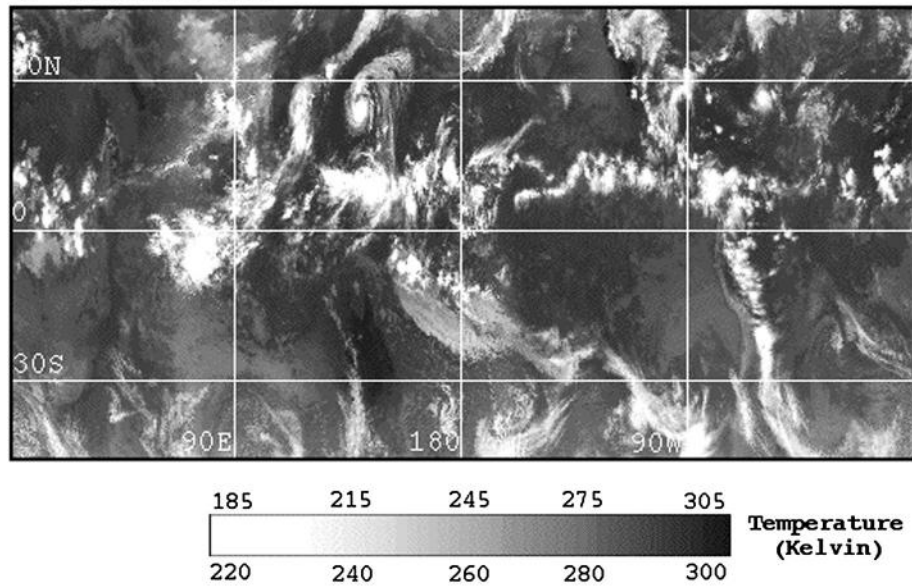
Mean Structure

On any given day in the tropics, there are usually a number of deep convective cloud systems that appear to be somewhat randomly distributed across the equatorial region. **Figure 1(a)** shows a satellite cloud image constructed from a number of geostationary and polar orbiting satellites for 7 September 1991. Bright areas denote cold temperatures and thus in this case indicate clouds whose tops are at or near the level of the tropopause, e.g., deep convective or cirrus clouds. Dark areas denote warm temperatures, which in this case implies clear skies. Evident throughout the tropical region, and aligned roughly parallel to the equator, are a number of cloud systems. Some of these systems exhibit horizontal scales on the order of a few hundred kilometers or less. Others, such as the large system in the Indian Ocean, have horizontal dimensions on the order of about 25 000 km. The vast difference in horizontal scales of these cloud systems can arise from a number of factors. Typically, the more mature a convective system is, the

larger its horizontal extent. This is mostly due to the development of high cirrus clouds in the outflow region of convective systems. In contrast to deep convective 'towers,' which typically have horizontal scales on the order of 1–10 km and when found in isolation usually indicate young or developing convective system, cirrus clouds can appear to extend over thousands of kilometers and encompass tens or hundreds of convective towers simultaneously. Thus, the size of the various convective systems shown in **Figure 1(a)** can be influenced by their maturity and their abundance in any one area and how these factors in turn influence the development of what appears as a common cirrus cloud. Furthermore, while most cloud systems in the tropics arise from simple convective instability, likely in conjunction with synoptic wavelike disturbances inherent to the equatorial region, the organization of some systems can be influenced by low-frequency phenomena which can increase their spatial extent. For example, the larger systems in the Indian and western Pacific Ocean may be influenced by the tropical Intraseasonal Oscillation or simply be larger due to the difference in climatological conditions (e.g., SST) between the Eastern and Western Hemispheres which is discussed below.

Other than the loose east–west orientation of the cloud systems in **Figure 1(a)**, there is no obvious systematic preference for the locations of these systems. Only upon averaging such observations over a time period relatively long (~months) compared to the lifetime of these systems (~hours–days) does a robust spatial preference become evident. **Figure 1(b)** shows a time-averaged satellite cloud image, constructed from daily cloud images, such as the one shown in **Figure 1(a)**, from 1 September to 31 November 1991. From this image, it is more apparent that for a given season particular regions of the tropics are favored for the development of tropical convective systems. The spatial structure of the deep convective cloud pattern shown in **Figure 1(b)** exhibits the spatial pattern roughly identified with the ITCZ, at least for the Northern Hemisphere fall season. Thus, while the cloud (or rainfall) pattern associated with the ITCZ is usually thought of as a continuous band of clouds (or rain), at any given time, this 'band' contains only a few disparate cloud systems. **Figure 2** shows the long-term mean rainfall pattern. The band(s) of high rainfall represent the mean, or archetypal, ITCZ spatial structure. Overall, this structure is roughly aligned

(a) 7 September 1991



(b) September–November mean

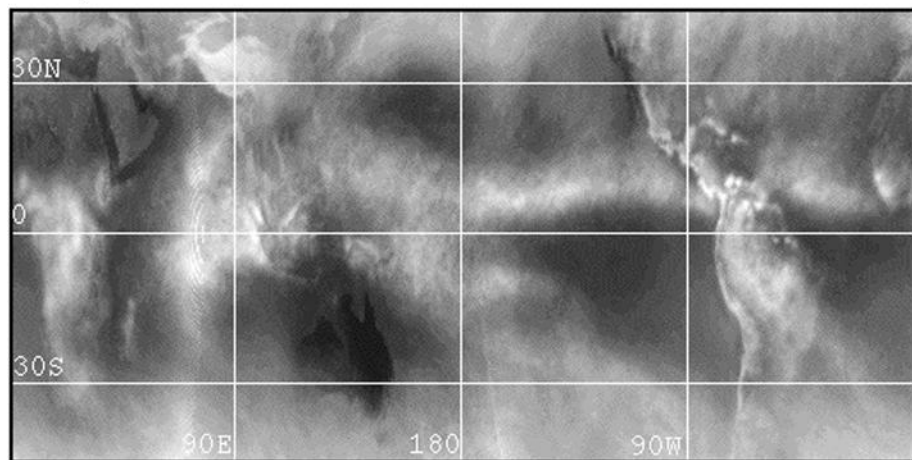


Figure 1 Cloud images constructed from 11- μm radiances measured from a number of geostationary and polar-orbiting weather satellites. (a) Instantaneous cloud field for 00 GMT 7 September 1991. (b) Time-mean 00 GMT cloud field for September through November 1991. Radiance values have been converted to equivalent blackbody temperature using the Stefan–Boltzman Law. Global Cloud Imagery courtesy of M. Salby, University of Colorado.

with the equator and it exhibits a significant amount of zonal symmetry relative to the rainfall pattern in the mid-latitudes. Apart from this zonal symmetry, the ITCZ rainfall distribution displays a fair amount of longitudinal variability as well. In the Atlantic and eastern Pacific Oceans, it is made up of very narrow, intense regions of rainfall that tend to lie just north of the equator. Over the South American and African continents, the mean rainfall distribution has a considerably larger latitudinal extent and tends to lie directly over the equator. Over the eastern Indian and western Pacific Oceans, the rainfall distribution is both broad in latitude and intense in magnitude. Two of the more notable zonal asymmetries in the ITCZ are the weak rainfall over the western Indian Ocean and the southeast extension of the ITCZ over the central Pacific Ocean. The latter, referred to as the South Pacific Convergence Zone (SPCZ), leads to an area of intense rainfall on either side of the equator with

a relatively dry region in between. Such a structure is often referred to as a ‘double ITCZ.’

The time-mean spatial structure of the ITCZ described above can be better understood by examining the time-mean sea surface temperature (SST), which is shown in [Figure 3](#). Due to the fact that on average the equatorial region receives the most solar irradiance, this region also tends to have the highest SSTs. While this tendency for very warm SST (i.e., greater than 25 °C) is mostly uniform with longitude, there are some deviations. These deviations are produced by a number of factors, including ocean basin geometry, ocean circulation properties – equatorial dynamics in particular, as well as the coupled interaction with the atmosphere, including the ITCZ itself. The relatively warm water of the equatorial region heats the air in the lower atmosphere making it less dense and buoyant relative to the air aloft. This buoyancy forcing leads to rising motion

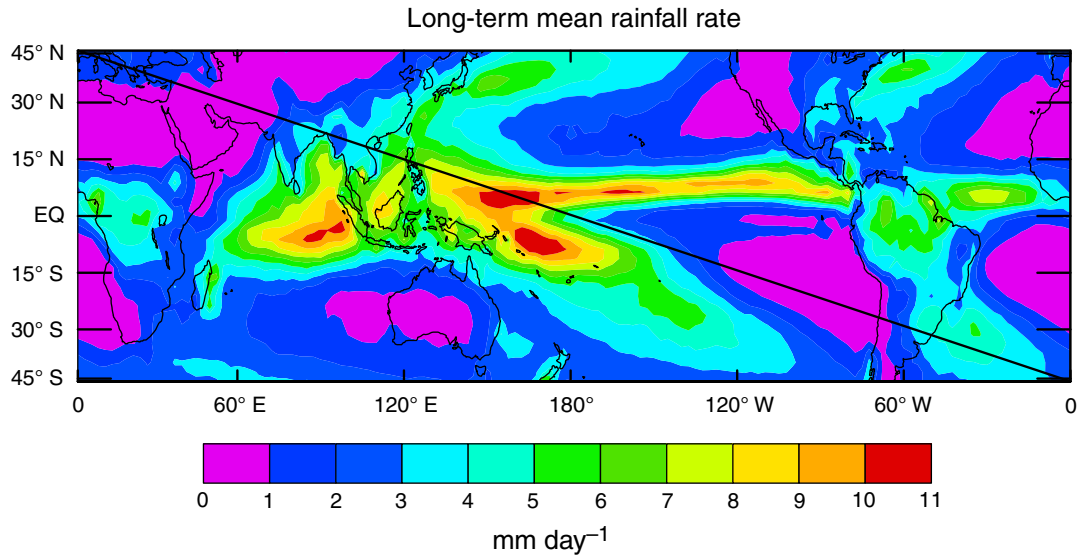


Figure 2 Long-term mean (1979–96) rainfall rate constructed from a combination of satellite-derived products and *in situ* observations. Source: National Oceanographic and Atmospheric Administration's (NOAA) Climate Prediction Center (CPC) Merged Analysis of Precipitation (CMAP).

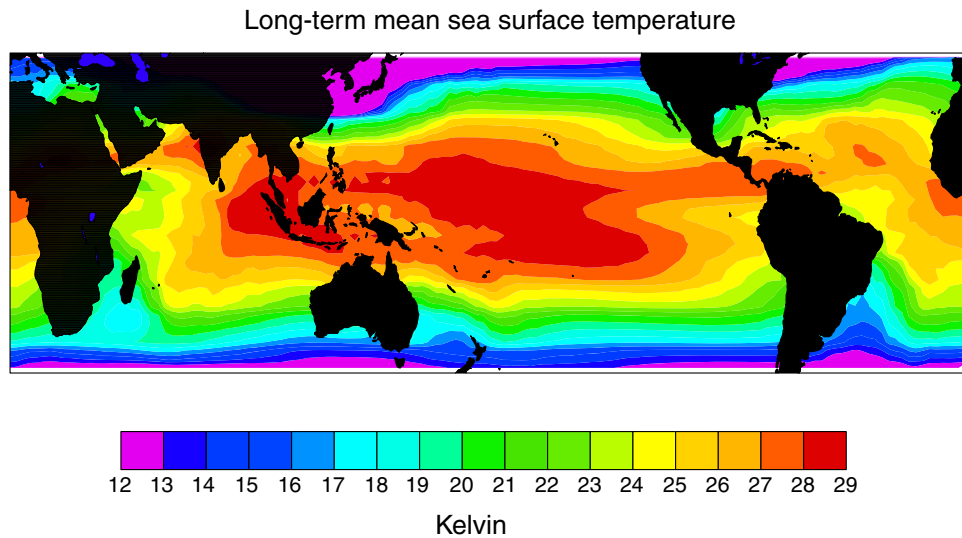


Figure 3 Long-term mean (1968–96) sea surface temperature (SST) constructed from a combination of satellite-derived values and *in situ* observations. NOAA's National Environmental Satellite, Data, and Information Service.

over the equatorial region. As the moist near-surface air rises, it cools adiabatically and begins to undergo condensation which releases the latent heat contained in the water vapor and produces rainfall at the surface. This latent heating enhances the buoyancy and associated upward motion of the air even further, which in turn enhances the adiabatic cooling, water vapor condensation, and surface rainfall. This process continues until nearly all the water vapor condenses out of the parcel and/or the parcel is no longer buoyant with respect to its environment. In either case, this usually happens when the parcel reaches the inversion associated with the tropopause, whereupon the air begins to move away from the equator. This divergent upper level air undergoes cooling through radiative

heat loss, causing it to lose buoyancy, and sinks in the subtropical regions.

Figure 4 shows the long-term mean outgoing thermal radiation leaving the top of the atmosphere. The regions of large radiative heat loss lie on the poleward edges of the regions of high rainfall (i.e., **Figure 2**) and extend into the subtropics. Upon reaching the surface, this sinking air is relatively dry but gains moisture again via surface evaporation as it converges toward the equator. **Figures 5 and 6** show the long-term mean surface wind and evaporation fields. The surface wind field shows that over most of the tropical regions surface air tends to converge into the areas of high rainfall. The evaporation field indicates that as this air converges toward these equatorial

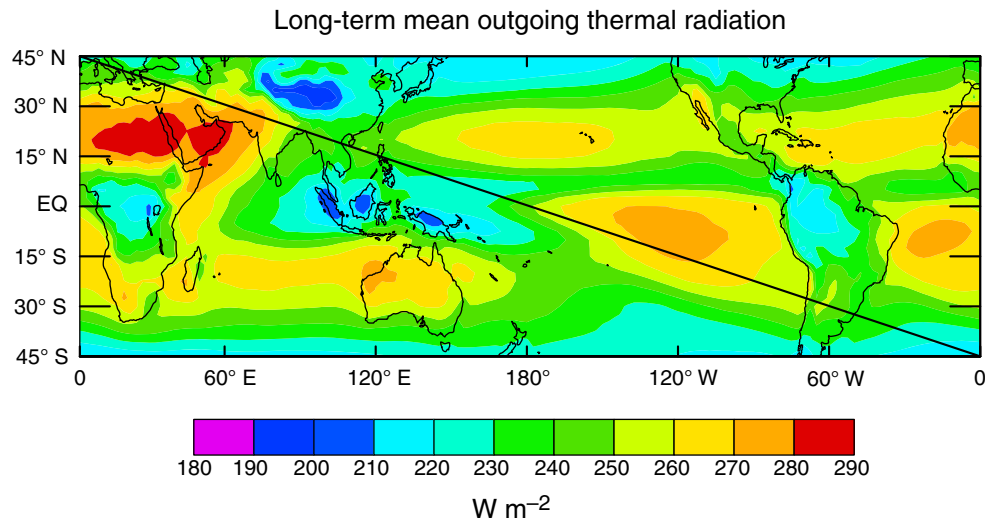


Figure 4 Long-term mean (1974–98) satellite-derived outgoing thermal radiation. Source: NOAA's National Environmental Satellite, Data, and Information Service.

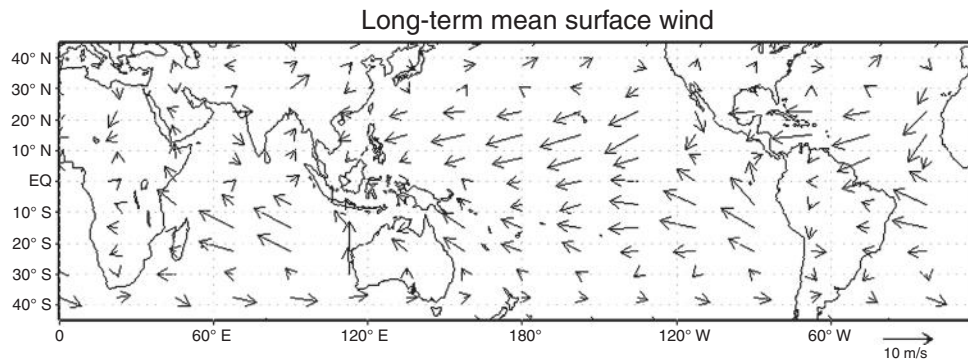


Figure 5 Long-term mean (1968–96) surface wind values derived from a combination of *in situ* observations and analyses using a numerical weather prediction model. Source: National Centers for Environmental Prediction/National Center for Atmospheric Research Reanalysis Project.

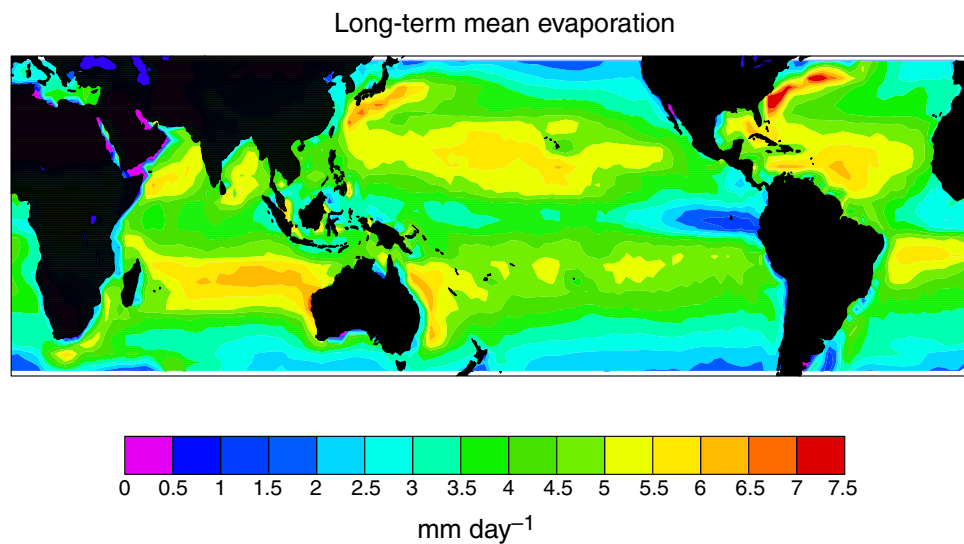


Figure 6 Long-term mean (1968–96) ocean surface evaporation values constructed from a combination of *in situ* values and analyses using a numerical weather prediction model. Source: National Centers for Environmental Prediction/National Center for Atmospheric Research (NCEP/NCAR) Reanalysis Project.

regions, it gains moisture from the ocean, particularly in the areas of the 'trade winds' where the wind speeds are higher. The combination of the above processes leads to a deep meridional circulation cell, extending over the depth of the troposphere, with air converging toward the equator at low levels, rising in the equatorial regions, diverging at upper levels, and sinking in the subtropics. The zonal mean of this circulation pattern is typically referred to as the 'Hadley Circulation.'

From the physical description above, it is evident that a close association exists between the spatial structure of the warmer SSTs and the rainfall pattern associated with the ITCZ. For example, **Figures 2 and 3** show that the narrow bands of rainfall over the Atlantic and eastern Pacific correspond well to the relatively warm bands of warm water north of the equator in these regions. Similarly, the very broad area of warm water in the Indian and western Pacific Oceans corresponds well to the more widespread area of intense rainfall of the ITCZ in these regions. In addition, the discussion above highlights the complex makeup of the water and energy cycles in the tropics and the role of the ITCZ within these cycles. The schematic diagram in **Figure 7** highlights important aspects of these water and energy cycles and illustrates how the physical processes associated with the ITCZ described above fit together in an idealized latitude-height diagram. The downward arrows at the top of the atmosphere depict the incoming solar energy from the sun and the fact that there is a reduction of solar energy as one moves poleward. As **Figure 1(a)** indicates, at any given time, most of the tropics exhibit clear skies. This allows a large portion of this solar energy to reach the surface and induce a pole-to-equator SST gradient, with the warmest SSTs in the near-equatorial region. The upward arrows at the surface of the ocean depict the ocean to atmosphere energy exchange, which takes place primarily through the transfer of heat and moisture from the ocean to the near-surface air via sensible heat and latent (i.e., evaporative) heat fluxes. As the air rises over the warmest water, a convergent circulation is induced at lower levels with the upper levels exhibiting divergence. The rising air experiences adiabatic cooling which leads to condensation of the moisture and the release of the stored latent heat. The

former falls back to the surface as precipitation, while the latter heats the air, leading to a further enhancement of the vertical motion. Now, the heat that was originally derived from incoming solar energy deposited in the ocean resides in the atmosphere. The upward arrows at the top of the atmosphere denote transfer of this energy back to space via radiative heat loss as the air diverges away from the equator and sinks back to the surface.

Seasonal Variations

Over the course of the annual cycle, seasonal changes occurring in the ITCZ modify the mean structure depicted in **Figure 2**. In general, the entire line-oriented convection band marches north in the Northern Hemisphere spring and summer and south in the Southern Hemisphere spring and summer. The differences in the amplitudes and phases of the ITCZ excursions at different longitudes are dictated in part by the different characteristics of the surface (i.e., land or ocean) and the local atmospheric circulation pattern. The ITCZ over land (e.g., Africa and South America) follows the annual march of the sun, while the migration of the ITCZ over extended ocean regions lags slightly behind by a month or two. This time lag is most apparent in the eastern Pacific and the Atlantic Oceans, where the ITCZ is furthest south in the Northern Hemisphere spring and furthest north in the Northern Hemisphere fall. The origin of this time lag is primarily due to the large thermal inertia of the ocean mixed layer compared to the land surface but also involves complex dynamical interactions that develop between the ocean and atmosphere.

While most of the seasonal changes in the ITCZ are associated with latitudinal migration, there are other significant structural changes. One of the more significant of these is over South America where large spatial differences exist between the 'ITCZs' of the Northern and Southern Hemisphere summers. During the Southern Hemisphere summer, the rainy season encompasses nearly the entire tropical area of the South American continent. This produces a latitudinally and

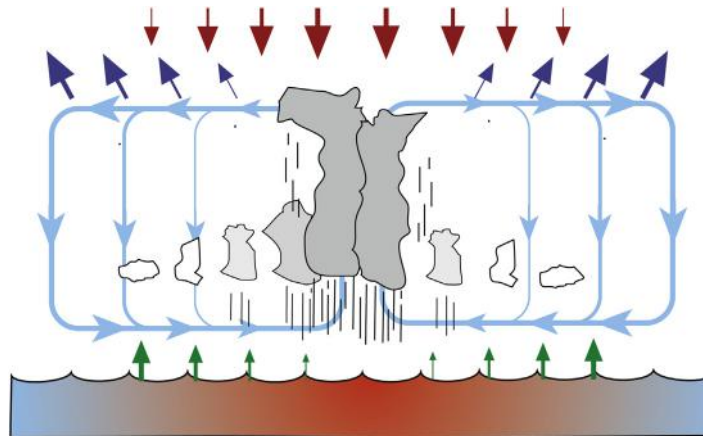


Figure 7 Schematic depiction of the ITCZ within the context of the water and energy cycles of the tropics. The downward arrows at the top of the atmosphere depict the incoming solar energy from the sun. The upward arrows leaving the surface of the ocean depict the transfer of heat and moisture from the ocean to the near-surface air via sensible heat and latent (i.e., evaporative) heat fluxes. The upward arrows at the top of the atmosphere denote this energy being transferred back to space via radiative heat loss.

longitudinally broad ITCZ. In the Northern Hemisphere summer, the ITCZ overlies the oceanic region north of the continent and has a structure more consistent with its oceanic counterparts to the east and west. Another dramatic seasonal change associated with the ITCZ occurs in the Indian Ocean region during the Asian summer monsoon. As the monsoon circulation develops and intensifies, the convection zone splits, with a very intense area of rainfall occurring over the Indian subcontinent and a weaker rainfall maximum remaining in the equatorial region. A similar intensification of rainfall occurs in the Southern Hemisphere summer over northern Australia; however, in this case, the equatorial component tends to be suppressed. During this same period, the convergence zones over Africa and the Indian Ocean become more continuous due to the reduced coastal ocean upwelling off the east African coast. Other modest seasonal deviations occur in the eastern Pacific during the northern spring, when the ITCZ occasionally separates into two zones of convection straddling the equator. This 'double ITCZ' results from a relaxation of the southeast trade winds, which greatly diminishes the equatorial and nearby coastal ocean upwelling leaving seasonably warm surface temperatures south of the equator. In a related manner, the two branches of convergence in the central Pacific oscillate in strength during the year with the southern and northern branches intensifying during their respective summer season.

To help illustrate and quantify some of the seasonal changes in the ITCZ described above, **Figure 8** shows time-latitude diagrams of rainfall over the course of the calendar year for a number of distinct tropical regimes. For example, **Figure 8(a)** shows that the annual cycle of the ITCZ over Africa exhibits a migration pattern that has a nearly sinusoidal nature. In this region, the ITCZ appears to closely follow the solar cycle of surface heating, with a lag of about a month. It has a fairly even intensity throughout the year ($\sim 5 \text{ mm day}^{-1}$) and migrates from about 15° S to 10° N . The difference in poleward extremes is associated with the Sahara Desert, the dryness of which inhibits the northern migration of the ITCZ. In contrast to this nearly sinusoidal case, all other regions (**Figure 8(b)** and **8(g)**) show an annual cycle that has seasonal dependencies in intensity, structure, and/or a larger phase lag relative to the solar cycle of surface heating. For example, the ITCZ migrations over the Indian (**Figure 8(b)**) and western Pacific (**Figure 8(c)**) Oceans show strong rainfall intensification associated with the summer monsoons. In particular, the Asian summer monsoon produces a significant enhancement to the rainfall in the Northern Hemisphere summer months over Southeast Asia and, as mentioned above, produces two bands of rainfall in the Indian Ocean region during this period. The annual cycle of the ITCZ in these two regions lags approximately 2 months behind the solar heating cycle.

The eastern Pacific (**Figure 8(e)**) and Atlantic (**Figure 8(g)**) Oceans have very similar annual cycles. As suggested earlier by **Figure 1(b)** and **Figure 2**, the ITCZ in these regions remains primarily in the Northern Hemisphere throughout the year, with some weak rainfall ($\sim 4\text{--}5 \text{ mm day}^{-1}$) occurring south of the equator in the Northern Hemisphere spring. During this time of year, warm water ($\sim 27^\circ \text{ C}$ or greater) occurs on both sides of the equator in this region and the ITCZ, in its southern most position, is split by a zone equatorial ocean upwelling (i.e., cool equatorial SSTs). The phase of the annual cycle in

these regions lags behind the surface solar heating cycle by approximately 2–3 months, and each produces the most intense ITCZ in the Northern Hemisphere fall. During this season, the surface water associated with the equatorial countercurrents is warmest and the low-level trade wind convergence is strongest.

The annual cycle of the central Pacific and South America shows very different characteristics than those described above. As illustrated earlier in **Figure 2**, the ITCZ in the central Pacific is composed of northern and southern convergence zones straddling the equator. While this large-scale 'double' convergence zone remains intact during the course of the annual cycle, the intensity of the summer hemisphere branch tends to dominate. The annual cycle of the ITCZ over South America displays the least amount of symmetry with respect to north–south migration and ITCZ intensity. The surface underlying the ITCZ is largely responsible for this asymmetry as mentioned above. In this region, the phase of the ITCZ is locked to the annual cycle during its northward propagation. However, after having reached the oceanic region north of South America, the convection diminishes slightly, and the cycle appears to lag slightly until the rainy season begins again over the Amazon Basin in November–December. The annual cycle of the global ITCZ has a modest resemblance to a sinusoidal pattern, with the intensity of the zonally averaged mean rainfall being strongest during the Northern Hemisphere summer and fall ($\sim 7 \text{ mm day}^{-1}$) and weakest during the equinoxes ($\sim 5 \text{ mm day}^{-1}$).

Interannual Fluctuations

Apart from the regular seasonal variations, the ITCZ undergoes interannual fluctuations in its position and intensity. **Figure 9** illustrates the range of interannual variability exhibited by the ITCZ over the period 1979–98 for three of the longitude sectors discussed in the previous section. This figure shows time-latitude diagrams of the seasonal anomalies of rainfall from the mean annual cycles presented in **Figure 8**. Note that each is plotted using the same color scale. Immediately evident is the fact that the interannual anomalies in the ITCZ position and intensity are weakest over Africa and strongest for the ITCZ over the central Pacific Ocean. Typical seasonal rainfall anomalies for the ITCZ over Africa are about $\pm 0.5 \text{ mm day}^{-1}$ and range up to about $\pm 1.5 \text{ mm day}^{-1}$ in the more extreme events. Depending on the location and intensity of the mean ITCZ rainfall band, these values represent variations on the order of 10–25% of the mean values. Overall, these anomalies illustrate that this region undergoes relatively weak, low-frequency rainfall variations with the early and late 1980s being relatively wetter than normal and the late 1970s and early 1990s being relatively drier than normal. Within this low-frequency variability are periods where the ITCZ exhibits short-lived variability in its intensity and latitude. For example, during the 1981–82 winter, the ITCZ extended anomalously southward, while in the 1982–83 winter, the rainfall associated with the ITCZ was about 25% stronger than normal. During the fall of 1986 and summer of 1987, the ITCZ extended anomalously northward, bringing rain to the normally dry Sahara Desert. Another notable anomaly in the ITCZ over Africa occurred in

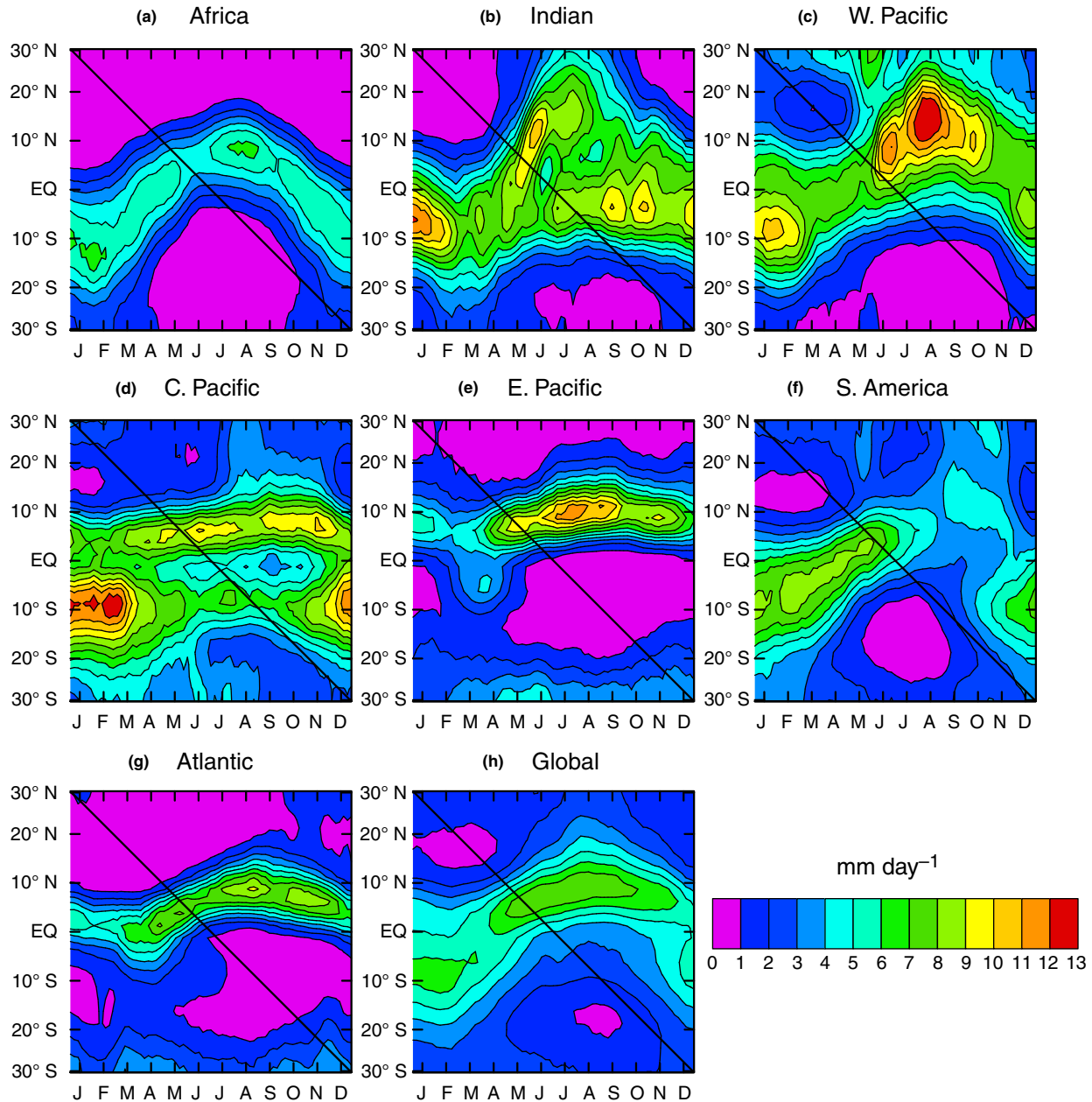


Figure 8 Time-latitude diagrams of the annual cycle of the ITCZ, in terms of rainfall, zonally averaged over eight different longitude sectors. (a) Africa, 10–40° E; (b) Indian, 60–100° E; (c) W. Pacific, 110–150° E; (d) C. Pacific, 160° E–60° W; (e) E. Pacific, 100–140° W; (f) S. America, 45–75° W; (g) Atlantic, 10–40° W; and (h) global, 0–360° E. Mean annual cycles were computed for the period from 1979 to 1996. Source: National Oceanographic and Atmospheric Administration's (NOAA) Climate Prediction Center (CPC) Merged Analysis of Precipitation (CMAP).

the winter of 1991–92 when the ITCZ was particularly weak and did not migrate as far south as normal.

The time-latitude rainfall anomaly diagram for the Indian Ocean region (Figure 9b) shows variability quite different than that exhibited by the ITCZ over Africa. First, the meridional extent of the anomalous excursions is much greater, extending to at least 30°. For the most part, this is simply related to the much broader latitudinal extent of the mean ITCZ pattern in this region (i.e., Figures 2 and 8). Second, the intensity of the fluctuations is slightly greater, ranging up to about $\pm 3 \text{ mm day}^{-1}$. However, given the larger mean rainfall

values for this region, this anomaly range represents deviations from the annual cycle on the order of 25% which is similar to the case for Africa. Third, even with the seasonal smoothing applied to the data, the ITCZ in this region exhibits considerably more variability at shorter time scales than for example the ITCZ over Africa. This shorter term variability is partly attributable to the Intraseasonal Oscillation that has been found to prevail most strongly over the Indian and western Pacific Oceans. Given the distribution of land and people within this sector, the most consequential of the rainfall anomalies occur in the Northern Hemisphere summer, north of about 15° N.

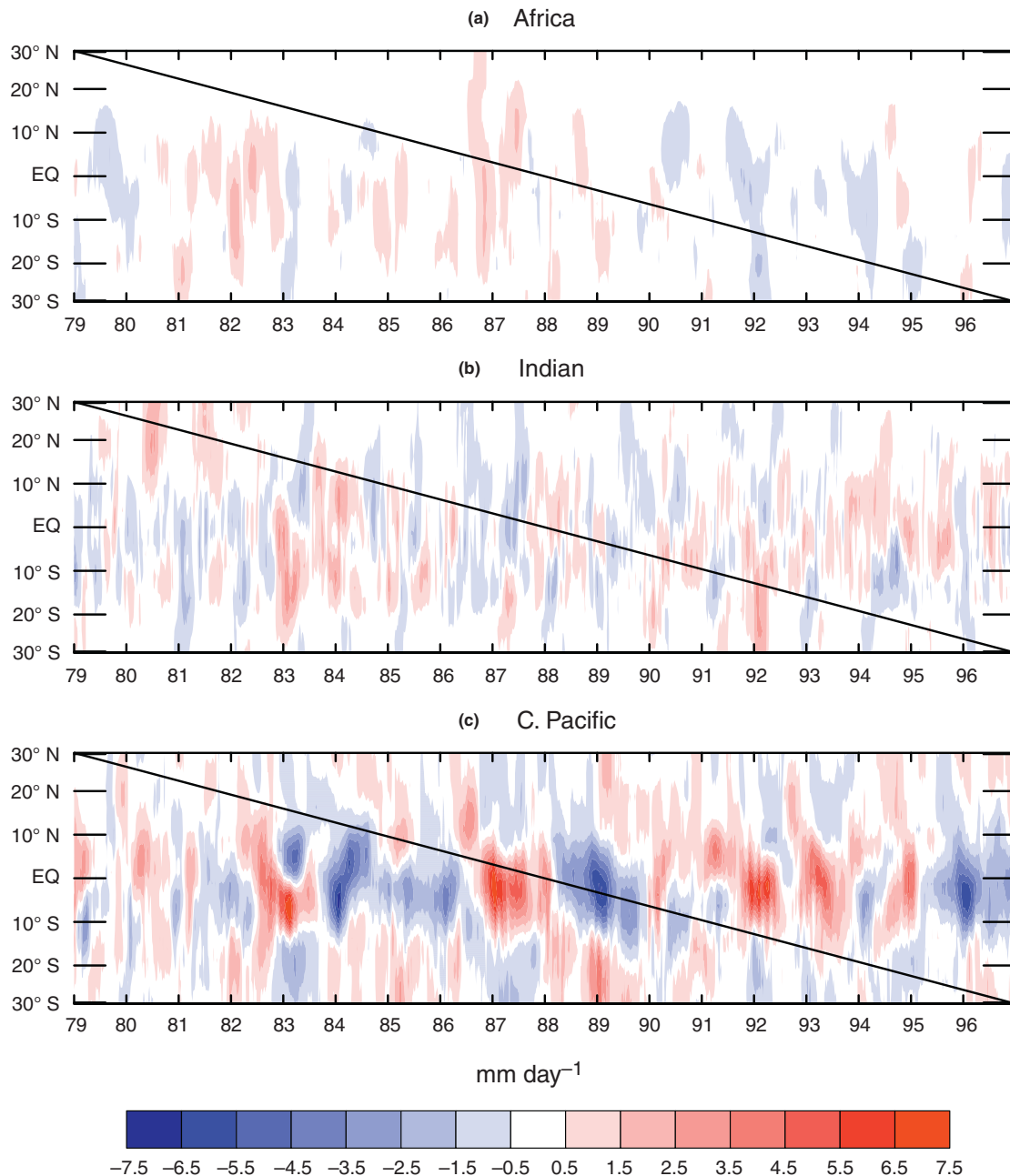


Figure 9 Time-latitude diagrams of the anomalies in rainfall associated with the ITCZ zonally averaged over three different longitude sectors: (a) Africa, 10–40° E; (b) Indian, 60–100° E; and (c) C. Pacific, 160° E–160° W. Departures are computed from the mean annual cycles shown in Figure 8. Source: National Oceanographic and Atmospheric Administration's (NOAA) Climate Prediction Center (CPC) Merged Analysis of Precipitation (CMAP).

From Figure 7, it can be seen that the maximum rainfall associated with the Indian summer monsoon occurs in July. Figure 8(b) illustrates some aspects of the variability associated with the timing and strength of the Asian monsoon as it relates to the ITCZ. For example, in 1980, the rainfall associated with the northward migration of the ITCZ and the development of the summer monsoon was particularly intense, while the 1981 summer monsoon appears to have come slightly earlier than normal. In contrast, the summer monsoons of 1983, 1987, and 1989 are examples of the monsoon-related ITCZ rainfall being

a bit weaker than normal. It is important to emphasize out that even though these anomalies ($\sim 1 \text{ mm day}^{-1}$) only represent about 10% of the total rainfall that typically occurs during the monsoon (Figure 8(b)), they represent very important departures for the people and industries (e.g., agriculture) that are affected by them.

For the case of the central Pacific (Figure 9c), the anomalous ITCZ rainfall is dominated by negative and positive anomalies on or near the equator. Note that these values are significantly larger than the ITCZ rainfall anomalies occurring

in either of the other longitude sectors discussed above (including those longitude sectors not discussed). In this region, the zonally averaged rainfall anomalies range up to at least $\pm 7 \text{ mm day}^{-1}$. In some instances, particularly for the large anomalies right on the equator, these values can exceed 100% of the mean values associated with the annual cycle (Figure 7(d)). These large variations in the position and intensity of the ITCZ in this region are associated with climate phenomena known as the El Niño – Southern Oscillation (ENSO). In warm phases of ENSO (i.e., El Niño), SST in the central and eastern equatorial Pacific Ocean can become anomalously warm by about $1\text{--}3^\circ\text{C}$, while in cold phases (i.e., La Niña), this region becomes anomalously cool by a similar magnitude. This has a dramatic effect on the organization of tropical convection in this region as well as in regions remote to the central Pacific. Evident from Figure 9(c) are the large positive rainfall anomalies associated the strong El Niño of 1982–83, moderate El Niño of 1986–87, and the prolonged and somewhat weaker El Niño(s) of the early 1990s. Typically, these events cause the two convergence zones which are located

slightly away from the equator in the central Pacific to merge into a single zone of convection and rainfall centered on or very near the equator. The large negative equatorial rainfall anomalies are associated with the La Niñas of 1984, 1988–89, and 1995–96. Central Pacific rainfall anomalies associated with La Niña typically cause the relatively dry equatorial zone near the dateline (Figure 2) to become even drier and to extend further west than normal. It is important to point out that these central Pacific ENSO-related ITCZ rainfall anomalies do not occur in isolation. Typically, these anomalies induce anomalies of the opposite sense and a somewhat weaker magnitude in the western Pacific sector and in some cases the Indian Ocean and South American sectors as well. In fact, these ENSO-related rainfall anomalies are so large that they are the only significant departure that appears to occur in the global mean ITCZ (data not shown). Anomalies in the global mean ITCZ rainfall are typically in phase with the anomalies in the central and eastern Pacific Ocean and have magnitudes that range up to about 1 mm day^{-1} . However, with respect to quantifying the size of this climate signal in the context of the ITCZ, it is important to

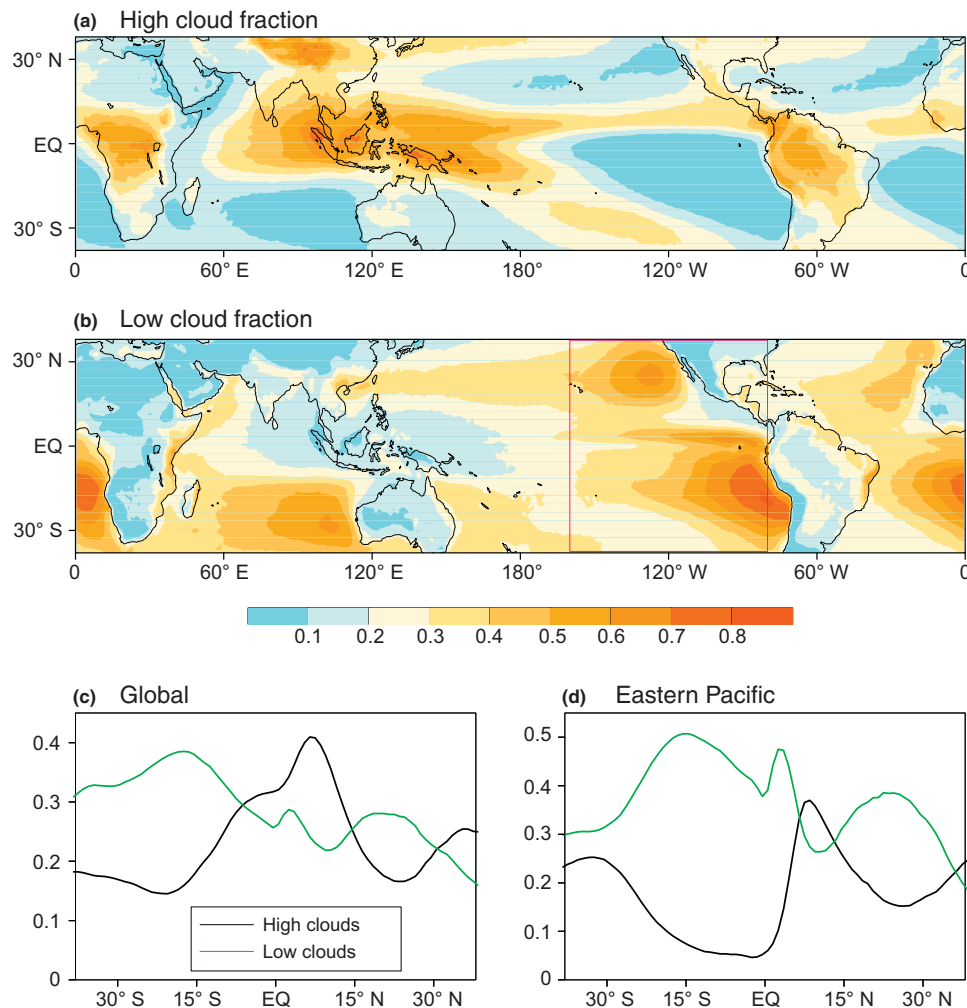


Figure 10 Long-term mean (2002–10) satellite-derived cloud fraction maps: (a) high clouds; (b) low clouds. Latitudinal profiles of high (black) and low (green) cloud fractions averaged over (c) global longitudes and (d) eastern Pacific longitudes (160°W – 80°W ; see red rectangle in (b)). Source: National Aeronautics and Space Administration's (NASA) Moderate Resolution Imaging Spectroradiometer (MODIS) instrument.

point out that accurate measurements of rainfall over the oceanic regions are possible only with satellite retrievals and this is still an area of active research.

Perspectives from New Satellite Observations

Recently launched satellite instruments provide an unprecedented opportunity to characterize the three-dimensional structure of the ITCZ. **Figure 10** depicts the long-term mean distribution of high and low cloud fractions derived from the Moderate Resolution Imaging Spectroradiometer (MODIS) instrument. The pattern of high cloud fractions (**Figure 10(a)**) largely follows the mean rainfall pattern in **Figure 2**, suggesting the development of deep convective clouds within the ITCZ. Meanwhile, low clouds prevail over subtropical regions over the eastern Pacific and Atlantic Oceans as well as in the southern Indian Ocean (**Figure 10(b)**). Maximum low cloud fractions are exhibited over the eastern portions of these ocean

basins, largely due to coastal oceanic upwelling induced by the trade winds and downward motion in the subtropical high region associated downward branch of the Hadley Circulation (**Figure 7**). Meridional asymmetry in cloud fraction about the equator is clearly evident in **Figure 10(c)** and **10(d)**. Maximum deep cloud fractions are found near 6° N in the globally zonal mean profile (**Figure 10(c)**), largely representing the northward displaced ITCZ over the eastern Pacific and Atlantic Oceans. Meanwhile, minimum high cloud fractions are observed over the subtropical latitudes in both hemispheres, where the maxima in low cloud fractions are evident. In addition to the two peaks over the subtropical regions in both hemispheres, there is a third maximum of low cloud fraction found just north of the equator, which captures the low cloud maximum over the eastern Pacific at this latitude belt (see **Figure 10(b)**). Similar features in high and low cloud profiles are found over the eastern Pacific sector (**Figure 10(d)**). High clouds are largely confined within a narrow latitude belt centered at 9° N. Over the broad region on both sides of the

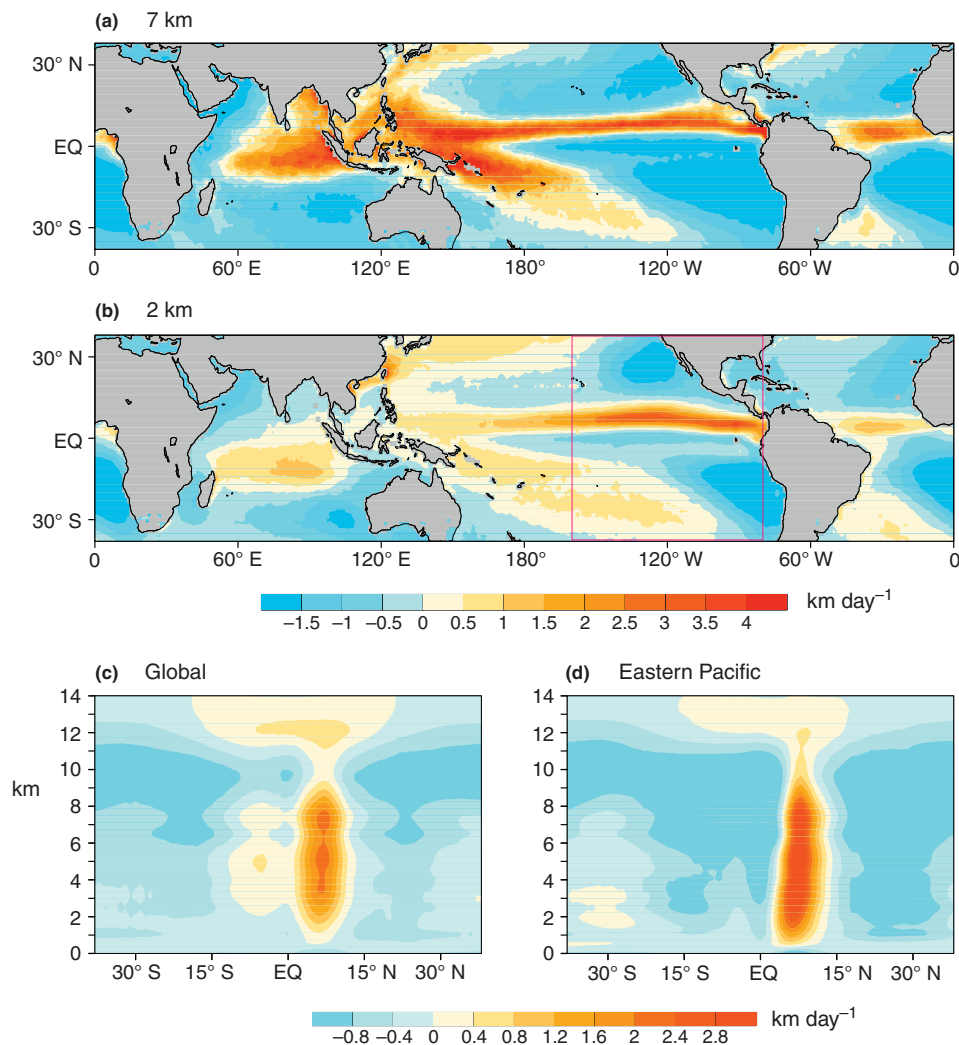


Figure 11 Long-term mean (1998–2008) satellite-derived total atmospheric diabatic heating at (a) 7 km and (b) 2 km. Latitude-vertical profiles of diabatic heating averaged over (c) global longitudes and (d) eastern Pacific longitudes (160° W– 80° W; see red rectangle in (b)). Source: NASA's Tropical Rainfall Measuring Mission (TRMM), courtesy of W. Olson, University of Maryland.

ITCZ, low clouds dominate whereas high clouds are largely suppressed.

Figure 11 displays spatial distribution of long-term mean total atmospheric diabatic heating at 7 and 2 km as well as its height-latitude profiles averaged over the global longitudes and eastern Pacific sectors. These heating fields are derived based on the measurements from microwave imager and precipitation radar aboard the Tropical Rainfall Measurement Mission (TRMM). The heating pattern at 7 km largely mirrors the mean rainfall pattern in **Figure 2**, suggesting the dominance of the condensational latent heating release in the total diabatic heating over these tropical convective regions. In convectively inactive regions, over regions where low clouds dominate (**Figure 10(b)**) in particular, the total diabatic heating at 7 km is characterized by strong net cooling due to the radiative heat loss in the absence of deep clouds. At 2 km, while strong radiative cooling is still evident over these low cloud regions, heating over the convective regions generally exhibits weaker amplitude compared to that at 7 km. A weak heating center is found over the southern Indian Ocean near 10° S at 2 km, in contrast to the maximum heating near the equator and northern Indian Ocean at 7 km.

The deep structure of the convection over the ITCZ regions is further evident in **Figure 11(c)**, which illustrates the globally averaged mean vertical-latitude diabatic heating profiles. A strong heating maximum is observed to the north of the equator centered at 8° N, stretching from lower troposphere to upper level near 10 km. Another relatively weaker heating center is found to the south of equator near 5° S, reflecting the condensational heating release over the SPCZ and ITCZ over the Indian Ocean. The heating amplitude over the southern branch is much weaker than its counterpart in northern

hemisphere, largely due to the inclusion of the cold SST tongue regions over the eastern Pacific and Atlantic Oceans in the global averages. The strong deep heating structure coupled with the ITCZ convection over the eastern Pacific is further displayed in **Figure 11(d)**, with broad cooling regions present on both sides of the heating area. The strong vertically developed heating over the ITCZ is crucial in driving the atmospheric circulation, which in turn plays a central role in the global water and energy cycles as depicted in **Figure 7**.

See also: Numerical Models: Parameterization of Physical Processes: Clouds. **Tropical Cyclones and Hurricanes:** Hurricanes: Observation. **Tropical Meteorology and Climate:** Hadley Circulation; Intraseasonal Oscillation (Madden–Julian Oscillation).

Further Reading

- Emanuel, K.A., 1994. *Atmospheric Convection*. Oxford University Press, New York.
- Gill, A.E., 1982. *Atmosphere–Ocean Dynamics*. Academic Press, Orlando, FL.
- Henderson-Sellers, A., Robinson, P.J., 1986. *Contemporary Climatology*. Longman Scientific and Technical, Essex, UK.
- Houghton, J.T., 1986. *The Physics of the Atmosphere*. Cambridge University Press, Cambridge, UK.
- Kidder, S.Q., Vonder Haar, T.H., 1995. *Satellite Meteorology: An Introduction*. Academic Press, San Diego, CA.
- Kraus, E.B., Businger, J.A., 1994. *Atmosphere–Ocean Interaction*. Oxford University Press, New York.
- Lindzen, R.S., 1990. *Dynamics in Atmospheric Physics*. Cambridge University Press, Cambridge, UK.
- Salby, M.L., 1996. *Fundamentals of Atmospheric Physics*. Academic Press, San Diego, CA.

Intraseasonal Oscillation (Madden–Julian Oscillation)

RA Madden, National Center for Atmospheric Research, Boulder, CO, USA

© Published by Elsevier Ltd.

Synopsis

The Intraseasonal Oscillation (Madden–Julian Oscillation) reflects the west-to-east movement of very large-scale atmospheric circulation cells in the equatorial plane. Locally, associated variables like surface pressure, winds, and convection oscillate with a typical timescale of 45 days. It is associated with much of equatorial weather variations on the intraseasonal timescale.

Introduction

The global trade winds of the Northern and Southern Hemispheres meet near the Equator in the Intertropical Convergence Zone (ITCZ). Here, the surface air rises in large convective storms mostly over preferred regions of South America, Africa, and the Malay Archipelago, and then moves north and south and east and west at upper levels. The circulations in the north and south directions form the Hadley Cells and those in the east and west directions which are centered on the Malay Archipelago form the Walker Cells. The first are named after George Hadley who described the circulation from tropics to midlatitudes in the 1700s. The Walker Cells are named after Sir Gilbert Walker, a British climatologist. Walker wrote in the 1920s and 1930s about an east-to-west oscillation in pressure between the Indian/West Pacific Oceans and the Southeast Pacific Ocean. It was later learned that this pressure oscillation was accompanied by a longitudinal shift in the main region of upward motion in the east–west circulation cells. They were named the Walker Cells. The pressure oscillation that Walker studied has a dominant timescale on the order of 2–7 years and is now called the ‘Southern Oscillation,’ a part of the El Niño–Southern Oscillation (ENSO). However, circulation in the Walker Cells changes on all timescales, and here is described a variation on an intraseasonal timescale.

Figure 1 captures the rudimentary features of the most important intraseasonal modifications of the average Walker Circulation. Each panel represents a snapshot taken in the equatorial plane about 7–15 days apart with time increasing from top to bottom. Note that convective areas indicated by the cartoon clouds and east–west circulation cells indicated by the streamlines move from the Indian to the Pacific Ocean. The clouds represent large areas of convection on the order of a few thousand kilometers across. The converging and diverging eastward- and westward-directed streamlines are meant to depict regions of horizontal convergence and divergence. Maximum convergence or divergence need not be at the null point of the streamlines as shown. In fact, convection is often shifted into the region of the low-level westerlies. All the features shown in **Figure 1** represent anomalies from the time-averaged tropical circulation. As they move eastward, they modify the clouds in the ITCZ on timescales shorter than a season but longer than a couple of weeks. The average period of the phenomenon is near 45 days. As a result, it has been referred to as the 40–50 day, the 30–60 day Oscillation, and the Intraseasonal Oscillation after its typical timescale, and the Madden–Julian Oscillation (MJO) after authors who described it in the 1970s. The space–

time spectrum of Wheeler (see **Tropical Meteorology and Climate: Equatorial Waves**) also points to the very large spatial scale of **Figure 1** and the typical 30–60 day period.

Even though its existence has been known since the 1970s, much about it is not fully understood. Here, known features and hypotheses about the oscillation are presented.

Description of the Oscillation

Figure 1(a) is meant to convey relatively low surface pressure accompanied by a convergence of moist, low-level air in the Indian Ocean which results in building convection. The tropopause height rises above the convection. The tropopause is the top of the lowest layer of the atmosphere called the troposphere. The weather is confined to the troposphere, which, near the Equator is about 18 km thick. To the east, over the Central Pacific, there is relatively high surface pressure, sinking motion, and suppressed convection. At this time, the circulation in the Walker Cell is increasing with its center of rising motion shifted to west of its average position over the Malay Archipelago. The low-pressure anomalies then move and spread eastward. Local anomalies are on the order of -1 hPa (wet region) to $+1$ hPa (dry region). The area of enhanced large-scale convection and the deeper troposphere also moves eastward at about 5 m s^{-1} to a place in the Western Pacific, close to its time average position (**Figure 1(b)**). Precipitation anomalies are, on average, $\pm 3 \text{ mm day}^{-1}$. Eventually (**Figure 1(c)**), rising motions and convection weaken over the colder equatorial waters of the Central and Eastern Pacific, and high pressure, sinking motion, and suppressed convection reside over the Indian Ocean. The tropopause height falls there too.

After the convection weakens and disappears over the Central Pacific, a wavelike disturbance often continues eastward in the upper troposphere at a faster speed ($10\text{--}20 \text{ m s}^{-1}$). It can propagate the full circumference of the Earth. The oscillation is not so easily detected in the lower troposphere from about 90° W eastward to 20° E , but related convection sometimes can be seen over South America and the Atlantic accompanying the upper-level divergence that is indicated in **Figure 1(d)**. At this time, there is relatively suppressed convection over the Indian and Pacific Oceans. Although the oscillation is irregular, it is present about half the time. It tends to be more active during northern winters.

The accompanying large-scale convective region contains many tall clouds which can be readily identified in infrared satellite measurements because of their very cold tops. **Figure 2** shows cold clouds (blues and greens) over the Indian Ocean in early November

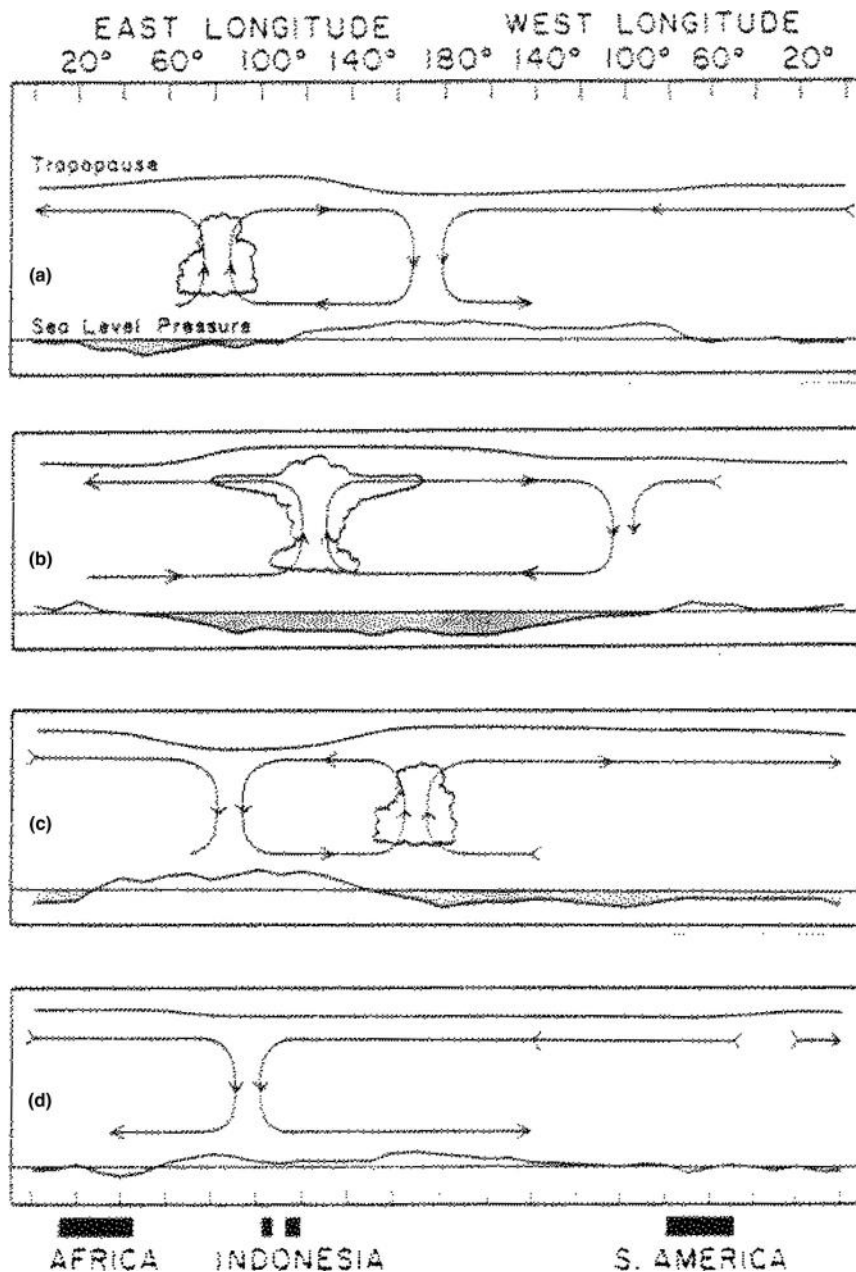


Figure 1 A schematic of the approximate structure of the oscillation in the equatorial plane. The situations summarized in each panel are 7–15 days apart with time increasing downward. Cartoon clouds indicate large regions of increased convection. Streamlines show the east–west circulation with low-level convergence into, and upper-level divergence out of, the convective areas. Wavy line at the top represents the tropopause and that at the bottom anonymous sea level pressure. Reproduced from Madden, R.A., Julian, P.R., 1972. Description of global-scale circulation cells in the tropics with a 40–50 day period. *Journal of the Atmospheric Sciences* 29, 1109–1123.

and again in late December of 2009 that move to the Central Pacific in late November and late January 2010, respectively.

Interestingly, satellite data with higher spatial resolution than that of [Figure 2](#) show that within the eastward-moving large-scale cloudy areas are organized convective cloud clusters of only few hundred kilometers across. These smaller scale features actually move westward.

Besides the eastward movement of the large-scale cloudy regions along the Equator shown in [Figures 1 and 2](#), there

sometimes is some poleward movement of alternately cloudy and fair weather. This poleward movement is most marked over India during the summer monsoon there. Another common variation is eastward propagation to about 100° E longitude and then northeastward or southeastward movement. The former is most common in northern summer and the latter in southern summer, reflecting a connection with clouds of the ITCZ which are typically displaced into the summer hemisphere.

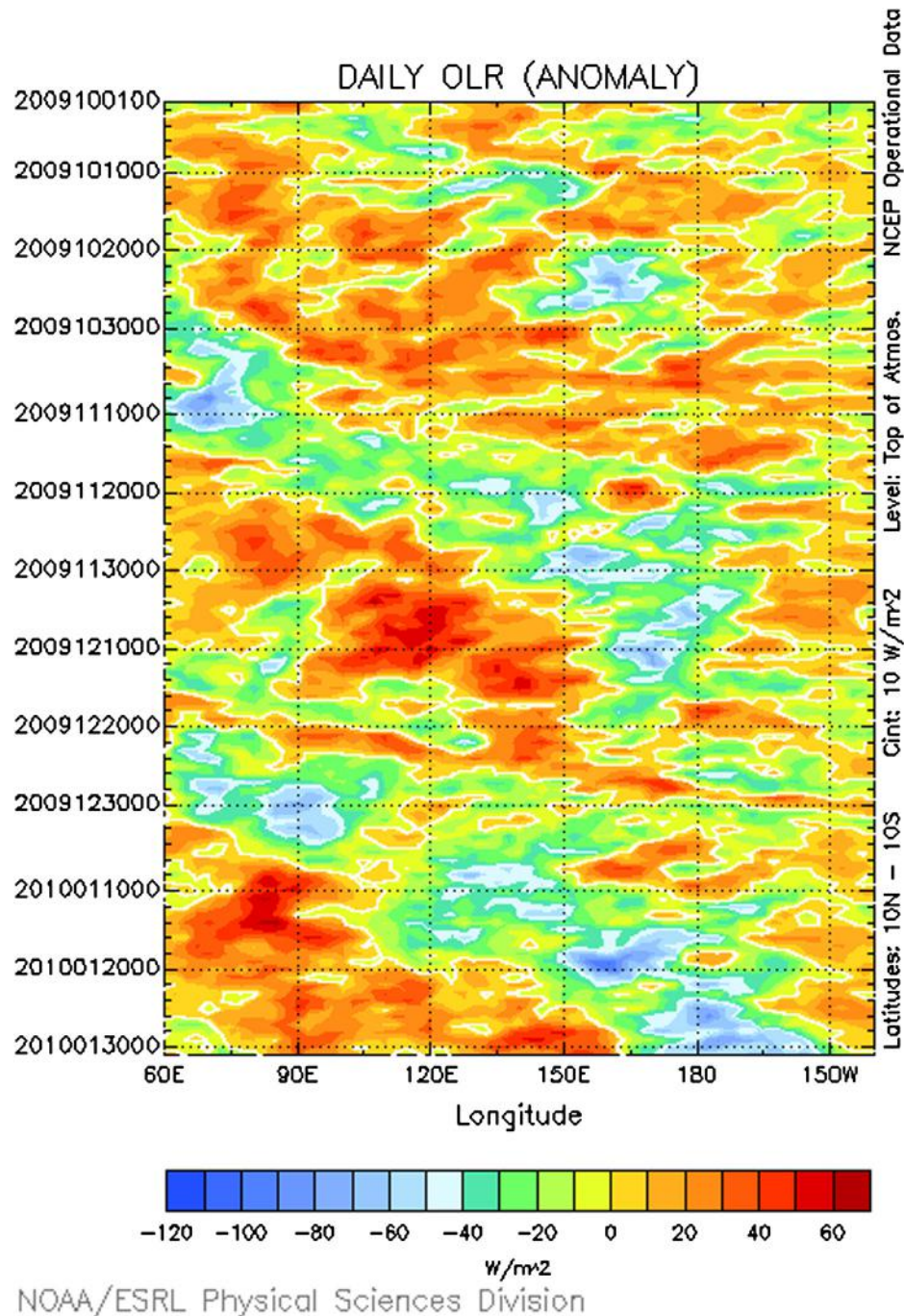


Figure 2 Anomalies from a time average of outgoing longwave radiation for the latitude band 10°S–10°N. Blues and greens are regions of cold temperature and high convective clouds. The vertical axis is time increasing downward from 1 October 2009, 00.00 GMT (2009100100) to 30 January 2010, 00.00 GMT. Patterns sloping diagonally from higher left to lower right imply eastward propagation. Image provided by the NOAA/ESRL Physical Sciences Division, Boulder Colorado from their Web site at <http://www.esrl.noaa.gov/psd/>.

A Partial Theoretical Model

The large-scale convective area associated with an oscillation produces considerable heat as rising moist air condenses to form cloud droplets. **Figure 3** shows the theoretically expected response to such heating at the Equator. To the east of the heating, winds blow from the east in the lower troposphere

into the heating region with lowest pressure along the Equator, and back to the east in the upper troposphere with highest pressure along the Equator. The response to the west is characterized by rotating eddies on either side of the Equator. They rotate cyclonically (clockwise in the Southern Hemisphere and counterclockwise in the Northern Hemisphere) about low pressure in the lower troposphere, and anticyclonically about

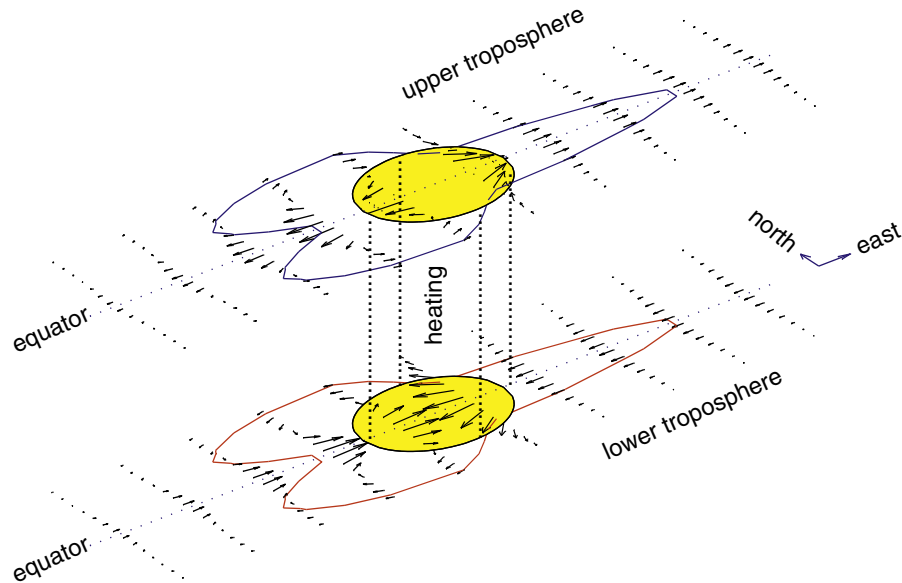


Figure 3 The theoretical response of winds in the lower and upper troposphere (indicated by the black arrows) to heating. The heating zone is indicated by the yellow ellipses and vertical dotted lines. Lowest relative pressure is within the red line in the lower troposphere, and highest relative pressure is within the blue line in the upper troposphere. Determined from equations in Gill, A.E., 1982. *Atmosphere-Ocean Dynamics*. Academic Press, New York.

high pressure in the upper troposphere. In addition, along the Equator, lower tropospheric winds blow from west to east, and upper tropospheric winds from east to west, in and out of the heating region.

The east side response is a Kelvin wave named after Lord Kelvin who studied water waves along a vertical side boundary. In this case, the Equator, where the vertical component of the Earth's rotation vector changes sign, serves as the vertical side boundary. The west side response is Rossby waves whose existence and behavior are due to the latitudinal variation of that same vertical component of the Earth's rotation vector. C-G Rossby was the first to clearly isolate the dynamics of these waves. Converging low tropospheric winds of the Kelvin and Rossby waves rise in the cumulonimbus clouds of the heating region, and diverge in the upper troposphere.

The observed structure of the oscillation, during its passage from the Indian to Pacific Oceans, is similar in many respects to the forced Kelvin-Rossby waves of Figure 3. It may be that convection building in the equatorial Indian Ocean forces the Kelvin-Rossby waves of Figure 3. The entire complex moves eastward, and upon reaching the cold water of the Central Pacific can no longer support the convection. A free Kelvin wave then continues eastward in the upper troposphere. This scenario would not, however, explain the eastward movement of the large-scale convective region, the typical timescale of the oscillation, nor why it organizes into such a large spatial scale. These important characteristics are not yet fully understood.

Effects of the Oscillation

Figure 2 points to the fact that the oscillation affects weather in the equatorial region. It has a wide range of other known effects as well. They are summarized by the Climate Variability and

Predictability (CLIVAR) Research Project, a part of the World Climate Research Program (WCRP) as follows:

1. Many observational studies have shown anomalously strong MJO activity prior to and during the eastward shifts of the Walker Cells that occur every 2–7 years (warm events of ENSO). Their exact role in timings of ENSO variations is under study.
2. The MJO modulates cyclone activity in many regions of the tropics. *Hurricane* (or *Typhoon*) genesis is two to four times more likely over the Indian, West Pacific, Caribbean, and Atlantic Oceans during its wet phase than during its dry phase.
3. The *Australian Monsoon* is the name of the major rainy season of January and February over Northern Australia. The monsoon onset typically occurs from mid- to late December. Onset is less likely to occur during a dry phase of the MJO. After onset, there are regularly occurring bursts (wet or active periods) and breaks (dry periods) in monsoon activity. Along Australia's North Coast, precipitation can vary as much as 6 mm day^{-1} between burst and break. A large portion of this variability is connected with MJOs.
4. The northern summer season *Asian Monsoon* is affected by MJOs. They also influence breaks and bursts here. The vast majority of monsoon lows occur during enhanced MJO precipitation and upper-level divergence.
5. *High latitude weather* is also affected. This is not surprising since the MJO influences large-scale precipitation in the tropics which, in turn, forces much of the higher latitude weather. Extreme precipitation amounts can fall along the West Coast of North America roughly coincident with MJO phase 1C (Figure 1). MJOs may also affect the Arctic Oscillation (AO) and the Southern Annular Mode (SAM). The AO and SAM are manifested by pressure variations with opposite sign between polar and midlatitudes of the Northern and Southern Hemispheres, respectively, that affect weather patterns.

6. Analysis of satellite measured ocean color data shows that the MJO produces systematic variation in *ocean surface chlorophyll*. Changing surface wind associated with the MJO is a contributing mechanism changing vertical entrainment of nutrients and thus the surface chlorophyll.
7. *African rainfall* variability at intraseasonal time scales is still poorly understood. However, an MJO-like signal with a period of 40–45 days can be seen over much of the continent. Depending on location and season, some 30–40% of intraseasonal rainfall variance can be explained by it.
8. *Variations in total column O₃* appear to be affected by the MJO. Resulting O₃ variations of ± 10 Dobson units are mainly evident in the subtropics, and they are as large as ones associated with annual and interannual timescales. Negative O₃ anomalies flank or lie to the west of MJO convection. This is consistent with a higher tropopause there (**Figure 1**), since the tropopause separates O₃-rich stratospheric air from that of the less rich troposphere.
9. There are easily identified variations in the relative angular momentum (RAM) of the atmosphere, which is the globally integrated west-to-east wind, associated with the oscillation. They are on the order of 10% of the total RAM. By virtue of the fact that total angular momentum of the atmosphere–ocean–Earth system remains very nearly constant, corresponding variations are evident in the Earth’s rotation and the length of day. Required momentum exchanges among atmosphere, ocean, and solid Earth take place through surface wind friction and pressure differences across mountains.

Implications

Improved understanding and more realistic simulations of the MJO (including onset and demise) will contribute to better

forecasting in the global tropics and (at times) in the extra-tropics on a multiple week timescale.

See also: **Climate and Climate Change:** Climate Variability: North Atlantic and Arctic Oscillation. **Dynamical Meteorology:** Kelvin Waves; Rossby Waves; Waves. **General Circulation of the Atmosphere:** Angular Momentum of the Atmosphere. **Stratosphere/Troposphere Exchange and Structure:** Tropopause. **Tropical Cyclones and Hurricanes:** Hurricanes: Observation. **Tropical Meteorology and Climate:** El Niño and the Southern Oscillation: Observation; El Niño and the Southern Oscillation: Theory; Hadley Circulation; Intertropical Convergence Zone; Monsoon: Overview; Walker Circulation.

Further Reading

- CLIVAR Madden–Julian Oscillation Working Group, 2009. MJO simulation diagnostics. *Journal of Climate* 22, <http://dx.doi.org/10.1175/2008JCLI2731.1>.
- Gill, A.E., 1982. *Atmosphere–Ocean Dynamics*. Academic Press, New York.
- James, I.N., 1994. *Introduction to Circulating Atmospheres*. Cambridge University Press, Cambridge, UK.
- Lau, K.M., Waliser, D.E. (Eds.), 2005. *Intraseasonal Variability in the Atmosphere–Ocean–Climate System*. Springer, Heidelberg, Germany.
- Madden, R.A., Julian, P.R., 1972. Description of global-scale circulation cells in the tropics with a 40–50 day period. *Journal of the Atmospheric Sciences* 29, 1109–1123.
- Madden, R.A., Julian, P.R., 1994. Observations of the 40–50-day tropical oscillation – a review. *Monthly Weather Review* 122, 814–837.
- US CLIVAR Project page www.usclivar.org/mjosci.php.
- Zhang, C., 2005. Madden–Julian oscillation. *Reviews of Geophysics* 43, 1–36.

Madden–Julian Oscillation: Skeleton and Conceptual Models

AJ Majda, New York University, New York, NY, USA

SN Stechmann, University of Wisconsin–Madison, Madison, WI, USA

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

The Madden–Julian Oscillation's (MJO's) 'skeleton' is its fundamental features on intraseasonal/planetary scales: (1) slow eastward phase speed of roughly 5 m s^{-1} , (2) peculiar dispersion relation with $d\omega/dk \approx 0$, and (3) horizontal quadrupole vortex structure. The MJO skeleton model is a minimal, nonlinear oscillator model that captures these three features together. The fundamental mechanism involves interactions between moisture, convective activity, and equatorial fluid dynamics. Further details beyond the MJO skeleton are referred to as the MJO's 'muscle' and include refined vertical structure and upscale convective momentum transport from subplanetary-scale convection and waves.

Introduction

The Madden–Julian Oscillation (MJO) is the dominant component of intraseasonal (≈ 30 – 60 days) variability in the tropics. This variability appears not only in the wind, pressure, and temperature fields, but also in water vapor and convection. The structure of the MJO is a planetary-scale ($\approx 10\,000$ – $40\,000 \text{ km}$) circulation cell with regions of enhanced and suppressed convection, and it propagates slowly eastward at a speed of roughly 5 m s^{-1} . The main regions of MJO convective activity are the Indian and western Pacific Oceans, and the MJO interacts with monsoons, tropical cyclones, El Niño–Southern Oscillation, and other tropical phenomena. Nevertheless, while the MJO is mainly a tropical phenomenon, it also interacts with the extratropics and can affect midlatitude predictability.

The fundamental features of the MJO on intraseasonal–planetary scales are referred to here as the MJO's 'skeleton':

1. Slow eastward phase speed of roughly 5 m s^{-1} ,
2. Peculiar dispersion relation with $d\omega/dk \approx 0$, and
3. Horizontal quadrupole vortex structure.

While these are the salient planetary–intraseasonal features of MJO composites, individual MJO events often have additional features, such as westerly wind bursts, and different varieties of convective features within the MJO envelope. Since these additional features add detailed character to each MJO's structure, and since these features often account for additional strength beyond the MJO's skeleton, they are referred to here as the MJO's 'muscle.'

While the MJO is often described as an intraseasonal–planetary scale phenomenon, it actually has a complex multiscale structure. The MJO's region of enhanced convection is composed of a menagerie of smaller-scale convective processes, including synoptic-scale convectively coupled equatorial waves (CCEWs), mesoscale convective systems (MCSs), and tropical cyclones. Hence the MJO is essentially an envelope of subplanetary-scale convective activity and is a multiscale phenomenon. In light of this, multiscale interactions are a key part of the theory for the MJO's skeleton and muscle.

The MJO skeleton model is a minimal dynamical model that captures the MJO's intraseasonal–planetary scale features 1–3 together. The model is a nonlinear oscillator model whose fundamental mechanism involves interactions between moisture, convective activity, and equatorial fluid dynamics. Before presenting the full MJO skeleton model, these main components are described individually.

Dry Equatorial Long-Wave Equations

The equatorial long-wave equations describe equatorial dynamics on planetary spatial scales of $O(10\,000 \text{ km})$ in the zonal x direction and intraseasonal timescales of $O(10 \text{ days})$ or longer. The equations take the form

$$\frac{\partial u'}{\partial t} - \gamma v' + \frac{\partial p'}{\partial x} = 0 \quad [1]$$

$$\gamma u' + \frac{\partial p'}{\partial y} = 0 \quad [2]$$

$$\frac{\partial p'}{\partial z} = \theta' \quad [3]$$

$$\frac{\partial u'}{\partial x} + \frac{\partial v'}{\partial y} + \frac{\partial w'}{\partial z} = 0 \quad [4]$$

$$\frac{\partial \theta'}{\partial t} + w' = 0 \quad [5]$$

Here u' , v' , and w' are the zonal, meridional, and vertical velocity anomalies, respectively; and p' and θ' are the pressure and potential temperature anomalies, respectively. Notice that the $\partial v'/\partial t$ term is small in this long-wave limit, and hence geostrophic balance is satisfied in the meridional y direction but not the zonal x direction. The Coriolis terms $-\gamma v'$ and $\gamma u'$ vanish at the equator ($y = 0$) and are written using the equatorial beta-plane approximation. The eqns [1]–[5] have been nondimensionalized using the standard equatorial reference scales described in Table 1.

The wave solutions of eqns [1]–[5] can be found by expanding in vertical and meridional basis functions. First, each variable is expanded in terms of vertical structure functions such as $u'(x, y, z, t) = \sum_j u'_j(x, y, t) \sqrt{2} \cos(jz)$, etc. This allows the eqns [1]–[5] to be separated into an infinite number of (long-wave) shallow water systems, where each mode j corresponds to a separate independent (long-wave) shallow water system. For the first baroclinic mode, $j = 1$, the system has the form

$$\frac{\partial u'}{\partial t} - \gamma v' - \frac{\partial \theta'}{\partial x} = 0 \quad [6]$$

$$\gamma u' - \frac{\partial \theta'}{\partial y} = 0 \quad [7]$$

$$\frac{\partial \theta'}{\partial t} - \frac{\partial u'}{\partial x} - \frac{\partial v'}{\partial y} = 0 \quad [8]$$

where the subscript $j = 1$ has been dropped from all variables.

Table 1 Constants and reference scales for nondimensionalization

Parameter	Derivation	Value	Description
β		$2.3 \times 10^{-11} \text{ m}^{-1} \text{ s}^{-1}$	Variation of Coriolis parameter with latitude
θ_0		300 K	Potential temperature at surface
g		9.8 m s^{-2}	Gravitational acceleration
H		16 km	Tropopause height
N^2	$(g/\theta_0)d\bar{\theta}/dz$	10^{-4} s^{-2}	Buoyancy frequency squared
c	NH/π	50 m s^{-1}	Velocity scale
L	$\sqrt{c/\beta}$	1500 km	Equatorial length scale
T	L/c	8 h	Equatorial timescale
	$HN^2\theta_0/(\pi g)$	15 K	Potential temperature scale
	H/π	5 km	Vertical length scale
	$H/(\pi T)$	0.2 m s^{-1}	Vertical velocity scale
	c^2	$2500 \text{ m}^2 \text{ s}^{-2}$	Pressure scale

From Stechmann, S.N., Majda, A.J., Khouider, B., 2008. Nonlinear dynamics of hydrostatic internal gravity waves. *Theoretical and Computational Fluid Dynamics* 22, 407–432. DOI [10.1007/s00162-008-0080-7](https://doi.org/10.1007/s00162-008-0080-7).

Next, through meridional expansion, this long-wave shallow water system can be separated into equations for an infinite set of zonally propagating, equatorially trapped waves. This is achieved by expanding the variables u' , v' , θ' in terms of meridional basis functions such as $u'(x, y, t) = \sum_m u'_m(x, t) \phi_m(y)$ etc. Here the appropriate meridional basis functions are the parabolic cylinder functions, the first few of which are $\phi_0(y) = \pi^{-1/4} \exp(-y^2/2)$, $\phi_1(y) = \pi^{-1/4} \sqrt{2}y \exp(-y^2/2)$, and $\phi_2(y) = \pi^{-1/4} 2^{-1/2} (2y^2 - 1) \exp(-y^2/2)$. Notice the $\exp(-y^2/2)$ factor in each of the parabolic cylinder functions, whose rapid decay away from the equator, $y = 0$, corresponds to the terminology ‘equatorially trapped waves.’ Through such a meridional expansion, one can derive the infinite set of equations

$$\frac{\partial K'}{\partial t} + \frac{\partial K'}{\partial x} = 0 \quad [9]$$

$$\frac{\partial R'_1}{\partial t} - \frac{1}{3} \frac{\partial R'_1}{\partial x} = 0 \quad [10]$$

$$\frac{\partial R'_2}{\partial t} - \frac{1}{5} \frac{\partial R'_2}{\partial x} = 0 \quad [11]$$

$$\frac{\partial R'_3}{\partial t} - \frac{1}{7} \frac{\partial R'_3}{\partial x} = 0 \quad [12]$$

⋮

Here K' , R'_1 , R'_2 , R'_3 , ... are the amplitudes of equatorial long-wave Kelvin and Rossby waves, respectively. The variables $u'(x, y, t)$, $v'(x, y, t)$, $\theta'(x, y, t)$ are recovered using the relationships between the primitive variables $u'_m(x, t)$, $v'_m(x, t)$, $\theta'_m(x, t)$ and the wave amplitudes $K'(x, t)$, $R'_m(x, t)$. For instance, $u'_0(x, t) = [K'(x, t) - R'_1(x, t)/2]/\sqrt{2}$, etc. To illustrate this, Figure 1 shows u' , v' , θ' for the two separate cases with either $K'(x) = -\cos(x)$ or $R'_1(x) = \cos(x)$. These illustrations can be thought of as the meridional structures associated with the wave amplitudes $K'(x, t)$ and $R'_1(x, t)$.

The dispersion relations for the waves from eqns [9] to [12] are seen to be $\omega_K(k) = k$, $\omega_{R1}(k) = -k/3$, $\omega_{R2}(k) = -k/5$, etc., which arise from assuming solutions of the form

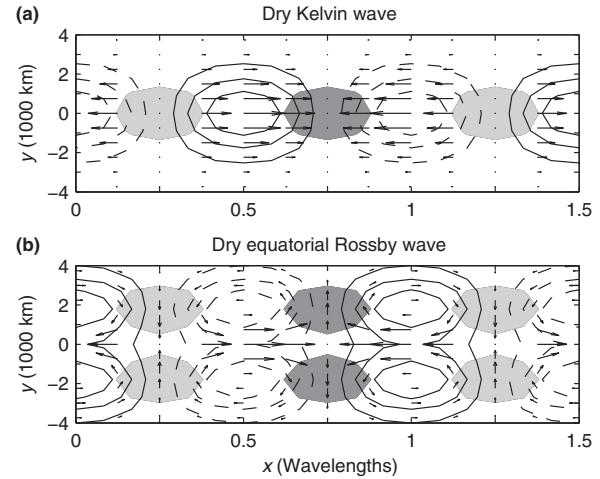


Figure 1 Horizontal structures of two unforced ‘dry’ equatorial long-waves: (a) Kelvin wave, and (b) the first equatorial Rossby wave, R'_1 . Contours show lower tropospheric pressure with positive (negative) anomalies denoted by solid (dashed) lines. The contour interval is one-fourth the maximum amplitude of the anomaly, and the zero contour is not shown. Anomalies of convergence (divergence) that are greater than two-thirds the maximum amplitude are shaded dark (light) gray. From Majda, A.J., Stechmann, S.N., 2009b. The skeleton of tropical intraseasonal oscillations. *Proceedings of the National Academy of Sciences U.S.A.* 106 (21), 8417.

$K'(x, t) = \exp \{i[kx - \omega_K(k)t]\}$, etc. Hence these waves are nondispersive in this long-wave limit.

Moisture–Convection Interactions on Intraseasonal–Planetary Scales

Besides the fluid dynamical core, the second main component of the MJO skeleton model is moisture–convection interactions. These quantities are represented by the variables q' and a , respectively:

- q' : lower tropospheric moisture anomaly,
- a : amplitude of the convection/wave activity envelope.

It is noteworthy that, for the MJO skeleton model, it is only the amplitude of the convection/wave activity envelope that is needed, not any of the details of the particular convection/waves that make up the envelope, although the specific details can be important for convective momentum transports (CMTs) or other features of the MJO's 'muscle.' Also notice that a , without a prime, is not an anomaly.

A key part of the q -and- a interaction is how the moisture anomalies influence the convection. The premise is that, for convective activity on planetary–intraseasonal scales, it is the *time tendency* of convective activity – not the convective activity itself – that is most directly related to the (lower tropospheric) moisture anomaly. In other words, rather than a functional relationship $a = a(q)$, it is posited that q mainly influences the tendency – i.e., the growth and decay rates – of the convective activity. The simplest equation that embodies this idea is

$$\frac{\partial a}{\partial t} = \Gamma q' a, \quad [13]$$

where Γ is a constant of proportionality: positive low-level moisture anomalies create a tendency to enhance the envelope of convection/wave activity, and negative low-level moisture anomalies create a tendency to decrease the envelope of convection/wave activity.

The basis for eqn [13], and the physics behind it, comes from a combination of observations, modeling, and theory. Generally speaking, it is well known that tropospheric moisture content plays a key role in regulating convection. On planetary–intraseasonal scales specifically, several studies have shown that the lower troposphere tends to moisten during the suppressed convection phase of the MJO, and lower tropospheric moisture leads the MJO's heating anomaly, which suggests the relationship eqn [13]. Furthermore, this relationship is also suggested by simplified models for synoptic-scale convectively coupled waves, which show that the growth rates of the convectively coupled waves depend on the wave's environment, such as the environmental moisture content. In eqn [13], the factor $\Gamma q'$ acts as a growth/decay rate for the convective activity, and the value of Γ has been estimated from these growth rate variations in idealized models of convectively coupled waves. Lastly, amplitude equations such as eqn [13] have been used in other areas of science and engineering, and they can sometimes be derived from the governing equations using systematic asymptotics.

MJO Skeleton Model

By combining the parameterization eqn [13] with the (long-wave-scaled) linearized primitive eqns [1]–[5], the MJO skeleton model is obtained:

$$\frac{\partial u'}{\partial t} - \gamma v' + \frac{\partial p'}{\partial x} = 0 \quad [14]$$

$$\gamma u' + \frac{\partial p'}{\partial y} = 0 \quad [15]$$

$$\frac{\partial p'}{\partial z} = \theta' \quad [16]$$

$$\frac{\partial u'}{\partial x} + \frac{\partial v'}{\partial y} + \frac{\partial w'}{\partial z} = 0 \quad [17]$$

$$\frac{\partial \theta'}{\partial t} + w' = \bar{H}a - s^\theta \quad [18]$$

$$\frac{\partial q'}{\partial t} - \bar{Q}w' = -\bar{H}a + s^q \quad [19]$$

$$\frac{\partial a}{\partial t} = \Gamma q' a. \quad [20]$$

An equation for the lower-tropospheric moisture q' is also included in the system in eqn [19]. The terms $\pm \bar{H}a$ represent a heat source and moisture sink due to convective activity, and s^θ and s^q represent radiative cooling and a moisture source, respectively.

Notice that this model contains a minimal number of parameters: $\bar{Q} = 0.9$, the (nondimensional) mean background vertical moisture gradient; and $\Gamma = 1$, or $\Gamma \approx 0.2 \text{ K}^{-1} \text{ d}^{-1}$ in dimensional units. The source terms s^θ and s^q must also be specified (see below). Also notice that the parameter $\bar{H}$ is actually irrelevant to the dynamics (as can be seen by rescaling a); it is written here for clarity of presentation: dimensionally, it gives $\bar{H}a$ the units of a heating rate while keeping a nondimensional. The dimensional value of $\bar{H}$ was chosen to be 10 K per day so that a typical value of a is ≈ 0.1 , similar to the nondimensional value of u .

The MJO skeleton model, eqns [14]–[20], is a *nonlinear oscillator* model for the MJO skeleton as a neutrally stable wave. In other words, the model includes neither damping nor instability mechanisms. The fundamental mechanism of the oscillation involves interactions between moisture, convection, and circulation. The nonlinear oscillator component can be seen in the terms $\partial q'/\partial t = -\bar{H}a$ and $\partial a/\partial t = \Gamma q' a$. However, there is also the important moisture convergence term, $\bar{Q}w'$, which couples this q -and- a oscillator to the circulation.

To obtain the simplest model for the MJO, truncated vertical and meridional structures are used. The procedure is analogous to the case of the dry equatorial long-wave equations described above, where equations for the dry zonally propagating, equatorially trapped waves $K'(x, t)$, $R'_m(x, t)$ were derived. For the vertical truncation, only the first baroclinic mode is used in the simplest model so that $u'(x, y, z, t) = u'(x, y, t)\sqrt{2}\cos(z)$, etc. The resulting equations are

$$\frac{\partial u'}{\partial t} - \gamma v' - \frac{\partial \theta'}{\partial x} = 0 \quad [21]$$

$$\gamma u' - \frac{\partial \theta'}{\partial y} = 0 \quad [22]$$

$$\frac{\partial \theta'}{\partial t} - \frac{\partial u'}{\partial x} - \frac{\partial v'}{\partial y} = \bar{H}a - s^\theta \quad [23]$$

$$\frac{\partial q'}{\partial t} + \bar{Q}\left(\frac{\partial u'}{\partial x} + \frac{\partial v'}{\partial y}\right) = -\bar{H}a + s^q \quad [24]$$

$$\frac{\partial a}{\partial t} = \Gamma q' a. \quad [25]$$

Next, for the meridional truncation, the variables are expanded in terms of the parabolic cylinder functions $\phi_m(y)$ as described above for the *dry* equatorial long-wave equations. It is also assumed that q and a have simple equatorial meridional structure proportional to $\exp(-y^2/2)$: $a(x, y, t) = [\bar{A}(x) + A'(x, t)]\phi_0(y)$, where $\bar{A}(x)$ is a background state. For the long-wave-scaled equations, such a meridional heating structure is known to couple with only Kelvin waves and the first symmetric equatorial Rossby waves R'_1 , and the resulting meridionally truncated equations can be written as

$$\frac{\partial K'}{\partial t} + \frac{\partial K'}{\partial x} = -\frac{1}{\sqrt{2}}\bar{H}A' \quad [26]$$

$$\frac{\partial R'}{\partial t} - \frac{1}{3}\frac{\partial R'}{\partial x} = -\frac{2\sqrt{2}}{3}\bar{H}A' \quad [27]$$

$$\frac{\partial Q'}{\partial t} + \frac{1}{\sqrt{2}}\bar{Q}\frac{\partial K'}{\partial x} - \frac{1}{6\sqrt{2}}\bar{Q}\frac{\partial R'}{\partial x} = \left(-1 + \frac{1}{6}\bar{Q}\right)\bar{H}A' \quad [28]$$

$$\frac{\partial A'}{\partial t} = \Gamma Q'(\bar{A} + A'), \quad [29]$$

where the subscript 1 has been dropped from R'_1 for brevity. The equations have been written in terms of anomalies from the state of radiative–convective equilibrium, $\bar{H}\bar{A}(x) = S^\theta(x) = S^q(x)$. Two simple cases are a uniform background state with constant $\bar{S}^\theta = 1$ K per day and a spatially varying state $S^\theta(x)$ that represents a warm-pool sea surface temperature distribution. While $S^\theta(x)$ and $S^q(x)$ are treated here as imposed functions, they could be treated as interactive functions of state variables in order to represent radiative feedback or surface-flux feedback.

An important point is that $K'(x, t)$ and $R'(x, t)$ are the amplitudes of the *structures* of Kelvin and Rossby waves, but these amplitudes in eqns [26] and [27] need not always *propagate* like ‘dry’ waves. In the absence of forcing in eqns [9] and [10], the ‘dry’ long-wave Kelvin and equatorial Rossby wave solutions are dispersionless waves that propagate at 50 and 17 m s^{−1}, respectively. However, in the presence of the coupled dynamical forcing A in eqns [26]–[29], the Kelvin and

equatorial Rossby wave *structures* can be coupled to each other and to Q and A ; and these coupled modes/structures can have propagation speeds very different from 50 or 17 m s^{−1}, and they can be dispersive. One such mode has a mixed Kelvin–Rossby wave structure and dispersion characteristics of the MJO and is summarized below.

Energetics

The nonlinear MJO skeleton model has two important energy principles, in the absence of source terms s^θ and s^q . First, the model eqns [21]–[25] conserve a vertically integrated moist static energy:

$$\partial_t(\theta + q) - (1 - \bar{Q})(u_x + v_y) = 0. \quad [30]$$

Second, this model conserves a positive total energy that includes a contribution from the convective activity a :

$$\begin{aligned} \partial_t \left[\frac{1}{2}u^2 + \frac{1}{2}\theta^2 + \frac{1}{2}\frac{\bar{Q}}{1-\bar{Q}}\left(\theta + \frac{q}{\bar{Q}}\right)^2 + \frac{\bar{H}}{\Gamma\bar{Q}}a \right] \\ - \partial_x(u\theta) - \partial_y(v\theta) = 0. \end{aligned} \quad [31]$$

This total energy is a sum of four terms: dry kinetic energy $u^2/2$, dry potential energy $\theta^2/2$, a moist potential energy proportional to $(\theta + \bar{Q}^{-1}q)^2$, and a convective energy $\bar{H}a/(\Gamma\bar{Q})$. Note that the natural requirement on the background moisture gradient, $0 < \bar{Q} < 1$, is needed to guarantee a positive energy.

Linear Theory Results

Now the linear modes of the model are presented. Since the model eqns [26]–[29] involve four dynamically coupled variables, there are four linear modes. The dispersion relation for the linear modes is shown in Figure 2. (Only the two low-frequency, intraseasonal modes are shown. The other two modes are high-frequency modes and are only weakly coupled to the wave activity.) Figure 2 shows that the skeleton model has eastward-propagating waves with phase speeds of roughly 5 m s^{−1} and the peculiar dispersion relation $d\omega/dk \approx 0$, in agreement with the MJO. Moreover, the phase

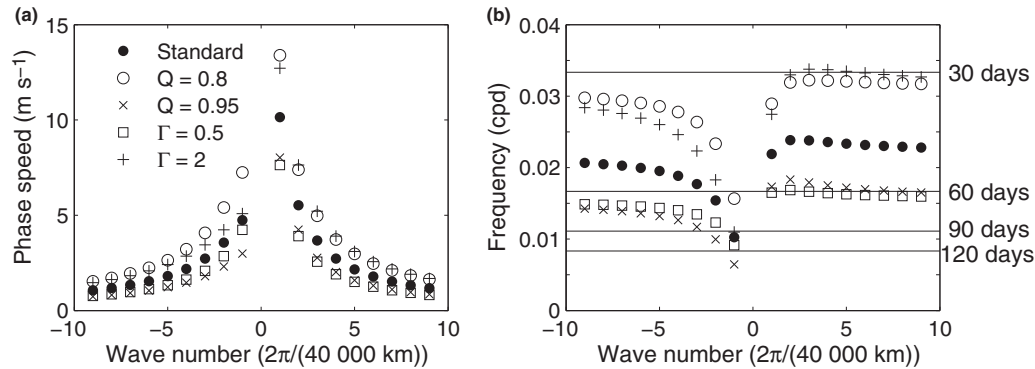


Figure 2 Linear wave phase speed ω/k (a) and oscillation frequency ω/k (b) as functions of wave number k for the low-frequency modes of eqns [26]–[29]. Filled circles denote results with the standard parameter values. Other markers denote results with one change made to the standard parameter values: $\bar{Q} = 0.8$ (open circles), $\bar{Q} = 0.95$ (crosses), $\Gamma = 0.5$ (squares), and $\Gamma = 2$ (pluses). Horizontal lines in (b) denote oscillation periods of 30, 60, 90, and 120 days. From Majda, A.J., Stechmann, S.N., 2009b. The skeleton of tropical intraseasonal oscillations. *Proceedings of the National Academy of Sciences U.S.A.* 106 (21), 8417.

speed and dispersion relation are robust over a wide range of parameter values, with the oscillation periods spanning the range of 30–60 days, which is the observed range of the MJO's oscillation period. The westward-propagating waves, on the other hand, which are plotted with positive ω and negative k , have frequencies $\omega(k)$ that vary significantly with k ; moreover, their oscillation periods are seasonal, not intraseasonal, for $k = 1$ and 2. This suggests the first piece of an explanation for the observed dominance of eastward-propagating intraseasonal variability: the westward-propagating modes have seasonal oscillation periods, on which timescales other phenomena are expected to dominate over modulations of convective activity.

The features in Figure 2 are fundamental features 1 and 2 of the MJO skeleton that were mentioned in the Introduction section. For a linear wave, these two features arise from the eigenvalues of the linear system. The second important piece from the linear system is the eigenvectors, which determine the structure of the linear waves and fundamental feature 3 from the Introduction section. When trying to match theory and observations, it is important that both the eigenvalues and the eigenvectors – i.e., all three fundamental features 1–3 – are matched together.

The physical structure of the wave number-2 MJO eigenmode is shown in Figure 3 for the standard parameter values. Horizontal quadrupole vortices are prominent, as in observations, and the maximum convective activity is collocated with the maximum in equatorial convergence. The lower tropospheric moisture leads and is in quadrature with the convective activity, which is also roughly the relationship

seen in observations. The pressure contours display the mixed Kelvin–Rossby wave structure of the wave: equatorial high pressure anomalies are collocated with the westerly wind burst as in Kelvin waves, and they are flanked by off-equatorial low pressure anomalies and cyclonic Rossby gyres, in broad agreement with the observational record. Rectification of the vertical structure and some of the phase relationships is likely due to effects of higher vertical modes as part of the MJO's 'muscle.'

The relative contributions of K , R , Q , and A to these linear waves are shown in Figure 4 for wave numbers 1–3. The MJO has significant contributions from both the Kelvin and Rossby components, whereas the westward modes are dominated by the Rossby component. In addition, the larger Q and A amplitudes suggest the second piece of an explanation for eastward-propagating rather than westward-propagating intraseasonal oscillations: the eastward-propagating modes are more strongly coupled to equatorial moist convective processes.

In addition to these illustrations, a formula for the intraseasonal oscillation frequency ω of the MJO skeleton can be obtained by considering the even simpler case of flow above the equator. In this case, v and γ are set to zero, and meridional derivatives are ignored. The result is a linear system of four equations for u' , θ' , q' , a' :

$$\frac{\partial u'}{\partial t} - \frac{\partial \theta'}{\partial x} = 0 \quad [32]$$

$$\frac{\partial \theta'}{\partial t} - \frac{\partial u'}{\partial x} = \overline{H}a' \quad [33]$$

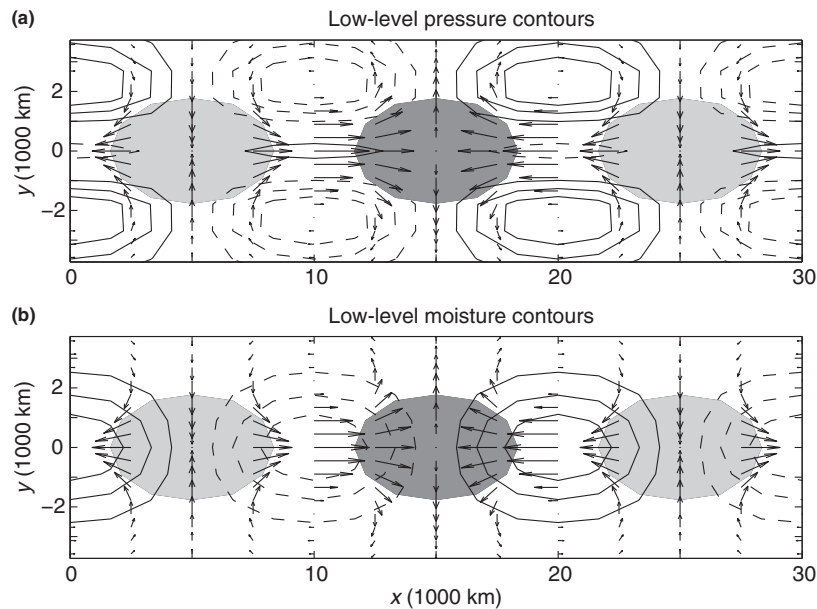


Figure 3 Physical structure of the wave number-2 Madden–Julian Oscillation mode of eqns [26]–[29] for the standard parameter values. Lower tropospheric velocity vectors are shown with contours of lower tropospheric pressure anomalies (a) and lower tropospheric moisture anomalies (b) with positive (negative) anomalies denoted by solid (dashed) lines. The contour interval is one-fourth the maximum amplitude of the anomaly, and the zero contour is not shown. Positive (negative) anomalies of convective activity A that are greater than one-half the maximum amplitude are shaded dark (light) gray. From Majda, A.J., Stechmann, S.N., 2009b. The skeleton of tropical intraseasonal oscillations. *Proceedings of the National Academy of Sciences U.S.A.* 106 (21), 8417.

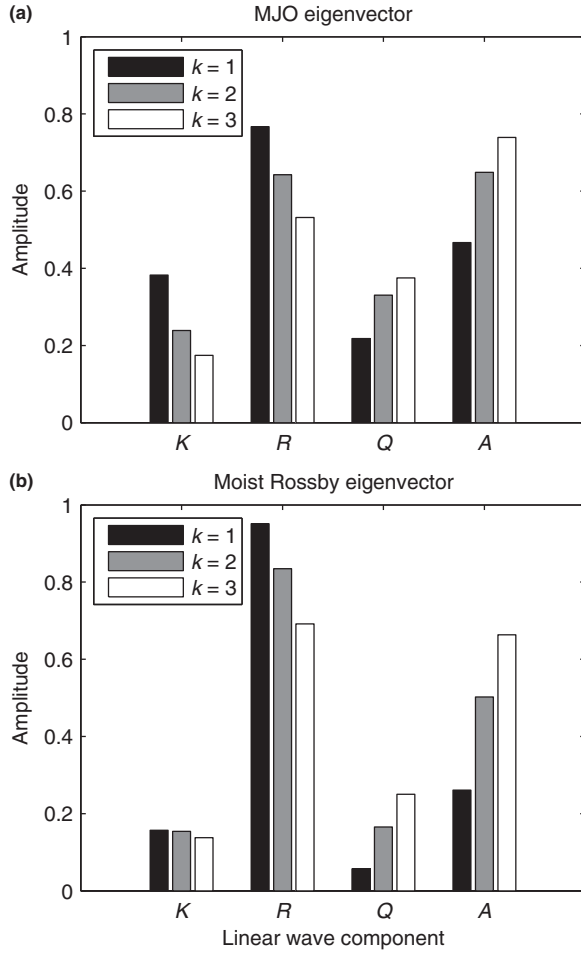


Figure 4 Contributions of each component K , R , Q , and A to the linear wave eigenvectors of the Madden–Julian Oscillation (a) and the low-frequency westward-propagating mode (b) for the standard parameter values. Results for wave numbers $k = 1, 2$, and 3 are shown in black, gray, and white, respectively. From Majda, A.J., Stechmann, S.N., 2009b. The skeleton of tropical intraseasonal oscillations. *Proceedings of the National Academy of Sciences U.S.A.* 106 (21), 8417.

$$\frac{\partial q'}{\partial t} + \tilde{Q} \frac{\partial u'}{\partial x} = -\bar{H}a' \quad [34]$$

$$\frac{\partial a'}{\partial t} = \Gamma \bar{a}q'. \quad [35]$$

This system can be solved exactly due to the perfect east–west symmetry:

$$2\omega^2 = \Gamma \bar{S}^\theta + k^2 \pm \sqrt{\left(\Gamma \bar{S}^\theta + k^2\right)^2 - 4\Gamma \bar{S}^\theta k^2 (1 - \tilde{Q})} \quad [36]$$

where k is the zonal wave number, and the radiative cooling rate $\bar{S}^\theta$ is described below eqn [29]. For the low-frequency waves, this is approximately equal to

$$\omega \approx \sqrt{\Gamma \bar{S}^\theta (1 - \tilde{Q})}, \quad [37]$$

For the standard parameter values used here, the oscillation period corresponding to eqn [37] is 45 days, in agreement with observations of the MJO. Notice that this formula is independent of the wave number k ; i.e., this model recovers the peculiar dispersion relation $d\omega/dk \approx 0$ from the observational record, and it relates the MJO frequency to the three parameters of the model.

Nonlinear Simulation

Figure 5 shows a numerical solution of the nonlinear eqns [26]–[29] for a warm-pool sea surface temperature. The convective activity is aligned over the warm pool, which is in the center of the domain from $x \approx 10\,000$ – $30\,000$ km. The MJO events have prominent phases of both active and suppressed convection, and each event has its own individual characteristics in terms of strength, lifetime, regional variations, etc. Furthermore, in addition to the prominent eastward-propagating disturbances, there are instances of localized standing oscillations throughout the domain. For instance, to illustrate this, two rectangular boxes are drawn in Figure 5: $x = 11\,000$ – $15\,000$ km, $t = 3400$ – 3470 days; and $x = 15\,000$ – $19\,000$ km, $t = 3440$ – 3530 days. Localized standing oscillations are prominent again, later, in the region $x = 15\,000$ – $19\,000$ km, $t = 3550$ – 3600 days (for comparison, no box added). The western end of the warm pool is a common location of the standing oscillations, which is in broad agreement with the visual appearance of standing oscillations in the Indian Ocean, sometimes at the beginning of an MJO event.

MJO’s ‘Muscle’

While the MJO skeleton model describes the fundamental planetary–intraseasonal features of MJO composites, individual MJO events often have additional features, such as westerly wind bursts, and different varieties of convective features within the MJO envelope. Since these additional features add detailed character to each MJO’s structure, and since these features often account for additional strength beyond the MJO’s skeleton, they are referred to here as the MJO’s ‘muscle.’

One contributor to the MJO’s muscle is CMT due to the convective systems that are embedded within the MJO’s envelope. These systems include various types of MCSs and CCEWs, and their CMT can potentially accelerate or decelerate the winds associated with the MJO skeleton.

To illustrate the CMT due to MCSs and CCEWs, an exactly solvable model is presented for their velocity anomalies (u', w') and momentum fluxes $\overline{u'w'}$ in a two-dimensional x – z setting above the equator. Consider the following exactly solvable model for a CCEW’s velocity (u', w'):

$$w'(x, z, t) = S'_\theta(x, z, t) \quad [38]$$

$$\frac{\partial u'}{\partial x} + \frac{\partial w'}{\partial z} = 0. \quad [39]$$

In this model, called the weak-temperature-gradient approximation, the wave’s vertical velocity w' is exactly in balance with the heating rate S'_θ , which we must specify. The wave’s horizontal velocity u' is then determined from the

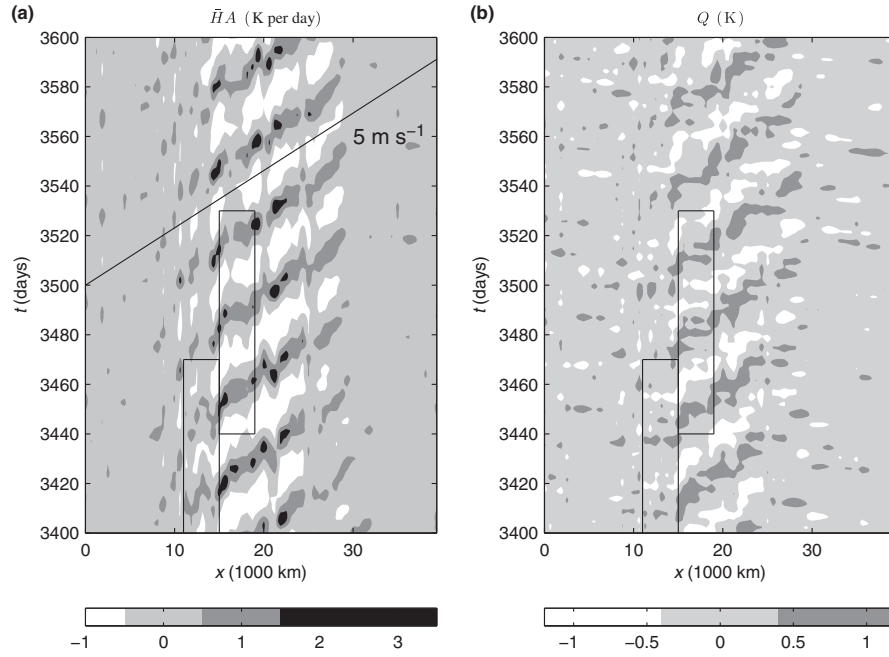


Figure 5 Numerical solution of the nonlinear eqns [26]–[29] for a warm-pool sea surface temperature. Contour plots of (a) $\bar{H}'A'(x, t)$ and (b) $Q(x, t)$. Rectangular boxes show regions of standing oscillations: $x = 11\,000$ – $15\,000$ km, $t = 3400$ – 3470 days; and $x = 15\,000$ – $19\,000$ km, $t = 3440$ – 3530 days. From Majda, A.J., Stechmann, S.N., 2011. Nonlinear dynamics and regional variations in the MJO skeleton. *Journal of Atmospheric Sciences* 68, 3053–3071.

incompressibility constraint in eqn [39]. Given this exact solution for u' and w' of the CCEW, its effect on the larger-scale mean flow is determined by

$$\frac{\partial \bar{u}}{\partial t} = -\frac{\partial}{\partial z} \overline{w'u'}, \quad [40]$$

where this is the horizontal spatial average of the horizontal momentum equation, $\partial_x u + \partial_x(u^2) + \partial_z(wu) + \partial_x p = 0$, and where bar and prime notations are used to denote a horizontal spatial average and fluctuation, respectively:

$$\bar{f}(z, t) = \frac{1}{L} \int_0^L f(x, z, t) dx \quad [41]$$

$$f'(x, z, t) = f - \bar{f}, \quad [42]$$

where periodic horizontal boundary conditions are assumed for simplicity. From eqn [40], it is seen that a CCEW will alter the mean flow if and only if $\partial_z \overline{w'u'} \neq 0$. In the context of convective motions where $w' > 0$, this effect on the mean flow is called CMT.

To illustrate CMT in some specific cases, consider a heat source with two phase-lagged vertical modes, $\sin(z)$ and $\sin(2z)$, which represent deep convective heating and congestus/stratiform heating, respectively:

$$S'_\theta = a_* \{ \cos(kx - \omega t) \sqrt{2} \sin(z) + \alpha \cos[k(x + x_0) - \omega t] \sqrt{2} \sin(2z) \}, \quad [43]$$

where k is the horizontal wave number and a_* is the amplitude of the heating. Two key parameters here are α , the relative

strength of the second baroclinic heating, and x_0 , the lag between the heating in the two vertical modes. Heating structures of this form are commonly associated with MCSs and CCEWs, due to variations in congestus, deep, and stratiform cloud populations. Figure 6 shows three cases for the lag x_0 : 0 (top), +500 km (middle), and –500 km (bottom) for a wave with wavelength 3000 km, heating amplitude $a_* = 4$ K per day, and relative stratiform heating of $\alpha = -1/4$. The lag determines the vertical tilt of the heating profile. Given this heating rate, the velocity can be found exactly from eqn [39]:

$$u'(x, z, t) = -\frac{a_*}{k} \{ \sin(kx - \omega t) \sqrt{2} \cos(z) + 2\alpha \sin[k(x + x_0) - \omega t] \sqrt{2} \cos(2z) \} \quad [44]$$

$$w'(x, z, t) = a_* \{ \cos(kx - \omega t) \sqrt{2} \sin(z) + \alpha \cos[k(x + x_0) - \omega t] \sqrt{2} \sin(2z) \} \quad [45]$$

With this form of u' and w' , the eddy flux divergence is

$$\frac{\partial}{\partial z} \overline{w'u'} = \frac{3}{2} \frac{\sin(kx_0)}{k} a_*^2 \alpha [\cos(z) - \cos(3z)] \quad [46]$$

Notice that a wave with first and second baroclinic components generates CMT that affects the first and *third* baroclinic modes. Also notice that eqn [46] is nonzero as long as $\alpha \neq 0$ (i.e., there are both first and second baroclinic mode contributions) and $x_0 \neq 0$ (i.e., there is a phase lag between the first and second baroclinic modes). MCSs and CCEWs in nature typically have this type of structure.

For illustrations of the above exact solutions, consider the three cases shown in Figure 6: upright updraft (top), westward

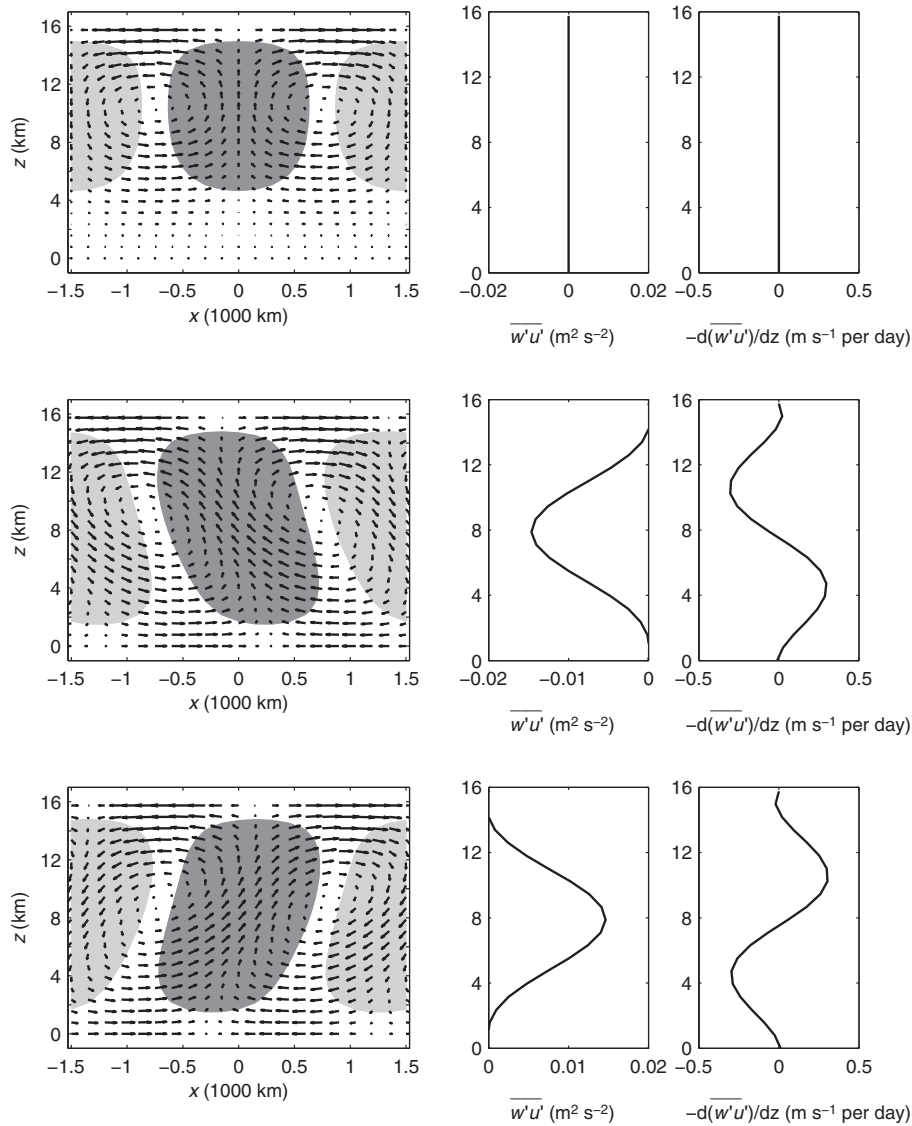


Figure 6 Solutions to the exactly solvable model eqns [38] and [39] for convectively coupled equatorial waves (CCEWs) structure and convective momentum transport in three cases: upright updraft (top), vertically tilted updraft of ‘eastward propagating’ CCEW (middle), and vertically tilted updraft of ‘westward propagating’ CCEW (bottom). Left: vector plot of (u', w') and shaded convective heating $S'_\theta(x, z)$. For vectors, the maximum u' is 6.0 m s^{-1} for the top and 4.0 m s^{-1} for the middle and bottom, and the maximum w' is 2.8 cm s^{-1} for the top and 2.2 cm s^{-1} for the middle and bottom. Dark shading denotes heating, and light shading denotes cooling, with a contour drawn at one-fourth the maximum and minimum values. Middle: vertical profile of the mean momentum flux $\overline{w'u'}$. Right: negative vertical derivative of the mean momentum flux $-\partial_z \overline{w'u'}$. From Stechmann, S.N., Majda, A.J., Skjorshammer, D., 2013. Convectively coupled wave–environment interactions. *Theoretical and Computational Fluid Dynamics* 27, 513–532. DOI [10.1007/s00162-012-0268-8](https://doi.org/10.1007/s00162-012-0268-8).

propagating’ CCEW (middle), and ‘eastward propagating’ CCEW (bottom). Although there is no inherent definitive propagation in the exactly solvable model eqns [38] and [39], propagation direction labels are assigned to the vertical tilt directions according to the structures of observed CCEW: heating is vertically tilted with leading low-level heating and trailing upper-level heating with respect to the CCEW propagation direction. Specifically, the middle row of Figure 6 corresponds to the observed structures of convectively coupled Kelvin waves, which propagate eastward, and the bottom row of Figure 6 corresponds to westward-propagating inertia-gravity waves (also called ‘two-day waves’). Also shown in

Figure 6 are the average vertical flux of horizontal momentum, $\overline{w'u'}$, and its vertical derivative, $\partial_z \overline{w'u'}$. These exact solutions show that upright updrafts have zero CMT, and tilted updrafts have nonzero CMT with the sign determined by the CCEW’s propagation direction. Note that the vertically averaged momentum would not be affected by CMT in this model, since $\overline{w'u'}$ is necessarily zero at the upper and lower rigid boundaries. Westerly wind bursts, which refer to lower tropospheric winds, can be accelerated by CMT due to eastward-propagating MCSs and CCEWs, as illustrated in the middle row of Figure 6.

One effect that is not included in the model eqns [38] and [39] is the effect of the larger-scale mean wind $\bar{u}(z, t)$ on the

MCSs or CCEWs. This could affect whether eastward- or westward-propagating convective system is favored in a given environment, which, in turn, affects the CMT. Moreover, the larger-scale mean wind $\bar{u}(z, t)$ is also needed, in addition to $\overline{w'u'}$, in order to classify momentum transport as either ‘upscale’ or ‘downscale.’ For example, according to eqn [40], if $-\partial_z \overline{w'u'} > 0$ and $\bar{u}(z, t) > 0$ at some level z , then the large-scale mean wind $\bar{u}(z, t)$ is accelerated, and the momentum transport is ‘upscale’ because the smaller-scale MCSs or CCEWs cause an increase in the large-scale mean wind. Conversely, in the same situation but with $\bar{u}(z, t) < 0$, the large-scale mean wind is decelerated, and the momentum transport is ‘downscale.’

See also: Dynamical Meteorology: Kelvin Waves; Primitive Equations; Rossby Waves; Waves. **Tropical Meteorology and Climate:** Equatorial Waves; Intraseasonal Oscillation (Madden–Julian Oscillation).

Further Reading

- Biello, J.A., Majda, A.J., 2005. A new multiscale model for the Madden–Julian oscillation. *Journal of Atmospheric Sciences* 62, 1694–1721.
- Khouider, B., Majda, A.J., 2006. A simple multicloud parameterization for convectively coupled tropical waves. Part I: linear analysis. *Journal of Atmospheric Sciences* 63, 1308–1323.

- Kiladis, G.N., Straub, K.H., Haertel, P.T., 2005. Zonal and vertical structure of the Madden–Julian oscillation. *Journal of Atmospheric Sciences* 62, 2790–2809.
- Majda, A.J., 2003. *Introduction to PDEs and Waves for the Atmosphere and Ocean*, Courant Lecture Notes in Mathematics, vol. 9. American Mathematical Society, Providence, RI, x + 234 pp.
- Majda, A.J., Stechmann, S.N., 2009a. A simple dynamical model with features of convective momentum transport. *Journal of Atmospheric Sciences* 66, 373–392.
- Majda, A.J., Stechmann, S.N., 2009b. The skeleton of tropical intraseasonal oscillations. *Proceedings of the National Academy of Sciences U.S.A.* 106 (21), 8417–8422.
- Majda, A.J., Stechmann, S.N., 2011. Nonlinear dynamics and regional variations in the MJO skeleton. *Journal of Atmospheric Sciences* 68, 3053–3071.
- Moncrieff, M.W., 2010. The multiscale organization of moist convection and the intersection of weather and climate. In: Sun, D.-Z., Bryan, F. (Eds.), *Climate Dynamics: Why Does Climate Vary?*, Geophysical Monograph Series, vol. 189. American Geophysical Union, Washington, D.C., pp. 3–26.
- Tian, B., Waliser, D., Fetzer, E., Lambrechts, B., Yung, Y., Wang, B., 2006. Vertical moist thermodynamic structure and spatial–temporal evolution of the MJO in AIRS observations. *Journal of Atmospheric Sciences* 63 (10), 2462–2485.
- Zhang, C., 2005. Madden–Julian oscillation. *Reviews of Geophysics* 43. RG2003, <http://dx.doi.org/10.1029/2004RG000158>.

Monsoon: Overview

J Slingo, University of Reading, Reading, UK

© 2003 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 3, pp 1365–1370, © 2003 Elsevier Ltd.

Monsoon derives from the Arabic word 'mausam', meaning season, and in its broadest definition describes those climates that are seasonally *arid*. As shown in Figure 1, many regions of the tropics and subtropics experience a rainy summer season and a dry winter season, although regions close to the Equator can often experience two rainy seasons; e.g., equatorial east Africa with its 'long' (March–May) and 'short' (October–December) rains. The main driver of this marked seasonality in rainfall is the change in the distribution of surface heating between winter and summer, primarily associated with seasonal variations in the position of the sun. Because of this close relationship with the solar seasonal cycle, the start of the rainy season often begins with remarkable regularity each year.

Although Figure 1 shows that many regions are seasonally arid, the more precise definition of a monsoon climate, as proposed by Ramage, identifies South Asia, Australia, and Africa as having distinct monsoons. Ramage's criteria for a monsoon to exist are as follows:

1. Prevailing wind direction shifts by at least 120° between January and July.
2. Prevailing wind direction persists for at least 40% of the time in January and July.

3. Mean wind exceeds 3 m s^{-1} in either month.
4. Fewer than one cyclone–anticyclone alternation occurs every 2 years in either month in a 5° latitude–longitude rectangle.

These criteria essentially demand that a monsoon be characterized by a wind regime that is steady, sustained and therefore inherently driven by the seasonally evolving boundary conditions, such as land or ocean surface temperatures. It excludes most extratropical regions that are characterized by synoptic weather systems with alternating cyclonic–anticyclonic circulations.

Based on these criteria, Figure 2 shows that this major reversal in the seasonal wind regimes only occurs over (1) India and South-East Asia, (2) northern Australia, and (3) West and central Africa. These three regions constitute the major monsoons of the global circulation. Although the Americas and South Africa also experience a strong seasonal cycle in rainfall, the prevailing wind direction is largely unchanged between winter and summer (Figure 2), and so strictly speaking cannot be classified as monsoon regions. The name 'monsoon' is often used to denote the rainy season, but in fact can relate to both extremes of the seasonal cycle. The term winter monsoon is used in South-East Asia to describe the dry, north-easterly

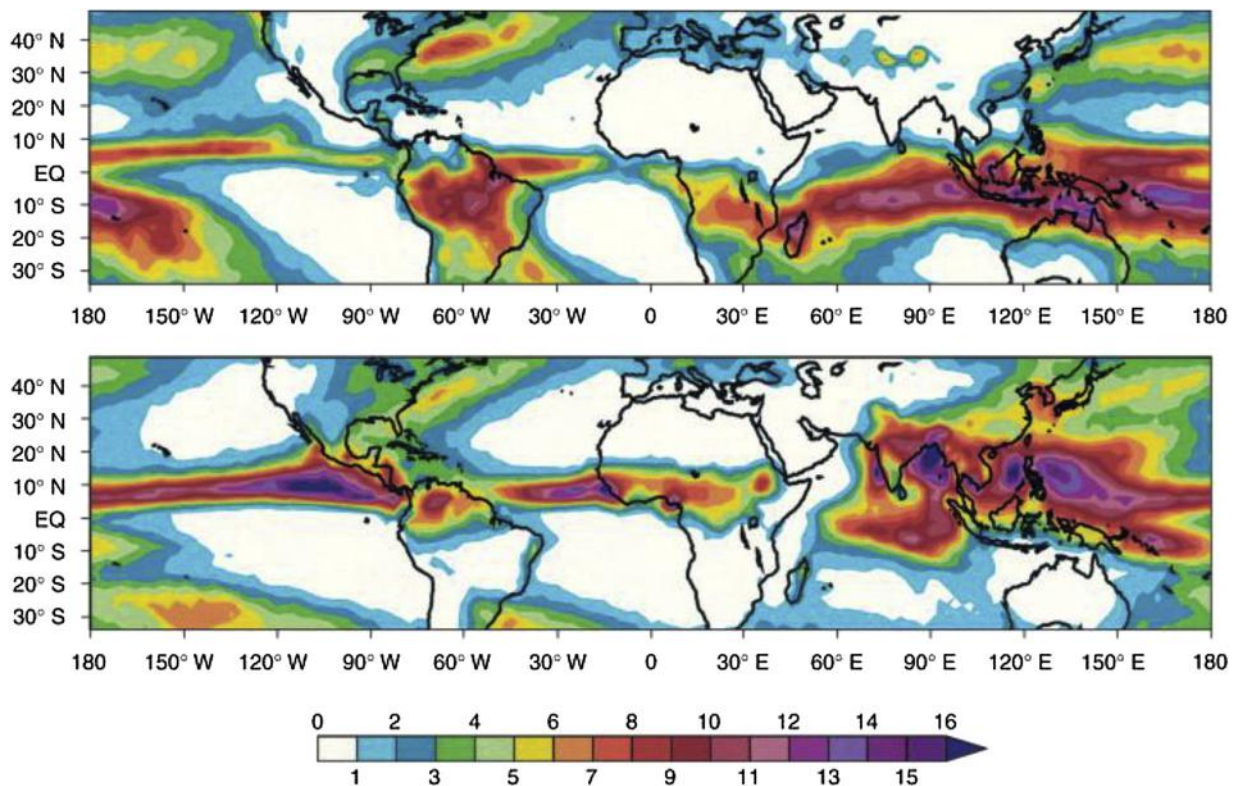


Figure 1 Mean precipitation distributions for northern winter (January–February; upper panel) and northern summer (July–August; lower panel) at the height of the monsoon season (units: mm day^{-1}). (Data source: Global Precipitation Climatology Project.)

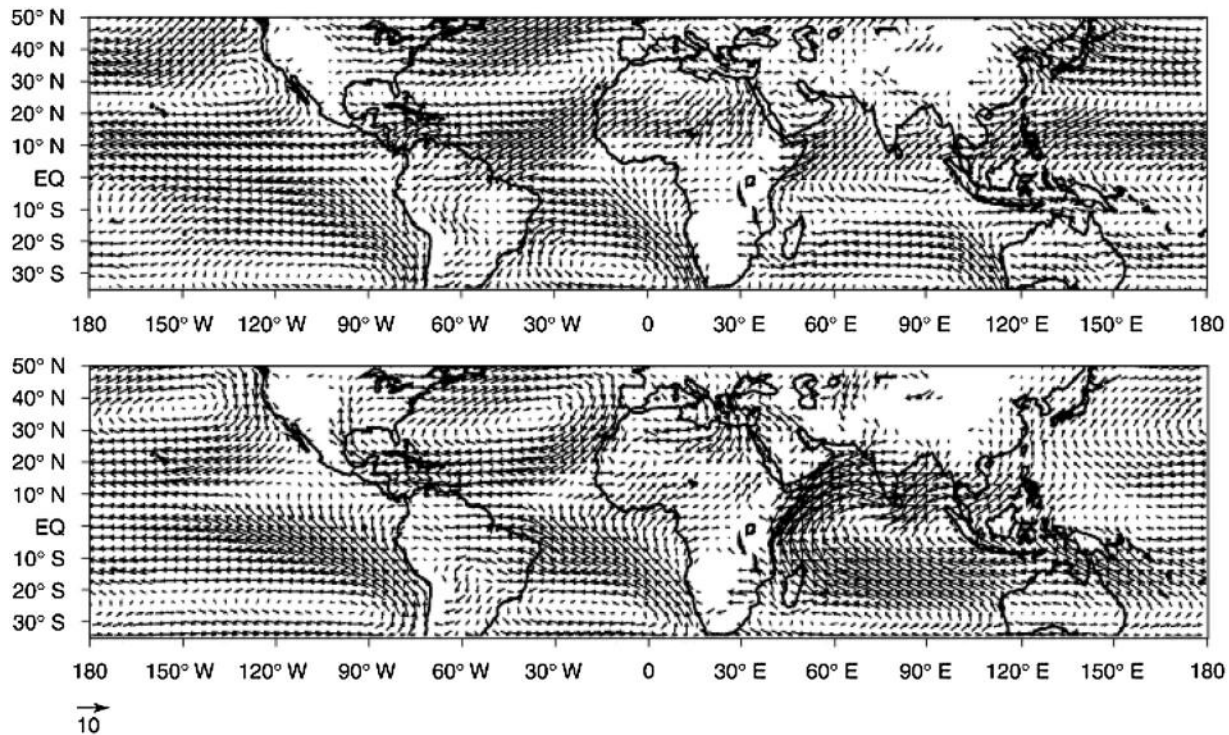


Figure 2 Mean winds at 925 hPa for northern winter (January–February; upper panel) and northern summer (July–August; lower panel) at the height of the monsoon season (units: m s^{-1}). (Data source: Reanalyses from European Centre for Medium Range Weather Forecasts for 1979–93.)

winds that prevail over the northern Indian Ocean and South China Sea during boreal winter (Figure 2, upper panel).

For a monsoon to be established, a thermal contrast between the land and ocean must exist. This occurs when large land masses, such as Asia, Africa, and Australia, heat up rapidly during the spring and summer (Figure 3). Since the thermal inertia of the land is much less than the surrounding oceans, the continents respond much more rapidly to the seasonal cycle in solar heating, setting up large temperature gradients. These hot land masses draw humid air in from the surrounding oceans, like a massive sea breeze (Figure 2). As the moisture-laden air reaches the warm land, it rises, the moisture condenses, and the rainy season begins. By contrast, in winter the land becomes much cooler than the surrounding oceans and cold, dry air then flows from the land out over the ocean. Often the two monsoons, winter and summer, are closely linked with the winter monsoon of one hemisphere feeding the summer monsoon of the other. For example, in the Asian–Australian monsoon system, the dry air from the winter continent flows across the Equator toward the summer hemisphere (Figure 2), picking up moisture from the warm oceans and feeding the monsoon rains over the summer continent.

A critical factor that determines the generation of a monsoon is the geographical orientation of the oceans and continents. The strongest monsoons occur where there is a pronounced north–south distribution in land and ocean that can take advantage of the north–south progression of the solar seasonal cycle. As Figure 3 shows, the largest land–sea temperature contrasts occur over the seasonally arid regions of North Africa, India, and Australia, during the months preceding

the summer monsoon. These very warm temperatures lead to the development of thermal lows which serve to pull in air from surrounding regions. Once the monsoon is established and the rains begin, the land surface temperatures tend to cool due to the increased soil wetness, but the atmospheric warming from latent heat release associated with the monsoon rains (Figure 1) maintains the low-pressure regions (Figure 4) which continue to drive the monsoon winds.

Although this classic description of monsoons provides the fundamental basis for their existence, there are important regional differences associated with the shape of continents, orography (particularly mountain barriers), and ocean temperatures. For the Asian summer monsoon, the Tibetan plateau acts as an elevated heat source which clearly influences the establishment and maintenance of the monsoon circulation. The seasonal heating of the plateau leads to a reversal of the meridional temperature gradient which extends throughout the troposphere (Figure 5). This reversal is instrumental in triggering the large-scale seasonal change in the circulation over East Asia, with the poleward transition of the subtropical jet and the onset of the monsoon over the Indian subcontinent.

The importance of this deep warm core over South Asia is demonstrated in Figure 6, which shows the winds in the free troposphere at 700 hPa. These winds can be compared with the boundary layer winds at 925 hPa in Figure 2. It is only for the domain of the Asian monsoon that the seasonal reversal of the winds is seen extending above the boundary layer into the free troposphere. This is important because it is only when there is advection of moist air through a substantial depth of the troposphere that sustained monsoon rainfall is achieved.

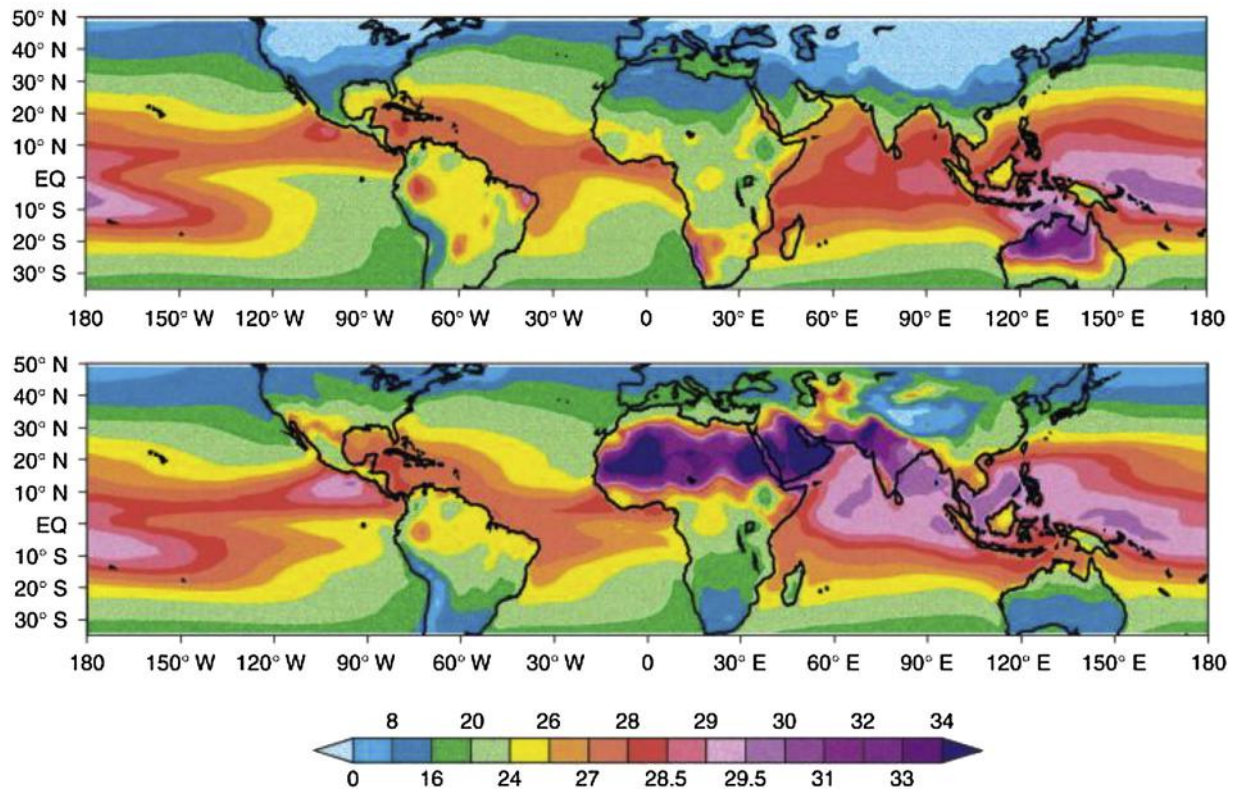


Figure 3 Surface temperatures for early austral summer (November–December; upper panel) and boreal summer (May–June; lower panel) at the onset of the monsoon (units: °C). (Data source: Reanalyses from European Centre for Medium Range Weather Forecasts for 1979–93.)

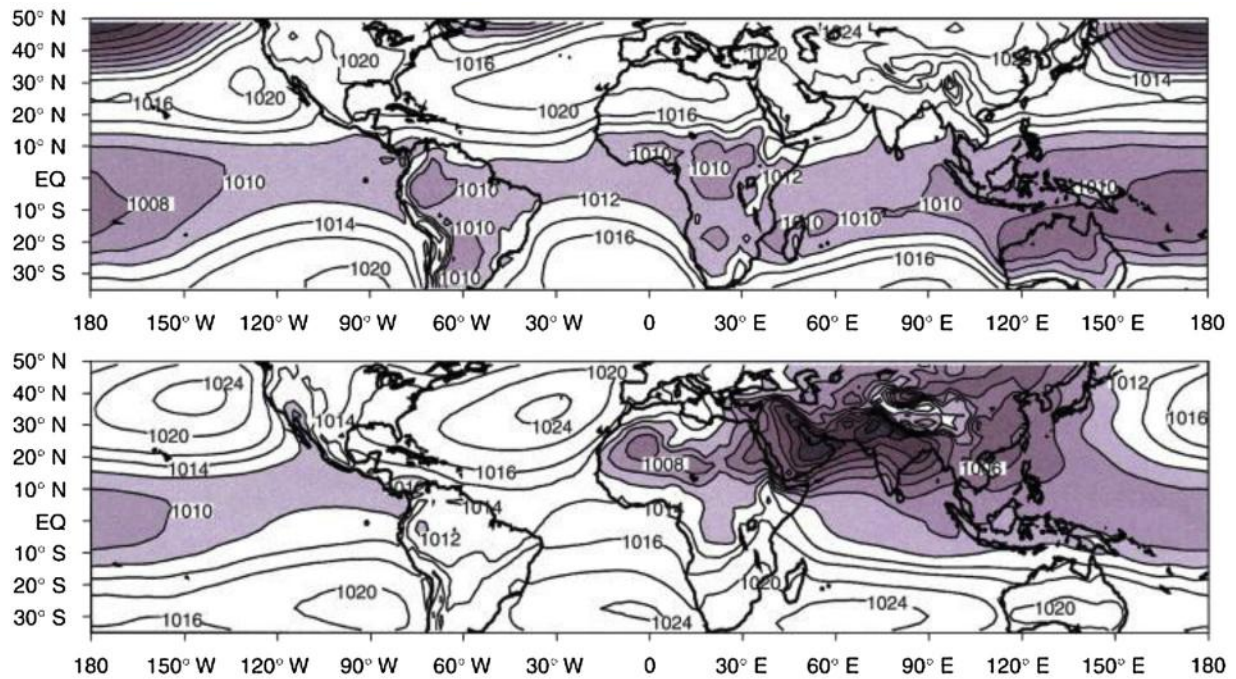


Figure 4 Mean sea-level pressure for northern winter (January–February; upper panel) and northern summer (July–August; lower panel) at the height of the monsoon season (units: hPa). (Data source: Reanalyses from European Centre for Medium Range Weather Forecasts for 1979–93.)

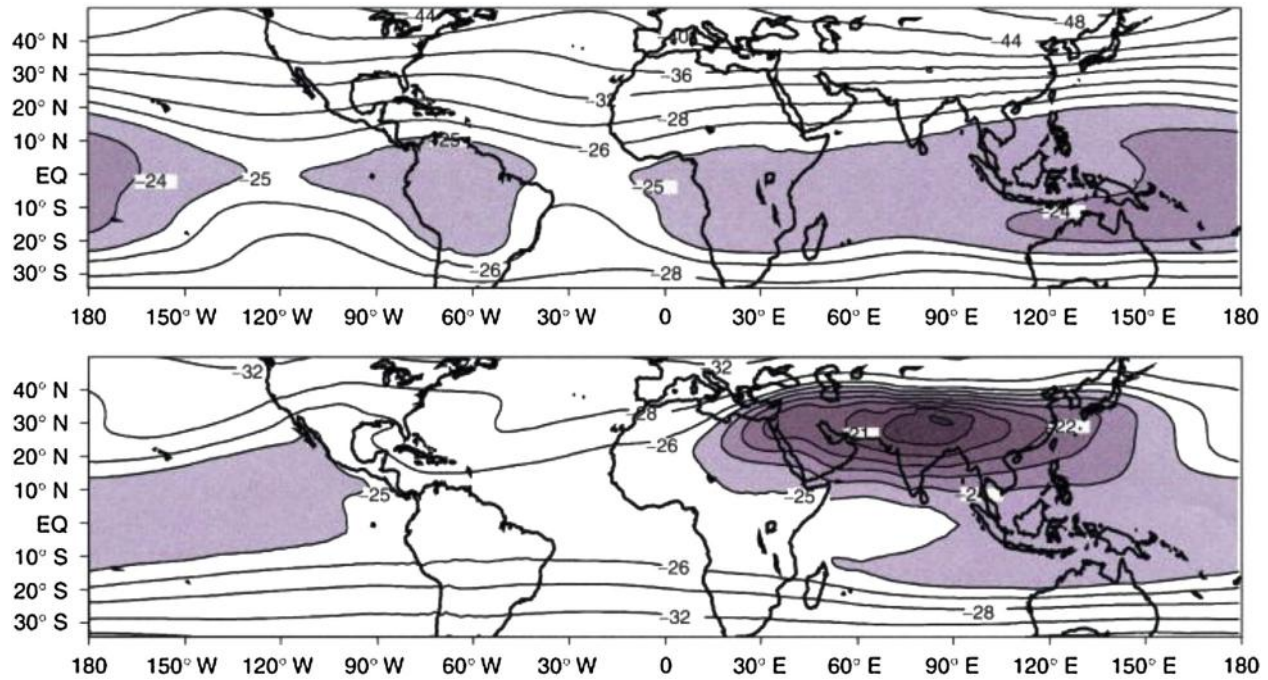


Figure 5 Mean temperature between 200 and 500 hPa for northern winter (January–February; upper panel) and northern summer (July–August; lower panel) at the height of the monsoon season (units: K). (Data source: Reanalyses from European Centre for Medium Range Weather Forecasts for 1979–93.)

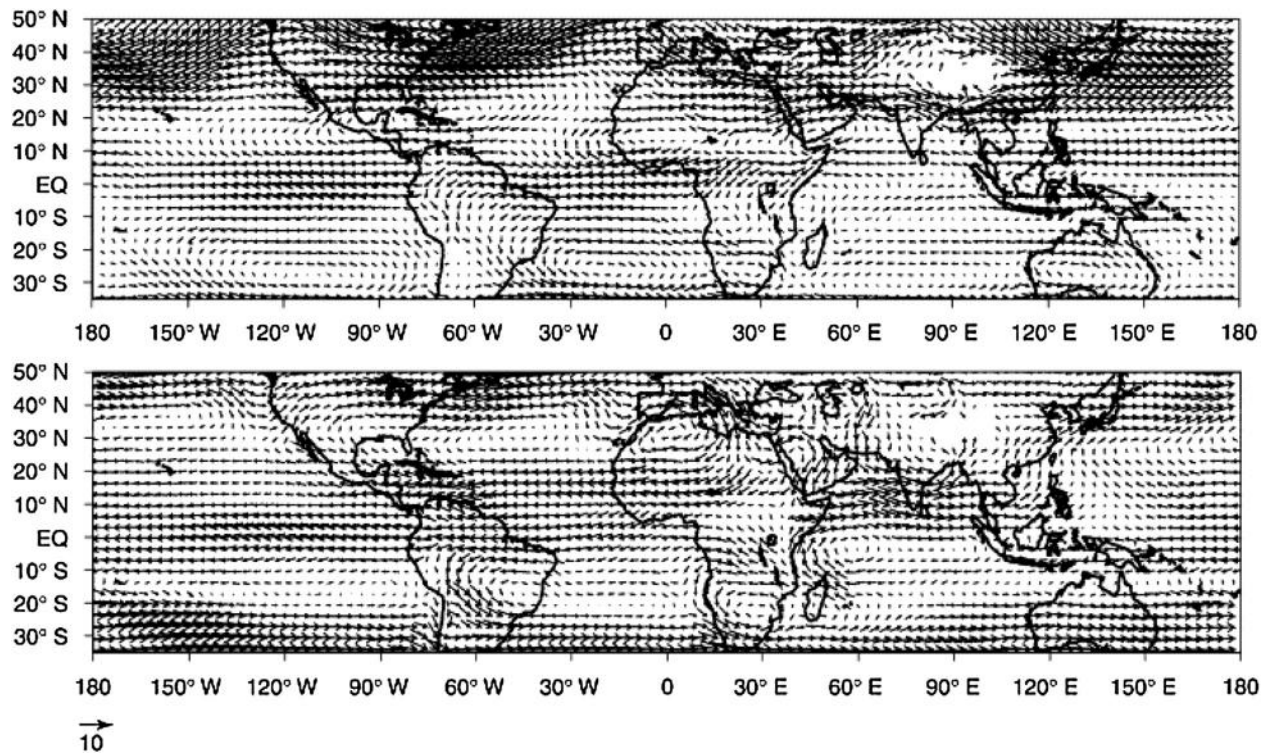


Figure 6 Mean winds at 700 hPa for northern winter (January–February; upper panel) and northern summer (July–August; lower panel) at the height of the monsoon season (units: m s^{-1}). (Data source: Reanalyses from European Centre for Medium Range Weather Forecasts for 1979–93.)

The significance of the Tibetan plateau is further demonstrated when a comparison is made between the summer monsoons of South Asia, northern Australia, and West Africa. In the absence of an orographic heat source, the seasonal reversal of the meridional temperature gradient through the depth of the troposphere is barely evident over Australia and West Africa. This is despite the very substantial surface warming of Australia and North Africa shown in [Figure 3](#). The effect of the confinement of the seasonal reversal in the meridional temperature gradient to the near-surface layers over Africa and Australia can be seen in the winds at 700 hPa ([Figure 6](#)). Unlike South-East Asia, there is no seasonal reversal of the winds in the free troposphere so that a deep moist layer is not established in the same way as over South Asia. Consequently, the monsoon rains of West Africa and Australia are not as intense, nor do they extend as far polewards.

Monsoons are crucial elements of the global circulation and monsoon rainfall provides the water needed by over 60% of the world's population. Understanding and predicting how monsoons may change from year to year, and the result of global warming are key scientific, economic and societal issues. The management of water resources is a top priority for monsoon-affected countries to enable the population to survive from one rainy season to the next. Food production in seasonally arid areas is also inherently risky. By the end of the

dry season, the soil is parched and planting cannot begin until the rains arrive. A late or weak monsoon can lead to a short or poor growing season and hence low yields. Agricultural failure has a profound effect on the economy of monsoon-affected countries, such as India, where farming accounts for 30% of the gross domestic product and 67% of the workforce.

See also: Tropical Meteorology and Climate: Monsoon: Dynamical Theory; Monsoon: ENSO–Monsoon Interactions; Tropical Climates.

Further Reading

- Fein, J.S., Stephens, P.L. (Eds.), 1987. *Monsoons*. Wiley-Interscience, New York, USA.
- Hastenrath, S., 1994. *Climate Dynamics of the Tropics: An Updated Edition of Climate and Circulation of the Tropics*. Kluwer, Norwell, MA.
- Pant, G.B., Rupa Kumar, K., 1997. *Climates of South Asia*. Belhaven Studies in Climatology. Wiley, Chichester.
- Ramage, C., 1971. Monsoon Meteorology. In: *International Geophysics Series*, vol. 15. Academic Press, San Diego, CA.
- Webster, P.J., Magana, V.O., Palmer, T.N., et al., 1998. Monsoons: processes, predictability, and the prospects for prediction. *Journal of Geophysical Research* 103 (C7), 14 451–14 510.

Monsoon: Dynamical Theory

PJ Webster and J Fasullo, University of Colorado – Boulder, Boulder, CO, USA

© 2003 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 3, pp 1370–1386, © 2003, Elsevier Ltd.

Elements of a Monsoon Circulation

A monsoon is a circulation system with certain well-defined characteristics. During summer, lower tropospheric winds flow toward heated continents away from the colder oceanic regions of the winter hemisphere. In the upper troposphere the flow is reversed, with flow from the summer to the winter hemisphere. Precipitation generally occurs during summer, centered in time on either side of the summer solstice and located over the heated continents and the adjacent oceans and seas in the vicinity of a trough of low pressure referred to as the 'monsoon trough'. Most summer rainfall is associated with synoptic disturbances that propagate through the region. However, these disturbances are grouped in periods lasting from 10 to 30 days. Such envelopes of disturbed weather and heavy rainfall are referred to as 'active periods of the monsoon'. The intervening periods of mini-drought are referred to as 'monsoon breaks'. The location of the monsoon trough and axis of heavy monsoon precipitation is generally well poleward of the position of the oceanic intertropical convergence zone (ITCZ), within which the majority of tropical oceanic precipitation occurs. For example, the rainfall associated with the South Asian monsoon falls at the same latitudes as the great deserts of the planet.

Monsoon systems are associated with colocated pairs of continents such as Asia and Australia, or continents straddling the Equator such as north-west and south-west Africa, and North and South America defining, respectively, the Asian–Australian monsoon system, the West African monsoon, and the American monsoon. Each system is different in terms of intensity and circulation characteristics. For example, the northern arm of the American monsoon is a relatively weak counterpart of the other major monsoon systems and there does not appear to be a discernible cross-equatorial component during the summer. In that sense, the North and South American monsoons may be thought of as almost separate entities. Rainfall that occurs over the continents that span the Equator (e.g., equatorial Africa and South America, and Indonesia) is not strictly monsoonal and possesses double rainfall maxima occurring with the equinoxes. Monsoon climates, on the other hand, possess a single solstitial rainfall maximum, while solstices demark the dry seasons for equatorial climates.

Basic Driving Mechanisms of the Monsoon

It is helpful to consider first a simple prototype geography that will allow us to identify the basic elements of a monsoon system. The geographical model we adopt is an oceanic planet with a continental cap extending from the subtropics to the pole in one hemisphere. After establishing the important processes that drive the monsoon for this simple geography, we will return to the consideration of local influences.

Monsoons arise from the development of cross-equatorial pressure gradients produced or modified by the following physical properties of, or processes associated with, the land–ocean–atmosphere system: differential heating of land and ocean produced by the different heat capacity of land and water; the different manner in which heat is transferred vertically and stored in the ocean and the land; modification of differential heating by moist processes; the generation of meridional pressure-gradient forces resulting from the differential heating; and the meridional transport of heat in the ocean by dynamical processes. Each of these processes and properties has to be considered relative to the rotation of the planet, and the influence of local effects such as the geography of the ocean and the land masses, and regional topography.

Differential Heating

Heat capacity differences

There is roughly a factor of 4 difference between the specific heat of water ($4218 \text{ J kg}^{-1} \text{ K}^{-1}$) and dry land (roughly $1300 \text{ J kg}^{-1} \text{ K}^{-1}$). Wet soil may have a heat capacity 30% higher than dry soil. For some net heating rate, the temperature of a mass of dry land the increment in temperature will be nearly four times greater than that of a similar mass of water. In the late seventeenth century, Halley (of Halley's comet) was the first to suggest that monsoon circulations were driven by heating gradients produced by the heat capacity differences between the land and the ocean and used his theory to explain aspects of the West African and South Asian surface monsoon winds that had been reported by explorers and traders. He also understood the role of the annual cycle of solar heating that produced the strong seasonality of the monsoon and the reversal of the circulation during the winter.

Halley had defined a basic factor that determines the existence of monsoons. However, to understand how different heat capacities produce motion and why the Halley's theory has to be expanded, it is necessary to delve deeper into the physics of the atmosphere and the ocean. If the heat flux into the surface layer is $F \text{ (W m}^{-2}\text{)}$ and if there is no heat flux out of the bottom of the layer at some depth $z = z_1 \text{ (m)}$, the heating rate of the layer will be determined by the flux divergence in the layer (eqn [1]).

$$\frac{dT}{dt} = -\frac{1}{\rho C_p} \frac{dF}{dz} = \frac{1}{\rho C_p} \frac{F_{z=0}}{\Delta z} \quad [1]$$

In eqn [1], $F_{z=0}$ is the net flux at the surface and Δz is the thickness of the layer. The surface energy balance is given by eqn [2], where I_{net} is the net radiation at the surface given by the sum of the net solar radiation, the upwelling infrared radiation, and the re-radiation from the atmosphere (the greenhouse effect), respectively.

$$F_{z=0} = I_{\text{net}} - H_s - H_e, \quad [2]$$

I_{net} is given by eqn [3], where S is the solar flux at the surface, a is the system albedo, ε is the emissivity of the atmosphere, T_g and T_a are the surface and atmospheric temperatures; H_s and H_e are the sensible and latent turbulent heat transports away for the surface as described in Figure 1.

$$I_{\text{net}} = S(1 - a) + \varepsilon\sigma T_g^4 - \sigma T_a^4 \quad [3]$$

From eqn [1] it is apparent that the heating rate of a slab will depend on the heat capacity the layer, its thickness, and the net energy flux into or out of the layer at the surface. The temperature of a motionless oceanic slab (i.e., no vertical mixing or horizontal advection) is determined by the net heating at the surface. South of the Equator, in the winter hemisphere, the slab ocean would cool by a combination of evaporative cooling and negative net radiational heating. To the north of the Equator, the ocean would heat if net radiational heating exceeded the evaporative cooling. During summer the land heats more rapidly than the adjacent ocean because of its smaller specific heat and shallow Δz . These factors easily compensate for the fact that dry land has a larger albedo than the ocean (20–40% versus about 10%). In the winter the land surface will cool much more quickly than the ocean simply because there is little available heat in the subsurface that can be made available to heat the surface on seasonal time scales because of the slowness of the diffusive processes. Given that the sensible heat exchange between the land surface and the atmosphere depends to a large degree on their temperature difference, the atmospheric column over the land will be warmer than over the ocean.

In the simple model described above, the differences between the heating rates of land and the ocean reside in their different heat capacities and densities, and the depth Δz of the slab that is defined as the depth over which the heating is spread. Over land, the Δz is very small because of the opacity of the soil to radiation, and the depth of the ‘active’ layer in which there is a discernible signal of the annual cycle is only a meter or so. This depth is constrained by the inefficiency of conductive heat transfer. If the ocean is assumed to be immobile, then its effective depth over which heating is spread may be defined as the e -folding depth of solar radiation. Here radiative transfer and the opacity of the ocean determine the effective depth. Observations suggest that the solar radiation e -folding depth is about 10 m. The ocean temperature variation will also lag the surface heating. This may be seen by setting the surface flux, $F(z = 0)$ in eqn [1], proportional to $\sin(\omega t)$. In this case the temperature variation will be proportional to $-\cos(\omega t)$, therefore lagging the forcing by a quarter period.

In summary, when the ocean is assumed to be immobile, an annual cycle of ocean–land temperature difference and meridional pressure-gradient force is achieved. Substitution of numbers into eqns [1], [2] and [3] shows that the variability of the ocean temperature is much larger than observed as long as a depth of about 10 m is used. More realistic amplitude can be achieved by ‘tuning’ the depth of the active ocean layer to be considerably greater. It turns out there are good physical reasons why we may expect a deeper active layer.

Mixing and storage of heat

So far, the fluid nature of the ocean has been ignored. In a fluid, wind forcing and gravitational instabilities formed by the

cooling of the surface layer may induce turbulence and mixing of the surface and subsurface water. Wind stress also can move a body of water horizontally, producing ocean currents that can advect heat and mass from one place in an ocean basin to another. The impact of lateral transports will be considered later. Stable layers near the surface can be produced by the freshening effect of precipitation. These fresh layers may reduce the impact of wind stirring. With these factors in mind, we can return to the consideration of the heat balances of the ocean and the land regions and the atmospheres above. These processes are shown schematically in Figure 1.

During the summer, when the net heating of the ocean surface is positive, wind-induced turbulence mixes the warm surface water downward. As long as the net surface flux of energy into the ocean is positive, wind mixing will increase the heat content (or heat storage) of an ocean column. The mixing is very effective and observations show that in the tropical ocean a constant-temperature mixed layer may extend down below the surface to depths of 50–100 m. During winter, when the net surface heating is negative, the colder surface water (formed by the negative heat balance at the surface) is mixed downward to be replaced by warmer subsurface water that had been mixed down into the ocean column during the previous summer. As long as the surface energy balance is negative, wind-induced turbulence will decrease the total heat content (i.e., reduce heat storage) in the ocean column. As turbulent mixing occurs over a much deeper layer than the e -folding depth of solar radiation, the heat absorbed in the surface layer of the ocean is spread through a depth greater than the e -folding penetration depth of solar radiation. That is, Δz is larger and the overall sea surface temperature (SST) changes are smaller in the presence of turbulence than if the ocean were immobile.

The impact of changes in heat storage on the ocean temperature is twofold. First, it moderates the SST, which in turn modulates the temperature and moisture content of the air adjacent to the ocean surface. Atmospheric turbulent mixing produced either mechanically by wind stress or by buoyancy effects extends the imprint of the SST into the troposphere. Second, the mixing processes in the ocean column produces the observed lags between the ocean temperature and the solar cycle. Land surface temperature tends to follow the solstices, although, because of moist processes, the maximum land temperature occurs before the onset of the summer rains.

The Generation of Monsoonal Pressure Gradient Forces

To account for the observed reversal of the monsoon circulation with height we require a basic driving force that changes in magnitude or reverses with height. The only force available is the horizontal pressure-gradient force. It is relatively simple to show that the horizontal pressure-gradient force between the summer and winter hemispheres may change with height and, under certain circumstance, even reverse. Eliminating density between the equation of state and the hydrostatic equation gives eqn [4].

$$\frac{\partial p}{\partial z} = -\frac{g}{R\bar{T}} \quad [4]$$

In eqn [4], $p(z)$ is the atmospheric pressure, g is the acceleration due to gravity, R is the gas constant, and $\bar{T}$ is the mean

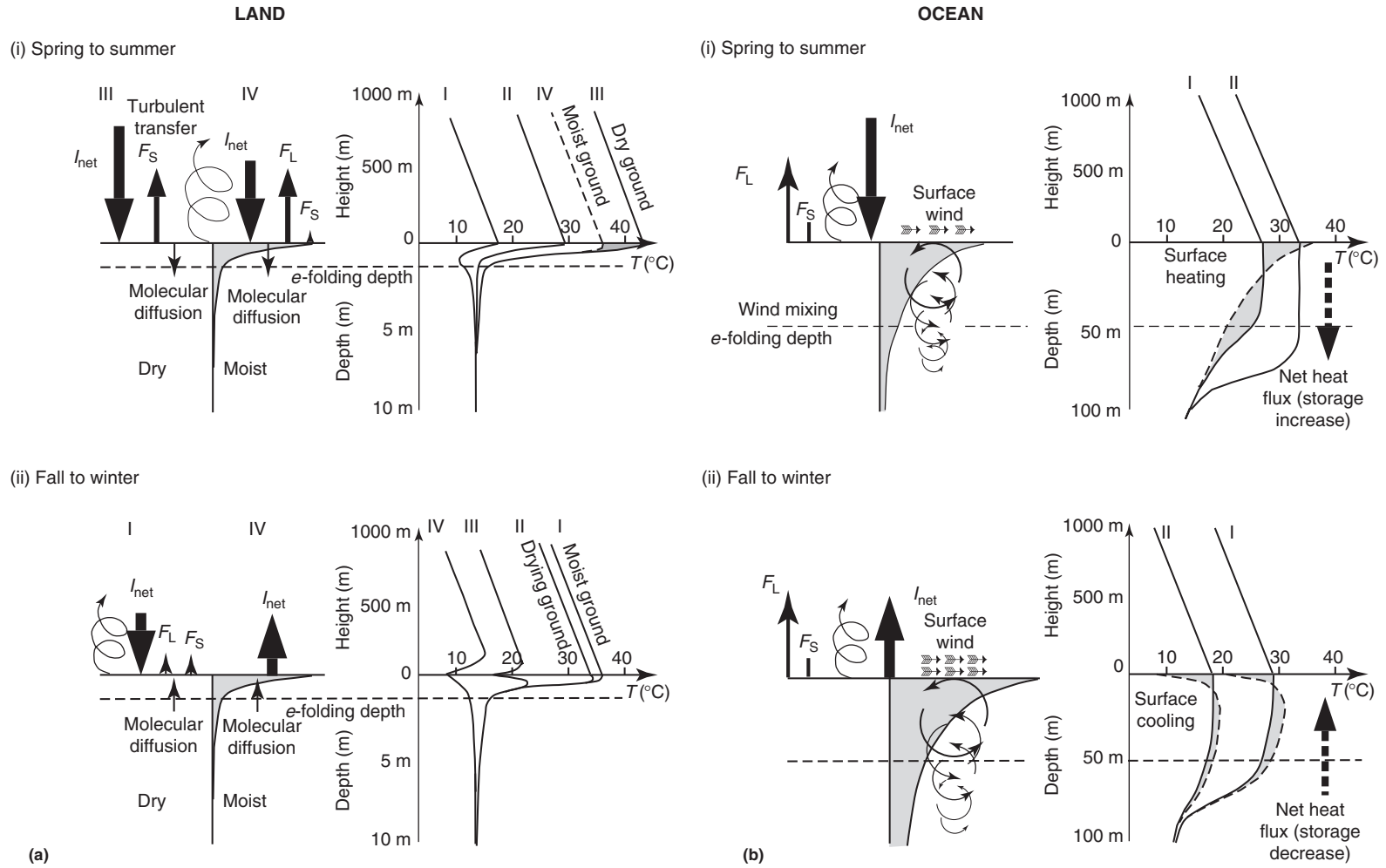


Figure 1 Schematic diagram of the dominant physical processes determining the surface temperature over (a) land, and (b) the ocean for the transitions from (i) spring to summer and (ii) fall to winter. The radiative and turbulent fluxes are labeled relative to eqns [1], [2] and [3]. Relative e -folding depths of diffusion, solar radiation penetration, and turbulent mixing are shown as horizontal dashed lines in each section. The net incoming radiation is attenuated in the first few millimeters of soil, from where it is transferred by molecular diffusion for a few meters. Wind stress at the ocean surface causes substantial transfers of heat (upward in winter, downward in summer) between the surface layers and the subsurface ocean by inducing turbulent mixing. Successive lower tropospheric temperature profiles are also shown (marked I, II, and so on). The shading in (a)(i) shows the change in temperature if the soil is moist. The shading in (b) shows the changes in the ocean temperature profiles when heated surface water moves downward in the summer or when cooled surface water is replaced by warmer subsurface water.

temperature of the atmospheric column. Equation [4] states that the change of pressure with height is inversely proportional to the mean temperature of the column. Therefore, over the warm summer continent, the pressure will decrease with height at a lesser rate than over the cold ocean, as shown in Figure 2. The relationship can be explored by integration in the vertical through the thickness of a slab of atmosphere between heights $z = 0$ and $z = z_1$. The difference in pressure $\Delta \ln p(z)$ at height $z = z_1$ between the warm and cold columns of Figure 2 can be expressed as eqn [5], where $\bar{T}_c$ and $\bar{T}_w$ are the mean temperatures of the atmospheric columns over the heated land and the cooler ocean.

$$\Delta \ln p(z_1) = \frac{g}{R} z_1 \left(\frac{1}{\bar{T}_c} - \frac{1}{\bar{T}_w} \right) + \Delta \ln p(0) \quad [5]$$

Here, Δ refers to the difference of a quantity between the warm and cold columns along a constant height surface. From eqn [5], the condition for $\Delta \ln p(z_1) > 0$, assuming that the surface pressure difference between the warm and cold columns is zero (i.e., $\Delta \ln p(0) = 0$), is that $\bar{T}_w > \bar{T}_c$. In this case air above the surface will be forced to flow from the summer to the winter hemisphere, with mass continuity providing a lower tropospheric return flow from the winter to the summer hemisphere.

However, in general, the surface pressure over the winter subtropics is higher than the surface pressure over the heated continent, perhaps by as much as 20 hPa. In fact, as the solar heating increases over the continent, the surface pressure is observed to fall so that $\Delta \ln p(0)$ becomes increasingly negative (Figure 3). Thus the criterion $\bar{T}_w > \bar{T}_c$ is not sufficient to ensure that there will be a reversed upper tropospheric pressure gradient and a return flow to the winter hemisphere. From eqn [5], a general condition for the temperature difference for $\Delta \ln p(z_1) > 0$ in the presence of a surface pressure gradient can be found (eqn [6]).

$$\bar{T}_w > \frac{gz_1 \bar{T}_c}{gz_1 + R \Delta \ln p(0) \bar{T}_c} \quad [6]$$

These simple arguments suggest that there may be a threshold in temperature difference between the winter hemisphere and the summer hemisphere in the presence of a surface pressure gradient, before a reverse pressure gradient in the upper troposphere is established. At that stage, pressure gradients throughout the troposphere are conducive for the maintenance of a direct thermal circulation. This threshold in mean tropospheric temperature gradient may be the reason for the observed sudden onset of the monsoon over South Asia in late May or early June, at which time deep convection and heavy precipitation occur.

The post-monsoon onset circulation, in the absence of moist processes, is shown in Figure 2(a). If it is assumed that the monsoon is in steady state, the amount of heat gained by surface heating must be balanced by heat lost to space by radiative processes. For a given stratification, the vertical extent of the dry monsoon is determined by the input of heat at the lower boundary. The longitudinal extent of the circulation is determined by the time it takes for a parcel to radiate away excess heat gained at the continental surface. If radiative processes were very efficient, the longitudinal scale of the

monsoon would be very small. However, radiative processes are slow, with e -folding dissipative time scales of about 20 days. Thus, the parcel takes a considerable time to cool and a parcel in the upper troposphere travels a considerable distance while cooling.

Moist Processes and the Monsoon Solar Collector

So far, moist processes have been ignored except for their implicit inclusion in surface evaporation. Moist processes change the character of the monsoon by moistening the land surface and being the agent of strong mid-tropospheric heating through the release of latent heat over the summer continent or adjacent marginal seas.

The source for summer monsoon rainfall is water evaporated from the ocean as air flows toward the heated continent under the action of the pressure-gradient forces discussed above. Figure 4 shows the source regions of moisture for the monsoons. The figure plots the vertically averaged moisture transport, B_q , defined as in eqn [7], where $q(z)$ and $\tilde{V}(z)$ are the specific humidity and the horizontal velocity vector, respectively.

$$B_q = \int_0^\infty q \tilde{V} dz \quad [7]$$

As moisture tends to decrease exponentially above the surface, the greatest contributions to B_q come from the surface boundary layer. Moisture accumulation zones can be seen to the south of Asia extending well into the winter hemisphere during the boreal summer (Figure 4(a)) and, to a lesser extent, to the north of Australia during the boreal winter (Figure 4(b)). Even though evaporation cools the surface of the ocean, the boundary layer air flows across a gradient of increasing SST due to the net positive radiation budget at low latitudes. Consequently, the boundary layer air becomes warmer along its trajectory and, as the surface saturated vapor pressure increases, the moisture content of the boundary is elevated.

One might imagine that the dry monsoon model shown in Figure 2(a) is applicable to the monsoons during spring. Between the spring equinox and the summer solstice, the temperature of the land increases, producing a low-level pressure gradient, causing a steady advection of moist air toward the continent. Eventually, sufficient water vapor will be imported over the land so that rising motion will result in the release of latent heat and an increase in temperature of the continental atmospheric column, eventually producing a reversed pressure gradient at higher levels. The strengthening of the monsoon occurs with the rapid development of the upper-tropospheric meridional temperature gradient. The increase in columnar temperature necessary to produce the reversal (see the previous Section) is directly attributable to the release of latent heat. At this point, the acceleration of the monsoon is substantial. Surface winds that were relatively weak prior to the onset of the monsoon exceed 10 m s^{-1} at the surface when the monsoon is established. Evaporation and increased ocean mixing accompany the strengthening winds and the SST of the North Indian Ocean drops rapidly by $1\text{--}2^\circ\text{C}$.

At the time of the year when large-scale precipitation occurs, two very important transitions occur in the monsoon. First,

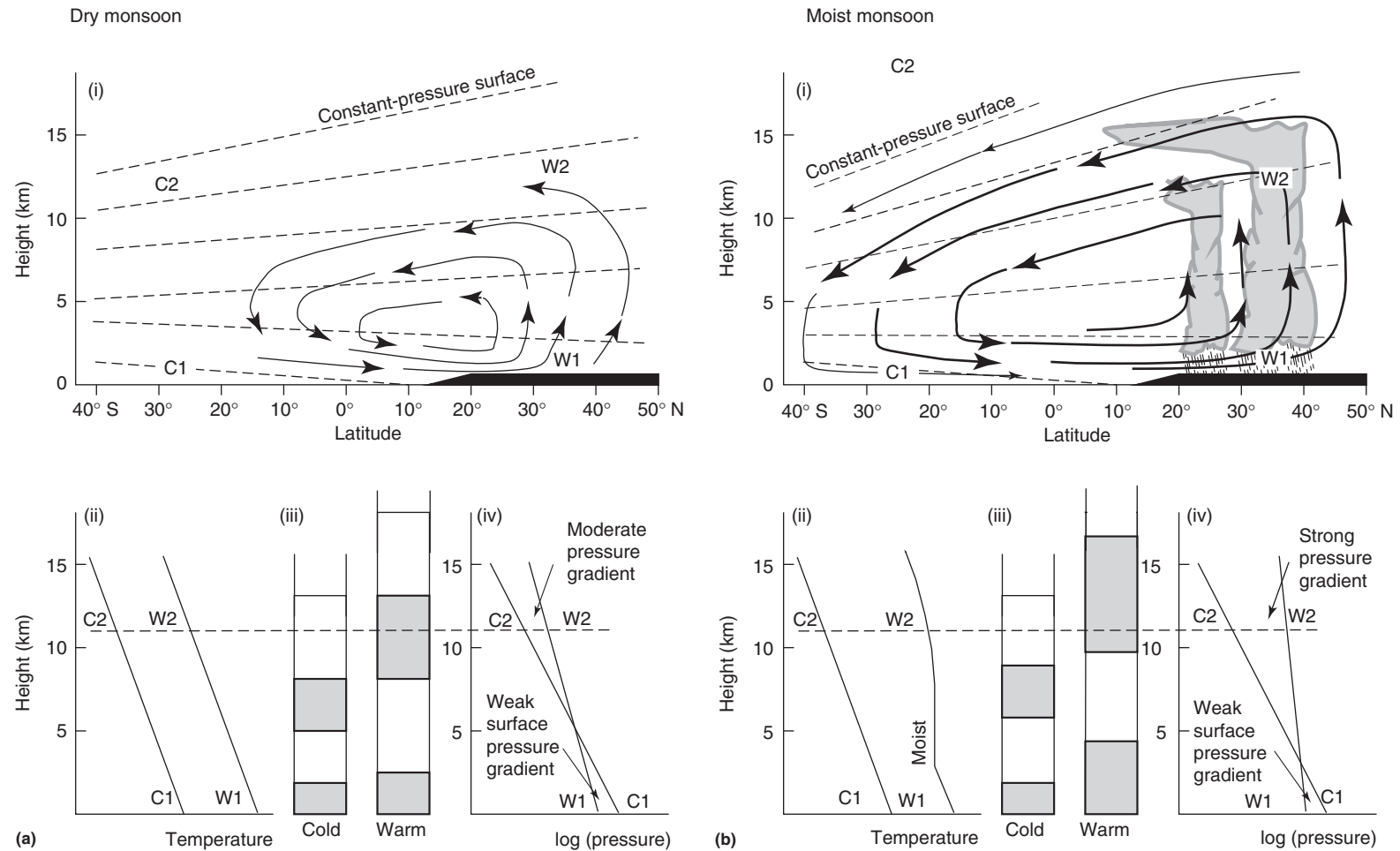


Figure 2 A mechanistic view of the development of the meridional monsoon circulation (a) when moist processes are ignored and (b) when moist processes are taken into account. The panels show (i) the resultant circulation, (ii) the temperature profiles, (iii) the distribution of mass in the vertical columns, and (iv) the change of pressure with height. Dashed lines in panel (i) show constant-pressure surfaces. Dashed lines in panels (ii)–(iv) denote a constant height. In both examples it is assumed that the difference in temperature of the warm and cold columns is sufficient to generate a reversing pressure gradient with height in the presence of the surface pressure gradient as described in eqn [6]. The figure is discussed extensively in the text.

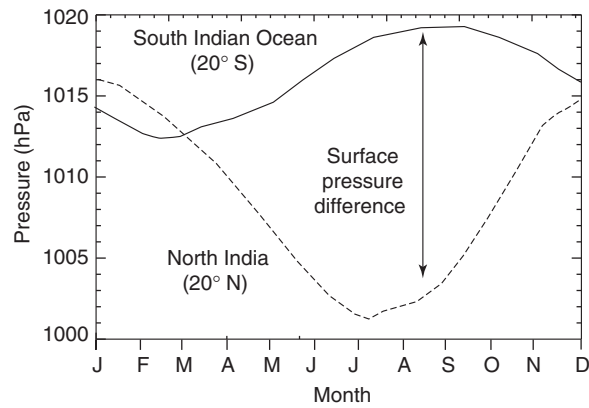


Figure 3 The mean annual cycle of sea-level pressure at 20° N and 20° S along 80° E, representing the monsoon trough and South Indian Ocean, respectively.

the dry land area becomes moist, sometimes so wet that it adopts many of the characteristics of a warm shallow ocean or lake (**Figure 1**). Accompanying the surface moistening is a substantial decrease of the surface land temperature. For example, the daily maximum surface temperature of New Delhi, India often exceeds 45 °C prior to the onset of the monsoon, while during active rainy periods following the monsoon onset the surface temperature maxima are between 30 °C and 32 °C. At the same time, the intensity of the monsoon increases substantially, with both the surface and upper tropospheric winds strengthening considerably.

It would seem that the simultaneous decrease in surface temperature of the land and the cooling of the adjacent ocean with the strengthening of the monsoon would act as a negative feedback and that the monsoon strength should decrease. However, the stronger winds also cause a very large increase in the amount of water vapor imported over the continental regions. The enhanced convergence of water vapor causes the release of latent heat to increase substantially. Thus, once the monsoon strengthens, the importance of the surface temperature gradient decreases and the overall driving of the monsoon

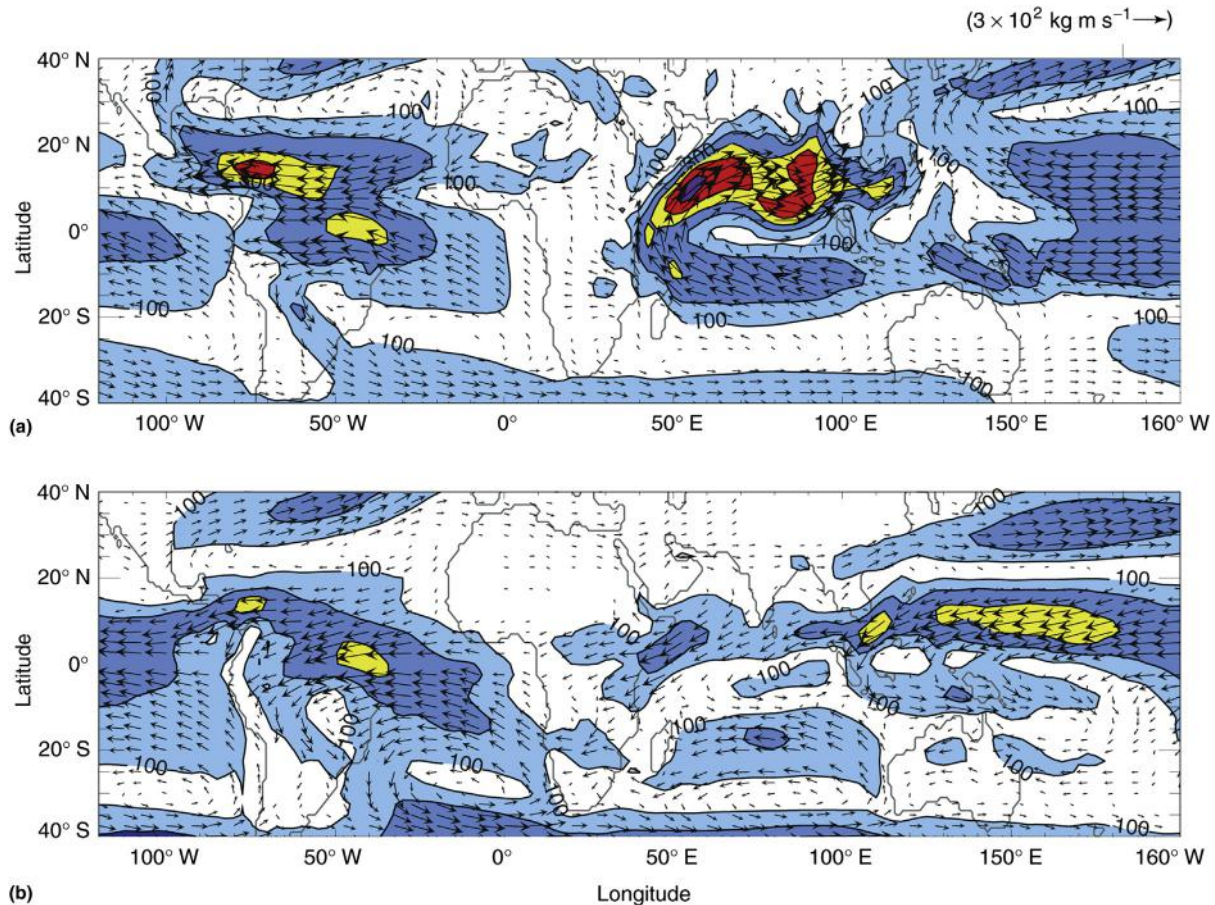


Figure 4 Distribution of mean vertically integrated moisture transport from eqn [7] for the period (a) June–September and (b) December–February. Viewed in the context of moisture transport, the Asian–Australian monsoon system appears in both (a) the boreal summer and (b) the boreal winter as strong interhemispheric systems with moisture sources clearly defined in the winter hemisphere. Both the African summer and winter monsoons are less clearly defined. Weak moisture fluxes into north-west Africa are evident, for example, but the region is dominated by strong westward moisture fluxes associated with the Trade Wind across the Atlantic. Furthermore, the moisture fluxes associated with the North and South American monsoons appear restricted to their respective hemispheres. Only the Asian–Australian monsoon possesses a truly interhemispheric solar collector.

is taken over by the heating of the troposphere over the continents through the release of latent heat. Also, prior to precipitation, the depth of the dry circulation is relatively shallow (Figure 2(a)). However, in the moist monsoon, the circulation occupies the entire troposphere as a result of the buoyant moist parcels releasing their latent energy as they ascend to great heights (Figure 2(b)). Furthermore, the vertical lapse rate decreases, moving closer to moist adiabatic. Thus, even though the surface temperature over the land decreases, the average tropospheric temperature increases.

The monsoon may be thought of as a great solar collector. The latent heat of condensation that is released over the heated continents, and which drives the established monsoon, is the summation of the evaporation occurring at the surface of the ocean between the winter hemisphere and the heated continents (Figure 4). As the evaporation results from a net positive radiation excess at the ocean surface or from the winds driven by the SST gradients, evaporation may be thought of as the integration of the solar heating across the ocean. Thus, the latent heating of the atmosphere over the continents is the accumulation of a large fraction of the solar energy incident over the vast ocean fetch of the monsoon winds. However, the effect of this solar heating is concentrated into a relatively small region of summer hemisphere continents and the adjacent seas where precipitation occurs.

The vertical and longitudinal scales of the moist monsoon are much larger than those of the dry monsoon as may be seen by comparing Figure 2(a) and 2(b). With moist processes, there is a very large amount of heat added to the system in the vigorous moist ascent over the heated continents. This is heat accumulated during the long trajectories over the oceans: the solar collector effect. The input of energy is far larger than in the dry monsoon and, as a result, the amount of time required for the upper tropospheric flow to radiate excess heat away is much longer. Thus, the upper tropospheric pressure gradient is maintained over a much larger scale and the moist monsoon acquires a correspondingly large longitudinal scale.

The Impact of Rotation

Halley was the first to understand the fundamental role of land-sea differences in producing onshore monsoon winds. Some 50 years, later in 1735, Hadley published a paper that considered the effects of the rotation of the planet and produced a general theory for the Trade Wind regime and the monsoons.

The monsoon model we have developed so far defines a circulation system in which air parcels move about under the action of body forces whose direction and magnitude are determined by the distribution of heating. In essence, we have added to Halley's monsoon model the physics of moist processes and a more realistic view of the ocean heat capacity. However, the planet is rotating and the effects of rotation are extremely important in the equatorial regions as the Coriolis force ($f = 2\Omega \sin \phi$) changes sign and the gradient of the Coriolis force ($\beta = 2\Omega (\cos \phi)/a$) is a maximum. As monsoon motion is cross-equatorial, the dynamics of low-latitude phenomena become very important. Here we restrict ourselves to the discussion of two immediate effects of rotation on the monsoon system: the general form of the flow under

the action of a constant cross-equatorial pressure gradient on a rotating planet, and the forcing of boundary layer mass and heat transports in the upper ocean.

Impacts on the atmosphere

The horizontal equation of motion on a rotating platform may be written as eqn [8], where α (s^{-1}) is a dissipation coefficient. In the boundary layer, α^{-1} is large and has an e -folding scale of about 2 days.

$$\frac{d\mathbf{V}}{dt} = -\frac{1}{\rho}\nabla p - f\mathbf{k} \times \mathbf{V} - \alpha\mathbf{V} \quad [8]$$

Consider now a constant pressure gradient in the lower troposphere extending from high pressure in the wintertime subtropics to the low pressure over the heated continent in the summer hemisphere. Without rotation, steady flow will occur when the frictional force (proportional to wind speed) balances the pressure gradient force. From eqn [8] this balance may be written as eqn [9], where v is the meridional component of the velocity vector.

$$\alpha v = \frac{1}{\rho} \frac{\partial p}{\partial y} \quad [9]$$

The resultant flow, which is purely meridional, is depicted schematically in Figure 5(a)(i). With rotation, steady flow comes about from a balance between the Coriolis force ($f\mathbf{k} \times \mathbf{V}$ in eqn [8]), and the pressure gradient force ($\partial p / \rho \partial y$). From eqn [8], the balance may be expressed as in eqn [10]

$$f\mathbf{k} \times \mathbf{V} + \alpha\mathbf{V} = -\frac{1}{\rho} \frac{\partial p}{\partial y} \quad [10]$$

Between the high-pressure region and the Equator the flow must cross the isobars, as shown in (Figure 5(a)(ii)) because of the frictional force eventually adopting the form given by eqn [9] at the Equator. At all latitudes in this simple example, the flow is across the pressure gradient, flowing from low to high pressure. The resultant surface flow diverges out of the winter hemisphere surface anticyclone, flows across the Equator, and finally converges into the continental low-pressure region over the heated continent. In the winter, when the polarity of the pressure systems is reversed, the flow is a mirror image to the flow shown in Figure 5(a).

As discussed earlier, the flow in the upper troposphere is in the opposite meridional direction through the action of a reversed pressure gradient force that increases with height. Furthermore, compared to the surface boundary layer, the dissipation coefficient (α) is an order of magnitude smaller. The resulting flow is closer to geostrophic than the surface boundary layer flow but still cross-gradient as it spirals out of the upper anticyclone and moves westward. The trajectory that an upper-tropospheric air parcel takes is such that it will move a much greater distance in the longitudinal direction (compared to the surface flow) before it crosses the Equator and eventually descends into the winter hemisphere. The strong upper tropospheric easterlies are referred to as the Easterly Jet that extends from South Asia across East and Central Africa and attains speeds of $40\text{--}50 \text{ m s}^{-1}$. The extent of the Easterly Jet can be seen in (Figure 6(a)), which shows the mean northern hemisphere 200 hPa flow.

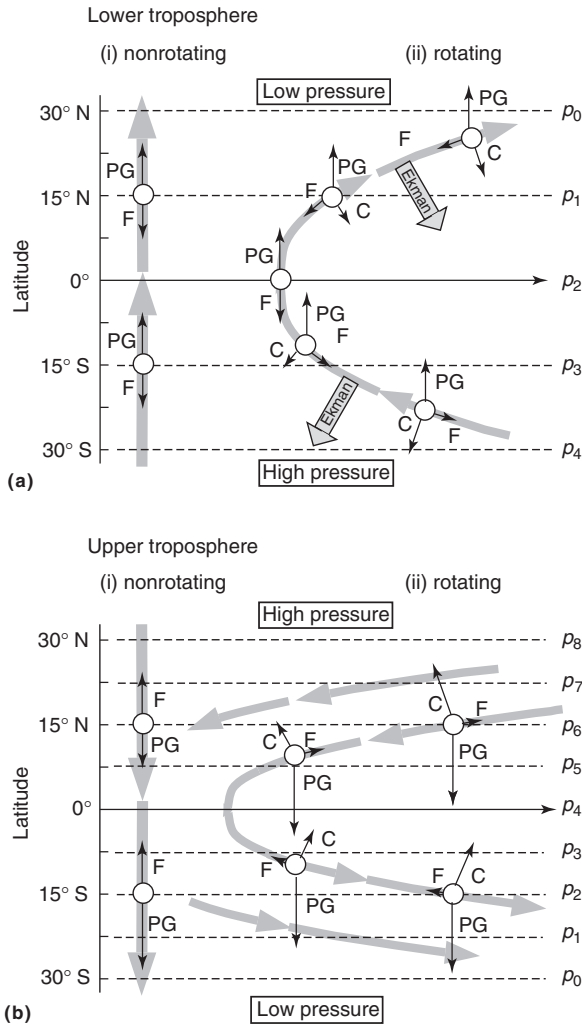


Figure 5 The circulation across the Equator in (a) the lower troposphere and (b) the upper troposphere on (i) a nonrotating planet, and (b) a rotating planet. Circulation is forced by a pressure-gradient force (PG) that is constant in latitude and subject to Coriolis effects (C) and frictional dissipation (F) that is assumed to be proportional to the parcel speed. The broad arrows denote the integrated Ekman mass ocean transport forced by the surface winds.

The vertical variation of the geostrophic flow is given by the thermal wind equation, which can be obtained by differentiating eqn [10] in the vertical, using the equation of state and neglecting frictional effects. This gives eqn [11], where $\bar{T}$ is the mean temperature of a layer.

$$\frac{\partial \mathbf{V}_g}{\partial z} = -\frac{g}{f\bar{T}} \nabla \bar{T} \times \hat{\mathbf{k}} \quad [11]$$

Thus if the mean temperature of a layer decreases toward the poles, then the vertical shear will be positive (e.g., lower tropospheric easterlies and upper tropospheric westerlies) in either hemisphere. However, in the monsoon regions, the shear is negative (lower level south-westerlies and upper level easterlies) so that temperature must increase towards the pole. Figure 7 plots the mean global upper tropospheric (500–200 hPa) temperature for the boreal summer and winter. During

the boreal summer, the upper troposphere of South Asia/Tibetan Plateau region is the warmest region on the planet. Importantly, the mean temperature is $>5^\circ\text{C}$ warmer than the Equator at the same longitude. During the boreal winter, a weak mean upper tropospheric temperature maximum exists in the vicinity of North Australia. The anomaly is about 1°C warmer than at the Equator. This reversed temperature gradient is sufficient to drive a weak upper tropospheric easterly jet stream over North Australia. Thermal ridges may be seen over other parts of the world (e.g., North America), but nowhere else does the temperature gradient reverse between the Equator and the pole except over South Asia and North Australia.

The Easterly Jet produced by the reversed temperature gradient has a pronounced effect on the North African monsoon. Figure 6(a) shows maximum jet speeds are attained over the central North Indian Ocean. Over central and western Africa, the flow decelerates in the exit region of the jet. The deceleration of the jet stream is important as it produces a secondary circulation over West Africa that is in the opposite sense to the local monsoon direct circulation. The circulation is produced in the following manner. Consider a parcel flowing through the Easterly Jet in geostrophic balance. From Figure 6(a) it is evident the parcel moves through the jet stream it will experience first an increase in easterly wind speed (i.e., $du_g/dt < 0$) in the entrance region, where u_g is the zonal component of the geostrophic wind. In the exit region of the jet, the easterly wind decreases in intensity (i.e., $du_g/dt > 0$). The zonal momentum equation can be written as eqn [12], where we have decomposed the meridional velocity component into a geostrophic and ageostrophic component (i.e., $v = v_g + v_a$).

$$\frac{du_g}{dt} = f v - \frac{1}{\rho} \frac{\partial p}{\partial x} = f(v_g + v_a) - \frac{1}{\rho} \frac{\partial p}{\partial x} = f v_a \quad [12]$$

Thus, in the entrance region of the jet there must be a southward displacement of the parcel. On leaving the jet, the parcel must move northward. This ageostrophic flow is shown in Figure 6(b). Over Africa, the result is the production of a secondary circulation made up of the ageostrophic northerly flow with descending air on the poleward side of the jet stream and ascending air on the equatorward side. This secondary circulation is in opposition to the monsoon meridional circulation that would have rising air over North Africa (Figure 6(c)). As a result, the secondary circulation restricts the northward extension of monsoon rainfall over the continent and may help explain why the West African monsoon is restricted to the coastal belt and why rainfall decreases so rapidly over the Sahel and the Sahara. On the north side of the entrance region of the Easterly Jet, the South Asian monsoon circulation is enhanced by the secondary circulation. This enhancement occurs in the Bay of Bengal region and perhaps explains the existence of the most intense rainfall in the Asian monsoon.

It is paradoxical that the limitations on the poleward extent of the African rainfall are the result of the South Asian monsoon rainfall falling so far to the north. Reasons for the anomalous latitudinal location of the South Asian monsoon rainfall will be discussed in the next section. These are the same reasons that determine the longitudinal extent of the jet stream and hence the limit of its effect on African rainfall.

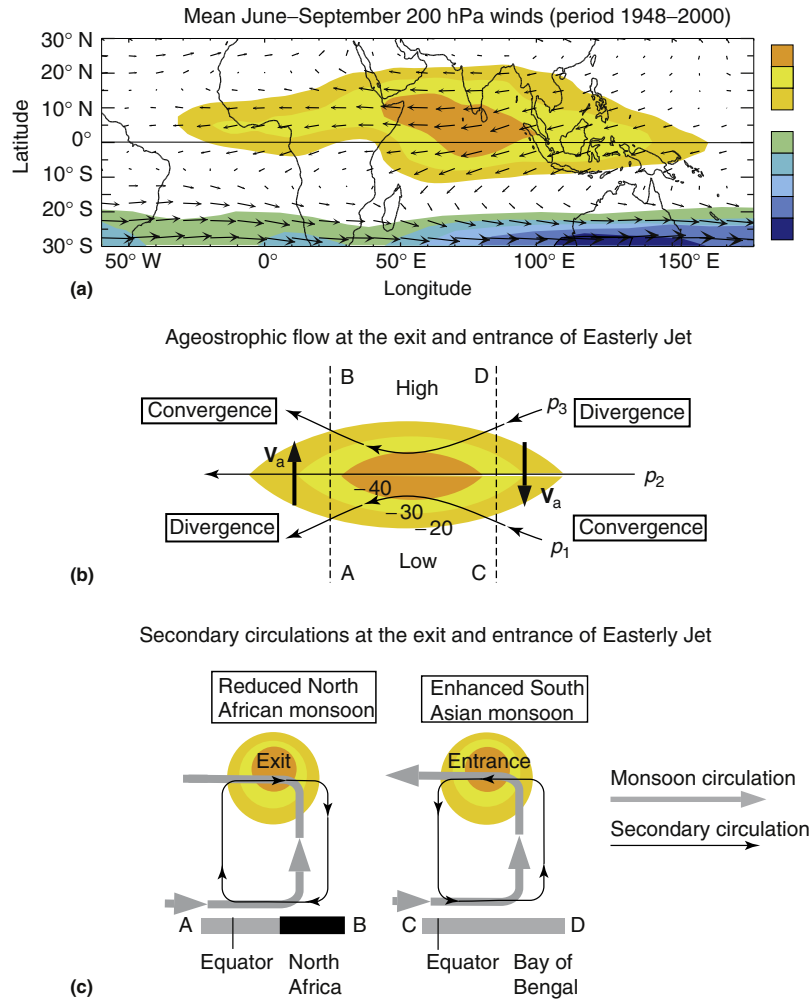


Figure 6 Impacts of the large monsoon heating over South Asia. (a) The upper tropospheric circulation. Note the extent of the Easterly Jet that commences over the north-eastern Indian Ocean and extends out across the Atlantic Ocean. (b) The secondary circulation associated with the entrance and exit regions of an easterly jet stream. Note that the ageostrophic flow produces a northward deflection of the winds that is consistent with subsidence on the northward side of the jet and rising surface pressure. (c) Schematic diagram showing the secondary circulation in opposition to the monsoon meridional cell.

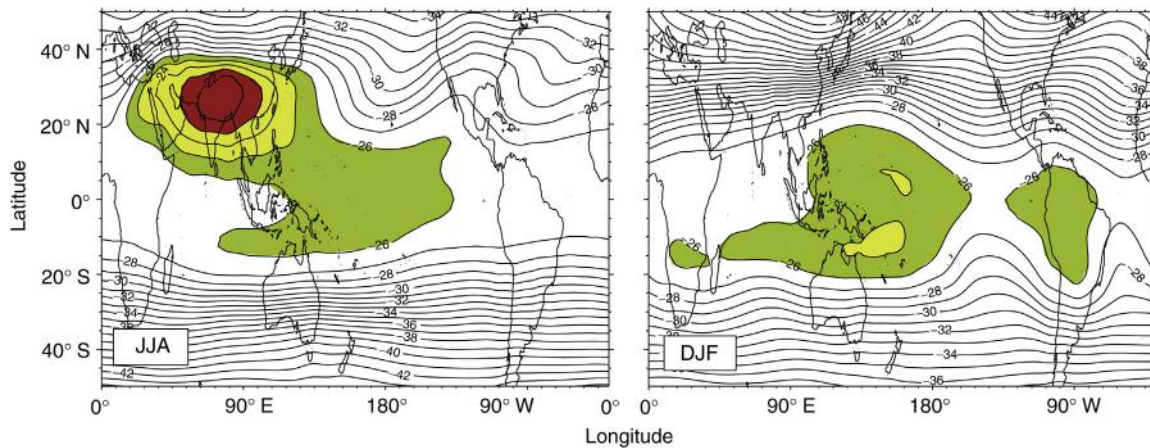


Figure 7 Distribution of the mean upper tropospheric temperature averaged between 200 and 500 hPa for the boreal summer (JJA) and winter (DJF). Note the two locations where the mean temperature is warmer than the equatorial temperature: over the Tibetan Plateau during the boreal summer and to the north-east of Australia in the austral summer.

Impacts on the ocean

Rotational effects are very important in the ocean and play a key role in determining both the overall character of the annual cycle of the monsoon and its interannual variability. The heat balance of the ocean mixed layer may be written as eqn [13].

$$C_w \frac{dhT}{dt} = F_{z=0} + F_{adv} \quad [13]$$

Here h and T are the thickness and temperature of the mixed layer, and $F_{z=0}$ is defined in eqn [2]. F_{adv} refers to the advection of heat in and out of the column. Equation [9] allows for the horizontal advection of heat by ocean currents. The depth of the mixed layer is determined, to a large degree, by wind-forced turbulent mixing as discussed earlier.

Observations in the North Indian Ocean suggest that the advective term may be important. In the period between March and May, the average net surface flux into the North Indian Ocean is $>100 \text{ W m}^{-2}$. However, the SST increases by only about 2°C during this period compared to about 9°C that would be expected if $F_{adv} = 0$. It is important to understand the reason for this SST modulation, as the monsoon would be very different if the temperature of the North Indian Ocean were much warmer. Modulation might occur by either heat storage increases (by deepening of the mixed layer) or by the horizontal advection of heat away from the summer hemisphere. As it turns out, both processes are important.

The instantaneous northward flux of heat across the Indian Ocean between two meridians x_1 and x_2 and the change in heat storage in the ocean column are given by eqns [14] and [15].

$$F_{adv} = \rho_w C_w \int_{x_2}^{x_1} \int_z^0 (HvT) dx dy dz \quad [14]$$

$$S_Q = \rho_w C_w \int_{x_1}^{x_2} \int_{y_1}^{y_2} \int_z^0 (HT) dx dy dz \quad [15]$$

Here ρ_w and C_w represent the density and specific heat of sea water, and the coordinates x , y , and z represent longitude, latitude, and depth, respectively. Figure 8 shows the total cross-equatorial flux of heat flux by oceanic processes from eqn [14], the change in oceanic storage of heat of the entire Indian Ocean north of the Equator from eqn [15], and the net surface heat flux into the North Indian Ocean from [2]. In addition, the atmospheric flux of latent heat across the Equator is also plotted. Compared to the magnitude of the net surface heat flux into the ocean (about 0.5 PW), the changes in storage and the heat flux across the Equator are much larger, with amplitude swings of $\pm 2 \text{ PW}$ through the annual cycle. Clearly, the surface heat flux is not driving either meridional heat transport or changes in heat storage. A large part of the changes in storage occur through cross-equatorial transports of heat. Figure 8 shows that transport has an extremely large annual cycle, with a maximum northward transport occurring in winter and early spring and a reverse transport occurring in the summer. Finally, it should be noted that the atmospheric latent heat flux is almost exactly out of phase with, and roughly equal in magnitude to, the lateral oceanic heat transport across the Equator. Clearly, the ocean and the atmosphere are working in

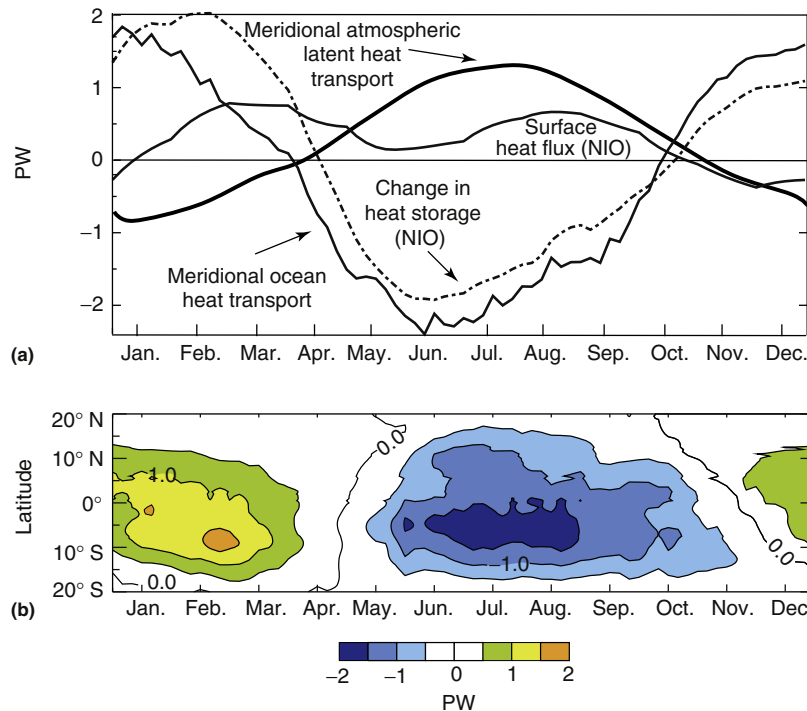


Figure 8 (a) Annual cycle of the meridional ocean heat transport across the equator, changes in the surface heat flux integrated over the North Indian Ocean, the change in total heat storage in the North Indian Ocean (NIO), and the atmospheric meridional transport of latent heat across the Equator. All quantities are averaged across the Indian Ocean. (b) Time-latitude plot of the cross-equatorial ocean flux of heat. All units are in PW (10^{15} W).

tandem in some manner to determine the overall heat budget of the monsoon.

If the atmospheric and oceanic components of the monsoon are coupled, what are the processes responsible for the coupling? The fundamental clue to understanding the coupling comes from noting that the majority of mass and heat transport in the ocean is wind-driven. Transport is a combination of geostrophic and Ekman transport such that $F_{adv} = F_{geos} + F_{Ek}$. The meridional Ekman transport is especially interesting and may be written as in eqn [16], where τ_x is the zonal wind stress and T' is the degree to which the temperature of the upper ocean is warmer than the lower ocean.

$$F_{Ek} = C_w \int \frac{-\tau_x}{f} T' dx \quad [16]$$

In the Northern Hemisphere ($f > 0$), the transport will be southward if the winds are westerly ($\tau_x > 0$) and northward for easterly winds ($\tau_x < 0$). As $f < 0$ in the Southern Hemisphere, the reverse associations are true. More generally, vertically averaged oceanic Ekman transports are to the right of the surface wind in the Northern Hemisphere and to the left of the wind in the Southern. Thus, in a monsoon system, the heat transport will be southward in both hemispheres in the boreal summer and northward in the winter hemisphere, as shown in Figure 5(a). This means that the overall oceanic heat transport has the opposite sign to the atmospheric heat transport that is in the direction of the lower tropospheric divergent wind. The overall effect of the wind-driven oceanic heat transports is to cool the SST in the summer hemisphere and warm the SST in the winter hemisphere, thus reducing the cross-equatorial SST gradient.

Finally, it should be noted that Ekman transports are not really applicable very close to the Equator as the 'spin-up' time for an Ekman transport to begin after winds goes as the inverse of the Coriolis parameter. At the Equator the spin-up time would be infinite. However, the approximation is valid a few

degrees from the Equator and ocean models show a smooth transition of heat transport across the Equator that matches the Ekman transport to the north and south.

Regional Effects

In reality, the geography of the monsoon regions is far more complicated than the idealized model discussed so far. Each monsoon region possesses different land-sea distributions and topography. It might therefore be expected that the monsoon flows will have strong regional character. Nowhere is topography more important than around the Indian Ocean basin where the East African Highlands and the Tibetan Plateau complex extend along the eastern boundary of the basin and over South Asia, respectively.

The meridionally oriented East African Highlands act as a mechanical barrier to the South Indian Ocean south-east trades and concentrate the low-level flow intercepting the barrier crossing the Equator as an intense low-level flow called the Somalia Jet. Wind speeds in the core of the jet exceed $20\text{--}25 \text{ m s}^{-1}$. The impact of the East African Highlands is shown schematically in Figure 9. The jet stream drives a strong cross-equatorial Somalia Current and farther up the coast produces intense coastal upwelling that is associated with the $3\text{--}4^\circ\text{C}$ drop in SST in the eastern Arabian Sea compared to $1\text{--}2^\circ\text{C}$ over the entire North Indian Ocean.

During the spring and summer, the role of the Tibetan Plateau complex thought to be thermal, acting as an elevated heat source rather than as a barrier. To understand the concept of an elevated heat source, one needs to consider the heat balance on the surface of the plateau and the heat balance of the free atmosphere at the same elevation but away from the plateau. During spring and summer, the interception of radiation on the plateau causes a transfer of sensible heat and latent heat to the atmosphere of about 100 W m^{-2} compared to a much smaller value in a layer of atmosphere over the plains to

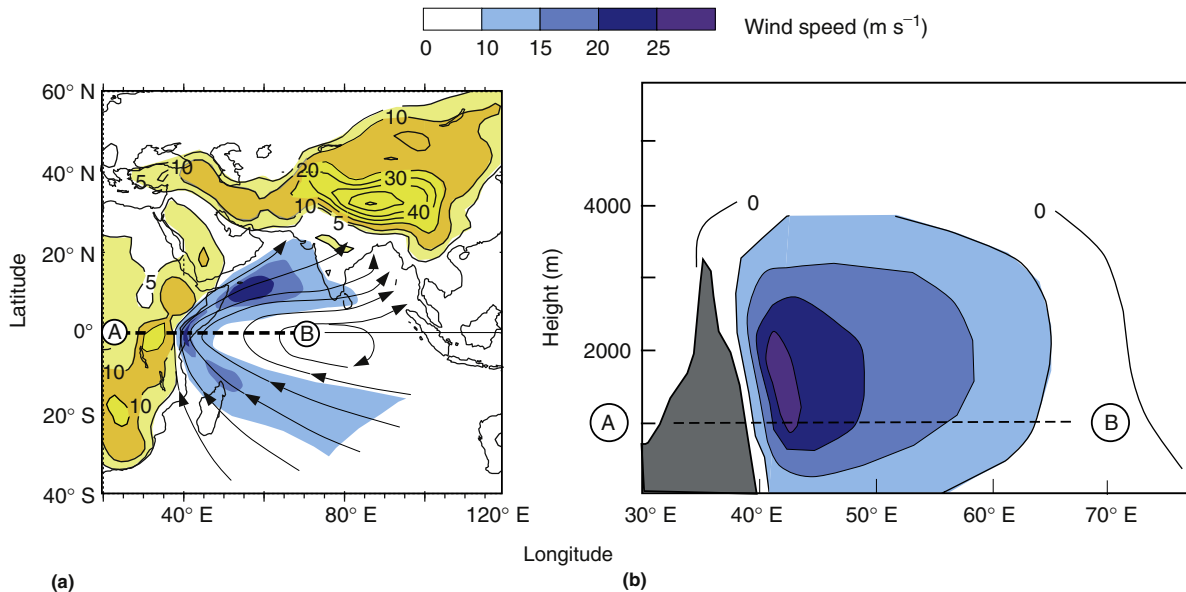


Figure 9 (a) Horizontal section of the Somalia Current at 1 km showing its lateral extent. Region north of the equator is a region of strong ocean upwelling forced by the wind. (b) Latitude–height section of the Somalia Jet along the equator. Units are m s^{-1} .

the south. The heating of the plateau produces a localized heating anomaly. **Figure 10** shows the heating that occurs throughout the troposphere at 10° N, 20° N, 30° N, and 40° N at 80° E after May 1. During the boreal winter, there is no evidence of heating over the plateau. Instead, the topography acts as a mechanical barrier to the strong boreal winter westerlies.

To a large degree, the Tibetan Plateau is major factor determining the character of the South Asian summer monsoon. The location of the elevated heat source to the north of 25° N is sufficiently strong to produce rainfall well into the subtropics, which, on a global perspective, are usually regions of low rainfall and deserts. The location of the upper tropospheric anticyclone is determined by Tibetan Plateau, which, in turn, sets the position of the upper tropospheric Easterly Jet and the region influenced by the secondary circulations in the exit region (**Figure 6**).

Regulation of the Monsoon System

One of the interesting features of the South Asian monsoon is its relative constancy from year to year. For example, the mean precipitation over the Indian subcontinent is 852 mm with a standard deviation of 84 mm. These statistics represent a high-rainfall region with very small variability about the mean. Furthermore, prolonged droughts or floods (i.e., mean summer rainfall outside the range of 852 ± 84 mm) are very rare. Between 1870 and 1998 there were 17 flood years and 22 drought years. Equally interesting is the lack of prolonged anomalies. Excessive rainfall occurred in successive years only twice (in 1892–94, and 1916–17), while deficient rainfall has occurred in successive years three times (1904–05, 1965–66, and 1985–87). The El Niño Southern Oscillation (ENSO) phenomenon explains about 40% of the variance, with El Niño

often being associated with drought and La Niña with flood. Also, it has been noted that there are periods where there is strong biennial trend in the rainfall record, with strong and weak monsoon years occurring successively. Overall, it would seem that the monsoon is regulated in some fashion to fall within rather narrow limits of variability. In the following paragraphs, it will be suggested that regulation of monsoon intensity occurs through the interaction of the ocean and the atmosphere in the manner introduced described under 'The Impact of Rotation'.

Regulation of the Monsoon Annual Cycle

Figure 11 shows a schematic of the atmospheric circulation and corresponding ocean heat transport for the boreal summer and winter. The impact of the wind-driven ocean heat transport is to homogenize the SST gradient across the Equator, lowering it in the summer hemisphere and warming it in the winter hemisphere. The reduction of the SST gradient will have two major impacts. First, the cross-equatorial pressure gradient will be reduced, which will have the effect of reducing the intensity of the monsoon winds. Second, the reduced SST in the summer hemisphere reduces the saturated vapor pressure of an air parcel approaching the continental region. As the SST is very warm, small changes in the SST will invoke large changes in the saturated vapor pressure, as determined by the nonlinear Clausius–Clapeyron equation. The reduced monsoon winds will lower the convergence of moisture into the continental regions. In tandem with the reduced saturated vapor pressure this reduces the release of latent heat in the precipitating regions. The overall impact of the two processes is to reduce the amplitude of the monsoon annual cycle. Without ocean heat transports, the SST extremes would produce a monsoon that would be very different from that which is observed today and would quite likely be much wetter.

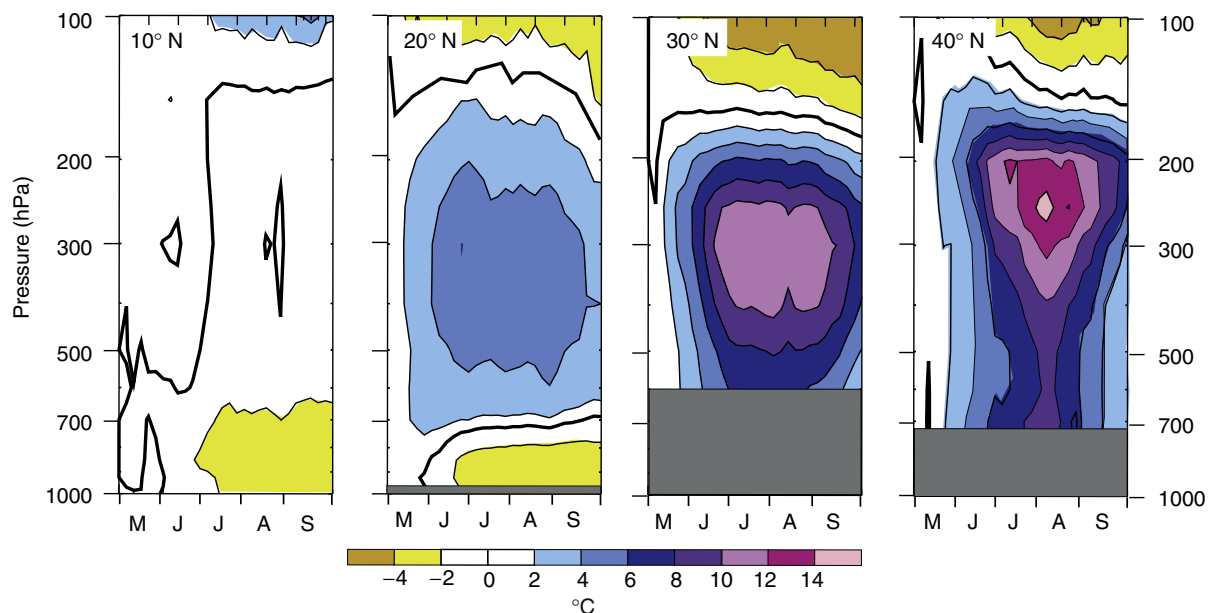


Figure 10 Time section of the change in atmospheric temperature over the Tibetan Plateau based on temperatures at 1 May.

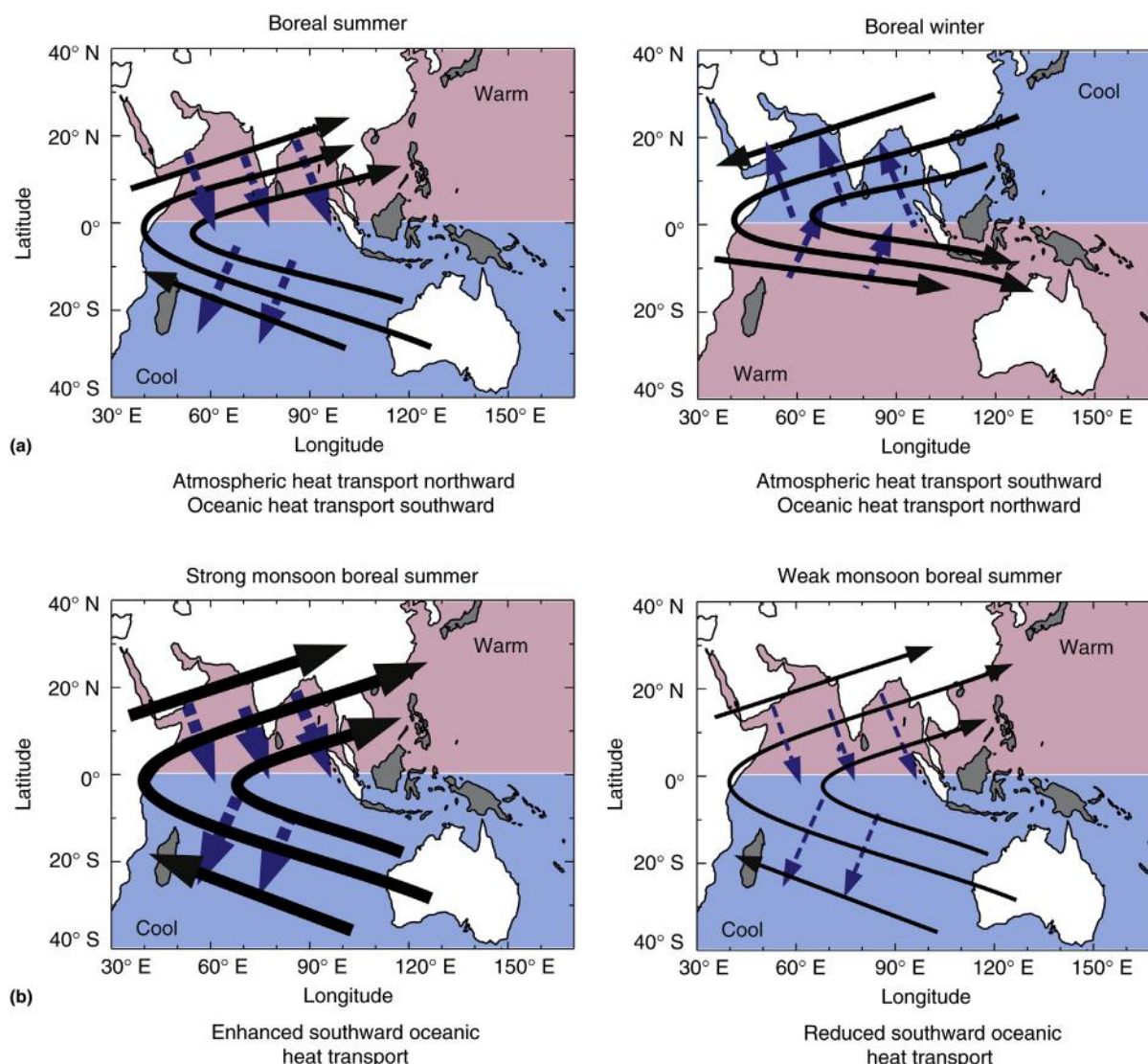


Figure 11 (a) A regulatory model of the monsoon for the annual cycle. Climatological lower tropospheric winds (solid black lines) are shown for summer (left panel) and winter (right panel). Blue dashed lines show the direction of Ekman mass transports which are to the right of the wind in the northern hemisphere and to the left in the southern hemisphere. Given the general sea-surface temperature gradient, heat is transported from the summer hemisphere to the winter hemisphere in both seasons. From Figure 8 it can be seen that these heat transports can be substantial enough to change the temperature of the entire North Indian ocean by 2–3°C. (b) Summer representations of the surface wind in strong and weak monsoon years. In a strong year, the Ekman transport will be large while in a weak year it will be smaller. Thus during a strong year the northern Indian Ocean will cool at a greater rate than during a weak year. Given that a strong monsoon is usually preceded by warmer than average sea-surface temperatures in the Indian ocean, a strong monsoon year is often followed by a weak monsoon and vice versa.

It is important to note that the annual cycle of oceanic heat transports, as described above, is a phenomenon that is restricted to regions that have strong cross-equatorial pressure gradients and a surface flow from the winter to the summer hemisphere. For example, the Pacific Ocean is characterized by a Trade Wind regime that converges into the western Pacific from both hemispheres. The Ekman transports of mass and heat are away from the tropics in both hemispheres: to the right of the north-easterly Trades and the left of the south-easterly Trades. The ocean heat transport produced by the Pacific Trade Winds is probably important in regulating the surface temperature of the Pacific warm pool. However, the strong

seasonality and change in sign of the heat transport is restricted to the Indian Ocean monsoon regime.

Regulation of Monsoon Interannual Variability

Imagine that, for some reason, that the SST in the northern Indian Ocean during spring was higher than normal. Such alterations to the climatology could be imposed by external factors such as ENSO, anomalous winter or spring snowfall over Eurasia, or the result of stochastic processes. Warmer than average SSTs would lead to a stronger than average monsoon winds. In fact, there is observational evidence for the

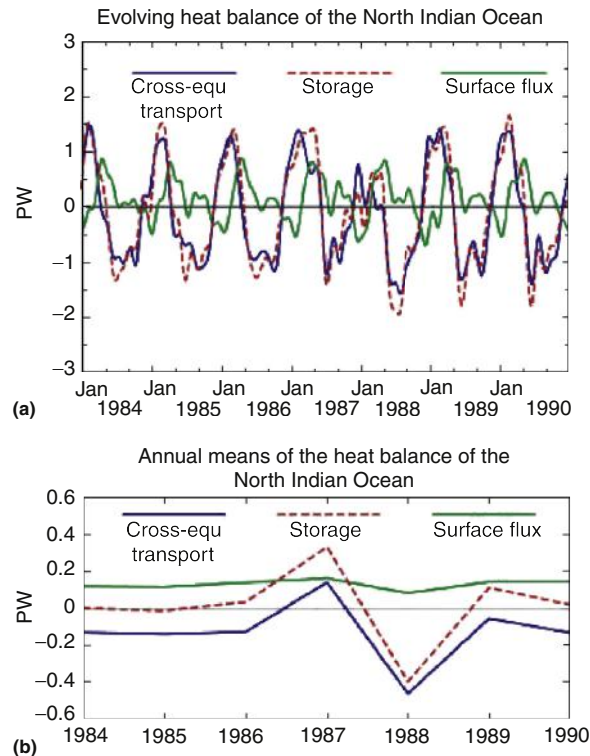


Figure 12 Time sections of the oceanic cross-equatorial heat flux, the net surface heat flux into the North Indian Ocean, and the change in ocean heat storage in the North Indian Ocean for the years 1983–89 as computed from an intermediate ocean model driven by observed winds and fluxes: (a) 5-day averages, (b) annual averages. 1987 and 1988 were weak and strong monsoon years associated, respectively, with El Niño and La Niña.

relationship. Indian Ocean SSTs during the previous winter and early spring correlate positively with the strength of the ensuing monsoon. Overall, the same processes that regulate the annual cycle of the monsoon ensure that the anomalous imposed conditions (e.g., high SST, weak winds) are eradicated.

For the case of higher than average springtime North Indian Ocean SSTs, the meridional pressure gradient would tend to be stronger and the near-surface saturated vapor pressure higher, causing a stronger than average monsoon to develop. A strong monsoon with greater than average winds would (1) cause deep mixing and increase the ocean storage of heat, and (2) drive stronger than average southward wind-driven oceanic heat transports. The impact of the strong monsoon would be to cool the North Indian Ocean so that the SSTs during the following winter would be lower than average. In turn, these lower SSTs would produce a weaker than average monsoon in the following spring and summer. Associated with the weaker than average monsoon winds would be a lowering of the southward heat transport.

Figure 12 shows an example of the monsoon interannual regulation process at work. In 1987 a strong El Niño occurred and was followed in 1988 by an equally strong La Niña. In 1987 the South Asian monsoon was weak, with Indian summer rainfalls 17% below normal. In 1988 the monsoon was strong, with rainfall over India 13% above normal. The figure shows

a time section of the annually averaged northward oceanic heat transport for these two years calculated using a stand-alone ocean model driven by observed winds. Figure 12 plots annually averaged values. The long-term average cross-equatorial heat transport is about -0.2 PW which matches the climatological net surface heating over the North Indian Ocean. However, during the weak monsoon year of 1987, the net surface flux was slightly higher and, associated with the weaker monsoon winds, the cross-equatorial ocean heat transport reversed from the climatological value of -0.2 PW to $+0.2$ PW. During the following year, the net surface flux into the North Indian Ocean was reduced while, responding to the much stronger monsoon winds, the oceanic heat transport was twice the long-term mean value of -0.2 PW at -0.45 PW. The two years 1987 and 1988 illustrate the manner in which the monsoon system responds to forcing anomalies and returns the system to equilibrium.

Summary

A rather different picture of the monsoon has been presented. Rather than describing the monsoon as a gigantic sea breeze driven by the differential heating between the ocean and land, a self-regulating system has been introduced in which dynamic aspects of the ocean play critical roles in modulating the strength of the monsoon. On this view, it is clear that if the variability of the monsoon is to be predicted, it will be necessary to consider the system as a fully coupled interactive system. As such, the numerical models that will be used to simulate and predict monsoon behavior will have to have interactive oceans.

See also: **Climate and Climate Change:** Climate Prediction: Empirical and Numerical. **Numerical Models:** Regional Prediction Models. **Oceanographic Topics:** General Processes. **Tropical Cyclones and Hurricanes:** Hurricanes: Observation; Overview and Theory. **Tropical Meteorology and Climate:** Equatorial Waves; Intertropical Convergence Zone; Monsoon: ENSO–Monsoon Interactions; Monsoon: Overview; Tropical Climates.

Further Reading

- Fein, J.S., Stephens, P.L. (Eds.), 1987. *Monsoons*. Wiley Interscience, New York.
- Hastenrath, S., 1991. *Climate Dynamics of the Tropics*. Kluwer Academic, Dordrecht.
- Tomczak, M., Stuart, J.S., 1994. *Regional Oceanography: An Introduction*. Pergamon, Oxford.
- Webster, P., 1994. The role of hydrological processes in ocean–atmosphere interaction. *Reviews of Geophysics* 32, 427–476.
- Webster, P.J., Palmer, T., Yanai, M., et al., 1998. Monsoons: Processes, predictability and the prospects for prediction. *Journal of Geophysical Research* 103 (C7), 14451–14510.
- Webster, P.J., Clark, C., Chirikova, G., et al., 2001. The monsoon as a self-regulating coupled ocean–atmosphere system. In: *Meteorology at the Millennium*. Academic Press, London, pp. 198–219.

Monsoon: ENSO–Monsoon Interactions

K-M Lau, NASA/Goddard Space Flight Center, Greenbelt, MD, USA

© 2003 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 3, pp 1386–1391, © 2003, Elsevier Ltd.

Introduction

The word ‘monsoon’ is derived from the Arabic word *mausam*, which means season. In monsoon regions, the change of season is accompanied by a reversal of the prevailing wind direction and abrupt changes in rainfall patterns. The most pronounced monsoon climate is found in the Asian–Australian region. Other regions that exhibit monsoonal characteristics include the maritime continent, western Africa, central America, south-western North America, and South America. These regions are often adversely affected by severe droughts and floods caused by the strong year-to-year-variability in monsoon rainfall. During the summer of 1997, southern China experienced severe flood while northern China was gripped by one of the driest seasons on record. In the summer of 1998, a monsoon depression in Bangladesh devastated the country, causing major floods in the Ganges and the Brahmaputra river, displacing over 30 million people, and resulting in property and agricultural loss of over \$US 3.4 billion. In the same summer, a flood of biblical proportion ravaged the Yangtze River basin and northeastern China, displacing over 220 million people, inflicting a huge economic loss of over \$US 12 billion. Understanding and predicting monsoon variability is therefore vitally important for benefit of society.

Scientists have long suspected that there is a connection between monsoon rainfall variability and components of the global circulation system. G. T. Walker in the 1920s found that the Indian monsoon rainfall anomalies may be foreshadowed by seasonal surface pressure variations over several ‘strategic’ points remote from India. Of the many planetary-scale patterns found by Walker, the most important is the Southern Oscillation (SO), which was described as ‘a swaying of pressure on a big scale backwards and forwards between the Pacific Ocean and the Indian Ocean’. However, his effort to translate the monsoon–SO relationship into seasonal and interannual prediction of Indian monsoon rainfall was largely

unsuccessful. Several decades later, Bjerknes first suggested that the SO as found by Walker is closely linked to El Niño, Bjerknes work and subsequent research have led to the development of various empirical monsoon forecast schemes based on the El Niño–Southern Oscillation (ENSO). However, results are mixed, because even though ENSO may be important in influencing monsoon rainfall variability, there are a large number of factors that may confound or limit monsoon predictability. With the advances in our understanding and prediction of ENSO, there is now renewed interest in reexamining the monsoon–ENSO relationship and its possible utilization for monsoon prediction. The discussion here will focus mainly on the most pronounced monsoon system, the Asian–Australian monsoon (AAM), unless specified otherwise.

The Basic Relationship

Figure 1 shows the time-series of all-Indian summer rainfall (June–September) from 1871 to 1998. Large interannual variability can be seen. There also appears to be interdecadal modulation of the amplitude of the anomalies. Relatively small anomalies are found in the 1880s–90s, the 1920s–40s, and in the 1990s. Large anomalies are found in the 1900s–20s, 1960s–80s. As indicated by the color shading, the Indian monsoon is generally below normal preceding the peak of a warm sea surface temperature (SST) event (El Niño) and above normal preceding a cold event (La Niña). However, there are warm or cold events that produce very weak signals or even anomalies of the opposite sign. Most important, a large number of major rainfall anomalies are not related to either El Niño or La Niña. Similarly relationship can be found for the austral monsoon (not shown). Hence, not all monsoon floods and droughts can be explained by ENSO.

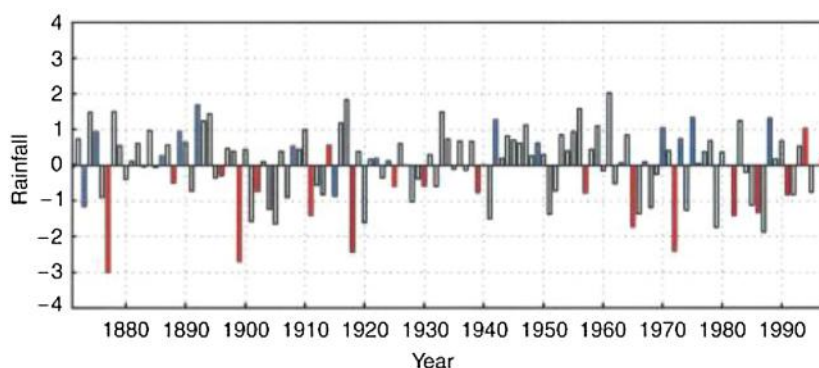


Figure 1 Time-series of All-India rainfall anomaly, normalized by standard deviation, with respect to the base period 1871–1998. Red bars indicate years with peak El Niño warming in following winter, and blue bars indicate years with peak La Niña cooling in following winter.

Table 1 Rainfall anomaly statistics for all-India summer rainfall and northern Australia summer rainfall, based on approximately 124 years of data

Rainfall	All India		Northern Australia	
	Total	El Niño (La Niña)	Total	El Niño (La Niña)
Below average	53	24 (2)	49	20 (4)
Above average	71	4 (19)	58	5 (17)
Drought	22	11 (2)	18	9 (0)
Flood	18	0 (7)	17	2 (5)

From Webster et al. (1998) Monsoon: processes, predictability, and the prospects for prediction. *Journal of Geophysical Research – Oceans* 103: 14451–14510. Reproduced by permission of American Geophysical Union.

Table 1 summarizes the monsoon rainfall statistics in relationship to the occurrence of El Niño and La Niña for a multidecadal period, for all of India and for northern Australia. During the 124-year period, out of the 53 events classified as below average, 24 events were associated with El Niño. Only 2 were associated with La Niña. For 71 above-average events, 4 belong to the El Niño group and 19 to the La Niña group. Altogether, El Niño explains fewer than 50% of the total below-average events and La Niña fewer than 30% of the above-normal events. For extreme events, the statistics shown in **Table 1** convey the same message. Out of the 22 extremely deficient-rainfall events, 11 were associated with El Niño and only 2 were found in La Niña. Conversely, for

extreme excessive-rainfall events, 7 out of 18 were found in La Niña, and none were associated with El Niño. For northern Australia, the statistics are very similar. Obviously, the relationship is far from perfect, but it is clear that in some fundamental way the two phenomena are related.

Spatial Patterns

To understand the physical underpinnings of the basic ENSO–monsoon relationship, we turn to the spatial distribution of rainfall and large-scale circulation anomalies associated with ENSO. **Figure 2** shows the seasonal climatologies (long-term mean) and composites of seasonal mean anomalies of rainfall and surface wind over the Asian–Australian monsoon region for December–January–February (DJF) and June–July–August (JJA) respectively. The anomalies are shown as the difference between composites of El Niño and La Niña. Climatologically, the boreal summer monsoon is associated with centers of heavy rainfall anchored to the west coast of India, the Bay of Bengal, Indo-China and the South China Sea/Western Pacific region (**Figure 2(a)**). The low-level circulation is dominated by a large-scale anticyclonic (clockwise in the Northern Hemisphere) gyre circulation over the Indian Ocean. This anticyclone gives rise to a strong low-level westerly flow across the Indian subcontinent and Indo-China, subsequently turning northward into East Asia and Japan. A dominant low-level anticyclonic circulation, known as the Subtropical High is located over the subtropical western

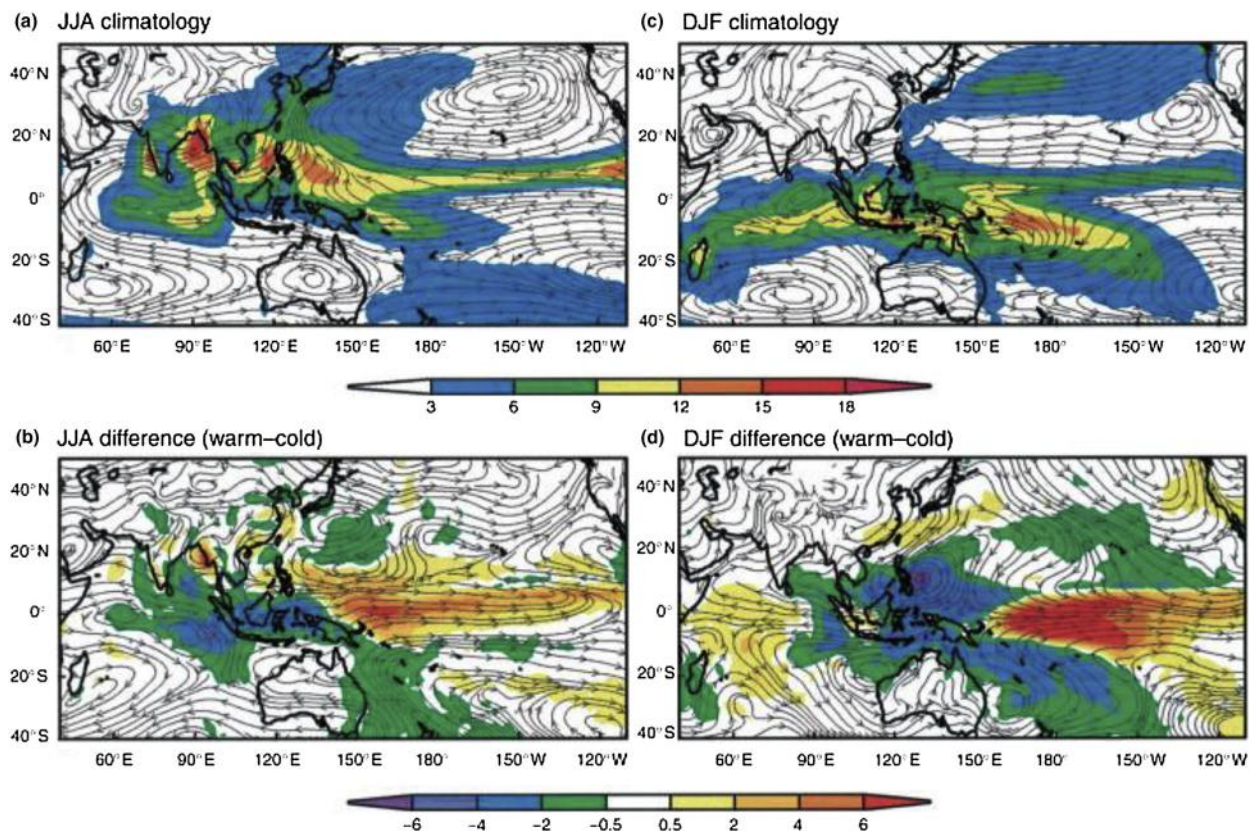


Figure 2 Climatology of rainfall and 850 mbar wind streamlines for (a) JJA, (c) DJF. Rainfall and 850 hPa streamline anomalies (warm events minus cold events) for (b) JJA and (d) DJF. Unit of rainfall in mm per day.

Pacific, with prevailing easterlies in its southern flank, between 10° S and 10° N. The westerlies and easterlies converge over the South China Sea/Philippines region, coinciding with the zone of heavy precipitation in this region. During El Niño (opposite for La Niña), enhanced rainfall and low-level westerlies are found over the equatorial central Pacific (Figure 2(b)). Reduced rainfall is found over the maritime continent and the subtropical western Pacific in both hemispheres. The rainfall pattern between 10° S–10° N over the Indian Ocean and the Western Pacific has been referred to as a ‘rainfall dipole’, which is one of the many key signatures of El Niño. Note that the responses over the monsoon land regions are not very well defined, compared to their oceanic counterparts. In particular, over the Indian subcontinent the signal is mixed. Over the Indian Ocean, the anomalous easterly flow between the equator and 15° N signals a weakening of the Indian Ocean gyre circulation. The rainfall dipole, which is made up of general suppression in rainfall (sinking motion) over the monsoon region and enhancement of rainfall (rising motion) in the equatorial central Pacific, is due to the eastward shift of the Walker circulation (*see Tropical Meteorology and Climate: Walker Circulation*) during El Niño.

During the austral summer, December–January–February (DJF), the zone of heavy rainfall shifts south of the Equator to the maritime continent, northern Australia and oceanic regions further east, near the dateline (Figure 2(c)). The large-scale circulation features low-level easterlies over the Northern Hemisphere western Pacific and much of the Indian Ocean just north of the Equator. Over the far western Pacific and the Indian Ocean, the flow turns southward and then eastward into the Southern Hemisphere, culminating in a belt of westerlies from the eastern Indian Ocean to 160° E. During an El Niño, rainfall anomaly occurs again in the form of an east–west dipole, with increased rainfall over the equatorial central Pacific and suppressed rainfall over the far western Pacific/maritime continent. In addition, a broad region of enhanced rainfall is found over the central and western Indian Ocean and an area of reduced rainfall over the eastern equatorial Indian Ocean. Together, these rainfall anomalies form a secondary dipole over the Indian Ocean. While the Indian Ocean dipole appears as part of the signal of ENSO, it has been suggested that the Indian Ocean dipole may also arise from regional coupled ocean–atmosphere processes that are independent of ENSO. Accompanying the anomalous rainfall are anomalous low-level westerlies over the central Pacific and easterlies over far western Pacific and Indian Ocean, consistent with the eastward shift of the Walker circulation during El Niño.

The relationship between rainfall over India and northern Australia and El Niño/La Niña noted earlier can be seen only as elements of a much larger shift in the tropical rainfall and circulation pattern. Note that composite pictures like those shown in Figure 2 only bring out common features for different El Niño/La Niña events, the actual rainfall anomalies may vary greatly regionally and for individual events.

Factors Affecting Monsoon–ENSO Relationships

Regional Ocean–Atmosphere Processes

Recent studies have shown that in addition to the basin-scale monsoon–ENSO mode associated with the shift of the

Walker circulation, there are intrinsic modes of variability in the monsoon–ocean–atmosphere system that are not directly connected with ENSO. These modes stem from regional coupled ocean–atmosphere processes in the Indian Ocean, the western Pacific, the South China Sea and Indonesian waters. The regional anomalies can amplify, oppose, and/or modulate ENSO-induced direct changes as the Asian–Australian monsoon responds and adjusts to the ENSO forcings. The regional ocean–atmosphere process regulates the SST and rainfall covariability in the monsoon region, through fluxes of heat and momentum at the ocean–atmosphere interface. This regional coupling covers a wide range of timescales ranging from the intraseasonal to interannual and beyond (see the discussion of intraseasonal variability later).

Snow Cover

A major factor that may confound the monsoon–ENSO relationship is the impact of snow cover on the evolution of the monsoon. In a report that dates back to the nineteenth century, Blanford in 1884 first showed that the strength of the Indian summer monsoon may be affected by snow cover over the Himalayas. There is now observational evidence suggesting that persistent winter snow cover over the Tibetan Plateau may delay or weaken the monsoon during the following summer. Increased snow cover increases the surface albedo, reflecting more solar radiation from the land surface. Heat energy that would otherwise be used to heat up the land will be used to melt the extra snow. As a result, the land surface heats up more slowly during the spring and summer, reducing the land–sea contrast and therefore weakening the monsoon. Conversely, reduced snow in the previous winter may lead to enhanced monsoon rainfall. However, while Eurasian snow cover changes may be related to internal atmospheric processes, increased snow cover over Eurasia may not be independent of ENSO on the interannual timescale.

Land–Atmosphere Hydrological Feedback

Land–atmosphere processes can affect monsoon and monsoon–ENSO relationship by altering the energy and water cycles within the monsoon regions, through surface heat fluxes and hydrological feedback mechanisms, as illustrated in the schematics in Figure 3. For example, if the soil moisture content of the Asiatic land mass is abnormally high during the start of a monsoon season, land surface evaporation will be increased. This will lead to increased moistening of the atmospheric boundary layer, more unstable air masses and hence more convection and rainfall, resulting in a positive feedback, leading to further moistening of the land region. However, the cloudy sky condition stemming from enhanced convection will shield and reduce the amount of solar radiation reaching the land surface, causing the land to cool. As the land mass cools, the resulting decreased land–sea thermal contrast can only support a weaker large-scale monsoon circulation, with reduced monsoon rainfall. This results in a negative feedback, halting further increase in soil moisture. These feedback mechanisms are dependent not only on local processes but also on the remote forcing such as forced large-scale descent or ascent over the AAM region by ENSO. The large-scale vertical

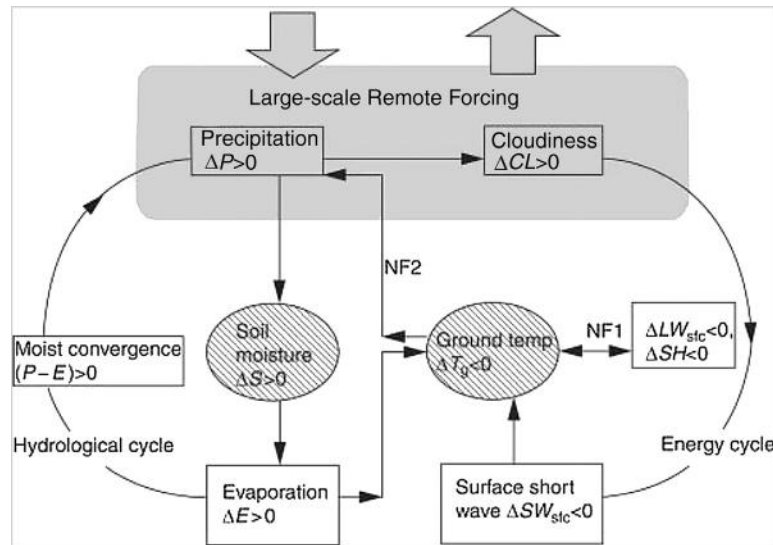


Figure 3 Schematic showing elements of land–atmosphere water and energy cycle feedback for the Asian–Australian monsoon (AAM) region. Negative feedback links are denoted by the symbol NF. The symbols used are for precipitation (P), evaporation (E), soil moisture (S), cloudiness (CL), ground temperature (T_g), surface long-wave radiation (LW_{stc}), surface short-wave radiation (SW_{stc}) and sensible heat flux (SH). The sign of the anomalies indicate the possible feedback mechanisms for a persistent cool, wet monsoon climate when the ENSO forcing favors upward motion over the AAM region.

motions provide a strong control on atmospheric stability and initiation of convection. Even though the ENSO remote forcing has relatively slow timescales, its impact may be sufficient to tip the delicate balance of the aforementioned local feedback processes, causing either the persistence of a given climate state or transition from one state to the other.

Intraseasonal Variability

One of the key characteristics of the monsoon is the presence of a rich spectrum of intraseasonal variability. These include quasi-periodic oscillations of 30–60 days or 10–20 days and transient waves of 3–5 days. The intraseasonal variabilities are generated by internal atmospheric dynamics but strongly modified by sea surface temperature and land surface processes. They are responsible for the modulation of monsoon onsets, breaks, and evolution regionally. Intraseasonal variabilities, especially those in the lower-frequency end of the spectrum, can have strong impacts on the seasonal mean monsoon climate. Over different regions, they can either strengthen or weaken the direct influence by ENSO on the monsoons. It has been suggested that, the near normal monsoon rainfall over India during the strong El Niño of 1997–98 may be due to the effects of pronounced intraseasonal variability, which brought copious rainfall to many parts of India in spite of the tendency of ENSO to weaken the AAM.

Biennial Variability

Rainfall records in many monsoon regions show a strong biennial tendency, i.e., a strong monsoon followed by a weak monsoon, and vice versa. Some aspects of this tendency can be discerned from Figure 1. The origin of the biennial tendency in the monsoon is still unclear, but it may be related to local air–sea interaction as well as basin-scale coupled processes. Coincidentally, a strong biennial tendency has also been

found in ENSO cycles. Except for the different timescale, the evolutionary features of the biennial oscillation in sea surface temperature, sea level pressure, wind, and precipitation are very similar to those of ENSO. Recent studies have suggested that strong monsoon–ENSO interactions may be manifested in a strong biennial tendency in ENSO cycles. Because the monsoon wind forcings have strong seasonality, and their directions and magnitudes are strongly tied to ENSO, it is possible that monsoon wind forcings may induce a biennial time scale in ENSO.

Interdecadal Variability

Because of multiple contributing factors, the relationship between monsoon and ENSO is likely to be nonstationary. Webster and colleagues showed that during the late 1800s and early 1900s the correlation between Indian rainfall and the Southern Oscillation Index (SOI) was relatively high (~ 0.6), but in the 1920s–40s the correlation dropped to less than 0.2–0.3. It rose again to 0.5–0.6 in 1960s–80s. In more recent decades, there is again a sharp decline in the correlation to less than 0.2. These changes appear to correlate with changes in sea level pressure and large scale circulation patterns. The interdecadal modulation of monsoon–ENSO relationship has led some authors to suggest that the recent weakening of the relationship may be linked to changes in the structure of ENSO, perhaps due to global warming. However, these assertions should be treated with extreme caution, because of the uncertainty in the historical data records and the sensitivity of the analysis to long-term trends.

Future Prospects

Year after year, occurrences of devastating monsoon-related droughts and floods serve as a reminder that improved

prediction of the monsoon rainfall is paramount for the well-being of millions living in the monsoon regions. Today, monsoon prediction is still a very challenging problem. With the better understanding of monsoon–ENSO interaction, and its relationship with other contributing factors, scientists are hopeful that they are a step closer to better understanding the causes and the predictability of monsoon rainfall anomalies and to devising better monsoon prediction schemes.

See also: Tropical Meteorology and Climate: El Niño and the Southern Oscillation: Observation; El Niño and the Southern Oscillation: Theory; Monsoon: Dynamical Theory; Walker Circulation.

Further Reading

- Bjerknes, J., 1969. Atmospheric teleconnections from the equatorial Pacific. *Monthly Weather Review* 97, 163–172.
- Kim, K.M., Lau, K.M., 2000. Mechanism of monsoon induced biennial variability in ENSO. *Geophysical Research Letters* 28, 315–318.
- Lau, K.-M., Bua, W., 1998. Mechanism of monsoon–Southern Oscillation coupling: insights from GCM experiments. *Climate Dynamics* 14, 759–779.
- Lau, K.M., Sheu, P.J., 1988. Annual cycle, quasi-biennial oscillation and Southern Oscillation in global precipitation. *Journal of Geophysical Research* 93, 10975–10988.
- Lau, K.M., Wu, H.T., 2000. Intrinsic coupled ocean-atmosphere modes of the Asian summer monsoon: a re-assessment of monsoon–ENSO relationships. *Journal of Climate* 14, 2880–2895.
- Meehl, G.A., 1994. Coupled land–ocean-atmosphere processes and South Asian monsoon variability. *Science* 265, 263–267.
- Rasmusson, E.M., Carpenter, T.H., 1983. The relationship between eastern equatorial Pacific seas surface temperature and rainfall over India and Sri Lanka. *Monthly Weather Review* 111, 517–528.
- Rasmusson, E.M., Wang, X., Ropelewski, C.F., 1990. The biennial component of ENSO variability. *Journal of Marine Systems* 1, 71–96.
- Shen, S.-H., Lau, K.M., 1995. Biennial oscillation associated with the East Asian summer monsoon and tropical sea surface temperature. *Journal of the Meteorological Society of Japan* 73, 105–124.
- Shukla, J., Paolino, D., 1983. The Southern Oscillation and the long-range forecasting of monsoon rainfall over India. *Monthly Weather Review* 111, 1830–1837.
- Walker, G.T., 1923. Correlation in seasonal variations in weather III: a preliminary study of world weather. *Memoir of Indian Meteorological Department* 24, 75–131.
- Walker, G.T., 1924. Correlations in seasonal variations of weather. IV: A further study of world weather. *Memoir of the Indian Meteorological Department* 24, 275–332.
- Webster, P.J., Yang, S., 1992. Monsoon and ENSO: selectively interactive systems. *Quarterly Journal of the Meteorological Society* 118, 877–926.
- Webster, P.J., Magana, V.O., Palmer, T.N., et al., 1998. Monsoon: processes, predictability, and the prospects for prediction. *Journal of Geophysical Research – Oceans* 103, 14451–14510.

Tropical Climates

S Hastenrath, University of Wisconsin, Madison, WI, USA

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

Radiative forcing and diurnal and local circulations are crucial in the tropical circulation and climate system. On the large scale there is prevailing subsidence and high surface pressure in the subtropics, in the lower layers heat and moisture transport equatorward, and convergence, ascending motion and abundant rainfall in the equatorial zone. Seasonality is very pronounced in the monsoon regions of southern Asia and Africa. Equatorial zonal circulations develop over the oceans. Interannual variability of circulation can have severe climatic implications in some tropical regions.

Introduction

For the understanding of the global climate system, processes in the tropics merit particular attention. Research on the low latitudes in recent decades has developed an increasing appreciation for the large time and space scales, and has turned its focus on the dynamics of climate, as compared with the emphasis on weather events and smaller-scale processes which prevailed throughout the earlier part of the past century. 30° N and 30° S are plausible approximate boundaries of the tropics, considering the location of the subtropical high-pressure cells; the domain of net radiative heat gain at the top of the atmosphere; the prevalence of diurnal over the annual cycle of solar radiation and temperature; and the latitudinal variation of the Coriolis parameter.

Diurnal and Local Processes

In the tropics, day-periodic processes and local to meso-scale circulations are much more vigorous than in higher latitudes. As a result of radiation geometry, the amplitude of the diurnal cycle of insolation and temperature is much larger than the annual cycle. Over the open ocean, the cloudiness–rainfall maximum occurs commonly in the late night to early morning hours. Over the tropical land areas, land- and seabreeze and mountain circulations dominate the diurnal march of cloudiness, rainfall, and weather. Anabatic flows lead to an afternoon maximum on the mountains, with compensating subsidence and clearing over the basins. Rainfall in the basins occurs with some preference at night. At the coasts, a land breeze may interact with large-scale flow to produce a convergence and rainfall maximum in the latter part of the night. Large inland lakes experience lower-tropospheric convergence, cloud cover and rainfall during the night, and subsidence and clear skies during the day. Diurnal mass exchanges between highlands and adjacent plains on the scale of the hundreds of kilometers may be responsible for the suppression of daytime convection and the origin of a nighttime cloudiness–rainfall maximum over the lowlands.

The marked diurnal and local controls in the tropics are important factors for the regional climates and the large-scale circulation. Thus in many tropical regions the annual rainfall distribution can be appreciated only from the interaction between diurnal factors, local circulations, and the large-scale

flow. The resultant spatial pattern of latent heating is in turn essential in the driving of the circulation on a planetary scale.

Planetary Context

The zonally averaged planetary scale circulation of the tropics consists of an easterly wind regime bounded by the anticyclonic axes of the subtropical highs and the two thermally direct mean meridional circulation cells (Hadley circulations), sketched in Figure 1. The tropical atmosphere is a source of westerly absolute angular momentum and of kinetic energy. The required poleward transports in low latitudes are performed primarily by the mean meridional circulation, while beyond the fringe of the tropics eddy mechanisms assume the dominant role. The intensity of the mean meridional circulation and the associated production and export of kinetic energy are largest in the respective winter. In mechanical terms, the tropical atmosphere thus is instrumental in the maintenance of the global circulation.

Further important constraints of the general circulation relate to the budgets of heat and moisture, and these can be adequately treated only for the atmosphere–hydrosphere

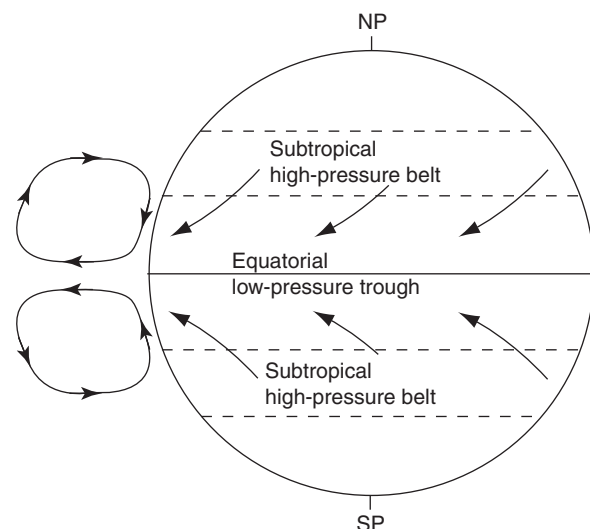


Figure 1 Schematic sketch of planetary high- and low-pressure belts and mean meridional circulations.

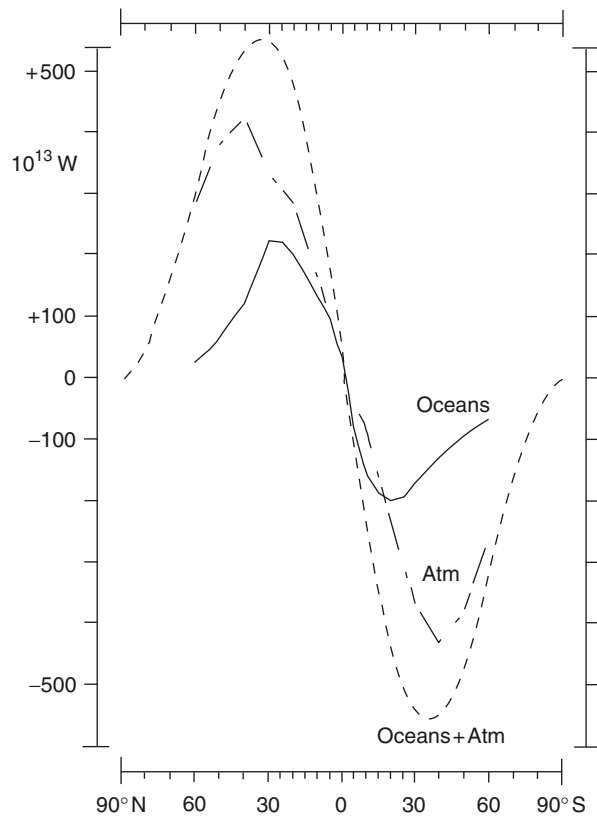


Figure 2 Annual mean meridional transports within the coupled atmosphere–hydrosphere system (dashed), within the oceans (solid), and within the atmosphere (dash-dotted). Positive values indicate northward heat flux. Adapted from Hastenrath, S., 1995. *Climate Dynamics of the Tropics*. Kluwer, Dordrecht, Boston, London.

system as a whole. Fundamental to the heat budget of the planet Earth is the net radiation at the top of the atmosphere. This defines the total required poleward transport of heat in the coupled atmosphere–ocean system. Through the analysis of atmospheric transports (from upper-air sounding) and of the heat fluxes across the ocean surface (from ship observations) this has been partitioned into the contributions from the atmosphere and the hydrosphere, respectively. Results are presented in **Figure 2**. The required total poleward transport in the combined system is largest around 30° N and 30° S. The atmospheric transport is largest in the mid latitudes. In the tropics, the oceans accounts for about half the total poleward transport, with the largest amounts around 30° N and 20° S. The transport of water vapor and latent heat in the atmosphere is related directly to the meridional pattern of precipitation and evaporation from the surface. In the mid latitudes transient eddies account for most of the poleward transport of the atmospheric water vapor. By contrast, in the tropics the mean meridional circulation carries water vapor and latent heat equatorward, and this transport is concentrated in the surface layer. Latent heat release and precipitation are concentrated in the equatorial zone, from where sensible heat and geopotential energy are exported poleward in the upper portion of the mean meridional circulation.

Regional Circulation Systems

Complementing the latitude-mean sketch in **Figure 1**, **Figure 3** portrays the spatial patterns of the lower- tropospheric circulation during the extremes of the annual cycle. Persistent features throughout the year are the three subtropical highs over the southern oceans, the two highs over the North Pacific and North Atlantic, and the enclosed near-equatorial trough extending around the globe. In austral summer, the thermally induced near-equatorial trough is centered to the south of the Equator, and thermal lows are found over the three southern

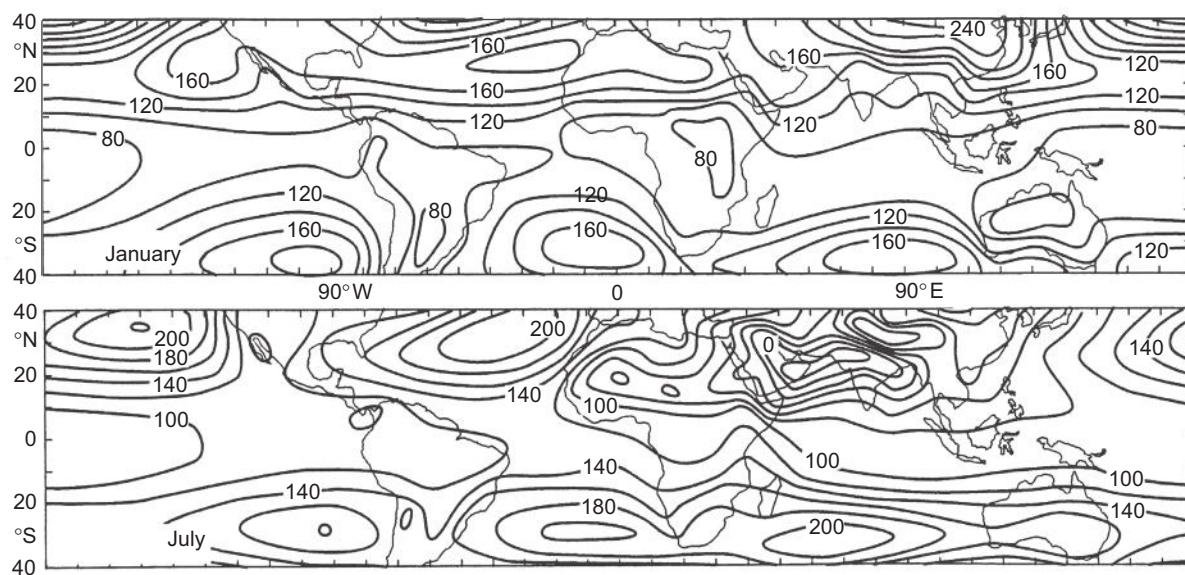


Figure 3 Mean patterns of 1000 mbar topography, with isoline spacing of 10 gpm in January and July. (Period 1958–97.)

continents. In boreal summer, the subtropical highs of both hemispheres and the enclosed near-equatorial trough appear displaced northward, and as part of this the monsoon heat low extends from northern hemispheric Africa to southern Asia. The domain of tropical circulation is meaningfully delineated by the subtropical high-pressure belts and anticyclonic axes of either hemisphere. From there the lower-tropospheric trade winds emanate to meet within a band of highest surface temperature and low-pressure trough near the Equator. The seasonality of heat-low processes is primarily responsible for the annual latitude migration of the equatorial trough.

Jet Streams

In the low latitudes various jet stream systems are developed at different times of the year. The upper-tropospheric Subtropical Westerly Jet is a feature of the respective winter hemisphere and owes its existence to the convergence of the poleward transport of absolute angular momentum in the upper and poleward portion of the Hadley cells, and to other factors. The likewise upper-tropospheric Tropical Easterly Jet, extending from South-East Asia over the Indian Ocean and Africa to the Atlantic, is confined to the height of the summer and is related to the thermal wind pattern associated with the then strongly heated subtropics and the cooler equatorial atmosphere. Through cross-circulations in the entrance and exit regions it exerts a control on the surface climate. The West African Mid-Tropospheric Easterly Jet is also limited to the boreal summer and related to the thermal wind pattern, in this instance associated with the hot desert air to the north and the cool monsoon air to the south of the Intertropical Discontinuity. The East African Low-Level Jet (or Findlater Jet) appears in boreal summer at about 1 km, extending with clockwise curvature from the Southern Indian Ocean across the Equator to the Arabian Sea. It has been called the 'backbone' of the south-west monsoon circulation. The Equatorial Mid-Tropospheric Easterly Jet, a recent discovery, is related to the tongue of cold surface waters in the equatorial Pacific.

Subtropical Highs and Trade Winds

The subtropical highs are the source of the trades and function as major centers of action for the tropical circulation. They are located farthest away from the Equator during the respective summer, but in both hemispheres they assume a westernmost position in boreal summer. The trades represent the lower-tropospheric portion of the Hadley cells. They pick up moisture (and to a lesser extent sensible heat) from the tropical oceans, accumulate it below the trade inversion, and carry it into the equatorial trough zone, where rainfall and latent heat release are concentrated. The trades thus serve an important role in global energetics. The trade inversion is lowest and best developed in the eastern equatorward sector of the subtropical highs, and rises and weakens both equatorward and toward the central and western part both of the Atlantic and the Pacific. Large-scale subsidence is the major factor in the origin and maintenance of the trade inversion. [Figure 4](#) illustrates the rise

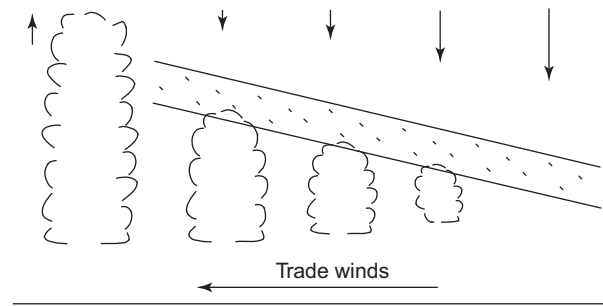


Figure 4 Sketch illustrating the rise of the trade inversion, change from subsidence to ascending motion, and increasing convective activity, along the tradewind trajectory.

of the trade inversion, the decreasing subsidence, and the increasing convective activity, along the trade wind trajectory.

Equatorial Trough Zone

The equatorial trough zone and associated quasi-permanent circulation features are in their development and annual latitude migration controlled by heat-low mechanisms. Embedded within broad and coincident bands of high surface temperature and low pressure are an axis of confluence between airstreams of northern- and southern-hemispheric origin and a belt of maximum convergence–cloudiness–rainfall. The latter is typically located well away from the confluence axis, the separation being particularly large over northern-hemispheric Africa and the adjacent Atlantic. Thus insolation in the latitude of highest surface temperature and lowest pressure is ensured by comparatively scarce cloudiness. The meridional-vertical cross-section in [Figure 5](#) illustrates the structure of the equatorial trough zone in the West African–tropical Atlantic sector. In the season and longitudes of strong development and far northerly location of the flow discontinuity, the moist cross-equatorial flow from the southern hemisphere undercuts the Northeast trades in a wedge

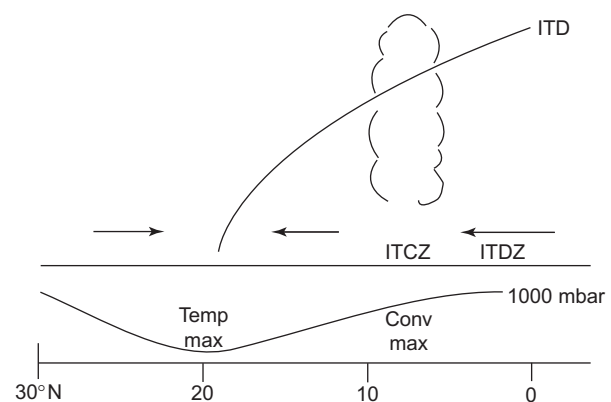


Figure 5 Schematic meridional-vertical cross-section across the Equatorial Trough Zone in the tropical Atlantic–West African sector, showing 1000 mbar topography, maxima of surface temperature (TEMP MAX) and convergence (CONV MAX); meridional component of north-east tradewinds and of cross-equatorial flow from the Southern Hemisphere; Intertropical Divergence Zone (ITDZ), and Intertropical Convergence Zone (ITCZ), and flow confluence (Intertropical Discontinuity, ITD).

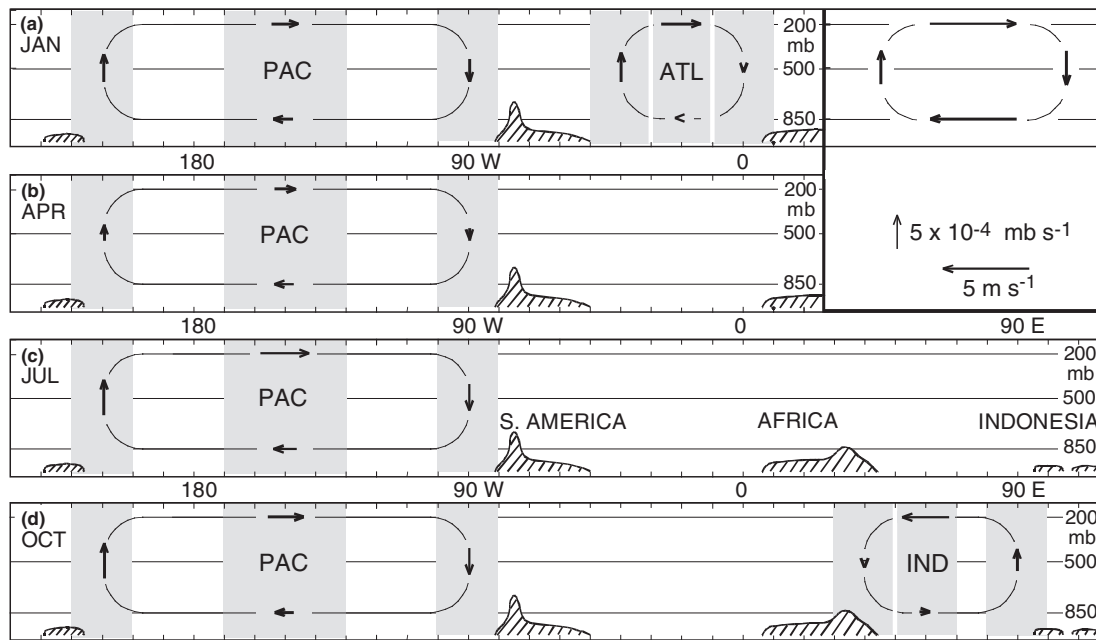


Figure 6 Zonal-vertical cross sections showing equatorial zonal circulation cells. Shading delimits the domains for which data were sampled, and horizontal lines mark the 200, 500 and 850 mb levels. Vertical arrows denote the omega vertical motion at 500 mb, and horizontal arrows horizontal divergent flow at 200 and 850 mb. Inset, upper right, explains the arrangement of arrows representing $5 \times 10^{-4} \text{ mb s}^{-1}$, and 5 m s^{-1} . (a) January, (b) April, (c) July, (d) October. Adapted from Hastenrath, S., 2007. Equatorial zonal circulations: historical perspectives. *Dynamics of Atmospheres and Oceans* 43, 16–24.

fashion. The dynamics of the cross-equatorial flow over the eastern Atlantic and Pacific are dominated by the latitude variation of the Coriolis parameter. Among the resulting characteristics are a recurvature of flow from south-easterly to south-westerly at about 5° N , a speed maximum at that latitude, a band of divergence between Equator and recurvature, the Intertropical Divergence Zone (ITDZ), and a band of convergence poleward from it, the Intertropical Convergence Zone (ITCZ). The Pacific Equatorial Dry Zone extending from the eastern to the central Pacific has ITDZ characteristics. Meridional climatic gradients from the Pacific Equatorial Dry Zone to the zonally oriented band of intense convergence–cloudiness–rainfall, the ITCZ, are among the steepest found on Earth in the absence of topographic effects. Over the Indian Ocean during the south-west monsoon the cross-equatorial flow recurves near the Equator, owing to a marked zonal component of the pressure gradient.

Monsoons

The monsoon area of the world is delineated primarily in terms of the complete annual reversal of wind regimes, thus encompassing the Indian Ocean sector and much of tropical Africa. Over sub-Saharan West Africa during boreal summer, a deep moist airstream from the southern hemisphere replaces and undercuts the dry Northeast trade winds originating from the Sahara. In the Indian Ocean sector during boreal winter, winds sweep from southern Asia across the Equator into the southern hemisphere. Of far greater proportions is the boreal summer Southwest monsoon. The establishment of a heat-low-induced monsoon trough over South Asia is instrumental in its development. In the South Indian Ocean Southeast trades

recurve, cross the Equator, and continue into the South Asian continent as the Southwest monsoon. Monsoon depressions and ‘breaks in monsoon’ are among the more important synoptic situations. On the Indian subcontinent and adjacent regions, the bulk of the annual precipitation falls during the Southwest monsoon. By contrast, the greater Indonesian region and parts of South-East Asia receive much of their annual rainfall during the boreal winter monsoon, with north-easterlies blowing over the South China Sea. More complicated regimes are found in equatorial East Africa, where the precipitation peaks are timed in the monsoon transition seasons. The bulk of the water vapor brought to condensation over South Asia during the boreal summer Southwest monsoon stems from south of the Equator. Atmosphere, hydrosphere, and lithosphere are all essential for the energetics of the monsoon.

Equatorial Zonal Circulations

The dynamics of zonal circulation cells in the near vicinity of the Equator are characterized by vanishing Coriolis acceleration, and thus balance between the pressure gradient and frictional accelerations. Pertinent are the coherence of vertical motion at the western and eastern extremities and the continuity of the divergent part of the zonal flow following the motion. **Figure 6** illustrates the existence of equatorial zonal circulation cells in the course of the annual cycle. A well-developed zonal circulation cell persists along the Pacific Equator all year round, the so-called ‘Walker circulation’, with ascending motion over the central Pacific, divergent to convergent eastward flow in the upper troposphere, subsidence over the eastern Pacific, and compensating divergent- to-convergent westward flow in the

realm of the Equatorial Mid-tropospheric Easterly Jet. Over the Indian Ocean, a distinct equatorial zonal circulation cell exists in boreal autumn only, with ascending motion over Indonesia, divergent-to-convergent westward flow in the upper troposphere, subsidence over the coast of East Africa, and compensating divergent-to-convergent eastward flow concentrated in the mid troposphere. In the Atlantic sector, a weak zonal cell is found in boreal winter, with ascending motion over the mouth of the Amazon, and upper-tropospheric eastward divergent outflow to a subsidence center over the eastern Atlantic.

Climatology of Weather Systems

Weather systems play an important role in the regional climate, and their occurrence is related to the large-scale circulation setting. In comparison with the higher latitudes there is a vast diversity of weather systems in the tropics. A few regionally important systems are considered here.

Most destructive are the Tropical Storms; systems beyond a certain intensity are called Hurricanes in the Americas, Typhoons in the Western Pacific, and Cyclones in the Indian Ocean. The major regions of storm formation are the western Atlantic and the Caribbean, the eastern Pacific, the eastern South Indian Ocean, the Arabian Sea and the Bay of Bengal.

Waves in the easterlies are the most common disturbances of the trade wind regime. Waves form with preference between about 15° E and 20° E to the south of the West African Mid-Tropospheric Easterly Jet. In boreal summer, a wave crosses the West coast of Africa about every 3–4 days. Passing on to the cold eastern Atlantic, the waves generally decay, but remnants survive to the western Atlantic and Caribbean, where some regenerate and account for about half of the Atlantic tropical Cyclones.

Squall Lines are among the more severe storms in the tropics. A system is typically hundreds of kilometers in meridional extent and consists of a line of active thunderstorms. These are particularly common over sub-Saharan Africa in boreal summer, but they have also been described from the western Atlantic, the equatorial central Pacific, and northern India. Over sub-Saharan West Africa they account for the bulk of the annual rainfall.

The Temporales of Pacific Central America originate as disturbances in the Intertropical Convergence Zone over the Eastern Tropical Pacific, especially during May–June and September–October. Temporales are feared for their abundant and sustained rainfall, and the hazards of landslides and flooding.

Wintertime cold-air intrusions from mid latitudes affect various regions of the global tropics, namely the Central American–Caribbean region, eastern South America, both northern- and southern-hemispheric Africa, and South-East Asia. Though of extratropical origin, they interact with tropical weather systems and play a distinct role in the regional climate.

Climatic Variability

The quasi-permanent regional circulation systems documented in a previous section undergo variations not only in the annual cycle but also from year to year and on longer time scales. A

direct consequence of such variability in the regional circulation is the variation in climate of tropical regions, with rainfall anomalies being of primary interest for the low latitudes. Much publicity has been given to the Southern Oscillation (SO), which has at its heart a pressure seesaw between the eastern and western extremities of the Pacific. To the extent that it is associated with changes in the quasi-permanent regional circulation systems, it may be reflected in regional rainfall anomalies; weak manifestations of the SO are indeed pervasive. Essential is the forcing by the regional circulation. Only a few regional themes are considered here.

Southern Oscillation and El Niño

The Southern Oscillation entails a large-scale pressure seesaw, on a time scale of 2–10 years, and with dipoles over the eastern South Pacific and the greater Indonesian-Australasian region, but spanning the global tropics. The low phase of the SO can be defined by anomalously low/high surface pressure at Tahiti/Darwin. Although these large-scale, long-term pressure variations have been studied since the latter part of the last century, their causal relation to the El Niño phenomenon on the West coast of South America has been recognized only since the 1970s.

During the high phase of the Southern Oscillation, both the eastern South Pacific high and the Indonesia low are strongly developed, and vigorous easterly trade winds sweep the equatorial Pacific, piling up waters at its western extremity. Accordingly, the eastward slope of the free ocean surface, the westward deepening of the oceanic mixed layer, and the Equatorial Undercurrent are pronounced, and waters are particularly cold off the South American West coast and in an extended zone immediately to the South of the Equator stretching from the coast of the Americas far into the open Pacific. Within the atmosphere, the Walker circulation along the Pacific Equator is strong, featuring not only vigorous surface easterlies but also an enhanced westerly return flow aloft, as well as pronounced convection and ascending motion over the Indonesian dipole and marked subsidence over the East Pacific dipole.

During the low phase of the Southern Oscillation, the atmosphere–ocean system in the Pacific operates in a remarkably different mode. Both the Eastern South Pacific high and the Indonesia low are weak; the slackened zonal pressure gradient entails weaker surface easterly winds in the equatorial zone, and accordingly the zonal slopes of the free ocean surface, of constant-pressure topographies at depth, and of the thermocline, diminish; the Equatorial Undercurrent slows down and may surface or vanish altogether. The relaxation of surface wind stress incites equatorial Kelvin waves which travel to the eastern extremity of the Pacific within 2–3 months, where they are manifest in a warming of surface waters, with maximum around the March–April peak in the annual march. The warm ocean and torrential rains at the otherwise deserts coast of Peru and Ecuador trigger an ecological catastrophe, including the mass death of fish and Guano birds, floods, destruction of roads and houses, and loss of human life. Within the atmosphere, the Walker circulation along the Pacific Equator is weak, as manifest in the slackened trade winds and westerly return flow aloft, as well as reduced convection,

rainfall, and ascending motion over the Indonesia dipole and lesser subsidence over the eastern Pacific.

Variations of the upper-air circulation are an essential part of the El Niño Southern Oscillation (ENSO) phenomenon. During the high or cold phase, the troposphere is anomalously cold throughout the global tropics, so that upper-tropospheric topographies are low, particularly in the lower latitudes, thus entailing departure easterlies and equatorward flow aloft. By contrast, during the low or warm phase, the tropical troposphere is warm, upper-tropospheric topographies are inflated – especially in the tropics – and departure westerlies and poleward flow prevail in the upper troposphere. These variations appear broadly in phase throughout the tropical belt, but the oceanic warming and cooling, and consequently the temperature variations in the overlying atmosphere, as well as the upper-tropospheric height variations, are most pronounced in the eastern to central Pacific. This allows for zonal flow departures in the upper troposphere over the equatorial zone of the Western Pacific to run essentially inverse to most of the remainder of the tropics. The aforementioned zonal flow variations in the upper troposphere over the equatorial western Pacific are an integral part of the modulations in the Walker circulation.

Indian Monsoon

Indian south-west summer monsoon rainfall is associated with characteristic circulation departure patterns from the pre-monsoon throughout the post-monsoon seasons. Abundant rainfall is heralded by strong heat low development over the continent, strong surface wind and a warm Arabian Sea preceding the monsoon onset. The negative correlation with pressure and the positive coupling with wind persist to the post-monsoon season. Enhanced upper-tropospheric easterlies throughout the summer half-year are further indicative of a good monsoon year. Abundant rainfall years tend to coincide with the high or cold phase of the Southern Oscillation.

Northeast Brazil

The droughts of northern Northeast Brazil (the Nordeste), which has its rainy season concentrated mainly in March–April, are characterized by an anomalously far poleward position of the near-equatorial trough and embedded confluence axis and convergence band; positive sea surface temperature departures in the tropical North Atlantic; and anomalously cold waters in the equatorial South Atlantic. Surface circulation features conducive to drought include the distant position of the convergence band; the cold south equatorial water and its resulting effects on moisture and instability of the boundary layer flow; and enhanced meridional temperature contrasts across the Equator, which drive a thermally direct meridional circulation cell in the atmosphere, featuring subsidence over the Nordeste. Departure patterns in the large-scale atmospheric and ocean fields are approximately inverse in the dry and wet years, and evolve during the half-year preceding the Nordeste rainy season, thus offering the prospect of climate prediction.

Sahel

Droughts in sub-Saharan Africa, which experiences its rainfall at the height of the boreal summer, are associated with an anomalously far equatorward position of quasi-permanent circulation features, in particular the surface confluence axis and convergence band, and negative sea surface temperature anomalies are found in a broad band across the tropical North Atlantic. Anomalously warm surface waters in the western Indian Ocean and in the equatorial Pacific also tend to be associated with deficient Sahel rainfall.

In addition to the interannual variability, the Sahel zone of West Africa has experienced since the middle of the past century a trend towards much drier conditions, in tune with the long-term evolution of the large-scale circulation setting. This is in part illustrated in Figure 7. Given the steep meridional gradient of annual rainfall totals from the Sahara desert to the coast of the Gulf of Guinea, the drastic downward trend of Sahel rainfall is commensurate with the observed southward displacement of the wind confluence over the adjacent Atlantic. A further characteristic of the evolution from the middle of the century to the mid 1980s was a cooling of the tropical North Atlantic. Since the mid 1980s there have been tenuous indications for a recovery from the prolonged drought regime, along with a northward shift of the wind confluence and the warming of the tropical North Atlantic waters. The long-term trend in Sahel rainfall must be seen in the context of an evolving general circulation, rather than as a mere response to the local alterations of surface conditions.

Time Scales of Variability

Interannual variability appears concentrated in various preferred time scales. A quasi-biennial oscillation is apparent in India's rainfall, in the circulation over the equatorial Atlantic and in rainfall over Northeast Brazil, as well as in other elements and areas. Northeast Brazilian rainfall and the equatorial Atlantic circulation further exhibit spectral power at around 13–14 years. By contrast, interannual variability in various regions of the outer tropics and subtropics of the Americas and Africa (Central American–Caribbean area, sub-Saharan Africa, Southern Africa, subtropical South America) is concentrated at a time scale of 2–3 decades. Positive-feedback mechanisms may play a role in the persistence of anomalous regimes in these semi-arid to semi-humid regions. The coexistence of various preferred time scales of interannual variability in the Atlantic and surrounding continents is re-markable.

Climate Prediction

Climate anomalies such as those discussed in the preceding sections can have a severe social and economic impact. Accordingly, it has long been found desirable to find ways of predicting such events well in advance. While the hazards posed by certain tropical weather systems, in particular the Tropical Storms and Squall Lines, are recognized, in the tropics seasonal forecasting has arguably even greater practical relevance than weather forecasting. The diagnostic understanding

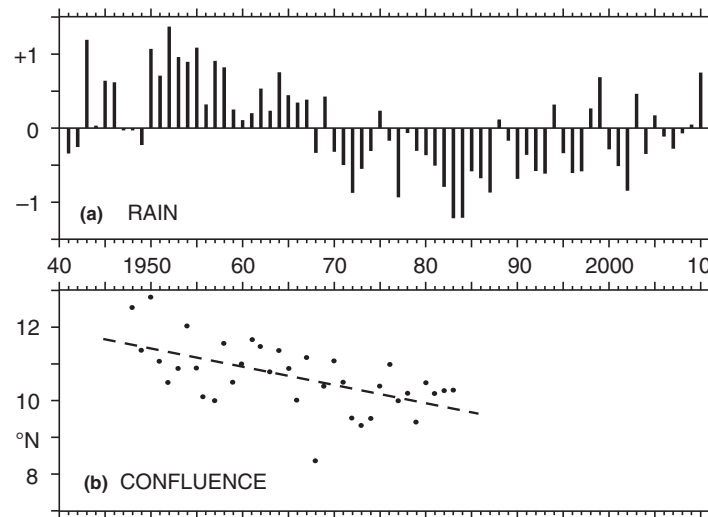


Figure 7 Long-term evolution of circulation and climate in the West African–Atlantic sector (a) Sahel rainfall; (b) latitude of wind confluence (zero meridional wind component) over the tropical Atlantic. Adapted from Hastenrath, S., 1995. *Climate Dynamics of the Tropics*. Kluwer, Dordrecht, Boston, London.

of circulation and climate variability is essential for the development of climate prediction methods. Empirical as well as numerical modeling approaches are being used to that end. There are prospects of success for certain tropical regions.

See also: **Boundary Layer (Atmospheric) and Air Pollution:** Diurnal Cycle. **Climate and Climate Change:** Climate Variability: Decadal to Centennial Variability; Climate Variability: Seasonal and Interannual Variability; Overview. **General Circulation of the Atmosphere:** Energy Cycle; Mean Characteristics; Overview. **Mountain Meteorology:** Land and Sea Breezes. **Numerical Models:** General Circulation Models. **Oceanographic Topics:** General Processes; Surface/Wind Driven Circulation. **Tropical Meteorology and Climate:** El Niño and the Southern Oscillation: Observation; El Niño and the Southern Oscillation: Theory; Equatorial Waves; Intertropical Convergence Zone; Monsoon: ENSO–Monsoon Interactions; Monsoon: Overview; Walker Circulation.

- Griffiths, J. (Ed.), 1972. *Climates of Africa*. World Survey of Climatology, vol. 10. Elsevier, Amsterdam.
- Hastenrath, S., 1995. *Climate Dynamics of the Tropics*. Kluwer, Dordrecht, Boston, London.
- Hastenrath, S., 2007. Equatorial zonal circulations: historical perspectives. *Dynamics of Atmospheres and Oceans* 43, 16–24.
- Hastenrath, S., Lamb, P., 1997. *Climatic Atlas of the Tropical Atlantic and Eastern Pacific Ocean*. University of Wisconsin Press, Madison.
- Hastenrath, S., Lamb, P., Greischar, L., 1989. *Climatic Atlas of the Indian Ocean: Part I: Surface Climate and Atmospheric Circulation; Part II: The Oceanic Heat Budget; Part III: Upper-Ocean Structure*. University of Wisconsin Press, Madison.
- Ramage, C., 1971. *Monsoon Meteorology*. Academic Press, London.
- Schwerdtfeger, W. (Ed.), 1976. *Climates of Central and South America*. World Survey of Climatology, vol. 12. Elsevier, Amsterdam.
- Takahashi, K., Arakawa, H. (Eds.), 1981. *Climates of Southern and Western Asia*. World Survey of Climatology, vol. 9. Elsevier, Amsterdam.
- Webster, P.J., Magana, V.O., Palmer, T.N., Shukla, J., Tomas, R.A., Yanai, M., Yasunari, T., 1998. Monsoons: processes, predictability, and the prospects for prediction. *Journal of Geophysical Research: Atmospheres* 103 (C7), 14451–14510.

Further Reading

- Arakawa, H. (Ed.), 1969. *Climates of Central and South America*. World Survey of Climatology, vol. 12. Elsevier, Amsterdam.
- Das, P.K., 1986. Monsoons. Fifth IMO lecture, WMO-No. 613. World Meteorological Organization, Geneva.

Walker Circulation

K-M Lau, NASA/Goddard Space Flight Center, Greenbelt, MD, USA

S Yang, NOAA/NWS/NCEP, Climate Prediction Center, Camp Springs, MD, USA

© 2003 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 6, pp 2505–2510, © 2003, Elsevier Ltd.

Introduction

The term Walker Circulation was first introduced in 1969 by Professor Jacob Bjerknes, referring to the large-scale atmospheric circulation along the longitude–height plane over the equatorial Pacific Ocean. The Walker Circulation features low-level winds blowing from east to west across the central Pacific, rising motion over the warm water of the western Pacific, returning flow from west to east in the upper troposphere, and sinking motion over the cold water of the eastern Pacific. Since Bjerknes's introduction of the Walker Circulation, there have been reports of similar east–west circulation cells spanning different longitudinal sectors along the Equator. Today, the Walker Circulation generally refers to the totality of the circulation cells as shown in Figure 1.

Bjerknes originally named the Pacific east–west circulation the Walker Circulation because he considered it the key part of Sir Gilbert Walker's Southern Oscillation (*see Tropical Meteorology and Climate: El Niño and the Southern Oscillation: Observation*). He interpreted the Walker Circulation as an atmospheric circulation driven by the gradient of sea surface temperature along the Equator and suggested that the characteristics of the Walker Circulation were largely determined by the coupling between the tropical atmosphere and oceans.

Bjerknes's work on the Walker Circulation marked an important milestone toward our basic understanding of the dynamics of zonal atmosphere–ocean coupling along the equatorial Pacific Ocean. Although his results were based on very limited data, Bjerknes's original conjecture that the year-to-year variation of the Walker Circulation is closely tied to that of the Southern Oscillation and El Niño has been confirmed by a large number of observational and modeling studies during the several decades since his first report.

Climatology and Variability

Annual Mean

Thanks to the advance in satellite observations and improved assimilation of observations into global general circulation models, now a much more detailed and quantitative description of the Walker Circulation is available. It is known that the tropical wind is made up of rotational and divergent components. The former is directly related to the effects of the rotation of the Earth and the latter to the overturning circulation, driven by atmospheric heating processes. The Walker Circulation and associated overturnings in the equatorial plane should refer only to the divergent component of the wind. Figure 2(a) shows the annual climatology (the mean state of all months) of the overturning circulations along the equatorial plane as streamlines constructed from the divergent zonal and vertical winds. It can be seen that the major rising branch of the Walker Circulation is found over the western Pacific and maritime continent, with a maximum in the upper troposphere (300–200 hPa) over Indonesia (115–120° E). A westward tilt with height of the ascending motion is also apparent. Over the eastern Pacific, there is a broad region of subsidence, with maximum descent in the coastal region of South America (~80° W). Connecting the ascending and descending branches are low-level easterlies and upper level westerlies over the central and eastern Pacific with strong low-level convergence near 160° E.

As is evident in Figure 2(a), the Walker Circulation also includes secondary circulation cells whose rising motions appear over the land regions of South American and Africa, with compensating subsidence over the Atlantic and the Indian Ocean. Compared with the Pacific branch of the Walker Circulation, these cells cover smaller longitude ranges and tend to have weaker vertical motions in the annual mean climatology.

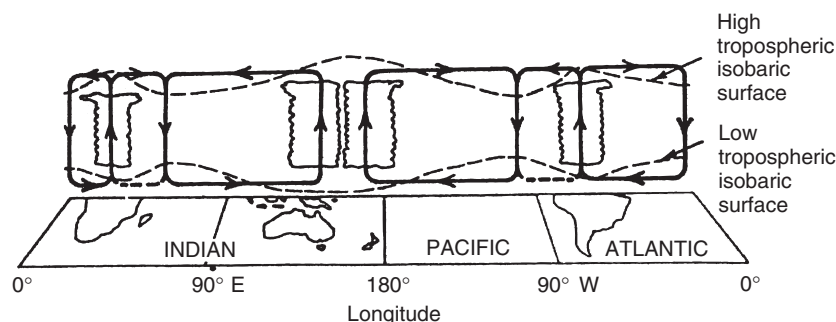


Figure 1 Schematic view of the east–west atmospheric circulation along the longitude–height plane over the Equator. The cell over the Pacific Ocean is referred to as the Walker Circulation. Adapted from Webster, P.J., 1983. The large scale structure of the tropical atmosphere. In: Hoskins, B.J., Pearce, R.P., (Eds.), *General Circulation of the Atmosphere*. Academic Press, London, UK, pp. 235–275.

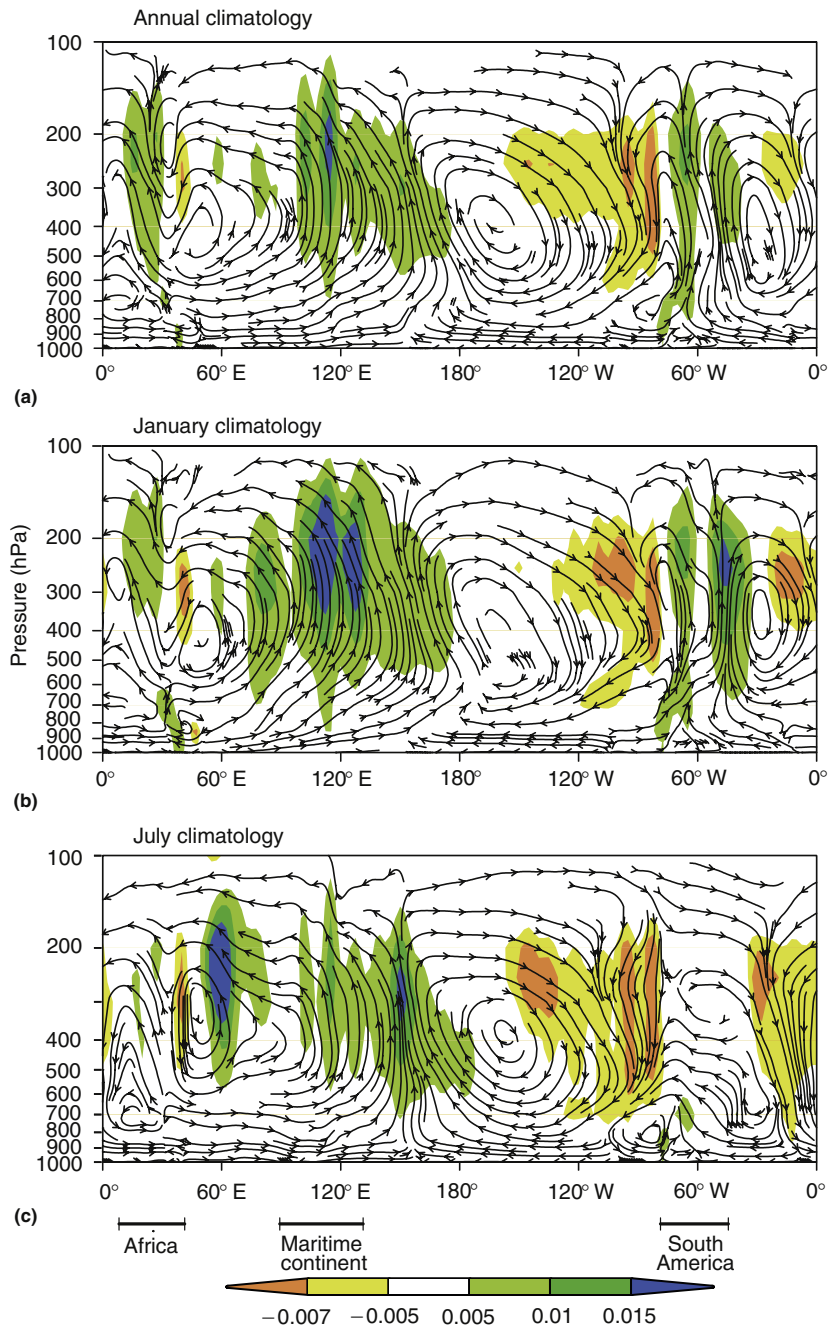


Figure 2 The equatorial east–west atmospheric circulation calculated from the reanalysis of data from the US National Centers for Environmental Prediction and National Center for Atmospheric Research. Shown are the climatologies of all months (top panel), January (middle panel), and July (bottom panel) for the period 1949–99. The streamlines are constructed from the divergent components of the winds (meters per second), with the values of the vertical motions multiplied by 30 times. Areas of strong upward (positive values) and downward motions (negative values) are colored.

Also closely linked to the Walker Circulation are the east–west circulations away from the equatorial region in the subtropics, associated with the large-scale monsoon circulation and the convergence of the trade winds. Some meteorologists consider these circulations part of the family of Walker cells.

Because of the prevailing subsidence, the atmosphere over the tropical eastern Pacific is highly stable. This stability is unfavorable to and limits the occurrence of deep clouds and precipitation. In contrast, over the warm pool (see

Oceanographic Topics: General Processes) of the western Pacific and Indonesian region, the air is unstable and deep convective clouds and heavy precipitation occur frequently. A significant portion of latent heat associated with cumulus convection (see **Thermodynamics: Moist (Unsaturated) Air**) that drives the global atmospheric circulation is released in the ascending region of the Walker Circulation over the warm pool. In addition, the Walker Circulation is associated with low sea level pressure in the west and high pressure in the east. The

basin-wide pressure gradient is the main driving force for the low-level zonal wind of the Walker Circulation. Consistent with the distribution of surface pressure and sea surface temperature, the lower troposphere is relatively warm in the ascending branch and cold in its descending branch. This means that the Walker Circulation is a thermally direct circulation, which converts available potential energy to kinetic energy of atmospheric motions.

Seasonal Variations

The location and intensity of the Walker Circulation undergo large fluctuations on seasonal to interannual timescales. Also affecting the Walker Circulation are intraseasonal fluctuations associated with the Madden and Julian Oscillation (see **Tropical Meteorology and Climate: Intraseasonal Oscillation (Madden–Julian Oscillation)**). The seasonal variation of the Walker Circulation is a reflection of the east–west swaying of the large-scale circulation in the tropical atmosphere in response to thermal contrasts between sea and land induced by the annual cycle of incoming solar radiation. In January (**Figure 2(b)**), the ascending branch of the Walker Circulation is very pronounced over the Indo-Pacific and maritime continent region ($60\text{--}120^\circ\text{ E}$) and the descending branch over the eastern Pacific ($150\text{--}90^\circ\text{ W}$). The South America–Atlantic cell is strong, with pronounced rising motion over the northeastern Brazil ($\sim 60^\circ\text{ W}$). However, the Africa–Indian Ocean cell is not very well established in January. In July (**Figure 2(c)**), the Walker Circulation intensifies and becomes well-defined, with the rising branch shifted eastward and concentrated near 150° E and the sinking branch over the eastern Pacific ($120\text{--}90^\circ\text{ W}$). The strengthening of the Walker Circulation is consistent with the seasonal development of the cold water over the equatorial eastern Pacific and the increased sea surface temperature gradient across the Pacific. Mid-tropospheric sinking motion is found over the eastern Indian Ocean. This sinking motion is connected by low-level westerly wind blowing from the Indian Ocean toward the western Pacific, joining the rising motion there. The overturning is completed by the returning upper level easterlies over the Indian Ocean. The Indian Ocean cell is associated with the development of the South Asian monsoon (see **Tropical Meteorology and Climate: Monsoon: Overview**). In July, the overturning motion over the South America–Atlantic sector is suppressed.

Interannual Variability

Figure 3(a) shows the Walker Circulation during January 1998 when an El Niño event was at its peak. Here, the east–west circulations comprising the Walker Circulation differed significantly from the climatology. Rising motions prevailed at almost all longitudes. In particular, strong ascent in the mid-troposphere replaced climatological descending motion over the central and eastern Pacific, where the water was anomalously warm due to El Niño. The Walker Circulation was weakened and became less organized. On the contrary, during January 1999 when a La Niña, or reverse El Niño, was at its peak, the Walker Circulation was enhanced and became very pronounced, with well-defined rising and sinking branches

(**Figure 3(b)**). While the Walker Circulation and South America–Atlantic cell intensified, the Indian Ocean cell weakened compared to the climatology (see **Tropical Meteorology and Climate: Monsoon: ENSO–Monsoon Interactions**). Changes in the Walker Circulation between El Niño and La Niña years shown above are accompanied by changes in cloud and rainfall patterns. In contrast, when the Walker Circulation weakens during an El Niño, clouds and precipitation increase over the central–eastern Pacific and decrease over the western Pacific. In contrast, when the Walker Circulation strengthens during a La Niña, clouds and precipitation are enhanced over the western Pacific and Indonesian region.

Ocean–Atmosphere Coupling

As Bjerknes postulated, the Walker Circulation and its variations are strongly coupled to fluctuations of the tropical sea surface temperature across the entire Pacific basin. A climatological Walker Circulation with strong surface easterlies maintains an equilibrium state in the tropical atmosphere and ocean in which the western Pacific is characterized by higher sea level, deeper thermocline, higher sea surface temperature, lower atmospheric surface pressure, and increased precipitation relative to the eastern Pacific. During an El Niño, a relaxation of the lower level easterlies, signaling a weakening of the Walker Circulation, is accompanied by weaker upwelling in the eastern Pacific, leveling of the thermocline, and reduction of sea surface temperature gradient across the Pacific. During a La Niña, changes of the opposite sign occur.

The apparently self-sustaining oscillations of the Walker Circulation stem from the interplay of various feedback processes associated with strong coupling of the tropical atmosphere and oceans. Warming of the western Pacific increases the convective potential energy of the atmosphere, resulting in deep convection and heavy precipitation. However, the more abundant cloud amounts associated with increased precipitation reduce the incoming solar radiation, causing the sea surface to cool, thus limiting the further development of atmospheric convection. This process tends to slow down the Walker Circulation and arrest the warming. On the other hand, the warm sea surface temperature and strong upward atmospheric motions will lead to stronger surface easterlies that induce further warm water in the western Pacific and cool water in the eastern Pacific through increased upwelling. The increased east–west sea surface temperature gradient can lead to a stronger Walker Circulation. The aforementioned process was interpreted by Bjerknes as an acceleration of the Walker Circulation that also provides for ‘an increase of the east–west temperature contrast that is the cause of the Walker Circulation in the first place’.

Impacts on World Weather and Climate

The Walker Circulation regulates global exchange of momentum, heat, and water vapor within the tropics via massive overturning motions. In doing so, it plays an important role in the balance of atmospheric energy in the equatorial region and in determining the characteristics of

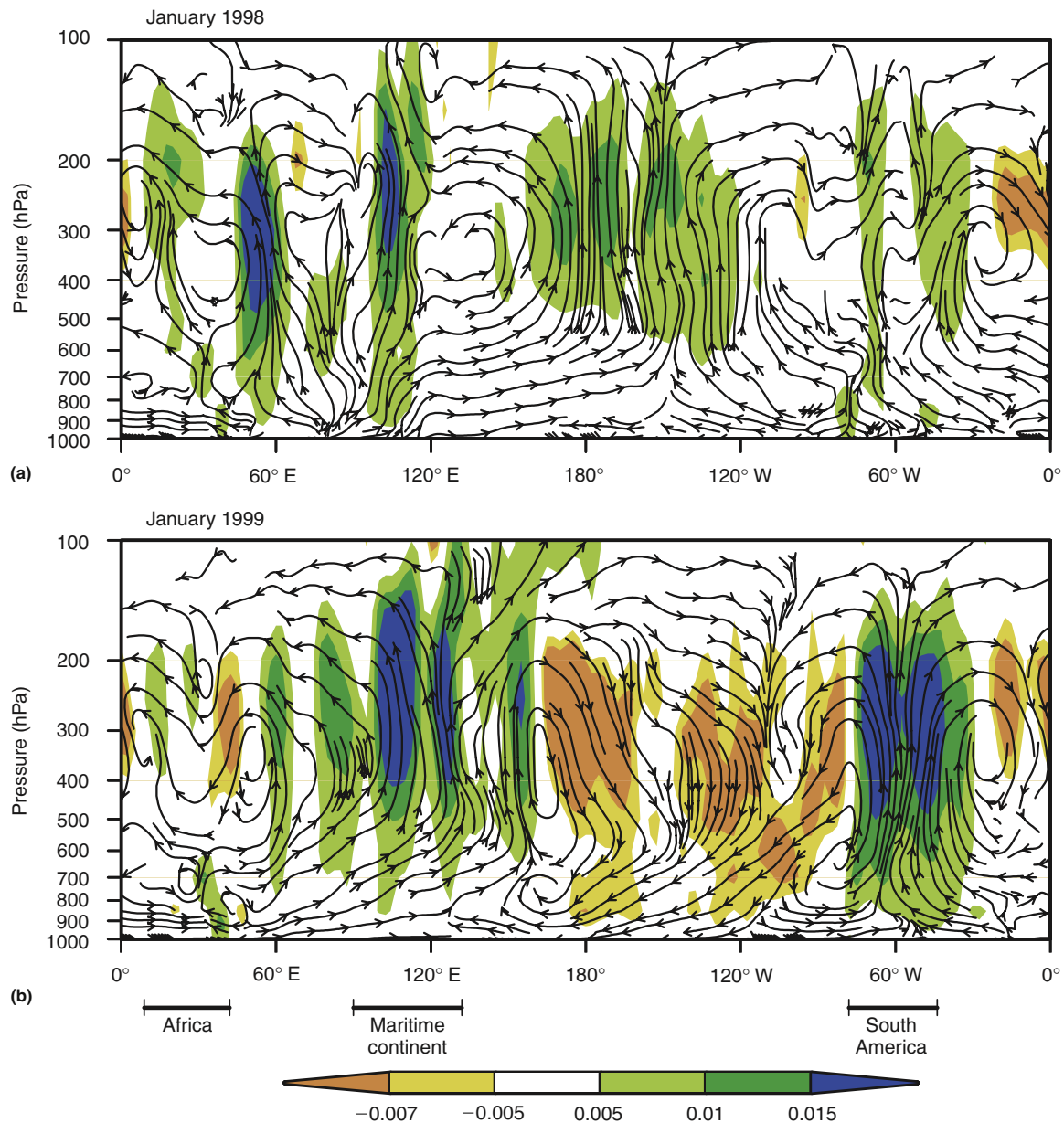


Figure 3 The equatorial east–west atmospheric circulation, calculated from the reanalysis of data from the US National Centers for Environmental Prediction and National Center for Atmospheric Research, for January 1998 (peak of an El Niño; panel (a)) and January 1999 (peak of a La Niña; panel (b)). The streamlines are constructed from the divergent components of the winds (meters per second), with the values of the vertical motions multiplied by 30 times. Areas of strong upward (positive values) and downward motions (negative values) are colored.

weather and climate in the tropics. The strongest atmospheric impacts associated with the fluctuations of the Walker Circulation are found over tropical and subtropical regions around the Pacific rim. During an El Niño, the weakening Walker Circulation causes widespread drought (*see Hydrology, Floods and Droughts: Drought*) in Indonesia-maritime continent, drought in northeastern Brazil, severe floods (*see Hydrology, Floods and Droughts: Flooding*) in Peru and Ecuador, and in southeastern Brazil and northern Argentina. During a La Niña, the Walker Circulation intensifies and leads to rainfall anomalies with reverse sign compared to El Niño. The Walker Circulation also represents

the fundamental link between the changes in sea surface temperature in the eastern Pacific and the variability of the Asian–Australian monsoon. The mechanisms that are responsible for the interactions between the monsoon and El Niño Southern Oscillation have been attributed, in part, to the changes in the Walker Circulation. Although individual cases may vary, in the summer preceding the peak phase of El Niño, which usually occurs in the northern winter, the Walker Circulation is weakened and shifted eastward owing to reduced east–west sea surface temperature gradient across the Pacific ocean. This suppresses broad-scale convection over the western Pacific and eastern Indian Ocean and leads to weaker

South Asian monsoon. There is some evidence that changes in land use, such as deforestation in Brazil and Indonesia, may also cause long-term changes in the Walker Circulation due to changes in land temperature, and therefore east–west thermal gradient along the tropical belt.

Besides modulating tropical weather and climate, the Walker Circulation is important as a driver of energy exchange between the tropics and higher latitudes. Bjerknes envisioned an interaction between the Walker Circulation and the Hadley Circulation in the form of an inverse variation between the two circulations. Bjerknes stated that “when the cold water belt along the Equator is well developed, the air above it will be too cold and heavy to join the ascending motion in the Hadley circulations. Instead, the equatorial air flows westward between the Hadley circulations of the two hemispheres to the warm west Pacific”. Such an interaction leads to changes in extra-tropical westerly jet streams (*see* **Synoptic Meteorology: Jet Streaks**) and the so-called Pacific–North American and Pacific–South American teleconnection patterns. These changes are the causes for severe weather and climate anomalies in the Asian–Pacific–American regions.

Summary

The Walker Circulation comprises east–west atmospheric circulation cells along the equatorial belt. The most dominant component is the Pacific branch, which consists of easterly winds at the lower troposphere, westerly winds at the upper troposphere, rising motion over the western Pacific, and subsidence over the eastern Pacific. The Walker Circulation possesses pronounced variability on seasonal intraseasonal and interannual timescales, and is an integral component of the El Niño–Southern Oscillation climate system. Fluctuations of the Walker Circulation can lead to extreme weather conditions in different parts of the world.

See also: **Hydrology, floods and droughts:** Drought; Flooding.
Oceanographic Topics: General Processes. **Synoptic**

Meteorology: Jet Streaks. **Thermodynamics:** Moist (Unsaturated) Air. **Tropical Meteorology and Climate:** El Niño and the Southern Oscillation: Observation; Intraseasonal Oscillation (Madden–Julian Oscillation); Monsoon: ENSO–Monsoon Interactions; Monsoon: Overview.

Further Reading

- Bjerknes, J., 1966. A possible response of the atmospheric Hadley circulation to equatorial anomalies of ocean temperature. *Tellus* 4, 820–829.
- Bjerknes, J., 1969. Atmospheric teleconnections from the equatorial Pacific. *Monthly Weather Review* 97, 163–172.
- Krishnamurti, T.N., 1971. Tropical east–west circulations during the northern summer. *Journal of the Atmospheric Sciences* 28, 1342–1347.
- Krishnamurti, T.N., Kanamitsu, M., Koss, W.J., Lee, J.D., 1973. Tropical east–west circulations during the northern winter. *Journal of the Atmospheric Sciences* 30, 780–787.
- Lau, K.M., Bua, W., 1998. Mechanisms of monsoon–Southern Oscillation coupling: Insights from GCM experiments. *Climate Dynamics* 14, 759–779.
- Troup, A.J., 1965. The ‘southern oscillation’. *Quarterly Journal of Royal Meteorological Society* 91, 490–506.
- Walker, G.T., 1923. Correlation in seasonal variations of weather, VIII: A preliminary study of world weather. *Memoirs of the Indian Meteorological Department, Calcutta* 24 (4), 75–131.
- Walker, G.T., 1924. Correlation in seasonal variations of weather, IX: A further study of world weather. *Memoirs of the Indian Meteorological Department, Calcutta* 24 (9), 275–332.
- Walker, G.T., 1928. *World weather III*. *Memoirs of the Royal Meteorological Society, London, UK* 2 (17), 97–106.
- Walker, G.T., Bliss, E.W., 1932. *World weather V*. *Memoirs of the Royal Meteorological Society, London, UK* 4 (36), 53–84.
- Wallace, J.M., Gutzler, D.S., 1981. Teleconnections in the potential height field during the Northern Hemisphere winter. *Monthly Weather Review* 109, 784–812.
- Webster, P.J., 1983. The large scale structure of the tropical atmosphere. In: Hoskins, B.J., Pearce, R.P. (Eds.), *General Circulation of the Atmosphere*. Academic Press, London, UK, pp. 235–275.
- Webster, P.J., Yang, S., 1992. Monsoon and ENSO: Selectively interactive systems. *Quarterly Journal of Royal Meteorological Society* 118, 877–926.
- Zhou, J., Lau, K.M., 2001. Principal modes of interannual and decadal variability of summer rainfall over South America. *International Journal of Climatology* 21, 1623–1644.

TROPOSPHERIC CHEMISTRY AND COMPOSITION

Contents

- Aerosols/Particles**
- Aliphatic Hydrocarbons**
- Aromatic Hydrocarbons**
- Biogenic Hydrocarbons**
- Cloud Chemistry**
- H₂**
- Hydroxyl Radical**
- Mercury**
- Oxidizing Capacity**
- Peroxyacetyl Nitrate**
- Sulfur Chemistry, Organic**
- Volatile Organic Compounds Overview: Anthropogenic**

Aerosols/Particles

JH Seinfeld, California Institute of Technology, Pasadena, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

Particles are ubiquitous in the atmosphere. The tropospheric aerosol is remarkably diverse in its composition, reflecting the wide range of particle sources in the atmosphere. Chemical components of tropospheric particles include inorganic materials such as sulfate, ammonium, nitrate, trace metals, and a wide array of carbonaceous compounds. Concentrations of airborne particles vary greatly over the globe, from the lowest concentrations in pristine areas to the highest levels in polluted urban centers. Atmospheric aerosols carry the chemical signature of the sources of direct particle emissions into the atmosphere as well as that of the conversion of gaseous molecules into particulate-phase species.

Introduction

Atmospheric particles (aerosols) range in size from a few nanometers to tens of micrometers in diameter. (An aerosol is strictly defined as a suspension of fine liquid or solid particles in a gas. In common usage, however, the particles themselves are referred to as the aerosol.) Primary aerosols are those particles emitted directly into the atmosphere. Particles or particulate materials are also formed in the atmosphere by nucleation or condensation of vapor species; such gas-to-particle conversion processes are an important source of atmospheric particulate material.

Particles are ubiquitous in the atmosphere; there is no region of it totally devoid of them. Particles are, in fact, the nuclei for the formation of clouds. Those particles that, in the presence of small amounts of water supersaturation, grow spontaneously to form cloud droplets are called cloud

condensation nuclei (CCN). The number of particles in a given ambient population that will form cloud droplets in a certain situation depends on the amount of ambient water supersaturation and the composition of the particles themselves. Particles containing water-soluble compounds preferentially act as CCN over those that contain largely insoluble compounds.

The troposphere is the lowest layer of the atmosphere, ranging from the Earth's surface to an altitude of about 10–15 km, depending on latitude and time of year. The temperature in the troposphere decreases with altitude – this leads to a dynamically unstable situation where mixing from the surface up to the top of the troposphere occurs on the order of a week or so. All of the atmosphere's weather and most of its water are in the troposphere. The atmospheric layer lying on top of the troposphere is the stratosphere, extending to an altitude of about 50 km. The temperature in the

stratosphere increases with altitude, leading to a dynamically stable layer in which vertical mixing processes are slow. The coldest point in the atmosphere between the surface of the Earth and the top of the stratosphere is at the troposphere–stratosphere boundary, the tropopause. Typical tropopause temperatures are about 220 K. When tropospheric air penetrates up into the stratosphere, water vapor in that air is frozen out at the tropopause; consequently, the stratosphere is very dry and generally cloudless. (Stratospheric clouds do form at the extremely cold temperatures of the winter stratosphere over the poles, where even the small amount of water vapor is frozen to form polar stratospheric clouds.)

Particles are eventually removed from the atmosphere by two routes: deposition at the Earth's surface ('dry' deposition) and incorporation into cloud droplets, either as the droplets form or as they fall ('wet' deposition). Because the troposphere is fairly vigorously mixed and because the incidence of precipitation is frequent, the lifetime of particles in the troposphere is relatively short, ranging from a few days to a few weeks. (Because the only route for the removal of stratospheric aerosols from the atmosphere is relatively slow transport down into the troposphere, where they are subject to removal by wet or dry deposition, stratospheric particles can have lifetimes of several years.) While not all particles released simultaneously into the atmosphere will remain in the atmosphere for an identical amount of time, there is an average residence time, or lifetime, for such particles. For a well-mixed atmospheric reservoir, the mean residence time of particles is the total mass of particles in the reservoir divided by their overall rate of removal, expressed in mass per time. Since there is a highly nonuniform geophysical distribution of particle sources over the globe, the relatively short residence time of particles leads to a highly nonuniform distribution of aerosols in the troposphere; that is, particles are simply not airborne long enough to become uniformly mixed throughout the troposphere. As a result, tropospheric aerosols tend to be categorized according

to the different areas of the Earth in which they are found, such as marine, desert, urban, rural, etc.

Chemical Components of Atmospheric Aerosols

This article addresses the composition of airborne particles in the troposphere. The tropospheric aerosol is remarkably diverse in its composition, reflecting the wide range of particle sources at the Earth's surface. Figure 1 depicts schematically the chemical components of tropospheric particles and their routes of incorporation into the aerosol phase. Chemical components of tropospheric aerosols include inorganic materials such as sulfate, ammonium, nitrate, sodium, chloride, trace metals, crustal elements, carbonaceous material, and water. Particles in different regions of the troposphere contain these species in differing amounts and proportions, but most of the mass of airborne tropospheric particles is represented by the collection of species depicted in Figure 1. In the remainder of this article, we discuss the sources of each of these classes of species, and how their absolute amounts vary in different regions of the troposphere.

Some aerosol constituents are relatively inert and nonvolatile, and their presence in the particulate phase does not depend on the composition of the atmosphere in which the particles are dispersed. Examples of such components are metals, like lead, and alkaline Earth elements, such as calcium and magnesium. The elemental composition of such materials in aerosols can be related directly to the sources of particles emitted into the atmosphere. Other aerosol species are more volatile, such that they are distributed between gaseous and aerosol (aqueous or solid) phases in the atmosphere, and this phase distribution depends on atmospheric conditions. Atmospheric acids and bases fall into this category, as do carbonaceous compounds with vapor pressures sufficiently low for the species to distribute themselves

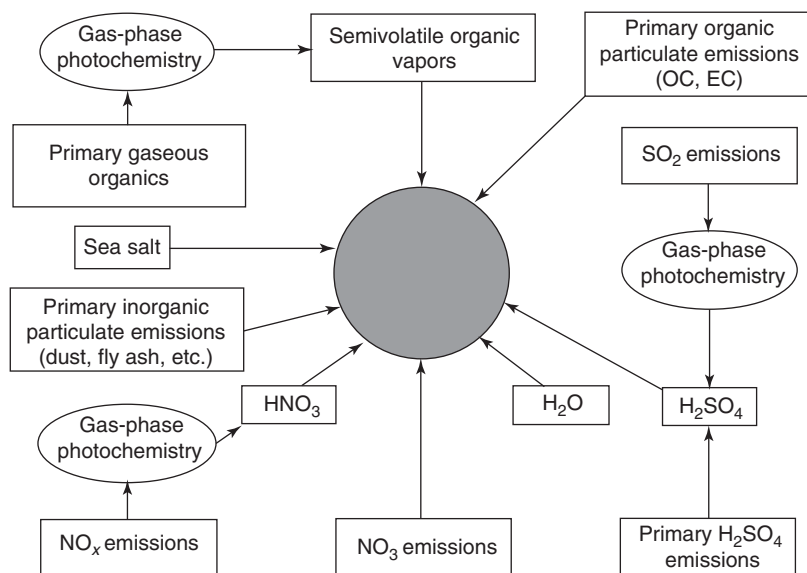


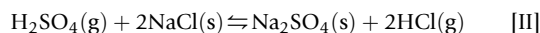
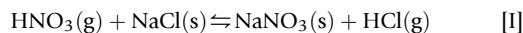
Figure 1 Chemical components of tropospheric aerosols and their routes of incorporation into the aerosol phase. EC, elemental carbon; OC, organic carbon.

between the gas and aerosol phases but not so low that the compounds lie entirely in the particulate phase. Compounds with vapor pressures in the range where they are distributed between the gas and aerosol phases are often referred to as semivolatile.

Inorganic Aerosols

Sulfur is a ubiquitous component of atmospheric aerosols. Whereas some particles are directly emitted, containing sulfur, the majority of the sulfur in atmospheric aerosols is the result of gas-to-particle conversion. The principal gaseous atmospheric precursor to aerosol sulfur is sulfuric acid (H_2SO_4), which is produced in the atmosphere from the oxidation of sulfur dioxide (SO_2). Sulfur dioxide has both natural and anthropogenic sources. The most important anthropogenic source of SO_2 is combustion of sulfur-containing fuels. A major natural source of SO_2 is the atmospheric oxidation of dimethyl sulfide (CH_3SCH_3) emitted from oceanic phytoplankton. Gaseous sulfur dioxide reacts with the hydroxyl (OH) radical through a series of steps to produce the H_2SO_4 vapor molecule. Sulfuric acid has an extremely low vapor pressure; consequently, it rapidly condenses on any available particles. In some instances, when sufficient aerosol surface area is not present to absorb the sulfuric acid, H_2SO_4 , in combination with water molecules, will spontaneously nucleate to form new ultrasmall particles. Sulfates have vapor pressures low enough for them to tend, when present, to reside overwhelmingly in the aerosol phase.

In addition to sulfuric acid, NH_3 , HNO_3 , and HCl are the major inorganic vapors involved in formation and growth of atmospheric aerosol. A principal source of chlorine in the atmosphere is sea salt particles. The NaCl in sea salt can react with sulfuric and nitric acids to produce gaseous HCl :



Gas-phase ammonia can react with acid vapors to form ammonium salts in the aerosol phase. Products include NH_4NO_3 , $(\text{NH}_4)_2\text{SO}_4$, NH_4HSO_4 , and NH_4Cl . The reversible reaction



is the principal source of nitrate in atmospheric aerosols. An analogous reaction is the equilibrium among NH_3 , HCl , and NH_4Cl :



When the relative humidity is sufficiently high, the particle phase becomes an aqueous solution, and the above salts dissociate into their corresponding ions: NH_4^+ , NO_3^- , H^+ , Na^+ , Cl^- , HSO_4^- , and SO_4^{2-} . These ions play an important role in governing the amount of liquid water in the aerosol particle at a given ambient relative humidity.

Deliquescence Behavior of Atmospheric Aerosols

With a single aqueous phase containing dissolved electrolytes, when compounds condense or evaporate, water must also

condense or evaporate to maintain the particle water activity equal to the ambient relative humidity. Thus, the water content of particles changes in response to condensation or evaporation of electrolytes, or in response to changes in ambient relative humidity itself (Figure 2). At a fixed electrolyte content, as the ambient relative humidity is lowered, water evaporates from the particle to maintain equilibrium. If the relative humidity is lowered sufficiently, a solid crystalline phase may form when one or more of the electrolytes reach saturation. For instance, if the particle contains sodium chloride, this solid phase will occur theoretically at 75% relative humidity. This defines the efflorescence point, which, according to equilibrium thermodynamics, is equivalent to the deliquescence point – the efflorescence point defines the relative humidity where the aqueous-to-crystalline phase transition occurs, whereas the deliquescence point defines that of the crystalline-to-aqueous transition. In general, the deliquescence point is a function of temperature, and this dependence can be derived from the Clausius–Clapeyron equation. The efflorescence and deliquescence points generally do not coincide, because nucleation of the new crystalline phase typically does not occur until the solution is substantially supersaturated (Figure 2).

For single-salt solutions, there is a single deliquescence point separating the higher relative humidities (where the particle is aqueous) from the lower relative humidities (where the particle is crystalline). Aqueous particles containing, for instance, sulfuric acid do not display a deliquescence point, because no solid phase of sulfuric acid exists for typical tropospheric temperature ranges.

If the particle contains multiple salts, or combinations of salts and acids, there is a range of relative humidities where aqueous and solid phases can coexist in equilibrium. Consider, for example, two salts in solution. As the relative humidity is lowered, eventually one of the salts becomes saturated and forms its crystalline phase. As the relative humidity continues

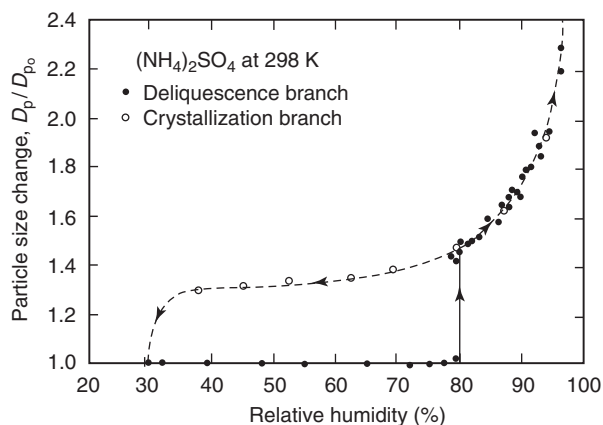


Figure 2 Deliquescence and crystallization behavior of atmospheric aerosols. Growth (filled circles) of an initially dry ammonium sulfate (namely, $(\text{NH}_4)_2\text{SO}_4$) particle as relative humidity (RH) is increased. The particle remains dry until about 80% RH, where it spontaneously takes on water (deliquesces). Subsequently, as the wet particle is exposed to decreasing RH (open circles), the particle remains wet below 80% RH and finally crystallizes at around 30% RH. This hysteresis behavior upon wetting and drying is typical of atmospheric inorganic salts.

to decrease, more of this salt precipitates, and the solution becomes simultaneously more concentrated in the non-precipitating salt. Eventually, the nonprecipitating salt reaches its saturation, and further decreases in relative humidity cause the particle to effloresce completely. If the particle contains a salt and an acid, the first deliquescence will occur, but the aqueous phase will continue to persist, owing to the absence of a deliquescence point for the acid. It can be shown that the deliquescence points of nonreacting mixtures are always lower than those of the single solutes.

Carbonaceous Aerosols

Carbonaceous material is a ubiquitous component of atmospheric aerosols. From combustion processes, carbon-containing compounds are emitted directly into the atmosphere in particulate form, a prime example of which is soot. In addition, during the gas-phase atmospheric oxidation of organic gases (by species such as OH, O₃, and NO₃), products are formed with oxygen-containing functional groups. These products tend to have lower vapor pressures than their parent organic molecules, and, as a result, can condense to the aerosol phase. Carbonaceous material incorporated into particles by this route is referred to as secondary organic aerosol. Natural hydrocarbons like the monoterpenes (a class of compounds having the chemical formula C₁₀H₁₆) lead to secondary organic aerosol when oxidized in the atmosphere by OH, O₃, and NO₃. In areas of high vegetation coverage, terpenes can contribute a substantial amount of organic aerosol.

The carbonaceous fraction of atmospheric aerosol is often divided according to elemental carbon (EC) and organic carbon (OC). Loosely speaking, EC refers to that fraction of the aerosol collected on a filter that cannot be vaporized upon heating. Chemically, EC is associated with sootlike compounds. By contrast, OC represents that fraction of the carbon-containing material that can be vaporized. Thus, the division of the carbonaceous material into EC and OC is essentially an operational one, defined on the basis of the technique used in measurement.

Degree of Chemical Mixing in Atmospheric Particles

All atmospheric particles of the same size do not have the same chemical composition. An 'externally mixed' aerosol population is defined as one in which different particles of the same size have different chemical compositions. In the extreme, an external mixture would consist only of particles containing individual pure substances, such as pure ammonium sulfate in one class of particles and pure EC in another class. Such complete segregation of substances does not occur. Rather, atmospheric aerosols consist of particles that initially have the chemical signature of their sources but, as they age in the atmosphere, take on other species from the gas phase or, possibly, coagulate with other particles.

Gas–Aerosol Equilibrium

The key thermodynamic properties of organic compounds that control their partitioning to condensed phases in the

atmosphere are the Henry's law constant (K_H), pure species vapor pressure (p^0), and, for substances that dissociate into ions in aqueous solutions, the dissociation constant (K_a). The atmospheric relative humidity and presence of other dissolved species in liquid aerosols also affect partitioning via the amount of condensed-phase water and the activity coefficient of the dissolved organic substance (in an aqueous or organic liquid phase). These relationships are briefly summarized here.

The solubility of an organic gas A in water is described by:



$$K_H = \frac{\gamma_A m_A}{p_A} \quad [1]$$

where K_H (mol kg⁻¹ atm⁻¹) is the Henry's law constant, m_A is the molality of A (moles of A per kilogram of liquid water), and p_A (atm) is the gas-phase partial pressure of A. The quantity γ_A is the activity coefficient of A in the aqueous phase and in an infinitely dilute aqueous solution has a value of unity. Aerosols can contain a large number of both inorganic and organic dissolved components, and γ_A is a function of the solution phase composition and the ambient relative humidity, which controls the concentrations of the dissolved components in the aqueous aerosol.

If the organic compound, for example, an organic acid RCOOH, dissociates significantly in aqueous solution,



then overall partitioning to the aerosol phase is enhanced, as the total molality of RCOOH in the aerosol phase is given by

$$m_{TOTAL} = m_{RCOOH} + m_{RCOO^-} \quad [2]$$

$$m_{TOTAL} = p_{RCOOH} K_H \left(\frac{1}{\gamma_{RCOOH}} + \frac{K_a}{m_H + \gamma_H \gamma_{RCOO}} \right) \quad [3]$$

where K_a (mol kg⁻¹) is the dissociation constant of the acid. It is apparent from eqn [3] that, in this case, partitioning is dependent upon the two equilibrium constants, aerosol pH (the hydrogen ion activity $m_{H^+} \gamma_H$) and the activity coefficients, γ_{RCOOH} and γ_{RCOO^-} .

Atmospheric condensed phases (clouds, fogs, aerosols) contain up to about 0.1 g water m⁻³ of atmosphere. It can be shown that, for an atmospheric gas to partition significantly into cloud water, the Henry's law constant (an 'effective' value, including the effects of any dissociation) must exceed 10⁴ mol kg⁻¹ atm⁻¹. This value is even greater if partitioning into aqueous aerosols (10⁻⁴–10⁻⁵ g water m⁻³) is considered.

Atmospheric Concentrations of Aerosols

The global tropospheric aerosol can be considered to consist of about five general categories of material: (1) sea salt; (2) soil dust; (3) inorganic salts (sulfate, nitrate, ammonium); (4) OC; and (5) EC. The relative proportions of each of these five classes vary over the globe. By sheer mass, sea salt has the largest flux into the atmosphere of any of the five categories, but the actual

mass of sea salt aerosol in the atmosphere at any time is actually less than that of soil dust. This is because sea salt aerosol does not rise vertically into the atmosphere very far above the ocean surface and falls back into the sea fairly readily. Soil dust, which has an estimated global emission rate about one-fifth that of sea salt, is actually estimated to have about twice the total mass loading in the atmosphere. This larger mass loading is a result of the fact that soil dust gets transported much higher in the atmosphere than sea salt and consequently has a much longer residence time in the atmosphere. Saharan dust is regularly measured in the air of Miami, Florida, for example. Inorganic salts, OC, and EC have principally anthropogenic sources, and these sources tend to be concentrated in the Northern Hemisphere. The average atmospheric residence time of these 'pollution' particles is on the order of a week. Given typical wind speeds, this means that such aerosols will tend to be located over and downwind of the heavily industrialized regions of the Northern Hemisphere. One exception to the overwhelming concentration of anthropogenic aerosols in the Northern Hemisphere is the carbonaceous aerosols emitted from burning of biomass, which occurs on a large scale in the Southern Hemisphere.

Concentrations of atmospheric aerosol components are usually reported in either of two ways. The first is the mass concentration, expressed in terms of micrograms of aerosol species per cubic meter of air, $\mu\text{g m}^{-3}$. The second is based on

the mole fraction of the species in air. A commonly used measure of the level of gaseous species in the atmosphere is mole fraction. For the trace atmospheric gases, important in atmospheric chemistry, mole fractions range from about 10^{-6} to 10^{-12} . These mole fractions are equivalent to the volume of species per volume of air. It has proved convenient to report such mole, or volume, fractions in terms of so-called mixing ratios, such as parts per billion (1 ppb: one part in 10^9 parts of air) or parts per trillion (1 ppt: one part in 10^{12} parts of air). Aerosol species concentrations can, likewise, be expressed in such mixing ratio units.

Figure 3 shows a distribution of aerosol chemical composition at many sites around the world, mainly in the Northern Hemisphere. These data were obtained with the Aerodyne Aerosol Mass Spectrometer. Aerosol composition is apportioned according to: sulfate, nitrate, ammonium, and organic compounds. A pie chart of composition is given for each site, and the mass concentration value below each pie chart is the annual average total mass concentration for that site. (The mass of water in the aerosol is not indicated, as water is evaporated from the particles in the course of measurement by the mass spectrometer.) Sites range from those in remote areas to those in major urban centers. The total aerosol mass concentration ratio between the most polluted cities and the most remote areas is as large as 70. Carbonaceous species (organics) constitute roughly one-half of the airborne particulate mass

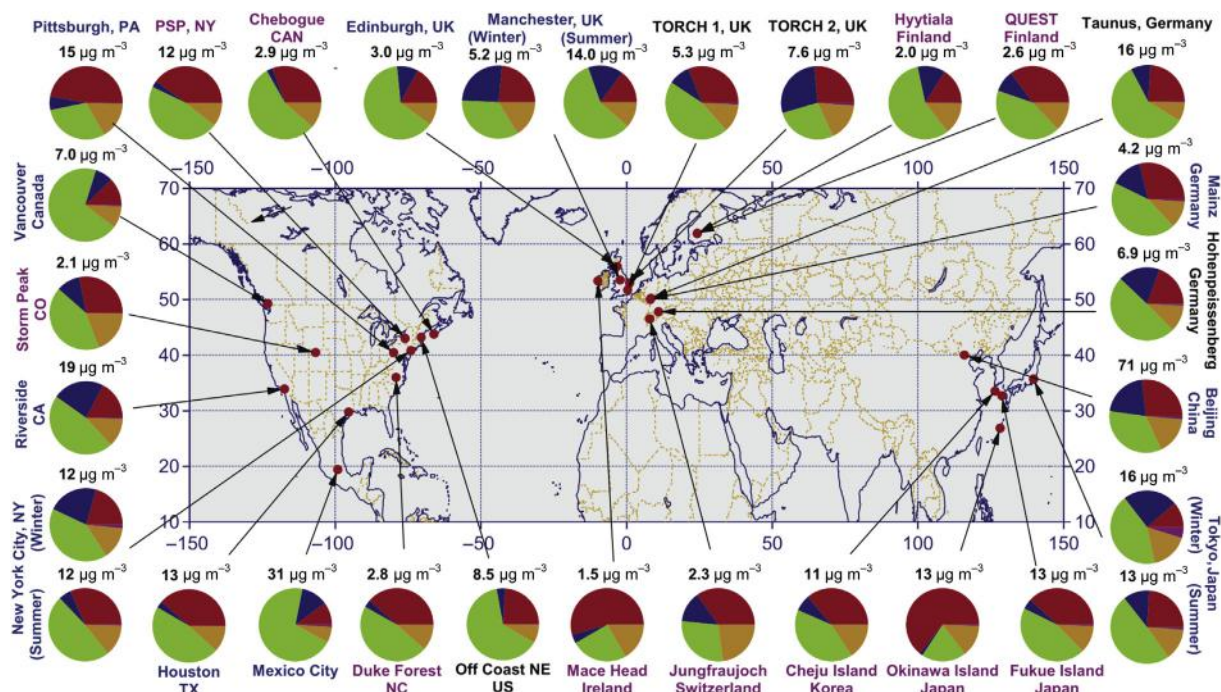


Figure 3 Aerodyne Aerosol Mass Spectrometer datasets of global aerosol chemical composition. Colors for the study labels indicate the type of sampling location: urban areas (blue), <150 km downwind of major cities (black), rural/remote areas >150 km downwind (pink). Pie charts show the average mass concentration and chemical composition: organics (green), sulfate (red), nitrate (blue), ammonium (orange), and chloride (purple) of nonrefractory particles of diameter less than $1 \mu\text{m}$. Adapted from Zhang, Q., Jimenez, J.L., Canagaratna, M.R., Allan, J.D., Coe, H., Ulbrich, I., Alfarra, M.R., Takami, A., Middlebrook, A.M., Sun, Y.L., Dzepina, K., Dunlea, E., Docherty, K., DeCarlo, P.F., Salcedo, D., Onasch, T., Jayne, J.T., Miyoshi, T., Shimojo, A., Hatakeyama, S., Takegawa, N., Kondo, Y., Schneider, J., Drewnick, F., Borrmann, S., Weimer, S., Demerjian, K., Williams, P., Bower, K., Bahreini, R., Cottrell, L., Griffin, R.J., Rautiainen, J., Sun, J.Y., Zhang, Y.M., Worsnop, D.R., 2007. Ubiquity and dominance of oxygenated species in organic aerosols in anthropogenically influenced Northern Hemisphere midlatitudes. *Geophysical Research Letters* 34, L13801. <http://dx.doi.org/10.1029/2007GL029979>.

worldwide. As depicted schematically in Figure 1, carbonaceous species are both emitted directly in the particle phase as well as transferred from the gas phase to the aerosol phase in the atmosphere itself as a result of gas-phase chemical reactions that generate less volatile species.

Organic aerosol comprises a major fraction (18–70%; average = 45%) of the nonrefractory submicron particle mass at 37 global locations (Figure 3), while sulfate (10–67%; average = 32%), nitrate (1.2–28%; average = 10%), ammonium (6.9–19%; average = 13%), and chloride (<Detection limit–4.8%; average 0.6%) account for the rest of the particle mass. Analyses of the 37 field studies in urban and anthropogenically influenced rural/remote area shown in Figure 3 indicate the ubiquity and dominance of oxygenated species in organic aerosols, of which a major fraction is likely to be secondary in nature.

Ranges of the mass of total fine particle (roughly those particles with diameters less than 2 μm) in the troposphere at ground level are:

Remote	<5 $\mu\text{g m}^{-3}$
Nonurban continental	15 $\mu\text{g m}^{-3}$
Urban	50 $\mu\text{g m}^{-3}$
Urban (heavily polluted)	>100 $\mu\text{g m}^{-3}$

As expected, highest global concentrations of atmospheric aerosols are found in large urban areas, where total mass concentration levels can routinely exceed 100 $\mu\text{g m}^{-3}$.

Conclusion

Atmospheric aerosols contain the chemical signature of the sources of direct particle emissions into the atmosphere as

well as that of the conversion of gaseous molecules into particulate-phase species. The broadest classification of the chemical components of atmospheric aerosols is into inorganic species (salts, metals), carbonaceous compounds, and water. Many aerosol species are nonvolatile; once deposited in the aerosol phase, they remain there until the particle is eventually removed from the atmosphere. Other species, such as some ammonium salts and a variety of organic compounds, distribute themselves between the gas and aerosol phases in accordance with local atmospheric conditions. Not all particles of the same size at the same location have the same chemical composition, as particles arise from different sources and have different histories in the atmosphere. Particle mass concentrations vary over the globe, from the order of 1 $\mu\text{g m}^{-3}$ in the cleanest air masses to more than 100 $\mu\text{g m}^{-3}$ in polluted urban areas. Understanding the dynamics of the chemical composition of the atmospheric aerosol remains one of the challenges of atmospheric science.

See also: **Aerosols:** Observations and Measurements. **Chemistry of the Atmosphere:** Observations for Chemistry (*In Situ*): Particles. **Tropospheric Chemistry and Composition:** Cloud Chemistry.

Further Reading

Seinfeld, J.H., Pandis, S.N., 2006. *Atmospheric Chemistry and Physics: From Air Pollution to Climate Change*, second ed. Wiley, New York. (1998).

Singh, H.B. (Ed.), 1995. *Composition, Chemistry, and Climate of the Atmosphere*. Van Nostrand Reinhold, New York.

Aliphatic Hydrocarbons

J Rudolph, York University, Toronto, ON, Canada

O Stein, IEK 8: Troposphere, Research Center Juelich, Juelich, Germany

© 2015 Elsevier Ltd. All rights reserved.

Nomenclature

NMHC Nonmethane hydrocarbons, all hydrocarbons except methane.

VOC Volatile organic compound.

Isoprenoides Compounds derived from and including isoprene and terpenes.

Tg 10^{12} g.

yr year.

Tg yr⁻¹ Teragram per year (unit for emission rates).

OH radical Hydroxyl radical.

NO_x Sum of NO and NO₂.

NO₃ Nitrate radical.

K Kelvin, units for absolute temperature.

hPa Hecto Pascal, unit for pressure.

OH_kalkane Rate constant for reaction of an alkane with OH-radicals.

OH_kalkene Rate constant for reaction of an alkene with OH-radicals.

O₃kalkene Rate constant for reaction of an alkene with ozone.

OH_k Rate constant for reaction of a VOC with OH-radicals.

NO₃k Rate constant for reaction of a VOC with NO₃-radicals.

Cl_k Rate constant for reaction of a VOC with Cl-atoms.

[OH] OH-radical concentration.

τ_{alkane} Atmospheric lifetime of an alkane.

τ_{alkene} Atmospheric lifetime of an alkene.

Synopsis

During the last decades it has been recognized that organic trace gases other than methane play an important role in the chemistry of the troposphere. Aliphatic nonmethane hydrocarbons (NMHCs) are one of the major groups among the wide range of compounds that constitute nonmethane volatile organic compounds (VOCs). Although the total global emission rate of aliphatic NMHC is only a small fraction of all VOC emissions, the uneven spatial and temporal distribution of aliphatic hydrocarbon emissions combined with their high reactivity makes them important contributors to atmospheric chemical reactions in areas with strong emissions such as urban and industrialized regions or areas with strong biomass burning activities. Aliphatic hydrocarbons are also valuable tracers for identifying important atmospheric processes and trace gas sources.

Introduction

Aliphatic hydrocarbons, similar to other volatile organic compounds (VOCs) and carbon monoxide, are major players in a variety of important chemical processes in the atmosphere. They react with hydroxyl (OH) radicals and thus influence the atmospheric balance of reactive radicals and thus the self-cleansing properties of the atmosphere. Their atmospheric oxidation results in the formation of a number of important secondary photooxidants, the arguably most important one being ozone. Other important secondary photooxidants derived from the atmospheric oxidation of aliphatic hydrocarbons include aldehydes, peroxides, ketones, and a wide range of organic nitrogen compounds including peroxyacetyl nitrate. The oxidation of aliphatic hydrocarbons also contributes to formation of secondary organic aerosols (SOAs).

Strictly speaking, the definition of aliphatic hydrocarbons includes a very wide variety of and an extremely large number of compounds. However, for the practical purpose of this article the range of substances will be limited to those aliphatic hydrocarbons that are most abundant in the troposphere or for some other reason of specific interest for the chemistry of

the atmosphere. Furthermore, some specific groups of hydrocarbons such as methane and isoprenoides are treated in separate articles and will therefore be excluded, although they chemically belong to the category of aliphatic hydrocarbons. For the same reason, aliphatic hydrocarbons that are primarily attached to atmospheric particulate matter will not be discussed here. Here the term aliphatic nonmethane hydrocarbon (aliphatic NMHC) will be used for the thus defined group of compounds. Even with these limitations, the number of compounds that fall into the thus defined category of aliphatic NMHC is potentially very large.

Nearly all aliphatic NMHCs that are relevant for the chemistry of the troposphere can be combined into two chemical groups, alkanes and alkenes. The only relevant exception is ethyne, the only alkyne that is of some significance for the chemistry of the troposphere. Alkanes that are relevant have seldom more than eight or nine carbon atoms; the important alkenes generally have less than five carbon atoms. This substantially reduces the number of aliphatic hydrocarbons that are relevant for the chemistry of the troposphere. Furthermore, there are many parallels between different compounds and the main features of the tropospheric

chemistry and distribution of aliphatic hydrocarbons can be derived from a few general principles.

With very few exceptions the atmospheric residence time of aliphatic hydrocarbons is a few days or less, a timescale that does not allow significant large-scale transport. Therefore, aliphatic hydrocarbons have very little direct relevance for the chemistry of the stratosphere and upper troposphere or regions remote from major sources of aliphatic hydrocarbons. This article will discuss the sources, sinks, chemical transformations, and distribution of aliphatic hydrocarbons in the troposphere.

Sources of Atmospheric Aliphatic Hydrocarbons

Overview

There are several global and regional emission inventories available on the Internet. While they are reasonably consistent in the main features such as global totals or global emission patterns, they differ in many details and also present information with varying levels of detail. Moreover, many of these inventories are designed with the use of numerical model simulations of air pollution in mind and therefore often combine different chemical species into lumped categories. Furthermore, establishing emission inventories typically is based on a significant amount of extrapolation from limited sets of observations. Consequently there are substantial uncertainties for total emission rates, speciation, geographical distribution, trends, and seasonal cycles. There is evidence, mainly based on ambient observations that several current inventories tend to underestimate anthropogenic emission rates. This should be kept in mind when using the overview over aliphatic NMHC emissions presented in this article.

For practical and regulatory reasons emission inventories often identify sources of atmospheric trace gases by the type of activity that results in emission, for example domestic heating, transportation, or electricity generation. Sometimes sources are also characterized by the type of material used as fuel or feedstock, for example natural gas, oil, coal, and wood. A further differentiation can be based on distinction between emissions from production, processing, distribution, storage, and usage. [Table 1](#) lists the most important types of sources for atmospheric aliphatic hydrocarbons and some of their subcategories.

Table 1 Sources of atmospheric aliphatic NMHC and their subcategories

Category	Subcategories
Biomass burning	Wildfires, prescribed fires, deforestation, agricultural waste burning, biofuels
Fossil fuel use	Road transportation (exhaust and fugitive emissions), nonroad transportation, fuel production, storage, distribution, solvent production and usage, industrial losses, electricity generation, domestic heating
Biological production	Vegetation, soil emissions, oceanic emissions
Geothermal vents	
Solid waste disposal	

The composition of emissions is determined by the type of process that results in emission as well as the chemical composition of the parent material. Emissions resulting from evaporation or leakage of material containing aliphatic hydrocarbons typically have a composition closely related to that of the parent material. In the case of partial evaporation, often fractionation, depending on vapor pressure of the individual component, occurs. Emissions resulting from incomplete combustion are highly dependent on type of fuel and the efficiency of the combustion process. Reduction of emissions of hydrocarbons from combustion processes is typically achieved by increased combustion efficiency, sometimes by adding a second oxidation step such as catalytic converters in cars. Differences in fuel and combustion process influence not only the total VOC emissions, but also have substantial impact on the composition of the emissions. [Figure 1](#) gives examples of differences in relative composition between some of the major sources of atmospheric aliphatic hydrocarbons.

Emission Rates and Source Distributions

Global average emissions of aliphatic NMHC are presently in the range of 90 to 150 Tg yr⁻¹ with a best estimate of approximately 120 Tg yr⁻¹. [Figure 2](#) gives an estimate of the average contributions of individual NMHC to total emissions. About 40–50% of these emissions are connected to fossil fuel usage, including production, storage, and distribution. Biomass burning contributes between 20 and 30%. However, only part of biomass burning emissions is from wildfires, significant emission also results from various agricultural practices such as burning of agricultural waste, controlled forest burning and use of biofuels, and therefore controlled by human activities. Biogenic contributions to aliphatic NMHC in the atmosphere are emissions from vegetation and oceans. It should be noted that emission rates of aliphatic NMHC from vegetation are small compared to the emission rates for isoprene and terpenoids, but nevertheless contribute to the atmospheric budget of aliphatic NMHC. Oceanic emissions of light alkenes and alkanes are in the range of a few teragram per year, a small fraction of the total global aliphatic NMHC emissions and thus of minor importance for the total budget. However, due to the absence of other sources over remote oceans regions, these emissions can impact the chemistry of the remote oceanic atmosphere. [Table 2](#) lists global average emission rates of light aliphatic NMHC for different source categories. It has been shown recently that geothermal vents are a source for light alkanes, but the magnitude of this source is highly uncertain.

With very few exceptions, source strength estimates for aliphatic NMHC are based on a so-called bottom-up approach where estimates are based on the evaluation of sources. For a few aliphatic NMHC so-called top-down estimates have been made. In these estimates, the emission rates are based on budget considerations and calculation of loss rates. This requires knowledge of representative atmospheric distributions of the individual NMHC, which exists only for the least reactive NMHCs such as ethane and propane. Source estimates derived from such top-down evaluations are, as far as available, included in [Table 2](#).

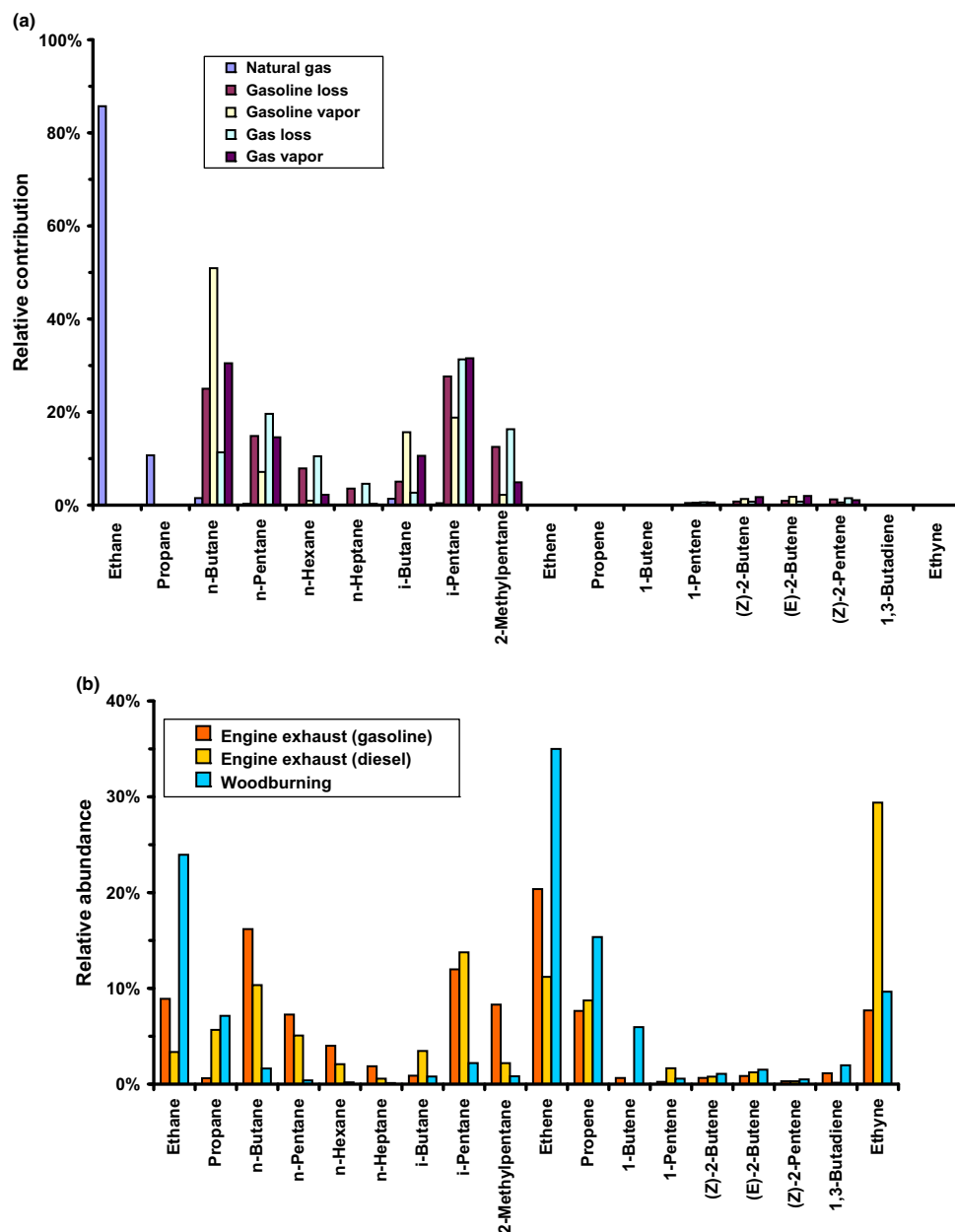


Figure 1 Comparison of the relative composition of different source categories. ‘Gas vapor’ refers to losses of liquified petroleum gas during distribution and usage. The individual emissions are normalized to the sum of all identified aliphatic NMHC emitted from the specific source category. Data are compiled from McLaren, R., Singleton, D.L., 1996. Analysis of motor vehicle sources and their contribution to ambient hydrocarbon distributions at urban sites in Toronto during the Southern Ontario Oxidants Study. *Atmospheric Environment* 30(12), 2219–2232; the GEIA/EDGAR emission database; Lobert, J.M., Scharffe, D.H., Hao, W.M., et al., 1991. Experimental evaluation of biomass burning emissions: nitrogen and carbon containing compounds. In: Levine, J. (Ed.), *Global Biomass Burning: Atmospheric, Climatic, and Biospheric Implications*. MIT Press, Cambridge, MA, London, GB, 290–304; von Czapiewski, K., 1999. Ph.D. Thesis, Universität zu Köln.

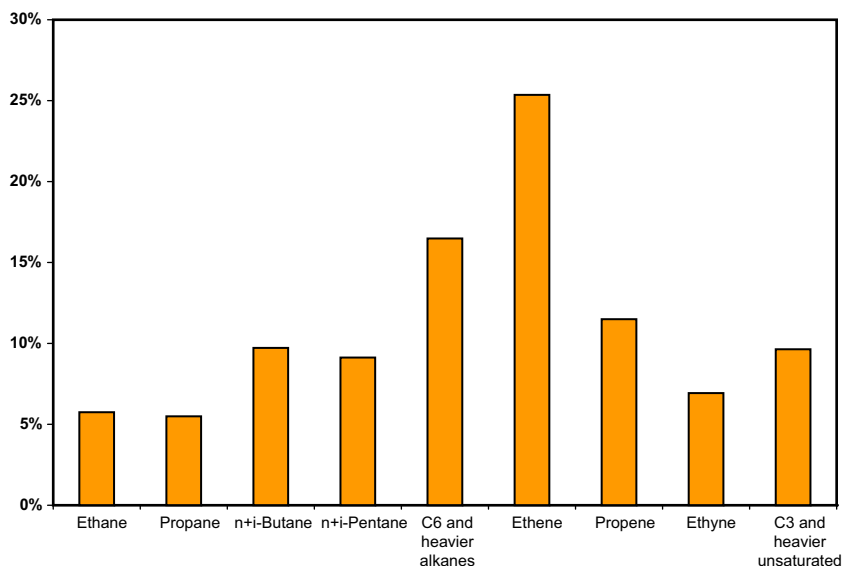


Figure 2 Average relative contribution of individual compounds to total global aliphatic NMHC emissions. Data are taken from: GEIA/EDGAR emission database; Ehhalt, D.H., 1999. Gas phase chemistry of the troposphere. In: Zellner, R. (Ed.), Global Aspects of Atmospheric Chemistry. Springer, New York, pp. 21–109; Middleton, P., 1995. Sources of air pollutants. In: Singh, H.B. (Ed.), Composition, Chemistry, and Climate of the Atmosphere. Van Nostrand Reinhold, London, pp. 88–119.

Table 2 Annual emission rates for light aliphatic NMHC for different source categories based on the MACC emission inventory

Species	Sector	Emission rate Tg yr ⁻¹
C ₂ H ₄	Anthropogenic	7.69
	Biogenic	16.6
	Ocean	1.40
	Biomass burning	4.30
	Total	30.0
C ₃ H ₆	Anthropogenic	3.50
	Biogenic	6.07
	Ocean	1.52
	Biomass burning	2.48
	Total	13.6
C ₂ H ₆	Anthropogenic	3.38
	Biogenic	0.14
	Ocean	0.98
	Biomass burning	2.26
	Total	6.76
C ₃ H ₈	Atmospheric loss rate	10–13 ^{a,b,c}
	Anthropogenic	3.99
	Biogenic	0.02
	Ocean	1.29
	Biomass burning	1.25
	Total	6.55
	Atmospheric loss rate	8.4 ^b

^aRudolph, J., 1995. The tropospheric distribution and budget of ethane. Journal of Geophysical Research 100, 11369–11381.

^bGupta, M.L., Cicerone, R.J., Blake, D.R., Rowland, F.S., Isakson, I.S.A., 1998. Global atmospheric distribution and source strength of light hydrocarbons and tetrachloroethene. Journal of Geophysical Research Atmosphere 103, 2829–2835.

^cXiao, I., Jacob, D.J., Hudman, R.C., et al., 2008. Global budget of ethane and regional constraints on U.S. sources. Journal of Geophysical Research Atmosphere 113, D21306. <http://dx.doi.org/10.1029/2007JD009415>.

There are no chemical mechanisms that result in the formation of aliphatic NMHC from other atmospheric trace gases and, with the exception of minor emissions from airplanes, all sources of aliphatic NMHC are located on the ground. Consequently, nearly all emissions occur in the lower levels of the atmospheric boundary layer. High-emission rates of aliphatic NMHC are observed for densely populated regions, major industrialized areas and regions with intense fossil fuel production. Emissions from biomass burning and biological sources are often spread over large areas. Emission of saturated NMHC is mainly associated with fossil fuel production, unsaturated NMHC mainly with incomplete combustion processes, including biomass burning, and areas of dense vegetation (Figure 3).

Temporal Changes in Emission Rates

Emission rates of aliphatic NMHC are tied to a number of factors that change with time. Periodic changes of emissions can be driven by environmental factors such as temperature or human activity patterns. Emissions from transportation peak during morning and afternoon rush hours, fuel and solvent evaporation rates depend on ambient temperature and biogenic emissions can depend on temperature and photosynthesis rate. In some cases, weekly cycles in emissions are created by differences in human activities between weekdays and weekends. Annual cycles are driven directly and indirectly by ambient conditions. Emissions from domestic heating peak during cold seasons, biogenic emission and biomass burning depend on growing cycles as well as temperature, humidity, and rainfall. In most parts of the Northern Hemisphere

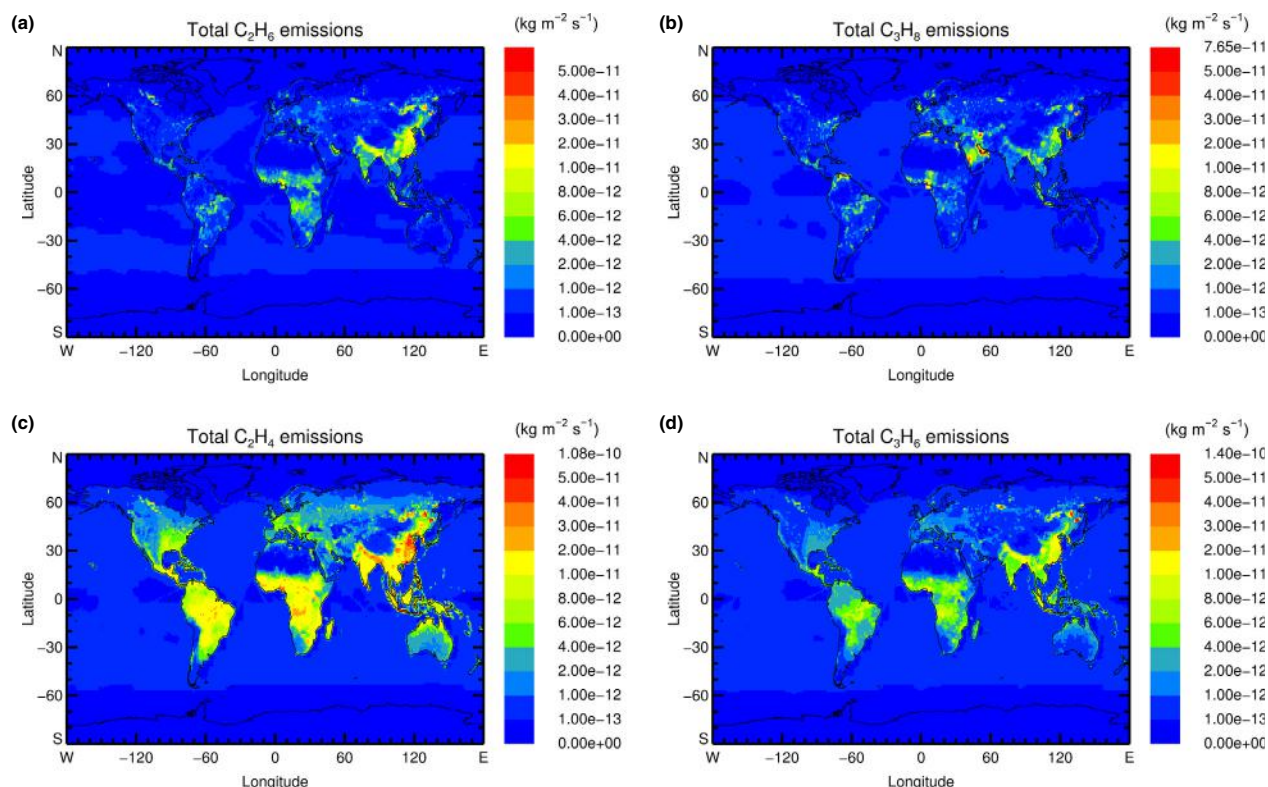


Figure 3 Global distribution of the total annual average of emission rates of several light aliphatic NMHC. Data are taken from the Granier, C., Bessagnet, B., Bond, T. et al., 2011. Evolution of anthropogenic and biomass burning emissions of air pollutants at global and regional scales during the 1980–2010 period. *Climate Change* 109, 163–190. doi:10.1007/s10584-011-0154-1.

biomass burning emissions from wildfires and prescribed fires peak between July and October, but in some areas such as Northern Hemisphere Africa and southern Asia the occurrence of dry seasons and agricultural practices cause peak emissions around December and January. The change in global distribution of ethene emission rates shown in **Figure 4** is predominantly the result of seasonal variations of biomass burning and biological activity.

Long-term trends in emission are caused by population growth as well as industrial and technological development. Varying intensity in forest clearing, changes in agricultural practices, and control strategies for wildfires also contribute to secular trends in emission rates. There is no doubt that aliphatic NMHC emissions increased very substantially between the beginning of the industrial age and the last decades of the twentieth century. However, during the last decades the overall

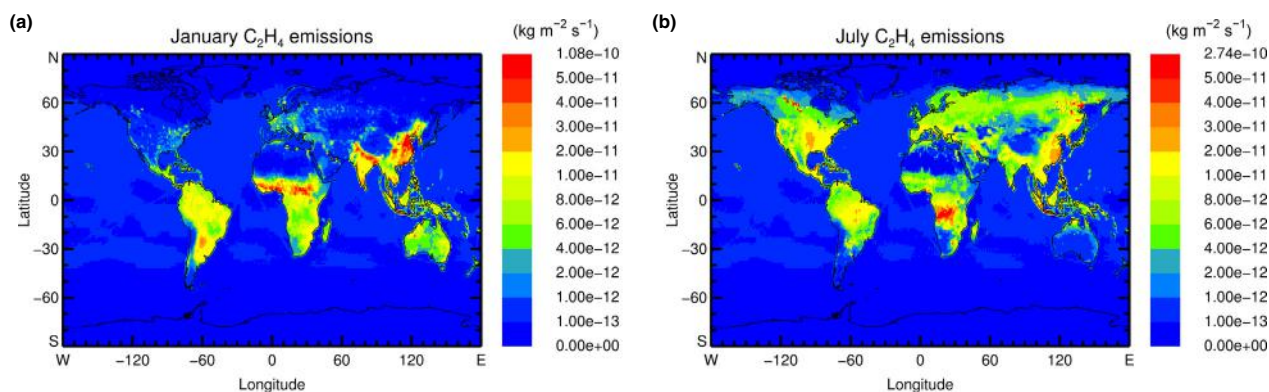


Figure 4 Seasonal change of the global distribution of ethene emission rates. Data are taken from Granier, C., Bessagnet, B., Bond, T., et al., 2011. Evolution of anthropogenic and biomass burning emissions of air pollutants at global and regional scales during the 1980–2010 period. *Climate Change* 109, 163–190. doi:10.1007/s10584-011-0154-1.

trend is no longer obvious. Factors contributing to increased emissions, such as worldwide increase in population, increase in fossil fuel consumption rates, increased worldwide industrialization, and rapid economic developments in India and China and other developing countries are to some extent compensated by investments in emission reduction technologies and improvements in efficiency of fossil fuel usage driven by increasing costs of oil and natural gas. Other factors contributing are the drastic economic and ecological changes in the area of the former Soviet Union and other state-directed economies and efforts to limit greenhouse gas emissions, which also impact emissions of aliphatic NMHC. Overall artificial aliphatic NMHC emissions peaked around 1990 with indication for a slight decrease early in the twenty-first century (Figure 5).

Emission rates from different source categories show varying temporal trends as a result of regulatory measures, development in technology as well as ecological and economical considerations. For example, emission from fossil fuel production increased during the last two decades, while at the same time transportation-related emissions significantly decreased. It is important to realize that the relatively moderate changes in global emission rates of artificial aliphatic NMHC are the result of substantially different trends in different countries. For most developed nations, emissions decreased significantly during the last two decades (Figure 6(a)). This is primarily due to reductions in emissions from road transport (Figure 6(b)). In exceptional cases emissions from developed countries increased substantially as result of emissions from increasing fossil fuel production outweighing reductions in emissions from road transport (Figure 6(b) and 6(c)). For developing countries often changes in emissions from several source categories contribute to overall increasing emission rates. It should be noted that for countries with very moderate changes in overall emission rates, this can be the result of compensating changes in emissions from different source

types. One of the potential consequences is that, even in case of little change in total emissions, speciation and spatial distribution change as can be seen in the change of the global distribution of anthropogenic ethene emissions between 1990 and 2008 (Figure 7). While there are large reductions in emissions from Europe and North America, due to the emission increases in Africa, India, and southern Asia, the change in total global anthropogenic ethene emissions is only about 10%, which is well below the uncertainty of the emission rate estimates.

Most changes in emission rates are gradual, resulting in smooth trends over several years, sometimes decades. A notable exception is emissions from countries with former state-directed economies. In these countries rapid and drastic economic and political changes resulted in large changes of emissions over timescales of less than a decade around 1990. The only source category with frequent large interannual variations in emission rates is biomass burning, especially emissions from wildfires. This is the result of substantial interannual variations in temperature, humidity, and rainfall distributions.

There are several emerging technologies such as mining of oil sands, hydraulic fracturing for natural gas mining, and nontraditional usage of biological materials as fuels, which result in emissions of aliphatic NMHC. Most likely these sources will become significant in the future. However, at present these emissions are not sufficiently well characterized to allow meaningful quantitative predictions.

Atmospheric Reactions

Tropospheric Removal

The only relevant removal of aliphatic NMHC in the troposphere occurs by gas phase reactions, predominantly initiated

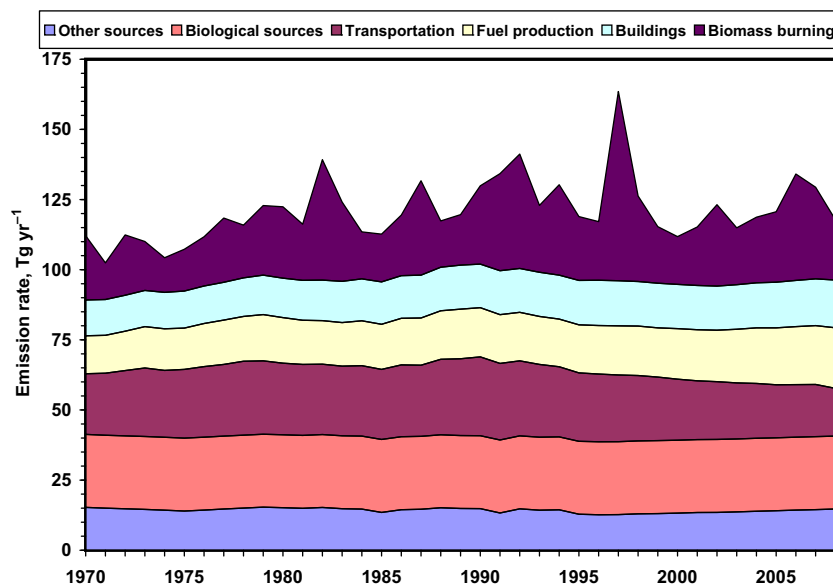


Figure 5 Trend of global annual average emission rates for aliphatic NMHC. 'Other' includes minor anthropogenic emissions such as solvent use as well as natural sources with the exception of wild fires, which are included in biomass burning. Data are taken from the EDGAR emission inventory.

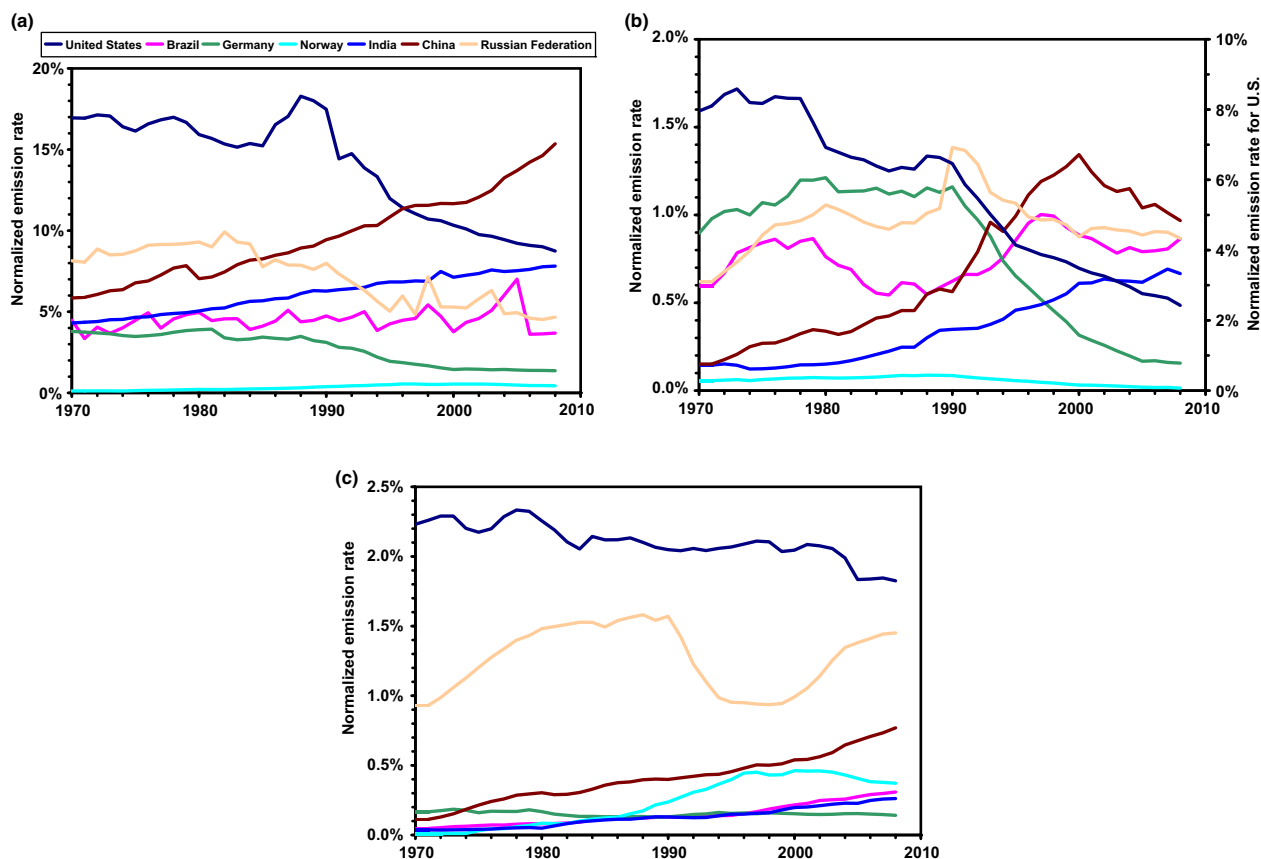


Figure 6 Example for temporal development of aliphatic NMHC emissions for selected countries. Shown are the changes of total annual emissions (a) as well as for two major source categories, road transportation (b), and fugitive emissions (c). The emissions are given as percentage of the total aliphatic NMHC emissions in 1970. Please note that emissions from road transportation for the United States are on a secondary axis. Data are taken from the EDGAR emission inventory.

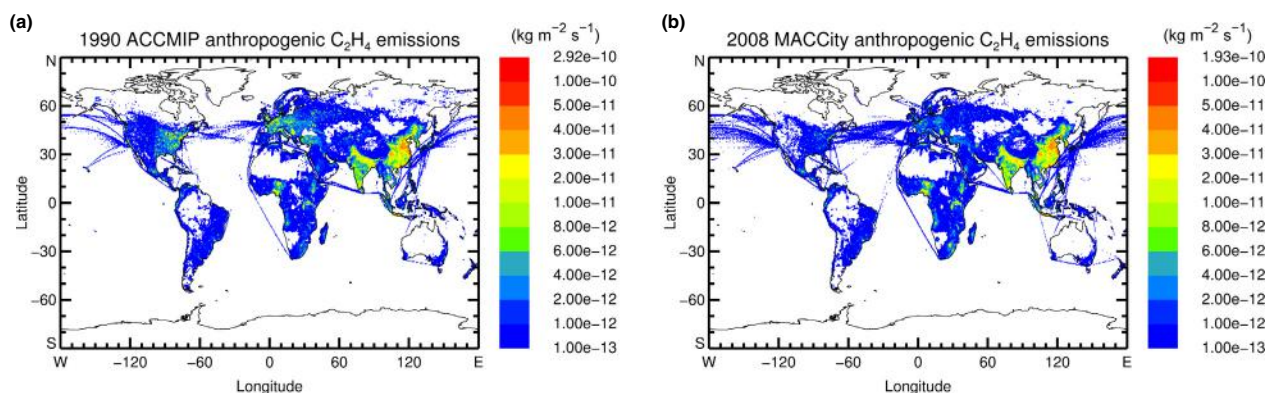


Figure 7 Change in global distribution of ethene emissions for anthropogenic sources between 1990 and 2008. Data are taken from the MACCity emission inventory.

by OH radicals. For alkanes and ethyne the reaction with OH radicals is on global average the dominant loss mechanism. Alkenes react also with ozone, although this reaction generally is less important than removal by OH radicals. To a minor extent the reaction with chlorine (Cl) atoms, bromine (Br) atoms, and the nitrate (NO_3) radicals contribute to tropospheric removal of aliphatic NMHC. Although these latter

reactions are on global average of minor importance they can be highly relevant under specific conditions.

In the lower polar troposphere during spring Cl atom concentrations frequently cause a clearly visible depletion in the concentration of nearly all aliphatic NMHCs. Similarly, under these conditions Br atom reactions contribute significantly to the removal of ethyne and light alkenes. Reactions with halogen

atoms can also have significance for loss processes of aliphatic NMHC in coastal regions heavily impacted by industrial emissions. Due to the extremely low concentrations of aliphatic NMHC in the stratosphere, stratospheric reactions of aliphatic NMHC contribute only marginally to the chemistry of the stratosphere and have very little impact on the overall budget of aliphatic NMHC.

For most aliphatic NMHCs, the rate constant for reaction with the nitrate radical is too slow to be of relevance for the troposphere. However, at high nitrate radical concentrations, which can occur at night in regions with substantial sources of nitrogen oxides (NO_x), this reaction can be an important loss mechanism for alkenes. The reactivity of alkenes toward the nitrate radical increases drastically with increasing alkyl substitution at the double bond and for these highly reactive alkenes' reaction with the nitrate radical can be an efficient loss process at night. In general, this is less important for the chemistry of the alkenes treated here, but can be highly relevant for reactions of isoprene, terpenes, and related compounds of predominantly biogenic origin.

The rate constants for atmospheric reactions of aliphatic NMHC depend on the type of reaction partner as well as on the chemical structure of the NMHC. These rate constants range from effectively zero to the collision-frequency limit. With very few exceptions, the relevant rate constants for reactions of aliphatic NMHC in the troposphere have been subject to detailed laboratory studies and are well understood. There are several algorithms that allow predictions of reaction rate constants based on structure-reactivity relations. Table 3

presents examples of rate constants for reaction of some of the most important aliphatic NMHCs with OH radicals, ozone, chlorine atoms, and the nitrate radical. The rate constants are given for standard conditions; that is a temperature of 298 K and a pressure of 1000 hPa. Most of these are temperature dependent. Since the temperature in the troposphere varies substantially, this has to be considered when using rate constants for calculations under nonstandard conditions. With extremely few exceptions the rate constants do not depend on pressure for the pressure range that is relevant for the troposphere.

As mentioned above, tropospheric removal of alkanes is nearly entirely due to reaction with OH radicals. Thus their atmospheric lifetime, τ_{alkane} , can be calculated from the rate constant for reaction with OH radicals, $^{\text{OH}}k_{\text{alkane}}$, and the OH-radical concentration, $[\text{OH}]$:

$$\tau_{\text{alkane}} = (^{\text{OH}}k_{\text{alkane}} * [\text{OH}])^{-1} \quad [1]$$

In the case of alkenes, the reaction with ozone also has to be considered and we obtain:

$$\tau_{\text{alkene}} = (^{\text{OH}}k_{\text{alkene}} * [\text{OH}] + ^{\text{O}_3}k_{\text{alkene}} * [\text{O}_3])^{-1} \quad [2]$$

The rate constants for reactions of individual aliphatic NMHC range over several orders of magnitude depending on chemical structure. Consequently atmospheric lifetimes of aliphatic NMHC cover a wide range, from several months for ethane to less than 1 h for some of the most reactive alkenes. Figure 8 gives an overview of the average atmospheric lifetime for some of the most abundant aliphatic NMHCs.

Table 3 Rate constants for the reaction of some alkanes and alkenes with OH radicals, ozone, chlorine atoms, and NO_3 radical at 298 K and 1 atm total pressure of air

Compound	OH $^{\text{OH}}k$ (10^{-12} cm^3 $\text{molecule}^{-1} \text{ s}^{-1}$)	Ozone $^{\text{O}_3}k$ (10^{-17} cm^3 $\text{molecule}^{-1} \text{ s}^{-1}$)	Chlorine $^{\text{Cl}}k$ (10^{-11} cm^3 $\text{molecule}^{-1} \text{ s}^{-1}$)	NO_3 $^{\text{NO}_3}k$ (cm^3 $\text{molecule}^{-1} \text{ s}^{-1}$)
Ethane	0.25		5.9	$1.4 * 10^{-18}$
Propane	1.1		14	$1.7 * 10^{-17}$
n-Butane	2.4		22	$4.6 * 10^{-17}$
i-Butane	2.2		14	$1.1 * 10^{-16}$
n-Pentane	4.0		28	$8.7 * 10^{-17}$
i-Pentane	3.7		22	$1.6 * 10^{-16}$
n-Hexane	5.5		34	$1.1 * 10^{-16}$
n-Heptane	7.0		39	$1.5 * 10^{-16}$
Cyclohexane	7.2		35	$1.4 * 10^{-16}$
Ethene	8.5	0.16	9.9	$2.1 * 10^{-16}$
Propene	26	1.0	23	$9.5 * 10^{-15}$
1-Butene	31	0.96	22	$1.4 * 10^{-14}$
2-Methylpropene	51	1.1		$3.3 * 10^{-13}$
1-Pentene	31	1.0		
2-Methyl-2-butene	86	40		$9.4 * 10^{-12}$
2,3-Dimethyl-2-butene	110	113		$5.7 * 10^{-11}$
1,3-Butadiene	66	0.63	42	$1.0 * 10^{-13}$
Ethyne	0.9		7.4	

Taken from Atkinson, R., 1994. Gas phase tropospheric chemistry of organic compounds. Journal of Physical Chemistry Reference Data Monographs 2, 11–216; Atkinson, R., 1997. Gas-phase tropospheric chemistry of volatile organic compounds: 1. Alkanes and alkenes. Journal of Physical Chemistry Reference Data 26, 215–290; Atkinson, R., Baulch, D.L., Cox, R.A., Hampson Jr., R.F., Kerr, J.A., Rossi, M.J., Troe, J., 1997. Evaluated Kinetic and photochemical data for atmospheric chemistry: supplement VI IUPAC Subcommittee on gas kinetic data evaluation for atmospheric chemistry. Journal of Physical Chemistry Reference Data 26, 521–1011; Stutz, J., Ezell, M.J., Ezell, A.A., Finlayson-Pitts, B.J., 1998. Rate constant and kinetic isotope effect in the reaction of atomic chlorine with n-butane and simple alkenes at room temperature. Journal of Physical Chemistry 102, 8510–8519.

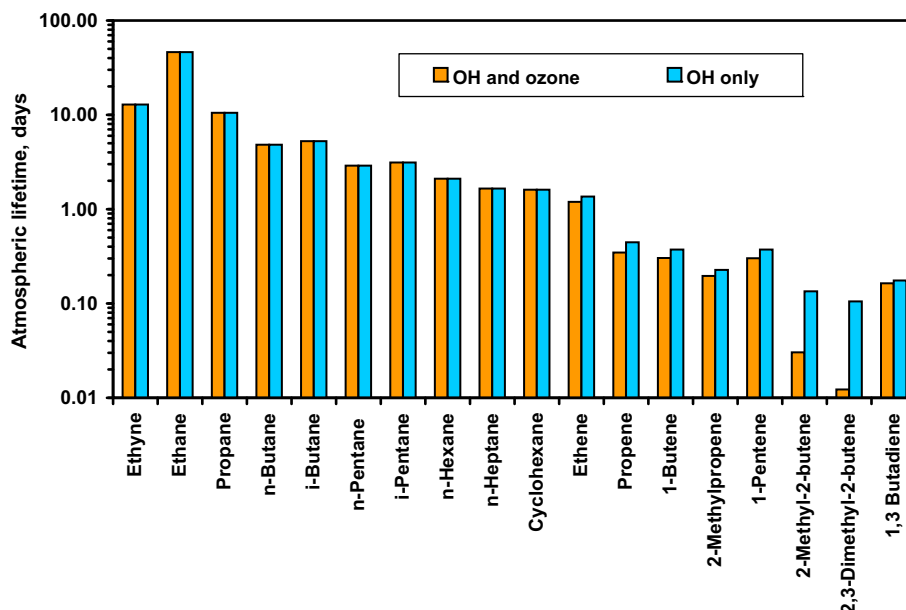


Figure 8 Atmospheric lifetimes of selected aliphatic NMHC. The calculations are based on rate constants from Table 3, average OH-radical concentration of 1×10^6 radicals cm^{-3} , and average ozone volume mixing ratio of 30 nmol mol^{-1} .

The concentrations of OH radicals and ozone in the troposphere exhibit a very high spatial and temporal variability. Furthermore, many of the rate constants for reaction of hydrocarbons with OH radicals or ozone have significant temperature dependence. Therefore, the atmospheric lifetimes of aliphatic NMHC are highly dependent on time and location and the values presented in Figure 8 can only serve as a general guideline to average values. Due to the short atmospheric residence time for most aliphatic NMHC, conditions representative for relatively small temporal and spatial scales often determine their removal rates. It is also important that OH-radical and ozone concentrations as well as temperature exhibit strong, systematic variations with season, latitude, and altitude, which result in corresponding systematic variations of the local atmospheric lifetimes of hydrocarbons.

Generally the atmospheric residence times of aliphatic NMHC decrease from high toward low latitudes and they are longer in winter than in summer. The summer to winter increase is more pronounced at high latitudes than at low latitudes. Furthermore, the removal rates of aliphatic NMHC exhibit a very pronounced diurnal variability, with the fastest losses occurring around noon and very slow, often negligible removal rates during night.

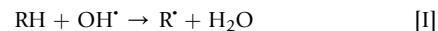
Formation of Secondary Pollutants

Ultimately all atmospheric NMHCs are oxidized to carbon dioxide and water. However, this occurs via complex reaction sequences comprised of many individual steps and several of the products formed are important for the chemistry of the atmosphere. The oxidation of aliphatic NMHC plays a major role in the formation of tropospheric ozone and other photooxidants such as peroxides and peroxyacetyl nitrate. Furthermore, aliphatic NMHCs are precursors for a variety of

oxygenated compounds such as aldehydes, ketones, and carboxylic acids as well as organic nitrates. Oxidation of aliphatic NMHC also contributes to formation of SOAs, but it should be noted that the yield of condensable organic compounds from atmospheric oxidation of most aliphatic NMHCs is very small and only oxidation of aliphatic NMHCs with six or more carbon atoms has significant yields for SOA.

Here we give a brief overview of the reactions that result in the formation of the most important oxidation products of aliphatic NMHC. Since the initial reactions of alkanes and unsaturated compounds are fundamentally different, these two groups of compounds will be treated separately, although the types of products and the subsequent reactions show similarities.

Alkanes. The reaction of alkanes (RH) with OH radicals is entirely by abstraction of a hydrogen atom from one of the carbon atom and results in the formation of an alkyl radical:



For most aliphatic NMHCs, this reaction can occur on different, not equivalent carbon atoms. Generally, the most stable alkyl radical, that is the one with the higher number of alkyl groups attached, is formed preferentially. For example, in the reaction of propane with OH radicals the 2-propyl radicals will be preferred over the 1-propyl radical.

In the atmosphere, this hydrogen atom abstraction initiates a chain reaction that, depending on the concentration of other atmospheric trace constituents, results in the formation of several other radicals and to some extent recycles the OH radical.

Following hydrogen abstraction is a rapid addition of an oxygen molecule:



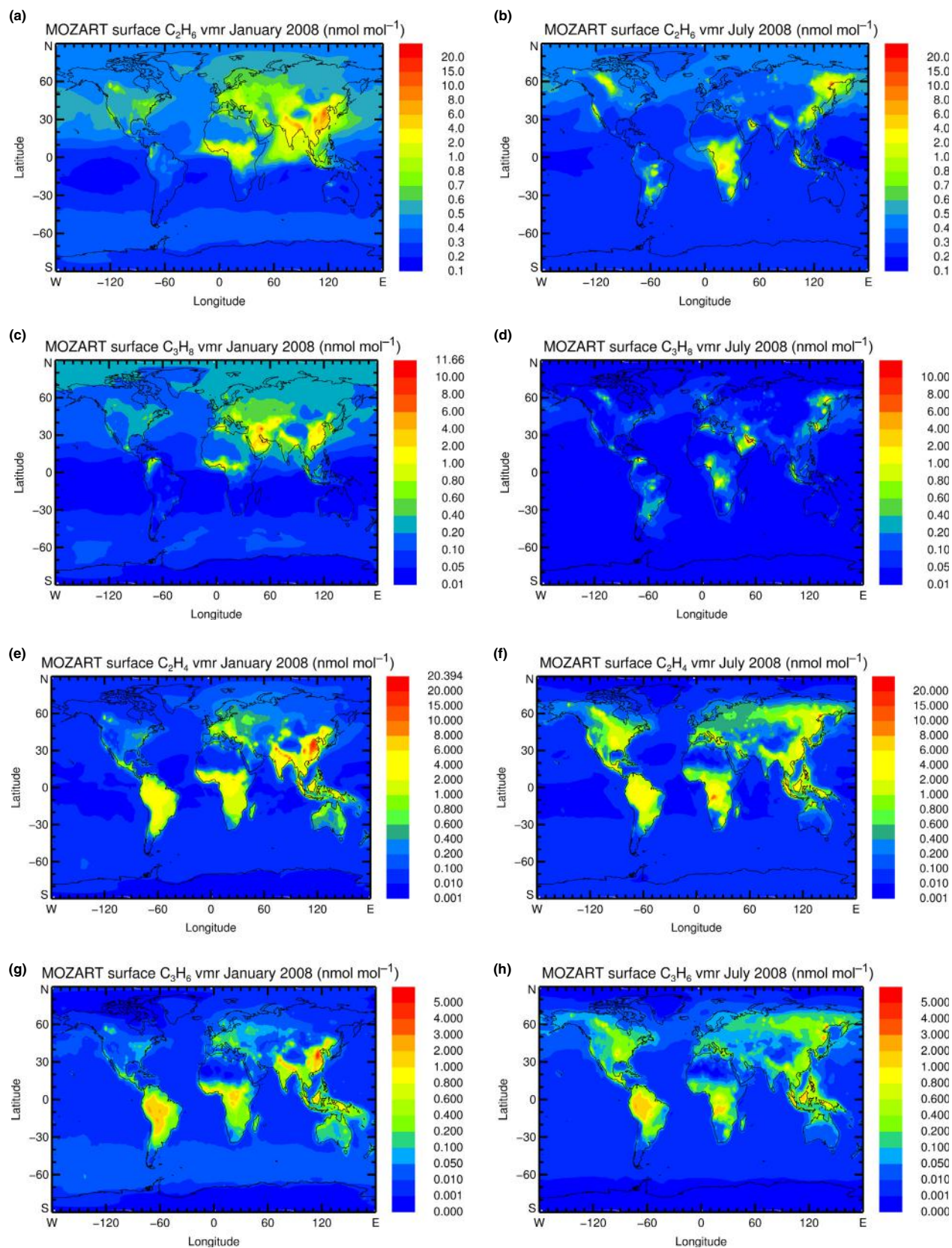
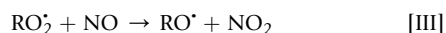


Figure 9 Modeled global distributions of light aliphatic NMHC in the lowest levels of the atmosphere.

In the presence of significant levels of NO, as a rule of thumb more than 40 pmol/mol, the reaction chain proceeds predominantly via [reaction \[III\]](#)



For peroxy radicals with larger alkyl groups to some extent addition of NO, followed by isomerization and formation of an alkyl nitrate is an alternative pathway.

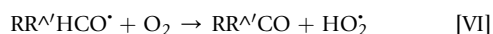


For small alkyl groups, e.g., ethyl or propyl, the [reaction \[IV\]](#) only contributes a few percent or less, but for large alkyl groups with six or more carbon atoms this pathway can contribute in the range of 10–35%. Thus [reaction \[IV\]](#) can be an important source for alkyl nitrates in the atmosphere.

At low NO levels the alkylperoxy radical reacts with other peroxy radicals ($\text{RO}_2^\bullet$ or $\text{HO}_2^\bullet$) as illustrated in [reaction \[V\]](#).

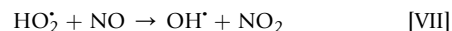


The reaction of the alkoxy radical (RO) is highly dependent on the structure of the alkyl group. For radicals of the $\text{RR}'\text{HCO}^\bullet$ or $\text{RH}_2\text{CO}^\bullet$ type, abstraction of a hydrogen atom by an oxygen molecule is the preferred route:

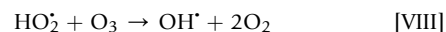


Alternatively, scission of a C–C bond forming a carbonyl compound and an alkyl radical or isomerization to a hydroxy alkyl radical can occur. In the atmosphere, the radical formed by both alternative processes will rapidly add an oxygen molecule and the newly formed peroxy radical will undergo a reaction sequence analogous to that outline above.

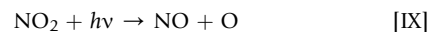
Depending on the concentration of the reactants, the $\text{HO}_2^\bullet$ radical formed by [\[VI\]](#) will react with NO, ozone, or other peroxy radicals. In the presence of NO, the reaction will be similar to [reaction \[III\]](#) and recycles the OH radical:



At very low NO concentrations the reaction sequence proceeds analogous to [reaction \[V\]](#) or by reaction with ozone:



[Reactions \[VII\] and \[VIII\]](#) regenerate the OH radical consumed in the initial reaction step. Consequently, the atmospheric oxidation of alkanes is an OH radical-initiated chain reaction. It is important that the conversion of NO into NO_2 by [\[III\]](#) and [\[VII\]](#) is not a net loss of NO since during daytime NO is regenerated by photolysis of NO_2 .



The oxygen atom formed in this reaction recombines in the atmosphere with an oxygen molecule resulting in the formation of ozone. The efficiency of the formation of ozone from atmospheric oxidation of hydrocarbons thus is highly dependent on NO concentration.

Alkenes. Although the initial step of the reaction of unsaturated aliphatic hydrocarbons is an addition to the double bond instead of abstraction of a hydrogen atom there are many similarities to the oxidation of saturated NMHC. The most obvious difference is that alkenes react very rapidly with OH radicals, often by orders of magnitude faster than the alkanes. The addition of the OH radicals to an alkene results in the formation of an β -hydroxyalkyl radical, which then, similar to alkyl radicals, in the atmosphere rapidly adds an

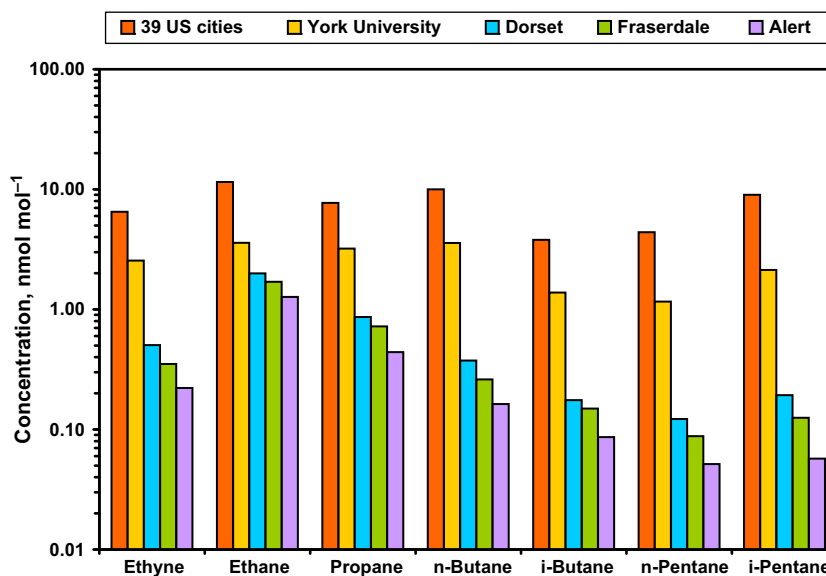
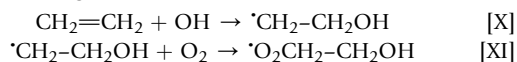
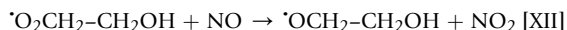


Figure 10 Average concentration of selected aliphatic NMHC in the 1980s at locations subject to different levels of anthropogenic emissions. '39 US cities' represents the average of studies in 39 major cities in the United States; York University is located in a suburban area in the Greater Toronto area, Ontario, Canada; Dorset is a semirural region in southern Ontario; Fraserdale is a remote station in central Ontario; Alert is a background station in the Canadian Arctic. Data are from: Seila, R.L., Lonneman, W., Meeks, S., 1989. Determination of C2 to C12 Ambient Hydrocarbons in 39 U.S. Cities From 1986 Through 1989, US Environmental Protection Agency Official Research Division Report EPA 600/058, pp. 53–89; Jobson, B.T., Wu, Z., Niki, H., Barrie, L., 1994. Seasonal trends of isoprene, C2–C5 alkanes, and acetylene at a remote boreal site in Canada. *Journal of Geophysical Research* 99, 1589–1599; Rudolph, J., unpublished results.

oxygen molecule forming a β -hydroxyalkylperoxy radical as shown here using ethene as example

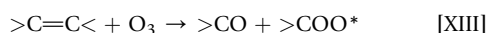


These β -hydroxyalkylperoxy radicals react very similar to the alkylperoxy radicals (reaction [III]). In the presence of NO, the preferred reaction is the formation of a β -hydroxyalkoxy radical.



Analogous to the alkane oxidation an alternative reaction pathway at very low NO concentration is the reaction with other peroxy radicals. In the atmosphere the β -hydroxyalkoxy radicals either react with molecular oxygen, decompose, or isomerize, similarly to the alkylperoxy radicals (see above).

One major difference between the atmospheric chemistry of alkanes and alkenes is that alkenes react with ozone. This reaction results in formation of a carbonyl compound and an excited Criegee intermediate ($>\text{COO}^*$).



In the troposphere the excited Criegee intermediate can undergo a variety of different reactions, the most important channels are collisional deactivation, decomposition, or isomerization. These reaction channels branch further into formation of a variety of more or less complex molecules and radicals, depending on the structure of the alkene precursor. One very important reaction channel is the formation of OH radicals. Especially under conditions when direct photochemical formation of OH radicals is slow, such as at night or in the early morning, alkene ozonolysis can be a relevant source of free radicals.

Qualitatively the effect of atmospheric oxidation of alkenes on the chemistry of the troposphere is very similar to that of alkanes. However, due to their higher reactivity, alkenes often play a more important role for the chemistry of the troposphere, especially in areas directly influenced by strong, local, or regional emissions. Furthermore, the

oxidation of alkenes generally results in more complex molecules and products, increasing the complexity of the chemistry of the troposphere.

Atmospheric Distribution

During the last four decades a variety of techniques to quantitatively determine hydrocarbons in air have been developed, providing the possibility for detailed studies of their atmospheric distribution. It should be noted that measurement techniques for aliphatic NMHC seldom specifically target this group alone, but generally aim at analysis of NMHC in general. The most widely used and versatile measurement technique for aliphatic hydrocarbons is gas chromatography in combination with cryogenic preconcentration of the hydrocarbons prior to the chromatographic separation. This method allows analysis of a broad range of hydrocarbons with detection limits in the lowest pmol mol^{-1} range, sufficient for reliable measurements of relevant concentrations of aliphatic hydrocarbons under most conditions. Generally, whole air samples are collected in airtight sample containers and transferred to the laboratory for analysis. During the last two decades *in situ* instrumentation has also been used in a substantial number of studies. In addition to modified gas chromatographic instrumentation mass spectrometric methods such as Proton Transfer Reaction Mass Spectrometry (PTRMS) are widely used. Mass spectrometric methods have become extremely valuable for measurements that require high time resolution such as mobile platforms or flux and gradient measurement.

Due to the numerous measurement series of aliphatic NMHC in the troposphere that have been conducted during the last decades, there is good understanding of their basic temporal and spatial distribution. However, due to the short atmospheric residence time of most aliphatic NMHCs there is often very large, sometimes seemingly random

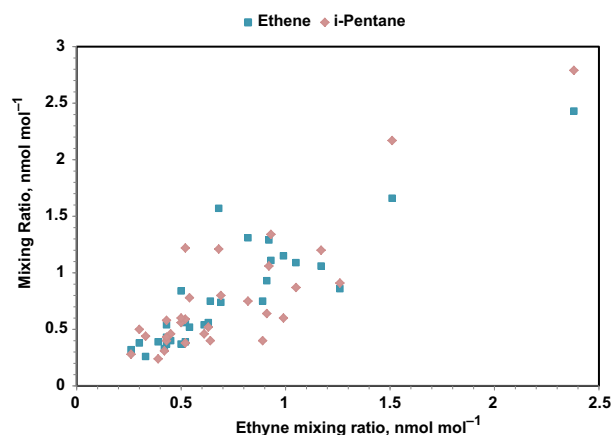


Figure 11 Example for correlation between mixing ratios of aliphatic NMHC. Data points represent the average mixing ratios for 28 US cities. Data are from Baker, A.K., Beyersdorf, A.J., Doeze, L.A. et al., 2008. Measurements of nonmethane hydrocarbons in 28 United States cities. *Atmospheric Environment* 42,170–182.

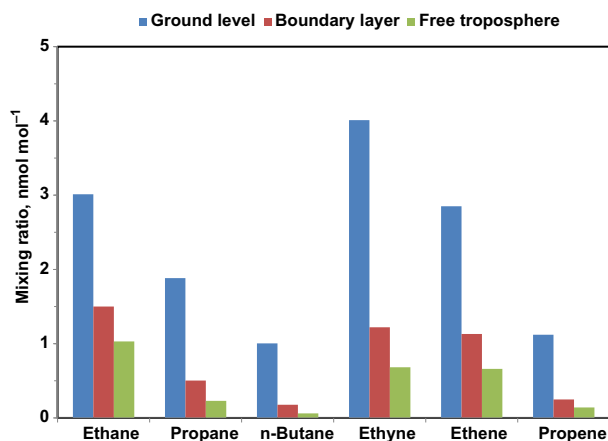


Figure 12 Comparison of mixing ratios of selected aliphatic NMHC at different altitudes over northeast China. Data are from Xue, L., Tao Wang, T., Simpson, I.J., et al., 2011. Vertical distributions of non-methane hydrocarbons and halocarbons in the lower troposphere over northeast China. *Atmospheric Environment* 45, 6501–6509.

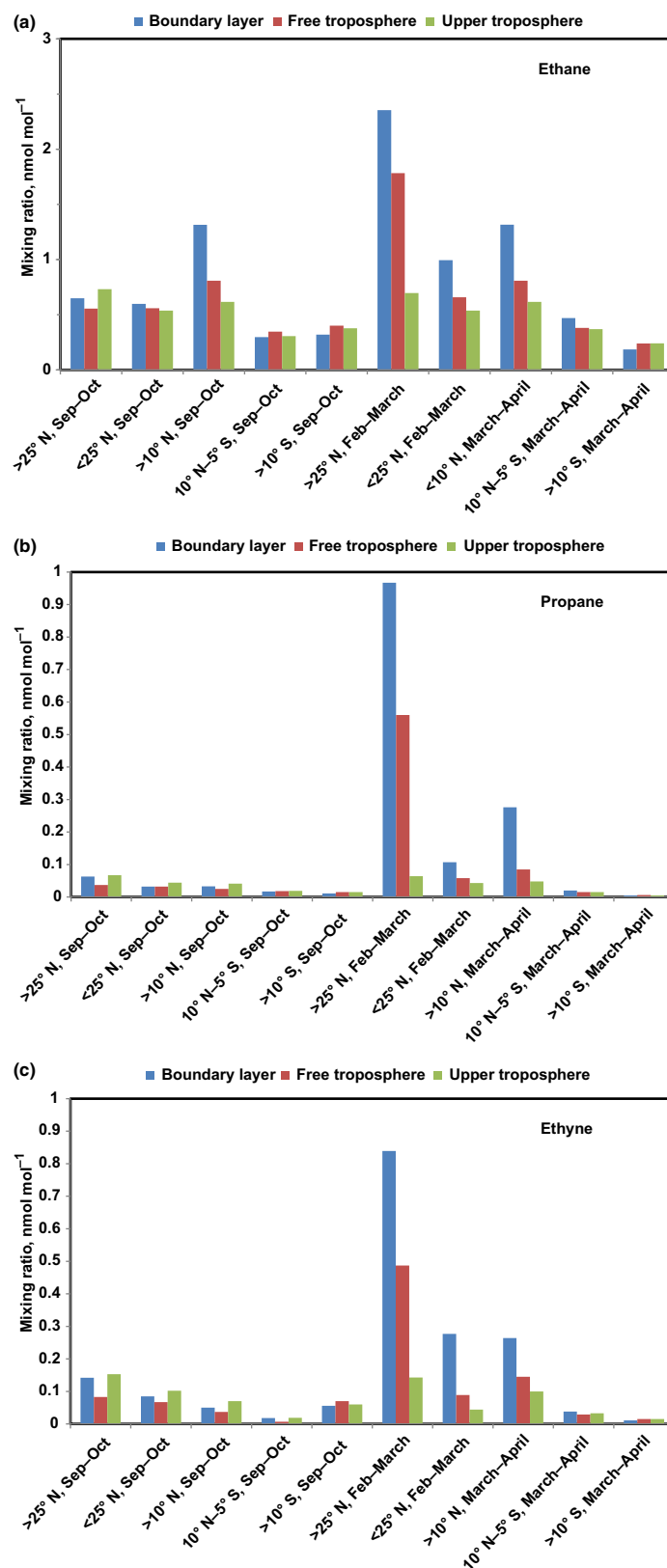


Figure 13 Comparison of mixing ratios of selected aliphatic NMHC between planetary boundary, free troposphere, and upper troposphere over the Pacific for different latitudes and seasons. Data are from Blake, N.J., Blake, D.R., Chen, T.-Y. et al., 1997. Distribution and seasonality of selected hydrocarbons and halocarbons over the western Pacific basin during PEM-West A and PEM-West B. *Journal of Geophysical Research Atmosphere* 102, 28315–28331; Blake, N.J., Blake, D.R., Simpson, I.J., et al., 2001. Large-scale latitudinal and vertical distributions of NMHCs and selected halocarbons in the troposphere over the Pacific Ocean during the March–April 1999 Pacific Exploratory Mission (PEM-Tropics B). *Journal of Geophysical Research Atmosphere* 106, 32627–32644.

Table 4 Relative trend in atmospheric concentrations of selected aliphatic NMHC (% yr⁻¹)

Location	Urban London	Rural England	Remote mountain site in Germany
Ethane	-4	Insignificant trend	Insignificant trend
Propane	-3	-3	-2
n-Butane	-17	-6	-4
n-Pentane	-12	-2	Insignificant trend
Ethene	-20	-6	-4
Propene	-20	-6	-10
Ethyne	-15	-7	-4

Data are taken from von Schneidmesser, E., Monks, P.S., Plass-Duelmer, C., 2010. Global comparison of VOC and CO observations in urban areas. Atmospheric Environment 44, 5053–5064.

variability. Consequently, measurements of their concentrations are generally only representative on local or regional scales. Only for some of the least reactive of the aliphatic NMHCs reasonably well-defined and representative global distribution have been derived from atmospheric observations.

Concentrations of aliphatic NMHC range from several ten nanomoles per moles in urban, polluted regions to a fraction of nanomoles per moles and less in the remote Southern Hemisphere. The distribution of the concentration of aliphatic NMHC in the troposphere is determined by the interaction of three processes, (1) emissions; (2) atmospheric transport, dilution, and mixing; and (3) removal. Since all three processes show a complex dependence on time and location, the very substantial variability of the concentrations of aliphatic NMHC in the troposphere is not surprising.

Nevertheless, a substantial part of this variability is systematic. Today, a reasonable overview of the basic features of the atmospheric distribution of aliphatic NMHC can be obtained from computer-based numerical model simulations. While these model calculations seldom are able to describe all details of the atmospheric variability of NMHC concentrations, they have developed to a stage where they can provide a good overview. Numerical models are used to calculate the atmospheric distribution of NMHC on a wide range of spatial and temporal timescales.

Large-scale distributions. Figure 9 presents modeled global ground level distributions for aliphatic NMHC with average atmospheric lifetimes ranging from about 2 month (ethane) to less than 1 day (propene). The differences between January and July are mainly due to the seasonal variability of the OH-radical concentration, but also changes in atmospheric transport patterns and seasonal variation of emission rates, for example from biomass burning or vegetation. Atmospheric concentrations of aliphatic NMHC rapidly decrease with increasing distance from the source regions. The concentration gradients are the result of atmospheric mixing and chemical loss processes. The strong gradients of the latitudinal distributions between midnorthern latitudes and the Southern Hemisphere reflect primarily the large-scale distribution of the sources, but it has to be remembered that atmospheric transport and removal processes also play

a major role in determining the shape of the latitudinal profiles. Specifically, the atmospheric degradation of aliphatic NMHC by reaction with OH radicals is faster at low latitudes, a consequence of the latitudinal gradient of the OH-radical concentration. Consequently, seasonal cycles in the Southern Hemisphere are shifted by 6 month relative to those in the Northern Hemisphere and the seasonal variability in the tropics is less pronounced than at midlatitudes. The steepness of the gradients between source regions and remote areas depends on reactivity of the trace gas and the change in concentration from areas with high emission rates toward background ratios is more pronounced for NMHC with high reactivity. The result is an increase in systematic as well as seemingly random variability of mixing ratios with increasing NMHC reactivity. Figure 10 shows an example for the change in observed mixing ratios for locations subject to different levels of artificial trace gas emissions.

One frequently observed consequence of differences in NMHC reactivity is the systematic change in the ratios of NMHC concentrations with increasing NMHC processing. This is used to estimate the extent of photochemical processing or photochemical age of the NMHC ('hydrocarbon clock'). However, due to the complex interaction of transport and processing the photochemical age derived from ratios of hydrocarbon concentrations is sometimes difficult to interpret. In the absence of significant photochemical loss, for example, in close proximity to strong NMHC sources, often a strong correlation between the mixing ratios of different aliphatic NMHC can be seen (Figure 11). This is the result of variability of emission rates, atmospheric mixing, and dilution. Since the relation between different trace gases ('fingerprint') is often characteristic for specific source types this can be used to

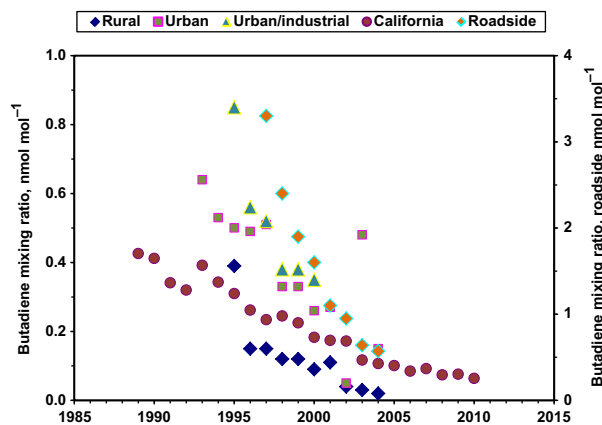


Figure 14 Trend of butadiene mixing ratios at different types of locations in the United Kingdom. For comparison also the average of observations at several locations in California are included. Please note that the scale for roadside observations is on a secondary axis. Average data for California are taken from the California Environmental Protection Agency Air Resources Board (<http://www.arb.ca.gov/adam/toxics/statepages/butastate.html>). Data for the UK are from Dollard, G.J., Dumitrean, P., Telling, S., Dixon, J., Derwent, R.G., 2007. Observed trends in ambient concentrations of C2–C8 hydrocarbons in the United Kingdom over the period from 1993 to 2004. Atmospheric Environment 41, 2559–2569.

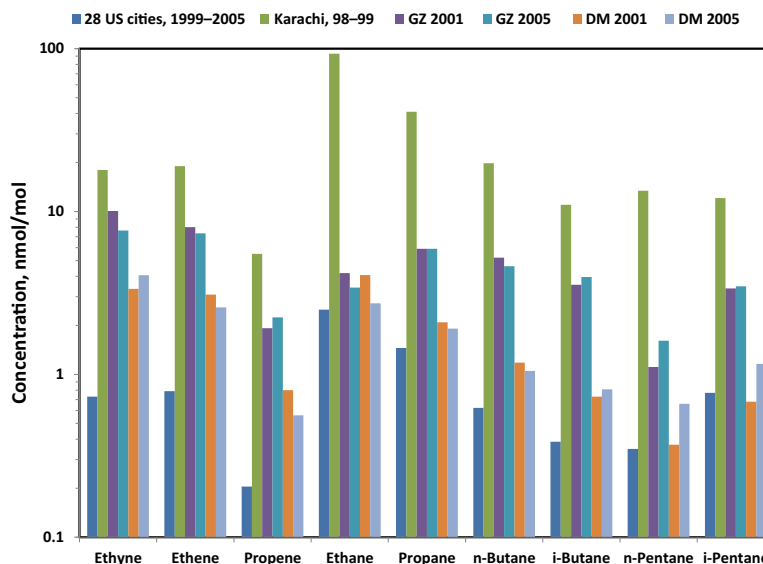


Figure 15 Mixing ratios of selected aliphatic NMHC at urban and rural locations in Pakistan (Karachi) and China (GZ, Guangzhou a highly industrialized large City; DM, Denghou Mountain, a rural area in the Pearl River Delta) between 1999 and 2005. Included for comparison are average mixing ratios for 28 US cities measured between 1999 and 2005. Data are taken from Baker, A.K., Beyersdorf, A.J., Doezema, L.A., et al., 2008. Measurements of nonmethane hydrocarbons in 28 United States cities. *Atmospheric Environment* 42, 170–182; Tang, J.H., Chan, L.Y., Chan, C.Y. et al., 2008. Implications of changing urban and rural emissions on nonmethane hydrocarbons in the Pearl River Delta region of China. *Atmospheric Environment* 42, 3780–3794; Barletta, B., Meinardi, S., Simpson, I.J. et al., 2002. Mixing ratios of volatile organic compounds (VOCs) in the atmosphere of Karachi, Pakistan. *Atmospheric Environment* 36, 3429–3443.

identify the contributions of different source types to air pollution ('source apportionment').

The vertical distributions of aliphatic NMHC depend strongly on local and regional emission rates and the atmospheric lifetime of the compound. Figure 12 shows examples for the decrease in mixing ratios of aliphatic NMHC between ground level and free troposphere over a region with high emission rates. For regions with low or insignificant sources the gradients are often less pronounced and for very remote regions vertical gradients of the least reactive aliphatic NMHC are often insignificant (Figure 13).

In urban areas, the concentrations of aliphatic NMHC often show distinct diurnal patterns with high values in the morning and lower concentrations in the afternoon. Although OH-radical concentrations have distinct diurnal variations, with high values around noon and in early afternoon, and very low values at night, in the late evening, and early morning, the diurnal variability of the chemical removal rate is very seldom the main driving force behind this diurnal variability. Very often the diurnal cycle of aliphatic NMHC is driven by diurnal variations in atmospheric transport or mixing processes and the emission rates, fully consistent with the finding of generally very similar composition pattern in morning and afternoon, independent of the reactivity of the aliphatic NMHC.

Secular trends. Due to the high spatial and temporal variability of most aliphatic NMHCs secular trends in their atmospheric mixing ratios differ between regions. For Western Europe and North America the significant decrease in emissions resulted in a significant reduction in the atmospheric levels of most aliphatic NMHCs (Table 4; Figure 14). For

most developing countries present-day levels of air pollution in major cities significantly exceed those in developed countries (Figure 15). There is some evidence that during the last decade increasing efforts in air pollution mitigation in developing countries reduce the upward trend in pollution levels. However, due to the limited extent of systematic, consistent monitoring of aliphatic NMHC in most developing countries the presently available data do not allow identification of representative secular trends.

Only for ethane, the least reactive of the aliphatic NMHC, observations allow deriving a meaningful global trend. During the last two decades the global average mixing ratio decreased by about $0.15 \text{ nmol mol}^{-1}$ from nearly $0.8 \text{ nmol mol}^{-1}$ in the mid 1980s.

See also: Aviation Meteorology: Aircraft Emissions. **Chemistry of the Atmosphere:** Chemical Kinetics; Principles of Chemical Change. **Tropospheric Chemistry and Composition:** Aromatic Hydrocarbons; Biogenic Hydrocarbons; Hydroxyl Radical; Oxidizing Capacity; Volatile Organic Compounds Overview: Anthropogenic.

Further Reading

- Brasseur, G.P., Orlando, J.J., Tyndall, G.S. (Eds.), 1999. *Atmospheric Chemistry and Global Change*. Oxford University Press, New York.
- Ehalt, D.H., 1999. Gas phase chemistry of the troposphere. In: Zellner, R. (Ed.), *Global Aspects of Atmospheric Chemistry*. Springer-Verlag, Berlin, pp. 21–109.
- Finlayson-Pitts, B.J., Pitts Jr., J.N., 2000. *Chemistry of the Upper and Lower Atmosphere*. Academic Press, San Diego.
- Graedel, T.E., Crutzen, P.J., 1993. *Atmospheric Change*. W.H. Freeman, New York.

- Hewitt, N. (Ed.), 1999. *Reactive Hydrocarbons in the Atmosphere*. Academic Press, San Diego.
- Inness, A., Baier, F., Benedetti, A., et al., 2013. The MACC reanalysis: an 8 year data set of atmospheric composition. *Atmospheric Chemistry and Physics* 13, 4073–4109. <http://dx.doi.org/10.5194/acp-13-4073-2013>.
- Jeffries, H.E., 1995. Photochemical air pollution. In: Singh, H.B. (Ed.), *Composition, Chemistry, and Climate of the Atmosphere*. Van Nostrand Reinhold, London, pp. 308–348.
- Koppmann, R. (Ed.), 2007. *Volatile Organic Compounds in the Atmosphere*. Blackwell Publishing, Oxford.
- Seinfeld, J.H., Pandis, S.N., 1998. *Atmospheric Chemistry and Physics*. Wiley, Chichester.
- Warneck, P., 1988. *Chemistry of the Natural Atmosphere*. Academic Press, London.

Relevant Websites

Rate Constants and Reaction Mechanisms
<http://jpldataeval.jpl.nasa.gov/>.
<http://kinetics.nist.gov/>.
<http://mcm.leeds.ac.uk/MCM/>.
Emission Inventories
<http://eccad.sedoo.fr/>.
<http://edgar.jrc.ec.europa.eu/>.
<http://join.iwk.fz-juelich.de/macc>.
<http://www.gmes-atmosphere.eu/fire/>.

Aromatic Hydrocarbons

I Barnes, University of Wuppertal, Wuppertal, Germany

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by I Barnes, K H Becker, volume 6, pp 2376–2385, © 2003, Elsevier Ltd.

Synopsis

Aromatic hydrocarbons are an important class of organic compounds found in the atmosphere, which are highly reactive and have large emission rates, particularly in urban environments. The article describes the gas-phase atmospheric chemistry of the atmospherically most important aromatic hydrocarbons, benzene, toluene, ethyl benzene, and the isomers of xylene that are collectively often referred to as BTEX. The article discusses the emission sources and atmospheric gas-phase degradation processes of BTEX, which have important implications for photooxidant and secondary organic aerosol formation in the troposphere.

Introduction

Aromatic hydrocarbons are an important class of organic compounds found in the atmosphere, which are highly reactive and have large emission rates. It has been estimated that aromatic species contribute about 10% to the total global anthropogenic nonmethane organic carbon (NMOC) emissions, the major source being car exhaust from gasoline-powered vehicles, with a significant contribution also from solvent usage. The observation that biomass burning represents a significant source of aromatics has attracted special attention in recent years, since such processes occur on a global scale. Natural, minor sources of aromatics such as emission from soils and plants have also been identified. The emission of aromatics will therefore have an impact on tropospheric processes on local, regional, and global scales.

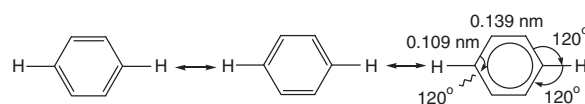
In many urban areas, emissions of volatile organic compounds (VOCs) and NO_x , when combined under meteorological conditions that favor smog formation, can lead to the formation of ground-level ozone and particulate matter (PM) at harmful levels. Aromatic hydrocarbons make a significant contribution to the formation of ozone and other photooxidants in urban atmospheres. It has been estimated that the percentage contribution of aromatic hydrocarbons to ozone production can be between 30 and 40%. Such a high contribution would make the aromatics the most important class of hydrocarbons with regard to photochemical ozone formation. However, it should be borne in mind that the oxidation mechanisms of aromatic hydrocarbons are still imprecisely known, and verification of such estimates is necessary. In addition to their high atmospheric photochemical reactivity and their consequent major influence on the formation of tropospheric ozone and on the oxidizing capacity of the atmosphere, it is now firmly established that the photochemistry of aromatic compounds leads to the formation of secondary organic aerosols (SOAs), which are known to be harmful to human health, reduce visibility, and can contribute to climate change. In fact, studies have shown that the atmospheric organic aerosol formation potential of whole gasoline vapor can be accounted for solely in terms of the aromatic fraction of the fuel. As emissions of aromatics are concentrated in urban areas, where many people live and work, the

formation of SOAs becomes a more acute problem. With regard to health effects, benzene, a ubiquitous aromatic environmental contaminant, is hematotoxic, clastogenic, and leukemogenic in humans and produces bone marrow toxicity and various types of cancer in animals. The reason for the carcinogenesis of benzene is still not entirely clear and it has been classified as a Group-A human carcinogen by the US Environmental Protection Agency. Laboratory studies have also shown that the photooxidation of other aromatic hydrocarbons leads to the formation of mutagenic products.

Presented here is a brief overview of the atmospheric composition and gas-phase chemistry of aromatic hydrocarbons concentrating mainly on benzene, toluene, the xylene isomers, and ethyl benzene, since these rank highly among the most important aromatic hydrocarbons emitted to the atmosphere. This group of aromatic hydrocarbons is often referred to as BTEX in the literature. Since the subject is complex, the reader is encouraged to consult the recent articles and books listed in the reading material for coverage in depth on particular aspects of aromatic hydrocarbon chemistry covered here.

Structure and Concentration in the Atmosphere

The simplest of the aromatic hydrocarbons is benzene, which has a hexagon ring structure with three 'double bonds' in the ring. In the aromatic benzene ring, these 'double bonds' are fully conjugated and consequently the compound is relatively stable. All carbons are sp^2 hybridized, and each π orbital overlaps, to an equal extent, its two neighbors. The consequently delocalized electrons form a π cloud above and below the ring. In the molecular structure of benzene all six C–C bonds are equal and all bond angles are 120° . The C–C bond length in benzene is 0.139 Å, between the values for the single (0.147 nm) and the double bond (0.134 nm) in butadiene. All of the three structures shown here for benzene are equivalent.



The generic term for substituted benzenes is arene. An arene, as a substituent, is referred to as an aryl group, abbreviated Ar. The parent aryl substituent is phenyl, C_6H_5 . There are three possible arrangements of disubstituted benzenes. These are described by the prefixes 1,2 (*ortho*-, or *o*-, Greek: straight) for adjacent substituents, 1,3 (*meta*-, or *m*-, Greek: transposed) for 1,3-disubstitution, and 1,4 (*para*-, or *p*-, Greek: beyond) for 1,4-disubstitution. The substituents are listed in alphabetical order. **Figure 1** lists the structural formula, International Union of Pure and Applied Chemistry (IUPAC), and common names of the more prevalent types of aromatic hydrocarbon found in the atmosphere.

Of the 32 most prevalent nonmethane hydrocarbons (NMHCs) observed in urban air, 7 are aromatic hydrocarbons. Toluene is often the most abundant aromatic compound in urban air accounting for around 6% of the observed NMHC. A wealth of data now exists on aromatic hydrocarbon composition and concentration in the atmosphere and the listed literature can be used to access this information. Atmospheric concentrations of aromatic hydrocarbons in urban areas in Europe and North America are often in the ppb range, for example, measured concentrations for benzene, toluene, the xylene isomers, and ethyl benzene are typically in the ranges 1–9, 1–17, 0.3–10, and 0.3–2 ppb, respectively. Recent measurements in Asian cities and metropolitan areas show that the BTEX levels in these areas are often much higher whereby a large fraction of the emissions comes from the use of coal as a domestic fuel. In rural and marine regions, the concentrations are generally below 0.5 ppb.

Table 1 shows an example of the typical VOC composition of Los Angeles morning air in October 1995, with the percentages of the contributions from alkanes, alkenes, aromatics, and other NMHCs. The percentage of OH reactivity (rate coefficient weighted) for each individual compound and compound class group is also shown. **Table 2** shows a comparison of total nonmethane organic compounds (TNMOCs) and percentile distribution of aromatic hydrocarbons for selected European locations. Both tables demonstrate

the high contribution of aromatic compounds to the NMOC mix, particularly in urban environments where the major sources occur. The emissions of aromatic hydrocarbons are overwhelmingly anthropogenic in origin; major sources are the evaporation of fuels, automobile tailpipe exhaust, and industrial processes – in particular, solvent utilization. In many instances, the measured aromatic hydrocarbon composition resembles very closely the aromatic hydrocarbon signature obtained from studies on the NMOC emissions from automobiles, earmarking on-road vehicles as the leading, if not major, source of urban aromatic hydrocarbons. In rural areas where emissions from biogenics dominate the contribution from aromatic hydrocarbons is much lower. In urban areas and suburban areas with a high biomass contribution the measured atmospheric NMOC often resembles a combination of the expected anthropogenic and biogenic emission signatures for the sources present in that particular region.

Biomass burning represents a small but significant emission source of aromatic hydrocarbons, in particular benzene. Plants are known to emit unsaturated compounds containing an aromatic ring and microorganisms can produce benzene and toluene. It has also been shown quite recently that some plants under stress situations emit toluene. At present it is not possible to quantify the emissions of simple aromatic hydrocarbons from the biosphere, but the indications are that the contribution is small.

Estimations of global emissions of aromatics range from 18.7 (17.1% of the total anthropogenic nonmethane volatile organic compounds (NMVOCs) emissions) to 25 teragrams per year, with benzene and toluene alone constituting 25% of the total anthropogenic NMVOC emissions. Regulations in industrialized countries are helping to reduce the aromatic content of fuel; however, even if the content is lowered to about 20 mass percent, the continually increasing global oil demand will keep the global emissions of aromatic hydrocarbons from fossil fuel combustion at a high level. The rapid urbanization rate, growth of megacities, and economic improvements in many parts of Asia are coupled with increasing levels of

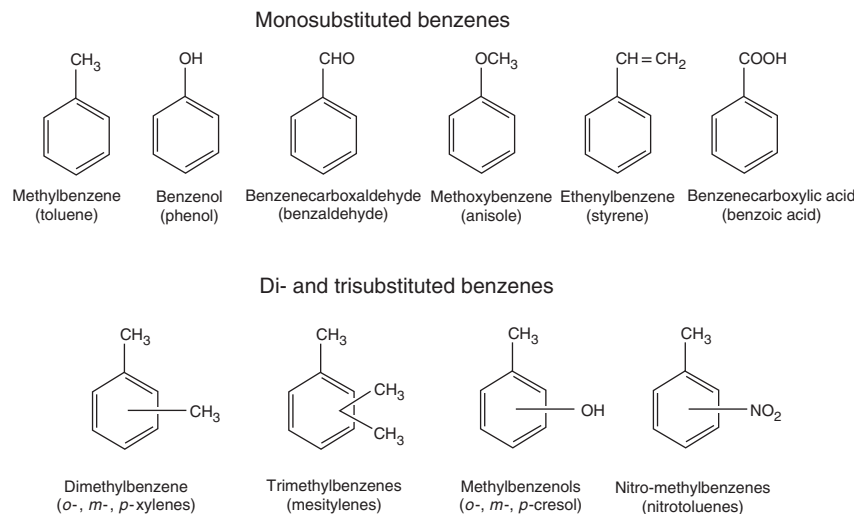


Figure 1 Structures and names of simple aromatic compounds commonly found in the atmosphere.

Table 1 Typical composition of Los Angeles morning air in October 1995

	Average ppb C	Min ppb C	Max ppb C	OH reactivity (%)
<i>Methane</i>	2000			1.7
<i>CO</i>	500			11.6
Alkanes (42% of NMHC)	211.2			14.6
2-Methylbutane	28.69	8.01	74.64	2.1
Propane	27.50	9.18	51.49	1.0
Ethane	20.23	7.92	47.12	0.2
<i>n</i> -Butane	16.58	5.01	32.58	1.0
<i>n</i> -Pentane	15.39	4.58	34.85	1.2
2-Methylpentane	12.26	3.99	35.78	1.1
Methylcyclopentane	8.73	2.43	25.96	1.0
<i>n</i> -Hexane	8.16	2.27	23.67	0.7
2*-Methylpropane	8.08	3.09	12.86	0.4
2,2,4-Trimethylpentane	7.86	2.11	24.08	0.3
3-Methylpentane	7.41	2.80	21.28	0.6
3-Methylhexane	6.72	1.97	16.68	0.7
Methylcyclohexane	6.28	1.32	13.95	0.9
<i>n</i> -Heptane	5.02	1.20	11.47	0.5
2-Methylhexane	4.94	1.45	14.58	0.5
2,3-Dimethylpentane	4.72	1.23	14.21	0.3
2,3-Dimethylbutane	3.27	0.91	9.57	0.3
Cyclohexane	3.14	1.07	8.65	0.4
2,4-Dimethylpentane	2.81	0.77	8.96	0.2
2-Methylheptane	2.29	0.70	6.62	0.2
2,3,4-Trimethylpentane	2.23	0.57	8.67	0.2
<i>n</i> -Octane	2.09	0.77	6.02	0.2
Cyclopentane	1.95	0.58	4.80	0.2
3-Methylheptane	1.94	0.54	6.59	0.2
<i>n</i> -Nonane	1.62	0.51	4.26	0.2
2,2-Dimethylbutane	1.24	0.42	3.40	0.0
Alkenes (7% of NMHC)	34.2			27.3
Ethene	14.91	4.31	45.67	6.2
Propene	5.88	1.48	20.22	5.0
2-Methylpropene and 1-butene	4.97	1.23	20.27	6.2
2-Methyl-2-butene	1.48	0.38	6.96	2.1
<i>trans</i> -2-Pentene	1.16	0.00	5.33	1.1
1-Pentene	0.92	0.14	2.75	0.6
<i>trans</i> -2-Hexene	0.80	0.13	2.31	0.9
2-Methyl-1,3-butadiene	0.68	0.14	2.47	1.3
<i>trans</i> -2-Pentene	0.66	0.13	2.78	0.8
<i>trans</i> -2-Butene	0.60	0.17	2.48	0.9
<i>cis</i> -2-Butene	0.56	0.11	2.55	0.8
Cyclopentene	0.38	0.07	1.77	0.5
3-Methyl-1-butene	0.35	0.09	1.10	0.2
2-Methyl-1-pentene	0.32	0.00	1.33	0.3
4-Methyl-1-pentene	0.29	0.00	0.61	0.2
<i>cis</i> -2-Hexene	0.23	0.05	1.04	0.2
Aromatics (19% of NMHC)	96.4			17.9
Toluene	33.02	9.36	91.01	2.7
<i>m</i> - and <i>p</i> -Xylene	19.92	5.77	68.20	4.6
1,2,4-Trimethylbenzene	12.12	3.90	32.72	4.2
Benzene	9.73	3.62	30.05	0.2
<i>o</i> -Xylene	7.14	2.17	23.75	1.2
Ethylbenzene	5.41	1.76	16.79	0.5
1,3,5-Trimethylbenzene	3.47	0.83	14.39	2.1
Styrene	3.06	0.79	10.30	2.2
<i>n</i> -Propylbenzene	1.60	0.57	4.40	0.1
<i>iso</i> -Propylbenzene	0.94	0.20	6.10	0.0

(Continued)

Table 1 Typical composition of Los Angeles morning air in October 1995—cont'd

	Average ppb C	Min ppb C	Max ppb C	OH reactivity (%)
<i>Methane</i>	2000			1.7
<i>CO</i>	500			11.6
Other (33% of NMHC)	165.2			27.0
Ethyne	17.35	5.76	54.86	0.7
MTBE	20.46	6.80	62.64	1.2
Other identified	73.86	23.22	222.93	14.6
Unidentified	53.57	19.38	166.14	10.6
Total	3007			100

NMHC, Nonmethane hydrocarbon; MTBE, Methyl tert-butyl ether.

Source: Fujita, E.M., Lu, Z., Sheetz, L., Harshfeld, G., Zielinska, B., 1997. Determination of Mobile Source Emission Fraction Using Ambient Field Measurements. Report available from the Coordinating Research Council, 3650 Mansell Road, Suite 140, Alpharetta, GA 30022-8246, USA.

pollution and unless stringent control regulations are enforced high levels of aromatic hydrocarbons will remain a problem in these areas for decades to come.

Chemical Transformation Mechanisms in the Atmosphere

Main Oxidants and Relative Importance

The gas-phase oxidation of organic compounds in the troposphere is initiated by reactions with OH and NO₃ radicals, O₃ and potentially Cl atoms and ground state oxygen O(³P) atoms, and in some instances by photolysis at wavelengths longer than 290 nm. Since benzene and the mono-, di-, and trialkyl-substituted benzenes absorb very little of the sunlight present in the troposphere, loss through photolysis is unimportant in the urban atmosphere. The reaction of aromatic hydrocarbons with O(³P) atoms is also of negligible importance. Average Cl atom concentrations will generally not be high enough, particularly in urban environments, to play any significant role in the degradation of aromatic hydrocarbons and consequently such reactions are not considered here.

The relative importance of the other reaction pathways depends on the atmospheric concentrations of the reaction initiators (O₃, OH, NO₃) and their rate coefficients with the individual aromatic hydrocarbon. Because of the high variability of pollutant concentrations and the diurnal cycle in solar radiation, the atmosphere concentrations of O₃, OH, and NO₃ can vary considerably in time and space. Table 3 lists rate coefficients for reactions of a selection of aromatic hydrocarbons, with O₃, OH, and NO₃ for 298 K and 1 atm total pressure. It is instructive to compare the lifetimes of the aromatic hydrocarbons for typical concentrations of the oxidant species, which can occur during daylight and nighttime hours in a polluted urban atmosphere; such a comparison is made in Table 4.

It is evident from Table 4 that the oxidation of benzene and the alkyl-substituted benzenes in the atmosphere is initiated almost exclusively by reaction with OH radicals, the

Table 2 Comparison of total nonmethane organic compounds (TNMOCs) and percentage distribution of aromatic hydrocarbons for selected European locations

Location	Characterization	Country	TNMOC	Aromatics (%)
(a) Data from several field studies in the late 1980s and early 1990s				
Milan	Urban	Italy	490.8 $\mu\text{g m}^{-3}$	43.7
Rome	Urban	Italy	583.5	34.9
Taranto	Urban	Italy	586	34.1
Montelibretti	Suburban/rural	Italy	360.5	6.0
Madrid	Suburban	Spain	104.3	34.5
Castelporzio	Forest/rural	Italy	88.7	2.6
Castelporzio	Pine-oak forest	Italy	95.5	3.9
Viols en Laval	Forest maquis	France	36.0	6.0
Monti Crimini	Pine forest	Italy	96.2	8.7
Storkow	Pine forest	Germany	91.2	25.8
Burriana	Orange field	Spain	25.5	5.0
Apenini mountains	Forest-rural	Italy	34.0	1.6
Svalbard Islands	Remote arctic	Norway	48.3	5.2
(b) Measurement made in 1997 and 1998 in Germany				
Wuppertal	Urban	Germany	25–184 ppb C	44
Wuppertal	Traffic tunnel	Germany	87–5528	47
Menz	Rural, Berlin plume	Germany	23.7	28
Menz	Rural, background	Germany	12	19
(c) Berlin, Germany. Summer 1996, 4-month average				
Frohnau	Suburban	Germany	31 $\mu\text{g m}^{-3}$	14.0
Nansenstrasse	Urban, roadside	Germany	88	25.0
Frankfurter Allee	Urban, roadside	Germany	175	31.8
(d) Munich, Germany. 15 March–3 April 1996. Urban				
All days	Mean values	Germany	112.3 ppb C	34.0
Weekdays	Mean values	Germany	117.2	35.3
Weekdays 5.00–21.00 h	Mean values	Germany	134.0	36.3

Sources: Ciccioli, P., Brancaleoni, E., Frattoni M., 1999. Reactive hydrocarbons in the atmosphere at urban and regional scales. In: Hewitt, C.N. (Ed.), *Reactive Hydrocarbons in the Atmosphere*. Academic Press, San Diego, CA, pp. 159–207; Kern, T., Metz, N., Kley, D., 1997. Untersuchungen von Verkehrsabgasemissionen insbesondere im Hinblick auf die Ozonbildung, Abschlußbericht Forschungszentrum Jülich, D-52425 Jülich und BMW AG, Energie und Umwelt, D-80788 München; Kurtenbach, R., Brockmann, K.J., Brust, et al., 1999. VOC-measurements in rural air within the BERLIOZ campaign. In: Becker, K.H. (Ed.), *The BERLIOZ Campaign/TFS-LT3 Annual Report 1998* (German), University of Wuppertal; Thijssen, T.R., van Oss, R.F., Lenschow, P., 1999. Determination of source contributions to ambient volatile organic compound concentrations in Berlin. *J. Air Waste Manage. Assoc.* 49, 1394–1404.

Table 3 Room temperature rate coefficients for the reaction of aromatic hydrocarbons with OH and NO₃ radicals and O₃ in air

Compound	$k(\text{OH}) \text{ cm}^3 \text{ molecule}^{-1} \text{ s}^{-1}$	$k(\text{NO}_3) \text{ cm}^3 \text{ molecule}^{-1} \text{ s}^{-1}$	$k(\text{O}_3) \text{ cm}^3 \text{ molecule}^{-1} \text{ s}^{-1}$
Benzene	1.22×10^{-12}	$<3 \times 10^{-17}$	$<1 \times 10^{-21}$
Toluene	5.63×10^{-12}	7×10^{-17}	$<1 \times 10^{-21}$
Ethylbenzene	7.0×10^{-12}	$\leq 5.7 \times 10^{-16}$	$<1 \times 10^{-21}$
<i>o</i> -Xylene	13.6×10^{-12}	4.1×10^{-16}	$<1 \times 10^{-20}$
<i>m</i> -Xylene	23.1×10^{-12}	2.6×10^{-16}	$<3 \times 10^{-21}$
<i>p</i> -Xylene	14.3×10^{-12}	5.0×10^{-16}	$<1 \times 10^{-21}$
1,2,4-Trimethylbenzene	32.5×10^{-12}	1.8×10^{-15}	1.3×10^{-21}
Benzaldehyde	12.9×10^{-12}	2.4×10^{-15}	–
Styrene	58×10^{-12}	1.5×10^{-12}	1.7×10^{-17}
Phenol	27×10^{-12}	3.6×10^{-12}	–
<i>o</i> -Cresol	41×10^{-12}	1.4×10^{-11}	3×10^{-19}

contribution from NO₃ being extremely small and that from O₃ negligible. Reaction with O₃ is of any significance only for aromatic alkenes such as styrene. Reaction with NO₃ radicals is important for aromatic alkenes and also aromatic hydrocarbons containing hydroxyl groups (e.g., phenol, cresol isomers, etc.). At night this removal process dominates the removal of

these compounds under the given conditions. However, under certain circumstances the NO₃ reaction can dominate the removal of hydroxybenzenes, and also contribute to the removal of aromatic alkenes during daytime. The removal process can be particularly important at low sun angles, when the NO₃ photolysis loss becomes slow enough to allow the

Table 4 Calculated lifetimes for selected aromatic hydrocarbons for reaction with OH and NO₃ radical and O₃ concentrations representative of a polluted atmosphere

Aromatic compound	Daytime			Nighttime		
	[OH] (0.16 pptv)	[NO ₃] (3 pptv)	[O ₃] (110 ppbv)	[OH] (0.0007 pptv)	[NO ₃] (100 pptv)	[O ₃] (80 ppbv)
Benzene	2.4 days	14.3 year	>11.7 year	548 days	157 days	>16.1 year
Toluene	12.5 h	6.1 year	>11.7 year	119 days	67 days	>16.1 year
Ethylbenzene	10.1 h	≥0.75 year	>11.7 year	96 days	≥8 days	>16.1 year
<i>o</i> -Xylene	5.2 h	1.16 year	>1.2 year	49 days	11 days	>1.7 year
<i>m</i> -Xylene	3.1 h	1.65 year	>3.9 year	29 days	18 days	>5.4 year
<i>p</i> -Xylene	4.9 h	0.86 year	>11.7 year	47 days	9 days	>16.1 year
1,2,4-Trimethylbenzene	2.2 h	87 days	9 year	21 days	2.6 days	12.4 year
Benzaldehyde	5.5 h	65 days	–	52 days	2 days	–
Styrene	1.2 h	2.5 h	6 h	11 days	4.5 min	8.3 h
Phenol	2.6 h	1.1 h	–	25 days	2 min	–
<i>o</i> -Cresol	1.7 h	16 min	14.3 days	16 days	29 s	19.7 days

NO₃ concentration to build up to levels where it can effectively compete with the OH reactions.

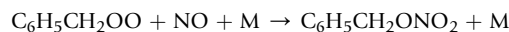
As can be seen in Table 4, with the exception of benzene most aromatic hydrocarbons react fairly rapidly with OH and have atmospheric lifetimes in the range of hours to days, so that atmospheric degradation will be initiated in fairly close proximity to the emission source. Benzene with a lifetime of several days can be transported to hundreds of kilometers.

Primary Reaction Steps

As indicated above, reaction with OH radicals dominates the tropospheric oxidation processes of aromatic hydrocarbons. As shown in Figure 2 the reaction proceeds by two pathways: (1) OH radical addition to the aromatic ring to form a hydroxycyclohexadienyl radical, i.e., an OH-aromatic adduct, which can thermally decompose back to reactants or undergo further reactions to products, and (2) H-atom abstraction from the C–H bonds of alkyl substituent groups.

The rate coefficient for attack at a methyl substitution is of the order of $6 \times 10^{-13} \text{ cm}^3 \text{ molecule}^{-1} \text{ s}^{-1}$. It generally accounts for ≤10% of the overall reaction and becomes increasingly less important with alkyl substitution, since the rate of OH addition to the ring increases. The radicals resulting from H-atom abstraction from alkyl groups of substituted

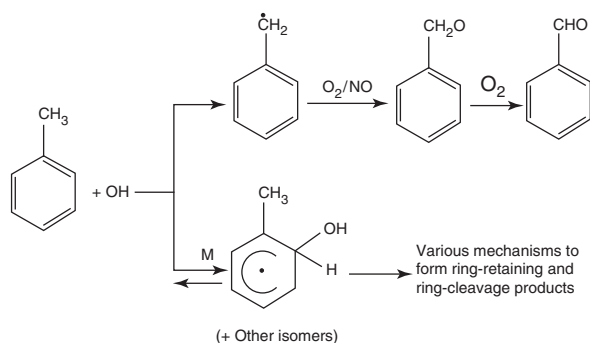
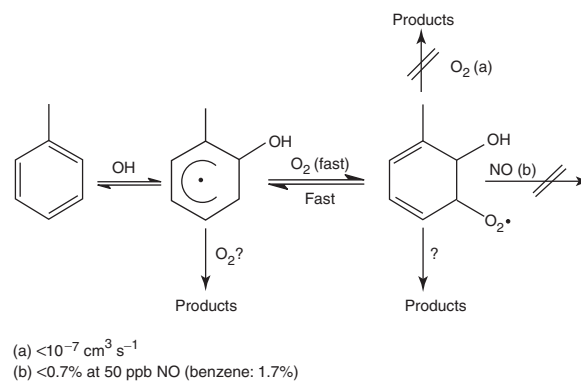
benzenes undergo reactions analogous to those for alkyl radicals generated from the oxidation of alkanes. The oxidation results mainly in the formation of ring-retaining aromatic aldehyde products (as shown in Figure 3 for toluene). Small yields of aromatic nitrate compounds can also be generated from reaction of the alkyl peroxy radicals with NO, e.g.,



Addition of OH to the aromatic ring is by far the major reaction pathway. Depending on the symmetry of the aromatic compounds, there can be up to six distinct addition points for the OH group. The position of attack relative to an existing substitution is favored in the following order: *ortho*- > *para*- > *ipso*- > *meta*-.

OH-aromatic adducts have been observed spectroscopically and the reactions of the OH-benzene, OH-toluene, and OH-xylene adducts with NO, NO₂, and O₂ have been studied. The adducts do not appear to react with NO, but reaction with O₂ and NO₂ is observed with rate coefficients between $(1.8\text{--}20) \times 10^{-16}$ and $(2.5\text{--}3.6) \times 10^{-12} \text{ cm}^3 \text{ molecule}^{-1} \text{ s}^{-1}$, respectively. In the atmosphere, reaction of the adduct with O₂ will dominate; however, in laboratory studies with high NO₂ concentrations, both reactions can be important.

The reactions of the OH-aromatic adducts with O₂ are complex, and various different mechanisms have

**Figure 2** Primary pathways in the OH-radical initiated oxidation of aromatic hydrocarbons using toluene as an example.**Figure 3** Reactions of the hydroxycyclohexadienyl radical (OH-aromatic adduct) with O₂.

been proposed to account for the observed products and experimental observations. The multitude of different hydroxy-peroxy radicals which can be generated by addition of O_2 to the OH-aromatic adduct are shown in Figure 4, taking toluene as an example.

Figure 5 summarizes various possible reaction pathways of the OH-aromatic-peroxy radicals, which have been postulated to explain the products observed in the OH-radical initiated oxidation of aromatic hydrocarbons using benzene (1) as an example. For reasons of simplicity and clarity, only formation of a β -hydroxy-peroxy radical (3) via the addition reaction of O_2 with the OH-benzene adduct (2) is shown, since this is considered to be the major pathway, not only for benzene but also for other alkyl-substituted benzenes.

Reaction of O_2 with the OH-aromatic adduct can also lead to products other than a hydroxy-peroxy radical (4). A phenolic product (4) and HO_2 can be formed by attack at the H-atom α - to the hydroxyl group. The attack of O_2 directly at the OH group may generate an aromatic oxide (5) and HO_2 . The aromatic oxide (5) is in equilibrium with the oxepin (6); photolysis of these compounds is also known to yield phenolic products (4), and reaction with OH is known to produce diunsaturated dicarbonyls (8), however, the route producing the aromatic oxide (5) is not considered to be important and has only been shown for completeness.

Such diunsaturated dicarbonyls can also be produced if the β -hydroxy-peroxy radical (3) reacts via 'conventional' atmospheric peroxy radical reactions, i.e., reaction with NO to form an oxy radical followed by ring opening to form (8). Recent work indicates that the reaction of OH-aromatic-peroxy radicals such as (3) with NO is slow and will not be

important under atmospheric conditions. In laboratory studies with high NO levels, the process may occur and these compounds have indeed been observed in some laboratory product studies.

The β -hydroxy-peroxy radical (3) can rearrange to form a peroxide-bridged bicyclic radical (9), which further adds O_2 to form a bicyclic peroxy radical. This reaction sequence is currently considered the major fate of β -hydroxy-peroxy radicals. In Figure 5, the peroxide-bridged bicyclic radical (9) reaction step leading to the oxy radical (10) involves the intermediateness of a bicyclic peroxy radical which reacts further with NO to form the oxy radical (10). Other reactions of the bicyclic peroxy radicals are possible and these are discussed below and shown schematically in Figure 6 for the bicyclic peroxy radical formed in the OH-initiated oxidation of toluene. Using a turbulent flow chemical ionization mass spectrometry technique the bicyclic peroxy radicals have now been detected and positively identified for a number of aromatic hydrocarbons and their reactions investigated under different NO_x conditions.

In the atmosphere reaction of the bicyclic peroxy radicals with NO to produce oxy radicals (10) will dominate. The major fate of these oxy radicals will be ring opening, as depicted in Figures 5 and 6, leading to monounsaturated dicarbonyls and α -dicarbonyls. The reaction of dienedials (such as (8) in Figure 5) with OH radicals is another possible route to 1,2-dicarbonyls and unsaturated dicarbonyl products; however, this route will be very minor, if not negligible, under atmospheric conditions.

Another option for the β -hydroxy-peroxy radical, (3) in Figure 5, is rearrangement to an epoxy-oxy radical (13), which can form either the unsaturated epoxy-ketone (14) or/and the ring-opened unsaturated epoxy dicarbonyl product (15). However, the epoxy-carbonyl species have only been observed in a few studies and there is still a lot of uncertainty as to the importance of this reaction route. Despite the uncertainty the channel is often included in aromatic oxidation schemes used in models to ensure carbon mass balance.

In recent years several new investigations, using mainly mass spectrometric techniques, have been reported that have attempted to address the question whether other important chemical processes are missing from the currently formulated aromatic hydrocarbon OH-radical initiated oxidation mechanisms. The results from these investigations support that apart from reactions of the bicyclic peroxy radicals with NO to form α -dicarbonyl and monounsaturated dicarbonyl products, products such as bicyclic nitrates, bicyclic peroxides, bicyclic carbonyls, and bicyclic alcohols (such as shown in Figure 6) may be formed, depending on the $RO_2(HO_2)/NO$ ratio in the experimental system. More investigations using different techniques are needed to verify and quantify these reaction channels.

In recent years, the possibility of ipso addition of the OH radical to alkylated benzenes, i.e., addition of OH at a position occupied by an alkyl substituent followed by ejection of the alkyl group (dealkylation) has been investigated. The process is illustrated in Figure 7 for the reaction of the OH radical with *o*-xylene. Whereas this pathway is known to occur in the liquid phase, it has always been assumed to be negligible in the gas phase with OH preferably adding to a nonsubstituted position.

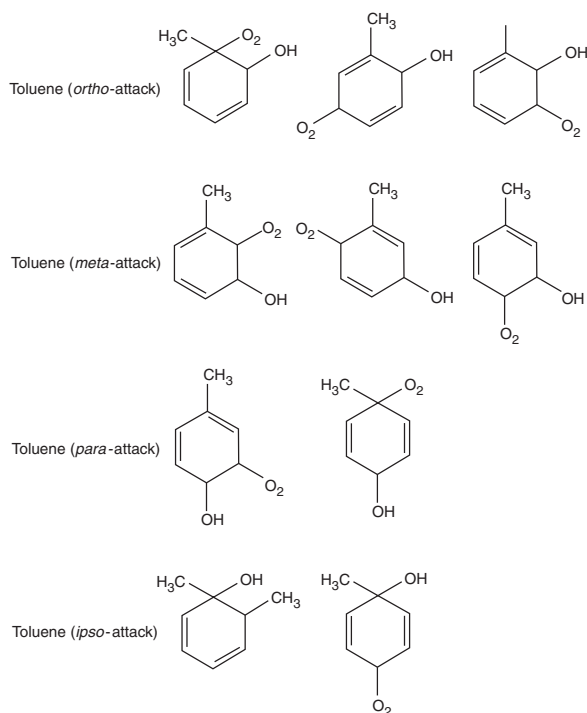


Figure 4 Possible peroxy radicals generated from OH radical addition to aromatic hydrocarbons taking toluene as an example.

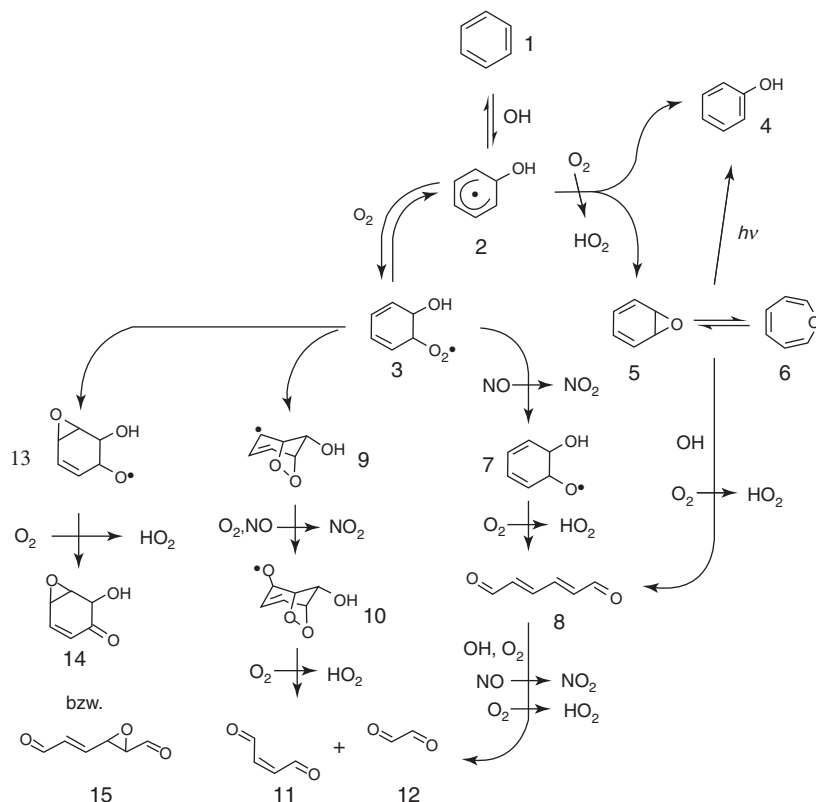


Figure 5 Representation of the various possible reaction pathways postulated to explain the products observed in the OH-radical initiated oxidation of aromatic hydrocarbons using toluene as an example. Only the *ortho* OH-radical addition pathway is shown.

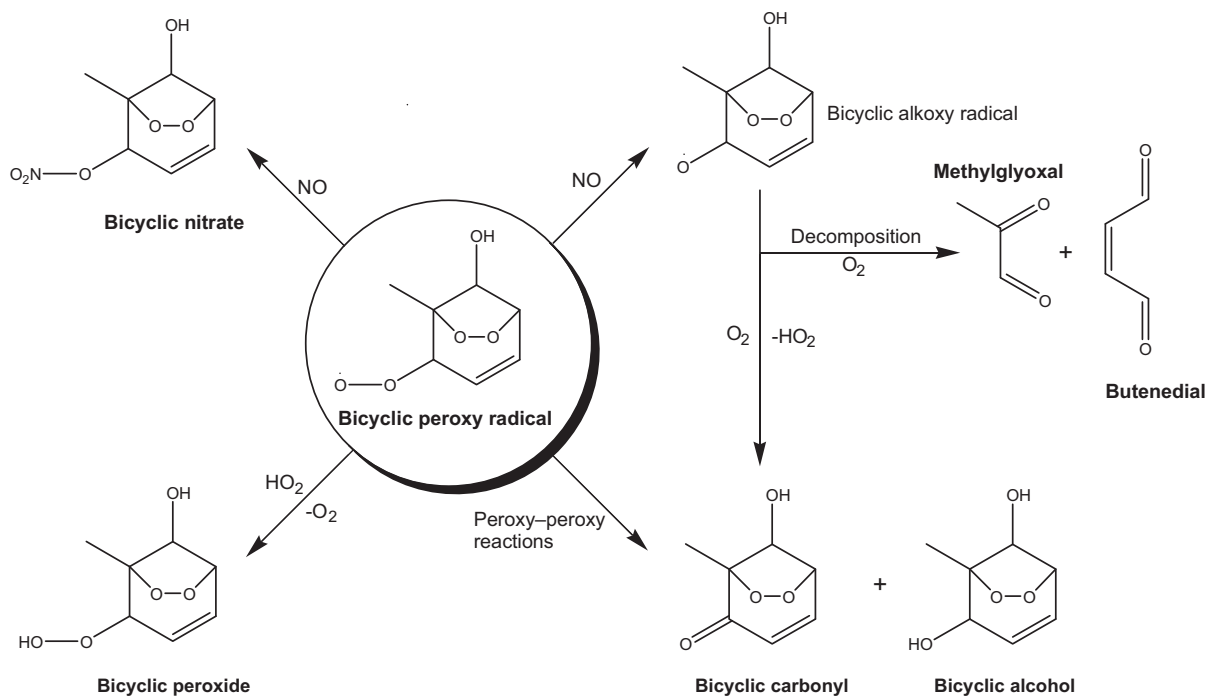


Figure 6 Simplified representation of the possible reactions of a bicyclic peroxy radical formed in the OH-radical initiated oxidation of toluene.

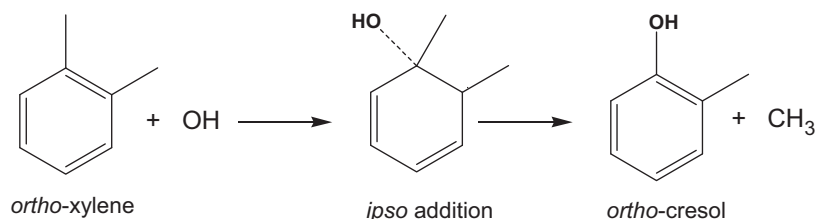


Figure 7 Ipso addition of OH to *o*-xylene followed by demethylation to form *o*-cresol.

However, for a fully substituted aromatic such as hexamethyl benzene OH ipso addition has been shown to be an important gas-phase reaction route. A flow tube study on the OH-initiated oxidation of toluene and the xylenes has reported yields of demethylated phenolic type compounds of between 5 and 10%, however, in an environmental chamber study of *m*-xylene and *p*-cymene an upper limit of 1% was reported for the dealkylation pathway. Therefore, it remains unclear whether the dealkylation pathway is a significant gas-phase oxidation pathway for moderately substituted and alkyl substituted aromatic compounds. However, with increasing degree in alkyl substitution evidence for this pathway should be searched for in experimental product investigations.

Product Yields

In this section, the typical products of the OH-initiated oxidation of aromatic hydrocarbons are discussed. A large number of product studies of the OH-initiated oxidation of aromatic hydrocarbons have been reported. Though many of these studies have focused on toluene, the most abundant aromatic hydrocarbon found in polluted air, numerous data exist for many other aromatic hydrocarbons, including some polycyclic aromatics. A large number of ring-opening products have been observed in the OH-initiated oxidation of aromatic hydrocarbons, but only some of these have been determined quantitatively. Table 5 shows the types of ring-retaining and ring-opening products identified and quantified in the OH-radical initiated oxidation of aromatic hydrocarbons taking toluene as example. Yields are not given, since these can vary quite substantially with the experimental conditions employed. Hydroxybenzenes (e.g., phenol from benzene and cresols from toluene) are major ring-retaining products formed by abstraction of a ring H atom with O₂ (Figure 5), with yields typically of around 60% for phenol from benzene, and 20% for

the cresols from toluene, and rank among the best-quantified products.

All of the possible monounsaturated dicarbonyls and α -dicarbonyls that can be formed in the OH-initiated oxidation of benzene, toluene, and the xylene and trimethylbenzene isomers have now been detected and quantified. All the current aromatic hydrocarbon OH oxidation mechanisms predict a 1:1 stoichiometric relationship between 1,2-dicarbonyl products (such as glyoxal (12) in Figure 5) and unsaturated dicarbonyl products (such as butenedial (11) in Figure 5). Some investigations, however, have found that the unsaturated dicarbonyl products are often observed in significantly lower yields than the 1,2-dicarbonyls, their presumed product partners. However, many of the unsaturated dicarbonyl products are highly reactive and it is thought that secondary chemistry is responsible for the nonstoichiometric product yields. Many of the multitude of reported products observed in low yield are secondary or tertiary in nature, being formed by the further oxidation of stable (nonradical) intermediates of the atmospheric oxidation of the parent aromatic hydrocarbons.

One problem encountered in all existing product studies of the OH-initiated oxidation of aromatic hydrocarbons is the deficient carbon balance. Generally, only between 50 and 60% of the carbon, which reacts with OH radicals is identified as products. The main reason for this is that all the proposed and identified products of the oxidation are highly reactive themselves; most of the products, react about one order of magnitude faster with OH radicals than with the parent aromatic hydrocarbon. This often results, as stated above, in low yields of these products, which makes them difficult to detect and even more difficult to quantify. Such large differences in the reactivity between products and reactants are not observed in any other class of VOCs present in the atmosphere. The high reactivity of aromatic hydrocarbon photo-oxidation products can be explained by the three 'double bonds' in the benzene ring. In the aromatic benzene ring, these 'double bonds' are fully conjugated and consequently very stable. Once the aromaticity is destroyed, however, nonaromatic double bonds are formed, making the products highly reactive.

Table 5 Types of products quantified in the OH-radical initiated oxidation of aromatic hydrocarbons: example for toluene

Ring-retaining products	Ring-opened products
Benzaldehyde	Maleic anhydride ^a
<i>o</i> -, <i>m</i> -, <i>p</i> -Cresol	α -Angelica lactone ^a
2-Methyl- <i>p</i> -benzoquinone	2-Oxo-2-pentenal
Benzyl alcohol	Methylglyoxal
<i>o</i> -, <i>m</i> -, <i>p</i> -Nitrotoluene	Glyoxal
Bicyclic carbonyls, alcohols, peroxides, nitrates	

^aProducts from reactions of straight-chain monounsaturated dicarbonyls.

Uses of Aromatic Hydrocarbon Oxidation Mechanisms and Limitations

Chemical oxidation mechanisms of VOCs are required for air quality simulation models, which are used to predict the formation of secondary pollutants such as ozone and also secondary aerosol formation. Due to computing limitations the model generally uses lumped species or highly

parameterized mechanisms and it is known that the mechanisms, in particular for aromatic hydrocarbons, are not fit for purpose in many respects. One way to obtain a simplified or lumped mechanism that is to use an explicit mechanism which can be systematically reduced such. The explicit mechanism has of course to be fit for purpose for the task to which it will be applied. The most explicit mechanisms for aromatic hydrocarbons are those contained in the Master Chemical Mechanism (MCM), which is maintained by the University of Leeds, UK (MCM: <http://mcm.leeds.ac.uk/MCM/>). The mechanisms have been largely developed from experimental kinetic and product data and are frequently updated to incorporate mechanistic new findings. The mechanisms are designed to provide an explicit representation of the initial OH reaction step for aromatics, as well as all subsequent reactions believed to play a role in the ultimate effect of aromatic oxidation chemistry on air quality.

The MCM aromatic oxidation mechanisms have been tested against environmental chamber experiments in which atmospheric conditions were simulated as closely as possible. Unfortunately, the MCM mechanisms fail to correctly predict ozone production rates, as well as OH radical levels. In the case of toluene, the MCM overpredicts the ozone concentrations by ~55% and underpredicts the OH radical production by ~44%. The simulation experiments highlight that some fundamental aspects of the oxidation mechanisms of aromatic hydrocarbons are not yet fully understood and that further improvement in the MCM are required before it can reliably be used to understand and reduce air quality problems related to elevated levels of aromatic compounds in the atmosphere.

Aerosol Formation from the Oxidation of Aromatic Hydrocarbons

Epidemiological studies have reported an association between daily mortality and fine particles. There are two general classes of PM: primary and secondary. Carbonaceous (organic) aerosol constitutes 20–90% of all submicron particles in the atmosphere, and up to 80% of this is classified as SOA generated in the photochemical oxidation of organic compounds present in the atmosphere. Air-quality models currently underpredict the concentrations of organic aerosol in the atmosphere, especially in the summer and in urban areas showing that our understanding of anthropogenic SOA formation is incomplete.

The fraction of atmospheric SOA originating from the oxidation of the aromatic hydrocarbons is not known with certainty; however, it has been concluded from field and laboratory observations that aromatic hydrocarbons probably play a significant and possibly domineering role in the formation of SOA in urban areas. Toluene (methylbenzene) and other light aromatics are thought to be the dominant anthropogenic SOA precursors. The emissions of toluene in the United States are approximately 1 teragram of carbon per year, and in addition to being an important SOA precursor itself, toluene is often used as a model system representative of the SOA formation from other aromatic VOCs.

Studies of SOA formation from VOCs are typically performed in large, Teflon film reactors (often referred to as smog or environmental chambers). Performing SOA formation experiments in chambers have, however, inherent difficulties

since gas-particle-wall partitioning can occur which can affect both the yield and the chemical composition of the SOA. Recent investigations suggest that wall processing of the SOA may be more of a problem than previously thought. The aerosol mass yields from the photooxidation of aromatic hydrocarbons reported in the literature vary widely and direct comparisons are difficult since, apart from potential wall complications, the experiments are conducted under different conditions and with different aerosol mass loadings which are now known to lead to differences in the measured aerosol yields. It is also still an open question as to whether the aerosol yields observed in laboratory experiments are directly transferable to atmospheric conditions.

Recent laboratory SOA formation yield measurements coupled to gas-phase yield measurements, indicate that approximately 20% of the SOA of benzene, toluene, and *m*-xylene can be ascribed to the secondary reactions of phenolic primary products. The yields for aromatic hydrocarbons as for other VOCs have been shown to be highly sensitive to the NO_x levels, with higher SOA yields being observed under low-NO_x conditions. This dependence of the SOA formation on the NO_x level has been proposed to be the result of differences in concentrations of different oxidants (OH, O₃, and NO₃) or in changes in the peroxy radical chemistry. At high RO₂(HO₂)/NO ratios peroxy-peroxy radical reactions will dominate the fate of the peroxy radicals rather than reaction with NO leading to products more susceptible to SOA formation such as organic hydroperoxides (ROOH). Measurements in studies performed under high-NO_x conditions suggest there may also be a 'rate effect' in SOA formation, in which SOA yields are higher when the oxidation rate is faster. The NO_x dependence of SOA yields is thus a crucial parameter for atmospheric modeling, which is currently very poorly understood. Other parameters, which play a role in determining the SOA yield from aromatics are relative humidity, the presence or absence of seed aerosols, and temperature.

There have been very few studies on the molecular composition of the SOA from aromatic hydrocarbons. Studies to date indicate, i.e., the reactions of ring-opened first-generation products, which are probably responsible for a large fraction of the aerosol formation observed in the photooxidation of aromatic compounds. The chemical nature of only a small fraction (15–30%) of the secondary aerosol mass has been identified to date. It is thought that organic acids formed from the oxidation of α,ω -dialdehydes initially to α,ω -oxyacids and further to α,ω -diacids may be important unidentified components of the aerosol. Evidence has now been found for the presence of bicyclic hydroperoxides in aromatic SOA. Since the organic compounds present in the aerosols have obviously a low vapor pressure, and many are not commercially available, there are considerable experimental hurdles with regard to collection, identification, and quantification, which still need to be overcome before a full aerosol composition specification will be possible.

The formation of SOA from the gas-phase oxidation of aromatic compounds discussed here refers to that which would take place in the first few hours after emission. In the atmosphere, however, SOA will continue to evolve chemically over its atmospheric lifetime of typically 5–12 days. Simulating chemical aging over an extended period of time in laboratory

investigations of SOA formation is a major experimental challenge. At the time of writing, procedures for studying SOA aging in laboratory chambers are in the early stages of development.

Outlook

Although much new and encouraging information concerning the mechanisms of the atmospheric oxidation of aromatic hydrocarbons has emerged in recent years, in particular from mass spectrometric techniques, vital information concerning the actual nature of the ring-opening process which is necessary for realistic model construction is still uncertain. The wealth of new kinetic and mechanistic information obtained, however, allows limits to be put upon some of the various possible ring-cleavage reaction channels. Since the database has been considerably broadened, it is likely that in the near future, with some further refinements in the database, it will be possible to obtain a reasonably accurate description for the photooxidation of aromatic hydrocarbons. This may be achieved either by an experimental breakthrough or by optimization of a 'multi-channel' reaction mechanism (utilizing new kinetic and mechanistic data) to fit a set of well-defined simulation experiments suitable for different atmospheric conditions.

See also: Tropospheric Chemistry and Composition: Aliphatic Hydrocarbons; Biogenic Hydrocarbons; Volatile Organic Compounds Overview: Anthropogenic.

Further Reading

- Andino, J.M., Smith, J.N., Flagan, R.C., Goddard, W.A., Seinfeld, J.H., 1996. Mechanism of atmospheric photooxidation of aromatics: a theoretical study. *Journal of Physical Chemistry* 100, 10967–10980.
- Arey, J., Obermeyer, G., Aschmann, S.M., Chattopadhyay, S., Cusick, R., Atkinson, R., 2009. Dicarbonyl products of the OH radical-initiated reaction of a series of aromatic hydrocarbons. *Environmental Science and Technology* 43, 683–689.
- Atkinson, R., 1994. *Gas-Phase Tropospheric Chemistry of Organic Compounds*. American Institute of Physics, Washington, DC. ISBN 9781563963407.
- Atkinson, R., 2000. Atmospheric chemistry of VOCs and NO_x. *Atmospheric Environment* 34, 2063–2101.
- Birdsall, A.W., Elrod, M.J., 2011. Comprehensive NO-dependent study of the products of the oxidation of atmospherically relevant aromatic compounds. *Journal of Physical Chemistry A* 115, 5397–5407. <http://dx.doi.org/10.1021/jp2010327>.
- Bloss, C., Wagner, V., Jenkin, M.E., Volkamer, R., Bloss, W.J., Lee, J.D., Heard, D.E., Wirtz, K., Martin-Reviejo, M., Rea, G., Wenger, J.C., Pilling, M.J., 2005. Development of a detailed chemical mechanism (MCMv3.1) for the atmospheric oxidation of aromatic hydrocarbons. *Atmospheric Chemistry and Physics* 5, 641–664.
- Bohn, B., Zetzsch, C., 1999. Gas-phase reaction of the OH-benzene adduct with O₂: reversibility and secondary formation of HO₂. *Physical Chemistry Chemical Physics* 1, 5097–5107.
- Calvert, J., Atkinson, R., Becker, K., et al., 2002. *The Mechanisms of Atmospheric Oxidation of Aromatic Hydrocarbons*. Oxford University Press, New York.
- Finlayson-Pitts, B.J., Pitts, J. (Eds.), 1999. *Chemistry of the Upper and Lower Atmosphere*. Academic Press, San Diego.
- Forstner, H.J.L., Flagan, R.C., Seinfeld, J.H., 1997. Secondary organic aerosol from the photooxidation of aromatic hydrocarbons: molecular composition. *Environmental Science and Technology* 31, 1345–1358.
- Hewitt, C.N. (Ed.), 1999. *Reactive Hydrocarbons in the Atmosphere*. Academic Press, San Diego, CA.
- Hildebrandt, L., Donahue, N.M., Pandis, S.N., 2009. High formation of secondary organic aerosol from the photo-oxidation of toluene. *Atmospheric Chemistry and Physics Discussions* 9, 693–733.
- Klotz, B., Barnes, I., Becker, K.H., Golding, B.T., 1997. Atmospheric chemistry of benzene oxide/oxepin. *Journal of the Chemical Society London Faraday Transactions* 93, 1507–1516.
- Klotz, B., Sørensen, S., Barnes, I., 1998. Atmospheric oxidation of toluene in a large-volume outdoor photoreactor: in situ determination of ring-retaining product yields. *Journal of Physical Chemistry* 102, 10289–10299.
- Loza, C.L., Chhabra, P.S., Yee, L.D., Craven, J.S., Flagan, R.C., Seinfeld, J.H., 2012. Chemical aging of *m*-xylene secondary organic aerosol: laboratory chamber study. *Atmospheric Chemistry and Physics* 12, 151–167.
- Nakao, S., Clark, C., Tang, P., Sato, K., Cocker III, D., 2011. Secondary organic aerosol formation from phenolic compounds in the absence of NO_x. *Atmospheric Chemistry and Physics Discussions* 11, 2025–2055. <http://dx.doi.org/10.5194/acpd-11-2025-2011>.
- Ng, N.L., Kroll, J.H., Chan, A.W.H., Chhabra, P.S., Flagan, R.C., Seinfeld, J.H., 2007. Secondary organic aerosol formation from *m*-xylene, toluene, and benzene. *Atmospheric Chemistry and Physics* 7, 3909–3922.
- Odum, J.P., Jungkamp, P.W., Griffin, R.J., et al., 1997. Aromatics, reformulated gasoline and atmospheric organic aerosol formation. *Environmental Science and Technology* 31, 1890–1897.
- Piccot, S.D., Watson, J.J., Jones, J.W., 1992. A global inventory of volatile organic compound emissions from anthropogenic sources. *Journal of Geophysical Research* 97, 9897–9912.
- Smith, D.F., Kleindienst, T.E., McIver, C.D., 1999. Primary product distributions from the reaction of OH with *m*-, *p*-xylene, 1,2,4- and 1,3,5-trimethylbenzene. *Journal of Atmospheric Chemistry* 34, 339–364.
- Yu, J.Z., Jeffries, H.E., Sexton, K.G., 1997. Atmospheric photooxidation of alkylbenzenes. 1 Carbonyl product analyses. *Atmospheric Environment* 31, 2261–2280.

Biogenic Hydrocarbons

A Guenther, Pacific Northwest National Laboratory, Richland, WA, USA

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

All living organisms produce volatile organic compounds (VOCs) including some that are emitted into the atmosphere where they have a significant impact on the chemical composition of the atmosphere including constituents, such as organic aerosol and ozone, that influence climate. These emissions respond to climate and land cover change, and there may be complex interactions with important climate change implications. Many VOCs are found in living organisms but relatively few are emitted at substantial rates, primarily from vegetation foliage. These include terpenoids (e.g., isoprene), hydrocarbons (e.g., ethene), and oxygenated VOCs (e.g. methanol). Some biogenic VOC emissions also have important ecological roles.

Introduction

All living organisms produce hydrocarbons, which consist solely of hydrogen and carbon, and other volatile organic compounds (VOCs) that can also contain elements such as oxygen, nitrogen, or sulfur. The term VOC will refer here to all VOCs other than methane. The major categories of VOC chemical species (terpenoid, oxygenated, and other) and the important sources (vegetation foliage) are described, and the atmospheric and ecological factors that control the various types of emissions are discussed. Most of the VOCs emitted into the atmosphere are of biogenic (produced and emitted by biological organisms) origin. The remainder, primarily from wood and fossil fuel combustion, is a small fraction of the global total but greatly dominates in urban and industrial areas and in remote areas where biomass burning occurs. The known and suspected ecological and physiological roles of biogenic VOC are noted and their impact on the environment is considered. Biogenic VOCs can have a significant impact on the chemical composition of the atmosphere and are strongly influenced by environmental conditions. As a result, it is likely that global emissions of these compounds will change in response to global climate and land cover change. This raises the potential for a feedback coupling that could significantly perturb the Earth system.

Compounds

Tens of thousands of organic compounds have been identified in plant tissues and more are discovered every year. Many of these compounds have very low volatility, which prevents or greatly limits their emission to the atmosphere. In addition, some organic compounds are stored in plant structures that present substantial barriers to emission to the atmosphere. As a result, a relatively small number of all organic compounds found within plants are emitted at rates that can significantly influence the atmosphere.

Terpenoid VOCs

The terpenoids are an important class of organic compounds that include hemiterpenes (containing 5 carbon atoms), monoterpenes (10 carbon atoms), sesquiterpenes (15 carbon atoms), and diterpenes (20 carbon atoms). Only two

hemiterpenes, isoprene and methyl butenol, are known to be emitted in significant quantities. Isoprene has the single largest contribution to the total global VOC emission. There are about a thousand monoterpene compounds but only less than 20 are emitted into the atmosphere at significant rates. The largest contributor to the global monoterpene emission is α -pinene. Other monoterpenes that are commonly observed as plant emissions include β -pinene, Δ^3 -carene, D-limonene, camphene, ocimene, myrcene, sabinene, β -phellandrene, and p-cymene. Relatively little is known about the emission of the approximately 3000 sesquiterpene compounds found in plants. Their emission is a relatively small component of global biogenic VOCs but may be important for aerosol production. None of the approximately 2000 diterpenes are known to be emitted into the atmosphere in significant amounts but this may be due to a lack of observations targeting these compounds.

Oxygenated VOCs

Biogenic VOCs include a large variety of oxygenated compounds including alcohols (e.g., methanol), aldehydes (e.g., acetaldehyde), ketones (e.g., acetone), acids (e.g., formic acid), ethers (e.g., 1,8 cineole), and esters (hexenyl acetate). The molecular weights of these compounds range from less than 32 g mol^{-1} (e.g., methanol and formaldehyde) to over 100 g mol^{-1} (e.g., decanal and methyl heptenone). The larger compounds are more likely to contribute to aerosol formation or growth. Methanol, ethanol, acetone, and acetaldehyde are thought to be the oxygenated VOCs with the highest global emission rates and have a significant impact on the chemical composition of the atmosphere. There are not enough information to determine whether there are substantial emissions of many oxygenated VOC species.

Other VOCs

There are a number of other organic compounds including alkanes (e.g., ethane), alkenes (e.g., ethene), arenes (e.g., toluene), sulfur compounds (e.g., dimethyl sulfide), and nitrogen compounds (e.g., hydrogen cyanide). Of these compounds, ethene has the highest global emission rate. Some of the other compounds are rarely emitted but emissions can be high for certain conditions and locations. Dimethyl sulfide is a volatile organic sulfur compound that is produced and

emitted by terrestrial vegetation and marine phytoplankton at rates that can have a significant impact on the chemistry of remote atmospheres.

Sources and Driving Variables

Vegetation is the major source of VOC emission into the atmosphere. These emissions are primarily from foliage (leaves and needles). The remainder is from other live tissues including stem, cortical, root, and reproductive tissues. Trees are the major source of global emissions although shrubs and nonwoody plants are important sources of at least some compounds. Microorganisms and animals are a minor source of biogenic VOC emission. Isoprene and acetone are the dominant VOCs emitted in human breath.

VOC Production and Emission Pathways

There are many different biogenic VOC production/emission sources including chloroplasts, specialized defense, unspecialized defense, metabolic by-products, decaying and drying vegetation, plant growth hormones, and floral scents. Each pathway is responsible for the emission of more than one type of VOCs and some VOCs are emitted by more than one pathway.

Chloroplast emissions dominate the global VOC flux. Even though only 30% of all plant species exhibit this type of emission, the emission rates are typically much higher than those of other emission processes. Isoprene is the major chloroplast emission from most landscapes, but some regions are dominated by chloroplast emissions of methyl butenol or monoterpenes (e.g., α -pinene). Isoprene is produced in chloroplasts by a unique enzyme, isoprene synthase, and depends on the availability of photosynthetic carbon, but is distinct from photorespiration. Isoprene penetrates the intercellular space of the leaf and then exits the plant via the stomata. Isoprene synthase activity is controlled primarily by leaf temperature and light (photosynthetically active radiation). Diurnal variations range over several orders of magnitude and are near zero in the dark. The temperature and light conditions that a leaf has been exposed to during the previous hours to days can significantly influence emissions. Leaf age, phenology, nutrient levels, carbon dioxide concentration, and water stress are other factors that control isoprene emission.

Many plants produce and store terpenoid compounds within specialized tissues that act as a physical barrier to insects and pathogens and/or as a feeding deterrent if consumed. These compounds are primarily monoterpenes (C_{10}) and diterpenoids (C_{20}) with a large number, but small quantity, of sesquiterpenes (C_{15}). The biochemistry of these compounds, and the genetic and ecological controls, has been investigated due to the importance of this plant defense mechanism for economically significant plants such as pine trees. Although there are thousands of terpenoids, the total emission from stored terpenoid pools is dominated by contributions from less than 10 compounds. Diurnal variations in emissions of these stored compounds range from about a factor of two to four and are primarily controlled by temperature. The relationship between emission and temperature is exponential and is dependent on the compound and the resistance properties in the plant.

Some of the VOCs produced in unspecialized plant tissues also play a role in defending plants against disease and herbivory. For example, emissions of ethane, ethanol, methyl salicylate, octanone, and methoxyphenol are associated with resistance of plants to herbivore infestations. These compounds can act by repelling pests or attracting predators. The hexenal family compounds (2-hexenal, 3-hexenol, 3-hexenal, 3-hexenyl acetate, hexanal, and hexanol) may have antibiotic properties and can be emitted by damaged plants at very high rates. There is a reasonably good understanding of the mechanisms controlling the production of hexenal family compounds, which are typically triggered by membrane damage due to mechanical wounding or the presence of pathogens.

Ethene is a volatile hormone that controls numerous aspects of plant growth and development, including fruit ripening, seed germination, flowering, and senescence. It can also trigger plant defense and internal concentrations are greatly increased by a variety of stresses. The biochemistry of ethene production is better understood than that of most VOCs but investigations of ethene emission have been primarily directed at understanding its role as a plant hormone and provide little information for numerical emission modeling. Ethene production is widespread in plants and is likely a significant emission from most landscapes. It may be particularly sensitive to changes in ecosystem stress.

Large VOC emissions are observed when vegetation is cut or when dried foliage is hydrated. Thus, the initial cutting of live vegetation as well as long-term emissions from drying and dead vegetation produces significant emissions. The dominant oxygenated VOC emissions include both reactive (e.g., acetaldehyde and hexenal family) and less reactive (e.g., methanol, acetone, and butanone) compounds. These fluxes dominate VOC emissions in agricultural areas during harvesting. Other human activities, including timber harvesting, lawn mowing, and rangeland management practices, may significantly enhance VOC emissions and should be characterized in order to understand the importance and impact of this emission source.

Floral scents are composed of a large variety of VOCs including alkanes, alcohols, esters, aromatics, nitrogen compounds, monoterpenes, and sesquiterpenes. Flowers are likely to be the dominant biogenic source of some of these compounds. Although flowers are expected to be a minor component of annual global VOC emissions, flowering may dominate VOC emissions in some locations at certain times.

At least some portion of the total emission of a number of VOCs (e.g., methanol, formaldehyde, acetone, propene, butene, isoprene, formic, and acetic acid) appears to be emitted by processes other than those described above. These emissions may simply represent the leakage of plant metabolites. For example, methanol can be produced in plants as a result of the enzymatic demethylation of a component of the pectin in plant cell walls. This process is an important step in the elongation process that allows plant cell walls to grow. The biochemical mechanisms for ethanol and acetaldehyde production are also reasonably well understood. Production of these two compounds by plants is expected under anaerobic conditions such as those experienced by roots in flooded soils, wet crevices of damaged trees, and by leaves under certain stressed conditions. Production pathways for acetic acid, acetone, formaldehyde, and formic acid in plant cells have been proposed but need to be confirmed.

Geographical and Seasonal Distributions

The global annual emission of biogenic VOC is estimated to be greater than 1×10^{15} g. Over 99% is emitted by terrestrial vegetation, primarily from foliage. A large fraction of this total is associated with tropical regions due to the high temperature and light regimes, often coupled with a high plant biomass, and emissions that occur throughout the entire year. Regions of higher (northern and southern) latitudes exhibit a pronounced seasonal cycle in the physiological activity of the vegetation resulting in a maximum emission during the summer time when temperature and light intensity are highest. Emission variations are based on environmental, e.g., light and temperature, conditions and land cover characteristics including foliar biomass, and plant species distributions. Emissions from landscapes without vegetation are negligible. For landscapes with vegetation, emissions vary more than the order of magnitude for different locations. Variations of this magnitude are also observed for different seasons at the same location.

Ecological and Physiological Roles

The production of VOC is a significant resource allocation for biological organisms, which leads to the question of why these compounds are lost into the atmosphere. Three possible reasons for biogenic VOC emissions are that they provide a significant benefit to the organism, they represent the loss of an unneeded metabolic by-product, or they are unintentional losses that result from having volatile compounds in a leaky structure. The role of each of these possibilities is discussed below.

Biogenic VOC emissions have an important role in at least some biological organisms. Flowering plants use VOC to attract pollinators. Insects and animals use VOC as pheromones for a variety of signaling activities. Plants can also use VOC to signal other parts of a plant, or even other plants, to begin or end specific growth phases such as flowering. Plants may also use VOC to repel pests, either as an antibiotic or by making the plant less appetizing. The emission of a volatile compound can also be a convenient way of removing a toxic substance produced by an organism. This is particularly useful for an immobile organism that does not have the ability to move away from the waste it has excreted. Examples of potentially toxic compounds that can be removed as VOC emissions include formaldehyde and acetone. Other beneficial roles that have been proposed for biogenic VOC include protecting plants from dangerously high leaf temperatures and modifying the chemical composition of the atmosphere in order to reduce the deposition of toxic compounds, such as ozone, or increase the deposition of needed nutrients, such as some nitrogen compounds.

Biological organisms metabolize organic substrates to produce needed cellular components. These processes often result in the production of metabolic by-products that are not needed by the organism. It is generally expected that organisms have pathways to recycle these by-products into other useful compounds but there may be situations where the organisms simply expel the compound into the atmosphere. The production of some VOCs may protect plants from extreme environmental conditions. This includes protection from high

light intensities, by dissipating excess reducing power, and high temperature, by stabilizing plant membranes.

All organisms exchange gases, e.g., oxygen, water vapor, and carbon dioxide, with their environment. In the process of doing so it is likely that at least some VOCs are unintentionally lost to the atmosphere. In addition to losses through the main gas transport pathways, i.e., the stomata of plants or the mouths of animals, there is the possibility for losses through other structures such as plant cell walls or the skin of animals. These losses would be especially large when a normally airtight structure becomes physically damaged and exposed to the environment. Some plant functions may benefit from having a highly volatile compound that can be moved rapidly to combat stressful situations that arise quickly such as rapidly fluctuating high temperatures or insect infestation. A VOC would be a good candidate for this function but would be susceptible to the unintentional loss of that compound.

Environmental Impact

Atmospheric Chemical Composition

The presence of biogenic VOC in the atmosphere has been recognized for centuries due to the noticeable odor of some of these compounds. For example, hexenol is characteristic of the smell of freshly cut grass and α -pinene contributes to the odor of pine trees. Scientific investigations of biogenic hydrocarbons in the atmosphere were initiated in the Soviet Union in the late 1920s. It was initially thought that these highly flammable compounds might have some military applications. Research in North America began several decades later and was initially focused on the potential role of biogenic hydrocarbons in atmospheric haze. This work continues to be an active area of research due to the recognition that the production of organic aerosol from biogenic VOC may greatly impact the global radiation balance. Of particular interest is the finding that anthropogenic pollution enhances the formation of organic aerosol from biogenic VOC and so contributes to radiative forcing. A major focus for investigations of biogenic VOC during the past four decades has been determining if these compounds have a role in tropospheric ozone formation. This work has demonstrated that a good understanding of biogenic VOC emissions is needed for developing regulatory control strategies for both rural and urban areas in some regions. Other studies have shown that oxygenated VOCs emitted from vegetation contribute to oxidant production in the upper troposphere. It has also been shown that biogenic sources of organic acids, from either primary sources or atmospheric oxidation, can contribute to precipitation acidity.

Carbon Balance

VOC emissions are a minor but significant pathway for the flow of carbon between an ecosystem and the atmosphere in at least some landscapes. Since the fate of most VOCs is oxidation to CO_2 , these CO_2 precursors may need to be considered for an accurate accounting of global CO_2 sources and sinks. The total annual VOC flux of over 1×10^{15} g of carbon is significant in comparison to the estimated annual global atmospheric CO_2 increase of about 3×10^{15} g of carbon. Net CO_2 fluxes between terrestrial

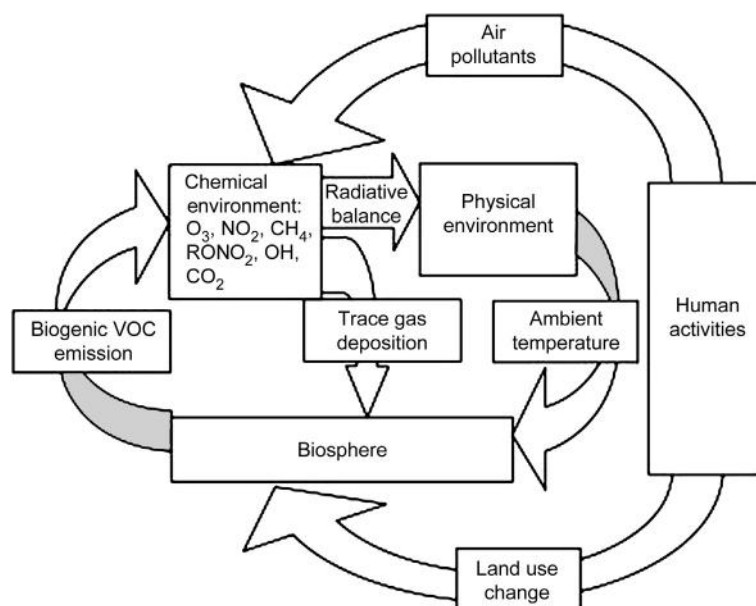


Figure 1 Illustration of a potential global feedback coupling driven by changes in biogenic VOC emissions.

landscapes and the atmosphere at regional scales are estimated using complex models of net ecosystem production, which treat explicitly carbon uptake by photosynthesis and losses through autotrophic and heterotrophic respiration. Some of these models are being developed to treat explicitly carbon losses through VOC emission since they are of similar magnitude, and have similar levels of uncertainty, as other terms that now receive explicit treatment (e.g., cement production).

Global Change and Possible Feedbacks

Biogenic VOC emissions can influence the chemical composition of the atmosphere, including radiatively active constituents that can modify the physical environment resulting in a global warming or cooling. The chemical and physical environment, in turn, can significantly influence the biosphere through nonlinear relationships with the ecological and physiological processes controlling VOC emission. This interaction may thus result in a global feedback coupling involving the chemical and physical climate system and operating over a wide range of time and spatial scales. The major components of this system are illustrated in **Figure 1** and are discussed below.

Interactions between VOC emissions and the global chemical (e.g., ozone, OH, methane, CO₂, NO₂, and RONO₂) environment are complex and nonlinear and therefore difficult to predict. Increases in emissions of some VOCs are likely to result in increases of gases that can increase radiative forcing (leading to warming) while others, such as those that can form organic particles, may decrease radiative forcing (leading to cooling). Due to the large differences in emission rates associated with different vegetation types, there is a substantial potential for land use change to influence biogenic emissions. Since woody plants tend to have much higher isoprene and monoterpene emission rates, compared to crops and grasses, it might be presumed that deforestation would greatly reduce biogenic

VOC emissions. However, there is a tendency for higher isoprene emissions from the woody plants (shrubs and sun tolerant trees) that replace a closed canopy forest. In addition, oxygenated VOC emissions tend to be higher from degraded and deforested landscapes. There is an equally large potential for perturbed biogenic VOC emissions as a result of climate change. Biogenic VOC emissions are very sensitive to temperature and an increase of a few degrees could lead to increase in emissions of more than 25%. The overall result of expected future land use and climate change is an increased biogenic VOC production that could result in significant perturbations in trace gas distributions and global biogeochemical cycles.

See also: Climate and Climate Change: Carbon Dioxide. Global Change: Biospheric Impacts and Feedbacks. Land-Atmosphere Interactions: Trace Gas Exchange. Tropospheric Chemistry and Composition: Volatile Organic Compounds Overview: Anthropogenic.

Further Reading

- Duhl, T.R., Helmig, D., Guenther, A., 2008. Sesquiterpene emissions from vegetation: a review. *Biogeosciences* 5, 761–777.
- Guenther, A.B., Jiang, X., Heald, C., Sakulyanontvittaya, T., Duhl, T., Emmons, L., Wang, X., 2012. The Model of Emissions of Gases and Aerosols from Nature version 2.1 (MEGAN2.1): an extended and updated framework for modeling biogenic emissions. *Geoscientific Model Development* 5, 1471–1492.
- Laohawornkitkul, J., Taylor, J.E., Paul, N.D., Hewitt, C.N., 2009. Biogenic volatile organic compounds in the Earth system. *New Phytologist* 183, 27–51.
- Niinemets, U., Monson, R. (Eds.), 2013. *Biology, Controls and Models of Tree Volatile Organic Compound Emissions*. Springer, Netherlands.
- Stavrakou, T., Guenther, A., Razavi, A., Clarisse, L., Clerbaux, C., Coheur, P.F., Hurtmans, D., Karagulian, F., De Mazière, M., Vigouroux, C., Amelynck, C., Schoon, N., Laffineur, Q., Heinesch, B., Aubinet, M., Rinsland, C., Müller, J.F., 2011. First space-based derivation of the global atmospheric methanol emission fluxes. *Atmospheric Chemistry and Physics* 11, 4873–4898.

Cloud Chemistry

P Herckes, Arizona State University, Tempe, AZ, USA

JL Collett, Jr., Colorado State University, Fort Collins, CO, USA

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

This article provides an overview of the chemical composition of clouds and the chemistry occurring within cloud and fog droplets. The processes by which chemical species enter cloud droplets from the gas and aerosol phase are summarized. The most prominent aqueous chemical processes occurring in clouds and fogs are detailed, including sulfur oxidation and complexation, oxidant chemistry, and secondary organic aerosol formation. Some emerging topics like biological materials are introduced. The article closes with a discussion of the role of clouds in atmospheric deposition mechanisms.

Introduction

Clouds occupy a small but important fraction of the total volume of the troposphere. In addition to playing a key role in the global hydrologic cycle and influencing atmospheric radiative transfer, clouds interact with a variety of chemical species. Together with gases and particles, clouds comprise a complex multiphase system. Clouds act both as reactors for the

production of new chemical species and as vectors for particle and trace gas removal, via wet deposition and direct deposition of cloud and fog drops to the surface.

Figure 1 provides an overview of several important processes in the multiphase atmospheric system. Clouds and fogs (clouds in contact with Earth's surface) interact with both aerosol particles and trace gases. The incorporation of particles and gases into cloud drops is the key step in determining the

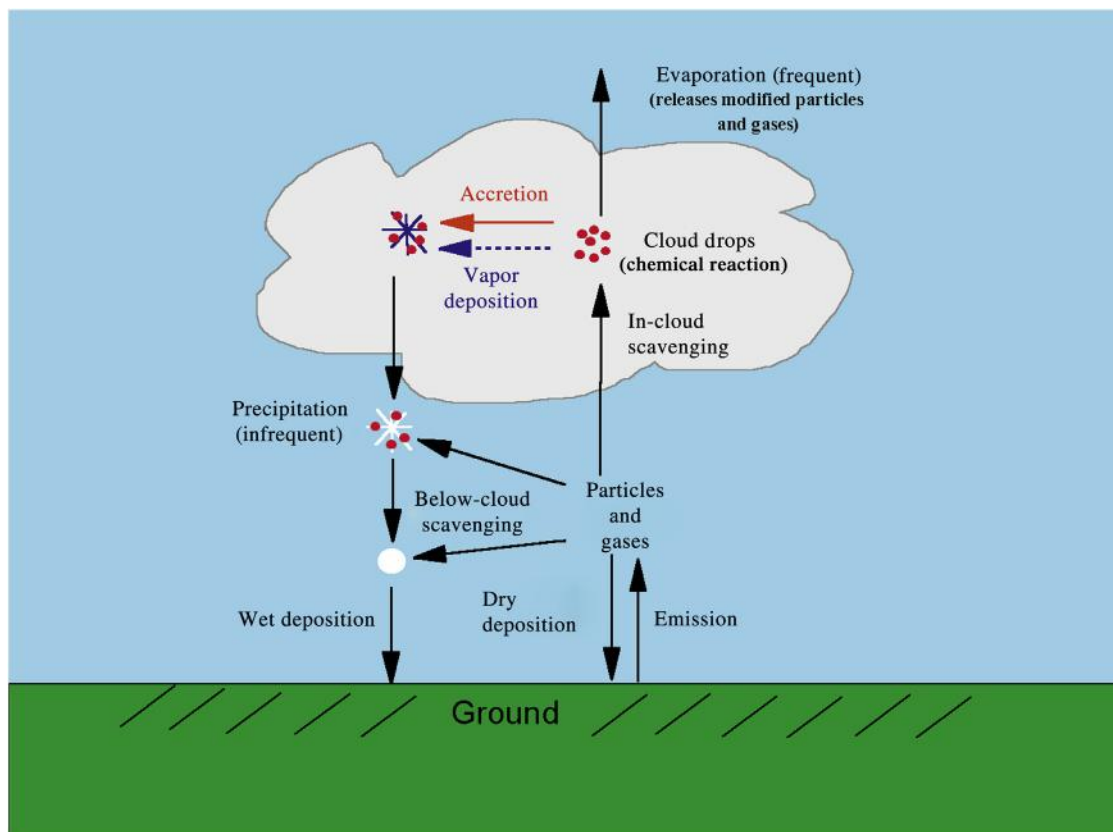


Figure 1 Schematic representation of the multiphase cloud-particle-trace gas system in the atmosphere. Included are processes of dry and wet deposition, particle and gas scavenging by cloud drops, chemical reaction, and precipitation formation in a mixed phase (ice-liquid water) cloud.

initial chemical composition of the cloud. In the sections below, we discuss several aspects of tropospheric cloud chemistry. While not discussed here, it is important to keep in mind that clouds also occur in the stratosphere where they play an important role in promoting heterogeneous reactions important to stratospheric ozone depletion and other processes.

Cloud Drop Formation

Cloud drops form via condensation of water vapor onto a subset of particles termed as cloud condensation nuclei (CCN), a process known as *heterogeneous nucleation*. In the absence of suitable particles, clouds would not form readily as enormous supersaturations are required for the homogeneous nucleation of water vapor. Many atmospheric aerosol particles are hygroscopic. They can take up or lose water in response to changes in the ambient relative humidity (see **Aerosols: Aerosol–Cloud Interactions and Their Radiative Forcing**). At high humidity these particles can take up water vapor. The point at which a dry salt particle spontaneously takes up water vapor to form a saturated salt solution is known as the relative humidity of deliquescence (RHD). The RHD depends on the salt composition. Not all types of particles exhibit deliquescence. Acidic particles, for example, generally have some water associated with them even at extremely low relative humidities.

Particles containing liquid water are often referred to as haze particles. As the humidity increases, above the RHD for deliquescent salt particles, haze particles take on additional water to maintain equilibrium with the partial pressure of water vapor in the atmosphere. The equilibrium partial pressure of water vapor above a haze drop depends on its size and composition. Increased drop curvature raises the equilibrium water vapor pressure via the Kelvin effect. Increased drop solute content lowers the equilibrium vapor pressure by displacement of water molecules near the drop surface. The combined effects of drop curvature and solute content are often described using Köhler theory (see **Aerosols: Aerosol–Cloud Interactions and Their Radiative Forcing**). Application of Köhler theory indicates that for a given particle size, there is a critical supersaturation above which drops will spontaneously take up water vapor and grow unstably. Such drops have been *activated*. The process of activation is also referred to as nucleation scavenging. Growth of activated drops is limited primarily by the availability of water vapor. Haze particles that do not reach their critical supersaturation will not activate and will remain in equilibrium with the ambient water vapor concentration.

Many types of atmospheric particles are capable of serving as CCN. In some clouds, particles as small as 0.1 μm diameter can activate and grow into cloud drops. In other cases, particularly when supersaturations are low and/or particle concentrations are high, the minimum particle size activated may be 0.5 μm or larger. CCN have traditionally been thought to be comprised mainly of sulfate particles, sea salt particles and, in some environments, nitrate particles. Many studies, however, demonstrate that some carbonaceous particles are also efficient CCN. Substantial research efforts are ongoing to elucidate interactions of carbonaceous particles with clouds and fogs.

The soluble fraction of the CCN determines the initial chemical composition of the cloud drop. In addition to

nucleation scavenging, unactivated (interstitial) aerosol particles can be incorporated into cloud drops by a variety of mechanisms including interception, inertial impaction, and diffusion. While nucleation scavenging can often be quite efficient, scavenging of interstitial particles tends to be relatively inefficient.

The concentration of a compound in the aqueous phase resulting from particle scavenging can be expressed as

$$C_w = \frac{\varepsilon_a c_a}{\text{LWC}} = C_a \frac{(\varepsilon_n + \varepsilon_B + \varepsilon_C)}{\text{LWC}} \quad [1]$$

where C_w is the aqueous phase concentration inside the droplets resulting from particle scavenging, C_a is the initial mass concentration of the particles, and ε_a is the mass fraction incorporated in the drops: the scavenging efficiency. LWC is the cloud's liquid water content. ε_n , ε_B , and ε_C represent the mass fraction of matter incorporated by nucleation (ε_n), Brownian motion (ε_B), and collision (ε_C). Overall mass scavenging efficiencies for soluble aerosol species, such as sulfate or chloride, may be only 20–30% for polluted radiation fogs but can approach 100% for clouds formed by vigorous updrafts in pristine environments. Similar results were recently reported for organic carbon with up to 90% scavenging efficiency for fogs in more remote locations while elemental carbon scavenging efficiencies are typically lower. The scavenging efficiency of organic aerosol particles has been demonstrated to vary depending on the source of particle production and associated organic matter composition.

Dissolution of Soluble Gases

In addition to particle scavenging, the composition of cloud drops can be significantly affected by dissolution of soluble gases. These processes are represented in **Figure 2**, where several key soluble gases are shown. Even in remote areas, cloud drops can be partially acidified by dissolution of carbon dioxide. In polluted areas, further acidification can occur by the uptake of nitric acid and sulfur dioxide. In certain environments hydrochloric acid can also be important. Low molecular weight carboxylic acids, especially formic and acetic acids, can be important contributors to drop acidity in both polluted and pristine environments. Sulfur dioxide is moderately soluble in most cloud drops. Its uptake is of particular interest because of the potential for rapid oxidation to sulfate in the aqueous phase. Gaseous oxidants such as hydrogen peroxide and ozone are important because they serve as effective oxidants of dissolved sulfur dioxide. A number of other organic gases can also dissolve into drops. Dissolution of formaldehyde is depicted in **Figure 2**. Carbonyl species, in particular formaldehyde and atmospherically abundant dicarbonyls such as glyoxal and methylglyoxal, are of interest because they may be present at high concentrations and can react with dissolved sulfur dioxide. The dissolution of volatile organics and their subsequent aqueous phase reactivity can yield nonvolatile material and has recently been suggested as an important mechanism for secondary organic aerosol (SOA) generation.

The extent to which a gas partitions into a cloud drop at equilibrium depends on its solubility. In some cases the time required to achieve equilibrium is long relative to the rate of

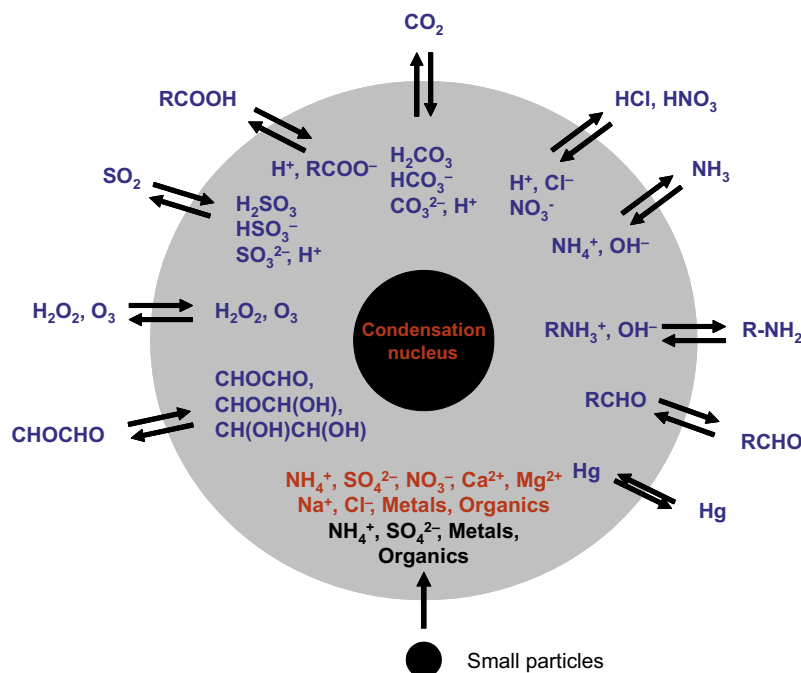


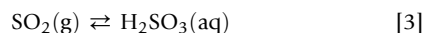
Figure 2 Schematic representation of the influence of particles and gases on cloud/fog drop composition. Reactions between several of the dissolved species shown here can also be important.

reaction of the dissolved gas or to the drop's lifetime, so that phase equilibrium may not be achieved. When the liquid and gas phases are in equilibrium, the concentration in the drop is given by Henry's law:

$$[X_{\text{aq}}] = H_X p_X, \quad [2]$$

where $[X_{\text{aq}}]$ is the concentration of the species X in solution ($\text{mol} \cdot \text{l}^{-1}$), p_X is the partial pressure of X in the atmosphere (atm), and H_X the Henry's law constant for X ($\text{mol} \cdot \text{l}^{-1} \cdot \text{atm}^{-1}$). The solubility of most gases increases with decreasing temperature.

As written above, Henry's law expresses the physical solubility of a gas. For many species, the overall solubility is further enhanced by dissociation or reaction in solution. An important example is the dissolution of sulfur dioxide.



If the compounds are reactive in solution, it is useful to define an effective Henry's law constant which takes into account the chemical reactions and includes the total amount of incorporated compound (e.g., $\text{H}_2\text{SO}_3 + \text{HSO}_3^- + \text{SO}_3^{2-}$)

$$H_{\text{SO}_2}^* = \text{H}_{\text{SO}_2} \left(1 + \frac{K_1}{[\text{H}^+]} + \frac{K_1 K_2}{[\text{H}^+]^2} \right) \quad [6]$$

In eqn [6] K_1 and K_2 are the acid dissociation constants corresponding to reactions [4] and [5], respectively. In case of acid-base reactions in solution, the effective Henry's law constant varies with pH. In case of sulfur dioxide dissolution,

the formation of the ionized forms provides extra reservoirs for sulfur dioxide in solution, thereby increasing the effective solubility of sulfur dioxide substantially when the drop pH exceeds one or both pKa's for dissolved sulfur dioxide. In addition to ionization, other reversible processes undergone by dissolving compounds are often considered in determining effective Henry's law constants. A good example is the aqueous hydration of formaldehyde to form a gem diol. The solubility of atmospheric trace gases in aqueous solution varies strongly, from slightly soluble species (e.g., O_3) to moderately soluble species (e.g., SO_2) to highly soluble species (e.g., H_2O_2 and HNO_3).

It is noteworthy that some field and laboratory observations of the partitioning of gases show a substantial sub- or supersaturation of volatile species in the aqueous phase relative to the gas phase. Hence, species distribution between phases does not always achieve equilibrium, at least on atmospherically relevant timescales. Such situations can occur for specific species as a result of mass transfer limitations between phases, high reactivity in the aqueous phase combined with limited rates of uptake from the gas phase, or, potentially, the presence of films on the surface of droplets.

Drop Composition

The combined effects of particle and gas scavenging, along with chemical reactions (discussed below), determine the chemical composition of cloud drops. Absolute concentration levels of individual species are also influenced in part by condensational growth or evaporation of drops. Major species commonly found to dominate fog and cloud composition include a number of ions, especially nitrate,

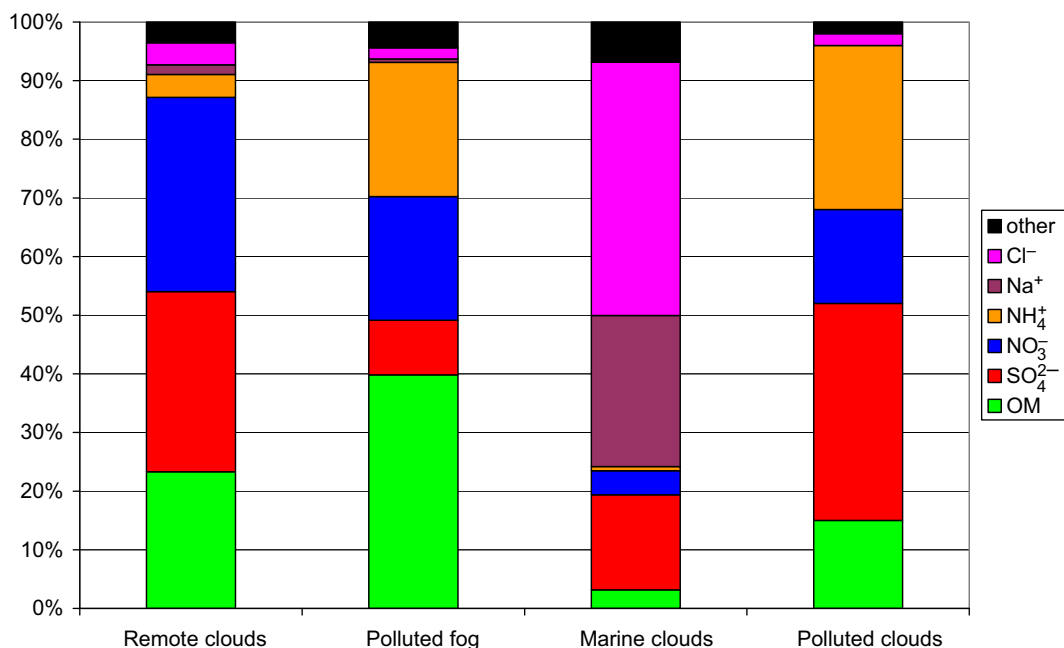


Figure 3 Average mass composition of bulk fog and cloud water samples. OM represents organic matter, assumed equal to 1.8 times the TOC concentration to represent mass contributions from hydrogen, oxygen, nitrogen, and other atoms in organic molecules. Other ions include H^+ , Mg^{2+} , Ca^{2+} , and K^+ . Polluted fog example is from Fresno (California), remote clouds are from Whistler (Canada), marine clouds are from the North Pacific, and polluted cloud example is from Mt. Tai, China.

sulfate, and ammonium. In coastal regions sea-salt ions are important contributors to the drop composition, while in some locations contributions of soil dust components can be important. Organic matter is also an important component of fog and cloud composition in many environments (Figure 3). In some polluted environments, concentrations of formaldehyde, glyoxal, methylglyoxal, acetate, and formate can reach levels similar to those observed for major inorganic ion species. Table 1 provides approximate concentration ranges of various chemical species observed in clean and polluted clouds and in polluted radiation fogs.

Table 1 Typical range of LWC and solute concentrations in remote and polluted clouds and fogs

Solute	Polluted cloud	Polluted fog	Remote cloud
LWC (g m^{-3})	0.05–1	0.02–0.8	0.05–2
pH	2–5	3–7	4–6
SO_4^{2-} ($\mu\text{eq l}^{-1}$)	50–2000	50–5000	5–50
NO_3^- ($\mu\text{eq l}^{-1}$)	10–2000	50–20 000	0–20
NH_4^+ ($\mu\text{eq l}^{-1}$)	50–1000	100–20 000	10–50
Cl^- ($\mu\text{eq l}^{-1}$)	0–500	0–100	0–500
Na^+ ($\mu\text{eq l}^{-1}$)	0–500	0–200	0–500
HCOO^- ($\mu\text{eq l}^{-1}$)	0–100	0–1000	0–20
CH_3COO^- ($\mu\text{eq l}^{-1}$)	0–100	0–500	0–10
HCHO (μM)	10–50	5–500	0–10
CHOCHO (μM)	5–50	5–200	0–10
TOC (ppmC)	2–30	5–50	0–8
Fe (μM)	0.01–10	1–50	0–5

Experimental observations reveal that only a fraction of the total organic carbon (TOC) content of fogs or cloud is comprised of low molecular weight compounds such as formate, acetate, formaldehyde, and other small carbonyls and dicarbonyls. Some efforts have been made to comprehensively characterize the higher molecular weight compounds in fogs and clouds. Initial studies revealed similarities to humic substances, although evidence shows that the higher molecular weight species in fogs and clouds only present limited similarities to terrestrial humic substances. Other fractions appear similar to fulvic acids or unique to atmospheric materials. The current consensus might be summarized as follows: organic matter is complex, comprised of thousands of species, many of them in the range of hundred to several hundred Dalton molecular weight with a high content of heteroatoms (O,N,S). Many studies have focused on measuring particular compound families because of their specific sources or their toxicity. Hundreds of organic compounds have been identified in fogs and clouds in recent years using advanced mass spectrometric analysis. Organic compounds measured in fogs and clouds include carboxylic acids (formic, acetic, pyruvic, propionic), aldehydes (formaldehyde, acetaldehyde, glyoxal, methylglyoxal, benzaldehyde), polycyclic aromatic hydrocarbons (chrysene, pyrene, fluoranthene, benzo(e)-pyrene, benzo(b)-fluoranthene), ketones (acetone), carbohydrates (levoglucosan, mannosan, galactosan), pesticides (carabryl, diazinon, malathion, lindane, atrazine), and various phenols (phenol, 2- and 4-nitrophenol, 2,4-dinitrophenol, methyl nitrocatechol, *m/p*-cresol, guaiacol, syringol). This list of compounds is illustrative but not exhaustive. Many organic compounds found in cloud drops are emitted during combustion while others are formed

as secondary products of atmospheric reactions. With the exception of the low molecular weight carboxylic acids and aldehydes, as well as levoglucosan in a strongly biomass burning impacted environment, cloud drop concentrations of most individual organic compounds are typically much smaller than those of the major inorganic species.

It is noteworthy that a part of the organic matter in clouds is of biological origin including amino acids, proteins, and other cellular materials, including RNA and DNA. There is substantial evidence for the presence of whole biological organisms (bacteria, virus, fungi) as well as spores in fogs and clouds (see **Tropospheric Chemistry and Composition: Aerosols/Particles**). The variability of organisms, their viability, their contribution to cloud organic matter, and the potential impact of biological reactions on cloud chemistry are currently under investigation.

Although clouds have traditionally been assumed to be comprised of populations of chemically homogeneous drops, there is now substantial evidence that a wide variety of drop compositions often exists within a given cloud or fog. These differences arise in part from the variability in the composition of the underlying CCN. Other differences arise from variations in condensational growth rates as a function of drop size and from nonequilibrium absorption of highly soluble trace gases. Increasingly, cloud chemistry models and measurements are being designed to account for the chemical heterogeneity among cloud drop populations. This heterogeneity can affect rates of in-cloud reactions as well as the efficiency of solute deposition through precipitation scavenging and direct cloud drop deposition.

Reactions Occurring in Cloud Drops

Many chemical species incorporated in cloud drops can react in the aqueous phase. Considerable attention has been focused on the aqueous oxidation of dissolved sulfur dioxide to sulfate, although numerous other reactions have also been shown to be important, including the formation of SOA. Several key chemical reactions are briefly discussed in the following.

Aqueous Phase Sulfur Oxidation and Complexation

Oxidation of SO_2 to H_2SO_4 is a key reaction in the atmosphere. The resulting sulfate contributes to acid deposition, visibility degradation, respiratory problems, and climate modification. Sulfate production occurs both in the gas phase and in the aqueous phase, but the latter is much faster.

As described above, dissolution of SO_2 in cloud drops is a function of drop pH (Figure 4). In the pH range usually encountered in cloud and fog drops (pH between 2 and 7), the dominant form of dissolved sulfur dioxide is bisulfite. As the pH climbs above 7, the speciation shifts toward sulfite.

Several S(IV) oxidation mechanisms occur in the aqueous phase; their relative importance depends on the drop pH and the availability of oxidants and catalysts. Three important aqueous S(IV) oxidation pathways are oxidation by hydrogen peroxide, ozone, and oxygen (autooxidation) catalyzed by Fe(III) and Mn(II). Figure 5 depicts the rate of sulfate production by these three pathways for 1-ppbv sulfur dioxide in the

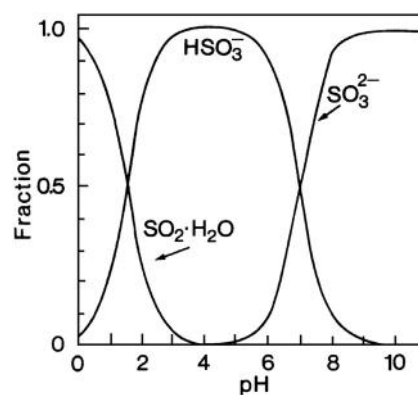


Figure 4 The speciation of dissolved sulfur dioxide as a function of pH.

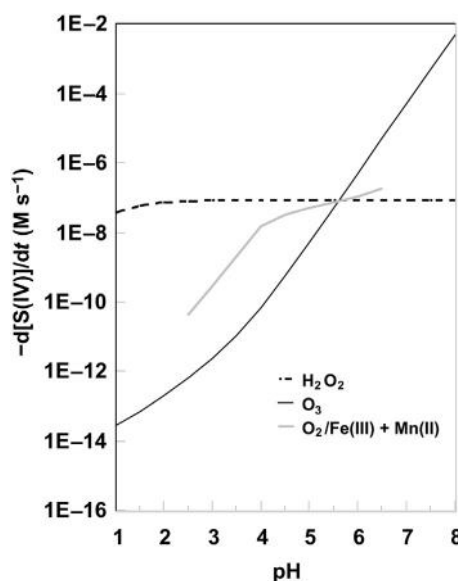


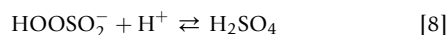
Figure 5 The pH dependence of the rate of aqueous S(IV) oxidation by hydrogen peroxide, ozone, and oxygen (catalyzed by Fe(III) and Mn(II)). Conditions represented in the figure are $T = 298 \text{ K}$, $p_{\text{SO}_2} = 1 \text{ ppbv}$, $p_{\text{O}_3} = 30 \text{ ppbv}$, $p_{\text{H}_2\text{O}_2} = 1 \text{ ppbv}$, $\text{Fe(III)} = 2.5 \times 10^{-6} \text{ M}$, and $\text{Mn(II)} = 1.0 \times 10^{-6} \text{ M}$.

presence of 30-ppbv ozone and 1-ppbv hydrogen peroxide. The rate of production is shown as a function of drop pH.

In many clouds hydrogen peroxide is assumed to be the dominant S(IV) oxidant. As shown in Figure 5, this oxidation pathway tends to be much faster than the others at lower pH values. While the ozone and autooxidation paths become slower with decreasing pH, the rate of the hydrogen peroxide pathway is essentially independent of pH over the range of interest. The lack of pH dependence results from the opposing pH dependence of two contributing factors. While the effective solubility of sulfur dioxide decreases with decreasing pH, the intrinsic rate of sulfate production increases at lower pH. It is believed that oxidation occurs by nucleophilic displacement of a water molecule by hydrogen peroxide attack on bisulfite



followed by reaction of the peroxymonosulfurous acid intermediate with hydrogen ion to yield sulfuric acid



Because the second step is rate limiting, the intrinsic rate of oxidation increases with increasing hydrogen ion concentration. Oxidation of S(IV) is also possible by organic peroxides, but those reactions are of lesser importance due to their lower atmospheric concentrations and their lower aqueous solubility.

At higher pH, or after available hydrogen peroxide has been consumed, S(IV) oxidation by ozone can be important. The rate of reaction of S(IV) with ozone in aqueous solution can be expressed as

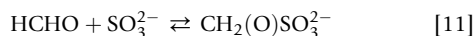
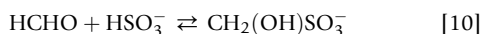
$$-\text{d}[\text{S(IV)}]/\text{d}t = (k_0\alpha_0 + k_1\alpha_1 + k_2\alpha_2)[\text{S(IV)}][\text{O}_3] \quad [9]$$

where k_0 , k_1 , and k_2 are the reaction rate constants for reaction of ozone with H_2SO_3 , HSO_3^- , and SO_3^{2-} , respectively; α_0 , α_1 , and α_2 represent the fractions of dissolved sulfur dioxide present as H_2SO_3 , HSO_3^- , and SO_3^{2-} , respectively. Because the effective solubility of sulfur dioxide increases with pH and because sulfite is oxidized much more rapidly than bisulfite, which is oxidized more rapidly than sulfurous acid, the rate of sulfate production by this pathway increases strongly with increasing pH. As shown in Figure 5, the rate of this pathway can exceed even that of the hydrogen peroxide pathway at pH values greater than 5 for typical conditions. The rate of this pathway becomes so rapid, in fact, that it tends to become mass transport limited for larger drops at high pH. This is especially likely when formaldehyde is present as a copollutant, since dissolved sulfur dioxide is rapidly consumed by reaction with both ozone and formaldehyde at high pH.

Aqueous S(IV) oxidation by oxygen is also possible but is very slow in the absence of catalysts. The reaction can be fast enough to be important when catalyzed by certain trace metals including Fe(III) and Mn(II). Of particular interest is the synergistic catalysis afforded by the simultaneous presence of Fe(III) and Mn(II). While several investigators have reported oxidation rate expressions for this pathway, there is far from uniform agreement among the reported expressions. Evaluation of the importance of this pathway in actual clouds and fogs is further complicated by the difficulty of making accurate measurements of iron and manganese speciation in the field.

In some circumstances, S(IV) oxidation by radicals, including OH, Cl_2^- , and Br_2^- , may also be important, especially in polluted conditions where the pH is low and sulfur dioxide concentration far exceeds the hydrogen peroxide concentration.

Field studies have found S(IV) to be present in cloud droplets at much higher concentrations than predicted by Henry's law, even accounting for ionization of dissolved sulfur dioxide in solution. An important reason accounting for this apparent discrepancy is the tendency for sulfite and bisulfite to form complexes with various aldehydes, especially formaldehyde:

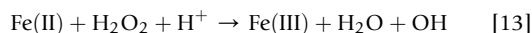
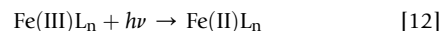


The product of the reaction with formaldehyde is hydroxymethanesulfonate (HMS). Formation of HMS is favored at high pH. Significant concentrations of HMS have been measured in high pH fogs in polluted environments, including California's San Joaquin Valley. Formation of HMS is of interest because it represents an additional sink for sulfur dioxide in high pH drops and because it is fast enough under those conditions to limit the amount of aqueous phase sulfate production. Although S(IV) complexation by other aldehydes can also occur, these reactions are generally of lesser importance than those with formaldehyde due to the lower solubilities of higher molecular weight aldehydes and their lower concentrations in the atmosphere. HMS is stable with respect to oxidation by O_3 and H_2O_2 , but may be oxidized by OH.

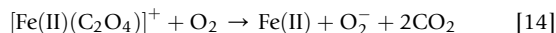
Oxidant Chemistry

A number of other reactions also occur in cloud drops involving radicals and other oxidants. Several oxidants, including H_2O_2 , O_3 , HO_2 radicals, and OH radicals, contribute importantly to aqueous phase atmospheric chemistry. Organic peroxides and organic peroxy radicals do not play an important role because of their lower atmospheric concentrations and/or lower aqueous solubilities.

H_2O_2 , O_3 , HO_2 , and OH can all be transferred to cloud or fog drops from the gas phase. H_2O_2 , HO_2 , and OH can also be generated photolytically in solution. Several photolytic sources exist for OH production, including photodissociation of H_2O_2 , iron-hydroxocomplexes, nitrate, and nitrite. OH can also be produced by the photo-Fenton reaction



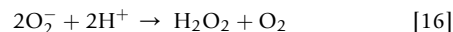
where L denotes an organic ligand such as oxalate. Experimental studies confirm that HO_2 is formed photochemically upon the illumination of cloud and fog samples, subsequently yielding H_2O_2 . Superoxide (O_2^-) is believed to be formed from reaction of Fe(II)-oxalate complexes



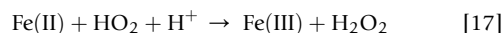
followed by protonation to yield HO_2 .



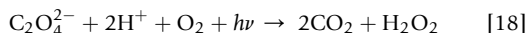
Superoxide can form peroxide in solution via



Fe(II) has also been postulated to react with HO_2 (or O_2^-) to form H_2O_2



Overall, the redox pair of Fe(II)/Fe(III) catalyzes the degradation of oxalate into CO_2 and H_2O_2 .



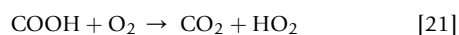
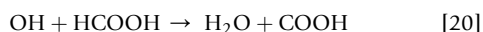
Laboratory studies of the effect of sunlight illumination on cloud samples suggest that these reactions can be important in both producing H_2O_2 and influencing the daytime cloud pH.

OH and HO₂ can also be formed in solution as a result of bimolecular reactions and radical interconversion processes. One reaction of particular interest is the oxidation of hydrated formaldehyde (methylene glycol) by OH to produce HO₂ and formic acid. The overall stoichiometry for this multistep reaction is



Formic acid is commonly observed in clouds and fogs formed in both pristine and polluted environments. Similar reactions are possible for oxidation of hydrated formaldehyde by other radicals, including NO₃, SO₄⁻, Cl₂⁻, and CO₃⁻. Unlike the radicals discussed above, which can be formed *in situ* or transferred from the gas phase, SO₄⁻, Cl₂⁻, and CO₃⁻ are only formed in the aqueous phase.

Formic acid and other higher molecular weight organic acids are also subject to oxidation by aqueous radicals. Reaction between formic acid and OH,



represents an efficient source of HO₂ and a strong sink for OH. OH is also capable of oxidizing formate anion, so that HO₂ formation by formic acid oxidation is not expected to vary strongly with pH.

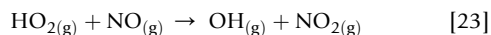
In addition to reacting with aldehydes and organic acids, aqueous phase radicals are capable of reacting with other organic substrates containing abstractable hydrogen atoms. Because we know relatively little about the organic speciation of fog and cloud drops, it is difficult to predict accurately how such reactions might impact the lifetimes of aqueous phase radicals.

Several aqueous radicals are believed to undergo reaction with one or more transition metals. In addition to reaction with iron, reactions are also possible with manganese, cobalt, and copper. Copper is of particular interest because it reacts quick enough with HO₂/O₂⁻ to compensate for its generally low concentrations in atmospheric waters, while manganese and cobalt are not expected to significantly influence aqueous phase radical concentrations.

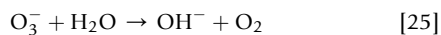
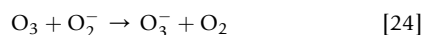
Reactions with nitrite may also represent important sinks for aqueous radicals including OH. The mechanism for the reaction of radical species X with nitrite is given by



Cloud drop scavenging of HO₂ from the gas phase and *in situ* HO₂ production via the mechanisms discussed above have received attention because of their potential effects on tropospheric ozone concentrations. One expects that depletion of HO₂ from the gas phase should influence ozone concentrations in the gas phase because of the suppression of the reaction:



Further, dissociation of HO₂ in solution to form superoxide can promote ozone destruction in the aqueous phase via



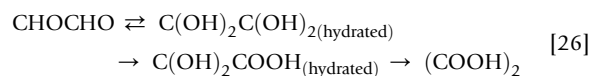
The overall effect of clouds on tropospheric O₃ concentrations remains a topic of discussion with some authors arguing for a potentially significant effect while others suggest it is probably not a major factor. Differing conclusions stem at least in part from differences in the reaction mechanisms considered.

Formation of Secondary Organic Aerosol Material

In recent years laboratory and fieldwork have provided substantial evidence that fogs and clouds can convert volatile organic gases into nonvolatile compounds via aqueous phase reactions and, hence, contribute to the formation of SOA. The number of precursor compounds and chemical processes considered important for SOA formation is ever increasing. We will discuss here the most studied molecular systems involving glyoxal and methylglyoxal, two species formed in the gas phase from biogenic and anthropogenic precursors (*see Aerosols: Aerosol Physics and Chemistry*). The SOA contributions from aqueous phase reaction of these species are likely to be especially important as they are some of the most abundant individual organic compounds in fogs and clouds as detailed earlier.

SOA formation processes can be separated into photochemical processes and *dark* (nonphotochemical) reactions. In the dark reactions, oligomerization of dicarbonyls with themselves or other species like amines is observed. Oligomers are small polymers containing only a few monomer units. The resulting daughter species have higher molecular weight than the precursor molecules and are considerably less volatile. These processes are sensitive to the precursor concentrations and hence are particularly prominent in highly concentrated solutions like haze droplets and are considered of lesser importance in more dilute (grown) cloud and fog water droplets.

In photochemical processes, the volatile precursors are oxidized to less volatile species. In the case of glyoxal, for example, glyoxal might be oxidized into glyoxilic acid and further into oxalic acid.



Photochemical experiments reveal a multitude of products, indicative of various simultaneous pathways. In addition photochemical experiments also show the formation of oligomers similar to those produced in dark reactions.

Deposition

The chemical and microphysical properties of clouds exert a significant influence on the composition of precipitation and the resulting wet deposition. Precipitation (rain drops and snow crystals) can form in a variety of ways. Outside the tropics, much precipitation is produced in mixed phase (ice-liquid water) clouds. In this environment precipitation is formed in the ice phase in the presence of supercooled cloud drops. Ice crystal growth can occur by water vapor deposition, inertial capture of cloud drops (*accretion* or *riming*), and ice crystal aggregation. Because the saturation vapor pressure of

water is higher over liquid water than over ice at the same temperature, ice crystals often take up water vapor at the expense of the cloud drops. As the cloud drops evaporate, their solute concentrations increase. Meanwhile, the condensation of water vapor on the ice crystal surface tends to dilute the concentrations of those species present. Consequently, it is common for precipitation to possess much lower solute concentrations than found in collocated cloud drops. For larger ice crystals and larger cloud drops, accretional growth of precipitation becomes more important. When crystals are heavily rimed, their composition has been shown to closely resemble that of the accreted cloud drops. Because accretional growth favors inertial capture of larger cloud drops, any drop size-dependence of cloud drop composition can influence the precipitation scavenging efficiencies of individual solute species. This is also true in warm rain formation, where collision and coalescence of large cloud drops leads to formation of rain drops.

In some environments, cloud and fog drops can be directly deposited on Earth's surface. Drops are removed by inertial impaction, interception, and sedimentation. The latter process dominates removal in radiation fogs while the former processes can be quite important for capture of cloud drops by montane forest canopies. Various studies in the United States and Europe have shown that for high elevation sites with a high cloud interception frequency, hydrologic input by direct cloud deposition is measurable but often low compared to the input by rain and snow. However, because cloud solute contents are often far higher than precipitation solute contents, cloud drop deposition can significantly enhance total wet deposition fluxes of many chemical species from the atmosphere to terrestrial ecosystems. Likewise, studies of radiation fogs reveal that they can represent an important vector for deposition of accumulation mode aerosol particles in polluted environments.

In certain parts of the world, including coastal northern California and parts of the Hawaiian Islands, the hydrologic input from cloud drop deposition appears to be an important contributor to the survival of some plant species. Likewise, cloud drop deposition has been harnessed in some arid regions as a means of supplementing the local water supply for small towns and villages. Large mesh screens are erected on ridge tops, where nonprecipitating clouds are frequently intercepted, to capture water for domestic use.

See also: **Aerosols:** Aerosol Physics and Chemistry; Aerosol–Cloud Interactions and Their Radiative Forcing. **Biogeochemical Cycles:** Sulfur Cycle. **Clouds and Fog:** Fog; Stratus and Stratocumulus. **Tropospheric Chemistry and Composition:** Aerosols/Particles; Sulfur Chemistry, Organic.

Further Reading

- Amato, P., Menager, M., Sancelme, M., Laj, P., Mailhot, G., Delort, A.M., 2005. Microbial population in cloud water at the Puy de Dome: implications for the chemistry of clouds. *Atmospheric Environment*. 39, 4143–4153.
- Bator, A., Collett Jr., J.L., 1997. Cloud chemistry varies with drop size. *Journal of Geophysical Research*. 102, 28071–28078.
- Collett Jr., J.L., Prevot, A.S.H., Staehelin, J., Waldvogel, A., 1991. Physical factors influencing winter precipitation chemistry. *Environmental Science and Technology*. 25, 782–789.
- Collett Jr., J.L., Bator, A., Sherman, D.E., Moore, K.F., Hoag, K.J., Demoz, B.B., Rao, X., Reilly, J.E., 2002. The chemical composition of fogs and intercepted clouds in the United States. *Atmospheric Research*. 64, 29–40.
- Ervens, B., Turpin, B.J., Weber, R.J., 2011. Secondary organic aerosol formation in cloud droplets and aqueous particles (aqSOA): a review of laboratory, field and model studies. *Atmospheric Chemistry and Physics*. 11, 11069–11102.
- Faust, B.C., 1994. Photochemistry of clouds, fogs and aerosols. *Environmental Science and Technology*. 28, 217A–222A.
- Finlayson-Pitts, B.J., Pitts Jr., J.N., 2000. *Chemistry of the Upper and Lower Atmosphere*. Academic Press, San Diego, 969 pp.
- Herckes, P., Valsaraj, K.T., Collett, J.L., 2013. A review of observations on organic matter in fogs and clouds: origin, processing and fate. *Atmospheric Research*. 132–133, 434–449.
- Herrmann, H., 2003. Kinetics of aqueous phase reactions relevant for atmospheric chemistry. *Chemical Reviews*. 103, 4691–4716.
- Herrmann, H., Tilgner, A., Barzaghi, P., Majdik, Z., Gligorovski, S., Poulain, L., Monod, A., 2005. Towards a more detailed description of tropospheric aqueous phase organic chemistry: CAPRAM 3.0. *Atmospheric Environment*. 39, 4351–4363.
- Munger, J.W., Collett Jr., J., Daube Jr., B.C., Hoffmann, M.R., 1989. Carboxylic acids and carbonyl compounds in southern California clouds and fogs. *Tellus* 41B, 230–242.
- Pruppacher, H.R., Klett, J.D., 1997. *Microphysics of Clouds and Precipitation*. Kluwer, Dordrecht, 954 pp.
- Ravishankara, A.R., 1997. Heterogeneous and multiphase chemistry in the troposphere. *Science* 276, 1058–1065.
- Seinfeld, J.H., Pandis, S.N., 2006. *Atmospheric Chemistry and Physics – From Air Pollution to Climate Change*, second ed. John Wiley, New York. 1232 pp.
- Weathers, K.C., Likens, G.E., Bormann, F.H., Bicknell, S.H., Bormann, B.T., Daube Jr., B.C., Eaton, J.S., Galloway, J.N., Keene, W.C., Kimball, K.D., McDowell, W.H., Siccama, T.G., Smiley, D., Tarrant, R.A., 1988. Cloudwater chemistry from ten sites in North America. *Environmental Science and Technology*. 22, 1018–1026.

Introduction

Molecular hydrogen (H₂) belongs to the group of the five most abundant trace gases in the troposphere. The mean tropospheric mixing ratio is roughly about 500 parts per billion by volume (ppbv), a volume fraction of 500×10^{-9} . In the atmosphere H₂ is also present in two isotopic forms: the stable compound HD (¹H²H) and the radioactive form HT (¹H³H). In contrast to carbon dioxide (CO₂), methane (CH₄), nitrous oxide (N₂O), and carbon monoxide (CO), H₂ is not a greenhouse gas. Scientific interest in its atmospheric distribution and budget has been mainly due to the following aspects.

1. During the nuclear weapons tests in the 1950s, radioactive tritium (³H) was released in the form of HT into the atmosphere in significant amounts. The need for information about possible removal mechanisms other than radioactive decay motivated the first detailed studies of the global budget of molecular hydrogen.
2. In addition to carbon monoxide (CO), H₂ is a major product of the photochemical oxidation of CH₄ and other hydrocarbons in the atmosphere. Therefore, the tropospheric cycles of these three trace gases are strongly coupled and information on the H₂ budget will also improve the knowledge about the tropospheric budget of the other species.
3. In addition to methane (CH₄) and water vapor (H₂O), H₂ contributes to the budget of total water, defined as

$\sum \text{H}_2\text{O} = 2\text{CH}_4 + \text{H}_2\text{O} + \text{H}_2$, in the stratosphere. Because the greenhouse effect of H₂O contributes to the negative forcing (cooling) in the lower stratosphere, and owing to the direct effect of water vapor in heterogeneous reactions that activate reactive chlorine, the temporal evolution of $\sum \text{H}_2\text{O}$ is of some importance for the future development of the stratospheric ozone layer. Therefore, the tropospheric H₂ budget must be understood in order to assess a possible long-term global trend of the H₂ mixing ratio and its contribution to the temporal trend of $\sum \text{H}_2\text{O}$ in the lower stratosphere.

This article summarizes the known information about the spatial and temporal distribution of molecular hydrogen in the troposphere. The most important production and removal terms that contribute to the global tropospheric H₂ budget are reviewed from the current literature.

Global Tropospheric Abundance

The first reliable observation of molecular hydrogen in tropospheric air was made in 1923 by P.M. Schuftan at an air liquefaction plant in Southern Germany. He found a mixing ratio of 500 ± 100 ppbv, which was first reported in the scientific literature by F.A. Paneth in 1937. Until 1980 only few measurements of H₂ in the nonurban atmosphere were available. These are summarized in Figure 1 together with the

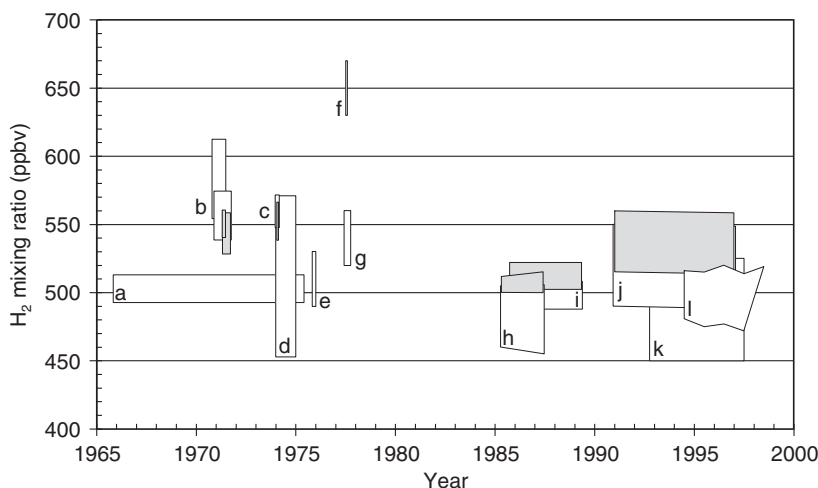


Figure 1 Time series of reliable observations of molecular hydrogen in nonurban tropospheric air. The data for the period from 1950 until 1977 are from (a) Ehhalt *et al.* (1977); (b) Schmidt (1974); (c) Schmidt (1978); (d) Heidt and Pollock (1976); (e) Seiler *et al.* (1978); (f) Herr and Barger (1978); and (g) Fabian *et al.* (1978). The data for the period from 1985 until 1997 were taken from the more systematic studies: (h) and (i) Khalil and Rasmussen (1990); (j) Novelli *et al.* (1999); (k) Francey *et al.* (1998); and (l) Simmonds *et al.* (2000). Boxes indicate the variability (vertical scale) and the time period (horizontal scale) of the various observations. Open and shaded boxes represent data for the Northern and Southern Hemispheres, respectively.

results obtained during some more extended series of observations that were made up to 2000.

Apparently, the available historical data set does not indicate any systematic long-term trend of the global H₂ mixing ratio. Two more recent studies have reported a slight increase over observational periods of a few years. However, the study in 1999 by Novelli and colleagues, which includes observations from a large number of globally distributed stations rather indicates a slight decrease of the H₂ mixing ratio of about 3% over the period from 1991 to 1996.

The early data plotted in Figure 1 suggest a larger variability than those obtained after 1980. In large part this is due to the individual absolute calibration scales employed during the different observational series reviewed in 1979 by Schmidt and colleagues. Later data reported by various workers (see Further Reading) indicate more consistent mean mixing ratios. The observed variability is mainly due to the seasonal variation of the H₂ mixing ratio, which was only noticed during these more systematic studies.

Latitudinal Distribution

A closer inspection of the historical data reveals another important feature of the global distribution. Data reported by Schmidt for 1971–72 indicated an H₂ level in surface air of the Northern Hemisphere that is about 6% higher than that observed in the Southern Hemisphere. The interhemispheric difference was only about 1.5% in the upper troposphere and a similar value (about 1.3%) was derived later by Schmidt from observations made in 1974.

In contrast, more recent observations show surface air mixing ratios that are consistently lower in the Northern Hemisphere than in the Southern Hemisphere. These observations are displayed in Figure 2. The most comprehensive global data set, that of Novelli and colleagues, obtained over the period from 1992 until 1995, indicates mean surface H₂ mixing ratios that are largest (about 540 ppbv) in the tropical region and decrease with increasing latitude to values close to 530 ppbv and 480 ppbv in the Southern and Northern Hemispheres, respectively. No new data for the free upper troposphere have become available in recent times.

Although there is no significant long-term trend in the historical record of the global mean mixing ratio (see Figure 1), the data plotted in Figure 2 suggest that the surface mixing ratios might have varied with time by some 3–4%. But the differences in the absolute H₂ levels of the three data sets, which amount to about 20–40 ppbv, might also be attributed to possible differences in the absolute calibration scales. However, there is clear evidence for a reversal of the relative, interhemispheric, difference of H₂ the mixing ratio over the last about three decades. This reversal is most probably due to a decrease of the mean H₂ mixing ratio in the Northern Hemisphere.

Seasonal Variability

The first systematic long-term measurements at monitoring stations revealed another cause of the observed natural

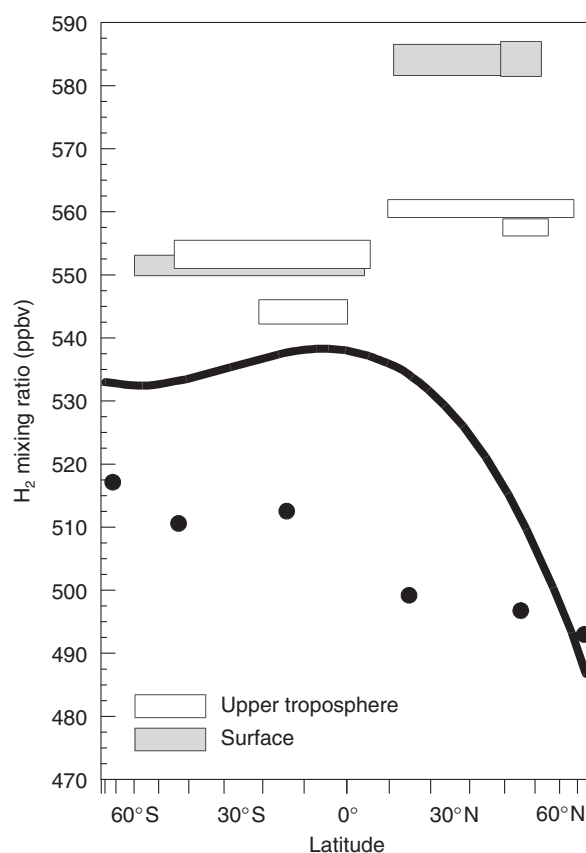


Figure 2 Observations of the latitudinal distribution of H₂ in the troposphere. Boxes represent average H₂ levels (and their 1 standard deviation uncertainty) in surface air and in the upper troposphere as observed by Schmidt (1974, 1978) between 1971 and 1974. Solid dots represent mean mixing ratios obtained by Khalil and Rasmussen (1990) from measurements at six sites over the period from 1985 to 1989. The heavy line shows the annual mean H₂ mixing ratios observed by Novelli *et al.* (1999) over the period 1992–95.

variability of H₂ in surface air: a pronounced seasonal variation. The amplitude of the seasonal cycle is about 20 ppbv in the Southern Hemisphere but is significantly larger, some 70 ppbv, in the Northern Hemisphere, in particular at higher northern latitudes. In fact, this seasonal variation is the cause of much of the observed asymmetry in hemispheric mean H₂ levels. The average latitudinal gradient displayed in Figure 2, derived from the comprehensive data collected at about 50 globally distributed sites, is due to the generally lower seasonal minima in the Northern Hemisphere.

In addition to this difference in amplitude, the seasonal cycles exhibit another interesting feature: a nonsymmetric timing of the seasonal minima and maxima. In the Northern Hemisphere, highest H₂ mixing ratios are observed in March and April (i.e., in early spring), while they are observed in the December to February period (i.e., in early summer) in the Southern Hemisphere; that is, the shift between both hemispheres is only 3 to 4 months. A typical example of the different timing of the seasonal cycles is shown in Figure 3, which compares the annual variation of H₂ at the stations Mace Head (53° N) and Cape Grim (42° S).

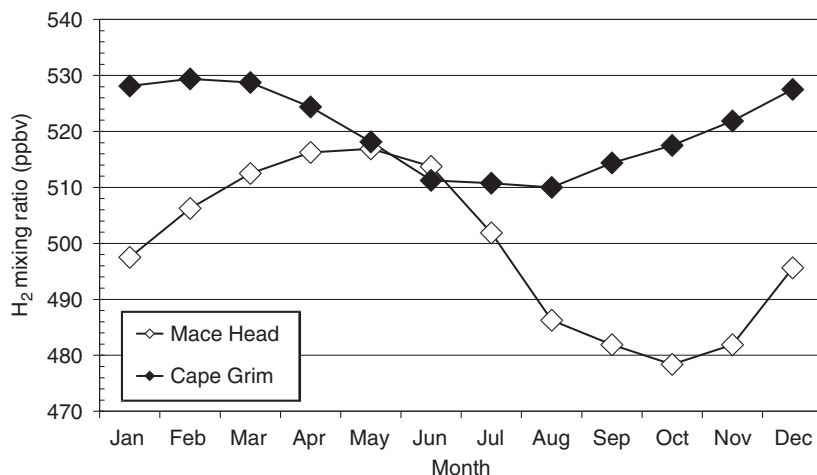


Figure 3 Seasonal variation of the H₂ mixing ratio in surface air at the monitoring sites Mace Head (53° N) and Cape Grim (42° S). Adapted from Simmonds *et al.* (2000).

H₂ Sources

The observed seasonal variation and its different timing in both hemispheres must be the consequences of a rather complex interaction of the production and destruction processes for atmospheric H₂. Moreover, the reversal of the interhemispheric H₂ distribution (see Figure 2) strongly suggests that the strength of some of the major source or sink processes must have changed over the last few decades. The global budget of H₂ is believed to be largely determined by natural processes. However, a number of important anthropogenic sources have also been identified. The sources are briefly discussed in the following sections. Updated estimates of the principal production rates were published by Ehhalt and by Novelli and colleagues.

Combustion Processes

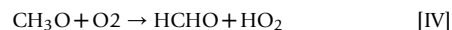
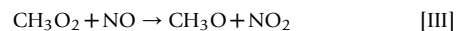
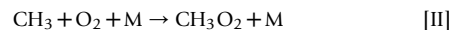
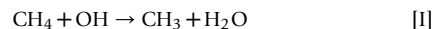
Observations in urban air or continental air masses generally show higher mixing ratios than those in background tropospheric air. This suggests that industrial combustion processes and motor vehicle exhaust in particular are major sources of H₂. Direct quantification of the strength of this source is difficult. Best estimates rely on the correlation of the observed mixing ratio increments above the H₂ background values relative to that of carbon monoxide (CO), a trace gas for which the anthropogenic source is much better known. Novelli and colleagues, as well as Simmonds and colleagues, reported correlation ratios significantly lower than those reviewed by Warneck in 2000. Novelli and colleagues estimate the anthropogenic emission of H₂ from automotive exhaust and other technological processes as $15 \pm 10 \text{ Tg year}^{-1}$. Most of this production (between 80% and 95%) should be released in the Northern Hemisphere. Simmonds and colleagues reported substantially lower ratios of the H₂ and CO increments in European air masses. If these lower values are shown to be representative for the present global H₂ emissions from technological sources, the above estimate of this source must be

strongly reduced. Because anthropogenic emissions from fossil fuel combustion continue to increase steadily with time, absolute H₂ emission rates from fossil fuel use must have dropped substantially over recent decades in order to cause the observed reversal of the interhemispheric difference in the H₂ mixing ratio.

Biomass burning is another anthropogenic process that releases H₂ to the atmosphere. Estimates of this source are based mainly on measurements of the H₂/CO₂ concentration ratio measured in natural fire plumes and during laboratory combustion experiments. Novelli and colleagues reviewed previous estimates and reported a strength of this source as $16 \pm 5 \text{ Tg year}^{-1}$. It is generally believed that at least 80% of all fires are caused by humans. Most fires occur in the tropical region and in the extratropics of the Northern Hemisphere. Therefore, 60% of the total H₂ emissions from biomass burning are expected to occur in the Northern Hemisphere.

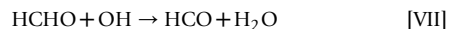
Photochemical Production

The photochemical oxidation of volatile hydrocarbons is by far the largest natural source of H₂. The reaction of methane (CH₄), the most abundant hydrocarbon in tropospheric air, with hydroxyl radicals (OH) initiates a reaction sequence ([I], [II], [III] and [IV]) that produces formaldehyde (HCHO).



Although the four reactions of this sequence do not fully describe the more complex chemical processes of CH₄ oxidation in the real natural troposphere, it may be assumed that each CH₄ molecule oxidized yields one HCHO molecule. The HCHO produced is either photolyzed in the sunlit atmosphere

via the two pathways ([V] and [VI]) or is destroyed by reaction with OH radicals as in reaction (reaction [VII]).



Only the second photolysis reaction (reaction ([VI])) leads to the formation of H₂. Therefore, the H₂ production rate depends on the relative rates of HCHO destruction via the three pathways, and will vary with altitude, daytime, and season. Assuming a mean global tropospheric OH concentration of about 10⁶ cm⁻³, as derived from global observations of methylchloroform (CH₃CCl₃), a global H₂ production through CH₄ oxidation of 26 ± 9 Tg year⁻¹ was estimated. Owing to the small interhemispheric gradient in the global distribution of CH₄, this H₂ source should be of the same magnitude in both hemispheres.

Similar reaction sequences produce HCHO during the oxidation of nonmethane hydrocarbons (NMHC); the oxidation of isoprene, for example, yields about 3 molecules of HCHO per isoprene molecule. However, the oxidation reaction sequences for the different compounds belonging to the group of NMHC species (e.g., isoprenes and terpenes) are much more complex than that for CH₄. Net HCHO yields depend on the amount of NO_x in the air mass and on the efficiency of heterogeneous reactions, in particular in the absence of sunlight. Furthermore, the global distributions, emission rates, and degradation processes for many NMHC compounds are only poorly understood, although these species are ubiquitous in the global troposphere. Therefore, global estimates of the H₂ production from these compounds have considerable uncertainties. Novelli and colleagues in 1999 estimated the global H₂ source from the oxidation of NMHC compounds in the range 14 ± 7 Tg year⁻¹. This production rate is about 30% lower than the estimate by Ehhlalt. Owing to the global partitioning of the NMHC emissions, about 70% of this H₂ production is released into the Northern Hemisphere.

Ocean Surface Waters

As a result of H₂ production by microorganisms, the surface waters of the Atlantic ocean were observed to be generally supersaturated by a factor of 3.0 ± 1.2 relative to surface air. If this value is assumed to be representative for all oceans, the global oceanic emission is estimated as 3 ± 2 Tg year⁻¹. According to the land-sea partitioning, about 60% of the total oceanic emission are released into the Southern Hemisphere.

Nitrogen Fixation

Another rather small source was identified by Conrad and Seiler, who observed that H₂ is released from soils that are covered by leguminous plants. H₂ production was correlated with N₂ fixation by bacteria that live in symbiosis in the root nodules of these plants. By extrapolating the measured correlation ratio to global conditions, these authors estimated the

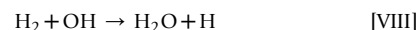
annual H₂ production from this process to be about 3 ± 2 Tg year⁻¹. Because this source is coupled to vegetation, the release rate should be largest in the Northern Hemisphere and also show an annual variation with a maximum during the main growth period.

H₂ Sinks

Because there is no unambiguous indication of any long-term trend of the global H₂ abundance, the total H₂ production must be balanced by H₂ consumption of the same global strength. To date, only two notable removal processes have been identified.

Photochemical Destruction

Like many other trace gases, H₂ reacts with OH radicals (reaction ([VIII]))



The global tropospheric distribution of H₂ is rather uniform and its seasonal variation is relatively small (less than ±5%). Since the rate coefficient of this reaction is well known from laboratory experiments and the mean global OH concentration can be reliably derived from global measurements of methylchloroform (CH₃CCl₃) the global H₂ removal rate can be estimated to a relatively good precision. For a mean OH concentration of about 10⁶ cm⁻³, the global strength of the photochemical sink has been estimated as 19 ± 5 Tg year⁻¹. This is only about 25% of the total H₂ production. Owing to the small interhemispheric gradient in the global distribution of H₂, the removal rate should be of the same magnitude in both hemispheres.

Soil Uptake

A large number of experimental studies have revealed that uptake by natural soils is most probably the largest loss process for atmospheric H₂. The deposition velocities for various types of soil determined by several authors have been reviewed; they cover a broad range from 0.01 to 0.1 cm s⁻¹ and, if they are extrapolated to global loss rates, estimates of this sink include a large uncertainty. Global uptake rates ranging from about 15 to 110 Tg year⁻¹ have been reported in the literature. The average global loss rate for soils reported by Novelli and colleagues, 56 ± 41 Tg year⁻¹, leads to an almost balanced budget of global H₂ sources and H₂ sinks. Measurements of the isotopic composition of tropospheric H₂ showed, that the global loss rate due to soil uptake should be limited to about 60 Tg year⁻¹. The content of the heavier stable isotope deuterium (D) in atmospheric H₂ is rather high and H₂ released from the known sources has a relatively low D content. Therefore, the enhanced D/H ratio in atmospheric H₂ must be due to isotope fractionation that occurs during its decomposition in the loss processes. Because a large isotopic enrichment factor was measured for the reaction of H₂ with OH but only very small isotope effects were observed during soil uptake, Ehhlalt and colleagues

concluded from available data of D/H ratio in atmospheric and source H₂ that the global strength of the two sinks should be comparable in magnitude. The photochemical removal by reaction with OH can be estimated to a reasonably good precision, and isotope data put a strong constraint on the magnitude of soil uptake. In fact, global removal by soil uptake may be overestimated by a factor of 3 if calculations are based only on H₂ deposition velocities published in the literature.

The Tropospheric H₂ Budget

A summary of the production and loss rates for atmospheric hydrogen is given in Table 1). Updated estimates derived by Ehhlalt and by Novelli and colleagues are listed for comparison. The total sources and sinks appear to be balanced within the given range of uncertainties, but for some of the individual terms the numbers given in both budgets differ substantially. In particular, the largest distinct term in the H₂ budget — the loss rate due to uptake by natural soils — is still only poorly defined.

The budget terms of the global hydrogen cycle are compiled in Table 2), again based on the work by Ehhlalt and of Novelli and colleagues. From their data set collected at 50 globally distributed sites, Novelli and colleagues derived a global mean tropospheric H₂ mixing ratio of 531 ± 6 ppbv. Owing to the deeper minima in the seasonal cycle observed in the Northern

Hemisphere, there is about 3% less H₂ in the Northern than in the Southern Hemisphere. Assuming no vertical H₂ gradient in the troposphere, the authors calculate the global tropospheric burden of H₂ to amount to 155 Tg. This is very close to the number of 150 Tg reported by Ehhlalt in 1999. Assuming a global turnover of the order of 70.0 ± 35 Tg H₂ year⁻¹, one obtains a residence time of molecular hydrogen between 1.5 and 4.3 years, with an average value of 2.3 years.

The relative strengths of the sources and sinks and their partitioning between the hemispheres provide some clues to the puzzling features of the global H₂ distribution. The major portion of the global fossil fuel emissions is released into the Northern Hemisphere. A significant decrease of the strength of this source could explain the reversal of the interhemispheric gradient of the mean H₂ mixing ratio suggested by the historical data.

The role of soil uptake as the dominant H₂ sink seems to be a plausible explanation for the deeper minima in the seasonal cycles observed in the Northern Hemisphere because about 70% of the global removal takes place on the continental land surface in this hemisphere.

No satisfactory explanation is yet available for the observation of the asymmetry in the timing of the seasonal cycles in both hemispheres. As noted by Simmonds and colleagues, the seasonal cycle in the Southern Hemisphere with a maximum during local summer 'more closely reflects the photochemical nature of the main hydrogen source term than measurements in Northern Hemisphere midlatitudes'.

Novelli and colleagues argue that the maximum in the Southern Hemisphere might additionally be driven by biomass burning in the tropics. These authors also examined the effect of the hemispheric partitioning of the major source and sink terms of the H₂ budget (see Table 1). Their results suggest that the H₂ maximum observed in the Northern Hemisphere during late winter and early spring is 'largely driven by the absence of a strong sink combined with seasonally independent technological emissions and a maximum in tropical biomass burning'. They also argue that the H₂ minimum observed in the Northern Hemisphere during fall results from the combined effects of still relatively high photochemical H₂ removal and strong soil uptake.

Although some progress has been made in the qualitative explanation of the main features of the spatial and temporal distribution of tropospheric H₂, considerable uncertainties remain in most of the terms of the global budget. Further investigations of the budget of the H₂ isotopomer HD and of the fractionation of D to H during the major production and loss processes could help to reduce the present uncertainties. More observations of the ratio of H₂ and CO increments in tropospheric background air will help to decide on the importance of decreasing H₂ emissions from fossil fuel combustion as a cause for the reversal of the interhemispheric gradient of the H₂ mixing ratio.

Table 1 Estimates of sources and sinks of tropospheric H₂

Type	Novelli et al. (1999)	Ehhlalt (1999)
Sources (Tg year ⁻¹)		
Fossil fuel use emissions	15 ± 10	20 ± 10
Biomass burning	16 ± 5	10 ± 5
Oceanic emissions	3 ± 2	3 ± 2
CH ₄ oxidation	26 ± 9	15 ± 5
NMHC oxidation	14 ± 7	20 ± 10
N ₂ fixation	3 ± 1	3 ± 2
<i>Total sources</i>	77 ± 16	71 ± 20
Sinks (Tg year ⁻¹)		
Oxidation by OH	19 ± 5	25 ± 5
Soil consumption	56 ± 41	40 ± 30
<i>Total sinks</i>	75 ± 41	65 ± 30

Table 2 The tropospheric H₂ budget

Type	Novelli et al. (1999)	Ehhlalt (1999)
Total sources (Tg year ⁻¹)	77 ± 16	71 ± 20
Total sinks (Tg year ⁻¹)	75 ± 41	65 ± 30
Mean tropospheric mixing ratio (ppbv)	531 ± 6	515 ^a
Mean tropospheric burden (Tg) ^b	155.0	150.0
Mean tropospheric lifetime (years)	2.0	2.3

^aQuoted from Khalil and Rasmussen (1990).

^bAssuming constant mixing ratios throughout the troposphere (see Schmidt (1974), Ehhlalt et al. (1977).)

See also: Chemistry of the Atmosphere: Isotopes, Stable; Methane. **Land-Atmosphere Interactions:** Trace Gas Exchange. **Tropospheric Chemistry and Composition:** Biogenic Hydrocarbons; Hydroxyl Radical.

Further Reading

- Conrad, R., Seiler, W., 1980. Contribution of hydrogen production by biological nitrogen fixation to the global hydrogen budget. *Journal of Geophysical Research* 85, 5493–5498.
- Ehhalt, D.H., 1999. Gas phase chemistry of the troposphere. In: Baumgärtel, H., Grünbein-Wand Hensel, F. (Eds.), *Global Aspects of Atmospheric Chemistry*. Dr. Dietrich Steinkopff Verlag, Darmstadt, pp. 21–110.
- Ehhalt, D.H., Schmidt, U., Heidt, L.E., 1977. Vertical profiles of molecular hydrogen in the troposphere and stratosphere. *Journal of Geophysical Research* 82, 5907–5911.
- Ehhalt, D.H., Davidson, J.A., Cantrell, C.A., et al., 1989. The kinetic isotope effect in the reaction of H₂ with OH. *Journal of Geophysical Research* 94, 9831–9836.
- Fabian, P., Borchers, R., Weiler, K.H., et al., 1979. Simultaneously measured vertical profiles of H₂, CH₄, CO, N₂O, CFC₁₃ and CF₂Cl₂ in the mid-latitude stratosphere and troposphere. *Journal of Geophysical Research* 84, 3149–3154.
- Francey, R.J., Steele, L.P., Langenfelds, R.L., et al., 1998. Atmospheric carbon dioxide and its stable isotope ratios, methane, carbon monoxide, nitrous oxide and hydrogen from Shetland Isles. *Atmospheric Environment* 32, 3331–3338.
- Heidt, L.E., Pollock, W.H., 1976. Measurements of N₂O, CH₄, H₂, CO and CO₂ in the non-urban troposphere. *Proceedings of the AGU/AMS Symposium on The Non-Urban Tropospheric Composition*. Hollywood, FL, USA.
- Herr, F.L., Barger, W.R., 1978. Molecular hydrogen in the near-surface atmosphere and dissolved in waters of the tropical North Atlantic. *Journal of Geophysical Research* 83, 6199–6205.
- Khalil, M.A.K., Rasmussen, R.A., 1990. Global increase of atmospheric molecular hydrogen. *Nature* 347, 743–745.
- Novelli, P.C., Lang, P.M., Masarie, K.A., et al., 1999. Molecular hydrogen in the troposphere: Global distribution and budget. *Journal of Geophysical Research* 104, 30427–30444.
- Prinn, R.G., Weiss, R.F., Miller, B.R., et al., 1995. Atmospheric trends and lifetime of CH₃CCl₃ and global OH concentrations. *Science* 269, 187–192.
- Schmidt U (1974, 1975) Molecular hydrogen in the atmosphere. *Tellus* 26: 78-90; 27: 93–94
- Schmidt, U., 1978. The latitudinal and vertical distribution of molecular hydrogen in the troposphere. *Journal of Geophysical Research* 83, 941–946.
- Schmidt, U., Kulessa, G., Röth, E.P., 1979. The atmospheric H₂ cycle. In: Nicolet, M., Aikin, A.C. (Eds.), *Proceedings of the NATO Advanced Study Institute on Atmospheric Ozone: Its Variation and Human Influence*. US-DOT, Washington, DC, pp. 307–322. FAA Report No FAA-EE-80-20.
- Seiler, W., Müller, F., Oeser, H., 1978. Vertical distribution of chlorofluoromethanes in the upper troposphere and lower stratosphere. *Pageoph* 116, 554–566.
- Simmonds, P.G., Derwent, R.G., O'Doherty, S., et al., 2000. Continuous high-frequency observations of hydrogen at the Mace Head baseline atmospheric monitoring station over the 1994–1998 period. *Journal of Geophysical Research* 105, 12105–12121.
- Warneck, P., 2000. *Chemistry of the Natural Atmosphere*, second ed. Academic Press, San Diego.

Hydroxyl Radical

KC Clemitshaw, Imperial College of Science, Technology, and Medicine, Ascot, UK

© 2003 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 6, pp 2403–2411, © 2003, Elsevier Ltd.

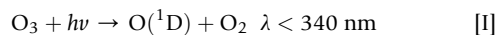
Introduction

The key roles played by hydroxyl radicals (OH) within tropospheric chemistry, including the major sources, sinks, and chemical processes involving OH, are discussed. The major daytime sources of OH include the photodissociation of ozone (O₃) and nitrous acid (HONO), the photo-oxidation of formaldehyde (CH₂O) and acetone (CH₃C(O)CH₃), together with the ozonolysis of alkenes. Nighttime processes leading to OH production also include the ozonolysis of alkenes and, more importantly, nitrate radical (NO₃)-facilitated decomposition of peroxyacyl nitrates (RC(O)O₂NO₂) and NO₃-initiated oxidation of alkenes. The major sink processes for OH are to initiate the oxidation of carbon monoxide (CO), methane (CH₄), and a wide range of reactive volatile organic compounds (VOCs), and to form nitric acid (HONO₂) via reaction with NO₂. In the presence of sufficient NO and NO₂ (collectively termed NO_x), these oxidation mechanisms are propagated by organic peroxy (RO₂), alkoxy (RO), and hydroperoxy (HO₂) radicals. They regenerate OH and proceed with the formation of oxidants such as O₃, carbonyls (e.g., CH₂O), peroxides (e.g., hydrogen peroxide, H₂O₂), and organic nitrates (e.g., peroxyacetyl nitrate, PAN, CH₃C(O)O₂NO₂). OH and HO₂ chemistry may also lead to O₃ depletion under conditions of low NO_x. Finally, a brief description of several spectroscopic techniques that have been developed for quantitative ambient measurements of OH, together with their recent applications and intercomparisons in ground-based and airborne field studies of the tropospheric chemistry of OH, is also given.

Daytime Sources of OH in the Troposphere

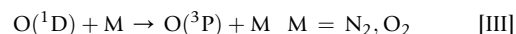
Photodissociation of O₃

The primary daytime source of OH radicals throughout much of the troposphere is the photodissociation of O₃ in its near-ultraviolet (UV) absorption band, followed by reaction of the O(¹D) photoproduct with water vapor (reactions [I] and [II]).



However, both the O₃ absorption cross sections and quantum yields of formation of O(¹D) atoms decrease with increasing wavelength. Consequently, the rate of O₃ photodissociation, $j(\text{O}(\text{}^1\text{D}))$, and therefore the rate of OH production, varies strongly with changes in atmospheric path length that not only accompany spatial variations in altitude and latitude but also temporal variations on diurnal and seasonal time-scales. Indeed, as illustrated in Figure 1, daytime measurements of $j(\text{O}(\text{}^1\text{D}))$ and OH often display a high degree of positive correlation, especially in clean air that is characteristic of

remote environments and has relatively simple photochemistry. The rate of production of OH via O₃ photodissociation is also influenced by quenching of O(¹D) to O(³P) atoms via collision with tropospheric N₂ and O₂ (reaction [III]).



The fractional conversion, f , of O(¹D) into OH is described approximately by eqn [1], where $P_{\text{H}_2\text{O}}$ is the partial pressure of water vapor and P is the total pressure.

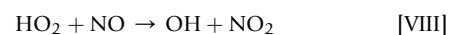
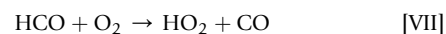
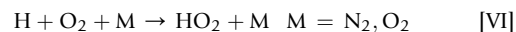
$$f = P_{\text{H}_2\text{O}} / [P_{\text{H}_2\text{O}} + 0.13(P - P_{\text{H}_2\text{O}})] \quad [1]$$

Photodissociation of HONO and Photo-Oxidation of CH₂O

Clearly, the factor, f , must be taken into account when comparing the relative importance of OH produced directly upon UV photodissociation of other precursor molecules. For HONO, photodissociation occurs on a time scale of approximately 15 min (reaction [IV]).



During the darkness of night, HONO concentrations may build up via heterogeneous hydrolysis of NO and NO₂, and as a result of direct emissions from motor vehicles. At sunrise, the photodissociation of HONO provides an early morning pulse of OH at a time of day when $j(\text{O}(\text{}^1\text{D}))$ is very low due to the long atmospheric path lengths and the low intensity of near-UV solar radiation. HO₂ is an important reactive intermediate of the photo-oxidation of CH₂O, and OH may be formed indirectly via the rapid reaction of HO₂ with reactions [V], [VI], [VII], and [VIII].



Urban environments often have high local emissions of CH₂O, whereas O₃ concentrations are generally suppressed due to rapid reaction with elevated levels of the primary pollutant, NO. Consequently, urban summertime rates of OH production via reactions [V], [VI], [VII], and [VIII] may exceed those via O₃ photodissociation by a factor of 4 or so for typical photodissociation rates and precursor concentrations.

Photo-Oxidation of CH₃C(O)CH₃

By contrast, the photo-oxidation of acetone, CH₃C(O)CH₃, which produces CH₂O as an intermediate species, has been suggested recently to be a significant source of HO₂ and OH (collectively termed HO_x) in the upper troposphere and

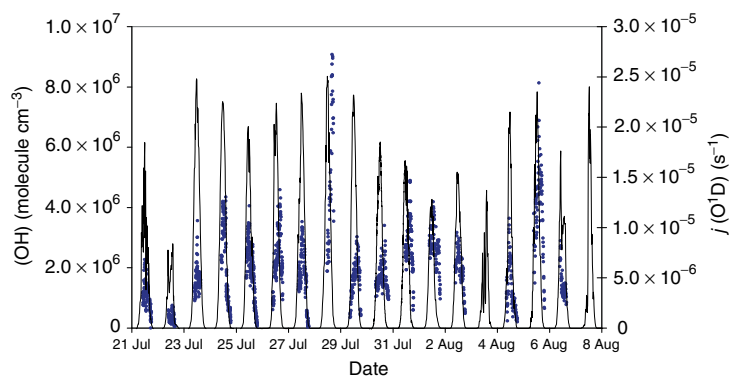
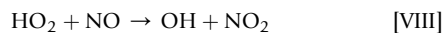
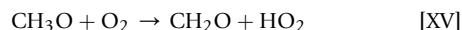
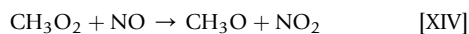
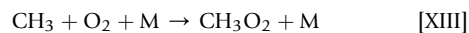
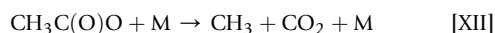
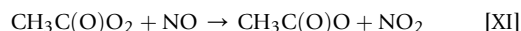


Figure 1 Simultaneous observations of OH (blue circles) and $j(\text{O}^1\text{D})$ (black line) made during the PRIME'99 campaign at the Silwood Park Atmospheric Research Station, near Ascot, UK. Data kindly provided by Professor M.J. Pilling and Dr D.E. Heard, University of Leeds, UK.

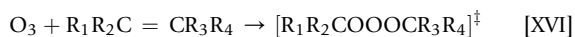
lower stratosphere (reactions [IX], [X], [XI], [XII], [XIII], [XIV], and [XV]).



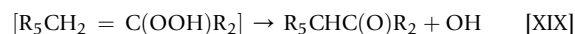
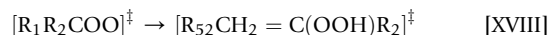
Inclusion of this process in numerical models of atmospheric chemistry reduces considerably the discrepancies between calculated and measured HO_x concentrations and O_3 production rates.

Ozonolysis of Alkenes

The ozonolysis of alkenes ($\text{R}_1\text{R}_2\text{C} = \text{CR}_3\text{R}_4$) represents another important process that produces OH, and may even dominate the photodissociation of O_3 , HONO, and CH_2O in polluted urban environments. The initial reactions of alkenes with O_3 are generally slow, but, as shown in Table 1 may proceed at rates that are comparable to corresponding reactions of alkenes with OH occurring during daytime and with NO_3 during nighttime. Although the details are not yet understood fully, kinetic and mechanistic laboratory studies indicate that the electrophilic addition of O_3 across the double bond leads to the initial formation of an energy-rich primary ozonide ($[\text{R}_1\text{R}_2\text{COOOCR}_3\text{R}_4]^\ddagger$), which decomposes rapidly to generate pairs of energy-rich, carbonyl-substituted Criegee biradicals ($[\text{R}_1\text{R}_2\text{COO}]^\ddagger$ and $[\text{R}_3\text{R}_4\text{COO}]^\ddagger$), and primary carbonyl compounds ($\text{R}_1\text{C}(\text{O})\text{R}_2$ and $\text{R}_3\text{C}(\text{O})\text{R}_4$), according to reactions (reactions [XVI], [XVIIa], and [XVIIb]).



The Criegee biradicals either collisionally stabilize, isomerize and/or decompose to form OH and other organic radicals, for example by reaction [XVIII] or reaction [XIX].

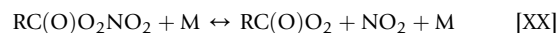


More recent laboratory studies have also shown that the fractional yields of OH produced upon ozonolysis of alkenes are significant under simulated tropospheric conditions, with approximate values ranging from 0.1 for ethene to 0.85 for α -pinene. The ozonolysis of alkenes therefore represents a potentially important daytime source of OH that may lead to the net production of O_3 under certain conditions.

Nighttime Sources of OH in the Troposphere

NO_3 -Facilitated Decomposition of $\text{RC}(\text{O})\text{O}_2\text{NO}_2$ and Oxidation of Biogenic VOCs

Clearly, OH may also be produced during nighttime upon ozonolysis of alkenes. However, of more significance at night are other non-photochemical thermal reactions that involve NO_3 . NO_3 has a lifetime with respect to photodissociation of approximately 5 s during the day, and its daytime oxidation chemistry is much less important than that of the more reactive OH. However, under conditions of low NO_x , nocturnal NO_3 chemistry may lead to OH production by facilitating the decomposition of peroxyacyl nitrates ($\text{RC}(\text{O})\text{O}_2\text{NO}_2$), as shown in reactions [XX] and [XXI].



NO_3 may also initiate the oxidation of VOCs such as alkane hydrocarbons (RH), although, as shown in Table 1, these are particularly slow reactions [XXII].

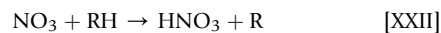
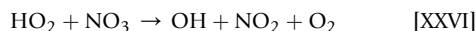
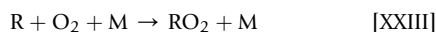


Table 1 Comparison of tropospheric lifetimes of a range of CO and VOCs at typical ambient rural concentrations with respect to reaction with 1.6×10^6 molecule cm^{-3} OH, 30 ppbv O_3 , and 10 pptv NO_3

VOC	OH	O_3	NO_3
Carbon monoxide	30 days		
Alkanes			
Methane	10 years		
Ethane	29 days		91 years
Propane	6.3 days		7.8 years
Butane	2.9 days		2.7 years
2-Methyl propane	3.1 days		1.5 years
Pentane	1.8 days		1.5 years
2-Methyl butane	1.9 days		1.3 years
Alkenes			
Ethene	20 h	9.7 days	7.3 months
Propene	6.6 h	1.5 days	4.9 days
1-Butene	5.5 h	1.6 days	3.5 days
2-Butene	2.9 h	2.4 days	2.9 h
3-Methyl propene	3.4 h	1.4 days	3.4 h
1-Pentene	5.5 h	1.5 days	3.5 days
2-Pentene	2.6 h	2.4 h	2.9 h
2-Methyl 1-butene	2.8 h	1.4 days	3.4 h
3-Methyl 1-butene	5.5 h	1.6 days	3.5 days
2-Methyl 2-butene	2.0 h	55 min	7.1 min
1,3-Butadiene	2.6 h	2.4 days	11 h
Isoprene	1.7 h	1.2 days	1.7 h
Aromatics			
Benzene	5.7 days		
Toluene	1.2 days		1.8 years
Ethyl benzene	23 h		
<i>o</i> -Xylene	12 h		4.1 months
<i>m</i> -Xylene	7.1 h		6.6 months
<i>p</i> -Xylene	12 h		3.4 months
Aldehydes			
Formaldehyde	18 h		2.7 months
Acetaldehyde	11 h		17 days
Sulfur-containing			
Dimethyl sulfide	1.5 days		1.0 h
Dimethyl disulfide	46 min		1.5 h

ppbv, parts per billion by volume; pptv, parts per trillion by volume. Adapted without permission from Jenkin, M.E., Clemitshaw, K.C., 2000. Ozone and other secondary photochemical pollutants: Chemical processes governing their formation in the planetary boundary layer. *Atmospheric Environment* 34 (16), 2499–2527.

Nevertheless, subsequent reactions ([XXIII], [XXIV], [XXIV], [XXV], and [XXVI]) lead to OH production.



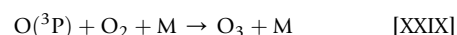
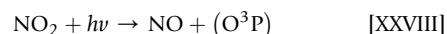
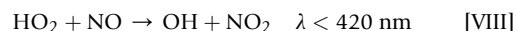
For dimethyl sulfide (DMS), which is of marine phytoplankton origin, and biogenic alkenes such as isoprene and terpenes, which are emitted by certain plants and trees, the primary step is the rapid addition of NO_3 rather than H atom abstraction. The subsequent reaction mechanisms are quite

complex, but in general, NO_3 oxidation of lesser alkyl-substituted alkenes leads to higher yields of HO_x (and bifunctional organic nitrates) and *vice versa*. Indeed, simultaneous observations of DMS, isoprene, terpenes, NO_3 , HO_2 , and RO_2 in the marine boundary layer and forested regions have provided increasingly convincing evidence for nighttime oxidation chemistry involving these species and, by implication, the production of OH. Unfortunately, confirmatory nighttime measurements of OH are relatively scarce, but concentrations of the order of $2\text{--}3 \times 10^4$ molecule cm^{-3} have been recorded on occasion. These data are a factor of $10^2\text{--}10^4$ less than typical maximum daytime OH values, and thus represent a significant observational achievement. Clearly, measurements of OH, NO_3 , HO_2 , and RO_2 , together with supporting data on O_3 , NO_x , HONO, HONO₂, and speciated VOCs and organic nitrates, would provide much improved understanding of the role of OH in nighttime oxidation chemistry.

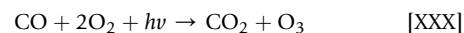
Sinks and Reservoirs of OH in the Troposphere

Oxidation of CO

Figure 2 illustrates schematically the production of OH via O_3 and HONO photodissociation, the ozonolysis of alkenes, and the photo-oxidation of CH_2O . The mechanism of the oxidation of carbon monoxide (CO) in the troposphere is also shown. It is initiated solely by reaction with OH, propagated by HO_2 , which reacts with NO to regenerate OH and produce NO_2 , and proceeds via the photodissociation of NO_2 with the formation of O_3 (reactions [XXVII], [VI], [VIII], [XXVIII], and [XXIX]).



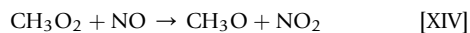
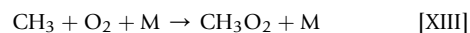
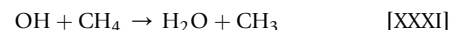
Reaction [XXX] is the overall net reaction.



This represents a chemical process that rapidly interconverts HO_x within seconds, NO_x within a few minutes, and does so without consuming either HO_x or NO_x .

Oxidation of CH_4

Similarly, as shown also in Figure 2, in the presence of sufficient NO, methylperoxy radicals (CH_3O_2), methoxy radicals (CH_3O), and HO_2 serve to propagate the OH-initiated oxidation of CH_4 . Once again OH is regenerated, in this case by reactions [XXXI], [XIII], [XIV], and [VIII], and O_3 is formed via reactions [XXXVIII] and [XXIX].



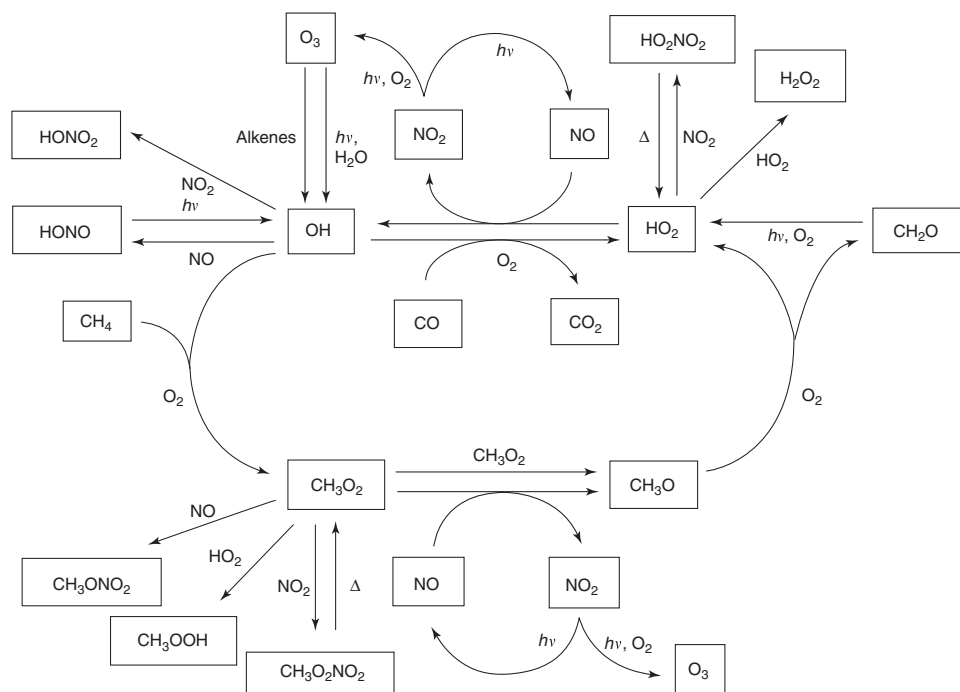
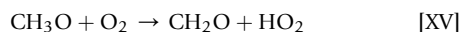
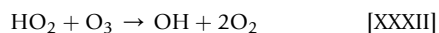


Figure 2 Daytime tropospheric chemistry illustrating sources of OH via photodissociation of O_3 and HONO, ozonolysis of alkenes, and the photo-oxidation of CH_2O . Reaction with OH initiates the oxidation of CO and CH_4 . In the presence of sufficient NO, OH is regenerated via reaction of NO with HO_2 , and O_3 is formed in cyclic reaction mechanisms. Reactions that interrupt or terminate these cycles represent loss processes for OH and HO_2 and include the production of $HONO_2$, H_2O_2 , CH_3OOH , and CH_3ONO_2 .

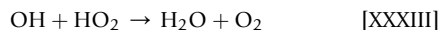


CH_2O is also produced and acts on a time scale of 5 h as a temporary reservoir species or secondary photochemical source of HO_2 (and thus OH) via reactions [V], [VI], [VII], and [VIII]. By contrast, in more remote environments with insufficient NO for reaction [VIII] to dominate the fate of HO_2 , O_3 is removed via reaction reaction [XXXII] with HO_2 , which yields OH.



Competition between reactions [VIII] and [XXXII] determines whether net production or loss of O_3 occurs.

The mutual reaction of OH and HO_2 , which is not shown in Figure 2, represents an important sink or termination process, which leads to a net loss of HO_x (reaction [XXXIII]).



By contrast, self- and analogous cross-reactions of HO_2 and CH_3O_2 lead to the formation of the HO_x reservoir species, hydrogen peroxide (H_2O_2), and methyl hydroperoxide (CH_3O_2H), via reactions [XXXIV] and [XXXV].

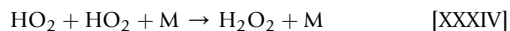
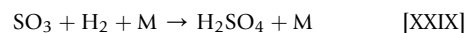
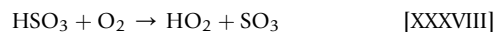
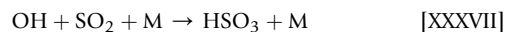


Figure 3 illustrates the percent contribution of the OH loss due to reaction with CO, CH_4 , and a wide range of biogenic and anthropogenic nonmethane hydrocarbons (NMHCs) as a diurnal cycle averaged over a field measurement campaign in

a forested region in Europe. In this case, isoprene was the dominant species reacting with OH during the day, whereas limonene and α - and β -pinene became more important during nighttime. These observations are consistent with their respective temperature- and light-dependent vegetative emission rates and OH reactivities.

Oxidation of NO_2 and SO_2

Apart from a wide range of VOCs and CO, the oxidation of inorganic compounds such as NO_2 and SO_2 is also initiated by reaction with OH, as shown by reactions [XXXVI], [XXXVII], [XXXVIII], and [XXIX].



The production of nitric acid in reaction [XXXVI] is the dominant loss mechanism for HO_x and NO_x in polluted atmospheres. Dry deposition and heterogeneous hydrolysis of NO_x are other important removal pathways. For SO_2 , however, reactions [XXXVII], [XXXVIII], and [XXIX] not only convert OH to HO_2 , but also lead to the formation of sulfuric acid (H_2SO_4), which rapidly condenses onto aerosol surfaces due to its low vapor pressure, thereby acting as a potential nucleus for new particle formation. SO_2 possesses high aqueous solubility and

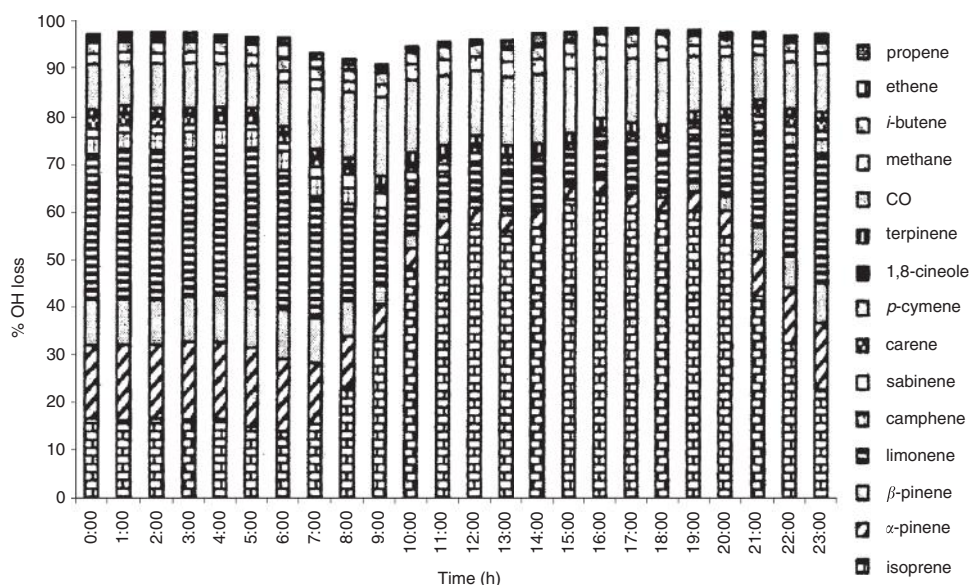


Figure 3 The percent contribution of OH loss due to reaction with CO, CH₄, and a wide range of biogenic and anthropogenic non-methane hydrocarbons (NMHCs) as a diurnal cycle averaged over a field measurement campaign in a forested region in Europe. Reproduced without permission from Carslaw N, Creasey DJ, Harrison D, *et al.* (2001) OH and HO₂ radical chemistry in a forested region of north-western Greece. *Atmospheric Environment* 35: 4725–4737.

Table 2 Global emission rates of trace gases and their proportion that is removed by reaction with a mean global OH concentration of 1×10^6 molecule cm⁻³

Trace gas	Global emission rate (Tg y ⁻¹)	Removal by OH (%)
Carbon monoxide	2800	85
Methane	530	90
Ethane	20	90
Isoprene	570	90
Terpenes	140	50
Nitrogen dioxide	150	50
Sulfur dioxide	300	30
Dimethyl sulfide	30	90

Reproduced without permission from Ehhalt DH (1999) Photooxidation of trace gases in the troposphere. *Physical Chemistry Chemical Physics* 1: 5401–5408.

is therefore also oxidized in cloud droplets and precipitation, particularly via reaction with H₂O₂. To illustrate the wide range of compounds that OH reacts with, Table 2 lists estimated global emission rates of several important trace gases and the proportion that is removed via their reaction with an OH concentration of 1×10^6 molecule cm⁻³.

Comparison between Measured and Calculated OH Concentrations

For remote, clean air environments, the high level of agreement typically observed between OH concentrations measured directly and those calculated from a hierarchy of numerical models of the chemical and physical sources and sinks of OH, indicate that these relatively simple systems are sufficiently well understood. However, perhaps it is not surprising that, in chemically more complex systems such as urban environments

with many large sources of stationary and mobile anthropogenic pollutants, and forested regions that are characterized by a variety of biogenic emissions, the level of agreement between observations and calculations is typically less good. As described above, OH plays a central role in the tropospheric chemistry of most gas-phase organic and inorganic pollutants. Quantitative measurements of OH are invaluable in the validation of the chemical mechanisms and kinetic parameters incorporated within the models. Indeed, correct predictions of previously unidentified sources and sinks of OH, and of erroneous rate coefficient measurements, have resulted from the application of numerical chemical models to the interpretation of observational data sets from several field studies.

Measurement Techniques for Field Studies of OH Radicals

High reactivity with a wide range of VOCs, rapid interconversion with HO₂, lifetimes of the order of 0.1–1 s and highly variable, diurnal, seasonal, and spatial concentrations between 10^4 and 10^8 molecule cm⁻³ mean that the quantitative detection of OH remains one of the most important yet difficult challenges in tropospheric chemistry. Nevertheless, several field-proven, spectroscopic measurement techniques have been developed over the last two decades. Each technique is described briefly below, together with recent applications and intercomparisons in ground-based and airborne field measurement studies of the tropospheric chemistry of OH.

DOAS, L-POAS, and MOAS

DOAS (Differential Optical Absorption Spectroscopy) and L-POAS (Long-Path Optical Absorption Spectroscopy) are

similar to each other but distinct from MOAS (Multipass Optical Absorption Spectroscopy). The former are usually employed to measure over integrated optical path lengths of several kilometers, whereas the latter utilizes an open, wall-less White cell with a base length of 6 m to yield data that are more comparable with local, *in situ* measurements. Concentrations of OH are derived by applying least-squares fits of OH reference spectra to the recorded differential optical density using singular value decomposition routines. A considerable advantage of these optical spectroscopic methods for ambient measurements of OH is that calibration is achieved using data for absorption cross-sections predetermined in the laboratory. Unlike other techniques, sophisticated *in situ* calibration systems are not required. Detection limits of the order of 2×10^5 molecule cm^{-3} are achievable for optical path lengths of 2–5 km, with measurement frequencies of approximately 1 to 5 min. There have been several recent reports of the application of DOAS/L-POAS and MOAS to measurements of OH in ground-based studies in remote, rural, and more polluted environments.

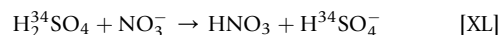
LIF and FAGE

LIF (Laser-Induced Fluorescence) and FAGE (Fluorescence Assay by Gas Expansion) are essentially identical techniques. Ambient air is expanded as a continuous, supersonic free jet through a nominal 1 mm nozzle into a low-pressure detection chamber maintained at approximately 1 mbar. A laser beam at 308 nm is used to promote ground state $X^2\Pi$ OH into the first electronically excited state, $A^2\Sigma$. At low pressure, the fluorescence lifetime of OH is increased and allows delayed-gated photon counting to capture the extremely weak OH LIF signal, while discriminating against the much more intense scattered light. Interferences arising from the photodissociation of O_3 to $\text{O}(^1\text{D})$ atoms followed by subsequent reaction of $\text{O}(^1\text{D})$ with H_2O to form OH radicals during the laser pulse is negligible using 308 nm excitation at low pressure. Note that HO_2 may also be measured by reaction with NO in the low-pressure gas expansion chamber to form OH, which are subsequently detected as described above. Calibration requires not only measurement of the response to independently quantifiable sources of OH and HO_2 coupled to the inlet system, but also determination of the efficiency of conversion of HO_2 to OH. In addition to their increasing utility in photochemical reactor chambers, there have been several recent reports of the application of LIF/FAGE to measurements of OH (and HO_2) in ground-based, shipboard, and airborne studies of tropospheric chemistry in a variety of remote, marine, rural, and urban environments.

CIMS and IMR-MS

The development of CIMS (Chemical Ionization Mass Spectrometry) for the detection of OH in the atmosphere began in the late 1980s and 1990s. IMR-MS (Ion Molecule Reaction-Mass Spectrometry) is identical. In each case, OH radicals are reacted with $^{34}\text{SO}_2$ in the presence of O_2 to produce $^{34}\text{SO}_3$, which reacts with water vapor to form $\text{H}_2^{34}\text{SO}_4$ according to reactions [XXXVII]–[XXXIX]. $\text{H}_2^{34}\text{SO}_4$ is then detected using quadrupole mass spectrometry as the highly

stable anion, $\text{H}^{34}\text{SO}_4^-$, following H-atom abstraction by NO_3^- ions (reaction [XL]).



OH concentrations of less than 1×10^5 molecule cm^{-3} can be measured in less than 1 min. These techniques have also found increasingly widespread application in recent ground-based and airborne studies of the sources and sinks of OH in the troposphere.

Other Methods

Field measurements of ambient OH concentrations may also be made using a simple, portable and inexpensive method. This method utilizes the rapid reaction of OH with *o*-hydroxybenzoic acid in a buffered solution to produce 2,5-dihydroxybenzoic acid, which is quantified by reverse-phase high pressure liquid chromatography with fluorescence detection. In clean air, a detection limit of $3\text{--}6 \times 10^5$ molecule cm^{-3} may be achieved for a sampling period of 45–90 min. Tropospheric OH concentrations may also be determined from measurements of the differential loss rates of many NMHCs, especially reactive alkenes, in well-defined, summertime urban plumes. A key assumption is that the observed hydrocarbon losses are due solely to OH radical reaction, and that plume dilution is correctly accounted for. Finally, globally averaged OH concentrations of approximately 1×10^6 molecule cm^{-3} may be estimated from measurements of the global budget of methyl chloroform (CH_3CCl_3), which is entirely of anthropogenic origin and almost exclusively removed by reaction with OH. Clearly, information about regional- and local-scale sources and sinks of OH is not provided from such estimates, but the important role of OH in tropospheric chemistry on a global scale may be deduced.

Intercomparisons of Ambient Measurements of OH

Participation in intercomparison exercises represents an often undervalued but vitally important stage in the development and validation of new instrumentation and measurement techniques. Despite the undoubted importance of making ambient measurements of OH in order to understand further its central role in tropospheric chemistry, there have been relatively few intercomparison exercises carried out over the last two decades in order to compare measurement techniques. This may have been due partly to the difficult challenges posed, but this situation may change over the next decade with further evaluation of the more recently developed techniques in order to quantify their selectivity, accuracy, precision, detection limits, sensitivity, and reliability. However, those OH intercomparison exercises that have been carried out have been informative and instructive.

For example, an intercomparison of airborne measurements of OH was recently carried out above the Pacific Ocean. LIF and CIMS data recorded on DC-8 and P3-B aircraft, respectively, during two brief close proximity flights in the marine boundary layer showed exceptionally good agreement. Flights at an altitude of 5.5 km resulted in an average concentration difference similar to the measurement uncertainties of each technique of 40%, but much less than the

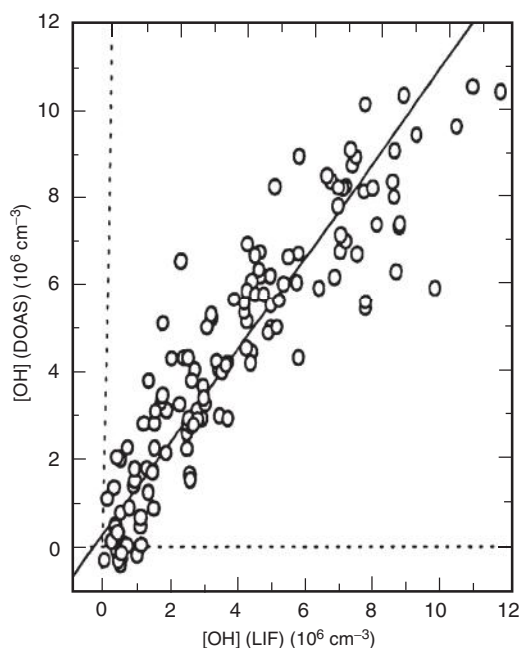


Figure 4 Bivariate plot of 137 OH data pairs measured by DOAS and LIF during the 1994 POPCORN campaign in a rural environment in north-east Germany. The correlation coefficient, r , is 0.90; the solid line represents a weighted linear fit with a gradient of 1.09 ± 0.04 and an intercept of $(0.28 \pm 0.15) \times 10^6 \text{ molecule cm}^{-3}$. Reproduced without permission from Hofzumahaus, A., Aschmutat, U., Brandenburger, U., et al., 1998. Intercomparison of tropospheric OH measurements by different laser techniques during the POPCORN campaign 1994. *Journal of Atmospheric Chemistry* 31 (1–2), 227–246.

combined uncertainties from all the contributing measurements. In each case, comparisons were made by normalizing the data with a photo-stationary state model. In addition, OH concentrations measured between 25° N to 25° S latitude over the entire longitude and altitude range of the study agreed to within 10%, although the ratio of DC-8:P-3B measurements increased at higher altitudes. These results illustrated no obvious measurement discrepancies and minimal common interferences and calibration errors for two quite distinct OH measurement techniques.

An intercomparison of extensive LIF and DOAS measurements of OH radicals in a rural environment in north-east Germany during August 1994 showed excellent agreement for the same air mass, thereby demonstrating impressively high degrees of specificity, accuracy and reliability. As shown in **Figure 4**, a linear relationship with a correlation coefficient, r , of 0.90, a gradient of 1.09 ± 0.04 and an intercept of

$0.28 \pm 0.15 \times 10^6 \text{ molecule cm}^{-3}$ was obtained from 137 data pairs. Earlier intercomparisons of ambient OH measurements using L-POAS and CIMS were carried out at Fritz Peak Observatory, Colorado, in 1991, and during the Tropospheric OH Photochemistry Experiment at Idaho Hill, Colorado, in 1993. In general, good agreement within 30% was obtained for similar chemical and meteorological conditions, which confirmed the absence of significant interferences or artefacts. A 1992 intercomparison of FAGE and an obsolete ^{14}CO radiocarbon technique not only demonstrated good correlation between data sets, with an r value of 0.86, but also revealed a significant unresolved calibration discrepancy of a factor of 2.9. Sadly, the first OH measurement intercomparison exercise in 1983–84 demonstrated the then unreliable airborne operation and insufficient detection sensitivities of LIF and the radiocarbon technique.

See also: Ozone Depletion and Related Topics: Photochemistry of Ozone; Stratospheric Ozone Recovery. **Radiation Transfer in the Atmosphere:** Ultraviolet Radiation. **Stratospheric Chemistry Topics:** HO_x .

Further Reading

- Abram, J.P., Creasey, D.J., Heard, D.E., Lee, J.D., Pilling, M.J., 2000. Hydroxyl radical and ozone measurements in England during the solar eclipse of 11 August 1999. *Geophysical Research Letters* 27 (21), 3437–3440.
- Atkinson, R., 2000. Atmospheric chemistry of VOCs and NO_x . *Atmospheric Environment* 34 (12–14), 2063–2101.
- Brauers, T., Hausmann, M., Bister, A., Kraus, A., Dorn, H.-P., 2001. OH radicals in the boundary layer of the Atlantic Ocean: 1. Measurements by long-path laser absorption spectroscopy. *Journal of Geophysical Research* 106 (D7), 7399–7414.
- Ehhalt, D.H., 1999. Photooxidation of trace gases in the troposphere. *Physical Chemistry Chemical Physics* 1 (24), 5401–5408.
- Eisele, F.L., Mount, G.H., Tanner, D., et al., 1997. Understanding the production and interconversion of the hydroxyl radical during the OH Photochemistry Experiment. *Journal of Geophysical Research* 102 (D5), 6457–6465.
- Eisele, F.L., Tanner, D.J., Cantrell, C.A., Calvert, J.G., 1996. Measurements and steady state calculations of OH concentrations at Mauna Loa observatory. *Journal of Geophysical Research* 101 (D9), 14 665–14 679.
- Finlayson-Pitts, B.J., Pitts Jr., J.N., 2000. *Chemistry of the Upper and Lower Atmosphere. Theory, Experiments and Applications*. Academic Press, San Diego, CA.
- Hofzumahaus, A., Aschmutat, U., Brandenburger, U., et al., 1998. Intercomparison of tropospheric OH measurements by different laser techniques during the POPCORN campaign 1994. *Journal of Atmospheric Chemistry* 31 (1–2), 227–246.
- Jaegle, L., Jacob, D.J., Brune, W.H., Wennberg, P.O., 2001. Chemistry of HO_x radicals in the upper troposphere. *Atmospheric Environment* 35 (3), 469–489.
- Jenkin, M.E., Clemitshaw, K.C., 2000. Ozone and other secondary photochemical pollutants: chemical processes governing their formation in the planetary boundary layer. *Atmospheric Environment* 34 (16), 2499–2527.

Mercury

J Munthe, IVL Swedish Environmental Research Institute, Göteborg, Sweden
J Sommar, Göteborg University, Göteborg, Sweden

© 2003 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 6, pp 2411–2415, © 2003, Elsevier Ltd.

Introduction

The main environmental concern associated with mercury is accumulation of methylmercury in fish, which may, if the fish is consumed, cause significant adverse effects on humans. The fetus is especially sensitive, and consumption by the mother of contaminated fish during pregnancy can lead to irreversible damage to the central nervous system of the unborn child. The biogeochemistry of mercury is complex and involves accumulation in soils and sediments, methylation processes, and finally uptake and accumulation in aquatic food chains.

Although mercury in the atmosphere is normally present at mixing ratios below parts per 10^{12} (ppt), i.e. far below any levels at which risks to human health may occur, much attention has been given to understanding the atmospheric emissions, transport, transformations, and deposition of mercury.

Mercury differs from all other metals because it is the only one that exists in liquid form at room temperature. Mercury also has a relatively high vapor pressure. In the atmosphere, the predominant form of mercury is elemental monoatomic vapor.

This species is relatively stable and has an atmospheric lifetime in the range 0.5–2 years.

Mercury is present at background atmospheric concentrations around $1\text{--}2\text{ ng m}^{-3}$. Slightly lower values are found in the Southern Hemisphere, where anthropogenic emissions are lower than in the North. During the last 10–20 years, environmental concern over mercury pollution has led to increased research efforts to describe the atmospheric pathways of mercury from emissions to deposition. **Figure 1** shows a schematic description of the main atmospheric processes involving mercury.

Emissions

Mercury is emitted from a variety of anthropogenic and natural sources. Main anthropogenic sources include coal combustion, the cement industry, chlorine manufacturing plants, and waste incineration. Source strengths will vary within each category depending on the mercury content in the raw material and the extent to which control techniques have been employed. Natural sources include volcanoes and diffuse emissions from mercury-containing mineralizations.

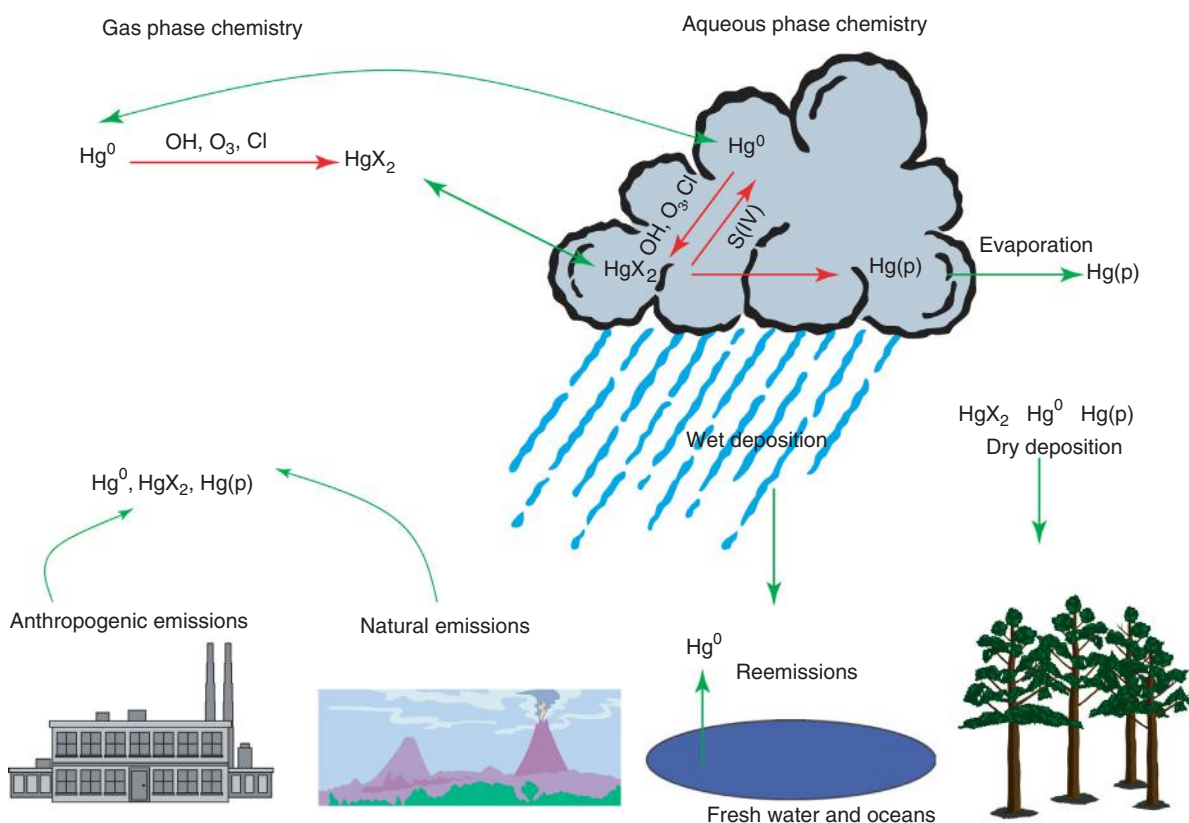


Figure 1 Schematic representation of the main atmospheric processes involving mercury.

Different emission sources emit different fractions of mercury species (see Speciation below). The global anthropogenic emissions of mercury have been estimated to be 1900 t, with Asia contributing more than 50% and Europe and North America less than 25% each.

Natural emissions and reemissions are exceedingly difficult to quantify. Emissions from natural surfaces (soils and water) may also originate from previously deposited anthropogenic mercury as well as from natural sources. The variability in time and with geographical location is also considerable. Most estimates suggest that the natural emissions are of the same order of magnitude as the anthropogenic emissions.

Speciation

A key parameter for the understanding of the atmospheric cycling of mercury is the speciation. The main form of atmospheric mercury is elemental vapor (Hg^0), making up at least 95% of the total concentration as interpreted from ground-based measurements. The remaining fraction consists of trace levels of gaseous divalent mercury compounds such as the chloride (HgCl_2) and mercury associated with particulate matter (Hg(p)). While Hg^0 is relatively stable and can be transported over hemispheric or even global scales, HgCl_2 and Hg(p) are more readily deposited and can be regarded as local or regional pollutants (i.e. deposited within 10–1000 km). Organomercury species such as monomethylmercury (CH_3HgX , XCl, OH) are present in the atmosphere at a few picograms per cubic meter. The detection of permethylated (CH_3)₂Hg in ambient air is normally restricted to biologically active marine regions of upwelling deep water.

Since the low concentrations of atmospheric mercury species frequently preclude chemical or physical identification, operational measurement techniques are often used. Measurement data are typically presented as total gaseous mercury (TGM), reactive gaseous mercury (RGM), or total

particulate mercury (TPM). These species are defined on the basis of the specific sampling or analytical procedures as follows.

- Total gaseous mercury (TGM) comprises all gaseous mercury species collected by, for example, a gold amalgam sampling device. It consists mainly of Hg^0 and traces of divalent gaseous species such as HgCl_2 .
- Reactive gaseous mercury (RGM) comprises gaseous mercury species collected by a specific sampling device such as KCl-coated denuder or mist chamber. It consists of water-soluble gaseous divalent species such as HgCl_2 .
- Total particulate mercury (TPM) comprises mercury collected by a filter device, usually quartz fiber or fluorocarbon.

The most common analytical methods are cold vapor atomic fluorescence spectrometers (CVAFS) and cold vapor atomic absorption spectrometers (CVAAS). In both cases, mercury in a sample is transformed into Hg^0 vapor and released (either thermally or by purging with an inert gas) into a measurement cell. CVAFS is a significantly more sensitive technique than CVAAS but it requires detection in argon or helium, whereas CVAAS detection can be made in air.

Transformations

A number of transformation processes of atmospheric mercury have been identified experimentally over the last 10 years. These experimental studies have demonstrated the ability of closed-shell molecules (e.g., O_3 , Cl_2) and radicals (e.g., OH, Cl, Br) to oxidize elemental mercury in the gaseous and aqueous phases. A general conclusion is that the gas-phase reactions are generally slow, while those occurring in the aqueous phase may approach the diffusion-controlled limit. Despite this, the overall atmospheric oxidation rate is relatively slow because of the low solubility of Hg^0 in water. Moreover, a reversible redox balance is present in the

Table 1 Chemical reactions and rate constants. The use of bold type indicates the reactions most relevant to the atmospheric cycling of mercury

Reaction	Rate constant ($\text{l mol}^{-1} \text{s}^{-1}$ at ambient temperature if not stated otherwise)	Medium: gaseous (g) or aqueous (aq)
$\text{Hg}^0 + \text{O}_3 \rightarrow \text{HgO} + \text{O}_2$	18 ± 12	g
$\text{Hg}^0 + \text{OH} + \text{M} \rightarrow \text{HgOH} + \text{M}$	$(5.2 \pm 2.7) \times 10^7$	g
$\text{Hg}^0 + \text{Cl} + \text{M} \rightarrow \text{HgCl} + \text{M}$	9.0×10^9 at 120–170°C	g
$\text{Hg}^0 + \text{Br} + \text{M} \rightarrow \text{HgBr} + \text{M}$	1.6×10^8 at 120–170°C	g
$\text{Hg}^0 + \text{BrO} \rightarrow \text{products}$	$\geq 2.7 \times 10^7$	g
$\text{Hg}^0 + \text{H}_2\text{O}_2 \rightarrow \text{products}$	≤ 480	g
$\text{Hg}^0 + \text{Cl}_2 \rightarrow \text{products}$	$\leq 24\,000$	g
$\text{Hg}^0 + \text{NO}_3 \rightarrow \text{HgO} + \text{NO}_2$	$\leq 2.4 \times 10^6$	g
$(\text{CH}_3)_2\text{Hg} + \text{OH} \rightarrow \text{CH}_3\text{HgOH} + \text{CH}_3$	$(1.2 \pm 0.1) \times 10^{10}$	g
$(\text{CH}_3)_2\text{Hg} + \text{Cl} \rightarrow \text{CH}_3\text{HgCl} + \text{CH}_3$	$(1.7 \pm 0.2) \times 10^{11}$	g
$(\text{CH}_3)_2\text{Hg} + \text{NO}_3 \rightarrow \text{HgO} + 2 \text{CH}_3 + \text{NO}_2$	$(4.5 \pm 1.6) \times 10^7$	g
$\text{Hg}^0 + \text{OH} \rightarrow \text{HgOH}$	$(2.4 \pm 0.4) \times 10^9$	aq
$\text{HgX} + \text{O}_2 \rightarrow \text{HgX}^+ + \text{O}_2^-$ (XOH, Cl etc.)	$\sim 10^9$	aq
$\text{HgSO}_3 + \text{H}_2\text{O} \rightarrow \text{Hg}^0 + \text{HSO}_4^- + \text{H}^+$	0.6 s^{-1}	aq
$\text{CH}_3\text{HgCl} + h\nu \rightarrow \text{CH}_3 + \text{HgCl}$	$\sim 1 \times 10^{-6} \text{ s}^{-1}$ in summer and at 60° N	aq
$\text{Hg}^0 + \text{O}_3 \rightarrow \text{HgO} + \text{O}_2$	$(4.7 \pm 2.2) \times 10^7$	aq
$\text{Hg}^0 + \text{HClO}/\text{ClO}^- \rightarrow \text{products}$	$\sim 2 \times 10^6$	aq

aqueous phase involving reduction of divalent mercury by S(IV) complexes and possibly HO₂ radicals, further lowering the overall oxidation rate. Monovalent mercury compounds are transient species formed in the reactions of mercury. Being a radical, monomeric monovalent mercury in the aqueous phase is very susceptible to oxidation by molecular oxygen. Hg(II) complexes that are formed undergo very rapid equilibration with the ligands present.

In polar atmospheres, Hg⁰ can exhibit a very different behavior. The specific photochemistry occurring during polar sunrise increases the oxidizing capacity to such an extent that the atmospheric boundary layer can be almost completely depleted of Hg⁰. The Hg⁰ depletion occurs simultaneously with that of surface ozone, which is strongly affected by autocatalytic release of bromine to the gas-phase. The strong temporal correlation between Hg⁰ and O₃ depletion suggest linked chemistry, possibly involving the same families of chemical species. Strongly oxidizing bromine monoxide is also abundant above the boundary layer, in the free troposphere, where it may influence the speciation of mercury.

The atmospheric fate of (CH₃)₂Hg is controlled by rapid chemical degradation to products that are mainly removed via dry and wet deposition. The major reaction products consist of CH₃HgX derivatives, while a complete cleavage of both Hg–C bonds appears to occur when the oxidizing agent can transfer oxygen atoms.

To evaluate the importance of different oxidizing reactions, the chemical lifetime of Hg⁰ ($= k^{-1}[\text{oxidant}]^{-1}$) can be calculated using the rate constants presented in Table 1 and the relevant concentration ranges in which the oxidants are present. The resulting lifetimes are presented in Figure 2.

Air–Surface Exchange

Interactions with water, vegetation, and soil surfaces are an important characteristic of the atmospheric cycling of mercury. Significant emissions (or reemissions of previously deposited mercury) occur from natural waters. In principle, oxidized mercury deposited via wet or dry processes can be reduced in surface waters and reemitted to the atmosphere. Mercury also interacts with vegetation via dry deposition and reemissions.

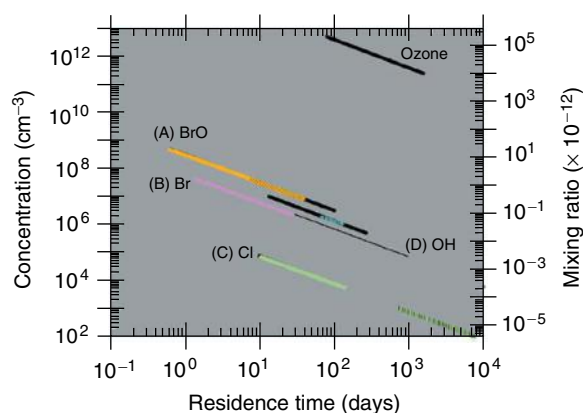


Figure 2 Residence time of Hg⁰ in the troposphere due to gasphase chemical degradation. The ranges of mixing ratios are taken from the literature. No temperature correction of the rate of reactions given in Table 1 has been made. (A) Light line corresponds to conditions during the Polar spring ozone depletions and hatched line to concentrations measured in the free troposphere. (B) Applies only to polar tropospheric ozone depletions. (C) Light line corresponds to conditions during the polar spring ozone depletions and hatched line to estimates of annual global mean concentrations. (D) Hatched line indicates annual global mean concentrations.

Wet Deposition

While dry deposition of gaseous divalent mercury and particulate phase mercury can contribute significantly to the overall deposition flux, the main removal process of atmospheric mercury is wet deposition. Gaseous divalent or particulate mercury is either incorporated directly into clouds or washed out by precipitation.

Occurrence

Mercury is a natural component of the atmosphere. Anthropogenic activities have increased the atmospheric concentrations and deposition fluxes by a factor of 2–5. Typical concentrations of atmospheric mercury species are presented in Table 2.

Table 2 Typical concentrations of mercury species in the planetary boundary layer

Species	Concentration range	Location
Hg ⁰	0.5–1.2 ng m ⁻³ 1.1–1.8 ng m ⁻³ 1.5–15 ng m ⁻³	Atlantic air, Southern Hemisphere Atlantic air, Continental background, Northern Hemisphere Continental air, urbanized, industrial
RGM	–30 pg m ⁻³ (very variable)	Background air, marine and continental, higher near sources
TPM	0.1–5 pg m ⁻³ 5 to > 50 pg m ⁻³	Background air Continental background, higher near sources.
CH ₃ HgX	0.1–5 pg m ⁻³	Background air
(CH ₃) ₂ Hg	<5 pg m ⁻³	Background air
Hg(II) in precipitation	–30 pg m ⁻³ (very variable and normally <5 pg m ⁻³) 1–20 ng l ⁻¹	Marine polar air

RGM, reactive gaseous mercury; TPM, total particulate mercury

See also: **Chemistry of the Atmosphere:** Volcanoes: Composition of Emissions. **Land-Atmosphere Interactions:** Trace Gas Exchange. **Tropospheric Chemistry and Composition:** Aerosols/Particles; Aliphatic Hydrocarbons; Aromatic Hydrocarbons; Biogenic Hydrocarbons; H₂; Hydroxyl Radical; Oxidizing Capacity; Peroxyacetyl Nitrate; Volatile Organic Compounds Overview: Anthropogenic.

Further Reading

- Lamborg, C., Fitzgerald, W.F., O'Donnel, J., Torgersen, T., 2002. A non-steady-state compartmental model of global-scale mercury biogeochemistry with interhemispheric atmospheric gradients. *Geochimica et Cosmochimica Acta* 66, 1105–1118.
- Lin, C.J., Pehkonen, S.O., 1999. The chemistry of atmospheric mercury: a review. *Atmospheric Environment* 33, 2067–2079.
- Mason, R.P., Fitzgerald, W.F., Morel, F.M.M., 1994. The biogeochemical cycling of elemental mercury – anthropogenic influences. *Geochimica et Cosmochimica Acta* 58, 3191–3198.
- Schroeder, W.H., Anlauf, K.G., Barrie, L.A., et al., 1998. Arctic springtime depletion of mercury. *Nature* 394, 331–332.
- Schroeder, W.H., Munthe, J., 1998. Atmospheric mercury – an overview. *Atmospheric Environment* 32, 809–822.
- US EPA Mercury Report to Congress. Volumes II and III dealing with atmospheric issues. Available at <http://www.epa.gov/airprog/oar/mercury.html>.

Oxidizing Capacity

DH Ehhalt, F Rohrer, and A Wahner, Forschungszentrum Jülich, Jülich, Germany

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by D H Ehhalt, A Wahner, volume 6, pp 2415–2424, © 2003, Elsevier Ltd.

Synopsis

The oxidizing capacity of the troposphere relies on oxidants such as O₃, NO₃, and OH. Of these, OH is by far the most important. The supply of all the oxidants is limited and thus the oxidizing capacity is finite. The factors that control OH, O₃, and NO₃ are reviewed and examples of their tropospheric distributions are presented. The resulting lifetimes of the most important tropospheric trace gases are listed. The possible change of the oxidant distributions under the impact of anthropogenic emissions is indicated. A quantitative measure of the oxidation capacity is derived from the total loss rate of all trace gases due to OH.

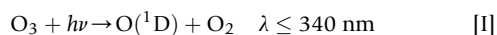
Introduction

The term 'oxidizing capacity' is widely used, but loosely defined. Its attraction is that it conveys in a succinct manner, two major features of atmospheric chemistry: namely, that the atmosphere actively oxidizes gaseous trace constituents and pollutants, and thereby initiates their removal from the atmosphere; and also, that this capacity is finite. The first feature is not unexpected; after all molecular oxygen, O₂, is a major constituent of the Earth's atmosphere. The second is somewhat surprising – considering that O₂ is present at about 21%, whereas the most abundant oxidizable trace gases are present at ppm level. The explanation is, of course, that O₂ does not directly react with those molecules. Tropospheric oxidation is rather initiated by a number of oxidants, above all by the hydroxyl radical, OH, but also by ozone, O₃, the nitrate radical, NO₃, and to a lesser extent by chlorine (Cl) and bromine (Br) atoms. Ground state atomic oxygen, O(³P), plays only a very minor role. In the liquid phase, for example in cloud droplets, hydrogen peroxide also acts as an oxidant. All these molecules and radicals are generated within the troposphere, and the rates of the respective generation processes limit their supply and thus the oxidizing capacity of the troposphere.

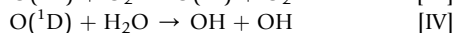
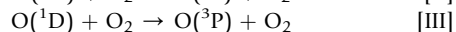
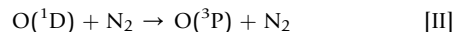
In the following, we briefly review the factors that control the tropospheric concentrations of the major oxidants, OH, O₃, and NO₃. We present examples of their tropospheric distributions, and indicate how these may have changed under the impact of anthropogenic emissions. From the superposition of the various loss processes, we derive the atmospheric lifetimes for the most important tropospheric trace gases. Finally, based on all this information we present a measure of the tropospheric oxidizing capacity through an approximate but simple expression.

The Tropospheric Chemistries of OH and O₃

The tropospheric chemistries of OH and O₃ are very closely interlinked. In fact, the primary production of OH is based on the photolysis of O₃. At wavelength below 340 nm, this photolysis yields electronically excited oxygen atoms, O(¹D) (eqn [I]):

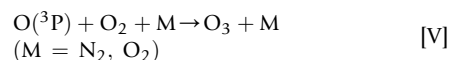


Most of the O(¹D) atoms produced in reaction [I] are quenched to ground state atomic oxygen, O(³P), in collisions with atmospheric nitrogen and oxygen molecules (eqns [II] and [III]), but a fraction collides with water vapor molecules, to form the very reactive hydroxyl radical, OH (eqn [IV]).

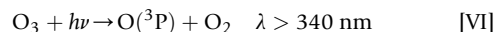


H₂O is ubiquitous in the troposphere. For instance, in surface air at midlatitudes, H₂O is present at a volume mixing ratio of about 1%. There about 10% of the O(¹D) generate OH, because reaction [IV] has a rate constant of approximately a factor of 10 higher than the quenching reactions [II] and [III].

The conversion of O₃ to OH via the reactions [I] and [IV] at the same time forms a major sink of tropospheric O₃, whereas the O(³P) formed by [II] and [III] returns to O₃ by recombination with O₂ (eqn [V]).



The same holds for the O(³P) formed directly in the photolysis of O₃ at wavelengths longer than 320 nm (eqn [VI]).



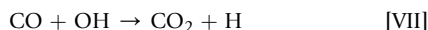
The generation of O(³P) mainly by photolysis (eqn [VI]) but also by reactions [II] and [III] and others is balanced by the loss due to recombination. During daytime this results in a steady-state concentration of a few thousand O(³P) per cm³. Given the relatively low reactivity of O(³P), this concentration is too small to be of significance for the oxidation of trace gases in the troposphere. The fact that the O atoms generated in reaction [VI] interact rapidly with one of the major constituents of air limits their possible importance.

Hydroxyl radicals, on the other hand, do not react with any of the major constituents of air. Rather, they have the ability to initiate chain reactions in an O₂-containing atmosphere. When reacting with trace gas molecules, OH is not consumed, but is regenerated in catalytic cycles. In this way, relatively large OH concentrations, up to 10⁷ cm⁻³, are maintained in the sunlit troposphere despite the high reactivity of OH toward most pollutants and other trace gases. These two properties, high

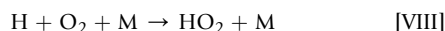
reactivity and relatively high concentrations, make OH the most important oxidizing agent in the troposphere.

OH Reaction with Molecules

The simplest example of such a reaction is given by the atmospheric oxidation of carbon monoxide. The reaction of CO with OH immediately forms the stable end product CO₂ (eqn [VII]).



The reaction also forms a hydrogen atom, which is very reactive. The hydrogen atom rapidly combines with O₂ to form a hydroperoxy radical, HO₂ (eqn [VIII]).

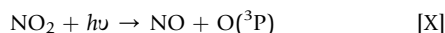


The addition of a hydrogen atom to O₂ weakens the bond between the oxygen atoms, and HO₂ reacts much more readily than O₂. In particular, HO₂ oxidizes nitric oxide, NO (eqn [IX]).



In the planetary boundary layer over the industrialized and highly populated continents, where daytime concentrations of NO exceed 0.1 ppb, reaction [IX] is by far the fastest HO₂ reaction. Most important for our present argument is the fact that the OH radical consumed in reaction [VII] is regenerated in reaction [IX]. This is generally true: virtually all reactions of OH with molecular species lead to chain reactions that recycle OH. Reaction [IX] also highlights one of the roles of the nitrogen oxides for the OH budget. In the form of NO they quickly return the less reactive HO₂ back to the highly reactive OH, thus increasing the OH concentration.

The nitrogen dioxide molecule, NO₂, generated in reaction [IX] photolyzes readily in the near-UV region and therefore contributes to tropospheric photochemistry (eqn [X]).



In the sunlit atmosphere, the lifetime of NO₂ against photolysis is a few minutes. The resulting O atom immediately combines with O₂ to form O₃ (reaction [V]). This NO-mediated process consisting of reactions [IX], [X], and [V], in which an O atom from a peroxy radical is passed along to O₃, is the major ozone formation mechanism in the troposphere. It illustrates the fact that in an NO-containing atmosphere, trace gas removal is invariably linked with the production of O₃ or other photo oxidants. Supplemented by reactions [X] and [V], the hydroxyl radical reactions [VII]–[IX] combine to the net reaction shown in eqn [XI].

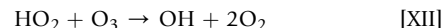


It does not consume OH, HO₂, NO, and NO₂, hence the cycle consisting of the reactions [VII]–[X] and [V] can be completed repeatedly before that chain reaction is interrupted by termination reactions. By this and other chain reactions, OH and HO₂ are interconverted within a matter of seconds. Therefore, both are often lumped together as HO_x.

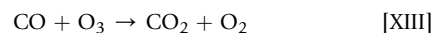
The oxidation of more complex molecules, such as hydrocarbons, leads to other peroxy radicals, RO₂, which also convert NO to NO₂ and augment the generation of O₃. In addition, intermediates such as formaldehyde (HCHO) are formed,

whose photolysis leads to formation of HO_x and thus introduces a degree of autocatalysis.

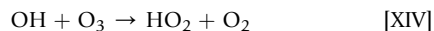
Parallel to reaction [IX], HO₂ reacts directly with O₃ to return OH (eqn [XII]).



Thus even without NO there is a certain measure of recycling. But in this case the net result of the reaction chain, [VII], [VIII], and [XII], is a destruction of O₃ (eqn [XIII]).

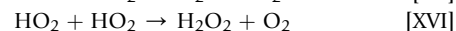
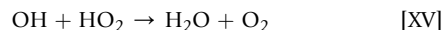


Together with reaction [XIV], reaction [XII] forms the second important loss mechanism of tropospheric O₃, again closely tied to the HO_x chemistry.

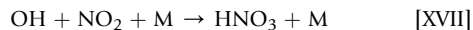


Destruction of HO_x

To produce a net loss of HO_x and terminate the reaction chains, HO_x radicals have to react with other radicals. Eqns [XV] and [XVI] are responsible for HO_x loss in clean air.



The addition of OH to NO₂, forming a nitric acid molecule, HNO₃ (eqn [XVII]), is the dominant HO_x loss reaction in the polluted atmosphere.



At the same time, reaction [XVII] provides the major loss mechanism for NO_x. It also illustrates the tendency of atmospheric oxidation to produce acidic end products, which are eventually removed from the atmosphere by rainout.

The OH Balance Equation

The various production and destruction reactions of OH can be combined in a local OH budget equation of the type shown as eqn [1], where P_{OH} and D_{OH} stand for the local production and destruction terms of OH.

$$\frac{d[\text{OH}]}{dt} = P_{\text{OH}} - D_{\text{OH}} \approx 0 \quad [1]$$

In the case of a very short-lived species like OH, contributions from transport, i.e., the flux divergence, are small compared to P_{OH} and D_{OH} , and can be neglected. The same holds for the temporal change, $d[\text{OH}]/dt$.

P_{OH} consists of two, systematically different, terms: the primary production of OH (P_{OH}) through the photolysis of precursors such as O₃ or H₂O₂ and the OH produced from recycling HO₂ (Π_{OH}).

By far the largest contribution to P_{OH} on a global scale comes from the photolysis of O₃, which can be derived from reactions [I]–[IV] to give eqn [2].

$$P_{\text{OH}} = \frac{2[\text{O}_3] \cdot [\text{H}_2\text{O}] \cdot J_1 \cdot k_4}{k_2[\text{N}_2] + k_3[\text{O}_2] + k_4[\text{H}_2\text{O}]} + \dots \quad [2]$$

J_1 is the photolysis frequency of reaction [I], and the k_i refers to the rate constants of reactions [III]–[IV]. Smaller contributions,

such as that from the photolysis of H_2O_2 , are indicated by the ellipsis '...'.

Π_{OH} is determined by reactions [IX] and [XII] as eqn [3].

$$\Pi_{\text{OH}} = [\text{HO}_2] \cdot (k_9[\text{NO}] + k_{12}[\text{O}_3]) \quad [3]$$

D_{OH} , the destruction of OH, is given by eqn [4].

$$\begin{aligned} D_{\text{OH}} &= [\text{OH}] \cdot (k_7[\text{CO}] + k_{14}[\text{O}_3] + \cdots \\ &\quad + k_{17}[\text{NO}_2] + k_{15}[\text{HO}_2] + \cdots) \\ &= [\text{OH}] \cdot \tau_{\text{OH}}^{-1} \end{aligned} \quad [4]$$

A complete chemical mechanism would contain contributions to the production and destruction of OH other than those listed above. Again, these are indicated by the dotted ellipsis '...'. These additional terms would, however, fall in the same categories distinguished here. The first line in eqn [4] contains the reactions of OH with molecules, i.e., those that eventually generate HO_2 ; the second line contains the reactions of OH with radicals that lead to a net HO_x loss. The whole term in parenthesis in eqn [4] represents the pseudo-first-order reaction frequency or inverse lifetime, τ_{OH}^{-1} , of an OH radical.

We now define the recycling ratio $\rho = \Pi_{\text{OH}}/P_{\text{OH}}$. This is the number of HO_2 radicals converted to OH divided by the number of OH generated directly. We note that the precursors for direct OH production, mainly O_3 but also H_2O_2 , are also produced by OH chemistry, although not necessarily at the same location, and, thus, contain an element of recycling. Because of their long lifetime their recycling is slow and incomplete and their local concentration is largely determined by transport. We further note that Π_{OH} also includes a slow component, namely HO_2 derived from aldehyde photolysis, besides the fast recycling of HO_2 generated immediately after the OH attack. Nonetheless, for most of the troposphere, ρ also reasonably approximates the chain length, that is, the number of times an OH is recycled via HO_2 before it is removed by a radical–radical reaction. Introducing τ_{OH} and ρ , eqn [1] takes the simple form of eqn [5a] or eqn [5b].

$$P_{\text{OH}}(1 + \rho) - \frac{[\text{OH}]}{\tau_{\text{OH}}} = 0 \quad [5a]$$

or

$$[\text{OH}] = P_{\text{OH}} \cdot \tau_{\text{OH}}(1 + \rho) \quad [5b]$$

Equation [5b] demonstrates nicely that the ability of OH for chain reaction and slow recycling enhances its concentration by a factor $(1 + \rho)$ over that of a nonchain-reacting radical, whose steady-state concentration would be given merely by the product of its production and lifetime, $P \cdot \tau$. Tropospheric values for ρ vary between 0.3 and 10 depending on NO_x in a manner similar to that of OH shown in Figure 1. We note that the factors ρ and τ_{OH} still depend on the HO_x concentration. Thus, eqn [5b] by itself is not sufficient for calculating the OH concentration. It helps, however, to categorize the action of the various reactions on OH. The nitrogen oxides, $\text{NO}_x = \text{NO} + \text{NO}_2$, for example, act on OH in two different ways. In the form of NO, they accelerate the recycling from HO_2 to OH (reaction [IX]; cf eqn [3]) and thus enhance the concentration of OH; in the form of NO_2 they enhance the net loss of OH via reaction [XVII], and thus decrease τ_{OH} and

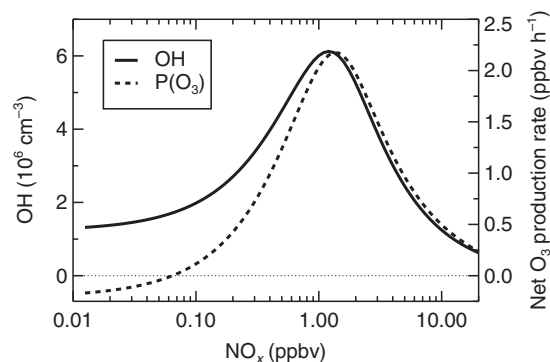


Figure 1 Dependence of the OH concentration and net O_3 production on NO_x calculated with a steady-state model for remote, rural conditions. Adapted from Ehhalt, D.H., 1999b. Photooxidation of trace gases in the troposphere. *Physical Chemistry Chemical Physics* 1, 5401–5408.

therefore the concentration of OH (cf eqn [4]). The opposing actions of the two processes lead to a highly nonlinear dependence of the OH concentration on NO_x (see Figure 1). At low NO_x , where recycling dominates, the addition of NO_x leads to an increase in OH concentration and thus to an increase in photochemical activity; at high NO_x , reaction [XVII] eventually becomes the dominant loss path of OH and further addition of NO_x reduces the OH concentration. The position of the maximum with respect to NO_x depends to some extent on the local mix of reactive trace gases. Except in polluted areas, however, tropospheric chemistry operates in NO_x regimes defined by the left flank of the OH curve in Figure 1; that is, an increase in global NO_x concentrations will lead to an increase in global OH.

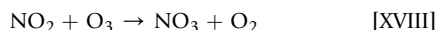
The dependence of OH on other parameters is more monotonic, though still not linear. For the NO_x regime of the global troposphere, an increase in CO or other oxidizable molecules will lead to a decrease in OH (cf reaction [VII]; eqn [4]). But a change in H_2O will lead to a change of equal sign in OH (cf eqn [2]). Solar radiation acts through several channels: through P_{OH} via the photolysis of O_3 , but also through ρ and τ_{OH} , because photolysis of NO_2 shifts the $[\text{NO}]/[\text{NO}_2]$ ratio. These factors all conspire to give a nearly linear relationship between tropospheric OH and solar UV radiation.

Figure 1 also shows the dependence on NO_x of the net formation rate of O_3 given by $d[\text{O}_3]/dt = P_{\text{O}_3} - D_{\text{O}_3}$, where the local O_3 production, P_{O_3} , is maintained by reaction cycles such as that consisting of [VII]–[X] and [V]. The O_3 destruction, D_{O_3} , is given by reactions such as [XIII], [XIV], or [I] combined with [IV].

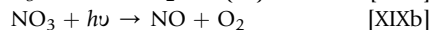
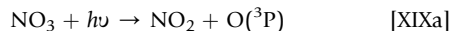
In the clean environment considered in Figure 1, the net O_3 production rate is relatively small; even at the maximum it amounts to no more than 2.3 ppbv h^{-1} . The profile of the net O_3 formation rate is similar to that of the OH concentration, i.e., to the photochemical activity. At low NO_x concentrations, however, it begins to deviate significantly, because there the O_3 -destroying reactions, such as reaction [I] combined with reaction [IV], remain active, whereas the O_3 -producing reactions, such as the sequence [VII]–[X], decrease with decreasing NO_x . Eventually, below an NO_x concentration of about 0.07 ppb, photochemistry leads to a net destruction of O_3 .

Chemistry of NO₃

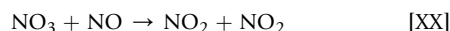
The nitrate radical, NO₃, is formed primarily by the relatively slow reaction [XVIII].



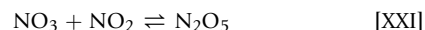
However, in daylight NO₃ is photolyzed within tens of seconds via reactions [XIXa] and [XIXb].



In addition, NO₃ reacts rapidly with NO (eqn [XX]) that is always present during daytime owing to photolysis of NO₂ (reaction [X]) such that during daytime NO₃ remains at very low concentrations.



During nighttime, however, NO₃ can build up to significant concentrations because the remaining loss reactions are much slower. These are (1) the heterogeneous uptake of NO₃ by moist aerosol surfaces, fog, or cloud droplets; and (2) the heterogeneous uptake of N₂O₅, which is formed from combination of NO₃ with NO₂ (eqn [XXI]).



This reaction is important especially at low temperatures since the thermal decomposition of N₂O₅ to NO₂ and NO₃ is strongly temperature dependent. N₂O₅ reacts with condensed water to form HNO₃. It thus reacts rapidly upon collision with the wet surfaces such as cloud droplets or aerosol particles at relative humidities above 60%. In nights with high relative humidities, this reaction path together with the direct uptake of NO₃ provides an efficient loss of NO_x.

Finally, NO₃ reacts with organic molecules in a manner similar to OH but for most molecules much more slowly than the corresponding reactions with OH. Nevertheless, in polluted regions it can provide a significant loss for organic molecules.

Global Distributions of OH, O₃, and NO₃

The tropospheric chemistry outlined above maintains global distributions of OH, O₃, and NO₃ that vary with latitude, longitude, daytime, and season. For reference, Figure 2 presents the mean zonal distributions of OH, O₃, and NO₃ for January

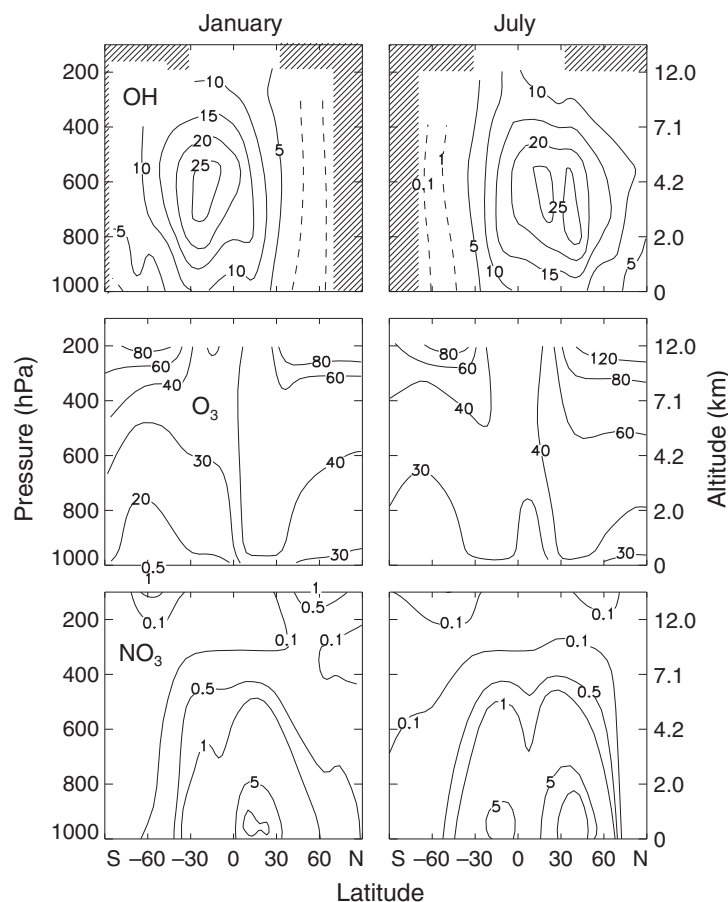


Figure 2 Mean zonal distributions of OH, O₃, and NO₃ for January and July based on 3D chemical transport models. OH distribution (top panels, contour lines $\times 10^5$ OH cm⁻³) adapted from Spivakovsky, C.M., Logan, J.A., Montzka, S.A., et al., 2000. Three-dimensional climatological distribution of tropospheric OH: update and evaluation. *J. Geophys. Res.* 105, 8931–8980. O₃ distribution (middle panels, contour lines ppbv) adapted from Wang, Y., Jacob, D.J., Logan, J.A., 1998. Global simulation of tropospheric O₃-NO_x-hydrocarbon chemistry. *Journal of Geophysical Research* 103, 10757–10767. NO₃ distribution (bottom panels, contour lines pptv), private communication Hauglustaine, D., 2000.

and July. They are based on 3D chemical transport models, since measurements are too sparse to allow a representative and consistent reconstruction of the average concentration fields. Nevertheless, the concentration fields shown are expected to closely approach reality. For O_3 , the measured seasonal variations, as well as height profiles, show good agreement with those modeled for the same sites. For OH, the globally averaged concentration agrees well with the value of $1.0 \times 10^6 \text{ cm}^{-3}$ derived empirically from the budget of methyl chloroform, CH_3CCl_3 .

The main oxidant, OH, shows a strong latitudinal variation at all altitudes, with a broad maximum in the tropics centered at the latitude of maximum solar radiation. It also varies with altitude with a broad maximum around 4 km at all latitudes. In addition, OH, as well as the other oxidants, O_3 and NO_3 , exhibits higher concentrations in the Northern Hemisphere whose continents provide most of the global emissions, whether natural or artificial. For the same reason, the longitude by latitude presentation (Figure 3) also indicates higher concentrations of OH over the continents.

The chemistry outlined above also predicts that the increase of the anthropogenic emissions of NO_x , CO, and CH_4 that accompanied the industrialization and population growth of the last century must have caused significant changes in the concentrations of the tropospheric oxidants. There is indeed

experimental evidence mainly from historical records over Europe that the concentration of O_3 has increased by about a factor of two over that period. Figure 4 presents the model-calculated change in the zonally and annually averaged distribution of O_3 and OH from preindustrial to present time. For O_3 the models predict a general increase of about 60%, which is, however, lower in the Southern Hemisphere (+50% in the case presented) than in the Northern Hemisphere (+80%). The model-calculated globally averaged OH decreased by about 10%, a relatively small value. However, as Figure 4 indicates, the changes are quite unevenly distributed, with large zones of substantial positive change. The OH concentration over the continents increased strongly, whereas it decreased over the oceans, mapping the nonuniform distribution of NO_x , which is short lived and mostly emitted over the continents. In fact, the global distribution of OH, as well as its trend, strongly depends on the assumed spatial and temporal distribution of the NO_x emissions. The most recent analysis of the CH_3CCl_3 budget suggests a relatively constant mean global OH concentration over the past 20 years with a small interannual variability of $2.1 \pm 1.8\%$ per year, but no significant long-term trend.

The difference between the global change for OH and O_3 is not unexpected. For O_3 the increased emissions of CO and NO_x reinforce each other, as more CO leads to higher formation of

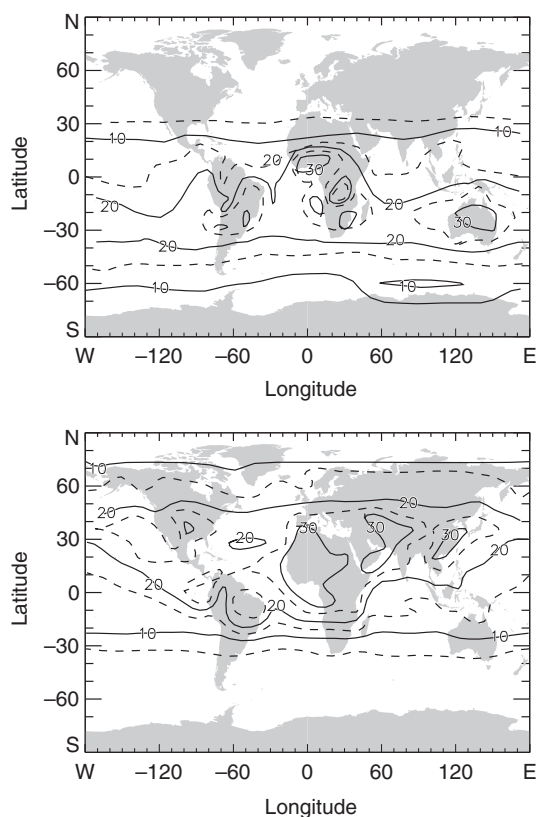


Figure 3 Global distribution of OH for January (top panel) and July (bottom panel) at 700 mb (contour lines $\times 10^5 \text{ OH cm}^{-3}$). Adapted from Spivakovsky, C.M., Logan, J.A., Montzka, S.A., et al., 2000. Three-dimensional climatological distribution of tropospheric OH: update and evaluation. *Journal of Geophysical Research* 105, 8931–8980.

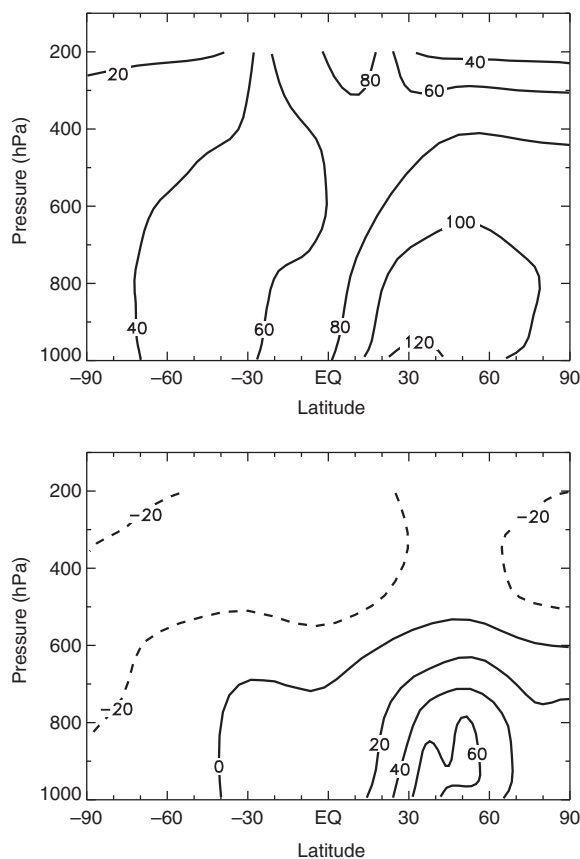


Figure 4 Calculated relative change in the zonally and annually averaged distribution of O_3 (top panel) and OH (bottom panel) from preindustrial to present times (contour lines in %). Adapted from Wang, Y., Jacob, D.J., 1998. Anthropogenic forcing on tropospheric ozone and OH since preindustrial times. *Journal of Geophysical Research* 103, 31123–31135.

HO₂, which in turn is efficiently converted to O₃ by the enhanced NO. For OH they counteract each other – increased NO_x induces a larger and increased CO a lower OH concentration. For NO₃ one would also expect a global increase in the concentration field since preindustrial times, since both the NO₂ and O₃ concentrations have increased globally. However, no hindcasts have been published for NO₃.

Obviously, as long as the emissions of NO_x, CO, and CH₄ continue to grow, further changes in the global distributions of OH, O₃, and NO₃ have to be expected. For example, for the future emission scenario A1F1 from IPCC 2001, which presents one of the worst cases and assumes global emissions of NO_x, CO, VOC, and CH₄ of 243, 193, 198, and 128%, respectively, of the values in the year 2000, model calculations predict a decrease in the mean global OH of –14% and an increase in O₃ of 47% by the year 2100.

The Global Loss Rate

The removal of a given trace gas from the troposphere is usually generated by several different processes. In addition to the tropospheric oxidation by OH, O₃, NO₃, and other radicals, direct photolysis, dry deposition at the Earth's surface, and export to the stratosphere may play a role. For a gas X_i, the global tropospheric loss rate, L_i, is given by the sum of all the individual loss rates integrated over the volume of the troposphere and averaged over a full year. Correspondingly, the global tropospheric lifetime of a trace gas, τ_i, is defined as its tropospheric burden, M_i, divided by its global tropospheric loss rate. The contribution of the major oxidants to the global loss rate of X_i is given by eqn [6], where the brackets denote the local concentrations in units of cm^{–3}, the k_i are the respective rate constants, which generally depend on temperature and thus also on location and season, ν stands for tropospheric volume, t for time, and a = 1 year.

$$L_i = \frac{1}{a} \int_{\nu, t} (k_{i,OH} \cdot [OH] + k_{i,O_3} [O_3] + k_{i,NO_3} \cdot [NO_3]) \cdot [X_i] d\nu \cdot dt$$

$$= M_i \tau_i^{-1} \quad [6]$$

Examples of the rate constants are given in Table 1 for a few selected organic trace molecules. Clearly, the rate constants vary over many orders of magnitude, reactions with OH having the highest and those with O₃ the lowest rate constants. We also note that the ratios k_{OH}:k_{O₃}:k_{NO₃} vary substantially between the various trace gases. Thus the relative contributions of the global OH, O₃, and NO₃ fields to the global oxidation rate vary greatly with the trace gas considered.

Generally, however, tropospheric oxidation is dominated by reaction with OH. This is emphasized by Table 2, which lists the trace gases with the highest turnover in the atmosphere. The most important three trace gases, namely CO, CH₄, and H₂, are not attacked by O₃ or NO₃. In fact, significant direct oxidation by O₃ is limited to the terpenes and NO_x; and that by NO₃ to the terpenes and dimethyl sulfide. In terms of global tropospheric removal, direct attack of O₃ and of NO₃ adds up to a loss rate of 3.75 Tmol per year, i.e., about 2% of the total loss rate of 180 Tmol per year given by the sum over the individual trace gas losses in Table 2 and the oxidation of their secondary products. This does not preclude that oxidation by NO₃ and O₃ becomes important on a local scale, or for certain molecules even on a global scale, but its impact on the total turnover of oxidizable molecules is small. Similarly, the tropospheric role of Cl and Br atoms appears small. Based on the global budgets of ethane and tetrachloroethene, the average Cl atom concentration in the Northern Hemisphere has been estimated to less than 1000 cm^{–3} and in the Southern Hemisphere less than 2000 cm^{–3}. This is three orders of magnitude smaller than the globally averaged OH concentrations. This difference is so large that it cannot be compensated by the generally higher rate constants of hydrocarbons with Cl atoms. Finally, the role of H₂O₂ is pretty much limited to the oxidation of SO₂ in the liquid phase (Table 2) and its overall impact is on the order of 1% of the total loss rate of 180 Tmol per year.

As a consequence, the tropospheric oxidizing capacity is generally viewed as synonymous with the OH chemistry in the troposphere. Moreover, at least implicitly, the mean global OH abundance has been, and occasionally still is, used as a quantitative measure for the oxidizing capacity. Note that according to this measure, the oxidizing capacity is predicted not to have changed much over the last 100 years. At first glance this choice appears plausible, because OH determines the loss rates of many

Table 1 Rate constants of reactions of various organic trace molecules with OH, O₃, and NO₃. The rate constants are given for 298 K in cm³ molecule^{–1} s^{–1}. The respective lifetimes are calculated on the basis of following concentrations: OH: 1 × 10⁶ cm^{–3}, O₃: 30 ppbv, NO₃: 1 pptv

Organic	Rate constants (at 298 K)			Lifetimes		
	k _{OH}	k _{O₃}	k _{NO₃}	τ _{OH}	τ _{O₃}	τ _{NO₃}
n-Butane	2.4 × 10 ^{–12}	<10 ^{–22}	4.6 × 10 ^{–17}	4.8 days	–	28 years
Ethene	8.6 × 10 ^{–12}	1.6 × 10 ^{–18}	2.1 × 10 ^{–16}	32 h	9.6 days	6 years
Propene	2.9 × 10 ^{–11}	1.0 × 10 ^{–17}	9.5 × 10 ^{–15}	9.6 h	37 h	49 days
Trans-2-butene	6.4 × 10 ^{–11}	1.9 × 10 ^{–16}	3.9 × 10 ^{–13}	4.3 h	1.9 h	28 h
Isoprene	1.0 × 10 ^{–10}	1.3 × 10 ^{–17}	7.0 × 10 ^{–13}	2.8 h	28 h	16 h
α-Pinene	5.3 × 10 ^{–11}	8.7 × 10 ^{–17}	6.2 × 10 ^{–12}	5.2 h	4.1 h	1.8 h
Formaldehyde	8.5 × 10 ^{–12}	–	5.6 × 10 ^{–16}	33 h	–	2.3 years
CH ₃ CHO	1.5 × 10 ^{–11}	–	2.7 × 10 ^{–15}	19 h	–	170 days
(CH ₃) ₂ S	5.0 × 10 ^{–12}	<10 ^{–18}	1.1 × 10 ^{–12}	56 h	>15 days	10 h

Atkinson, R., 1997. Gas-phase tropospheric chemistry of volatile organic compounds: 1. Alkanes and alkenes. J. Phys. Chem. Ref. Data 26, 215–290; Atkinson, R., Baulch, D.L., Cox, R.A., et al., 2006. Evaluated kinetic and photochemical data for atmospheric chemistry: volume II – gas phase reactions of organic species. Atmospheric Chemistry and Physics 6, 3625–4055.

Table 2 Global turnover of the major tropospheric trace gases and attribution to the major oxidants

Trace gas	Global lifetime	Global loss rate (Tmol per year)	Removal (%)			
			OH ^a	O ₃ ^b	NO ₃ ^c	Other
CO	1.5 months	100	85	–	–	15 (Soil uptake)
H ₂	2 years	38	25	–	–	75 (Soil uptake)
CH ₄	8 years	36	90	–	–	10 (Soil uptake; stratos.)
Isoprene	Hours ^d	8	80	7	13	–
SO ₂	Days ^d	5	30	–	–	70 (Hetero. in clouds)
NO _x	0.3–5 days ^d	3	50	40	–	10 (Soil uptake)
Terpenes	Hours ^d	1	20	25	55	–
C ₂ H ₆	2 months	0.7	80	–	–	20 (Cl reaction) ^e
N ₂ O	120 years	0.6	–	–	–	100 (Stratosphere)
(CH ₃) ₂ S	Days ^d	0.5	70 ^f	–	30 ^f	–

^aUsing a mean global OH concentration of $1 \times 10^6 \text{ cm}^{-3}$.^bUsing a mean global O₃ concentration of 30 ppbv.^cUsing a mean global NO₃ concentration of 1 pptv.^dOrder of magnitude or range of local lifetimes. These are too short and variable to make a global lifetime meaningful.^eUpper limit using a mean global Cl concentration of $1 \times 10^3 \text{ cm}^{-3}$.^f3D model; Isaksen, private communication. Due to a strong anticorrelation induced by the oceanic emission of (CH₃)₂S and continental emission of NO₂ removal by NO₃ is less than that calculated from the global means (footnote a), (footnote c) which would assign 85% of the (CH₃)₂S loss to NO₃.

Adapted and expanded from Ehhalt, D.H., 1999a. Chapter 2: Gas phase chemistry of the troposphere. In: Zellner, R. (Guest ed.), Baumgärtel, H., Grünbein, W., Hensel, F. (eds.), Global Aspects of Atmospheric Chemistry, Topics in Physical Chemistry, vol. 6. Steinkopff, Darmstadt, pp. 21–109.

trace gases (see Table 2), such that the knowledge of global OH allows a straightforward estimate of the mean global lifetimes of the individual trace gases. However, taken by itself, the OH concentration says nothing about the rates of the photochemical system and thus nothing about the absolute strength of the oxidizing capacity. A regional example illustrates the point: The OH concentrations in the lower and upper troposphere are about equal (Figure 2). Nevertheless, the loss rate of CH₄ in the upper troposphere is more than a factor of 10 lower than in the lower troposphere, because of the lower temperature and lower CH₄ concentration of the former. In addition, this measure provides little immediate insight into the parameters controlling the oxidizing capacity of the troposphere.

Therefore, more recently, another more appropriate measure for the oxidizing capacity has been introduced. This consists of the total annual oxidative loss of trace gases in the troposphere, L_t (eqn [7], with L_i defined by eqn [6]).

$$L_t = \sum_i L_i \quad [7]$$

Restricting, as above, the losses considered to those facilitated by OH, the integrand for $\sum_i L_i$ becomes identical to D_{OH} (eqn [4]), the local loss rate of OH, with one exception: the term $k_{15} [\text{HO}_2][\text{OH}]$ does not appear among the L_i . This term, however, contributes only about 1% to the sum defining D_{OH} , so that this difference can be safely neglected. Thus we obtain eqn [8], where L_{OH} is the annual mean global loss rate of OH.

$$L_t \equiv \frac{1}{a} \int_{v,t} D_{OH} dv dt = L_{OH} \quad [8]$$

Using eqn [5b], we can rewrite eqn [8] as eqn [9].

$$\begin{aligned} L_t \equiv L_{OH} &= \frac{1}{a} \int_{v,t} P_{OH}(1 + \rho) dv dt = (1 + \bar{\rho}) \frac{1}{a} \int P_{OH} dv dt \\ &= (1 + \bar{\rho}) P_{OH,g} \end{aligned} \quad [9]$$

The total oxidative loss rate of trace gases is given by the total loss rate of OH, which in turn is identical to the annual global primary production rate of OH, $P_{OH,g}$, multiplied by an appropriately averaged feedback factor $(1 + \bar{\rho})$. In other words, this measure for the oxidizing capacity is identical to the total global supply of OH. This definition has several advantages:

1. It quantifies the total trace gas loss initiated by OH in the troposphere.
2. It indicates directly the essential dependences of the oxidizing capacity on its controlling parameters. Since P_{OH} is essentially given by the photolysis of O₃, i.e., $P_{OH} \sim J_1[\text{H}_2\text{O}][\text{O}_3]$ (see eqn [2]), it is clear that $P_{OH,g}$ is dominated by the tropospheric fields of O₃, H₂O, and solar UV radiation. All of these are influenced to some extent by human activities – for example, O₃ by tropospheric chemistry acting on anthropogenic emissions, H₂O by a warming climate, and solar UV by the loss of stratospheric O₃. Clearly ρ is also influenced via tropospheric chemistry, but see point (4) below.
3. This measure also suggests that the oxidizing capacity of the troposphere depends via O₃ and ρ on the emissions of NO_x and volatile organic compounds and thus may respond with some resilience to the load placed on it by anthropogenic emissions, for example, by increasing the burden of tropospheric O₃.
4. It directly indicates that the oxidizing capacity will always remain finite, since both, $P_{OH,g}$ and $\bar{\rho}$ remain finite for any foreseeable emission scenario. For the present troposphere, $P_{OH,g}$ and $\bar{\rho}$ can be estimated from 3D model calculations to about 108 Tmol per year and 0.74, respectively; the corresponding values for the preindustrial troposphere were 71.5 Tmol per year and 0.62. For very high loads of certain pollutants, $\bar{\rho}$ can become negligibly small. Thus, if the global emission flux of NO_x was to exceed $P_{OH,g}$, the global OH abundance would be successively titrated away by reaction [XVII], while the NO_x burden would build up until checked by other losses. It has been pointed out that such

events of overwhelmed oxidation capacity can be observed regionally over areas with high NO_x emissions. There, during wintertime when the flux of solar UV is low, the levels of oxidants such as O_3 and H_2O_2 are greatly reduced in favor of a buildup of the primary pollutants, NO_x and hydrocarbons, quite in contrast to summertime conditions. In such a case, the oxidizing capacity is limited to $P_{\text{OH},g}$. Generally, however, emissions of NO_x and oxidizable molecules have increased together causing an increase in $\bar{p}$.

The new measure, L_v , predicts a substantial increase of the tropospheric oxidizing capacity, about 60%, over the past 100 years, mainly because of the increase of global O_3 by about 60%, but also due to an increasing $\bar{p}$ as a consequence of the global increase in NO_x . The same conclusion can be reached directly from eqn [8], since the predicted global abundance of OH decreased by only 10%, whereas the predicted concentrations of the major molecules CH_4 and CO increased by a factor of 2–3, greatly enhancing the total global loss rate of trace gases.

The model-derived value of L_t for the oxidizing capacity of the present troposphere is 188 Tmol per year. This value includes the oxidation of secondary products such as HCHO. It is uncertain, because the actual emissions of the various tropospheric trace gases are uncertain and some of the minor trace gases may not be treated in the models at all. Moreover, the degradation pathways of some hydrocarbons are still not fully understood and may include OH recycling processes, not considered so far. Such seems to be the case for isoprene. Thus, the above value for L_t may rather represent a lower limit for the current oxidizing capacity of the troposphere.

See also: Ozone Depletion and Related Topics: Photochemistry of Ozone. **Tropospheric Chemistry and Composition:** Aliphatic Hydrocarbons; Aromatic Hydrocarbons; Hydroxyl Radical; Peroxyacetyl Nitrate; Sulfur Chemistry, Organic; Volatile Organic Compounds Overview: Anthropogenic.

Further Reading

- Ehhalt, D.H., 1999a. Chapter 2: Gas phase chemistry of the troposphere. In: Zellner, R. (Guest ed.), Baumgärtel, H., Grünbein, W., Hensel, F. (eds.), *Global Aspects of Atmospheric Chemistry*, Topics in Physical Chemistry, vol. 6. Steinkopff, Darmstadt, pp. 21–109.
- Ehhalt, D.H., 1999b. Photooxidation of trace gases in the troposphere. *Physical Chemistry Chemical Physics* 1, 5401–5408.
- Kleinmann, L.I., 1991. Seasonal dependence of boundary layer peroxide concentrations: the low and high NO_x regimes. *Journal of Geophysical Research* 96, 20721–20733.
- Lelieveld, J., Dentener, F.J., Peters, W., Krol, M.C., 2004. On the role of hydroxyl radicals in the self-cleansing capacity of the troposphere. *Atmospheric Chemistry and Physics* 4, 2337–2344. SRef_ID: 1680-7324/acp/2004-4-2337.
- Lelieveld, J., Butler, T.M., Crowley, J.N., et al., 2008. Atmospheric oxidation capacity sustained by a tropical forest. *Nature* 452, 737–740. <http://dx.doi.org/10.1038/nature06870>.
- Montzka, S.A., Krol, M., Dlugokencky, E., et al., 2011. Small interannual variability of global atmospheric hydroxyl. *Science* 331, 67–69. <http://dx.doi.org/10.1126/science.1197640>.
- Prinn, R.G., 2003. The cleansing capacity of the atmosphere. *Annual Review of Environmental Resource* 28, 29–57. <http://dx.doi.org/10.1146/annurev.energy.28.011503.163425>.

Peroxyacetyl Nitrate

HB Singh, NASA Ames Research Center, Mountain View, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

Peroxyacetyl nitrate (PAN; $\text{CH}_3\text{C}(\text{O})\text{OONO}_2$) is an important constituent of photochemical smog. It is also a ubiquitous chemical present throughout the global troposphere. A unique property of PAN is that it is very stable at cold temperatures and easily decomposes to release NO_x at warm temperatures. In remote atmospheres, it can act as a carrier and a reservoir of NO_x , which is necessary for ozone (O_3) formation. Peroxyacetyl nitrates (PANs) are known to be eye irritants (lachrymators), phytotoxins, and bacterial mutagens. The most serious biological effects of PANs are of a phytotoxic nature resulting in injury to plants and vegetation.

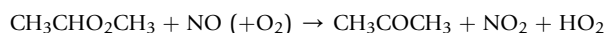
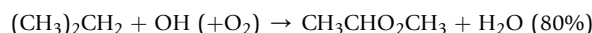
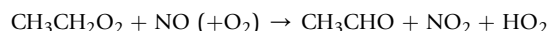
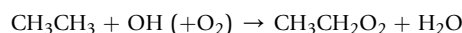
Introduction

Peroxyacetyl nitrate (PAN; $\text{CH}_3\text{C}(\text{O})\text{OONO}_2$), also known as peroxyacetic nitric anhydride, was first identified as a component of Los Angeles smog in the 1950s. Among the large number of peroxyacetyl nitrates (PANs; $\text{RC}(\text{O})\text{OONO}_2$) that have been found to be present in the ambient air, PAN is the most abundant. Other commonly observed but much less abundant chemicals in this series are peroxypropionyl nitrate (PPN; $\text{C}_2\text{H}_5\text{C}(\text{O})\text{OONO}_2$), peroxybenzoyl nitrate (PBzN; $\text{C}_6\text{H}_5\text{C}(\text{O})\text{OONO}_2$), peroxyacryloyl nitrate (APAN; $\text{CH}_2=\text{CHC}(\text{O})\text{OONO}_2$), peroxyisobutyryl nitrate (PiBN; $(\text{CH}_3)_2\text{CHC}(\text{O})\text{OONO}_2$), and peroxymethacryloyl nitrate (MPAN; $\text{CH}_2=\text{C}(\text{CH}_3)\text{C}(\text{O})\text{OONO}_2$). A unique property of PANs is that they are not directly emitted from any known source. They are all products of atmospheric photochemical reactions involving hydrocarbons and nitrogen oxides (NO_x). This makes them excellent indicators of photochemical activity. PANs are chemical pollutants that can cause damage to agricultural crops and are often the reason for eye irritation felt by many people on smoggy days. In recent decades, it has been shown that PAN is not merely a pollutant but is a ubiquitous chemical present throughout the global troposphere. In remote atmospheres, it can act as a carrier and a reservoir of NO_x , which is necessary for ozone (O_3) formation. Here, we discuss mechanisms for the atmospheric formation and destruction of PAN, its measurement methods, concentrations in polluted and remote air, role in atmospheric chemistry, and potential biological effects.

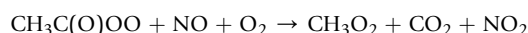
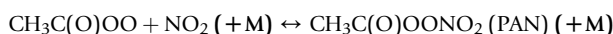
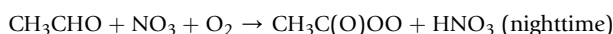
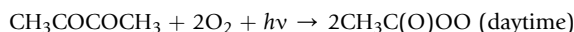
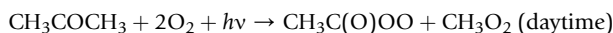
Physical and Chemical Properties

At standard temperature and pressure, PAN exists as a colorless liquid with a melting point of -48.5°C and a boiling point of 104.5°C . Its vapor pressure closely follows the equation $\ln P (\text{hPa}) = -4586/T(\text{K}) + 19.04$. It is nearly insoluble in water (Henry's coefficient is 3.5 M atm^{-1} at 22°C ; $\text{M} = \text{mol l}^{-1}$) but is highly soluble in nonpolar organic solvents and hydrolyzes rapidly in basic solution. It is explosive in pure form and finds no commercial application. In the atmosphere, it is formed from the reaction of peroxyacetyl (PA; $\text{CH}_3\text{C}(\text{O})\text{OO}$) radicals with NO_2 . Although PA radicals can be formed in

many ways, their most common precursors are carbonyls (such as acetaldehyde, acetone, and biacetyl). These carbonyls are often secondary products of atmospheric reactions involving hydrocarbons and oxidants such as O_3 and hydroxyl (OH) radicals. They can also be directly emitted from natural and artificial sources. Hydrocarbons in general are oxidized by reaction with OH radicals and the resulting peroxy (RO_2) radicals react with NO to produce a variety of carbonyls as a product. Below is an example of the formation of acetaldehyde (CH_3CHO) and acetone (CH_3COCH_3) from the oxidation of ethane and propane, respectively.

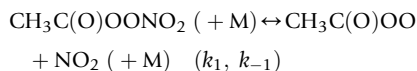


It should be noted that each of the above reactions involves multiple steps. The carbonyls thus formed are highly reactive and break down to produce PA radicals, which can react with ambient NO_2 to form PAN. While PAN formation primarily requires sunlight, its synthesis is also possible at night. Not all PA radicals produce PAN, however, some also react with NO and HO_2 in ways that ultimately produce formaldehyde and peracetic acid, respectively. Molecules such as acetone are uniquely effective in forming PA radicals (and PAN) directly in the upper troposphere.



Atmospheric abundance of PAN is strongly affected by its loss processes. The most important of these is thermal dissociation (TD). This loss rate is an extremely strong function of temperature and hence varies greatly in the atmosphere. The

various rate constants that apply to the equilibrium of PAN with NO_2 are listed below. At warm temperatures, as may be encountered in the boundary layer, PAN dissociates rapidly (hours) and the equilibrium is shifted toward NO_2 . At colder temperatures, as may be encountered in the upper troposphere, PAN is very stable and this reaction is shifted greatly in favor of PAN.



$$k_1 = k_0[\text{M}]\{1 + k_0[\text{M}]/k_\infty\}^{-1}0.3\{1 + (\log k_0[\text{M}]/k_\infty)^2\}^{-1}$$

$$k_0 = 4.9 \times 10^{-3} \exp(-12\,000/T) \text{ cm}^3 \text{ molecule}^{-1} \text{ s}^{-1};$$

$$k_\infty = 4.0 \times 10^{16} \exp(-13\,600/T) \text{ s}^{-1}; k_1/k_{-1} = 1.11 \times 10^{28} \exp(-14\,000/T) \text{ molecule cm}^{-3}.$$

PAN absorbs UV radiation in the troposphere and its absorption cross sections have been measured from 195 to 345 nm wavelengths at several representative temperatures. The absorption is considerably slower at colder temperatures. Breakdown via photolysis removes PAN with an e-fold ($1/e$) time (lifetime) of approximately 3 months. PAN also reacts with OH radicals, but this rate has been found to be extremely slow ($k \leq 3 \times 10^{-14} \text{ cm}^3 \text{ molecule}^{-1} \text{ s}^{-1}$) and provides an ineffective removal process. Figure 1 shows the loss rates of PAN in the troposphere calculated for these processes for July 1 and 30° N conditions and a standard atmosphere. It is clear that thermal loss completely dominates in the lower troposphere. Above 7 km, PAN loss is slow and is dominated by photolysis. PAN has extremely low solubility in water ($H = 3.5 \text{ M atm}^{-1}$ at 22° C) and removal by washout or rainout processes is unimportant. Near the surface of the earth, PAN is also lost by dry deposition. Typical deposition velocities over soil/grass and oceans/water surfaces of 0.25 and 0.01 cm s^{-1} , respectively, have been measured. PPN removal rates from all these processes are similar to those of PAN but have been less accurately measured.

As is evident from Figure 1, PAN in the upper troposphere has a lifetime of many months and thus can be transported over very long distances. Cold air masses containing PAN can

release NO_2 under warmer conditions. By removing NO_2 during its formation phases, transporting it in colder regions, and then rereleasing it in warmer environments, PAN provides an important mechanism for the global transport of NO_x thereby influencing processes of tropospheric ozone formation.

Detection and Measurement Methods

PANs are thermally unstable and are otherwise easily destroyed in contact with surfaces. Therefore, PAN measurements must be performed with minimum delay between sampling and analysis. PAN in the atmosphere has been measured by a variety of techniques. The first laboratory and atmospheric measurements of PAN were made using Fourier transform infrared spectroscopy. While extremely useful for confirmation of PAN, these techniques are best used under controlled laboratory conditions and when PAN concentrations are high. The most commonly used method for the routine measurement of PAN and its homologues has been with gas chromatography using electron capture detector (GC-ECD). Being a good electron absorber, PAN is sensitively detected by the ECD. A number of GC columns can be used to separate PAN from a complex atmospheric mixture of chemicals. Typically DB-210, SPB-1701, and Rtx-200-fused silica capillary columns (0.25–0.5 mm id, 5–30 m length, and $0.25 \mu\text{m}$ film thickness) or packed columns (3 mm id, 40 cm length) with 10% Carbowax 1000 on Supelcoport have been used to separate PAN from other atmospheric constituents. A direct 1–5 ml injection of air is generally adequate to measure PAN in urban and rural environments where PAN is nearly always present at a mixing ratio of $>0.1 \text{ ppb}$ (10^{-9} v/v ; nmol/mol ; ppbv).

In remote atmospheres, PAN concentrations can be extremely low and preconcentration of air samples is often necessary. A 100 ml sample of air freeze trapped at -140° C can provide a measurement sensitivity of 1 ppt (10^{-12} v/v ; pmol/mol ; pptv) in a GC-ECD system. In addition to PAN, a number of other important constituents can also be simultaneously measured. Because of their high sensitivity and easy availability, GC-ECD systems have been extensively used for the simultaneous detection of PANs, alkyl nitrates, and halogen tracers (C_2Cl_4) in airborne- and ground-based studies. Other techniques involving luminol chemiluminescent detection, pulsed discharge detection, and chemical ionization mass spectrometry (CIMS) have also been used on a more limited basis. All these techniques provide PAN measurements in a discreet mode and are relatively slow in response.

Recently, techniques have been developed for near continuous measurements of PAN and its homologues. In one widely accepted technique, PANs are thermally dissociated into acylperoxy radicals, which are made to react with I^- to produce a carboxylate ion that is unique for each parent species and is easily detected by mass spectrometry. Instruments based on TD-CIMS can measure a series of PANs with extremely fast response (1 s) and high sensitivity (10 ppt). Figure 2 shows a spectrogram of PANs from a sample of ambient air over Colorado. Fast response PAN detection is useful for airborne studies and for measurements of its deposition and flux over surfaces. Progress continues to be made to derive a global

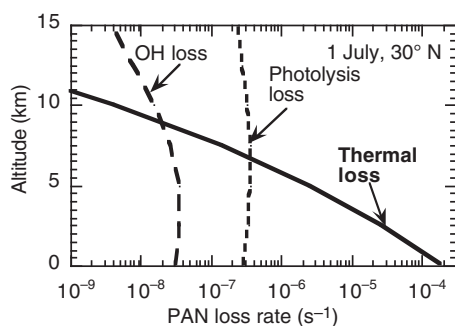


Figure 1 Atmospheric loss rate of PAN due to thermal dissociation, photolysis, and reaction with OH calculated for 1 July and 30° N . Adapted from Talukdar, R.K., Burkholder, J.B., Schmoltner, A.M., Roberts, J.M., Wilson, R.R., Ravishankara, A.R., 1995. Investigation of the loss processes for peroxyacetyl nitrate in the atmosphere: UV photolysis and reaction with OH. *Journal of Geophysical Research* 100, 14163–14173.

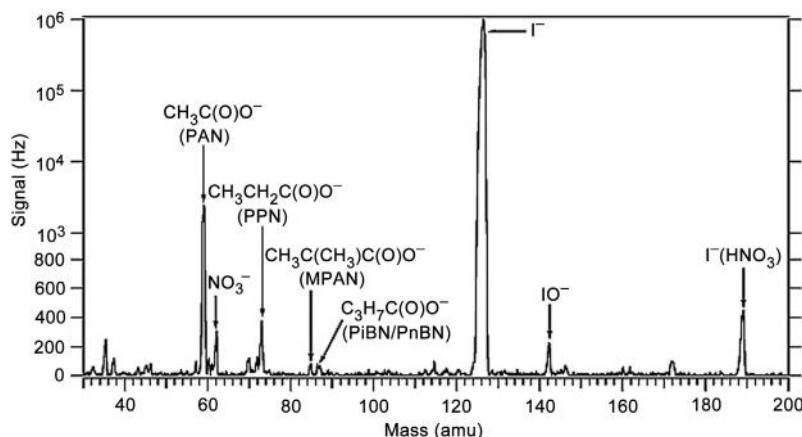


Figure 2 Mass spectrum of ambient air in Boulder, Colorado, on 16 October 2002 using I^- . Note that the vertical scale is linear up to 1000 Hz and logarithmic above 1000 Hz. The PAN signal corresponds to 400 ppt. From Slusher, D.L., Huey, L.G., Tanner, D.J., Flocke, F.M., Roberts, J.M., 2004. A thermal dissociation–chemical ionization mass spectrometry (TD-CIMS) technique for the simultaneous measurement of peroxyacetyl nitrates and dinitrogen pentoxide. *Journal of Geophysical Research* 109, D19315. <http://dx.doi.org/10.1029/2004JD004670>.

climatology of PAN from satellite observations in the upper troposphere and lower stratosphere.

A major difficulty in the quantitative determination of PAN has been the preparation of accurate calibration standards. This largely arises from the fact that PAN is highly unstable and must be prepared in pure form utilizing secondary synthesis methods. In early years, PAN (and PPN) was synthesized by photolysis of alkyl nitrite in oxygen.



PANs produced by this method were quite impure and extensive preparative chromatography was required for purification. PAN was then analyzed by infrared spectroscopy using known IR band strengths typically at 793.9 and 1165 cm^{-1} . These methods were tedious and standards could be prepared only at very high concentrations. In recent years, better calibration methods have been devised. A widely used method utilizes the fact that highly pure PAN can be synthesized in a nonvolatile liquid matrix, such as *n*-tridecane, by nitration of peracetic acid. It is possible to keep a PAN/*n*-tridecane liquid mixture stable for a period of many weeks at dry ice ($-78^\circ C$) temperatures. For routine calibrations, a capillary diffusion tube (2 mm id) is filled with this PAN/*n*-tridecane solution and held at ice-water temperature ($0^\circ C$). Air flows over this tube at a known rate. A nylon filter is inserted in line to remove any traces of nitric acid. The PAN in the exit stream, typically 1–50 ppb, is measured by using a standard chemiluminescent NO_y detector. Trace quantities of CH_3ONO_2 and NO_2 are often present in the exiting air stream, but tests have shown that these impurities are quite small ($<5\%$). These standard mixtures can be further diluted with a dynamic dilution system to generate ppt level mixing ratios in air.

Atmospheric Concentrations

PAN and PPN have been extensively measured in urban and rural locations around the globe. Although PAN concentrations in excess of 200 ppb have been reported from Southern California from the 1960s, these must be considered extremely rare

occurrences. Since 1970, reported concentrations have never exceeded 60 ppb even under extreme polluted and stagnant conditions. **Figure 3** shows the daytime monthly mean and monthly maximum concentrations measured at Riverside in Southern California. A clear seasonal cycle with highest mixing ratios in summer and lowest in winter is seen. Monthly mean mixing ratios never exceeded 10 ppb.

The photochemical nature of PAN formation is also evident in its diurnal cycle with highest concentrations in the late afternoon (**Figure 4**). In rural locations, PAN mixing ratios rarely exceed 5 ppb and average values are in the vicinity of 1–3 ppb. There is no reliable information on the long-term trends in urban PAN, but there are indications that PAN levels have declined over the last four decades. In large part, this is attributable to the dramatic reductions that have been achieved in the emissions of hydrocarbons and NO_x from automobiles.

In remote atmospheres, PAN is nearly always present at sub-ppb levels. Its mixing ratios are generally higher in the middle and upper troposphere compared to the surface where it has a short lifetime. In many regions of the atmosphere, PAN is an important component of total reactive nitrogen often referred to as NO_y . **Figure 5** shows the partitioning of reactive nitrogen throughout the troposphere as measured over North America. PAN concentrations are lowest in the tropics and increase

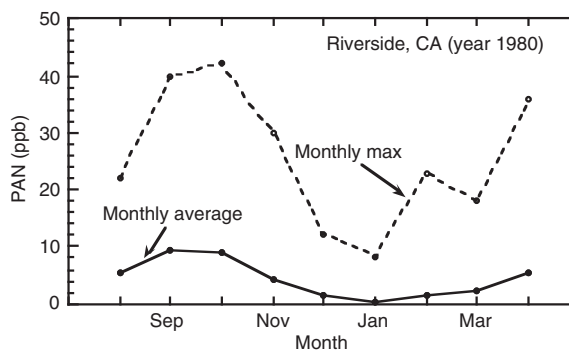


Figure 3 Seasonal cycle of daytime PAN mixing ratios at Riverside, California.

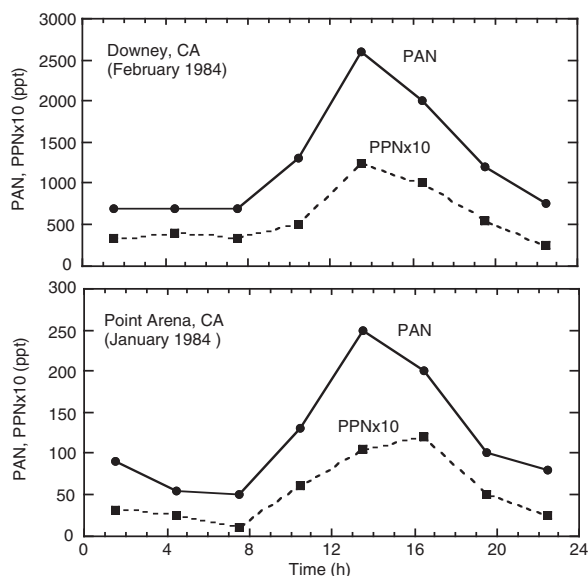


Figure 4 Diurnal variation of PAN and PPN at an urban (Downey) and a coastal/rural (Point Arena) site in California.

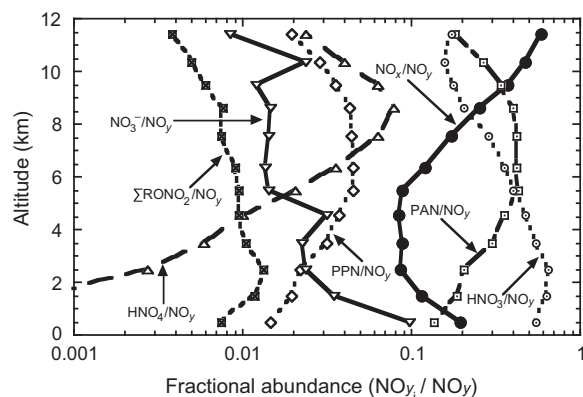


Figure 5 Partitioning of reactive nitrogen in the troposphere over midlatitude North America. NO_y stands for sum of all reactive nitrogen (ΣN) available in the atmosphere. Adapted from Singh, H.B., Salas, L., Herlth, D., Kolyer, R., Czech, E., Avery, M., Crawford, J.H., Pierce, R.B., Sachse, G.W., Blake, D.R., et al., 2007. Reactive nitrogen distribution and partitioning in the North American troposphere and lowermost stratosphere. *Journal of Geophysical Research* 112, D12S04. <http://dx.doi.org/10.1029/2006JD007664>.

rapidly toward higher latitudes. This is due to the greater abundance of precursors as well as the greater stability of PAN at the colder high latitudes. Although PAN concentrations can vary widely, it has been found to be globally ubiquitous.

Many PAN homologues have been detected in urban and rural environments. These are generally present only in the vicinity of the sources of their precursors and do not appear to play any significant role in global atmospheric chemistry. In urban atmospheres, PPN concentrations are typically 5–20% of PAN (Figure 4). However, as air masses are transported over long distances this ratio continues to decline. Measurements of APAN, PBzN, and MPAN are extremely sparse and often below detection limits. Limited data suggest that APAN and PiBN mixing ratios are 3–5% of PAN in urban areas. Upper limit PBzN measurements

suggest that its urban mixing ratios are in the sub-ppb range and almost never exceed 2 ppb. Similarly, MPAN, a product of biogenic hydrocarbon oxidation, has been measured at relatively low concentrations (10–350 ppt) in forested rural environments. Even when present at low concentrations, these PAN homologues can be used as unique photochemical tracers of a variety of anthropogenic and biogenic sources.

Biological/Toxic Effects

PANs are known to be eye irritants (lachrymators), phytotoxins, and bacterial mutagens. Potential phytotoxic episodes in which visible injury to susceptible plants occurs (e.g., pinto beans) are likely only when PAN concentration in excess of 15 ppb is sustained for several hours. Both the eye irritation potential and phytotoxicity of higher PAN homologues (e.g., PPN, PBzN) are substantially greater than that of PAN but their atmospheric concentrations are also much lower. A number of studies have been done on mice to assess the toxic affects of PAN. These studies used PAN concentrations that were often 10^2 – 10^3 times larger than what is typically encountered in polluted air. Based on existing data, it appears that the concentrations of PAN observed in urban atmospheres around the globe are too low to cause injury to human health. Long-term studies on the health effects of exposure to low concentrations of PAN and its homologues have not been performed. The most serious biological effects of PANs are of a phytotoxic nature resulting in injury to plants and vegetation.

See also: Chemistry of the Atmosphere: Chemical Kinetics. Tropospheric Chemistry and Composition: Aliphatic Hydrocarbons; Hydroxyl Radical; Volatile Organic Compounds Overview: Anthropogenic.

Further Reading

- Altshuler, A.P., 1993. PANs in the atmosphere. *Journal of the Air and Waste Management Association* 43, 1221–1230.
- Gaffney, J.S., Marley, N.A., Prestbo, E.W., 1989. Peroxyacetyl nitrates (PANs): their physical and chemical properties. In: Hutzinger, O. (Ed.), *The Handbook of Environmental Chemistry*, vol. 4. Part B Springer-Verlag, New York, pp. 1–38.
- Kleindienst, T.E., 1994. Recent developments in the chemistry and biology of peroxyacetyl nitrate. *Research on Chemical Intermediates* 20, 335–384.
- Moore, D.P., Remedios, J.J., 2010. Seasonality of peroxyacetyl nitrate (PAN) in the upper troposphere and lower stratosphere using the MIPAS-E instrument. *Atmospheric Chemistry and Physics* 10, 6117–6128.
- Moxim, W.J., Levy II, H., Kasibhatla, P.S., 1996. Simulated global tropospheric PAN: its transport and impact on NO_x . *Journal of Geophysical Research* 101, 12621–12638.
- Singh, H.B., 1987. Reactive nitrogen in the troposphere: chemistry and transport of NO_x and PAN. *Environmental Science and Technology* 21, 320–327.
- Singh, H.B., Salas, L., Herlth, D., Kolyer, R., Czech, E., Avery, M., Crawford, J.H., Pierce, R.B., Sachse, G.W., Blake, D.R., et al., 2007. Reactive nitrogen distribution and partitioning in the North American troposphere and lowermost stratosphere. *Journal of Geophysical Research* 112, D12S04. <http://dx.doi.org/10.1029/2006JD007664>.
- Slusher, D.L., Huey, L.G., Tanner, D.J., Flocke, F.M., Roberts, J.M., 2004. A thermal dissociation–chemical ionization mass spectrometry (TD-CIMS) technique for the simultaneous measurement of peroxyacetyl nitrates and dinitrogen pentoxide. *Journal of Geophysical Research* 109, D19315. <http://dx.doi.org/10.1029/2004JD004670>.
- Talukdar, R.K., Burkholder, J.B., Schmoltner, A.M., Roberts, J.M., Wilson, R.R., Ravishankara, A.R., 1995. Investigation of the loss processes for peroxyacetyl nitrate in the atmosphere: UV photolysis and reaction with OH. *Journal of Geophysical Research* 100, 14163–14173.

Synopsis

This article reviews the sources and sinks of the most important organic sulfur that are released to the atmosphere mainly in the reduced form. The reduced sulfur compounds are analogous to organic oxygen compounds and consist of thiols (RSH), sulfides (RSR), and disulfides (RSSR) whereby dimethyl sulfide (DMS, CH_3SCH_3) is by far the most abundant and atmospherically important of all the reduced organic sulfur compounds. The atmospheric chemistry of DMS and its relevance is discussed at length.

Introduction

Sulfur plays an important role in both the tropospheric and stratospheric budgets of atmospheric gases and investigations of the atmospheric sulfur cycle have been a subject of intense scientific interest for many years. The reasons for the interest is the need to parametrize the contributions of atmospheric sulfur, which originate from both natural processes and anthropogenic activities, to environmental problems. The most important of these problems is an assessment of the magnitude of the involvement of the atmospheric chemistry of the organic sulfur compounds dimethyl sulfide (CH_3SCH_3 ; DMS) and dimethyl sulfoxide (CH_3SOCH_3 ; DMSO) in climate regulation or modification.

In industrialized regions such as the United States and most of the Europe, anthropogenic sulfur emissions (mainly comprising SO_2) exceed natural emissions by about one order of magnitude. On a global scale, biogenic emissions become important, with contributions to the sulfur budget of 15–20% and 50–60% in the Northern and Southern Hemispheres, respectively. Of the biogenic contribution, one compound, namely DMS, constitutes approximately 50% of the emissions. This article is restricted to the sources, sinks, and relevant gas-phase atmospheric interactions of the main organic sulfur compounds in the troposphere.

Organic Sulfur Compounds in the Atmosphere

The gas-phase chemistry of atmospheric sulfur involves nearly a score of compounds, representing sulfur valence states from -2 to $+6$. The important organic sulfur compounds that are released to the atmosphere are mainly in the reduced form. The common organic sulfur compounds are divalent and analogous to organic oxygen compounds such as alcohols and ethers: RSH, thiols; RSR, sulfides (thioethers); and RSSR, disulfides.

The names and structures of the organic sulfur compounds known to be involved in atmospheric sulfur chemistry are listed in [Figure 1](#). With the exception of hydroxymethanesulfonic acid and the highly reactive methanesulfenic and methanesulfinic acids, all of these compounds have been detected in the atmosphere. The more oxidized sulfur compounds are less volatile than the reduced

forms and as a result are partitioned to varying degrees into condensed phases. Other important inorganic sulfur compounds involved in the atmospheric sulfur cycle include hydrogen sulfide (H_2S), carbonyl sulfide (OCS), carbon disulfide (CS_2), sulfur dioxide (SO_2), and sulfuric acid (H_2SO_4). Although these compounds are not treated specifically in this article, the reader is strongly encouraged to consult the literature to obtain an insight into their role in atmospheric sulfur chemistry. The roles of sulfur dioxide and sulfuric acid in relation to organic sulfur atmospheric chemistry will become evident during the course of the article.

Over 20 thiols and nearly 40 sulfides/disulfides are known to be emitted into the atmosphere. The sources for most of these compounds are natural, but wood pulping, papermaking processes, and other industrial processes are also sources. Thioethers are widely used in various industries, for example, aryl, alkenyl, and alkyl thioethers are important synthetic reagents and intermediates in organic synthesis. They have been employed in the synthesis of novel biologically active compounds, polymer materials and used as extracting reagents. Because of the frequent unpleasant smells associated with them, even at very low levels, thioethers are mixed into liquefied petroleum gas and fuel cells as leakage sensor odorants. Alkyl thioethers, which can be found in garlic, are synthesized and used as forage additives. These uses may result in their potential release into the atmosphere. Alkyl thiols, dialkyl sulfides, and polysulfides are present in animals, animal manure, kerogens and are associated with sewage treatment plants and the wood pulping industry. Consequently, it is therefore not unreasonable to expect that some of the more volatile alkyl sulfides will be released to some extent to the atmosphere. Alkyl sulfides like diethyl sulfide ($\text{CH}_3\text{CH}_2\text{SCH}_2\text{CH}_3$, DES) and ethyl methyl sulfide ($\text{CH}_3\text{CH}_2\text{SCH}_3$, EMS) have both natural and industrial sources and have been detected over seaweeds fields, and in soil atmospheres.

Among all the organic sulfur compounds emitted into the atmosphere only methyl mercaptan, DMS, and dimethyl disulfide (DMDS) are of any real significance for atmospheric sulfur chemistry. Of these three organic sulfur compounds, it is now firmly established that DMS is by far the most important and abundant. The ranges in the literature estimations of the emission source strengths of the most

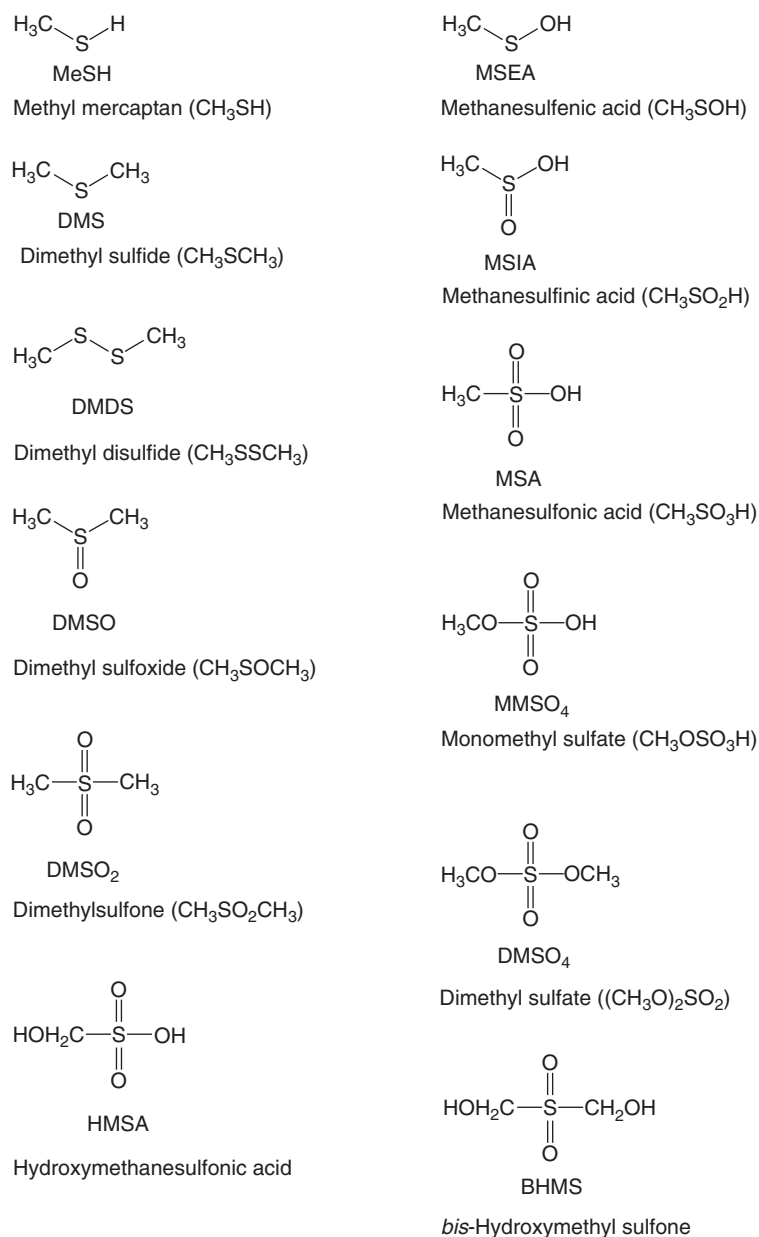


Figure 1 Structural and empirical formulas and names of the major important sulfur compounds in the atmospheric sulfur cycle. Commonly used abbreviations are also indicated.

important organic and inorganic sulfur compounds involved in the atmospheric sulfur cycle are listed in Table 1. The current 'best estimate' emission source strengths are also given.

DMS is a product of biological processes involving marine phytoplankton thus high concentrations of DMS are found in areas where microorganisms are abundant, however, the amounts released vary with species and the stage of their life cycle. Seawater is found to be highly supersaturated with respect to atmospheric levels of DMS, which indicates a net flux of this gas from the ocean to the atmosphere of the order of 15–45 Tg (S) per year. The fluxes to the atmosphere are dependent not only on concentration but also on wind speed,

which makes determination of fluxes difficult and accounts for the large variation in reported values. With the increasing number and sophistication of field studies somewhat lower DMS fluxes in the range 10–20 Tg (S) per year are now more favored. Detailed seasonal maps of DMS in seawater in the world's oceans are available, and highlight the biological origin of the compound. Some DMS in seawater is probably oxidized to DMSO, however, since DMSO is much more soluble than DMS it is unlikely to be released from seawater in any significant amounts.

DMS is estimated to account for approximately 50–60% of the total natural sulfur gas released to the atmosphere. Numerous measurements have been made and continue to

Table 1 Ranges of estimated global sulfur emissions as reported in the recent literature

Substance	Total emission rate (Tg (S) per year)	
	Literature ranges	Current best estimate
Methyl mercaptan, CH ₃ SH	<i>a</i>	<i>a</i>
DMS, CH ₃ SCH ₃	15–45	24.45 ± 5.30
DMDS, CH ₃ SSCH ₃	<i>b</i>	<i>b</i>
Hydrogen sulfide, H ₂ S	5–60	7.72 ± 1.25
Carbonyl sulfide, OCS	0.8–1.7	1.31 ± 0.25
Carbon disulfide, CS ₂	1.1–1.3	0.66 ± 0.19
Sulfur dioxide, SO ₂	100–120 ^c	100–120

DMS, Dimethyl sulfide; DMDS, Dimethyl disulfide.

^aNo good estimates. Probably <10% of biogenic sulfur emissions.

^bNo good estimates. Probably a few percent of biogenic sulfur emissions.

^cMainly of anthropogenic origin.

be performed on the atmospheric concentration of DMS. The atmospheric concentration of DMS is highly variable. In the free troposphere, concentrations are very low, probably <3 ppt. In the marine boundary layer (MBL), it varies from <10 ppt in polluted air to 50–100 ppt over oligotrophic waters and can be as high as 1 ppb in coastal and upwelling waters. At sea surface level the global DMS average is of the order of 100 ppt. In clean marine conditions, atmospheric DMS concentrations typically exhibit minima in the afternoon and maxima at night due to daytime oxidation by the OH radical. Under polluted conditions the diurnal cycle can be affected by oxidation of DMS at night by the NO₃ radical. Atmospheric measurements of DMSO are not as frequent as those for DMS because of measurement difficulties. Mixing ratios of DMSO have been measured in the ranges 0.4–5.8 ppt in the southern Indian Ocean, 0.4–57 ppt in coastal Antarctica, and 0.02–10.1 ppt over northeastern Crete.

There are presently no good estimations of the emission strengths of both methyl mercaptan and DMDS. Methyl mercaptan is probably only of importance in the vicinity of industrial sources and reported mixing ratios over the surface ocean and in tropical forests are typically <5 ppt. No reliable measurements exist on the atmospheric concentrations of DMDS. Owing to its high atmospheric reactivity, atmospheric concentrations of DMDS are very low and of importance only in the proximity of the source. Opinions are currently divided concerning the importance of the contribution of DMDS to the atmospheric sulfur cycle and budget.

Chemical Transformation Mechanisms in the Atmosphere

The atmospheric chemistry of DMS and the other organic sulfur compounds is very complex. However, common intermediates are involved in the atmospheric oxidation processes for all three of the most common organic sulfur compounds – CH₃SH, DMS, and DMDS. Several reviews listed in the literature give good detailed accounts of laboratory kinetic and product investigations on the atmospheric reactions of CH₃SH, DMS, and DMDS.

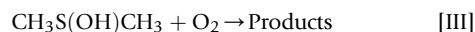
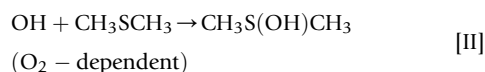
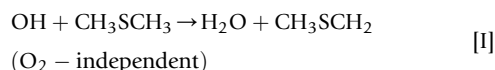
Main Oxidants

For all the compounds, it is generally accepted that reaction with the OH radical is the most important daytime initiator of the oxidation process. During the night and under highly polluted conditions, reactions with NO₃ radicals will be of importance. Concurrent measurements of OH and NO₃ radicals, for example, during a campaign over the northeastern coast of Crete indicated that the NO₃ levels could explain most of the observed diurnal variation of DMS. Similarly, measurements of DMS and NO₃ radicals in a polluted marine environment off the New England Coast indicated that, depending on NO₃ mixing ratios, between 65 and 90% of the DMS oxidation was due to NO₃. The area over which DMS oxidation by NO₃ was at least as strong as by OH was estimated to extend as far as 3000 km from the coastal anthropogenic NO_x sources. Under pristine background marine conditions with very low NO_x, reactions with NO₃ are relatively unimportant.

Models of global sulfur chemistry have been interpreted as indicating that DMS sulfur chemistry is too rapid to be accounted for entirely by known sink reactions with OH and NO₃ radicals. It has been suggested that reactions with halogens or halogen oxides could be a hitherto neglected sink for DMS in the MBL. Evidence is currently emerging that chlorine and bromine chemistry is more important in the troposphere than had been previously thought, although more field experiments are needed for definite validation. In recent field campaigns, midcontinental observations of the chlorine atom precursor nitryl chloride (NOCl) have been reported at a distance of 1400 km from the nearest coastline. It has also been suggested that a significant fraction of tropospheric chlorine atoms may arise directly from anthropogenic pollutants. The BrO radical has now been detected in many field campaigns in both polluted and nonpolluted marine environments and in some environments the oxidation of DMS by BrO has been found to take place at a comparable rate to oxidation by OH. In summary, the current evidence from field campaigns supports that it is highly probable that Cl atoms and BrO radicals are quite important oxidants for DMS under certain conditions and reactions of DMS with these oxidants need to be included in models of DMS tropospheric chemistry. Table 2 lists the typical atmospheric concentrations of potential oxidants for DMS, their rate coefficients with DMS, and remarks on their potential importance based on these parameters.

Initiating Oxidation Steps

The DMS–OH reaction is now known to proceed via both O₂-independent and O₂-dependent pathways (eqns [I]–[III]).



The O₂-independent pathway dominates at room temperature and involves hydrogen abstraction. The O₂-dependent

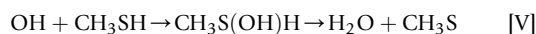
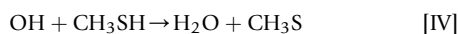
Table 2 Possible processes for removal of DMS in the atmosphere along with their relative contributions to the overall lifetime (τ) of DMS

Removal process	$k_{298\text{ K}}$ (cubic centimeter per molecule per second)	Typical reactant concentration ranges (molecule cm^{-3})	τ_{DMS}^a
$\text{OH} + \text{DMS} \rightarrow \text{products}$	6.3×10^{-12}	$\sim 2 \times 10^6$ (12 h daytime average)	0.92 days important ^b
$\text{HO}_2 + \text{DMS} \rightarrow \text{products}$	$< 10^{-16}$	$\sim 1 \times 10^8$ (average)	3.2 years unimportant
$\text{O}(^3\text{P}) + \text{DMS} \rightarrow \text{CH}_3\text{SO} + \text{CH}_3$	5×10^{-11}	Negligible	– unimportant
$\text{O}_3 + \text{DMS} \rightarrow \text{products}$	1×10^{-18}	7×10^{11} (24 h average)	16.5 days unimportant
$\text{NO}_3 + \text{DMS} \rightarrow \text{products}$	1.1×10^{-12}	$\sim 1 \times 10^8$ (concentration in remote marine areas)	2.3 h important ^b
$\text{Cl} + \text{DMS} \rightarrow \text{products}$	2×10^{-10}	10^3 – 10^5	58–0.6 days possibly regionally important, needs validation
$\text{ClO} + \text{DMS} \rightarrow \text{products}$	9.5×10^{-15}	10^6 – 10^8	3.3–0.33 years unimportant
$\text{Br} + \text{DMS} \rightarrow \text{products}$	$< 1 \times 10^{-13}$	10^7 – 10^8	11.6–1.2 days probably unimportant
$\text{BrO} + \text{DMS} \rightarrow \text{products}$	2.7×10^{-13}	10^9 – 10^{10}	8.7–0.87 h possibly regionally important, needs validation ^c
$\text{IO} + \text{DMS} \rightarrow \text{products}$	1.2×10^{-14}	10^6 – 10^7	2.6–0.26 years unimportant

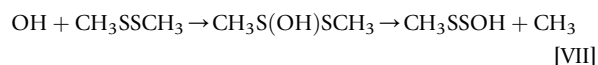
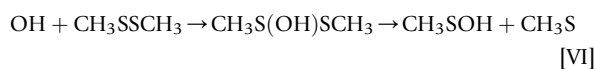
DMS, Dimethyl sulfide.

^aThe lifetime τ_{DMS} is defined as $1/k[\text{reactant}]$ where k is the rate coefficient (cubic centimeter per molecule per second) for the reaction of the reactant with DMS and $[\text{reactant}]$ is the typical atmospheric concentration (molecule cm^{-3}) of the species.^bConcentration is very variable.^cProbably most important in the Arctic, but recent field measurements indicate a possible global importance for BrO; ongoing field/satellite measurements should resolve this issue in the near future.

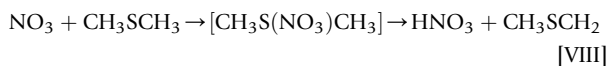
pathway involves formation of an adduct $((\text{CH}_3)_2\text{S}-\text{OH})$ with a binding energy of $\sim 45 \text{ kJ mol}^{-1}$. The fate of the adduct is reaction with O_2 in competition with decomposition back to reactants. The adduct lifetime is strongly temperature dependent and lowering the temperature favors the addition over the abstraction channel. As discussed below, DMSO is now known to be a major product (over 50%) of the reaction of the adduct with O_2 . The mechanism of OH reaction with methyl mercaptan (CH_3SH) ($k_{\text{OH}}(298 \text{ K}) = 3.3 \times 10^{-11} \text{ cm}^3 \text{ per molecule per second}$) is unclear; however, it appears that the reaction leads to CH_3S radical formation. Although a direct hydrogen atom abstraction pathway is operative, the observation of a negative temperature dependence supports formation of adducts. Formation of an OH adduct ($\text{CH}_3\text{S}(\text{OH})\text{H}$) that rapidly eliminates H_2O can therefore not be excluded (eqns [IV] and [V]).



The kinetics of the reaction of OH with DMDS ($k_{\text{OH}}(298 \text{ K}) = 2.3 \times 10^{-10} \text{ cm}^3 \text{ per molecule per second}$) also shows a negative temperature dependence, which again suggests that the mechanism involves complex formation. Formation of CH_3S and CH_3SOH has been observed, however, a measured yield of 30% for CH_3S indicates that other pathways leading to CH_3 and CH_3SSOH are also operative (eqns [VI] and [VII]).



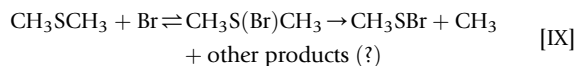
The available evidence suggests that reaction of DMS with NO_3 proceeds initially by NO_3 addition to the sulfur atom followed by hydrogen intramolecular transfer, i.e., an overall hydrogen abstraction reaction (eqn [VIII]).



For the reactions of NO_3 with CH_3SH ($k_{\text{NO}_3}(298 \text{ K}) = 9.2 \times 10^{-13} \text{ cubic centimeter per molecule per second}$) and DMDS ($k_{\text{NO}_3}(298 \text{ K}) = 7 \times 10^{-13} \text{ cm}^3 \text{ per molecule per second}$), a mechanism involving initial addition of NO_3 to the sulfur atom followed by rearrangement or fragmentation of the adduct has been inferred from the scant information that is available. Because of the fast reaction of OH radicals with DMDS, reaction between NO_3 and DMDS is probably of only marginal importance in the atmosphere.

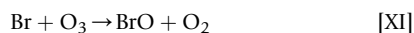
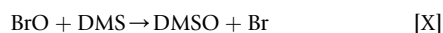
There is increasing evidence that reactive halogen chemistry involving X and XO, where X = Cl, Br, or I, plays an important role in tropospheric chemistry in the MBL. The role of halogen species, especially BrO and Cl, in the oxidation of DMS is still unclear and may be very regionally dependent, as demonstrated in Table 2. For the reaction of Cl with DMS, a direct hydrogen transfer pathway leading to formation of HCl and CH_3SCH_2 radicals is operative along with an adduct-forming channel. The products of the adduct-forming channel have not yet been characterized. The reaction of Br with DMS

proceeds via reversible adduct formation but there also appears to be an adduct decomposition channel leading to formation of CH_3SBr and CH_3 radicals.



The overall reaction rate is too slow (see Table 2) for the process to be of any significance in the atmosphere. The reaction of iodine atoms with DMS is negligibly slow and also not important.

The kinetics of the reactions of ClO , BrO , and IO with DMS have been measured and the rate coefficients take the order $\text{BrO} > \text{ClO} \sim \text{IO}$. All of the reactions are thought to result in oxidation of DMS to DMSO and release of the halogen atom. Based on the available kinetic data, atmospheric model calculations indicate that oxidation by BrO can be a relevant sink of DMS, which could account during daytime for 24–38% loss of DMS. Very recent kinetic data indicate that the reaction of BrO with DMSO at high pressure may be significantly faster than previously determined, which would further enhance the importance of BrO . This remains to be validated. In addition, the reaction of BrO with DMS can promote $\text{O}_3 \rightarrow \text{O}_2$ conversion through the sequence of eqns ([X] and [XI]) where DMS is oxidized and no BrO is lost.



Thus, BrO can be an important sink of DMS in the presence of O_3 , and at the time of writing field campaigns are in progress to quantify the possible contribution to the DMS budget.

Reactions of Other Organic Sulfur Compounds with Oxidants

In comparison to the many studies performed on the reactions of CH_3SH , DMS, and DMDS, there have been very few studies on the atmospheric chemistry of other organic sulfur compounds. There have been several room-temperature kinetic studies of the gas-phase reactions of OH radicals with the dialkyl sulfides, DES and EMS, and the cyclic compounds, tetrahydrothiophene and thiophene. The measured rate coefficients for reactions of OH with both DES and EMS show an O_2 partial pressure dependence and the experimental observations are consistent, a four-step mechanism involving a direct H-atom abstraction channel, and a reversible OH addition channel forming an OHSR_2 adduct, which can react with molecular oxygen has been postulated to explain the complex kinetics of the reaction of DMS with OH radicals in the presence of oxygen.

High yields of SO_2 are observed in the reaction systems indicating that the H-atom abstraction channel is the most important pathway for the reactions of OH radicals with both DES and EMS. From kinetic investigations, it has been estimated under atmospheric conditions approximately 73 and 76% of the reactions of OH with DES and EMS, respectively will proceed via the abstraction channel. The reactions form mainly RSCHCH_3 radicals. These react with molecular oxygen to form alkylthio peroxy radicals ($\text{CH}_3\text{CH}(\text{OO}\cdot)\text{SR}$), which are further converted to alkylthio alkoxy radicals ($\text{CH}_3\text{CH}(\text{O}\cdot)\text{SR}$)

the major fate of which is decomposition to give an aldehyde and an alkyl thiyl radical (RS) as discussed in detail for DMS below. The subsequent reactions of the alkyl thiyl radicals will lead to a large extent to the formation of SO_2 and an aldehyde. The yields of aldehydes and SO_2 measured in the product studies support that reaction channels other than decomposition are probably also operative for the alkylthio alkoxy radicals. Reactions of OH with DES and EMS with O_2 may lead to formation of thioesters: s-ethyl thioacetate ($\text{CH}_3\text{CH}_2\text{SC}(\text{O})\text{CH}_3$), s-methyl thioacetate ($\text{CH}_3\text{SC}(\text{O})\text{CH}_3$), and s-ethyl thioformate ($\text{CH}_3\text{CH}_2\text{SCHO}$). In contrast to the chemistry of the alkoxy radicals formed in the reaction of OH with DMS (see below), the alkoxy radicals $\text{CH}_3\text{CH}_2\text{SCH}(\text{O}\cdot)\text{CH}_3$ and $\text{CH}_3\text{CH}_2\text{SCH}_2(\text{O}\cdot)$ formed in the $\text{OH} + \text{DES}$ and EMS reactions, respectively, could isomerize by a 1,5-H shift through a six-member transition state. The sulfur and carbon balances in gas-phase product studies of the oxidation of dialkyl sulfides are currently only of the order of 50%.

Room-temperature rate coefficients have been reported for the gas-phase reactions of Cl atoms, DES, EMS, di-n-propyl sulfide, and di-n-butyl sulfide. A very slow rate coefficient of $2.77 \times 10^{-19} \text{ cm}^3$ per molecule per second has been reported for the reaction O_3 with DES and it is probable that the reactions of O_3 with other alkyl sulfides are equally slow and of no atmospheric importance. Apart from DMS, there do not appear to have been any kinetic studies of the gas-phase reactions of NO_3 radicals or halogen oxides (OX , $\text{X} = \text{Cl}, \text{Br}, \text{I}$) with alkyl sulfides. Reactions of NO_3 and BrO radicals with the alkyl sulfides are likely to be of the same order of magnitude or faster than the corresponding reactions of the radicals with DMS and thus of potential importance as atmospheric sinks for the compounds.

Observed DMS Oxidation Products and Possible Formation Pathways

Although the rate coefficients for the initiation reactions of organic sulfur compounds are reasonably well established, the reactions responsible for the observed products and their yields are not always clear. Product information has been obtained using absolute methods, but much of the information stems from studies in so-called photochemical reactors. Many of these product studies have employed highly variable NO_x levels, and the chemistry that occurs in these high NO_x systems can be very different from that occurring in the marine atmosphere where the NO_x levels are generally very low. Consequently, there is a large variation in the yields of the products reported in the literature and it is still not possible to make reliable quantitative predictions of the distribution of DMS oxidation products for specific sets of atmospheric conditions.

In laboratory photoreactor studies at room temperature, SO_2 , methanesulfonic acid (MSA), DMSO, dimethyl sulfone (DMSO_2), and methylsulfonyl peroxyxynitrate (MSPN ; $\text{CH}_3\text{SO}_2\text{OONO}_2$) have been observed as products of both the OH and NO_3 radical-initiated oxidation of DMS. In the reaction of NO_3 with DMS, unity yield of HNO_3 is also observed, formed via a hydrogen-atom abstraction reaction from one of the CH_3 entities. Formation of methanesulfinic acid (MSIA) in

significant yield has been reported very recently for OH + DMS. Both laboratory and field observations support that SO₂ is a major oxidation product of DMS oxidation, and recent photoreactor studies have demonstrated that it may be formed via both the addition and abstraction channels in OH + DMS. Under NO_x conditions approaching those of the atmosphere, SO₂ molar yields of 70–80% have been reported for photoreactor experiments on OH + DMS. Similar yields have been deduced from field measurements of DMS and its products. Apart from gas-phase oxidation with OH radicals to form H₂SO₄(g), the main fate of SO₂ is uptake in the liquid aerosol (condensed phase) and oxidation of H₂SO₄(l).

As indicated earlier, results from room-temperature laboratory end-product studies show that DMSO is a major product of the addition pathway in the OH + DMS reaction. The absolute molar yield is still uncertain, but it is of the order of 50–70% and may be higher. In the gas phase, DMSO is known to be oxidized rapidly by OH radicals ($k_{\text{OH}}(298\text{ K}) \sim 10^{10}\text{ cm}^3\text{ per molecule per second}$). The results from OH + DMSO product investigations are contradictory. DMSO₂, SO₂, and MSA have been observed as products in photoreactor studies; however, there are considerable discrepancies in the yields and secondary chemistry is thought to strongly influence the product distribution in such studies. In a recent time-resolved tunable diode laser spectroscopy study on the OH + DMSO reaction, CH₃ radicals were detected in near unity yield at low pressure. This supports a mechanism in which the main reaction pathway is the formation of CH₃ radicals with methanesulfinic acid as coproduct. At this writing, photoreactor experiments using ion chromatography to detect MSIA have been reported that substantiate the tunable diode laser spectroscopy study.

The only product observed in the reaction of NO₃ with DMSO is DMSO₂. DMSO and DMSO₂ have been identified in field measurements in the gas/aerosol phase and also in rain-water samples. The fate of these compounds is poorly understood with regard to their removal by heterogeneous processes or by deposition. There is evidence that in the condensed phase DMSO is oxidized initially to MSIA, which is then further oxidized to MSA.

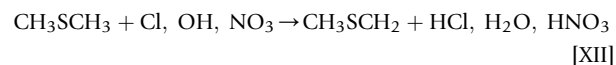
Carbonyl sulfide (OCS) has been detected in chamber product studies of the OH-initiated oxidation of DMS under low NO_x conditions. A formation yield of $0.7 \pm 0.2\%$ S was reported. The mechanism of formation is not known. Although the OCS yield is low, because of the relatively high global DMS source strength (15–45 Tg (S) per year), the result suggests that the oxidation of DMS could represent a substantial source of atmospheric OCS, contributing 0.10–0.28 Tg (OCS) per year.

In summary, although the major products of the OH- and NO₃-initiated oxidation are known, detailed quantitative information on their distribution is not yet available, particularly with regard to the effects of temperature and variable O₃ and NO_x concentrations. More information on the fate of DMSO is important because at high latitudes the O₂-dependent addition branch of the OH + DMS reaction forming DMSO in high yield dominates the overall DMS reactivity. Thus with decreasing temperature DMSO chemistry, both gas and liquid phases (DMSO is highly water soluble), will increasingly control the branching ratio for SO₂, SO₄²⁻, and acid formation (MSA and MSIA). Thus an exact knowledge of

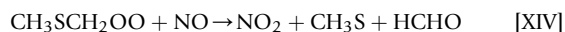
the formation yield of DMSO with temperature is an important parameter in atmospheric organic sulfur chemistry.

Reaction Pathways Leading to Products

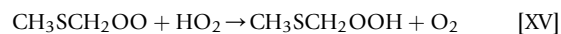
As indicated above, generation of the CH₃SCH₂ radical is an important pathway for the Cl-, OH-, and NO₃-initiated oxidation of DMS (eqn [XII]). In the atmosphere, the CH₃SCH₂ radical will add O₂ to form the peroxy radical CH₃SCH₂OO (eqn [XIII]).



It is known that reaction with NO will produce NO₂, CH₃S, and formaldehyde (eqn [XIV]).



It has been suggested that in the remote MBL, where NO_x levels are low (typically 2–8 ppt), the main daytime removal process for CH₃SCH₂OO could be reaction with HO₂ to form a hydroperoxide (eqn [XV]).



It is argued that if the removal of this hydroperoxide by photolysis and reaction with OH (see Figure 2 for pathways) is similar in magnitude to that of CH₃OOH, then CH₃SCH₂OOH may be a significant reservoir species for sulfur in the remote MBL. As shown in Figure 2, similar pathways may also exist for other oxidized forms of CH₃SCH₂OOH. This present hypothesis remains to be verified by either field observations or laboratory experiments. It is interesting to note that if the surface deposition and/or washout and rainout processes of CH₃SCH₂OOH and its oxidized forms are significant, which is probably the case, then these processes represent a potential short circuit for DMS atmospheric chemistry, preventing formation of the gas-phase oxidized forms.

It is now well established that CH₃S is an important intermediate in the oxidation of DMS and also of CH₃SH. Despite the key role that the CH₃S radical plays in atmospheric sulfur chemistry, the reactions that convert it to S(vi) end products, H₂SO₄ and CH₃SO₃H (MSA), are poorly understood. A weakly bound CH₃S–O₂ adduct is known to exist and it is quite probable that CH₃SO_x–O₂ adducts ($x = 1, 2$) may also be important. At the time of writing, the available laboratory data indicate that reaction of O₃ with CH₃S will probably be a dominant removal pathway for this species in the atmosphere and that NO₂ could also play a role under polluted conditions. Reactions of O₃ and NO₂ with CH₃SO_x ($x = 1$ or 2) may also be important. Table 3 compares the first-order loss rates for reactions of CH₃S and CH₃SO_x ($x = 1$ or 2) with O₂, O₃, and NO₂ for different O₃ and NO₂ levels.

The products of these reactions are not clear, but the available evidence supports the yield of CH₃SO₂ being substantial. It is often argued that thermal decomposition of CH₃SO₂ radical to CH₃ and SO₂ is the most important reaction in determining the formation of SO₂ in DMS oxidation. Important competing steps will be reaction with O₂ or O₃, which will

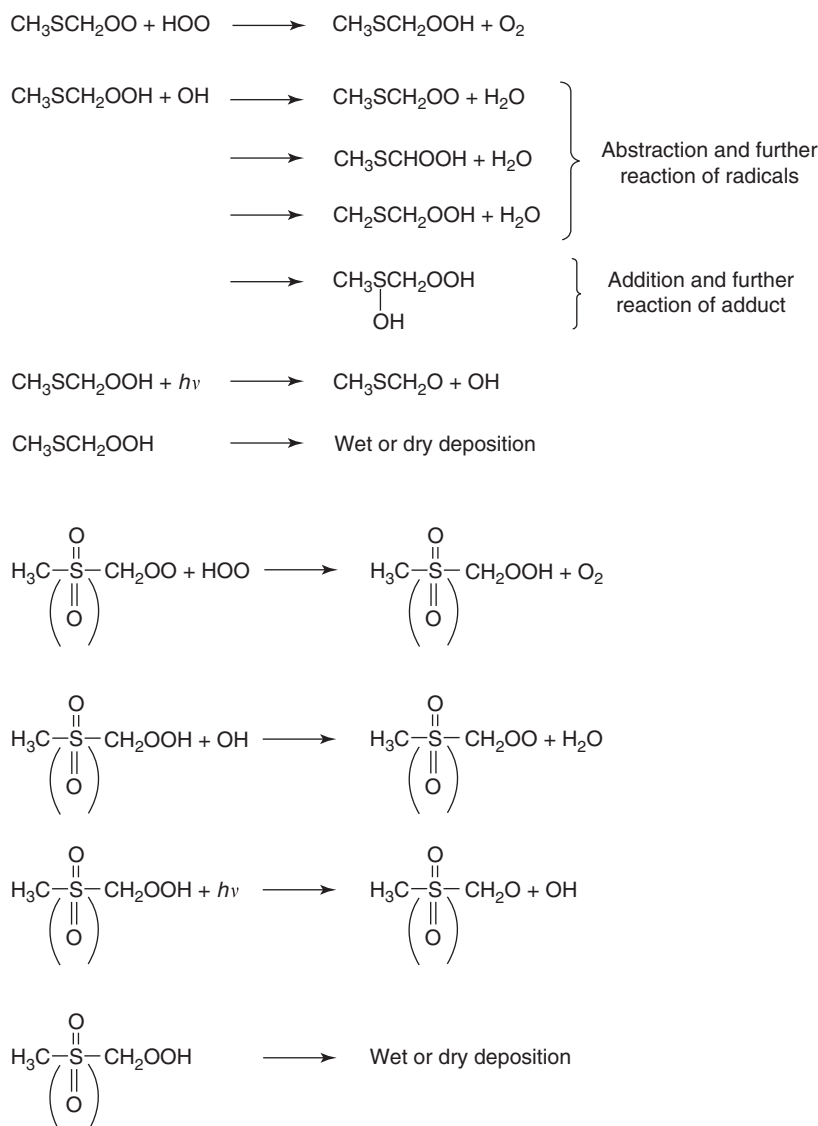


Figure 2 Possible atmospheric reactions of the $\text{CH}_3\text{SCH}_2\text{OO}$ radical and its oxygenated analogs $(\text{CH}_3\text{S}(\text{O})_x\text{CH}_2\text{OO}, x = 1 \text{ or } 2)$. The large parentheses around the $\text{S}=\text{O}$ bonds denote two different initial reactants and subsequent products, i.e., a compound with an $\text{S}=\text{O}$ entity and another with an SO_2 entity.

result in the formation of $\text{CH}_3\text{S}(\text{O}_2)\text{OO}$ and CH_3SO_3 . Since the majority of SO_2 is oxidized in the condensed phase, the thermal decomposition will result in a low yield of $\text{H}_2\text{SO}_4(\text{g})$; however, unity conversion of CH_3S to nonsea-salt sulfate would occur. Thermal decomposition of CH_3SO_3 to CH_3 and SO_3 would lead to a high yield of $\text{H}_2\text{SO}_4(\text{g})$ via further reaction of SO_3 with water. However, at the time of writing it is thought that bimolecular reaction with hydrogen-containing species to form MSA dominates. $\text{CH}_3\text{S}(\text{O}_2)\text{OO}$ can add NO_2 to form a short-lived peroxyxynitrate, but further reaction to generate MSA will also occur.

The unimolecular decomposition rates of CH_3SO_x radicals and their O_2 adducts are expected to be strongly temperature dependent. Competition between the thermal decomposition of these adducts and their bimolecular reactions with atmospheric species is the most probable reason for much of the

variation not only in field observations of DMS oxidation products but also in the variations of products reported in laboratory studies.

Representation of DMS Oxidation in Models

Accurate and detailed models of all aspects of the atmospheric chemistry of DMS are necessary to assess any role that DMS may potentially play in the formation of new cloud condensation nuclei (CCN) formation in oceanic regions and thus climate regulation. As evidenced in this article the gas-phase chemistry of DMS is highly complex and uncertain in many aspects and in addition the nucleation ability of DMS-derived sulfuric acid in the MBL is controversial and much debated in the scientific literature. Many DMS oxidation mechanisms have been

Table 3 Comparison of the first-order loss rates ($k[A]$) for the reaction of CH_3SO_x ($x=1-2$) radicals with O_2 , O_3 , and NO_2 at 298 K under various atmospheric conditions

Reaction $\text{CH}_3\text{S} + \text{A}$	$k_{298\text{ K}} (\text{cm}^3 \text{ per molecule per second})$	$[\text{A}] (\text{molecules cm}^{-3})$	$k[\text{A}] (\text{s}^{-1})$
$\text{CH}_3\text{S} + \text{O}_2 \rightarrow \text{products}$	$<3.0 \times 10^{-18\text{a}}$	5.17×10^{18}	<15.5
$\text{CH}_3\text{S} + \text{O}_3 \rightarrow \text{products}$	5.6×10^{-12}	9.84×10^{11} (40 ppb)	5.5
		2.46×10^{12} (100 ppb)	13.8
$\text{CH}_3\text{S} + \text{NO}_2 \rightarrow \text{products}$	5.6×10^{-11}	2.46×10^{10} (1 ppb)	1.4
		1.23×10^{12} (20 ppb)	28.9
$\text{CH}_3\text{SO} + \text{O}_2 \rightarrow \text{products}$	$(7.7 \times 10^{-18})^{\text{b}}$	5.17×10^{18}	(40)
$\text{CH}_3\text{SO} + \text{O}_3 \rightarrow \text{products}$	6×10^{-13}	9.84×10^{11} (40 ppb)	0.6
		2.46×10^{12} (100 ppb)	1.5
$\text{CH}_3\text{SO} + \text{NO}_2 \rightarrow \text{products}$	8×10^{-12}	2.46×10^{10} (1 ppb)	0.2
		1.23×10^{12} (20 ppb)	9.8
$\text{CH}_3\text{SO}_2 + \text{O}_2 \rightarrow \text{products}$	$(2.6 \times 10^{-18})^{\text{b}}$	5.17×10^{18}	(13)
$\text{CH}_3\text{SO}_2 + \text{O}_3 \rightarrow \text{products}$	3×10^{-13}	9.84×10^{11} (40 ppb)	0.3
		2.46×10^{12} (100 ppb)	0.7
$\text{CH}_3\text{SO}_2 + \text{NO}_2 \rightarrow \text{products}$	4×10^{-12}	2.46×10^{10} (1 ppb)	0.1
		1.23×10^{12} (20 ppb)	4.9

^aThe rate coefficient is for the reaction $\text{CH}_3\text{S} + \text{O}_2$ going to products other than CH_3SOO . The upper limit is not corrected for the $\text{CH}_3\text{S} + \text{O}_2 \leftrightarrow \text{CH}_3\text{SOO}$ equilibrium.

^bThe rate coefficient is estimated, only the reaction $\text{CH}_3\text{S} + \text{O}_2$ going to products other than $\text{CH}_3\text{S} + \text{O}_2 \leftrightarrow \text{CH}_3\text{SOO}$ is considered. The rate is not corrected for the $\text{CH}_3\text{SO}_x + \text{O}_2 \leftrightarrow \text{CH}_3\text{SO}_x\text{OO}$ equilibrium.

formulated from the available experimental evidence for use in various types of atmospheric models used to assess the role of DMS in climate. The schemes range from very detailed descriptions of the DMS oxidation with up to 65 reactions to a very simple scheme with only six reactions suitable for global modeling. The Master Chemical Mechanism (<http://mcm.leeds.ac.uk/MCM/>) is a near-explicit chemical mechanism, which describes the detailed gas-phase chemical processes involved in the tropospheric degradation of a series of primary emitted volatile organic compounds (VOCs). Currently, the degradation of methane and 142 nonmethane VOCs is represented and recently a very simplified DMS oxidation mechanism has been added. There have been several assessments of the various mechanisms with respect to their capability of predicting observed end products of the DMS oxidation at climatologically different locations in the remote MBL. In the most recent intercomparison of seven commonly used DMS oxidation chemical mechanisms for the remote MBL, the conclusion was reached that several uncertainties limiting our understanding of atmospheric oxidation of DMS with implications for climate still exist.

Important Atmospheric Interactions of DMS

Aerosols in the 0.001–10 μm size range play a crucial role in regulating climate since they scatter and absorb radiation in the solar and thermal infrared ranges, which modifies the Earth's radiation budget. A significant fraction of the tropospheric aerosol mass in the submicrometer-size particles is estimated to be derived from homogeneous and in-cloud oxidation of gaseous sulfur compounds, which are both of anthropogenic and biogenic origin. There are three main sources of aerosol in the remote MBL: sea-salt and nonsea-salt sources, and entrainment of free tropospheric aerosol. The SO_2 produced from DMS oxidation is the predominant source of *in situ*

nonsea-salt sulfate. SO_2 is oxidized to H_2SO_4 in the gas phase by reaction with OH and in sea-salt aerosols and cloud droplets. H_2SO_4 contributes either to aerosol growth or to the formation of new sulfate particles. The formation and subsequent growth of small particles can affect climate via direct scattering and absorption of solar radiation (direct effect), and by modifying the frequency, lifetime, and optical properties of clouds (indirect effect).

According to the CLAW hypothesis, the aerosol formation stemming from the degradation of DMS could play an important role in the radiation budget and possibly climate regulation. In the hypothesis, long-term climate is seen as part of a feedback loop in which warmer temperatures would increase phytoplanktonic activity in the oceans, which in turn would increase emissions of DMS. It is suggested that higher DMS concentrations in the atmosphere would increase sulfate containing CCN and thus cloudiness and albedo and lead to lower temperatures. Although the CLAW hypothesis is an interesting and once very popular vision of the biological mediation in global climate, evidence from ice cores has yet to offer convincing proof for this concept of climate control. In addition many studies are now highlighting significant nonlinearity in DMS aerosol–cloud interactions.

Considering the potential importance of DMS in the regulation of the Earth's climate, there have been relatively few laboratory investigations of DMS photooxidation systems aimed at quantifying the contribution of DMS-derived nonsea-salt sulfate to CCN. Model studies of global SO_2 indicate that there may be as yet unaccounted oxidants involved in the DMS oxidation. The Intergovernmental Panel on Climate Change (1995) has classified the coupling between DMS and aerosols as an important component of the planetary climate system that needs to be understood in detail.

The evaluation of the role of DMS in climate regulation has been complicated recently by two important findings. First, the conventional view has generally been that nonsea-salt sulfate is

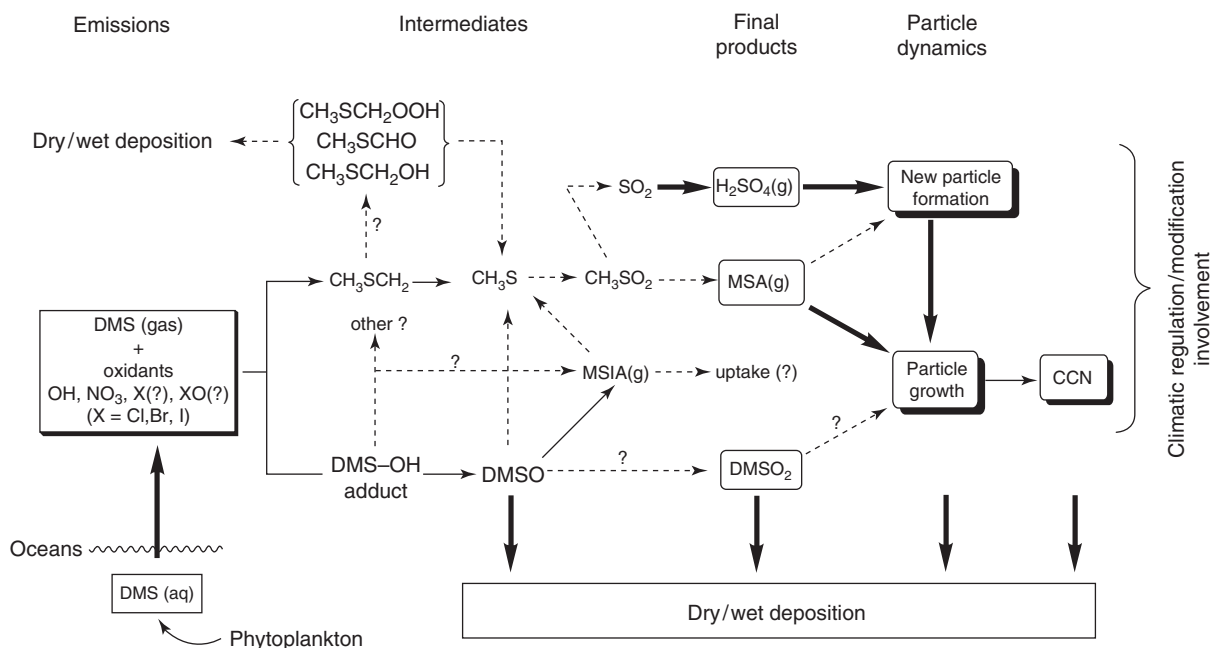


Figure 3 Simplified diagram of the processes involved in the degradation of DMS in the marine atmosphere. The dotted lines represent processes where considerable uncertainty still exists concerning the branching ratios and/or kinetics and the question marks signify that either the existence or the importance of the channel is not known.

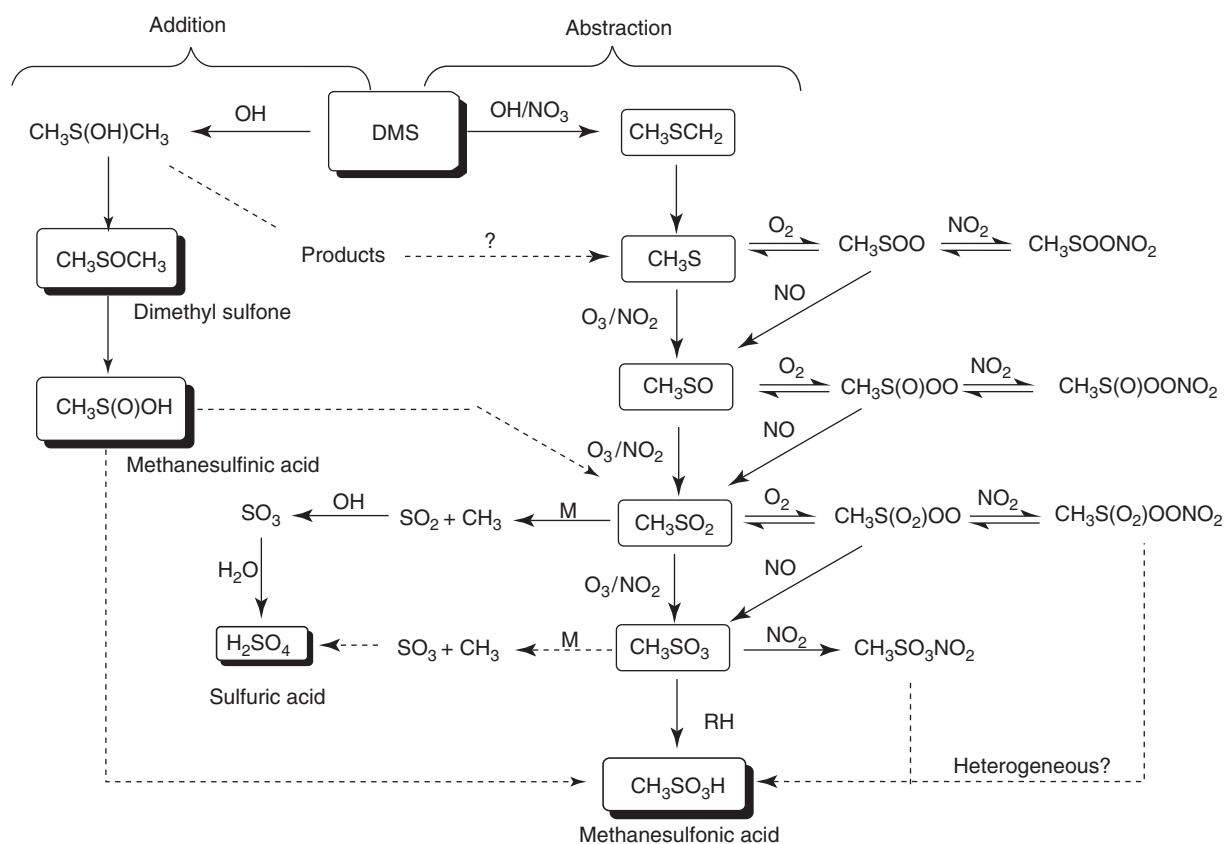


Figure 4 Simplified chemical mechanism for the atmospheric gas-phase reaction of OH radicals with DMS.

the primary source of CCN in the remote MBL and that sea salt contributes little to this CCN population. However, recent evidence suggests that the contribution from sea salt to CCN can often be quite substantial and under conditions of high wind speed may even dominate, and also that scavenging of SO₂ and H₂SO₄ by sea-salt aerosol is significant. These findings show that nonsea-salt sulfate and sea-salt aerosols cannot be considered as distinct populations and imply that any predicted climate changes based solely upon changes in nonsea-salt sulfate CCN concentrations are likely to be substantially overstated. Second, a recent comprehensive, geographically resolved inventory of ship emissions has shown that in large sections of the Northern Hemisphere, and in some regions of the Southern Hemisphere, anthropogenic sulfur emissions from ships are comparable to biogenic DMS emissions, the dominant source of sulfur from the ocean. Ship emissions have the potential to affect air quality in many coastal and port regions along heavily traveled international trade routes. The results suggest that emissions of sulfur and particulate matter from the international shipping industry need to be considered in the study of marine and coastal atmospheres.

As evidenced by the material presented here, further progress in the elucidation of the DMS oxidation mechanism and its implications for the Earth's climate will require both further advances in field studies and detailed kinetic studies combined with modeling of the field data (Figures 3 and 4).

See also: **Aerosols:** Aerosol–Cloud Interactions and Their Radiative Forcing. **Biogeochemical Cycles:** Sulfur Cycle. **Clouds and Fog:** Classification of Clouds; Climatology; Cloud Microphysics; Cloud Modeling; Noctilucent Clouds. **Mesoscale Meteorology:** Cloud and Precipitation Bands. **Tropospheric Chemistry and Composition:** Cloud Chemistry.

Further Reading

- Berresheim, H., Wine, P.H., Davis, D.D., 1995. Sulfur in the atmosphere. In: Singh, H.B. (Ed.), *Composition, Chemistry, and Climate of the Atmosphere*. Van Nostrand Reinhold, New York, pp. 251–307.
- Finlayson-Pitts, B.J., Pitts, J., 1999. *Chemistry of the Upper and Lower Atmosphere*. Academic Press, London.
- Hansen, L.D., Eatough, D.J., 1991. Organic oxysulfur compounds in the atmosphere. In: Hansen, L.D., Eatough, D.J. (Eds.), *Organic Chemistry of the Atmosphere*. CRC Press, Boca Raton, FL, pp. 199–232.
- Katoshevski, D., Nenes, A., Seinfeld, J.H., 1999. A study of processes that govern the maintenance of aerosols in the marine boundary layer. *Journal of Aerosol Science* 30, 503–532.
- Keene, W.C., Sander, R., Pszenny, A.A.P., 1998. Aerosol pH in the marine boundary layer – a review and model evaluation. *Journal of Aerosol Science* 29, 339–356.
- O'Dowd, C., Smith, M.H., Consterdine, I.E., Lowe, J.A., 1997. Marine aerosol, sea-salt, and the marine sulfur cycle: a short review. *Atmospheric Environment* 31, 73–80.
- Ravishankara, A.R., Rudich, Y., Talukdar, R., Barone, S.B., 1997. Oxidation of atmospheric reduced sulfur compounds: perspective from laboratories. *Philosophical Transactions of the Royal Society of London* 332, 171–182.
- Saltzman, E.S., Cooper, W.J. (Eds.), 1989. *Biogenic Sulfur in the Environment*. American Chemical Society, Washington, DC.
- Seinfeld, J.H., Pandis, S.N., 1998. *Atmospheric Chemistry and Physics*. Wiley, New York.
- Warneck, P., 1999. *Chemistry of the Natural Atmosphere*, second ed. Academic Press, London.
- Yin, F., Grosjean, D., Seinfeld, J.H., 1990. Photooxidation of dimethyl sulfide and dimethyl disulfide. I: mechanism development. *Journal of Atmospheric Chemistry* 11, 309–364.
- Yin, F., Grosjean, D., Flagan, C.R., Seinfeld, J.H., 1990. Photooxidation of dimethyl sulfide and dimethyl disulfide. II: mechanism evaluation. *Journal of Atmospheric Chemistry* 11, 365–399.

Volatile Organic Compounds Overview: Anthropogenic

RG Derwent, rdscientific, Newbury, UK

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

The term anthropogenic volatile organic compounds (VOCs) refers to those organic compounds other than methane that arise from human activities and are capable of producing photochemical oxidants such as ozone by atmospheric chemical reactions with nitrogen oxides in the presence of sunlight. VOCs are an important class of air pollutants commonly found in the atmosphere at ground level in all urban and industrial centers. There are many hundreds to thousands of compounds, which come within the category of VOCs. Not all VOCs contribute equally strongly to photochemical oxidant formation. Some VOCs show a high propensity to form oxidants whilst others are largely unreactive.

Definition

The term anthropogenic volatile organic compounds (VOCs) refers to those organic compounds other than methane that arise from human activities and are capable of producing photochemical oxidants such as ozone by atmospheric chemical reactions with nitrogen oxides in the presence of sunlight.

Background

The role and importance in atmospheric chemistry of VOCs produced by human activities was established about 60 years ago by Haagen-Smit in his pioneering studies of Los Angeles smog (Haagen-Smit et al., 1953). He identified the key role of hydrocarbon (HC) oxidation, in the presence of sunlight and oxides of nitrogen (NO_x), as a photochemical source of ozone (O_3) and other oxidants. Detailed understanding of the mechanism of photochemical smog formation has developed since then through the combination of smog chamber, laboratory chemical kinetics, field experiment, air quality monitoring, and computer modeling studies.

Since these early pioneering studies in Los Angeles, photochemical smog and the elevated ozone levels associated with it, have subsequently been detected in almost all of the world's major urban and industrial centers, at levels which exceed internationally agreed criteria set to protect human health (World Health Organisation, 2006). Despite the importance given now to VOCs, their routine measurement in the atmosphere has only recently become commonplace. Furthermore, there are few detailed emission inventories for the major urban and industrial centers for which anthropogenic VOC emissions are fully resolved by species.

Properties of VOCs

VOCs are an important class of air pollutants commonly found in the atmosphere at ground level in all urban and industrial centers. There are many hundreds to thousands of compounds, which come within the category of VOCs and the situation is yet further complicated by different definitions and nomenclature. Strictly speaking, the term VOC refers to those organic

compounds, which are present in the atmosphere as gases but which under normal conditions of temperature and pressure would be liquids or solids. A VOC is by definition, an organic compound whose vapor pressure at say 20 °C is less than 760 torr (101.3 kPa) and greater than 1 torr (0.13 kPa). Many common and important organic compounds would be ruled out of consideration if these upper and lower limits were strictly adhered to.

Here, this strict definition is not applied and the term VOC is taken to mean any carbon-containing compound found in the atmosphere, excluding elemental carbon, carbon monoxide, and carbon dioxide. This definition is deliberately wide and encompasses both gaseous carbon-containing compounds and those similar compounds adsorbed onto the surface of atmospheric suspended particulate matter. The definition here includes substituted organic compounds so that oxygenated, chlorinated, and sulfur-containing organic compounds would come under the definition of VOCs. Some of the most strongly and widely emitted VOCs are listed in Table 1. Although this top 50 list has been compiled using emissions data for the United Kingdom, the major VOC sources, such as motor vehicle exhausts and solvent usage, are common across much of northwest Europe and so the list has a wider relevance than just to the United Kingdom.

Other terms used to represent VOCs are HCs, reactive organic gases (ROGs), and nonmethane VOCs. The use of common names for the organic compounds is preferred here since these are more readily understood by industry and more commonly used in the air pollution literature. IUPAC names are, however, provided in all cases in Table 1 where they differ significantly from the common names.

VOCs play a crucial role in ground level photochemical oxidant formation since they control the rate of oxidant production in the presence of sunlight in those areas where NO_x levels are sufficient to maintain ozone production. The contribution that organic compounds make to the exceedance of environmental criteria for ozone is now widely recognized. Long-range transboundary transport of ozone is an important feature of the problem. Organic compounds come within the scope of the Geneva Protocol to the United Nations Economic Commission for Europe International Convention on Long-range Transboundary Air Pollution (http://www.unece.org/env/lrtap/vola_h1.html). Ground level ozone is of concern

Table 1 Common names, IUPAC names, percentage emission by mass, and POCPs of the 50 most prolifically emitted VOCs in the United Kingdom

Common name	IUPAC name	Percentage emission by mass, %	POCP
<i>n</i> -butane	butane	8.9	31
ethanol		7.4	34
ethane		5.0	8
propane		3.9	14
toluene	methylbenzene	3.6	44
<i>n</i> -pentane	pentane	3.5	40
<i>i</i> -pentane	2-methylbutane	3.2	34
ethylene	ethene	2.8	100
<i>n</i> -hexane	hexane	2.6	40
methanol		2.0	13
<i>m</i> -xylene	1,3-dimethylbenzene	2.0	86
trichloroethene		2.0	29
2-methylpropane		1.9	28
formaldehyde	methanal	1.7	46
acetone	propanone	1.6	6
<i>n</i> -heptane	heptane	1.4	35
ethylbenzene		1.3	46
propylene	propene	1.2	117
<i>n</i> -octane	octane	1.2	34
benzene		1.1	10
methylethylketone	2-butanone	1.1	32
<i>o</i> -xylene	1,2-dimethylbenzene	1.0	78
1,2,4-trimethylbenzene		1.0	110
dichloromethane		1.0	3
butyl acetate		1.0	26
acetylene	ethyne	0.9	7
<i>p</i> -Xylene	1,4-dimethylbenzene	0.9	72
2-propanol		0.8	18
2-methylpropene		0.8	63
ethyl acetate		0.8	19
<i>n</i> -decane	decane	0.8	36
tetrachloroethene		0.5	1
4-methyl-2-pentanone		0.5	52
2-butene		0.5	113
1-butanol		0.4	52
<i>n</i> -nonane	nonane	0.4	34
2-butoxyethanol		0.4	45
1,3,5-trimethylbenzene		0.4	107
acetaldehyde	ethanal	0.4	55
1,3-butadiene		0.4	89
2-methylpentane		0.4	41
methyl acetate		0.4	7
undecane		0.3	36
2-pentene		0.3	111
1-propanol		0.3	48
1-methoxy-2-propanol		0.3	34
1,2,3-trimethylbenzene		0.3	105
methylethylbenzene		0.3	73
2-methylhexane		0.3	32
3-methylpentane		0.2	37

Percentage emissions by mass taken from <http://naei.defra.gov.uk/>.
 POCPs taken from Derwent, R.G., Jenkin, M.E., Passant, N.R., Pilling, M.J., 2007. Environmental Science Policy 10, 445–453.

not only with respect to human health but also because of its effects on crops, plants, and trees. Elevated ozone concentrations during summertime photochemical pollution episodes may exceed environmental criteria set to protect human health and natural ecosystems. It is these concerns, which led to the formulation of the Geneva Protocol and which underpin the reductions in emissions and control actions, which it stipulates.

Not all VOCs contribute equally strongly to photochemical oxidant formation. Some VOCs are highly reactive and show a high propensity to form oxidants. Others are largely unreactive and make little or no contribution to ground level ozone formation. Reactivity scales have been constructed to provide an indication of each VOCs' propensity to contribute to ground level ozone formation. One such scale is the photochemical ozone creation potential (POCP) scale (Derwent et al., 2007). POCPs are expressed relative to ethylene (=100) and give the formation potential of each VOC on a mass emitted basis. The POCPs of the top 50 most strongly and widely emitted VOCs are listed in Table 1. POCPs range from 1 for the most unreactive chlorinated solvent, tetrachloroethylene, a widely used dry cleaning agent to 117 for the highly reactive propylene, an important component of motor vehicle exhaust. Representatives can be found in Table 1 of some of the more important classes of organic compounds including alkanes, alkenes, aromatic, and oxygenated organic species.

Sources of VOCs

There are a wide range of sources in urban and industrial areas that give rise to VOC emissions. Motor vehicle traffic is an important source of VOCs, which arise from the exhaust and evaporative emissions of petrol-engined road vehicles. Motor spirit has a high vapor pressure and so emissions of VOCs can occur from moving and stationary vehicles, from the petrol stations when vehicle and storage tanks are being filled, and from the petrol distribution chain that supplies petrol station forecourts. VOC emissions arise from oil refineries and industrial facilities that produce bulk organic chemicals. VOCs are often valuable solvents and are widely used in the home and in industry in paints, lacquers, varnishes, inks, and degreasing agents (<http://www.ceip.at/>). Some of the most widely emitted VOC species are listed in Table 1, together with their fractional emission rates and POCPs.

To reduce exposure levels to elevated ozone levels, actions have been taken to control photochemical ozone formation in many countries across the world. Because ozone is not emitted into the atmosphere and all the ozone present at ground level has been formed there by atmospheric chemical reactions, these control actions have to act on the emissions of the ozone precursors: VOCs and NO_x. Actions to control VOC emissions have usually involved fitting three-way exhaust gas catalyst and evaporative canister systems to petrol-engined motor vehicles and by tackling emissions from oil refineries and the petrol distribution chain. Substitution of water-based paints, for example, has led to reductions in solvent emissions from homes. In this way, much progress has been achieved to reduce episodic peak ozone levels and hence exceedances of the environmental criteria set to protect human health, crops, and vegetation.

References

- Derwent, R.G., Jenkin, M.E., Passant, N.R., Pilling, M.J., 2007. Environmental Science Policy 10, 445–453.
- Haagen-Smit, A.J., Bradley, C.E., Fox, M.M., 1953. Industrial and Engineering Chemistry Research 45, 2086.
- World Health Organisation, 2006. Air Quality Guidelines. Global Update 2005. WHO Regional Office for Europe, Copenhagen, Denmark.

Further Reading

- Finlayson, B.J., Pitts, Jr., J.N., 2000. Chemistry of the Upper and Lower Atmosphere. Academic Press, San Diego, CA.
- Seinfeld, J.H., Pandis, S.N., 2006. Atmospheric Chemistry and Physics. John Wiley and Sons, Hoboken, NJ.

TURBULENCE AND MIXING

Contents

Overview

Turbulence, Two Dimensional

Turbulent Diffusion

Overview

P Haynes, University of Cambridge, Cambridge, UK

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

The three distinct processes of transport, stirring, and mixing are defined and described. The connection between relative dispersion in backward time and mixing is explained. Mixing in three-dimensional flows (e.g., boundary layer turbulence) and large-scale chaotic advection flows (e.g., in the free troposphere and stratosphere) is compared. A brief account is given of the processes which lead to mixing in the real atmosphere, particularly outside of the boundary layer, including how those processes have been constrained by chemical observations and modeling.

Introduction

The atmospheric flow has an important effect on the distribution of chemical species by moving them away from the location of their sources and toward the location of their sinks. This process is *transport*. Unless the flow is uniform in space, it also distorts the geometric structure of chemical concentration fields so as to bring air parcels with different chemical character into closer and closer proximity. This process may be called *stirring*. The stirring processes typically draw out a volume of air (perhaps air with an anomalous chemical concentration) into thin filaments or sheets. Ultimately, the distance between air parcels of differing chemical concentration is so small that molecular diffusion may act rapidly to homogenize the chemical concentration fields. This latter process is *mixing*. The three processes of transport, stirring, and mixing are depicted schematically in [Figure 1](#).

Without the stirring effect of the flow, timescales for diffusive homogenization would be extremely long. Consider the advection–diffusion equation for the concentration $\chi(\mathbf{x}, t)$ of some chemical species.

$$\frac{D\chi}{Dt} = \frac{\partial\chi}{\partial t} + \mathbf{u} \cdot \nabla\chi = \kappa \nabla^2 \chi \quad [1]$$

$$\text{Estimated size } \frac{\chi U}{L} \quad \frac{\kappa \chi}{L^2}$$

where $\mathbf{u}(\mathbf{x}, t)$ is the velocity field, with $\mathbf{x}$ a vector representing spatial position and t time, and κ is the molecular diffusivity. Suppose that a characteristic length scale for the concentration field and the velocity field is L and the characteristic velocity scale is U . These scales might, for example, be imposed by boundary conditions, geometry, or forcing. The corresponding estimates for the advection and diffusion terms (respectively the second term on the left-hand side and the term on the right-hand side) in [eqn \[1\]](#) are shown below the equation. Note first that the ratio of the advection term to the diffusion term is UL/κ , generally called the Peclet number (Pe). If this number is large then advection dominates diffusion. If it is small then diffusion dominates advection. It also follows from the above that the timescale for diffusion is given by L^2/κ .

In the atmosphere the molecular diffusivity of most gaseous chemical species varies from about $10^{-5} \text{ m}^2 \text{ s}^{-1}$ in the troposphere to $10^{-3} \text{ m}^2 \text{ s}^{-1}$ in the upper stratosphere to higher values still in the mesosphere and thermosphere. The corresponding estimate for the time needed for molecular diffusion to homogenize chemical concentrations in the troposphere and lower stratosphere over a distance of, say, 100 m would therefore be more than 10^8 s , i.e., a few years. In reality, the effect of molecular diffusion is considerably enhanced by the stirring effect of the flow, which reduces the length scale of the concentration field, thereby increasing the size of the diffusive term on the right-hand side of [eqn \[1\]](#),

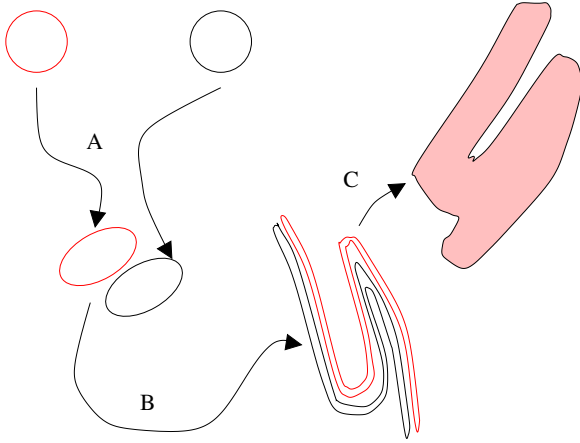


Figure 1 Schematic showing three-stage process leading to mixing. Two air masses containing different chemicals are initially well separated. In stage A *transport* brings the two air masses into the same region. (One might more accurately call this *relative transport*.) However, length scales are still relatively large and diffusive timescales are long. In stage B, stirring distorts the chemical distributions into filaments or sheets, thereby reducing their characteristic length scales. As a result diffusive timescales are reduced and in stage C *mixing* between the air masses can occur, leading to homogenization (at the molecular scale) of chemical concentrations. Note that stages A and B have been separated here for clarity, but in many cases it may not be appropriate to distinguish between A and B.

which has magnitude $\kappa\chi/l^2$ when the length scale of the concentration field is l . The diffusive term therefore becomes comparable to the advective term when $\kappa/l^2 \sim U/L$, i.e., when $l \sim L(\kappa/UL)^{1/2} = LPe^{-1/2}$.

Stretching, Relative Dispersion, and Mixing

The stirring process involves the geometrical deformation of material lines and surfaces. One useful measure of the stirring effect of the flow is the rate at which material lines or surfaces are stretched. For example, the equation for a material line element $l(t)$ (where $l(t)$ represents the difference in position between two nearby marked fluid particles) is

$$\frac{dl}{dt} = l \cdot (\nabla \mathbf{u}) \quad [2]$$

This equation is valid providing that the separation distance $|l|$ is smaller than the length scale on which the flow varies. If there is a systematic tendency for nearby particles to separate then, for any pair of particles with finite initial separation, eqn [2] will eventually become invalid. Note that velocity gradient tensor on the right-hand side of this equation must be evaluated following one of the fluid particles defining the line element. It is therefore the time history of this tensor following the flow that determines the stretching. A very simple example is that where the flow is two dimensional and the velocity gradient tensor is steady, taking the form

$$(\nabla \mathbf{u})_{ij} = \left(\frac{\partial u_i}{\partial x_j} \right) = \begin{pmatrix} a & b-c \\ b+c & -a \end{pmatrix} \quad [3]$$

(chosen to match the requirement that the flow is incompressible, i.e., $\partial u_1/\partial x_1 + \partial u_2/\partial x_2 = 0$). Then it is straightforward to solve eqn [2] to show that if $a^2 + b^2 > c^2$ the solution is $l(t) = l_+ \mathbf{e}_+ \exp(\sqrt{a^2 + b^2 - c^2}t) + l_- \mathbf{e}_- \exp(-\sqrt{a^2 + b^2 - c^2}t)$, where l_+ and l_- are constants and $\mathbf{e}_+$ and $\mathbf{e}_-$ are constant vectors. The first term dominates at large time implying that the length of line elements increases exponentially with time and hence that there is effective relative dispersion of nearby fluid particles. On the other hand, if $a^2 + b^2 < c^2$, then the solutions of eqn [2] are oscillatory, i.e., there is no systematic increase with time.

The model above in which the velocity gradient tensor is constant in time is not very relevant to real atmospheric flows. (Note the model corresponds to the tensor being constant following an air parcel, not simply constant at fixed points in space.) In many realistic flows, particularly turbulent flows, the velocity gradient tensor experienced by an air parcel has a complex time variation and might be represented as a random function of time. A model where the velocity gradient tensor is represented as a random function, with certain specified statistical properties, is called a random-straining model. Such models predict that the exponential stretching is robust, although there is an important dependence of the rate of increase on the correlation time for the random variation of the velocity gradient tensor. More detailed investigation of more realistic models for a wide variety of atmospheric flows shows that there is very often exponential stretching of line elements at a rate that roughly corresponds to the size of the components of the velocity gradient tensor. Exceptions might be in interior of long-lived eddies, where the regime for oscillatory solutions of eqn [2] might be relevant, and in cases where the correlation time for the velocity gradient tensor is very short, when the stretching rate will be reduced.

To emphasize the implications of material line lengthening and relative dispersion for stirring and mixing, it is useful to consider the evolution of a small material surface (assumed smaller than the length scale on which the velocity field varies) that is initially a sphere (or, in two dimensions, a small material contour that is initially a circle). The tendency of line elements to stretch, as described by eqn [2], implies that the sphere is deformed into an ellipsoid, at least one axis of which systematically increases in time. In an incompressible flow, the volume of the sphere remains constant with time, therefore the systematic increase in length of one axis is inevitably accompanied by the systematic decrease in length of another axis. This is a manifestation of the scale reduction that leads to mixing. In a compressible flow, there is no absolute constraint on the volume of the sphere, but nonetheless it is the case that in almost all flows the density will not systematically reduce, implying again that one axis must systematically reduce in length. The geometry of the ellipsoidal material surface becomes more complicated when its maximum dimension becomes as large as the length scale on which the velocity field varies. The surface is then strongly distorted and folded as different parts of the surface sample very different velocity gradients.

The relevance of deformation of material surfaces or curves to the evolution of the concentration of a chemical species is emphasized by noting that a similar picture holds in backward time (see Figure 2). Neglecting the effects of diffusivity for the present, the values of concentration in a small spherical region will be the values that were present in the same material region

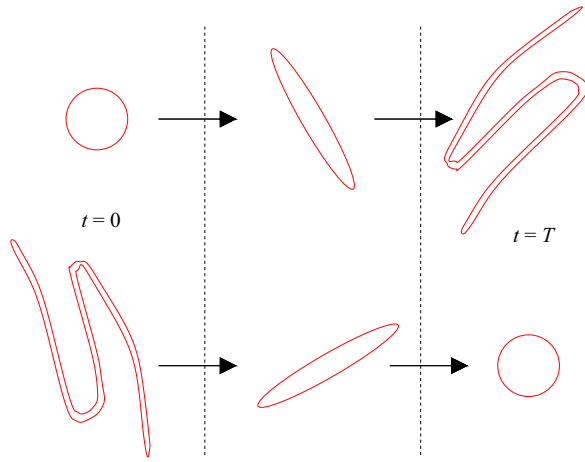


Figure 2 Schematic of the deformation of two material curves/surfaces. The top panel shows a small circle at time $t = 0$, which is then deformed into an ellipse (while its maximum dimension is less than the characteristic scale of the flow) and then into a more complex structure (as different parts of the curve experience very different velocity fields). The bottom panel shows a small circle at time $t = T$, which originated from a complex filamental structure at $t = 0$. This structure may be obtained by deforming the circle in backward time. The values of chemical concentration inside the circle at $t = T$ are just those sampled by the filamental structure at time $t = 0$. (Note that the two panels do not imply any kind of reversibility – the lower panel corresponds to a particular choice of initial condition that involves into a circle at time T . If the evolution was continued after time T the circle would stretch and eventually become geometrically complex, much as in the top panel.)

at the initial time. If that material region is stretched (in backward time) to length scales greater than those on which the concentration varies in the initial condition, then the small spherical region will contain a wide range of different concentration values, and it can be safely assumed that the effect of diffusivity will eventually be to homogenize those values over the region. The intimate relation between relative dispersion, i.e., the separation of nearby particles, and mixing has been exploited in many theoretical studies of the mixing problem.

The stirring and mixing process has so far been described as completely generic. One could equally well be considering the mixing of a smoke plume from a factory into the surrounding boundary layer air, or the mixing into the upper troposphere of boundary layer air that has been lofted in a convective cloud or a convective complex, or the mixing of stratospheric ozone-depleted Antarctic air into midlatitudes as the polar vortex breaks up in the late spring. These examples range in scales from a hundred meters or so to several thousand kilometers. But the flows that are responsible for stirring and mixing in each of the cases are very different and that has important implications for the stirring and mixing process.

Mixing in Three-Dimensional Turbulence

One important category of flow that gives rise to effective mixing is three-dimensional turbulence that occurs, for example, in the boundary layer and in convective clouds and

also in intermittent internal mixing events in the troposphere and the stratosphere. The dynamics of three-dimensional turbulence is characterized by energy input at some large scale, L . If velocities at the large scale are of typical size U , then the energy input rate ε satisfies $\varepsilon \sim U^3/L$. The dynamics of the turbulent eddies transfer energy from the input scale to smaller and smaller scales through a so-called inertial range. The classical Kolmogorov scaling predicts that in the inertial range the velocity fluctuations at length scale l vary as $\varepsilon^{1/3} l^{1/3} = U(l/L)^{1/3}$ and therefore that the typical size of components of the velocity gradient tensor is $(U/L)(L/l)^{2/3}$. The energy transfer is terminated at the dissipation scale l_ν , often called the *Kolmogorov scale*, where molecular diffusion of momentum becomes competitive with other flow processes. l_ν may be estimated by assuming that stretching and diffusion timescales are equal, implying that $l_\nu \sim \nu^{3/4} \varepsilon^{-1/4}$, where ν is the momentum diffusivity. The rate of energy dissipation at the Kolmogorov scale must, in a steady state, equal ε , the rate of energy input at the large scale.

It follows that the largest contributions to the stretching come from the velocity field at the dissipation scale and furthermore the stretching rate at those scales is considerably larger (by a factor of $(UL/\nu)^{1/2}$) than that estimated using velocity and length scales at the energy input scale. (Note the velocity gradient is a good estimate for the inverse of the timescale for line stretching since, for three-dimensional turbulence, the time series of velocity gradient experienced by an air parcel is essentially random and, furthermore, the correlation time is of order the inverse of the velocity gradient at the dissipation scale.)

If the diffusivity κ for the chemical species is much less than ν , then the Peclet number at the Kolmogorov scale may be large and the eddies at this scale must stir the chemical concentration field to a smaller scale (often called the *Batchelor scale*) for diffusive homogenization to occur. However, for most chemical species of interest in the atmosphere we may assume $\kappa \sim \nu$ and therefore diffusive homogenization occurs at the Kolmogorov scale. The timescale required for a chemical concentration field that is initially on the energy input scale, i.e., large scale, to be geometrically deformed until molecular diffusion is important may be estimated by assuming that $(L/U)(l/L)^{2/3}$ is roughly the time required to reduce the scale from l to $l/2$. Summing the times needed to reduce the scale from L to l_ν gives a convergent geometric series whose sum is estimated by L/U multiplied by some moderate factor that is independent of ν and hence κ , if these are small. The timescale required for homogenization on the molecular scale is therefore the same order of magnitude as the time required for advection around the turbulent region on the large scale. Thus, the timescale for mixing within a region containing three-dimensional turbulent flow is the same order of magnitude as the timescale for transport across that region. The implication is that three-dimensional turbulence is very effective at mixing and that, to some reasonable approximation, if there is time for transport of air parcels across a turbulent region then there is also time for mixing within that region.

The fact that stretching is dominated by velocity fluctuations at the Kolmogorov scale implies that the geometrical structure of the chemical concentration field also varies on this scale. An example of a typical chemical concentration field is shown in **Figure 3** (top panel). (This field has been obtained by

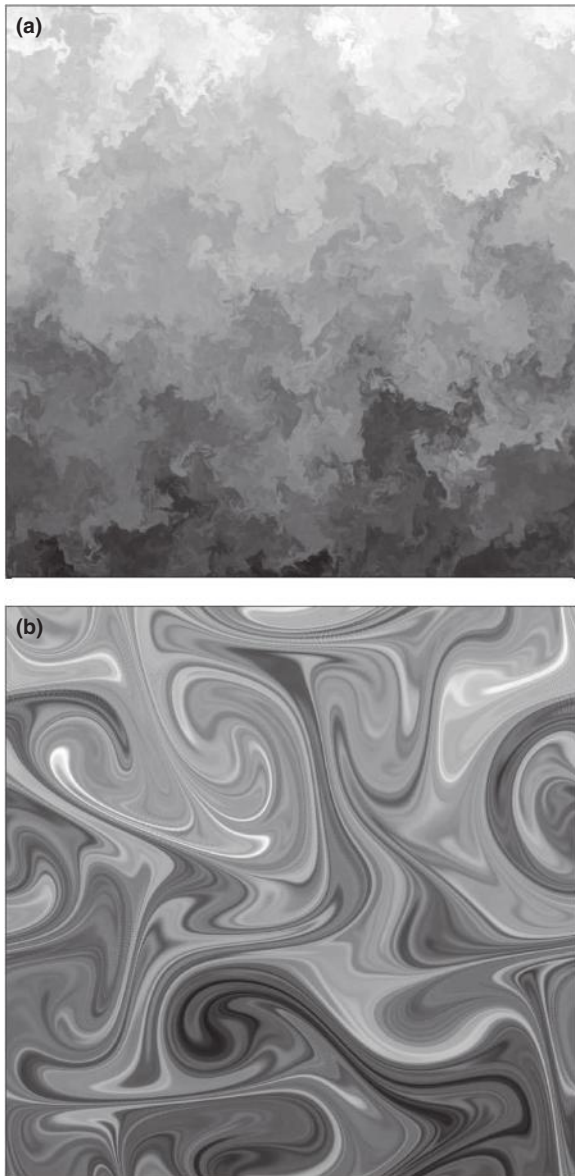


Figure 3 (a) Chemical concentration field resulting from numerical solution of the advection-diffusion equation for a two-dimensional flow in which there is small-scale energy input. This gives rise to an energy spectrum similar to that predicted by Kolmogorov scaling for three-dimensional turbulence and many aspects of the chemical concentration field shown here, particularly the small-scale geometry, are as they would be in three-dimensional turbulence. (The energy transfer in the two-dimensional case is upscale rather than downscale, but this by itself is not particularly important for the chemical concentration field.) (b) Chemical concentration field resulting from numerical solution of the advection-diffusion equation for specified large-scale flow in which the velocity is a smoothly but randomly varying function of time.

numerical solution of the advection-diffusion equation in a model flow, not a three-dimensional turbulent flow but an artificially forced two-dimensional flow that also exhibits Kolmogorov scaling.) The orientation of the contours of concentration changes on the scale of the smallest eddies in the flow.

Mixing in three-dimensional turbulent flows has been much studied in the laboratory and is of great practical significance because of its relation to the dispersion of pollutants in the atmospheric boundary layer. Research has focused on considering the mixing problem via the problem of relative dispersion, and stochastic models have carefully been formulated and refined to describe this relative dispersion process.

Large-Scale Flows: Quasi-Two-Dimensional Turbulence and Chaotic Advection

Rotation and buoyancy stratification both tend to inhibit three-dimensional turbulence on sufficiently large scales. On scales larger than a few tens of kilometers transport is accomplished by frontal systems, cyclones, and anticyclones and, in the stratosphere, propagating and breaking planetary-scale Rossby waves. The stable stratification and geometric constraints ensure that air parcel trajectories are along weakly sloping surfaces, so that horizontal displacements are generally much larger than vertical displacements.

These flows have a dual character, with some aspects of their behavior appearing organized and wavelike and other aspects exhibiting considerable nonlinearity and randomness. In the latter respect, these flows might therefore be regarded as a kind of turbulence, analogous to the two-dimensional turbulence studied in idealized numerical simulations and laboratory experiments. However, the strong difference of these flows from three-dimensional turbulence is that strong vortex stretching is inhibited and therefore the cascade of energy to small scales is inhibited. The velocity field essentially has a finite spatial scale and hence there is no strong increase of velocity gradients as scale shrinks.

It is now realized that flows such as these, with a relatively simple structure in space and time, may be highly effective at stirring and mixing through a phenomenon known as ‘chaotic advection.’ The idea here is that velocity fields with a very simple structure in space and time may lead to complex and irregular, i.e., chaotic, particle trajectories with, for example, the distance between initially nearby particles increasing exponentially in time. As explained above, this separation of nearby particles implies that chemical concentration fields rapidly become complex, with variations on scales much smaller than that of the advecting flow.

In large-scale chaotic advection type flows the magnitude of velocity gradients is constant with scale and may be estimated as L/U , where L is a typical length scale and U is a typical velocity magnitude of the large-scale flow. In this case, diffusive mixing takes place on a length scale of $(\kappa L/U)^{1/2}$. The time to reach this scale from the large scale is $(L/U) \log (UL/\kappa)^{1/2}$, i.e., considerably larger than the timescale for transport by the large-scale flow. The implication is that in these flows advected chemical species may be spread across the flow domain on the large scale, but for a substantial time may remain unmixed on the small scale (in contrast to the case in three-dimensional turbulence). Furthermore, the fact that the velocity field itself varies only weakly on the scale $(\kappa L/U)^{1/2}$ suggests that the chemical concentration fields tend to be aligned in smooth (but thin) filaments or sheets. Such filaments or sheets of anomalous chemical species, that have not yet been mixed with

their environment, are widely observed in the atmosphere. An example of chemical concentration fields arising in a two-dimensional chaotic advection flow is shown in [Figure 3](#) (bottom panel). (Contrast with the top panel of [Figure 3](#).)

Mixing Processes in the Real Atmosphere

In the real atmosphere three-dimensional turbulence is inevitably affected by the presence of stable density stratification, which implies that energy input is needed to move fluid particles in the vertical. Broadly speaking, the effect of the stratification is to limit the vertical length scales of the turbulence and hence the vertical length scales over which mixing can occur. The dynamics of the mixing also tends to lead to the formation of interfaces at the upper or lower limits of the turbulent regions across which there are strong density gradients. The stability of these interfaces implies that mixing across them occurs relatively infrequently, often through complex phenomena such as the breaking of waves propagating on the interfaces themselves.

For the planetary boundary layer, the role of stable stratification is to limit the height of the turbulent layer. The strong stable stratification often found at the top of the boundary layer has an important effect on exchange between the turbulent boundary layer and the free atmosphere above it. But when considering transport within the boundary layer itself stratification may often be neglected.

In the 'free' atmosphere above the boundary layer, there is certainly no permanent 'background' field of three-dimensional turbulence. In the troposphere convection gives rise to turbulence, but the fraction of the atmosphere that is actively convecting at any instant is relatively small. In the upper troposphere and stratosphere, there is little or no convection, but localized regions of turbulence arise through shear instability or through breaking of inertia-gravity waves. These turbulent regions form sporadically, mix in a localized region, and then decay. Thus, outside the planetary boundary layer, turbulent mixing occurs through the net effect of localized, intermittent turbulent events.

Evidence of the effect of turbulence mixing in the free troposphere and stratosphere comes, for example, for combined observational and modeling studies of layered structures in chemical species, such as ozone, water vapor, and carbon monoxide. Layered structures, with vertical scales of a few hundred meters, have commonly been observed in the troposphere, for example, from *in situ* observations from aircraft. The fact that these structures have survived for several days, deduced from the distance between the observation and the likely region of origin, implies that turbulent mixing cannot be too strong. On the other hand, careful comparison between observations (sometime multiple observations of the same air mass as it is transported over several days) and predictions from Lagrangian models (where air parcels are transported by

the wind fields defined by meteorological analysis data sets) shows that some mixing certainly takes place, typically corresponding to a vertical diffusivity of $1 \text{ m}^2 \text{ s}^{-1}$ in the troposphere and $0.1 \text{ m}^2 \text{ s}^{-1}$ in the lower stratosphere.

The route by which air masses with different chemical characteristics are mixed in the free atmosphere can be summarized as follows. Long-range transport and stirring on the large scale is accomplished by quasi-horizontal motion that draws chemical fields out into thin sloping sheets (which appear as thin filaments on a quasi-horizontal surface). If there was no three-dimensional turbulence, then mixing would eventually be achieved by molecular diffusion alone, but this mixing would occur at very small scales. In practice the effects of molecular diffusion are enhanced, probably through intermittent encounters of air parcels with three-dimensional turbulence. This enhancement appears to be substantial in the troposphere, where convection and other mechanisms for generation of active three-dimensional turbulence are common, weaker in the lower stratosphere, and then to increase again in the upper stratosphere and mesosphere, where three-dimensional turbulence associated with the breaking of gravity waves is more widespread.

See also: Aviation Meteorology: Clear Air Turbulence. **Boundary Layer (Atmospheric) and Air Pollution:** Coherent Structures; Complex Terrain; Convective Boundary Layer; Modeling and Parameterization; Observational Techniques *In Situ*; Observational Techniques: Remote; Ocean Mixed Layer; Overview; Stably Stratified Boundary Layer; Surface Layer. **Dynamical Meteorology:** Kelvin–Helmholtz Instability. **Mountain Meteorology:** Lee Waves and Mountain Waves. **Numerical Models:** Parameterization of Physical Processes: Clouds; Parameterization of Physical Processes: Gravity Wave Fluxes; Parameterization of Physical Processes: Turbulence and Mixing. **Turbulence and Mixing:** Turbulence, Two Dimensional; Turbulent Diffusion.

Further Reading

- Fernando, H.J.S., 1991. Turbulent mixing in stratified fluids. *Annual Review of Fluid Mechanics* 23, 455–493.
- Ivey, G.N., Winters, K.B., Koseff, J.R., 2008. Density stratification, turbulence, but how much mixing? *Annual Review of Fluid Mechanics* 40, 16984.
- Ottino, J.M., 1990. Mixing, chaotic advection, and turbulence. *Annual Review of Fluid Mechanics* 22, 207–253.
- Pierrehumbert, R.T., Yang, H., 1993. Global chaotic mixing on isentropic surfaces. *Journal of Atmospheric Sciences* 50, 2462–2480.
- Pisso, I., Real, E., Law, K.S., Legras, B., Bousserez, N., Attie, J.L., Schlager, H., 2009. Estimation of mixing in the troposphere from Lagrangian trace gas reconstructions during long-range pollution plume transport. *Journal of Geophysical Research* 114, D19301.
- Sawford, B., 2001. Turbulent relative dispersion. *Annual Review of Fluid Mechanics* 33, 289–318.
- Shraiman, B.I., Siggia, E.D., 2000. Scalar turbulence. *Nature* 405, 639–646.

Turbulence, Two Dimensional

P Bartello, McGill University, Montréal, QC, Canada

© 2003 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 6, pp 2451–2455, © 2003, Elsevier Ltd.

Introduction

Historical interest in two-dimensional (2D) turbulence in the atmospheric and oceanic sciences stems from the fact that the equation describing the conservation of potential vorticity in a thin layer of fluid on a rotating sphere (based on the shallow-water equations) reduces to the much simpler equations for a 2D incompressible fluid (the Navier–Stokes equations) when the characteristic horizontal length scale, L , is small enough such that the flow does not feel the effect of the Earth's sphericity and yet large enough such that the quasi-geostrophic approximation is valid. In this case, the governing equations reduce to

$$\frac{\partial \zeta}{\partial t} + \mathbf{u} \cdot \nabla \zeta = \nu \nabla^2 \zeta + F, \quad \nabla \cdot \mathbf{u} = 0 \quad [1]$$

where $\zeta = \hat{\mathbf{k}} \cdot \nabla \times \mathbf{u}$ is the vertical vorticity, $\mathbf{u}$ is the velocity, $\hat{\mathbf{k}}$ is the unit vector normal to the flow plane, and F represents an unspecified vorticity forcing. In the limit where the Reynolds number, $Re = UL/\nu$ (U is a characteristic velocity), is large, the behavior is often highly nonlinear and turbulent. Although the range of length scales over which this is a good approximation to geophysical fluid flow is at most a factor of 10–100, eqn [1] has been used as a simple conceptual model. This is partially justified by the fact that, even though large-scale atmospheric flow is not exactly 2D, this is the simplest setting in which to explore the nature of large nonlinearity. In addition, the strong stratification in the atmosphere inhibits vertical velocity, yielding approximately layerwise 2D flow. Since analytical techniques are lacking, laboratory experimentation and numerical simulation have provided a large part of our understanding today. Clearly, a much wider range of active length scales can be simulated in 2D than in 3D.

Cascade Phenomenologies

Theoretical approaches to turbulence have traditionally been based on the fact that, no matter how nonlinearly complicated the flow is, in the absence of forcing and molecular viscosity, it must conserve its full set of invariants, such as energy. 2D turbulence is distinguished from its 3D counterpart by the lack of vortex tube stretching, implying that the vorticity is conserved by each fluid parcel. A distinguished role in 2D turbulence theory is played by the mean-square vorticity or enstrophy, $Z = \frac{1}{2} \langle \zeta^2 \rangle$, where angle brackets are averages over a statistical ensemble of such flows. Simultaneous conservation of enstrophy and energy, $E = \frac{1}{2} \langle \mathbf{u} \cdot \mathbf{u} \rangle$, yield turbulent behavior in 2D fluids which is quite different from that in 3D. This is best illustrated by expressing the fields in terms of their Fourier coefficients. If we consider flow in a doubly periodic domain of length L and then take the limit $L \rightarrow \infty$, we can describe turbulence that is statistically homogeneous,

meaning that statistics do not depend on the position at which they are measured. In such a domain, a field ζ can be expressed as

$$\zeta(\mathbf{r}, t) = \sum_{\mathbf{k}} \hat{\zeta}_{\mathbf{k}}(t) e^{i\mathbf{k} \cdot \mathbf{r}} \quad [2]$$

where $\mathbf{r} = (x, y)$ is the position vector and $\mathbf{k} = (k_x, k_y)$ is the wave vector. The enstrophy is

$$Z = \frac{1}{2} \sum_{\mathbf{k}} \langle \hat{\zeta}_{\mathbf{k}}(t) \hat{\zeta}_{\mathbf{k}}^*(t) \rangle \quad [3]$$

where $*$ denotes complex conjugation. After taking the limit $L \rightarrow \infty$, the sums can be expressed as integrals. The 2D wave vector integration can be performed using polar coordinates $k = (k_x^2 + k_y^2)^{1/2}$ and $\theta_k = \tan^{-1} k_y/k_x$. Since eqn [1] is statistically isotropic, the integration in θ_k simply gives $2\pi k$, yielding the enstrophy as

$$Z = \int Z(k) dk \quad [4]$$

where the enstrophy spectrum is $Z(k) = \pi k \langle \hat{\zeta}_{\mathbf{k}} \hat{\zeta}_{\mathbf{k}}^* \rangle$. The same operations on the velocity field yield

$$E = \int E(k) dk \quad [5]$$

where the energy spectrum is $E(k) = \pi k \langle \hat{\mathbf{u}}_{\mathbf{k}} \cdot \hat{\mathbf{u}}_{\mathbf{k}}^* \rangle$ and, owing to the relationship between vorticity and velocity, $Z(k) = k^2 E(k)$.

Consider a set of initial conditions characterized by a thin energy spectrum localized near some intermediate wavenumber k_i . If the initial spectral distribution is thin enough, we can quantify k_i by using the first moment of $E(k)$, i.e.,

$$k_i = \frac{\int k E(k) dk}{\int E(k) dk} \quad [6]$$

How will the energy spectrum evolve in time? One of the fundamental characteristics of turbulence is that it mixes the fluid. A very thin energy spectrum, implying that only a narrow range of scales is excited, is a rather ordered low-entropy state. As the nonlinear term spreads the energy around from mode to mode in the complex vortex interactions discussed below, we expect that the width, W , of the energy spectrum will increase. Defining W via

$$W^2 = \frac{\int (k - k_i)^2 E(k) dk}{\int E(k) dk} \quad [7]$$

we require that $dW^2/dt > 0$, implying

$$\frac{dW^2}{dt} = -\frac{dk_i^2}{dt} > 0 \quad [8]$$

taking into account the invariance of the energy and enstrophy along with the definition of k_i . If the width of the initial energy spectrum increases, then the characteristic wavenumber, k_i , must decrease. In other words, the typical length scale of the eddy motions must increase with time. This is in marked

contrast to 3D turbulence where the transfer is to eddies of ever-diminishing size.

At the same time as energy is transferred to larger scales, the enstrophy is transferred to smaller scales. This can be illustrated by the fact that each fluid parcel conserves its vorticity, implying that as turbulence mixes the fluid, vorticity contours are stretched out into ever-longer curves (see Figure 2). This has the consequence of reducing mean distances between contours, thereby increasing gradients. The variance of vorticity gradients can be expressed as $\int k^2 Z(k) dk$. If the enstrophy ($\int Z(k) dk$) is constant, but vorticity gradients increase, vorticity must be transferred to larger wavenumbers or smaller scales.

The notion that energy is transferred to larger scales while enstrophy is transferred to smaller scales in 2D turbulence led to the formulation of cascade phenomenologies analogous to that of Kolmogorov for 3D flow. A major difference between the 3D and 2D phenomenologies is that an *a posteriori* analysis shows the 2D result to be inconsistent with one of the hypotheses. This will be discussed further below. Assume that energy is injected by external forcing at a rate ε and also that this forcing is active only over a narrow range of wavenumbers centered on k_f . The forcing also injects enstrophy at a rate $\eta \sim k_f^2 \varepsilon$. We also assume that there is a long 'inertial' range of wavenumbers where neither the forcing nor the dissipation are important. This can only be true if the nonlinear transfer is local in k . In other words, it must proceed via numerous small jumps in wavenumber. In each step of the process some information on the nature of the forcing is lost, such that eventually at intermediate wavenumbers the result is independent of the forcing. In order to achieve statistical stationarity, the enstrophy injection at k_f at rate η must be balanced by its dissipation at small scales at the same rate. It therefore follows that there is a constant downscale flux η of enstrophy across the entire inertial range. Dimensional analysis then yields the energy spectrum

$$\eta = \text{const} \sim \frac{\text{enstrophy}(k)}{\text{time}(k)} \sim \frac{k^3 E(k)}{[k^3 E(k)]^{-1/2}} \quad [9]$$

where the quantity $\text{time}(k)$ refers to the time scale of enstrophy transfer at k . Here it was assumed that the enstrophy near

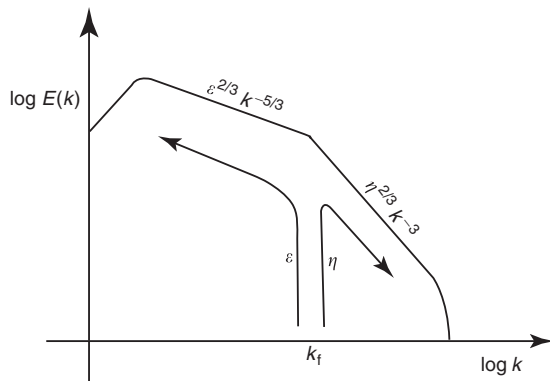


Figure 1 Energy spectrum of 2D turbulence at statistical stationarity showing the direct enstrophy cascade and the inverse energy cascade.

wavenumber k is $kZ(k)$ and that $\text{time}(k) \sim L(k)/U(k) \sim [k^3 E(k)]^{-1/2}$, since $U(k) \sim [kE(k)]^{1/2}$ and $L(k) \sim k^{-1}$. These relations assume the transfer to be local, i.e., the transfer into k comes from other modes near k , such that all quantities are evaluated locally. Solving for $E(k)$ gives

$$E(k) \sim \eta^{2/3} k^{-3} \quad [10]$$

for the energy spectrum in the downscale (or direct) enstrophy cascade range. As mentioned above, an *a posteriori* analysis of this theoretical spectrum reveals the nonlinear transfer in the inertial range to be nonlocal. In other words, instead of the enstrophy transfer proceeding by numerous small jumps in length scale, as required if the intermediate scales' dynamics are independent of F , the transfer takes rather larger jumps. In the upscale energy cascade range, the reasoning is as for Kolmogorov and

$$E(k) \sim \varepsilon^{2/3} k^{-5/3} \quad [11]$$

Strictly speaking, a mechanism is required to dissipate energy at scales much larger than the forcing scale in order for eqn [11] to reach the statistical stationarity assumed here. One physical possibility is the scale-independent dissipation inherent in the Ekman layer, provided that it is negligible over the inertial range. Both cascade ranges are shown schematically in Figure 1.

Numerical Simulations

As stated above, 2D turbulence is less computationally intensive than its 3D counterpart. In the early 1970s, numerical simulations began to be possible at low Re . They showed energy spectra consistently steeper than the k^{-3} prediction in the enstrophy cascade. By 1981, it was clear that a major surprise of these simulations was the development of intermittency in the vorticity field. Typically decaying (i.e., $F = 0$) turbulence was simulated starting from a random initial field with Gaussian statistics. Theory also predicts the energy spectrum [10] for the decay case. The numerical results showed that like-signed extrema had a tendency to merge and form larger vortices. In addition, close approaches of vortices stretched out long vorticity filaments whose width decreased until the dissipation scale was reached. The net result of these complicated vortex interactions is the growth of relatively large quiescent regions between a diminishing number of intense vortices. A typical snapshot of a simulation vorticity field is displayed in Figure 2. During the 1980s and 1990s, resolutions continued to increase and it became clear that the spectra were steeper than the k^{-3} prediction and that if (by some numerical trick such as scrambling the complex phases of the $\hat{\zeta}_k$) the intermittency was reduced, the evolution was more consistent with the k^{-3} phenomenology.

It must be stated that any inertial range argument assumes a sufficiently large range of wavenumbers such that forcing and dissipation are negligible. Unfortunately, it is not entirely clear how large this range must be for the approximation to be valid. It may be that resolutions much larger than those currently used (4096^2) are required. If this is the case, the relevance to atmospheric and oceanic dynamics is not obvious. It may also

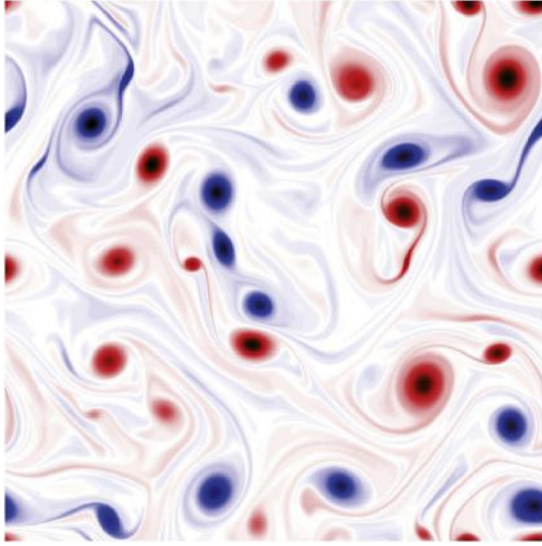


Figure 2 Vorticity field in a high-resolution simulation of decaying 2D turbulence from random initial conditions.

be that the nonlocalness of enstrophy transfer is responsible for the failure of eqn [10], although a theoretical attempt to account for this yields a spectrum (with a logarithmic correction factor) not much different from eqn [10], whereas the simulation spectra are considerably steeper.

In addition to simulations of decaying turbulence, high-resolution simulations of forced turbulence have been performed. They also tend to show spectra steeper than k^{-3} in the enstrophy cascade range as well as a dependence on the nature of the forcing mechanism throughout the resolved range, indicating that we have not yet achieved the resolutions required for universal (forcing-independent) behavior. Simulations of the enstrophy cascade are performed by forcing at a low wavenumber, whereas simulations of the inverse energy cascade result when k_f is larger. The latter have generally shown reasonable agreement with eqn [11], although this is not the case in all studies.

Vortex Dynamics

Current numerical and theoretical research on this subject focuses almost entirely on the vortices and their dynamics. The cascade phenomenologies discussed above are crude examples of statistical theories where the flow variables are assumed to be independent random oscillators weakly coupled by their contribution to the global invariants. More sophisticated theories based on this idea also fail in the presence of intermittent vortices. The reason is presumably that the Fourier transform of a thin isolated peak in x covers a wide range in k . As a result, the Fourier modes are not weakly coupled but rather describe long-lived structures moving around in space. This has caused some to advocate abandoning the traditional Fourier approach.

The intermittency of the vorticity in simulations of decaying 2D turbulence suggests that a model based on a collection of point vortices might be a good approximation. A vortex which

persists much longer than the time it takes a fluid parcel to circle around its center must be an essentially nondissipative structure. As a result, a Hamiltonian formulation has been employed and compared favorably with full simulations of eqn [1]. Ballistics theory, in which the vortices behave much like billiard balls, has also been applied. Unfortunately, these theories are not a natural setting to describe the observed mergers of like-signed vortices which lead to the growth of the characteristic length scale. Merger rules must be formulated and inserted into the theories by appealing to comparison with experiment or numerical simulation. The extent to which these empirically derived rules are independent of such details as the initial conditions is currently a matter of debate.

Since merger rules form an intrinsic part of discrete vortex theories, recent attention has been directed at the vortex interactions themselves. It has long been observed that the most intense vortices at the end of a long simulation can be traced all the way back to the initial conditions, where they were undistinguished extrema in a field with Gaussian statistics. Since the viscous term in eqn [1] can only diffuse vorticity down vorticity gradients, and since the other terms ensure conservation of vorticity by each fluid parcel, the central vorticity in a vortex can only decrease with time. The most intense vortices in the later stages are therefore those having decayed the least and those having sustained the least damage in their interactions with other vortices. A closer look at these interactions shows that, while some vortices do not survive, those that do are rendered steeper by the interactions. That is to say, the vorticity becomes more uniform in the center while the vorticity gradient is increased at the edges. This process has been called vortex erosion or stripping and it leads to a rather robust central value of vorticity which changes little with time.

Perhaps the first step in formulating vortex merger rules is to determine the statistics of the decay of ρ , the total number of vortices per unit area in decaying 2D turbulence. Numerical evaluations of this quantity differ from simulation to simulation as a function of the Reynolds number and the initial conditions. An interesting idea proposed to analyze a simulation with a narrow range of vortex amplitudes, ζ_{ext} , and radii, a . The total energy can be estimated dimensionally as

$$E \sim \rho a^4 \zeta_{\text{ext}}^2 \quad [12]$$

and the enstrophy is

$$Z \sim \rho a^2 \zeta_{\text{ext}}^2 \quad [13]$$

Since energy is transferred to larger scales and is therefore not dissipated significantly, E is held fixed. Based on the observed tendencies for the most intense vortices' central vorticities to be robust to interactions, ζ_{ext} is held constant too. Assuming ρ to decay as a power of time, $\rho \sim t^{-\xi}$, the other statistics follow as $a \sim t^{\xi/4}$ and $Z \sim t^{-\xi/2}$. The simulation discussed here gave $\xi \approx 0.75$. While these temporal scaling relations are based on the measured value of $\rho(t)$, they also provided good estimates of the simulation $a(t)$ and $Z(t)$. How well they apply when vortex properties are not narrowly distributed is still a matter of debate.

Vortex theories take no account of the low-amplitude vorticity filaments which can always be observed between the intense coherent structures. This background 'sea' of vorticity clearly interacts with the vortices in that they emerge from it at

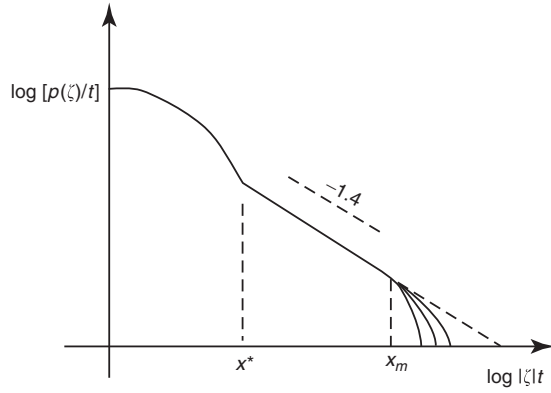


Figure 3 Schematic of the scaled vorticity probability density function. The dashed line represents the universal function, f (see text), while the solid lines denote the probability at several times such that the hyperbolic range widens with time.

small times. Their close interactions have also been observed to stretch out weak filaments which then decay into the background sea. It remains to be seen how dynamically important the sea is to the evolution of the turbulence. A schematic of the vorticity's probability density function for decaying 2D turbulence is shown in Figure 3. If the probability of finding a vorticity between ζ and $\zeta + d\zeta$ is denoted as $p(\zeta)d\zeta$, then a self-similar behavior has been observed in decaying 2D turbulence such that

$$p(\zeta) \sim t f(\zeta t) \quad [14]$$

except at large $|\zeta|t$, where the curve drops quickly to zero owing to the fact that the largest vorticities in the field are relatively constant in time. In eqn [14] f is a universal dimensionless function. Figure 3 shows that for low vorticities ($|\zeta|t < x^* \approx$) the probability density drops off rapidly with increasing ζ . The weak background sea of vorticity filaments is described by this range. When $x^* < |\zeta|t < x_m$ the probability density decays as

$$p(\zeta) \sim \frac{t}{(|\zeta|t)^{1.4}} \quad [15]$$

This is referred to as a hyperbolic distribution and it describes fields which are very intermittent indeed, corresponding to the isolated coherent vortices. At $|\zeta|t > x_m(t)$ the probability drops rapidly to zero. If extreme vorticities are conserved as in the simple scaling above, then $x_m(t) \sim t$ and the probability

density provides a more standard statistical definition of the quantity ζ_{ext} . How $x_m(t)$ behaves with time and as a function of the Reynolds number needs further investigation and is essentially a function of the vortex dynamics. The self-similar decay of vorticity shows an increasingly wide hyperbolic vortex range and a diminishing background sea as time proceeds.

Conclusions

Over the last 30 years numerical simulations of 2D turbulence have surprised researchers a great deal. The failure of the classical cascade phenomenologies and its attribution to the development of isolated coherent vortices has been the subject of much attention in the scientific literature. Clearly, a lot has been learned from the study of this simplest of models of a highly nonlinear turbulent fluid.

It was stated in the introduction that one of the reasons 2D turbulence has sparked such interest in the scientific community is that it was possible to test out at least some of the ideas numerically on even the relatively primitive computers of the early 1970s. Since then, the history of the subject parallels that of computational science. In spite of the fact that current computers allow scientists to move on to more realistic atmospheric and oceanic flows, many of the surprising elements of early simulations of 2D turbulence are still unexplained. The development of isolated vortices in the decay case is still somewhat mysterious and there exists no theory to describe their interactions once they have developed. There is still a lot more work to be done.

See also: Dynamical Meteorology: Vorticity. Numerical Models: Parameterization of Physical Processes: Turbulence and Mixing.

Further Reading

- Batchelor, G.K., 1953. *The Theory of Homogeneous Turbulence*. Cambridge University Press, Cambridge.
- Lesieur, M., 1992. *Turbulence in Fluids*. Kluwer, Dordrecht.
- Salmon, R., 1998. *Lectures on Geophysical Fluid Dynamics*. Oxford University Press, Oxford.

Turbulent Diffusion

A Venkatram, University of California – Riverside, Riverside, CA, USA

S Du, California Air Resources Board, Sacramento, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 6, pp 2455–2466, © 2003, Elsevier Ltd.

Glossary

1 AU Astronomical Unit = the mean Earth–Sun separation.

Dobson Unit Unit for total column ozone

(1 DU = 2.69×10^{16} molecule cm^{-2}). A column amount of 300 Du, which is a typical global average that corresponds to a 3-mm layer of pure ozone at standard temperature and pressure.

RAF Radiative amplification factor.

SZA Solar zenith angle: the angle between the local vertical and the Sun (SZA = 90° – solar elevation).

SPF Sun protection factor (for sunscreens).

TOMS Total Ozone Mapping Spectrometer (a satellite-borne ozone sensor).

Nomenclature

$\tilde{C}$ Instantaneous concentration.

C Mean concentration.

c Concentration fluctuation defined as $\tilde{C} - C$.

C_{peak} Peak concentration.

$\overline{C^y}$ Crosswind-integrated concentration.

D Molecular diffusivity of species in air.

h_s Source height.

H_0 Surface heat flux.

K Eddy diffusivity.

L Monin–Obukhov length.

l Length scale characterizing turbulent fluctuations.

N Brunt–Vaisala frequency.

Q Source strength for continuous emissions.

Q_m Total mass of a puff.

R_{Lw} Lagrangian autocorrelation for crosswind velocity.

s Puff dimension (or spread).

T_{av} Averaging interval.

T_c Timescale of concentration fluctuations.

T_{Lw} Lagrangian timescale for crosswind velocity.

T_{Lw} Lagrangian timescale for vertical velocity.

t Travel time from source to receptor.

$\tilde{U}$ Instantaneous wind speed.

U Average wind speed.

u Wind speed fluctuation defined as $\tilde{U} - U$.

u_* Surface friction velocity.

w_s Gravitational settling velocity of a heavy particle.

δ Deviation of time-averaged mean about the ensemble mean.

ϵ Rate of dissipation of turbulent kinetic energy.

ρ Density of air.

σ_c Standard deviation of concentration fluctuations.

σ_v Standard deviation of crosswind velocity component.

σ_w Standard deviation of vertical velocity component.

σ_y Lateral dimension (or spread) for Gaussian plume model.

σ_z Vertical dimension (or spread) for Gaussian plume model.

τ Surface shear stress.

τ_{Leff} Effective Lagrangian timescale of a heavy particle.

τ_{particle} Relaxation timescale of a heavy particle.

Introduction

Turbulent diffusion or, more correctly, turbulent dispersion addresses the problem of estimating the concentration field in a turbulent flow. Our ignorance of the details of the turbulent velocity field translates into uncertainty in estimating the details of the concentration field. Thus, as in other problems related to turbulence, the goal of turbulent dispersion is limited to understanding the ensemble-averaged statistics of the concentration field and the associated fluxes that govern the field.

Most of our understanding of turbulent dispersion is couched in terms of semiempirical models that have been developed by fitting tentative theories to observations. For example, models for dispersion in the surface boundary layer are based on tracer experiments conducted in Prairie Grass, Nebraska, USA, in the 1950s. The development of models for dispersion in the convective boundary layer (CBL) has been guided by laboratory experiments conducted by Willis and Deardorff in the 1970s. Thus, to a large extent, our understanding of turbulent dispersion consists of a patchwork of semiempirical models, each of which describes a limited set of observations. The primary goal of this

article is to provide the reader an understanding of these semi-empirical models. It is only over the past 10 years that attempts have been made to develop methods that can be applied to a large class of problems. These methods will be discussed in the last part of this article.

This article assumes that the reader is familiar with the fundamentals of turbulence in the atmospheric boundary layer. To keep this article to a manageable size, we will not deal with several important topics including the effects of turbulence on chemical reactions. We begin with the statement of the problem that the techniques of turbulent dispersion attempt to solve.

The Problem of Turbulent Dispersion

The evolution of the concentration field of a species is governed by the mass conservation equation, eqn [1], where the 'squiggle' overbar refers to the instantaneous field, and D is the molecular diffusivity of the chemical species in the fluid.

$$\frac{\partial \tilde{C}}{\partial t} + \frac{\partial}{\partial x_i} (\tilde{U}_i \tilde{C}) = D \frac{\partial^2 \tilde{C}}{\partial x_i \partial x_i} \quad [1]$$

Because the instantaneous velocity field is unknown, we have to limit ourselves to estimating the ensemble-averaged mean and the associated statistics of the concentration field.

To define an ensemble, we express the instantaneous velocity field as the sum of a known velocity field, U_i , and the unknown deviation, u_i , as in eqn [2].

$$\tilde{U}_i = U_i + u_i \quad [2]$$

Then, an ensemble is defined as the infinite set of possible concentrations corresponding to a given U_i . Note that the definition of an ensemble is arbitrary because it depends on what we know about the velocity field. Each member of this ensemble corresponds to the component of the unknown velocity field, u_i ; we assume that we know something about the statistics of this unknown field. The concentration associated with each member of the ensemble can be written as eqn [3], where C is the average obtained over all possible concentration fields in the ensemble.

$$\tilde{C} = C + c \quad [3]$$

Substituting eqn [3] into eqn [1] and averaging over the ensemble yield eqn [4], where the overbar refers to the ensemble average, the subscripted index refers to coordinate direction, and repeated indices imply summation. This equation can be solved only by modeling the turbulent flux term, $\overline{u_i c}$, using known properties of the velocity field. Before we take up models for this term, let us consider the problem in turbulent dispersion that has to a large extent motivated the development of the field.

$$\frac{\partial C}{\partial t} + \frac{\partial}{\partial x_i} (U_i C + \overline{u_i c}) = D \frac{\partial^2 C}{\partial x_i \partial x_i} \quad [4]$$

The Point Source in the Atmospheric Boundary Layer

The classical problem of turbulent dispersion in the atmospheric boundary is that of a continuous source emitting material at some height above the ground (see Figure 1). If the

ground is taken to be the reference height, with the x -axis of the coordinate system aligned along the wind direction at the source, empirical evidence indicates that the time-averaged (typically 1 h) concentration field can be described in terms of the Gaussian distribution by eqn [5], where y is the cross-wind coordinate, Q is the source strength (mass/time), h_s is the height of the source above ground, U is the time-averaged wind speed at source height, and σ_y and σ_z are the plume spreads corresponding to the Gaussian distribution.

$$C(x, y, z) = \frac{Q}{2\pi\sigma_y\sigma_z U} \exp \left[-\frac{(z - h_s)^2}{2\sigma_z^2} - \frac{y^2}{2\sigma_y^2} \right] \quad [5]$$

The effect of the ground on concentrations is accounted for by making sure that there is no flux of material through the plane at $z = 0$. The mathematical trick to achieve this is to place an 'image' source at a distance $z = -h_s$; the upward flux from this image source essentially cancels out the downward flux from the real source. Then, the concentration can be described by eqn [6].

$$C(x, y, z) = \frac{Q}{2\pi\sigma_y\sigma_z U} \exp \left[-\frac{y^2}{2\sigma_y^2} \right] \times \left\{ \exp \left[-\frac{(z - h_s)^2}{2\sigma_z^2} \right] + \exp \left[-\frac{(z + h_s)^2}{2\sigma_z^2} \right] \right\} \quad [6]$$

In the real atmosphere, dispersion in the upward direction is limited by the height of the atmospheric boundary layer. This limitation of vertical mixing is incorporated into the Gaussian formulation by reflecting off the top of the mixed layer. When the material is 'reflected' from both the ground as well as the top of the mixed layer, it is necessary to account for the infinite set of 'reflections' from the two surfaces. This can be readily accounted for in the Gaussian formulation.

The point source solution is the kernel for the integral that is used to estimate dispersion from a variety of sources, including line and volume sources. Most dispersion models that apply to spatial scales of tens of kilometers are based on eqns [5] and [6]. The plume spreads or σ values are empirically derived from observations. It turns out that we continue to use plume-spread formulations first recommended by Pasquill in the 1960s. These formulations, largely based on the Prairie Grass experiment, relate plume spread to surface meteorological measurements such as wind speed and cloud cover. The advances in micrometeorology during the 1970s provided the incentive to develop dispersion models that relied on theoretical understanding of dispersion.

The preceding expression for the concentration field is essentially an empirical description of observations. Equation [5] is a formal solution to the point source problem only when the turbulence field is homogeneous, and the velocity distribution is Gaussian. Under these circumstances, the concentration field can be analyzed using a statistical approach, first proposed by Taylor.

Statistical Analysis of Dispersion

Consider a source located at ' s ' emitting particles continuously into a turbulent flow. If the mean flow and turbulence are

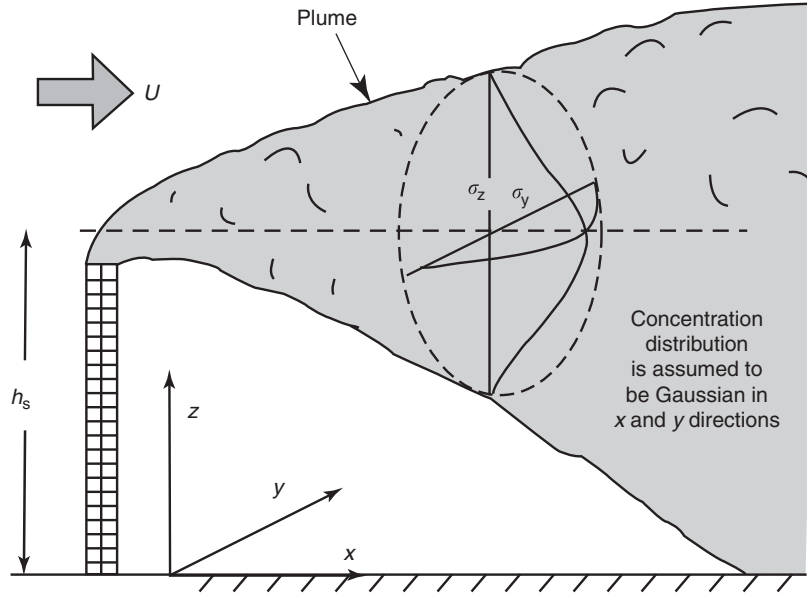


Figure 1 Turbulent dispersion from a point source.

steady, the ensemble-averaged concentration at ' r ,' $C(\mathbf{r})$ can be shown to be

$$C(\mathbf{r}) = Q \int_0^{\infty} p(\mathbf{r}|\mathbf{s}, t) dt \quad [7]$$

where Q is the mass emission rate of particles, and $p(\mathbf{r}|\mathbf{s}, t)dV$ is the probability that a particle released at ' s ' will be found in a volume dV surrounding ' r ' after a travel time t from release. Then, the problem of calculating the concentration reduces to estimating the probability density function, $p(\mathbf{r}|\mathbf{s}, t)$, of particle positions as a function of travel time from the source. A good approximation for this function is the Gaussian distribution. Placing the origin of our coordinate system at the source, we can express the distribution as

$$p(\mathbf{r}|\mathbf{s}, t) = \frac{1}{(2\pi)^{3/2} \sigma_x \sigma_y \sigma_z} \times \exp \left(-\frac{1}{2} \left(\frac{(x_r - \bar{x})^2}{\sigma_x^2} + \frac{(y_r - \bar{y})^2}{\sigma_y^2} + \frac{(z_r - \bar{z})^2}{\sigma_z^2} \right) \right) \quad [8]$$

where σ_x , σ_y , and σ_z are the standard deviations of particle positions about their mean positions $\bar{x}$, $\bar{y}$, and $\bar{z}$ after a travel time, t , from release, and x_r , y_r , and z_r are the receptor coordinates. These statistics are derived by averaging over an infinite number of particles for a flow with fixed mean and turbulent characteristics; the statistics are functions of travel time, t .

The probability distribution function, eqn [8], represents an ensemble average over all possible particle positions for a fixed travel time from the source; the travel time is the difference between the arbitrary time at which the particle is released and time at which the particle is observed at a location. In principle, it can be constructed experimentally by releasing particles serially from a source, and recording the coordinates of these particles at specified travel times. Thus, eqn [8] does not

describe the distribution of particle positions within a 'puff,' which usually describes an entity at an instant of time.

To make progress, we need expressions for σ_x , σ_y , and σ_z . While the mean particle positions are determined by the mean flow, the standard deviations depend on the characteristics of the turbulent flow. Taylor derived expressions for the variance of particle positions as a function of travel time from a fixed release point in a steady flow in which turbulent statistics do not depend on location. His expressions for the asymptotic behavior of plume spread are

$$\begin{aligned} \sigma_y &= \sigma_v t \quad \text{for } t \ll T_{Lv} \\ \sigma_y &= \sigma_v (2tT_{Lv})^{1/2} \quad \text{for } t \gg T_{Lv} \end{aligned} \quad [9]$$

where T_{Lv} is the Lagrangian timescale, which can formally defined in terms of the statistics of the turbulent flow. For our purposes, it is sufficient to interpret the timescale as roughly the time over which a particle retains its initial velocity. For small travel times, a particle's velocity remains essentially unchanged from its value at the release point, and the particle trajectory is a straight line. This explains the result that, for small travel times, the spread of particles is proportional to the travel time from the source (eqn [9]). On the other hand, when the travel time is large compared to the Lagrangian timescale, the plume spread is proportional to the product of the 'average' step size, $\sigma_v T_{Lv}$, and the square root of the number of steps, t/T_{Lv} , taken by the particle.

In order to obtain an expression for the concentration, we still have to integrate eqn [7] after inserting eqn [8] with appropriate expressions for plume spread. Let us first consider an idealized flow that is used to model dispersion in the atmospheric boundary layer. In this flow, the mean wind U is along the x -axis, and the turbulence is homogeneous and stationary. These assumptions lead to

$$\bar{x} = Ut, \quad \bar{y} = 0, \quad \text{and} \quad \bar{z} = 0 \quad [10]$$

If we make the assumption that along-wind dispersion, σ_x , is small compared to transport by the mean wind, the

exponential term in eqn [8], associated with downwind dispersion, becomes a Dirac delta function in the limit of σ_x going to zero. This allows us to integrate eqn [7] for arbitrary σ_y and σ_z to obtain

$$C(\mathbf{r}) = \frac{Q}{2\pi\sigma_y\sigma_zU} \exp\left(-\frac{1}{2}\left(\frac{y^2}{\sigma_y^2} + \frac{z^2}{\sigma_z^2}\right)\right) \quad [11]$$

where the plume spreads are evaluated at $t = x/U$. This equation is identical to the empirical expression presented earlier. We can obtain an expression for the concentration even when downwind dispersion is not small if we can express plume spreads in terms of the asymptotic limits of eqn [9].

Observations of plume spread from elevated releases are often summarized in the form

$$\sigma_y = \frac{\sigma_{yt}}{(1 + t/2T_{Lv})^{1/2}} \quad [12]$$

to ensure consistency with theory of eqn [9]. For ground-level releases, there is no simple way of relating travel time to distance because the velocity varies rapidly with height near the ground. Under these circumstances, we need to use methods that are discussed next.

Dispersion in an Inhomogeneous Boundary Layer

The theory presented thus far applies to a boundary layer in which the mean and turbulent properties are constant in space and time. To apply it to a real boundary layer in which the properties are highly inhomogeneous, we can use one of two approaches. The first is to average the turbulence and mean properties over the region of interest, and use the average properties in the (homogeneous) formulations discussed earlier. This is the most straightforward approach, except that the averaging procedure is necessarily arbitrary. The validity of the method needs to be established by comparing the results obtained from the formulations with observations or theory that accounts for inhomogeneity more explicitly. In general, empirical knowledge derived from observations plays a major role in the development of practical models of dispersion. As in most turbulence research, theory can suggest plausible forms for a dispersion model, but the model almost always contains parameters that have to be estimated from observations.

Even if we could treat the boundary layer as vertically homogeneous, the presence of boundaries, such as the ground and the top of the mixed layer, makes it difficult to estimate the Lagrangian timescale, T_{Lv} , from a priori considerations. Thus, the timescale is often treated as an empirical parameter that is derived by fitting eqn [12] to observations. Alternatively, we can postulate an expression for T_{Lv} in terms of a known length scale l as shown in eqn [13].

$$T_{Lv} = \alpha l / \sigma_v \quad [13]$$

The parameter α has to be obtained by fitting estimates of plume spread to observations. In unstable conditions, l , usually scales with the depth of the boundary layer, while in stable conditions, the relevant length scale is taken to be σ_w/N , where N is the Brunt–Vaisala frequency.

The second approach to accounting for inhomogeneity in the boundary layer is based on the solution of the species

conservation equation. Let us examine this approach in some detail because it yields useful results for dispersion in the surface boundary layer.

Solving the Species Conservation Equation

The species conservation equation (eqn [4]) can be rewritten as eqn [14].

$$\frac{\partial C}{\partial t} + \frac{\partial}{\partial x_i}(U_i C) = \frac{\partial}{\partial x_i}(\overline{u_i c}) + D \frac{\partial^2 C}{\partial x_i \partial x_i} \quad [14]$$

One way of modeling the turbulent flux term is to postulate the concept of eddy diffusivity. It is based on an analogy with molecular diffusion, in which the flux of material in any direction is proportional to the gradient of the concentration. For example, the turbulent flux of species is according to eqn [15], where K_i is the so-called eddy diffusivity.

$$\text{turbulent flux} = \overline{u_i c} = -K_i \frac{\partial C}{\partial x_i} \quad [15]$$

where the bar over i negates the summation convention. This relationship cannot be justified rigorously for turbulent transport. However, it has heuristic value, and is useful for developing semiempirical models of turbulent transport. The use of the eddy diffusivity in the mass conservation equation is often referred to as K -theory.

With eqn [15], eqn [14] can be rewritten in the form of eqn [16],

$$\frac{\partial C}{\partial t} + \frac{\partial}{\partial x_i}(U_i C) = \frac{\partial}{\partial x_i}\left(K_i \frac{\partial C}{\partial x_i}\right) \quad [16]$$

where we have neglected the molecular diffusion term in comparison to turbulent diffusion. While molecular diffusion can often be ignored in calculating the ensemble mean, it plays a major role in determining the statistics of concentration fluctuations, as we will see later.

One way of checking whether the use of the eddy diffusivity is plausible is to see whether eqn [16] yields solutions that are compatible with observations. We will apply eqn [16] to the point source problem. If we assume that transport along the mean wind dominates over the corresponding diffusion term, and turbulent properties are homogeneous, eqn [16] can be reduced to the form given by eqn [17].

$$U \frac{\partial C}{\partial x} = K_z \frac{\partial^2 C}{\partial z^2} + K_y \frac{\partial^2 C}{\partial y^2} \quad [17]$$

It turns out that eqn [17] yields the empirical Gaussian solution of eqn [5] if we ensure that the eddy diffusivity is related to the plume spread according to eqn [18].

$$K_i = \frac{1}{2} U \frac{d\sigma_i^2}{dx_i} \quad [18]$$

If we take plume spread to follow the behavior described in eqn [9], we see from eqn [18] that the eddy diffusivity is proportional to the travel time, x/U , from the source, for travel times less than the governing Lagrangian timescale. What this means is that the eddy diffusivities corresponding to two different sources displaced along the wind will have different values at the same location. It is only at large travel times that

eddy diffusivities become independent of travel time (eqn [19]), where T_{Lv} and T_{Lw} are the Lagrangian timescales for the horizontal and vertical velocity fluctuations, respectively.

$$K_y = \sigma_v^2 T_{Lv} \quad \text{and} \quad K_z = \sigma_w^2 T_{Lw} \quad [19]$$

The eddy diffusivity, K_z , can be related to turbulent flow properties by appealing to 'mixing length' theory, which suggests the relationship [20], where σ_w is the standard deviation of the vertical velocity fluctuations, and l_z is the 'length scale' of turbulence for vertical transport, defined by eqn [21], which is consistent with eqn [13].

$$K_z = \sigma_w l_z \quad [20]$$

$$l_z = \sigma_w T_{Lw} \quad [21]$$

We are now in a position to make some additional statements on the applicability of the eddy diffusivity concept. We saw earlier that eqn [19] is valid when the travel time is much larger than the Lagrangian timescale, expressed by the relationship [22].

$$t = \frac{x}{U} \gg T_{Lw} \quad [22]$$

If we combine this condition with the expression for plume spread, eqn [9], and use eqn [21], we obtain eqn [23].

$$\sigma_z \gg l_z \quad [23]$$

Thus, the eddy diffusivity concept is most applicable when the scale of concentration variation, σ_z , is much larger than the scale of the eddies responsible for plume spreading.

Equation [20] is useful because we can guess at the appropriate form of l_z , and then see whether the consequences of our assumption agree with observations. Over the years, we have developed enough experience with different types of flows to prescribe useful forms for the mixing length (or eddy diffusivity) for these flows. Our initial guesses for K are usually based on measurement of fluxes and the associated gradients for a limited set of situations. This K is then extrapolated to situations different from those used to derive it. For example, we can derive a K for heat flux, and find out whether it works for pollutant transport. It is this type of semiempirical arguments that form the basis of practical calculations for turbulent flows.

Equation [15] represents only one possible approach to expressing the turbulent flux. In principle, we can write conservation equations for the turbulent fluxes, but these equations contain 'third-order' terms that are essentially the fluxes of the second-order terms. These third-order terms have to be parameterized using some sort of flux-gradient approximation. At this point, there is no compelling evidence to suggest that these approaches yield much better results than the closure of eqn [15]. One way of improving upon a simple prescription of the eddy diffusivity is to formulate semiempirical conservation equations for the components of eddy diffusivities: the turbulent velocity and the length scale in eqn [20]. In practice, the turbulent velocity is related to the turbulent kinetic energy, k , and the length scale is related to the turbulent dissipation rate, ε . While the $k - \varepsilon$ approach is popular in modeling turbulent flows, it has found limited application in modeling dispersion.

The eddy diffusivity formulation is almost exclusively used in comprehensive Eulerian air quality models that include details of atmospheric process such as gas- and aqueous-phase chemistry. The main reason is that the species conservation equation, formulated in terms of the eddy diffusivity, is a convenient framework to incorporate a number of processes, including nonlinear chemistry. The resulting mass conservation can be solved using numerical methods. Results from comprehensive air quality models indicate that modeling dispersion with the eddy diffusivity model has some practical value, even though the underlying theoretical justification is weak.

We saw earlier that the eddy diffusivity concept is likely to be most applicable when the turbulent length scales are smaller than or comparable to the concentration space scales. Thus, we might expect it to apply to dispersion in the surface boundary layer, where plume spread in the vertical spread is comparable to the length scale of the eddies responsible for vertical transport. It turns out that K -theory provides useful results for dispersion in the surface boundary layer even though it is characterized by steep gradients in temperature and velocity. However, the gradients of fluxes and turbulence levels are negligible. In the surface boundary layer, semiempirical theories, referred to as Monin-Obukhov similarity, provide useful relationships between velocity and temperature gradients and the corresponding heat and momentum fluxes. These relationships are cast in terms of length and velocity scales, which are the surface friction velocity u_* and the Monin-Obukhov length, L , defined by eqns [24],

$$u_* = \sqrt{\frac{\tau}{\rho}} \quad [24]$$

$$L = -\frac{T_0 u_*^3 \rho C_p}{g \kappa H_0}$$

where τ is the surface shear stress, ρ is the air density, C_p is the specific heat of air, T_0 is the surface temperature, κ is the von Karman constant, g is the acceleration due to gravity, and H_0 is the surface heat flux. These relationships can be used to derive eddy diffusivities for heat and momentum.

Using the eddy diffusivity for heat in the mass conservation equation has provided concentration estimates that compare well with observations made in field experiments conducted in Prairie Grass, Nebraska, in the 1950s. We note that data from this experiment, conducted with relatively primitive equipment, are still the most complete for the analysis of surface layer dispersion.

The solutions of the mass conservation equations, using the eddy diffusivity, have a number of useful asymptotic forms for the crosswind-integrated concentration, $\bar{C}'$, as shown by eqn [25], where $\bar{x} = x/|L|$, $\bar{C} = \bar{C}' u_* |L| / Q$, with Q representing the source strength. These asymptotic forms are useful because they can be patched together to obtain analytic expressions that span the entire range of stability. Unknown parameters in these expressions have been obtained by fitting them to observations from Prairie Grass.

$$\begin{aligned} \bar{C} &\sim \bar{x}^{-1} \text{ for neutral conditions} \\ &\sim \bar{x}^{-2/3} \text{ for stable conditions} \\ &\sim \bar{x}^{-2} \text{ for unstable conditions} \end{aligned} \quad [25]$$

These expressions for crosswind-integrated concentrations can be converted to yield centerline concentrations through

eqn [26], where σ_y is the crosswind spread, and the crosswind distribution is taken to be Gaussian.

$$C(x, y, 0) = \frac{\bar{C}'}{\sqrt{2\pi}\sigma_y} \exp\left(-\frac{y^2}{2\sigma_y^2}\right) \quad [26]$$

Equation [25] can be used to derive expressions for the vertical plume spread. These expressions depend on distance from the release location because travel time has little meaning near the ground. In fact, most dispersion models used in practical applications express the concentration in terms of a Gaussian distribution, where the plume spreads are empirically derived functions of source–receptor distance and micrometeorology.

Puff Dispersion

In the preceding sections, we have discussed dispersion of plumes, which refers to a continuous release from a source. Often we are interested in concentrations associated with puffs of material that are released almost instantaneously, as in an explosion. The concentration of material in the plume is determined by the spread of the material about the center of mass of the moving puff. The analysis of such puffs is more complicated than that of plumes because the spread depends on both space and time correlations between particles in the puff; as Taylor's analysis indicates, plume dispersion can consider the motion of particles to be independent of each other. These space–time correlations depend on the properties of the turbulent eddies that contribute to puff spread at any instant of time. The length scale of the relevant eddies is roughly proportional to the size of the puff. Eddies smaller than this length scale disperse the material within the puff, while eddies larger than the puff transport the puff as a whole.

When the puff size is of the order of the Kolmogorov microscale, the puff spreads by molecular diffusion, which implies that the puff spread, s , is proportional to the square root of time. When the puff size is comparable to eddies in the inertial subrange, dimensional considerations suggest that the rate of puff spread is represented by eqn [27], where ε is the dissipation rate of turbulent kinetic energy.

$$\frac{ds}{dt} \sim (\varepsilon s)^{1/3} \quad [27]$$

Integration yields expression [28].

$$s \sim \varepsilon^{1/2} t^{3/2} \quad [28]$$

This rapid growth phase ends when the puff size is comparable to the largest eddy of dimension L . Then, the puff spread can be written as [29], where σ_v refers to the standard deviation of the turbulent velocity fluctuations in the direction of the spread.

$$s \sim (\sigma_v L)^{1/2} t^{1/2} \quad [29]$$

These three regimes of puff growth can, in principle, be patched together to provide a continuous description of puff spread. But this is rarely done in practice, and one usually resorts to empirical descriptions of puff spread. The actual concentration in a puff is usually estimated with a Gaussian distribution about the puff center of mass.

The simplest puff model is the Gaussian puff model that relates concentration at receptor (x, y, z) at time t due to a puff released from origin at time 0 by eqn [30], where Q_m is the total mass of the puff and s is the puff spread corresponding to the Gaussian distribution.

$$C(x, y, z, t) = \frac{Q_m}{(2\pi)^{3/2} s^3} \exp\left[-\frac{(x - Ut)^2 + y^2 + z^2}{2s^2}\right] \quad [30]$$

In practical applications, a continuous release in a wind field that varies in space and time can be modeled through a series of puffs, each of which is allowed to follow a different trajectory. The concentration at a receptor at any given time is calculated by summing the contributions from these puffs. One advantage of this approach is that it can deal with situations when the mean wind is calm.

Dispersion of Heavy Particles

The previous discussions implicitly assumed that the material being dispersed by turbulence has the same density as air. This assumption is clearly not valid for aerosol particles, whose densities are typically over 1 g cm^{-3} . Two effects influence dispersion of such particles. One is related to the finite time required by the particle to respond to turbulent velocity fluctuations. The other is the so-called trajectory crossing effect related to particle trajectories being different from fluid parcel trajectories because of gravitational settling. Let us consider each of these effects.

The particle inertia effect is related to the difference between the fluid velocity, u_f , and the particle velocity, u_p , whose difference is proportional to the reaction time to turbulent velocity fluctuations. This can be expressed symbolically by eqn [31], where the relaxation timescale of the particle is given by eqn [32].

$$\frac{(u_f - u_p)}{u_f} \sim \frac{\tau_{\text{particle}}}{\tau_{\text{turbulence}}} \quad [31]$$

$$\tau_{\text{particle}} \sim \frac{w_s}{g} \quad [32]$$

Here, w_s is the gravitational settling velocity, which is a function of the size and density of the particle and the viscosity of the fluid. The turbulence timescale can be expressed as [33], where l is a measure of eddy size and σ_w is the associated velocity fluctuation.

$$\tau_{\text{turbulence}} \sim \frac{l}{\sigma_w} \quad [33]$$

Thus, inertia effects can be neglected if the ratio of these two scales is small (see eqn [34]).

$$\frac{w_s \sigma_w}{gl} \ll 1 \quad [34]$$

If we take small to mean 0.1, and consider $100 \text{ }\mu\text{m}$ particles with settling velocities of the order of 1 m s^{-1} , and take the turbulent velocity to be 1 m s^{-1} , the inertia effect is not likely to be important for length scales greater than 1 m . However, it could play a role in the dispersion of large particles under very stable conditions, close to the ground.

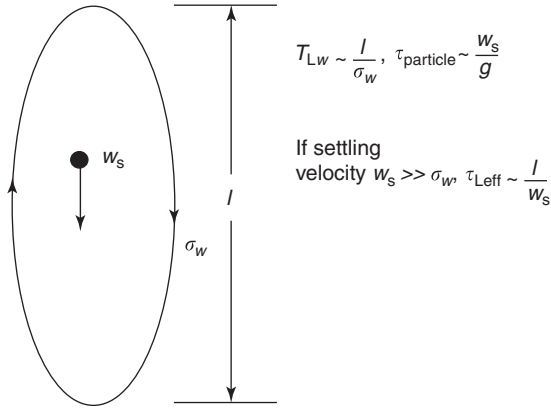


Figure 2 Timescales that affect dispersion of heavy particles.

The trajectory crossing effect can be examined by considering the extreme case when the time taken for a falling particle to traverse an eddy is much smaller than the Lagrangian timescale for vertical dispersion. Then, the effective Lagrangian timescale becomes this traversal time because it corresponds roughly to the decorrelation time. The effective Lagrangian timescale and the associated eddy diffusivity for the heavy particle can now be represented by eqn [35].

$$\tau_{\text{Leff}} = \frac{l}{w_s} \quad \text{and} \quad K_{\text{eff}} = \frac{\sigma_w^2 l}{w_s} \quad [35]$$

Figure 2 shows the timescales that govern dispersion of heavy particles. In practice, the effects of particle settling are not important in determining plume spread because settling velocities for most particles are generally much smaller than turbulent velocities. However, even particle velocities of the order of a few centimeters per second lead to mean downward motion of the plume, and hence increase concentrations at ground level. This effect can be described approximately by ‘tilting’ the plume toward the ground, represented by eqn [36].

$$h_s(\text{with particles}) = h_s - \frac{w_s x}{U} \quad [36]$$

Alternatively, the entire concentration profile can be moved ‘into’ the ground by a distance $w_s x/U$ after computing the concentration corresponding to passive dispersion. It is simple to account for mean motion of particles in the eddy diffusivity formulation through the advection term, $w_s(\partial C/\partial z)$.

Other Approaches to Modeling Dispersion

The eddy diffusivity approach does not generally apply to sources far removed from the ground. For example, it is difficult to justify its application to the convective atmospheric layer, where the turbulent length scales are large compared to the spatial scales of the concentration field. The pdf approach can provide useful results under these circumstances. Studies show that dispersion in the CBL is strongly influenced by the relative longevity of convective downdrafts and updrafts. The majority of particles released in downdrafts travel continuously downward until they

reach the vicinity of the ground. Numerical experiments indicate that to a useful degree of approximation, the velocity of these particles can be taken to be constant at the value at the release point.

This assumption allows one to express the crosswind-integrated ground-level concentration, $\bar{C}^y$, in terms of the pdf of the vertical velocity, $p(w|h_s)$, at the height of release, h_s , by eqn [37], where the vertical velocity corresponds to that required to bring material from the release point to the ground-level receptor at x , given by [38].

$$\bar{C}^y = \frac{2Q}{x} P(w|h_s) \quad [37]$$

$$w = -Uh_s/x \quad [38]$$

The simple formulation, which can be readily modified to account for the presence of the mixed layer, provides an excellent description of laboratory observations of dispersion in the CBL. Note that eqn [37] reproduces the empirical Gaussian distribution if we make the reasonable assumption that the pdf of vertical velocity fluctuations is normal, giving eqn [39], where σ_z is expressed by eqn [40].

$$\bar{C}^y = \sqrt{\frac{2}{\pi}} \frac{Q}{U\sigma_z} \exp\left[-\frac{h_s^2}{2\sigma_z^2}\right] \quad [39]$$

$$\sigma_z = \sigma_w x/U \quad [40]$$

The Gaussian formulation is not reliable in the CBL because the pdf is positively skewed. The associated negative mode of the pdf leads to the descent of the plume centerline when the release is elevated, and leads to concentrations that are about 30% higher than that predicted with the Gaussian formulation.

In principle, if we could simulate all the scales of turbulent motion, there would be no need for models of turbulent dispersion. We could use the Navier–Stokes equations to generate an ensemble of flows, obtain the corresponding concentration fields from the species conservation equation, and average over them to obtain the ensemble average as well as the statistics of concentration fluctuations. With the rapid increases in computing power, direct numerical simulation (DNS) is becoming a reality, and we have been able to obtain useful information for low Reynolds number flows. However, it will be some time in the future before we will be able to use the technique for routine applications.

The large eddy simulation (LES) technique avoids the computational demands of DNS by only simulating the energy-containing eddies. The effects of the unresolved scales of motion are modeled using a variety of parameterizations. It is believed that the most important features of flow are insensitive to these parameterizations, because the subgrid scales contain a small fraction of the total energy. This assumption has been vindicated by LES of CBLs that was pioneered by Deardorff in the 1970s. Velocity fields generated by LESs compare well with observations, and continue to provide information that is difficult to obtain in the field. Lamb used the velocity fields generated by Deardorff to simulate dispersion in the CBL. The simulation consisted of releasing a large number of particles and tracing their motion using the LES velocity field. Then, the crosswind concentration averaged over

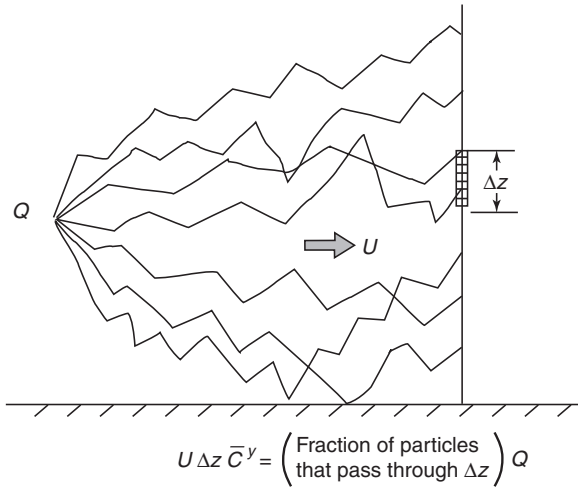


Figure 3 Calculation of concentrations using Lagrangian stochastic simulation of particle trajectories.

a vertical distance Δz is given by eqn [41], where f is the fraction of the particles released that pass through Δz .

$$\bar{C}^y = \frac{Q}{U \Delta z} f \quad [41]$$

Figure 3 justifies this equation.

Lamb's simulations provided valuable insight into dispersion in the CBL, including the observation that the locus of the maximum concentration descended toward the ground. This behavior is related to the negative mode of the pdf of the vertical velocity fluctuations.

A technique that is not as computationally demanding as DNS or LES is called Lagrangian stochastic simulation (LSS). It is attractive because it only uses the statistics of the velocity field such as velocity variance, and dissipation rate of turbulent kinetic energy. Because this technique is being used to examine routine dispersion problems, we discuss it in some detail in the next section.

Lagrangian Stochastic Models

In Lagrangian stochastic modeling, turbulent dispersion is modeled by tracing the motion of a large number of fluid particles that are tagged at the source; these particles are treated mathematically as points. The evolution of the velocity of each particle depends on turbulence properties at the current location of the particle. To illustrate the basic ideas, let us trace a particle that is only affected by vertical velocity fluctuations. Then, the vertical velocity of a parcel at time $t + \Delta t$ is related to the velocity at time t through eqn [42], where w' is a random velocity drawn from the distribution of vertical velocity fluctuations.

$$w(t + \Delta t) = \alpha w(t) + \beta w' \quad [42]$$

Thus, the future velocity of the particle consists of a deterministic component that depends on its current velocity, and a random component that depends on the turbulence at the location of the particle. We can show that

eqn [42] is equivalent to the Langevin equation in which a particle is subject to a linear viscous force and a random pressure force.

The particle position is traced through eqns [43].

$$\begin{aligned} \Delta z &= w(t + \Delta t) \Delta t \\ \Delta x &= U \Delta t \end{aligned} \quad [43]$$

In homogeneous turbulence, we can show that α and β are given by eqns [44],

$$\alpha = \left(1 - \frac{\Delta t}{T_{Lw}}\right) \quad \text{and} \quad \beta = \left(\frac{2 \Delta t}{T_{Lw}}\right) \quad [44]$$

where the Lagrangian timescale, T_{Lw} , can be related to the turbulent dissipation rate according to eqn [45], where C_0 is a constant.

$$T_{Lw} = \frac{2 \sigma_w^2}{C_0 \varepsilon} \quad [45]$$

While this simple model has produced useful results, it is not applicable to boundary layers with gradients in turbulence properties. New developments in LSS are best described by recasting eqn [42] as eqn [46], where a is given by eqn [47], and the $d\zeta$ is a normal random variable with zero mean and variance dt .

$$dw = aw \, dt + (C_0 \varepsilon)^{1/2} d\zeta \quad [46]$$

$$a = \frac{C_0 \varepsilon}{2 \sigma_w^2} \quad [47]$$

During the 1980s, several ad hoc formulations for the term ' a ' in eqn [46] were proposed to account for flow complexities such as inhomogeneity and non-Gaussianity of the turbulent velocities. This unsatisfactory situation was remedied by Thomson in 1987 when he proposed a systematic method for constructing formulations for ' a ' by invoking the constraint that the model for this term should preserve a well-mixed concentration field. This is equivalent to insisting that the Lagrangian pdf for fluid particles marked at the source should become identical to that of unmarked fluid particles at large travel times. The evolution of the pdf is governed by the Fokker-Planck equation corresponding to eqn [46]. Using the solution of this equation, Thomson was able to derive formulations for the term ' a ' that accounted for inhomogeneity and non-Gaussianity of the turbulent velocities. For example, the modified form of eqn [46] can be represented by eqn [48].

$$dw = -\frac{C_0 \varepsilon}{2 \sigma_w^2} w \, dt + \frac{1}{2} \left(1 + \frac{w^2}{\sigma_w^2}\right) \frac{\partial \sigma_w^2}{\partial z} dt + (C_0 \varepsilon)^{1/2} d\zeta \quad [48]$$

It turns out that ' a ' can be expressed uniquely only for one-dimensional turbulence or isotropic turbulence; for two- and three-dimensional turbulent flows, ' a ' is not a unique function of the velocity field. This nonuniqueness problem can be serious because for a given turbulent flow, substantially different concentration fields can be obtained from two models, both of which satisfy the well-mixed condition.

One way of alleviating this problem is to drive Lagrangian stochastic models with velocity fields obtained from large eddy simulation because large eddies that contribute to inhomogeneity and anisotropy of the velocity field are treated explicitly in both LES and LSS. The subgrid-scale eddies can be considered

locally homogeneous and locally isotropic. Because only the particle motion associated with subgrid energy is modeled with LSS, errors associated with the nonuniqueness of the LSS model can be minimized.

Concentration Fluctuations and Model Uncertainty

Any given observed concentration will deviate from the ensemble average. This deviation is caused by the intrinsic variability of the concentration field called concentration fluctuations. In principle, the statistics of these concentration fluctuations can provide insight into the expected deviation of the model-predicted ensemble mean from a corresponding observed concentration.

To understand the effect of concentration fluctuations, consider the following model of the concentration time series in which the concentration is a peak value, C_{peak} , or zero. Then, we can show that the variance of the instantaneous concentration about the ensemble-averaged mean is given by eqn [49].

$$\frac{\sigma_c^2}{C^2} = \frac{C_{\text{peak}}}{C} - 1 \quad [49]$$

Because the peak-to-mean concentration ratio can be as large as 10 or even 100, especially close to an elevated source, the standard deviation of the concentration fluctuations can be several times the mean. If we are interested in predicting a time-averaged concentration, we have to estimate the timescale, T_c that governs the concentration fluctuations. This then allows us to estimate the number of independent concentration events that we are likely to observe during the averaging interval T_{av} , as T_{av}/T_c . Then, the deviation, δ , of the time-averaged mean about the ensemble mean is given by eqn [50].

$$\frac{\delta}{C} \sim \left[\frac{T_c}{T_{\text{av}}} \right]^{1/2} \frac{\sigma_c}{C} \quad [50]$$

T_c roughly corresponds to the time taken for the instantaneous plume to pass a fixed observer. If we take the instantaneous plume size to be 100 m and the wind speed to be 5 m s^{-1} , the timescale, T_c , is 20 s. Assuming that we are interested in an averaging time of 1 h, δ works to be about 25% of the mean for a peak-to-mean ratio of 10. This exercise identifies the variables that might govern the deviation between model estimates and observations. Actual comparisons between model estimates and observations indicate that the error is much larger because of model formulation and input errors. A model is generally considered adequate if its estimates are consistently within a factor of two of the observations.

We can use a simple model to discuss the role of molecular diffusion in determining concentration fluctuations. Assume that we release Q_m mass units in an initial volume of V_0 , so that the initial concentration is Q_m/V_0 . As the released material is stretched over a larger volume, V , in space, the volume marked by the material remains unchanged if the fluid is incompressible and molecular diffusion is negligible (see Figure 4). The mean concentration, corresponding to material spread over V , is proportional to $1/V$, while the peak concentration remains unchanged from the initial value. Then, the peak-to-mean ratio, which determines the concentration variance, is simply V/V_0 . In the presence of molecular diffusion, the material is no longer confined to its initial volume, and the peak concentration has to decrease with time. Thus, molecular diffusion decreases concentration fluctuations relative to the mean value.

The meandering plume model, used to estimate the variance of concentration fluctuations, is an extension of the previous concept. Here, the time-averaged plume is assumed to be composed of instantaneous plumes whose dimensions are determined by relative dispersion. Then, the peak concentration is determined by the concentration within the instantaneous plume, which is inversely proportional to s^2 , where s refers to the spread by relative dispersion. Then, if σ is the time-averaged spread of the plume, the peak-to-mean ratio, and the

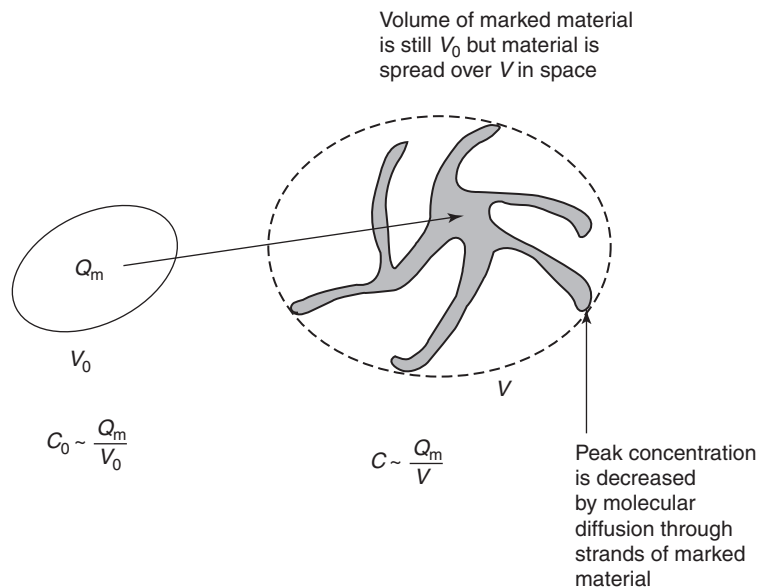


Figure 4 Factors that affect concentration fluctuations. $C_{\text{peak}} \sim C_0$ in the absence of molecular diffusion, and $\sigma_c^2/C^2 \sim C_{\text{peak}}/C \sim V/V_0$.

normalized concentration variance is of the order of σ^2/s^2 as long as we can neglect molecular diffusion within the instantaneous plume. Molecular diffusion will eventually smear the concentration over the instantaneous plume and thus decrease the normalized concentration variance.

The statistics of concentration fluctuations can be used in estimating the air quality impacts of species for which small time exposures are important. These statistics can be used to estimate the probability that a certain threshold is exceeded. Thus, there is great interest in formulating models for concentration fluctuations. Some of the more recent models have combined a version of LSS with LES velocity fields to model relative dispersion and thus concentration fluctuations. In the future, we are likely to see more use of DNS in understanding turbulent dispersion. This does not mean that our ability to predict concentrations will improve substantially, because the nature of turbulence places practical limits on predicting individual realizations of concentrations.

See also: **Aerosols:** Observations and Measurements; Aerosol Physics and Chemistry. **Boundary Layer (Atmospheric) and Air Pollution:** Modeling and Parameterization; Overview. **Aviation Meteorology:** Clear Air Turbulence. **Numerical Models:** Parameterization of Physical Processes: Turbulence and Mixing. **Turbulence and Mixing:** Overview; Turbulence, Two Dimensional.

Further Reading

- Arya, S.P., 1999. *Air Pollution Meteorology and Dispersion*. Oxford University Press, New York.
- Csanady, G.T., 1973. *Turbulent Diffusion in the Environment*. Reidel, Dordrecht, Holland.
- Nieuwstadt, F.T.M., van Dop, H. (Eds.), 1982. *Atmospheric Turbulence and Air Pollution Modeling*. Reidel, Dordrecht, Holland.
- Pasquill, F., Smith, F.B., 1983. *Atmospheric Diffusion*, third ed. Ellis Horwood Limited, John Wiley & Sons, New York.
- Rodean, H.C., 1996. *Stochastic Lagrangian Models of Turbulent Diffusion*. American Meteorological Society, Boston, MA.
- Seinfeld, J.H., Pandis, S., 1999. *Atmospheric Chemistry and Physics of the Atmosphere*. Wiley-Interscience, New York.
- Taylor, G.I., 1921. Diffusion by continuous movements. *Proceedings of the London Mathematical Society* 20, 196–211.
- Tennekes, H., Lumley, J., 1972. *A First Course in Turbulence*. MIT Press, Cambridge, MA.
- Venkatram, A., Wyngaard, J. (Eds.), 1988. *Lectures on Air Pollution Modeling*. American Meteorological Society, Boston, MA.

WEATHER FORECASTING

Contents

Marine Meteorology

Operational Meteorology

Seasonal and Interannual Weather Prediction

Severe Weather Forecasting

Wildfire Weather

Marine Meteorology

L Xie and B Liu, North Carolina State University, Raleigh, NC, USA

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

Marine meteorology is a subfield of meteorology, which deals with the weather and climate as well as the associated ocean conditions in marine, island, and coastal environments. Marine meteorologists collect and analyze weather and oceanographic data, make marine weather forecasts, and provide marine meteorological services to support marine and coastal activities, including but not limited to shipping, fishing, tourism, offshore oil drilling and mining operations, oil spill control, offshore wind and tidal energy harvesting, search and rescue at sea, and naval operations. The interactions between the atmosphere and the ocean create a unique dynamical framework for the fundamental principles of marine meteorology, and the need for employing air–sea coupled models for marine weather prediction.

Introduction

Marine meteorology is a subfield of meteorology, which deals with the weather and climate as well as the associated oceanographic conditions in marine, island, and coastal environments. The physical and dynamical foundations of marine meteorology are no different from other areas of meteorology, but the fundamental processes, which distinguish marine meteorology from other subfields of meteorology are the interactions between the ocean and the atmosphere. Therefore, the part of the physical oceanography dealing with the upper and coastal oceans, which are directly affected and influenced by weather is also considered an integral component of marine meteorology.

Marine meteorology is also unique in that it is a science geared toward the understanding and production of weather information in support of marine and coastal activities, including shipping, fishing, tourism, offshore oil drilling and mining operations, oil spill control, offshore wind and tidal energy harvesting, search and rescue at sea, and naval operations.

The physical processes associated with marine weather encompass not only complex marine environments from the open ocean to the coastal zone and from islands to marginal seas, but also broad dynamic scales. Small-scale phenomena

include turbulence and eddies that occur within the boundary layers on both sides of the air–sea interface. These small-scale features play a major role in determining the air–sea momentum and heat and moisture fluxes that, in turn, affect the atmospheric and oceanic processes at other scales. At meso- and synoptic scales, marine meteorology deals with phenomena such as coastal fronts, coastal low-level jets, land–sea breezes, coastal and sea fogs, coastal cyclones and tropical cyclones (including hurricanes and typhoons), as well as hazardous ocean conditions including storm surge, wind waves, and rip currents. At large and global scales, marine meteorology deals with the weather and climate associated with seasonal, intraseasonal, and interannual variability in maritime environment, such as monsoons, trade wind systems, Madden–Julian oscillation, and El Niño/Southern Oscillation.

In the following, we will present the fundamental principles of marine meteorology and describe the observation, prediction, and application systems of marine weather and climate.

Fundamental Principles

Marine meteorology deals with two forms of geophysical fluids, the atmosphere and the ocean. Therefore, the fundamental physical principles for marine meteorology include the

conservation laws and governing equations for the motions of both geophysical fluids and their interactions. The conservation laws include the conservations of mass, momentum, and energy. The governing equations for atmosphere can be expressed as (Holton and Hakim, 2012):

$$\frac{D\mathbf{U}}{Dt} = -2 \times \boldsymbol{\Omega} \times \mathbf{U} - \frac{1}{\rho} \nabla p + \mathbf{g} + \mathbf{F}_r \quad [1]$$

$$\frac{D\rho}{Dt} + \rho \nabla \cdot \mathbf{U} = 0 \quad [2]$$

$$c_v \frac{DT}{Dt} + p \frac{D\alpha}{Dt} = J \quad [3]$$

$$p = \rho RT \quad [4]$$

where $\mathbf{U}$ is the three-dimensional velocity vector, $\boldsymbol{\Omega}$ is the angular velocity of the earth, ρ is air density, α is specific volume, p is air pressure, $\mathbf{g}$ is gravity (vector variable), $\mathbf{F}_r$ is frictional force, T is air temperature, J is the heating rate due to radiation, conduction, and latent heat release, c_v is the specific heat at constant volume, and R is the gas constant for dry air.

The governing equations for ocean circulation are similar to those for the atmosphere. Ignoring any source/sink terms, the nonhydrostatic primitive equations for the ocean can be written as (Haidvogel and Beckmann, 1999):

$$\frac{D\mathbf{V}}{Dt} = -2 \times \boldsymbol{\Omega} \times \mathbf{V} - \frac{1}{\rho} \nabla p + \mathbf{g} + \nabla \cdot \boldsymbol{\tau} \quad [5]$$

$$\frac{D\rho}{Dt} + \rho \nabla \cdot \mathbf{V} = 0 \quad [6]$$

$$\rho c_p \frac{DT}{Dt} - \beta T \frac{Dp}{Dt} = 0 \quad [7]$$

$$\frac{DS}{Dt} = 0 \quad [8]$$

$$\rho = \rho(T, S, p) \quad [9]$$

Here, $\mathbf{V}$ is the current velocity vector, p is pressure, ρ is water density, T is water temperature, S is salinity, $\mathbf{g}$ is gravity, $\boldsymbol{\tau}$ is the stress tensor, c_p is the specific heat of water at constant pressure, and β is the thermal expansion coefficient $-\frac{1}{\rho} \left(\frac{\partial \rho}{\partial T} \right)_p$.

Sea surface waves are ubiquitous at ocean surface, with air and water on each side of the air-sea interface. The dynamics of surface waves is also determined by the Navier-Stokes equation for a two-layer fluid. It is so complicated that one cannot solve it directly either analytically or numerically. Fortunately, the density of air is much smaller than that of water. Sea surface waves can be described in good approximation by the linearized Navier-Stokes equations for a one-layer fluid in a gravitational field (Komen et al., 1994):

$$\frac{\partial \mathbf{u}}{\partial t} = -\frac{1}{\rho} \nabla p \quad [10]$$

$$\frac{\partial w}{\partial t} = -\frac{1}{\rho} \frac{\partial p}{\partial z} - g \quad [11]$$

$$\nabla \cdot \mathbf{u} + \frac{\partial w}{\partial z} = 0 \quad [12]$$

$$\frac{\partial \eta}{\partial t} = w, z = 0 \quad [13]$$

$$p = p_{\text{atm}}, z = 0 \quad [14]$$

where $\mathbf{u}$ is the horizontal velocity vector, w is the vertical velocity component, and p_{atm} is the air pressure at the sea surface. Assuming wavelike solutions, one can obtain the dispersion relation for sea surface waves:

$$\omega_{\pm} = \pm \sigma, \sigma^2 = gk \tanh kh \quad [15]$$

where ω is the wave frequency, k is wave number, g is the scalar variable, and h is water depth. The general solution can then be expressed as

$$\begin{pmatrix} \eta \\ \mathbf{u} \\ w \\ p + \rho g z \end{pmatrix} = a \begin{pmatrix} 1 \\ \omega_{\pm} \mathbf{k} / k e^{kz} \\ -i \omega_{\pm} e^{kz} \\ \rho g e^{kz} \end{pmatrix} e^{i(\mathbf{k} \cdot \mathbf{x} - \omega_{\pm} t)} + \text{c.c.} \quad [16]$$

in which, a is an arbitrary constant and c.c. is the complex conjugate term. In reality, sea surface waves can be thought of as a superposition of many such linear waves. However, in practice one never considers the deterministic initial value problem for a realistic sea basin, because it is virtually impossible to determine the correct phases for the wave components. Instead, we resort to the homogeneous and stationary statistical theory of linear random waves and introduce wave spectrum $F(\mathbf{k}, \mathbf{x}, t)$ and wave action $N(\mathbf{k}, \mathbf{x}, t) = F(\mathbf{k}, \mathbf{x}, t) / \sigma(\mathbf{k}, \mathbf{x})$, which characterize the statistical properties of sea surface waves, to describe the wave energy and its propagation and dissipation. The wave action balance equation can be written as (Komen et al., 1994):

$$\left\{ \frac{\partial}{\partial t} + (\mathbf{c}_g + \mathbf{U}) \cdot \frac{\partial}{\partial \mathbf{x}} - [\nabla_{\mathbf{x}}(\sigma + \mathbf{k} \cdot \mathbf{U})] \cdot \frac{\partial}{\partial \mathbf{k}} \right\} N = \frac{S_{\text{in}}}{\sigma} + \frac{S_{\text{nl}}}{\sigma} + \frac{S_{\text{ds}}}{\sigma} + \frac{S_{\text{bot}}}{\sigma} \quad [17]$$

where $\mathbf{U}$ is the background current velocity and $\mathbf{c}_g$ is the wave group velocity. The right-hand side terms are the source terms describing wind input (S_{in}), white-capping dissipation (S_{ds}), bottom friction (S_{bot}), and a nonlinear interaction term (S_{nl}). The above equation is typically adopted as the governing equation for various surface wave models. Surface wave characteristics such as significant wave height can be obtained from the wave action (spectrum).

The above equations form the basic governing equations for the atmosphere, ocean circulation, and sea surface waves. However, these equations are not complete without the relationships governing their interactions. Atmospheric circulation, ocean circulation, and sea surface waves interact through various air-sea exchanges across the air-sea interface, and wave-current interactions within the ocean. The atmosphere and the ocean exchange mass, momentum, and energy, as well as other chemical properties primarily through turbulent fluxes, forming an air-sea coupled system. Parameterizations of these turbulent fluxes are usually either through the bulk formula or through the sea surface roughness lengths. The bulk formula for estimating air-sea momentum (τ), sensible heat (H_s), water vapor (E), and latent heat (H_l) fluxes can be expressed as:

$$\tau = \rho_a C_d U^2 \quad [18]$$

$$H_s = \rho_a c_p C_H U (\theta_s - \theta_a) \quad [19]$$

$$E = \rho_a C_E U (q_s - q_a) \quad [20]$$

$$H_l = L_v E = \rho_a L_v C_E U (q_s - q_a) \quad [21]$$

where U is the surface wind speed relative to the ocean surface currents; θ_a , q_a , and θ_s , q_s are potential temperature and specific humidity in air and at sea surface, respectively; ρ_a is air density; c_p is the specific heat of air at constant pressure; L_v is the latent heat of evaporation; C_d , C_H , and C_E are the air-sea transfer coefficients: the drag coefficient, Stanton number, and Dalton number, respectively. Based on various measurements under low to moderate wind conditions, the drag coefficient usually increases with surface wind, and the Stanton and Dalton numbers are generally weakly dependent upon wind speed. According to the Monin-Obukhov similarity theory, under neutral stable conditions, air-sea transfer coefficients can be related to sea surface roughness lengths:

$$C_d = \kappa^2 \left[\ln \left(\frac{z}{z_0} \right) \right]^{-2} \quad [22]$$

$$C_H = \kappa^2 \left[\ln \left(\frac{z}{z_0} \right) \right]^{-1} \left[\ln \left(\frac{z}{z_T} \right) \right]^{-1} \quad [23]$$

$$C_E = \kappa^2 \left[\ln \left(\frac{z}{z_0} \right) \right]^{-1} \left[\ln \left(\frac{z}{z_q} \right) \right]^{-1} \quad [24]$$

where κ is the Karman constant, z_0 is the sea surface aerodynamic roughness length, and z_T and z_q are the sea surface scalar roughness lengths. Thus, another method to parameterize air-sea fluxes is through these roughness lengths.

Under low to moderate wind conditions, air-sea momentum flux is usually estimated through the Charnock relation (Charnock, 1955):

$$gz_0/u_*^2 = \alpha \quad [25]$$

where u_* is the friction velocity and α is the Charnock parameter which was traditionally thought of as a constant. A constant Charnock parameter means the drag coefficient increases linearly with surface wind, and the effects of sea surface waves are implicitly included in this relation (Guan and Xie, 2004). However, it has been recognized in recent years that sea state including wave state and sea sprays could have important impacts on sea surface wind stress, especially under high wind conditions (Powell et al., 2003; Donelan et al., 2004; Liu et al., 2012; Green and Zhang, 2013).

As for the air-sea heat and moisture fluxes, many field observations show that the air-sea exchange coefficients for sensible heat and water vapor are independent of wind speed, corresponding to the decrease of the scalar roughness parameters with the increasing wind speed. The parameterization of sea surface scalar roughness parameters can be related to sea surface aerodynamic roughness length, e.g., through Coupled Ocean-Atmosphere Response Experiment (COARE) algorithm V3.1 (Fairall et al., 2003):

$$z_T = z_q = \min(1.1 \times 10^{-4}, 5.5 \times 10^{-5} \text{Re}_*^{-0.6}) \quad [26]$$

where $\text{Re}_* = z_0 u_* / \nu$ is the Reynolds number of sea surface aerodynamic roughness. However, under high wind conditions,

as surface wind speed increases, wave breaking and wind tearing of the wave crests disrupt the air-sea interface and generate sea sprays. The existence of sea sprays will, in turn, mediate the air-sea heat and water vapor fluxes, and impact air-sea momentum flux as well, especially under high wind conditions. Recent studies showed that the drag coefficient levels off or even decreases with increasing wind speed under high wind conditions, which is likely due to the existence of sea sprays. Although it is still controversial regarding whether sea sprays have significant impact on air-sea heat and moisture fluxes under low to moderate winds, it is believed that sea sprays have substantial contribution to air-sea heat and moisture transfers under high wind conditions.

Marine Weather Prediction System

A weather prediction system typically contains a data acquisition system, data analysis and assimilation system, a modeling system, and an application product development, dissemination, and decision support system (Figure 1). By collecting quantitative data about the current atmospheric and oceanic states through the observing system, marine weather forecasts predict how the atmospheric and oceanic states will evolve based on scientific understanding of the atmospheric, oceanic, and air-sea interaction processes. What makes marine weather forecasting different from terrestrial weather forecasting is that it provides predictions not only for weather conditions over the ocean and coastal regions, but also for the weather-induced oceanic conditions, such as sea surface waves, rip currents, and storm surge. Thus, a marine weather prediction system consists not only atmospheric but also oceanic observation, forecasting, and application systems as shown in Figure 1.

Marine meteorology and oceanographic observing systems focus on simultaneous observations of meteorological variables above the sea surface and oceanographic variables at and below the sea surface. Table 1 lists the major marine meteorology and oceanographic observation programs coordinated internationally and in the United States. In addition to *in situ* and ground-based observations, remote sensing from satellites of both the atmosphere and the sea surface constitutes a large component of marine meteorology and oceanographic data. These satellites are launched and maintained by international organizations such as the European Space Agency or individual countries such as the United States National Aeronautics and Space Agency, United States National Oceanic and Atmospheric Administration, and China Meteorological Administration.

A marine weather forecast system typically consists of a suite of weather, ocean circulation, and sea surface wave forecast models. Traditionally, these models are stand-alone without taking into account the mutual couplings among them. An atmospheric model predicts weather conditions and provides atmospheric forcing to drive the sea surface wave and ocean circulation models. A sea surface wave model, driven by surface winds from an atmospheric model, provides forecasts for sea states, including significant wave heights and mean wave periods, etc. An ocean circulation model, driven by the atmospheric forcing from an atmospheric model, predicts oceanic conditions including storm surge and inundation as well as

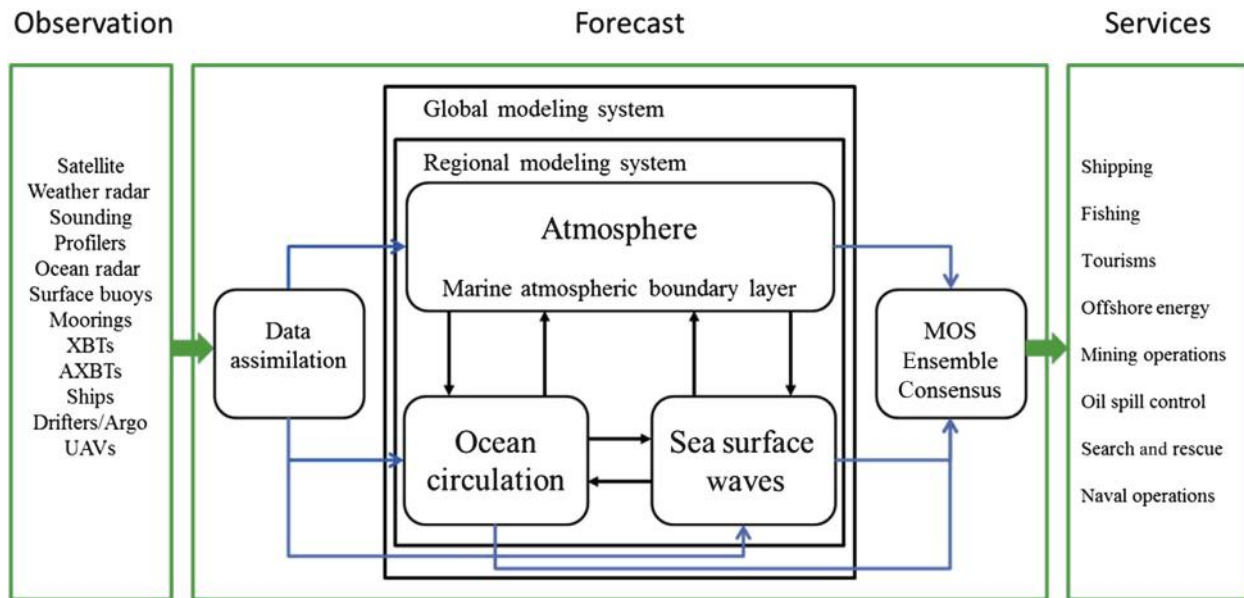


Figure 1 Illustration of a typical marine weather prediction system. XBTs, expendable bathythermographs; AXBTs, airborne expendable bathythermographs; UAVs, autonomous underwater vehicles; MOS, model output statistics.

Table 1 List of major marine meteorology and ocean observation programs

GOOS: GOOS is a permanent Global Ocean Observing System designed to provide observations, modeling, and analysis of surface meteorological and oceanic variables to support operational marine meteorology, climate change, and ocean service applications worldwide (http://www.ioc-goos.org). In the United States, the Integrated Ocean Observing System (IOOS) serves as the US component of GOOS (http://www.ioos.noaa.gov). It is a vital tool for tracking, predicting, managing, and adapting to changes in the US ocean, coastal, and Great Lakes environment. It comprises 11 regional associations. The EuroGOOS is an association of national governmental agencies and research organizations committed to European-scale operational oceanography within the context of GOOS (http://www.eurogoos.org). VOS: The World Meteorological Organization's (WMO) Voluntary Observing Ships (VOSs) scheme collects surface weather data from over 4000 voluntary ships and disseminated to weather service centers worldwide (https://www.wmo.int/pages/prog/amp/mmop/JCOMM/OPA/SOT/vos.html). The US component of VOS (http://www.vos.noaa.gov/vos_scheme.shtml). SOT: The Ship Observations Team under Joint Technical Commission for Oceanography and Marine Meteorology (JCOMM), providing ship-based surface meteorological observations, aero-logical profiles, and subsurface oceanographic observations (http://www.wmo.int/pages/prog/amp/mmop/sot.html). DBCP: The Data Buoy Cooperation Panel was established as a joint WMO/Intergovernmental Oceanographic Commission (IOC) initiative in 1985 to provide data from drifting buoys (Lagrangian drifters), moored buoys (meteorological moorings, tropical moorings, ocean reference stations), as well as ice buoys (http://www.wmo.int/pages/prog/amp/mmop/dbcp.html). GLOSS: The global sea-level network under JCOMM, maintaining the tide gauge network (http://www.wmo.int/pages/prog/amp/mmop/gloss.html). Argo: It is an international program to deploy and maintain a global array of around 3000 autonomous profiling floats measuring the profiles of ocean temperature and salinity to depths of around 2000 m, as well as ocean currents at that depth. Before it becomes operational, Argo is managed by a science team, but once operational, Argo will become a component of the integrated ocean observing system coordinated through JCOMM (http://www.argo.net/). OceanSITES: The Ocean Sustained Interdisciplinary Time-series Environment observation System provides data from deep ocean multidisciplinary reference stations (mainly moored buoys). It is an integral part of GOOS (http://www.oceansites.org/). GHRSSST: The group for high-resolution sea surface temperature (SST) program provides global high-resolution (<10 km) SST products to the operational oceanographic, meteorological, climate, and general scientific community (https://www.ghrsst.org/). NDBC: The United States National Data Buoy Center maintains 103 buoys and 47 fixed C-MAN stations in the oceans and the Great Lakes (http://www.ndbc.noaa.gov/).

sea surface variables. These stand-alone model runs can be considered as one-way coupling among these three model components, without feedback effects being considered.

However, as illustrated in **Figure 2**, the atmosphere, surface waves, and ocean circulation are coupled through air-sea (Kraus and Businger, 1994) and wave-current interaction processes (Xie et al., 2001, 2003). The weather systems are

coupled to the upper ocean mixed layer and surface waves through momentum, heat, and moisture exchanges at the air-sea interface. While the atmospheric forcing drives sea surface waves and underlying ocean currents, ocean exchanges heat and moisture with the atmosphere through air-sea heat and moisture fluxes. The existence of sea surface waves modifies the structures of both the atmospheric and marine boundary

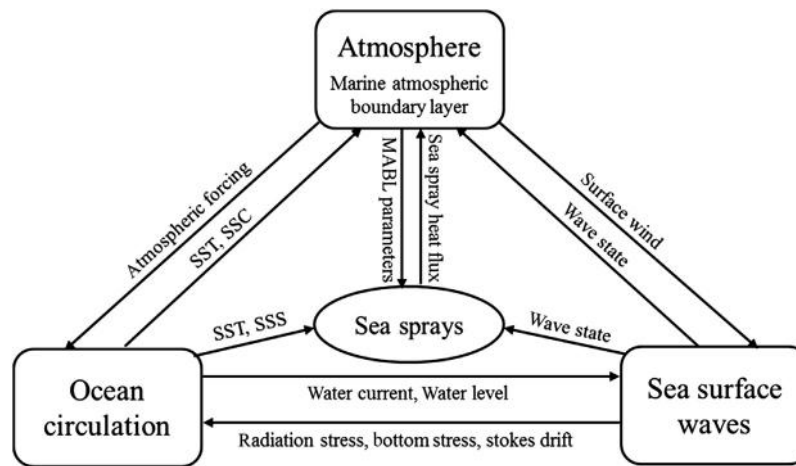


Figure 2 Illustration of a coupled atmosphere–ocean circulation–sea surface wave modeling system. SST, sea surface temperature; SSC, sea surface current; SSS, sea surface salinity. Adapted from Liu, B., Liu, H., Xie, L., Guan, C., Zhao, D., 2011. A coupled atmosphere–wave–ocean modeling system: simulation of the intensity of an idealized tropical cyclone. *Monthly Weather Review* 139, 132–152.

layers, and thus influences air–sea momentum and heat fluxes. Therefore, coupled atmosphere–wave–ocean modeling systems taking into account various air–sea and wave–current interaction processes have been developed and utilized in marine weather forecasting, especially for tropical cyclone systems which are associated with high winds and strongly coupled to the upper ocean mixed layer and surface waves (Liu et al., 2011).

Another important component in the marine weather forecasting system is data assimilation. As the observation system proceeds, various data assimilation techniques, including optimal interpolation, three-dimensional variational method (3DVAR), four-dimensional variational method (4DVAR), and ensemble Kalman filter (EnKF), are utilized to obtain better initial conditions for atmospheric (Zhang et al., 2011), sea surface wave, and oceanic models (Peng et al., 2006). There is also a limitation for the current data assimilation techniques for a coupled air–sea modeling system for marine weather forecasting. Data assimilation techniques are generally only applied to individual model components in a marine weather forecasting system. In this situation, the benefit of adjustments in one model by assimilating observational data can be compromised by errors from the other model (Peng et al., 2007). Unified data assimilation techniques for coupled air–sea modeling systems, which simultaneously assimilate various atmospheric and oceanic observations into coupled air–sea modeling systems are of increasing interest for marine weather forecasting.

In addition, because of the chaotic nature of the atmosphere and ocean, the error in the initial conditions, the massive computational power required, and an incomplete understanding of atmospheric, oceanic, and air–sea interaction processes, numerical prediction model forecasts become less accurate as the forecast time increases. The methods of ensemble forecast and model consensus can help to reduce forecast errors and to identify the most likely outcome.

The historical and near real-time marine weather data along with marine weather and sea state forecasts are disseminated to

various users in marine and coastal environments through international, regional, and national marine weather service centers. As human activities grow in the coastal zone, and as tourism, shipping, fishing, and exploitation of marine resources expand, there is a growing need for the marine meteorological services and information to adequately plan coastal zone marine activities and provide early detection and warning of marine-related hazards.

See also: Air Sea Interactions: Momentum, Heat, and Vapor Fluxes; Surface Waves. **Data Assimilation and Predictability:** Data Assimilation; Ensemble Prediction. **Dynamical Meteorology:** Primitive Equations. **Numerical Models:** Regional Prediction Models. **Oceanographic Topics:** General Processes. **Synoptic Meteorology:** Extratropical Cyclones. **Tropical Cyclones and Hurricanes:** Overview and Theory. **Tropical Meteorology and Climate:** El Niño and the Southern Oscillation: Theory; Intraseasonal Oscillation (Madden–Julian Oscillation). **Weather Forecasting:** Severe Weather Forecasting.

Further Reading

- Charnock, H., 1955. Wind stress on a water surface. *Quarterly Journal of the Royal Meteorological Society* 81, 639–640.
- Donelan, M.A., Haus, B.K., Reul, N., Plant, W.J., Stiassnie, M., Graber, H.C., Brown, O.B., Saltzman, E.S., 2004. On the limiting aerodynamic roughness of the ocean in very strong winds. *Geophysical Research Letters* 31, L18306. <http://dx.doi.org/10.1029/2004GL019460>.
- Fairall, C.W., Bradley, E.F., Hare, J.E., Grachev, A.A., Edson, J.B., 2003. Bulk parameterization of air–sea fluxes: updates and verification for the COARE algorithm. *Journal of Climate* 16, 571–591.
- Green, B.W., Zhang, F., 2013. Impacts of air–sea flux parameterizations on the intensity and structure of tropical cyclones. *Monthly Weather Review* 141, 2308–2324.
- Guan, C., Xie, L., 2004. A unified linear parameterization of the drag coefficient over the surface of the ocean. *J. Phys. Oceanogr.* 34, 2847–2851.
- Haidvogel, D.B., Beckmann, A., 1999. *Numerical Ocean Circulation Modeling*. Imperial College Press, London, 318 pp.

- Holton, J.R., Hakim, G.J., 2012. *An Introduction of Dynamic Meteorology*, fifth ed. Academic Press, Burlington, MA, 552 pp.
- Komen, G.J., Cavaleri, L., Donelan, M., Hasselmann, K., Hasselmann, S., Janssen, P.A.E.M., 1994. *Dynamics and Modeling of Ocean Waves*. Cambridge University Press, Cambridge, 532 pp.
- Kraus, E.B., Businger, J.A., 1994. *Atmosphere-Ocean Interaction*, second ed. Oxford University Press, Oxford, 362 pp.
- Liu, B., Liu, H., Xie, L., Guan, C., Zhao, D., 2011. A coupled atmosphere-wave-ocean modeling system: simulation of the intensity of an idealized tropical cyclone. *Monthly Weather Review* 139, 132–152.
- Liu, B., Guan, C., Xie, L., 2012. The wave state and sea spray related parameterization of wind stress applicable to from low to extreme winds. *Journal of Geophysical Research* 117, C00J22. <http://dx.doi.org/10.1029/2011JC007786>.
- Peng, S., Xie, L., Pietrafesa, L.J., 2007. Correcting the errors in the initial conditions and boundary conditions of storm surge simulation using an adjoint optimal technique. *Ocean Modelling* 18, 175–193.
- Peng, S.-Q., Xie, L., 2006. Effect of determining initial conditions by four-dimensional variational data assimilation on storm surge forecasting. *Ocean Modelling* 14, 1–18.
- Powell, M.D., Vickery, P.J., Reinhold, T.A., 2003. Reduced drag coefficient for high wind speeds in tropical cyclones. *Nature* 422, 279–283.
- Xie, L., Wu, K., Pietrafesa, L.J., Zhang, C., 2001. A numerical study of wave-current interaction through surface and bottom stresses: Part I: Wind-driven circulation in the South Atlantic Bight under uniform winds. *Journal of Geophysical Research* 106, 16841–16856.
- Xie, L., Pietrafesa, L.J., Wu, K., 2003. A numerical study of wave-current interaction through surface and bottom stresses: coastal ocean response to Hurricane Fran 1996. *Journal of Geophysical Research* 108 (C2, 3049).
- Zhang, M., Zhang, F., Huang, X.-Y., Zhang, X., 2011. Intercomparison of an ensemble Kalman filter with three- and four-dimensional variational data assimilation methods in a limited-area model over the month of June 2003. *Monthly Weather Review* 139, 566–572.

Operational Meteorology

DR Novak, Weather Prediction Center, College Park, MD, USA

© Published by Elsevier Ltd.

This article is a revision of the previous edition article by J V Cortinas Jr., W Blier, volume 4, pp 1567–1576, © 2003, Elsevier Ltd.

Synopsis

The field of operational meteorology involves the generation and widespread dissemination of weather information on a consistent and reliable basis. Operational meteorology plays a critical role in the safeguarding of lives and property, as well as providing highly valued information to both general public and commercial enterprises. This article provides an overview of operational meteorology, including the forecast process, information on the structure and function of modern weather services, interactions between the various national and private weather services, and on the interrelationships between operational meteorology and science and technology.

Introduction

Operational meteorology involves the generation and widespread dissemination of weather information on a consistent and reliable basis. In its modern application, this process typically involves the gathering of vast quantities of meteorological observations and execution of complex numerical forecast models, and thus tends to be extremely computer intensive. Operational forecasters examine and interpret the raw numerical forecast output, and use scientific knowledge about how the atmosphere behaves to develop finalized forecast information. This finalized information is disseminated to society, and plays a critical role in the safeguarding of lives and property, as well as providing highly valued information to both general public and commercial enterprises.

The content of the disseminated weather information varies according to the issuing organization and the type of weather information required by its customers; typically, they include gridded, graphical, and narrative forecasts and analyses of weather conditions over particular geographic regions and notification of any anticipated or observed hazardous weather conditions. Although broadcast meteorologists are perhaps the most recognized face of weather information, other forms of dissemination include governmental networks, radio, the Internet, and mobile applications.

This article provides an overview of operational meteorology, including the forecast process, information on the structure and function of modern weather services, interactions between the various national and private weather services, and on the interrelationships between operational meteorology and science and technology.

The Operational Forecast Process

Modern operational forecasting involves multiple aspects, which are shown schematically in Figure 1. It all begins with observations of the atmosphere, oceans, and land surface.

Observations are gathered from *in situ* platforms such as surface weather stations and weather balloons, or remotely sensed sensing platforms, such as Doppler and

dual-polarization radars and sophisticated satellite systems. These observations are then converted into a form that numerical computer models can use. These computer models are based upon the physical laws of atmospheric science. Atmospheric models are extremely complex, and predict variables such as temperature, moisture, and wind over the globe at time increments on the order of tens of seconds. These models have to consider factors such as snow cover, soil moisture, vegetation, sea surface temperature, solar radiation, and many others. As you can imagine, these models are computationally intensive, and continually test the limits of computer science. Figure 2 shows the growth of operational supercomputer resources since 2000 at the United States (US) National Weather Service (NWS), as measured by the number of calculations per second. A teraflop is a trillion floating-point calculations per second. Soon, more powerful petaflop computers will be used. These computers are capable of a thousand trillion calculations per second. As this computing power increases, the model representation of the atmosphere, ocean, and land surface becomes more and more realistic, resulting in improved forecasts.

Increased computing power has also allowed the advent of 'ensembles.' Ensembles are multiple model forecasts derived using slightly different initial conditions and representations of atmospheric processes. These systems account for the fact that the observations have errors, and our representation of the atmosphere also has errors. These small errors grow through the forecast period at different speeds, depending on the weather regimes. Ensembles quantify just how fast errors are growing in the forecast, and help quantify the uncertainty of the forecast. For example, each member of an ensemble provides a different possible forecast. If the differences between these possible forecasts are small, then the forecast scenario is fairly certain. However, if the members have widely different solutions, then the probability of any one scenario is small, and it is an uncertain forecast.

Output from these computer models is visualized and examined by human forecasters, who use their expert knowledge to make modifications. For example, the model may show the expected precipitation for tomorrow afternoon. The human

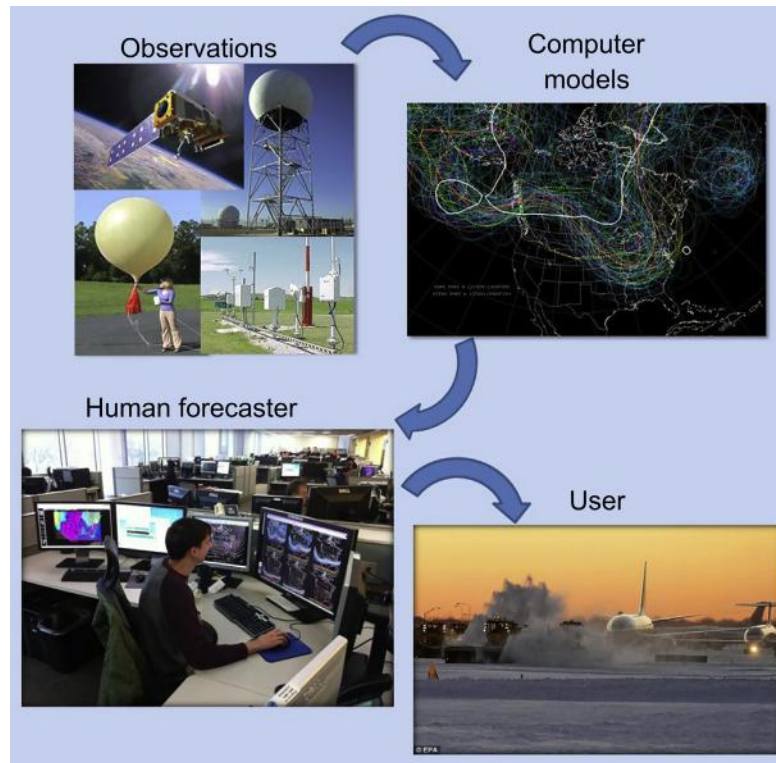


Figure 1 Schematic diagram of the operational forecast process showing (upper left) observational instruments including satellites, radar, weather balloons, and surface stations, (upper right) computer model output, (lower left) human forecasters interpreting the model output, and (lower right) users, such as airlines making flight decisions.

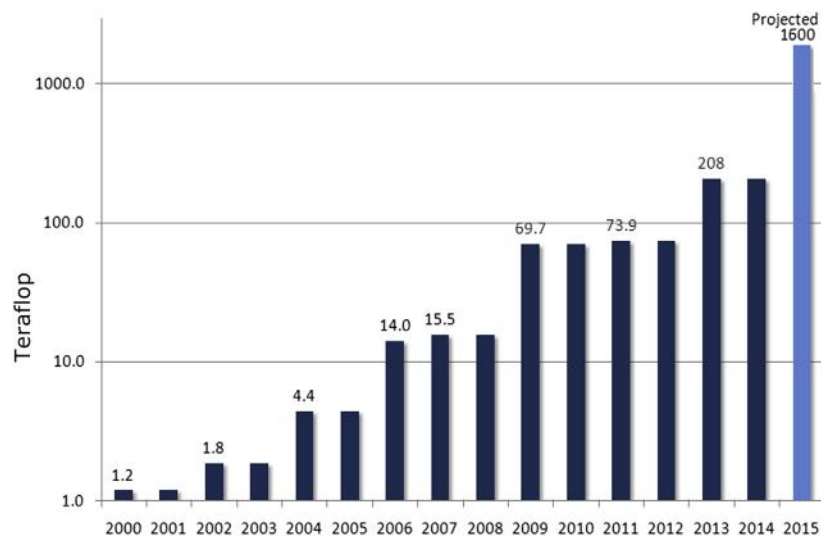


Figure 2 (Top) Growth of operational supercomputer resources since 2000 at the National Weather Service, as measured by the number of calculations per second (1 teraflop = 1 trillion floating-point operations per second).

forecaster considers factors such as model bias, physical realism, past verification, and consensus among various modeling systems to develop a most-likely forecast. The procedures and form in which forecasts are generated vary around the world, but it is becoming a common practice to create a database of grids of surface weather elements, which

can be graphically displayed as well as queried to create narrative text descriptions (**Figure 3**).

The final forecast information is then disseminated to a wide variety of users. Human forecasters may help interpret the impact of the weather on a particular user's need, such as whether airport runways will need to be cleared of snow, or

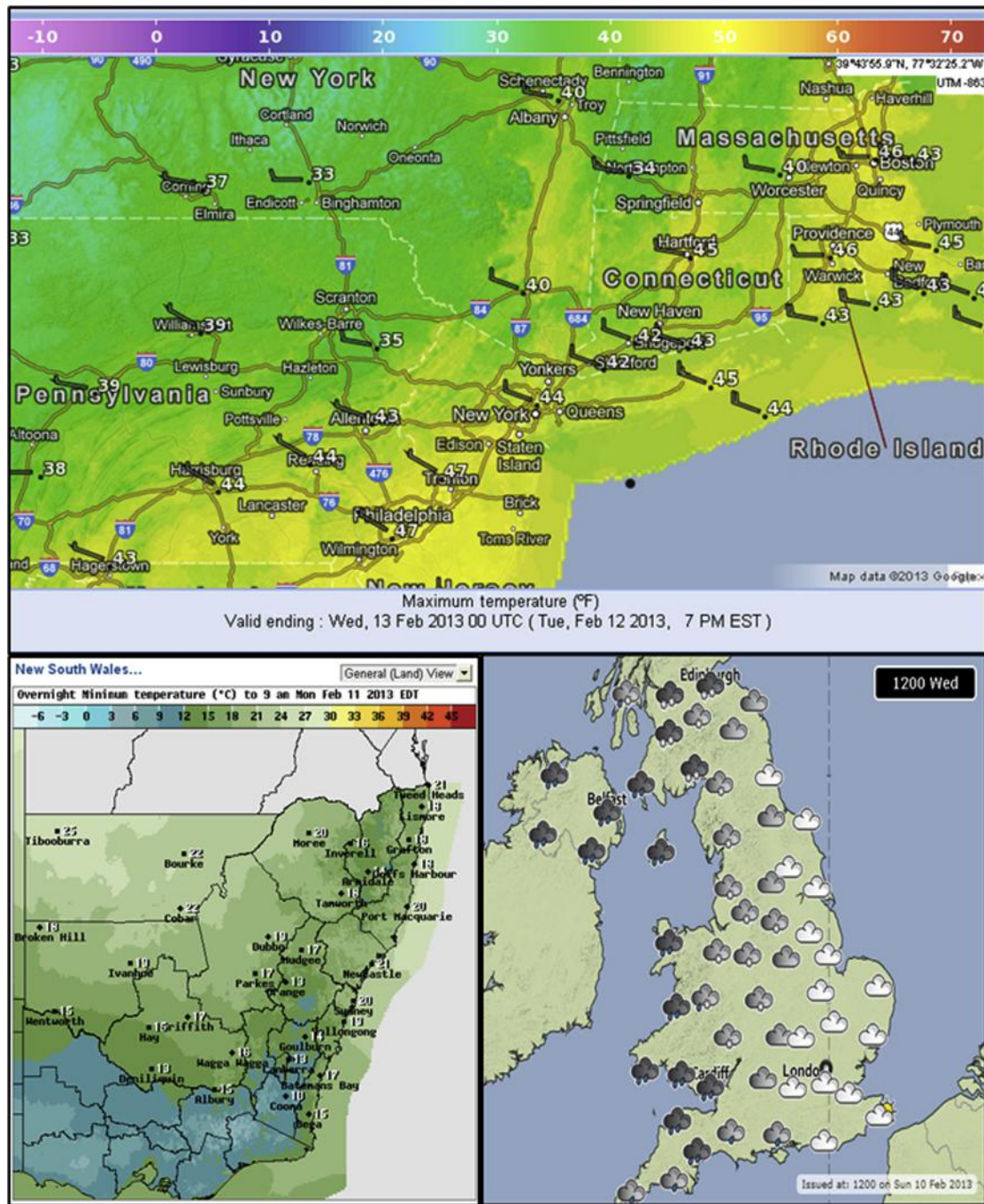


Figure 3 Examples of gridded and graphical forecasts from the (top) US National Weather Service, (bottom left) Australian Bureau of Meteorology, and (bottom right) United Kingdom Met Office.

more importantly whether evacuations are necessary for public safety. This decision support activity is a growing area of effort.

Modern Practice – United States

In this article, we describe the modern practice of operational meteorology using the United States as an example. This

includes public sector, military and civilian services, a vibrant private sector, and a supporting research community. The United States has one of the largest and most highly developed weather enterprises in the world. Since a number of other countries have weather enterprises somewhat similar to that of the United States, the latter provides a basic understanding of the present state of operational meteorology in the world. The US public, private, and research sectors are discussed next.

Public Sector

The public sector is shared by military and civilian efforts. The military forecasts are tailored to military needs, such as route forecasts for military ships and aircraft, and are generally not publicly available. The NWS is the public civilian effort, and is responsible for forecasting the weather, warning of weather hazards, and taking and recording observations of meteorological and hydrological conditions. This section will focus on the NWS.

The NWS accomplishes these tasks through an organizational structure comprising local forecast offices, regional offices, and national forecast centers. Working together, these offices provide the public with: (1) specific forecasts of expected weather conditions for the next 7 days, (2) general climate predictions up to 1 year in advance, (3) forecasts of potentially hazardous weather (called a 'watch'), and (4) warnings for imminent or already occurring hazardous weather.

The meteorological forecasts and services provided by the NWS can be divided into the following four basic programs:

1. *Warning Program.* Perhaps the most important function of operational meteorology is alerting society of dangerous weather. The warning program provides warnings to the public of hazardous weather and weather-related conditions. Warnings are issued for hurricanes and tropical storms, high winds, tornadoes and severe local storms, winter weather, river flooding, flash flooding, coastal flooding, high waves, and space weather anomalies, as shown in [Table 1](#). Warnings are generally issued based on

the likelihood of the event occurring and the impact the expected conditions may have.

2. *Public Weather Service Program.* This program provides the public with current weather and climate data, and forecasts ([Table 2](#)). These products also serve as the starting point for most interpretive and applied forecast services, including weather services provided by private sector meteorologists.
3. *Aviation Forecast Program.* This program provides area, route, and terminal weather information to the domestic and international aviation community to ensure safe and efficient flight operations.
4. *Marine Warning and Forecast Program.* This program provides routine forecasts, as well as watches and warnings for hazardous weather-related marine conditions.

The NWS relays weather information to the public by several methods, including the mass media, National Oceanographic and Atmospheric Administration (NOAA) Weather Radio, and other federal and international data transmission services. The media, in particular, play a critical role in disseminating weather information, including warnings of life-threatening storms.

Local weather offices

The NWS Weather Forecast Offices (WFOs) are the primary governmental providers of local weather and climate information, including warnings. There are a total of 122 WFOs, each of which has complete forecast responsibility for its particular geographic subregion during the first 7 days of the forecast. Together, these WFOs provide coverage for all of the

Table 1 Select types of hazardous weather warnings issued by National Meteorological Services and government partners

Hazard	Australia	United Kingdom	Canada	United States
Severe local storm	Severe thunderstorm		Severe thunderstorm	Severe thunderstorm
Winter weather	Severe weather	Snow Ice Cold	Tornado Snowfall Blizzard Freezing rain Windchill Cold wave	Tornado Winter storm Blizzard Ice Windchill
Flooding	Severe weather Flood	Rain Flood	Heavy rain	Flash flood Flood
Marine	Coastal waters wind Ocean wind		Marine wind	Gale Storm Hurricane force wind
Hurricane	Tropical cyclone		Hurricane	Hurricane

Table 2 Primary public forecast products issued by a US Weather Forecast Office

Product	Issuance frequency	Description
National Digital Forecast Database	At least twice daily with updates as necessary	2.5 km grids of sensible weather elements
Point-and-click forecasts (via web interface)	At least twice daily with updates as necessary	Detailed forecasts for a small area (2.5 km grid box). Typically include information on expected weather, probabilities of precipitation, the highest and lowest temperatures, and wind speed and direction. Covers periods up to 7 days.
Area forecast discussions	Four times per day	Discussion explaining the reason behind the current zone forecasts.

United States, including Puerto Rico, the U.S. Virgin Islands, and the US territories of the western Pacific.

All WFOs operate 24 h everyday of the year. Each office has a staff of ~10 forecasters, and is managed by a Meteorologist-in-Charge, a Warning Coordination Meteorologist (WCM), and a Science and Operations Officer (SOO). The WCM is responsible for developing the WFO's Warning Preparedness Program, which strives to reduce loss of life and property, and to ease the social and economic impacts of weather-related natural disasters. The WCM accomplishes this by assuring warning system efficiency, increasing public response to warnings, and providing technical help for the development of response plans by government and private institutions. The SOO serves as the primary scientific advisor to the WFO's staff, assuring the scientific integrity of all hydrometeorological products and services provided by the WFO, and is responsible for transferring new and emerging scientific technologies and techniques from the research community to the operational weather forecast and warning environment. Each office also has its own electronics and information technology staff, which is responsible for maintenance of both internal and external electronic equipment (including, among other things, WFO computer-based data acquisition and display systems, surface weather sensors, and weather radars).

Regional offices

River Forecast Centers

River Forecast Centers (RFCs) specialize in water resource information, with information on current and forecast river conditions, floods, and droughts. There are 13 RFCs in the NWS, organized by regional river systems.

Center Weather Service Units

The Center Weather Service Units (CWSUs) specialize in providing the Federal Aviation Administration Air Route Traffic Control Centers specialized weather information for regional aviation. These units promote the safe and efficient flow of aircraft within the National Airspace System. There are 21 CWSUs in the NWS.

National centers

There are nine national forecast centers forming the National Centers for Environmental Prediction (NCEP). NCEP is comprised of two core support centers:

Environmental Modeling Center – develops and improves the operational suite of model and ensemble output for weather, climate, and hydrologic forecasting.

NCEP Central Operations – runs the operational suite of model and ensemble output and disseminates this information.

The information provided by the core support centers is used by seven specialized service centers:

Aviation Weather Center – provides aviation warnings and forecasts of hazardous flight conditions at all levels within domestic and international.

Climate Prediction Center – monitors and forecasts short-term climate fluctuations and provides information on the effects climate patterns can have on the nation.

National Hurricane Center – provides forecasts of the movement and strength of tropical weather systems and issues watches and warnings for the United States and surrounding areas.

Ocean Prediction Center – issues weather warnings and forecasts out to 5 days for ocean areas poleward of 30° N in the eastern Pacific and 31° N in the western Atlantic.

Space Weather Prediction Center – provides space weather alerts and warnings for solar disturbances that can affect people and equipment working in space and on earth.

Storm Prediction Center (SPC) – provides tornado and severe weather watches for the contiguous United States along with a suite of hazardous weather forecasts.

Weather Prediction Center – provides nationwide analysis and forecast guidance products out through 7 days.

An important mission of these centers is to provide the local and regional forecast offices with guidance products and services to assist them in meeting their watch, warning, and forecast responsibilities. Additionally, some centers have watch and warning responsibility.

To illustrate the function of the national centers, and the interaction between them and the local forecast offices, we proceed with an examination of the SPC, located in Norman, Oklahoma. The SPC is responsible for issuing forecasts, watches, and watch status messages for severe thunderstorms and tornadoes across the contiguous United States. (In the United States, a severe thunderstorm is defined as one that results in one or more of the following: a tornado; straight-line winds of at least 25 m s⁻¹ (50 knots); and/or hail with a diameter of at least 1.9 cm (3/4 inches).) In addition, several times each day SPC meteorologists prepare 'convective outlooks' of expected areal coverage of nonsevere and severe thunderstorms for the next few days. A convective outlook consists of a map of the United States depicting these areas of expected thunderstorm activity (Figure 4), accompanied by text describing the underlying meteorological reasoning (see Weather Forecasting: Severe Weather Forecasting for more information about forecasting severe weather). The SPC also monitors fire weather across the country, and issues specific products for these hazards.

When it appears that severe weather will occur within the next several hours, the SPC meteorologists coordinate with WFO forecasters by an internal instant messaging service and conference calls to issue a severe thunderstorm or tornado watch for the area of concern (Figure 4). The watch is an important component of the severe weather program in the United States and places local and state government offices, as well as the public in the watch area on higher alert for the possibility of severe weather. When severe weather develops, SPC meteorologists continue to provide assistance to local offices by issuing regularly updated mesoscale discussions and watch status messages. Local offices interpret radar data, examine environmental data, and monitor storm spotter reports to issue local warnings (Figure 4). These warnings trigger sirens, mobile phone alerts, and TV crawl messages to alert the public.

Private Sector

The need for forecasts tailored to user decision making, particularly in the energy and transportation industries, has increased the role of the private sector in several industrialized countries. For example, some private sector forecasters provide

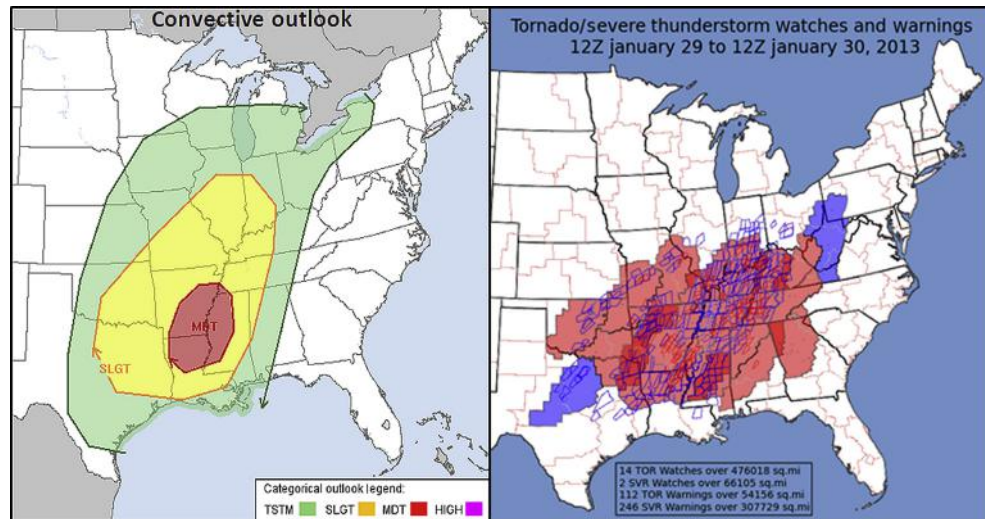


Figure 4 Example of products issued by the SPC and local WFOs for the 29–30 January 2013 outbreak. (Left) Convective outlook issued 2 days prior to the event. Areas to the right of the solid lines with arrows indicate where forecasters expect thunderstorms (severe and nonsevere) to occur during the outlook period. Areas identified as ‘SLGT’ or ‘MDT’ indicate where there is a slight and moderate risk of severe thunderstorms (those that produce either a tornado, winds $>25 \text{ m s}^{-1}$ (50 knots), or thunderstorm-generated damage, and/or hail with a diameter of at least 1.9 cm (3/4 inches).) The risk level indicates the number and/or coverage of severe thunderstorms within the identified areas. (Right) Watches (red is tornado and blue is severe thunderstorm) issued by the SPC and the hollow polygons are warnings issued by the local WFOs. Figure provided by the NOAA/NWS/Storm Prediction Center.

energy companies with detailed and specially tailored weather forecasts for particular regions to help them anticipate future energy consumption.

Many private companies use the fundamental weather forecast information from the NWS, and provide meteorological information catered to a wide range of customers in the United States, including broadcast and the Internet media, energy, and transportation companies. Private sector operational meteorologists play an important role in providing the public with reports of current and forecast weather conditions; they

also contribute significantly to the rapid and widespread dissemination of hazardous weather information from the NWS. A 2010 survey of the US public found that television, the Internet, and radio are the public’s primary sources of short-term weather information, especially during hazardous weather events (Figure 5). Commercial broadcast stations and their meteorologists work in partnership with the NWS to quickly transmit official advisories, watches, and warnings (Figure 6).

Private sector operational meteorologists are also employed by companies that provide fee-based forecasts specifically

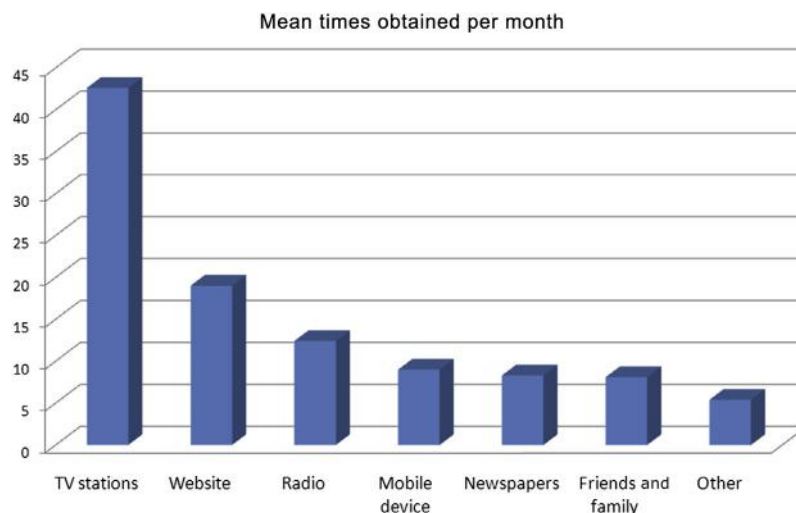


Figure 5 Sources of weather forecasts from a national US survey of over 2000 people. The survey was conducted in 2010. Adapted from Demuth, J., Lazo, J.K., Morss, R.E., 2012. Assessing and Improving the NWS Point-and-click Webpage Forecast Information. <http://opensky.library.ucar.edu/collections/TECH-NOTE-000-000-000-859>.



Figure 6 A broadcast meteorologist presenting weather information during a local news broadcast. The weather maps seen by television viewers behind the meteorologist are electronically generated using a process called 'chroma key,' which uses a blank neon green or electric blue 'chroma wall' to create the effect. © Times Union.

tailored to the particular needs of their customers, thus potentially adding value to the more general forecasts issued by the NWS. A citrus farmer, for example, may want to know whether temperatures will go below freezing in his field on a particular night, or a major business franchise may want to know whether the weather will affect shipments.

Other companies that employ operational meteorologists include those that provide services in the areas of air quality, climatology, training and education, forensics, research and development, instrumentation, data processing, remote sensing, and weather modification. Whatever the specific task, and whether working in the public or private sector, all operational meteorologists must have a good scientific understanding of how the atmosphere behaves, the ability to synthesize data from multiple sources quickly, and effective communication skills.

Research Sector

Operational meteorology is based on scientific understanding of the atmosphere and oceans. It is critical to infuse new research findings into applied services. Most of the applied atmospheric science research is conducted by universities and federal labs, such as the National Center for Atmospheric Research, National Severe Storms Lab, and Atlantic Oceanographic and Meteorological Laboratory. Nearly 100 colleges and universities offer Atmospheric Science or Meteorology degrees in the United States, providing a robust educational and research foundation for operational meteorology. The integration of this new science and technology into operations is discussed below in the [Integration of Science and Technology into Operations](#) section.

International Weather Services

The role of the national meteorological services in other countries is similar to that in the United States: to generate and disseminate weather observations, forecasts, and hazardous

weather advisories for the benefit of public safety and national commerce. In addition to periodic reports of weather conditions, the most common type of information provided by national meteorological services is a general weather forecast for the next several days ([Figure 3](#)). Many national meteorological centers also provide warnings for hazardous weather, similar to those issued in the United States ([Table 2](#)).

In addition to providing warnings, reports of weather conditions, and general weather forecasts, some national meteorological services also provide detailed weather information to particular industries, often those associated with transportation and energy. Frequently, these customers pay a fee to the government to receive this specifically tailored information. This fee-based model is unlike that used in the United States, where such specialized services are generally provided by individual private companies rather than government agencies.

In countries with limited resources, international efforts, such as those coordinated by the World Meteorological Organization (WMO), are necessary to help provide basic meteorological services. The WMO, a branch of the United Nations, was formed to help coordinate international cooperation among the approximately 35 regional and 190 national meteorological centers in the world. Formed in 1951, the WMO helps Member States with worldwide coordination of national meteorological activities through the establishment of cooperative agreements; the development of standards for measuring and reporting atmospheric conditions; the establishment of meteorological observation sites and training of meteorological personnel in developing countries; and the implementation of various research programs (including the World Climate Program, the Hydrology and Water Resources Program, and the Atmospheric and Environment Program).

The WMO also develops and coordinates procedures for sharing observational data and meteorological services. The Global Telecommunication System is the coordinated global system of telecommunication facilities and arrangements for the rapid collection, exchange, and distribution of meteorological and related data, forecasts, and alerts. This communication network enables real-time exchange of information, critical for forecasting and warnings of hydrometeorological hazards in accordance with established procedures. This is an especially important function, as meteorological phenomena cross geopolitical boundaries and can affect many countries at the same time. For example, when a hurricane threatens the coasts of the United States, Cuba, and Mexico, the national meteorological services in these countries work together by sharing meteorological information and resources to ensure all three countries receive adequate warning of any associated threats to life and property. The planning for this important type of international interaction is coordinated by the WMO.

Another international collaborative enterprise is the European Centre for Medium-Range Weather Forecasting (ECMWF), located in Reading, England. Formed in 1973 by a multinational convention, the ECMWF is an international organization supported by 23 European states. The objectives of the ECMWF are: (1) to develop numerical methods for medium-range (3–10 days) weather forecasting; (2) to prepare, on a regular basis, medium-range weather forecasts for distribution to the meteorological services of the Member States; (3) to conduct scientific and technical research directed to the

improvement of these forecasts; and (4) to collect and store appropriate meteorological data. Using state-of-the-art computer systems and having its own dedicated research staff, the ECMWF is the premier operational and research institutions for medium-range forecasting in the world.

Integration of Science and Technology into Operations

The evolution of operational meteorology is dependent on the capabilities of the available technology and on advances in the science of meteorology. In recent years, technological developments in four key areas have played critical roles in the advancement of operational meteorology: remote sensing of the atmosphere, computing speed and capacity, communications, and graphical display software. The first of these has resulted in the acquisition of far more comprehensive and detailed observations of the state of the atmosphere. Recent advancements in meteorological radar and satellite-borne instruments, in particular, have enabled increasingly detailed observation of the structure of the atmosphere in regions distant from the physical locations of the actual observing equipments. For example, the upgrade of the US radar network to include dual polarization now allows better detection of tornadoes, hail, heavy rain, and winter weather. With the resulting improved data coverage, numerical weather forecasts have also improved (since a prerequisite for determining the future state of the atmosphere is an accurate rendition of its present structure). These improved forecasts are related in part to the extraordinary recent increases in computing speed have also been of great significance (e.g., [Figure 2](#)). One of the primary challenges (and limitations) in numerical weather

prediction lies in trying to predict the future state of the atmosphere faster than the atmosphere itself evolves. The faster the computing devices available for carrying out the enormously voluminous and complex calculations, the more detail that can be included in the same calculation time.

Rapid increases in the quantity of observational and computer model data place increasing demand on the communication systems that carry this information to the weather enterprise. Commensurate development also becomes necessary in the equipment and techniques used to organize and display these data. Data management and display issues have been addressed through the recent development and implementation of sophisticated workstation-based meteorological graphics display software.

An example of the latter is the Advanced Weather Interactive Processing System, and its follow-on upgrade being commissioned in all forecast offices in the US NWS by 2015 ([Figure 7](#)). This system, and others like it, allows the forecaster to quickly and comprehensively examine graphical displays of both meteorological observations from multiple sources and the predicted future state of the atmosphere from numerical forecast models. The system also includes the capability for examination, generation, and dissemination of gridded, graphical, and text-based products. So, for example, a forecaster could overlay color-enhanced displays of satellite imagery, surface weather reports, weather radar data, and remotely sensed lightning strikes to quickly determine where the most severe conditions were occurring within a large thunderstorm complex, and then write and disseminate any requisite warnings.

Traditionally, the transfer of promising applied science and technology advances has been slow and challenging. This is due to the high bar operational meteorology sets; that is, the



Figure 7 A meteorologist working at an Advanced Weather Interactive Processing System workstation with a situational awareness monitor overhead at the New York City Weather Forecast Office. Adapted from NOAA.

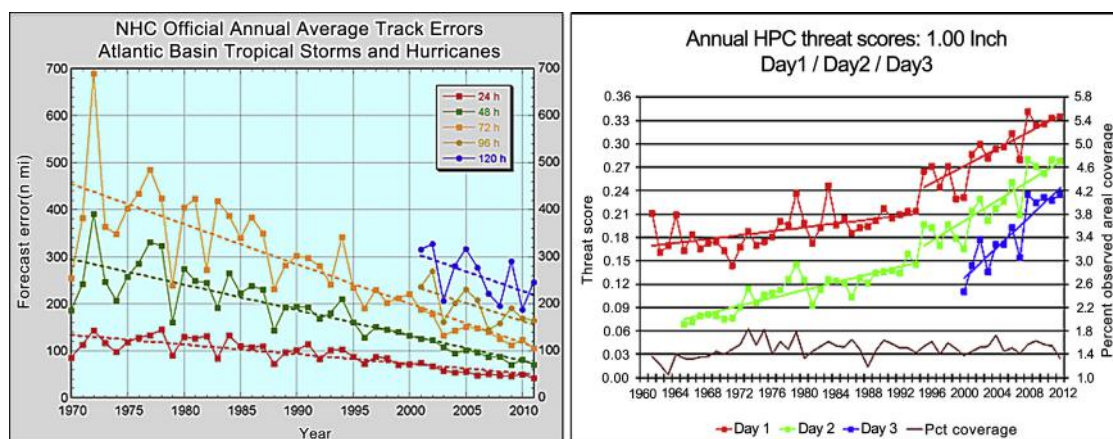


Figure 8 (Left) Trends in NWS National Hurricane Center track errors for Atlantic Basin Tropical Storms and Hurricanes and (right) trends in NWS Weather Prediction Center rainfall skill. Improvement is noted for both phenomena. In fact, a day 3 forecast today is similar in skill to a day 1 forecast in 1990. Adapted from NOAA.

new advancement not only has to be an improvement over the current system, but also has to be efficient, sustainable, and easily adaptable to the existing operational system. To foster the transfer of promising applied science and technology advances, testbeds are playing an increasingly important role. Testbeds facilitate the testing of new advances in science and technology in a pseudo-real-time operational environment, thereby exposing the technique to the rigors of operational forecasting, including feedback from forecasters. This engagement of researchers and developers early in the testing process has resulted in an acceleration of advancements.

The combined improvements of remote sensing of the atmosphere, computing speed and capacity, communications, and graphical display software have resulted in tremendous improvement in weather forecasts. For example, a day 3 forecast of 1 inch of precipitation or the track of a tropical cyclone today is similar in skill to a day 1 forecast of these events in 1990 (Figure 8). These improvements are a testament to the power of science and technology advancements, and to the weather enterprise that has applied these advances for the benefit of society.

Emerging Trends

As in many fields, advancements in technology have provoked the question of what the role of the human should be. Traditionally, human forecasters have interpreted computer model output, and made adjustments that resulted in improved final forecasts; however, this human edge has eroded in the recent decade as models improve. Debate within the community has focused on whether the forecaster's skills are most aptly applied to the short-range forecast and warnings (first 24 h of the forecast), high-impact events at any lead time, or to the interpretation of the forecast for users.

Regardless of how involved the forecaster is in the creation of the forecast, consensus is building that human forecasters can play an important role in providing interpretive decision

support services. Users are particularly interested in how the weather will affect their critical and specific decisions. These decisions can range from whether to stage road crews for winter weather, to adjusting energy production, to considering whether mandatory evacuations are necessary. To address these societal needs, the public sector is embedding meteorologists within the Emergency Operations Centers of large cities to help the emergency management community understand the weather forecast and the impacts relative to their decisions. An example from the private sector is the close consultation with energy providers on the ramp-up and ramp-down of energy production to meet demand.

As meteorologists provide these interpretive services, new challenges are arising. For example, what is the best way to communicate uncertain weather information? What forms of communication (for example, verbal or graphical) are most effective in fostering understanding? How can one motivate individuals to evacuate in life-threatening weather situations? Operational meteorologists are partnering with social scientists to help answer these types of questions.

A key area of focus has been on the use of weather information to minimize cost and risk. Although in its infancy, cost-loss models are being used in various industries. Consider the example of resource protection, where a citrus farmer has to decide whether to take mitigating action to save the crop (Figure 9). In this case, there is a cost to the mitigating action, but also a potential large loss if mitigating action is not taken and a damaging freeze occurs. The farmer needs to know at what risk of a damaging hard freeze the benefit of taking mitigating action outweighs the cost. Cost-loss models consider the probabilities of a loss (damaging hard freeze), and the cost of taking the mitigating action (such as water, fuel, and overtime). In other words, they help determine the user's risk tolerance. If the risk of an event is greater than this risk tolerance, mitigating action should be taken. A complementary example of resource acquisition is shown in Figure 9. These cost-loss models may not take into account other social factors, such as social credibility, local politics, and policies, and so may not be

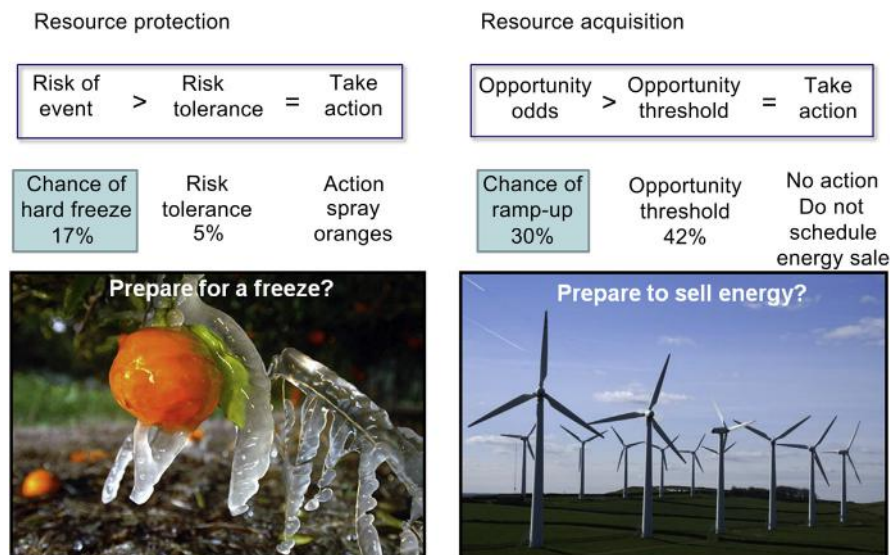


Figure 9 Examples of using cost-loss concepts for resource protection (left) and resource acquisition (right). Adapted from NOAA.

completely calibrated, but cost-loss models are an emerging tool for operational meteorologists conducting decision support.

Additionally, as weather forecasts improve, there are emerging areas of application including air quality, water quality, ecosystem health, and human health. For example, air quality is sensitive to temperature, humidity, atmospheric stability, sunshine, and wind. Detailed forecasts of these parameters are critical to anticipate harmful air quality days. Similarly, heavy rainfall can wash pollutants and fertilizers into watersheds, disrupting the natural ecosystem. The weather also impacts human health. For example, wet, warm conditions with warm nights can increase the mosquito population, leading to increased spread of disease.

Despite the enormous advances made in operational weather forecasting during the twentieth century, theoretical predictability studies suggest significant further improvement remains possible. As this evolution continues, operational meteorology will remain the bridge between our scientific understanding of how the atmosphere behaves and the application of this knowledge for the safety and benefit of society.

See also: Agricultural Meteorology and Climatology. **Data Assimilation and Predictability:** Data Assimilation; Ensemble Prediction. **Dynamical Meteorology:** Overview. **General Circulation of the Atmosphere:** Weather Regimes and Multiple Equilibria. **Mesoscale Meteorology:** Overview. **Numerical Models:** Regional Prediction Models. **Synoptic Meteorology:** Forecasting; Weather Maps. **Weather Forecasting:** Seasonal and Interannual Weather Prediction; Severe Weather Forecasting.

Further Reading

- Clark, A.J., et al., 2012. An overview of the 2010 hazardous weather testbed experimental forecast program spring experiment. *Bulletin of the American Meteorological Society* 93, 55–74.
- Demuth, J., Lazo, J.K., Morss, R.E., 2012. Assessing and Improving the NWS Point-and-Click Webpage Forecast Information. <http://opensky.library.ucar.edu/collections/TECH-NOTE-000-000-859>.
- Lazo, J.K., Morss, R.E., Demuth, J.L., 2009. 300 billion served. *Bulletin of the American Meteorological Society* 90, 785–798.
- Mass, C.F., 2003. IFPS and the future of the national weather service. *Weather Forecasting* 18, 75–79.

Relevant Websites

- <http://www.bom.gov.au> – Australian Bureau of Meteorology. This site contains information about weather services in Australia as well as Australian weather data.
- <http://www.ecmwf.int> – European Centre for Medium-Range Weather Forecasts (ECMWF). This site contains information about the types of research conducted at the ECMWF.
- <http://www.metoffice.gov.uk/> – The Meteorological Office of the United Kingdom (UK). This site contains weather services of the United Kingdom (UK).
- <http://www.msc-smc.gc.ca> – Meteorological Service of Canada. This site contains information about the national weather services in Canada as well as Canadian weather data.
- <http://www.nws.noaa.gov> – National Oceanic and Atmospheric Administration/National Weather Service. This site contains information about national weather services in the USA as well as American weather data.
- www.wmo.ch – World Meteorological Organization (WMO). Information about the activities of the WMO.

Seasonal and Interannual Weather Prediction

JP Li and RQ Ding, Institute of Atmospheric Physics, Chinese Academy of Sciences, Beijing, China

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by M Hoerling and A Kumar, volume 6, pp 2562–2567, © 2003, Elsevier Ltd.

Synopsis

This article presents a brief review of the current state of knowledge and capability in seasonal and interannual prediction. The review covers several relevant major issues, including (1) the definition and scientific basis of seasonal and interannual prediction; (2) the nonlinear local Lyapunov exponent predictability theory and its application in quantifying the seasonal predictability limit; (3) current approaches to seasonal and interannual predictions; and (4) current practice in seasonal and interannual predictions. Within this context, some of the major challenges remaining in interannual climate prediction are addressed.

Introduction

Climate predictions are inherently probabilistic statements about the future climate conditions on timescales ranging from seasons to decades or longer, and on spatial scales ranging from local to regional and global. Specifically, predictions of seasonal and interannual weather (i.e., short-term climate) are predictions of the departures from the average (normal) climate for upcoming seasons, and these are distinct from short-range weather predictions that attempt to deterministically forecast day-to-day weather behavior. Such predictions may provide some statistics on the seasonal or annual mean anomaly together with a measure of its probability of occurrence, and such information is useful for governmental, nongovernmental, and private agencies in making long-term decisions and planning in various fields (e.g., farming, early warning of potential hazards, drought mitigation, disaster prevention, insurance policy, and other economic activities). In the past two decades, there has been substantial progress in seasonal predictions, and now many operational and research centers around the world routinely make such predictions. The success of seasonal predictions has arisen from an improved understanding of the sources and limits of seasonal predictability as well as advances in climate models.

The sources of predictability vary with the timescale, influenced by the timescales of the predictands (i.e., the variables to be predicted). Slowly varying boundary conditions, such as sea surface temperatures (SSTs), sea ice, soil moisture, and snow cover at the surface, are common sources of predictability on seasonal and interannual timescales. One of the most important sources of seasonal and interannual climate predictability is the El Niño–Southern Oscillation (ENSO) phenomenon, which is the dominant mode of variability in the tropical Pacific at interannual timescales and that has a widespread effect on the global climate system. Recent research has indicated that the influence of the stratosphere on the troposphere is also an important source of seasonal predictability. Despite these sources, the utility of seasonal and interannual climate prediction is limited by errors in initial and boundary conditions, and by deficiencies in prediction models. The time-averaged seasonal anomalies in current prediction models can only be accurately predicted for a lead time of a few

months, thereby indicating that the predictability of such anomalies is inherently limited. It is useful to know the limits of seasonal and interannual predictability, as such information could be used to guide improvements in prediction models.

Various methods have been proposed to estimate seasonal and interannual predictability. Most of these methods assess the relative strength of the potentially predictable ‘climate signal’ due to the boundary forcings and of the unpredictable ‘climate noise’ due to intrinsic random variations, or the ‘signal-to-noise’ (S/N) ratio. However, the method employed to calculate the S/N ratio provides only a qualitative measure of seasonal and interannual predictability, such as identifying regions as being either predictable or unpredictable, but without quantifying the limits of predictability.

Recently, nonlinear methods have been introduced to quantitatively measure seasonal and interannual predictability. For example, the nonlinear local Lyapunov exponent (NLLE), which is a nonlinear extension of the traditional Lyapunov exponent concept, can quantitatively determine the limit of seasonal and interannual predictability by exploring the divergence evolution of the distance between initially local dynamical analogs (LDAs) from the observational time series. Using the NLLE method, it is possible to determine how far ahead we can predict the time-averaged seasonal anomalies in different regions. In addition to the NLLE, other nonlinear methods, such as the condition nonlinear optimal perturbation, have been introduced and applied to analyses of ENSO prediction and predictability.

Basic approaches to seasonal to interannual predictions include empirical approaches trained on observational data, dynamical approaches using general circulation models (GCMs), and combinations of these approaches. Regardless of the procedure, these predictions often show relatively low skill due to the influences of random weather fluctuations and other sources of unpredictability, particularly in areas away from the tropics. Given this limited predictability, it is more appropriate to make probabilistic seasonal and interannual predictions rather than deterministic predictions, since a probabilistic prediction yields estimates of the probability that the seasonal or annual mean temperature and precipitation will be above, near, or below normal, and provides quantitative information regarding prediction uncertainty.

Seasonal prediction products are now available from World Meteorological Organization (WMO)-designated Global Producing Centres (GPCs) of Long-Range Forecasts and are produced on multiple spatial scales by using GCMs to serve a wide range of users for various purposes. However, validation of these products is limited by the predictability of predictands. Empirical approaches based on an understanding of physical mechanisms on how predictors impact predictands are also sometimes employed to provide more skillful seasonal forecasts. Interannual prediction products currently have relatively low skill levels and are not available at most forecasting centers; consequently, there is a need for vast improvements in dynamical or empirical models for interannual predictions.

Quantifying Seasonal and Interannual Predictability

Previously, seasonal and interannual predictability has mainly been estimated using the S/N ratio method. The 'climate signal' arises from the slowly varying boundary forcings and represents the potentially predictable component of seasonal variability. The natural variability, or 'climate noise,' due to internal dynamics, represents the unpredictable component of seasonal variability. Under the assumption that the nonlinear interactions between signal and noise are small, evaluating the ratio of their variances will provide a qualitative assessment of the seasonal predictability. However, this method has several limitations. First, the climate signal and noise are not entirely independent or linearly separable, and their nonlinear interactions are sometimes strong and might make an important contribution to predictability. Second, the natural variability is not completely unpredictable. The internal dynamics and persistence of the atmosphere could also provide some predictability. Finally, the S/N ratio method provides only a qualitative (not quantitative) measure of seasonal and interannual predictability. In fact, we are sometimes more concerned with how far ahead we can predict the time-averaged seasonal anomalies. Here we quantitatively estimate seasonal and interannual predictability, using the NLE method that has recently been introduced to quantitatively measure the predictability of dynamical systems.

For any dynamical system, the NLE, λ , is defined as:

$$\lambda[x(t_0), \delta(t_0), \tau] = \frac{1}{\tau} \ln \frac{\|\delta(t_0 + \tau)\|}{\|\delta(t_0)\|}, \quad [1]$$

where $x(t_0)$ denotes an initial state in phase space, $\delta(t_0)$ the initial error, t_0 initial time, τ evolution time, and $\delta(t_0 + \tau)$ the evolution error. The ensemble mean NLE over the global attractor of the dynamical system is given by

$$\begin{aligned} \bar{\lambda}[\delta(t_0), \tau] &= \int_{\Omega} \lambda[x(t_0), \delta(t_0), \tau] d\Omega \\ &= \langle \lambda[x(t_0), \delta(t_0), \tau] \rangle_N, \quad (N \rightarrow \infty) \end{aligned} \quad [2]$$

where Ω represents the domain of the global attractor of the system and $\langle \lambda[x(t_0), \delta(t_0), \tau] \rangle_N$ denotes the ensemble average of samples of sufficiently large size N ($N \rightarrow \infty$). The ensemble mean NLE reflects the global evolution of mean error growth over an attractor and can measure global mean predictability.

For nonlinear dynamical systems, we can directly calculate the mean NLE via the numerical integration of their error

evolution equations. In addition, if large amounts of observational or experimental data are available for dynamical systems, we can estimate the mean NLE by making use of these data when the evolution equations of the systems are either unknown or incomplete. The general idea of the algorithm to estimate the mean NLE based on observational data is to find LDAs of evolution patterns from observational time series. The LDA is based on the initial information and evolution information at two different time points in the time series. Considering that the initial information of chaotic systems will gradually decay at an exponential rate, the error within the evolution time interval, which is defined as the time at which autocorrelations of observational time series drop to below the 95% significance level, is multiplied at each step by a weighting coefficient that decreases at an exponential rate with time. The exponential rate can be estimated from autocorrelations of observational time series. For observational time series at different grid points, the evolution time interval and exponentially decreasing weighting coefficients are taken as their average over all grid points. The mean NLE is then estimated by calculating the mean exponential divergence rate of the LDAs, and the mean error growth is obtained by calculating the mean root-mean-square error (RMSE) for all LDAs. The mean RMSE is chosen as a measure of error because it bridges a close relationship between the error saturation value and the standard deviation (SD) of the observational time series, as will be shown later.

The mean error growth obtained by calculating the mean RMSE between LDAs initially increases quickly then slows down and finally reaches a saturation value (Figure 1). According to the dynamical systems theory, the error saturation value represents the average distance between two randomly chosen points over an attractor, which is also referred to as the attractor radius (AR). The AR is equal to $\sqrt{2}$ times the SD of the observational time series. The saturation level of error growth implies that almost all information on initial states is lost and the prediction becomes meaningless. The potential predictability limit is then determined as the time at which the mean error reaches the AR, while the practical predictability limit is determined as the time at which the mean error reaches the SD (Figure 1). The potential predictability limit may only be

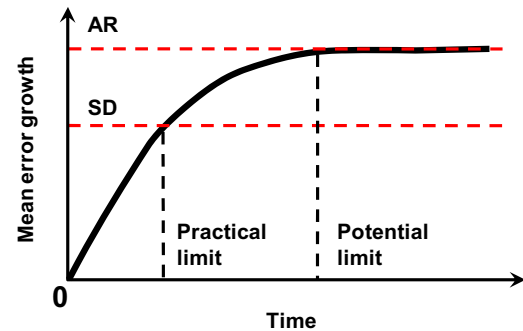


Figure 1 Schematic illustration of the determination of practical and potential predictability limits via the mean error growth as a function of time, obtained by calculating the mean root-mean-square error (RMSE) between local dynamical analogs (LDAs). The upper and lower red dashed lines represent the amplitudes of attractor radius (AR) and standard deviation (SD) of observational time series, respectively.

attained when the models are able to realistically predict dynamical systems with very small initial and model errors. The practical predictability limit may be attainable through recently achievable enhancements to existing models of nonlinear dynamical systems. The error growth will enter a strong nonlinear phase with a steadily decreasing growth rate when the magnitude of the mean error exceeds the SD, which is always poorly represented in models of nonlinear dynamical systems. Therefore, modelers undoubtedly face big challenges in improving their models to attain the potential predictability limit.

Predictability is generally dependent on time and space and is controlled by multiple forcing factors, which leads to considerable difficulty in understanding the sources of predictability. Therefore, it is necessary to separately investigate the effect of each forcing factor on predictability. Conditional predictability can be considered as an estimate of the predictability of a certain variable of dynamical systems due to forcing by, or explained by, one or more factors. For example, ENSO is one of the primary boundary forcings for the monthly and seasonal predictability of the atmosphere. The conditional NLE is introduced to study the conditional predictability of the atmosphere associated with ENSO. Let $\mathbf{x}$ and $\mathbf{y}$ be one-dimensional or multi-dimensional vectors of dynamical systems, and assume that the vector $\mathbf{y}$ acts as an external forcing factor for the vector $\mathbf{x}$. To investigate the effect of $\mathbf{y}$ on the predictability of $\mathbf{x}$, the conditional NLE $\lambda_{\mathbf{x}|\mathbf{y}}$ of $\mathbf{x}$ given $\mathbf{y}$ is defined as:

$$\lambda_{\mathbf{x}|\mathbf{y}}[\mathbf{x}_{\mathbf{y}}(t_0), \delta_{\mathbf{x}|\mathbf{y}}(t_0), \tau] = \frac{1}{\tau} \ln \frac{\|\delta_{\mathbf{x}|\mathbf{y}}(t_0 + \tau)\|}{\|\delta_{\mathbf{x}|\mathbf{y}}(t_0)\|}, \quad [3]$$

where $\mathbf{x}_{\mathbf{y}}(t_0)$ denotes the initial state of $\mathbf{x}$ given $\mathbf{y}$, $\delta_{\mathbf{x}|\mathbf{y}}(t_0)$ the initial error, t_0 initial time, τ the evolution time, and $\delta_{\mathbf{x}|\mathbf{y}}(t_0 + \tau)$ denotes the nonlinear error of $\mathbf{x}$ at the time $t_0 + \tau$ under the influence of $\mathbf{y}$. For observational time series of $\mathbf{x}$ and $\mathbf{y}$, linear regression can be used first to extract the components of $\mathbf{x}$ explained by $\mathbf{y}$. Through searching the LDAs between the regressed time series and the original time series of $\mathbf{x}$, the conditional NLE is then estimated by calculating the mean exponential divergence rate of initially analogous points.

The NLE method may be applied to quantitatively estimate the predictability limit of different atmospheric, oceanic, or other climate variables at different timescales. Figure 2 shows the spatial distribution of the annual mean practical predictability limit of daily 500-hPa geopotential height (GHT), as obtained using the NLE method. The practical predictability limit of the National Centers for Environmental Prediction (NCEP)/National Center for Atmospheric Research (NCAR) reanalysis data appears to have a zonal distribution, with a maximum of 14–15 days over the tropics and Antarctic, followed by 13–14 days over the Arctic, 8–13 days over mid-high latitudes in the Northern Hemisphere (NH), and the lowest limit of 7–12 days over midlatitudes in the Southern Hemisphere (SH). The existence of storm tracks and strong baroclinicity over the midlatitudes may explain the relatively low predictability limit in these latitudes. In contrast, the practical predictability limit of the CFSv2 forecasts has a very different distribution, with a maximum limit of 7–9 days over the NH mid-high latitudes, followed by nearly 7 days over the SH mid-high latitudes, and the lowest limit of 3–6 days over the tropics. The results suggest that there is considerable potential for numerical models to increase their predictability through model improvement, especially in the tropics.

Figure 3 shows the spatial distributions of annual mean limits of practical and potential predictability of monthly and seasonal 500-hPa GHT and monthly mean SST. The practical predictability limits of monthly and seasonal mean 500-hPa GHT both appear to be highest over the tropics between 15° S–15° N (a maximum over the Western Pacific warm pool), with values greater than 4 months and 2 seasons, respectively. Correspondingly, the practical predictability limit of monthly mean SST is also the greatest in the tropics, including the tropical Central-Eastern Pacific, and tropical Indian and Atlantic Oceans, where values exceed 8 months (the maximum exceeds 10 months over the tropical Central-Eastern Pacific). However, relatively lower predictability of SST is observed over the Western Pacific warm pool. In comparison, the distributions of the practical predictability limit for monthly mean 500-hPa GHT in the tropics are more uniform than those for monthly mean SST, possibly because the entire tropical atmosphere may respond quickly to SST forcing in the tropical

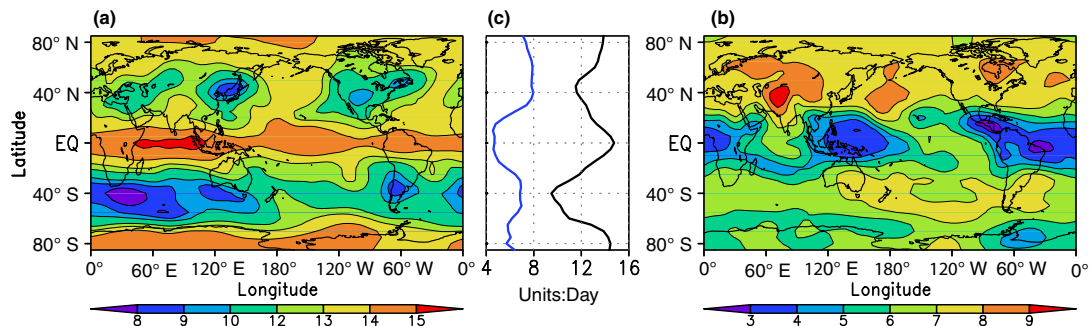


Figure 2 Spatial distribution of the annual mean practical predictability limit (days) of daily 500-hPa geopotential height (GHT), obtained using the NLE method. (a) The NCEP/NCAR reanalysis data (1948–2010). Daily data are obtained as a 1-day running mean of the 6-hourly daily 500-hPa GHT. (b) The NCEP CFSv2 daily forecasts. The predictability in the CFSv2 is estimated by verifying the 45-day retrospective forecasts of CFSv2 against the NCEP/NCAR reanalysis data for the period 1999–2000. (c) The zonal mean profiles of practical predictability limit. The black and blue lines are for the reanalysis and model, respectively.

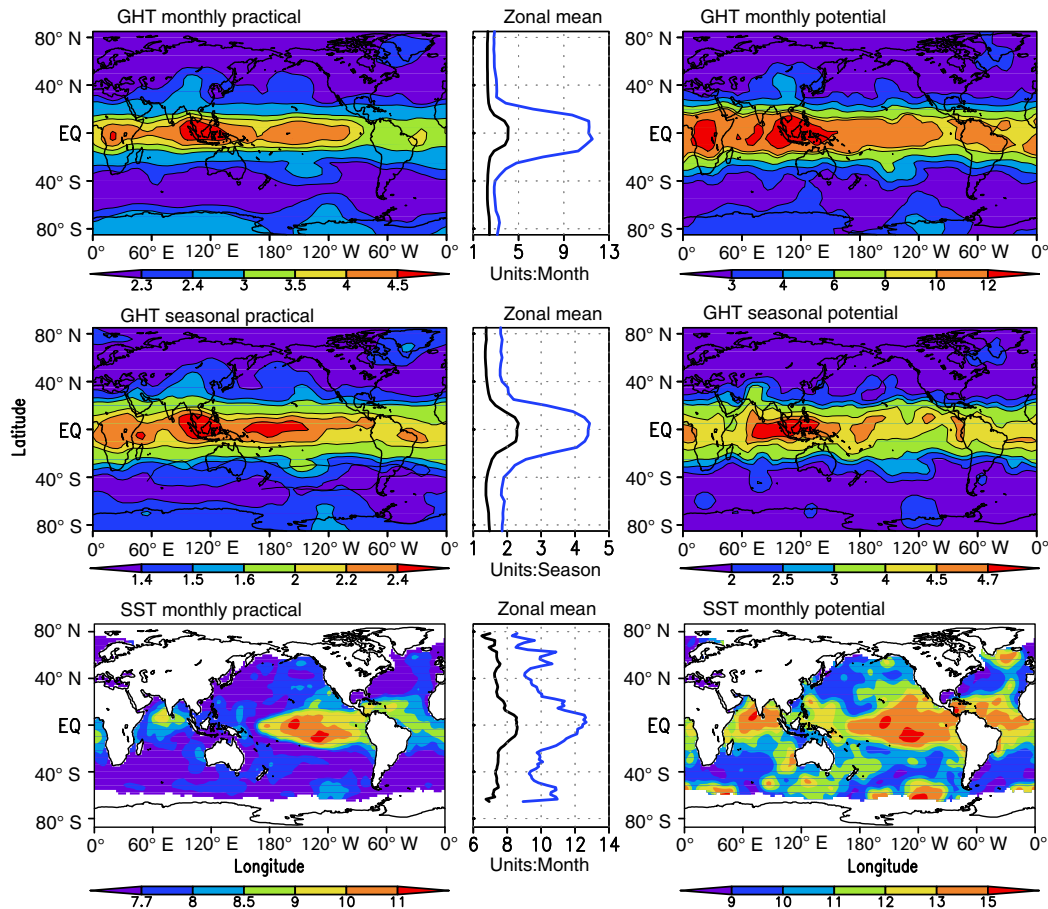


Figure 3 Spatial distribution of the annual mean practical (left panel) and potential (right panel) predictability limits of 500-hPa GHT from the NCEP/NCAR reanalysis (1948–2010) and SST from the HadISST (1900–99) at monthly and seasonal scales and their corresponding zonal mean profiles (middle panel), obtained using the NLE method. Upper, middle, and lower panels are for monthly, seasonal, and monthly scales, respectively, and units in the panels are month, season and month, respectively. The black and blue lines in the middle panels of zonal mean profile are the practical and potential predictability limits, respectively. Monthly (seasonal) data of 500-hPa GHT are obtained as 30-day (90-day) running means of the 6-hourly 500-hPa GHT. Monthly mean SST data are used directly.

Central–Eastern Pacific, and tropical Indian and Atlantic Oceans. The practical predictability limits for 500-hPa GHT and SST gradually decrease from the tropics to mid–high latitudes. At mid–high latitudes, the practical predictability limits of monthly and seasonal mean 500-hPa GHT are only 2 months and 1.5 seasons, respectively, while the limit for monthly mean SST is mostly below 8 months.

Compared with the practical predictability limits, the potential predictability limits of monthly and seasonal mean 500-hPa GHT and monthly mean SST have similar distributions but with much higher values over the tropics. The higher potential predictability for both the atmosphere and SST in the tropics may arise from the strong ocean–atmosphere feedback and slowly varying SST forcing, which can result in a small error growth rate even after a few months.

As mentioned above, ENSO is an important source of atmospheric predictability on seasonal and interannual time-scales. Figure 4 shows the distributions of annual mean limits of practical and potential predictability for monthly 500-hPa GHT explained by the ENSO index, obtained using the

conditional NLE. Both the ENSO-related practical and potential predictability limits are largest in the tropics and exceed 2.5 months. In particular, the two regions with the highest predictability limits are the Central–Eastern equatorial Pacific and Western Pacific warm pool, where the two corresponding poles of the Southern Oscillation are located. There, over 50% of the monthly atmospheric predictability is related to monthly variations in tropical Pacific SST. In contrast, ENSO makes little contribution to predictability outside the tropics, other than along the two great circle routes spanning the Pacific–North American (PNA) and Pacific–South American (PSA) teleconnection patterns, where about 1 month of practical predictability and 2 months of potential predictability are related to ENSO. The stronger response to ENSO of the PSA pattern than the PNA pattern results in a higher conditional predictability in the PSA region than in the PNA region. It is these two teleconnection patterns, constructing a remote relationship between mid–high latitude atmospheric circulation and tropical ENSO-related SST, that provide useful information for seasonal and interannual predictions over North and South America.

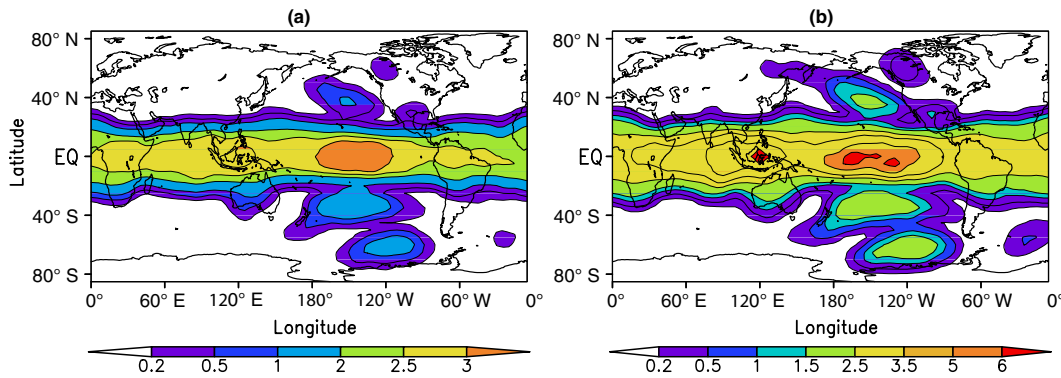


Figure 4 Spatial distributions of the annual mean practical (a) and potential (b) predictability limits (months) of monthly 500-hPa GHT explained by the ENSO component, obtained using the conditional NLLE method.

Approaches to Seasonal and Interannual Predictions

Currently, seasonal to interannual climate predictions are based on dynamical and empirical approaches as well as combinations of such approaches. In the past, statistical approaches were most commonly used for climate predictions. However, with the rapid development of computing technologies and great improvements in model parameterization, dynamical approaches have become effective tools for climate predictions. To date, operational climate prediction products from WMO GPCs of Long-Range Forecasts (e.g., European Centre for Medium-Range Weather Forecasts (ECMWF) and NCEP's Climate Prediction Center (CPC)) are based on dynamical approaches, which are likely to be the main method of climate prediction in the coming decades.

Dynamical seasonal prediction approaches utilize GCMs, including atmospheric GCMs, coupled ocean–atmosphere GCMs, or more complex coupled climate system models that consider the roles of atmosphere, ocean, land surface, and sea ice. External boundary forcing conditions and the extent to which the boundary conditions can be predicted are important factors for successful climate predictions by dynamical models. SSTs, especially tropical Pacific SSTs, are one of the most important boundary conditions for atmospheric models, since the predictability of SSTs in the tropics is higher than that in other regions, as mentioned before. The highly predictable SSTs in the tropics contribute to the high predictability limit of the overlying atmosphere through strong air–sea interaction, which is essential for predicting the ENSO cycle. Another factor that influences model predictions is uncertainties in the initial conditions. Since the climate system is strongly nonlinear, regional or global climate systems may contain multiple equilibria coexisting under the same external forcing, and the final equilibrium state depends on the initial conditions. Consequently, climate prediction models must consider errors in the initial and boundary conditions in coupled GCMs. In addition, the reference climatology used for the definition of climate anomalies, which is generally the most recent 30-year average, is also an impact factor that influences predictions. This factor is also relevant to empirical approaches to climate prediction.

Dynamical models used for seasonal predictions can be classified as simple models, intermediate-complexity coupled models, hybrid coupled models, and coupled GCMs,

depending on the degree of complexity of the models. Simple dynamical models are the simplest of these models and are useful tools for studies of climate dynamics. Such models have the potential to yield skillful predictions, as the validity of the prediction is not entirely dependent on the degree of complexity of the models. To reduce the uncertainty associated with errors in the initial conditions, ensemble prediction techniques, which conduct multiple numerical predictions using slightly different initial conditions to attempt to generate a representative prediction, are employed. To reduce uncertainties associated with the physical parameterization scheme, which differs among models, multimodel ensemble (MME) techniques are employed and have proven to be extremely effective. MME approaches are the best currently available for quantifying and resolving uncertainties, and have proven to produce better predictions than any single model alone.

Dynamical approaches have a number of advantages. They are not only unconstrained by linear relationships between the predictors and predictands of interest, but can be used to predict unprecedented conditions, especially for a nonstationary climate. In addition, dynamical approaches can help us to understand the physical mechanisms that underlie the statistical relationships established by purely empirical approaches. However, dynamical approaches are limited by intensive computing requirements and large model biases in physical parameterizations, which limit the seasonal forecast skill. Moreover, dynamical approaches also suffer from errors introduced because of imperfections in the model formulation as well as the numerical methods to approximately solve the evolution equations of the dynamical system. Empirical approaches are sometimes used as a supplement to dynamical approaches.

Empirical approaches, which are the simplest way to make climate predictions, fall into two categories: purely statistical relations between possible predictors and predictands of interest, and physically based linkages between predictors and predictands based on an understanding of the underlying physical mechanism. The latter methods have a physical basis and generally show better forecasting performance. Many statistical methodologies, including linear and nonlinear methods, have been applied to identify physical links between predictors and predictands. A climate variable is usually characterized by multitimescale variabilities that are controlled or driven by different climate factors at different timescales.

Consequently, a timescale decomposition (TSD) downscaling approach is used for statistical downscaling, which estimates how local or regional climatic variables (e.g., the temperature, precipitation, and wind speed at a particular site or over a region) are affected by large-scale climatic conditions (e.g., atmospheric surface pressure, GHT, SST, ENSO), in order to separate predictands with distinct multitimescale features into different timescale components that are associated with different predictors. The TSD model makes use of different downscaling models that correspond to different timescale variability linked to various mechanisms. The TSD model has been successfully applied to seasonal predictions of summer (July–August) rainfall over North China.

There are some quite successful cases of empirical predictions. For example, the connections the East Asian summer monsoon (EASM) with ENSO and spring (April–May) North Atlantic Oscillation (NAO) have been established. The EASM variability is strongly linked to not only the winter (December–January–February (DJF)) ENSO but also the spring NAO. NAO anomalies in spring can trigger a tripole SST pattern in the North Atlantic that persists into the subsequent summer. The tripole SST pattern in summer impacts EASM through two pathways: (1) it can excite an Atlantic–Eurasian (AEA) teleconnection pattern over mid–high latitudes, which may strengthen (or weaken) the East Asian subtropical front (Meiyu–Baiu–Changma), leading to a strong (or weak) EASM; and (2) the tropical Atlantic can strongly modulate the western subtropical high, which is an important linkage between the EASM and the DJF ENSO. Based on the underlying physical mechanism, an empirical NAO–ENSO-based seasonal prediction model of the EASM has been constructed and has been shown to produce skillful seasonal predictions of the EASM and summer rainfall over the Yangtze River valley. Another example of a successful empirical prediction is the connection between the spring (March–April–May) precipitation in the Nordeste region of Brazil and tropical Atlantic SST in the months before the rainy season.

Empirical approaches to seasonal predictions have a number of advantages. Compared with dynamical approaches, empirical approaches are easy to apply, as they are much less computationally expensive. In addition, they are unaffected by model errors in parameterizations. However, there are some limitations in empirical approaches. The relationships between predictors and predictands established by such approaches are usually linear, which ignores the nonlinear processes between them. Although some nonlinear statistical methods can be used, they are prone to overfitting noise. Furthermore, empirical approaches might have difficulty predicting abrupt changes in climate, and they are limited by the length of time series of observed data. Nonetheless, recent investigations show that the reliability of physically based empirical approaches has been enhanced by an improved understanding of the physical and dynamical processes that underlie the way in which forcing (predictors) impacts on predictands.

Climate predictions at an interannual timescale have received increasing attention in recent years. Such predictions bridge the gap between seasonal and decadal predictions, and have the potential to provide information on the climate conditions and climate risks in upcoming years, which would be valuable for decision making in various sectors such as

agriculture, energy, economics, and infrastructure. However, in contrast to seasonal predictions, interannual climate predictions are at an early stage of development, and the effective sources and limits of interannual predictability remain unclear. As the strongest signal of interannual climate variability, ENSO can be predicted only at lead times of up to 1 year, beyond which predictions become meaningless. In addition, many potential sources of predictability are not considered in current models and approaches. Thus, the refinement of interannual predictions faces great challenges. In the context of current efforts to improve interannual climate predictions, it is necessary to examine the sources and limits of interannual predictability. Much work is required to improve the level of prediction skill on such a timescale.

Current Practice in Seasonal and Interannual Predictions

Due to the importance of climate predictions in long-term decision making, and the possibility that such predictions can provide an early warning of potential future hazards, climate prediction products have become the focus of increasing public interest. Climate predictions can be produced at global, regional, or local scales to serve users with different purposes; however, their use must take into account the predictability limits of the predictands. That is, predictions that are made beyond the predictability limits of the predictands have no practical value.

There are several centers currently generating global climate predictions, and these include the WMO GPCs of Long-Range Forecasts; e.g., CPC/NCEP/National Oceanic and Atmospheric Administration (NOAA), ECMWF, the Australian Bureau of Meteorology, the Beijing Climate Center of China Meteorological Administration, and the UK Met Office. Regional climate prediction products are produced by the WMO Regional Climate Centers and also the Regional Climate Outlook Forums with the aim of delivering better climate services and improving the ability to meet national climate information needs related to reducing climate-related risks and supporting sustainable development. In addition, many nations generate national climate prediction products at the seasonal scale, which are the responsibility of the national meteorological and hydrological services.

Seasonal prediction products from the CPC and ECMWF are generated using their own forecasting systems. The CPC uses the so-called US National MME (NMME), which is an experimental multimodel seasonal forecasting system comprising coupled models from US modeling centers, including NOAA/NCEP, NOAA/Geophysical Fluid Dynamics Laboratory (GFDL), International Research Institute for Climate and Society (IRI), NCAR, National Aeronautics and Space Administration (NASA), and Canada's Canadian Meteorological Center (CMC). The ECMWF employs the European multimodel seasonal to interannual prediction (EUROSIP) system, which is based on coupled models from the ECMWF, UK Met Office, Météo-France, and NCEP. The CPC seasonal predictions are issued with probabilities calibrated against past cases by statistical forecast tools, but the ECMWF predictions are not calibrated in this way, primarily because of the limited number of past predictions.

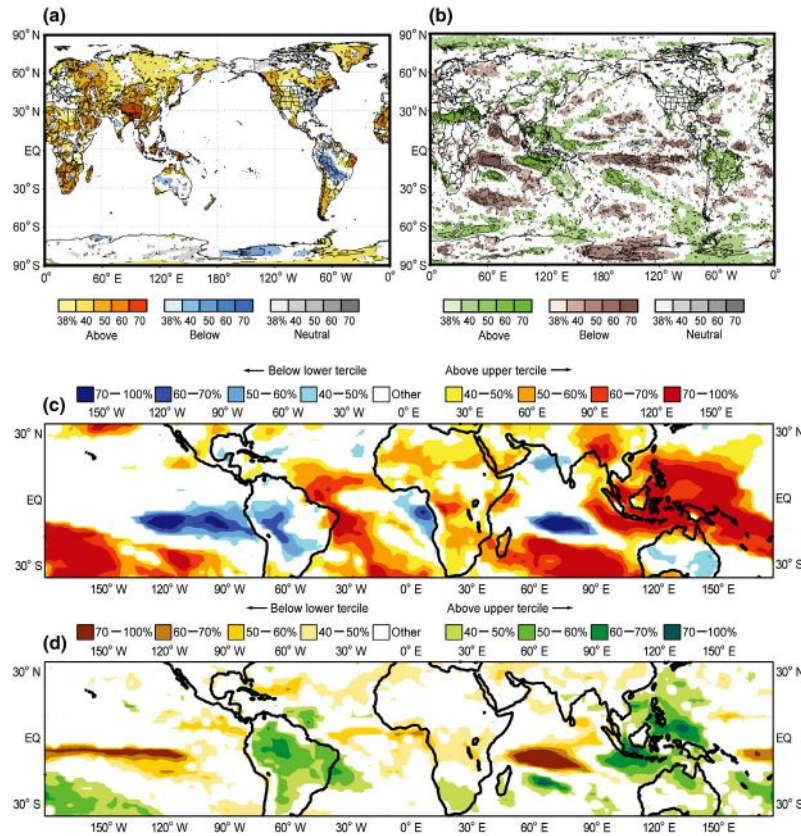


Figure 5 Examples of MME tercile categorical seasonal prediction. (a) Global seasonal mean 2 m temperature. With permission from CPC NMME. (b) As for (a) but for precipitation. (c) As for (a) but for the tropics. With permission from ECMWF EUROSIP. (d) As for (c) but for precipitation. With permission from ECMWF EUROSIP. These prediction products are based on a start date of July 2013 and show the 3-month outlooks (August–September–October (ASO) 2013) for a lead time of 1 month. The expected changes in probabilities (%) of three categories are shown: above normal, below normal, and neutral for CPC, and below lower-tercile, above upper-tercile, and between the lower and upper terciles for ECMWF.

Figure 5 shows some examples of MME tercile categorical seasonal predictions by the CPC and ECMWF for 2 m temperature and precipitation. The predictions from the CPC were produced at the given lead times from 1 to 5 months, but for the ECMWF ranged from 1 to 3 months. The CPC produces global seasonal predictions (predictions for 2 m temperature are available over the land), while the ECMWF predictions only cover the tropics (possibly due to the higher predictability in the tropics), implying a difference in seasonal prediction products between the ECMWF and CPC. For this case, the two centers show a similar seasonal probabilistic prediction of the mean 2 m temperature in the tropics over land; i.e., a warming trend in Africa, East Asia, and the Maritime continent, and a cooling trend in South America. The seasonal precipitation predictions over the tropics from the two centers are also similar, but for the region from West Africa to Arabia, there is a wetting trend in the CPC prediction, but a neutral condition in the ECMWF prediction.

Verification of these predictions is required to provide an indication for users of the extent to which the predictions are credible. Figure 6 is an example of the verification of the seasonal precipitation predictions for the continental United States at a lead time of 0.5 months using the Heidke Skill Score from the CPC. A score of 100 denotes a perfect prediction,

a score of -50 indicates a completely incorrect prediction, and 0 denotes a random prediction. As shown in Figure 6(a), although there are a few cases with higher scores, the mean value of all manuals is less than 5.0, indicating the low skill of the seasonal prediction. There are seasonal fluctuations in this verification; i.e., a higher score in winter–spring and a lower score in summer–autumn. Seasonal variations in the predictability of seasonal precipitation may account for this, and the more deep-rooted causes of such fluctuations are related to seasonal changes in the predictability of predictors. Furthermore, the skill score shows a distinct spatial distribution (Figure 6(b) and 6(c)); i.e., higher scores in southern continental United States, and lower scores in northern continental United States.

ENSO is one of the key prediction objectives in operational seasonal predictions. Figure 7 shows two examples of the seasonal prediction products of Niño3.4 SSTs (5° N– 5° S, 120° – 170° W) from the ECMWF EUROSIP and CPC NMME. The EUROSIP and NMME produce predictions for up to 6 and 8 months ahead, respectively. For the predictions at the start time of September 2012, the two predictions have large uncertainties, and their MMEs have larger prediction errors, showing the low skill of the predictions for this case. The observational Niño3.4 index reaches a peak in August 2012, and enters its

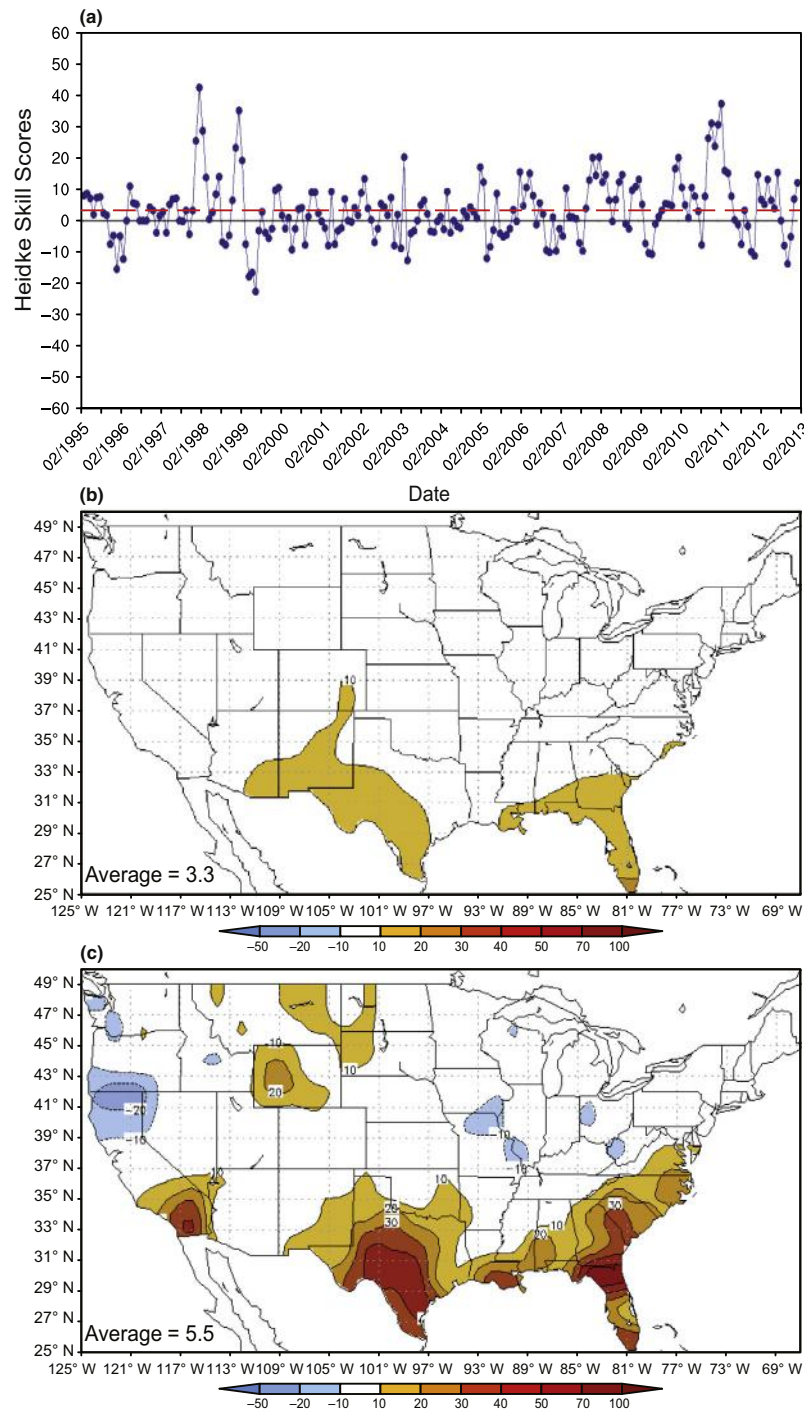


Figure 6 Example of verification of seasonal precipitation prediction for the continental United States at a lead time of 0.5 months using the Heidke Skill Score (ranging from -50 to 100) from the CPC. (a) Seasonal precipitation Heidke Skill Score for all categories from January 1995 to March 2013. The blue dashed line is the average skill score. (b) and (c) are the distributions of the skill score of all manuals, and the November–December–January (NDJ) manuals, respectively.

negative phase in September 2012. However, the two predictions did not capture this phase transition. This example shows that against a background of ongoing climate change, the complexity of the climate system may increase, which will make it increasingly difficult to predict ENSO, even though it has

a higher predictability. The predictions at the start time of July 2013 from the two systems suggest that the Niño3.4 index may go to a positive phase during the next half year.

Physically based empirical approaches are often used to develop seasonal prediction products. For example, the TSD

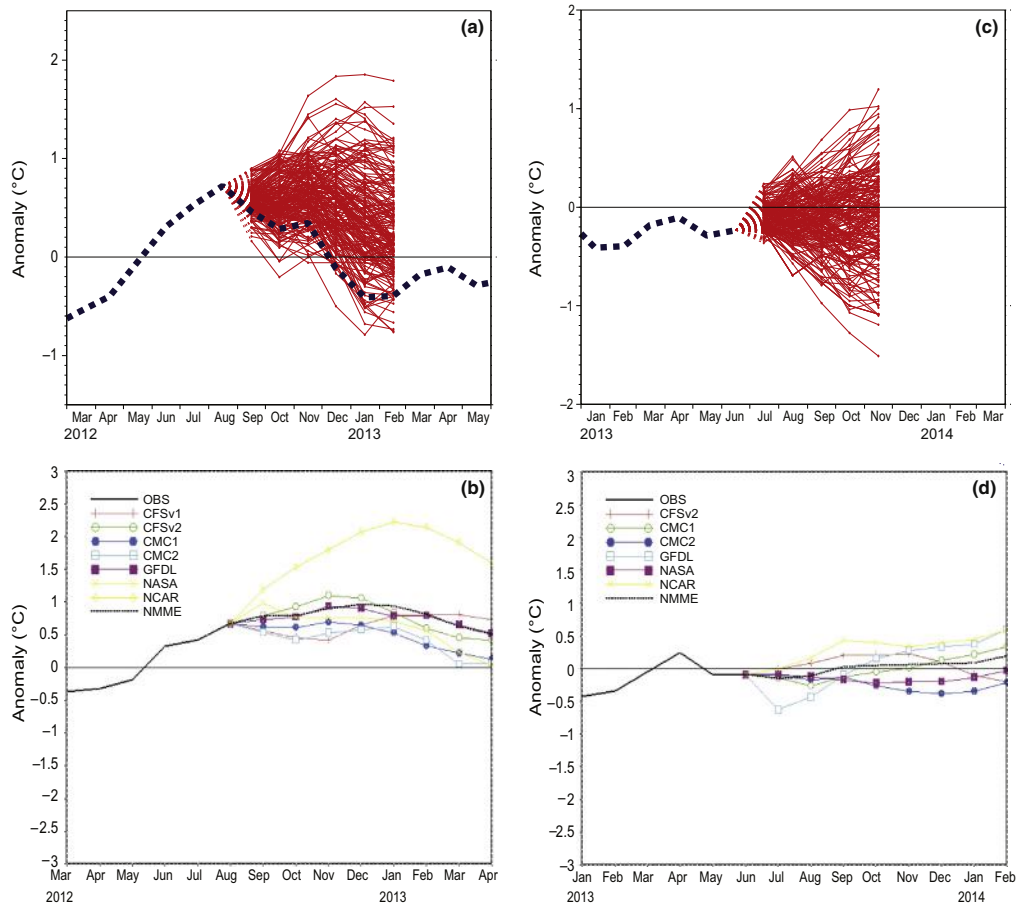


Figure 7 Examples of seasonal forecast of Niño3.4 SSTs (5°N – 5°S , 120° – 170°W). (a, c) with permission from ECWMF EUROSIP and NMME (b, d) with permission from CPC NMME. (a) and (b) indicate the predictions made at a start time of September 2012. (c) and (d) denote the predictions made at a start time of July 2013. Predicted monthly mean anomalies from each individual ensemble member of EUROSIP (NMME) are relative to the NCEP Olv2 1981–2010 climatology (OISST, 1982–2010 climatology) and shown as red dotted lines (colored lines). The observations are represented by the thick dashed blue lines in (a) and (c), and black solid lines in (b) and (d). The dashed black lines in (b) and (d) indicate the MMEs of all members.

downscaling model is used in the seasonal prediction of summer (July–August) rainfall over North China. This model of interannual summer rainfall variability over North China is linked to the June ENSO and June AEA teleconnection, while that for interdecadal summer rainfall variability over North China is related to the decadal variability of sea level pressure over the southwest Indian Ocean. Taking the downscaled interannual and interdecadal components together, the downscaled total rainfall was obtained. The independent validation shows a better prediction skill for the TSD approach. The TSD model was used by the National Climate Center of the CMA to make seasonal predictions of summer rainfall over North China in 2010, 2011, and 2012, and these predictions were consistent with observations. The summer rainfall anomaly for 2013 over North China by the TSD model was above normal (not shown).

Further successful examples of empirical predictions are the empirical NAO–ENSO-based seasonal prediction models of the EASM index (EASMI) and summer precipitation (RI_{YR}) over the middle and lower reaches of the Yangtze River valley, as mentioned above. These empirical seasonal prediction models

were derived using a linear regression method for the period 1979–2006, as follows:

$$\begin{aligned} \text{EASMI} = & 11.56 + 0.60 \cdot \text{NAOI} + 0.22 \cdot \text{ENSO}_{\text{develop}} \\ & - 0.43 \cdot \text{ENSO}_{\text{decay}}, \end{aligned} \quad [4]$$

$$\begin{aligned} RI_{YR} = & -5.92 - 0.41 \cdot \text{NAOI} - 0.04 \cdot \text{ENSO}_{\text{develop}} \\ & + 0.22 \cdot \text{ENSO}_{\text{decay}}. \end{aligned} \quad [5]$$

Here, NAOI denotes the mean NAO index value for April–May, $\text{ENSO}_{\text{develop}}$ the Niño3.4 index difference between April–May and February–March, and $\text{ENSO}_{\text{decay}}$ the Niño3.4 index for the preceding winter (DJF). The seasonal predictions of the EASMI and RI_{YR} generated using these models (Figure 8) show that the performance of seasonal predictions for both EASMI and RI_{YR} between 2007 and 2012 is good, implying that the empirical models are effective tools for the seasonal prediction of EASM. The predictions for 2013 by the models show a somewhat stronger EASM, and reduced summer precipitation over the middle and lower reaches of the Yangtze River valley.

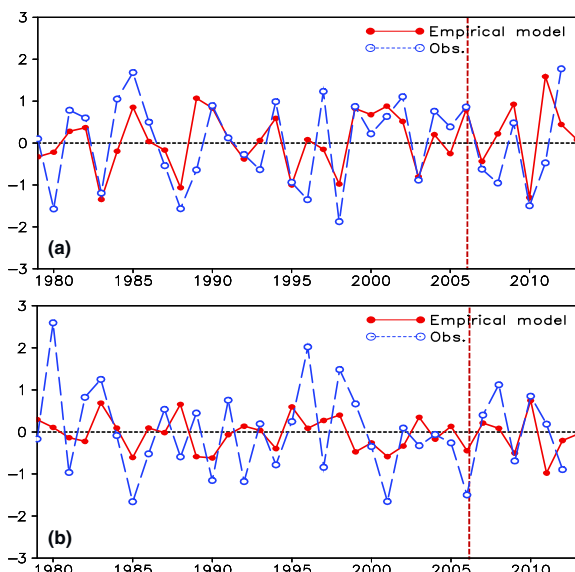


Figure 8 Example of seasonal prediction of the EASMI (a) and summer (JJA) precipitation $R1YR$ over the middle and lower reaches of the Yangtze River valley (b) using the NAO–ENSO-based seasonal prediction model. All values are normalized. The observations are indicated by dashed lines, and the model results by solid lines. The model fit period is 1979–2006, and the prediction period is 2006–13.

Interannual predictions are extremely difficult and relevant research in this area remains limited; consequently, very few interannual prediction products are currently available. Therefore, additional research is required to develop such products in the future.

See also: **Air Sea Interactions:** Sea Surface Temperature. **Climate and Climate Change:** Climate Prediction: Empirical and Numerical; Climate Variability: Nonlinear and Random Effects; Climate Variability: Seasonal and Interannual Variability; Overview. **Data Assimilation and Predictability:** Ensemble Prediction; Predictability and Chaos. **General Circulation of the Atmosphere:** Overview; Weather Regimes and Multiple

Equilibria. **Land-Atmosphere Interactions:** Overview. **Numerical Models:** General Circulation Models; Methods; Regional Prediction Models. **Tropical Meteorology and Climate:** El Niño and the Southern Oscillation: Observation; El Niño and the Southern Oscillation: Theory.

Further Reading

- Ding, R.Q., Li, J.P., 2007. Nonlinear finite-time Lyapunov exponent and predictability. *Physics Letters A* 364, 396–400.
- Goddard, L., Mason, S.J., Zebiak, S.E., et al., 2001. Current approaches to seasonal-to-interannual climate predictions. *International Journal of Climatology* 21, 1111–1152.
- Guo, Y., Li, J.P., Li, Y., 2012. A time-scale decomposition approach to statistically downscale summer rainfall over North China. *Journal of Climate* 25, 572–591.
- Hoerling, M.P., Kumar, A., 2003. Weather prediction-seasonal and interannual weather prediction. In: *Encyclopedia of Atmospheric Sciences*, first ed. Academic Press, London.
- Kirtman, B., Pirani, A., 2009. The state of the art of seasonal prediction: outcomes and recommendations from the First World Climate Research Program Workshop on seasonal prediction. *Bulletin of the American Meteorological Society* 90, 455–458.
- Kumar, A., Chen, M.Y., Wang, W.Q., 2013. Understanding prediction skill of seasonal mean precipitation over the tropics. *Journal of Climate* 26, 5674–5681.
- Li, J.P., Chou, J.F., 1997. Existence of atmosphere attractor. *Science China (Series D)* 40, 215–224.
- Li, J.P., Ding, R.Q., 2011. Temporal-spatial distribution of atmospheric predictability limit by local dynamical analogues. *Monthly Weather Review* 139, 3265–3283.
- Li, J.P., Ren, R.C., Qi, Y.Q., et al., 2013. Progress in air–land–sea interactions in Asia and their role in global and Asian climate change. *Chinese Journal of Atmospheric Sciences (in Chinese)* 37, 518–538.
- Lorenz, E.N., 1993. *The Essence of Chaos*. University College London Press, London.
- Mu, M., Duan, W.S., Wang, B., 2003. Conditional nonlinear optimal perturbation and its applications. *Nonlinear Process Geophysics* 10, 493–501.
- National Research Council, 2010. *Assessment of Intraseasonal to Interannual Climate Prediction and Predictability*. The National Academies Press, Washington, DC.
- Robock, A., 2001. Stratospheric forcing needed for dynamical seasonal prediction. *Bulletin of the American Meteorological Society* 82, 2189–2192.
- Stockdale, T.N., Alves, O., Boer, G., et al., 2010. Understanding and predicting seasonal-to-interannual climate variability – the producer perspective. *Procedia Environmental Science* 1, 55–80.
- Wu, Z.W., Wang, B., Li, J.P., Jin, F.-F., 2009. An empirical seasonal prediction model of the East Asian summer monsoon using ENSO and NAO. *Journal of Geophysical Research* 114, D18120.

Severe Weather Forecasting

DJ Stensrud, HE Brooks, and SJ Weiss, National Oceanic and Atmospheric Administration, Norman, OK, USA

© Published by Elsevier Ltd.

Synopsis

Severe weather is any type of violent or extreme weather event that represents a hazard to public safety and welfare, including tornadoes, damaging windstorms, flash floods, lightning, and hailstorms. The societal costs of severe weather events are very high, involving injuries, lost lives, and property damage. Accurate severe weather forecasts help reduce the loss of life and mitigate economic loss. The severe weather forecasting process requires knowledge of climatology and incorporates the techniques of pattern recognition and parameter evaluation as applied to both observations and numerical weather prediction model forecasts.

Introduction

Severe weather is any type of violent or extreme weather event that represents a hazard to public safety and welfare. In the United States, the country that arguably has the largest severe weather threat, the term severe weather is applied only to those events associated with severe local thunderstorms, such as tornadoes, hail, lightning, and damaging surface winds. In other countries, it is more common to include hurricanes, flash flooding, fire weather, or winter storms in the definition of severe weather, since these weather-related events mostly affect public welfare. Hurricanes, in particular, are a significant weather hazard, but owing to their longer lifetimes and limited geographic region of influence, they are not considered here. The specific events that are covered under the umbrella of severe weather are not as important as the recognition that weather events influence our daily lives and can be both a blessing and a hazard.

The human and economic costs of severe weather events can be very high, involving injuries, lost lives, and significant property damage. Accurate severe weather forecasts help reduce the loss of life and mitigate economic loss by encouraging people to seek shelter or limit their exposure to the threat, by encouraging businesses to alter their plans, and by helping governments to prepare to respond to any potential damage and requests for assistance. It is estimated that over 15 000 deaths occur annually around the world as a result of severe weather. The National Climatic Data Center reports 15 tornado outbreaks in the United States that individually produced over \$1 billion in damage over a 10-year period ending in 2009. Flash floods are a significant hazard in Europe, with numerous individual events in the past decade again causing over \$1 billion in damage. Damage from the wildfires in Russia during August 2010 is estimated to be in excess of \$15 billion. Increasing worldwide population, urbanization, and greater dependence on complex technology-based infrastructure suggest that our risk exposure to the effects of severe weather is increasing. A single severe weather event that damages a critical piece of infrastructure could negatively impact an area for a long period of time.

Global estimates of the severe weather threat can be produced from reanalysis data. The conditions that favor significant severe thunderstorms are identified in the data for locations across the globe and the frequency at which these

favorable conditions occur is calculated. The resulting map highlights regions with the most persistent severe weather potential (Figure 1), such as central North America, southern South America, western and southern Africa, the Mediterranean region of Europe, southern and eastern Asia, and eastern Australia. While these areas have higher severe weather frequency, reports indicate that severe weather occurs on every continent except Antarctica, so no country is completely safe from the threat.

The global nature of the severe weather threat, and the disruption it produces, further underscores the value of providing severe weather forecasts that allow for actions to be taken to protect life and property. The forecast process is difficult, however, because severe weather events often occur on very small scales that are not well observed and the physical processes that produce severe weather are not completely understood. This combined lack of observational detail and physical understanding is partially compensated for by the

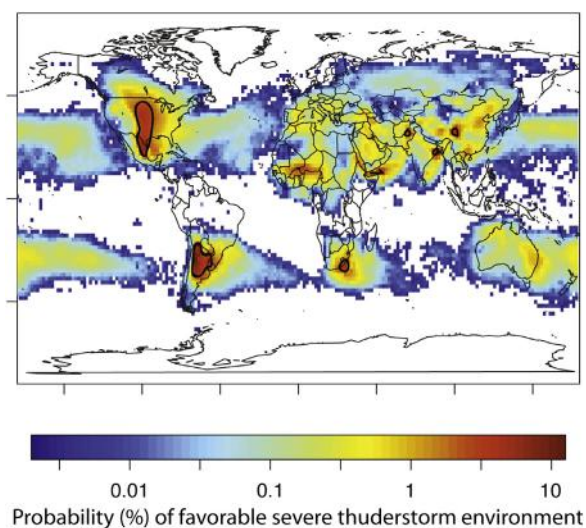


Figure 1 The yearly probability (%) of environments favorable for severe thunderstorms calculated from global reanalysis data over the years 1948–2008. Areas surrounded by a thick black line have a 5% or larger probability of favorable environments for severe thunderstorms, which equates to an average of 18 days per year.

presence of repeatable synoptic-scale atmospheric features for many of these events. Thus, the forecasting of severe weather is based largely upon knowledge of climatology, pattern recognition, and meteorological parameter evaluation, in which the human forecaster plays the crucial role. Pattern recognition and parameter evaluation approaches are applied to both observational data and numerical weather prediction model forecasts. The recent availability of forecasts from convection-allowing numerical weather prediction models, which can predict individual thunderstorms and their evolution, shows great promise although the optimal role these predictions will play in the forecast process is still being evaluated.

The primary missions of the Storm Prediction Center, a national center within the US National Weather Service, are to forecast severe weather events associated with thunderstorms and to provide fire weather outlooks. The severe weather forecasting mission has been part of the weather service since 1952 and is supported because of the rarity of severe weather events in any one locality and the recognition that these events span geopolitical boundaries. In some regions of the country, a local weather forecaster may experience a significant severe weather event once during an entire career. This requires the forecaster to make all the right decisions, based upon limited understanding and limited observations, the very first time a significant threat is encountered. In contrast, a forecaster at a center with national responsibility experiences significant severe weather events on many days each year. Thus, the experience base upon which to guide decisions grows rapidly and can be shared easily with local forecasters and the general public. Countries of smaller physical size than the United States may benefit greatly from a multicountry severe weather forecasting center.

In this article, the process of forecasting severe weather events prior to and during their development is discussed. This is not the same process as nowcasting, in which the present weather situation is monitored and linear extrapolation is used to predict future events as illustrated in Figure 2. The forecasting process does not necessarily involve a longer time scale,

but nonlinear methods and forecaster experience are used in the process. Severe weather forecasts typically extend from 0 h out to a week. Short-range forecasts (0–24 h) provide detailed guidance on the expected hazards, their location, and evolution. These forecasts use both observational data and numerical weather prediction model output in their creation. Forecasts for the next day and extending out to a week provide less detailed guidance, such as the expected coverage and intensity of the threat, and are based almost entirely upon output from numerical weather prediction models. While output from numerical weather prediction models contributes significantly to the forecast process, one should never lose sight of the fact that the human forecaster makes the process successful.

Elements of Forecasting

The forecasting process typically begins by defining a broad region of potential severe weather threat and then refining this area of potential threat so that the final forecast product is as specific as possible. This forecasting process requires knowledge of climatology and incorporates the techniques of pattern recognition and parameter evaluation. We assume that the forecaster has access to a standard suite of meteorological observations, including at a minimum data from surface stations and upper-air soundings, and also has access to numerical weather prediction model output. Real-time data from radars, satellite, radar wind profilers, lightning detection systems, and aircraft are also very useful. Observational data are typically analyzed first, to develop a picture of the atmosphere at the most recent observational time, followed by an analysis of numerical weather prediction model data to help gain an appreciation of possible scenarios and extend the forecast into the future.

Climatology

One of the first elements used in the forecasting of severe weather is knowledge of the climatology of the various types of severe weather events. Climatology increases awareness of where and when severe weather has occurred in the past, so that the forecaster is prepared and looking for indications of a severe weather threat on days when severe weather is known to occur. A comparison of the patterns of yearly cycles of mean monthly tornado and heavy rainfall (25.4 mm h^{-1} , which is used as a proxy for flash floods) over the United States indicates that the regions of highest probabilities for these two events are different (Figures 3 and 4). The area of highest heavy rainfall frequency is located to the south, southeast, or east of the area of highest tornado frequency. The event frequencies also are different, with heavy rainfall much more common than a tornado within 40 km of a given location. Knowledge of these patterns and their differences helps forecasters better anticipate when severe weather is most likely. There also are days when the potential for severe weather is high but the expected type of severe weather event (e.g., tornadic supercell or damaging windstorm) is unclear. Climatological information can help determine which event is most likely on a given day. However, the usefulness of climatology is dependent on the robustness of the severe weather reports used to create the climatology. Since

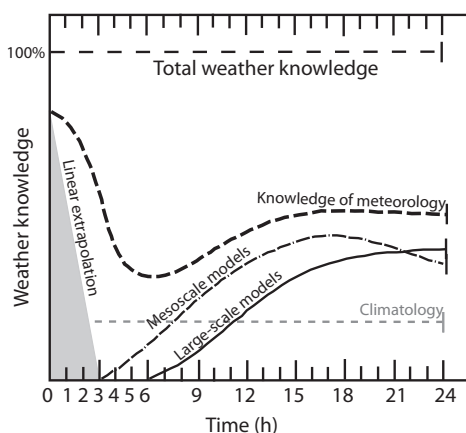


Figure 2 Schematic of the utility of various techniques used in weather forecasting as a function of time. After Doswell III, C.A., 1986. Short-range forecasting. In: Ray, P. (Ed.), *Mesoscale Meteorology and Forecasting*. American Meteorological Society, Boston, pp. 689–719.

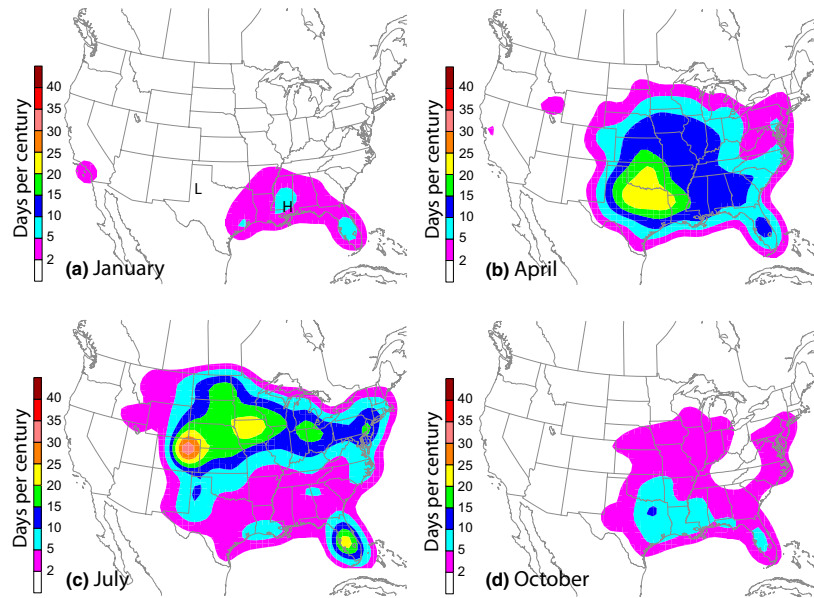


Figure 3 Tornado days per century over the United States for the months of (a) January, (b) April, (c) July, and (d) October, illustrating the seasonal cycle of tornadic activity. Fifteen years of data were used to create these analyses. Locations of Lubbock, TX (L), and Hattiesburg, MS (H), are indicated by the letters in (a).

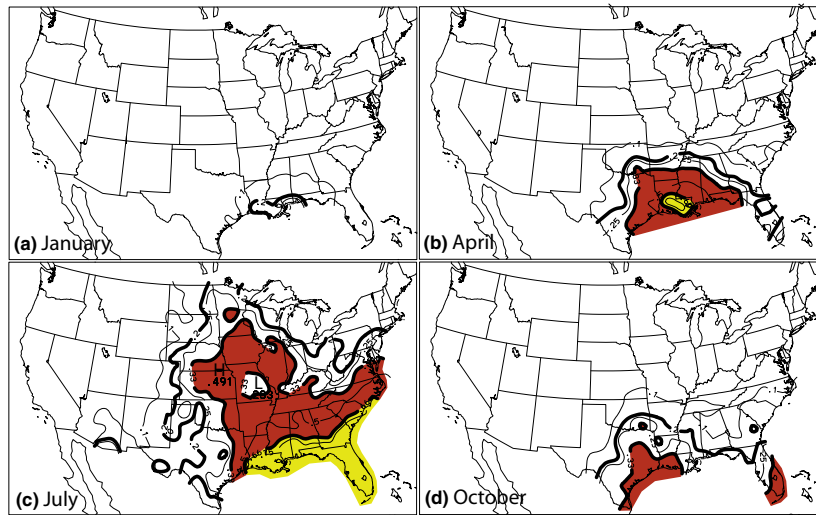


Figure 4 Isolines of 25.4 mm h^{-1} rainfall events per year over the United States for the months of (a) January, (b) April, (c) July, and (d) October, illustrating the seasonal cycle of heavy rainfall. Isolines of 0.1, 0.2, 0.25, 0.33, 0.5, 0.66, and 0.75 events per year are shown. Areas with greater than 0.33 events per year are shown in red, whereas areas with greater than 0.66 events per year are shown in yellow. Forty years of data were used to create these analyses. From Brooks, H.E., Stensrud, D.J., 2000. Climatology of heavy rain events in the United States from hourly precipitation data. *Monthly Weather Review* 128, 1198–1199.

many severe weather events can be identified only by a human being watching the event unfold, public recognition and knowledge of severe weather influences the accuracy of the climatology. If the public is told that hail never occurs in their city, then it is very likely that if a hail event happens either it will not be reported or the report will not be accepted as accurate.

Climatological analyses further can be used to address the variability in the timing and location of severe weather events.

Annual cycles of various severe weather events can be generated for specific locations. These cycles highlight both the year-to-year reliability and the year-to-year variability of the severe weather threat at different locations (**Figure 5**). This information helps forecasters to anticipate when severe weather is more likely and also helps them to provide information to the public about why an apparently unlikely event – such as a tornado during the winter in the state of Mississippi (**Figure 5(b)**) – is not unusual. This region of the United States has

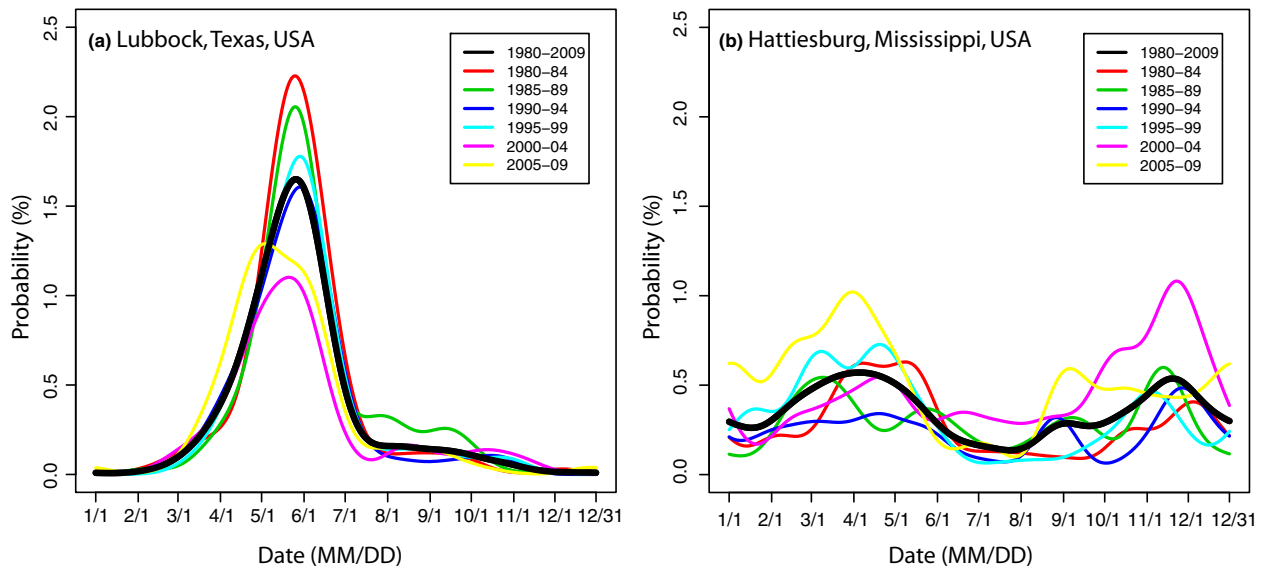


Figure 5 Probability (%) of a tornado within 40 km for (a) Lubbock, Texas, USA, and (b) Hattiesburg, Mississippi, USA, plotted versus day of the year (month/day). The probabilities were calculated using six different 5-year data sets, illustrating the robustness or the variation of the threat. The black line is the mean probability for the complete 30-year period. Locations of Lubbock and Hattiesburg shown in [Figure 3\(a\)](#).

a persistent low level of threat that peaks slightly in the early summer and the winter.

Pattern Recognition

Many repeatable atmospheric features have been associated with severe weather events. These synoptic-scale atmospheric patterns are observed by our present data networks and are used routinely by forecasters as the next step in assessing the severe weather threat. Composite charts, in which specific features found at different pressure levels within the troposphere are plotted on a single chart, are often used. These presentations allow forecasters to view the atmosphere in a three-dimensional perspective and can highlight the vertical relationships between synoptic features and selected parameter fields considered important for the type of event being forecast. Composite charts plotted for typical tornadic thunderstorm, damaging windstorm (derecho), and flash flood days illustrate that the synoptic and mesoscale features associated with these events are different, although there are also some similarities ([Figure 6](#)). Individual events do not always contain all of the features seen in these composite analyses, but a number of these features are likely to be observed. Composite analyses assist in defining the broad region of potential threat, which often can cover a large geographic region. More information is required to then develop the tailored forecast product the public needs.

Parameter Evaluation

Parameter evaluation techniques help to refine the region of severe weather threat determined from pattern recognition. These techniques are based upon our understanding of the physical processes that produce severe weather and have changed significantly over time as our understanding of these processes has advanced. Through numerical simulations of

severe weather events, special field experiments, and forecaster experience, the knowledge of the specific types of local environments in which severe weather occurs is growing. This knowledge leads to the development of various parameters that are used to assess and refine the severe weather threat.

Whereas pattern recognition focuses primarily upon the synoptic-scale features plotted on a map, parameter evaluation uses the vertical thermodynamic and wind profiles at a point location to provide guidance on the type of severe weather that can be expected. For example, thunderstorms require large values of convective available potential energy to produce updrafts strong enough to form large hail. However, we also know that updraft strength is not the only factor influencing hailstones and that the depth over which the hailstone is falling through temperatures above freezing is also important. A parameter could be designed to combine the value of convective available potential energy with the height of the freezing level into a single number. Tornadic thunderstorms are often associated with environments having very moist boundary layers topped by an inversion and steep mid-level lapse rates ([Figure 7](#)). However, tornadic thunderstorms also often have a low-level veering wind profile in which the scalar product of the velocity and vorticity vectors (called the storm-relative environmental helicity) is large. A parameter could be designed to combine information on the boundary layer moisture content, mid-level lapse rate, and storm-relative environmental helicity into a single number. Candidate parameters are often tested within a large database that contains the vertical environmental profiles associated with various severe weather events to determine if the parameters can discriminate between different types of severe weather. Successful parameters that quantify specific aspects of the vertical thermodynamic and wind profiles known to be associated with a certain type of severe weather event help further refine the severe weather threat region.

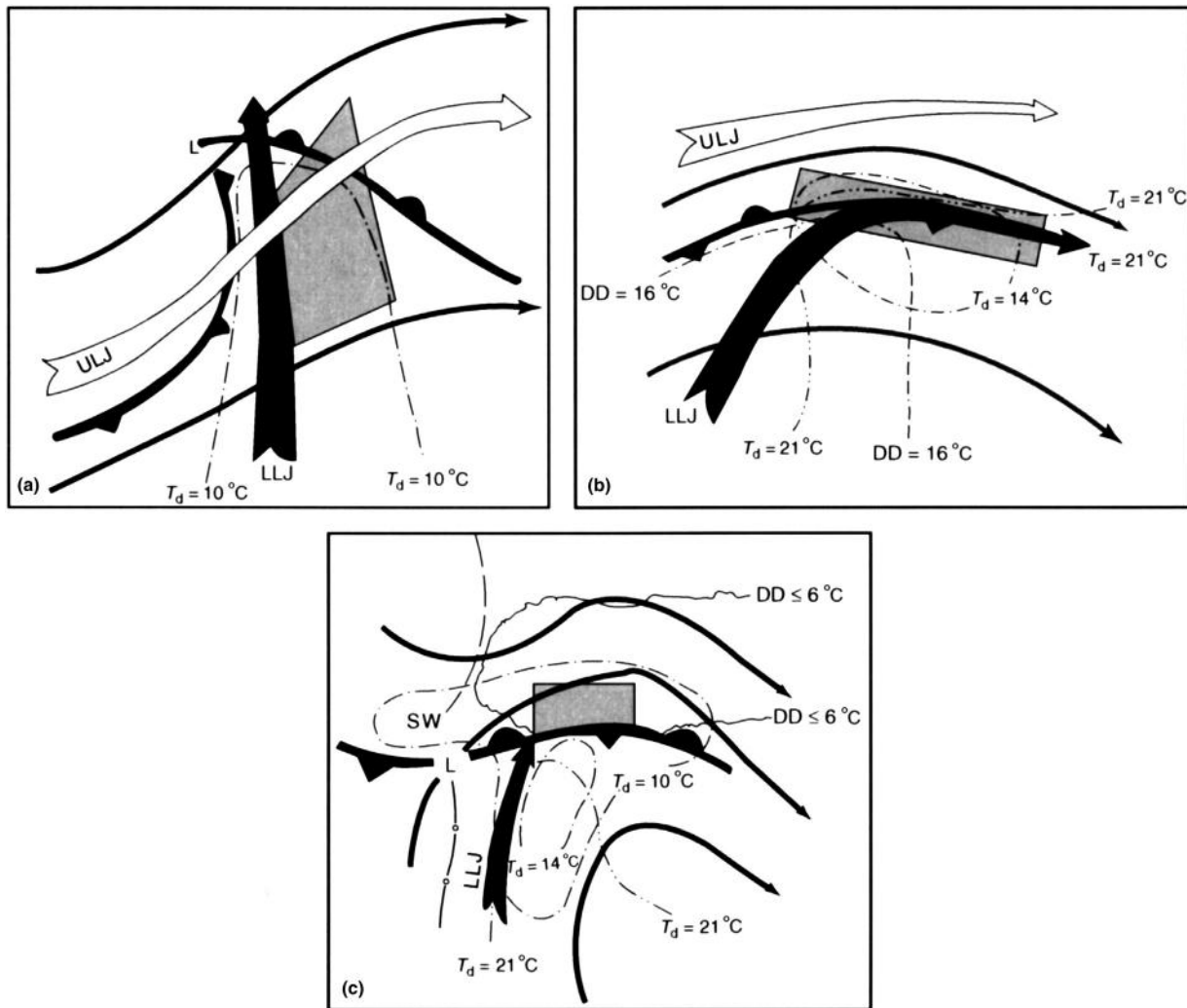


Figure 6 Composite charts for (a) tornadic thunderstorm, (b) derecho, and (c) flash flood days. Dark wind vectors indicate the position of the low-level jet (LLJ); white wind vectors indicate the position of the upper-level jet (ULJ); and smaller dark wind vectors indicate the 500 hPa flow patterns. SW denotes 500 hPa short-wave location, and the wavy line denotes 500 hPa dewpoint temperature depression (DD) isoline in ($^{\circ}\text{C}$). L denotes the surface low-pressure center. Dash-dot line indicates the 850 hPa dewpoint temperature (T_d) isoline; dash-dot-dot line indicates the surface dewpoint temperature isoline; dash-dash-dot line indicates the 700 hPa dewpoint temperature depression isoline; and dash-circle line indicates position of the dryline. Cold and warm frontal boundaries are denoted using standard convention. Polygon-shaped shaded area indicates where the threat of the event is focused. Part (a) after Barnes, S.L., Newton, C.W., 1986. Thunderstorms in the synoptic setting. In: Kessler, E. (Ed.), *Thunderstorm Morphology and Dynamics*, second ed. University of Oklahoma Press, Norman, pp. 75–112. Part (b) after Johns, R.H., Howard, K.H., Maddox, R.A., 1990. Conditions associated with long-lived derechos: an examination of the large-scale environment. Preprints, 16th Conference on Severe Local Storms, Kananaskis Park, Alberta. American Meteorological Society, Boston, pp. 408–412. Part (c) after Maddox, R.A., Chappel, C.F., Hoxit, L.R., 1979. Synoptic and meso- α aspects of flash flood events. *Bulletin of the American Meteorological Society* 60, 115–123.

Parameter evaluation is used to reduce or increase both the areal extent and the magnitude of the severe weather threat within the broad threat region defined from pattern recognition and an awareness of climatology. If one or more parameters are unfavorable for severe weather, then the region of severe weather threat may be reduced or these factors may be discussed in the forecast text. Parameter evaluation can also lead to areas being identified as having a larger threat if many of the parameters are favorable. Through this approach, the threat region is modified and often reduced in size. Unfortunately, a variety of environmental factors influence

the development of severe weather. It is not always clear in advance whether one marginal ingredient (such as instability) can be offset by sufficient vertical wind shear to develop an isolated severe storm or will instead result in more widespread severe weather.

Convection Initiation

Convection initiation is one of the most difficult aspects of severe thunderstorm forecasting. It is not uncommon for all the

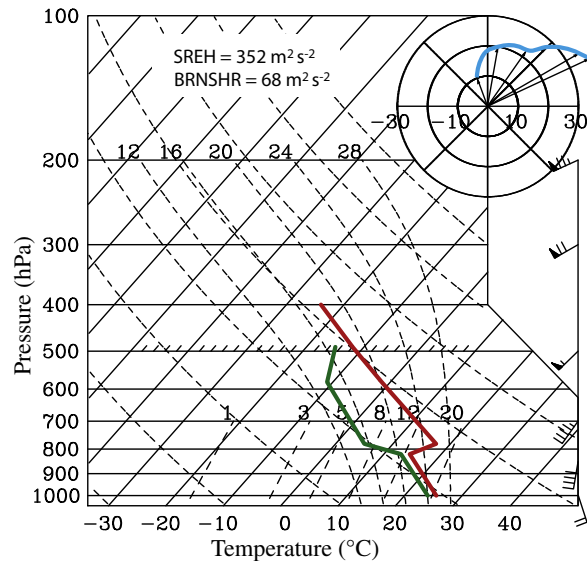


Figure 7 Skew- T log p diagram of mean temperature (red line), dewpoint temperature (green line), and winds from tornado proximity soundings. Full wind barb is 5 m s^{-1} . Hodograph plotted in upper right for winds at surface and at 850, 700, 500, 300, and 200 hPa levels showing significant curvature of the wind profile with increasing height. Values of storm-relative environmental helicity (SREH) and bulk Richardson number shear (BRNSHR) calculated from this wind profile are shown. SREH calculated using a climatological storm motion based upon the mean wind in the 850–200 hPa layer. Thermodynamic sounding after Fawbush, E.J., Miller, R.C., 1953. A method for forecasting hailstone size at the earth's surface. *Bull. Am. Meteorol. Soc.* 34, 235–244; wind profile after Maddox, R.A., 1976. An evaluation of tornado proximity wind and stability data. *Monthly Weather Review* 104, 133–142.

above factors to indicate that the storm environment is favorable for severe weather, but determining whether or not storms will develop is very challenging to forecast accurately. Three elements are necessary for the development of thunderstorms: low-level moisture, steep mid-level lapse rates, and lift. Many of the above analysis techniques evaluate the low-level moisture content and steep mid-level lapse rates, but whether or not a parcel of boundary layer air can reach its level of free convection, above which it is warmer than the environment and rises freely to near the tropopause, is not easy to assess. Even when we believe it is likely that deep convection will develop, specifying the time and location of initiation can prove very elusive. The difficulty of forecasting convection initiation is apparent in both numerical weather prediction model forecasts and human forecasts.

Convection initiation is a small-scale process and is influenced by small-scale features such as boundary layer thermals and rolls, terrain, and vegetation contrasts; by mesoscale features such as cold surface outflows from previous convective activity, drylines, gravity waves, bores, and sea breeze fronts; and by synoptic-scale features such as cold fronts, warm fronts, jet streaks, and upper-tropospheric troughs. A single feature may control convection initiation for one event, while a number of these features may interact synergistically to produce the lift needed to initiate deep

convection for another event. Several features also may interact synergistically to suppress convection initiation or may act in opposing directions, leading to great uncertainty in whether or not convection will develop. It is no wonder that predicting convection initiation is difficult. In one case, strong surface outflows of evaporatively cooled downdraft air from two separate areas of ongoing thunderstorms were observed to collide and a new area of convection was forecast to occur in the region of collision. No new convection developed. Postanalysis with special data indicated that a collision of upper-level anvil outflow also occurred over a large region above the surface, producing sinking motion that was strong enough to inhibit further convection initiation (**Figure 8**). The multiscale processes often associated with convection initiation are beyond our present ability to observe using current observational facilities, while many of these processes are not understood completely. It should be no surprise that accurate forecasts of convection initiation continue to elude even the best forecasters.

The difficulty of forecasting the exact details of convection initiation was one factor that led to the development of the severe local storm watches for both tornadic thunderstorms and severe (nontornadic) thunderstorms in the United States. A severe local storm watch means that severe local storms are likely if storms develop. These watches are usually issued for periods of 4–6 h over a region of 52 000 km^2 and often have a quadrilateral shape. They are intended to focus attention on a smaller region in which the forecaster believes there is a significant threat of severe weather and in which convection initiation is most likely to occur first within a broader region of threat. Watches are an effective way to alert the public to stay abreast of current developments so that they can take appropriate action if necessary.

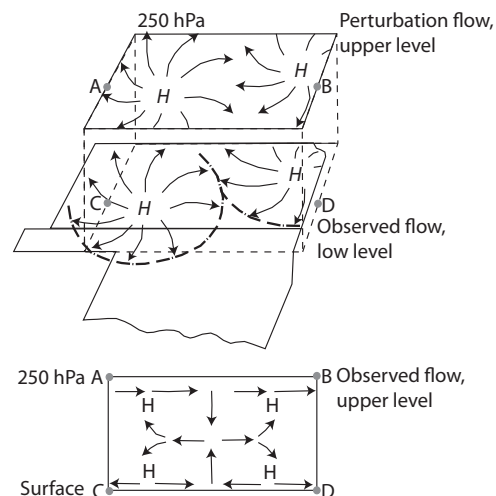


Figure 8 Schematic of hypothesized mesoscale features and circulations showing observed colliding low-level storm outflows and the perturbation upper-level converging anvil outflows. Lower schematic depicts hypothesized vertical motions driven by the convergence of both upper-level and low-level outflows through a vertical cross-section connecting points A, B, C, and D in the upper schematic. After Stensrud, D.J., Maddox, R.A., 1988. Opposing mesoscale circulations: a case study. *Weather Forecasting* 3, 189–204.

Evolutionary Concerns

Once convection develops, or a severe weather event is underway, it is necessary to determine how long the severe weather threat will last, what additional areas will be affected by the event, and how the structure of the developing system will evolve. In many cases, it is not unusual for the type of severe weather threat to change. The initial threat may be from isolated tornadic thunderstorms, but over time these thunderstorms may merge and form a line of convection or a mesoscale convective system. At this later time, the severe threat may be flash flooding or damaging straight-line surface winds. This change in the severe weather threat often requires a different type of response from local officials such as emergency managers and is another important component of severe weather forecasting.

Forecasting thunderstorm evolution after initiation is challenging. A number of factors influence whether or not isolated storms will grow upscale into a convective line and, if they do, how quickly this evolution will occur. These factors include the evolving environmental conditions; the number, separation, and alignment of storms; the strength of thunderstorm cold pools; interactions between cold pools; and the ability of cold pools to initiate new storms along their leading edge. Results from convection-allowing numerical models suggest that while the models can provide useful information on convective mode, predictions of upscale growth are difficult.

Numerical Models and Ensembles

Analyses and forecasts from numerical weather prediction models provide the forecaster with valuable information on the current and possible future states of the atmosphere. Model analyses are important as they combine, in a dynamically consistent manner, all the available observations with a previous model forecast to form a three-dimensional snapshot of the atmosphere at a particular time. Some models produce these analyses at hourly intervals, allowing forecasters to assess the rapid changes in atmospheric structure often associated with severe weather. The ability to predict how the atmosphere will evolve over the next few hours also is very important, because patterns and parameters that affect severe weather potential are rarely static. They continuously undergo change, and small changes in some parameters can lead to either severe weather occurring or no storms at all. Since model-generated analyses and forecasts are available at regular time intervals and over a three-dimensional grid that covers the forecast area of interest, the same techniques used for assessing the severe weather threat with observational data can be used to assess the severe weather threat using model analyses and forecasts. These techniques include pattern recognition and parameter evaluation. Model forecasts also provide guidance on where and when convection will develop, which can help the human forecaster refine the region of threat. However, the added difficulty is that numerical weather prediction models are not perfect and often have biases, so one must account for these model errors in the preparation of the forecasts. An additional challenge arises because known biases and systematic errors often change as numerical models are updated. This

model improvement process requires forecasters to continuously monitor model performance.

Measures of numerical model skill indicate that the operational models in use today are much better at forecasting the synoptic-scale features of the atmosphere than were the numerical models in use 20 years ago. Unfortunately, the skill of numerical models to forecast the small-scale and mesoscale features so important in the forecasting of severe weather has not seen a similar increase. Thus, even though numerical models predict details such as convection initiation and evolution, these smaller-scale fields cannot be treated with the same certainty as the synoptic-scale patterns. Convection is one of the most difficult forecast fields to predict accurately, since it is influenced strongly by all the other physical processes in the numerical model and is sensitive to mesoscale variability in initial conditions that are not resolved by standard observational data networks.

One recent addition to the operational model suite in several countries is a short-range ensemble forecast system. An ensemble is a group of model forecasts valid over the same time period, with each ensemble member starting from slightly different initial conditions. Some ensemble systems also use several different models or different physical process parameterization schemes within the same model. Ensembles allow forecasters to assess in what ways the model predictions contained within the ensemble are the same and in what ways they are different. This assessment provides information that can be used to develop greater confidence in specific outcomes, say when all the model predictions agree, and to determine how alternative outcomes seen in only a few of the predictions would influence the severe weather threat. If these alternative outcomes lead to significant changes in the expectations of severe weather, then the forecaster can determine how best to communicate this information to the public and will carefully monitor the observations to determine if this outcome becomes more likely over time. Ensembles have revolutionized the forecast process as they allow for a richer exploration of possible future atmospheric states and their relation to severe weather, thereby reducing the number of severe events that are not anticipated.

Due to the large computational cost of running an ensemble system, the ensemble model grid spacing is usually larger than in single deterministic forecast models. Thus, forecasters have both an ensemble forecast system and a high-resolution deterministic forecast system available for their use. This situation gives the forecaster the ability to view the high-resolution deterministic forecast from the perspective of the ensemble output and determine how they agree or disagree. This assessment can increase or decrease forecaster confidence in the solutions provided by the single high-resolution forecast. However, high-resolution ensembles in which convection is no longer parameterized and instead occurs on the grid scale are being explored in operations. One such convection-allowing ensemble forecast has been produced experimentally for the National Oceanic and Atmospheric Administration's Hazardous Weather Testbed by the Center for Analysis and Prediction of Storms at the University of Oklahoma. An example from this experimental storm-scale ensemble is shown in a case study. In addition, a storm-scale ensemble of opportunity currently is being

produced in real-time that is comprised of forecasts from seven deterministic convection-allowing model runs. While most of these forecasts are nonoperational and subject to occasional outages, the results from this ensemble are encouraging.

Forecasting Severe Weather Events

The forecaster develops a likely evolution of the severe weather situation through an examination of all the observational and model forecast information using pattern recognition and parameter evaluation techniques as well as past experience. Conceptual models of various severe weather types also are very helpful in this forecast process. Close monitoring of the evolving atmospheric conditions using observations and analyses helps the forecaster determine if any changes to the severe weather forecast are warranted. Forecasters also assess the accuracy of the model forecasts by comparing the model fields with all available observational data. Computer workstations that can display simultaneously a multitude of observations (such as surface and upper-air observations, and radar and satellite imagery) and model analyses and forecasts are very helpful. This nearly continuous situational awareness process, which combines observational assessment with model forecast evaluation, may lead to adjustments in the severe weather forecasts. The forecast process continues as each new piece of information is made available to the forecaster.

It is difficult to illustrate the process of severe weather forecasting without examining an event that shows the complexities one encounters when using observations and numerical model output. Therefore, a severe weather event that illustrates the diversity of the forecast issues is discussed. Since this is a brief analysis of an event that has already happened – so that the ‘answer’ to the forecast problem is already known – the importance of particular synoptic features

and meteorological parameters is typically clearer now than prior to the event.

Tornadoic Thunderstorm Event: 17–18 June 2010

Since strong and violent tornadoes are almost always associated with supercell thunderstorms, a subset of all thunderstorms that have strong, persistent mid-level rotation coincident with the storm updraft, the forecasting of significant tornadoes is very similar to the forecasting of supercell thunderstorms. Environmental factors that are favorable for supercell thunderstorm development are moderate to large values of convective available potential energy and vertical wind shear, often evaluated by calculating the storm-relative environmental helicity. Many studies have demonstrated that these two factors lead to the development of mid-level rotation in thunderstorms, the key characteristic of supercells. It has recently been suggested that the development of low-level rotation in thunderstorms is influenced by the wind shear and moisture near the ground. Thus, once a prediction of supercell thunderstorms is made, forecasters examine the wind shear over the lowest 1 km and cloud base height to refine the tornado threat.

There were reports of 66 tornadoes across the states of Iowa, Minnesota, and North Dakota on 17 June 2010, with numerous additional reports of hail and damaging winds (Figure 9). Climatology indicates that the frequency of tornado reports in Minnesota reaches its yearly maximum in late June, so a forecaster would be very aware of any potential tornado threat on this day. The annual cycle of the hail threat parallels the tornado threat, and damaging winds are also possible even though this threat actually peaks in July. The threat for all three of these severe events is relatively high during late June in Minnesota.

A composite analysis of conditions during the event from 0000 UTC 18 June (Figure 10) shows many similarities to

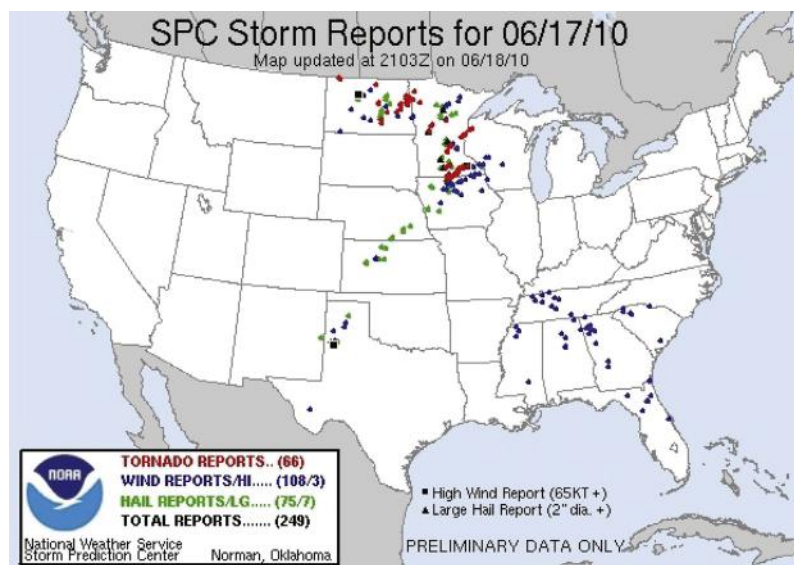


Figure 9 Reports of tornadoes (red), hail (green), and damaging surface winds (blue) during the 24-h period from 1200 UTC 17 June 2010 through 1200 UTC 18 June 2010 over the United States collected by the Storm Prediction Center of the United States National Weather Service. Image courtesy of the Storm Prediction Center.

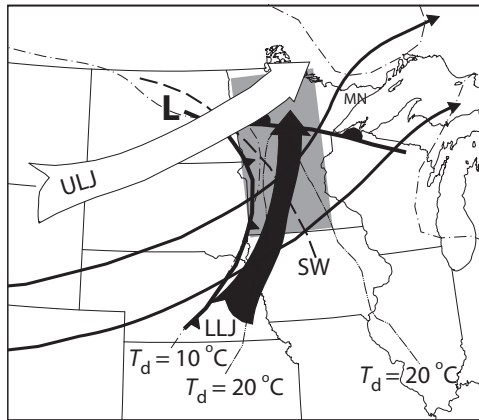


Figure 10 Composite chart for atmospheric conditions over the northern United States valid 0000 UTC 18 June 2010. Dark wind vector indicates the position of the low-level jet (LLJ); white wind vector indicates the position of the upper-level jet (ULJ); and smaller dark wind vectors indicate the 500 hPa flow pattern. SW denotes 500 hPa short-wave location, the dash-dot line denotes 850 hPa dewpoint temperature (T_d) isoline ($^{\circ}\text{C}$), and dash-dot-dot line denotes the surface dewpoint temperature isoline. L denotes the surface low-pressure center, with cold and warm fronts indicated using standard conventions. Polygon-shaped shaded area indicates the region of threat for tornadic supercell thunderstorms. MN indicates the state of Minnesota.

tornado composite analyses. These features include a cold front advancing from the west, deep low-level moisture to the east of the front, a strong southerly low-level jet that advects moisture northwards, and an upper-level southwesterly jet with an associated short-wave trough. Parameter evaluation techniques further support the threat for supercell thunderstorms. Values of convective available potential energy exceed 1000 J kg^{-1} across the threat area, while the storm-relative environmental helicity values exceed $300 \text{ m}^2 \text{ s}^{-2}$ and are well above the threshold values typically associated with

supercell thunderstorms. Now that the threat of supercell thunderstorms has been diagnosed, parameters are examined to refine the tornado threat. Model analyses and observations indicate the wind shear in the lowest 1 km above ground is very strong and the cloud base height is low, suggesting that tornadoes are likely if supercell thunderstorms develop. Thus, the threat for tornadic supercell thunderstorms is supported by both the composite analysis and parameter evaluation.

Numerical weather prediction model forecasts of this event also support the conclusion that supercell thunderstorms are likely. Both composite analysis and parameter evaluations applied to model forecasts valid at 0000 UTC 18 June suggest a threat for tornadic supercell thunderstorms (not shown). However, even more impressive is a 22-h forecast of radar reflectivity from a convection-allowing model run, provided by the NOAA National Severe Storms Laboratory, that compares reasonably well with radar observations at 2200 UTC 17 June 2010 (Figure 11). The forecast shows a combination of isolated storms and more contiguous lines of convection that are located $\sim 50 \text{ km}$ to the west of the observed thunderstorm locations. This size of placement error is very good for a 22-h forecast, while the type of convection produced in the model forecast closely mirrors the reflectivity structures seen in the radar observations. The information provided on the type of convection is particularly useful to forecasters and further supports the idea that isolated thunderstorms are likely on this day.

Further support for supercell thunderstorms on this day comes from a 24-h convection-allowing ensemble forecast valid at 0000 UTC 18 June provided by the Center for Analysis and Prediction of Storms at the University of Oklahoma. Model forecasts of updraft helicity, a field that is used to identify rotating thunderstorms, show high probabilities across Minnesota (Figure 12). This result indicates that supercell thunderstorms are present in many of the ensemble members, thereby reinforcing the conclusions based upon a single model forecast, composite analysis, and parameter evaluation. The

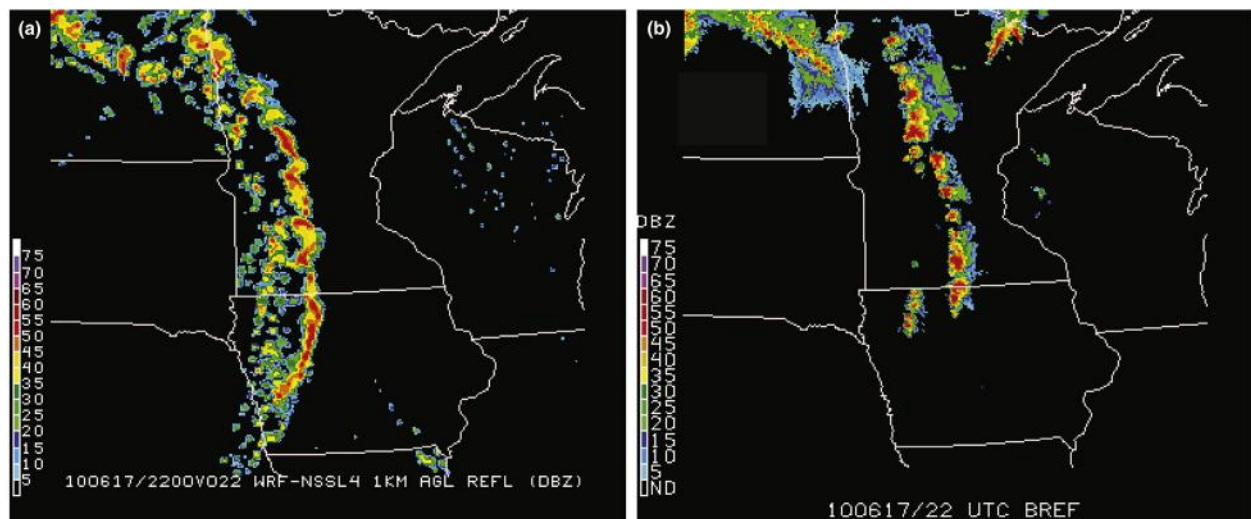


Figure 11 Radar reflectivity (dBZ) at 2200 UTC 17 June 2010 from (a) a 22-h numerical model forecast produced by the National Severe Storms Laboratory and (b) observations from the National Weather Service radars.

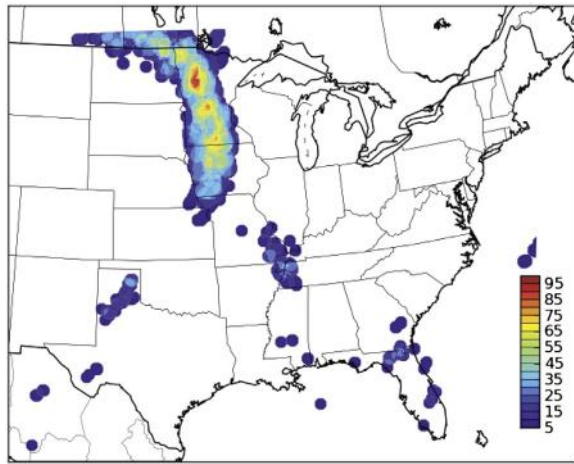


Figure 12 Probability (%) of updraft helicity values exceeding $25 \text{ m}^2 \text{ s}^{-2}$ from a storm-scale ensemble forecast valid 0000 UTC 18 June 2010 produced by the Center for Analysis and Prediction of Storms at the University of Oklahoma. Updraft helicity is a parameter that is used to diagnose rotating (supercell) thunderstorms in numerical weather prediction model forecasts.

consistency in these evaluations of the severe weather threat provides the forecaster with very high confidence that the event will unfold as anticipated.

Future Directions

As computer power continues to increase, it is likely that numerical weather prediction models will be able to produce convection-allowing ensemble forecasts over large regions of the globe that can explicitly resolve individual thunderstorms in great detail and account for imperfections in the model initial conditions and physics. New observational systems, such as geostationary satellites, phased array radars, national surface mesonetworks, and lightning mapping arrays, will improve our ability to sample the atmosphere and construct more accurate analyses of atmospheric conditions. These improvements in both observations and model forecast ensembles will help human forecasters define the severe weather threat and communicate it effectively to the public. The advent of smart phones is already changing how weather information is disseminated to the public, and continued rapid changes in this area are expected.

Some have taken the enhanced observation and numerical modeling capability to mean that the need for human

forecasters will be reduced, particularly since numerical model guidance will have so much detail. However, the atmosphere is a complex system and this complexity may be greatest for severe weather events where nonlinear processes play a significant role in their evolution. Processes like convection initiation will continue to be difficult to forecast accurately by both numerical models and forecasters. Advances in our scientific understanding of the processes that lead to the development of severe weather will continue to alter the parameter evaluation methods mentioned and may even lead to improved composite analyses. Therefore, the need for human forecasters and their ability to assimilate vast amounts of information quickly and incorporate recent scientific advances into their work will remain for the foreseeable future.

See also: **Electricity in the Atmosphere:** Lightning. **Hydrology, Floods and Droughts:** Flooding. **Mesoscale Meteorology:** Convective Storms: Overview; Hail and Hailstorms; Microbursts; Overview; Severe Storms. **Numerical Models:** Regional Prediction Models. **Radar:** Polarimetric Doppler Weather Radar. **Synoptic Meteorology:** Forecasting.

Further Reading

- Bluestein, H.B., 1993. Synoptic-Dynamic Meteorology in Midlatitudes. In: Observations and Theory of Weather Systems, vol. 2. Oxford University Press, New York.
- Browning, K.A. (Ed.), 1982. Nowcasting. Academic Press, New York.
- Cotton, W.R., Anthes, R.A., 1989. Storm and Cloud Dynamics. Academic Press, New York.
- Doswell III, C.A. (Ed.), 2001. Severe Convective Storms. American Meteorological Society, Boston.
- Houze Jr., R.A., 1993. Cloud Dynamics. Academic Press, New York.
- Johns, R.H., Doswell III, C.A., 1992. Severe local storms forecasting. *Weather Forecasting* 7, 588–612.
- Kessler, E. (Ed.), 1983. The Thunderstorm in Human Affairs. University of Oklahoma Press, Norman.
- Kessler, E. (Ed.), 1986. Thunderstorm Morphology and Dynamics. University of Oklahoma Press, Norman.
- Markowski, P.M., Richardson, Y.P., 2010. Mesoscale Meteorology in Midlatitudes. Wiley-Blackwell, Hoboken.
- Pielke, R.A., 1984. Mesoscale Meteorological Modeling. Academic Press, New York.
- Ray, P.S. (Ed.), 1986. Mesoscale Meteorology and Forecasting. American Meteorological Society, Boston.
- Saucier, W.J., 1983. Principles of Meteorological Analysis. Dover, New York.

Relevant Websites

- <http://www.nssl.noaa.gov>.
- <http://www.nssl.noaa.gov/hazard>.
- <http://www.spc.noaa.gov>.

Wildfire Weather

J Coen, National Center for Atmospheric Research, Boulder, CO, USA

© 2015 Elsevier Ltd. All rights reserved.

Synopsis

Weather across timescales influences wildland fires, from microscale temperature and humidity effects that determine whether a fire ignites, to thermal diurnal wind and humidity patterns and topographic flow effects on the wildfire spread rate and direction, to seasonal and interannual weather patterns favoring large fire growth. Dynamic feedback of heat and momentum fluxes between a fire and the atmosphere creates universally observed fire phenomena as fires create their own weather. Fires affect weather and the environment, as fire emissions reduce air quality through increased particulate levels and gases that chemically evolve to degrade regional air quality. Fires contribute to climate, as they release greenhouse gases and particulates that affect cloud microphysics and radiative properties.

Introduction

Weather is one of the most significant factors influencing wildland fires, as weather conditions from the seasonal to the instantaneous play a role in where fire occurs, how it spreads, the intensity of the burning, and whether or not the fire displays extreme and erratic behavior. Wildfires in turn make their mark upon the atmosphere, releasing as much as several megawatts of heat per square meter and releasing smoke composed of aerosol particles, water vapor, and other gases as well as re-releasing pollutants captured by the biomass. The connection between wildfires and their environment is even more complicated, however, because complex dynamical interactions can occur when the atmospheric winds caused by the fire feed back to influence the fire itself. Before examining the complex interplay between a fire and the atmosphere, the basic elements required for a fire will be discussed. Three elements (the 'fire triangle') must be present for a fire to occur and continue. There must be fuel (any material, living or dead, that can burn), oxygen for the fuel, and the fuel condition is a result of the recent weather as well as the year's pattern, especially that of precipitation; atmospheric winds may fan the fire with oxygen-rich air; and heat is often provided for ignition by lightning strikes or transferred to new fuel by wind, for example, by the lofting of burning embers.

Three environmental factors affect wildland fire behavior: weather, fuel characteristics, and topography. Fuel factors include the type, moisture, size, shape, amount, and arrangement. Topography factors include the orientation toward the sun, the slope, and features such as narrow canyons and barriers such as creeks, roads, and unburnable fuel. Of the three environmental factors, weather (including factors such as wind, temperature, relative humidity (RH), and precipitation) is the most rapidly changing. Weather phenomena that bring changes such as cold fronts, foehn winds, thunderstorm downdrafts, sea and land breezes, and diurnal slope winds can be particularly dangerous, as they can suddenly change the fire's direction and behavior. Weather also influences other factors of fuel and topography, by controlling the fuel moisture through precipitation, RH, and winds, and by complicating the fire-accelerating effect of slopes with topographically induced flows. Additionally, though weather can influence wildfire behavior through many

pathways, it has long been recognized that intense fires 'create their own weather.' That is, the heat and moisture created by the fire feed back into the atmosphere, creating intense winds that drive the fire's behavior, sometimes overwhelming the effect of ambient winds. Research studies have only begun to investigate this dynamic interaction between wildfires and the atmosphere.

Effects of Weather on Fires

Fire is the release of energy as heat and light when, through a chemical reaction, oxygen combines with a combustible (burnable) substance at a sufficiently high temperature, involving both heat transfer and fluid dynamics. Weather affects fires on scales ranging from persistent, seasonal patterns related to global circulation patterns down to small-scale, rapidly changing conditions in the fire's environment by altering the fluid dynamics of the air in which the fire occurs and factors that control the combustion of wildland fuels.

Overall, human-caused fires are the most prevalent sources of fires around the world, often set to clear land for agriculture. However, in the United States, about 80% of fires are caused by natural sources, nearly all by lightning strikes. When lightning strikes or an active firebrand lands on fuel, ignition is not certain. The probability of ignition (the 'lightning ignition efficiency') depends on the duration of the lightning strike or contact, fuel moisture and temperature, as well as other non-weather-related elements such as fuel depth.

Wildland fuel consists of ground material such as duff and forest litter, as well as needles, leaves, twigs, and larger woody materials of trees. Wood is made up of cellulose (cell walls) and hemicellulose (which is readily pyrolyzed), lignin (which mainly forms char), extractives (primarily hydrocarbons), lipids (a ready source of combustible volatiles), and ash content (which exerts a suppressing effect). Solid wildland fuels such as wood do not burn directly; ignition has several stages. During a pre-ignition phase, fuel must be heated by an outside source and its temperature rises toward the ignition temperature (about 390 °C for fuels composed mainly of cellulose); thus, the ignitability depends on the fuel's initial temperature, as well as its thermal properties. This ignition temperature is

well above the boiling point of water, so all the water present in the fuel must also be vaporized before ignition will occur. At higher temperatures (200–340 °C), fuel undergoes pyrolysis, the thermal decomposition into an envelope of flammable gases, semivolatile tar, and a solid char. These volatile gases (volatile organic compounds (VOCs)) burn in flaming combustion in a narrow envelope of flame around the fuel-rich zone only where the oxygen (diffusing inward from free air) and fuel (diffusing outward) are mixed in certain proportions above the ignition temperature to produce rapid oxidative chemical reactions that emit visible light called a flame. Some VOCs volatilized ahead of the flame front escape the combustion process and are found unaltered in the smoke plumes. Ignition is the transition from pre-ignition to combustion, the temperature at which external heating is no longer required. Once ignition is reached, the heat generated by combustion brings other fuel to ignition, continuing the process. Finer fuels such as grass carry the flame front (Figure 1); the speed at which this interface between burning and unburned fuel travels

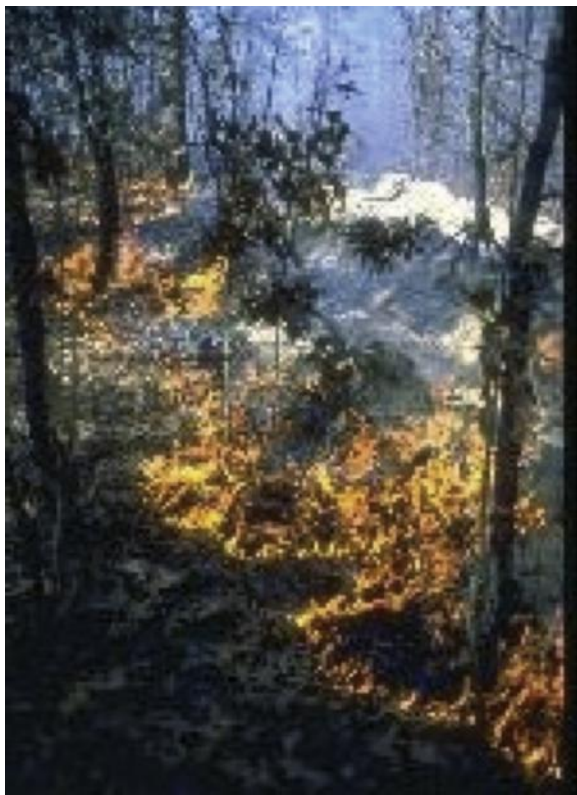


Figure 1 The fire process. Fuels are usually a mixture of dry fine and heavy fuel. Heat preheats the fine fuel. Moisture is boiled off. Tar appears as visible smoke. A cloud of combustible gases is formed. Heat ignites the combustible gases. Flames from burning fine fuel particles preheat larger fuel particles. Gases from larger fuel particles ignite. The fine fuel begins to burn by glowing combustion. Ash is formed. Pre-ignition, ignition, and flaming and glowing combustion are occurring in different parts of the fire. Smoke particles and CO result from incomplete combustion. Wood collapses because the applied heat weakens cellulose. Gray mineral ash coats the fuel surface. Ash must be knocked off to prevent smothering. The fire is extinguished. Most of the cellulose fuel has reacted with oxygen to form CO₂ and water. Photograph courtesy of the U.S.D.A. Forest Service.

is called the fire spread rate. However, combustion may not always involve a flame. After flaming combustion has consumed most of the volatile substances, the remaining solid carbon may burn by smoldering (surface oxidation, also called glowing combustion).

The thoroughness with which combustion consumes the fuel (the ‘combustion efficiency’) varies. Most laboratory studies of gaseous combustion have been carried out with well-mixed (‘premixed’) fuel–air mixtures, but in natural fires, mixing of fuel and air is an integral part of the burning process – these flames are known as diffusion flames. Consequently, combustion is less efficient, burning only where there is a sufficient mixture of fuel and oxygen. In a free-burning fire in the atmosphere, enough air is usually entrained into the flame to burn all the combustible gases; however, some products of incomplete combustion survive and are released to the atmosphere. If combustion is not complete, some of the volatile products remain as very small, suspended drops of fluid, along with minute carbonaceous particles formed in the flame under conditions of low oxygen and high temperature. These make up the particulate portion of smoke that reduces visibility. Carbon monoxide (CO) is also a product of incomplete combustion, and the ratio of CO to total carbon content including CO₂ (a product of complete combustion), indicates the degree of complete combustion. Extinction is the termination of combustion and occurs when one of the three elements of the fire triangle is removed (e.g., there is no more fuel, the oxygen supply is cut off, or the heat required to bring fuel to ignition is removed by dousing with water).

Small-Scale Effects of Temperature and RH on Combustion

Atmospheric temperature, RH, and wind directly affect combustion rates or the fluid dynamics of the air in which combustion occurs. Higher fuel and ground temperatures make it easier for fire both to ignite and to burn, because less heat must be added and less time is required to bring the fuel’s temperature to the ignition temperature. This is one reason why fire activity usually peaks in the late afternoon. Solar radiation can also add to the fuel temperature – in some climates, there can be a 25 °C difference in fuel temperatures between sunny and shaded spots. Cloud cover or dense smoke in the fire’s own smoke plume can shade the fire and reduce fuel temperatures.

A fire’s potential is very sensitive to the dead fuel moisture, defined as the percentage of mass of water to the mass of dry fuel. The equilibrium moisture content is the value the moisture content would approach if the fuel were exposed to constant temperature and RH for an infinite time. However, atmospheric conditions surrounding fuel seldom remain the same for long, and different fuels can be classified according to the time it takes them to adjust to changes in the atmosphere. The fuel moisture in dead fuels (fuels that are no longer part of a living plant or tree) is determined by ambient RH, temperature, and precipitation, and responds to changes in its environment with a time lag proportional to the fuel’s diameter; this time lag is loosely defined as the time it takes a fuel particle to reach two-thirds of the way to equilibrium with its local environment. The moisture content of fine fuels such as twigs, needles, and grass adjusts quickly to ambient conditions, with a timescale of minutes to an hour. Larger diameter fuels take longer to adjust

to ambient moisture conditions; for example, logs that are 15 cm in diameter have an average time lag of about 36 days. (For clarity, this time lag is generally expressed in hours, and fuels are often characterized as having 10-, 100-, or 1000-h fuel moistures.) For this reason, large precipitation rates over a short time will not raise fuel moisture as much as lesser rainfall rates over a longer time. Wood that may ignite easily at fuel moistures of a few percent may smolder at fuel moistures of 13%, for example, and not burn at all if the fuel moisture rises to 15–20%. Thus, lightning strikes in relatively cool, moister evenings can cause fire starts that smolder overnight (or even for weeks) before becoming active when conditions for fires improve, as daytime temperatures rise and humidity drops, and fuels dry to the point where flaming spread is possible. In the western United States during the summer, dry lightning storms – lightning storms in which negligible precipitation reaches the ground – commonly ignite many fires. These storms, characterized by deep, dry adiabatic layers and midtropospheric moisture, support high-based thunderstorms that nevertheless produce many lightning strikes.

Wind, either ambient or induced by the fire itself, encourages combustion and fire spread by drying the fuel (by increasing the evaporation rate), by ventilating the fire by increasing the supply of oxygen, by influencing the direction of fire spread (generally directing it toward fresh fuel), and by tilting the flame forward and closer to unburned fuel ahead of the fire that may be warmed and dried by radiant heat, a process very dependent on proximity to the flame. The relative importance of these convective and radiative effects is still unclear.

Atmospheric stability is recognized as an indicator of fire behavior, as fires purportedly burn less intensely when air is stably stratified. A physical basis for this consensus is that a stable atmosphere is averse to vertical motions and thus limits the updraft strength over a fire (limiting the convergence at its base, limiting the winds ventilating the fire), while a super-adiabatic surface layer would enhance any perturbation caused by a fire. However, this may also partially be attributed to atmospheric conditions that accompany stable atmospheric stratifications, an example of which is seen during inversions. An atmospheric inversion is a layer in the atmosphere where the atmospheric temperature is steady or increases with altitude. Inversions often form at night near the surface and are common during calm, settled weather. They are an important factor in fire behavior, because when an inversion starts to lift or break, fire behavior can change abruptly, as winds pick up, surface temperatures rise, and RH drops. Similar to inversions, other atmospheric temperature structures called thermal belts play into fire behavior. Thermal belts of relatively warm air may form midway upslopes in mountainous areas. The top of the night inversion is usually below the main ridges. The height of the warmest air temperature is at the top of the inversion. From the top of the inversion, the temperature decreases as one goes farther up or down the slope. This region of warmer air (with accompanying low RH) is called the thermal belt. Within the thermal belt, wildland fires can remain active throughout the night.

Subsidence, the large-scale sinking of air often associated with high-pressure systems, can also play a role in fire behavior. As air from higher elevations in high-pressure systems descends

to lower elevations, it warms and dries. Although the air is stable, subsidence can lead to increased fire activity as this warmer and drier air sinks to the surface.

A negative vertical wind shear – that is, winds that decrease with height – has long been cited as a contributing factor to intensification of fires, but the dynamics through which this may occur have not been explained.

Indicators of the potential severity of wildfires can be designed by combining known factors. For example, the widely used Haines Index (HI) is calculated by combining the vertical temperature difference between the 70- and 50-kPa atmosphere levels (as a proxy for lapse rate) and the temperature and dew point difference at 70 kPa (indicating the dryness) to produce a number that indicates the potential for fire growth. The stability and moisture terms are each assigned values between 1 and 3 that, when added together, result in an index that varies from 2 to 6. An HI of 2 or 3 indicates a relatively moist, stable air mass with a very low potential for large wildfire growth, while a value of 6 indicates a dry, unstable air mass with high potential for large growth and extreme behavior of wildfires. HI values of 4 and 5 suggest low and moderate potential, respectively. Even though the effects of wind are not factored into the HI, the index has been shown to correlate well with rapid fire growth.

Diurnal Variability and Changing Weather Patterns

The daily diurnal cycle in temperature and RH, characterized approximately by a sinusoidal pattern with a maximum RH and minimum temperature in early morning, is a crucial window of opportunity in fire suppression. This nighttime recovery in temperature and RH is the norm, especially in high, dry regions, where nighttime temperatures can drop as much as 25 °C as radiative cooling leads to a cool, stable nocturnal boundary layer. This period is known as the ‘nocturnal lay-down,’ as fires subside at night in low temperature, high RH conditions, offering an opportunity for suppression. Exceptions can occur if there is nighttime cloud cover, which warms the surface through radiative heating, or winds that dissipate the stable layer and encourage vertical mixing.

Changes in weather conditions can lead to unpredictable fire behavior, as they can suddenly change the fire spread rate, direction, and fire intensity. Such changes include the following:

- Pressure gradient winds related to synoptic-scale weather patterns.
- Cold fronts, which bring erratic, gusty winds, and wind shifts after the frontal passage.
- Regional downslope winds, such as foehn winds in the Rockies and Santa Ana winds in California. These strong, dry winds may blow for days with gusts up to 50 m s⁻¹, exacerbating fire behavior while decreasing fuel moisture.
- Thunderstorms that can cause strong downdrafts, gusty surface outflows of 30 m s⁻¹ (higher for microbursts), and lightning.
- Marine-related features such as sea and land breezes that affect the winds, humidities, and temperatures in coastal areas.
- The breakup of a temperature inversion.

- Diurnal slope winds. In mountainous areas, heating of the slopes by solar radiation during the day leads to upslope winds of several meters per second during the afternoon, and cool drainage winds that move downslope during the night and early morning.

In summary, fires benefit from sustained high temperatures, low RH (single digits), dry fuels, strong winds, no RH recovery overnight, dry lightning, passing dry thunderstorms, and gusty winds. These are conditions that often coincide, resulting in extreme behavior such as spotting, crowning, torching of trees, uphill 'runs,' rolling debris, and sudden flare-ups that push fires across firebreaks and rapidly increase their size. Fires are hindered by afternoon cloud cover, moisture, still air, and sustained rain. In the United States, if weather conditions are forecast to include hot, dry days combined with high winds that could result in extreme fire behavior, or if extensive fire starts within the next 24 h are likely (particularly in the worst-case scenario of developing thunderstorms with dry, gusty winds and dry lightning), a fire weather forecaster issues a 'Red Flag Warning.' A Fire Weather Watch is issued when Red Flag conditions are expected within the following 72 h.

Seasonal and Interannual Weather Effects on Fire

Patterns and variability in weather on longer timescales affect wildfire activity primarily by affecting fuel moisture – prolonged periods of hot, dry weather can leave vegetation cured like tinder. The role of climate is most apparent in the regionally synchronized outbreaks of wildfires in particular years that have occurred for centuries. Broad regions of fire-enhancing weather are often associated with the biggest resource losses due to demand for suppression simultaneously in many spots. For example, 6 out of the 10 largest fires of 1999 in the United States were located in northern Nevada, 5 of which began over a 3-day period. Even though the connection of seasonal climate variability to wildland fires appears intuitively obvious – that droughts are correlated with wildland fires – the predictive nature of this relationship is poorly understood. For example, in some ecosystems, the drought–fire relationship also involves time lags with previous moisture conditions. In the Southwest United States, when an extremely dry winter–spring follows a wet spell (when the growth and accumulation of fuels is encouraged), exceptionally large areas burn the following summer. In some areas of the Northwest United States and Nevada, the fire risk at the season's peak is only weakly correlated with current climate conditions, and is more strongly linked to conditions 10–18 months before.

The interannual variability arising from the El Niño/La Niña episodes of the Southern Oscillation (SO) phenomenon is often cited as a factor in weak/severe wildfire seasons, respectively, for the United States (the reverse for Indonesia and Australia), but usefulness as a predictive tool is limited to certain regions. Interdecadal changes in fire–climate relationships in the Southwest United States parallel evidence of shifts in the frequency or amplitude of the SO over the past three centuries, where high correlations between fire and climate signals reflect periods of high amplitude in the SO and rapid

switching from extreme wet to dry years. Weak correlations are associated with a decrease in SO frequency or amplitude.

A research challenge is to separate the climate impacts on wildfire areal extent from anthropogenic effects that confound the climate relations, such as how decades of fire suppression and associated fuel buildups have compounded the effect of drought, and how some land use policies have allowed flammable non-native species to intrude into shrublands.

Dynamic Interactions between a Fire and the Environment

A fire is a continually propagating phenomenon and spreads by igniting new fuels along its outer perimeter. The leading part of the fire line is called the 'head' and is the fastest spreading. The slowest-spreading part is called the backing region, where the fire creeps against the wind. The fire spread rate is measured from a point on the fire line in a direction perpendicular to the perimeter. The spread rates can range over 3 orders of magnitude, from a barely creeping fire to about 6.7 m s^{-1} (the equivalent of a 4-min mile running pace). Fires spread most rapidly in the direction of the local wind and, in the absence of other factors, upslope on terrain. Slope and wind increase the spread rate in a similar way, by decreasing the angle between the flames (and the strong convective updraft in the fire circulation) and the fuels ahead of the fire. This could cause an increase in the flame spread rate by the increased radiant flux preheating and drying fuel ahead of the flames or repeated contact by flames; thus, the effect is particularly effective if the fire is spreading up a steep slope.

Wind enhances the spread of a fire front in a similar way. This effect increases approximately exponentially with wind speed up to a critical wind speed, at which extinction, or blowout, occurs. Although blowout is not believed to occur naturally in wildland fires, it is the basis of some fire suppression strategies, and works by distorting the reaction zone within a flame by stretching it thin so that fuel vapors have a much shorter period in which to react. If the reaction zone is too thin, combustion is incomplete and the flame is effectively cooled ultimately below the ignition temperature where it cannot be sustained.

Fire can also grow by igniting fuel outside its perimeter by spotting, in which burning embers are carried distances from inches up to many kilometers from the fire either by the ambient wind or by the fire's convection column before being dropped. This mechanism is particularly prevalent in Australian fuel complexes, where eucalypt bark provides copious numbers of flaming brands that can land up to 30 km ahead of the main fire.

Examples of Atmosphere–Fire Coupling

Fires sometimes exhibit extreme fire behavior, when the fire behaves in ways that preclude direct fire suppression. Examples of this are high spread rates, crowning (when the fire climbs into the tree canopies, where it travels rapidly), spotting, the presence of fire whirls (vortices of fire often seen along fire lines), a strong convection column indicative of intense burning, torching of trees (a sudden ignition of an entire tree),

and tall flames (sometimes 65 m above tree height). In these situations, it is difficult to predict what a fire may do next, because the fires often behave erratically and sometimes dangerously. Research suggests that the cause of extreme behavior is an interaction of the fire with its atmospheric environment. Numerous examples of interesting behavior result from this type of interaction.

Fire whirls

The intense buoyancy released into the atmosphere by a fire causes strong convective updrafts that create a strong convergence into its base, creating intense winds near the surface that drive the fire's behavior, sometimes overwhelming the winds in the fire's environment. Vorticity is produced by these buoyancy gradients, which is tilted and stretched in the updraft column. Fire whirls, which resemble dust devils, are often generated by fires (Figure 2) and can vary from common small whirls up to 5 m tall filled with flame to giant fire vortices approximately 0.5 km in diameter and several thousand meters tall, in which the entire smoke-filled convection column rotates. They can lift large burning embers and loft them far across the fire line, starting spot fires. They can be intense, and have been known to even break off trees that are more than 1 m in diameter. They may have a vorticity on the order of $1\text{--}10\text{ s}^{-1}$, though their radii are usually small and thus their rotational speeds do not reach the speeds within tornados. Fire whirls frequently occur where flow eddies occur, such as on the lee side of a fire, particularly near the outside edges of a fire front or when a fire is burning on the lee side of a ridge.

Forward bursts

A crown fire is one that spreads through the tree canopy, with or without a surface fire beneath. Flame lengths over 50 m along with spread rates over 3.5 m s^{-1} have been observed. If conditions for sustained spread through the crowns are not favorable, the trees will burn out, but a running crown fire can



Figure 2 A fire whirl is a spinning, moving column of ascending air containing flames that lifts smoke and debris. When seen along fire lines, whirls apparently range in diameter from tens of centimeters to small tornados in size and intensity, and are documented as having uprooted trees. Photograph from the International Crown Fire Modeling Experiment, 1997, courtesy of the Canadian Forest Service.

result when winds increase and the flames from one burning tree canopy are driven into adjacent trees. Crown fires can produce dramatic forward bursts of flame (Figure 3), allowing them to jump across firebreaks or rush uphill. Analysis of such an event captured in infrared imagery (Figure 4) by a research



Figure 3 Forward bursts from a crown fire prescribed burn during the International Crown Fire Modeling Experiment, 1997. Courtesy of the Canadian Forest Service.

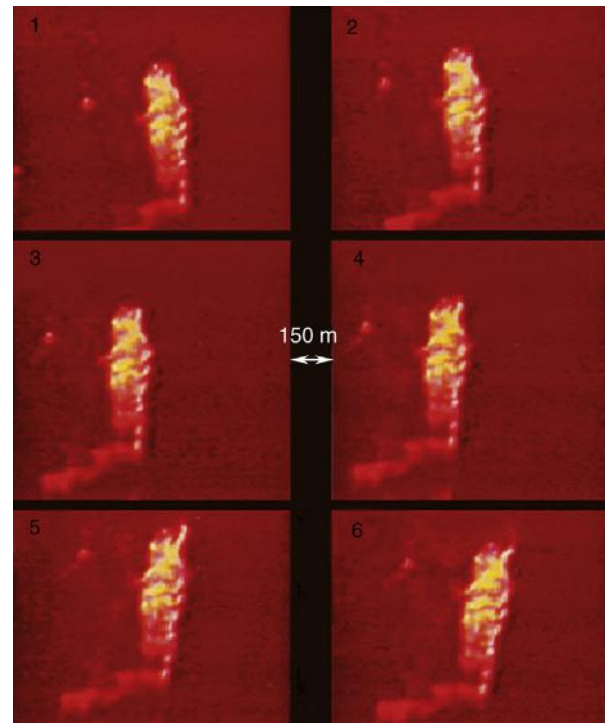


Figure 4 A sequence of infrared images looking down on a crown fire climbing a slope (toward the upper right) from a research aircraft. Colors represent temperature, where yellow is hottest, and darker reds are cooler. The six images were taken 0.75 s apart. Note the hot yellow finger that shoots out 100 m suddenly in 1–2 s (frames 3–6); it then disappears. From Radke, L.F., Clark, T.L., Coen, J.L., et al., 2000. The wildfire experiment (WiFE): observations with airborne remote sensors. *Canadian Journal of Remote Sensing* 26, 406–417.

aircraft flying overhead showed a finger of flame burst forward 100 m at an estimated speed of 50 m s^{-1} and then dissipate within 2 s, leaving burning embers in its path. This type of explosive behavior results from an interaction between the atmospheric winds and the vorticity and winds produced by the fire.

Horizontal roll vortices and tree crown streets

After a fire has passed, crown fires are often seen to have left distinctive, sometimes unusual, patterns of burned and unburned trees. One example is 'streets,' where long strips of unburned tree vegetation are left amidst otherwise blackened areas; in one case one side of a row of trees was scorched, and the other side had few or no burn marks, indicating outward airflow from within the unburned street. These patterns have not been explained completely. It has been hypothesized that downward and outward flow of relatively cool air from horizontal roll vortices, perhaps from a tilting convection column, kept fire out of the crowns, and as it spreads outward in both directions near the ground, it caused the fire to leave these burned/unburned patterns. Sometimes the burn patterns are far more complex (Figure 5) with streets and swirl patterns, and may involve a sequence of formation and collapse of such rotating columns along the enlarging fire. Other explanations suggest that patterns such as this are the result of fluctuating wind speeds or directions.

Convective fingers

Fire lines moving across the landscape do not normally move in a straight line when driven by moderate ambient winds (about $3\text{--}4 \text{ m s}^{-1}$) but bow forward into parabolic fingers. The parabolic shape can be understood by examining how the convergence pattern produced under the fire's updraft is projected upon a straight fire line (Figure 6). In the absence of an ambient wind, a fire would create a vertically oriented



Figure 5 Burn scars ('fire streets') in yellow, unburned fuel in green, left after a fire burned through a North Carolina forest. Photograph courtesy of U.S.D.A. Forest Service.

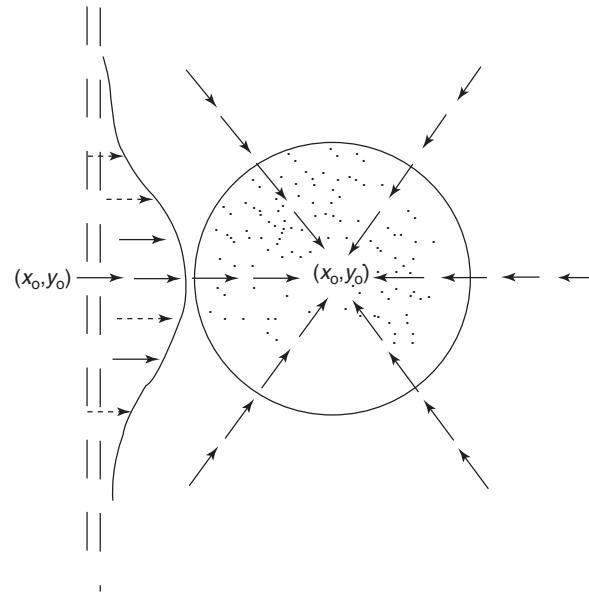


Figure 6 The idealized structure of the near-surface wind convergence pattern beneath the convective updraft (stippled) of a fire line (driven by winds from the left, moving to the right). The convergence beneath the convective updraft draws the fire line forward in a bow shape. From Clark, T.L., Jenkins, M.A., Coen, J., Packham, D., 1996. A coupled atmospheric-fire model: convective feedback on fire-line dynamics. *Journal of Applied Meteorology* 35, 875–901. Copyright American Meteorological Society.

convective cell centered over itself, and draw air equally from all sides. However, in an ambient wind, shown in Figure 6 as coming from the left, assuming the convective column is tied to the fire at the surface and tilts downstream with height, the effect of downstream tilting is to shift the center of the low-level convergence pattern ahead of the fire front, drawing air at low levels (levels important to driving a fire) from different angles along the fire line, forming a curved inflow region at the fire front. Within a range of wind speeds, the stronger the wind, the stronger the tilt, and the farther forward the center of convergence lies. Stronger ambient winds may cause the convective column to break away and reform, complicating these interactions. Longer fire lines may form many protrusions, or 'fingers' along their length, as in the Onion fire, shown in Figure 7. The Onion fire produced fingers about 1 km wide apart only during times when an ambient wind was present. When the mean wind died down, the fire burned out the fuel between the fingers, once again forming a more linear and stalled smoldering fire line.

Blowups

A blowup is a sudden intensification and acceleration of burning during a fire. The burn rate may increase 500% and flame lengths can stretch 65 m above the trees. The convective updraft strengthens, pulling in surrounding air and further ventilating the fire. These firestorms create their own wind patterns, which have been strong enough to uproot trees and loft embers to start spot fires miles away. There is no definitive explanation of the cause of blowups, although they typically occur during the afternoon, with strong environmental wind



Figure 7 Convective fingers (spaced about 1 km apart) during the Onion fire. Photograph courtesy of Charles George, U.S.D.A. Forest Service. From Clark, T.L., Jenkins, M.A., Coen, J., Packham, D., 1996. A coupled atmospheric-fire model: convective feedback on fire-line dynamics. *Journal of Applied Meteorology* 35, 875–901. Copyright American Meteorological Society.

and fire-generated buoyancy-induced flow effects, and often in steep terrain. Current hypotheses suggest that interactions between combustion chemistry and fluid dynamics may be an essential part of blowup events, involving sudden mixing of fresh air with excess pyrolysis gases produced by radiative and convective heating within and ahead of the main flame zone.

Wildfire Impacts on Weather and the Environment

As well as responding to its atmospheric environment and changes in the weather, fire exerts its own impact. Wildland fires release heat ($1.74 \times 10^7 \text{ J kg}^{-1}$ of dry cellulose fuel burned), particulates (primarily composed of carbon), water vapor, carbon dioxide (CO_2), and an extraordinary spectrum of other gases including carbon monoxide (CO), hydrocarbons (HC), nitrogen oxides (NO_x), and sulfur oxides (SO_x). The released particulates are composed of both inorganic 'black' carbon and organic carbon (including fatty acids and toxic polycyclic aromatic hydrocarbons). The rate, composition, and emission factor for each chemical constituent (the mass release of each particular species per unit mass of fuel consumed) depends on the phase of combustion (whether the early, high-intensity flaming period, or a later low-intensity smoldering period), but these combustion characteristics themselves are affected by the plant species amount of biomass present, the size distribution of the fuel, precipitation history at the site, ambient atmospheric conditions, and other factors.

The smoke from fires can sometimes be traced thousands of kilometers in satellite images (see [Figure 8](#)). In 1998, smoke from thousands of Mexican fires reached the Southern United States. In other cases, pollutants from burns in Alberta, Canada, have been tracked to the Southeastern United States. Thus, the effect of wildland fires on the atmosphere, whether air pollution, impacts on cloud droplet distributions, or as a source of greenhouse gases, is not just a short-term local problem, but has regional or global impacts.

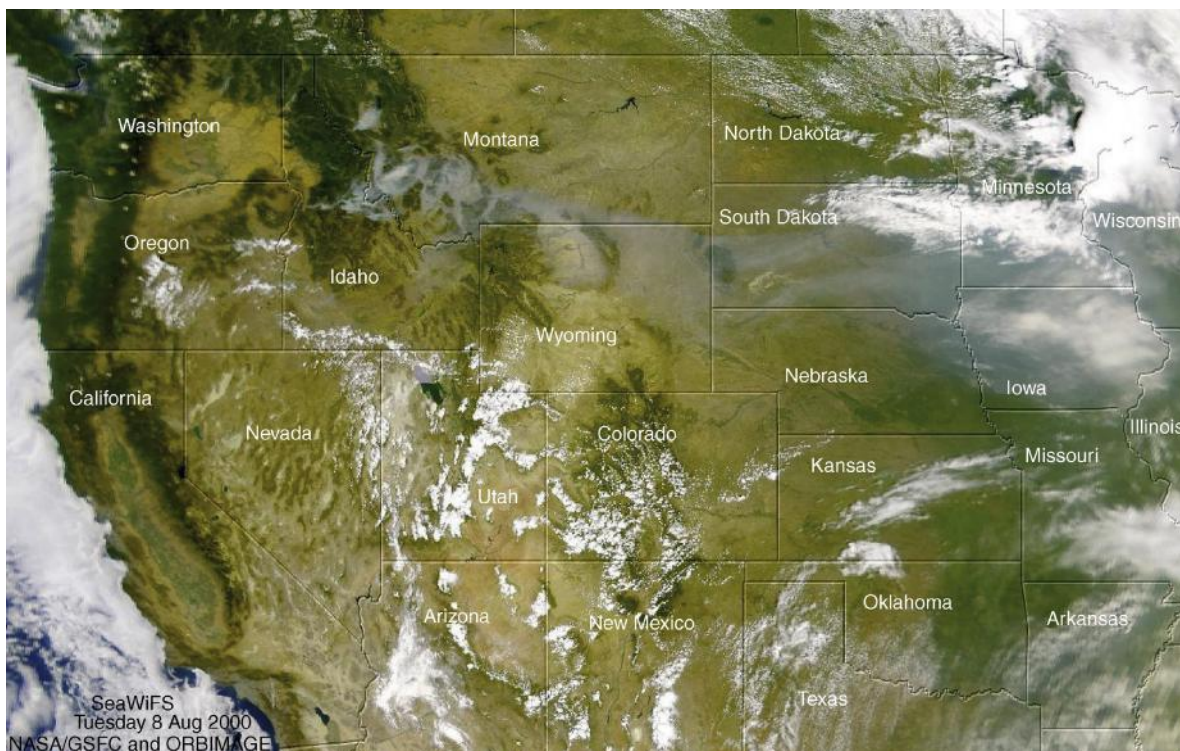


Figure 8 A SeaWiFS (Sea-viewing Wide Field-of-view Sensor) satellite image of the smoke from wildfires in Montana and Idaho on 8 August 2000. The smoke can be identified at least as far east as Illinois. Image provided by ORBIMAGE Orbital Imaging Corporation and processed by NASA Goddard Space Flight Center.

Emissions Components

Carbonaceous aerosol particles

Wildland fires vary widely in intensity (heat released per unit time) and severity (heat released per unit area). Shortly after ignition, a flaming phase continues for several minutes during which time the maximum release rates of heat, particles, and gases occur. Rate of fuel consumption is high, while the duration is very short. After flaming, some woody materials begin smoldering, with smoldering emissions decreasing exponentially with time. (In places with deep layers of duff or rotten wood, smoldering can continue for several hours to days, although in some cases fires that entrench in duff can smolder for months, depending on the depth and moisture of the duff and amount of rotten material, until snow falls.) The particle size distribution also varies between fires, with the most intense fires releasing the highest proportion of large particles and the most buoyant plumes, consuming 70–95% of available fuel. Many fires, however, only smolder, are of low intensity, burn a much smaller fraction of the available biomass, and emit nearly all fine particles in only slightly buoyant plumes. Biomass fires produce smoke particles in a bimodal distribution, with one peak in the 0.1–1.0 μm diameter range that spans the wavelengths of visible light (0.4–0.7 μm), and another in large particles over 10 μm in diameter.

Pollutants, nutrients, greenhouse gases, and VOCs

Wildland fires emit a variety of chemical gases, depending on fuel type, fuel chemistry, and fire behavior. These include long-lived greenhouse gases, e.g., CO_2 or CH_4 , as well as short-lived chemicals such as carbon monoxide (CO), nitrogen oxides (NO_x), hydrocarbons ($\text{C}_1\text{--C}_{12}$), partially oxidized hydrocarbons ((POHC) alcohols, aldehydes, ketones, and furans), halocarbons (CH_3X), sulfur compounds (carbonyl sulfide, dimethyl sulfide, dimethyl disulfide), and alkyl nitrates. Some of these emissions have important effects on the atmosphere: CO , non-methane hydrocarbons, and POHC influence the oxidative capacity of the atmosphere, and in the presence of NO_x lead to photochemical ozone formation on local and regional scales. Alkyl nitrates serving as a reservoir for NO_x might cause additional ozone formation in especially warm and pristine environments. Halocarbons, methyl chloride, and methyl bromide, released from fires are estimated to be 25% of the global budget, and are precursors to stratospheric ozone destruction. COS released by wildfires is a minor part of the global budget but is a precursor for sulfuric acid in stratosphere.

Mercury is another species released by wildfires. About half the atmospheric mercury comes from natural sources (in soil, oceans, and volcanoes), and the other half from human activity (including industrial pollution such as from coal-fired plants). Mercury is removed from the atmosphere through oxidation, creating ionic mercury in the troposphere and in clouds, where it then rains or dry deposits out to the surface. Contrary to the belief that this mercury is ‘lost,’ it may be taken up in the biomass and, when the biomass is consumed by fire, 95–99% of the mercury is emitted, 95% as gaseous mercury, and 5% as particulates. Studies estimate that mercury released by wildland fires may represent 25% of all anthropogenic sources.

Emission products including mercury can be lofted by the ‘smokestack effect’ deep into the atmosphere, where winds are

generally stronger, increasing the distance the emission products travel from their source. Intense wildland crown fires can produce heat fluxes on the order of 10 MW m^{-2} and a well-defined, strong updraft (50 m s^{-1} has been measured) over surface convergence. Aided by the water vapor released by combustion, intense fires can rapidly produce deep ‘pyrocumulus’ clouds over the fire (Figure 9), sometimes reaching the tropopause, with smoke trails visible from satellite flowing over 2000 km (Figure 8). These clouds can produce rain (thereby scavenging some of the smoke), hail, lightning, and strong surface winds such as microbursts, complicating fire behavior at the ground. As the fire burns less intensely and does not reach the critical temperature required to release flammable gases and sustain flaming combustion, it produces less heat, the air over the fire is less buoyant, smoke is not lofted as high, and it settles in valleys, where it can be trapped for days, improving only as it mixes with cleaner air.

Effect of Aerosol Particles on Cloud Microphysical and Radiative Transfer Properties

An important aspect of an extremely large portion of the aerosol particles released by fires is their ability to serve as cloud condensation nuclei (CCN) for droplets that evolve to form haze, fog, and clouds. When a fire is present, the concentrations released by fires can easily overwhelm the background concentration by factors of 2–50. *In situ* measurements of smoke-modified clouds in Mexico and Indonesia showed the cloud droplet concentrations were more than double those of unaffected clouds, increasing the droplet concentrations from 400 cm^{-3} to over 1000. Globally, they are estimated to be as plentiful as the other primary CCN sources – anthropogenic processes and the oceans. These particles increase cloud droplet



Figure 9 Smoke plume from a California wildfire. Photograph courtesy of the National Interagency Fire Center media library.

concentration, reducing cloud droplet size and narrowing the cloud droplet size distribution, and thus coalescence rate between droplets (the 'Twomey effect'), and perhaps the cloud's precipitation efficiency. This effect of reducing the average cloud particle radius also has been shown to increase the cloud's albedo, altering its radiative properties. This leads to the possibility of regional feedback, as widespread fires release particles that might in turn decrease the precipitation efficiency of clouds, exacerbating the dry conditions that favored the fires.

Hydrology and Land Surface Effects

Intense fires have had severe hydrological consequences, through their ability to influence the physical and chemical properties of soil, the ground flora, water-holding properties, and the capacity of soil to resist erosion. The combination of combustion and heat transfer during wildland fires produces steep temperature gradients in the surface layers of soil, as flame temperatures of 700–800 °C can occur. Heat produced by combustion of the material on the soil surface vaporizes organic substances, which are then moved downward in the soil until they reach cooler underlying soil layers, where they condense. After a fire, a layer of water-repellent soil resembling a waxy film may be found on the soil surface or a few centimeters below the burned area. Fire severity, behavior, soil properties, and temperature gradients developing in the soil during a fire all affect the formation of a water-repellent layer.

Watersheds damaged by intense fires can suffer increased runoff, greater erosion, and greater sedimentation rates, resulting in mudslides, silted-out reservoirs and streams, and the potential for flooding. The longevity of these effects from weak fires and prescribed burns is generally limited to less than a year, but water repellency caused by a severe fire burning through large fuel accumulations can extend over several years. Current research is investigating whether fires can change the surface albedo and other properties enough to affect mesoscale land surface processes and change precipitation patterns.

See also: **Aerosols:** Aerosol–Cloud Interactions and Their Radiative Forcing; Role in Radiative Transfer. **Boundary Layer (Atmospheric) and Air Pollution:** Complex Terrain. **Chemistry of**

the Atmosphere: Principles of Chemical Change. **Clouds and Fog:** Cloud Microphysics. **Dynamical Meteorology:** Static Stability. **Hydrology, Floods and Droughts:** Modeling and Prediction. Inadvertant Weather Modification. **Mesoscale Meteorology:** Overview. **Mountain Meteorology:** Downslope Winds; Valley Winds. **Numerical Models:** Methods. **Radiation Transfer in the Atmosphere:** Absorption and Thermal Emission. **Tropical Meteorology and Climate:** El Niño and the Southern Oscillation: Observation. **Tropospheric Chemistry and Composition:** Aerosols/Particles; Mercury; Volatile Organic Compounds Overview: Anthropogenic.

Further Reading

- Drysdale, D., 1998. *An Introduction to Fire Dynamics*, second ed. Wiley, Chichester.
- Johnson, E.A., Miyanishi, K. (Eds.), 2001. *Forest Fires: Behavior and Ecological Effects*. Academic Press, New York.
- Levine, J.S., 1991. *Global Biomass Burning*. MIT Press, Cambridge, MA.
- Pyne, S.J., Andrews, P.L., Laven, R.D., 1996. *Introduction to Wildland Fire*. Wiley, New York.
- Roberts, C.F., 1974. Weather and forest fires. *Weatherwise* 27, 100–105, 115.
- Swetnam, T.W., Betancourt, J.L., 1998. Mesoscale disturbance and ecological response to decadal climatic variability in the American southwest. *Journal of Climate* 11, 3128–3147.
- Williams, F.A., 1982. Urban and wildland fire phenomenology. *Progress in Energy and Combustion Science* 8, 317–354, 406–417.

Relevant Websites

- <http://www.epa.gov/region09/naturalevents/wildfire/index.html> – Environmental Protection Agency, Region 9: Natural Disasters and Weather Emergencies.
- <http://www.globalfiredata.org/index.html> – Global Fire Emissions Database.
- http://www.nasa.gov/mission_pages/fires/main/index.html – NASA Fire and Smoke.
- http://www.nifc.gov/preved/comm_guide/wildfire/ – National Interagency Fire Center, Communicator's Guide for Wildland Fire Management.

Inadvertent Weather Modification

SA Changnon, University of Illinois, IL, USA

© 2003 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 6, pp 2533–2537, © 2003, Elsevier Ltd.

Introduction

Inadvertent weather modification refers to any changes in the weather resulting from human actions done for purposes other than changing the weather. Human activities have accidentally been able to change all facets of the weather at and near the Earth's surface. For example, major cities change every facet of their weather, including the number of thunderstorms; large irrigated areas create clouds and rainfall when conditions are right; and jet aircraft often create cirrus clouds. The list of accidental weather changes is almost endless. However, many such changes in the atmosphere from human changes in the biosphere have been relatively localized, and their effects typically create weather changes occurring on scales ranging from 1 km² up to 10 000 km².

Knowledge that humans were altering the weather became commonplace about 200 years ago as residents of many of Europe's larger cities recognized that urban areas had poorer visibility due to the increase in haze and smoke and higher temperatures than did their surrounding rural areas. Serious studies of urban effects on weather developed in Europe in the nineteenth century and intensified as networks of surface weather stations yielded data that allowed scientists to define the patterns of temperature, humidity, and precipitation in and around cities. The industrial revolution began in the nineteenth century and led to many more weather changes. For example, by 1880 scientists in England discovered that more rainfall occurred on work days, Monday through Friday, than on nonwork days (Saturday and Sunday), a result correctly assessed as an effect of industrial pollutants.

The growth of meteorological knowledge about the atmosphere's behavior, which expanded rapidly during the twentieth century, brought forward investigations of how other land use changes by humans and the emissions from factories and vehicles altered the weather. Progress in detecting and understanding inadvertent weather modification came rapidly in the latter half of the twentieth century as new remote sensors, including satellites and radars, made it much easier to detect and measure conditions. Major advances in meteorological knowledge now allow use of numerical modeling to estimate urban heat islands and other weather anomalies. Interest in and concern over inadvertent weather changes grew as the recognition of the widespread and significant magnitude of these changes developed among the citizenry and public officials. The weather changes accidentally wrought by human activities translate into climate changes when the weather change becomes frequent and significantly different from the climate values expected in a natural, unaffected environment. **Figure 1** shows the average summer pattern of thunderstorms around St. Louis, MO, revealing a localized increase from about 16 storm days per summer in the region's rural areas to a peak of 28 days in and east of the city. Studies of individual summer days revealed those days when the weather conditions were

changed sufficiently by St. Louis to initiate thunderstorms. For example, in one summer 7 added storm days occurred over the city, whereas 16 added storm days were created in the following summer, revealing the large year-to-year variability that exists in most inadvertent weather changes. Many inadvertent weather changes are intermittent and depend on the type of natural weather conditions existing.

Atmospheric Processes Changed

Changes that humans have made to the biosphere that translate into effects on various atmospheric processes include

- surface color (albedo);
- surface topography (roughness) and orientation;
- subsurface materials (soils) including their thermal conductivity, thermal capacity, and hydraulic conductivity;
- artificial sources and sinks of heat, moisture, and mass;
- heat released naturally or artificially at the ground;
- amount of water and dust;
- biota.

Basically, the above changes result from two major human activities: changes in the existing natural land use, and emissions of gases and particles as a result of combustion. Acting individually or in concert, these characteristics affect certain basic processes of exchange between the biosphere and the atmosphere, which, in turn are themselves interactive. These exchange processes include

- the sensible heat budget;
- the moisture budget;
- momentum exchange;
- mass (other than water) exchange;
- atmospheric electrical exchange properties.

Alterations in these exchange processes affect a wide variety of weather conditions. Those commonly altered are at the surface and in the low levels of the atmosphere and include

- radiation;
- temperature;
- humidity;
- wind speed and direction;
- visibility (fog and haze).

These changes are further associated with changes in conditions extending farther aloft in the atmosphere including clouds, precipitation, and storminess. Microphysical processes in clouds are changed by urban and industrial emissions of various particles, including soluble aerosols and ice nuclei, as well as by additional moisture and enhanced vertical air motions. Emissions of some trace gases, such as sulfur dioxide from the burning of coal, can also lead to new particle formation and altered cloud properties. In some instances,

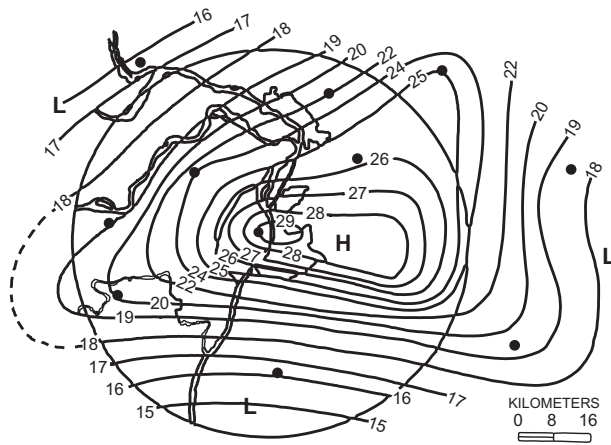


Figure 1 Pattern based on the average number of summer thunderstorms in and around the St. Louis, MO, urban area.

emissions are injected into the upper atmosphere by high-flying aircraft.

Effects of Changed Land Use Conditions

The description of what has been found about inadvertent weather modification is organized around land use types and emissions.

Urban Areas

A major means whereby humans have affected weather has been through the creation of urban environments. In the thirteenth century, residents of London first noted that the city had temperatures and fogs different from those in surrounding rural areas. The industrial revolution and associated rapid growth of cities that began in the nineteenth century resulted in significant local- and regional-scale changes in weather both in and near urban locales. Even small towns of a few thousand population create an effect on temperatures and have an 'urban heat island,' and large cities like New York create much higher temperatures ranging from 3 to 7 °C above rural values.

Cities are major converters of energy and matter, and their products affect the heat and moisture budgets as well as the exchange of mass into the atmosphere. Hence, urban areas alter all forms of weather. The list of changes includes increased contaminants in the air, less solar radiation, higher temperatures, decreased visibility, lower humidity, altered wind speed and direction, greater cloudiness, added precipitation, altered atmospheric electricity, added severe weather events, and changes to mesoscale synoptic weather conditions including fronts. The degree of change in any of these elements at any given time depends upon (1) the areal extent of the urban complex, (2) its industrial types, (3) its juxtaposition to large water bodies and topographic features such as mountain ranges, (4) the time of day, (5) the type of natural background weather system present at any given time, (6) the season of the year, and (7) the climate. Urban areas often act as mechanical barriers to low-level airflow and can even retard the action of

fronts. Temporal differences produced by cities can be illustrated by solar radiation, which is decreased by urban aerosols, and the decrease is much greater in winter than in summer, is decreased more on weekdays than weekends, and is decreased more in the morning than in the afternoon.

Figure 2 shows the long-term frequencies of smoke and haze days as measured at several US Midwestern cities of varying sizes. These visibility-related data are quantitative surrogates of the effects of growing cities and are the only type of pollutant-related data systematically collected since the start of the twentieth century. Chicago, the largest of the cities considered, was the first to show an increase and one that rapidly became sizable; smaller industrial cities like Moline, Springfield, and Peoria also showed increases beginning later in the 1930s and 1940s. The only small nonindustrial city, Cairo, has no apparent increase.

Weather changes resulting from urbanization are summarized in **Table 1**. Changes that are easy to measure include visibility, temperature, and winds. Visibility in and downwind of a major city is usually diminished. The most radical change produced by a city is in the heat balance. In-city temperatures are generally higher than nearby rural values, the difference being greatest at night, but in the early morning the urban temperatures are sometimes slightly lower than rural values. Heat-island effects have been measured in hundreds of cities of widely varying sizes and climates, and can now be numerically modeled based on the urban size and shape. The daytime urban-generated heating can extend vertically up to several thousand meters above a large city, and in some situations a heat plume extends downwind of the city. Urban areas also act as physical obstacles to low-level winds, decreasing their speeds in the city and altering wind directions near the surface. When low wind speeds are present, a rural-to-urban circulation pattern often develops.

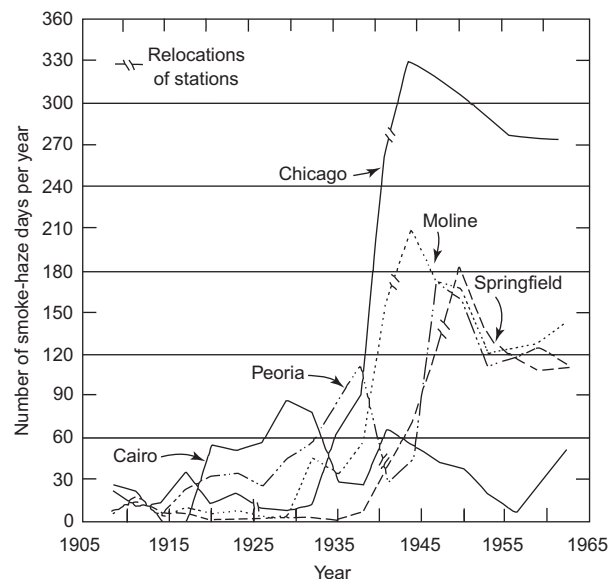


Figure 2 Temporal distribution of the annual number of days with smoke and/or haze affecting visibility at selected US Midwestern cities of varying sizes.

Table 1 Average weather changes resulting from urbanization, as measured at many cities in varying climatic zones

	<i>Annual</i>	<i>Cold season</i>	<i>Warm season</i>
Contaminants (volume)	+1000%	+2000%	+500%
Solar radiation (Langleys)	−22%	−34%	−20%
Temperature (°C)	+2	+3	+1
Humidity (relative)	−6%	−3%	−8%
Visibility (frequency)	−26%	−34%	−17%
Fog (frequency)	+60%	+100%	+30%
Wind speed	−25%	−20%	−30%
Cloud frequency	+8%	+5%	+10%
Rainfall amount	+14%	+13%	+15%
Snowfall amount	+10%	+10%	−
Thunderstorm frequency	+16%	+5%	+30%

Urban weather changes that are more difficult to measure include alterations in humidity, fogginess, cloudiness, precipitation, snowfall, solar radiation, atmospheric electricity, and severe weather. Humidities over cities are generally decreased because there is less evapotranspiration from urban areas than from rural areas owing to the presence of hard surfaces and the lack of plants. However, certain urban-located industrial activities release large volumes of moisture from combustion and cooling devices.

Various urban–industrial complexes affect clouds, precipitation, and related severe weather conditions. These effects are difficult to measure and often extend 5–50 km beyond the city. In general, only large cities, typically those with populations exceeding one million, create localized changes in clouds and related precipitation processes. A major field study centered on St. Louis in the 1970s defined how a city alters clouds and precipitation. Clouds and the formation of precipitation depend on vertical motions in the atmosphere, the presence of adequate moisture, and small particles that serve as cloud and raindrop nuclei. St. Louis often deflected the low-level airflow, resulting in a convergence zone in the lee of the city, a condition that enhanced vertical motions, as did the heat being released by the city. The city was also a source of various aerosols, including cloud condensation nuclei essential to cloud droplet formation, and certain industrial emissions were found to add freezing nuclei, a key to raindrop formation. Cities are complex rainmakers. In the warm season, the enhanced convection also led to more thunderstorms and hail in and east of the city.

Industrial Complexes

Some industrial complexes are isolated away from cities, and certain such rurally located industrial centers that produce large releases of moisture, heat, and particles have been found to create measurable localized weather changes. For example, power generating plants, largely through the release of moisture and latent heat from cooling lakes or towers, have been found to create frequent fogs, icing, and clouds, and in some instances, rain and snow showers. These effects are most common in winter. Large fires from natural gas fields have been noted to create clouds and sometimes rain.

Nonagricultural Rural Areas

This land use category embraces large portions of the Earth's land surface and includes deserts, swamps, grasslands, forests, and high-altitude barren areas. Deserts have been modified by overgrazing, warfare, and overcultivation, and these activities have loosened the surface, causing soil deflation. The resulting dust added to the atmosphere alters the regional radiation budget. Deforestation leads to increases in solar radiation at the surface, altering surface temperatures and evapotranspiration. Deforestation in eastern Europe was found to have led to less rainfall, and in more tropical areas deforestation has led to greater temperature extremes and altered rainfall. Human alteration of natural grasslands, largely by converting them to pasture lands, has led to overgrazing and some dramatic changes in weather, including less cloudiness and decreases in rainfall. The drainage of wetlands and swamps has produced small-scale effects on surface conditions, including higher surface temperatures.

Agricultural Areas

All forms of agricultural land use affect the weather, including cultivated fields, pastures, orchards, and planned forests. The burning of agricultural residues in certain areas of the world adds heat and particles, and in turn these reduce visibility, reduce incoming solar radiation, change clouds, and, in some instances, decrease rainfall from the addition of too many cloud condensation nuclei. Irrigation over large areas changes the surface albedo and adds moisture to the atmosphere. In some climate zones such as the High Plains of the United States, large-scale irrigation has increased clouds and rainfall in and beyond the irrigated area. Different crops have different transpiration rates, leading to soil moisture differences, and when different crops are adjacent and planted over large areas the resulting moisture differences affect the boundary-layer structure. This can alter mesoscale wind conditions and can lead to convection, clouds, and rainfall. Other weather changes from agricultural activities include the use of hedgerows, which decrease wind speeds, and crop spraying, which adds micro-sized particles capable of affecting cloud formation.

Marine Areas

The creation of large reservoirs in certain drier climates has led to increased clouds under certain conditions. The potential for purposefully modifying the weather in dry areas using reservoirs has been explored and modeled to ascertain the possible effects on clouds and rainfall.

Transportation Corridors

The exhaust from high-flying aircraft often leads to contrails, which in high-frequency flight corridors leads to shields of cirrus clouds covering thousands of square kilometers. In parts of the central United States, contrail-generated cirrus has increased the number of cloudy days by 30%. These added clouds have also decreased surface daytime temperatures. Heat and particulate emissions from

vehicles on high-density transportation corridors sometimes lead to fogs and decreased visibility along and adjacent to the corridors.

See also: **Aviation Meteorology:** Aircraft Emissions. **Boundary Layer (Atmospheric) and Air Pollution:** Microclimate. **Climate and Climate Change:** Overview. **Clouds and Fog:** Fog. **Hydrology, Floods and Droughts:** Deserts and Desertification. **Land-Atmosphere Interactions:** Overview.

Further Reading

- Bornstein, R.D., Oke, T.R., 1981. Influence of pollution on urban climatology. *Advances in Environmental Science and Engineering* 2, 171–202.
- Cermak, J.A., Davenport, A., Plate, E., Viegas, D., 1993. *Wind Climate in Cities*. Kluwer Academic, Dordrecht.
- Changnon, S.A., Semonin, R., Auer, A., Braham, R., Hales, J., 1981. *Metromex: A Review and Summary*. American Meteorological Society, Boston, MA.
- Cotton, W.R., Pielke, R.A., 1995. *Human Impacts on Weather and Climate*. Cambridge University Press, New York.

This page intentionally left blank